Компьютерная лингвистика и интеллектуальные технологии

по материалам ежегодной Международной конференции «Диалог 2009»

Выпуск 8 (15)

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference "Dialogue 2009"

Issue 8 (15)

УДК 80/81; 004
ББК 81.1
К63

Программный комитет конференции выражает искреннюю благодарность
Российскому фонду фундаментальных исследований за финансовую поддержку,
грант № 09-01-06015-г

Редакционная коллегия: *А. Е. Кибрик* (главный редактор), *В. И. Беликов, Б. В. Добров, Д. О. Добровольский, Л. М. Захаров, И. М. Зацман, Л. Л. Иомдин, И. М. Кобозева* (ответственный секретарь), *Е. Б. Козеренко, М. А. Кронгауз, Н. И. Лауфер, Н. В. Лукашевич, А. С. Нариньяни* (зам. гл. редактора), *Г. С. Осипов, Н. В. Перцов, Т. В. Черниговская, И. В. Сегалович, В. П. Селегей*

К63 **Компьютерная** лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). — М.: РГГУ, 2009. — XII, 620 с.

ISBN 978-5-7281-1102-3

Сборник включает 93 доклада международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2009», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

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

УДК 80/81; 004
ББК 81.1

© Коллектив авторов, 2009

© Институт проблем информатики РАН, 2009

© Российский государственный гуманитарный университет, 2009

ISBN 978-5-7281-1102-3

Предисловие

Восьмой выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит материалы 15-й Международной конференции «Диалог»

В программу конференции было отобрано 93 доклада, охватывающие все традиционные направления конференции:

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

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

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

По традиции Программный комитет предложил участникам тему — доминанту конференции. В этом году такой темой стало «Создание и применение компьютерных лексических ресурсов». Таким образом, в фокусе внимания «Диалога 2009» находятся проблемы создания и использования лексических баз данных разного типа: от компьютерных словарей и тезаурусов до лексико-семантических ресурсов типа WordNet, FrameNet и типологических баз данных.

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

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

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

Организаторы

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

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

При поддержке Российской ассоциации искусственного интеллекта

Международный программный комитет

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

Организационный комитет

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

Секретариат

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

Рецензенты

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

Содержание

Алиев Р. М., Янь Цзинбинь, Хейдоров И. Э.
Поиск ключевых слов с использованием решетки слогов 1

Апресян В. Ю.
Семантические источники уступительности 6

Баглей С. Г., Антонов А. В., Мешков В. С., Суханов А. В.
Статистические распределения слов в русскоязычной текстовой коллекции 13

Баранов А. Н.
О некоторых семантических коррелятах формального варьирования идиом (операция замены) 19

Беликов В. И., Ахметова М. В.
Статистическая оценка функциональных свойств лексики по материалам интернета 25

Богатырёв М. Ю., Тюхтин В. В.
Построение концептуальных графов как элементов семантической разметки текстов 31

Богданова Н. В., Асиновский А. С., Русакова М. В., Рыко А. И., Степанова С. Б., Шерстинова Т. Ю.
Звуковой корпус как способ мониторинга и фиксации разных форм естественного языка 38

Большаков И. А.
КроссЛексика — большой электронный словарь сочетаний и смысловых связей русских слов 45

Бугаков О. В.
Создание семантического словаря предложных конструкций на основе Украинского национального лингвистического корпуса 51

Васильев В. Г.
Выделение фрагментов в текстах при классификации 57

Воскресенский А. Л., Гуленко И. Е., Хахалин Г. К.
Словарь RuSLED как инструмент семантических исследований 64

Гольдин В. Е., Мартьянов А. О., Сдобнова А. П.
Электронный русский ассоциативный словарь школьников 69

Григорьян Е. Л.
О характере синтаксической полисемии 75

Гришина Е. А.
К вопросу о соотношении слова и жеста (вокальный жест О в устной речи) 80

Диконов В. Г., Богуславский И. М.
Универсальный словарь концептов 91

Добровольский Д. О., Левонтина И. Б.
Русское *нет*, немецкое *nein*, английское *no*: сопоставительное исследование семантики на базе параллельных корпусов 97

Ермаков А. Е., Плешко В. В.
Обработка естественно-языковых запросов к поисковой машине на основе их лингвистического анализа 102

Зализняк Анна А.
О понятии семантического перехода 107

Занегина Н. Н.
<strike>Я этого не говорил</strike>: о литуративах, зачеркиваниях или мнимых текстах 112

Захарова И. В., Городечный П. П.
Об одном подходе к автоматическому построению онтологии для задач анализа текстов 116

Зобнин А. И., Сахарова А. В.
Универсальная система синтаксической разметки текстов ObjectATE 120

Иомдин Б. Л.
Терминология быта. Поиски нормы 127

Иомдин Л. Л., Лобанов Б. М.
Синтаксические корреляты просодически маркированных элементов предложения и их роль в задачах синтеза речи по тексту 136

Кибрик А. А., Кодзасов С. В., Худякова М. В.
Просодическая транскрипция: уровни детализации 143

Кибрик А. Е.
К проблеме вариативности языка: метод многофакторного исчисления второго порядка 149

Клышинский Э. С.
Некоторые сложности автоматизированной лемматизации несловарных словоформ 165

Кобзарева Т. Ю.
Синтаксическая несовместимость как свойство линейной организации русского предложения 170

Кобозева И. М.
Семантика глагола понимать: от пропозиционального отношения к межличностному 176

Кодзасов С. В., Архипов А. В., Захаров Л. М., Кривнова О. Ф.
База данных «интонация русских повествовательных текстов» 181

Кожунова О. С.
Выявление номинализованных конструкций в параллельных текстах патентных документов на русском и немецком языках 185

Скобки в русских идиомах 192

Козеренко А. Д.
Parentheses in Russian idioms 192

Комарова А. Д.
Методика корпусных исследований паузации на примере изучения паузации в японском языке в контексте послелогов и топика 197

Коротаев Н. А.
Отсутствие пауз на границах элементарных дискурсивных единиц: опыт корпусного исследования 204

Котов А. А.
Паттерны эмоциональных коммуникативных реакций: проблемы создания корпуса и перенос на компьютерных агентов 211

Котта Рамузино П.
Непосредственный институциональный диалог. Опыт прямой линии с президентом В. В. Путиным. Дискурсивные стратегии 219

Крейдлин Г. Е.
Невербальное поведение людей разных культур в диалоге I: финская и русская жестовые системы 224

Кретов А. А., Рафаева А. В.
Программа семантической классификации лексики — ПроСеКа: теоретические и прикладные аспекты 230

Крылов С. А.
Квазикорпусное изучение словарной продуктивности лексико-синтаксических разрядов в русском языке по словарю С. И. Ожегова 236

Крылова Т. В.
Прилагательные со значением высокой и низкой температуры и наивно-языковая оценка температуры 243

Кудинов М. С., Гришина Е. А.
Инструменты полуавтоматической разметки для Мультимедийного русского корпуса (МУРКО) 248

Кузнецов И. П., Ефимов Д. А.
Средства настройки процессора Semantix на предметную область 262

Кустова Г. И.
Электронный семантический словарь глагольных прилагательных: структура и типы информации 271

Ландэ Д. В., Жигало В. В.
Подход к созданию многоязычных параллельных корпусов веб-публикаций 278

Лебедев А. С.
Редактор расширенных сетей переходов с графическим интерфейсом пользователя 284

Лобанов Б. М.,
Проблема разрешения «Ё»-омографов при синтезе речи по тексту 291

Лукашевич Н. В., Добров Б. В.
Автоматическое аннотирование новостных кластеров на основе тематического представления 299

Ляшевская О. Н., Кузнецова Ю. Л.
Русский фреймнет: к задаче создания корпусного словаря конструкций 306

Махова А. А., Ляшевская О. Н., Десятова А. В.
Части тела с точки зрения топологии: корпусное исследование 313

Митрофанова О. А., Захаров В. П.
Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике 321

Муталов Р. О.
Опыт создания корпусов дагестанских языков 329

Недолужко А.
Разметка кореференции на синтаксически аннотированном корпусе чешских текстов 332

Николаева Ю. В.
Сегментация устного нарратива и изобразительные жесты: кинетические признаки границ и связей между сегментами дискурса 340

Окатьев В. В., Ерехинская Т. Н., Скатов Д. С.
Модели и методы учета пунктуации при синтаксическом анализе предложения русского языка 346

Орлова С. В.
Перевод немецкой частицы *doch* на русский язык (в контексте констативов): *ведь, же, всё же* или *всё-таки*? 352

Остапова И. В.
Лексикографическая структура этимологического словаря и его представление в цифровой среде 359

Падучева Е. В.
Посессивы и имена способа действия 365

Пазельская А. Г.
Модели деривации и синтаксическая позиция отглагольных существительных по корпусным данным 373

Палько М. Л.
Просодия обращений в немецком языке в сопоставлении с русским 379

Переверзева С. И.
Невербальный коммуникативный акт утешения: материалы к построению словаря невербальных коммуникативных актов 384

Подлесская В. И., Кибрик А. А.
Дискурсивные маркеры в структуре устного рассказа: опыт корпусного исследования 390

Поляков А. Е., Бергельсон М. Б., Пильщиков И. А.
Конкорданс к текстам Ломоносова — концепция и реализация 396

Потапов М. В.
Синтактически инвариантный метод идентификации семантики информации 405

Потемкин С. Б.
Неконтролируемый синтаксический анализ 409

Продан А. И., Корольков Е. А., Опарин И. В., Таланов А. О.
Особенности использования многоуровневой разметки звукового корпуса *unit selection* в системе гибридного синтеза «Живой голос» 415

Рахилина Е. В., Карпова О. С., Резникова Т. И.
Модели семантической деривации многозначных качественных прилагательных: метафора, метонимия и их взаимодействие 420

Розина Р. И.
Так называемый: семантика вводных метаязыковых оборотов 426

Романов А. С., Мещеряков Р. В.
Идентификация автора текста с помощью аппарата опорных векторов в случае двух возможных альтернатив 432

Рыко А. И., Степанова С. Б.
Стратегии членения спонтанной речи на синтаксические единицы 438

Семенова С. Ю.
Информация энциклопедического характера в прикладном семантическом словаре 444

Сидорова Е. А., Кононенко И. С.
Подход к извлечению фактов из текста на основе онтологии 451

Скатов Д. С., Ерехинская Т. Н., Окатьев В. В.
Модели и методы анализа иерархически структурированных текстов 458

Соколова Е. Г., Загорулько Ю. А., Кононенко И.С.
Опыт систематизации знаний и интернет-ресурсов для портала знаний по компьютерной лингвистике 465

Тер-Аванесова А. В., Крылов С. А.
Использование лексико-грамматических баз данных в русской диалектной лексикографии 471

Тимошенко С. П. Цинман Л. Л.
Лексические функции и возможности оптимизации поиска информации в интернете (на материале параметрических слов) 476

Тихомиров И. А., Смирнов И. В.
Применение методов лингвистической семантики и машинного обучения для повышения точности и полноты поиска в поисковой машине «Exactus» 483

Труб В. М.
К проблеме вариативности видовых форм императива 488

Урысон Е. В.
Разговорные словечки как бы и конкретно 493

Усачёва М. Н.
Значения предлогов «по» и «к» в русском языке: кодирование сирконстантов и семантических ролей 499

Фёдорова О. В., Шаврыгина А. С.
Влияние места словесного ударения на распознавание слов в русской устной речи 504

Хуршудян В. Г., Даниэль М. А., Левонян Д. В., Плунгян В. А., Поляков А. Е., Рубаков С. В.
Восточноармянский национальный корпус www.eanc.net 509

Циммерлинг А. В.
Синтаксические нули в теории грамматики 519

Цирульник Л. И., Барбук С. Г., Лобанов Б. М.
Статистический анализ и контекстуальные правила разрешения графической омонимии при синтезе речи по тексту 530

Шарапов Р. В., Шарапова Е. В.
Алгоритм обнаружения ссылочного спама 537

Шаронов И. А.
Коммуникативы и методы их описания 543

Шмелёва Е. Я., Шмелёв А. Д.
Вариативность, продолжение и серийность анекдотов: проблемы построения базы данных 548

Юдина М. В., Фёдорова О. В.
Разрешение синтаксической неоднозначности: эффекты прайминга и самопрайминга 554

Яворская М. В., Азарова И. В.
Структура атрибутивных значений в тезаурусе RussNet (на материале перцептивных прилагательных) 559

Ягунова Е. В.
Формирование наборов опорных слов в разных типах восприятия речи 566

Янко Т. Е.
Русские обращения: словарная информация и вокативные конструкции 574

Gornostay T., Aker A.
Создание и использование многоязычного корпуса объектно-ориентированных топонимических текстов для оптимизации задачи автоматического генерирования описания изображений 580

Meelis Mihkla, Indrek Kiissel, Tõnis Nurk, Liisi Piits
Транскрибирование, структурирование и временной анализ речевого корпуса эстонского языка при выборе единиц в системе синтеза (текст-речь) 588

Partee Barbara H.
Динамический характер значения прилагательных 593

Petrenko M.
Онтологическая семантика и абдукция: обработка эллипсиса 598

Abstracts 605

Авторский указатель 616

Поиск ключевых слов с использованием решетки слогов

Keyword search using syllable lattice

Алиев Р. М. (RomanAliyev@gmail.com),
Янь Цзинбинь (yjbemail@163.com), **Хейдоров И. Э.** (igorhmm@mail.ru)

Белорусский государственный университет, Минск, Беларусь

Методы поиска ключевых слов, основанные на решетке слогов могут исключить проблему отторжения слов, не входящих в словарь (СВС), и компенсировать потерю качества поиска ключевых слов, обусловленную ошибками распознавания. В данной работе предлагается для повышения точности поиска ключевых слов вычислять апостериорную вероятность для различных последовательностей фрагментов слов (слогов). Эксперименты показали высокую эффективность предложенного алгоритма для поиска ключевых слов в потоке слитной речи.

1. Введение

Поиск ключевых слов в речевом потоке является одной из наиболее сложных задач в области обработки речи. На основе этой технологии можно реализовать системы аудио индексации, поиска речевой информации по образцу в мультимедиа-архивах, автоматический контроль речевых сообщений в системах безопасности и т. д. [1] [2].

В настоящий момент используются несколько методов поиска ключевых слов. Первый и наиболее простой метод поиска ключевых слов использует распознаватель со словарем для перевода непрерывной речи в текст. Для поиска ключевого слова осуществляется поиск в полученном тексте с использованием традиционных алгоритмов поиска текста. Проблема этого метода состоит в том, что из-за ограниченного множества слов в распознавателе, невозможно распознать слова, отсутствующие в словаре, например, имена, акронимы и слова из иностранных языков.

Другой метод поиска ключевых слов основан на скрытых Марковских моделях (СММ). Он использует СММ для каждого ключевого слова и одну «модель мусора» для всех остальных слов [3]. Временная последовательность символов ключевых слов и мусор символов формируется в результате распознавания речевой последовательности. Этот метод не имеет ограничений при условии, что определено множество ключевых слов, которые необходимо найти. Но для каждого нового ключевого слова, которое нужно найти, нужно не только обучать новую СММ модель, но также нужно заново обучать модель мусора, поэтому использование этого метода при определенных условиях вызывает серьезные затруднения.

В последнее время достаточно широкое распространение приобрела идея построения и использования решетки фрагментов речи для решения различных задач [4]. Для поиска ключевых слов каждый узел решетки ассоциируется с моментом времени произнесенной речи. Основное преимущество этого метода в том, что он обладает большой гибкостью: даже если фонема ключевого слова не есть лучшая гипотеза между узлами решетки, она все равно сохраняется в результате распознавания. Результат поиска не зависит от словаря распознавателя, поскольку поиск можно организовывать для любой фонемной последовательности запрашиваемого ключевого слова, поэтому в этом методе проблема СВС решается наиболее естественным и эффективным способом.

2. Структура системы поиска ключевых слов на основе решетки

Введем понятие решетки $L = (N, A, n_{start}, n_{end})$ как направленного непериодического графа, где N — множество его узлов, A — множество связей между узлами и $n_{start}, n_{end} \in N$ — начальный и конечный узлы решетки соответственно. Представим связь между узлами в виде $a = (S[a], E[a], I[a], w[a])$, где $S[a], E[a] \in N$ — начальный и конечный узлы; $I[a]$ — фрагмент речи (слог или фонема); $w[a] = p_{ac}(a)^{1/\lambda}$ — весовой коэффициент связи; $p_{ac}(a)$ — акустическое сходство; λ — весомый коэффициент.

Предположим, что мы имеем последовательность слогов $l_1, l_2, \ldots, l_K$, слияние которых дает ключе-

вое слово *Kw*. В результате поиска нужно получить множество *W* наиболее вероятных ветвей *w* ориентированного графа, которые соответствуют запрашиваемому ключевому слову. Этот поиск осуществляется методом последовательного прокладывания пути в структуре решетки. В первую очередь среди всех узлов множества N необходимо найти начальный узел, соответствующий первому слогу ключевого слова, затем граничных узлов производится поиск следующего вероятного узла, который будет соответствовать следующему слогу. Процедура повторяется *K* раз до тех пор, пока не будет обработан последний слог ключевого слова.

Для полученной последовательности наблюдений *O* производим вычисление апостериорной вероятности для полученных результатов поиска:

$$P(Kw \mid O) = \sum_{\forall w \in W} P(w \mid O), \qquad (1).$$

которая является суммой апостериорных вероятностей всех возможных путей в графе, сопоставляемых ключевому слову *Kw*.

Представим апостериорную вероятность пути следующим образом:

$$P(w \mid O) = P(O, w) / P(O). \qquad (2)$$

Для вычисления $P(w|O)$ воспользуемся алгоритмом прямого и обратного хода:

Шаг 1 **.** Вычисление значений прямых переменных и обратных переменных $\alpha(v)$ и $\beta(v)$:

$$\alpha(v) = P(O_0^{t(v)}, v) = \sum_{w \in W_v^-} P(O_0^{t(v)}, w), \qquad (3)$$

$$\beta(v) = P(O_{s(v)}^{T} \mid v) = \sum_{w \in W_v^+} P(O_{s(v)}^{T}, w) / b(v). \qquad (4)$$

Исходя из уравнений (3), (4):

$$\alpha(v) = \sum_{v_1,\ldots,v_I, v \in W_v^-} [\prod_{i=1}^{I-1} b(v_i) q(v_i, v_{i+1})] b(v_I) q(v_I, v) b(v), \qquad (5)$$

$$\beta(v) = \sum_{v, v_1, \ldots v_I \in W_v^+} q(v, v_1) [\prod_{i=1}^{I} b(v_i) q(v_i, v_{i+1})]. \qquad (6)$$

Очевидно, что $\alpha(v)$ и $\beta(v)$ нельзя вычислить прямо, но их можно выразить через рекурсивные соотношения:

$$\alpha(v) = \sum_{(u,v) \in E} \alpha(u) q(u, v) b(v), \qquad (7)$$

$$\beta(v) = \sum_{(u,v) \in E} q(u, v) b(v)\ \beta(v), \qquad (8)$$

$$\alpha(v) = b(v), \text{ если } v \in V_0, \qquad (9)$$

$$\beta(v) = 1, \text{ если } v \in V_T. \qquad (10)$$

Таким образом, начиная с начального узла, можно вычислить значение $\alpha(v)$ каждого узла и, начиная с конечного узла, можно вычислить значение $\beta(v)$ каждого узла.

Шаг 2. Вычисление вероятности *P(O)*:

$$P(O) = \sum_{w_g \in W_G} P(O, w) = \sum_{v \in V_T} \alpha(v). \qquad (11)$$

Из выражения (11) видно, что $P(O)$ вычисляется непосредственно по прямым переменным всех ключевых слов.

Шаг 3. Вычисление апостериорной вероятности P(w|O):

$$P(w \mid O) = \alpha(v_1)[\prod_{k=1}^{K} q(v_k, v_{k+1}) b(v_{k+1})] \beta(v_K) / P(O). \qquad (12)$$

На основе описанной выше решетки была предложена и разработана следующая схема поиска ключевых слов (рис. 1). Работа системы осуществляется в три этапа. Первый этап — выделение мелкепстральных признаков речевого сигнала на основе вейвлет-преобразования. В работе предлагается использовать вейвлетное преобразование для построения вектора признаков ввиду его хорошего разрешения по времени для спектрального диапазона речевого, что особенно важно для автоматической сегментации [5]. Второй этап — обучение распознавателя с помощью языковой и акустической баз данных, в результате чего формируется решетка фрагментов слов. Третий этап — поиск в решетке возможных ключевых слов с подтверждением с помощью вычисления апостериорной вероятности.

3. Эксперимент

Предложенная схема поиска ключевых слов на основе вейвлетного преобразования и решетки фрагментов слов была реализована на языке программирования C++ с использованием библиотеки VCL (рис.2). База данных акустических моделей была получена с помощью отсегментированной чистой речи продолжительностью 124 часа. Для проведения эксперимента по поиску ключевых слов была использована записанная речь диктора радиопередачи «Новости» продолжительностью 3,54 часа. В качестве вектора признаков был использован 39-компонентный вектор, включающий в себя энергию, мелкепстральные характеристики и их производные, полученные на основе вейвлет-преобразования. Обучение распознавателя на основе СММ было осуществлено с использованием библиотеки HTK [6].

Результаты поиска 40 наиболее часто встречаемых в новостях ключевых слов с использованием предложенной системы по сравнению с обычной системой на основе алгоритма *N-bes* [6], представ-

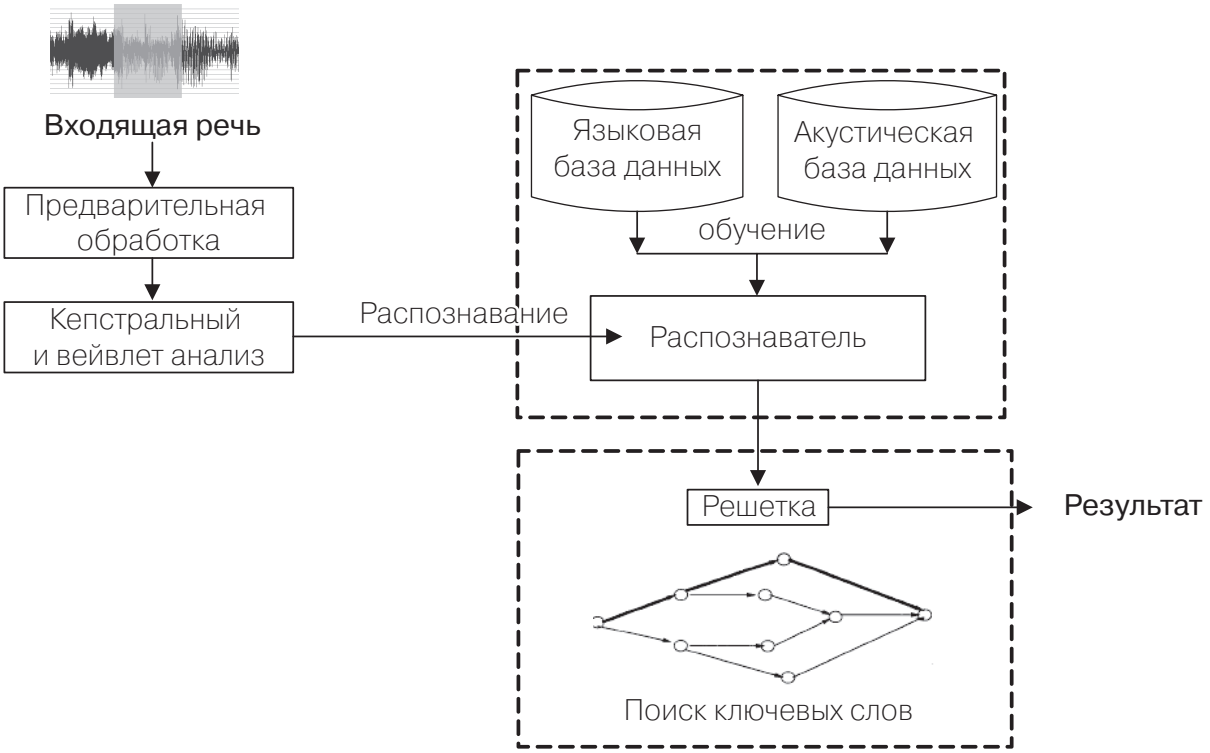


Рис. 1. Схема работы системы поиска ключевых слов на основе решетки фрагментов слов.

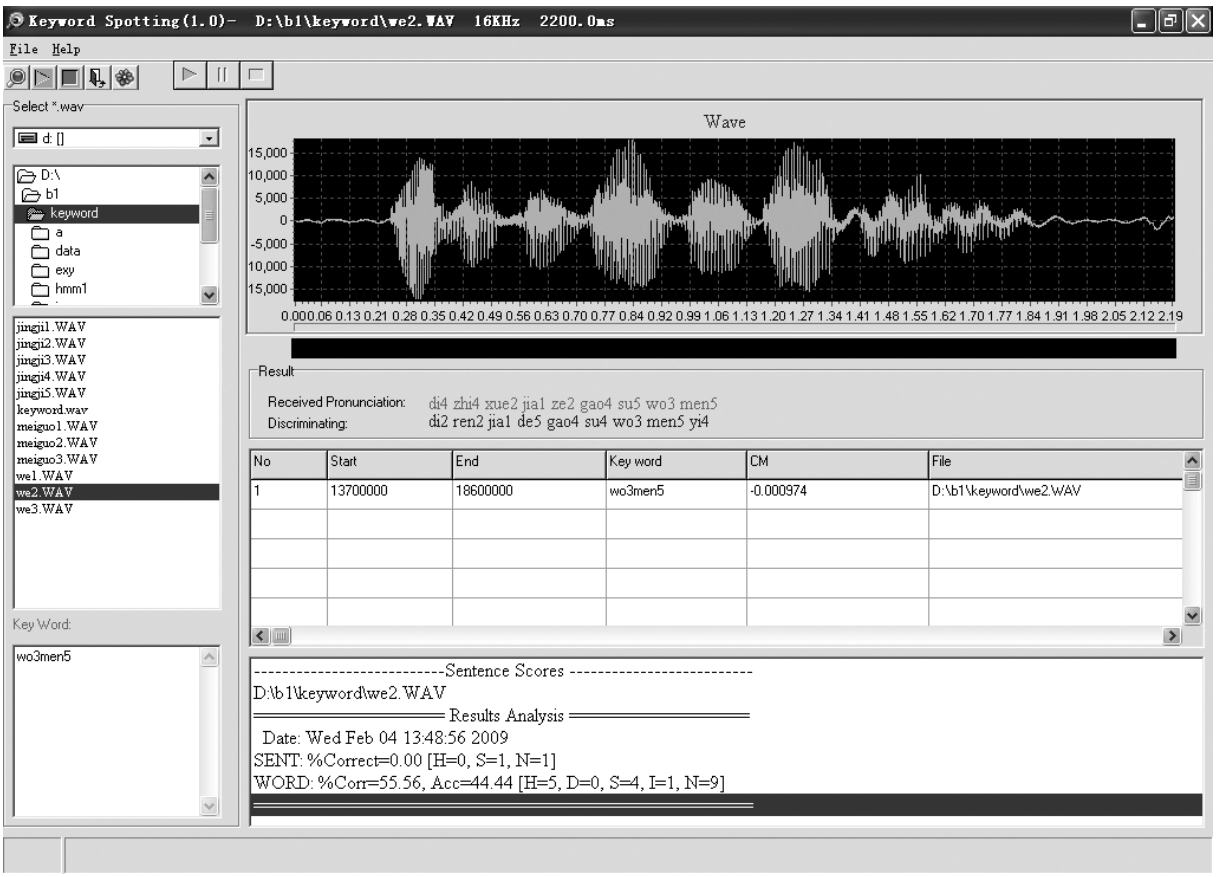


Рис. 2. Интерфейс системы поиска ключевых слов на основе решетки слогов.

лены в табл. 1 и 2. 40 слов для поиска были выбраны таким образом, чтобы их встречаемость в новостях была не менее 100 раз. Это сделано для того, чтобы обеспечить статистическую устойчивость оценки точности, поскольку при меньшем количестве встречаемых слов каждая ошибка поиска очень сильно влияет на оценку характеристик системы. В таблице 1 приводятся средние значения характеристик системы обнаружения ключевых слов: точность, представляющей собой отношение количества правильно найденных слов к общему количеству слов, классифицированных как ключевые, и чувствительность, представляющее отношение количества правильно найденных слов к общему количеству ключевых слов, присутствующих в записи. В таблице 2 для каждого из 13 ключевых слов представлено число их правильных обнаружений в потоке речи при помощи алгоритма N-best (N=1, 3, 5, 7, 9, 15, 20) и на основе решетки слогов.

Как видно из таблицы, использование алгоритма решетки позволяет улучшить точность поиска ключевых слов в среднем на 9 процентов по сравнению с алгоритмами на основе N-best поиска.

4. Заключение

В данной работе предложена и реализована система поиска ключевых слов на основе решетки частей слов. Результат эксперимента показал, что использование решетки для слогов, сгенерированных на основе СММ, позволяет улучшить точность поиска в среднем на 9 процентов.

Таблица 1. Сравнение точности распознавания и чувствительности для систем на основе решетки слогов и N-best.

Система распознавания ключевых слов	**Точность**	**Чувствительность**
N-best (N=20)	69.5%	77.6%
Система на основе решетки слогов	86.2%	88.3%

Таблица 2. Сравнительная таблица количества правильно обнаруженных слов различными методами

Ключевое слова	**Всего слов**	**Обнаружено слов**							
		N=1	N=3	N=5	N=7	N=9	N=15	N=20	Решетка
fang1mian4	180	116	119	122	125	133	136	137	163
guan1xi4	172	114	114	116	119	124	128	130	158
guo2jia1	486	312	334	345	354	360	369	369	435
jing1ji4	576	378	384	401	412	427	444	447	498
mei3guo2	232	156	158	157	157	159	159	160	220
qi3ye4	418	250	270	273	287	302	316	316	353
sheng1chan3	252	176	177	180	182	186	189	190	222
shi4jie4	182	80	110	114	120	137	147	148	154
ta1men5	420	268	277	285	296	303	319	320	385
wo3men5	676	360	407	430	478	512	523	525	589
yi2ge4	528	324	344	365	387	400	411	412	457
zhong1guo2	620	433	438	440	449	457	465	465	566
zi4ji3	228	140	153	154	157	159	168	170	208
Всего	4970	3107	3285	3382	3523	3659	3774	3789	4408

Литература

1. *Young S. J., Brown M. G., Foote J. T., Jones G. J. F., Jones K. S.* Acoustic indexing for multimedia retrieval and browsing // IEEE International Conference on Acoustics, Speech and Signal Processing, 1997.
2. *Jones G. J. F., Foote J. T., Jones K. S., Young S. J.* Retrieving spoken documents by combining multiple index sources // Proc. SIGIR, 1996. P. 30–38.
3. *Wilpon J. G., Rabiner L. R., Lee C. H., Goldman E. R.* Automatic recognition of keywords in unconstrained speech using Hidden Markov Models // IEEE Transactions on Acoustics, Speech and Signal Processing, 1990.
4. *Soong F. K., Lo W. K., Nakamura S.* Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words // Proceeding of SWIM2004, 2004. P. 127–128.
5. *Кухарчик П. Д., Хейдоров И. Э., Бовбель Е. И., У Ши, Янь Цзинбинь.* Определение патологий голосового тракта путем анализа речевого сигнала на основе вейвлетного преобразования и метода опорных векторов // Электроника, Минск, БГУ, 2008. № 12, С. 44–49.
6. *Young S., Evermann G., Kershaw D.* The HTK Book // 2008.

Семантические источники уступительности[1]

Semantic sources of concession

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

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

Уступительность — сложный смысл, и обычно слова с уступительным значением развиваются из слов с более простыми, первичными смыслами — например, условия (*если и, даже если*), желания (*хотя, хоть*), одновременности (*в то время как, тогда как*), степени (*тем не менее, по крайней мере*), соответствия действительности (*правда*), вероятности (*конечно*) и другими.

Уступительность связана с перечисленными смыслами, а также с некоторыми другими, не только диахронически (как, например в случае наречия *мало*, имевшего еще в XVIII веке уступительное значение), но и синхронно — на уровне смыслов отдельных лексем, а также полисемии.

Так, в семантический инвариант уступительности, в первую очередь, в значение основных уступительных лексем *хотя, хоть, несмотря на* входит указание на условие и вероятность, а в многие уступительные единицы — также указание на желание (*только бы, лишь бы*), высокую или низкую степень (*несмотря ни на что, хоть*), вероятность (*даже если*), оценку говорящего (*добро бы*).

И если у *хотя* и *хоть*, у которых уступительное значение является основным, в синхронной полисемии смысл ‘желание’ как самостоятельное значение не вычленяется, то в синхронной полисемии слова *правда* переносное уступительное значение присутствует наряду с основным — значением соответствия действительности, а в полисемии *только* — переносное уступительное значение присутствует наряду с основным значением количественной оценки.

Образцом переплетения смыслов уступительности, малой и высокой степени, желания, вероятности может служить полисемия слова *хоть*, у которого есть значение и классического концессива — нарушение нормы и обманутые ожидания, и риторические уступительные значения, содержащие полемику с потенциальным оппонентом. Ниже приводятся некоторые значения с краткими толкованиями, иллюстрирующие данное утверждение:

хоть 1.1. ‘имеет место P, имеет место Q; обычно если имеет место ситуация типа Q, то не имеет места ситуация типа P’: *Он поел, хоть не был голоден; Он хоть и был пьян, но все же что-то соображал* [классическое уступительное значение — говорящий отмечает необычность одновременного существования некоторых двух ситуаций]

хоть 1.2. ‘имеет место P, говорящий признает, что имеет место противоположно оцениваемое Q; говорящий считает, что P важнее’: *Домик хоть и небольшой, а свой; Он стал победителем, хоть и с минимальным отрывом* [риторическое уступительное значение — говорящий отмечает, что имеет место некоторая, по его мнению, важная ситуация, причем признает, что имеет место и ослабляющее важность этой ситуации дополнительное обстоятельство]

хоть 1.3. ‘даже если бы имело место P в высокой степени, имело бы место Q’: *Не пущу, хоть зарежьте; Хоть даром отдай, никто у тебя эту развалюху не возьмет* [значение уступительности и высокой степени — говорящий сообщает, что некоторая ситуация не будет иметь места даже при исключительно благоприятных сопутствующих обстоятельствах]

хоть 2.1. ‘говорящий или субъект готов удовлетвориться очень малым P’: *Хоть денек отдохнуть удалось; Хоть глоток воды подай!* [значение уступительности, желания и малого количества — говорящий сообщает, что в отсутствие желаемого готов удовлетвориться чем-то очень малым].

[1] Эта работа написана при финансовой поддержке следующих грантов: грант Президента РФ для ведущих научных школ НШ-3205.2008.6, грант РГНФ 07-04-00-202а “Системообразующие смыслы русского языка” и грант Программы фундаментальных исследований историко-филологического отделения РАН «Русская культура в мировой истории».

У слова *хоть* основное значение, а также большинство вторичных — уступительные, поэтому мотивировать развитие его полисемии не столь сложно.

Однако есть много слов, у которых уступительное значение является одним из многих переносных значений, а основное значение какое-то другое. При анализе таких уступительных лексем бывает полезно обращение к их исходному значению, во-первых, для того, чтобы понять логику возникновения переносных уступительных значений, а также для того, чтобы различить их между собой.

Всякому, кто работает с толкованиями слов, особенно лексикографу, хорошо известно, как трудно бывает сделать толкование достаточно общим, чтобы оно охватывало весь круг употреблений данной лексемы, но при этом достаточно специфичным, чтобы оно было способно отразить различия между нею и ее близкими синонимами. Особенно это касается переносных значений, семантика которых часто бывает менее четко очерчена, чем семантика основных значений. В некоторых случаях именно обращение к исходному значению может способствовать более тонкому различению между синонимичными лексемами с близкими значениями, поскольку переносные значения сохраняют некий семантический и прагматический «отсвет» исходного.

Рассмотрим две уступительных лексемы — производный от частицы *только* союз *только* (в вариантах *вот только, только* и *только вот*) и производную от существительного *правда* единицу смешанной синтаксической природы *правда*, которая может употребляться как вводное слово и как союз. В значении, представленном в примерах ниже, они синонимичны:

(1а) Он способный, правда очень ленивый

(1б) Он способный, только очень ленивый

(2а) Было солнечно, правда холодновато

(2б) Было солнечно, вот только холодновато

(3а) Нам платят регулярно, правда немного

(3б) Нам платят регулярно, только вот немного

Эти лексемы настолько близки, что часто употребляются вместе: *Завтра же он покончит со всем этим и придёт в музей за расчётом […] Вот только, правда, Дашу жалко немного* (Ю. Домбровский, Факультет ненужных вещей).

На первый взгляд кажется, что к обеим лексемам подходит толкование, предлагаемое в работе [Санников 1989, 180] для союза *только*, которое выглядит следующим образом:

X, только Y = ‘X;

воздействие X-а на описываемую ситуацию (или общую положительную оценку) ослаблено, но незначительно, наличием Y-а’.

То есть, *только* (и, возможно) *правда* описывают такое соотношение ситуаций, когда упоминаемая первой и оцениваемая положительно ситуация X является более важной для общей оценки положения вещей, чем упоминаемая второй и оцениваемая отрицательно ситуация Y. На самом деле, как мы увидим, это толкование адекватно описывает большое количество (но не все) употребления союза *только*, а также многие употребления лексемы *правда*. Кроме того, при дальнейшем рассмотрении материала становится видно, что в ряде контекстов эти лексемы не взаимозаменимы. Обращение к их исходным значениям позволяет понять природу их различий в переносных значениях.

Семантические и прагматические различия между *только* и *правда*

I. Во-первых, и это отмечается в работе [Апресян 2006:701], *правда* не всегда описывает доминирование положительно оцениваемой ситуации над отрицательно оцениваемой — распределение может быть и обратным; ср. (4а) и (4б), где доминирует плохое:

(4а) Нам платят мало, правда, регулярно

(4б) Он жуткий лентяй, правда, довольно способный

Для *только* такое употребление невозможно:

*(5а) *Нам платят мало, только регулярно*

*(4б) *Он жуткий лентяй, только довольно способный*

Таким образом, фиксируем первое различие между этими лексемами — для лексемы *правда* возможно доминирование плохой ситуации над хорошей, для лексемы *только* — невозможно. Интересно, что для близких к союзу *только* союзов *только что* и *вот только что* такое распределение ситуаций (плохая доминирует над незначительной хорошей) является единственно возможным: *Нам платят мало, только что <вот только что> регулярно*, при невозможности **Нам платят регулярно, только что <вот только что> мало*.

II. Однако *только* может употребляться в контекстах, где ситуации вообще не оцениваются как положительные или отрицательные, а союз имеет чисто ограничительную интерпретацию, а именно,

вводит обстоятельство, которое является исключением из общего положения вещей:

(5) *У нас все по-прежнему, только я перешел на другую работу*
[Имеет место некоторое положение вещей — ‘У нас все по-прежнему’; единственное обстоятельство, которое является исключением из статуса-кво, — это ‘Я перешел на другую работу’]

Иногда подобные ограничительные контексты требуют некоторого прагматического домысливания, чтобы понять, почему некоторое обстоятельство является исключением из общей ситуации:

(6) *Она выглядела совершенно спокойной, только глаза блестели ярче обычного*
[Имеет место некоторая ситуация — ‘Она выглядела совершенно спокойной’; единственным обстоятельством, которое мешает такой оценке, является ‘Глаза ее блестели ярче обычного’, т.к. блеск глаз может указывать на волнение]

(7) *Темная ночь, только пули свистят по степи*
[Имеет место некоторая ситуация — ‘Темная ночь’; предположительно, ночь — это такое время, когда все тихо и ничего не происходит; единственное обстоятельство, которое нарушает тишину — это ‘Пули свистят по степи’]

В Малом академическом словаре русского языка употребления типа *Он способный, только ленивый* и *У нас все по-прежнему, только я перешел на другую работу* подаются как отдельные значения — уступительное и ограничительное; в Большом академическом словаре русского языка подобные употребления рассматриваются внутри одного общего значения.

Кажется, что для выделения двух разных значений нет достаточных оснований, т.к. нет контекстов, где были бы возможны обе интерпретации — чисто ограничительная и чисто уступительная; семантически уступительное употребление является частным случаем ограничительного. И в том, и в другом случае вторая часть сообщения в какой-то мере ослабляет собой первую. Прагматическая оценка ситуаций как желательной или нежелательной, которая ассоциируется с уступительным употреблением, привносится контекстом, как в примерах (1б), (2б), (3б), или же возникает «по умолчанию», когда невозможна никакая другая интерпретация. Так, фразы типа *Платье длинное, только красное* могут восприниматься либо как аграмматичные, либо как оценочные. Так как цвет платья и его длина — параметры независимые, непонятно, каким образом сообщение об одном могло бы ослаблять важность сообщения о другом. Единственная интерпретация, при которой можно это предложение осмыслить — оценочная; если подразумевается, что субъект хотел бы получить длинное платье, причем не красного цвета, то сообщение о длине будет восприниматься как желательное и положительное, а сообщение о цвете — как ослабляющее его нежелательное и отрицательное: *Платье длинное, это хорошо, только, к сожалению, красное — мне не идет*.

При этом, как сказано в предыдущем пункте, для *только* оценочная интерпретация может быть лишь однонаправленной: доминирующее положение вещей всегда оценивается положительно, исключение из него всегда оценивается отрицательно. Поэтому обратное распределение оценок невозможно; ср. неправильность **Платье длинное, только, к счастью, красное*.

Для *правда* такая оценочная интерпретация предпочтительна, однако не обязательна; так, фраза *Платье длинное, правда красное*, при всей своей прагматической неловкости, может иметь две оценочных интерпретации: *Платье длинное, правда, к сожалению, красное* [субъект хотел получить длинное не красное платье; доминирует положительная оценка]; *Платье длинное, правда, к счастью, красное* [субъект хотел получить не длинное красное платье; доминирует отрицательная оценка].

Интересно, что эта особенность, обязательное доминирование положительного при наличии оценки у *только*, распространяется только на союз, но не на ограничительную частицу; ср. контрастные примеры:

(8а) *Учится он ужасно, только по биологии пятерки = Учится он ужасно, пятерки только по биологии*
[*Только* — частица; доминирует отрицательная оценка]

(8б) **Учится он ужасно, только увлекается биологией*
[*Только* — союз; доминирование отрицательной оценки невозможно]

По параметру оценочности одиночный союз *только* отличается также от *вот только* и *только вот*; если для *только*, как показано выше, возможны безоценочные, чисто ограничительные контексты, то союз *только* в сочетании с частицей *вот* всегда прагматически окрашен и вносит какую-то утилитарную оценку в самые нейтральные фразы. Так, фраза *Темная ночь, вот только <только вот> пули свистят по степи* подразумевает какое-то недовольство говорящего тем, что тишина ночи нарушается свистом пуль. Во фразы, где вторая, менее важная ситуация в силу каких-то обстоятельств не может оцениваться отрицательно, *вот только* и *только вот* не подставляются; ср. пример из МАС’а: *Несколько секунд стояла тишина, только*

вода тихо и ласково звенела в шлюзах (Короленко, Слепой музыкант), где *вот только* и *только вот* были бы прагматически неуместны: ?*Несколько секунд стояла тишина, вот только <только вот> вода тихо и ласково звенела в шлюзах*.

По-видимому, это связано с наложением семантики и прагматики дейктической частицы *вот* на семантику и прагматику частицы *только*. *Только*-частица имеет два основных значения — исключения, описанное, в частности, в работе [Богуславский 1996:92][2] (*Только дети могут так оригинально мыслить*), а также значение количественной оценки, описанное, в частности, в работе [Апресян 1995:68] (*Он съел только два арбуза*) — 'X, и говорящий считает, что это мало'. Союз *только* наследует от частицы *только* и семантическое указание на исключительность, и, в какой-то мере, прагматическое указание на низкую количественную оценку (последнее является слабым смыслом и реализуется не во всех контекстах одиночного *только*).

Количественная оценка вообще легко трансформируется в качественную, и словам со значением малого количества и низкой степени свойственно развивать отрицательную прагматическую оценку; ср., например, наречие *мало*, у которого есть значение 'недостаточно', содержащее отрицательную утилитарную оценку (*Для визы одного приглашения мало*). В сочетании с семантикой частицы *вот*, фокусирующей внимание на выделенном объекте, уже заключенное в значении *только* указание на исключительность и оценку дополнительно «кристаллизуется» и закрепляется; поэтому безоценочные контексты, возможные для одиночного *только*, невозможны для *вот только* и *только вот*.

III. Еще одно различие между *только* и *правда* касается соотношения важности рассматриваемых двух ситуаций. Согласно первоначальному предположению и приводившемуся толкованию, обе лексемы указывают на то, что первая из упоминаемых ситуаций важнее, а вторая, вводимая союзом, менее важна. Однако анализ примеров показывает, что это не вполне соответствует действительности. Рассмотрим следующие примеры:

(9а) *Телевизор-то мы купили, только он не работает*

(9б) *Лечить его лечили, только вот не вылечили*

(9в) *Он парень обаятельный, только вот вор и наркоман*

(9г) *Обещали платить по 500 евро в день, только вот не заплатили ни копейки*

Необычность этих примеров состоит в том, что в них нарушается привычное для *только* соотношение ситуаций — первая ситуация важнее, вторая ситуация, вводимая союзом, менее важна и ею можно пренебречь. Прагматически ситуации, вводимые союзом *только* (и *только вот*), а именно 'не работает', 'не вылечили', 'вор и наркоман', 'не заплатили' явно важнее для оценки общего положения вещей, чем упоминаемые первыми ситуации 'купили', 'лечили', 'обаятельный', 'обещали заплатить'. Кроме того, в первых двух примерах относительная неважность первой части сообщения подчеркивается дополнительно — употреблением уступительной частицы *–то* и уступительно-противительной фразеосхемы ИНФИНИТИВ + ЛИЧНАЯ ФОРМА ГЛАГОЛА[3], в смысл которых входит указание на неважность ситуации, вводимой ими, по сравнению с ситуацией, которая замыкает сообщение.

В этих примерах лексема *правда* прагматически странна:

(9а) *?Телевизор-то мы купили, правда, он не работает*

(9б) *?Лечить его лечили, правда, не вылечили*

(9в) *?Он парень обаятельный, правда, вор и наркоман*

(9г) *?Обещали платить по 500 евро в день, правда, не заплатили ни копейки*

Если предложенные в начале толкование *только* было бы абсолютно точным, союз *только* также не мог бы употребляться с частицей *-то* и с уступительно-противительной фразеосхемой, т. к. возникала бы семантическая аномалия — противоречие модальных рамок: *-то* и фразеосхема предполагают, что первая ситуация не важна, а важна вторая, а *только*, согласно толкованию, предложенному в [Санников 1989, 180], имеет обратное распределение оценок по важности и указывает на то, что важна первая ситуация, а вторая не важна. Однако прагматический запрет на подобное употребление свойствен лишь лексеме *правда*, но не лексеме *только*.

В чем же разница между *только* и *правда*? Значит ли это, что указание на относительно большую важность ситуации, вводимой союзом, есть в значении *правда*, но отсутствует в значении *только*?

---

[2] '*только* (P, Q, R)' = 'среди элементов множества P ни один, отличный от Q, не обладает свойством R'.

[3] Уступительно-противительные фразеосхемы подробно рассматриваются в работах [Булыгина, Шмелев 1997, Шмелев 2002]. Они по преимуществу диалогичны и предполагают скрытую полемику с собеседником — «возражение под видом согласия» [Булыгина, Шмелев 1997: 313]. Как отмечается в работе [Шмелев 2002: 426], они выражают смыслы 'хотя и' или 'хотя и не'.

Для того, чтобы ответить на это вопрос, обратимся к основным значениям слов *только* и *правда*.

Различия в их переносных значениях в какой-то степени являются отражением различий в их основных значениях. У *только* переносное уступительное значение «наследует» от основных значений два элемента — элемент ‘исключения’ — (‘среди элементов множества P ни один, отличный от Q, не обладает свойством R’) и элемент количественной оценки ‘X, и говорящий считает, что это мало’. Последний и создает в большинстве контекстов союза *только* оценку второй ситуации как ‘неважной’, то есть смысл ‘мало’ трансформируется в смысл ‘пренебрежимо’ или ‘неважно’ (см. также выше). Однако элемент количественной оценки является слабым смыслом и нейтрализуется, когда контекст этому противоречит, как в вышеприведенных примерах (9а)–(9г).

Что касается элемента ‘исключения’, то сам по себе он недостаточен для импликации смысла ‘мало’. Безусловно, в большинстве контекстов эти два смысла естественно комбинируется друг с другом; ср. — *Тебя кто-нибудь поздравлял с днем рожденья? — Да никто не поздравлял, только Ваня* [‘Поздравлял Ваня, и это мало’].

Однако ‘исключение’ может имплицировать и совсем другую прагматику. А именно, в силу своей единственности, элементы, являющиеся исключением, привлекают внимание адресата и, соответственно, приобретают значимость. Это очень заметно в основном значении частицы *только*; ср. следующий пример, где элемент-исключение является единственным значимым:

(11) Мне это принесет только радость

Контексты, где элемент, являющийся исключением, на самом деле является самым важным, встречаются у всех лексем со значением исключения, не только у частицы *только*; ср. также предлог *кроме*:

(12) Я от тебя не видела ничего, кроме горя

Таким образом, в значении союза *только* элемент ‘исключение’ может приобретать совершенно разную прагматическую окраску — либо ‘P является единственным исключением, и поэтому на P можно не обращать внимания’, либо ‘P является единственным исключением, и поэтому P привлекает к себе внимание’.

В сочетании с фокусирующей внимание частицей *вот* вторая интерпретация более естественна, поэтому *вот только* и *только вот* часто вводят важные ситуации, которыми нельзя пренебречь при оценке общего положения вещей:

(13) И стать бы ему непременно героем Отечественной войны, да вот только война-то кончилась (М. Пришвин, Кладовая солнца)

(14) Перенимать опыт профессионалов — дело нужное, только вот голову терять при этом не следует («Солдат удачи», 2004.08.04)

Вообще же главное в употреблении *вот только* и *только вот* — непременная оценочность, о которой говорилось выше, в разделе II. Первая ситуация всегда оценивается, как нечто хорошее, вторая ситуация — как нечто плохое, а вот количественная оценка, относительная важность ситуаций может варьироваться от контекста к контексту. Часто в употреблении *вот только* и *только вот* есть некоторая сознательная игра, некоторое риторическое преуменьшение обстоятельства, которое на самом деле является самым важным, и к которому говорящий желает привлечь внимание.

Кроме того, есть дополнительный синтаксический фактор, способствующий привлечению внимания к ситуации, вводимой союзами *только*, *вот только* и *только вот*. А именно, *только*, *вот только* и *только вот* не могут начинать предложение, они всегда его заканчивают, то есть всегда употребляются в позиции фокуса, ремы. Фраза с *только*, *вот только* и *только вот* выделяется фразовым ударением: *Я понимаю вашу ситуацию, только <только вот, вот только> упреки ваши — не по ↓адресу*.

Таким образом, одна из основных прагматических функций *только*, особенно в сочетании с *вот* — привлечение внимания к некоторому обстоятельству Q, которое ослабляет важность ситуации P и часто нежелательно для говорящего или третьего лица.

На основании вышеприведенного анализа употреблений союза *только*, можно предложить для него следующее уточненное толкование:

P, только Q ‘Имеет место P, часто желательное; имеет место Q, часто нежелательное; говорящий сообщает, что важность P не ослабляется ничем, кроме Q’ [в зависимости от ситуации Q может быть очень важным или неважным]

Для *только* в сочетании с частицей *вот* предлагается следующее толкование:

P, только вот Q,
P, вот только Q ‘Имеет место желательное P; имеет место нежелательное Q; говорящий сообщает, что важность P не ослабляется ничем, кроме Q; говорящий хочет привлечь внимание к Q’ [Q чаще важно]

Семантика и прагматика лексемы *правда* устроена несколько иначе. У *правда* желательность/нежелательность относятся не к самим ситуациям P и Q, а к коммуникативным намерениям говорящего, ко-

торый хочет привлечь внимание к существованию ситуации P (которая сама по себе может быть как желательной, так и нежелательной для говорящего или третьего лица).

Таким образом, прагматика лексемы *правда* в каком-то смысле противоположна прагматике *только*, особенно в сочетании с *вот*; говорящий признает, что некоторое обстоятельство Q имеет место, однако хотел бы отвлечь от него внимание, «спрятать» его, поскольку его целью является сообщение о существовании ситуации P, важность которой обстоятельство Q ослабляет. «Из честности», или же уступая оппоненту, говорящий мельком упоминает ситуацию, вступающую в противоречие с его основным коммуникативным намерением, однако представляет ее как менее важную: *Правда, машина не новая, но очень надежная*.

Синтаксически *правда* может употребляться и в постпозиции, и в препозиции к фразе, вводящей основную ситуацию P, и, кроме того, может сопровождаться сообщением, усиливающим P; интонационно, фраза, вводимая *правда*, не несет главного фразового ударения, им выделяются фразы, вводящие ситуацию P, а также усиливающее сообщение о P:

(19а) Я ↓рада тебя видеть [P], правда, сейчас уже поздновато [Q], но ты ↓приходи [усиление P] — важно то, что говорящий хочет видеть адресата, а позднее время, мешающее этому, не важно — в целом предложение является приглашением прийти

(19б) Правда, сейчас уже поздновато [Q], но я ↓рада тебя видеть [P], ты ↓приходи [усиление P].

Если подставить в эти фразы *только*, то, во-первых, перестановка P и Q становится невозможной, во-вторых, меняются фразовые ударения, в-третьих, меняется коммуникативное намерение:

*(20а) *Только сейчас уже поздновато [Q], но я ↓рада тебя видеть [P], ты ↓приходи [P'] — семантическая аномалия, т. к. только подразумевает, что первая ситуация важнее, а но — что вторая*

(20б) Я ↓рада тебя видеть [P], только сейчас уже ↓поздновато [Q] — тот факт, что время позднее, важнее, чем желание говорящего увидеть адресата; в целом, предложение является объяснением, почему приглашение не состоится

Чем можно объяснить эти особенности лексемы *правда* по сравнению с *только*? Представляется, что обращение к исходному значению существительного *правда* может быть полезным. Для существительного *правда 1* в работе [В. Апресян 2009][4] предлагается следующее толкование:

> *правда об A1* (*Я хочу знать правду об этом деле*; *До сих пор неизвестна правда об убийстве этого журналиста*; *Правда оказалась ужасной*):
> 'То, каким A1 является в действительности;
> люди не знали, какой A1, или думали, что A1 другой;
> говорящий считает, что знание того, каким A1 является в действительности, важно и ценно для людей;
> говорящий считает, что какие-то люди не хотят этого признавать или не хотят, чтобы другие это знали'.

Уступительная лексема *правда* наследует от существительного *правда 1* компонент отношения к действительному положению вещей как к чему-то нежелательному; таким образом, общим для *правды*-существительного и уступительной *правды* является смысл 'говорящий считает, что какие-то люди не хотят этого признавать или не хотят, чтобы другие это знали'. В случае уступительной *правды* человеком, который не хочет признавать и упоминать действительное положение вещей, является сам говорящий, и ситуацию Q он упоминает только «для полноты картины», либо отчасти соглашаясь с оппонентом (адресатом), который упомянул ее ранее.

Предложим скорректированное толкование уступительной лексемы *правда*:

> *P, правда Q*
> 'Имеет место P;
> имеет место Q;
> говорящий хочет сделать утверждение, что P;
> говорящий признает, что имеет место Q, ослабляющее важность P;
> говорящий считает, что в данной ситуации P несколько важнее, чем Q'.

Суммируя сказанное, отметим, что *правда*, *только* и *вот только* <*только вот*> имеют следующие различия в употреблении, вызванные нижеследующими семантическими и прагматическими причинами, в первую очередь определенным фокусом внимания и наличием разных модальных рамок:

---

[4] Из-за недостатка места у нас нет возможности приводить аргументы в пользу каждого из элементов толкования слова *правда* в его основном значении, но в работе [В. Апресян 2009] каждый из семантических компонентов подробно мотивируется посредством сопоставления лексемы *правда* с лексемой *факт*.

	«Объективные» употребления, не содержащие оценки ситуаций P и Q по желательности/нежелательности вида: *Тихо* [P], *только дождь стучит по крыше* [Q]	Может ли вводить желательную ситуацию Q, над которой доминирует нежелательная ситуация P; ср. *Платят мало* [P], *правда регулярно* [Q]	Может ли ситуация Q, вводимая союзом, быть важнее ситуации P: *Приятно-то приятно* [P], *только неполезно* [Q]; *Обещать обещал* [P], *только ничего не сделал* [Q]
правда	**невозможны**, т. к. *правда* содержит указание на коммуникативные **желания** говорящего — он хочет подчеркнуть важность P и преуменьшить важность Q, поэтому какая-то оценка всегда присутствует	**допускает**, т.к. желания говорящего связаны не с содержанием ситуаций, а с содержанием его сообщения — говорящий хочет сообщить о P и по возможности отвлечь внимание от Q, поэтому P для него всегда имеет приоритет, независимо от того, каковы P и Q	**невозможно**, т. к. цель говорящего — сделать сообщение о наличии P, которую он считает более важной, и отвлечь внимание от Q, которую он считает менее важной, и упоминает «против воли»
только	**возможны**, так как у *только* оценка является слабым смыслом и может нейтрализоваться контекстом	**не допускает**, т.к. в значение *только* входит оценка P как желательной ситуации, Q как нежелательной ситуации, которая может нейтрализоваться контекстом, но не меняться на противоположную	**возможно**, т.к. исключение Q, на которое указывает говорящий, может быть важнее общего правила
вот только, <*только вот*>	**невозможны**, т. к. *вот* способствует фокусированию внимания на компоненте оценки, из-за чего он теряет статус слабого смысла	**не допускает**, т. к. в значение *только вот* и *вот только* входит оценка P как желательной ситуации, Q как нежелательной ситуации, которая может не нейтрализоваться контекстом	**возможно**, т. к. исключение Q, специальное внимание к которому привлекает говорящий, может быть важнее общего правила

Литература

1. *Апресян В. Ю.* Уступительность в языке // Лингвистическая картина мира и системная лексикография. Под ред. Ю. Д. Апресяна. М.: 2006. С. 615–712.
2. *Апресян Ю. Д.* Лексическая семантика (синонимические средства языка). 2-е изд., испр. и доп. М., 1995.
3. *Богуславский И. М.* Сфера действия лексических единиц. М.: 1996.
4. *Булыгина Т. В., Шмелев А .Д.* Языковая концептуализация мира (на материале русской грамматики). М.: 1997.
5. *Санников В. З.* Русские сочинительные конструкции. Семантика. Прагматика. Синтаксис. М.: 1989.
6. *Шмелев Д. Н.* О «связанных» синтаксических конструкциях в русском языке // Д. Н. Шмелев. Избранные труды по русскому языку. М.: 2002.

Статистические распределения слов в русскоязычной текстовой коллекции

Statistical distributions of words in a collection of Russian texts

Баглей С. Г. (baglei@galaktika.ru),
Антонов А. В. (alexa@galaktika.ru),
Мешков В. С. (meshkov@galaktika.ru),
Суханов А. В. (sukhanov@galaktika.ru)

Корпорация «Галактика», Москва

Изучение статистических свойств текстов является предметом большого количества работ в области прикладной математики и лингвистики. Подобно многим предыдущим исследованиям, в нашей работе мы придерживались рамок допущения о порождении текста, основанном на случайном процессе Бернулли. Развивая направление, мы исследовали статистические распределения отдельных слов в документах русскоязычной новостной коллекции текстов.

В статье описаны виды распределений слов, относящихся к различным частотным диапазонам. Для наиболее интересующих нас частотных уровней слов проведена аппроксимация графиков распределений, получены коэффициенты функций распределений и величины стандартных отклонений.

Мы рассматриваем полученные статистические данные как основу, на которую можно опираться для получения более реальной оценки вероятности появления некоторого слова в произвольном тексте русского языка. Данная оценка может использоваться для выявления адекватности соответствия слова некоторой текстовой коллекции.

1. Введение

Статистические и вероятностные распределения слов представляют достаточный интерес и являются предметом многих исследований в различных областях: криптографии, статистической теории игр, лингвистике, молекулярной биологии. Несмотря на довольно существенные различия между данными областями, общей для объекта исследования является модель порождения исследуемого текста. В зависимости от области знаний таким текстом может являться естественно-языковой текст, последовательность кодовых слов в цепочке ДНК или передаваемых криптографических данных.

В существующих работах, в основном, описываются два подхода к статистическому представлению текста: один из них основан на Марковском процессе порождения, другой — на модели Бернулли. Использование той или иной модели обусловлено различной природой рассматриваемых предметных областей. В каждом из случаев применения модели формирование данных подчиняется различным принципам и может соответствовать одному из вероятностных подходов в большей или меньшей мере.

В нашем случае предметом рассмотрения был большой набор текстов на русском языке. Каждый из текстов представлял собой реальное новостное сообщение. Мы выбрали для использования модель порождения текста Бернулли, в рамках которой вероятность отдельного испытания (порождения слова) не зависит от результатов остальных испытаний (порождения остальных слов в тексте). В отличие от данной модели, в Марковской цепи порядка N вероятность отдельного испытания зависит от N предшествующих событий.

Для обработки текстов на естественном языке модель Бернулли является более подходящей с точки зрения удобства представления об их формировании. Данная модель лежит в основе многих известных вероятностных алгоритмов обработки текста, таких, как, например, байесовский классификатор текстов [McCallum, 1998.]. При этом, в областях,

не связанных с естественно-языковой обработкой, существует достаточное количество методов, работа которых основана на Марковской модели порождения текста [Schbath, 2000].

Отличие нашего подхода от большинства исследований, опирающихся на модель Бернулли, состоит в том, что мы рассматриваем в качестве некоторого конечного текста не всю текстовую коллекцию, а отдельный документ, являющийся частью набора. Иначе говоря, коллекция в этой модели представляет собой не единый «мешок слов», а объединение нескольких «мешочков». Преимуществом использования данной модели является возможность исследования распределений слов и их частотных характеристик во всем наборе новостных текстов, с учетом элементов набора — отдельных документов.

В своей работе мы попытались получить ответы на некоторые вопросы — например, о том, какие виды распределений наиболее характерны для слов, составляющих большое множество реальных текстов на русском языке. Нам представляется возможным определение достоверности соответствия некоторого подмножества текстов всей текстовой коллекции, основываясь на сравнении соответствия распределения слов в данном подмножестве распределениям этих же слов в коллекции.

Одна из проблем в области корпусной лингвистики, обсуждение которой натолкнуло нас на проведение исследования, и решение которой не удалось определить на «круглом столе» конференции «Диалог'2008», была связана с определением «выбросов» в текстовых корпусах и формулировалась следующим образом. Какая выборка текстов представляется более адекватной в качестве контекста употребления некоторого слова: та, где некоторый термин встречается по 3 раза в 7 текстах, или та, где он встречается по 7 раз в 3 текстах? Данная работа представляется первым этапом для создания инструмента, который может помочь дать ответ на этот вопрос.

2. Распределения слов в текстах

В известных работах в области теории информации [Bayesa-Yates, Ribeiro-Neto, 1999] показано, что частотное распределение слов в большом множестве текстов имеет степенной вид, K/j^{Θ}. Закон Ципфа [Zipf, 1949], описывающий соотношение частоты встречаемости слова и его ранга, в свою очередь, является частным случаем степенного закона распределения, при котором значение Θ близко к единице. В модели большого текстового массива, которую описывает закон Ципфа все множество слов, входящих в коллекцию, рассматривается, как единое целое.

В рамках данного представления проведены исследования [Reinert, Schbath, Waterman, 2000], [Rennie, 2005] о распределениях слов в последовательностях большой размерности n. Было установлено, что распределение частоты встречаемости $N(w)$ некоторого слова w в общем случае подчиняется Гауссову (нормальному) закону распределения. Вместе с тем, оказалось, что такое распределение не наблюдается для тех слов, математическое ожидание появления которых $(n-l+1)\mu(w)$ очень мало. Другими словами, нормальное распределение не наблюдается для редких слов w. Было обнаружено, что более подходящим распределением для таких слов является распределение Пуассона. В качестве объяснения подобного различия в виде распределений можно принимать свойство независимых переменных Бернулли, сумма которых, в зависимости от асимптотического поведения предполагаемого значения, может быть аппроксимирована и в виде нормального распределения, и распределения Пуассона.

Использованная в вышеуказанных работах модель «мешка слов» позволила выявить характеристики распределений, важные для исследований. При этом можно отметить, что за рамками исследований остались характеристики встречаемости слов от текста к тексту внутри коллекций. С точки зрения информационного массива, полностью соответствующего модели Бернулли, данная необходимость, возможно, и отсутствует. Вместе с тем, известно, что естественно-языковые тексты соответствуют модели Бернулли с учетом некоторого приближения. Как правило, слова в текстах на естественном языке распределены неравномерно и исследование подобных «неравномерностей» представляет собой отдельную задачу для изучения.

Распределения слов в отдельных текстах, составляющих текстовые коллекции, также исследовались ранее, например, в следующих работах: [Gotoh, Renals, 2003], [Blake, 2006]. Полученные данные аппроксимировались авторами в различных пространствах, среди прочих, и в пространстве Ципфа «ранг-документ». Было установлено, что у часто используемых общеупотребительных слов распределение частот их встречаемости имеет биномиальный вид. При этом, распределения по документам редких слов, обладающих при этом достаточной статистической базой для построения таких распределений, имеют вид распределений Пуассона, так же, как и при использовании модели единого «мешка слов». Статистическая база для исследований представляла собой в первом случае подборку расшифровок звуковых новостных сообщений объемом 2583 документа, во втором — три словарных корпуса на английском языке объемом, соответственно, 1400, 100 000 и 1 000 000 статей.

Опираясь на результаты, полученные в данных работах, мы исследовали характеристики распределений на большом массиве текстов русского языка. Для нас представляли интерес как задача выявления распределений в применении к текстовому масси-

ву на русском языке, так и анализ распределений на большом массиве данных. Кроме того, нашей целью было получение значений аппроксимирующих функций для проведения дальнейшего исследования в рамках поставленной задачи сравнительного анализа распределений слов в различных текстовых массивах.

3. Результаты экспериментов

Для проведения экспериментов была взята база, содержащая коллекцию новостных сообщений на русском языке за 14-летний период, с 1995 по 2008 год. Основные характеристики базы:

- объем базы — 14,5 миллионов документов;
- средняя длина документа в базе — 503 слова;
- общее количество словомест — 7,3 миллиарда;
- объем словаря базы — 15,6 миллиона слов;
- количество источников-СМИ — около 2 тысяч.

Из всего словаря базы мы отобрали список слов, количество словомест которых превысило 500 вхождений. Данное ограничение было использовано для того, чтобы статистических данных употребления слова было достаточно для построения соответствующего распределения. Отобранные 122,5 тысячи слов были упорядочены по частоте их встречаемости в коллекции.

Далее для каждого слова был сформирован вектор количеств документов, в которых это слово встречается 1, 2, 3, …, 255 и свыше раз. Каждому классу документов в векторе был присвоен свой вес — коэффициент, равный количеству вхождений в него заданного слова. Все случаи, при которых слово встречалось в документах 255 раз и более, были объединены в один общий класс документов, ввиду его априорно достаточно низкого веса. При подсчете количества словоупотреблений мы учитывали все различные морфологические формы, в которых может употребляться некоторое слово.

В результате присвоения весов по количеству вхождений, мы получили для каждого слова еще один взвешенный вектор распределения слова по документам. После получения всех значений была проведена нормировка обоих сформированных векторов.

На рис. 1 в качестве примера приведены распределения значений, полученных для взвешенных векторов слов «прямота», «значение», «быть». Слова были выбраны из различных частотных диапазонов употребления в коллекции: количество словоупотреблений для данных слов отличается от одного к другому на 1,5–2 порядка.

Анализ полученных результатов показал, что распределения по документам группы наиболее частотных слов в коллекции графически соответствуют

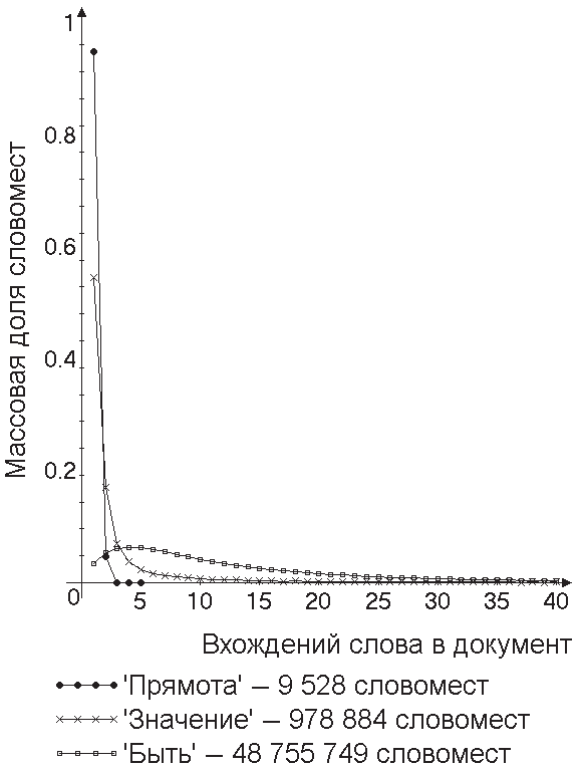


Рис. 1. Распределения взвешенных векторов частот некоторых слов

виду распределения Пуассона с *параметром* λ, значение которого находится в диапазоне 3,5…6,5. При этом усредненное значение количества появлений таких слов в некотором отдельно взятом документе базы также соответствует параметру распределения Пуассона λ, который можно представить в виде математического ожидания количества появлений некоторого слова в документе [Вентцель, 2001]

Распределения прочих слов, не относящихся к наиболее частотным, интересовали нас главным образом на начальном участке координаты вхождений слова в документ. Поскольку на данном участке содержится подавляющая часть значений всего распределения, точность аппроксимации именно в этом диапазоне представлялась наиболее важной с точки зрения практической применимости результатов.

Нами было замечено, что для не самых частотных слов функция распределения Пуассона на начальном участке не обладает достаточной точностью и имеет существенное отклонение в некоторых диапазонах значений. Мы попытались подобрать более подходящую аппроксимирующую функцию, для чего видоизменили график значений распределения частот. Вектор значений частот слов был преобразован к накапливающемуся виду: каждое следующее значение функции стало включать в себя сумму всех предыдущих ее значений.

Далее, чтобы проанализировать влияние частоты словоупотребления на изменение параметров функции, мы: — упорядочили все выбранные слова базы, основываясь на частоте их появления в кол-

лекции; — выделили ряд произвольных интервалов в полученном списке, принимая в качестве критериев выделения следующие условия. Во-первых, диапазон должен иметь статистическую представительность, достаточную для построения распределения. Во-вторых, частоты употребления слов, находящихся на нижней и верхней границах диапазона, не должны существенно отличаться. Допустимый порог для второго из условий составил около 5%. Учитывая данные условия отбора, интересующие нас диапазоны частот соответствовали уровню 10 000, 100 000, 150 000, 500 000, 1 000 000 вхождений слов; — для каждого из интервалов в упорядоченном списке выбранных слов базы была отобрана достаточно объемная выборка слов — по одной тысяче представителей, находящихся в вышеуказанных частотных диапазонах. Чтобы обеспечить равномерность распределения и сгладить погрешности при возможных выбросах, слова выбирались подряд, только с учетом их ранга в упорядоченном списке. Чтобы минимизировать влияние на статистические характеристики выбранного набора слов, мы не учитывали разбиение по прочим признакам — частям речи, ролям в предложениях, формам словоизменения.

После проведения отбора слов мы получили усредненные характеристики распределений слов для заданных частотных интервалов. Мы провели сравнение нескольких методик усреднения и остановились на среднем арифметическом. В отличие от среднего гармонического, среднее арифметическое не вносит своей погрешности в точках нулевых усредняемых значений. По сравнению со средним степенным, среднее арифметическое достаточно удобно с точки зрения его вычисления.

Таким образом, для каждого из списков в 1000 слов были найдены средние арифметические значения элементов взвешенных векторов частот. На рис. 2 приведены графики распределений средних значений по всем исследованным нами интервалам. Для сравнения приведен график, полученный аналогичным образом, для верхушки из 50 слов в коллекции, частота которых превысила 12 миллионов словомест.

Далее с помощью регрессионного анализа значений мы аппроксимировали полученные графики, используя функцию степенного вида:

$$y = 1 - \frac{k}{j^{\theta}} \qquad (1)$$

Выбор функции данного вида был сделан, во-первых, исходя из небольших коэффициентов стандартных (среднеквадратических) отклонений, полученных при аппроксимации. Во-вторых, мы стремились к наибольшей точности аппроксимации на начальном участке графика, так как подавляющая доля значений была представлена именно в этом диапазоне. Пример сравнения графиков значений распределения и аппроксимирующей функции приведен на рис.3.

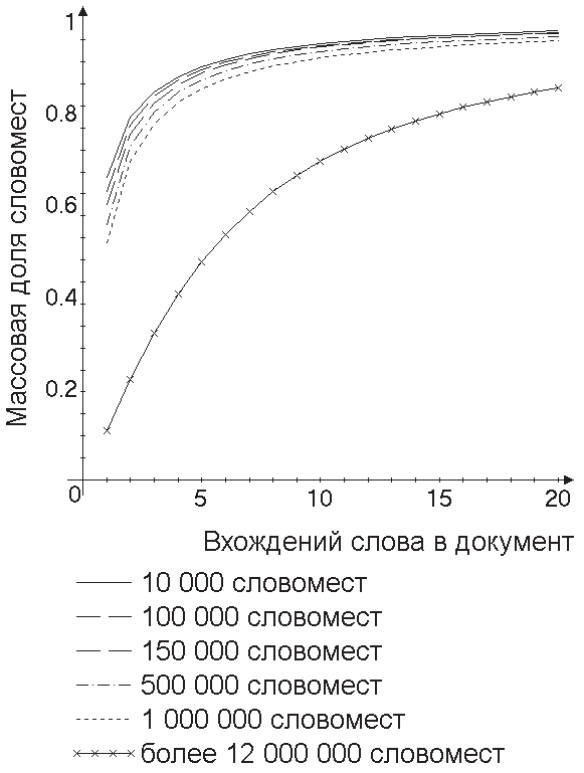


Рис. 2. Распределения накапливающихся взвешенных частот для подмножеств из 1000 слов в некоторых частотных диапазонах

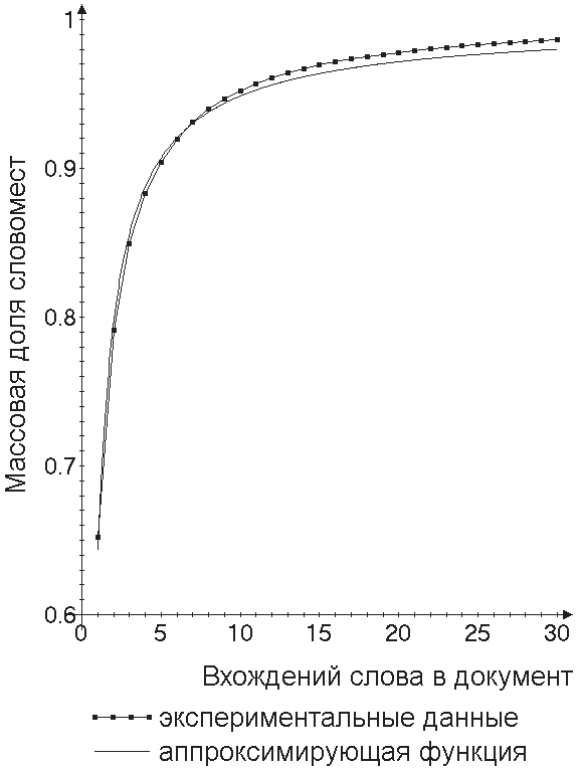


Рис. 3. Аппроксимация распределения накапливающихся взвешенных частот для 1000 слов из диапазона в 50 000 словомест

Аналогичным образом были рассчитаны аппроксимирующие функции для остальных выбранных диапазонов частот. Графики полученных функций приведены на рис. 4.

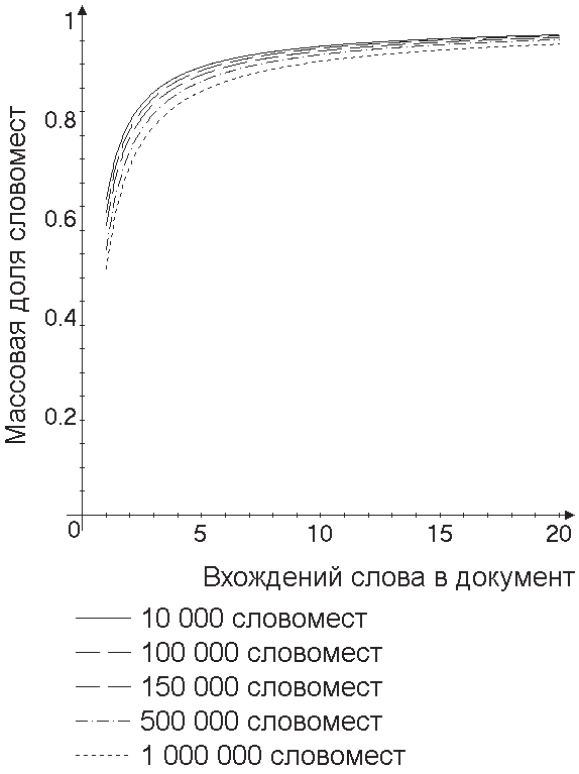


Рис. 4. Аппроксимация распределения накапливающихся взвешенных частот для 1000 слов из заданных диапазонов

В таблице 1 приведены значения коэффициентов функций степенного вида (1), полученных с помощью регрессионного анализа. Коэффициент *s* представляет собой значение стандартного (среднеквадратического) отклонения каждой из функций.

Таблица 1. Значения коэффициентов аппроксимирующих функций

Уровень частот слов	*k*	*Θ*	*s*
10 000	0,370	0,854	0,0109
50 000	0,369	0,857	0,0084
100 000	0,397	0,851	0,0096
150 000	0,424	0,839	0,0106
500 000	0,473	0,838	0,0112
1 000 000	0,512	0,794	0,0101

Очевидно, что коэффициент *k* степенной функции заметно возрастает при увеличении частоты слов в базе. Коэффициент *Θ* при этом убывает. Из полученных данных можно заметить, что для заданных диапазонов частот степенная функция соответствует распределению взвешенных значений частот с приемлемым уровнем отклонений.

4. Выводы

На основе полученных в ходе экспериментов данных можно сделать следующие основные выводы:

- мы получили усредненные параметры распределения некоторых частотных диапазонов слов в большой новостной русскоязычной коллекции текстов;
- на основе анализа списка слов, упорядоченных по частоте употребления в базе, был выделен диапазон частот словоупотреблений, свойственный значимым словам, несущим информацию о различных предметных областях, и имеющий достаточную статистическую базу для построения распределения;
- для некоторых уровней в рамках выделенного диапазона была проведена аппроксимация значений распределения, с помощью которой были найдены коэффициенты функции распределения словомест для заданной частоты словоупотребления в базе.

Полученные результаты не являются окончательными для данной работы. Мы планируем базироваться на них в своих будущих исследованиях, которые можно условно разделить на два этапа.

На первом этапе мы планируем получить более реальную оценку вероятности появления слова в произвольном новостном тексте на русском языке. На втором этапе — использовать функции распределения, полученные в данной работе в сочетании с оценками вероятностей появления слов для задачи расчета оценки соответствия этих слов некоторой произвольной коллекции текстов. Одной из составляющих перечисленных этапов работы может являться исследование влияния грамматических частей речи, а также, возможно, некоторых других признаков слов, на характеристики распределений в текстах.

На наш взгляд, основной целью подобного исследования может являться создание инструмента для получения сравнительных характеристик распределений слов в текстовых массивах. Полученная информация подобного рода может быть применена для выявления наличия или отсутствия аномалий в статистике употребления отдельно взятых слов или словарных групп на основе анализа их распределений в различающихся коллекциях.

Умение правильно использовать такие данные может стать ключевым с точки зрения решения задачи определения приоритета значимости слов, пример постановки которой был приведен во введении данной работы.

Литература

1. *Blake, C.* A Comparison of Document, Sentence, and Term Event Spaces // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006. Pages: 601–608.
2. *Régnier, M.* A Unified Approach to Word Occurence Probabilities // Discrete Applied Mathematics, 2000. Volume 104, Issue 1–3, Pages: 259–280.
3. *Reinert, G., Schbath, S., Waterman, M.* Probabilistic and Statistical Properties of Words: An Overview // Journal of Computational Biology, 2000. Volume 7, Number 1/2, Pp. 1–46.
4. *Schbath, S.* An Overview on the Distribution of Word Counts in Markov Chains // Journal of Computational Biology, 2000. Volume 7, Number 1/2, Pp. 193–201.
5. *Rennie J.* A Better Model for Term Frequencies // 2005.
6. *Zipf, G.* Human behaviour and the principle of least effort // An introduction to human ecology, 1949. 1st edn., Addison Wesley.
7. *Вентцель, Е.* Теория вероятностей // 7-е изд. стер. М.: Высшая школа, 2001. Т. 2, с. 106–115.
8. *Baeza-Yates, R., Ribeiro-Neto, B.* Modern Information Retrieval // Addison Wesley, 1999.
9. *Gotoh, Y, Renals, S*. Statistical Language Modelling // Lecture Notes in Computer Science, 2003. Springer, Volume 2705, Pages: 78–105.
10. *Régnier, M., Denise A.* Rare events and conditional events on random strings // Discrete Mathematics and Theoretical Computer Science, 2004. Vol. 6, n°2, Pages: 191–214.
11. *Church, K., Gale, W.* Poisson Mixtures // Journal of Natural Language Engineering, 1995.
12. *McCallum, A., Nigam K.* A Comparison of Event Models for Naive Bayes Text Classification // In AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41–48. Technical Report WS-98-05. AAAI Press. 1998.

О некоторых семантических коррелятах формального варьирования идиом (операция замены)[1]

Semantic correlates of formal variation in the field of idiomatics (the operation of substitution)

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

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

В докладе обсуждается проблема формального варьирования идиом. Основной предмет анализа — операция замены компонентов идиомы. Разработана классификация типов замены, основанная на эмпирическом исследовании материала. Основное теоретическое допущение заключается в том, что формальное варьирование сопровождается относительно регулярными преобразованиями семантики идиом, причем в реальном употреблении эти семантические модификации выполняют в тексте определенные дискурсивные функции. В докладе показано, что одна из важнейших дискурсивных функций операции замены — языковая игра.

1. Введение в проблему

Вариативность представляет собой весьма распространенное явление в системе языка. Она проявляется на всех уровнях языковой системы — от фонетики и морфологии до варьирования макроструктуры текста. Описание варьирования языковых выражений часто оказывается затрудненным из-за неясности соотношения варьируемых форм с изменениями в их плане содержания, а также из-за сложности определения базовой формы, от которой, собственно, и следовало бы «отсчитывать» варьирование. Особенно сложным описание варьирования оказывается для сферы фразеологии. Из-за особенностей устройства фразеологизмов — формально они выглядят как словосочетания и предложения, а содержательно являются единицами словаря — варьирование идиом, коллокаций и пословиц столь разнообразно, что в ряде случаев неясно, имеем ли мы дело с варьированием одного фразеологизма или речь идет о двух (и более) разных фразеологизмах. Ср., например, случаи типа *вынуть сердце — вынуть душу; стоять за спиной — стоять за плечами; как с неба свалиться — как с луны свалиться.*

Формальные преобразования могут быть описаны как операции, применяемые к стандартной форме идиомы, в результате которой появляется ее модификации. Несмотря на то, что формальное варьирование во фразеологии изучается уже очень давно (см., например, [Абрамович 1964; Диброва 1981; Лавров 1983]) и даже имеются специальные словарные издания, ставящие задачу исчисления способов варьирования фразеологизмов [Мелерович, Мокиенко 1997], связь между формальным варьированием и его влиянием на семантику остается не раскрытой в должной мере. Кроме того, важно установить и дискурсивную функцию операции замены. Иными словами, что это — ошибка в речи в условиях дефицита времени, языковая игра или попытка скрытого воздействия на собеседника? Данный доклад посвящен описанию одного из фрагментов этой многообразной системы варьирования — операции замены компонента идиомы — и тем следствиям в плане содержания идиомы, к которым она приводит, а также

[1] Работа выполнена при поддержке гранта РГНФ 07–04–12117в

дискурсивным функциям формального варьирования указанного типа[2].

2. Операция замены: подобие мишени замены источнику замены

Операция замены может быть охарактеризована по-разному. Например, с точки зрения мишени замены (МЗ)— того, что заменяется (словосочетание, слово, часть слова), с точки зрения источника замены (ИЗ) — того, чем заменяется мишень замены. Определенный интерес представляет анализ синтаксических и морфологических свойств компонентов идиом — как заменяемых, так и тех, которые используются для замены. Однако куда более интересным оказываются не чисто формальные характеристики, а выявление семантических отношений между объектом и источником замены. В самом общем случае МЗ и ИЗ могут быть связаны отношениями семантического **подобия** и **различия**.

В имеющемся материале отношение подобия реализуется в следующих вариантах.

2.1. Синонимическое преобразование МЗ

В качестве ИЗ для объекта замены выступает его полный или частичный синоним. Это наиболее очевидное выражение отношения семантического подобия между МЗ и ИЗ, ср. примеры группы (1):

(1) **а**. *Впрочем, выставка, прошедшая тут недавно, делалась совсем не для того, чтобы* ***бросить еще один булыжник*** *[вместо: бросить камень] в несчастного, из которого бесчеловечная эпоха сотворила миф — восковую фигуру с заострившимся носом, провалившимися щеками и яростно горящим слепым взором. [Столица.]* **б**. *У детей* ***тормоза*** *страха и совести* ***работают*** *еще* ***плохо*** *[вместо: тормоза отказали], и поэтому даже матерые волки преступного мира побаиваются попасть в лапки расшалившихся детенышей. [Столица]* **в**. *Очи, за реку глядя, проглядел я все, право, / Хоть бы пиктограмму он предпослал, / родственник милый, / Так-то, мол, и так-то, такие-де у меня планы, / Иначе же и кинуться куда — ума не приложишь, / Вечно с дядей таким* ***шиворот-наоборот*** *[вместо: шиворот-навыворот] происходит. [Саша Соколов]*

Дискурсивный потенциал этого преобразования четко выделить не удается, поскольку во многих случаях процедура синонимической замены может интерпретироваться адресатом просто как невольная оговорка — так часто бывает и на самом деле. Однако иногда возникает отчетливый игровой эффект. Интересно, что чем менее тривиален синоним, тем больше вероятность игровой интенции говорящего, ср. выше контекст (1в), в котором в качестве ИЗ используется не вполне тривиальный синоним. Ср. также (1г, д):

(1) **г**. *Дерзай и ты, мое непритязательное перо, воздай обольстительнице за содеянное:* ***сорви*** *с нее, образно говоря,* ***фарисейские облачения*** *[вместо: сорвать покровы] <…> [Саша Соколов]* **д**. *При этом так ругался по-латыни, / Что скифы эти* ***корчились в гробах***. *[вместо: переворачиваться в гробу] [В. Высоцкий]*

Важно отметить, что дискурсивные функции операции замены на синоним (и ряда других видов операции замены) в определенном смысле противопоставлены категории значимого варьирования [Баранов, Паршин 1986], позволяющей по-разному описывать одну и ту же ситуацию (ср. *бутылка наполовину полна* vs. *бутылка наполовину пуста*). Дело в том, что значимое варьирование как средство скрытого воздействия на адресата предполагает не только аспект различения (то есть внесение смысловых различий в варьируемые формы), но и аспект отождествления — принципиальную возможность отождествления варьируемых форм. Только в этом случае варьирование оказывается инструментом скрытого воздействия. Однако аспект отождествления при варьировании формы идиом существенно ослаблен, поскольку базовая, исходная форма идиомы в большинстве случаев фиксирована в сознании носителя языка, входя в его «ментальный лексикон». Иными словами, замена слова в идиоме на синоним практически всегда воспринимается как отклонение от стандарта, что затрудняет возможность отождествления варьируемых форм как «одного и того же». Следовательно, и дискурсивный потенциал такого вида варьирования как средства скрытого воздействия весьма и весьма ограничен.

Иногда игровой эффект возникает при необходимости замены слова в идиоме на квазисиноним для сохранения рифмы, ср. характерный пример из В. Высоцкого: Но если ***туп, как дерево*** [вместо: *туп как бревно, туп как пень*] — родишься баобабом / И будешь баобабом тыщу лет, пока помрешь. Впрочем, в последнем случае ИЗ лексема *дерево* (или *пень*) семантически поддержана материализацией метафоры — рассуждениями о баобабе.

Среди случаев синонимического преобразования МЗ (также с игровым эффектом) можно рассмотреть сравнительно редкий вариант использования

[2] В качестве источника использовалась база данных по современной идиоматике, включающая более 50 тыс. контекстов употребления идиом в художественной литературе и публицистике 60–90 гг. База данных разработана в Отделе экспериментальной лексикографии Института русского языка РАН.

в идиоме слова другого языка, аналогичного по значению МЗ, ср. ***Кто есть ху,*** *как сказал Горбачев* (заголовок) [*Известия*].

2.2. Замена на «функциональный» синоним

Эта группа преобразований непосредственно примыкает к предшествующей. В качестве ИЗ выступает слово, явно не синонимичное МЗ в общем случае, но аналогичное ему по функции в данной ситуации. Ср. (1ж) Чтобы стать цивилизованными не по наряду, не на словах, а на деле, нужно прежде всего сбросить эту гирю, по капле ***выдавить из себя «совка»*** [вместо: *выдавливать из себя раба*] — иждивенца и нытика, волевым усилием попытаться, по крайней мере, перестроить собственную ментальность, не щадя и того, что глубоко в нас засело, настолько глубоко, что воспринимается уже либо как знак национальной идентичности, либо как черта характера. *[Столица]*

Дискурсивные функции замены такого рода аналогичны синонимическому преобразованию МЗ, рассмотренному выше.

2.3. Стилистическая модификация МЗ

В качестве ИЗ выступает слово, находящееся в другом стилистическом регистре — стилистический синоним:

(2) **а**. *Ты надежды на скорый отход мой покинь, еще помаячу там-сям, еще* ***помозолю гляделки*** *[вместо: мозолить глаза] некоторым чуток, постою над душой у некоторых, послезит еще к небу бельмо мое. [Саша Соколов]* **б**. *И сказал Господь: «Да* ***восчешутся руки*** *[вместо: руки чешутся] мои, да возложутся на ребра твои, и сокрушу я их». Так и с недругом будет моим! — мне врач обещал, что к четвергу так и будет. [В. Высоцкий]* **в**. *И классику опять охота надеяться. Во-первых, на то, что читатель кинется изучать его классические первоисточники. Во-вторых, что в этой связи можно будет отлично* ***выспаться на лаврах*** *[вместо: почивать на лаврах]. [М. Мишин]*

С дискурсивной точки зрения примеры данного типа по большей части носят явно игровой характер, поскольку очевидно, что выбор стилистического синонима для замены в идиоме — это не случайная ошибка, вызванная дефицитом времени в языковом оформлении своей мысли, а намеренное преобразование базовой формы идиомы в стремлении создать эффект «мерцания» семантики: от базовой формы к модификации и наоборот. Причем не следует думать, что во всех случаях такая игра преследует какие-то глубокие цели речевого воздействия. Часто эффект мерцания и становится самоцелью, поскольку и в этом случае достигается важнейшая особенность языковой игры — создание нескольких слоев смысла высказывания.

Некоторым умельцам удается достичь аналогичного игрового эффекта и при замене одного слова сниженного регистра речи — на другое с такой же сниженной стилистической характеристикой, ср. контекст из Саши Соколова: Верим-верим, ты у нас марафонить известный мастак, вон мослы-то себе отрастил — первый сорт, и сухие и долгие, нам ли с нашими ***бестолковыми моськами в калашный ряд*** [вместо: *со свиным рылом*], а тем паче Илье-Безобразнику.

2.4. Словообразовательная модификация МЗ

В этот раздел попадают, в частности, те случаи, когда источник замены представляет собой словообразовательный вариант МЗ, образованный, например, с помощью уменьшительных суффиксов, ср. (3а, б)

(3) **а**. *Весь вывернусь наружу я — / И* ***голенькую правду*** *[вместо: голую правду]/ Спою других не хуже я / Про милое оружие… [В. Высоцкий]* **б**. *У детей тормоза страха и совести работают еще плохо, и поэтому даже матерые волки преступного мира побаиваются* ***попасть в лапки*** *[вместо: попасть в лапы] расшалившихся детенышей. [Столица]*

Дискурсивный потенциал варьирования такого рода весьма разнообразен — от явно игрового употребления в (3а), до семантически мотивированного употребления *лапки* — вместо *лапы* — по отношению к «детенышам». Но опять-таки использование данного способа для скрытого речевого воздействия маловероятно.

2.5. МЗ и ИЗ — элементы отношения ‘род vs. вид’, ‘вид vs. род’

Семантическое подобие между элементами родо-видового отношения обеспечивается нахождением в одном таксоне семантического тезауруса, ср. (4), (5)

(4) [замена вида на род] О, Марина, ***первая птичка*** *[вместо: первая ласточка] Запада, залетевшая по запаху на оттепель в наш угол! [В. Аксенов] Но идеал* ***связать не мог*** */ В археологии* ***двух строк****, [вместо: не мочь связать двух слов] — / И Федя его снова закопал. [В.Высоцкий]*

(5) [замена рода на вид] **а**. *Но вы сами **извилиной пошевелите** [вместо: шевелите мозгами], куда я с данной вакансии соскочу, где еще дармового горючего вам всем неприкаянным, нацежу. [Саша Соколов]* **б**. *Нет, я лучше — от и до, / Что и как случилось: / Здесь **гадючее гнездо** [вместо: змеиное гнездо], / "Юность", получилось. [В. Высоцкий]* **г**. *Показываются те стрекулисты, недомерки в шестнадцать мальчишеских, и ведут мою профурсетку на **лобную поляну** [вместо: лобное место] любви. [Саша Соколов]*

В качестве сравнительно редкого варианта встречается такой тип замены, при котором и МЗ, и ИЗ принадлежат одному таксону тезауруса, находясь на одном уровне абстракции, ср. (5) **д**. *Сказать после этого очень короткого представления: «Я не понимаю обэриутов» — способен разве что инопланетянин, **свалившийся с Золотой Звезды*** [вместо: *с Луны свалиться*] *имени Семена Бабаевского. Ведь А. Введенский — через Пономарева — все объяснил. Довел до ума — игривого, но все же не заигравшегося. [Столица]* Разумеется, в последнем примере замена далеко не симметричная, поскольку *звезда* выступает здесь не как небесное тело, а как наградной знак. Тем не менее, в процессе игрового употребления актуализуется и прямое значение слова *звезда*. Близкий способ игрового преобразования обсуждается ниже в 2.7.

2.6. МЗ и ИЗ — элементы отношения ‘объект vs. материал объекта’

Таких примеров почти нет, хотя очевидно, что такой тип замены семантически совершенно прозрачен. В эту группу попадает уже упоминавшийся выше пример (6) Но если ***туп, как дерево*** [вместо: *туп как бревно, туп как пень*] — родишься баобабом / И будешь баобабом тыщу лет, пока помрешь. *[В. Высоцкий]*. Однако такое отнесение правомерно, если *дерево* понимается как материал, а не как объект (этому противоречит последующая материализация метафоры).

2.7. МЗ и ИЗ — элементы отношений ‘объект X vs. части X-а’, ‘часть X-а vs. другие части X-а’

Ср. (7).

(7) **а**. *Он туда, сюда, уж и ЦК партии подключил — все-таки он генерал армии, а **не хвост собачий** [эвфемистическая замена обсценной идиомы не х.. собачий] — и, значит, в конце концов, своего добился, дали ему билет на какой-то несуразный, строго закрытый поезд специального назначения. [В. Пьецух]* **б**. *Не замечать за собой ничего дурного. Пусть **левая** твоя **ноздря не ведает, куда сморкнулась правая** [вместо: левая рука не ведает, что творит правая]. [В. Ерофеев]* **в**. *«А вы не **ошиблись** случайно **замком**?» спросил Модерати. «Не думаю», отвечал Палисандр. «Это ведь Мулен де Сен ЛУ?» [Саша Соколов]* **г**. *Не вызывает сомненья, что тут и спал, и слегка погодя, во избежание недомолвок с держащими власть, промыслил патент. Каким Макаром — **статья одиннадцатая** [вместо: статья десятая], но заказов набрал — кот наплакал, приходилось вертеться по местным составам и перелицовывать кой-что из старья. [Саша Соколов]* **д**. *Центр постарался сделать все возможное, чтобы в ходе подготовки к референдуму сыграть на патриотических чувствах людей и, образно говоря, **убить** сразу даже не **двух**, а **трех зайцев** [вместо: убить двух зайцев] [Столица]*

В примерах (7г,д) в качестве целого — X-а — выступает натуральный ряд.

Целое (X) может быть представлено и общим таксоном семантического тезауруса, в который входят мишень замены и источник замены, ср. (8)

(8) **а**. *Мой первый заместитель по "Щиту" полковник Кудинов обещает **съесть свою папаху** [вместо: съесть свою шляпу], если генерал-полковник Ачалов сумеет хоть раз подтянуться на перекладине… [Столица]* **б**. *…я **и пару ломаных юаней**, /будь я проклят, / За эти иксы-игреки **не дам**! [вместо: и ломаного гроша не дать] [В. Высоцкий]* **в**. *А каково чувствовать себя мышкой, к которой играет гигантская кошка — будь то ГАИ, паспортный стол и т.д. и т.п., — и метаться по неустроенной этой, несчастной стране, постоянно рискуя попасть в ловушку, постоянно чувствуя над собой занесенную когтистую лапу закона! Даже предполагая, в случае чего, избежать опасности, эту **лапу позолотив** [вместо: позолотить ручку]. Впрочем, это уже — следующая ловушка. [Столица]*

Отношение между элементами таксона X в примерах типа (8) не всегда симметричны. Так, хотя и *шляпа* и *папаха* входят в таксон ‘головные уборы’, *папаха*, конечно, менее стандартна как пример головного убора, чем *шляпа*. Аналогично, *юань,* с точки зрения носителя русского языка, менее парадигмальный пример денег, чем *грош*[3]. На этом, собственно и строится языковая игра. С другой стороны, в примере (8в) сам таксон (нечто вроде ‘ко-

[3] Не очевидно, что так будет и в будущем.

нечности людей и животных') достаточно уникален, что тоже служит неплохой основой для игрового экспериментирования.

2.8. МЗ фонетически подобна ИЗ

Как показывает функционирование языка, особенности устройства формы выражения отнюдь не автономный параметр языкового знака. Структурализм в этом смысле явно упрощал реальную ситуацию. Например, фактор рифмы оказывает существенное влияние на формирование устойчивых фразеологических оборотов [Баранов, Добровольский 2008]. С другой стороны, и форма печатного знака существенна для выражения смысла в культурной традиции: так, специфика иллокутивной семантики текста рекламы определяет репертуар гарнитур, которые могут быть использованы для передачи нужной семантики [Баранов, Паршин 1989]. Фонетическое уподобление широко используется как вариант операции замены при варьировании формы идиом, причем чаще всего с игровым эффектом, ср. (9).

(9) **а**. *Неожиданно Гурин произнес: — Сколько же они народу передавили? — Кто? — не понял я. — Да эти барбосы… Ленин с Дзержинским.* ***Рыцари без страха и укропа****… [вместо: рыцари без страха и упрека] [С. Довлатов]* **б**. *Ох, проявите интерес к моей персоне! / Вы, в общем, сами тоже — форменная соня: / Без задних ног уснете — ну-ка добудись, — / Но здесь сплю я —* ***не в свои сони не садись****! [вместо: не в свои сани не садись] [В. Высоцкий]* **в**. *Признайтесь-ка, кстати, скольких вы совратили, бесчестный оборотень. Доверьтесь, доверьте нам наше число по секрету. Исключительно антер ну — да ну же,* ***честное пенсионерское****, мы никому не скажем. [Саша Соколов]* **г**. *Не худой, все заметят, Илья себе челночек прикукобил, те так себе. А вы как думали, я скажу, полагали —* ***луком*** *я* ***шит*** *[вместо: не лыком шит], дулей делан? [Саша Соколов]* **д**. *Десяти лет не прошло, как Алексей Максимович ругательски ругал Ленина за злостные опыты над Россией, а уже он бичевал мягкотелую интеллигенцию и ее «****кочку зрения****» [вместо: точка зрения], восхищался темпами сноса Иверской часовни, укорял в мещанстве Канта, Толстого и Достоевского <…> [Столица]*

В отличие от синонимических и квазисинонимических замен, фонетическое уподобление вполне пригодно для речевого воздействия, поскольку сходство означающих МЗ и ИЗ предполагается как основное условие данного преобразования. Фонетическое сходство может использоваться для того, чтобы ввести МЗ и ИЗ в один ассоциативный ряд[4]. Последствия этого могут быть самыми различными — от снижения ценностного статуса МЗ (ср. *рыцарь без страха и упрека* vs. *рыцарь без страха и укропа, точка зрения* vs. *кочка зрения*[5]) до усложнения плана содержания языкового выражения за счет актуализации потенциальной омонимии (ср. *не в свои сони не садись* vs. *не в свои сани не садись*). В последнем случае возникающий второй смысловой план часто связан с предшествующей семантикой текста, с описываемой ситуацией, теми или иными прагматическими факторами коммуникативного взаимодействия и т. д.

Интересно, что иногда введение в ассоциативный ряд не требует фонетического уподобления: в этих случаях оказывается достаточным фактор устойчивости самой структуры идиомы, ср. *И страшно даже подумать, что было бы, если бы страна прислушалась к советам демократов типа Явлинского, и вместо того, чтобы принять Конституцию, передала бы «этому» парламенту функции Учредительного собрания. Тогда бы* ***узнали***, ***почем фунт Жириновского*** [вместо: *почем фунт лиха*] [Московский комсомолец].

Фонетическое подобие часто оказывается причиной ошибки, сбоя в коммуникации, который тем не менее может интерпретироваться другими участниками ситуации общения как языковая игра, ср. (10) Наиболее яркими <…> были стилистические погрешности и опечатки: «В октябре Мишутке кануло тринадцать лет…» (Рассказ «Мишуткино горе».); «Да ***будет*** ему ***земля прахом***! [вместо: *пухом*] — кончил свою речь Одинцов…» (Рассказ «Дым поднимается к небу».) *[С. Довлатов]*.

3. Операция замены: различие между источником замены и мишенью замены

Что касается отношения **различия**, то на этот вид замены приходятся только антонимические преобразования. Это объясняется тем, что незна-

---

[4] Подробнее о приеме «замазывания», или введения в отрицательно оцениваемый контекст/ассоциативный ряд, см. [Баранов 2007, с. 179 и далее].

[5] В данном случае варьирование совершенно сознательное. Противопоставление словарного и авторского употребления идиомы даже вынесено автором в название статьи, а статья эта — «О кочке и о точке» (1933) — принадлежит М. Горькому: «Есть кочка зрения и точка зрения. Это надобно различать ‹…› С высоты кочки не много увидишь. Точка зрения — нечто иное: она образуется в результате наблюдения, сравнения, изучения литератором разнообразных явлений жизни». Автор благодарит В. И. Беликова за данное уточнение.

чительная отклонения от семантики МЗ и ИС были уже описаны выше в отношениях сходства. Под **антонимическим преобразование МЗ** естественно понимать замену лексемы-МЗ на антоним в широком смысле, ср. (9а-г).

(9) **а**. *Я мажу джем на черную икру, / Маячат мне и близости и дали, — / **На жиже** — не на гуще мне **гадали**, — [вместо: гадать на кофейной гуще]/ Я из народа вышел поутру — / И не вернусь, хоть мне и предлагали. [В. Высоцкий]*
в. *Потому что потом у вас же и в остальных империях в результате все тех же противоречий сгорело всякого барахла на мильярды драхм: ипподромы и велодромы, кунсткамеры и рейхстаги, мосты и механические мастерские. А уж библиотекам **сам черт велел** [вместо: сам бог велел]: ведь — папирус. Отвлекитесь от ваших потусторонних забот и взгляните окрест: пепелища. [Саша Соколов]* **г**. *Погодите немного, говорил он [Член], построят тут продовольственные и другие культурно-просветительные учреждения, и тогда вас отсюда **палкой не загонишь** [вместо: палкой не выгонишь]. [А. Зиновьев]*

4. Заключение

Исследование варьирования в идиоматике не должно ограничиваться описанием и исчислением возможных форм модификации идиом. Большой интерес представляет семантика варьирования — те преобразования значения, которые сопровождают формальное варьирование, а также его дискурсивные функции. Операция замены компонентов идиомы демонстрирует разнообразие формальных вариантов, которые, однако, по большей части сводятся к игровым употреблениям (разумеется, небольшой процент приходится на ошибки и оговорки). Это объясняется в первую очередь тем, что наличие (в большинстве случаев) базовой формы, хорошо ощущаемой носителями языка, затрудняет полное отождествление вариантов и тем самым ограничивает возможность использования этого механизма для скрытого речевого воздействия. Однако это не является препятствием для порождения дополнительных «слоев» семантики, на которых и основывается феномен языковой игры. В вырожденном случае игра просто сводится к феномену «мерцания смысла» — от одного смыслового плана к другому и наоборот — даже при отсутствии семантических связей этих планов друг с другом.

Литература

1. *Абрамович И. М.* Об индивидуально-авторских преобразованиях фразеологизмов и отношении к ним фразеологического словаря // Проблемы фразеологии, Л., 1964.
2. *Баранов А. Н.* Лингвистическая экспертиза текста. М., 2007.
3. *Баранов А .Н., Добровольский Д. О.* Аспекты теории фразеологии. М., 2008.
4. *Баранов А. Н., Паршин П. Б.* Языковые механизмы вариативной интерпретации действительности как средство воздействия на сознание // Роль языка в средствах массовой коммуникации. М.: ИНИОН, 1986.
5. *Баранов А. Н., Паршин П. Б.* Воздействующий потенциал варьирования в сфере метаграфемики // Проблемы эффективности речевой коммуникации. М.: ИНИОН, 1989.
6. *Диброва Е. И.* Вариантность фразеологических единиц в современном русском языке, Ростов н/Д., 1981.
7. *Лавров Н. И.* Факультативные элементы фразеологического значения диалектных ФЕ // Актуальные проблемы русской фразеологии. Л., 1983.
8. *Мелерович А. М., Мокиенко В. М.* Фразеологизмы в русской речи. М., 1997.
9. *Крейдлин Г. Е.* Невербальная семиотика // М.: Новое литературное обозрение, 2002.
10. *Якобсон Р. О.* О лингвистических аспектах перевода // Вопросы теории перевода в зарубежной лингвистике. М.: 1978. С. 16–24.

Статистическая оценка функциональных свойств лексики по материалам интернета

Www statistical estimation of the functional properties of lexical items

Беликов В. И.
Институт русского языка им. В. В. Виноградова РАН, Москва

Ахметова М. В.
«Живая старина», Москва

В докладе рассматриваются возможности использовать статистику числа сайтов для объективной оценки устаревания лексики, ее территориального распространения и стилистического статуса. Индикатором функциональных различий лексических единиц служат расхождения в частоте их появления в текстовых массивах разных типов (классическая vs. сетевая литература, официальные тексты, блоги и т. п.), а также в сегментах Интернета с различной доменной или территориальной привязкой.

Принято считать, что норма, зафиксированная в толковых словарях, отражает немаркированный узус носителей литературного языка. Однако в действительности «кодификаторы ориентируются в первую очередь на собственный узус, во вторую — на узус своего круга, но лишь настолько, насколько этот узус пассивно знаком самим лексикографам ‹…› Словоупотребление “социально неблизких” высокообразованных слоев населения на узус лексикографов не влияет и никак не учитывается» [Беликов, в печати$_1$: 343–344]. То же касается жаргона и других типов ненормативной лексики. Остановимся на фактах, очевидных при объективном анализе реального узуса, но плохо соотносящихся со словарными описаниями лексики и фразеологии. Цель нашей статьи не в сборе очередной коллекции ошибочных интерпретаций, а в том, чтобы показать, что в настоящее время существуют легкодоступные и достаточно простые способы объективного выявления разнородных функциональных свойств лексики.

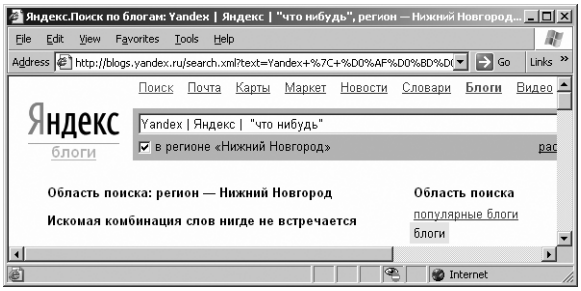
В силу ограниченности объема мы не будем типологизировать функциональные свойства лексики, лишь поясним, что речь идет о любых нетривиальных грамматических и стилистических особенностях слова или фразеологизма, упоминание которых оказалась бы полезным в словарной статье (оговоримся, что не всякая такая информация реально содержится в словарях). В рамках небольшой статьи удастся доказательно остановиться на интернет-верификации лишь некоторых типов функциональных свойств лексических единиц, так что текст невольно будет иметь несколько очерковый характер.

Охарактеризуем материал, **текстовые массивы Интернета**. Хорошим подспорьем для разного рода лингвистических разысканий стал Национальный корпус русского языка, но для детального исследования лексики его объема часто не хватает. Выручают некоторые сегменты Интернета, которые корпусами можно называть лишь метафорически. Корпуса формируются не стихийно, а создаются намеренно; их параметры задаются и контролируются, технические затруднения при поиске носят случайный характер и поисковые возможности при развитии корпуса могут меняться лишь в лучшую сторону. Не то с текстовыми массивами. Здесь мы знаем лишь самые общие характеристики содержащихся в них текстов, однако и их часто вполне достаточно. Важнейшими текстовыми массивами для работы с лексикой оказываются Библиотека Максима Мошкова (БМ) — большое собрание литературных и стилистически близких к ним текстов, и стихийно формирующаяся русскоязычная блогосфера.

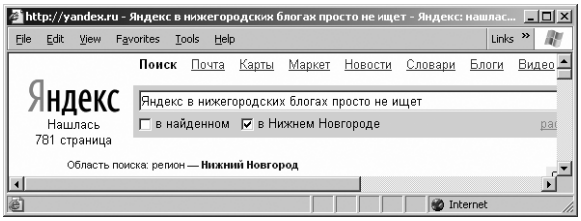
Важным достоинством БМ является ее разделение на несколько подмассивов, стилистика которых существенно различна. В первую очередь значимо противопоставление разделов «Собрание классики» (az.lib.ru), «Современная русская проза» (lib.ru/PROZA) и «Самиздат» (zhurnal.lib.ru). «Современность» в соответствующем разделе БМ трактуется достаточно широко и охватывает значительную часть советского периода. Собранием действительно современного профессионального литературного творчества является другой сегмент Интернета — «Журнальный зал» (magazines.russ.ru), где сосредоточены журнальные публикации с 1990-х гг.

«Самиздат» представляет собой очень большое по объему собрание современных самодеятельных произведений разного жанра, обычно невысоких художественных достоинств; многие авторы имеют достаточно смутные представления о литературной норме, доля разговорной и просторечной лексики в авторском тексте этого массива заметно выше, чем в только что упомянутых, поэтому в целом лексикон «Самиздата» близок к современному разговорному узусу.

Язык блогов во многом является отражением повседневного молодежного словоупотребления. Теоретически Яндекс допускает поиск в блогах с заданием отдельных параметров и их комбинаций: региона, пола и возраста блоггеров, а также с выделением конкретного фрагмента блогосферы (livejournal.com, diary.ru и т. п.). На практике же Яндексу не удаются поиски в блогах одного из крупнейших городов России.



Столь же безнадежны поиски чего бы то ни было у блоггеров Великого Новгорода, Нижнего Тагила, Нового Уренгоя и других городов, пишущихся с пробелом. Проблема теоретически вряд ли серьезна, поскольку при общем (не блоговом) поиске, где регион также можно ограничить, Яндекс без труда справляется с двусловными топонимами:



Не особенно дружественная политика Яндекса по отношению к любителям лексической статистики постоянно ужесточается. С самого начала появления поиска по блогам Яндекс вынуждал пользователей получать информацию блоками по 10 записей, уже года два, как он отказывается показывать 1001-ю и последующие найденные единицы, а с лета 2008 г. поиск с разделением по возрасту стал невозможен — со второй страницы результатов Яндекс сбивается с ограничения по возрасту и выдает лишь общий результат.

Не так уж редко кодифицированные **грамматические характеристики** отдельных единиц противоречат повседневному узусу.

Во всех толковых словарях фигурирует слово *корректив* (м. р.). Анализ текстовых массивов Интернета показывает, что если в 1930-е гг. и ранее *вносился корректив*, то к середине XX в. вноситься стали *коррективы*, в дальнейшем нейтрализация по числу способствовала смене рода на женский. К настоящему времени использование этого слова в единственном числе мужского рода стало безусловной архаикой, сейчас в словаре стоило бы писать: «**корректива**, *ж*. (*р е ж е* **корректив**, *м*.), обычно *мн*. …» [см. подробнее: Беликов, в печати$_2$].

Глаголы *лазить* (*лажу*, *лазишь*,…) и *лазать* (*лазаю*, *лазаешь*,…) всеми словарями признаются синонимичными и описываются обычно в одной статье; в московском словаре Шведовой [ТСРЯ] второй снабжается пометой *разг*., что имеет естественное объяснение: «на слух» *лазать* в Москве говорят заметно реже, чем в Петербурге. Но в действительности положение с отдельными словоформами этих глаголов различно. Судя по блогам (2007–2008), в петербургском узусе преобладают личные формы «от *лазать*», соотношение: *лазаешь/лазишь* — 52/30, *лазает/лазит* — 204/112; но с заметно более частотными инфинитивами положение обратное: 797/1120. В московских блогах преобладание «строго нормативного» инфинитива выражено очень явственно: за IV квартал 2008 соотношение: *лазить/лазать* составило 696/167, но с личными формами происходят странные вещи, за 2007–2008 гг. *лазаешь/лазишь* — 115/140, а *лазает/лазит* — 647/590. «Правильная форма» *лажу* быстро устаревает и не всегда используется даже в старшем поколении; показательна реакция одного известного ученого (не русиста), чл.-корр. РАН: «Говорю *лазию*, пишу *лазаю*». — «А *лажу*?» — «Ну, это просторечие, *Из кичмана не вылажу* какое-то».

У глагола *определиться* в ОШ было значение ‘определить своё местонахождение, положение’ с пометой *спец*. и примером *Лётчик определился с помощью приборов*; в обновленном издании [ТСРЯ] это значение дополнено: «…; вообще установить, решить что-н. для себя» и снабжено речением *О. в своих планах, целях, намерениях*. Это значение ввел в общерусский обиход М. С. Горбачев, у которого стандартным управлением было *определиться по чему-л*., однако в общем употреблении сразу стали конкурировать разнообразные модели (тип управления иногда связан с зависимым существительным, но во многих случаях взаимозаменяем [Беликов 2004]). В бумажных СМИ в самом начале 1990-х основной стала модель *определиться в чём-л*., но уже к середине 1990-х она стала вытесняться моделью *определиться с чем-л*., которая к настоящему времени оказалась вне конкуренции. Это легко выявляется в блогосфере, поиск на *определиться /+4 (планы | цели | намерения)* по 2008 г. выявил 157 случаев *определиться с* и 25 *определиться в*. «Исконное» *определиться по* с этими словами в блогах не встретилось, хотя в СМИ оно еще попадается: *Госу-*

дарству российскому необходимо четко определиться по своим внешнеполитическим целям и задачам («Время новостей»; 4.07.2006); *Г. Бежуашвили считает необходимым для обоих государств* [Грузии и Украины] *определиться по своим интеграционным планам* («Интерфакс», 2.04.2007). С другими управляемыми словами соотношение оказывается иным, но при возможности конкуренции «с-управление» заведомо преобладает, соотношение *с-* и *в-*управления при аналогичном поиске с набором *понятия* | *термины* | *определения* составляет 56/23, с теми же словами в ед. ч. — 22/2[1]. Указание на тип управления — необходимая часть словарной статьи глагола, но при этом должны быть проиллюстрированы все возможности, при ограничениях на объем единственно упоминаемым может быть только наиболее универсальное[2] и частотное. Резон для соединения «старого» и «нового» словоупотреблений в одно значение неясен; причиной могли, конечно, быть интроспективно полученные конструкты типа **Лётчик определился в своих планах с помощью приборов* или **Лётчик определился со своими целями по приборам*, но среди многих сотен просмотренных примеров сходные структуры не встретились.

Обращение к интернет-массивам позволяет довольно точно определить время и темпы конкретных словарных изменений. Изменения эти могут иметь разный характер: лексическая единица может «просто» устареть и выйти из употребления, может, наоборот, проявить территориальную или социальную экспансию, а может замениться другой, внешне сходной. Словарями эти процессы фиксируются далеко не всегда адекватно.

В словаре под редакцией Д. Н. Ушакова имелись слова *бестоварье* ‘недостаток, отсутствие товаров’ с пометой *газет.*, *ведьмак* ‘знахарь, колдун, оборотень’ с пометой *обл.* и *решебник* ‘учебное пособие, содержащее подробные решения задач, помещенных в каком-н. задачнике; ключ к задачнику’ с пометой *школьн. арго*. Последнее слово позднее в толковые словари не включалось, то ли как устаревшее, то ли как имеющее слишком узкогрупповое распространение. *Ведьмак* в достаточно объемном (130 тыс. слов) БТС отсутствует, но в новом БАСе представлен, по-прежнему с пометой *обл.* А *бестоварье* в двух названных академических словарях получило помету *разг.* Между тем *бестоварье* в БМ встречается лишь в 8 текстах, датированных 1917–1928 годом (три художественных: «Разгром» Фадеева, «Третья столица» Пильняка, «Возвращение Мюнхгаузена» С. Кржижановского), а также в два современных, но в обоих случаях цитируются документы той же эпохи. В блогосфере это «разговорное слово» появилось лишь четырежды, причем только в цитатах 1920-х гг. Что касается *ведьмака*, то обнаружившаяся в последние десятилетия всенародная любовь ко всему мистическому и потустороннему, сделала это слово чрезвычайно частотным: если в разделе «Классика» БМ оно встретилось лишь в 4 произведениях, то в современной литературной периодике («Журнальный зал» по 2008 г. вкл.) — в 19 текстах; в повседневном узусе ведьмаки упоминаются чаще многих общеизвестных политиков: в блогах за первую неделю декабря 2008 г. *ведьмаки* встретились 203 раза. Разумеется, никакого «областного» налета у этого слова уже нет. *Решебник* же достаточно давно стал упоминаться вне связи со школой, ср. у Аркадия Штейнберга о рыбаке: *Он перелистывает, как решебник, / Волну, волну… Ответа нет как нет* («Взморье», 1932); позднее это слово метафорически использовалось в стихах Борисом Слуцким и Сашей Соколовым. Что касается основного значения, то оно ежедневно по нескольку раз фигурирует в блогосфере, причем и в связи с вузовским обучением.

В словарях до сих пор отсутствует наречие *пешкодралом* и до самого последнего времени не было фразеологизма (*не*) *брать в голову*,[3] устаревающее *пешедралом* и практически вышедшее из употребления (*не*) *забирать в голову* не имеют в словарях ограничительных помет. *Пешкодралом* впервые появляется в конце XIX в., в современной профессиональной прозе *пешкодралом* и *пешедралом* равночастотны, в «Самиздате» преобладание нового варианта уже вполне выражено (55/35), а в блогосфере можно говорить о его победе (по 2008 г. вкл. соотношение составляет 2114/337). Экспансия нового варианта фразеологизма была гораздо более быстрой: он появился и победил в течение второй половины XX в.: в БМ-классике соотношение (*не*) *забирать* и (*не*) *брать в голову*, составляет 22/1 (единственный пример с *не брать в голову* находится в относительно случайной для этого раздела литературоведческой повести о Лескове Л. А. Аннинского «Несломленный»), а в «Журнальном зале — 3/150.

Детальный анализ блогов позволяет увидеть разницу в скорости лексических изменений в различных территориальных и социальных группах. Как известно, в настоящее время происходит переход от выражений типа *как нé_фига делать*, *мне это пó_фигу* к *как нé_фиг делать*, *мне на это пó_фиг*[4]; знак «_» символизирует здесь нерелевантность пробела в орфографической записи[5]. Подтверждается не только вполне естественное предположение, что инновация

---

[1] Примеры типа *определяться с семантикой основных понятий*, *определиться в значении базовых понятий*, естественно, в результат не включены.

[2] *Определиться в* несовместимо, например, с временными отрезками, ср. *определиться с отпуском* (**…в отпуске*).

[3] Впервые фиксируется в Словаре-тезаурусе… [2007].

[4] По материалам СМИ этот процесс описан в [Беликов 2008].

[5] В действительности носители новой модели существенно сильнее ощущают наречный характер выражения, что отражается в орфографии, ср. статистику по московским блогам 2004 г.: *по фигу*: 426, *пофигу*: 923; *по фиг*: 153, *пофиг*: 1667.

в бóльшей степени характерна для младших возрастов, но также выявляется такой нетривиальный факт: женщины используют новую модель чаще мужчин. Проиллюстрировать это придется лишь «архивными» поисками июня—июля 2008 г., (как говорилось, выявление возрастной статистики по блогам с середины 2008 стало практически невозможным). Вот данные по Петербургу, где смена *по_фигу* на *по_фиг* идет несколько быстрее, чем в Москве.

блоггеры Петербурга	**2005, весь год**		**2008, январь–май**	
	по_фиг(у)	в т. ч. *по_фиг*	*по_фиг(у)*	в т. ч. *по_фиг*
все блоггеры	1388	935 (67,4 %)	2529	1847 (73,0 %)
указавшие пол:				
женщины	655	454 (69,3 %)	1602	1189 (74,2 %)
мужчины	607	397 (65,4 %)	792	546 (68,9 %)
указавшие возраст:				
11–19 лет	34	23 (67,6 %)	640	484 (75,6 %)
20–30 лет	614	406 (66,1 %)	549	381 (69,4 %)
старше 30 лет	95	57 (60,0 %)	117	70 (59,8 %)

Среди прочего эта таблица иллюстрирует необщеизвестные факты о быстрых переменах в половозрастной структуре блоггерского сообщества: за неполных три года доля младшей возрастной группы выросла с 5% до половины, а женщины, лишь незначительно превосходившие мужчин, стали составлять 2/3 (при анализе стилистически различной лексики структура блоггеров может меняться существенно, но эти данные близки к средним показателям для единиц повседневного узуса). И не следует забывать, что возраст всегда указывает лишь около половины блоггеров, причем среди неуказавших доля женщин выше, чем в блогосфере в целом. Так — в Петербурге в конкретные годы; в других местах и в другое время состав блоггеров будет иным.

Кроме собственно лексических перемен, Интернет помогает отследить **изменения в сочетаемости**. Любители футбола заметили, что с недавних пор судьи все чаще не *назначают* пенальти, а *ставят*. В блогах обсчитывались вхождения глаголов *ставить/поставить* и *назначать/назначить* на расстоянии «/3» от *пенальти*. В 2006 г. из общего числа в 1106 таких контекстов с глаголом *(по)ставить* было 55 (5%), в 2007 — 7% от 1371, в 2008 — 15% от 2079. В течение 2008 г. поквартальный рост в процентах составлял 11–15–16–18. В некоторых контекстах постановка пенальти уже преобладает, так, соположение *поставить* vs. *назначить* и *«левый пенальти»* менялось следующим образом: 2006 — 3:13, 2007 — 3:4, 2008 — 8:3.

Похожими темпами идет экспансия глагола *(по)ставить* в спортивной прессе. В двух наиболее массовых изданиях, «Советский Спорт» и «Спорт-Экспресс», рост соотношения глаголов *(по)ставить* и *назначить* (*назначать*) в соответствующих контекстах был следующим: 1999–2000: 4/1091; 2001–2002: 13/1005; 2003–2004: 26/963; 2005–2006: 40/895; 2007–2008: 67/720. Поначалу *(по)ставить* обычно заключалось в кавычки (ср. *Гусев не «поставил» пенальти ни в наши, ни в их ворота, распределив все поровну* — Советский Спорт; 31.08.1999), но в последние годы кавычки редки, новая конструкция стала использоваться в заголовках: *На Украине за «локомотивский» фол против Кержакова пенальти ставят без раздумий* («На страже Родины», СПб; 14.05.2005), *В Манчестере такие пенальти не ставили* (Советский спорт; 15.05.2008).

Выражение *левый пенальти* появляется в СМИ в конце 1990-х (самый ранний пример в базе «Интегрум»: *В последнем матче, который якутяне выиграли у "Атома" из Железногорска, судьи дважды высосали левые пенальти, ни за что ни про что удалили нашего игрока с поля, но все равно «Автомобилист» сумел отстоять свое хрупкое преимущество и победил 3:2* — «Якутия», 29.11.1997), но употребляется очень редко. Статистика *левых пенальти* мизерна, но отмахнуться от нее нельзя: соотношение *поставленных* и *назначенных* по всем газетам таково: 1999–2004: 1/14, 2005–06: 5/10, 2007–08: 4/1.

Примеры на *(по)ставить штрафной* в прессе до сих пор единичны[6], но блоггеры так пишут все чаще: за 2005–2007 на *(поставить | ставить) /1 !штрафной* находится, соответственно, 1, 2 и 3 релевантных примера, а в 2008 — уже 15 (при 113 на *назначить /1 !штрафной*).

Выше речь шла о лексике, равномерно распространенной по русскоязычному пространству. Но существуют тысячи регионально маркированных лексических и фразеологических единиц различного стилистического статуса (нормативные, разговорные, жаргонные). Уточнения ареалов их распространения без обращения к различным сегментам Интернета невозможно.

Принято считать, что в паре синонимов *студень/холодец* (в знач. 'холодное кушанье из сгустившегося желеобразного мясного отвара с мелкими кусочками мяса') для Петербурга более характерно первое; это подтверждается и толковой лексикографией северной столицы: при немаркированном *студне*, *холодец* в МАСе имел помету *разг.*, а в БТС был «понижен» до *нар.-разг.* Между тем Интернет довольно отчетли-

---

[6] Один из наиболее ранних: *Все-то у него идет под классические 0:0, как вдруг на 95-й* ***ставят*** *ему* ***штрафной****, не смертельный, но сильно неприятный — эх, не всегда ладится законтачить с последним судейским поколением!* (Футбол-review; 16.06.1999).

во указывает на явную экспансию холодца в Петербурге; в младших возрастах в этом значении холодец уже более употребим. В блогах по окт. 2008 вкл. соотношение слов *студень* и *холодец* в Москве составляло 1772:**4634**, в Петербурге — **1330**:1026. Но при этом семантический шум для слова *холодец* в обоих городах равномерен и невелик, а за *студнем* в молодежном узусе скрываются также значения ‘студент’ и ‘студенческий билет’, при этом в Москве доля этих студенческо-жаргонных единиц сравнительно невелика, а в Петербурге составляет около половины всех студней. Результаты выборочного анализа словоупотреблений за 2005–2008 гг. в двух городах таковы:

	«студень», всего	*студень* ‘студент’	*студень* ‘студенческий билет’
Блоги Санкт-Петербурга:			
октябрь 2005	18	1	8
октябрь 2006	27	3	18
октябрь 2006	33	4	14
октябрь 2008	40	5	16
«всего»	118	13	**56**
Блоги Москвы (те же 4 месяца):			
«всего»	165	14	16

Имея в виду различия в объемах блогосферы двух городов, вполне очевидно, что центром распространения нового жаргонизма, особенно в значении ‘студенческий билет’, является Петербург.

Интернет-блоги во многих случаях являются наиболее эффективным инструментом выявления ареалов распространения регионально маркированных единиц: *чойс* ‘любая лапша быстрого приготовления’ и *оптарь* ‘оптовый рынок’ находятся только в омских блогах, *садоогород* ‘садово-огородное товарищество или участок в нем; используется и в официальных контекстах — почти исключительно в Удмуртии, *ссобойка* (также *собойка*) ‘набор продуктов на работу, в дорогу; школьный завтрак, взятый из дома’ — практически только в Белоруссии.

Однако в ряде случаев семантические и стилистические особенности лексической единицы таковы, что статистический анализ блогов не дает нужных результатов.

Недавний фразеологизм «с канцелярским оттенком» *мокрая печать* (поставленная непосредственно на документ, не ксерокопированная), на первый взгляд, распространен достаточно широко. Но анализ его появления в текстах, размещенных на разных доменах первого уровня, однозначно указывает на то, что фразеологизм привязан в первую очередь к Украине; вот статистика сайтов с релевантными документами[7] за отдельные годы:

Мокрая печать	…–2004	2005	2006	2007	2008	всего
Всего	44	36	68	164	450	762
В домене *.ru*	16	12	23	53	112	216
в том числе про Украину домене *.ru*	10	3	14	19	41	87
В домене *.ua*	7	13	22	62	187	291
В домене *.md*						11

Как видим, фразеологизм чаще появляется в русскоязычных текстах домена *.ua*, а в домене *.ru* значительная часть релевантных текстов касается Украины.

На другом стилистическом полюсе находится детская лексика: тут анализ только блогов дает очень приблизительный результат в силу достаточно редкого спонтанного появления соответствующих единиц в блогах и неравномерности распространения блогосферы по русскоязычной территории. Но совмещение анализа блогов с непосредственным опросом в Интернете позволяет делать достаточно четкие выводы. Ср., например, распределение синонимичных именований игры, самым распространенным названием которой является *вышибалы*: широко распространенное в недавнем прошлом название *круговая лапта* сохранилось только в Азербайджане, в обширном регионе от Красноярска до Хабаровска преобладает именование *выжигалы*, в Смоленской обл. — *высекалы* и т. п.[8]

В некоторых случаях региональное развитие блогосферы пока слишком незначительно, чтобы в Интернете можно было получить релевантную информацию. Так, слово *чебэшка* с орфографическими вариантами ‘дом, не полностью обеспеченный коммунальными удобствами; квартира в таком доме’ (от сокр. *ч/б* = частично благоустроенный) встретилось в пяти газетах (30 текстов), причем только из Якутии [Ахметова 2008], но анализ блогов практически ничего не дает: за 2007–2008 гг. слово *чебэшка* (*чэбэшка*, *чебешка* и т. п.) встретилось у 17 блоггеров в значении ‘черно-белая фотография, черно-белая фотопленка’ и лишь однажды в значении жилища — про якутский поселок Багатай, но у московского блоггера.

Приведенные выше примеры можно легко умножить. Интернет позволяет подтверждать или опровергать имеющиеся в толковых словарях сведения о функциональных свойствах лексики, выявлять новые, трудноуловимые традиционными методами особенности употребления слов и фразеологизмов.

Огрехи словарей прошлого вполне объяснимы, но в XXI веке лексикограф не имеет права работать по старинке и манкировать легкодоступными текстовыми массивами.

---

[7] Нерелевантными считались тексты с упоминанием *мокрой печати* как традиционного (не компьютерного) способа фотопечати.

[8] См. подобнее на форуме «Городские диалекты» — http://forum.lingvo.ru/actualthread.aspx?tid=110505.

Литература

1. *Ахметова М. В.* Региональная вариативность терминов, связанных с городской недвижимостью (по материалам электронной базы периодики «Интегрум») // Компьютерная лингвистика и интеллектуальные технологии. Вып. 7 (14). По материалам Международной конференции Диалог 2008. М.: ИПИ РАН, 2008.
2. *БАС:* Большой академический словарь русского языка / Гл. ред. К. С. Горбачевич. — М.—СПб: Наука. Тт. 1–9. А–Медь. 2004–2007.
3. *Беликов В. И.* «… Справочники устаревают и требуют корректив» // Материалы XIV Филологических чтений имени проф. Р. Т. Гриб. Лесосибирск. (В печати$_2$).
4. *Беликов В. И.* Yandex как лексикографический инструмент // «Компьютерная лингвистика и интеллектуальные технологии». Труды Международной конференции Диалог'2004. М., «Наука», 2004.
5. *Беликов В. И.* Динамика утраты флексии в полузнаменательных наречных выражениях // Язык современного города. Тезисы докладов международной конференции Восьмые Шмелевские чтения. — М.: ИРЯ РАН, 2008.
6. *Беликов В. И.* Стереотипы в понимании литературной нормы // Стереотипы в языке, коммуникации и культуре. М.: РГГУ. (В печати$_1$).
7. *БТС:* Большой толковый словарь русского языка. / Под ред. С. А. Кузнецова. — СПб., 1998.
8. *МАС:* Словарь русского языка: В 4 т. / Под ред. А. П. Евгеньевой. — 2-е изд., испр. и доп. — М.: Рус. яз., 1981–1984.
9. *ОШ:* Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка. М., 1992; изд. 2-е, М., 1994; изд. 4-е, М., 1997.
10. *Словарь-тезаурус* современной русской идиоматики: около 8 000 идиом современного русского языка / Под ред. А. Н. Баранова, Д. О. Добровольского — М.: Мир энциклопедий Аванта+, 2007.
11. *ТСРЯ:* Толковый словарь русского языка с включением сведений о происхождении слов / Ин-т рус. яз. РАН. Отв. ред. Н. Ю. Шведова. — М.: Азбуковник, 2007.
12. *Ушаков:* Толковый словарь русского языка. Т. 1–4 / Под ред. Д. Н. Ушакова. М., 1935–1940.

Построение концептуальных графов как элементов семантической разметки текстов[1]

Creating conceptual graphs as elements of semantic texts labeling

Богатырёв М. Ю. (okkambo@mail.ru), **Тюхтин В. В.**

Тульский государственный университет

В работе рассматриваются возможности применения концептуальных графов в качестве средства семантической разметки корпусов текстов. Такая разметка образует метаданные, позволяющие эффективно решать некоторые задачи Text Mining. Предлагается алгоритм автоматического построения концептуальных графов, приводятся результаты экспериментов на текстах аннотаций научных статей.

1. Введение

Концептуальные графы являются одной из семантических моделей текста, относящейся к классу семантических сетей. Впервые концептуальные графы были предложены в работах Дж. Совы, обобщенные результаты которых представлены в его монографии [1], и в настоящее время играют важную роль как средство моделирования *структур, наделенных смыслом*, в таких областях как математическая лингвистика, биоинформатика, математическая логика.

Концептуальный граф — это двудольный направленный граф, состоящий из двух типов узлов: *концептов* и *концептуальных отношений*, или просто *отношений*. Для концептуальных графов разработан стандарт их представления (см. также работу Дж. Совы [2]) и языки описания, среди которых наиболее популярны CGIF (Conceptual Graph Interchange Form) и XML — представление концептуальных графов.

Концептуальные графы применяются в задачах анализа текстовой информации, относящихся к направлению, обозначаемому термином *Text Mining*. В одной из последних обзорных российских работ М.С. Куприянова и др. [3] термин Text Mining именно так и переводится: «*анализ текстовой информации*». В англоязычной литературе Text Mining — это разновидность анализа данных *Data Mining* (иногда называемая text Data Mining), причем существенная особенность анализа в обоих вариантах состоит в том, что он направлен на *извлечение знаний* из данных, текстовых или иных. Направления Text Mining и Data Mining объединяет другое направление — *поиск знаний в базах данных* (Knowledge Discovery in Databases, KDD), которое в последнее время чаще называют *поиск знаний в данных,* используя ту же аббревиатуру. Для краткости, и учитывая неоднозначность трактовки, мы будем употреблять термин Text Mining без перевода.

В задачах Text Mining строятся *кластеры*, *ассоциации*, анализируются *особенности* текстов, *подобие* текстов и т.д. с целью извлечения знаний из текстовых данных. При этом термин «знание» трактуются как некоторая «овеществленная» модель знаний: процесс извлечения знаний приводит к нахождению конкретных значений параметров заранее заданной модели знаний. Все модели знаний условно можно разделить на два класса: модели в виде правил и модели в виде структур. Структурные модели образуют иерархию. Вершину ее составляют модели в виде формальных онтологий, а элементарной структурной моделью, соответствующей каждому предложению текста, является концептуальный граф.

Формально семантика концептуальных графов задается логическими выражениями, формируемыми на графе. Соответственно, логика предикатов первого порядка остается к настоящему времени основным математическим аппаратом исследования концептуальных графов. Здесь получено много результатов, касающихся фундаментальных свойств концептуальных графов. Стиль работ данного направления хорошо иллюстрирует работа [4].

[1] Работа выполнена при поддержке РФФИ, грант № 07-07-00276-а.

Традиционные методы обработки текстов используют ключевые слова или векторные модели текста, что требует значительных вычислительных ресурсов при обработке больших текстовых коллекций. Концептуальный граф как модель сложнее, чем набор ключевых слов, но компактнее, чем, например, вектор, построенный на тексте в методах латентно — семантического анализа. Это позволяет эффективно применять концептуальные графы в задачах Text Mining. Примером является работа [5], где на концептуальных графах решаются задачи кластеризации и выявляются отклонения сравниваемых текстов. В работе [11] мы применили концептуальные графы к построению ассоциативных правил, извлекаемых из текста.

Несмотря на признание концептуальных графов в качестве семантической модели текста и постоянный интерес к ним (см., например, электронный ресурс [12]), практическое применение концептуальных графов требует решения ряда проблем. Среди них одной из важнейших является проблема автоматизации построения концептуальных графов. Другой проблемой является поддержка концептуальных графов в реальных системах Text Mining. Появление размеченных корпусов текстов как элементов информационных систем открывает здесь новые направления исследований и, по-видимому, позволит получать более эффективные решения задач Text Mining

В данной работе представлены некоторые результаты, относящиеся к алгоритмизации построения концептуальных графов и их поддержке в пилотном исследовательском проекте электронной библиотеки.

2. Автоматическое построение концептуальных графов

Автоматизация построения концептуальных графов является в целом нерешенной проблемой. Сложность данной проблемы обусловлена смысловым многообразием, присутствующем в любом тексте естественного языка. Поэтому каждому предложению текста может соответствовать несколько концептуальных графов. Данный факт приводит к идее связать построение концептуальных графов с решаемыми при их помощи задачами. Другими словами, строить графы под задачу. На практике так и происходит: из всего многообразия задач автоматизации построения концептуальных графов решают некоторое подмножество задач, актуальных с точки зрения конечной цели применения концептуальных графов. Данный принцип применялся нами. Например, в рассматриваемом ниже алгоритме из всех знаков пунктуации анализируются только запятые.

Существует несколько подходов к построению концептуальных графов. Согласно *вербоцентрическому* подходу, в каждом предложении фиксируется центральный концепт — глагол, задающий главный смысл предложения; детализация смысла задается другими концептами и отношениями. При этом необходимо обеспечить корректную работу алгоритма в достаточно сложных случаях — в предложениях без глаголов или имеющих сложные глагольные формы. Другим важным известным решением является применение *семантических ролей для* построения отношений, которое рассмотрено ниже.

Известны системы, использующие автоматическое построение концептуальных графов для англоязычных текстов. В работе Ангеловой и др. [6] изложены принципы такого построения на основе вербоцентрического подхода. Развитием данной работы является работа S. Hensman [7], где для построения концептуальных графов привлекаются известные ресурсы VerbNert и WordNet.

Нами разрабатывается подобная система для англо — и русскоязычных текстов. В системе применяется алгоритм построения концептуальных графов, главные пункты которого состоят в следующем.

1. Анализ языка предъявленного текста; выбор между русским и английским языками. Данный этап необходим, ввиду принципиальных различий в обработке англо — и русскоязычных текстов алгоритмом.
2. Разделение предъявленного текста на предложения. Разделение предложений на слова, знаки пунктуации и иные символы.
3. Определение синтаксических элементов предложения. Построение нормальных форм для синтаксических единиц.
4. Определение морфологических признаков элементов предложения.
5. Формирование концептов из списка элементов предложения. В качестве концептов выбираются основные части речи, исключая частицы, союзы, предлоги, вводные слова.
6. Определение концептуальных отношений и акторов. *Актором* называется внешнее отношение, смысл которого назначается, а не извлекается из предложения.

Реализация пунктов 1, 2 не вызывает проблем. Синтаксический анализ в п.3 выполняется на основе известных алгоритмических решений АОТ [13]. Как показали эксперименты, данных решений недостаточно и в алгоритм добавлены значительное число как синтаксических правил, так и новых, в ряде случаев эвристических решений.

Морфологический анализ в п.4 выполняется с привлечением системы DWARF [14] и ее словарных ресурсов.

Самой сложной задачей при построении концептуальных графов является задача построения отношений. Известны подходы к решению данной

задачи, основанные на *разметке семантических ролей* предложения [8]. В нашей системе развивается подобный подход для русскоязычных текстов, основанный на применении шаблонов.

2.1. Выделение семантических ролей

Семантической ролью называется совокупность черт, общих для одинаково кодируемых элементов предложения. Сложность разметки семантических ролей обусловлена тем, что роль может не совпадать с элементом предложения и определяться не одним, а несколькими элементами. Разметка семантических ролей требует не только синтаксического, но и морфологического анализа текста.

При выделении семантических ролей применяются четыре множества объектов:

- множество атрибутов элементов предложения для русского языка,
- множество шаблонов в виде правил,
- множество атрибутов шаблонов и
- множество семантических ролей.

Шаблон содержит список специальных атрибутов, характеризующих сочетания слов анализируемого предложения. Шаблон определяет семантическую роль, но не тождественен ей. Среди атрибутов шаблонов имеются, например, такие: *название связи*, *тип связи*, *направление поиска главного слова* и т.д.

Имеется несколько типов шаблонов:

- *двухуровневый шаблон* — проверяет два рядом стоящих элемента предложения;
- *трёхуровневый шаблон* — проверяет три рядом стоящих элемента предложения;
- *грамматический шаблон* — проверяет три элемента и дополнительно наличие конца или начала предложения или соседнего знака пунктуации.

Множество шаблонов составляется на основе правил русского языка, но может пополняться новыми шаблонами с учетом конкретной лексики, например, научной.

Множество семантических ролей состоит из имен и описаний семантических ролей. В настоящее время в системе применяются роли: *агенс*, *пациенс*, *генетив*, *реципиент, атрибут, модификатор, объект, тема, источник, цель.*

После выполнения синтаксического разбора алгоритм получает список элементов предложения и каждому элементу соответствует код, собранный на множестве атрибутов элементов предложения для русского языка. Данный код используется далее при подборе шаблона, обрабатывающего элементы предложения. Последовательной проверкой соответствия шаблонов сочетаниям элементов предложения и применением соответствующих правил шаблонов на списке элементов предложения алгоритм добивается полного разбора предложения. Примененные для обработки словосочетаний шаблоны порождают семантические роли, которые тождественны концептуальным отношениям. При этом словосочетания, прошедшие обработку каждым шаблоном правил, удаляются и далее не рассматриваются. Поэтому порядок применения шаблонов имеет существенное значение и, как показывают эксперименты, влияет на результаты построения концептуальных отношений.

Например, применение одного из правил АОТ слева направо и справа налево дает различные результаты, причем правильным является второй результат, показанный на рис. 1б.

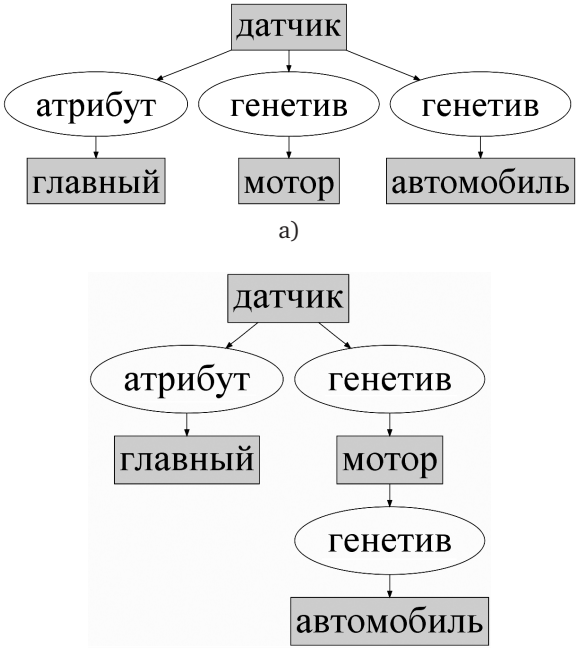


Рис. 1. Варианты концептуальных графов при разборе фразы «главный датчик мотора автомобиля» слева направо (а) и справа налево (б).

В стандартном варианте применения правила на рис. 1-а концепты *главный*, *мотор*, *автомобиль* равноправны, а в инверсном варианте на рис. 1б имеет место иерархия концептов, которая позволяет правильно интерпретировать смысл фразы «главный датчик мотора автомобиля»: мотор принадлежит автомобилю.

При определении морфологических признаков в алгоритме применяется ряд новых правил, позволяющих строить концептуальные графы более корректно. Так к правилу обработки однородных прилагательных добавлено правило построения отношения «*модификатор*» между прилагательным и местоимением в дательном падеже. Результат иллюстрируется примером на рис.2.

На рис. 2 показаны примеры моделирования фразы «предоставить необходимую им информацию» до введения правила построения отношения «модификатор» (рис. 2а) и после введения правила (рис. 2б).

На рис. 2-а показаны «бездомные» концепты, не связанные никакими отношениями, и выявляемые системой при интерпретации графов. Добавление отношения «модификатор» исправляет ошибку.

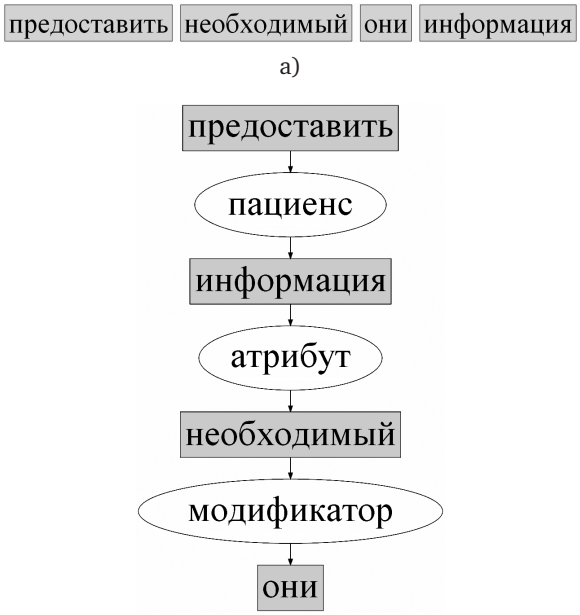


Рис. 2. Варианты моделирования фразы «предоставить необходимую им информацию».

Появление «бездомных» концептов связано с важной проблемой регулярности алгоритма.

2.2. Проблема регулярности.

Рассмотренный здесь алгоритм является сходящимся, то есть всегда приводит к построению концептуального графа. Однако, важной алгоритмической проблемой, возникающей при построении подобных систем, остается *проблема регулярности*. Нестрого, суть проблемы сводится к следующему: гарантирует ли алгоритм однотипные, регулярные решения на множестве данных, также являющихся однотипными, или отличающихся от однотипных применением к ним регулярных правил? Другими словами, построив при помощи алгоритма правильный концептуальный граф для одного предложения, вправе ли мы ожидать, что для другого предложения концептуальный граф будет построен так же корректно?

Данная проблема для задач Text Mining общего алгоритмического решения не имеет.

В нашем случае признаком нерегулярности алгоритма является появление «бездомных» концептов, показанных на рис. 2а. Увеличение числа правил в алгоритме приводит к исключению «бездомных» концептов, как это видно из рис. 2б, однако проблема регулярности остается.

С целью исследования данной проблемы в системе используется контролируемое увеличение числа правил алгоритма. В результате в интерфейсе системы введена дополнительная опция выбора «*глубины смысла концептуального графа*», получаемого из предложения. Данная опция имеет численное значение. При установке опции в наибольшее её значение мы получаем наиболее полный результирующий граф, а при установке опции в наименьшее значение (0), получаем граф, содержащий основной (вербальный) смысл предложения. «Глубина смысла» имеет также промежуточные значения. При увеличении значения опции в граф добавляются различные типы связей, причём таким образом, что значимость добавляемых связей обратно пропорциональна уровню глубины смысла. То есть, чем глубже исследуется смысл, тем более незначительные для основного смысла предложения элементы и связи добавляются в концептуальный граф.

Отметим одно полезное свойство алгоритма, связанное с проблемой регулярности. При разборе предложений с ошибками, включающими несогласованные, посторонние элементы, алгоритм порождает «бездомные» концепты. В этом случае их индикация позволяет выявить необычные особенности текста.

3. Применение концептуальных графов в проекте электронной библиотеки

Рассмотренный алгоритм реализован в исследовательском проекте электронной библиотеки, ориентированной на хранение текстов научных публикаций. Отметим две специфические функции, которые планируется реализовать в проекте дополнительно к стандартным библиотечным функциям.

1. **Диагностика новой информации, появляющейся в библиотеке**. Данная функция необходима, когда ресурсы библиотеки пополняются из сети Интернет. При реализации функции требуется решение задачи Text Mining, известной как *извлечение фактов — Fact Extraction*. На концептуальных графах данная задача может быть решена методом выявления отклонений в сравниваемых текстах [5].
2. **Концептуальная трассировка**. Используя данную функцию, пользователь системы, введя текст — запрос, получает в качестве выходных данных иерархическую структуру понятий, *имеющих отношение* к данному запросу. Функция полезна для обучения. Ее реализация требует построение онтологий.

Тексты научных публикаций загружаются в систему из сети Интернет, из открытого источника [15]. Практически все тексты данного источника англоязычные. Для проведения экспериментов на русскоя-

зычных текстах используется источник [16] — труды конференций RCDL за 10 лет в объеме 546 единиц.

Концептуальные графы строятся только для аннотаций статей, поскольку аннотации статей призваны сжато и точно отражать их содержание. Кроме того, аннотации имеют ограниченную лексику, что важно для повышения регулярности работы алгоритма построения концептуальных графов. Построенные графы обрабатываются подсистемой интерпретации. Среди ее функций разрабатывается функция принятия решения о загрузке текста статьи в библиотеку по результатам анализа ее аннотации. Здесь как раз необходима диагностика новой информации.

Реализация рассмотренных функций основана на решении задач *агрегирования* и *кластеризации* на концептуальных графах, однако не сводится только к ним. Рассмотрение указанных задач выходит за рамки данной работы, их более детальное описание можно найти в работе [9].

Построение концептуальных графов выполняет специальная подсистема. В ней имеется диалоговый режим и режим автоматического построения концептуальных графов по находящимся в базе данных текстам аннотаций.

В диалоговом режиме реализовано полное управление процессом построения концептуальных графов: можно корректировать результаты работы алгоритма (с заданием упомянутой выше «*глубины смысла концептуального графа*»), изменять и вводить новые концепты и отношения, а также акторы, использовать визуализацию (см. рис. 1, 2), удалять графы, конвертировать их в разные форматы.

В автоматическом режиме соответствующие аннотациям генерируемые графы пополняют базу данных графов в формате XML. При этом происходит отбраковывание «неправильных» графов, имеющих «бездомные» концепты (см. выше). Также особо фиксируются несвязные графы.

Сравнение работы алгоритма на русскоязычных и англоязычных текстах дало, как и ожидалось, значительно большее число «неправильных» графов для русскоязычных текстов, чем для англоязычных. Это объясняется двумя причинами:

- известной сложностью русского языка по сравнению с английским — большим числом правил и их нерегулярностью;
- несоблюдением принципа компактности в русскоязычных аннотациях — наличием длинных и очень длинных предложений, что повышает вероятность генерации «неправильных» графов.

3.1. Поддержка контекстов.

Главной проблемой построения концептуальных графов, которая решается в настоящее время, является *поддержка контекстов*.

Каждый концептуальный граф строится для одного предложения, но предложения могут быть связаны по смыслу. Часто страдательный залог, употребленный в предложении, является индикатором важной ссылки на внешнюю информацию — *контекст*, содержащуюся в других предложения или вне анализируемого текста. Построение концептуальных графов для предложений в страдательном залоге является в настоящее время самой трудной задачей. Однако, получаемые здесь «неправильные» графы могут быть полезны в задаче поддержки контекстов.

Проиллюстрируем данную ситуацию следующим примером. На рис. 3 показан концептуальный граф, соответствующий предложению «Много статей посвящено данной теме».

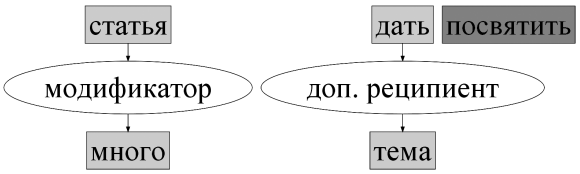


Рис. 3. Концептуальный граф предложения «Много статей посвящено данной теме»

Полученный граф является несвязным, что будет отслежено подсистемой построения графов. Отношение *дополнительный реципиент* как раз и введено для применения в подграфах несвязных графов. Правильно ли построен данный граф? С одной стороны, страдательный залог, имеющийся в предложении, никак не отражен. С другой стороны, отдельный подграф [*дать*] → (*доп. реципиент*) → [*тема*] задает особенность, которую мы можем дальше интерпретировать. Информация (знание), представляемая данным графом, сводится к следующему:

- существует много статей на некоторую тему;
- эта тема не описана в данном предложении, но, возможно, описание дано в других предложениях.

Таким образом, тема в анализируемом предложении определена в контексте. Мы можем сопоставить с графом рис. 3 индикатор: «Дать тему!», определив его буквально как правый подграф графа. Данный индикатор может служить признаком контекста и инициировать его построение.

Формализм концептуальных графов определяет контекст как *концепт графа, с которым связан некоторый непустой другой граф* [2]. В нашем примере концепт *тема* и является контекстом. Остается построить по остальному тексту граф (его может не быть!), описывающий тему, которой посвящено много статей.

Рассмотренный пример является частным случаем, индуцирующим практически важную, на наш взгляд, задачу: имея концептуальные графы как структуры микро-уровня (уровня одного предложения), строить онтологии как структуры макро-

уровня. Строя таким образом онтологии *снизу вверх*, в отличие от применяемых сейчас в системах построения онтологий методов *сверху вниз*, мы обеспечим большую адекватность онтологии той информации, которая содержится в соответствующих ей данных. Построение онтологий данным способом при помощи концептуальных графов связано с решением двух задач — задачи агрегирования концептуальных графов и задачи поддержки контекстов на концептуальных графах.

3.2. Концептуальные графы как элементы разметки корпусов текстов

Идеология разрабатываемой в данном проекте системы соответствует идеологии *корпусов текстов*. В самом деле, корпус текстов — это способ компьютерного хранения текстовых данных нового поколения, который, кроме собственно текстов, содержит их *разметку*. Разметка представляет собой *метаданные*, отражающие как лингвистическую, так и экстралингвистическую информацию, касающуюся хранимых текстов.

Разметка определяется задачами, решаемыми на текстах корпуса. Примерами двух принципиально разных классов задач, решаемых на корпусах, являются *лингвистические исследования текстов* и *извлечение знаний из текстов корпусов*. Последний класс задач соответствует назначению разрабатываемой в данном проекте системы.

Семантической разметкой назовем отображение текста корпуса на некоторую семантическую модель, например, концептуальный граф. Элементы текста — слова и предложения — отображаются в элементы модели — концепты, и отношения концептуального графа.

Применение концептуальных графов в качестве полноценного средства разметки текстов неразрывно связано с организацией корпуса как информационной системы, что сводится к следующему:

- автоматизация построения концептуальных графов;
- организация хранения концептуальных графов;
- алгоритмическая и вычислительная поддержка задач, решаемых при помощи концептуальных графов.

Данные элементы являются составными частями рассмотренной здесь системы. Ее развитие может быть связано с решением задач корпусной лингвистики.

4. Выводы и дальнейшие исследования

Автоматическое построение концептуальных графов как семантических моделей текста является алгоритмически сложной задачей. Применение для ее решения разметки семантических ролей предложения позволяет получить приемлемые результаты. Однако, данного решения недостаточно для построения графов для предложений в страдательном залоге. Поддержка контекстов является возможным решением в данном случае, не выводящим за пределы формализма концептуальных графов.

Кроме того, построенные концепты и соответствующие им графы — контексты могут служить основой для построения онтологий. Такой способ построения онтологий *снизу вверх*, по-видимому, будет способствовать сохранению той информации, которая содержится в порождающих онтологии данных. Однако, данная идея требует более тщательной проработки.

Дальнейшие исследования по данной теме планируется выполнить в следующих направлениях:

- исследование и доработка алгоритма построения концептуальных графов в части реализации поддержки контекстов;
- переход к задаче агрегирования концептуальных графов с последующей разработкой инструментов интерпретации графов — агрегатов как элементов онтологий;
- выполнение вычислительных экспериментов с алгоритмом построения концептуальных графов в режиме реального времени.

Литература

1. *Sowa, J. F.* Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA. 2000.
2. *Sowa J.F. Conceptual Graphs: Draft Proposed American National Standard, //International Conference on Conceptual Structures ICCS-99, Lecture Notes in Artificial Intelligence 1640, Springer 1999.*
3. *Технологии* анализа данных: Data Mining, Visual Mining, Text Mining, OLAP/ А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. — СПб: БХВ-Петербург, 2008, — 384 с.
4. *Chein M., Mugnier Marie-Laure.* Conceptual Graphs: fundamental notions // Revue d'Intelligence Artificielle, Vol. 6, n 4, 1992, pp 365–406.
5. *Montes-y-Gomez, M., Gelbukh, A., Lopez-Lopez, M.* Text Mining at Detail Level Using Conceptual Graphs. //Lecture Notes In Computer Science; Vol. 2393. P. 122–136.
6. *Boytcheva, S. Dobrev, P. Angelova, G.* CGExtract: Towards Extraction of Conceptual Graphs from Controlled English. //Lecture Notes in Computer Science № 2120, Springer 2001.
7. *Hensman, S.* Construction of conceptual graph representation of texts. In Proceedings of Student Research Workshop at HLT-NAACL, Boston, 2004, p.p. 49–54.
8. *Gildea D., Jurafsky D.* Automatic labeling of semantic roles //Computational Linguistics, 2002, v. 28, p.p. 245–288.
9. *Богатырев М. Ю., Латов В. Е., Столбовская И. А., Тюхтин В. В.* Эволюционный подход к задаче кластеризации на концептуальных графах и его применение в системах поддержки электронных библиотек. — Математические методы распознавания образов. 13 Всероссийская конференция. Сб. докладов. — М.: МАКС Пресс, 2007. — 668 с. — С. 464–468.
10. *Богатырев М. Ю., Латов В. Е., Столбовская И. А.* Применение концептуальных графов в системах поддержки электронных библиотек. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Девятой Всероссийской научной конференции RCDL'2007 (Переславль-Залесский, Россия, 15–18 октября 2007). — Т. 2, С. 104–110.
11. *Богатырев М. Ю., Тюхтин В. В.* Решение некоторых задач Text Mining при помощи концептуальных графов. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Десятой Всероссийской научной конференции RCDL'2008 — Дубна, 2008. — 415 с. — С. 31- 36.
12. *Электронный* ресурс: A World of Conceptual Graphs: http://conceptualgraphs.org/
13. *Электронный* ресурс: Автоматическая Обработка Текста http://aot.ru/
14. *Электронный* ресурс: Cognitive Technologies — Интеллектуальные технологии управления. http://www.cognitive.ru
15. *Электронный* ресурс: Scientific Literature Digital Library http://citeseer.ist.psu.edu/
16. *Электронный* ресурс: Труды конференций Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции http://rcdl.ru/

Звуковой корпус как способ мониторинга и фиксации разных форм естественного языка[1]

A speech corpus as a tool for monitoring and fixation of various forms of natural language

Богданова Н. В. (nvbogdanova_2005@mail.ru),
Асиновский А. С. (a.s.asinovsky@gmail.com),
Русакова М. В. (mvrusakova@gmail.com),
Рыко А. И. (aryko@yandex.ru), **Степанова С. Б.** (stsvet_2002@mail.ru),
Шерстинова Т. Ю. (sherstinova@gmail.com)

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

Доклад посвящен разработке методов мониторинга и фиксации звукового материала естественного языка, принципов организации информационной среды и программного инструментария для нужд интегрального моделирования, а также описанию первых готовых блоков Звукового корпуса русского языка (ЗКРЯ).

1. Введение

Язык современного города представляет собой, вне всякого сомнения, фактор социальной и психологической дифференциации — настолько в нем переплетены и взаимосвязаны различные социолекты и идиолекты. Это и делает его столь интересным объектом для самых разных исследований. О «лингвистической истории большого города» много писал еще Б. А. Ларин, усматривая в ней «борьбу языков, отражающую непрестанное столкновение и скрещивание <…> разнородных культур» [6: 177]. По мнению Б. А. Ларина, языковое разнообразие города характеризуется, с одной стороны, *многоязычием*, обусловленным «встречей разноязычных коллективов», а с другой – *многодиалектностью*, поскольку «в каждом слое городского населения, кроме первичного, “своего”, наречия, необходимо располагают еще каким-либо универсальным языковым типом, приобщающим к большой социальной среде» [6: 190]. Наилучшим образом все это многообразие речи горожан может быть представлено с применением корпусного подхода к собранию речевого материала. Именно такой подход был реализован при создании Звукового корпуса русского языка (ЗКРЯ), сбалансированного как лингвистически, так и психо- и социолингвистически.

С другой стороны, одной из основных задач современной лингвистики является накопление и систематизация новых данных о том, как соотносятся лингвистические описания, сделанные на основе традиционных понятийных и терминологических систем, и физически наблюдаемая звуковая материя естественной устной русской речи. Поворот лингвистики от структуралистского описания языковой системы и обслуживания прикладных областей знания к говорящему и слушающему человеку выразился прежде всего в понимании бесперспективности дальнейшего моделирования методом «черного ящика» в процессе решения прикладных задач и «дальнейшего шлифования метода классических дефиниций» [7: 232]. «Нужна принципиально другая “дефинициология”, чем та, которая досталась нам от аристотелевой и математической логики». Необходимо понять, «что такое точное описание в применении к языковым явлениям. Такое понимание нельзя импортировать из других областей знания, включая математику. Чтобы понять операционную природу лингвистической точности, надо предварительно углубить наши неформальные знания о *природе* (курсив автора – *Авт.*) языковых явлений, а уж затем выработать соответствующие формализмы точного их представления» [7: 232].

Идея о существовании, наряду с грамматикой языка, своеобразной грамматики речи, обладающей своими собственными единицами и специфическими правилами их функционирования и сочетания (своей парадигматикой и синтагматикой), в лингвистике отнюдь не нова и в течение последнего столетия высказывалась неоднократно. Так, еще Ф. де Соссюр писал, что «лингвист должен также рассматривать

[1] В настоящее время работа проводится при поддержке РГНФ — проект 07-04-12163в «Разработка информационной среды для мониторинга устной русской речи».

взаимоотношения книжного языка и обиходного языка (по сути, взаимоотношения языка и живой повседневной речи — *Авт.*), ибо развитие всякого литературного языка, продукта культуры, приводит к размежеванию его сферы со сферой разговорного языка» [11: 44]. Подобную мысль находим и в трудах Л. В. Щербы, ср.: «нужно прежде всего различать у русских, т. е. у говорящих и пишущих на общерусском литературном языке, два языка: один слышимый и произносимый (снова, по-видимому, имеется в виду живая речь — *Авт.*), а другой написанный, которые находятся один к другому в известных отношениях, но не тождественны — элементы одного не совпадают с элементами другого» [15: 1]. И далее автор пишет, что «если надо различать эти два языка, то надо, очевидно, различать и их грамматики» [15: 12].

Одним из первых при таком подходе встает вопрос об инвентаре тех единиц, которые должны стать стержнем любой грамматики, в том числе и грамматики речи. Уже первые наблюдения над спонтанным материалом показали, что на нем преломляются все исходные языковые метапонятия, на которых строится обычно его анализ. Привычные понятия *фонемы*, *морфемы*, *слова* и *предложения* оказываются неприложимыми или плохо приложимыми к спонтанной речи. Фактически на этом материале все традиционные метапонятия языка (единицы его описания) так или иначе разрушаются, на их месте создается нечто новое, что не всегда легко поддается определению и описанию. Ср.: «…происходящее при переходе от чтения к говорению “переключение кодов” приводит к расширению допустимых пределов варьирования, т. е. размытости фонетических характеристик таких единиц, как фонема, слово (на уровне фонетической реализации), синтагма, размытости, степень которой возрастает по мере возрастания степени неофициальности общения» [13: 133].

Несмотря на значительные успехи в различных областях лингвистики, связанных (в самом общем смысле) с функционированием языка в процессе коммуникации, на сегодняшний день не существует сколько-нибудь целостного *многоуровневого* описания звучащей речи. Настоящий проект направлен на исследование разнообразных закономерностей разворачивания естественной устной речи в первую очередь на русском языке и представляет собой один из первых подходов к исследованию звучащей речи с применением новых методологических возможностей и с формированием новых терминологических решений в перспективе.

2. Материал корпуса

Материалом настоящего исследования является естественная звучащая русская речь. Формат языкового материала может быть определен как *Звуковой корпус русского языка* в его естественной, звучащей, форме. К настоящему времени подготовлены 3 основных модуля, состоящие из 6 речевых подкорпусов (см. табл. 1).

Таблица 1. Блоки речевого материала

Один речевой день			
ORD	«Один речевой день» (Повседневная речь)	34 информанта + 560 коммуникантов	235 часов звучания
Сбалансированный материал			
MED	Речь медицинских работников	32 диктора 210 текстов	6 часов звучания
JUR	Речь юристов	40 дикторов 322 текста	16 часов звучания
RKI	Речь преподавателей РКИ	20 дикторов 70 текстов	3,5 часа звучания
STUD	Речь студентов	5 блоков	7 часов звучания
Интерферированная речь			
RIA	Интерферированная речь (русская речь американцев)	48 дикторов	8 часов звучания

А. *ORD. «Один речевой день»*

Принципиальное значение для нашего проекта имеет модуль «Один речевой день», занимающий в нем центральное место. *Целью* создания данного блока ЗКРЯ явилось изучение речевого поведения носителя русского языка в течение дня (с использованием методики 24-часовой записи) в зависимости от ряда факторов:

- его социально-психологические характеристики,
- его коммуниканты (= социальная роль),
- место, где протекает общение,
- время суток.

Выборка информантов при записи материала ORD — пока несбалансированная, хотя, вероятно, она в некоторой степени отражает социальный и психологический срез современного общества. Из 34 участников эксперимента было 12 мужчин и 22 женщины, в возрасте от 15 до 63 лет. Одновременно была осуществлена запись около 560 их коммуникантов (приблизительность цифры объясняется невозможностью в некоторых случаях определить принадлежность голоса одному или разным коммуникантам). Все они относятся к различным социальным группам и имеют разный уровень образования — от среднего специального до высшего, в том числе с ученой степенью. Сферы деятельности информантов также оказались разнообразны — это и учеба, и занятость на производстве, и научно-преподавательская деятельность, и бизнес, и торговля, и юриспруденция, и медицина, и строительство, и некоторые другие[2].

[2] Подробнее о модуле ОРД см. [12].

*В. **Сбалансированный материал***

Иные принципы были положены в основу создания второго блока Звукового корпуса русского языка [1]. Этот блок изначально достаточно строго сбалансирован по разным параметрам — социологически, психологически и собственно лингвистически.

Социологическая балансировка материалов корпуса заключалась в изначальном подборе информантов с разными социальными характеристиками, среди которых, помимо традиционных признаков пола, возраста, образования и проч., следует отметить *уровень речевой компетенции* (УРК) говорящего, который характеризует умение человека решать разные коммуникативные задачи, свободу его в выборе речевых средств и навыки построения устного монолога различного характера. Определяют УРК в основном два признака говорящего индивида: уровень образования и *профессиональное* или *непрофессиональное* его *отношение к речи*. Предполагается, что непрофессионально относятся к речи те люди, для кого язык/речь является лишь *средством общения* (большинство так называемых «наивных» носителей) или средством общения и *объектом изучения* (школьники или «кабинетные» ученые-филологи). Их профессиональная деятельность обычно не связана с активной публичной речевой практикой. Профессиональное отношение к речи характеризует тех, для того язык/речь – это не только средство общения или даже объект изучения, но еще и *орудие труда* (актеры, дикторы, лекторы, публичные и общественные деятели, преподаватели, особенно — преподаватели-филологи). Высшее образование и профессиональное отношение к речи нормально формируют высокий уровень речевой компетенции говорящего, высшее образование и непрофессиональное отношение к речи — средний УРК, отсутствие высшего образования и непрофессиональное отношение к речи – низкий УРК.

Психологическая балансировка данного блока Звукового корпуса (осуществленная лишь частично) заключалась в подборе информантов с разными психологическими характеристиками. В основе такого подбора лежал психологический тест Г. Айзенка на определение интровертности-экстравертности и эмоциональной неустойчивости личности[3].

Лингвистическая балансировка материала заключалась в том, что все тексты, составляющие Корпус, построены в рамках комплекса *коммуникативных сценариев*, обычно реализующихся в нашей повседневной бытовой речи:

- *чтение* (сюжетный/несюжетный исходный текст);
- *пересказ* (сюжетный/несюжетный исходный текст);
- *описание изображения* (сюжетное/несюжетное);
- *свободный рассказ на заданную тему* (знакомая/незнакомая тема).

[3] Подробнее об использовании данного теста в наших исследованиях см. [5].

Представляется, что все характеристики подобных бытовых монологов определяются двумя их признаками, находящимися в отношениях обратно пропорциональной зависимости: степень *лингвистической мотивированности* текста неким исходным стимулом и степень *спонтанности* вторичного речевого произведения. Чем более монолог мотивирован тем или иным первичным текстом, тем он менее спонтанен, и наоборот.

Дополнительными характеристиками исходного стимула, способными повлиять на свойства спонтанного монолога, стали в данном проекте *сюжетность-несюжетность предтекста или изображения* и *степень знакомства* говорящего *с темой свободного монолога*, заданной вопросом. Эти дополнительные характеристики не меняют степени лингвистической мотивированности и, соответственно, степени спонтанности вторичного текста, но все же оказывают влияние на выбор говорящим тех или иных речевых средств и в целом на лингвистическую природу вторичного текста. Можно предположить, что в данном случае решающими являются характеристики уже не (или не только) первичного текста, но и самого говорящего — уровень его речевой компетенции или психологический тип личности.

Думается, что соблюдение принципов, положенных в основу создания данного блока Звукового корпуса русского языка, позволяет получить достоверный и представительный речевой материал, пригодный для анализа в различных аспектах и дающий представление об особенностях речи того или иного социума.

3. Программное обеспечение исследования

Формирование звуковых корпусов и их многоуровневая разметка стали возможны в последнее время благодаря развитию информационных технологий в гуманитарных науках. Работа, связанная с интерпретацией звучащей речи, исключительно трудоемка, но благодаря специальным программам она принципиально выполнима.

При создании нашего корпуса используются следующие программные средства:

- программа профессионального фонетического анализа Praat;
- профессиональный аудиоредактор Sound Forge 8.0;
- программа многоуровневого лингвистического аннотирования ELAN;
- программа лексикографической обработки данных KartaTeka (собственная разработка);
- программа для создания баз данных Accsess.

Исследовательский интерфейс связывает все модули в единую информационно-исследовательскую среду.

3.1. Программа профессионального фонетического анализа Praat

Программа Praat, созданная сотрудниками факультета фонетики Амстердамского университета П. Бёрсма и Д. Вининком (Paul Boersma, David Weenink)[4], предназначена для лингвистов, исследующих звучащую речь. Она предоставляет ряд возможностей для сегментации звукового потока, анализа и синтеза речи, для манипуляций со звуком в целях проверки различных гипотез, связанных с организацией звуковой формы языка, а также дает возможность создания иллюстративного материала для публикации результатов исследований.

Поскольку Praat позволяет осуществлять многоуровневую разметку звучащей речи, именно эта программа использовалась нами для соположения и интеграции информации, относящейся к собственно акустическому, фонетическому, словесному и фразовому уровням[5].

3.2. Программа многоуровневого лингвистического аннотирования ELAN

Программа ELAN является средой профессионального лингвистического аннотирования аудио- и видеоматериалов, которая поддерживает фактически неограниченное количество уровней аннотации, любые шрифты и кодировки данных, сложные иерархические структуры связей между данными, экспорт и импорт аннотаций в основные форматы представления данных. Программа разработана в институте психолингвистики им. Макса Планка в Голландии специально для исследования звучащего языка, речевого поведения и жестикуляции и является удобным средством для обработки, документирования и аннотирования разнообразных мультимедийных корпусов[6].

ELAN поддерживает:

- визуализацию аудио- и/или видеосигналов одновременно с полученными аннотациями;
- временну́ю привязку аннотаций к медийному потоку;
- сложные связи аннотаций друг с другом;
- неограниченное количество задаваемых пользователем уровней аннотации (Tiers);
- различные шрифты и кодировки;
- экспорт данных в виде текстовых файлов табличного вида (tab-delimited text);
- импорт и экспорт между ELAN, PRAAT, ToolBox, Shoebox и другими популярными лингвистическими программами;
- поисковые опции.

На рис. 1 представлен звуковой фрагмент *Я настолько себя плохо чувствую, я так устал* с разметкой формата ELAN.

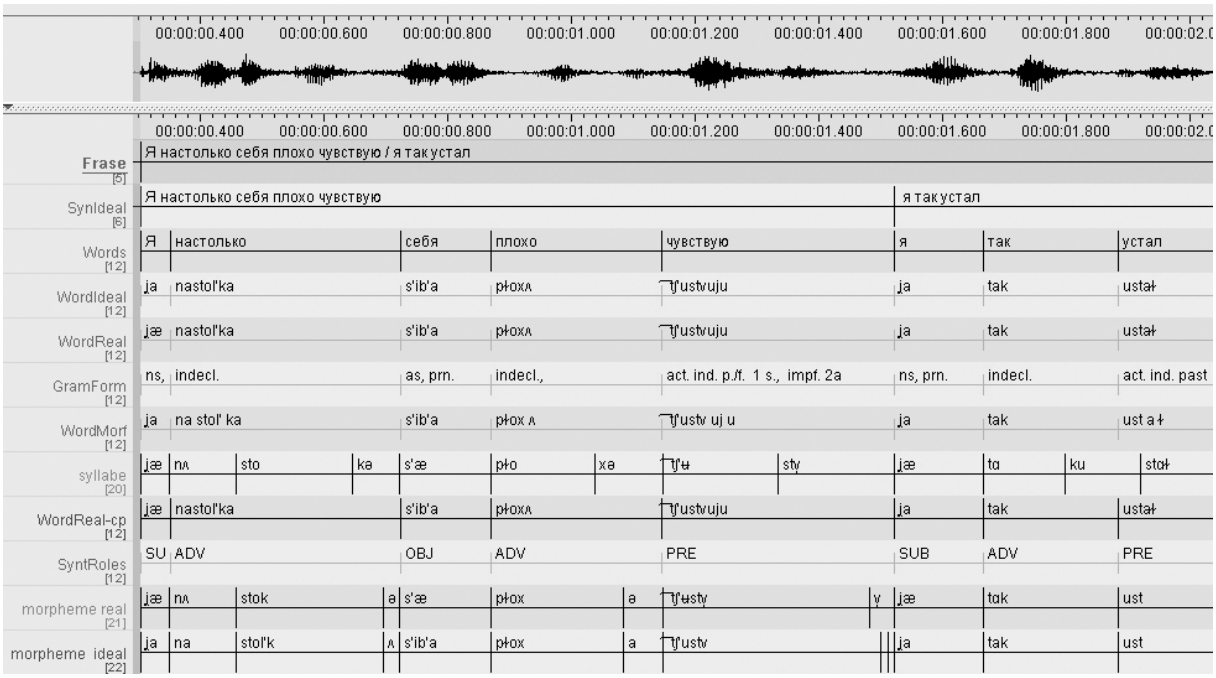


Рис. 1. Разметка данных в формате ELAN

[4] Создатели Praat регулярно обновляют версии своей программы на сайте www.praat.org и предлагают бесплатное ее использование для некоммерческих целей.

[5] Подробное описание принципов выделения и аннотирования уровней см. в [10].

[6] Программа является свободно распространяемой и может быть скачана с сайта института психолингвистики им. Макса Планка: http://www.lat-mpi.eu/tools/elan/.

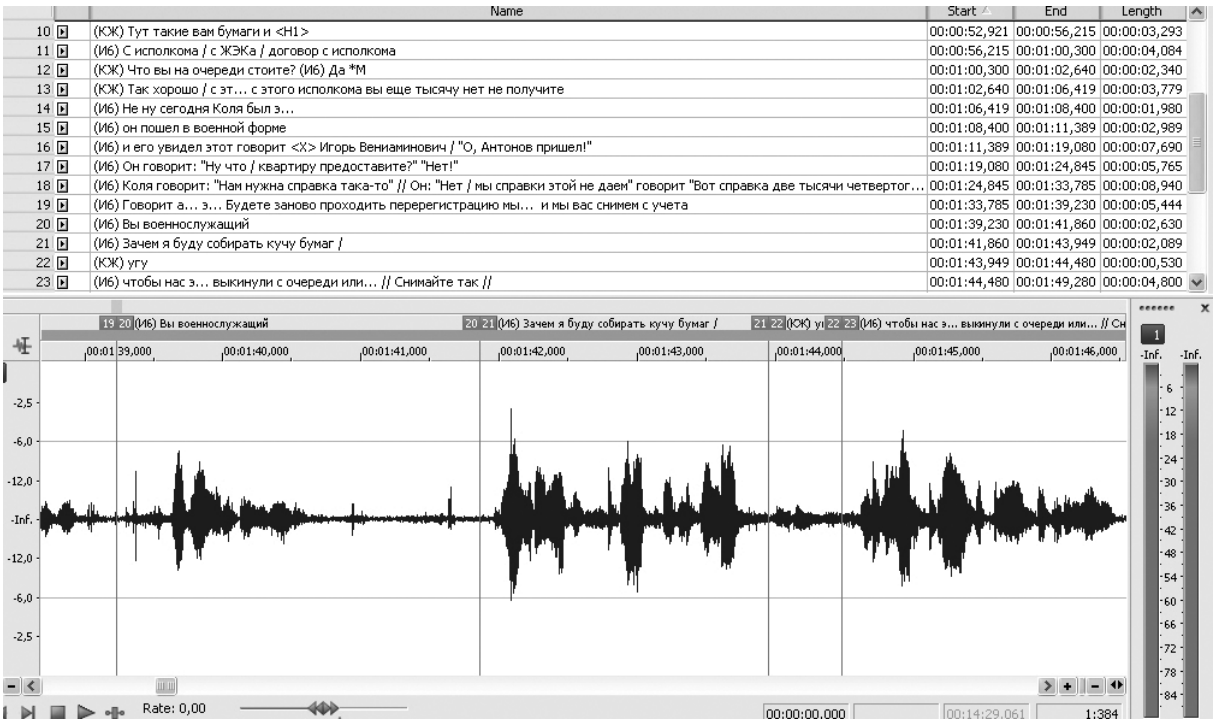


Рис. 2. Разметка данных в формате Sound Forge

3.3. Сравнение форматов аннотирования данных

Использование разных форматов аннотирования обусловлено спецификой используемых программных средств. Так, программа лингвистического аннотирования ELAN является крайне удобной для представления многоуровневой разметки разнородного материала, хранит данные в удобном для их последующей автоматической обработки виде. Именно формат ELAN принят за основной для разметки нашего корпуса. Программа PRAAT используется для фонетической обработки материала — заполнения аннотаций на фонетических уровнях, а программа Sony Sound Forge хорошо работает с аудиофайлами большой длительности и позволяет совместить аннотирование с их точной сегментацией (см. рис. 2). Все используемые форматы аннотирования данных являются совместимыми и могут быть взаимно перекодированы.

3.4. Электронная картотека E-Kar

Программа E-Kar автоматически создает конкорданс по выбранным текстам и позволяет решать многообразные задачи классификации и описания языковых единиц. В частности, она позволяет собрать по тексту все словоформы, в нем встречающиеся, посчитать их частоты (см. рис. 3), предъявить для этих словоформ любое лингвистическое расширение по тексту или группе текстов (см. рис. 4), имеющихся в электронной коллекции.

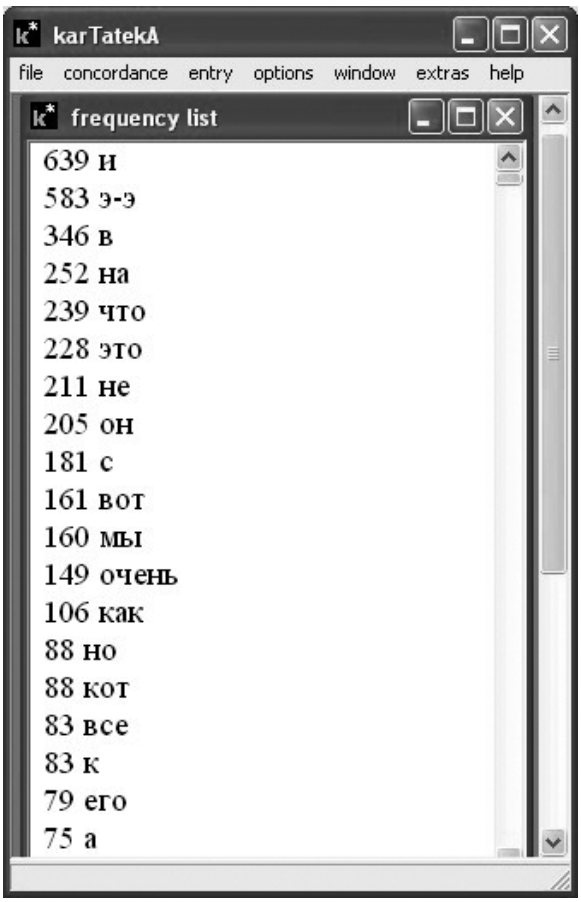


Рис. 3. Частотный словник словоформ

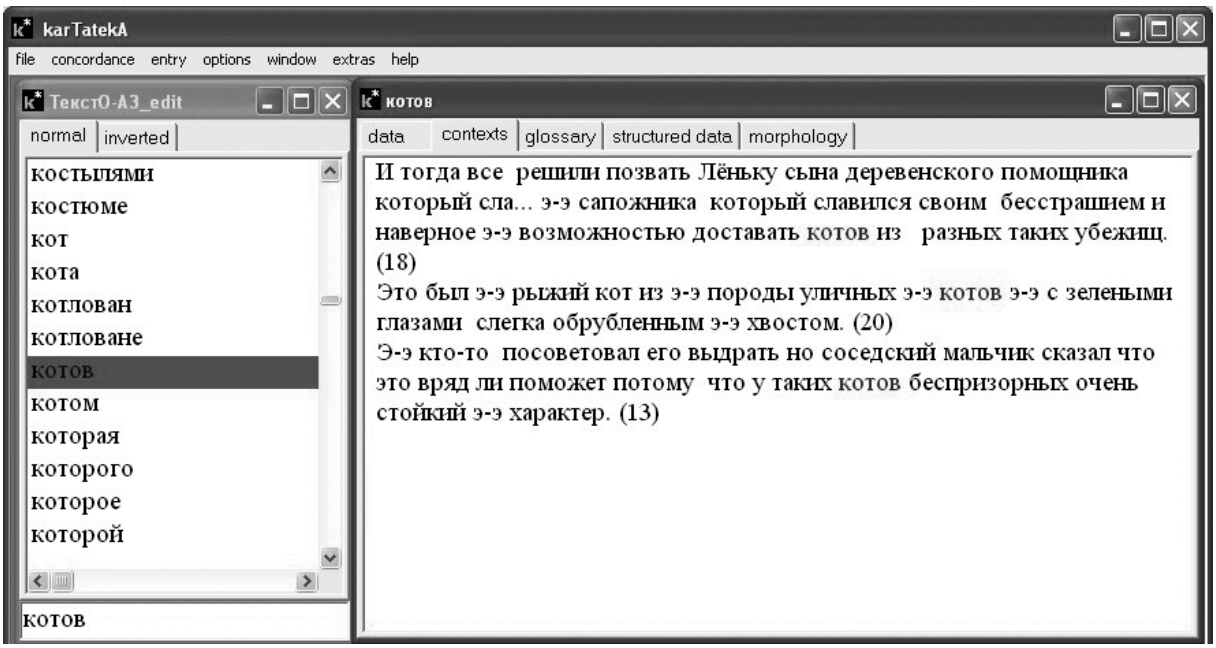


Рис. 4. Словоформа с контекстом

Результаты автоматической работы программы (конкорданса) и последующей работы эксперта дают возможность новых содержательных форм интерпретации лингвистического материала.

Конкорданс дает необходимую для лингвиста информацию, которая помогает классифицировать словоформы по морфологическим признакам и проводить лемматизацию, а также идеографическую или тематическую классификацию.

3.5. Специализированная база данных

Для самого крупного звукового модуля исследовательской среды «Один речевой день» (ORD), включающего в себя около 260 часов аудиоматериала, подготовлена специализированная база данных SpeechDay, реализованная в среде MS Access 2003.

На настоящий момент база данных состоит из 7 заполненных таблиц. Все таблицы можно условно разделить на 2 группы: *фактические данные* и *результаты научно-исследовательской работы и их интерпретация*. Некоторые таблицы содержат «смешанные» данные[7].

4. Некоторые результаты исследований на материале ЗКРЯ

Хотя работа над созданием ЗКРЯ еще далека от своего завершения (впрочем, вряд ли её можно и нужно завершать, учитывая основную цель проекта — отражение и фиксация постоянно изменяющейся повседневной русской речи), материал корпуса и процесс его аннотирования уже подвергся исследованиям, о результатах которых стоит здесь упомянуть.

Так, специальный анализ материала [14] показал, что бóльшая часть записей ORD (44 %) была сделана информантами на работе или учебе, со значительным отрывом далее следуют записи семейных разговоров по вечерам (10 %), разговоры в кафе или ресторанах (10 %), по дороге куда-либо (9 %), утром за завтраком (5 %). Остальные разговоры (за обедом, во время спортивных или культурных мероприятий, в гостях и т. п.) в жизни наших испытуемых занимают гораздо меньше времени. При этом различие между мужчинами и женщинами заключается в основном в том, что мужчины потратили на разнообразные мероприятия почти на 9 % больше времени, чем женщины; естественно и времени на дорогу у них ушло больше (почти на 5 %). Женщины же «это время» потратили на разговоры дома вечером (на 7 % больше), на вечеринках и за ужином (примерно на 2 % больше по каждой категории) и утром (на 3 % больше). Впрочем, с психологической и социологической точек зрения такой результат не является неожиданным.

Интересные результаты дала разметка части материала ЗКРЯ с точки зрения отклонений от нормативной речи на всех лингвистических уровнях. Оказалось, что отклонения в русской речи являются весьма обыденным явлением, по частотности сопоставимым с употреблением имен существительных и гораздо более частотным, чем, например, употребление прилагательных [9].

Статья [2] посвящена анализу зависимости способа передачи чужой речи от уровня речевой компетенции говорящего (материал MED).

[7] См. описание в [12].

В работе [8] приводятся результаты наблюдений над речевыми особенностями проявления агрессивности в спонтанной речи (материал ORD), в [4] анализируется лингвистическая структура высказывания с точки зрения проявления в ней психологических характеристик говорящего (ORD).

Проведено исследование специфики пересказа как вида речевой деятельности (на материале блока JUR). Анализ показал, что на порождение репродуктива оказывает влияние ряд факторов как лингвистического (характер первичного текста), так и экстралингвистического (социальные или психологические характеристики говорящих) толка [5].

В работе [4] показано, как профессия говорящего проявляется в спонтанной речи на лексическом уровне (блок MED).

Материалы Звукового корпуса по мере их обработки передаются в Национальный корпус русского языка.

Литература

1. *Богданова Н. В., Бродт И. С., Куканова В. В., Павлова О. В., Сапунова Е. М., Филиппова Н. С.* О «корпусе» текстов живой речи: принципы формирования и возможности описания // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). П материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 57–61.
2. *Богданова Н. В., Бродт И. С.* О способах передачи чужой речи (на материале звукового корпуса русского языка) // Материалы XXXVII международной филологической конференции. Выпуск 21. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 10–15 марта 2008 года / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2008. С. 3–16.
3. *Иванова О. А.* К характеристике внутриязыкового контакта между литературной и профессиональной речью носителя русского языка // Материалы XXXVII международной филологической конференции. Выпуск 21. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 10–15 марта 2008 года / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2008. С. 25–35.
4. *Королева И. В.* Индивидуальные состояния и свойства языковой личности: влияние на лингвистическую структуру высказываний // Материалы XXXVII международной филологической конференции. Выпуск 21. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 10–15 марта 2008 года / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2008. С. 36–46.
5. *Куканова В. В.* Специфика пересказа как вида речевой деятельности: эндо- и эзоединицы звучащего спонтанного монолога // Вестник Санкт-Петербургского университета. Филология. Востоковедение. Журналистика. Серия 9. Вып. 4. Часть 2. СПб., 2008. С. 135–145.
6. *Ларин Б. А.* История русского языка и общее языкознание. М., 1977.
7. *Кибрик А. Е.* О «невыполненных обещаниях» лингвистики 50–60-х годов // Московский лингвистический альманах. Спорное в лингвистике. Вып. 1, М., 1996.
8. *Маркасова Е. В.* Риторическая энантиосемия в корпусе русского языка повседневного общения «Один речевой день» // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 352–355.
9. *Русакова М. В.* Сбои при порождении словоформы в устной речи как результат спонтанного взаимодействия стратегий и механизмов // Материалы XXXVI международной филологической конференции. Выпуск 20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 12–17 марта 2007 года / Отв. ред. А. С. Асиновский, Н. В. Богданова. СПб., 2007. С. 59–71.
10. *Рыко А. И., Степанова С. Б.* Многоуровневая лингвистическая разметка звукового корпуса русского языка // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог» (2008). М.: 2008. С. 460–465.
11. *Соссюр Ф. де.* Курс общей лингвистики. М., 1933.
12. *Степанова С. Б., Асиновский А. С., Богданова Н. В., Русакова М. В., Шерстинова Т. Ю.* Звуковой корпус русского языка повседневного общения «Один речевой день»: концепция и состояние функционирования // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 488–494.
13. *Фонетика* спонтанной речи / Под ред. Н. Д. Светозаровой. Л., 1988.
14. *Шерстинова Т. Ю.* «Один речевой день» на временно́й шкале: о перспективах исследования динамических процессов на материале звукового корпуса // Филология. Востоковедение. Журналистика. Серия 9. СПб., 2009 (в печати).
15. *Щерба Л. В.* Избранные работы по русскому языку. М., 1957. С. 11–20.

КроссЛексика — большой электронный словарь сочетаний и смысловых связей русских слов

Crosslexica: a large electronic dictionary of collocations and semantic links between russian words

Большаков И. А. (bolshakov34@mail.ru)

Национальный политехнический институт, Мехико, Мексика

Большой русский электронный словарь содержит словник из 185 тыс. титулов, 1,75 млн. словосочетаний, 2 млн. смысловых связей между словами, английские переводы титулов, их морфопарадигмы. Работает в диалоге (редактирование текстов, обучение языку) и доступен из программ парсинга, разрешения омонимии, обнаружения/исправления смысловых ошибок, стеганографии.

1. Введение

За последние 20 лет в русском письменном языке произошли существенные сдвиги.

- Изменилась и пополнилась лексика. Накапливавшиеся ранее разговорные слова и жаргонизмы выплеснулись на страницы изданий, на телевидение, в Интернет. Появилось много новых заимствований.
- Соответственно изменился и пополнился состав словосочетаний, которыми, по формулировке И. Мельчука [6], только и говорит человек.
- В части владения языком ситуация поляризовалась. На одном полюсе, возросло число авторов (обозревателей, журналистов, ученых-гуманитариев), виртуозно владеющих языком и не стесненных советскими речевыми штампами. На другом полюсе, появилась масса полуграмотных «афторов», демонстрирующих Интернете убогую приблатненную или англизированную лексику и попирающих нормативную орфографию. Между этими полюсами сохранилась группа научно-технических авторов, не блещущих стилем и разнообразием лексики, но обеспечивающих языковую преемственность в своей сфере.

В итоге академические словари русского языка заметно устарели. Изданные в последние годы «большие» словари, напр., [1, 2, 3, 5], успевают истолковывать новации, но не отображают увеличенной массы словосочетаний.

В те же десятилетия радикально усовершенствовалась вычислительная техника. Типовой объем дискового накопителя увеличился в тысячи раз. В памяти десктопа, лаптопа, мобильника уже умещаются словари любого объема. При выдаче словарных статей на экран уже не обязательно повторять их привычный бумажный формат.

Данный доклад отражает результаты 19-летней работы над электронным русским словарем КроссЛексика, который объединяет подъязыки разных групп пользователей без навязывания норм и лексических крайностей. Словарь

- содержит около 185 тысяч слов и неразрывных выражений, с особым упором на их 1,75 млн. сочетаний;
- отражает 2 млн. смысловых связей (синомимы, антонимы, гипонимы-гиперонимы, меронимы-холонимы, семантические дериваты);
- включает более 0,5 млн. паронимических связей (буквенного либо морфемного сходства);
- тематически универсален, т.е. содержит политическую, научную, экономическую, политехническую и общежитейскую лексику;
- комбинаторикой различает слова из нескольких тысяч омонимических групп;
- с помощью отношений синонимии и род-вид оперативно порождает более 1,5 млн. словосочетаний, в словаре непосредственно не представленных;
- управляет порядком и полнотой выдачи и по требованию отсеивает ненужную данному пользователю информацию;
- дает английские переводы для титулов словника и частотных словосочетаний;
- приводит морфопарадигмы большинства титулов словника.

Структура КроссЛексики состоит из алфавитно упорядоченного словника и матрицы классифици-

рованных связей между его элементами. Грамматически правильное сочетание восстанавливается через связь синтагматического типа, если ввести любой знаменательный компонент сочетания.

Основной режим работы КроссЛексики — диалоговый, в нем можно редактировать русские тексты, обучаться русскому языку, получать всевозможные лексические и грамматические справки. Можно обращаться к словарю и из внешних программ, осуществляющих парсинг, разрешающих омонимию, обнаруживающих и исправляющих смысловые ошибки и др.

Конкретно, в КроссЛексике представлена следующая тематика:

- Экономика и бизнес;
- Политика и политология;
- Различные разделы техники: радиоэлектроника, компьютеры, программирование, автомобили, бытовая техника, строительство и др.;
- Точные и естественные науки: математика, физика, химия, биология, география и др.;
- Гуманитарные науки (напр., лингвистика), искусство, религия;
- Медицина (преимущественно бытовая);
- Бытовой язык, включая бранную лексику без мата.

2. Общие параметры и лингвистическая информация словаря

Титулы словника делятся на существительные (31%), глаголы (21%), прилагательные (27%) и наречия (21%). Омонимических групп 2,3 тыс. с 5,4 тыс. разных смыслов. Раскрываются 3,6 тыс. склеек типа *физфак = физический факультет,* рассматриваемые и как титулы словника, и как словосочетания.

Суммарное количество словосочетаний — около 1,75 млн. (В [4] и [2] их примерно 100 тыс. и 200 тыс.) Количество семантических связей более 2 млн., паронимических связей — более 0,5 млн.

Титулы делятся на:

- Субстантивные (раздельно единств. и множеств. число);
- Глагольные (инфинитив + личные формы, два вида берутся раздельно);
- Адъективные (прилагательные или причастия двух видов раздельно);
- Адвербиальные (наречия или деепричастия двух видов раздельно).

Отказ от чисто лексемного принципа представления существительных и глаголов диктовался необходимостью отразить различия в комбинаторике подпарадигм. Титулы подпарадигм рассматриваются как взаимные семантические дериваты.

Служебные слова (предлоги, союзы) встроены в словосочетания и своих статей обычно не имеют. Предикативные высказывания типа *а пошел ты* отнесены к наречиям.

Субстантивная статья имеет титулом либо отдельное существительное (*абажур, абберация, аббревиатура, абзац, битва…*), либо устойчивое именное словосочетание (*алкогольные напитки, ближнее зарубежье, сельское хозяйство, точка зрения, уровень жизни, болеутоляющие средства…*).

Глагольная статья имеет титулом либо одиночный глагол (*говорить, идти, обсуждать, спать, демонстрировать…*), либо глагол с возвратным местоимением (*вести себя…*), либо глагольный оборот (*наводить страх, свалиться как подкошенный, залить фары, испытывать стремление…*).

Адъективная статья имеет титулом либо отдельное прилагательное (*абстрактный, авансовый, авантюрный, автономный, воздушно-реактивный…*), либо отдельное причастие, м.б. переходящее в прилагательное (*агонизирующий, вдвинутый, коррумпированный…),* либо адъективный оборот (*хорошо одетый, большой дальности, бросающийся в глаза, в елочку, как сталь, в денежном выражении, без подкладки, большого ума…*).

Адвербиальная статья имеет титулом либо отдельное наречие *(абстрактно, адски, долго, плохо, по-мужски, удовлетворительно…*), либо отдельное деепричастие *(базируясь, дрожа, надев, успев…),* либо адвербиальный оборот *(аккуратным образом, без воодушевления, более или менее, будто обухом по голове, как выжатый лимон, в особой степени, куда попало, мелкой дрожью, на цыпочках, долгое время…).*

Связи между статьями словника делятся на:

- **Синтагматические**, формирующие словосочетания;
- **Семантические**, связывающие слова со смысловым сходством;
- **Паронимические,** связывающие внешне сходные слова.

Словосочетание — это два знаменательные слова, синтаксически связанные и устойчиво совместимые по смыслу. В синтаксической связи между двумя знаменательными словами может стоять служебное слово (предлог или союз) согласно формуле

знамен. слово1 → (служебное слово) → → знамен. слово2,

например, *сотрудничество → ради → мира, пойти → на → курсы, уверенный → в → победе.*

Каждое словосочетание доступно с двух сторон. Доступ с одной из сторон дает одностороннюю связь, и таких связей в словосочетаниях ровно вдвое больше, чем самих словосочетаний.

Наиболее частотны в словаре следующие типы словосочетаний:

- **Существительное — прилагательное** или **глагол/прилагательное/наречие — наречие**, образующие определительные пары (*краснокочанная капуста, резко высказаться, полностью ясный, ужасно страшно*);
- **Причастие/прилагательное** — его прямое, косвенное или предложное **дополнение-существительное**, включая ходовые обстоятельства (*рассмотревший вопрос, ковырявший в носу, оставшийся из-за погоды, красный от гнева, купленный на рынке*);
- **Глагол** — его прямое, косвенное или предложное **дополнение-существительное,** включая ходовые обстоятельства (*рассмотреть вопрос, остаться из-за погоды, купить на рынке, отличаться сдержанностью*);
- **Деепричастие/наречие** — его прямое, косвенное или предложное **дополнение-существительное** (*рассмотрев вопрос, ковыряя в носу, купив на рынке, близко от города*);
- **Существительное-подлежащее —** его **сказуемое** в виде личной формы глагола или краткого прилагательного/причастия (*внимание (было) привлечено, доклад (был) краток, враг напал, глазки бегают);*
- **Существительное** — подчиненное ему **существительное** (*сердце матери, наложение взыскания, отличия в произношении, борьба с бюрократизмом*).

Менее частотны в словаре следующие типы словосочетаний:

- **Глагол** — его **инфинитивное дополнение** (*собраться поехать, мечтать выкупаться, хотеть перекусить)*;
- **Существительное** — его **инфинитивное дополнение** (*соблазн сказать, желание уйти, проблема выжить*);
- **Прилагательное/причастие** — его **инфинитивное дополнение** (*готовый действовать, желающий начать*);
- **Деепричастие/наречие** — его **инфинитивное дополнение** (*мечтая сказать, пожелав уйти, собравшись купаться*);
- **Глагол** — его **адъективное дополнение** (*вернуться здоровым, найти мертвым, считать выполненным*);
- **Прилагательное/причастие** — его **адъективное дополнение**: *найденный нормальным, вернувшийся здоровым*);
- **Деепричастие/наречие** — его **адъективное дополнение**: *найдя нормальным, считая выполненным*).
- Устойчивые **сочиненные пары** из одинаковых частей речи (*ясный и четкий, быть или не быть, власть и бизнес, в срок и в полном объеме, базы и склады, наука и техника*).

Статистика по базе словосочетаний выявляет следующие существительные с наибольшим числом управляющих глаголов:

514 *работа*1	387 *место*1	365 *руки*
463 *деньги*	377 *дом*1	339 *дело*1
413 *ребенок*	367 *дети*	333 *дорога*

Существительные с наибольшим числом определений:

1189 *человек*	548 *глаза*	520 *вид*1
715 *лицо*1	536 *женщина*	507 *режим*2
557 *работа1*	534 *взгляд1*	499 *голос*1

Глаголы с наибольшим числом дополнений:

2284 *быть*	1270 *стать*	963 *считать*
2185 *иметь*	1095 *начать*	959 *вести*
1442 *находиться*	1068 *получить*	951 *оказаться*

Прилагательные — наиболее частые определения:

2958 *большой*	1604 *новый*	1321 *явный*
2047 *крупный*	1463 *постоянный*	1202 *огромный*
1749 *небольшой*	1407 *полный*1	1183 *сильный*

Смысловые связи делятся на следующие группы:

- **Синонимы (**например, *дурак — болван*) в виде 19 тыс. синонимических групп в среднем по 5,6 элементов; односторонних синонимических связей 1,12 млн.
- **Семантические дериваты** — это группы типа {*Москва1, москвичи; московский…*} или {*извлечение; извлекать, извлечь; извлеченный, извлекший; извлекая, по извлечении, путем извлечения*}, односторонних связей 0,88 млн.
- **Часть/целое** (например, *террариум — зоопарк*), связей 21 тыс.
- **Род/вид** (например, *диплом1 — документ*), связей 14 тыс.
- **Антонимы** (например, *длинный — короткий*), связей 12 тыс. Они дополняются выдачей антонимов для синонимов данного слова и синонимов для антонимов.

Смысловые связи важны не только сами по себе. Во-первых, толковательные синонимы (а их в словаре много) способны разъяснить смысл слова, и это особенно важно при наличии глоссов только для омонимов. Так, у *кошерный* есть синоним *отвечающий иудейским нормам*. Во-вторых, с помощью семантически связанных слов можно построить из титулов словника новые правдоподобные словосочетания, непосредственно в базе не представленные, напр., получить новое словосочетание по формуле

(букет цветов) *&* ***(астры*** *IS_A* ***цветы)*** *→*
→ ***(букет астр)****.*

Паронимические связи в словаре представлены:

- **Буквенными** паронимами, т.е. словами той же части речи, отличающимися от данного слова на одну букву, например, *кадка*: {*кака, ка**с**ка, ка**ч**ка, ка**ш**ка, кладка…*}.
- **Морфемными** паронимами, имеющими ту же часть речи и общий корень, но иное сочетание аффиксов, напр., {***бег, бег**ун, **бег**а, **бег**лость, при**беж**ище, про**беж**ка…*}.

Практически для каждого изменяемого титула словаря, включая многословные, дается его морфологическая парадигма.

У титулов словника есть английские переводы. Собранные вместе, они образуют отдельный словник, через который можно войти в русскую часть словаря.

Для словосочетаний введены три градации фигуральности (идиоматичности). Если пометы фигуральности нет, словосочетание понимается как есть: *идти в школу, вызвать слесаря*. При помете **idiom** словосочетание понимается только фигурально (идиоматически): *сесть в галошу, висеть на волоске.* При помете **mb idiom** словосочетание понимается либо фигурально, либо в прямом смысле: *сесть в лужу, первая ракетка*.

Большее число градаций имеет стиль (степень разговорности) слов и словосочетаний.

- Если пометы стиля нет, слово (словосочетание) является нейтральным и достаточно употребительным, так что его полезно знать (*стена, окно, книга, налоги…*).
- Помета в виде зеленого буллита указывает специальное, книжное или забытое слово (словосочетание). Предлагается пользоваться им, если нет опасности непонимания слушателями: *абсцесс, парадигма, экзистенциальный,…*
- Желтый буллит указывается чисто разговорное слово или выражение, которым предлагается не пользоваться в официальных документах: *мотать нервы, жевать сопли,….*
- Красный буллит указывается бранное слово или выражение, которым нельзя пользоваться при дамах, детях и в официальной обстановке: *говно, жопа, засранец, мудак, взять за яйца,…*
- Черный буллит указывает нейтральное бытующее выражение, смысл которого лингвисты рекомендуют передавать более нормативно: *оплатить за проезд, проплатить операцию.…*

Пользователю КроссЛексики предлагаются некоторые опции, а именно

- Можно выбрать язык обращения со словарем:
  - **Русский** (компоненты меню, названия разделов выдачи, толкования омонимов и справочная информация даются по-русски), либо
  - **Английский** (все вышеуказанное дается по-английски).
- В процессе работы с КроссЛексикой можно:
  - Выбрать алфавитный порядок выдачи основных типов словосочетаний либо частотный порядок (словосочетания с более частотными в словаре элементами выдаются первыми);
  - Установить порог отсечения словосочетаний с низкочастотными в словаре титулами;
  - Отменить выдачу бранной, разговорной и/или специальной лексики вместе с соответствующими словосочетаниями;
  - Ввести запрос (1) с клавиатуры, (2) подведя указатель к нужной строке в окне словника, (3) выбрав строку в обновляющемся списке *История,* либо (4) выбрав строку в списке словосочетаний на экране. Последний вариант начинает навигацию по словнику.

3. Приложения словаря

Приложения словаря возможны:

- **Диалоговые** (интерактивные), когда пользователь обращается к словарю в диалоговом режиме и использует результаты, например, при параллельном редактировании текста или при обучении русскому языку;
- **Недиалоговые** (неинтерактивные), когда внешняя программа обращается к словарю за информацией и использует результаты для своих целей.

Вот примеры диалоговых запросов русскоязычного пользователя:

- Как можно выразиться глаголом о *плате за проезд? платить, оплатить, оплачивать проезд* либо *заплатить за проезд* (*проплатить проезд* и *оплатить за проезд* тоже даются на экране, но снабжены черным буллитом).
- Как можно еще назвать *бразильских женщин?* — *бразильянки*. *А* как *иракских женщин*? — Да никак иначе! (Но *иракец, иракцы* допустимы.)
- Как «запустить» иск? — *внести, возбудить, вчинить, подать* или *предъявить иск,* а также *обратиться с иском.*
- Как управляет существительными глагол *забыть?*
  - **забыть что/кого?** *забыть адрес, багаж, вкус, времена, время, вчерашнее,…* (101 словосочетание)
  - **забыть о чем/о ком?** *забыть о времени, обо всем, о вчерашнем, о главном,…* (37)
  - **забыть про что/про кого?** *забыть про все, про главное, про детей, про диссертацию, про семью,…* (22)
  - **забыть в чем/в ком/где?** *забыть в вагоне, в гостях, в комнате, в кафе, в ресторане, в спешке…* (10),
  - **забыть на чем/на ком/где?** *забыть на диване, на кресле, на кровати…* (7)

  - **забыть при чем/при ком?** *забыть при декларировании, при зачтении…* (3)
  - **забыть по чему/по кому?** *забыть по рассеянности, по невнимательности* (2)
  - **забыть из-за чего/из-за кого/почему?** *забыть из-за волнения, из-за спешки* (2)
  - **забыть за чем/за кем?** *забыть за давностью* (1),
  - **забыть от чего/от кого/откуда?** *забыть от волнения* (1)
- С какими существительными сочетаются морфемные паронимы **вероятный** Vs. **вероятностный**?

  вероятный определяет существительные *адрес, альтернатива, вариант, версия, встреча…*, а

  вероятностный — *автомат, алгоритм, анализ, анализатор, аспекты…*, и пересечение этих множеств ничтожно мало.
- С какими существительными сочетаются омографы **доменный1** Vs. **доменный2**?

  доменный1 определяет существительные *адрес, аукцион, бизнес, границы, зона, имена…*, а

  доменный2 — *газы, кокс, конструкция, мастера, печь…*, и пересечение этих множеств пусто.
- С какими существительными сочетаются квазиомографы **личный** Vs. **личной**?

  личный определяет существительные *автомашина, адъютант, амбиции, антипатии, архив…*, а

  личной — *карман, крем, напильник, нашивки, полотенце, салфетка…*, и пересечение этих множеств незначительно.
- Что означают и что определяют прилагательные *ретроактивный, проактивный, адвалорный, халяльный…*? На это отвечают их синонимы и связанные существительные.

Диалоговые запросы для уже продвинутого в русском языке иностранца включают все средства, предложенные русскоязычному пользователю, и многое иное:

- Среди сведений по орфографии и морфологии слов можно, например, узнать, что ***Христос*** склоняется особо.
- Можно увидеть сферы применения синонимов *малый* и *маленький*. Так, равно допустимы *малые дети* и *маленькие дети*, но возможны только *малый бизнес* и *маленькие апельсины*.
- При обращении через английский словарь в виде глагола ***pay*** будут получены русские глаголы ***обращать, обратить, окупать, окупить, оплатить, оплачивать, платить, уделить, уделять, уплатить, уплачивать***, и далее можно справиться о любом из них.

Среди недиалоговых приложений КроссЛексики отметим в первую очередь:

- **Облегчение парсинга.** В предложении ищутся все возможные словосочетания, имеющиеся в КроссЛексике, и чем больше обнаружено таких словосочетаний в данном варианте разбора предложения, тем вероятнее этот вариант.
- **Разрешение неоднозначности слово.** Ищутся словосочетания и семантические связи для отдельных омонимов, и выбирается омоним, для которого в контексте найдено наибольшее число синтаксически и семантически сочетающихся соседей.
- **Стеганография и стеганализ.** Сочетания и синонимы слов, встреченные в тексте, используются для регулируемой замены одних синонимов другими, так чтобы в этих заменах закодировать стороннюю информацию, тем самым тайно передающуются несущим текстом без изменения его смысла.
- **Идиоматичный перевод английских словосочетаний**. Например, в ответ на введенное *strong woman* словарь выдает *крепкая баба, сильная женщина…*
- **Информационный поиск**. Предполагается автоматически обогащать запрос не только семантически связанными словами, но и словами, формирующими высокочастотные словосочетания со словами запроса.

4. Источники и метод пополнения словаря, покрытие им текстов

Базовым методом подбора материала был ручной. На момент начала разработки ни корпусов русских текстов, ни интернетовских поисковиков, ни идей работы с ними просто не существовало. Однако на каждом этапе уже наличествующая версия словаря выявляла необходимость очередных его пополнений.

Основными источниками словосочетаний явились:

- Двуязычные словари (особо отметим словарь под ред. Ю.Д. Апресяна и русско-испанский словарь Г.Я. Туровера и Х. Ногейры);
- Академический четырехтомный словарь русского языка;
- Множество специализированных словарей по экономике, бизнесу, электронике, вычислительной технике и др.;
- Наблюдаемый шесть лет поток новостей, политических и научных статей портала *газета.ру*;
- Многочисленные справки по комбинаторике слов в Яндексе и Гуггле;
- Систематические сканирования текстов в рекламных буклетах, объявлениях по ремонту и строительству, в журналах для автомобилистов, в гламурной журналистике, в спаме.

Из Национального корпуса русского языка не было взято ничего. Он появился слишком поздно и вначале был очень небольшим, а для свободного поиска со статистическими оценками результатов недоступен и сейчас.

Методы автоматического извлечения коллокаций из корпусов и Интернета начали разрабатываться лишь в последние годы [8, 9]. Но они прямо не приложимы к высокофлективному языку, да и не дали новых словарей английских коллокаций.

Для проверки через интернет, стоит ли включать в КроссЛексику данное словосочетание, полученное откуда угодно, была предложена количественная мера [10], успешно применяемая к новым пополнениям КроссЛексики.

В части оценок покрытия КроссЛексикой отдельных слов и словосочетаний автоматических средств пока не разработано, но проводились ручные эксперименты. Если исключить названия организаций и географических объектов и личные имена, то покрытие знаменательных слов уже несколько лет близко к 100%. В части же покрытия словосочетаний произошедшее за последние 12 лет увеличение их числа в КроссЛексике в три раза привело к сдвигу примерно с 60% до 75%. Однако последняя цифра резко колеблется от текста к тексту, и требуются новые массовые независимые оценки. Двумерный закон Ципфа неумолим, и гарантированное покрытие хотя бы 80% словосочетаний потребует новых колоссальных усилий. Стопроцентное же покрытие едва ли возможно из-за авторской свободы употребить *ответственный за шишки, минюстовский блин комом* и под.

5. Заключение

Предложен новый словарный ресурс — комбинаторный словарь КроссЛексика, по объему и структуре не имеющий аналогов ни для одного языка. Он оставляет далеко позади единственный для английского языка словарь коллокаций [7] и существенно превышает русские словари [2] и [4].

При высочайшем покрытии лексики и сравнительно высоком покрытии словосочетаний, а также при простом доступе КроссЛексика предназначается для широкого круга пользователей.

Литература

1. *Квеселевич Д. А.* Толковый словарь ненормативной лексики русского языка. Москва: Астрель АСТ, 2005, 1022 стр.
2. *Комплексный* словарь русского языка. Под ред. А.Н. Тихонова. Москва: Русский язык медиа, 2007, 1230 стр.
3. *Крысин Л. П.* Толковый словарь иноязычных слов. Москва: Эксмо, 2008, 942 стр.
4. *Словарь* сочетаемости слов русского языка. Под ред. П. Н. Денисова и В. В. Морковкина. Москва: Русский язык, 1983, 686 стр.
5. *Толковый* словарь русского языка начала XXI века. Под ред. Г.Н. Скляревской. Москва: Эксмо, 2007, 1132 стр.
6. *Mel'čuk, I.* Phrasemes in Language and Phraseology in Linguistics. In: M. Everaert et al. (Eds.) Idioms: Structural and Psychological Perspectives. Lawrence Erlbaum Associates Publ., Hillsdale, NJ / Hove, UK, 1995, p. 169–252.
7. *Oxford* Collocations Dictionary for Students of English. Oxford University Press, 2003.
8. *Lin, Dekang.* Extracting Collocations from Text Corpora. First Workshop on Computational Terminology, Montreal, Canada, August, 1998.
9. *Kilgarriff, A., P. Rychlý, P. Smrz, D. Tugwell.* The Sketch Engine. Practical Lexicography: A Reader, Oxford University Press, UK, 2008, p. 297–306.
10. *Bolshakov, I. A., E. I. Bolshakova, A. P. Kotlyarov, A. Gelbukh.* Various Criteria of Collocation Cohesion in Internet: Comparison of Resolving Power. In: A. Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. Proc. 9th Intern. Conf. on Computational Linguistics CICLing-2008, Haifa, Israel. LNCS 3878, Springer, 2008, p. 95–116.

Создание семантического словаря предложных конструкций на основе Украинского национального лингвистического корпуса

Creating a semantic dictionary of prepositional constructions on the basis of the Ukrainian National Linguistic Corpus

Бугаков О. В. (ovbugakov@gmail.com)

Украинский языково-информационный фонд НАН Украины, Киев, Украина

Рассматриваются поисковые возможности Украинского национального лингвистического корпуса, а также создаваемые на его основе лингвистические базы данных. Описана структура электронного семантического словаря предложных конструкций, построенного в соответствии с теорией лексикографических систем.

В последние десятилетия наблюдается тенденция к лексикографированию языковых единиц, формально не являющихся единицами лексического уровня [4, 7]. Попытки лексикографирования семантических, синтаксических, когнитивных и других структур более высокого уровня, чем лексический, не только отражают общую тенденцию к лексикографическому описанию всех языковых явлений, но и отвечают требованиям практики по разработке усовершенствованных систем лингвистического обеспечения [6].

Целью нашего исследования являлось лексикографическое проектирование и создание электронного семантического словаря предложных конструкций в соответствии с принципами теории лексикографических систем. Основой для его создания служил Украинский национальный лингвистический корпус (УНЛК), созданный в Украинском языково-информационном фонде НАН Украины (УЯИФ). Объем корпуса — около 54 млн с/у. Корпус представлен текстами разных стилей и жанров без соблюдения пропорций. В случае необходимости исследователь может самостоятельно создавать подкорпуса отдельных стилей с учетом статистических параметров.

В УНЛК предусмотрены два типа поиска. Первый — по библиографическим реквизитам, второй — полнотекстовый поиск с использованием современных лингвистических технологий. Поиск по библиографическому описанию предназначен, в первую очередь, для отбора подмассива информации для последующей обработки.

Полнотекстовый поиск осуществляется после предварительной процедуры индексирования текстов в кодировке UNICODE, сопоставленных с объектами хранения электронной библиотеки. Для проведения полнотекстового поиска необходимо ввести поисковое словосочетание и задать параметры полнотекстового поиска. Полнотекстовый поиск может быть выполнен с учетом следующих параметров:

- поиск по текстам из текущей корзины;
- с учетом порядка слов;
- с лемматизацией;
- с учетом синонимии;
- по синонимическим рядам;
- по грамматическим параметрам;
- без учета расстояния между словами.

После проведения полнотекстового поиска пользователю предоставляется возможность просмотра локализаций поисковых фраз в выбранном тексте. При выборе одного из объектов результатов поиска происходит поиск контекстов внутри проиндексированного текста.

Поисковые слова контекста в тексте выделяются определенным цветом, например в локализации поисковой фразы *робити добро* красным цветом выделены словоформы *робити* и *добро*, отвечающие поисковой фразе при поиске с лемматизацией (рис. 1).

На материале УНЛК было проведено комплексное исследование функционирования украинских предлогов в украинском тексте на трех уровнях — морфологическом, синтаксическом и семантическом.

Достижение поставленной цели предусматривало решение ряда задач: 1) уточнение реестра предлогов на основе анализа украинских текстов УНЛК; 2) анализ омонимии предлога с другими частями речи в тексте; 3) установление текстовых

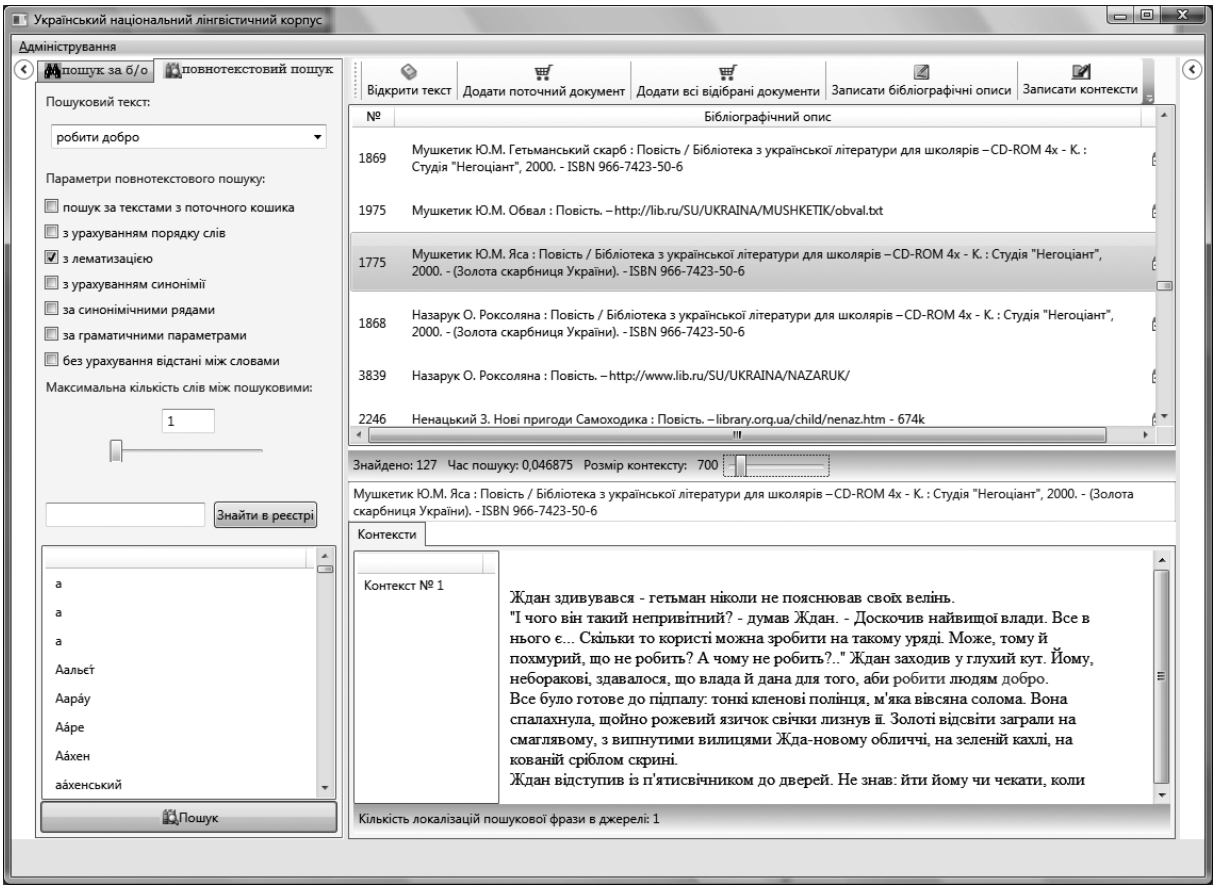


Рис. 1. Полнотекстовый поиск в УНЛК

условий снятия омонимии предлога с другими частями речи в тексте; 4) разработку алгоритма разграничения составных предлогов от сочетаний простого предлога с полнозначным словом (*с помощью, с целью*); 5) разработку алгоритма определения зон предложных связей в тексте как отдельного модуля автоматического синтаксического анализа; 6) установление семантических отношений между компонентами зоны предложных связей в системе автоматического семантического анализа; 7) создание семантического словаря предложных конструкций.

Исследование функционирования предлогов, как и любых других единиц языка, в тексте предусматривает использование статистических методов и метода контекстной диагностики. В связи с этим возникла необходимость проведения анализа на репрезентативном материале с целью обеспечения надлежащего уровня достоверности результатов. Таким материалом и служил УНЛК.

Программное обеспечение УНЛК позволяет создавать специализированные субкорпуса, ориентированные на решение поставленных заданий. С помощью специально разработанной в УЯИФе программы отмеченные субкорпуса переводятся в форматы баз данных с определенной структурой, ориентированной на проведение конкретных лингвистических исследований. Лингвистические базы данных (ЛБД), выполняющие функцию инструмента и материала исследования языкового явления, структурированы по следующему принципу: текстовые сегменты (контексты), которые содержат конкретную языковую единицу (предлог), ставятся в соответствие заранее определенным дифференциальным признакам, по которым осуществляется анализ. Структурирование ЛБД по полям, отвечающим множеству параметров анализа диагностирующих контекстов, и организация доступа к этим полям позволяют автоматически классифицировать материал по каждому из параметров и любой их комбинации.

В соответствии с поставленными выше задачами в УЯИФе создано три предложные ЛБД: лингвистическую базу предложных сочетаний — претендентов на роль составного предлога (ЛБСП), лингвистическую базу грамматических омографов с предложным компонентом (ЛБОП) и лингвистическую базу зон предложных связей (ЛБЗПЗ).

Первая из них (ЛБСП) построена на подкорпусе УНЛК объемом 23 млн с/у. Общая длина ЛБСП — 51025 контекстов [3]. Исходным материалом для второй базы — ЛБОП — служили морфологически размеченные тексты трех стилей (научный, художественный, публицистический), каждый из которых представлен выборкой в 1 млн с/у. Общий объем базы — 200123 контекста [1].

Исходным текстовым материалом для формирования третьей базы — ЛБЗПЗ — служил подкорпус текстов публицистического стиля объемом 6 млн с/у. Общий объем полученной базы — 20768 контекстов [2]. Поскольку именно на основе последней базы и был разработан электронный семантический словарь предложных конструкций, рассмотрим подробнее ее структуру.

База данных была создана для исследования функционирования предлогов на синтаксическом и семантическом уровнях. Она структурирована по полям, соответствующим множеству параметров, предварительно определенных прогнозирующими текстовыми признаками для алгоритмического установления зон предложных связей (ЗПС) при автоматическом семантико-синтаксическом анализе: 1) «Контекст», 2) «Длина ЗПС», 3) «Первая позиция предлога», 4) «Постпозиция ГС», 5) «Контактность ГС», 6) «Главное слово», 7) «Код ГС», 8) «Семантический класс ГС», 9) «Контактность ЗС», 10) «Зависимое слово», 11) «Код ЗС», 12) «Семантический класс ЗС», 13) «Отношение», 14) «Ремарки».

Зона предложных связей включает предлог, главное слово (слово, управляющее предложно-именной синтаксемой) и зависимое слово (слово, подчиняющееся главному с помощью предлога) [2]. На семантическом уровне зоны предложных связей рассматриваются с точки зрения семантической интерпретации синтаксической связи между ГС и ЗС.

При формировании словаря из базы были отброшены строки, в которых: 1) ГС расположено в постпозиции по отношению к предлогу, 2) ГС отсутствует — в случае вхождения предлога в эллиптическую конструкцию, 3) ЗС выражено именем существительным или другой частью речи, являющейся названием книги, газеты, организации и т. д., 4) предлог вошел в состав устойчивого сочетания. В результате использования созданного запроса количество строк в ЛБД сократилось до 13261. Именно столько словарных статей и содержит разработанный словарь.

Теоретическими предпосылками создания словаря является исследование семантики предлога в формализме теории семантических состояний [5], поскольку объектом лексикографирования выступают предложные конструкции, а элементами интерпретационной части — их семантические состояния.

Согласно теории семантических состояний, любое слово (языковая единица) контекста или языкового потока находится в определенном семантическом состоянии, которое для единиц лексического уровня является суммой признаков грамматической и лексической семантики [5]. Семантическое состояние предлога определяем как реализацию конкретного семантического отношения в тексте между главным и зависимым словами, обусловленную семантическими состояниями последних, которые представляют объекты внеязыковой действительности.

Проведение исследования с целью выявления множества типичных семантических состояний класса предлогов является необходимой предпосылкой создания указанного словаря. Установлению семантических состояний предлогов предшествует определение совокупности семантических состояний, которые выражают предлоги с синтаксически связанными с ними словами, с учетом семантических атрибуций ГС и ЗС в ЗПС.

По результатам нашего исследования, проведенного на ЛБД, было выделено 20 типов семантических отношений, которые могут выражать предлоги в тексте:

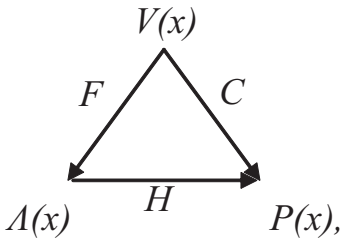
- Объектные (отношение действия к предмету, на который это действие направлено, или предмета к другому предмету, который является объектом действия первого: *за, на, о, замість, на_зміну, між, стосовно, щодо* и др. Этот тип отношений зафиксирован в 16,65% от общего количества ЗПС в ЛБД).
- Пространственные (отношение действия или предмета к месту, пространству, где происходит это действие или находится предмет: *в/у, перед, над, під, біля, поруч_з* и др. — 12,23%).
- Временные (отношение действия ко времени, в которое оно происходит, или явления, предмета ко времени своего существования: *перед, під_час, протягом, після* и др. — 9,29%).
- Условные (отношение действия, признака или предмета к обстоятельствам, условиям, при которых происходит действие или существует предмет: *з_урахуванням, залежно_від, на_випадок, у_випадку, у_разі* и др. — 8,51%).
- Причинные (отношение действия, предмета, явления к причине их возникновения или к их следствию: *унаслідок, у_результаті, у_зв'язку_з, на_ґрунті* и др. — 7,69%).
- Лимитивные (отношение действия к сфере его распространения или предмета к сфере его деятельности: *у галузі, у_межах, у_рамках* и др. — 7,5%).
- Отношения цели (отношение действия к другому действию, явлению, предмету, являющихся целью выполнения этого действия, или к предмету, в интересах которого происходит действие: *в_ім'я, в_інтересах, для, заради* и др. — 7,41%).
- Отношения направления движения (отношение движения в указанном направлении: *до, у_напрямі, усередину, назустріч, мимо* и др. — 5,17%).
- Комитативные (отношение: 1. действия к другому действию, сопровождающему первое, 2. действия к предмету, действием которого сопровождается первое действие, 3. между двумя предметами (лицами), являющихся общими исполнителями или объектами определенного действия: *одночасно_з, паралельно_з, при, спільно_з* и др. — 4,02%).

- Субъектные (отношение действия или предмета к другому предмету, который выполняет действие или является его потенциальным исполнителем: *від_імені, з_боку, за* и др. — 3,67%).
- Отношения способа действия (отношение действия к средству или способу его осуществления: *за_допомогою, за_рахунок, через* и др. — 3,31%).
- Коррелятивные (отношение соответствия действия, состояния, проявления признака к предмету или явлению: *відповідно_до, стосовно_до, згідно_з* и др. — 3,14%).
- Атрибутивные количественные (отношение предмета, свойства, действия к количественному признаку: *біля, близько, понад, коло* и др. — 2,89%).
- Сравнительные (отношение признака, предмета или состояния к другому признаку или предмету, с которыми сравниваются первые: *на_зразок, подібно_до, понад, порівняно_з* и др. — 2,74%).
- Атрибутивные качественные (отношение предмета к его качественному признаку: *з, у_вигляді, з_погляду* и др. — 1,57%).
- Отношения назначения (отношение предмета или явления быть назначением для другого предмета или действия: *для, у_справах* и др. — 1,4%).
- Генеративные (отношение предмета или лица к другому предмету или лицу, указывающих на происхождение первых, или отношения действия к предмету или лицу, являющихся источником этого действия: *від, з, з-під, з-поза* и др. — 1,38%).
- Партитивные (отношение части к целому: *до, в/у, від, з, з-поміж* — 0,89%).
- Функциональные (отношение действия к предмету, на выполнение функций которого направлено действие: *у_ролі, в/у, за* — 0,29%).
- Трансгрессивные (отношение действия, обозначающего преобразование, к предмету, который является результатом или источником этого преобразования: *до, в/у, на, з* — 0,26%).

В тексте каждый из выделенных типов реализует, как правило, определенное множество конкретных отношений в зависимости от семантического состояния конкретного предлога, определяющегося в тексте согласно теории семантических состояний шестью параметрами: релятивной семантикой самого предлога (квазилексическое значение), падежом ЗС, которым управляет предлог (квазиграмматическое значение), а также лексической и грамматической семантикой ГС и ЗС. С учетом указанных параметров в пределах базы было выделено 131 семантическое отношение.

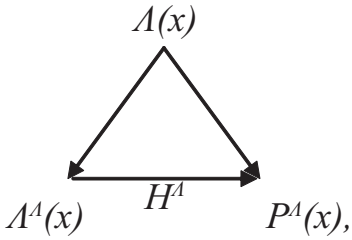
Полученная система семантических отношений предлогов легла в основу электронного семантического словаря предложных конструкций. Интерпретация конкретных предложных сочетаний при создании словаря выводилась путем сопоставления грамматической и семантической информации главного и зависимого слов и потенциальных возможностей предлога передавать семантические отношения.

Словарь представляет собой реализацию определенной лексикографической системы [8]. Структуру словарной статьи $V(x)$ семантического словаря можно представить в виде:



(1)

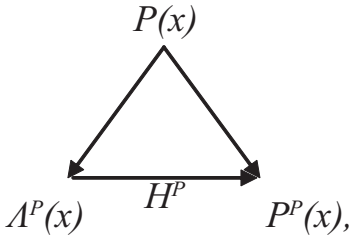
где x — реестровая единица словаря (предложная конструкция); F и C — операторы, выделяющие в тексте формальную и содержательную части описания реестровой единицы x — $\Lambda(x)$ и $P(x)$, соответственно; H — оператор, обеспечивающий соответствие между $\Lambda(x)$ и $P(x)$; элемент $\Lambda(x)$ играет роль левой (реестровой), а $P(x)$ — правой (интерпретационной) частей словарной статьи $V(x)$. Каждая из указанных частей, в свою очередь, делится на левую и правую части, т.е. происходит рекурсивная редукция второго порядка. Структура левой части получает вид:



(2)

где $\Lambda^{\Lambda}(x)$ содержит последовательность символов, а $P^{\Lambda}(x)$ содержит информацию об изъятии из этой последовательности символов трех знаковых репрезентантов компонентов предложной конструкции — k_1 (ГС), k_2 (предлог) и k_3 (ЗС).

Структура правой части имеет вид:



(3)

где в $A^p(x)$ фиксируются грамматические компоненты семантических состояний ГС, предлога и ЗС, а также квазиграмматический компонент семантического состояния предлога, в $P^p(x)$ — лексические и квазилексические компоненты их семантических состояний.

Связи между компонентами семантических состояний ГС, предлога и ЗС можно представить так:

$$\begin{gathered} H(k_1) \\ G(k_1) \rightarrow L(k_1) \\ H(k_2) \\ G(k_2) \rightarrow L(k_2) \\ H(k_3) \\ G(k_3) \rightarrow L(k_3), \end{gathered} \qquad (4)$$

где $G(k_1)$, $G(k_2)$, $G(k_3)$ — грамматические и квазиграмматические компоненты семантических состояний ГС, предлога и ЗС, $L(k_1)$, $L(k_2)$, $L(k_3)$ — лексические и квазилексические компоненты семантических состояний ГС, предлога и ЗС, $H(k_1)$, $H(k_2)$, $H(k_3)$ — операторы, обеспечивающие связи между этими компонентами семантических состояний составляющих предложной конструкции.

В процессе разработки словарной статьи было решено подавать грамматические и квазиграмматические компоненты семантических состояний ГС, предлога и ЗС в левой части словарной статьи, что обусловлено принципом экономичности.

В соответствии с этим схему словарной статьи словаря можно представить следующим образом:

ГС *ч.р.* **ПРЕДЛОГ ЗС** *ч.р. в …пад.* || семантический класс ГС; семантический класс ЗС; семантическое отношение.

ГС подается в начальной форме, а ЗС — в форме того падежа, которого требует конкретный предлог. Компонент, отвечающий грамматическому значению предлога, а именно, падеж зависимой именной формы, подается не возле предлога, а в блоке грамматического значения ЗС. Для увеличения визуального эффекта левая часть отделена от правой двумя прямыми черточками. Вместо семантических отношений подаются только их названия, а сами толкования поданы в отдельном файле. Такая форма представления будет упрощать автоматический поиск предложных конструкций по параметру семантического отношения.

Проиллюстрируем изложенное на примере словарной статьи:

ПРИХОДИТИ *дієсл.* **НА БАЗАР** *ім. у знах. в.* || рух; місце; Відношення напрямку руху 1.

Структурогенными компонентами левой части являются:
k_1 — **ПРИХОДИТИ** (ГС), k_2 — **НА** (предлог), k_3 — **БАЗАР** (ЗС).

Структурогенными компонентами правой части являются:

$G(k_1)$ — «глагол» (грамматический компонент семантического состояния ГС), $G(k_2)$ — *знах. в.* — «винительный падеж» (квазиграмматический компонент семантического состояния предлога), $G(k_3)$ — *ім.* — «имя существительное» (грамматический компонент семантического состояния ЗС).

$L(k_1)$ — рух (лексический компонент семантического состояния ГС), $L(k_3)$ — місце (лексический компонент семантического состояния ЗС), $L(k_2)$ — Відношення напрямку руху 1 (квазилексический компонент семантического состояния предлога).

Структурирование словарной статьи разработанного нами словаря предусматривает возможность поиска информации по всем выделенным структурогенным компонентам: по знаковому представлению предлога, ГС и ЗС, по их грамматическим показателям, то есть по частеречной принадлежности, по падежу ЗС, по семантическим классам ГС и ЗС и по семантическим отношениям. Поиск возможен как по отдельным параметрам, так и по их совокупности.

Словарь представляет собой открытую систему, которая может постоянно пополняться новыми предложными конструкциями.

Понятно, что возникает вопрос в целесообразности создания такого типа словаря. Кроме того, что словарь можно использовать в дальнейших исследованиях, в частности для изучения синонимии и антонимии предлогов, словарь может быть использован в системах автоматической обработки текста, в частности в лингвистических анализаторах как источник лексической информации при идентификации ЗПС в тексте. Рассмотрим возможность использования словаря в синтаксическом анализаторе. На этапе синтаксического анализа обращение к словарю происходит в случае, когда на основе грамматических, позиционных и семантических признаков не удается однозначно установить главное слово в ЗПС из-за присутствия нескольких формальных претендентов на роль ГС. Например, в предложении:

«Але<SC> ***проходження<NA>*** *програми<FB> дій<FI> уряду<MB>* ***через<PD>*** ***парламент<MD>*** *показало<VT>, що<SS> про<PD> взаєморозуміння<ND> та<SC> взаємопідтримку<FD> не<Z0> йдеться<YO>.<e>»*

у предлога *через* есть несколько претендентов на роль ГС — имена существительные в препозиции *проходження, програм, дій, уряду* и глагол-сказуемое в постпозиции *показало*, которые входят в пределы интервала поиска ГС. Идентификация претендента *проходження* с конструкцией в словаре указывает на то, что именно эта словоформа является главным словом в зоне связей предлога *через*.

Литература

1. *Бугаков О. В.* Аналіз граматичної омонімії прийменників у мові й у тексті // Мовознавство. К.: 2004, № 5–6, С. 87–98.
2. *Бугаков О. В.* Зони прийменникових зв'язків у синтаксичній структурі українського речення // Мовознавство. К.: 2005, № 5, С. 75–87.
3. *Бугаков О. В., Грязнухина Т. А., Рабулец А. Г.* Формирование предложных текстоориентированных баз данных на корпусе украинских текстов // Труды международной конференции «MegaLing'2005. Прикладная лингвистика в поиске новых путей». СПб.: Изд-во «Осипов», 2005. С. 11–16.
4. *Золотова Г. А.* Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. М.: Наука, 1988.
5. *Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна* та ін. К.: Довіра, 2005.
6. *Рабулець О. Г., Сухарина Н. М., Широков В. А., Якименко К. М.* Дієслово в лексикографічній системі. К.: Довіра, 2004 .
7. *Русский* семантический словарь. Толковый словарь, систематизированный по классам слов и значений / Под ред. Н. Ю. Шведовой. М., 1998. Т. 1.
8. *Широков В. А.* Інформаційна теорія лексикографічних систем. К.: Довіра, 1998.

Выделение фрагментов в текстах при классификации

Markup of text fragments during classification

Васильев В. Г. (vvg_2000@mail.ru)

Институт проблем информатики РАН

В работе проводится сравнительный анализ подходов к выделению значимых фрагментов в текстах в процессе автоматической классификации. Рассматриваются новые алгоритмы на основе скрытой марковской модели, покрытия текста специальным иерархическим множеством фрагментов и предварительной сегментации текстов.

1. Введение

При решении задач автоматической классификации текстовых данных одной из важных задач является представление и объяснение результатов классификации пользователю. В частности, достаточно важным для понимания причин отнесения текста к определенной рубрике является выделение в нем релевантных ей фрагментов (особенно это актуально в случае классификации политематических документов).

В случае использования лингвистического подхода (т. е. правила отнесения текстов к рубрике описываются с помощью некоторого информационно-поискового языка) такое выделение является достаточно простой задачей, которая решается путем отбора предложений, удовлетворяющих введенному правилу. Однако при использовании статистического подхода к классификации ее решение значительно усложняется. Это связано с тем, что в данном случае документ обычно представляется в виде одного вектора весов информационных признаков для всего текста, а также с отсутствием обучающих массивов с эталонным делением текстов на фрагменты.

Введем необходимые обозначения. Пусть $\omega_1, ..., \omega_k$ рубрики иерархического классификатора, задающие темы, которые представляют интерес и которые требуется автоматически выделять в текстах, $X = (X_1, ..., X_n)$ — текст на естественном языке, состоящий из *n* предложений, X_i — вектор весов слов в предложении $i = 1, ..., n$ размерности *m*, где *m* — общее число слов в тексте. Требуется для каждой рубрики ω_j, $j = 1, ..., k$, определить факт наличия в тексте информации по ней и в случае положительного решения найти предложения, ей соответствующие. Иными словами, для рубрики ω_j, $j = 1, ..., k$, требуется найти вектор: $t_j = (t_{j1}, ..., t_{jn})$, где

$$t_{ji} = \begin{cases} 1, & X_i \in \omega_j, \\ 0, & X_i \notin \omega_j. \end{cases}$$

Задача выделения значимых фрагментов тесно связана со следующими классическими задачами анализа текстов:

- классификация текстов — определение принадлежности текстов к рубрикам классификатора (в данном случае производится оценка отдельных предложений);
- сегментация текстов — разделение текстов на тематически однородные фрагменты (в данном случае производится выделение в тексте групп предложений, относящихся к одной теме);
- реферирование текстов — выделение значимых предложений в тексте с целью построения его краткого изложения (в данном случае производится выделение не всех значимых предложений, а только относящихся к определенной тематике).

Основной сложностью при выделении значимых фрагментов в текстах является то, что в общем случае оценку принадлежности предложения к рубрике нельзя проводить без учета соседних предложений. Например, возможна ситуация, когда фрагмент текста X, состоящий из предложений $X_i, X_{i+1}, ..., X_{i+s}$, относящийся к рубрике

ω_j, $j = 1, ..., k$, при классификации целиком будет отнесен к данной рубрике, а при классификации предложений $X_i, X_{i+1}, ..., X_{i+s}$ по отдельности ни одно из них может быть не отнесено к данной рубрике ω_j. Также возможна и обратная ситуация, что предложение X_i относится к рубрике ω_j, но при рассмотрении вместе с соседними предложениями оно уже не будет отнесено к данной рубрике.

Рассмотрим более подробно методы сегментации текстов, которые являются наиболее близкими по своему содержанию к задаче, решаемой в настоящей работе. Можно выделить следующие основные подходы к их построению:

- процедурный подход;
- структурный подход;
- вероятностный подход;
- оптимизационный подход.

Процедурный подход основан на построении правил, учитывающих различные элементы текста: отступы строк, знаки препинания, ключевые слова, референтные связи между словами, а также различные элементы оформления документов (заголовки, разделы, параграфы). Данный подход оказывается эффективным только в том случае, если формат обрабатываемых документов является известным.

Структурный подход основан на использовании различных мер близости между предложениями или фрагментами текста. При этом возможно как вычисление простейших статистик совместной встречаемости слов в различных блоках текста, так и использование методов кластерного анализа. При этом наибольшее распространение получил метод скользящего окна или «перекрывающегося текста» (text-tiling) [2], основанный на нахождении мест в тексте, где мера близости между двумя соседними блоками предложений минимальна. Также интересным является подход, основанный на использовании дивизимного алгоритма иерархического кластерного анализа [3].

Вероятностный подход для сегментации текстов основан на построении различных вероятностных моделей порождения слов в текстах. На практике наибольшее распространение получило представление текстов с помощью скрытых марковских моделей [5, 1]. В частности, в работе [5] рассматривается задача разделения составного текста, представляющего запись передач новостей по радио, на отдельные новостные сообщения. В данном случае открытым состояниям скрытой марковской модели, формальное определение которой будет дано далее, соответствуют отдельные слова в тексте, а скрытым состояниям — позиции слов в отдельных сообщениях, т. е. первым словам каждого сообщения будет соответствовать состояние с номером 1.

Оптимизационный подход основан на задании некоторого показателя качества разделения текста на фрагменты и нахождении такого разделения, которое обеспечивало бы его максимум. В частности, в работе [4] показатель качества зависит от длины фрагмента текстов и степени близости соседних фрагментов текстов. Для нахождения максимума показателя используется алгоритм на основе методов динамического программирования.

Недостатком приведенных подходов к сегментации является то, что при разделении текста на фрагменты в них не учитывается известная информация о тематиках, которые интересуют пользователя (в нашем случае его интересы заданы в виде набора рубрик классификатора). Рассмотрим несколько возможных подходов к решению задачи разметки текстов, которые, с одной стороны, используют идеи, приведенных выше подходов к сегментации текстов, а с другой стороны, лишены указанного недостатка и опираются на использование ранее обученных классификаторов.

2. Выделение фрагментов путем построения иерархического покрытия текста

Пусть текст $X = (X_1, ..., X_n)$ представляется в виде множества векторов

$$F = \left\{ \sum_{t=l_1}^{l_2} X_t \mid 1 \le l_1 \le l_2 \le n \right\},$$

где X_i — вектор весов слов (словосочетаний) предложения $i = 1,...,n$, n — число предложений в тексте. Данное множество включает в себя множество всех непрерывных фрагментов (последовательностей предложений без пропусков), содержащихся в тексте. Для оценки степени соответствия предложения текста рубрике можно использовать следующие выражения:

$$w_{ij} = \max_{Y \in F, X_i \in Y} g_j(Y) \text{ или}$$

$$w_{ij} = \underset{Y \in F, X_i \in Y}{mean}\, g_j(Y),\ i = 1, ..., n,\ \ j = 1, ..., k,$$

где $g_j(Y)$ — функция, осуществляющая вычисление степени соответствия вектора Y рубрике ω_j, $j = 1, ..., k$, построенная в результате обучения некоторого статистического классификатора.

Таким образом, вес w_{ij} предложения X_i, $i = 1, ..., n$, для рубрики ω_j равен максимальному (среднему) значению степени близости к данной рубрике ω_j всех фрагментов, содержащих данное предложение. Несложно заметить, что в данном случае для выделения значимых фрагментов в тексте для каждой рубрики требуется выполнить классификацию $\frac{n \cdot (n+1)}{2}$ фрагментов, т. е. вычислительная сложность составляет порядка $O(n^2)$.

Для снижения вычислительной сложности можно воспользоваться представлением текста X в виде следующего иерархически упорядоченного множества векторов фрагментов (иерархического покрытия):

$$H = H_0 \cup \left(\bigcup_{t=1}^{\lceil \log_2(n) \rceil} H_t \right),$$

$$H_t = \left\{ \sum_{i=1+l2^{t-1}}^{\min\left(l2^{t-1}+2^t, n\right)} X_i \mid l = 0,..., \left\lceil \frac{n}{2^{t-1}} - 1 \right\rceil \right\},$$

$$H_0 = \{X_1, ..., X_n\},$$

где $\lceil x \rceil = \min(l \in Z \mid l \geq x)$.

Несложно заметить, что для мощности множества H справедливы следующие соотношения:

$$|H| = |H_0| + \sum_{t=1}^{\lceil \log_2(n) \rceil} |H_t| \leq n + \sum_{t=1}^{\lceil \log_2(n) \rceil} \left(\frac{n}{2^{t-1}} + 1 \right) \leq$$

$$\leq \log_2(n) + n + 1 + n \sum_{t=1}^{\infty} \frac{1}{2^{t-1}} \leq \log_2(n) + 1 + 3n$$

Таким образом, при использовании иерархического покрытия H вычислительная сложность нахождения степени соответствия текста отдельной рубрике составляет порядка $O(n)$.

Для построенного иерархического покрытия H справедлива следующая теорема, которая говорит о качестве аппроксимации полного множества фрагментов F с помощью множества H.

Теорема. (Об иерархическом покрытии.)

Для любого фрагмента $Y \in F$ существует фрагмент $Z \in H$ такой, что

$$\frac{|Y \Delta Z|}{|Y|} \leq \frac{1}{2}.$$

Доказательство. Рассмотрим следующие три случая для числа предложений во множестве $Y \in F$.

1. Пусть $|Y| = 2^l$, $l \in \{0, 1, ..., \lceil \log_2(n) \rceil\}$.

Тогда $Y = \sum_{i \in [s, s+2^l-1]} X_i$, где $s \in \{1, ..., n - 2^l + 1\}$,

и существует $Z = \sum_{i=1+v2^{l-1}}^{\min\left(v2^{l-1}+2^l, n\right)} X_i \in H$ такое,

что $\left| s - \left(1 + v2^{l-1}\right) \right| \leq 2^{l-2}$,

v — неотрицательное целое, т. е. начало фрагмента соответствующего Z расположено не далее чем в 2^{l-2} предложениях от начала фрагмента Y.

Отсюда получаем, что $|Y \Delta Z| \leq 2 \cdot 2^{l-2} = \frac{1}{2}|Y|$.

2. Пусть $|Y| \geq 2^l + 2^{l-1}$ и $|Y| < 2^{l+1}$.

$$Y = \sum_{i \in [s, s+2^l+2^{l-1}+\delta-1]} X_i \text{, где}$$

$$s \in \{1, ..., n - 2^l - 2^{l-1} - \delta + 1\},$$

$$\delta \in [1, ..., 2^{l-1} - 1],$$

и существует $Z = \sum_{i=1+v2^{l-1}}^{\min\left(v2^{l-1}+2^l, n\right)} X_i \in H$ такое,

что $Z \subset Y$, v — неотрицательное целое.

Отсюда получаем, что

$$|Y \Delta Z| = |Y \setminus Z| \leq 2^{l-1} + 2^{l-1} = \frac{1}{2}|Y|.$$

3. Пусть $|Y| < 2^l + 2^{l-1}$ и $|Y| > 2^l$.

Тогда $Y = \sum_{i \in [s, s+2^l+\delta-1]} X_i$,

где $s \in \{1, ..., n - 2^l - \delta + 1\}$,

и существует $Z = \sum_{i=1+v2^{l-1}}^{\min\left(v2^{l-1}+2^l, n\right)} X_i \in H$ такое,

что $\left| s - \left(1 + v2^{l-1}\right) \right| \leq 2^{l-2}$, v — неотрицательное целое.

Отсюда получаем, что

$$|Y \Delta Z| \leq 2 \cdot 2^{l-2} - \delta \leq \frac{1}{2}|Y|. \blacksquare$$

Таким образом, для каждого фрагмента Y из F существует фрагмент из H, который отличаются от него не более чем на половину его длины. В целом схема алгоритма классификации текста принимает следующий вид.

3. Схема работы алгоритма на основе иерархического покрытия

1. Построение иерархического покрытия множества предложений текста.
2. Выполнение независимой классификации фрагментов текста с использованием ранее обученного классификатора.
3. Вычисление итоговых весов предложений в тексте путем объединения результатов классификации фрагментов, входящих в покрытие.
4. Отбор предложений, вес которых выше некоторого порога. ■

4. Выделение фрагментов с использованием скрытой марковской модели

В соответствии с работой [6] скрытая марковская модель с дискретным временем определяется как набор следующих элементов:

$S = \{s_1, ..., s_N\}$ — множество скрытых состояний;

$q_1, ..., q_n \in S$ — последовательность скрытых состояний;

A — матрица переходных вероятностей размера $N \times N$, где $a_{ij} = p(q_t = s_j \mid q_{t-1} = s_i)$.

U — пространство наблюдаемых состояний;

$y_1, ..., y_n \in U$ — последовательность наблюдаемых состояний;

$f_1(u), ..., f_N(u)$ — условные функции распределения для состояний s_i, $i = 1, ..., N$;

$\pi = (\pi_1, ..., \pi_N)$ — вектор начальных вероятностей скрытых состояний.

Рассмотрим теперь как с использованием аппарата скрытых марковских моделей можно производить выделение в тексте *X* предложений, соответствующих отдельной рубрике ω_r, $r = 1, ..., k$.

В данном случае пространство наблюдаемых состояний $U = R^m$, где m — размерность словаря признаков для обучающей выборки, а элементы последовательности наблюдаемых состояний $y_i = X_i$, $i = 1, ..., n$.

Множество скрытых состояния $s = \{s_1, s_2, s_3, s_4\}$ определим следующим образом:

s_1 — предложение находится внутри фрагмента, соответствующего рубрике w_r;

s_2 — предложение находится в начале фрагмента, соответствующего рубрике w_r;

s_3 — предложение находится в конце фрагмента, соответствующего рубрике w_r;

s_4 — предложение не относится к рубрике ω_r.

Так на этапе обучения эталонное распределение текста на фрагменты отсутствует, то матрицу переходных вероятностей A зададим априори. В данной работе при проведении экспериментов использовалась следующая матрица

$$A = \begin{pmatrix} 0.8 & 0 & 0.2 & 0 \\ 0.8 & 0.2 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0 & 0.5 \end{pmatrix}.$$

Начальные вероятности скрытых состояний положим равными друг другу, т. е. $\pi_j = 1/4$, $j = 1,...,4$.

Основную сложность при построении данной модели составляет задание условных функций распределения (плотности) $f_j(u)$ для состояний s_j, $j = 1,...,4$,
и вычисление их значений $w_{ij} = f_j(X_i)$.

В настоящей работе значения данных функций определяются следующим образом. Каждому предложению X_i ставятся в соответствие следующие вектора:

$$y_i = \sum_{t=-b}^{b} w_t X_{i+t},\ y_i^- = \sum_{t=-b}^{0} w_t X_{i+t},\ y_i^+ = \sum_{t=0}^{b} w_t X_{i+t}$$

где b — константа, задающая размер блока (число учитываемых предложений справа и слева от предложения с номером i), w_t — вес предложения с индексом $i + t$, $i = 1, ..., n$,

$w_t = (1 + |t|)^{-1}$, $w_t = 1$, t — целое.

Отсюда w_{ij}, $i = 1,...,n$, определяются следующим образом:

$$w_{i1} = p(y_i \mid \mathsf{w}_r),$$
$$w_{i2} = p(y_i^+ \mid \mathsf{w}_r),$$
$$w_{i3} = p(y_i^- \mid \mathsf{w}_r),$$
$$w_{i4} = (1 - p(y_i \mid \mathsf{w}_r)).$$

Для определения последовательности скрытых состояний используется стандартный алгоритм динамического программирования (алгоритм Витерби [6]), который находит наиболее правдоподобную последовательность из всех возможных. Таким образом, Общая схема работы алгоритма выделения и классификации фрагментов в результате приобретает следующий вид.

5. Схема работы алгоритма на основе скрытой марковской модели

1. Выделить предложения в тексте и сопоставить им векторы информационных признаков.
2. Произвести классификацию отдельных предложений с помощью обученного ранее статистического классификатора.
3. С использованием алгоритма Витерби [6] произвести оценивание скрытых состояний цепи Маркова по наблюдаемым состояниям.
4. Произвести разметку текста с использованием найденных скрытых состояний. ■

6. Выделение фрагментов путем независимой сегментации

В данном случае сначала выделение фрагментов производится путем сегментации текста на тематически однородные фрагменты без учета информации о структуре рубрик классификатора. Затем проводится классификация выделенных фрагментов и отбираются те из них, которые удовлетворяют рубрикам классификатора. В настоящей работе для сегментации текстов было решено остановится на методах, описанных в работах [2] и [3].

В первом методе, который называется Text Tiling, сегментация текста проводится следующим образом. Для каждого промежутка между предложениями вычисляется косинусная мера близости блока из r предложений справа и слева от него. В результате формируется вектор s размерности n, где n — число предложений, r — размер блока (в экспериментах использовалось значение $r = 4$). Далее значения вектора s сглаживаются с использованием метода скользящего среднего с различными параметрами и находятся точки локального минимума, которые и используются в качестве границ фрагментов.

Второй метод основан на использовании иерархического алгоритма кластерного анализа, который последовательно формирует разбиения данных на кластеры, начиная с ситуации, когда в одном кластере содержаться все наблюдения, и заканчивая ситуацией, когда каждое наблюдение образует отдельный кластер. Схема вычислений в данном случае следующая.

7. Схема алгоритма сегментации на основе иерархического кластерного анализа

1. Вычисляется матрица $C = (c_{ij})$ размерности $n \times n$,

элементы которой $c_{ij} = \dfrac{X_i^T X_j}{\|X_i\|_2 \|X_j\|_2}$,

$i, j = 1,\ldots,n$.

2. Строится матрица локальных весов $R = (r_{ij})$ размерности $n \times n$, где

$$r_{ij} = \frac{\left|\forall p \in [1, i-r, i+r], q \in [j-r, j+r], p, q \in 1,\ldots,n : m_{ij} > m_{pq}\right|}{(2r+1)^2}$$

$$i, j = 1,\ldots,n,$$

r — константа, задающая размер окна, в рамках которого вычисляются локальные веса.

Осуществляется рекурсивное разбиение существующих фрагментов на части, начиная с ситуации, когда все предложения относятся к одному фрагменту, и, заканчивая ситуацией, когда перестанет возрастать функция среднего значения внутрифрагментной близости μ_T, которая определяется следующим образом

$$\mu(T) = \frac{\sum_{k=1}^{|T|} \sum_{i \in t_k} \sum_{j \in t_k} r_{ij}}{\sum_{k=1}^{|T|} |t_k|^2}, \text{ где}$$

$T = \{t_1,\ldots,t_{|T|}\}$ — множество фрагментов в тексте. ■

В целом схема работы алгоритма классификации фрагментов текста в данном случае принимает следующий вид.

8. Схема работы алгоритма на основе независимой сегментации

1. Выделить предложения в тексте и сопоставить им векторы информационных признаков.
2. Произвести разбиение предложений текста на непересекающиеся фрагменты с использованием алгоритма сегментации.
3. Выполнить классификацию выделенных фрагментов с использованием обученного ранее классификатора и отобрать фрагменты, удовлетворяющие хотя бы одной рубрике классификатора.
4. Произвести разметку текста по результатам классификации фрагментов. ■

9. Экспериментальная оценка эффективности подходов к выделению фрагментов

Для оценки влияния процедур выделения фрагментов на итоговое качество классификации были проведены два эксперимента с массивом нормативно-правовых документов, используемом в рамках семинара РОМИП в 2004 году.

10. Эксперимент по оценке качества классификации с учетом деления на фрагменты

В первом эксперименте из-за отсутствия эталонного массива с размеченными фрагментами производилась оценка качества классификации текстов целиком с учетом выделения фрагментов различными алгоритмами. Для обучения и оценивания

использовались документы, входящие обучающее множество коллекции РОМИП-2004. Данное множество было преобразовано следующим образом. Сначала были отобраны 44 рубрики, содержащие не менее 50 документов. Затем множество отобранных документов было разбито в пропорции 80% на 20% для построения нового обучающего и тестового множеств.

Для классификации текстов использовались метод вероятностной классификации на основе смеси распределений фон Мизеса-Фишера (VMF) и метод на основе машин опорных векторов (SVM). Фрагменты размером менее 4 предложений при классификации не учитывались. Множество рубрик, к которым относится текст целиком, формировалось путем объединения множеств рубрик, к которым были отнесены отдельные фрагменты. Таким образом, в результате классификации текста ему ставится некоторое множество рубрик. Оценка качества в данном случае производится с использованием стандартных коэффициентов точности, полноты и F-меры с использованием микро усреднения.

Результаты оценки качества классификации приведены в следующей таблице. В ней используются следующие обозначения: NONE — классификация без выделения фрагментов; HIER — выделение фрагментов на основе построения иерархического покрытия; HMM — выделение в тексте фрагментов с использованием скрытой марковской модели; TILE — выделение фрагментов путем предварительной сегментации текста с помощью алгоритма Text Tile; DIV — выделение фрагментов путем предварительной сегментации текста с использованием алгоритма дивизимного кластерного анализа. Для каждого метода через запятую приводятся показатели: P — точность, R — полнота, F — F-мера.

Таблица 1. Качество классификации с выделением фрагментов

Метод	SVM (P, R, F)	VMF (P, R, F)
NONE	36%, 60%, 43%	22%, 78%, 32%
HIER	29%, 70%, 38%,	15%, 85%, 24%
HMM	26%, 67%, 36%	9%, 91%, 16%
TILE	30%, 61%, 40%	15%, 83%, 25%
DIV	34%, 63%, 41%	16%, 81%, 26%

Анализ результатов эксперимента, приведенного в таблице 1 позволяет сделать вывод, что использование практически всех методов выделения фрагментов приводит к повышению полноты классификации, что является следствием рассмотрения большего количества элементов текста. При этом наиболее повышение полноты достигается в случае использовании метода HIER, а наименьшее при использовании метода TILE.

11. Эксперимент по оценке качества выделения фрагментов в текстах

Во втором эксперименте для оценки качества выделения фрагментов было сформировано искусственное тестовое множество $\widetilde{T} = \{\widetilde{T}_1,...,\widetilde{T}_N\}$, элементы которого $\widetilde{T}_j = [T_{j_1}, T_{j_2},..., T_{j_h}]$ получаются в результате конкатенации случайного набора из h текстов из множества T. В данном случае N — число текстов во множестве $\widetilde{T}$, $j_1,..,j_h$ — индексы случайно выбранных с возвращением текстов из множества T, T — множество таких текстов из тестового множества, используемого в первом эксперименте, которые относятся только к одной рубрике.

Каждому тексту $\widetilde{T}_j = [T_{j_1}, T_{j_2},..., T_{j_h}]$, $j = 1,..., N$, была поставлена в соответствие матрица эталонной классификации предложений

$$C_j^e = \left(c_{jki}^e\right)$$

и матрица автоматической классификации

$$C_j^a = \left(c_{jki}^a\right),$$

где $c_{jki}^a, c_{jki}^e \in \{0,1\}$ — признак принадлежности предложения $i = 1,..., n_j$ к рубрике ω_r, $r = 1,..., k$, k — число классов. Необходимо отметить, что при формировании данной матрицы предполагалось, что все предложения из текстов T_{j_l}, $l = 1,..., h$, относятся к одному классу.

Для оценки качества выделения фрагментов использовались следующие показатели:

$$P_{jr} = \sum_{i=1}^{n_j} c_{jki}^e c_{jki}^a \Big/ \sum_{i=1}^{n_j} c_{jki}^a$$

и $R_{jr} = \sum_{i=1}^{n_j} c_{jki}^e c_{jki}^a \Big/ \sum_{i=1}^{n_j} c_{jki}^e$ — точность и полнота классификации предложений в тексте $j = 1,..., N$ для класса $r = 1,..., k$;

$$P = \frac{1}{Nr}\sum_{r=1}^{k}\sum_{j=1}^{N} P_{jr} \text{ и } R = \frac{1}{Nr}\sum_{r=1}^{k}\sum_{j=1}^{N} R_{jr} —$$

средние значения точности и полноты классификации предложений по всем классам.

В следующей таблице приводятся результаты экспериментов по оценке качества выделении фрагментов с использованием двух методов классификации. При проведении эксперимента использовались $h = 5$ и $N = 100$.

Таблица 2. Качество выделения фрагментов в текстах

Метод	SVM (P, R, F)	VMF (P, R, F)
NONE	20%, 23%, 22%	15%, 46%, 22%
HIER	**62%**, **72%**, **66%**	22%, **72%**, 33%
HMM	46%, 4%, 7%	3%, 3%, 3%
TILE	52%, 57%, 54%	29%, **72%**, **41%**
DIV	51%, 59%, 54%	33%, 48%, 37%

Таким образом, можно сделать следующий вывод, что качество выделения фрагментов оказывается на достаточно высоком уровне, учитывая недостатки, которые присущи массиву текстов РОМИП. При этом наилучшие показатели качества выделения фрагментов достигаются при использовании метода на основе иерархического покрытия. При этом использование алгоритма SVM оказывается предпочтительнее.

12. Заключение

Таким образом, в данной работе рассмотрены подходы к разметке текстов по результатам автоматической классификации, основанные на построении специального иерархического покрытия документов фрагментами, использовании скрытых марковских моделей и сегментации текстов. Проведенные эксперименты показали, что за счет выделения фрагментов в текстах можно повысить полноту классификации.

Актуальной задачей для дальнейших исследований является проведение более подробных исследований по оценке качества определения границ фрагментов в тексте, а также проведение с другими алгоритмами сегментации текста на фрагменты.

Работа выполнена при поддержке гранта Президента Российской Федерации для государственной поддержки молодых российских ученых МК-12.2008.10.

Литература

1. *Blei D., Moreno P. J.* Topic Segmentation with an Aspect Hidden Markov Model // SIGIR'01, September 9–12, 2001, New Orleans, Louisiana, USA. — 6 p.
2. *Chual T., Chang S., Chaisorn L., Hsu W.* Story Boundary Detection in Large Broadcast News Video Archives — Techniques, Experience and Trends // MM'04, October 10–16, 2004, New York, USA. — pp. 656–659.
3. *Choi F., Wiemer-Hasting P., Moore J.* Latent semantic Analysis for Text Segmentation // Proceedings of NAACL'01, Pittsburgh, PA, 2001. — pp. 109–117.
4. *Fragkou P., Petridis V., Kehagias Ath.* Linear Text Segmentation using a Dynamic Programming Algorithm, 2003. — 8 p.
5. *Greiff W., Morgan A., Fish R., Richards M., Kundu A.* Fine-grained hidden markov modeling for broadcast-news story segmentation // Proceedings of the First international Conference on Human Language Technology Research (San Diego, March 18 — 21, 2001). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 2001. — pp 1–5.
6. *Rabiner L.* A tutorial on hidden markov models and selected applications in speech recognition // Proc. IEEE, 77 (2), 1989. — pp. 257–286.

Словарь RuSLED как инструмент семантических исследований

RuSLED dictionary as tool for semantic study

Воскресенский А. Л. (avosj@yandex.ru), **Гуленко И. Е.** (gig@yandex.ru), **Хахалин Г. К.** (gkhakhalin@yandex.ru)

Специальная (коррекционная) общеобразовательная школа-интернат № 101 I и II вида

Описывается использование словаря русского жестового языка в качестве индикатора различных значений слов русского языка, что позволяет более целенаправленно вести анализ контекста для разрешения многозначности.

Введение

Одним из необходимых свойств системы искусственного интеллекта должна быть способность поддерживать коммуникацию с другими интеллектуальными системами и людьми, для чего необходимо понимать воспринимаемые речь и тексты. Понимание текста необходимо также для осуществления обучения и самообучения системы.

Создаваемая система автоматизированного сурдоперевода [1], поскольку между жестами жестового языка глухих и словами речи слышащих во многих случаях нет однозначного соответствия, также должна обладать способностью понимать вводимый текст (который может быть результатом работы подсистемы распознавания речи), чтобы на его основе формировать адекватные жестовые высказывания. Но в случае перевода со словесного на жестовый язык проблемы разрешения многозначности отличаются от подобных задач при переводе с одного словесного языка на другой. Некоторые понятия, однозначно воспринимаемые в словесном языке, в жестовом языке приобретают несколько значений, которые должны быть выделены и разделены для генерации правильного перевода.

Задача выбора семантического значения слова осложняется субъективным восприятием исследователя, на что мы обращали внимание ранее [2]. Как показано ниже, сопоставление слов русского языка и соответствующих жестов языка глухих России позволяет более объективно подойти к задаче разрешения многозначности и использования контекста для ее решения.

1. Краткое описание словаря

Словарь русского жестового языка RuSLED (Russian Sign Language Explanatory Dictionary) включает в себя функции толкового словаря, как для введенного слова, так и для его жестового представления. На вход словаря подается произвольная форма слова, а на выходе демонстрируются варианты жестового толкования данного слова [3]. Основные особенности словаря были описаны ранее [3], здесь приводятся данные, отличающие текущую версию.

Словарь содержит 2372 слова (с толкованиями их значений) и 2537 видеоизображений жестов (включая различные варианты исполнения), передающих значения этих слов. Для 1592 жестов (63 % от общего числа, вошедших в словарь) даны дополнительные пояснения, относящиеся к манере исполнения жеста или описывающие смысловые нюансы, передаваемые жестом. В словаре представлены жесты, используемые в Санкт-Петербурге и его окрестностях, что дало повод назвать данный словарь «Петербургский диалект». Видеоряд словаря составлен на основе видеокурса, изданного Межрегиональным центром реабилитации (МЦР), г. Павловск [4] (рис. 1).

Поставленная ранее цель: использование для демонстрации жестов виртуального персонажа (аватара) пока не достигнута из-за сложности представления мимики, сопровождающей жесты и выполняющей весьма важную роль в жестовом языке глухих. Так, например, слова *милый, симпатичный* передаются одним жестом, но отличаются движениями губ, проговаривающих фрагменты соответствующих слов.

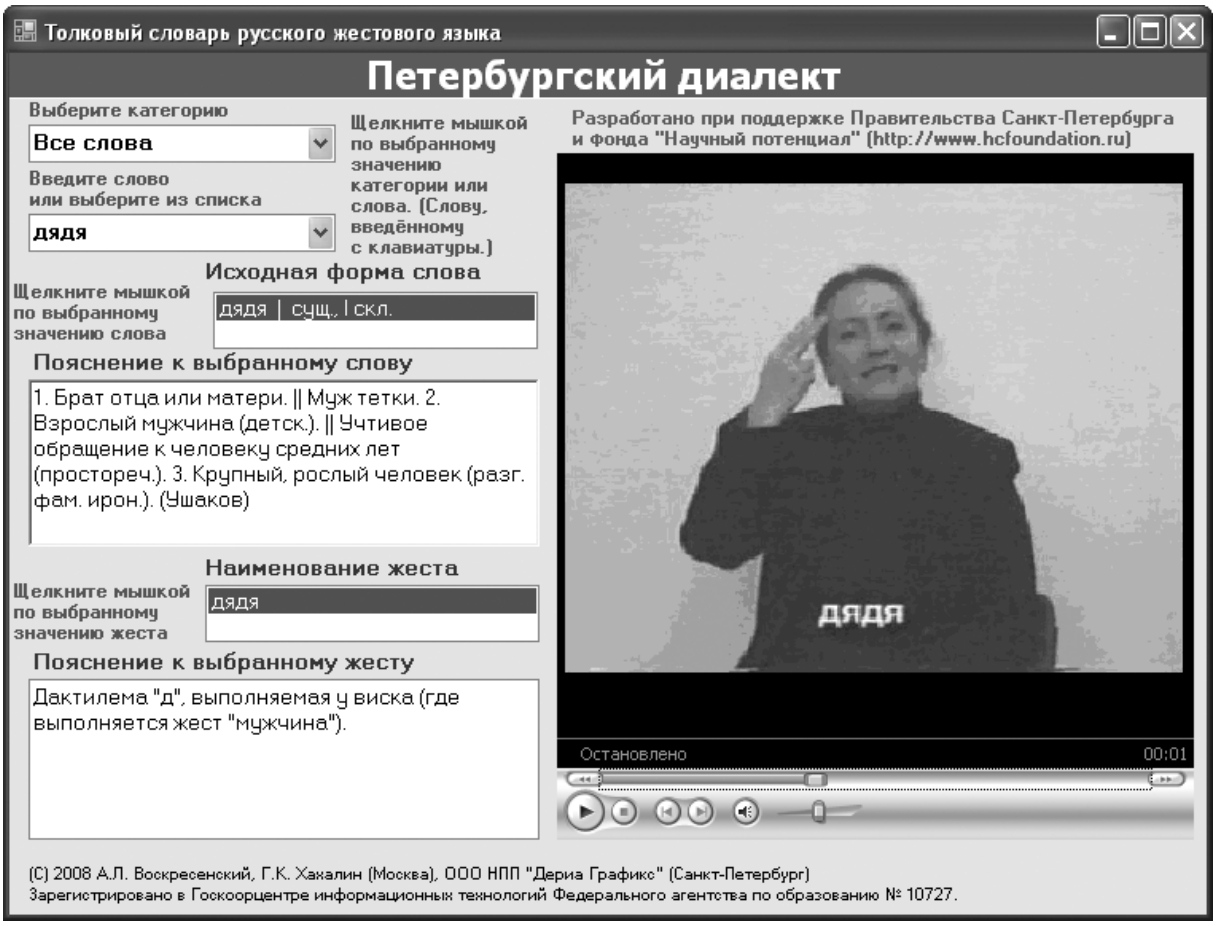


Рис. 1. Экранная форма словаря RuSLED

При составлении пояснений к некоторым жестам использовались пояснения из словаря «Говорящие руки» Фрадкиной [5], составленном на основе московского варианта жестового языка.

При составлении пояснений к словам использованы более 30 словарей и энциклопедий, доступ к которым осуществлялся через Интернет, используя в основном службу «Словари» портала Яндекс, за исключением нескольких словарей, в частности, одной из версий Толкового словаря русского языка Ушакова, размещенной на портале ГРАМОТА.РУ.

По рекомендациям сурдопедагогов обеспечена возможность фильтрации словника словаря по грамматическим категориям (существительные, глаголы, прилагательные, наречия, предлоги, частицы, числительные, местоимения). Это отличает словарь от известных словарей жестового языка: там деление материала осуществляется по темам, а не по грамматическим категориям слов, значения которых передают жесты. Именно поэтому данный словарь мы считаем не имеющим аналогов «мостиком» между словесным и жестовым русским языком, облегчая глухим пользователям понимание смысловых различий между словами русского языка.

Для просмотра всего содержимого словаря нужно выбрать категорию «Все слова».

Программная оболочка словаря зарегистрирована Госкоорцентром информационных технологий (ОФАП Минообразования и науки РФ) № 10727 от 30.05.2008. Дистрибутив словаря на DVD выполнен и распространяется ООО НПП «Дериа графикс» (г. Санкт-Петербург).

2. Слова и их семантические значения

При разработке интеллектуальных систем для обработки текста во многих случаях явно или косвенно подразумевается, что слова, из которых состоят обрабатываемые тексты, имеют постоянное семантическое значение. При этом, используя процедуры снятия омонимии, можно определить действительные значения слов и, соответственно, фраз и текста в целом. Соответственно, опираясь на фиксированные семантические значения ключевых слов (семантических примитивов), делаются попытки создать (или выявить) универсальный семантический код [6], оперируя которым можно определить значение любого высказывания.

Однако еще в 1940 г. академик Л. В. Щерба показал [7], что используемые в процессе речевого

общения понятия не имеют строго определенного значения. Аналогично, строя вероятностную модель языка, В. В. Налимов полагает, что «слова, на которых основана наша культура, не имеют и не могут иметь конкретных значений. Возможно, и даже необходимо, рассматривать слова как обозначения семантических полей с нечеткими границами, по которым строится функция распределения вероятностей…» [8].

Для случая перевода с одного словесного языка на другой разрешение многозначности переносных значений слов, например в случае синекдохи, не столь существенно, т. к. синекдоха разрешается слушателем подсознательно. В случае же перевода на жестовый язык необходимо осуществить явную подстановку требуемого значения слова, т. к. мышление глухих более конкретно, обобщение понятий у них происходит с большим трудом, чем у слышащих, что выражается набором жестов, образующих жестовой язык. Это приводит к большей сложности процедур автоматизированного перевода со словесного языка на жестовый, чем в случае перевода для словесных языков.

К сожалению, на настоящий момент нет ни одной автоматической системы, понимающей введенный текст. Наши подходы к решению данной проблемы требуют для своего описания обширного изложения, превышающего лимиты, допускаемые условиями публикации материалов Диалога. Кратко некоторые положения, лежащие в основе нашего подхода, описаны ниже.

Известные методы снятия многозначности, опирающиеся на частотное сопоставление контекстов, окружающих анализируемые слова [9, 10], дают полезные результаты в случаях омографии и омонимии, но во многих случаях не могут разрешить многозначность переносных значений слов, например в случаях синекдохи, метафоры. В этом случае необходимо учитывать не только значения отдельных слов текста, но и текста в целом. Как показано в [11, 12], система понимания текста, помимо средств, обеспечивающих лингвистический разбор текста, должна включать в себя блок, отслеживающий изменения характеристик описываемых в тексте объектов (пространственное положение, размеры, облик, возраст и т.п.) и хранящий значения этих характеристик в привязке к времени текста и к астрономическому времени. Подтверждение этих положений дают и примеры многозначности слов, приведенные в разделе 3, требующие при сурдопереводе выбора жестов, соответствующих конкретному значению понятия, передаваемому словом в текущем контексте.

Сопоставление в заданные моменты времени текста текущих значений характеристик для различных объектов позволяет создавать описания текущих ситуаций, причем с учетом объектов, не упоминаемых в обрабатываемом предложении текста. Это отличает данный подход от предлагаемого в [13], где описание ситуации создается на основе текущего предложения.

3. Примеры некоторых случаев многозначности при сурдопереводе

При разрешении случаев многозначности, вызванных омографией словоформ или омонимией различных понятий, эффективны частотные методы обработки контекста, опирающиеся дополнительно на знания соотношений слов, описываемых тезаурусом [10].

Однако, как указывалось ранее, для некоторых случаев полисемии, существенных при сурдопереводе, они не применимы. Покажем это на некоторых примерах.

Слово *земля* в русском языке имеет ряд значений, из которых в словаре RuSLED встречаются значения *планета*, *почва*, *берег*. Рассмотрим последний случай.

Для жеста, передающего значение *берег*, в словаре [5] приводится пояснение: *«"Земля!" — закричали матросы».* Различные программы-переводчики, доступные в Интернете, дают следующие варианты перевода (примеры 1, 2, 3):

(1) *«Ground!» — sailors cried.* (Cognitive Translator, http://cs.isa.ru:10000/ct/)

(2) *«The Earth!» — sailors have cried*. (PROMT© Translator, http://www.translate.ru/)

(3) *«Land!» — cried the sailors*. (Переводчик Google©, http://translate.google.com/)

Общаясь с помощью словесной речи, мы каждый раз решаем задачу распознавания информации, передаваемой нам собеседником. При этом происходит подстановка понятий, хранящихся в нашей памяти, т. е. воспринятый смысл текста не является точным аналогом слов, составляющих фразы текста. Там, где это возможно, воспринятое содержание фразы внутренне дополняется (и корректируется) в соответствии с общим содержанием текста и имеющимися знаниями об окружающем мире, не вызывая проявляемых внешне затруднений и протеста. Поэтому варианты (1) и (3) могут быть признаны допустимыми для случая словесного языка, а вариант (2) — нет, поскольку *«The Earth»* означает планету Земля, которую матросы не могут увидеть как цельный объект ни при каких обстоятельствах.

Но отметим, что ни в одном из случаев не получено значение *coast (берег)*, необходимое для задачи сурдоперевода.

Поясним ход рассуждений, приводящих к распознаванию синекдохи, когда обобщающим словом *земля* обозначается и часть земли, граничащая с водой (*берег*): матросы находятся на корабле, находящемся в открытом море → корабль со всех сторон окружен водой → граница воды и суши (земли) называется берег → если матросы закричали «Земля!», это означает, что они увидели границу между водой и сушей (землей), т. е. *берег*.

Представленные рассуждения соответствуют традиционной системе логических умозаключений, известной со времен Аристотеля. Из известных прототипов систем искусственного интеллекта, использующих подобную логику, можно назвать, например, системы NARS, Novamente [14].

При сурдопереводе возникают и ситуации многозначности, вызванные особенностями общения на жестовом языке. Так, например, для понимания текста возникает задача разрешения референциальных ссылок личных местоимений *он, она*. При сурдопереводе эта задача осложняется тем, что в случае присутствия в месте осуществления диалога объектов, на которые ссылаются эти местоимения, они передаются жестом *это* (при этом жест указывает на соответствующий объект, т. е. разрешение дейктической референции требует знания о пространственном положении объекта [15]), в случае же их отсутствия (при анафорической референции) используются жесты *он* или *она* (к сожалению, эти жесты выпали из рассмотрения в [15]).

Таким образом, система сурдоперевода должна решать задачу выбора и подстановки нужных жестов, исходя из общего содержания переводимого текста, представленного в виде данных в привязке ко времени текста взаимодействий основных и второстепенных действующих лиц, а также объектов фона [12].

Эти значения не всегда, как показывают приведенные примеры, будут очевидными, поэтому такая задача с полным основанием может считаться интеллектуальной.

При разрешении метафоры необходимо опираться на сведения о возможных действиях объектов, хранимые в онтологии системы. Так, например, фраза *Машина летела по дороге* должна передавать понятие *быстро ехать*, тогда как фраза *Душа летела к ней* должна передавать понятие *стремиться*.

Выявление метафоры может осуществляться при задании в онтологии для концептов, описывающих сущности, допустимых для этих концептов действий (атрибут «делает») и соответствующил правил обработки контекста. Если действие объекта не совпадает с возможными, заданными в его онтологическом описании, это может свидетельствовать о метафорическом обороте, требующем специальной обработки для выявления значения, передаваемого в жестовом высказывании.

Заключение

Разработка толкового словаря русского жестового языка RuSLED является первым шагом в создании автоматизированной системы сурдоперевода. Сопоставление двух ресурсов, объединенных в этом словаре: слов русского языка и жестов русского жестового языка позволило обратить внимание на некоторые особенности передачи семантики высказываний с помощью словесного и жестового языков, которые не всегда очевидны.

Наличие этих особенностей требует, чтобы система сурдоперевода не просто сопоставляла слово исходной фразы соответствующему жесту, а понимала обрабатываемый текст для его преобразования в последовательность понятий, передаваемых с помощью жестовых высказываний.

Возможно, разрабатываемые для нужд системы сурдоперевода средства семантической обработки текста окажутся полезными и для задач словесных языков, в частности, для задач машинного перевода.

В настоящее время работа по созданию автоматизированной системы сурдоперевода находится в начальной стадии. Обращение к имеющимся в общем доступе ресурсам могло бы содействовать ее успешной реализации, но, к сожалению, семантическая разметка внешних источников информации далеко не всегда отвечает требованиям решаемой задачи. В [10] семантическая разметка внешних источников информации также не используется, возможно, по этим же причинам. Например, как видно из рис. 2, 3, семантическая разметка в Национальном корпусе русского языка (www.ruscorpora.ru) не позволяет принять решение о конкретном значении слова в текущем контексте (хотя и имеется пометка, что омонимия снята).

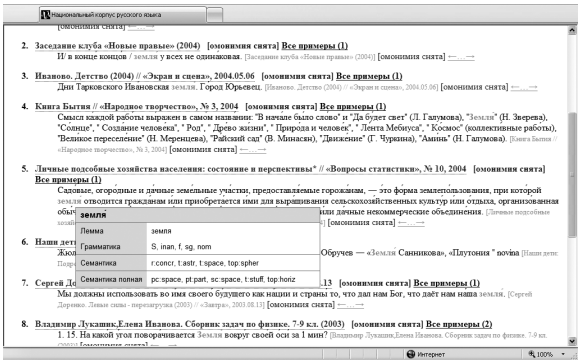


Рис. 2. Слово «земля» в значении «земельный участок»

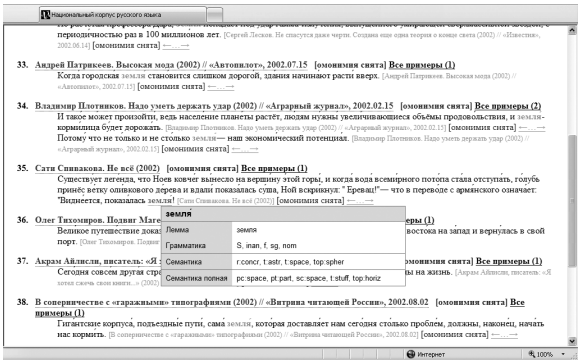


Рис. 3. Слово «земля» в значении «берег, край суши»

Литература

1. *Voskressenski A.* Signs and speech: two forms of human communication // Proceedings of the Ninth International Conference «Speech and Computer» SPECOM'2004. Saint-Petersburg, Russia, 2004, P. 666–669.
2. *Воскресенский А. Л., Хахалин Г. К.* Средства семантического поиска. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая — 4 июня 2006 г.). — М.: Изд-во РГГУ, 2006. — С. 100–104.
3. *Воскресенский А. Л.* Сопоставительное лексикографическое описание слов русского языка и жестов языка глухих России в словаре RuSLED // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). — М.: РГГУ, 2008. С. 91–96.
4. *Специфические* средства общения глухих: Видеокурс: В 3 частях // СПб — Павловск: МЦР, 2002.
5. *Фрадкина Р. Н.* Говорящие руки: Тематический словарь жестового языка глухих России // М.: Изд-во «Сопричастность» ВОИ, 2001. — 598 С.
6. *Мартынов В. В.* Основы семантического кодирования. Опыт представления и преобразования знаний. // Мн.: ЕГУ, 2001. — 140 С.
7. *Щерба Л. В.* Опыт общей теории лексикографии. // «Изв. АН СССР, ОЛЯ», № 3, 1940, С. 100.
8. *Vasily Nalimov.* Realms of the Unconscious; The Enchanted Frontier. // ISI Press, 1982.
9. *Жигалов В. А., Жигалов Д. В., Жуков А. А., Кононенко И. С., Соколова Е. Г., Толдова С. Ю.* Система ALEX как средство для многоцелевой автоматизированной обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002. Т. 2: Прикладные проблемы. — М.: Наука, 2002. — С. 192–208.
10. *Лукашевич Н. В., Чуйко Д. С.* Автоматическое разрешение лексической многозначности на базе тезаурусных знаний. // Интернет-математика 2007: Сборник работ участников конкурса. — Екатеринбург: Изд-во Урал. ун-та, 2007. — С. 108–117.
11. *Воскресенский А. Л., Хахалин Г. К.* О структуре системы, понимающей текст. // Вторая Международная конференция «Системный анализ и информационные технологии» САИТ-2007 (10–14 сентября 2007 г., Обнинск, Россия): Труды конференции. В 2 т. — М.: Издательство ЛКИ, 2007. — Т. 1. — С. 111–114.
12. *Voskresenskij A.* Text Disambiguation by Educable AI System // The First Conference on Artificial General Intelligence / P. Wang et al. (Eds.), AGI-08, 1–3 March, 2008, Memphis. IOS Press, 2008.
13. *Леонтьева Н. Н.* Автоматическое понимание текстов. Системы, модели, ресурсы. — М.: Издательский центр «Академия», 2006. — 304 с.
14. *Artificial* General Intelligence / B. Goertzel, C. Pennachin (eds). — Springer, 2007.
15. *Кибрик А. А., Прозорова Е. В.* Референциальный выбор в русском жестовом языке. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.) / Под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. — М.: Изд-во РГГУ, 2007. — С. 220–230.

Электронный русский ассоциативный словарь школьников

The digital russian associative dictionary of schoolchildren

Гольдин В. Е. (goldinve@yandex.ru),
Мартьянов А. О. (comrad-mao@mail.ru),
Сдобнова А. П. (sdobnovaap@yandex.ru)

Саратовский государственный университет имени Н. Г. Чернышевского

В докладе обсуждаются возможности решения психолингвистических, социолингвистических и культурологических проблем с опорой на материалы электронного «Ассоциативного словаря школьников Саратова и Саратовской области».

Введение

Создание крупных ассоциативных словарей, включающих прямые статьи, обратные и ряд обобщающих приложений, опирается, как правило, на использование электронных баз данных (см., например: РАС, САС). Поскольку ассоциативные словари являются лингвистическими источниками сугубо научного характера, то их электронная форма должна обеспечивать возможность реализации запросов, соответствующих современному пониманию сущности вербальных ассоциаций, актуальным исследовательским задачам, часто используемым приемам анализа ассоциативных данных. При этом на базе о т д е л ь н ы х (хотя бы даже достаточно крупных) ассоциативных словарей существенный прогресс в изучении ассоциативно-вербальных сетей, по-видимому, невозможен, так как для обоснованных выводов психолингвистического, социолингвистического, лингвокультурологического характера необходимо сопоставление результатов, полученных в разное время и в различных условиях, полученных в экспериментах с представителями различных социальных и культурных групп, полученных с применением различных исследовательских процедур и т.д., — необходима, другими словами, более или менее сложившаяся а с с о ц и а т и в н а я л е к с и к о г р а ф и я я з ы к а, в рамках которой накапливаются не только материалы, но и некоторые становящиеся стандартными формы их сбора, анализа, словарного воплощения. Задача стандартных форм — поддерживать содержательную сопоставимость данных.

Русская ассоциативная лексикография, начало которой относится к 70-м гг. XX века, успешно развивается. Реализация таких проектов, как РАС и САС, публикации других ассоциативных словарей русского языка, учет известных зарубежных достижений в области ассоциативной лексикографии (например, Эдинбургского ассоциативного тезауруса) уже позволяют формулировать основные требования к функциональной стороне электронных ассоциативных словарей исследовательского характера, однако ими пока еще не обеспечена выработка достаточно полной системы таких форм и процедур, поэтому важно учитывать и подвергать осмыслению все имеющиеся сегодня практические решения в данной сфере. Ниже рассматривается одно из них: «Ассоциативный словарь школьников Саратова и Саратовской области» (АСШС). Статья посвящена структуре содержания АСШС и тем из главных итогов изучения материала АСШС, которые представляются значимыми для русской ассоциативной лексикографии в целом.

1. Состав и функциональная специфика словаря

«Ассоциативный словарь школьников Саратова и Саратовской области» создан на основе свободных ассоциативных экспериментов, проводившихся в устно-письменном формате (устное предъявление стимулов и письменная фиксация реакций испытуемыми) в течение 10 лет (с 1998 по 2008 г. включительно). Он реализован как Access-приложение. Стимульный ряд АСШС включает 1125 вербальных единиц. На конец декабря 2008 г. в БД АСШС сосре-

доточено более 900 000 пар «стимул-реакция» (S-R), ими представлено 62 500 различных реакций. Данные о количестве реакций, полученных от школьников различных возрастных групп, от школьников городских и сельских, от мальчиков и девочек, приведены в табл. 1 и табл. 2.

Анкетирование имело анонимный характер, однако параллельно со сбором материала для АССШС А.П. Сдобновой проводился лонгитюдный ассоциативный эксперимент с отдельными конкретными школьниками. Материала обоих экспериментов демонстрирует одни и те же основные закономерности возрастной динамики лексикона школьников.

Словарь поддерживает фильтрацию материала по возрасту, полу испытуемых, месту их проживания, типу учебного заведения (школа или гимназия, лицей), дате проведения эксперимента и некоторым другим признакам. АССШС генерирует в табличной форме прямые статьи на заданные стимулы и обратные статьи по заданным реакциям, сопровождая статьи необходимыми количественными данными; словарь предлагает стандартные перекрестные запросы, позволяет сопоставлять между собой множества реакций на выделенные пользователем группы стимулов, сопоставлять множества стимулов, вызывающих указанные пользователем группы реакций, получать общие перечни реакций и перечни реакций, ранжированные по убыванию так называемых «входящих» связей, что необходимо при исследовании «ядра языкового сознания».

В процессе перенесения реакций в БД АССШС орфографические ошибки, допущенные школьниками, исправлялись (например, ответ **смышлённый* фиксировался как *смышлёный*, ответ **ширенга* — как *шеренга*, ответ **Алладин* как *Аладдин* и т.д.), но представленные в ответах единицы с любыми особенностями морфологического, морфонологического или синтаксического характера, просторечные, жаргонные и иные нелитературные единицы, включая обсценные слова и выражения, многочисленные окказиональные образования, экспрессивные написания, слова других языков, записанные в соответствующей графике, иноязычные слова и выражения, переданные школьниками в русской транслитерации, неполные и контаминированные написания, различного рода небуквенные знаки (*ушибнуть, крикучий, залазиют, курей, задуманность, ерундый, многостранный, тырчик, мобилы, мозголомня, пофигень, туть-туть, щас!, ничё не делать, не знай где, днюха, upset, screen, трабл, не идти в школу:-))), кувыр.com* и под.) сохранялись в базе в аутентичном виде наряду с нормативной русской лексикой, нормативными грамматическими и графическими формами.

Существенной особенностью АССШС является то, что в нем сохраняется и доступен для анализа образ каждой анкеты, введенной в базу. Это дает возможность рассматривать индивидуальное использование типовых стратегий реагирования (стратегия оценивания, Я-стратегия, стратегия противоречия и др.), оценивать относительную значимость общих ассоциативных стратегий и ассоциативных потенций самих стимулов, а также возрастную динамику стратегий. Материал свидетельствует о том, что в анкетах школьников младшей возрастной группы действие частных стратегий реагирования проявляется сильнее, чем в анкетах школьников других возрастных групп; к старшим

Таблица 1. Распределение материала (пар S-R) в соответствии с полом и возрастом испытуемых

Группы испытуемых по полу	**Возрастные группы школьников**			
	1–4 классы	5–6 классы	7–8 классы	9–11 классы
Всего пар S-R в ответах мальчиков	134290	104128	93888	102428
Всего пар S-R в ответах девочек	135683	108369	98923	124980
ВСЕГО	269973	212497	192811	227408

Таблица 2. Распределение материала (пар S-R) в соответствии с местом жительства и возрастом испытуемых

Группы испытуемых по месту жительства	**Возрастные группы школьников**			
	1–4 классы	5–6 классы	7–8 классы	9–11 классы
Всего пар в ответах сельских школьников	45736	81387	74358	97438
Всего пар в ответах школьников районных центров	106043	55053	46123	33954
Всего пар в ответах учащихся школ областного центра	118194	76057	72330	96016
ВСЕГО	269973	212497	192811	227408

классам существенно возрастает влияние ассоциативных потенций каждого отдельного стимула на формирование ассоциативных реакций, и «длинные стратегии» реагирования практически перестают использоваться.

Анкеты с преобладанием асемантических реакций и реакций, не мотивированных стимулами (такие анкеты традиционно отбраковываются создателями ассоциативных словарей), сохраняются в базе АСШС, но со специальной меткой и доступны для анализа как в качестве отдельного подмножества, так и, при необходимости, — в составе общего корпуса данных.

2. Основные результаты и перспективы исследования материалов АСШС

1. Материалы АСШС сопоставимы с материалами, полученными в ассоциативных экспериментах со школьниками других регионов: с пермскими, омскими, курганскими, московскими (см.: Береснева и др. 1995; Овчинникова и др. 2000; Гуц 2004; Федченко 2006). Установлено, что местные особенности нашли отражение лишь в незначительной части ассоциативных данных АСШС, что выявляемые в экспериментах ассоциативно-вербальные сети соответственных групп школьников различных регионов в основном совпадают и что, таким образом, АСШС фактически не имеет регионального характера и репрезентирует общие особенности языкового сознания российских русскоязычных школьников конца XX — начала XXI века. Этим определяются широкие возможности использования данных АСШС как научного источника.
2. На материале АСШС для каждой из возрастных групп выявлены ядерные зоны языкового сознания, образуемые словами-реакциями с наибольшим количеством ассоциативных связей (Сдобнова $2007_а$). Ядро в целом определяется по сложившейся традиции в составе 75 единиц; они распределяются по смысловым зонам ‘Человек’, ‘Оценка’, ‘Деятельность’, ‘Жизнь’, ‘Природный и животный мир’ (Сдобнова $2007_б$).

Ядро лексикона школьников отличается высокой стабильностью: почти 60% его состава у школьников всех возрастных групп (с 1 по 11 классы) составляют следующие константные единицы: *большой*, *быстро*, *вода*, *делать*, *дело*, *день*, *деньги*, *дерево*, *дом*, *дорога*, *друг*, *думать*, *животное*, *жизнь*, *игра*, *идти*, *класс*, *книга*, *красивый*, *люди*, *маленький*, *мальчик*, *мама*, *машина*, *много*, *мой*, *нет*, *он*, *плохо*, *плохой*, *работа*, *радость*, *ребенок*, *собака*, *стол*, *ум*, *умный*, *урок*, *учитель*, *хороший*, *хорошо*, *человек*, *школа*, *я*. Часть из них (*большой, быстро, идти, книга, маленький, мальчик, хороший, хорошо, человек, я*) входят в ядро лексикона уже у дошкольников 3–6 лет (Соколова 1999).

3. В лексиконе школьников, в том числе и в его ядерной зоне, протекают достаточно сложные процессы, и материал АСШС позволяет исследовать лексикон школьников как активно формирующуюся систему. В ней представлены средства политической (*Евросоюз, СНГ, СССР, Единая Россия, Белый дом, террорист, террористический* и др.), административно-правовой (*президент, Госдума, парламент, Российская Федерация, премьер, МВД, МЧС, губернатор; конституция, контракт, правонарушение, Уголовный кодекс, указ, юридический* и др.), образовательной, сакральной и других современных сфер коммуникации, лексика наддиалектных социально-функциональных вариантов языка: книжно-литературная, литературно-разговорная, профессиональная, просторечная, жаргонная (см.: Сдобнова 2008).

В реакциях школьников находят отражение процессы и тенденции, характерные для современной русской речи:

- активность использования заимствований (богатый → *олигарх*, система → *хакер*, кабинет → *boss*, песня → *cool*, точность → *accuracy!*, кошмар → *nightmare*, катастрофа → *plane crash* и под.);
- усиление в коммуникации игровой, импровизационной манеры самовыражения и экспрессивности речи, поддерживаемое современной рекламой и расширяющимся интернет-общением, в том числе применение разнообразных небуквенных графических средств и сочетаний кириллических и латинских букв в написании слов (стыд → *но*, различный → *небез*, куча → *!!! Супер!!!*, голосование → *Ура!!!!!*, личность → ???; выставка → *а билет?*, имя → *Шеff;* теперь → *TEPERЬ,* идея → *иди в Икею:-)* и под.);
- жаргонизация современной речи (материалы АСШС содержат почти 1000 различных жаргонных единиц, а включающие их реакции составляют в БД АСШС около 2% всего множества реакций);
- укрепление в языковом сознании и речевой практике молодого поколения новых словообразовательных микросистем, моделей, например с корнями *супер-*, *фото-*, *шоу-* и другими, жаргонных микросистем с корнями *лох-*, *прикол-*, жаргонных образований способом усечения (*ботан, недоум, препод, скин, инет, комп, универ, калашник, мерс, неформал, экстремал, гомосек, эпилепс, синхрон, эксклюзив, экстрим, мобил, клик, напряг, оттяг, облом, отпад, прочёс, угар, в лом, в хлам, внапряг, невдогон, невозмога*).

Состав ассоциативных реакций, как свидетельствует материал АСШС, связан и с полом, и с возрастом школьников, и с местом их постоянного проживания, и со временем проведения экспериментов, и с форматом экспериментов (устно-письменным в отличие, например, от письменно-письменного формата), и даже с обычным или «элитарным» типом школ, в которых проводятся эксперименты. Однако более всего обнаруживаемые различия соотносимы с возрастными группами испытуемых, и АСШС дает богатый сопоставительный материал для изучения динамики ассоциативно-вербальных сетей школьников с 1 по 11 класс. В возрастной динамике ассоциаций проявляется, конечно, не возраст сам по себе, а соответствующие ему, связанные с ним социокультурные и психофизиологические особенности учащихся. В настоящее время АСШС — единственный из русских ассоциативных словарей, содержащий данные экспериментов со школьниками всех возрастных групп.

Обнаруживаются общие закономерности динамики ассоциативно-вербальной сети учащихся в школьные годы, и вместе с тем изменения каждого из ассоциативных полей обладают своей спецификой и не всегда подчиняются единой линии изменений. Так, по мере взросления учащихся доля отказов («нулевых» реакций) в корпусе их ответов постоянно уменьшается. Это общая тенденция. На рис. 1 представлены доли отказов в ответах мальчиков и девочек с 1 по 11 класс, и результат, полученный при делении материала по гендерному признаку, подтверждает существование сформулированной выше общей закономерности.

Однако при этом в составе ассоциативных полей отдельных стимулов доля отказов уже в ответах

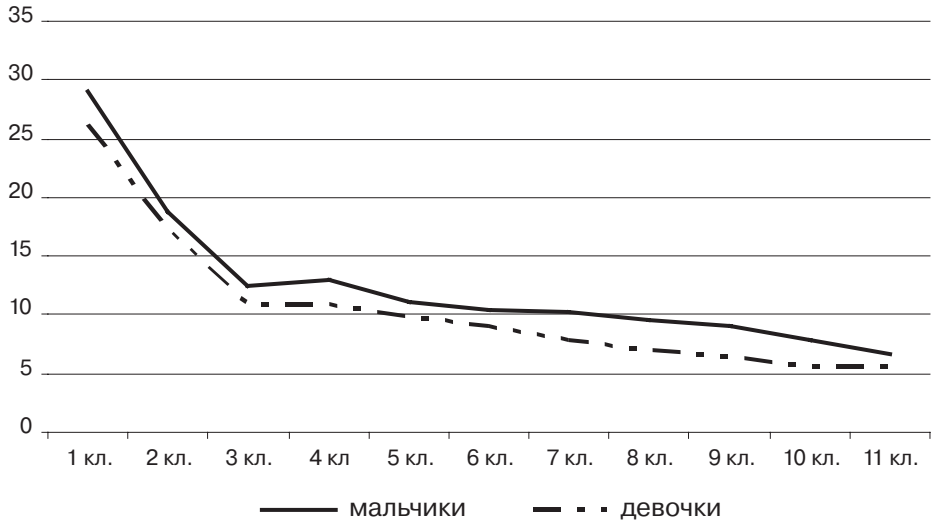


Рис. 1. Динамика «*нулевых» реакций*

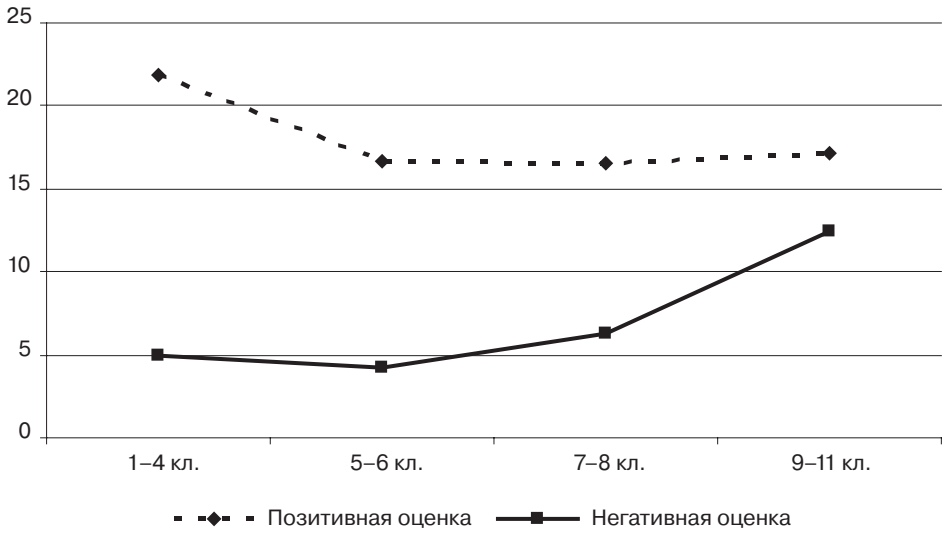


Рис. 2. Динамика оценочных реакций в ассоциативном поле ЖИЗНЬ

учащихся 1–4 классов не превышает 1–1,5% и остается на том же уровне в течение всех лет обучения в школе (стимулы берёза, поле, лиса, стакан); в ассоциативных полях других стимулов (удивление, изящный, зависеть) доля отказов колеблется на уровне 20–30% у школьников всех возрастных групп без значительного снижения в ответах старшеклассников; ассоциативные поля таких стимулов, как скупой, скорбь, корысть, наивность, гравий, передовой, демонстрируют значительное уменьшение доли отказов, но в ответах на одни стимулы (корысть, скорбь, наивность) резкое уменьшение доли отказов фиксируется в 7–8 классах, а в ответах на другие стимулы (передовой, скупой) — только в самой старшей возрастной группе, у учащихся 9–11 классов. За этими и подобными различиями стоят многие факторы собственно лингвистического и психологического характера (лексическая и грамматическая специфика слова-стимула, степень его речевой освоенности испытуемыми, парадигматические, синтагматические, фразеологические связи слова-стимула, гендерные особенности развития лексикона и др.). Учесть результаты их совокупного действия можно, выделяя характерные типы возрастной динамики ассоциативных полей.

Материал АСШС позволяет выделить четыре наиболее противопоставленных типа возрастной динамики ассоциативных полей. Это, во-первых, «тип стандартизации», характеризуемый увеличением совокупной доли «главных ассоциатов» (к «главным ассоциатам» относятся вербально выраженные реакции, составляющие не менее 5% всех ответов на данный стимул) при относительной стабильности содержательной структуры поля, во-вторых, — «тип усложнения поля», для которого характерны возрастные изменения противоположной направленности, связанные с усложнением представлений испытуемых о соответствующих фрагментах мира, в-третьих, — «тип вхождения единицы в лексикон» и «тип периферийного развития», отражающие развитие языковой компетенции школьников (см.: Гольдин, Сдобнова 2007). Основные типы реализуются множеством частных разновидностей и переходных случаев, поразному представленных в ядерной и периферийной зонах внутреннего лексикона школьников.

Наиболее характерный тип возрастной динамики ассоциативных полей ядерной зоны лексикона школьников — тип стандартизации. Так, динамический ряд совокупных долей главных ассоциатов в ассоциативных полях стимула ЖИЗНЬ имеет следующий вид: 6,41% (1–4 кл.) — 13,53% (5–6 кл.) — 21,34% (7–8 кл.) — 22,51% (9–11 кл.). При последовательном росте совокупной доли главных ассоциатов основная структура ассоциативного поля ЖИЗНЬ сохраняется в реакциях школьников всех четырех возрастных групп: большую часть ассоциатов составляют оценочные реакции (*хороша, красота, радость, трудная* и др.), значительное место в структуре поля принадлежит парадигматической реакции *смерть*, представлены группы реакций, отмечающие течение и длительность жизни, ее наполнение (*любовь, дружба, учёба* и др.), реакции, выделяющие человека как субъекта жизни. Качественные изменения в поле ЖИЗНЬ также имеют место: от младших групп к старшим несколько уменьшается доля позитивных оценочных реакций и соответственно растет доля реакций негативного характера, но преобладание оценок позитивного характера сохраняется и в ответах старшеклассников (см. рис. 2).

Существуют различные понимания стандартизации ассоциативных полей. Если некоторую структуру поля, например представленную реакциями образованных взрослых испытуемых, принять за норму, то стандартизация определяется как мера приближения к данной норме. В нашем исследовании стандартизация рассматривается поиному: как степень общности ассоциативных связей конкретных стимулов, обеспечивающая взаимопонимание в однородных группах испытуемых. Материалы АСШС свидетельствуют о том, что некоторое усиление стандартизации ассоциативных полей у испытуемых-старшеклассников имеет место, но это не глобальный и не однонаправленный процесс, требующий дальнейшего изучения, в том числе и на данных АСШС.

Заключение

«Ассоциативный словарь школьников Саратова и Саратовской области» заполняет существовавший до его создания пробел в совокупном массиве данных русской ассоциативной лексикографии о динамике ассоциативно-вербальной сети в школьные годы и демонстрирует необходимость использования в электронных ассоциативных словарях ряда дополнительных функций, в том числе доступа к образу каждой анкеты и сортировки реакций по количеству входящих связей.

Литература

1. *Береснева Н. И., Дубровская Л. А., Овчинникова И. Г.* Ассоциации детей от шести до десяти лет. Пермь, 1995.
2. *Гольдин В. Е., Сдобнова А. П.* «Дом» в ассоциациях школьников (динамический аспект) // Вестник Пермского университета. Сер. Филология. 2006. Вып. 3(3). — С. 43–50.
3. *Гольдин В. Е., Сдобнова А. П.* Динамика языкового сознания современной молодежи по данным «Ассоциативного словаря школьников Саратова и Саратовской области» // Известия Саратовского университета. Новая серия. 2007. Т. 7. Серия Филология. Журналистика. Вып. 1. — С. 24–30.
4. *Гольдин В. Е., Сдобнова А. П.* Русская ассоциативная лексикография. Саратов, 2008.
5. *Гуц Е. Н. Ассоциативный словарь подростка. Омск, 2004.*
6. *Овчинникова И. Г., Береснева Н. И., Дубровская Л. А., Пенягина Е. Б.* Лексикон младшего школьника (характеристика лексического компонента языковой компетенции). Пермь, 2000.
7. *РАС* — Русский ассоциативный словарь. В 2 т. / Ю. Н. Караулов, Г. А. Черкасова, Н. В. Уфимцева, Ю. А. Сорокин, Е. Ф. Тарасов. М., 2002.
8. *САС* — Славянский ассоциативный словарь: русский, белорусский, болгарский, украинский / Н. В. Уфимцева, Г. А. Черкасова, Ю. Н. Караулов, Е. Ф. Тарасов. М., 2004.
9. *Сдобнова А. П.* Единицы ядра языкового сознания современного школьника // Вопросы психолингвистики. — $2007_{а}$. — № 5. — С. 99–104.
10. *Сдобнова А. П.* Лексикон современного российского школьника // Образ России: извне и изнутри: Сб. стат. / Под ред. Е. Ф. Тарасова, Н. В. Уфимцевой, Е. А. Аршавской / Институт языкознания РАН. М.; Калуга, 2008. — С. 207–218.
11. *Сдобнова А. П.* Ядро языкового сознания школьников: функционально-семантическая структура // Языковое сознание: парадигмы исследования. М.; Калуга, $2007_{б}$. — С. 169–192.
12. *Соколова Т. В.* Ассоциативный тезаурус ребенка 3–6 лет. Автореф. дис. … д-ра филол. наук. М., 1999.
13. *Федченко А. В.* Языковое сознание русских и американских подростков. М., 2006.

О характере синтаксической полисемии

On the nature of syntactic polysemy

Григорьян Е. Л. (elena_grigorian@yahoo.co.uk)

Южный федеральный университет, Ростов-на-Дону

На материале некоторых диатезных преобразований отмечается, что большинство структур могут использоваться для выражения различных смыслов и что диатезы, противопоставляемые в одних контекстах, могут не нести никаких семантических различий в других и различаться на уровне прагматики.

Сопоставление употребления конструкций, реализующих различные диатезы одного глагола и описывающих идентичные или сходные денотативные ситуации, позволяет сделать ряд наблюдений, проливающих свет на особенности собственно синтаксической семантики. Во-первых, большинство структур могут использоваться для выражения различных смыслов, и с каждой из них связан ряд семантических признаков, не обязательно присутствующих в каждом конкретном употреблении. Любой из них может быть достаточным основанием для выбора соответствующей конструкции. Во-вторых, нередко структуры, соотносящиеся как диатезные преобразования, оказываются равнозначными и могут быть взаимозаменимы в одном контексте без всякого ущерба для содержания.

Иначе говоря, при противопоставлении исследуемых структур мы сталкиваемся, с одной стороны, с явным контрастом, подчёркиванием дифференциального признака, а с другой стороны, встречаются употребления, в которых соответствующий признак отсутствует.

Далее это будет показано на примере (1) пассивных конструкций в противопоставлении некоторым другим диатезам; (2) безличных конструкций в сопоставлении с личными; (3) производных диатез действительного залога, в которых подлежащим является не агенс, а средство или актанты сходной ролевой семантики, в сопоставлении с агентивной конструкцией. Для соответствия регламенту доклада мы вынуждены в разделе 1.3 ограничиться английским языковым материалом, а в разделе 3 — русским, хотя упомянутые языки в исследуемых аспектах полностью аналогичны.

1.1. В английском языке пассивные конструкции могут быть в некоторых отношениях сближены с конструкциями с симметричными глаголами (т.е. глаголами, которые не меняют своей формы, если подлежащим является объект действия) в непереходном употреблении. Последние соотносятся с переходными конструкциями тех же глаголов так же, как пассив соотносится с активом (*The boy broke the window — The window broke*), и второй случай (иногда называемый в грамматиках «немаркированный пассив») в некоторых контекстах заменим обычным пассивом, т.е. возможна нейтрализация различий; вместе с тем во многих контекстах значения этих форм расходятся и могут даже противопоставляться. Так, в примере (1), где описываются действия таможенников, и в контрпримере неравнозначность очевидна:

(1) *At Kermanshah the machine came down for passport examinations and customs. A bag of Mr. Parker Pyne's* ***was opened****. A certain small cardboard box was scrutinized with some excitement. Questions were asked…* (A. Christie).

Ср. *A bag opened,* что описывало бы спонтанное происшествие.

(2) *Submarines in the Channel. A merchant ship* ***had been sunk*** *that morning* (Aldington).

Ср. *a merchant ship had sunk,* что совместимо с описываемой ситуацией, но не предполагает внешнего воздействия и может описывать и такое событие, когда судно никто не топил.

(3) *Scotland has large monuments to two women martyrs drowned for their faith, in spite of the fact that they* ***weren't drowned*** *at all and neither was a martyr anyway* (J. Tey).

В примерах (1) — (3) наглядно выступает дифференциальный признак **«внешнее воздействие vs самопроизвольность (автономность)».**

В то же время мы сталкиваемся с полной равнозначностью тех же единиц и конструкций в других контекстах:

(4) *He* ***has*** *just fallen into the well and* ***been drowned*** (G.B. Shaw).

(5) *The boat capsized and over 20 passengers* ***were drowned*** (Пример из [Huddleston, Pullum 2002: 1466]).

Характерно, что в подобных примерах практически без изменения смысла форма пассива может быть заменена личной формой соответствующего симметричного глагола (*over 20 passengers* ***drowned***), при этом они не могут быть преобразованы в актив и, разумеется, в них невозможно агентивное дополнение. Ср. пример (3) с собственным значением пассива (и с тем же глаголом).

1.2. Кроме того, пассив нередко противопоставляется рефлексивам, прежде всего собственно-возвратному значению. Характерен следующий пример из русского языка, где пассив и рефлексив явно противопоставлены:

(6) *Федор. Князь Шуйский удавился? Иван Петрович? Лжешь!* ***Не удавился — Удавлен он****! Ты удавил его! …Убийца! Зверь!* (А. К.Толстой).

Ср. также английские примеры, где пассив противопоставлен рефлексиву и подчеркивается дифференциальная сема 'д р у г и м и':

(7) *Cromwell. Then know that the King commands me to charge you in his name with great ingratitude! And to tell you that there was nor never could be so villainous a servant nor so traitorous a subject as yourself.*
More. So I am brought here at last.
Cromwell. Brought? ***You brought yourself*** *to where you stand now!*
More. Still, in another sense, ***I was brought*** (R. Bolt).

(8) *Without realizing it she judged people as much by standards of Walter Scott and Jane Austen as by empirically arrived at; seeing those around her as fictional characters and making poetic judgements of them. But alas what she had thus* ***taught herself*** *had been very largely vitiated by what she* ***had been taught*** (J. Fowles).

Однако наряду с этим встречаются употребления, которые не предполагают обязательного несовпадения деятеля с пациенсом: в следующих примерах состояние человека или его положение в пространстве является результатом его собственных действий, а не воздействия извне, как в приведенных выше примерах.

(9) *И* ***одет*** *он* ***был*** *проще, чем всегда* ***одевался*** (М. Горький).

(10) *The host* ***was turned*** *towards her with an anticipatory smile* (Maugham).

Примечательно, что в этом отношении не все глаголы одинаковы — и, по-видимому, не все языки. О человеке, который причесался (умылся, оделся), можно сказать, что он причесан (умыт, одет) и т. д., но о самоубийце нельзя сказать *он убит (отравлен, повешен, застрелен*). Эти различия, возможно, связаны с глаголами как лексическими единицами: так, например, *он отравлен* также не может значить *он отравился* в смысле «съел что-то ядовитое или недоброкачественное», т.е. в данном случае обязательно присутствует значение «в результате действий других»; таким образом, причастия указывают не просто на некоторое положение дел (в том числе возникшее в результате какого-либо действия), но и на само действие. Примечательно, что английский язык в этом отношении проявляет некоторые отличия от русского: ср. примеры (4) и (5).

1.3. Внутри пассива также часто противопоставляются собственно пассив (акциональный пассив), статив и результатив, причём последние два случая иногда расцениваются не как разновидности пассива, а как самостоятельные залоги [Буланин 1978]. Действительно, семантическое различие отражается и в ряде формальных различий: несовпадении значений видо-временных форм, совместимости с различными модификаторами и т.д. В английском и французском языках эти два значения могут дифференцироваться видо-временными формами, ср.:

(11) *The cotton which* ***half-filled*** *the bed of the pickup truck* ***had been covered*** *for the night with tarpaulin, so he didn't even need the quilt (Faulkner).*

В первом случае указывается на наличное состояние, синхронное времени восприятия персонажа, во втором — на предшествующее действие, которое восстанавливается на основе наблюдаемого результата: поверх хлопка был брезент[1].

[1] Формы Past Perfect Passive могут указывать не только на предшествующее действие, но и на предшествующее состояние, ср. (12) *The jug* ***had evidently been once filled*** *with water, as it was covered inside with green mould* (Wilde).

Примечательно, однако, что в подобных случаях формы могут быть взаимозаменимы при указании на одну и ту же денотативную ситуацию:

(13) The paper had the appearance of a rough map. By much folding it ***was creased and worn*** *to the pitch of separation, and the second man held the discoloured fragments together where they* ***had parted*** *(Wells).*

Стоит отметить последнюю форму *had parted*: в том же контексте при сохранении содержания могло быть *parted,* хотя способ показа был бы другой.

Состояние, совпадающее с временным планом текста, может описываться с указанием на некоторые характеристики предшествующего действия, что довольно часто встречается в пассивных конструкциях:

(14) The paper had the appearance of a rough map. ***By much folding*** *it* ***was creased and worn*** *to the pitch of separation…* (Wells).

Описывается внешний вид карты на момент повествования.

(15) I know where it ***is hidden****, and* ***for what purpose*** (Wilde).

В последнем примере *is hidden* указывает на местонахождение, т.е. на статическое положение, актуальное в момент речи (иначе было бы *was hidden),* тогда как упомянутая далее цель относится к предшествующему действию.

Таким образом, аналогично рассмотренным выше случаям, пассив и статив могут как противопоставляться, так и совмещаться или не дифференцироваться, т.е. могут допускаться оба прочтения без изменения смысла в одном и том же выражении.

2. Безличные предложения, в отличие от сопоставимых с ними личных структур, представляют ситуацию как результат стихийного процесса или как «перцептивное происшествие» (термин заимствован из работы [Розина 2000]) и подчёркивают пациентивное значение (affectedness) соответствующего актанта, причём собственная семантика конструкции ярче всего проявляется в нестандартных употреблениях, ср.:

(16) С тобой ***нас*** *нынче* ***затесало*** *В толпу глазеющих зевак* (А. Блок).

В отличие от стандартного *мы затесались*, употребление безличной конструкции указывает на неконтролируемый характер ситуации: она возникла не по воле людей, которые оказались в толпе, а сложилась стихийно.

С другой стороны, семантика безличных конструкций этого типа предполагает перцептивную составляющую; ср. эпизод из романа М. Булгакова «Роковые яйца», в котором описывается нашествие гигантских рептилий:

(17) Но огромная пружина, оливковая и гибкая, сзади, выскочив из овального окна, перескользнула двор, заняв его весь пятисаженным телом, и во мгновение обвила ноги Щукина. ***Его швырнуло*** *вниз на землю, и блестящий револьвер отпрыгнул в сторону* (Булгаков).

Описывается ощущение Щукина: хотя упоминается тварь, которая его швырнула, показано не её действие, а то, что случилось с персонажем.

Однако встречаются примеры, в которых личная и соответствующая безличная конструкция полностью равнозначны в плане семантики и различаются только на уровне коммуникативной структуры:

(18) Три года назад ***во время сильной бури вывернуло*** *с корнем высокую старую сосну, отчего и образовалась эта яма* (Чехов).

Ср. *сильная буря вывернула с корнем сосну*, где *буря* получила бы больший коммуникативный вес, чем в приведенном примере.

Понижение коммуникативного статуса актанта, обозначающего действующую силу, причину, связано с тем, что внимание сосредоточено в первую очередь на самом процессе, зрительном или ином ощущении наблюдателя, ср.

(19) Запирая за охотником дверь, лесник видел, как лужи на просеке, ближайшие сосны и удаляющуюся фигуру гостя ***осветило молнией*** (Чехов).

Ср. *осветила молния.*

В последнем случае молния явно привлекает большее внимание, нежели приведенном примере, где *осветило молнией* представляет собой нерасчленную рему; связь такого способа изображения с восприятием подтверждается словами *видел, как*.

Вместе с тем встречаются контексты, где личные и безличные конструкции абсолютно равнозначны:

(20) Перекатывающийся стук покрыл лестницу, и в ответ ему, как оглушительная зингеровская швейка, завыл и затряс всё здание пулемёт. Стёкла и рамы ***вырезало*** *в верхней части, как ножом, и тучей пудры понеслась штукатурка по всей бильярдной* (Булгаков).

(21) Вслед за ними выскочил Коротков и очень вовремя, потому что ***пулемёт*** *взял ниже и* ***вырезал*** *всю нижнюю часть рамы (Булгаков).*

Варьирование конструкций связано, по-видимому, с требованиями формальной связности текста и стремлением разнообразить синтаксис.

3. Аналогичную картину представляют производные диатезы действительного залога, в которых подлежащим является не агенс, а средство (орудие и актанты сходной ролевой семантики). Структуры такого типа, во-первых, широко используются в художественной литературе как чисто изобразительный прием, что может определяться восприятием наблюдателя (рассказчика или персонажа): часто устраняется партиципант, не входящий в поле его зрения.

(22) Через полчаса зверь высунул из травы мокрый черный нос, похожий на свиной пятачок. ***Нос*** *долго* ***нюхал*** *воздух и дрожал от жадности* (Паустовский).

В поле зрения наблюдателей попадает только нос, в то время как сам обладатель носа не виден в траве. Ср.: *Зверь нюхал воздух…*

(23) Далеко за лесом ***играла свирель*** *возвращавшегося пастуха* (Чехов).

Такое предложение вряд ли применимо для ситуации, когда пастух виден. Ср. возможное для описания обеих ситуаций: *играл на свирели пастух.*

(24) Ему удалось очень плотно и ладно ударить куда-то, — куда — он не видел, — но тотчас ***множество кулаков****, справа, слева, куда ни сунься,* ***продолжало*** *его* ***обрабатывать*** (Набоков).

В приведённых примерах дан фрагмент ситуации, видимый, доступный непосредственному ощущению или слуху; описываются не столько события или действия, сколько ощущения: не «что произошло», а «что увидел (услышал, почувствовал)» рассказчик или персонаж, т.е. в большей степени описывается зрительный образ, ощущение или звук, нежели само событие. Действие осмыслено как перцептивное происшествие.

Устраняться из «кадра» может не только невидимое, но и просто не вызывающее внимания наблюдателя. В художественных текстах часто устраняется участник ситуации, не попадающий в фокус внимания наблюдателя, хотя и видимый ему. При этом в позиции подлежащего находится элемент, привлекающий внимание; часто это подвижный элемент [Апресян 1970], кроме того, пример (25) отражает широко отмечаемую в психолингвистической литературе тенденцию к совпадению подлежащего с наиболее ярким, внешне заметным ситуации [Osgood, Bock 1975], [MacWhinney 1977]:

(25) Макар ясно видел острые уши лисицы; ее пушистый ***хвост вилял*** *из стороны в сторону, как будто заманивая Макара в чащу* (Короленко).

Вне подобного контекста едва ли возможно *хвост лисицы вилял*; но в связи с отмеченной особенностью нейтральный вариант *лисица виляла хвостом* мало подходит для данного контекста.

В подобных употреблениях данная структура уже не указывает на характер денотативной ситуации, а только распределяет актанты в зависимости от их значимости с точки зрения говорящего (рассказчика) или наблюдателя.

Однако аналогичное диатезное преобразование может быть связано со значением ослабленного контроля:

(26) И тут я сам не заметил, как ***руки*** *мои* ***открыли*** *ящик, где лежал злополучный роман* (Булгаков).

Ср. *…как я открыл ящик.*

(27) И когда наконец я возвращаюсь к себе в кабинет, моё ***лицо*** *всё ещё* ***продолжает улыбаться****, должно быть, по инерции* (Чехов).

Ср. *я всё ещё продолжаю улыбаться.*

Между отмеченными значениями просматривается связь: субъект неконтролируемых действий часто находится в фокусе эмпатии и собственное действие представляется как неожиданное для него самого, уподоблено восприятию внешнего события.

Кроме того, в структурах этого типа характерное для подлежащего каузальное значение может приписываться не фактическому каузатору действия, а другому актанту, который, соответственно, оформлен как подлежащее и в связи с этим может приобретать даже оттенок агентивного значения:

(28) Но как беспощадно-неблагодарно было всё то, что выходило из-под его кисти! ***Кисть*** *невольно* ***обращалась*** *к затверженным формам, руки складывались на один заученный манер…* (Гоголь).

Ср. *художник невольно обращался…*

В то же время встречаются примеры употребления подобных деагентивных конструкций как чисто изобразительного приёма, в каком-то смысле парафраза, метонимии на уровне текста: вместо целой ситуации упоминается её фрагмент.

(29) Изломанный ***веер закрывает*** *хорошенькое личико. Писатель подпирает кулаком свою многодумную голову, вздыхает и с видом знатока-психолога задумывается* (Чехов).

Ср. *Дама закрывает лицо веером.*

(30) Работы производятся по команде. Обыватели разом нагибаются и выпрямляются; сверкают лезвия кос, ***взмахивают грабли, стучат заступы, сохи бороздят*** *землю — все по команде* (Салтыков-Щедрин).

В таких контекстах этот прием иногда служит созданию визуального образа, но чаще, по-видимому, просто позволяет разнообразить синтаксис или, наоборот, создавать параллелизм, служить компрессии текста или поддерживать его связность; в таких случаях употребление той или другой структуры в основном формально.

Е. В. Падучева [Падучева 1994], обращаясь к этому типу диатез, отмечает, что за ними стоит переосмысление типа ситуации; как видно из рассмотренного выше, это действительно часто имеет место, но это не единственная возможная функция данного диатезного преобразования.

Перечисленные значения могут проявляться независимо друг от друга, ни одно из них не обязательно и ни для одного из них не обязательна данная форма выражения. Все эти значения должны быть признаны равноправными: любого из них достаточно для употребления маркированной диатезы. Ограничений на их проявление или тенденции к предпочтительному прочтению выявить не удалось.

4. Анализируя частные случаи, мы обнаруживаем, что с одними и теми же структурами могут связываться различные смыслы, причем между ними нет четкой границы и то, что противопоставляется в одних употреблениях, совмещается или остается неопределенным в других. Таким образом, семантика, связываемая с тем или иным типом конструкций, вовсе не обязательно реализуется в каждом конкретном употреблении и для многих контекстов противопоставляемые диатезы оказываются равнозначными, т. е. синтаксические структуры могут и не нести никаких из указанных значений, а служить для аранжировки компонентов в соответствии с коммуникативным заданием или поддерживать формальную связность текста, а могут просто разнообразить синтаксис, помогая избежать монотонности, и формировать ритмическую структуру текста.

На данном этапе исследования не проступает никаких явных закономерностей, предопределяющих, какое из возможных значений реализуется в данном контексте; вопрос о возможностях их выявления и тем более формализации правил следует оставить на дальнейшую перспективу.

Литература

1. *Апресян Ю. Д.* Семантические преобразования и синтагматические фильтры // Машинный перевод и прикладная лингвистика. Вып. 14. 1970.
2. *Буланин Л. Л.* К соотношению пассива и статива в русском языке //Проблемы теории грамматического залога. Л.: 1978.
3. *Гаврилова В. И.* Краткое причастие на -н/-т как форма статального вида страдательного залога // Типология вида. Проблемы, поиски, решения. М.: 1998.
4. *Князев Ю. П.* Результатив, пассив и перфект в русском языке // Типология результативных конструкций. Л.: 1983
5. *Падучева Е. В.* Типы каузальных отношений в семантической структуре лексемы // Russian linguistics. 1994. Vol. 18. № 1.
6. *Розина Р. И.* От происшествий к действиям (Семантическая деривация как способ пополнения общего жаргона) // Русский язык сегодня. 1. М.: 2000.
7. *Huddleston R., Pullum G. K.* The Cambridge Grammar of the English Language. Cambridge: 2002.
8. *MacWhinney B.* Starting Points // Language. 1977. Vol. 53. № 1.
9. *Osgood C., Bock R.* Salience and Sentencing: some production principles // Sentence Production: Developements in Research Theory. 1975.

К вопросу о соотношении слова и жеста (вокальный жест *О* в устной речи)

On gesture–word correlation (vocal gesture *Oh* in spoken Russian)

Гришина Е. А. (rudi2007@yandex.ru)

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

В статье на материале Мультимедийного русского корпуса (МУРКО) анализируется употребление вокального жеста *О*. Для анализа привлекаются сведения о типе *телесной* жестикуляции, которая сопровождает этот вокальный жест в речи. В результате исследования у вокального жеста *О* обнаруживается три типа употреблений — дейктическое (*О* в значении указательной частицы), эмоциональное (*О* в значении устного междометия) и физиологическое (*О* как возглас).

Что скрывает в себе междометие «О»?
О, мое естество! О, мое Божество!
О, спасибо! О, сжалься! О, лучше всего
Междометье восторга и трепета — «О».

Нонна Слепакова

Фактически, язык жестов значительно легче понять, чем вокальную часть сообщений, и для меня всегда было загадкой, почему <…> современные исследователи не концентрируются на языке жестов <…> По крайней мере, у лошадей жесты гораздо более упорядоченны, чем звуки, и намерения лошади легко могут быть поняты по ее жестам.

Henry Blake. Talking with Horses. A Study of Communication between Man and Horse // http://filly.msk.ru/articles/vocab/blake5.htm, пер. TriA)

1. Постановка задачи

В ходе разработки общей структуры Мультимедийного русского корпуса (МУРКО)[2], создаваемого в рамках Национального корпуса русского языка, накопились некоторые любопытные лингвистические данные, которые будут представлены в настоящей статье. Напомним, что МУРКО предполагает выдачу пользователю морфологически и семантически аннотированных расшифровок устных текстов параллельно со звуковой дорожкой (а для фильмов — также параллельно с видеорядом). Это позволит в ближайшем будущем ставить и решать задачи, связанные с фонетикой, орфоэпией, а также с жестовым сопровождением устной речи, — в дополнение к стандартным задачам, которые можно решать на материале НКРЯ. В частности, МУРКО открывает новые перспективы для изучения системы междометий в русском языке.

В исследованиях последних лет часто звучит мысль о том, что системы междометий в русской устной и в русской письменной речи, будучи безусловно связанными между собой, тем не менее не только функционируют достаточно автономно, но и, вполне вероятно, имеют лишь частично пересекающиеся множества единиц. В работе [Шаронов 2006] автор прямо предлагает различать — в том числе и терминологически — соответствующие устные и письменные единицы: «То, что принято называть междометием, в языке имеет две различные реализации —

[1] Исследование проведено при поддержке гранта РФФИ 08–06–00371а и программы ОИФН РАН 2009–2011 гг. «Генезис и взаимодействие социальных, культурных и языковых общностей».

[2] О МУРКО см. [Гришина, Савчук 2008], [Гришина 2009], а также [Кудинов, Гришина 2009].

устную (называемую в этологии *вокальным жестом*) и его письменную фиксацию, которую мы и будем называть собственно *междометием*».

Представляется, что это верный взгляд на предмет. Но если проводить такое разделение, то следует признать, что система междометий как предмет исследования находится в данный момент в гораздо более выгодном положении, чем система вокальных жестов, поскольку для второй отсутствует достаточно представительный общедоступный корпус, а следовательно, на вокальные жесты вынужденно переносятся свойства междометий (т. е. на устные единицы переносятся особенности письменных единиц)[3]. В настоящей статье мы хотели бы с помощью данных, полученных из предварительных материалов для МУРКО, проанализировать такой относительно частотный вокальный жест, как *О*.

Отметим при этом, что не будут рассматриваться некоторые типы контекстов: 1) контексты, в которых вокальный жест *О* произносится без паузы перед следующим словом, т.е. конструкции типа *О господи! О черт! О нет! О* + звательный падеж, *О* + императив (такого типа контексты требуют отдельного анализа, поскольку в них вокальный жест *О* не является самостоятельным высказыванием[4]); 2) контексты, в которых вокальный жест *О* произносится не на выдохе, как большая часть междометий, а на вдохе[5]; 3) контексты со слитными (т. е. без пауз) повторами данного вокального жеста.

2. Методика исследования

Обратим внимание на тот факт, что исходные данные для анализа междометия (письменного) *О* и вокального жеста (устного) *О* у нас принципиально разные.

Для **междометия** это — ближайший контекст, который позволяет 1) проанализировать описываемую в тексте ситуацию, 2) определить (в частности, методом внутреннего проговаривания), какой именно тип произнесения междометия уместен в описываемой ситуации и 3) вывести из этого способа произнесения посредством интроспекции значение междометия[6].

Что касается **вокального жеста**, то здесь ситуация несколько иная. Помимо конситуации, а также интонационной, тембровой, фонетической и иных характеристик вокального жеста, он в большинстве случаев[7] сопровождается обычным телесным жестом (сочетанием жестов)[8]. Телесные жесты (этим словосочетанием в данной статье мы обозначаем жесты, осуществляемые посредством человеческого тела, и противопоставляем вокальным жестам) обычно не привязаны к конкретному слову, образуют самостоятельную знаковую систему и, следовательно, могут рассматриваться как некоторые независимые маркеры значения того или иного вокального жеста. Таким образом, вполне осмысленной является задача — попытаться проанализировать данный конкретный вокальный жест, опираясь на сопровождающие его телесные жесты[9].

При этом следует иметь в виду, что теоретически между любым словом во фразе (в т. ч. и вокаль-

---

[3] Последнее связано также с тем, что «взрослые осознают как слова родного языка только те междометия, которые встречались им в написанном виде» [Протасова 2005: 165].

[4] Лингвистическая и педагогическая литература не определилась с частеречной принадлежностью *О* в такого типа контекстах. Некоторые словари (например, [Ушаков]), исследователи (см. [Шаронов 2006]) и учебные пособия (см. [Учебник Грамоты]) считают эту частицу междометием, другие — (усилительной, местоименной) частицей (например, [Розенталь 2005: 233]). Мы не беремся предлагать окончательный ответ на этот вопрос, представляется, однако, довольно очевидным, что решения здесь должны приниматься отдельно для письменного и для устного функционирования этой единицы.

[5] Считается, что все слова русского языка, в т.ч. и междометия, произносятся на выдохе (см., например, [Протасова 2005: 165]). Обратим, однако, внимание на то, что это неверно не только для вокального жеста *О*, но и для вокальных жестов *А, И, У*. Все они в ряде конситуаций могут произноситься на вдохе, и у всех у них в этом случае проявляется один общий смысловой компонент ‘внезапный ужас, испуг’ или ‘возмущение’. Безусловно, появление этого компонента связано с таким явлением, как задержка или перебив дыхания при сильном испуге или сильном возмущении. Каждый из перечисленных выше гласных звуков трансформирует это основное чувство по-своему (например, *И* на вдохе тяготеет к возмущению, *О* — к испугу), но исследовать и описывать эти единицы следует в составе группы, а не поодиночке.

[6] См. в работе [Шаронов 2006а]: чтобы понять значение междометия, «необходимо распознать стоящий за междометием вокальный жест, реальное звучание, и воспроизвести его хотя бы про себя. Если мы не распознаем вокальный жест, то вряд ли поймем эмоциональное состояние персонажа».

[7] Из послуживших материалом данного исследования 236 случаев употребления вокального жеста *О* в кинофильмах — только в 39-ти (т.е. только в 16% случаев) вокальный жест не сопровождался телесным жестом (естественно, мы не принимаем во внимание те ситуации, когда организация кадра не давала нам возможности увидеть телесное сопровождение вокального жеста — например, говорящий был вне камеры).

[8] На важность этого обстоятельства указывал И. А. Шаронов: «В устной речи вокальный жест реализует возможности тона, тембра и долготы звучания, выступает не в одиночестве, а «аккомпанируется» соответствующей мимикой и симптоматическими жестами, что позволяет относительно легко идентифицировать передаваемую эмоцию» [Шаронов 2006а].

[9] Идея использовать для толкования междометий сопровождающие их жесты не нова, мы лишь попытались воплотить ее в жизнь более или менее последовательно. Безусловно, для такой постановки задачи исследователю положено иметь если не законченную классификацию, то достаточно полное описание значительного числа телесных жестов. В процессе подготовки МУРКО, нами (с опорой на работы Г.Е.Крейдлина и его соавторов [Крейдлин 2002] и [Григорьева и др. 2001]) было разработано такое описание — на основе сплошной росписи жестикуляции и ее анализа по ряду единообразных параметров в нескольких фильмах (см. об этом подробнее [Гришина 2009] и [Кудинов, Гришина 2009]). В данной работе мы используем именно этот материал.

Таблица 1

+	0	–
значения слова и жеста совпадают (т. е. либо слово дублирует значение жеста, либо наоборот, жест дублирует значение слова)	жест и слово независимы друг от друга (частный случай — использование жеста без слова или использование слова без жеста)	жест и слово противоречат друг другу (т.е. значение каждой из единиц опровергает значение другой)

ным жестом) и любым сопровождающим его телесным жестом могут складываться три типа отношений (см. табл. 1).

Понятно, что использовать значение телесного жеста для анализа значения вокального жеста возможно только в первом случае, т.е. при положительном, плюсовом отношении между словом и жестом. Но для этого нужно иметь какие-то основания для признания данного отношения между словом и жестом плюсовым, а не нулевым или минусовым. Представляется, что таким основанием может служить повторяемость: если между вокальным и телесным жестом существуют смысловые совпадения, то их сочетание в потоке речи не может быть случайным, а следовательно, одиночным. Напротив, независимость значений вокального и телесного жеста (нулевое отношение) или их противоречие друг другу (минусовое отношение) должны приводить к отсутствию повторов, к случайности сочетания вокального и телесного жеста[10].

3. Основные типы телесных маркеров, сопровождающих вокальный жест О

Для того, чтобы описать типы телесных жестов, которые сопровождают вокальный жест *О*, будем использовать понятие (семантически) производного жеста. Производным жестом считается жест, значение которого включает в качестве семантической составляющей значение другого жеста, играющего в данном случае роль своего рода семантического примитива.

I. Например, в качестве базового жеста может рассматриваться ***указательный жест*** (указание пальцем, рукой, обеими руками, подбородком, кивок и проч.). Как разновидность указательного жеста следует рассматривать также такие жесты, как *демонстрация* или *предъявление* слушающему некоторого объекта, которые предполагают, что слушающий обязательно обратит на них внимание. На базе этого семантического примитива выстраиваются следующие производные жесты (классы жестов):

1) жест ‘похвала’=показать большой палец (включает в себя *демонстрацию* большого пальца слушающему)[11]
2) жесты ‘фиксация объекта’ (жесты, вычленяющие в окружающей действительности те или иные объекты, на которые в дальнейшем говорящий может *указать* слушающему, например, пристукнуть по какому-либо объекту, являющемуся предметом разговора, или поднять указательный палец вверх для фиксации важной темы, появившейся в разговоре)
3) жесты ‘привлечь внимание’ (жесты, которые привлекают внимание слушателя к некоторому объекту, т. е. *указание* на некоторый объект, но без использования базового указательного жеста, например, коснуться кого-л., протянуть руку к кому-л.)
4) жесты ‘обратить внимание’ (жесты, с помощью которых говорящий вычленяет некоторые объекты в окружающей действительности для себя, в своих собственных интересах, т. е. как бы *указывает* на них себе самому, например, проводить взглядом, оглянуться на что-л.)

Анализ материала показывает, что порядка 40% случаев употребления вокального жеста *О*, которые сопровождаются теми или иными телесными жестами, являются в своей основе дейктическими, т. е. аккомпанируются либо базовым, либо производным дейктическим жестом. Приведем некоторые примеры: (1)–(5).

II. Следующим «блоком» телесных жестов, существенных для анализа вокального жеста *О*, с нашей точки зрения, является блок, основанный на жестах, выражающих ***удивление***[12] (например, поднять бро-

---

[10] Здесь уместно будет подчеркнуть, что материал, использованный в данной работе, — расшифровки кинофильмов в сочетании с соответствующими видеофрагментами, — ни в коем случае не может рассматриваться как исчерпывающий. Уже не говоря о том, что обследованный объем кинотранскриптов недостаточен для фундаментальных выводов, он еще и отстает от современной жизни. Например, пока нам не удалось зафиксировать в фильмах примеров заимствованного двойного междометия *о-О*, с ударением на втором *о*, с твердым приступом перед обеими частями и с понижением тона на кварту (это междометие широко используется сейчас в молодежной среде в значении ‘ничего себе! ой-ёй-ёй!’ и сходных).

[11] Сходным образом устроен жест ‘отказ’ = показать кукиш, однако на нашем материале он встретился только в сопровождении непосредственного дейксиса *во* (ср. ниже).

[12] На это и ранее указывали исследователи, см., например, [Борисова 2004]: «*О* — … может выражать удивление и другие оттенки, связанные с реакцией на что-то новое».

(1) Базовый указательный жест:

[Афоня/Леонид Куравлев] [поет] *А распашу ль я!* [Егоза/Савелий Крамаров] [поет] *А распашу ль я!* [Афоня/Леонид Куравлев] [поет] *Пашенку!* [Егоза/Савелий Крамаров]:

Речевой ряд	*Пашенку!*		*О!*	*Иван Иваныч Иванов! Помнишь?*
Событийно-жестовый ряд	[поет]	[выглядывает в окно]	[показывает пальцем]	

[Афоня/Леонид Куравлев] *А то!* *«Афоня» (Г. Данелия)*

(2) ‘похвала’:

[Макарыч/ Алексей Смирнов] *Кузнечик.* [Кузнечик/ Сергей Иванов] *Я.* [Макарыч/ Алексей Смирнов]:

Речевой ряд	*Иди к командиру. Настроение /*	*о!*
Событийно-жестовый ряд		[показывает большой палец]

«В бой идут одни “старики”» (Л. Быков)

(3) ‘фиксация объекта’:

[Голос по радио] *…трудовых вахт. Труженики Липецкого машиностроительного комбината взяли на себя новое обязательство по повышению производительности труда.* [Крымов/Станислав Говорухин] *Ой / до чего ж всё это надоело! Ну какая производительность труда? Чем больше горбатишься / тем меньше тебе платят.*

Речевой ряд	*Крути дальше!*		*О!*	*О / оставь.*
Событийно-жестовый ряд	[обращается к Алике]	[по радио поёт Джо Дассен]	[поднимает вверх указательный палец]	[поднимает вверх указательный палец]

«Асса» (С. Соловьев)

(4) ‘привлечь внимание’:

[Андрей/ Геннадий Гарбук] *Ну а теперь насчёт хат. У нас на работе участки дают за городом. Свою я туда… перевезу под дачу…*

Речевой ряд	*О!*	*И насчёт твоей договорился.*
Событийно-жестовый ряд	[похлопывает отца по руке]	

Наш шеф купит. На снос. *«Белые Росы» (И. Добролюбов)*

ви, отпрянуть, развести руками и мн. др.)[13]. Этот базовый комплекс жестов также порождает несколько производных.

1) жесты ‘высокой оценки’ (жесты, обозначающие настолько высокое качество некоторого объекта, что это вызывает у говорящего *удивление*, например, мотать головой, откинуть голову назад, закрыть глаза и мн. др.)
2) жесты ‘приветствия’ (жесты, сопровождающие этикетную ситуацию приветствия, когда говорящий впервые после некоторого перерыва видит адресата, что вызывает у говорящего либо *удивление*, либо *высокую оценку*, либо сочетание этих эмоций, — например, раскинуть руки, вскинуть руки, встать, рукопожатие и др.)[14]

Анализ материала показывает, что порядка 43% случаев употребления вокального жеста *О*, которые сопровождаются теми или иными телесными жестами, аккомпанируются либо базовыми жестами со значением ‘удивление’, либо производными. Примеры (6а–г)–(8).

III. Третьим блоком телесных жестов, существенных для анализа вокального жеста *О*, является блок жестов, которые условно можно назвать ***физиологическими***. Здесь можно выделить две группы

[13] Как известно, удивление может сопровождаться как положительными, так и отрицательными эмоциями, соответственно, в базовой группе собраны случаи как удивления-удовлетворения, так и удивления-недовольства (пример последнего — разочарование). Кроме того, удивление может сопровождаться усмешкой, улыбкой, смехом — возникает ситуация насмешки, в которой используются соответствующие жесты (повести подбородком вбок, склонить голову набок).

[14] Таким образом, наш материал входит в противоречие с работой [Борисова 2004], в которой утверждается, что в междометии *о* при «сравнении с *а* очевиден дополнительный оттенок важности, серьезности того, на что реагируют. Этот оттенок может давать и значение скорби (серьезность огорчения): *О-о-о, как это ужасно!* Радость получается не вполне естественная: *О, мы встретились!*». Это рассогласование может объясняться, среди прочего, ориентацией цитированной работы на письменные источники (за неимением других), а нашей работы — исключительно на устную речь.

(5) ‘обратить внимание’:

[Женский голос] [зовет] *Маша!* [Маша/ Татьяна Мычко]:

Речевой ряд	*О!*	*Нас зовут.*
Событийно-жестовый ряд	[оглядывается]	[встает, уходит]

[Женский голос] [зовет] *Саша!* *«Долгие проводы» (К. Муратова)*

Базовый жест ‘удивление’:

(6а) ‘приятное удивление’:

[Тютюрин/ Георгий Вицин] [показывает на жену] *Это моя Зина!*

Речевой ряд	[Переводчик принца Бурухтана/ Борис Сичкин] *О /*	*Зина!* [Тютюрин / Георгий Вицин] *А рядом / Эдита Пьеха!*	[Переводчик принца Бурухтана/ Борис Сичкин] *О /*	*Пьеха! Не может быть.*
Событийно-жестовый ряд	[откидывает голову назад]		[откидывает голову назад]	

«Неисправимый лгун» (В. Азаров)

(6б) ‘неприятное удивление’:

[Нервный / Владимир Большов] *Ну вот. Вот такая история. Сплошное недоразумение. И что? Что теперь?!*

Речевой ряд	[Сухонький / Василий Мищенко] *О /*	*ничего себе!*
Событийно-жестовый ряд	[качает головой]	

[Нервный / Владимир Большов] *Что? Что теперь-то? А? Что?* *«А поутру они проснулись» (С. Никоненко)*

(6в) ‘разочарование’:

Речевой ряд		[Оксана/ Лидия Федосеева-Шукшина] *О…*	снова каша?
Событийно-жестовый ряд	[Оксана подходит к столу]	[надувает губы]	

[Колхозник/ Виктор Филиппов] *Оксаночка / садитесь.* *«Из жизни отдыхающих» (Н. Губенко)*

(6г) ‘насмешка’:

Речевой ряд		[Дон Сезар/ Михаил Боярский] *О!*	*Вот это дело!*
Событийно-жестовый ряд	[Дон Хосе выхватывает шпагу]	[поводит подбородком вбок, усмехается]	

«Дон Сезар де Базан» (Я. Фрид)

(7) ‘высокая оценка’:

[Гуськов/ Алексей Миронов] [на берегу у костра] *Оставайтесь / Андрей Егорыч. Вот на утренней зорьке… заведём…*

Речевой ряд	*От рыба-то пойдёт!*	*О!*	*Пойдёт! Пойдёт!*
Событийно-жестовый ряд		[закрывает глаза, запрокидывает голову]	

«Простая история» (Ю. Егоров)

(8) ‘приветствие’:

Речевой ряд	[Сергей/ Олег Даль] *А где наша мама?*	*О!*	*Вот наша мама!*
Событийно-жестовый ряд	[выходит из машины, оглядывается]	[вскидывает руку]	

«В четверг и больше никогда» (А. Эфрос)

телесных жестов, которые имеют между собой мало общего. Но их объединяет тот факт, что все они находятся на грани физиологии и социологии.

1) жесты ‘больно’, ‘неприятно’ (жесты, которые выражают физическую боль или физиологическое отвращение к чему-л., например, сморщиться, скривиться, схватиться за больное место и пр.)
2) жесты ‘интенсивность действия’ (жесты, которые иллюстрируют силу какого-л. действия, мощность какого-л. события, например, трясти крепко сжатым кулаком). Этот базовый тип жестов имеет *производные* — жесты, которые иллюстрируют силу, интенсивность не внешнего действия, а психологического состояния

(9а) ‘больно’:

[Тимофей кладет Федосу на спину горячий кирпич] [Федос/ Всеволод Санаев] *Сымай скорей! Ой-ой!* [Тимофей/ Борис Новиков] *Терпи*.

Речевой ряд	[Федос/ Всеволод Санаев] *О… О…*	[Тимофей/ Борис Новиков] *Терпи.*	[Федос/ Всеволод Санаев] *О!*	*Ой!*
Событийно-жестовый ряд	[морщится]		[морщится]	[сбрасывает кирпич, встает и бежит за Тимофеем]

«Белые Росы» (И. Добролюбов)

(9б) ‘неприятно’:

Речевой ряд			[Король/ Юрий Богатырев] *О!*
Событийно-жестовый ряд	[Дон Сезар передает королю цепь дона Хосе]	[король кривится, осматривает цепь, бросает ее на пол]	[кривится, оглядывается на дона Сезара]

«Дон Сезар де Базан» (Я. Фрид)

(10а) ‘физическое напряжение’:

[Потёмкин/ Борис Ливанов] *А мне флотоводец надобен / а не ябедники / не фискалы.*

Речевой ряд	*Осьмнадцать доносов на него настрочили / а он знай себе турку лупит!*	*О!*
Событийно-жестовый ряд		[вытягивает вперед руку, трясет кулаком]

«Адмирал Ушаков» (М. Ромм)

(10б) ‘интенсивность чувства’:

Речевой ряд	[Плющихин/ Николай Трофимов] *Мадам!*	[Хозяйка гостиницы/ Людмила Гурченко] *О!*	[Плющихин/ Николай Трофимов] *От так!*
Событийно-жестовый ряд	[поднимает нож, подает его хозяйке]	[вытягивает по направлению к Плющихину шею, запрокидывает голову]	

«Табачный капитан» (И. Усов)

говорящего (например, вытянуть вперед шею и одновременно запрокинуть назад голову как выражение интенсивности какого-л. чувства)

Анализ материала показывает, что этот класс употребления вокального жеста *О* не очень обширен, только 4% случаев. Примеры (9а–б), (10а–б).

4. Фонетические особенности разных типов вокального жеста О

Таким образом, сочетание вокального жеста *О* с разными типами телесных жестов позволяет нам выделить три основные группы употребления этого вокального жеста — *дейктическая* группа, группа, в которой базовым типом являются жесты удивления (назовем для краткости эту группу *эмоциональной*) и *физиологическая* группа. Разумеется, эти три группы не исчерпывают случаи употребления вокального жеста *О*[15], однако объем обследованного материала пока достаточно невелик, поэтому не исключено, что при его расширении возникнут еще некоторые группы. В качестве потенциальных можно рассматривать, например, группу ‘усталость’ (говорящий кладет голову кому-л. на плечо), которую можно обобщить до группы ‘расслабленность’, а также группу ‘подтверждение, согласие’, когда говорящий согласно кивает и при этом произносит *О*, и т. д.: т. е. те примеры, которые сейчас выглядят изолированными и случайными, могут превратиться в самостоятельные группы.

Теперь следует рассмотреть материал с фонетической точки зрения (учитывая, разумеется, проведенное разбиение по группам). Очевидно, что поскольку 1) вокальные жесты все-таки не в 100% случаев сопровождаются телесными и 2) общение отнюдь не всегда устроено таким образом, что слушающий видит говорящего, вокальный жест, в частности, *О*, должен иметь некоторые фонетические характеристики, которые, в отсутствие телесной поддержки, должны помогать слушающему понимать говорящего. В стандартном случае употребления, т. е. при сочетании аудио- и видеоряда, семиотическая мощность вокального жеста достаточно вели-

[15] Это очевидно из того факта, что описанные в предыдущем разделе три группы в сумме не дают 100% употреблений вокального жеста *О*.

(11)

[Сваха/ Лидия Смирнова] *А в одном доме хотели надо мной насмешку сделать.* [Мать Бальзаминова/ Людмила Шагалова] *Да?* [Сваха/ Лидия Смирнова] *Вместо водки поднесли рюмку ладеколону.*

Речевой ряд	[Мать Бальзаминова/ Людмила Шагалова] *О!*	*Ну скажите!*
Событийно-жестовый ряд	[поднимает вверх палец (дейксис: фиксация объекта), поводит подбородком вбок (эмоция: сочувствие)]	

[Сваха/ Лидия Смирнова] *Ничего. Я выпила. Да ещё поблагодарила.* *«Женитьба Бальзаминова» (К. Воинов)*

ка, но и в отсутствие видеоряда аудиоряд должен хоть как-то справляться с задачей более или менее однозначной передачи смысла.

Начнем с третьей, *физиологической* группы. Здесь ситуация достаточно проста и однозначна: для большинства случаев употребления вокального жеста *О* в сочетании с соответствующими жестами характерно сочетание двух фонетических особенностей — 1) относительно длительное звучание и 2) форсаж звука, причем вокальный жест начинается с форсированного (усиленного) звучания (с добавлением хрипа, свиста, голосовых смычек или шипения), которое постепенно ослабевает.

Сложнее обстоит дело с дейктической и эмоциональной группой. На первый взгляд представляется, что разница здесь — между кратким (дейксис) и долгим (эмоция) звучанием (это различие отражается и в достаточно частом способе записи эмоционального *О* посредством удвоенного или утроенного «о» — *о-о* или *о-о-о*). Так, если мы сравним процентное содержание удлиненного/краткого *О* среди дейктических и эмоциональных вокальных жестов, то получим следующую картину:

Таблица 2

	всего	**дейктическая группа**	**эмоциональная группа**
долгое *О*	51%	*14%*	**79%**
краткое *О*	49%	**86%**	21%

Как видим, для дейктической группы характерна краткость вокального жеста *О* (отклонение от среднего — почти в два раза), а для эмоциональной группы — долгота (отклонение от среднего значения — в полтора раза). Проблема, однако, в том, что определить долготу/краткость вокального жеста в потоке речи объективно довольно затруднительно, поскольку долгота — в значительной степени относительный параметр, зависящий от индивидуального темпоритма речи (под темпоритмом понимаются индивидуальные особенности речи, связанные со скоростью и энергетикой речи данного говорящего в данных обстоятельствах), от ситуации общения и т.д. Желательно поэтому найти некоторый более независимый от индивидуальных особенностей речи параметр. Как представляется, таким параметром может служить наличие у дейктического *О* твердого приступа и отсутствие такового у эмоционального *О*:

Таблица 3

	всего	**дейктическая группа**	**эмоциональная группа**
О без твердого приступа	61%	*18%*	**93%**
О с твердым приступом	39%	**82%**	*7%*

Как видим, «крестообразное» распределение значений сохранилось — для дейктической группы значительно более характерно произнесение с твердым приступом, а для эмоциональной группы — без него[16].

Таким образом, твердый приступ перед вокальным жестом *О* может рассматриваться как признак, отличающий дейктическое *О* от эмоционального. Следовательно, в тех случаях, когда у вокального жеста нет поддержки в виде телесного жеста, именно наличие/отсутствие твердого приступа выступает как источник информации о выборе в пользу либо дейктического, либо эмоционального восприятия вокального жеста *О*. Обратим при этом внимание на то, что в тех случаях (относительно нечастых), когда вокальный жест *О* сопровождают телесные жесты из обеих групп, т.е. из дейктической и из эмоциональной, обычно «побеждает» дейктическая составляющая, т. е. вокальный жест *О* имеет при себе твердый приступ (см. пример 11).

Что касается фонетических характеристик вокального жеста *О* внутри группы ‘удивление’, то они достаточно разнообразны и представляют собой отдельный предмет для изучения. В частности, в этой группе широко применяется дугообразное изменение тона, понижение тона, растяжка без перепадов тона выше или ниже основного уровня тона для данного отрезка речи и проч. Нам, однако, представляется важным тот факт, что от двух других смысловых групп вокальных жестов *О* группа ‘удивление’ отличается двумя легко уловимыми

[16] Высокая степень корреляции между таблицами 2 и 3 объясняется тем, что *О* с твердым приступом может быть только кратким, а *О* без твердого приступа — и долгим, и кратким, но с явным преобладанием долгого.

и легко идентифицируемыми фонетическими признаками — отсутствием форсажа звука и отсутствием твердого приступа.

Фонетическое соотношение этих трех групп тем самым выглядит следующим образом:

Таблица 4

	‘удивление’	**‘дейксис’**	**‘физическое/психологическое напряжение’**
твердый приступ	–	+	–
форсаж	–	–	+

Таким образом, наши данные позволяют лишь частично согласиться с утверждением в [Шаронов 2006], что «релевантными фонетическими признаками вокальных жестов являются характер тона, долгота и лишь отчасти — артикуляционные признаки издаваемого звука». Характер тона чрезвычайно важен для того, чтобы мы могли отличить, например, приятное удивление от разочарования, но если нам не удастся провести — по любой причине — такого различения, то на наше восприятие высказывания в целом это роковым образом не повлияет, поскольку благодаря отсутствию твердого приступа или форсажа звука у вокального жеста *О* мы точно поймем, что это удивление, а не дейксис и не репрезентация физического напряжения. Для иллюстрации обратимся к цитате из Н. Слепаковой, предпосланной статье в качестве эпиграфа. Очевидно, что употребление вокального жеста *О* в сочетании *О, спасибо!* по крайней мере трехзначно: 1) значение, которое мы не рассматриваем в данной статье, — усилительная частица *О*, примыкающая без паузы к следующему слову, т.е. *О спасибо!;* 2) дейктическое значение — в ситуации, когда в поле внимания говорящего при участии слушающего появляется некоторый нужный говорящему объект и говорящий фиксирует его появление с помощью вокального жеста *О* с твердым приступом и одновременно благодарит слушающего с помощью *спасибо*; 3) эмоциональное значение, когда слушающий производит некоторое действие, которое может вызвать у говорящего а) удивление, б) высокую оценку, в) разочарование, г) насмешку, и все эти эмоции могут влиять тем или иным образом на характер тона вокального жеста *О*; слушающий при этом, в зависимости от его внимательности, фонетической подготовленности и тысячи иных факторов может понять говорящего правильно или неправильно в части выбора эмоции, но он никогда не воспримет в данном случае вокальный жест *О* как дейксис.

Что касается долготы/краткости вокального жеста *О*, то они, как представляется, самостоятельного значения не имеют (и это естественно — странно было бы полагаться на столько зависимый от индивидуальных особенностей говорения признак), а являются производными от наличия твердого приступа, который решительно не предполагает долготы вокального жеста, а только его краткость (что касается вокального жеста ‘физическое/психологическое напряжение’, то и в этом случае долгота/краткость имеют служебное значение и зависят от долготы и интенсивности форсированного начального звучания).

В заключение этого раздела в качестве версии выскажем предположение, что существенные фонетические особенности всех трех выделенных групп употребления вокального жеста *О* и стандартно сопровождающие их телесные жесты связаны иконическими отношениями, т.е. фонетические особенности вокального жеста «похожи» на физиологические особенности телесного жеста. Так, вокальный жест *О* в дейктическом значении имеет фонетически «точечный» характер — краткость произнесения в сочетании с точно обозначенным с помощью твердого приступа резким началом. Такой же точечный характер имеет базовый для этой группы употреблений *О* указательный жест *показать пальцем* — его объект, цель, на который указывает палец, также воспринимается как некоторая точка в окружающем пространстве. Вокальный жест *О* в физиологическом значении, как мы уже писали, имеет форсированное, усиленное по сравнению с нормой звучание — и большая часть сопровождающих его телесных жестов также «исполняется» на фоне повышенного физического напряжения. Что касается вокального жеста *О* в эмоциональном значении, то он подлежит дальнейшему исследованию — и, вероятней всего, не на слух, а с привлечением тех или иных технических средств, — но уже сейчас можно сказать о том, что в значительной части его употреблений мы имеем дело с выгнутым дугообразным движением тона ∩. При этот самыми частотными жестовыми сопровождениями в этой зоне являются жесты *поднять брови/бровь, запрокинуть голову, вскинуть руку/руки, широко раскрыть глаза*, *привстать*, *подскочить,* т.е. жесты, отчетливо включающие в качестве компонента подъем, движение вверх, и это заставляет нас подозревать иконические отношения между вокальным и телесным жестом и в этой смысловой зоне.

5. Заключение

Как видим, при опоре на телесные жесты совокупность употреблений вокального жеста *О* распадается на три зоны, каждую из которых можно считать либо омонимом, либо одним из значений. На данный момент нам представляется наиболее приемлемым следующее разделение: $О^1$=вокальный жест *О* в дейктическом значении, $О^2$=вокальный жест *О* в эмоциональном значении, $О^3$=вокальный жест *О* в физиологическом значении. Основанием для

такого разделения на омонимы является тот факт, что эти три варианта вокального жеста *О* должны иметь разную частеречную характеристику. Например, $О^1$, т. е. дейктический вокальный жест, с нашей точки зрения, должен характеризоваться как указательная частица. В работе [Гришина 2008], анализирующей соотношение в устной речи вариантов указательной частицы *вот — во* и *о*, — нам уже приходилось об этом писать. $О^1$ (*О* с твердым приступом) является таким же непосредственным дейксисом[17], как и *во*, и в большом количестве контекстов они могут заменять друг друга (так, в примерах 1–5, 11 в настоящей работе *о* легко заменяется на *во*, чего не скажешь о примерах из эмоциональной и физиологической зоны). В работе [Гришина 2008] было предложено считать, что различие между $о^1$ и *во* определяются разной степенью учета говорящим интересов слушающего: *во* осуществляет указание для говорящего и для слушающего (говорящий вычленяет для себя некоторый объект в окружающей действительности и одновременно указывает на него слушающему), а $о^1$ осуществляет указание прежде всего в интересах говорящего (т.е. говорящий с помощью $о^1$ вычленяет некоторый объект в окружающей действительности прежде всего для себя, удовлетворение интересов слушающего в данном случае — задача второго эшелона).

Анализ материала для настоящей статьи показал, что между $о^1$ и *во* существует еще и гендерное различие: из двух этих непосредственных дейксисов женщины употребляют *во* почти в 3 раза *реже*, чем в среднем.

Таблица 5

	всего	***о***	***во***
мужчины	87%	75%	95%
женщины	13%	**25%**	5%

Кроме того, очевидно различие между $о^1$ и *во* по социальному параметру. На т. н. «костюмные» фильмы, т. е. фильмы, имеющие отношение к прошлому, приходится только 3% всех употреблений *во*, и все говорящие в этих случаях относятся — по сюжету — к «социальным низам». Что касается $о^1$, то здесь процент употребления в «костюмных» фильмах существенно выше — 14%, и нет никакого ограничения по социальному статусу говорящих (естественно, мы далеки от того, чтобы считать такое распределение *во/о* отражением реального словоупотребления в соответствующие эпохи, но это распределение хорошо отражает представления об уместности того или иного дейксиса с точки зрения современного носителя русского языка). Из этого следует, что *во*, в отличие от нейтрального $о^1$, стилистически снижено (и как результат своего рода «положительного» шовинизма, зафиксированного в языке, в меньшей степени подобает женщинам, чем мужчинам). Объяснить сниженность *во* по сравнению с $о^1$, как представляется, можно именно тем, что *во* принуждает слушающего к немедленному восприятию объекта указания, а «интеллигентное» $о^1$, напротив, лишь фиксирует некоторый объект, не вовлекая слушающего в ситуацию указания и оставляя ему свободу выбора.

Что касается $О^3$, то этот вокальный жест не является ни указательной частицей, как $О^1$, ни собственно устным междометием, как $О^2$, но должен классифицироваться как физиологический возглас, балансирующий на грани между речевой единицей и доречевым вокализованным выражением того или иного физиологического или психологического состояния говорящего организма[18]. Сказывается это, в частности, в том, что при расшифровке фильмов и устных текстов очень часто расшифровщики пропускают вокальный жест *О* в этой функции, не включают его в текст. И действительно — будучи включенными в письменный текст без сопутствующих пояснений, возгласы $О^3$ истолковываются читателем как указательная частица $О^1$ и/или как устное междометие $О^2$, а для того, чтобы $О^3$ воспринималось адекватно, т.е. как физиологический возглас, текст расшифровки должен обязательно включать метатекстовые описания данного возгласа. Да и телесные жесты, сопровождающие возглас $О^3$, имеют минимальную социализацию, минимальную степень условности (например, *потереть ушибленное место, сморщиться* и под.).

Таким образом, предварительное словарное описание вокального жеста *О,* с нашей точки зрения, может выглядеть следующим образом.

$О^1$, указательная частица, чаще всего произносится с твердым приступом, кратко. **1.** Указывает на некоторый объект, появившийся в поле зрения или зоне внимания говорящего; сопровождающие телесные жесты: демонстрация; показать пальцем; показать подбородком; показать руками/рукой; предъявить что-л. [Четвёртый эцилопп] *И радуйся / радуйся!* [Машков/ Станислав Любшин] *Радуюсь!* [смотрит вверх, показывает подбородком] *О! Смотри / Пэжэ пришёл.* [Четвёртый эцилопп] *Где?* [смотрит вверх] [Г. Данелия. Кин-дза-дза (1989)] **2.** Фиксирует, вычленяет некоторый объект в окружающей действительности в качестве потенциального объекта указания; сопровождающие телесные жесты: двинуть головой вперед; двинуть кисть к кому-л.; поднять палец вверх; ткнуть пальцем. [Автор аутотренинга] *Я прислушиваюсь к своим мышцам. Они свободны / расслаблены…*

[17] См. о непосредственных дейксисах [Николаева 2004: 60].

[18] Именно к $О^3$ в наибольшей степени относится определение междометия как «реактивного эмоционального возгласа», «не обладающего признаками интенциональности и адресованности», использование которого «не имеет жесткой привязки к диалогу» (ср. [Шаронов 2008]).

[Коля/ Всеволод Шиловский] *О* [тычет пальцем вперед] */ это по делу. Расслабимся.* [разливает портвейн] [С. Микаэлян. Влюблен по собственному желанию (1982)] **3.** Сопровождает ситуацию обращения внимания говорящим на некоторый объект; сопровождающие телесные жесты: оглянуться; проводить взглядом. [Горпина Дормидонтовна/ Зоя Федорова] *Нечипор! Нечипор!* [Дед Нечипор/ Евгений Лебедев] [оглядывается] *О! Что-то моя… Гапуся… летит.* [А. Тутышкин. Свадьба в Малиновке (1967)] **4.** Привлекает внимание слушающего к какому-л. объекту; сопровождающие телесные жесты: поднять палец вверх; коснуться кого-л.; положить руки на чье-л. плечо; протянуть руку к кому-л.; ткнуть пальцем. [Лагутин/Сергей Донцов] [сыну] *Красиво говоришь!* [видит отца Майи, кладет руки на плечо сыну и Майе] *О / аке!* [отцу Майи] *Салем / аке!* [Ю. Мамин. Фонтан (1989)]. **5.** Сопровождает условные жесты (показать большой палец, показать кукиш), демонстрируя слушающему орган тела, осуществляющий данный жест. [Тютюрин/ Георгий Вицин] *Не-не-не / не надо / не надо! Закон есть закон! О!* [показывает большой палец] [В. Азаров. Неисправимый лгун (1973)]

О², устное междометие, может произноситься долго (чаще) или кратко (реже), обычно с изменением высоты тона. **1.** Выражает удивление, испытываемое говорящим; сопровождающие жесты: вскинуть руки; выдвинуть подбородок; качать/качнуть головой; надуть губы; откинуть голову; отпрянуть; повести подбородком вбок; поднять брови; упереться руками в бока; проводить взглядом; развести руками; стоп!; широко открыть глаза. [Антонио/ Алигьеро Носкезе] *Она красивая девушка. А вам она не нравится?* [Андрей/ Андрей Миронов] *О* [поводит подбородком вбок, поднимает брови] */ совсем наоборот*. [Э. Рязанов. Невероятные приключения итальянцев в России (1974)]. **2.** Выражает насмешку; сопровождающие жесты: повести подбородком вбок; склонить голову набок. [Искра/ Ирина Чериченко] *Уши. Он в детстве застудил уши. А теперь его не берёт медкомиссия.* [Зина/ Наталья Негода] *О…* [наклоняет голову набок] *Всё-то ты знаешь / и про модели / и про уши…* [Ю. Кара. Завтра была война (1987)]. **3.** Выражает высокую оценку говорящим некоторого объекта; сопровождающие жесты: вскинуть руку; двинуть голову вперед; закрыть глаза; качнуть головой; мотать головой; откинуть голову; повести подбородком вбок; поднять брови; поднять глаза к небу; потереть ладони; раскинуть руки. [Виктор/ Андрей Панин] *НИИ бросил. Занимаюсь воспитанием дитя богатых родителей. Писят долларов… в день.* [Вера/ Ирина Розанова] *О!* [Виктор/ Андрей Панин] *Да. Два раза в неделю.* [П. Тодоровский. Жизнь забавами полна (2002)]. **4.** Сопровождает приветствие; сопровождающие жесты: вскинуть руку; встать; прижать руку к груди; протянуть руки к кому-л.; раскинуть руки; снять головной убор; снять очки; рукопожатие. [Привалов / Владимир Володин] [протягивает руку для рукопожатия] *Тренер по боксу / Привалов.* [Кошелев / Сергей Блинников] [пожимает Привалову руку] *О / знаю / читал!* [А. Фролов. Первая перчатка (1946)]

О³, возглас, произносится долго (чаще) или кратко (реже) с форсированным звуком в начале произнесения. **1.** Выражение боли; выражение физиологического отвращения к чему-л.; сопровождающие жесты: скривиться; сморщиться; схватиться за больное место. [Мишка Кисель/ Станислав Садальский] *Разуй глаза! Дочь это моя! Моя! Галюня! Моя! Галюнька! Моя!* [Вася сажает Киселю на лицо пчелу, которая кусает Киселя] *А! О!* [морщится, хватается рукой за лицо] *А!* [И. Добролюбов. Белые Росы (1983)]; [см. также выше пример (9б)]. **2.** Выражение интенсивности какого-л. физического действия; сопровождающие жесты: трясти кулаком. [см. выше пример (10а)]. **3.** Выражение интенсивности некоторого чувства, ощущения; сопровождающие жесты: вытянуть шею вперед, запрокинув голову; прикрыть глаза. [Андрей/ Никита Михалков] *Ой / соскучился!* [Вера/ Людмила Гурченко] *А я-то!* [Андрей целует Веру] [Вера/ Людмила Гурченко] *О!* [прикрывает глаза, потом опускает взгляд] [Андрей/ Никита Михалков] *Ну чё / пойдём?* [Э. Рязанов. Вокзал для двоих (1982)]

В заключение хотелось бы соотнести пробную словарную статью междометия *О*, предложенную в работе [Шаронов 2006] и созданную, судя по приведенным цитатам, на основе письменных источников, со словарной статьей вокального жеста *О* по нашим данным (распределением по значениям мы пренебрегли, поскольку интереснее сравнить собственно наборы смысловых параметров, а не их группировки).

Как видим, пересечение имеет место в основном в эмоциональной зоне ($О^2$). Дейктическая зона либо не отражается в письменных источниках, либо не учтена автором статьи сознательно (представляется, что скорее первое, поскольку, как мы уже писали в работе [Гришина 2008], складывается впечатление, что $О^1$, функционируя исключительно в устной речи, до сих пор не отрефлектировано ни в словарях, ни в лингвистических исследованиях). Из физиологической зоны $О^3$ в описание в статье И. А. Шаронова попал только возглас боли. Что касается «стона утомления, пресыщенности», зафиксированного в [Шаронов 2006], то, как мы уже писали выше, это значение на нашем материале пока не набирает достаточного количества примеров, чтобы его рассматривать как самостоятельное (что до «риторического обращения», то мы с самого начала вывели его из рассмотрения, см. выше). Таким образом, устные источники предоставляют, как кажется, гораздо более разнообразный и обширный материал для анализа функционирования первообразных вокальных жестов, соответственно, на выходе получаем и более разнообразный результат описания.

Таблица 6

Настоящая статья	Работа [Шаронов 2006]
указание на объект фиксация объекта говорящий обращает внимание на что-л. привлечение внимания слушающего к какому-л. объекту сопровождение условных жестов (показать большой палец, показать кукиш)	
выражение удивления	возглас заинтересованности при неожиданном восприятии, обнаружении чего-л.
выражение разочарования	
выражение насмешки	
выражение высокой оценки чего-л.	возглас удовлетворения достигнутым результатом; возглас впечатления от признака, степени чего-л.; эмоциональное восклицание при введении значимой темы
сопровождает приветствие	
выражение боли	стон боли
выражение физиологического отвращения	
выражение интенсивности физического действия выражение интенсивности чувства	
	стон утомления, пресыщенности
	риторическое обращение

Литература

1. *Борисова Е. Г.* Междометия склоняются или спрягаются? // http://www.dialog-21.ru/Archive/2004/Borisova.htm
2. *Григорьева С. А., Григорьев Н. В., Крейдлин Г. Е.* Словарь языка русских жестов. М. — Вена: 2001
3. *Гришина Е. А.* Частица вот: варианты, используемые в непринужденной речи // Инструментарий русистики: корпусные подходы (Slavica Helsingiensia 34). Хельсинки, 2008, с. 63–91 [2008]
4. *Гришина Е. А.* Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006–2008: новые результаты и перспективы (в печати) [2009]
5. *Гришина Е. А., Савчук С. О.* Корпус звучащей русской речи в составе Национального корпуса русского языка. Проект // http://www.dialog-21.ru/dialog2008/materials/html/19.htm
6. *Крейдлин Г. Е.* Невербальная семиотика. М.: 2002
7. *Кудинов М. С., Гришина Е. А.* Инструменты полуавтоматической разметки для Мультимедийного русского корпуса (МУРКО) (в настоящем сборнике) [2009]
8. *Николаева Т. М.* Функции частиц в высказывании. М., 2004 (первое издание — 1985)
9. *Протасова Е.* Эмоциональная регуляция в общении взрослого и ребенка (на примере усвоения междометий и частиц) // Эмоции в языке и речи. М., 2005, с. 161–177
10. *Розенталь Д. Э.* Русский язык. Справочник-практикум. М., 2005
11. *Учебник* Грамоты: пунктуация // http://www.gramota.ru/class/coach/punct/45_180
12. *Толковый* словарь русского языка. Под ред. Д. Н. Ушакова. Т. II. М., 1938
13. *Шаронов И. А.* О новом подходе к классификации эмоциональных междометий // http://www.dialog-21.ru/dialog2006/materials/html/Sharonov.htm [2006]
14. *Шаронов И. А.* К вопросу об универсальности и национальной специфичности междометий // Международный симпозиум «Инновации в исследованиях русского языка, литературы и культуры", Пловдив, 1–3 ноября 2006 года». http://www.russian.slavica.org/down/SBORNIK-1.doc [2006a]
15. *Шаронов И. А.* К вопросу о разграничении эмоциональных междометий и модальных частиц // http://www.dialog-21.ru/dialog2008/materials/html/87.htm [2008]

Универсальный словарь концептов[1]

Universal dictionary of concepts

Диконов В. Г. (dikonov@iitp.ru)
ИППИ РАН

Богуславский И. М. (bogus@iitp.ru)
UPM/ИППИ РАН

Москва-Мадрид, 2008

В статье представлен универсальный словарь концептов, разрабатываемый в рамках проекта по созданию семантического языка-посредника для глобального обмена информацией. Описываются основные принципы и содержание создаваемого словаря, который может стать общедоступным лингвистически нейтральным ресурсом.

1. Введение

Статья посвящена созданию нового лингвистического ресурса — Универсального словаря концептов (UDC), также именуемого словарём UNL. Он является частью более широкого международного проекта по разработке семантического языка-посредника UNL (Universal Networking Language) [2, 8]. UDC будет использоваться в качестве лексикона этого языка. Хотя словарь тесно связан с языком UNL, он имеет значительную ценность в качестве самостоятельного ресурса и может использоваться для решения различных научных и практических задач, не имеющих отношения к UNL.

2. Универсальный словарь концептов

Основной единицей языка UNL и UDC является концепт — абстрактная семантическая единица, совпадающая со значениями слов, которые выделяются толковыми словарями. Например, согласно данным Merriam-Webster, Collins Cobuild, Oxford и других словарей английского языка слово *baby* может иметь пять значений:

человеческий младенец,
детёныш млекопитающего,
привлекательная девушка,
ребячливый человек,
любимая вещь, идея или проект.

Каждое из них является отдельной лексической единицей UNL и получает уникальный идентификатор — универсальное слово (UW). Обычно для каждого концепта имеется только одно UW.

Словарь не допускает омонимии, то есть ситуации, когда одно UW применялось бы для обозначения разных концептов.

Все концепты заимствуются из естественных языков, а не создаются искусственно. Существование каждого концепта должно быть подкреплено лексикографическими данными какого-нибудь естественного языка или практической необходимостью, например выразить абстрактное грамматическое значение или ввести нетерминальный символ для организации концептов в словаре.

Универсальный словарь концептов стремится включить в себя концептуальные лексиконы всех естественных языков и установить между ними взаимные связи. Если в словаре недостает нужного концепта для описания полисемии слова естественного языка, то следует добавить его, создав новое UW, и определить его связи с другими концептами. Также следует отметить, что каждый концепт имеет свой определённый набор семантических валентностей.

[1] Авторы благодарны РФФИ за частичную финансовую поддержку данной работы (гранты 08-06-00367 и 08-06-00344). В основе данного материала лежит доклад [3] на семинаре проекта MONDILEX (http://www.mondilex.org) «Lexicographic Tools and Techniques» в Москве (на английском языке).

3. Структура словаря

Универсальный словарь концептов должен включать в себя три основных компонента:

1. список концептов, обычно называемый словарём UW;
2. сеть связей между концептами, известная как база знаний UNL (UNLKB)[2];
3. локальные словари, которые связывают концепты со словами различных естественных языков.

3.1. Список концептов

Список концептов включает в себя все имеющиеся в словаре и используемые в языке UNL концепты и существует в виде перечня UW. Различий между UW для полученных из разных языков концептов не проводится. **Все концепты равноправны как отдельные лексические единицы** UNL и включаются в единый список. Вместе с тем, словарь позволит установить источник появления каждого концепта и языки, в которых он имеет прямое лексическое выражение.

Согласно общему принципу каждый концепт должен быть представлен только одним UW. Однако, едва ли возможно полностью избежать ситуаций, когда создаётся несколько разных UW для одного и того же концепта. Такое может происходить по техническим и организационным причинам в децентрализованном сообществе. Словарь должен иметь средства для разрешения подобных коллизий.

В простейшем из случаев уже существующее UW изменяется для того, чтобы исправить ошибку, обеспечить лучшее описание концепта или дополнить UW недостающей информацией. Прежнюю версию UW нельзя удалить немедленно, потому что она может быть использована в существующих UNL-текстах (или на неё могут ссылаться другие лингвистические ресурсы). Простое удаление сделало бы такие тексты несовместимыми со словарём. Словарь должен иметь механизм хранения истории изменения каждого UW, позволяющий отслеживать каждую зарегистрированную версию UW и не допускать повторного введения устаревших UW в словарь.

Ещё одним источником нескольких UW для обозначения одного концепта является сама природа человеческого языка и процессов категоризации. Каждый естественный язык содержит определённое количество полных синонимов, которые могут со временем разойтись в своём значении, например *everyone* и *everybody* в английском языке. Составить их исчерпывающий и точный список чрезвычайно трудно. В результате, на основе таких слов неизбежно будут возникать дополнительные UW для концептов, которые уже имеют своё UW.

Оба процесса создают группы UW, напоминающие синсеты в ресурсах семейства Wordnet [6]. Такие группы следует выделять среди массы синонимов, рассматриваемых как сходные, но разные концепты.

3.2. Семантическая сеть

Концепты образуют семантическую сеть, связанную отношениями гиперонимии, меронимии, конкретизации, синонимии, антонимии, ассоциации и отношений, которые описывают семантические валентности концептов. Назначение семантической сети — предоставить по возможности правильное и объективное описание связей между концептами, которые существуют в человеческом языке и сознании.

Семантическая сеть состоит из трёх различных структур, формируемых а) онтологическими отношениями, которые организуют концепты в группы согласно различным семантическим классификациям, б) семантическими отношениями, которые фиксируют подобие или различие между концептами, и в) аргументными отношениями, которые указывают набор валентностей каждого концепта и возможные классы их заполнителей.

3.2.1. Онтологическая структура

Онтологическая структура состоит из UNL-отношений **icl** (гиперонимия), **pof** (меронимия) и **iof** (конкретизация). Дополнительно могут быть использованы некоторые другие отношения, в частности **val** (значение параметра) и **fld** (область знаний).

Отношения *icl* и *iof* имеют привилегированный статус, так как хотя бы одно из них обязательно присутствует в каждом UW. С их помощью фиксируется принадлежность выражаемого UW концепта к одному или нескольким общим онтологическим классам. Каждый концепт должен быть связан со всеми классами, непосредственным представителем которых он является. Результатом является обладающая свойством иерархичности сеть онтологических отношений, встроенная в сеть из прочих отношений. Гиперонимические классы иерархичны по своей природе. Со значительной степенью упрощения они могут быть выстроены в форме дерева, хотя реальные связи между классами могут быть более сложными (см. рисунок

[2] В более ранних публикациях на связанные с UNL темы UNLKB может именоваться «Master entries dictionary» (словарь полных вариантов). Это название связано с идеей ввести развёрнутые варианты UW (Master Definition), которые бы включали в свой список ограничений все связи концепта с любыми другими концептами. В настоящее время полные варианты UW не используются, но их легко будет получить из UNLKB.

5 ниже). Словарь UDC предлагает более гибкий и реалистичный метод представления отношений между концептами, чем обычное дерево. Возникающая в результате структура оказывается гибридной. В ней совмещаются свойства дерева и сети. Цепи отношений могут разделяться и затем соединяться вновь, как показано на рисунке 1, но вместе с тем структура имеет общую исходную точку или корень.

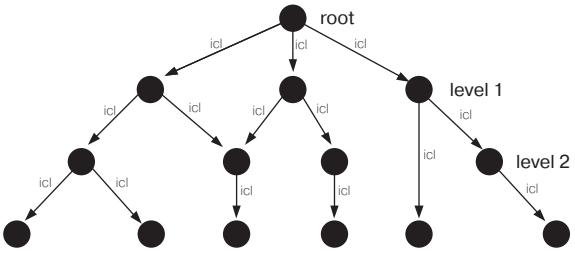


Рис. 1. Онтологическая структура

Абстрактный корневой класс называется «uw» (произвольный концепт). Он подразделяется на более узкие абстрактные классы объектов, свойств, действий, событий и т.п. Можно говорить об уровнях онтологической структуры в рамках одной цепи гиперонимов между концептом и корневым узлом, но концепт может одновременно относиться к нескольким уровням или ветвям структуры.

Онтологические отношения позволяют прослеживать соотношения объёмов понятий и находить обобщающие термины при деконверсии, если в целевом языке нет точного переводного эквивалента Например: при переводе русского слова *жениться*, которое буквально означает *«обрести жену»* и не имеет точного эквивалента в английском, следует заменить соответствующий концепт на более общий *«вступить в брак»*, который имеет прямой перевод на английский язык (см. выше UW с заглавным словом *marry*) .

3.2.2. Семантическая структура

Семантическая структура устроена иначе. Она состоит из семантических отношений **equ** (синонимия), **ant** (антонимия) и **com** (ассоциация). Отношение equ не позволяет различить полные и квази-синонимы. Поэтому его можно дополнить другим выразительным средством, позволяющими маркировать группы UW, которые обозначают один и тот же концепт. Семантические отношения связывают концепты в группы и не образуют никакой иерархии. Возникающая в результате структура, как показано на рисунке 2, является децентрализованной сетью.

В отличие от онтологической структуры, семантическая не обязательно должна быть связной. Она может состоять из многих изолированных фрагментов.

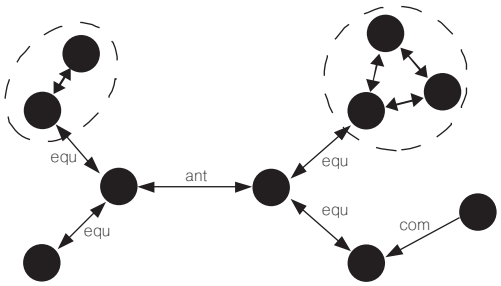


Рис. 2. Фрагмент семантической структуры

3.2.3. Аргументная структура

Аргументная структура является набором аргументных отношений, например **agt** (агент), **obj** (объект), **ptn** (партнер), **ben** (бенефициар), **plt** (место назначения), **src** (источник), **gol** (конечное состояние) и т.п. Эти отношения связывают каждый концепт с концептами абстрактных классов, представители которых обычно заполняют соответствующую отношению валентность. В большинстве случаев аргументные отношения указывают на концепты, которые принадлежат к сравнительно компактной группе наиболее абстрактных онтологических классов, расположенных близко к корню онтологической структуры (рисунок 3).

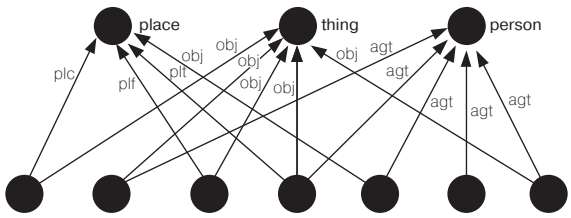


Рис. 3. Аргументная структура

Все три структуры объединяют одни и те же концепты и накладываются друг на друга. Вместе они составляют семантическую сеть UDC.

3.3. Локальные словари

Локальные словари хранят информацию о соответствии концептов и лексики конкретного естественного языка. Для каждого поддерживаемого инфраструктурой UNL языка должен существовать свой локальный словарь. Простейший локальный словарь может быть просто списком пар концептов и их переводов на естественный язык. К словам естественного языка может быть добавлена морфологическая и грамматическая информация, а также любые другие полезные сведения.

Перевод концепта на естественный язык может быть не одним словом, а словосочетанием или целой фразой. Некоторые концепты могут выражаться однословно в одном языке и словосочетаниями или аббревиатурами в другом, например *старшеклассник — senior pupil* или *важная персона — VIP*.

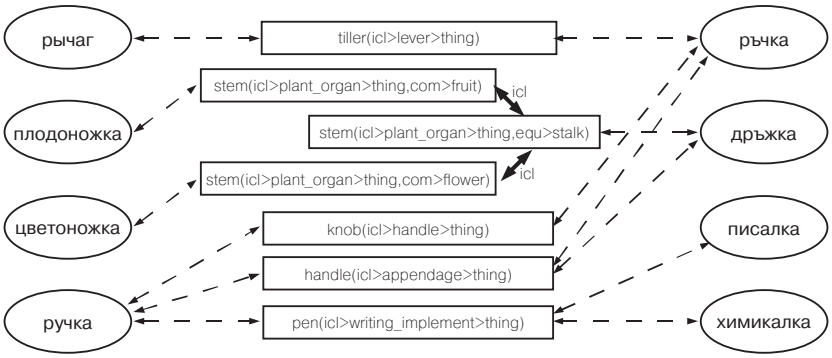


Рис. 4. Концепты и возможные связи нескольких русских и болгарских слов

Однако, перевести некоторые концепты на отдельные языки не удаётся даже описательно. Если необходимо перевести такой концепт, следует при помощи семантической сети найти ближайший более общий термин или наоборот — более узкий. На рисунке 4 представлен подобный пример. На рисунке показаны связи между русскими (слева) словами *ручка*, *рычаг*, *плодоножка*, *цветоножка* и их болгарскими эквивалентами (справа), где UW используются в качестве промежуточного средства представления значений этих слов. В русском языке нет прямого соответствия болгарскому слову *дръжка* в значении *орган растения поддерживающий цветок или плод*. Для нахождения перевода нужно проследить онтологические (icl) связи, которые ведут к концептам *плодоножка* и *цветоножка*. Кроме того, концепт *ручка для письма* имеет два возможных перевода на болгарский.

4. Универсальный словарь концептов и Wordnet

Универсальный словарь концептов похож на ресурсы семейства Wordnet во многих важных аспектах. Оба словаря используют базовое понятие концепта и определяют совпадающие типы связей между своими словарными единицами. Большая часть данных универсального словаря на момент написания статьи была получена из Princeton Wordnet [1]. Ещё больше информации, включая новые концепты и связи между ними [5], можно импортировать из других ресурсов семейства Wordnet. Однако, между UDC и ресурсами Wordnet имеются и важные различия.

4.1. Связь с естественными языками

Каждый Wordnet описывает лексику определённого естественного языка. Разные ресурсы семейства Wordnet могут быть связаны между собой при помощи межъязыковых индексов (ILI). Эти индексы описывают связи между синсетами определённых версий Princeton Wordnet (часто старых версий 1,5 или 1,6) и Wordnet-ресурсами других естественных языков. Однако, ILI-индексы играют вспомогательную роль. Только некоторые из неанглийских Wordnet-ресурсов имеют привязку к Princeton Wordnet. Кроме того, эти связи устаревают с выходом его новых версий.

Универсальный словарь концептов можно сравнивать с несколькими связанными посредством ILI словарями Wordnet, но индексы ILI связывают не все языки со всеми, а только отдельные пары языков, причем, как правило, один язык в такой паре — английский. В универсальном словаре концептов нет явного предпочтения концептуального лексикона одного из естественных языков как исчерпывающего эталона для всех прочих. Вместо этого основное внимание направлено на формирование единого общего набора концептов и установление связей между ними. Связи со словами естественных языков определяются в рамках локальных словарей. Они не будут теряться при изменениях списка концептов и структуры семантической сети.

UW состоят из заглавного слова и набора ограничителей. В качестве заглавных слов, как правило, используются слова английского языка. Однако, это не значит, что словарь использует английский как посредник или эталон при описании других языков. Он был выбран в качестве **основного источника** заглавных слов по сугубо практическим причинам — как единственный язык, знания которого можно потребовать от всех членов коллектива разработчиков. В тоже время, когда для концепта иного языка нет точного эквивалента в английском, с помощью ограничителей можно модифицировать значение английского слова и обеспечить точное описание. Кроме того, не все заглавные слова UW происходят из английского языка.

Концепты происходящие из любых языков могут быть непосредственно связаны друг с другом и служить основой для создания новых UW или ограничителей для описания любых других концептов. Например:

samovar(icl>boiler>concrete_thing,com>tea)
tula_samovar(icl>***samovar***>concrete_thing,com>tula(iof>city))

sauna*(icl>sweating_room>place,com>finnish,com>dry)*
parilka*(icl>sweating_room>place,com>russian,com>steam)*
*venik(icl>massage_tool>...com>****parilka(icl>sweating_room)****)*

Если число специфичных для других языков мира концептов будет расти, оснований для утверждения об особой роли английского в UDC будет становиться всё меньше.

4.2. Иерархические структуры

Словари семейства Wordnet организуют именные и глагольные концепты в гиперонимические деревья. Структуры такого рода удобны для поиска и анализа, однако древесная классификация в чистом виде не поддерживает частично пересекающиеся классы. Деревья могут без оговорок и упрощений применяться только для самых верхних уровней полной лингвистической онтологии. Например, в Princeton Wordnet имеются концепты *(теннисной) ракетки* и *(хоккейной) шайбы*, а также класс «*спортивный инвентарь*». При этом *ракетка* является членом класса спортивного инвентаря, а *шайба* — нет. Вместо этого она включена в класс «*дискообразные предметы*». Перемещение концепта *шайба* в класс «*спортивный инвентарь*» в чисто древесной структуре приведёт к потере информации о том, что этот предмет имеет дискообразную форму.

Универсальный словарь концептов стремится реализовать другой менее формально ограниченный подход. Базовая онтологическая структура является сетевым графом, с некоторыми чертами древесности. Наличие у концепта нескольких родительских узлов является допустимым. Это позволяет давать более подробное описание каждого отдельного концепта и реализовывать более полные и детальные классификации множеств концептов. Каждый концепт должен быть связан со всеми возможными непосредственными гиперонимами. Например, слово *суши* в Wordnet непосредственно входит в класс *блюдо* (приготовленная пища). Предположим, что мы хотим ввести дополнительную классификацию блюд по национальности (суши — блюдо японской кухни) и основному ингредиенту (суши делается из рыбы). Определить, какой из двух новых классов выше в иерархии, невозможно, потому что они соответствуют пересекающимся множествам объектов (Рисунок 5)[3].

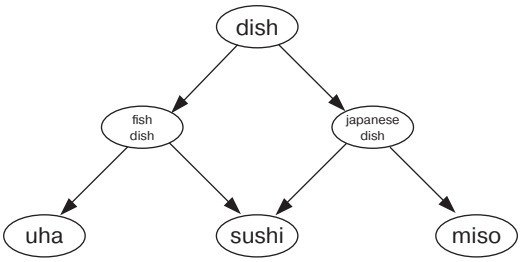


Рис. 5. Несколько родительских классов

Использование сетевой структуры вместо древесной имеет некоторые последствия. Древесная структура позволяет с полной уверенностью проследить цепочку гиперонимов каждого концепта до корня дерева даже при наличии петель, как в Wordnet. Гибридная сетевая структура допускает множество равнозначных цепочек, приводящих к различным и взаимоисключающим классам на высших уровнях иерархии гиперонимов. Это может привести к неопределённости и путанице. Например, класс «*functional thing*», одним из представителей которого является концепт *hammer*, может одновременно входить в классы «*abstract thing*» и «*concrete thing*», тем самым допуская гипотезу, что *молоток* является нематериальным объектом! Для UW эта проблема может сниматься путём добавления дополнительной связи с надлежащим вершинным классом.

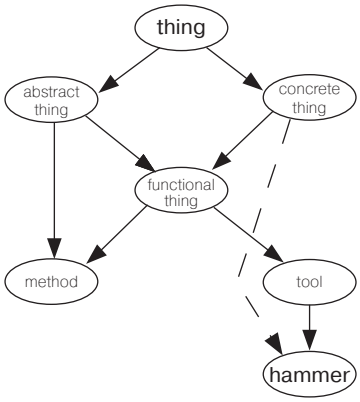


Рис. 6. Дополнительная связь с правильным вершинным классом

Согласно рисунку 6 UW концепта молоток должно выглядеть так: *hammer(icl>tool>****concrete_thing****)*. Когда известны оба конца цепи гиперонимов, становится возможным прослеживать онтологические отношения в гибридной сетевой структуре между любым концептом и соответствующим вершинным классом. Это позволяет получить полную иерархию классов.

4.3. Прочие особенности

В отличие от Wordnet универсальный словарь концептов не ограничивается определёнными частями речи. Он включает в себя полный набор концептов, соответствующих предлогам, союзам и словам со специальными грамматическими функ-

[3] Princeton Wordnet предлагает способ включить синсет одновременно в несколько классов на одном уровне иерархии, но такое расширенное описание не стало повсеместным. Например, слово *key* в значении «*килограмм наркотического вещества*» одновременно включен в находящиеся на одном уровне классы «*единица массы*» и «*единица измерения метрической системы*», а уже на следующем уровне две гиперонимические цепи объединяются в класс «*единицы измерения*».

циями, например, модальным глаголам. Это связано с тем, что в UNL нет явного деления на части речи и каких-либо ограничений круга значений, которые могут быть выражены с помощью UW.

Универсальный словарь концептов предоставляет информацию о семантических валентностях своих единиц. Валентности обозначаются отношениями языка UNL. Для каждой из них указывается наиболее общий онтологический класс, представители которого обычно заполняют эту валентность. В ресурсах семейства Wordnet могут присутствовать сведения о типичном контекстном окружении членов синсета, но единого подхода не существует. Так, Princeton Wordnet описывает модели управления глаголов (sentence frames) без какой-либо семантической классификации связей глагола с актантами, а сами актанты подразделяются только на два класса «somebody» и «something». Однако, в пока ещё неопубликованном словаре Russnet [7], который является наиболее многообещающим аналогом Wordnet для русского языка, описание семантических валентностей ожидается [6].

Некоторые словари семейства Wordnet описывают также синтаксические свойства слов, такие как часть речи, род, одушевлённость и т.п. [9], в то время как другие опираются на программы морфологического анализа и синтеза. В универсальном словаре концептов такой информации нет, так как для универсального семантического языка она не имеет смысла. Соответствующие сведения о словах естественных языков могут быть включены в локальные словари.

5. Развитие словаря

Важно, чтобы процесс развития словаря следовал принципам **разделения труда, постепенности, использования уже накопленных данных и децентрализации**. Поскольку ни один исследователь или коллектив не имеет достаточных ресурсов и знаний для решения задачи в целом, наилучшей формой организации представляется модель открытого сообщества.

Каждый раз, когда накапливается значительный пакет изменений, и нет формальных возражений против них, следует делать очередной срез словаря и публиковать его в качестве новой версии. С этого момента все участники проекта должны обновить собственные производные ресурсы для использования новой версии словаря концептов.

Универсальный словарь концептов будет опубликован под свободной лицензией, как только будет завершена работа над первой версией. Это подразумевает, что данные можно будет распространять и использовать для любых научных и личных целей. Каждый будет иметь право расширять ресурс и исправлять ошибки при условии, что все изменения будут возвращены сообществу пользователей и редакторов словаря. Качество предлагаемых к включению в словарь новых данных должно проверяться экспертами.

Литература

1. *Bekios J., Boguslavsky I., Cardeñosa J., Gallardo C.* An Efficient Method for Building Multilingual Lexical Resources // Proceedings of the Fifth International Conference Information Research and Applications i.TECH 2007, T. 1. Sofia.: 2007. C. 39–45.
2. *Boguslavsky I., Cardeñosa J., Gallardo C., Iraola L.* The UNL Initiative: An Overview // Computational Linguistics and Intelligent Text Processing.2005.
3. *Boguslavsky I. M., Dikonov V. G.* Universal Dictionary of Concepts // Proceedings of the first MONDILEX workshop «Lexicographic Tools and Techniques». M.: 2008. C.31-42.
4. *Fellbaum, C.* WordNet: An Electronic Lexical Database // MIT Press.1998.
5. *Iraola L.* Using WordNet for linking UWs to the UNL UW // International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies.Alexandria.: 2003
6. *Азарова И. В.* Схемы управления в грамматике и рамки валентностей в RussNet // http://project.phil.pu.ru/RussNet. 2005
7. *Азарова И. В., Митрофанова О. А., Синопальникова А. А.* Компьютерный тезаурус русского языка типа WordNet // Материалы конференции Диалог 2003. М.: 2003.
8. *Веб-страница* проекта UNL http://www.undl.org.
9. *Сухоногов А. М., Яблонский С. А.* Разработка русского WordNet // RCDL2004.Пущино.: 2004.

Русское *нет*, немецкое *nein*, английское *no*: сопоставительное исследование семантики на базе параллельных корпусов

Russian *net*, german *nein*, english *no*: contrastive semantic analysis with parallel corpora

Добровольский Д. О. (dm-dbrv@yandex.ru; Dmitrij.Dobrovolskij@assoc.oeaw.ac.at), **Левонтина И. Б.** (irina.levontina@mail.ru)

Институт русского языка им. В. В. Виноградова РАН

Слова *нет*, *nein* и *no* оказываются эквивалентными далеко не во всех случаях, причем расхождения в их употреблении системно возникают во вполне определенных типах контекстов. Основная причина таких расхождений в том, что русское *нет* способно отсылать не только к диктуму, но и к модусу, соотноситься с разными слоями смысла высказывания, в частности с «невысказанным».

1. Описание семантики в целом в значительной степени опирается на разного рода перефразирования, синонимические преобразования и установление смысловых различий между близкими единицами (ср. понимание смысла как инварианта синонимических преобразований в модели «Смысл-Текст»). В первую очередь, конечно, рассматриваются случаи, когда появление той или иной единицы в одних контекстах возможно, а в других — нет. Иными словами, описание значения основывается на употреблении, возможности/невозможности взаимозамен и т. п. В сопоставительных исследованиях этот принцип выглядит так: в одних случаях слово одного языка переводится на другой язык с помощью X, а в других — с помощью Y. И взаимозамены X и Y невозможны.

Однако часто различия в употреблении имеют характер предпочтений, тенденций, статистических закономерностей. Подобные различия, тем не менее, часто основаны на разнице в семантике единиц. Соотношения такого рода исследовать гораздо труднее. Их достоверное описание было практически невозможно до появления корпусов параллельных текстов. До этого лингвисты могли делать утверждения о предпочтениях, лишь основываясь на собственной семантической интуиции и отдельных наблюдениях.

Конечно, было бы наивно думать, что само наличие параллельных корпусов обеспечивает нас новым уровнем знаний. Корпус является лишь инструментом, который позволяет проверить гипотезы, возникающие в результате семантического анализа. Установление статистических закономерностей не должно подменять собою обнаружение семантических механизмов, лежащих в основе функционирования языковых единиц.

Кроме того, специфика языка как объекта исследования такова, что предполагает критическое отношение исследователя к материалу, то есть к узусу.

Наконец, реально существующие корпуса имеют множество ограничений, связанных с их объемом и составом.

2. Некоторое время назад мы обратили внимание на различия в употреблении русского слова *нет* и его немецкого эквивалента *nein*. Мы предположили, что такие различия семантически мотивированы, и исследовали их в основном путем опроса носителей языка. Многочисленные примеры несовпадений заставили нас заподозрить, что если русское *нет* имеет очень широкое значение, которое может быть связано как с диктумом, так и с модусом высказывания, то немецкое *nein* несколько у́же и в целом более тесно связано с диктумом, с пропозицией.

Чтобы проверить эту гипотезу, мы применили следующую процедуру:

1. Мы построили ряд естественных русских высказываний с *нет* в контексте императива;
2. Мы перевели их на немецкий язык, употребив *nein* в соответствии с русским *нет*;
3. Мы предъявили полученные фразы носителям немецкого языка, а затем обобщили и проанализировали их реакцию.

Мы начали с того, что взяли следующий фрагмент со словом *nein* из перевода «Волшебной горы» Томаса Манна (пер.: В. Станевич) как бы в обратном переводе на немецкий: — *Мне пора.* — ***Нет****, подожди, постой!* — *Ich muss schon gehen.* — *?****Nein****, warte!*

(в оригинале: *Halt, warte!*). Некоторые из опрошенных информантов сочли фразу *Nein, warte!* неудачной, предложив заменить ее на *Warte mal* или *Warte doch.* Другие сказали, что в разговорном языке она допустима.

Мы обнаружили определенные семантические закономерности, которые описали в работе [Добровольский, Левонтина, в печати]. Однако мы столкнулись с тем, что в тех случаях, когда различия в употреблении носят характер предпочтений, опрос информантов представляет определенные трудности. Когда различия не имеют формы явного запрета, одни информанты склонны к тому, чтобы любую фразу счесть возможной, другим же, напротив, все начинает казаться сомнительным. Кроме того, в языке часто бывает так, что отдельно взятую фразу сказать можно, но в реальном дискурсе предпочитается другая форма выражения.

Сложность здесь состоит еще и в том, что многие семантические механизмы носители языка не осознают, так же, как они не слышат, например, что оглушают согласные на конце слова (ср. соответствующие наблюдения Пинкера в [Pinker 1994]).

Сознавая ограничения, связанные с опросом информантов, мы решили продолжить исследование, основываясь на материале параллельных корпусов (мы воспользовались Корпусом Австрийской академии — ААС[1] и НКРЯ). На этом этапе мы решили добавить к сопоставлению английский язык, обратив внимание на то, что в рассматриваемом отношении английский и немецкий языки представляются похожими, хотя и не совпадающими.

Параллельные корпуса (доступные нам) — исключительно художественные тексты. Это нам кажется положительным моментом, потому что литературные переводы художественных текстов (особенно классических) выполняются, как правило, на хорошем уровне, хорошим литературным языком. С этой точки зрения богатый материал дают такие тексты, которые имеют по несколько различных профессиональных переводов. Так, например, мы рассматривали текст «Идиота» Достоевского (из корпуса ААС) и три его наиболее известных перевода (самый первый полный перевод Э.К. Разин — псевдоним Лесс (Элизабет) Керрик –, и два последних: Х. Хербота и С. Гайер).

(1) ***Нет****, Рогожин на себя клевещет; у него огромное сердце, которое может и страдать и сострадать.*

(2) ***Ja****. Rogoschin verleumdet sich selbst; er hat ein großes Herz, das zu leiden und mitzuleiden vermag. (пер.: С. Гайер)*

(3) ***Doch*** *Rogoshin verleumdete sich selbst; er hatte ein weites Herz, das sowohl leiden als auch mitleiden konnte. (пер.: Х. Хербот)*

(4) ***Nein****, Rogoshin verleumdet sich selbst: er hat ein großes, ein so großes Herz, ein Herz, das Leid und auch Mitleid zu empfinden versteht. (пер.: Э. К. Разин)*

Совершенно естественное в русской фразе *нет* у немецких переводчиков, особенно современных, явно вызвало затруднения, и буквальный перевод с помощью *nein*, во всяком случае в современном немецком языке, по каким-то причинам избегается[2].

3. В настоящей работе мы используем немецко-русский, англо-русский и русско-немецкий и русско-английский параллельные корпуса. В результате семантического анализа расхождений в употреблении *нет, nein, no* удалось установить, что они системно возникают во вполне определенных типах контекстов. Разумеется, объем статьи не предполагает в данном случае полного описания сходств и различий русского *нет*, немецкого *nein* и английского *no*. Здесь мы представим только несколько «очагов» расхождений между ними. В дальнейшем результаты можно будет проверить на более широком материале. Однако некоторые выводы о семантике русского *нет*, немецкого *nein* и анлийского *no* уже можно сделать. Мы пока оставляем в стороне вопрос о том, что у *nein* и *no* есть дополнительные значения, которых нет у русского *нет* (например, значение ‘не может быть’[3]).

Отправной точкой нашего исследования является русское слово *нет*. Поиск в корпусах осуществляется именно на это слово. При работе с параллельными корпусами обращает на себя внимание, что во многих случаях *nein* и *no* переводятся не с помощью *нет*, а как-то иначе, а *нет* далеко не всегда переводится как *nein* или *no*.

4. Наиболее бросающееся в глаза различие между тремя рассматриваемыми языками состоит в том, что они принципиально по-разному устроены с точки зрения выражения опровержения. В русском и английском языках представлена двучленная система: *да/yes* и *нет/no*, в то время как в немецком система трехчленная: также есть *ja* и *nein*, но кроме того, есть отдельное слово *doch* (нечто вроде ‘нет да’).

---

[1] Этот параллельный корпус создан в Австрийской академии наук в Вене.

[2] Здесь необходимо учитывать, что Светлана Гайер, автор последнего, наиболее точного и художественно совершенного перевода, в целом стремится переводить по возможности максимально близко к оригиналу, и то, что она сочла необходимым поменять *нет* на *да*, показательно.

[3] Ср. следующий диалог по телефону:
— Guten Tag! Hier Dima.
— Nein! Dima? Sind Sie in Deutschland?
— Здравствуйте! Это Дима.
— Не может быть! (букв.: Нет!) Дима? Вы в Германии?

При этом в английском и русском языках утвердительное и отрицательное слово распределены поразному. Английское *yes* может, в отличие от русского *да*, использоваться не только для подтверждения, но и для опровержения. Возможен и, более того, совершенно типичен диалог: — *You can't do it! — Yes, I can*. По-русски же неправильно: — *Вы этого не можете! — Да, могу*. Тут надо сказать: — *Нет, могу*. Наше *да* выражает скорее согласие с собеседником, а не удостоверение правильности содержания высказывания. И в немецком нельзя здесь употребить слово *ja* 'да'. Нужно использовать *doch* (ср. *Doch, ich kann es*).

Так, в «Вини-Пухе» Пятачок настаивал, что надпись на обломанной табличке у его дома «ПОСТОРОННИМ В» — это имя его дедушки. Дальше поанглийски так:

(5) *Christopher Robin said you couldn't be called Trespassers W, and Piglet said* ***yes, you could****, because his grandfather was.*

А в заходеровском тексте так:

(6) *Кристофер Робин сказал, что не может быть такого имени — Посторонним В., а Пятачок ответил, что* ***нет, может, нет, может,*** *потому что дедушку же так звали! (пересказ Б. Заходера)*

Естественно, в этом месте немецкого перевода читаем:

(7) *Christopher Robin sagte, man könne nicht Betreten V heißen, und Ferkel sagte,* ***doch, das könne man****, sein Großvater habe ja so geheißen. (пересказ Х. Ровольта)*

Итак, в целом можно сказать, что немецкому *doch* в русском языке регулярно соответствует *нет*, а в английском — *yes*. Анализ англо-русского корпуса показывает, что в данном типе контекстов *yes* вполне системно переводится как *нет*. Ср. примеры из «Прощай, оружие!» Эрнеста Хэмингуэя в переводе Е. Калашниковой:

(8) *— Still I would probably have been killed. — Not in this ambulance business. —* ***Yes****, even in the ambulance business.*
— Но я бы, наверно, погиб. — Ну, в санитарном отряде едва ли. — ***Нет****, даже и в санитарном отряде.*

(9) *— There is no finish to a war. —* ***Yes*** *there is.*
— Война не имеет конца. — ***Нет****, конец есть.*

(10) *— You don't like Rome? —* ***Yes****, I love Rome.*
— Вам не нравится Рим? — ***Нет****, я люблю Рим.*

Впрочем, анализ русско-английского корпуса показывает, что данному типу русских контекстов в английском не всегда соответствуют именно фразы с *yes*. Очень типичен здесь и вариант с выносом вспомогательного глагола, иногда с участием других средств смыслового подчеркивания. Ср. контекст из «Мертвых душ» (пер.: Д. Дж. Хогарт):

(11) *— Ты, однако ж, не сделал того, что я тебе говорил, — сказал Ноздрев, обратившись к Порфирию и рассматривая брюхо щенка, — и не подумал вычесать его? —* ***Нет****, я его вычесывал.*
— I can see that you haven't done what I told you to do, he continued to Porphyri after an inspection of the animal's belly — you have quite forgotten to brush him. — I ***DID*** *brush him, protested Porphyri.*

Как видно из приведенных примеров, у русского слова *нет* обнаруживаются удивительно глубокие отличия от английского, а также и немецкого (см. подробнее в работе [Добровольский, Левонтина, в печати]) аналогов. Эти различия можно описывать в терминах сферы действия. Несколько огрубляя, можно сказать, что английские *yes* и *no* ориентированы в первую очередь на внутреннюю сферу действия, то есть на пропозициональное содержание самого высказывания, в котором они находятся. В то же время русские *да* и *нет* ориентированы скорее на внешнюю сферу действия, то есть выражают согласие или несогласие с тем, что говорилось или даже подразумевалось в других высказываниях.

Если бы дело было только в примерах такого типа, можно было бы считать, что мы имеем дело с некоторым регулярным различием, которое можно описать одним простым правилом. Однако есть много других примеров несовпадения употребления слова *нет* с английскими и немецкими эквивалентами, которые, будучи даже не столь очевидными в каждом отдельном случае, в своей совокупности доказывают, что речь действительно должна идти о фундаментальных различиях в семантике, которые и лежат в основе этих расхождений.

Так, для русского языка очень характерна ситуация, когда в споре оба говорящих начинают свои реплики с *нет*, независимо от того, отрицается ли что-либо в данной реплике (естественно, в английском или немецком переводе эти *нет* чаще всего опускаются); ср.:

(12) *—* ***Нет****, ты не можешь отказаться, — говорил Ноздрев, горячась, — игра начата! — Я имею право отказаться, потому что ты не так играешь, как прилично честному человеку.*
— ***Нет****, врешь, ты этого не можешь сказать!*
— ***Нет****, брат, сам ты врешь! (Н. В. Гоголь. Мертвые души)*

*— **But** you can't refuse to, — said Nozdrev, growing heated, you see, the game has begun.*
— Nevertheless, I have a right not to continue it, seeing that you are not playing as an honest man should do. — You are lying — you cannot truthfully say that. — Tis you who are lying.[4]

Другая важная особенность русского *нет* состоит в том, что оно часто выражает, так сказать, ответ на невысказанное. Здесь возможны самые разные виды смысловых связей, которые мы сейчас рассмотреть не можем. Некоторые типы подобных употреблений мы описали в статье [Добровольский, Левонтина, в печати], в связи со случаями несоответствия в употреблении русского *нет* и немецкого *nein*.

Ср. следующий пример из «Мертвых душ», где *нет* используется для понижения иллокутивной силы, значимости вопроса:

*(13) — Извинительней сходить в какое-нибудь непристойное место, чем к нему. — **Нет**, я спросил не для каких-либо, а потому только, что интересуюсь познанием всякого рода мест, — отвечал на это Чичиков.*
— A man had far better go to hell than to Plushkin's.
— Quite so, — responded Chichikov, my only reason for asking you is that it interests me to become acquainted with any and every sort of locality.

Чичиков спешит оправдаться, опасаясь, что Собакевич заподозрит его в чрезмерной заинтересованности в информации о Плюшкине, и *нет* служит для того, чтобы заранее снять все возможные подозрения.

Чрезвычайно характерна для русского языка и ситуация, когда *нет* относится к собственным колебаниям говорящего, о которых он до этого не упоминал (*Нет, я все-таки скажу…*), и в этом случае ни ни *по*, ни *nein* невозможны. Ср. следующий пример:

*(14) Письмо начиналось очень решительно, именно так: «**Нет**, я должна к тебе писать»! (Н. В. Гоголь. Мертвые души)*
*Beginning abruptly with the words «I **MUST** write to you»!*

Наиболее экзотическую с точки зрения немецкого и английского языка группу примеров употреблений русского *нет* составляют характерные для русского непринужденного дискурса контексты, где *нет* вообще никак не связано с идеей отрицания (*Нет, ну что за молодец!; Нет, ты просто гений!*). В уже цитированной совместной работе мы описывали эксперимент: русские фразы с *нет* были записаны в реальном диалоге и переведены на немецкий, ср:

*(15) Почему ты не сдал работу вовремя? **Нет**, это просто дурь какая-то!*
*Warum hast du die Arbeit nicht rechtzeitig abgegeben? ?**Nein**, das ist doch blöd!*

Мы также обнаружили, что в русских переводах немецких художественных текстов *нет* очень часто появляется в таких случаях — при том что в немецком оригинале никакого отрицания не было. Ср. примеры из «Волшебной горы» Томаса Манна в переводе Веры Станевич:

(16) Zu einem Sterbenden! Das ist doch stark!
*Это умирающему-то! **Нет**, это уж слишком!*

(17) «Sie war unter die Decke gekrochen!» sagte Joachim. «Stelle dir meine Empfindungen vor!»
*— Она заползла под одеяло! — сказал Иоахим. — **Нет**, ты представь, что я испытывал!*

(18) Es ist etwas Bedenkliches um die Musik, meine Herren.
***Нет**, господа, в музыке есть что-то подозрительное.*

Во всех этих примерах русское *нет* взаимодействует не с пропозицией, а с представлениями говорящего о тех или иных свойствах ситуации в целом. Иными словами, отрицается не сказанное, а «невысказанное». Так, говоря *Нет, это уж слишком!*, мы, естественно, не даем отрицательный ответ на поставленный вопрос (никто ни о чем не спрашивал), а выражаем свое неприятие обсуждаемого положения вещей. Сходным образом, во многих ситуациях *Нет, ты представь себе!* звучит по-русски естественнее, чем просто *Ты представь себе!* Говорящий как бы возражает против самой возможности того, что адресат не проявит к его словам должного сочувствия.

Вообще, случаи, в которых переводчик добавляет слово, которого не было в оригинале, очень показательны. Видимо, хороший переводчик идет на это, только если ему кажется, что без такого добавления высказывание будет звучать по-русски неестественно или бессвязно.

В указанном отношении с русскими переводами английских текстов дело обстоит так же. Ср. контексты из «Прощай, оружие!»:

(19) «He was really a lovely horse,» Catherine said.
*«**Нет**, правда, чудесная лошадь», сказала Кэтрин.*

(20) I tell you this war is a bad thing.
***Нет**, в самом деле, скверная штука война.*

[4] Английское *but* довольно регулярно соответствует русскому *нет* в функции «ответа на невысказанное».

(21) — He's just a big-hearted joker. — Really he was very nice.
*— Он шутник, твой доктор. — **Нет**, в самом деле, он очень славный.*

Надо заметить, что для русского дискурса вообще очень характерны сочетания типа *нет, правда/нет, точно/нет, конечно* и даже *нет, да (Нет, да, я согласен)*, в то время как *nein, stimmt/nein, ja/no, sure/no, yes* явно нетипичны.

Русское *нет* очень сильно отличается от немецкого и английского аналогов в контексте побудительных высказываний (ср. подробнее в цитированной совместной статье). Особенно это заметно в ситуации повторного требования.

*(22) — Да зачем, я и так вижу: доброй породы! — отвечал Чичиков. — **Нет**, возьми-ка нарочно, пощупай уши! (Н. В. Гоголь. Мертвые души)*
*— Why should I? — without doing that, I can see that he is well-bred. — **Nevertheless**, catch hold of his ears and feel them.*

Значение этой дискурсивной частицы *нет* в других языках распределяется между несколькими частицами и союзами: *aber, doch, sondern, but, nevertheless, however*.

Вообще повторное побуждение в русском языке практически обязательно маркируется — часто при помощи частиц *ну, да, же* и др.; см. [Левонтина 1999]. В случае эксплицитного отказа повторное требование часто начинается с *нет*, а в немецком и английском это обычно недопустимо. Ср. следующий контекст, содержащий повторное побуждение: после полученного отказа говорящий повторяет свое требование, не желая признавать права адресата на отказ.

*(23) «Kommen Sie mit», sagte K., «zeigen Sie mir den Weg, ich werde ihn verfehlen, es sind hier so viele Wege.» «Es ist der einzige Weg», sagte der Gerichtsdiener nun schon vorwurfsvoll, «ich kann nicht wieder mit Ihnen zurückgehen, ich muss doch meine Meldung vorbringen und habe schon viel Zeit durch Sie versäumt». — «**Kommen Sie mit!**» wiederholte K. jetzt schärfer, als habe er endlich den Gerichtsdiener auf einer Unwahrheit ertappt. (Ф. Кафка. Процесс)*
*— Пойдемте со мной, — сказал К, — покажите мне дорогу, не то я запутаюсь, здесь столько входов и выходов. — Нет, это единственный выход, — уже с упреком сказал служитель, — а вернуться с вами я не могу, мне еще надо передать поручение, я и так потерял с вами уйму времени. — **Нет пойдемте!** — уже резче сказал К, словно наконец уличил служителя во лжи. (пер.: Р. Райт-Ковалева)*

Этот тип контекстов также иллюстрирует те свойства русского *нет*, которые отличают его от английского и немецкого аналогов — такие, как ориентация на внешний контекст (то есть на аспекты ситуации, внешние по отношению к пропозитивному содержанию высказывания), способность отсылать не только к диктуму, но и к модусу, соотнесенность с разными слоями смысла высказывания, в частности с «невысказанным», то есть всевозможными предполагаемыми следствиями и импликатурами дискурса. Все эти особенности слова *нет* хорошо согласуется с характерной для русского языка структурой дискурса: его повышенной связностью, особенно типичной для диалога, постоянной апелляцией к разным слоям высказывания и вниманием к взаимоотношениям с собеседником, которые говорящий непрерывно выстраивает в ходе диалога.

Литература

1. *Добровольский Д. О., Левонтина И. Б.* 500 способов сказать «нет» (русско-немецкие соответствия) // Логический анализ языка. Ассерция и негация. М.: Индрик, в печати.
2. *Левонтина И. Б.* Стратегии уговаривания: частицы в повторных просьбах // Язык. Культура, Гуманитарное знание. Научное наследие Г. О. Винокура и современность. М.: Научный мир, 1999. С. 188–201.
3. *Pinker S.* The language instinct. How the mind creates **language.** New York: HarperCollins, 1994.

Обработка естественно-языковых запросов к поисковой машине на основе их лингвистического анализа

Natural language query processing for search engine based on linguistic analysis

Ермаков А. Е. (ermakov@metric.ru), **Плешко В. В.** (vp@rco.ru)

ООО «ЭР СИ О» (www.rco.ru), Москва

Описывается новый способ преобразования запросов на естественном языке в языки запросов поисковых машин, основанный на машинном анализе синтаксических связей между словами и их отображении на соответствующие операторы языка поисковой машины с максимальным сохранением смысла исходного запроса.

1. Введение

Языки запросов современных поисковых машин, используемых для поиска текстов в базах данных или полнотекстовых хранилищах документов, разрешают задавать различные ограничения на искомые комбинации слов в тексте, определяя обязательность или необязательность присутствия тех или иных слов, допустимое расстояние между словами и порядок их следования в тексте, а также позволяя искать слова во всех грамматических формах, что дает возможность в принципе формулировать очень сложные запросы, точно и полно описывая возможные способы выражения в тексте искомого смысла. Такие возможности поддерживают, к примеру, поисковые машины в СУБД Oracle и СУБД Microsoft SQL Server, поисковые машины компаний Google и Яндекс.

Проблема создания хороших информационно-поисковых систем на базе поисковых машин заключается в том, что пользователь системы часто желает формулировать свой запрос в виде простого набора слов, словосочетаний или фразы на естественном языке, ожидая от системы понимания хотя бы элементарных способов того, в какой форме соответствующий смысл может быть выражен в тексте. Так, большинство поисковых запросов, по которым пользователь может найти требуемые тексты, состоят более чем из одного слова. Здесь начинаются проблемы — как обработать запрос из нескольких слов, в каком виде транслировать его поисковой машине?

Во-первых, не ясно, как искать вхождение слов в документ — как цепочку подряд следующих слов (используя оператор языка запросов для поиска «по фразе»), как набор близко расположенных слов (используя оператор типа «рядом»), просто как набор встречающихся совместно в одном документе слов (используя оператор И), или как набор таких слов, из которых лишь некоторые должны обязательно встретиться в документе (используя оператор ИЛИ).

Во-вторых, не ясно, как расширять слова запроса грамматическими формами. Если искать все формы для каждого слова запроса, то точность поиска оказывается не высока по двум причинам. Во-первых, без учета грамматических связей слов запроса нет возможности разрешить омонимию: например, при обработке запросов *решение суда* и *грузовые суда* все варианты словоформы *суда* следует строить в первом случае от слова *суд*, а во втором — от слова *судно*. Во-вторых, при поиске словосочетаний допустимыми являются не все грамматические формы слов: например, при обработке запроса *президент России* слово *президент* стоит искать во всех вариантах, а слово *Россия* следует искать только в заданной форме родительного падежа, иначе можно найти фрагменты текста следующего вида: *к встрече американского президента Россия готовилась заблаговременно*. Кроме того, поиск слова во всех грамматических формах обычно увеличивает нагрузку на поисковую машину.

В итоге, обычно информационно-поисковые системы инициируют поиск всех слов запроса по ИЛИ либо по И, допуская каждое слово во всех грамматических формах, используя ту особенность поисковых машин, что те обычно ранжируют найденные документы по релевантности таким образом, что первыми в результатах поиска выдаются документы, со-

держащие наибольшее количество слов из запроса, в которых эти слова расположены наиболее близко в тексте. Поскольку при этом никак не учитывается связанность слов в запросе, результаты поиска могут содержать ошибки, вызванные случайной близостью в тексте не связанных по смыслу слов. Так, например, все слова словосочетаний *президент России* и *российский президент* целесообразно искать в тексте только рядом, поскольку большинство других случаев их близкого положения будут соответствовать совершенно иным смыслам. Напротив, слова словосочетания *зарегистрировать изобретение* могут находиться в тексте рядом в любом порядке, будучи разделенными другими словами, например: *изобретение способа преобразования запросов, которое так и не было зарегистрировано*. Помимо невысокой точности, избыточный поиск по ИЛИ обычно также увеличивает нагрузку на поисковую машину.

Для повышения точности поиска обычно используют информацию о частоте встречаемости слов запроса в найденных документах и во всей коллекции, по которой ведется поиск [1]. Наиболее ярким примером отечественной системы, воплотившей данный подход, является Галактика-Zoom [3].

Для повышения точности поиска в академических коллективах разрабатываются методы, основанные на предварительном лингвистическом разборе текстов [4–6]. Для эффективного практического применения такие методы требуют сохранения полученных описаний грамматической или семантической структуры в специальном индексе, который затем должен использоваться при поиске для сравнения со структурой запроса, получаемой лингвистическим анализатором. При доступных сегодня вычислительных мощностях данный подход не является промышленным, так как требует, во-первых, значительных вычислительных затрат для лингвистического анализа индексируемой коллекции текстов, а во-вторых, разработки специализированной поисковой машины, вследствие чего, в частности, не может быть универсально применен к любой базе данных. Поэтому практические попытки применения лингвистических методов к искомым текстам ограничиваются созданием мета-поисковых систем, которые лишь пытаются переупорядочить документы, найденные другой информационно-поисковой системой, на основании анализа небольших фрагментов текста, выданных поисковиком в качестве рефератов по запросу.

Другие известные анонсированные методы связаны с применением тезаурусов, например [2,7,8], и предназначены для повышения не точности, а полноты поиска (за исключением случаев, когда тезаурус используется для снятия омонимии).

Полезный результат, достигаемый при использовании описываемого способа поиска, заключается в повышении точности поиска при сохранении его высокой полноты, а также в снижении нагрузки на поисковую машину.

Настоящий способ поиска основан на использовании лингвистических знаний о грамматике того естественного языка, на котором формулируется поисковый запрос, и предлагает использовать синтаксические связи между словами поискового запроса для выбора оптимального выражения на языке запросов поисковой машины, а также, при отсутствии результата поиска документов по этому выражению, для формирования последовательности поисковых выражений с уменьшающейся степенью строгости поисковых ограничений и с максимально возможным сохранением смысла исходного запроса, что обеспечивает последовательное повышение полноты поиска с минимальной потерей точности. Соответствие операторов языка запросов синтаксическим связям между словами устанавливается на основании того принципа, что более сильно связанные в запросе слова должны искаться на более близком расстоянии в тексте и с более жесткими ограничениями на допустимые грамматические формы.

2. Базовый способ преобразования запроса

Конкретное соответствие типов синтаксических связей между словами поискового запроса и операторов языка запросов поисковой машины может быть различным, поскольку зависит от:

а) обрабатываемого естественного языка (русский, английский и др),
б) используемого синтаксического анализатора и типов выделяемых им синтаксических связей,
в) используемой поисковой машины и поддерживаемых ей операторов языка запросов.

В наиболее сложном случае возможно отображение связей между словами на операторы из следующего множества: И, РЯДОМ, РЯДОМ_УПОРЯДОЧЕННО, ФРАЗА, ВО_ВСЕХ_ФОРМАХ.

Ниже в таблице 1 приведены примеры установления соответствия между основными синтаксическими связями в русском языке и перечисленными операторами. В этих и следующих примерах будет использована нотация, абстрагированная от формальных особенностей языка какой-либо конкретной поисковой машины, но позволяющая записать поисковые выражения с использованием всех общепринятых операторов: — следующие друг за другом слова должны искаться в тексте «как фраза» (в соседних слово-местах); — оператор *and* означает, что слова должны встречаться в одном тексте в любых местах; — операторы *near* и *near_ord* означают, что слова должны находиться в тексте на небольшом расстоянии друг от друга, при-

Таблица 1. Пример установления соответствия между основными синтаксическими связями в русском языке и основными операторами поисковых машин

Тип синтаксической связи	Оператор запроса	Форма подчиненного слова	Пример фрагмента запроса
прилагательное или причастие в составе именной группы (*рыжий <- конь, пишущее -> устройство*)	ФРАЗА	как у главного	`(m:РЫЖИЙ m:КОНЬ)` `(m:ПИШУЩЕЕ m:УСТРОЙСТВО)`
приложение (*царь -> Иван*)	ФРАЗА	как у главного	`(m:ЦАРЬ m:ИВАН)`
генитив (*отношение -> принадлежности*)	ФРАЗА	исходная	`(m:ОТНОШЕНИЕ ПРИНАДЛЕЖНОСТИ)`
деепричастие (*двигаться, -> гремя*)	РЯДОМ	любая	`ДВИГАТЬСЯ near ГРЕМЕТЬ`
инфинитив (*попытка -> заставить -> работать*)	ФРАЗА	исходная	`(m:ПОПЫТКА ЗАСТАВИТЬ РАБОТАТЬ)`
предлог (*под <- капотом*)	ФРАЗА	исходная	`(ПОД КАПОТОМ)`
аргумент предиката при глаголе (*клиент <- арендует -> у предприятия)*	РЯДОМ	исходная, если с предлогом, иначе — любая	`m:КЛИЕНТ near m:АРЕНДОВАТЬ near (У m:ПРЕДПРИЯТИЕ)`
аргумент предиката при существительном (*аренда -> земли, договор -> (о) разоружении*).	РЯДОМ_УПОРЯДОЧЕННО	исходная, возможно с предлогом	`m:АРЕНДА near_ord ЗЕМЛИ` `m:ДОГОВОР near_ord (О РАЗОРУЖЕНИИ)`
обстоятельство (*отчет -> (при) упрощенке*)	И	исходная с предлогом	`m:ОТЧЕТ and (ПРИ УПРОЩЕНКЕ)`
синтаксически ничему не подчиненные слова (*платежи перечисление*)	И	любая	m:ПЛАТЕЖ and m:ПЕРЕЧИСЛЕНИЕ

чем второй маркер дополнительно указывает, что порядок слов в тексте должен соответствовать порядку слов в поисковом выражении. Соответствующие операторы в различных поисковых машинах имеют свои особенности реализации; — оператор *m:* означает, что соответствующее слово должно искаться во всех грамматических формах (иначе слово ищется только в указанной форме). — оператор = означает эквивалентность указанных слов при поиске

Рассмотрим пример конструирования запроса к поисковой машине для запроса *авансовые платежи налог на прибыль предприятий*.

В терминах приведенной выше нотации по этому запросу может быть сконструировано следующее выражение:

(m:АВАНСОВЫЙ m:ПЛАТЕЖ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ ПРЕДПРИЯТИЙ))).

Соответствующее выражение на языке запросов, поддерживаемом в СУБД Oracle, где оператор *near* требует указания максимального расстояния между словами, (например, 5) и флаг учёта порядка слов (TRUE, что реализует оператор *near_ord*), будет выглядеть так:

($АВАНСОВЫЙ $ПЛАТЕЖ) and (near(($НАЛОГ, НА, ПРИБЫЛЬ ПРЕДПРИЯТИЙ), 5, TRUE)).

Соответствующее выражение на языке запросов, поддерживаемом в СУБД MS SQL Server, где не существует оператора, эквивалентного *near_ord* (используется только *near*), а оператор, эквивалентный *m:*, может применяться только ко всем словам фразы, будет выглядеть так:

FORMSOF(INFLECTIONAL, "АВАНСОВЫЙ ПЛАТЕЖ") AND (FORMSOF(INFLECTIONAL, НАЛОГ) NEAR НА NEAR "ПРИБЫЛЬ ПРЕДПРИЯТИЙ").

3. Построение последовательности поисковых выражений

В информационно-поисковой системе, если полнота поиска по исходному запросу оказалась неудовлетворительной (найдено недостаточно документов), запрос может быть преобразован с некоторым ослаблением поисковых ограничений и вновь передан поисковой машине, что может в итоге привести к построению целой последовательности поисковых выражений до тех пор, пока не будет достигнута требуемая полнота поиска. Ниже описываются те преобразования, которые ослабляют поисковые ограничения, одновременно пытаясь максимально сохранить смысл исходного запроса.

Во-первых, при конструировании поискового выражения возможно из запроса на естественном языке исключать слова, синтаксически или семантически подчиненные другим словам.

Так, при исключении зависимых слов *авансовый* и *предприятие* может быть сконструировано выражение *(m:ПЛАТЕЖ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ)))*. А при исключении из исходного запроса зависимого слова *на* может быть получено *(m:АВАНСОВЫЙ m:ПЛАТЕЖ) and (m:НАЛОГ near (m:ПРИБЫЛЬ ПРЕДПРИЯТИЙ))*.

Однако, часто именно синтаксически зависимые слова наиболее точно определяют предмет поиска и их удаление делает запрос бессмысленным, например, *общий принцип изменения и расторжения договора*. Поэтому, критерий синтаксической зависимости при исключении слов из запроса является менее приоритетным, чем следующий критерий, который учитывает лексическую значимость слов.

Во-вторых, при конструировании поискового выражения возможно из запроса исключать слова, входящие в заданный стоп-словарь, с сохранением грамматики фразы. Например, существительное исключается вместе с согласованными определениями.

Так, при конструировании выражения для запроса *высокий рост детской смертности в Никарагуа* можно исключить общеупотребимое слово *рост* вместе со своим определением *высокий*. В результате получим следующее выражение: *(m:ДЕТСКИЙ m:СМЕРТНОСТЬ) and m:НИКАРАГУА*.

И напротив, возможно из поискового запроса исключать слова, не входящие в заданный словарь терминов предметной области. Так, при исключении из запроса *система современного налогообложения в малом бизнесе* слов *система* и *современный*, не входящих в словарь юридических терминов, получим следующее выражение: *m:НАЛОГООБЛОЖЕНИЕ and (m:МАЛЫЙ m:БИЗНЕС)*.

В-третьих, при конструировании поискового выражения возможно любое из слов или словосочетаний запроса заменить на слово или словосочетание из словаря синонимов, гипонимов или других связанных по смыслу слов, объединяемых в поисковом запросе оператором =, *or* или другим эквивалентным по смыслу оператором.

Так, для запроса *рост доходов населения в России* соответствующее выражение при расширении синонимами будет выглядеть следующим образом:

(m:(РОСТ=ПОВЫШЕНИЕ=ПОДЪЕМ) ДОХОДОВ=ПРИБЫЛИ НАСЕЛЕНИЯ=ЖИТЕЛЕЙ) and m:(РОССИЯ=РФ).

В-четвертых, если в результате поиска по любому из приведенных способов найдено недостаточно документов, возможно для каждого поискового выражения строить несколько похожих выражений с меньшей степенью строгости путем замены одних операторов поиска на другие *(ФРАЗА -> РЯДОМ -> И -> ИЛИ)*.

Рассмотрим пример построения последовательности поисковых выражений для приведенного выше запроса *авансовые платежи налог на прибыль предприятий*. Исходное выражение

(m:АВАНСОВЫЙ m:ПЛАТЕЖ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ ПРЕДПРИЯТИЙ)))

может быть далее преобразовано в следующие последовательности выражений:

1.1. удаление зависимых слов: *(m:ПЛАТЕЖ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ)))*;
1.2. расширение синонимами: *(m:ПЛАТЕЖ=ПЛАТА=ПЛАТИТЬ=ЗАПЛАТИТЬ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ=ДОХОД)))*;

или

2.1. замена части операторов на "более слабые": *(m:АВАНСОВЫЙ m:ПЛАТЕЖ) and (m:НАЛОГ near (m:ПРИБЫЛЬ ПРЕДПРИЯТИЙ))*;
2.2. замена части операторов на "более слабые": *(m:АВАНСОВЫЙ m:ПЛАТЕЖ) or (m:НАЛОГ and (m:ПРИБЫЛЬ ПРЕДПРИЯТИЙ))*;
2.3. удаление зависимых слов: *(m:ПЛАТЕЖ) or (m:НАЛОГ and m:ПРИБЫЛЬ)*.

4. Заключение

Описанный способ обработки поисковых запросов на естественном языке был успешно апробирован в одном из проектов компании «ЭР СИ О» (http://www.rco.ru) на базе специализированной поисковой машины заказчика и показал заметное повышение точности поиска на многословных запросах. Дальнейшее практическое исследование, в том числе сравнение качества поиска с другими информационно-поисковыми системами, планируется провести в рамках очередного семинара РОМИП (http://romip.narod.ru/). На изобретение подана и зарегистрирована патентная заявка «Способ выполнения поиска в компьютерной системе» №2008138379 от 26.09.2008.

Литература

1. *C. J. Van Rijsbergen.* Information Retrieval, 2nd edition. — Butterworths, London, 1979.
2. *Солтон Дж.* Динамические библиотечно-информационные системы. — Пер. с англ. — М.: Мир, 1979. — 558 с.
3. *Антонов А. В., Курзинер Е. С.* Новые возможности поисково-аналитической системы «Галактика-Zoom» (ранжирование документов по значимости). // Материалы конференции «Диалог-2003». (http://www.dialog-21.ru/archive/2003/Antonov.htm)
4. *Осипов Г. С.* и др. Проблемы обеспечения точности и полноты поиска: Пути решения в интеллектуальной мета-поисковой системе «Сириус». // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2005. — Москва, Наука, 2005. — С. 390–394.
5. *Тихомиров И. А.* Вопросно-ответный поиск в интеллектуальной поисковой системе Exactus // Российский семинар по Оценке Методов Информационного Поиска. Труды четвертого российского семинара РОМИП'2006. — Санкт-Петербург: НУ ЦСИ, 2006, — С. 80–85.
6. *Окатьев В. В., Баркалов К. А.* Патент № 2320005 на изобретение «Способ поиска информации», опубликовано 20.03.2008, ООО «Диктум».
7. *Лукашевич Н. В., Добров Б. В.* Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002 / Под ред. А. С. Нариньяни — М.: Наука, 2002. Т. 2. С. 338–346.
8. *Брин С., Гомес Б., Тонг С.* Патент № 2324220 на изобретение «Оснащение пользовательского интерфейса расширением поисковых запросов», опубликовано 10.05.2008, Гугл Инк. (US).

О понятии семантического перехода[1]

On the notion of semantic shift

Зализняк Анна А. (anna-zalizniak@mtu-net.ru)

Институт языкознания РАН

В статье уточняется понятие семантического перехода как объекта семантической типологии и единицы «Каталога семантических переходов в языках мира», создаваемого в настоящее время коллективом авторов в Институте языкознания РАН.

В настоящей статье я изложу некоторые соображения относительно понятия ***семантического перехода*** как объекта семантической типологии[2] (обоснованного в [Зализняк 2001]), возникшие в ходе работы над «Каталогом семантических переходов в языках мира», которая ведется на протяжении последних семи лет коллективом авторов (М. С. Булах, Д.С.Ганенков, И. А. Грунтов, Т. А. Майсак, М. М. Руссо) под моим руководством. О концепции Каталога и структуре базы данных, в формате которой он реализуется, см. [Зализняк 2006: 392–402, Грунтов 2007, Zalizniak 2008, Zalizniak, Bulakh, Ganenkov, Gruntov, Maisak, Russo (in print)].

Под «семантическим переходом» понимается факт совмещения, в пределах одного слова, двух разных значений — в форме либо синхронной полисемии, либо диахронической семантической эволюции. Так, например, значения ‘вести счет’ и ‘иметь мнение’ выражаются синхронно русским глаголом *считать*; значения ‘схватить’ и ‘понять’ сосуществуют диахронически в русском глаголе *понять*. Понятие семантического перехода в настоящий момент оказалось, по-видимому, одним из центральных для семантики. Это обусловлено, с одной стороны, развитием диахронической когнитивной семантики (см. [Traugott, Dasher 2002, Sweetser 1990, Haser 2000, Bybee *et al.* 1994, Blank 2000; Koch 2001, 2004] и др.), объектом которой является семантическая эволюция, т.е. семантические переходы в диахронии; с другой — это бурное развитие грамматической типологии, в частности, в области теории грамматикализации, объектом которой являются переходы лексических значений в грамматические (см. [Bybee *et al*. 1994, Heine, Kuteva 2002, Плунгян 1998, Майсак 2005]). С третьей стороны, это работы в области систематизации описания лексической многозначности (Падучева 2004, Кустова 2004, Розина 2005, Толстая 2008 и др.): обнаруживаемые здесь устойчивые соотношения и отношения семантической деривации между значениями и классами значений также естественно концептуализуются как семантические переходы. Понятие семантического перехода используется также в семантических исследованиях самого широкого спектра — ср., например, статью [Зализняк, Торопова, Янин 2005], авторы которой пользуются понятием «сходного семантического перехода» при установлении значения слов в древних текстах.

Несколько замечаний о самом термине. Слово *переход* представляется наиболее удачным для обозначения обсуждаемого явления, хотя никакого «перехода» здесь нет (напомню: семантический переход — это факт совмещения двух значений). Тем самым «переход» — это метафора, но эта метафора мне кажется более эффективной, чем любой другой, более техничный термин — например, *корреляция* или *деривация*. Надо сказать, что «перехода» нет даже в диахронии. Как известно, семантическая эволюция обычно происходит по такой схеме: у некоторого слова, имеющего значение A, появляется (обычно контекстно обусловленное) значение B, затем значение A исчезает, а B освобождается от контекстной зависимости, и это выглядит похоже на то, что значение A «переходит» в значение B (см., в частности, [Traugott, Dasher 2002: 11; Evans, Wilkins 2000: 549]).

Семантическая деривация (см. определение и постановку задачи ее описания в Падучева 2004:

[1] Работа выполнена при поддержке ИНТАС, грант № 05-1000008-7917.

[2] О соотношении лексической, семантической и лексико-семантической типологии см. [Brown 2001, Рахилина, Плунгян 2007, Koptjevskaja-Tamm 2008].

147–154) — это правило порождения производного значения определенного типа из исходного значения определенного типа. Между тем семантический переход — это факт совмещения в пределах одного слова двух индивидуальных значений.

Термин *семантический сдвиг*, иногда применяемый к обсуждаемому явлению, имеет коннотацию, что результат такого процесса — это некоторое «отклонение» от прямого или буквального значения (например, метафорический сдвиг). Слово *переход* этого смысла не содержит, т.е. представляет собой более широкое понятие, чем *сдвиг*; впрочем, в качестве английского эквивалента для термина *семантический переход* мы используем *semantic shift*; в литературе в обсуждаемом значении используются также термины *semantic change* и *semantic extension*; по-французски тот же феномен обозначается термином, имеющим совсем другую внутреннюю фому, — *association sémantique* (см. [Vanhove 2008]).

Семантический переход обнаруживает себя в реализациях, т. е. словах или парах родственных слов, в которых он представлен. На сегодня мы выделяем шесть типов реализаций семантического перехода: 1) **полисемия** (A и B — значения одного слова одного языка в синхронии), 2) **семантическая эволюция** (A и B — значения одного слова одного языка, или языка-предка и языка потомка, в диахронии), 3) **деривация** (B — значение морфологического деривата слова, имеющего значение A, ср. итал. *contare* ‘вести счет’ и *raccontare* ‘рассказывать’); 4) **когнаты** (значения A и B принадлежат словам двух близкородственных языков, ср. фр. *espérer* ‘надеяться’ и исп. *esperar* ‘ждать’; 5) **заимствование** (иностранное слово со значением A в языке-источнике в заимствующем языке приобретает значение B, ср. румын. *a munci* ‘работать’ из ст.-слав. мѫчити ‘мучить’; 6) **грамматикализация** (значение B — грамматическое).

Таким образом, реализацией семантического перехода может быть либо одно слово, имеющее оба значения, либо пара родственных (в том или ином смысле) слов, одно из которых имеет значение A, а другое — значение B. Тем самым, семантический переход, это, по существу, расширение понятие полисемии — за счет расширения понятия «одного и того же слова», а именно, за счет включения сюда диахронической семантической эволюции, когнатов в родственных языках и морфологических дериватов (ср. Толстая 2008: 20)[3].

Итак, понятие семантического перехода представляет собой способ отразить тот факт, что некоторые два смысла выражаются одним и тем же словом. Факт совмещения двух значений в пределах одного слова, в свою очередь, очень важен, так как свидетельствует о их близости, имеющей довольно существенные последствия[4]. Если сходное совмещение значений наблюдается в нескольких словах (одного или разных языков), возникает ***семантическая параллель***. Так, например, семантический переход ‘достигать’ → ‘быть достаточным’ имеет следующие реализации (каждая из них является семантической параллелью к двум остальным):

1) русск. *достать*:
значение A: *достать до потолка*
значение B: *достаточно*, ср. также устар. *достанет* (*если достанет сил*)

2) русск. *хватать*:
значение A: *хватать за руку*
значение B: *мне не хватает денег*

3) нем.*reichen:*
значение A: *er reicht mit dem Kopf fast bis zur Decke* ‘он достает головой почти до потолка’; *erreichen das Ufer* ‘достигнуть берега’,
значение B. *das Geld reicht nicht* ‘не хватает денег’

Наличие семантических параллелей говорит о том, что данное совмещение значений отражает некоторые общие принципы языковой концептуализации[5]. Поскольку Каталог ориентирован на установление закономерностей, в него включаются только такие переходы, которые имеют как минимум две реализации, т.е. фактически это собрание семантических параллелей.

Идентификация семантических переходов сталкивается с рядом трудностей. Мы здесь рассмотрим лишь одну из них, которую назовем проблемой синкретизма. Формулируется она так: следует ли, на основании того, что в языке X смыслы P и Q выражаются разными словами, а в языке Y —

---

[3] Включение в число параметров варьирования морфологической деривации (словообразования) требует одной существенной оговорки: естественно, речь не идет о случаях, когда значение B принадлежит словообразовательной модели; так, пара *писать — переписать* не порождает семантического перехода ‘писать’ — ‘переписать’.

[4] См. об этом подробнее [Зализняк 2005].

[5] Это безусловно так, если языки не контактируют; если же они контактируют, то может оказаться, что данный семантический переход был заимствован (т.е. имеет место семантическая калька), ср. слова: англ. *to find*, нем. *finden*, франц. *trouver* итал. *trovare*, русск. *находить*: все они имеют значение ‘(вновь) обретать <некий предмет>’ и значение ‘иметь мнение’; маловероятно, что сходное семантическое развитие во всех этих языках было полностью независимым. Однако факт калькирования сам по себе представляет интерес в разных отношениях и, в частности, он косвенным образом тоже свидетельствует о его когнитивной значимости. О проблеме калькирования семантических переходов и обосновании включения семантических калек в Каталог семантических переходов см. [Зализняк 2006: 413–417; Грунтов 2007].

одним словом (т.е. «колексифицируются» — если воспользоваться термином из [François 2008]), признавать существование семантического перехода P → Q? Например, есть ли такие переходы: ‘кисть руки’ — ‘часть руки от плеча до кисти’, ‘синий’ — ‘голубой’, ‘синий’ — ‘зеленый’, ‘плыть (активно, о человеке)’ — ‘плыть (пассивно, о предмете)’, ‘теща’ — ‘свекровь’, ‘старший брат’ — ‘младший брат’, ‘теленок до года’ — ‘теленок от года до двух’, ‘быть знакомым’ — ‘владеть информацией’, ‘правда’ — ‘истина’, ‘свобода’ — ‘воля’ и т.д.? Для всех этих пар есть языки, где данные смыслы выражаются разными словами и языки, где они выражаются одним словом.

Трудность в том, что этот вопрос, будучи, безусловно, ключевым для построения типологии семантических переходов, т.е. фактов совмещения в одном слове двух значений, в общем виде не может быть решен — ни положительно, ни отрицательно. Действительно, положительное решение этого вопроса означает, прежде всего, совершенно неправомерное дробление значений (даже если считать это не полисемией, а синкретизмом). Так, ни в каком смысле, по-видимому, нельзя сказать, что русское слово *брат* «совмещает» два значения — ‘старший брат’ и ‘младший брат’.

Можно было бы считать, что семантический переход в обсуждаемой ситуации следует усматривать лишь тогда, когда слово, обслуживающее оба значения, является полисемичным, т.е. именно «совмещает» два отчетливо различных значения; в этом случае семантический переход возникает по определению (естественно, при условии наличия хотя бы еще одной реализации данного перехода, см. выше). Такое решение в теоретическом отношении безупречно, но оно натыкается на непреодолимое препятствие, состоящее в том, что реально отграничить полисемию от синкретизма невозможно[6]. Единственным критерием, который практически может быть применен (особенно по отношению к труднодоступным языкам) здесь может быть принятая лексикографическая практика, хотя он и не слишком надежен[7]. Так или иначе, определение понятия семантического перехода, очевидно, не должно зависеть от противопоставления полисемия vs. синкретизм.

С другой стороны, однозначно отрицательное решение этого вопроса тоже выглядело бы странно, так как тот факт, что имеются языки, где некоторые два смысла выражаются одним словом, без всякого сомнения указывает на близость данных двух смыслов, т.е. именно на то, что является объектом типологии семантических переходов.

Наше решение таково. В вопросе о разграничении полисемии и синкретизма достаточно следовать толковым словарям. С другой стороны, одного только факта раздельной лексификации некоторых двух значений в каком-то одном языке недостаточно, чтобы считать все случаи их колексификации в других языках полисемией (и даже синкретизмом) и, главное, постулировать соответствующий семантический переход. Для этого должны быть выполнены какие-то еще дополнительные условия. А именно, здесь должны учитываться два фактора. Первый — количественный: если языков, различающих значения P и Q, много, то это аргумент «за» (как в случае с ‘hand’ — ‘arm’, см. [Haspelmat et al. 2005]). Второй аргумент «за» — если тот же семантический переход обнаруживается в виде реализации другого типа, например, в форме семантической эволюции или в рамках морфологической деривации.

Рассмотрим, например, гипотетический семантический переход ‘сыр’ — ‘творог’[8], возникающий за счет того, что во французском языке оба продукта называются одним словом *frommage* (творог обозначается как *frommage blanc*, т.е. трактуется как вид сыра), между тем в русском языке имеется два разных слова — *сыр* и *творог*. Как уже говорилось, для постулирования семантического перехода этого недостаточно. Однако мы располагаем следующими фактами русского языка. Словарь Срезневского дает для слова *сыръ* в качестве 1-го значения ‘творог, сыр’ (т.е. русское слово *сыр* — это реализация обсуждаемого семантического перехода класса «эволюция»); слово *сырник* (кулинарное изделие из творога, ср. другой, более поздний вариант его обозначения — *творожник*); словосочетание *сыр домашний* (вид творога, ср. другое его название *творог зернистый*). Слово *сырник* демонстрирует реализацию обсуждаемого семантического перехода класса «деривация». Заметим, что аналогичная ситуация имеет место в немецком языке, где для сыра и творога есть два разных слова — соответственно, *Käse* и *Quark*, однако творожный торт называется *Käsekuchen*. В испанском языке одно из обозначений для творога (*requesón*) производно от названия сыра (*queso*); франц. *fromage blanc*, англ. *cottage cheese*, — это устойчивые сочетания, т.е. тоже дериваты. Тем самым, обсуждаемый переход может быть признан и включен в Каталог.

---

[6] В статье [Koptjevskaja-Tamm 2008: 8] относительно семантической структуры русского глагола *плыть/плавать* на фоне английских глаголов *float*, *swim*, *sail*, обозначающих три разных вида плавания, отмечается, что здесь методологически имеется три возможности: “semantic generality, polysemy and agnosticism”, при этом наиболее правильным решением автору представляется последнее.

[7] Так, слово *рука*, согласно русской лексикографической практике, не различает значений ‘arm’ и ‘hand’ (т.е. это не полисемия, а синкретизм); ср., однако аргументы в пользу признания такой полисемии для слова *рука* в [Wierzbicka 2007].

[8] Вопрос о направлении данного перехода составляет отдельную проблему.

В связи с проблемой синкретизма следует упомянуть метод типологического исследования принципов колексификации, отчасти альтернативный к нашему: метод ***семантических карт,*** разработанный М.Хаспельматом — см. Haspelmath 2003, где предлагается, в частности, семантическая карта значений ‘дерево’ — ‘древесина’ — ‘дрова’ — ‘лес’ в нескольких европейских языках. Этот метод успешно применяется также, например, в статье [François 2008], где исследуются возможности колексификации многочисленных производных значения ‘дышать’. Несколько иной вариант данного метода представляют собой семантические таблицы в [Koch 2001: 1145–1146]. Приведем, в существенно модифицированном виде, одну из таких таблиц (ср. также таблицу ‘человек’ — ‘мужчина’ — ‘муж’ в [Зализняк 2006: 411]):

Табличный метод обладает двумя преимуществами: во-первых, в таблице в отражены парадигматические отношения между несколькими значениями; во-вторых, для таблицы несуществен характер отношений (полисемия vs. синкретизм) между колексифицируемыми значениями. Поэтому метод семантических таблиц является удобным дополнительным — относительно принятого в «Каталоге» метода бинарных семантических переходов — инструментом описания для соответствующих семантических зон.

	‘волосы на голове’	**‘волосы в бороде’**	**‘волосы на теле человека’**	**‘шерсть у животного’**	**‘шерсть (материал)’**
англ.	*hair*				*wool*
нем.	*Haar*				*Wolle*
рус.	*волосы*				
лат.	*capellus*		*pilus*		*vellus, lana*
фр.	*cheveu*	*poil*			*laine*
итал.	*capello*	*pelo*			*lana*

Литература

1. *Грунтов И. А.* «Каталог семантических переходов» — база данных по типологии семантических изменений // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог-2007. М.: 2007. С. 157–161.
2. *Зализняк А. А., Торопова Е. В., Янин В. Л.* Берестяные грамоты из раскопок 2004 г. в Новгороде и Старой Руссе // Вопросы языкознания, 2005, № 3.
3. *Зализняк Анна А.* Семантическая деривация в синхронии и диахронии: проект создания «Каталога семантических переходов» // Вопросы языкознания, № 2, 2001. С. 13–25.
4. *Зализняк Анна А.* Проблема внутренней формы слова в типологическом аспекте //Язык. Личность. Текст. Сборник статей к 70-летию Т. М. Николаевой. М.: 2005. С. 87–108.
5. *Зализняк Анна А. Многозначность в языке и способы ее представления // М.: «Языки славянских культур», 2006.*
6. *Майсак Т. А.* Типология грамматикализации конструкций с глаголами движения и глаголами позиции // М.: 2005.
7. *Плунгян В. А.* Проблемы грамматического значения в современных морфологических теориях // Семиотика и информатика. Вып. 36, М.: 1998. С. 324–386.
8. *Рахилина Е. В., Плунгян В. А.* О лексико-семантической типологии // Глаголы движения в воде: лексическая типология. М.: 2007. С. 9–26.
9. *Розина Р. И.* Семантическое развитие слова в русском литературном языке и современном сленге // М.: 2005.
10. *Толстая С. М.* Пространство слова. Лексическая семантика в общеславянской перспективе // М.: 2008.
11. *Blank A.* Polysemy in the Lexicon // R. Eckarolt, K. von Heusinger (eds.). Meaning change — meaning variation. Workshop held at Konstanz, Feb. 1999. Konstanz, 2000.
12. *Brown C.* Lexical typology from an anthropological point of view // Language Typology and Language Universals. An International Handbook. M. Haspelmath, E. König, W. Oesterreicher, W. Raible (eds.),. Berlin — N.-Y.: Walter de Gruyter. Vol. 2., P. 1178–1189

13. *Bybee, Joan L.; Perkins, Revere; and William Pagliuca.* The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World. Chicago: University of Chicago Press, 1994.
14. *François A.* Semantic maps and the typology of colexification. // Vanhove M. (ed.) 2008. P. 163–216.
15. *Haser V.* Metaphor in semantic change // Metaphor and Metonymy at the Crossroads. A Cognitive Perspective. Ed. by A. Barcelona. Topics in English Linguistics, 30/ Mouton de Gruyter: Berlin — New York, 2000. P. 171–193.
16. *Haspelmath M.* The geometry of grammatical meaning: Semantic map and cross-linguistic comparison // Tomasello M. (ed.) The new psychology of language, vol 2. Mahwah, NJ: Lawrence Erlbaum, 2003. P. 211–242.
17. *Haspelmath M., Matthew S. Dryer, David Gil, and Bernard Comrie* (eds.) 2005. The World Atlas of Language Structures. Oxford: Oxford University Press.
18. *Heine B., Kuteva T.* World Lexicon of Grammaticalization. Cambridge Univ. Press, 2002.
19. *Koch P.* 2001. Lexical typology from a cognitive and linguistic point of view // Language Typology and Language Universals. An International Handbook. M. Haspelmath, E. König, W. Oesterreicher, W. Raible (eds.), Vol. 2, 1143–1175. Berlin — N.-Y.: Walter de Gruyter.
20. *Koch P.* Diachronic onomasiology and semantic reconstruction // W. Mihatsch, R. Steinberg (eds.), Lexical Data and Universals of Semantic Change, Tübingen: Stauffenburg, 2004. P.79–106.
21. *Koptjevskaja-Tamm M.* Approaching lexical typology // Vanhove M. (ed.) 2008. P.3 52.
22. *Sweetser E. E.* From Etymology to Pragmatics // Cambridge Studies in Linguistics 54. Cambridge: Cambridge Univ. Press, 1990.
23. *Traugott E. C., Dasher R. B.* Regularities in semantic change. Cambridge Univ. Press, 2002.
24. *Vanhove M.* (ed.) 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, 2008.
25. *Vanhove M.* Semantic associations between sensory modalities, prehension and mental perception // Vanhove M. (ed.) 2008 P.341–370.
26. *Wierzbicka A.* Bodies and their parts: An NSM approach to semantic typology // Language Sciences, 2007, Vol. 29. P. 14–65.
27. *Wilkins D. P.* Natural tendencies of semantic change and the search for cognates. // M. Durie, M. Ross (eds.). The Comparative Method Reviewed. Regularity and irregularity in Language Change. N.Y., Oxford: Oxford Univ. Press, 1996.
28. *Zalizniak Anna A*. A Catalogue of Semantic Shifts: towards a Typology of Semantic Derivation // Vanhove M. (ed.) 2008. P. 217–232.
29. *Zalizniak Anna A., M. Bulakh, D. Ganenkov, I. Gruntov, T. Maisak, M. Russo.* The Catalogue of Semantic Shifts as a Database for Semantic Typology. // Linguistics (in print).

<strike>Я этого не говорил</strike>: о литуративах, зачеркиваниях или мнимых текстах[1]

<strike>I've never told that</strike>: about lituratives, strikeouts or imaginary texts

Занегина Н. Н. (zanegina@list.ru)

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

Статья посвящена рассмотрению языковых (семантических и синтаксических) особенностей явления из сферы интернет-коммуникации, при котором часть текста выделяется особым шрифтом — зачеркивается.

0. В статье речь пойдет о коммуникативных причинах и языковых особенностях распространенного в интернет-коммуникации (реже — вообще в печатном тексте) явления, при котором часть сообщения выделена специальным шрифтом — зачеркнута[2]. Явление получило особое распространение вместе с новыми техническими возможностями[3].

Попытки филологического описания и интерпретации явления была сделана в работе [Гусейнов 2008]. Его семантика (или, точнее, причины его использования) была описана здесь силами самого явления:

«Семантика литуратива включает следующие уровни:

— *«~~Всё-то вам, дуракам, приходится объяснять, а то ведь сами бы не поняли, ага!~~»*

— *«~~На самом деле, я хотел сказать вот это!~~»*

— *«~~Всё дозволено~~»*

— *«~~Хоть я и понимаю, что так говорить не принято~~, современные технологии позволяют мне ненадолго и почти безболезненно обойти требования приличий (политической корректности, логики и т.п.), но потом вернуться к привычной повестке дня».*

— *«~~Хоть я и понимаю, что кому-нибудь хотелось бы сказать это~~, по здравом размышлении нельзя не признать вот это»».*

В этой же работе явление получило название литуратив (от litura (от глагола lino) — «зачеркивание»). Можно предложить еще одно название — мнимый текст.

Если целью зачеркиваний в бумажном тексте *обычно* является избавление от нежелательного фрагмента, исключение его из поля зрения читателя, то цель электронного зачеркивания — *всегда* обратить внимание на зачеркнутый текст. К такому «удаленному» тексту бывает необходимо привлечь повышенное внимание по двум основным причинам: 1) текст был подвергнут правке, и теперь хорошо бы сделать так, чтобы читатель заметил, что именно было поправлено[4]; 2) автору важно подчеркнуть мысль, содержащуюся в зачеркнутом тексте[5]. Действительно, «текст, который хотят уничтожить, просто стирают. Зачеркивают как раз то, что имеют в виду. А затем как бы произносят вслух другое

[1] Автор выражает признательность за помощь в подготовке настоящей статьи Н. И. Лауфер.

[2] Ср.: «6. Лекция: Оформление текста. Шрифт <…> Зачеркнутый текст используется, в основном, при оформлении документов приватного характера, в частности, при создании записей в блогах» (Спиридонов О. В. Работа в Microsoft Word 2007, http://www.intuit.ru/department/office/msword2007/6/3.html).

[3] По словам Г. Гусейнова, они «расцвели именно в блогосфере с первых 2000-х гг. Зачеркивание как индивидуальный прием можно встретить хоть у Лоренса Стерна в «Тристраме Шенди» или у Гофмана» (http://arno1251.livejournal.com/291761.html?view=4427697#t4427697).

[4] Ср.: «Well-known PR blogger, Jeremy Pepper, has a credo of his own, which states: "When I correct mistakes — if they are beyond spelling and grammar mistakes — I will note that a post has been updated." The strikethrough function is the HTML tag normally associated with this type of "letting the record stand" form of editing» [Chaney 2008].

[5] Ср.: «…in Internet culture, the strike-through has already taken on an ironic function, as ~~a ham-fisted way of having it both ways in type~~ a witty way of simultaneously commenting on your prose as you create it» [Cohen 2007].

(незачеркнутое). Возникает игра и новое измерение текста. Наряду с тем, что говорится, появляется то, о чем автор думает» [Кронгауз 2009].

Подробное рассмотрение истории явления и схожих с ним приемов не входит в задачи настоящей работы. Для исследования был собран корпус минимальных контекстов, в которых встречается зачеркнутый текст. Источник — записи в «Живом журнале», сделанные в последние два года одиннадцатью авторами. На данный момент обработано более 150 контекстов. Литуративы проанализированы с точки зрения прагматики (возможные причины их использования) и синтаксиса (синтаксические характеристики зачеркиваемых единиц).

1. Прагматика

Среди зачеркиваемых текстов устойчиво выделяются пять частотных, постоянно воспроизводимых типа, которые условно можно назвать следующим образом: «не буду говорить непринятое», «не буду говорить неправду», «не буду говорить правду», «не буду говорить банальности», «не буду говорить» (менее частотные типы, а также варианты, не попавшие в данный корпус контекстов, требуют отдельного рассмотрения[6]).

1.1. Автор подвергает порождаемые им тексты **внутренней цензуре**, не прошедшие цензуру, т. е. противоречащие каким-либо нормам («не буду говорить непринятое»), зачеркиваются. Не пропускаются цензурой:

- слова, не удовлетворяющие культуроречевым требованиям допустимости: относящиеся к сниженной или обсценной лексике:

(1) *Да! Дайте я вам расскажу, какая я ~~дура~~балда.*

(2) *Чуваки исследуют сывороточный альбумин и на этом основании говорят человеку, есть ли у него рак. Из спектров электронного парамагнитного резонанса (ЭПР) аликвот со спиновым зондом и этанолом, на основе компьютерной имитации и фитинга спектров ЭПР, определяют степень изменения параметров связывания спинового зонда в центрах связывания альбумина в зависимости от концентрации спинового зонда и концентрации этанола, ~~не хрен собачий.~~*

(3) *Вдохновение, оно, ~~самка собаки~~, ну точь-в-точь как здоровье* (в данном примере произошло двойное зачеркивание: обсценное слово заменено на эвфемизм, который все равно зачеркнут).

Слово может заменяться на стилистически близкое, но не имеющее отрицательной коннотации:

(4) *Здравствуй, профессиональная ~~шизофрения~~ деформация. Сегодня ночью мне снились имена. Не люди, а именно имена. Сами по себе. Толпой* [автор работает с именами собственными] (деформация — это лучше, чем название болезни).

При этом их все-таки произносят, поскольку такие слова являются способом выражения и, как следствия, нейтрализации отрицательных эмоций.

- слова, характеризующие человека с отрицательно оцениваемой стороны:

(5) *Можно же простить такую фигню, раз уж я такая ~~дотошная~~ внимательная <…>* (нехорошо быть чрезмерно внимательным, но можно быть просто внимательным).

(6) *нет, чесслово, просто хотела ~~вылить ему на голову бутылку кетчупа и сплясать голая на его столе~~ поговорить…* (нехорошо вести себя подобным образом).

При этом их все-таки произносят, поскольку такие слова позволяют сказать о желаемом, но противоречащем традиционным моральным нормам.

- слова, затрагивающие табуированную для данного человека тему (замужество, одиночество, внешний вид, возраст, плохое настроение, сентиментальность):

(7) *Пять минут назад узнала, что последняя моя однокурсница вышла замуж. Я одна осталась в ~~вечных~~ девках* (страшно не выйти замуж).

(8) *НГ мне светит встретить ~~и провести~~ в полном одиночестве. По статистике таких людей всего 1%. Как и женщин, выходящих замуж в первый раз после 25 лет* (страшно провести Новый год или новый год в одиночестве).

(9) *Да, и спасибо всем за то, что моя аська взорвалась сообщениями, телефон не успевает записывать смски, а я — читать ваши пожелания ~~потому что мешают набежавшие слезы щщастья. Щорт, стала сентиментальной на старости лет~~* [благодарность за поздравления с днем рождения] (страшно быть сентиментальной).

При этом их все-таки произносят, поскольку такие слова позволяют в безопасной форме проговорить то, что вызывает страх, и тем самым отчасти нейтрализовать его.

1.2. Автор сообщает нечто **идеальное**, желаемое, но невозможное, зачеркивает его, а затем сообщает реальное, **правду** («не буду говорить неправду»).

(10) *Я вчера <…> успела сфотографироваться. На паспорте буду ~~красивая~~ опознавабельна*

[6] В частности, отдельного рассмотрения требует попытка записи пьесы техническими средствами блогов — запись комедии А. С. Грибоедова «Горе от ума» в виде блога с комментариями (зачеркнутыми в этом блоге оказываются реплики героев в сторону) (http://community.livejournal.com/gore_ot_uma).

(вряд ли получится красиво, будет достаточно, если просто смогут узнать).

(11) *Есть все же ~~много~~ несколько людей, к которым я помчусь ночью, если они позвонят, что у них беда* (вряд ли таких людей много, в действительности их всего лишь несколько).

(12) *вы как хотите, а я пошла вносить в свой вишлист ~~КВАРТИРУ с ровными полами прекрасной встроенной КУХНЕЙ и шикарной встроенной духовкой!!!~~ стотыщ разных форм для выпечки* (хорошо было бы иметь квартиру, но пока достаточно будет формы для выпечки).

1.3. Автор, наоборот, сначала сообщает **правду** либо в сниженной стилистической форме, либо неприглядную по сути, а затем **приукрашивает** ее стилистически или фактически («не буду говорить правду»):

(13) *А еще я была счастлива, когда взяла машинку и мы с Темкой состригли ему ~~лохмы~~ золотые локоны почти под ноль* (было бы хорошо, если бы это были красивые волосы).

(14) *Уже год я ~~мучительно стараюсь не есть; когда очень хочется, ем курицу~~ не ем мяса* (было бы хорошо, если бы просто не ела мяса).

1.4. Автор сначала вспоминает некую **устойчивую конструкцию,** языковой шаблон (устойчивое сочетание, крылатое выражение, прецедентный текст), поскольку именно эта единица может быть сразу извлечена из памяти, а затем зачеркивает его как банальный («не буду говорить банальности») или заменяет его часть или его весь единицей, описывающей конкретную ситуацию.

(15) *Некоторым (не показываю на них пальцем~~, хотя, конечно, это был Слоненок~~) сказочно, просто сказочно повезло* (цитата из мультфильма).

(16) *<…> наша стиральная машинка <…> берет и ТУТ ЖЕ НАКРЫВАЕТСЯ МЕДНЫМ ТАЗОМ!!! Вот совсем. ~~Пепельницу из мерседеса вытряхивали, по колесам мерседеса стучали!~~ Фильтры чистили. Накрылся, похоже, мотор* (цитата из анекдота)

(17) *Самое интересное, что этот паук выжил и как ни в чем не бывало продолжал ползать в ванной по стенке. Гвозди бы делать из этих ~~людей~~ пауков!* (цитата из стихотворения[7]).

(18) *Ну вот, коротенько ~~минут на сорок~~, я вам все рассказала :)* (цитата из фильма[8]).

1.5. Автор эксплицитно формулирует причину зачеркивания: не хочется / не стоит об этом говорить («не буду говорить»):

(19) *Но вот перевести все не успела уже <…> ~~Про то, что перевод безвоздмездный, потому что от начальницы, я писать не буду! :)))~~*

(20) *Днем раньше Нюра поила меня чаем с имбирем ~~про пирожные не скажу~~* [вместе с чаем были съедены пирожные].

2. Синтаксические особенности

С точки зрения синтаксической организации этого явления можно выделить пять типов в зависимости от того, какая составляющая текста зачеркивается и на что она заменяется.

2.1. Вербально не выраженная единица зачеркивается и ничем не заменяется («ноль заменяется на ноль»):

(21) *Вдохновение, оно, ~~самка собаки~~, ну точь-в-точь как здоровье* (обсценное слово заменено на эвфемизм, который все равно зачеркнут).

2.2. Часть слова зачеркивается («часть заменяется на ноль»). Чаще всего это недописанное обсценное слово:

(22) *даже если бы в сутках было 25, 26 или 28 или ну 36 часовя бы все равно сумела их куда-то ~~прое…~~*

2.3. Единица текста (слово/словосочетание или часть предложения) зачеркивается («единица заменяется на ноль»):

(23) *Был тут у меня разговор с братом. Братик младший, понимаете, и вот спрашивает он у меня, ~~мол, про бабочек и мух-дрозофилл я знаю, а вот как оно у людей?~~*

2.4. Единица текста (слово/словосочетание или часть предложения) заменяется на единицу текста («единица заменяется на единицу»):

(24) *А что ~~лучше~~ — хуже: быть завистливым или злопамятным?*

---

[7] «Из стихотворения «Баллада о гвоздях» (1922) советского поэта Николая Семеновича Тихонова (1896—1979), где повествуется о человеческой стойкости. Видимо, этот образ вызван ассоциацией с выражениями «железный характер», «железная воля» и т. п.: Гвозди бы делать из этих людей, Крепче б не было в мире гвоздей. Иносказательно: о человеческой стойкости или упрямстве (ирон.)» (В. Серов. Энциклопедический словарь крылатых слов и выражений // http://bibliotekar.ru/encSlov/4/11.htm).

[8] «Из кинофильма «Карнавальная ночь» (1956), снятого режиссером Эльдаром Рязановым по сценарию Бориса Савельевича Ласкина (1914–1983) и Владимира Соломоновича Полякова (1909–1979). Слова начальника Дома культуры бюрократа Огурцова (актер Игорь Ильинский). Шутливо-иронически о выступлении, которое вовсе не обещает быть коротким» (В.Серов. Энциклопедический словарь крылатых слов и выражений // http://www.bibliotekar.ru/encSlov/10/151.htm).

Здесь возможны более сложные варианты замен:

(25) *Это я ~~кисти~~ объектива* [имя, фамилия] — *<…> замечательного фотографа* (устойчивое сочетание *X кисти Y* не может быть использовано, потому что функцию кисти выполняет объектив).

(26) *Экс ~~сматываетсворачивает в это время удочки~~ матрас* (автору кажется более удачным не воспроизведение фразеологизма *сматывать удочки* целиком с последующей его заменой на необходимое сочетание, а последовательная замена составных частей фразеологизма, позволяющая сделать языковую игру более сложной).

2.5. Целый текст зачеркивается («текст заменяется на ноль»).

(27) *~~Что бы такого написать с новым юзерпиком? Черт, ничего в голову не идет. Может, про любовь что ли. Тут нате вам — колечко, сердечко, книжка. Или про Дину, которая сделала это фото? Или ваще про добавление новых юзерпиков, которыми толком не пользуешься, просто рассматриваешь сам по себе и радуешься — такое вот коллекционирование. Или Про брильянты? — котльцо -то на юзерпике с бриллиантами. Правда, не мое, не мое кольцо, у меня другое обручальное. Или про …~~ У МЕНЯ НОВЫЙ ЮЗЕРПИК!*

Встречается в том случае, если текст потерял свою актуальность, если он представлял собой список дел, которые теперь выполнены, если целью записи в блоге был не текст, а использование нового юзерпика (см. пример выше).

Данный анализ собранных литуративов пока не позволяет сделать значительные выводы о сущности явления, однако демонстрирует разнообразие его проявлений и условий бытования. В дальнейшем было бы интересно рассмотреть семантические и синтаксические ошибки в построении подобных мнимых текстов, бытование зачеркиваний в остальной интернет-коммуникации, в особенности в той, которая требует максимально быстрого порождения текста (коммуникация в мессенджере), в печатных СМИ, связь литуративов с иронией и чтением между срок, а также сопоставить с постулатами Грайса.

Литература

1. *Chaney P.* Using strikethrough function in blogs… you may know «how», but do you know «why»? // http://www.conversationalmediamarketing.com/2008/05/using-strikethr.html. 08.05.2008.
2. *Cohen N.* Crossing Out, for Emphasis // http://www.nytimes.com/2007/07/23/business/media/23link.html?_r=2. 23.07.2007.
3. *Гусейнов Г.* Неполная коммуникация в блогосфере: эрративы и литуративы // http://www.speakrus.ru/gg/liturative.htm.
4. *Кронгауз М.* Язык и коммуникация: новые тенденции. Лекция // http://www.polit.ru/lectures/2009/03/19/communication.html. 19.03.2009.

Об одном подходе к автоматическому построению онтологии для задач анализа текстов

An approach to automated ontology building in text analysis problems

Захарова И. В. (iren@csu.ru),
Городечный П. П. (petr.gorodechnyj@edu.csu.ru)

Челябинский государственный университет, математический факультет

В статье описан метод автоматического построения онтологии для сложных задач классификации, аннотирования и поиска текстовых документов.

1. Введение

Развитие индустрии систем электронного документооборота, сопровождающееся ростом массивов обрабатываемых полнотекстовых документов, требует новых средств организации доступа к информации, многие из которых следует отнести к разряду систем искусственного интеллекта — систем обработки знаний. Одним из эффективных подходов к выявлению и обработке смысла текстовых документов является использование онтологий.

Онтология определяет термины, используемые для описания и представления знаний той или иной предметной области. Она необходима для людей, для приложений систем баз данных и различных других информационных систем, которые совместно используют специфическую информацию в какой-либо предметной области. Онтологии включают доступные для компьютерной обработки определения основных понятий предметной области и связи между ними[2].

2. Модель онтологии, специализированная для задач семантического поиска и классификации

Формально определим онтологию как множество

$O = (L, C, F_l, F_c, R_h)$, где

$L = \{(w_i, x_i)\}_{i=1,n}$ — словарь терминов предметной области,

w_i — термин, возможно более одного слова

x_i — его рейтинг относительно других терминов в концепции.

$$C = \{c_i\}_{i=1,m}$$

C — набор понятий (концепций),

$F_l(L) -> C$ — Функция интерпретации терминов
Сопоставляет набору терминов из словаря подмножество концепций.

$F_c(C_i) -> L$ — Функция интерпретации концепций;
сопоставляет концепции набор терминов из словаря.

R_h — Отношения иерархии между концепциями [4].

$$P(c_i \mid u)$$

В качестве функции интерпретации терминов возьмем — вероятность выбора концепции при условии запроса *u*.

Применив формулы полной вероятности и формулы Байеса[3], получим

$$F_l(u) = \left\{ c_i \mid P(c_i \mid u) = \right.$$

$$\left. = \max_{c_j \in C} \left(\sum_{w \in u} \left(\frac{x_w^j}{\sum_{c_k \in C} x_w^k} \bullet \frac{count(w, L)}{\sum_{w' \in u} count(w', L)} \right) \right) \right\},$$

$$i = \overline{1,n}$$

Определим обратную функцию интерпретации как множество терминов, относящихся к данной концепции с весом большим, чем средний вес всех терминов для данной концепции.

Функцию интерпретации концепций определим как

$$F_c(c_i) = \left\{ w_j \mid x_w^j \geq \frac{\sum_{w \in L_i} x_w}{\sum_{w \in L_i} 1}, j = \overline{1,k} \right\}, \text{ где}$$

$L_i = \bigcup_j w_j^i$ — множество всех терминов, соответствующие концепции c_i .

3. Метод построения онтологии

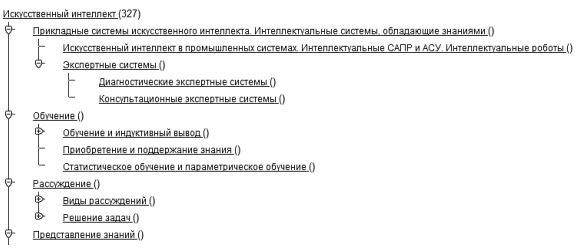
Для реализации эффективного семантического поиска необходима онтология, которая, по сути, описывает не одну какую-либо предметную область, а классифицирует все виды сущностей и связи между ними. Создание подобной системы возможно как минимум двумя путями.

Специалисты в некоторой предметной области создают для собственных целей онтологию. Объединяя эти предметно — ориентированные онтологии и добавляя, возможно, при этом дополнительные связи, получаем «обобщенную онтологию». Метод, очевидно, долгий и требующий работы множества экспертов по многим предметным областям.

Второй способ — построить онтологию автоматически, используя для этого имеющиеся коллекции информационных ресурсов и библиографических баз данных, представленных в Интернет.

В 1962 г. в стране в качестве единой обязательной классификации принята Универсальная десятичная классификация (УДК), и введено обязательное индексирование всех публикаций, т. е. все информационные материалы в области естественных и технических наук издаются с индексами Универсальной десятичной классификации.

Пример дерева УДК для «ветки» 004.8.



В результате, мы имеем экспертную базу, на многих языках, где для каждого классификационного кода определено подмножество различных публикаций, содержащих знания по данной теме. Наша задача выделить эти знания и представить их в виде набора терминов, наиболее характерных для данной рубрики[5].

Рассмотрим библиографическую запись об одной книге:

Ирбенек В. С. Алгоритмы проектирования топологии электрических соединений в САПР электронной аппаратуры // Зарубежная радиоэлектроника. Успехи современной радиоэлектроники. — 2002. — № 7. — С. 71–79

Ключевые слова

автоматизация; автоматизированное проектирование; алгоритмы; деревья Краскала-Прима; деревья Штейнера; ортогональная метрика; проектирование автоматизированное; САПР; электроника; электронная аппаратура.

Код УДК

004.896

Сам метод выделения терминов из ББД можно представить в виде схемы

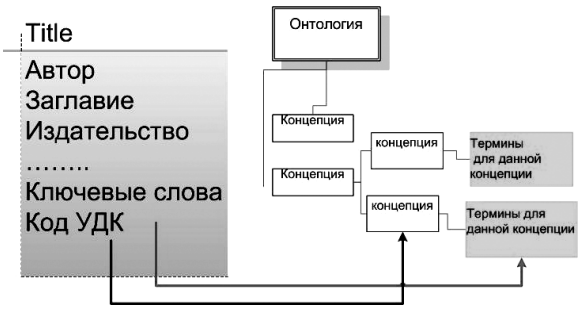


Рис. 1

В настоящее время в России в библиотечном сообществе широко распространена идея создания сводных каталогов, объединяющих отдельные библиотечные каталоги участников либо в единый физический каталог (путем копирования данных на один сервер), либо в распределенный каталог (поиск и работа с которым осуществляется распределенно). Управление доступом к распределенным информационным ресурсам и взаимодействие электронных библиотек осуществляется по принципу распределенных систем на базе открытых стандартов обмена данными. Для реализации электронных библиотек используются в основном два протокола: Z39.50 и HTTP. В качестве подсистемы построения онтологии был выбран протокол Z39.50, изначально ориентированный на информационно-поисковые задачи именно в библиографических базах данных [6,7].

Общая архитектура приведена на рисунке 2.

С помощью программы были просканированы сводные и распределенные каталоги Ассоциации Региональных Библиотечных Консорциумов (АРБИКОН) и выделено 133 151 концепции, содержащие от 5 до 100 терминов для каждой концепции.

Рис. 2

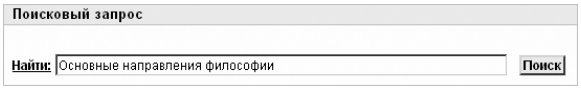
4. Применение онтологии

Полученную онтологию предполагается использовать в аналитической системе BIOAP (Basic Integrated Ontological Analytic Processor) 1.0 для:

- Классификация/рубрицирования (определения типа документа)
- Реферирования/аннотирования (извлечения краткого содержания из текста)
- Семантического поиска по коллекции документов

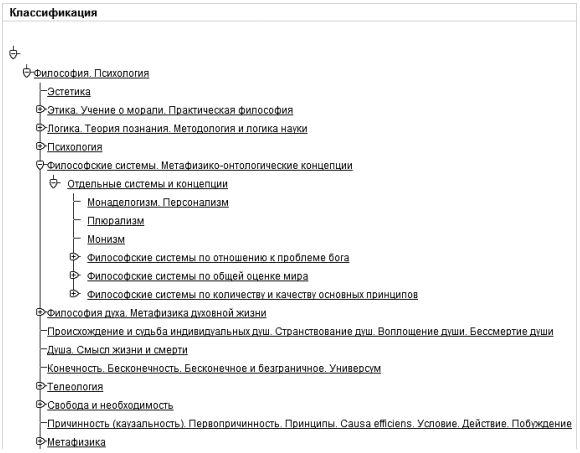
На данный момент реализованы алгоритмы семантического поиска.

Ввод запроса пользователем осуществляется как в любой поисковой системе:



Дальнейшая работа системы осуществляется по следующей схеме:

1. Запрос делится на термины
2. К терминам применяется функция, выделяющая корень слова (стемминг)
3. Для каждого термина рассчитывается его вес в контексте онтологии
4. Применяем функцию интерпретации терминов к запросу.
5. Получаем список концепций, наиболее релевантных запросу по смыслу. Например, для указанного выше запроса получим концепцию «*Философия. Психология*».
6. Для каждой концепции выполняем поиск подчиненных концепций и выводим их на экран в виде дерева. Например, для указанного запроса будет получено следующее дерево:



7. Пользователь уточняет запрос, указывая конкретно, какая тема его интересует.

В данном случае выбирается «Философские системы. Метафизико-онтологические концепции» и далее «отдельные системы и концепции», потом — «монизм»

8. Исходный запрос дополняется терминами из онтологии, семантически связанными с этой концепцией.

В данном случае запрос был дополнен следующими терминами: *Гуманизация образования, Тейяр Де Шарден, одаренность, формирование личности, всеединство, преджизнь, ноосфера*.

9. Расширенный запрос передается поисковой системе «Yandex Standard»
10. На экран выводится список найденных документов.

Например для указанной концепции «отдельные системы и концепции. Монизм» был получен следующий список документов :

1. Алешин А. И. Русская философия: Малый энциклопедический словарь.
2. Малахов В. С., Филатов В. П. Современная западная философия: Словарь.
3. Канке В. А. Основные философские направления и концепции науки. Итоги XX столетия.
4. Реале, Джованни. Западная философия от истоков до наших дней.
5. Суханов К. Н. Динамика тематической направленности философствования в XIX–XX столетиях.
6. Люсый А. К двумирной полноте — через терминологический монизм
7. Бородин Е. Т. Монизм и плюрализм в современной общественной науке.

5. Заключение

Предполагается дальнейшее использование онтологии для решения задач классификации и реферирования больших коллекций электронных полнотекстовых документов.

Литература

1. *Baeza-Yates R., Ribeiro-Neto B.* Modern Information Retrieval. ACM Press, 1999.
2. *Gruber T. R.* A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 1993.
3. *Гнеденко Б. В.* Курс теории вероятностей. — М.: Наука, 1988.
4. *Zakharova I. V., Melnikov A. V., Vokhmitsev J. A.* «An approach to automated ontology building in text analysis problems». // Workshop on computer Science and Information Technologies CSIT'2006, Karlsruhe, Germany, 2006. P. 177–178.
5. *Melnikov A. V., Zakharova I. V.* «Method of automatic ontology creation based on bibliographic databases». // Workshop on computer Science and Information Technologies CSIT'2005, Ufa, Russia, 2005. P. 270–272.
6. *Глухов В. А., Голицына О. Л., Максимов Н. В.* Электронные библиотеки. Организация, технология и средства доступа // НТИ. — Сер. 1, — 2000, — №10.
7. *Жижимов О. Л.* Введение в Z39.50. — Новосибирск: Изд-во НГОНБ, 2000.

Универсальная система синтаксической разметки текстов ОбъектATE

Universal syntax annotation system ОбъектATE

Зобнин А. И. (Alexey.Zobnin@gmail.com)
Московский государственный университет им. М. В. Ломоносова, Москва, Россия, Институт русского языка им. В. В. Виноградова РАН

Сахарова А. В. (nenen@mail.ru)
Институт русского языка им. В. В. Виноградова РАН

Представлена система универсальной разметки текста ObjectATE, основанная на принципах объектно-ориентированного проектирования. Она используется в Отделе лингвистического источниковедения Институте русского языка им. В. В. Виноградова РАН для морфологической и синтаксической разметки древнерусских текстов (переводных памятников и летописей) в полуавтоматическом режиме. Система является гибкой и позволяет пользователю самому как задавать макет разметки в терминах шаблонов и надстроек (классов), так и описывать способы ее визуализации.

1. Предпосылки создания системы

На данный момент отсутствуют многофункциональные средства создания лингвистических текстовых корпусов, позволяющие заниматься лингвистической разметкой корпуса начиная с того уровня (морфологического, поверхностно-синтаксического, семантического и т. п.), который выбирает разметчик и по тем параметрам, которые он задает сам. Однако именно такое средство необходимо для создания лингвистически размеченного корпуса древних письменных памятников. Поскольку лексика и грамматика древних памятников не изучена в полном объеме, а сами тексты не свободны от разного рода ошибок и темных мест, их грамматическая разметка должна быть ручной. Однако было бы хорошо, если бы применяемая для такой разметки информационная система позволяла частично автоматизировать эту разметку.

Создаваемая система обработки текста ObjectATE (Object-oriented ancient text editor) призвана решить эти проблемы. Она разрабатывается и используется в Отделе лингвистического источниковедения Институте русского языка им. В. В. Виноградова РАН с 2006 г. [1–5]. Она пришла на смену предыдущей системе ATE, с помощью которой велась ручная и полуавтоматическая разметка морфологии в древнерусских текстах — переводных памятниках и летописях (переводная антология «Пчела», Киевская летопись по Ипатьевскому списку, Новгородская первая летопись и др.).

Новая система создавалась с целью вести прежде всего ручную синтаксическую разметку этих текстов. Однако рутинную часть работы в ней можно автоматизировать с учетом имеющейся морфологической разметки и формулируемых пользователем правил (морфологических и формально-синтаксических). При этом она призвана быть максимально гибкой и многофункциональной, позволяющей создателю корпуса строить в принципе любые единицы лингвистического анализа по своим собственным (а не только по тем или иным общепринятым) моделям. Было предложено решение на основе объектно-ориентированного подхода, широко применяемого в программировании. На разработку программы оказала большое влияние информационно-аналитическая система «Манускрипт»[1]. Уже в процессе создания ObjectATE авторы познакомились с такими системами обработки текста, как Emdros[2] и GATE[3]. Эти системы имеют свою специфику, и их трудно приспособить к решению поставленной задачи с максимальной общностью и универсальностью. Так, система «Манускрипт» довольно сложна и не дает пользователю

[1] http://manuscripts.ru.

[2] Emdros — the database engine for analyzed or annotated text. http://emdros.org.

[3] GATE — General Architecture for Text Engineering. http://gate.ac.uk.

возможности гибко настраивать макет разметки. Система GATE предусматривает множество различных обработчиков текста и позволяет его *аннотировать*, но, к сожалению, не предусматривает непосредственной работы с синтаксическими конструкциями. Больше всего соответствует поставленной задаче система Emdros, в которой всякий объект разметки представлен как множество элементарных составляющих (называемых монадами) и может содержать атрибуты и ссылки на другие объекты. Такой подход позволяет описывать синтаксис предложения как структуру составляющих. Запросы в Emdros строятся как описания таких структурных включений, но они не рассматриваются как типы данных.

В отличие от всех перечисленных наш подход обладает большим уровнем абстракции, а также механизмом ограничений, которые позволяют контролировать и даже частично автоматизировать разметку. Отметим, что рекомендации TEI[4], в первую очередь предназначенные для описания структуры текста, также не позволяют вести разметку на достаточном уровне абстракции. При этом объектно-ориентированный подход не предполагает особой формы хранения данных, которое может быть основано как на реляционной базе данных, так и на xml-подобном языке. Таким образом, сомнений в необходимости создания собственной разработки не возникало, однако знакомство с идеями, заложенными в этих системах, оказалось очень полезным и обязательно найдет применение в будущих версиях ObjectATE.

2. Функциональные возможности

Система ObjectATE разрабатывается как программное средство для создания, хранения и обработки текстов, проанализированных на любом лингвистическом уровне. Она позволяет заниматься в ручном режиме морфологической разметкой предварительно уже разделенного на словоформы текста, т. е. присваивать словоформам значения морфологических категорий (полей словоформ); при этом пользователь может сам создавать или редактировать списки этих категорий и их значений. В настоящее время в систему внедряются механизмы лемматизации, автоподстановки морфологических параметров, создания словников и указателей (хотя пока что для этих целей в Отделе лингвистического источниковедения применяется предыдущая версия редактора ATE).

Даже при отсутствии морфологической информации (т.е. если текст только разделен на словоформы) система ObjectATE обеспечивает возможность ручной синтаксической разметки текста, т.е. создания в базе данных новых объектов — единиц синтаксического анализа. Такие объекты могут быть организованы как угодно: состоять из одной, двух или большего количества словоформ, равноправных по отношению друг к другу или нет. При этом, разумеется, пользователь может описывать синтаксис и стандартным образом.

Так, чтобы разбирать текст по зависимостям, он должен создать список типов синтаксических связей и начать связывать друг с другом наборы словоформ. После того, как пользователь сформулирует, какой узел у каждой связи является вершинным, а какой — подчиненным, становится возможным построение ориентированного дерева зависимостей. Система также позволяет создавать вспомогательные для синтаксического анализа узлы, функционирующие как аналоги словоформ, например нулевые подлежащие личных глаголов или фантомные эллиптические нули с указанием на опущенную словоформу.

Чтобы маркировать некую синтаксическую группу, пользователь может просто одинаковым образом выделить словоформы или нули, входящие в нее, и создать соответствующий объект. Но система предоставляет возможность также заниматься полноценным синтаксическим анализом по группам, образующим иерархическую структуру. Это значит, что можно создавать синтаксические объекты из других уже существующих, которые только в частном случае представляют собой словоформы. Для этого, создав класс синтаксических объектов (в терминологии грамматики составляющих — фразовую категорию), пользователь затем должен оговорить, чем он может быть представлен (например, сформулировать, что сказуемое предложения может представлять собой одну словоформу, восстановленный ноль, аналитическую конструкцию и т.п.). Именно для того чтобы описать это явление, в системе предусмотрено применение механизма надстроек, позволяющих описывать множество разнородных объектов, удовлетворяющих определенным условиям. Например, надстройка «Глагол-связка» включает в себя как словоформу (глагол *быти* в личной форме), так и синтаксический объект под названием «Аналитическая личная форма» *(еси был, был бы)*. Создав такую надстройку «Глагол-связка», мы должны оговорить, что синтаксическая группа «Глагол-связка» должна образовываться только из объектов, входящих в эту надстройку.

Если морфологическая информация о словоформах для разбираемого текста уже имеется, система может использоваться для упрощения и частичной автоматизации синтаксической разметки. Предположим, можно создать надстройку «Сказуемое», в которую будут входить только все личные глаголы, и надстройку «Подлежащее», куда будут входить все субстантивы в именительном падеже. Вхождение в надстройку «Подлежащее» окажется в данном слу-

[4] TEI — Text Encoding Initiative. http://www.tei-c.org.

чае не достаточным, а только необходимым условием создания связи «Подлежащее–Сказуемое», поскольку, как известно, имя в именительном падеже может играть и другую синтаксическую роль.

Также можно оговорить не только условия вхождения словоформы или синтаксического объекта в надстройку, но и ограничения самого синтаксического объекта (например, согласование подлежащего и сказуемого по лицу: если лицо сказуемого — первое, то его подлежащее — либо ноль, либо местоимение первого лица). При необходимости можно оговаривать порядок словоформ относительно друг друга. В подобной системе также можно заниматься и ручной разметкой текста на более глубоких языковых уровнях, вводя специальные метки (коммуникативный статус, семантическая роль и т.п.) для синтаксических объектов или отрезков предложений. Наконец, для переводных текстов разрабатываются макеты разметки для описания соответствий между оригиналом и переводом.

3. Объектная модель данных

Как уже было сказано, система построена на основе объектно-ориентированного подхода, широко применяемого в программировании. Этот подход тесно связан с понятием онтологии в информатике. Весь размеченный документ представляется как набор объектов. Процесс разметки состоит в создании и модификации объектов.

В начале работы пользователь задает метаданные, то есть данные о структуре будущих объектов. Метаданные состоят из шаблонов и надстроек над ними. Шаблон следует понимать как абстрактный тип данных, определяющий вид объекта. Например, в стандартных текстах, с которыми работает система разметки, предполагаются такие шаблоны, как «Страница», «Строка», «Словоформа». Напротив, конкретные страница, строка или словоформа в тексте — это объекты соответствующих шаблонов. Всякий шаблон имеет уникальное имя.

Каждому шаблону приписан определенный набор полей и ограничений. С помощью полей одни объекты в документе могут быть связаны с другими. Так, строка текста относится к какой-то странице, слова расположены в определенных строках, а всякая словоформа обладает определенной частью речи. Поля шаблона — это набор типов признаков, которые могут быть у объекта этого шаблона. Соответственно, каждому полю шаблона приписано имя, а также указано, какие объекты могут выступать в качестве значения этого поля у объектов данного шаблона. Так, пользователь может определить шаблон «Главные члены предложения» с полями «Подлежащее» и «Сказуемое». Ограничения могут относиться и к типу данных значений полей (ясно, что подлежащее не может быть «Страницей», «Строкой» или «Частью речи»), и на сами значения полей и их подполей (например, если подлежащее это отдельная словоформа, имеющая падеж, то этот падеж должен быть именительным). Ограничения последнего вида можно накладывать на весь шаблон в целом. Такие ограничения записываются в виде логических условий на поля (и их подполя с любым уровнем вложенности). Истинность этих ограничений зависит от потенциального набора значений полей. Предполагается, что для всякого объекта данного шаблона эти ограничения должны превращаться в тождественно истинные выражения.

Шаблоны могут выстраиваться в иерархии наследования. Эта возможность оказывается очень удобной при описании метаданных. Шаблон-наследник приобретает все свойства (поля и ограничения) шаблона-предка и может их расширять. Шаблон-предок может быть объявлен абстрактным. Это значит, что он используется только как общий предок для других шаблонов-наследников, а создавать объекты такого шаблона нельзя. Например, если пользователь хочет наделить все объекты синтаксической разметки полем «Комментарий», он может определить это поле у общего абстрактного шаблона «Синтаксический объект» и вывести из него другие шаблоны.

В системе реализован механизм множественного наследования, позволяющий включать один и тот же шаблон в различные иерархии. Благодаря механизму надстроек в системе можно описывать условное наследование, когда набор базовых типов, вообще говоря, зависит от выполнения условий для конкретного объекта.

Надстройка отдаленно напоминает абстрактный шаблон. Она строится над уже существующими шаблонами или надстройками, которые называются кандидатами на вхождение в эту надстройку. Каждому кандидату надстройки может быть приписано условие на его вхождение в надстройку. Как и ограничение шаблона, это условие представляет собой логическое выражение, зависящее от конкретного объекта, его полей, подполей и т. д. Можно индуктивно определить понятие реализации объектом надстройки или шаблона. Во-первых, всякий объект O реализует свой собственный шаблон и все шаблоны-предки этого шаблона. Далее, пусть K — кандидат надстройки H и объект O реализует K. Тогда считается, что O реализует надстройку H, если для объекта O выполнено условие на вхождение кандидата K в H.

Надстройки появились в модели по двум причинам. Во-первых, этот механизм позволяет детально задать условия на поля шаблонов, а во-вторых, надстройки позволяют описывать простые запросы к данным. Рассмотрим эти возможности подробнее. Ранее полю шаблона строго сопоставлялся его тип — другой шаблон. Считалось, что только объек-

ты этого другого шаблона могут являться значениями полей. Это вызывало определенные трудности, прежде всего с «нулевыми» синтаксическими объектами. Нужно было сделать так, чтобы синтаксические нули наравне со словоформами могли быть полями других синтаксических объектов. Однако в случае, когда такие поля выражены словоформами, должно было выполняться дополнительное условие. В предлагаемой модели типом поля шаблона может быть или шаблон, или надстройка. Соответственно, объект может быть значением такого поля, если он реализует его тип. Такой подход позволяет более гибким образом описать модель разметки. С другой стороны, в программе имеется возможность проверить, реализует ли данный объект указанную надстройку, вывести список надстроек, реализуемых данным объектом, а также вывести все объекты, реализующие данную надстройку. Сами эти объекты могут иметь разные шаблоны; их объединяет лишь то, что при выполнении условий вхождения мы относим их к данной надстройке. Поэтому надстройки удобно рассматривать как описания простых запросов к данным, то есть таких запросов, которые возвращают отдельный список объектов.

Надстройка, как уже было сказано, задает достаточные условия для отнесения объекта к некоторой категории. В системе предусмотрен простой механизм, позволяющий показать, что для данного объекта данная надстройка задает и необходимые условия.

Всякий объект имеет обязательную текстовую компоненту «Содержание». Содержание объекта может задаваться пользователем, либо вычисляться по определенным правилам через содержания полей. Объекты имеют также два поля для сортировки и сравнения: дескрипторы начала и конца объекта. Считается, что все объекты данного фиксированного шаблона можно естественным образом упорядочить по их дескрипторам. Если дескрипторы начала и конца различаются, то объект считается «протяженным». Так, естественный порядок имеется на страницах, строках и словах текста. Удобно считать слово «атомарным» объектом, а дескрипторы начала и конца строки приравнивать к дескрипторам первого и последнего слова в строке. Аналогично, дескрипторы начала и конца страницы приравниваются соответственно к дескрипторам первой и последней строки в этой странице. Правила назначения дескрипторов новым объектам можно задавать при описании метаданных. Дескрипторы позволяют строить запросы и ограничения на порядок слов (например, найти все связи «Субстантив–атрибут», в которых субстантив находится раньше атрибута).

Поля шаблонов могут быть трех видов: обычные поля, коллекции и диапазоны (архитектурно предусмотрен четвёртый вид — коллекция диапазонов, но он пока не реализован, так как на данный момент не востребован). Поле-коллекция отличается от обычного поля тем, что предполагает сразу несколько различных значений. Диапазон — это «связная» коллекция, то есть множество объектов, идущих подряд в смысле упорядочения по дескрипторам. Для диапазона достаточно задать начальный и конечный объект. Типичный пример диапазонов — строки в странице или какие-либо естественные связные большие фрагменты текста (например, блоки, части, прямая речь и т. д.).

Поля шаблонов делятся на обязательные и опциональные. Обязательное поле заполняется при создании объекта (например, при синтаксической разметке). Для опциональных полей предлагается список возможных вариантов заполнения. Данный список формируется на основе ограничений шаблона и уже заполненных полей.

Если все кандидаты надстройки имеют общие поля, то при записи условия на поле типа этой надстройки такие поля можно использовать в выражениях. Кроме того, надстройки могут иметь свои поля. Собственные поля надстройки всегда являются опциональными. Объект приобретает такое поле только в том случае, если он реализует надстройку. С помощью такого механизма можно удобно описывать морфологическую разметку. Именно так была организована морфологическая разметка в базе данных «Новгородская первая летопись». В этой модели, например, словоформа имела лишь поле «Часть речи», а другие морфологические поля появлялись у нее только если она реализовывала какие-либо надстройки. Так, поле «Падеж» появлялось, если словоформа реализовывала надстройку «Имя», и т. д.

Условия и ограничения в метаданных задаются на специальном языке, который интерпретируется программой. Пользователь может создавать их как с помощью конструктора ограничений, так и записывать вручную. Язык содержит основные логические операторы AND, OR, NOT, операторы равенства (=), неравенства (<>), принадлежности (IN) и непринадлежности множеству (NotIN). В выражениях могут участвовать поля и их подполя с любым уровнем вложенности. Имена подполей задаются в квадратных скобках и разделяются точкой. Поле-коллекция всегда рассматривается как множество; кроме того, множество может быть задано явно с помощью фигурных скобок и перечислением входящих в него объектов. По умолчанию сравнение объектов производится по их содержанию. Вот пример ограничения на шаблон «Связь с согласованным атрибутом»:

```
([Атрибут].[Часть речи] IN {'прилагательное',
'причастие'})
OR ((([Атрибут].[Часть речи] = 'местоимение')
  AND ([Атрибут].[Лицо] NotIN {'1-е', '2-е', '3-е'})
  AND ([Атрибут].[Лексема] NotIN {'и'}))
OR ([Атрибут].[Часть речи] = 'числительное').
```

(Здесь так записанное условие на лицо атрибута просто означает, что это лицо отсутствует.)

Перечислим еще некоторые важные операторы этого языка:

1. Оператор проверки реализации IS. Он позволяет проверить, что данное поле объекта реализует указанную надстройку или шаблон. Например, «[Атрибут] IS Словоформа». Также в синтаксис языка ограничений добавлено ключевое слово Me, обозначающее сам проверяемый объект. В условиях на вхождение в надстройку удобно писать выражения вроде «Me IS Субстантив».

2. Условный оператор IF. С его помощью можно корректно обращаться к полям объектов, которые, вообще говоря, не являются общими. Вместе с оператором IS он частично заменяет механизм надстроек, обеспечивая большую гибкость. Пусть, например, поле «Подлежащее» может быть выражено как словоформой, так и нулем. Пусть шаблоны «Словоформа» и «Ноль» безусловно входят в некоторую надстройку. У шаблона «Ноль» нет поля «Падеж»; к падежу можно обратиться только у «Словоформы». Поэтому условие на подлежащее можно записать так:

```
IF ([Подлежащее] IS Словоформа, [Подлежащее].
        [Падеж] = 'именительный').
```

3. Операторы сравнения <= и >= позволяют сравнивать объекты по их дескрипторам сортировки и, в частности, строить запросы на порядок слов.

4. Интерфейс программы

Программная оболочка ObjectATE представляет собой графический пользовательский интерфейс для работы с размеченным текстом по описанной объектной модели. Она позволяет просматривать текст и отдельные объекты разметки с их свойствами, редактировать, создавать и удалять объекты, выполнять запросы. Программа имеет мощные и гибкие средства графической визуализации разметки и синтаксических связей, которые описываются внешним образом в xml-файле. Это тоже своего рода «метаданные», относящиеся к интерфейсу. С их помощью можно визуализировать в тексте результаты запроса и связанные объекты, строить графы иерархических зависимостей (синтаксические деревья, словоуказатели) и т. д. Фрагменты окон работающей программы приведены на рисунках 1, 2 и 3.

Текущая версия системы реализована на платформе Microsoft .NET Framework с использованием реляционных баз данных Microsoft Access и Microsoft SQL Server.

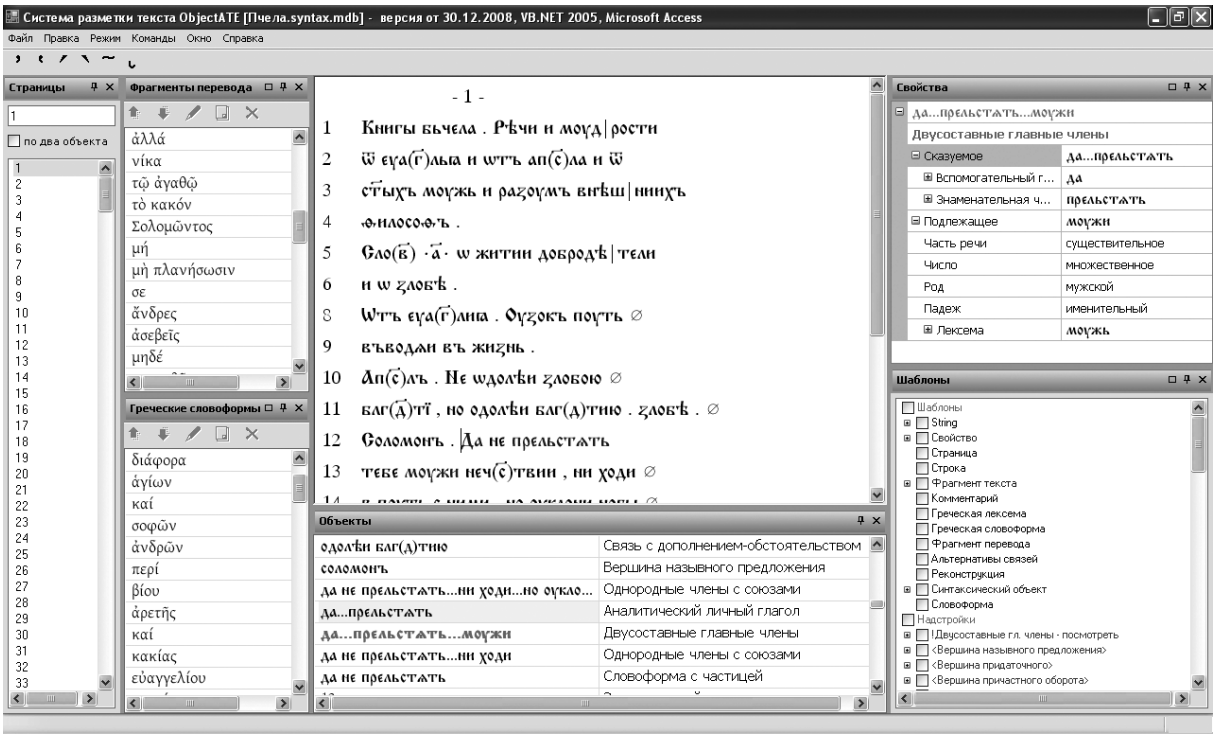


Рис. 1. Панели графического интерфейса пользователя программы ObjectATE

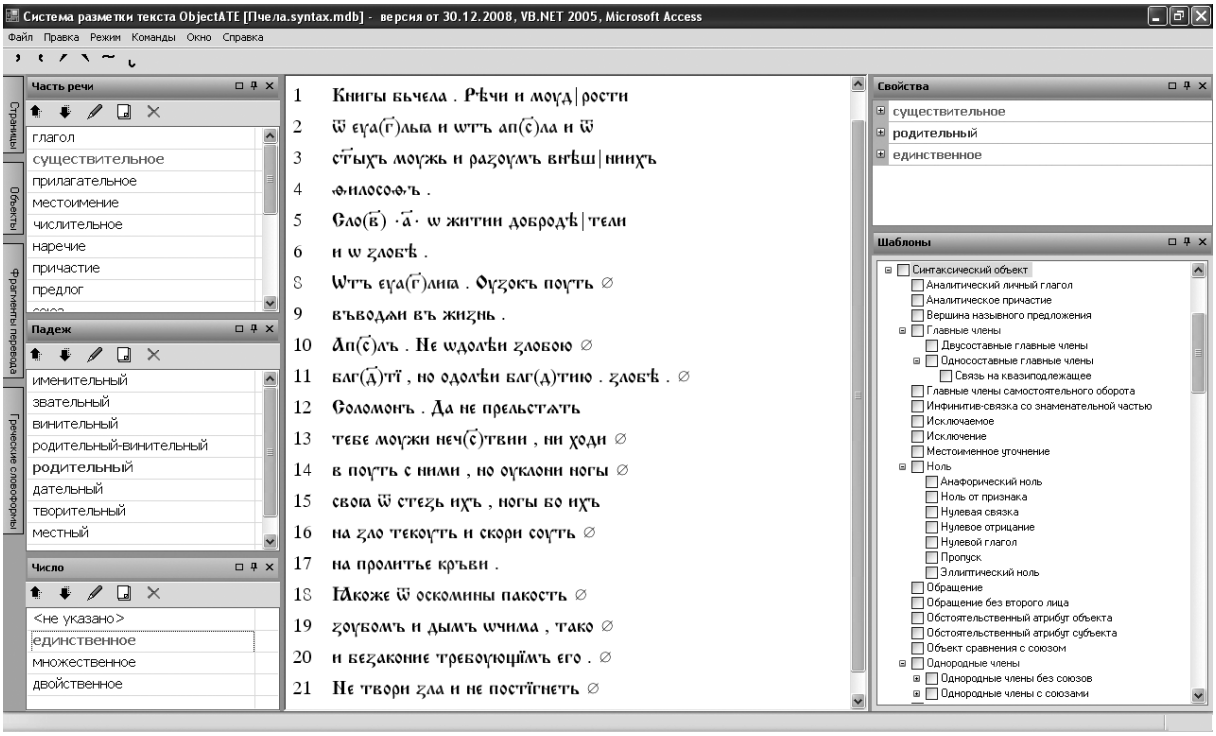


Рис. 2. Выполнение простейших запросов с помощью подсветки
(запрос на существительные в единственном числе и в именительном падеже)

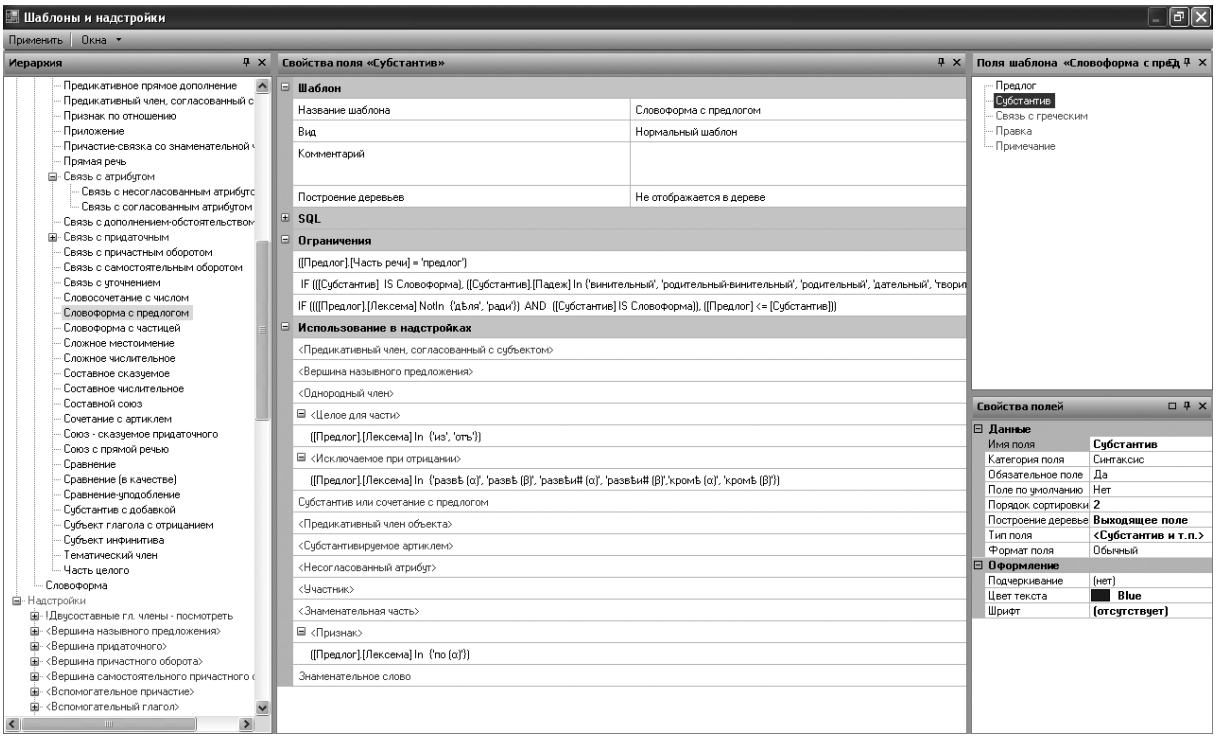


Рис. 3. Диспетчер шаблонов и надстроек.
Показаны свойства шаблона «Словоформа с предлогом»

Литература

1. *Зобнин А. И., Маркелова А. В.* Универсальная система разметки текста ATE-2 // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам. Материалы международной научной конференции. Ижевск, 2006. С. 51–55.
2. *Зобнин А. И., Маркелова А. В.* Универсальная система разметки текста ObjectATE // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам. Материалы международной научной конференции. Казань, 2008. С. 114–117.
3. *Сахарова А. В.* Возможности применения универсальной системы синтаксической разметки текста ObjectATE // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам. Материалы международной научной конференции. Казань, 2008. С. 247–249.
4. *Зобнин А. И., Пичхадзе А. А.* Корпус древнерусских переводов XI–XII веков: результаты и перспективы // Научно-техническая информация. Информационные процессы и системы. М., 2005. Сер. 2, № 3, С. 44–47.
5. *Зобнин А. И., Сахарова А. В.* Универсальная система разметки текста ObjectATE // Национальный корпус русского языка: 2006–2008. Результаты и перспективы. М., 2008. С. 283–297.

Терминология быта. Поиски нормы[1]

Everyday terminology. In pursuit of standards

Иомдин Б. Л. (iomdin@ruslang.ru)

Институт русского языка им. В. В. Виноградова РАН

Работа посвящена лексике, описывающей бытовые артефакты. Показано, что ее представления в словарях, в нормативных документах и в узусе существенно различаются, а единые толкования принципиально невозможны. Предложен проект составления толково-энциклопедического словаря-тезауруса бытовой терминологии.

Вводные замечания

Предметная лексика, в частности слова, называющие различные созданные человеком материальные элементы быта, составляет большую часть нашего активного словаря. Однако современная теоретическая семантика и лексикография обычно уделяют такой лексике меньше внимания, чаще сосредоточиваясь на предикатной лексике, словах, описывающих внутренний мир человека, лексемах с нетривиальной сферой действия и другом материале, обладающем лингвистически интересными особенностями. Например, в НОСС из 354 статей лишь несколько посвящены предметной лексике (*дом*, *забор*, *колокольчик*, *памятник* и отчасти некоторые другие). В то же время такая лексика также представляет собой достойный изучения материал и ставит перед лексикографом непростые задачи. В настоящей работе рассмотрены три группы слов, относящихся к разным предметным областям (посуда, галантерея и одежда). На основе анализа этого материала показано, какие проблемы возникают при описании такого рода лексики, и предложен подход к ее адекватному представлению в словаре. Материалом работы послужили толковые словари русского языка, официальные нормативные документы, корпусы текстов, результаты текстового и мультимедийного поиска в интернете и опросы носителей.

[1] Работа выполнена при финансовой поддержке Программы фундаментальных исследований отделения историко-филологических наук РАН «Генезис и взаимодействие социальных, культурных и языковых общностей» и гранта НШ-3205.2008.6 для поддержки научных исследований, проводимых ведущими научными школами РФ. В работе использован Национальный корпус русского языка (www.ruscorpora.ru). Автор также благодарит за помощь в подготовке доклада В. И. Беликова, Л. Л. Иомдина и Ю. В. Халееву.

1. Материал словарей

1.1. Посуда

В этой группе рассматриваются три слова с близким значением: *рюмка*, *бокал* и *фужер*. Вот их толкования, приводимые в девяти авторитетных толковых словарях русского языка:

	СУш	**БАС**	**МАС**	**БТС, НБАС**[2]	**СШ, СОШ**[3]	**ЛЛ**	**ТСИ**
рюмка	Стеклянный суживающийся книзу стаканчик на ножке, употр. преимущ. для спиртных напитков	Небольшой сосуд для вина, обычно стеклянный, на ножке	Небольшой, обычно стеклянный сосуд на ножке, употребляемый для питья спиртных напитков	БТС: Небольшой, обычно стеклянный сосуд на ножке для питья спиртных напитков	Небольшой на тонкой ножке сосуд для вина	Небольшой сосуд на ножке, употр. обычно для питья спиртных напитков	НЕТ
бокал	Посуда для вина, похожая на рюмку, но большего размера	Высокий стеклянный или металлический сосуд для вина на ножке в форме рюмки, но большего размера	Сосуд для вина в форме рюмки, но большего размера	БТС, НБАС: Сосуд для вина в виде большой рюмки	Сосуд для вина в виде большой рюмки	Сосуд для вина больших размеров, чем рюмка	1. Сосуд для вина в виде большой рюмки 2. прост. Высокая чашка цилиндрической формы без ручки
фужер	Бокал для вина	Большая и широкая рюмка, используемая обычно для шипучего вина, прохладительных напитков	Высокий бокал для прохладительных напитков, вина и т.п.	БТС: Большой бокал на высокой ножке для прохладительных напитков, шипучих вин и т.п.	Большой бокал на высокой ножке	Большой бокал на высокой ножке	Большой бокал на высокой ножке

1.2. Галантерея

Ниже приводится аналогичная таблица для слов *бумажник*, *кошелек* и *портмоне*.

	СУш	**БАС**	**МАС**	**БТС, НБАС**	**СШ, СОШ**	**ЛЛ**	**ТСИ**
бумажник	Карманный портфельчик для бумажных денег и мелких бумаг	Род кожаной или матерчатой карманной сумки для ношения и хранения деловых бумаг, бумажных денег и т. п.	Карманный портфельчик для бумажных денег и документов	БТС: Мужской плоский складывающийся книжечкой кошелек с несколькими отделениями для бумажных денег, документов и т.п. НБАС: Складывающийся плоский кошелек с отделениями для бумажных денег, документов и т.п.	Складывающийся вдвое карманный плоский портфельчик с несколькими отделениями, без ручки и обычно без запора, для ношения бумажных денег, документов	Карманная папочка (из кожи и т. п. материалов) с отделениями — вместилище для бумажных денег и мелких бумаг	НЕТ
кошелек	Мешочек, чаще кожаный, с металлическим затвором, для ношения денег в кармане	Мешочек, сумочка для денег	Мешочек или карманная сумочка для денег	БТС: Мешочек или карманная сумочка для денег. НБАС: Небольшая сумочка для денег. // устар. Кожаный, вязаный и т.п. мешочек для денег	Небольшая с запором [СОШ: карманная] сумочка для денег	Небольшая, обычно карманная сумочка для денег	НЕТ
портмоне	Небольшой кошелек для денег	Небольшой кошелек для денег	устар. Небольшой кошелек для денег	БТС: Кошелек, бумажник	Небольшой кошелек	НЕТ	устар. Кошелек с несколькими отделениями для денег и документов

[2] Пока опубликованы только первые 9 томов НБАС. Имеющиеся в НБАС толкования близки к толкованиям БТС и потому помещены в тот же столбец.

[3] Толкования в СОШ и СШ всех рассмотренных слов практически идентичны; два имеющихся различия мы специально отметили.

1.3. Одежда

Ниже приводится аналогичная таблица для слов, называющих одежду для верхней части тела.

	СУш	**БАС**	**МАС**	**БТС, НБАС**	**СШ, СОШ**	**ЛЛ**	**ТСИ**
свитер	Теплая вязаная фуфайка без застежек с высоким загнутым воротником, надеваемая через голову	Вязаная фуфайка без застежек с высоким воротником, плотно обтягивающая торс и шею	Теплая вязаная фуфайка без застежек с высоким воротником, надеваемая через голову	БТС: Теплая вязаная фуфайка без застежек с высоким воротником, надеваемая через голову	Теплая вязаная фуфайка без застежек с высоким воротом	Предмет теплой вязаной одежды для верхней части тела, без застежек, с высоким воротом	Теплая вязаная фуфайка без застежек, надеваемая через голову
джемпер	(нов.). Вязаная шерстяная или шелковая кофта без застежек, надевающаяся через голову и плотно облегающая корпус	Вязаная кофта, фуфайка, без воротника, надеваемая обычно через голову и плотно облегающая фигуру	Вязаная кофта без воротника и без застежек, надеваемая через голову	БТС: Тонкий вязаный свитер без воротника (обычно с вырезом на груди) и без застежек НБАС: Вязаная кофта без воротника и застежек, надеваемая через голову	Вязаная фуфайка без воротника, надевающаяся через голову	Предмет вязаной одежды для верхней части тела без воротника и застежек	Вязаная фуфайка без воротника, надеваемая через голову
пуловер	Трикотажная фуфайка, без воротника и без застежек, плотно облегающая корпус	Род свитера без воротника	Трикотажная фуфайка без воротника и без застежек, плотно облегающая фигуру	БТС: Род джемпера (обычно трикотажного или вязаного) без воротника и без застежек	Вязаная теплая [СОШ: трикотажная] фуфайка без воротника и застежек	Вязаный предмет одежды для верхней части тела, без воротника, надеваемый поверх рубашки	Вязаная фуфайка без воротника и без застежек
кофта	Верхняя часть женского костюма. || Короткое теплое женское пальто (простореч.).	1. Короткая верхняя женская одежда, обычно просторная || Женская одежда, надеваемая на ночь поверх рубашки 2. Простореч. Короткое теплое женское пальто. 3. Короткая верхняя одежда, часть национального костюма китайцев, японцев и др.	1. Короткая (обычно до пояса или до бедер) женская одежда 2. Прост. Короткое теплое женское пальто	БТС, НБАС: Короткая (обычно до пояса или до бедер) женская или детская одежда [БТС: носимая с юбкой или брюками] [НБАС: // устар. Домашняя женская одежда, надевавшаяся на ночь поверх рубашки // Короткая просторная мужская одежда из мягкой ткани]	1. Короткая женская одежда, обычно носимая с юбкой или брюками. *Ночная к.* (надеваемая поверх ночной рубашки) 2. Короткая просторная верхняя одежда (устар)	Предмет одежды, преимущ. женской, для верхней части тела, надеваемый поверх блузки, рубашки (в 1 знач.), платья (во 2 знач.)	Короткая женская одежда с застежкой, вязаная или из ткани, которую носят обычно с юбкой или брюками
кофточка	Кофта из легкой ткани	Уменьш.-ласк. к *кофта* (в 1-м и 2-м знач.)	Уменьш. к *кофта*; кофта из легкой ткани	БТС, НБАС: Уменьш.к *кофта* [БТС: 2. Блуза из легкой ткани].	уменьш. от *кофта*.	1. см. *кофта*. 2. Блузка.	1. уменьш. от *кофта*. 2. разг. то же, что кофта.
фуфайка	Теплая вязаная рубашка без рукавов или с рукавами, одеваемая (Sic!) вниз для тепла или надеваемая сверху	1. Теплая вязаная, шерстяная или байковая рубашка, безрукавка. || Майка с рукавами. 2. Стеганая куртка	1. Теплая вязаная рубашка или безрукавка 2. Стеганая ватная куртка; ватник	БТС: 1. Теплая вязаная шерстяная или байковая рубашка или безрукавка; свитер 2. Стеганая ватная куртка; ватник	1. Теплая вязаная рубашка. 2. То же, что ватник (прост.)	Теплый вязаный предмет одежды для верхней части тела	НЕТ
водолазка	НЕТ	НЕТ	НЕТ	БТС: разг. Тонкий обтягивающий джемпер с высокой горловиной НБАС: разг. Тонкий обтягивающий тело свитер	разг. Тонкий обтягивающий свитер	Легкий свитер, носимый вместо рубашки или блузки	НЕТ
банлон	НЕТ	НБАС: 1. Разновидность полиамидного синтетического волокна; тонкий эластичный трикотаж из такого волокна. 2. Тонкий свитерок с глухим стоячим воротником из такой ткани; водолазка.		БТС: 1. Разновидность полиамидного синтетического волокна; тонкая, легко растягивающаяся ткань из такого волокна. 2. Джемпер с высоким стоячим воротником из такого материала; водолазка.	НЕТ	НЕТ	НЕТ

2. Материал нормативных документов

В энциклопедии и энциклопедические словари рассматриваемая нами лексика не включается[4], однако частично описывается в официальных стандартах. ГОСТа для первой группы понятий нам найти не удалось, а две других содержатся в нормативных документах. Приведем соответствующие фрагменты:

Бумажник — изделие мелкой кожгалантереи для документов и бумажных денег.

Кошелек — изделие мелкой кожгалантереи для монет.

Портмоне — изделие мелкой кожгалантереи для бумажных денег и монет. [ГОСТ 2000].

Свитер — трикотажная плечевая одежда с длинными рукавами, без застежки, с высоким воротником (более 5 см), покрывающая туловище и частично бедра.

Джемпер — трикотажная плечевая одежда с рукавами, без застежки или с застежкой вверху, покрывающая туловище и частично бедра.

Пуловер — разновидность джемпера.

Кофта — недопустимый термин.

Кофточка — швейная или трикотажная плечевая одежда с рукавами, застежкой, покрывающая туловище частично или полностью, для новорожденных, детей ясельных или дошкольных групп.

Фуфайка — трикотажная плечевая одежда, покрывающая туловище частично или полностью, надеваемая на корсетные изделия или непосредственно на тело.

Водолазка — НЕТ. [ГОСТ 1988].

3. Анализ материала

Легко видеть, что рассмотренные словарные толкования обладают рядом недостатков. Отсутствует система в выборе родовых слов: так, в СУш три очень похожих предмета толкуются через три разных genus proximum (‘стаканчик’, ‘посуда’ и ‘бокал’); в нескольких словарях *свитер* толкуется через ‘фуфайку’, а очень близкий по значению *джемпер* — через ‘кофту’; *кошелек* толкуется как ‘небольшая сумочка’, а *портмоне* — как ‘небольшой кошелек’ и т.п. Слова, используемые в толкованиях, часто имеют менее прозрачное и определенное значение, чем сами толкуемые слова (ср. два значения у *фуфайки* и явную стилистическую отмеченность и несовременность этого слова). В разных толкованиях одно и то же называется по-разному (ср. *ворот* и *воротник* в толкованиях ЛЛ). Кроме того, эти толкования не соответствуют официальной терминологии, содержащейся в нормативных документах. Так, ни в одном словаре для рассмотренной одежды нет различительного признака «наличие и длина рукавов»; по ГОСТу у *джемпера* допускается наличие застежки; слово *кофта* ГОСТ запрещает вообще и использует лишь слово *кофточка* (которое отсутствует в ТЭС, самом большом справочном издании России, объемом в 147 000 слов); в [ГОСТ 2000] у *бумажника*, *кошелька* и *портмоне* нет различий по размеру, форме и наличию замка, а *кошелек* считается использующимся только для ношения монет. Однако наиболее интересно то, что ни толкования словарей, ни определения ГОСТов не соответствуют современному узусу.

3.1. Посуда

Так, *рюмками* чаще называют небольшие емкости (в том числе и без ножки), если они используются для водки или других крепких спиртных напитков. Ср. толкования на специальных сайтах, посвященных этикету: *Рюмка — это небольшой, чаще всего стеклянный или хрустальный сосуд на ножке для крепкого спиртного*; *Рюмка — используется для алкогольных напитков, в основном, для водки. Гораздо меньше бокала и фужера*. В [Черникова 2003] *рюмкой* называется посуда для коньяка, ликеров, коктейлей и крепких вин. Поиск в НКРЯ *рюмки вина* дает 96 (12) примеров, *рюмки водки* — 535 (167), *рюмки коньяка <коньяку>* — 202 (116) (в скобках даны цифры по текстам, созданным после 1990 года, и соотношение еще более убедительно). У *бокала* соотношение обратное: 152 (93) — 8 (8) — 8 (7). Поиск на сайте журнала «Винομания» дает 8 упоминаний *рюмок* (для водки, текилы и коктейлей) и 90 упоминаний *бокалов*. На сайте «Журнала о людях и вине» *рюмка* упоминается 15 раз (в сочетаниях *ликерная <коньячная> рюмка* и *рюмка водки*), а *бокал* — 277 раз. *Фужер* в текстах упоминается в основном как емкость для шампанского. Поиск в интернете изображений *рюмки* выдает большое количество фотографий и рисунков емкостей, не имеющих ножки; бокалы и фужеры без ножек в результатах соответствующего поиска практически не встречаются.

Таким образом, можно выделить следующие различительные признаки, релевантные для этой группы слов в узусе:

	Основная функция	**Размер**	**Наличие ножки**
рюмка	для крепких спиртных напитков	небольшой	необязательно
бокал	обычно для вин	средний или большой	обязательно
фужер	обычно для шампанского	средний или большой	обязательно

[4] За исключением ТЭС, толкования которого близки к толкованиям БТС.

3.2. Галантерея

Если по словарям *портмоне* отличается от *кошелька* и *бумажника* наименьшим размером, то в узусе этого различия нет (что ярко демонстрируют, в частности, результаты поиска изображений в интернете и изучение каталогов магазинов и фирм-производителей кожгалантереи). Вопреки пометам устар. в МАС и ТСИ, это слово сейчас очень распространено, но указывает на то, что обозначаемый предмет более дорогой, модный, стильный; ср. *Конечно, всем знакомое слово «кошелек» звучит для иного уха по-деревенски. Куда привлекательнее французское словечко, хотя и с тем же самым значением* («Карелия», №134, 19.12.2002); *Слово «кошелек» уже недостаточно модно? Купите портмоне и блесните новым модным аксессуаром на публике* (реклама). Несмотря на происхождение этого слова (от франц. *porte-monnaie* '[предмет для] ношения монет'), его первоначальное значение ('кошелек для ношения денег (в кармане), особенно металлических (монет)' [Черных 1999], 'нечто среднее между бумажником и кошельком, для мелких денег' [Смирнов 1908]) и требования ГОСТа, сейчас портмоне содержат несколько отделений для кредитных карточек, удостоверений и т.п. Ср. рекламные тексты: *Кожаные портмоне — это вместительный, удобный кошелек на все случаи жизни. В нем есть отделения для купюр, кредитных карточек и всякой мелочи, которая обычно проживает в кармане*; *Портмоне — это уникальное и наиболее полезное изобретение человечества, так как позволяет хранить все самое необходимое в компактном мини-портфеле. В нем найдется место для всего, начиная от бумажных купюр и звонких монет, заканчивая мед. полисом, техосмотром, правами и другими визитками и карточками*.

Для данной группы слов релевантными оказываются следующие различительные признаки:

	Функция: хранение и ношение…			**Стиль**
	монет	**купюр**	**документов**	
бумажник	нет	да	да	любой
кошелек	да	да	нет	любой
портмоне	да	да	да	модный

3.3. Одежда

Вопреки словарям и ГОСТу, застежка свитеру никак не противопоказана: *свитер на молнии* встречается в Яндексе[5] 25 000 раз. При этом *свитер* необязательно имеет ворот, что также хорошо видно при поиске изображений[6]. Поиск в Яндексе *свитера с воротом* дает больше тысячи результатов, *свитера с воротником* — больше 3000 (большую часть результатов составляют разделы каталогов одежды) — это было бы странно, если бы ворот был неотъемлемой частью любого свитера; ср. необычность *холодильник с дверцей <с дверью>* (меньше 100 результатов). *Пуловер* и *джемпер*, различия которых никак не описаны в ГОСТе и в большинстве словарей (даже в БТССин, содержащем синонимический ряд «*джемпер, пуловер*», между ними не усмотрено ни одного различия), в узусе обычно различаются наличием выреза и его формой. Правда, различия при этом иногда описываются противоположным образом; ср. толкования с интернет-форумов, посвященных вязанию и моде: *Классический пуловер — это тот же джемпер с вырезом и без застежки*; *Отличие пуловера от джемпера чисто академическое — горловина выполнена в виде полукруга, а не галочки*; *Пуловер — как джемпер, но с круглым вырезом горловины*; *Джемпер — с круглой горловиной «под горлышко», пуловер — с v-образной*; *Пуловер — это как бы джемпер, имеющий низкий вырез углом*. Наряду с женскими широко распространились и мужские *кофты*. *Фуфайками*, в словарях описываемыми как теплая одежда, в модных магазинах сейчас обычно называют легкие футболки.

Кратко приведем наиболее яркие данные проведенного нами эксперимента, в рамках которого 70 респондентов, средний возраст 27 лет, живущие преимущественно в Москве и Санкт-Петербурге, а также в других городах, описывали изображения по-разному одетых людей. Цель эксперимента не была известна его участникам, поэтому можно предположить, что многие из них не задумывались специально о выборе наиболее точной номинации для каждого предмета одежды. Для 25 из 26 изображений, найденных нами в интернете по ключевым словам *свитер, кофта, джемпер, пуловер, водолазка* нашелся хотя бы один носитель, назвавший его *свитером* (несмотря на наличие или отсутствие застежек, ворота, выреза); при этом у половины предметов номинация *свитер* по частотности оказалась на первом месте, у остальных — на втором (это верно даже для изображения собаки в специальной одежде для животных). У 72% респондентов наиболее частое слово для описания этих предметов — *свитер* (у 18% — *кофта*, у 2% — *джемпер*). Всего слово *свитер* встретилось 524 раза, *кофта* — 312, *водолазка* — 185, *джемпер* — 50, *пуловер* — 32, остальные слова — менее 20 раз (так, слово *фуфайка*, ключевое для словарей, встретилось всего

[5] Здесь и далее дается цифра с первой страницы выдачи на 31.03.2009; конечно, на нее можно ориентироваться лишь для определения порядка величины.

[6] Интересно, что в словарной статье Википедии, где *свитер* толкуется как «предмет вязаной одежды для верхней части тела без застежек, с длинными рукавами и характерным высоким двухслойным воротом», среди иллюстраций есть значительное количество свитеров с застежками и/или без ворота.

2 раза). Из всех встретившихся слов наиболее четко выделяется *водолазка*: для каждого предмета это название (или его петербургский аналог) выбирали либо очень часто (более 75% опрошенных), либо очень редко (менее 5%). Все названные так предметы имеют ворот, не имеют застежки и надеваются через голову. Остальные наименования образуют достаточно диффузную зону. С помощью регрессионного анализа наших данных[7] выявились следующие различительные признаки: *свитер* достаточно толстый и надевается через голову; *джемпер* не имеет ворота; *пуловер* надевается через голову, не имеет выреза и нехарактерен как мужская одежда; *кофта* не надевается через голову. Все остальные признаки оказались несущественными. Для описания одежды мужчин слово *кофта* было использовано 178 раз (напомним, что по словарям *кофта* — предмет женской одежды)[8].

4. Проблемы

4.1. Изменения жизни

Обсуждаемые слова относятся к лексическому пласту, постоянно и быстро меняющемуся, и особенно это касается названий одежды и аксессуаров, подверженных влиянию моды. В [Смирнов 1908] отсутствуют *свитер*, *джемпер*, *пуловер*, а из названий одежды, не отмеченных как устаревшие или специальные, есть *бекеша*, *канзу*, *ватерпруф*, *кафтан*, *салоп* (все эти слова сейчас не употребляются). Среди названий одежды, получивших широкое распространение в последние годы, — *топ*, *боди*, *комбидрес(с)*, *шраг* (последнее слово отсутствует во всех словарях, включая ТЭС и Скляревская 2006). В некоторых словарях находим слово *монетница*, истолкованное как ‘коробка, род кошелька с отделениями для монет различного достоинства’. В НКРЯ оно не встретилось ни разу (!), однако поиск в интернете показывает, что сейчас это слово распространилось в новом значении: ‘фирменная кассовая тарелочка под мелочь. Один из видов печатной сувенирной рекламной продукции’ [Стефанов 2004]. Появляются новые предметы, для которых еще не существует устоявшихся названий. Так, в словарях отсутствует слово *маниклип* (‘зажим для денег’, современная модная замена кошельку). Предмет для ношения кредитных, клубных, дисконтных карточек еще не получил никакого названия, хотя говорящие ощущают такую необходимость; ср. *Вообще-то это не кошелек, а скорее сумочка для хранения документов и карт (товарищи, дайте идею названия, ничего в голову не приходит)*; *Бумажник для карточек. Или это не бумажник, короче, просто держатель карт. … И куда податься за такой зверушкой в Москве?*; *Это не бумажник… Это книжка с кредитками… Так что это пластикник…* [с интернет-форумов].

4.2. Социально обусловленные различия

Среди обсуждаемой лексики много слов, имеющих хождение лишь в определенных регионах. В Словаре «Языки городов» (см. также Беликов 2004, 2009) подавляющее большинство слов относится именно к названиям разного рода артефактов. В частности, в нем представлено региональное значение слова *бокал* (‘фаянсовый или стеклянный сосуд в форме стакана с ручкой для холодных и горячих напитков’), региональные синонимы слов *кошелек* (*гомонок*) и *водолазка* (*банлон*[9], *битловка*, *гольф*, *роллинг*). Некоторые слова имеют еще более ограниченное распространение, например, в определенном социолекте или в семейном лексиконе (ср. Занадворова 2000).

4.3. Размытость значений

Вообще, существование множества предметов с похожими функциями и внешним видом неизбежно приводит к тому, что разные говорящие называют их по-разному или владеют не всеми названиями. Различия семейных лексиконов и идиолектов часто вызывают споры (ср. характерные реплики в интернет-форумах, посвященных обсуждению художественной фотографии: *Это ж не рюмка, а бокал для мартини*; *Вроде это не рюмка, а бокал или фужер*; *Какой же это бокал, это рюмка — бокал на тонкой ножке*; *Это не рюмка, а бокальчик*[10]). Множество тем на форумах посвящено обсуждению вопросов вроде *Что такое пуловер? Что такое джемпер? Чем отличаются? Хочу выяснить*; *Запуталась в названиях одежды, помогите разобраться*; *Я занимаюсь составлением каталога одежды и постоянно мучаюсь*; *Подскажите, пожалуйста, чем отличается кошелек от портмоне?* Знание различий между обсуждаемыми словами часто оце-

---

[7] Исследование проведено Ю. В. Халеевой.

[8] Полностью данные эксперимента доступны на izjumis.livejournal.com. Автор приносит благодарность всем участникам эксперимента.

[9] О слове *банлон/бадлон/бодлон*, приводимом без помет лишь в словарях, издаваемых в Санкт-Петербурге, см. подробнее в [Беликов 2009]. В нашем эксперименте это слово употребляли исключительно жители Санкт-Петербурга, при этом некоторые из них также употребляли и слово *водолазка*.

[10] Последние две реплики свидетельствуют о том, что в идиолекте их авторов существенным является отнюдь не различие в размере, представляемое словарями как основное.

нивается говорящими как специальное и сложное: *Нам, Денис, этого не понять. Я тоже между свитером и кофтой не вижу разницы*; *Ничего себе, какие у тебя, однако, глубокие познания! Может, скажешь еще, чем отличается свитер от пуловера?*; *Получала вещи из химчистки — меня отругала приемщица. Она мне объяснила, что мы с ее сменщицей неверно заполнили бланк. Я тогда записала все ее рекомендации по названиям одежды.* Ситуация осложняется тем, что словоупотребление специалистов часто отражает их профессиональный жаргон или какие-либо служебные инструкции и дезориентирует носителей литературного языка[11]. Если во многих других случаях (например, при изучении той или иной науки) можно овладеть необходимой лексикой, обратившись к словарям или энциклопедиям, то в сфере бытовых реалий непонятно, где именно содержится «истина в последней инстанции»[12].

5. Выводы

5.1. Несводимость к единым толкованиям

Представляется, что попытки свести воедино данные словарей, предписания нормативных документов и предпочтения узуса не окажутся плодотворными: получившиеся описания будут чересчур подробными, вариативными или максимально неконкретными, не отразят ничьего реального словоупотребления и не будут иметь никакого практического применения.

5.2. Доминанта

Несмотря на то, что в номенклатуре каждое название должно соответствовать своему референту, представляется, что в литературном языке группы слов, описывающих сходные предметы, как и ряды синонимов, имеют доминанту (ср. о ней, например, НОСС): в данном случае это не только наиболее употребительное слово, но и такое, которым можно заменить остальные слова во всех контекстах, где не подчеркиваются различия между соответствующими предметами. Так, в рассмотренных группах доминантами являются слова *рюмка*, *свитер* и *кошелек*. Они более частотны[13], особенно в неспециальных текстах последнего времени. Показательно также употребление в текстах словарных речений: из всех обсуждаемых слов в речениях ЛЛ 20 раз встречается *свитер*, 10 раз *кошелек*, 2 раза *рюмка*, 1 раз *кофта*; в речениях СОШ похожая ситуация. Именно доминанты предпочитаются в некоторых типах контекстов. Скажем, при указании на «квант» спиртного *рюмка* естественнее, чем ее синонимы; ср. *хватить по рюмке*, но *?хватить по бокалу <??по фужеру>*, *пьянеть от одной рюмки <?от одного бокала, ??от одного фужера>*. Рекомендация *взять с собой теплый свитер* выглядит значительно естественнее и встречается в десятки раз чаще, чем та же рекомендация с упоминанием *кофты*, *джемпера* или *пуловера*. Характеристика *на любой кошелек* встречается в Яндексе 59 000 раз, *на любой бумажник* — 14 раз, *на любое портмоне* — ни разу. У доминант шире сочетаемость: так, в НКРЯ глагол *натянуть* 7 раз сочетается со словом *свитер* и ни одного раза — со всеми остальными словами этой группы. От доминант гораздо легче образуются и чаще употребляются диминутивы: ср. 795 упоминаний в НКРЯ *рюмочек* против 49 *бокальчиков* и 9 *фужерчиков*; 82 *свитерка* и только 2 *джемперка*, а *пуловерок <пуловерчик>* и *водолазочка* отсутствуют вовсе[14]. У доминант более развита фразеология (ср. *на рюмку чая, тугой кошелек, кошелек или жизнь*), именно от них чаще образуются новые слова (ср. *интернет-кошелек*, но не **интернет-бумажник* или **интернет-портмоне*). Ср. также очевидное преобладание номинации *свитер* в результатах нашего эксперимента.

Ни в одном словаре, насколько нам известно, информация о доминанте в таких случаях не приводится. Между тем кажется, что при учете ее существования многие тексты можно понять правильнее, не приписывая им более узкой интерпретации, чем та, которую имел в виду автор; скажем, упоминание *свитера* совсем не обязательно подразумевает наличие ворота и отсутствие застежки. В таком упрощенном режиме употребления рассмотренные слова имеют приблизительно следующие, квазигиперонимические значения: *свитер* — ‘вязаный или сходный предмет неуличной одежды для торса’, *рюмка* — ‘емкость, предназначенная для питья спиртных напитков’, *кошелек* — ‘переносной предмет, предназначенный для хранения денег’.

---

[11] Еще один яркий пример несоответствия бытового и профессионального словоупотребления представляет собой пара *рубашка/сорочка*. Согласно ГОСТ 1988, *рубашка* — недопустимый термин (правильные термины — только *нижняя сорочка*, *верхняя сорочка* и *рубашечка*). В нашем эксперименте слово *рубашка* было употреблено респондентами для описания картинок 158 раз, слово *сорочка* — 2 раза.

[12] В случаях, когда лексика не относится к сфере быта, сомнения и заблуждения неспециалистов также часто носят массовый характер (ср. удивительную распространенность ошибочного употребления слова *литавры* [Левонтина 2004]), однако здесь мнение специалистов имеет безусловный приоритет.

[13] Так, из всей рассмотренной группы слов-обозначений одежды только *свитер* попал в список самых частотных, ср. [Ляшевская, Шаров 2008].

[14] *Кофточек*, разумеется, значительно больше, но этот диминутив давно получил отдельное значение; ср. также разные значения *блузы* и *блузки*.

5.3. Проект словаря

Проведенное исследование, как кажется, говорит о целесообразности составления толково-энциклопедического словаря-тезауруса бытовой терминологии. Словарная статья такого словаря должна включать в себя по крайней мере следующие зоны: (1) группа слов с близким значением; (2) доминанта группы; (3) различительные признаки, релевантные для данной группы; (4) данные нормативных документов, если они существуют; (5) предпочтения узуса; (6) информация о наличии региональных различий и особенностях разных социолектов; (7) иллюстрации — фотографии или схематические изображения. При этом необходимо в каждом семантическом описании выявлять ядро и периферию и по возможности четко различать профессиональное и общеязыковое употребление. Такой словарь мог бы быть востребован и специалистами, составляющими каталоги товаров или, скажем, таможенные декларации[15], и создателями или пользователями баз данных, например, при подборе товара по параметрам (ср. Яндекс.Маркет), и переводчиками при подборе наиболее адекватных эквивалентов, и всеми, кто изучает язык (как иностранный или как родной) и хочет получить правильное и всестороннее представление о современном словоупотреблении.

[15] Ср. «крик души» на одном из интернет-форумов: *Помогите, пожалуйста, срочно — вопрос жизни и смерти. Делал инвойсы для официальной растаможки одежды и в русском варианте перевел jackets как жакеты. А при проходе границы провели досмотр и выяснили, что это куртки… Чем это чревато? Что нужно делать? И вообще где прописано, чем жакеты от курток отличаются? Вот например — кто мне объяснит, чем джемпер от свитера отличается?*

Литература

1. *БАС* — Словарь современного русского литературного языка в семнадцати томах. М.–Л.: Изд-во АН СССР, 1950–1965.
2. *Беликов* 2004 — Беликов В. И. Сравнение Петербурга с Москвой и другие соображения по социальной лексикографии // Русский язык сегодня. Вып. 3: Проблемы русской лексикографии. М.: Ин-т рус. яз. РАН, 2004. С. 23–38.
3. *Беликов* 2009 — Беликов В. И. Стереотипы в понимании литературной нормы // Стереотипы в языке, коммуникации и культуре. М.: РГГУ, 2009. С. 339—359.
4. *БТС* — Большой толковый словарь русского языка / Сост., гл. ред. С. А. Кузнецов. СПб.: Норинт, 1998.
5. *БТССин* — Большой толковый словарь синонимов русской речи / Под ред. Л. Г. Бабенко. М.: АСТ-Пресс, 2008.
6. *ГОСТ* 1988 — Государственный стандарт Союза ССР. Изделия швейные и трикотажные. Термины и определения. ГОСТ 17037-85. М.: Издательство стандартов, 1988.
7. *ГОСТ* 2000 — Межгосударственный стандарт. Изделия кожгалантерейные. Термины и определения. ГОСТ 28455-90. М.: Стандартинформ, 2000.
8. *Занадворова* 2000 — Занадворова А. В. Узус семейного речевого общения: особенности номинаций // Языковая личность: институциональный и персональный дискурс: Сб. науч. тр. Волгоград: Перемена, 2000.
9. *Левонтина* 2004 — Левонтина И. Б. Медный барабан // Еженедельный журнал, № 123, 8.06.2004.
10. *ЛЛ* — Лопатин В. В., Лопатина Л. Е. Русский толковый словарь. М.: Русский язык, 1997.
11. *Ляшевская, Шаров* 2008 — Ляшевская О. Н., Шаров С. А. Частотный словарь национального корпуса русского языка: концепция и технология создания // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). М.: РГГУ, 2008. С. 345–351.
12. *МАС* — Словарь русского языка в четырех томах. / Под ред. А. П. Евгеньевой. 2-е изд., испр. и доп. М.: Русский язык, 1981–1984.
13. *НБАС* — Большой академический словарь русского языка. Тт. 1–9. / Гл. ред. К. С. Горбачевич. М.–СПб.: Наука, 2004–2008.
14. *НОСС* — Новый объяснительный словарь синонимов русского языка. Второе издание, исправленное и дополненное. / Под общим руководством акад. Ю. Д. Апресяна. М.-Вена: Языки славянской культуры; Wiener Slawistischer Almanach. Sonderband 60, 2004.
15. *Скляревская* 2006 — Толковый словарь русского языка начала XXI века. Актуальная лексика / Под ред. Г. Н. Скляревской. М.: Эксмо, 2006.
16. *Словарь* «Языки городов» / ABBYY Lingvo Клуб, http://www.lingvo.ru/goroda/articles.asp
17. *Смирнов 1908 — В. Смирнов. Полный словарь иностранных слов, вошедших в русский язык. М., 1908.*

18. *СОШ* — Ожегов С.И., Шведова Н.Ю. Толковый словарь русского языка. Изд. 4-е. М.: Русский язык, 1997.
19. *Стефанов* 2004 — Стефанов С.И. Реклама и полиграфия: опыт словаря-справочника. М.: Гелла-принт, 2004.
20. *СУш* — Толковый словарь русского языка / Под ред. Д. Н. Ушакова. М., Гос. ин-т «Сов. энцикл.»; ОГИЗ; Гос. изд-во иностр. и нац. словарей, 1934–1940.
21. *СШ* — Толковый словарь русского языка с включением сведений о происхождении слов / Отв. ред. Н. Ю. Шведова. М.: Издательский центр «Азбуковник», 2007.
22. *ТСИ* — Крысин Л. П. Толковый словарь иноязычных слов. М.: Русский язык, 1998.
23. *ТЭС* — Толково-энциклопедический словарь. / Под ред. С. М. Снарской. СПб.: Норинт, 2006.
24. *Черникова* 2003 — Черникова О. Стеклянная посуда. Дипломная работа по специальности «повар, бармен, официант». Краснодар, 2003.
25. *Черных* 1999 — Черных П. Я. Историко-этимологический словарь русского языка. 3-е изд. М.: Русский язык, 1999.

Синтаксические корреляты просодически маркированных элементов предложения и их роль в задачах синтеза речи по тексту[1]

Syntactic correlates of prosodically marked elements of the sentence and their role in the tasks of text-to-speech synthesis

Иомдин Л. Л. (iomdin@iitp.ru)
Институт проблем передачи информации РАН им А. А. Харкевича, Москва

Лобанов Б. М. (lobanov@newman.bas-net.by)
Объединенный институт проблем информатики НАН Беларуси, Минск

Работа посвящена экспериментальному исследованию возможности использования синтаксического анализа письменного текста на начальном этапе алгоритма синтеза речи по тексту. Произведена попытка установить корреляции между элементами построенной автоматически синтаксической структуры предложения в виде дерева зависимостей, и просодически выделенными элементами этого предложения. Первые результаты эксперимента показывают, что данный подход имеет хорошие перспективы.

Введение

Синтез речи по тексту предполагает наличие автоматической процедуры формирования текущих контуров мелодии, силы звука, фонемной длительности и длительности пауз на основе анализа определенных свойств входного текста и его просодической разметки. Просодическая разметка текста заключается в его членении на синтагмы, разметке синтагм на акцентные единицы и маркировке интонационного типа синтагм в соответствии с определёнными правилами. В [1] и более подробно в [2] были описаны правила просодической разметки текста на основе его частичного синтаксического анализа (анализа словосочетаний) и указывалось, что в достаточно полной степени эта проблема может быть решена лишь с использованием глубокого синтаксического анализа. Приемлемой лингвистической теорией, на которой может строиться система такого анализа, представляется теория «Смысл ⇔ Текст» И. А. Мельчука (см., например, [3]).

В данной работе исследуются связи между элементами синтаксической структуры предложения и экспертной просодической разметкой предложений на примере текстов новостных телевизионных передач на русском языке. Исследование таких связей стало возможным благодаря разработке системы «ЭТАП-3» [4,5] и созданию базы данных «Интонация русских информационных текстов» [6]. Опираясь на относительно небольшой фрагмент этой базы данных, мы попытались с помощью системы ЭТАП-3 получить предварительные ответы на следующие вопросы:

1. Существуют ли статистически значимые синтаксические корреляты **просодического выделенния слов** в синтагмах?
2. Существуют ли статистически значимые синтаксические корреляты членения предложений на **предпаузальные и беспаузальные синтагмы**?
3. Существуют ли статистически значимые синтаксические корреляты особенностей членения предложений на предпаузальные и беспаузальные синтагмы, а также просодического

[1] Авторы благодарны Российскому фонду фундаментальных исследований (грант № 08–06–00373) и Белорусскому фонду фундаментальных исследований (грант № Ф08Р-016) за частичную финансовую поддержку настоящего исследования. Мы хотели бы также выразить свою признательность С.В. Кодзасову и Л.М. Захарову за предоставленную нам возможность использовать в процессе работы базу данных «Интонация русских информационных текстов».

выделения слов в синтагмах **для различных дикторов?**

4. Существуют ли статистически значимые синтаксические корреляты особенностей членения предложений на предпаузальные и беспаузальные синтагмы, а также просодического выделенния слов в синтагмах **для различных стилей речи?**
5. Если таковые закономерности существуют, то какова **числовая оценка их частотности**?

1. Синтаксический анализатор системы ЭТАП-3

Синтаксический анализатор, или парсер, многоцелевого лингвистического процессора ЭТАП-3 разработан в Лаборатории компьютерной лингвистики ИППИ РАН им А.А.Харкевича и используется в различных приложениях, в том числе в системе машинного перевода с русского языка на английский, в системе синонимического перифразирования, а также для построения синтаксически размеченного корпуса русского языка SynTagRus. [5. 7].

Этот парсер, опирающийся в значительной мере на уже упомянутую лингвистическую теорию «Смысл ⇔ Текст», строит для каждого предложения письменного текста его синтаксическую структуру (СинтС) в виде дерева зависимостей, т.е. связного ориентированного графа без циклов. Каждый узел такого дерева соответствует некоторому слову предложения, а его дуги помечены именами синтаксических отношений (СинтО).

В СинтС каждого предложения имеется единственная вершина, которой непосредственно или опосредованно подчиняются все остальные узлы. Имена СинтО эксплицируют различные типы синтаксических связей между словами; в текущей версии парсера используется свыше 65 различных СинтО. Например, связь между глагольным сказуемым в качестве вершины и именным подлежащим при нем в качестве зависимого члена (*мальчик ← читает*) представляется **предикативным СинтО**; связь между предикатным словом и первым дополнением при нем (*читает → книгу, чтение → книги*) представляется **1-ым комплетивным СинтО**; связь между существительным и определяющим его прилагательным (*детская ← книга*) оформляется **определительным СинтО**, а связь между глаголом и наречным обстоятельством (*читает → вслух*) задается **обстоятельственным СинтО.**

СинтС предложения, генерируемая парсером ЭТАП-3, является упорядоченным деревом зависимостей — оно сохраняет информацию о порядке следования слов в предложении.

Алгоритм русского синтаксического анализа обращается к лингвистическим ресурсам двух основных типов: набору бинарных синтаксических правил, или синтагм[2], и так называемому комбинаторному словарю, содержащему богатую и разнообразную информацию о каждом входящем в него слове. Парсер работает пофразно и может функционировать в нескольких режимах, в частности, 1) в полностью автоматическом дежурном режиме, при котором для каждого предложения строится ровно одна СинтС; 2) в режиме множественного анализа, когда пользователь может потребовать от системы построить для неоднозначного предложения несколько СинтС или даже все возможные СинтС; 3) в интерактивном режиме, когда в определенных точках алгоритма парсер, встретив неоднозначную лексическую единицу или омонимичную синтаксическую конструкцию, предлагает пользователю выбрать ту или иную морфологическую, лексическую и/или синтаксическую интерпретацию элементов предложения и тем самым направить работу по некоторому конкретному пути.

Система ЭТАП-3 в целом и ее синтаксический анализатор рассчитаны в первую очередь на тексты нейтрально-деловой прозы. Это, в частности, означает, что систему нецелесообразно применять к стилистически окрашенному материалу, к авторской художественной прозе, поэзии или же к разговорной речи.

2. Экспериментальный текстовый и аудиоматериал, отобранный для исследования

Ниже приводится экспериментальный текст, маркированный в соответствии с [3] знаками интонационной транскрипции (без указания тональных акцентов). Из общей базы данных [3] отобраны небольшие фрагменты, включающие записи (1–32) из новостных передач РТР и записи (33 — 37) из передач НТВ.

1. *[м1] Полтора часа *назад | из *Вены пришло *сенсационное *известие |, которое грозит *крупным международным *скандалом | и должно *повлиять | на судьбу арестованного в *Австрии | сотрудника || международного управления *РосКосмоса |||.*
2. *[м1] Австрийский *МИД | официально *признал |, что гражданин *России |, задержанный по подозрению в *шпионаже |, *имеет дипломатический *иммунитет |, а это *значит |, что по нормам международного *права | он не **мог быть *арестован |||.*

[2] Тем самым термин «синтагма» используется здесь иначе, чем это принято в литературе, посвященной автоматической обработке устной речи (в том числе и в настоящей статье).

3. *[ж1] Москва требует немедленного *освобождения| нашего *гражданина.*
4. *[ж1] Австрийскому *послу вручена *нота |||.*
5. *[ж1] А в официальном *заявлении| сказано, что этот *шаг || властей *Австрии| расценивается как *недружественный,| наносящий *ущерб| двусторонним *отношениям || и что он не *укрепляет авторитета *Австрии,| как места *расположения *штаб-квартир || ряда международных *организаций |||.*
6. *[ж1] *Сейчас| на прямом эфире из *Вены| к нам присоединяется наш *специальный *корреспондент| Иван *Родионов |||.*
7. *[ж1] Иван, *здравствуйте |||!*
8. *[ж1] *Как из всей этой неловкой *ситуации |собираются *выходить австрийские *власти |||?*
9. *[ж1] И *почему такая *проволочка | с *подтверждением ||правового *статуса нашего *соотечественника |||?*
10. ***[м2] *Здравствуйте, *коллеги |||.***
11. ***[м2] Действительно, *сегодня| наступило || *решающее *развитие| в ситуации вокруг *арестовано российского || *сотрудника || *РосКосмоса ||.***
12. ***[м2] *Сегодня || пришло *подтверждение|| от|| правового управления || *ООН|| в *Нью-йорке |||.***
13. ***[м2] Я *позволю себе| коротко процитировать эту *бумагу || со слов| *официального представителя || *австрийского| *МинЮста| Томаса *Тайблензгера |||.***
14. ***[м2] А *австрийский *МИД || *сослался на то, что **теперь || *решение должно принимать| *юридическое *ведомство |||.***
15. ***[м2] *Позиция *нашего| дипломатического ведомства *известна:| оно, с *самого *начала ||, назвало *задержание || сотрудника *РосКосмоса| *«нарушением || международных *прав» |||.***
16. ***[м2] *Они будут *готовы *отпустить || нашего *гражданина |, как *только поступит *официальная || *реакция || российской *стороны |, и | *цитата |, как сказал Герхард *Ярыш |: «мы *готовы отпустить || *завтра ||, *самое *позднее | в *пятницу» |||.***
17. ***[м2] И вот *сейчас ||, уже перед *самым *эфиром |, поступила *информация |, что || российский гражданин *переведен все-таки | с *зальцбуржского следственного *изолятора |, где он был еще || сегодня *утром |||.***
18. ***[м2] Я *говорил с представителями этого *изолятора ||, так *вот, он сегодня *вечером | уже оказался в *Вене ||, где || и | в течение *двух ближайших *дней |, если верить представителю прокуратуры ||, должен быть *освобожден |||.***
19. ***[м2] *Сейчас || российский *посланник |, *представитель российского *посольства ||, *приглашен || в австрийский *МИД для *консультаций ||, по | поводу этой *новой | теперь уже *ситуации вокруг|| российского *гражданина |||.***
20. *[ж2] В *России сегодня *зафиксированы | сразу *два *случая массовых *отравлений |||.*
21. *[ж2] Пятьдесят два *ребенка | попали в больницу | из подмосковного лагеря «*Смена» ||, и шестьдесят *четыре *человека | *госпитализированы в *Биробиджане |||.*
22. *[ж2] В *причинах разбирается *прокуратура |||.*
23. *[ж2] А в *Красноярске |, как раз *завершился суд | по похожему *делу |||.*
24. *[ж2] Оглашен *приговор в отношении *распорядителей и *поваров |, *обслуживавших в *марте губернаторский *бал ||, который больше *двухсот *гостей | покинули в *сплошном *расстройстве |||.*
25. *[ж2] В *кухне этих *происшествий | *разбирался Дмитрий *Кайстра |||.*
26. ***[м3] *Смена еще не *закончилась |, а в лагере с одноименным *названием | в *Рузском районе *Подмосковья | разыгралась настоящая *драма |||.***
27. ***[м3] Почти *одновременно | *пятьдесят *детей | с диагнозом острая кишечная *инфекция | *были госпитализированы в *больницу | *поселка *Тучково |||.***
28. ***[м3] *Решением *суда повар был *оштрафован |, а директор фирмы *поставщика | лишен *права || заниматься организацией общественного *питания | в течение *года и девяти *месяцев |||.***
29. ***[м3] Теперь не *ясно |, *поставщики или *повара | были лучше осведомлены о качестве *сосисок ||, которыми как-то в *мае | *отобедали ученики одной из *школ | в *Нефтеюганске |||.***
30. ***[м3] Следом за *трапезой, целый *класс | в полном *составе отправился на больничные *койки | инфекционного *отделения ||, где, к *слову сказать, и *встретил || *последний школьный *звонок |||.***
31. ***[м3] *Врачи *говорят |, что *случаев отравления становится все *больше ||, и *причины *каждое *лето || — *одни и *те же |||.***
32. ***[ж3] «*Чаще всего это бывает или *вода ||, опять же непригодная для *питья || *зараженная вода ||, *или | *нарушение работы | холодильного *оборудования ||, или нарушения *технологий *приготовления || пищевых *продуктов» |||.***
33. *[м4] *Тысячи *пассажиров | "застряли" этой *ночью | в американском аэропорту *Лос-Анджелеса |||.*

34. *[м4] Из-за компьютерного *сбоя | в системе *идентификации || люди были вынуждены | находиться в *самолетах | более *семи *часов |||.*
35. *[м4] *Неполадки произошли | именно в *тех *компьютерах |, которые *отвечают || за предоставление *информации | о личных *данных ||, например, *сведений о нахождении в *розыске |||.*
36. *[м4] В *результате иностранные *пассажиры | не *могли пройти таможенный *контроль |||.*
37. *[м4] *Пока системные *администраторы | *устраняют *проблемы | в базе *данных ||, *прибывающие международные *рейсы | отправляют на *посадку || в калифорнийский аэропорт *Онтарио | и *Лас-Вегас |||.*

В этих предложениях представлены стенограммы трех женских голосов (ж) и четырех мужских — (м), а также двух стилей речи: спонтанная речь (предложения 10–19 и 26–32, они выделены полужирным курсивом) и стиль чтения (остальные). Концы предложений отмечены знаками |||, концы предпаузальных синтагм — знаками ||, а концы интонационно выделенных беспаузальных синтагм — знаками |. Последние проставлены на основе аудитивного анализа соответствующих звуковых файлов. Просодически выделенные слова в синтагмах помечены звездочкой (*). В каждом предложении вручную проставлены знаки препинания.

3. Синтаксический эксперимент

Для целей настоящего исследования парсер ЭТАП-3 обрабатывал все предложения экспериментального корпуса в дежурном режиме: для каждого предложения строилась единственная синтаксическая структура (СинтС).

Все предложения подавались на вход парсера в практически непрепарированном письменном виде[3].

Почти для всех предложений парсер ЭТАП-3 построил адекватную СинтС. В одном случае (предложение 37) парсер неверно интерпретировал синтаксически неоднозначное высказывание и перепутал подлежащее глагола с прямым дополнением; кроме того, в ряде ситуаций были неточно установлены синтаксические хозяева предложно-именных групп. Эти погрешности, как впоследствии выяснилось, не оказали влияния на результаты эксперимента.

Ниже даются некоторые примеры полученных парсером СинтС и комментарии к ним, представляющие собой синтаксическую характеристику просодически выделенных слов в предложениях. На рис. 1 и 2 представлены СинтС предложений 1 и 5 из корпуса РТР, соответствующих записям дикторов М1 и Ж1 в стиле чтения текста, а на рис. 3 — СинтС предложения 32, соответствующего записи диктора Ж3 в стиле спонтанной речи. На рис.4 дается СинтС предложения 35 из корпуса НТВ, записанного диктором М4 в стиле чтения текста.

Просодически выделенные слова предложения на рисунке 1 синтаксически интерпретируются следующим образом: *Вены* — самый правый элемент группы первого дополнения; *назад* — самый правый элемент группы обстоятельства; *известие* — правый элемент группы подлежащего, от которой «отрезано» придаточное; *скандалом* — самый правый элемент группы первого дополнения; *Австрии* — самый правый элемент группы причастного оборота; *Роскосмоса* — самый правый элемент группы первого дополнения. Просодические выделения слов *сенсационный, крупный* и *повлиять* с точки зрения СинтС представляются случайными.

В предложении на рисунке 2 просодически выделенные слова получают следующую синтаксическую интерпретацию: *заявлении* — самый правый элемент группы обстоятельства; *Австрии* — самый правый элемент группы дополнения; *недружественный* — правый элемент группы дополнения, от которого отрезана сочинительная группа; *ущерб* — правый элемент группы дополнения, от которой отрезана группа дополнения; *отношениям* — самый правый элемент группы дополнения; *укрепляет* — вершина второго однородного предложения; *Австрии* — правый элемент группы квазиагентивного дополнения, от которой отрезан сравнительный оборот; *штаб-квартир* — правый элемент группы квазиагентивного дополнения, от которой отрезана группа дополнения; *организаций* — самый правый элемент группы дополнения. Два акцентуированных слова и здесь представляются случайными с точки зрения СинтС: *шаг* и *расположения*.

В предложении на рисунке 3 просодически акцентуированные слова имеют такую интерпретацию: *вода* — правый элемент группы присвязочного дополнения, от которой отрезана вся сочинительная цепочка; *питья* — самый правый элемент группы дополнения; *оборудования* — правый элемент группы дополнения, от которой «отрезана» сочинительная цепочка; *технологий* — правый элемент группы дополнения, от которой «отрезана» группа внутреннего дополнения; *приготовления* — правый элемент группы дополнения, от которой «отрезана» группа внутреннего дополнения; *продуктов* — самый правый элемент группы дополнения. Акцентное выделение четырех слов — *чаще, зараженная, или, нарушение* и здесь представляется случайным. Характерно, что одно слово — *всего*, с точки зрения авторов, должно быть выделено, но фонетическая запись этого не подтверждает.

[3] Исключение составляло предложение (16) *Они будут готовы отпустить э… нашего гражданина…*, где заполняющий паузу неструктурный элемент *э…* был опущен, поскольку мешал бы построению приемлемой синтаксической структуры.

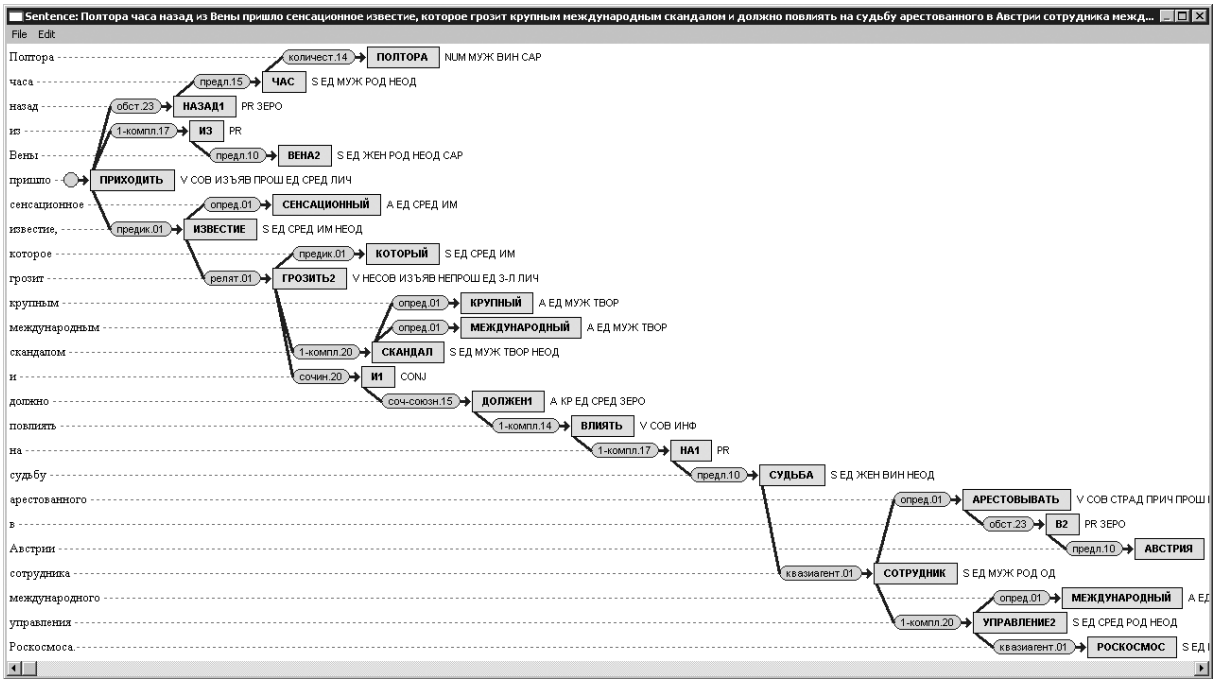


Рис. 1. СинтС предложения 1: *Полтора часа *назад из *Вены пришло *сенсационное *известие, которое грозит *крупным международным *скандалом и должно *повлиять на судьбу арестованного в *Австрии сотрудника международного управления *РосКосмоса.*

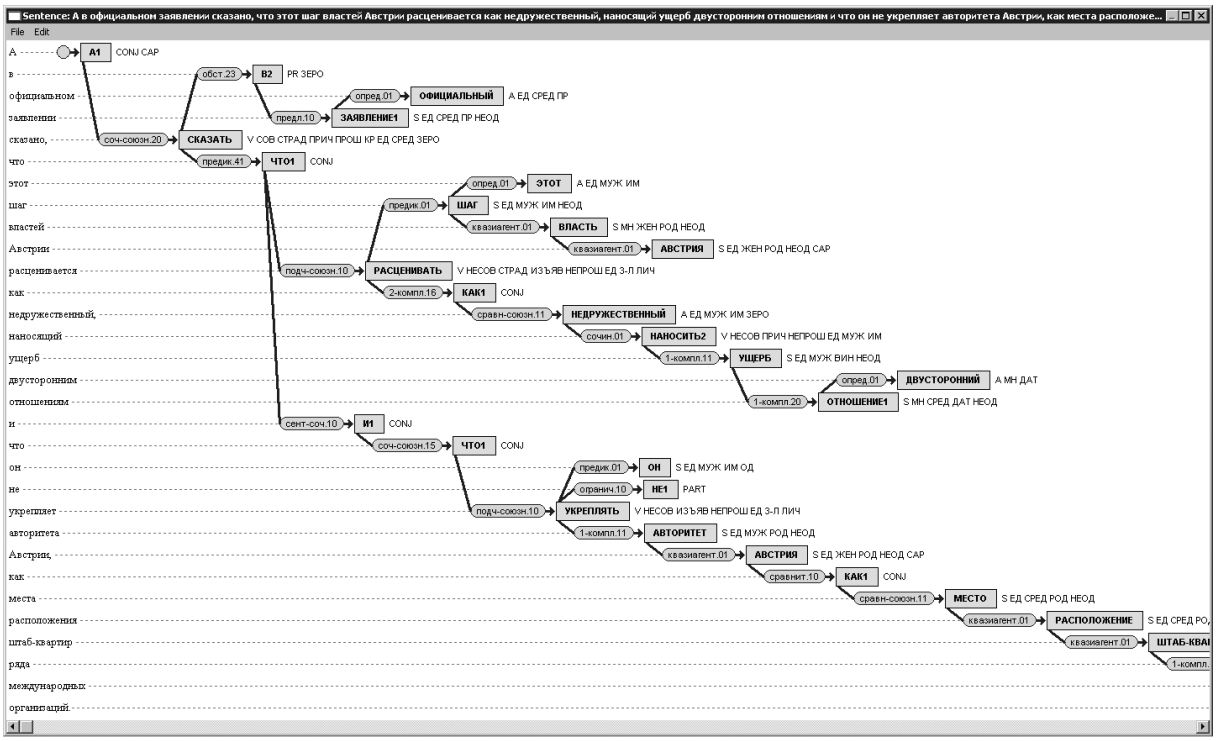


Рис. 2. СинтС предложения 5: *А в официальном *заявлении сказано, что этот *шаг властей *Австрии расценивается как *недружественный, наносящий *ущерб двусторонним *отношениям, и что он не *укрепляет авторитета *Австрии, как места *расположения *штаб-квартир ряда международных *организаций.*

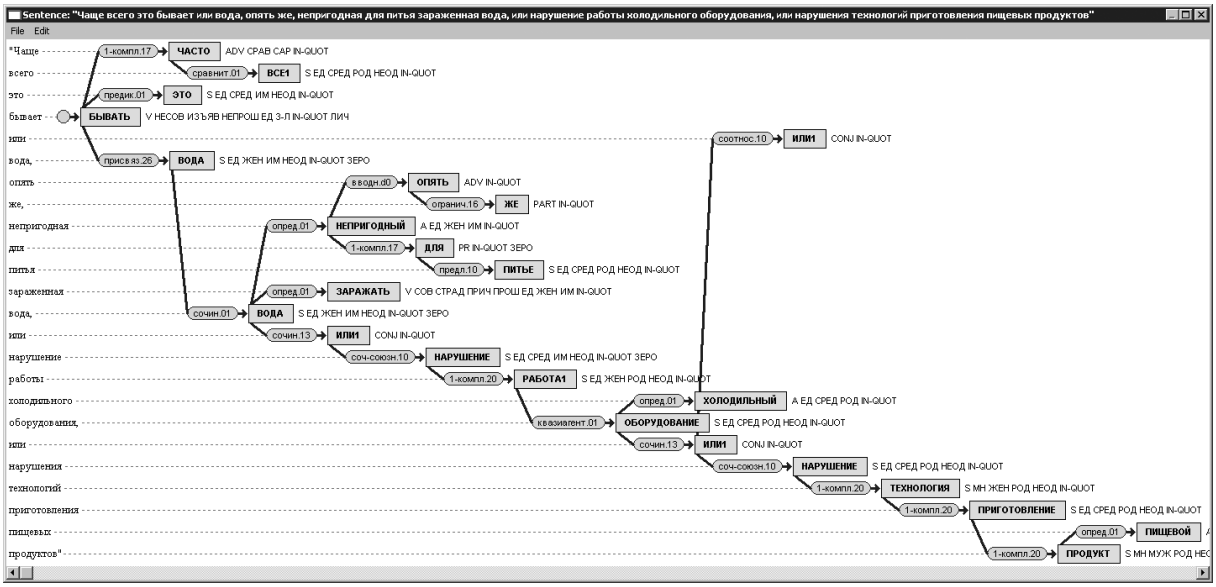


Рис. 3. Предложение 32: **Чаще всего это бывает или *вода, опять же непригодная для *питья, *зараженная вода, *или *нарушение работы холодильного *оборудования, или нарушения *технологий *приготовления пищевых *продуктов.*

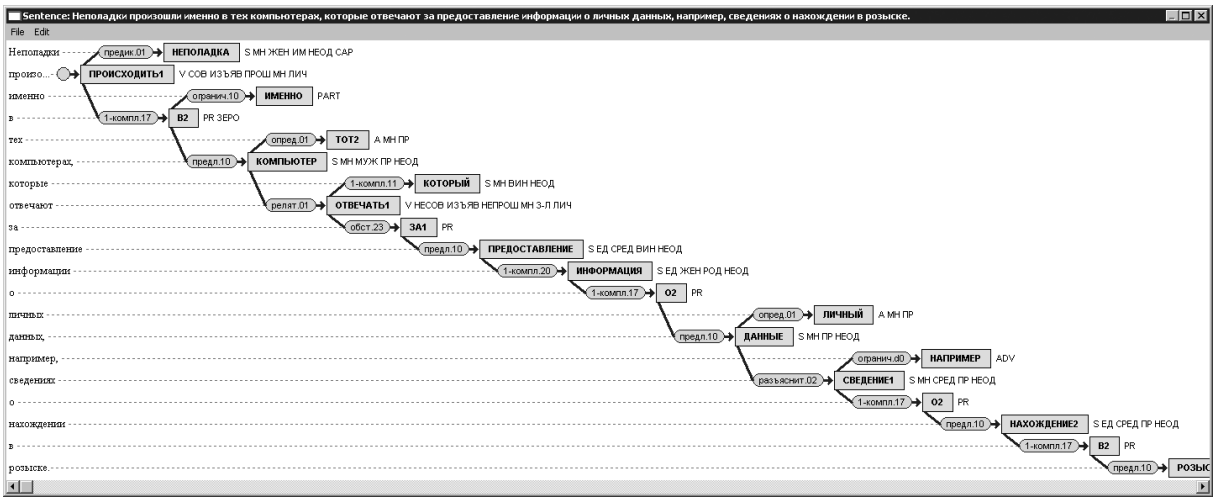


Рис. 4. Предложение 35. **Неполадки произошли именно в *тех *компьютерах, которые *отвечают за предоставление *информации о личных *данных, например, *сведений о нахождении в *розыске.*

Таблица 1. Статистические характеристики акцентуации и членения предложения на синтагмы

Стиль речи	Кол-во слов (всего)	Кол-во выделенных слов	Кол-во синтагм (всего)	Кол-во синтагм перед паузой	Кол-во знаков препинания	Среднее кол-во слов в синтагме
Стиль речи «чтение»	294	111 (38%)	70	31	76	4,2
«Спонтанная речь»	287	134 (47%)	105	60	55	2,7
Весь корпус	581	245 (42%)	175	91	131	3,3

В предложении на рисунке 4 просодически выделенные элементы интерпретируются следующим образом: *неполадки* — самый правый элемент группы подлежащего *компьютерах* — правый элемент группы дополнения, от которой отрезано длинное придаточное; *отвечают* — вершина придаточного; *информации* — правый элемент группы дополнения, от которой отрезана внутренняя группа дополнения; *данных* — правый элемент группы дополнения, от которой отрезана разъяснительная группа;

розыске — самый правый элемент группы дополнения; Два слова — *тех* и *сведений* — акцентуированы с точки зрения структуры случайно.

4. Обсуждение результатов

Внимательное исследование итогов эксперимента позволило авторам сформулировать, в первом приближении, несколько простых правил идентификации просодически маркированных элементов предложения[4].

А. Правила просодического выделения.

Просодически выделенными словами являются:

1) абсолютная вершина предложения;
2) вершины всех частей сложносочиненного предложения;
3) вершины всех придаточных предложений;
4) самые правые субстантивные элементы группы подлежащего, дополнения или обстоятельства при вершинах, перечисленных в пп. 1–3;
5) самый правый субстантивный элемент первой именной подгруппы в группах, перечисленных в п. 4;
6) отдельные классы лексических единиц и конкретные лексические единицы, стоящие в определенной позиции (наречия-детерминанты в начале слов, числительные и количественные существительные).

Б. Правила членения предложения на просодические синтагмы

1) Как следует из анализа изученных текстов, примерно в 90% случаев граница синтагмы выставляется непосредственно после конца просодически акцентированного слова. Остальные 10% приходятся на индивидуальную, синтаксически немотивированную установку границы синтагмы после неакцентированного слова.
2) Хорошим признаком конца синтагмы является наличие знака препинания (в более чем 90 % случаев появление границы синтагмы коррелирует с присутствием такого знака).
3) Появление границы синтагмы подчиняется некоторым статистическим закономерностям и в основном зависит от стиля речи (см. табл. 1).

5. Заключительные замечания

Из полученных результатов, по нашему мнению, логически вытекает следующий план дальнейшей работы: пополнение синтаксического анализатора правилами маркировки просодически выделенных элементов предложения, а также правилами его синтагматического членения. Тем самым будет сделан важный шаг в сторону совершенствования системы синтеза речи за счет блока высокоуровневого синтаксического анализа.

[4] Класс просодического выделения, скажем, противопоставление тематического и рематического интонационных контуров, на данном этапе эксперимента не учитывался.

Литература

1. *Лобанов Б. М.* Алгоритм сегментации текста на синтаксические синтагмы для синтеза речи // Труды Международной конференции «Компьютерная лингвистика и интеллектуальные технологии» (Диалог'2008). — М.: Наука, 2008. С. 323–329.
2. *Лобанов Б. М.* Компьютерный синтез и клонирование речи / Б. М. Лобанов, Л. И. Цирульник // Минск: Белорусская Наука, 2008. 342 с.
3. *Мельчук И. А.* Опыт теории лингвистических моделей «Смысл ⇔ Текст». Семантика, синтаксис. Отв. ред. А. А. Холодович. М. Наука. ГРВЛ. 1974 г. 314 с.
4. *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л.* и др. Лингвистический процессор для сложных информационных систем. М.: Наука, 1992. 256 с.
5. *Богуславский И. М., Иомдин Л. Л., Валеев Д. Р., Сизов В. Г.* Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов // Труды Международной конференции <Корпусная лингвистика — 2008>. СПб.: Санкт-Петербургский государственный университет, 2008. С. 56–74.
6. *Кодзасов С. В., Архипов А. В., Захаров Д. М., Кривнова О. Ф.* База данных «интонация русских информационных текстов // Труды Международной конференции «Компьютерная лингвистика и интеллектуальные технологии» (Диалог'2008). — М.: Наука, 2008. С. 206–209.
7. *Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л., Санников А. В., Санников В. З., Сизов В. Г., Цинман Л. Л.* Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка 2003–2005 г. (результаты и перспективы). М: «Индрик», 2005. С. 193–214.

Просодическая транскрипция: уровни детализации[1]

Prosodic transcription: levels of detail

Кибрик А. А. (kibrik@comtv.ru)
Институт языкознания РАН

Кодзасов С. В. (sankod@yandex.ru), **Худякова М. В.** (mariya.kh@gmail.com)
Московский государственный университет им. М. В. Ломоносова

В книге Кибрик и Подлесская (ред.) 2009 была предложена система просодически ориентированной дискурсивной транскрипции для устной русской речи. В докладе предлагается ряд расширений к транскрипции, в том числе различие между выделительными и тональными акцентами, более подробная разметка тональных акцентов, интервал движения тона в акценте, динамическое раздвоение гласных и др.

1. Вводные замечания

В настоящее время очень актуальна проблема создания корпусов устной речи. Все больше распространяется точка зрения о том, что ограничение языкового материала лишь письменным текстом не дает возможности сформировать адекватные представления о естественном языке. При этом устная речь не может исследоваться «как таковая», лишь в виде звукового сигнала. Для ее объективного и подробного анализа необходима репрезентация в графическом виде. Такую репрезентацию обычно называют *дискурсивной транскрипцией* (Du Bois et al. 1992; Edwards 2001).

В книге Кибрик и Подлесская (ред.) 2009 (ниже — К&П2009) была предложена система транскрибирования устного дискурса для русского языка. Эта система ориентирована на то, чтобы специально подготовленные транскрайберы могли последовательно и надежно транскрибировать исходный дискурс в звуковой форме. Более краткое и предварительное описание этой системы содержится также в статье Кибрик и Подлесская 2003.

2. Транскрипционная система К&П2009

Система транскрипции К&П2009 не является чисто просодической или тем более фонетической. Ее целью была репрезентация основных аспектов устного дискурса, в том числе дискурсивной семантики (в частности, иллокутивные и фазовые значения). Однако просодический компонент занимает в этой системе большое место. Основные просодические компоненты, отражаемые в данной системе, следующие:

- паузы, в том числе абсолютные и заполненные, а также их длительность
- акценты
- тоны в акцентах
- некоторые дискурсивно значимые движения тона за пределами акцентов
- сегментные нелексические явления (смычки, придыхание)
- удлиненная реализация фонем
- тональный регистр
- темп произнесения
- фонетическая редукция
- эмфатическая просодия

Необходимо отметить очень важный принцип дискурсивной транскрипции К&П2009. Эта транскрипция отражает локальную дискурсивную структуру, которая носит квантованный характер. Дискурс порождается в виде последовательность квантов, или шагов — так называемых *элементарных дискурсивных единиц* (ЭДЕ). ЭДЕ идентифицируются транскрайбером главным образом на основе следующих просодических признаков:

- единый тональный контур
- наличие, как правило, одного главного акцента

[1] Данное исследование поддержано грантом РГНФ №08-04-00165а. Работа А. А. Кибрика также поддерживалась грантом Фонда поддержки отечественной науки.

- темповый паттерн: ускорение в начале, замедление к концу
- громкостный паттерн: громче в начале, тише к концу
- пограничная паузация

Эти признаки не являются обязательными и далеко не всегда присутствуют одновременно, но кумулятивно они определяют просодический прототип ЭДЕ.

В транскрипте ЭДЕ отображаются как отдельные строки.

3. Примеры транскриптов, созданных в соответствии с системой К&П2009

В книге К&П2009, помимо описания системы транскрибирования устного дискурса, содержится также корпус устной речи «Рассказы о сновидениях», полностью затранскрибированный на основе этой системы. Корпус состоит из 129 рассказов, общая длительность звучания — около 2 часов.

Приведем два рассказа из книги К&П2009, записанных при помощи этой транскрипционной системы. Отметим, что каждый рассказ в корпусе отражается в виде следующих компонентов:

- идентификационного кода, включающего инициалы рассказчика, указание на пол и возраст
- общего описания, указывающего длительность рассказа, просодическую характеристику и качество записи
- колонки с временно́й разметкой — время от начала рассказа до начала данной ЭДЕ
- колонки с номерами ЭДЕ в рамках рассказа
- колонки, содержащей транскрипционную запись ЭДЕ
- колонки комментариев.

034z АМ м 11 лет

Длительность рассказа 40.8 с.

Склонность к интонации многоточия. В первой половине рассказа фрикативное г.

Качество записи высокое. На заднем плане слышны голоса, визг.

0.0	1.	>>/сон вот,	Начало произнесения первой фонемы не попало в запись.
0.3	2.	когда я \потеря-ался.	
1.5		···(0.9) *Угу*.	Тихо.
3.2	3.	···(0.5) Вотh-h я-а вышел из /дома,	
5.2	4.	были-иh ==	
5.5	5.	··(0.4) ээ(0.4) *никого \не было –дома*,	
7.3	6.	*мне оставили \ключи*.	
8.4	7.	····(1.0) Я-а ··(0.1) /погулял чуть-чуть,	Внутри начальной паузы посторонний скрип.
10.9	8.	··(0.4) /потом-м вот захотел /–е-есть...	
12.7	9.	··(0.4) пошёл /домой,	
13.7	10.	там никого \не-е было...	
14.8	11.	··(0.4) /потом начал искать своих \роди-ителей...	
17.1	12.	···(0.5) потом ··(0.2) /в-вышел ··(0.1) ээ(0.2) ’из \подъезда,	
19.7	13.	никого н-нигде \не-ет...	
21.0	14.	···(0.9) /о-обошёл вотh-h ··(0.2) где-то пол-/–у-улицы...	
24.2	15.	··(0.2) никого \не было.	
25.2	16.	···(0.5) Потом я-а /испугался,	
27.1	17.	···(0.7) /побежал вот куда гла-а= \глаза глядят,	В оборванном фрагменте гла-а= фрикативное г, далее во фрагменте глаза глядят оба раза смычное.
29.9	18.	··(0.4) /вдруг увидел какого-то /мальчика,	На увидел накладывается техническая помеха. В слове вдруг последнюю фонему произносит смычно.
31.9	19.	··(0.1) я у него \спросил:	
33.1	20.	··(0.4) «/Где мои \родители?»	
34.4	21.	···(0.5) Ну-у’,	
35.1	22.	··(0.2) он \сказал:	Начало строки на фоне постороннего стука.
35.9	23.	···(0.5) «\Возвращайся /домой,	
37.5	24.	ты \увидишь их \–там.»,	
38.5	25.	··(0.3) я вернулся /домой,	
39.7	26.	и их-х \увидел.	Тихо. Последний слог — шепотом.

084n ДК м 11 лет

Длительность рассказа 58.2 с.

В паузах часто шумные вдохи. В конце рассказа слышно, что у рассказчика насморк. Часто встречается скрипучий голос. Отчетливая артикуляция и просодия. Не выговаривает твердое р.

Качество записи высокое.

0.0	1.	Я-а ···(0.7) ''(0.1) у своего /\дру-уга,	На первые три слова накладывается техническая помеха. Слово друга произносит как дыруга.
2.7	2.	···(0.5) т-там ··(0.1) ээ(0.2) наш /дом ··(0.1) отрем-м= || ре-е= || ··(0.2) на \ремонт \поста-авили.	Заполненная пауза и фрагмент ре-е — скрипучим голосом.
7.9	3.	···(0.8) И-и /–всех ··(0.2) /\переслали ээ(0.1) в другой \дом,	Несколько слогов на протяжении этой строки произносит скрипучим голосом.
12.0	4.	··(0.4) какой-то /**гря-азный**...	Удлиненный слог — со скрипом. В этой и двух следующих строках говорит с отвращением.
13.6	5.	/**се-ерый**...	
14.4	6.	···(0.7) там /–**ле-естница** побитая...	
16.8	7.	····(1.1) И вот там ещё моя \мама ~	
19.7	8.	···(0.5) {ЧМОКАНЬЕ 0.2} эээ(0.8) ····(1.2) И-и /\мы-ы' ···(0.5) /гуля́ть собра́лись,	
24.9	9.	··(0.4) а я /\куртку /забыл одеть,	На слове одеть слышен тихий стук.
27.1	10.	а мой друг /\вы́бежал,	
28.2	11.	<ему показали>	
28.6	12.	··(0.4) /\музыка неприятная,	
30.1	13.	на площадке какие-то /\пья-аницы были...	
32.4	14.	····(1.4) И я /\испуга-ался,	Фрагмент и я — скрипучим голосом.
35.3	15.	···(0.9) и-и там-м ··(0.1) так и-и ос= ···(0.6) =–тался там-м,	Слог тал — скрипучим голосом.
39.6	16.	···(0.6) {ШМЫГАНЬЕ 0.3} ''(0.2) пытался /–две-ерь запереть к парадному там...	Слово дверь произносит со смехом.
43.4	17.	····(1.0) (*Не \запиралась.*)	Тихо.
45.0	18.	*потом* ··(0.4) в зв= в /–звоно-ок звонил...	Первое слово — тихо.
47.5	19.	(Никто не /–открывал...)	
48.9	20.	{ШМЫГАНЬЕ 0.5} ···(0.9) {ЧМОКАНЬЕ 0.4} *потом там* ···(0.6) ээ(0.3) \ещё громче музыку /включили,	
54.2	21.	··(0.3) и я \просну-улся.	В этой и следующей строке есть акценты на первом безударном слоге акцентированных слов.
55.8	22.	··(0.4) И /больше вот **ничего** не помню.	Слово ничего произносит по слогам.

4. Возможные расширения просодического компонента транскрипции

Не существует никакого объективно заданного набора признаков, которые должны отражаться в дискурсивной транскрипции. Каждая система транскрипции имеет тот уровень детализации, который соответствует исследовательским целям ее авторов. Кроме того, каждая такая система всегда является прагматичным компромиссом. С одной стороны, полнота и подробность — это плюс. С другой стороны, чем более подробна транскрипция, тем больше она содержит различных конвенций и обозначений и тем труднее читается. Кроме того, увеличение степени подробности всегда означает увеличение трудо- и времязатрат. В транскрипционной системе К&П2009 имеется определенный баланс между этими противонаправленными критериями. В целом, эта система была ориентирована на описание локальной дискурсивной структуры, а не собственно фонетики устной речи.

Тем не менее, существует целый ряд направлений, в которых эта транскрипционная система может быть расширена и углублена. Опираясь на работы Кодзасов1996, Кодзасов 2009, Кодзасов и др. 2008, в данном докладе мы предлагаем несколько *расширений* системы К&П2009. Большинство из них касаются акцентных признаков; возможность некоторых из этих расширений была указана уже в Кибрик и Подлесская (ред.) 2009, раздел 4.5.4. В первую очередь, в отличие от К&П2009, мы различаем два типа акцентов — выделительные и тональные.

4.1. Выделительные акценты

Артикуляционно выделительные акценты представляют собой дыхательные толчки, сопровождающие ударные слоги некоторых слов, входящих

в ЭДЕ. Просодическая роль выделительных акцентов — создание ритмической схемы высказывания. В расширенном варианте транскрипции выделительный акцент обозначается символом * перед словом, на которое он падает. Например:

(1) *034z*

5.5 5. ‥(0.4) ээ(0.4) **никого/ не\ было *до*(2)*ма*,

Выделительный акцент не обязательно предполагает изменение тона — ср., например, акцент на ударном слоге слова дома в (1). Чаще всего, однако, выделительные акценты одновременно являются тональными (о них см. ниже); такое сочетание имеет место в слове никого в примере (1).

4.2. Тональные акценты

Тональные акценты представляют собой отклонения от некоторого базового уровня тона. Тональные акценты размечены в К&П2009, однако лишь для ударных слогов слов. Между тем, тональные акценты не всегда привязаны к ударному слогу; кроме того, тональных акцентов в слове может быть более одного. Поэтому в расширенном варианте транскрипции знаки тональных акцентов (/, \ и др.) пишутся не перед словом, а после той фонемы, на которой размещено значимое движение тона.

Чаще всего, конечно, падение или повышение тона происходит в пределах ударного слога, см. пример (2):

(2) *084n*

16.8 7. ‥‥(1.1) И вот\ та/м ещё/ моя *ма\ма ~

Однако достаточно часто оно бывает не локализовано или происходит на целом слове или группе слов, см. комментарий о падении тона в слове ремонт в строке 2 рассказа 084n в разделе 5 ниже. Особый интерес представляют случаи, когда тональный акцент размещен на предударном слоге, как, например, на первом слоге слова собра́лись в (3):

(3) *084n*

19.7 8. ‥‥(0.5) {ЧМОКАНЬЕ 0.2} эээ(0.8)
‥‥(1.2) И-и/– (2) мы-ы/\(2)’
‥‥(0.5) *гуля//(2)ть со\бра́(2)лись,

4.3. Интервалы изменения тона

В расширенном варианте транскрипции была введена градация интервалов движения тона в тональных акцентах. Изменения могут быть средние (примерно от 1/2 до 2/3 октавы), большие (порядка 1 октавы) и малые (порядка 1/3 октавы). Соответствующие обозначения для изменений тона (восходящий и нисходящий соответственно):

- средние /, \
- большие //, \\
- малые ⸗, ⸗.

Ровный тон (–) обозначается, если он входит в состав тонального комплекса с раздвоением гласной, см. ниже.

4.4. Динамическое раздвоение гласных

Динамическое раздвоение гласных — это распадение звуковой волны гласного на два компонента по амплитуде. Раздвоение обозначается символом (2) после соответствующей гласной (и после знака тонального акцента, если он имеется). Это явление широко представлено в спонтанной устной речи. Так, в следующем примере представлено три таких случая в одной ЭДЕ:

(4) *084n*

7.9 3. ‥‥(0.8) И-и/–(2) *все/–(2)х ‥(0.2)
пересла/ли\ ээ(0.1) в друго/й *до\(2)м,

5. Транскрипты, оформленные на основе расширенной просодической транскрипции

Ниже представлены транскрипты тех же двух рассказов из корпуса «Рассказы о сновидениях», которые были приведены выше в соответствии с системой К&П2009. Здесь транскрипты реализованы согласно расширениям, описанным в разделе 4. Несколько дополнительных признаков, таких как степень раствора рта, размечены в колонке комментариев. Попутно с внесением расширений в транскрипте были исправлены несколько неточностей, замеченных в предыдущем варианте. Читатель может самостоятельно сравнить два варианта транскрипта.

034z АМ м 11 лет

Длительность рассказа 40.8 с.

Склонность к интонации многоточия. В первой половине рассказа фрикативное г.

Качество записи высокое. На заднем плане слышны голоса, визг.

0.0	1.	>>*сон во\(2)т,	Начало произнесения первой фонемы не попало в запись.
0.3	2.	*когда я *потеря-а\(2)лся.	
1.5		•••(0.9) *Угу*.	Тихо.
3.2	3.	•••(0.5) Во/тh-h я-а вы/\шел из до/ма,	
5.2	4.	бы/ли-и\(2)h ==	
5.5	5.	••(0.4) ээ(0.4) **никого/ не\ было *до*(2)*ма*,	
7.3	6.	*мне *оста*\(2)*вили *ключи*\.	
8.4	7.	••••(1.0) *Я-а/\(2) ••(0.1) погуля/л чуть-чу\ть,	Внутри начальной паузы посторонний скрип.
10.9	8.	••(0.4) *пот/ом-м вот захоте\л *е-есть…	
12.7	9.	••(0.4) пош/ё\л *домо/й,	
13.7	10.	та\м никого *не-е\ было\…	Слово там произносится с узким раствором рта.
14.8	11.	••(0.4) по–т/о–м начал искать своих *роди-ителей…	Деклинация тона на словах начал искать своих родителей.
17.1	12.	•••(0.5) потом ••(0.2) *в-вы/шел ••(0.1) ээ(0.2) ’и\з по\дъе\зда,	
19.7	13.	никого\ н-нигде\ не-е\т…	
21.0	14.	•••(0.9) *о-обо/шё/л во\тh-h ••(0.2) где\-то *пол-у-улицы\…	
24.2	15.	••(0.2) ни\кого *не\ было.	Деклинация тона.
25.2	16.	•••(0.5) Потом я-а *испуга/лся,	
27.1	17.	•••(0.7) *побежа\л вот куда гла-а= *глаза\ глядя\т,	В оборванном фрагменте гла-а= фрикативное г, далее во фрагменте глаза глядят оба раза смычное.
29.9	18.	••(0.4) *вдру/г увидел како\го-то *ма/льчика,	На увидел накладывается техническая помеха. В слове вдруг последнюю фонему произносит смычно.
31.9	19.	••(0.1) я у него *спроси\л:	
33.1	20.	••(0.4) «*Где\ мои *роди\тели?»	
34.4	21.	•••(0.5) Ну-у’\(2),	
35.1	22.	••(0.2) он *ска\зал:	Начало строки на фоне постороннего стука.
35.9	23.	•••(0.5) «*Возвраща\йся *домо/й¡,	
37.5	24.	ты *уви\дишь их та\м.»,	
38.5	25.	••(0.3) я верну\лся *домо\й,	
39.7	26.	и их-х *уви\дел.	Тихо. Последний слог — шепотом.

084n ДК м 11 лет

Длительность рассказа 58.2 с.

В паузах часто шумные вдохи. В конце рассказа слышно, что у рассказчика насморк. Часто встречается скрипучий голос. Отчетливая артикуляция и просодия. Не выговаривает твердое р.

Качество записи высокое.

0.0	1.	Я-а •••(0.7) ’’(0.1) у своего *дру-у//га\,	На первые три слова накладывается техническая помеха. Слово друга произносит как даруга.
2.7	2.	•••(0.5) т-там ••(0.1) ээ(0.2) наш *до//м ••(0.1) отрем-м= || ре-е= || ••(0.2) на/ *ремо\нт по\ста-авили.	Заполненная пауза и фрагмент ре-е — скрипучим голосом. В слове ремонт падение тона не локализовано. В слове поставили упереднение ударного а.
7.9	3.	•••(0.8) И-и/–(2) *все/–(2)х ••(0.2) пересла/ли\ ээ(0.1) в друго/й *до\(2)м,	Несколько слогов на протяжении этой строки произносит скрипучим голосом. Слово дом — узкий раствор рта.
12.0	4.	••(0.4) какой-то ***гря-а**/(2)**зны\й**…	Удлиненный слог — со скрипом. В этой и двух следующих строках говорит с отвращением.
13.6	5.	***се-е**/(2)**ры\й**…	
14.4	6.	•••(0.7) там ***ле-е**/(2)**стница** поби\тая…	

16.8	7.	····(1.1) И вот та/м ещё/ моя *ма\ма ~	
19.7	8.	···(0.5) {ЧМОКАНЬЕ 0.2} эээ(0.8) ····(1.2) И-и/– (2) мы-ы/\(2)' ···(0.5) *гуля//(2)ть со\бра́(2)лись,	
24.9	9.	··(0.4) а я/– *к/у\ртку *забы//л оде/(2)ть,	На слове одеть слышен тихий стук.
27.1	10.	а мо/й дру–г *вы/б–е\жа/л,	
28.2	11.	<ему показали>	
28.6	12.	··(0.4) *му//зы\ка неприя/\(2)тная,	
30.1	13.	на площадке какие-то *пья-а/\ницы были...	
32.4	14.	····(1.4) И я *испуга-а/лся\,	Фрагмент и я — скрипучим голосом.
35.3	15.	···(0.9) и-и та/\м-м ··(0.1) так и-и ос= ···(0.6) =*тался там-м,	Слог тал — скрипучим голосом.
39.6	16.	···(0.6) {ШМЫГАНЬЕ 0.3} ''(0.2) пытался *две-е/рь запереть к парадному там...	Слово дверь произносит со смехом.
43.4	17.	····(1.0) (*Не запирался.*)	Тихо.
45.0	18.	*потом* ··(0.4) в зв= в *звоно-о/\(2)к звонил...	Первое слово — тихо.
47.5	19.	(Никто\ не открыва–л...)	
48.9	20.	{ШМЫГАНЬЕ 0.5} ···(0.9) {ЧМОКАНЬЕ 0.4} *потом там* ···(0.6) ээ(0.3) ещё\ гро\мче му\зыку *включ\ч/и\ли,	
54.2	21.	··(0.3) и\ я просну-у\(2)лся.	В этой и следующей строке есть акценты на первом безударном слоге акцентированных слов.
55.8	22.	··(0.4) И бо/–(2)льше во\т ***ничего*** не по\мню.	Слово ничего произносит по слогам.

6. Заключительные замечания

В данном докладе предлагается ряд расширений к транскрипционной системе устной русской речи, описанной в книге Кибрик и Подлесская (ред.) 2009. Была добавлена разметка таких просодических параметров, как выделительные акценты, интервал движения тона, раздвоение гласных, была уточнена разметка тональных акцентов.

На дальнейших этапах мы предполагаем добавить в эту разметку такие параметры, как тайминг тональных акцентов, долготные акценты, типы фонаций и некоторые другие.

Литература

1. *Кибрик А. А., Подлесская В. И.* 2003. К созданию корпусов устной русской речи: принципы транскрибирования // Научно-техническая информация. Серия 2-6. С. 5–11.
2. *Кибрик А. А., Подлесская В. И. (ред.)* 2009. Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК.
3. *Кодзасов С. В.* 1996. Комбинаторная модель фразовой просодии // Просодический строй русской речи. М.: Ин-т русского языка РАН. С. 85–123.
4. *Кодзасов С. В.* 2009. Исследования в области русской просодии. М.: ЯСК.
5. *Кодзасов С. В., Архипов А. В., Захаров Л. М., Кривнова О. Ф.* 2008. База данных «Интонация русских информационных текстов» // Компьютерная лингвистика и интеллектуальные технологии. Вып. 7 (14). По материалам международной конференции «Диалог 2008». М.: ИППИ РАН. С. 206–209.
6. *Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., Paolino, D.* 1992. Discourse transcription // Santa Barbara Papers in Linguistics, 4. Santa Barbara: UCSB.
7. *Edwards, J. A.* 2001. The transcription of discourse. // D. Schiffrin, D. Tannen, H. Hamilton (eds.) The handbook of discourse analysis. Oxford: Blackwell. P. 321–348.

К проблеме вариативности языка: метод многофакторного исчисления второго порядка

Towards the problem of linguistic variability: a multifactor second-order calculus method

Кибрик А. Е.

Московский государственный университет им. М. В. Ломоносова

Исследуется клаузальное сочинение в 23-х родственных дагестанских идиомах, отличающееся чрезвычайным разнообразием. В исследуемой выборке не найдено ни одной пары идиомов с тождественными сочинительными конструкциями. Создается впечатление полной произвольности и хаотичности выбора формальных средств оформления этой конструкции в конкретных идиомах. Такая ситуация создает непреодолимые теоретические трудности. Ни традиционный метод классификации, ни структурный метод исчисления не дают удовлетворительного результата.

В статье применяется принципиально другой подход. Его можно назвать методом многофакторного исчисления второго порядка: исчисление сочинительных конструкций обеспечивается не на уровне самих конструкций, а на уровне тех параметризованных принципов и стратегий, которые лежат в основе построения конкретных сочинительных конструкций.

1. Трудности описания вариативности сочинительных конструкций

Около тридцати лет назад я провел эмпирические сопоставительное исследование базового синтаксиса двух десятков дагестанских языков. А именно, на основе единой анкеты был собран языковой материал по структуре элементарного предложения, сентенциальным актантам, рефлексивизации и клаузальному сочинению и опубликован в форме препринтов очень ограниченным тиражом [Кибрик 1979–1981][1]. Позднее я попытался обобщить собранные ранее данные и выявить реальное многообразие конструкций с сентенциальными актантами [Kibrik 1987]. На описательном уровне достичь этого удалось, но имеющихся в то время теоретических инструментов было явно недостаточно для перехода на объяснительный уровень.

Значительно хуже обстояло дело с обобщениями в области *клаузального сочинения*. Под клаузальным сочинением понимается конструкция (состоящая обычно из двух клауз), отражающая находящиеся в поле зрения когнитивно сопряженные[2] события, происходящие одно за другим или одновременно. Следует подчеркнуть, что клаузальное сочинение определяется *не синтаксически, а семантически*. С синтаксической точки зрения семантическое сочинение может выражаться как средствами сочинительной синтаксической связи (например, двумя финитными клаузами, соединенными сочинительным союзом), так и подчинительной[3]. В дагестанских языках клаузальное сочинение

[1] Позднее это исследование было опубликовано в [Кибрик 2003].

[2] Когнитивной сопряженностью называется непосредственная (максимально тесная) связь между когнитивными единицами. В нашем случае между двумя единицами (событиями) имеется отношение непосредственной временной близости, не разрываемой никаким другим событием. Эти события обычно связаны также наличием одних и тех же участников.

[3] М. Хаспельмат основательно обсуждает эту проблему в типологической перспективе [Haspelmath 2004: 33–37] и приходит к аналогичному выводу о возможном несоответствии между семантическими и синаксическими свойствами конструкции, когда предложение семантически является сочинительной структурой, а синтаксически — подчинительной [Haspelmath 2004: 35]. Он также вполне правомерно заключает, что «Имеется много конструкций, представляющих собой их смесь [синтаксическое сочинение и подчинение — АК] и мы находимся лишь самом начале пути к пониманию того, какие ограничения могут быть наложены на такое смешение» [Haspelmath 2004: 37].

регулярно выражается с помощью подчинительной техники. Сочинительная техника является позднейшим заимствованием и не влияет на прочие свойства сочинительной конструкции.

Серьезные трудности анализа сочинительных конструкций возникали уже на уровне описательных обобщений. В материале не встретилось **ни одной пары языков с идентичной сочинительной конструкцией**. Более того, в диалектах одного языка тоже часто отсутствовали идентичные конструкции. Наконец, в немалом числе языков использовалось более одной конструкции. Ситуация выглядит парадоксальной: при наличии некоторых схожих локальных формальных средств в большом количестве родственных языков, наблюдается практически полное отсутствие повторяемости конструкций. Такая безрадостная картина привела к тому, что я не возвращался к ней несколько десятилетий. Настоящая работа является попыткой взглянуть на эту проблему в рамках новой парадигмы, которую можно было бы назвать *когнитивно ориентированной типологией*. Однако с целью облегчения ее понимания я начну с рассмотрения конкретного языкового материала.

2. Примеры клаузального сочинения в пяти языках

В каждом языке материал по клаузальному сочинению собирался методом анкетирования. Анкета представляла собой минимальную группу примеров сочиненных клауз, одна из которых содержала непереходный, а другая — переходный глагол. В непереходной клаузе представлен одноместный глагол движения, а в переходной — агентивный глагол ‘бить’, управляющий агентивной ИГ в эргативе и пациентивной ИГ в номинативе. При этом ИГ в непереходной клаузе кореферентна одной из ИГ в переходной клаузе. Примеры различаются порядком следования непереходной и переходной клаузы (отражающим последовательность событий). В нотации первая клауза выделена квадратными скобками. Кореферентные ИГ коиндексированы. Кроме того, контролер анафорической связи обозначен затемнением.

Эти предложения семантически соответствуют следующим русским эквивалентам (см. (1)–(2)).

Синтаксическое сочинение

(1) а. Мальчик$_i$ девочку ударил и ∅i ушел.
б. Мальчик девочку$_i$ ударил, и она$_i$ ушла.
в. Мальчик$_i$ пришел и ∅$_i$ ударил девочку.
г. Девочка$_i$ пришла, и мальчик ее$_i$ ударил.

Синтаксическое подчинение

(2) а. [∅$_i$ ударив девочку], мальчик$_i$ ушел.
б. *[Мальчик ∅$_i$ ударив], девочка$_i$ ушла.
в. [∅$_i$ придя], мальчик$_i$ ударил девочку.
г. *[∅$_i$ придя], мальчик девочку$_i$ ударил.

В русском языке клаузальное сочинение организовано по аккузативной схеме. А именно:

- при сочинении клауз происходит опущение повторного упоминания референта во второй клаузе, если в обеих клаузах имя этого референта находится в подлежащной позиции;
- если в одной из клауз имя референта находится в любой иной позиции, повторное упоминание в другой клаузе оформляется местоимением.

Теперь обратимся к собственно дагестанскому материалу.

2.1. Бежтинский язык[4]

В бежтинском языке имеются две техники построения сочинительной конструкции. Начнем с рассмотрения 1-й конструкции, см. (3).

Во всех четырех предложениях первая клауза оформлена конвербным показателем *-na / -nä*[5] при глаголе. Из двух ИГ, находящихся в анафорической связи, первая ИГ редуцируется до нуля. Она является мишенью, сигнализатором (анафором) анафорической редукции, а вторая ИГ является контролером (антецедентом) и оформлена как полная ИГ. Иными словами, фактически в этом случае реализована катафорическая схема редукции. Кроме того, в первой, переходной клаузе полная ИГ в номинативе имеет при себе показатель сочинения, совпадающий с показателем конверба, см. (3а).

В примерах (3а-б) наличие нулевой анафоры в первой (переходной) клаузе однозначно восстанавливается, поскольку ее второй актант оформлен полной ИГ.

В примерах (3в-г) во второй клаузе имеется два кандидата на роль контролера, и однозначный выбор может быть осуществлен лишь благодаря согласованию глагола с нулевой ИГ. В (3в) это согласование по 1-му классу, а в (3г) — по 2-му[6].

[4] Бежтинский язык относится к цезской группе. В данной работе используются данные тлядальского диалекта.

[5] Выбор алломорфа зависит от правила гармонии, согласно которому ряд гласного корня контролирует ряд гласной суффиксов.

[6] Подавляющее большинство дагестанских языков имеют согласовательные именные классы. В четырехклассных системах принято их цифровое обозначение. Обычно к 1-му классу относятся имена мужчин, ко 2-му — имена женщин, распределение имен по 3-му и 4-му классам семантически неочевидно.

1. Мальчик девочку побил и ушел (букв. Девочку побив, мальчик ушел).
   а. [∅$_i$ kid-nä j=äɬ'en-nä,] öže$_i$ eɬ'erö.
   ERG 2.девочка.NOM-& 2=бить-CNV 1. мальчик. NOM 1.уйти.AOR

2. Мальчик девочку побил и она ушла (букв. Мальчик побив, девочка ушла).
   б. [öždi ∅$_i$ j=äɬ'en-nä,] kid$_i$ j=eɬ'erö.
   мальчик.ERG 2.NOM 2=бить-CNV 2. девочка. NOM 2=уйти.AOR

(3)

3. Мальчик пришел и побил девочку (букв. Придя, мальчик девочку побил).
   в. [∅$_i$ õq'o-na,] öždi$_i$ kid j=äɬ'ellö.
   1.NOM 1.прийти.CNV мальчик.ERG 2. девочка. NOM 2=бить.AOR

4. Девочка пришла и мальчик ее побил (букв. Придя, мальчик девочку побил).
   г. [∅$_i$ j=õq'o-na,] öždi kid$_i$ j=äɬ'ellö.
   2.NOM 2=прийти-CNV мальчик.ERG 2. девочка. NOM 2=бить.AOR

Патимат пришла и побила Аминат / Аминат пришла и Патимат ее побила. (4)
[∅i/j j=õq'o-na,] pat'imati$_i$ aminat$_j$ j=äɬ'ellö.
2.NOM 2= прийти.CNV Патимат .ERG 2. Аминат .NOM 2=бить.AOR

1. Мальчик девочку побил и ушел.
   а. öže$_i$ [∅$_i$ kid-nä j=äɬ'en-nä,] eɬ'erö.
   1. мальчик .NOM ERG 2.девочка.NOM-& 2=бить-CNV 1.уйти.AOR

2. Мальчик девочку побил и она ушла.
   б. kid$_i$ [öždi ∅$_i$ j=äɬ'en-nä,] j=eɬ'erö.
   2. девочка .NOM мальчик.ERG 2.NOM 2=бить-CNV 2=уйти.AOR

(5)

3. Мальчик пришел и побил девочку.
   в. öždi$_i$ [∅$_i$ õq'o-na,] kid j=äɬ'ellö.
   мальчик .ERG 1.NOM 1.прийти-CNV 2.девочка.NOM 2=бить.AOR

4. Девочка пришла и мальчик ее побил.
   г. kid$_i$ [∅$_i$ j=õq'o-na,] öždi j=äɬ'ellö.
   2. девочка .NOM 2.NOM 2.прийти.CNV мальчик.ERG 2=бить.AOR

А что будет, если обе ИГ относятся одному и тому же классу? См. (4).

Такие предложения не запрещены, но они создают референциальную неоднозначность.

Рассмотрим теперь 2-ю конструкцию[7], см. (5).

Отличие 2-й конструкции от 1-й состоит в том, что контролер анафорической связи продвигается в левую позицию предложения. В итоге во всех предложениях значение референта опущенной ИГ может быть восстановлено.

Итак, между 1-й и 2-й конструкцией, с точки зрения базового принципа различительности, лежит пропасть. Допустимость предложений типа (4) является для дагестанских языков почти уникальной.

С точки зрения базового принципа синтаксической нейтральности данные бежтинского языка доказывают, что ролевые характеристики именных групп не влияют на выбор конструкции.

Отдельного обоснования требует интерпретация предложения (5г). С точки зрения падежного кодирования 'девочка' в главном и зависимом предложении имеет форму номинатива. Поэтому можно было бы считать, что выдвижения влево в этом случае нет, и *kid* находится в зависимой клаузе, а нулевое кодирование представлено во второй, главной клаузе. Однако тогда (5г) строилось бы без учета принципа структурного приоритета, что противоречит как 1-й, так и 2-й конструкции. Принятый анализ (5г) является единственно возможным из системных соображений. Вместе с тем последовательность двух номинативных ИГ, одна из которых не имеет материального выражения, создает определенные трудности восприятия, и в *арчинском* языке реализована другая, уникальная в нашем корпусе языков, техника.

[7] Для облегчения восприятия все последующие примеры используют те же самые предложения в том же порядке. Поэтому при отсылке к конкретному типу предложения ниже используются их порядковые номера, данные курсивом: *(1), (2), (3), (4)*.

1. Брат сестру побил и ушел.

а.		ušdu$_i$,	[∅$_i$	došdur	dāɬerɬ̄i-li,]	uqIa
		брат.NOM	ERG	сестра.NOM	бить-CNV	уйти.AOR

2. Брат побил сестру, и она ушла.

б.		došdur$_i$,	[ušmi	∅$_i$	dāɬerɬ̄ili,]	o=r=qIa.
	2.	сестра .NOM	брат	NOM	бить-CNV	2=уйти.AOR

3. Брат пришел и сестру побил. (6)

в.		ušmi$_i$,	[∅$_i$	qI$_o$a-li,]	došdur	dāɬerɬ̄i.
		брат .ERG	1.NOM	1.прийти-CNV	сестра.NOM	бить.AOR

4. Сестра пришла и брат ее побил.

г.	*	došdur$_i$,	[∅$_i$	da=qIa-li,]	ušmi	dāɬerɬ̄i.
	2.	сестра .NOM	2.NOM	2=прийти-CNV	брат.ERG	бить.AOR
г′.		ušmi,	[došdur$_i$	da=qIa-li,]	∅i	dāɬerɬ̄i.
		брат.ERG	сестра .2.NOM	2=прийти-CNV	NOM	бить.AOR

3. Брат пришел и сестру ударил.

в.	ušmi$_i$,	[∅$_i$	qI$_o$a-li,]	došmis	daxdi.	
	брат .ERG	1.NOM	1.прийти-CNV	сестра.DAT	ударить.AOR	(7)

4. Сестра пришла и брат ее ударил.

г.	došmis$_i$	[∅$_i$	da=qIa-li,]	ušmi	daxdi.
	сестра .DAT	2.NOM	2=прийти-CNV	брат.ERG	ударить.AOR

1. Брат, побив сестру, ушел.

а.		wac$_i$,	[∅$_i$	jac-la	č’īn,]	w=eχa wu=na.	
	1.	брат .NOM	ERG	сестра.NOM-&	бить.CNV	1=уйти.AOR	

2. Сестра, (как) брат (ее) побил, ушла.

б.		jac$_i$,	[wacud-la	∅$_i$	č’īn,]	j=eχa ji=na.	
	2.	сестра .NOM	брат.ERG-&	2.NOM	бить.CNV	1=уйти.AOR.	(8)

3. Брат пришел и побил сестру.

в.	[wac$_i$	wi=ʔa,]	∅$_i$	jac	č’in.	
	брат .NOM	1=прийти.CNV	erg	сестра.NOM	бить.AOR	

4. Сестра пришла и брат (ее) побил.

г.	[jac$_i$	ni=ʔa,]	wacud	∅$_i$	č’in.	
	2. сестра .NOM	1=прийти.CNV	брат.ERG	2.NOM	бить.AOR	

1. Мальчик девочку побил и ушел.

а.		uže$_i$,	[∅$_i$	kad-ɨn	šuk’u-n,]	ẽʟ’i.	
	1.	мальчик .NOM	ERG	2.девочка.NOM-&	2.бить-CNV	1.уйти.AOR	

2. Мальчик девочку побил и она ушла.

б.		kad$_i$,	[uža-n	∅$_i$	šuk’u-n],	j=ẽʟ’i.	
	2.	девочка .NOM	мальчик.ERG-&	NOM	2.бить-CNV	2=уйти.AOR	(9)

3. Мальчик пришел и побил девочку.

в.	uže$_i$	ot’q’ɨ	[∅$_i$	kad-ɨn	šuk’u-n].	
	1. мальчик .NOM	1.прийти.AOR	ERG	2.девочка.NOM-&	2.бить-CNV	

4. Девочка пришла и мальчик ее побил.

г.	kad$_i$	j=ot’q’ɨ	[uža-n	∅$_i$	šuk’u-n].	
	2. девочка .NOM	2=прийти.AOR	мальчик.ERG-&	NOM	2.бить-CNV	

1. Мальчик девочку побил и ушел.

а. [vašas$_i$, jaša ʟ'ap'žigi ʟ'ap'u,] ʁa$_i$ || vaša$_i$ ʟunʟ'a.
мальчик.1.ERG девочка.2.NOM бить.CNV* PRON.1 || мальчик.NOM уйти.AOR

2. Мальчик девочку побил и она ушла.

б. [vašas$_i$, jaša ʟ'ap'žigi ʟ'ap'u,] ʁe$_i$ || jaša$_i$ ʟunʟ'a.
мальчик.1.ERG девочка.2.NOM бить.CNV PRON.2 || девочка.NOM уйти.AOR

(10)

3. Мальчик пришел и побил девочку.

в. [vaša$_i$ şelžigi solu,] ∅$_i$ || ʁas$_i$ jaša ʟ'ap'ri.
мальчик.1.NOM прийти.CNV ERG || PRON.1.ERG девочка.NOM бить.AOR

4. Девочка пришла и мальчик ее побил.

г. [jaša$_i$ şelžigi solu,] vašas ∅$_i$ || ʁe$_i$ ʟ'ap'ri.
девочка.2.NOM прийти.CNV мальчик.ERG NOM || PRON.2.NOM бить.AOR

* Здесь и ниже конверб выражен аналитически редупликацией глагольного корня.

В арчинском языке схема клаузального сочинения в предложениях типа (6а–в) идентична аналогичным бежтинским в (5а–в), а предложение типа (5г) в арчинском языке блокировано, см. (6г). Грамматичным является предложение (6г´).

Утверждение, что выбор техники построения предложения (6г´) мотивирован тождеством падежей контролера и мишени, доказывается примерами типа (7г), где, как и в (7в), тождества падежей контролера и мишени нет.

Дело в том, что, в отличие от глагола «бить», имеющего аргументы в номинативе и эргативе, глагол «ударить» оформляет аргументы дативом и эргативом. Поэтому в (7г) падеж контролера *došmis* не совпадает с падежом мишени. Такая конструкция не вызывает сложностей восприятия и идентична конструкции (5г).

2.2. Чамалинский язык[8]

Сочинительная конструкция в чамалинском языке в ряде отношений также напоминает 2-ю конструкцию *бежтинского* языка. Рассмотрим эмпирический материал (по данным говора сел. Нижнее Гаквари), см. (8).

В (8) мишень во всех случаях оформляется нулем и контролер всегда предшествует мишени. Остаются три вопроса.

Во-первых, чем регулируется позиция контролера анафорической связи? В (8а–б) он находится в главной клаузе, а в (8в–г) — в зависимой.

Во-вторых, почему в (8а–б) имеет место *выдвижение ИГ-контролера влево*?

В-третьих, почему выдвижение влево ИГ-контролера имеет место только в (8а–б)?

Наконец, в (8в) можно видеть, что в зависимой клаузе показатель сочинения *-la* имеется не только на номинативной полной ИГ, но и на эргативной, в отличие от бежтинского языка.

2.3. Хваршинский язык[9]

Рассмотрим примеры (9).

Примеры (9а–б) по структуре идентичны *чамалинским* примерам (8а–б). Однако (9в–г) отличаются от всего, что мы наблюдали ранее. А именно, зависимая клауза следует за главной. Почему?

2.4. Аварский язык[10]

Аварский язык отличается от ранее рассмотренных языков тем, что в нем интенсивно используются разные способы кодирования мишени анафорической редукции. Рассмотрим примеры (10)[11].

Прежде всего бросается в глаза, что мишень всегда находится во второй, главной клаузе и выражается нулем, личным местоимением или полной ИГ. При этом в (10в–г) возможно местоимение или нуль, а в (10а–б) — местоимение или полная ИГ. Поскольку местоимение возможно во всех случаях, различаются нуль и полная ИГ.

Как соотносятся данные аварского языка с данными ранее рассмотренных языков?

---

[8] Чамалинский язык относится к андийской группе дагестанских языков.

[9] Хваршинский язык, как и бежтинский, относится к цезской группе. Рассмотрим примеры (данные говора сел. Инхоквари).

[10] Данные по аварскому языку относятся к одному из идиомов анцухского диалекта (селение Чадаколоб), в ряде отношений далекому от литературного языка.

[11] В аварском языке по классам изменяются только глаголы, начинающиеся с гласного, поэтому в нижеследующих примерах глаголы не имеют классных показателей. Местоимение 3-го лица в номинативе различает классы.

1. Мальчик сестру побил и ушел.

а.	baj$_i$	[∅$_i$	či	d.uvčč-unu,]	guši.
	мальчик.NOM	ERG	сестра.NOM	бить-CNV	уйти.AOR

2. Мальчик сестру побил и она ушла.

б.	[bali	či$_i$	uvčči-gan,]	dumu$_i$	guši.
	мальчик.ERG	сестра.NOM	бить-TEMP	PRON.NOM	уйти.AOR

(11)

3. Мальчик пришел и сестру побил.

в.	bali$_i$,	[∅$_i$	d.uf-nu,]	či	guvčči
	мальчик.ERG	NOM	прийти-CNV	сестра.NOM	бить.AOR

4. Сестра пришла и мальчик ее побил.

г.	[či$_i$	gi-gan,]	bali	dumu$_i$	guvčči.
	сестра.NOM	прийти-TEMP	брат.ERG	PRON.NOM	бить.AOR

2.5. Табасаранский язык[12]

Завершим рассмотрение первичного языкового материала данными дюбекского говора, см. (11).

В дюбекском говоре в сочинительной конструкции используется несколько кодирующих техник.

Во-первых, в (11а, в) мишень выражается нулем, а в (11б, г) — местоимением.

Во-вторых, глагол зависимой клаузы в (11а, в) оформлен показателем конверба *-unu*, а в (11б, г) — показателем темпорального конверба *-gan*.

В-третьих, в (11а, в) имеет место вынос антецедента анафорической связи влево, а в (11б, г) этой техники нет.

Итак, все кодирующие средства распределены в соответствии с принципом *синтаксической аккузативности*. Данные дюбекского говора имеют для наших целей принципиальное значение. Это, пожалуй, единственный случай в нашем материале, когда имеются веские основания считать, что синтаксическая аккузативность в данном языке реализована. Заслуживает, правда, внимания тот факт, что аккузативная схема реализуется не одним кодирующим средством, а системой самостоятельных средств, действующих в одном направлении. Этот факт указывает на то, что такая техника складывалась постепенно, хотя отдельные ее компоненты присутствуют в различных комбинациях в разных дагестанских языках.

Мы рассмотрели данные пяти языков, демонстрирующие вариативность клаузального сочинения. В принципе, каждый конкретный язык может быть формально описан в терминах наблюдаемых конструкций. Однако любое такое описание не дает ключа к пониманию того,

- какой механизм лежит в их основе,
- каким образом он мотивирует их феноменальное разнообразие,
- какие свойства этого механизма естественным образом ограничивают потенциальное множество сочинительных конструкций, как не зарегистрированных, так и таких, которые могут быть впоследствии обнаружены или возникнуть в процессе развития дагестанских языков.

За внешней непредсказуемостью сочинительных конструкций и кажущейся их произвольностью кроется стройный многофакторный механизм, управляющий построением сочинительных конструкций. Идея многофакторного механизма была предложена Джоанной Николс [Nichols 1981] для описания падежного оформления предикатных имен (predicate nominals) в русском языке. Проблема выбора падежа предикатного имени[13] имеет длительную, но безуспешную традицию изучения в русистике. Джоанна предположила, что на этот выбор влияет множество одновременно действующих факторов, и, используя метод минимальных пар и психолингвистического эксперимента, некоторые из этих факторов выделила. Это пионерское исследование создавало предпосылки для совершенно нового взгляда на архитектуру языка. Насколько мне известно, такая идеология не получила до сих пор распространения в западной лингвистике.

[12] Табасаранский язык представлен говором сел. Дюбек, относящегося к северному диалекту табасаранского языка. Данные дюбекского говора взяты из работы [Бергельсон и Кибрик, 1982]. Примечательно, что в отношении сочинительной конструкции дюбекский говор принципиально отличается от говора кондикского языка, относящегося к южному диалекту.

[13] речь идет о согласовании предикатного имени с контролером по падежу или маркированием этого имени творительным падежом, например, *она (NOM) была* ***бестолковая ученица*** *(NOM)* vs. *она (NOM) была* ***бестолковой ученицей*** *(INSTR)*.

3. Стратегии и принципы

Предлагаемое ниже типологическое описание основывается на гипотезе о многофакторном механизме образования сочинительных конструкций в дагестанских языках. В работе проводится различение двух типов факторов: *стратегий* и *принципов*. Как стратегии, так и принципы организованы параметрически и принимают в дагестанских языках некоторое множество значений.

- Стратегии — это локальные факторы, предопределяющие построение конструкций определенного типа.
- Принципы — это такие факторы, которые относятся к базовым синтаксическим выборам языка, во многом предопределяющим самые разные компоненты языковой структуры, в частности, выбор конкретных стратегий построения тех или иных конструкций.

Стратегии более наглядно представлены в синтаксических конструкциях, поэтому удобнее начать с них.

3.1. Стратегии построения сочинительной конструкции

Рассмотрим основные стратегии оформления конструкций с семантическим клаузальным сочинением. Напомню, что в дагестанских языках синтаксически сочинение оформляется средствами синтаксического подчинения, то есть последовательностью зависимой и главной клаузы, поэтому на этой стратегии я специально останавливаться не буду.

3.1.1. Кодирование зависимой клаузы

В дагестанских языках зависимая клауза маркируется при ее вершине (глаголе), иногда дополнительно при полной именной группе (ИГ). Используются следующие стратегии маркирования зависимой клаузы:

- CNV конверб
- TEMP временной конверб
- STEM форма, совпадающая с основой глагола
- VF финитная форма глагола
- & показатель сочинения на NP в переходной зависимой клаузе

Наиболее употребителен показатель конверба (типа русского деепричастия или английского gerund) при глаголе, см. примеры (3–10, 11а,в). Возможен также маркер временного конверба (английский аналог — временные союзы типа *when*), см. пример (11б,г), использование финитной формы глагола или формы, совпадающей с глагольной основой. Наряду с приглагольным маркером подчинения в некоторых языках показатель сочинения прибавляется также к вершине именной группы переходной клаузы, см. (2а, 5а, 8а–б, 9а–г).

3.1.2. Линейная последовательность клауз

Обычно в сочинительной конструкции клаузы следуют одна за другой и иконически отражают временную последовательность событий. В некоторых языках при линейном порядке *dependent clause* + *main clause* в некоторых контекстах наблюдается вынос одной ИГ[14] главной клаузы в абсолютно левую позицию. Главная клауза в таком случае разрывается, обрамляя зависимую клаузу, см. (5а–г, 6а–г, 7в–г, 8а–б, 9а–б, 11а, в).

Итак, возможна такая стратегия:

- вынос влево ИГ-(контролера) второй клаузы

3.1.3. Кодирование мишени анафорической редукции

Ввиду того, что семантическое сочинение мотивировано когнитивной сопряженностью событий, неудивительно, что очень часто эти события имеют общих участников (одного или более). Говорят, что номинации этих участников связаны анафорическим отношением. Одна из номинаций вводит в рассмотрение некоторый референт (антецедент), а другая повторно его называет. Эта вторая номинация обычно выражается редуцированно. Она называется мишенью (анафором), а связанная с ней первая номинация — контролером (антецедентом)[15] анафорической редукции. Контролер выражается полной ИГ и неотличим от прочих полных ИГ. В определенных контекстах это может создавать трудности однозначного установления контролера.

В дагестанских языках используются следующие стратегии редукции мишени:

- ∅ отсутствие фонетического материала
- PRON личное местоимение
- REFL возвратное (эмфатическое) местоимение
- FULL полная ИГ (повтор ИГ[16])

Наиболее частотна максимальная редукция, при которой мишень редукции не имеет материального выражения (так называемая *нулевая анафора*: обычно для наглядности на место мишени в примерах условно вводится нулевой знак). Таких случаев в приведенном материале большинство. Прочие виды выражения мишени: личным местоимением (*прономинальная анафора*), см. (10а–г, 11б, г), возвратным местоимением (*рефлексивная анафора*)

---

[14] Как станет ясно в дальнейшем, эта ИГ обычно, хотя и не всегда, см. *арчинский* язык в (6г′), является контролером (антецедентом) анафорической связи. Процесс выноса ИГ мотивирован теми принципами построения сочинительной конструкции, которые релевантны для данного языка.

[15] Эта терминология основана на предположении, что антецедентом является первое упоминание референта. В действительности возможен и их обратный порядок (катафора). Однако и в этом случае используется тот же термин антецедент.

[16] То есть отсутствие редукции.

или без использования редукции, см. (10а-б), — обычно используются как дополнительные стратегии в специальных контекстах, часто наряду с нулевой анафорой.

Таково пространство варьирования стратегий оформления сочинительной конструкции. Для нас особый интерес представляет вопрос о выборе способа выражения кореферентности (=анафорической связи) между именными группами в сочиненных клаузах:

- имеет ли место редукция именной группы в одной из клауз,
- какая из кореферентных именных групп является контролером (антецедентом), а какая — мишенью редукции,
- имеются ли ограничения на ролевые (падежные) характеристики мишени и контролера при анафорической редукции.

В этой сфере наблюдается наибольшее разнообразие, даже между генетически наиболее близкими языками. Чтобы ответить на эти вопросы, следует перейти к рассмотрению пространства варьирования принципов, в определенной степени управляющих выбором стратегий.

3.2. Принципы построения сочинительной конструкции

В данной работе в центре нашего внимания являются те принципы, которые ограничивают набор стратегий при кодировании анафорических отношений между клаузами в сочинительной конструкции.

3.2.1. Идентификация ИГ, связанных анафорическим отношением

Если сочинены клаузы с одноместными глаголами, то при кореферентности ИГ в этих клаузах анафорическое отношение между ИГ-ми устанавливается однозначно. Если же в одной из клауз имеется двухместный глагол с двумя ИГ-ми — агентивом и пациентивом, то идентификация ИГ, вступающей в анафорическое отношение с ИГ другой клаузы, требует специальных кодирующих средств. С этой ситуацией связан базовый принцип различительности.

- Принцип различительности: обеспечивать однозначное нахождение той ИГ в переходной клаузе, которая кореферентна ИГ непереходной клаузы.

Принцип различительности в общей формулировке[17] является одним из общих семиотических принципов организации языка. Он ориентирован на слушающего, декодирующего сообщение, и накладывает на форму сообщения требование восстановимости заданного говорящим смысла. В частности, важным свойством каждой номинации является вычислимость ее референта (возможность приписать ей тот референт, который имел в виду говорящий). Так, в русском языке структура переходной клаузы с субъектно-объектной оппозицией системно допускает такую возможность. Например, в предложении ***Саша*** *любит* ***Машу*** однозначно указывается на референт-экспериенцер (*Саша*) и референт-стимул (*Маша*) благодаря падежному маркированию. Однако в предложении ***Болящий дух*** *врачует* ***песнопенье*** принцип различительности не срабатывает, поскольку обе номинации (*болящий дух* и *песнопенье*) морфологически не различают именительного и винительного падежей. Поэтому принцип различительности не является абсолютным, он указывает на универсальную тенденцию организации языковой структуры. В дагестанских языках представлены обе возможности, ср. выше (3–4) и (5) в *бежтинском* языке, хотя следование принципу различительности существенно доминирует.

3.2.2. Синтаксические свойства базисной конструкции предложения

Широко известно, что дагестанские языки отличаются от европейского стандарта наличием в них так называемой эргативной конструкции. В ряде своих работ я подробно останавливался на теории элементарного предложения и особенностях эргативной конструкции в дагестанских языках (см. обобщение моей концепции в [Kibrik 1997]. Поэтому здесь я укажу очень кратко только самые существенные положения, необходимые для понимания данной работы.

С точки зрения формальной кодировки ИГ в непереходной и переходной клаузах языки мира используют несколько базисных типов конструкций, исходно ориентированных на роли участников ситуации.

Если ограничиться двумя важнейшими элементарными ролями агенса и пациенса, то наиболее иконичной конструкцией, объединяющей с помощью некоторых кодирующих средств непереходную и переходную клаузу, является так называемая активная конструкция. В ней агенс переходной и непереходной клаузы имеет одну кодировку, а пациенс — другую.

Привычная для нас аккузативная конструкция объединяет агенс переходной клаузы с единственной ИГ непереходной клаузы. Это объединение, как я полагаю, мотивировано особой гиперролью принципала, *главного участника ситуации, в первую очередь ответственного за то, что событие имеет место*. В этой конструкции маркированной (например, аккузативом) является гиперроль пациентива (пациенс и другие элементарные роли типа темы, цели и др.), а немаркированной — гиперроль принципала (номинатив).

[17] которую я здесь не привожу, так как это увело бы нас далеко от темы данной работы.

Возможно также объединение ИГ непереходной клаузы с пациентивом переходной клаузы. В этом случае реализуется эргативная конструкция, в которой маркированным является участник с гиперролью агентива (объединяющего агенс с другими элементарными ролями, например, экспериенцером), а немаркированным — участник с гиперролью абсолютива (кодируется в дагестанских языках номинативом[18]). Гиперроль абсолютива объединяет ИГ непереходной клаузы с ролью переходного пациенса. Ее значение: *это наиболее непосредственный, ближайший участник события, наиболее затрагиваемый им или вовлеченный в него*[19]. В дагестанских языках немаркированная гиперроль абсолютива кодируется прямым падежом номинативом, обычно совпадающим с корнем имени, а гиперроль агентива — косвенным эргативным падежом.

Иногда встречается также так называемая трехчленная конструкция, противопоставляющая агентив и пациентив переходной клаузы единственному участнику непереходной.

Отмечу, что все четыре конструкции следуют принципу различительности, о котором говорилось выше. Однако, возможна также и пятая, нейтральная конструкция, не различающая ролей переходной и непереходной клаузы: <S = A = P>. Эта конструкция находится в противоречии с принципом различительности. Возможно поэтому она встречается крайне редко.

Весь этот экскурс в типологию элементарной клаузы опирался на морфологические тесты. Более детальное межъязыковое исследование эргативной конструкции в 70-х годах позволило обнаружить, что языки, имеющие морфологическую эргативную конструкцию, очень по-разному ведут себя синтаксически, особенно при построении поликлаузальных структур. Возникла гипотеза, см. [Anderson 1976], что в языках с эргативной конструкцией их эргативность ограничена морфологическим уровнем, а их синтаксис организован по аккузативной схеме. В этой связи сочинительная конструкция дагестанских языков представляет интерес с точки зрения того, как оформляются в этой конструкции анафорические отношения. Предвосхищая дальнейшее, можно сказать, что гипотеза Андерсона в данном случае не проходит, хотя в некоторых языках она как будто бы реализуется.

Универсальная для дагестанских языков семантически эргативная конструкция[20] на синтаксическом уровне[21] потенциально может соотноситься со следующими альтернативными синтаксическими принципами:

- **Принцип синтаксической нейтральности** (использовать одинаковые схемы анафорического кодирования независимо от ролевых характеристик именных групп — S, A и P).
- **Принцип синтаксической аккузативности** (различное анафорическое кодирование в зависимости от ролевых характеристик именных групп, а именно S и A vs. P).

У читателя может возникнуть вопрос, почему на синтаксическом уровне не действует принцип синтаксической эргативности. Ответ на него простой. Дагестанские языки являются семантически эргативными (ролевыми) языками (в отличие, например, от синтаксической аккузативности европейских, следующих морфологической кодировке аккузативной схемы далеко не всегда[22]). Падежное маркирование в дагестанских языках однозначно связано с гиперролевыми характеристиками ИГ, поэтому, в частности, в них отсутствуют залоги, поскольку они нарушают это однозначное соответствие[23]. Для семантически эргативных языков ожидаемой является синтаксическая нейтральность.

Что касается синтаксической аккузативности, ее существование означало бы сдвиг в сторону языков другого типа, а именно языков морфологически эргативных, а синтаксически аккузативных[24] (что и утверждается, как говорилось выше, некоторыми лингвистами). Именно вопрос о наличии этой альтернативы является для нашей темы важнейшим.

---

[18] Отмечу, что я принципиально не согласен с внедрившейся в последние тридцать лет практикой усматривать особый морфологический падеж абсолютив. Против этого имеется масса возражений, но здесь не место об этом подробно говорить, см. [Kibrik 1997: сноска 19]. Термин абсолютив я использую для семантической гиперроли.

[19] Замечу, что участник непереходной клаузы в перспективе аккузативной конструкции интерпретируется как принципал, а в перспективе эргативной конструкции — как абсолютив, и это никоим образом не является семантическим парадоксом, потому что это — единственный участник события.

[20] кодирующая предикатно-аргументную структуру клаузы в терминах гиперролей агентива и абсолютива.

[21] с точки зрения моделей организации кодирования кореферентных отношений.

[22] Аналогичным образом, язык дирбал, также использующий эргативную конструкцию, является синтаксически эргативным: в нем кореферентные связи ограничены номинативом, поэтому при кореферентности с ИГ в эргативе в зависимой клаузе необходимо антипассивное преобразование, повышающее эту ИГ до статуса номинатива. Эту интерпретацию языка дирбал см. в [Kibrik 1997: 319–320].

[23] Так, в русском пассиве *Чины людьми даются* номинативом маркируется пациентив, а не агентив, а агентив кодируется не номинативом, а творительным падежом.

[24] При этом вставал бы вопрос о сохранении или утрате семантической эргативности, который мы пока затрагивать не будем.

3.2.3. Синтаксические позиции кореферентных ИГ

При выборе контролера (и, соответственно, мишени) анафорического отношения могут действовать следующие принципы:

- **Принцип структурного приоритета** (контролер в главной клаузе (МС)).
- **Принцип линейного приоритета** (контролер линейно предшествует мишени).
- **Принцип приоритета непереходной клаузы** (контролер в непереходной клаузе).

Дагестанские языки различаются тем, какому из принципов они следуют.

Принцип структурного приоритета мотивирован престижной синтаксической позицией в сочинительной конструкции: элемент главной клаузы выше по синтаксическому дереву элемента зависимой клаузы.

Принцип линейного приоритета мотивирован первым упоминанием референта в линейной структуре конструкции, что соответствует общей идее связи между кореферентными ИГ слева направо.

Наконец, последний принцип мотивирован однозначностью нахождения контролера, поскольку в непереходной клаузе такая ИГ единственная.

3.2.4. Линейная последовательность главной и зависимой клауз (при синтаксическом подчинении)

Как указывалось выше, в нарративах линейная последовательность клауз иконически указывает на временную последовательность развертывания событий. При подчинительной технике оформления семантического сочинения необходимо указать, какая из клауз является главной, а какая зависимой. Это предопределяется базисным синтаксическим принципом:

- **Принцип левостороннего ветвления.**

Этот принцип (зависимая составляющая предшествует главной) не является абсолютным. В некоторых случаях он может не действовать или локально нарушаться. При соблюдении этого принципа последовательность главной и зависимой клауз следующая:

Зависимая клауза + Главная клауза, то есть:
$[_{C2}[_{C1} DC] + MC]$

В противном случае последовательность такая:

Главная клауза + Зависимая клауза, то есть: $[_{C1} MC + [_{C2} DC]]$

Итак, мы рассмотрели репертуар стратегий и принципов, которые участвуют в построении сочинительной конструкции в дагестанских языках и приводят к ее разнообразию.

Теперь покажем, как такая система стратегий и параметров описывает рассмотренные выше примеры.

4. Варьирование сочинительной конструкции в терминах стратегий и принципов

Теперь проинтерпретируем тот же языковой материал с точки зрения предлагаемой теории параметризованных стратегий и принципов.

4.1. Бежтинский язык

Принципы и стратегии, предопределяющие 1-ю сочинительную конструкцию, см. (3–4), представлены на схеме 1.

Принцип левостороннего ветвления проявляется в том, что во всех предложениях зависимая клауза предшествует главной. Следует отметить, что этот принцип действует в подавляющем большинстве рассмотренных конструкций (в 21-м из 23-х идиомов). Более того, этот принцип не является специфической характеристикой клаузального сочинения, он в значительной степени предопределяет синтаксис дагестанских языков в целом, в частности, базовый порядок слов предикатно-аргументной структуры: <A + P + VERB>[25], и в именной группе: <Modifier + NP>[26]. Этот принцип не является чисто формальным. Он мотивирован тем, что вершина составляющей дефолтно является ее фокусом, а зависимая составляющая — темой, естественный порядок которых ТЕМА + ФОКУС. Принцип синтаксической нейтральности проявляется в том, что ролевые характеристики ИГ в сочиненных клаузах не влияют на построение сочинительной конструкции. Этот принцип также является статистически доминирующим (действует в 19-ти идиомах).

В этой конструкции маркирование подчинения в зависимой клаузе подчиняется базовой стратегии для так называемых (в соответствии с терминологией [Nichols 1986]) dependent-marking languages. Выбор единой формы конверба, а также нулевой анафоры для всех четырех предложений (3а–г) мотивирован принципом синтаксической нейтральности. Наличие контролера анафорической связи в главной клаузе, а мишени — в зависимой диктуется принципом структурного приоритета.

Система принципов и стратегий для этой конструкции полностью ее описывает. Простота данной системы создает неоднозначность в (3в–г) и (4), по-

[25] Я сознательно не использую традиционную нотацию SOV, поскольку она исходит из ложной презумпции универсальности ядерных синтаксических отношений (субъекта и объекта), см. [Kibrik 1997].

[26] Вершинная ИГ занимает конечную позицию, которой предшествует ИГ в генитиве, относительное предложение и прочие модификаторы.

Схема 1

Бежтинский язык. 1-я конструкция

Принципы:	**Стратегии:**
Синтаксическая нейтральность, левостороннее ветвление (DC + MC), структурный приоритет (контролер в MC)	Кодирование подчинения в DC: CNV + & при ИГ*, Кодирование мишени: ∅

* Это полная ИГ переходной клаузы. В бежтинском языке такая полная ИГ должна быть в номинативе.

Схема 2

Бежтинский язык. 2-я конструкция

Принципы:	**Стратегии:**
Различительность, синтаксическая нейтральность, левостороннее ветвление, структурный приоритет, линейный приоритет	Вынос влево ИГ-контролера второй клаузы, Кодирование подчинения в DC: CNV + & при ИГ, Кодирование мишени: ∅

Схема 3

Арчинский язык

Принципы:	**Стратегии:**
Различительность, синтаксическая нейтральность, левостороннее ветвление, ±структурный приоритет*, линейный приоритет	±вынос влево ИГ второй клаузы, кодирование подчинения в DC: CNV кодирование мишени: ∅

* Знак ± означает, что данный принцип может при некоторых условиях не действовать.

скольку принцип различительности для данной системы нерелевантен.

Чтобы избежать референциальной неоднозначности, в бежтинском языке имеется также 2-я конструкция, см. (5).

Эта конструкция строится на несколько отличной системе принципов и стратегий, см. схему 2. А именно, добавляются принцип различительности и линейного приоритета, а также мотивируемая ими стратегия выноса влево ИГ-контролера из второй клаузы.

Легко видеть, что принципы структурного и линейного приоритета всочетаниисприципом левостороннего ветвления вступают в конфликт: согласно принципу левостороннего ветвления и структурного приоритета ИГ-контролер должен находиться во второй, главной клаузе, а согласно принципу линейного приоритета он должен предшествовать мишени, находящейся в первой, зависимой, клаузе. Преодолевается этот конфликт стратегией выноса влево ИГ-контролера главной клаузы. В результате активированный принцип различительности соблюдается не только в (5а–б), но и в (5в–г).

Принципы и стратегии, действующие в арчинском языке, представлены на схеме 3.

Специфичным в арчинском языке является частичное блокирование принципа структурного приоритета, возможность выноса влево ИГ, не являющейся контролером и отсутствие показателя сочинения на полной ИГ в зависимой клаузе.

4.2. Чамалинский язык

Чамалинские данные приведены в (8). Принципы и стратегии, активированные в чамалинском языке, сформулированы в схеме 4.

Чамалинский язык отличается от 2-й бежтинской конструкции лишь наличием принципа приоритета непереходной клаузы вместо принципа структурного приоритета.

Остаются два вопроса.

Во-первых, почему в (8а-б) имеет место выдвижение ИГ-контролера влево? С точки зрения принципа различительности это избыточно. Однако данная стратегия необходима ввиду действия принципа линейного приоритета.

Схема 4

Чамалинский язык

Принципы:	**Стратегии:**
Различительность, синтаксическая нейтральность, левостороннее ветвление линейный приоритет, приоритет непереходной клаузы	вынос влево ИГ-контролера второй клаузы, кодирование подчинения в DC: cnv + & при ИГ, кодирование мишени: ∅

Схема 5

Хваршинский язык

Принципы:	**Стратегии:**
Различительность, синтаксическая нейтральность, структурный приоритет, линейный приоритет, приоритет непереходной клаузы	вынос влево ИГ-контролера второй клаузы*, кодирование подчинения в DC: cnv + & при ИГ, кодирование мишени: ∅

* Если такой ИГ нет, данная стратегия не реализуется.

Схема 6

Аварский язык

Принципы:	**Стратегии:**
Различительность, синтаксическая нейтральность, левостороннее ветвление, линейный приоритет	Кодирование подчинения в DC: cnv, кодирование мишени: pron, см. (а–г), ∅, см. (в–г), full, см. (а–б), если классное согласование не различает контролеров

Во-вторых, почему в (8в–г) нет выдвижения влево ИГ-контролера? Это объясняется тем, что ИГ-контролер в (8в–г) находится в первой клаузе и уже удовлетворяет принципу линейного приоритета.

Таким образом, принцип линейного приоритета играет в этой системе важную роль и особым образом ограничивает стратегию выдвижения ИГ-контролера.

Наконец, можно видеть, что, в отличие от бежтинского языка, в зависимой клаузе показатель сочинения имеется не только на номинативной полной ИГ, как в (8а), но и на эргативной, как в (8б).

4.3. Хваршинский язык

Действующая в хваршинском языке система принципов и стратегий представлена на схеме 5.

В хваршинском языке не задействован принцип левостороннего ветвления, что делает возможным в (9б-в) порядок клауз MC + DC, ранее не встречавшийся. В свою очередь необходимость такого порядка диктуется одновременной активацией принципов линейного приоритета, структурного приоритета и приоритета непереходной клаузы: непереходная клауза должна быть первой и главной клаузой в предложении. По тем же причинам стратегия выноса влево имеется только в (9а–б).

4.4. Аварский язык

Поскольку в аварском языке действует принцип линейного приоритета, во всех случаях контролер находится в зависимой, а мишень — в главной клаузе.

4.5. Табасаранский язык

Полагаю, что теперь данные табасаранского языка в 2.5 (см. выше) могут быть без труда интерпретированы в систему принципов и стратегий, как это показано на схеме 7.

Все стратегии в табасаранском языке дополнительно распределены в (11) между предложениями (а, в) и (б, г) в соответствии с принципом синтаксической аккузативности. Принцип структурного приоритета таким же образом действует в предложениях (а, в).

Схема 7

Табасаранский язык (говор сел. Дюбек)

Принципы:	**Стратегии:**
Синтаксическая аккузативность, различительность, левостороннее ветвление, ±структурный приоритет, см. (а, в) vs. (б, г), линейный приоритет	Вынос влево ИГ-контролера второй клаузы, см. (а, в), кодирование подчинения в DC: cnv, см. (а, в), temp, см. (б, г), кодирование мишени: ∅, см. (а, в) pron, см. (б, г)

5. Заключение

1. Языковое разнообразие. Подведем итоги рассмотрения организации клаузального сочинения двадцати дагестанских языков (точнее, 23-х идиомов). На нескольких примерах был продемонстрирован феномен разнообразия сочинительных конструкций в дагестанских языках. Полное рассмотрение всех изученных языков см. в [Кибрик 2007]. Важный эмпирический факт состоит в том, что в выборке из 23 идиомов *не встретилось ни одной пары идиомов, имеющих идентичную систему сочинительных конструкций*. Более того, некоторые идиомы имеют более одной сочинительной конструкции. При этом анкетный метод сбора материала не гарантирует, что в действительности дагестанские языки не используют других сочинительных конструкций, по тем или иным причинам не обнаруженных в процессе сбора материала.

Очевидно, что в такой ситуации традиционный таксономический метод классификации зарегистрированных конструкций неприменим. Более того, обычно эффективный структурный метод исчисления в терминах самих этих конструкций также является недостаточным, поскольку он не позволяет ограничить множество всех потенциально возможных конструкций, как не зарегистрированных, так и таких, которые могут быть впоследствии обнаружены или возникнуть в процессе развития дагестанских языков.

В качестве альтернативы в статье предложен принципиально другой подход, а именно метод многофакторного исчисления второго порядка, который непосредственно *исчисляет не конструкции, а объекты когнитивной природы*: факторы, ответственные за образование конкретных конструкций.

2. Метод многофакторного исчисления второго порядка. Этот метод строится на презумпции, что при всем многообразии сочинительных конструкций различия между ними не безграничны. Это многообразие объяснятся ограниченным множеством альтернативных принципов и стратегий, способных объединяться в различные комбинации. Иными словами, ограничение сочинительных конструкций обеспечивается не на уровне самих конструкций, а на уровне тех параметризованных принципов и стратегий, которые лежат в основе построения конкретных сочинительных конструкций. Их репертуар описан в разделе 3, и он действительно покрывает все разнообразие зафиксированных конструкций.

3. Неспецифический характер принципов и стратегий. Особо следует подчеркнуть общеязыковой статус принципов, а также большинства стратегий: сфера их действия не ограничена семантическим сочинением, они проявляются также в построении многих других синтаксических конструкций. Такое понимание архитектуры естественного языка опирается на принцип многофакторности, лежащий в основе выбора средств для построения языковых конструкций. Иными словами, семантическое сочинение отображается в синтаксическую конструкцию исходя из ограниченного репертуара параметризованных элементарных концептов, встроенных в грамматику данного языка. Этот репертуар в свою очередь является подмножестовом универсального репертуара, ограничивающего вариативность естественных языков.

Существенно, что в целом для дагестанских языков *ни один из параметров и ни одна из стратегий не являются обязательными,* но выбор некоторого параметра (и его значений) часто предопределяет последующие выборы. Возможно, техника клаузального сочинения есть цепочка частично детерминированных выборов из ограниченного репертуара альтернатив[27].

4. Мотивированность принципов. Следует обратить внимание на то, что введенные в рассмотрение конкретные принципы и стратегии, являясь формальными конструктами, напоминающими

[27] Такая идеология реализована в ряде других моих работ, затрагивающих реляционную структуру элементарного предложения [Кибрик 2003: 133–187]; «аномалии» личного спряжения, мотивированные дейктической иерархией; семантические роли и когнитивную маркированность [Кибрик 2003: 270–304]; внешний посессор [Кибрик 2003: 307–319]; многофакторные стратегии согласования [Кибрик 2003: 400–450].

конструкты генеративной теории принципов и параметров[28], в то же время глубоко мотивированы когнитивной структурой и имеют содержательную интерпретацию.

Так, *принцип различительности* характеризует не только (и не столько) сочинительные конструкции, а базовый семиотический принцип — обеспечивать существующими в языке знаковыми средствами различение релевантных смыслов языковых выражений.

Принцип синтаксической нейтральности является необходимым свойством *чистых* ролевых языков, последовательно концептуализирующих различия между участниками ситуаций в терминах соответствующих гиперролей. Дагестанские языки следуют эргативной схеме, основанной на гиперролях — агентиве и абсолютиве, подробнее см. [Kibrik 1997, Кибрик 2003: Part 3. Теория элементарного предложения], и относятся к типу семантически эргативных языков. Это означает, что для них гиперролевые характеристики участников ситуации имеют фиксированное кодирование и не могут быть изменены ни в каком контексте. Поэтому, в частности, синтаксический уровень «прозрачен», нейтрален в отношении ролевых характеристик. Действительно, нейтральность синтаксиса в области сочинительных конструкций проявляется в подавляющем большинстве рассмотренных языков.

Принцип синтаксической аккузативности как будто бы предполагает кардинальное отклонение от дагестанского ролевого прототипа — переход на другую систему гиперролей (принципала и пациентива). Однако более внимательный анализ данных этого не подтверждает. Активизация данного принципа имеет другую природу. Известно, что аккузативная конструкция (объединение **A** и **S**) гармонирует с концептом независимого коммуникативного измерения — *топиком*. Зачатки этого измерения проявляются в принципе левостороннего ветвления, см. ниже, являющимся общедагестанской характеристикой. В целом, примерно в половине языков намечается формирование относительно новых для дагестанских языков концептов коммуникативного измерения и их грамматикализация. Это обстоятельство играет существенную роль в кодировании анафорических отношений в сочинительных конструкциях и постепенном наращивании черт синтаксической аккузативности. Множественность типов конструкций позволяет почти пошагово восстановить этот исторический процесс и его мотивацию. Типологически важно, что такой сдвиг происходит в результате кластеризации концептов ролевого и коммуникативного членения, не затрагивая по существу исконной техники кодирования концептов ролевого членения (падежное кодирование актантов и классное согласование).

Принцип левостороннего ветвления, как было отмечено выше, также мотивирован коммуникативным измерением. А именно, вершина составляющей гармонирует с ее *фокусом*, естественная линейная позиция которого — в конце составляющей. Можно отметить, что в широкой типологической перспективе этот принцип реализуется в языках с конечной позицией глагола.

В списке принципов имеется три конкурирующих фактора, влияющих на выбор ИГ-контролера. *Принцип структурного приоритета* ограничивает позицию контролера главной клаузой, а мишень — зависимой. Синтаксическая оппозиция главной и зависимой клаузы гармонирует с коммуникативными концептами *фигура* и *фон* соответственно. *Принцип линейного приоритета* опирается на когнитивно-дискурсивное различие между первичным и повторным упоминанием референта. При первичной номинации референт впервые вводится в рабочую память, а прочие его вербальные упоминания являются побочным эффектом линейности речи. В когнитивной структуре (в рабочей памяти) имеется только один концепт (метка референта), и отсылка осуществляется к нему, а не к вербальному антецеденту. *Принцип приоритета непереходной клаузы* мотивирован тем, что в этом случае нахождение контролера тривиально и требует минимальных вычислительных усилий. Таким образом, этот принцип следует из более общего поведенческого принципа — *принципа экономии усилий.*

5. Мотивированность стратегий. Что касается стратегий, то в большинстве случаев они также формальны только в техническом смысле слова, будучи наделены когнитивной, семиотической или системной мотивацией.

Например, стратегия *выноса влево ИГ-контролера второй клаузы* обычно мотивирована системно. Такая линеаризация не случайна, она вызвана необходимостью снять конфликт между независимыми принципами линейного приоритета и структурного приоритета (или приоритета непереходной клаузы). В некоторых особых случаях сфера действия этой стратегии расширяется, будучи мотивирована более локальным ограничением на линейную цепочку находящихся в разных клаузах контролера и мишени, кодируемых одним и тем же поверхностным падежом, см. конструкцию (6г′) арчинского языка.

Стратегия *кодирования мишени анафорической связи* допускает в дагестанских языках определенный репертуар их оформления. Этот репертуар соответствует универсальному набору средств оформления повторной номинации. Нулевое кодирование мишени в наибольшей степени соответствует когнитивной интерпретации анафорической связи, а также принципу синтаксической нейтральности. Не случайно оно представлено как единственная возможность в половине идиомов.

[28] Хотя я хочу указать, что эти понятия не являются специфическими для данной теории. Например, они используются также в работе [Comrie 2004] без придания им строго терминологического значения.

Как показывает анализ конкретных языков, выбор альтернативных кодирующих средств в определенной степени мотивирован Иерархией маркированности средств повторной номинации:

∅ < PRON < REFL/EMPH < FULL NP

Нулевое кодирование в этом случае означает максимально немаркированную ситуацию, а оформление повторной номинации полной именной группой — максимально маркированную. Личное местоимение кодирует коммуникативно более маркированные номинации, чем нуль. Возвратное (точнее, эмфатическое) местоимение более маркировано, чем личное[29], так как оно дополнительно указывает на смену ожиданий адресата, см. [Кибрик, Богданова 1995, Лютикова 1999].

6. Типология родственных языков. Данное исследование убедительно демонстрирует важность типологического подхода к родственным языкам. Во-первых, он позволяет *избежать ошибок в интерпретации языковых выражений каждого из языков,* что часто неизбежно происходит при их автономном изучении из-за явного или скрытого недостатка аргументов для отсева конкурирующих описательных альтернатив. Типологический подход позволяет тестировать правильность описания языковых выражений конкретного языка и обнаруживать их наиболее адекватную интерпретацию. Во-вторых, он *показывает опасность типологических обобщений, формулируемых на основе статистически нормализованной выборки из языков различных семей,* ограниченной их одинаковой представительностью. Так, если в такую выборку будет включен в качестве представителя дагестанских языков лезгинский язык, то окажется, что при клаузальном сочинении дагестанские языки характеризуются синтаксической аккузативностью (а также не имеют именных классов, латеральных согласных, фарингализации и т. д.).

Полное описание синтаксиса клаузального сочинения всех двадцати трех идиомов дополнительно продемонстрировало, что давнее утверждение С. Андерсона [Anderson 1976], подкрепленное случайными и поверхностно проанализированными примерами, об универсальной синтаксической аккузативности языков с эргативной конструкцией не выдерживает эмпирической проверки языковыми данными дагестанских языков. Лишь в трех языках (*лезгинский, крызский, агульский*) наблюдаются сочинительные конструкции, похожие на аккузативную схему, и лишь в одном языке (*табасаранский*), правда, только в одном из двух проанализированных идиомов, зафиксирована собственно аккузативная схема. Однако при их сравнении с другими языками наглядно видно, что формально аккузативная схема не выделяется дискретно на множестве языков, а выступает как элемент континуума и результат постепенного наращивания средств кодирования концептов коммуникативного измерения, автономного по отношению к ролевому измерению. Более того, в пределах одного идиома агульского языка сосуществуют две конструкции, относящиеся к разным синтаксическим типам.

7. Сочинительные конструкции в исторической перспективе. Если взглянуть на варьирование сочинительной конструкции с исторической точки зрения, то бросается в глаза тот факт, что это варьирование не коррелирует с генетической и ареальной близостью языков. Такая дистрибуция свойств по языкам свидетельствует, что сочинительные конструкции в дагестанских языках неустойчивы, находятся в процессе их формирования. К общедагестанскому фонду могут быть отнесены в первую очередь принципы (исключая коммуникативное измерение[30]), а также ряд стратегий, типа синтаксического подчинения и использования нулевой анафоры[31].

[29] Влияние этой иерархии проявляется и в русском языке (примеры Е. А. Лютиковой): (i) Маша$_i$ встретила гостью$_j$ и ∅$_i$ пошла на кухню. (ii) Маша$_i$ встретила гостью$_j$ и она$_j$ / та$_j$ пошла на кухню. (iii) Маша$_i$ встретила гостью, а сама$_i$ пошла на кухню.

[30] Однако коммуникативно мотивированный принцип левостороннего ветвления является базовым.

[31] Выражаю благодарность Андрею Кибрику, Екатерине Лютиковой и Якову Тестельцу, ознакомившимся с рукописью и сделавшим ряд ценных замечаний.

Литература

1. *Бергельсон М. Б. и Кибрик А. Е.* 1982. Сочинительное сокращение в табасаранском языке // А. Е. Кибрик (ред.) Табасаранские этюды. М.: Изд. МГУ, 66–73.
2. *Кибрик А. Е.* 1979–1981. Материалы к типологии эргативности. Препринты Института русского языка АН СССР. М.
3. *Кибрик А. Е.* 2003. Константы и переменные языка. СПб: Алетейя.
4. *Кибрик А. Е. и Богданова (Лютикова) Е. А.* 1995. САМ как корректор ожиданий адресата // ВЯ, № 3, 28–47.
5. *Кибрик А. Е.* 2007. Принципы и стратегии клаузального сочинения в дагестанских языках. // ВЯ №3: 107–149.
6. *Лютикова Е. А.* 1999. Эмфатическое местоимение wuǯ // А. Е. Кибрик (ред.) Элементы цахурского языка в типологическом освещении. М.: Наследие, 617–629.
7. *Anderson, Stephen.* 1976. On the notion of subject in ergative languages // Ch. Li (ed.) Subject and topic. New York: Academic Press, 1–23.
8. *Comrie, Bernard.* 1988. Coreference and conjunction reduction in grammar and discourse // J. A. Hawkins (ed.) Explaining language universals. New York: Basil Blackwell, 186–208.
9. *Haspemath, Martin.* 2004. Coordinating constructions. An overview // M. Haspelmath (ed.) Coordinating constructions. Amsterdam: Bebjamins, 3–39.
10. *Kibrik, Aleksandr.* 1987. Constructions with clause actants in Daghestanian languages // R. M. W. Dixon (ed.) Studies in ergativity. Amsterdam: Elsevier Science Publishers, 133–178.
11. *Kibrik, Aleksandr.* 1997. Beyond subject and object: Toward a comprehensive relational typology. // Linguistic typology 1, 279–346.
12. *Nichols, Johanna.* 1981. Predicate nominals: a partial surface syntax of Russian. Berkeley: Univ. of California Press.
13. *Nichols, Johanna.* 1986. Head-marking and dependent marking grammar. // Language 62(1), 56–119.

Некоторые сложности автоматизированной лемматизации несловарных словоформ

Some difficulties in automated lemmatization of word forms not contained in the dictionary

Клышинский Э. С. (klyshinsky@mail.ru)

Институт прикладной математики им. М. В. Келдыша РАН

В статье рассмотрены результаты машинного эксперимента по лемматизации несловарных словоформ. Рассматриваются некоторые сложности, возникающие в процессе лемматизации. В заключении делается вывод о невозможности на данный момент полностью автоматизировать процесс включения новых слов в морфологический словарь.

1. Введение

На данный момент создано большое количество компьютерных программ словарной морфологии. Некоторые из них сочетают сразу несколько функций (см., например, [1, 2]). Но даже в этом случае в первую очередь в словарь должны быть занесены лексемы, которым в дальнейшем будет привязываться другая информация.

В связи с активным наполнением словарей последнее время начали активно развиваться методы автоматизированной лемматизации словоформ [3, 4]. Каждая серьезная система морфологического анализа обладает возможностью порождения гипотез относительно нормальной формы и набора параметров незнакомого слова. Однако на практике данные методы не позволяют полностью автоматизировать процесс лемматизации несловарных словоформ, так как количество порождаемых гипотез зачастую слишком велико.

В данной работе мы рассмотрим результаты машинного эксперимента по автоматической лемматизации несловарных (то есть отсутствующих в данном словаре) словоформ. Также мы попытаемся рассмотреть некоторые трудности, которые мешают успешной автоматизации данного процесса.

2. Метод исследования

В данной работе была поставлена цель оценить эффективность использования системы автоматической лемматизации для задач наполнения морфологического словаря. Исследования проводились на подсистеме морфологического анализа системы «Кросслятор 2.0» [2]. Объем словаря — почти 160 тыс. словооснов. Для тестирования использовалось несколько корпусов текстов, представленных в открытом доступе в сети Интернет:

1. фрагмент Национального корпуса русского языка, объем более 900 тыс. словоформ по которым было порождено более 36 тыс. несловарных лексем и 1,7 млн. словарных лексем;
2. информация с сайта bash.org.ru, объем более 620 тыс. словоформ по которым было порождено более 65 тыс. несловарных лексем и более миллиона словарных лексем;
3. текущие новости с портала lenta.ru за период с марта 2005 по декабрь 2008, объем более 24,4 млн. словоформ по которым было порождено более 1,1 млн. несловарных лексем и около 46,6 млн. словарных лексем;
4. текущие новости с портала rbc.ru за период с января 2003 по декабрь 2008, объем более 17,3 млн. словоформ по которым было порождено около 2,1 млн. несловарных лексем и более 32,8 млн. словарных лексем;
5. литературный портал lib.ru, объем более 688,5 млн. словоформ по которым было порождено около 37,1 млн. несловарных лексем и более 1,3 млрд. словарных лексем;
6. материалы конференции Диалог с 2003 по 2008 год, объем около 730 тыс. словоформ по которым было порождено более 41 тыс. несловарных лексем и около 1,3 млн. словарных лексем;

Выделение словоформ проводилось простейшим путем, то есть не учитывалась возможность

разрыва слова знаком препинания, например, дефисом. В связи с этим такие слова, как «как-то», «кто-нибудь», «Аддис-Абеба» и другие, рассматривались по частям и, как следствие, могли попасть в список несловарных, хотя слово целиком находится в словаре. Однако фрагменты общеупотребительных слов (например, «нибудь») и фрагменты, являющиеся самостоятельными словами («как», «то»), к началу эксперимента были представлены в словаре.

Омонимия никак не разрешалась, то есть одна словоформа могла участвовать в полученном результате несколько раз. Однако, как это видно из приведенных выше цифр, средний уровень омонимии не превышал двух.

В ходе выдвижения гипотез о лемме несловарной словоформы проводилась активная фильтрация полученных результатов. При этом использовалось несколько сильных, но интуитивно верных положений.

1. Гипотезы, порожденные на основе редковстречающихся парадигм, в рассмотрение не брались. Под редковстречающейся понималась парадигма, количество слов которой не превышало заданного порога. При среднем количестве слов в одной парадигме около 50, брались два значения этого порога: 5 и 10 слов.
2. Для словарных слов, принадлежащей одной парадигме, определялся список букв, заканчивающих их псевдоосновы. В случае, если для словоформы выдвигалась гипотеза о ее принадлежности к данной парадигме, и если при этом ее псевдооснова не оканчивалась ни на одну из полученных букв, то такая гипотеза отвергалась.
3. Отсеивались гипотезы, образованные от словоформы, встретившейся единственный раз в исследуемом корпусе и являющиеся единственной словоформой, использованной в данной парадигме. Предполагалось, что подобная словоформа скорее всего написана с ошибкой. Исключение делалось для парадигм не изменяющихся слов (то есть содержащих единственную позицию в парадигме).
4. Считалось, что псевдооснова несловарных словоформ, объединяемых в рамках одной парадигмы, должна содержать хотя бы один символ.

Применение этих положений позволило сократить количество анализируемой информации до приемлемого уровня, позволившего перейти к процессу кластеризации словоформ [5].

При кластеризации объединялись все словоформы с одинаковой псевдоосновой и образованные по единой парадигме. При этом считалось, что одна и та же словоформа может быть омонимичной, и ей разрешалось входить в несколько гипотез. Подобный подход позволил несколько улучшить результат, оставив в нем правильные варианты. При этом, однако, количество оставляемых гипотез заметно возросло.

После кластеризации проводилось отсеивание полученных лемм по критерию максимальной встречаемости словоформ, вошедших в лемму. То есть для каждого слова определялось сколько раз оно встретилось в тексте. Далее эти значения суммировались по парадигмам и оставлялись лишь парадигмы с максимальной сумой.

В итоге генерировалось два списка лемм (словарных и несловарных) с привязанными к ним словоформами. Для каждого списка генерировалась статистика заполнения парадигм, то есть количество гипотез в зависимости от процента позиций, занятых в парадигме. Для данной статистики анализировалось количество парадигм, заполненных более чем на половину. Предполагалось, что подобные парадигмы дают приемлимый результат с точки зрения количества порождаемых гипотез. На самом же деле эффективная работа специалиста, оценивающего результаты лемматизации, возможна лишь при заполнении парадигм более чем на 80%.

3. Результаты исследования

По результатам эксперимента было выяснено, что большой процент слов, отсутствующих в словаре — это словоформы, написанные с ошибкой. Однако применение алгоритмов орфокоррекции для отсеивания ошибочных результатов здесь невозможно, так как многие новые слова отличаются на одну-две буквы от имеющихся в словаре. Не нарушая общности рассуждений предположим, что в словаре уже имеется слово «угол», но отсутствует слово «уголь». Тогда при анализе второго слова с учетом ошибок, слово «уголь» не будет добавлено к незнакомым словоформам, а увеличит встречаемость слово «угол» на единицу. Вторую большую группу составляли имена собственные. И лишь на последнем месте по количеству находятся слова из специальной или редко используемой общеупотребительной лексики. Для корпуса текстов из библиотеки Мошкова дополнительный хаос внесли тексты на белорусском и украинском языке. Точная оценка относительных размеров групп не проводилась в связи с большим объемом исследуемых корпусов.

Отдельную проблемы составляют слова, омонимичные как словарным, так и несловарным словоформам. В этом случае словоформа будет отнесена к словарным, а в парадигме соответствующей несловарной леммы будет «дырка». Данная проблема не исследовалась в нашей работе и нуждается в отдельной проработке.

В ходе исследования полученных результатов было выяснено, что практически ни одна группа гипотез, объединенных одним списком словоформ, не содержала в себе единственную и верную гипотезу. Среди прочего, это связано с тем, что в русском языке встречаются парадигмы, объединяющие один и тот же набор флексий, однако приписывающие

им различные наборы параметров. Так, для слова «админ», встретившегося на сайте www.bash.org.ru, порождались следующие леммы. «-» означает пустой постфикс. В скобках написаны словарные представители парадигмы.

Единственное число	**им. п.**	**род. п.**	**вин. п.**	**дат. п.**	**тв. п.**	**пред. п.**
АДМИН (ТЕЛЕФОН) м.р., неодуш	—	А	—	У	ОМ	Е
АДМИН (ТОН) м.р., неодуш	—	А	—	У	ОМ	Е
АДМИН м.р., неодуш	—	А	—	У	ОМ	Е/У
АДМИН (АКТИВИСТ) м.р., одуш	—	А	А	У	ОМ	Е
АДМИН (ОПЕР) м.р., одуш	—	А	А	У	ОМ	Е
Множественное число	**им. п.**	**род. п.**	**вин. п.**	**дат. п.**	**тв. п.**	**пред. п.**
АДМИН (ТЕЛЕФОН) м.р., неодуш	Ы	ОВ	Ы	АМ	АМИ	АХ
АДМИН (ТОН) м.р., неодуш	А/Ы	ОВ	А/Ы	АМ	АМИ	АХ
АДМИН м.р., неодуш	Ы	ОВ	Ы	АМ	АМИ	АХ
АДМИН (АКТИВИСТ) м.р., одуш	Ы	ОВ	ОВ	АМ	АМИ	АХ
АДМИН (ОПЕР) м.р., одуш	А/Ы	ОВ	ОВ	АМ	АМИ	АХ

Таким образом, даже один и тот же набор словоформ может быть различным образом размещен в различных парадигмах.

Для каждого из корпусов исследовалось количество лемм в зависимости от процента заполнения их парадигм. Дело в том, что по единственной словоформе довольно сложно корректно предсказать всю лемму. В связи с этим большое количество не полностью заполненных парадигм позволяет говорить о большом количестве ручной работы, которую придется проделать для лемматизации.

Для словарных слов была получена статистика, представленная на Рис. 1. На данном рисунке представлено распределение лемм по степени заполненности их парадигм для различных исследованных корпусов. Из нее видно, что с ростом объема корпуса растет и количество лемм, для которых встретилась большая часть их словоформ. То же самое можно видеть и на графике, представленном на Рис. 2. Здесь можно увидеть процент парадигм, заполненных не менее чем на 50%, в зависимости от логарифма объема корпуса. Данный график представляет собой, по всей видимости, сигмоиду, и из него видно, что на какой-то момент наступает насыщение. Для получения более точных результатов необходимо провести еще несколько экспериментов в промежуточных точках и точках на концах интервалов. Слова с не полностью заполненной парадигмой могут относиться, например, к специальной редковстречающейся лексике или к словам с дефектной парадигмой. В связи с этим дальнейшее увеличение объема корпуса не приведет к значительному изменению результатов. Так, на корпусе текстов из библиотеки Мошкова, процент парадигм, заполненных более чем на 50%, составил 99%.

Рис. 1. Зависимость относительной встречаемости парадигм от их заполнения (для словарных словоформ)

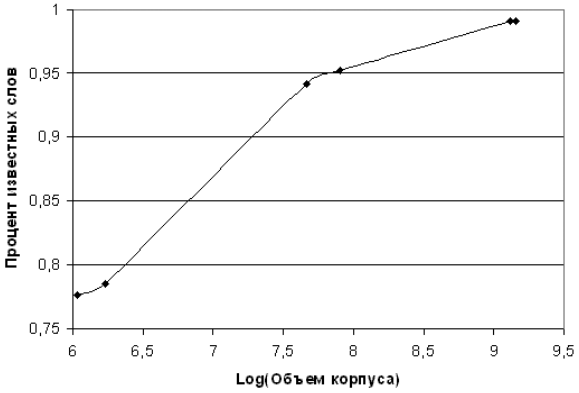


Рис. 2. Зависимость количества парадигм с заполнением >50% от логарифма объема базы (для словарных словоформ)

Рис. 3. показывает распределение лемм по степени заполненности их парадигм для несловарных словоформ. Легко видеть, что оно характеризуется противоположными тенденциями. Большая часть словоформ размещаются в парадигмах с заполнением менее 50%. При увеличении объема корпуса парадигмы постепенно перетекают из левой половины графика в правую, однако общая тенденция сохраняется. Это может быть объяснено, например, сохранением процента слов, содержащих ошибку, при увеличении объема базы. Также можно предположить, что с увеличением объема базы увеличива-

ется и число имен собственных, никогда до сих пор не встречавшихся, слов, относящихся к специальной лексике, сленгу, «новоязу» (образованию новых слов, значение которых ясно из контекста или используемой основы: «даунлоадить», «сторублируйте», «мазелин», …) и другим явлениям языка. Пик при значении 100% заполнения парадигмы связан с наличием неизменяемых слов, но в гораздо большей мере определяется словами с полностью заполненной парадигмой. Это говорит о том, что слова чаще всего либо встречаются лишь в нескольких формах на протяжении всего текста, либо (значительно реже) употребляются во всех своих словоформах. Следует заметить, что без предварительной фильтрации результатов, процент слабо заполненных парадигм был бы еще выше.

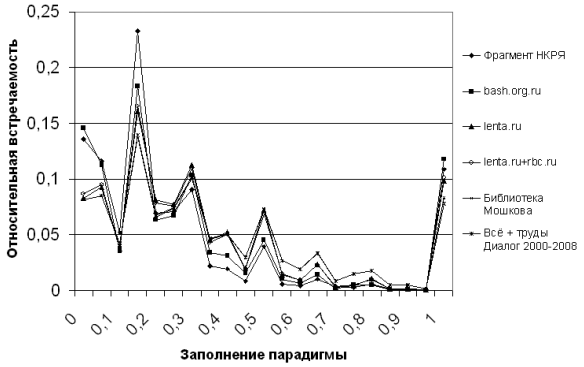


Рис. 3. Зависимость относительной встречаемости парадигм от их заполнения (для несловарных словоформ)

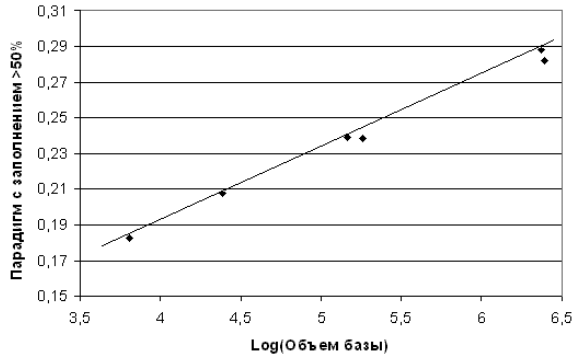


Рис. 4. Зависимость количества парадигм с заполнением >50% от логарифма объема базы (для несловарных словоформ)

Перетекание парадигм в область с заполнением >50% является довольно значительным (с 18 до 28%), но требует экспоненциального роста объема базы. Рис. 4. показывает зависимость количества таких парадигм от логарифма объема базы. Полученные данные неплохо апроксимируются прямой линией, из которой, однако, выпадают две точки, соответствующие относительно небольшому изменению объема базы. Можно предположить, что отклонение основывается на резком изменении используемой лексики. Следует заметить, что отсеивание белорусских и украинских текстов из библиотеки Мошкова могло изменить вид аппроксимирующей функции. Данный вопрос нуждается в дополнительных машинных экспериментах.

4. Обсуждение результатов

Анализ результатов показывает, что полностью автоматизировать процесс лемматизации несловарных словоформ на данный момент невозможно. Исключение может составить морфология, основанная на стемминге, когда нас больше интересует связь словоформы с нормальной формой или ее псевдооснова [4]. В этом случае можно пренебречь некоторыми нюансами. В противном случае, как это было показано выше, существует очень низкая вероятность получить по набору словоформ единственную включающую их парадигму.

Большое количество ошибок, встречающихся в любых текстах, зашумляет выход системы лемматизации и требует длительного ручного труда по отделению корректных вариантов от ошибочных. К счастью, возможностей для ошибки предоставляется очень много, и поэтому большинство ошибок встречаются один или два раза и отсеиваются на этапах фильтрации или кластеризации. Однако некоторые ошибочные словоформы могут войти в состав других парадигм, изменив тем самым результаты кластеризации не в лучшую сторону. Кроме того, у многих авторов существуют, так сказать, «любимые» ошибки, когда одна и та же ошибка допускается многократно в различных словоформах. Использование отредактированных источников должно облегчить труд, однако количество таких источников мало. Как показал эксперимент, даже в НКРЯ имеются отдельные лексические недостатки, не говоря об использовании разговорной лексики, не облегчающей труд системы лемматизации и специалистов.

И, наконец, дальнейшее развитие морфологического словаря упирается в экспоненциальный рост объема обучающего корпуса, что влечет за собой рост «шума» в результатах работы системы. Все вместе это ведет к постепенному замедлению работы по лемматизации. Общие трудозатраты на заполнение морфологического словаря имеют экспоненциальную форму.

Однако качественный скачок может быть получен путем применения ряда других подходов. Так, например, анализ окружения неизвестного слова существенно снижает омонимию, возникающую при выдвижении гипотез о принадлежности данного слова той или иной парадигме. Одним из методов, которые здесь можно применить, является метод триграмм [6]. Сам вопрос применения метода сня-

тия омонимии при работе с неизвестными словами в английском языке уже исследовался, например, в работе [7]).

Для получения лучших результатов можно предложить использовать специализированные корпуса научной направленности. Во-первых, количество ошибок в них существенно ниже, чем в большинстве современных литературных источников. А во-вторых, пополнение словарей ведется в основном за счет специальной лексики, которая как раз и расположена компактным образом в научных корпусах.

Резюмируя, следует сказать, что даже в текущем состоянии автоматизированный (а не автоматический) процесс лемматизации позволяет существенно сэкономить время специалиста, пополняющего морфологический словарь.

Литература

1. *Сидорова Е. А.* Многоцелевая словарная подсистема извлечения предметной лексики // Труды международного семинара Диалог'2008 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2008. С. 475–481.
2. *Елкин С. В., Клышинский Э. С., Стекляннников С. Е.* Проблемы создания универсального морфосемантического словаря // Сб. трудов Международных конференций IEEE AIS'03 и CAD-2003, том 1, Дивноморское. 2003
3. *Сегалович И., Маслов М.* Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // Труды Международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям. Казань: ООО «Хэтер», 1998. Т. 2. С. 547–552
4. *Ляшевская О. Н., Кобрицов Б. П., Сичинава Д. В.* Автоматизация построения словаря на материале массива несловарных словоформ // Интернет-математика 2007
5. *Черненьков Д. М.* Автоматизированное пополнение морфологического словаря на массиве текстовых документов // Труды научно-практического семинара «Новые информационные технологии-12». М.: МИЭМ, 2009. С. 138–141.
6. *Сокирко А. В., Толдова С. Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп) // Интернет-математика-2005
7. *Mikheev A.* Automatic rule induction for unknown word guessing // Computational Linguistics, 23(3): 405–423, 1997

Синтаксическая несовместимость как свойство линейной организации русского предложения[1]

Syntactic incompatibility as a property of the linear organization of a Russian sentence

Кобзарева Т. Ю. (stamstam@mtu-net.ru)

Российский государственный гуманитарный университет

Обсуждается одно из свойств организации линейной структуры русского предложения, важное для автоматического анализа, — синтаксическая несовместимость: невозможность одновременного появления некоторых компонент предложения в его фрагментах, заданных знаками препинания и сочинительными союзами. Рассматривается использование этого свойства на разных этапах автоматического анализа.

1. Введение

При синтаксическом анализе мы интерпретируем линейную структуру предложения (S): «В основе всего структурного синтаксиса лежит соотношение между структурным порядком и порядком линейным» [1]. Используя словарную информацию и информацию, которая закодирована порядком слов и знаков препинания (ЗП), мы ставим в соответствие анализируемым фрагментам линейной последовательности слов и ЗП некоторую грамматическую структуру.

Речь пойдет об особенностях линейной организации русского S, которые используются в системе поверхностно-синтаксического анализа русского S, разрабатываемой в настоящее время в РГГУ, и которые, как нам представляется, важны для любой системы, ориентированной на максимальное использование информации порядка слов и ЗП при минимизации лексико-семантической информации.

Специфику системы определяет впервые использованная при синтаксическом анализе русского предложения иерархия процедур анализа, рассмотренная и обоснованная в [2]. Самой важной ее особенностью является то, что сегментация — поиск границ сегментов (простых и придаточных S, деепричастных, причастных и др. обособленных оборотов) с одновременным элиминированием разрывов, возникающих при вложении сегментов в сегменты — предшествует моделированию внутренней структуры сегментов и отношений между ними, т. е. до построения большей части синтагматических связей.

Система состоит из 6 процедурно независимых модулей:

1) постморфология — несловарные проблемы морфанализа: обработка имен собственных и названий, окказиональных аббревиатур, числительных [13];
2) разрешение частичной омонимии совпадения форм разных частей речи [3];
3) предсегментация — построение проективных фрагментов именных и предложных групп [14], т.е. поиск хозяина необособленного согласованного определения, выраженного прилагательным или причастием, конструкций с числительными, сложных сказуемых и др., т.е. связей, определяющих единицы линейной структуры, необходимые для сегментации;
4) сегментация — построение сегментов [9];
5) моделирование структуры синтагматических связей внутри сегментов;
6) построение связей сегментов [15];.

На каждом этапе встают проблемы разрешения неоднозначности интерпретации S, которые часто порождаются потенциальной неоднозначностью некоторой компоненты его линейной структуры — отдельного слова [3,4], ЗП [5] или некоторого фрагмента предложения — последовательности слов и\ или ЗП [6,7]. Иногда это имплицирует истинную неоднозначность, то есть возможность нескольких правильных с т. зр. носителей языка пониманий всего S, но чаще правильным бывает только одно из гипотетических значений.

[1] Доклад подготовлен при частичной поддержке РФФИ грант № 09-06-00275-а

Ниже мы рассмотрим одно из свойств линейной структуры русского S — свойство синтаксической несовместимости — и покажем, что это свойство на разных уровнях анализа — в разных модулях — может помочь в определенных ситуациях найти правильную интерпретацию линейной структуры S.

2. Используемые понятия

В основе предлагаемого подхода лежат представления Люсьена Теньера о структуре ситуаций, представленных простыми S: «Глагольный узел, который является центром предложения в большинстве европейских языков, выражает своего рода маленькую драму. Действительно, как в какой-нибудь драме, в нем обязательно имеется действие, а чаще всего также действующие лица и обстоятельства» [1].

В каждом языке есть много способов объединения этих «маленьких драм» в одно S, где несколько ситуаций образуют синтаксическое единство. При этом в русском S исходные ситуации, каждая из которых могла бы быть выражена отдельным простым S с некоторой вершиной–предикатом, представлены, во-первых, простыми S, составляющие основу любого S: простые в составе сложносочиненных и простые–главные в сложноподчиненных. Во-вторых, — трансформами исходных простых S — придаточными S и разными оборотами — деепричастными, причастными и другими, которые подчинены простым-главным или таким же трансформам. В качестве обобщающего названия для всех таких частей S будем использовать уже упомянутый выше термин сегмент [8,9].

В русском письменном языке исторически сложилась традиция задавать границы сегментов при помощи ЗП. В настоящее время существуют правила пунктуации, которые регламентируют соответствующее использование ЗП. В некоторых ситуациях границами сегментов могут быть и сочинительные союзы (СС) или комбинации ЗП и СС [9]. ЗП, СС и морфологически автономные [12] комбинации ЗП и СС будем называть операторами (F).

При сочинении двух слов в тексте между ними тоже обязательно есть сочиняющий их F. Одна из проблем анализа связана с омонимией операторов: F могут не только быть границами сегментов, но и манифестировать сочинение слов и\или сегментов [5].

3. Проективность и линейная организация предложения

Свойство проективности, открытое Теньером применительно к подчинительным связям слов в простых S, хорошо исследовано [10]. Если, изображая граф связей слов, мы по горизонтали сохраняем порядок слов в S, а слова — узлы графа располагаются на разных уровнях соответственно иерархии подчинительных связей, то «для правильных синтаксических структур, изображенных в виде дерева, <…> перпендикуляры, опущенные из узлов дерева, не пересекают его ветвей».

Легко заметить, что в проективном фрагменте текста между двумя связанными словами могут находиться только слова, прямо или опосредованно им подчиненные. Из проективности графа зависимостей слов внутри сегмента вытекает проективность сегментов: непосредственные или опосредованные слуги предикативных вершин сегментов не могут находиться внутри линейного пространства сегмента, имеющего другую вершину. Это означает, что каждая из исходных ситуаций, которую манифестирует один сегмент, локализуется в своем участке линейного пространства S: S делится операторами на сегменты, и каждый сегмент представлен отдельной частью линейного представления S, не пересекающейся с другими сегментами.

Проективность связей внутри сегментов в каких-то случаях нарушается: нарушение внутрисегментной проективности обычно не ведет к непониманию или, во всяком случае, не очень сильно его осложняет. Но проективность сегментов практически никогда не нарушается. Соответственно, сегмент — часть линейной структуры предложения, где могут находиться только непосредственные или опосредованные слуги вершины этого сегмента.

У каждого сегмента обязательно есть левая и правая границы, задаваемые операторами (кроме левой границы сегмента в самом начале S). При сочинении двух слов в тексте между ними тоже обязательно есть сочиняющий их F. Эта функциональная омонимия операторов осложняет анализ [5].

Линейная структура сегмента — простого-главного или придаточного S может быть дополнительно осложнена сочинением предикатов, при этом в сегменте появляется несколько вершин.

Как было показано в [11], операторы, сочиняющие предикаты, являются границами зон влияния этих предикатов: Fk, «сочиняющий» два предиката, делит отрезок между ними на две зоны — части S, где могут находятся только непосредственные или опосредованные слуги одного из них и не могут — слуги другого.

Таким образом, в S

- между любыми двумя предикатами (будь то вершины разных сегментов или сочиненные сказуемые одного сегмента) **обязательно есть Fk,** разделяющий в линейной структуре зоны их влияния;
- в сегменте не могут находиться непосредственные или опосредованные слуги вершины другого сегмента;

- в зоне влияния одного из предикатов, сочиненных внутри простого или придаточного предложения, не могут находиться слуги других предикатов.

Из сказанного следует и для анализа существенно, что мы еще до построения сегментов знаем, что между каждыми двумя вершинами сегментов обязательно должен найтись хотя бы один оператор.

Выделим предикативные вершины сегментов:

(1) *Хлестаков **порхает** по пьесе, не **желая** толком понять, какой он **поднял** переполох, и жадно **стараясь** урвать все, что **подкидывает** ему счастливый случай. (В. Набоков, далее (Н))*

(2) *Иван, **зная** все это, заблаговременно **запасся** двумя вязками бубликов и колбасою и, **спросивши** рюмку водки, в которой не **бывает** недостатка ни в одном постоялом дворе, **начал** свой ужин, **усевшись** на лавке перед дубовым столом, вкопанным в глиняный пол. (Н. Гоголь)*

На этапах, когда сегменты еще не построены и функции операторов еще не определены, мы можем использовать тот факт организации линейной структуры, что между каждыми двумя вершинами обязательно должен быть Fk.

Безусловными вершинами сегментов являются глаголы в личной форме, деепричастия, краткие причастия и краткие прилагательные. Последние — с учетом возможности их вхождения в сложные сказуемые. Для полного причастия и прилагательного необходимо убедиться, что оно является вершиной сегмента — обособленного определительного оборота.

4. Свойство синтаксической несовместимости

Рассмотренные выше особенности задают простое, но важное свойство организации линейной структуры русского S, которое мы назовем принципом **синтаксической несовместимости**: внутри сегмента, а при наличии сочинения вершин — внутри зоны влияния одной из вершин сегмента — не могут одновременно находится слова, относящиеся к ситуациям разных предикативных вершин.

5. Использование принципа синтаксической несовместимости при синтаксическом анализе

Принцип синтаксической несовместимости может быть использован на разных этапах анализа. Рассмотрим его использование в двух модулях: 1) модуле разрешения морфологической неоднозначности и 2) в модуле сегментации.

1) Использование свойства синтаксической несовместимости при разрешении омонимии частей речи

Этап морфологического анализа является необходимым для синтаксической интерпретации S. На этом этапе порождается, в частности, неоднозначность морфологической интерпретации слов из-за случайных совпадений отдельных словоформ, принадлежащих лексемам разных частей речи.

В русском языке существует около 60 типов омонимии частей речи [2,3]. Важными являются типы омонимии, где одна из омонимичных частей речи — потенциальная вершина сегмента. К таким типам относятся, например:

- личная форма глагола vs. существительное (*белил, берегу, бури, вил…*)
- краткое прилагательное \ краткое причастие vs. наречие (*совершенно, дико, двусмысленно, забавно…*)
- краткое прилагательное \ краткое причастие (vs. существительное (*весел, весом, гол, долги…*)
- деепричастие vs. предлог (*для, благодаря, включая…*)
- деепричастие vs. полнозначное или местоименное полное прилагательное, \полное причастие (*скупая, строгая, заезжая, моя…*)

и др.

Один из способов разрешения омонимии — анализ грамматического контекста: проверка того, какие части речи и ЗП и в каком порядке окружают омоним. Именно такой способ разрешения частичной омонимии принят в описываемой системе. Анализ грамматического контекста хорош тем, что, задавая не очень большой, но достаточно представительный набор конфигураций, разрешающих омонимию каждого типа, мы создаем для русского языка фрагмент грамматики линейной структуры, который не требует лексической конкретизации и который можно далее пополнять и унифицировать.

Из свойства синтаксической несовместимости вытекает одно из условий такого фрагмента грамматики, действительное для всех перечисленных типов омонимии. Простота его проверки и довольно высокая частота линейных ситуаций, когда оно позволяет разрешить омонимию частей речи является веским аргументом для использования его перед более сложным анализом контекста.

Условие синтаксической несовместимости для разрешения омонимии Li= потенциальная предикативная вершина vs. Ln= не предикативная вершина: если у нас есть частичный омоним=Lo, одно из значений которого потенциально имеет синтаксическую функцию предикативной вершины, а в исследуемом контексте есть неомонимичный морфологический предикат=Pr и между Lo и Pr нет оператора, Lo не может иметь значение предикативной вершины.

Таким образом, если мы встречаем частеречные омонимы во фрагментах типа (омонимы подчеркнуты, а вершины выделены) *странно на меня* ***смотрит; мог*** *с тех пор стать совершенно другим; существенно* ***изменив****; решение, значительно с этого времени* ***пересмотренное****; на краю села к тому времени уже* ***построили****…; немедленно* ***взяв*** *в руки жгут;,* ***нет*** *у меня клея****; поднялась*** *буря* и т.д., где между омонимом указанных типов — а группы таких типов в словаре велики и в текстах часты — и неомонимичной предикативной вершиной нет F, алгоритм на основании условия синтаксической несовместимости легко разрешает омонимию.

Ниже в примерах (3) и (4) подчеркнуты интересующие нас омонимы и выделены полужирным неомонимичные Pr. В скобках справа от омонима приводится тип омонимии в результате морфанализа, в данном случае — «краткое прилагательное (Abr) vs. наречие (D)», а справа от стрелки — результат разрешения этой омонимии в этих контекстах в силу выполнения условия синтаксической несовместимости.

(3) Рядом стоял воспитатель, и, когда серый резиновый мяч, которым играли в футбол, ***подкатился*** *случайно (Abr\D → D) к его ногам, учитель словесности, инстинктивно продолжая очаровательное предание, сделал вид, что хочет его пнуть, неловко (Abr\D → D)* ***потоптался****, чуть не потерял голову и рассмеялся с большим добродушием.(H)*

(4) Наплакавшись вдоволь, он поиграл с жуком, нервно (Abr\D → D) ***поводившим*** *усами, и потом давил его камнем, стараясь повторить первоначальный сдобный хруст (H)*

2) Использование свойства синтаксической несовместимости при построении сегментов — при разрешении функциональной омонимии ЗП.

Синтаксическая несовместимость позволяет в некоторых случаях чрезвычайно упростить процедуру анализа и на этапе сегментации.

Для того, чтобы иметь возможность находить сегменты в S любой сегментной структуры, удобно использовать, как уже было показано в [2,9] рекурсивную процедуру, строящую сегменты в S справа налево. Для того, чтобы было понятно использование при этом принципа несовместимости, кратко опишем эту процедуру.

Используется следующая модель: русское S состоит из цепочки **β-сегментов** — простых (сочиненных и главных) предложений, каждое из которых может быть разорвано вложением в него некоторого числа **α-сегментов:** обособленных согласованных определений, деепричастных, предложных, вводных и сравнительных оборотов и придаточных S. Каждый **α-сегмент** в свою очередь может быть разорван следующего уровня вставлениями **α-сегментов,** причем количество вставлений, как и количество уровней вставлений, теоретически не ограничено.

Построение сегментов включает в себя 3 этапа:

(1) определение левых границ **α-сегментов: поиск α-отрезков —** безусловных минимальных левых компонент **α-сегментов**;

(2) построение **α-сегментов — поиск** правых границ **α-сегментов** с одновременным восстановлением целостности **α-сегментов,** разорванных вложениями;

(3) определение границ **β-сегментов** с элиминированием разрывов.

Правые границы определяет рекурсивный алгоритм, который, рассматривая поочередно справа налево в S минимальные отрезки **α-сегментов —** отрезки, определенные как безусловные фрагменты **α-сегментов** при установлении левых границ, и двигаясь слева направо от обрабатываемого **α-отрезка,** присоединяет еще не идентифицированные β-отрезки, если это грамматически допустимо, к **α-отрезку,** построенному к текущему моменту анализа.

Процедура повторяется соответственно правилам присоединения с учетом проективности **α-сегментов,** уже построенных правее анализируемого отрезка, пока не находим его правую границу.

Эта процедура начинает построение сегментных матрешек с самого глубокого вложения и позволяет анализировать на основе небольшого базисного синтаксического словаря возможных линейных конфигураций любые допустимые комбинации.

Рассмотрим построение сегментов на следующем примере. Определяем на первом шаге **α-** и **β**-отрезки:

(5) **β**=[*Едва уловимую особенность*], $\boldsymbol{\alpha}_7$=[***отличавшую*** *его сына от всех тех детей*], $\boldsymbol{\alpha}_6$=[*которые по его мнению должны были стать людьми*], $\boldsymbol{\alpha}_5$=[*ничем не* ***замечательными***], **β**=[*он* ***понимал*** *как тайное волнение таланта*], *и*], $\boldsymbol{\alpha}_4$=[*твердо* ***помня***], $\boldsymbol{\alpha}_3$=[*что* ***покойный тесть*** *был композитором*], **β**=[***он*** *в приятной мечте*], $\boldsymbol{\alpha}_2$=[*похожей на литографию, спускался ночью со свечой в гостиную*], $\boldsymbol{\alpha}_1$=[*где вундеркинд в белой рубашонке до пят играет на огромном черном рояле*]. (H)

На втором этапе строятся **α**-сегменты. Ниже (Рис.1) приведена условная схема движения по исходным отрезкам примера (5) при построении в нем **α**-сегментов.

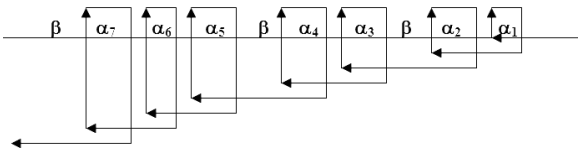


Рис. 1

β-отрезок — это часть S между двумя ЗП, в которой нет слов, маркирующих **α**-отрезки — минимальные правые составляющие **α**-сегментов. При анализе **β**-отрезок, ближайший к **α**-отрезку справа, может быть присоединен к нему тогда и только тогда, если хотя бы одно слово этого отрезка связано подчинительной или сочинительной связью с некоторым словом уже построенной части сегмента. Однако проверка того, имеется ли такая связь, сложна, громоздка и не всегда однозначна.

Использование принципа несовместимости позволяет во многих (но, естественно, не во всех) ситуациях этих проверок избежать. Если в очередном **β**-отрезке, который мы, удлиняя строимый сегмент, пытаемся присоединить, есть слово, которое по своим морфологическим характеристикам не может принадлежать строимому **α**-сегменту, это — в силу проективности сегментов [9] — означает, что построение очередного **α**-сегмента закончено.

Рассмотрим в (5) ситуации, где работает принцип несовместимости. При попытке присоединить к $\boldsymbol{\alpha}_4$=[*твердо* ***помня***] отрезок **β**=[***он*** *в приятной мечте*] мы видим, что в **β**-отрезке есть свободный неомонимичный Им.п. На этапе предсегментации мы уже построили именные группы с существительными в Им. п. в роли определений. Поэтому свободный Им.п. не может появиться внутри деепричастного оборота: деепричастие и свободное (по результатам анализа в модуле предсегментации) существительное в Им. п. не могут относиться к одному сегменту, т. е. синтаксически несовместимы.

Синтаксическая несовместимость деепричастия и существительного в Им.п. позволяет в некоторых случаях разрешить падежную омонимию Им.п.\ Вин п. В примере (2) в деепричастном обороте ***зная*** *все это* можно снять Им.п. у слов *все* и *это*, что в каких-то случаях может предупредить ошибки при анализе сочинения существительных.

При построении согласованного определения $\boldsymbol{\alpha}_5$=[*ничем не* ***замечательными***] ближайший справа **β**=[*он* ***понимал*** *как тайное волнение таланта*], свободный к моменту построения этого сегмента, не м.б. присоединен, так как предикативная вершина — глагол в личной форме — не может находиться в одном сегменте с ***замечательными*** — предикативной вершиной обособленного определения. По тем же причинам этот **β**-отрезок не может быть присоединен и к $\boldsymbol{\alpha}_7$.

Таким образом, если мы строим деепричастный или причастный оборот, **β**-отрезок справа от него при его удлинении не может быть к нему присоединен, если в этом **β**-отрезке есть глагол в личной форме или свободный Им.п.

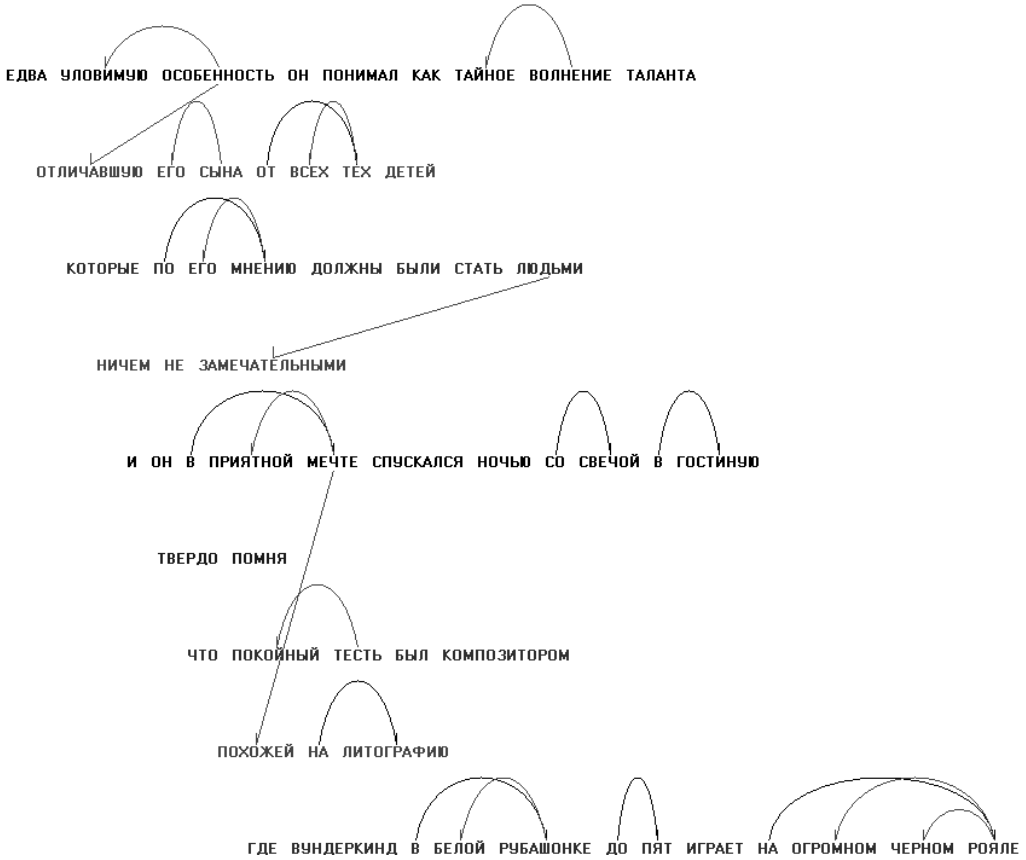


Рис. 2. Визуализация результата анализа примера (5), использующего в ходе рекурсивной процедуры построения сегментов принцип несовместимости (экспериментальная реализация модулей предсегментации и сегментации И. М. Ножова)

Таким образом, принцип несовместимости определят одно из простых и часто работающих в тексте условий присоединения. Для каждого типа **α**-сегментов можно задать список компонент, наличие которых в ближайшем справа **β**-отрезке означает, что этот отрезок не может быть присоединен и при этом — в силу проективности сегментов — построение рассматриваемого **α**-сегмента закончено.

6. Заключение

Рассмотрено свойство синтаксической несовместимости — важное свойство линейной организации русского S, которое можно использовать на разных уровнях синтаксического анализа для разрешения потенциальных неоднозначностей интерпретации S. Это свойство определяется семантикой линейной организации русского предложения и, в частности, является следствием исторически сложившейся в русском языке традиции чрезвычайно семантикализованного использования ЗП в письменном русском языке. Рассмотрение свойства синтаксической несовместимости еще раз показывает, насколько осмысление функциональной семантики русских ЗП, отражающей особенности смысловой организации линейной структуры русского S, может быть полезно для автоматического анализа.

Литература

1. *Теньер Люсьен,* Основы структурного синтаксиса. — М.: Прогресс, 1988.
2. *Кобзарева Т. Ю.* Иерархия задач поверхностно-синтаксического анализа русского предложения // НТИ, Сер.2, №1, 2007 — С 23–35.
3. *Кобзарева Т. Ю., Афанасьев Р. Н.* Универсальный модуль предсинтаксического анализа омонимии частей речи в русском языке на основе словаря диагностических ситуаций // Труды международного семинара Диалог'2002 Протвино 2002. Т. 2. С 258–268.
4. *Зинькина Ю. В., Пяткин Н. В., Невзорова О. А.* Разрешение функциональной омонимии в русском языке на основе контекстных правил // Труды межд. конф. Диалог'2005. — М.: Наука, 2005. С. 198–202.
5. *Кобзарева Т. Ю.* Омонимия и синонимия знаков препинания в русском тексте // Труды Международной конференции Диалог'2005. — М.: Наука, 2005. — С. 233–237.
6. *Иорданская Л. Н.* Синтаксическая омонимия в РЯ (с точки зрения автоматического анализа и синтеза). НТИ, сер. 2. 1967, №5 — С 9–17
7. *Дрейзин Ф. А.* Синтаксическая омонимия // Машинный перевод и прикладная лингвистика. М., 1988
8. *Мельчук И. А.* Автоматический синтаксический анализ. Т. 1. — Новосибирск.: Ред.-изд. отдел Сибирского отделения АН СССР, 1964.
9. *Кобзарева Т. Ю.* Принципы сегментационного анализа русского предложения. Московский лингвистический журнал // М., 2004. Т. 8. №1 — С. 31–80
10. *Иорданская Л. Н.* Автоматический синтаксический анализ. Т. 2. // Новосибирск: Наука, 1967
11. *Кобзарева Т. Ю.* Рекурсивность и проективность сочинительных связей в русском тексте // Компьютерная лингвистика и интеллектуальные технологии Труды Международной конференции Диалог 2006, Бекасово, 31 мая — 4 июня 2006 г. — М.: Наука, 2006. — С. 223–229.
12. *Плунгян В. А.* Общая морфология. Введение в проблематику. М., 2003.
13. *Кобзарева Т. Ю.* Морфанализ in vivo // Труды Международной конференции Диалог'2004, — М.: Наука, 2004 — С. 286–291.
14. *Кобзарева Т. Ю.* Некоторые свойства линейной структуры именных и предложных групп (Поверхностно-синтаксический анализ русского предложения)// Вестник РГГУ. № 8/07, Серия «Языкознание» (Московский лингвистический журнал № 9/2), Москва 2007. — С. 113–130.
15. *Кобзарева Т. Ю.* Построение графа связей сегментов (поверхностно-синтаксический анализ русского предложения) // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог, М.: Наука, 2008 — С. 192–198.

Семантика глагола ***понимать***: от пропозиционального отношения к межличностному

Semantics of the verb ***ponimat'***: from propositional towards interpersonal attitude

Кобозева И. М. (kobozeva@list.ru)

Московский государственный университет им. М. В. Ломоносова, Москва, Россия

Рассматривается глагол *понимать* с прямым дополнением — именем лица. Выделяется и описывается 6 его значений, отражающих разные интенциональные состояния: рациональные, эмоциональные, и межличностные. Развитие у *понимать* ненталных значений объясняется действием механизма, противоположного механизму социо-когнитивного конфликта.

Счастье — это когда тебя понимают.

Глагол *понимать* — один из самых употребительных ментальных глаголов русского языка. По частоте встречаемости в Национальном корпусе русского языка (НКРЯ) его превосходят только *знать* и *думать*, а *сознавать*, *осознавать*, *постигать*, *уяснять*, которые рассматриваются как его синонимы и используются в толковых словарях для описания его значения, уступают ему на порядок. Приведем общее количество вхождений этих слов в НКРЯ (со снятой омонимией):

Знать / узнать	288571
Думать / подумать	180360
Понимать / понять	165060
Считать / посчитать / счесть	92474
Сознавать	6349
Осознавать/осознать	5843
Уяснять / уяснить	1211

Сфера употребления глагола *понимать* существенно шире той, которую имеет его основной переводной эквивалент в английском языке — *understand*. Беглый просмотр параллельного англо-русского подкорпуса НКРЯ показал, что практически всегда (за исключением, пожалуй, только вводных употреблений), глаголу *understand* соответствует *понимать*. Обратное же неверно. Мы употребляем глагол *понимать* не только в тех случаях, где в английском используется *understand* или *comprehend*, но и в тех, где употребляются *know* ‘знать’, *realize* ‘осознать’, *follow* ‘следовать’, *see* ‘видеть’, а в контексте отрицания — *wonder* ‘удивляться’, *be surprised* ‘быть удивленным’ и др., см. например:

(1) а. *His wife said, “I don't* ***know*** *why I should be so hungry.*
б. *— Не* ***понимаю****, почему мне так хочется есть, — сказала его жена.*

(2) а. *I don't think you* ***realize*** *how important you are, to our happy world as it stands now.*
б. *Вы даже не* ***понимаете****, как вы нужны в этом счастливом мире сегодняшнего дня.*

Глагол *понимать* не обделен вниманием семасиологов. Однако различные его значения и употребления исследованы очень неравномерно. А о том, что таковых много, косвенно свидетельствует многообразие его синтаксических свойств. Так, он употребляется и как предикат с сентенциальным актантом (3), и как двухвалентый переходный глагол (4), и как трехвалентный глагол (5), и как вводный глагол (6):

(3) *Однако учёный* ***понимал****, что подобные языки по-прежнему ориентированы на машину.*
(4) *Я вас* ***понимаю*** *и уважаю.*
(5) *Мы* ***понимали*** *под взяткой совершение неких действий должностными лицами…*
(6) *В руке он держал дипломат и напоминал слесаря-сантехника, который шёл себе «по вызову», прихватив «струменты», зашёл в помещение чинить краны, а тут,* ***понимаешь****, люди стихи читают.*

Наиболее изученным в семантическом аспекте является глагол *понимать* с сентенциальным актантом (*X понял / понимает, что P*)[1], который обозначает в форме СВ событие, состоящее в достижении Х-ом знания посредством определенных мыслительных операций, а в форме НСВ — ментальное свойство Х-а, являющееся результатом этого достижения. Вслед за М. А. Дмитровской назовем это употребление «пропозициональным» [Дмитровская 1985, 106]. Наиболее эксплицитное толкование этого употребления (далее — Т_А) дал Ю. Д. Апресян:

> (Т-А) *A понимает, что Q* = ‘В момент t_0 А **знает** или **представляет**, что *Q*; это знание или представление возникло в результате того, что до t_0 А **знал** что-то о ситуациях, связанных с Q, и **думал** о чем-то, связанном с Q; **знание**, что Q, делает возможным **знать** или **представить**, что может произойти после t_0’ [Апресян 1995, 414]

Другие употребления глагола *понимать* исследованы в меньшей степени. Некоторые из них были очень тонко, хотя и неформально, описаны в упомянутой работе М. А. Дмитровской.

Мы предприняли исследование именно непропозициональных употреблений глагола *понимать,* причем только тех из них, в которых он является двухвалентным и его вторая валентность заполнена именем лица. Эта конструкция интересна тем, что в ней *понимать* может обозначать не только ментальные действия и состояния, но и, как мы попытаемся показать, эмоциональное состояния, и межличностные отношения. Материалом исследования служили контексты из НКРЯ, а целью — построение эксплицитных толкований, на основе которых можно было бы понять причины развития у данного ментального предиката значений, относящихся к сфере межличностных отношений.

Итак, в конструкции $N_{им}$ *понимает* $N_{вин}$, где N — имена лиц реализуются следующие семантические варианты глагола *понимать*.

I. Понимание как знание субъектом Х свойств объекта Y, полученное в результате накопления информации о нем и опыта взаимодействия с ним — вариант *понимать*$_{когн}$, ср.:

(7) а. *Через куначество русские и горцы ближе узнавали и лучше* ***понимали*** *друг друга…*
б. *Он не любил и не* ***понимал*** *крестьян, хотя вырос в деревне и был сыном фермера.*

Этот вариант *понимать* близок к тому варианту пропозиционального *понимать*, который в [Дмитровская 1985] рассматривается как предикат пропозиционального достижения *понимать*-5. И. М. Богуславский даже предлагает для пропозиционального *понимать, что P* и переходного *понимать* в случаях типа *понимать музыку (детей)* единое толкование (далее Т_Б):

> (Т_Б) *X понимает Y* = ‘То, что Х обработал или обычно обрабатывает компонентом своей психики W, обычно разумом, факты, связанные с У-ом, каузировало то, что Х (а) имеет или (б) начинает иметь в сознании истинную информацию Z о существенных свойствах Y-а’ [Богуславский 1984, 623]

Очевидна близость, но не тождественность (Т_А) и (Т_Б), и мы согласны с Ю. Д. Апресяном (см. [Апресян 1995, 414, сноска 8]) в том, что (Т_Б) репрезентирует другое значение *понимать*. Синтезируя идеи М. А. Дмитровской, Ю. Д. Апресяна и И. М. Богуславского, мы получаем для *понимать* в примерах типа (7) такое толкование:

> (Т1) ***X понимает***$_{когн}$ ***Y-а*** *= X понимает, какой Y (на самом деле)* = ‘в результате того, что Х располагал большим количеством фактов, связанных с Y-ом, и думал о чем-то, связанном с У-ом, Х знает существенные свойства Y-а, то есть знает об У-е то, что позволяет ему правильно оценивать и предсказывать поведение У-а’.

В этом значении глагол можно в некотором смысле считать фактивным. Ср. аномальность фраз типа *?Он (не)правильно понимает горцев, ?Он понимает горцев, но они совсем не такие.* Оценочное отношение Х-а к Y-у не фиксировано: Y может быть плох, хорош или аксиологически безразличен для Y-а. *Понимать*$_{когн}$ имеет обе видовые формы, обозначая в СВ ментальное событие, а в НСВ — ментальное свойство и сочетается с обстоятельствами степени, которые свидетельствуют о градуированности этого типа понимания: с накоплением фактов Х может узнавать все больше существенных свойств Y-а и все точнее предсказывать его поведение. В качестве объекта *понимать*$_{когн}$ часто выступают нарицательные ИГ во множественном числе с родовым статусом — *дети*, *горцы*, *крестьяне* и т. п., поскольку выделить существенные свойства у категорий (в случае лиц — возрастных, социальных, гендерных и т. п.) с некотором смысле проще: за сравнительно короткое время, наблюдая разные экземпляры одной категории, можно выявить их общие не необходимые, но существенные свойства. В случае индивидов для выявления таких свойств обычно требуется более продолжительное время (ср. *пуд соли вместе съесть*).

II. Понимание Х-ом текста, произведенного Y-ом, на любом из уровней глубины его интерпретации— вариант *понимать*$_{ком}$. Примеры:

[1] Интересно, что в словаре Ожегова это употребление не представлено.

(8) а. *Он был полиглотом и умёл говорить, кажется, на всех языках мира, в том числе на русском и польском, — и на всех ужасно плохо, еле понятно. Но мы с ним* ***понимали*** *друг друга*.

б. *Маркос слушал индейцев — и* ***не понимал*** *их! Так же, как индейцы* ***не понимали*** *студиозусов-революционеров. Не потому, что не знали языка: марксистская теория ничего общего не имела с реалиями Лакандонской сельвы*

Реализуется в контексте описания коммуникативного взаимодействия X-а и Y-а, который, как правило, включает слова из семантического поля языка и речевой деятельности. При утверждении имплицируется уверенность говорящего в том, что полученная X-ом информация именно та, которую Y имеет в виду, но эта импликация легко погашается при помощи наречий с прямым значением AntiVer и AntiBon (ср. *Он неправильно / плохо понял оратора*). Оценочные импликации в отношении лица Y, как и в варианте I, отсутствуют. Другая импликация, фоновая, связывает X-а и Y-а некоторым слабым отношением: ‘представления X-а и Y-а о мире частично совпадают’.В этом варианте *понимать* также имеет обе видовые формы, обозначая в СВ ментальное событие, а в НСВ —ментальное свойство и сочетается с обстоятельствами степени, которые свидетельствуют о градуированности этого типа понимания (по его точности или глубине). Типичный объект здесь, в отличие от ЛСВ I, — конкретно-референтный, но другие денотативные статусы, разумеется, также возможны.Для этого варианта предлагается толкование (Т4):

(Т2) ***X понимает$_{ком}$ Y-а*** = ‘когда X воспринимает высказывание Y-а, X получает ту информацию, которую Y имел в виду’ Импликация: ‘знания и представления X-а и Y-а о мире частично совпадают’

III. Понимание X-ом причин или мотивов поведения Y-а — *понимать*$_{пов}$. Примеры:

(9) а. — *Не могу не спросить… о нашумевшем деле военного журналиста Григория Пасько… В прессе на этот счёт говорилось немало. — Отчасти я* ***понимаю*** *журналистов: корпоративность, цеховая солидарность … Но здесь коленкор иной.*

б. *Рита, конечно, была права — Мике не следовало без приглашения заходить в ее комнату. Но, с другой стороны, Ганин прекрасно* ***понимал*** *сына. Ведь … он торопился вернуться к отцу с книгой…*

В контексте всегда присутствует упоминание некоторого неоднозначного в оценочном отношении поступка или желания, намерения Y-а, а также тех сведений, которые позволили X-у понять (в смысле Т-А), почему или зачем Y так поступил или намерен поступить. Это употребление в [Дмитровская 1985] включается вместе с некоторыми пропозициональными употреблениями, вводящими косвенный вопрос (КВ), в класс «понимание как уяснение». Импликацией уяснения мотивов поступка Y-а является признание X-ом поведения Y-а по меньшей мере осмысленным, объяснимым. Аксиологическая оценка в *понимать*$_{пов}$ не фиксирована и раскрывается только при помощи дополнительных указаний. Наречие *прекрасно,* которое в ЛСВ I и II выполняет только функцию интенсификатора, в данном ЛСВ (в отличие от своего квазисинонима *отлично*) не только маркирует высокую степень уверенность X-а в правильности найденного объяснения поступку Y-а, но и сдвигает отношение к нему в положительную сторону. Наречие *отчасти* (совершенно не характерное для двух рассмотренных выше ЛСВ) сдвигает оценку поступка в отрицательную сторону. Общее отрицание (*X не понимает Y-а*) при данном ЛСВ выражает не только тот факт, что X-у не удалось объяснить для себя поведение Y-а, но и имплицирует слабо отрицательную оценку Y-а со стороны X-а: если X не понимает, почему или зачем Y так себя повел, то обычно X склонен «винить» в этом не свой интеллект, а Y-а, ведущего себя как-то странно[2]. В качестве толкования можно предложить (Т3):

(Т3) ***X понимает$_{пов}$ Y-а*** = ‘X, получив информацию о действиях Y-а и связав ее с известными ему фактами об Y-е, находит объяснение этим действиям’ Импликация: ‘X оценивает поведение Y-а как осмысленное’.

IV. Понимание как одобрение X-ом действий Y-а — *понимать*$_{оцен}$. Примеры:

(10) а. … *вот уже 2 раза я хотела этого в начале отношений, но они не проявляли инициативы, а дальше отношения сходили на нет… Поэтому то, что вы через 3 недели это сделали первый раз: я вас* ***понимаю*** *и считаю это нормальным…*

б. *Я* ***не понимаю*** *людей, которые покупают путёвки и едут «отдыхать». Мне надо именно путешествовать, чтобы была цель, какие-то самостоятельные действия.*

В этом случае в контексте так же, как и в случае *понимать*$_{пов}$, всегда присутствует упоминание некоторого поступка или поведения Y-а. Но если в *понимать*$_{пов}$ наряду с этим присутствовала также ссылка на факты, связанные с самим Y-ом, опираясь на которые X устанавливает мотивы поведения Y-а, то в *понимать*$_{оцен}$ это-

[2] В основе этой импликации лежит фундаментальная оппозиция свой-чужой, ср. намек через дополнительность в [Кобозева, Лауфер 1988].

го нет, да и мотивы поведения Y-а лежат на поверхности (часто это просто желание Y-а). Зато упоминаются действия или желания самого X-а в аналогичной ситуации. При этом понимание X-ом поведения Y-а утверждается, если X обнаруживает сходство поведения Y-а со своим, и отрицается, если не обнаруживает. В таком контексте нет оснований считать, что глагол *понимать* обозначает результат ментальной деятельности X-а по поиску причин или мотивов поведения Y-а. (*Не*)*понимание* в таких случаях сводится к оценке X-ом поведения Y-а: X одобряет или не одобряет поведение Y-а в зависимости от того, как он в подобных ситуациях ведет себя сам. Оценка поведения Y-а имплицирует позитивное или негативное отношение X-а к Y-у (ср. семантический компонент 'Y (не)желателен для X-а по причине Z' в толкованиях глаголов чувства [Иорданская 1968]). *Понимать*$_{оцен}$ обычно имеет форму НСВ, обозначая отношение X-а к Y-у, хотя СВ тоже возможен. Ср. произносимое с угрозой *Не понял(а)!* (= 'мне не нравится то, что ты сделал'). Объект такого понимания может быть как конкретным, так и родовым. В последнем случае частотна конструкция *X (не) понимает Y-ов, которые P*, где P — оцениваемое поведение.

Для этого ЛСВ предлагается следующее толкование:

(Т6) ***X понимает$_{оцен}$ Y-а*** = 'X, зная о поведении Y-а в определенной ситуации и зная, что он сам ведет или повел бы себя так же в подобной ситуации, признает поведение Y-а естественным'

Импликация: 'X считает, что Y в социально-психологическом плане ему ближе, чем те лица, которые в подобных ситуациях ведут себя иначе'

V. Понимание как переживание X-ом сходства внутреннего состояния Y-а с его собственным — *понимать*$_{эмоц}$. Примеры:

(11) а. *Прибираю квартиру, стряпаю. И как **понимаю** жену, когда ей хочется, чтоб заметили и оценили уборку и еду.*
б. *Walzer написал (а): «OFF: Опять конъюнктивит в детском здоровье, рррррр сейчас зарычу. Давала себе зарок не ходить в такие топики ». :-) :-) :-) О-о-о, если бы Вы знали, КАК я Вас **понимаю**!*

В контексте всегда присутствует упоминание некоторого психического состояния Y-а, обусловленного определенной ситуацией, а также отсылка к личному психическому опыту X-а в аналогичной ситуации. X, обнаруживая сходство между психическим состоянием Y-а и своим состоянием в такой же ситуации, уже не столько оценивает состояние Y-а как естественное, сколько эмоционально переживает этот факт. Импликацией этого употребления является ощущение X-ом эмоциональной близости с Y-ом. Только для этого ЛСВ характерно употребление модификатора *как* в «восклицательном» значении, выражающего проявление признака в такой высокой степени, что говорящий выходит из состояния эмоционального равновесия. И данный ЛСВ с модификатором *как*, действительно, как правило, встречается в тех случаях, когда X совпадает с говорящим или с персонажем, которому автор эмпатизирует, высокая степень тут относится или к сходству своих чувств с чувствами другого и / или к силе собственного чувства X-а. Форма глагола всегда НСВ, тип предиката — состояние. Данный ЛСВ сродни «пропозициональному» Пониманию-4 в [Дмитровская 1985], ср. *Петя глубоко чувствовал душевное состояние отца, он понимал, что Василий Петрович как-то особенно мучительно переживает смерть Толстого*. Оба эти состояния *понимания* достигаются практически без помощи рациональной ментальной деятельности, благодаря «вчувствованию». Но пропозициональное *понимать*-4 может обозначать обретение субъектом нового знания о внутреннем мире другого и потому употребляется и в СВ, а наш непропозициональный вариант V имеет внутреннее состояние другого в качестве пресуппозиции и утверждает только переживание субъекта, вызванное сходством его внутреннего состояния с состоянием другого. Толкование этого употребления — Т5:

(Т5) ***X понимает$_{эмоц}$ Y-а*** = 'X, зная о внутреннем состоянии Y-а, вызываемом некоторой ситуацией, сам испытывает или вспоминает, что испытывал то же состоянии в аналогичной ситуации'. Импликация: 'X ощущает душевную близость с Y-ом'

VI. Понимание как социальное отношение X-а к У-у — понимать$_{соц}$

(12) а. *Карталов был единственным человеком, который ее **понимал**. … А если тебя **понимает** такой человек, разве этого мало? Он один ценил ее умение владеть собою… Он любил в ней сдержанность, которая не позволяет изображать страсть пошлыми жестами.*
б. *Я … сижу около смертельно больного папы, кругом такие же еле живые старые люди… — Дорогие мои товарищи, друзья! Я вам очень благодарна за внимание к моему дорогому папе, за то, что уважаете и **понимаете** друг друга*
в. *Он был одержимый, понимаете? У одержимых не бывает личной жизни. — Иногда бывает, — не согласился Крячко. — Когда жена **понимает** мужа…*
г. *— Я люблю воров. Они смелые. — А я люблю тебя, что ты образованный и понимаешь нашего брата* (т. е. воров. — *И. К.*).

Этот семантический вариант реализуется в контексте, который характеризуется отрицательно — от-

сутствием всех тех «экспонентов», которые позволяли бы усмотреть в нем какой-либо другой вариант: не упоминаются ни условия, обеспечивающие накопление X-ом фактов для познания Y-а (*понимать*$_{когн}$), ни речевые действия Y-а, которые X мог бы интерпретировать (*понимать*$_{ком}$), ни поступки Y-а, которые X мог бы объяснять для себя или оценивать (*понимать*$_{пов}$ и *понимать*$_{оцен}$), ни чувства Y-а, которые X мог бы сравнить со своими (*понимать*$_{псих}$). Могут упоминаться свойства Y-а и эмоционально-оценочное отношение X-а к этим свойствам и их носителю — *любить*, *ценить* (см. примеры 12а,г). Часто в этом контексте *понимать* оказывается в одном ряду с глаголами из семантической группы чувств-отношений (*любить*, *уважать*). Субъект и объект такого понимания обычно связаны друг с другом теми или иными социальными отношениями — семейными (12в), профессиональными (12а), приятельскими (12г), соседскими (12б) и т. п. Во всяком случае, в отличие от всех рассмотренных выше ЛСВ, это не могут быть два незнакомых друг с другом человека. Характерно, что такое понимание рассматривается как благо, ресурс для объекта: им гордятся (20а), за него благодарят (20в), им счастливы (см. эпиграф). И еще одна характерная деталь: чаще этот вид *понимания* утверждается не его субъектом (*Я вас / его понимаю*), а его объектом (*Ты меня понимаешь*; *Он меня понимает*), что говорит о том, что именно объекту «виднее», *понимают* его (в этом смысле) или нет. Соответственно, в предложениях типа *Жена понимала мужа, но он считал, что она судит о нем совершенно неверно* реализуется не вариант VI, а вариант I — *понимать*$_{когн}$: жена знала существенные свойства мужа, а он сам имел о них превратное представление. Все это приводит нас к следующей экспликации данного варианта:

> (Т6) ***X понимает***$_{соц}$ ***У-а*** = X, имеющий регулярные контакты с с Y-ом, знает благодаря этому свойства Y-а, мотивы его поведения и чувства; оценки X-ом этих свойств, мотивов и чувств совпадают с оценками Y-а; при этом X хорошо относится к Y-у из-за его положительных свойств и мирится с его отрицательными свойствами.

Импликация: ‘X-у свойственно вести себя по отношению к Y-у желательным для Y-а образом’В (21) ментальная, «рассудочная» составляющая (знание об Y-е) составляет пресуппозицию, а утверждается совпадение этих знаний с представлениями Y-а и позитивное отношение к Y-у. Поскольку все ассертивные составляющие значения содержат предикаты свойств или отношений (совпадения, оценочного отношения, стереотипа поведения), то и глагол обозначает нелокализованное во времени положение дел и употребляется в этом значении только в НСВ.

Итак, мы видим, что в зависимости от контекста глагол *понимать*, прямым объектом которого является имя лица, может выражать целую гамму интенциональных состояний человека, направленных на другого человека, начиная с чисто «рассудочных» (см. ЛСВ I и II), переходя ко все более тесно связанным с эмоциональоно-оценочной сферой и кончая преимущественно оценочным отношением к личности другого, влекущим за собой определенное поведение по отношению к этой личности.

Переход от «рассудочного» к «социальному» *пониманию* имеет под собой социально-психологическую базу. В конфликтологии установлено: когда мнения и убеждения X-а оказываются несовместимыми с чьими-либо другими, X склонен ощущать, что его самого ставят под сомнение, что к нему проявляют враждебное отношение. В связи с этим было введено понятие **социо-когнитивного конфликта** [Light, Perret-Clermont 1994], которое в [Greco Morasso 2008] фигурирует как когнитивная база развития двух значений у английского глагола *conflict* — значения пропозициональной несовместимости и значения межличностной враждебности. Естественно предположить, что в противоположном случае, то есть когда мнения и убеждения X-а, а тем более его чувства совпадают с чьими-либо другими, X склонен ощущать, что к нему проявляют благорасположение. Можно назвать это эффектом **социо-когнитивной гармонии**. На основе этого и эффекта и возникает возможность метонимического сдвига пропозиционального и коммуникативного *понимать* через ряд ступеней к *понимать* межличностному.

Литература

1. *Апресян Ю. Д.* Проблема фактивности знать и его синонимы // Вопросы языкознания 1995, № 4
2. *Богуславский И. М.* Словарная статья глагола понимать // И. А. Мельчук, А. К. Жолковский. Толково-комбинаторный словарь современного русского языка. Опыты семантико-синтаксического описания русской лексики. Вена, 1984
3. *Дмитровская М. А.* Механизмы понимания и употребление глагола понимать // Вопросы языкознания 1985. № 3.
4. *Иорданская Л. Н.* Попытка лексикографического толкования группы русских слов со значением чувства // Машинный перевод и прикладная лингвистика. М., 1970. Вып. 13.
5. *Кобозева И. М., Лауфер Н. И.* Известия АН СССР. Об одном способе косвенного информирования // Серия литературы и языка. 1988. № 5.
6. *Greco Morasso S.* The ontology of conflict // Pragmatics and cognition, 2008, Vol.16, №3.
7. *Light P., Perret-Clermont A. N.* Social context effects in learning and testing // A. Gellatly, D. Rodgers and J. A. Sloboda (eds) Cognition and Social Worlds. Oxford, 1989.

База данных «интонация русских повествовательных текстов»[1]

The database on intonation of russian narrative texts

Кодзасов С. В. (sankod@philol.msu.ru),
Архипов А. В. (arxipov@philol.msu.ru),
Захаров Л. М. (leon@philol.msu.ru),
Кривнова О. Ф. (okri@philol.msu.ru)

Московский государственный университет им. М. В. Ломоносова, Москва, Россия

В докладе сообщается о результатах 2-го этапа работы по созданию БД «Интонация русских информационных и повествовательных текстов». Этот этап открывал 2-й трехлетний цикл исследований русской интонации.

Напомним, что **1-й трехлетний цикл** работ нашей группы (2004–2006 гг.) был связан с интонацией **диалогического** дискурса (интонация вопросов, побуждений, сообщений и реакций на них). Результаты исследований авторского коллектива в этой области сообщались на прошлых конференциях «Диалог» (см. список публикаций).

2-й трехлетний цикл посвящен интонации **недиалогического** дискурса. Он также состоит из трех этапов:

2007 г. — интонация информационных текстов (на материале телевизионных новостей);

2008 г. — интонация литературных повествовательных текстов (на материале записей современной прозы: Булгаков, Пелевин, Маринина), а также интонация фольклорных повествовательных текстов в исполнении профессиональных чтецов;

2009 г. — интонация спонтанных нарративов (рассказы о случаях из жизни).

Проект связан с быстро развивающейся областью отечественной лингвистики, которую можно назвать дискурсивной грамматикой русского языка. В ее задачи входит описание дискурсивной лексики, дискурсивного синтаксиса и дискурсивной интонации.

Как было сказано выше, до сих пор нашей группой изучалась в основном интонация диалогического дискурса. В осуществляемом ныне проекте ставится задача создания компьютерной базы для изучения интонационной специфики информационного и повествовательного дискурса. При этом просодия текстов исследуется в тесной связи с их иллокутивным и модальным содержанием, а также с их фазовой структурой. Особое внимание уделяется просодическим техникам маркировки жанровых и регистровых характеристик текстов и их компонентов. Это позволит уточнить системы семантических, прагматических и стилистических дескрипторов, которые используются в отечественной текст-лингвистике в настоящее время.

Результаты проекта могут использоваться как при дальнейших лингвистических исследованиях грамматики текста, так и в прагматических целях (оптимизация учебных и инструктивных текстов, рекомендации по использованию разных произносительных техник). В число потенциальных пользователей таким образом попадают не только лингвисты, но также дикторы и составители текстов в средствах массовой информации.

Проект предполагает создание трех представительных корпусов звучащих текстов для трех базовых недиалогических областей дискурса: информационные тексты (2007 г.), нарративные тексты разных жанров (современные литературные тексты, фольклорные тексты) (2008 г.), спонтанные нарративы (2009 г.). Характерные фрагменты для каждого типа дискурса вводятся в компьютерную базу данных. Каждый файл снабжается просодической транскрипцией, включающей все просодические характеристики: тональные акценты и интегральные уровни тона, выделительные акценты и громкости, долготы и темпы, фонации и просодические тембры. План работ по годам соответствует трем указанным областям дискурса.

Ниже приводятся фрагменты базы данных для фольклорных и нарративных текстов, прочитанных профессиональными чтецами.

[1] Работа поддержана грантом РГНФ № 07-04-12160в.

1. Сказка. *Кот, козёл да баран*

Код	Орфография	Просодия	Комментарии
КОТ.0.			
КОТ.0.1.	Кот, козел да баран.		
КОТ.1.			
КОТ.1.1.	Жили-были в одном дворе кот, козел да баран; жили они дружно: сена клок и тот пополам.	*Жили(\\)-*были(//) в *одном(\/) *дворе(/) | *кот(-/), *козел(-/) да(/-) *баран(-\); || *жили(/-) они *дружно(-\\): || *сена(\\) *клок(\) | и *тот(\) пополам(/~).||	Пословная акцентуация интродуктивного предложения — типичная фигура при чтении сказок.
КОТ.1.2.	А коли вилы в бок — так одному коту Ваське!	А коли *вилы(/) в бок | — так одному коту *Ваське(\')! ||	
КОТ.1.3.	Он такой вор и разбойник, каждый час на промысле, и где что плохо лежит, туда и глядит.	*Он *такой(/) *вор(/)(:) и *разбойник(\)(:), || *каждый(/) *час(\) на промысле, || и где что плохо *лежит(//), || *туда(\н) и *глядит(\в).||	Экспрессивное удлинение гласных: *вор* (:), *разбойник* (:)
КОТ.2.			
КОТ.2.1.	Вот однажды лежат себе козел и баран и *разговаривают промеж собой.	*Вот(//) *однажды(//) || *лежат(//-) себе козел и *баран(//-) *и *разговаривают(/)(\) промеж *собой(\\). ||	Лексическая безакцентность слова *козел.* Слово *коза* в этом контексте скорее имело бы выделительный акцент.
КОТ.2.2.	Откуда ни возьмись — котишко-мурлышко, серый лобишко, идет да таково жалостно плачет.	*Откуда ни *возьмись — || *котишко(/)-*мурлышко(\), ||*серый(\) *лобишко(\),|| *идет(/\) || да *таково(/\) *жалостно(:) *плачет(:). ||	Иконические долготы в словах *жалостно плачет.*
КОТ.3.			
КОТ.3.1.	Козел да баран спрашивают:	*Козел(/) да *баран(/) *спрашивают(\): ||	
КОТ.3.2.	Кот-коток, серенький лобок, о чем плачешь, на трех ногах скачешь?	*Кот-коток(\ — \), *серенький(\) *лобок(/)(\), о *чем(/) *плачешь(-\), на *трех(/) *ногах(\) *скачешь(\)?	

2. Литературный текст в жанре детектива (А. Маринина)

Код	Орфография	Просодия	Комментарии
МАР1.1.			
МАР1.1.1.	Субботнее утро Настя Каменская провела за своим любимым занятием.	*Суббо́тнее(/) *у́тро(/') | *На́стя(/') *Ка́менская(/) *провела́(/) | за *свои́м(/) *люби́мым(/) *заня́тием(/'). ||	Пословная акцентуация, характерная для экспозиции повествования.
МАР1.1.2.	Она ленилась.	Она *ле́нилась(/). ||	Восходящий тон на сонорном. Направление тона трудно объяснить иллокутивной семантикой.
МАР1.1.3.	Еще вчера вечером на вопрос мужа: «Чем собираешься завтра заниматься?» — она честно ответила: «Буду лениться».	*Е́ще(//)(\) вчера *ве́чером(/) на вопрос *му́жа(\): «*Че́м(\) *соби́раешься(\) *за́втра(\) *занима́ться(/)?» | — она *че́стно(/\) *отве́тила(/): | «*Бу́ду(//) *ле́ни́ться(\)(\)». ||	Специфическая тональная огласовка слова *еще* нуждается в интерпретации.

Код	Орфография	Просодия	Комментарии
MAP1.2.			
MAP1.2.1.	И вот теперь она валялась в постели, прихлебывая крепкий горячий кофе, слушала музыку и предавалась неспешным размышлениям.	И *в<u>о</u>т(/\) *т<u>е</u>п<u>е</u>рь(\)(\) она *ва-л<u>я</u>лась(\H) в *пост<u>е</u>ли(/), | прихлебывая *кр<u>е</u>пкий(/') горячий *к<u>о</u>фе, | *слу-шала(/\) *м<u>у</u>зыку(//) | и *предава<u>а</u>лась *неспе<u>ш</u>ным(/) *размышл<u>е</u>ниям(\). ||	Непонятно отсутствие тонального акцента на слове *кофе.* Требует истолкования восходяще-нисходящий тон на слове *слушала.*
MAP1.2.2.	Правда, надо отдать ей должное — размышления были все-таки связаны с работой.	[*Пр<u>а</u>вда(/), *н<u>а</u>до(\) *отд<u>а</u>ть(\) ей *д<u>о</u>лжно<u>е</u>(\)(\) | — *размышле-ния(\) *б<u>ы</u>ли(\) все-таки *св<u>я</u>заны(\) с *раб<u>о</u>той(\)](H). ||	Низкий регистр тона всего фрагмента.
MAP1.2.3.	Во-первых, она думала об исчезновении вещественных доказательств по делу об убийстве пятнадцатилетнего подростка.	*Во-первых(\), | она *думала(\) об ис-ч<u>е</u>зновении(\) вещественных *до-каз<u>а</u>тельств(\) по *д<u>е</u>лу(\) об *убий-стве(\`) *пятн<u>а</u>дцатил<u>е</u>тнего(\`) *под-р<u>о</u>стка(/\) ||	Цепочка нисходящих акцентов, в том числе и в позициях незавершенности высказывания.
MAP1.2.4.	Этим убийством их отдел занимался вот уже четыре месяца.	Этим *уб<u>и</u>йством(/) | их *отд<u>е</u>л(/) *занима<u>л</u>ся(/) | *в<u>о</u>т(\) *уже(\) *четы-ре(\) *месяца(\). ||	Цепочка восходящих тонов сменяется цепочкой нисходящих. Можно проинтерпретировать как различие тематического и рематического компонента высказывания.
MAP1.2.5.	Во-вторых, она думала о свалившемся на них два дня назад убийстве пятерых человек — целой семьи известного московского художникапортретиста.	*В<u>о</u>-вторых(/-), | она *думала(/) о *свалившемся(/) на них *два *дня *назад *уб<u>и</u>йстве(\) *п<u>я</u>тер<u>ы</u>х(\)(\) *ч<u>е</u>лов<u>е</u>к(\)(/) | — *ц<u>е</u>лой(\) *с<u>е</u>-мь<u>и</u>(\)(/) *изв<u>е</u>стного(/) *моск<u>о</u>вско-го(\) *художника -*портрет<u>и</u>ста(\). ||	Мотивация направления тона для большинства слов неочевидна. Простую интерпретацию допускают лишь начальный восходящий и конечный нисходящий акценты.
MAP1.2.6.	В-третьих, Анастасия Каменская с раздражением думала о том, что нужно получать новый комплект форменной одежды, а для этого нужно найти старые ордера, по которым она так и не получила форму в прошлый раз.	*В-тр<u>е</u>ть<u>и</u>х(\)(/), | *Анастас<u>и</u>я(\) *Кам<u>е</u>нская(/) с *р<u>а</u>здраж<u>е</u>нием(\) (/) думала о *т<u>о</u>м(/), | что *н<u>у</u>жно(/) *получ<u>а</u>ть(\) *н<u>о</u>вый(/) *компл<u>е</u>кт(\) | форменной *од<u>е</u>жды(/)(/) |, а для *<u>э</u>того(//) *н<u>у</u>жно(/) найти старые *ордер<u>а</u>(//), || по *котор<u>ы</u>м(\) она *так(/) и *н<u>е</u>(\) *п<u>о</u>луч<u>и</u>ла(\ \ \) *ф<u>о</u>р-му(/-) в *пр<u>о</u>шлый(/-) *р<u>а</u>з(-\). ||	Также и здесь почти все слова внутри фрагмента могут быть огласованы разными тональными акцентами. Кроме того, тональные акценты для большинства слов необязательны.
MAP1.2.7.	Куда она эти ордера засунула, Настя вспомнить не могла, стало быть, придется сочинять покаянный рапорт об их утере.	*Куд<u>а</u>(//-) она *<u>э</u>ти(\) ордера засунула, | *Н<u>а</u>стя (/-) вспомнить не *могла<u>(</u>-\), || стало быть, *приде<u>т</u>ся (\) сочинять покаянный *р<u>а</u>порт(/) об их *ут<u>е</u>ре. (\) ||	Также и этот фрагмент допускает разные прочтения. Например, возможен восходящий акцент на слове *засунула.* Также, возможен выделительный акцент на слове *сочинять.*

Как и результаты предшествующего этапа работы, полученная информация представлена в виде базы данных в формате MICROSOFT ACCESS. Программа позволяет прослушивать файл, видеть его осциллограмму и интонограмму, а также анализировать его просодическую дескрипцию. Какие интонационные характеристики фиксирует эта дескрипция? Прежде всего, это акцентная структура текстов (размещение семантически нагруженных выделительных акцентов — они обозначены звездочками). Далее, это тональная структура компонентов сообщения (тоны обозначены слешами) и средства их экспрессивного подчеркивания. Производился также упрощенный анализ фазовых и тематических компонентов текста, что облегчает пользователю анализ функций просодических средств его оформления.

Мы приводим в Приложении расшифровку знаков используемой в данной работе просодической транскрипции, поскольку она необщепринята и может вызвать затруднения у потенциального читателя. Эта же система использовалась нами при создании баз данных по интонации русских диалогических реплик (см. список публикаций).

Приложение

Интонационная транскрипция: знаки для тональных акцентов

\ — нисходящий нейтральный (4–7 полутонов)
\’ — нисходящий малый (3–4 полутона)
\” — нисходящий сверх-малый (менее 3 полутонов)
\\ — нисходящий увеличенного интервала
\в — нисходящий в высоком регистре
\н — нисходящий в низком регистре
\с — нисходящее движение происходит на начальном согласном слога
\~ — нисходящий тон растянут на слово
/ — восходящий нейтральный
/’ — восходящий малый
/” — восходящий сверх-малый
// — восходящий увелич. интервала
/в — восходящий в высоком регистре
/н — восходящий в низком регистре
/с — восходящее движение происходит на начальном согласном слога
/~ — восходящий тон растянут на слово
/-\ — восходящий — ровный — нисходящий
/\ — восходяще-нисходящий тон внутри гласного
/с\г — восходящий тон на согласном и нисходящий на гласном
\/ — нисходяще-восходящий тон внутри гласного
/- — восходящий плюс ровный без падения

Нетональные акценты

: — долготный акцент на гласном
* — громкостный акцент на подчеркнутом гласном слова

Знаки для синтагменных просодий

инк — инклинация тона
дек — деклинация тона
ТИ — низкая громкость (тихо)
ГР — высокая громкость (громко)
Н — низкий регистр
В — высокий регистр
Б — быстрый темп
НПР — напряженная фонация
ПДХ — придыхательная фонация
ФЦТ — фальцетный регистр

Литература

1. *Кодзасов С. В., Бонч-Осмоловская А. А., Захаров Л. М., Кобозева И. М., Кривнова О. Ф.* База данных «Интонация русского диалога»: вопросительные реплики // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2005». М., 2005. С. 245–249.
2. *Кодзасов С. В., Архипов А. В., Бонч-Осмоловская А. А., Захаров Л. М., Кривнова О. Ф.* База данных «Интонация русского диалога»: побудительные реплики. // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2006». М., 2006. С. 236–242.
3. *Кодзасов С. В., Архипов А. В., Захаров Л .М., Кривнова О. Ф.* База данных «Интонация русского диалога»: реплики-сообщения // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007». М., 2007. С. 269–277.
4. *Кодзасов С. В., Архипов А. В., Захаров Л. М., Кривнова О. Ф.* База данных «Интонация русских информационных текстов» // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог ’2008». М., 2008. С. 206–209.

Выявление номинализованных конструкций в параллельных текстах патентных документов на русском и немецком языках

Detection of nominalized structures in parallel patent texts in Russian and in German

Кожунова О. С. (okozhunova@ipiran.ru)

Институт проблем информатики Российской академии наук

Исследуется явление номинализации в двуязычной ситуации (русский-немецкий) с привлечением результатов сопоставительных исследований для трех языков: русский — английский — немецкий, а также способ идентификации параллельности текстов для патентной сферы и типы трансформаций.

1. Введение

На современном этапе создания систем обработки естественного языка для информационных технологий особое значение приобретает разработка методик анализа параллельных текстов на нескольких языках. При этом возникает множество задач, связанных с их адекватной интерпретацией и применением, в первую очередь, задачи машинного перевода и обработки знаний [1]. Проблема извлечения и обработки знаний открывает перспективы развития интеллектуальных направлений компьютерной лингвистики, поскольку ее основной акцент смещен в сторону глубинных представлений языка, в которых используются как грамматические (морфологические и синтаксические), так и семантические атрибуты для описания языковых объектов. В данной работе исследования параллельных текстов рассматриваются именно в ракурсе этой проблемы.

Как показано в [2] ключевой задачей при разработке методов сопоставления параллельных текстов является выявление и детальное описание тех языковых трансформаций, которые имеют место при переводе естественно-языковых конструкций с одного языка на другой, поскольку далеко не всегда некоторое содержание передается структурно-подобными средствами в текстах на разных языках. При этом особое внимание должно уделяться выявлению признаков «глагольности» и «номинативности» [2] в сопоставляемых языковых структурах параллельных текстов, поскольку это важно для последующей разработки алгоритмов анализа текстов с использованием методов машинного обучения. Сравнительное исследование употребления различных частей речи в параллельных текстах на разных языках дает основу для выявления и описания языковых трансформаций, при этом центральной трансформацией является *номинализация*. Явление номинализации было исследовано в ряде работ отечественных и зарубежных лингвистов [3, 4, 5, 6]. Ближе всего к нашему пониманию этого явления определения, данные в работе [4]: «конструкции… называются номинализованными в том смысле, что их естественно рассматривать как результат номинализации конструкций с предикативным употреблением глаголов и прилагательных» и [5]: «номинализация — это синтаксический процесс, который соотносит предложения с именными группами». Основная цель наших исследований выявление номинализованных конструкций в патентных текстах на русском и немецком языках, прежде всего, в формулах изобретений и сопоставительное описание глагольно-именных межъязыковых трансформаций. Центральная задача наших исследований — выявление тех ситуаций, когда фраза с личным глаголом в тексте на одном языке превращается в именную фразу в тексте на другом языке и обратно.

2. Определение эквивалентности языковых структур при сопоставлении патентных текстов и выявление номинализации

Тексты патентных изобретений, по существу, представляют собой особый жанр, поскольку к ним предъявляются строго регламентированные требования как на содержательном уровне (тематика, сте-

пень новизны и актуальности текста определенной предметной области, иллюстративность, практическое значение и т.п.), так и в вопросе оформления, к соблюдению определенной структуры, прозрачности и четкости изложения и т. д. Так, например, необходимым условием патентного описания является четкое перечисление характеристик изобретения по нумерованным пунктам. Казалось бы, это условие должно облегчать их анализ при сопоставлении с аналогичным описанием на другом языке. Однако жесткость структуры патентных формул влечет за собой осложнение процедуры их сравнения, так как нумерация пунктов часто не совпадает и условие параллельности текстов не соблюдается.

Поэтому при анализе формул изобретения патентных текстов на русском, немецком и английском языках на основании отобранного массива текстов патентов 2006 года (455 текстовых представлений) была выявлена необходимость в разработке подхода к определению параллельности таких текстовых описаний и предварительной типологии анализируемых текстов.

Для первичного отбора параллельных текстов на трех языках был использован инструмент аккумуляции и индексации патентов MIMOSA V5, который представляет собой базу данных кратких патентных описаний. Внутри каждого описания размещены ссылки на внешние ресурсы, содержащие формулы изобретения и соответствующие тексты патентов. С целью проведения синтаксического разбора патентных формул изобретения на русском и английском языках применялся программный пакет Dwarf 2.0. Этот синтаксический анализатор позволяет проводить грамматический разбор текстов и строить деревья разметки с выделением соответствующих атрибутов. Далее в тексте приведены примеры, в которых используются результаты работы Dwarf 2.0 (пример 1).

(1) ***Формула изобретения на немецком языке**:*

*1. *Vorrichtung zum Sterilisieren von Getränkebehältern durch Strömungsbeaufschlagung mit einem Gemisch aus Luft und H202-Dampf, mit einem H202-Verdampfer, der einen luftdurchströmten Verdampfungsraum (4) mit beheizten Wänden aufweist*, gegen dieH202 von einer Düseneinrichtung (7,9) gesprüht wird, dadurch gekennzeichnet, dass die Düseneinrichtung eine einen engen Luftstrahl (8)einblasende Luftdüse (7) und eine H202Düse (9) aufweist, die zur Erzeugung eines H202-Strahles (10) im Abstand zum Luftstrahl und auf diesen gerichtet ausgebildet ist.*

***Фрагмент синтаксического разбора[1]*:**

1	*Top*	*Top*	1	_	/
2	vorrichtung	vorrichtung	1	vp	/nn_verb, sg, nom
3	zum	zum	4	prep	/prep_dativ
4	sterilisieren	sterilisieren	6	vp	/nn_verb, sg, dat
5	von	von	6	prep	/prep_dativ
6	getraenkebehaeltern	getraenkebehaelter	8	np	/nn, pl, dat
7	durch	durch	8	prep	/prep_accus
8	Stromungsmungsbeauf-schlagung	Stromungsbeauf-schlagung	11	np	/nn_verb, sg, acc
9	mit	mit	10	prep	/prep_dativ
10	einem	einem	11	det	/det, indef, sg, dat
11	gemisch	gemisch	13	np	/nn, sg, acc
12	aus	aus	13	prep	/prep_dativ
13	luft	luft	11	np	/nn/sg/trd/heu/
14	und	und	15	cnj	/ cnj
15	h202-dampf	h202-dampf	18	np	/nn_comp, sg, acc
16	mit	mit	17	prep	/ prep_dativ
17	einem	einem	18	det	/det, indef, sg, dat
18	h202-verdampfer	h202-verdampfer	21	np	/nn_comp, sg, acc
19	der	der	20	det	/det, def, sg, acc
20	einen	einen	21	det	/det, indef, sg, acc
21	luftdurchstroemten	luftdurchstromte	22	adj	/ adj, sg, acc
22	verdampfungsraum	verdampfungsraum	21	np	/nn, sg, acc
23	mit	mit	24	prep	/ prep_dativ, управл датив
24	beheizten	beheizt	25	adj	/ adj, pl, dativ
25	waenden	wand	24	np	/nn/sg/trd/heu/
26	aufweist	aufweisen	11	vp	/vb, sg, 3p

[1] Разбор выполнен вручную с использованием грамматических атрибутов и средств разметки Dwarf 2.0: http://cs.isa.ru:10000/troll/

В ходе эксперимента по анализу немецких, русских и английских патентных текстов было выявлено, что формулы изобретения существенным образом трансформируются структурно, что также отражается на порядке описания содержательной компоненты.

Например, отдельные фрагменты формулы изобретения на одном из рассматриваемых языков расширяются, включают в себя более детальные описания, или, напротив, сужаются, часто происходит удаление части контента или его дополнение:

В параллельных текстах примера 1 можно проследить следующие русско-немецкие лексико-семантические и синтаксические трансформации:

(a) Getränkebehältern -> емкостей для напитков
N [dat, masc, pl] -> N [acc,fem, pl] + Prep+ N [gen, masc, pl]

(b) durch Strömungsbeaufschlagung -> путем воздействия
Prep + N [acc, fem, sg] -> Prep + N [gen, neutr, sg]

(c) luftdurchströmten -> омываемую воздухом
Comp_part [acc, masc, sg] -> Part [acc, fem, sg] + N [instr, masc, sg]

(d) Verdampfungsraum -> испарительную камеру
N [acc, masc, sg] -> Adj [acc, fem, sg] + N [acc,fem, sg]

Формула изобретения на русском языке:

*1. *Устройство для стерилизации емкостей для напитков путем воздействия на них потоком смеси воздуха и пара перекиси водорода, содержащее испаритель H2O2, имеющий омываемую воздухом испарительную камеру (4) с обогреваемыми стенками,* которые с помощью соплового устройства (7, 9) обрызгиваются перекисью водорода, отличающееся тем, что сопловое устройство содержит только одно воздушное сопло (7), выполненное с возможностью вдувания тангенциально в камеру узкой воздушной струи (8), и только одно сопло (9) для перекиси водорода, выполненное с возможностью формирования сплошной струи жидкости, имеющей диаметр примерно в пределах равных нескольким десятым миллиметра, причем сопло для перекиси водорода установлено в корпусном блоке (3) с возможностью формирования струи жидкости (10) в поперечном к воздушной струе (8) направлении и на расстоянии от нее.*

***Фрагмент синтаксического разбора[2]*:**

1	*Top*	*Top*	1	_	/
2	устройство	устройство	1	np	/nn/sg/neu/nom/
3	для	для	4	prep	/prp/
4	стерилизации	стерилизация	2	prepnp	/nn/sg/fem/gen/
5	емкостей	емкость	4	gen	/nn/pl/fem/gen/
6	для	для	7	prep	/prp/
7	напитков	напиток	5	prepnp	/nn/pl/msc/gen/
8	путем	путь	9	prep	/nn/sg/msc/ins/
9	воздействия	воздействие	7	prepnp	/nn/sg/neu/gen/
10	на	на	11	prep	/prp/
11	них	они	9	prepnp	/prn/pl/acc/trd/
12	потоком	поток	1	np	/nn/sg/msc/ins/
13	смеси	смесь	12	gen	/nn/sg/fem/gen/
14	воздуха	воздух	13	gen	/nn/sg/msc/gen/
15	и	и	14	conj	/cnj/
16	пара	пар	14	homo	/nn/sg/msc/gen/
17	перекиси	перекись	16	gen	/nn/sg/fem/gen/
18	водорода	водород	17	gen	/nn/sg/msc/gen/
20	содержащее	содержать	1	adj	/ptp/sg/neu/nom/prs/act/
21	испаритель	испаритель	20	acc	/nn/sg/msc/nom/
22	н2о2	н2о2	1	misc	/
25	имеющий	иметь	23	adj	/ptp/sg/msc/nom/prs/act/
26	омываемую	омывать	29	ptp	/ptp/sg/fem/acc/prs/psv/
27	воздухом	воздух	26	ins	/nn/sg/msc/ins/
28	испарительную	испарительный	29	adj	/adj/sg/fem/acc/
29	камеру	камера	25	acc	/nn/sg/fem/acc/
33	с	с	35	prep	/prp/
34	обогреваемыми	обогревать	35	adj	/ptp/pl/ins/prs/psv/
35	стенками	стенка	23	prepnp	/nn/pl/fem/ins/

[2] Разбор выполнен при помощи инструмента для синтаксического разбора Dwarf 2.0: http://cs.isa.ru:10000/troll/

Трансформации типов *(a)–(d)* назовем *предложно-номинативными*. Они встречаются в патентных описаниях формул изобретений среди русско-немецких двоек параллельных текстов с доминирующей частотой (~60 %). Таким образом, преобразование составных отглагольных существительных (немецкий язык) во фразовые номинативные структуры (русский язык) в формулах изобретения патентных параллельных текстов является наиболее распространенным. При чем подобные трансформации происходят не только на уровне семантики различных языковых структур, но и на лексическом уровне (как это видно из структурных формул приведенных выше), поскольку, как правило, одна языковая единица немецкого языка преобразуется в несколько языковых единиц русского языка.

(e) mit beheizten Wänden -> с обогреваемыми стенками
Prep + Part [dat, fem, pl] + N [dat, fem, pl] -> Prep + Part [instr, fem, pl] + N [[instr, fem, pl]

Менее частотные трансформации типа *(e)* назовем *адъективно-номинативными*. Причем встречаемость таких трансформаций, как показал анализ экспериментального массива формул изобретений, как правило, зависит от исходной предметной области описываемых изобретений. Например, в описаниях патентов в области химии и медицины доминируют именно *адъективно-номинативные* трансформации:

die lösliche Guanylatcyclase -> растворимой гуанилатциклазы
verzweigtes Halogenalkoxy -> разветвленный галогеналкил

(f) der einen [a] aufweist (который имеет) -> имеющий [a]
Pro [rel, masc, sg] + Art [indef, masc, acc] + [a] + Vb [3 ps, sg, pres] _-> Part [nom, masc, sg] + [a]

Что касается трансформаций типа *(f)* (назовем их *структурно-вербальными*), то они носят другой функционально-семантический характер. То, что было относительным местоимением и личной формой глагола в немецком языке становится причастием в русском языке. Таким образом, в структурно-вербальных трансформациях осуществляются не только лексико-семантические преобразования, но и синтаксические.

(2) ***Формула изобретения на немецком языке:***

2. Vorrichtung nach Anspruch 1, dadurch gekennzeichnet, dass der Verdamp fungsraum (4) mit dem Auslass(1) des Verdampfers über beheizte Kanäle (5) verbunden ist.

Формула изобретения на русском языке:

3. Устройство по п.1, отличающееся тем, что испарительная камера (4) соединена с выходным отверстием (1) испарителя обогреваемыми каналами (5).

В примере 2 также присутствуют выделенные в примере 1 трансформации. Тем не менее, остановимся на наиболее интересном преобразовании:

[a] соединена [b] -> [a] [b] verbunden ist

[a] + *Part [short, nom, fem, sg]* + [b] -> [a] + [b] *+ Part [II, sg, masc] + Vb [3 ps, sg, pres]*

Аналогичная трансформация была рассмотрена в примере 1, случай (f). Здесь также присутствует не только лексико-семантическое преобразование, но и синтаксическое: краткое причастие на русском языке трансформируется в составную глагольную структуру на немецком языке.

Большое количество формул изобретения аналогичных приведенным примерам позволяет сделать вывод о том, что их сопоставительный анализ следует проводить на основе выделения содержательных объектов. То есть при первичном отборе параллельных текстов и определении их параллельности необходимо опираться не на синтаксическую структуру, а на выделенные семантические объекты [7, 8]. Поэтому на стадии предварительного анализа текстов мы подразделяем их на собственно параллельные, непараллельные и *концептуально--сопоставимые*[3]. Под *параллельными текстами* в данном случае понимаются тексты, которые являются эквивалентными и адекватными переводам с одного языка на другой, то есть совпадающими по объему фактической информации в них представленной. Тексты, в которых описывается один и тот же контент, но которые не являются эквивалентными переводами и не совпадают по объему, мы называем *концептуально-сопоставимыми текстами*.

В зависимости от того, с какими текстами мы имеем дело, варьируется и методика их анализа. В частности, при сопоставлении параллельных текстов наиболее интересен последовательный контрастивный анализ предложений с целью обнаружения трансформаций, которые, как правило, регулярные и предсказуемые:

(3)

- ***Формула изобретения на немецком языке:*** Verfahren zur Epoxidierung einer organischen Verbindung mit wenigstens einer C C-Doppelbindung mit Wasserstoffperoxid in Gegenwart wenigstens einer katalytisch aktiven Verbindung und wenigstens eines Lösungsmittels, dadurch gekennzeich-

[3] Термин *концептуально-соотносимые* введен в данной работе впервые и является результатом наших исследований структурных особенностей параллельных текстов формул изобретения в патентных документах.

net, dass ein Produktgemisch umfassend a-Hydroperoxyalkohole unter Einsatz wenigstens eines Reduktionsmittels reduziert wird.

- ***Формула изобретения на русском языке:*** *Способ эпоксидирования пропена путем взаимодействия с перекисью водорода в присутствии, по меньшей мере, одного каталитически активного соединения, включающего цеолитный катализатор, и, по меньшей мере, одного органического растворителя с использованием разделения продуктов, отличающийся тем, что смесь продукта, содержащую а -гидропероксипропанолы, восстанавливают с применением, по меньшей мере, одного восстановителя, причем а -гидропероксипропанолы восстанавливают до соответствующего пропиленгликоля.*
- ***Формула изобретения на английском языке:*** *A process for the epoxidation of an organic compound having at least one C-C double bond by means of hydrogen peroxide in the presence of at least one catalytically active compound and at least one solvent, wherein a product mixture comprising [alpha]-hydroperoxyalcohols is reduced using at least one reducing agent.*

(a) Verfahren zur Epoxidierung –> Способ эпоксидирования –> A process for the epoxidation
N [verb, nom, neutr, sg] + Prep [zu+der, dat, comp, fem, sg] + N [dat, fem, sg] –> N [nom, masc, sg] + N [gen, neutr, sg] –> Art [indef, sg] + N [com, sg] + Prep + Art [def, 0] + N [com,sg]

(b) ein Produktgemisch –> смесь продукта –> a product mixture
Art [indef, masc, nom, sg] + N [comp, nom, neutr, sg] –> N [nom, fem, sg] + N [gen, masc, sg] –> Art [indef, sg] + N [com, sg] + N [com,sg]

(c) dadurch gekennzeichnet –> отличающийся тем –> wherein
Pron + Part [II f, masc, sg] –> Part [nom, masc, sg] + Pron [instr, masc, sg] -> Adv

В примере 3 приведены трансформации, которые обнаружены в тройке параллельных текстов. Для троек параллельных текстов патентных формул изобретений наиболее частотными типами трансформаций оказались преобразования (a) и (b), то есть *предложно-номинативные* и *адъективно-номинативные* (порядка 55 %), реже встречаются трансформации типа (c), которые мы назовем *предложно-адъективными* (около 30 %). Как уже было отмечено в комментариях к примерам 1 и 2, *предложно-номинативные* и *адъективно-номинативные* трансформации являются наиболее частотными для русско-немецких параллельных текстов. В данном примере такие трансформации наиболее характерны для параллельных троек текстов, что объясняется сходством грамматических структур в немецком и английском языках.

Ниже приведен пример не параллельных, а сопоставимых текстов на немецком и русском языках. В таких текстах уже сложнее выявлять трансформации, следуя жестко регламентированной структуре патентных описаний (нумерация пунктов не совпадает, перевод текста может содержать более детальную или, напротив, более сжатую информацию, ключевые понятия могут быть заменены на синонимичные оригинальным терминам и т.д.). В частности, в примере 4 пункт 6 формулы изобретения на русском языке соответствует пункту 5 формулы изобретения на немецком языке, причем немецкий вариант более сжат по сравнению со своим русскоязычным аналогом.

(4) ***Формула изобретения на русском языке:***

4. Способ по одному из пп.1-3, отличающийся тем, что измеренное отклонение передают на управляющее устройство, которое рассчитывает необходимое изменение расстояния и инициирует изменение расстояния.

5. Способ по п.1, отличающийся тем, что экструдируют пластмассовые трубы.

6. Способ изготовления экструдированных изделий путем экструдирования расплавленного жгута из экструдера, формования расплавленного жгута в инструменте в расплавленный шланг, охлаждения и калибрования экструдированного изделия и его перемещения с помощью тянущего гусеничного устройства, содержащего по меньшей мере две гусеничные ленты, а также резки экструдированного изделия на отдельные части, отличающийся тем, что расстояние между гусеничными лентами и продольной осью гусеничного тянущего устройства в соответствии с данными измеренного изменения диаметра экструдированного изделия автоматически устанавливают в зависимости от изменения таким образом, чтобы ось симметрии экструдированного изделия совпала с продольной осью гусеничного тянущего устройства.

Формула изобретения на немецком языке:

4. Verfahren nach einem der Ansprüche 1 bis 3, dadurch gekennzeichnet, dass die gemessene Auslenkung einer Steuerung zugeführt wird, die die notwendige Abstandsänderung errechnet und die Abstandsänderung veranlasst.

5. Verfahren zum Herstellen von Extrusionsprodukten, insbesondere von Kunst- stoffrohren, mittels Extrudieren eines Schmelzestrangs aus einem Extruder, Formen des Schmelzestrangs in einem Werkzeug, insbesondere in einem Rohrwerkzeug zu einem Schmeizeschlauch, Abkühlen und Kalibrieren des Extrusionsprodukts und Abziehen desselben mit einem Raupenabzug mit min- destens zwei Raupenketten, sowie Trennen des Extrusionsprodukts in einzel- ne Stücke, dadurch gekennzeichnet, dass der Abstand der Raupenketten von einer Längsachse des Raupenabzugs nach Massgabe einer gemessenen Durchmesseränderung des Extrusionsprodukts in Abhängigkeit von dieser automatisch so eingestellt wird, dass die Symmetrieachse des Extrusions- produkts mit der Längsachse des Raupenabzugs übereinstimmt.

Поэтому если параллельные тексты имеет смысл сопоставлять и анализировать последовательно, по предложениям, то концептуально-сопоставимые тексты анализируются с опорой на содержательные объекты, то есть концепты, релевантные для данной предметной области (сами концепты, их связи, характеристики и т.п.) [7]. Такой подход оправдывает себя уже потому, что в первой сотне патентных описаний (из 455 за один год) на трех языках

большинство составили именно концептуально-сопоставимые тексты (порядка 70 %).

В программе индексации и поиска патентов (MIMOSA V5) такие тексты отмечены как эквивалентные, однако, при этом из-за различной длины и содержания текстов существенно затруднен поиск, анализ и сравнение значимых концептов. Поэтому при таком сопоставлении формул изобретения патентов на разных языках предлагается использовать предварительно сформированное лексико-семантическое представление, в котором содержатся соответствующие концепты со всеми возможными вербальными реализациями, синонимическим рядом и релевантными связями с другими концептами [8].

В примере 5 приведены фрагменты концептуально-сопоставимых текстов на немецком и русском языках:

(5) ***Формула изобретения на немецком языке:***

5. Verfahren zum Herstellen von Extrusionsprodukten, insbesondere von Kunst- stoffrohren, mittels Extrudieren eines Schmelzestrangs aus einem Extruder, Formen des Schmelzestrangs in einem Werkzeug, insbesondere in einem Rohrwerkzeug zu einem Schmeizeschlauch, Abkühlen und Kalibrieren des Extrusionsprodukts und Abziehen desselben mit einem Raupenabzug mit min- destens zwei Raupenketten, sowie Trennen des Extrusionsprodukts in einzel- ne Stücke, dadurch gekennzeichnet, dass der Abstand der Raupenketten von einer Längsachse des Raupenabzugs nach Massgabe einer gemessenen Durchmesseränderung des Extrusionsprodukts in Abhängigkeit von dieser automatisch so eingestellt wird, dass die Symmetrieachse des Extrusions- produkts mit der Längsachse des Raupenabzugs übereinstimmt.

6. Raupenabzug zum Abziehen von Extrusionsprodukten, insbesondere von Kunststoffrohren (8), mit mindestens zwei Raupenketten (10,10'), wobei die Raupenketten (10,10') mindestens während des Abziehens symmetrisch zu i einer Längsachse (12) des Raupenabzugs (6) anordenbar und an das Extru- sionsprodukt andrückbar sind, sowie mit Mitteln zum Ändern des Abstands zwischen einer Anzahl von Raupenketten (10,10') und der Längsachse (12) des Raupenabzugs (6), dadurch gekennzeichnet, dass Mittel vorgesehen sind, um Auslenkungen von Raupenketten (10, 10') orthogonal zur Längsachse (12) zu messen

Формула изобретения на русском языке:

5. Способ по п. 1, отличающийся тем, что экструдируют пластмассовые трубы.

6. Способ изготовления экструдированных изделий путем экструдирования расплавленного жгута из экструдера, формования расплавленного жгута в инструменте в расплавленный шланг, охлаждения и калибрования экструдированного изделия и его перемещения с помощью тянущего гусеничного устройства, содержащего по меньшей мере две гусеничные ленты, а также резки экструдированного изделия на отдельные части, отличающийся тем, что расстояние между гусеничными лентами и продольной осью гусеничного тянущего устройства в соответствии с данными измеренного изменения диаметра экструдированного изделия автоматически устанавливают в зависимости от изменения таким образом, чтобы ось симметрии экструдированного изделия совпала с продольной осью гусеничного тянущего устройства.

В примере 5 явно прослеживается несовпадение контента одинаково пронумерованных пунктов при практически полном соответствии пункта 5 немецкоязычного варианта и пункта 6 русскоязычного варианта формулы изобретения. В русском варианте доминируют назывные предложения с распространениями, в то время как в немецком тексте в основном представлены глагольные структуры:

- *Verfahren zum Herstellen von Extrusionsprodukten, insbesondere von Kunst- stoffrohren,……*
- *Способ изготовления экструдированных изделий путем экструдирования расплавленного жгута из экструдера……..*

Таким образом, подобные глагольно-именные соответствия, выраженные синтаксическими структурами на разных языках, можно назвать *функционально-синонимическими* [9, 10].

3. Заключение

В работе исследуется вопрос выявления номинализации в параллельных и концептуально-соспоставимых текстах формул изобретения патентов на трех языках (русском, английском и немецком), а также способ идентификации параллельности текстов для патентной сферы и типы трансформаций для русско-немецких и русско-немецко-английских текстовых представлений. Кроме того, анализируется структура патентных представлений с точки зрения извлечения лингвистической информации, а именно, концепты различных предметных областей, их характеристики, связи с другими концептами, языковые структуры, в которые они включены, трансформации, в которых они участвуют чаще всего и т. д.

Было выявлено, что наиболее частотными русско-немецкими трансформациями являются глагольно-именные трансформации. В различных типах текстов (параллельных и концептуально-сопоставимых, двойках и тройках параллельных текстов) доминируют определенные типы преобразований. В частности, наиболее распространенным является преобразование составных отглагольных существительных (немецкий язык) во фразовые

номинативные структуры (русский язык). Кроме того, было обнаружено, что в русско-немецких параллельных текстах осуществляются не только лексико-семантические преобразования, но и синтаксические. Для троек русско-немецко-английских патентных параллельных текстов, в свою очередь, наиболее характерными оказались *предложно-номинативные* и *адъективно-номинативные* трансформации, которые также являются самыми частотными для русско-немецких параллельных текстов. При этом для всех типов текстов в русскоязычных формулах изобретения характерно превалирование назывных конструкций с распространениями, в то время как в немецкоязычных текстах доминируют глагольные структуры (на всех уровнях анализа предложения).

Благодарности

Автор выражает благодарность Елене Борисовне Козеренко за конструктивное обсуждение данной работы.

Литература

1. *Козеренко Е. Б.* Лингвистическое моделирование для систем машинного перевода и обработки знаний // Информатика и ее применения, №1, том 1. — М.: Торус, 2007. — С.54–65.
2. *Козеренко Е. Б.* Глагольно-именные трансформации при англо-русском машинном переводе // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» / Под ред. Л. Л. Иомдина, Н. И. Лауфер, А.С. Нариньяни, В. П. Селегея. — М.: Изд-во РГГУ, 2007. — С. 286–294.
3. *Жолковский А. К., Мельчук И. А.* О семантическом синтезе «Проблемы кибернетики», вып. 19. М, 1967.
4. *Падучева Е. В.* О семантике синтаксиса. Материалы к трансформационной грамматике русского языка. Изд. 2-е. М: КомКнига, 2007, 296 с.
5. *Jacobs, Roderick A. and Peter S. Rosenbaum.* English Transformational Grammar. Blaisdell, 1968.
6. *Балли, Ш.* Общая лингвистика и вопросы французского языка. Изд. 2-е, М.: УРСС, 2001.
7. *Кузнецов И. П., Мацкевич А. Г.* Семантико-ориентированные системы на основе баз знаний. — М.: из-во МТУСИ, 2007. — 173 с., 59 ил.
8. *Козеренко Е. Б.* Проблема эквивалентности языковых структур при переводе и семантическом выравнивании параллельных текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» / Под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. — М.: Изд-во РГГУ, 2006. — С. 252–258.
9. *Nivre J., Boguslavski I., Iomdin L.* Parcing the SynTagRus Treebank of Russian \ Proceedings of the International Conference COLING'2008, Manchester, UK, 2008.
10. *Macken L., Lefever E., Hoste V.* Linguistically-based sub-sentential alignment for terminology extraction from a bilingual automotive corpus \ Proceedings of the International Conference COLING'2008, Manchester, UK, 2008.

Скобки в русских идиомах[1]

Parentheses in Russian idioms

Козеренко А. Д. (akozerenko@mail.ru)

Институт русского языка им. В. В. Виноградова РАН

В работе проводится анализ значения идиом русского языка, содержащих слово *скобки*. Делается наблюдение, что идиомы *взять в скобки* и *вынести за скобки* парадоксальным образом имеют одинаковое значение. Дается объяснение этому явлению; рассматриваются остальные идиомы, содержащие слово *скобки*.

«Человек, взятый под стражу, подобен тексту, взятому в скобки: он отчуждается.»

Палисандр Дальберг. Саша Соколов. (Палисандрия)

0. В этой работе мы хотим рассмотреть несколько идиом семантического поля ВАЖНОСТЬ — НЕВАЖНОСТЬ в русском языке применительно к проблеме толкования идиом с учетом их внутренней формы. Подобный опыт уже был произведен в работе [6], где нами рассматривался ряд идиом этого же семантического поля и обсуждался набор семантических компонентов, оказывающихся важными при их толковании. Ср. также работы [1] и [5], в которых вводится тематика обсуждения внутренней формы идиомы и включения ее в толкование, а также работу [2], посвященную теоретическим проблемам описания идиоматики. Описания представленных ниже идиом являются частью Фразеологического объяснительного словаря русского языка[2] и приводятся в формате, используемом в Словаре: после базовой формы идиомы указываются ее обязательные синтаксичесие валентности, стилистические, временные и т. п. пометы, приводится толкование идиомы, однословный синоним, примеры употребления идиомы в современной литературе различных жанров[3], а также в речевом общении и текстах Интернета[4].

1. В данном случае речь пойдет об идиомах, содержащих слово *скобки*: *оставить за скобками* (что-л.) *книжн.; заметить/отметить в скобках* (что-л.) *книжн.; взять в скобки* (что-л.) *книжн.; вынести/вывести за скобки* (что-л.) *книжн.* Все они отражают идею отделения несущественной части от существенной. Во внутренней форме этих идиом мы можем наблюдать скобки как знак, служащий для осуществления этого отделения. Рассмотрим следующие примеры употребления идиом *взять в скобки* (что-л.) *книжн.*; *вынести/вывести за скобки* (что-л.) *книжн.*:

(1) Культура обладает для России не только внутренней самоценностью. Культура может и должна стать экспортным товаром. (Прошу все негативное, что связано с двумя последними словами, *взять в скобки*). Экспорт культуры — стратегический фактор, значение которого трудно переоценить. Корпус Публ.

[1] Работа выполнена при финансовой поддержке РГНФ, грант 07-04-12117в.

[2] Фразеологический объяснительный словарь русского языка — проект Отдела экспериментальной лексикографии Института русского языка РАН им. В. В. Виноградова (далее Словарь). Фразеологический объяснительный словарь русского языка [4] выпущен издательством Эксмо в марте 2009г.

[3] Примеры употребления идиом в современной художественной литературе и публицистике подбирались в Машинном корпусе текстов Отдела экспериментальной лексикографии ИРЯ им. В. В. Виноградова РАН. Также использовался Национальный корпус русского языка.

[4] Более подробно об особенностях оформления леммы, принципах приписывания валентностей и системе помет см. во вступительной и заключительной статьях Словаря-тезауруса современной русской идиоматики [3]

(2) В основе мемуаров будут лежать не только личные воспоминания и впечатления автора, но и та оперативная информация, которая становилась известна Коржакову по роду его службы. Вопрос о юридической ответственности за распространение секретной оперативной информации мы *вынесем за скобки*, как и проблемы этики, поскольку в сегодняшней России политика и мораль — вещи, по-видимому, несовместные. Новая газета

Легко заметить, что в обоих случаях значение идиом примерно одинаково, его можно передать квазисинонимичными выражениями *не рассматривать что-л., не обсуждать что-л.* Ср.: <…> Прошу не рассматривать все негативное, что связано с двумя последними словами; <…> Вопрос о юридической ответственности за распространение секретной оперативной информации мы обсуждать не будем, как и проблемы этики <…>. Однако во внутренней форме этих двух идиом описываются прямо противоположные действия — *взять в скобки* и *вынести за скобки*. Внутренняя форма обеих идиом достаточно прозрачна и однозначно интерпретируется носителем языка. Как же получилось, что две идиомы с прямо противоположной внутренней формой имеют одинаковое актуальное значение? И как это сочетается с гипотезой о влиянии внутренней формы на значение и употребление идиомы?

Решение этой загадки достаточно просто. Дело в том, что в первом случае (*взять в скобки*) речь идет о грамматических скобках — парном знаке препинания, отделяющем несущественную часть текста от существенной. В случае грамматических скобок несущественная часть текста оказывается внутри скобок, а основной, существенный текст — вне скобок, снаружи. Во втором же случае (*вынести за скобки*) используются математические скобки, которые, напротив, содержат в себе существенное, тогда как несущественное выносится за их пределы (ср. математическое действие «вынести за скобки», когда внутри скобок остаются компоненты, подлежащие какому-л. дальнейшему математическому преобразованию, а за скобки выносятся те компоненты, над которыми эти преобразования производиться не будут). Этим и объясняется, что в одном случае скобки заключают в себе несущественное, а в другом — существенное.

2. Приведем толкования этих двух идиом с несколькими илюстрирующими примерами. Как мы уже говорили, актуальное значение этих идиом совпадает, поэтому отличия наблюдаются преимущественно в той части толкования, которая передает внутреннюю форму идиомы (в приведенных ниже толкованиях эта часть выделена курсивом и вводится оператором *осмысляется*[5]).

ВЗЯТЬ В СКОБКИ (что-л.) *книжн* не учитывать что-л. в процессе рассуждения, считая это вообще важным, но в данном случае не имеющим отношения к высказываемой мысли, *что осмысляется как обособление части текста парным знаком препинания, указывающим на факультативность* ❖ не рассматривать что-л., не обсуждать что-л. 🗐 Лично у меня такое впечатление, что ему было наплевать. Скорей всего он даже рад был воспользоваться моей вспыльчивостью. — Агатов предупреждающе поднял палец. — Но последнее мое утверждение — сугубо субъективное. Прошу *взять* его *в скобки*, поскольку я строго придерживаюсь фактов. Д. Гранин. Иду на грозу • Россия пополнила собою клуб великих стран, потерпевших поражение в мировой войне. Поражение России в одном отношении есть ее великий выигрыш в другом — российский народ обрел свободу и возможность этой свободой распорядиться (опять-таки *беру в скобки* отсталость нашего национального сознания, его тягостную инертность). Корпус Публ.

ВЫНЕСТИ/ВЫВЕСТИ ЗА *СКОБКИ* (что-л.) *книжн.* не учитывать что-л. из того, что находится в сфере рассмотрения, в процессе рассуждения, считая это вообще важным, но в данном случае не имеющим отношения к высказываемой мысли, *что осмысляется как вынесение чего-л. за парный знак, указывающий пределы сферы действия математических преобразований* ❖ не рассматривать что-л., не обсуждать что-л. 🗐 Отмеченные биографами и воспоминателями черты его характера, казавшиеся основными, чаровавшие миллионы сердец и умов, оказались случайными для хода истории; история государства российского не отобрала эти человечные и человеческие черты характера Ленина, а отбросила их как ненужный хлам. <…> А всё, *вынесенное за скобки*, как временное, случайное, возникшее в силу особых обстоятельств подполья и ожесточения борьбы первых советских лет, оказалось непреходящим, определяющим. В. Гроссман. Все течет • Официально предполагалось, что Хелен добилась дружбы с сенатором, заманив его самой передовой в мире марксистско-ленинской идеологией. Как она ухитрилась это сделать, не раскрывая принадлежности к советской разведке, было непонятно и потому *выводилось за скобки* и оставлялось без рассмотрения, как все непонятное. Корпус Детект.

Таким образом, в обоих случаях скобки присутствуют в той части толкования идиомы, которая передает ее внутреннюю форму, но в одном случае они описываются как грамматический знак, а в другом — как математический. Игровые примеры употребления этих идиом также обыгрывают соответствующие свойства разных скобок, ср.:

(3) <…> Курехин предложил прокатиться до Пскова. Поскольку Оля временами жила у меня, а временами ночевала у матери (не столько из своих номадических привычек, сколько из молчаливо-

[5] Более подробно об этом и других операторах внутренней формы см. во вступительной статье к Фразеологическому объяснительному словарю русского языка [4].

го обоюдного уговора — чтобы иногда разгонять кровь и давать друг другу повод для пустяковой ревности), мне порой выпадал беспризорный, скрытый от разноцветных Олиных глаз досуг, так что я легко согласился, — сегодня был как раз такой случай. Что касается *запертой в скобки* пустяковой ревности, то упоминание о ней отнюдь не значит, будто разгулу полнокровных страстей я предпочитаю всякие эрзац-страстишки. П. Крусанов. Перекуем орала на свистела

(4) Бездумно мама не бросала слов ни на ветер, ни в безветренную погоду. Она выстраивала мысли с алгебраической точностью, *вынося за скобки* всё лишнее. И почти никогда не меняла свои твёрдые точки зрения на какие-либо точки с запятыми или многоточия. А. Алексин. Раздел имущества

Совпадающая часть актуального значения этих двух идиом — ‘не учитывать что-л. в процессе рассуждения, считая это вообще важным, но в данном случае не имеющим отношения к высказываемой мысли’ — указывает на то, что обсуждаемый фрагмент (что-л.) не относится к основной линии повествования, хотя и является важным сам по себе (и может относиться, допустим, к какой-либо боковой линии повествования).

Актуальное значение идиомы *вынести/вывести за скобки* (что-л.) *книжн.* содержит дополнительный компонент ‘(что-л.) из того, что находится в сфере рассмотрения’. Это обусловлено тем, что в ситуации, описываемой во внутренней форме идиомы фокус внимания находится внутри скобок. Скобки как бы ограничивают сферу рассмотрения, за пределы которой выносится все несущественное для рассмотрения (в данном случае — математического преобразования). В ситуации же, описываемой во внутренней форме идиомы *взять в скобки* (что-л.) *книжн.* фокус внимания говорящего находится на основном тексте за скобками. В скобки заключается то, что не должно попасть в сферу рассмотрения говорящего.

Рассмотрим остальные идиомы с математическими и грамматическими скобками.

3. К идиомам с математическими скобками во внутренней форме относятся идиомы *оставить за скобками* (что-л.) *книжн., вынести/вывести за скобки книжн. 1,2*.

Идиома *вынести/вывести за скобки книжн.* помимо значения *вынести/вывести за скобки* (что-л.) имеет второе значение — *вынести/вывести* (что-л.) *за скобки* (чего-л.), ср.:

(5) Поэт — не профессия, а особый душевный строй, воплощенный не только в строках, рифмах и ритмах, но и в образе жизни, в каждом движении и поступке. Отсюда странноватость обличья и шокирующие нарушения хорошего тона, как бы *выносящие* работу поэта *за скобки* нормального общества. Корпус Публ.

В этом случае скобки как бы разграничивают некоторые области социальной жизни на непересекающиеся части. Внутри скобок остается бóльшая и более значимая область (обозначаемая дополнением в род.п.), за скобки выносится что-л., не вписывающееся в эту область по каким-либо причинам. Ср. толкование и иллюстрирующие примеры:

ВЫНЕСТИ/ВЫВЕСТИ (что-л.) ЗА *СКОБКИ* (чего-л.) исключить что-л. (обычно явление социальной жизни) из класса явлений, к которому оно принадлежит, *что осмысляется как вынесение чего-л. за парный знак, указывающий пределы сферы действия математических преобразований* Необходимой для страны мобилизации и необходимой консолидации не может сегодня дать ни Гайдар со всей своей командой, ни Чубайс и т.д. И прежде всего в силу их либеральных убеждений — стремления *вывести* государство *за скобки* общественной жизни. Независимая газета • Навязший в зубах остаточный принцип финансирования культуры связан не с тем, что денег нет. Просто привыкли, что культура *вынесена за скобки* социальной политики. Она иррациональная и, как только начинает сама пытаться зарабатывать, превращается в свою меркантильную и циничную противоположность. Новая газета

Идиома *оставить за скобками* (что-л.) *книжн.* имеет то же актуальное значение, что идиомы *взять в скобки* (что-л.) *книжн.* и *вынести/вывести за скобки* (что-л.) *книжн.* Отличие во внутренней форме от идиомы *вынести/вывести за скобки* (что-л.) *книжн.* заключается только в том, что в одном случае несущественное выносится из ограниченной скобками области, где содержится существенное, а в другом — не вносится в эту область. В том и другом случае внутри скобок остается только существенное. Ср. толкование этой идиомы и иллюстрирующие примеры:

ОСТАВИТЬ ЗА *СКОБКАМИ* (что-л.) *книжн.* не учитывать что-л. в процессе рассуждения, считая это вообще важным, но в данном случае не имеющим отношения к высказываемой мысли, *что осмысляется как невведение чего-л. в сферу действия математических преобразований, ограниченную соответствующим парным знаком* ❖ не рассматривать что-л., не обсуждать что-л. Что ждет меня впереди? Вправе ли я был в двадцать три года так резко, радикально менять свой жизненный путь? За моей спиной к тому времени была хоть и небольшая, но все-таки, как теперь говорят, рабочая биография. *Оставим за скобками* военное детство с его бомбежками, обстрелами <…> как-никак, к тому времени я успел закончить горно-металлургический техникум в Липецке <…>. Н. Пеньков. Была пора • Российские верхи по-прежнему исходят из статуса своей страны как великой державы, одного из столпов нынешнего и будущего миропорядка. При этом они, конечно, хорошо сознают, но *оставляют за скобками* как нечто преходящее нынешнюю слабость Рос-

сии, фактически не позволяющую ей играть такую роль. Корпус Публ.

Допустимо употребление этой идиомы в пассивной форме, ср.:

(6) Встречи продавца и покупателя обычно проходят строго в присутствии маклера. Продавец в соглашении оговаривает с фирмой цену своей квартиры, но за какие деньги она продается на самом деле — *остается за скобками*. А это источник неконтролируемого «навара» для фирмы. Корпус Публ.

Встречаются единичные употребления с дополнением в родительном падеже (*оставить* (что-л.) *за скобками* (чего-л.)) в значении не рассматривать что-л., не обсуждать что-л., однако они крайне редки и такое употребление находится на грани языковой нормы, ср.:

(7) Однажды мне снилось, что по какому-то досадному стечению обстоятельств я стал ангелом на шпиле Петропавловского собора и, спасаясь от пронизывающего ветра, пытаюсь застегнуть пиджак, пуговицы которого никак не желают пролезать в петли, — при этом удивляло меня не то, что я вдруг оказался высоко в ночном петербургском небе, а то, что мне никак не удается эта привычная операция. Нечто похожее я испытывал и сейчас — нереальность происходящего *оставалась* как бы *за скобками* моего сознания; сам же вечер был вполне обычным, и если бы не легкое покачивание вагона, вполне можно было бы предположить, что мы сидим в одном из маленьких петербургских кафе и мимо окна проплывают фонарики лихачей. В. Пелевин. Чапаев и пустота

4. К идиомам с грамматическими скобками во внутренней форме относятся *заметить/отметить в скобках (что-л.) книжн.; взять в скобки (что-л.) книжн*. Идиома *заметить/отметить в скобках (что-л.) книжн.* употребляется преимущественно перформативно, ср.:

(8) Просыпались поздно. Иногда дневное время посвящалось делам, причем многие довольно много работали, как это ни странно, но работали незаметно, не сознаваясь в этом другим. Помню, Осетинский заклинал меня: никому не рассказывай, что я пишу сценарий за три дня. *В скобках заметим*, что при удачном стечении обстоятельств и запуске в производство на гонорар от одного сценария можно было широко жить пару лет. Н. Климонтович. Далее — везде

Вот как выглядит толкование этой идиомы в Словаре:

ЗАМЕТИТЬ/ОТМЕТИТЬ В *СКОБКАХ* (что-л.) *книжн.* сказать или написать что-л. не относящееся к теме обсуждения, но вообще важное, *что осмысляется как обособление этого высказывания парным знаком препинания, указывающим на факультативность* ❖ добавить что-л. 🗐 <…> за вечерним чаем на веранде Иван Дмитриевич становился слезлив, сентиментален и даже зачем-то рассказывал о таких преступлениях, которые не сумел раскрыть. Все это для мемуаров было совершенно лишним, но Сафронов понимал: хотя лирическая влага разжижает сюжет, ее можно отжать из текста лишь вместе с кровью героя. *Заметим в скобках*, что так он в итоге и поступил. Л. Юзефович. Дом свиданий • Не то тебя, скажу тебе, сгубило, / <…> не то, что — за печатями семью — / Елизавета Англию любила / сильней, чем ты Шотландию свою / (*замечу в скобках*, так оно и было); <…> Они тебе заделали свинью / за то, чему не видели конца / в те времена: за красоту лица. И. Бродский. 20 сонетов к Марии Стюарт

Все идиомы, содержащие слово скобки, в особенности это относится к идиоме *заметить/отметить в скобках (что-л.) книжн.*, часто встречаются в предложении, обособленном скобками.

5. Итак, мы рассмотрели четыре идиомы, содержащие слово *скобки* в русском языке, одна из них представлена в двух значениях. Интересно, как идиомы, включающие слово *скобки*, описаны в существующих современных фразеологических словарях. Словарь Молоткова [ФСРЯ 1986] содержит только одну идиому со словом *скобки*: в скобках — между прочим, попутно, кстати (сказать, заметить и т.п.). Словари Телии [Телия 1995, Телия 2006] не содержат идиом со словом *скобки*. Словарь Яранцева [Яранцев 1997] содержит одну идиому со словом *скобки*: в скобках — между прочим, кстати, попутно (сказать, заметить, упомянуть, напомнить и т.п.)(ср. описание этой идиомы в [ФСРЯ 1986]). В словаре Мокиенко [Мокиенко 1997] нет идиом со словом *скобки*. В словаре Лубенской [Лубенская 1997] представлены две идиомы со словом *скобки*, снабженные более точными и развернутыми толкованиями на английском языке: В СКОБКАХ сказать, заметить, прибавить, упомянуть и т.п. — (to note, add, say stc sth.) incidentally, as an aside from the main topic, и ВЫНОСИТЬ/ВЫНЕСТИ ЗА СКОБКИ — to separate a certain question, problem, occurance etc from the context in which it belongs.

Таким образом, можно заключить, что эти идиомы недостаточно полно описаны в существующих словарях. Более полное и подробное описание, которое было представлено в настоящей статье, вошло в новый лексикографический источник.

Литература

1. *Баранов А. Н., Добровольский Д. О.* Внутренняя форма идиом и проблема толкования // Известия АН. Серия литературы и языка, 1998. Том 57, № 1, с. 36–44.
2. *Баранов А. Н., Добровольский Д. О.* Аспекты теории фразеологии. М., 2008
3. *Баранов А. Н., Добровольский Д. О., Киселева К. Л., Козеренко А. Д.* при участии М. М. Вознесенской и М. М. Коробовой, под редакцией А. Н. Баранова и Д. О. Добровольского. Словарь-тезаурус современной русской идиоматики, М., 2007
4. *Баранов А. Н., Вознесенская М. М., Добровольский Д. О., Киселева К. Л., Козеренко А. Д.* Фразеологический объяснительный словарь русского языка, М., 2009
5. *Добровольский Д. О.* Образная составляющая в семантике идиом // ВЯ 1996, N 1.
6. *Козеренко А. Д.* Идиомы семантического поля ВАЖНОСТЬ — НЕВАЖНОСТЬ в русском языке // Труды Международного семинара Диалог'2006 по компьютерной лингвистике и ее приложениям. Москва, 2006. С. 248–251.

Словари

- *Лубенская* 1997 — Лубенская С.И. Русско-английский фразеологический словарь. М., 1997.
- *Мокиенко* 1997 — Мелерович А. М., Мокиенко В. М. Фразеологизмы в русской речи. Словарь. М., 1997
- *Телия* 1995 — Т. С. Аристова, М. Л. Ковшова, Е. А. Рысеева и др.: под ред. В. Н. Телия. Словарь образных выражений русского языка. М., 1995
- *Телия* 2006 — Большой фразеологический словарь русского языка. Значение. Употребление. Культурологический комментарий/ Отв. Ред. В. Н. Телия. М., 2006
- *ФСРЯ* 1986 — Фразеологический словарь русского языка / Под. ред. А. И. Молоткова. М., 1986.
- *Яранцев* 1997 — Яранцев Р. И. Русская фразеология. Словарь-справочник: Ок. 1500 фразеологизмов. М., 1997

Методика корпусных исследований паузации на примере изучения паузации в японском языке в контексте послелогов и топика[1]

Pauses after postpositions and topical particle *wa* in Japanese: a corpus study

Комарова А. Д. (komarovochka@gmail.com)

Российский государственный гуманитарный университет

В данной работе описывается паузация в японском языке до и после первичных послелогов и топикальной частицы. Рассматриваются такие параметры, как вероятность, длина и заполнение данных пауз, а также случаи, когда они оказываются на разных типах синтаксических границ.

В данной работе рассматривается паузация в устном японском дискурсе в позициях до и после приименных послелогов и топикализующей частицы.

Материалом для исследования служил корпус из 45 тестов — «рассказов по картинкам». Всего от двенадцати говорящих было записано по четыре текста с рассказами и пересказами несложных сюжетов по наборам картинок, однако в связи с плохим качеством записи 3 текста не использовались.

Для каждого из тестов было посчитано абсолютное число вхождений того или иного послелога (частицы), количество пауз после данного показателя, а также случаи, когда в непосредственной близости встречались речевые сбои — заполненные паузы или коррекции.

Паузация представляет интерес как одно из явлений, присущих устному дискурсу. В устной речи паузы обладают рядом функций, например, дают говорящему возможность сделать вдох. В данной работе принят когнитивный подход, в соответствии с которым позиции, в которых возникают паузы, обусловлены, в частности, синтаксическими критериями.

Так, в [Chafe 1994] было показано, что интонационные единицы (т. е., условно говоря, фрагменты от паузы до паузы) в 60% случаев соответствуют клаузам. Аналогичные выводы для русского языка были сделаны в [Кривнова, Чардин 1999]. А в [Кибрик, Подлесская 2009] также указываются такие характерные позиции «планирования», т. е. синтаксические границы, на которых говорящий делает паузу, позволяющую ему спланировать новый отрезок, как границы предложений.

Данные японского языка представлены в [Iwasaki 1993] и [Sakura, Fuji 2006], причем в первой работе утверждается, что интонационные единицы в японском языке мельче, чем в английском, а во второй показывается, что, как правило, интонационные единицы, меньшие, чем клаузы — это топикальные группы.

Паузация в японском языке на границах предложений и клауз, входящих в сложные предложения, рассматривалась в [Комарова 2008]. Результаты этой работы показали, что паузы на границах предложений в устном японском дискурсе возникают в 94% случаев и часто (в 23% случаев) являются заполненными, а паузы на границах клауз возникают приблизительно в 60% случаев.

Паузы после топикализующей частицы и послелогов относятся, как правило, к паузам внутри клауз, но в некоторых случаях могут совпадать с паузами на границах клауз и предложений.

1. Паузы после топикализующей частицы *wa*

Частица *wa* маркирует топик, т. е. ту именную группу, о которой в предложении что-то сообщается, тему высказывания (см. [Алпатов, Аркадьев, Подлесская 2008] часть 2, стр.36–55). Топикализующая частица *wa* встретилась в корпусе всего 205 раз.

[1] Работа выполнена при финансовой поддержке РГНФ, грант №09-04-00106а

В 72 случаях после данного показателя были незаполненные паузы. Еще в 27 случаях данные паузы были заполненными или смешанными (т. е. содержали в себе и лексически или долексически заполненный фрагмент, и незаполненный). Таким образом, всего паузы после топикализующей частицы *wa* встретились в 48% случаев (99 раз), причем в 27% случаев эти паузы были заполненными (здесь и в дальнейшем под заполненными паузами будут пониматься и заполненные, и смешанные, т. к. в данном случае принципиально именно наличие «филлера» — заполнителя).

Что касается длины пауз после *wa*, то для незаполненных пауз среднее статистическое составило 0,5 сек, медиана — 0,3 сек, мода — 0,1 сек. Для смешанных пауз эти показатели сильно отличались и составили, соответственно, 1,2 сек, 1,3 сек и 1,4 сек. В данном случае интересно сравнить показатели для незаполненных и смешанных пауз. Сопоставление показывает, что длина стандартной незаполненной паузы значительно (почти на 1 сек.!) меньше, чем длина паузы с заполнением. Это показывает, что наличие заполнителя связано с когнитивными затруднениями говорящего и удлиняет паузу. Если рассматривать все (незаполненные и смешанные) паузы вместе, то общие показатели составляют: среднее статистическое — 0,7 сек, медиана — 0,4 сек и мода — 0,1 сек.

С синтактико-семантической точки зрения позиции *wa* в тексте были не равноценными. Так, наиболее характерно употребление топика в начале предложения, например:

(1) WAK T-1

4	男の人	は	町	へ	出掛けて、
	Otokonohito	wa	machi	e	dekake-te,
	мужчина	TOP	город	в	выйти-CNV

Мужчина вышел в город…

Впрочем, начальное положение не обязательно, топику могут предшествовать союзные наречия, другие актанты глагола и т. п.:

(2) KEN R-2, 1

ある	朝	たろうくん	は		目	が	覚めて
Aru	asa	tarou-kun	wa	…(0,6)	me	ga	same-te
один	утро	Таро-кун	TOP		глаза	NOM	открыть-CNV

Однажды утром Таро открыл глаза…

В примере (2) топикальной именной группе предшествует обстоятельство времени «однажды утром», однако функция топика здесь та же, что в примере (1) и принципиально, что пауза после *wa* по типу относится к «паузам внутри клауз».

Однако в корпусе также встречаются примеры, когда именная группа, маркированная показателем *wa,* оказывается не внутри, а на границе клауз. Таких примеров в корпусе всего 18, их можно разделить на 4 типа в зависимости от синтаксической структуры.

В первую очередь положение топикализующей частицы на границе клауз связано с цитационными конструкциями, см. пример (3):

(3) KEN T-1

13	お子供たち			は
	o-kodomo-tachi			wa
	HON-ребенок-PL			TOP
14		十	円	出したら
	..(0,3)	juu	en	dash-itara
		десять	йена	вынимать-COND
15	教えて			あげる
	oshie-te			age-ru(0,2)
	объяснять-CNV			дать-PRS
16	と			言いました。
	to			i-imash-ita.
	QUOT			сказать-ADR-PST

Дети сказали: «Если дашь 10 йен, объясним».

Цитируемое предложение фактически представляет собой вложенную предикацию, которая одновременно заполняет валентность глагола «говорить» на прямое дополнение (говорить что?). Цитируемый фрагмент сам по себе представляет сложное предложение из двух клауз. Если «выкинуть» данный фрагмент из главного предложения, то топикализующая частица *wa* окажется в стандартной позиции внутри клаузы.

В пяти из десяти (т. е. в 50%, как и в целом для позиции после wa) случаев подобных конструкций с передачей прямой речи возникают паузы после *wa*. Как правило, они довольно краткие, медиана длин данных пауз составляет 0,3 сек., а среднее арифметическое — 0,74 сек. Лишь в одном случае длина паузы достигает 2,7 сек и, видимо, свидетельствует о когнитивных затруднениях говорящего, хотя у этого носителя в целом паузы длиннее, чем у других.

По своей длине паузы после *wa* перед цитацией скорее близки к паузам после *wa* внутри клаузы, чем к паузам на границе клауз (см. подробнее [Комарова, 2008]). Заполненных пауз в данной позиции не встретилось.

На цитирование похож еще один тип конструкций, в которых показатель топика *wa* оказывается в позиции на границе клауз, — это определительные придаточные. В японском языке как в языке с левым ветвлением определительное придаточное предшествует определяемому слову. Оно может разрывать главное предложение и оказываться, как и цитируемый фрагмент, вложенной предикацией, «отрывающей» топик от смыслового глагола.

(4) YAM T-1

30		これ	は
	…(0,8)	Kore	wa
		Это	TOP

31	君	の	ほしがって	いた
..(0,2)	kimi	no	hoshigat-te	i-ta
	ты	GEN	хотеть-CNV	AUX.PRG-PST
32	車		だ	よ
	kuruma		da	yo”
	машина		COP	PRT

Это машина, которую ты хотела.

В примере (4) 2-я строка представляет собой относительное придаточное с мишенью релятивизации *kuruma* (машина). Без этого фрагмента главное предложение представляет собой грамматически и семантически законченную предикацию, в отличие от цитационных конструкций, где осталась бы незаполненной валентность глагола говорения на содержание.

Всего в корпусе показатель топика перед определительными придаточными встретился 4 раза. В трех случаях после *wa* были паузы длиной 0,1, 0,2 и 0,2 сек. Заполненных пауз в данной позиции не встретилось.

Длина пауз после топикализующей частицы перед определительными придаточными свидетельствует о том, что данные паузы скорее относятся к «паузам внутри клауз после *wa*» , чем к паузам на границах клауз.

Еще один тип конструкций, при котором граница после *wa* совпадает с границей клауз, — это случаи, когда *wa* маркирует союзные имена. При этом сами союзные имена оформляют предикативные обстоятельства, поэтому данный тип границ относится непосредственно к границам между клаузами. Например:

(5) YAM R-2

6		ワイン	を	飲んだ	後	は
	…(0,9)	Wain	o	non-da	ato	wa
		вино	ACC	пить-PST	после	TOP
7	再び	スキ	に	行く	の	です	が
	futatabi	suki	ni	ik-u	no	des-u	ga
	второй.раз	лыжи	DAT	идти-PRS	NML	COP.ADR-PRS	но

После того, как (он) выпил вина, второй раз пошел на лыжах…

В примере (5) показатель *wa* относится к союзному имени *ato* со значением «после», а в целом *ato wa* присоединяется к финитной форме глагола *non-da* «выпил». Таких случаев в корпусе было 3. В одном из них после топикализующей частицы была незаполненная пауза 0,2 сек.

Данная пауза достаточно коротка, однако описанный тип границ невозможно отнести к типу «внутри клаузы», т. к. союзное имя с показателем топика маркирует глагольную форму, которая в японском языке занимает конечное положение в клаузе.

Наконец, в корпусе встретился один пример, когда показатель *wa* оказался на границе клауз в связи с речевым сбоем:

(6) SHU R-2

1	二つめ	の	話	は==
	futatsume	no	hanashi	wa ==
	второй	GEN	рассказ	TOP
			男の子 = =	
….(1,5)	ee(0,3)		otokonoko ==	
			мальчик	

Второй рассказ… Мальчик…

В примере (6) говорящий, видимо, хочет сформулировать, о чем будет рассказ в целом, но отказывается от этой идеи и начинает передавать сюжет поэтапно. О когнитивных затруднениях говорящего свидетельствуют довольно длительная незаполненная и заполненная паузы.

Итак, из 18 случаев, когда топикализующая частица оказалась на границе клауз в 14 случаях по типу данную границу можно отнести к «паузам внутри клауз». В 8 случаях из 14, т. е. в 57% случаев на данных границах были паузы. Этот показатель близок к общему показателю «пауз после *wa*» (48%), хотя несколько выше. Заполненных пауз в данных примерах не было, а длина незаполненных, как было показано, также по медиане (0,2 сек) близка к общему показателю для *wa* (0,3 сек).

Оставшиеся 4 примера также не влияют на статистику, т. к. единственная пауза после союзного имени довольно короткая (0,2 сек), а пауза после обрыва — заполненная.

2. Паузы после ga

Послелог *ga* — показатель номинатива, маркирующий подлежащее. *Ga* относится к первичным послелогам, т. е. выражает исключительно синтаксическую связь, в данном случае — предиката с первым актантом (см. подробнее [Алпатов, Аркадьев, Подлесская, 2008] стр. 190–194) и может вытесняться показателем топика *wa*.

В корпусе данный послелог встретился 112 раз. В 18 случаях (т. е. в 16%) после него были паузы, лишь одна из них была заполненной, см. пример (7)

(7) MAY R-1, 1

ええと		お父さん	が		ええと	家族	の
eeto	..(0,2)	o-too-san	ga	…(0,5)	eeto	kazoku	no
HEZ		HON-отец-сан	NOM		HEZ	семья	GEN
プレゼント	を	買い	に	行きました。			
purezento	o	ka-i	ni	ik-imash-ita			
подарок	ACC	покупать-CNV	DAT	идти-ADR-PST			

Ну, отец пошел покупать подарок семье…

Так как в корпусе встретился всего один пример с заполненной паузой после *ga*, в данном случае пока-

затели длины измерялись для всех пауз после *ga* в целом и составили: среднее статистическое — 0,44 сек., медиана и мода — 0,4 сек. Из этих данных видно, что примере (7) пауза после *ga* — стандартной длины для данной позиции и не влияет на статистику, а на когнитивные затруднения говорящего указывает лексический маркер хезитации — *eeto* («ну»).

Как и в случае с топикализующей частицей *wa* часть пауз после *ga* формально оказались на границе, а не внутри клауз. Все эти примеры были связаны с цитационными конструкциями, см. пример (8):

(8) YUK R1

14		そうすると	女の子	が
	..(0,4)	Sousuruto	onnanoko	ga
		Тогда	девочка	NOM
15		車	が	ほしい
	..(0,3)	“Kuruma	ga	hoshi-i”
		машина	NOM	желаемый-PRS
16		と	言います。	
		to	i-imas-u.	
		QUOT	говорить-ADR-PRS	

Тогда девочка говорит: «Хочу машину».

Всего таких случаев в корпусе встретилось 7, причем в 6 из них (т.е. в 86%) перед прямой речью были паузы. Заполненных пауз в данной позиции не встретилось, длина незаполненных по своим показателям была близка к длине пауз после ga в целом (стоит, однако, учитывать, что паузы после *ga* в цитационных конструкциях составили треть от всех пауз после *ga*). Среднее статистическое составило 0,42, медиана — 0,45 и мода — 0,5 сек.

Интересно, что перед прямой речью после показателя *ga* паузы возникают чаще, чем в такой же позиции после топикализующей частицы.

Можно составить такую таблицу:

Показатель	На границе цитационных конструкций	Количество пауз
wa	10	5
ga	7	6

Итак, паузы после *wa* возникают лишь в 50% случаев (в 5 из 10), а после *ga* — в 86%. Возможно, это связано с тем, что топикализующая частица в большей степени выполняет выделительную функцию, а после *ga* требуется небольшая пауза, чтобы отделить прямую речь от главного предложения.

На это указывает также тот факт, что вне конструкций с цитацией паузы после *ga* встречаются значительно реже, чем после *wa* (ср. 16% и 48% соответственно).

3. Паузы после *o*

Послелог *o* в японском языке маркирует второй актант глагола — прямое дополнение. Как и *ga*, он относится к первичным послелогам и отражает исключительно синтаксическую связь между предикатом и его актантом.

O — самый частотный из рассматриваемых послелогов, он встретился в корпусе 293 раза. В 22 случаях (т.е. 7,5%) после него были паузы, 4 из них (т. е. 18%) оказались заполненными. Среднее статистическое длин данных пауз — 0,4 сек., медиана — 0,2 сек. и мода — 0,1 сек.

В абсолютном большинстве случаев послелог *o* встретился в позиции непосредственно перед предикатом, например:

(9) WAK R-2, 4

男の人	は	朝ご飯	を	食べました。	
Otokonohito	wa	asagohan	o	tabemashita.	
мужчина	TOP	завтрак	ACC	есть-ADR-PST	

Мужчина позавтракал…

(10) KIM T-2, 5

その後		スキ	を	し	に	出かけました。
....(1) Sonoato	...(0,5)	suki	o	shi	ni	dekake-mash-ita.
Потом		лыжи	ACC	делать	DAT	выйти-ADR-PST

После этого (он) пошел кататься на лыжах.

В примере (9) *o* маркирует прямой объект простого глагола, а в примере (10) — прямой объект целевой конструкции. Однако в обоих случаях он ничем не отделен от предиката. Возможно, именно вследствие близости к предикату вероятность пауз перед *o* минимальна.

В некоторых случаях достаточно длительная пауза после *o*, вероятно, может свидетельствовать о хезитации:

(11) DAI T-2

7					でも	その	まま
	su=	ee(0,4)	….(1,7)		demo	sono	mama
					но	такой	как.есть
8			スキー	を		した	ので
	….(1,2)		sukii	o	….(1,0)	shi-ta	node(0,8)
			лыжи	ACC		делать.PST	так как

Но, так как (он) в том состоянии катался на лыжах…

В примере (11) видно, что уже в начале клаузы у говорящего возникают сложности, и он не может сразу породить нужный фрагмент. Он обрывает неудачный фрагмент, за этим следует смешанная пауза. Перед именной группой, оформленной *o*, снова возникает довольно длинная пауза и, наконец, сделав паузу после *o*, говорящий тянет союз, следующий

за предикатом. Все это отражает проблемы планирования данного фрагмента.

В двух случаях послелог *o* оказался в позиции на границе клауз:

(12) YAM R-1, 16

	子供	たち	を
….(1,3)	Kodomo	tachi	o:
	ребенок	PL	ACC

	「母	さん	は	何	を‖	母	さん	に	は
…(0,8)	“kaa	san	wa	nani	o‖	kaa	san	ni	wa
	мама	сан	TOP	что	ACC	мама	сан	DAT	TOP

何	を	買ったら	
nani	o	kat-tara…	
что	ACC	купить-COND	
機嫌	を	直して	もらえる
…(0,7) kigen	o	naosh-ite	mora-e-ru,
настроение	ACC	починить-CNV	получить-POT-PRS
と		思う?」	
to		omo-u?”	
QUOT		думать-PRS	

Детям: «Что бы купить маме, чтобы улучшить (ее) настроение, как вы думаете?»

(13) KEN T-1, 37

	それで	そういう=	そういう	こと を==	
….(9,2)	Sorede(0,3)	sooyu=	sooyuu	koto	o==
	тогда		вышеназванный	NML	ACC
家	へ	戻って		から	
uchi	e	modot-te		kara(0,3)	
дом	в	возвращаться-CNV		после	

Это= После того, как (он) вернулся домой…

В примере (12) *o* вводит прямую речь. Интересно, что глагол говорения (в данном случае, это мог бы быть *tazuneru* «спрашивать») в этом фрагменте опущен. В качестве следующей предикации приводится ответная реплика детей. Как и в примере с *ga* (ср. пример (8)) говорящий делает паузу после *o*, вероятно, чтобы выделить прямую речь. В примере (13) рассказчик отменяет фрагмент перед *o*, с порождением которого у него возникли трудности (в данной клаузе уже был один обрыв), и произносит другой фрагмент, причем он не делает паузы перед откорректированным вариантом.

4. Паузы после ni

Послелог *ni* маркирует косвенное дополнение глагола. Он, так же как показатели *ga* и *o* относится к первичным послелогам, однако чаще, чем они, может отражать несинтаксический компонент значения, т. е. выполнять вторичные функции.

Показатель *ni* встретился в корпусе 207 раз. В 48 случаях (т. е. 23%) после него были паузы. 5 из них (т. е. 10%) оказались заполненными. Среднее значение длины пауз после *ni* составило 0,5 сек., медиана — 0,4 сек. и мода — 0,2 сек.

В 8 примерах позиция после *ni* совпала с границей клауз. В первую очередь, как и с другими послелогами, это было связано с началом прямой речи. Интересен пример (14):

(14) YUK T1

	で	とりかか=		あ	自分	の	子供	を=
…(0,7)	De	..(0,4) torikaka=		..(0,2)	a jibun	no	kodomo	o=
	И				сам	GEN	ребенок	ACC
			に					
..(0,4)			ni					
			DAT					

何	が	ほしい	か
“nani	ga	hoshi-i	ka”
Что	NOM	желаемый-PRS	Q
を	一所懸命	尋ねる	と
o	ishshokenmei	tazune-ru	to
ACC	изо.всех.сил	спросить-PRS	TEMP

Своих детей: «Что вы хотите» — изо всех сил спрашивал…

В данном случае говорящий маркирует адресата локутивного глагола послелогом прямого объекта *o*, однако вспоминает о том, что более грамотно с точки зрения модели управления использовать *ni*. Он отменяет неправильный фрагмент и после паузы заменяет его на более подходящий. При этом паузы после *ni* он не делает, возможно, именно потому, что она возникла перед послелогом (отдельно о паузах перед послелогами будет сказано ниже). Саму же прямую речь как прямой объект глагола говорения говорящий «законно» маркирует послелогом *o*.

Еще один пример аналогично примеру (12) иллюстрирует эллипсис глагола говорения.

(15) SHU T-1

6	….(2,4)	で	子供たち	に
		de	kodomo-tachi	ni
		Тогда	ребенок-PL	DAT
7	….(1,0)		子供たち	は
			kodomo-tachi	wa
			ребенок-PL	TOP

Тогда к детям… Дети…

В примере (15) описывается ситуация, когда отец обращается к детям и дальше следует их прямая речь. Однако говорящий опускает локутивный предикат, не заканчивая 6-ю строку. В данном случае непонятно, насколько серьезные когнитивные затруднения возникли у рассказчика. То ли он не смог закончить фразу и подобрать нужный глагол, то ли ре-

шил, что и так ясно, что он имел в виду, и заканчивать данное предложение не обязательно. Незаполненная пауза перед следующей репликой и отсутствие попыток произнести какой-нибудь глагол говорения свидетельствуют в пользу второго предположения.

Наконец, интересный пример, когда показатель *ni* оказался не внутри клаузы, связан со случаем парцелляции:

(16) KIM T-2

3	紅茶	と		ピザ	を	食べました。
	koocha	to	..(0,3)	pizza	o	tabe-mash-ita.
	черный.чай	и		пицца	ACC	есть-ADR-PST
4		朝ご飯				に。
		Asagohan				ni.
		Завтрак				DAT

Съел пиццу с черным чаем. На завтрак.

В примере (16) говорящий выносит обстоятельство за глагол, хотя в японском языке глагол должен занимать конечное положение в предложении. Это может быть связано с тем, что рассказчик забыл нужный элемент и не успел поставить его перед глаголом. Или он хочет коммуникативно выделить этот фрагмент, например, указывая, что пицца не самый стандартный вариант завтрака.

Всего в корпусе встретилось 7 случаев, когда показатель *ni* использовался для введения прямой речи. При этом, кроме случая, рассмотренного в примере (15), где пауза после *ni* составила 1,0 сек., еще трижды перед прямой речью возникали короткие паузы (0,1, 0,2 и 0,2 сек.).

В примере (16) пауза после парцеллированного элемента составила 1,0 сек. По своим длине и позиции она относится к типу «пауз на границе предложений».

5. Паузы после no

Послелог *no* маркирует, в отличие от рассмотренных ранее *ga, o, ni*, атрибутивную, а не актантную связь. Он встретился в корпусе 96 раз. Всего в 4-х случаях (т.е. в 4%) после *no* были паузы, причем две из них были заполненными.

(17) ISI R-2, 26

私	の	ええと	左	腕	と	右	足	は
watashi	no	eeto	hidari	ude	to	migi	ashi	wa
я	GEN	HEZ	левый	рука	и	правый	нога	TOP

Мои левая рука и правая нога…

В примере (17) проблемы говорящего маркирует лексический показатель хезитации *eeto* (аналог русского «ну»). Видимо, когнитивные затруднения рассказчика связаны с использованием таких понятий, как «правый» и «левый», и он задумывается, пытаясь по картинке понять, о каких именно руке и ноге идет речь.

Приведем еще один пример с паузой после *no*:

(18) YAM T-1, 23

こちら	の		車	の	方		け=	え	
....(2,5) Kochira	no	...(0,9)	kuruma	no	hoo	...(0,5)	k=	...(0,6)	
Здесь	GEN		машина	GEN	сторона				

Это… Эта машина…

В примере (18) пауза после *no* — незаполненная, однако ее длина и наличие коррекции в последующем фрагменте свидетельствуют о когнитивных затруднениях говорящего. Еще в одном случае незаполненная пауза после *no* составила 0,1 сек. За одну именную группу перед ней была заполненная пауза, что также позволяет предположить некоторые трудности с порождением текста у рассказчика, однако, длина паузы незначительна, так что доказать это невозможно.

Таким образом, среднее арифметическое для длин данных пауз составило 0,8 сек., медиана — 0,9 сек., однако эти данные не показательны, т. к. высчитывались на основе 3-х примеров (четвертая пауза была заполненной, без незаполненного или долексически заполненного фрагмента, а для лексических заполнителей время не высчитывается).

6. Паузы перед послелогами

Отдельно хочется остановиться на позиции перед послелогами. В корпусе всего один (!) раз встретилась пауза перед одним из рассмотренных послелогов, см. пример (19)

(19) ISI T-2, 23

左	腕		を	骨折	して
hidari	ude	..(0,2)	o	kossetsu	shi-te
левый	рука		ACC	перелом	VRB-CNV

(Он) сломал левую руку…

Еще в одном случае говорящий тянул слово перед послелогом *ni*. В примере (19) нет речевых сбоев и пауза перед послелогом достаточно краткая, однако ее исключительно малая вероятность (это единственная пауза на 913 вхождений послелогов) позволяет говорить о возможных проблемах планирования.

Итак, данные корпуса свидетельствуют о том, что паузы перед послелогами практически невозможны.

Что касается позиции после послелогов, то:

1) Паузы после *no* возникают всего в 4% и в 50% случаев — заполненные, т. е. почти всегда свидетельствуют о хезитации;

2) Паузы после *o* возникают в 7,5% случаев, в 18% случаев заполненные, таким образом, вероятность этих пауз также минимальна;
3) Паузы после *ga* возникают в 16% случаев, в лишь 5% случаев они заполненные, что свидетельствует о возможности данных пауз;
4) Паузы после *ni* возникают еще чаще — в 23% случаев, в 10% случаев они заполненные, т.е., видимо, паузы в данной позиции «законны».
5) Наконец, паузы после *wa* встречаются в 48% случаев. В 27% случаев они бывают заполненными.

Итак, частота пауз в позиции после топика и вероятность наличия заполнителя указывают на то, что эта позиция является одной из «позиций планирования», как, например, граница предложений (где паузы встречаются в 94% случаев, и 23% из них — заполненные, см. [Комарова, 2008]). Это лишний раз подтверждает тот факт, что топик относится ко всему предложению в целом, т. е., говорящий может сформулировать, о чем будет идти речь, но еще не знать, что именно и как он скажет.

Кроме того, данные по послелогам *ga, ni, o* могут свидетельствовать о существовании некоторой иерархии свободы актантов. Так, наименее вероятны паузы после *o*, несколько выше их вероятность после *ga* и, наконец, после *ni* они уже довольно частотны.

Литература

1. *Алпатов В. М., Аркадьев П. М., Подлесская В. И.* Теоретическая грамматика японского языка, М. 2008
2. *Кибрик А. А., Подлесская В. И. (ред.)* Рассказы о сновидениях: корпусное исследование устного русского дискурса. Москва, 2009.
3. *Комарова А. Д.* Паузация в японском языке на границах синтаксических единиц разного уровня: корпусное исследование // международная конференция Диалог 2008
4. *Кривнова О. Ф., Чардин И. С.* Паузирование при автоматическом синтезе речи // Теория и практика речевых исследований (АРСО-99). Москва, 1999.
5. *Chafe. W.* Discourse, consciousness, and time. Chicago, 1994.
6. *Iwasaki S.* The Structure of the Intonation Unit in Japanese // Japanese/Korean Linguistics, vol. 3, ed. by Soonja Choi. Stanford, 1993, 39-53
7. *Sakura C., Fuji S.* Intonation units, information structure, and grammatical constructions in Japanese ans English.//Oral presentations. University of Tokyo, 2006.

Отсутствие пауз на границах элементарных дискурсивных единиц: опыт корпусного исследования[1]

A corpus study of pausation at syntactic boundaries: why pauses do not always appear where we expect them

Коротаев Н. А. (n_korotaev@hotmail.com)

Российский государственный гуманитарный университет

Произнесение последовательности элементарных дискурсивных единиц без пограничной паузы связано с повышенной степенью семантической интеграции, что отражается в просодических и синтаксических свойствах анализируемых сочетаний. По данным корпуса устных рассказов, доля границ без пауз может существенно варьировать от рассказа к рассказу.

1. Введение

Какими исследовательскими установками мы бы ни пользовались при анализе устного дискурса, невозможно пройти мимо того очевидного факта, что живая речь порождается не плавно, а сегментно, определенными минимальными шагами, или квантами. Такого рода шаги известны в литературе под разными названиями — интонационная группа, синтагма, фраза, ритмическая группа, интонационная единица, минимальная дискурсивная единица и др. (см., в частности, Светозарова и др. 1988, Chafe 1994: 57, Хитина 2004, Кривнова 2007: раздел 2.3, Щерба 1955, Degan, Simon 2005). Разнятся и приводимые определения этих единиц, но в целом все авторы признают, что такие кванты речи должны обладать некоторым семантико-синтаксическим и просодическим единством. В работах У. Чейфа была также предложена когнитивная интепретация: такие дискурсивные единицы считаются выразителями одного «фокуса сознания» (см. Chafe 1994), т. е. совокупности информации, которую селективное человеческое сознание может одновременно удерживать в активном состоянии. Иными словами, произнесение одной такой единицы отражает одно когнитивное усилие говорящего.

Между составляющими дискурс единицами имеются границы. Весьма часто эти границы сопровождаются паузами — периодами времени, используемыми говорящим для обдумывания следующего шага дискурса (см. Levelt 1989: 258). Такие паузы мы называем *пограничными*. В работе Clark, Clark 1977 было введено понятие «идеального речепорождения» (ideal delivery): в прототипической ситуации говорящий плавно, без сбоев и задержек, произносит одну минимальную единицу дискурса, затем делает паузу, чтобы полностью спланировать новый отрезок речи, который, в свою очередь, столь же успешно порождает. От этой идеальной модели возможны два отклонения: во-первых, в речи нередко наблюдаются паузы внутри минимальных дискурсивных составляющих (по большей части, они связаны с явлениями хезитации — см. Подлесская, Кибрик 2006, 2009, Campione, Véronis 2005); во-вторых, на границе составляющих может не быть паузы. Именно случаям второго типа посвящено настоящее исследование.

2. Материал исследования, понятийный аппарат

Материалом данной работы послужил корпус «Рассказы о сновидениях», состоящий из аудиофайлов и транскриптов 129 устных рассказов детей и подростков об увиденном ими во сне. Возраст информантов варьирует от 7 до 17 лет, суммарная продолжительность звучания составляет около 2 часов. Более полную информацию о корпусе, а также под-

[1] Работа выполнена при финансовой поддержке РФФИ, грант 07-06-00061.

робное изложение принимаемой нами концепции описания устного монологического дискурса см. в книге Кибрик, Подлесская (ред.) 2009[2].

Минимальные кванты дискурса, о которых шла речь во введении, мы называем *элементарными дискурсивными единицами* (ЭДЕ). Преимущество этого термина, на наш взгляд, состоит в том, что он, с одной стороны, наиболее точно отражает функциональную природу данных единиц, с другой стороны, позволяет единым образом описывать структуру устного и письменного модусов дискурса.

Под *паузой* мы понимаем временной отрезок, превышающий 0,05 с, во время которого либо не происходит никакой вокализации (*абсолютная пауза*), либо говорящий производит некоторый «долексический» сегмент (*простая заполненная пауза*; это может быть редуцированный гласный «шва», сонорный носовой или глоттальный скрип), либо имеется сочетание заполненной и абсолютной пауз (*смешанная заполненная пауза*). В используемой нами системе дискурсивной транскрипции каждая ЭДЕ записывается в отдельную строку; абсолютные паузы отмечаются при помощи определенного количества поднятых точек (от двух до четырех в зависимости от длительности паузы), заполненные — при помощи такого же количества графем э («шва»), м (сонорный носовой сегмент) и ’ (глоттальный скрип); после обозначения типа паузы в скобках указывается ее длительность с точностью до десятой доли секунды. Так, в примере (1), представляющем собой начальный фрагмент рассказа, содержится 3 ЭДЕ; на границах между ними имеются абсолютные паузы длительностью 0,5 с и 0,8 с соответственно; кроме того, в каждой из них есть и внутренние паузы разных типов — абсолютные (в ЭДЕ 2 и 3), простые заполненные (в ЭДЕ 3) и смешанные заполненные (в ЭДЕ 1 и 2):

084n[3]

0.0 1. Я-а ···(0.7) ’’(0.1) у своего /\дру-у-га,

2.7 2. ···(0.5) т-там ··(0.1) ээ(0.2) наш /дом ··(0.1) отрем-м= || ре-е= || ··(0.2) на \ремонт \поста-авили.

7.9 3. ···(0.8) И-и /–всех ··(0.2) /\переслали ээ(0.1) в другой \дом,

Как видно из приведенного примера, пограничные паузы записываются в транскрипте не в конце, а в начале строки. Этот выбор не случаен: функционально пограничная пауза теснее связана не с предшествующей, а с последующей ЭДЕ.

Для дальнейшего изложения крайне важно сделать следующее замечание. Определение границ ЭДЕ — многофакторная процедура, предполагающая использование целого ряда семантико-синтаксических и просодических критериев. Однако существенно, что данные о паузации в перечень этих критериев не входят и при делении рассказов на ЭДЕ не учитываются. Таким образом, в рассуждениях о наличии или отсутствии паузы на границе ЭДЕ мы избегаем порочного круга, который был бы неминуем, если бы параметры «наличие границы ЭДЕ» и «наличие паузы» были бы логически связаны. Подробнее о процедуре деления на ЭДЕ и связанных с ней теоретических и практических проблемах см. Кибрик, Подлесская (ред.) 2009: 45–388.

3. Границы ЭДЕ и паузы

Всего в исследованном корпусе содержится 3724 ЭДЕ, принадлежащих рассказчикам[4], соответственно, за вычетом 129 первых ЭДЕ каждого рассказа мы имеем 3595 границ ЭДЕ. Из них содержат паузы 2250, т. е. 62,6 %. Несложно подсчитать, что это значение ощутимо превышает вероятность появления паузы в произвольном месте между словами. Такая вероятность составляет всего 0,253 (3509 пауз на 13879 потенциальных границ между словам в корпусе), т. е. почти в 2,5 раза ниже зафиксированной частоты пограничных пауз. Иными словами, появление паузы на границе ЭДЕ — факт далеко не случайный (см. также Кибрик, Подлесская (ред.) 2009: 64–72).

И все же в 37,4 % случаев (1345 примеров) границы ЭДЕ не содержат пауз. Здесь налицо отклонение от прототипической ситуации, описанной в модели идеального речепорождения, но в то же время это явление слишком частотное, чтобы его можно было счесть маргинальным. Очевидно, что эта просодическая техника связана с достаточно стандартной когнитивной потребностью говорящего. В наиболее общей форме ее можно определить как стремление (разумеется, далеко не обязательно

[2] Стоит также упомянуть, что корпус состоит из двух частей: 69 рассказов были записаны от детей и подростков с невротическими расстройствами, 60 — от информантов из контрольной группы. В рамках проекта «Рассказы о сновидениях» между двумя частями корпуса выявлены определенные языковые различия, однако в настоящей работе мы рассматривали корпус как гомогенный и не ставили цель установить корреляции между интересующими нас параметрами и неврологическим статусом рассказчиков, хотя такая задача, безусловно, может представлять немалый исследовательский интерес.

[3] Здесь и далее после номера указывается его идентификационный код в корпусе; слева от транскрипционной записи каждой ЭДЕ указывается время начала ее произнесения (за нулевую отметку принимается начало соответствующего звукового файла) и ее номер в рамках рассказа. Подробнее о формате транскрипции см. Кибрик, Подлесская (ред.) 2009.

[4] Еще 52 ЭДЕ принадлежат интервьюеру и другим участвовавшим в записи лицам.

осознанное) выразить более высокую, чем обычно, степень семантической интеграции ЭДЕ. Ниже мы подробнее проанализируем ряд свойств, демонстрируемых последовательностями ЭДЕ без пограничной паузы.

4. Отсутствие пограничной паузы: доминирующие модели

Рассмотрим один из 1345 случаев отсутствия паузы на границе ЭДЕ — пример (2), в котором без пограничной паузы произносятся ЭДЕ 7 и 8:

061n
15.4 6. ···(0.9) Он показывал кр-расный \све-ет.
17.9 7. ···(0.6) Я /испуга̲лся,
19.2 8. и вышел н-на \у̲лицуw.
20.9 9. ····(2.2) Но это ещё не-е **о-о̲чень** страшно было.

Пример (2) показателен сразу с нескольких точек зрения. В нем в сконцентрированном виде реализуется набор из четырех типичных свойств последовательностей ЭДЕ без пограничной паузы: ни в одной из двух соседствующих ЭДЕ не содержится внутренних пауз; соседствующие ЭДЕ входят в одно предложение; с синтаксической точки зрения они относятся к простым глагольным клаузам; дискурсивный отрезок, произносимый без пограничных пауз, ограничивается двумя ЭДЕ. Разберем эти четыре свойства по очереди.

4.1. Внутренние паузы

Просодическое единство отрезка 7—8 из примера (2) подчеркивается не только отсутствием пограничной паузы, но также и тем, что ни в строке 7, ни в строке 8 нет внутренних пауз. Это стандартное положение дел для пар ЭДЕ без пограничной паузы: оно имеет место примерно в 65 % случаев. Иными словами, отклонение от модели идеального речепорождения сразу в обе стороны встречается весьма редко. Впрочем, оно все же возможно. В частности, относительно регулярна ситуация, при которой внутренняя пауза имеется во второй по порядку ЭДЕ и располагается непосредственно после начального служебного слова — союза, коннектора и проч.:

104n
0.0 1. Мне стало-о /сни̲ться,
1.0 2. что-о ···(0.7) ээ(0.3) ··(0.4) /уже начал появляться \све̲т.

Примеры типа (3) свидетельствуют о том, что говорящий может планировать синтаксический тип конструкции (в данном случае — конструкцию с сентенциальным дополнением) раньше, чем ее лексическое наполнение. Недвусмысленно определив формальную структуру своего высказывания, рассказчик берет передышку, прежде чем произнести заключительную часть конструкции[5].

Практически с той же частотой (примерно по 15 % случаев) внутренние паузы приходятся на первую ЭДЕ. В этом случае говорящий, взяв время на обдумывание посередине строки, успевает спланировать не только окончание этой ЭДЕ, но и всю следующую, см. пример (4):

105n
65.8 37. и зашла в какую-то ··(0.4) \ко̲мнату,
66.9 38. и там увидела \НЛО̲,

Наконец, в отдельных случаях (менее 5 % от всех рассмотренных примеров) с внутренними паузами произносятся обе соседствующие ЭДЕ, см. следующий пример:

108n
89.1 50. мм(0.2) ··(0.2) и-и ээ(0.3) ··(0.4) /я очен-нь ···(0.6) обратила /\внимание ··(0.2) на-а /\де̲вушку,
93.9 51. которая сидела ··(0.4) мм(0.1) ··(0.1) в \синем /кост̲юме,

Несложно заметить, что во фрагменте (5) говорящая испытывает серьезные трудности в планировании и порождении своего дискурса. Любопытно при этом, что проще всего ей дается построение каркаса определительной конструкции: без какой-либо хезитации (и без пограничной паузы!) произносится сочетание вершинного имени и начала относительного придаточного.

4.2. Паузы и границы предложения

Вопрос о том, насколько применимо к устному дискурсу одно из центральных понятий традиционной грамматики — понятие предложения, остается дискуссионным (см., например, Земская (ред.) 1973: 217–225, Chafe 1994: Ch.4, Miller, Weinert 1998). В рамках проекта «Рассказы о сновидениях» этот термин используется в следующем смысле. Под *предложением* понимается последовательность, завершающаяся так называемой *финальной* ЭДЕ — ЭДЕ, в которой происходит завершение иллокуции. Для монологического нарративного дискурса это в пер-

[5] В данном случае уже неважно, планирует ли говорящий содержательное наполнение заключительной части конструкции или только ищет подходящую вербализацию для уже составленного плана. Важно, что хезитационная пауза случается уже после того, как определена синтаксическая структура высказывания.

вую очередь иллокуция сообщения. В транскрипции конец иллокуции сообщения (и, соответственно, конец предложения) отмечается при помощи пунктуационного знака «точка», стандартный не-конец — при помощи пунктуационного знака «запятая», см. приведенные выше примеры[6]. ЭДЕ, закрывающиеся запятыми, мы называем *нефинальными* и считаем, что они входят в одно предложение с последующей финальной. Начало нового предложения, как и в кодифицированной письменной речи, отмечается заглавной буквой.

Определение границ предложений, как и определение границ ЭДЕ, представляет собой отдельную аналитическую процедуру. Основной критерий тут — просодический: финальная ЭДЕ предложения-сообщения должна произноситься с падением в главном акценте, причем не с любым падением, а с падением в определенный для конкретного говорящего нижний уровень частоты основного тона (четкую формулировку этого критерия см. в работе Кибрик 2008; о различии между падениями в самый низкий и менее низкий уровни см. также Оде 1995: 206 и сл.)[7]. Данные о паузации тут, как и при делении на ЭДЕ, в расчет не принимаются. Реализация упомянутой процедуры привела к следующим результатам. Было выяснено, что в устном русском нарративе могут встречаться предложения весьма различной длины — от 1 ЭДЕ до 30. При этом есть основания полагать, что имеется зависимость между длиной рассказа и преобладающей в нем моделью построения предложения: чем длиннее рассказ, тем выше в нем вероятность появления длинных предложений, и наоборот (см. Коротаев, Кибрик 2008).

Очевидно, что членение рассказа на предложения — один из способов указать на различия в степени семантической интеграции: ЭДЕ, входящие в одно предложение, связаны между собой теснее, чем ЭДЕ, принадлежащие разным предложениям. Иначе говоря, граница предложения в общем случае представляет собой точку разрыва на некотором уровне дискурсивной и когнитивной связности (ср. употребляемое в этом контексте понятие «суперфокуса сознания» в Chafe 1994: 148). Поэтому неудивительно, что границы предложений, определяемые вне зависимости от наличия или отсутствия паузы, в подавляющем большинстве случаев паузы все же содержат. В исследованном корпусе подобным образом дело обстоит в 86,7 % случаев.

Верно и обратное: пара ЭДЕ, произносимая без пограничной паузы, практически всегда входит в одно предложение — см. приведенные выше примеры (2)–(5). Эта тенденция проявляется на материале корпуса с еще большей четкостью — 90,1 % случаев.

4.3. Синтаксический тип

ЭДЕ 7 и 8 примера (2) представляют собой простые глагольные клаузы. Такие ЭДЕ мы называем *каноническими*, они составляют около 50 процентов всех ЭДЕ корпуса. Среди пар ЭДЕ, произносимых без пограничной паузы, их доля несколько выше — 58,4 %. Интереснее следующее наблюдение. Примерно 60 % этих случаев представляют собой не простое соположение клауз, а жесткую полипредикативную конструкцию определенного типа — сложносочиненную (см. примеры (2), (4)), объектную (3), определительную (5). Если же мы пойдем с противоположной стороны и рассмотрим весь массив сложноподчиненных конструкций в корпусе, окажется, что пограничная пауза наличествует в них лишь в 32,9 % случаев (против, напомним, 62,6 % для всего корпуса). Это еще одно свидетельство того, что отсутствие пограничной паузы указывает на повышенную семантическую интеграцию — в данном случае закрепленную в синтаксической форме (см. также Коротаев 2009). В связи с этим стоит еще раз обратить внимание на пример (5) выше. В нем пауза между определяемым словом и началом определительной клаузы отсутствует даже несмотря на то, что весь контекст насыщен внутренними хезитационными паузами.

Еще один весьма типичный контекст связан с использованием так называемых *регуляторных* ЭДЕ, не содержащих пропозициональной информации, но играющих важную роль в организации и регулировании речевого потока. Такие ЭДЕ состоят из дискурсивных маркеров, самый часто встречающийся из них — акцентированный маркер вот. Функция этой единицы состоит в указании на то, что некоторый отрезок дискурса (ЭДЕ или группа ЭДЕ) завершен и говорящий переходит к следующему (см. Дараган 2003)[8]. Любопытно при этом, что около половины всех примеров с вот (или его вариантом ну вот) устроено по следующей схеме: между предшествующим отрезком и маркером пауза есть (это верно для подавляющего числа случаев, не менее 90 %), а между маркером и последующим отрезком — нет. Таким образом, просодически (ну) вот

---

[6] Отметим, что используемые в транскрипции пунктуационные знаки зачастую имеют другие значения, чем в стандартной письменной пунктуации. За подробностями отсылаем читателя к книге Кибрик, Подлесская (ред.) 2009: 45—388.

[7] Главный, или *несущий*, акцент отмечается в транскрипции подчеркиванием ударной гласной акцентированного слова, направление тона в акценте — иконическими значками /, \ и проч. перед началом слова.

[8] Подчеркнем еще раз, что речь идет только о случаях акцентированного употребления *вот*, составляющего отдельную синтагму. О ряде других значений этой многогранной частицы см., в частности, работу Кобозева 2007.

оказывается ближе к посттексту, а не к посттексту. Именно эта ситуация представлена в следующем примере (строки 10–12):

088n
9.1 8. ••(0.3) и /я была в таких \жёлтых таких /туфл<u>я́</u>х таких,
11.7 9. ••(0.2) с-с такими с \шнур<u>о</u>чка↑ми,
12.9 10. *м<u>а</u>ма купи↑ла*,
13.6 11. •••(0.8) \<u>вот</u>,
14.4 12. и /я пошла по \метр<u>о</u>w.

Отсюда можно сделать вывод, что *вот* используется прежде всего для поддержания дискурсивной связности рассказа — в том случае когда необходимо нейтрализовать эффект затянувшейся паузы и *немедленно* перейти к новому дискурсивному отрезку.

4.4. Длина серии границ ЭДЕ без паузы

В примере (2) отрезок дискурса, произносимый без пограничных пауз, ограничивается двумя ЭДЕ. Это наиболее типичный случай, хотя встречаются и целые серии ЭДЕ без пограничных пауз, см. следующий пример, в котором начальная пауза отсутствует в 9 ЭДЕ подряд (это абсолютный рекорд в корпусе):

091n
28.5 21. ••••(1.6) /\Н<u>у</u> значит,
30.5 22. в \ц<u>е</u>лом,
30.9 23. такая \тётька на меня /<u>по</u>смотре<u>л</u>а,
32.5 24. я /<u>с</u>мотрю,
33.0 25. у неё лишний /гл<u>а</u>з оказался,
34.1 26. и какое-то **–лицо**...
34.9 27. >\варще< \мертвячье!
35.9 28. Я Чингиза /позвал<u>а</u>,
36.9 29. как-то еле-еле /в<u>ы</u>манила,
38.2 30. и \дал<u>а</u> \дёру.

В целом, однако, с каждой новой ЭДЕ вероятность того, что следующая граница будет лишена паузы, все более понижается, см. таблицу 1.

Данные таблицы 1 поддаются достаточно тривиальной интерпретации: поддерживать высокую степень просодической интеграции на таком длительном отрезке, как (7), тяжело как с физиологической точки зрения, так и с точки зрения планирования. И все же разброс возможностей — от двух ЭДЕ без пограничной паузы до десяти подряд — наводит на мысль о том, что отсутствие пограничной паузы, вероятно, стоит рассматривать не как изолированное явление, а как отражение общей просодической стратегии конкретного рассказа и конкретного говорящего. Один из методов выявления такого рода стратегии представлен в следующем разделе.

Таблица 1. Серии границ ЭДЕ без пауз

Число границ без пауз подряд	Всего серий в корпусе	Суммарное число границ без пауз в сериях данного типа
1	534	534
2	198	396
3	72	216
4	29	116
5	8	40
6	2	12
7	2	14
8	1	8
9	1	9
Всего	847	1345

5. Варьирование по рассказам и говорящим

Как уже было отмечено, 62,6 % всех границ ЭДЕ в корпусе сопровождаются паузой. Среднее значение этого параметра по 129 рассказам (назовем его *коэффициентом пограничной паузации рассказа*) составляет 64,2 %, медианое — 63,3 %. Основное ядро рассказов имеет коэффициент пограничной паузации от 50 % до 75 %, однако встречаются и случаи более существенного отклонения от срединных значений.

На одном полюсе находятся рассказы, в которых паузы имеются в каждой границе ЭДЕ. Всего таких рассказов пять, причем ни в одном из них не содержится более 12 ЭДЕ. Однако близкие показатели возможны и для более длинных рассказов. К примеру, рассказ 043z состоит из 52 ЭДЕ, и 43 из 51 границы в нем сопровождаются паузами, т. е. коэффициент пограничной паузации в этом рассказе равен 84,3 %. Всего же в корпусе насчитывается 23 рассказа с крайне высоким, от 80 %, значением этого параметра. Причем, что характерно, авторами 9 из них являются два говорящих. Оба этих информанта все свои рассказы строят по модели, предполагающей высокий коэффициент пограничной паузации: у одного он не опускается ниже 79,3 %, у другого — ниже 82,4 %.

На противоположном полюсе располагаются рассказы с чрезвычайно низким коэффициентом пограничной паузации. Наиболее примечателен тут рассказ 091n, из которого взят пример (8): в нем паузы имеются только в 7 из 30 границ ЭДЕ, т. е. всего лишь в 23,3 % возможного максимума. В целом же подобная (или близкая к ней) модель построения дискурса представлена в корпусе не слишком широко. Всего рассказов с коэффициентом пограничной паузации менее 45 % насчитывается 15. Закономерности, связанные с отдельными говорящими, тут не так бросаются в глаза, как в случае с коэффициентом более 80 %: здесь представлены в основном те рассказчики, от которых записывалось всего по одному рассказу.

6. Заключение

На материале устного корпуса «Рассказы о сновидениях» мы рассмотрели сочетания ЭДЕ, произносимые без пограничной паузы. Это явление, отклоняющееся от модели так называемого идеального речепорождения, на всем массиве корпуса встречается относительно нечасто (37,4 % случаев), однако для определенного типа конструкций такое просодическое оформление оказывается, наоборот, весьма характерным. Речь идет, в первую очередь, о сложноподчиненных конструкциях и конструкциях с дискурсивным маркером вот. В целом же отсутствие пограничной паузы является одним (возможно, самым очевидным) из сигналов повышенной степени семантической интеграции. Об этом свидетельствует не только синтаксический тип конструкций, но также и преимущественное отсутствие наряду с пограничными паузами и внутренних, почти обязательное вхождение последовательностей ЭДЕ без пауз в одно предложение и — в определенной мере — тот факт, что такие последовательности редко содержат более двух ЭДЕ. В числе других просодических сигналов семантической интеграции можно указать, в частности, отсутствие ресета (т. е. сброса частоты основного тона на уровень, стандартный для начала каждой новой ЭДЕ) и произнесение пары ЭДЕ с одним несущим акцентом (см. Коротаев, Кибрик, Подлесская 2009: 317–331).

Для анализа отдельных рассказов было введено понятие коэффициента пограничной паузации — отношения числа границ ЭДЕ с паузами к общему числу границ ЭДЕ в рассказе. Для конкретных рассказов этот параметр варьирует от 23,3 % до 100 %. Как можно объяснить подобный разброс в значении коэффициента пограничной паузации? Нам известны по крайней мере два подхода к анализу неравномерностей появления пауз в дискурсе. Согласно одному из них (см. Erbaugh 1996), удельный вес пауз различен в различных элементах нарративной схемы (макроэпизодах); в другом (см., например, Butterworth, Goldman-Eisler 1979) используется хуже формализуемое понятие когнитивного ритма. Выявление того, каким образом коэффициент пограничной паузации соотносится с этими параметрами дискурсивной структуры, а также с прочими особенностями просодической организации рассказа (например, общим темпом речи) — дело дальнейших исследований.

Литература

1. *Дараган Ю .В.* Паразитизм или симбиоз: механизм преодоления коммуникативных сбоев и обслуживающие его вербальные средства // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2003». М.: Наука, 2003. С. 166–178.
2. *Земская Е. А. (ред.) Русская разговорная речь. М.: Наука, 1973.*
3. *Кибрик А. А.* Есть ли предложение в устной речи? // А. В. Архипов, Л. М. Захаров, А. А. Кибрик и др. (ред.) Фонетика и нефонетика. К 70-летию Сандро В. Кодзасова. М: ЯСК, 2008. С. 104–115.
4. *Кибрик А. А., Подлесская В. И.* (ред.) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК, 2009.
5. *Кобозева И. М.* Полисемия дискурсивных слов и попытка ее разрешения в контексте предложения (на примере слова вот) // Труды международной конференции «Диалог 2007». Бекасово, 2007. С. 250–255.
6. *Коротаев Н. А.* Просодическая организация сложноподчиненных конструкций // Кибрик, Подлесская (ред.) 2009. С. 488–522.
7. *Коротаев Н. А., Кибрик А. А.* Иллокуция сообщения в устных рассказах: опыт корпусного исследования // Труды международной конференции «Корпусная лингвистика — 2008». Санкт-Петербург: СПбГУ, 2008. С. 214–220.
8. *Коротаев Н. А., Кибрик А. А., Подлесская В. И.* Осложнения канонической структуры: на стыке моно- и полипредикативности // Кибрик, Подлесская (ред.) 2009. С. 219–332.
9. *Кривнова О. Ф.* Ритмизация и интонационное членение текста (опыт теоретико-экспериментального исследования). Дис. … д-ра филол. наук. М.: МГУ, 2007.
10. *Оде С.* Интонационная система русского языка в свете данных перцептивного анализа // Л. Л. Касаткин (ред.) Вопросы фонетики II. М.: Институт русского языка РАН. С. 200–215.
11. *Подлесская В. И., Кибрик А. А.* Коррекция в устной монологической речи по данным корпусного исследования // Русский язык в научном освещении, 12–2. 2006. С. 7–55.
12. *Подлесская В. И., Кибрик А. А.* Речевые сбои и затруднения // Кибрик, Подлесская (ред.) 2009. С. 177–218.
13. *Светозарова Н. Д., Вольская Н. Б., Павлова А. В., Шитова Л. Ф.* Просодическая организация русской спонтанной речи // Н. Д. Светозарова (ред.) Фонетика спонтанной речи. Л.: Изд-во Ленинградского университета, 1988. С. 141–182.
14. *Хитина М. В.* Делимитативные признаки устноречевого дискурса. М.: МГЛУ, 2004.
15. *Щерба Л. В.* Фонетика французского языка. М.: Изд-во литературы на иностранных языках, 1955.

16. *Butterworth, B., Goldman-Eisler, F.* Recent studies on cognitive rhythm // A. W. Siegman, S. Feldstein. Of speech and time. Hillsdale: Lawrence Erlbaum, 1979.
17. *Campione, E., Véronis, J.* Pauses and hesitation in French spontaneous speech // DiSS-2005. P. 43–46.
18. *Chafe, W.* Discourse, consciousness, and time. Chicago: University of Chicago Press, 1994. P.
19. *Clark, H. H., Clark, E. V.* 1977. Psychology and Language. N. Y: Harcourt Brace Jovanovich, 1977.
20. *Degand, L., Simon, A. C.* 2005. Minimal Discourse Units: Can we define them, and why should we? // M. Aurnague, M. Bras, A. LeDraoulec, L. Vieu (eds.) Proceedings of SEM-05. Connectors, discourse framing and discourse structure: from corpus-based and experimental analyses to discourse theories. 2005. P. 65–74.
21. *Erbaugh, M. S.* A uniform pause and error strategy for native and non-native speakers // Proceeding of the International Conference on Spoken Language Processing. Philadelphia, 1996.
22. *Levelt, W. J. M.* Speaking. From intention to articulation. Cambridge, Mass.: MIT Press, 1989.
23. *Miller, J., Weinert, R.* Spontaneous spoken language: Syntax and discourse. Oxford: Clarendon Press, 1998.

Паттерны эмоциональных коммуникативных реакций: проблемы создания корпуса и перенос на компьютерных агентов

Patterns of emotional reactions in communication: problems of corpora studies and application to computer agents

Котов А. А. (kotov@harpia.ru)

Российский государственный гуманитарный университет

В докладе приводится архитектура компьютерных агентов, которые имитируют эмоциональное речевое поведение, в частности, изменения настроения во времени. В мультимодальном корпусе (запись экзаменов) мы рассматриваем последовательности контрастных эмоциональных реакций и возможность перенести эти реакции на компьютерных агентов.

Развитие робототехники и компьютерных интерфейсов предъявляет новые требования к исследованию коммуникативного взаимодействия. Роботы или виртуальные компьютерные персонажи должны иметь возможность (а) успешно распознавать намерения и эмоциональное состояние человека-собеседника и (б) при необходимости — адекватно имитировать речевые и невербальные реакции человека в диалоге. Прежде всего, интерес представляют эмоциональные коммуникативные реакции и, в целом, поведение в эмоциональной ситуации. Одна из важных проблем в этой области — это анализ паттернов эмоциональных реакций, распределённых во времени. Человек может «сложно» переживать негативное событие: внутренне расстраиваясь, ругаясь на собеседника, а затем — «приходя в себя» и успокаиваясь. То же относится и к некоторым позитивным реакциям. «Узнавание» таких паттернов в поведении собеседника позволяет нам судить о его эмоциях и характере — именно поэтому данное явление так важно для исследования в связи с разработкой компьютерных агентов и бытовых роботов.

Для детального анализа коммуникативного поведения человека создаются мультимедийные корпуса: поведение актёров или реальных людей записывается на одну или несколько видеокамер и сопровождается лингвистической разметкой. Паттерны поведения, выделенные при анализе этих корпусов, могут переноситься на компьютерных агентов [Rehmand, Andre, 2008], и многие из этих проектов специально ориентированы на дальнейшее применение в области робототехники и создания интерфейсов: здесь возможны разные методологические подходы.

Один подход состоит в том, чтобы предоставить людям возможность взаимодействовать с роботом или друг с другом и автоматически выявлять повторяющиеся коммуникативные паттерны [Campbell, 2008] — в этом случае мы рассматриваем все поступающие данные как «корпус» и автоматически выделяем типы явлений («ярлыки» разметки). Противоположный подход состоит в том, чтобы задать типы явлений, которые должны быть представлены в базе — по этому принципу создаются мультимодальные корпуса эмоций: актёров просят произнести текст с некоторой эмоцией из списка, заданного экспериментатором [Bänziger, Scherer, 2007]. Центральное положение занимают проекты, в которых испытуемые или информанты сняты в реальных или экспериментальных ситуациях, например, корпус HUMAINE [Douglas-Cowie, Cowie et al., 2007]. Достаточно подробный анализ современных проектов мультимодальных корпусов приводится в работах [Cowie, Douglas-Cowie et al., 2005; Martin, Paggio et al., 2008; López, Cearreta et al., 2009]. Кроме того, список эмоциональных мультимодальных баз данных приведён на странице http://emotion-research.net/wiki/Databases

В нашем случае — для исследования паттернов эмоциональных коммуникативных реакций мы собрали видеозаписи студентов, устно сдающих различные задания в рамках зачётов или экзаменов. Студенты были предупреждены о видеосъемке, но (как видно по некоторым признакам) достаточно быстро переставали обращать внимание на камеру.

Сдача экзамена — это сравнительно редкая ситуация; в жизни каждого человека она наступает

не более нескольких десятков раз. Экзамен представляет из себя особый ритуал, элементы которого почти не встречаются в жизни вне учебных заведений. Вместе с тем, экзамен — это ситуация вопросно-ответного взаимодействия, в которой одна сторона даёт задание, а другая должна его выполнить или иным образом показать себя с наилучшей стороны. С этой точки зрения — ситуация экзамена похожа на типовую ситуацию, в которой должен будет действовать интерфейс или бытовой робот. Кроме того, экзамен — это очень эмоциональная ситуация, представляющая конфликт между сильнейшей мотивацией и строгими социальными ограничениями. В этой ситуации мы ожидаем увидеть характерные паттерны проявления эмоциональных состояний. **Мы благодарим всех студентов, согласившихся на видеосъемку для создания корпуса!**

1. Модель речевого поведения эмоциональных компьютерных агентов

Мы разрабатываем модель эмоционального агента, которая, в частности, реализована в виде программы — прототипа для «анимирования» компьютерных персонажей (компьютерных агентов — Рис. 1) [Котов, 2008]. Компьютерный агент должен реагировать на события внешнего мира или обращённые к нему высказывания, меняя свою реакцию в зависимости от типа события и от своего «настроения». Так, агент должен реагировать негативно: (а) если само событие — плохое, или (б) если агент находится в плохом настроении. Агент принимает на вход предикативные структуры: набор из предиката и некоторого количества актантов. Предикат и актанты — это множества признаков, их состав и значение переменны и задаются входящим событием или содержанием поступившего текста. Каждая предикативная структура обрабатывается с помощью набора *сценариев* — отношений типа «если-то». Сценарии используются для имитации реакций агента: для негативных или позитивных эмоциональных реакций используются *д-сценарии*, для «рациональных» реакций — *р-сценарии*. Каждый сценарий обладает переменной активизацией — входящая предикативная структура активизирует в разной степени все сценарии; на активизацию сценариев также влияет текущее «настроение» агента. Сценарий с максимальной активизацией формирует речевой и поведенческий выход модели: агент произносит высказывания, связанные с этим сценарием, и демонстрирует заложенный в базу жест.

Агент может не только отвечать на событие одним высказыванием, но и имитировать изменения в настроении и речи, происходящие в течение 10–40 секунд после события: если мы ‘стукнем’ агента, он будет сначала расстраиваться и ругаться, потом может начать винить себя за невнимательность и, наконец, сможет перейти к «рациональному рассуждению» о том, как избежать подобных ситуаций (имитируется не сам процесс рационального рассуждения, а только характерные речевые реакции). При этом агент как бы «испытывает сложное переживание» и проходит через несколько коротких *микросостояний*, характеризуемых некоторой эмоцией и некоторым единым способом выражения (Рис. 2). Каждое микросостояние связано со множеством сценариев. Если микросостояние активно, агент предпочитает реагировать с помощью этих сценариев. Микросостояния могут активизироваться входящими событиями (через связанные с ними сценарии), причём одно событие может в разной степени активизировать несколько микросостояний. После этого агент начинает реагировать на событие, перебирая несколько самых активных микросостояний, начиная с большего по степени уменьшения.

«Неприемлемые» высказывания могут подавляться модулем фильтрации. Если агент не может прямо выразить в речи выбранные фразы, он может переключаться в коммуникации (обращаться к другому собеседнику) или заменять подавленные высказывания лицемерием или иронией. При «иронии» агент выбирает высказывания, характерные для сценария с наибольшей активизацией и с противоположным микросостоянием — и сопровождает их «смайликом». Если мы ‘стукнем’ агента, он может подавить ответные ругательства и с сарказмом ответить «*Хорошо, что Вы обратили на меня внимание* ☺». При этом агент использует выход позитивного д-сценария ВНИМАНИЕ, который в данной ситуации получает сравнительно слабую активизацию (больше остальных позитивных сценариев, но меньше многих негативных).

Одна из главных задач этого подхода состоит в том, чтобы сделать последовательности коммуникативных реакций агента максимально естественными. Как видно из архитектуры агента — последовательности его реакций не являются фиксированными: микросостояния могут получать разную активизацию и выстраиваться в разном порядке. Однако, пополняя базы жестов, высказываний и других способов реагирования, а также настраивая состав и чувствительность микросостояний, мы можем приблизиться к адекватной картине, когда в эмоциональной ситуации агент будет внешне демонстрировать правдоподобную картину человеческих переживаний. Для этого необходимо исследовать паттерны эмоционального коммуникативного реагирования на реальных случаях поведения людей в эмоциональных ситуациях.

Для описания эмоциональных высказываний агентов (но не для разметки корпуса) мы используем типологию коммуникативных целей, схожую с типологией целей, предложенных Р. Шенком для анализа рассказов историй в дружеских компаниях [Schank, 2000].

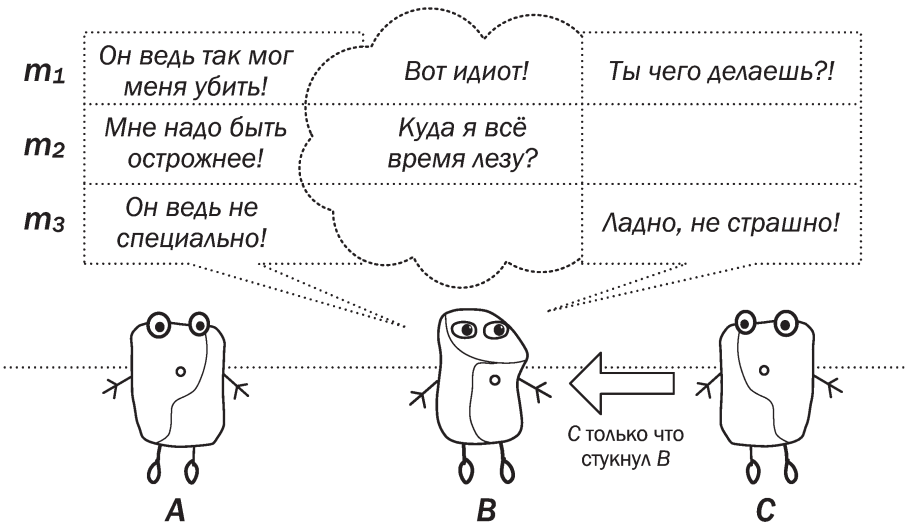


Рис. 1. Имитация речевого поведения агента B с переключениями в коммуникации

Мы имитируем речевое поведение агента ***B*** в ситуации, когда его только что 'стукнул' агент ***C***. Агент ***B*** переживает несколько коротких состояний (m_1-m_3), при этом он думает что-то «про себя», обращается к агенту ***C*** или к присутствующему при этом агенту ***A***.

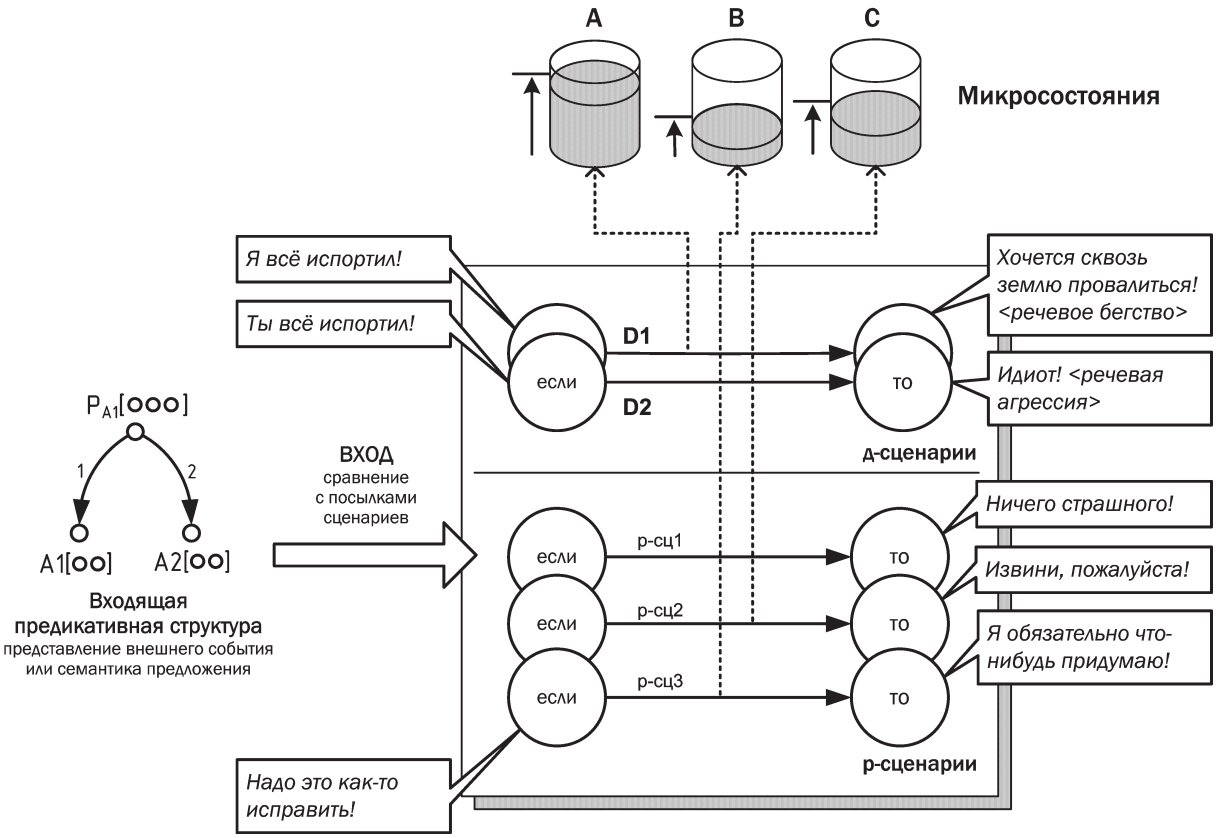


Рис. 2. Устройство эмоционального агента, имитирующего речевую эмоциональную динамику во времени

Входящее событие сравнивается с посылками сценариев и активизирует сценарии, а также связанные с ними микросостояния. Здесь активны оказались микросостояния A-C-B, агент будет переживать их последовательно, что сформирует следующий речевой выход: *Я всё испортил! Хочется сквозь землю провалиться! <пауза> Извини, пожалуйста! <пауза> Надо это как-то исправить! Я обязательно что-нибудь придумаю!*

0-цель («ноль-цель») соответствует высказываниям, вызванным эмоциональным состоянием и плохо поддающимся контролю, например, это ругательства и возгласы радости (эта цель отсутствует в исходной типологии Шенка).

Я-цель соответствует высказываниям, которые вызваны эмоциональным состоянием говорящего, причём говорящий намерен получить дальнейшее удовлетворение, привлекая внимание адресатов. Говорящий может говорить *Я очень умный!*, потому что он ждёт восхищения адресата, или *На улице очень холодно!* — потому что он замерз и хочет поделиться эмоционально-значимым событием.

Преследуя **ты-цель**, говорящий намерен добиться удовлетворения от эффекта на адресата: он может говорить *Я очень умный!* на собеседовании — чтобы его взяли на работу, или *На улице очень холодно!* — чтобы заставить адресата остаться дома.

Наконец **цель коммуникации** вынуждает говорящего к некоторому высказыванию — например, говорящий вынужден ответить на комплимент встречным комплиментом или вынужден что-либо сказать, чтобы оживить увядшую беседу. Эта типология важна для анализа дальнейшего материала.

2. Корпус и проблемы описания паттернов эмоциональных коммуникативных реакций

Для анализа паттернов эмоциональных коммуникативных реакций мы собрали корпус видеофрагментов, где студенты различных факультетов РГГУ защищают перед экзаменатором письменные работы или «отвечают» определения терминов по билетам. На данный момент в корпусе находится 236 фрагментов продолжительностью от 2 минут до получаса.

Для разметки видеофрагментов мы используем программу ELAN, разработанную в Институте психолингвистики им. Макса Планка (Неймеген, Нидерланды).[1]

Основным полем для разметки является шкала времени (timeline), на которой можно определить ряд параллельных «дорожек» для разметки речи, жестов и мимики людей в кадре.

Для разметки фрагментов мы используем 14 дорожек; записываются речь основного героя в кадре, речь основного собеседника и других собеседников, жесты и эмоциональные движения, выполняемые головой, глазами, ртом и руками, движения корпусом или изменения позы; 2 дорожки отведены для разметки острот и иронии, 2 — для разметки микросостояний.

Для речи основного героя, жестов, острот и микросостояний мы используем по 2 дорожки — на 1-й дорожке записываются действия/высказывания коммуниканта, а на второй дорожке — фазы этих действий. Для устной речи «фазами» считаются междометия, паузы хезитации (сомнения) и поправки — то есть речевые элементы, нарушающие «нейтральный» плавный характер речи. Для жестов в качестве фаз размечаются паузы при выполнении жеста, «зацикливание», в существенных случаях — фазы экскурсии и рекурсии жеста; для микросостояний и острот — на 2-й дорожке при возможности размечаются составные элементы: конкретные микросостояния или компоненты остроты.

В материале корпуса информанты сталкиваются с различными эмоциональными ситуациями: в первую очередь, это «сложные» вопросы экзаменатора и указания экзаменатора на ошибки в ответе. В этих эмоциональных ситуациях информанты могут демонстрировать различные речевые и поведенческие реакции. При анализе примеров из корпуса нас, в первую очередь, интересовали такие случаи, где информанты в качестве реакции демонстрируют ряд противоположных жестов или речевых реакций. Такие последовательности могут иметь разную природу — они или указывают на сложное выражение одного эмоционального состояния, либо являются следствием быстрой смены эмоциональных состояний (0-цели) или речевых стратегий (ты-цели).

2.1. Множественные жесты и действия

Мы ожидали, что в эмоциональной ситуации информанты могут демонстрировать эмоциональный жест, например, хвататься за голову или стучать по столу. Предыдущая версия используемой нами компьютерной модели предоставляла агенту возможность при наступлении некоторого микросостояния продемонстрировать один жест. Тем не менее, в материале корпуса можно найти достаточно много фрагментов, когда говорящий демонстрирует в качестве реакции сразу несколько последовательных жестов, соответствующих одному микросостоянию. Причём множество из этих жестов (или форм поведения) традиционно не рассматривались как формы коммуникативного взаимодействия. Рассмотрим один из наиболее показательных примеров.

(1) ***А*** (жен.) отвечает значение термина «антономазия»; до начала этого фрагмента она думает над вопросом 35 секунд; данный фрагмент имеет продолжительность 58 секунд; продолжительность каждого действия в коммуникации размечена в миллисекундах.[2]

[1] Программа доступна в рамках лицензии GNU по адресу: http://www.lat-mpi.eu/tools/elan/

[2] Разметка в миллисекундах позволяет оценить продолжительность пауз или, наоборот, беглый характер диалога (в том случае, когда в записанном виде он обогащён деталями и выглядит громоздко).

[5800] ***А*** (прикладывает палец левой руки к крылу носа): *Не когда выпадение одного или нескольких звуков?… Например=*
[2035] ***Б***: *Антономазия?*
[1616] ***А*** (перекладывает палец к виску): *Да, я пытаюсь вспомнить.* ***Б*** (одновременно с *А*): *Нет.*
[1489] ***Б***: *Это не выпадение звуков.*
[23220] пауза; ***А***, приложив левую руку к голове, смотрит вбок, в билет и потом — вниз.
[2696] ***А*** (левой рукой демонстрирует жест, как если бы держала в руке шар): *Неправильная грамматическая организация предложения?*
[1984] пауза
[1160] ***Б***: *Антономазия? Нет.*
[1152] ***А*** смеётся и откидывается на стуле. **Б**: *Опять не попали.*
[1649] ***А*** поправляет волосы.
[1144] ***А*** пододвигает свой стул ближе к столу.
[1415] ***А*** поправляет кофту: повернув руки большими пальцами к адресату, одёргивает кофту вниз по бокам.
[1766] ***А*** «потягивается» — подняв и сжав плечи, вытягивает вниз обе руки.
[1274] пауза; ***А*** наклоняется к столу и смотрит на билет.
[2400] ***А*** поправляет волосы.
[1422] ***А***: *У меня есть ещё два варианта.*
[2338] ***А*** смеётся, двигая при этом корпусом.
[3640] ***А*** прикладывает указательный палец левой руки к носу, а большой — к подбородку и продолжает смеяться, раскачиваясь при этом корпусом.

По этому примеру видно, что после указания на неверный ответ ***А*** поправляет волосы, одежду, двигает стул и меняет положение тела, то есть демонстрирует по меньшей мере шесть различных действий, прежде чем ответить речевым высказыванием. При этом само высказывание является ироничным и сопровождается дальнейшими жестами. Первичный анализ корпуса демонстрирует, что ряд вполне нейтральных жестов обнаруживают тесную связь с эмоциональными ситуациями: некоторые информанты регулярно поправляют одежду (точно как в примере 1), поправляют волосы или слегка облизываются. Например, информант С07-17.7.2008 (жен.) во время защиты письменных работ перед комиссией в течение 9 минут 37 раз слегка высовывает язык, облизывая губы, и 7 раз манипулирует волосами, перемещая заколотый «хвост» на плечо или убирая за спину.

2.2. Рудиментарное проявление негативных эмоций и ирония

Агрессия или гнев — это эмоции, неприемлемые на экзамене. Тем не менее, в базе мы обнаружили целый ряд примеров, когда информанты демонстрируют жесты или элементы поведения, соответствующие целому ряду эмоциональных состояний: «агрессии», «гневу», «переживанию боли», «истерическому поведению» или «детскому капризу». Интересно то, что эти жесты имеют не изолированный характер, а организованы в общие синтагмы с речевыми фрагментами и другими жестами, в частности, в большинстве случаев в стадии завершения (рекурсии) они смягчаются улыбкой или переходят в смех.

Поведение, сходное с элементами истерического поведения, представлено в следующем примере:

(2) ***А*** (жен.) отвечает значение термина «антономазия»; до этого она 2,5 минуты предлагала разные (неверные) варианты определения или держала паузу.
[2300] ***А***: *А не это — «глух» - «глуп»?*
[1800] пауза
[1140] ***Б***: *Это не антономазия.*
[1360] ***А*** содрогаясь корпусом и тряся руками откидывается вперёд, смеётся (или демонстрирует дыхательные спазмы): *Да что ж такое-то?*
[589] ***А***, продолжая смеяться, делает обеими руками такие жесты, как если бы снимала со своей одежды и выбрасывала волосы или нитки.
[2811] ***А***, продолжая улыбаться, поправляет волосы.
[871] ***А*** демонстрирует мимику осуждения: растягивает губы, поднимает брови, расширяет глаза и слегка поворачивает головой в разные стороны.
[6809] пауза; ***А*** двигает нижней челюстью, смотрит на билет и на экзаменатора.
[далее ***А*** просит разрешить посмотреть, в каком месте списка находится термин]

В этом примере высказывание *Да что ж такое-то?* являлось симптомом эмоционального состояния и преследовало 0-цель. То же самое или сходные высказывания могут использоваться в других контекстах для достижения ты-цели — демонстрируя наш гнев и осуждение (возможно — мнимые), мы стремимся изменить поведение адресата. Рудиментарное проявление такого гнева или осуждения присутствует в следующем примере:

(3) ***А*** (жен.) сдаёт определение «метафоры». Фрагмент 26,8 с.
[1121] ***Б***: *Приведите пример метафоры какой-нибудь.*
[5479] ***А*** (поворачивает голову в сторону, взгляд — вверх, поворачивает головой из стороны в сторону как при возмущении или осуждении, говорит еле слышно): *Боже мой!* <шумно выдыхает>
[801] ***А*** (показывает пальцем в сторону собеседника): *Давайте лучше=*
[1139] ***Б***: *На лекции* [*их*] *были десятки.*

[3180] ***А*** (говорит со значительными перепадами тона): *Ну да, так не могу сейчас вспомнить!*
[1180] ***Б***: *А что делать?* [*У нас ведь –*] *зачёт!*
[5120] ***А*** наклоняет голову чуть вбок, прямо смотрит на адресата, демонстрируя осуждающее выражение лица, губы сжаты и несколько растянуты в улыбку; два раза шумно выдыхает (даже сдувая со стола лист бумаги);
[6900] ***А*** поворачивает голову вбок и вверх, размышляя над ответом;
[1840] ***А*** (поворачивает голову обратно к собеседнику): *Какие-нибудь там «смешные зайчики».*

Мы считаем, что в данном примере присутствует сложное взаимодействие между коммуникативными целями. В результате сложного вопроса информант испытывает эмоциональное состояние и проявляет его в речи — это соответствует 0-цели. Однако информант использует форму выражения, обычно преследующую ты-цель. Очевидно, информант не хочет «призвать адресата к порядку» или «осудить адресата» — то есть ты-цель в данном случае в полной мере не присутствует. Таким образом, переживая эмоциональное состояние и будучи вынужденным его выразить (0-цель), говорящий выбирает средства, обычно преследующие ты-цель, возможно, поскольку эти формы выражения для говорящего более освоены в других ситуациях.

В следующем примере выражение рудиментарной агрессии вызвано 0-целью: коротким эмоциональным состоянием говорящего. Вместе с тем, это выражение носит игровой характер и для него нельзя определить объект агрессии — нельзя однозначно утверждать, что агрессия направлена на собеседника или на самого говорящего.

(4) ***А*** (жен.) в очередной раз сдаёт определения терминов. ***А*** объясняет различие в типах метафоры — in praesentia и in absentia (по [Дюбуа, Пир et al., 1986]); фрагмент — 22 с.
[660] ***Б*** (подтверждая предыдущий фрагмент ответа ***А***): *Да.*
[5000] ***А***: *Ну!!! In praesentia — это когда цель есть, in absentia — это когда* [*области*] *цели нет.*
[420[3]] ***Б***: *Как же* [*области*] *цели нет, а перенесение* [*признаков*] *есть?*
[4289] ***А*** (кладёт руки на колени; зажмуривает глаза, выпрямляет корпус, смотрит в сторону, трясётся и фарингально рычит; в рекурсии — подобие улыбки): *Опять-опять. Я не знаю!*
[4771] ***А*** смотрит в билет, чешет затылок, несколько раз пробует начать говорить.
[2480] ***А*** (смотрит в билет): *Нет, цель-то наверно есть, но она скрытая!*
[3019] ***А*** пристально смотрит на адресата, ожидая его реакции.
[1321] ***А*** (смотрит в билет, повторяет тихо и очень спокойно): *Скрытая цель.*

В этом примере форма выражения агрессии более прототипична, однако информант использует специальные средства, чтобы сделать проявление агрессии «игровым»: он отворачивается от адресата и демонстрирует подобие улыбки. По контексту видно, что информант далее быстро переключается на рациональное рассуждение и даёт верный ответ, то есть в полной мере не испытывает состояние агрессии.

Схожим образом в примерах корпуса информанты имитируют переживания боли или детское капризное поведение. При имитации боли испытуемые сжимают губы, морщат лоб и «скулят», или зажмуриваются и обнажают зубы (иногда — шумно вдыхая сквозь зубы). Информанты отвечают этой формой поведения на сложный вопрос, однако после «имитации боли» следует рациональное рассуждение информанта — он перебирает разные варианты ответа и часто правильно выполняет задание. В одном случае информант демонстрирует имитацию боли в конце интервала размышления (длиной 18 секунд) и непосредственно перед правильным ответом (см. таблицу).

При имитации детского каприза информанты «хныкают», что сопровождается резким движением тела — подбородок движется назад (голова опускается), корпус — вперёд, руки — в стороны. Это действие предшествует просьбе: информант просит смягчить задание или сориентировать его в вопросе. Интересно то, что эта форма поведения также сопровождается улыбкой.

По-видимому, эти рудиментарные проявления эмоций обладают двойственной функцией: с одной стороны, говорящий действительно испытывает отрицательную эмоцию (некоторое подобие агрессии или боли), с другой стороны, выбирая способ выражения этой эмоции, говорящий маркирует свой жест улыбкой — это не даёт собеседнику права обижаться и отвечать на «несерьёзное» внешнее проявление эмоции, но при этом сравнительно хорошо обозначает внутреннее состояние говорящего.

Применительно к архитектуре агентов это означает, что даже при рациональной реакции на инструкцию человека, агент может достаточно свободно выражать негативные эмоции (точнее — использовать средства выражения для негативных сценариев, получивших максимальную активизацию при анализе входа), при этом улыбаясь или применяя другие маркеры иронии или игрового поведения. Проявления «крайних» форм эмоций, смягчённые улыбкой, служат для того, чтобы показать более спокойную реальную эмоцию говорящего.

[3] Здесь отмечено время отступа до начала следующей фразы.

Описание поведения; высказывание	Возможная функция
[2314] ***А*** (улыбаясь, размахивая одной и двумя руками перед корпусом): *Можно я к Вам буквально через пять минут=*	**эмоц.:** ***А*** намерена спровоцировать симпатию; **рац.:** ***А*** намерена договориться ещё раз сдать задание;
[1612] ***А*** (наклоняется к столу; продолжает размахивать руками, при этом размахивает у висков; иконически демонстрируя собственную неадекватность): *=подойду? У меня просто это.*	**эмоц.:** ***А*** намерена спровоцировать снисхождение; **рац.:** ***А*** намерена сгладить негативное впечатление от своей неудачной попытки;
[8580] ***А*** (садится ровно, сопровождает свою речь иконическими жестами левой рукой): *Я просто хотела с «ассонансом»… просто не «нагнетение», а какое-то другое слово подобрать? Да? Про= Ну, я имею в виду…*	***А*** старается действовать рационально; **рац.:** ***А*** уточняет задание, возможно, пытается сократить задание при следующей попытке ответить;
[4680] ***А*** (сидит ровно; двумя руками, повернув ладони к себе делает энергичные махи от корпуса в направлении адресата; говорит «холодным» голосом): *В= Всё= В чём… Я просто не совсем понимаю, в чём как бы вопрос?*	**эмоц./рац.:** ***А*** выдвигает претензию, обвиняет адресата в неточности задания;
[893] ***А*** (наклоняется к столу): <смеётся> *Ну, в смысле=*	**эмоц./рац.:** ***А*** стремится снизить негативный эффект от своих предшествующих слов;

2.3. Координация стратегий воздействия при достижении ты-цели

В ситуации экзамена информант может предпринимать множество рациональных (или ритуальных) действий: просить ответить на вопросы билета в другом порядке, просить ещё об одной попытке сдать другой билет (в нашем случае это допускалось процедурой), выдвигать претензии к экзаменатору в некорректном ведении экзамена и т. д. Эти действия рациональны, но в коммуникации на каждое из них может накладываться разная эмоциональная роль: информант может просить ещё об одной попытке ответить, потому что (а) у него есть право это сделать, и он требует уважать это его право, (б) он очень милый и вызвал симпатию экзаменатора, (в) он плохо себя чувствует и провоцирует снисхождение и т. д. Таким образом, коммуникативная ситуация является двухуровневой: на одном уровне коммуникант обсуждает рациональные шаги, а на другом — предъявляет для себя эмоциональную позицию (маску) в обоснование этих шагов. Такая эмоциональная роль обычно используется для достижения ты-цели: говорящий стремится получить преимущество в эмоциональном взаимодействии, чтобы достичь рациональной цели. Для достижения успеха говорящий может подряд перебирать несколько стратегий воздействия на адресата.

(5) ***А*** (жен.) ранее пыталась определить «ассонанс» как «нагнетение гласных» — экзаменатор попросил уточнить ответ. ***А*** 78 секунд держит паузу — при этом нагибается над столом, чешет переносицу, прикладывает палец к губам, откидывается на стуле, смотрит по сторонам, потом опять смотрит в билет. Фрагмент 18 с.

В этом фрагменте видно, что говорящий очень часто меняет стратегию, с помощью которой он пытается воздействовать на адресата — на каждую стратегию приходится в среднем 3,6 секунды времени. Из-за этого речевое поведение может показаться сбивчивыми. Такой пример может демонстрировать более общую ситуацию, когда в арсенале говорящего имеется несколько стратегий достижения успеха в коммуникации, и при неудаче одной стратегии говорящий обращается к следующей стратегии. Однако частая перемена стратегий говорит о том, что этому выбору в данном случае недостаёт координации. При создании компьютерного агента мы можем учитывать эту особенность, заставляя его быстро менять стратегии эмоционального взаимодействия (при ориентации на ты-цель), тем самым имитируя усталость в ситуации напряжения.

3. Заключение

Работа с корпусом позволила сформулировать следующие принципы речевого поведения для реализации в компьютерных агентах.

1. В результате переживания эмоционального состояния человек может демонстрировать ряд последовательных жестов или движений (прим. 1). Человек может слегка облизываться, поправлять одежду, потягиваться, манипулировать каким-нибудь объектом и т. д. Эти движения вызваны эмоциональным состоянием, но не являются коммуникативными жестами в полном понимании: адресат может их даже не замечать, но в целом, поведение, сформиро-

ванное набором этих действий, может вызывать у адресата определённое впечатление об эмоциональном состоянии оппонента. Ранее агенты демонстрировали только один жест — этот жест сопровождал фразу или микросостояние. Сейчас в агентов добавлена возможность демонстрировать последовательности жестов и действий при активизации определённого микросостояния.

2. Даже при вполне формальной коммуникации человек может демонстрировать рудиментарные проявления агрессии, гнева или боли (прим. 2, 3, 4). Эти знаки по природе двойственны: с одной стороны, они отражают внутреннее состояние человека, с другой стороны, они не являются «полнозначным» выражением негативных эмоций — поэтому в рекурсии они маркируются улыбкой. Для агентов это означает возможность более открыто выражать негативные сценарии в речи; выражение этих сценариев должно сопровождаться «двойственной» улыбкой или другими маркерами иронии.
3. Коммуникация может предполагать формальное взаимодействие — формальный ответ на вопрос или точное выполнение инструкции. Однако, если со стороны агента инструкция не может быть выполнена, агент может переживать из-за неуспеха до начала своих действий (демонстрировать высказывания, преследующие 0-цели), затем — пытаться рационально выполнить инструкцию, а затем — пытаться воздействовать на адресата, чтобы вызвать положительное впечатление (демонстрировать высказывания, преследующие ты-цели). Эта последовательность действий видна в примере (5). В других случаях некоторые части этой последовательности могут быть опущены.
4. Если агент должен выполнить некоторые действия для адресата, то успешность взаимодействия зависит не от успешности выполнения действий, а от итогового удовлетворения адресата. Если агент не может выполнить инструкцию, он может использовать целый ряд стратегий (преследующих ты-цели), чтобы воздействовать на адресата и вызвать его удовлетворение. Если одна из стратегий «проваливается», агент может переходить к другой стратегии. Этот механизм перехода и выбора стратегий может давать сбой в напряжённой ситуации: стратегии будут быстро меняться без должной координации (прим. 5) — внешне это может служить симптомом напряжения и усталости агента.

Литература

1. *Дюбуа Ж., Пир Ф., Тринон А.* Общая риторика.- М.: Прогресс, 1986.
2. *Котов А. А.* Управление динамикой речевого поведения виртуальных компьютерных агентов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 7 (14). — М.: РГГУ, 2008. — С. 241–247.
3. *Bänziger T., Scherer K. R.* Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus // ACII 2007, LNCS 4738 / Ed. A. Paiva, R. Prada, R. W. Picard. — Berlin, Heidelberg: Springer-Verlag, 2007. — С. 476–487.
4. *Campbell N.* Technology and Techniques for Talking Together // Третья международная конференция по когнитивной науке: Тезисы докладов. — М., 2008. — С. 533–534.
5. *Cowie R., Douglas-Cowie E., Cox C.* Beyond emotion archetypes: Databases for emotion modelling using neural networks // Journal of Neural Networks. — 2005. — 18. — С. 371–388.
6. *Douglas-Cowie E., Cowie R., Sneddon I.* et al. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data // ACII 2007, LNCS 4738 / Ed. A. Paiva, R. Prada, R. W. Picard. — Berlin, Heidelberg: Springer-Verlag, 2007. — С. 488–500.
7. *López J. M., Cearreta I., Garay-Vitoria N.* et al. A Methodological Approach for Building Multimodal Acted Affective Databases // Engineering the User Interface. — London: Springer-Verlag, 2009. — С. 1–17.
8. *Martin J.-C., Paggio P., Kuehnlein P.* et al. Introduction to the special issue on multimodal corpora for modeling human multimodal behavior // Language Resources & Evaluation. — 2008. — 42. — С. 253–264.
9. *Rehmand M., Andre E.* From Annotated Multimodal Corpora to Simulated Human-Like Behaviors // Modeling Communication, LNAI 4930 / Ed. I. Wachsmuth, G. Knoblich. — Berlin, Heidelberg: Springer-Verlag, 2008. — С. 1–17.
10. *Schank R. C.* Tell me a story: narrative and intelligence. — Evanston, Illinois: Northwestern University Press, 2000 (1990).

Непосредственный институциональный диалог. Опыт прямой линии с президентом В. В. Путиным. Дискурсивные стратегии

Citizen-institution non-mediated dialogue: the Russian direct line case

Котта Рамузино П. (paola.cottaramusino@unimi.it)

Миланский государственный университет, Милан, Италия

Прямая линия изучается в настоящей работе как своеобразный пример особого вида институционального дискурса. В частности, рассматриваются дискурсивные стратегии, применяемые «слабой» стороной данного вида коммуникации, т. е. гражданином/телезрителем.

1. Прямая Линия, введенная в 2001 году бывшим Президентом РФ В.В. Путиным в качестве нового и своеобразного средства политической коммуникации, привлекала и привлекает к себе внимание лингвистов. Представляя собой «метаморфозу речевого жанра интервью» (Паршина 2007: 31), являясь дистанционным, публичным диалогом между телезрителями и представителем власти, она приобрела за последние годы особую физиономию и жанровую структуру (там же: 32–33). Ее также можно включить в теоретические рамки т.н. институционального дискурса (institutional discourse) (Drew Heritage 1992), несмотря на то, что в ней сосуществуют неоднородные элементы: диалогический жанр является некоторым образом «неполным», так как у телезрителя нет реплики, как это происходит в обычных интервью. Из жанра интервью, с другой стороны, к ней перешли некоторые стратегии, или тактики, межличностных отношений (к примеру, прагматическая категория вежливости), сосуществующие с риторическим развертыванием дискурса, более характерным для политического дискурса монологического типа. Институциональный дискурс, по своей природе целенаправленный, организован вокруг центрального задания, характеризующего его и определяющего его основные характеристики. Как уже отмечено в литературе (Drew Heritage 1992), участники разных видов институциональной коммуникации проявляют метапрагматическую компетенцию, т.е. умение «вести себя» в соответствии с данной коммуникативной ситуации, хотя не обязательно в качестве высококвалифицированных участников. Они умеют, по словам Ауэра (1992: 26), включаться в коммуникативную обстановку согласно «default assignment of context», или, используя термин Гоффмана, в нужный «фрейм» (т. е. знание участников коммуникации о том, что они делают или должны делать в определенной коммуникативной ситуации). При этом можно упомянуть о том, что отличает институциональный дискурс от обыденного дискурса, как не раз отмечали представители конверсационного анализа на основе исследований таких видов интеракции, как беседы «врач/пациент», «начальник/подчиненный»: не обязательно (и как правило редко) все участники институционального дискурса преследуют те же цели (Caffi 2000: 171).

1.2. Что касается Прямой Линии, исследователи склонны, с одной стороны, выявлять черты демократизации, в ней якобы проявляющиеся (Колокольцева 2004), а с другой стороны внимательно указывают на другую характеристику, частично противоречащую первой, т. е. на иерархическое неравновесие, асимметричность между участниками коммуникации (Паршина 2007, Колокольцева 2004). В преобладающем большинстве институциональных дискурсов проявляется прямая связь между «статусом» и ролью участников, с одной стороны, и их дискурсивными «правами и обязанностями» с другой (Drew Heritage 1992: 48–49). В такой коммуникативной обстановке, чаще чем в других контекстах, важную роль играют два из трех факторов, в свое время указанных Брауном и Левинсоном (1987), как угрожающих «лицу»[1] собеседника, т. е. социальное расстояние между Говорящим и Слушателем, и власть одного над другим. Равным образом,

[1] Понимаем «лицо» в значении Брауна и Левинсона (1987: 61) “the public self-image that every member wants to claim for himself, consisting in two related aspects: negative face (…) and positive face (…)

Линелл и Лакманн (Linell e Luckmann 1991), измеряют асимметричность коммуникативной обстановки, основываясь на четырех параметрах, которые ими определены как количественное, интеракциональное, семантическое и стратегическое доминирование, и по этим же параметрам, статус телезрителя в ПЛ является, хотя не всегда, в основном слабым[2].

1.3 Изучению таких "unequal encounters" (неравных встреч) посвящены, в частности, работы Томас (Thomas 1985, 1989), в которых удачно описаны те "pragmatic features [which] were striking in the speech of all the dominant participants and entirely absent from the speech of the subordinate participants" (Thomas 1985: 767). Довольно-таки часто оставались вне внимания исследователей «слабые» участники коммуникации, т.е. те участники, которые имеют меньше власти, в силу чего в их распоряжении меньше "metapragmatic acts" (*там же*). Это непрофессиональные участники коммуникативной интеракции, опосредованной, и это тоже важно, таким мощным средством массовой информации, как телевидение.

1.4 Наша цель — исследовать в рамках данного вида институционального дискурса роль «непрофессионального», наивного, интервьюера. Наивного не только потому, что российские граждане, которые задают вопросы Президенту, скорее всего не знают и не умеют пользоваться стратегиями «нейтральности», употребляемыми профессионалами (Clayman 1992), но еще и потому что мы будем рассматривать лишь первые выпуски ПЛ, за 2002 и 2003 гг.[3]: цельный и сравнительный анализ всех существующих выпусков последующих лет, показывает, как при укреплении жанра ПЛ, не только ответы приобрели стандартную структуру (Паршина 2007: 34), но и вопросы стали более стандартными и предсказуемыми, ограничивая тем самым индивидуальные коммуникативные стратегии.

2. По мнению Томас (1985: 766), мы должны учесть, что «the power relationship obtaining between the participants in an interaction and the institutional norms within which that interaction takes place are central to the way in which the discourse is developed and individual utterances interpreted». Неоднократно исследователи замечали, как, при «неравных коммуникативных интеракциях» сильный участник располагает целым рядом стратегий, при помощи которых «the dominant participant effectively denies his/her interlocutor the possibility of escaping into indirectness and 'pragmatic ambivalence'» (Thomas 1985: 767). Нас интересует «слабый» участник, а, именно то, каким образом он решает, почти не осознанно, острый вопрос об отношении к власти, который лежит в основе институциональной коммуникации и предопределяет коммуникативное поведение участников.

Вообще, при такой асимметричности, когда ярко проявляется социальное расстояние и иерархическая подчиненность одного коммуниканта, ожидается почти без исключений, что слабый говорящий осуществляет свои речевые акты *off record* (Brown e Levinson 1987: 17–21), т.е. косвенно, двусмысленно, чтобы минимизировать последствия собственных речевых поступков[4], хотя данное речевое поведение предполагает искусного говорящего, умеющего пользоваться целым рядом риторических и стилистических фигур. Как увидим, высказываясь, наоборот, *on record*[5], Г в ПЛ в любом случае находит разные способы ослабленного выражения критики, осуждения и т.д., смягчая тем самым акт, угрожающий «лицу» собеседника. Лишь в редких и особых случаях он выражается эксплицитно.

Примеры из ПЛ будем рассматривать в следующих секциях, учитывая уровень эксплицитности выражения вопросов (от полностью *on record* стратегий до протиположной крайности).

2.1. Вопросы *on record*, без митигации.

В исследованных прямых линиях были обнаружены лишь три случая такого рода и, как увидим, они скорее всего, связаны со средством коммуникации.

В первом примере (1), крайне прямой вопрос, завершающий длинное введение о личном опыте телезрителя и вообще о пенсионной ситуации сограждан[6], является явной критикой и косвенным директивным актом (Вы не должны были подписывать):

---

[2] Согласно этой классификации, интервьюер ПЛ располагает частичным семантическим доминированием, так как он выбирает тему вопроса.

[3] Как известно, в сети имеются в принципе все версии, печатные и видео, Прямых Линий. К моему большому удивлению, оказалось, что отсутствуют (они якобы существуют, но нельзя их открыть по техническим причинам), именно первые видеозаписи. Несмотря на многочисленные попытки и на обращение к вебмастеру, не удалось их посмотреть. Естественно данное условие не позволяет утверждать, что стенограммы этих выпусков полностью соответствуют оригиальным речевым формулировкам выступающих.

[4] "By going off record (…) a speaker can profit in the following ways: he can get credit for being tactful, non coercive and he can avoid responsability for the potentially face-damaging interpretation" (Brown Levinson 1987:71).

[5] Вообще "to go on record" это когда Г совершает речевой акт, чья коммуникативная интенция не двусмысленная, ясная всем (*там же*: 68); в данном случае Г имеет перед собой две возможные стратегии: он может выступать полностью эксплицитно ("without redressive action, baldly") или компенсируя угрозу другим способом ("with redressive action") (*там же*: 69).

[6] "Мне 53 года, скоро пенсионный возраст. И глядя на жизнь пенсионеров, на ту пенсию, которую они получают, начинаешь задумываться, что же делать. Два года назад я заключил договор с негосударственным пенсионным фондом и теперь ежемесячно из своей заработной платы выплачиваю данному фонду. И в это время Государственная Дума принимает закон о пенсиях депутатам и бывшим депутатам. Размер этой пенсии составляет 16 тысяч рублей, что несоизмеримо с пенсией, которую получают сейчас пенсионеры по России».

(*1*) *Вопрос:* почему*, Владимир Владимирович,* Вы подписали данный закон*? (ПЛ 2003)*

В следующем примере, благодаря анонимности, гарантируемой самым средством (Интернет) и защищающей его от последствий (ответ и упрек Президента[7]), Говорящий позволяет себе вопрос полностью *on record*:

(2) *Слышали, что его скоро снова переименуют в Сталинград. Вы что там, все …* (Дальше слово, которое я не могу произнести в прямом эфире, я найду более мягкий аналог[8]) *Вы что там, все с ума посходили?* (po internetu, ПЛ 2002)

В третьем примере резко отрицательный оценочный элемент является преобладающим благодаря пресуппозиции:

(3) *Владимир Владимирович, когда наша страна начнет относиться к проблемам инвалидов так же,* как в цивилизованных странах*?* Спасибо. (ПЛ 2002, телефонный центр)

Анонимность защищает и следующего гражданина, который осуществляет открытый директивный акт, смягчая его лишь частично вводным риторическим вопросом (*как Вы считаете*):

(4) *Я хочу Вас спросить,* как Вы считаете*,* не пора ли заканчивать с приватизацией и начинать национализацию*?* (ПЛ 2003, телефонный центр)

2.2 Другая стратегия осуществляется при выражении прямых вопросов, являющихся речевыми актами осуждения, чей оценочный компонент ослабляется при помощи переноса ответственности с собеседника на другое, третье лицо (или на государство).

(5) *Объясните мне, пожалуйста, почему* государство так много говорит и так и не решило проблему *ветхого жилья?* (ПЛ 2003)

Такая же стратегия повторяется в следующем примере, в котором осуждение смягчается тем, что ответственность перекладывается на другого политика, освобождая от нее непосредственного собеседника, не ставя его в тупик:

(6) *Теперь как* наше родное государство *позволяет делать такие дела?(…) Прошу:* как быть с реформой Чубайса? К чему это приведет*?* (ПЛ 2002)

В примере (7) еще больше усиливается оценочный элемент: ответственность за распродажу, воровство госсобственности (обозначаемые инкапсулатором «это», анафорически отсылающим к общему фонду знаний участников коммуникации), возлагается на неопределенное местоимение «кто-то» (которое приобретает в сочетании с наречием *дешево* еще и пренебрежительную коннотацию)

(7) *Как* это *проходило? Как* это *без денег,* дешево кто-то скупил*, и куда* это *все девалось? Теперь как наше родное государство позволяет делать такие дела?* (ПЛ 2002)

В последнем примере осуждение скрывается под просьбой о помощи, угроза лицу собеседника смягчается частично употреблением местоимением «мы» (Мы-государство, мы сибирские люди, мы-родина)

(8) *Вот скажите, пожалуйста,* как нам быть*?* Как нам дальше жить*? Я Вас приглашаю в гости на нашу сибирскую землю* (ПЛ 2002)

Обращаем внимание на употребление личных и притяжательных местоимений (мы, наш/е) в данных высказываниях, явно направленное на создание общего пространства, то же самое отмечается и в следующем примере:

(9) *Владимир Владимирович,* мы хотим быть нужными *России.* Мы *хотим жить и работать в* нашем *городе* (ПЛ 2003)

В примере (10) критика касается не самого Президента, а государственных органов: пропозициональное содержание основывается на пресуппозиции общих знаний, за счет которой приобретает сильно оценочный характер, но опять же говорящий, хотя не «лицом к лицу», а по интернету, снимает ответственность со своего собеседника, давая ему возможность выйти из ситуации (Обладаете ли Вы полной информацией…)

(10) *ОВД, ГАИ, эти службы* явно или скрытно занимаются бизнесом. *Безопасность гарантируется только для «служащих» этих структур.* Обладаете ли Вы полной информацией *об истинном положении дел в этих структурах?* (ПЛ 2003, по Интернету А. Н. из Мурманской области)

Ослабление прямой критики осуществляется еще при помощи перенесения ответственности за сказанное на третье лицо (синтаксически реализуется неопределенно-личными предложениями (11)), или на общепризнанный, авторитетный, хотя совершенно не определенный, источник (12):

(11) Говорят*, что в Российской армии генералов в два раза больше, чем в Советской.* (ПЛ 2002)

---

[7] "Я бы тоже мог употребить слова, которые нельзя использовать в эфире. Очень прошу пользоваться той лексикой, которая употребима в режиме той работы, которую мы сейчас с вами вместе проводим».

[8] Данное замечание журналиста напоминает нам о его роли «режиссера» в изучаемом виде коммуникации: это именно журналисты на месте, которые давая слово, выбирая вопросы, предопределяют ход диалога.

(12) *Довольно-таки часто мы слышим* с экранов телевизоров *о том, что сотрудники милиции у нас бывают связаны с криминальной деятельностью.* (ПЛ 2002)

2.3 Иной раз говорящий прибегает к модализации эпистемической задачи (Caffi 2000: 451), оперируя модальными категориями при помощи формул, выражающих сомнение, снимающих таким образом ответственность с адресата (13), или лексически — употребляя модальные частицы (пр. 14–18), или еще при помощи субъективизации пропозиционального содержания (19–21); как показывают примеры (20–21) данное ограничение следует за противительными союзами или модальными частицами, выражающими сомнение, или предшествует им, тем самым их смягчая (21)

(13) А Вы уверены*, что информация о состоянии дел в стране, которую Вы получаете как глава государства и которую Вы используете для принятия решений, является достоверной и объективной?* (ПЛ 2002)
(14) *Меня интересует, что на сегодняшний день происходит в Чечне,* может*, пора начинать переговоры*? (ПЛ 2002)
(15) *Это* разве *возможно*? (ПЛ 2002)
(16) *Реально* ли *это?* (ПЛ 2002)
(17) *Правильно* ли *это*? (ПЛ 2002)
(18) Похоже, *наше государство забыло о науке* (ПЛ 2002)
(19) *Она как бы заявлена, но реальных действий на местах не видно.* Я вот не вижу… *То есть квартплату поднимают* … (ПЛ 2002)
(20) *Но* мне кажется*, что в этом вопросе наше государство должно нас защитить* (ПЛ 2002)
(21) *Но вот,* на наш взгляд*,* …. (ПЛ 2002)

2.4 *Off-record* высказывания

Как отмечалось выше, выражаясь не прямо Г избегает определенных последствий, хотя одновременно данный способ коммуникации требует более искусного выражения. В последующих двух примерах увидим, как Г прибегают к стилистическим средствам, для того чтобы смысл высказывания прозвучал размытым, растворенным.

В первой части примера (22) Г якобы растворяет критику политических деятелей в пространном тексте, оценочный компонент которого просматривается через призму иронии: Г описывает картину, драматически актуализирует быт политической элиты (*наверху, происходит)*, амплифицируя её нанизыванием синонимов (*встречи, совещания, заседания, законы новые принимают*), придающим картине явный иронический оттенок. Критика эксплицитно направлена якобы только на чиновников (вторая часть примера):

(22) *У вас, наверху, происходит много всего интересного: встречи, совещания, заседания, законы новые принимают, а в стране и у нас, в Дальневосточном Федеральном округе, все больше и больше появляется чиновников. И только заниматься народом они не хотят, их только становится все больше и больше. Что с этим делать?* (ПЛ 2002)

Следующий пример (23) является аллюзией, в нашей перспективе аллюзия входит в число стратегий, направленных на определение общего пространства, общих знаний, или на смягчение речевого акта:

(23) Ермолаева Венера Лутфиевна, глава сельсовета. *Здравствуйте, уважаемый Владимир Владимирович! У меня вопрос: как дальше будет с пенсиями? И самое главное: как будет сохраняться накопительная часть пенсий?* Залезут ведь, точно, залезут… (ПЛ 2002)

Ответ Президента показывает, что адресат понял аллюзию, и хотя не сразу, на нее реагирует:

В. В. Путин: Вы знаете, это очень важный и интересный вопрос, который волнует миллионы людей (…)
Вы спрашивали о сохранности и использовании этой накопительной части. Очень важная тема. Должен Вам сказать, что от тривиального жульничества*, конечно, закон оградит любого человека. Закон этот достаточно качественный и все возможные защиты от, повторяю, элементарного жульничества там предусмотрены.*

2.5 В последнюю группу включим примеры, в которых говорящие не выражают никакой критики, а, наоборот, намереваются установить с адресатом эмпатию, развертывая все те стратегии, которые, как установлено Брауном и Левинсоном, направлены на удовлетворение “positive face” собеседника (…), ссылаясь на общий опыт, на частные, но доступные факты личной жизни Президента. Можно заметить, что именно эта стратегия стала преобладающей с течением времени. Среди множества примеров можно выделить (27), который показывает провал выбранной тактики, что видно из ответа Президента (который не мог иначе ответить).

(24) *И, кстати, можно спросить: пользуетесь ли Вы такими лекарствами*? (ПЛ2002)

(25) *Я, кстати, знаю, что Вы активно занимаетесь спортом. Может быть, надо ввести специальный «путинский стандарт»? Не думаю, что половина наших генералов сможет подтянуться хотя бы 10 раз.* (ПЛ 2002)

(26) *абсолютно уверены, что Вы владеете проблемами, которых великое множество на деревне*. (ПЛ 2002)

(27) *Мне кажется, там будет второй Вьетнам. Мне кажется, просто все побегут оттуда, и оставшийся там хаос ударит по всем.* (ПЛ 2003)

(28) *Я знаю, что в ночь перед выборами в Госдуму у Вашей любимой собаки Кони родились щенки. Я бы хотел узнать, как щенки себя чувствуют и какова их дальнейшая судьба?* (ПЛ 2002)

Итак, проведенный анализ показывает, что ПЛ, по крайней мере в первых выпусках, являлась особым видом политического институционального дискурса Особым, потому что несмотря на явную асимметричность участников, оставляла некоторую свободу и слабому участнику, свободу выбирать предмет коммуникации (семантическое доминирование). Все-таки неравновесие заставляло слабых, наивных интервьюеров, прибегать, даже когда они намерены выступить открыто (*on record*), ко всем им доступным стратегиям смягчения, ослабления речевых актов, чтобы минимизировать последствия собственных высказываний.

Литература

1. *Auer P. Di Luzio A.* (eds) The contextualization of language // Amsterdam, Philadelphia: Benjamins, 1992.
2. *Caffi C.* La Mitigazione, Pavia: C.L.U., 2000 (английский перевод: Caffi C., Mitigation, The Netherland: Elsevier, 2007)
3. *Clayman S. E.* Footing in the achievement of neutrality: the case of news interview discourse // Drew P. Heritage J. Talk at work. Interaction in istitutional settings // Cambridge: Cambridge University Press 1992, p. 163–198.
4. *Drew P. Heritage J.* Talk at work. Interaction in istitutional settings // Cambridge: Cambridge University Press 1992
5. *Гаврилова М. В.* Языковые средства актуализации значимой информации в выступлениях Президента Путина // http://www.philol.msu.ru/~rlc2004/ru/participants/psearch.php?pid=17345
6. *Голанова Е. И.* Публичный диалог: коммуникативный узус и новые жанровые характеристики // http://fixed.ru/prikling/conf/stilsist1/publijeatzxb.html
7. *Колокольцева Т. Н.* Новые жанры диалогической коммуникации: теледиалог Президента с гражданами России // http://www.philol.msu.ru/~rlc2004/ru/participants/psearch.php?pid=17345
8. *Linell P. Luckmann T.* Asymmetries in Dialogue: Some Conceptual Preliminaries // Markova I. Foppa K. Asymmetries in Dialogue // Harvester Wheatsheaf: Hemel Hempstead, p. 1–20.
9. *Паршина О. Н.* Российская политическая речь // Москва: УРСС, 2007.
10. *Прямая* Линия 2003 http://www.linia2003.ru/
11. *Прямая* Линия 2002 http://www.linia2002.ru/
12. *Thomas J. A.* The Language of Power. Towards a dynamic pragmatics // Journal of Pragmatics 9 (1985), p. 765–783.

Невербальное поведение людей разных культур в диалоге I: финская и русская жестовые системы[1]

The nonverbal behavior of people of different cultures in a dialog I: Finnish and Russian gesture systems

Крейдлин Г. Е. (gekr@iitp.ru)

Российский государственный гуманитарный университет,
Институт лингвистики, Москва

В работе излагаются некоторые размышления внешнего, так сказать, стороннего, наблюдателя о финской невербальной знаковой культуре, некоторых невербальных знаках и моделях диалогического поведения финнов. В целях сопоставления привлекается соответствующий русский материал.

1. Общая характеристика финской жестовой системы

До настоящего времени финские жесты систематическому анализу не подвергались (жесты я понимаю здесь широко, т. е. как включающие в себя жесты рук, ног, головы и плеч, позы, мимические жесты, знаковые телодвижения, знаки-касания, знаки-взгляды и комплексные вербально-невербальные знаковые формы поведения — манеры). Больше того, мне не известны даже работы, которые были бы посвящены каким-то отдельным сторонам финской невербальной культуры или отдельным финским жестам. Между тем мои наблюдения над коммуникацией финнов, проведенные интервью и обсуждение с довольно широкой молодежной аудиторией[2] невербальных знаков, которые жители финских городов (Хельсинки, Тампере, Иматра, Турку) часто используют в повседневной коммуникативной практике, позволяют утверждать следующее. В финской культуре, прежде всего городской[3], есть целый ряд непривычных для нас и весьма интересных для научного анализа невербальных феноменов.

Хотя в диалоге финны, разумеется, пользуются соматическими знаками, их жестовый код сильно редуцирован в физиологическом отношении и довольно скромен в отношении количественном. Лексическая система языка финских жестов состоит в основном из смеси шведских, немецких и русских знаков, и все эти знаки бесконфликтно живут в культуре и устной коммуникации финнов.

На функционирование финских жестов очевидное влияние оказывает протестантская культура, осуществляющая негласный контроль над телесной этикой (о телесной этике и невербальном контроле см. Крейдлин 2002а; Крейдлин 2005) и невербальным этикетом интерактивного поведения (о невербальном этикете см. Крейдлин 2002а; Крейдлин 2002б; Морозова 2004). Это влияние проявляется, прежде всего, в разных способах **сокрытия тела**, что, в свою очередь, выражается в очевидной минимизации телесных манифестаций, в упрощении — вплоть до полного бездействия и молчания — материальных знаковых средств выражения смысла, и в явном **культивировании понятий стыда и скромности**. Последнее отражается, в частности, в стремлении (1) избегать в диалоге невербальных действий, в частности движений, в результате которых партнер теряет лицо в глазах собеседника и, возможно, в глазах других людей, и в склонности (2) к постоянной и осознанной рефлексии своего невербального поведения. Иными словами, если воспользоваться семантическим языком А. Вежбицкой, финны все время 'думают о том, что другие могут подумать о твоем поведении сейчас'.

Финская культура, и не только невербальная, отличается простотой и самоограничением. Финны, если это не необходимо, с вами не будут разговари-

---

[1] Работа поддержана коллективным грантом РГНФ — проект 07-04-00203а.

[2] Широкое обсуждение жестов и разных особенностей невербального поведения финнов проводилось мной, в основном, в двухнедельный период пребывания в Хельсинки в ноябре 2006 года, когда я читал лекции студентам отделения семиотики Александрова института Хельсинского университета. Однако коллекция жестов создавалась и пополнялась также в другие, как более ранние, так и более поздние, поездки в Финляндию.

[3] За время пребывания в Финляндии я провел (в общей сложности) пять дней за пределами финских городов, на хуторах, и, насколько я могу судить, невербальное поведение живущих там финнов обладает по сравнению с поведением городскими жителями своей спецификой.

вать, и уж тем более не станут без необходимости активно жестикулировать. В своей массе они послушны и скромны в поведении. Особенно внимательно они относятся к людям, которые старше по возрасту или по социальному положению; кажется даже, что финны уважают старших больше, чем себя, и со старшими молодые финны ведут себя особенно сдержанно. Вера финнов в свою политическую систему, социальную защищенность (что, безусловно, абсолютно оправданно) и религию заставляет их быть чрезвычайно восприимчивыми к механизмам контроля диалогического поведения не только своего, но и партнера и тонко чувствовать изменения в способах и моделях такого поведения.

Люди многих культур приписывают финнам такие качества, как задумчивость, медлительность, чрезмерную серьезность и молчаливость, а серьезность и в еще большей степени молчание связывают с аскетизмом финнов и в какой-то мере с асоциальностью. В фильмах режиссера, получившего мировую известность, Аки Каурисмяки, герои ведут своеобразные длительные диалоги молчания без особой жестикуляции, причем, по мнению кинокритиков, эти диалоги занимают важное место в этической и эстетической системе режиссера. Историческая традиция протестантской этики учит финнов уделять больше внимания духовным и когнитивным параметрам и измерениям жизни, а не телесным, а в молчании, как говорят финны, проявляются близость людей и мудрость.

Сильную конкуренцию внутреннему влиянию на коммуникативное поведение финнов протестантской культуры и этики оказывает влияние внешнее. Оно идет, по всей видимости, от англосаксонской культуры, главным образом, американской. Современный финский жестовый язык постоянно пополняется из телепередач, электронных СМИ, кино и театра, где число американских и английских текстов превосходит, по мнению моих, в основе своей критически настроенных финских студентов, все допустимые нормы.

Впрочем, внутреннее влияние протестантской этики пока еще явно преобладает. Неслучайно поэтому многие финны, в особенности молодые, говорят о том, что их поведение до сих пор часто интерпретируется носителями других культур как изоляционизм и аскетизм, а то и как нежелание общаться или простое занудство — в общем, как нечто крайне неприятное и досаждающее другим, что приходится терпеть.

2. Некоторые финские жесты и модели коммуникативного поведения

Теперь я остановлюсь на некоторых финских жестах и рассмотрю отдельные способы и модели невербального и смешанного, вербально-невербального, диалогического поведения финнов.

План дальнейшего изложения таков.

Сначала речь пойдет об особых дейктических жестах двумя руками и характерном использовании финнами так называемых жестовых удлинителей. Удлинители — это такие материальные объекты, которые, не будучи частями тела человека, могут служить составным элементом того или иного жеста. Например, русский жест **показать пальцем** часто осуществляется ручкой или другим подобным предметом (карандашом, веточкой и т. п.), как бы составляющими продолжение руки. Между тем название жеста, будучи конвенциональным и вполне устойчивым, при этом сохраняется: жеста **показать ручкой** в русском языке жестов нет. Аналогично, исполняя жест **помахать рукой** <на прощание>, женщина может воспользоваться платком или шалью, а мужчина — шляпой, шапкой, шейным платком и др., и всё это будут удлинители руки[4].

После разбора дейктических жестов я остановлюсь (б) на одном интересном финском невербальном феномене — так называемых прерванных жестах и поговорю (в) о финских жестах-взглядах. Однако для этого мне понадобится ввести или напомнить ряд понятий и терминов науки невербальной семиотики.

***

В невербальной семиотике выделяются три семиотических класса жестов. Это (а) **эмблематические жесты**, или **эмблемы**, имеющие самостоятельное лексическое значение и способные выражать и передавать его в диалоге отдельно и независимо от вербального контекста. Это (б) **иллюстративные жесты**, или **иллюстраторы**, выступающие в диалоге только вместе с речью и указывающие или выделяющие в ней или каких-то других элементах диалога некоторый объект или фрагмент. И, наконец, это (в) **регулятивные жесты**, или **регуляторы**, управляющие ходом диалога — устанавливающие, поддерживающие или завершающие диалог.[5]

Некоторые эмблемы и иллюстраторы содержат в своей семантике указание на участников актуальной ситуации общения, на объект и его признаки, на место и время, релевантные для данной ситуации. Например, для невербального выражения смысла 'я' европейцы и американцы часто показывают указательным пальцем или рукой на область сердца или груди, а китайцы — на нос. Такого рода эмблемы или иллюстраторы, не важно, — называют **дейктическими**, или **указательными**, **жестами**. Примерами русских дейктических жестов являются единицы **подзывать рукой**, **показать рукой**, **показать пальцем**, **показать глазами**, **поманить пальцем**.

---

[4] Подробно об удлинителях см. в СЯРЖ 2001; 3].

[5] Об этих классах жестов см. в книгах.

Различие между дейктическими эмблемами и дейктическими иллюстраторами состоит только в том, сопровождают ли они речь и, если да, то с какой степенью обязательности. Впрочем, граница между этими двумя видами жестов не является жесткой.

Дейктические иллюстраторы дополнительно к вербальному сообщению указывают на человека, объект, размер, место даже том случае, когда референт в актуальном пространстве диалога отсутствует (ср. фразы *Я только что оттуда* и *Подойдите ко мне вот настолечко поближе* с соответствующими жестами). Кроме того, они могут указывать на время или на стадии события, о котором идет речь в повествовании, ср. последовательность движений, сопровождавшую устный рассказ: *Сначала* — рука идет в сторону — *он встал, потом* — рука отводится еще дальше — *пошел к двери*<…>. А слова *вон там* часто сопровождаются комплексом иллюстраторов — **кивком головы** и **взглядом** в сторону человека или предмета, которые могут указывать также на объекты, отсутствующие в поле зрения собеседников. Для невербальной характеристики диалога присутствие/отсутствие в нем третьего лица является важным фактором.

В работах Крейдлин 2007, 2008 было показано, что форма дейктических жестов определяется значениями трех признаков, а именно, (1) 'каков активный орган жеста и/или какова его рабочая часть'; (2) 'каково направление этого органа (части) в данном жесте' и тем (3) 'какова (для мануальных жестов) ориентация ладони'.

***

Русские дейктические жесты осуществляются чаще всего рукой (причем, и это важно (!), обычно одной рукой) или пальцами руки — указательным, большим и мизинцем[6]. Кроме того, русские иногда указывают головой и глазами. А в других культурах и даже окказионально в русской культуре для указаний могут использоваться также другие части тела. Например, в ряде стран Латинской Америки распространена указательная жестовая комбинация: **подбородок — взгляд**, ср. предложение, демонстрирующее возможность указания подбородком у русских: *Что все это значит? — и он строго указывает подбородком на злополучную расписку* (*С. Гандлевский*). У некоторых народов есть указательных жесты ноги (De Jorio 2000, 70–74), а масаи в Африке (Sherzer 1972), индейцы Куна в Панаме (Enfield 2002) и некоторые другие народы широко пользуются указательным жестом, совершаемым нижней губой.

***

Теперь перейду к финским дейктическим жестам.

У финнов в городах (на хуторах я его не видел) можно встретить свой, крайне необычный для нас указательный жест, совершаемый сразу двумя руками. Его можно назвать «**махать двумя руками**». Человек держит руки почти горизонтально перед собой на уровне плеч. Локти чуть прижаты к корпусу. Пальцы вытянуты и раскрыты, руки поднимаются и опускаются приблизительно в той же манере, с какой ведется речь. Скорость, частота и амплитуда движения меняются в зависимости от того текста, который в данный момент произносится, и от того, в какой позе выполняется жест, например, стоит человек или сидит — в положении стоя размах и объем движения бóльшие. Жест **махать двумя руками** — иллюстративный, поскольку всегда сопровождает речь, и относится к подклассу аккомпаниаторов, а не демонстраторов, так как является дополнительным средством экспрессии, а не служит для показа формы, размера, места или какого-то другого параметра текста или ситуации общения.

Это именно указательный жест: финны указывают им то место в произносимом тексте, когда излагается некий факт или событие. Жест является контекстно жестко обусловленным. Сразу за ним, а чаще вместе с ним, следует предложение или более короткая синтаксическая группа, содержание которой является оговоркой или противоречием только что произнесенной фразе, той, что шла до жеста. Как мне объяснили финны, данный жест очень часто исполняется вместе и одновременно с произнесением финского аналога русского слова *но*.

Вот финн, которого я просил найти хельсинский адрес одного человека, произносит по-английски *We've got an ad'dress* 'Мы достали адрес'. И, произнося эту фразу, человек машет обеими руками. Руки поднимаются вверх на артикле и опускаются вниз на ударном слоге слова *ad'dress*. Сразу вслед за жестом идет продолжение: *But it is written in Finnish* 'Но он написан на финском'. Это как раз и есть та самая оговорка, о которой я говорил выше. Она показывает, что воспользоваться бумажкой, на которой записан адрес, мне не удастся. Жест демонстрирует то, что адресат (здесь я сам) должен понять вторую часть текста как результат отбрасывания первой. Данным жестом первая часть текста как бы иконически выводится за пределы пространства говорящего и, что очень важно, руки при отбрасывающем движении отходят от корпуса чуть в сторону.

При жесте «**махать двумя руками**» возможны удлинители — в данном случае это был листочек бумажки с адресом, но могут быть и другие удлинители — ручка, карандаш и т. п.

Во время изучения данного жеста у меня возникло предположение, имеющее общий характер,

---

[6] Мизинцем часто показывают маленькие дети. Гораздо реже им пользуются для указания взрослые; делают это, главным образом, женщины, причем в игровой, «заговорщицкой» или шутливой ситуации/ манере, например, когда взрослый ведет себя, как ребенок.

а именно, что иногда информации о наличии при каком-то жесте удлинителя и о характере удлинителя, информации, которая имеется в Словаре языка русских жестов (СЯРЖ 2001), может оказаться недостаточной для правильной реализации (исполнения) жеста. В ряде случаев необходимо указывать также место, куда во время исполнения жеста помещается удлинитель, а этой информации в СЯРЖ нет. Так, в жесте **махать двумя руками**, по утверждению финнов, предмет-удлинитель в норме помещается между пальцами, как та бумажка, которую держал студент, чтобы полная форма жеста в большой степени напоминала каноническое исполнение. Разумеется, пальцы тут уже не будут раскрытыми, однако предпочитаемое сходство форм остается.

Теперь остановлюсь на **прерванных жестах**.

Под прерванным жестом я имею в виду любой жест, реализация которого внезапно прекратилась внутри текущего коммуникативного или когнитивного процесса. Прерываться могут почти все жесты, но процесс прерывания легче всего увидеть на эмблемах. Прерванные жесты больше характерны для финской бытовой коммуникации, чем для коммуникации профессиональной, социальной или деловой.

Прерывание жеста выражает смыслы 'сбивчивости', 'несвязности' или отражает внезапный переход говорящего/жестикулирующего к другой мысли — переход, столь характерный для финской культуры. Прерванный жест подчеркивает изолированность, отдельность или, может быть, даже точнее, отделённость, одного человека от другого[7]. А потому не удивительно, что носителями многих других культур подобное невербальное поведение представляется как агрессия или как нарушение одного из кооперативных законов телесной коммуникативной этики. И собеседник финна, не знакомый или плохо знакомый с финской культурой, каковым был и я сам, обычно в недоумении заканчивает или на какое-то время прекращает разговор.

Прерывание жеста может происходить на разных стадиях его реализации — экскурсии, воспроизведении, рекурсии — и, естественно, может обозначать разные вещи. Прерванные жесты находятся в стороне от магистральной линии употребления знаков и выражаемых ими значений, но отражают соотношение между ментальностью, то есть когнитивными и психофизиологическими процессами, происходящими в мозгу финна, с одной стороны, и принципами знаковой коммуникацией финнов, с другой. Эти жесты тесно связаны с культурой стыда, то есть с тем понятием, которое, я убежден, относится к основным этнокультурным характеристикам финского народа.

Приведу пример употребления прерванного жеста.

Отойдя на некоторое расстояние после вполне дружеского разговора со мной, женщина, повернувшись лицом, начинает махать мне на прощание рукой, исполняя хорошо знакомый мне прощальный жест. Однако в самом его начале, едва успев взмахнуть один, от силы два раза, рукой, женщина вдруг резко опускает руку, как бы о чем-то внезапно подумав или испытав внезапную боль, и полностью прекращает жестикуляцию. Она заметила, что на нее в этот момент кто-то смотрит.

Финны, с которыми я общался по-английски, когда я рассказал им про увиденное и поделился с ними тем, что в этот момент почувствовал, использовали для характеристики поведения женщины глагол *desert*. Этот глагол обозначает негативно окрашенное действие или поведение, ср. его русские эквиваленты: 'бросать, покидать (кого-л.); бежать (от кого-л.); а также сочетания *to desert a friend* 'оставить друга (в беде)'; *to desert one's family* 'бросить семью (на произвол судьбы)'; *to desert one's leader* 'оставить своего командира'.

Жест как элемент диалога оказался здесь совершенно бесполезным знаком, так как не нашел своего адресата. И все же не только этот, но и любой другой прерванный жест, понимается как значимый соматический нулевой знак, как знаковый отказ от, так сказать, материально выраженного жеста. Прерванный жест следует, на мой взгляд, интерпретировать как позитивный знак в системе культурной идентификации финнов — как коммуникативный знак присущих финнам скромности, стыда и аскетизма, а не как отражение полного безразличия к происходящему.

Прерванные жесты наряду со знаковым молчанием создают невербальную драматургию устного текста. Они говорят о том, что перед тем, как действовать, следует подумать, и если нужно, то подавить действия и уменьшить количество невербальных знаков либо изменить их качество, редуцируя (полную) стандартную форму.

Теперь несколько слов о знаках-взглядах.

В финской культуре есть один необычный взгляд, который имеет знаковую природу и который по-фински называется **katseluasento** 'увидел море'. На английский язык он переводится как *admiring the scenery*.

Визуальные наблюдения и анализ контекстов употребления **katseluasento**, а также последующий опрос информантов по поводу обнаруженных мной особенностей его употребления и употребления некоторых его номинаций в письменных текстах, позволяют выделить следующие смыслы жеста (ограничусь тут однословной характеристикой семанти-

[7] Это disconnectedness of individuals, как выразился во время обсуждения понятия «прерванный жест» один из моих слушателей — финский студент, подтвердив мое предположение.

ки жеста, см. о ней в СЯРЖ 2001, Крейдлин 2002). Это 'концентрация внимания', 'восхищение увиденным', 'уединенность в пределах своего личного пространства', 'психологический отдых', или, по словам одного финна, 'resting mentally'.

Вот человек идет по улице, внезапно останавливается, голова чуть вскидывается вверх, руки опущены, локти чуть отведены назад, глаза уставлены в одну точку. Поза чуть похожа на русскую позу военных **Смирно,** или **Вытянуться по струнке**: глаза человека широко открыты, взгляд застывший, но в этой позе он не видит ничего вовне себя: взгляд устремлен вовнутрь, и сконцентрирован **только на том, что в этот момент увидено внутренним взором**. Взгляд застыл, но ноги человека находятся в привычном и удобном положении, он может совершать мелкие движения туловищем на месте или пальцами, например, чуть постукивать ими по бокам или бедрам. Человек стоит так примерно 15 — 20 сек. Он находится на психологически далеком расстоянии от других людей.[8] Эмблематический комплекс «застывший взгляд **katseluasento** + застывшая поза» сигнализирует, что никто и ничто не должно сейчас мешать человеку. Его невербальное поведение говорит о том, что ему сейчас нужно уединение и пространство — и не только физическое, но и психологическое. Именно самой этой актуально принимаемой позой такое личное пространство и устанавливается.

Пространство это закрыто для окружающих, так как жестикулирующий не хочет, чтобы его беспокоили, но оно легко открывается и становится доступным для проникновения внутрь. Если человека потревожат, то он не злится — по крайней мере, мне не было заметно, чтобы злился. И уж точно жестикулирующий человек не становится от этого враждебным к потревожившему его лицу. Мелкие, незначительные события (какой-то шум, встречный уличный поток, разговоры вокруг, говорящая реклама и др.), которые как помехи теоретически могли бы повлиять на поведение жестикулирующего, отвлечь его и вынудить изменить позу и взгляд, обычно этого не делают. Но если человека окликают, как это было в трех из примерно двадцати случаев употреблений описываемого жестового комплекса, отмеченных мной в общей сложности за примерно месяц пребывания в Финляндии, то такое событие уже не является незначительным. Человек, если к нему в этот момент обращаются, будто встряхивается, приходит в себя и «возвращается к обычной жизни».

«Перебивать» эти взгляд и позу, например, уставившись на жестикулирующего, или нарочно громко позвать его, считается социально неприемлемым поведением. Окружающие сочтут вас невежливым, нарушающим законы этики, а то и бесчувственным и недружелюбным — у вас должна быть веская причины, чтобы помешать другому человеку вести себя таким образом. Напротив, не обращать на жестикулирующего внимания, считается проявлением уважения к личности и к личному пространству человека. Не смотреть на другого, когда на то нет особой необходимости, — это типичная черта финской культуры и коммуникативного поведения финнов. Вот еще один пример из моей собственной практики. Когда вы входите в аудиторию, где должны читать лекцию, коренных финнов среди студентов других национальностей выделить довольно просто: финские студенты как будто не замечают вас. Они просматривают какие-то бумаги, читают газеты и совсем не смотрят в вашу сторону, будто вас тут нет вообще. Они не обращают на вас никакого внимания до тех пор, пока вы не поздороваетесь со всеми слушателями и не скажете, что лекция начинается. По окончании ее финны тоже молча собираются и также молча покидают аудиторию, не смотря на вас и буркая в сторону нечто вроде «До свидания». В то же время, если вы обращаетесь к финну, будь то в помещении или на улице, с каким-то вопросом или просите помочь, на вас обязательно посмотрят и обязательно (!) помогут, если смогут

Финны говорят, что взгляд **katseluasento** чаще можно встретить весной и летом. Это, как мне кажется, можно объяснить тем, что весной и летом больше тепла и света, люди ведут себя более активно, и растет число жестов. Кроме того, людям в силу повышенной активности чаще нужна релаксация и требуется больше времени на отдых. Зимой же — холодно, темно, двигательная активность у людей, в том числе жестовая, снижена, и отдыхать по времени можно меньше.

В названии жеста присутствует элемент иронии, но за ним стоит известная любовь финнов к воде, к водной среде. Финны очень любят гулять по берегу моря, стоять на пирсе или у дамбы и просто смотреть на воду или поверх воды. Они часто проводят свободное время, например, отпуск, на берегу моря, озера или реки. Финны любят ходить в сауну, охлаждаться после парилки, сидя или лежа у края бассейна, стоять полуобнаженными с полотенцем поверх плеча или в руке и с каким-нибудь напитком и т. п. Наслаждаясь этим, финны могут ни на кого не смотреть, они как бы застывают в позе и взгляде. Именно так происходит невербальная реализация соответствующих смыслов.

В диалоге, однако, такой взгляд невозможен. Он недопустим даже в диалоге равноправных партнеров, не говоря уже о диалоге с ранжированными коммуникативными или социальными ролями. Одна студентка из Австрии, долгое время живущая в Хельсинки, прокомментировала это мое наблюдение следующим образом (цитирую в своем переводе

[8] О видах расстояний, о физическом и психологическом расстояниях и о науке проксемике, которая изучает пространственные знаки и их использование в коммуникации, см., например, Крейдлин 2002а.

с английского прямо по записи): «Гость тоже не может так смотреть на хозяина. Когда тот ему что-нибудь показывает, там дом или сад, то гость по этикету должен выказывать удивление и восхищение». Взгляд **katseluasento** выглядел бы тут аномалией, психической диспропорцией, «дисбалансом», или *mental imbalance*, по замечанию той же студентки.

3. Заключение

Семантические области, в которых существуют и проявляют себя жесты, в значительной степени являются культурно обусловленными. Телесная символика, включающая в себя жесты, позы, знаковые телодвижения, взгляды, касания, мимику и манеры, а также конкретные модели и приемы невербального выражения смысла, помогает реконструировать тот тип мышления, который стоит за ними и который обеспечивает существование и функционирование в данной культуре этих единиц, моделей и приемов. Она позволяет также обнаружить скрытые культурные смыслы, которые прошли мимо внимания тех лингвистов, которые исследуют исключительно вербальные компоненты коммуникации, и выявить некоторые механизмы и правила диалогического взаимодействия.

Литература

1. *Крейдлин* 2002а — Крейдлин Г. Е. Невербальная семиотика. Москва: «Новое литературное обозрение». 2002.
2. *Крейдлин* 2002б — Kreydlin, G. E. Ethics and etiquette in nonverbal signs ('Этика и этикет в невербальных знаках') // Лингвистический беспредел. Сборник статей к 70-летию со дня рождения проф. А. И. Кузнецовой (составители Т. Б. Агранат, О. А. Казакевич, под общей редакцией А. Е. Кибрика). Москва: Изд-во МГУ, 2002. С. 310–320.
3. *Крейдлин* 2005 — Крейдлин Г. Е. Невербальный контроль в диалоге: единицы, модели, правила / Труды международной конференции «Диалог' 2005: компьютерная лингвистика и интеллектуальные технологии». Москва, 2005.
4. *Крейдлин* 2007 — Крейдлин Г. Е. Механизмы взаимодействия невербальных и вербальных единиц в диалоге: II А. Дейктические жесты и их типы // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» Москва, вып. 6(13), 2007. С. 300–327.
5. *Крейдлин Г. Е.* Механизмы взаимодействия невербальных и вербальных единиц в диалоге: II Б. Дейктические жесты и речевые акты // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2008» (Бекасово, 4–8 июня 2008 г.) — М., вып. 7 (14), 2008. С. 248–253.
6. *Морозова Е. Б.* О понятии невербального этикета (к постановке проблемы)» // Агрессия в языке и речи. Сборник научных статей (составитель И. А. Шаронов). Москва: изд-во РГГУ, 2004. С. 67–80.
7. *СЯРЖ* 2001 — Григорьева С. А., Григорьев Н. В., Крейдлин Г. Е. Словарь языка русских жестов. Москва — Вена: «Языки русской культуры», Wiener Slawistischer Almanach, Sonderband 49, 2001.
8. *De Jorio, A.* Gesture in Naples and Gesture in Classical Antiquity. A translation of "La mimica degli antichi investigate nel gestire napoletano" (1832), and with an Introduction and Notes, by A. Kendon. Bloomington: Indiana Univ. Press, 2000
9. *Enfield, N.* 'Lip-pointing': A discussion of form and function with reference to data from Laos. Gesture, v. 2, № 1, 2002. С. 185–211.
10. *Sherzer, J.* Verbal and nonverbal deixis: The pointed lip gesture among the San Blas Cuna. Language in society, 2, № 1, 1972. С. 117–131.

Программа семантической классификации лексики — ПроСеКа: теоретические и прикладные аспекты

On the semantic classification program ProSeCa: theoretical and practical aspects

Кретов А. А. (a_a_kretov@rambler.ru)
Воронежский государственный университет, Воронеж

Рафаева А. В. (anna_raf@rambler.ru)
Московский государственный университет им. М. В. Ломоносова, Москва

Для семантической классификации лексем предлагается использовать модифицированный метод словарной идентификации Э.В.Кузнецовой, ориентированный на лексическую, а не грамматическую семантику. Описывается компьютерная программа ПроСеКа, облегчающая процесс семантической классификации лексики.

Исследование лексико-семантических процессов и состояний лексико-семантической системы, предполагает знание существа, скорости и направления этих процессов, что предполагает хронологическую и семантическую привязку слов и их значений. Хронологическая привязка осуществляется через датировку обрабатываемых текстов. Для осуществления семантической привязки необходимо явным образом описать семиосферу и определить в ней место каждого значения. Наиболее близкими к предлагаемому подходу являются работы по построению ресурсов типа WordNet для разных языков, в которых лексические значения описываются в виде семантической сети [Fellbaum 1998; Азарова и др. 2004]. Но этот подход отличен от нашего. Одно дело — когда каждая дефиниция (и соответствующее ей значение слова) является траекторией в пространстве метаслов. Другое — когда значения лексем представляют собой не цепочку, а узел графа. Узлы графа объединяются отношением «толкуемое-толкующее» или «частное-общее». Таким образом, значение у нас — не семантическая цепочка, а одно из её звеньев — единица классификации.

Синопсисы идеографических словарей отражают как строение семиосфер, так и мировоззрение составителей. Стремлением уменьшить произвол в идеографии было продиктовано алгоритмизованное обращение к дефинициям толковых словарей Э. В. Кузнецовой [Кузнецова 1969], Ю. Н. Караулова и других [Караулов 1982].

Опыт коллектива, созданного Э. В. Кузнецовой [Кузнецова 1988] и руководимого Л. Г. Бабенко [Бабенко 1999], показал, что глагольная семантика по природе своей грамматична, и чем далее мы идём по цепочке глаголов-идентификаторов, тем меньше лексического остаётся в глаголе и тем очевиднее «грамматичность» его семантики.

Например, *плестись* — «*идти* медленно, устало, с трудом передвигая ноги» > идти «*двигаться*, передвигаться ступая ногами» > двигаться «*совершать* движение» > совершать «*делать*, осуществлять, производить» > делать «*совершать*, выполнять, производить». Круг замкнулся. *Производить* > осуществлять «приводить в исполнение, *воплощать* в действительность» > воплощать «*делать* реальностью, осуществлять». Таким образом, от *идти* остаётся лишь ‘каузировать быть’ (осуществлять перемещение ног) или на языке семантических функций «Caus перемещение ног», соответственно, *плестись* — «медленно, устало, с трудом Caus перемещение ног». Поскольку ноги используют преимущественно для перемещения, формулу можно переписать как «Caus Func ноги». Как видим, лексическая семантика сконцентрирована в существительном *ноги*, а ‘функционировать’ и ‘каузировать’ — значения грамматические.

Итак, метод словарной идентификации Э. В. Кузнецовой необходимо переориентировать с метаслов-глаголов на метаслова-существительные.

Для этого надо организовать привязку лексических значений к семантическому пространству. Сложность задачи состоит в том, что имеющиеся описания семантического пространства по разным причинам непригодны, а «опробованный в экспе-

рименте Ю.Н. Караулова метод автоматического извлечения тезауруса из толкового словаря может использоваться только при условии дальнейшего редактирования полученных данных человеком» [Кобозева 2000: 134] или при условии предварительного редактирования дефиниций: ведь лексикографы не предполагали, что их дефиниции послужат семантической классификации лексики.

Для описания семиосферы с опорой на дефиниции нужна специальная компьютерная программа, избавляющая исследователя от повторения уже выполненной работы по семантизации ЛСЕ (значений слов и фразеологизмов) в произвольно выбранном тексте.

Такая потребность возникает в силу ряда причин: 1) открытости словаря, 2) необходимости описывать синхронное состояние лексико-семантической системы, 3) необходимости определять тематическую принадлежность текстов, 4) необходимости проекции лексической семантики текста на систему семантических координат.

Осуществление полной семантизации ЛСЕ текста или корпуса текстов открывает целый ряд новых возможностей: 1) создание частотно-семантических словарей, названное П. М. Алексеевым «оценкой толкового словаря по тексту» [Алексеев 1973], 2) создание исторической лексикологии русского (а в перспективе — любого другого языка, имеющего письменную традицию), 3) типологические исследования лексической семантики и т.д.

Предлагаемое решение задачи базируется на следующих положениях:

1) лексическая семантика не зависит от грамматики данного языка («части речи» — разные по форме сосуды, наполняемые одной и той же лексико-семантической субстанцией);
2) в индоевропейских языках лексическая семантика концентрируется в именах (существительном и прилагательном) и глаголе (в отличие от WordNet'а, мы не работаем с наречиями), а в общем случае — в лексических морфемах — корнях.
3) лексическая семантика глагола и прилагательного в конечном итоге сводится к семантике существительного и может быть описана через неё (например, *белый* — 'цвета *снега, мела* или *молока*');
4) собственно глагольная семантика при ближайшем рассмотрении оказывается грамматической: процессуальной (действие, отношение, состояние, погружённые во время), инхоативной-каузативной (*ослепнуть* — 'начать не иметь зрения' и *ослепить* — 'каузировать начать не иметь зрения'), фазовой (*расцвести* 'начать быть — о цветке', *увянуть* 'кончить быть — о цветке), утвердительной-отрицательной (*ослепить* — каузировать кого-л. начать не иметь зрения, *воочесить* — 'каузировать кого-л. начать иметь зрение'), акционсъартной (*петь — попеть, запеть, пропеть, допеть, распеться, отпеть, отпеться* и т. д.);
5) семантические функции И.А.Мельчука — А.К.Жолковского — Ю.Д.Апресяна являются грамматической надстройкой над лексической семантикой; аналогичный статус имеют и «семантические примитивы» А.Вежбицкой.
6) лексическая семантика не выразима вне частеречного оформления; следовательно, при анализе лексической семантики следует ориентироваться на наименее маркированную часть речи — существительное и те значения, которые им выражаются. Наименьшая маркированность существительного как части речи обоснована В.Г.Руделёвым [Руделёв 1995], а также выводится из сближения маркированности с рецессивностью, а немаркированности с доминантностью, предложенного Вяч. Вс. Ивановым и Т.В. Гамкрелидзе [Гамкрелидзе, 1984]. Наиболее многочисленный член оппозиции является доминантным и немаркированным, а наиболее многочисленная часть речи в известных нам словарях — имя существительное;
7) специфика лексической семантики может быть выявлена и описана только в результате последовательного снятия грамматических надстроек и напластований, составляющих «грамматику семантики»;
8) всё регулярное должно выноситься из словаря в грамматику [Щерба, 1974; Морковкин 1990];
9) графически лексико-семантическое пространство может быть представлено в виде ориентированного графа, исходными (при развёртывании) и конечными (при классификации значений) узлами которого являются базовые понятия человеческого языка.

Проектами, подобными данному, являются: [Паллас 1787–1789; Шишков 1832; Roget 1986; Lorge, Thorndike, 1938; Прокопов 1945; Караулов 1982; Морковкин 1984; Кузнецова 1988; Баранов 1995; Шведова 1998–2007, Fellbaum 1998; Бабенко 1999; Лукашевич, Добров 2002].

Идеографические словари отражают языковую реальность на нижних уровнях обобщения (синонимические ряды, гипо-гиперонимические отношения), а на высших уровнях обобщения количество и качество выделяемых таксонов зависит от исследователя, на что указывали [Задорожный 1983] и [Караулов 1976, 1981]. Особенно хотелось бы отметить значительную частеречную независимость «Тезауруса» [Roget 1986], хотя и обусловленную морфологической бедностью английского языка, но принципиально верную.

«Русский семантический словарь» [Шведова 1998–2007] — в соответствии со взглядами Н. Ю. Шведовой — ориентирован на частеречную

семантику, поэтому *белый, белеть-белить* и *белизна* в нём оказываются в разных местах и разных томах, а сбор этих лексически тождественных значений воедино по трудозатратам близок к созданию нового идеографического словаря.

Словарь [Lorge, Thorndike 1938] свидетельствует о принципиальной возможности тотальной семантизации больших корпусов текстов, правда, опыт, накопленный при создании этого словаря, практически недоступен, равно, как и проверка обоснованности решений, принятых его составителями. В цифры, полученные исследователями, остаётся только верить или оценивать их достоверность, исходя из соображений общего плана.

В этом отношении содержательнее «Русский семантический словарь» [Караулов 1982]. К сожалению, слова-аттракторы в нем задавались a priori, а не получались в ходе исследования. Кроме того, этот словарь, как и словарь екатеринбургского коллектива [Кузнецова 1988], показал, что *словарные дефиниции — независимо от их качества — лишь полуфабрикат для семантизации лексики*.

Опыт уральских лингвистов [Кузнецова 1988; Бабенко 1999] позволил увидеть: ориентация на толкующие глаголы при семантизации глагольной лексики порой приводит к созданию чисто грамматических (фазовых или каузативных) группировок глаголов, весьма разнородных по своей лексической семантике.

Поскольку слово или словосочетание текста на любом языке может быть семантизировано по-русски, мы в перспективе получаем инструмент семантического анализа текстов на любых языках и соответственно — анализа лексико-семантических систем любого языка.

Автоматическая семантизация иноязычного текста существенно облегчается, если он входит в корпус параллельных текстов, одним из которых является русский. В таком случае исследователю останется установить соответствие между входным и русским словом и связать это слово со словарём.

Выделение в дефинициях метаслов-идентификаторов может быть частично формализовано. Так, в сочетании «глагол+существительное» идентификатором, как правило, является существительное. В конструкциях типа «*Молочные железы женщины*», «*Физиологическое состояние человека*», «*Непрерывное движение крови*» надо выбирать идентификатором последнее слово — «женщина», «человек», «кровь». В конструкции «*Совокупность жизненных отправлений организма*» — «организм».

Опыт [Караулов 1982] свидетельствует также о необходимости создания (с опорой на имеющиеся) особого типа дефиниций и особого метаязыка (на базе русского), ориентированных на компьютерный анализ и приспособленных для него: предполагается снятие неоднозначности (асимметрии) единиц — как в виде омонимии-полисемии, так и в виде синонимии. Этот метаязык должен быть ориентирован скорее на компьютер, чем на человека. Хорошая дефиниция — та, что автоматически приводит к верной классификации входной ЛСЕ.

В качестве базовых предполагается взять дефиниции МАС-2 с корректировкой по БТС и БАС-3. многоступенчатые толкования МАС предполагается обрабатывать следующим образом: например, в случае глагола *идти* (первое значение: «Передвигаться, перемещаться в пространстве»): а) каждая отдельная дефиниция принимается за отдельное значение; б) идентификаторами оказываются существительные: *ноги, средства передвижения, почтовые отправления/грузы, облака, вода, воздух, льдины, плоты, бревна*; в) для значения «г) Перемещаться массой, потоком, вереницей. 1) О движении облаков, воды, воздуха и т. п. 2) О движении льдин, плотов, бревен и т. п. по воде. || Передвигаться стаей, косяком и т. п. (о рыбе, мелком пушном звере)» интерпретация двуступенчата: рыба -> косяк -> движение косяка; лемминг -> стая леммингов -> движение стаи леммингов. Исходное — льдина, плот, бревно, рыба или мышь, к которой применяется семантическая функция «множество», а затем к данному конкретному множеству применяется семантическая функция «движение».

Для компьютеризованного построения лексико-семантического пространства русского языка (а в перспективе и других языков) в виде ориентированного графа необходимо:

- снять неоднозначность единиц метаязыка, представив их в виде: *1лемма1 <номер омонима, лемма, номер значения>* (синонимия метаязыка может быть устранена целенаправленной редукцией синонимических рядов до их доминанты);
- каждой ЛСЕ дать дефиницию;
- в дефиниции выделить метаслово (по сложно формализуемому принципу — сохранению важнейшего лексического значения). Именно это метаслово будет служить определяемым следующего уровня;
- повторять процесс до тех пор, пока последовательность (цепочка) подобных пар вида ЛСЕ — словарная дефиниция не дойдёт до одного из финальных узлов или пока не возникнет противоречие с уже введёнными данными. Одним из признаков, позволяющих определить, что конкретная ЛСЕ является финальной, служит появление цикла в толковании: ЛСЕ в конечном итоге толкуется сама через себя или через ЛСЕ, отличающуюся лишь номером значения (например, *1существо2 (то, что существует как живой организм, **животное**) — 1животное1 (живое **существо**, способное чувствовать и передвигаться) — 1существо1)*;
- для этого последнего узла словарная дефиниция может и не быть заданной, что приближает концы цепочек к неопределяемым понятиям математики.

Каждая такая последовательность (цепочка) ЛСЕ является путём в ориентированном графе, представляющем лексико-семантическое пространство языка, а само ЛСП строится как объединение таких путей, заданных пользователем на основе анализа словарных дефиниций.

Исходная задача программы ПроСеКа (ПРОграмма СЕмантической Классификации) — служить инструментом создания, проверки и сохранения цепочек ЛСЕ, заданных пользователем, т.е. фактически быть редактором этих цепочек. Программа написана на языке C++ в среде Borland C++ Builder. Основная задача программы — сохранять цепочки и текстовые примеры к ним, введённые пользователем, и создавать словарь всех встретившихся в этих цепочках ЛСЕ. Кроме того, в функции программы входит контроль за согласованностью данных: каждая ЛСЕ может иметь не более одной дефиниции (финальные узлы цепочек могут не иметь дефиниции), и последовательность ЛСЕ во всех цепочках должна сохраняться неизменной. Кроме того, на множество цепочек могут накладываться дополнительные ограничения, например, пользователь может ограничить длину цепочек.

Цепочки могут создаваться как в программе (в Мастере цепочек или в редакторе, позволяющем вводить несколько цепочек одновременно), так и в текстовом редакторе или электронной таблице, в виде текстовых файлов с разделителями. На Рис. 1 приведён пример цепочки, построенной пользователем в Мастере цепочек для ЛСЕ *ездить*:

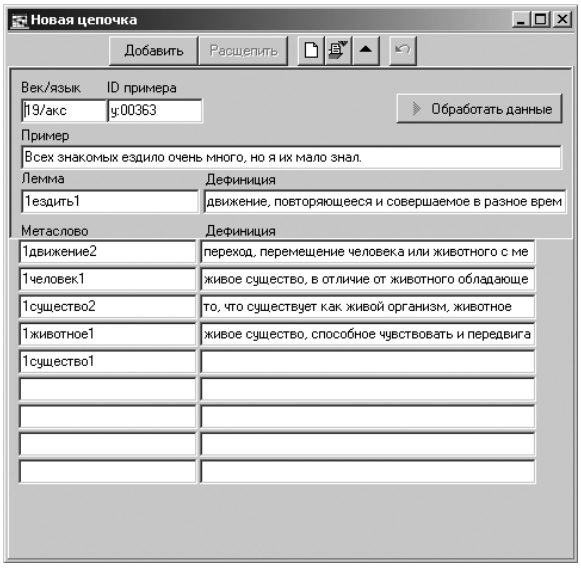


Рис. 1. Режим редактирования цепочки в Мастере цепочек

Создание цепочек частично автоматизировано. Так, если в цепочке используется слово, уже внесённое в словарь, продолжение цепочки достраивается автоматически.» В том случае, когда пользователь вводит многозначное слово, одно или несколько значений которого уже содержатся в словаре, программа предлагает пользователю достроить цепочку по уже имеющимся данным или ввести новое значение рассматриваемого слова. Выбор нужного значения — дело пользователя, а не программы, задачи программы сводятся к следующему: 1) предложить варианты автоматического продолжения цепочек для тех слов (значений), которые уже содержатся в словаре, и 2) проверить соответствие каждой новой цепочки установленным правилам (т.е. выполнение наложенных ограничений). Помимо существенной экономии сил и времени, такой режим исключает ошибки ручного ввода, в том числе несогласованность различных цепочек, ошибок и опечаток в дефинициях и т.п.

На множество цепочек накладываются следующие ограничения:

- Ограничение на длину. Цепочка должна содержать хотя бы два узла, максимальная длина цепочки может быть ограничена пользователем (по умолчанию ограничение отсутствует);
- Ограничение на единственность цикла. В цепочке не может присутствовать более двух одинаковых узлов, т. е. допустимо появление не более одного слова, толкующегося через себя само (в том же или другом значении). При этом второе вхождение данного слова служит сигналом конца цепочки;
- Ограничение на единственность толкования. Каждое значение слова или словосочетания может иметь не более одного толкования (финальные узлы могут не иметь толкования вовсе). Поскольку используемые словарные толкования редактируются, представляется достаточно важным, чтобы в дальнейшем не возникло расхождений между данными, внесёнными в различное время.
- Ограничение на согласованность. Множество цепочек должно быть согласованным, т. е. от каждого значения слова существует ограниченное число путей к финальному узлу или узлам, за исключением случаев толкования его через себя или другое значение этого же слова.

Первые два ограничения проверяются для каждой конкретной цепочки отдельно; третье и четвертое, очевидно, требуют сравнения новой цепочки с уже введёнными данными.

На Рис. 2 приведён пример одного из вариантов автоматического дополнения, предложенного программой. Исходная цепочка, созданная при помощи программы MS Excel, противоречила введённым ранее данным (см. Рис. 3).

Результат обработки введённых пользователем данных помещается в файлы данных (в настоящее время это текстовые файлы с разделителями, что позволяет просматривать и анализировать результат работы программы в электронной таблице, в дальнейшем возможен переход на другой или другие форматы хранения данных). Кроме того, результат выводится на экран (Рис. 4).

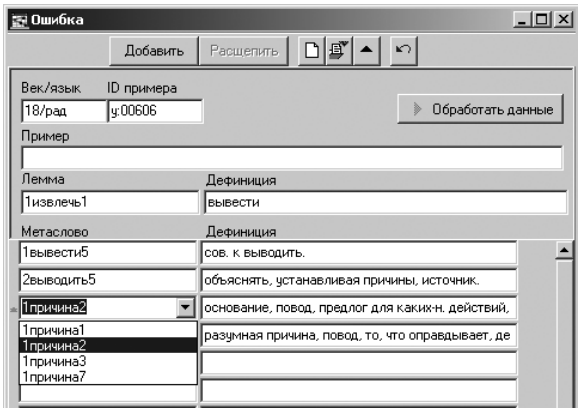


Рис. 2. Автоматическая проверка и автоматическое дополнение цепочки

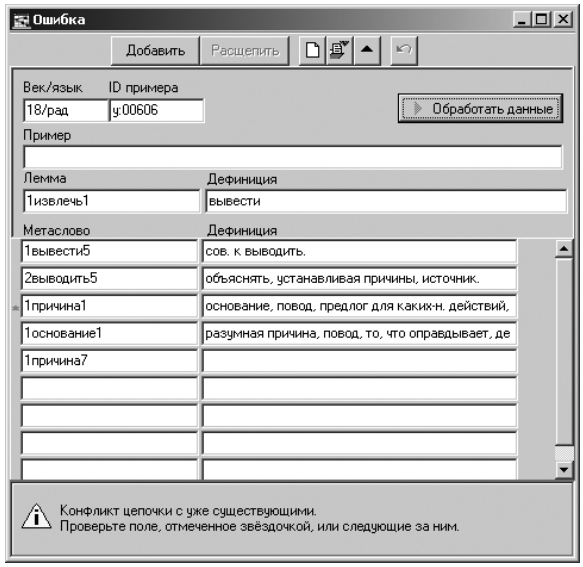


Рис. 3. Результат автоматической проверки цепочки: сообщение об ошибке

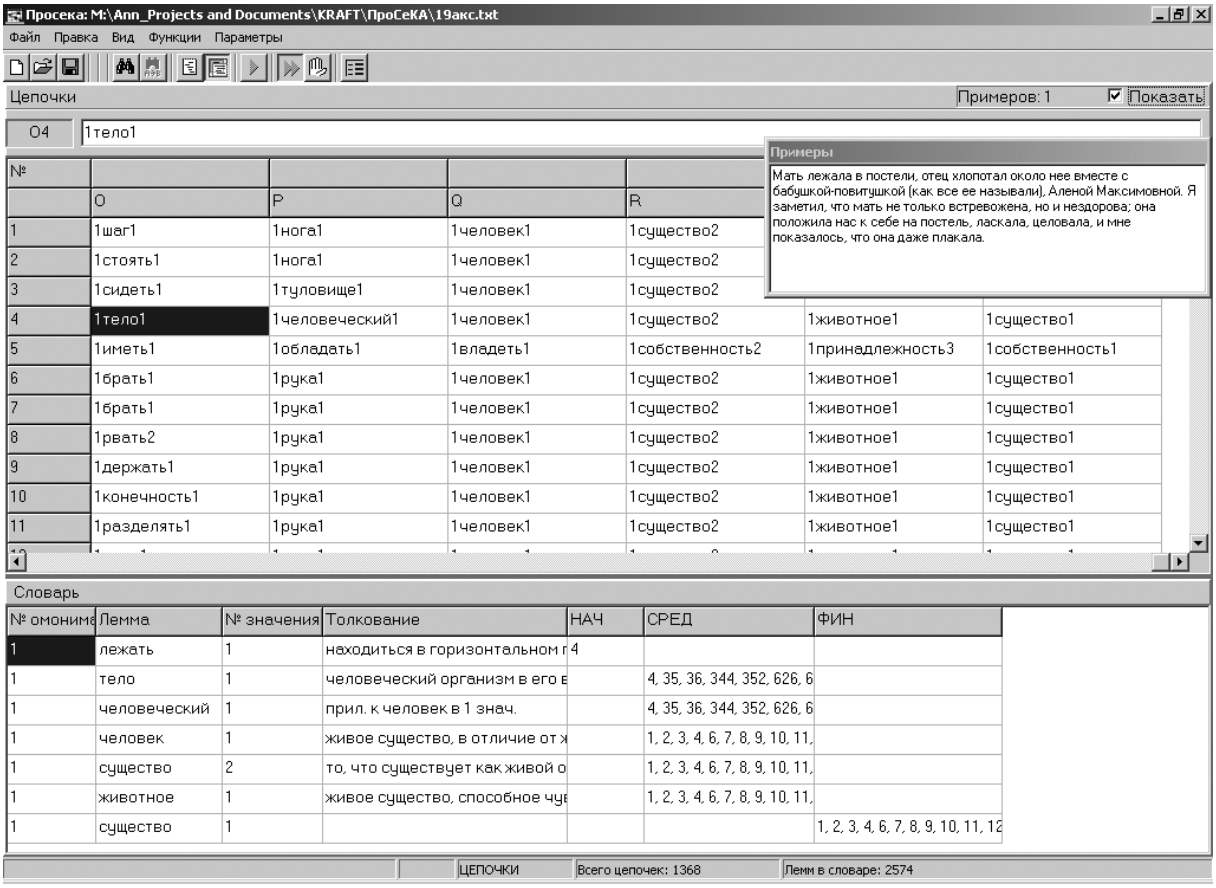


Рис. 4. Просмотр цепочек и словаря

В верхней части экрана даны все существующие в настоящее время цепочки, выровненные по правому краю, а в нижней — все ЛСЕ, входящие в выбранную цепочку, причём показаны не только сами ЛСЕ, но и часть словарной статьи (дефиниция, а также идентификаторы цепочек, в которые входит заданная ЛСЕ). В отдельном окне показываются все примеры, приписанные к выбранной цепочке. Пользователь может включать и отключать просмотр примеров.

Работа с программой выявила необходимость внесения следующих дополнений:

1) Возможность задавать произвольное количество наборов правил для проверки цепочек. Как уже говорилось, ЛСП строится эмпирически, как множество допустимых «путей» в этом

пространстве. При этом некоторые первоначальные предположения о виде ЛСП не подтвердились, другие требуют дальнейшей проверки. Эта возможность в настоящее время реализована на уровне абстрактного класса правил и ряда конкретных правил;

2) Возможность изменять вид и представление данных (абстракция данных). В настоящее время программа фактически хранит не только сами созданные пользователем цепочки, но и порядок их построения. На первоначальном этапе построения ЛСП эта возможность, безусловно, полезна, однако в дальнейшем, особенно при переходе к изучению вида ЛСП, она едва ли окажется нужной;
3) Дополнительные возможности по классификации, сортировке и обработке данных. В частности, сейчас цепочки располагаются в том порядке, в каком они введены в программу, что является не самым удачным способом хранения с точки зрения быстродействия программы и удобства просмотра.
4) Возможно, целесообразным окажется просмотр ЛСП не только в виде цепочек, построенных пользователем, но и в виде графа, описывающего допустимые переходы между узлами.

Реализация этих возможностей — дело будущего.

Литература

1. *Азарова И. В., Синопальникова А. А., Яворская М. В.* Принципы построения wordnet-тезаруса RussNet. — http://www.dialog-21.ru/Archive/2004/Sinopalnikova.htm.
2. *Алексеев П. М.* Семантические частотные словари // Статистика речи и автоматический анализ текста. 1972. — Л., 1973, с.20-36.
3. *Бабенко Л. Г.* (ред.) Толковый словарь русских глаголов: Идеографическое описание. Английские эквиваленты. Синонимы. Антонимы. — М., 1999.
4. *Баранов О. С.* Идеографический словарь русского языка / О. С. Баранов. — М.,1995.
5. *БАС-3* Большой академический словарь русского языка, тт. 1–10. — М.-СПб, 2004–2008.
6. *БТС-1998* Большой толковый словарь русского языка / под ред. С. А. Кузнецова, — СПб, 1998.
7. *Гамкрелидзе Т. В.* Индоевропейский язык и индоевропейцы: Реконструкция и историко-типологический анализ праязыка и протокультуры / Гамкрелидзе Т. В., Иванов Вяч. Вс. — Тбилиси, 1984.
8. *Задорожный М. И.* Два подхода к построению идеографического словаря // О преподавании русского языка и литературы в киргизской школе. Вып. 10. — Фрунзе, 1983.
9. *Караулов Ю. Н.* Общая и русская идеография. М., 1976.
10. *Караулов Ю. Н.* Лингвистическое конструирование и тезаурус литературного языка, М., 1981.
11. *Караулов Ю. Н.* Русский семантический словарь. Опыт автоматического построения тезауруса: от понятия к слову / Ю. Н. Караулов, В. И. Молчанов, В. А. Афанасьев, Н. В. Михалев. — М.: Наука, 1982.
12. *Кобозева И. М.* Лингвистическая семантика. — М.: Удиториал УРСС, 2000.
13. *Кузнецова Э. В.* Метод ступенчатой идентификации в описании лексико-семантических групп слов. //Учен. зап. / Тартус. ун-т. — 1969. — Вып. 228.
14. *Кузнецова Э. В.* (ред.) Лексико-семантические группы русских глаголов. — Свердловск, 1988.
15. *Лукашевич Н. В., Добров Б. В.* Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Труды Международного семинара Диалог 2002 по компьютерной лингвистике и ее приложениям, Т. 2. — М.: Наука, 2002.
16. *МАС-2* Словарь русского языка в 4-х тт. / под ред. А. П. Евгеньевой, Изд. 2-ое, испр. и доп. — М., 1981–1984.
17. *Морковкин В. В.* (ред.) Лексическая основа русского языка. Комплексный учебный словарь. — М., 1984.
18. *Морковкин В. В.* Основы теории учебной лексикографии. Дисс. в форме науч. докл. .. докт. филол. наук. М., 1990.
19. *Паллас П. С.* Сравнительные словари всех языков и наречий, собранные десницею всевысочайшей особы. Отделение перьвое, содержащее в себе Европейские и Азиатские языки. Ч. 1–2, СПб, 1787–1789.
20. *ПрокоповВ.В.* Основныелексико-семантические группы русского глагола. Дисс. … канд. филол. наук, — Самарканд, 1945. — 197 с.
21. *Руделёв В. Г.* Вначале было слово / Руделёв В. Г., Руделёва О. А. — Тамбов , 1995.
22. *Шведова Н. Ю.* (ред.) Русский семантический словарь, тт. I–IV. — М., 1998–2007.
23. *Шишков А. С.* Собрания языков и наречий с примечаниями на оныя. //Собр. соч. и переводов Адмирала Шишкова, Ч. XV, СПб, 1832.
24. *Щерба Л.*В. Языковая система и речевая деятельность. Л.: Наука, Ленингр. отд-ние, 1974.
25. *Fellbaum, C.* (1998), ed. «WordNet: An Electronic Lexical Database». MIT Press, Cambridge, MA.
26. *Lorge Irving, Thorndike Edvard E.* A Semantic Count of English Words. — New York, 1938.
27. *Roget* 1986: The Penguin Roget`s thesaurus of English words and phrases /New edition copletely revized, updated and abridged by Susan M. Lloyd, Pengin books, 1986. — 776 p.

Квазикорпусное изучение словарной продуктивности лексико-синтаксических разрядов в русском языке по словарю С. И. Ожегова[1]

«Quasi-corpus» investigation of lexical productivity of non-trivial basic diatheses of Russian with special regard to S. I. Ozhegov's dictionary of russian

Крылов С. А. (krylov-58@mail.ru)

Институт востоковедения РАН, Институт системного анализа РАН

«Квазикорпусная» лингвистика исследует не только первичные, но и вторичные источники информации о языке (напр., грамматики и словари). Изучена статистика лексико-синтаксических разрядов (подклассов частей речи, отличающихся лексико-синтаксическими пометами) в словаре С. И. Ожегова 1989 г.

В отличие от «корпусной» лингвистики в узком смысле слова, «квазикорпусная» лингвистика (составляющая часть корпусной лингвистики в широком смысле слова) допускает исследование статистики употребления некоторых единиц в составе источников не только первичного характера, но также и вторичного (метаязыкового) характера, в частности, в составе текстов реально существующих грамматик и словарей.

Одной из таких задач стало изучение статистики грамматических помет в словаре С. И. Ожегова 1989 г. (далее СО-1989). В рамках данной задачи можно выделить более узкую — подсчёт частотности лексико-синтаксических помет (далее ЛСП).

На основе помет СО-1989 был создан инвентарь условных маркеров, включающих наряду с исходной ЛСП некоторые обобщённые символы со значением «синтаксических форм», а именно:

A (беспредложный вин. п.)

C (содержание мысли)

D (беспредложный дат. п.)

E (элативная форма)

F (лативная форма)

G (беспредложный род. п.)

I (беспредложный твор. п.)

Inf (инфинитив)

L (эссивная форма)

N (нуль)

R («прочие» синтаксические формы)

S (комитативная форма)

T (делиберативная форма)

V (маршрут //транслативная форма).

С помощью интегрированной информационной среды StarLing, созданной С. А. Старостиным [2], была создана специальная лексическая база данных на основе СО-1989. Количество лексических единиц (словозначений) в этой базе составляет 64346 (для простоты ЛСП фразем и идиом не учитывались).

---

[1] Настоящее исследование выполнено при поддержке РГНФ (гранты № 08-04-00190а, № 08-04-12126в, № 08-04-12132в, № 07-04-00161а).

[2] См.: *Крылов С. А.* Стратегии применения интегрированной информационной среды StarLing в корпусной лингвистике и в компьютерной лексикографии // Смирнов И. С. (ред.). Orientalia et classica. Труды Института восточных культур и античности. Выпуск XIX. Аспекты компаративистики. 3. М.: РГГУ, 2008. — С. 649–668.

Идентификация лексико-синтаксических разрядов (ЛСР) производилась на основании сочетания «общекатегориальной» пометы (далее ОКП), указывающей на часть речи или на некоторый грамматический подкласс внутри этой части речи, с ЛСП, указывающей на «исходную диатезу», то есть на тот тип соотношения между смысловыми и синтаксическими валентностями данной лексической единицы (далее ЛЕ), который присущ ей в синтаксически немаркированном употреблении [3].

Диатезы (в других терминах, «модели управления» или «конструкции») бывают исходные и производные (трансформированные). Исходные диатезы (ИД) бывают тривиальные и нетривиальные.

Тривиальные ИД непосредственно выводятся из ОКП, а потому в словарях задаются фигурой умолчания; как правило, это абсолютные (непереходные) ИД. Их словарная продуктивность максимальна (см. в табл. 1 ЛСР №№ 1–3, 6–7, 9, 12, 14, 20–22, 24–25, 28, 38, 41, 43, 47, 49, 52, 68–69, 78, 87, 89, 97–98, 110, 119–120, 122–123, 125, 149, 157, 178–181, 187 и т. п.).

Содержательно каждая ЛСП отражает некоторую нетривиальную ИД. Нетривиальные ИД, зафиксированные в ЛСП, имеют меньшую словарную продуктивность (см. в табл. 1 ЛСР №№ 4–5, 8, 10–11, 13, 15–19, 23, 26–27, 29–30 и т. п.).

Всего в СО-1989 зафиксировано 980 ЛСР (в ранговом словаре они упорядочены от № 1 до № 980).

Столбец A указывает на ранги; столбец B — на относительную продуктивность ЛСР (т. е. результат деления абсолютной продуктивности на 6,4346); столбец C — на продуктивность ЛСР в абсолютном измерении (количество ЛЕ данного ЛСР). Столбец D указывает на условные маркеры ЛСР.

Табл. 1. Список 246 ЛСР, имеющих относительную продуктивность выше 0.31 (а абсолютную — выше 2).

A ранг	B доля на 10 000	C абсолютное число	D обозначение ЛСР
1	4469,90	28762	сущ.
2	1606,32	10336	глаг.
3	1526,75	9824	прил.
4	730,58	4701	**глаг. A [*что*]**
5	444,63	2861	**глаг. A [*кого-что*]**
6	144,53	930	нареч.
7	115,47	743	первая часть сложных слов
8	100,39	646	**глаг. A [*кого (что)*]**
9	47,87	308	приставка
10	42,43	273	**глаг. A [*что*] // G [*чего*]**
11	30,30	195	**сущ. G [*чего*]**
12	27,97	180	частица
13	27,35	176	**глаг. I [*чем*]**
14	20,51	132	в знач. сказ.
15	20,05	129	**глаг. S [*с кем*]**
16	17,72	114	**глаг. F [*на кого-что*]**
17	16,32	105	**глаг. F [*во что*]**
18	15,23	98	**глаг. D [*кому*]**
19	13,99	90	**предлог с род. п. G [*кого-чего*]**
20	13,21	85	вводн. сл.
21	12,59	81	мест. нареч.
22	11,50	74	союз
23	10,72	69	**глаг. S [*с кем-чем*]**
24	10,57	68	неизм.
25	10,26	66	прил. притяж.
26	9,79	63	**глаг. I [*кем-чем*]**
27	9,64	62	**глаг. D [*кому-чему*]**
28	9,64	62	межд.
29	9,32	60	**глаг. G [*чего*]**
30	9,17	59	глаг. R [Inf]
31	8,70	56	**глаг. A [*что*] & F [*во что*]**
32	8,55	55	**глаг. F [*к кому-чему*]**
33	8,55	55	**предлог с род. п. G [*чего*]**
34	6,99	45	**глаг. L [*в чём*]**
35	6,53	42	**глаг. E [*от кого-чего*]**
36	6,37	41	глаг. N [Impers.!]
37	6,22	40	**глаг. F [*на что*]**
38	6,22	40	числит. колич.
39	5,91	38	**глаг. A [*что*] & I [*чем*]**
40	5,75	37	**глаг. A [*что*] & D [*кому*]**
41	5,59	36	**мест. неопр.**
42	5,44	35	**глаг. E [*от чего*]**
43	5,44	35	сущ. обращение
44	5,28	34	**глаг. G [*кого-чего*]**
45	4,97	32	**глаг. F [*на кого (что)*]**
46	4,82	31	**глаг. S [*с чем*]**
47	4,82	31	мест. определит.
48	4,51	29	**сущ. G [*кого-чего*]**
49	4,20	27	числит. собир.
50	3,89	25	**глаг. F [*к чему*]**
51	3,73	24	**глаг. F [*до чего*]**
52	3,73	24	межд. звукоподр.
53	3,57	23	**глаг. F [*в кого-что*]**
54	3,26	21	в знач. сказ. R [Inf]
55	3,26	21	глаг. C []
56	3,26	21	**предлог с тв. п. I [*кем-чем*]**
57	3,11	20	**глаг. A [*что*] // R [Inf]**

[3] Синтаксически немаркированное употребление глагола обычно представляет собой его употребление в составе так называемой «ядерной конструкции» (см.: Апресян Ю. Д. Экспериментальное исследование семантики русского глагола. М., Наука, 1967, с. 48).

A ранг	B доля на 10 000	C абсолютное число	D обозначение ЛСР
58	3,11	20	**предлог с вин. п. A [*кого-что*]**
59	3,11	20	**сущ. G [*кого*]**
60	2,95	19	**глаг. F [*к кому*]**
61	2,95	19	**глаг. L [*над кем-чем*]**
62	2,95	19	**глаг. N [Impers.!] & D [*кому*]**
63	2,80	18	**глаг. A [*кого-что*] & I [*чем*]**
64	2,80	18	**глаг. F [*за кого-что*]**
65	2,80	18	предлог с вин. и предл. п. A [] // L []
66	2,64	17	мест. нареч. и союзн. сл.
67	2,64	17	предлог с дат. п. D []
68	2,64	17	числит. порядк.; сущ.
69	2,49	16	мест. указат.
70	2,49	16	**предлог с вин. п. A [*что*]**
71	2,33	15	**глаг. A [*что*] & S [*с чем*]**
72	2,33	15	**глаг. A [*что*] // T [*о чём*]**
73	2,18	14	**в знач. сказ. D [*кому*]**
74	2,18	14	**глаг. A [*кого-что*] & F [*во что*]**
75	2,18	14	**глаг. G [*кого (чего)*]**
76	2,18	14	**глаг. I [*кем*]**
77	2,18	14	сущ. Adj
78	2,18	14	числит. порядк.; прил.; сущ.
79	2,02	13	**глаг. A [*кого (что)*] & I [*чем*]**
80	1,86	12	**глаг. N [Impers.!] & A [*кого-что*]**
81	1,86	12	**предлог с тв. п. I [*чем*]**
82	1,71	11	**глаг. A [*кого-что*] // G [*чего*]**
83	1,71	11	**глаг. A [*кого (что)*] & F [*во что*]**
84	1,71	11	**глаг. A [*что*] & F [*на кого-что*]**
85	1,71	11	**глаг. D [*чему*]**
86	1,71	11	**глаг. L [*за кем-чем*]**
87	1,71	11	мест. притяж.
88	1,71	11	**предлог с дат. п. D [*кому-чему*]**
89	1,71	11	союз и частица
90	1,55	10	**глаг. A [*кого-что*] & F [*во что*]**
91	1,55	10	**глаг. A [*кого-что*] & D [*кому*]**
92	1,55	10	**глаг. F [*за что*]**
93	1,55	10	**глаг. S [*с чем*] // R [Inf]**
94	1,55	10	нареч. и предлог с род. п.
95	1,40	9	**глаг. A [*что*] // F [*через что*]**
96	1,40	9	**глаг. T [*о ком-чём*]**
97	1,40	9	мест. отриц.
98	1,24	8	возглас
99	1,24	8	**глаг. A [*кого*]**
100	1,24	8	**глаг. A [*кого (что)*] & F [*во что*]**
101	1,24	8	**глаг. A [*кого (что)*] & I [*чем*]**
102	1,24	8	**глаг. A [*что*] & F [*на что*]**
103	1,24	8	**глаг. A [*что*] // I [*чем*]**
104	1,24	8	**глаг. E [*из чего*]**
105	1,24	8	**глаг. F [*до кого-чего*]**
106	1,24	8	**глаг. G [*кого-чего*] // R [Inf]**
107	1,24	8	**глаг. L [*в ком-чём*]**
108	1,24	8	**глаг. N [Impers.!] & A [*кого (что)*]**
109	1,24	8	глаг. N [Impers.!] & R [Inf]
110	1,24	8	мест. вопросит. и союзн. сл.
111	1,24	8	**предлог с вин. и тв. п. A [*кого-что*] // I [*кем-чем*]**
112	1,09	7	**глаг. A [*кого-что*] & D [*кому*]**
113	1,09	7	**глаг. A [*что*] & F [*к чему*]**
114	1,09	7	**глаг. E [*с кого-чего*]**
115	1,09	7	**глаг. E [*с чего*]**
116	1,09	7	**глаг. F [*на что*] // R [Inf]**
117	1,09	7	**глаг. G [*чего*] // R [Inf]**
118	1,09	7	**глаг. N [Impers.!] & D [*кому*] & R [Inf]**
119	1,09	7	нареч.; сущ.
120	1,09	7	пов.
121	1,09	7	**сущ. F [*на что*]**
122	1,09	7	числит. неопр.-колич.
123	1,09	7	числит. порядк.
124	0,93	6	**в знач. сказ. D [*кому*] & R [Inf]**
125	0,93	6	вводн. сл. и частица
126	0,93	6	**глаг. A [*кого-что*] & D [*кому-чему*]**
127	0,93	6	**глаг. A [*кого-что*] & S [*с кем-чем*]**

A ранг	B доля на 10 000	C абсолютное число	D обозначение ЛСР
128	0,93	6	**глаг. A [*что*] & E [*из чего*]**
129	0,93	6	**глаг. D [*кому*] & R [Inf]**
130	0,93	6	глаг. N [Impers.++]
131	0,93	6	**нареч. G [*кого-чего*]**
132	0,93	6	предлог с вин. п. A[]
133	0,93	6	**предлог с предл. п. L [*ком-чём*]**
134	0,93	6	**прил. D [*кому-чему*]**
135	0,78	5	**в знач. сказ. G [*кого-чего*]**
136	0,78	5	**в знач. сказ. G [*кого-чего*] // R [Inf]**
137	0,78	5	**глаг. A [*кого-что*] // (A [*что*] & F [*на кого-что*])**
138	0,78	5	**глаг. A [*кого-что*] // I [*чем*]**
139	0,78	5	**глаг. A [*кого-что*] // R [Inf]**
140	0,78	5	**глаг. A [*кого-что*] // устар. G [*кого-чего*]**
141	0,78	5	**глаг. A [*кого (что)*] & L [*в чём*]**
142	0,78	5	**глаг. A [*кого (что)*] F [*к чему*]**
143	0,78	5	**глаг. A [*что*] & D [*кому-чему*]**
144	0,78	5	**глаг. A [*что*] & F [*на кого (что)*]**
145	0,78	5	**глаг. A [*что*] // G [*кого-чего*]**
146	0,78	5	**глаг. A [*что*] // T [*о ком-чём*]**
147	0,78	5	**глаг. E [*от кого*]**
148	0,78	5	**глаг. F [*к чему*] // R [Inf]**
149	0,78	5	нареч. и в знач. сказ.
150	0,78	5	предлог с предл. п. L []
151	0,78	5	**предлог с предл. п. L [*чём*]**
152	0,78	5	**предлог с род. п. G [*кого*]**
153	0,78	5	**прил. F [*к чему*]**
154	0,78	5	**прил. F [*на что*] // R [Inf]**
155	0,78	5	**прил. G [*чего*] // I [*кем-чем*]**
156	0,78	5	**прил. I [*чем*]**
157	0,78	5	сравн.
158	0,78	5	сущ. (обычно в обращении).
159	0,62	4	**глаг. (A [*кого-что*] // I [*чем*]) & V [*по чему*]**
160	0,62	4	**глаг. A [*кого-что*] & F [со словами «*куда*», «*куда-то*», «*некуда*»]**
161	0,62	4	**глаг. A [*кого-что*] & F [*до чего*]**
162	0,62	4	**глаг. A [*кого-что*] // T [*о ком-чём*]**
163	0,62	4	**глаг. A [*кого (что)*] & (F[к чему] // R[Inf])**
164	0,62	4	**глаг. A [*кого (что)*] & (F [*на что*] // R [Inf])**
165	0,62	4	**глаг. A [*кого (что)*] & E [*от кого-чего*]**
166	0,62	4	**глаг. A [*кого (что)*] & R [Inf]**
167	0,62	4	**глаг. A [*кого (что)*] & F [*на что*]**
168	0,62	4	**глаг. A [*что*] & F [*в кого-что*]**
169	0,62	6	**глаг. D [*кому*] & R [Inf]**
170	0,62	4	**глаг. E [*с кого*]**
171	0,62	4	**глаг. F [*за что*] // R [Inf]**
172	0,62	4	**глаг. F [*на кого*]**
173	0,62	4	**глаг. L [*за кем-чем*] // (устар.) A [*что*]**
174	0,62	4	**глаг. N [Impers.!] & A [*что*]**
175	0,62	4	**глаг. N [Impers.!] & D [*кому*]**
176	0,62	4	**глаг. T [*о чём*]**
177	0,62	4	**глаг. V [*по кому-чему*]**
178	0,62	4	звукоподр.
179	0,62	4	мест. личн. 1 л. мн. ч.
180	0,62	4	мест. личн. 3 л.
181	0,62	4	нареч. и союз
182	0,62	4	**предлог с дат. п. D [*чему*]**
183	0,62	4	**предлог с род. п. G [*кого (чего)*]**
184	0,62	4	**прил. F [*к кому-чему*]**
185	0,62	4	**прил. S [*с чем*]**
186	0,62	4	прил. в знач. сказ. R [Inf]
187	0,62	4	сущ.; нареч.; неизм.
188	0,62	4	сущ. (обычно в обращении)
189	0,62	4	**сущ. G [*чего*] // Adj [*какая*]**
190	0,62	4	**сущ. L[*в чём*]**
191	0,62	4	частица С глаголами прошедшего времени образует сослагательное наклонение

A ранг	B доля на 10 000	C абсолютное число	D обозначение ЛСР
192	0,62	4	частица и в знач. сказ.
193	0,47	3	**в знач. сказ. D [*кому-чему*]**
194	0,47	3	**в знач. сказ. D [*кому*] & R [Inf]**
195	0,47	3	в знач. сказ. и нареч.
196	0,47	3	восклицание
197	0,47	3	**глаг. A [*кого-что*] & D [*чему*]**
198	0,47	3	**глаг. A [*кого-что*] & F [*к кому-чему*]**
199	0,47	3	**глаг. A [*кого-что*] & E [*от кого-чего*]**
200	0,47	3	**глаг. A [*кого-что*] & E [*от чего*]**
201	0,47	3	**глаг. A [*кого-что*] & F [*за кого-что*]**
202	0,47	3	**глаг. A [*кого-что*] & F [*на кого-что*]**
203	0,47	3	**глаг. A [*кого-что*] & S [*с кем-чем*]**
204	0,47	3	**глаг. A [*кого (что)*] & F [*на что*]**
205	0,47	3	**глаг. A [*кого (что)*] & T [*о чём*]**
206	0,47	3	**глаг. A [*кого (что)*] & E [*от кого-чего*]**
207	0,47	3	**глаг. A [*кого (что)*] & E [*от чего*]**
208	0,47	3	**глаг. A [*кого (что)*] & F [*за что*]**
209	0,47	3	**глаг. A [*кого (что)*] & I [*кем*]**
210	0,47	3	**глаг. A [*что*] & E [*от чего*]**
211	0,47	3	**глаг. A [*что*] & F [*подо что*]**
212	0,47	3	**глаг. A [*что*] & I [*кем-чем*]**
213	0,47	3	**глаг. A [*что*] & S [*с кем*]**
214	0,47	3	**глаг. A [*что*] // A [*что*] & S [*с чем*]**
215	0,47	3	**глаг. A [*что*] // C [с союзом «*что*»]**
216	0,47	3	**глаг. A [*что*] // F [*на что*]**
217	0,47	3	**глаг. A [*что*] // V [*через что*]**
218	0,47	3	**глаг. D [*кому*] & (A [*что*] // R [Inf])**
219	0,47	3	**глаг. D [*кому*] & A [*что*]**
220	0,47	3	**глаг. D [*кому*] // F [*к кому*]**
221	0,47	3	**глаг. D [*кому*] // L [*перед кем*]**
222	0,47	3	**глаг. E [*от чего*] // R [Inf]**
223	0,47	3	**глаг. F [*во что*] // R [Inf]**
224	0,47	3	**глаг. F [*подо что*]**
225	0,47	3	**глаг. G [*кого-чего*] // A [*что*]**
226	0,47	3	**глаг. L [*на чём*]**
227	0,47	3	**глаг. L [*перед кем-чем*]**
228	0,47	3	**глаг. N [Impers.!] & A [*кого-что*]**
229	0,47	3	**глаг. S [*с кем-чем*] & L [*в чём*]**
230	0,47	3	**глаг. S [*с кем (чем)*]**
231	0,47	3	**глаг. T [*о ком-чём*] // C[с союзом «*что*»]**
232	0,47	3	мест. вопросит.
233	0,47	3	мест. нареч., с последующим отрицанием.
234	0,47	3	мест. указат. и определит.
235	0,47	3	**нареч. G [*чего*]**
236	0,47	3	**нареч. S [*с кем-чем*]**
237	0,47	3	**предлог с вин. и твор. п. A [*что*] // I [*чем*]**
238	0,47	3	**предлог с вин. п. F [*кого-что*]**
239	0,47	3	**предлог с вин. п. F [*что*]**
240	0,47	3	**прил. S [*с кем-чем*]**
241	0,47	3	сущ.; неизм.
242	0,47	3	**сущ. D [*кому*]**
243	0,47	3	**сущ. F [*на кого-что*]**
244	0,47	3	**сущ. Poss [*чей*] // G [*кого*]**
245	0,47	3	частица при вопросе
246	0,47	3	числит.

Остальные 734 ЛСР (№№ 247–980) имеют ещё меньшую словарную продуктивность (относительную менее 0.47, а абсолютную менее 3).

В числе уникальных ЛЕ (входящих в одноэлементные ЛСР, то есть с относительной словарной продуктивностью 0.31, а абсолютной 1) оказались представители разных частей речи и грамматических разрядов, а именно: 321 — глаголы; 55 — существительные; 40 — частицы; 35 — прилагательные; 28 — наречия; 23 — союзы; 21 — неглагольные предикативы (помета «в знач. сказ.»); 21 — местоимения; 15 — местоименные наречия; 13 — предлоги; 12 —

междометия; 5 — вводные слова; 4 — числительные; 2 — связки; 2 — сравн.; 1 — союзное слово; 1 — субст. мн.; 1 — нескл.; 1 — возглас; 1 — восклицание; 1 — зват.; 1 — окрик; 1 — ответ; 1 — отклик; 1 — уверение; 1 — форма пов.; 1 — форма 3 л. ед.ч.; 1 — форма 3 л. мн.ч.; 1 — форма род. п. в знач. притяж.

Классификация ЛЕ может основываться не на простом пронумерованном перечне ЛСР, а на определённом ограниченном наборе синтаксических дифференциальных признаков (СДП), характеризующих каждый ЛСР. Такой набор складывается из двух частей.

Во-первых, это (1) ОКП (включая указание (1А) на «часть речи» данной ЛЕ и (1Б) на тот «внутричастеречный» разряд этой ЛЕ, который зафиксирован в СО-1989).

Во-вторых, это (2) ЛСП (характеристика ИД), складывающаяся из (2А) пометы о безличности/личности данной ЛЕ; (2Б) РП (характеристика типа управления). В РП входят следующие характеристики (СДП):

(2Б.1) синтаксическая абсолютность/релятивность (отсутствие/наличие дополнений);

(2Б.2) общее количество дополнений;

(2Б.3) указание на форму выражения каждого дополнения, то есть на синтетический падеж (A, G, D, I), на предложно-падежную форму (F, L, E, S, T), на инфинитив (Inf), на прилагательное (Adj: *какой*; в том числе притяжательное — Poss: *чей*), на «пропозитив» (изъяснительное придаточное предложение) (C: с союзом *что*). Среди предложно-падежных форм выделяются эссивные (L: *на кого-что*, *в кого-что*, *под кого-что*, *за кого-что*, *перед кем-чем*), лативные (F: *на кого-что*, *в кого-что*, *к кому-чему*, *под кого-что*, *до кого-чего*), элативные (E: *от кого-чего*, *из кого-чего*, *с кого-чего*,), транслативные (V: *через кого-что, по кому-чему*), комитативные (S: *с кем-чем*), делиберативные (T: *о ком-чём*) [4].

(2Б.4) указание на семантическую одушевлённость (= «одухотворённость») дополнений. Этот признак принимает 4 значения:

(а) отчётливая «одухотворённость»: A (*кого*) и т п.
(б) тенденция к «одухотворённости»: A (*кого* (*что*) и т п.
(в) нейтральность (диффузность, немаркированность): A (*кого-что*) и т п.;
(г) отчётливая «неодухотворённость»: A (*что*) и т п.

Из таблицы 1 (в её полном варианте) нетрудно получить данные о словарной продуктивности любого из этих СДП и любого пучка нескольких СДП (При этом между СДП, входящими в один пучок, можно задавать отношения конъюнкции и дизъюнкции; любой СДП может быть подвергнут логическому отрицанию).

Отдельной интересной задачей было бы соотнесение рангов словарной продуктивности ЛСР с рангами их текстовой частотности, что требует уже применения собственно корпусных методов.

Ранги текстовой частотности ЛСР можно измерять несколькими способами с разной степенью погрешности.

«Корпусные» методы измерения базируются на обследовании достаточно представительного корпуса предварительно размеченных текстов, где каждому слову приписан некоторый ЛСР (разумеется, для измерения частотности реальных диатез каждой текстовой словоформе (каждому вхождению каждой словоформы в текст) следовало бы приписать ту реальную диатезу, в которой эта словоформа употреблена).

Достижение большей или меньшей точности в таком измерении напрямую зависит от того, производится ли разметка (1) вручную или (2) автоматически (ручная разметка обеспечивает более высокую точность измерения).

Автоматическая разметка может базироваться (2а) на моносемизированном корпусе (со снятой омонимией и со снятой полисемией), (2б) на частично моносемизированном корпусе (только со снятой омонимией, но без снятой полисемии) или (2в) на максимально «сыром» (не моносемизированном) корпусе (в котором не снята даже омонимия).

Степень точности (а значит, и адекватности) измерения падает в направлении от способа (1) (где она максимальна) к способу (2в) (где она ми-

[4] Категориальные признаки, лежащие в основе классификации синтаксических форм, заимствованы из разных источников, в том числе из номенклатуры пространственных значений в классификации А. Е. Кибрика (Кибрик А. Е. К типологии пространственных значений // Звегинцев В. А. (отв. ред.). Язык и человек. М.: Изд-во Моск. ун-та, 1970).
Некоторые из выделяемых классов имеют аналоги в классификации синтаксических форм Г. А. Золотовой (см.: Золотова Г. А. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. Изд. 3-е. М.: УРСС, с. 430–432), однако отличаются от синтаксических форм Г. А. Золотовой тем, что используемые в настоящей модели синтаксические формы выделяются не на семантической, а на сугубо формальной (морфологической) основе. Такой подход позволяет описывать систему предложно-падежных форм, являющихся выразителями тех или иных семантических ролей, безотносительно к тому, в какой степени решён вопрос об инвентаризации самих этих семантических ролей. Дело в том, что наиболее естественная номенклатура синтаксических форм, базирующаяся на их основных (главных, первичных) синтаксических функциях, по определению имеет право не принимать во внимание их вторичные функции. Применительно к системам синтетически выражаемых падежей такой «формалистический» подход является практически общепринятым, так что в данном случае речь идёт всего лишь о том, чтобы последовательно применять его к системам падежей, выраженных аналитически — сочетаниями слов-реляторов, или «прилогов» (в частности, в русском языке — предлогов) с формами синтетически выражаемых падежей (см.: Яхонтов С. Е. Классы глаголов и падежное оформление актантов // Храковский В. С. (отв. ред.). Проблемы теории грамматического залога. Л.: Наука, ЛО, 1978, с. 102–108).

нимальна); однако коэффициент трудоёмкости разметки, напротив, очевидным образом оказывается минимальной при выборе способа (2в) и достигает максимума при выборе способа (1).

Трудозатраты на разметку равны результату произведения коэффициента трудоёмкости разметки на объём исследуемого корпуса.

Было бы ошибочным полагать, что всякое измерение частотности ЛСР, обладающее менее чем 100%-ной адекватностью, является бесполезным. Ведь степень адекватности такого измерения находится в прямой зависимости от степени моносемичности разметки, а последняя зависит не только от качества самой разметки, но и от качества исследуемого корпуса: чем меньше доля омонимичных и полисемичных словоформ в исследуемом корпусе, тем более адекватным окажется и само измерение текстовой частотности ЛСР.

Измерения системной мощности (относительной словарной продуктивности) и текстовой мощности (относительной встречаемости в корпусе) ЛСР имеет разнообразные приложения. Результаты этих измерений могут быть использованы в нижеследующих сферах.

(1) Сопоставительная характеристика языков и квантитативная типология.

Так, во французском языке (как показал В. Г. Гак) встречаемость транзитивных диатез в речи сравнительно с интранзитивными оказывается значительно выше, чем в русском языке.

Синтаксическая типология (ср. различение эргативных vs. аккузативных языков, *быть*-языков vs. *иметь*-языков и т. п.) также оказывается неполной без измерения встречаемости разных конструкций (и выделения шкалы постепенных переходов между синтаксическими типами языков).

(2) В преподавании языка (как родного, так и неродного), необходима иерархизация ЛСР по квантитативному параметру ядерности/периферийности: в более кратких описаниях языка можно ограничиваться лишь более ядерными типами ЛСР, а в более полных — включать также более периферийные типы ЛСР.

(3) При разработке моделей АОТ упорядочение ЛСР по степени их словарной и текстовой мощности предполагает построение синтаксических анализаторов, выстроенных в цепочку последовательных аппроксимаций (где начальная модель учитывает лишь основные типы конструкций, а каждая последующая модель охватывает постоянно растущее число более маргинальных конструкций), что повышает степень адекватности модели.

Прилагательные со значением высокой и низкой температуры и наивно-языковая оценка температуры[1]

The adjectives with meaning of high and low temperature and linguistic estimation of temperature

Крылова Т. В.

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

В этой статье рассматриваются прилагательные со значением температуры: холодный, прохладный, горячий, жаркий, теплый. В первом разделе анализируется разбиение этих прилагательных на группы, во втором рассматривается их сочетаемость с наречиями со значением степени. При этом высказывается предположение, что многие различия в употреблении данных прилагательных обусловлены различием в языковой оценке высокой и низкой температуры. В заключении та же мысль иллюстрируется на материале глаголов со значением температуры.

1. Введение

В ходе работы над Активным толковым словарем нами были рассмотрены прилагательные со значением температуры: *холодный, прохладный, теплый, горячий, жаркий.* Эти слова представляют большой интерес в двух отношениях. Во-первых, они дают ценный материал для реконструкции наивно-языковых представлений о восприятии температуры и, в частности, о языковой оценке низкой и высокой температуры. Во-вторых, рассматриваемые слова обладают колоссальным метафорическим потенциалом, причем соотношение их прямых и переносных значений позволяет сделать ряд выводов о механизмах осмысления эмоций в языке.

Рассматриваемые слова неоднократно становились объектом лингвистического изучения. В частности, они были рассмотрены в работе Е. В. Рахилиной [Рахилина 2000], где данные прилагательные описывались в соответствие с идеями когнитивной лингвистики. Специфика нашего подхода к анализу семантики этих слов, обусловленная установкой на их лексикографическое описание, состоит в подробном изучении структуры их полисемии. Так, в отличие от работы [Рахилина 2000], где у рассматриваемых прилагательных усматривается всего два значения — прямое и переносное, в данной работе делается попытка разграничить в рамках этих блоков отдельные лексемы и истолковать их, по возможности избежав кругов в толковании.

2. Разбиение прилагательных со значением температуры на группы

Слова семантического поля температуры обнаруживает много общего в структуре полисемии. Рассмотрим хотя бы блок прямых значений. Если свести воедино все основные значения, которые присутствуют у этих слов, получится следующее.

Характеристика температуры объекта или связанных с ней ощущений: 1) температура объекта, воспринимаемая посредством контакта с ним (*холодная <горячая, теплая, прохладная> вода*); 2) субъективные температурные ощущения, которые обычно бывают вызваны воздействием на человека той или иной температуры воздуха (*Ему стало холодно <тепло, жарко>*); 3) объективная температура воздуха каком-л. месте или в какой-л. промежуток времени (*В комнате холодно <тепло, жарко>; холодный <теплый, жаркий> день*). **Физические свойства объекта, связанные с температурой**: 4) количество тепла, выделяемого источ-

[1] Данная работа осуществлена при поддержке гранта РГНФ №06-04-00289а «Разработка словника и проспекта Активного словаря русского языка» на 2006–2008 гг., гранта Программы фундаментальных исследований ОИФН РАН «Русская культура в мировой истории», гранта Президента РФ для поддержки научных исследований, проводимых ведущими научными школами РФ № НШ-5611.2006.6.

ником тепла (*холодные <теплые, горячие, жаркие> лучи солнца*); 5) наличие в помещении отопления (специальное *холодный <теплый> гараж,* уходящее *холодный карцер*); 6) способность теплоизолирующих объектов — в частности, одежды, реже — различных материалов — сохранять тепло (*теплая куртка,* разговорное *холодная куртка,* специальное или разговорное *Этот линолеум очень теплый, теплая пряжа*); 7) способ осуществления действий, в плане наличия / отсутствия нагревания (или его интенсивности) (*холодное <горячее> копчение,* специальное *горячий <холодный> отжим*).

Весь комплекс перечисленных значений присутствуют только у прилагательного *холодный,* которое имеет наиболее развитую многозначность.

Рассмотрим значения, характеризующие собственно температуру. Они (или хотя бы два из них) присутствуют у всех рассматриваемых слов. Хотя во всех этих значениях *прохладный* и *теплый* по характеру температуры сближаются соответственно с *холодным* и *горячим,* они одновременно объединяются между собой, противопоставляясь *холодному, горячему* и *жаркому.*

Во-первых, оба прилагательных первой группы обозначают умеренную степень признака: *прохладный* обозначает умеренно высокую температуру, *теплый* — умеренно высокую.

Во-вторых, они всегда или очень часто (как *прохладный*)[2] предполагают положительную оценку данного признака, указывая на то, что соприкосновение с объектом (или нахождение в соответствующих условиях) обычно приятно. (См. об этом также в [Рахилина 2000: 218-219]).

Приведем толкования.

Прохладный 1.1 — ‘Такой, температура которого ниже температуры тела, однако не составляет с ней резкого контраста, так что соприкосновение с объектом обычно вызывает приятное ощущение, особенно если человеку жарко’; ср. *прохладный душ, прохладные руки.*

Прохладный 1.2 — ‘Такой, который характеризуется в меру низкой температурой воздуха, так что человек, который находится в этих условиях, испытывает приятное ощущение, особенно если ему жарко’ [о месте, промежутке времени или в сочетании со словами *погода, климат*]; ср. *прохладная гостиная; В комнате прохладно.*

Теплый 1.1 — ‘Такой, температура которого близка к температуре тела, так что соприкосновение с объектом вызывает приятное ощущение, или выше нормы, существующей для такого рода объектов’ [в последнем случае имеются в виду объекты, которые обычно бывают холодными]; ср. *теплая ванна, теплое течение Гольфстрим*.

Теплый 1.2 — ‘Человек А1 испытывает ощущение температурного комфорта’ [в роли А1 может выступать часть тела]; ср. *В пуховой куртке мне было тепло; Ногам тепло.*

Теплый 1.3 — ‘Такой, который характеризуется температурой воздуха, достаточно высокой, чтобы человек, который находится в этих условиях, испытывал ощущение температурного комфорта, или которая немного выше нормы’ [о месте, промежутке времени или в сочетании со словами *погода, климат*]; ср. *теплый день, Сегодня тепло.*

Из-за сходства оценок *теплый* и *прохладный* чаще всего противопоставляются не друг другу, а прилагательным, обозначающим высокую степень противоположного признака, тем более что для последних очень типична отрицательная оценка.

Вследствие антонимичности оценок *теплый* в физическом блоке обычно противопоставляется *холодному*, *прохладный — горячему* или *жаркому*. Ср. странное *На улице тепло, а в доме прохладно* при стандартном *На улице жарко, а в доме прохладно.*

Иная ситуация — в блоке переносных значений; здесь *прохладный* объединяется с *холодным;* они вместе указывают на отрицательное чувство или отношение и противопоставляются *теплому* и *горячему*, которые указывают на положительное чувство или отношение: ср. *теплый прием, горячий прием* VS. *холодный прием, прохладный прием.* (Тот факт, что положительная оценка в случае *прохладный* не наследуется в переносных значениях, отмечается также в работе [Рахилина 2000: 219]).

Причина возникновения такого рода значений, на наш взгляд, — метафорическое сближение положительных эмоций и их органов (души и сердца) с огнем; ср. метафорические употребления типа *горячее <пламенное> сердце, сердечный <душевный> жар; сердечный огонь; душевное тепло, дарить тепло*. Вследствие этого все слова со значением низкой температуры, вместе взятые, указывают на отрицательные эмоции, слова со значением высокой температуры — на положительные.

В частности, *горячий,* несмотря на типичность для него в основном блоке компонента «неприятное ощущение», указывает на положительные эмоции (чаще всего — любовь), причем подчеркивает интенсивность последних. Компонент ‘положительные чувства’ входит в значение лексемы *горячий 2.6* (*горячий прием*), которая толкуется как ‘такой, с помощью которого человек А1 показывает, что он испытывает сильную любовь, восхищение или другие сильные положительные чувства к человеку А2’. Кроме того, употребление лексемы *горячий 2.4* (*горячая любовь, горячее сочувствие*)[3] также указывает на тяготение этого прилагательного к сфере положительных эмоций.

---

[2] См. об этом подробнее ниже.

[3] Хотя в последнем случае горячий указывает только на интенсивность чувств или отношений (Горячий 2.4 — ‘Очень интенсивный’), тем не менее эта лексема обычно употребляется применительно к положительным чувствам и отношениям — ср. более типичное горячая любовь, горячая симпатия при менее стандартном горячая ненависть, горячая антипатия.

Аналогичным образом, *прохладный,* который в основном блоке часто содержит компонент «приятное ощущение», в метафорическом блоке тем не менее указывает на негативные эмоции — ср. *прохладный прием, прохладные отношения, прохладный отзыв.* (В этом случае реализуется значение *прохладный 2*: ‘такой, с помощью которого человек А1 показывает, что он не испытывает радости от общения с человеком А2 или не испытывает восхищения от человека или другого объекта А2’).

Таким образом, признак значения температуры становится для рассматриваемых прилагательных в метафорическом блоке основным группирующим признаком. Признак степени, на основании которого в основном блоке *теплый* и *прохладный* противопоставляются *горячему* и *холодному,* для метафорического блока менее значим и не играет группирующей роли. Это проявляется в отсутствии сближения между *теплым* и *прохладным* в переносном значении, о чем свидетельствует, в частности, возможность их использования в контексте противопоставления: *Он заранее приготовился к прохладному приему, но встреча оказалась неожиданно теплой.*

Однако вернемся к блоку прямых значений. При том, что *прохладный, теплый* в этом блоке сближаются между собой, между ними фиксируются различия в сочетаемости, которые, на наш взгляд, свидетельствуют о разной оценке в языке высокой и низкой температуры.

Рассмотрим сочетаемость *прохладного* и *теплого* с наречиями малой степени.

Теплый 1.1 (теплая вода) и *теплый 1.3 (теплый день; На улице тепло)* в положительной степени не сочетаются с такими наречиями малой степени, как *немного* и *слегка*[4]*,* что для соответствующих значений прилагательного *прохладный* вполне нормально. Ср. нормальное *Ветер немного прохладный (прохладный 1.1); Сегодня немного прохладно (прохладный 1.2)* при неудачном *Ветер немного теплый; Сегодня немного тепло.*

На наш взгляд, это может объясняться несовпадением оценок, выражаемых вышеназванными наречиями, с одной стороны, и прилагательным *теплый,* с другой.

Немного и *слегка* (опять-таки в сочетании с прилагательными положительной степени) тяготеют к отрицательно оцениваемым свойствам: стандартно *немного <слегка> грустный <глупый, усталый>* при неудачном *немного <слегка> веселый <умный, бодрый>*.

Между тем, обнаруживается, что *прохладный* и *теплый* в отношении оценок не вполне идентичны. Если *теплый 1.1* и *1.3* практически во всех случаях выражают положительную прагматическую оценку (за исключением ситуации, когда *теплый 1.1* используется применительно к напиткам — *теплое пиво, теплая вонючая вода*), то *прохладный 1.1* и *1.2* выражают отрицательную оценку достаточно часто, хотя и реже, чем положительную[5]. В случае, когда прилагательное *прохладный* получает отрицательную оценку, оно сближается с *холодным*: *Надень куртку, на улице прохладно* ≈ ‘немного холодно’; ср. также сочетания *прохладный ветерок* и *прохладный ветер*, которые различаются оценками.

Бо’льшая способность *прохладного* к выражению отрицательной оценки как раз и является причиной того, что лексемы *прохладный 1.1* и *1.2,* в отличие от *теплого 1.1* и *1.3*[6]*,* могут сочетаться с наречиями *немного* и *слегка* — не случайно в контекстах этих наречий *прохладный* всегда используется в негативном значении. (Ср. странное *В саду хорошо и немного прохладно* и стандартное *Застегнись, здесь немного прохладно*).

В свою очередь, *теплый 1.1* может сочетаться с наречием малой степени *еле,* что для *прохладного 1.1* нетипично.

Это связано с тем, что наречие *еле* указывает на отрицательную оценку говорящим малой степени признака и на нарушение ожиданий[7]. Не случайно *еле теплыми* могут быть названы только такие объекты, которые в норме должны быть горячими — это, прежде всего, еда и напитки (*Чай <суп> еле теплый, надо подогреть*) и, кроме того, отопительные приспособления, нагревательные приборы и пр. (*Батареи еле теплые; Чайник <утюг> еле теплый*).

Во всех этих случаях *еле теплый* описывает нарушение прагматической нормы, что сближает случаи типа *Суп еле теплый* с уже описанными выше случаями типа *Суп совсем холодный.* Такие фразы практически синонимичны, тем более что они часто используются в разговорной речи с оттенком преувеличения, когда на самом деле температура объекта достаточно высока. Разница между ними — в том, как оформляется смысл ‘недостаточно горячий’. В случае *Суп совсем холодный* он оформляется через приписывание объекту противоположного признака (холодного), в случае *Суп еле теплый* — с помощью приписывания ему признака теплого в минимальной степени (при том, что сам признак ‘теплого’ представляет собой признак ‘горячего’ в малой степени).

Для прохладного 1.1 использование в контексте нарушения прагматической нормы невозможно; ср.

---

[4] В форме сравнительной степени теплый допускает такие сочетания; ср. Стало немного теплее.

[5] Из-за этого прохладный 1.2 может использоваться в контексте что-то, тогда как для теплый 1.3 это невозможно: ср. нормальное Что-то сегодня прохладно при неудачном Что-то сегодня тепло.

[6] В тех редких случаях, когда теплый выражает отрицательную оценку (теплое пиво), в его семантике содержится элемент ‘выше нормы’, что затрудняет сочетаемость с немного.

[7] См. об этом в [НОСС 2004: 648].

неправильное **Квас еле прохладный, надо поставить в холодильник*. Можно предложить следующее объяснение этого факта: хотя в жизни норма может быть связана как с высокой, так и с низкой температурой, в языке высокая температура чаще оценивается положительно, чем отрицательная. По этой причине выражение смысла 'недостаточно холодный', в отличие от 'недостаточно горячего' является нетипичным; ср. прагматически не вполне стандартное *Пиво совсем теплое* при абсолютно стандартном *Каша совсем холодная*, а также странное *Нос собаки совсем горячий* при стандартном *Руки у тебя совсем холодные.*

3. Заключение

Итак, была высказана гипотеза о том, что для наивно-языкового сознания высокая и низкая температура не вполне равноценны: представление о прагматической норме связывается прежде всего с высокой температурой. В подтверждение этой гипотезы можно привести самые различные факты.

Один из них состоит в том, что на фоне многочисленных глаголов, описывающих неприятное ощущение низкой температуры (*замерзнуть, озябнуть, продрогнуть, застыть, закоченеть*), практически отсутствуют глаголы со значением неприятного ощущения высокой температуры, если не считать разговорных *запариться* и *зажариться*, сфера употребления которых весьма ограничена. (Представление о том, что тепло для человека полезней холода, отражается также в поговорке *Пар костей не ломит*).

Сказанное подтверждается также несовпадением оценок антонимичных слов со значением высокой и низкой температуры.

Сравним оценки в случае *жаркий 1.2 (жаркий день)* и *холодный 1.3 (холодный день)*. Несмотря на то, что обе лексемы содержат в толковании компонент 'неприятное ощущение', оценки в этих случаях не совсем идентичны. В случае *холодный 1.3* отрицательная оценка преобладает, а положительная практически невозможна. Это отчетливо ощущается, в частности, при сопоставлении сочетаний *холодный день* и *морозный день:* если второе может использоваться в контексте положительной оценки (*Стоял чудесный морозный день; морозный денек*), то для первого такое использование нетипично; ср. нестандартное *Стоял чудесный холодный день; холодный денек.*

Что касается *жаркий 1.2,* то его оценка менее однозначна. Краткая форма *жарко,* которая используется в безличном употреблении, всегда выражает отрицательную оценку; ср. *В городе летом жарко; Пес залаял, как на луну, потом чихнул — на чердаке было пыльно и жарко от раскаленной железной крыши* (Е.и В. Гордеевы).

Между тем, для полной формы *жаркий,* наряду с преобладающими примерами типа *невыносимо жаркий день; И снова потянулся длинный, мучительный, жаркий, бессмысленный день* (Ю. Домбровский), возможно и использование в контексте положительной оценки. Ср. примеры типа *Воскресенье. Чудный жаркий день. Отправляюсь в Ручьи и снова симфония воздуха, отдыха и Великой Пищи* (А. Болдырев); *Дальше шла запись по дням. […] «Чудный жаркий день. Вечером ездил на велосипеде»* (В. Набоков); *Дивная погода! Жаркий день. Теплый, душистый, упоительный май* (И. Е. Репин).

Аналогичное несовпадение оценок мы встречаем в парах *теплый 1.1* VS. *прохладный 1.1 (теплая <прохладная> вода)* и *теплый 1.3* VS. *прохладный 1.2 (теплый <прохладный> день)*; см. об этом выше.

Похожее соотношение — между лексемой *согреться 2,* производной от *теплый 1.1* (*Вода согрелась),* и ее антонимом *остыть.* Первая из них всегда выражает положительную оценку, для второго более типична отрицательная оценка.

Рассмотрим глагол *остыть.* Он толкуется следующим образом, *А1 остыл* — 'Вследствие исчезновения причины высокой температуры объекта А1 и контакта А1 с окружающей средой, имеющей более низкую температуру, его температура понизилась, и он перестал быть горячим или теплым'. Ср. *Песок медленно остывал; Вода остывает медленнее, чем воздух; Суп <чай, утюг> совсем остыл, надо снова греть; Печка совсем остыла.*

Для остыть очень характерна отрицательная оценка, особенно в сочетании с названиями объектов, которые в норме должны быть горячими или теплыми (*Суп <утюг> остыл; Вода в ванне остыла)*. Особенно типична такая оценка для форм настоящего времени несовершенного вида в конструкции *А1 остывает*[8]; ср. *Обед остывает; Скорее садись за стол, суп остывает; Давай лучше чай допьем, остывает. У тебя есть варенье?* (М. Милованов). В предложениях такого типа *остывать* обычно указывает на то, что объект становится более холодным, чем нужно, сближаясь с лексемой *стыть 1 (Обед остывает ≈ Обед стынет)*. Предложения такого типа редко используются в сочетании с названиями объектов, которые в норме должны быть холодными, т. е. когда речь идет о «правильном» понижении температуры, соответствующем желанию человека. Ср. не вполне стандартное *Компот остывает; Холодец остывает.*

[8] Такая интерпретация становится необязательной, если остывать имеет зависимые слова (Панель плиты Hansa Prestige очень быстро остывает — через минуту после выключения к ней можно прикасаться («Домовой», 2002.06.04); Компот остывает на окне) или если в предложении есть указания на то, что человек заинтересован в развитии процесса и контролирует его (Пусть лепешка остывает; Пока заварное тесто остывает, приготовьте крем).

Теперь рассмотрим лексему *согреться 2*. Она допускает два режима употреблений.

В первом из них *согреться* (обычно в форме совершенного вида) обозначает процесс, каузированный человеком и развивающийся под его контролем; ср. *Кофе <Чай> согрелся; Чайник согрелся; Вода для стирки согрелась; Постель согрелась, можешь убирать грелку.*

Во втором режиме употреблений *согреться* не указывает на то, что процесс повышения температуры каузирован человеком, однако предполагает, что данный процесс является желательным или приятным для последнего[9]; ср. *Вода в пруду согрелась, можно купаться; В самом деле, воздух тотчас начал согреваться* (Н. А. Дурова); *Как только сходит снег, согревается земля, […] в тайге просыпается самый страшный зверь — гнус* («Вокруг света», 2004); *Камни были холодными на ощупь, но быстро согревались в руке, некоторые из них переливались, казалось, что они меняют свой цвет* (М. Львова); *Пройдя через полость носа, воздух согревается и очищается* («Семейный доктор», 2002); *Как лед кругом шеи-то [бусы], и не согреваются нисколько* (П. П. Бажов).

В сочетании с названиями объектов, которые в норме должны быть холодными, вследствие чего повышение их температуры является нежелательным, *согреться* не используется, что отличает его от *нагреться*. Ср. нормальное *Мокрое полотенце на лбу у больного нагрелось, и сиделка поменяла его; Пиво нагрелось на солнце, поставь его в холодильник* при странном *Мокрое полотенце на лбу у больного согрелось, и сиделка поменяла его; Пиво согрелось, поставь его в холодильник*. Это связано с тем, что *согреться* в данном режиме указывает на достижение прагматического стандарта теплого и всегда оценивается положительно.

Можно выстроить следующее толкование *согреться 2,* которое будет охватывать оба круга употреблений.

Согреться 2. A1 согрелся ‘В результате близости к источнику тепла или контакта с ним температура неодушевленного объекта A1 повысилась, и он стал горячим или теплым, что соответствует целям какого-л. человека или людей или является для них желательным’.

Описанное различие в оценках между *согреться* и *остыть,* на наш взгляд, лишний раз свидетельствует о большей значимости для наивно-языкового сознания высокой температуры, по сравнению с низкой (а также указывает на особую выделенность в этом отношении того участка температурной шкалы, который соответствует представлению о теплом).

[9] Аналогичным образом, лексема остудить 1.2, обозначающая воздействие (Ветер остудил ее горящие щеки), обычно выражает положительную прагматическую оценку.

Литература

1. *НОСС 2004* — Новый объяснительный словарь синонимов русского языка, Москва-Вена, 2004.

2. *Рахилина 2000* — Рахилина Е. В. Когнитивнй анализ предметных имен: семантика и сочетаемость. М., 2000, 211–233.

Инструменты полуавтоматической разметки для Мультимедийного русского корпуса (МУРКО)[1]

Semiautomatic marking tools for the Russian multimedia corpus (MURCO)

Кудинов М. С. (peshka1@mail.ru)
Филологический факультет МГУ

Гришина Е. А. (rudi2007@yandex.ru)
Институт русского языка им. В. В. Виноградова РАН, Москва

В статье описываются рабочие места разметчика речевых действий и жестов для Мультимедийного русского корпуса (МУРКО), которые позволяют быстро и единообразно аннотировать такой исключительно разнообразный материал, как русская жестикуляция и различные речевые акты, используемые в устной русской речи.

1. Введение

Приступая к созданию того или иного корпуса, его конструктор, среди прочих, должен ответить для себя на следующие два важных вопроса: 1) какие типы аннотаций (разметки) предполагается использовать в данном корпусе, 2) какова предполагаемая методика аннотирования. Для Мультимедийного русского корпуса (МУРКО) предварительные ответы на первый вопрос уже даны: помимо стандартных (для устного подкорпуса Национального корпуса русского языка) типов аннотации (морфологической, семантической, акцентологической, социологической) предполагаются дополнительные типы — орфоэпическая, жестовая, разметка речевых действий (см. [Гришина 2009а], [Гришина 2009б]). В настоящей статье мы предложим ответ на второй вопрос — какая именно методика аннотирования предполагается для МУРКО.

Говоря в целом, здесь у нас есть возможность выбора из трех вариантов: 1) автоматическая, 2) полуавтоматическая, 3) ручная разметка. Так, например, морфологическая и семантическая разметка в НКРЯ осуществляются полностью в автоматическом режиме[2]. Социологическая разметка предполагает предварительную ручную обработку материалов с последующим автоматическим приписываем тем или иным фрагментам текстов социологических параметров (имя, пол, возраст, профессия говорящего). Акцентологическая разметка, напротив, предполагает первоначальную обработку текста полностью автоматическим акцентуатором с последующей ручной коррекцией ударений в соответствии с реальным произнесением. Таким образом, и социологическая, и акцентологическая разметки осуществляются в полуавтоматическом режиме. Что касается орфоэпической разметки в МУРКО, то она будет осуществляться автоматически с опорой не на звуковой/фонематический состав слова, а на буквенный состав в сочетании с ориентацией на место постановки ударения[3].

Что касается разметки речевых действий и жестовой разметки, то здесь говорить об автоматическом аннотировании не приходится. Для разметки речевых действий в автоматическом режиме потребовались бы сведения об интонационной, фонетической, пунктуационной, синтаксической, семантической, морфологической характеристике того или иного речевого действия, а для получения таких сведений как раз и создаются корпуса. Что касается жестовой разметки, то и здесь проблема аналогична — для того, чтобы научить машину автоматически отделять один жест от другого, ее нужно «натренировать» на достаточно больших корпусах, не говоря уже о потребности в солидной теоретической базе. Кроме того, в МУРКО кор-

[1] Статья написана при поддержке программы ОИФН РАН 2009–2011 гг. «Генезис и взаимодействие социальных, культурных и языковых общностей».

[2] См. [Ляшевская и др. 2005]. В подкорпусе со снятой грамматической омонимией результаты автоматической морфологической разметки корректируются в ручном режиме, соответственно, морфологическая разметка в т. н. «снятом» корпусе должна оцениваться как полуавтоматическая (ср. [Сичинава 2005]).

[3] Необходимость и общая структура орфоэпической аннотации подробнее излагаются в [Гришина 2009б].

пусной материал будет представлен фрагментами фильмов, а не специально подготовленными (с использованием 4–6 камер) видеозаписями, так что подобная задача неразрешима технически (в связи с необходимостью программного отслеживания положения частей человеческого тела в течение всей сцены).

Таким образом, при аннотировании речевых действий и жестов в МУРКО мы обречены на выбор между полностью ручной аннотацией и аннотацией в полуавтоматическом режиме. Были принято решение достичь полуавтоматического режима работы. При этом за человеком в паре «человек–машина» остается самое важное — принятие содержательного решения по тому или иному жесту, тому или иному речевому действию; программа же «принимает обязательства» а) предлагать одинаковые аннотационные формулировки для одинаковых содержательных решений, б) размещать принятые человеком содержательные решения (в единообразных формулировках) в единой базе в соответствии с единообразной матрицей.

Именно по таким принципам были созданы два рабочих места разметчика (PMP) — PMP речевых действий Marker и PMP жестов GesturesMarker. Таким образом, работа PMP представляет собой последовательное заполнение описательной таблицы (на форме — компонент DataGridView .NET Framework) для каждого клипа. Работа разметчика заключается в правильном выборе из ряда параметров,заранее заданных ему в диалоговом окне.

2. Пользовательский интерфейс PMP Marker

Главное окно PMP Marker выглядит следующим образом (см. рис. 1). Как видим, левый (серый) столбец содержит строки, каждая из которых соответствует тому параметру, по которому описывается как целый клип (имя файла, пол говорящих, язык, на котором говорят, социальная ситуация, отраженная в клипе), так и каждое речевое действие в данном клипе (типы речевых действий, полнота речевого действия, манера говорения, типы повторов, типы вокальных жестов, зафиксированные в клипе, — последние 4 параметра не видны на рисунке и проявляются при скроллинге главного окна).

Работу логично начать с нажатия кнопки «Следующий шаг». Нажатие генерирует вывод диалоговой формы, соответствующей той или иной стадии заполнения.Так, например, на втором шаге (см. рис. 2) после нажатия перед пользователем открывается окно, в котором он может выбрать 1) количество говорящих, 2) пол говорящих, 3) язык, на котором говорят в клипе, 4) тип социальной ситуации, отраженной в клипе (если не выбрана никакая социальная ситуация, то она считается неспецифичной).

Как видим, разметка на рис. 2 утверждает, что в клипе отражен фрагмент лекции, которую на русском языке читает лектор-мужчина.

После нажатия кнопки «Завершить» перед разметчиком открывается следующее окно (рис. 3), в котором он может выбрать тип (*вопрос, согласие, отрицание, этикетные высказывания* и т. д.) и конкретную разновидность речевого действия, которое совершает в данном клипе говорящий[4].

Аналогичным образом построены окна и на всех последующих шагах (разметка полноты речевого действия, указание на наличие и тип повторов, указание на манеру говорения, на наличие и тип вокальных жестов в высказывании). Когда разметка данного клипа проходит весь предназначенный путь, разметчику предоставляется возможность проверить результат и внести изменения, после чего сохраненные данные записываются в таблицу Microsoft Excel. Результирующая Excel-таблица содержит следующие столбцы:

1. Имя файла
2. Количество говорящих
3. Пол говорящих
4. Язык, на котором говорят
5. Социальная ситуация, отраженная в клипе
6. Типы речевых действий, отраженные в клипе
7. Конкретное речевое действие, отраженное в клипе
8. Полнота речевого действия
9. Наличие и типы повторов
10. Манера говорения
11. Типы вокальных жестов и междометий

Каждый из этих столбцов (кроме первого) задает параметр, по которому можно производить отбор клипов из базы данных корпуса. Так, можно найти клипы, содержащие отрывки из диалога двух женщин на русском языке (с акцентом или без оного), который включает общие и частные вопросы, содержит однословные неоднократные повторы и вокальные жесты удивления. В сочетании со стандартной для НКРЯ лексической, морфологической, семантической, акцентологической и социологической информацией, а также с орфоэпической разметкой, характерной для МУРКО, это предоставит пользователю корпуса значительные исследовательские возможности.

[4] В Приложении 1 к данной статье читатель может найти список речевых действий, сгруппированных по типам, — так, как последние сформировались на начало 2009 г. PMP Marker позволяет разметчику как пополнять/трансформировать список, так и работать с ним вне зависимости от распределения речевых действий по типам, которое, очевидно, является весьма условным.

Рис. 1. Главное окно PMP Marker

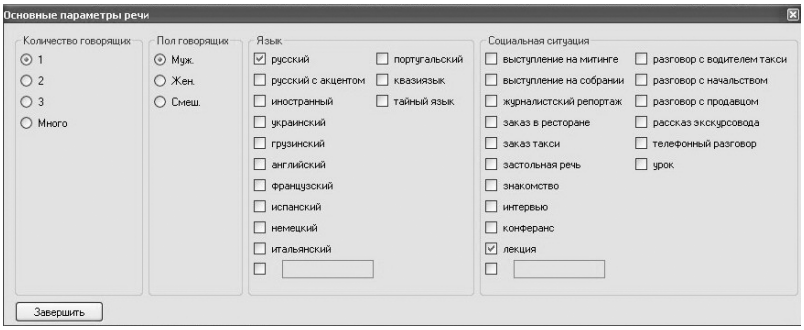


Рис. 2. Второй шаг

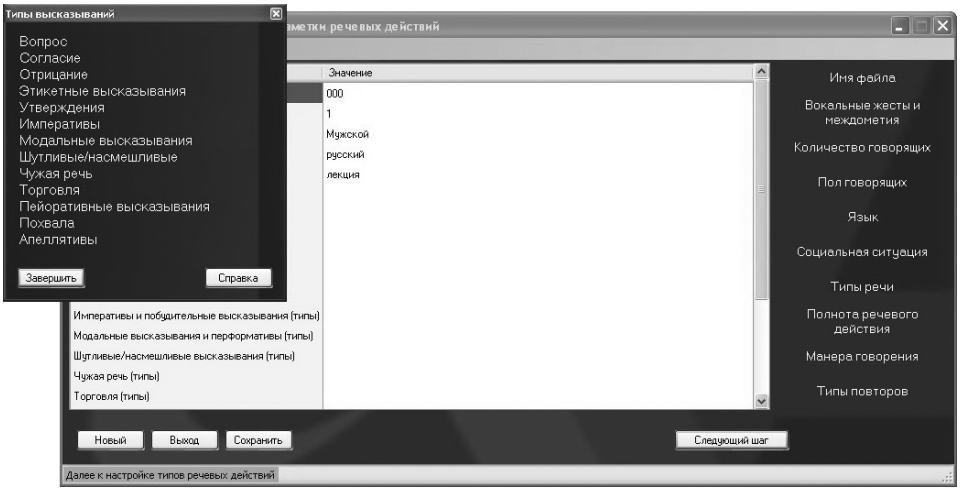


Рис. 3. Третий шаг

3. Пользовательский интерфейс PMP GesturesMarker

На аналогичных принципах строится работа PMP GesturesMarker. Главное окно программы выглядит следующим образом (см. рис. 4).

После заполнения поля ввода для имени клипа активируется кнопка «Далее», которая приводит пользователя к следующему этапу описания жеста (см. рис. 5). Одновременно начинается заполнение основной таблицы — введенная информация сохраняется в полях «Имя файла», «Имя говорящего», «Пол говорящего», «Пол персонажа», «Возраст говорящего», «Возраст персонажа», «Социальная ситуация»[5] (см. рис. 5).

[5] Поле «Социальная ситуация» в стандартном случае заполняется в PMP Marker, т. е. при разметке речевых действий. В PMP GesturesMarker поле «Социальная ситуация» заполняется только в том случае, если описываемый клип не содержит речевой составляющей, а включает только жестовый материал.

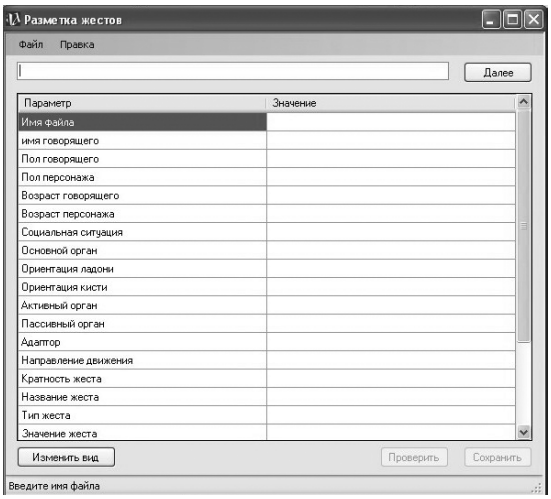


Рис. 4. Главное окно PMP GesturesMarker

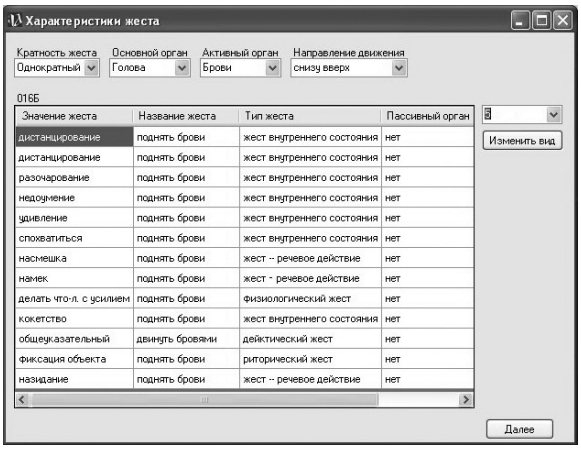


Рис. 6. Третий шаг разметки жестов

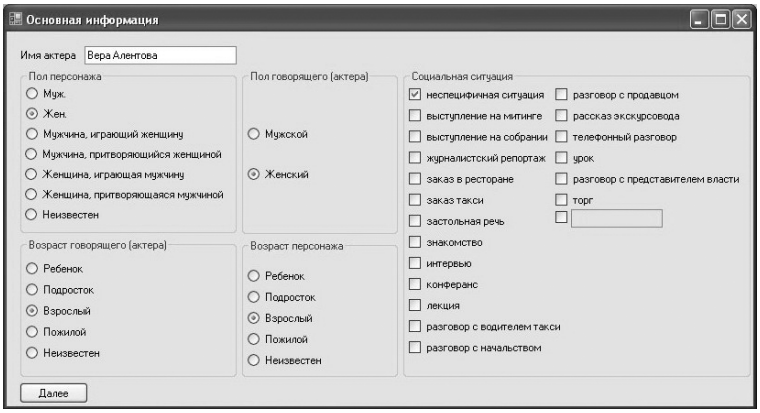


Рис. 5. Второй шаг разметки жестов

Как видим, на втором этапе описания жеста пользователю предлагается ввести имя актера, его пол, пол персонажа, возраст актера и персонажа, а также типовую социальную ситуацию, в которой происходит действие (в данном случае, актриса Вера Алентова, женского полу, играет взрослую женщину в неспецифичной ситуации). После нажатия кнопки «Далее» соответствующие строки в столбце на рис. 4 заполняются, а пользователь переходит к следующему шагу, где производится описание собственно жеста (см. рис. 6).

Последовательным выбором значений в окошках «Кратность жеста» → «Основной орган» → «Активный орган» → «Направление движения» разметчик приходит к списку однократных жестов, для которых основным органом[6] является голова, активным органом — брови, движущиеся снизу вверх. Выбирая строчку 5, пользователь тем самым выбирает жест *поднять брови* от *удивления*, который является *жестом внутреннего состояния*. Таким образом, на данном этапе, нажав кнопку «Далее», пользователь заполнит строчки «Кратность жеста», «Основной орган», «Активный орган», «Направление движения», «Значение движения», «Название жеста», «Тип жеста», «Пассивный орган» и некоторые другие, которые на рис. 4 не видны. После чего пользователь перейдет к следующему шагу.

Проделав последовательно все действия, предусмотренные PMP, пользователь перейдет к следующему жесту в данном клипе или к следующему клипу, а заполненная им таблица, как и в случае PMP Marker, трансформируется в очередную строку результирующей таблицы Excel.

Результирующая таблица содержит следующие столбцы:

1. Имя файла
2. Имя говорящего
3. Пол говорящего
4. Пол персонажа
5. Возраст говорящего
6. Возраст персонажа
7. Социальная ситуация
8. Основной орган
9. Ориентация ладони
10. Ориентация кисти
11. Активный орган
12. Пассивный орган

[6] Основной орган — зона человеческого тела, в которой осуществляется данный жест (см. [Крейдлин 2002]).

13. Адаптор
14. Направление движения
15. Кратность жеста
16. Название жеста
17. Тип жеста
18. Значение жеста
19. Наличие удлинителя
20. Наличие спойлера
21. Эмоции
22. Полнота жеста
23. Аутентичность жеста
24. Аксессуары

Эти столбцы, как и в случае PMP Marker, задают параметры поиска в будущем корпусе. Некоторые из этих позиций требуют краткого пояснения. ***Адаптор*** – это необходимый участник жеста (человек или предмет), например, объект указания для дейктического жеста, собеседник для этикетного жеста, галстук для декоративного жеста *поправить галстук* и т. д. ***Удлинитель*** — предмет, который выступает заместителем того или иного органа человеческого тела при осуществлении жеста, например, ручка или карандаш при указании вместо указательного пальца (см. [Крейдлин 2002]). ***Спойлер*** — предмет, который мешает полноценному исполнению данного жеста, например, сумка в руке, которая мешает в полном объеме исполнить жест *развести руками*. ***Аксессуары*** — это предметы, так или иначе вовлеченные в осуществление жеста. Аксессуары могут быть адапторами, удлинителями и спойлерами. Под ***эмоциями*** в данном случае понимаются явные проявления эмоций (улыбка, смех, плач), которые сопровождают данную жестикуляцию. При описании ***полноты*** жеста учитываются ситуации, которые ведут к прерыванию жестикуляции (добровольному или принудительному). ***Аутентичностью*** жеста мы называем параметр, который учитывает, с одной стороны, искренность жеста (различаются аутентичные и притворные жесты), а с другой стороны, — степень принадлежности жеста жестикулирующему (например, говорящий может показывать на себе чужие жесты, отражать зеркально жестовое поведение собеседника).

Как видим из приведенного списка параметров, часть их имеет технический, объективный характер (пп. 2–15, 19–24), в чрезвычайно высокой степени не зависящий от разметчика, а следовательно, и от пользователя будущего корпуса. Первоначально у создателей МУРКО был соблазн ограничиться только этими объективными показателями, чтобы элиминировать субъективный момент в разметке. Однако в дальнейшем было принято решение ввести в разметку содержательные пункты 16–18, которые бы давали истолкование (хотя бы приблизительное) жеста и тем самым расширяли бы возможности поиска информации в МУРКО.

Истолкование жеста строится по трем зонам. Во-первых, жесту приписывается некоторое ***значение*** (п. 18), которое становится ясным разметчику из контекста данного конкретного клипа. Набор значений очень велик, но, по-видимому, все-таки не бесконечен. Значения жестов группируются в ***типы***. На данный момент предлагается выделить следующие типы жестов:

1. Дейктические — жесты, которые выражают указание
2. Декоративные — жесты, которыми говорящий улучшает свой внешний вид
3. Речевые действия — жесты, которые обозначают тип речевого действия, осуществляемого говорящим
4. Жесты внутреннего состояния — жесты, демонстрирующие внутреннее состояние говорящего
5. Изобразительные — жесты, изображающие некоторый предмет или событие
6. Корпоративные — условные жесты, характерные для некоторой относительно узкой социальной группы
7. Пейоративные — жесты, выражающие критику в чей-либо адрес
8. Поисковые — жест, сопровождающие поиск или получение какой-либо информации
9. Регулирующие — жесты, регулирующие поведение собеседника (или претендующие на осуществление такой регуляции)
10. Риторические — жесты, не имеющие самостоятельного значения и просто сопровождающие речь в целях повышения ее воздействия
11. Условные — жесты с утерянной или сомнительной этимологией, свойственные носителям данного языка и культуры
12. Физиологические — жесты, употребление которых определяется физиологией человека
13. Этикетные — жесты, связанные с речевым и поведенческим этикетом, характерным для носителя данного языка и культуры
14. Заимствованные — жесты, использование которых позволяет говорящему имитировать свою принадлежность к чужой культуре

Очевидно, что этот набор типов весьма условен и ни в коей мере не претендует на то, чтобы быть логически выверенной классификацией, — это всего лишь набор содержательных меток, который позволяет быстрее ориентироваться во множестве зафиксированных жестов.

Кроме того, каждый жест (кроме большей части изобразительных жестов) имеет некоторое ***название***. В большинстве случаев это название, уже устоявшееся в языке[7], однако для некоторых жестов (например, для значительного числа риторических

[7] Очевидно, что жесты, имеющие одно и то же название, например, *кивнуть, махнуть рукой,* могут иметь самые разные значения и, соответственно, относиться к самым разным типам.

жестов, которые ввиду своего «служебного», «подчеркивающего» употребления очень часто не отрефлектированы в языке) приходилось изобретать свои названия, часто весьма условные и стилистически ущербные. Приложение 2 содержит предварительный список жестов, упорядоченный по типам — значениям — названиям, так, как он сложился на начало 2009 г.

4. Редактирование PMP

Значительным удобством при работе с обоими PMP является тот факт, что оба рабочих места задумывались и действительно сделаны как гибкие, лабильные, т.е. пользователю предоставлена возможность дополнить или поправить некоторые параметры и группы параметров в ходе работы.

Например, очевидным представляется, что отраженный на рис. 2 список типичных социальных ситуаций не полон. Именно поэтому последней строкой в этом списке идет пустая ячейка, которую разметчик может заполнить самостоятельно (например, добавить «Разговор с представителем власти», которого в данном списке отчетливо недостает). В этом случае добавленное значение данного параметра будет направленно в соответствующий столбец результирующей таблицы Excel (столбец «Социальная ситуация»), однако само значение параметра («Разговор с представителем власти») будет в результирующей таблице помечено специальным знаком ($), чтобы супервайзер разметки мог проверить, какое именно значение, не предусмотренное первоначальным планом, было введено разметчиком.

В случае, если дополнение признается существенным, оно может быть внесено в данный список окончательно с помощью окошка редактирования (см. рис. 7).

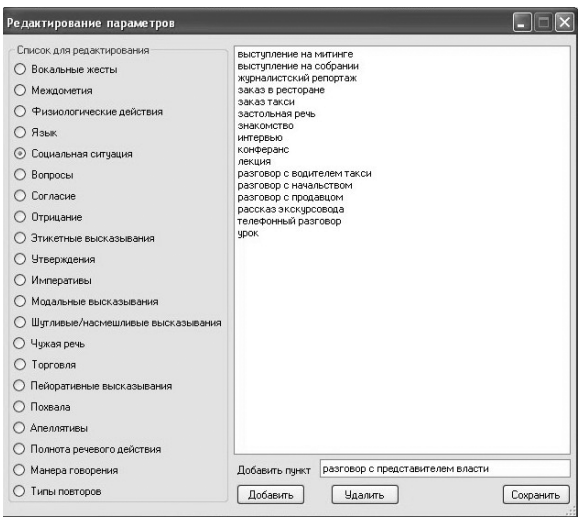


Рис. 7. Окно редактирования параметров PMP Marker

Похожим образом устроено редактирование параметров в PMP GesturesMarker. Так, например, в случае, если мы хотим пополнить список удлинителей при описании жестов, мы вносим название удлинителя в список, который можно видеть на рис. 8.

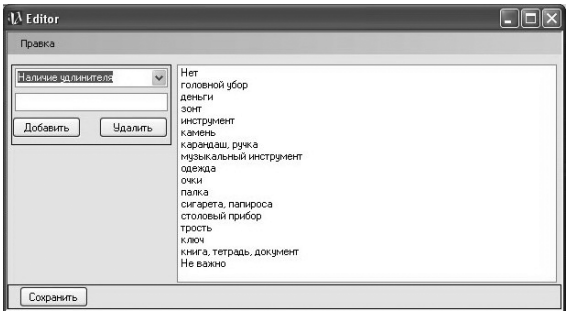


Рис. 8. Окно редактирования параметров PMP GesturesMarker (списки)

Поскольку характеристика жестов содержит гораздо большее число параметров, чем описание речевых действий (и некоторые из них никогда не фиксируются вместе), описание жестов формирует нечто похожее на «дерево заполнения», терминальными вершинами которого являются таблицы-списки, в которые можно вписывать недостающие варианты:

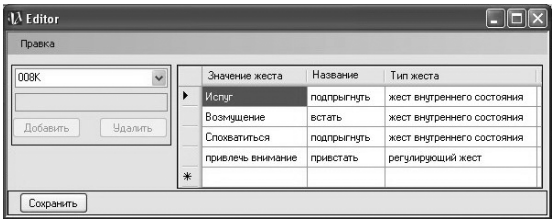


Рис. 9. Окно редактирования параметров PMP GesturesMarker (таблицы)

5. Заключение

Итак, положительными моментами в описанных рабочих местах разметчика (PMP) являются следующие:

1) удобный и интуитивно понятный интерфейс, который позволяет быстро освоить PMP
2) быстрота обработки больших массивов клипов
3) единообразие описания чрезвычайно разнородного материала
4) возможность трансформировать PMP в зависимости от вновь полученных данных

Разработка этих программных оболочек, тем самым, позволяет в максимальной степени оптимизировать сроки и трудозатраты, необходимые для создания МУРКО и, возможно, других мультимедийных корпусов.

Приложение 1. Список речевых действий, используемых для разметки с помощью PMP Marker

Тип речевых действий	Речевые действия
Вопрос	общий, частный, косвенный, контактный, обратной связи, нечленораздельный, переспрос
Согласие	подтверждение, понимание, признание, разрешение, согласие, подчинение
Отрицание	возражение, дистанцирование, запрет, недоверие, недовольство, незнание, непонимание, опровержение, отказ, отрицание, поправка, сомнение, спор, удивление
Этикетные высказывания	благодарность, извинение, пожалуйста, пожелание, поздравление, предложить помощь, представиться, приветствие, приглашение, проводы, прощание, соболезнование
Утверждения	аналогия, аргумент, вспомнить, вывод, догадаться, заявление, комментарий, констатация, объявление, объяснение, перечисление, подсказка, предположение, предсказание, предупреждение, рапорт, рассказ, сентенция, совет, сообщение, указание на кого/что, уговор, утверждение, уточнение
Императивы	баюканье, заказ, инструкция, команда, настаивать, поучение, предложение, предостережение, приказ, распоряжение, стоп!, торопить, требование, уговоры, успокаивать, утешение
Модальные высказывания	беспокойство, ввод информации, восклицание, горе, жалоба, клятва, молитва, намерение, напоминание, ничего!, обвинение, обещание, просьба, раскаяние, сочувствие, уверенность
Шутливые/ насмешливые высказывания	ирония, намек, насмешка, шутка
Чужая речь	пересказ, повтор подсказки, цитирование
Торговля	реклама, торг
Пейоративные высказывания	брань, оскорбление, критические замечания, порицание, проклятье, стыдить, угроза, упрек
Похвала	похвала, похвальба, тост, удовлетворение
Апеллятивы	зов, обращение, обращение к животному, отклик на обращение, отклик на пароль, пароль, привлечение внимания, призыв, призыв к порядку

Приложение 2. Список жестов, используемых для разметки с помощью PMP GesturesMarker[8]

Значение жеста	Название жеста	Тип жеста
идентификация собеседника	ткнуть пальцем	дейктические жесты
общеуказательный	двинуть бровями, демонстрация, касаться чего-л., кивнуть, коснуться кого-л., коснуться чего-л., обвести рукой собеседников, подбросить что-л., показать глазами, показать на себя, показать пальцами, показать пальцем, показать подбородком, показать руками, показать рукой, показывать на себя, показывать пальцем, показывать руками, положить руку на что-л., постукивать пальцем по чему-л., похлопывать по чему-л., предъявить что-л., предъявлять что-л., пристукнуть по чему-л., провести рукой над чем-л., ткнуть пальцем, трясти чем-л., шлепнуть кого-л. по какой-л. части тела	
самоидентификация	показать пальцем на себя, показывать на себя	
фиксация объекта	двинуть кисть к кому-л., двинуть ладони к собеседнику, коснуться чего-л., подбросить что-л., поднять брови, поднять палец, показать пальцем, постукивать пальцем по чему-л., пристукнуть по чему-л., ткнуть пальцем, ткнуть рукой	

[8] В названиях жестов различаются совершенный вид (как обозначение однократных жестов) и несовершенный вид (как обозначение многократных жестов).

Значение жеста	**Название жеста**	**Тип жеста**
общедекоративный	откинуть голову, погладить шею, подтянуть брюки, поправить галстук, поправить одежду, поправить прическу, поправлять галстук, поправлять одежду, поправлять прическу	декоративные жесты
возражение	качнуть головой	жесты — речевые действия
договор	трясти чью-л. руку	
клятва	бить себя в грудь, вскинуть руки, прижать руку к груди, прижимать руки к груди, сплести пальцы	
назидание	двинуть головой вперед, повести подбородком вбок, поднять брови, поднять палец, ткнуть пальцем, ткнуть рукой	
намек	кивнуть, поднять брови, поднять бровь	
насмешка	двинуть головой вперед, запрокинуть голову, мотать головой, натянуть головной убор на глаза собеседника, отвернуться, повести подбородком вбок, подмигнуть, поднять брови, поднять бровь, потереть лицо, смерить глазами, широко раскрыть глаза	
одобрение	выдвинуть подбородок, двинуть головой вперед, качать головой, кивать, кивнуть, махать рукой, похлопывать кого-л. по корпусу, трепать кого-л. по голове	
отказ	двинуть кисть к кому-л., заслониться рукой, качать головой, качать пальцем, качнуть головой, коснуться кого-л., махать рукой, отвернуться, отмахнуться, отрезать рукой, оттолкнуть кого-л., оттолкнуть что-л., показать кукиш, положить руку на чью-л. руку, прикрыть посуду рукой, стоп!, шлепнуть кого-л. по губам	
отрицание	двинуть кисть к кому-л., качать головой, качнуть головой, махать рукой, отрезать рукой, повести подбородком вбок, показать кукиш, стоп!	
подтверждение	закрыть глаза, кивать, кивнуть, толкнуть кого-л.	
подчинение	кивнуть, поклониться	
поздравление	аплодировать	
понимание	выдвинуть подбородок, закрыть глаза, запрокинуть голову, кивать, кивнуть, откинуть голову	
похвала	аплодировать, показать большой палец	
предложение	протянуть руки к кому-л., раскинуть руки	
просьба	броситься к кому-л. на грудь, взять чью-л. руку, гладить кого-л. по голове, качнуть головой, положить руку на плечо, поцеловать в щеку, прижать руку к груди, притянуть к себе кого-л., протянуть руку к кому-л., сплести пальцы	
секретничать	взять под локоть, наклониться к собеседнику, прикрыть рукой телефонную трубку	
согласие	кивать, кивнуть, махнуть рукой	
требование	махнуть рукой, стукнуть по чему-л., стучать кулаками, топать ногами	
угроза	грозить кулаком, грозить пальцем, замахнуться, качать головой, лицо дернулось, погрозить кулаком, прищуриться	
упрек	двинуть головой вперед, качнуть головой, повести подбородком вбок	
я так думаю	коснуться пальцем виска	
беспокойство	прижать руку ко рту	жесты внутреннего состояния
благоговение	прижать к груди что-л.	
вожделение	коснуться кого-л., обводить руками контуры чьего-л. тела	
возмущение	всплеснуть руками, встать, качать головой, оттолкнуть что-л., поднимать глаза к небу, посмотреть в сторону, потрясать руками, потрясать рукой, прижать руки к груди, развести руками, стучать кистями друг о друга, схватиться за сердце, топнуть ногой, широко раскрыть глаза	
вызов	засунуть руки в карманы, опереться спиной на что-л., положить руки на бедра	

Значение жеста	Название жеста	Тип жеста
высокая оценка	вдохнуть, вскинуть руки, вскинуть руку, закрыть глаза, качнуть головой, коснуться кого-л., мотать головой, отвернуться, откинуть голову, отодвинуться от собеседника, повести подбородком вбок, поднять руку к небу, посмотреть отодвинувшись, прижать к груди что-л., прижать руку к груди, раскинуть руки, стоп!, схватиться за голову	жесты внутреннего состояния
горе	закрыть лицо руками, закусить губу, схватиться за голову	
готовность	заложить руку за спину, засучить рукава, лизнуть палец, надвинуть головной убор, натянуть перчатку, опереться руками на что-л., поправить галстук, поправить одежду, поправлять одежду, сплести пальцы, стукнуть себя по бедру	
дистанцирование	барабанить пальцами по чему-л., засунуть руки в карманы, качнуть головой, мерить глазами, надвинуть головной убор, обводить глазами собеседника, опереться головой на руку, отвернуться, отвести руку в сторону, отряхивать руки, поднимать глаза к небу, поднять брови, поднять бровь, поднять глаза к небу, пожать плечами, пожать плечом, поигрывать чем-л., развести руками, сжать губы, склонить голову, скрестить ноги, сложить руки за спиной, сложить руки на груди, смерить глазами, соединить пальцы домиком	
догадаться	махнуть рукой, прикрыть рот рукой, стукнуть себя по голове, схватиться за голову	
дружелюбие	положить руку на плечо, хлопнуть кого-л. по колену	
задуматься	барабанить пальцами по чему-л., взяться за подбородок, закусить губу, облизнуть губы, поглаживать мочку уха, поднести руку ко рту, поправлять одежду, посмотреть вдаль, потереть лицо, потереть подбородок, потирать голову, потирать лицо, почесывать голову, проводить взглядом, сдвинуть головной убор назад, сжать губы, сложить руки за спиной, теребить что-л., тереть лицо, тереть подбородок, чесать за ухом	
игривость	погладить кого-л. по лицу, поигрывать чем-л., поцеловать кого-л., щелкнуть по носу	
интерес	коснуться чьего-л. подбородка, оглянуться, положить руки на бедра, посмотреть поверх очков, склонить голову набок	
испуг	зажмуриться, закрыть лицо рукой, заслониться, отпрянуть, отступить, подпрыгнуть, прикрыть рот рукой, смерить глазами, схватиться за голову, схватиться за сердце, улыбка сбежала с лица, широко раскрыть глаза	
кокетство	воздушный поцелуй, вытянуть руки, задрать ногу назад, запрокинуть голову, отвести голову, повернуться к собеседнику, повести подбородком вбок, подмигнуть, поднять брови, покоситься, склонить голову набок, сморщить нос, щелкнуть по носу	
лукавство	брови домиком, качать головой, поднимать брови, посмотреть искоса	
манерность	оттопырить мизинец	
недоверие	двинуть головой вперед, зажмуриться, мотать головой, смерить глазами	
недовольство	выпятить губу, закусить губу, качать головой, махнуть рукой, нахмуриться, отвернуться, повести подбородком вбок, поднять бровь, поморщиться, потупиться, прищуриться, сбросить чью-л. руку, сжать губы, скривиться, смерить глазами, улыбка сбежала с лица	
недоумение	выпятить губу, закусить губу, коснуться чего-л., моргать, открыть рот, повести подбородком вбок, поднять брови, поднять бровь, пожать плечами, прикрыть рот рукой, развести руками	
неожиданность	вздрогнуть, оглянуться, отпрянуть	
нервозность	барабанить пальцами по чему-л., вытереть пот, ломать пальцы, облизнуть губы, опереться головой на руку, поигрывать чем-л., потирать руки, похлопывать по чему-л., похлопывать чем-л., почесываться, прижать руку к груди, теребить что-л.	
нетерпение	поправить ремешок часов, притоптывать	

Значение жеста	Название жеста	Тип жеста
облегчение	закрыть глаза, поднять глаза к небу, положить голову на чье-л. плечо	жесты внутреннего состояния
огорчение	качать головой, опереться головой на руку, посмотреть вдаль, стукнуть по чему-л.	
ожидание	барабанить пальцами по чему-л., положить руки на бедра, проводить взглядом, сложить руки за спиной, сложить руки на груди, упереться рукой в бок	
озабоченность	качать головой, покоситься, улыбка сбежала с лица	
опасение	взяться за подбородок, оглянуться	
осторожность	взяться за подбородок, закусить губу, заслониться, надвинуть головной убор, оглядываться, оглянуться, отвернуться	
отвращение	поморщиться	
отчаяние	закрыть глаза, запрокинуть голову, махнуть рукой, прижать руку ко рту, сплести пальцы, стукнуть кулаками по чему-л., стукнуть по чему-л., схватиться за голову	
печаль	склонить голову	
подобострастие	присесть	
потрясение	обводить глазами окружающих, смотреть по сторонам	
предвкушение	облизнуть губы, потирать руки	
пренебрежение	дергать уголком рта, качать головой, качнуть головой, махнуть рукой, отшвырнуть что-л., повести подбородком вбок, поводить рукой, пожать плечом, предъявить что-л., сжать губы, скривиться	
привязанность	гладить кого-л. по голове, гладить кого-л. по плечу, коснуться чьего-л. лица, обнять кого-л., погладить по голове, положить голову на чье-л. плечо, положить руку на плечо, поправить чужую одежду, поправлять кому-л. волосы, поправлять чужую одежду, притянуть к себе чью-л. голову, трепать кого-л. по щеке	
радость	встать на цыпочки, расправить плечи, склонить голову набок, сплести пальцы, широко раскрыть глаза	
раздражение	двинуть кисть к кому-л., дернуть уголком рта, закатить глаза, отвернуться, подбросить что-л., поднимать глаза к небу, поднять глаза к небу, поставить что-л. с размахом	
разочарование	закрыть глаза, отодвинуться от собеседника, поднять брови, сжать губы, улыбка сбежала с лица, шлепнуть себя по бедру	
расслабленность	опереться спиной на что-л., поигрывать чем-л., похлопывать чем-л., скрестить ноги	
растерянность	обводить глазами окружающих, осматривать себя, ощупывать себя, потирать голову, потирать грудь, потирать шею	
решимость	выдохнуть, заложить руку за спину, засучить рукава, надвинуть головной убор, опереться руками на что-л., поправить галстук, рубануть воздух кулаком, сдуть волосы со лба, сплести пальцы, стукнуть себя по бедру, упереться рукой в бок	
скука	барабанить пальцами по чему-л., зевнуть, опереться головой на руку	
смущение	втянуть голову в плечи, отвернуться, отвести взгляд, пожать плечами, пожать плечом, поигрывать чем-л., покоситься, потупиться, прикрыть рот рукой, склонить голову набок, сплести пальцы, сплетать пальцы, теребить что-л., чертить что-л.	
солидарность	взять за руку, оглянуться, подмигнуть	
сосредоточенность	наклонить голову, нахмуриться, прищуриться, сжать губы	
сочувствие	качнуть головой, притянуть к себе чью-л. голову, трепать кого-л. по голове	
спохватиться	вскинуть руки, махнуть рукой, поднять брови, подпрыгнуть, прикрыть рот рукой, схватиться за голову	
стыд	закрыть глаза, закрыть лицо руками, отвернуться, потупиться	
уверенность	потирать руки, сплести пальцы	

Значение жеста	Название жеста	Тип жеста
удивление	вскинуть руки, всплеснуть руками, двинуть головой вперед, кивнуть, коснуться чего-л., мерить глазами, наклониться к собеседнику, остановиться, откинуть голову, открыть рот, отодвинуться от собеседника, отпрянуть, переглядываться, переглянуться, повести подбородком вбок, поднять брови, поднять бровь, пожать плечами, покоситься, посмотреть пристально, посмотреть пристально, прищуриться, проводить взглядом, развести руками, смерить глазами, сплести пальцы, широко раскрыть глаза	жесты внутреннего состояния
удовлетворение	отодвинуться от собеседника, отряхивать руки, повести подбородком вбок, подбочениться, положить руки на бедра, потирать нос, потирать руки, похлопывать по чему-л., расправлять усы, склонить голову набок, стоп!, упереться рукой в бок, хлопнуть кого-л. по плечу	
что тут поделаешь!	качнуть головой, махнуть рукой, опереться спиной на что-л., отвести руку в сторону, повести подбородком вбок, пожать плечами, положить руки на бедра, развести руками	
ясно без слов	отвести руку в сторону, развести руками	
действие (бить)	взмахнуть рукой	изобразительные жесты[9]
действие (встать)	поднять руку вверх	
действие (задуматься)	поднять палец	
действие (закрыть)	махнуть рукой	
действие (играть на музыкальном инструменте)	перебирать пальцами	
действие (отменить)	отрезать руками	
действие (отодвинуться друг от друга)	разводить руками	
действие (писать)	рисовать в воздухе	
действие (смерть)	высунуть язык, закатить глаза	
действие (смерть)	провести пальцем у горла	
действие (схватить)	сжать кулаки	
действие (трепетать)	перебирать пальцами	
качество (красота)	показать руками	
качество (крепко)	взмахнуть рукой, нахмуриться, сжать кулак	
качество (отсутствие чего-л.)	вывернуть карманы, отрезать рукой	
качество (размер)	показать уровень, провести рукой по горлу, развести руки в сторону	
качество (точность)	двигать рукой, соединив кончики пальцев, соединить кончики пальцев	
качество (уродство)	скорчить рожу	
количество (много)	показать уровень	
количество (пять)	растопырить пальцы	
направление (прочь)	махнуть рукой	
объект (вода)	перебирать пальцами	
объект (парта)	расставить руки	
топология (пространство)	отвести руку в сторону	
молитва	поклониться, стоя на коленях	корпоративные жесты
пароль	показать три пальца	
пионерское приветствие	салютовать	
воинское подчинение	отдать честь	

[9] В таблице оставлены только те немногие жесты, для которых удалось подобрать названия.

Значение жеста	Название жеста	Тип жеста
дразнить	показать язык	пейоративные жесты
дурак!	крутить пальцем у виска, постукивать пальцем по виску, сплюнуть, стучать кулаком по голове	
критические замечания	взмахнуть рукой, грозить пальцем, качать головой, махнуть рукой, поморщиться, прижать руку к груди, протянуть руку к кому-л.	
передразнивание	мотать головой	
подумаешь!	вскинуть руку	
пошел вон!	раскинуть руки	
сумасшедший!	крутить пальцем у виска, постукивать пальцем по виску	
искать что-л.	оглядываться, оглянуться, тереть лицо, шарить по карманам	поисковые жесты
обратить внимание	мерить глазами, оглянуться, переглядываться, переглянуться, придвинуться к чему-л., проводить взглядом	
оценивать обстановку	оглядываться	
оценка веса	подбрасывать в руке что-л., подбросить в руке что-л.	
оценка внешнего вида	смерить глазами	
оценка времени	посмотреть на часы	
оценка температуры	ощупывать что-л., пощупать что-л.	
поиск слова	трясти рукой	
узнавание	выдвинуть подбородок	
достаточно!	остановить жест собеседника	регулирующие жесты
задавать ритм	барабанить пальцами по чему-л., дирижировать	
замолчи!	прикрыть кому-л. рот	
запрет	грозить пальцем, заслонить руками что-л., качнуть головой, стоп!	
иди!	вскинуть руку, выдвинуть подбородок, качнуть головой, кивнуть, манить пальцем, махать рукой, махнуть рукой, повести кого-л. за собой, поманить, толкнуть кого-л., хлопнуть кого-л. по плечу	
начали!	кивнуть, коснуться кого-л., махать рукой, махнуть рукой	
не трогай!	придержать что-л. рукой	
остановить кого-л.	дернуть кого-л. за какую-л. часть тела, придержать кого-л., развернуть кого-л. лицом к себе	
остановить машину	голосовать	
подбодрить	коснуться кого-л., похлопывать по плечу, шлепнуть кого-л. по спине	
прекрати!	грозить пальцем, дать кому-л. подзатыльник, натянуть головной убор на глаза собеседника, остановить жест собеседника, оттолкнуть кого-л., придержать кого-л., скрестить руки, стоп!, толкнуть кого-л., хлопнуть кого-л. по плечу, шлепнуть кого-л. по руке	
привлекать/привлечь внимание	дергать кого-л. за одежду, махать руками, подталкивать кого-л., постукивать кого-л. по какой-л. части тела, постукивать пальцем по какой-л. части тела, протягивать руки к кому-л., стучать по спине, трясти кого-л., указывать пальцем, хлопать в ладоши, выдвинуть подбородок, дать пощечину кому-л., дернуть кого-л. за какую-л. часть тела, дернуть кого-л. за одежду, коснуться кого-л., махнуть рукой, обвести рукой собеседников, подмигнуть, поднять палец, поднять руки, положить руки на плечи кому-л., положить руку на плечо, привстать, протянуть руку к кому-л., стоп!, ткнуть пальцем, толкнуть кого-л., толкнуть кого-л. локтем, хлопнуть в ладоши, хлопнуть кого-л. по плечу	
призыв к порядку	дернуть кого-л. за какую-л. часть тела, качать пальцем, коснуться кого-л., поднять палец, посмотреть строго, стукнуть по чему-л., стучать по чему-л., ткнуть пальцем, толкнуть кого-л. локтем	
самоуспокоение	схватиться за сердце	
торопить	похлопывать по спине, стучать по часам	
успокаивать	коснуться кого-л., махнуть рукой, поглаживать кого-л., положить руки на плечи кому-л., положить руку на плечо, похлопывать по плечу, поцеловать кого-л., протянуть руку к кому-л., стоп!, хлопнуть кого-л. по колену	
утешение	положить руку на плечо	

Значение жеста	Название жеста	Тип жеста
интенсификация действия	отдать что-л., размахнувшись	риторические жесты
материализация речи: аргумент	двинуть кисть к кому-л., двинуть кистью на себя, ткнуть пальцем	
материализация речи: брань	бодать собеседника, боднуть собеседника, двинуть кулаками, ритмические биение, ткнуть пальцем	
материализация речи: возражение	боднуть собеседника, двинуть кистью от себя, махнуть руками, махнуть рукой, развести руками, ритмическое биение	
материализация речи: вопрос	вскинуть руку, выдвинуть подбородок, двинуть головой вперед, двинуть кисть к кому-л., двинуть ладони к собеседнику, кивнуть, коснуться кого-л., наклониться к собеседнику, отвести руку в сторону, развести руками, ритмическое биение	
материализация речи: отрицание	мотать головой	
материализация речи: перечисление	загибать пальцы при счете, кивнуть, пристукивать по чему-л., ритмическое биение, рубить воздух ладонями	
материализация речи: побуждение	выдвинуть подбородок, двинуть кисть к кому-л., махнуть рукой, постукивать пальцем по чему-л., похлопывать по спине	
материализация речи: предложение	двинуть ладони к собеседнику	
материализация речи: просьба	выдвинуть подбородок, двинуть кисть к кому-л., двинуть кистью на себя, двинуть ладони к собеседнику, повести подбородком вбок, ритмическое биение	
материализация речи: требование	ритмические биение, ритмическое биение, рубануть воздух руками	
материализация речи: убеждение	двинуть головой вперед, коснуться кого-л., махать рукой, наклониться к собеседнику, присесть, ритмические биение, ритмическое биение, рубить воздух ладонью	
материализация речи: утверждение	выдвинуть подбородок, двигать головой вперед, двинуть головой вперед, двинуть кисть к кому-л., кивать, кивнуть, махнуть рукой, наклониться к собеседнику, отрезать рукой, пристукнуть по чему-л., разрубить воздух руками, ритмические биение, ритмическое биение, рубануть воздух рукой, рубить воздух ладонью, стукнуть кулаками по чему-л., стукнуть по чему-л.	
материализация речи: финал	вскинуть руку, двинуть кисть к кому-л., кивнуть, пожать чью-л. руку, пристукнуть пальцем по чему-л., пристукнуть по чему-л., рубануть воздух руками, рубануть воздух рукой, стукнуть кулаками по чему-л., факел	
материализация речи: чужая речь	двигать кистью на себя, двинуть кистью на себя	
новая тема	запрокинуть голову	
потребность в поддержке	смена собеседника	
предвосхищение отрицания	мотать головой	
предвосхищение согласия	кивать, кивнуть	
пьяный	щелкнуть пальцем по горлу	условные жесты
сдаваться	руки вверх	
тост	чокнуться	
больно	вздрогнуть, держаться за больное место, дунуть на больное место, потереть больное место, сморщиться, схватиться за больное место, тереть больное место, трясти рукой	физиологические жесты
горько	выдохнуть, помахивать рукой у рта, сморщиться	
делать что-л. с усилием	поднять брови	
душно	ослабить ворот, ослабить галстук	
жажда	облизнуть губы	

Значение жеста	Название жеста	Тип жеста
кричать	двинуть головой вперед, оттянуть локти назад, приставить руку ко рту	физиологические жесты
не видно	вытянуть шею	
не слышно	двинуть головой вперед, нахмуриться, придвинуть ухо	
неприятно	сморщиться	
нервно	руки дрожат	
стереть чужое прикосновение	потереть рукой	
ударить	замахнуться	
усталость	вытереть пот, закрыть глаза, опереться головой на руку, тереть лицо	
холодно	вздрогнуть	
чешется	почесываться	
чисто	вытереть нос, утереть рот	
благодарность	аплодировать, выдвинуть подбородок, двукратный поцелуй, кивать, кивнуть, коснуться кого-л., поклониться, положить руку на чью-л. руку, поцеловать руку, прижать руку к груди, рукопожатие, тряхнуть кого-л. за плечи	этикетные жесты
вы правы!	стоп!	
зевать	похлопывать рукой по рту	
извинение	кивнуть, повести подбородком вбок, прижать к груди руку собеседника, прижать руку к груди	
не стоит благодарности	качнуть головой, кивнуть, отмахнуться	
помощь	поддержать кого-л.	
при кашле	прикрыть рот рукой	
при смехе	закрыть лицо рукой, прикрыть рот рукой	
приветствие	вскинуть руку, вскинуть руку, вскинуть руку к голове, вскинуть руку к голове, вскинуть руку к голове, встать, кивнуть, коснуться кого-л., махать рукой, махать рукой, махнуть рукой, обнять кого-л., поклониться, поцеловать кого-л., поцеловать руку, протянуть руки к кому-л., раскинуть руки, рукопожатие, снять головной убор, снять очки	
приглашение	показать рукой	
прощание	воздушный поцелуй, вскинуть руку, вскинуть руку к голове, кивнуть, махать рукой, махнуть рукой, поцеловать кого-л., рукопожатие, троекратный поцелуй	
спросить разрешения	кивнуть	
я слушаю	кивнуть	

Литература

1. *Гришина Е. А.* Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006–2008: новые результаты и перспективы (в печати) [2009а]
2. *Гришина Е. А.* Национальный корпус русского языка как источник сведений об устной речи // Речевые технологии (в печати) [2009б]
3. *Крейдлин Г. Е.* Невербальная семиотика. М.: 2002
4. *Ляшевская О. Н., Плунгян В. А., Сичинава Д. В.* О морфологическом стандарте Национального корпуса русского языка // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: 2005, с. 111–136
5. *Сичинава Д. В.* Обработка текстов с грамматической разметкой: инструкция разметчика // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: 2005, с. 136–154

Средства настройки процессора Semantix на предметную область

Means for tuning of linguistic processor «Semantix» on subject field

Кузнецов И. П. (igor-kuz@mtu-net.ru)
Институт проблем информатики Российской академии наук

Ефимов Д. А. (d.efimov@synsys.ru)
ЗАО «Синергетические Системы»

Рассматривается семантико-ориентированный лингвистический процессор, осуществляющий автоматическую формализацию потоков текстов на естественном языке. В качестве исходного материала использован корпус текстов, связанный с описанием памятников. Из таких текстов выделяются информационные объекты: памятники, места их расположения, лица, их ролевые функции, события с указанием участия в них лиц и др. Рассматривается инструментальная среда, позволяющая быстро находить ошибки процессора и устранять их, подстраивая лингвистические знания.

Введение

С каждым годом возрастает совокупный объем цифровой информации. В основном, это тексты естественного языка (ЕЯ). Важная проблема — выбор из этих текстов информации, необходимой для профессиональной деятельности той или иной категории пользователей, и предоставление ее в форме, удобной для восприятия.

Следует учитывать, что большинство пользователей интересуются лишь конкретными вещами. Например, следователям важны фигуранты, их места жительства, телефоны и др. Специалистов по кадрам интересуют организации, где человек работал, кем и когда это было. Других интересуют памятные места, их местонахождение, кто автор, архитектор и т.д. Подобную информацию будем называть ***информационными объектами***, которые различаются по типам. Например, лица и фигуранты — это объекты одного типа, адреса — другого и т.д. Одно из важнейших направлений автоматической обработки потоков текстов — выявление информационных объектов и связей из текстов ЕЯ с ориентацией на определенную категорию пользователей. Это направление связано с формализацией текстов и относится к области «извлечение знаний» (Knowledge Extraction). При этом результаты должны быть представлены в формах, к которым привык пользователь (или же в формах, удобных для последующей обработки — поиска, экспертных оценок). Эта область в силу ее актуальности привлекает все больше исследователей, [4,5 и др.].

Отметим, что связи между объектами могут иметь высокую степень разнообразия. Например, памятники могут быть связаны не только с местом или лицами, которым они посвящены, но и с действиями или событиями — инициированием работ, проектированием, архитектурными работами, изготовлением отдельных компонент (постамента, фигур,...) и многое другое. Такие события привязаны к времени, месту, связаны с лицами, участвующие в создании памятников. Одни события могут быть составной частью других. Они могут быть связаны причино-следственными и временными отношениями. Таким образом, события — это тоже информационные объекты, связанные между собой и с другими информационными объектами. В ЕЯ такие связи выражаются с помощью глагольных форм, форм с отглагольными существительными, различными оборотами. Возникают сложные структуры.

Для представления подобных структур и их формализации были разработаны ***расширенные семантические сети*** (РСС). Для автоматического анализа текстов ЕЯ с их отображением на РСС в рамках проектов ИПИ РАН была разработана научная база для построения ***семантико-ориентированных лингвистических процессоров*** (ЛП): методики представления сложных видов знаний, ***инструментальная***

среда ДЕКЛ для обработки структур знаний, сетевые позиционные грамматики, онтологии, морфологический анализ на основе обобщенных окончаний, базы знаний (БЗ) на РСС, различные виды объектных поисков []. На этой основе разработан ряд прикладных систем [3,10,11]. Последний вариант ЛП, изготовленного совместно с ЗАО «Синергетические Системы» в виде модуля SDK, получил название Semantix и иллюстрировался на предыдущей конференции Диалог-08 [11]. Рассматривались различные аспекты организации подобных ЛП и особенности их работы.

Целью настоящей статьи является анализ важной компоненты семантико- ориентированных ЛП — средств настройки на предметную область. Здесь важную роль играют следующие факторы: как устроены лингвистические знания (ЛЗ), а также средства и методики их отладки или подстройки под предметную область пользователя. В качестве примера использован корпус текстов, связанный с описанием памятников. Данная предметная область достаточно интересная — со своими особенностями. В ней имеют место стандартные объекты (лица, даты, организации, профессии), а также объекты — памятники (компоненты их описания), места расположения, связанные с ними лица, события.

С точки зрения лингвистики область не является тривиальной. Пользователю важно знать, кто является инициатором памятника, архитектором, скульптором и др. Когда произошла закладка памятника, его установка, открытие и т.д. Чтобы выделить эти сведения, необходим глубинный анализ текста с выявлением имеющих место событий, для которых ищется привязка ко времени и месту.

1. Особенности семантико-ориентированных процессоров

Семантико-ориентированные ЛП, осуществляющие выявление информационных объектов и связей, основаны на правилах выделения компонент текста (слов, словосочетаний), из которых составляются информационные объекты. Такие правила в той или иной степени учитывают наличие ключевых слов, признаки слов (лексические, морфологические, семантические), взаимное расположение слов (их позиции), а также контекст.

В настоящее время развиваются два основных направления, связанных с построением семантико-ориентированных ЛП. Первое — когда правила «вшиваются» в программы. Создаются блоки, анализирующие слова, признаки, согласованность слов, наличие комбинаций цифр, знаков и др. Из таких блоков строятся правила. Каждое правило — это программа анализа с выделением объектов. Такие информационные объекты, как лица (ФИО), даты, адреса и организации с достаточной степенью надежности выделяются программными средствами. Анализ предложений сводится к выявлению наличия в них объектов, ключевых слов, значимых глаголов, например, *СОЗДАТЬ… ПРОЕКТ…*<лицо> и др. Лингвист задает правила анализа (модели), на основе которых из блоков строятся программы. При настройке на новые объекты лингвисту нужно разрабатывать новые правила, а программисту строить новые программы. Если лингвист чего-то не учел, то и программы будут давать ошибочные результаты. Нужно снова обращаться к лингвисту и т.д. В силу многообразия языковых конструкций, используемых при описании объектов (даже на корпусах текстов сравнительно небольших объемов), учесть все варианты представляется крайне трудной проблемой. Поэтому процесс настройки будет многоэтапным с определенной степенью сходимости.

В тоже время, программные средства постоянно совершенствуются (C#, .DOT и др.). Правила могут быть оформлены как программные объекты с экземплярами, учитывающими различные детали. В связи с этим процесс построения правил упрощается. Данный подход эффективно применяется в тех случаях, когда не требуется сложных видов анализа, например, связанных с выделением семантически связанных слов, событий с их атрибутами и др. Это направление развивается в ряде организаций, в том числе, в ЗАО «Синергетические Системы».

Второе направление, когда программа ***лингвистического процессора*** (ЛП) отделяется от ***лингвистических знаний*** (ЛЗ). Последние состоят из правил выделения объектов, включают в себя предметные словари, а также другие средства, определяющие всю процедуру анализа. ЛЗ имеют вид декларативных структур, которые легко менять и настраивать. В нашем случае роль таких структур выполняют фрагменты РСС [1,2]. Настройка ЛП осуществляется только за счет разработки ЛЗ, определяющих набор выделяемых объектов и связей. Задача ЛП — поддерживать ЛЗ, в том числе, процесс применения правил. При использовании подобных ЛП облегчается настройка на корпуса текстов, особенности предметной области. Корректировать ЛЗ может человек, обученный формализму РСС и знакомый с элементами математической лингвистики. Ему не нужно уметь программировать. Тогда возникает вариант, когда один человек может настраивать ЛП — находить ошибки и устранять их.

В данной статье речь будет идти о таких ЛП, к которым относятся процессоры системы «Аналитик» (с ее приложениями — системами «Криминал», «Обработчик резюме» и др.), а также процессор Semantix. Они все работают по одному принципу. Будем называть их **процессорами типа Semantix** (или просто Semantix). Отметим, что перечисленные системы отличаются только ЛЗ, которые определяют область приложений [3,6].

Основные компоненты семантико-ориентированного ЛП, основанного на ЛЗ:

1.1. Блок лексико-морфологического анализа

Выделяет из документа слова и предложения и выдает в виде семантической сети (***ПС-документа***), представляющей последовательность компонент (слов в нормальной форме, чисел, знаков) и их основные признаки, Использует набор предметных словарей (словарь стран, регионов России, имен, профессий и др.) для придания словам и словосочетаниям дополнительных семантических признаков [9].

1.2. Блок синтактико-семантического анализа

Путем анализа ПС-документа выделяет объекты и связи. На их основе строит другую семантическую сеть, представляющую семантическую структуру (***СС-документа***), называемую ***содержательным портретом*** [2,6,7]. Блок управляется ЛЗ, за счет которых обеспечивается: — Извлечение информационных объектов (лиц, организаций, событий, их места, …). — Выявление связей объектов. Например, как лица связаны с организациями, адресами и др. — Анализ глагольных форм, причастных и деепричастных оборотов с выявлением фактов участия объектов в тех или иных действиях. — Выявление связей действий с их местом или временем (где и когда имело данное действие или событие). — Анализ причино-следственных и временных связей между действиями и событиями.

1.3. Блок построения каталогов объектов

Выделяет из СС-документов объекты определенного типа, которые упорядочиваются по алфавиту и образуют каталог. Например, таким способом создаются каталоги лиц (их ФИО), дат, адресов и др. — только тех, которые встретились в документах.

1.4. База лингвистических знаний (ЛЗ)

Содержит правила анализа текста во внутреннем представлении (РСС). Они определяют работу ЛП.

Отметим, что с каждым годом растет количество теоретических работ в данной области (в том числе, см. последние сборники «Диалог»). Для многих из них доведение до конечного продукта может оказаться проблематичным в силу сложности возникающих проблем. В данной работе рассматривается работающий ЛП, отлаженный на различных предметных областях [11]. Процессоры такого типа регулярно демонстрировались на конференциях «Диалог» в течении 10 лет [2].

2. Проблемы настройки на предметную область

При настройке на предметную область возникают следующие трудности. Во-первых, при наличии разнотипных объектов требуются соответствующие правила их выделения. Качество ЛП определяется трудоемкостью построения таких правил. В процессоре Semantix выделение всех объектов (их может быть более 40 типов) и событий осуществляются правилами, которые конструктивно оформлены одинаковым образом, и соответственно, которые работают однотипным методикам. Поэтому трудоемкость построения не высокая. Важно, что изменяются только ЛЗ, но не программы.

Во-вторых, при настройке возникают частые случаи ***коллизии*** правил выделения: одни правила могут захватывать слова, которые относятся к другим объектам и которые должны обрабатываться другими правилами. В связи с этим правила должны иметь средства их быстрой подстройки, ограничивающие возможность применения. В процессоре Semantix такая подстройка осуществляется за счет изменения списков, задающих допустимые признаки слов, стоящих на тех или иных позициях. В общем случае такие списки организованы в виде И-ИЛИ графов.

В-третьих, важный фактор — это **избирательность правил** и процедур идентификации: ***коэффициент шумов и потерь***. Под шумами понимается наличие лишних слов в объектах. Потери — это когда объект не выявлен или выявлен частично (в тексте есть слова, которые не вошли в объект). В процессоре Semantix правила (составляющие ЛЗ) имеют все средства для повышения степени избирательности правил и минимизацию шумов и потерь при большом количестве выделяемых объектов. С помощью ЛЗ обеспечивается настройка на особенности языка — на типовые конструкции и формы языка с учетом признаков, которые даются словам. Имеются все необходимые удобства в плане создания и корректировки правил.

В-четвертых, определенные трудности вызывает выделение связей. Это не только глубинный анализ глагольных и других форм. Многие связи даются с помощью анафорических ссылок, а также по умолчанию. Требуется организация сложного процесса их поиска. Такие процессы организуются, чтобы связать лицо с его местом проживания или местом работы, идентифицировать слова (*ПАМЯТНИК, ФИГУРА*,..) с объектами (типа *памятник*) и т.д. Эти слова и подразумеваемые объекты могут стоять в тексте на значительном расстоянии. Важно не захватить посторонний объект. В процессоре Semantix для этой цели используются специальные фильтры.

В-пятых, для настройки ЛП на корпуса текстов необходим специальный **комплекс инструментальных средств**, обеспечивающих следующие функции:

- последовательную обработку множества документов (корпуса текстов) с формализацией каждого из них, формированием СС-документов и построением общей БЗ;
- формирование списков выделенных объектов (для каждого типа объектов свой список), осуществляемое в процессе обработки множества документов; такие списки будем называть ***каталогами***;
- возможность выделения в каталоге любого объекта с быстрым поиском документов, из которых выделен данный объект;
- подача на вход ЛП найденного документа с анализом процесса его формализации (формирования СС-документа);
- визуализацию процесса применения правил и осуществляемых ими преобразований;
- трассировку работы каждого правила с указанием, в какой последовательности захватываются слова (благодаря каким признакам), а также, почему и на каком слове процесс применения правила закончился;
- выдачу в одно окно СС-документа и сам документ для их сравнения;
- обращение к ЛЗ с выбором любого правила и его изменением.

Эти средства позволяют быстро находить ошибки в работе ЛП и корректировать ЛЗ. Методика достаточно проста. Берется корпус текстов и включается в последовательную обработку с формированием общей БЗ и каталогов объектов. Они просматриваются. С их помощью легко находить объекты с ***шумами*** (лишними словами). Во многих случаях сразу видны ***потери*** — если по смыслу объект не соответствует своему статусу. Труднее находить потери, когда объект имеется в документах, но не найден. Тогда нужно просматривать каталог «лишних» слов и их комбинаций (которые не вошли ни в какие объекты). Или же просматривать СС-документов и сравнивать их с самими документами. Конечно, идеальный вариант, когда кто-то (лингвист) выделяет из корпуса текстов объекты определенного типа, и они сравниваются с построенным каталогом. Но этот вариант крайне трудоемок, когда имеют место корпуса текстов большого объема, которые постоянно обновляются.

Следует отметить, что подобные инструментальные средства отсутствуют в процессоре Semantix, изготовленного в виде модуля SDK (там выключена интерфейсная компонента). Но они имеются в системах «Аналитик» ("Криминал»), где реализованы различные виды объектных поисков, ответ на запросы в свободной форме, развита интерфейсная компонента. Поэтому процесс разработки и отладки ЛЗ для новой предметной области идет в рамках этих систем. Отлаженные ЛЗ переносятся в модуль SDK.

3. Выделяемые объекты и связи

Набор выделяемых объектов и связей определяется задачами пользователя. В рамках выполнения плановых работ была взята предметная область, связанная с описанием памятников. Работа в этой области явилась еще одним примером достаточной универсальности процессора Semantix. Корпус текста имеет вид:

6. *Памятник руководителю первой русской кругосветной экспедиции Ивану Фёдоровичу Крузенштерну находится на набережной Лейтенанта Шмидта. Решение об его установке было принято в 1869 году, в канун столетия со дня рождения адмирала. Памятник находится напротив здания Морского кадетского корпуса. Торжественная закладка памятника состоялась 8 ноября 1870 года, в день 100-летнего юбилея Крузенштерна. Фигура из бронзы была отлита в декабре 1872 года по модели И. Н. Шредера. Гранитный постамент спроектировал архитектор И. А. Монигетти. Открытие памятника состоялось 6 ноября 1873 года.*

7. *23 мая 1909 года в центре Знаменской площади был открыт конный памятник Александру III. Автор памятника Паоло Трубецкой выполнил его откровенно как карикатуру, что вызвало довольно сильный скандал. В 1937 году памятник перенесли во двор Михайловского дворца. Перенос памятника объяснили тем, что он якобы мешал трамвайному движению, хотя к тому времени трамвай по Невскому проспекту ходил уже около трёх десятилетий.*

8. *Памятник Екатерине II создан в 1862–1873 годах по проекту …*

В качестве основных типов информационных объектов и связей были выбраны следующие:

- памятники (монументы, скульптуры и т.д.);
- лица (кому посвящен памятник, кто участвовал в его создании и др.);
- ролевые функции или профессии (скульптор, архитектор, дизайнер, зодчий, …) ;
- места расположения памятников;
- события с указанием участия в них информационных объектов («памятник создан …», «работа выполнена …»);
- даты, время, интервалы времени; — организации (связанные с созданием памятников);
- связи между различными типами информационных объектов (время и место событий, ролевые функции лиц и др.).

Отметим, что выявление ролевых функций лиц и связей требует анализа различных форм ЕЯ (с однородными членами: отглагольными существитель-

ными, причасными: деепрчаастными оборотами и др.) для выделения событий с их привязкой к времени и месту. Требуется сложный (многоуровневый) лингвистический анализ с учетом различных признаков, в том числе семантических.

4. Правила выделения объектов

Правила, осуществляющие выделение лиц, дат, интервалов времени, профессий, организаций, событий и связей были взяты из предметных областей, на которые уже был настроен процессор Semantix. Они отлаживались на текстах — «Документы о терроризме», «Автобиографии», «Сводки происшествий» и др.

Новые информационные объекты — это памятники и (в значительной степени) места их расположения. Ранее не встречались описания типа: *возле Каменного моста, напротив здания Моссовета, на площади перед Михайловским замком* и др. Например, в сводках происшествий таких описаний не встречается — всегда фигурируют названия мест.

Были разработаны правила выделения памятников (их несколько). Правила имеют левую часть (условие применения) и правую (действия). К примеру, одно из них выглядит следующим образом:

MUSTBE(MONUM~2,1)
STR_OR(*БЮСТ,СТАТУЯ,ФИГУРА,СОБОР,ЦЕРКОВЬ,ХРАМ*/1+)
STR_OR(*WORK_K,NAT_K,ИМПЕРАТОР,ЦЕСАРЕВИЧ,ГРАФ,КНЯЗЬ,ГЕРЦОГ,..*/2+)
STR_OR(*КОГО,КВЧ,MONUM_K,FIO,FAM*/3+)
CONTEXT(1-,2-,3-/MONUM~2)
P_P(MONUM~2,MON~2L)
MONUM_(1,2,3/MON~2L) MON~2L(*MONUM*,ADD_)
MAYBE(MONUM~2,2)

Данное правило записано в формализме РСС и означает следующее. Вызов правила — по его идентификатору MONUM~2. Фрагмент P_P(MONUM~2,MON~2L) разделяет левую и правую части, т.е. CONTEXT(1-,2-,3-/MONUM~2), который задает условие применения, и MONUM_(1,2,3/MON~2L) — что формировать.

Применять правило нужно с 1-й позиции — MUSTBE(MONUM~2,1). Нужно искать ключевые слова, отмеченные 1+ , т.е. *БЮСТ, СТАТУЯ,*... На следующей позиции должно быть одно из слов списка 2+. Это может быть слово с признаком *WORK_K* (т.е. входить в словарь професссий) или с признаком *NAT_K* (словарь национальностей), или же одно из перечисленных далее слов. Фрагмент MAYBE(MONUM~2,2) указывает, что эта позиция факультативная — перечисленных слов может не быть в тексте).

На следующей позиции должно быть одно из слов списка 3+. Это может быть слово с признаком *КОГО* (род. падеж), или слово в кавычках, или слово с признаком *MONUM_K* (словарь памятников с уникальными названиями), или фамилия, или лицо (ФИО). Если условие выполняется, то правило будет применимым. Формируется объект MONUM_(1,2,3) с признаком *MONUM*. Аргументами являются слова (или объекты), которые оказались на позициях 1,2,3. Сформированный объект замещает эти слова и занимает свою позицию: три позиции замещаются на одну.

Правило будет применяться, когда в тексте встречаются описания типа *статуя императора Николая I, Фигура Петра Великого* и т.д.

Для уникальных памятников, которые невозможно выделить через ключевые слова, создан предметный словарь MONUM_K.SLV, фрагмент которого имеет вид:

. . .
Александровская колонна
«Железный Феликс»
«Лысый Камень»
Вандомская колонна
Воин-освободитель
Демидовский столп
<Медный Всадник>
Миноносец «Стерегущий»
Родина-мать
Собор Парижской богоматери
Соловецкий камень
«Шалаш»
«Царь-плотник»
. . .

Если в тексте встретилось одно из этих слов (или словосочетаний), то ему (им) присваивается признак *MONUM_K*, который учитывается правилами из ЛЗ. Следует отметить, что словарь *MONUM_K* сравнительно небольшой, как и набор ключевых слов. Основная часть описания памятника выделяется из текста ЕЯ. Это слова, составляющие окрестность ключевых слов , например, ФИО, слова в родительном падеже и др. (см. правило MONUM~2).

В результате применения правил формируется СС-документа (содержательный портрет), где все слова приведены в нормальную форму, а объекты и связи представлены в виде фрагментов РСС. Например, для 1-го документа (из корпуса текстов) он будет иметь вид;

ДОК_(1,MONUM.TXT,"ПАМЯТНИКИ;")
FIO(КРУЗЕНШТЕРН,ИВАН,ФЕДОРОВИЧ," "/1+)
РАБ_(1-,РУКОВОДИТЕЛЬ,ПЕРВЫЙ,РУССКАЯ,КРУГОСВЕТНЫЙ,ЭКСПЕДИЦИЯ/2+)
MONUMENT_(ПАМЯТНИК,2-/3+)
НАХОДИТЬСЯ(3-/4+)

```
АДР_(НАБ.,ЛЕЙТЕНАНТ,ШМИДТ/5+)
Где(4-,5-)
РЕШЕНИЕ(О,3-/6+)
ПРИНЯТЬ(6-,УСТАНОВКА/7+)
ДАТА_(1869,ГОД/8+)
Когда(7-,8-)
ОРГ_(МОРСКОЙ,КАДЕТСКИЙ,КОРПУС/9+)
PLACE_(НАПРОТИВ,ЗДАНИЕ,9-/10+)
Где(4-,10-)
. . .
```

Первый фрагмент говорит, документ взят из файла MONUM.TXT и имеет номер 1. Последующие три фрагмента представляют *«Памятник руководителю первой русской кругосветной экспедиции Ивану Фёдоровичу Крузенштерну»*. Следующие три фрагмента — *"он находится на набережной Лейтенанта Шмидта»* и т.д. Коды 1+ и 1- (2+ и 2- и т.д.) обозначают один и тот же объект — лицо FIO (соответственно, профессию — РАБ_). Более подробное описание того, как устроена СС-документа, см. в [2,8,9]. СС-документов составляют базу знаний и служат для решения задач. В частности, СС-документов являются исходным материалом для автоматического порождения различных сведений — кто автор, архитектор, когда памятник установлен, открыт и т.д. Это делается с помощью экспертных программ на языке ДЕКЛ, который создан для обработки структур знаний на РСС [2].

5. Каталоги объектов

Как уже говорилось, отладка правил и ЛЗ ведется в рамках систем «Аналитик» («Криминал»). Они обеспечивают последовательную обработку множества документов (корпуса текстов из заданного файла) с формализацией каждого из них, формированием СС-документов и автоматическим построением общей БЗ и каталогов объектов. Такие каталоги — это списки выделенных объектов, где слова представлены в нормальной форме (необходимо для поиска). Например, каталог выделенных памятников (когда обработано 30 документов) имеет вид:

АЛЕКСАНДРОВСКАЯ КОЛОННА
АЛЛЕГОРИЧЕСКИЙ ЖЕНСКИЙ ФИГУРА МУДРОСТЬ
В ПАМЯТЬ ЖЕРТВА РЕПРЕССИЯ ПОЛИТИЧЕСКИЙ
В ПАМЯТЬ О ПРИБЫТИЕ ЛЕНИН
ВАНДОМСКАЯ КОЛОННА
СКУЛЬПТУРА ВЕРБЛЮД
ГРАНИТНЫЙ ПОСТАМЕНТ
ЕКАТЕРИНИНСКИЙ ДВОРЕЦ
КОННЫЙ ПАМЯТНИК АЛЕКСАНДР III
МИНОНОСЕЦ СТЕРЕГУЩИЙ
МИХАЙЛОВСКИЙ ДВОРЕЦ
МИХАЙЛОВСКИЙ ЗАМОК
МОНУМЕНТАЛЬНЫЙ ПАМЯТНИК ЛЕРМОНТОВ М. Ю.
ОБЕЛИСК ГОРОДУ ГЕРОЮ ЛЕНИНГРАДУ
ПАМЯТНИК-БЮСТ ЗНАМЕНИТЫЙ АРХИТЕКТОР XVIII В. РАСТРЕЛЛИ ФРАНЧЕСКО Б
ПАМЯТНИК-БЮСТ МУСОРГСКИЙ М. П.
ПАМЯТНИК-БЮСТ НЕКРАСОВ Н. А.
ПАМЯТНИК-БЮСТ ПРЖЕВАЛЬСКИЙ Н. М.
ПАМЯТНИК-БЮСТ СЕМЕНОВ-ТЯН-ШАНСКИЙ П. П.
ПАМЯТНИК АДМИРАЛ КРУЗЕНШТЕРН И. Ф.
ПАМЯТНИК АЛЕКСАНДР I
ПАМЯТНИК ГОРЬКИЙ А. М.
ПАМЯТНИК ЕКАТЕРИНА II
ПАМЯТНИК ИВАН ФЕДОРОВИЧ
ПАМЯТНИК КРЫЛОВ ИВАН АНДРЕЕВИЧ
ПАМЯТНИК ЛОМОНОСОВ М. В.
ПАМЯТНИК МЕДНЫЙ ВСАДНИК ПЕТР I
ПАМЯТНИК МИНОНОСЕЦ СТЕРЕГУЩИЙ
ПАМЯТНИК НАЗЫВАТЬСЯ ГЕРОЙ КРАСНОДОНА
ПАМЯТНИК НИКОЛАЙ I
ПАМЯТНИК РУКОВОДИТЕЛЬ ПЕРВАЯ РУССКАЯ КРУГОСВЕТНАЯ ЭКСПЕДИЦИЯ ..
. . .

Каталог выделенных мест:
АЛЕКСАНДРОВСКИЙ ПАРК ПРОСП. КАМЕННООСТРОВСКОЙ
БОЛЬШОЙ ПРОСП. САМПСОНИЕВСКИЙ
В ЦЕНТР ПЛ. ДВОРЦОВЫЙ
В ЦЕНТР ПЛ. ЗНАМЕНСКАЯ
В ЦЕНТР ПЛ. СЕНАТСКИЙ
*ВЕРЕБЬИНСКИЙ МОСТ ЖЕЛЕЗНЫЙ ДОРОГА ГОРОД САНКТ-ПЕТЕРБУРГ — МОСКВА**
ДВОР МИХАЙЛОВСКИЙ ДВОРЕЦ
ИОАННОВСКИЙ МОСТ
НАБ. ЛЕЙТЕНАНТ ШМИДТ
НАБ. РЕКА НЕВА НАПРОТИВ МОРСКОЙ КАДЕТСКИЙ КОРПУС
НАБ. РЕКА ФОНТАНКА
НАБ. РЕКА ФОНТАНКА СКВЕР ЛОМОНОСОВСКИЙ
НАПРОТИВ ЗДАНИЕ МОРСКОЙ КАДЕТСКИЙ КОРПУС
ПАРК ИМЕНИ 30 ЛЕТИЯ ВЛКСМ
ПЕРЕД ЗДАНИЕ УЧИЛИЩЕ ПРОСП. ЛЕРМОНТОВСКОЙ
ПЕТЕРГОФСКИЙ ПАРК АЛЕКСАНДРИЯ
ПЛ. ОСТРОВСКИЙ СКВЕР ПЕРЕД АЛЕКСАНДРИНСКИЙ ТЕАТР
ПРОСП. КАРЛ МАРКС
САД ПРУДОК НА УГОЛ УЛ. БАССЕЙНОЙ
САНКТ ПЕТЕРБУРГ
СКВЕР ЕКАТЕРИНА ПЕРЕД ТЕАТР
СКВЕР ПЕРЕД АЛЕКСАНДРИЙСКИЙ ТЕАТР
СКВЕР ПЛ. ТРОИЦКАЯ
УГОЛ КРОНВЕРКСКИЙ ПРОСП. КАМЕННООСТРОВСКИЙ НА НЕБОЛЬШОЙ ..
ЦЕНТР ГЛАВНЫЙ ПЛ. ГОРОД САНКТ ПЕТЕРБУРГ

Выявление шумов и потерь сводится к просмотру таких каталогов с поиском неполных или бессмысленных описаний. Выбрав любую строчку и нажав ENTER, будет найден документ, из которого выделен соответствующий объект. Этот документ подается на вход ЛП для повторного анализа.

Еще раз отметим, что такие каталоги строятся автоматически в процессе анализа корпусов текстов с выделением объектов. Сформированный каталог можно поместить в соответствующий предметный словарь, расширив его. Однако, это не приведет к существенному улучшению работы ЛП, так как имеющиеся в каталоге объекты и так устойчиво выделяются ЛП.

6. Просмотр процесса применения правил

Когда найден документ, из которого не правильно выделен объект, нужно найти соответствующее правило и скорректировать его. Для этого служит **режим просмотра** процесса анализа документа, где для каждого правила указывается, какие оно осуществило преобразования. Процесс визуализации изображен на рис.1.

На рис.1 показано следующее. Правило FF~5 применилось и выделило лицо *МОНИГЕТТИ И. А.* Следующее правило объединило фамилию *КРУЗЕНШТЕРН* с выделенным лицом — *КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ*. Слово РАБ_ указывает на профессию, а OBJ_ — на памятники. Правила GG~1 и GG~2 выделяют группы согласованных слов — генетивные цепочки. Правила MONUM~1, MONUM~4 выделяют группы слов — описания памятников. В режиме просмотра легко увидеть, что сделало каждое правило и где имела место ошибка. Более того, все правила имеют встроенные **механизмы трассировки**. Их можно активизировать для каждого интересующего правила. Трассировка визуализирует процесс, выводя на экран, в какой последовательности захватываются слова и почему правило оказалось не применимым [11,12].

7. Выдача результатов

Главным в работе ЛП является формирование СС-документов (см. п. 4), которые пользователь не должен видеть. Но на основе СС-документов с помощью достаточно простых ДЕКЛ-программ могут строиться различные формы или описания, необходимые для пользователя. Например, это может быть XML-файл, где для каждого объекта дается набор

.....
ФИО: *КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ* << FF~4
ФИО: *МОНИГЕТТИ И. А.* << FF~5
ФИО: *КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ & ФИО: КРУЗЕНШТЕРН* — одно лицо << FIO_UN
+++ Уровень +++ LEVEL_T3
+++ Уровень +++ LEVEL_T4
WORD_C: *КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << GG~1
WORD_C: *РУССКАЯ КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << GG~1
WORD_C: *ПЕРВЫЙ РУССКАЯ КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << GG~1
РАБ_: *РУКОВОДИТЕЛЬ ПЕРВЫЙ РУССКАЯ КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << GG~2
WORD_C: *ТОРЖЕСТВЕННЫЙ ЗАКЛАДКА* << GG~1
WORD_C: *КАНУН СТОЛЕТИЕ* << GG~2
WORD_C: *100-ЛЕТНИЙ ЮБИЛЕЙ* << GG~2
РАБ_: *ШРЕДЕР И. Н. МОДЕЛЬ* << WORK~1A
РАБ_: *МОНИГЕТТИ И. А. АРХИТЕКТОР* << WORK~1A
РАБ_: *КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ РУКОВОДИТЕЛЬ ПЕРВЫЙ РУССКАЯ КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << WORK~1A
OBJ_: *ПАМЯТНИК КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ РУКОВОДИТЕЛЬ ПЕРВЫЙ РУССКАЯ КРУГОСВЕТНЫЙ ЭКСПЕДИЦИЯ* << MONUM~1
OBJ_: *ФИГУРА ИЗ БРОНЗЫ* << MONUM~4
OBJ_: *ГРАНИТНЫЙ ПОСТАМЕНТ* << MONUM~4
+++ Уровень +++ LEVEL_T5
ОН — это *ПАМЯТНИК КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ* ... << ID_2MM
ПАМЯТНИК — это *ПАМЯТНИК КРУЗЕНШТЕРН ИВАН ФЕДОРОВИЧ* ... << ID_2W
.....

ОТЛИТЬ: ФИГУРА ИЗ БРОНЗЫ ПО ШРЕДЕР И.Н. МОДЕЛЬ << VV~1

Когда: ДАТА_: *ДЕКАБРЬ 1872* . . .

Рис. 1. Визуализация процесса анализа

составляющих его слов в нормальной форме (для поиска) и его описание, взятое из текста [11] . Описания выделенных объектов могут служить основой для заполнения БД или полей какого-либо сайта (формы), как это сделано в системе формализации резюме.

Ниже приведен еще один пример, когда для лиц даются из ролевые функции, а для времени и места указываются события, к которым они относятся.

```
<DOCUMENT DOC_NUM="0">
    <OBJECT ID="1» TYPE="FIO">
        <ARG TYPE=" -- Посвящен:"/>
        <SOURCE> Ивану Федоровичу Крузенштерну</SOURCE>
    </OBJECT>
    <OBJECT ID="2» TYPE="USER_OBJECT">
        <SOURCE> Памятник руководителю первой русской кругосветной экспедиции Ивану
                 Федоровичу Крузенштерну</SOURCE>
    </OBJECT>
    <OBJECT ID="3» TYPE="ADDRESS">
        <ARG TYPE=" -- ГДЕ ... находится:"/>
        <SOURCE> на набережной Лейтенанта Шмидта</SOURCE>
    </OBJECT>
    <OBJECT ID="4» TYPE="DATE">
        <ARG TYPE=" -- КОГДА ... решение об его установке было принято:"/>
        <SOURCE> в 1869 году</SOURCE>
    </OBJECT>
    <OBJECT ID="5» TYPE="PLACE">
        <ARG TYPE=" -- ГДЕ ... находится:"/>
        <SOURCE> Напротив здания Морского кадетского корпуса </SOURCE>
    </OBJECT>
    <OBJECT ID="6» TYPE="DATE">
        <ARG TYPE=" -- КОГДА ... закладка памятника состоялась:"/>
        <SOURCE> 8 ноября 1870</SOURCE>
    </OBJECT>
    <OBJECT ID="7» TYPE="USER_OBJECT">
        <SOURCE> Фигура из бронзы</SOURCE>
    </OBJECT>
    <OBJECT ID="8» TYPE="FIO">
        <ARG TYPE=" -- Создатель модели:"/>
    <SOURCE> И.Н. Шредера</SOURCE>
    ...
```

Формирование такого XML-файла осуществляется с помощью ДЕКЛ-программ, которые анализируют СС-документа и присваивают новые свойства объектам. Отметим, что отдельные компоненты описания взяты из текста и поэтому могут быть в различных падежных формах, например, *И. Н. Шредера*. В настоящее время разработана программа их корректировки, которая в данной работе не использована.

Заключение

В настоящее время лингвистический процессор Semantix настроен на автоматическую обработку потоков текстов на естественном языке (ЕЯ), представляющих собой: резюме на русском и английском языке, сообщения СМИ (о терактах), тексты описания достопримечательностей (памятников), сводки происшествий, справки по уголовным делам. Процессор может быть использован для обработки архивных и информационно-рекламных материалов, почтовых сообщений и т. д. Достоинства этого процессора — высокая избирательность при выделении объектов и связей, наличие правил глубинного анализа с выделением событий и их привязкой к времени и месту, а также наличие средств быстрой настройки на новую предметную область. Как показывает опыт, время такой настройки исчисляется не годами, а неделями, месяцами: зависит от количества и сложности новых объектов и допустимыми коэффициентами шумов-потерь.

ДЕМО-версия процессора Semantix — http://www.semantix4you.com

Литература

1. *Кузнецов И. П.* Семантические представления // М. Наука. 1986 г. 290 с.
2. *Кузнецов И. П., Мацкевич А. Г.* Семантико-ориентированные системы на основе Баз Знаний.// Монография, МТУСИ. М.: 2007. 173 с.
3. *Кузнецов И. П.* Методы обработки сводок с выделением особенностей фигурантов и происшествий // Труды международного семинара Диалог-1999 по компьютерной лингвистике и ее приложениям. Том 2. Тарусса 1999.
4. *Cunningham, H.* Automatic Information Extraction. // Encyclopedia of Language and Linguistics, 2cnd ed. Elsevier, 2005.
5. *Han J. and Kamber, M.* Data Mining: Concepts and Techniques // Morgan Kaufmann, 2006.
6. *Igor Kuznetsov, Elena Kozerenko.* The system for extracting semantic information from natural language texts // Proceeding of International Conference on Machine Learning. MLMTA-03, Las Vegas US, 23–26 June 2003, p. 75–80.
7. *Кузнецов И. П., Мацкевич А. Г.* Англоязычная версия системы автоматического выявления значимой информации из текстов естественного языка // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2005», Звенигород, 2005.
8. *Kuznetsov I. P., Kozerenko E. B.* Linguistic Processor «Semantix» for Knowledge extraction from natural texts in Russia and English. Proceeding of International Conference on Machine Learning, ISAT-2008. 14–18 July, 2008 Las Vegas, USA// CSREA Press, 2008, p.835–841.
9. *Кузнецов И. П., Сомин Н. В.* Средства настройки семантико-ориентированного лингвистического процессора на выделение и поиск объектов. Сб. ИПИ РАН, Вып.18. 2008 г., стр. 119–143 .
10. *Кузнецов И. П.* Объектно-ориентированная система, основанная на знаниях в виде XML-представлений.// Сб. ИПИ РАН, Вып.18.М.: 2008. С. 96–118.
11. *Кузнецов И. П., Ефимов Д. А.* Особенности извлечения знаний семантико-ориентированным лингвистическим процессором Semantix.// Сб. Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам конференции «Диалог 008»..РГГУ, М.:2008., С. 281–291.

Электронный семантический словарь глагольных прилагательных: структура и типы информации[1]

The semantic database of verbal adjectives: structure and types of information

Кустова Г. И. (galinak03@gmail.com)

Московский педагогический государственный университет

Доклад посвящен созданию электронного семантического словаря (базы данных) глагольных прилагательных (ср. *входной, лечебный, осветительный* и под.). Рассматриваются темы: а) связь прилагательного с глагольной ситуацией и возможности выражения глагольных актантов, ср.: *стиральная машина* (инструмент), *стиральный порошок* (средство); б) роль семантического класса и функционального предиката существительного в формировании семантической модели сочетания «глагольное прилагательное + существительное»; в) типы информации в базе данных; г) уточнение семантической разметки словаря Национального корпуса русского языка.

1. Материал, цели и задачи

Словарь глагольных прилагательных, лингвистическим параметрам которого посвящена данная работа, является фрагментом общего электронного словаря прилагательных, но здесь мы будем рассматривать его как самостоятельную базу данных.

Термин «глагольные прилагательные» будет пониматься широко: это не только прилагательные, образованные от глаголов, причастий и отглагольных существительных (*стиральный*, *испытательный*, *моющий*, *вязаный*, *дежурный*, *обзорный*, *сварочный*), но и «супплетивные» прилагательные с событийной семантикой, которые не образованы от глагола, но отсылают к некоторому событию, ситуации (ср. *обеденный зал* и *банкетный зал*). Рассматриваются только относительные прилагательные, т. е. из материала данной работы исключаются качественные прилагательные, генетически связанные с глаголами (ср. *нестерпимый, почтительный, оправданный, спорный* и под.)

Глагольные прилагательные имеют «двунаправленные связи» — с глаголом и с существительным. В силу происхождения и семантики глагольное прилагательное отсылает к той же ситуации, которую обозначает глагол; далее будем называть эту ситуацию «глагольной». С другой стороны, прилагательное, в прототипическом случае, является членом конструкции «A (прилагательное) + S (существительное)», которую далее будем называть «атрибутивное сочетание (конструкция) с глагольным прилагательным», или просто «атрибутивное сочетание (атрибутивная конструкция)». Таким образом, в ходе анализа атрибутивных конструкций необходимо проследить и (а) сохранение / редукцию глагольных свойств прилагательного и (б) его семантическое взаимодействие с существительным. Будут рассматриваться только конкретные существительные.

Атрибутивные конструкции с глагольным прилагательным будут рассматриваться

(1) с точки зрения создания электронного семантического словаря (базы данных) глагольных прилагательных: необходимо определить, какие типы информации должны быть представлены в таком словаре;

(2) с точки зрения совершенствования семантической разметки Национального корпуса русского языка (далее — НКРЯ / Корпус; о семантической разметке в Корпусе см. [Кустова и др. 2005]):

(а) глаголы в Корпусе имеют семантическую разметку (‘воздействие’, ‘движение’, ‘восприятие’ и т. п.). В ходе анализа атрибутивных сочетаний необходимо выяснить, должны ли глагольные прилагательные автоматически наследовать класс «производящих» глаголов или для них нужно разработать специальную разметку, отражающую их собственные языковые свойства и поведение именно как прилагательных;

[1] Работа выполнена при поддержке РГНФ, проект № 08-04-00183а. Примеры атрибутивных сочетаний взяты из Национального корпуса русского языка.

(б) конкретные существительные в Корпусе также имеют семантическую разметку, в частности, у них выделяются такие классы, как «лицо», «вещество», «инструменты», «мебель», «одежда», «оружие», «транспорт», «текст» и др., которые отражают их свойства (ср. «вещество») или тип использования (ср. «одежда»). В ходе анализа атрибутивных сочетаний необходимо выяснить, достаточно ли существующей разметки для описания моделей семантического взаимодействия существительных с глагольными прилагательными или она должна быть уточнена и усовершенствована.

2. Глагольные прилагательные и глаголы

Поскольку глагольное прилагательное коррелирует с глаголом, в его семантике «заложен», в принципе, тот же самый набор актантов с теми же самыми семантическими ограничениями на их заполнение. С другой стороны, это редуцированный, пониженный в ранге предикат, поэтому далеко не все глагольные свойства и отношения он наследует. В теоретическом плане — в плане изучения процессов редукции, смысловой компрессии, — безусловно, интересно проследить, что сохраняется от «глагольного состояния», а что меняется. Однако это важно и в практическом плане — во-первых, для семантического словаря: какие глагольные признаки приписывать прилагательным?; а во-вторых, для автоматического анализа текста: семантические модели сочетаний глагольных прилагательных и существительных в семантическом представлении предложения должны эксплицироваться точно так же, как и собственно глагольные структуры.

Отметим важнейшие отличия глагольного прилагательного от глагола, которые существенны для интерпретации атрибутивных конструкций.

(1) В контексте глагола референты связанных с ним существительных мыслятся как участники актуальной ситуации (*X стирает Y порошком в тазу*), тогда как отглагольное прилагательное выражает постоянное отношение предмета к глагольной ситуации, т.е. представляет глагольную ситуацию как постоянную характеристику некоторого предмета — например, как предназначение, ср. *стиральный порошок*.

(2) Глагол может синтаксически выражать при себе всех участников ситуации[2], причем одновременно; у прилагательного одна-единственная позиция для существительного (мы не берем двухместные прилагательные типа *X склонен к чему*; *X готов на что* и под.), т.е. оно может одновременно выразить лишь одного участника глагольной ситуации. Теоретически в эту позицию могут «по очереди» попадать все участники глагольной ситуации, практически же глагольное прилагательное сочетается лишь с некоторыми актантами, ср.: **стиральная женщина* (агенс), **стиральные вещи* (пациенс), *стиральная машина* (инструмент), *стиральный порошок* (средство), *стиральный таз* (место)[3].

(3) В глагольной конструкции семантическая роль существительного выражается падежом. В атрибутивном сочетании семантическую роль существительного в соответствующей глагольной ситуации нельзя выразить падежом, тем не менее, получая «на входе» такие сочетания, слушающий для каждого существительного устанавливает семантическую роль, — в противном случае он просто не понимает смысла данного атрибутивного сочетания. Следовательно, в атрибутивной конструкции на семантическом уровне выражается то, что глагол выражает синтаксически: роль данного предмета в «глагольной» ситуации. А поскольку при прилагательном могут выражаться разные актанты, то и ролевая семантика в атрибутивных сочетаниях будет разная. Это не разные значения прилагательного при разных существительных, но это тоже особого рода значение. Мы будем называть его семантическая модель атрибутивного сочетания «A + S».

Семантическая модель отражает структуру глагольной ситуации и показывает роль референта существительного атрибутивного сочетания как актанта глагольной ситуации:

стиральный порошок: кто стирает что ЧЕМ → порошком
стиральный таз: кто стирает что ГДЕ/В ЧЕМ → в тазу

Семантическую модель сочетания «A + S» слушающему позволяет реконструировать, с одной стороны, знание ситуации, обозначаемой глагольным прилагательным, а с другой — установление корреляции между глагольной ситуацией и семантическим классом определяемого существительного. Поэтому прилагательное налагает определенные «номинативные» требования на способ выражения участника глагольной ситуации.

---

[2] Мы исходим из того набора участников (ролей), который изложен и проанализирован в работе [Апресян 1974: 125–129]; например, в ситуации физического действия типичными участниками являются: Агенс, Пациенс, Инструмент, Средство, а также (иногда) Место.

[3] Здесь напрашивается аналогия с актантными существительными, которые тоже могут выражать отношение предмета или лица к некоторой ситуации, причем каждое существительное соответствует какой-то одной роли. Теоретически, поскольку в языке существует механизм актантной деривации (описанный в теории лингвистических моделей «Смысл⇔Текст», см. [Мельчук 1974], [Апресян 1974]), «глагольное» (актантное) обозначение (в том числе — супплетивное) могли бы получить все участники ситуации, однако реально в языке существуют не все актантные существительные: *лечить* — Субъект: *врач* (ср. *лекарь*), Объект: *пациент*, Средство: *лекарство*; *отравить* — Субъект: *отравитель*, Объект: — , Средство: *яд*.

Для глагола обычно не существенно, как назван участник. Например, в случае *Прачка стирает* участие лица в ситуации стирки в качестве агенса выражается падежом и дублируется семантикой существительного, а в случае *Клава стирает* такого дублирования нет, и оно не нужно, т.к. для выражения роли достаточно падежа (поэтому для обозначения участника можно использовать даже местоимение: *кто стирает?*). Для прилагательного способ выражения актанта небезразличен, т.к. адресату необходимо реконструировать семантическую модель атрибутивного сочетания с опорой на семантический класс существительного, ср.: *купальный бассейн* ['вместилище, емкость'] = 'где купаются'; *купальная шапочка* ['одежда: головной убор'] = 'в чем купаются'.

(4) Четвертое отличие связано с изменением / сохранением значения глагола и прилагательного. У глагола обычно бывает несколько значений, и изменение значения в типичном случае связано с изменением модели управления (падежной рамки) и / или изменением семантического класса актанта (актантов), ср. *болеть чем — болеть за кого*; *потерять кошелек* (конкр. сущ.) — *потерять терпение* (абстр. сущ.). Как показано в [Толдова, Кустова, Ляшевская 2008], на этих двух параметрах (падежи и семантические классы актантов) можно построить так называемые семантические фильтры, используемые программой автоматического разрешения неоднозначности глаголов в текстах НКРЯ.

У прилагательного в его единственную позицию попадают существительные разных семантических классов, но, в отличие от глагола, это не приводит к изменению значения прилагательного (меняется только, как уже было сказано, интерпретация всего сочетания «A + S»): глагольное прилагательное, в типичном случае, имеет одно значение, — соответствующее первому (основному) значению глагола, — и сочетается с существительными, которые относятся к этому значению глагола (входят в его «падежную рамку»): *строить* ('возводить сооружения') — *строительный рабочий*, *строительная площадка*, другие значения — *строить планы*, *строить глазки* — не имеют прилагательного (**строительная девушка*); *освещать* ('снабжать светом') — *осветительный прибор*, ср. *освещать события* — **осветительная статья*.

Если семантику прилагательного считать заданной и фиксированной, то решающую роль в семантическом анализе атрибутивных конструкций будет играть семантика существительного.

3. Глагольные прилагательные и существительные

Для интерпретации сочетаний «A + S» важен не только семантический класс существительного ('лицо', 'инструмент', 'одежда' и т.д.). Многие существительные характеризуют объект через отношение к некоторой ситуации и содержат соответствующий семантический предикат, который тоже участвует в интерпретации атрибутивных конструкций. Среди конкретных существительных это прежде всего имена деятеля, ср. *артист* — 'играть', *офицер* — 'служить, воевать', и названия артефактов. Обозначения артефактов принято называть функциональными именами (см. [Арутюнова 1980]; [Рахилина 2000]), т.к. функция, для выполнения которой они создаются, встроена в семантику соответствующих существительных, ср.: *игла* — 'чтобы шить, прокалывать', *молоток* — 'чтобы забивать, ударять'; *нож* — 'чтобы резать', *мыло* — 'чтобы мыть, стирать', *лекарство* — 'чтобы лечиться', *матрас* — 'чтобы лежать' и т.д. Семантический предикат, отражающий функцию предмета (или деятельность лица), будем называть функциональным предикатом.

В атрибутивном сочетании может возникать два основных типа отношений между семантическими предикатами прилагательного и существительного: согласованность или несогласованность.

Существительное может иметь семантический предикат, «родственный» глагольной ситуации, выражаемой прилагательным, ср. *стрелковое оружие* ('стрелять' — 'поражать, в том числе выстрелами'), или являться нейтральным обозначением, не препятствующим «актантной интерпретации» атрибутивного сочетания, ср. *строительный рабочий* (агенс), *стиральная машина* (инструмент), *моющий порошок* (средство). Такие атрибутивные сочетания мы будем называть семантически согласованным, а прилагательные — функциональными.

Однако прилагательное может быть семантически не согласованным с существительным и обозначать в большей или меньшей степени случайные, второстепенные, дополнительные признаки, не связанные с функциональным предикатом существительного: *бродячие артисты* «бродят», перемещаются, что не связано с функцией 'играть'; *складной нож* можно сложить, но есть и другие ножи, которые не складываются, к их функции ('резать') это не имеет отношения; *надувной матрас* — случайный признак, не связанный с предикатом 'лежать' и указываемый лишь потому, что обычно матрасы не надувают, а набивают. В принципе, такие прилагательные, как и большинство относительных прилагательных, выполняют «классифицирующую» функцию (см. [Павлов 1960]) — выделяют подкласс из общего класса: подкласс *бродячих артистов* из всех *артистов*, подкласс *надувных матрасов* из всех *матрасов*; и обозначаемая ими «побочная» характеристика подходит для этого не хуже других. Но поскольку это все-таки побочная характеристика, не связанная с семантическим предикатом су-

ществительного, мы будем называть их нефункциональными[4].

Семантически не согласованные прилагательные могут быть субъектными — если референт существительного является субъектом ситуации, обозначаемой глагольным прилагательным: *проезжий генерал* (он проезжает), *дежурный офицер* (он дежурит), а могут быть объектными — если референт существительного является объектом, ср. *складной нож* (его складывают)[5]. Для семантически несогласованных прилагательных семантический класс существительного, как и его функциональный предикат, не важен: они сочетаются не только с функциональными именами, но и с именами натуральных классов, ср. *лежачий камень*, *привозные огурцы*.

Функциональное и нефункциональное прилагательные (если они оба представлены у одного и того же глагола), как правило, различаются по морфологической структуре и словообразовательной модели, ср.: *доильный аппарат — дойная корова*; *вязальная машина — вязаная шапка*. Однако есть и случаи «омонимии»: *подъемный кран* (функция: поднимать) vs. *подъемный мост* (его поднимают). Вообще, морфологические показатели сами по себе не могут служить надежным критерием для отнесения прилагательного к классу функциональных, ср., например, прилагательные с одним и тем же суффиксом *-тельн-*: *наблюдательная труба* (функциональное: через нее что-то наблюдают), *наблюдательный человек* (качественное), *внушительное превосходство* (‘большое’; вообще утрачена связь с глагольной семантикой). Поэтому соответствующая характеристика должна приписываться прилагательному уже в словаре.

Далее будут рассматриваться только функциональные глагольные прилагательные.

Атрибутивные сочетания глагольных прилагательных с существительными разных семантических классов

ВЕЩЕСТВА. С точки зрения сочетаемости с глагольными прилагательными среди веществ выделяются две группы: вещества-артефакты (*порошок*, *средство*, *раствор*, *капли* и под.) и «натуральные» (природные) вещества.

Большинство сочетаний глагольных прилагательных с существительными первой группы, которые встретились в Корпусе, являются аналитическим обозначением Средства той ситуации, которая названа прилагательным (участник «Средство» бывает только в агентивных ситуациях, см. [Кустова 2004: 63–78]): *взрывной порошок* (=порох), *дезинфицирующий порошок*, *моющий порошок*, *полировальный порошок*, *стиральный порошок*; *лечебное средство*, *моющее средство*, *чистящая присадка*, *горючесмазочные материалы*. Обычно объект, на который воздействует Средство, не выражается, но иногда прилагательное, в силу особой словообразовательной модели, включает и этот объект, ср. *жаропонижающее средство*.

Если существительное не содержит в семантике идеи «артефактности», то сочетание имеет другую семантическую формулу: оно приписывает веществу свойство оказывать то воздействие, которое названо глагольным прилагательным: *красящие вещества*, *отравляющие вещества*.

ИНСТРУМЕНТЫ,ПРИСПОСОБЛЕНИЯ,УСТРОЙСТВА, МЕХАНИЗМЫ. Ситуация с инструментами похожа на ситуацию со средствами. Многие атрибутивные сочетания являются своего рода аналитическими наименованиями инструментов и устройств: *игровой / разменный / раскройный / торговый автомат*; *посадочный автомат* (ср. пример из Корпуса: *Осуществляется деятельность по созданию посадочных автоматов к сажалкам СБН-1 А и ЛМД-1* [«Лесное хозяйство», 2003]); *разведывательная аппаратура*; *копировальный / сварочный аппарат*; *морозильная / холодильная камера*; *вычислительная / вязальная / стиральная / швейная машина*; *нагревательный / осветительный прибор*; *метательный снаряд*; *взрывное / запорное / подслушивающее / подъемное устройство*; *пусковой механизм*; *ткацкий / точильный* и т.д. *станок*. Есть обозначения с включенным объектом: *маслобойный, металлорежущий, посудомоечный, снегоочистительный* (таким образом, если бы *подъемный кран* назывался *грузоподъемным*, не было бы омонимии с *подъемным мостом*).

Глагольное прилагательное используется и тогда, когда нужно обозначить нетипичную функцию, не совпадающую с функциональным предикатом существительного, или специальную функцию: *сигнальная лампочка* (лампочка должна светить, это ее стандартная функция, а у сигнальной лампочки свет — лишь средство подачи сигнала), *увеличительное стекло* (сквозь любое стекло можно смотреть, но через увеличительное можно рассмотреть мелкие объекты).

Место-приспособление (обычно емкость): *заварочный чайник*, *стиральный таз / бак*.

ОРУЖИЕ: *стрелковое оружие*.

ТРАНСПОРТ: *летательный аппарат*; *прогулочный катер*; *разведывательный самолет*.

ТЕКСТ и РЕЧЬ: *сопроводительное письмо*; *объяснительная записка*; *описательная диссертация*; *обзорная статья*; *поминальная молитва*.

---

[4] Сочетания типа *лечащий врач* или *лежачий больной* мы не рассматриваем, т.к. они идиоматичны. Например, *лежачий больной* не просто лежит — время от времени лежат все больные, а не только лежачие, для больного это «нормальное» состояние (ср. выражение *лежать в больнице*), — лежачий же больной вообще не может встать, что вовсе не вытекает из семантики компонентов атрибутивного сочетания.

[5] Субъектные и объектные прилагательные до некоторой степени являются аналогами активных и пассивных причастий глаголов, хотя полного соответствия здесь нет.

ДОКУМЕНТ: *входной / пригласительный / проездной билет*; *посадочный талон*; *избирательный бюллетень*.

ОДЕЖДА и ОБУВЬ: *беговая обувь*; *спальный чепец*; *рабочая одежда*; *купальный костюм / халат*, *купальная шапочка*.

МЕБЕЛЬ: *обеденный / письменный стол*; *спальная тахта*.

МЕСТА-АРТЕФАКТЫ, ПОМЕЩЕНИЯ, СООРУЖЕНИЯ: *беговая дорожка*; *взлетная полоса*; *зрительный / читальный / танцевальный зал*; *купальный / плавательный бассейн*; *сборный / пропускной / наблюдательный пункт*; *испытательный полигон*.

Как показывает приведенный материал, функциональные прилагательные практически не сочетаются с главными участниками ситуации — агенсом и пациенсом[6], а сочетаются со «вспомогательными» участниками — инструментом, средством, местом-приспособлением, а также с сирконстантами — то же место, но уже в роли обстоятельства. Причем в качестве обозначений этих участников выступают названия артефактов.

Хотя названия артефактов уже содержат в себе функциональный предикат, при определенных условиях появляется необходимость использовать еще и функциональное прилагательное.

Здесь можно выделить следующие случаи.

(1) В контексте параметрического существительного — основная функция.

Существительное может быть семантически вырожденным, «родовым», с общей «классифицирующей» семантикой (‘средство / приспособление для какой-либо деятельности’) и не содержать никакого конкретного предиката, ср. *порошок, состав, смесь, материал, средство, машина, прибор, устройство, механизм*. Тогда глагольное прилагательное обозначает основную, главную функцию артефакта, для которой он создавался и которую он должен выполнять: *стиральный порошок*; *лечебное средство*; *смазочные материалы*. Поскольку в «потребительском» обществе происходит процесс взрывного увеличения разного рода средств и приспособлений для осуществления разных типов деятельности, а однословных обозначений для них заведомо не хватает или они неизвестны говорящему в силу отсутствия специального образования, то аналитические обозначения артефактов оказываются востребованными и получают широкое распространение в профессиональных языках, в технической документации, в языке прессы.

(2) В контексте стандартного обозначения артефакта.

(а) Функциональный подтип.

Если артефакт обозначен «нормальным» функциональным существительным, и тем самым его основная функция уже выражена, тогда глагольное прилагательное может использоваться для выделения функционального подкласса: *стиральное мыло* vs. *банное*, *туалетное*; *стрелковое оружие* (каким способом поражает) (ср. *холодное*, *ядерное*).

(б) Связанная функция.

До сих пор речь шла о случаях, когда референт существительного является непосредственным участником той ситуации, которую обозначает глагольное прилагательное, например, *стиральный порошок* — средство в ситуации «стирка», *стрелковое оружие* — инструмент в ситуации «стрельба». Однако есть и другие случаи, когда референт существительного непосредственно не участвует в глагольной ситуации, но способствует ее реализации. Речь идет об объектах типа *обеденный стол* или *купальная шапочка*.

Для того чтобы пообедать, нужен не стол, а обед, а для того чтобы искупаться, нужна не шапочка, а вода. Стол не является участником ситуации «обедать», поскольку человек не обедает «столом». Человек просто ‘**сидит** за столом’, хотя и это не является семантическим предикатом стола как предмета мебели. Для стола существенна поверхность, на которой можно что-то разместить. Но поскольку в нашей культуре человек, как правило, обедает (и вообще принимает пищу) сидя за столом, то у стола появляется постоянное отношение к ситуации приема пищи, которое и выражается прилагательным, более того, имеется специальный вид столов, которые связаны именно с данной ситуацией (ср. *письменный стол*). Аналогично с одеждой: *купальную шапочку* надевают не «чтобы купаться», а «когда купаются», т.е. купаются не «с помощью» шапочки, а просто в шапочке.

Таким образом, для существительных классов мебель, одежда, посуда, помимо основной, прямой функции, которая встроена в их семантику, характерна еще «связанная» («метонимическая») функция, т. е. в сочетаниях типа *купальная шапочка* предмет (референт существительного) не является непосредственным участником глагольной ситуации, но связан с ней «регулярными», системными отношениями, способствует ее реализации и специально предназначен для использования в этой ситуации. Можно сказать и по-другому: такие объекты, как обеденный стол или купальная шапочка, участвуют в ситуациях обеда и купания, но не в качестве основных, а в качестве «вспомогательных», «сопутствующих» участников.

---

[6] Пациенс вообще имеет мало глагольных обозначений, как аналитических, так и синтетических (ср., однако, существительные типа *еда* (то, что едят), *ноша* (то, что несут)). Агенс в русском языке «обслуживается», в основном, существительными (ср. *учитель*, *пловец, стрелок*).
Впрочем, агентивные сочетания с глагольными прилагательными встречаются, хотя они крайне немногочисленны. Их можно назвать функционально-классифицирующими: они нужны не столько для того, чтобы обозначить функцию (она и так уже содержится в семантике существительного), сколько для того, чтобы выделить из общего класса некоторый подкласс, например: *боевой генерал* (ср. *паркетный генерал*).

(3) В контексте пространственных существительных.

Пространственные объекты, созданные человеком (дороги, аэродромы и под.), а также помещения и сооружения являются функциональными (как и соответствующие существительные): они создаются для определенного типа использования, и этот тип использования можно считать их основной функцией — такой же, как стирка для стирального порошка. Сооружения и помещения, как и другие артефакты, могут иметь однословное название, содержащее указание на их предназначение (ср. *баня*, *ресторан*). Но если существительное не содержит конкретного семантического предиката (*зал, пункт, помещение*), то конкретное предназначение выражается глагольным прилагательным: *читальный зал*; *пропускной пункт*; *караульное помещение*. А если основная функция включена в семантику существительного, то прилагательное может ее конкретизировать: *беговая дорожка*; *сидячие места* (в вагоне).

Для полноты картины надо добавить, что глагольные прилагательные сочетаются не только с предметными существительными, но и с событийными (обозначающими ситуацию или временной период). Здесь уже требуется установление отношений не между предметом и ситуацией, а между двумя ситуациями, и у таких сочетаний, естественно, будут свои собственные семантические модели, например: V — цель субъекта, осуществляющего P (*карательная экспедиция, ознакомительная поездка*); V — содержание ситуации P, обозначенной родовым существительным (*сварочные работы, розыскные мероприятия*); V — ситуация, которая осуществляется в период P (*испытательный срок, отопительный сезон*) и т.п.

В дальнейшем будут соединены базы данных глагольных и неглагольных прилагательных, и в связи с этим встанут вопросы о соотношении «глагольных» и «неглагольных» сочетаний, имеющих сходные семантические модели: *проездной билет* дает право на проезд, а *театральный* — на посещение спектакля (что должно извлекаться из семантических представлений слов *театр* и *билет*); сочетания *телефонный*, *музыкальный* и *кофейный автомат*, хотя и не содержат глагольных прилагательных, по семантической модели не отличаются от *игрового* и *торгового автоматов*.

4. Выводы

Практические выводы по итогам анализа сочетаний глагольных прилагательных и существительных таковы.

(а) В базе данных (а также в семантическом словаре Корпуса) глагольные прилагательные кроме пометы «глагольное» должны иметь помету «нефункциональное» (в том числе — «субъектное», ср. *ходячий*, и «объектное», ср. *привозной*, *вязаный*) или «функциональное» (*швейный*, *гладильный*, *нагревательный*).

(б) Семантическая разметка функциональных глагольных прилагательных не должна быть механическим отражением семантической разметки соответствующих им глаголов. Например, у сочетаний *стиральная машина* и *наблюдательная труба* семантические модели будут одинаковые: ‘S — инструмент ситуации V’, значит, семантические различия глагольных признаков (классы производящих глаголов — ‘физическое воздействие’ и ‘восприятие’) для этих прилагательных не существенны. С точки зрения семантической модели эти (и мн. др., ср. *вязальный*, *чистящий*, *подслушивающий*, *осветительный* и под.) прилагательные попадут в одну группу: активное воздействие на объект для изменения его состояния / признака, создание / разрушение, приложение усилий для достижения цели / получения результата. В другую большую (и весьма разнородную с точки зрения глагольной классификации) группу попадут прилагательные от глаголов положения (*висячий*), движения (*беговой*), времяпрепровождения (*прогулочный*) и др.

(в) Важную роль при анализе атрибутивных сочетаний с глагольными прилагательными (в том числе, автоматическом) играет характер участия в глагольной ситуации референта существительного. Поэтому в базу данных глагольных прилагательных необходимо поместить информацию об актантной структуре (падежной рамке) производящего глагола и семантических ограничениях на заполнение его валентностей, т.к. она используется, чтобы реконструировать семантические модели атрибутивных сочетаний.

(г) Для функциональных прилагательных в базе данных должны быть специальные поля, чтобы указать семантические классы определяемых существительных и семантические модели атрибутивных сочетаний, связанные с этими семантическими классами. При этом должны быть предусмотрены не только основные, но и метонимические контексты. Например, для прилагательного *осветительный* это будут не только названия приборов, приспособлений и систем, выполняющих функцию «освещать» (*лампа, прибор, оборудование, аппаратура, техника, осветительные средства, устройство, сеть, система*), но и контексты типа *осветительная опора*, *мачта* (то, на чем закреплен осветительный прибор);

(д) Для целей семантического анализа атрибутивных сочетаний необходимо уточнить семантическую разметку существительных в словаре Корпуса: с точки зрения сочетания с глагольными прилагательными в классе артефактов можно выделить два подкласса — «активные», которые воздействуют (пусть и с помощью человека) или «действуют» (подобно человеку): инструменты, механизмы и приборы, оружие (*вязальная машина* вяжет, *чистящее средство* чистит, *подслушивающее устройство* подслушивает), и «пассивные», которые используются: мебель, посуда, одежда, места и помещения.

Литература

1. *Апресян Ю. Д.* Лексическая семантика. Синонимические средства языка. М.: 1974.
2. *Арутюнова Н. Д.* К проблеме функциональных типов лексического значения // Аспекты семантических исследований. М.: 1980. С. 156–249.
3. *Кустова Г. И.* Типы производных значений и механизмы языкового расширения. М.: 2004.
4. *Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В.* Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: 2005. С. 155–174.
5. *Мельчук И. А.* Опыт теории лингвистических моделей «Смысл ⇔ Текст». М., 1974.
6. *Толдова С. Ю., Кустова Г. И., Ляшевская О. Н.* Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии. Диалог'2008. М.: 2008. С. 522–529.
7. *Павлов В. М.* О разрядах имен прилагательных в русском языке // Вопр. языкознания, 1960, № 1.
8. *Рахилина Е.В.* Когнитивный анализ предметных имен: Семантика и сочетаемость. М.: 2000.

Подход к созданию многоязычных параллельных корпусов веб-публикаций

The approach to creation of multilingual parallel corpora of web publications

Ландэ Д. В. (dwl@visti.net), **Жигало В. В.** (vladlen@visti.net)
Информационный центр «ЭЛВИСТИ», Киев, Украина

Описан метод построения двуязычного параллельного корпуса веб-публикаций, базирующийся на использовании частотных морфологических словарей, а также эмпирико-статистических алгоритмов. Предложен подход к преодолению омонимии в родственных флективных языках, позволяющий отбирать наиболее частотные нормальные формы. Алгоритм реализован в качестве программного комплекса и интегрирован в систему контент-мониторинга InfoStream. На основе предложенного метода был создан двуязычный русско-украинский параллельный корпус текстов веб-публикаций объемом свыше 450 000 пар документов.

Большое место в документальных информационно-поисковых системах занимают алгоритмы выделения ключевых слов, с помощью которых выполняются многие процедуры, охватываемые концепцией Text Mining, например, поиск подобных документов, выявление дубликатов, построение сниппетов, информационных портретов, дайджестов и т. д.

Заметим что проблема поиска подобных документов — одна из важнейших проблем современного информационного поиска, так как важные сообщения многократно дублируются. В данной статье описан метод, с помощью которого реализуется выявление информационных дубликатов, представленных на разных языках (русском и украинском). В результате применения этого метода авторами построен параллельный по информационному содержанию документальный корпус, который можно назвать «квазипараллельным», однако, он может также считаться параллельным в понимании [8], так как оснащен некоторыми автоматически сформированными тегами и переводами выделенных лексем на 2 языка. Выравнивание данного корпуса по предложениям или словам, а также морфологическая разметка корпуса отнесена к перспективам выполненной работы и выходит за рамки данной публикации.

На сегодняшний день существуют алгоритмы создания параллельных корпусов документов, которые можно условно разделить на две группы: традиционные и статистические.

К первой группе можно отнести алгоритмы, с помощью которых создавались такие параллельные корпусы, как Корпус CRATER [1]; Параллельный корпус переводов «Слова о полку Игореве» [2]; параллельный русско-английский корпус входящий в состав Национального корпуса русского языка [3]; параллельный русско-словацкий корпус[4] и т. д. Создание данных корпусов связано с тем, что исходные данные заведомо параллельные.

Ко второй группе можно отнести параллельные корпусы, созданные с помощью статистических алгоритмов, такие как [5–8], основанные на анализе страниц многоязычных веб-сайтов, объединении заранее подготовленных фрагментарных массивов и т. д.

Авторами предлагается новый подход к созданию параллельных корпусов документов, основанный на алгоритме поиска дубликатов документов на разных языках. Подход дает возможность отыскать похожие документы на разных языках в большом массиве документов. В результате можно убедится в том что в корпус попали параллельные документы из разных источников. Методы, основанные на анализе сайтов со страницами на разных языках, не позволяют определить дубликаты на разных источниках (сайтах), не указав специально параллельность этих источников. Традиционные же методы построения параллельных корпусов используют заведомо параллельные данные, что делает их в данном случае непригодными для использования.

Предложенный подход позволил создать двуязычный украинско-русский параллельный корпус текстов из веб-публикаций на русском и украинском языках объемом свыше 450 000 пар документов. Оцененная экспертами точность предложенного алгоритма составляет 98%.

Одной из основных проблем при автоматическом анализе текста является омонимия. Существующие подходы разрешения омонимии можно разделить на два основные типа: детерминированные и вероятностные. К детерминированным можно отнести методы, применяемые, например, в системе «ЭТАП» [9], где используется «фильтровый метод» синтаксического анализа, система «Диалинг» [10], или морфологический анализатор английского языка ENGTWOL [11], которые основаны на правилах снятия неоднозначности на основе контекстных правил. Вероятностный подход к преодолению омонимии широко обсуждался в работах российских исследователей [12–14], в применялся еще в 80-х годах XX века в системе М. Харста [15] для снятия неоднозначности у существительных путем использования размеченных вручную текстовых корпусов и выбора лексических и грамматических ключей.

Предложенный авторами подход к вычислению опорных слов документов (именно так будем обозначать ключевые слова, имея в виду, возможно, более узкую сферу применения) основаны на векторном представлении текста и используют статистические свойства текстов.

В данной работе описываются процедуры создания частотного словаря на основе морфологического словаря (МС) с использование тестового корпуса документов, построения алгоритма вычисления опорных слов с использованием частотного МС и модификации общеизвестного подхода TF IDF [16–13], а также статистического подхода к преодолению омонимии. На основе созданного алгоритма был построен программный комплекс, который интегрирован в систему контент-мониторинга InfoStream [17].

Реализован алгоритм построения параллельного корпуса документов, который учитывает не только статистические свойства текстов, но и некоторые морфологические признаки. В соответствии с этим алгоритмом построение параллельного корпуса происходит в несколько основных этапов:

- создание морфологических словарей;
- создание частотных морфологических словарей;
- создание словарей переводов;
- создание процедуры определения опорных слов в документах;
- определение разноязычных дубликатов.

Для русского и украинского языков были использованы свободно доступные электронные словари: ispell с набором более 1 млн. словоформ и «Словники України» [18], с набором более 4 млн. словоформ, а также словарь Зализняка, который насчитывает порядка 100 тыс. слов.

Эксперты дополнили морфологические словари неологизмами, названиями известных фирм, брендов и известными фамилиями, которых не было в исходных словарях.

Для обучения частотных морфологических словарей были взяты электронные публикации новостей, полученные из Интернет с помощью системы контент-мониторинга InfoStream. Количество публикаций составило 3 млн. 700 тыс. документов, 1 млн. 300 тыс. на украинском языке и около 2 млн. 400 тыс. на русском языке, за период с 01.01.2007 по 31.12.2007.

В соответствии с предложенным методом, «обучение» словарей проводится в несколько этапов. Первый этап заключается в разделении документов на словоформы и сохранении полученных словоформ с информацией о номерах соответствующих документов.

На втором этапе, созданный файл словоформ сортируется, после чего подсчитывается количество вхождений каждой словоформы, и количество документов в которых она встретилась. Найденные частоты записываются в частотный словарь, на основании которого определяется вероятная нормальная форма каждого слова.

Для выявления омонимии, в выходной файл записываются все нормальные формы соответствующие словоформе, т. е. если одной словоформе соответствует сразу несколько нормальных форм, сохраняются подсчитанные частоты со всеми найденными нормальными формами. На третьем этапе происходит заключительный подсчет количества нормальных форм и сохранение результатов в частотный словарь.

Для решения задачи построения параллельных текстовых корпусов в результирующие словари отбираются все словоформы имен существительных.

Описанный подход предусматривает использование алгоритма разрешения контекстной неоднозначности, так как омонимия является существенной проблемой при определении опорных слов документа, например, слово «села», которое в практике русского языка может быть множественным числом от слова «село», а также производной от глагола «садиться», может некорректно переводиться и использоваться на украинском языке, так как слово «село» переводится на украинский язык как «село», а слово «садиться» — «сідати». Неправильный выбор нормальной формы может привести к тому, что в одинаковых по информационному содержанию документах на разных языках будут использованы различные опорные слова. Для решения этой проблемы использовался, как оказалось позднее, эффективный и достаточно быстрый алгоритм, что особенно важно, так как этап обучения частотных словарей и этап их использования связаны с обработкой больших объемов текстовой информации.

В Табл. 1 показан пример обучения частотного словаря для слов «садиться» и «село». Предложено правило, в соответствии с которым, если в систему поступила словоформа, которая на практике может приводить к нескольким нормальным формам (например, для словоформы «села» допустимы нормальные формы «село» и «садиться»), то так называемые «индексы нормальных форм» для этой словоформы увеличиваются на единицу. В табл. 1 показан пример, когда в текстовом корпусе словоформа «села» встретилось 20 раз, словоформа «село» — 50 раз, словоформа «сели» — 10 раз, а словоформа «селом» — 30 раз. В результате обучения, в словари попадают слова «село» с индексом нормальной формы 100 и «садиться» с индексом 80, соответственно, в дальнейшем при отборе опорных слов предпочтение будет отдано слову «село».

В рамках данного исследования использовались словари переводов с русского на украинский и с украинского на русский язык. Данные словари были получены путем перевода наиболее частотных нормальных форм имен существительных с помощью бесплатных онлайн-словарей переводов в Интернет [19–21]. В случае, если одной словоформе соответствовало несколько переводов, то выбиралось наиболее употребляемые словоформы языка перевода в соответствии с частотным словарем. Полученный таким образом русско-украинский словарь насчитывал 80 тыс. наиболее частотных нормальных форм имен существительных, украино–русский — 90 тыс. наиболее частотных нормальных форм имен существительных.

Табл. 1. Пример обучения системы

Словоформа	Количество	Индекс нормальных форм
села	20	садиться → +20 село →+20
село	50	садиться → +50 село → +50
сели	10	садиться → +10
селом	30	село → +30
		село = 100 садиться = 80

Одним из эффективных подходов к выделению опорных слов из текста является векторная модель, в рамках которой, каждому слову документа присваивается его весовой коэффициент. Чем больше коэффициент слова, тем больше это слово характеризует документ. Для выявления опорных слов в тексте была использована модификация метода TF IDF — формула Okapi BM25 [22], которая в отличии от общепринятого подхода TF IDF позволяет учитывать среднюю длину документа в корпусе.

При использовании морфологических словарей предусмотрено, что отсеиваются все нормальные формы, соответствующие словам, находящихся в «стоп-словарях».

Для создания параллельного корпуса были взяты электронные публикации из Интернет, полученные с помощью системы InfoStream, за период с 01.01.96 по 28.02.2009, с общим количеством документов 60 млн., по всем политематическим источникам.

При реализации алгоритма происходит считывание текстового документа из входного потока, после чего выполняется выделение словоформ и поиск нормальной формы для каждой из них. В случае омонимии, выбирается наиболее частотная (с наибольшим индексом) по словарю нормальная форма словоформы. После вычисления соответствующих весовых коэффициентов с помощью формулы Okapi BM25 происходит ранжирование нормальных слов и выбирается двенадцать наиболее «весомых». Полученные двенадцать опорных слов переводятся на другой язык с помощью словарей переводов. Все опорные слова и слова-переводы приписываются к документу и выдаются в выходной поток.

Уже несколько лет в системе InfoStream используется механизм поиска дубликатов, который позволяет с помощью опорных слов находить подобные документы, представленные на одном языке. В этом механизме 6 опорных (наиболее весомых) слов исследуемого документа, сравниваются с 12-ю опорными словами каждого из документов корпуса вебпубликаций (рис. 1).

Именно таким же путем проводился поиск разноязычных дубликатов. Кроме того, данная процедура была дополнена рядом эвристических критериев, например:

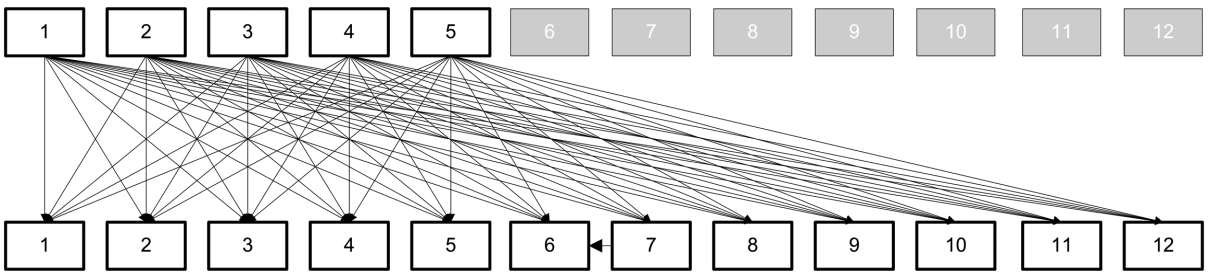


Рис. 1. Сравнение опорных слов

- общее количество слов в переведенном варианте не должно отличаться от оригинала более чем на 10%;
- количество чисел в документах не должно отличатся больше чем на два.

Анализа полученных результатов проводился путем изучения экспертами случайных выборок документов, определенных как разноязычные дубликаты. «параллельных» документов. Проведенный таким образом анализ показал, что в среднем 98% содержания каждого документа имеют разные дополнения: например, ссылки на другое издательство, или же название издательства издавшего документ.

На базе системы InfoStream был разработан программный комплекс для работы с параллельным корпусом в поисковом режиме [23]. Данный программный комплекс позволяет производить поиск по корпусу документов как на русском так и на украинском языках, а также поддерживает одновременный вывод параллельных текстов, релевантных запросам пользователей. На рис. 3 приведен интерфейс, на котором представлены результаты поиска по «экономический кризис» (в результате было выбрано 157 параллельных текстов, релевантных данному запросу).

Для такого большого полученного корпуса возникает проблема ручной проверки, в таком случает было решено использовать метод случайной выборки документов, по которым эксперты смогли определить точность соответствия документов в 98%.

Был произведен детальный анализ корпуса параллельных документов и получены такие результаты:

Общее количество слов в корпусе составляет более 192,7 млн., из которых 96 млн. в украинских документах, 96.7 млн. — в русских документах.

Средняя длина документа в корпусе составляет 195 слов для украинского и 196 слов для русского.

Количество источников документов на украинском языке содержащихся в корпусе — 997. Количество источников документов на русском языке — 1768. Наиболее частотные источники приведены в Табл. 2.

На рис. 2 представлен пример вывода заголовков и аннотаций параллельных документов, содержащихся в корпусе, найденных по ключевым словам «экономический кризис». Полный текст пары параллельных документов приведен на рис. 3.

Указанный параллельный корпус расположен по адресу http://ling.infostream.ua и свободно доступен через поисковую систему. Корпус постоянно расширяется (по мере мониторинга новостей из Интернет) и в данный момент уже содержит более 450 тыс. пар документов на русском и украинском языках. Также выложен для скачивания и использования в научных и учебных целях параллельный корпус объемом около 30 тыс. пар документов.

Используя приведенный подход можно создавать не только русско-украинский параллельный корпус, но и, вероятно, подобные корпусы для любых языков входящих в славянскую группу языков. Авторами планируется построение корпуса параллельных украинско-английских, русско-английских корпусов и украино-русско-анлийских корпусов, однако, для перехода к работе с нефлективными языками необходим пересмотр некоторых из приведенных алгоритмов.

К перспективам данной работы также можно отнести:

- расширение разнообразия много языковых корпусов;
- расширение украинско-русского параллельного корпуса;
- совершенствование программной оболочки для просмотра параллельных корпусов, а также выравнивание данных корпусов по предложениям;
- создание автоматических переводчиков на основании построенных корпусов.

Табл. 2. Наиболее частотные источники

№ п.п.	Украино-язычные источники	Кол-во публикаций	Русскоязычные источники	Кол-во публикаций
1.	ForUm	33547	ForUm	30903
2.	УНІАН	31573	УНИАН	26509
3.	РБК-Екраїна	21517	УКРИНФОРМ	25838
4.	УТРО-Україна	20019	РБК-Украина	21849
5.	УКРІНФОРМ	19031	Корреспондент.net	21646
6.	Оглядач	18460	УТРО-Украина	19769
7.	ProUa	14090	ICTV	19719
8.	Кореспондент.net	13505	ProUa	15189
9.	Укроп	12346	Обозреватель	14844
10.	ГлавРед	8905	ГлавРед	10475
11.	Новинар	8159	NewsRu.ua	6284
12.	NewsRu.ua	7377	Форпост	5621
13.	УКРИНФОРМ	6518	Подробности	4204
14.	Форпост	6017	Київ-Прес-Інформ	3385
15.	Вголос	5535	Zaxid.net	3081

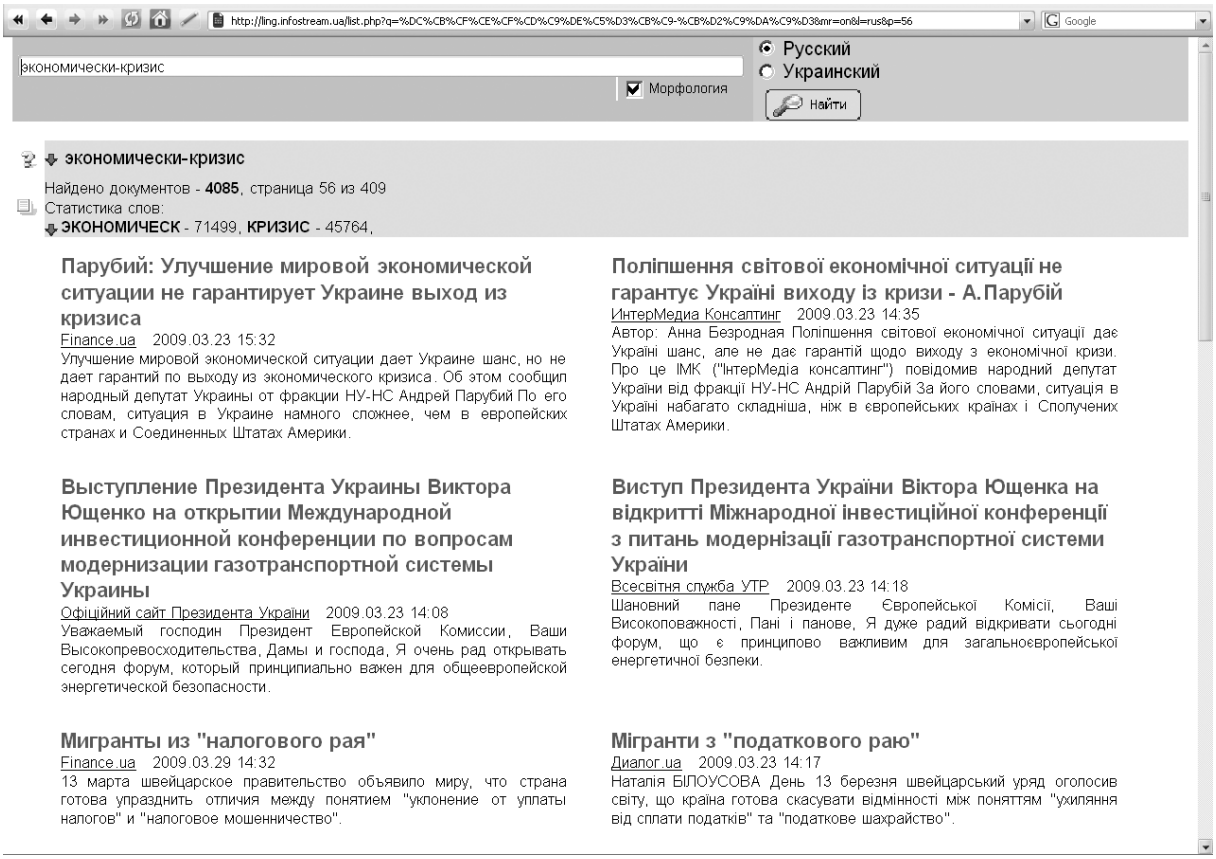


Рис. 2. Поисковый интерфейс для параллельного корпуса

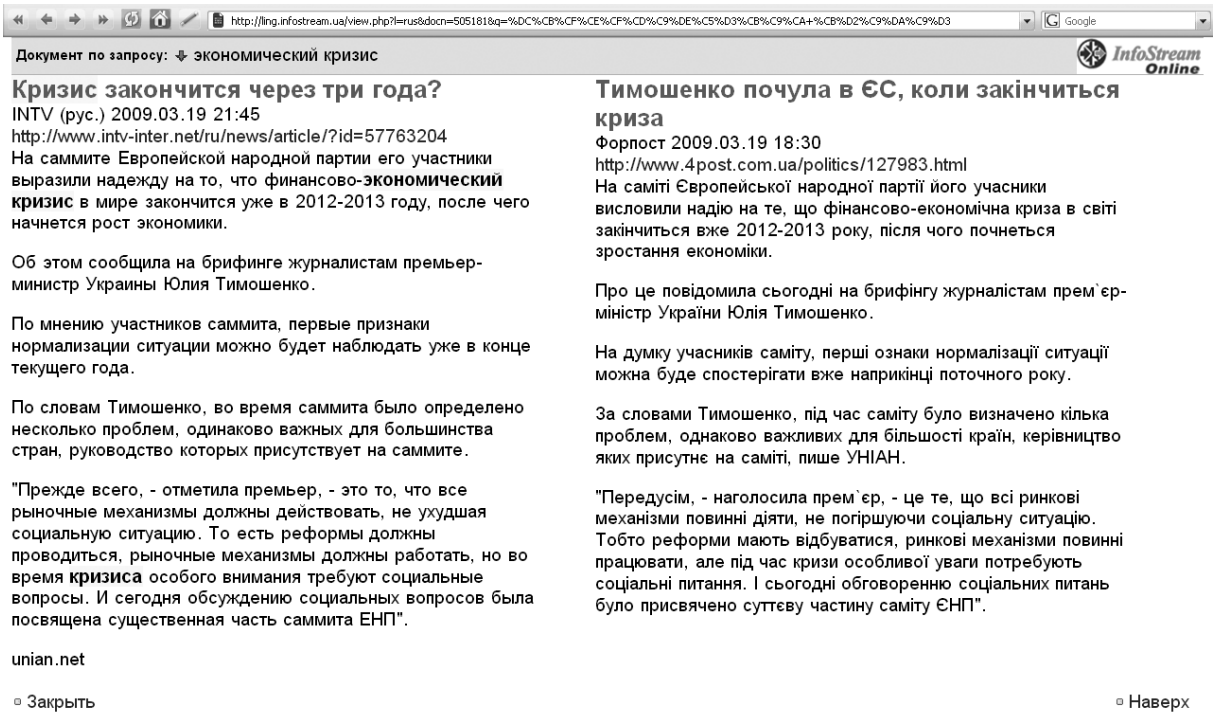


Рис. 3. Пример параллельных документов

Литература

1. *В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна.* Корпусна Лінгвістика — К.: Довіра, 2005. — 471 с.
2. http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html(сайт CRATER Multilingual Aligned Annotated Corpus)
3. http://nevmenandr.net/slovo/ (сайт Параллельного корпуса переводов «Слова о полку Игореве»).
4. http://www.ruscorpora.ru/corpora-biblio.html (сайт Национального корпуса русского языка).
5. *Гарабик Р., Захаров В.* Параллельный русско-словацкий корпус//Труды международной конференции Корпусная лингвистика — 2006. Sankt-Petersburg: St. Petersburg University Press 2006, s. 81–87.
6. *Resnik P.* Parallel strands: a preliminary investigation into mining the web for bilingual text. In D. Farwell, L. Gerber and E. Hovy (eds) Machine Translation and the Information Soup, Springer, Berlin, pp. 72–82.
7. *Resnik, P. and Smith, N. A.* 2003. The Web as a parallel corpus. Comput. Linguist. 29, 3 (Sep. 2003), pp. 349–380.
8. *Xiaoyi Ma, Mark Y. Liberman.* BITS: A Method for Bilingual Text Search over the Web// http://papers.ldc.upenn.edu/MTSVII1999/BITS.pdf
9. *Цинман Л. Л. , Сизов В. Г.* Лингвистический процессор ЭТАП: дескрипторное соответствие и обработка метафор // Труды межд. семинара Диалог‘2000. — М.: Изд-во РГГУ, 2000. — С. 366–369.
10. *Сокирко А. В., Ножов И. М.* Описание МаПоста // АОТ :: Технологии :: Описание МаПоста: http://www.aot.ru/docs/mapost.html (17 октября 2005 г.)
11. *Jurafsky D., Martin J. H.* Speech and Language Processing: An ntroduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall PTR, Upper Saddle River, NJ, 2000.
12. *Зеленков Ю. Г., Сегалович И. В., Титов В. А.* Вероятностная модель снятия морфологической омонимии на основе нормализубющих подстановок //Труды межд. конф. Диалог'2005. — М.: Наука, 2005.
13. *Баглей С. Г., Антонов А. В., Мешков В. С., Титов А. В.* Вероятностный подход к задаче разрешения омонимии слов и словарных пар // Труды межд. конф. Диалог'2007. 2007. С. 23–28.
14. *Зинькина Ю. В., Пяткин Н. В., Невзорова О. А.* Разрешение функциональной омонимии в русском языке на основе контекстных правил // Труды межд. конф. Диалог'2005. — М.: Наука, 2005. С. 198–202.
15. *Hearst M. A.* Noun homograph disambiguation using local context in large text corpora// Processing of the 7th conference on Research and Development in Information Retrieval ACM/SIGIR, pp. 36–47. — UW Centre for the New OED & Text Research Using Corpora, Pittsburgh, PA., 1991.
16. *Salton, G., Buckley, C.* Term-Weighting Approaches // Automatic Text Retrieval. Information Processing and Management, 24(5), pp. 513–523, 1988.
17. http://www.infostream.ua
18. http://www1.ulif.org.ua/ulif/
19. http://perevod.uaportal.com/
20. http://www.trident.com.ua/rus/online.php
21. http://translate.google.com/
22. http://www.xapian.org/docs/bm25.html
23. http://ling.infostream.ua

Редактор расширенных сетей переходов с графическим интерфейсом пользователя

The editor of the augmented transition networks with graphical user interface

Лебедев А. С. (andremoniy@gmail.com)

Московский государственный институт электроники и математики

В докладе описывается разработанный визуальный редактор расширенных сетей переходов, позволяющий облегчить работу эксперта-программиста лингвистического процессора, основанного на использовании таких сетей. Возможности редактора иллюстрируются на простых примерах.

1. Введение

Центральной задачей при создании лингвистического процессора является выбор модели и разработка семантического анализатора. В данной работе рассматривается модель на основе расширенных сетей переходов.

Расширенная сеть переходов, также известная под названием «ATN-сеть (augmented transition network)» — давно известный инструмент семантического анализа [1]. Технологии, основанные на использовании расширенных сетей переходов, используются, например, таким крупным и известным разработчиком программного обеспечения, как компания «ПРОМТ»[2]. Расширенная сеть переходов в описываемой реализации представляет собой однонаправленный граф без циклов, где дугам графа поставлены в соответствие наборы морфологических свойств лексем, а узлам — наборы семантических ролей, которые можно выделить на данном этапе прохождения по графу. Алгоритм работы предлагаемого ATN-анализатора будет рассмотрен ниже. Расширенные сети представляют собой, таким образом, один из механизмов, с помощью которых эксперты предметной области и эксперты-лингвисты могут принимать непосредственное участие в программировании лингвистического процессора.

Такая модель использует парадигму, когда знания конструируются человеком, т. е. человек-эксперт постоянно сопровождает систему, внося в нее по мере необходимости новые знания. Для этой цели важно иметь хороший инструментарий для разработчика. [1]

В данной работе ставится задача создания приложения с удобным пользовательским интерфейсом и набором необходимого инструментария, посредством работы с которым можно было бы облегчить процесс программирования лингвистического процессора за счет привлечения экспертов-лингвистов предметной области. Данная задача была решена путем создания **визуального редактора расширенных сетей переходов**, работа с которым не требует знаний в области программирования.

2. Модель предлагаемого ATN-анализатора

Для более глубокого понимания поставленной задачи и способа ее решения опишем модель разработанного анализатора на основе сети переходов.

База знаний в разработанном ATN-анализаторе представлена в виде набора расширенных сетей переходов. Каждая сеть в данной модели соответствует одной части речи, с которой начинается разбор предложения. Дуги в таких сетях помечены частью речи и набором морфологических признаков. Все пути в такой сети, ведущие от начального состояния к конечному, соответствуют некоторому правилу для разбора предложения. В узлах сети могут находиться некоторые правила и команды, управляющие работой ATN-анализатора. Правила представляют собой набор семантических связей, выделенных на данном этапе разбора. Команды позволяют изменять состояние внутренних переменных анализатора или использовать в качестве семантической

связи неопределенные формы глаголов (в случае, если требуется выражение семантики с помощью конкретной глагольной формы).

Разбор предложения происходит следующим образом. Морфологический анализатор[1] получает на входе словоформу, а на выходе выдает данную словоформу и набор ее морфологических признаков, ключевым из которых является часть речи. В случае если рассматриваемому слову соответствует несколько разных лексем (с разными частями речи), то на выходе морфологический анализатор выдает массив соответствующих наборов морфологических признаков для каждой возможной лексемы. В соответствии с частью речи первого слова предложения ATN-анализатор открывает соответствующую расширенную сеть переходов. Если найдено несколько частей речи, то для каждой создается копия семантического представления и открывается соответствующая расширенная сеть переходов.

Для иллюстрации рассмотрим простейший случай, когда каждой словоформе соответствует одна часть речи (т. е. отсутствуют морфологические многозначности по признаку части речи). Пример предложения:

(1) Старушка вынула из рабочего ящика нательный золотой крестик Наташи (Достоевский).

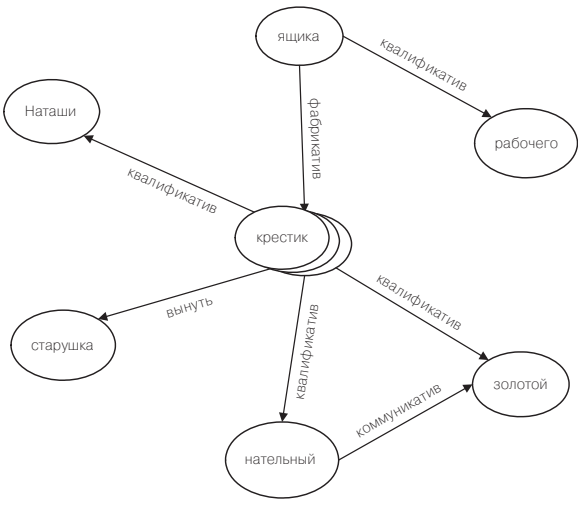


Рис. 1. Семантическое представление примера (1)

Семантическое представление (СП) предложения (1), полученное с помощью разработанного лингвистического процессора представлено на рис. 1. Для удобства разработчиков и экспертов система способна отображать схемы полученных семантических представлений в виде звезды. Ближе к центру рисунка располагаются лексемы, имеющие наибольшее число семантических и синтаксических связей.

При построении данного семантического представления ATN-анализатором использовался следующий фрагмент расширенной сети переходов — сеть1 (рис. 2). Числа в вершинах графа — это уникальные идентификаторы вершин.

Как видно на рис. 2. сеть перегружена неопределенными формами слов, и представляет собой синтаксически-семантический разбор конкретного примера (1) и поэтому не применима для других предложений. В качестве универсализации можно ослабить морфологические свойства ребер и исключить из них неопределенные словоформы. Тогда данная сеть (рис. 3) приобретет более облегченный вид, что позволит использовать ее для анализа целого класса предложений, имеющих такую же структуру.

Неопределенная форма у глагола оставлена, так как глагол определяет в данном контексте лингвистическое отношение между подлежащим *старушка* и дополнением *крестик*. На рис. 3 также представлен список морфологических признаков, назначенных ребру *180-18001*, которое описывает переход в случае, если следующее слово из входного потока является именем собственным, находящимся в родительном или именительном падеже, женского рода. Данная расширенная сеть уже более универсальна, и с помощью нее можно разобрать такие предложения, как:

(2) Девочка вынула из потертого рюкзака большой тяжелый учебник Маши.

(3) Гувернантка вынула из старого сундука белое свадебное платье Анны.

Следующим шагом универсализации сети является удаление из фреймов таких морфологических свойств, как «род», «число» и т.п. Таким образом, мы получим сеть3, которая позволяет разбирать уже более широкий класс предложений, в том числе и таких как:

(4) Мальчик вынул из синего моря ветхую рыбацкую сеть Петра.

Итак, для программирования ATN-анализатора требуется редактор, в возможности которого входят следующие основные функции:

- внесение в базу знаний новых расширенных сетей переходов;
- редактирование имеющихся сетей: добавление, удаление и изменение узлов сети, редактирование фреймов, назначаемых ребрам и узлам сети;
- поиск по имеющимся сетям.

---

[1] В работе применялись для исследований два морфологических анализатора:

а) ABBYY Retrieval & Morphology 4.0 Engine — инструментарий разработчика программного обеспечения;

б) Парсер mystem (http://company.yandex.ru/technology/mystem/) [3].
Наиболее положительный результат морфологического анализа показал парсер mystem.

Рис. 2. Фрагмент варианта расширенной сети переходов для примера (1) (Сеть1)

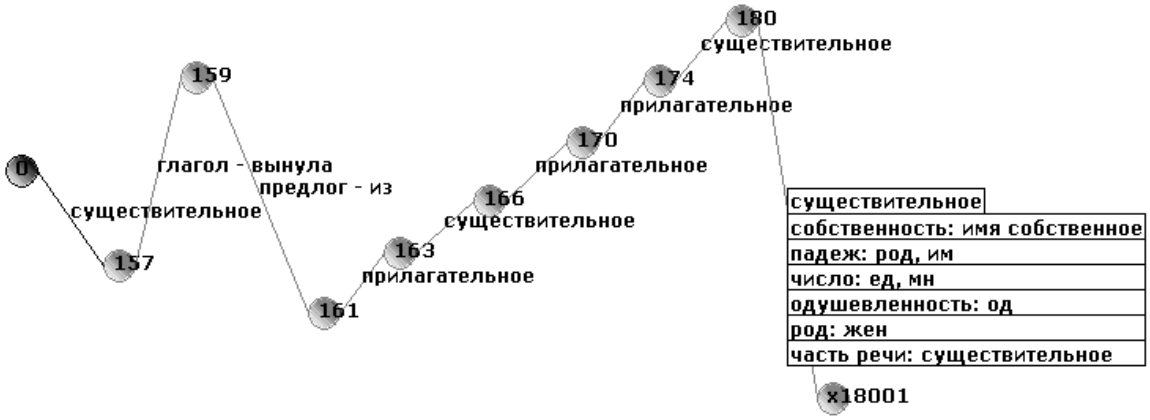


Рис. 3. Фрагмент расширенной сети переходов без словоформ для примера (1) (Сеть2).

Рис. 4. Сеть3

3. Визуальный редактор расширенных сетей переходов

Расширенные сети переходов представляют собой *семантическую память* (СП) [4]. В приведенном выше примере разработки расширенной сети заменимость морфологического анализатора достигается за счет его реализации в отдельном пакете.

Интерфейс редактора представляет собой окно, позволяющие отобразить в отдельных вкладках несколько расширенных сетей в виде графов (см. рис. 5).

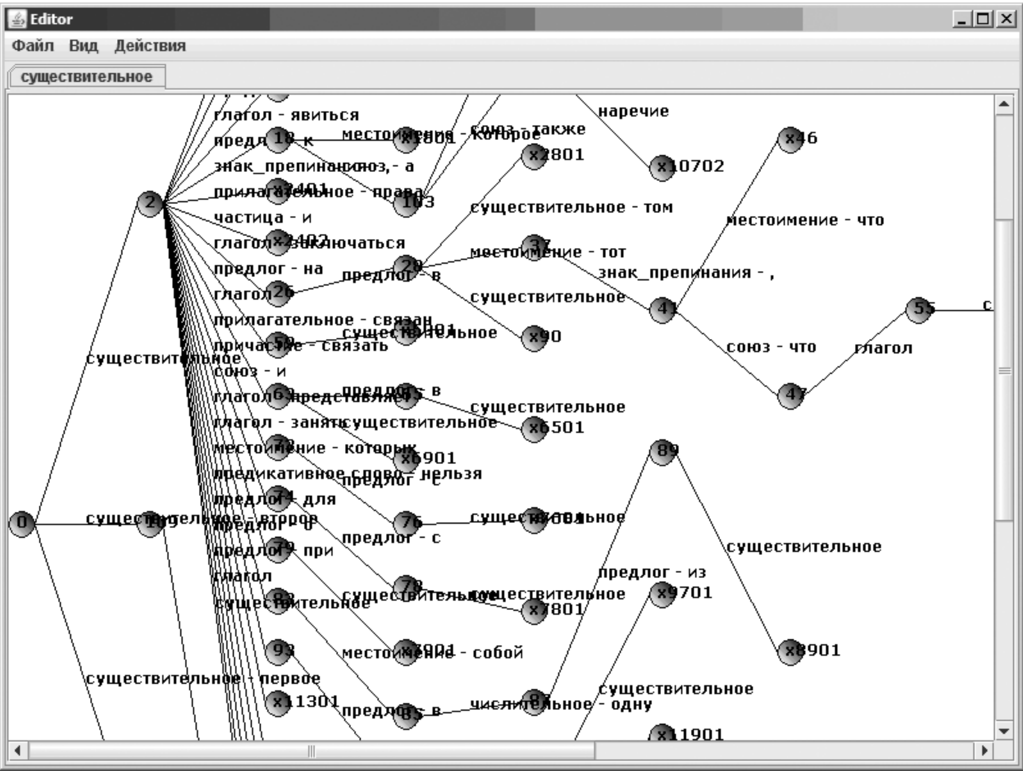


Рис. 5. Экран визуального редактора расширенных сетей переходов

переходов рассматривался простой случай, где не учитывалась многозначность слов и словоформ[2]. Кроме того, в узких предметных областях только эксперт предметной области может грамотно выделить те ключевые термины, которые особенно важно обработать исключительным образом при разборе предложения, т.е. термины, исключительным образом влияющие на семантику текста (и предложения в частности). В данном докладе описывается именно такой инструментарий, представленный в виде визуальной среды для редактирования расширенных сетей переходов (в модели, описанной в п. 2).

Редактор представляет собой оконное приложение, написанное на языке Java 1.5. В качестве морфологического анализатора теоретически может использоваться любой, практически же один из рассмотренных выше (программа mystem или морфологический анализатор Abbyy). Практическая

Конструирование сети производится с помощью мыши. Правым щелчком мыши у выбранного узла сети создаются новые ребра. Нумерация узлов производится автоматически. Признаком конца сети является латинская буква «x» в начале имени узла. Морфологические свойства, которыми обладают ребра, могут быть заданы путем выбора нужного ребра, двойным щелчком мыши на выбранном ребре и последующим редактированием окна ввода свойств ребра (см. рис. 6). Во вкладке «Собственные свойства» задаются парные морфологические свойства в виде «свойство — значение», например: «падеж — вин».

Аналогичным образом осуществляется редактирование свойств узла (см. рис. 7). Под свойствами узла также понимаются парные совокупности, типа «свойство — значение». Название свойств не регламентировано, что позволяет при разработке ATN-анализатора не привязываться к фиксированным именам-константам: например, в качестве имени свойства, описывающего семантическое отношение в использованной реализации ATN-анализатора, выбрано имя «*role*». Семантическая роль описывается как заключенная в круглые скобки последовательность числовых значений, выражающих смещение

---

2 Например, морфологический анализатор воспримет словоформу «рабочий» как существительное и прилагательное одновременно, выдавая на выходе описание этих двух лексем. Следовательно, и расширенную сеть переходов необходимо конструировать с учетом многозначности слов.

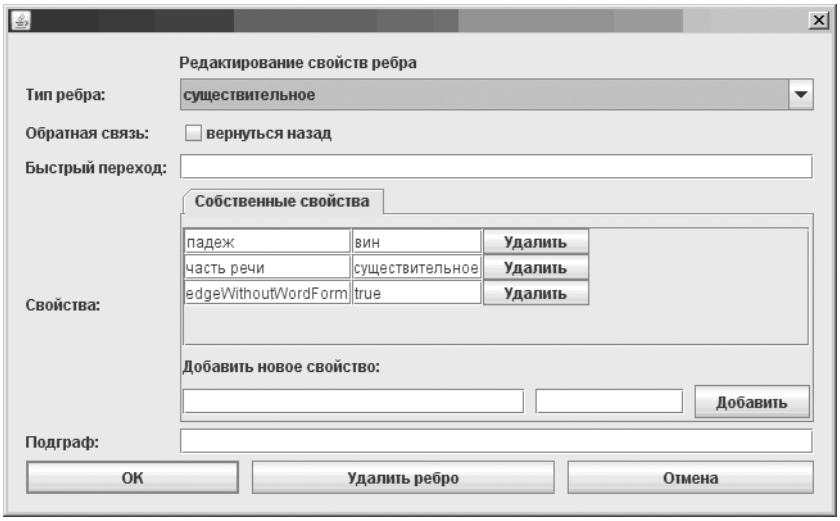


Рис. 6. Экран редактирования свойств ребра

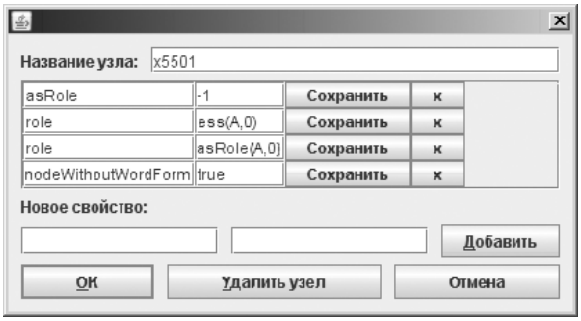


Рис. 7. Экран редактирования свойств узла

влево относительно текущего узла сети. Например, если в семантическое соотношение нужно включить слово, связанное с текущим ребром (т.е. ребром, «правым» концом которого является текущий узел), то его смещение будет равно «0», предыдущее ребро — «−1» и т.д. Порядок указания смещений соответствует порядку в описании семантической роли. Пользователь может выбрать узлы, участвующие в создаваемой роли, путем щелчка мыши по нужным узлам в соответствующем порядке. Затем достаточно выбрать узел для создания роли и указать ее название.

Редактор предоставляет пользователю широкие возможности по визуальному конфигурированию графа на экране, в том числе автоматическое выстраивание узлов сети, обеспечивающие удобное зрительное восприятие картинки, а именно, исключает пересечения ребер графа и представляет его в виде древа. В редакторе реализована функция прокрутки экрана, что позволяет редактировать сети практически не ограниченного объема[3].

Для ускорения конструирования сетей в редакторе представлена возможность автоматического построения ветвей графа по введенному предложению. Данный процесс выглядит следующим образом: пользователь вводит предложение, которое разбивается на отдельные лексические единицы (слова, знаки препинания, цифры), и каждая единица обрабатывается встроенным морфологическим анализатором. В строгой последовательности, в соответствии с введенным предложением, система строит ветвь сети, где каждому ребру приписывается набор морфологических признаков, взятых у соответствующего слова. Пользователю редактора остается только удалить лишние морфологические свойства и добавить нужным узлам сети свойства или семантические роли.

При большом числе вершин графа картинка может оказаться весьма громоздкой и трудной для восприятия. Для устранения этого эффекта предлагается использование фильтров, в частности:

- «только со словоформами» (т. е. отображаются только те дуги, чьи фреймовые структуры содержат конкретные словоформы)
- «только универсальные» (т. е. отображаются только те дуги, чьи фреймовые структуры не содержат конкретных словоформ).

В функционал программы включен также поиск по сетям. Экран окна поиска представлен на рис. 8.

Поиск можно производить в обычном режиме, а также в комбинации режимов «строгий поиск со словоформами» и «полное совпадение морфологических признаков».

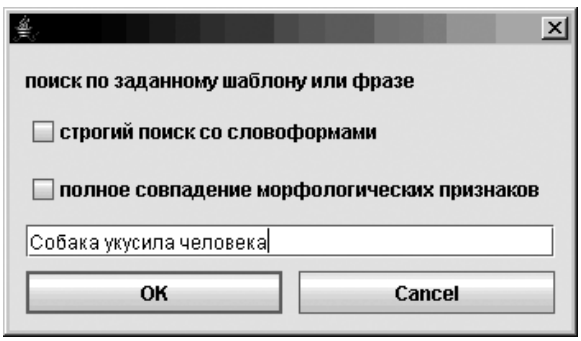


Рис. 8. Экран окна поиска

[3] Объемы ограничены только аппаратными характеристиками ЭВМ, на которой эксплуатируется редактор.

Функция поиска также использует морфологический анализатор. В обычном режиме будут найдены все пути графа, позволяющие произвести разбор данного предложения. В режиме строго поиска отобразятся те пути графа для разбора данного предложения, дуги которых содержат точное совпадение по словоформам введенной строки поиска. В режиме полного совпадения морфологических признаков отобразятся те пути графа, чьи дуги содержат морфологические признаки, которые совпадают с соответствующими признаками слов строки поиска, однако совпадение словоформ в данном режиме не обязательно.

В системе также реализован механизм внутреннего поиска, который запускается каждый раз при вызове автоматической перестройки графа. Данный алгоритм позволяет автоматически отслеживать одинаковые фрагменты сети, расположенные в разных частях графах и при возможности их объединять. Под двумя одинаковыми фрагментами в данном случае понимаются такие, которые имеют началом либо корневой узел графа, либо оканчиваются терминальными узлами (помеченными префиксом «x»), и имеют совпадающий по длине и соответствующим свойствам ребер и узлов некоторый путь. В частности, данный механизм удобен в случае использования автоматического построения «скелета» сети по шаблону-строке, когда после окончания редактирования свойств ребер требуется перестроить граф, объединив одинаковые фрагменты.

Сконструированные сети сохраняются в виде XML файлов, что упрощает применение созданных файлов в стороннем программном обеспечении благодаря легкости обработки языка XML. Классы, описывающие объекты графов, предоставляются в открытом доступе; таким образом, их можно использовать непосредственно при программировании на языке Java для использования сконструированных расширенных сетей переходов. Ниже приведен фрагмент XML файла, описывающий один узел сети:

```
<void property="node1">
  <object id="GraphXNode0" class="g.GraphXNode">
    <void property="name">
      <string>0</string>
    </void>
    <void property="parent">
      <object idref="GraphX0"/>
    </void>
    <void property="properties">
      <void method="add">
        <object class="g.GraphXProperty">
          <void property="name">
            <string>editor_x</string>
          </void>
          <void property="value">
            <string>0</string>
          </void>
        </object>
        </void>
        <void method="add">
          <object class="g.GraphXProperty">
            <void property="name">
              <string>editor_y</string>
            </void>
            <void property="value">
              <string>320</string>
            </void>
          </object>
        </void>
      </void>
    </object>
  </void>
```

4. Сравнение с известными графическими редакторами графов

Было произведено сравнение созданного визуального редактора с несколькими известными бесплатными и общедоступными графическими редакторами, применяемых специалистами для создания и редактирования ATN-сетей и Синтаксических деревьев.

1) Augmented Syntax Diagram (ASD) Editor and Parser — редактор Расширенных Синтаксических Диаграмм, предложенный профессором Джеймсом А. Масоном (James A. Mason). Сравнение ASD и ATN сетей произведено проф. Д. Масоном в статье [5]. Отметим схожие черты ASD сетей, их редактора и предложенной в данной статье модели ATN сетей и описываемого визуального редактора:

- использование числовых индексов, или номеров этапов, в узловых метках для различения между разными узлами, которые помечены такими же словарными элементами;
- графический интерфейс, позволяющий конструировать графы на экране, а также изменять местоположение элементов с помощью мыши.

К недостаткам данного ASD-редактора можно отнести:

- недостаточно понятный интерфейс визуального отображения сетей и функций их редактирования;
- отсутствие функций поиска и автоматического построения заготовок сети;
- система адаптирована большей частью для обработки английского языка.

В целом, можно отметить, что данный ASD-редактор отражает в большей степени видение модели автором (проф. Д. Масоном), что накладывает некоторые ограничения на его широкое применение.

2) Linguistic Tree Constructor – визуальный редактор для анализа текстов с помощью синтакси-

ческих деревьев [6]. Программа имеет платформонезависимую реализацию, поддерживает любые языки для описания синтаксических деревьев.

К недостаткам системы можно отнести:

- слишком упрощенную модель отображения деревьев;
- обилие англоязычных сокращений-терминов;
- неудобный графический инструментарий;
- собственные форматы файлов данных, что затрудняет использование их в стороннем ПО.

3) TreeForm Syntax Tree Drawing Software — мощный инструмент для создания и редактирования синтаксических деревьев [7]. Программа имеет платформо-независимую реализацию, поддерживает различные языки при конструировании деревьев.

Можно выделить следующие преимущества системы:

- оригинальный и удобный интерфейс для построения синтаксических деревьев;
- возможность настройки цветовых схем интерфейса пользователя;
- экспорт деревьев в графических форматах (JPEG и PNG);

К недостаткам системы TreeForm можно отнести следующее:

- обилие англоязычных сокращений-терминов;
- отсутствие функции поиска;
- отсутствие нумерации узлов сети.

5. Заключение

Созданный визуальный редактор расширенных сетей переходов позволяет облегчить работу эксперта-программиста лингвистического процессора, основанного на использовании ATN-анализатора. Он включает в себя все необходимые функции, связанные с обработкой графов, назначением свойств ребер и вершин. Включение в состав редактора морфологического анализатора позволяет строить «скелеты» сетей по заданному предложению-шаблону, а алгоритм внутреннего поиска отслеживает наличие подобных структур в уже имеющейся базе знаний, что упрощает создание и расширение существующих сетей.

На использовании морфологического анализатора также основана функция поиска, которая позволяет эксперту быстро находить нужные фрагменты сетей для дальнейшего анализа или редактирования.

Фильтрация отображения позволяет отделить на визуальном уровне универсальные части сети от частей, предназначенных для обработки определенных особых словоформ и конструкций.

В комплексе с данным редактором разработан ATN анализатор, использующий файлы с описанием сетей, созданные с помощью редактора.

В настоящий момент разработанный редактор в совокупности представляемых им набором функций и удобным пользовательским интерфейсом является уникальным инструментом подобного типа.

Литература

1. *Люгер, Джордж, Ф.* Искусственный интеллект: стратегии и методы решения сложных проблем, 4-е издание. Пер. с англ. — М.: Издательский дом «Вильяме», 2003. — 864 с.
2. *Технологии* компании ПРОМТ. [Электронный ресурс] — Режим доступа: http://www.promt.ru/company/technology/promt/ — Загл. с экрана.
3. *A fast* morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. [Электронный ресурс] — Режим доступа: http://company.yandex.ru/articles/iseg-las-vegas.xml — Загл. с экрана.
4. *Кузнецов И. П.* Семантические представления. М.: Наука, 1986. — 296 с.
5. *James A. Mason.* Augmented Syntax Diagram Grammars. [Электронный ресурс] — Режим доступа: http://www.yorku.ca/jmason/asdgram.htm — Загл. с экрана.
6. *Linguistic* Tree Constructor: About LTC. [Электронный ресурс] — Режим доступа: http://ltc.sourceforge.net/about.html — Загл. с экрана.
7. *TreeForm* Syntax Tree Drawing Software, Version 1.0.3. [Электронный ресурс] — Режим доступа: http://www.ece.ubc.ca/~donaldd/treeform.htm — Загл. с экрана.

Проблема разрешения «Ё»-омографов при синтезе речи по тексту

The problem of the «Ё»-homographs resolution in text-to-speech synthesis

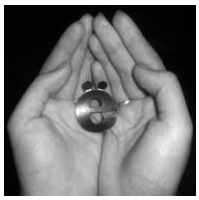
Лобанов Б. М. (lobanov@newman.bas-net.by),

Объединенный институт проблем информатики НАН Беларуси, Минск, Беларусь

В статье рассматривается проблема адекватного разрешения неопределенностей в системах синтеза речи по тексту, связанных с частным случаем омонимии — графической «Ё»-омонимией. Рассмотрены статистические характеристики омографических пар, в том числе «Ё»-омографов. Исследованы статистические характеристики распределений внутри наиболее часто встречающихся пар «Ё»-омографов. Обсуждаются пути разрешения наиболее частотной омографической пары «ВСЁ» и «ВСЕ».

“Когда же расставите точки над «ё»? Ё моё!!!”

LobanoPhone — 2000



Введение

Проблема адекватного разрешения неопределенностей, связанных с омонимией, играет существенную роль в решении задач распознавания и синтеза речи. Наиболее важное значение эта проблема приобретает при решении задач преобразования «речь — текст» (распознавание речи), когда существенным является разрешение почти всех видов омонимии: синтаксической, грамматической, лексической, словообразовательной и фонетической (см. словарь лигвистических терминов [1]). Только один вид омонимии — графическая омонимия, не играет роли в решении задач распознавания речи. Зато этот единственный вид омонимов, называемых омографами, играет весьма существенную роль в задачах преобразования «текст — речь» (синтез речи). Игнорирование существования омографов нарушает смысловое восприятие синтезированной речи и дополнительно ухудшает её естественность. Нам не известно ни одной работы, направленной на анализ и решение проблемы адекватного разрешения неопределенностей при синтезе русской речи по тексту, связанных с существованием омографов. В данной работе мы попытаемся в какой-то степени заполнить этот пробел, опираясь на фактический материал, представленный в словаре омографов русского языка [2].

В русском языке существуют два источника графической омонимии: вариативность *словесного ударения*, местоположение которого в письменной речи не указывается (СУ-омографы), и письменная традиция не обязательного проставления необходимых точек на *букве «Ё»* («Ё»-омографы). Литера «Ё» была предложена княгиней Екатериной Дашковой в 1783 году, а в печати употреблена в 1795 году. Отдельной буквой она долгое время не считалась и в азбуку официально не входила. В русском языке буква «Ё» используется, чаще всего в тех позициях, где произношение [(j)o] образовалось из [(j)e], чем и объясняется производная от «Е» форма буквы, хотя с точки зрения фонетики логичней было бы поставить точки не над «Е», а над «О». Букве «Ё» — 225 лет. Хотя она родилась в Санкт Петербурге, однако 20 октября 2001 года в Ульяновске открылся единственный в мире памятник букве «Ё» (см. фото).

Существует много различных мнений, как в пользу, так и против непременного исполь-

зования буквы «Ё» в печатном тексте (см. http://www.yomaker.ru/). С нашей позиции — позиции разработчиков систем синтеза речи по тексту — отсутствие в тексте «Ё» влечёт за собой дополнительные трудности, которые должны быть разрешены в той или иной степени. Простейшее решение — игнорирование проблемы — влечёт за собой дополнительные трудности в восприятии синтезированной речи и к раздражающему слух Е-канию. Данная работа посвящена исследованию статистических закономерностей проявления «Ё»-омонимии в различных текстах, а также обсуждению вопросов разрешения связанных с ней неопределённостей.

1. Статистические характеристики омографических пар

Статистические исследования проводились с использованием специально разработанной программы «HOMOGRAPH STATISTICS» и электронного словаря омографов, созданного на основе книжного словаря [2]. Целью исследования являлось определение статистической значимости «Ё»-омографов в общем списке «СУ»- и «Ё»-омографов [2], а также выявление особенностей статистических распределений только внутри подкласса «Ё»-омографов. Общее количество омографов, в соответствии с приведенными в [2] данными, составляет 3894 пар, из них «Ё»-омографов — только 232 пары.

Статистические характеристики определялись в отдельности для достаточно представительных и различных типов текстов:

- А. С. Пушкин — стихотворные произведения,
- Л. Н. Толстой — роман «Анна Каренина,
- Б. Акунин, Д. Рубина, Л. Петрушевская — современная проза,
- Труды конференции «ДИАЛОГ-2006» — научная проза.

В таблице 1 приведены интегральные статистические характеристики этих текстов по всей совокупности омографов, содержащихся в словаре [2].

Как видно из таблицы 1, выбранные тексты различных жанров имеют примерно одинаковый объём, в среднем — около 300 тыс. слов. Средний процент вхождения омографов составил 3,15%. Если считать, что среднее число слов на странице равно 650, то около 20-ти слов могут оказаться омографами. В случае их неадекватного раскрытия, как показывает опыт, это приводит к весьма негативному впечатлению при прослушивании синтезированной речи. Из таблицы видно также, что наибольшее количество омографов встречается в современной прозе, а наименьшее — в научном тексте. Очень интересный факт вытекает при рассмотрении 4-го столбца таблицы: всего только порядка 20% от общего многообразия всех омографических пар встречается в проанализированных текстах! Это указывает на первостепенную важность этого подмножества в решении задач разрешения омографии.

В таблице 2 приведены статистические характеристики 4-х классов текстов по совокупности пар «Ё»-омографов, содержащихся в словаре [2].

В сравнении с данными таблицы 1, средний процент вхождения «Ё»-омографов значительно ниже и составил 0,59%, что соответствует их общему количеству. Однако, если сравнить отношение количества всех пар омографов к количеству «Ё»-омографов: ***3894/232*=16,8** и соответствующее отношение процентов их вхождения в тексты: **3,15/0,59=5,3**, то можно отметить более чем 5-ти кратную частотность «Ё»-омографов, а, следовательно, существенную важность разрешения этого вида омографии при синтезе речи. Как и в случае таблицы 1, только порядка 30% от общего многообразия всех «Ё»-омографических пар встречается в проанализированных текстах.

В таблице 3 приведены дифференциальные характеристики статистического анализа текстов по всей совокупности омографов (первые 15 наиболее частотных пар омографов), содержащихся в словаре [2]. Как видно из таблицы, во всех художественных текстах пара «Ё»-омографов слова «***все***» выдвинулась на 1-е место. В специфическом научном тексте «Диалог-06» омограф «***все***» уступил 1-е место, к нашему удовольствию, омографу «***слова***». Из таблицы видно также, что и некоторые другие «Ё»-омографы вошли в число наиболее частотных: ***«перед, всем».*** На рисунке 1 графически представлены распределения количества встречаемости в различных текстах 10-ти наиболее частотных пар омографов. Из рис. 1 видно, что пары омографов наиболее равномерно распределены (*а, следова-*

Таблица 1. Результаты теста по всем омографам

Тип текста	Общее количество слов в тексте	Общее количество пар омографов	Число различных пар омографов
Словарь омографов [2]	—	3894 (100%)	3894 (100%)
А. С. Пушкин	266,726 (100%)	9,421 (3,53%)	827 (21,2 %)
Л. Н. Толстой	279,448 (100%)	8,747 (3,13%)	680 (17,5%)
Б. Акунин и др.	379,277 (100%)	13,630 (3,59%)	1088 (27,9%)
«ДИАЛОГ-2006»	305,742 (100%)	7,195 (2,35%)	563 (14,5%)
Среднее количество	307,775 (100%)	3,15%	20,3%

Таблица 2. Результаты теста по «Ё»-омографам

Тип текста	Общее количество слов в тексте	Общее количество пар «Ё»-омографов	Число различных пар «Ё»-омографов
Словарь омографов [2]	—	232 (100%)	232 (100%)
А. С. Пушкин	266,726 (100%)	1,411 (0,53%)	71 (30,6%)
Л. Н. Толстой	279,448 (100%)	2,276 (0,81%)	56 (24,1%)
Б. Акунин и др.	379,277 (100%)	2,935 (0,77%)	82 (35,3%)
«ДИАЛОГ-2006»	305,742 (100%)	810 (0,26%)	49 (21,1%)
Среднее количество	307,77 (100%)	0,59%	27,8%

Таблица 3. Результаты теста по всем омографам

А. С. Пушкин		Л. Н. Толстой		Б. Акунин		Диалог-06	
все	458	***все***	1670	***все***	1963	*слова*	735
уж	436	*уже*	601	*уже*	811	***все***	433
уже	361	*надо*	376	*потом*	555	*уже*	247
перед	214	*потом*	229	*уж*	345	*связи*	184
моя	204	*глаза*	211	*глаза*	328	*части*	133
всем	132	*слова*	173	*надо*	319	*корпуса*	133
глаза	126	*уж*	164	*руки*	270	*стороны*	125
сердца	120	*тому*	144	***перед***	265	*правила*	124
слова	113	*голову*	143	*голову*	198	*правило*	118
потом	112	*руки*	143	*дома*	168	*оно*	114
ночи	108	***всем***	125	***всем***	141	***перед***	105
тому	98	*дома*	124	*самом*	137	*тона*	103
пора	95	*лица*	112	*слова*	127	*рода*	101
души	95	*дела*	100	*моя*	123	*второй*	93
мою	92	*должно*	91	*двери*	109	*свойства*	91

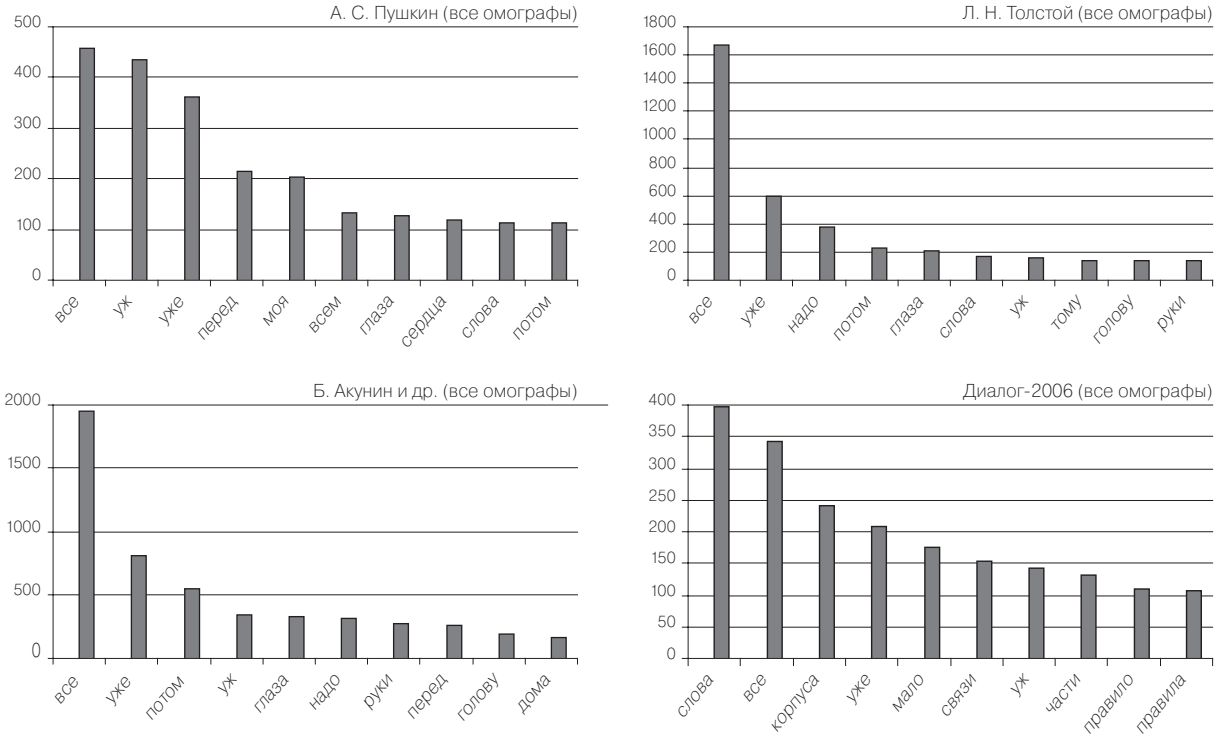


Рис. 1. Распределения встречаемости пар омографов в различных текстах

тельно, наиболее информативны!) в стихотворных произведениях А.С. Пушкина и в научных трудах участников «ДИАЛОГа».

В таблице 4 приведены дифференциальные статистические характеристики текстов — первые 15 наиболее частотных пар «Ё»-омографов, содержащихся в словаре [2]. Как и ожидалось 1-е места во всех текстах заняла пара омографов ***«все»***. Соответствующие таблице графичекие распределения представлены на рис. 2.

2. Статистические характеристики распределений внутри пар «Ё»-омографов

Для определения статистических характеристик распределений внутри пар «Ё»-омографов использовались результаты описанного выше статистического анализа дифференциальных характеристик пар «Ё»-омографов и данные Интернет ресурса [3] «*Поиск по акцентуированному корпусу*». Вначале

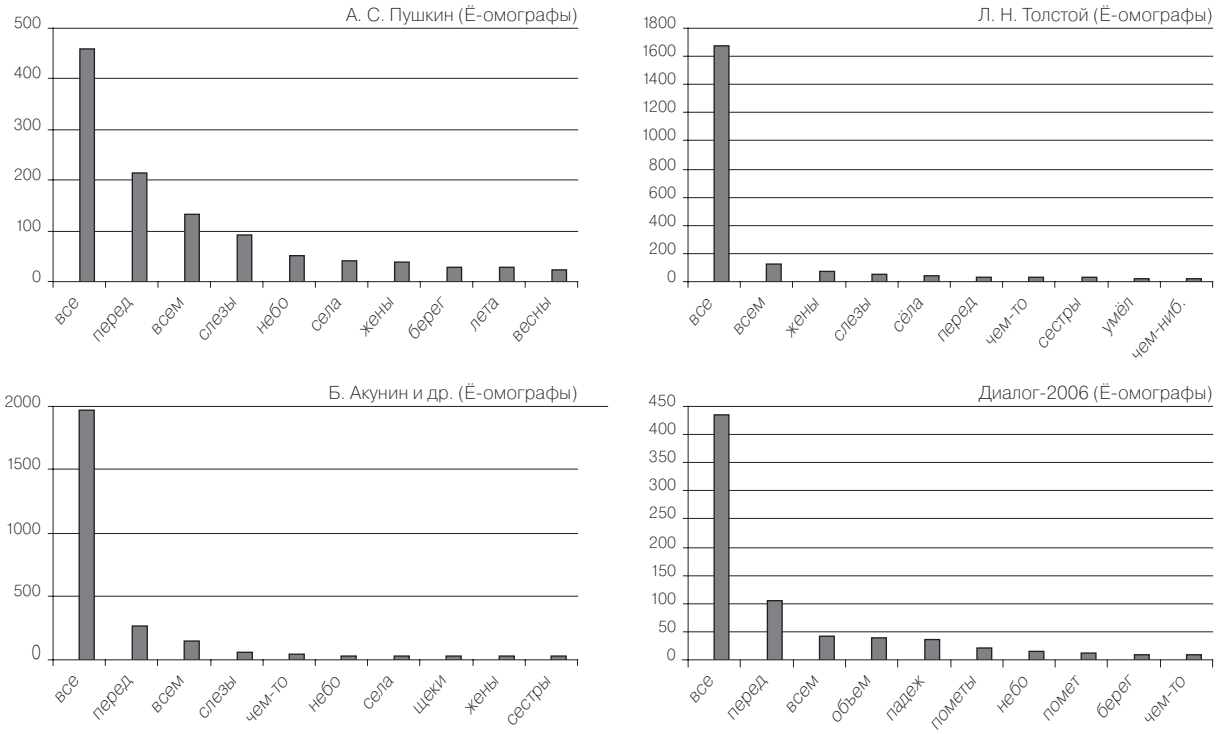


Рис. 2. Распределения встречаемости пар «Ё»-омографов в различных текстах

Таблица 4. Результаты теста по «Ё»-омографам

А.С. Пушкин		Л.Н. Толстой		Б. Акунин		Диалог-06	
все	458	***все***	1670	***все***	1963	***все***	433
перед	214	***всем***	125	***перед***	265	***перед***	105
всем	132	***жены***	78	***всем***	141	***всем***	41
слезы	92	***слезы***	53	***слезы***	56	*объем*	39
небо	50	***села***	45	***чем-то***	37	*падеж*	36
села	42	***перед***	35	***небо***	34	*пометы*	20
жены	37	***чем-то***	32	***села***	32	***небо***	14
берег	29	***сестры***	27	*щеки*	28	*помет*	12
лета	28	*умел*	22	***жены***	26	***берег***	10
весны	23	*чем-нибудь*	17	***сестры***	23	***чем-то***	9
умел	22	***небо***	14	*счета*	21	*ребра*	7
небо	21	*щеки*	11	*умел*	17	*запрет*	7
смел	18	*звезды*	9	*стекла*	17	***жены***	6
небом	15	*весел*	9	*небе*	14	*полет*	6
небе	14	*весны*	8	*осел*	14	*черта*	6
берет	13	*черта*	8	*черта*	13	*села*	4

Таблица 5. Парная и внутрипарная встречаемость «Ё»-омографов

Пара «Ё»-омографов	Количество пар в 4-х текстах	Количество пар в Корпусе	Количество Ё-слов в корпусе	Количество Е-слов в корпусе	Соотношение внутри пар	
					% кол. «Ё»	% кол.«Е»
1	*2*	*3*	*4*	*5*	*6*	*7*
все	4524	5970	4143	1826	**100**	**44**
перед	620	640	0	640	0	100
всем	440	505	109	362	**28**	**100**
слезы	200	60	60	1	100	2
села	120	64	2	62	3	100
небо	100	126	0	126	0	100
чем-то	80	123	53	70	**75,7**	**100**
жены	64	49	14	35	**40**	**100**
сестры	52	34	24	10	**100**	**42**
берег	40	85	4	81	5	100

из таблицы 4 были отобраны 10 наиболее частотных пар «Ё»-омографов по всем рассмотренным выше ***4-м текстам*** (помечены жирным шрифтом в табл. 4) и подсчитаны суммарные количества их встречаемости (см. столбец 2 таблицы 5 и рис.3). Затем для этих слов с помощью Интернет ресурса [3] в ***Корпусе*** текстов по драматургии, беллетристике, публицистике и научно-популярной литературе определены суммарные количества их встречаемости (см. столбец 3 таблицы 5 и рис. 3). В столбцах 4, 5 приведены результаты встречаемости в Корпусе [3] «Ё» и «Е» слов (см. также рис. 4), в столбцах 6, 7 — соотношение количества слов с «Ё» и «Е» в процентах внутри пар «Ё»-омографов (см. также рис. 5).

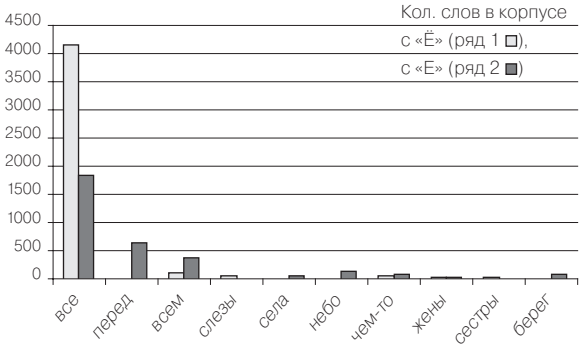


Рис. 4. Распределения кол. слов с «Ё» (ряд 1) и «Е» (Ряд 2) внутри пар «Ё»-омографов

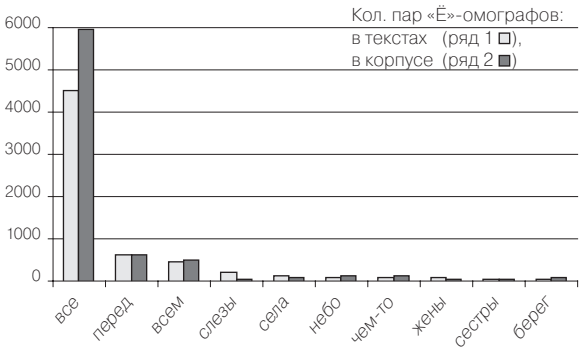


Рис. 3. Распределения встречаемости 10-ти наиболее частотных пар «Ё»-омографов

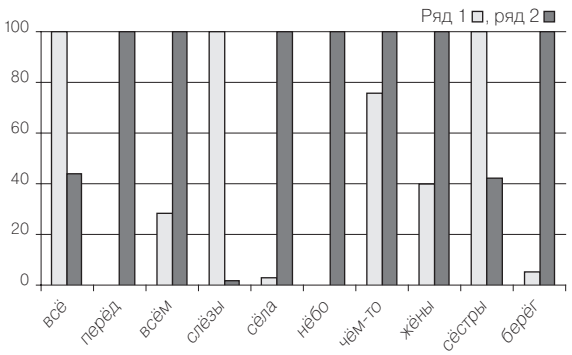


Рис. 5. Соотношения кол. слов в % с «Ё» (Ряд 1) и «Е» (Ряд 2) внутри пар «Ё»-омографов

3. Некоторые правила разрешения «Ё»-омографической неопределённости

Анализируя результаты, приведенные в таблице 5 и на рис. 3 — 4, можно сделать следующие выводы.

1. Как видно из табл. 5 (столбцы 2 и 3) использованная для статистического анализа выборка ***Текстов*** (А.С. Пушкин — стихотворные произведения, Л.Н. Толстой — роман «Анна Каренина, Борис Акунин, Дина Рубина, Людмила Петрушевская — современная проза, Труды конференции «ДИАЛОГ-2006» — научная про-

за) является достаточно представительной и сравнимой по объёму с ***Корпусом*** текстов по драматургии, беллетристике, публицистике и научно-популярной литературе, представленном в [3].

2. Полученные распределения встречаемости 10-ти наиболее частотных пар «Ё»-омографов в изученных ***Текстах*** и в ***Корпусе*** в высокой степени подобны (см. рис. 3), что говорит о достаточной степени достоверности полученных данных.
3. Из рис. 3 и 4 следует, что подавляющее количество «Ё»-омографов как ***Текстах,*** так и в ***Корпусе*** приходится на пару омографов «ВСЕ», что подчёркивает исключительную важность нахождения правил их разрешения при синтезе речи.
4. Из табл. 5 (столбцы 6, 7), а также из рис. 5 видно, что в 5-ти из 10-ти наиболее частотных пар «Ё»-омографов появление той или иной реализации омографа в паре более или менее равновероятно (пары: *ВСЁ_ВСЕ, ВСЁМ_ВСЕМ, ЧЁМ-ТО_ЧЕМ-ТО, ЖЁНЫ_ЖЕНЫ, СЁСТРЫ_СЕСТРЫ*). В оставшихся 5-ти парах с высокой степенью достоверностью можно выбирать варианты: *ПЕРЕД, СЛЁЗЫ, СЕЛА, НЕБО, БЕРЕГ*.
5. Для пар омографов: *ВСЁМ_ВСЕМ, ЧЁМ-ТО_ЧЕМ-ТО,* слова с «Ё» с высокой степенью достоверностью могут быть определены по наличию перед ними предлогов *«о», «об»* или *«обо»*.
6. Для пар омографов: *ЖЁНЫ_ЖЕНЫ, СЁСТРЫ_СЕСТРЫ,* слова с «Ё» могут быть определены по их пренадлежности к существительным множественного числа.
7. Наибольшую трудность представляет разрешение омографической неопределённости для слов *ВСЁ_ВСЕ*.

3.1. «ВСЁ» или «ВСЕ»?

Для разрешения омографической неопределённости пары ***ВСЁ_ВСЕ*** можно использовать некоторые эмпирически найденные контекстуальные правила, работающие с достаточно высокой степенью достоверностью. Для этой цели был проведен выборочный анализ встречаемости слов ***ВСЁ*** и ***ВСЕ*** в сочетании с другими словами в романе Б. Акунина «Азазель», содержащего 55 тыс. слов. Было подсчитаны количество сочетаний слова *ВСЁ* с различными словами или знаками препинания при условии, что слово *ВСЕ* ни разу не встретилось в тех же сочетаниях. Получены следующие наиболее частотные сочетания этого вида:

- *ВСЁ+Любой Знак Препинания* — 24 раза
- *ВСЁ+РАВНО* — 21раз
- *ВСЁ+ ЭТО* — 11 раз
- *ВСЁ+ТАК(ТОТ, ТЕМ) ЖЕ* — 9 раз
- *ВСЁ ВРЕМЯ* — 5 раз
- *ВСЁ ЕЩЁ* — 4 раза
- *ВСЁ БЫЛО* — 3 раза
- *ВСЁ МОЖЕТ* — 3 раза.

Определено также около 30 других сочетаний такого рода, встретившихся от 1-го до 2-х раз в проанализированном тексте.

Для более глубокого анализа возможностей разрешения омографической неопределённости пары ***ВСЁ_ВСЕ*** на том же тексте были проведены эксперименты с использованием синтаксического разбора предложений с использованием разработанной в Институте проблем передачи информации РАН системы ЭТАП-3, которая для каждого предложения строит синтаксическую струкутуру в виде дерева зависимостей [4]. На рис. 6 — 8 приведены примеры правильного синтаксического разбора предложения со словом *ВСЁ*. При правильном разборе омограф ***ВСЁ*** маркируется либо как местоимение-существительное (S) единственного числа среднего рода (рис.6), либо как местоимение-прилагательное (A) единственного числа среднего рода (рис. 7), либо как частица (PART), играющая роль ограничителя (рис. 8).

На рис. 9–10 приведены примеры правильного синтаксического разбора предложения со словом ***ВСЕ.*** При правильном разборе омограф ***ВСЕ*** маркируется всегда как местоимение-существительное (A) множественного числа.

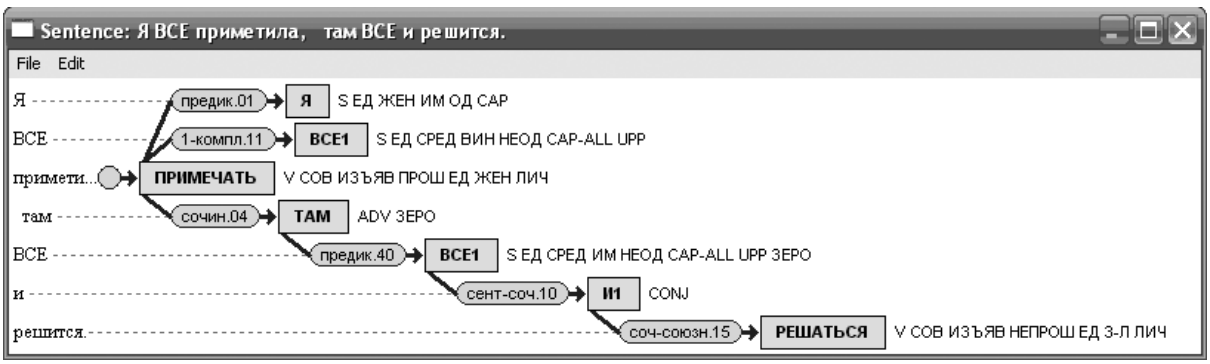


Рис. 6. Пример 1 правильного синтаксического разбора предложения со словом **ВСЁ**

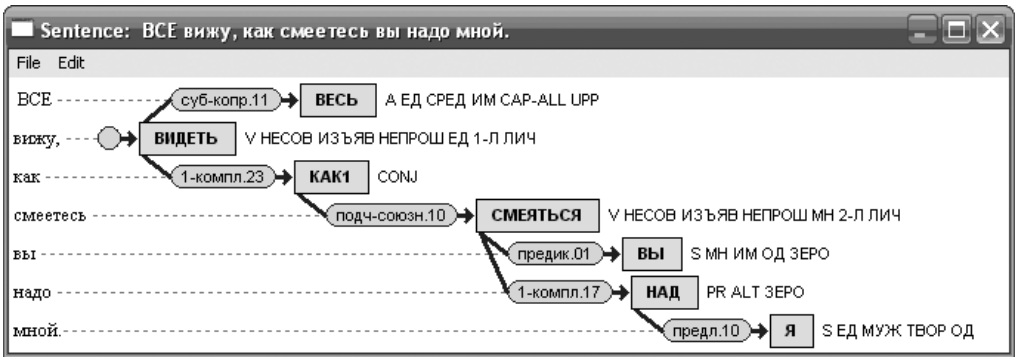


Рис. 7. Пример 2 правильного синтаксического разбора предложения со словом **ВСЁ**

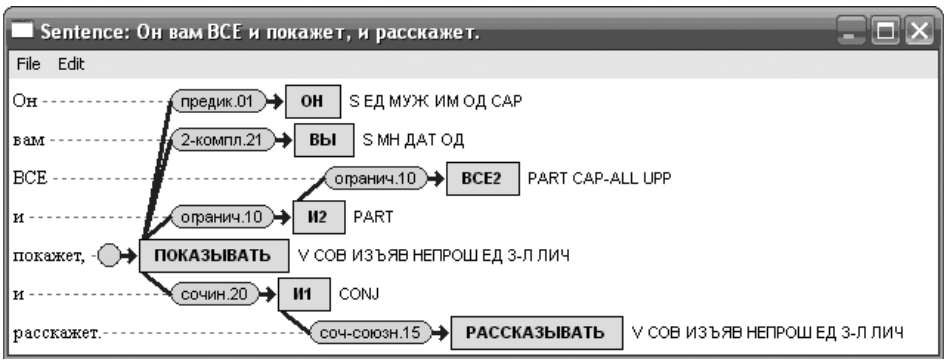


Рис. 8. Пример 3 правильного синтаксического разбора предложения со словом **ВСЁ**

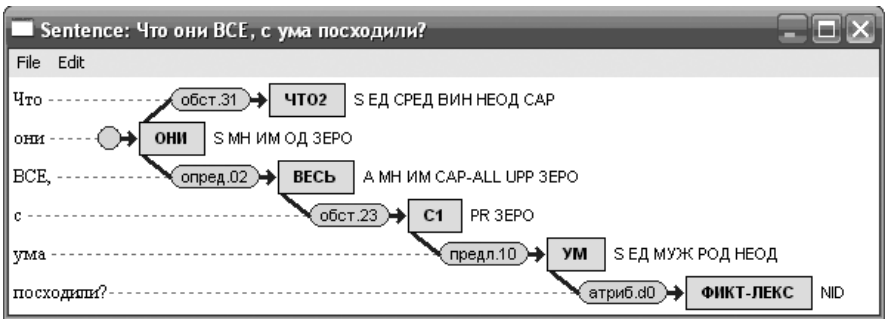


Рис. 9. Пример 1 правильного синтаксического разбора предложения со словом **ВСЕ**

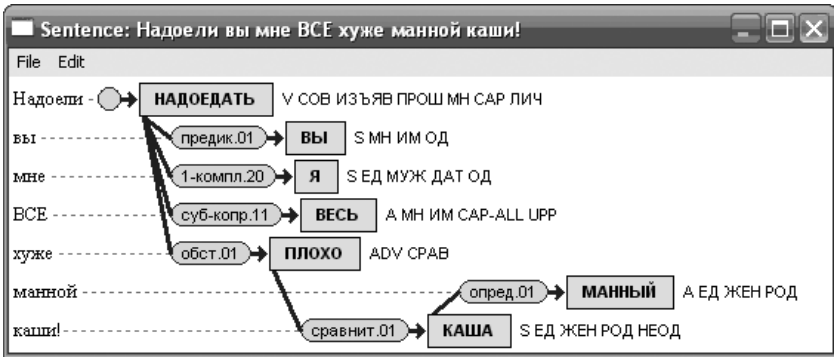


Рис. 10. Пример 2 правильного синтаксического разбора предложения со словом **ВСЕ**

На рис. 11 и 12 приведены примеры неправильного синтаксического разбора предложения со словом *ВСЁ*. В этих примерах слово ***ВСЁ*** ошибочно распознано как ***ВСЕ*** , т.е. как местоимение-прилагательное (рис.11), либо как местоимение-существительное (рис.12) множественного числа.

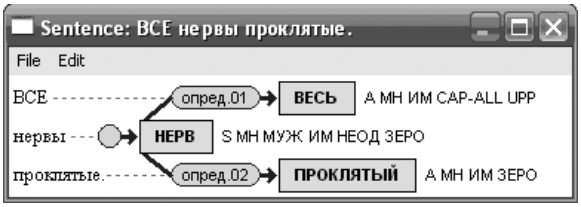


Рис. 11. Пример 1 неправильного синтаксического разбора предложения со словом **ВСЁ**

В заключение заметим, что при использовании системы ЭТАП-3 на всём протестированном тексте (роман Б. Акунина «Азазель»), в котором присутствовало 123 вхождения омографа ***ВСЕ,*** обнаружено лишь 5 ошибочных отнесений слова ***ВСЁ*** к слову ***ВСЕ,*** т.е. только 4% ошибочного распознавания!

Заключение

Однозначного ответа на вопрос, поставленный в качестве эпиграфа к этой статье, пока не существует. Однако, с уверенностью можно сказать, что полное алгоритмическое решение задачи расстановки недостающих точек над «Ё» наступит не ранее, чем в полной мере будут решены проблемы морфологического, синтаксического, семантического и прагматического анализа текстов. Например, как понять: *ВСЁ ДЕРЬМО,* или *ВСЕ ДЕРЬМО?* Система «ЭТАП» говорит, что *ВСЁ*.

В заключение хочу выразить искреннею благодарность ***Елене Ягуновой*** за предоставление словаря омографов [2] и за подсказку использовать в работе Интернет ресурс [3]. И, наконец, но не в последнюю очередь, ***Леониду Иомдину*** за предоставленную мне возможность использования синтаксического анализатора «ЭТАП-3» в ходе выполнения данной работы.

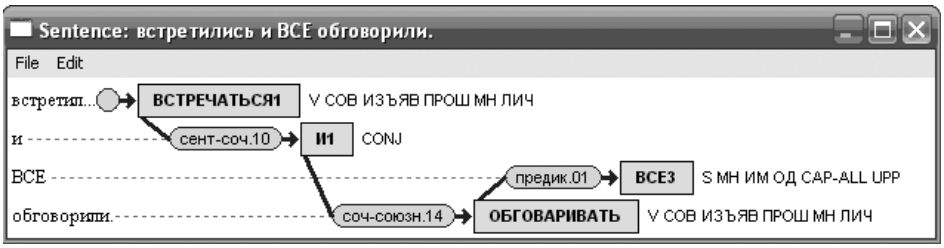


Рис. 12. Пример 2 неправильного синтаксического разбора предложения со словом **ВСЁ**

Литература

1. *Д. Э. Розенталь, М. А. Теленкова.* Словарь-справочник лингвистических терминов // Изд. «Просвещение», М. 1976, 543 с..
2. *А. В. Венцов* и др. Словарь омографов русского языка // Изд. СПбГУ, Санкт-Петербург, 2004, 160 с.
3. *Национальный* корпус русского языка “Поиск по акцентуированному корпусу” // Интернет ресурс: http://www.narusco.ru
4. *И. М. Богуславский, Л. Л. Иомдин, Д. Р. Валеев, В. Г. Сизов.* Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов // Труды Международной конференции <Корпусная лингвистика — 2008>. СПб.: Санкт-Петербургский государственный университет, 2008. С. 56–74.

Автоматическое аннотирование новостных кластеров на основе тематического представления

Summarization of news clusters based on thematic representation

Лукашевич Н. В. (louk@mail.cir.ru), **Добров Б. В.** (dobroff@mai.cir.ru)

Научно-исследовательский вычислительный центр МГУ им. М. В. Ломоносова; АНО Центр информационных исследований

Представлен метод автоматического построения аннотации для новостного кластера на основе тематического представления кластера, моделировании лексической связности текста и тезаурусном описании лексических значений, что позволяет улучшить связность и полноту аннотации, а также снизить повторы.

1. Введение

Современные технологии обработки новостных потоков обычно включают в себя краткое представление содержания новостного кластера в виде аннотации (обзорного реферата). Так как в широких предметных областях результаты по автоматической генерации текста не являются устойчивыми, то стандартным способом аннотирования кластера является составление обзорного реферата, включающего заголовок и несколько предложений, обычно извлеченных из разных текстов кластера [11].

В настоящее время предложено достаточно большое количество методов автоматического обзорного реферирования [7]. Существенными проблемами при составлении аннотации новостного кластера являются [1, 7, 11]:

- обеспечение полноты представления информации, в том числе наиболее свежей информации,
- снижение повторов при представлении информации,
- обеспечение связности и понятности представляемой информации.

Для определения избыточности в порождаемых аннотациях используются различные меры сходства между предложениями. Одним из распространенных подходов является предварительная кластеризация — выделение близких по содержанию кластеров предложений [10]. Другим подходом для уменьшения избыточности является сравнение предложений-кандидатов с предложениями, уже попавшими в аннотацию, и оценка новой (непохожей) информации, например, подход Maximal Marginal Relevance (MMR) [6].

Для повышения связности аннотации и снижения повторов в работах [5, 8, 12] предлагается использовать построение аннотаций на основе так называемых лексических цепочек, то есть групп слов текста, связанных между собой по смыслу. Связность аннотации повышается на основе моделирования лексической связности текста. Суть подхода заключается в том, что повтор устраняется за счет оперирования не отдельными словами, а набором близких по смыслу слов. Основным источником информации о взаимосвязи слов при анализе англоязычных текстов служит обычно тезаурус WordNet.

В нашей работе [3] 1997 года был предложен подход к построению аннотаций отдельных документов на основе лексических цепочек специального вида — тематических узлов [4]. В качестве лексической базы для создания тематических узлов использовался двуязычный Информационно-поисковый тезаурус для автоматического концептуального индексирования по общественно-политической тематике (далее Тезаурус) [2].

Данная статья посвящена описанию расширения исходного подхода по автоматическому аннотированию одного документа на случай создания аннотации для новостного кластера.

2. Построение тематического представления текста

Основными критериями для построения лексических цепочек в большинстве подходов являются следующие [5, 8, 12]:

- наличие и сила связей между лексемами, описанных в некотором ресурсе,
- расстояние между вхождениями лексем в тексте, измеряемое обычно в предложениях. Если расстояние от текущего слова до предшествующих вхождений лексической цепочки больше некоторого порога, то лексическая цепочка прерывается и начинается новая,
- лексические цепочки строятся по порядку появления слов в тексте,
- для оценки важности лексической цепочки используется ряд параметров, таких как длина цепочки, плотность и др.

В нашем подходе мы обосновываем (подробнее см. [2, 3, 4]), что для того, чтобы лексическая цепочка соответствовала тематической структуре текста, нужно выполнение следующих условий:

- лексическая цепочка должна иметь внутреннюю структуру узла — к одному выделенному элементу относятся все другие элементы лексической цепочки,
- лексические цепочки в виде тематических узлов строятся на основе обработки целого текста, анализа частотностей концептов тезауруса, системы отношений между ними,
- значимость тематического узла для отражения содержания текста определяется не столько длиной, покрытием и другими характеристиками цепочки, а тем, насколько часто элементы этого тематического узла встречались с элементами других тематических узлов в одних и тех же предложениях текста, то есть насколько много пропозиций конкретных предложений текста было посвящено обсуждению отношений между элементами выявленных тематических узлов.

В совокупности построенные тематические узлы составляют тематическое представление текста.

В качестве примера рассмотрим автоматически создаваемое тематическое представление для следующего текста (представлен начальный фрагмент):

Китай и Тайвань установили авиасообщение после 60-летнего перерыва

После почти 60-летнего перерыва открылось регулярное авиасообщение между Тайванем и материковым Китаем. Первый чартерный рейс с 250 пассажирами уже прибыл в столицу Тайваня из китайского города Гуанчжоу, передает «Би-би-си». Ожидается, что аэропорты острова будут принимать рейсы из пяти китайских городов: Пекина, Шанхая, Гуанчжоу, Сямэня и Нанкина. Договоренность о прямых регулярных авиарейсах была достигнута в середине июня 2008 года на переговорах между руководством Тайваня и Китая. Восстановление авиасообщения произошло не в последнюю очередь благодаря победе на выборах главы администрации Тайваня в марте 2008 года сторонников тесного сотрудничества с материковым Китаем…

Приведем примеры тематических узлов, созданных в процессе обработки этого текста (главное понятие тематического узла выделено сдвигом влево; указана частота употребления понятия в тексте):

КИТАЙ	8
ГАНЧЖОУ	2
ШАНХАЙ	2
НАНКИН	1
ПЕКИН	1
ТАЙВАНЬ	7
ТАЙБЕЙ	1
АВИАЦИОННЫЕ ПЕРЕВОЗКИ	*2*
АВИАРЕЙС	1
АЭРОПОРТ	1
ПОЛИТИЧЕСКАЯ ПАРТИЯ	1
КОММУНИСТ	*1*
ПРАВИТЕЛЬСТВО	1
ПУБЛИЧНАЯ ВЛАСТЬ	1

Для многих приложений важно определение основной темы документа, которая моделируется совокупностью основных тематических узлов. Если рассматривать основную тему документа как пропозицию, то каждый основной тематический узел соответствует отдельному тематическому элементу, входящему в состав этой пропозиции.

В нашей модели мы предполагаем, что понятия основных тематических узлов постоянно встречаются рядом друг с другом (связаны по тексту). Для оценки совместной встречаемости понятий (текстовых связей понятий) используется линейное окно понятий длиной 3.

После того как созданы тематические узлы, текстовые связи понятий каждого тематического узла суммируются и определяются текстовые связи между тематическими узлами.

Основными тематическими узлами, которые соответствуют основной теме документа, в первую очередь, являются такие тематические узлы, которые:

- все связаны между собой текстовыми связями;
- сумма частот текстовых связей между ними максимальна.

В рассматриваемом примере тематического представления основными тематическими узлами стали узлы с главными понятиями: *КИТАЙ*, *ТАЙВАНЬ, АВИАЦИОННЫЕ ПЕРЕВОЗКИ, ГОРОД, РЕЙС*

3. Использование тематического представления для составление аннотации одного текста

Для составления аннотации отдельного документа используются два принципа.

Во-первых, важными (информативными) и, следовательно, возможно включенными в аннотацию считаются те предложения текста, которые содержат по крайней мере два понятия, входящих в состав разных основных тематических узлов текста.

Во-вторых, для каждой пары выявленных основных тематических элементов текста в аннотацию выбираются предложения, содержащие первое вхождение этой пары, следуя по порядку текста.

Не все основные темы начинают обсуждаться в тексте сразу, с первого предложения -- часть из них возникает в последующих предложениях. Чтобы сохранить связность и последовательность изложения текста автор именно в этом первом предложении новой темы должен наиболее точно указать связь новой темы со всем предшествующим текстом. Таким образом, следуя за автором вводе нового тематического элемента, можно повысить общую связность аннотации.

Нужно отметить, что при хорошем покрытии предметной области Тезаурусом появление в очередном предложении новой темы выявляется весьма точно, а это означает, что связность получаемой аннотации в среднем весьма высока.

Построение аннотации реализуется следующим образом:

1) Для построения аннотаций сначала формируется множество «аннотационных» фрагментов, которые являются целыми предложениями исходного текста, содержат в своем составе глагол в изъявительном наклонении или краткое прилагательное, и не являются вопросительными или восклицательными предложениями.
2) Перед построением аннотации создается таблица всех возможных пар основных тематических узлов.
3) Начиная с начала текста, отбираются такие предложения, которые содержат еще не упоминавшуюся пару разных тематических узлов.

Таким образом для текста примера получаем следующую аннотацию, в которой упомянуты все основные тематические узлы данного документа:

Китай и Тайвань установили авиасообщение после 60-летнего перерыва.

Первый чартерный рейс с 250 пассажирами уже прибыл в столицу Тайваня из китайского города Гуанчжоу, передает Би-би-си. Ожидается, что аэропорты острова будут принимать рейсы из пяти китайских городов Пекина, Шанхая, Гуанчжоу, Сямэня и Нанкина.

Отметим, что в аннотации пропущено первое предложение, которое не содержит новой пары тематических узлов по сравнению с заголовком текста.

4. Построение тематического представления для новостного кластера

Новостной кластер представляет собой совокупность тематически близких документов.

Поэтому тематическую структуру новостного кластера так же, как и отдельного элемента можно выявить за счет построения тематического представления этого кластера, и это представление можно будет использовать для управления набором предложений в аннотацию кластера, а именно для обеспечения полноты, снижения повторов, а также обеспечения связности аннотации кластера.

Построение тематического представления новостного кластера осуществляется простым способом: все тексты кластера склеиваются в единый текст, для которого производится стандартный тематический анализ одного документа и строится тематическое представление.

Результат этой процедуры, а затем и результат построения аннотации в некоторой степени зависит от порядка просмотра документов в кластере. Мы используем следующий метод объединения документов кластера в единый текст, используемый для построения аннотации.

Сначала в новостном кластере определяется «центр кластера» — документ, наиболее близкий к центру тяжести множества документов кластеров в метрическом пространстве нормализованных лемматическом и концептуальном (по тезаурусу) индексов. Определяется «ядро» кластера — документы достаточно близкие к центру (по некоторому порогу). Затем «центр кластера» сдвигается в документ из ядра кластера, который был опубликован последним по времени. Пересчитываются веса связей документов кластера к новому центру. С учетом задаваемого интервала времени по убыванию веса сначала заполняются документы за последнее время, затем все остальные. Так как отбирается всего несколько предложений, то имеется общее ограничение на количество отбираемых в «единый документ» документов.

Приведем примеры тематических узлов для кластера, в который входит текст примера (справа указана частотность концепта в кластере):

КИТАЙ	*103*
ПЕКИН	21
ГУАНЧЖОУ	13
ГОСУДАРСТВО	9
ЮАНЬ	7
ШАНХАЙ	6
КИТАЙЦЫ	5
НАНКИН	5
ГУАНДУН	1
АВИАЦИОННЫЕ ПЕРЕВОЗКИ	*33*
АВИАЦИОННАЯ КОМПАНИЯ	9
САМОЛЕТ	9
АВИАРЕЙС	7
ТРАНСПОРТНАЯ СФЕРА	4
АЭРОПОРТ	3
ТРАНСПОРТНЫЕ ПЕРЕВОЗКИ	2
АЭРОБУС	1
АВИАЛИНИЯ	1
ТУРИСТ	*12*
ЧЕЛОВЕК	62
ТУРИЗМ	2
ПОЕЗДКА	1
ПРАВИТЕЛЬСТВО	*6*
РУКОВОДИТЕЛЬ	6
ОРГАН ПУБЛИЧНОЙ ВЛАСТИ	3
РУКОВОДСТВО	2
ОРГАН ИСПОЛНИТЕЛЬНОЙ ВЛАСТИ	2
ПУБЛИЧНАЯ ВЛАСТЬ	1

Таким образом, по основным тематическим узлам тематического представления могут быть определены основные элементы, обсуждаемой в кластере темы.

Как видно, тематические узлы включают понятия достаточно разной частотности.

Низкочастотные концепты тематического узла могут быть ошибочно включены в тематический узел, кроме того, представительность ими основной темы документа невелика. Поэтому можно задать выделение ядра тематических узлов, которое определяется как коэффициент от 0 до 1. Этот коэффициент определяет, какая доля от общей частотности тематических узлов будет включена в ядро. Для построения аннотации кластера используются только ядра основных тематических узлов (используется коэффициент 0,7).

5. Построение аннотации новостного кластера

Аннотация новостного кластера состоит из заголовка и нескольких предложений из разных документов новостного кластера.

Зная ядра тематических узлов, полноту изложения содержания кластера мы обеспечиваем тем, что должны отбирать для аннотации предложения, содержащие пары этих тематических узлов — именно тогда эти предложения будут описывать взаимоотношения между основными тематическими элементами кластера.

При отборе заголовка для аннотации ищется заголовок, содержащий пару наиболее частотных тематических узлов. Если таких заголовков нет, то ищутся заголовки, содержащие понятия из одного наиболее частотного тематического узла.

Для выбора очередного предложения в списке основных тематических узлов отмечаются все тематические узлы, которые уже были упомянуты. Очередное предложение должно содержать пару основных тематических узлов.

Для обеспечения связности требуется, чтобы очередное предложение содержало либо уже упомянутый тематический узел, либо уже упоминавшееся слово с большой буквы.

Кроме того, делается ряд дополнительных проверок:

- предложение не должно являться вопросительным или отрицательным предложением;
- предложение не должно содержать в заданном числе первых слов местоимение;
- начало предложения не должно совпадать с началами заголовка и предложений, уже взятых в аннотацию;
- число слов предложения, совпадающего со словами предшествующих предложений, не должно превышать некоторой доли длины предложения.

Понятно, что даже при проверке вышеупомянутых условий может найтись еще достаточно много подходящих предложений-кандидатов. Кроме того, оценка предложений на основе понятий тезауруса не является достаточной без учета упоминаемых именованных сущностей, которые могут быть и не описаны в тезаурусе.

Поэтому вводится еще и общая оценка предложения с помощью вычисления веса предложения, которая складывается из двух компонентов: весов упомянутых понятий Тезауруса, которые были получены в тематическом представлении [2], а также весов содержащихся в предложении слов с большой буквы, не считая первого слова предложения.

Для вычисления весов слов с большой буквы (далее Слов), сначала вычисляется вес самого частотного Слова W_{max_word} в документе кластера:

$$W_{max\_word} = \min (1, W_{max\_conc} * (Fr_{max\_word} / Fr_{max\_conc}))$$

где W_{max_conc} — максимальный вес понятия тезауруса в тематическом представлении, Fr_{max_conc} — частотность в тексте понятия тезауруса с максимальным весом, Fr_{max_word} — частотность самого частотного Слова.

Остальные веса Слов (W_{word}) вычисляются пропорционально их частотности:

$$W_{word} = W_{max\_word} * (Fr_{word} / Fr_{max\_word})$$

Просмотр предложений-кандидатов начинается с начала документа кластера, то есть предложения набираются сначала из главного документа кластера и наиболее близких к нему по содержанию. Каждое следующее предложение берется из другого документа.

Для кластера примера была получена следующая аннотация (в скобках указан источник новости и время публикации):

Предложения	Тематические узлы
Китай и Тайвань установили авиасообщение после 60-летнего перерыва *(Новые Известия — лента новостей, 04.07.2008 11:08:45)*	*КИТАЙ, ТАЙВАНЬ, АВИАЦИОННЫЕ ПЕРЕВОЗКИ (авиасообщение)*
Первый чартерный рейс с 250 пассажирами уже прибыл в столицу Тайваня из китайского города Гуанчжоу. (Lenta.ru — главные новости, 04.07.2008 9:47:25)	*КИТАЙ, ЧАРТЕРНЫЕ ПЕРЕВОЗКИ (чартерный рейс), ГОРОД, ПАССАЖИР*
С 4 июля самолеты с материкового Китая на остров Тайвань и обратно будут летать каждую неделю с пятницы по понедельник. (РегKURSCITY.RU — Курс, 04.07.2008 9:35:34)	*КИТАЙ, ТАЙВАНЬ, АВИАЦИОННЫЕ ПЕРЕВОЗКИ (самолет), ОСТРОВ*
Перед прибывающими в ближайшие выходные 600 туристами из Китая будет расстилаться красная ковровая дорожка. (BBCRussian.com (Главная), 04.07.2008 1:18:25)	*КИТАЙ, ТУРИСТ*
По завершении в 1949 году гражданской войны в Китае и изгнания правительства Гом-Инь-Дана на Тайвань, отношения между двумя сторонами Тайваньского пролива были заморожены. (РегЛІГАБізнесІнформ — Украины — Новости за рубежом, 04.07.2008 9:14:00)	*КИТАЙ, ТАЙВАНЬ, ПРАВИТЕЛЬСТВО*

Каждое предложение аннотации содержит не менее двух понятий из разных основных тематических узлов, как минимум один из которых новый (выделен подчеркиванием в правом столбце таблицы), а другие были упомянуты ранее.

6. Оценка качества аннотаций новостных кластеров

Оценка качества аннотаций может быть проведена вручную или с помощью автоматических процедур [7].

При ручной оценке относительно аннотации могут быть заданы такие вопросы с оценкой по 5 бальной шкале:

- является ли предложения аннотации грамматически правильными,
- является ли текст аннотации связным,
- содержит ли аннотация все основные обсуждаемые темы исходного документа (документов) и др.

Однако, проведение ручной оценки весьма трудоемко и требует привлечения независимых экспертов.

Для автоматической оценки качества аннотаций может использоваться метод ROUGE (Recall Oriented Understudy for Gisting Evaluation), который подсчитывает число перекрытия (n-граммы слов) автоматической аннотации с «идеальными» аннотациями, составленными людьми [9].

6.1. Оценка качества аннотаций отдельных документов

Качество описываемой технологии автоматического аннотирования англоязычных текстов тестировалось на конференции SUMMAC (Summarization conference), организованной DARPA в 1998 году. Программа использовала английский перевод Общественно-политического тезауруса.

Суть соревнования заключалась в следующем: каждый участник соревнования получил на две недели 1000 документов и представил две аннотации — аннотацию наилучшей длины (то есть система сама определяла длину аннотации) и 10-процентную аннотацию, т.е. аннотацию, составляющую 10 процентов длины исходного текста.

Описываемая технология имела лучший показатель для аннотаций наилучшей длины и показатели на 10-процентную аннотацию были лучше, чем средние [13].

6.2. Оценка качества аннотаций новостных кластеров

В качестве метрики для оценки мы использовали метрики ROUGE-1-cir и ROUGE-2-cir, которые вычисляли следующим образом:

$$ROUGE - N - cir(A_i) = \frac{\sum_{M_{ij}} count(Ngram(A_i) \cap Ngram(M_{ij}))}{\sum_{M_{ij}} count(Ngram(M_{ij}))},$$

где A_i — оцениваемая обзорная аннотация i-того кластера, M_{ij} — ручные аннотации i-того кластера,

Ngram(D) — множество всех n-грамм из лемм соответствующего документа *D*. При сравнении отдельных документов в расчет берутся только уникальные n-граммы, присутствующие в обоих документах — не поощряется многократный повтор одного и того же предложения. При рассмотрении нескольких аннотаций, наоборот, повторение одинаковых элементов поощряется. Биграммы в наших оценках учитывались с перестановками.

Для оценки качества построенных аннотаций мы воспользовались данными, любезно предоставленными С.Д. Тарасовым (Военмех, Спб.). В проведенных С.Д. Тарасовым экспериментах группе студентов было предложено построить ручную аннотацию для новостных кластеров, которые брались из системы Google.Новости в период с 01 по 05 декабря 2008 года. Ручная аннотация должна была быть составлена из четырех предложений. Ограничений на выбор предложений из разных текстов не накладывалось.

Мы выбрали достаточно случайным образом из полученных данных 15 новостных кластеров разной тематики, включая новости о погоде, спорте, финансах и политике, для которых имелось от 18 до 40 ручных аннотаций (всего 462).

В качестве «базовой оценки», следуя [7], мы рассматривали следующие варианты искусственных аннотаций:

- первый документ кластера;
- заголовки первых четырех документов;
- первые предложения первых четырех документов;
- последний документ кластера.

В качестве автоматической аннотации рассматривались аннотации из заголовка и трех предложений, взятых из разных текстов.

Мы получили следующие результаты (в таблице приведены результаты для разных параметров ядра кластера — см.раздел 5):

Следует отметить, что некоторые ручные аннотации совпадали с первым или последним документом кластера. Определенным недостатком используемых данных является то, что некоторые кластеры содержали документы за несколько дней, поэтому ручные аннотации чаще содержали предложения из последних документов кластера.

Существует определенная критика использования метрик ROUGE для оценки качества аннотирования. Метрика чувствительна к длинам сравниваемых документов, не учитывает связность аннотаций. В целом, существует большое разнобразие между ручными аннотациями разных экспертов. В нашем случае нам лишь важно было оценить близость построенных автоматических и ручных аннотаций для оценки перспективности предложенного подхода.

Заключение

В статье представлена технология автоматического построения аннотации для новостного кластера, которая использует свойство лексической связности текста и знания о лексических значениях, которые описаны в Общественно-политическом тезаурусе, для обеспечения связности, полноты и снижения повторов аннотации кластера.

Описанный метод построения обзорных рефератов позволяет в широких пределах варьировать представление кластера при сохранении уровня отображения содержания и связности. Можно задавать как количество документов (исходящих ссылок), отражаемых в аннотации, так и количество предложений из каждого документа. В частности, могут быть смоделированы аннотации, формируемые в ресурсе Яндекс.Новости (до трех-четырех документов по одному-два предложения), или аннотации, формируемые в ресурсе Google.Новости (три-четыре предложения из одного документа и два заголовка из других документов), или Рамблер.Новости (три предложения из одного документа и два-три предложения из других документов).

Вид аннотации	ROUGE-1-cir	ROUGE-2-cir
первый документ кластера	0,219	0,083
заголовки первых четырех документов	0,162	0,056
первые предложения первых 4 документов	0,269	0,107
последний документ кластера	0,278	0,168
автоматическая аннотация с ядром 0,20	0,331	0,150
автоматическая аннотация с ядром 0,40	0,328	0,140

Литература

1. *Абрамова Н. Н, Абрамов В. Е.* Автоматическое составление обзорных рефератов новостных сюжетов // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2007. Переславль-Залесский: 2007.
2. *Добров Б. В., Лукашевич Н. В.* Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ // Третья Всероссийская конференция по Электронным Библиотекам «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Петрозаводск: 2001. С. 78–82.
3. *Лукашевич Н. В.* Автоматическое построение аннотаций на основе тематического представления текста // Труды международного семинара Диалог'97. М.: 1997 — С. 188–191.
4. *Лукашевич Н. В., Добров Б. В.* Исследования тематической структуры текста на основе большого лингвистического ресурса // Труды международного семинара «Диалог 2000». М.: 2000. Том 2, С.252–258.
5. *Barzilay R., Elhadad M.* Using Lexical Chains for Text Summarization // ACL/ EACL Workshop Intelligent Scalable Text Summarization. Madrid: 1997.
6. *Carbonell J., Goldstein J.* The use of MMR, diversity-based reranking for reordering documents and producing summaries // In Proceedings of the 21st Annual International ACM SIGIR Conference. 1998. pp. 335–336.
7. *Dang H. T.* National Institute of Standards and Technology (NIST) «Overview of DUC 2006» // In Proceedings of DUC 2006.
8. *Doran W., Stokes N., Dunnion J., Carthy J.* Comparing Lexical Chain-based Summarisation Approaches using an Extrinsic Evaluation // In the Proceedings of the Global WordNet Conference(GWC 2004). Brno: 2004.
9. *Lin Chin-Yew.* ROUGE: a Package for Automatic Evaluation of Summaries // In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). Barcelona: 2004.
10. *Radev D., Jing H., Budzikowska M.* Centroid-based summarization of multiple docuemtns: sentence extraction, utility-based evaluation, and user studies // In ANLP/NACCL Workshop on Summarization. Seattle: 2000.
11. *Radev D., McKeown K., Hovy E.* Introduction to the Special Issue on Summarization // Computational linguistics. 2002. 28 (4). P.399–408.
12. *Silber G., McCoy K.* Efficiently computed lexical chains as an intermediate representation for automatic text summarization //. Computational Linguistics.2002. 28 (4).P.487–496.
13. *Tipster* SUMMAC Text Summarization Evaluation. Final report. — MITRE Technical report MTR 98000138. — October, 1998.

Русский фреймнет: к задаче создания корпусного словаря конструкций

Russian framenet: towards a corpus-based dictionary of constructions

Ляшевская О. Н. (olesar@mail.ru),
Кузнецова Ю. Л. (julia.kuznetsova@uit.no)

University of Tromsø, Тромсе, Норвегия

В российской компьютерной лингвистике до сих пор нет ресурсов, аналогичных FrameNet, VerbNet и PropBank, в которых можно было бы получить иллюстрации глагольного и именного управления, а также периферийных лексически ориентированных конструкций. В работе описываются задачи создания и структура FrameNet-ориентированной системы, предназначенной для исследования морфологических, синтаксических, семантических и лексических ограничений в русских конструкциях.

1. Введение

В последнее время благодаря появившимся поисковым системам (таким как Яндекс и Google) и корпусам текстов с возможностью поиска по ним (в частности, НКРЯ (http://ruscorpora.ru) и многим другим) возможности лингвиста значительно расширились. В лингвистике теперь стало можно ставить такие задачи, которые прежде решить было невозможно или для исследования которых требовалось необозримое количество времени и ресурсов. Например, теперь легко и быстро можно узнать, сочетается ли некоторый глагол с некоторым объектом, и с каким из двух объектов он встречается чаще, чем с другим.

Вместе с тем, при решении ряда задач наши поисковые возможности ограничены. С одной стороны, в запросах к поисковым системам и к большинству корпусов мы можем задавать только линейный порядок слов, поэтому лингвисту, исследующему конструкции, приходится перебирать все возможные комбинации элементов. С другой стороны, зачастую не хватает информации о синтаксической связи слов. Например, сложности возникают при поиске примеров конкретного варианта управления глагола или примеров конкретной конструкции (такой как конструкция с квазиимперативом долженствования *Они едят в ресторанах, а я плати*). Как правило, в результатах поиска присутствует столько «шума», что требуется множество дополнительных усилий для того, чтобы отделить нужные примеры от примеров, случайно попавших в выданные результаты.

По-видимому, лучше всего указанную задачу решает синтаксически размеченный корпус НКРЯ (http://www.ruscorpora.ru/search-syntax.html), в котором можно искать синтаксические зависимые, в том числе без учета порядка слов, но пока что этот ресурс не очень представителен в связи с малым размером. Кроме того, при создании этого корпуса не ставилась задача обеспечить достаточную выборку примеров для отдельных лексических единиц; в частности, даже частотные русские глаголы в нем представлены неравномерно. Например, обнаружив в синтаксическом корпусе, что в четырех из пяти имеющихся примеров на глагол *гнать* этот глагол употреблен с прямым объектом, мы не можем сделать вывод ни о том, что это отражение наиболее частотной модели, ни что эта модель встречается в среднем в 80% случаев.

В российской компьютерной лингвистике до сих пор нет специализированных ресурсов, аналогичных FrameNet (Johnson, Fillmore et al. EE), VerbNet (Kipper et al. 2006) или PropBank (Palmer 2005). В этих источниках исследователи английского языка могут получить данные о типах глагольного управления, их распределении и вариативности в разных лексических единицах, просмотреть иллюстративный материал – причем на примерах из реальных текстов. Система FrameNet, кроме того, содержит аналогичные данные об именах существительных и прилагательных, и, что примечательно, в настоящее время эволюционирует в сторону словаря конструкций («New Constructicon», см. Fillmore 2008).

Таким образом, речь идет о создании русского фреймнет-ориентированного ресурса, спроектированного с учетом традиций отечественной лексической семантики и специфики русского языка, где информация о предложно-падежной реали-

зации управления предикатов и поверхностно-синтаксических свойствах других конструкций имеет особую ценность. Эта компьютерная система должна решать не только задачи словаря (ср. систему «Лексикограф», www.lexicograph.ru, бумажные словари Апресян, Палл 1982, Сазонова 2008, лексикографические проекты Азарова и др. 2004, Апресян 2008 и др.), но и представлять аннотированный корпусной материал. Этот ресурс реализует принципы гибридных систем, в которых авторитетные лексикографы видят будущее словарей (Atkins 1992, Kilgarriff et al. 2006): словарь в выходом в корпус. С одной стороны, это словарь – но с возможностью расширения иллюстраций за счет поиска в корпусе. С другой стороны, это корпус, но с экспертным отбором примеров. Наконец, это лексически ориентированный ресурс, где выборка примеров строится с ориентацией на конкретные лексемы, однако в каждом предложении разметке подлежат все предикатные слова и связанные с ними конструкции.

В результате появляется возможность проследить, как реализуются активные и пассивные валентности глаголов, реляционных имен типа *ненависть* или *пациент*, прилагательных (ср. *готов к выступлению*), наречий и т. п., как влияют на это «нелексические» грамматические конструкции, например, инфинитивная или компаративная, как это связано с порядком слов, пунктуацией и так далее. Тем самым, создаются предпосылки для прорыва в еще одной важной и малоисследованной области лингвистики – в изучении взаимодействия различных конструкций на пространстве текста.

Технологию создания и структуру такого корпусного словаря конструкций и описывает данная работа.[1]

2. Формирование банка предложений

В первую очередь, необходимо собрать представительный массив данных для исследований управления и сферы действия предикатов (глаголы, имена существительные, имена прилагательные, неизменяемые части речи) и грамматики конструкций. Наиболее тщательно описанная область — глаголы, с них мы и предполагаем начать сбор данных. В дальнейшем банк примеров будет пополняться с целью сбалансированного представления имен существительных, а также прилагательных с предложным управлением. Наименее полно описаны конструкции «вне управления»: в настоящее время нет даже их сколько-нибудь полного инвентаря. Мы собираемся выявить и разметить такие конструкции в собранных примерах, а затем, на последнем этапе, целенаправленно собрать материалы для составленного «словника» конструкций.

Таким образом, формирование банка предложений будет проходить в три стадии: «глаголы» — «имена» — «малый синтаксис» (прочие конструкции). Ниже мы опишем процедуру сбора данных для первой стадии.

Список «целевых» глаголов (1000 единиц) формируется по признаку их частотности, а также разнообразия и вариативности управления. На каждую лексическую единицу должно быть собрано 100 предложений НКРЯ (в дальнейшем, с учетом собранных данных, выборка может быть доведена до 200 примеров). Выборка должна быть сбалансирована по следующим показателям:

1) метатекстовые признаки:
   - время создания текста (после 1950 г., первая половина XX в., XIX в., XVIII в.);
   - (в идеале) не более 1 примера из 1 автора;
   - жанр (художественная литература, публицистика, прочая нехудожественная литература, устная речь);
2) характеристики предложения:
   - длина;
   - сложность (простое, в т. ч. осложненное знаками препинания, сложносочиненное, сложноподчиненное, парцелляты);
   - место предложения в тексте;
3) место предиката в предложении:
   - начало — середина — конец.

Для особо частотных глаголов выборка может быть дополнительно сбалансирована относительно контекстных маркеров, а именно, слов открытых лексических классов, которые наиболее часто встречаются в предложении с данным глаголом, ср. *речь идет, подписать… договор* и т. д. Если их частотность в выборке будет не слишком велика, это послужит некоторой гарантией тематического разнообразия иллюстраций.

Банк предложений будет включать отобранные предложения и их ближайший контекст (3 предложения справа и слева от них в тексте).

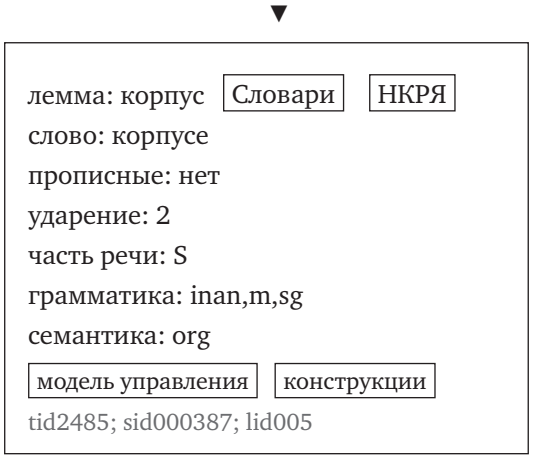
3. Разметка предложения

Каждое предложение проходит несколько этапов разметки. Во-первых, в банк данных заносится информация о ближайшем контексте (ее можно просмотреть, нажав на знак ←…→, см. пример ниже). Во-вторых, на основании мета-текстовых данных НКРЯ заполняется паспорт текста (⌂ автор, его пол и год рождения, время создания текста, жан-

[1] Проект выполняется в рамках программы фундаментальных исследований ОИФН РАН (2009–2011 гг.), направление «Лингвистические аспекты исследования текста».

рововая принадлежность). В-третьих, размечаются характеристики предложения: длина, сложность, место в тексте (начало, середина, конец), состав и основные характеристики лексических элементов и пунктуационных знаков.

(1) Он служил в конном корпусе Гая.



[Анатолий Рыбаков. Тяжелый песок (1975–1977)] ←...→ ⌂

Рис. 1. Пример предложения и паспорт слова корпус

Каждое слово в предложении также получает свой паспорт: поля паспорта заполняются автоматически, в частности, с использованием данных лексико-грамматической и лексико-синтаксической разметки НКРЯ. В дальнейшем эта информация может пополняться и редактироваться. Предусмотрена возможность посмотреть информацию о слове в словарях МАС и Ожегова-Шведовой (интернет-версии), а также получить справку об употреблении слова в НКРЯ (основной и синтаксический корпус).

Последний этап разметки предложения — определение элементов, которым будут должны быть приписаны модели управления (МУ) и словоцентричные конструкции. В примере (1) МУ приписываются глаголу *служить,* существительному *корпус*, прилагательному *конный* и предлогу *в*. Из словоцентричных конструкций следует отметить устойчивый оборот *конный корпус*.

4. Модели управления: словарь

Информация о моделях управления содержится в двух частях системы: в словаре (основные МУ) и при конкретных лексических элементах в предложении (реализация МУ в тексте). Словник словаря МУ имеет две версии: целевой (все единицы, для которых целенаправленно собирались данные) и полный (все предикаты, которые встретились в предложениях). Словарь также делится на частеречные разделы: глаголы, абстрактные и предметные имена, прилагательные, наречия, предлоги и прочее.

Пилотная версия глагольной части словаря (инвентарь МУ и их характеристики) базируется на работе Апресян, Палл 1982, в котором собрана информация о более чем 1300 глаголах. В дальнейшем список моделей управления может быть сокращен или расширен с учетом информации из других источников («Лексикограф», толковые словари, ТКС, НОСС, синтаксический корпус НКРЯ, RussNet, материалы проекта дизамбигуации глагольных значений (Толдова 2008) и др.), а также новых эмпирических данных. По мере необходимости будут формироваться словарные входы имен существительных и других частей речи.

В пределах словарной статьи все модели управления получают условное название (например, «служить в министерстве», «служить царю» и т.д.) и связаны между собой отношениями «синтаксической деривации» (Падучева 2004). При каждой МУ указывается число примеров, ассоциированных в банке предложений с данной моделью. Просмотреть эти примеры можно, перейдя по гиперссылке. Таким образом, мы получаем иерархию МУ от основных к производным и далее к частным МУ в примерах.

Способ представления МУ в словаре заимствован нами в основных чертах из системы «Лексикограф» (Падучева 2004). МУ включает в себя следующие сведения: сокращенное обозначение участника, стандартный способ морфосинтаксического выражения, синтаксический ранг, экспликация участника, семантические ограничения на заполнение валентности (см. рис. 2).[2]

«служить в министерстве» (примеров: 4) ▶				
Имя	Экспликация	Морфосинтаксис	Ранг	Семантические ограничения
X	тот, кто служит	NPnom	Субъект	hum ▶
Y	место, где служат	в + NPloc	Периферия	org ▶

Рис. 2. Фрагмент статьи глагола служить, МУ «служить в министерстве»

[2] Здесь и далее используются сокращенные обозначения грамматических категорий и лексико-семантических признаков, принятые в основном корпусе НКРЯ, в частности, следующие пометы частей речи: A (имя прилагательное), S (имя существительное), V (глагол), SPRO (местоимение-существительное); лексико-семантические пометы inan («неодушевленное»), abstr («абстрактное»), hum («лицо»), org («организация»), food («пища»), famn («фамилия»), persn («имя»). Также используются принятые в грамматике составляющих обозначения групп: NP (именная группа), VP (глагольная группа), PP (предложная группа).

Экспликация — это адаптация понятия «семантическая роль» к разным типам предикатов. Как показал опыт развития FrameNet, традиционный список семантических ролей неизбежно разрастается при расширении круга размеченных глаголов и увеличении тематического разнообразия описываемых ими ситуаций, а применительно к именам существительным и прилагательным инвентарь ролей вообще не разработан. В силу этих обстоятельств было принято соглашение, что в поле «Экспликация» может быть указана семантическая роль или стоять иное пояснение, помогающее отождествить участника, например, в МУ имени *слуга* (ср. *слуга Петра*) участник Петр может быть обозначен как «хозяин» или «тот, кому служат»). Мы предполагаем, что в дальнейшем инвентарь этих маркеров будет пересмотрен и систематизирован в особую иерархическую структуру (ср. граф фреймовых элементов во FrameNet).

По желанию пользователя «имя» участника может быть представлено в двух вариантах (X, Y, Z… или 1, 2, 3…), а способы поверхностно-синтаксического выражения — в традициях грамматики зависимостей (в + S_{loc}, в + N_{loc}, ср. практику ТКС, Апресян, Палл 1982, системы «Лексикограф» и др.) или грамматики составляющих (в + NP_{loc}; PP, см. ниже). Поле синтаксического ранга (Субъект, Объект, Периферия, Инкорпорированный участник) присутствует только в МУ глаголов. Семантические ограничения на заполнение валентности записываются в стандарте тегов семантической разметки НКРЯ; в случае жестких лексических ограничений здесь может быть перечислен список конкретных лексем русского языка (ср. другое значение глагола *служить*, в котором объектная валентность может быть заполнена именами *служба, молебен, панихида, обедня*).

В качества образца МУ предметного имени на рис. 3. представлен фрагмент статьи слова *корпус*, где указана модель, релевантная для употребления этого существительного в примере (1).

«корпус Тухачевского» (примеров: 1) ▶			
Имя	Экспликация	Морфосинтаксис	Семантические ограничения
Y	тот, кто командует	NPgen	hum, famn ▶

Рис. 3. Фрагмент статьи имени корпус, МУ «корпус Тухачевского»

5. Модели управления: реализация в тексте

Задача разметки управления в предложении — определить конкретные способы реализации валентностей в тексте, а также указать сопутствующие сирконстанты и «новых» участников (например, бенефицианта). В первую очередь, разметчик должен связать предикат с определенной МУ в словаре. Сделав это, он получает шаблон, включающий МУ из словаря и новые поля для заполнения, см. рис. 4.

Во вторую очередь, разметчик должен указать в предложении элементы, соответствующие каждому участнику. Шаблон заполняется в двух стилях: грамматики составляющих (ГС) и грамматики зависимостей (ГЗ). Например, заполняя шаблон глагола *служить* (см. пример выше) в стиле ГС, разметчик должен выделить в предложении словосочетание «в конном корпусе Гая» и связать его с полем <phrase>; а заполняя аналогичную позицию в стиле ГЗ, он должен выделить два элемента («в» и вершину именной группы, «корпусе») и связать их с полем <word>.

На следующем этапе в действие вступает программа автозаполнения, которая вносит информацию в остальные поля шаблона на основе имеющейся морфологической и семантической информации в паспорте лексемы. Затем разметчик может вручную откорректировать данные в шаблоне.

В случае, если в предложении присутствуют сирконстанты предиката или элементы, которые обозначают дополнительных (необязательных) участников ситуации, разметчик должен расширить шаблон и добавить данные о новых членах. Напротив, если кто-либо из обязательных участников не упомянут в предложении, следует указать причину его опущения. В частности, это может быть пассивизация; опущение субъекта в нефинитных формах; парцелляция; участник может быть известен из контекста (упомянут в предконтексте или постконтексте); управление может быть передано другому предикату (конструкция контроля) и т. д. Точно так же следует указывать конструкцию, «ответственную» за изменение синтаксического ранга участника.

Рассмотрим пример (2), в котором модели управления двух глаголов, *собрать* и *поесть*, реализуются нестандартно. Субъект ситуации, обозначаемой глаголом *собрать*, известен и упомянут как обращение в предшествующем контексте (*Оля*), однако он не может быть реализован при глаголе в форме императива. Пациенс обозначается инфинитивной группой — это развитие переходной МУ (*собрать корзинку с едой*). Кроме того, в ситуации появляется бенефициант, который вводится дативной группой (дитранзитивная конструкция), а также указана цель действия (обстоятельство *в дорогу*).

Субъект ситуации, обозначаемой глаголом *поесть*, известен и упомянут в предложении (*нам*), однако он синтаксически зависит от другого глагола (*собрать*) и по правилам грамматики не может быть реализован при глаголе в форме инфинитива. Пациенс у *поесть* генерический (любая еда) — это свойство конструкции «собрать поесть». Вообще, если

(1) Он служил в конном корпусе Гая. *[Анатолий Рыбаков. Тяжелый песок (1975–1977)]* ←...→ ⌂

«служить в министерстве»		lid004		(примеров: 4) ▶
Имя	Экспликация	Морфосинтаксис	Ранг	Семантические ограничения
X	тот, кто служит	NP_{nom}	Субъект	hum ▶
группа	<phrase>	<style NPnom>	▼	<semantic tags>
вершина	<word>	<style Snom>	<rank>	<semantic tags>
Y	место, где служат	в + NP_{loc}	Периферия	org ▶
группа	в конном корпусе Гая	PP: в + NPloc	Стандартный ▼	org
вершина	в корпусе	в + Sloc	Периферия	org
+				

Рис. 4. Пример шаблона с частично заполненными полями

(2) Оля! ***Собери*** *нам поесть в дорогу.* *[Фазиль Искандер. … (…)]* ←...→ ⌂

«собрать поужинать»	собрать V,2p,act,imper,pf,sg		lid001	(примеров: 14) ▶
Имя	Экспликация	Морфосинтаксис	Ранг	Семантические ограничения
X	агенс	NP_{nom}	Субъект	hum ▶
группа			Императив ▼	
вершина			Нет	hum
Y	то, что собирают	VP_{inf}	Периферия	inan ▶
группа			Стандартный ▼	
вершина			Периферия	
Z	бенефициант	NP_{dat}	Периферия	hum ▶
группа	нам	NPdat	Дитранзитив ▼	
вершина	нам	SPROdat	Периферия	hum persn
W	цель		Сирконстант	▶
группа	в дорогу	в + NPacc	Генерический ▼	
вершина	в дорогу	в + Sacc	Нет	abstr
+				

(2) Оля! Собери нам ***поесть*** *в дорогу.* *[Фазиль Искандер. … (…)]* ←...→ ⌂

«есть рыбу»	поесть V,act,inf,pf,v		lid003	(примеров: 2) ▶
Имя	Экспликация	Морфосинтаксис	Ранг	Семантические ограничения
X	агенс	NP_{nom}	Субъект	hum ▶
группа			Инфинитив ▼	
вершина			Нет	hum
Y	пациенс	NP_{acc}	Периферия	food ▶
группа			Генерический ▼	
вершина			Нет	food
+				

Рис. 5. Реализация МУ у глаголов собрать и поесть

бы пациенс ситуации *поесть* был специфицирован, он бы был «захвачен» глаголом *собрать*, ср. *Вот, собрала вам пирожков поесть*; однако чаще объект в этой конструкции опускается. Таким образом, конструкция «собрать поесть», с одной стороны, входит как элемент в периферийную МУ глагола *собрать*, а с другой стороны, является частным случаем конструкции контроля объекта, обладая в то же время собственными синтаксическими свойствами.

6. Словарь конструкций

Понятие «конструкции» мы понимаем максимально широко, в традициях основного направления Грамматики Конструкций (Fillmore et al. 1988, Goldberg 1995, см. также обзор Кузнецова 2007). В частности, конструкциями мы называем:

- тривиальные синтаксические конструкции: A + S, на + SPRO и т. д;
- общие синтаксические конструкции: сочинение и подчинение, перестановки порядка слов, конструкции повтора, пассив, генитив при отрицании, безличная конструкция, дитранзитивная конструкция (*испекла Пете* пирог), локативная трансформация (*грузить баржу лесом*), аппозитивная адъективная (*валялся пьяный*), компаративная (*гиены трусливее зайцев*) и т. д. – многие их этих конструкций основаны на «трансформациях» моделей управления;
- модели управления глаголов, предикатных имен (*ненависть*), реляционных предметных имен (*пациент*), имен прилагательных (ср. *готов к выступлению*), наречий, предлогов и т. д.;
- неоднословные лексические элементы (обороты): например, *иметь в виду*; *несмотря на*, вводные обороты, конструкции с лексическими функциями (*оказывать влияние*), а также их «малый синтаксис»;
- периферийные словоцентричные конструкции, например, сериальная (*сижу смотрю*), конструкции типа *гулять так гулять, взял да и помер* и т. д.

Как уже было указано ранее, мы хотим учесть опыт развития системы FrameNet и собрать данные о конструкциях последнего типа, т. е. ориентированных вокруг отдельных лексических элементов или фразем. Источниками словаря являются Грамматика 1980, Шведова 1960, Золотова 1980, лингвистические описания отдельных конструкций, а главное, конструкции, обнаруженные при разметке предложений. В общих чертах структура словарного входа сходна со словарем МУ, представляя перечень структурных моделей конструкции. В каждой модели перечисляются элементы конструкции, как постоянные, так и переменные, и указаны их лексические, грамматические и семантические ограничения. В отличие от МУ, отдельное поле посвящено порядку слов и разрывности несущих элементов конструкции.

7. Заключение

Предлагаемая система должна иметь несколько индексов (словарь лексем, словарь типов моделей управления, словарь конструкций). Выбрав лексический вход, можно увидеть набор МУ данного слова, получить список конкретных поверхностно-синтаксических вариантов ее реализации и далее просмотреть аннотированный корпусной материал. Каждая строка в МУ также содержит гиперссылку, которая дает возможность получить список конкретных вариантов реализации данной валентности в предложениях.

Кроме того, корпусный словарь должен предоставлять возможности поиска. Пользователь может задать шаблон модели управления или конструкции, оговорить ограничения на лексические, грамматические и лексико-семантические признаки; ограничить длину предложения или его тип (сложноподчиненные и т. д.). В системе можно искать без учета порядка слов, или же, напротив, задать ограничения на линейный порядок элементов. В результатах запроса будет содержаться также информация о частотности того или иного явления в материалах корпусного словаря.

Естественное ограничение данной системы в том, что здесь нельзя проследить распределение моделей управления в пределах целого текста. Впрочем, в дальнейшем мы не исключаем возможности составления экспериментального корпуса со сплошной фреймнет-разметкой.

Литература

1. *Азарова И. В., Синопальникова А. А., Яворская М. В.* Принципы построения wordnet-тезауруса RussNet // Кобозева И. М., Нариньяни А. С., Селегей В. П. (ред.), Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2004. М.: 2004. С. 542–547.
2. *Апресян Ю. Д.* О проекте активного словаря (АС) русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). М.: РГГУ, 2008. С. 23-31.
3. *Апресян Ю. Д., Палл Э.* Русский глагол — венгерский глагол. Управление и сочетаемость. Будапешт, 1982.
4. *Грамматика 1980:* Русская грамматика. М., 1980.
5. *Золотова Г. А.* Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. М., 1980.
6. *Кузнецова Ю. Л.* Грамматика Конструкций. Обзор // НТИ. Сер. 2, № , 2007?.
7. *МАС: Евгеньева А. П.* Словарь русского языка: В 4 т. 2-е изд. М., 1981–1984 г.
8. *НОСС: Апресян Ю. Д.* и др. Новый объяснительный словарь синонимов. М., 2003.
9. *Ожегов С. И., Шведова Н. Ю.* Толковый словарь русского языка. М., 1992.
10. *Падучева Е. В.* Динамические модели в семантике лексики. М., 2004.
11. *Падучева Е. В., Кустова Г. И., Розина Р.* «Лексикограф». http://lexicograph.ru.
12. *Сазонова И. К. Толково-грамматический словарь русских причастий. М., 2008.*
13. *Толдова С. Ю., Кустова Г. И., Ляшевская О. Н.* Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка: глаголы // Труды международной конференции «Диалог 2008». М., 2008. С. 522–529.
14. *ТКС: Мельчук И. А., Жолковский А. К.* Толково-комбинаторный словарь современного русского языка. Вена, 1984.
15. *Шведова Н. Ю.* Очерки по синтаксису русской разговорной речи. М., 1960.
16. *Atkins B. T.* Tools for computer-aided corpus lexicography: the Hector project // Papers in Computational Lexicography: Complex'92, F. Kiefer, G. Kiss and J. Pajsz (eds.) Hungarian Academy of Sciences, Budapest, 1992. P. 1–60.
17. *Fillmore C., Kay P., O'Connor K. T.* Regularity and idiomaticity in grammatical constructions: the case of let alone // Language, №64, 1988. С. 501–538.
18. *Fillmore Ch.* Border conflicts: FrameNet meets Construction Grammar // EURALEX 2008.
19. *Goldberg A.* Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press, 1995.
20. *Johnson C., Fillmore C., Petruck M, Baker C., Ellsworth M., Ruppenhofer J. and Wood E.* FrameNet: Theory and Practice. [Electronic resource]. Mode of access: http://www.icsi.berkeley.edu/framenet.
21. *Kilgarriff A., Rundell M. and Uì Dhonnchadha E.* Efficient Corpus Creation for Lexicography: building the New Corpus for Ireland' // Language Resources and Evaluation, 40, 2006. P. 127–152.
22. *Kipper K., Korhonen A., Ryant N., Palmer M.* Extending VerbNet with novel verb classes // Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy. June, 2006. См. также http://verbs.colorado.edu/~mpalmer/projects/verbnet.html.
23. *Palmer M., Gildea D., Kingsbury P.* The Proposition Bank: A corpus annotated with semantic roles // Computational Linguistics Journal, 31:1, 2005. См. также http://verbs.colorado.edu/~mpalmer/projects/ace.html.
24. *Ronald D., Jurafsky D., Menn L., Gahl S., Elder E., Riddoch C.* Verb subcategorization frequency differences between business-news and balanced corpora: the role of verb sense // ACL Workshop on Comparing Corpora, 2000.

Части тела с точки зрения топологии: корпусное исследование[1]

Names of body parts from the viewpoint of topology

Махова А. А. (discourse@yandex.ru), **Ляшевская О. Н.** (olesar@mail.ru)
University of Tromsø, Норвегия

Десятова А. В. (patine@gmail.com)
Российский государственный гуманитарный университет, Москва, Россия

Работа посвящена исследованию имен, называющих части тела, с точки зрения понятия топологического типа, введенного Л. Талми. На основе анализа сочетаемости данных имен с прилагательными формы и размера определяется их место в топологической классификации и рассматриваются особенности их пространственной семантики.

1. Введение

Имена, называющие части тела, представляют особую тему в лексической и когнитивной семантике. Среди работ по русскому языку, непосредственно относящихся к нашей теме, можно назвать статью Рахилина 1989, в которой затрагивается проблема частей и целого в отношении частей тела. В частности, автор замечает, что в русском языке имена частей тела ведут себя специфическим образом: невозможно такое сочетание, как, например, **пальцы правой ступни*, но возможно *пальцы правой ноги*. Таким образом, пальцы представляются русскоговорящим частями ноги (или руки), но не ступни (или кисти).

В работе Е.В. Рахилиной и В.И. Подлесской (2000) речь идет об ориентации в пространстве объемных объектов и об особой конструкции с творительным падежом, описывающей эту ситуацию: *лицом к стене*, *(упасть) носом в землю* и т.п. Авторы статьи опираются на понятие топологического типа, которое будет положено также в основу нашего доклада.

Непосредственно форма частей тела служит предметом исследования в рамках теории жестовой коммуникации (Крейдлин 2002; Крейдлин, Летучий 2007; Аркадьев и др. в печати). В указанных работах в основном обсуждается ориентация частей тела в пространстве в зависимости от конкретного жеста. Также проводится классификация частей тела: выделяются так называемые составные части тела, например, рука, состоящая из собственно руки и ладони, и т. д.[2].

В нашей работе мы обратимся к проблеме восприятия и языкового описания формы частей тела, и рассмотрим ее с точки зрения теории топологических типов. Идея такого исследования родилась в ходе изучения прилагательных формы. Оказалось, что имена, называющие части тела, ведут себя особенным образом в сочетании с прилагательными этой семантической группы. Так, некоторые из этих имен могут сочетаться с большим количеством разнообразных прилагательных формы (например, *нос* или *лицо*), а другие — лишь с ограниченным набором (*ноги*, *щеки*). Некоторые части тела могут характеризоваться прилагательными, которые обычно не сочетаются с именами объектов подобной формы (например, *круглые руки*). И, наконец, части тела, между которыми, на первый взгляд, нет ничего общего, могут описываться одними и теми же прилагательными (*выпуклый лоб — выпуклые глаза*). На наш взгляд, объяснить подобные случаи можно,

[1] Работа выполнена при поддержке РГНФ, грант № 07-04-00240а.

[2] «Тело человека не представляет собой некоего единого целого, оно состоит из отдельных частей, а эти части сами могут состоять их отдельных частей. Назовем все такие части, части частей и т.д. составными частями тела. <…> С языковой точки зрения противопоставлены, например, рука и ладонь, при том, что биологически ладонь — часть руки…» (Переверзева 2007).

используя понятие топологического типа. Таким образом, задача статьи — отнести имена, называющие части тела человека, к различным топологическим классам и проанализировать их сочетаемость с прилагательными формы и размера.

2. Топологическая классификация предметных имен

Понятие топологического типа было предложено Леонардом Талми (Talmy 1983/2000). Топологическими типами он назвал эталонные формы (например, контейнеры, поверхности, линии и т.д.), с которыми человек соотносит все материальные объекты в процессе восприятия. От того, к какому топологическому типу в языковом сознании человека принадлежит объект, зависит и тот набор параметров, по которому он будет оцениваться как *длинный/широкий/большой* и т.д. Говорящий оценивает объект не по отдельным признакам (длине, ширине, высоте), а целиком, как относящийся к какому-либо топологическому типу (Рахилина 2000).

Идея топологического типа непосредственно связана с антропоцентричностью языка. Не каждому материальному объекту могут быть сопоставлена некоторая выбранная форма и/или размер: это произойдет только в том случае, если они **функционально** важны для человека. Например, невозможны сочетания **толстый потолок* или **прямоугольная книга*, так как объем потолка для нас несуществен, как и привычная форма книги. В обычной ситуации мы не скажем *?длинный подоконник* или *?короткий подоконник*: длина подоконника для нас не релевантна, зато важна площадь его поверхности, на которую можно что-либо поместить. Поэтому частотными будут сочетания *широкий* или *узкий подоконник*. Для языкового сознания не имеет значения реальные форма и размер объектов, их расположение в пространстве: мы обращаем внимание только на те их измерения, которые функционально значимы для нас.

Определить принадлежность какого-либо объекта к одному из топологических типов можно, опираясь на сочетаемость имени, называющего данный объект. Одним из ярчайших показателей является сочетаемость с предлогами. Так, имя, сочетающееся с предлогами *в, из, внутри, изнутри* может быть с определенной долей вероятности отнесено к топологическому классу контейнеров, если этому не противоречит его функция. Другим показателем является адъективная сочетаемость: например, к топологическому классу пластин может быть отнесено имя, сочетающееся с прилагательными *толстый* и *плоский*; к контейнерам — *глубокий* и т.д.

Анализ существительных, называющих части тела, мы провели, основываясь на их сочетаемости с прилагательными формы и размера. Рассматривались следующие имена: *голова, грудь, спина, плечи, локти, руки, пальцы, ноги, бедра, колени, пятки, живот, лицо, лоб, щеки, брови, нос, глаза, губы (рот), подбородок* и др. Во внимание не принимались, в частности, *уши* как объекты особого типа; *волосы*, которые не могут быть однозначно отнесены к частям тела; *ресницы*, *зубы* и другие части тела, представляющие собой множество объектов. Основным материалом исследования и источником примеров послужил корпус НКРЯ.

3. Топологические типы частей тела по данным языка: сочетаемость с прилагательными формы

Начиная наше исследование, мы провели априорную топологическую классификацию частей тела, основанную на интуитивном сходстве этих объектов с другими окружающими предметами (инвентарь топологических типов приведен в работе Десятова и др. 2008). В частности, основными типами здесь оказались:

1) **Выступы** (объемные предметы, имеющие выступающую часть, а также сами выступающие части): *локти, колени, нос, подбородок, плечи, горб*.
2) **Стержни-столбы** (вытянутые объекты жесткой формы, ориентированные вертикально): *ноги*.
3) **Стержни-веревки** (вытянутые объекты, не имеющие жесткой формы): *хвост*, *волосы, руки, пальцы*.
4) **Дуги** (объекты или полосы, имеющие форму полуокружности): *бровь*.
5) **Полосы** (вытянутые объекты, имеющие поверхность): *пояс*.
6) **Вертикальные поверхности:** *спина*, *лоб*.
7) **Пластины** (объекты, имеющие небольшую толщину и широкую поверхность): *ладонь*, *подошва*.
8) **Вместилища** (объекты, имеющие полость): *рот, рука*.
9) **Шары** (объемы, имеющие шарообразную форму): *голова*.
10) **Круги**: лицо.

Для проверки нашей гипотезы мы составили базу данных на основе выборки из материалов НКРЯ. В табл. 1 приводится фрагмент статистической части базы, которая дает наглядное представление о типичности словосочетаний имен частей тела с теми или иными прилагательными.

Как оказалось, многие наши предположения подтвердились — и в то же время некоторые факты оказались неожиданными. Ниже будет дано опи-

Табл. 1. Частотность сочетаний некоторых имен частей тела с прилагательными формы и размера (по данным выборки НКРЯ).

	толстый	тонкий	длинный	короткий	широкий	узкий	высокий	низкий	круглый	впалый	выпуклый	плоский
бровь	10	99	18	4	25	7	14	7				
голова	13	2	11	1	6	19	2		137			9
грудь	15		8	1	176	19	146		9	106	41	33
живот	31					2	9	1	37	26	14	19
лоб					71	58	244	96				
нос	53	99	282	29	68	13						
палец	137	224	212	79	4	1			2			7

сание двух групп имен: в одной имена «послушно» сочетаются с набором прилагательных, предсказанных для их топологического класса, другие же (например, *нос*), выступают в сочетаниях, характерных для разных классов.

3.1. Имена частей тела, принадлежащие одному топологическому типу

Категоризация имен частей тела по топологическим типам позволила некоторые из них достаточно определенно отнести к тому или иному классу: таковы слова *грудь*, *живот*, *бедра*, *голова*, *ноги*. Прилагательные, с которыми сочетаются эти имена, обычно сочетаются с существительными, имеющими определенную форму: *выпуклый узор* — выступ, *ровный пол* — поверхность, *узкий луч* — полоса.

Наш анализ показал, что практически все части тела, кроме, пожалуй, ног и рук, кажутся выступами или в том числе выступами (ср. *плечи, колени, бока, нос, подбородок, щеки* и др.). Очевидно, это связано с тем, что в сознании говорящих части тела важны именно как **выступающие** части — они приковывают внимание, позволяют отождествить форму и запомнить отличительные особенности человека.

Как может показаться, *грудь* и *живот* должны прежде всего осмысляться носителями русского языка как поверхности, но на самом деле они, как правило, фигурируют как выступы. Так, *грудь* может быть *впалой*, *выпуклой*, *выгнутой*, *плоской*, *крутой*, *покатой*, *пологой*; имя *живот* сочетается почти с таким же набором прилагательных, ср. (1–3):

(1) *Угловатый мужчина с **впалой грудью** и аккуратной рыжей бородой* [Д. Симонова. Сорванная слива].

(2) *Смоляные кудри ниспадали на лоб спящего бога торговли, воровства и адюльтеров, на **впалом животе** дремал кот* [О. Некрасова. Платит последний].

Несмотря на отсылку к плоской или вогнутой форме, прилагательные *плоский* и *впалый* характеризуют не поверхности (ср. **плоский пол*, *впалая стена*), а нечто недостаточно выступающее по сравнению с ожидаемым (ср. *плоский затылок, впалые щеки*). *Крутой, покатый* и *пологий* — это прилагательные, описывающие наклонные поверхности (ср. *крутой/пологий спуск*) и особого рода выступы — ступени (ср. *крутой берег, крутой обрыв*) и горы (ср. *крутая скала*). Хотя топологических типов несколько, все они функционально связаны с реальным или воображаемым подъемом/спуском. Тем самым, *крутой* — это тот, на который трудно подняться. Среди имен частей тела названные прилагательные также выбирают выступы (ср. *крутые бока, бедра, плечи, подбородок*), в частности, *крутая грудь* метафорически представляется как высокая ступень:

(3) *Откинув голову, готовая обнять мир, с улыбкой счастья, девушка возносилась из бега морской волны, и свет влажно мерцал на **крутой груди**, ветер откидывал невесомые волосы, и вся она была порывом открытой солнцу юности* [Д. Биленкин. Вечный свет].

Существительное *ноги* относится к другому топологическому классу — это столбы. *Ноги* могут быть *длинными*, *толстыми*, *худыми*, *прямыми*, *ровными* и т. д. Столбы — вообще довольно узкий топологический тип, ограниченный сразу несколькими параметрами: объекты, входящие в него, должны быть вытянутой цилиндрической формы, при этом довольно широкими (чтобы отличаться от стержней и веревок) и главное — вертикально ориентированными. Под эти критерии из всех частей человеческого тела подходят только *ноги*, поскольку именно они являются тем, на чем «держится» человеческое тело.

Еще одна «особенная» часть тела — это *голова*. Она тоже относится к одному топологическому типу, к которому больше ни один из рассматриваемых объектов отнести нельзя, — к шарам. Существительное голова сочетается с большим количеством

прилагательных формы, самая типичная сочетаемость слова *голова — круглая*, *округлая*, *вытянутая*, *овальная*, *шарообразная голова*. Но многие из прилагательных, сочетающихся со словом *голова*, заключают в себе яркий оценивающий или описательный компонент: *грушевидная, клинообразная, остроконечная, приплюснутая, продолговатая, угловатая, фигуристая, яйцевидная голова*:

(4) *Он не облысел, а совершенно облез, и* ***фигуристая голова*** *его с выносом на затылок напоминала мозговую кость с колбасного завода* [В. Астафьев. Веселый солдат].

Интересно, что все вышеперечисленные прилагательные описывают форму, являющуюся ярким отступлением от «правильной» шарообразной, каковой и должна быть голова в представлении человека (или, если говорить точнее, эллипсоидной): слишком похожая на шар — *сплюснутая*, *приплюснутая*, слишком вытянутая — *клиновидная*, *яйцевидная*, имеющая лишние выступы, деформированная — *фигуристая*, *грушевидная*, *угловатая*.

3.2. Имена частей тела, принадлежащие нескольким топологическим типам

Многие имена частей тела могут относиться сразу к нескольким топологическим категориям. В этом случае существительные сочетаются с разными рядами прилагательных — например, с теми, которые обычно характеризуют выступы, и с теми, которые обычно относятся к поверхностям.

В основном, части тела принадлежат к двум топологическим типам, один из которых является выступом. Второй же тип часто характеризует видимый контур объекта: например, круг или овал, ограничивающий *глаз* (ср. *круглые, миндалевидные глаза*), полосу *носа* или *губ*, дугу или полосу *бровей*. Как мы уже сказали, выступы выделены в сознании человека — по ним мы замечаем отличительные особенности внешности. Контуры не менее важны в этом отношении, в особенности это касается тех частей, которые составляют лицо. Можно сказать, что лицо — это картина, на которой «нарисованы» круги и полосы.

Типичным представителем сдвоенного типа «выступ + контур» являются *глаза*. Чаще всего воспринимается именно контур глаз, о чем говорит количество и разнообразие соответствующих прилагательных, сочетающихся с этим словом: *изогнутые, овальные, продолговатые, удлиненные, миндалевидные глаза:*

(5) *Её маленькое скуластое лицо и* ***продолговатые глаза*** *с большими веками заворожили меня.* [Б. Минаев. Детство Левы].

Довольно ограниченное количество прилагательных характеризуют глаза как выступы: *вздутые, впалые, выпуклые, плоские глаза:*

(6) ***Плоские глаза*** *Долгина метнулись поверх голов, потом на бумажку, которую он держал в руках.* [Д. Гранин. Искатели].

Такое распределение по типам легко объяснимо: основной особенностью внешности человека является цвет и форма его глаз (и, соответственно, их размер). Характеристика глаз как выступов встречается довольно редко, причем во всех таких сочетаниях есть негативный оценочный компонент, из чего следует, что «нормальные» глаза в традиционном представлении вообще не должны быть выступающими.

Похожим образом воспринимаются *губы* — они также могут быть выступами и контурами-полосами, причем контурами чаще. Среди прилагательных, описывающих губы как выступы, также доминируют слова с негативной оценкой: *вздутые, вогнутые, выпуклые, надутые, плоские губы*, ср. (7), кроме одного контекста — *пухлые* или *пухленькие* губы, ср. (8):

(7) *А Ася ласково улыбалась ярко накрашенным ртом, ее* ***выпуклые губы*** *походили на резиновый потрескавшийся шланг капельницы* [И. Полянская. Сельва].

(8) *Хозяин молодой, — игриво сказала она, покусывая яркие,* ***пухлые губы*** [Б. Екимов. Пастушья звезда].

Губы как контуры могут быть описаны как *волнистые, вытянутые, длинные, извилистые, изогнутые, прямые, тонкие, узкие*:

(9) *Господин в бархатном картузе показывал Семену Ивановичу лавку, любовно притрагивался к пыльным полированным плоскостям, мудрено вытягивал* ***извилистые губы*** [А. Н. Толстой. Похождения Невзорова, или Ибикус][3].

Брови — еще одна часть тела, относящаяся одновременно к двум топологическим типам. Видимым контуром бровей является дуга. Контекстов с прилагательными, описывающими брови как выступы, довольно мало: *выпуклые брови* и *круглые брови*[4]:

[3] У существительного *рот* такое же соотношение в сочетаемости: рот — выступ (*надутый*, *выпуклый*, *впалый*, *плоский рот*), рот — контур (*прямой*, *вытянутый*, *извилистый*, *подковообразный рот*).

[4] О неоднозначности сочетания *круглые брови* и других сложных случаях см. в разделе 4.

(10) *У него был такой строгий сухой нос, и резкие* ***выпуклые брови****, и седина, нелгущая седина в волосах.* [С. Н. Сергеев-Ценский. Печаль полей].

Зато очень много прилагательных описывают брови как арки (дуги): *гнутые*, *дугообразные*, *изогнутые*, *продолговатые*, *овальные брови:*

(11) *Сзади их, то дергая мужей за рукав, то снимая через их плечо стакан с водкой, — для себя, разумеется, — сидели рослые женщины с* ***гнутыми бровями*** *и руками круглыми, как булыжник.* [А. С. Грин. Алые паруса].

Особый интерес представляют части тела, которые могут относиться более чем к двум топологическим типам. Один из таких примеров — нос. Он является, что вполне естественно, выступом, а также стержнем и контуром. Нос является самой яркой деталью человеческого лица, и говорящий может видеть его в разных ракурсах. Во-первых, просто как выступ довольно сложной формы: *грушеобразный, закругленный, плоский, сплюснутый, крутой, выгнутый, острый, остроконечный, курносый нос*, ср.

(12) *Лицо прокурора было мясистое и красное, с толстыми губами и широким,* ***грушеобразным носом*** [А. Р. Беляев. Человек-амфибия].

Другая ипостась носа — стержень, прикрепленный к лицу (вообще же, стержень — это объект вытянутой цилиндрической формы, достаточно узкий, свободно ориентированный в пространстве): *длинный, короткий, прямой, утолщенный, широкий, тонкий нос*, ср.

(13) *Из-за козырька я видел только его короткий* ***широкий нос*** *и узкие губы* [В. Скворцов. Каникулы вне закона].

Наконец, если говорящий смотрит на нос в профиль, его могут заинтересовать очертания носа: *вздернутый, орлиный, крючковатый, крючкообразный нос.* Этот контур нельзя отнести ни к одному из ранее перечисленных топологических типов, ср.

(14) *Оживленный, хорошо откормленный, при усах под небольшим* ***крючковатым носом****, он показался ей зеленоглазым, настоящей одесской красоты мужчиной* [В. Беломлинская «…Где пасешь ты? Где отдыхаешь в полдень?»].

Еще одна часть тела, принадлежащая сразу к нескольким топологическим типам, — это лицо. Оно также является одной из важнейших воспринимаемых частей тела человека, причем оно как бы складывается из нескольких других — носа, глаз, рта, щек и т.д. — то есть лицо является местом, где все они располагаются.

По сочетаемости с прилагательными *лицо* можно отнести к следующим топологическим типам — выступ, поверхность, контур. Это значит, что лицо является одновременно и двухмерным, и трехмерным объектом. Существительное *лицо* сочетается с прилагательными *вздутый*, *вогнутый*, *впалый*, *выпуклый*, *плоский*, и в этих случаях лицо является выступом — трехмерным объектом:

(15) *Искаженно-****выпуклое лицо*** *расплывалось* [Е. Парнов. Третий глаз Шивы].

С другой стороны, лицо может быть *овальным*, *продолговатым*, *удлиненным*, *остроугольным*, *угловатым*, *прямоугольным*, *круглым* и, следовательно, быть двухмерным объектом и характеризовать контур:

(16) *Ее* ***остроугольное лицо*** *было насыщено какой-то нервной дрожью* [Б. К. Зайцев. Голубая звезда].

Ровное, *узкое*, *широкое лицо* — словосочетания, характеризующие лицо как поверхность:

(17) ***Широкое лицо*** *его светилось безмятежно довольной улыбкой* [В.В. Крестовский. Панургово стадо].

Можно обобщить сказанное тем, что все части тела, «находящиеся» на лице, выступают в двух, почти равноправных ипостасях — как контуры и как выступы. Очевидно, для описания человеческой внешности принципиально важны и абрис, и рельеф таких частей тела, как нос, глаза, губы, брови.

4. Трудные случаи категоризации имен частей тела

В первую очередь, нужно отметить, что сочетаемость имени с предлогами и с прилагательными не всегда полностью соотносится: например, если адъективная сочетаемость имени говорит о том, что оно относится к одному топологическому типу, это не значит, что оно не может сочетаться с предлогом, характерным для других типов. Особым образом ведет себя предлог *на*: он может сочетаться почти со всеми существительными, называющими части тела, даже если они не относятся к типу поверхностей или относятся к нему в меньшей степени (*на щеках*, *на животе*). Очевидно, это связано с тем, что части тела, будучи обычно в той или иной степени могут быть задействованы человеком в разных целях, особенно легко могут функционировать в качестве поверхностей.

4.1. Случаи двойной категоризации

В некоторых случаях существительное, обозначающее часть тела, может вести себя неоднозначным образом. К таким случаям относится сочетаемость слова *лоб* с прилагательными формы. С одной стороны, *лоб* может быть *широким* и *узким*, с другой стороны — *высоким* и *низким*. В обоих случаях лоб — это поверхность (хотя лоб может быть также и выступом — *выпуклый*, *вогнутый*, *плоский*, *покатый*, *крутой лоб*), но в одном случае она ориентирована горизонтально (полоса), а в другом вертикально:

На его выпуклом ***широком лбу*** *быстро проступал пот* [Е. Лукин. В стране заходящего солнца].

Одни глаза были те же — волнующие и беспокойные; в них-то и тонула мать, смеясь и плача, изредка трогая морщеной, блеклой ладонью прямые коротко остриженные волосы сына и белый его ***узкий лоб*** [М. А. Шолохов. Тихий Дон].

Только трепетала от дыхания пышная рыжая челка над ***высоким лбом*** [А. Берсенева. Полет над разлукой].

Эти цельность и безоговорочность достойны зависти, как и гладкие ***низкие лбы****, не отягощенные сомнением и раздумьем* [В. Валеева. Скорая помощь].

В этих случаях прилагательные формы одновременно выполняют роль прилагательных размера: *широкий* и *высокий лоб* противопоставлены по размеру *узкому* и *низкому лбу*. Однако негативные коннотации имеет только словосочетание *низкий лоб* — оно обычно относится к людям недалеким, грубым.

Еще один сложный случай — такая часть тела, как брови. Сложность связана с тем, что брови состоят из двух равноценных частей, находящихся на некотором расстоянии друг от друга под углом друг к другу, то есть это парная часть тела. При этом сама бровь является, как мы уже говорили выше, дугой, то есть тоже состоит из двух симметричных частей. Существительное брови часто сочетается с прилагательными: *угловатые, разлатые, крутые*, ср.

(18) *Она была похожа на башню: огромная, высоченная женщина с розовым лицом, ярко накрашенными губами и* ***крутыми бровями*** [А. Иванов. Географ глобус пропил].

Все эти прилагательные тем или иным образом характеризуют угол, но без специальных указаний автора не представляется возможным понять, имеется ли в виду угол между двумя бровями или угол между двумя частями одной брови, то есть является ли, например, *крутой* каждая бровь или они поднимаются по направлению к друг другу под крутым углом.

4.2. Особенные прилагательные: *круглый*

Прилагательное *круглый* само по себе привлекало к себе внимание лингвистов (см. Рахилина 2000, Гилярова 2002 и библиографию там). В частности, Е. В. Рахилина показала, что это прилагательное характеризует объекты самой разной формы, а не только те, которые в нашем сознании ассоциируются с кругом. Из этого она заключает, что значение прилагательного *круглый* состоит в характеризации не формы объекта, а изменения этой формы: «Единство значения прилагательного *круглый* оказывается не в общности результата этой своеобразной семантической трансформации — т. е. не в том, чтобы все объекты, к которым оно применяется, стали шарами, а в общности самих изменений, которые претерпевают разные формы, оказываясь круглыми» (Рахилина 2000: *135*). Как оказалось, в сочетании с именами, называющими части тела, это прилагательное также ведет себя необычным образом. Первая особенность *круглого* заключается в том, что оно сочетается с очень большим количеством таких имен, почти со всеми. Среди наиболее очевидных — *голова* и *лицо*. Однако голова относится к топологическому типу шаров (сюда мы приписываем ее относительно условно, так как «образцовая» голова имеет форму вытянутого шара), а лицо — к выступам, поверхностям и контурам. В данном случае лицо интересует нас как контур (контур мы понимаем в широком смысле и называем так любую линию, которая может быть проведена на поверхности объекта), имеющий форму круга (примером подобного сочетания может служить также сочетание *круглые глаза*). Проще говоря, голова является трехмерным объектом, а лицо — двухмерным. Круг — это тоже плоская фигура. Однако частотное сочетание *круглая голова*, конечно, означает «голова правильной шарообразной формы», то есть в данном случае мы имеет дело с переходом объемных фигур в плоские. Аналогичный пример — *овальная голова* — менее частотное, но все же распространенное словосочетание.

Здесь уместно сделать небольшое отступление, касающееся собственно существительного *голова*. В принципе, его сочетаемость не представляет особого интереса: в основном, оно сочетается с прилагательными, обозначающими разные виды «деформаций» — отступлений от правильной шарообразной формы (*грушевидная, вытянутая, приплюснутая, яйцевидная голова*). Необычно сочетание *удлиненная* и *длинная голова*:

(19) *Орлович теперь своими глазами видел на лестничной площадке англичанку, если не немку или не шведку, в черном пальто-дутике и в теплых наушниках на* ***удлиненной голове*** [В. Аксенов. Негатив положительного героя].

В современных текстах мы встретим только *удлиненную*, но не *длинную голову*. Однако в текстах

XIX века и первой трети XX века такое сочетание встречается достаточно часто:

(20) *Через минуту он входил в комнату, нагибая свою **длинную голову**, чтобы не удариться о косяк низкой двери* [С.М. Степняк-Кравчинский. Андрей Кожухов (1898)].

(21) *Пьяненький мужичок, мотая взъерошенной **длинной головой**, заглядывал в колодец и тянул: — И-их* [Ф. Сологуб. В толпе (1907)].

Очевидно, такое сочетание практически полностью вышло из употребления на сегодняшний день.

Возвращаясь к прилагательному *круглый*, перейдем к другим именам частей тела, с которым оно сочетается. Это в основном те части тела, которые мы относим к выступам. Среди них можно выделить выступы шарообразной формы (в нашей общей классификации топологических типов мы не делаем таких тонких разграничений). Например, это щеки, живот и лоб. *Круглые щеки* и *круглый живот* — это, что вполне очевидно, щеки и живот полного человека либо человека с такой конституцией тела или лица, при которой эти части тела хорошо заметны, выпуклы. Однако с существительным *лоб* прилагательное *круглый* сочетается совсем не так активно. Этому можно дать следующее объяснение: лоб и его «круглость» никак не связаны с полнотой, грубо говоря, лоб всегда круглый, в отличие от щек и живота, которые могут быть втянутыми, впалыми или плоскими, поэтому в дублировании этой информации он не нуждается, несмотря на то, что и щеки, и живот, и лоб всегда сохраняют круглую форму как контуры. Таким образом, в случае с прилагательным *круглый* мы всегда имеем дело с некоторой контаминацией поверхностей и выступов (кругов и шаров).

Заметим еще одну интересную сочетаемостную особенность прилагательного *выпуклый* в отношении этих же имен. *Выпуклый живот* и *выпуклый лоб* — вполне частотные сочетания, в отличие от ?*выпуклых щек*. Возникает вопрос: почему лоб, который редко бывает *круглым*, достаточно часто бывает *выпуклым*? Это можно объяснить тем, что, как уже было замечено выше, *круглый лоб* не значит 'лоб толстого человека', а *круглый* (об этом еще будет идти речь ниже) в отношении частей тела всегда несет с собой такой подтекст. Таким же образом о животе толстого человека скорее скажут *круглый*, а не *выпуклый*. То есть *выпуклый* характеризует более устойчивую форму, ту, которая часть тела имеет от природы, а не вследствие каких-то естественных изменений, происходящих с человеческим телом. Что касается ?*выпуклых щек*, то щеки всегда выпуклы в той или иной степени (как и лоб, который всегда круглый). Щеки могут интересовать нас именно с точки зрения полноты, поэтому мы можем говорить о круглых или пухлых щеках или, наоборот, о впалых.

Еще один антоним *выпуклого — вогнутый*. Это слово в сочетании с именами частей тела иллюстрирует способность прилагательных формы вносить в описание человека элемент оценки. *Вогнутый лоб* как яркое отступление от нормы всегда несет негативную оценку:

(22) … *конвойные волокли <…> громадного, костлявого, длинноносого парня с потным лицом, побитым оспой, с длинными обезьяньими руками, **низким вогнутым лбом** и далеко выдающимся дегенеративным затылком, известного налетчика, грабителя, разбойника…* [В. Катаев. Трава забвенья].

Обратим внимание на сочетаемость существительного *лоб* с *выпуклый* и прилагательными размера. Существует *низкий вогнутый лоб* (см. пример выше), но при этом не встречается ?*высокий вогнутый лоб*. Известно, что высокий лоб всегда считался достоинством внешности:

(23) *Юля подняла лицо, и Дмитрий, как всегда, подумал, какая она милая с этим своим курносым носом, зелеными глазищами и рыжеватыми кудряшками над **высоким лбом**.* [Н. Катерли. Дневник сломанной куклы // «Звезда», 2001], — тем самым, отрицательная и положительная оценка не могут соседствовать в одном описании.

Обратимся теперь к наиболее «маргинальной сочетаемости» прилагательного *круглый*. Это касается следующих типов: выступов — спина, бедра, плечи, подбородок, нос, брови — и стержней — рук. Все они могут быть круглыми, хотя ни одна из этих частей тела не является ни кругом, ни шаром. Конечно, сочетаемость такого рода не является самой распространенной, скорее — авторской. И именно она подтверждает выводы Е. В. Рахилиной: *круглый* в данном случае значит 'округлившийся, ставший более округлым, приблизившийся к форме круга (или, точнее, шара)', что в отношении частей тела значит 'полный, пополневший'.

(24) *И уже — одновременно с матерью — другой рукой обнимая отца, поглаживая и похлопывая его по мягкой **круглой спине** и что-то ласково и ободряюще говоря в ответ на их бессвязные — счастливые и испуганные — восклицания, — он услышал вдруг легкие и быстрые, странно не знакомые этому дому шаги…* [С. Бабаян. Господа офицеры].

(25) *У нее были розовые щеки, **круглые бедра** и пышные плечи* [А. Мариенгоф. Роман без вранья].

(26) *— Таня радостно хлопнула подругу по **круглому плечу**, отчего та ойкнула и присела* [Е. Романова, Н. Романов. Дамы-козыри].

Неоднозначность может возникать в связи с сочетанием *круглый подбородок.* Обычно оно обозначает, что подбородок просто не имеет ярко выраженной квадратной формы (что характерно скорее для мужчин), то есть он округлый, «женский»:

(27) *Получалось не очень внушительно: чёлка светлых волос, вздёрнутый нос,* **круглый подборо́док** *— далеко до силача* [Б. Екимов. Фетисыч].

(28) *Сперва он увидел ее шею, с ямочкой у груди, потом* ***круглый подбородок****, потом глаза, ждущие, неестественно застывшие* [Д. Гранин. Искатели].

Между тем, в некоторых контекстах может иметься в виду именно подбородок полного человека:

(29) *Мягкий свет убывающей луны струился сзади, освещая часть городской стены и двух неизвестных, чуть склонившихся над парапетом. Оба с падающими на плечи волосами, крупными мясистыми носами и* ***круглыми подбородками*** [С. А. Еремеева. Лекции по истории искусства].

Круглый нос также, конечно, значит не ‘толстый нос’, а ‘нос картошкой’, то есть имеющий кончик, отдаленно напоминающий нечто шарообразное:

(30) *Выступление высокого, чернооkого, красивого, несмотря на* ***круглый нос*** *и круглое лицо, Кости Горелика было всегда дивертисментом.* [И. Дьяконов. Книга воспоминаний].

И, наконец, под *круглыми бровями*, вероятно, имеются в виду кустистые, выпуклые брови (ср. 31–32) или же дугообразные (ср. 33–34):

(31) *Исподлобья, из-под* ***круглых бровей*** *глядел недопесок на скачущего и бормочущего дядю Мишу.* [Ю. Коваль. Недопесок].

(32) *Да и слушать его не надо было, а только смотреть на эти прыгающие щеки, вспотевшие* ***круглые брови****, всю эту колыхающуюся искренним смехом тыкву, чтобы самому почувствовать, как вдруг щеки начинают расползаться и в груди что-то пищать — хи-ы!* [Н.А. Тэффи. Святой стыд].

(33) *Алька удивленно выгнула* ***круглую бровь*** *— бровями она была в тетку, — спрыгнула с табуретки, приподняла подол платья.* [Ф. Абрамов. Пелагея].

(34) *Рябков из щели в свою очередь заметил, до какой степени* ***круглые брови*** *начальника противоречат прямоугольной форме его устойчивых очков.* [О. Славникова. Стрекоза, увеличенная до размеров собаки).

Здесь, как и во многих других случаях наш анализ, увы, осложнен тем, что иногда мы можем только догадываться о той форме объекта, которую автор не счет нужным проиллюстрировать более подробно. Но не исключено и то, что для частей тела очень важным является гендерный аспект, и в одном и том же сочетании значение прилагательных может быть разным в зависимости от того, о ком идет речь — о мужчине или о женщине (как в случае с *круглыми бровями* и с *круглым подбородком*).

Литература

1. *Talmy L.* How language structures space // Talmy L. Toward a Cognitive Semantics. Vol. I. Cambridge, 2000.
2. *Аркадьев П. М., Крейдлин Г. Е., Летучий А. Б.* Семиотическая концептуализация тела и его частей // Вопросы языкознания (в печати).
3. *Гилярова К. А.* Языковая концептуализация формы физических объектов. Дисс… канд. филол. наук. М.: МГУ, 2002.
4. *Десятова А. В., Ляшевская О. Н., Махова А. А.* Конструкция с творительным формы «XY-ом» // Труды международной конференции «Диалог 2008». М., 2008. С. 133–139.
5. *Крейдлин Г. Е.* Невербальная семиотика: Язык тела и естественный язык. М.: Новое литературное обозрение, 2002.
6. *Крейдлин Г. Е., Летучий А. Б.* Концептуализация частей тела в русском языке и в невербальных семиотических кодах // Русский язык в научном освещении. М.: Языки славянских культур, 2006. № 2 (12). С. 80–116.
7. *Переверзева С. И.* Признак «ориентация» в системе частей тела и его языковое выражение. Дипломная работа. — М.: РГГУ, 2007.
8. *Рахилина Е. В.* К основаниям лингвистической мереологии // Семиотика и информатика. М.: Языки русской культуры, 1989. Вып. 30. С. 75–79.
9. *Рахилина Е. В.* Когнитивный анализ предметных имен: семантика и сочетаемость. М.: Русские словари, 2000.
10. *Рахилина Е. В., Подлесская В. И.* «Лицом к лицу» // Логический анализ языка: Язык пространства. М.: Языки русской культуры, 2000. С. 98–107.

Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике[1]

Automatic analysis of terminology in the Russian text corpus on corpus linguistics

Митрофанова О. А. (alkonost-om@yandex.ru)
Санкт-Петербургский государственный университет (СПбГУ)

Захаров В. П. (vz1311@yandex.ru)
Санкт-Петербургский государственный университет (СПбГУ),
Институт лингвистических исследований РАН (ИЛИ РАН)

В докладе рассматриваются результаты анализа русскоязычной терминологии корпусной лингвистики, полученные при совмещении ручной и автоматической обработки специального корпуса текстов. Особое внимание уделяется выявлению однословных и неоднословных терминов, использованию лексико-грамматических шаблонов для описания внутренней структуры терминов, а также терминообразующих контекстов.

1. Постановка проблемы, цели и задачи исследования

В многообразии жанров корпусов текстов особое место занимают корпусы специальных, прежде всего, научных текстов, отражающие знания по отдельным предметным областям. Особенности данных корпусов — наличие жёстких ограничений по типу и тематике текстов, входящих в их состав; формализованность содержания текстов, опирающегося на логико-понятийную схему предметной области; высокая структурированность словаря текстов за счёт насыщенности терминами; очевидное влияние научного стиля на лексико-семантические, морфологические, синтаксические параметры текстов в корпусе [Герд 2005]. Сочетание указанных особенностей специальных текстов делает их хотя и сложным, но всё же весьма привлекательным материалом для исследования. Многие проблемы, возникающие при работе со специальными корпусами текстов, не имеют очевидных и однозначных решений. Таковы вопросы о том, что считать термином той или иной области знаний, как описать, представить значения и связи терминов в терминосистеме, как разработать специальный корпус текстов, как выделить термины из текстов в таком корпусе и др. Следует подчеркнуть, что полное решение данных вопросов выходит за рамки нашего исследования; в процессе работы с терминологией мы используем нестрогое понимание термина как лексической единицы, характерной для некоего текста или множества текстов.

Результаты анализа корпусов текстов, сформированных для отдельных предметных областей, имеют высокую прикладную ценность. Специальные корпусы текстов и извлечённые из них данные востребованы как в научно-технической лексикографии (при составление терминологических словарей, классификаторов, рубрикаторов), так и в сфере автоматической обработки текстов (при автоматическом индексировании и реферировании документов, автоматической классификации и кластеризации документов, в информационном поиске и машинном переводе). На основе специальных корпусов текстов создаются и пополняются терминологические базы и банки данных, терминологические тезаурусы, формальные онтологии для отдельных предметных областей, многоязычные терминологические ресурсы.

Особенно важны исследования специальных корпусов текстов для развивающихся областей знаний, и одной из таких областей является сама кор-

[1] Работа выполнена при частичной финансовой поддержке гранта РГНФ (проект номер 07-04-00161а).

пусная лингвистика. Существуют системные описания терминологии корпусной лингвистики для английского [Baker et al. 2006], а также для ряда других языков, в том числе и славянских: см., например, соответствующий раздел в терминологической базе данных для словацкого языка, разрабатываемой в Институте языкознания Л. Штура (Братислава, Словакия) (URL: https://data.juls.savba.sk/std/) [Levická 2007; Šimková 2006]. Однако в русскоязычных терминологических ресурсах данная предметная область до недавнего времени не была представлена.

С 2002 г. на кафедре математической лингвистики СПбГУ и в ИЛИ РАН осуществляется проект, **целью** которого является создание корпуса русскоязычных текстов по корпусной лингвистике и разработка лингвистических ресурсов на основе данного корпуса. В рамках проекта проводится многоаспектное исследование содержания и структуры текстов в корпусе, что предполагает решение ряда **задач**, среди которых

- извлечение, анализ и систематизация терминологии корпусной лингвистики,
- классификация терминов в корпусе,
- разработка формальной онтологии по корпусной лингвистике,
- тематическая рубрикация текстов в корпусе,
- подготовка данных для компьютерного тезауруса по корпусной лингвистике.

Отдельные результаты работы, полученные к настоящему времени, освещены в ряде публикаций: см., в частности, [Виноградова, Митрофанова, Паничева 2007; Виноградова, Митрофанова 2008; Mitrofanova et al. 2007]. В данной статье обсуждается один из аспектов данного проекта, а именно, проблема автоматизации извлечения терминов, анализа и систематизации терминологии корпусной лингвистики.

2. Исходные лингвистические данные

В состав русскоязычного корпуса текстов по корпусной лингвистике входят тексты различной тематики, отражающие широкий спектр проблем корпусной лингвистики: определение корпусной лингвистики как особой области научной деятельности, противопоставление её другим направлениям лингвистики и языковой инженерии; определение корпуса в соотнесённости с другими типами лингвистических данных; различные аспекты создания и использования корпусов; процедуры, выполняемые при работе с корпусом (разметка, типы разметки, поиск в корпусе); типология корпусов; корпусы текстов с позиций разработчиков и пользователей; взаимодействие корпусов и корпусориентированных лингвистических ресурсов и пр. Ядро корпуса составляют материалы научных конференций по корпусной лингвистике [КЛ и ЛБД 2002, КЛ 2004, КЛ 2006, КЛ 2008], отдельные статьи, учебные пособия, монографии и другие научные материалы. Корпус периодически пополняется новыми документами. Материалы корпуса хранятся в текстовом формате, наряду с этим у разработчиков корпуса существует доступ к файлам с оригинал-макетами. В ходе подготовки текстов статей к размещению в корпусе производится 1) графематический анализ, направленный на выделение и удаление нетекстовых элементов (таблиц, рисунков, формул, гиперссылок, числовых данных и пр.) и иноязычных вкраплений, 2) морфологический анализ (лемматизация, полная морфологическая разметка), 3) метаразметка, которая предполагает фиксацию основных параметров каждой статьи в её паспорте. Наряду с библиографическим описанием эксперты включают в число параметров статьи и наборы из 10 выделяемых вручную терминов-дескрипторов, позволяющих диагностировать тематическую принадлежность текста и проверить данные автоматического анализа. Например:

Текст:

И. С. Николаев, А. С. Герд, И. В. Азарова. «Корпус данных в проекте “Комплексная модель формирования культурного ландшафта и историко-культурной зоны Ингерманландии на Северо-Западе России по данным топонимики”» (*КЛ 2006*).

Набор терминов-дескрипторов:

[*данные, источник, картотека, корпус, культурный, ландшафт, поиск, словарь, топоним, топонимический*]

При формировании наборов терминов-дескрипторов учитывались не только частотность терминов в тексте, но и их содержательный вес. Термины-дескрипторы представлены в нормализованном виде: в наборе присутствует лемма, которая соотносится со входящими в текст словоформами, например: **корпус** (*корпус, корпуса, корпусу, корпусом, корпусе, корпусы, корпусов, корпусам, корпусами, корпусах*) и пр.

Связи терминов-дескрипторов в текстах корпуса исследовались с помощью инструмента автоматической классификации лексики (АКЛ) [Виноградова, Митрофанова, Паничева 2007]. Основным принципом АКЛ является возможность определения содержательной близости лексических единиц при сопоставлении их синтагматических свойств (иначе говоря, их сочетаемости с другими элементами контекста, дистрибуции). Программа АКЛ предусматривает предварительную обработку текстов, представление множества контекстов употребления исследуемых лексем как точек или векторов дистрибуций в N-мерном пространстве, вычисление семантических расстояний между исследуемыми лексемами, кластерный анализ, при котором используются дан-

ные о семантических расстояниях. Чем ближе синтагматические свойства лексем (а стало быть, чем ближе их значения), тем меньше расстояние между векторами их дистрибуций и тем больше вероятность их объединения в один кластер. Сформированные таким образом кластеры лексем допускают дальнейшую лингвистическую интерпретацию. При работе с текстами корпуса по корпусной лингвистике процедуры АКЛ производились в двух режимах: структурирование терминов-дескрипторов в наборах и выявление классов условной эквивалентности для каждого из терминов-дескрипторов.

В ходе экспериментов производилась иерархическая кластеризация терминов-дескрипторов в наборах для каждой из статей в корпусе; в качестве меры расстояния использовался косинус угла между векторами дистрибуций (*Cos*). Результаты кластеризации выводятся в виде многоуровневого списка слов в виде скобочной записи, которая отражает последовательность объединения терминов-дескрипторов в кластеры. Наряду с этим пользователь получает данные о частотности исследуемых лексем в обрабатываемом тексте, а также значения расстояний во всевозможных парах лексем из анализируемого набора. Например:

Текст:
Е. Л. Алексеева, А.М. Лаврентьев, И. В. Азарова, Л. А. Захарова «Разметка корпуса древнерусских агиографических текстов» (*КЛ 2004*)

Кластерная структура набора терминов-дескрипторов:
[*корпус*, *разметка*] *Cos* = 0,375

[*агиографический*, *русский*] *Cos* = 0,284

[*житие*, *текст*] *Cos* = 0,277

[[*агиографический*, *русский*] [*житие*, *текст*]] *Cos* = 0,259

[[*корпус*, *разметка*] [[*агиографический*, *русский*] [*житие*, *текст*]]] *Cos* = 0,251

[*представление* [[*корпус*, *разметка*] [[*агиографический*, *русский*] [*житие*, *текст*]]]] *Cos* = 0,219

[[*представление* [[*корпус*, *разметка*] [[*агиографический*, *русский*] [*житие*, *текст*]]]] *электронный*] *Cos* = 0,258

[*рукопись* [[*представление* [[*корпус*, *разметка*] [[*агиографический*, *русский*] [*житие*, *текст*]]]] *электронный*]] *Cos* = 0,171

[*словоформа* [*рукопись* [[*представление* [[*корпус*, *разметка*] [[*агиографический*, *русский*] [*житие*, *текст*]]]] *электронный*]]] *Cos* = 0,138

Абсолютные частоты терминов-дескрипторов:
агиографический (*f* = 4), *житие* (*f* = 13), *русский* (*f* = 7), *текст* (*f* = 47), *корпус* (*f* = 8), *электронный* (*f* = 8), *рукопись* (*f* = 15), *словоформа* (*f* = 15), *представление* (*f* = 7), *разметка* (*f* = 5)

С помощью программы АКЛ для каждого из терминов-дескрипторов в наборах производится автоматическое формирование классов условной эквивалентности, включающих слова с близкой дистрибуцией в тексте. Близость дистрибуции также оценивается на основе значений *Cos*. Например:

Текст:
В. П. Захаров. Корпусная лингвистика *(Захаров 2005)*

Классы условной эквивалентности термина-дескриптора *разметка* (объем классов — 20 слов):

Обработка текста с лемматизацией		**Обработка текста без лемматизации**	
РАЗМЕТКА	Cos	***разметка***	*Cos*
ПРОСОДИЧЕСКИЙ	0,375	*просодическая*	0,362
БОЛЬШИНСТВО	0,288	*фиксирует*	0,285
АНАФОРИЧЕСКИЙ	0,288	*документа*	0,280
ВВОДИТЬСЯ	0,252	*абзацев*	0,280
ДОКУМЕНТ	0,251	*выделение*	0,279
ВЫДЕЛЕНИЕ	0,250	*местоименные*	0,271
МНОЖЕСТВО	0,240	*референтные*	0,270
ИНТОНАЦИЯ	0,226	*предложений*	0,265
РЕФЕРЕНТНЫЙ	0,214	*annotation*	0,255
РЕАЛЬНО	0,213	*анафорическая*	0,254
УДАРЕНИЕ	0,212	*разговорной*	0,253
РАЗ	0,198	*структурная*	0,251
МЕСТОИМЕННЫЙ	0,198	*корпусах*	0,250
ИНОСТРАННЫЙ	0,197	*просодических*	0,230
УПОТРЕБЛЯТЬСЯ	0,196	*интонацию*	0,224
НАЛИЧИЕ	0,185	*частеречная*	0,210
ДОСЛОВНО	0,180	*ударение*	0,207
ОГОВОРКА	0,167	*описывающие*	0,189
ПОВТОР	0,167	*оказаться*	0,168

По-видимому, последовательность формирования кластеров терминов-дескрипторов, а также состав выделенных для них классов условной эквивалентности отражает важнейшие парадигматические и синтагматические связи элементов исследуемых текстов. Тем самым, в процессе создания модели предметной области корпусной лингвистики производится обобщение выявленных связей терминов-дескрипторов до родовидовой иерархии понятий. В целях уточнения характера связей между понятиями, выраженными исследуемыми терминами, была проведена отдельная серия экспериментов. Процедуры отбора и кластеризации дескрипторов, характеризующих корпусную лингвистику, позволяют перейти с терминологического уровня на онтологический и сформировать упорядоченное множество категорий, которые необходимо вклю-

чить в формальную онтологию рассматриваемой области знаний. Формальная онтология по корпусной лингвистике относится к классу терминологических онтологий [Sowa]. В качестве представителей онтологических категорий были отобраны те из терминов-дескрипторов, которые оказались релевантны не только для отдельных текстов, но для предметной области в целом, обладают наибольшей частотой, попадают в ядра полученных кластеров, соответствуют исходным понятиям, выделенным на основе экспертных описаний. Всего было зарегистрировано 335 различных терминов-дескрипторов. Вероятно, такие термины-дескрипторы, как *корпус*, *текст*, *данные*, *разметка*, *тег*, *поиск*, *слово*, *лемма*, *словоформа*, *контекст* и пр. представляют понятийное ядро предметной области.

- *Предметная область «Корпусная лингвистика»*
- *корпус данных*
  - *корпус текстов*
  - *тип корпуса*
    - *работа с корпусом*
      - *разработка корпуса*
        - *отбор данных*
        - *оцифровка данных*
        - *разметка корпуса*
        - *корпус-менеджер*
          - *использование корпуса*
            - *поиск по корпусу*
              - *запрос к корпусу*
                - *терминальная цепочка символов*
                - *регулярное выражение*
                - *лемма*
                - *тег*
              - *результат работы с корпусом*
                - *конкорданс*
                - *контекст*
                - *словоуказатель*
                - *статистика*

Формальная онтология по корпусной лингвистике реализована в онторедакторе Protégé [Виноградова, Митрофанова 2008]. Выше приведены важнейшие категории формальной онтологии, упорядоченные в иерархию.[2] В отдельных полях формальной онтологии даются общепринятые дефиниции терминов-дескрипторов, фиксируются синонимические отношения между терминами-дескрипторами (например, *разметка*, *аннотация*, *аннотирование* и пр.). Кроме того, каждая категория формальной онтологии имеет атрибут *тексты*. Этот атрибут необходим для того, чтобы формальная онтология могла быть использована для тематической рубрикации документов из русскоязычного корпуса текстов по корпусной лингвистике. В качестве экземпляров данного атрибута приведены библиографические сведения о тех статьях из корпуса, в которых встретились термины-дескрипторы, соответствующие онтологическим категориям. Например:

Категория: *алгоритм*
Тексты: *П. Макагонов, М. Александров, А. Гельбух* «Формулы проверки подобия слов с обучением на примерах: построение и применение» (*КЛ 2004*); *К. Р. Пиотровская, Р. Г. Пиотровский, Ю. В. Романов* «Вторая когнитивная революция — инженерная и корпусная лингвистика» (*КЛ и ЛБД 2002*).

Тем самым, применение формальной онтологии предметной области корпусной лингвистики при работе с соответствующим корпусом текстов должно повысить эффективность поиска данных.

С расширением русскоязычного корпуса текстов должно происходить пополнение списка уже зарегистрированных терминов и обновление существующей формальной онтологии, на основе которой в дальнейшем планируется создание тезауруса по корпусной лингвистике. В связи с этим было принято решение изучить возможности частичной автоматизации терминологической работы и затем оптимизировать процедуру обработки документов из корпуса текстов по корпусной лингвистике.

3. Методы и инструменты анализа терминологии

Существует три основных класса методов извлечения терминологии из специального корпуса текстов: лингвистические методы, статистические методы и комбинированные методы.

Лингвистические методы в основном предполагают ручную обработку документов в специальном корпусе текстов, в ходе которой эксперты выявляют выражения, рассматриваемые как предполагаемые однословные термины и терминосочетания. Для выделения терминосочетаний рекомендуется использовать лексико-грамматические шаблоны однословных и неоднословных терминов. Целесообразно также использовать систему фильтров (стоп-словарь) для отсеивания нетерминов.

Применение статистических методов опирается на представление о том, что термины, как правило, это наиболее частотные слова и словосочетания, встречающиеся в специальных текстах и выражающие понятия предметной области. Терминосочетания обычно соотносятся с *n*-граммами (двух-, трех-, четырехчленными сочетаниями), характеризуются высокой степенью устойчивости. В качестве мер, пригодных для оценки устойчивости словосочетаний в специальных текстах, следует упомянуть *MI-score*, *t-score*, *Log-Likelihood*, *C-value*, критерий χ^2 и ряд других.

[2] В рамках данной статьи не ставится задача полного описания иерархии категорий формальной онтологии в силу её объёмности.

Во многих исследованиях, проводимых для русского и других славянских языков (см., например: [Браславский, Соколов 2006, 2007, 2008; Добров и др. 2003; Kupść 2007; Urbańska, Piechociński 2007] и др.) практикуется комбинированный подход, заключающийся в (полу)автоматической обработке специальных корпусов текстов. Комбинированные методы анализа терминологии предполагают совместное использование аппарата лексико-грамматических шаблонов, методов сборки терминосочетаний, системы фильтров, а также статистического аппарата.

Сочетание лингвистических и статистических приемов анализа документов в корпусе применяется в автоматизированной лексикографической среде Alex+ [Сидорова 2008(а), 2008(б)]. Alex+ представляет собой технологический комплекс для создания и поддержки предметно-ориентированных словарей, позволяющий выделять термины и терминосочетания из текстов по лексико-грамматическим шаблонам, получать статистические данные о встречаемости терминов и терминосочетаний в обрабатываемых текстах, автоматически пополнять словарь на основе обучающей выборки. В состав комплекса Alex+ входят модуль морфологического анализа системы Диалинг, модуль сборки терминосочетаний по шаблонам, модуль просмотра конкорданса, модуль тематизации, модуль выявления стоп-слов. Преимущества подготовки словарей в системе Alex+ залючаются в возможности разнообразного наполнения словарей, допускающих включение однословных и неоднословных терминов, в возможности представления нескольких типов данных о терминах (терминообразующие признаки, семантические признаки — соотнесенность с понятиями в иерархии классов, статистические признаки) и др. В Alex+ допускается построение формальной онтологии (или задание иерархии тем) параллельно со словарем, при этом словарь и иерархия тем могут применяться для автоматической классификации текстов. Существует также возможность обработки несловарных словоформ и др. Тем самым, параметры автоматизированной лексикографической среды Alex+ соответствуют целям обсуждаемого исследования, в связи с чем некоторые функции данного комплекса были задействованы при анализе терминологии в русскоязычном корпусе текстов по корпусной лингвистике.

4. Описание однословных и неоднословных терминов с помощью лексико-грамматических шаблонов

В ходе анализа однословных терминов и терминосочетаний были применены лексико-грамматические шаблоны (ср. морфологические шаблоны [Сидорова 2008(а), 2008(б)], лексико-синтаксические шаблоны [Большакова и др. 2007; Васильева 2004; Рабчевский и др. 2008]). Лексико-грамматические шаблоны служат для описания классов языковых выражений. В отдельном лексико-грамматическом шаблоне указываются существенные характеристики множества лексем, которые входят в языко вые выражения, принадлежащие классу, также приводятся возможные морфологические формы лексем и, при возможности, синтаксические условия употребления языкового выражения, построенного в соответствии с шаблоном (например, правила согласования морфологических признаков лексем).

Лексико-грамматические шаблоны были задействованы при выделении однословных и неоднословных терминов в автоматизированной лексикографической среде Alex+ [Сидорова 2008(а), 2008(б)].

Например, в результате обработки текста [Захаров 2005] с последующим отсеиванием стоп-слов (служебных слов, местоимений, числительных и др.), а также слов, не являющихся терминами (например, *миро*), в списке однословных терминов можно обнаружить существительные, прилагательные, глаголы:**N**: *выборка, выдача, данные, грамматика, документ, единица, жанр, запрос, инструмент, классификация, кодирование, лемма, массив, метаданные, метка, морфология, неоднозначность, поиск, пользователь, разметка, репрезентативность, составитель, текст, частота* и др.;

Adj: *автоматизированный, информационно-поисковый, корпусной, корпусный, лингвистический* и др.;

V: *автоматизировать, размечать* и др.

Среди неоднословных терминов обнаружены словосочетания, соответствующие следующим основным лексико-грамматическим шаблонам:

Adj+N: *автоматизированная система, автоматическая обработка / разметка / система, автоматический анализ / режим, анафорическая / морфологическая / семантическая / синтаксическая / структурная / просодическая разметка, совместная встречаемость, программное обеспечение, формальный язык, языковой корпус, языковая единица* и др.;

Adj+N+N: *автоматическая обработка текста, компьютерная база данных, компьютерная модель языка, лингвистический корпус текстов, представление корпуса текстов, формальный язык разметки* и др.;

N+Adj+N: *банк синтаксических структур, массив языковых данных, обработка типовых запросов* и др.;

N+Prep+Adj+N: *корпус с синтаксической разметкой, тексты на естественном языке, тексты на машинном носителе* и др.;

N+Prep+N: *доступ к корпусу, наука о языке, поиск в корпусе, сведения об авторе* и др.;

N+Prep+N+N: *поиск с указанием контекста* и др.;

N+N: *обучение языку, база данных, массив данных / текстов, вид разметки, источник данных, кодирование информации, корпус данных / текстов, модель языка, параметр разметки / кодирования / текста, разметка корпуса / документа / текста, размер корпуса, распознавание речи, тип корпуса / данных / разметки / текста, формат выдачи / данных* и др.;

N+N+N: *вывод результатов поиска, стандарт представления метаданных / данных* и др.;

Adj+Adj+N: *устная разговорная речь* и др.

Справедливо будет отметить, что данные словосочетания различаются не только по степени сложности (двух-, трёх-, четырёхкомпонентные терминосочетания), но также по устойчивости (особенно это касается трёх- и четырёхкомпонентных сочетаний, которые сами по себе содержат однословные термины и двухкомпонентные терминосочетания). Для определения устойчивости сочетаний также необходимо обращаться к статистическим критериям [Браславский, Соколов 2006, 2007, 2008; Добров и др. 2003; Захаров, Хохлова 2008; Чанышев 2008; Khokhlova 2008]. Самый важный вопрос, возникающий при анализе массивов однословных и неоднословных терминов — это вопрос об оценке степени терминологичности рассматриваемых единиц. Один из путей — определение индекса специфичности для данной совокупности текстов [Шайкевич 2003]. Решающее слово, вместе с тем, остаётся за специалистами-терминоведами и — в нашем случае — за экспертами в области корпусной лингвистики.

5. Описание терминообразующих контекстов с помощью лексико-грамматических шаблонов

Расширенные лексико-грамматические шаблоны успешно используются для выявления и описания терминообразующих контекстов. Терминообразующие контексты, как правило, содержат термин и его толкование, синонимы, переводные эквиваленты и т.д., при этом в контексте существуют определенные маркеры, позволяющие опознать сам термин и связанную с ним информацию.

Структура и типовое наполнение контекстов, содержащих толкования терминов, могут быть представлены, например, в следующих лексико-грамматических шаблонах:

NP(term) <*понимать/пониматься*> NP(def):

Под репрезентативностью понимается необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т.п. [Захаров 2005];

NP(def) <*называть/называться/иметь название*> NP(term):

Это кодирование информации ***имеет название*** *метаразметка…* [Захаров 2005];**NP(term) <*заключаться в*> NP(def):**

Разметка (tagging, annotation) ***заключается в*** *приписывании текстам и их компонентам специальных меток (tag, tags): внешних, экстралингвистических (сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика; сведения об авторе могут включать не только его имя, но также возраст, пол, годы жизни и многое другое* [Захаров 2005];

NP(term) <*представлять собой*> NP(def):

…устойчивые словосочетания ***представляют собой*** *с семантической точки зрения неделимую смысловую единицу…* [Захаров 2005].

Контексты, выражающие различные отношения между терминами, могут быть обобщены, например, в следующих лексико-грамматических шаблонах:

NP(term) <*, или*> NP(term) (синонимия):

…синтаксического анализа, ***или*** *парсинга…* [Захаров 2005];

NP(term) <являться результатом> NP(term) (отношение «процесс — результат»):

…синтаксическая разметка, ***являющаяся результатом*** *синтаксического анализа, или парсинга (англ. parsing)…* [Захаров 2005];

NP(term) <*обеспечивать*> NP(term) (отношение «объект — назначение»):

…конвертирование размеченных текстов в структуру специализированной лингвистической информационно-поисковой системы (corpus manager), ***обеспечивающей*** *быстрый многоаспектный поиск и статистическую обработку…* [Захаров 2005];**NP(term) <*включать (в себя)*> NP(term)** (количественные, гипонимические, мереологические, импликативные и др. отношения):

количественные отношения: *Корпусы нового поколения* ***включают*** *сотни миллионов слов, поэтому выдвигаются принципы разработки систем, которые бы минимизировали вмешательство человека* [Захаров 2005];

гипонимические отношения: *Метаописание текстов корпуса* ***включает*** *как содержательные элементы данных (библиографические данные, признаки, характеризующие жанровые и стилевые особенности текста, сведения об авторе), так и формальные (имя файла, параметры кодирования, версия языка разметки, исполнители этапов работ)* [Захаров 2005].

Тем самым, анализ терминообразующих контекстов способствует установлению системных связей терминов в терминосистеме, что позволяет

уточнять состав словника и пополнять блок дефиниций терминологического тезауруса.

В блок дефиниций тезауруса включаются толкования стандартных и авторских терминов, зафиксированные в текстах корпуса (как в экспертных, так и в исследовательских описаниях) или в других источниках энциклопедического характера. Вместе с тем, «готовые» толкования удается подобрать лишь к наиболее распространённым терминам, для остальных необходимо составлять дефиниции, и в подобных случаях обращение к лексико-грамматическим шаблонам также весьма уместно, так как это позволяет сохранить единообразие структуры толкований.

В дальнейшем при решении задач поиска в корпусе текстов и автоматизированного пополнения формальной онтологии возможно использование специализированного языка для записи лексико-грамматических шаблонов, например, языка LSPL (Lexical-Syntactic Pattern Language) [Большакова и др. 2007; Васильева 2004; Рабчевский и др. 2008].

6. Итоги исследования и направления дальнейшей работы

В ходе исследования были оценены возможности различных стратегий автоматизации работ по извлечению и систематизации терминологии из русскоязычного корпуса текстов по корпусной лингвистике.

Применение инструмента АКЛ, реализующего процедуры кластерного анализа в двух режимах, позволило выявить структурную организацию терминов-дескрипторов в корпусе текстов по корпусной лингвистике. Полученные данные легли в основу формальной онтологии предметной области, охватывающей базовые понятия и термины корпусной лингвистики.

Пополнение базового списка терминов и формирование списка терминосочетаний успешно проведено с помощью автоматизированной лексикографической среды Alex+. Проанализированы основные лексико-грамматические шаблоны для однословных и неоднословных терминов, встречающихся в текстах корпуса. Аппарат лексико-грамматических шаблонов также использовался в изучении структуры терминообразующих контекстов.

Результаты, полученные на нынешнем этапе работы, будут использованы при разработке тезауруса по корпусной лингвистике. Данный лингвистический ресурс планируется включить в состав портала знаний по компьютерной лингвистике, создаваемого коллективом российских учёных (Москва, Новосибирск, Санкт-Петербург) [Соколова и др. 2008].

Литература

1. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007. URL: http://www.dialog-21.ru/dialog2007/materials/html/11.htm
2. *Браславский П. И., Соколов Е. А.* Автоматическое извлечение терминологии с использованием поисковых машин интернета // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007. URL: http://www.dialog-21.ru/dialog2007/materials/html/14.htm
3. *Браславский П. И., Соколов Е. А.* Сравнение пяти методов извлечения терминов произвольной длины // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008. URL: http://www.dialog-21.ru/dialog2008/materials/html/11.htm
4. *Браславский П. И., Соколов Е. А.* Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2006». М.: 2006. URL: http://www.dialog-21.ru/dialog2006/materials/html/Braslavski.htm
5. *Васильева Н. Э.* Шаблоны употреблений терминов и их использование при автоматической обработке научно-технических текстов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2004». М.: 2004. URL: http://www.dialog-21.ru/Archive/2004/Vasiljeva.htm
6. *Виноградова Н. В., Митрофанова О. А.* Формальная онтология как инструмент систематизации данных в русскоязычном корпусе текстов по корпусной лингвистике // Труды международной конференции «Корпусная лингвистика — 2008». СПб.: 2008.
7. *Виноградова Н. В., Митрофанова О. А., Паничева П. В.* Автоматическая классификация терминов в русскоязычном корпусе текстов по корпусной лингвистике // Труды девятой Всероссийской научной конференции «Электронные

библиотеки: Перспективные методы и технологии, электронные коллекции» (RCDL–2007). Переславль-Залесский: 2007. URL: http://www.rcdl.ru//papers/2007/paper_31_v1.pdf

8. *Герд А. С.* Язык науки и техники как объект лингвистического изучения // А.С. Герд. Прикладная лингвистика. СПб.: 2005.
9. *Добров Б. В., Лукашевич Н. В., Сыромятников С. В.* Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой Всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции» (RCDL–2003). СПб.: 2003. URL: http://www.cir.ru/docs/ips/publications/2003_rcdl_thes_creation.pdf
10. *Захаров В. П.* Корпусная лингвистика / Учебно-методическое пособие. СПб.: 2005.
11. *Захаров В. П., Хохлова М. В.* Статистический метод выявления коллокаций // Языковая инженерия в поиске смыслов. XI Всероссийская объединенная конференция «Интернет и современное общество». Санкт-Петербург: 2008.
12. *КЛ и ЛБД 2002* — Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». СПб.: 2002.
13. *КЛ 2004* — Труды международной конференции «Корпусная лингвистика — 2004». СПб.: 2004.
14. *КЛ 2006* — Труды международной конференции «Корпусная лингвистика — 2006». СПб.: 2006.
15. *КЛ 2008* — Труды международной конференции «Корпусная лингвистика — 2008». СПб.: 2008.
16. *Рабчевский Е. А., Булатова Г. И., Шарафутдинов И. М.* Формализм записи лексико-синтаксических шаблонов в задаче автоматизации процесса построения онтологий // Труды десятой Всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции» (RCDL–2008). Дубна: 2008. URL: http://rcdl2008.jinr.ru/pdf/103_106_paper10.pdf
17. *Сидорова Е. А.* Многоцелевая словарная подсистема извлечения предметной лексики // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008(а). URL: http://www.dialog-21.ru/dialog2008/materials/html/74.htm
18. *Сидорова Е. А.* Подход к построению предметных словарей по корпусу текстов // Труды международной конференции «Корпусная лингвистика–2008». СПб.: 2008(б).
19. *Соколова Е. Г., Кононенко И. С., Загорулько Ю. А.* Проблемы описания компьютерной лингвистики в виде онтологии для портала знаний // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008. URL: http://www.dialog-21.ru/dialog2008/materials/html/75.htm
20. *Чанышев О. Г.* Автоматическое построение терминологической базы знаний // Труды десятой Всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции» (RCDL–2008). Дубна: 2008. URL: http://rcdl2008.jinr.ru/pdf/085_092_paper8.pdf
21. *Шайкевич А. Я.* Статистический словарь языка Достоевского. Введение. 2003. URL: http://nature.syktsu.ru/cfrl/dost_cd0/introdw.htm
22. *Backer P., Hardie A., McEnery T.* A Glossary of Corpus Linguistics. Edinburgh University Press: 2006.
23. *Khokhlova M.* Extracting Collocations in Russian: Statistics vs. Dictionary // Proceedings of 9th International Conference on Textual Data Statistical Analysis (JADT 2008). Lyon: 2008.
24. *Kupść A.* Extraction automatique de termes à partir de textes polonais // TALN 2007. Toulouse: 2007. URL: http://llf.linguist.jussieu.fr/llf/Gens/Kupsc/kupsc-taln07.pdf
25. *Levická J.* Terminology and Treminological Activities in the Present-Day Slovakia // Computer Treatment of Slavic and East European Languages: Fourth International Seminar. Bratislava, Slovakia, 25–27 October 2007. Proceedings. Bratislava: 2007.
26. *Mitrofanova O., Panicheva P., Savitsky V.* Automatic Word Clustering in Russian Texts based on Latent Semantic Analysis // Computer Treatment of Slavic and East European Languages: Fourth International Seminar. Bratislava, Slovakia, 25–27 October 2007. Proceedings. Bratislava: 2007.
27. *Šimková M.* Výberový slovník termínov z počítačovej a korpusovej lingvistiky. 2006. URL: http://korpus.juls.savba.sk/publications/block1/2006-simkova-vyberovy%20slovnik%20terminov/2006-simkova-vyberovy%20slovnik%20terminov.pdf
28. *Sowa J. F.* Building, Sharing, and Merging Ontologies. URL: http://www.jfsowa.com/ontology/ontoshar.htm
29. *Urbańska D., Piechociński D.* Automatic Term Recognition in Polish Texts // Computer Treatment of Slavic and East European Languages: Fourth International Seminar. Bratislava, Slovakia, 25–27 October 2007. Proceedings. Bratislava: 2007.

Опыт создания корпусов дагестанских языков[1]

An experience of creation of the national corpus of Dagestan languages

Муталов Р. О. (mutalovr@mail.ru)

Дагестанский государственный университет,
Махачкала, Республика Дагестан

В докладе рассматриваются проблемы и перспективы национальных корпусов текстов шести литературных языков Дагестана, работы по созданию которых ведутся в Дагестанском государственном университете. Особое внимание уделено проблемам создания системы автоматической разметки текстов и переводу печатных текстов в электронный формат.

Корпусная лингвистика в настоящее время является одной из быстроразвивающихся областей компьютерной лингвистки. Созданы корпуса текстов большинства распространенных языков мира; по многим языкам создаются параллельные корпуса, диалектные корпуса, корпуса устной речи, корпуса поэтической речи и т.д. Однако, как в России, так и за рубежом, работы по созданию корпусов текстов языков малочисленных народов находятся пока на начальной стадии. Исследовательской группой под руководством А.Е.Кибрика разработаны конкретные требования к современным корпусам текстов для малых языков [Кибрик и др. 2007]. Высокий уровень результатов достигнут при составлении корпусов хиналугского и алюторского языков, электронной грамматики и электронного словаря арчинского языка.

В дагестанском государственном университете ведутся работы по созданию Национальных корпусов шести литературных языков Дагестана — аварского, даргинского, лезгинского, лакского, кумыкского, табасаранского [Муталов 2007]. Разработку национальных корпусов дагестанских языков можно отнести к первым опытам создания корпусов текстов малых языков России. Необходимость скорейшего ввода в информационное поле материала данных языков заключается в том, что эти языки отнесены специалистами к языкам, обреченным в будущем на исчезновение; поэтому одной из важных задач создания корпусов является содействие в решении проблемы сохранения и развития национальных языков Дагестана. Следует также отметить, что дагестанские языки, обладая сложной морфологической системой и являясь языками эргативного типа, представляют для лингвистической типологии особый интерес. Тексты с лингвистической разметкой станут базой для создания современных научных грамматик, словарей и учебников. Корпуса текстов в перспективе станут также основой для разработок параллельных русско-дагестанских корпусов.

Основную часть текстов национальных корпусов дагестанских языков — до 75%, будут составлять художественные произведения дагестанских авторов. Будут также представлены и другие тексты: драматургия, мемуарно-биографическая литература, журнальная публицистика и литературная критика, газетная публицистика, научные, учебные, религиозные, юридические тексты, деловые и бытовые тексты. Они должны стать базой создания национальных корпусов. Предполагается, что объем создаваемых корпусов будет составлять от 3 до 5 млн. словоупотреблений по каждому языку.

Создателям корпусов приходится сталкиваться с трудностями, связанными как с малым количеством текстов в электронном формате, так и отсутствием механизмов разметки текстов каждого языка. Создатели корпусов освоили навыки оцифровки текстов, их сканирования и последующей обработки, вычитки и исправления ошибок, овладели технологией создания корпусов текстов. Работа по созданию электронных коллекций текстов ведется в двух направлениях: сбор имеющих в электронном виде текстов и перевод в электронный формат печатных текстов. Собраны воедино тексты, уже имеющие электронный формат; это, как правило, произведения авторов последних лет, созданные в компьютерную эру. Одна-

[1] Работа поддержана РФФИ, грант № 07-06-00460-а.

ко подавляющее большинство текстов на дагестанских языках написано в 60–90-е годы прошлого века и представлено в печатном виде. Трудоемкую техническую работу по их сканированию удалось ускорить посредством применения высокоскоростного планшетного сканера А3, позволяющего сканировать, помимо стандартных изданий, широкоформатные литературные журналы-альманахи. Для распознавания отсканированных текстов была применена последняя, девятая версия FineReader. Распознаванию сложных графем современной графики дагестанских языков, состоящих из нескольких, порой 3–4 знаков, способствовало применение созданных для решения данной проблемы специальных шрифтов. При вычитке текстов особое внимание обращалось удалению в словах знаков переноса и устранению «жестких» концов строк. Создаются электронные архивы текстов, представляющих собой отсканированные, но не обработанные тексты. Вычитка данных текстов и исправление имеющиеся в них ошибок и опечаток составляют значительную по объему часть работы.

В составлении инвентаря грамматических помет, созданного на основе латинской графики, и в сокращениях, принятых для обозначения определенных грамматических категорий, создатели корпусов, хотя и следовали рекомендациям известных «Лейпцигских правил глоссирования», значительно дополнили и уточнили список морфологических категорий, встречаемых в дагестанских языках.

Проведены работы по созданию электронных библиотек и электронных словарей дагестанских языков, что в комплексе должно стать основой для проведения грамматической разметки текстов. Подготовлена также информация, необходимая для проведения метаразметки текстов — собраны сведения об авторах, внешних параметрах текстов, проведена их типизация.

Поскольку наиболее важной частью создания корпусов является разметка текстов, основное внимание уделяется разработке механизмов лингвистической разметки текстов. Разметка текстов малых языков, как в России, так и за рубежом, делается большей частью вручную. По корпусам больших языков, количество словоупотреблений которых составляет несколько сот миллионов слов, разрабатываются специальные парсеры. Для создания парсеров по каждому корпусу дагестанских языков с объемом словоупотреблений в 3–5 млн. словоупотреблений, аналогичных парсеру Национального корпуса русского языка или корпусов других языков, естественно, не хватает ни людских, ни финансовых ресурсов. С другой стороны, не представляется также реальным проводить ручную разметку всех текстов корпуса. Поэтому была предпринята попытка создания собственной, упрощенной системы автоматической разметки текстов.

Для начала были созданы электронные словари, содержащие функционирующие в языке основные лексемы. Исходя из лексического значения и параметров словоизменения, все слова были распределены на несколько групп. В одну группу объединялись слова с полностью идентичными грамматическими признаками и близкие по семантике. Каждой словоформе приписываются следующие морфологические значения: исходная форма слова; принадлежность к той или иной части речи; семантическая группа. Затем даются словоизменительные признаки словоформы; для именных частей речи и наречия — это информация о классе, числе и падеже. Например, даргинскому слову *дудешлис* «отцу» приписывается значение мужского класса, дательного падежа, единственного числа. Для глаголов указывается информация о классе, числе, лице, виде, переходности, времени, наклонении; для отглагольных образований — причастия, деепричастия (деепричастия места) и масдара, помимо перечисленных признаков, указывается также и падеж. Здесь же дается информация о нестандартных словоформах; таковы, например, даргинские глаголы *гес* «дать», *хес* «принести», *кес* «привести».

Разработанная специально программа заменяет словоформу в тексте другой словоформой, имеющей морфологическую и семантическую разметку. При запросе нужного слова появляются предложения с данной словоформой, имеющими метаразметку, а морфологические и семантические значения слова можно извлечь при нажатии курсора мыши на словоформу — всплывают ее словоклассифицирующие и словоизменительные признаки. Здесь же появляется также информация о принадлежности слова к той или иной семантической группе.

Естественно, при автоматической разметке текстов и приписывании каждой словоформе морфологической информации возникает ряд проблем. К примеру, одна из них связана с омонимичностью классно-числовых показателей дагестанских языков. Следует отметить, что среди шести классных индикаторов в даргинском языке имеются три пары омонимичных показателей: классификатор *б* служит для обозначения среднего класса единственного числа и 3-го лица множественного числа мужского и женского классов; показатель *д* служит классификатором 1 и 2-го лиц множественного числа непереходных глаголов и 3-го лица среднего класса множественного числа. Омонимию классно-числовых показателей разметчик должен снимать вручную. В ряде дагестанских слов на современном этапе развития языков классные показатели окаменели; информация о выражении ими значения грамматического класса в таких случаях не дается.

При указании информации по категории числа имен существительных внимание обращается на группы слов, имеющих лишь форму единственного или множественного числа. Существительные, всегда функционирующие в форме единственного числа, указываются как слова «без формы множе-

ственного числа». Существительные же, имеющие лишь форму множественного числа, а также «собирательные существительные» указываются как существительные «нерасчлененной совокупности».

Хотя при разметке возникают проблемы, связанные со сложностью морфологической структуры дагестанских языков, или проблемы, создаваемые орфографическими правилами, (такими, как, например, слитное написание некоторых слов служебных частей речи с предшествующим словом), применение механизма автоматической разметки текстов для создания корпусов малых языков представляется обоснованным и эффективным. Она позволяет решать первостепенные задачи поиска лингвистической информации о слове. Система автоматической разметки текста была применена к нескольким текстам различных дагестанских языков, после чего пробные образцы размеченных текстов были перенесены на сервер Лаборатории лингвистических исследований ДГУ.

В перспективе предполагается продолжить работы по пополнению имеющихся электронных библиотек новыми текстами, а также по переводу в электронный формат печатных текстов. Оцифрованные тексты будут вычитаны, имеющиеся ошибки исправлены. Предстоит усовершенствовать и доработать систему автоматической разметки текстов. Будут продолжены работы по морфологической, семантической и экстралингвистической разметке текстов, а также ручное снятие омонимии. Размеченные подобным образом национальные корпуса и электронные словари дагестанских языков предполагается разместить в режиме открытого доступа в Интернет-сети.

Литература

1. *Кибрик А.Е., Архипов А. В., Даниэль М. А., Кодзасов С. В., Майерс Том, Нахимовский А. Д.* Технологии обработки языковых данных в документировании малых языков // М.: Материалы Международной конференции «ДИАЛОГ 2007» «Компьютерная лингвистика и интеллектуальные технологии», 2007.
2. *Муталов Р. О.* Корпусная лингвистика и перспективы ее развития в Дагестане // Махачкала: Современные проблемы кавказского языкознания, 2007. Вып. 7, С. 160–173.
3. *Плунгян В. А.* Зачем нужен Национальный корпус русского языка? неформальное введение. // М.: Национальный корпус русского языка: 2003–2005. Результаты и перспективы, 2005.

Разметка кореференции на синтаксически аннотированном корпусе чешских текстов[1]

Coreference annotation in Prague dependency treebank

Недолужко А. (nedoluzko@ufal.mff.cuni.cz)

Карлов университет, Прага, Чехия

В докладе представлена схема разметки кореференции на синтаксически аннотированном корпусе чешских текстов PTD. Рассматриваются три этапа разметки — разметка грамматической кореференции, где антецедент высчитывается на основе грамматических правил данного языка, разметка прономинальной текстовой кореференции и расширенная схема разметки именной текстовой кореференции и ассоциативной анафоры. Разметка грамматической и прономинальной кореференции была проделана на всем корпусе PTD, разметкой именной кореференции и ассоциативной анафоры занимается автор данного доклада в настоящее время. В докладе рассматриваются некоторые трудности классификации примеров, приводятся первые результаты.

1. Общие сведения

Синтаксически аннотированный корпус чешского языка (PDT) — это проект лингвистической разметки текстов, разрабатываемый в Институте формальной и прикладной лингвистики физико-математического факультета Карлова университета в Праге. Разметка проводится частично автоматически на трех уровнях — морфологическом, поверхностно-синтаксическом и глубинно-синтаксическом (подробнее с проектом можно ознакомиться Hajičová 2006, Недолужко 2007). В данном докладе речь пойдет о разме тке кореференции, реализуемой вручную и частично автоматически на глубинно-синтаксическом уровне.

В настоящее время аннотирование кореференции проводится с различной степенью подробности в большинстве синтаксически размеченных корпусов. Прономинальная кореференция представлена в американском PennTreebank (http://www.cis.upenn.edu/~treebank), концепции разметки именной кореференции представлены в проектах MUC-7 (Hirschman, 1998), MATE (Poesio, 2004), DRAMA (Passonneau, 1997), PoCoS (Chiarcos, Krasavina 2005), аннотация ассоциативной анафоры проводится в рамках проектов GNOME (на основе MATE), DRAMA, планируется в PoCoS и т. д.

В аннотации PDT 2.0 кореференция делится на грамматическую и текстовую. Кроме того, аннотируется т. наз. ассоциативная анафора (bridging) и некоторые особые случаи (экзофорическая отсылка и отсылка к большему, чем одно предложение, сегменту текста). Для аннотирования кореференции используется `id` антецедента, к которому отсылает `id` узла анафоры. Разметка кореференции приводилась в три этапа. Первый этап — разметка грамматической кореференции (см. 2), второй этап — разметка т.наз. текстовой прономинальной кореференции (см. 3), третий этап состоит из разметки именной кореференции и ассоциативной анафоры (см. 4). Далее будут представлены эти три этапа с особым акцентом на последний, которым автор доклада занимается в настоящее время.

2. Разметка грамматической кореференции

В случае грамматической кореференции антецедент высчитывается на основе грамматических правил языка. Грамматическая кореференция практически никогда не переходит границ предложения, ее всегда можно представить как отсылку одного узла к другому, следовательно ее аннотирование легко автоматизируется. К грамматической кореференции относится:

[1] эта работа была поддержана грантом GACR 405/09/0729

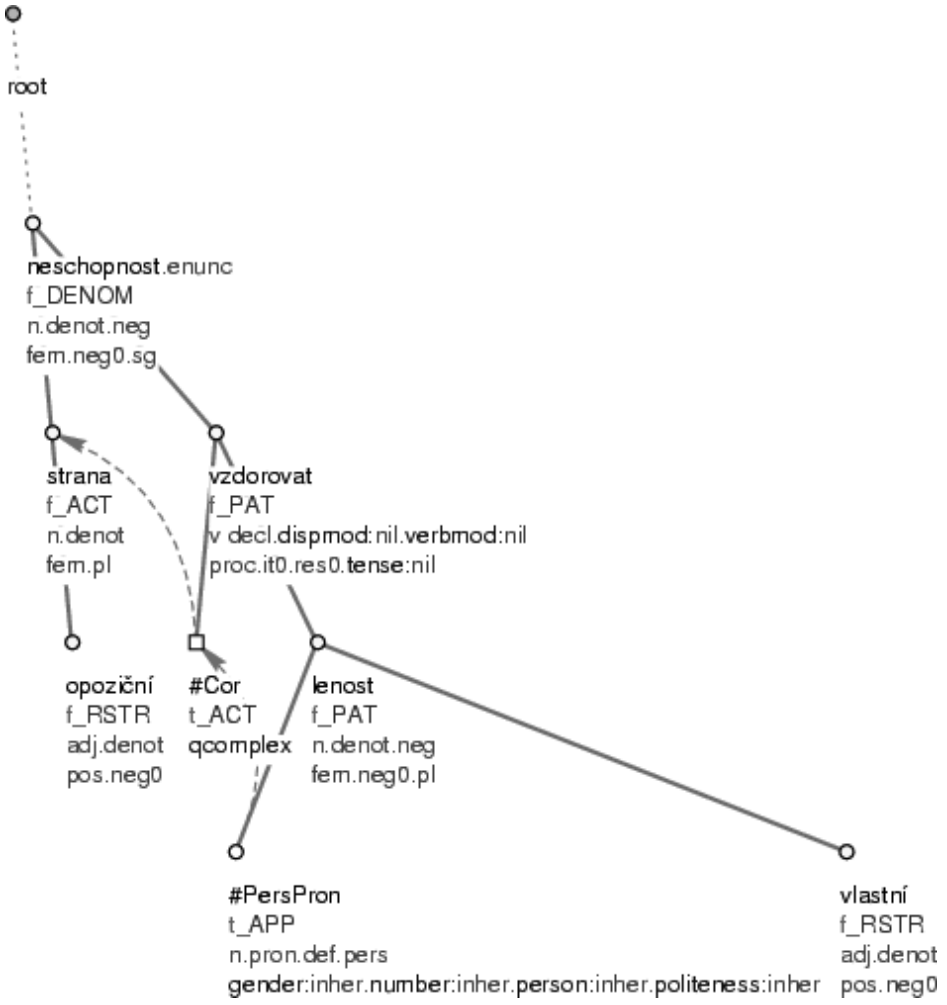


Рис. 1. Грамматическая кореференция

(1) *Neschopnost opozičních stran vzdorovat své vlastní lenosti.* — Неспособность оппозиционных партий бороться с собственной (со своей) ленью.

- кореференция возвратных местоимений, в случае, если они являются самостоятельным членом предложения (возвратное *se* («*ся*»)[2], лексемы *sebe* (*себя)* и *svůj* (*свой)*). Все возвратные местоимения имеют общую лемму *se* («ся») и отсылают к субъекту предложения, к ближайшему узлу с функтором ACT (агенс) — первично к агенсу той же клаузы, в случае, если он там отсутствует — к агенсу главного предложения. (см. рис. 1)
- кореференция относительных средств. К ним относятся относительные местоимения и наречия, относительные придаточные предложения и т. д. Ср. *člověk, který pije (человек, который пьет)*; *ve městě, kde se mi tak líbilo* (*в городе, где мне так понравилось*) и др.). В глубинно-синтаксическом дереве стрелка грамматической кореференции ведет от относительного местоимения *(который, где)* к управляющей именной группе (соответственно *человек, город*).
- кореференция в т. наз. контролирующих конструкциях (у некоторых глаголов, заданных списком в документации по разметке глубинно-синаксического уровня (Mikulová, 2005), напр. *стесняться, забыть, хотеть, научить* и др., один из актантов которых обязательно кореферентен с определенным актантом зависимого от них глагола в инфинитиве — напр. *zapomenout přečíst* (*забыть прочитать*)). При восстановлении модели управления зависимого глагола, его невыраженный кореферентный актант имеет лемму #Cor, от которого ведет стрелка грамматической кореференции к соответствующему актанту управляющего глагола. (см. рис. 1)
- кореференция актантов в реципрокальных конструкциях. Один из актантов имеет восстановленную лемму #Rcp, откуда ведет стрелка грамматической кореференции на лексически выраженный кореферентный актант (см. рис. 2).

[2] В чешском языке возвратное местоимение «ся» всегда является отдельной лексемой (клитикой).

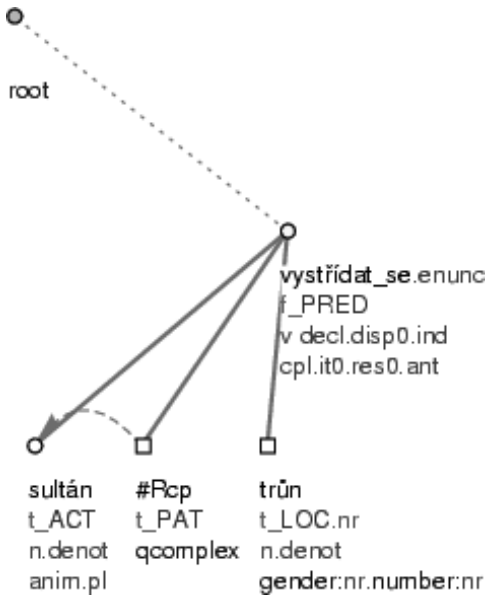


Рис. 2. Грамматическая кореференция

(2) *Sultáni se vystřídali na trůnu.* — Султаны поменялись местами (поменялись) на троне

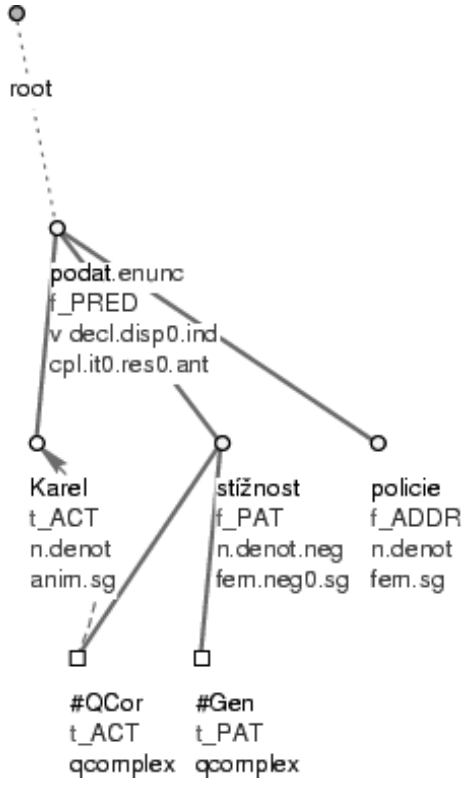


Рис. 3. Грамматическая кореференция

(3) *Karel podal stížnost policii.* — Карл подал жалобу в милицию.

- кореференция в т.наз. квазиконтролирующих конструкциях (в случае составного предиката, именной частью которого является имя существительное, имеющее модель управления, напр. *подать жалобу в милицию*). При восстановлении модели управления зависимого существительного, его невыраженный агенс имеет лемму #QCor, от которого ведет стрелка грамматической кореференции к агенсу управляющего глагола. (см. рис. 3)
- кореференция у дополнений с двойной зависимостью, выраженных формой глагола. Отношением кореференции связан восстановленный актант дополнения, выраженного формой глагола (причастием, деепричастием или инфинитивом) с актаном управляющего предиката. (см. рис. 4)

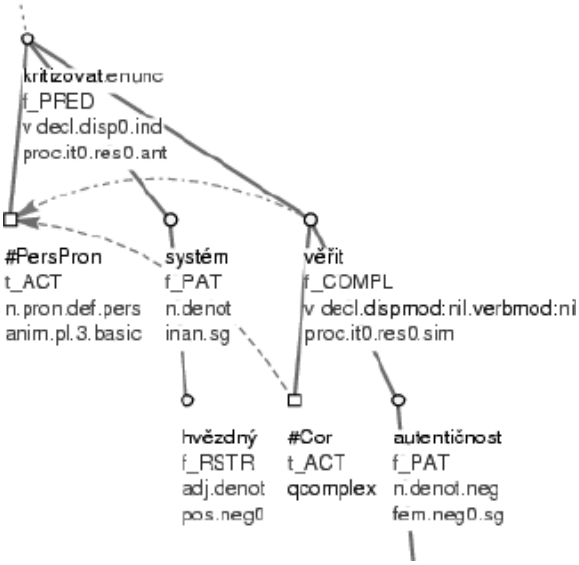


Рис. 4. Грамматическая кореференция

(4) *Kritizovali hvězdný systém, věříce v autentičnost…* — Они критиковали звездную систему, веря в истинность…

3. Разметка прономинальной текстовой кореференции

Текстовая кореференция понимается как использование различных языковых средств для анафорической (реже катафорической) отсылки. Эта отсылка реализуется не только за счет грамматических средств языка, но и на основании знания контекста. Текстовая кореференция может легко переходить границы предложения. Разметка текстовой кореференции проводилась вручную на всем корпусе текстов PDT. Текстовая прономинальная кореференция размечена в PDT 2.0 в следующих случаях:

- в качестве анафора выступают личные и притяжательные местоимения третьего лица. Кореференция местоимений первого и второго лица не размечается. Местоимения (в том числе эл-

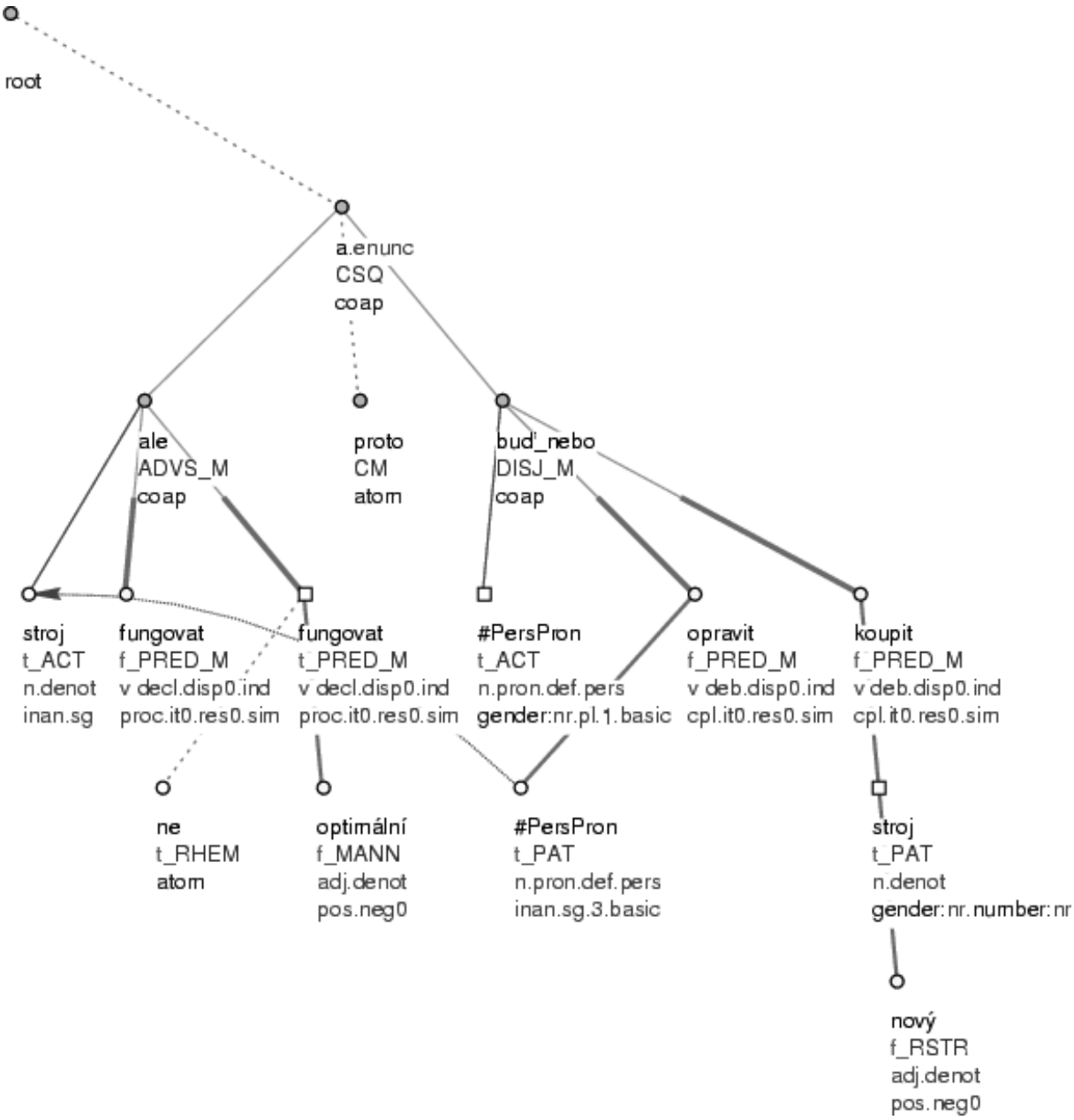


Рис. 5. Текстовая прономинальная кореференция

(5) *Stroj funguje, ale ne optimálně, a proto ho musíme buď opravit, nebo koupit nový.* — Прибор работает, но не оптимально, поэтому его нужно либо починить, либо купить новый.

липтированные) на глубинно-синтаксическом уровне имеют лемму `#PersPron` (см. рис. 5).

- в качестве анафора выступает указательное местоимение *этот* в субстантивной функции.
- в качестве анафора выступает эллиптированное и восстановленное на глубинно-синтаксическом уровне местоимение 3-го лица. Являясь языком pro-drop, чешский язык имеет сильную тенденцию опускать личные местоимения в анафорических конструкциях (ср. *cz O Nechtěl to říkat* vs. rus. *Он не хотел этого говорить*). На глубинно-синтаксическом уровне в PDT эти местоимения воостанавливаются, и им присваивается тектограмматическая лемма `#PersPron`. Информация о (не)выраженности этой леммы на поверхностном уровне содержится в атрибуте `is_generated`.

3.1. Отсылка к сегменту текста

Отсылка к сегменту текста имеет место в случае, когда либо антецедент местоимения состоит из более чем одного предложения, либо высчитывается на основании общего контекста. Информация об отсылке к сегменту текста фиксируется значением `segm` атрибута `coref_special`.

3.2. Дейксис

Отсылка к объектам внеязыковой действительности обозначается значением `exoph` атрибута `coref_special`.

4. Разметка именной текстовой кореференции и ассоциативной анафоры

4.1. Разметка именной текстовой кореференции

На данном этапе размечаются референциальные цепочки, где в качестве анафора выступают в основном имена существительные и некоторые наречия (*там*, *тогда* и др.). В некоторых случаях в отношении кореферентности могут участвовать прилагательные (притяжательные прилагательные и прилагательные, образованные от имен собственных) и числительные (выступающих в субстантивной функции и релевантных для связности текста). Технически разметка именной текстовой кореференции является частью предшествующей ей разметки прономинальной кореференции (используется `id` антецедента, к которому отсылает `id` узла анафоры, атрибут `coref_text.rf` содержит `id` кореферентного узла), однако добавляется информация о типе кореферентного отношения (атрибут `informal-type`). Отношение текстовой кореференции не фиксируется между субъектом и именной частью составного именного сказуемого, а также между узлами, находящимися в отношении аппозиции. Идентичность их референтов следует из синтаксической структуры дерева зависимостей.

При разметке именной текстовой кореференции используется 4 типа отношений:

- дефолтный тип 0 (значение `0` атрибута `informal-type`). Отношение между конкретнореферентными ИГ, причем анафор не является гиперонимом или синонимом ИГ антецедента. К этому типу относятся повторы ИГ антецедента (*женщина — женщина*), повторы ИГ антецедента с идентификатором (*женщина — эта женщина*), ИГ с существительным, антецедентом которого является местоимение или эллипсис, являющиеся звеном цепи прономинальной кореференции (таким образом достраиваются цепочки прономинальной кореференции, ср. женщина — она — женщина), частичные повторы ИГ антецедента (общество — акционерное общество) и др.
- синонимия в широком смысле (значение `SYN` атрибута `informal-type`). Обозначается, если анафорический член и ИГ антецедента — различные номинации. Помимо действительной синонимии, к этой группе относятся напр. такие случаи, как имя собственное — имя нарицательное (*Петя — раздолбай*), сокращение — полное название (*НДС — налог на добавленную стоимость*) и др.
- гиперонимия (значение ER атрибута `informal-type`). Этот тип не совсем соответствует своему названию, т.к. в процессе аннотирования его наиболее типичные пары (*яблоко — фрукт*) в результате нечеткой границы с предыдущим типом перешли в тип `SYN`. На настоящий момент тип `ER` приписывается в основном отсылкам на ситуацию (*Начальник заставил нас приходить вовремя. Это решение никому не понравилось*) и в случае т. наз. автонимной анафоры (отношения между ИГ *Адольф Гитлер — это имя, радуга — это слово* и т. д.)
- кореференция нереферентных и родовых ИГ (значение `NR` атрибута `informal-type`). Этот тип несколько проблематичен, т.к. решение связывать кореференцией ИГ, которые не обладают конкретной референцией, не является полностью интуитивным. Тем не менее зачастую нереферентные ИГ способны вступать в анафорические отношения наравне с референтными, в том числе являться антецедентами местоимений (Падучева 1985), поэтому не могут быть исключены из кореферентных цепочек. Пример пары кореферентных ИГ типа NR в (6):

(6) *Paláce neznamenají přepych. Ač se to na první pohled nezdá , obývání klasických renesančních a barokních palácůs velikými , řetězovitě propojenými místnostmi není žádné terno.* — Дворец не значит роскошь. На первый взгляд так не кажется, но обитание в о дворцах {`coref_text`, тип `NR` на «дворец»} в стиле барокко или ренессанса с огромными комнатами, расположенными анфиладой, не так уж безоблачно прекрасно.

Проблематичным является тот факт, что в произвольном корпусе текстов встречается большое количество нереферентных ИГ, отсылающих в принципе к одному и тому же, но не вступающих между собой в анафорические отношения. На данный момент мы не можем предложить алгоритм проведения четкой границы между нереферены (родовыми) ИГ, кореферентность которых является релевантной для связности текста, и просто повторяющимися ИГ с родовым статусом и отсылающими к одному и тому же, поэтому мы отдаем предпочтение аннотации кореферентности перед наличием анафорического отношения и связываем такие ИГ текстовой кореференцией с типом `NR`. Проблематичным также часто оказывается вопрос

о кореферентности ИГ с неконкретнореферентным денотативным статусом — при вторичном просмотре пар с отмеченной кореферентностью этого типа находится множество примеров, где кореференция не должна была бы быть обозначена.

Среди нереферентных ИГ не проводится различие на чистый повтор, синонимичные и гиперонимичные номинации. Это различие касается только ИГ с конкретной референцией. См пример (7):

(7) *Na telefonní číslo 855 44 33 bude jistě volat mládež s různými problémy. Doufejme, že linka si časem vydobude mezi dětmi takovou autoritu, aby se na ni obracely i ty, které jsou skutečně ohrožovány.* — По телефонному номеру 855 44 33 молодежь будет звонить с различного типа проблемами. Будем надеяться, что этот номер со временем достигнет такой популярности среди ребят {`coref_text`, тип `NR` на «молодежь»}, что по нему будут звонить и дети, которым действительно что-то угрожает.

Отдельную проблему представляют **абстрактные имена**. Проблематично уже само разделение имен на конкретные и абстрактные (Степанов 2004, Падучева 1986 и др.) Однако даже если предположить, что эта проблема решена, вопрос определения их денотативного статуса остается открытым. В нашей разметке кореференция абстрактных имен обозначается по умолчанию типом `NR`, однако не совсем последовательно. Если ИГ обладает абстрактной семантикой, но при этом очевидно конкретной денотативностью, разметчик вправе обозначить и дефолтный тип `0`. Эта конвенция однако является спорной и находится в стадии обсуждения. Ср. тип `NR` в (8) и тип `0` в (9):

(8) *Tímto faktorem je podnikatel — inovátor , který se snaží o zisk , a proto logicky nemůže existovat ve stavu statiky, která nezná ani zisk , ani ztrátu.* […] *Na konci tohoto procesu se systém vrátí ke statické rovnováze, v níž nebudou opět ani zisky, ani ztráty.* — Этим фактором является предприниматель-инноватор, который пытается получить прибыль и потому не может находиться в статичном состоянии, которому неизвестны ни прибыль {`coref_text`, тип `NR` на «прибыль»}, ни убыток. […] В конце этого процесса система снова возвращается к статическому равновесию, в котором снова не будет ни прибыли {`coref_text`, тип `NR` на «прибыль»}, ни убытка.

(9) *Televize dává příležitosti k podnikání.* […] … *nevyužité možnosti stále má televize zejména při regionálním vysílání.* — Телевидение располагает к предпринимательству. […] … неиспользованными возможностями обладает телевидение {`coref_text`, тип `0` на «телевидение»} прежде всего в региональном вещании.

Похожим образом разрешается разметка кореференции **имен действий**. Имена действий чаще бывают конкретны и соотносимы с реальной ситуацией, однако возникает проблема временной локализации действий и возможности кореференции ИГ расположенных на различных участках временной оси (см Падучева 1986). В данном случае решение о наличии стрелки кореференции часто бывает основано на языковой интуиции разметчика.

При разметке грамматической и текстовой кореференции выдерживается принцип сохранения референциальной цепочки, контролируемый частично автоматически. Если разметчик устанавливает отношение кореферентности с узлом, к которому уже ведет стрелка, новое отношение автоматически устанавливается с последним (самым правым) узлом.

4.2. Разметка ассоциативной анафоры (т. наз. bridging anaphora)

Параллельно с разметкой именной текстовой кореференции проводится разметка т.наз. ассоциативной анафоры (bridging anaphora). Анафорический член и антецедент в данном случае уже не кореферентны, но между ними имеется семантическое отношение определенного типа.

При аннотации PDT действуют некоторые конвенции выбора той или иной связи в сомнительных случаях. Одной из основных конвенций является предпочтение текстовой кореференции перед ассоциативной анафорой.

Наличие разметки ассоциативной анафоры связано с общей структурой дерева зависимостей глубинно-синтаксического уровня PDT. Ассоциативная анафора не аннотируется, например, если узел участника отношения является непосредственным потомком антецедента с определенным функтором (PAT, APP, AUTH и др.[3]), если отношения между участниками отношения уже выражены грамматическим функтором или синтаксической структурой дерева и т.д. (Nedoluzhko 2007)

В отличие от текстовой кореференции, разметка ассоциативной анафоры затрагивает практически только те узлы, которые соответствуют в тексте полнозначным лексемам. Ссылка на эллиптированные местоимения, союзы и знаки препинания возможна только в том случае, если другого не позволяет структура дерева.

- С технической точки зрения разметка ассоциативной анафоры — это отсылка узла анафора к `id` антецедента, информация о связи со-

[3] к описанию значения функторов см (Hajičová и др. 2006)

держится в атрибуте bridging. Информация о типе отношения отображается в атрибуте informal-type. Аннотация ассоциативной анафоры не является дополнением референциальной цепочки, состоящей из отношений грамматической и текстовой кореферентности, а существует параллельно. Референциальная цепочка ассоциативной анафоры не удерживается (по крайней мере, не удерживается последовательно).

При разметке PDT выделяются и размечаются следующие типы ассоциативной анафоры:

- отношение множество-подмножество/элемент множества (значения SUB_SET и SET_SUB атрибута informal-type в зависимости от направления отношения). Типичные примеры: *мушкетеры — Атос, Портос, Арамис; семинары — первый семинар, последний семинар*. Ср. также (10):

(10) *Na rozdíl od dobře vybaveného FS dnes nikdo z téměř dvou stovek poslanců kromě předsedy a místopředsedů sněmovny nemá svou kancelář, pracovní stůl, židli a telefon.* — В отличие от хорошо оснащенной Федерального парламента, сегодня ни у кого из почти двухсот депутатов, кроме председателя {bridging, typ=SET, на «poslanec (депутат)»} парламента и зампредседателей {bridging, typ=SET, на «poslanec (депутат)»} нет своего кабинета, рабочего стола, стула и телефона.

- отношение часть — целое (значения PART_WHOLE и WHOLE_PART атрибута informal-type в зависимости от направления отношения). Типичные примеры: *комната — потолок, рука — палец* и др. Как часть — целое аннотируются также неотделимые части в географических названиях, напр. *ФРГ — Бавария — Мюнхен*. Граница между отношениями «часть — целое» и «множество — подмножество» не всегда является достаточно отчетливой. Во многих случаях решение зависит только от исчисляемости объектов, входящих в данное отношение (напр. *заграница — Германия vs. иностранные государства — Германия; текст — предложение* и др.). Возможно, в дальнейшем эти два типа можно совместить (ср. проекты PoCoS, MATE и др.), но пока мы размечаем их отдельно.
- отношение дискурсивного контраста, имеющего значение для связности текста (значение CONTRAST атрибута informal-type). Этот тип частично пересекается с размеченным на всем корпусе PDT актуальным членением (Hajičová 2006, коротко также в Недолужко 2008), но не полностью его копирует. Члены отношения ассоциативной анафоры типа CONTRAST могут находиться в предложении как в позиции контраста, так и в позициях топика и фокуса; кроме того, ассоциативный контраст не ограничен рамками предложения. Ср. пример (11), где ИГ *коровы* расположена в фокусе:

(11) *Lidi nežvýkají, to jenom krávy*. — Люди не жуют, жуют только коровы {bridging, тип CONTRAST на узел «человек»}.

- отношение объекта и его функции/позиции (значения FUNCT_P и P_ FUNCT атрибута informal-type в зависимости от направления отношения). Напр. *школа — учитель, министр — министерство* и др.
- остальное (значение REST атрибута informal-type). В эту группу включаются отношения, которые не были описаны выше, но которые, возможно, будут позже уточнены и выделены в новые группы. Предполагается, что лингвисты-аннотаторы не будут загромождать этот тип парами, которые просто как бы то ни было семантически связаны, а помещать туда только потенциально классифицируемые случаи. В частности к ним относятся отношения место — житель (*Москва — москвич*), автор — творение, вещь — хозяин, родственные отношения (*дед — внук*), некоторые предикатно-аргументные отношения (*предпринимательство — предприниматель, спор — участник конфликта* и др.) а также некоторые релевантные для связности текста равнолексемные некореферентные пары (*случайность — еще одна случайность*)

* * *

Разметка именной текстовой кореференции и ассоциативной анафоры проводится в настоящее время автором данного доклада и тремя аннотаторами с лингвистическим образованием и знаниями в области теории референции и дискурса. Разметка проводится с помощью программы для аннотирования корпусных данных TrEd (од *tree editor*), разработанная на ÚFAL MFF UK, с использованием специально созданных приложений для разметки кореференции. Разметка проводится в основном вручную непосредственно на дереве зависимостей или на тексте (по желанию разметчика). Кроме того, было разработано несколько программ, упрощающих и ускоряющих процесс аннотирования: предварительное выделение лемм, совпадающих с актуальной, указание кореферентных связей данного узла и др. К концу 2008 года было размечено 7000 предложений.

Литература

1. *Hajičová E., Hajič J., Hlaváčová J., Klimeš V., Mírovský J., Pajas P., Štěpánek J., Vidová-Hladká B., Žabokrtský Z.* PDT 2.0 — Guide. UFAL & CKL, 2006 Доступно на http://ufal.mff.cuni.cz/pdt2.0/
2. *Hirschman L.* MUC-7 coreference task definition version 3.0. // Proc. of the 7th Message Understanding Conference под ред. Chinchor N. 1998. Доступно на http://acl.ldc.upenn.edu/muc7/co_task.html
3. *Kučová L.* и др. Anotování koreference v Pražském závislostním korpusu. ÚFAL/CKL Technical Report TR-2003-19. 2003
4. *Mikulova M.* и кол. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. — Institute of formal and applied linguistics, Charles University, Prague, 2005.
5. *Nědolužko A.* Zpráva k anotování rozšířené textové koreference a bridging vztahů v Pražském závislostním korpusu. (Report about the annotation of the extended text-coreference and bridging relations in Prague Dependency Treebank.). Technical report. Institute of formal and applied linguistics, Charles University, Prague. 2007 Доступно на http://ufal.mff.cuni.cz/~nedoluzko/koref_anot/manual_RK_kratky.pdf
6. *Passonneau R. J.* Instructions for Applying Discourse Reference Annotation for Multiple Applications (DRAMA), 1997
7. *Poesio M.* The MATE/GNOME Scheme for Anaphoric Annotation, Revisited // Proc. of SIGDIAL, Доступно на Boston, April. 2004. Доступно на http://cswww.essex.ac.uk/staff/poesio/publications/SIGDIAL04.pdf
8. *Недолужко А., Гаич Я.* Синтаксически аннотированный корпус чешского языка. // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Выпуск 7 (14) 2008 С.400-406 Падучева Е. В. О референции языковых выражений с непредметным значением. // НТИ, сер. 2, N 1, 1986.
9. *Падучева Е. В.* Высказывание и его соотнесенность с действительностью. М.: Наука, 1985.
10. *Степанов Ю. С.* Имена, предикаты, предложения (семиологическая грамматика). М.: Едиториал УРСС, 2004

Сегментация устного нарратива и изобразительные жесты: кинетические признаки границ и связей между сегментами дискурса

Segmentation of oral narrative discourse and illustrative gestures: visual clues as segment markers

Николаева Ю. В. (lis_julia@list.ru)

Московский государственный университет им. М. В. Ломоносова

Это исследование изучает взаимосвязь иллюстративных жестов, сопровождающих речь, и структуры дискурса. Задача этой работы — показать, какие отдельные признаки жестов-иллюстраторов указывают на наличие границ сегментов — таких единиц дискурса, которые крупнее, чем ЭДЕ, но меньше, чем весь дискурс.

Введение

Для целей исследования нами был создан корпус устных рассказов. Стимулом для них послужил 6-минутный видеосюжет, т. н. «Фильм о грушах» — «Pear film» (Chafe 1980). Было записано 8 описаний этого фильма, сделанных студентами МГУ, общей продолжительностью около 20 мин. Всего в корпусе было 594 ЭДЕ (элементарных дискурсивных единицы, которые обычно совпадают с простым предложением; подробнее об этом см. Кибрик, Подлесская 2006) и примерно 370 иллюстративных жестов.

При изучении корпуса главным вопросом был следующий: можно ли только на основании жестов построить по каким-то признакам структуру дискурса. Для ответа на этот вопрос были рассмотрены ряд жестовых характеристик, их взаимосвязь, а также некоторые другие дополнительные факторы.

1. Основные понятия.

Сферой данного исследования являются иллюстративные жесты и то, как они используются в процессе коммуникации.

Что такое жесты? Слово интуитивно понятно и очевидно, но тем сложнее дать ему научное определение. Следует учитывать также, что понятие «жест» используется во многих контекстах и для самых разных аналитических и практических целей.

В обычном смысле слово «жест» понимается как телодвижение, преимущественно рукой, сопровождающее речь для усиления ее выразительности или имеющее значение какого-либо сигнала, знака и т. п. (см. Словарь русского языка в 4-х тт. под ред. Шевелевой). Чтобы сузить число рассматриваемых явлений, большинство исследователей (см. напр. McNeill 1992, Goldin-Meadow 2003, Kendon 2004) рассматривают только движения рук, не всегда уточняя это эксплицитно. В нашем исследовании также речь идет только о движениях рук (хотя движения и изменение положения корпуса, головы и ног тоже имеют значение в дискурсе, но основная доля информации, передаваемой в общении, относится к движениям рук (см. Крейдлин 2002)).

Второй важный компонент в этом определении — намерение говорящего, который использует этот жест с целью выделить или сообщить что-то, в дополнение или вместо речи. Большинство авторов работ по данной теме прямо или косвенно выделяют этот аспект, обозначая область своего исследования. Вот какое определение жестам дает Адам Кендон: «Движения, распознаваемые другими как совершенные намеренно с целью выразить что-то в процессе коммуникации» (Kendon 1996: 9). В определении, которое предлагает Г. Е. Крейдлин, понимание жестов как коммуникативного средства отделено от намерения говорящего — жесты понимаются как носители информации, выступая в качестве «знаковых кинетических единиц выражения и передачи информации» (Крейдлин 2002: 10). Движения, не передающие ин-

формацию, например, продиктованные физиологическими потребностями (затекли ноги или чешется затылок), таким образом, не входят в сферу интересов исследователей жестов. Проведение границы между коммуникативными движениями и невольными и физиологическими движениями неизбежно сталкивается с некоторыми дополнительными вопросами, поскольку приходится разделять некоторый континуум, приписывая все множество возможных движений к одному из двух полюсов; еще один вопрос — кто наблюдатель, кто оценивает, является ли данное поведение знаковым или нет. В контексте данной работы эта проблема приобретает дополнительную сложность, поскольку дальше мы покажем, что движения, на которые раньше не обращалось внимания, и которые посторонние наблюдатели склонны были бы объяснять субъективными факторами, тесно связаны с потоком речи и даже могут добавлять некоторую новую информацию к передаваемым словам.

В этой же работе Г. Е. Крейдлина описана наиболее распространенная на данный момент классификация жестов. Они подразделяются на три основных семиотических класса: (а) имеющие самостоятельное лексическое значение и способные передавать смысл независимо от вербального контекста, (б) выделяющие какой-то речевой или иной фрагмент коммуникации, и (в) управляющие ходом коммуникативного процесса, то есть устанавливающие, поддерживающие или завершающие коммуникацию. Первый вид — эмблематические жесты, или эмблемы. Второй вид — это иллюстративные жесты, или иллюстраторы, а третий — регулятивные жесты, или регуляторы. Т.о. иллюстраторы понимаются как жесты, которые по своей природе не могут передавать значения независимо от вербального контекста и никогда не употребляются отдельно от него. Наряду со словом иллюстраторы, также используются термины «жестикуляция» (Kendon 1983) и «движения, ориентированные на речь» (Butterworth, Beattie 1979).

2. Характеристики жестов, связанные со структурой дискурса

Чтобы определить, какие особенности жестов и каким образом связаны с переходом от одной локальной темы, события или эпизода в мире дискурса к другому, мы обращали внимание на следующие признаки кинетических знаков:

1. Смена положения рук в нейтральной позиции (в положении покоя).
2. Изменение места производства жеста, а также сохранение ментального пространства [1] (для нескольких последовательных или не следующих друг за другом жестов и повторное обращение к предыдущим жестам).
3. Изменение числа рук, участвующих в жесте.
4. Повторяющиеся жесты или отдельные их характеристики, при изменении других.
5. Также учитывались форма, семантика и тип жеста.

В некоторых случаях какие-то характеристики жестов могут служить показателями границ сегментов разного уровня (более и менее крупных), их функций (вступление — основная часть — завершение; комментарии, относящиеся к метатекстовому уровню; второстепенные замечания, сбои, самоповторы и переформулировки) и даже связей сегментов дискурса между собой.

3. Положение рук в нейтральной позиции (в положении покоя)

Как показывает анализ корпуса устных рассказов, иногда даже минимальные изменения (говорящий держит руки на коленях ладонями вниз, потом переворачивает правую руку ладонью вверх, потом снова ладонью вниз) служат явными показателями границы между сегментами. Однако надо отметить, что в таких тонких вещах, как кинетические показатели структуры дискурса, рассматриваемые нами, очень многое определяется индивидуальными стратегиями каждого человека. У некоторых из 8 изученных человек нейтральное положение рук менялось каждые 2–3 ЭДЕ, у некоторых оно сохранялось неизменным на протяжении всего рассказа. Однако однозначно можно сказать, что этот фактор является признаком нового смыслового абзаца в большинстве случаев, что удивительно, если учесть, что движения могут быть едва заметными даже для внимательного наблюдателя. Уже достаточно хорошо изучено, что смена позы говорящего, приближение или отдаление корпуса могут указывать на смену темы разговора или даже на изменение стиля беседы (например, это может быть переход от обсуждения формальных вопросов к дружескому общению). Однако если говорящий кладет локоть на стол или разворачивает руку ладонью вверх, это тоже может говорить о некоторой ло-

[1] Ментальное пространство — понятие, предложенное Г. Фоконнье и М. Тернером (Turner, Fauconnier 1995) как описание представления множества возможных миров. Особенность ментального пространства — в том, что оно не является полным соответствием реальности, а является идеализированной когнитивной моделью. Говорящий, описывая некоторую ситуацию, создает воображаемое пространство, в котором находятся описываемые лица и происходят какие-то события. Это ментальное пространство находит свое отражение и в жестах, когда говорящий, например, с помощью указательных жестов размещает перед собой героев рассказа и затем показывает их действия и перемещения в соответствии с сюжетом.

кальной границе в дискурсе: например, имеет место появление нового действующего лица при сохранении места и времени описываемых событий.

Еще один момент, который мог бы служить знаком окончания эпизода — возвращение рук в положение покоя после окончания непрерывной серии жестов, т.н. жестовой фразы (об иерархии жестовых движений, предложенной Кендоном, см. Крейдлин 2002: 73 и McNeill 1992: 82). В нашем корпусе встречались рассказы, в которых говорящая опускала руки на колени после каждого жеста. Можно сделать вывод, что этот признак очень сильно зависит от индивидуальных особенностей жестикулирующего, так что опираться на него при выделении дискурсивных сегментов не всегда возможно. С большой степенью вероятности можно утверждать, что серии непрерывных жестов — жестовые фразы — находятся внутри одного сегмента.

4. Место производства жеста

Очевидно, что у каждого человека есть некоторое комфортное положение рук, в котором будут производиться все жесты, сопровождающие данный сегмент. Кажется логичным предположение, что по ходу рассказа жестовое пространство будет распределено между разными референтами — действующими лицами, объектами и пр., которые будут располагаться в разных местах, и потом использоваться при следующих упоминаниях. Однако, как следует из примеров в нашем корпусе, такие случаи очень редки. Место производства жеста зависит скорее от темперамента и других индивидуальных или ситуативных характеристик. На это может влиять, например, реакция слушающего, если он демонстрирует непонимание обращенных к нему слов. Так, в одном из рассказов из-за путаного изложения событий (частые возвраты и пояснения к уже рассказанному) слушающая не смогла идентифицировать антецедентов местоимения. Тогда рассказчица вернулась к началу этого эпизода, при этом подняла руки, чтобы жесты были видны адресату (до этого все ее жесты были закрыты от слушающей столом), и жестикулировала гораздо больше, энергичнее и подробнее, чем в первом варианте. При этом ее руки или она сама представляла только того персонажа, о котором она говорила в данный момент, без соотнесения с другими действующими лицами и без присутствия закрепленных за ними точек в пространстве рядом с говорящей.

Однако мы встретили у этой же рассказчицы примеры, когда жест указывал на некую точку в пространстве, где был выполнен один из жестов раньше — таким образом, происходила отсылка к описанным ранее событиям (пример 1, рис. 1).

(1) *ЭДЕ № 143–145*

Мимо проходят мальчики,
которые <u>едят груши</u>. [2]
правая рука у лица, ладонью к себе, пальцы полусогнуты — «едят груши», см. рис. 1.

<u>Эти же груши</u>
форма руки сохраняется, на той же высоте движение сначала вперед, к слушающей, затем в ту же точку у лица — т.о. получается сопоставление «тех» груш, о которых было рассказано раньше (собранных садовником) и которые должна вспомнить слушающая, и тех, которые были показаны в жесте у лица одновременно с предыдущей ЭДЕ.

В данном примере в ментальном пространстве у лица помещены не груши, а вся ситуация — мальчишки идут мимо садовника и едят груши. И сопоставление идет не с каким-то другим объектом, а с грушами же, но в других обстоятельствах, т. е. с другой ситуацией.

Рис. 1. «Едят груши»

[2] Здесь и далее в примерах используются следующие обозначения: жест сопровождал слова, подчеркнутые сплошной линией; место явно выделяющегося пика жеста, если он есть, или наложения жеста-удара показано вертикальной линией. Под этой строчкой шрифтом — описание жеста. Чтобы различить речь и описание жеста, речь набрана полужирным шрифтом. В скобках — заполненные и незаполненные паузы, их продолжительность указана в секундах с точностью до 0.1 сек. Вклинившаяся в рассказ реплика слушающего — в квадратных скобках.

5. Число рук, участвующих в жесте

Как отмечалось в работе Quek et al. 2001, «рукость» (handedness) — один из возможных признаков, указывающих на единство некоторого дискурсивного сегмента.

Кроме того, этот признак (одна или две руки участвуют в жесте) может закрепляться за отдельными референтами. Так, в нашем корпусе все участники изображали большую и тяжелую корзину с грушами, как бы держа ее за две ручки. Это уже похоже на эмблематический жест («корзина» — две руки ладонями друг к другу, сжатые в кулаки, как на рис. 2). Так, рассказ о том, как мальчик ставит эту корзину себе на велосипед, сопровождается несколькими похожими жестами: говорящий поднимает воображаемую корзину за две ручки и переносит ее в сторону и чуть выше, и один и тот же жест повторяется на протяжении всего сегмента, посвященного этому эпизоду, см. пример 2, рис. 2:

(2) *ЭДЕ №181–188*
потом решает, что нужно украсть | целую = целую корзину.
две руки симметрично сжаты в кулаки, ладонями друг к другу, с согнутыми под прямым углом локтями — «держит корзину»; движение справа налево по дуге, немного вверх — «поднимает и ставит корзину на велосипед»
(.. а 1.0) Он значит ее берет, с трудом поднимает эту корзину на велосипед,
жест повторяется, с большей амплитудой
(....1.2) Он поднимает, ставит эту корзину на велосипед,
жест повторяется, руки уже более расслаблены
(...0.8) и уезжает. (см. рис. 3)

Рис. 2. Жест «поднимает корзину»

6. Повторяющиеся жесты или сохранение отдельных особенностей жестов

Дэвид МакНил, первым обративший внимание на такой интересный феномен, называет повторяющиеся жесты catchments (McNeill et al. 2000).[3] Д.МакНил описывает это явление так: «Кэтчмент присутствует, когда какие-то характеристики сохраняются в двух или более (не обязательно последовательных) жестах. Смысл этого понятия в том, что повторение визуального представления в мышлении говорящего будет порождать повторяющиеся характеристики жеста. Выявляя кэтчменты, созданные определенным говорящим, мы можем увидеть, что именно этот говорящий объединяет в более крупные фрагменты — какие понятия рассматриваются им как сходные или относящиеся к одной группе, а какие значения помещены в разные кэтчменты или изолированы, и таким образом воспринимаются говорящим как имеющие разные или менее связанные значения.»

Эта тема дает самые большие возможности для обобщений. В примере 3 повторение формы жеста показывает тесную связь между соседними ЭДЕ.

Однако поскольку главным вопросом в нашей работе было построение структуры всего нарратива, то этого фактора, взятого отдельно, оказалось недостаточно. Повторяющиеся жесты — не постоянное явление в процессе речи, сопровождаемой жестикуляцией. Они помогают выделить какие-то отдельные сегменты, связать их друг с другом, как в примере 1, но для построения целостной структуры их решительно недостаточно.

7. Взаимодействие говорящего и слушающего и его влияние на жестикуляцию говорящего

Уже достаточно хорошо известно, что далеко не все жесты говорящего ориентированы на слушающего. В нашем эксперименте и говорящий, и слушающий сидели у стола, разделенные его краем, и примерно половина жестов были вне зоны видимости адресата.

В статье Bavelas 2007 было показано, что информация, передаваемая адресату, очень сильно зависит от эмоционального состояния адресата, направления его внимания и отношений между участниками коммуникации. О важности включения микросоциального контекста в анализ дискурса также писали Müller 2007 и Braten 2007. В нашем корпусе было несколько примеров такого взаимодействия. Один из частых случаев — говорящий, чаще всего невербально (интонацией или

[3] В русскоязычной литературе мы пока не встречали перевод или аналог этого термина, поэтому далее он остается без перевода.

жестами) запрашивает у слушающего обратную связь. Это может быть выражение неуверенности в точности пересказа фильма, или ожидание подтверждения того, что слушающий внимательно следить за рассказом и понимает, что там собственно происходило. Вот один из примеров такого общения, в котором вопрос задан говорящим, но с помощью невербальных сигналов, и а ответ дан слушающим, но посредством речи.

У одной из студенток, пересказывавших фильм, дважды появился жест, когда она убирала левую руку к лицу: в середине рассказа (пример 3, рис. 3) и в конце (пример 4).

(3) *ЭДЕ №15–23*
(…0.6) Он остановился рядом с корзинами,
(…0.5) ии по всей видимости украл одну из них.
(..0.4) То есть <u>он не говоря ни слова</u>,
Жест «метафора передачи» правой рукой, левая рука у лица.
[- которая была наполнена?] да, он не говоря {ни слова, просто (ммм .7) слез с велосипеда, взял эту корзину, поставил к себе на велосипед, и уехал.

Рис. 4. «То есть не говоря ни слова»

(4) *ЭДЕ №48–50*
И он (..мм 0.8) никак в общем не выразил свое неудовольствие, просто был в удивлении.
(…. 1.2) <u>Все</u>.
Жест «метафора передачи» правой рукой, левая рука у лица.

В обоих случаях (в № 17 и 20 и в последней, завершающей ЭДЕ №50) рука у лица сообщает о неуверенности (которая выражена в №16 словами «по всей видимости», а в конце проявляется также в длинной паузе и вопросительном взгляде).

Выводы

В нашей статье поднята пока малоизученная тема о связи структуры устного нарратива и иллюстративных жестов. На примере небольшого корпуса устных рассказов показано, как отдельные признаки жестов и положения рук могут добавлять дополнительную информацию касательно организации дискурса, состояния говорящего и процесса коммуникации. Так, изменение положения покоя рук с большой вероятностью указывает на границу сегмента — «абзаца» устной речи. Повторяющиеся жесты указывают на сохранение некоторой локальной темы или на возвращение референта, упомянутого ранее. Использование жестового пространства для повторной референции, по нашим наблюдениям, имеет свои особенности: в таких случаях жест напоминает не столько об объекте, сколько о целой ситуации из мира дискурса или о моменте в коммуникации.

Литература

1. *Кибрик А. А., Подлесская В. И.* Проблема сегментации устного дискурса и когнитивная система говорящего // Соловьев В. Д. (ред.) Когнитивные исследования. М.: Институт психологии РАН. 2006. Вып. 1. С. 138–158.
2. *Крейдлин Г. Е.* Невербальная семиотика // М.: Новое литературное обозрение, 2002.
3. *Николаева Ю. В.* Функциональные и семантические особенности иллюстративных жестов в устной речи (на материале русского языка) // «Вопросы языкознания», №4, 2004, с. 48–64
4. *Bavelas J.* Face-to-Face Dialog as a Micro-Social Context: The Example of Motor Mimicry // Gesture and the Dynamic Dimension of Language. Essays in honor of David McNeill. Eds. S. Duncan, J. Cussell, E. Levy. Pp. 127–146. Amsterdam/Philadelphia, 2007.
5. *Butterworth B., Beattie G.* Gestures and silence as indicators of planning in speech // R. N. Campbell, P. T. Smith (eds.) Recent advantages in the psychology of language: Formal and experimental approaches. NY: Plenum Press, 1978, 347–360.
6. *Braten E. S.* On being moved: From mirror neurons to empathy. // Gesture and the Dynamic Dimension of Language. Essays in honor of David McNeill. Eds. S. Duncan, J. Cussell, E. Levy. Pp. 109–116. Amsterdam/Philadelphia, 2007.
7. *Chafe W.* (ed.) The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production. // Norwood, New Jersey, Ablex, 1980.
8. *Goldin-Meadow, S.* Hearing gesture: How our hands help us thin. Cambridge, MA: Harvard University Press. 2003.
9. *Kendon A.* Gesture and speech: How they interact. // J. M. Wiemann, R. P. Harrison (eds.) Nonverbal interaction. Beverly Hills, CA: Sage Publications, 1983, 13–45.
10. *Kendon A.* An Agenda for Gesture Studies. Semiotic Review of Books. Vol. 7 (3) 1996. 8–12.
11. *Kendon A.* Gesture: Visible action as utterance. Cambridge: Cambridge University Press. 2004.
12. *McNeill D.* Hand and mind: what gestures reveal about thought. // Chicago: Chicago Univ. Press, 1992.
13. *McNeill D., Quek F., McCullough K.-E., Duncan S., Furuyama N., Brill R, Ma X.-F., Ansari R.* Catchments, prosody and discourse. // Gesture, 2000.
14. *Müller C.* A Dynamic View of Metaphor, Gesture and Thought. // Gesture and the Dynamic Dimension of Language. Essays in honor of David McNeill. Eds. S. Duncan, J. Cussell, E.Levy. Pp. 109–116. Amsterdam/Philadelphia, 2007.
15. *Quek F., McNeill D., Brill R, Duncan S., Ma X.-F., Kirbas C., McCullough K.-E., Ansari R.* Gesture and speech multimodal conversational interaction // VISLab Report-01-01, 2001.
16. *Turner M., Fauconnier G.* Conceptual Integration and Formal Expression // Metaphor and Symbolic Activity. 1995. Vol. 10. № 3.

Модели и методы учета пунктуации при синтаксическом анализе предложения русского языка

Models and methods of punctuation use in Russian language syntax parsing

Окатьев В. В. (oka@dictum.ru),
Ерехинская Т. Н. (te@dictum.ru), **Скатов Д. С.** (ds@dictum.ru)
ООО «Диктум», г. Нижний Новгород

Рассматривается функциональная омонимия знаков препинания в русском языке. Предложена формальная модель обособлений и рядов. Определены отношения между ними. Предложена математическая постановка задачи анализа пунктуации в контексте задачи машинного синтаксического анализа и алгоритм ее решения.

В настоящей работе рассматривается задача учета пунктуации при синтаксическом анализе предложения. Одной из основных проблем компьютерной лингвистики является омонимия, в частности функциональная омонимия знаков препинания. Знание функций знаков препинания в анализируемом предложении помогает определить структуру подчинительных связей слов, что, в конечном счете, позволяет более качественно решать задачу синтаксического анализа предложений на естественном языке.

На наш взгляд, в настоящее время задача синтаксического анализа предложения с учетом пунктуации недостаточно формализована. В работе [1] отмечается, что знаки препинания не только организуют текст, но иногда оказываются единственным доступным средством выбора правильной интерпретации текста. К сожалению, в этой работе не предлагаются конкретные алгоритмы реализации пунктуационного разрешения неоднозначности для систем автоматической обработки текста, т.к., по мнению ее авторов, эти алгоритмы сильно зависят от общего устройства конкретной системы.

Разными исследователями информация о знаках препинания в предложении используется поразному. При статистических подходах пунктуацию часто игнорируют. Интересен подход, основанный на использовании пунктуации — фрагментационный анализ [3-5]. При фрагментационном анализе заданное предложение разбивается по знакам препинания и союзам на отрезки, а затем с помощью специальных правил и словарей выполняются операции объединения, сочинения и подчинения этих отрезков. При построении анализатора на основе правил знак пунктуации может считаться токеном, и включаться в состав правил, как это делается в работах [2] и [10]. Кроме того, в работе [10] критикуется разделение синтаксического анализа на два уровня: обработка фрагментов, не содержащих знаков препинания, и затем определение отношений между этими фрагментами.

Много внимания уделяется пунктуации в работах Кобзаревой Т. Ю. В работе [6] описывается вложение сочиненных групп — «матрешка» сочинительных конструкций, ограничения по размещению отдельных компонентов. В работе [8] рассматривается структура вложений придаточных предложений и обособленных оборотов. В работе [7] описана омонимия знаков препинания и их способность выполнять несколько функций одновременно.

В настоящей работе описана формальная модель пунктуации и предложена математическая постановка задачи анализа пунктуации в контексте задачи синтаксического анализа. Формальная модель синтаксической конструкции, содержащей знаки препинания, понятия композиции конструкций и покрытия знаков препинания предложены Окатьевым В. В. Им же предложена формальная постановка задачи анализа пунктуации. Алгоритм решения задачи, его эвристические улучшения и интеграция с методами синтаксического анализа разработаны Ерехинской Т. Н.

1. Обзор функций знаков препинания

В грамматике русского языка [9] различаются:

1. Знаки выделяющие. Сюда относятся: две запятые (как единый парный знак), два тире (то же самое), скобки, кавычки;
2. Знаки отделяющие. Сюда относятся: точка, вопросительный знак, восклицательный знак, запятая, точка с запятой, двоеточие, тире, многоточие.

Использование знаков препинания в качестве формальных признаков при автоматической обработке текста осложняется их многофункциональностью. Например, запятая может выполнять как отделяющую функцию, так и выделяющую. Более того, одна и та же запятая может выполнять эти функции одновременно. В частности, запятая может закрывать причастный оборот и одновременно разделять однородные члены.

Проведенные исследования позволяют сделать вывод о наличии весьма строгих закономерностей, которым подчиняются правила пунктуации русского языка. Обнаруженные закономерности допускают их формализацию с помощью математических моделей. Эти модели могут использоваться для разработки компьютерных программ синтаксического анализа, в частности, они используются в системе синтаксического анализа DictaScope компании «Диктум» [11, 12].

2. Моделирование обособлений и рядов

В данной работе из всех синтаксических конструкций, в построении которых используются знаки препинания, будем рассматривать только обособления и ряды однородных членов. При машинном синтаксическом анализе эти конструкции невозможно рассматривать отдельно, поскольку в их построении применяются одни и те же знаки пунктуации, прежде всего – запятая.

При разработке методов анализа обособлений и рядов необходимо учитывать наличие сочинительных союзов, участвующих в построении союзных рядов однородных членов. Поскольку сочинительные союзы выполняют отделяющую функцию, то их в рамках данного исследования целесообразно рассматривать наравне с запятыми.

Отдельно следует сказать о двоеточии. В рамках данной работы будем его игнорировать. Разработанный подход к анализу пунктуации допускает включение этого знака препинания в модель, однако это привело бы к усложнению описания. Такое решение не приводит к существенному ухудшению результатов, поскольку основные сложности при анализе пунктуации связаны с запятыми, встречаемость которых на порядок выше, чем у двоеточий [12].

Перейдем к формальному описанию обособлений и рядов. Для этих целей пронумеруем слова в анализируемом предложении по порядку слева направо, начиная с единицы. Каждому знаку препинания припишем номер слова, после которого он непосредственно стоит, увеличенный на 0.5. Таким образом, знакам препинания будут соответствовать дробные номера. Например, если между вторым и третьим словами в предложении находится запятая, то у нее будет номер 2.5. В рамках разрабатываемой модели имеет значение линейный порядок слов и знаков препинания, а не величина их номеров.

Будем рассматривать ряд как множество пар однородных членов, между которыми можно предположить наличие сочинительной связи. Чтобы сделать такое предположение для пары слов b и d, необходимо указать их общего предка, от которого они зависят, и разделитель. Функцию разделителя может выполнять сочинительный союз, либо знак препинания: запятая или точка с запятой. Таким образом, сочинительная связь моделируется четверкой чисел:

$$f =< a, b, c, d >,$$

где a – главное слово; b, d — зависимые слова; c — разделитель однородных членов (сочинительный союз, запятая, точка с запятой). Для двойных сочинительных союзов будем указывать позицию второго союза.

Для f выполняется следующее соотношение:

$$b < c < d \ \& \ (a < b \ \vee \ d < a).$$

Вариант $a < b$ соответствует препозиции главного слова, $d < a$ — постпозиции.

Будем говорить, что сочинительная связь f содержит разделитель с, а разделитель с принадлежит сочинительной связи f.

(1) Запрещается[1] повреждать[2] или[3] загрязнять[4] покрытие[5] дорог.[6]

Союз *или* принадлежит сочинительной связи

$$f = \langle 1,\ 2,\ 3,\ 4 \rangle,$$

показанной на рис. 1. Стрелка обозначает подчинительную связь. Символ | обозначает разделитель однородных членов.

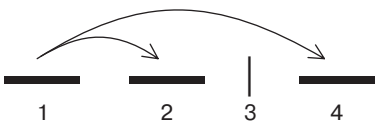


Рис. 1. Пример сочинительной связи.

Для моделирования сочинительной связи между предложениями в составе сложносочиненного предложения будем использовать виртуальное (воображаемое) главное слово с нулевым номером, у которого в подчинении находятся сказуемые b и d сочиняемых предложений:

$$f = \langle 0, b, c, d \rangle,$$

где *с* — союз или запятая, разделяющая предложения в составе сложноподчиненного.

(2) Пришел, увидел, победил.

В предложении присутствуют конструкции

$$f_1 = \langle 0,\ 1,\ 1.5,\ 2 \rangle \text{ и } f_2 = \langle 0,\ 2,\ 2.5,\ 3 \rangle.$$

Рассмотрим модель обособления. Обособление является общим понятием, обозначающим деепричастный, причастный, адъективный и другие обороты. Для целей настоящего исследования к обособлениям будем также относить и придаточные предложения. Для обособления обязательным является требование парности знаков препинания. Как правило, используются запятые, открывающая и закрывающая. В случаях, когда конец обособления совпадает с концом предложения, закрывающим знаком препинания для обособления служит знак конца предложения, например, точка.

При моделировании обособления существенными являются позиции знаков препинания — открывающего и закрывающего, а также позиции опорного и главного слов обособления, между которыми устанавливается подчинительная связь (по направлению от опорного к главному). Например, для обособляемого причастного оборота главным является причастие, а опорным – существительное в препозиции.

Таким образом, обособление также моделируется четверкой чисел:

$$g = \langle w, x, y, z \rangle,$$

где w — опорное слово обособления;
y — главное слово обособления;
x — открывающий разделитель (запятая);
z — закрывающий разделитель (запятая, точка с запят ой или знаки конца предложения).

Кроме того, *x* и *z* могут быть парными тире, скобками или кавычками.

Для g выполняется следующее соотношение:

$$x < y < z \ \&\ (w < x \ \vee\ z < w).$$

Вариант $w < x$ соответствует препозиции опорного слова, $z < w$ — постпозиции.

Будем говорить, что обособление g *содержит* разделители x и z, а разделители x и z *принадлежат* обособлению g.

(3) Водители[1] транспортных[2] средств[3] с[4] включенным[5] проблесковым[6] маячком[7] синего[8] цвета,[9(выполняя[10] неотложное[11] служебное[12] задание,[13)] могут[14] отступать[15] от[16] требований[17] настоящих[18] Правил[19].

Запятые принадлежат обособлению

$$g = \langle 14,\ 9.5,\ 10,\ 13.5 \rangle,$$

показанному на рис. 2. Скобки на рисунках и в примерах условно обозначают границы обособления.

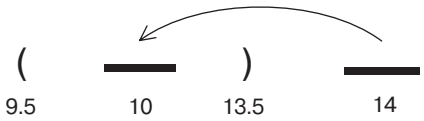


Рис. 2. Пример обособления.

Для моделирования обособления в начале предложения будем использовать виртуальную запятую с номером 0.5, расположенную между нулевым (виртуальным, воображаемым) и первым словами:

$$\mathrm{g} = \langle \mathrm{w}, 0.5, \mathrm{y}, \mathrm{z} \rangle.$$

Обособление, расположенное в конце предложения, будем моделировать четверкой чисел

$$g = \langle w, x, y, n + 0.5 \rangle,$$

где n — количество слов в предложении.

В предлагаемой модели к сочинительным союзам и ко всем знакам препинания, используемым в обособлениях и сочинительных связях, применяется общее понятие – разделитель. Будем называть *внутренним* разделитель с номером x, $1 < x < n$, где n — количество слов в предложении.

3. Формирование множества синтаксических конструкций

Обособления и сочинительные связи для краткости будем называть конструкциями. Обозначим C множество возможных (предполагаемых) конструкций, построенных автоматически на основе заданного предложения и базы правил русского языка. Описание алгоритмов построения множества предполагаемых конструкций выходит за рамки данной работы. Ознакомиться с ними можно в отчете [12].

В силу морфологической и синтаксической омонимии множество C будет содержать как истинные (те, которые построил бы человек при правильном разборе предложения), так и ложные конструкции. Истинные конструкции T, где $T \subseteq C$, соответствуют смыслу предложения и являются объектом поиска для анализатора.

Для целей дальнейшего изложения представим множество конструкций C в виде $C = R \cup I$, где R — множество сочинительных связей, I — множество обособлений.

(4) К[1] пешеходам[2] приравниваются[3] лица[4], передвигающиеся[5] в[6] инвалидных[7] колясках[8] без[9] двигателя[10], ведущие[11] велосипед[12], мопед[13], мотоцикл[14], везущие[15] санки[16], тележку[17], детскую[18] или[19] инвалидную[20] коляску[21].

Далее приводится список предполагаемых конструкций для данного предложения. Истинные конструкции помечены в списке символом V.

I — множество обособлений:	*R* — множество сочинительных связей
⟨4 4.5 15 21.5⟩	⟨4 5 10.5 11⟩ V
⟨4 4.5 11 21.5⟩	⟨11 12 12.5 13⟩ V
⟨4 4.5 15 17.5⟩	⟨3 12 12.5 13⟩
⟨4 4.5 15 16.5⟩	⟨3 4 12.5 13⟩
⟨4 4.5 11 14.5⟩	⟨11 13 13.5 14⟩ V
⟨4 4.5 11 13.5⟩	⟨3 13 13.5 14⟩
⟨4 4.5 11 12.5⟩	⟨3 4 13.5 14⟩
⟨4 4.5 5 10.5⟩ V	⟨11 14 14.5 16⟩
⟨4 10.5 15 21.5⟩	⟨11 14 14.5 17⟩
⟨4 10.5 11 21.5⟩	⟨11 14 14.5 21⟩
⟨4 10.5 15 17.5⟩	⟨4 11 14.5 15⟩ V
⟨4 10.5 15 16.5⟩	⟨15 16 16.5 17⟩ V
⟨4 10.5 11 14.5⟩ V	⟨11 16 16.5 17⟩
⟨4 10.5 11 13.5⟩	⟨11 14 16.5 17⟩
⟨4 10.5 11 12.5⟩	⟨15 17 17.5 21⟩ V
⟨4 14.5 15 21.5⟩ V	⟨11 17 17.5 21⟩
⟨4 14.5 11 21.5⟩	⟨11 14 17.5 21⟩
⟨4 14.5 15 17.5⟩	⟨21 18 19 20⟩ V
⟨4 14.5 15 16.5⟩	

4. Отношения между синтаксическими конструкциями

Рассмотрим отношения между конструкциями, вытекающие из их взаимного расположения в предложении. В общем случае две конструкции из *C* могут находиться в одном из трех отношений:

- быть соседними;
- быть вложенными одна в другую;
- быть несовместимыми.

Отношение вложенности конструкций можно разделить на четыре группы:

- сочинительная связь вложена в другую сочинительную связь;
- обособление вложено в сочинительную связь;
- обособление вложено в другое обособление;
- сочинительная связь вложена в обособление.

Дадим формальное определение указанных отношений.

Определение: Конструкции

$$f = \langle a,b,c,d \rangle \text{ и } \mathrm{g} = \langle \mathrm{w},\mathrm{x},\mathrm{y},\mathrm{z} \rangle$$

находятся в отношении соседства S, $S \subseteq C \times C$, если

$$max(a,b,c,d) \le min(w,x,y,z) \quad \& \quad f, g \in R \to d \ne x \quad \vee$$
$$min(a,b,c,d) \ge max(w,x,y,z) \quad \& \quad f, g \in R \to z \ne b.$$

Определение: Конструкции f и g находятся в отношении вложенности M, $M \subseteq C \times C$, в следующих случаях:

a) Пусть $f = \langle a,b,c,d \rangle$ и $\mathrm{g} = \langle \mathrm{w},\mathrm{x},\mathrm{y},\mathrm{z} \rangle$ — сочинительные связи.

Тогда g вложена в f, если

$$a \le w \le b \ \& \ a < x \ \& \ z \le b \ \& \ (z = b \to a = w) \vee$$
$$b \le w < c \ \& \ b < x \ \& \ z < c \ \vee$$
$$c < w \le d \ \& \ c < x \ \& \ z < d \ \vee$$
$$d \le w \le a \ \& \ d \le x \ \& \ z < a \ \& \ (d = x \to a = w).$$

Например, конструкции f и g могут располагаться так, как это показано на рис. 3.

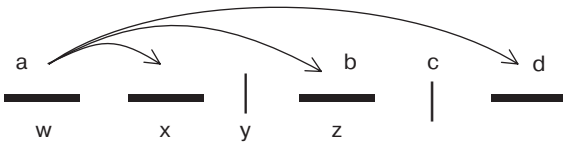


Рис. 3. Сочинительная связь вложена в сочинительную связь.

b) Пусть $f = \langle a,b,c,d \rangle$ — сочинительная связь,

$\mathrm{g} = \langle \mathrm{w},\mathrm{x},\mathrm{y},\mathrm{z} \rangle$ — обособление.

Тогда g вложена в f, если

$$a \le w \le b \ \& \ a < x \ \& \ z < b \ \vee$$
$$b \le w < c \ \& \ b < x \ \& \ z \le c \ \vee$$
$$c < w \le d \ \& \ c \le x \ \& \ z < d \ \vee$$
$$d \le w \le a \ \& \ d < \mathrm{x} \ \& \ \mathrm{z} < a.$$

Одно из возможных расположений показано на рис. 4.

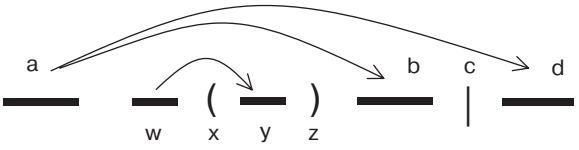


Рис. 4. Обособление вложено в сочинительную связь.

c) Пусть $f = \langle a,b,c,d \rangle$ и $\mathrm{g} = \langle \mathrm{w},\mathrm{x},\mathrm{y},\mathrm{z} \rangle$ — обособления.

Тогда g вложена в f, если

$$a \le w < b \ \& \ a < x \ \& \ z \le b \vee$$
$$b < w \le c \ \& \ b < x \ \& \ z < c \vee$$
$$c \le w < d \ \& \ c < x \ \& \ z \le d \vee$$
$$d < w \le a \ \& \ d \le x \ \& \ z < a.$$

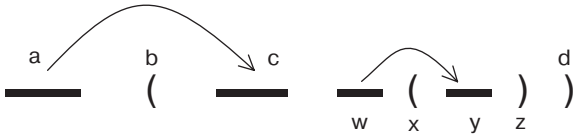


Рис. 5. Обособление вложено в обособление.

d) Пусть $f = \langle a,b,c,d \rangle$ — обособление,

$g = \langle w, x, y, z \rangle$ – сочинительная связь,

Тогда g вложена в f, если

$$a \le w < b \ \& \ a < x \ \& \ z < b \ \vee$$
$$b < w \le c \ \& \ b < x \ \& \ z < c \ \vee$$
$$c \le w < d \ \& \ c < x \ \& \ z < d \ \vee$$
$$d < w \le a \ \& \ d < x \ \& \ z < a \vee$$
$$a = w \ \& \ b < x \ \& \ z < d.$$

Наиболее характерные возможные размещения показаны на рис. 6, 7.

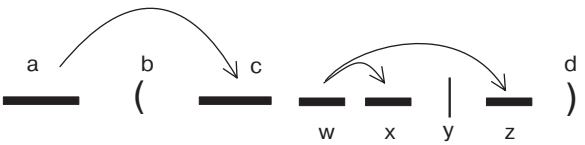


Рис. 6. Сочинительная связь вложена в обособление

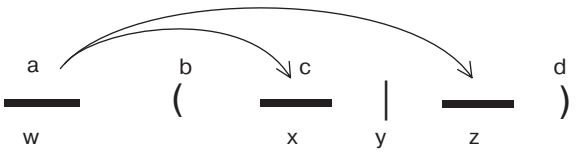


Рис. 7. Сочинительная связь вложена в обособление «Обособленный ряд».

Из определений следует, что отношения S и M не пересекаются.

Пары (f, g) из $C \times C / (S \cup M)$ составляют несовместимые конструкции. Таким образом, любые две конструкции в предложении находятся строго в одном из трех указанных отношений. При этом любая пара истинных конструкций принадлежит строго одному из отношений S или M: $T \times T \subseteq S \cup M$.

5. Формальная постановка задачи анализа пунктуации

Определим рекурсивно понятие композиции конструкций.

Определение. Комп озицией конструкций является:

1) Любая конструкция f *из* C;
   Множество конструкций
   $\{f_i\} \cup f$, $1 \le i \le |\{f_i\}|$, где $\{f_i\}$ — композиция и любая пара (f_i, f) находится в одном из отношений:
   а) $f_i \, S \, f$ б) $f_i \, M \, f$ в) $f \, M \, f_i$

Определение. Композиция конструкций содержит разделитель, если как минимум одна конструкция из композиции содержит этот разделитель.

Определение. Назовем покрытием такую композицию конструкций, которая содержит все внутренние разделители предложения.

Из определения следует, что композиция является покрытием независимо от того, содержит ли она виртуальный разделитель в начале предложения и знак препинания в конце предложения.

В рамках предлагаемой модели задача определения функций знаков препинания для заданного предложения русского языка сводится к построению покрытия. Следует заметить, что в общем случае покрытие строится неоднозначно.

6. Алгоритм построения покрытия

Рекурсивный алгоритм построения покрытия вытекает из определения композиции.

Исходные данные:

C — множество возможных конструкций;

$|C|$ — мощность C;

Q — композиция конструкций, изначально $Q = \varnothing$;

k — номер конструкции в C, с которой начинается поиск, изначально $k = 1$;

Возвращаемое значение:

true, если удалось построить композицию, *false* — иначе.

Поиск покрытия (C, Q, k)

Цикл $i = \overline{k, |C|}$ {

Если $(\forall K \in Q: \ (K, C[i]) \in M \cup S)$ {

$$Q_1 = Q \ \cup \ \{C[i]\};$$

Если (Поиск покрытия $(\tilde{N}, Q_1, i+1)$),
то вернуть *true*;
}
}
Если (Q содержит все внутренние разделители), то вернуть *true* (Q является покрытием);
Иначе вернуть *false*.

Данный алгоритм является переборным и представляет лишь теоретический интерес. В работе [12] приведены эвристики, применимые на практике.

7. Учет пунктуации в синтаксическом анализе

Как известно, синтаксический анализ сводится к выбору подчинительных связей, образующих дерево, из числа потенциально возможных связей. Простейший способ учета пунктуации в синтаксическом анализе основан на использовании покрытия в качестве фильтра: имея покрытие, можно провести удаление значительной части ложных синтаксических связей.

Рассмотрим эту идею отдельно для обособлений и рядов, составляющих покрытие. Пусть конструкция $f = \langle a, b, c, d \rangle$ является обособлением, т. е. *b* и *d* — это границы обособления, *(a, c)* — подчинительная связь. Тогда следует считать ложными и удалить все связи, конфликтующие со связью *(a, c)*, а также все связи, охватывающие одну из границ обособления — *b* или *d*, кроме, разумеется, *(a, c)*.

Пусть конструкция $f = \langle a, b, c, d \rangle$ является сочинительной связью, т.е. *c* – это разделитель, *(a, b)* и (*a, d*) – подчинительные связи. Тогда следует считать ложными и удалить все связи, конфликтующие со связями *(a, b)* и *(a, d)*, а также все связи *(x, y)*, где $b<x<c<y<d$. Направление связи *(x, y)* не имеет значения.

Отметим, что встраивание разработанных методов анализа пунктуации в синтаксический анализ является самостоятельной задачей и сильно зависит от применяемых подходов к синтаксическому анализу.

В рамках подхода, реализованного в системе синтаксического анализа DictaScope [11, 12], покрытие строится одновременно с деревом разбора. Анализатор позволяет получить в качестве результата не только структуру подчинительных связей, но и информацию о рядах и обособлениях. При этом общая вычислительная сложность алгоритма составляет $O(n^3)$, где n — количество слов в анализируемом предложении. Кроме того, благодаря учету пунктуации удалось существенно поднять качество синтаксического анализа.

8. Заключение

Проведенные исследования позволяют сделать следующие выводы:

- Раздельный анализ обособлений и рядов однородных членов невозможен в силу функциональной омонимии запятой;
- Структура обособлений и рядов образует покрытие в любом грамматически правильном предложении русского языка;
- Разработанные модели и методы анализа пунктуации применимы к языкам, составляющим индоевропейскую группу.

Исследование проводилось малым инновационным предприятием ООО «Диктум» при поддержке Фонда содействия развитию малых форм предприятий в научно-технической сфере.

Список литературы

1. *Бердичевский А. С., Иомдин Б. Л.* Роль пунктуации в разрешении неоднозначности. //Труды Международной конференции Диалог'2007.
2. *Мальковский М. Г., Старостин А. С.* Модель синтаксиса в системе морфосинтаксического анализа «Treeton» // Труды Международной конференции Диалог'2006.
3. *Кобзарева Т. Ю., Лахути Д. Г., Ножов И. М.* Модель сегментации русского предложения // Труды Международной конференции Диалог'2001. Т. 2. Аксаково 2001.
4. *Ножов И. М.* Морфологическая и синтаксическая обработка текста (модели и программы) // М., 2003.
5. *Гершензон Л. М., Панкратов Д. В.* Фрагментационный анализ русского предложения в системе Artefact // Труды Международного семинара Диалог'2002.
6. *Кобзарева Т. Ю.* Рекурсивность и проективность сочинительных связей в русском тексте // Труды Международной конференции Диалог'2006.
7. *Кобзарева Т. Ю.* Омонимия и синонимия знаков препинания в русском тексте //Труды Международной конференции Диалог'2005.
8. *Кобзарева Т. Ю.* Построение графа связей сегментов // Труды Международной конференции Диалог'2008.
9. *Розенталь Д. Э., Голуб И. Б., Теленкова М. А.* Современный русский язык// Айрис Рольф. — М.: Пресс 2001 г
10. *Bernard E. M. Jones.* Exploring the role of punctuation in parsing natural text //Proceedings of the 15th conference on Computational linguistics'1994. — V. 1, pp. 421–425.
11. *Электронный* ресурс www.dictum.ru
12. *Окатьев В. В., Гергель В .П., Алексеев В. Е., Таланов В. А., Баркалов К. А., Скатов Д. С., Ерехинская Т. Н., Котов А. Е., Титова А. С.* Отчет о выполнении НИОКР по теме: «Разработка пилотной версии системы синтаксического анализа русского языка» (инвентарный номер ВНТИЦ 02200803750) // М.: ВНИТЦ, 2008

Перевод немецкой частицы *doch* на русский язык (в контексте констативов): *ведь*, *же*, *всё же* или *всё-таки*?

Translation of German particle *doch* used in statements into Russian (in statements): *ved'*, *že*, *vse že* or *vse-taki*?

Орлова С. В. (svetlachok-star@yandex.ru)
Московский государственный университет им. М. В. Ломоносова

Данная работа представляет собой сравнительный анализ семантики немецкой частицы *doch* в контексте констативов и ее словарных переводных эквивалентов в русском языке — частиц *ведь*, *же*, *всё же* и *всё-таки*. Этот доклад является очередным в серии докладов, посвященных проблеме перевода немецких модальных частиц на русский язык.

Одной из наибольших трудностей, с которой сталкивается переводчик с немецкого языка на русский, является перевод частиц. Обращение к имеющимся двуязычным словарям для выбора подходящего переводного эквивалента в русском языке лишь частично позволяет справиться с этой трудностью. Однако контрастивные исследования немецких и русских частиц дают возможность выявить столь важные для перевода, но не зафиксированные в словарях сходства и различия в семантике этих дискурсивных слов. Примером подобного анализа является проведенное нами сопоставительное исследование семантики немецкой частицы *ja* и ее словарных переводных эквивалентов *же* и *ведь* ([3]), а также данная работа, посвященная анализу частицы *doch* и русских *же*, *ведь*, *всё же* и *всё-таки*, предлагаемых для ее перевода немецко-русскими словарями.

Материалом для анализа послужили 1) данные двуязычных словарей по переводу немецкой частицы *doch* на русский язык, 2) анализ употреблений немецкой частицы *doch* и русских частиц *же*, *ведь*, *всё же* и *всё-таки* в семантических словарях и специальных исследованиях, 3) корпус параллельных фрагментов (около 200), содержащих частицу *doch*, из немецких художественных произведений[1] и их переводов на русский язык профессиональными переводчиками.

Как показывают данные немецко-русских словарей, в качестве переводного эквивалента для частицы *doch* в констативах могут выступать частицы *ведь*, *же* (и их сочетание), *всё-таки*, *всё же*. Приведем пример словарной статьи для *doch*:

doch I. cj (…) 2) всё-таки, всё же
Ich kann ihm das nicht antun, er bleibt doch immer mein Bruder. Я не могу так поступить с ним, всё-таки он мой брат.
Ich will es doch wagen. Всё же я хочу рискнуть.
Das müsste ich doch wissen. Я должен был бы всё-таки знать это.

II. prtc (…) 2) выражает усиление; ведь, же, уж (часто не переводится)
Er ist doch sehr unglücklich. Он ведь очень несчастен.
Ich habe es dir doch gesagt! Я же / ведь тебе это сказал! (…) [1]

В словарях не представлены условия выбора той или иной частицы при переводе, а кроме того, не делается различия между ударной и безударной *doch*[2], хотя, по данным немецких лингвистов, они выполняют отличные друг от друга функции в предложении. Анализ примеров реального использования частицы в контексте подтвердил необходимость семантически разграничивать случаи употребления *doch* в безударной и ударной позиции. Далее мы рассмотрим инвариантное значение немецкой частицы *doch* и ее русских переводных эквивалентов *же*, *ведь*, *всё же* и *всё-таки*, представленное в различных

[1] В корпусе представлены фрагменты из произведений Э.-М. Ремарка (пер. И. Шрайбера), Ф. Кафки (пер.С. Апта и А. Махлиной), Т. Манна (пер. Т. Исаева) и сказок Братьев Гримм (пер. под ред. П. Полевого и Г. Петникова).

[2] Мы здесь не рассматриваем *doch* в функции самостоятельной реплики-реакции, когда, безусловно, эта частица является ударной.

исследованиях, и определим области пересечения семантики этих частиц.

1. Инвариантные значения частицы *doch* и ее словарных переводных эквивалентов *всё же, всё-таки, ведь* и *же*

Общее значение частицы *doch*, согласно Хельбигу ([6]), „заключается в противительном компоненте (в противоречии двух точек отсчета), (…) иногда противопоставление лишь предполагается, но не выражено эксплицитно. С помощью *doch* Говорящий подтверждает свою установку по отношению к сказанному или подтверждает существование/отсутствие некоторого положения дел (в противоположность предшествующему контексту или установке собеседника)“. Как видно из данной формулировки, основным компонентом в инвариантном значении *doch* является сема „противопоставление“, „противоречие". Речь идет о несоответствии одного положения дел (о котором сообщает говорящий) другому или позиции говорящего и слушающего. Поэтому говорящий, используя частицу *doch*, во-первых, всегда делает отсылку к предшествующему контексту, а во-вторых, вступает в некоторую оппозицию (или даже конфликт): он опровергает, противоречит, не соглашается с чем-либо.

(1) *A: Mach das Fenster zu!*
A: Закрой окно!
B: Es ist doch viel zu warm im Zimmer.
B: Но (ведь) в комнате (же) слишком тепло!
(примеры из [6])

Так, в примере (1) говорящий отклоняет просьбу закрыть окно, так как считает ее неуместной, противоречащей обстоятельствам — в комнате и так слишком тепло.

Теперь обратимся к семантике частиц *всё-таки* и *всё же*. В словаре Шимчук и Щур [5] можно найти перечень конкретных случаев употребления этих частиц и описание значения частиц как вклада, который они вносят в семантику предложения. Для обеих частиц важным являетя именно элемент противопоставления. Так, *всё-таки* может выражать существование события, признака или точки зрения на некоторое положение дел, которые оцениваются как „нежелательные, невозможные или маловероятные“ в силу того, что им противопоставляется другое событие, признак или точка зрения. Иначе говоря, ’P имеет место, несмотря на то, что P нежелательно’, или ’P имеет место, несмотря на то, что P маловероятно или невозможно, так как Q’.

(2) *Обещал, а всё-таки сделал по-своему.* (пример из [5])

В семантике *всё же* также прочно закреплен противительный компонент: некоторое событие, факт или признак имеют место вопреки чему-либо (’P имеет место несмотря на то, что Q’) или же некоторое положение дел противоречит, не соответствует представлениям, установкам и желаниям говорящего (’я считаю, что P, несмотря на то, что Q’).

(3) *Петров не мог так поступить: что ни говори, он всё же порядочный человек.* (пример из [5])

В отличие от рассмотренных выше частиц, инвариантные значения *ведь* и *же* не содержат подобного компонента противопоставления. Согласно Бонно и Кодзасову ([2]), «*ведь* указывает на то, что вводимая информация, будучи адекватной, является одновременно релевантной для правильной интерпретации ситуации адресатом речи. Гарантом адекватности является действительность, гарантом релевантности — говорящий». Или, как отмечено в работе Левонтиной [4], говорящий, используя частицу *ведь*, тем самым имеет ввиду: ‘Я знаю и считаю, что это нужно принять во внимание’.Общим элементом значения *ведь* и *doch* является, таким образом, заинтересованность говорящего в передаче слушающему правильной интерпретации положения дел.

Инвариантное значение частицы *же*, по мнению Бонно и Кодзасова, состоит в том, что она «маркирует сохранение точки слежения в сфере введенного в предтексте когнитивного объекта». Исходя из этой формулировки, трудно сделать какие-либо выводы о наличии общих компонентов семаники *же* и *doch*. Однако на уровне конкретных употреблений *же* и *doch* обнаруживаются сходства в семантике этих частиц. Аналогично частицам *doch* и *ведь*, *же* может вводить в диалог некоторый антитезис, аргумент "против", при этом происходит обязательная отсылка к предшествующему контексту и подчеркивается несоответствие ему.

(4) *Да как он вернется?! Дорогу же размыло!* (пример из [2])

(5) *Ты что форточку открыл — ребенка ведь простудишь!* (пример из [2])

(6) *Ich weiß nicht, warum du immer Kaffee trinkst. Milch ist doch viel gesünder.*
Я не знаю, почему ты всегда пьешь кофе. Молоко ведь / же намного полезнее. (пример из [9])

Итак, рассмотрев инвариантные значения частицы doch и ее возможных переводных эквивалентов *ведь*, *же*, *всё-таки* и *всё же*, мы выделили некоторые семантические сходства этих частиц,

иными словами, основу или предпосылку для выбора указанных русских частиц в качестве перевода для немецкой *doch*. Теперь обратимся к конкретным случаям употребления немецкой частицы и проанализируем, что общего и различного в функционировании *doch* и русских *ведь, же, всё-таки* и *всё же*.

2. Употребление частицы *doch* в безударной позиции

2.1. Doch в монологическом отрезке дискурса

Один из распространенных случаев употребления немецкой частицы *doch* — в утверждениях, где она, согласно Хельбигу, выполняет функцию усиления установки говорящего и способствует преодолению существующего в момент речи противоречия. Используя *doch*, говорящий (далее — Г) напоминает слушающему (далее — С) о чем-то ему известном, однако не учтенном в данный момент (чаще всего — имевшим место в прошлом и потому забытом). Тем самым Г актуализирует некоторую информацию в сознании слушающего, с целью изменить его представление о положении дел и получить с его стороны согласие или одобрение. Наличие противоположной установки является предпосылкой для введения реплики с частицей *doch*, что отличает *doch* от похожей по семантике частицы *ja* , поскольку *ja* уже изначально предполагает согласие С с вводимой информацией. Сравн. примеры из ([6]):

(7) *Wir wollten doch heute Abends ins Theater gehen. Мы ведь / же хотели сегодня пойти в кино.* (Мы договорились раньше об этом, но ты забыл)

(8) *Wir wollten ja heute Abends ins Theater gehen. Мы ведь / же хотели сегодня пойти в кино.* (Ты полностью согласен со мной, поскольку знаешь и помнишь о том, что мы договорились)

Что касается коммуникативной структуры высказывания, то, используя в речи частицу *doch*, Г актуализует уже известную С информацию, однако вводит ее в сознание С заново. Таким образом, статус информации в реплике с *doch* можно обозначить как одновременно «известное» и «новое» для С.

В данной функции *doch* ведет себя аналогично русской частице *ведь* в случае введения некоторого аргумента «против» установки или поведения слушающего. Актуализационный статус высказывания с *ведь* отличается однако возможным варьированием признака «известность» информации: Г может вводить как известный, так и неизвестный для С аргумент. Тем не менее, посредством *ведь* информация вводится как нечто «новое», она впервые или заново актуализуется в сознании С, что позволяет использовать *ведь* в качестве возможного перевода для *doch*.

(9) »Nun, was ist dir in die Quere gekommen, alter Bartputzer?« sprach der Esel.

»Wer kann da lustig sein, wenn's einem an den Kragen geht«, antwortete die Katze, »weil ich nun zu Jahren komme, meine Zähne stumpf werden und ich lieber hinter dem Ofen sitze und Spinne als nach Mäusen herumjage, hat mich meine Frau ersäufen wollen; ich habe mich zwar noch fortgemacht, aber nun ist guter Rat teuer: wo soll ich hin?«
»Geh mit uns nach Bremen, du verstehst dich ***doch*** *auf die Nachtmusik, da kannst du ein Stadtmusikant werden.«*

— Ну, что, старина, Кот Котофеич, беда, что ли, какая с тобой приключилась? — спрашивает его осел.
— Да как же мне быть веселым, когда дело о жизни идет, — отвечает кот, — стал я стар, зубы у меня притупились, сидеть бы мне теперь на печи да мурлыкать, а не мышей ловить, — вот и задумала меня хозяйка утопить, а я убежал подобру-поздорову. Ну, какой дашь мне добрый совет? Куда ж мне теперь деваться, чем прокормиться? — Пойдем с нами в Бремен, — ты ***ведь*** *ночные концерты устраивать мастер, вот и будешь там уличным музыкантом.*

Следует отметить, что частица *ведь* отличается от немецкой *doch*, а также и от других рассматриваемых нами частиц тем, что подаваемая информация расценивается Г как абсолютно правильная, поскольку соответствует действительному положению дел. Пользуясь терминами Бонно и Кодзасова, Г в высказывании 'ведь Р' сообщает некоторую адекватную информацию, причем гарантом адекватности выступает именно действительность. В связи с этим частица *ведь* не употребляется в императивах[3] (сравн. примеры (10) и (11).

(10) Напиши *ведь / же / всё же / всё-таки письмо отцу!

(11) Schreib doch deinem Vater!

Поэтому выбор *ведь* в качестве перевода для *doch* ограничен случаем введения аргумента «против» установки слушающего, когда говорящий пода-

[3] За исключением особого вида императивов, о которых писала Левонтина в [4], см. *Ведь ты учти, что это уже не в первый раз!*

ет аргумент как некоторое знание, в правильности которого он уверен (как в примере (12).

(12) *»Wo sind Sie denn nur so geisterhaft hergekommen? Ich habe doch die ganze Zeit die Tür beobachtet.«*
— Откуда это вы появились, словно призрак? Ведь я всё время следил за дверью.

Частица *же* может также вводить аргумент «против», однако вводимая информация имеет актуализационный статус «данного», то есть, по мнению говорящего, имеется не просто в фонде знаний собеседника, но и в его текущем сознании. Г при этом ведет себя с так называемой «риторической активностью» ([2]). Иными словами, использование *же* в данной функции показывает желание Г воздействовать на С, нередко для достижения собственных интересов. Поэтому переводчик выбирает *же* в качестве переводного эквивалента для *doch* именно при активной позиции Г, его заинтересованности в воздействии на С: в ситуации убеждения, уговаривания, упрека, оправдания.

(13) »Machen wir eine Probefahrt, Herr Blumenthal«, schlug ich schließlich, schon stark abgekämpft, vor. »Probefahrt?« erwiderte er, als hätte ich Bahnhof gesagt.
»Ja, Probefahrt. Sie müssen ***doch*** *sehen, was der Wagen leistet.*

— Давайте сделаем пробную поездку, господин Блюменталь, — предложил я наконец, уже основательно измочаленный.
— Пробную поездку? — переспросил он так, словно я предложил ему искупаться.
— Ну да, проедем. Вы ***же*** *должны сами убедиться, на что способна машина.*

Отметим, что употребление частицы *doch* в констативах не ограничивается случаем введения в речевой акт аргумента. Ведь *doch* используется Г с целью напоминания С о чем-либо, так или иначе противоречащем представлениям или установкам С. Высказывание ’*doch* Р’ имеет в качестве предпосылки компонент ’Слушающий, вероятно, думает, что не Р’. Такое положение дел охватывается сферой употребления частиц *всё же* и *всё-таки*.

Так, *всё же* имеет место в одном из двух частных случаев противоречия: 1) Г сообщает о некотором событии, факте или признаке, существующем вопреки другим обстоятельствам; 2) Г выносит некоторую оценку, опровергающую те или иные обстоятельства. Действительно, если обстоятельства, о которых С уже известно, противоречат Р, то С предполагает ’не Р’. Г посредством *всё же* опровергает его представления (ожидания). Поэтому переводчик может использовать в данном контексте *всё же* в качестве перевода для немецкой частицы *doch*:

(14) *Jetzt wird es bald drei Jahre her sein, da war ja mein Freund bei uns zu Besuch. Ich erinnere mich noch, daß du ihn nicht besonders gern hattest. Wenigstens zweimal habe ich ihn vor dir verleugnet, trotzdem er gerade bei mir im Zimmer saß. Ich konnte ja deine Abneigung gegen ihn ganz gut verstehn, mein Freund hat seine Eigentümlichkeiten. Aber dann hast du dich* ***doch*** *auch wieder ganz gut mit ihm unterhalten. Ich war damals noch so stolz darauf, daß du ihm zuhörtest, nicktest und fragtest.*

Уже скоро три года тому, как мой друг не был у нас в гостях. Насколько я помню, ты его особо не жаловал. Не меньше двух раз я скрыл от тебя, что он сидел у меня в комнате. Я вполне мог бы понять твою к нему неприязнь, у моего друга есть свои странности. Но как-то раз вы ***все же*** *вели очень приятную беседу. Я был еще так горд тем, что ты слушаешь его, киваешь и задаешь вопросы.*

Частица всё-таки также вносит в семантику предложения элемент противоречия. Говоря ’всё-таки Р’, мы сообщаем о том, что Р имеет место вопреки обстоятельствам, хотя Р оценивается как нежелательное, невозможное или маловероятное. Тем самым, Г в русском языке посредством всё-таки, а в немецком посредством doch опровергает интерпретацию С, что ’не Р’. Это хорошо видно из (15):

(15) *Aber selbst wenn die Schwester, erschöpft von ihrer Berufsarbeit, dessen überdrüssig geworden war, für Gregor, wie früher, zu sorgen, so hätte noch keineswegs die Mutter für sie eintreten müssen und Gregor hätte* ***doch*** *nicht vernachlässigt werden brauchen. Denn nun war die Bedienerin da.*

Но даже когда сестре, измученной службой, надоело заботиться, как прежде, о Грегоре, матери не пришлось заменять ее, но без присмотра Грегор ***все-таки*** *не остался. Теперь пришел черед служанки.*

Как *всё же*, так и *всё-таки* могут вводить известную или неизвестную для С информацию, при этом подают ее как нечто новое для С. Тем самым коммуникативная структура высказывания с *всё же* и *всё-таки* не противоречит структуре высказывания ’*doch* Р’.

Нередко переводчики оставляют частицу *doch* без перевода (см. пример (16) или обращаются к иным языковым средствам. Всё зависит от языковой интуиции переводчика, от того, насколько он ощущает уместность использования той или иной частицы в качестве переводного эквивалента.

(16) »Wollen Sie jetzt etwas fahren?« fragte ich. »Es macht Ihnen doch sicher Spaß.«

— Хотите немного поводить? — спросил я. — Это вам доставит удовольствие.

(В данном примере для использования *ведь* и *же* в функции аргумента "против" нет оснований, так как в вопросе не содержится в явном виде тезис, который бы допускал введение аргумента, а тем более в таком контексте не выражены обстоятельства или противоположные установки слушающего, которые можно было бы отвергать при помощи *всё же* или *всё-таки*. Г лишь предполагает возможную позицию С о том, что вождение не доставит удовольствия, и торопится опровергнуть эту позицию, чтобы предупредить отказ от своего предложения).

2.2. Doch в репликах-реакциях

Согласно Хельбигу, типичным для употребления частицы *doch* является введение репликиреакции на предшествующий речевой акт или положение дел. При этом Г подвергает критике или не принимает (отклоняет) речевой акт собеседника. Причиной несогласия Г может быть нарушение вторым участником диалога пресуппозиций (см. пример (17), условий успешности речевого акта (см. пример (18) (тогда реплика приобретает статус корректирующей),

(17) *A: Das neue Kleid passt dir so gut!*
B: Das Kleid ist doch nicht neu…Ich habe es von dir als Geschenk vor einem Jahr bekommen..
A: Тебе так идет новое платье!
B: (Но) платье ведь / же не новое… Ты подарил мне его год назад…

(18) *A: Du hast aber wenig Fleisch gekauft.*
B: Ich konnte doch nicht wissen, dass wir Besuch bekommen.
A: Ты купил мало мяса.
B: (Но) я ведь / же не знал, что у нас будут гости. (пример из [6])

или явная оппозиция Г по отношению к установке, выраженной адресатом:

(19) *A: Mit kleinen Hunden hat man keine Probleme.*
B: Sie sind doch so laut!
A: С маленькими собаками нет проблем.
B: (Да) от них ведь / же столько шума! (пример из [9])

Чаще всего подобная реплика-реакция приобретает сильную эмоциональную окраску и по иллокутивному типу представляет собой протест, возмущение, упрек, оправдание. Поэтому наиболее подходящим переводом для *doch* в этом случае является *же*, вносящая в семантику предложения благодаря «риторической активности» говорящего элемент конфликтности.

(20) *„Entschuldigung, Herr Doktor. Haben Sie absichtlich keine Krawatte umgebunden?“*
„Wieso? Fragte Hagedorn. Ich war ***doch*** *extra deswegen noch einmal in meinem Zimmer!“*

— Извините, господин кандидат. Вы умышленно не повязали галстук?
— Как так? — удивился Хагедорн. — Я ***же*** *из-за этого специально вернулся в номер!*

В случае выбора частицы *ведь* в качестве переводного эквивалента для *doch* переводчик компенсирует недостаточную степень конфликтности с помощью противительного союза *но* или частицы *да*, возможно также сочетание *ведь же* (см. пример (21). Тем самым привносится компонент противопоставления, противоречия, отсутствующий в семантике *ведь*.

(21) *»Wo liegt sie denn?« frage ich.*
»Im Luisenhospital«, sagt mein Vater.
»In welcher Klasse?«
»Dritter. (…) Sie wollte selbst dritter liegen. Sie sagte, dann hätte sie etwas Unterhaltung. Es ist auch billiger.«
»Dann liegt sie ***doch*** *mit so vielen zusammen. Wenn sie nur nachts schlafen kann.«*

— Где же она лежит? — спрашиваю я.
— В госпитале святой Луизы, — говорит отец.
— В каком классе?
— В третьем. (…) Она сама хотела, чтоб ее положили в третий. Она сказала, что там ей будет не так скучно. К тому же, это дешевле.
— ***Но ведь*** *там столько народу в одной палате! Она, пожалуй, не сможет спать по ночам.*

Для частиц *всё же* и *всё-таки* оказалось нетипичным употребление в качестве перевода немецкой *doch* в контексте корректирующей репликиреакции, однако они часто встречаются в контексте реакции, выражающей несогласие говорящего с предшествующим речевым актом и соответствуют *doch* в ударной позиции.

3. Употребление частицы *doch* в ударной позиции

Немецкая частица *doch* в контексте констативов способна нести на себе фразовое ударение (далее — *DOCH*). Некоторые немецкие лингвисты относят *DOCH* в этом случае к категории наречия (см. [6]), некоторые — к категории модальной частицы

(см. [8], [9]). Сообщая *'DOCH* P', говорящий делает отсылку к исходному положению 'P' и при этом корректирует эксплицитно высказанное утверждение о том, что 'не P' или же опровергает очевидный вывод о том, что 'не P'. Таким образом, выстраивается цепочка: 'P'— 'не P'— 'не не P'.

В качестве перевода для ударной *DOCH* из четырех ранее рассмотренных частиц используются только *всё же* и *всё-таки*. Частица *всё же* является более нейтральным вариантом, в то время как *всё-таки* подчеркивает наличие противоположной точки зрения о том, что P маловероятно или нежелательно.

(22) *Der eine rief: "Ei, da führt er die Königstochter vom goldenen Dache heim." "Ja", antwortete der zweite, "er hat sie noch nicht." Sprach der dritte: "Er hat sie* ***dóch****, sie sitzt bei ihm im Schiffe."*

И молвил один из воронов:
— Э, да это он везет к себе домой королевну с Золотой Крыши.
— Да, — ответил второй ворон, — но она ему еще не принадлежит.
А третий ворон сказал:
— А ***все-таки*** *она его, ведь она находится у него на корабле.*

(23) *Один ворон воскликнул:*
— Эге, вот он и везёт к себе королевну с золотой крыши.
— Да, — сказал другой, — везёт-то везёт, да довезёт ли? Третий вступился:
— А ***всё же*** *она у него в руках и сидит у него в каюте* (в другом переводе).

Как видно из примеров, употребление *ведь* или *же* в подобного рода контекстах неуместно, что объясняется отсутствием у обеих частиц функции выражения «двойного отрицания» ('не не P').

Сопоставительный анализ поведения *doch* и ее словарных переводных эквивалентов *ведь*, *же*, *всё же* и *всё-таки* в контексте констативов выявил, что все указанные частицы способны выражать некоторое противопоставление или противоречие в высказывании. Важность противительного компонента сближает по семантике немецкую частицу *doch* с русскими *всё же* и *всё-таки*. Так, при помощи этих частиц вводится 1) сообщение или оценка, которые противоречат некоторым обстоятельствам (ранее упомянутым в контексте), а вместе с тем, ожиданиям или установкам слушающего; 2) реплика-реакция на отрицание собеседником исходного высказывания говорящего (*doch* в данном случае ударная). Выбор частицы *всё-таки* маркирует оценку сообщаемого как нежелательного или маловероятного. С другой стороны, немецкая частица *doch* и русские *же* и *ведь* выполняют также общие функции, а именно: при введении 1) аргумента против позиции или поведения слушающего; 2) корректирующей реплики; 3) реплики-реакции, выражающей резкое несогласие говорящего с высказыванием собеседника. При этом степень конфликтности реплики с *doch* на порядок выше, чем в случае с *ведь*, что часто компенсируется при переводе добавлением к *ведь* противительного союза или частицы (в инициальной реплике *ведь* всегда требует присутствия такого дискурсивного слова). Использование же частицы *же* повышает риторическую активность говорящего, то есть его заинтересованность в воздействии на позицию или поведение слушающего (аналогично при переводе *ja* с помощью *же*, см. [3]).

Таким образом, макет словарной статьи для немецкой частицы doch в контексте констативов может выглядеть приблизительно так:

doch (…) в утвердит. предл., при описании или констатации некоторого положения дел
1. при введении аргумента против, исправлении собеседника, несогласии с его позицией а) *ведь* (часто *но ведь*, *да ведь*), б) *же* (повышает конфликтность, показывает стремление говорящего изменить позицию или поведение собеседника);
2. при введении сообщения или оценки, противоречащей обстоятельствам, упомянутым ранее а) *всё же*, б) *всё-таки* (показывает, что положение дел имеет место, несмотря на оценку его как нежелательного или маловероятного);
3. ударн. при исправлении собеседника, отрицающего исходное сообщение или оценку говорящего а) *всё же*, б) *всё-таки* (см. предыд. пункт)

В целом, результаты данного исследования могут использоваться при переводе, составлении немецко-русских словарей, а также при преподавании и изучении русского языка как иностранного.

Литература

1. *Большой* немецко-русский словарь в 3 томах. Под ред. О. И. Москальской. // М.: Русский язык, 2001.
2. *Бонно К., Кодзасов С. В.* Семантическое варьирование дискурсивных слов и его влияние на линеаризацию и интонирование (на примере частиц же и ведь) // Дискурсивные слова русского языка: опыт контекстно-семантического описания. Под ред. К. Киселевой и Д. Пайара. М.: 1998, 382–443.
3. *Кобозева И. М., Орлова С. В.* Одноклеточные организмы общения под микроскопом: немецкая частица ja в сопоставлении с ее переводными эквивалентами ведь и же. // Компьютерная лингвистика и интеллектуальные технологии, 7/14, М.: 2008.
4. *Левонтина И.* Об одной загадке частицы ВЕДЬ // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2005», М.: 2005.
5. *Шимчук Э., Щур М.* Словарь русских частиц. // Berliner slavistishe Arbeiten, B. 9, Frankfurt am Main: 1999.
6. *Helbig G.* Lexikon deutscher Partikeln // Leipzig: Langenscheidt Verlag Enzyklopädie, 1998.
7. *Langenscheidts* Grosswörterbuch Deutsch-Russisch // Berlin und München: Langenscheidt KG, 1997.
8. *Meibauer J.* Modaler Kontrast und konzeptuelle Verschiebung // Studien zur Syntax und Semantik deutscher Modalpartikeln, Tübingen: Max Niemeyer Verlag, 1994.
9. *Weydt. H, Harden Th., Hentschel E. und Rösler D.* Kleine deutsche Partikellehre // Stuttgart: Ernst Klett Verlag, 1983.

Лексикографическая структура этимологического словаря и его представление в цифровой среде

Etymological dictionary: lexicographic structure and representation in digital environment

Остапова И. В. (iros@zeos.net)

Украинский языково-информационный фонд НАН Украины, Киев, Украина

На основе формальной модели лексикографической системы Этимологического словаря украинского языка разработана технология построения инструментальной системы для поддержки функционирования словаря в цифровой среде. Основное внимание уделяется механизму языковой индексации словаря.

Этимологический словарь украинского языка (далее — ЭСУЯ), представляет собой фундаментальный лексикографический труд, который создаётся в рамках проекта формирования национальной словарной базы Украины [1]. Первый том был издан в 1982 году, выход шестого тома ожидается в 2009 году; седьмой том должен будет представлять индекс ко всему словарному массиву. Для этимологических словарей наиболее действенным поисковым инструментом является индекс по языковой принадлежности слов, с которыми устанавливается генетическая связь в каждой словарной статье.

На сегодняшний день (на материале 5-ти вышедших из печати томов) представлены уже 232 различных языка. Для каждого языка в словаре необходимо создать отдельный индекс с идентификацией всех точных текстовых локализаций каждого слова этого языка. Для всего массива словаря предполагаемая размерность индекса — около 120 тысяч единиц. Трудоемкость работы по построению такого индекса столь велика, что задача создания его в «ручном» режиме не представляется технологически оправданной. Поэтому была поставлена задача разработки специальной цифровой лексикографической среды, адаптированной к структурам ЭСУЯ и ориентированной на создание моногоязыкового индекса в автоматическом режиме.

Цифровая среда представляет собой качественно новый уровень сервиса для исследовательской работы с лингвистической информацией, представленной в словарной форме. И в первую очередь это относится к индексным системам. Под индексацией словаря мы понимаем набор формализованных правил и процедур, на основании которых можно получить информацию об определённых языковых фактах, зафиксированных в словаре. Реализуются эти правила в форме пользовательских интерфейсов. Однако следует учитывать тот факт, что эффективность автоматического построения индексных схем для цифрового словаря возможна только в достаточно формализованной среде.

При выполнении работы по созданию цифровой версии ЭСУЯ использовались методы, которые уже был успешно опробованы для решения подобных задач, в частности, для создания компьютерной лексикографической базы данных нового толкового Словаря украинского языка [3].

Мы рассматриваем словарь как информационную систему особого типа — лексикографическую. Согласно теории лексикографических систем это абстрактный языково-информационный объект, ориентированный на реализацию комплексного информационного описания лексико-грамматических структур определённого языка или совокупности языков [3].

Архитектура системы отвечает стандартной трёхуровневой архитектуре информационных систем ANSI/SPARK, согласно которой в информационной системе выделяются концептуальный, внутренний и внешний уровни данных [2].

В качестве концептуальной модели используется лексикографическая модель данных [3]. Ниже мы приводим её в несколько упрощённом виде:

$$\{I_0(D),\ V(I_0(D)),\ \beta,\ \delta[\beta],\ Red[V(I^Q(D))]\},$$

где D — объект моделирования — Этимологический словарь украинского языка; $I_0(D) = \{x_1\}$ множество

реестровых единиц словаря, в теории лексикографических систем его принято называть множеством *элементарных информационных единиц*; $V(I_0(D))$ — множество описаний (интерпретаций) элементарных информационных единиц, то есть текстов словарных статей: $V(I_0(D)) = \{V(x_i)\}$ — словарная статья с заголовковым словом (реестровой единицей) x_i; β — множество структурных элементов, которые были абстрагированы в результате анализа текста словаря; δ[β] — структура, которая порождается на β оператором δ; ограничения δ[β] на $V(x)$ порождает микроструктуру словарной статьи δ(x); $Red[V(I_0(D)]$ — механизм *рекурсивной редукции лексикографической системы*. Он даёт возможность последовательно выявлять всё более тонкие детали структуры лексикографической системы, в частности — осуществлять распределение структурных элементов словарной статьи на реестровую и интерпретационную части.

Концептуальная модель словаря строится на основе анализа полиграфической версии ЭСУЯ, то есть анализируется типографское оформление, организация и структура печатных текстов словарных статей, которые интерпретируются как идентификаторы соответствующих элементов лексикографических структур β и $\delta(x)$.

В качестве базового структурного элемента лексикографической системы ЭСУЯ мы определяем *этимологический класс*, который представляет собой блок линейного текста словарной статьи, в котором описаны определённые генетические связи реестрового украинского слова. Вычленение этимологических классов выполняется по формальным признакам: структурная единица идентифицируется как этимологический класс, если в тексте словарной статьи можно выявить уникальные знаковые последовательности, используемые в качестве разделителей. Для ЭСУЯ нами выделены следующие типы этимологических классов: *класс реестрового слова* (обозначим *HEAD*), *класс дериватов* (*DER*), *класс славянских соответствий* (*SLAV*), *языковой класс* (*LANG*), *библиографический класс* (*BIBL*), *класс ссылок* (*LINK*). Отметим, что тип описания «языковой класс» используется только как наименование структурного элемента концептуальной модели словаря, а не как лингвистический термин. Каждый из этих классов имеет уникальную структуру текста, что дало нам возможность построить процедуру идентификации типа каждого этимологического класса в словарной статье по формальным признакам.

Дадим краткую характеристику каждого класса. *Класс реестрового слова* содержит собственно реестровое (заголовковое) слово и определённые его параметры. Заголовковым может быть слово как литературное, так и диалектное, а также имя собственное. Этот класс уникальный и обязательно входит в состав словарной статьи.

К *языковому классу* относим следующие описания: а) реконструируемые формы реестрового слова или их основы на разных этапах развития праславянского языка, представленные в антихронологическом порядке; б) этимологически связанные с реестровым словом слова других индоевропейских языков, начиная с ближайших к праславянскому фонетических и словообразовательных форм; в) этимологически связанные с реестровым словом слова семито-хамитских или урало-алтайских языков; г) этимологическая связь слова не установлена, например «этимология неясная». Анализ ЭСУЯ показал, что таких классов максимум два в словарной статье, но мы не ограничиваем их количество в нашей модели. В состав словарной статьи должен входить хотя бы один класс данного типа.

Класс реестрового слова и *языковый класс* составляют минимальную структуру словарной статьи. Этимологические классы других типов являются факультативными.

Класс дериватов содержит родственные с реестровым словом слова украинского языка, то есть ближайшие этимологически значения. В тексте словарной статьи может быть не более одного этимологического класса этого типа.

Класс славянских соответствий содержит соответствия реестрового слова из всех славянских языков, в которых они зафиксированы. В словарной статье может быть не более одной структурной единицы этого типа.

Библиографический класс — блок текста, содержащий информацию о научных трудах, в которых рассматривается этимология соответствующего украинского слова или связанных с ним слов других языков. Такой класс может быть только один.

К *классу ссылок* относим те текстовые блоки, где описываются связи с другими статьями словаря.

Проиллюстрируем сказанное на примере двух небольших, но достаточно репрезентативных с точки зрения структуры словарных статей. Тексты приводим в форме, максимально приближенной к печатной версии.

*Пример 1 (словарная статья с заголовковым словом **абетка**):*

абетка, [*абетло*] Пі, *абетний* (заст.) «елементарний»**;—** власне українська назва азбуки, утворена за вимовою перших двох букв алфавіту (*а, бе*), очевидно, під впливом назв *азбука*, *альфабет* і п. abecadło «тс.» (від вимови перших трьох букв a, be, ce)**.—** Sadn. — Aitz. VWb. I 42**.—** Пор. **азбука, алфавіт.**

*Пример 2 (словарная статья с заголовковым словом **абзац**):*

абзац;— р. бр. *абзац*, болг. *абзац*, схв. *абзац*;**—** запозичення з німецької мови**;** нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sezzen, дангл. settan,

англ. set і спорідненого з псл. saditi, укр. *садити*. — CIC 7; Фасмер I 56; Paul DWb 8, 10; Kluge — Mitzka 705. — Див. ще **абажур, садити.** — Пор. **обцас.**

Пример 3 (этимологические классы для словарной статьи ***абетка;*** *тексты классов подаются в угловых скобках):*

HEAD ≡ <**абетка**>
DER ≡ <[*абетло*] Пі, *абетний* (заст.) «елементарний»>
LANG ≡ <власне українська назва азбуки, утворена за вимовою перших двох букв алфавіту (*а, бе*), очевидно, під впливом назв *азбука*, *альфабет* і п. abecadło «тс.» (від вимови перших трьох букв a, be, ce)>
BIBL ≡ <Sadn. — Aitz. VWb. I 42>
LINK ≡ <Пор. **азбука, алфавіт**>

Пример 4 (этимологические классы для словарной статьи ***абзац****):*

HEAD ≡ <**абзац>**
SLAVIA ≡ <р. бр. *абзац*, болг. *абзац*, схв. *абзац*>
LANG ≡ <запозичення з німецької мови; нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sezzen, дангл. settan, англ. set і спорідненого з псл. saditi, укр. *садити*>
BIBL ≡ <CIC 7; Фасмер I 56; Paul DWb 8, 10; Kluge — Mitzka 705>
$LINK_1$ ≡ <Див. ще **абажур, садити>**
$LINK_2$ ≡ <Пор. **обцас>**

Приведём пример словарной статьи, минимальной как структурно, так и содержательно:

Пример 5 (этимологические классы для словарной статьи ***андріяк****):*

[андріяк] «опій»; — походження неясне.
HEAD ≡ <**[андріяк]** «опій»**>**
LANG ≡ <походження неясне>

В тексте каждого этимологического класса устанавливаются связи реестрового слова с определёнными словами других языков. Все эти слова, включая реестровые, мы будем называть *этимонами.* При анализе текстов этимологических классов было выявлено восемь параметров, посредством которых описываются этимоны: *маркер языковой принадлежности* (обозначим P_L), *ремарка к маркеру языковой прнадлежности* (P_{RL}), *знаковое представление этимона* (P_A), *принадлежность к диалектной лексике* (P_D), *маркер омонимии* (P_o), *толкование* (P_S), *ремарка* (P_R), *библиография* (P_B). Мы перечислили параметры в том порядке, в котором они, как правило, следуют в тексте соответствующего этимологического класса. Два параметра являются обязательными: это P_L (*маркер языковой принадлежности*) и P_A (*знаковое представление этимона*). Эти два параметра обеспечивают уникальность каждого этимона словарной статьи: этимоны с одинаковой знаковой формой могут иметь разную языковую принадлежность, или этимоны с одинаковой языковой принадлежностью могут иметь разные знаковые формы. Остальные параметры — факультативные. Для каждого параметра определена формальная процедура, которая позволяет вычленить соответствующий параметр из текста для каждого этимологического класса.

Набор параметров $\{P_L, P_{RL}, P_A, P_D, P_O, P_S, P_R, P_B\}$ мы будем называть *этимон-структурой* и будем обозначать символом $ETYM(e_i)$, где e_i — соответствующий этимон; индекс *i* — порядковый номер данного этимона в тексте. Порядок следования параметров в этимон-структуре полагаем не существенным.

Не все параметры актуальны для каждого этимологического класса. Текст, который мы идентифицируем как этимологический класс, использует свое подмножество параметров; не каждый этимон обязан описываться полным набором параметров. Однако для достижения структурной однородности для каждого класса строится один тип этимон-структуры; если определённый параметр не задействован или не может быть выделен по формальным признакам, то его значению соответствует пустая строка текста. Этимон-структура строится только в том случае, если удалось вычленить P_A. Формально мы полагаем, что каждому этимологическому классу соответствует этимон-структура. Если языковой класс не имеет ни одного этимона (или не удалось его выявить формальной процедурой), то мы считаем его вырожденным этимологическим классом и ему соответствует пустая этимон-структура. Примером такого класса служит языковой класс для словарной статьи из примера 5.

Проиллюстрируем этимон-структуры на примерах текстов этимологических классов:

Пример 6 (этимон-структуры для класса реестрового слова):

HEAD (**абзац**) ≡ <абзац**>**
ETYM (e_1) ≡ $\{P_L$ = <укр.>, P_A = <абзац>$\}$

Пример 7 (этимон-структуры для класса дериватов):

DER (**абетка**) ≡ <[*абетло*] Пі, *абетний* (заст.) «елементарний»>
ETYM (e_1) ≡ $\{P_L$ = <укр.>, P_A = <абетло>, $P_D = 1, P_B$ = <Пі>$\}$
ETYM (e_2) ≡ $\{P_L$ = <укр.>, P_A = <абетний>, P_R = <(заст.)>, P_S = <«елементарний»>$\}$

Параметр омонимии P_O для этимона e_1 принимает значение 1, так как квадратные скобки указывают на принадлежность слова к диалектной лексике. По умолчанию для всех этимонов значение этого параметра 0.

Пример 8 (этимон-структуры для класса славянских соответствий):

SLAV(**абзац**) ≡ <р. бр. *абзац*, болг. *абзац*, схв. *абзац*>

ETYM (e_1) ≡ {P_L= <р.>, P_A = <абзац>}

ETYM e_2) ≡ {P_L= <бр.>, P_A = <абзац>}

ETYM (e_3) ≡ {P_L= <болг.>, P_A = <абзац>}

ETYM (e_4) ≡ {P_L= <схв.>, P_A = <абзац>}

Пример 9 (этимон-структуры для языкового класса):

LANG (**абзац**) ≡ <запозичення з німецької мови; нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sezzen, дангл. settan, англ. set і спорідненого з псл. saditi, укр. *садити*>

ETYM (e_1) ≡ {P_L= <нім.>, P_A = <Absatz>, P_S = <«перерва, пауза, уступ, абзац»>}

ETYM (e_2)≡ {P_L= <нім.>, P_A = <absetzen>, P_S = <«відсувати, відставляти»>}

ETYM (e_3) ≡ {P_L= <нім.>, P_A = <ab->, P_S = <«від-, з-»>}

ETYM (e_4) ≡ {P_L= <гот.>, P_A = <af>, P_S = <«від»>}

ETYM (e_5)≡ {P_L= <лат.>, P_A = <ab>, P_S = <«тс.»>

ETYM (e_6)≡ {P_L= <нім.>, P_A = <setzen>, P_S = <«тс.»>}

ETYM (e_7)≡ {P_L= <двн.>, P_A = <sezzen>}

ETYM (e_8) ≡ {P_L= <дангл.>, P_A = <settan>}

ETYM (e_9) ≡ {P_L= <англ.>, P_A = <set>}

ETYM (e_{10}) ≡ {P_L= <псл.>, P_A = <saditi>}

ETYM (e_{11}) ≡ {P_L= <укр.>, P_A = < садити >}

Основная проблема создания компьютерных словарей, исходя из их печатных версий, — это формирование соответствующей базы данных в автоматическом режиме непосредственно из текста словаря (парсинг). Опыт убеждает, что формирование лексикографических баз данных «вручную» из больших и сложных словарных текстов практически невозможно. Основная задача парсинга — автоматическое выделение определенных нами структурных элементов непосредственно из текста словаря, поскольку именно они выполняют роль элементов лексикографической базы данных.

Перед конверсией тексты всех томов были переведены в формат HTML и унифицированы как относительно структуры файлов, так и относительно знаковой системы. Словарь был подготовлен к печати различными издательскими технологиями. Первые три тома — в технологии монотайп, докомпьютерной. Поэтому печатные тексты сначала были отсканированы, распознаны программой FINEREADER, а затем вычитаны. Тексты 4-го и 5-го томов уже готовились в издательской системе, т. е. в цифровом формате. Знаковая система всех текстов словаря была унифицирована согласна кодировке UNICODE 3.0. Это позволяет выполнить инвентаризацию символов алфавита для представления этимонов каждого языка.

В результате этих операций были получены специальным образом препарированные тексты томов Этимологического словаря, полностью готовые для автоматической конверсии в лексикографическую базу данных.

Для поддержки цифровой версии словаря построен инструментальный комлекс, который обеспечивает такие основные функции [4]:

1) автоматическую конверсию текстов этимологического словаря в компьютерную базу данных;
2) традиционный вход в систему по реестровому слову и отображение текста словарной статьи;
3) редактирование любого структурного элемента словарной статьи;
4) построение этимон-структуры для словарной статьи в ручном режиме;
5) автоматическое построение этимон-структуры для словарной статьи;
6) создание словарной статьи с определенной структурой.

На рис. 1 показано одно из окон редактирования словарной статьи. На левой панели словарная статья представлена в виде дерева структурных элементов. Для каждого этимологического класса выводится упорядоченный список этимонов, тем самым графическими средствами визуализируется глубина этимологического исследования. С помощью кнопок на средней панели структурные элементы можно добавлять, удалять и менять порядок их следования. Функции кнопок варьируются в зависимости от выбранного структурного элемента. Так, например, кнопка «Додати» (добавить) при выборе этимона позволяет добавить только этимон. Для каждого структурного элемента разработана своё окно редактирования, которое отражает специфику этого элемента. Для каждого этимона выводится и текст соответствующего этимологического класса, однако с запретом его редактирования. Это даёт возможность верифицировать параметризацию этимона, выполненную автоматически.

Для автоматического построения языковых индексов разработано специальный инструментарий, который позволяет:

1) в интерактивном режиме формировать любое количество языковых регистров на множестве всех языков словаря;
2) задавать спектры индексации, учитывая структуру словарной статьи.

На рис. 2 показано диалоговое окно пользователя для формирования языкового регистра.

Левая панель предназначена для выбора уже сформированных регистров в качестве неизменяемых шаблонов. Правая панель используется для редактирования существующих и формирования новых регистров.

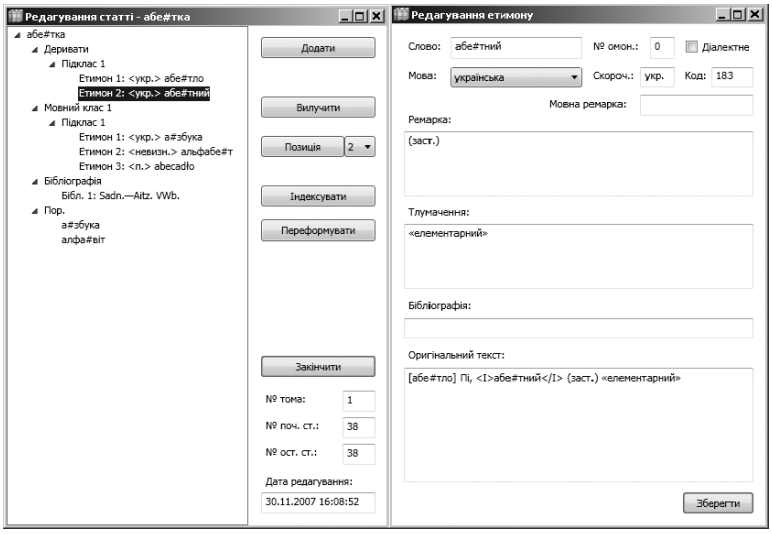


Рис. 1. Окно редактирования для словарной статьи с заголовковым словом ***абетка***

На рис. 3 показано окно главного пользовательского интерфейса словаря с индексом, построенным по сформированному регистру.

На левой панели из предложенного набора языков был выбран польский (возможен выбор как всех языков в регистре, так и определённого подмножества языков). На правой панели в реестровое окно выведен список всех этимонов, которые идентифицированы как слова польского языка.

В окно реестра могут быть также выведены заголовковые слова тех словарных статей, в которых зафиксированы этимологические связи с польским языком. Сформированный индекс по команде пользователя выводится в текстовый файл с указанием локализации каждого этимона.

Инструментальная система позволяет задать локализацию индексируемых элементов с точностью до структурного элемента — типа этимологического класса — словарной статьи (с помощью

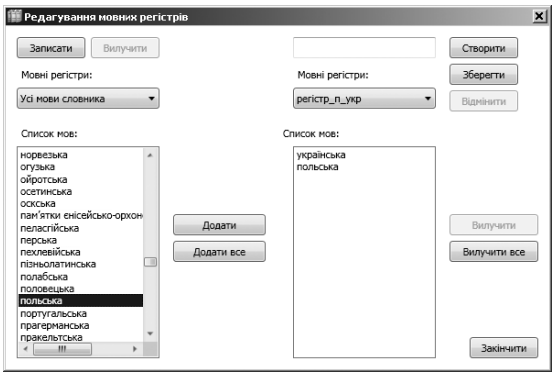


Рис. 2. Окно формирования языкового регистра

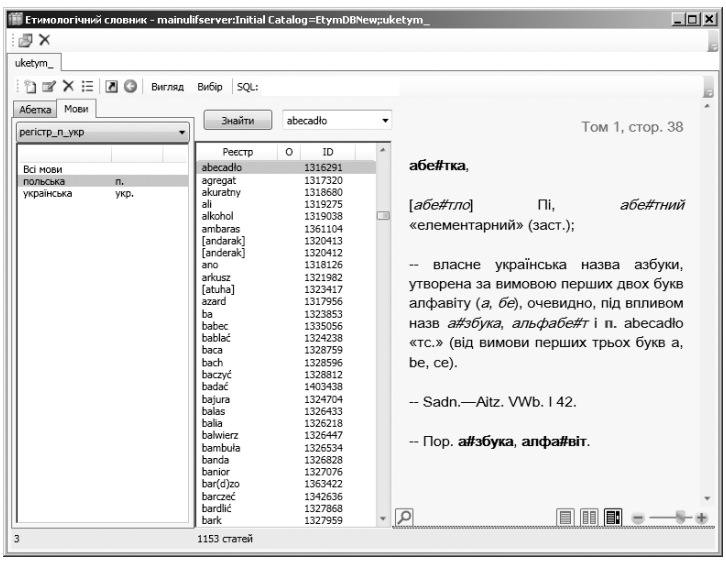


Рис. 3. Языковой индекс по заданному регистру

меню «Вибір» (Выбор) на верхней панели). В нашем случае был задан только *языковой класс*.

При активизации любого элемента реестра визуализируется текст словарной статьи.

Текст словарной статьи для вывода формируется из соответствующих полей базы данных. Полиграфическое оформление статьи практически сохранено полностью.

Описанный метод представления этимологического словаря в цифровой среде дал возможность построить для совокупности словарных статей словаря соответствующую совокупность этимон-структур как формальных репрезентантов описаний генетических связей реестровых единиц. Все схемы индексации строятся только на основе этимон-структур. Такой подход обеспечивает как возможность построения структур, имплицированных в текст словарных статей, так и отображения на цифровую среду аутентичного текста словаря, что делает цифровой словарь открытым для дальнейших интерпретаций.

Разработанные технологии и интерфейсы предлагаются как базовые для цифровых репрезентаций этимологических работ.

Литература

1. *Етимологічний* словник української мови: В 7 т. Київ: Наукова думка, 1982–2006. Т. 1–5.
2. *ANSI/X3/SPARK DBMS study group* interim report. FDT-Bull. ACM SIGMOD. 1975. V. 7. № 2.
3. *Широков В. А.* Елементи лексикографії. Київ: Довіра, 2005.
4. *Остапова И. В., Якименко К. Н.* Инструментальная лексикографическая система Этимологического словаря украинского языка // Прикладна лінгвістика та лінгвістичні технології. Київ: Довіра, 2008. С. 276–291.

Посессивы и имена способа действия[1]

Possessives and mode of action nouns in russian

Падучева Е. В. (elena708@gmail.com)
ВИНИТИ РАН

Концептуальный аппарат модели «Смысл–Текст» предлагает тщательно разработанную классификацию предикатных имен. Различаются имена действия, субъекта, объекта и второго объекта действия; имена инструмента, места и других обстоятельств. В работе речь идет об именах способа. Показано, что имена способа, мотивированные переходным глаголом, имеют специальную диатезу, в которой субъект выражен не твор. падежом, а посессивом.

Данная работа продолжает сопоставление (начатое в Падучева 1984) семантики посессивов со значениями генитива, который тоже имеет — в позиции при имени — прежде всего посессивное (иначе — притяжательное) значение.

Посессивы в русском языке можно разделить на две группы.

1) Притяжательные местоимения: личные *мой, твой, ваш, наш* и, с оговорками, местоимения 3 лица (*его, ее, их*); возвратное *свой*; вопросительно-относительное *чей*; неопределенные *чей-то, чей-нибудь, чей-либо, кое-чей*; отрицательное *ничей*.

2) Притяжательные прилагательные на *-ов, -ин, -овский*. (Как справедливо отмечается в Шмелев 2008, прилагательные на *-ий — лиса-лисий, баба-бабий* — обладают иными свойствами и должны изучаться отдельно; они не обладают основным значением посессивов — значением посессивности: *лисий хвост*, но не **лисий сыр*; *обезьянья ловкость*, но не **обезьяний банан*.)

Обе группы посессивов обстоятельно изучены в статье Шмелев 2008, которая сосредоточена на посессивах в контексте предметных имен. Меня же будут интересовать посессивы в контексте предикатных имен, т. е. отглагольных (как *интерпретация* от *интерпретировать*) и глагольных, как *концепция* от предполагаемого глагола со значением ‘создавать концепцию’. Будет показано, что посессивная диатеза позволяет выделить класс предметных имен — условно названных именами способа действия — обладающих нетривиальными общими свойствами.

1. Притяжательное местоимение

Соответствующие притяжательные имеются у всех субстантивных местоимений — за одним показательным исключением: отсутствует притяжательный коррелят у субстантивного *это* и соответствующего ему относительного *что* (как в контексте *… но робость поступить несовременно чуть-чуть мешала ей,* что *было зря*, Б. Ахмадулина).

Это отсутствие естественно вытекает из семантики притяжательности: в своем основном значении притяжательное местоимение выражает принадлежность лицу — т. е. обозначает лицо как посессора, а *это* не может обозначать лица.

Можно различить в значении притяжательного местоимения два компонента: а) денотативный компонент — референция к лицу; б) синтаксический компонент — принадлежность (в широком смысле) этому лицу.

Что касается значения принадлежности (принадлежность, притяжательность, посессивность — это все синонимы), то оно, как известно, возникает лишь в прототипическом для притяжательного местоимения контексте — в позиции при имени предмета (*моя кружка, ее расческа*). В контексте предикатного имени притяжательное местоимение не выражает принадлежности — и не всегда сохраняет свое свойство референции к лицу. Особенно это касается местоимений 3 лица: *ее отсутствие* может быть ‘отсутствие воды’, ‘отсутствие свободы’ и т. д. — хотя *в ее отсутствие* как *в мое*.

[1] Работа финансировалась грантом РГНФ, проект № 08-04-00181а.

Совмещение собственно притяжательных и субъектно-объектных функций у притяжательных местоимений широко распространено в самых разных языках, что представляет интерес с типологической точки зрения.

Остается еще сказать о референциальном статусе: у притяжательных местоимений имеется референциальная общность — они конкретно-референтны. Личное местоимение может относиться к имени родового статуса (*в нем* [сонете] *жар любви Петрарка изливал*), а притяжательное как правило не может.

Указанная референциальная специфика притяжательного местоимения проявляется в том, что во многих конструкциях генитив имени нельзя заменить на притяжательное местоимение, которое могло бы иметь соответствующий терм своим антецедентом[2]:

1) в конструкции с генитивом в значении сравнения: *золото волос* — °*их золото*;

2) генитив с характеризующим значением: *день радости* — °*ее день*;

3) в конструкции с определением в составе генитивной группы: *дома современной архитектуры* — °*ее дома*;

4) в конструкции со значением «свойственности» (термин из Земская 1992: 76): *взгляд Наполеона* = ‘взгляд как у Наполеона’; *ловкость обезьяны* = ‘ловкость как у обезьяны, с в о й с т в е н н а я обезьяне’; сочетания *его взгляд*, *ее ловкость* не допускают такого понимания

Если именная группа в такой конструкции конкретно-референтна, значение разрушается; сочетание *ловкость этой обезьяны* выражает просто отнесение свойства к его носителю.

Как легко видеть, все эти конструкции невозможны для притяжательного местоимения уже потому, что оно может иметь только конкретно-референтный статус (и антецедент).

2. Притяжательные прилагательные

Теперь о притяжательных прилагательных на *-ов (-ев), -ин (нин)* и *-овский*. В Грамматике 1960: 299 про прилагательные на *-ов* и *-ин* сказано, что они выражают «значение принадлежности одному лицу и не могут обозначать принадлежности группе лиц». Точнее будет сказать, что в составе прилагательных на *-ов* и *-ин* мотивирующее существительное, если оно имело референцию к конкретному лицу, сохраняет эту референцию: притяжательное прилагательное *-ов* и *-ин* — это как бы притяжательная форма конкретно-референтного имени. (Что же касается прилагательных на *-ий* (*бабий*), то они мотивированы существительным в родовом статусе.)

Родовой статус для слов на *-ов* и *-ин* невозможен, ср. пример из Шмелев 2008: сочетание *портрет девушки* будет переведено на английский язык как *a portrait of a girl*, а *девушкин портрет* — только как *a portrait of the girl*.

Имеется весьма широкий и постоянно пополняемый класс притяжательных прилагательных на *-овский*: *вендлеровская классификация, рихтеровское исполнение* и т. д. В отличие от модели на *-ов*, которая непродуктивна (в Земская 1992: 76 отмечен только один пример новообразования — *Гулливеров*), прилагательные на *-овский* с притяжательным значением образуются от иностранных фамилий, попадающих во 2-е склонение, практически без ограничений — так же, как прилагательные на *-ин* от имен 1-го склонения (*Ванин*, *Шарлоттин*). Прилагательные на *-овский*, как и притяжательные местоимения, выполняют собственно притяжательную функцию и функцию субъекта. Так что мы вправе говорить об общей посессивной семантике у местоимений и прилагательных. NB. В сочетаниях *апостольское послание, читательское мировоззрение, богословское толкование*, *обывательское восприятие действительности*, в которых основа имеет родовой, а не конкретно-референтный статус, суффикс *-ский*, а не *-овский*.

Если имя, мотивирующее посессив, конкретно-референтное, возможна синонимическая замена посессивной конструкции на генитивную [3]:

(1) Иногда *свиридовские взгляды на грядущее* (да и на современное ему настоящее тоже) просто донельзя черны и мрачны. [*Станислав Золотцев. Духовный подвиг исполина // «Наш современник», 2004.06.15*] = *взгляды Свиридова на грядущее*.

Впрочем, притяжательное прилагательное может быть употреблено «в значении свойственности»:

(2) В «Чайке» никто не стеснялся современных манер, отчаянных бросков преодолевая известную *чеховскую недосказанность характеров*, сложную разветвленную речь и объективность *чеховского взгляда на человека*. [*Анатолий Смелянский. Концерт для скрипки с оркестром (1990–2000)*]; здесь *чеховский* = ‘свойственный Чехову’.

[2] Знак ° выражает тот факт, что сочетание хотя и возможно, но не имеет ожидаемого значения.

[3] Все примеры с указанием источника — из Национального корпуса русского языка, адрес в Интернете — www.ruscorpora.ru .

3. Имена способа действия и их посессивно-генитивная диатеза

Имеется контекст, в котором общая семантика посессивов дала о себе знать. Речь идет о конструкции, в контексте которой притяжательное местоимение и прилагательное, включая прилагательное на *-овский*, полностью вытесняют генитив как средство выражения субъектного значения. Рассмотрим пример (на базе Падучева 1984, 1974: 203):

(3) а. *твое* исполнение Шопена;
б. *Рихтеровское* исполнение Шопена;
в. *исполнение *Рихтера* Шопена.

Как мы видим, в контексте примера (3) субъект предикатного имени может быть выражен притяжательным местоимением, см. (3а), и притяжательным прилагательным, см. (3б), но не родительным падежом имени — при том, что в принципе генитив имени используется для выражения субъекта достаточно широко (*после приезда брата, в исполнении Рихтера* и т.д.) и соподчинение генитивов отнюдь не невозможно (*лишение брата наследства*).

Задача работы — описать семантику притяжательной конструкции, представленной примером (3б), и уточнить класс предикатных имен, в контексте которых возникают эффекты такого рода, как в примере (3).

Как известно, для участника субъекта при предикатном имени в русском языке имеется специальный способ выражения — творительный падеж:

(4) разрушение Новгорода московским *князем*.

Генитивом чаще оформляется объект. Однако существуют ситуации, когда генитивом может, а иногда и должен, оформляться субъект. Например, субъект всегда выражается генитивом, если имя образовано от непереходного глагола:

(5) приезд *брата*;

(6) возвращение *Ивана* на родину.

Если же имя мотивировано переходным глаголом, то оформление его субъекта генитивом возможно лишь при определенных условиях. Так, возможен генитив субъекта у имен, образованных от глаголов *любить, уважать*, которые, будучи переходными, не лицензируют, однако, генитива объекта у производных имен (*любить апельсины* дает *любовь к апельсинам*, а не *любовь апельсинов*).

Два других условия гентивного оформления субъекта формулируются с обращением к понятию диатезы.

1. Если предикатное имя относится к типу имен объекта[4] (т.е. таких слов, как *совет, предложение, рассказ, прегрешение, намерение, воспоминание, замысел, жертва*), то генитив при нем однозначно интерпретируется как имя субъекта:

(7) предложение *Каспарова*.

Это естественно, поскольку устранение участника объекта входит у имен объекта в структуру их словообразовательной модели: валентность на объект заполняется, так сказать, самим именем.

2. Если у имени редуцированная диатеза (Падучева 1977) — безобъектная, то тоже генитив однозначно интерпретируется как субъектный — объект выражен в контексте (в примере ниже подразумеваемый объект подчеркнут):

(8) Машина не прошла проверки *экспертов*;
Он поддался на обман *злопыхателей*;
Это обеспечило ему поддержку *коллег*.

Если же при данном имени есть генитив, который выражает объект, то субъект принципиально не может быть выражен тоже генитивом — он должен быть выражен творительным падежом, как в примере (4).

Теперь о притяжательном местоимении. Оно отличается от субъектного генитива ровно в одном пункте — оно не исключено (в субъектном значении) в контексте имени, при котором генитив выражает объект[5], см. невозможное (3в) и безупречное (3а).

При этом, что замечательно, все посессивы ведут себя единообразно: не только притяжательное местоимение допустимо в контексте генитива субъекта, но и притяжательное прилагательное:

(9) геделево *доказательство* теоремы о полноте;
папино *истолкование* моей просьбы;
шаляпинское *исполнение* «Блохи».

Другое дело, что конструкция «субъектный посессив + предикатное имя + генитив объекта» допустима не при всех предикатных именах. Так, (10а), (11а) неправильно — нужно сказать (10б), (11б):

(10) а. *мое соблюдение тайны;
б. соблюдение мною тайны;

(11) а. *Все зависит от его соблюдения тайны;
б. Все зависит от соблюдения им тайны.

---

[4] Мы опираемся здесь на концепцию семантики предикатных имен как она представлена в работах Мельчук 1974, Апресян 1974.

[5] Пример, выявляющий это различие между генитивом и притяжательным местоимением был приведен в Иорданская 1967: 23.

Класс имен, располагающих диатезой с посессивным субъектом, представляет интерес — в контексте имен этого класса указанная конструкция не просто допустима: посессив является тут единственной возможностью выразить субъект.

В Падучева 1984 имена, располагающие диатезой с посессивом субъекта и генитивом объекта, были отнесены к именам способа. Имена способа предусмотрены общей классификацией отпредикатных имен из Мельчук 1974: 87, Апресян 1974: 48, 199. Но в числе примеров упоминаются преимущественно имена, образованные от непереходных глаголов (*походка, поведение*), и/или имена с родовым объектом, обозначающие узуальное действие или даже свойство на базе узуальных действий (*почерк, произношение*). Для них притяжательно-генитивная диатеза невозможна.

В то же время, в Апресян 1974: 199, где речь идет о регулярной многозначности имен действия и имен способа, в числе имен способа упомянуты слова *перевод, редакция*, и отмечается, что «существительные со значением способа иногда синкретически выражают и значение результата». А в таком случае есть основания отнести к именам способа слова, которые можно было бы назвать также и именами результата: четкая граница здесь отсутствует.

Имена способа, определенные таким образом, образуют весьма широкий класс. Особый интерес представляют имена способа, мотивированные переходными глаголами (такие как *трактовка*, *исполнение, понимание, осмысление, изображение, употребление, постановка, толкование*), поскольку это имена, обладающие тем замечательным свойством, что у них актантная структура включает субъект и объект, хотя они не являются именами ситуации.

Имена способа образуются от глаголов с валентностями «Кто?», «Кого/Что?» и «Как? / Каким способом/образом?» — последняя и отличает имена способа от имен ситуации.

Например:

X трактует Y так-то ⇒ X-ова трактовка Y-а <такова>.

Общее правило состоит в том, что имя *i*-го актанта /сирконстанта глагола V замещает свою *i*-ю валентность самим собой.

Парадоксальное место в актантной структуре глагола занимает участник Результат (Падучева 1999). У глаголов создания он выражается объектом (*построить дом, создать теорию*) и, в соответствии со словообразовательной моделью имени объекта, из аргументной структуры имени объекта выпадает; так, в сочетании *Петра творенье* слово *творенье* — имя объекта от глагола создания, и генитив при нем однозначно понимается как субъектный. Между тем во всех других классах переходных глаголов позицию прямого объекта занимает Пациенс (например: *проколол подошву*) или какой-то другой участник (например, Цель, Goal, как в *решил задачу*), а участник Результат — *прокол, решение* — не отражен в актантной структуре глагола. Поэтому при образовании имени результата от такого глагола генитив выражает Пациенса или Цель, а участник Результат хоть и выбывает из аргументной структуры имени, но это никак не сказывается на синтаксических потенциях слова.

Отсюда и происходит тот замечательный факт, что имена способа (равно как и имена результата не от глаголов создания), не будучи именами ситуации, подобно именам ситуации, способны сохранять обе основных синтаксических валентности мотивирующего переходного глагола — и субъектную, и объектную.

Итак, имена способа в этом широком смысле обладают следующими свойствами. Их актантная структура наследует от глагола обоих главных участников, субъект и объект, поскольку образуются по словообразовательной модели, которая не поглощает ни участника субъекта, ни объекта. Что объединяет их с именами ситуации. Но валентность субъекта у имен способа не может быть замещена творительным падежом, как у имен ситуаций, а только притяжательным местоимением или притяжательным прилагательным (типа *мамин, андреев*), в том числе — прилагательным на *-овский* (*отцовский, расселовский*). Примеры (из Падучева 1984); в (12) притяжательное местоимение, в (13) — притяжательное прилагательное:

(12) Его *выбор* секунданта был неудачен; Чье *исполнение* Шопена Вам понравилось больше?; А какое Ваше *объяснение* причин этой ссоры? Его *употребление* метафор вызывает недоумение;

(13) *расселовская* трактовка дескрипций; *геделево* доказательство теоремы о полноте; Вот как обстоит дело в *мамином* понимании.

Если валентность субъекта заполняется твор. падежом, существительное перестает пониматься как имя способа:

(14) а. Его *понимание* этой проблемы не совпадает с моим [*понимание* — имя способа];
б. *Понимание* им этой проблемы свидетельствует о его искушенности в таких делах [*понимание* — имя факта].

(15) а. Его *исполнение* Шопена было великолепно [*понимание* — имя способа];
б. *Исполнение* им Шопена было неуместно [*понимание* — имя действия].[6]

[6] В Падучева 1984 слово *исполнение* в примере (13б) было ошибочно названо именем факта (что отмечено в Schoorlemmer 1998): неуместным является, конечно, не факт, а действие.

Среди имен способа есть имена актанта и имена сирконстанта. Большая часть имен способа — это имена актанта Результат. Но у слов *почерк*, *походка*, в семантику которых входит идея Как?, Какой?, этот участник наследуется от сирконстанта мотивирующего глагола. Именем сирконстанта является имя способа *удар* в контексте *У него сильный удар*. Особо отметим слово *поведение*, у которого способ — Какое? — является актантом.

Имена сирконстантов обычно не наследуют участника Субъект; ср. *выход, выезд* как имена места. Но это потому, что они прошли через «опредмечивание» как отдельный этап семантической деривации.

Сколь важно различать имена ситуации и имена способа, показывает следующий пример [7]:

(16) *Описание Томсоном условий, определяющих выбор падежа, остается одним из лучших.

Сама по себе ИГ *описание Томсоном условий, определяющих выбор падежа* правильно построенная и не вызывает никаких возражений. Но в данном контексте она оказывается неуместной — неуместной потому, что *описание* должно быть оформлено (в контексте характеризующего предиката — *остается одним из лучших*), как имя способа, а не как имя события или факта. В контексте имени способа следует употребить не твор.падеж, а посессив:

(17) *Томсоновское* описание условий, определяющих выбор падежа, остается одним из лучших.

Сочетание *перевод сонетов Шекспира Маршаком* вполне законно. Нельзя, однако, сказать *Перевод сонетов Шекспира Маршаком* не знает себе равных. Единственно возможной является посессивно-генитивная диатеза — *Маршаковский перевод сонетов Шекспира*, ср. в этой связи Апресян 1974: 199.

Верно и обратное — если имя не может быть понято как имя способа, то посессив неуместен для оформлении его субъектного актанта. Так, в (16) нельзя заменить твор. падеж на притяжательное местоимение, поскольку имя *понимание*, в данном контексте, не является именем способа:

(18) Необходимое условие перевода — понимание *ею* <машиной> текста в полном объеме.

Почему же имена типа *исполнение* так естественно интерпретируются в значении способа — характера, манеры — исполнения? Дело в том, что упоминание субъекта переносит акцент на способ, которым данный субъект осуществляет это действие и, тем самым, на результат. Неудивительно, что имена способа тяготеют к контексту характеризующего предиката, см. (17), а также (19):

(19) *Его изображение Наполеона* слишком человечное.

Для имен способа характерна также диатеза с подъемом генитивного объекта — так сказать, атрибутивная:

(20) а. его изображение *Наполеона*;
б. *Наполеон* в его изображении.

Хотя имена способа мотивированы глаголом действия, их субъект, выраженный посессивом, не совсем Агенс: указание субъекта действия сообщает результату какой-то признак, характеристику, которую слушающий может вычислить на базе знакомства с субъектом. Иначе говоря, указание субъекта намекает на какую-то характеристику результата.

Заметим, что притяжательное прилагательное на *-овский* и на *-ин* не в точности равны: сочетание *папина фотография* может быть понято в двух смыслах (*папа* субъект и *папа* объект), а *веласкесовский портрет* — только в одном: *Веласкес* — субъект.

Ниже следуют примеры, иллюстрирующие посессивную диатезу имен способа. Нас будут интересовать не все имена способа, а только те, которые мотивированы переходными глаголами и, следовательно, имеют диатезу <Субъект-Посессив, Объект-Генитив>.

4. Примеры

Примеры ниже показывают, как посессив (притяжательное местоимение или прилагательное) выражает субъект имени способа в контексте генитива объекта при том же слове.

АНАЛИЗ
Он отмечает, что *сахаровский анализ опасности* мирового ядерного самоубийства в точности совпадает с документами, опубликованными американскими учеными. [*Г. Горелик., А. Сахаров. Наука и свобода (2004)*]

ВИДЕНИЕ
… перенос в современность действия романа Оруэлла лишь банализирует трагедию *оруэлловского видения тоталитаризма* XX века… [*М.Стуруа. Взорвать Оруэлла. Пародия на «Ферму животных» вызвала литературный скандал // «Известия», 2003.01.10*]

[7] Реальная фраза из работы нерусскоязычного слависта, прекрасно владеющего русским языком.

ВОСПРИЯТИЕ

Эти отношения помогают выявить некоторые чрезвычайно существенные аспекты *шолоховского восприятия* жизни. [*Ф. Раскольников. Статьи о русской литературе* (1986–2000)]

У слова *восприятие* имеется — необъяснимая — диатеза с твор. субъекта, как если бы *восприятие* было именем действия, а не способа:

Впрочем, убежден руководитель фонда «Экспертиза» Марк Урнов, в *восприятии власти населением* очень силен эмоциональный фактор. [*А. Корня. Портрет власти: отличник на фоне двоечников // «Время МН», 2003.11.24*]

Откровенно признаюсь, крепкие сомнения у меня возникли по поводу адекватности *восприятия действительности соседушкой*... [*В. Степанычев. Я памятник себе воздвиг? // «Богатей» (Саратов), 2003.10.23*]

Это ставит под сомнение отнесение слова *восприятие* к именам способа.

ВЫБОР

Слово выбор имеет косвенную диатезу в примерах (1), (2) и прямую в (3), где *выбор* это ‘предпочтение’ (Падучева 2004):

(1) Видимо, контакт с авиапромышленностью и повальная увлеченность авиацией 30 годов сказались на *отцовском выборе* будущей *профессии* сына. [*Валерий Родиков. Михаил Вальденберг: «МИГи» жизни // «Российское оружие: война и мир», 1997.01.28*]

(2) работает уже 23 года, и ни разу не пожалела о *своем выборе профессии*! [*Алла Ягодкина. Отворите скорей: Почтальон у дверей! // «Приазовский край», 2004.10.07*]

(3) Если человек лишен возможности выбора зла, то **его выбор добра** полностью девальвируется. [Виталий Куренной. Этика вида и генная инженерия // «Отечественные записки», 2003].

ИЗОБРАЖЕНИЕ

(1) Например, к *его изображению ревнивца*, в лице Отелло, нельзя прибавить ни одной черты: так оно полно. [*В. Н. Майков. Статьи из вып. 1 «Карманного словаря иностранных слов» (1845)*]

ИЗЛОЖЕНИЕ

(1) Русскому же читателю, одолевшему *роллановское изложение* этого циркового *вранья*, напомню лишь, [*Б. Носик. «Кто ты? — Майя» // «Звезда», № 4, 2001*]

ИНТЕРПРЕТАЦИЯ

(3) И здесь снова приходит на память «Анджело» и *его лотмановская интерпретация*. [*Г. Кружков. Поэт и эхо // «Дружба народов», №6, 1999.06.15*]

Чаще всего притяжательные на *-овский* мотивированы фамилиями иностранного звучания. Однако в тот же семантический класс входит, например, *отцовский*:

(5) Бесы, ведьмы и привороты — это уже сугубо *отцовская интерпретация*. [*А. Ткачева. Приворот* (1996)]

ИСПОЛНЕНИЕ

(2) прослушивалась магнитофонная запись «Петушков» в *Веничкином исполнении*, [*Н. Шмелькова. Последние дни Венедикта Ерофеева (2002)*] [безобъектная диатеза]

В (4) генитив объекта опущен, но должен быть ясен из предтекста:

(4) Но, пожалуй, именно *Рунино исполнение* до сих пор представляется мне наиболее точным. [*А. Городницкий. «И жить еще надежде» (2001)*]

ИСТОЛКОВАНИЕ

чувство юмора ей неведомо — что она доказала *своим истолкованием слова* «трубадур». [*М.Чулаки. Примус // «Звезда», № 1–2, 2002*]

КОНЦЕПЦИЯ

То, что с ним там происходит, вполне укладывается в *бессоновскую концепцию* французского *кино*. [*Французские поцелуи // «Известия», 2002.01.17*]

приближался к *толстовской концепции искусства*. [*Год Сергея Прокофьева // «Российская музыкальная газета»*, 2003.02.12]

Но вернемся к *немцовской концепции правительства* — уборщицы. [*Андрей Пионтковский. Вотум недоверия: убить мутанта // «ПОЛИТКОМ.РУ», 2003.06.18]* [прямая диатеза] *платоновская концепция любви* как восхождения по лестнице прекрасного [*Д.Голынко-Вольфсон. Культура интим-сервиса и кризис иронии // «Искусство кино», 2003.06.30*]

В *Дарвиновской концепции* эволюция состоит в том, что имеет место механизм приспособления. [*Лекция А. Раппапорта «Антиномии не чистого воображения»*]

ОЦЕНКА

А нас всех в классе поразила *его оценка* моих скромных *успехов*. [*И. А. Архипова. Музыка*

жизни (1996)]

ПЕРЕВОД

В *пастернаковском переводе Шекспира*, как и в тексте Грибоедова, постоянно происходит перемол вторгающихся в него кусков прозы. [*Наум Берковский. Шекспир у Эфроса. К гастролям московского театра на Малой Бронной (1990–2000)*]

Лютеровский перевод «Нового Завета» (1522) маркирует начало принципиально нового отношения к Писанию. [*Виталий Куренной. Медиа: средства в поисках целей // «Отечественные записки», 2003*]

ПЛАН

Чубайсовский план реструктуризации РАО «ЕЭС России» прокремлевские фракции поручили проводить не «команде» самого Чубайса, а правительству М.Касьянова. [*Александр Нагорный. Все на выборы, всё на продажу // «Завтра», 2003.02.25*]

В Париже царь заказал кондуктору Дюменю копию *леблоновского плана Петербурга*, заплатив за нее 100 ливров [Т. А. Базарова. План петровского Петербурга. Источниковедческое исследование (2003)]

Устиновские планы укрепления передовой феноменальны по масштабам, по разнообразию применяемых средств и детальности проработки всего этого разнообразия. [Виктор Некрасов. В окопах Сталинграда (1946)]

ПОДБОР

Я ловлю себя сейчас на том, что *мой подбор свидетельств* и аргументов слишком тенденциозный, односторонний, умышленный. [*В. Соловьев. Три еврея, или Утешение в слезах. Роман с эпиграфами (1975–1998)*]

ПОНИМАНИЕ

Нравственные основы были в меня заложены в семье, так что *мое понимание* ***греха*** — не церковное, а житейское. [*С. Тхоржевский. Поздние записи // «Звезда», № 5, 2002*] я … стараюсь дать себе отчет, что нового внесла эта фраза в *мое понимание текста* и как перестроила старое. [*М.Л. Гаспаров. «Снова тучи надо мною…" Методика анализа (1997)*]

Если это *мое* широкое *понимание слова* «революция» показалось 0г. Астафьеву неправильным, то он мог бы прямо на это возразить… [К.Н. Леонтьев. Культурный идеал и племенная политика (1890) [прямая диатеза]

Конрад считали оппозиционную культуру, по своему смыслу близкую к культуре критического *дискурса в гоулднеровском понимании*. [Б. М. Фирсов. Интеллигенция и интеллектуалы в конце XX века // «Звезда», № 8, 2001] *юнговское понимание* душевной *жизни* человека «как внутренней драмы со множеством персонажей», [С. Г. Бочаров. Вокруг «Носа» (1988)]

ПОСТАНОВКА

Мандельштам, очевидно, уловил карнавально сторону революционного «действа», начавшегося с *мейерхольдовской постановки «Маскарада»* [*М. Л. Гаспаров, Омри Ронен. Похороны солнца в Петербурге // «Звезда», № 5, 2003*]

Чисто *мандельштамовская постановка проблемы* поведения. [*Э. Герштейн. Вблизи поэта (1985–1999)*]

ПРОЧТЕНИЕ

Столь же сомнительно *его прочтение приватизации*. [*И. Федюкин. Новые американские книги о постсоветской России // «Неприкосновенный запас», 2003.01.15*]

С *«набоковским» прочтением Гоголя* можно не соглашаться. [*В. Сердюченко. Между выморочностью и гениальностью // «Лебедь» (Бостон), 2003.06.23*]

РЕШЕНИЕ

Свое решение проблемы предлагают несколько американских компаний, которые выпускают камуфляжные контейнеры для базовых станций. [*Маскировка для антенны //* «Computerworld», № 29, 2004] найти единственный ключ к пониманию рассказа или даже найти *чеховское решение вопросов*, поставленных в нем» [Феликс Раскольников. Статьи о русской литературе (1986–2000)]

СПОСОБ

это — лишь *лужковский способ оградить* россиян от распределения общенационального дохода. [*А. Лебедев. Слезам Москвы не верят. Теракт на Дубровке не уменьшил ненависти провинции к столице // «Известия», 2002.11.03*]

ТРАКТОВКА

Коммунистические бюрократы возмущались *мейерхольдовской трактовкой «Ревизора»* — незамаскированной сатирой на них самих. [*Ю. Анненков. Дневник моих встреч*

(1966)]

Сравнение примеров (а) и (б) показывает, что и в (а) речь идет скорее не о трактовках Чехова, а о трактовках в его стиле:

(а) И в этом году жанровая палитра *чеховских трактовок* простиралась от нежной лирики до оголтелой буффонады. [*Два сюжета // «Театральная жизнь», 2003.08.25*]

(б) Получилось забавно: после разгула «**контрчеховских**» **трактовок** вновь вернуться в милое лоно традиционных смыслов. [*Два сюжета // «Театральная жизнь», 2003.08.25*]

ТОЛКОВАНИЕ

Интересно, … что, они думают о *моем толковании Апокалипсиса*? [*Л. Сергиевский. Письмо в дискуссии (2000)*]

ФОРМУЛИРОВКА

Ибо этой «свободы духа» никогда и не могло быть у нас именно из-за первых двух положений *аксаковской формулировки «русского пути»*. [*Г. Киреев. От национальной утопии к национальной идее // «Лебедь» (Бостон), 2003.06.16*]

Заключение

Итак, мы обнаружили класс предикатных имен, которые обладают определенной семантической общностью и специфической — посессивной — диатезой: объект — генитив, субъект — посессив (т.е. притяжательное местоимение, притяжательное прилагательное на *-ов* и *-ин* или прилагательное на *-овский*). Модель, порождающая притяжательные прилагательные на *-овский*, обладает, в определенных пределах, абсолютной продуктивностью, и новообразования этого рода весьма широко используются в современном языке — в том числе, для выражения субъекта предикатного имени.

Мы не настаиваем на том, что термин «имена способа» является вполне удовлетворительным для увиденного класса слов. Можно надеяться, что лучший термин не заставит себя ждать.

Литература

1. *Апресян Ю. Д.* Лексическая семантика: Синонимические средства языка. М.: Наука, 1974.
2. *Виноградов В. В.* Русский язык: Грамматическое учение о слове. М.; Л.: Учпедгиз, 1947.
3. *Грамматика 1960* — Грамматика современного русского литературного языка: т. I. Фонетика и морфология. М.: Наука, 1960.
4. *Грамматика 1980* — Русская грамматика. Т. 1–2. Отв. ред. Н. Ю. Шведова. М.: Наука, 1980.
5. *Земская Е. А.* Словообразование как деятельность. М.: Наука, 1992.
6. *Иорданская Л. Н.* Автоматический синтаксический анализ. Наука, Сибирское отделение. Новосибирск 1967.
7. *Падучева Е. В.* О семантике синтаксиса. М.: Наука, 1974.
8. *Падучева Е. В.* О производных диатезах отпредикатных имен в русском языке. Проблемы лингвистической типологии и структуры языка, Л.: Наука, 1977.
9. *Падучева Е. В.* Притяжательное местоимение и проблема залога отглагольного имени. // Проблемы структурной лингвистики, М.: Наука, 1984, 50–66.
10. *Падучева Е. В.* Аспектуальные свойства глаголов с семантическим актантом Результат. Вопросы филологии, 1999, N 3, 19–26.
11. *Пешковский А. М.* Русский синтаксис в научном освещении. 6-е изд. М., 1938.
12. *Шмелев А. Д.* Посессивы в современной русской грамматике // Динамические модели. Слово, предложение, текст. М.: ЯСК, 2008, 927–942.
13. *Grimshaw 1990* — Grimshaw J. Argument Structure. L. etc.: MIT Press, 1990.
14. *Schoorlemmer M.* Complex event nominals in Russian: Properties and readings. Journal of Slavic linguistics 6(2), 205–254.

Модели деривации и синтаксическая позиция отглагольных существительных по корпусным данным

Derivational patterns and syntactic positions of deverbal nominals (on corpus data)

Пазельская А. Г. (anna_pz@abbyy.com)

ABBYY Software

Данная статья продолжает исследование различий в поведении русских имён ситуаций, образованных по разным деривационным моделям, на корпусном материале, почерпнутом по большей части из Национального корпуса русского языка. В этой работе исследуются различия, касающиеся синтаксической позиции отглагольного имени в предложении.

1. Введение

Имена действия в русском языке образуются от глаголов при помощи довольно большого количества суффиксов. Так, в [Шведова и др. 1982: 157–166] приводятся следующие достаточно продуктивные суффиксы: ***-нић*** (с вариантами ***–тић***, ***-енић***, ***-анић***: *чтение*, *отплытие*); ***-ић*** (*доверие*); ***-к(а)*** (*плавка*); ***-ција*** (*стилизация*); ***-ств(о)*** (*строительство*); ***-ø*** (*беседа*); ***-от(а)*** (*зевота*); ***-б(а)*** (*пальба*); ***-ня*** (*стрекотня*); ***-аж*** (*массаж*); ***-ёж*** (*грабёж*); ***-ок*** (*толчок*); ***-ость*** (*жалость*); ***-изм*** (*уклонизм*); ***-ур(а)*** (*режиссура*); ***-ищ(е)*** (*побоище*); ***-ч(а)*** (*передача*); ***-ль*** (*гибель*); ***-ыш*** (*проигрыш*); ***-от*** (*топот*); ***-ух*** (*голодуха*); ***-он*** (*выпивон*); ***-ин(ы)*** (*крестины*). Такое многообразие средств деривации (и в том числе многообразие продуктивных средств деривации, таких как суффиксы *-нић*; *-ић*; *-к(а)*; *-ција*; *-ств(о)* и *-ø*) вызывает желание понять, чем они различаются между собой.

Исследователи отглагольных имён существительных в различных языках неоднократно замечали, что отглагольные существительные, образованные по разным моделям, имеют разные свойства. Чаще всего отмечались различия в акциональных свойствах различных существительных, хотя и актантные различия тоже иногда попадали в поле зрения исследователей — см., например, Brinton 1995, Martin 2008, Пешковский 1956, Иванникова 1972.

Данное исследование отличается от предыдущих сразу в двух аспектах: во-первых, оно проводится не на априорном, а на корпусном материале, и во-вторых в нём рассматриваются различия отглагольных имён по синтаксическом параметру, а именно, по признаку позиции, занимаемой отглагольным именем в предложении.

Ожидается, что это обогатит наши знания об отглагольных именах, моделях их деривации, отличии одних моделей от других, а также о том, как синтаксическая позиция имени ситуации в предложении связана с другими его характеристиками, насколько она определяется его синтаксическими и семантическими свойствами, насколько — свойствами того слова, от которого имя ситуации зависит.

1.1. Модели деривации, выборка, значения признака «синтаксическая позиция отглагольного имени»

Настоящее исследование учитывает только имена ситуаций, образованные по наиболее продуктивным для русского языка моделям. В качестве условной «меры продуктивности» был взят следующий критерий: модель деривации имён ситуаций считается продуктивной, если в рабочей выборке из примерно 3500 отглагольных существительных, составленной на основании словника Учебного русско-французского словаря А. А. Зализняка, оказалось 200 или больше имён, образованных по этой модели. Вот какие суффиксы соответствуют этому критерию:

- нулевой суффикс, образующий имена мужского рода (*ударить — удар*)[1] — 468 существительных в рабочей выборке;

[1] Нулевой суффикс отглагольного имени, вообще говоря, позволяет поставить вопрос о направлении деривации: от глагола к имени, как во всех остальных отглагольных существительных, или от имени к глаголу, как это диктуется принципом монотонности деривации: морфемы могут только добавляться, но не удаляться, поэтому деривация должна идти от единиц с меньшим количе-

- суффикс *-к(а)* (*обработать — обработка*) — 285 существительных;
- суффикс *-нщ(е)/тщ(е)* (*разрушить — разрушение*, *прибыть — прибытие*) — 2085 существительных.

Для проведения настоящего исследования было выбрано по 10 достаточно частотных русских отглагольных существительных, образованных по этим трём моделям и достаточно часто употребляющихся для обозначения ситуаций (а не их участников). Из существительных на *-ние/тие* использовались следующие десять имён[2]: *отношение* (2766), *движение* (1744), *развитие* (1610), *положение* (1511), *состояние* (1503), *событие* (1179), *управление* (921), *желание* (847), *исследование* (839), *создание* (810). Из существительных мужского рода с нулевым суффиксом в нашу выборку попали следующие: *труд* (1504), *шаг* (1407), *ход* (1159), *сон* (1057), *страх* (890), *рост* (880), *перевод* (759), *удар* (700), *выход* (676), *обед* (676). И наконец, вот десять отобранных нами отглагольных существительных на *–ка*: *улыбка* (764), *оценка* (716), *ошибка* (679), *попытка* (493), *разработка* (422), *поддержка* (410), *подготовка* (380), *выставка* (368), *поездка* (360), *постановка* (273).

Затем при помощи НКРЯ для каждого отобранного существительного было собрано по 110 контекстов его употребления. Учитывались только те контексты, в которых отглагольное имя обозначает ситуацию, а не участника ситуации и не её результат. Не учитывались также употребления имени в устойчивых выражениях. Все собранные примеры (по 1100 на каждый суффикс отглагольного имени) были размечены по интересующему нас признаку «синтаксическая позиция отглагольного имени в предложении». В данном исследовании мы различали такие его значения:

1. Субъект предложения, в том числе субъект подчинённого предложения, бытийного предложения, предложения идентификации (1).

*(1) В расчёт принимаются и другие параметры: наличие патентов и лицензий (либо вероятность их получения); является ли **разработка** самостоятельным проектом или компонентом другого большого проекта и т. д. <…>[Европейский форум бухгалтеров // «Бухгалтерский учёт», 2003.06.16]*[3]

2. Прямой объект в предложении, в том числе выраженный родительным отрицания (2).

(2) Сразу после собрания акционеров Карпухин пишет председателю Госдумы Селезневу письмо, в котором слезно просит спикера не допустить перевода уголовного дела Ларина и Харитонкина в Тверскую область. [Евгений Толстых. Пивка для рывка // «Совершенно секретно», 2003.09.01].

3. Непрямой (дативный) объект (3).

(3) При этом исследованию сути потребительских запасов, по мнению автора, уделяется недостаточно внимания. [Потребительские запасы — сущность и подход к анализу // «Вопросы статистики», 2004].

4. Прочие приглагольные зависимые (4).

(4) ЮКОСу предложат дистанцироваться от попыток воздействия на государство и политико-экономического влияния на Думу. [ЮКОС — что дальше? Ходорковский и 'кремлевская доля' // «Газета», 2003.07.07].

5. Генитивное зависимое в именной группе (5).

(5) Эта стадия сна знакома лишь млекопитающим. [Кто заснул первым? // «Знание — сила», №1", 2003].

6. Предложное зависимое в именной группе (6).

(6) Тогда Минобразования РФ по ходатайству ассоциации Сибирское соглашение и администрации Красноярского края приняло решение о подготовке специалистов для этой отрасли на базе университета. [Грустные каникулы // «Поиск», 2003.09.12].

7. Независимое употребление — заголовок, парцелляция, часть разбитого на абзацы длинного перечисления (7).

(7) С. В. Моржухина, В. А. Балакин, С. В. Котенко, Л. П. Чермных, М. С. Хозяинов, О. Э. Вавина, В. М. Шкинев, Т. В. Данилова Е. В. Шкинева. Экологическое состояние р. Москвы на территории Раменского района Московской области [Экологическое состояние р. Москвы на территории Раменского района Московской области // «Геоинформатика», 2004.09.26].

8. Именная часть сказуемого (8).

(8) Внутри России, безусловно, убийство Сергея Юшенкова было самым значимым, печально-значимым событием. [Комментарий В. Третьякова на «Эхе Москвы» (2003.04.19)].

Разбиение примеров на эти восемь групп, как представляется, может дать достаточно показательную картину распределения существительных, образованных по различным моделям.

ством морфем к единицам с большим их количеством. Однако, из системных соображений и вслед за сложившейся традицией, мы будем считать, что имена с нулевым суффиксом также образуются от глаголов — или, по крайней мере, от глагольных основ.

[2] Здесь и далее в скобках после слова приводится его частотность.

[3] Все примеры получены из НКРЯ, поиск производился в сентябре–октябре 2008 г. Ссылки на источники примеров даны в формате Корпуса; слово, о котором идёт речь в данном примере, выделяется жирным шрифтом.

1.2. Актантные и акциональные свойства имён, образованных по разным моделям

Рассмотрим сначала актантные и акциональные свойства имён, образованных по интересующим нас моделям, и попробуем на основании их сформулировать гипотезы о том, какие синтаксические позиции эти имена будут занимать в предложении.

Основные свойства имён, образованных по трём вышеперечисленным моделям, были исследованы в Пазельская, в печати. Повторим здесь главные результаты этой работы. Вот как можно обобщить свойства трёх моделей деривации отглагольных существительных.

1. имена на *-ние/тие*
   - 1.1. чаще других употребляются в форме множественного числа;
   - 1.2. чаще, чем другие, имеют при себе внешний аргумент, выраженный притяжательным прилагательным;
2. имена мужского рода с нулевым суффиксом
   - 2.1. реже имён двух других типов имеют при себе выраженный внутренний аргумент, и ещё реже он выражается родительным падежом;
   - 2.2. чаще выражают при себе периферийных участников ситуации, отличных от внешнего и внутреннего аргументов, причём чаще всего они выражаются предложной группой;
3. отглагольные имена на *-ка*
   - 1.1. чаще других употребляются в контекстах, когда один или более участников ситуации, обозначаемой именем, контролируется какой-нибудь другой именной группой (или другими именными группами) в том же предложении;
   - 1.2. реже имеют при себе атрибуты, указывающие на локализацию ситуации во времени;
   - 1.3. существенно превосходят остальные существительные по тому, как часто внешний аргумент выражается родительным падежом.

Эти свойства позволяют сделать следующие выводы. Суффикс *-ка* и суффикс *-ние* противопоставлены таким образом, что первый гораздо чаще второго (и чаще всех остальных суффиксов отглагольных имён) образует отглагольные существительные с только каузативной семантикой, без возможности декаузативации, то есть устранения из обозначаемой существительным ситуации внешнего аргумента. Именно поэтому: имена на *-ка* чаще, чем другие имена, имеют при себе внешний аргумент, который никуда не устраняется не только семантически, но и синтаксически, и либо выражается родительным падежом (свойство 3.3 выше), либо контролируется какой-либо другой именной группой в предложении (свойство 3.1).

Более того, имена на *-ние*, по сравнению с существительными на *-ка*, не только реже выражают внешний аргумент, но и даже в этом случае выражают его таким способом, который оставляет возможность выразить внутренний аргумент при том же имени — а именно, притяжательным прилагательным. Имена на *-ка*, напротив, чаще всего выбирают генитивную стратегию выражения внешнего аргумента, которая исключает одновременное выражение внутреннего аргумента самым частотным для него способом — родительным падежом. Тем самым подтверждается, что имена на *-ка* более склонны к описанию каузативных ситуаций и к фокусированию на каузирующей подситуации, чем имена, образованные по другим моделям.

Отглагольные существительные мужского рода оказались слишком непохожи на исходные глаголы — они реже других отглагольных имён имеют при себе выраженный внутренний аргумент (п. 2.1). Зато они чаще других существительных имеют при себе периферийных участников, выраженных предложными группами (п. 2.2).

Это позволяет усомниться в исходной гипотезе, что эти отглагольные существительные наследуют свою аргументную структуру от глагола — как, впрочем и любые другие, что возвращает нас к широко обсуждавшейся в литературе проблеме наличия у отглагольных имён собственных аргументов и связи между этими аргументами и аргументами исходного глагола (ср., например, Grimshaw 1990 и дискуссию, вызванную этой книгой). Но если в случае с другими отглагольными существительными можно найти какие-то подтверждения наследования аргументов от глагола (состав участников, категориальные ограничения, регулярность деривации, аспектуальная структура, наследование оформления участников неструктурными падежами), то в случае имён с нулевым суффиксом таких подтверждений не видно.

Исследование такой ограниченной выборки позволяет сказать не слишком много об акциональном измерении отглагольных существительных. Тот факт, что отглагольные существительные на *-ка* реже других имеют при себе атрибуты, указывающие на локализацию во времени (3.2), позволяет предположить, что имена, образованные по этой модели, чаще обозначают обобщённые деятельности, что согласуется с присущим именам на *-ка* акцентом на каузативном компоненте ситуации. Наконец, относительно частое употребление имён на *-ние* в форме множественного числа (1.1) подтверждает то, что эти существительные в наименьшей степени подвержены каким-либо ограничениям на образование и употребление.

Исходя из этого, можно предположить, что отглагольные существительные с нулевым суффиксом чаще будут употребляться независимо — в назывных предложениях, заголовках и т. п., а также в контек-

стах, характерных для предметных имён, не связанных ни с какими глаголами. С другой стороны, имена на *-ка*, как наиболее склонные обозначать каузативные и процессуальные компоненты ситуаций, ожидаются прежде всего в двух типах контекстов: во-первых, при глаголах, подчиняющих отглагольные имена в виде прямых дополнений, — т. е. так называемых полувспомогательных глаголов, или лексических функций (таких как *производить обработку*, *давать оценку*, *осуществлять разработку* и т.п.), и, во-вторых, в контексте предикатов знания, мнения, эмоционального восприятия. От имён на *–ние/тие* и в этом случае, как и во всех остальных, мы ожидаем промежуточных, центральных, нейтральных свойств.

Посмотрим, насколько корпусной материал оправдывает эти ожидания.

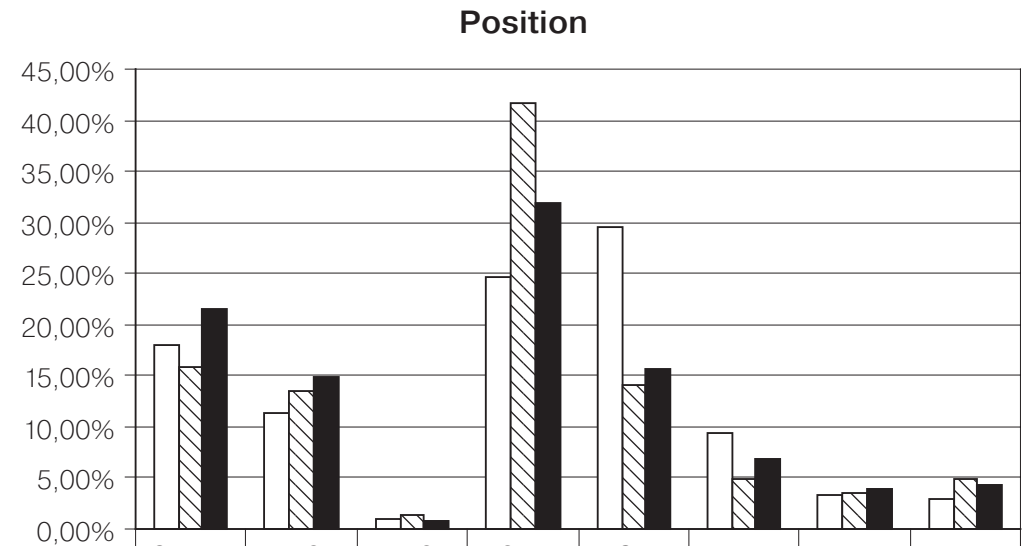
2. Модели деривации и синтаксическая позиция

Распределение отглагольных существительных, образованных по трём исследованным моделям, по признаку синтаксической позиции в предложении, показано в Таблице 1.

Для наглядности изобразим эти же данные (в процентном представлении) на графике (рис. 1).

Таблица 1. Распределение моделей деривации отглагольных существительных по признаку синтаксической позиции в предложении

	-ние		zero_m		-ка	
свойство	кол-во	%	кол-во	%	кол-во	%
субъект (Subject)	197	17,91%	175	15,91%	237	21,55%
прямой объект (DO)	124	11,27%	149	13,55%	163	14,82%
непрямой объект (IDO)	11	1%	15	1,36%	9	0,82%
другое приглагольное (Obliq)	271	24,64%	459	41,73%	351	31,91%
приименной генитив (Gen)	326	29,64%	156	14,18%	173	15,73%
другое приименное (PP)	103	9,36%	53	4,82%	75	6,82%
независимое употребление (Indep)	36	3,27%	39	3,55%	44	4%
именная часть сказуемого (Praed)	32	2,91%	54	4,91%	48	4,36%
всего	1100	100%	1100	100%	1100	100%



	Subject	DO	IDO	Obliq	Gen	PP	Indep	Praed
ние	17,91%	11,27%	1,00%	24,64%	29,64%	9,36%	3,27%	2,91%
zero_m	15,91%	13,55%	1,36%	41,73%	14,18%	4,82%	3,55%	4,91%
ка	21,55%	14,82%	0,82%	31,91%	15,73%	6,82%	4,00%	4,36%

Рис. 1. Распределение моделей деривации отглагольных существительных по признаку синтаксической позиции в предложении (в процентах)

Из Таблицы 1 и Графика 1 можно сделать следующие обобщения:

1. Имена на *-ние/тие* вдвое превосходят остальные существительные по использованию в качестве генитивных прииименных зависимых (Gen) и несколько чаще остальных существительных бывают прочими прииименными зависимыми (PP).
2. Отглагольные имена с нулевым суффиксом сильно превосходят остальные существительные по использованию в качестве косвенных приглагольных зависимых (Obliq), и несколько реже, чем другие имена существительные, используются в качестве отличных от генитивных прииименных зависимых (PP).
3. Отглагольные имена на *-ка* несколько чаще других отглагольных имён выступают в качестве субъекта и прямого дополнения при глаголах.
4. Другие различия между именами, образованными по разным деривационным, моделям, представляются слишком незначительными и не заслуживающими внимания.

3. Обсуждение результатов

Рассмотрим теперь, как наши обобщения согласуются с уже имеющимися знаниями о моделях деривации отглагольных имён и с предсказаниями на основе этих знаний.

Самым ожидаемым образом ведут себя отглагольные существительные на *-ка*: действительно, мы ожидали, что они, в силу своей склонности к обозначению процессов, часто будут выступать в контексте полувспомогательных глаголов. А поскольку такие глаголы чаще всего подчиняют отглагольное имя в форме прямого дополнения, совершенно неудивительно, что имена на *-ка* чаще других выступают в виде прямых дополнений. Это же позволяет объяснить и относительно высокую частотность употреблений существительных на *-ка* в качестве субъекта предложения: тексты, стилистика которых допускает отглагольные имена, изобилуют также и пассивными конструкциями, в которых прямое дополнение становится подлежащим (9). Именно такие конструкции отчасти и обеспечивают именам на *-ка* частотность в качестве подлежащего.

(9) О. ДМИТРИЕВА: В медицине у нас делается попытка обязательной системы страхования. [Беседа Н. Болтянской с О. Дмитриевой в прямом эфире «Эха Москвы» (2003.04.06)].

Второй частотный контекст, в котором имена на *–ка* выступают в роли подлежащего, сходен с первым: это тоже полувспомогательные глаголы, но такие, которые имеют имя ситуации единственным, подлежащным, актантом (10).

(10) Сейчас идёт их инвентаризация и параллельно-разработка методики разграничения. [Анастасия Корня. Реформа местного самоуправления ложится на карту // «Время МН», 2003.08.01].

Свой вклад в оба типа употреблений отглагольных имён на *-ка* — в виде подлежащего и в виде прямого дополнения — вносят и предикаты знания, мнения и эмоций (11).

(11) Бесперспективными представляются попытки декорирования изделий пейзажной и жанровой росписью, сводящей их до уровня китча. [Луховицкие узоры // «Народное творчество», 2003.08.18].

Остальные две модели деривации отглагольных имён, однако, скорее отклоняются от ожиданий, но эти отклонения представляется возможным объяснить исходя из уже известных нам свойств этих имён. Имена с нулевым суффиксом вовсе не лидируют по количеству независимых употреблений, и только употребление в качестве приглагольных зависимых, отличных от прямого объекта, сближает их с употреблениями предметных существительных. Действительно, именно употребление в качестве предложных актантных и неактантных участников ситуации, в отличие от прямого объекта и субъекта, определяется прежде всего глаголом, а не именем, и потому больше характерно для предметных имён, предъявляющих в целом меньше требований к подчиняющим их глаголам.

Возможно, употребление отглагольного существительного в качестве вершины отдельно стоящей именной группы (то, что мы называли здесь независимым употреблением) в принципе зависит скорее не от грамматических свойств отглагольного имени, а от прагматических факторов, таких как тематика и сфера употребления имени.

Существительные же на *-ние/тие*, показавшие себя скорее нейтральными в плане актантных и акциональных свойств, по признаку синтаксической позиции в предложении оказались в нескольких отношениях противопоставлены остальным двум моделям. Больше всего внимания здесь заслуживает употребление в качестве прииименного генитивного зависимого. Большинство случаев такого употребления имён на *-ние/тие* — употребления при другом имени ситуации, когда имя на *–ние/тие* выражает одного из участников этой ситуации (12).

(12) Свою роль играли общие интересы, спортивные увлечения, совместные занятия, но главным содержанием возникновения собственно дружеских отношений была увлекательность общения. [М. Э. Боцманова, Триггер Р. Д. Изучение психологии подростка в лаборатории Д. Б. Эльконина // «Вопросы психологии», 2004.02.10] [омонимия снята].

Здесь необходимо напомнить, что модель на *-ние/тие* — самый продуктивный в русском языке способ образования отглагольных существительных, и имена с этим суффиксом являются своего рода дефолтным способом выразить действие в виде имени. Именно поэтому, по-видимому, в случае, если возникает необходимость обозначить одно действие в качестве актантного зависимого другого действия, выбираются именно существительные на *-ние/тие*. То же можно сказать и об использовании отглагольных существительных на *-ние/тие* в качестве других, не генитивных, прииименных зависимых.

Фундаментальное свойство имён ситуаций — расхождение между глагольной семантикой и именной формой, и имена, образованные по различным моделям деривации, сочетают глагольные и именные свойства в разной степени (и в разной степени сохраняют свойства конкретного исходного глагола). Так, из трёх рассмотренных здесь моделей деривации существительные с нулевым суффиксом оказываются наиболее близкими к именам, а существительные на *-ка* — к исходным глаголам. Такие выводы можно было сделать ещё на основании корпусного исследования зависимых отглагольных имён — так, в пользу близости имён с нулевым суффиксом к предметным существительным говорит то, что они достаточно редко выражают при себе внутренний аргумент, а свойство имён на *-ка* часто встречаться в контекстах с контролем участника сближает их с инфинитивами и, тем самым, с глагольными формами.

Исследование синтаксической позиции отглагольных существительных в предложении в целом подтверждает эти выводы, но при этом даёт новое знание о самих этих синтаксических позициях. Так, мы узнали, что независимое употребление существительного определяется скорее не синтаксическими и вообще не грамматическими его характеристиками, а прагматикой. А употребление существительного в качестве предложных актантных и неактантных участников ситуации, как оказалось, зависит в большей степени не от подчиняемого существительного, а от подчиняющего его глагола.

Литература

1. *Grimshaw J.* Argument Structure. MA: MIT Press, 1990.
2. *Martin F.* The Semantics of Eventive Suffixes in French // Schäfer F. (ed.), 'SinSpec', Working Papers of the SFB 732. Stuttgart: University of Stuttgart, 2008. Vol. 1.
3. *Пазельская А. Г.* Модели деривации отглагольных существительных: взгляд из корпуса. // Киселёва К. Л., Плунгян В. А., Рахилина Е. В., Татевосов С. Г. (ред.) Корпусная грамматика русского языка. М.: ИМЛИ РАН, в печати.
4. *Пешковский А. М.* Русский синтаксис в научном освещении. М.: Учпедгиз, 1956.
5. *Иванникова Е. А.* К вопросу об аспекте изучения категории вида у отглагольных существительных в русском языке // Известия АН СССР, сер. лит. и языка. М., 1972. Том 31, № 2, с. 113–123.
6. *Шведова Н. Ю.* и др. (ред.). Русская грамматика. М., 1982. Т. I: фонетика, фонология, ударение, интонация, словообразование, морфология.

Просодия обращений в немецком языке в сопоставлении с русским[1]

Prosody of the German vocative NPs in contrast to the Russian ones

Палько М. Л. (m_palko@mail.ru)

Институт языкознания РАН

Исследуются просодические особенности обращений в немецком языке в сопоставлении с русским. Сравнительный анализ акцентирования обращений и в целом именных групп позволяет выдвинуть гипотезу о типологической особенности немецкого языка, состоящей в невозможности продвижения акцента в начало именной группы, которое характерно для большого класса обращений в русском языке.

Целью данной работы служит исследование просодических характеристик более чем однословных обращений в немецком языке и связанных с этим особенностей выбора акцентоносителя — словоформы-носителя коммуникативно релевантного тонального пика. Предметом нашего исследования служат именные группы с различной синтаксической структурой: *Herr Janzen!; Frau Müller!; Doktor Kozak!; Liebe Zuschauer!; Liebe Kollegen!*, сложные личные имена типа *Hans-Peter*. В условиях пространственной и психологической близости между говорящим и слушающим в русских более чем однословных обращениях акцент сдвигается в начало: ***Доктор*** *Козак, ну как же так?* (в данном примере и далее полужирным шрифтом выделяется словоформа-акцентоноситель.) Здесь акцентоносителем служит начальная словоформа *доктор*. Возникает вопрос: возможен ли аналогичный сдвиг в немецких обращениях.

Интонация русских и немецких обращений уже служила предметом лингвистического анализа, ср. работы [Кодзасов, 1996, 1999; Кравченко 1973: 128-133; Янко (в печати); Essen 1956: 32–36, 53–57]. Анализ различных комбинаций значений признаков психологической и пространственной удаленности говорящего от слушающего и средств их выражения, а также принципов выбора акцентоносителя на материале русских обращений был проведен Т. Е. Янко [Янко 2008].

Анализу просодии немецких более чем однословных обращений, таких как *Дорогие коллеги!*, *Ханс Петер!, Герр Мюллер!* посвящен раздел 1 ниже. Выбору акцентоносителя, анализ которого потребовал обращения к более широкому языковому материалу, в частности к выбору акцентоносителя в других немецких именных конструкциях посвящен раздел 2.

Материалом исследования служат записи немецких теле- и радиопередач, телесериалов, художественных фильмов, аудиозаписи диагностических предложений, прочитанных информантами носителями немецкого языка.

1. Просодия более чем однословных обращений в немецком языке

Точкой отсчета для анализа немецкого материала нам послужил русский язык. В русских обращениях акцентный пик может приходиться как на первый, так и на второй компонент линейной цепочки. Первый компонент имени слушающего акцентируется в случае, когда между собеседниками отсутствует психологическая дистанция, т.е. общение происходит на равных. Иначе говоря, центр тяжести переносится на начало имени слушающего. Так, в примере ***Галина*** *Петровна, Вам чайку налить?* акцентный пик фиксируется на начальной словоформе *Галина*, а финальная словоформа *Петровна* просодически редуцируется: она произносится на низких частотах голоса говорящего и в убыстренном темпе, релевантные движения тона на ней отсутствуют. Обратим внимание на то, что вне контекста обращения в именной группе *Галина Петровна* в соответствии с базовым принципом выбора акцентоносителя рус-

[1] Работа над темой финансируется Российским Государственным гуманитарным фондом, проект 09-04-00106а.

ского языка акцентоносителем служит словоформа *Петровна*, ср., например, *Это Галина* ***Петровна*** с акцентом на *Петровна*. ср. также ***Молодой*** *человек! Вы мне не поможете?* с акцентом на *молодой*. О базовом принципе выбора акцентоносителя см. [Янко 1991; 2001: 190; 2008: 38-42] и цитированную там литературу. Второй компонент в русских обращениях тоже может служить акцентоносителем в условиях соблюдения личностной дистанции, т. е. при общении в официальном тоне (*Марья* ***Ивановна****, зайдите к начальнику!*), или при гневе говорящего (*Галина* ***Петровна****! Вы что себе позволяете?!*; *Молодой* ***человек****! Прекратите это безобразие!*).

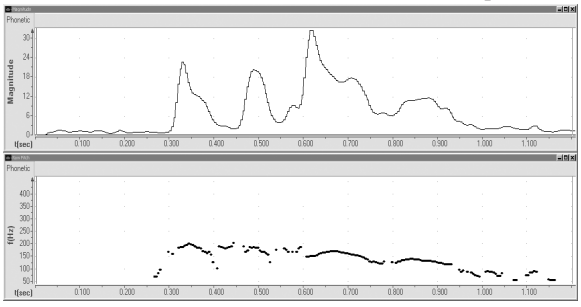
Итак, в русском языке имеется возможность перемещения акцента в зависимости от условий общения и целей говорящего, т.е. при формировании обращений русский язык отходит от своего основополагающего принципа при выборе акцентоносителя ритмических групп, ориентированного на синтаксические иерархии и фактор активации, и использует линейную иерархию, при которой акцентоносителем становится начальная (при неофициальном обращении к слушающему) или финальная (при официальном обращении к слушающему) словоформа в линеаризованной цепочке, представляющей имя (титул) слушающего. Акцентоноситель здесь выбирается в соответствии с принципом линейной структуры, т. е. синтаксическая структура на выбор акцентоносителя не влияет.

Анализ массива записей немецкой диалогической речи показал, что в атрибутивных именных группах с прилагательным и приложением[2] акцент фиксируется на синтаксической вершине словосочетания, т. е. что просодические конструкции с акцентом на первом компоненте сложного имени в немецком языке невозможны. В различных контекстах, независимо от типа общения в немецком обращении, состоящем из двух словоформ, акцентоносителем служит конечная словоформа.

Обратимся к примерам (1)–(4) и соответствующим тонограммам, которые иллюстрируют это положение. В связи с анализом тонограмм поясним, что для анализа звучащей речи мы используем компьютерную систему Speech Analyzer. Тонограммы ниже отражают изменение частоты тона в герцах. Тонограммы на рисунках ниже расположены на нижней, а в примере (5) — на единственной панели. При анализе примеров (1)–(4) и (6)–(8) используется также график изменения интенсивности, который расположен на панели над графиком частоты.

(1) *Doktor Kozak! Ich hab' ein Dokument das ich Ihnen zeigen möchte.*
'Доктор Козак! У меня есть один документ, который я бы хотел Вам показать.'

Тонограмма 1



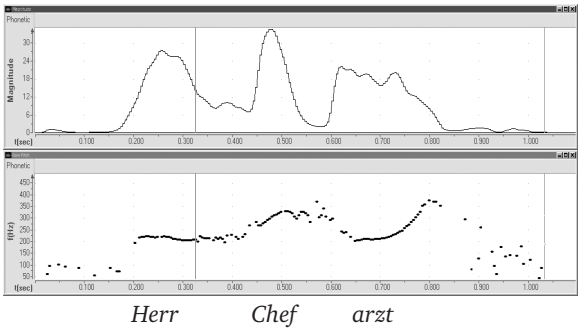
Doc tor Ko zak

Это пример из художественного фильма. Коллеги адвокаты обсуждают предстоящее судебное заседание и один из участников процесса в достаточно доверительном тоне обращается к другому. Тонограмма показывает, что на ударный слог первого слова *Doktor* приходится подъем тона, на второй словоформе *Kozak* фиксируется достаточно высокий, ровный, слегка нисходящий тон и повышенная интенсивность. Просодической редукции, которой, исходя из анализа русского материала, можно было бы ожидать на втором слове, нет.

Обратимся к примеру (2) из телефильма о работе врачей.

(2) *Herr Chefarzt, ich brauche ihre Unterschrifft… 'Господин начальник, мне необходима ваша подпись…'*

Тонограмма 2



Herr Chef arzt

Медсестра обращается к доктору с просьбой, используя при этом просительный тон. Первая словоформа характеризуется ровным тоном и достаточно высокой интенсивностью, как и в примере (1), вторая словоформа *Chefarzt* (на тонограмме она выделена курсорами) характеризуется восходящим движением тона на ударном слоге в сочетании с повышенной интенсивностью. Иначе говоря, акцентный пик и пик интенсивности приходится на вторую словоформу *Chefarzt*. Во второй словоформе вслед за подъемом на ударном слоге *Chef-* наблюдается

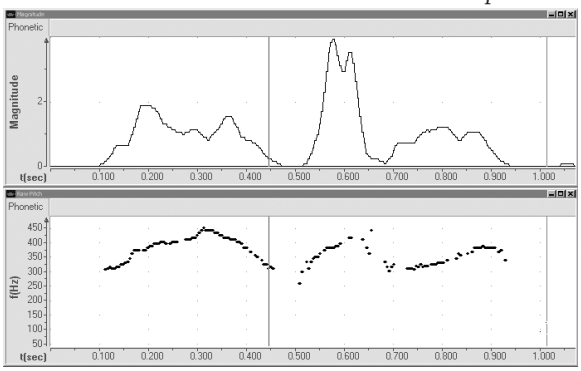
[2] В связи с синтаксической структурой именных групп поясним, что в именных группах с несогласованным определением акцентоносителем служит не синтаксическая вершина, а постпозитивное несогласованное определение, ср. русский пример *Саша с Уралмаша* с акцентоносителем несогласованным определением *Уралмаша*.

существенный подъем тона на *-arzt,* который объясняется незавершенностью высказывания. Таким образом, данный пример служит интересным образцом комбинации коммуникативных признаков вокатив плюс незавершенность коммуникации, см. также пример (3) ниже.

В примере (1) и в примере (2) пик интенсивности наблюдается на второй словоформе. Рассмотрим пример (3).

(3) Lilija Lvovna!
'Лилия Львовна!'

Тонограмма 3



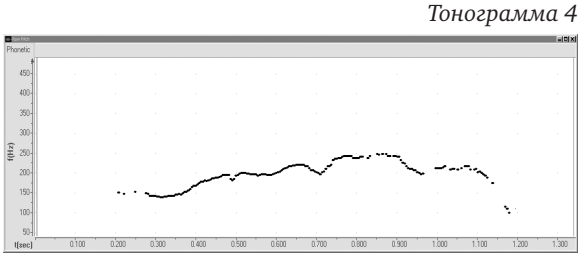
Lilija Lvo vna

Это пример из телесериала, в котором хозяйка отеля в Берлине обращается к экскурсоводу русской группы. На первой словоформе фиксируется подъем тона на ударном слоге, на второй словоформе — подъем тона и высокая интенсивность. Пик интенсивности, как и в предыдущих примерах, наблюдается на ударном слоге второй словоформы. На заударном слоге второй словоформы наблюдается подъем тона, который, как и в примере (2), маркирует значение незавершенности.

Проблема выбора носителя акцентного пика релевантна и для анализа сложных имен типа Marie-Luise, Anne-Elisabeth, Hans-Peter, выступающих в роли обращений. Акцентоносителем здесь тоже служит второй компонент сложного имени. Пример (4) является фрагментом интеллектуального шоу, ведущий телепередачи приветствует участницу.

(4) Marie-Luise! Herzlich willkommen!
'Мари-Луиза! Добро пожаловать!'

Тонограмма 4



Marie- Luise

Тонограмма показывает, что в данном случае на ударный слог первого имени *Marie* приходится подъем тона, акцентный пик достигается на ударном слоге второго имени — *Luise*.

Итак, немецкие примеры (1)–(4) показывают, что акцентный пик и пик интенсивности фиксируется на второй словоформе имени. При этом обращение может содержать указание на то, что сеанс коммуникации не заканчивается обращением. Примеры (2) и (3) показывают, что в контексте незавершенности на втором компоненте обращения фиксируется подъем частоты основного тона. При завершенной коммуникации на второй словоформе имени фиксируется падение.

Анализ немецких обращений позволил предположить, что типологической особенностью немецкого языка является невозможность продвижения акцента в начало именной группы, которое характерно для русского языка. В немецком языке в атрибутивных группах с прилагательным и приложением акцент фиксируется на синтаксической вершине словосочетания, т.е. просодические конструкции с акцентом на первом компоненте сложного имени в немецком языке невозможны. Так, поиск гипотетических примеров типа **Liebe Kollegen!* (ср. ***Молодой*** *человек!*) с акцентом на прилагательном дал нулевые результаты. Совершенно аналогично, отсутствуют и другие примеры с акцентом на первом компоненте, которые предположительно в немецком языке могли бы быть, если бы немецкий язык использовал просодическую стратегию неофициального общения, аналогичную русской, **Herr Buschmann!* с акцентом на компоненте *Herr* и **Doktor Kozak!* с акцентом на компоненте *Doktor.*

2. Выбор акцентоносителя в именных группах

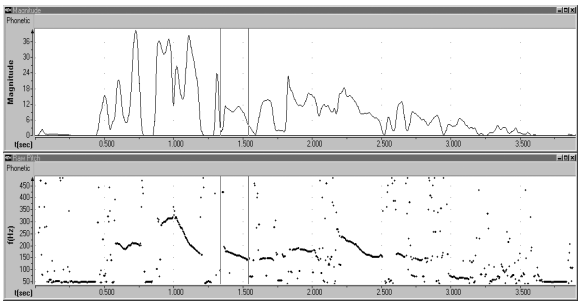
Для подкрепления нашей гипотезы обратимся к более широкому языковому материалу. Нашу гипотезу о невозможности переноса акцента в начало именной группы в немецком языке подтверждает анализ высказываний с контрастной темой с контрастом на прилагательном. Контраст на прилагательном — это особый контекст, который в русском языке требует особого акцентного выделения препозитивного прилагательного в структуре именной группы. Поясним также, что для анализа здесь выбраны высказывания с контрастной темой, а не с контрастной ремой, потому что в начале русского предложения заударная редукция, которая наступает на синтаксической вершине, следующей за контрастно выделенным прилагательным, ощущается особенно отчетливо. В русском языке, контрастная тема с контрастом на прилагательном

выражается подъемом тона на акцентоносителе контрастной коммуникативной составляющей с бóльшим перепадом частот и большей интенсивностью, чем при обычной теме. Вслед за контрастной темой идет интонационный спад, проявляющийся в пониженной интенсивности и в низком уровне частоты. Так, например, в предложении (5) *Его первая книга была самая интересная* контрастно выделенной оказывается словоформа *первая*, за ней следует интонационный спад на словоформе *книга*, определяющийся нисходящим тоном и небольшой интенсивностью. В немецком языке интонирование контрастной темы с контрастом на прилагательном отличается от русского. Различие ощущается на слух и состоит в отсутствии редукции слогов, следующих за контрастно выделенным. Акцент принимают и прилагательное — носитель контраста, и опорная словоформа — синтаксическая вершина словосочетания, то есть в немецком эквиваленте предложения *Его первая книга была самая интересная* выделению подвергается и эквивалент словоформы *первая*, и эквивалент словоформы *книга*.

Тонограммы (5) и (6) представляют русский пример и его немецкий эквивалент соответственно.

(5) Его первая книга была самая интересная.

Тонограмма 5



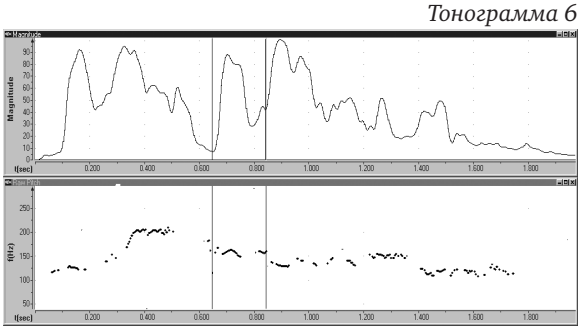
Его первая кни га была самая интересная

Пример (5) заимствован из [Янко 2008: 61]. На ударном слоге словоформы *первая* мы видим подъем тона и высокую интенсивность, маркирующие контраст, на ударном слоге словоформы *книга* наблюдается сравнительно небольшая интенсивность и низкое, ровное, слегка нисходящее движение тона. Ударный слог словоформы *книга* взят в курсоры. На словоформе *книга* наблюдается просодическая редукция.

Обратимся к немецкому примеру (6).

(6) Sein erstes Buch war das interessanteste.
‘Его первая книга была самая интересная.’

Тонограмма 6



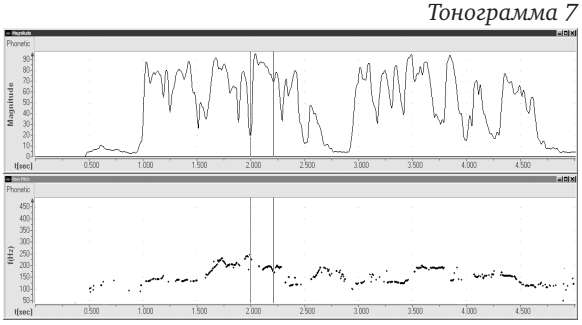
Sein erstes Buch war das interessanteste

Тонограмма показывает, что ударный слог словоформы *erstes* ‘первая’, как и в русском примере, несет высокий, крутой и интенсивный подъем тона, выражающий контраст, а на синтаксической вершине словосочетания (она выделена курсорами) по сравнению с русским языком редукция отсутствует. Словоформа *Buch* ‘книга’ характеризуется относительно ровным движением тона и высокой интенсивностью, что видно на графике интенсивности.

Рассмотрим еще один немецкий пример контрастной темы — пример (7).

(7) Von den amerikanischen Weinsorten, mag er besonders die aus Kalifornien.
‘Среди американских вин он предпочитает калифорнийские’.

Тонограмма 7



Von den ameri ka nischen Weinsorten mag er besonders die aus Kali fornien

Как видно на тонограмме, ударный слог словоформы *amerikanischen* ‘американские’ несет высокий и крутой подъем тона. На ударном слоге словоформы *Weinsorten* ‘вина’ *-wein* (он выделен курсорами) фиксируется ровный тон, однако, ударный слог характеризуется высокой интенсивностью, что видно на графике интенсивности. Выделение словоформы *Weinsorten*, таким образом, в этом случае определяется высокой интенсивностью и сравнительно высоким тоном.

Таким образом, в немецком языке в именных группах в контексте контрастной темы и в контексте обращения конечная словоформа характеризуется тональным выделением и/или высокой интенсивностью.

Следует сказать, что данные, полученные в исследованиях Т. Е. Янко [2008: 60–70] и Н. А. Фаустовой [2006], говорят о невозможности переноса акцентного пика в начало именной группы с согласованным определением и об отсутствии заударной редукции и в других западных языках: в английском, французском, датском и польском языках. Так Т. Е. Янко при анализе контрастных тем заключила, что в датском языке отсутствие редукции на опорном слове словосочетания прежде всего ощущается на слух, возникает впечатление, что интонационно выделенными оказываются обе словоформы: и прилагательное-носитель контраста, и синтаксическая вершина. Аналогично, в польском языке синтаксическая вершина также получает явственно ощутимое на слух акцентное выделение, которое преимущественно состоит в повышенной интенсивности. Во французском языке в подобных высказываниях акцентное выделение на прилагательном, независимо от порядка слов, вообще отсутствует: несмотря на контраст интонационно выделенной оказывается синтаксическая вершина. В наших примерах — это эквивалент русской словоформы *книга* из примера (5). Наиболее близким русскому в рассмотренном отношении оказывается английский язык. Словоформа-вершина словосочетания произносится на низких частотах и перепад частот между контрастной темой и соответствующей синтаксической вершиной очень существен. Однако интенсивность падает не столь явно, как в русском.

Отсутствие в немецком языке аналогов вокативных типов русского языка с акцентом на первом компоненте обращения типа русских ***Молодой*** *человек!* и ***Марья*** *Ивановна!* подкрепляется общей особенностью типологического портрета немецкого языка, который состоит в том, что в именной группе конечный элемент линейно-акцентной структуры акцентно выделен — тонально и/или средствами повышенной интенсивности, а заударная редукция отсутствует.

Литература

1. *Кодзасов С. В.* Комбинаторная модель фразовой просодии // Просодический строй русской речи. М., 1996.
2. *Кодзасов С. В.* Уровни, процессы и единицы в интонации // Проблемы фонетики. Вып. III. М.: Наука, 1999. С. 196–216.
3. *Кравченко М. Г., Зыкова М .А., Светозарова Н. Д.* Ударение и интонация в немецком языке. Л.: Просвещение, 1973.
4. *Фаустова Н. А.* Сопоставительный анализ английской, французской и русской интонации // Труды международной конференции «Диалог-2006», 2006. С. 527–533.
5. *Янко Т. Е.* Интонационные стратегии русской речи в сопоставительном аспекте. М.: Языки славянской культуры, 2008.
6. *Янко Т. Е.* Коммуникативная структура с неингерентной темой // Научно-техническая информация, 1991. Сер. 2. №7.
7. *Янко Т. Е.* Коммуникативные стратегии русской речи. М.: Языки славянской культуры, 2001.
8. *Янко Т. Е.* (в печати) Стереотипы звучащей речи: интонация обращений. // Логический анализ зыка. Моно-, диа- и полилог в различных языках, веках и культурах.
9. *Essen O. von.* Grundzüge der hochdeutschen Satzintonation. Düsseldorf: Ratingen, 1956.

Невербальный коммуникативный акт утешения: материалы к построению словаря невербальных коммуникативных актов[1]

Nonverbal communicative act of consolation: materials for a dictionary of nonverbal communicative acts

Переверзева С. И. (P_Sveta@hotmail.com)

Российский государственный гуманитарный университет, Москва

Настоящая работа продолжает серию исследований, посвящённых механизмам и способам взаимодействия невербальных и вербальных знаковых кодов в коммуникативном акте. В докладе обсуждаются некоторые проблемы построения словарей русских речевых и неречевых актов. В качестве примера словарного описания невербального коммуникативного акта приводится предварительный вариант статьи одного невербального акта утешения.

Введение

Одной из актуальных задач современной лингвистики является построение словаря русских речевых актов. Значительная часть работы для решения этой задачи уже проделана. Составлены толкования основных глаголов и некоторых других единиц, вводящих речевые акты, описаны условия их употребления, исследованы семантика, прагматика и синтаксические особенности этих единиц. Однако некоторые аспекты, связанные с взаимодействием в устной коммуникации знаков естественного языка и языка тела, в теории речевых актов практически еще не рассматривались.

В настоящей работе речь идёт о двух таких аспектах, а именно о невербальных знаках, сопровождающих речевой акт, и о невербальных аналогах речевого акта. Утверждается, что информация об этих единицах должна быть представлена в словаре русских речевых актов, причём в отдельных зонах. Рассматривается понятие невербального коммуникативного акта[2], которое является расширением понятия невербального аналога речевого акта. Ставится задача построить словарь русских невербальных коммуникативных актов и приводятся материалы для словарной статьи невербального коммуникативного акта утешения.

1. Невербальные единицы, сопровождающие речевой акт

Имеется много работ, посвящённых анализу и толкованию речевых актов (см., например, Гловинская 1993, Вежбицкая 1987, Шатуновский 2000). При этом знакам языка тела, сопровождающим речевые акты в устной коммуникации, обычно уделяется крайне мало внимания. О невербальных знаках либо не упоминают вовсе; либо, в лучшем случае, их приводят в виде краткого списка жестов[3] внутри словарной статьи, описывающей синонимический ряд основного глагола, которым вводится тот или иной речевой акт (см., например, описания синонимических рядов с вершинами *просить*, *предсказывать* и *обещать* в НОСС 2003).

Между тем, если мы ходим изучить и описать функционирование речевого акта в устном диало-

[1] Работа выполнена в рамках проекта «Части тела в русском языке и русской культуре» при поддержке Российского гуманитарного научного фонда (грант РГНФ № 07-04-00203а).

[2] Об этом понятии см. в книге Крейдлин 2002, с. 76.

[3] Здесь и далее слово *жест* понимается в широком смысле, то есть *жестами* называются собственно жесты (знаковые движения рук, ног, головы и плеч), мимика, или выражения лица, знаковые статические положения (позы), знаковые телодвижения, взгляды и комплексные формы поведения — манеры (см. Крейдлин 2002).

ге, то одного лишь списка жестов, в норме сопровождающих данный речевой акт, недостаточно. Необходимо также знать хотя бы то, каковы правила взаимодействия вербальных и невербальных единиц в пределах одного коммуникативного акта[4] и каковы правила или более общие закономерности, определяющие выбор говорящим того или иного знакового кода в коммуникации. Дело в том, что в очень многих диалогических ситуациях невербальные компоненты играют важную роль. Покажу это на двух примерах.

Во-первых, некоторые высказывания невозможно понять, не зная, какими жестами они сопровождаются, ср. *Вон там*; *Он вот такого роста*; *Отец вчера вот такую рыбу поймал*. Во всех этих высказываниях присутствует явление, называемое по-английски *mixed syntax*, — соединение синтаксических элементов естественного языка и языка тела.

Во-вторых, неисполнение жеста, сопутствующего данному речевому акту, может привести к коммуникативной неудаче. Например, речевые акты уверения и обещания в норме сопровождаются **взглядом в глаза адресату**. Если, произнося слова *Уверяю тебя* или *Клянусь тебе*, человек не смотрит в глаза собеседнику, тот может заподозрить его в неискренности, и акты уверения или обещания не произведут ожидаемого перлокутивного эффекта. При описании речевых актов приказа и просьбы следует учитывать специфические для каждого из них тон, каковым они в норме произносятся (ср. сочетания *умоляющий тон* и *приказной тон*). Сведения о том, какие невербальные единицы в норме сопутствуют данному речевому акту и какие из них обязательно должны сопутствовать ему, чтобы он был успешен, следует, по моему мнению, включить в будущий словарь речевых актов в качестве отдельной зоны каждой словарной статьи.[5]

2. Невербальные аналоги речевых актов. Невербальные коммуникативные акты

В акте устной коммуникации знаки естественного языка могут полностью замещаться знаками языка тела. Так, отвечая на вопрос *Тебе известно что-нибудь о наших планах?*, человек может сказать *нет*, а может **покачать головой**, ничего при этом не говоря. Встретив в многолюдном городе хорошего знакомого, мы можем приветствовать его словами или радостным восклицанием. Но, если наш знакомый находится далеко от нас, хотя и в зоне видимости, мы поступим иначе, поскольку правила этикета не позволяют громко кричать на улице. А именно, **встретившись глазами** с нашим знакомым, мы **улыбнёмся** и **помашем ему рукой**. Этот поведенческий комплекс является **невербальным аналогом речевого акта** приветствия.

Вот ещё несколько примеров невербальных аналогов речевых актов. Жест **погладить кого-л. по голове** (ср. жестовый фразеологизм *погладить по головке*) в одном из своих значений передаёт тот же смысл, что речевой акт похвалы. Разные жесты из класса **поклонов** являются аналогами таких речевых актов, как приветствие, прощание или благодарность. Омонимичные жесты, имеющие в русском языке номинацию *кивнуть*, замещают речевые акты приветствия и согласия (ср. *приветствовать кивком* и *кивнуть в знак согласия*). А речевому акту, сопровождаемому высказываниями *Я проиграл* или *Признаю своё поражение*, в языке жестов соответствует совокупность телесных знаков разной природы. Это жесты **опустить руки** и **склонить голову**, это опустошённый, холодный взгляд (часто тоже устремлённый вниз), это сдержанная мимика (люди стараются не проявлять свои чувства), это также смиренная, несколько сгорбленная поза и медленная, тяжёлая походка навстречу победителю. Побеждённые могут нести в руках белый флаг — предметный знак, символ сдачи.

Для некоторых речевых актов в качестве невербального аналога может использоваться нулевой знак **молчания**[6]. Рассмотрим следующую ситуацию: лектор обращается к аудитории со словами *Есть ли вопросы?…Ладно. Если вопросов нет, то мы перейдём к следующей теме*. Здесь молчание слушателей в ответ на инициирующую реплику *Есть ли вопросы?* является стандартным способом выражения смысла ‘вопросов нет’. Вербальная реакция слушателей в такой ситуации (например, если бы все хором ответили *Нет!*) могла бы показаться лектору по меньшей мере странной. В диалогах другого рода молчание собеседника в ответ на вопрос может быть интерпретировано спрашивающим как отсутствие ответа. И в этом случае произошёл бы коммуникативный провал, поскольку ответная реплика является иллокутивно вынужденной реакцией на вопрос[7]. Партнёр по диалогу, задавший вопрос, может упрекнуть собеседника в том, что тот *уходит* или *уклоняется от ответа*, а может повторить свой вопрос, как в следующем примере: —

---

[4] Механизмам взаимодействия вербальных и невербальных знаков в диалоге посвящены работы Крейдлин 2007, Крейдлин 2008.

[5] Замечу, что в Словаре языка русских жестов имеется зона «Звуковое и речевое сопровождение жеста», где приводятся те звуковые последовательности, которые с необходимостью или сочень высокой степенью обязательности воспроизводятся вместе с данным жестом. Подробнее об этом см. в СЯРЖ 2001.

[6] Подробно о знаковом молчании см. книгу Крейдлин 2005, с. 26–30.

[7] Об иллокутивно вынужденных репликах (речевых актах) см. статью Баранов, Крейдлин 1992.

Здесь нет Бима? — спросил Толик. — … — Не было его тут? — переспросил Алеша (Г. Троепольский. Белый Бим Чёрное ухо). Отмечу, что в приведённом примере коммуникативный акт ответа был бы успешен, даже если бы в нём присутствовал только невербальный, но материально выраженный компонент (например, жест **покачать головой**).

Невербальные аналоги речевых актов являются подклассом множества **невербальных коммуникативных актов (НКА)**. По аналогии с речевым актом, НКА можно определить как намеренное целенаправленное неречевое действие, совершаемое в соответствии с принципами и правилами телесного поведения, принятыми в данном обществе. НКА, в свою очередь, входят во множество **поведенческих актов**. К числу поведенческих актов относятся такие ненамеренные и нецеленаправленные действия, как **скакать на одной ножке**, **ходить из угла в угол**, манеры **поведения за столом** и др. В настоящей работе речь будет идти только о НКА и невербальных аналогах речевых актов, некоммуникативные поведенческие акты здесь рассматриваться не будут.

Между речевыми актами и их невербальными аналогами существуют достаточно прихотливые отношения. Так, неверно, что для всех речевых актов существуют невербальные аналоги. Например, речевые действия, обозначаемые глаголами *информировать, прогнозировать, ябедничать* не имеют стандартных соответствий в русском языке тела. Обратно, не всякий НКА может быть успешно замещён речевым актом. Так, одна из разновидностей **кивка**, называемая иногда *академический кивок*[8], является фатическим невербальным знаком, или жестом-регулятором[9]. Он исполняется для того, чтобы поддержать речь говорящего, и означает следующее: ‘жестикулирующий показывает, что он воспринимает то, что говорит адресат, и готов слушать его дальше’. В русском языке есть много слов, выполняющие фатическую функцию: слова *угу*, *ага*, *так*, *именно*, *конечно* и др., но их использование вместо академического кивка для официальной обстановки не характерно.

Отличие невербальных аналогов речевого акта от невербальных единиц, сопровождающих речевой акт, состоит в следующем: невербальные аналоги не являются обязательными компонентами речевого акта и могут замещать его в процессе общения. Поэтому в будущем словаре русских речевых актов должны быть предусмотрены как зона «невербальные единицы, сопровождающие данный речевой акт», так и зона «невербальные аналоги данного речевого акта»[10]. В них предположительно будут включены не только жесты, но и отдельные незнаковые движения — те, которые в определенных, четко очерченных контекстах обязательно означиваются.

Можно поставить и в каком-то смысле обратную задачу — задачу построения словаря русских невербальных коммуникативных актов. Перечислю некоторые типы информации, которые, как я полагаю, должны быть представлены в таком словаре. На настоящем этапе работы я не могу привести полный перечень всех словарных зон, а привожу лишь некоторые примеры.

3. Типы информации в словаре невербальных коммуникативных актов

(1) Вход (стандартная языковая номинация НКА);
(2) толкование НКА;
(3) типовые невербальные единицы, входящие в НКА (список единиц языка тела, используемые жестикулирующим при совершении данного НКА);
(4) сведения об участниках НКА (о жестикулирующем и адресате) и об их социальных статусах;
(5) типовое речевое сопровождение НКА (список единиц естественного языка, с высокой степенью обязательности употребляющихся вместе с данным НКА);
(6) типовые речевые аналоги НКА (список речевых актов, синонимичных данному НКА);
(7) типовые реакции адресата НКА. Они могут быть двух типов — вербальные и невербальные. В описание вербальных реакций входит список типовых речевых актов, вынуждаемых данным НКА, а в описание невербальных реакций — список типовых вынуждаемых НКА;
(8) правила, регулирующие взаимодействие участников НКА (в частности, регулирующие выбор жестикулирующим инициирующего НКА и выбор адресатом типовой реакции).

4. Невербальный коммуникативный акт «утешение»: материалы к построению статьи словаря невербальных коммуникативных актов

(1) Стандартная языковая номинация НКА: *утешение*
(2) Толкование НКА: *X утешает У-а посредством Q в P по S* = ‘Зная, что человек У испытывает некоторое плохое чувство Р, потому что произошло или с большой вероятностью может про-

[8] Описание этого жеста см. в СЯРЖ 2001, с. 60 — 61.

[9] О жестах-регуляторах см. в Крейдлин 2002.

[10] В СЯРЖ 2001, например, имеются зона жестовых аналогов и зона речевых аналогов описываемого жеста.

изойти некоторое плохое событие S, человек X хочет, чтобы У перестал чувствовать P, и для этого X совершает действие Q'.

(3) Типовые невербальные единицы, входящие в НКА.

В НКА утешения входят разные виды знаков языка тела.

Во-первых, это мануальные жесты утешения, в которых в качестве активного органа участвуют рука или руки жестикулирующего. В качестве пассивных органов в этих жестах могут выступать плечи адресата (ср. **положить руку на плечо**, **потрепать/похлопать по плечу**, **обнять за плечи/за плечо**), голова или волосы (**гладить по голове/по волосам**), лицо (**поцеловать**), руки (**взять за руку**, **пожать руку**) или тело в целом (**обнять**). Отмечу, что в очень многих жестах утешения участвуют плечи адресата, и это не случайно. Дело в том, что одна из функций плеч в русской культуре — переносить тяжёлый груз (ср. устойчивые выражения *гора с плеч свалилась*, *ему это не по плечу*)[11], а плохое чувство, которое испытывает адресат жеста, метафорически представляется как физическая тяжесть, лежащая у него на плечах.

Во-вторых, это выражение лица, а именно ласковая, ободряющая или успокаивающая **улыбка;** ср. *Так улыбаются плачущему ребенку — до свадьбы заживёт* (А. Генис. Темнота и тишина).

В-третьих, это исполненный сострадания знаковый **взгляд**, ср. *Лавиния смотрела на меня с ужасом и состраданием. За этот взгляд можно было и умереть! Она подошла и положила руку мне на плечо.* (Б. Окуджава. Путешествие дилетантов (Из записок отставного поручика Амирана Амилахвари).

В-четвёртых, это параязыковой знак — тихий, мягкий голос, несколько пониженный и приглушённый по сравнению с обычным голосом человека. В русском языке для него есть специальные названия — *утешающий голос* или *утешающий тон*. Ср. *Утешающим тоном старшей, очень ласково она стала говорить вещи, с детства знакомые и надоевшие Самгину*; *Сама она говорила мало, очень просто и всегда мягким, как бы утешающим тоном* (М. Горький. Жизнь Клима Самгина. Часть 2).

В-пятых, это знаковые и незнаковые телодвижения, приводящие к сокращению коммуникативной дистанции между собеседниками: **подойти к кому-л.**, **присесть рядом с кем-л.**, **наклониться к кому-л.** и т. д. Ср. *Я знал, чего он ждёт, я сразу же почувствовал, что этот упрямец <…> не встанет с места до тех пор, пока я не ободрю его. Прежде всего, надо попытаться успокоить старика. <…> Наклонившись к нему, я начал говорить* (С. Цвейг. Нетерпение сердца); *Пал Палыч подошел к маленькому плачущему человеку, встал подле него, положил руку на плечо* (О. Павлов. Карагандинские девятины, или Повесть последних дней).

В-шестых, это комплексные поведенческие акты. Если человек чем-то сильно расстроен и плачет навзрыд, собеседник может **предложить** ему **стакан воды**. Если же человек ведёт себя более спокойно и в состоянии говорить, то партнёр по диалогу слушает его, **демонстрируя внимание** и **сочувствие** позой, мимикой, жестами и взглядами. И такое поведение часто оказывает не меньший эффект, чем слова утешения. Ср. *Утешаю не словами, а тем, что слушаю, сочувствую* (В. Крупин. Выбранные места из дневников 70-х годов). Особую роль в таком поведении играет **молчание**, ср. *Всё это он отбарабанил впопыхах как по писаному, а потом замолчал и вдруг положил руку мне на плечо жестом сожаления* (А. Волос. Недвижимость).

Есть специфические жесты и манеры поведения взрослых в стереотипной коммуникативной ситуации утешения детей. Среди них жесты **взять на руки** (ср. выражение *взять на ручки*), **баюкать**, **целовать** и др. Ср. *На звезде, на планете — на моей планете по имени Земля, — плакал Маленький принц, и надо было его утешить. Я взял его на руки и стал баюкать* (А. де Сент-Экзюпери. Маленький принц)

Следует отметить, что при описании НКА утешения важно фиксировать не только телесное поведение жестикулирующего (утешителя), но и телесное поведение адресата НКА. Если человек чувствует что-то плохое, он может показать это собеседнику, причём не словами, а типичными для русской культуры формами поведения. Так, он может принять **грустную**, **задумчивую позу**, **низко опустить голову** и **закрыть лицо руками** (целью этих жестов является этикетное сокрытие проявлений эмоций на лице). Если чувства настолько сильные, что человек не может их контролировать, он **плачет** или даже **рыдает**. Все такие знаки являются симптоматическими, не коммуникативными, однако они служат окружающим сигналом того, что этот человек нуждается в утешении.

НКА утешения представляет собой частный случай коммуникации лицом к лицу (так называемой *face-to-face communication*). В коммуникации такого рода тела, головы и глаза участников в норме ориентированы друг на друга. Особенностью коммуникативного акта утешения является разная степень обязательности указанной ориентации для собеседников. Человек, погружённый в собственные переживания, может быть не в настроении вступать в диалог, даже если диалог уже начат; он может отрешённо слушать, **глядя вниз**, **перед собой**, куда-то **в сторону**, но не **на собеседника**. Такое поведение, впрочем, не нарушает условий успешности НКА утешения. Напротив, тот, кто утешает, не просто может, а должен **смотреть на собеседника**, **развернув к нему корпус**. Поведение человека, который ласковым тоном говорит добрые, утешительные слова

[11] Подробно о наивном представлении части тела «плечи» и её основных функциях см. в статье Крейдлин, Летучий 2006.

в сторону, **стоя спиной к адресату**, кажется странным и неестественным; НКА утешения в таком случае, скорей всего, будет неуспешным.

(4) Сведения об участниках НКА и об их социальных статусах.

Участниками НКА утешения могут быть мужчины, женщины и дети, представители разных профессий и разных социальных статусов. Ограничений на распределения между ними ролей (утешитель и утешаемый) нет, есть только частотные предпочтения. Например, ребёнок, конечно, может утешать взрослого, мужчину или женщину, не важно, однако гораздо чаще бывает так, что взрослые утешают детей.

Замечу, что степень психологической близости участников диалога и их социальные статусы влияют на то, какой из семиотических кодов (вербальный или невербальный) окажется преобладающим в НКА утешения. Если человек более низкого социального статуса утешает человека более высокого статуса (например, подчинённый утешает начальника), то преобладать будет естественно-языковой код, в противном случае оба кода равноправны.

(5 и 6) Типовое речевое сопровождение и типовые речевые аналоги НКА.

Есть несколько видов типовых высказываний, сопровождающих или замещающих НКА утешения. Среди них (а) устойчивые сочетания, цитаты, поговорки, ср. *До свадьбы заживёт!*; *То ли ещё будет!*; *Что было, то сплыло*; (б) языковые сочетания с глаголом в повелительном наклонении типа *Не бойся*; *Не робей*; *Не плачь*; *Не печалься*; *Не унывай*; *Мужайся!*; *Крепись!* (в) языковые единицы *Полно!*; *Будет!* и др. Следует отметить высказывания *Мужайся!* и *Крепись!*: они стереотипно адресованы мужчине (об этом говорит их внутренняя форма) и часто употребляются в ситуации извещения о смерти близкого человека, ср. *Подполковник подошёл ко мне и тихо говорит: «Мужайся, отец! Твой сын, капитан Соколов, убит сегодня на батарее. Пойдем со мной!»* (М. Шолохов. Судьба человека)

(7) Типовые реакции адресата НКА.

Типовые реакции адресата НКА утешения бывают двух видов: (а) продолжение или установление контакта с собеседником (и в результате, как правило, утешение) и (б) прерывание контакта, отказ от утешения. В первом случае человек **поднимает глаза**, **поднимает голову**, **поворачивается** к партнёру (то есть ориентирует своё тело на него, что, напомню, характерно для обычных ситуаций из разряда *face-to-face communication*). На его лице появляется **грустная** или **слабая** ответная **улыбка**, он старается сократить коммуникативную дистанцию до минимума, может **прильнуть** к тому, кто его утешает. Во втором случае человек ведёт себя противоположным образом: он **отворачивается** от собеседника, старается **избегать его взгляда**, иногда даже **отшатывается** от него и **уходит**.

(8) Правила, регулирующие выбор адресатом НКА той или иной типовой реакции.

На выбор адресатом определённого типа реакции влияет множество факторов. В частности, человек реагирует на смысл слов, обращённых к нему, а также на то, как произнесены слова утешения, какими жестами, тоном и мимикой они сопровождаются. Важно также, ожидает ли адресат НКА именно такого поведения от человека, который пытается его утешить. Ср. следующий фрагмент диалога:

Любовь Андреевна: Пожалейте меня, хороший, добрый человек.
Трофимов: Вы знаете, я сочувствую всей душой.
Любовь Андреевна: Но надо иначе, иначе это сказать…
(А. Чехов. Вишнёвый сад)

Мы не знаем в точности, как повёл себя Трофимов в описываемой ситуации, ограничился он только словами или сопроводил их какими-то жестами утешения. Очевидно лишь то, что его реакция не соответствовала ожиданиям Любови Андреевны. Возможно, она ждала более глубокого сопереживания её страданиям, но этой глубины её собеседник не выразил ни речью, ни поведением.

Есть и другие случаи, когда невербальные знаки утешения и особенно сопровождающие их слова могут вызвать более резкое раздражение или желание прервать контакт. Например, человек настолько не хочет мириться с происходящим, что его могут утешить только события или действия, существенно меняющие ситуацию. Ср. следующий пример: *Когда мы с Ольгой Бокшанской стали говорить Ляле какие-то слова утешения, она, сверкнув своими прекрасными глазами, вдруг сказала мне «Вот когда умрет ваш муж, тогда поймете!»* (С. Пилявская. Грустная книга). Ср. также *Не утешайте меня, мне слова не нужны, / Мне б отыскать тот ручей у янтарной сосны* (Ю. Визбор). Отмечу, что в последнем примере не случайно на первый план выходят именно слова утешения: в русской культуре стереотипно и отчетливо противопоставлены слово и дело.

Заключение

В настоящей работе был введен ряд понятий, которые могут использоваться в словарном описании основных русских речевых актов. Это понятия невербальных единиц, сопровождающих речевые акты, и невербальных аналогов речевых актов. Была сформулирована задача построения словаря основных русских невербальных коммуникативных актов, намечены структура и содержание такого словаря и приведен пример словарной статьи невербального коммуникативного акта утешения.

Литература

1. *Баранов, Крейдлин 2002* — Баранов А. Н., Крейдлин Г. Е. Иллокутивное вынуждение в структуре диалога // Вопросы языкознания, 1992, № 2. С. 84–99.
2. *Вежбицкая 1987* — Wierzbicka A. English Speech Acts Verbs: A semantic dictionary // Sydney etc.: Academic Press, 1987.
3. *Гловинская 1993* — Гловинская М. Я. Семантика глаголов речи с точки зрения теории речевых актов // Русский язык в его функционировании: Коммуникативно-прагматический аспект. М.: Наука, 1993.
4. *Крейдлин 2002* — Крейдлин Г.Е. Невербальная семиотика: Язык тела и естественный язык // М.: Новое литературное обозрение, 2002.
5. *Крейдлин 2005* — Крейдлин Г. Е. Мужчины и женщины в невербальной коммуникации // М.: Языки славянской культуры, 2005.
6. *Крейдлин 2007* — Крейдлин Г. Е. Механизмы взаимодействия невербальных и вербальных единиц в диалоге: II А. Дейктические жесты и их типы // Труды международной конференции «“Диалог 2007”: компьютерная лингвистика и интеллектуальные технологии». М., 2007. С. 300–327.
7. *Крейдлин 2008* — Крейдлин Г.Е. Механизмы взаимодействия невербальных и вербальных единиц в диалоге: II Б. Дейктические жесты и речевые акты // Компьютерная лингвистика и интеллектуальные технологии (по материалам ежегодной Международной конференции «Диалог» (Бекасово, 4 — 8 июня 2008 г.)). М., 2008. Вып. 7 (14). С. 248 — 253.
8. *Крейдлин, Летучий 2006* — Крейдлин Г. Е., Летучий А. Б. Части тела в русском языке и в невербальных семиотических кодах // Русский язык в научном освещении, 2006, № 12 (2). С. 80–115.
9. *Мосс 1935/1936* — Мосс М. Техники тела // Мосс М. Общества. Обмен. Личность: Труды по социальной антропологии. М.: Восточная литература, 1996 (ориг. изд. — 1935). С. 242–263.
10. *НОСС 2003* — Новый объяснительный словарь синонимов русского языка (2-ое издание) // М.: Языки славянской культуры, 2003.
11. *СЯРЖ 2001* — Григорьева С. А., Григорьев Н. В., Крейдлин Г. Е. Словарь языка русских жестов // М.; Вена: Языки русской культуры; Венский славистический альманах, 2001.
12. *Шатуновский 2000* — Шатуновский И. Б. Речевые акты разрешения и запрещения в русском языке // Логический анализ языка: Языки этики. М.: Языки русской культуры, 2000. С. 319–324.

Дискурсивные маркеры в структуре устного рассказа: опыт корпусного исследования[1]

The role of discourse markers in local discourse structure: a corpus study

Подлесская В. И. (podlesskaya@ocrus.ru)
Российский государственный гуманитарный университет

Кибрик А. А. (aakibrik@gmail.com)
Институт языкознания РАН

На материале корпуса с просодической разметкой исследуется характер интеграции дискурсивных маркеров в иерархическую, линейную и просодическую структуру устного рассказа. Разграничиваются случаи автономного употребления дискурсивных маркеров и случаи, когда маркер встроен в объемлющую пропозициональную единицу.

1. Введение

В данной работе исследуется характер интеграции в иерархическую, линейную и просодическую структуру устного рассказа дискурсивных маркеров — незнаменательных слов или словосочетаний, регулирующих дискурсивный процесс между говорящим и адресатом (см. Schiffrin 1987, Fraser 1999, Schouroup 1999, Баранов и др. 1993, Киселева, Пайар (ред.) 1998, Дараган 2000). Исследование опирается на данные устного корпуса «Рассказы о сновидениях». Одновременно с обсуждением особенностей поведения этого класса слов мы сформулируем определенные конвенции, которые мы предлагаем использовать для его отображения в дискурсивной транскрипции — системе последовательной графической репрезентации устной речи. Работа является частью более объемного корпусного проекта, с результатами которого можно ознакомиться в книге Кибрик и Подлесская (ред.) 2009.

2. Дискурсивные маркеры и элементарные дискурсивные единицы

В используемой нами системе представления локальной структуры дискурса центральную роль играет понятие элементарной дискурсивной единицы. Элементарная дискурсивная единица (ЭДЕ) — это квант устного дискурса, минимальный шаг, при помощи которого говорящий продвигает дискурс вперед. Понятие ЭДЕ подробно обосновывается в книге Кибрик и Подлесская (ред.) 2009. Здесь достаточно отметить, что ЭДЕ идентифицируются в первую очередь на просодических основаниях — при помощи таких критериев, как наличие единого тонального контура, одного главного акцента, темпового и громкостного паттернов, а также паузации. С семантико-синтаксической точки зрения прототипическая ЭДЕ представляет собой клаузу, но в особых (и нередких) случаях ЭДЕ может быть по объему как меньше, так и больше клаузы. В транскриптах устного дискурса, которые читатель увидит в примерах ниже, ЭДЕ графически изображаются как строки. Самый своеобразный тип ЭДЕ, меньших, чем клауза — это регулято́рные ЭДЕ (термин из книги Chafe 1994: 63ff.). Функционально регуляторные ЭДЕ отличаются от всех других тем, что они не несут пропозициональной информации и состоят из дискурсивных маркеров, функции которых находятся в сфере организации и регулирования дискурсивного потока. Так, примером регуляторных ЭДЕ (хотя и нечастым в корпусе «Рассказы о сновидениях») являются так называемые «актуализаторы», например да/нет; см. исследование этого феномена в работе Земская (ред.) 1973: 349–365. Одна из функций этого класса дискурсив-

[1] Работа выполнена при поддержке гранта РГНФ 08-04-00165а.

ных маркеров — эшелонировать поток речи, периодически запрашивая у адресата подтверждение, что соответствующая информация поступила своевременно и понята правильно:

(1) 017z[2]

7.1 4. (*Ну у нас в де-е= в \деревне /дом такой,
9.0 5. /да ,
9.2 6. ··(0.3) {ЧМОКАНЬЕ 0.2}
··(0.2) вот /здесь \дверь,
10.5 7. ра= || раньше была на /улицу,
12.0 8. там \ступеньки.

В предлагаемой нами системе дискурсивной транскрипции, дискурсивные маркеры с регуляторной функцией отделяются в особую строку, т. е. считаются отдельной ЭДЕ, если они акцентированы и не расположены внутри синтагмы с пропозициональным значением. Оба эти условия являются необходимыми: если хотя бы одно из них не выполняется (т.е. маркер не акцентирован и/или расположен внутри синтагмы), он не считается отдельной — регуляторной — ЭДЕ. В следующем примере в строке 17 дискурсивный маркер значит не имеет акцента и расположен внутри синтагмы с пропозициональным значением, в данном случае —между подлежащим и сказуемым клаузы, объединенной единым просодическим контуром; поэтому маркер значит не выделяется в особую ЭДЕ. И напротив, строки 21 и 22 — это регуляторные ЭДЕ ну значит и в целом, включающие акцентированные дискурсивные маркеры:

(2) 091n

24.6 17. Я значит /обозлилась,
25.5 18. и \кричу:
26.2 19. «\Зачем вы его /кормите там?,
27.4 20. туда \утягиваете?».
28.5 21. ····(1.6) /\Ну значит ,
30.5 22. в \целом,
30.9 23. такая \тётька на меня /посмотрела,

Вместе с тем, сформулированные выше два условия не являются достаточными для того, чтобы дискурсивный маркер мог сформировать регуляторную ЭДЕ. Для этого требуется еще, чтобы между ним и соседними синтагмами имелись просодические симптомы границы ЭДЕ: начало движения тоновой кривой с уровня ресета, относительное изменение темпа и громкости и ряд других. Если транскрайбер сомневается в наличии просодического «шва», то действует дефолтное правило не отделять дискурсивный маркер.

3. Автономизация дискурсивных маркеров

Дискурсивные маркеры в разной степени склонны к автономизации, т. е. к формированию отдельных регуляторных ЭДЕ (см. наблюдения Н. Н. Розановой (1983) о так называемой «динамической неустойчивости» несамостоятельных слов в разговорной речи и сходные рассуждения в Francis, Hunston 1992:159 о том, что просодическая автономность слов типа well в английском языке может в существенной степени варьировать).

Так, например, вариативным поведением обладают дискурсивные маркеры, сигнализирующие об эмоциональной реакции говорящего на неожиданно возникшие трудности вербализации или обнаруженную в своей речи ошибку, нуждающуюся в исправлении. Такими маркерами являются, например, ой, фу и ряд других единиц, в основном, междометийного характера. Слова этого класса могут формировать отдельную регуляторную ЭДЕ, отделенную от соседних пропозициональных ЭДЕ выраженными просодическими границами — как в примере (3), но могут и встраиваться внутрь пропозициональной ЭДЕ, утрачивая просодическую автономность, —как в примере (4):

(3) 022z

31.2 12. ····(1.4) ээ(0.2)
Он \пошёл,
33.6 13. ···(0.5)
ээ(0.2) \о-ой-й,

Междометие произносит приблизительно как эйхь и слегка в нос, выражая досаду по поводу возникшей трудности вербализации.

34.6 14 ··(0.4) я-а уже
/пошёл *куда-то*,
36.2 15. *я забыл* *отошёл*,

(4) 009z

29.2 12. ···(0.8) И ··(0.3)
с-сс ····(1.2) когда
я /вышла,

После союза и в паузе шумный вдох, а затем с-сс — имплозивный свист междометийного характера — как сигнал затруднения.

Продемонстрированные выше маркеры эмоциональной реакции на речевой сбой, актуализаторы типа да/нет, маркеры перехода к обобщению типа в целом, представлены в корпусе единичными случаями. Тем самым, они позволя-

[2] Здесь и далее в заголовке примера указывается номер того рассказа в корпусе «Рассказы о сновидениях», из которого взят данный пример. В начале каждой строки примера указывается время ее начала в секундах (отсчитывая от абсолютного начала рассказа), а затем номер данной ЭДЕ в рамках рассказа. Числа в скобках обозначают длительность пауз в секундах. При необходимости строка может сопровождаться неформальным комментарием транскрайбера, для этого используется четвертая, дополнительная колонка транскрипта. О других конвенциях дискурсивной транскрипции см. Кибрик и Подлесская (ред.) 2009.

ют охарактеризовать феномен относительной автономизации лишь с качественной стороны. Для того чтобы дать представление о количественном распределении автономизированных и неавтномизированных дискурсивных маркеров, в следующих разделах мы обратимся к двум наиболее частотным дискурсивным маркерам в корпусе —частицам вот и ну.

4. Вариативность автономизации маркера вот

Самым частотным дискурсивным маркером в нашем корпусе является вот (см. Кобозева 2007 о значениях этой частицы). Всего в корпусе имеется 437 вхождений слова вот. Из них 257 вхождений являются неакцентированными, см следующий пример, где в составе одной ЭДЕ имеется три неакцентированных вот:

(5) 127n

17.0 9. ···(0.5) И /меня при этом /провожала вот /вся вот эта вот моя \семейка.

Такие употребления дискурсивного маркера вот не связаны с образованием регуляторных ЭДЕ. Далее, в 8 случаях слово вот используется не как дискурсивный маркер, а как указательная частица. В этих употреблениях вот входит в состав клаузальной ЭДЕ, а также является ее акцентным центром — на нем располагается несущий (главный фразовый) акцент:

(6) 077n

39.6 26. ····(1.4) /\Вот что я чувствовал.

Разумеется, подобное употребление вот по просодическим и синтаксическим причинам не может быть выделено из состава ЭДЕ и, тем самым, не образует отдельной регуляторной ЭДЕ (хотя и удовлетворяет сформулированным выше необходимым условиям —находится в левой периферии синтагмы и является акцентированным).

В оставшихся 172 вхождениях вот мы имеем дело с регуляторной ЭДЕ, состоящей либо из единичного вот (153 случая), либо реже (в 19 случаях) из вот в сочетании с ну. За редкими исключениями вот в этом качестве реализуется с нисходящим, ровным или нисходяще-ровным акцентом (158 случаев из 172). Регуляторные ЭДЕ, основанные на дискурсивном маркере вот, представляют собой самую большую группу регуляторных ЭДЕ в корпусе. Из 172 случаев регуляторного, т. е. автономного, употребления этого маркера он дважды употреблен в особой функции —как средство передачи угрозы и осуждения в чужой речи:

(7) 128n

47.9 39. ···(0.7) \тоже на меня ’ /посмотрела ,
49.7 40. «/\Вот!
49.9 41. Я тебя /**–в-выгоню-у** и-из-зз ··(0.2) этой из ш= ··(0.2) ’из ··(0.2) ’ /школы!»,

Изысканность этого употребления состоит в том, что цитирование происходит в форме прямой речи, однако сам маркер не мог принадлежать исходной реплике — при непосредственном выражении угрозы в диалогическом режиме вот употреблено быть не может. Таким образом, маркер вот, подобно более нейтральным маркерам чужой речи типа мол, может служить средством перевода прямой речи в «несобственно-прямую». Заметим при этом, что мол — в отличие от вот — не способно к автономизации, эта частица никогда не акцентируется и всегда встроена в пропозициональную ЭДЕ. Еще одно важное обстоятельство состоит в том, что средством передачи угрозы и осуждения в чужой речи является даже не сам маркер вот, а реализуемая на нем особая просодическая фигура с восходяще-нисходящим тоном. Эта же фигура, с тем же значением, может реализоваться и на ряде других маркеров, например, на частице а, тоже способной к автономному употреблению. В следующем примере эти два маркера образуют две последовательных регуляторных ЭДЕ:

(8) 128n

20.5 19. ···(0.7) ’И \каждое н-на \меня=:
22.2 20. ··(0.1) «/\’А-а!
22.7 21. /\Вот!
23.0 22. Ты /\двоечница!
23.9 23. \С-смотри на своих оценки!»

Во всех остальных случаях, за исключением вышеупомянутых, функцию автономно употребленного маркера вот можно определить как делимитативную — указание на то, что некоторый дискурсивный фрагмент (ЭДЕ или группа ЭДЕ) завершен, и говорящий переходит к следующему (Дараган 2000, 2003, Прокуровская 1977, Васильева 1964). Так, в примере ниже вот маркирует возвращение к основной линии повествования после отступления, или вставки (строки 15–17 в скобках):

(9) 054z

23.8 14. ····(1.4) /потом подъехали к \/пляжу,
26.3 15. ···(0.5) (/Я вообще плавать не \уме-ею.
28.1 16. ··(0.4) Не \умела тогда,
29.3 17. *когда мне это \снилось*.)
30.2 18. ····(1.0) \вот,
31.5 19. и-и’ ····(1.1) /я почему-то \поплыла.

Делимитативная функция единиц этой группы проявляется, в частности, и в том, что они могут становиться локусом хезитации. Косвенно об этом

свидетельствует тот факт, что в 163 случаях из 172 употреблений вот и ну вот в качестве регуляторных ЭДЕ им предшествует пауза, притом, что в половине случаев пограничная пауза **после** этих регуляторных ЭДЕ отсутствует. Кроме того, в 62 случаях имеется удлинение ударной гласной (во-от) и/или начальной согласной (в-вот), скорее всего также хезитативного характера:

(10) 050z

54.8 30. ····(1.4) Но’ ··(0.3) они всё же /доехали,
57.9 31. /я’ ··(0.2) тоже оказалась на первом /этаже ,
59.9 32. ····(1.4) \в-вот,
61.8 33. ··(0.3) и-и’ ····(1.0) \встретили мы этих ··(0.1) двух /девочек ,

(11) 100n

26.5 19. ··(0.3) /\Я-аа ··(0.1) с= || \/поняла-а ,
28.1 20. ··(0.2) что это \па-апа-а.
29.4 21. ··(0.4) Ну \во-от.
30.4 22. ··(0.3) ’О= || /’открыла ему \/\две-ерь.

5. Вариативность автономизации маркера ну. Сравнение ну и вот

Вторым по частотности дискурсивным маркером, формирующим регуляторные ЭДЕ в нашем корпусе, является ну. Эта частица имеет весьма широкий спектр значений (см., например, Баранов, Кобозева 1988, Дараган 2000, 2002, 2003, Шмелев 2004, 2005 и др.), инвариантом которых, по мнению А. Д. Шмелева (2004), является «вынужденное говорение», т. е. говорящий сигнализирует о том, что его высказывание мотивировано условиями конкретной речевой ситуации, известными адресату речи, и выражает установку на взаимопонимание и кооперацию. Частица ну в нашем корпусе, как и частица вот, может выступать и в составе пропозициональных ЭДЕ, и формировать отдельную регуляторную ЭДЕ, однако она гораздо в меньшей степени, чем частица вот, склонна к автономизации. В таблице 1 показано сравнительное распределение акцентированных и неакцентированных употреблений этих частиц в составе ЭДЕ.

Таблица 1. Употребления частиц вот и ну в корпусе

	Без акцента	**С акцентом**			**Всего**
		Отдельная ЭДЕ	Внутри другой ЭДЕ	Всего	
вот	257	172	8	180	457
ну	200	30	57	87	287

Как видим, доля акцентированных употреблений от общего числа вхождений сопоставима для вот (39%) и для ну (30%), однако отдельную регуляторную ЭДЕ формируют лишь 34% акцентированных употреблений ну (30 из 87), тогда как для вот эта цифра составляет 95% (172 из 180).

Частица ну в качестве отдельной регуляторной ЭДЕ, как правило, реализуется с нисходящим, восходяще-нисходящим или ровно-нисходящим акцентом (24 случая из 30) и сигнализирует о хезитации и поиске вербальной формы («ну поиска» согласно работе Баранов, Кобозева 1988). Например:

(12) 115n

26.2 21. ··(0.4) папы /нет,
27.2 22. а ==
27.5 23. ···(0.5) \ну-у ,
28.4 24. ··(0.3) мне почему-то казалось
29.5 25. что я должна всё /решить,

В качестве неавтономного маркера поиска употребляется в нашем корпусе и большинство акцентированных ну в составе пропозициональной ЭДЕ, если они размещены не в ее абсолютном начале. В этой позиции они также реализуются преимущественно с нисходящим, восходяще-нисходящим или ровно-нисходящим акцентом (26 случаев из 28). Такие ну обычно соседствуют с другими симптомами речевых затруднений — заполненными паузами, повторами, самоисправлениями:

(13) 023z

2.0 3. и /подружился̱ ка̱к бу̱дто̱ с одн-ной де= ··(0.1) –\ну-у ··(0.3) –\девочкой.

И, наконец, особый класс — и по просодии, и по функции — составляют случаи употребления акцентированной частицы ну в абсолютном начале синтагмы, но без просодического шва после частицы. В таких контекстах частица ну является сигналом «естественного вывода» с апелляцией к известным слушателю обстоятельствам, ср. упомянутые выше работы А. Д. Шмелева (2004, 2005) и Ю. В. Дараган (2000, 2002, 2003), и реализуется преимущественно с восходящим, восходяще-ровным или ровным акцентом. В таких случаях, согласно принятой нами конвенции, частица не выделяется в особую ЭДЕ, т. е. признается неавтономной:

(14) 126n

16.1 12. и он начал ещё –больше меня \унижать.
18.0 13. ····(1.5) То есть ещё **ху́же**!
20.8 14. ····(1.2) –Вот,
22.3 15. /ну я естественно вот быстро /ушла ,

(15) 119n

32.6 22. ···(0.8) ''(0.4) У меня в /школе вот сейчас мне дали классное \руководство.
35.8 23. ··(0.4) Мне дети /завопили,
37.1 24. ··(0.4) «Нам Алёну /\И-игоревну!
38.4 25. Нам Алёну /\И-игоревну!»,
39.3 26. ···(0.7) /ну –пришлось...

Отметим, что в связи с тем, что наш корпус ограничен жанром личного рассказа, в нем не встретились такие автономные употребления частицы ну, как, например, ответные реплики согласия /подтверждения (Ты, оказывается, уже все сделал! — Ну! [= А ты как думал!]), более характерные для диалогического дискурса.

6. Заключение

Итак, дискурсивные маркеры в разной степени склонны к автономизации, т. е. к выделению в особую, регуляторную, ЭДЕ. Вариативное поведение, сходное с «ну поиска», демонстрируют, например, такие показатели речевых затруднений, как маркеры препаративной подстановки типа это, это самое, как его (см. о них подробнее Подлесская, Кибрик 2006, Podlesskaya 2006, 2007, In press). Они могут выступать и в составе пропозициональной ЭДЕ, как в примере (16), и автономно, как в примере (17):

(16) 070n

64.7 23. ····(2.8) 'И-и ····(3.0) онh ····(1.4) \это ··(0.3) каким-то ···(0.5) 'образом убежал в \Монголию.

(17) 016z

0.0 1. /Мы с мамой поех= ==
0.9 2. ···(0.9) \Это,
2.0 3. /мы с мамой поехали в А'= ==
3.3 4. ···(0.5) /Мне приснился /сон,
4.6 5. будто мы с /\мамой поехали в \Америку.

С другой стороны, имеются маркеры, которые практически всегда выступают в безударном варианте и не формируют регуляторных ЭДЕ. Такова частица там — «маркер несущественной детали» (Шмелев 2007). Под номером (19) ниже приведен полностью короткий рассказ о сновидении, в котором представлено восемь безударных вхождений этой частицы (девятое вхождение, в строке 9 — это не частица, а акцентированное указательное местоимение):

(18) 019z

0.0 1. /\Сон снился ,
0.5 2. как яw ····(1.3) в \классе был,
2.9 3. ··(0.2) там и про класс про /–школу...
4.3 4. про' || ··(0.3) про /–друзей...
5.5 5 ···(0.7) /Ну я там /помню,
6.9 6. ··(0.2) что-о' я там ···(0.9) ну с моим /–другом ···(0.5) \разговаривал я'...
11.0 7. ····(2.9) Ну –вот,
14.2 8. /потом || ····(3.4) ну –потом мы там ····(3.0) чего-то на \физкультуре были,
22.9 9. \играли там.
23.5 10. ····(1.4) Ну \вот ··(0.1) чего-то'.
25.8 11. ····(2.4) /Ну мы там вот \так:
29.1 12. ···(0.7) \школа там,
30.4 13. я /класс ' || ··(0.1) –класс <ну вь=> ··(0.1) <НРЗБЗ> видел вот \там...
32.6 14. ···(0.7) там /–друзе-ей...

Таким образом, можно заключить, что принципиальная способность или неспособность к автномизации является индивидуальным, словарно закрепленным свойством того или иного дискурсивного маркера, тогда как реализация этого свойства определяется конкретными контекстными условиями. Поведение незнаменательной лексики в спонтанной речи — это интересный, но пока недостаточно разработанный сюжет дискурсивных исследований, и мы надеемся, что просодически размеченный корпус может сыграть важную роль в изучении функций такого рода единиц.

Литература

1. *Баранов А. Н., Кобозева И. М.* Модальные частицы в ответах на вопрос // Прагматика и проблемы интенсиональности. М.: ИВАН СССР, 1988.
2. *Баранов А. Н., Плунгян В. А., Рахилина Е. В.* Путеводитель по дискурсивным словам русского языка. М.: Помовский и партнеры, 1993.
3. *Васильева А. Н.* Частицы разговорной речи. М.: Изд-во Московского университета, 1964.
4. *Дараган Ю. В.* Функции слов-«паразитов» в русской спонтанной речи // Труды Международного семинара «Диалог'2000» по компьютерной лингвистике и ее приложениям. Т. 1. Теоретические проблемы, 2000. С. 67–73.
5. *Дараган Ю. В.* Риторическая структура текста и маркеры порождения речи // Труды Международного семинара «Диалог'2002» по компьютерной лингвистике и ее приложениям. Т. 1. Теоретические проблемы, 2002. С. 114–127.

6. *Дараган Ю. В.* Паразитизм или симбиоз: механизм преодоления коммуникативных сбоев и обслуживающие его вербальные средства // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2003». Протвино, 11–16 июня, 2003 г. М.: Наука, 2003. С.166–178.
7. *Земская Е. А. (ред.)* Русская разговорная речь. М.: Наука, 1973.
8. *Кибрик А. А., Подлесская В. И. (ред.)* Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК, 2009.
9. *Киселева К. Л., Пайар Д. (ред.)* Дискурсивные слова русского языка. М.: Метатекст,1998
10. *Кобозева И. М.* Полисемия дискурсивных слов и попытка ее разрешения в контексте предложения (на примере слова вот) // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007». Бекасово, 30 мая — 3 июня 2007 г., 2007. С. 250–255.
11. *Подлесская В. И., Кибрик А. А.* Коррекция в устной монологической речи по данным корпусного исследования // Русский язык в научном освещении, 12-2, 2006. С. 7–55.
12. *Прокуровская Н. А.* Незнаменательная лексика в русской разговорной речи (состав и функции). Автореф. … дис. канд. филол. наук. Саратов, 1977.
13. *Розанова Н. Н.* Суперсегментная фонетика // Е. А. Земская (ред.) Русская разговорная речь. Фонетика. Морфология. Лексика. Жест. М.: Наука, 1983. С. 5–79.
14. *Шмелев А. Д.* «Заполнители пауз» как коммуникативные маркеры // Жанр интервью: Особенности русской устной речи в Финляндии и Санкт-Петербурге. Tampere (= Slavica Tamperensia VI), 2004. С.205–222.
15. *Шмелев А. Д.* Показатели хезитации в устной русской речи // Язык. Личность. Текст. Сб. ст. к 70-летию Т. М. Николаевой. М.: ЯСК, 2005. С. 518–529.
16. *Шмелев А. Д.* Частица там как маркер «несущественной детали» // Язык как материя смысла. Сб. ст. к 90-летию академика Н. Ю. Шведовой. М.: ИЦ Азбуковник, 2007. С. 208–218.
17. *Chafe W.* Discourse, consciousness, and time. Chicago: University of Chicago Press, 1994.
18. *Francis G., Hunston F.* Analysing everyday conversation // M. Coulthard (ed.) Advances in spoken discourse analysis. L.: Routledge, 1992. P. 123–161.
19. *Fraser B.* What are discourse markers? Journal of Pragmatics, 31, 1999. P. 931–952.
20. *Podlesskaya V.* Speech disfluencies as a window on syntax: Demonstratives as hesitation markers // Syntax of the World Languages (SWL2). September 14–17, 2006, Lancaster University, UK, 2006. P. 42.
21. *Podlesskaya V. I.* Parameters for typological variation of placeholders // 10th International Pragmatics Conference. Goteborg, Sweden, 8–13 July 2007. Abstracts. International Pragmatics Association, 2007. P. 26.
22. *Podlesskaya V. I.* In press. Parameters for typological variation of placeholders // N. Amiridze, B. Davis (eds.) Fillers in discourse and grammar. Amsterdam: John Benjamins.
23. *Schiffrin D.* Discourse markers. Cambridge: Cambridge University Press, 1987.
24. *Schouroup L.* Discourse markers // Lingua, 107, 1999. P. 227–265.

Конкорданс к текстам Ломоносова — концепция и реализация

Lomonosov concordance — concept and implementation[1]

Поляков А. Е. (pollex@mail.ru)
НТЦ «Информрегистр»

Бергельсон М. Б. (mirabergelson@gmail.com)
МГУ имени М. В. Ломоносова

Пильщиков И. А. (pilshch@yandex.ru)
ИМК МГУ имени М. В. Ломоносова

В докладе уточняются понятия и термины, связанные с разработкой полного электронного Конкорданса к текстам Ломоносова и обсуждаются практические решения, необходимые для реализации этого лексикографического продукта. Конкорданс строится на основе корпуса авторских текстов, снабженных структурной, филологической и грамматической разметкой. Описывается технология построения корпуса и конкорданса, принципы разметки корпуса, структура словарной статьи конкорданса, а также возможности его применения для лингвистических исследований.

0. Введение.

0.1. Цели и принципы

Конкорданс к произведениями и письмам М. В. Ломоносова, над которым работают участники настоящего проекта, строится на основе электронного корпуса текстов М. В. Ломоносова, представляющего собой филологически корректную цифровую версию академического Полного собрания сочинений и писем Ломоносова в 11-ти томах (1950–1983) и ряда дополнительных изданий. Цель конкорданса — представить авторское словоупотребление во всей его широте и во всей полноте его языковой специфики.

Конкорданс к текстам Ломоносова является частью проекта по созданию электронного научного издания «Ломоносов», которое позволит предоставить широкому кругу пользователей программно-информационную среду для изучения литературного и научного наследия, языка и биографии Ломоносова. Проект предусматривает создание открытого интернет-ресурса, который будет включать в себя:

1) корпус текстов Ломоносова, построенный на основе наиболее авторитетных изданий;
2) биографические, литературно-критические и историко-научные работы о Ломоносове;
3) полный алфавитно-частотный конкорданс к текстам Ломоносова.

Ломоносовский конкорданс строится на основе принципиально новой методологии и технологии подготовки, отвечающей современному уровню филологической науки. В основе его лежит корпус филологически выверенных авторских текстов, снабженных богатой структурной, филологической и грамматической разметкой.

0.2. Основные определения

Ключевыми для данного проекта являются понятия **конкорданса**, **корпуса**, а также их уточнения, как-то — электронный (сетевой) конкорданс,

---

[1] В докладе изложены результаты работы трех участников проекта «Электронное научное издание “Ломоносов”: корпус текстов, справочная информация, алфавитно-частотный конкорданс» (грант РГНФ 08-04-12120в; руководитель — чл.-корр. РАН В. А. Виноградов). Первоначально докладчики предполагали выступить с двумя сообщениями на взаимосвязанные темы, но по предложению оргкомитета объединили оба доклада в один.

полный корпус, дифференциальные словари. Корректное определение этих ключевых понятий возможно только на фоне анализа соответствующей концептуальной области.

Анализ современных лексикографических продуктов и аналитических работ по авторской и — шире — общей лексикографии показывает, что термин «словарь» представляет собой родовой обозначение, соответствующее общеязыковому употреблению слова. Фактически, в широком смысле слова, это некоторым образом упорядоченный список символов, выбранный из некоторого множества текстов.

Если говорить более подробно, то **словарь** в широком смысле подразумевает определенный способ выборки словника, множество текстов, из которого делается эта выборка, характеристику вокабул, дополнительную информацию, которая сопровождает каждую вокабулу, возможности филиации значений и их толкований. Таким образом, ключевые различия проявляются как на уровне микроструктуры — того, что представляет собой единица описания плюс структура словарной статьи, — так и на уровне макроструктуры — пространства текстов, являющегося базой для составления словаря. На основе описания этих параметров можно говорить и о принципиальном различии между конкордансами и собственно словарями (в узком смысле слова); оно состоит в принципах отбора, подачи и описания лексических единиц, в постановке различных целей описания и следовании им.

Минималистское определение конкорданса с необходимостью включает представление о корпусе (наборе текстов, отобранных и препарированных с определенной целью). Тогда **конкорданс** к некоторому корпусу — это список словоупотреблений (элементов корпуса) с отсылкой ко всем контекстам. Противопоставление между словарями и конкордансами идет сразу по нескольким шкалам — репрезентативности, ориентации на инвариант, смыслового или грамматического анализа. Можно сформулировать следующие противопоставления:

- Словарный подход к описанию лексики ориентирован на репрезентативность и тем самым на **нормативность**, а корпусный вариант (конкорданс) — на **исчерпывающее** описание.
- Словарь, анализируя различные употребления лексемы в разных значениях, стремится к нахождению **инварианта**, конкорданс — к **вариативности** и ставит своей первейшей задачей отразить все случаи употребления слова. Поэтому представление слова в конкордансе ставит во главу угла примеры (контексты словоупотреблений), а в словаре — словарную статью.
- На различном понимании термина **полнота** базируется еще одно принципиальное противопоставление словаря и конкорданса: полнота словаря определяется стремлением к исчерпывающему описанию значений, полнота конкорданса определяется исчерпывающим характером описания соответствующего корпуса.
- Из этого вытекает принципиальная необходимость грамматической (морфологической) информации в конкордансе, что помогает охарактеризовать и — если нужно — различить формы, и необходимость семантического анализа (описания или толкования значений) в словаре.

Итак, можно сформулировать представление об «идеальном» (прототипическом) словаре и прототипическом конкордансе. Словарный подход стремится выделить некоторую сущность и истолковать ее. Прототипический словарь — это нормативный толковый словарь, содержащий большое количество семантической информации, в частности, семантических и стилистических помет. Прототипический конкорданс обладает полнотой тезаурусного типа (словаря, в котором максимально полно представлены слова языка с примерами их употребления в тексте, что в полном объеме осуществимо, да и то с оговорками, лишь для мертвых языков), он является первой производной корпуса, «нарезанным» корпусом, так что в идеале совокупность примеров всех употреблений составляют корпус. Он обязан быть полным в отношении учета абсолютно всех словоупотреблений. От списка слов (словника) его отличает наличие морфологической характеристики словоупотреблений. В конкордансе морфологическое описание реализуется как лемматизация, позволяющая приписать каждой словоформе граммемную характеристику. Это, в свою очередь, позволяет различить омонимичные формы. Реальной такая постановка задачи и стремление хоть сколько-нибудь приблизиться к данному прототипу возможны только в электронной форме сетевого ресурса. Именно такой конкорданс должен быть реализован в данном проекте.

Из указанных различий между прототипическим словарем и прототипическим конкордансом следует и то, что именно является для каждого из них объектом и единицей описания. Для словаря это значение и словарная статья, для конкорданса — словоупотребление и корпус. Таким образом, различие между конкордансом и толковым словарем заключается в том, что конкорданс не предполагает установления структуры (филиации) значений регистрируемых слов и не обязательно включает толкование этих значений. Его полнота, в частности, требует, чтобы он был словарем регистрирующего типа. Задача Конкорданса состоит в том, чтобы отразить все возможные различия и особенности употреблений, так как заранее неизвестно, какие из них могут оказаться значащими и значимыми. В этом отношении Конкорданс представляет собой базу для созданий различного рода **дифференциальных**, в том числе и **толковых**, словарей.

1. Общефилологические аспекты создания Конкорданса

1.1. Выбор источника

Поскольку всякий конкорданс есть особого рода лингвистически препарированный указатель к конкретному корпусу текстов, перед составителями встает проблема отбора филологически корректных текстов данного автора. Для электронного корпуса последнее означает оптимальное соответствие выбранному печатному изданию, для печатного — соответствие тем задачам, которые ставит перед собой конкорданс.

Из имеющихся изданий Ломоносова 11-томное академическое издание (**ПСС**) в наибольшей степени пригодно для намеченных целей, однако по целому ряду параметров оказывается неудовлетворительным и оно. Академическое издание непоследовательно отражает ломоносовское правописание: для разных томов был выбран различный орфографический режим, что привело к неоднородности корпуса, положенного в основу Конкорданса. Однако остальные издания еще менее пригодны с текстологической и лингвистической точек зрения, а попытка исправить 11-томное издание стала бы попыткой подготовки нового критического издания; между тем такая задача явно выходит за рамки обсуждаемого проекта.

1.2. Произведение и текст. Проблема неустойчивости текста

Совокупность авторских текстов исторически делится на произведения. Произведением считается автономный текст, выделенность которого определена автором или (достаточно часто!) его редакторами. Текстуальный состав произведения не является диахронически неизменным: сам текст еще при жизни автора и по воле автора может варьироваться. Современные научные издания (и в том числе ПСС Ломоносова) стремятся представить текст с учетом его авторской вариативности. Альтернативные фрагменты текста (то есть фрагменты, **не** сосуществующие ни в одном его синхронном срезе) представлены либо как другие редакции произведения (связные тексты), либо как варианты (набор разночтений к основной либо иной редакции).

Для задач настоящего Конкорданса принято решение учитывать в **алфавитной** части все словоупотребления, зафиксированные в ПСС и в дополнительных источниках. Вопрос о том, как учитывать вариативность в **частотном** словаре, остается открытым. Ясно, что нужно дать пользователю возможность, как минимум, учитывать частоту употребления того или иного слова в основном корпусе текстов (то есть в наборе «окончательных редакций»). Другие редакции (связные тексты) можно принимать за отдельные произведения и давать две статистики: с учетом и без учета других редакций. Однако в целом ряде случаев полный текст другой редакции нам неизвестен. Абсолютную частотность слов, появляющихся в вариантах (разночтениях) можно было бы учитывать таким же образом, однако в таком случае неясно, к чему мы приравниваем общее количество слов в данном произведении.

1.3. Проблема неоднородности корпуса

Не всякая наблюдаемая в корпусе вариативность имеет авторскую природу. Серьезную проблему для словаря представляет неоднородность, созданная непоследовательностью или множественностью редакторских подходов, реализованных в используемом корпусе.

По отношению к ПСС это, в первую очередь, проблемы, связанные с орфографическим режимом представления разных произведений: как и в других изданиях послевоенного времени, ПСС модернизирует текст Ломоносова. Эта модернизация проходит 3 стадии: (1) замена символов (букв), отмененных реформой 1918 г. (ять, фита, ижица, и десятиричное), буквами современного алфавита (*е*, *ф*, *и*, *и* либо *й*); (2) замена современными морфемами морфем, отмененных реформой 1918 г. (флексии *-аго*, *-яго*; префиксы *из-*, *без-*, *воз-* etc. перед глухими консонантами; и др.); (3) иные замены начертаний слов, приближающие ломоносовское написание к современному (например, *горкий* VS *горький*, *не льзя* VS *нельзя* — как видим, сказанное относится не только к буквенному составу слов, но и к слитному/раздельному написанию, прописным/строчным буквам и т.д.).

Мало того, что модернизированные написания не соответствуют ломоносовским — в ПСС отсутствует единый режим орфографической модернизации. Так, в материалах по русской грамматике в примерах не проведена модернизация типа (1) и (2). Модернизация типа (3) не проведена в 8-м томе, содержащем поэтические и риторические произведения, но проведена в остальных томах. Например, в 7-м томе ПСС в «Предисловии о пользе книг церьковных» прилагательное *церьковный*, вопреки всем источникам текста, дано в твердом варианте (*церковный*), а в поэме о Петре Великом, напечатанной в 8-м томе, то же прилагательное выглядит как *церьковный* (в соответствии с последним прижизненным изданием 1761 г.).

Единственный выход из ситуации — не учитывать вариативности буквенного состава слов, слитного/раздельного написания и написания с прописной/строчной буквы при сведении словоформ в лек-

сему (вокабулу). Случаи типа слитного/раздельного написания *не* с глаголами и другими частями речи потребуют специального лингвистического анализа, а в отдельных случаях, возможно, и текстологической проверки ПСС. К сожалению, в рамках данного проекта такая проверка не может быть проведена тотально. Однако в тех случаях, когда мы можем с уверенностью исправить текст ПСС, мы должны сделать в Конкордансе соответствующую помету и учесть верное чтение.

2. Разметка

2.1. Определение и классификация видов разметки

Электронный текст имеет не линейную структуру, а включает несколько параллельных слоев информации. С одной стороны, текст состоит из языковых элементов различного уровня (слова, фразы, предложения), с другой стороны, он состоит из структурных сегментов различных типов (заголовки, сноски, ремарки, стихи, цитаты, таблицы, формулы, страницы). Обычно при построении корпусов учитывается только языковое членение текста, но мы считаем, что структурное членение является не менее важным. Многие структурные элементы текста имеют яркие языковые особенности, которые требуют специальной формы представления в корпусе.

Разметка — это расстановка в тексте документа специальных маркеров (тегов), которые эксплицируют «скрытые» элементы информации, присутствующие в тексте. В зависимости от типа этой информации можно выделить следующие виды разметки:

1) **Метатекстовая** разметка.

Включает параметры, характеризующие текст в целом, в частности:

- автор (фамилия, имя, отчество);
- название произведения (заголовок, подзаголовок, incipit);
- короткое имя (используется для цитирования в примерах);
- дата создания произведения;
- жанрово-тематический класс произведения; и т. д.

2) **Структурная** разметка.

Эксплицирует логическую структуру текста, в частности:

- деление текста на структурные элементы (части, главы, действия, явления, реплики);
- заголовки структурных элементов;
- сноски, ремарки;
- стихотворные элементы (строфы, стихи);
- таблицы, формулы, рисунки, и т.д.

Многие из этих элементов имеют яркие языковые особенности и требуют специальной разметки.

3) **Форматная** разметка.

Описывает параметры оформления текста, включая:

- параметры шрифта (размер, жирность, курсив, разрядка, верхние/нижние индексы);
- параметры абзаца (выравнивание, отступы, межстрочный интервал);
- полиграфические и декоративные элементы (колонтитулы, виньетки);
- номера страниц и стихов; и т.д.

Форматная разметка нужна для адекватного представления текста в электронной библиотеке, а для лингвистического анализа многие элементы этой разметки оказываются ненужными или, наоборот, недостаточно информативными. Например, шрифтовое выделение заголовка носит чисто декоративный характер, а шрифтовое выделение отдельных слов может означать самые разные сущности: цитата, пример употребления, формула, часть слова и т.д. Абзацные отступы и выравнивание могут маркировать структурные элементы текста: заголовок, подпись, стихотворная строка определенного размера и т.д. Следовательно, для анализа текста нельзя просто игнорировать оформление, но нужно дать ему семантическую интерпретацию и поставить соответствующий структурный тег.

4) **Грамматическая** разметка.

Описывает признаки, характеризующие конкретное словоупотребление, включая:

- лексема (словарная форма);
- грамматические признаки лексемы (часть речи, одушевленность, переходность);
- грамматические признаки словоформы (число, падеж, наклонение, время, лицо).

Параллельно с грамматической информацией, эта разметка эксплицирует членение текста на языковые элементы (токены) различного типа — предложения, слова, знаки препинания, цифры.

2.2. Формат и технология разметки

Формат разметки текстов для корпуса должен учитывать многоплановость текста и наличие в нем нескольких параллельных слоев информации — метатекстовой, структурной и собственно лингвистической. Формат должен быть открытым, компактным, расширяемым, он должен быть совместим с существующими форматами разметки и легко интегрироваться с программами обработки. Ключевыми свойствами здесь являются открытость и со-

гласованность формата на всех этапах обработки. Именно это позволяет связать все операции в единую технологическую цепочку, на входе которой находится неразмеченный текст, а на выходе — размеченный корпус, из которого автоматически получается конкорданс.

Формат разметки текстов для корпуса был разработан на базе существующих стандартов представления текстовой информации для интернета (HTML+CSS) и стандартов кодирования лингвистической разметки для корпусов (TEI, XCES). После детального анализа существующих стандартов разметки корпусов был сделан вывод, что универсальные стандарты типа TEI или XCES являются слишком сложными, избыточными и неудобными для массовой разметки текстов. Напротив, формат HTML позволяет адекватно представить структурную, форматную и метатекстовую информацию, а также допускает использование нестандартных тегов. Поэтому в качестве формата разметки для корпуса было выбрано подмножество HTML плюс некоторые элементы TEI/XCES для кодирования грамматической разметки.

Технологическая цепочка подготовки корпуса включает следующие этапы обработки:

1) первичная разметка текста для представления в электронной библиотеке;
2) дополнительная структурная разметка и сегментация текста для корпуса;
3) грамматическая разметка и ее ручная постобработка (снятие омонимии, исправление разборов);
4) преобразование в базу данных, построение конкордансов и других производных.

Первоначально корпус текстов Ломоносова подготавливается в формате HTML со специальной разметкой, ориентированной на представление в электронной библиотеке. Этот формат включает достаточно полную метатекстовую, структурную и форматную разметку, необходимую для точного воспроизведения содержания и внешнего вида текста, но недостаточную для корпуса.

На следующем этапе в текст вносится дополнительная структурная разметка, которая маркирует фрагменты текста, требующие специальной обработки (заголовки, цитаты, примеры, комментарии, иноязычный текст и т.д.). Иногда такие фрагменты легко распознаются по специфическому оформлению (курсив, выравнивание), но часто их приходится определять и размечать вручную.

Далее текст пропускается через морфологический анализатор (парсер), который порождает для каждого слова некоторое множество вариантов разбора. После этого необходимо вручную проверить и исправить ошибки разбора, удалить неправильные варианты и добавить недостающие.

2.3. Элементы текста, требующие специальной обработки

1) Тексты/фрагменты на иностранных языках

Фрагмент на иностранном языке должен быть размечен при помощи специального тега, при этом должен быть определен язык фрагмента (латинский, немецкий, французский, греческий и др.). В идеале каждое слово должно быть приведено к словарной форме, для этого необходимо найти морфологический парсер для соответствующего языка. При этом приходится решать сложные филологические проблемы, связанные с тем, что орфография текста отличается от современной или непоследовательна. В крайнем случае, иноязычные слова могут быть даны как простой список словоформ без приведения к лексеме.

2) Переводы

Многие иноязычные тексты имеют переводы на современный русский язык и образуют как бы корпус параллельных текстов. Для удобства использования корпуса желательно связать оригинальные тексты с их переводами хотя бы на уровне предложений.

3) Цитаты из других авторов

Цитаты должны быть размечены в тексте при помощи специального тега, чтобы отличить авторский текст от заимствованного. Цитаты могут быть выделены по формальным признакам (курсив, кавычки и т.д.), но эти признаки довольно расплывчаты, поэтому требуется ручная разметка.

Заметим, что не всегда возможно отличить точную цитату от неточной, измененной автором при цитировании. В таком случае действует презумпция, что цитата считается авторским текстом, если не доказано обратное.

4) Примеры употребления (слова, фразы в грамматике)

Примеры употребления необходимо выделить при помощи специального тега, чтобы отличить от нормального употребления слова. Такие фрагменты похожи на цитаты, только взятые не из конкретного текста, а из некоторой модели языка. Формально примеры обычно выделяются курсивом, как и цитаты.

Семантическое различие между языковыми примерами и основным текстом достаточно очевидно даже для редакторов академического собрания. Так, в «Российской грамматике» в примерах сохраняется оригинальная орфография (с ятем, ером и т. д.), тогда как в основном тексте орфография модернизирована.

5) Фрагменты слов (слоги, буквы в грамматике)

Такие фрагменты очень похожи на примеры: они также оформляются курсивом и в них часто сохраняется оригинальная орфография. Однако, в от-

личие от примеров, такие фрагменты не считаются словами, а выделяются в отдельное подмножество (не-слова), аналогично формулам и цифрам. Проблема в том, что фрагмент слова может быть омонимичен нормальной лексеме (-у, -а, -ой, раз-), и поэтому должен быть специально размечен, чтобы случайно не попасть в список лексем.

6) Сокращения

Сокращения должны быть раскрываться, если этого возможно, например:

м. г. => м<илостивый> г<осударь>
муж. р. => муж<ескаго?> р<ода>
Е. И. В. => Е<го> И<мператорское> В<еличество>

Если раскрытие неоднозначно, после него ставится знак вопроса.

Некоторые сокращения уже раскрыты в редакторском тексте при помощи угловых скобок. Поэтому для разметки корпуса нужно выбрать другой знак, отличный от редакторских символов.

Сокращения представляют большую проблему для парсера и грамматической разметки. Парсер не может разобрать неполные или разорванные слова или порождает нелепые разборы, которые в любом случае приходится исправлять вручную. Не всегда возможно однозначно восстановить полную форму и привести ее к лексеме. В корпусе такое словоупотребление должно иметь признак, что форма сокращенная и тем самым недостоверная.

7) Цифровые комплексы

Цифровые комплексы обычно не разбираются и не заменяются на словесную запись, за исключением тех случаев, когда содержат фрагменты флексий (в 1754-м году, в 1-ой части). Это обусловлено тем, что не всегда возможно восстановить словесную форму числа, записанного цифрами. Все цифровые комплексы собираются в отдельное подмножество, аналогично фрагментам слов.

8) Формулы (в физике, химии, математике).

Формулы, переменные и другие элементы научной нотации должны размечены при помощи специального тега, чтобы они не смешивались с нормальными словами. Формулы могут содержать фрагменты, совпадающие с обычными словами (a quadratus plus b quadratus), но не должны интерпретироваться как слова. Все формулы собираются в отдельное подмножество, аналогично фрагментам слов.

2.4. Грамматическая разметка

Морфологический анализатор (парсер) — программа, выполняющая грамматический разбор текста, который включает в себя следующие задачи:

1) **токенизация** — разбиение текста на элементарные знаки (токены) и определение типа для каждого токена: слово, знак препинания, цифровой комплекс, тег разметки и т.д.
2) **сегментация** текста на предложения (а также клаузулы и др. виды сегментов);
3) **морфологический анализ** для слов, присутствующих в грамматическом словаре;
4) построение гипотез для нераспознанных слов (если это возможно).

Парсер является достаточно универсальной программой и мало зависит от конкретного языка и словаря. Вся конкретно-языковая информация о словоизменении записывается во внешних файлах в специальном формате и включает в себя две таблицы — таблицу парадигм + грамматический словарь.

Грамматический словарь — список лексем языка с приписанной им информацией о словоизменении. Каждая лексема в словаре содержит, как минимум, следующую информацию:

1) основа с указанием чередований;
2) постоянные признаки лексемы (часть речи, род, одушевленность, переходность, и т. д.);
3) номер парадигмы.

Модель словоизменения для русского языка основана на «Грамматическом словаре русского языка» А. А. Зализняка, который представляет собой наиболее авторитетный стандарт в данной области. В процессе создания парсера оказалось, что описание словоизменения в данном словаре недостаточно формально для программной реализации, и его пришлось детализировать и формализовать. В частности, пришлось расширить номенклатуру парадигм, разработать специальную нотацию для описания чередований в основе, точно описать схемы чередований, и т.д. В целом электронный словарь для парсера представляет собой отдельный продукт, заметно отличающийся от печатного издания по структуре и составу информации.

Парсер анализирует каждую словоформу по отдельности, без учета синтаксического контекста, и приписывает ей множество вариантов разбора (которое может быть пустым). Каждый вариант разбора содержит следующую информацию:

1) лексема (словарная форма);
2) грамматические признаки лексемы (часть речи, род, одушевленность, переходность);
3) грамматические признаки словоформы (число, падеж, наклонение, время, лицо);
4) номер парадигмы.

Основные проблемы при морфологическом анализе таковы:

Парсер дает только предварительную грамматическую разметку, которую необходимо проверить и исправить вручную. Основные проблемы, которые приходится решать при ручной обработке, таковы:

1) омонимия (грамматическая и лексическая), при которой словоформа получает несколько вариантов разбора;

2) отсутствие разбора, если лексема отсутствует в словаре или имеет нестандартную форму;
3) неправильный разбор.

Основная задача — устранить межлексемную омонимию, чтобы сгруппировать словоформы в словарные статьи. Что касается внутрилексемной омонимии, то устранить ее очень трудно, поскольку в русском склонении много омонимичных форм. Поэтому было принято решение, что грамматические формы внутри лексемы не размечаются и не различаются (к дому=в дому), за исключением особых случаев.

В процессе обработки текста приходится решать сложные лингвистические проблемы, связанные с различиями между языком эпохи Ломоносова и современным русским языком. Ломоносов широко использует церковнославянские формы и слова, отсутствующие в современном языке. Орфография текстов также весьма разнообразна и непоследовательна. Чтобы улучшить качество распознавания, необходима ручная настройка словаря и парсера, а также использование эвристических приемов и построение гипотез по аналогии.

3. Структура словарной статьи

3.1. Основные понятия

Лексема — множество словоформ с одинаковым лексическим значением.

Словоформа — вариант лексемы, имеющий определенное грамматическое значение.

Словоупотребление — конкретная словоформа в конкретном месте в корпусе текстов.

Словарная статья включает следующие основные зоны:

1) Заголовочное слово.
2) Грамматические пометы (часть речи, род, вид, переходность).
3) Краткая дефиниция (при необходимости).
4) Суммарная частота по всем текстам (возможно разбиение по типам/жанрам).
5) Примеры употребления с адресами и гиперссылками.

3.2. Заголовочное слово

Заголовочное слово представляет все варианты данной лексемы в компактном виде. Форма заголовочного слова выбирается по общепринятым правилам: для глаголов — инфинитив, для существительных — им. падеж ед. числа, для прилагательных — им. падеж ед. числа муж. рода и т. д.

В случае омонимии к заглавной форме добавляются цифровые индексы, в основном соответствующие словарю А.А. Зализняка, например: град[1] [осадки] — град[2] [город], мир[1] [спокойствие] — мир[2] [вселенная]. Индексы добавляются также в том случае, если омонимы имеют различные грамматические признаки (часть речи), например: знать[1] *с.ж.неод.* — знать[2] *г.нсв.* — знать[3] *вводн.* Многозначные слова при необходимости также могут быть разделены на подзначения, например: свет[1] — свет[2] [мир, бомонд], двор[1] — двор[2] [окружение монарха].

При группировке словоформ в словарную статью возникает много нетривиальных проблем, обусловленных широкой вариативностью исходного материала, в частности:

- какие формы можно считать вариантами одной лексемы?
- как удобнее группировать варианты?
- какой вариант выбрать в качестве основного (заголовочного слова)?

Исходя из соображений удобства, мы приняли следующие практические правила.

Орфографические и морфологические варианты по возможности нужно объединять в одну статью: знание=знанье, вариант=варьянт, кресло=кресла, зал=зала=зало.

Форма заголовочного слова должна включать все варианты: знание, знанье; или основной вариант и другие варианты в скобках: знание (знанье). В качестве основного варианта выбирается самый частотный или совпадающий с современным (для удобства пользователя).

Слова, которые могут писаться слитно и раздельно, желательно считать единой лексемой, а не разбивать на отдельные слова: вслед=в след, наверное=на верное, также=так же.

Исключениями из этого правила являются:

- частица не с финитными формами глагола: незнает = не+знает;
- клитики ж(е), ли(ль), б(ы): тыж = ты+же, онже = он+же, егоже = его+же или егоже (от иже), тыль = ты+ль, яб = я+б.

3.3. Грамматические признаки лексемы

Каждая лексема имеет признак «грамматический класс» (часть речи), который в основном соответствует словарю Зализняка и грамматической традиции. Далее указываются грамматические признаки, характерные для данного класса лексем: для глаголов — вид, для существительных — род и одушевленность, а также число для singularia и pluralia tantum. Грамматические признаки записываются с использованием достаточно понятных сокращений.

Группировка грамматических форм по лексемам соответствует грамматической традиции. Так, причастия и деепричастия включаются в парадигму глагола наравне с финитными формами. Степени сравнения включаются в парадигму

соответствующих прилагательных и наречий, кроме некоторых специфических форм (*больше, меньше, дальше*), которые считаются отдельными наречиями.

По практическим соображениям в ряде случаев мы решили игнорировать незначительные смысловые и синтаксические различия между парами омонимичных (полисемичных?) слов, которые в традиционной грамматике относятся к разным грамматическим классам, например:

1) многие субстантивированные прилагательные (*знакомый, приезжий, русский*) не отличаются от исходных прилагательных, поскольку они образованы по регулярной семантической модели;
2) не различаются мелкие оттенки значения для некоторых неизменяемых слов, например: *еще, уже* — наречие и частица, *ли, разве, даже* — частица и союз, и др.
3) прилагательные, не отличающиеся по смыслу от соответствующих причастий (*открытый, одетый, занятый* и т.д.), обычно считаются причастиями и включаются в парадигму соответствующего глагола.

3.4. Краткая дефиниция

Конкорданс не является толковым словарем и ориентирован на человека, владеющего русским языком. Общеязыковые слова в конкордансе даются без дефиниций и без филиации значений, поскольку эту информацию легко получить из любого толкового словаря.

Краткая дефиниция используется только для объяснения редких и устаревших слов, специальных терминов, а также для различения омонимов типа град[1] [осадки] — град[2] [город]. При необходимости дефиниция может быть расширена или дополнена ссылкой на словарь. Основные источники дефиниций для неизвестных слов:

- списки устаревших слов и комментарии в Собрании сочинений Ломоносова;
- Словарь русского языка XVIII века;
- Словарь Академии Российской.

Краткая дефиниция не пытается подменить полноценную энциклопедическую статью, а служит только для получения общего представления об описываемом предмете. Конкорданс предназначен для исследования языка, а не истории реалий.

3.5. Частота

Указывается абсолютная частота лексемы, т. е. количество всех ее употреблений во всех рассматриваемых текстах. При необходимости дается разбиение по основным типам/жанрам текстов.

3.6. Примеры употребления

Включает полный список всех употреблений данной лексемы, включая высокочастотные слова (предлоги, союзы). Каждое словоупотребление включает следующий набор признаков.

1) Контекст

Контекст — это связный фрагмент исходного текста, включающий данную словоформу и достаточный для понимания ее смысла и синтаксического окружения. Обычно контекст представляет собой целое предложение или клаузулу (связная часть сложного предложения), а при необходимости расширяется за указанные пределы. Сама словоформа выделяется жирным шрифтом.

2) Адрес

Адрес данного словоупотребления в корпусе, достаточный для его идентификации при цитировании и предназначенный для человека. Адрес включает в себя:

1) короткое имя произведения;

2) внутренний адрес — названия или номера явно выраженных структурных элементов текста (действие, явление, реплика, ремарка, заголовок) или меток (номер страницы, стиха).

Для драматических произведений внутренний адрес включает в себя: номер действия, номер явления, имя говорящего.

Для поэтических произведений внутренний адрес включает номер строфы и стиха.

При отсутствии структуры внутренний адрес может включать в себя номер страницы.

Адрес не всегда обеспечивает однозначную идентификацию текстового фрагмента (может быть несколько реплик одного лица на одной странице, ремарки не имеют имени), но вполне достаточен для цитирования и согласуется с принятой традицией.

3) Ссылка на текст

Прямая гипертекстовая ссылка на соответствующее место в корпусе текстов.

Предполагается, что в тексте все предложения пронумерованы независимо от структурной разметки и имеют метки, на которые можно сразу перейти.

Прямая ссылка на текст необходима, чтобы получить более широкий контекст для данного словоупотребления, если пример в конкордансе недостаточно информативен.

4) Тип фрагмента

Тип текстового фрагмента, присвоенный при первоначальной разметке, например:

- заголовок структурного элемента (часть, глава, реплика);
- ремарка;
- сноска;

- цитата;
- пример употребления слова;
- тип речи (проза/стихи);
- подпись под рисунком, и т.д.

Разные типы текстовых фрагментов имеют языковые особенности, которые должны быть предметом специального изучения.

4. Возможности использования конкорданса

Размеченный конкорданс представляет собой словарную базу данных, из которой путем различной проекции и группировки данных можно получать различные виды словарей и проводить объективные исследования авторского языка.

Форма базы данных открывает целый ряд возможностей, недоступных в традиционных бумажных словарях.

- динамический выбор примеров по любым параметрам,
- динамическая сортировка и группировка,
- быстрый переход из словаря в корпус текстов,
- просмотр и выдача словарной информации в различных форматах,
- генерация печатных словарей.

4.1. Динамический выбор примеров

Пользователь может создать для себя рабочее подмножество (проекцию) корпуса по любым текстовым и метатекстовым параметрам:

- жанр и тематика текста,
- дата написания,
- название произведения,
- тип текстового фрагмента (заголовок, сноска, цитата, проза/стихи),
- класс лексической единицы (русское/латинское/немецкое слово, цифры, фрагменты),
- грамматические признаки лексемы (часть речи, род, вид, переходность),
- контекст (соседние слова),
  и т. д.

4.2. Динамическая сортировка

Стандартной формой представления информации является алфавитно-частотный конкорданс, где лексемы отсортированы в алфавитном порядке, а примеры внутри статьи — по словоформе. Эта форма выбрана как наиболее нейтральная и понятная. На самом деле примеры внутри статьи могут сортироваться по любым существующим параметрам, например:

- метатекстовые параметры (жанр, автор, название, дата),
- тип текстового фрагмента,
- контекст, т.е. слово справа/слева (так называемый KWIC — keyword in context).

При необходимости можно сделать сортировку по нескольким параметрам, например: словоформа + контекст+жанр, жанр+словоформа+слово справа, и т.д. Видимо, наиболее удобной является сортировка примеров по признаку словоформа+контекст (слово справа+слово слева).

4.3. Группировка

Путем различной сортировки и группировки данных из одной и той же словарной базы можно получить следующие виды словарей:

- алфавитный конкорданс, где лексемы отсортированы в алфавитном порядке, а словоупотребления внутри статьи — по грамматической форме или по KWIC;
- частотный словарь, где лексемы сгруппированы в порядке убывания частоты,
- обратный алфавитный словарь,
- грамматический словарь, где лексемы сгруппированы по грамматическим признакам,
- словари отдельных произведений или типов речи, и т. д.

Кроме того, исследователь всегда сможет получить нужную ему проекцию словаря по запросу. Возможность динамического получения новых видов словарей (в том числе, не предусмотренных первоначальным замыслом) является совершенно новым словом в практике филологической работы.

Синтактически инвариантный метод идентификации семантики информации

Syntactically the invariant method of identification of semantics of the information

Потапов М. В. (potapov_mv@rgrta.ryazan.ru)

Государственное образовательное учреждение высшего профессионального образования «Рязанский государственный радиотехнический университет», Рязань, Россия

Содержится описание практически апробированного метода оценки смыслового содержания информационных потоков, основанного на статистико-лингвистическом способе их представления и обработки с использованием подходов теории распознавания образов при анализе многомерных признаков.

Для оценки состояния контролируемых на удалении сложных технических комплексов актуальна задача распознавания и идентификации генерируемых информационных сообщений. Для этого предлагается использовать текстовое представление информационных потоков:

$$\vartheta = (A_X, \Theta), \quad (1)$$

где A_X — алфавит сообщений, Θ — отношения между элементами текста. Текст Θ — это множество элементарных знаков, между которыми установлены отношения $\Theta=\{\Theta_1, \Theta_2, \Theta_3\}$, определяемые правилами функционирования объектов, генерирующих информацию, принятой системой интерпретации и целями исследований. Синтактика Θ_1 связывает с текстом некоторую структуру отношений между знаками независимо от их содержания. Отношения между объектами и их обозначениями рассматривают семантика Θ_2 и прагматика Θ_3. Каждое из слов текста ϑ состоит из символов A_X. Пусть все слова текста ϑ образуют множество:

$$X_\Gamma(t) = \{x_1, x_2, ..., x_{N_\Gamma}\}, \quad (2)$$

где $N_\Gamma=\mathrm{Card}(X_\Gamma)$, а различные словоформы (2) образуют множество (алфавит):

$$Ax = \{a_1, a_2, ..., a_n\},\ a_i \neq a_j,\ i, j \in \{1, 2, 3, ..., n\}, \quad (3)$$

где $n = \mathrm{Card}(A_X)$ — мощность алфавита, причём $A_X \subset X_\Gamma(t)$.

Выбор типа словаря (3) зависит от особенностей формирования и дальнейшего использования информационного процесса (2). Место каждого знака a_i внутри каждого слова определяется смысловым содержанием и моделью функционирования объекта, порождающего эту информацию. Когда известна структура (синтактика), режимы и программы функционирования объекта (семантика и прагматика), анализ смыслового содержания состоит в оперативной оценке Θ знаковой системы (1). Это возможно при отсутствии помех естественного и искусственного характера, кодирующих преобразований, которые, разрушая структуру передаваемых данных, приводят к неоднозначности (1) — к потере их смыслового содержания. В этих условиях качественное и эффективное решение рассматриваемой задачи обеспечивает статистико-лингвистический подход [1, 2]. В качестве элементов (словоформ) алфавита (3) выступают слова сообщений:

$$A_X = \{a_1=x_1, a_2=x_2, a_3=x_3, \dots\} \quad (4)$$

В условиях неизвестной структуры сообщений, в качестве словоформ предлагается использовать длины блоков одноименных, подряд следующих одноимённых кортежей “0” и “1”:

$$A_X = \{a_1=\{0, 1\};\ a_2=\{00, 11\};\ a_3 = \{000, 111\}; \dots; a_n=\{0 \dots 0,1 \dots 1\}\} \quad (5)$$

Установлено, что осмысленные, информационно наполненные сообщения, содержат достаточно длинные кортежи $n>>30$. Это позволяет идентифицировать эти сообщения на фоне помех.

Предлагаемый метод сводится к последовательности процедур. Вначале, по словарю A_X (4) или

(5) строится эмпирический ряд частот появления словоформ:

$$f(x) = \{f(x)_i = Q^a_i / N_\Gamma,\ i=1, 2, 3, \ldots, n\}, \quad (6)$$

где Q^a_i — число вхождений слов $x \in A_X$ в $X_\Gamma(t)$; N_Γ — объём выборки.

Отметим, что для неискажённых, семантически нагруженных данных распределение (6) является негауссовым. Ранговым распределением назовём функцию $\Phi(r)$, которая ставит в соответствие номеру (рангу) r слова $x \in A_X$ частоту его появления f(x):

$$\Phi(r) = \{\Phi(r)_i \leftrightarrow f(x)_i,\ \Phi(r)_i > \Phi(r)_i + 1,\ i = 1, 2, 3, \ldots, n-1\} \quad (7)$$

Эмпирический ряд (7) прологарифмируем и аппроксимируем модифицированной зависимостью Ципфа-Мандельброта [1, 2] вида:

$$\Phi(r) = C_n r^{-\gamma_o \mathrm{Exp}(dr)}, \quad (8)$$

где C_n — константа, зависящая от n, γ_0 — начальный показатель рангового распределения $\gamma(r)=\gamma_o \mathrm{Exp}(dr)$, d — коэффициент его прироста. Их начальные значения определим как:

$$\begin{aligned} &C_n = \mathrm{MAX}\,\Phi(r) = \Phi(0);\\ &\gamma_r = (\mathrm{Lg}(\Phi_{max}) - \mathrm{Lg}(\Phi_r)) / \mathrm{Lg}(r)\ ;\\ &\gamma_0 = \gamma_r / \mathrm{Exp}(dr)\ ;\\ &d_{ij} = (\mathrm{Ln}(\gamma_i) - \mathrm{Ln}(\gamma_j)) / (r_i - r_j);\ i \neq j. \end{aligned} \quad (9)$$

Первичная аппроксимация (8–9) в последующем улучшается последовательно-итерационным подбором коэффициентов C_n, γ_0 , d . Критерием наилучшего приближения является минимум суммы квадратов разностей эмпирического $\Phi^Э(r)$ и аппроксимирующей зависимости теоретического $\Phi^Т(r)$ законов распределения:

$$S_\phi = 1/n \sum_{i=1}^{n} (\Phi^Э(r)_j - \Phi^Т(r)_j)^2 \quad (10)$$

При достижении Sф заданного порогового уровня $S^П_ф$ по критерию

$$\lim_{n\to\infty} S_\phi \to S^П_\phi \quad (11)$$

фиксируются — записываются в базу данных коэффициенты модели (8–9) и формируется точка в признаковом пространстве Далее проводится проверка гипотезы о принадлежности анализируемой информации к одной из эталонных $U_{\gamma d,\, i}$ с использованием критерия «минимаксного» расстояния, по результату которого либо дополняется эталонная база, либо отвергается выдвинутая гипотеза. В первом случае по расстоянию между полученными координатами (C_n, γ_0, d) и центром «тяжести» $Z_{u,\, i}$ наиболее близкого кластера $U_{\gamma d,\, i}$ оценивается мера качества, например отношение сигнал/шум:

$$\mathrm{MIN}_i(R_i = |(\gamma_0, d) - Z_{u,\, i}|). \quad (12)$$

Проведённые исследования показали, что модель (8–9) является точной и хорошо отражает закономерности эмпирического ряда в выборках различного объёма (рис. 1) разнородных информационных сообщений. Выпуклые области изменений коэффициентов γ_0 и d, построенные по интервальным оценкам при $P_Д = 0{,}95$ и фиксированном объёме выборки 20 Кбайт. Области возможных изменений γ_0 и d представляют статистические образы различных информационных процессов и могут использоваться для их идентификации.

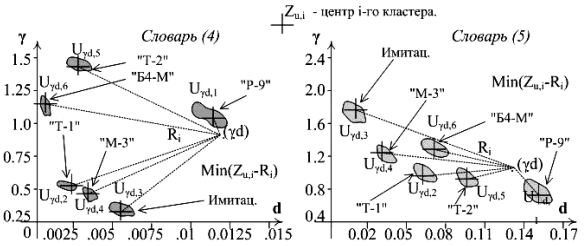


Рис. 1. Области коэффициентов аппроксимации ранговых распределений

Искажение шумами, шифрование, изменение смыслового содержания в целом или какой-либо его части приводят к изменению соответствий (8–9) и, как следствие, к трансформированию частотно-рангового распределения и увеличению размеров областей изменения коэффициентов γ_0 и d, что подтверждается результатами проведённого моделирования (рис. 2). Таким образом, можно выделить области $U_{\gamma d}=\{\gamma_{min}<\gamma_0<\gamma_{max};\ d_{min}<d<d_{max}\}$ для сообщений с нарушенной структурой сообщений $U^Ш_{\gamma d}$ и области $U^Э_{\gamma d}$, характерные для смысловой информации. Используя эти свойства и зная $U_{\gamma d,\, i} = \{U^Ш_{\gamma d}, i, U^Э_{\gamma d}, i\}$ словарей (4), (5) для каждого i-го типа данных можно строить алгоритмы оценки качества, отбраковки недостоверных участков, оценки отношения сигнал/шум идентификации сообщений и др.

Для сопоставления (идентификации) процессов предлагается использовать меры, предоставляемые теорией распознавания образов, рассматривая в качестве признаков коэффициенты γ и d. Статистико-лингвистический способ оценки достоверности и смыслового содержания информации с использованием геометрической интерпретации, кластеризации объектов, получения решающих функций с учётом весов и мер расстояний, устранения незначимых параметров и уменьшения размерности признакового пространства практически апробирован на разнообразных данных. Результаты быстрой и надёжной кластериза-

ции при наполнении эталонной базы и последующее гарантированное распознавание информации позволяют в качестве меры качества использовать расстояние между оценками (γ_0, d) и центрами «тяжести» $Z_{u,i}$ кластеров $U_{\gamma d,i}$ в признаковом пространстве γ, d.

При аппроксимации частотно ранговых распределений некоторых информационно-выраженных процессов зависимостью (8) был выявлен эффект получения неудовлетворительной оценки ($R \to max$) при $\Phi(r \to max) = 0$, $n_i < n_{max}$. Положительный эффект даёт исключения из анализа элементов $\Phi(r) = 0$, а также более точный подбор коэффициентов апроксимирующей зависимости γ_0 и d. Отмечено также, что для выполнения надёжного оценивания, необходимо выбирать одинаковые тип словаря (4) или (5), $N_Г$, $R_{i\,зад}$, $S^{П}_{ф} < 1\%$, $P_Д$.

Для гарантированной идентификации в алгоритм включены процедуры получения оценок тяжести «хвостов» [3], а также анализа законов распределения и статистических характеристик коэффициентов частотно-ранговых распределения (8). Кроме этого, дополнительными идентифицирующими критериями могут быть использованы факты выхода элементов частотно-рангового распределения анализируемого процесса из статистически накопленных эталонных допусковых коридоров разброса (6–12). Статистико-лингвистический алгоритм позволяет получать оценки качества при анализе как перекрывающихся, так и не перекрывающихся выборок из информационного потока (2).

Частотно-ранговые распределения (6–13) являются характеристиками неискажённого сигнала и сигнала с разрушенной информационной структурой кодирующим псевдослучайным преобразованием или шумами (рис. 2). Эти распределения и отмеченные закономерности их параметров являются важной характеристикой, отражающей статистические и структурные свойства анализируемой информации.

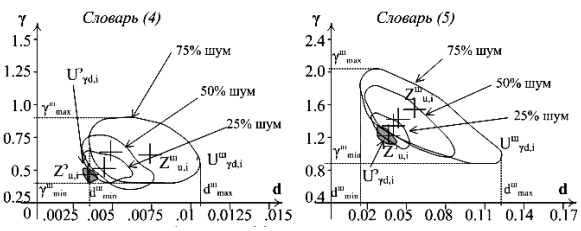


Рис. 2. Изменения областей коэффициентов аппроксимации ранговых распределений при зашумлении

На рис. 3 приведены результаты обработки, представляющие собой аппроксимацию на основе зависимости (8) эмпирического ряда распределений для различных по информативности источников информации.

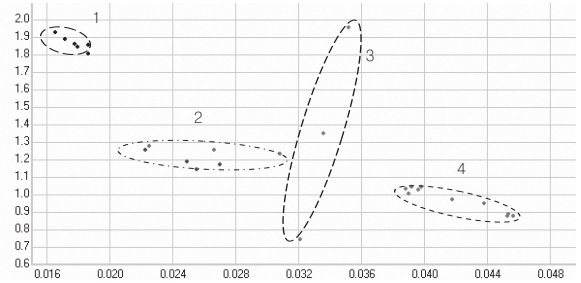


Рис. 3. Области возможных вариаций коэффициентов аппроксимации γ_0 и d

Области возможных вариаций коэффициентов аппроксимации γ_0 и *d* на рис. 3 подразделяются на четыре группы систем: 1 — малоинформативные (50 кбит/с), 2 — среднеинформативные (220 кбит/с), 3 и 4 — высокоинформативные (750, 2300 кбит/с). Решение о принадлежности значений {γ_0, d}тому или иному типу систем принимается на основании критерия качества

$$F = \ln (d / (c \cdot \lambda)), \qquad (14)$$

где c — расстояние между значениями {γ_0 , d} внутри группы, d — расстояние между значениями {γ_0, d} разных групп, λ — мера «одинаковости структуры» групп.

Точка со значениями {γ_0, d} относится к тому типу систем, где максимально значение критерия качества F. Частоты встречаемости словоформ для информации каждой из групп занимают устойчивые положения. Их можно использовать в качестве статистико-лингвистических образов систем, порождающих эту информацию, они характеризуют качество проектирования структур передаваемых данных телеизмерений. По тенденциям изменения коэффициентов γ_0 и d, используя свойство стабильности проявления закона распределения словоформ, можно идентифицировать характерные участки жизненного цикла одной и той же системы (рис. 4).

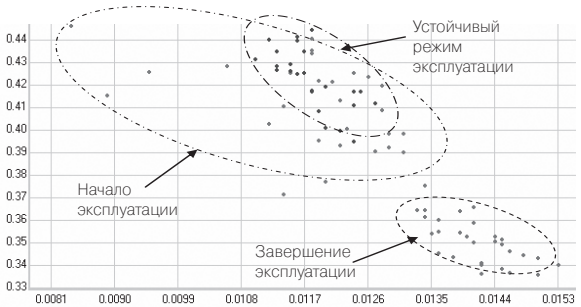


Рис. 4. Результаты анализа словоформ {0–1} словарь (5)

Использование предлагаемого подхода позволяет решить задачу классификации без выделения конкретных элементов (параметров), идентифицировать информационные фрагменты. Перспек-

тивным прикладным направлением полученных результатов является оценивание достоверности измерительной информации в нештатных, аварийных ситуациях и при наличии шумов и сбоев. Достоинства предлагаемого метода: работоспособность в условиях семантической неопределённости; синтаксическая инвариантность; снижение размерности анализируемых признаков; высокое быстродействие; возможность автоматизации процесса анализа; инвариантность к фазовым характеристикам анализируемых данных; высокая чувствительность; эффективная идентификация аномальных явлений в потоке данных; точность анализа вплоть до бита.

Предлагаемый метод статистико-лингвистического и графического анализа приводит к порождению новой информации о характере анализируемых данных — знаниепорождающей (когнитивной) графики [4]. Он практически апробирован на реальной разнородной информации и дал положительные результаты. Этот метод использован для построения высокоразвитых комплексов средств автоматизации распределённой автоматизированной системы, реализующей сквозную технологию регистрации, обработки, анализа, представления, хранения, тиражирования и архивирования потоковой циклической информации [5, 6].

Литература

1. *Мандельброт Б.* Теория информации и психолингвистика// Математические модели в социальных науках. — М.: Наука, 1973. С. 316–322.
2. *Орлов Ю. К.* Обобщенный закон Ципфа—Мандельброта и частотные структуры информационных единиц различных уровней// Вычислительная лингвистика. — М.: Наука, С. 179–194.
3. *Кукушкин С. С., Потапов М. В.* Статистико-лингвистические методы оценки смыслового содержания информации. / Проектирование ЭВМ. Межвузовский сборник научных трудов. Рязань, 1994. 112 с.
4. *Потапов М. В.* Когнитивная графика в задачах анализа телеметрической информации. 12-я международная научно-техническая конференция «Проблемы передачи и обработки информации в сетях и системах телекоммуникаций»: Тез. докл. /Рязан. гос. радиотехн. акад. Рязань, 2004. -180 с. ISBN 5-7722-0209-X.
5. *Везенов В. И., Гусев И. А., Потапов М. В., Селецкий О. Б., Юдин В. И.* Комплекс средств автоматизации ввода, регистрации, обработки, анализа и представления телеметрической информации. 3-я международная научно-техническая конференция «Космонавтика. Радиоэлектроника. Геоинформатика»: Тез. докл. /Рязан. гос. радиотехн. акад. Рязань, 2000. — 358с. ISBN 5-7722-0147-6.
6. *Потапов М. В., Кузин В. А.* Анализ характеристик телеметрической информации в условиях недостаточных априорных сведений. 4-я международная научно-техническая конференция «Космонавтика. Радиоэлектроника. Геоинформатика»: Тез. докл. /Рязан. гос. радиотехн. акад. Рязань, 2003. — 360с.

Неконтролируемый синтаксический анализ

Unsupervised parsing

Потемкин С. Б. (potemkin@philol.msu.ru)

Филологический факультет МГУ, Москва, Россия

Представлен статистический подход к синтаксическому анализу необработанных текстов в формализме дерева зависимостей. Приведен алгоритм, выполняющий парсинг зависимостей за время, квадратично зависящее от длины предложения, после обучения на неразмеченном корпусе.

1. Введение

Решение проблемы автоматического синтаксического анализа без предварительной ручной настройки и обучения на размеченном корпусе имеет большое теоретическое и практическое значение. Правила разбора, полученные в результате, могут пролить свет на процессы освоения языка людьми и на общую структуру языка, обеспечить предварительную обработку текстов при синтаксической разметке больших корпусов и, в перспективе, качественный анализ текстов для прикладных задач обработки естественного языка. Интерес к этой проблеме существенно повысился благодаря доступности огромных корпусов текстов, росту вычислительных мощностей и новым алгоритмам машинного обучения.

В последнее время прилагались большие усилия по использованию для синтаксического анализа размеченных корпусов, которые позволяют проводить проверку гипотез, выдвигаемых грамматическими теориями, а также формировать сами правила синтаксиса. Этот процесс называется «тренировкой» формальной грамматики и должен завершиться при достижении некоторого малого процента ошибок. Для тренировки грамматики составляются синтаксически аннотированные корпуса, которые получили название «treebank». В настоящее время имеются три-банки для языков: болгарского (BulTreeBank), польского (Проект CRIT-2), русского (ЭТАП-3, ИППИ РАН), и наиболее продвинутый для чешского языка (Prague Dependency Treebank). Ведутся работы для балканских (сербохорватского, словенского, боснийского) языков.

Большинство работ по синтаксическому анализу основано либо на правилах, либо на управляемом обучении. Хорошие синтаксические анализаторы в формализме непосредственных составляющих (НС) доступны для английского и некоторых других языков [3]. Также имеются работы, основанные на формализме дерева зависимостей [6, 7, 8, 10].

Для большинства языков мира, однако, отсутствуют хорошие синтаксические анализаторы, либо вообще какие-либо анализаторы. Это связано с тем фактом, что для большинства языков отсутствуют ресурсы, необходимые как для построения парсеров на основе правил (полные компьютерные грамматики), так и парсеров, обучаемых на примерах из три-банка. Поскольку создание таких ресурсов требует больших материальных и трудовых затрат, желательно разработать достаточно точные методы для выполнения грамматического разбора без обучения на три-банке, или для автоматического или полуавтоматического создания три-банка.

В течение последних лет наблюдается устойчивый прогресс в области неконтролируемого синтаксического анализа, но большая часть работ основана на НС-грамматиках, тогда как для описания синтаксиса русского языка традиционно используется модель зависимостей [11].

2. Достигнутый уровень разработок

Один из подходов к упрощению и облегчению синтаксической разметки национального корпуса заключается в использовании размеченного кор-

пуса другого, хорошо документированного языка, с применением специально создаваемых для этой цели алгоритмов «перевода разметки». В качестве опорного размеченного корпуса обычно выступает PennTreeBank английского языка. Поскольку для большинства языков имеются хотя бы переводные английские словари, задача разметки, в общем, упрощается, хотя результаты и неидеальны. Это особенно заметно для славянских языков с относительно свободным порядком слов, грамматика которых обычно описывается формализмом зависимостей, тогда как в PennTreeBank использован HC формализм.

Другой, чисто статистический подход, имеет определенные преимущества — необходимо иметь лишь ограниченный (около 1 млн. словоупотреблений) неразмеченный национальный корпус, без корпуса параллельных текстов и даже без двуязычного переводного словаря. Это особенно важно для малых и исчезающих языков, для описания которых отсутствуют материальные и людские ресурсы.

Статистический подход к синтаксическому разбору предложений применен в нескольких, взаимосвязанных методиках, включая DLM (dependency language model) (Gao, Suzuki, 2003) [4], U-DOP (unsupervised data-oriented parsing) (Bod, 2006) [2], CCL (common cover links) (Seginer 2007) [9].

В рамках метода Бода для выполнения парсинга необходимо:

- Построить все возможные деревья разбора для всех предложений корпуса и все поддеревья каждого дерева разбора.
- Найти наилучшее (наиболее вероятное) дерево разбора для данного предложения.

Фактически, при реализации метода возникают значительные вычислительные трудности, поскольку рост числа поддеревьев превышает экспоненциальный (каталанские числа) с удлинением предложения. Для решения этих проблем предложены методы представления поддеревьев в виде вероятностной контекстно-свободной грамматике, [5] и записи всех деревьев в виде «совместного леса» [1], что сводит задачу к обозримому, однако очень значительному, объему вычислений.

В подходе Сегинера общепринятое представление структуры предложения в виде дерева зависимостей заменяется совокупностью «общих покрывающих связей» (Common Cover Links, CCL). Разбор предложения проводится последовательно, пословно, путем анализа начальной последовательности слов предложения. Результаты такого частичного анализа в дальнейшем не изменяются, а лишь дополняются. Каждая новая связь добавляется, если она не нарушает определенные правила, заданные априори и если она обладает максимальным (среди допустимых) весом. Для определения весов создается лексикон, содержащий для каждого встреченного в тексте слова список левых и правых, связанных с ним, соседей и частоту таких связей.

По сравнению со структурой зависимостей CCL структура обладает определенными преимуществами: во-первых, для предложения типа «I know the boy sleeps» со структурой зависимостей [[I][know] [[the boy][sleeps]]] CCL не устанавливает направления связи в отношении [the boy]. Аналогично, для русского языка не будет установлено направление управления в предложно-падежной группе.

Второе отличие более существенно. В традиционном методе к моменту прочтения предложения до слова boy будет установлена зависимость между know и boy, однако после прочтения предложения до конца, придется удалить эту связь и установить новые – [know sleeps] и [sleeps boy] Эта проблема известна в психолингвистике как проблема повторного анализа. В CCL структуре эта проблема обойдена путем назначения каждой связи значения «глубины» этой связи. Этим достигается однозначность восстановления скобочной структуры, без необходимости удаления ранее установленных связей. Парсер на основе CCL, настроенный на английский язык, доступен для некоммерческого использования, http://staff.science.uva.nl/~yseginer/ccl/.

Наконец, Гао и Судзуки также предложили инкрементный подход к парсингу, при котором структура зависимостей строится последовательно, после ввода очередного слова предложения и вычеркивания связей, нарушающих ацикличность и проективность. Их метод был применен не к анализу структуры предложения, а к восстановлению иероглифической записи японского предложения (кана-кандзи) на основании слоговой записи (кана) – эта проблема и метод ее инкрементного решения характерны также для задачи распознавания речи.

Настоящая работа опирается в основном на методику Гао и Судзуки, однако для анализа зависимостей предложения разработан алгоритм, строящий покрывающее дерево всего предложения, без вычеркивания связей, работающий за время $O(n^2)$, при сохранении классического вида структуры зависимостей.

На основе представленного метода возможна автоматическая синтаксическая разметка неаннотированного корпуса, как для русского языка, так и для других языков, имеющих достаточные объемы электронных текстов, с преобладанием проективных предложений.

3. Модель локальных связей (МЛС)

В предлагаемой нами модели локальных связей стру ктура зависимостей строится снизу вверх. Вначале устанавливаются связи между соседними словами (локальность), которые объединяются в юниты, затем устанавливаются связи между соседними юнитами, и так далее, пока не достигает-

ся последний, верхний уровень объединения, чем и завершается построение дерева зависимостей. Существенным в этом процессе является выбор последовательности объединения юнитов, который определяется весом связи между ними. Аналогично модели грамматики связей (LG) [12] в нашей модели установленные связи являются ненаправленными, но в отличие от указанной работы, связи не помечаются и их установка не требует заранее подготовленного лексикона моделей управления.

3.1. Определения

Для более формального описания модели введем обозначения:

W — последовательность слов предложения;
W={w1, w2,..., wn}

T — дерево зависимостей, построенное над W;
T = {(i, j)}, где i, j – номера слов, связанных зависимостью. T является проективным деревом.

U — юнит, поддерево T над неразрывной подпоследовательностью W; U_{k0}*=wk, либо* U_{kl} *= {wk, wk+1, …, w+l}, где каждая пара слов связана ветвью, принадлежащей дереву T.*

Открытой вершиной юнита U назовем такую вершину wm, для которой не существует принадлежащей U ветви (i, j), i < m < j. Иначе вершина wm является закрытой.

Смежными юнитами U_{al}={wa, wa+1,…,wap} U_{bm} = {wb, wb+1,…, wbq} назовем такие юниты, для которых b=ap+1, то есть начало второго юнита непосредственно следует за концом первого юнита.

В принципе, модель языка должна определять вероятность предложения W по всем возможным деревьям T над W, то есть

$$P(W) = \Sigma P(W, T) \text{ по всем } T. \quad (1)$$

Практически, для оценки P(W) используется единственный член суммы, а именно P(W,T*):
где T* — наиболее вероятная структура зависимостей предложения, которая доставляет максимум выражению P(W,T):

$$T^* = \operatorname{argmax} P(W,T) \quad (2)$$

Цель парсинга состоит в том, чтобы найти самый вероятный разбор T* данного предложения W, максимизирующий вероятность P(T|W). Предполагая, что связи (i,j) независимы друг от друга (очень сильное допущение), имеем

$$P(T|W) = \Pi P((i,j)|W) \quad (3)$$

где P((i,j)|W) является вероятностью зависимости (i,j) в конкретном предложении W. Вероятность P((i,j)|W) невозможно оценить непосредственно, поскольку мы предполагаем, что корпус не содержит, или содержит очень мало тождественных предложений. Поэтому в качестве приближения P((i,j)|W) берется вероятность P(i,j), которая зависит только от встречаемости слов wi, wj в предложениях корпуса и, возможно, от расстояния (j–i).

Вероятность P(i,j) оценивается как

$$P(i,j) = C(wi, wj, R) / C(wi, wj) \quad (4)$$

где C(wi,wj,R) – число обнаружений связи R между словами wi и wj в корпусе, а C(wi,wj) — число обнаружений слов wi и wj в одном и том же предложении корпуса.

Вероятность зависимости P(i,j) можно рассматривать как вес связи d(i,j), то есть, связь с более высокой вероятностью имеет больший вес.

Проблема разреженности данных решается использованием приближения, описанного в работе [3], а именно, используется следующая оценка:

$$d(i,j) = E = \lambda 1 E1 + (1 - \lambda 1)(\lambda 2 E23 + (1 - \lambda 2) E4) \quad (5)$$

где

$$E1 = CR1/C1;\ E23 = (CR2+CR3)/(C2+C3)\ E4 = CR4/C4$$
$$CR1 = C(wi, wj, R);\ C1 = C(wi, wj),$$
$$CR2 = C(wi, *, R);\ C2 = C(wi, *),$$
$$CR3 = C(*, wj, R);\ C3 = C(*, wj),$$
$$CR4 = C(*, *, R);\ C4 = C(*, *).$$

где * означает любое слово.

Параметры λ1 и λ2 лежат в диапазоне (0,1) и определяются экспериментально. Нами приняты значения, приведенные в работе [4], а именно λ1=0,7, λ2=0,3.

3.2. Алгоритм парсинга

Традиционные методы парсинга используют алгоритм динамического программирования, который требует $O(n^5)$ операций. Для парсеров с использованием биграмной модели лексических зависимостей разработаны $O(n^3)$ алгоритмы [8].

Приведенный ниже алгоритм строит проективное дерево T* над последовательностью вершин {1, …, n} за время $O(n^2)$ при заданных значениях d(i,j), он очень эффективен и прост в реализации.

Функция *склеить*(Uap, Ubq, i, j) удаляет юниты Uap, Ubq, создает на их месте новый юнит Ua(p+q+1) и закрывает в нем все вершины, лежащие в интервале между i и j.

Предлагаемый алгоритм парсинга (рис. 1) локальных зависимостей (ПЛЗ) требует $O(n^2)$ операций для разбора предложения длиной n слов. Докажем это утверждение.

ПАРСИНГ ЛОКАЛЬНЫХ ЗАВИСИМОСТЕЙ (W)

```
1  n = длина(W)
2  do while n>0
3  dmax = max d(i, j), где i, j есть открытые вершины смежных юнитов Uap, Ubq
4  поместить = (wi, wj) в T*
5  Ua(p+q+1) = склеить(Uap, Ubq, i, j))
6  n = n–1
7  end do
8  return(T)
```

Рис. 1. Алгоритм парсинга локальных зависимостей

На последнем шаге цикла требуется установить связь между двумя юнитами, покрывающими все предложение. Для этого требуется найти максимальную по весу связь между открытыми вершинами этих юнитов. В наихудшем случае юниты имеют равную длину (или их длины отличаются на 1) и все их вершины открыты. Для выбора максимальной связи потребуется выполнить n/2 * n/2 т.е. $n^2/4$ сравнений.

На предпоследнем шаге каждый из юнитов делится пополам и потребуется выполнить $2*n^2/16$ сравнений. На шаге n-i потребуется выполнить $2^i *(n^2/2^{2i}) = n^2/2^i$ сравнений.

Суммируя по i, получаем общее число сравнений для наихудшего случая разбора:

$$n^2 * \Sigma\, 1/2^i$$

Ряд сходится к 1, а общее число операций = $O(n^2)$

3.3. Создание корпуса для обучения

В этом разделе описаны два метода, которые использовались, чтобы разметить необработанный текстовый корпус для обучения МЛС:

(i) сбор статистики грамматических признаков n-грам, n=3, и

Грамматические признаки кодировались согласно Грамматическому словарю Зализняка. Морфологическая омонимия не снималась, вместо этого грамматические признаки омонимичной словоформы расщеплялись: если словоформе может быть приписано m различных грамматических кодировок, статистика каждой из этих кодировок увеличивалась на 1/m.

(ii) сбор статистики k-буквенных окончаний n-грам, k=4, n=5.

Поскольку русский язык относится к флективным языкам, статистика k-буквенных окончаний использовалась параллельно статистике грамматических признаков, а также для внутренней проверки метода.

Для сбора статистики использовалась часть коллекции www.lib.ru объемом около 2 Гбайт.

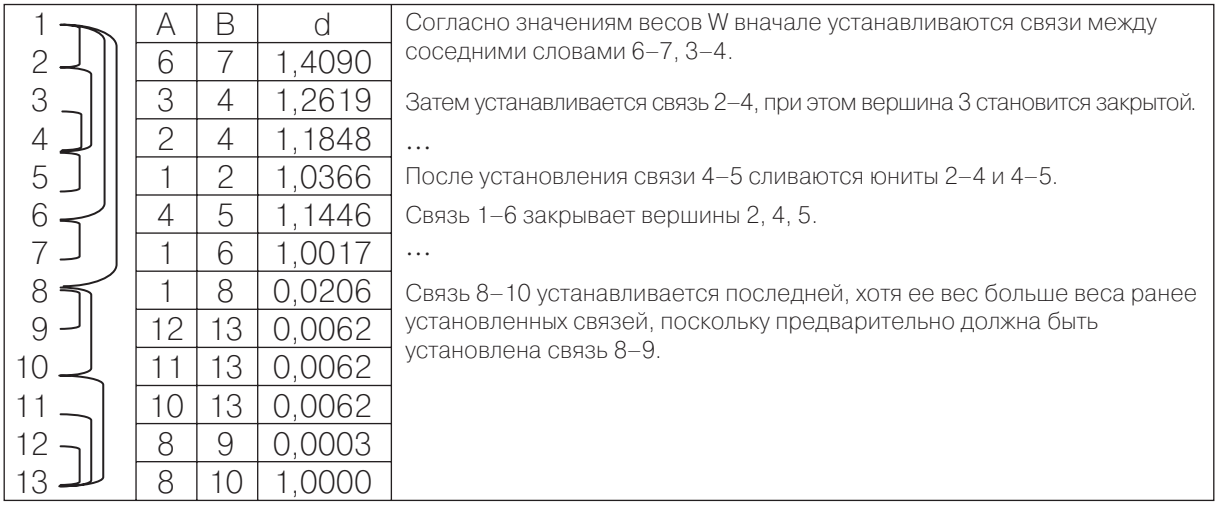


Рис. 2. Пример работы алгоритма

Итеративное обучение модели.

1. Каждое предложение тренировочного корпуса подвергается синтаксическому анализу согласно алгоритму Рис. 1. В качестве начальных значений веса зависимостей приняты величины $P(d(i,j)) = C(wi, wj, R) / C(wi, wj)$, $|i–j| < 5$ на основе собранной статистики (i) или (ii)

2. По результатам парсинга согласно (5) подсчитываются и записываются новые значения для E1, E23, E4 и E.

Выполняется парсинг каждого предложения с новыми значениями весов зависимостей.

Шаг 2 повторяется до тех пор, пока изменение весов зависимостей не станет меньше заданного порога.

3.4. Результаты экспериментов

В качестве исходного корпуса текстов выбрано собрание коротких рассказов А. П. Чехова объемом около 1 Мбайт. Общее число размеченных предложений — 12255 (исключены предложения длиной менее 3 слов). Средняя длина предложения — 15 слов. Самое длинное предложение состояло из 95 слов.

Все знаки препинания опускались.

Вид размеченного предложения представлен на Рис. 3 (рассказ «Драматург»).

Выводятся: слова предложения и номер слова, установленные связи и таблица «ЗАВИСИМОСТИ» с перечислением связей. Столбцы A и B содержат номера связанных вершин, d — вес связи (вычисленный по формуле 5), в правом столбце - флажок разрешения данной связи. Флажок позволяет исключать ложные связи в интерактивном режиме.

Приведенный пример представляет получение правильного разбора после небольшого числа итераций. Разбор большинства предложений, однако, содержит ложные зависимости, которые не устраняются после 10-й итерации.

Подсчет правильных и ложных связей выполняется обычно сравнением с «золотым стандартом», т.е. с корпусом безусловно правильно разобранных предложений. К сожалению, в настоящее время такой золотой стандарт для русского языка отсутствует в свободном доступе. Поэтому предполагается выполнить экспертную проверку деревьев разбора. При этом группе экспертов будет предложено вычеркивать ложные связи, не внося других исправлений. По результатам проверки можно будет выполнить оценку работы алгоритма, и, главное, внести изменения в веса связей, что позволит улучшить результаты анализа.

4. Заключение

Представлена модель локальных зависимостей, в которой имплицитно учтены лингвистические ограничения структуры предложения — вероятностные зависимости, которые выражают отношения между словами в предложении в виде ненаправленного проективного дерева, а также проективный характер предложения. Предложен новый алгоритм грамматического разбора, выполняющий поиск дерева зависимостей снизу вверх, устанавливая локальные связи между соседними словами и группами слов.

В цикле итерации после разбора всех предложений корпуса выполняется уточнение весов связей, затем вновь выполняется разбор и т.д.

Результаты проведенных экспериментов показывают, что результаты разбора улучшаются после нескольких итераций работы алгоритма, однако не для всех вариантов грамматической структуры и лексического состава предложений.

Имеется несколько возможностей для совершенствования модели.

В частности, при образовании очередного юнита, можно проверять, является ли он устойчивым или терминологическим словосочетанием, и обрабатывать его соответственно. Предполагается также эксплицитно включить в алгоритм проверку грамматических ограничений (напр., согласование существительного и прилагательного, запрет на связь предлога с более чем одним существительным и т.п.)

	A	B	d	
Доктор 1				
мгновенно 2	5	6	14,136	✔
проникается 3	3	4	1,7116	✔
уважением 4	8	9	1,0711	✔
к 5	2	3	1,0730	✔
пациенту 6	1	3	1,7056	✔
и 7	6	7	0,7046	✔
почтительно 8	4	6	0,7260	✔
улыбается 9	7	9	0,4719	✔
	—	—	—	**OK**

	A	B	d	
Доктор 1				
мгновенно 2	5	6	14,136	✔
проникается 3	3	4	1,7116	✔
уважением 4	8	9	1,0711	✔
к 5	2	3	1,0730	✔
пациенту 6	1	3	1,7056	✔
и 7	6	7	0,7046	✔
почтительно 8	4	6	0,7260	✔
улыбается 9	7	9	0,4719	✔
	—	—	—	**OK**

Рис. 3. Разметка предложения после 1-й и 4-й итерации

Далее, возможно преобразование ненаправленного дерева зависимостей в направленное — путем поочередного назначения каждой открытой вершины дерева (то есть, вершины, над которой не проходит ни одна связь) — корнем дерева, подсчета статистики образованных направленных связей и выбора наиболее вероятного варианта.

Поскольку модель локальных зависимостей применима к языку, основная часть предложений в котором — проективные, и благодаря высокой скорости парсинга, эту модель и алгоритм ПЛЗ можно использовать для языков с ограниченными лингвистическими ресурсами, даже в отсутствии морфологического анализатора.

Литература

1. *Billot S., Lang B.* The Structure of Shared Forests in Ambiguous Parsing // Proceedings ACL 1989.
2. *Bod R.* An all-subtrees approach to unsupervised parsing // Proceedings of COLINGACL
3. *Collins M., Hajic J., Brill E., Ramshaw L., Tillmann C.* A statistical parser for czech // Proceedings of the 37th Meeting of the Association for Computational Linguistics (ACL), pp. 505–512
4. *Gao J., Suzuki H.* Unsupervised learning of dependency structure for language modeling // ACL 2003, pp. 521–528.
5. *Goodman J.* Efficient algorithms for parsing the DOP model // Proceedings Empirical Methods in Natural Language Processing 1996, Philadelphia, PA: 143–152.
6. *McDonald R., Satta G.* On the complexity of non-projective data-driven dependency parsing // Proceedings of the International Conference on Parsing Technologies (IWPT)
7. *Nivre J.* An efficient algorithm for projective dependency parsing. // Proceedings of International Workshop on Parsing Technologies, pp. 149–160
8. *Smith D.A., Eisner J.* Bootstrapping feature-rich dependency parsers with entropic priors // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 667–677
9. *Seginer Y.* Fast Unsupervised Incremental Parsing // Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 384–391, Prague, Czech Republic, June 2007.
10. *Ножов И. М.* Реализация автоматической синтаксической сегментации русского предложения // Дисс. Канд. Техн. Наук. — М.: РГГУ, 2003
11. *Mel'čuk I.* (1988) Dependency Syntax: Theory and Practice // Albany, N.Y.: The SUNY Press, 428 pages.
12. *Ginter F., Pyysalo S., Boberg J., Salakoski T.* Regular Approximation of Link Grammar // T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 564–575, 2006.

Особенности использования многоуровневой разметки звукового корпуса unit selection в системе гибридного синтеза «Живой голос»[1]

Multi-tier markup of speech corpus for hybrid russian tts system «Vitalvoice»

Продан А. И. (prodan@speechpro.com),
Корольков Е. А. (korolkov@speechpro.com),
Опарин И. В. (ilya@speechpro.com),
Таланов А. О. (andre@speechpro.com)

ООО «Центр речевых технологий», Санкт-Петербург, Россия

Рассматривается система многоуровневой разметки звукового корпуса и её использование в системе гибридного синтеза «Живой голос» ООО «Центр речевых технологий» (ЦРТ). Система включает в себя взаимосвязанные уровни разметки, создаваемые и используемые как независимо друг от друга, так и в комплексе.

1. Введение

Система многоуровневой разметки речевого корпуса фонограмм используется при подготовке речевых данных для синтезатора речи по тексту «Живой голос» ЦРТ [1, 2].

Существует несколько подходов к организации автоматического синтеза речи по тексту. К основным можно отнести синтез по правилам (формантный синтез), артикуляторный синтез, компилятивный синтез, синтез на основании статистических моделей (HMM-синтез). На данный момент наилучшие результаты достигаются с использованием технологии Unit Selection. Данная технология позволяет достичь максимальной естественности синтезированной речи. В рамках работы по созданию новой системы синтеза русской речи, осуществляемой ЦРТ, создан синтезатор на основе использования технологии Unit Selection, совмещенной с аллофонным синтезом. Гибридный характер системы позволяет осуществлять масштабирование всей системы синтеза в зависимости от доступных ресурсов. Полноценный синтез Unit Selection, обеспечивающий наилучшее качество синтезированной речи, предполагается использовать на стационарных компьютерах; для мобильных решений предложен компромисс между качеством звучания и используемыми ресурсами памяти при помощи технологии аллофонного синтеза.

Одна из основных особенностей системы синтеза «Живой голос» — совмещение положительных сторон двух подходов — Unit Selection и компилятивного аллофонного синтеза. Таким образом, для синтеза каждым голосом подготавливаются два звуковых корпуса: аллофонный, в котором хранятся аллофоны во всех возможных контекстах, и репрезентативный речевой корпус для выбора звуковых единиц методом Unit Selection. В статье рассматривается система многоуровневой разметки корпуса для Unit Selection.

Характерной особенностью синтеза методом Unit Selection является его критическая зависимость от состава и полноты речевого корпуса. Качественный синтез возможен только на основе полного, сбалансированного и корректно размеченного речевого корпуса.

С ростом объема корпуса достигается темповая и интонационная вариативность речи диктора. Иными словами, чем больше корпус, тем больше вероятность того, что в нем найдется элемент в необходимом контексте с необходимой длительностью и контуром частоты основного тона (ЧОТ). Как следствие, меньше искажения от вынужденной модификации сигнала, а значит — выше естественность синтезируемой речи.

В целом, использование корректно размеченного, сбалансированного корпуса, есть необходимое условие для достижения высокого качества синте-

[1] Работа выполнена в рамках федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2012 годы»

зируемой речи. Известно, что качество синтеза Unit Selection не является постоянной величиной и зависит от состава синтезируемого текста. Такое свойство заложено в самой технологии. Действительно, когда выходной сигнал синтезатора составлен из оригинальных (немодифицированных) крупных фрагментов непрерывной речи, то качество речи практически совпадает с естественной. С другой стороны, когда требуется синтезировать речь по тексту, фрагменты которого представлены в корпусе лишь отдельными аллофонами, то в этом случае качество синтеза решающим образом определяется составом уровней разметки и точностью установки границ на этих уровнях. Ошибки в составе и границах обычно приводят к тому, что никакие дальнейшие усилия, связанные с модификацией речевого сигнала не в состоянии сделать синтезированную речь близкой к естественной.

2. Уровни разметки речевого корпуса Unit Selection

Основным принципом разметки корпуса является возможность учета всей информации, которая может потребоваться для синтеза. На разных уровнях разметки присутствуют как обозначения сегментных единиц: аллофонов, слогов, слов, пауз и их характеристик, так и информация более высокого уровня — об интонационном оформлении синтагм и отдельных слов, отметки об эмоциональной составляющей, и выделение неречевых явлений: смеха, кашля, заполненных пауз хезитации и т. п.

Такая система показала свою эффективность при подборе наиболее подходящих звуковых единиц, т.е. с наименьшими величинами стоимости замены и стоимости связи [3–6]: при необходимости по такой разметке из речевого корпуса можно извлечь любые единицы с заданными характеристиками.

Разметка корпуса производится в два этапа. На первом — разметка выполняется вручную с использованием специализированного звукового редактора Wave Assistant, каждый уровень хранится в отдельном текстовом файле. На рис. 1 представлен пример окна с сигналом в программе Wave Assistant, с принятыми уровнями разметки:

На втором этапе, разметка выполняется автоматически, при этом часть корпуса, размеченная вручную, используется для обучения акустических моделей системы автоматической разметки.

Далее рассмотрим каждый уровень и возможности его использования в системе синтеза более подробно.

2.1. Уровень разметки периодов основного тона

Первый и самый низкий уровень разметки — уровень периодов основного тона (ОТ). На нём каждый период основного тона обозначен метками, благодаря которым точно известна частота основного тона аллофона и скорость её изменения. Любая из меток может иметь специальный идентификатор, с помощью которого выделяются особые свойства периода. Отмечаются периоды, которые по каким-либо причинам (щелчок, какой-то посторонний короткий шум) лучше не использовать. Характер-

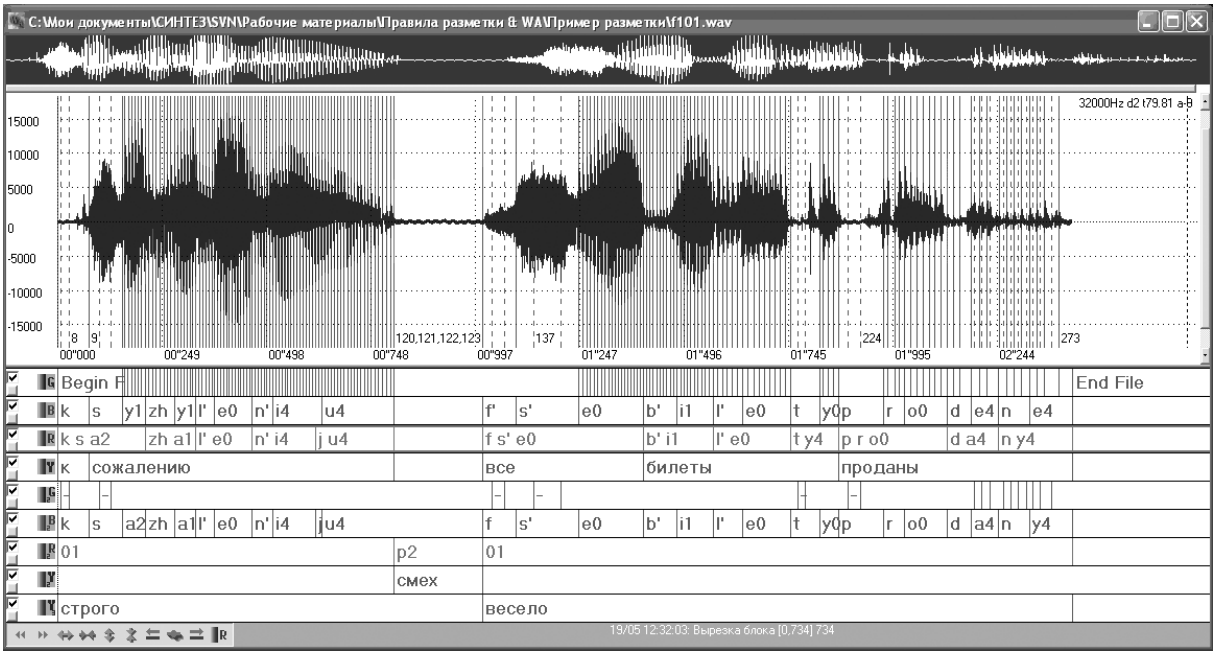


Рис. 1. Пример окна с сигналом, размеченным на всех используемых уровнях разметки

ные для звука периоды, которые при модификации речевого сигнала нельзя удалять или повторять, так же отмечаются специальным идентификатором.

Первый период после паузы и последний перед паузой также отмечаются особыми знаками, эти периоды захватывают часть паузы, благодаря чему звук при синтезе не обрывается, а естественным образом начинается или заканчивается.

На этом же уровне отмечается начало участка оглушения у звонких согласных, чтобы фрагмент после метки не принимался за обычный период ОТ, а так же, чтобы при выборе аллофона учитывалось, что согласный частично оглушён.

2.2. Уровень меток, использующихся для модификации речевого сигнала

Несмотря на то, что технология Unit Selection подразумевает выбор из звукового корпуса максимально длинных и хорошо стыкующихся между собой цепочек аллофонов без модификации, всё же в некоторых случаях минимальная модификация исходного речевого сигнала становится необходимой. Например, если длительность аллофона значительно (больше заданного порога) превышает предсказанную, подсчитанную с учётом средней длительности аллофона в корпусе, или наоборот значительно меньше её, включается модификация по длительности и этот аллофон соответственно укорачивается или удлиняется. Для этого необходимы специальные пометки, обозначающие относительно стационарные участки аллофона, которые можно сократить или наоборот частично повторить. Модификация производится и в том случае, если аллофон по частоте основного тона значительно отличается от соседних в выбранной цепочке. Обычно, если используется аллофон из аллофонного корпуса, то модификация по частоте ОТ нужна практически всегда.

Метки возможной модификации речевого сигнала используются для того, чтобы обозначить те зоны, в которых можно модифицировать исходный сигнал по длительности или частоте ОТ. Такие метки ставятся на двух уровнях: для модификации вокализованных звуков используются описанные выше метки на периодах основного тона и метки оглушения на уровне меток основного тона, а для глухих согласных или оглушённых частей звонких согласных используются дополнительные метки возможной модификации по длительности на специальном уровне.

Кроме того, на этом уровне есть возможность отметить аллофоны, которые по каким-либо причинам лучше не использовать для синтеза или те, которые лучше не брать по отдельности, а только в контексте их непосредственных соседей (например, сильно оглушённые звонкие согласные). Также на этом уровне устанавливаются другие специальные отметки, характеризующие способ сочетания элементов.

2.3. Уровни реальной и идеальной транскрипции

С целью совместить в себе возможность выбора длительных последовательностей аллофонов — при совпадении целых словосочетаний или даже фраз — и точный выбор необходимых более коротких цепочек в один-два аллофона для транскрипции используются сразу два уровня: уровень реальной и уровень идеальной транскрипции. На уровне реальной транскрипции устанавливаются реальные границы аллофонов и обозначаются именно те аллофоны, которые произнёс диктор. На уровне же идеальной транскрипции находятся аллофоны идеальной транскрипции в таком составе и порядке, в каком их генерирует автоматический транскриптор синтезатора. Метки идеальной транскрипции ставятся в соответствие меткам реальной. Если в действительности звук не реализован, то он отмечается только в идеальной транскрипции.

При значительном несовпадении — не соответствующее предсказанному место ударения, оговорка диктора — исправления вносятся в идеальную транскрипцию так, чтобы каждому аллофону на уровне идеальной транскрипции соответствовали только его возможные варианты произнесения на уровне реальной (и пропуск).

Кроме того, на уровне реальной транскрипции отмечаются такие явления, как назализованный нейтральный гласный в начале или конце фразы.

Поиск требуемых аллофонов производится по идеальной транскрипции, это позволяет найти максимально длинные последовательности аллофонов для заданного текста, причём отличия между реальной и идеальной транскрипцией учитываются в разных случаях с большим или меньшим весом. В том случае, когда требуется взять отдельный аллофон, то сразу идёт проверка по уровню реальной транскрипции.

2.4. Уровень слов

На уровне разметки слов устанавливаются метки на границах слов, идентификатором метки является само слово в орфографической записи. Данный уровень разметки создается автоматически с использованием текста, прочитанного диктором. Уровень используется для подбора аллофона по одному из параметров при расчете стоимости замены.

2.5. Уровень слогов

Уровень разметки слогов также генерируется автоматически. Слова делятся на открытые слоги. Уровень разметки на слоги используется при расчете стоимости замены. Учитывается положение нужного аллофона в слоге, количество слогов от начала синтагмы и число слогов до конца синтагмы.

2.6. Уровень интонации и пауз

В системе синтеза «Живой голос» используется список основных мелодических типов и их вариантов в звучащем тексте, созданный в СПбГУ на кафедре фонетики и методики преподавания иностранных языков [7]. В качестве основы взята расширенная классификация типов интонации Е.А.Брызгуновой [8]. Всего выделяется 13 типов интонационных конструкций с различными подтипами в зависимости от различных мелодических типов, места синтагматического ударения и т. д.

Случаи логического и эмфатического ударения, а также переноса синтагматического отмечены специальными знаками на уровне слов, что добавляет точности интонационной разметке.

В модели различается шесть типов пауз, в зависимости от завершённости или незавершённости предшествующей синтагмы, знаков препинания в исходном тексте и т.п.

Таким образом, информацию об интонационном оформлении можно получить как для отдельного аллофона, так и для всей синтагмы нужного типа в целом. Это полезно при выборе конкретного варианта интонационного оформления синтагмы: как именно он был произнесён диктором и в то же время даёт возможность получить «усреднённый» по всему корпусу вариант реализации того или иного подтипа интонационной модели. Это придаёт синтезированной речи большую естественность за счёт разнообразия её интонационного оформления.

2.7. Уровни разметки эмоциональной окрашенности

Для разметки эмоциональных модальностей и различных неречевых явлений, которые могут понадобиться для повышения естественности синтезированной речи, предназначены ещё два уровня системы разметки. На первом отмечаются явления смеха, кашля, причмокивания и т.п. На втором отмечаются эмоциональные модальности. Локализованные эмоции выделяются метками внутри синтагмы, если эмоция нелокализованная, то она задается для всей синтагмы целиком.

3. Система проверки разметки звуковой базы

Звуковой корпус размечается как вручную, так и автоматически. Лучшим компромиссом является автоматическая разметка, подстроенная под определённого диктора, то есть обученная на части материала, размеченной вручную [9]. Но ни ручная, ни автоматическая разметка никогда не дают стопроцентной точности и правильности. Неизбежны опечатки, неточно установленные границы, просто случаи, где необходимо указать вручную какие-либо особенности произнесения (для автоматической разметки) или какие-либо другие нестандартные ситуации.

Множество подобных ошибок можно найти автоматически. В ЦРТ специально для этого была разработана программа «MarkupChecker». При помощи неё проверяются на допустимость названия меток на разных уровнях и соответствия между ними. Программа не только даёт указания на явные ошибки, но также предупреждает о местах, где по каким-либо причинам ошибка является вероятной.

В данный момент автоматическая проверка ведётся по следующим параметрам:

- Проверка соответствия идеальной транскрипции в корпусе и транскрипции, полученной на выходе автоматического транскриптора, используемого в составе синтезатора.
- Проверяется наличие необходимых уровней разметки для звукового файла.
- На каждом уровне разметки производится проверка на допустимость присутствующих там обозначений (по списку).
- Производится проверка на соответствие меток начала слов и меток начала аллофонов, меток начала синтагм и меток начала слов, меток уровней идеальной и реальной транскрипции.
- Производится проверка на наличие зон возможной модификации по длительности для глухих согласных и оглушённых участков звонких.
- Производится проверка на наличие разметки по периодам основного тона для гласных и звонких согласных.
- Производится проверка на резкие изменения по длине периодов основного тона (слишком длинные или короткие периоды по сравнению с соседними).
- Производится проверка на наличие отметок пауз на уровнях слов и интонации (по соответствию меткам конца аллофона).
- Производится проверка на наличие метки в начале фразы.
- Производится проверка на наличие в слове нескольких ударных гласных и правильной расстановки степеней редукции по отношению к ударению.
- Производится проверка на допустимую разницу идеальной и реальной транскрипции (по подгружаемой таблице вариативности).
- Удаляются лишние пробелы в идентификаторах меток.
- Выдаются предупреждения о слишком больших зонах оглушения звонких звуков.

Как видно из приведённого выше списка, бóльшая часть ошибок выделяется именно благодаря сопоставлению различных уровней разметки.

4. Выводы

Предложенный способ разметки звукового корпуса для системы гибридного синтеза с использованием технологии Unit Selection является исчерпывающим, поскольку для каждой звуковой единицы корпуса и её соседей будь то период основного тона, аллофон, синтагма или фраза, обеспечивается доступ к информации на всех уровнях в комплексе. Данный способ разметки в сочетании с разработанным набором параметров для расчета минимальной стоимости замены и связи звуковых элементов позволяет получить высокое качество синтезируемой речи.

Литература

1. *Корольков Е.А., Главатских И. А., Таланов А. О., Киселев В. В., Опарин И. В.* Синтез естественной русской речи при помощи метода Unit Selection // Материалы XXXVI Международной филологической конференции.
2. *Oparin I., Talanov A.* Outline of a New Hybrid Russian TTS System // Proc. of the 12th International conference on Speech and Computer, SPECOM 2007, Moscow, Russia, 2007. Pp. 603–608.
3. *Black A. W., Hunt A. J.* Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database // In Proceedings of ICASSP 96. Atlanta, Georgia, 1996. Vol. 1, pp. 373–376.
4. *Vepa J.* Join Cost for Unit Selection Speech Synthesis // University of Edinburgh, 2004.
5. *Clark R. A. G., Richmond K., King S.* Multisyn: Open-domain unit selection for the Festival speech synthesis system // Speech Communication, 2007. Vol. 49, issue 4, pp. 317–330.
6. *Vepa J., King S.* Subjective evaluation of join cost functions used in unit selection speech synthesis // In Proceedings of the International Conference on Speech and Language Processing 2004. Jeju, Korea, 2004. Pp. 1181–1184.
7. *Вольская Н. Б., Скрелин П. А.* Моделирование интонации для синтеза речи по тексту // Уфа: 1998.
8. *Брызгунова Е. А.* Интонация // Русская грамматика. М.: 1980.
9. *Tatham M, Morton K.* Developments in Speech Synthesis // John Wiley & Sons Ltd, 2005.

Модели семантической деривации многозначных качественных прилагательных: метафора, метонимия и их взаимодействие

Semantic-derivational models of polysemous adjectives: metaphor, metonymy and their interaction

Рахилина Е. В. (rakhilina@gmail.com)
Институт русского языка РАН

Карпова О. С. (o_k_inbox.ru)
Российский государственный гуманитарный университет

Резникова Т. И. (tanja.reznikova@gmail.com)
ВИНИТИ РАН

В статье отражены результаты корпусного исследования семантических сдвигов русских качественных прилагательных в атрибутивных конструкциях. Обсуждаются регулярные модели метафорических и метонимических переносов и некоторые нестандартные случаи семантической деривации.

1. Многозначность прилагательных как прикладная задача и теоретическая проблема

Настоящая работа отражает результаты семантического исследования многозначных качественных прилагательных на материале Национального корпуса русского языка. В задачи анализа входит описание системы значений для каждой рассматриваемой лексемы, установление семантических отношений между отдельными значениями внутри лексемы, т. е. типов семантических сдвигов, и — на базе сопоставления полученных деривационных моделей для разных единиц — выявление регулярной многозначности и нестандартных семантических переходов в системе адъективной лексики.

Изучение моделей полисемии имеет богатые традиции в отечественной лингвистике, хотя прилагательные, в отличие, например, от глаголов, всегда оставались на периферии основных интересов исследователей. Среди работ, прослеживающих системную регулярность в многозначности прилагательных, отметим [Апресян 1974, Кустова 2004]. Особенность настоящего анализа заключается, однако, в том, что предметом рассмотрения для нас является вся частотная адъективная лексика со значением физических свойств и человеческих качеств, тем самым осуществимой становится задача исчерпывающей инвентаризации моделей семантической деривации русских качественных прилагательных. Кроме того, как мы надеемся показать ниже, в ходе такого «сплошного» изучения материала вскрываются нестандартные, малоизученные в лингвистической теории механизмы семантических сдвигов, — их анализ мог бы существенно расширить имеющиеся в современной науке представления о деривационных процессах в области лексики.

Итак, объектом нашего исследования являются частотные многозначные прилагательные русского языка. К частотным мы относим единицы с встречаемостью выше 2000 раз в НКРЯ (т.е. в массиве текстов объемом 160 млн. словоупотреблений). В соответствии с этим критерием было отобрано 300 полисемичных прилагательных. Теоретическая задача описания моделей семантической деривации этих лексем стала естественным развитием задачи прикладной, освещавшейся в ряде докладов на конференции «Диалог» (см. [Рахилина и др. 2006, Шеманаева и др. 2007]), — снятия многозначности в текстах НКРЯ. Эта задача в Корпусе решается при помощи создания правил-фильтров. Действие таких правил основано на том, что в контексте слово имеет только одно значение. Соответственно, если для каждого из значений данного слова удастся задать

параметры контекста, в котором это значение реализуется, то эту информацию можно затем использовать для автоматического определения значения слова в тексте.

Теоретической базой для «правилового» подхода к разрешению неоднозначности можно считать любую теорию, связывающую семантику лексемы с ее поверхностной сочетаемостью — например, исследования представителей Московской семантической школы или теорию Construction Grammar (см. [Fillmore et al. 1988, Goldberg 2005] и др.), ср. также термин *coercion* — «вынуждение» к сдвигу значения лексемы в составе комплексного выражения [Pustejovsky 1991] или когнитивный принцип единства концептуального домена (*domain unity*, см. [Croft 1993]).

Для прилагательных в атрибутивных конструкциях (*Adj + N*) контекст задается семантикой определяемого существительного (*N*). Тем самым прикладная задача разрешения многозначности в сфере адъективной лексики подразумевает, во-первых, установление набора значений каждого прилагательного и, во-вторых, определение для каждого значения допустимых при нем семантических классов существительных.

Однако с теоретической точки зрения интерес представляет не только перечисление всех засвидетельствованных в Корпусе значений лексемы, но и описание семантических отношений между этими значениями, или — иначе — типов семантических сдвигов, приводящих к появлению производных значений. Именно обсуждению этих переходов и будет посвящена основная часть данной статьи.

Согласно общепринятой теории, канонических, «законных» механизмов сдвига значения всего два — метафора и метонимия. Ниже на материале качественных прилагательных мы рассмотрим некоторые метафорические и метонимические модели семантической деривации, проследим основные закономерности развития значений в этой зоне, а также обсудим случаи, не укладывающиеся в традиционные представления о метафоре и метонимии.

2. Метафорические сдвиги в сфере качественных прилагательных

Как хорошо известно из теории концептуальной метафоры [Lakoff, Johnson 1980], метафора основана на взаимодействии двух концептуальных зон — области-источника и области-цели и заключается в проекции (mapping) элементов структуры одной области на структуру другой (ср. традиционное определение метафоры как переноса «по сходству»). В применении к прилагательным метафорический перенос, как правило, означает, что аналогия проводится между признаками объектов двух разных таксономических классов (ср. трактовку метафоры в глаголе как мену таксономического класса его актантов в [Падучева 2004, Кустова 2004]). Классы объектов, признаки которых могут сопоставляться в нашем материале, согласуются с известными и, по-видимому, типологически широко представленными моделями метафорических переходов, ср.:

- ‘человек’ — ‘механизм/инструмент’ (ср. *надежный: друг|велосипед*, *чувствительный: юноша|прибор*)
- ‘пространство’ — ‘время’ (ср. *длинный: коридор|день*, *далекий: станция|юность*)
- ‘жидкость’ — ‘множество’ (ср. *жидкий: каша|толпа*)
- ‘жидкость’ — ‘эмоциональное/ментальное состояние’ (ср. *глубокий: река|депрессия*)
- ‘конкретный предмет’ — ‘событие’ (ср. *крупный: кусок|успех*)

К частотным сдвигам относится метафорический перенос различных физических свойств неодушевленных объектов на нефизические свойства человека или его проявлений:

- *холодный* (*ветер|сердце*)
- *мягкий* (*диван|характер*)
- *яркий* (*свет|талант*)

В целом метафорические переходы, наблюдаемые в адъективной зоне, в значительной степени предсказуемы. Действительно, один и тот же признак не может применяться к сущностям разной природы, он должен «подстроиться» под семантику новой концептуальной зоны. Но сама возможность этой «подстройки» свидетельствует о «когнитивной проницаемости» границ между взаимодействующими при метафоризации концептуальными областями.

3. Метонимические сдвиги в сфере качественных прилагательных и транскатегориальная метонимия

Как в классических исследованиях [Ullman 1967], так и в работах самого последнего времени [Feyaerts 1999, Dirven 2002, Peirsman, Geeraerts 2006] метонимия описывается как семантический переход, возникающий на базе «смежности» двух ситуаций. При метонимическом сдвиге семантическое отношение между двумя значениями возникает в границах общей для них концептуальной области, и назначение механизма метонимии — в выделении, акцентировании некоторого фрагмента этой области (ср. в терминологии У. Крофта *domain highliting* [Croft 1993], ср. также синтактико-семантическую трактовку метонимии применительно к глагольной лексике в работах Е. В. Падучевой и Г. И. Кустовой как помещение в фокус, «продви-

жение» разных участников одной и той же ситуации [Падучева 2004; Кустова 2004, 2005]).

Поскольку традиционное понятие «смежности» носит довольно размытый характер, во многих работах, посвященных метонимии, предпринимаются попытки исчислить засвидетельствованные в языке модели метонимических сдвигов (из исследований последних лет наиболее полный список представлен в [Peirsman, Geeraerts 2006]). В адъективной лексике обнаруживаются реализации всех тех основных схем, которые известны по другим лексико-грамматическим классам слов, приведем примеры некоторых из них:

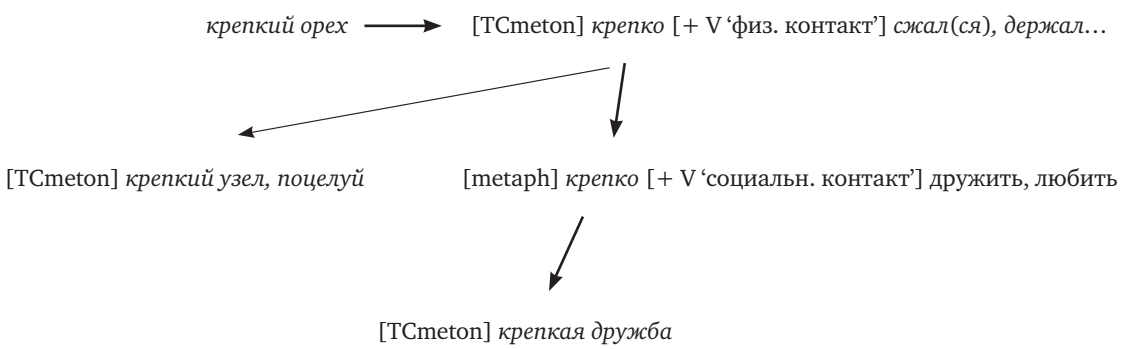
- признаки ‘части’ — ‘целого’, в т. ч.
  - ‘человека’ — ‘части тела’ (*сильный: человек|руки, добрый: человек|сердце*)
  - ‘элемента’ — ‘множества’ (*толковый: начальник|начальство*)
  - ‘материала’ — ‘изделия из этого материала’ (*ржавый: железо|гвоздь*)
- признак (какой-л. сущности) — время, когда реализуется этот признак (*голодный: человек|годы, теплый: погода|день*)
- признак (какой-л. сущности) — место, где реализуется этот признак (*голодный: человек|край, жаркий климат|страны*)
- причинно-следственные отношения
  - состояние — каузатор состояния (*радостный: человек|событие*) (ср. [Апресян1974, Кустова 1998])
  - состояние — результат состояния (*голодный: человек|обморок*) (ср. в когнитивной традиции эффект Goal-bias, который обычно связывается с глаголами движения и предпочтением в них конечной точки перед исходной, см. [Stefanowitsch, Rohde 2004])

Кроме подобных стандартных типов метонимических переходов в нашем материале мы отмечаем более сложный вид метонимии, почти не привлекавший до сих пор внимание исследователей. Речь идет о межчастеречных, или **транскатегориальных**, метонимических сдвигах (ср. [Кустова 2004: 312 и далее; Radden, Kövecses 1999]). В случае прилагательных наиболее очевидными оказываются деривационные отношения с наречиями. Действительно, во многих случаях мотивированность деривационных моделей обнаруживается только тогда, когда значения прилагательного и наречия рассматриваются в совокупности[1]. Так, *мягкий упрек* естественно связывать не с *мягкая подушка* или *мягкий характер*, а с *мягко упрекнул*. Аналогичным образом *тугая клавиша* наиболее тесно соотносится с *туго нажиматься*. В свою очередь, и для адвербиальных употреблений необходимо установить источник деривации. Например, если сочетание наречия с глаголом обозначает проявление свойства человека (ср. *строго сказать*, *щедро одарить*), то такие употребления логично считать производными от соответствующих атрибутивных конструкций, в которых реализуется значение ‘свойство человека’ (ср. *строгий начальник, щедрый меценат*).

Таким образом, последовательно восстанавливая отношения семантической деривации между отдельными значениями прилагательного и наречия, мы строим цепочки переходов, которые позволяют проводить обобщения относительно регулярных моделей многозначности в рассматриваемом фрагменте лексической системы. Примеры закономерностей такого рода мы рассмотрим в следующем разделе.

4. Цепочки семантических переходов и закономерности деривационных процессов

Приведем в качестве примера фрагмент цепочки семантической деривации для прилагательного *крепкий*:



В сочетании с конкретными объектами *крепкий* подразумевает свойство объекта, проявляющееся при внешнем воздействии, т. е. при физическом контакте

[1] Ср. в этой связи практику традиционных словарей, объединяющих прилагательные и наречия в одной словарной статье.

с воздействующим на него субъектом. Закономерно поэтому, что для наречия *крепко*, которое образуется посредством транскатегориальной метонимии, сочетаемость ограничена глаголами со значением физического контакта. Результат такого действия может описываться атрибутивной конструкцией (опять же посредством транскатегориальной метонимии), ср. *крепкий узел* — узел, который крепко завязали. Кроме того, физический контакт может метафорически переноситься в социальную сферу, тем самым порождается метафорическое значение наречия *крепко* в конструкции с глаголами социального контакта (*крепко дружить, любить*). Далее, социальный контакт может опять же описываться атрибутивно, порождая еще одну транскатегориальную метонимию — на сей раз от метафорического значения наречия.

Мы привели фрагмент схемы семантической многозначности для прилагательного *крепкий* в качестве иллюстрации того, как для каждого прилагательного в нашем материале восстанавливается последовательность семантических переходов. Анализ этих цепочек позволяет выявлять различные закономерности в устройстве деривационных моделей адъективной лексики. Обсудим некоторые правила такого рода.

- Если прилагательное в атрибутивной конструкции может выступать и в сочетании с наименованием неодушевленного конкретного объекта, и в сочетании с наименованием человека, то данные конструкции представляют разные значения прилагательного, причем эти значения связаны между собой метафорическим отношением (напрямую или через цепочку семантических переходов, включающих метафорический), ср. *вялый: цветок|мальчик, грязный двор|личность, широкий: улица|натура, слабый: человек|мотор*.

Исключение составляют конкретные неодушевленные существительные со значением частей тела (т. е. существительные класса неотторжимой принадлежности), которые обычно образуют метонимический сдвиг, ср. *добрый: человек|сердце.*

- Если исходным для деривационной цепочки является прилагательное со значением человеческого состояния или качества, то сочетаемость с абстрактным именем возникает посредством или метафоры (ср. *больной: мальчик|самолюбие*), или цепочки транскатегориальных метонимий (*строгий начальник — строго приказать — строгий приказ*).

Исключение составляет особый подкласс абстрактных существительных, ср. *голос, взгляд, походка* и под.: в конструкции с этими именами наблюдается простой метонимический сдвиг прилагательного (не через посредство наречия), ср. *больной голос, взгляд* (‘свойственный больному’). Примечательно, что эту группу можно считать аналогом класса неотторжимой принадлежности в зоне абстрактной лексики.

- Все русские многозначные прилагательные с исходным значением звука соотносятся с характеристикой человеческого голоса или с поведением массы людей, поэтому если исходным для деривационной цепочки является значение звука и при этом прилагательное допускает сочетаемость с наименованием человека, то два значения связаны метонимическим (не метафорическим!) преобразованием, ср. *громкий: голос|посетитель*, *хриплый: голос|мужик*, *шумный: компания|болельщик*.

Итак, зная исходное значение деривационной цепочки, во многих случаях можно строить довольно надежные гипотезы об устройстве модели многозначности данного слова. В этом месте, однако, естественно задаться вопросом, бывают ли такие переходы, которые невозможно предсказать, то есть сдвиги, отклоняющиеся от стандартных моделей семантических преобразований. Ответ на этот вопрос, безусловно, положительный. И здесь важно различать несколько типов таких отклонений.

Первый класс случаев представляют переходы, обусловленные внешними по отношению к современной лексической системе факторами. Это могут быть кальки или устаревшие употребления (в т. ч. фразеологизованные конструкции). Например, если в некоторой цепочке семантических переходов выпадает одно или несколько звеньев (т.е. если соответствующие сочетания выходят из употребления), то на месте возникающей лакуны образуется нестандартная для системы модель сдвига. Все эти переходы, соответственно, носят одиночный и несистемный характер (хотя каждый из них семантически мотивирован, ср. в этой связи [Dobrovol'skij, Piirainen 2005]), тем самым они в меньшей степени интересны для нашего исследования.

Гораздо больший интерес для нас представляет второй класс случаев. Эти сдвиги обусловлены действием внутрисистемных механизмов, однако модели таких переходов не укладываются в вышеописанные стандартные схемы метафорических и метонимических переносов. К рассмотрению случаев такого рода мы переходим.

5. Особый тип семантических сдвигов: *ребрендинг*

Для начала рассмотрим несколько примеров.

Прилагательное *грубый* выражает свойство человека. Метонимически это значение может переноситься на характеристику его действий, ср. *грубо разговаривать, вести себя, оскорблять*. Вполне предсказуемо в этой конструкции выступают глаголы, относящиеся к классу агентивных процессов. Напротив, сочетаемость с глаголами неконтролируемых ситуаций для *грубо* в основном невозможна,

ср. *грубо *чихает / *падает / *умирает*. Но в отдельном случае это ограничение неожиданно нарушается, ср. *грубо ошибся*. Нарушение контекстных ограничений естественным образом сопровождается резкой сменой семантики: в сочетании *грубо ошибся*, с одной стороны, пропадает идея грубости поведения, с другой — появляется значение большой степени.

Следующий пример — прилагательное *кривой*. В исходном значении оно обозначает форму объекта, отклоняющуюся от прямой линии (*сабля, улица*). При транскатегориальном метонимическом переносе наречие характеризует такое действие, при котором (или в результате которого) положение субъекта/объекта или его форма отклоняется от ожидаемой в этой ситуации прямой линии (*криво сидеть, положить, повесить, пришить, отрезать*). Тем самым при данном наречии ожидаются глаголы местоположения, помещения объекта или физического воздействия — так или иначе, результат этого действия имеет зрительный эффект, что предписывает исходная семантика прилагательного. Однако *криво* может выступать и при глаголе ментального воздействия, ср. *криво объяснять*. В принципе сам перенос из физической сферы в ментальную довольно частотен для метафорических сдвигов, однако конкретная визуальная идея, присутствующая в исходном значении прилагательного, в данном случае затрудняет стандартную метафорическую интерпретацию, требуя дополнительных когнитивных усилий. Их результатом становится исчезновение семантического компонента формы и развитие значения отрицательной оценки.

Наконец, последний пример — прилагательное *дружный*. Исходно оно характеризует множество людей, связанных определенными социальными отношениями (*дружный коллектив, семья*). Соответственно, транскатегориальный метонимический перенос дает сочетаемость с глаголами контролируемых социальных действий (*дружно играть, жить*). Но *дружно* обнаруживается и в конструкциях с глаголами неконтролируемых эмоциональных состояний и других индивидуальных ситуаций, ср. *дружно ненавидеть, врать, аплодировать*. И здесь опять же пропадает компонент исходного значения (идея дружеских отношений) и появляется новая идея одновременности действия.

Итак, во всех трех примерах мы наблюдали сходные явления: происходит нарушение ожидаемых сочетаемостных ограничений, утрачивается часть исходной семантики, возникает новое значение. Чем же мотивируется появление новой семантики? В каждом из примеров возникающее значение является выводом, или **импликатурой**, исходного значения. Так, *грубо разговаривать* предполагает идею громкости, агрессивности, из чего рождается представление об интенсивности действия и, соответственно, большой степени. *Криво пришить* означает неровность результата, из чего следует отрицательная оценка результата, и этот вывод и переносится в ментальную сферу. *Дружно играть* подразумевает успешную кооперацию в некоторой области социальной деятельности, что предполагает, что каждый из участников совершает сходные действия, откуда и рождается идея одновременного выполнения действий. Таким образом, в этих примерах перенос из одной концептуальной области в другую происходит не просто на основе сравнения, как при метафоре, это сравнение здесь осложняется действием механизма импликатуры. Такие переходы комплексной природы мы называем **ребрендингом** («*re-branding*», ср. [Резникова и др. 2008]).

Примечательно, что все отмеченные в наших примерах явления (нарушение сочетаемостных ограничений, стирание исходного значения, возникновение нового на базе конвенциализованной импликатуры) являются конституирующими для процесса грамматикализации (ср. Hopper&Traugot 1993, Bybee et al. 1994 и др.). Более того, получающаяся при таких переходах семантика (‘интенсивность’, ‘одновременность’, ‘оценка’) представляет значения из Универсального грамматического набора. Вместе с тем наш материал показывает, что грамматикализация является лишь частным случаем значительно более общих процессов, имеющих место внутри лексической системы.

Литература

1. *Апресян Ю. Д.* Лексическая семантика (синонимические средства языка). М.: Наука, 1974.
2. *Кустова Г. И.* Производные значения с экспериенциальной составляющей //Семиотика и информатика. М., 1998, N 36.
3. *Кустова Г. И.* Типы производных значений и механизмы языкового расширения. М.: Языки славянской культуры, 2004.
4. *Кустова Г. И.* О семантическом потенциале слов энергетической и экспериенциальной сферы // Вопросы языкознания, 2005, № 3.
5. *Падучева Е. В.* Динамические модели в семантике лексики. М.: Языки славянской культуры, 2004.
6. *Рахилина Е. В., Кобрицов Б. П., Кустова Г. И., Ляшевская О. Н., Шеманаева О. Ю.* Многозначность как прикладная проблема: Лексико-семантическая разметка в Национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая — 4 июня 2006 г.) / Под ред. Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. М.: Изд-во РГГУ, 2006.
7. *Резникова Т. И., Бонч-Осмоловская А. А., Рахилина Е. В.* Глаголы боли в свете Грамматики конструкций // НТИ, сер. 2, 2008, № 4. С. 7–15.
8. *Шеманаева О. Ю., Кустова Г. И., Ляшевская О. Н., Рахилина Е. В.* Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка: прилагательные // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.) / Под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. — М.: Изд-во РГГУ, 2007.
9. *Bybee J. L., Perkins R., Pagliuca W.* The evolution of grammar: Tense, aspect and modality in the languages of the world. Chicago, 1994.
10. *Croft W.* The role of domains in the interpretation of metaphors and metonymies // Cognitive Linguistics 2003, No. 4. Pp. 335–70.
11. *Dirven R.* Metonymy and metaphor: Different mental strategies of conceptualisation //R. Dirven and R. Pörings (eds.) Metaphor and Metonymy in comparison and contrast. Berlin/New York: Mouton de Gruyter, 2002. Pp. 75–111.
12. *Dobrovol'skij D., Piirainen E.* Figurative language: cross-cultural and cross-linguistic perspectives. Amsterdam; Heidelberg: Elsevier, 2005.
13. *Feyaerts K.* Metonymic Hierarchies: The Conceptualization of Stupidity in German Idiomatic Expressions // K.-U. Panther and G. Radden (eds). Metonymy in Language and Thought. Amsterdam & Philadelphia: John. Benjamins, 1999.
14. *Fillmore Ch.J., Kay P., O'Connor K. T.* Regularity and Idiomaticity in Grammatical Constructions: the Case of LET ALONE // Language, 1988, No. 64. Pp. 501–538.
15. *Goldberg A.E.* Constructions: A Construction Grammar Approach to Argument Structure. Chicago: Chicago University Press, 1995.
16. *Hopper P. J., Traugott E. C.* Grammaticalization. Cambridge, 1993.
17. *Lakoff G., Johnson M.* Metaphors We Live By. Chicago: University of Chicago Press, 1980.
18. *Peirsman Y., Geeraerts D.* Metonymy as a Prototypical category // Cognitive Linguistics 2006, 17(3). Pp. 269–316.
19. *Pustejovsky J.* The Generative Lexicon // Computational Linguistics, 2001, No. 17.4.
20. *Radden G., Kövecses Z.* Towards a Theory of Metonymy // K.-U. Panther and G. Radden (eds). Metonymy in Language and Thought. Amsterdam & Philadelphia: John. Benjamins, 1999. Pp. 17—59.
21. *Stefanowitsch A., Rohde A.* The goal bias in the encoding of motion events // K.-U. Panther and G. Radden (eds). *Motivation in Grammar.* Berlin and New York: Mouton de Gruyter, 2004. Pp. 249–268.
22. *Ullman S.* Semantics: an introduction to the science of meaning. Oxford: Blackwell, 1967.

Так называемый: семантика вводных метаязыковых оборотов[1]

The so-called: semantic analysis of Russian parenthetical metalinguistic phrases

Розина Р. И. (rarozina@yandex.ru)

Институт русского языка им. В. В. Виноградова РАН

Доклад посвящен семантике и текстовым функциям группы вводных оборотов, выражающих отношение говорящего к манере речи. В докладе доказывается, что эти обороты многозначны и что их функции в тексте изменяются в диахронии.

Доклад посвящен семантике группы вводных оборотов, выражающих «отношение к стилю, к манере речи, к характеру и способу изложения» («Грамматика — 80»). Это такие обороты, как *лучше сказать*, *иначе говоря*, *прямо сказать*, *грубо говоря*, *если можно так выразиться*, *вернее сказать*, *короче говоря*, *короче*, *одним словом*, *другими словами*, *проще сказать, попросту говоря*, *иными словами*, *как говорится*, *как теперь принято говорить*, *что называется*, *простите за выражение*, *извините*.

В посвященной этим оборотам отечественной литературе рассматриваются, главным образом, их функции в тексте.

В. В. Виноградов выделяет несколько групп вводных оборотов, модальных слов и частиц, характеризующих манеру изложения (Виноградов 1947: 736–737):

1. Модальные слова и частицы, обозначающие чужой стиль выражения, передачу чужой речи и ее оценку говорящим: (*мол*, *де*, *дескать*, *будто бы* и т.п.) В эту группу входят также продуктивные синтагмы с предлогом *по* и существительным в дательном падеже с определяющим его родительным падежом от названий лица (*по мнению такого-то*, *по словам такого-то*, *по рассказам*) и безличные и неопределенно-личные глагольные формы или синтагмы типа *говорят*, *говорит*, *передают*. Сюда же он относит встретившиеся нам в современных текстах обороты *как говорится* и *что называется*.
2. Модальные слова, содержащие оценку самого стиля, способа выражения: *буквально*, *собственно говоря*, *коротко* (*откровенно*) *говоря*. В эту же группу В.В. Виноградов включает те слова и выражения, которыми обозначается переход от одного стиля к другому: *то есть*, *иначе говоря* и т.п.
3. Модальные слова и устойчивые словосочетания, которыми обозначается характер речевой экспрессии или эмоциональный тон высказывания: *не в обиду будь сказано*, *шутка сказать*, *признаться сказать* и т.п.

В работе (Баранов, Кобозева 1984) описываются некоторые семантико-синтаксические свойства вводных слов и оборотов, — в частности их поведение в составе перформативных высказываний и способность относиться ко всему высказыванию в целом или к его части. Слова и обороты, характеризующие стиль выражения, легко употребляются в составе перформативных высказываний, не меняя их коммуникативной направленности. Что касается второй характеристики, то интересующие нас вводные слова и обороты делятся на две группы: некоторые (*иными словами*, *короче говоря*, *одним словом*) относятся ко всему высказыванию в целом, в то время как большинство слов и оборотов, характеризующих стиль изложения, относятся только к части предложения и указывают на осознание говорящим того, что для выражения своей мысли он выбирает не самые подходящие языковые средства.

[1] В статье использованы примеры из Национального корпуса русского языка.

В работе (Харченко 1984) рассматривается функционирование одного типа вводных оборотов, которые автор называет квазиплеонастическими, в научных текстах. В целом семантика оборотов и слов *иными словами*, *другими словами*, *иначе говоря*, *по-иному*, *иначе*, *грубо говоря*, *грубо*, *точнее говоря*, *образно*, *вернее говоря*, *вернее*, *лучше сказать* характеризуется как пояснительная, в рамках которой выделяются три типа значений: значение тождества, уточняющее-конкретизирующее и уточняющее-оговорочное. Для научных текстов характерно значение тождества, а для научно-популярных — другие два значения. В языке науке квазиплеонастические обороты имеют текстообразующую функцию: они тяготеют к началу абзаца, где служат средством связи между ним и предшествующим абзацем, и к концу абзаца, где устанавливают причинно-следственную связь с предшествующим контекстом.

В зарубежной лингвистике аналогичные и некоторые другие обороты и наречия (*very* 'очень', *a little bit* 'немного', *largely* 'в большой степени', *rather* 'скорее', *kind of* , sort of 'вроде' и т.п.) принято объединять термином «ограничители» или «загородки» (hedges) (Lakoff 1972; перевод по Баранов А.Н., Добровольский Д.О. и др. 2001). У загородок выделяется целый ряд функций: смягчить высказывание, например *I think you're wrong* 'Думаю, ты неправа' вместо категоричного *You're wrong* 'Ты неправа'; сделать членство в множестве менее определенным, например *A penguin is sort of a bird* 'Пингвин вроде бы птица'; указать на отношение говорящего к содержанию пропозиции, например *This hypothesis might be / perhaps is too far-fetched* 'Эта гипотеза, возможно / может быть преждевременна', однако главная функция «загородок» — в какой-то степени снять с говорящего ответственность за высказывание (Hedging: 1996).

В докладе будет показано, что исследуемые русские обороты могут выполнять в тексте различные функции и что им свойственна многозначность.

Проще говоря

Оборот *проще говоря* выступает в качестве связующего звена между двумя различающимися по стилю синонимами. При этом *проще говоря* может сочетаться с союзами: *или*, *то есть* и *а.* Таким образом, есть четыре варианта конструкции с оборотом проще *говоря*:

1. бессоюзная: *А, проще говоря В*, например:

(1) Мама была педиатром. *Проще говоря*, детским врачом (Алексин Анатолий. Сигнальщики и горнисты (1985)).

2. с союзом *или*: *А или, проще говоря, В*, например:

(2) Но у них есть предел моральных устремлений, высокий моральный потолок, или, *проще говоря*, идеал (Лидия Гинзбург. Записные книжки. Воспоминания. Эссе (1920–1943)).

3. с союзом *то есть*: *А , то есть, проще говоря, В*:

(3) Сколько я вас понимаю, вы смотрите на них как на претендентов. .. то есть, *проще говоря*, как на женихов? (Горький Максим. Мужик (1899))

4. С союзом *а*: *А, а проще говоря, В*:

(4) Если тот или иной гаджет способен воспроизводить звук, то его источник (а *проще говоря* — динамик) должен выделяться из корпуса и, желательно, зримо вибрировать. (Полина Стечкина. Гаджеты и гаджетины // «Бизнес-журнал», 2004)

Эти конструкции несколько различаются с точки зрения характера синонимов,

переход между которыми осуществляет оборот *проще говоря*: в конструкции с союзом *или* левый синоним — дескрипция, имеющая высокую (книжную) окраску, а правый — нейтральный универб, например:

(5) Либеральному строю обществ и либеральному движению умов всегда и везде сопутствует в сфере экономической господство ***подвижного капитала*** или, *проще говоря,* ***денег***. (К. Н. Леонтьев. А. И. Кошелев и община в московском журнале «Русская мысль» (1880)).

В конструкции с *то есть* левое слово — книжная косвенная номинация, а правое — нейтральная прямая (ср. тот же пример из Горького):

(6) Сколько я вас понимаю, вы смотрите на них как на ***претендентов***. то есть, *проще говоря*, как на ***женихов***? [Горький Максим. Мужик (1899)]

В конструкции с союзом *а* в качестве левого члена также может выступать дескрипция, а правого — ее однословный эквивалент, но особенность этой конструкции — в том , что в ней эквивалентом левого синонима может выступать не нейтральное , а разговорное или даже сленговое слово, ср.:

(7) Книга целиком и полностью посвящена ***внутрисемейной сексуальности,*** *а проще говоря — последствиям* ***траха с папашами, сестрами и прочими родственниками***. (Чтиво // «Хулиган», 2004)

Как кажется, функции оборота *проще говоря* в разных конструкциях различаются. В конструкциях с союзами *или* и *то есть* оборот *проще говоря* служит сближению синонимов, он устанавливает эквивалентность между ними. Между тем в конструкции с союзом *а* и в бессоюзной конструкции оборот *проще говоря* усиливает противопоставление между синонимами.

Интересно рассмотреть функционирование оборота *проще говоря* в диахронии. Если условно разделить все время существования этого оборота на три периода — досоветский (1800–1917 гг.), советский (1918–1985 гг.) и постсоветский (1986–2008), оказывается, что в досоветский период употреблялись только две конструкции — с союзом *или* и с союзом *то есть*. Оборот *проще говоря* использовался при этом для перехода от книжной лексики к нейтральной, например:

(8) Сколько я вас понимаю, вы смотрите на них как на ***претендентов…*** то есть, *проще говоря*, как на ***женихов***? [Горький Максим. Мужик (1899)]

от терминологии к обыденной лексике:

(9) Либеральному строю обществ и либеральному движению умов всегда и везде сопутствует в сфере экономической господство **подвижного капитала** или, *проще говоря*, денег. [К. Н. Леонтьев. А. И. Кошелев и община в московском журнале «Русская мысль» (1880)]

В советский период на первый план выступает бессоюзная конструкция, конструкция с *или* и *а* встречается редко, а конструкция с *то есть* вообще перестает употребляться. *Проще говоря* продолжает использоваться для перехода от терминологии к обыденной лексике, ср.:

(10) Мама была ***педиатром***. *Проще говоря*, ***детским врачом*** [А. Алексин. Сигнальщики и горнисты (1985)]

Но гораздо чаще *проще говоря* осуществляет переход от дескрипции, имеющей книжный характер, к нейтральному однословному обозначению, ср. :

(11) ***Средство существования***, *проще говоря* — ***деньги***. [Владимир Толстой, Иван Богачев. Нелишние деньги // «Студенческий меридиан», 1984]

(12) ***Героическим деянием***, или, *проще говоря*, ***подвигом***, мы с вами назовем только такой экстраординарный по своей смелости поступок, который служит нравственно близкой нам цели. [Крон Александр. Капитан дальнего плавания (1983)].

В основе всех этих противопоставлений мог лежать протест против советского бюрократического стиля. Противоположную тенденцию — стремление употреблять официальную лексику вместо нейтральной — высмеял в свое время К. Чуковский в книге «Живой как жизнь», ср. :

(13) Баня не работает, т.е. не *функционирует*.

В постсоветский период продолжают использоваться все эти три конструкции, но особенностью их функционирования является то, что в качестве правого члена конструкции может использоваться сниженная лексика и сленг, как в примере (8).

Попросту говоря

Примерно такую же, хотя менее четкую, картину дает очень близкий обороту *проще сказать* оборот *попросту говоря*. Он также употребляется в четырех конструкциях:

1. бессоюзной: *А, попросту говоря В,* например:

(14) Пол имел форму вогнутой поверхности, *попросту говоря*, ямы (Александр Житинский. Лестница (1972)).

2. с союзом *или*: *А или, попросту говоря, В*, например:

(15) На практике это означает, что национальной сборной придали статус второй, а бывших «лучших из лучших всея Руси» низвергли в тартарары вместе с отцом-основателем. Или, *попросту говоря*, кинули.

3. с союзом то есть: А , то есть, попросту говоря, В:

(16) Василий Борисыч, помазав власы своя елеем, то есть, *попросту говоря*, деревянным маслом, надев легонький демикотоновый кафтанчик и расчесав реденькую бородку, петушком прилетел в келарню добродушной Виринеи. [П. И. Мельников-Печерский. В лесах. Книга первая (1871–1874)

4. С союзом *а*: *А, а попросту В*:

(17) Но особо опасны массивные алкогольные эксцессы, а *попросту говоря*, запои . (А. Ф. Блюгер. Мишень для алкоголя // «Химия и жизнь», 1985).

И в диахронии конструкции с *попросту говоря* распределяются примерно так же, как конструкции с *проще говоря*.

В конструкции с союзами *то есть* и *или* оборот *попросту говоря* встречается только в досоветский период. В советский и постсоветский период встречается бессоюзная конструкция и конструкция с союзом *а*. Для постсоветского времени характерно общее снижение стиля: в качестве второго члена конструкции часто выступает сленг. Кроме того, в постсоветское время появляется конструкция без первого члена пары, а второй при этом является сленговым или жаргонным словом, ср.:

(18) Если без романтического налета, то шли ребятки, *попросту говоря*, на «мокрое дело» — бабушку Оленьки укокошить. (Как внучка бабушку заказала // "Сельская новь", 2003)

Грубо говоря

Первые примеры оборота *грубо говоря* встречаются в 19-м веке, но до 1917 г. он употребляется крайне редко (по данным Национального корпуса русского языка — всего дважды, у Вельтмана (1835 г.) и у Л. Н. Толстого (1894). В этот период он имеет значение ‘упрощая, представляя ситуацию более простой, чем есть на самом деле’, ср.:

(19) Дело просто, ясно и, *грубо говоря*, ведь вот в чем: Ты вошла в исключительно близкие отношения, в те отношения, в которые входят только с людьми, которых любят любовью [Л. Н. Толстой. Письма. 1894 (1894)]

Оборот *грубо говоря* становится гораздо более частотным в советское время. Главным образом, о н продолжает употребляться в прежнем значении упрощения ситуации, ср. (21), но кроме того, приобретает еще одно значение — ‘предупреждает о том, что будет использовано стилистически сниженное выражение’, ср. (22):

(20) *Грубо говоря*, все это вместе могло быть названо счастьем. [Юрий Визбор. Завтрак с видом на Эльбрус (1983)]

(21) Эти предприятия культурно-бытового обслуживания, *грубо говоря*, тащатся за хвостом у государства. [И. Грекова. Дамский мастер (1963)] .

Значение упрощения ситуации сохраняется и в постсоветский период, ср.:

(22) *Грубо говоря*, на входе имеем тушку птицы, а на выходе получаем готовый продукт: мясные полуфабрикаты, колбасы, сосиски, пельмени, копчености. [Господин строитель // «Пермский строитель», 2004.04.27]

но кроме того, оборот *грубо говоря вводит* в речь нецензурную лексику и сленг, ср.:

(23) То есть они вот отсюда выезжают / и вот пошли / *грубо говоря* / на.

в годы реформ так называемых / мягко выражаясь и *грубо говоря* / вот этого беспредела / это невозможно. [Беседа с социологом на общественно-политические темы, Москва // ФОМ (2004)].

Эти два значения оборота *грубо говоря* связаны метонимически по модели ‘упростить ситуацию — упростить средства описания ситуации’. Насколько уникальна такая связь значений?

Обратимся снова к обороту *попросту говоря*. Есть контексты, в которых значение оборота *попросту говоря* размыто и совмещает идею упрощения средств выражения и упрощения ситуации, например:

(24) Чтобы понимать смысл аксиомы, мы должны иметь представление о смысле участвующих в аксиомах понятиях — *говоря попросту*, понимать, что эти понятия означают. [Владимир Успенский. Из книги «Что такое аксиоматический метод?» (2002)].

Таким образом, здесь, скорее всего, мы имеем дело с регулярной полисемией, оформившейся в случае *грубо говоря* и неоформленной в случае *попросту говоря*.

Так называемый

Словосочетание *так называемый* отличается от оборотов, которые мы рассмотрели выше: оно помогает осуществить переход не «сверху вниз», от более сложного, книжного, к простому — нейтральному или сниженному (разговорному или сленговому) слову, а «снизу вверх», от нейтральной лексики к книжной.

По данным Национального корпуса русс кого языка первые употребления словосочетания *так называемый* встречаются, начиная с 1750 г. Чаще всего оно выполняет функцию перехода от общелитературной лексики к терминологической, ср.:

(25) Рядом с нею стоял плетёный *сарайчик, так называемый амшаник*, куда ставят улья на зиму. [И. С. Тургенев. Живые мощи (1874)]

Во многих примерах того времени словосочетание *так называемый* вводит в речь термин, не предваряя его общелитературным синонимом. Функция словосочетания в этом случае — предупреждение читателя о том, что в общелитературном контексте будет употреблен термин, ср.:

(26) А главное — крокодил есть собственность, стало быть, тут уже *так называемый экономический принцип* в действии [Ф. М. Достоевский. Крокодил (1865)].

Словосочетание *так называемый* сигнализирует и о том, что автор не считает термин принадлежащим своему языку, то есть о том, что в текст вводится чужая речь. Значение словосочетания *так называемый* в этом случае — 'так принято говорить / так говорят другие; я так не говорю'. Хорошая иллюстрация такого употребления — следующий пример из Бехтерева, критически относившегося к спиритическим сеансам:

(27) При этом выяснилось, что такое отгадывание, если и может быть осуществлено, то эта способность обнаруживается обыкновенно в особом состоянии человека, которое принято называть гипнотическим, ибо и *так называемый транс спиритов* должен быть понимаем как состояние гипнотическое, или гипноидное. [В. М. Бехтерев. Внушение и его роль в общественной жизни (1898–1925)]

Другая функция словосочетания *так называемый* в досоветский период — переход от имени нарицательного к имени собственному в тех случаях, когда оно известно только узкому кругу людей, например:

(28) Теперь уже эта посадка — *старый березовый, так называемый Абрамовский лес*. [Т. Л. Сухотина-Толстая. Детство Тани Толстой в Ясной поляне (1910–1950)]

(29) На другой день я съехал с крейсера «Генерал Корнилов» и переехал в приготовленное для меня помещение в город, *так называемый «Малый дворец»* [П. Н. Врангель. Записки (1916–1921)]

В некоторых случаях функция словосочетания так называемый — показать, что говорящий оценивает денотат отрицательно и считает его не соответствующим его принятому обозначению, например:

(30) Среди всех их, должно быть в насмешку над богом и Христом, стоит также длинноволосый и в парчовой ризе *так называемый священник*. [Л. Н. Толстой. Не могу молчать (1-я редакция) (1908)]

(31) После 17 октября во время нашей, *так называемой революции*, когда забрали громадную силу в России « черносотенцы », эти архиреакционеры, состоящее большею частью из лиц sans foi ni loi, и образовался, *так называемый, союз русского народа*, с его различными подотделами, союзом Михаила Архангела и т. д. [С. Ю. Витте. Воспоминания (1911)] .

Словосочетание *так называемый* встречается в этот период и в научных и научно-популярных текстах, и в этом случае мы сталкиваемся с еще одной его функцией: оно указывает на метафоричность термина, часто противопоставляя «народный» термин научному, ср.:

(32) Очень вреден вследствие огромного содержания окиси углерода *так называемый «водяной»* (гидрокарболовый) газ [Ф. Ф. Эрисман. Профессиональная гигиена (1871–1908)].

(33) — Если вы приставите стетоскоп к груди больного, — объясняет ассистент, — и в то же время будете постукивать рядом ручкою молоточка по плессиметру, то услышите ясный, металлический, *так называемый* «амфорический» звук… [В. В. Вересаев. Записки врача (1895–1900)]

Кроме того, словосочетание *так называемый* может вводить в текст заимствованное слово или же слово чужого языка:

(34) Наловил Проклятов много красной рыбы на веку своем; <…> случалось ему попадать на царское багренье, с которого отправляют, по древнему обычаю, ежегодно на почтовых тройках царский кус, или *так называемый презент* [В. И. Даль. Уральский казак (1843)]

В советский период некоторые функции этого словосочетания сохраняются. Оно по-прежнему сигнализирует об употреблении в тексте термина:

(35) …молодой человек с зеркальным пробором и лицом сукина сына, *так называемый* ***крупье***, раскладывал лопаткой с длинной ручкой ставки и запускал белый шарик в карусель крутящегося рулеточного аппарата с никелированными ручками([Валентин Катаев. Алмазный мой венец (1975–1977))

Так же, как и в предыдущий период, оно может выражать отрицательную оценку денотата и указывать на то, что автор текста или говорящий не считает его соответствующим его названию', ср.:

(36) Но вот начал трещать по швам стянутый оковами Варшавского Договора и державшийся на штыках Групп советских войск *так называемый* «социалистический лагерь». [Светлана Алексиевич. Цинковые мальчики (1984–1994)]

Оно также вводит заимствованное слово или просто слово чужого языка, ср.:

(37) …арена с большими трибунами, бассейн и *так называемый* «рикриэйшн сентр» — своеобраз-

ный клуб для плаванья, игр и валянья на траве (Василий Аксенов. Круглые сутки нон-стоп // «Нов. мир», № 8, 1976).

В постсоветское время эти функции словосочетания сохраняются, но на их фоне появляется и кардинально новая, противоречащая всем, до сих пор свойственным ему функциям: словосочетание *так называемый* начинает использоваться для введения в речь сниженной лексики, ср.:

(38) Я вернулся в свою камеру. ***«Третьяк»*** *так называемый*, для таких, тяжело-статейных людей, и буквально через час мне через кормушку дежурный сказал — Савенко, утром на завтра — с вещами.

(39) Нужно разбирать каждый случай и так называемого *наезда* и так не называемого *рейдерства* и захвата разбирать (ЭМ 18.10.2008).

Развитие этой функции у словосочетания *так называемый* можно связать с общей тенденцией развития вводных и модальных оборотов, которые служили для перехода от книжной лексики к нейтральной, а в последние годы стали вводить сленг..

Итак, можно говорить о некоторых общих тенденциях в развитии семантики модальных оборотов и их функционирования в тексте. У всех рассмотренных оборотов можно выделить два типа значений. В значении одного типа эти обороты относятся к ситуации, о которой идет речь; в значении другого типа — к средствам описания ситуации. В досоветский период они служат переходу от книжной лексики к нейтральной и часто имеют функцию объяснения. В постсоветский период они используются для перехода к сниженной лексике и сленгу.

Литература

1. *Баранов А. Н., Кобозева И. М.* Вводные слова в семантической структуре предложения // Системный анализ значимых единиц русского языка. Синтаксические структуры. Красноярск, 1984, с. 83–93.
2. *Баранов А. Н., Добровольский Д. О., Михайлов М. Н., Паршин П. Б., Романова О. И.* Англо-русский словарь по лингвистике и семиотике. М, 2001.
3. *Вежбицка А.* Метатекст в тексте // Новое в зарубежной лингвистике. Сост. Т. Николаева. М., 1978, вып. 8.
4. *Виноградов В. В.* Русский язык (грамматическое учение о слове) // М.: Государственное педагогическое издательство Министерства просвещения РСФСР, 1947.
5. *Русская* грамматика. Под ред. Н.Ю. Шведовой. М.: Наука, 1980.
6. *Харченко Н. П.* Квазиплеонастические обороты в языке науки // Системный анализ значимых единиц русского языка. Синтаксические структуры. Красноярск, 1984, с. 93–98.
7. *Lakoff G.* «Hedges: A study of meaning criteria and the logic of fuzzy concepts» // Papers from the Eighth Regional Meeting of Chicago Linguistic Society. Chicago, 1972, p. 183–228.
8. *Hedging:* A Challenge for Pragmatics and Discourse Analysis. Ed. by R.Markkanen, H. Schröder . Berlin: De Gruyter, 1996.

Идентификация автора текста с помощью аппарата опорных векторов в случае двух возможных альтернатив

Authorship identification with support vector machine in case of two possible alternatives

Романов А. С. (ras@ms.tusur.ru), **Мещеряков Р. В.** (mrv@keva.tusur.ru)

ГОУ ВПО «Томский государственный университет систем управления и радиоэлектроники», Томск

В статье проблема идентификации автора текста рассматривается как задача классификации. Обоснована важность решения задачи бинарной классификации для идентификации автора. Приведено описание и результаты экспериментов по идентификации автора текста с помощью аппарата опорных векторов в случае двух возможных альтернатив.

1. Постановка задачи

Проблему идентификации автора текста при ограниченном наборе альтернатив сформулируем следующем образом. Имеется множество текстов $T = \{t_1,\ldots,t_k\}$ и множество авторов $A = \{a_1,\ldots,a_n\}$. Для некоторого подмножества текстов $T' \subseteq T$ авторы известны $D = \{(t_i, a_i)\}_{i=1}^{\ell}$. Необходимо установить, кто из множества A является истинным автором остальных текстов (анонимных или спорных) $T'' = \{t_{|T'|+1},\ldots,t_k\} \subseteq T$.

В данной постановке задачу идентификации автора можно рассматривать как задачу классификации с несколькими классами [1, 2]. В этом случае множество A составляет множество предопределенных классов и их меток, D — обучающие примеры, а множество T'' — классифицируемые объекты. Целью является построение классификатора, решающего данную задачу, т.е. нахождение некоторой целевой функции $F : T \times A \to [0,1]$, относящей произвольный текст множества T к его истинному автору. Значения функции интерпретируется как степень принадлежности объекта классу: 1 соответствует положительному решению, 0 — отрицательному.

Задачу многоклассовой классификации можно свести к решению нескольких бинарных задач. Для этого существуют следующие основные стратегии выбора решения [3]:

- «Один против всех» (one-against-all). Для решения задачи строится n классификаторов таким образом, что каждый класс a_j сопоставляется с остальными $(n-1)$ классами, т.е. в каждом из j случаев выбор осуществляется из двух вариантов: «класс a_j» и «не класс a_j». Итоговое решение по всем классам принимается по схеме «победитель забирает всё» (winner takes all) — победителем считается класс, имеющий максимальное значение функции F.
- «Каждый против каждого» (one-against-one). Классификаторы строятся для каждой пары классов для того, чтобы можно было однозначно разделить любые два класса из множества A. Количество классификаторов в этом случае равно $n(n-1)/2$. После подачи на входы каждого из обученных классификаторов тестового образца получаем ответы, содержащие информацию о его принадлежности одному из двух классов, участвовавших в обучении. К полученному множеству ответов применяется схема мажоритарного голосования и класс, выбранный большинством классификаторов, принимается как итоговое решение.
- Ориентированный ациклический граф (DNA). На этапе обучения работает также как стратегия «каждый против каждого». На этапе тестирования и непосредственной классификации используется корневой бинарный ориентированный ациклический граф (ориентированное дерево) с $n(n-1)/2$ внутренними узлами — обученными бинарными классификаторами, и n листьями. Классифицируемый объект проходит путь от корня до одного из листьев, при этом в зависимости от результатов классификации в каждом узле один из классов отвергается, и дальнейшие действия продолжаются по ветке, соответствующей второму классу. После выполнения $(n-1)$ подобных операций, алгоритм достигает листа, который принимается как итоговое решение классификатора.

Таким образом, для того, чтобы классификация по нескольким классам проходила успешно, необходимо в первую очередь добиться высокой точности при решении задач бинарной классификации. Важными этапами при этом являются выбор алгоритма классификации и его параметров, количества обучающих примеров, а также выбор характеристик текста для анализа и необходимого объема выборок.

2. Классификатор на основе машины опорных векторов

В исследованиях используется классификатор на основе метода «машины опорных векторов» (Support Vector Machine, SVM), математический аппарат которого был предложен В.Н. Вапником в работах [4, 5] и одна из его популярных реализаций — библиотека libsvm [6]. Исследования отечественных и зарубежных авторов [1, 7] показывают, что SVM, на сегодняшний день, является одним из лучших методов классификации.

Пусть имеется помеченное тренировочное множество примеров $D = \{(x_i, y_i)\}_{i=1}^{\ell}$, $x_i \in X \subset R^d$, метки могут принимать значения $y_i \in Y = \{-1,+1\}$. SVM строит линейный классификатор в пространстве признаков с высокой размерностью таким образом, чтобы зазор между граничными точками двух классов, называемых опорными векторами, был максимальным. Для отображения исходных данных в пространство, в котором разделяющая их поверхность будет линейной, используются ядровые преобразования — некоторая функция

$$(\Phi(x), \Phi(x')) = k(x, x').$$

Классифицирующая функция, реализуемая SVM, записывается следующим образом:

$$f(x) = \left\{ \sum_{i=1}^{\ell} \alpha_i y_i k(x_i, x) \right\} + b$$

Чтобы найти оптимальный коэффициент α достаточно максимизировать функционал

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j k(x_i, x_j)$$

в положительном квадранте $0 \le \alpha_i \le C$, $i = \overline{1, \ell}$. Условие максимизации:

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0.$$

Параметр регуляризации *C* отвечает за соотношение между величиной зазора и количеством ошибок обучающего множества.

Необходимыми условиями решения задачи нелинейного программирования являются условия Каруша-Куна-Такера:

$$\alpha_i = 0 \Rightarrow y_i f(x_i) \ge 1,$$
$$0 < \alpha_i < C \Rightarrow y_i f(x_i) = 1,$$
$$\alpha_i = C \Rightarrow y_i f(x_i) \le 1.$$

Эти условия удовлетворяют множеству допустимых множителей Лагранжа $\alpha^0 = \{\alpha_1^0, \alpha_2^0, ..., \alpha_\ell^0\}$, максимизирующих целевую функцию $W(\alpha)$. Параметр смещения *b* выбирается, чтобы обеспечить выполнение второго условия Каруша-Куна-Такера для всех входных образцов, соответствующих множителям Лагранжа, лежащим не на границах. В общем случае только часть множителей Лагранжа α будет иметь ненулевые значения, они и составляют опорные вектора.

Пусть *I* — множество индексов образцов, относящихся к множителям Лагранжа, лежащим внутри границ:

$$I = \{i : 0 < \alpha_i^0 < C\},$$

а *J* множество индексов со значениями множителей Лагранжа, лежащих на верхней границе *C*:

$$J = \{i : \alpha_i^0 = C\},$$

тогда можно переписать следующим образом:

$$f(x) = \left\{ \sum_{i \in \{I,J\}} \alpha_i^0 y_i k(x_i, x) \right\} + b.$$

В отличие от искусственных нейронных сетей, применявшихся авторами ранее [8], SVM лучше подходит для работы с большим признаковым пространством, что важно при использовании *N*-граммных признаков текста. Нет необходимости в выборе количества скрытых элементов, скорость работы SVM существенно выше, чем нейронных сетей.

Программная реализация метода интегрирована в общую программную оболочку, описанную в работе [9].

3. Описание экспериментов

Для оценки точности классификатора в случае двух предполагаемых авторов были проведены эксперименты на корпусе, составленном из 47 текстов 11 русских писателей (см. табл. 1), взятых с Интернет-ресурса [10]. Количество обучающих примеров выбиралось исходя из потребностей при решении реальных задач идентификации авторства, когда количество материала ограничено. Использовались выборки объемом 1000–100 000 символов (~200–20 000 слов), количество обучающих примеров для каждого автора бралось равным 3, для тестирования использовалось по 1 выборке автора. В табл. 2 представлены признаки текста (частоты встречаемости тех или иных групп символов и слов), использованные в экспериментах.

Таблица 1. Корпус текстов для исследований

Автор	Название произведения
Айтматов Ч. Т.	«Белое облако Чингисхана»
	«Пегий пес, бегущий краем моря»
	«Плаха»
	«Прощай, Гульсары!»
Акунин Б.	«Азазель»
	«Пелагея и Бульдог»
	«Внеклассное чтение»
	«Статский советник»
Астафьев В. П.	«Печальный детектив»
	«Так хочется жить»
	«Царь-рыба»
	«Ода русскому огороду»
	«Жизнь прожить»
Беляев А. Р.	«Голова профессора Доуэля»
	«Остров погибших кораблей»
	«Светопреставление»
	«Последний человек из Атлантиды»
	«Человек Амфибия»
Булгаков М. А.	«Мастер и Маргарита»
	«Собачье сердце»
	«Театральный роман»
	«Белая гвардия»
Достоевский Ф. М.	«Братья Карамазовы»
	«Преступление и наказание»
	«Идиот»
	«Бесы»
Горький М.	«Дело Артамоновых»
	«Мать»
	«Фома Гордеев»
	«Коновалов»
Тургенев И. С.	«Дворянское гнездо»
	«Отцы и дети»
	«Накануне»
	«Рудин»
Набоков В. В.	«Лолита»
	«Защита Лужина»
	«Дар»
	«Король, дама, валет»
Распутин В. Г.	«Деньги для Марии»
	«Пожар»
	«Прощание с Матерой»
	«Живи и помни»
Булычев К.	«Лиловый шар»
	«Любимец»
	«Марсианское зелье»
	«Подземелье ведьм»
	«Смерть этажом ниже»

Таблица 2. Исследованные признаки текста

Обозначение признака	Расшифровка
УНИГРАММЫ	Буквы русского алфавита
УСЛОВНЫЕ	Условные вероятности появления одной буквы после другой
БИГРАММЫ	Пары букв русского алфавита
БИГРАММЫ_ГЛ	Биграммы, состоящие только из гласных
БИГРАММЫ_СГЛ	Биграммы, состоящие только из согласных
БИГРАММЫ_ВЧ	Биграммы с высокой частотой встречаемости
БИГРАММЫ_СЧ	Биграммы со средней частотой встречаемости
БИГРАММЫ_НЧ	Биграммы с низкой частотой встречаемости
БИГРАММЫ_100	100 наиболее частых биграмм
ТРИГРАММЫ	Тройки букв русского алфавита
ТРИГРАММЫ_100	100 наиболее частых триграмм
ТРИГРАММЫ_500	500 наиболее частых триграмм
ТРИГРАММЫ_1000	1000 наиболее частых триграмм
ТРИГРАММЫ_2000	2000 наиболее частых триграмм
ТРИГРАММЫ_3000	3000 наиболее частых триграмм
ТРИГРАММЫ_ВЧ	Триграммы с высокой частотой встречаемости
ТРИГРАММЫ_СЧ	Триграммы со средней частотой встречаемости
ТРИГРАММЫ_НЧ	Триграммы с низкой частотой встречаемости
ШАРОВ	Частоты всех слов из словаря Шарова [11]
ФОМЕНКО	Частоты «опорных слов» Фоменко [12]
ШАРОВ_100	100 наиболее частых слов из словаря Шарова
ШАРОВ_500	500 наиболее частых слов из словаря Широва
ШАРОВ_1000	1000 наиболее частых слов из словаря Шарова
ШАРОВ_2000	2000 наиболее частых слов из словаря Шарова

При использовании метода на основе частотного словаря С. А. Шарова все слова были приведены к нормальной форме с помощью алгоритма стемминга Snowball для русского языка [13].

Для получения характеристик с разделением по частоте встречаемости был проведен их частотный анализ для всех имеющихся текстов. Би-

граммы, триграммы и слова были упорядочены по частоте встречаемости в убывающем порядке. Часть биграмм и триграмм символов была отсеяна как нехарактерная для русского языка и как шумы, связанные с автоматической обработкой текстов. Границами для характеристик с высокой частотой (ВЧ) выбраны квантили уровней 0,66 и 1, для характеристик с низкой частотой (НЧ) — квантили уровней 0 и 0,33, и квантили уровней 0,33 и 0,66 — для характеристик со средней частотой (СЧ).

Параметры обучения моделей SVM были выбраны следующие:

- ядро на основе радиальных базисных функций (RBF): $k(t,t') = e^{-\gamma\|x-x'\|^2}$;
- значение параметра гамма $\gamma = 0{,}5$; — значение параметра регуляризации $C = 1$.

Последовательность шагов проведения экспериментов для оценки точности классификации по двум авторам приведена ниже.

1. Выбор параметров обучения моделей SVM, параметров текста для исследований.
2. Применение к каждому тексту операции «склеивания»: все слова приводятся к нижнему регистру, буква «ё» заменялась буквой «е», из текста удаляются все символы форматирования и пунктуации, включая пробел (это позволяет учитывать при анализе также и соединительные биграммы на границе двух слов).
3. Формирование пар классов из всего множества авторов (в данном случае количество пар классов равно $C_{11}^2 = 55$).
4. Для каждого автора из текущей пары формируется по 3 обучающих выборки необходимого объема и одна тестовая. Выборки извлекаются из разных текстов автора.
5. Подсчет параметров в выборках.
6. Нормирование параметров выборок в диапазон [−1..1].
7. Обучение модели SVM на данных пары выборок.
8. Подача на вход обученной модели SVM данных тестовых выборок, работа классификатора, считывание результатов.
9. Замена для каждого автора тестовой выборки на одну из обучающего множества.
10. Повтор с шага 8 до тех пор, пока каждая из четырех выборок автора не будет использована в качестве тестовой.
11. Увеличение объема выборки на заданный шаг, если предел не достигнут. Повтор с шага 5.
12. Повтор с шага 4 для следующей пары классов.

Для каждого объема выборки было проведено по 280 экспериментов (учитывались все сочетания авторов и текстов). В качестве результирующей оценки точности по данному признаку и объему выборки подсчитывалась средняя частота правильных классификаций.

4. Результаты экспериментов, обсуждение, выводы

По сформулированной выше методике были проведены эксперименты, результаты которых представлены на рис. 1–3.

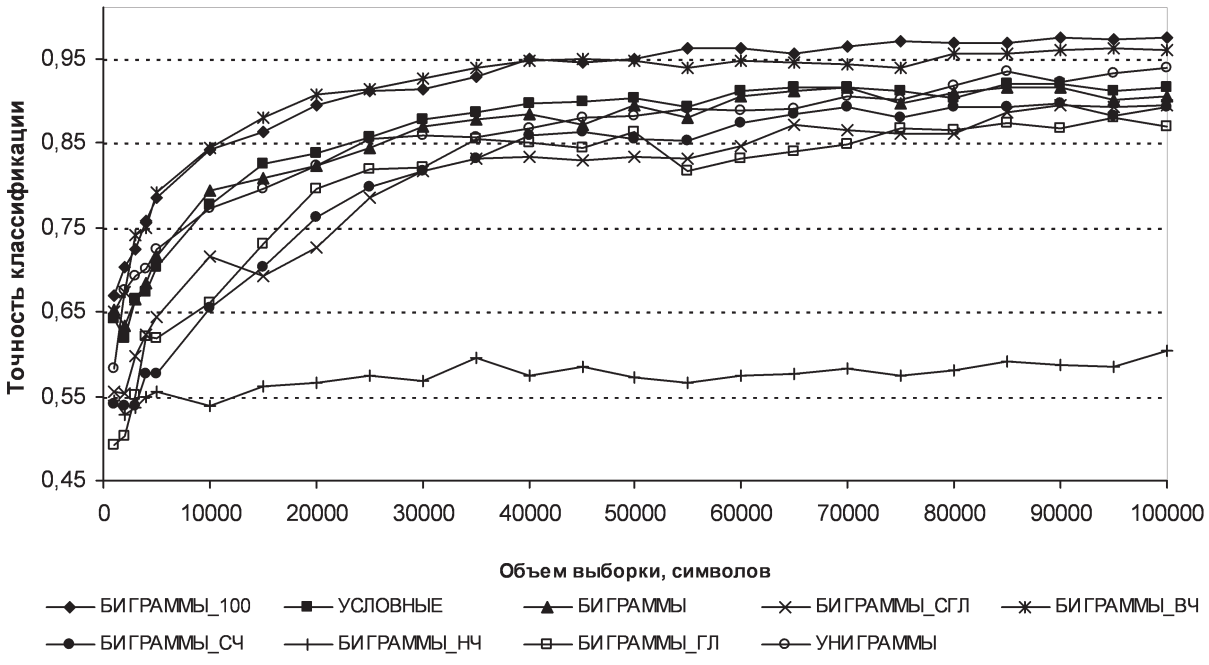


Рис. 1. Результаты исследований по унигрампмам и биграммам символов

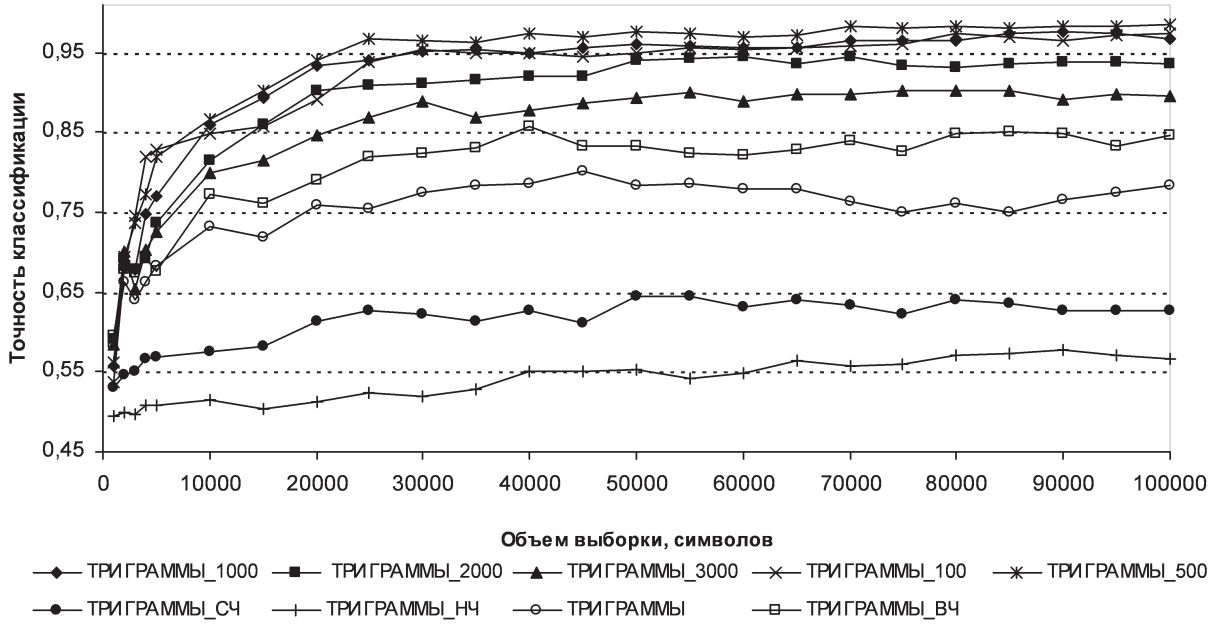


Рис. 2. Результаты исследований по триграммам символов

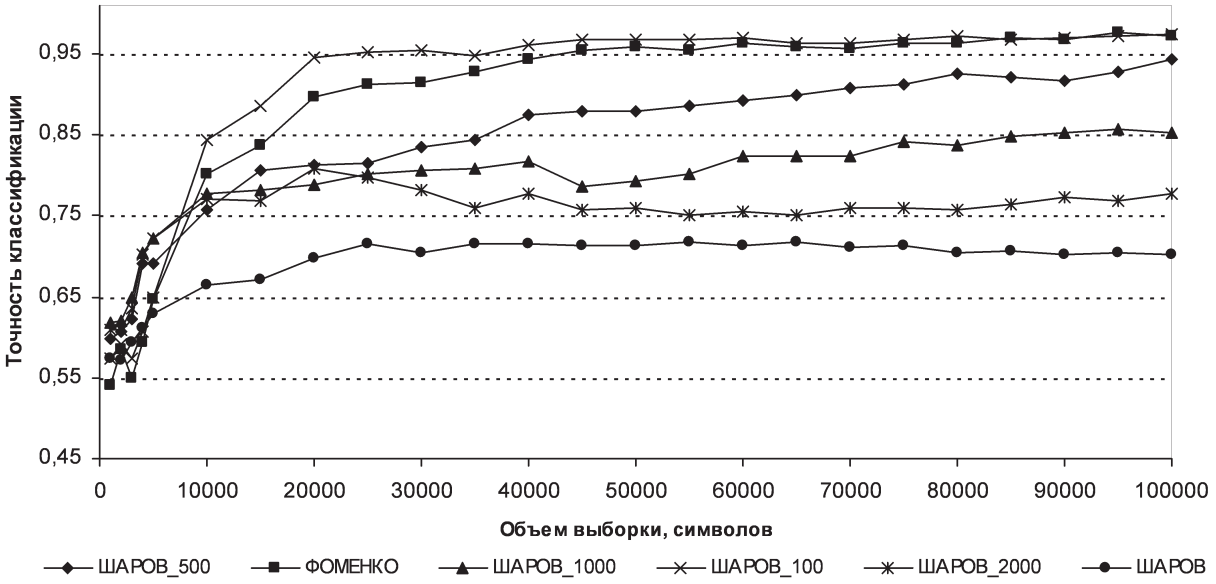


Рис. 3. Результаты исследований по частотному словарю русского языка

Из первой группы признаков (см. рис.1) наиболее точно классификация проходит при использовании признаков БИГРАММЫ_100 и БИГРАММЫ_ВЧ. Значение точности стабилизируется при объеме выборки равном 40 000 символов и далее колеблется около 0,96. Дальнейшее увеличение количества признаков и использование биграмм со средними и низкими частотами встречаемости ведет к снижению точности, к аналогичным результатам приводит использование биграмм, составленных отдельно из гласных и согласных букв (средняя точность 0,77). Приблизительно одинаковые результаты дает использование признаков БИГРАММЫ и УНИГРАММЫ (средняя точность 0,83), немного выше точность классификации по признаку УСЛОВНЫЕ — 0,84.

Среди признаков, основанных на триграммах символов (см. рис.2), наиболее точной оказывается классификация по признаку ТРИГРАММЫ_500, стабилизация наступает при объеме выборки равной 25 000 и далее точность колеблется около 0,97. При увеличении размерности признаков, аналогично экспериментам с биграммами символов, наблюдается снижение качества классификации.

Использование 100 наиболее частых слов русского языка предпочтительнее предложенного Фоменко набора служебных слов за счет более высо-

кого качества классификации на выборках любого объема, средняя точность классификации соответственно 0,87 и 0,86 для признаков ШАРОВ_100 и ФОМЕНКО (см. рис. 3). Стабилизация при использовании признака ШАРОВ_100 наступает при объеме выборки равном 20 000 символов и далее точность колеблется около 0,96. Повышение количества признаков до 500 и более снижает качество классификации.

В целом по результатам экспериментов можно сделать вывод, что идентификация автора с помощью аппарата SVM в случае двух альтернатив возможна при объеме выборки 20 000 символов и больше. При этом нецелесообразно использовать более 500 признаков.

В известных авторам работах по автоматическому определению авторства текста на русском языке приводятся результаты исследований при количестве классов, равном 10 и более, задача идентификации автора в случае двух возможных альтернатив не рассматривается. Сравнение же с аналогичными исследованиями для других языков проводить некорректно в силу особенностей строения каждого языка.

Высокая точность бинарной классификации позволит в дальнейшем применять наиболее эффективные наборы признаков для идентификации автора в случае трех и более предполагаемых авторов.

Однако необходимый для точной идентификации объем текста пока слишком велик для решения большинства практических задач. В дальнейших работах авторами планируется исследовать техники сглаживания [14] для уменьшения требуемого объема выборок, а также продолжить тему поиска статистически устойчивых характеристик на малых текстовых фрагментах и провести эксперименты с наиболее эффективными характеристиками на более представительном корпусе текстов.

Работа поддержана грантом ФСРМПНТ.

Литература

1. *Sebastiani F.* Machine learning in automated text categorization // ACM Computing Surveys, 2002. Vol. 34, № 1, P. 1–47.
2. *Шевелев О. Г.* Методы автоматической классификации текстов на естественном языке: Учебное пособие. Томск:ТМЛ-Пресс, 2007. 144 с.
3. *Hsu C.-W., Lin C.-J.* A comparison of methods for multi-class support vector machines // IEEE Transactions on Neural Networks, 2003. № 13(2). P. 415–425.
4. *Vapnik V. N.* Statistical Learning Theory // Wiley, New York, 1998. 732 pages.
5. *Vapnik V. N.* The nature of statistical learning theory // Springer-Verlag, New York, 2000. 332 pages.
6. *Hsu C.-W., Chang C.-C., Lin C.-J.* A practical guide to support vector classification. — Режим доступа: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, свободный.
7. *Васильев В. Г.* Комплексная технология автоматической классификации текстов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). — М.: РГГУ, 2008. С. 83–91.
8. *Романов А. С.* Подходы к идентификации авторства текста на основе n-грамм и нейронных сетей // Молодежь и современные информационные технологии. Сборник трудов VI Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых. Томск, 26–28 февраля 2008 г. Томск: Изд-во ТПУ, 2008. С. 145–146.
9. *Романов А. С.* Структура программного комплекса для исследования подходов к идентификации авторства текстов // Доклады Томского государственного университета систем управления и радиоэлектроники. Томск: Изд-во ТУСУР, 2008. Ч. 1. №2(18). С. 106–109.
10. *Библиотека* Максима Мошкова. — Режим доступа: http://www.lib.ru, свободный.
11. *Шаров С. А.* Частотный словарь русского языка. — Режим доступа: http://www.artint.ru/projects/frqlist.asp, свободный.
12. *Фоменко В. П., Фоменко Т. Г.* Авторский инвариант русских литературных текстов // Фоменко А. Т. Новая хронология Греции: Античность в средневековье. М.: Изд-во МГУ, 1996. Т. 2. С. 768–820.
13. *Porter M. F.* Russian stemming algorithm. — Режим доступа: http://snowball.tartarus.org/algorithms/russian/stemmer.html, свободный.
14. *Katz S. M.* Estimation of probabilities from sparse data for the language model component of a speech recognizer // IEEE Transactions on Acoustics, Speech and Signal Processing, 1987. № 35(3). P. 400–401.

Стратегии членения спонтанной речи на синтаксические единицы[1]

Strategies of delimitation of syntactic units in spontaneous speech

Рыко А. И. (aryko@mail.ru), **Степанова С. Б.** (stsvet_2002@mail.ru)

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

На материале Звукового корпуса русского языка проведен перцептивный эксперимент по членению спонтанных монологов на предложения. Анализируются стратегии экспертов, принимавших участие в эксперименте.

1. Введение

Нет ничего проще, чем выделить предложение в письменном тексте. Знаки препинания — точка, вопросительный, восклицательный знаки, многоточие — почти однозначно отмечают конец предложения. И дальше можно определять его тип: сложное или простое, сложносочиненное или сложноподчиненное, с союзной связью или бессоюзное, односоставное или двусоставное и т. п., выявлять связи между членами предложения, анализировать интонацию чтения того или иного типа предложений и т. д. Казалось бы, те же процедуры легко проделает любой лингвист и со спонтанной звучащей речью. Однако, как показывает практика и как пишут исследователи, занимающиеся живой, «непричесанной», записанной в полевых условиях речью, выделить в ней предложение (фразу, высказывание — в данном случае для нас это синонимы) — задача весьма сложная, а может быть, даже невыполнимая.

Именно поэтому многие исследователи отказываются от этого понятия при работе со спонтанной речью.

Так, при аннотировании корпуса устных рассказов детей о сновидениях создатели корпуса сегментируют речь на элементарные дискурсивные единицы (ЭДЕ), определяя их, вслед за У. Чейфом, как некие «минимальные кванты устного дискурса» [Подлесская, Кибрик 2004: 1], «вербализующие "фокус сознания" говорящего, т. е. совокупность информации, которую селективное человеческое сознание может одновременно удерживать в активном состоянии» [Кибрик и др. 2008: 1]. И хотя авторы пишут, что «типичная ЭДЕ совпадает по объему с предикацией (элементарным предложением, clause)» [там же], примеры показывают, что ЭДЕ московских ученых ближе к тому, что мы назвали бы «синтагмой».

В другом исследовании С.В. Андреева предлагает в качестве наименьшей структурной единицы синтаксиса устной речи принять конструктивно-синтаксические единицы (КСЕ), которые, с ее точки зрения, позволяют «охватить как предложенческие, так и непредложенческие, как предикативные, так и непредикативные речевые факты» [Андреева 2005: 9]. С. В. Андреева не разрабатывала специально принципы выделения КСЕ, лишь указала, что «при вычленении вспомогательных единиц речевой коммуникации учитывались следующие моменты: 1) коммуникативное, интонационное их выделение и образование синтагмы, 2) возможность при определенных условиях быть самостоятельной единицей» [там же: 45], что опять же говорит о близости этой единицы к синтагме.

Некоторые исследователи, определяя границы между предложениями в спонтанном тексте, прибегают к методу «пунктирования»: большая группа носителей русского языка ставит точки в орфографической расшифровке устных монологов, в которой убраны как знаки препинания, так и знаки пауз [Богданова, Бродт 2007]. Этот метод вполне применим тогда, когда исследователи интересуются прежде всего синтаксическими связями слов внутри предложений.

Мы же предполагали, что наличие звука поможет экспертам точнее и единодушнее определить границы такой привычной для всех единицы лингвистического анализа, как предложение. В связи с этим нас

[1] В настоящее время работа проводится при поддержке РГНФ — проект 07-04-12163в «Разработка информационной среды для мониторинга устной русской речи».

заинтересовал сам разброс мнений экспертов, принимавших участие в нашем эксперименте, мы попытались установить причины, вызывающие различия в стратегиях, и найти те факторы, которые определяют однозначное выделение границ предложения.

2. Материал и методика

Материалом в нашем эксперименте послужили спонтанные монологи и реплики из записей Звукового корпуса русского языка. Расшифрованные и представленные в орфографической форме без каких-либо знаков пауз или препинания, тексты были разосланы шести коллегам-лингвистам, имеющим достаточно большой опыт работы со спонтанной речью. Они могли неоднократно прослушивать записи. Задание было сформулировано так: поставить знаки конца предложений.

Выбор именно лингвистов на первом этапе проведения эксперимента обусловлен тем, что мы предполагали наличие у них сходного понимания термина «предложение» (в том числе и предложение в устной речи) и ожидали более-менее однотипных способов членения потока речи. Однако, как видно из таблицы 1, при большом количестве совпадений достаточно много и границ, поставленных одним-двумя испытуемыми.

Таблица 1

	Т1	Т2	Т3	Т4	Т5	Т6	Т7	Все тексты
Количество границ	4	14	14	6	11	16	91	156
6 совпадающих границ		1	1	2	4	4	23	31
5 совп. гр.	1	4	1	2		2	14	24
4 совп. гр.			4			4	12	20
3 совп. гр.		1	2		1	2	9	15
2 совп. гр.	2	4	3	2	3	3	9	26
1 совп. гр.	1	4	3		3	1	24	36

3. Обсуждение результатов

3.1. Стратегии испытуемых

При анализе полученных результатов выяснилось, что испытуемые, которых мы привлекали к членению спонтанного текста на предложения, реализуют различные стратегии, которые можно условно назвать «максималистской» и «минималистской».

При реализации «максималистской» стратегии в качестве предложений выделяются длинные, многосинтагменные структуры. В этом варианте членения текста в большом количестве представлены бессоюзные предложения, сложносочиненные и сложноподчиненные предложения. При следовании «максималистской» стратегии испытуемый ориентируется на письменный текст, стремится увидеть аналог «литературного» предложения, отталкивается в первую очередь от синтаксической структуры, зачастую игнорируя при этом интонацию. «Минималистская» стратегия позволяет провести границу везде, где только это можно сделать: два и более простых предложений вместо одного бессоюзного сложного, ряд простых предложений с эффектом парцелляции вместо сложноподчиненных и сложносочиненных. Очевидно, что «минималистская» стратегия предполагает ориентацию на интонацию текста, допуская расширение набора возможных синтаксических структур. Фактически любая затянувшаяся межсинтагменная пауза в рамках такой стратегии — повод для проведения границы между предложениями. Любопытно, что один и тот же испытуемый придерживается то одной, то другой стратегии — ни один из них не был последователен на протяжении всего эксперимента.

«Максималистская» стратегия:

(1) я вот так смотрю на такую вот машину сразу думаю не директор ли мой едет у него по-моему «Вольво» //

«Минималистская» стратегия:

(2) я вот так смотрю на такую вот машину // сразу думаю не директор ли мой едет // у него по-моему «Вольво» //

Проведение некоторых границ определенно вызывает различную интерпретацию разными испытуемыми. Так, «пограничные маркеры» (*вот, как бы, ага, ну и вот*) то выделяются только с одной стороны, присоединяясь к предложению справа или слева, то с двух сторон, т. е. как отдельные предложения.

(3) ***приходила когда я уже позавтракаю поиграю [3] вот [4] оставляла мне еду в термосе:***
приходила когда я уже позавтракаю поиграю вот // оставляла мне еду в термосе горячую
vs
приходила когда я уже позавтракаю поиграю // вот // оставляла мне еду в термосе горячую

На наш взгляд, не имеет особого значения, с одной или с двух сторон выделяется такой маркер — его пограничный характер очевиден в любом случае, и, по большому счёту, испытуемые, считаю-

щие такой маркер отдельной единицей или служебным средством, включенным в другую единицу, друг другу не противоречат.

По-разному могут трактоваться и случаи обрывов или самокоррекции: наличие синтаксического сбоя (перестройки синтаксической структуры) приводит к тому, что испытуемые воспринимают его как границу между предложениями:

(4) *у тебя нету этого ничего [6]* ***те-е [2] тебе*** *не надо ничего продавать понимаешь*

(5) ***с ним кане… [4] ну блин*** *[1] ты помнишь как я парился-то с этим*

Очевидно, что не все участки спонтанного текста членятся на предложения с одинаковой степенью легкости. Чем больше количество обрывов, словпаразитов, пауз, чем труднее уловить смысл сказанного, тем больше усилий затрачивается на членение такого фрагмента на предложения и тем больше возникает разногласий у испытуемых.

Ср. «легкий» фрагмент:

(6) *ну у нас тоже на лекциях я вот помню что все сидели вообще никто ничего не понимал* ***[6]*** *самые примерные просто успевали списывать там (нрзб)* ***[6]*** *ну я на практике еще что-то как-то понимала [2] а на теории так это вообще [1] никак* ***[6]*** *Котов у нас был этот прикольный*

и «трудный» фрагмент:

(7) так хорошо [4] просто Вик(?) а что́ это знаешь [3] а она что-то [2] такая [2] всем понятно [1] такая

По-видимому, это можно считать аргументом в пользу того, что предложение в восприятии испытуемых по большей части является единицей семантического плана: если смысл от слушающего ускользает, то не принципиально, где именно провести границу и проводить ли ее вообще.

По-разному испытуемые членят на предложения монологи, произнесенные в разных условиях. В качестве особого случая можно рассматривать речь на ходу (во время прогулки по городу) при прерывистом дыхании с длительными паузами. Здесь при последовательной реализации «минималистской» стратегии как предложение интерпретируется практически любой отрывок, произнесенный на одном выдохе и имеющий предикацию (то есть предложение фактически совпадает с синтагмой). При последовательной реализации «максималистской» (синтаксически ориентированной) стратегии паузы, как правило, игнорируются и членение «монолога на ходу» по количеству выделенных предложений не отличается от монолога, произнесенного в спокойных условиях.

3.2. Характеристика синтаксических условий границ между «предложениями»

В результате нашего эксперимента оказалось, что некоторые границы между «предложениями» независимо от приверженности той или иной стратегии испытуемые установили единодушно. Проанализируем те случаи, которые большая часть испытуемых интерпретировала как границы между предложениями (4–6 совпадений).

А. Смена плана повествования

1) Переход от общего рассуждения к рассказу о конкретных событиях

(8) *то есть учила естественно по советскому букварю другого не было* ***[5]*** *и там где Красная площадь на первой странице храм Василия Блаженного она тогда меня спрашивала*

Характеристика ситуации (в широком ее понимании — вплоть до характеристики эпохи) «другого не было» > рассказ о конкретном событии (бабушка задает вопрос по конкретной картинке).

(9) *нормальное отношение достаточно между родителями так что [2] что такого* ***[5]*** *мне еще сегодня предстоит (нрзб)*

Общее рассуждение (не очень понятно даже о чем оно) > сообщение о своем намерении.

(10) *самые примерные просто успевали списывать там (нрзб)* ***[5]*** *ну я на практике еще что-то как-то понимала*

Сообщение обо всех > сообщение о себе.

(11) *и там большая такая кирха [1] но это не кафедральный… собор был но какой-то большой* ***[8[1]]*** *я зашел просто [3] там идет не служба крестины были*

Общее описание сменяется рассказом о конкретном событии.

2) «Авторское резюме»: переход от частного к общему

[1] Часть материала прослушали 8 экспертов, в этом разделе мы рассматривали границы, единодушно определенные 7–8 из них.

Как правило, включает и какую-либо оценку событий говорящим; имеет показатели —слова с обобщающим значением (*то есть, ну в общем, ну, ну там, короче, ну знаешь, получилось, ну так*).

(12) *хъям бозий [2] храм божий [смех]* ***[4]*** *то есть это вот всё с ее слов реально вот помню до сих пор*

Рассказ о конкретном событии > оценка его (характеристика сохранности эпизода в памяти).

(13) *и причем как бы [1] Люба в термосе картошечка там* ***[4]*** *ну в общем чтобы подлиннее еще было*

(14) *у нас сегодня эта твоя ссылка на мужиков в красных труселях просто весь департамент валялся [2] я всем разослала* ***[5]*** *ну это надо [1] это [2] вы не поймете [2] это надо видеть*

(15) *ну вот [2] ты знаешь сколько они в Омске стоят [1] по Серёгиным данным [3] самый простой шкаф вообще без всяких-всяких наворотов просто* ***[4]*** *ну там же ну можно до бесконечности наворачивать всё зависит от фурнитуры от всего*

(16) *ну там же ну можно до бесконечности наворачивать всё зависит от фурнитуры от всего* ***[5]*** *короче стоит пять тысяч погонный метр [3] рублей*

(17) *ну и вот [3] пять тысяч [1] погонный метр [1] то есть в среднем шкаф выходит там тысяч тридцать* ***[5]*** *ну знаешь они там обычно пятиметровые*

(18) *и они всех провожают прихожан [5] у двери стоят и провожают* ***[7]*** *получилось что я первый выхожу* ***[8]*** *он меня спросил как мне понравилась служба [4] то есть всех в лицо знают [3] они своих прихожан всех знают и спрашивают…*

«Однозначными» границами с двух сторон выделен фрагмент, который несколько выбивается из контекста повествования (перечисления последовательных действий участников общей ситуации «окончание службы»): фрагмент *получилось что я первый выхожу* включает взгляд извне ситуации, некоторое авторское резюме.

(19) *и рассказал что туда каждое первое воскресенье… каждого месяца приезжает какая-то большая община из Петербурга [3] м-м [5]* ***(8[2])*** *ну так хорошее впечатление* ***[5]*** *спросил как впечатление [3] так ненавязчиво но тем не менее…* ***[4]*** *спросил поговорили [1] и тут же следующие пошли*

Снова в изложение последовательных событий включается «авторское резюме» (левая его граница усилена хезитационной паузой, правая граница размыта).

3) «Смена сюжета»

Характеристика макроситуации с разных сторон; переход к повествованию о новом этапе развития сюжета. Видимо, для всех примеров этой группы характерна смена субъекта повествования. Во многих случаях подобный переход отмечается пограничными маркерами (*ага, ну и вот, так вот, причем* (= «с другой стороны»).

(20) *ну я на практике еще что-то как-то понимала [2] а на теории так это вообще [1] никак* ***[5]*** *Котов у нас был этот прикольный*

О своем восприятии учебной дисциплины > о манере преподавателя.

(21) *такая что-то сама с собой ла-ла-ла-ла-ла* ***[4]*** *ага* ***[3] (8)*** *вся аудитория просто тупо сидит так (2) о-о-о*

Описание ситуации с разных сторон, характеристика разных ее участников. Пограничный «финальный» маркер *ага* может выделяться с обеих сторон.

(22) *надо это всё узнать* ***[4]*** *я вот так смотрю на такую вот машину [1] сразу думаю не директор ли мой едет [3] у него по-моему «Вольво»* ***[5]*** *ну и вот* ***[3]*** *пять тысяч [1] погонный метр*

Комментарий к тому, что попало в поле зрения > возвращение к основной теме монолога. Пограничный маркер *ну и вот* часто выделяется с двух сторон.

(23) *то есть ты с людей берешь тридцатник [1] а тратишь на это допустим тысяч десять* ***[4]*** *ну надо только потом кого-нибудь мастеров [1] поначалу там можно самим вот допустим Вован Аримов будет собирать*

Переход к новому этапу развития сюжета.

(24) *первый шкаф конечно будет сложный* ***[4]*** *вот Серёга говорит что он дал рекламу [1] в Омске [1] там сейчас кому-то он что-то будет делать*

Планы на будущее > рассказ об аналогичном опыте приятеля.

---

[2] Очевидно, что все аудиторы отметили здесь границу между «предложениями». Мы сочли безразличным, где именно они ее отметили — до паузы хезитации или после нее, и посчитали это единой границей.

(25) *там фурнитуры ну эти эти всякие [1] ролики [1] где там конструктор всё собирается* ***[5]*** *ну а люди которые в этом не понимают они готовы за это платить*

Характеристика ситуации с разных сторон — с технической и со стороны восприятия потенциальных клиентов.

(26) *а потом вот ты понимаешь нанимаешь уже плотника [2] который* ***[4]*** *просто я в детстве работал на сборках*

Б. «Формальные» границы

4) Граница между ответом на предшествующий вопрос собеседника и собственно повествованием

(27) *а-да-да* ***[4]*** *а потом вот ты понимаешь нанимаешь уже плотника*

(28) *да понравилось* ***[7]*** *интересно [1] конечно [2] и разница большая между Хельсинки и Стокгольмом*

5) Как правило, с двух сторон выделяются вопросы

Это происходит благодаря специфическому интонационному оформлению и характерному синтаксическому построению высказывания. Часто граница справа оказывается «механической» (т. е. дальше идет реплика собеседника), поэтому нами не рассматривается. Границы слева и справа появляются в случае наличия риторического вопроса или при передаче вопросов диалога.

(29) *ну я там говорил Кремль допустим [1] "р" не выговаривал [1] Кйемль* ***[4]*** *а это что [1] за собор?****[4]*** *хьям бозий [2] храм божий [смех]*

(30) *мне еще сегодня предстоит (нрзб)* ***[5]*** *может хочешь со мной пойти?* ***[4]*** *а это тебе не важно*

(31) *я задумался [1] стал более развивать* ***[5]*** *знаешь что такое шкафы-купе?*

(32) *тебе не надо ничего продавать понимаешь?* ***[4]*** *ты рекламу даёшь [1] люди тебе звонят*

(33) *да так просто задумался [1] хотя мне бизнес на хер не нужен* ***[5]*** *Как у тебя с Леной?*

В. Формальные показатели границ

7) «Пограничные маркеры» (вот, как бы, ага, ну и вот)

Видимо, можно различать финальные маркеры («как бы», «вот так»), которые сигнализируют о конце предложения, начальные маркеры («ну», см. примеры в разделе «Авторское резюме»), которые вводят новое предложение, и пограничные маркеры («вот», «ага», «ну и вот», «и всё»), которые, как правило, отделяются значительными паузами с обеих сторон и часто интерпретируются испытуемыми как самостоятельные единицы.

(34) *меня мама учила читать [3] она уходила на работу с утра [1] ну пока я спала и там на несколько часов [2] приходила когда я уже позавтракаю поиграю [2] вот* ***[4]*** *оставляла мне еду в термосе [2] горячую*

Пограничный маркер «вот» обычно «финальный», может выделяться с двух сторон, сопоставим с резюме («свернутое резюме»).

(35) *чтобы я читала да [2] как бы* ***[4]*** *причем писала крупными буквами*

«Как бы» — финальный «маркер неуверенности», предполагающий, видимо, затяжную паузу.

(36) *ну даже дело не в закалке наверно [2] както ну у них это все объясняется как бы* ***[4]*** *нормальное отношение достаточно между родителями так что [2] что такого*

(37) *такая что-то сама с собой ла-ла-ла-ла-ла* ***[4]*** *ага* ***[3]*** *вся аудитория просто тупо сидит так (2) о-о-о*

(38) *я вот так смотрю на такую вот машину [1] сразу думаю не директор ли мой едет [3] у него по-моему «Вольво»* ***[5]*** *ну и вот* ***[3]*** *пять тысяч [1] погонный метр*

(39) *говорят что зимой даже лучше чем летом* ***[5]*** *вот так*

4. Интонационный комментарий

Не будет преувеличением сказать, что интонационное оформление речи — самое вариативное из всех лингвистических явлений. Особенно это касается спонтанной речи. Во многом именно этим объясняется то разнообразие стратегий, которое продемонстрировали наши эксперты при членении речевого потока на предложения.

Возможно, у каждого носителя русского языка есть представление об «интонации точки» — интонации завершения высказывания. Акустически «интонация точки» обычно выражается в падение тона ниже среднего, за которым часто следует пауза. Если эти условия соблюдены и с семантико-синтаксической точки зрения граница представля-

ется вполне уместной, то в этом случае эксперты единодушно ставят ее. Например:

(40) оставляла мне еду в термосе [2] горячую [4] а термос прятала [6] а оставляла записку на столе где она спрятала термос

На слове *прятала* наблюдается падение ЧОТ с 240 Гц до 177 Гц на ударном гласном и падение продолжается до 100 Гц на заударных слогах, затем следует пауза длительностью в 433 мс. Эта «“интонационная” граница» поддерживается возможностью постановки здесь и синтаксической границы. Все эксперты ее отметили.

Практически все «уверенные» границы в эксперименте характеризуются аналогичными изменениями ЧОТ и наличием после них либо паузы, либо смены говорящего.

Сомнения начинаются тогда, когда те же самые просодические особенности появляются в местах, где с синтаксической точки зрения точки быть не должно. В примере (40) падение ЧОТ на слове *в термосе* не меньше, чем в слове *прятала*, пауза после него — 1776 мс, однако слово *горячую* явно связано со словом *еду*, поэтому только двое экспертов увидели в данном контексте эффект парцелляции и выделили *горячую* в отдельное предложение, остальные сочли такую длительную паузу чисто хезитационной.

Не меньше сомнений вызывают и случаи, когда синтаксическая граница вполне возможна, однако вместо падения тона слышен ровный тон или даже его повышение. Именно так часто оформляются следующие друг за другом простые предложения:

(41) *слушай **Славик** [3] я тут придумал вид **бизнеса** [3] мне кажется достаточно **перспективный** [2] и не такой **геморный** [6]*

Кроме последнего слова все остальные выделенные слова характеризуются повышением тона, что вызывает разночтение у наших испытуемых.

5. Основные итоги

Наш эксперимент показал, что профессиональные лингвисты, сталкиваясь при анализе звучащей спонтанной речи с неполной согласованностью семантических, синтаксических и интонационных характеристик отрезков звучания, прибегают к разным стратегиям вычленения синтаксических единиц. Некоторые предпочитают интерпретировать отдельные синтагмы как самостоятельные простые предложения, некоторые объединяют их в сложные.

Анализ «уверенных» границ помог выявить те формальные признаки, которые помогут в дальнейшем расшифровщикам спонтанной речи единообразно выделять в устных текстах предложения.

Литература

1. *Андреева С. В.* Элементарные конструктивно-синтаксические единицы устной речи и их коммуникативный потенциал: Автореф. дис. … докт. филол. наук // Саратов, 2005.
2. *Кибрик А. А., Подлесская В. И., Дараган Ю. В., Ефимова З. В., Коротаев Н. А., Литвиненко А. О., Цуканова В. Л.* Проблема сегментации устного дискурса и когнитивная система говорящего // www.ksu.ru/ss/cogsci04/science/cogsci04/107.doc
3. *Богданова Н. В., Бродт И. С.* Спонтанный монолог: синтаксические характеристики текста и уровень речевой культуры говорящего // Вестник Санкт-Петербургского государственного университета. Серия 9 (Филология). Вып. 1 (Часть I). Март 2007 г. С. 35–43.
4. *Подлесская В. И., Кибрик А. А.* Транскрипция устного дискурса для нужд корпусных исследований // http://www.dialog-21.ru/Archive/2004/Podlesskaja.pdf

Информация энциклопедического характера в прикладном семантическом словаре

On encyclopaedic data in an applied semantic dictionary

Семенова С. Ю. (Sonya_sem@mail.ru)
ИНИОН РАН, Москва

Включение в лингвистический словарь сведений об обозначаемых реалиях, практикующееся в современной лексикографии, отвечает духу когнитивного направления, для которого характерен интерес к целостному восприятию информации. В статье вопрос об энциклопедической информации обсуждается применительно к АОТ-ориентированному семантическому словарю, в котором лексика представлена структурированно, и круг сведений, как лингвистических, так и расширенных, жестко связан с форматами словаря. Рассматривается система энциклопедических функций словаря РУСЛАН; обсуждаются вопросы их заполнения и расширения их круга. Также рассматривается ряд лексических классов и единиц, для которых в словаре целесообразно указание конкретных видов энциклопедической информации.

1. Об энциклопедической информации в традиционных словарях

Разграничение лингвистической информации о слове и внелингвистической информации об обозначаемой реалии составляет одну из традиционных проблем в лексикографии. Оно связано с более общей и «вечной» проблемой вычленения собственно предмета лингвистики и такими ее гранями, как соссюровское противопоставление внутренней и внешней лингвистики, или, например, более локальное, возникшее в прикладных задачах разделение участников ситуации на семантические и прагматические. В лексикографической деятельности наряду с разграничением стоит и вопрос об объединении данных обоего рода. Например, в одной из последних своих работ В. Г. Гак отмечал тенденцию в общей лексикографии, обратную тенденции к рассеянию информации по разным словарям — включение разнообразных сведений о реалиях в толковые или переводные словари классического типа, стирание грани между языковым и энциклопедическим словарем, шаги к созданию своего рода «энциклопедии языка» [3]. Как отмечено в более давней работе этого ученого, соединение обоих видов информации может осуществляться двумя способами: 1) разработка двух словарей (примерно для одного и того же словника) — языкового и энциклопедического; 2) соединение сведений обоего рода в одной словарной статье. Последнее в традиционном словаре может быть осуществлено либо путем структурированного выделения энциклопедической информации, в специальной зоне, в виде стандартизованных пояснений, определений, либо более свободно, в том числе — в виде примеров, цитат, дающих представление о конкретных воплощениях описываемой сущности [2].

В отечественной лексикографии включение энциклопедических данных практикуется не только в словарях, предназначающихся широкой читательской аудитории и выполняющих просветительские функции, но и в словарях, адресованных в первую очередь лингвистам и базирующихся на четком очерчивании сферы данной науки. Например, такая информация содержится в ряде статей, подготовленных для Интегрального словаря русского языка [13]; в рамках его концепции этнолингвистические (культурологические) аспекты рассматриваются как лингвистически существенные для интегрального описания лексемы [там же, с. 4]. Энциклопедические сведения приведены для таких лексем, как *письмо* (в значении «текст») — данные о способах почтовой коммуникации; *журнал* (как «периодическое издание») — данные о видах периодических изданий, типовом оформлении. Для лексемы *месяц* (как «часть года») указываются сведения астрономического характера; для лексем *месяц* (как «видимый образ луны») и *небо* (в значении «высшие силы») приводятся культурологические материалы;

для лексемы *облако* («скопление в небе водяных капель») даны определенные естественнонаучные сведения; для имени *пробка* («крышка») указывается типовой материал изготовления. В некоторых статьях информация, расширенная по сравнению с минимальным толкованием, введена в поле значения. Например, для слова *забор* названы его части, способ строительства, обычные размеры и материалы.

2. Энциклопедическая информация в автоматических словарях

Соотношение двух видов данных актуально и в компьютерной лексикографии, в том числе в разработках, использующих формализмы для представления сведений о слове. Энциклопедическая информация может «официально» включаться в лексическую базу данных, образуя соответствующую зону, может принципиально оставаться «за кадром», может принимать в себя части толкования, которые не отражаются в лингвистической зоне в силу ограниченности ее форматов, а может, наоборот, отразиться в собственно лингвистической зоне.

В структурированной базе данных предметных имен системы «Лексикограф» [5], не нацеленной на явное представление энциклопедической информации и оперирующей, в основном, с членами толкования, можно, тем не менее, увидеть некоторые следы взаимодействия двух видов данных. Так, судя по примерам в [5], выбор типового предиката, указываемого при описании артефакта, может опираться на некое более расширенное, «энциклопедичное» видение смысла, чем то, которого требует, скажем, «узкое» традиционное толкование предметного имени через родовое; сама задача указания такого предиката предполагает обращение к онтологическому знанию, пусть и обыденному. Так, идеи укрытия, защиты для имени *палатка* (например, туристическая) наследуются от свойств родового имени *помещение* и опускаются при «именном» толковании с интегрирующим и дифференцирующим компонентами («*палатка1* — временное помещении из натянутой на остов ткани, шкуры»; см. словарь

С.И. Ожегова), а задача указания предиката требует эксплицировать их. С другой стороны, некоторые части толкования, не отражаемые в формализме базы (например, «топологический» аспект — наличие двух или трех колес у *велосипеда*; в МАСе и словаре С.И. Ожегова этот факт включен в толкование), либо должны быть вообще опущены, либо уйти в зону комментария, что фактически означает отнесение этой информации к энциклопедии.

Тенденция к сращиванию лингвистической и внелингвистической информации находится в русле когнитивного направления, для которого важны механизмы понимания и их моделирование, и менее существенно разделение понимаемого на языковые и бытийные составляющие.

Идея сочетания в словарных продуктах лингвистических сведений о предметных именах и энциклопедических сведений о пространственных свойствах обозначаемых предметов, высказывалась, например, в статье [4], относящейся к серии работ, посвященных грамматике вербализации зрительных впечатлений.

Концепция сочетания в прикладной системе сведений двух видов — терминологического словаря и декларативных знаний о предметной области, слияния тезауруса (именно как лингвистического продукта) и онтологии, обсуждается в [8].

Определенный тип объединения лексических и понятийных связей в рамках одного формализма реализуется в тезаурусных системах класса WordNet, в частности, в компьютерном тезаурусе русского языка, в существенной мере опирающегося на идеологию WordNet [1 и др.]. В данном случае традиционные тезаурусные связи («выше-ниже», «часть-целое») несут больше онтологическую нагрузку, а лексические свойства описываются дополнительными по сравнению с классическим тезаурусом связями — валентностными, деривационными и некоторыми другими.

Основная направленность прикладных компьютерных словарей — автоматическая обработка текста, в конечном счете, понимание текста и использование содержащейся в нем информации. Очевидно, что «интеллектуальность» обработки, возможность интерпретирования текста, раскрытия аллюзий, логического вывода должна увеличиваться при расширении словарной информации, при опоре процессоров на энциклопедические данные наряду с собственно словарными.

3. Энциклопедическая информация в словаре РУСЛАН

3.1. Состав энциклопедических функций

Оба вида сведений, языковых и экстралингвистических, представленных в одном словарном продукте, предполагается использовать в рамках Информационно-лингвистической модели обработки текста, предложенной Н. Н. Леонтьевой. Основным семантическим словарным ресурсом системы, разрабатываемой в соответствии с этой моделью, является словарь РУСЛАН [6, 7 и др.; 12 и др.]. Работы в рамках данной модели неоднократно поддерживались грантами РГНФ и РФФИ; в настоящий момент в РГНФ подан проект 09–04–00302а «Взаимодействие словаря и текста в процессе построения ситуативных структур».

В словаре РУСЛАН, параллельно с зонами семантических характеристик, валентностей, тезаурусных связей, лексической сочетаемости, формализованного представления ситуаций и некоторых других, предусмотрена зона энциклопедической информации. Как и лингвистические данные, энциклопедическая информация отражается в формализованном виде, путем использования системы жестких шаблонов.

Это осуществляется с помощью ряда энциклопедических функций, аналогичных лексическим. В настоящее время в метаязыке словаря, разработанном Н.Н. Леонтьевой, предусмотрены функции: Anti (антоним — функция, которая для одних лексем может интерпретироваться как энциклопедическая, для других как обычная лексическая); Cap (главный, выделенный элемент обозначаемой сущности); Param (типовой параметр, по которому характеризуется сущность); Mesur (типовая единица измерения — например, если сущность представляет собой измеряемую величину); Sing (единичный элемент, если сущность концептуализируется как множество или масса). Имеются также функции БОЛЬШОЙ_РАЗМЕР; ИСХОДНАЯ_ТОЧКА; КОНЕЧНАЯ_ТОЧКА; ПРИНАДЛЕЖНОСТЬ; СОСТАВНАЯ_ЧАСТЬ. Примеры использования функций при описании слов или лексем: Anti (*гуманитарный*) = *естественный* (о науках); Anti (*холодный*) = *горячий*; Cap (*страна*) = *столица*; Param (*потолок*) = *высота*; Mesur (*температура*) = *градус*; Sing (*снег*) = *снежинка*; БОЛЬШОЙ_РАЗМЕР (*камень*) = *глыба*; ИСХОДНАЯ_ТОЧКА (*жизнь*) = *рождение*; КОНЕЧНАЯ_ТОЧКА (*жизнь*) = *смерть*; ПРИНАДЛЕЖНОСТЬ (*воин*) = *оружие*; СОСТАВНАЯ_ЧАСТЬ (*диссертация*) = *глава*.

Кроме того, в словаре есть «географические» энциклопедические функции для нарицательных имен соответствующего семантического поля; это функции ВЫШЕ_ГЕО, НИЖЕ_ГЕО, ЖИТЕЛЬ: ВЫШЕ_ГЕО (*море*) = *океан*; НИЖЕ_ГЕО (*море*) = *озеро*; ЖИТЕЛЬ (*город*) = *горожанин* и т.п. (Топонимическая лексика, для которой эти функции также актуальны (ЖИТЕЛЬ (*Тула*) = *туляк*; ЖИТЕЛЬ *(Германия*) = *немец* и т.п.), составляет отдельную, отраслевую базу данных). Функции ВЫШЕ_ГЕО и НИЖЕ_ГЕО позволяют не только обозначать иерархии, но и отражать буквальную пространственную вложенность географических объектов: ВЫШЕ_ГЕО (*Сибирь*) = *Российская Федерация*).

При значительном количестве функций, заложенных в метаязык РУСЛАНа, в настоящее время энциклопедическая информация регулярным образом указывается лишь для небольшой части лексики, и объем энциклопедических сведений предполагается расширить. Это требует осмысления сфер применения имеющихся функций, разграничения между ними, введения новых функций, доработки форматов, а также выявления лексических классов, для которых целесообразно указание онтологической информации. Далее в статье мы коснемся этих вопросов; рассмотрение будет иметь пока предварительный характер.

3.2. Энциклопедическая информация и часть речи

Энциклопедическая информация в разной мере актуальна для разных частеречных классов. Традиционно ее приводят, в основном, для номинативной лексики, в том числе для предметных имен. Круг словарных статей в [13], для которых указаны энциклопедические данные, подтверждает это. В самом деле, номинативная лексика, с одной стороны, с ее относительно устоявшимся дроблением на бытийные классы, а с другой стороны, с ее соответствием декларативному, статичному знанию, на котором обычно основываются онтологии, более всего подходит для отработки методик энциклопедических описаний.

Кстати, что также соответствует выбору лексем в [13], в энциклопедической информации часто «нуждаются» существительные, семантические полнозначные (а не тяготеющие к местоименности или метаязыковым единицам, типа *вопрос*, *предмет*, *цель* и т.п.), несколько выходящие за рамки самой общей лексики и отчасти, «наполовину», принадлежащие сферам техники, культуры, экономики, других отраслей. (Специальных, внелингвистических сведений также может требовать отраслевая номенклатура, но узкоспециальные термины предметных областей мы здесь не рассматриваем, и в словарь РУСЛАН такая лексика попадает в очень незначительном количестве).

Для неноминативной лексики вопрос требует отдельного исследования. Что касается глагольной лексики, то круг глаголов, «требующих» энциклопедическую информацию, по-видимому, невелик. Для определенных пластов производной глагольной лексики, сложившейся по механизмам словообразовательных лексических функций (*финансы -> финансировать*), энциклопедические данные подчас можно приводить при описании мотивирующего слова. Названия базовых, «исконных» действий и состояний (*бежать, говорить, понимать* и др.) и их приставочные дериваты, в основном, не требуют энциклопедических пояснений; базовые глаголы интересны именно лексическим значением. В конкретном словаре РУСЛАН для многих глаголов желательные информативные элементы описания могут быть размещены в перспективных и пока малоосвоенных зонах ситуаций, прагматических импликатур; тем самым зона энциклопедии разгружается. Но энциклопедическая информация все же полезна при описании неких специальных действий, типа названий ремесел, технологий — *сваривать, верстать* и др.; она, например, может отражать аспекты, параметры технологических процессов.

Для прилагательных и наречий в РУСЛАНе энциклопедическая информация, в основном, ограничивается пока указанием антонимии (см. п. 3.4), хотя для ряда лексических единиц, кроме того, могут быть указаны внеязыковые сведения — культурологические, например, описывающие социальные сте-

реотипы (для слов *культурный, интеллигентный, взаймы, ва-банк* и других), либо элементы научных сведений (например, для составных слов научной сферы, типа *электромагнитный, злокачественный, полнозначный* и т.п.).

3.3. Энциклопедическая функция «Параметр»

Одной из важных в принципе и относительно частотных в РУСЛАНе разновидностей энциклопедической информации является информация о типовых параметрах обозначаемых объектов. В основном эта функция применяется при описании номинативой лексики. Изначально функция Param предназначалась для указания только количественных параметров: Param (*холм*) = *высота*; Param (*штанга*) = *вес* и т.п. В дальнейшем целесообразным представляется расширить понимание этой функции, описывать с ее помощью также неколичественные, не градуируемые в числовой форме признаки сущностей, но способные, как и количественные, принимать варьирующиеся значения при варьировании в объекте: Param (*плод*) = *спелость*; Param (*краска*) = *цвет*; Param (*автомобиль*) = *марка* и т.п. Расширение понятия «параметр» в неколичественную сферу, с сохранением идеи варьируемости значений внутри данного класса объектов, является естественной ступенью абстрагирования от исходной, количественной интерпретации данного понятия [11]. Функция Param удобна для «анкетирования» описываемой лексики, помогает лексикографу структурировать онтологические знания об обозначаемом предмете, уяснять, какие свойства для данного объекта существенны (так, *абзац* как элемент письменного текста характеризуется *содержанием, координатами на странице*, *длиной*, *шрифтовым оформлением*).

Указание в зоне энциклопедической информации типовых количественных параметров для данной сущности позволяет, например, при текстовом анализе уточнять конкретный вид параметра в конструкциях с составными количественными прилагательными: *пятилетний ребенок* = > *возраст*; *пятилетняя дискуссия* = > *продолжительность*; *пятилетний план* = > *срок*. Указание типовых неколичественных признаков также способствует распознаванию смысловых связей между текстовыми единицами.

В дальнейшем представляется полезным разделение в словаре энциклопедических свойств на постоянные и переменные для обозначаемой сущности: свойства ее как таковой и свойства, отличающие данный экземпляр в ряду ему подобных. Так, для сущности *август* постоянными свойствами будут продолжительность в 31 день, принадлежность к летнему сезону, наличие погодных предвестников осени (при этом восьмой порядковый номер в календарном году считается элементом толкования; см. МАС или словарь С.И. Ожегова), а в качестве переменного свойства, например, может быть указан признак «оценка по погоде», принимающий разные значения в разные годы («теплый», «дождливый» и т.п.).

Переменная информация естественным образом выражается при помощи функции Param. Если эта функция обозначает признак с небольшим и «перечислимым» кругом значений, то эти значения полезно добавить в ее формат. Так, для признака «род» (грамматический) целесообразно перечислить значения «мужской», «женский», «средний»; для бинарного признака «истинность» — указать варианты «истина» и «ложь».

Совокупность постоянных свойств будет распределена между разными энциклопедическими функциями, в том числе может быть задействована та же функция Param, но с указанием фиксированного значения. Различие между постоянными и переменными свойствами аналогично различиям между прямой и параметрической диатезами некоторых глаголов с валентностью на содержание, в том числе глаголов принятия решений типа *выбирать, рисковать* [10 и др.]. Нотацию для отражения постоянных и переменных значений в РУСЛАНе еще предстоит разработать.

К функции Param тематически близка функция Mesur. Последняя применяется для представления имен количественных параметров *температура, вес* и др., но уместна также при описании сущностей, не параметризовавшихся до статуса величин, но связанных с величинами, так называемых «квазипараметров» [11]. Например, единица измерения *градус* характеризует не только *температуру*, но и погодные явления, такие как *мороз, жара, заморозки*: Mesur (*мороз*) = *градус* (ср. атрибутивные конструкции типа *мороз 20 градусов*), а мерами веса характеризуется не только сам параметр *вес*, но и предметные сущности *груз*, *багаж*, имеющие валентность на значение данного параметра: Mesur (*груз*) = *килограмм; тонна*. Иногда описание требует указания прототипического масштаба единиц измерения: вряд ли *груз* будет характеризоваться *граммами*, а *гиря* — *тоннами*; *размах* скорее всего будет измеряться в *метрах*, а, например, не в *сантиметрах* и не в *световых годах*. Определенное множество имен параметров (*расстояние, время, ток, частота* и др.) нейтрально к масштабу величин — либо в силу своей общезначимости (*время*), либо в силу научной «беспристрастности» к масштабу величины (*ток*), и для этого множества в словаре указываются либо основные «антропоцентрические» единицы — *метр, килограмм, минута, час,* либо основные (т.е. не кратные 10 в n-ой степени) физические единицы — *ампер, ватт, герц* и т.п.

Детализированное описание средств выражения параметрической информации полезно для

сплошного анализа текста, а также для задач типа Information extraction, для извлечения параметрической информации из текста и представления в виде структур, подобных традиционным информационным триадам или объектно-характеристическим таблицам, для совершения математических и логических операций над извлеченными из текста данными [11].

Вообще, параметрическими структурами (понимаемыми широко, а не только количественно) охватываются значительные слои знаний, выражаемых естественным языком; в определенном смысле параметрическая структура, мыслимая либо как двухчастная (признак — - значение), либо как трехчастная /«триада»/ (объект-признак — - значение), либо четырехчастная (объект-признак-условие — - значение), но с неизменным разделением на главные, разнородные по своей природе части — «левую» и «правую» — на признак и его значение, гомоморфна структуре суждения, предложения, также в главном своем членении двухчастной — субъектно-предикатной; темо-рематической.

3.4. Функция «Антоним» — лексикологические и энциклопедические аспекты

Широко практикующимся при ведении словаря РУСЛАН является указание антонимов. В данном словаре не предусмотрено деление антонимов на лексические и энциклопедические; для простоты оба типа лексем, находящихся в противопоставлении к данной, вносятся в энциклопедическую зону. Лексические антонимы понимаются традиционно — это лексемы, толкования которых на уровне примитивов отличаются отрицанием одного из семантических компонентов или расположением значения соответственного компонента симметрично на противоположных полюсах шкал «много-мало», «хорошо-плохо» (*горячий — холодный*; *добрый — злой*). Неточные лексические антонимы, кроме того, отличаются дополнительными оттенками значений, например, незначительным нарушением симметрии на полюсах (*горячий — прохладный*; *добрый — недоброжелательный* и т.п.[9, с. XV]).

Антонимы энциклопедические составляют некий дополнительный, объемлющий пласт лексических пар, для которых нарушаются те или иные элементы приведенного определения. Для них, например, вместо прототипически статичного положения на шкалах может отмечаться противоположная направленность процессов: *анализ — синтез, индукция — дедукция* и др. Противопоставляться могут символические представители подлинных оппозиций: *зеленый — красный* (о сигналах светофора); *белый — красный* (о политических силах в Гражданскую войну); за первой парой стоит бувальное противопоставление разрешения и запрета; за второй — основное политическое противостояние (за частную собственность или против нее). Противопоставляемые единицы могут быть разными по семантической природе; по своей прототипической концептуализации; так, в паре *ядерный — обычный* (о вооружениях) прилагательное *ядерный* означает «относящийся к процессам, происходящих в атомном ядре, к использованию энергии атомного ядра» (МАС), а *обычный1* — «такой, как всегда, постоянный, привычный» (МАС).

Для энциклопедических антонимов важно противопоставление референтов, практикующееся в реальной действительности.

В отдельных случаях противопоставления, как энциклопедические, так и лексические, могут быть исчерпывающими, когда в данном классе сущностей возможны только две данные альтернативы: *рай — ад*; *дескриптивный — нормативный* (о словаре); *цветной — черно-белый* (об изображениях); *аристократия — плебс* (т.е. не-аристократия).

В подавляющем большинстве случаев полной логической взаимодополняемости не отмечается (например, кроме *ядерного* и *обычного* оружия, еще имеется *химическое*, *биологическое* и др.; кроме *начала* и *конца* имеется *середина* и т.д.). В качестве антонимичных энциклопедических пар выбираются те, для которых противопоставления прагматически актуальны; подспудно учитывается тот факт, что действительность, исходно многообразная, способна выстраиваться под некое главное бинарное противостояние. Так, противопоставление *либерал — коммунист,* по-видимому, является главным для современной российской политической сцены, хотя имеются также сторонники других мировоззрений. В сфере наук наиболее представительной, видимо, является оппозиция *гуманитарные — естественные*, связанная с противопоставлением природы и общества, хотя существуют, кроме того, науки *точные*, *технические, прикладные*. Кроме *морского* и *сухопутного* транспорта имеется также *речной, воздушный*, но оппозиция *морской — сухопутный*, содержащая идею отрицания («флот» либо «нефлот»), является при классификации транспорта наиболее принципиальной.

Приведение антонимов в словаре полезно для структуризации возможных контекстов описываемой единицы, для раскрытия ее парадигматики.

Проблему при описании антонимии составляет обусловленная ею потребность мелко дробить слово на значения; аристотелевский принцип, согласно которому разные антонимы указывают на разные значения исходной лексемы, едва ли полностью реализуем в прикладном словаре, в котором значения должны распознаваться процессором, и потому их не должно быть много. В РУСЛАНе разделение слова на значения, конечно, практикуется, но оно является укрупненным.

Заметим, что при описании антонимии выход за рамки чистой лексикологии практикуется и в других прикладных системах. Так, в русском тезаурусе типа WordNet антонимия рассматривается на уровне синсетов, т.е. по-видимому возможны ситуации, когда антонимия не будет точной для всех абсолютно лексических пар внутри противопоставляемых синсетов [1].

3.5. О других энциклопедических функциях

Функции Cap, СОСТАВНАЯ_ЧАСТЬ при описании предметной лексики могут, в частности, использоваться для отражения топологических знаний о предмете: Cap (*кепка*) = *козырек*; СОСТАВНАЯ_ЧАСТЬ (*цирк* /как помещение/) = *арена; купол*. Пример с именем *кепка* показывает, что энциклопедическая зона может вмещать элементы толкования. Подобная информация в РУСЛАНе не отражается в семантической зоне, где описание смысла представляет собой пересечение таксономических классов.

При работе с функциями Cap, СОСТАВНАЯ_ЧАСТЬ, ПРИНАДЛЕЖНОСТЬ встают проблемы их различения: какой элемент обозначаемой сущности считать выделенным (допускается только один выделенный элемент), а какие — «рядовыми»? Не является ли, например, *арена* не просто составной частью, а главным элементом *цирка*? Функция СОСТАВНАЯ_ЧАСТЬ уместна, если объект концептуализируется как составной; например, если это техническое устройство: СОСТАВНАЯ_ЧАСТЬ (*компьютер*) = *монитор*. Функция ПРИНАДЛЕЖНОСТЬ, например, используется, если описываемым объектом (ее аргументом) является некий действующий субъект, «собственник»: ПРИНАДЛЕЖНОСТЬ (*художник*) = *палитра*.

Разграничение сфер употребления этих близких энциклопедических функций входит в более общую задачу разграничения единиц метаязыка. Ряд разграничений среди семантических характристик и смысловых отношений, составляющих основной инструментарий для описания смысла, приведен в [12]; различительная работа с энциклопедическими функциями еще предстоит.

Анализ лексики показывает, что круг энциклопедических функций, структурированно отражающих знания о мире, может быть несколько расширен в РУСЛАНе. Например, может указываться материал, из которого из которого изготавливается предмет; материал может быть постоянным и переменным: Subst (*сабля*) = *сталь* (постоянное свойство); Subst (*посуда*) = *фарфор; металл; пластик* и т.п. Могут указываться химические составляющие: Content (*абрикос)* = *каротин*. Может отражаться сфера применения сущности:

Usage (*барий*) = *рентгенодиагностика;* Usage (*французский* /язык/)= *дипломатия*. Может указываться типовая манифестация: Manif (*недовольство*) = *социальный протест*; Manif (*воспаление*) = *отек*.

Наряду с функциями Cap и СОСТАВНАЯ_ЧАСТЬ может быть введена функция «разновидность»: РЗВДН (*ткань*) = *ситец; шелк; драп* и т.п.; РЗВДН (*оружие*) = *холодный; огнестрельный; ядерный* и т.п. Отчасти эта функция будет дублировать тезаурусные связи, но в то же время, она даст возможность «собирать» терминологические словосочетания, в частности, адъективные, с помощью которых обозначается большое число понятий, иерархически подчиненных данному. (Для РУСЛАНа включение словосочетаний актуально, поскольку в настоящее время тезаурусная зона словаря в основном отражает связи между изолированными словами).

Быть может, целесообразно также введение иерархических связей по предметным областям, подобных связям среди географических объектов: ВЫШЕ_ЛИНГВ (*французский*) = *романский*; ВЫШЕ_БИОЛ (*мандарин*) = *цитрусовый* и т.п. Этот вопрос зависит от дальнейших путей развития системы, от того, с какие отраслевые задачи могут оказаться возложенными на нее. Отраслевые иерархии могут отличаться от обыденных, определяемых подчас не научными классификациями, а прагматикой. Например, для обыденной картины мира тезаурусная связь ВЫШЕ (*собака*) = *домашнее животное* представляется более значимой, чем связь ВЫШЕ_БИОЛ (*собака*) = *млекопитающее.*

Энциклопедическая зона, помимо фиксации определенного круга бытийных знаний, снабжает словарь рядом дополнительных лексических связей, и эти связи обогащают его как поисковую систему, гипертекстовую систему, например, увеличивают количество потенциальных входов. В массе своей лексические связи между лингвистической и энциклопедической частями словаря могут поставлять материал для различных информационных экспериментов, выявления новых соотношений. Так, интересный пример несимметричности лексических единиц есть в работе [5], где сравниваются имена *медведь* и *берлога*: для имени *медведь берлога* входит в энциклопедическую информацию (место зимней спячки), а для имени *берлога медведь* — явный семантический компонент.

Конечно, АОТ-ориентированный словарь, создаваемый небольшим коллективом, в настоящее время может вобрать лишь ограниченные сведения онтологического характера. В перспективе, для более существенного расширения информации о мире, в словаре должен быть задействован развернутый аппарат представления знаний.

И очевидно, что энциклопедическая зона связана с прагматикой; отбирая энциклопедические сведения для словаря, лексикограф, с одной стороны, должен исходить из того, что возможно выразить имеющимися средствами, а с другой стороны, должен предсказывать информационную значимость помещаемого материала для потенциального пользователя или потенциальных приложений.

Литература

1. *Азарова И. В., Митрофанова О. А., Синопальникова А. А.* Компьютерный тезаурус русского языка типа WordNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Межд. конференции Диалог'2003 (Протвино, 11–16 июля 2003 г.). — М.: Наука, 2003. — С. 43–50.
2. *Гак В. Г.* Проблема создания универсального словаря (энциклопедический, культурно-исторический и этнолингвистический аспекты) // Национальная специфика языка и ее отражение в нормативном словаре: Сб. статей. — М.: Наука, 1988. — С. 119–125.
3. *Гак В. Г.* От лингвистического словаря к энциклопедии языка // Проблемы русской лексикографии. Тезисы докладов международной конференции Шестые Шмелевские чтения 24–26 февраля 2004 г. — М., 2004. — С. 17.
4. *Кобозева И. М.* Как мы описываем пространство, которое видим: типы и ранги объектов // Труды Межд. Семинара Диалог'96 по компьютерной лингвистике и ее приложениям. Пущино, 4–9 мая 1996 г. — М., 1996. — С. 109–111.
5. *Красильщик И. С., Рахилина Е. В.* Предметные имена в системе «Лексикограф» // НТИ. Сер. 2. — М., 1992. — № 9. — С. 24–31.
6. *Леонтьева Н. Н.* К теории автоматического понимания естественных текстов. Часть 2. Семантические словари: состав, структура, методика создания — М.: Изд-во МГУ, 2001.
7. *Леонтьева Н. Н.* Автоматическое понимание текстов: системы, модели, ресурсы: Учеб. пособие для студ. лингв. фак. вузов — М.: Изд. Центр «Академия», 2006. — 304 с.
8. *Нариньяни А. С.* Кентавр по имени Теон: тезаурус + онтология // Труды Межд. семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. — Аксаково, 2001. — Т. 1. — С. 184–188.
9. *Новый* объяснительный словарь синонимов русского языка. Первый выпуск. 2-е изд., испр. — М.: Школа «Языки русской культуры», 1999.
10. *Падучева Е. В.* «Риск — благородное дело»: о системе значений глагола рисковать // Труды Межд. семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. — Аксаково, 2001. — Т. 1. — С. 189–194.
11. *Семенова С. Ю.* Параметризация как метод познания и как языковой механизм // Логический анализ языка. Квантификативный аспект языка — М.: Индрик, 2005. — С. 466–476.
12. *Семенова С. Ю.* Если семантический класс широк для слова (К представлению лексики в машинном словаре) // Компьютерная лингвистика и интеллектуальные технологии: Труды Межд. конф. «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.) — М.: Изд. центр РГГУ, 2007. — С. 492–501.
13. *Семиотика* и информатика. Выпуск 32. Материалы к Интегральному словарю современного русского литературного языка (образцы словарных статей). — М., 1991.

Подход к извлечению фактов из текста на основе онтологии[1]

An ontology-based approach to fact extraction

Сидорова Е. А. (lena@iis.nsk.su), **Кононенко И. С.** (irina_k@cn.ru)
Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск

Предлагается подход к решению задачи извлечения из текста фактической информации. Используемая база знаний включает онтологию предметной области, словари предметной лексики, модель сегментации документов и схемы извлечения фактов, которые связывают термины словаря с элементами онтологии.

1. Введение

Многие информационные системы (ИнС) предъявляют весьма схожие требования к сервису анализа текста, которые сводятся к задаче преобразования слабо-структурированного текста к хорошо структурированной информации. Отличия заключаются в предметной области и структуре извлекаемых знаний.

Предлагаемая технология ориентирована на анализ документов жанра деловой прозы, для которой характерны следующие особенности: ограниченность предметной области и языка документов, наличие строгой модельной ситуации (определяемой характером автоматизации или назначением ИнС), четкость функций каждого сообщения, что позволяет сконцентрировать анализ вокруг наиболее значимых понятий предметной области. Именно такие документы являются важнейшими с точки зрения компьютерной обработки для самых различных ИнС.

Тексты деловой прозы выделяются в Национальном Корпусе Русского Языка (НКРЯ) [1] в рамках системы метапризнаков: значения признака «сфера функционирования» позволяют извлекать из корпуса и исследовать официально-деловые, публицистические, производственно-технические и учебно-научные тексты. Дополнительное привлечение признака «тип текста» позволяет дифференцировать документы по жанру (деловое письмо, научная статья и т.д.).

2. Роль онтологии предметной области при обработке текста

Основная особенность предлагаемого подхода состоит в том, что процесс анализа организован под управлением онтологии, которая расширяется за счет полученной в результате анализа информации, что, в свою очередь, является основой пополнения лингвистической базы знаний — цикличность этого процесса отмечается, например, в [2].

Отличительной чертой такого подхода является ориентация используемых лингвистических описаний на конкретные предметные знания. Словари, помимо универсальной и жанровой, содержат предметную лексику: однословные и многословные термины (словокомплексы), а также лексические шаблоны, которые представляют имена понятий и информационных объектов данной предметной онтологии (с помощью системы семантических классов, признаков и отношений). Кроме того, выбор правил анализа (как поверхностно-синтаксических, используемых для сборки словокомплексов, так и правил сборки фактов) непосредственно определяется спецификой предметной области и структурой целевой онтологической информации. Онтология определяет формат данных, которые хранятся в ИнС, и то, какую именно информацию необходимо извлекать из текста документа. Результат анализа документа представляется в виде семантической сети информационных объектов, являющихся экземплярами понятий и отношений, заданных онтологией предметной области.

[1] Работа выполняется при финансовой поддержке Президиума РАН (ИП СО РАН № 2/12 «Формальные языки и методы спецификации, анализа и синтеза информационных систем» в рамках программы фундаментальных исследований Президиума РАН № 2), РФФИ (проект № 09-07-00400).

Онтологический подход является прямым продолжением и развитием семантически ориентированного подхода к анализу и пониманию отдельного запроса и связного текста, в течение ряда лет разрабатывавшегося коллективом ИСИ СО РАН [3]. Преимущественное использование лексико-семантической информации не исключает применения частичного синтаксического анализа и синтаксических ограничений, накладываемых на семантический каркас концептуальных схем фактов (см. разделы 4.2 и 5). Известные системы различает полнота и роль синтаксического анализа в процессе извлечения фактической информации из текстов. Так, технология [4] предполагает построение полного семантико-синтаксического дерева предложения, к которому применяются шаблоны (своего рода фильтры), описывающие искомые факты. В нашем подходе, как и в [5], синтаксический анализатор применяется локально (при обнаружении ключевых единиц и их конфигураций), в частности, предусмотрено определение актантных позиций предикатных слов, ср. [6].

3. Иллюстрирующий эксперимент

Предлагаемый подход иллюстрируется на примере разработки сервиса анализа документов для информационного ресурса «Хроники СО АН».

3.1. Описание предметной области

Базовая задача анализа сообщений из архива хроник заключается в извлечении названий организаций — структурных подразделений академии наук, упоминаний персон, их ученых званий и степеней, выявление связей между персонами и организациями, а также их изменения во времени. В частности, факт переименования в онтологии отражается с помощью введения для соответствующих атрибутов множества имен с датами начала и конца их действия. Предметная онтология ИнС «Хроники СО АН» включает:

struct Data(Data_begin: data, Data_end: data);
struct Naming(Name: string, Время_действия: Data);
class *Персона* (Фамилия: Naming; Имя: string; Отчество: string; Инициалы: string; ПолноеИмя: string; Звание: domen_Звания; Степень: domen_Степени)
class *Организация* (Название: Naming; Аббревиатура: Naming)
class Институт : Организация
class Экспедиция : Организация **…**
relation *Сотрудник* <Персона, Организация> (Должность: domen_Должности; Дата: Data) **…**

3.2. Особенности подъязыка документов

Данный эксперимент основан на электронной коллекции документов (http://chronicle.iis.nsk.su/catalogue.aspx), единицами которой являются тексты-описания исторических событий, связанных с деятельностью Сибирского отделения Академии наук. В текстах архива излагаются наиболее существенные факты научной и научно-организационной деятельности Сибирского отделения АН СССР.

В настоящее время в архиве содержатся описания 1242 событий. Ниже приведены примеры из электронного архива, в которых выделены фрагменты, соответствующие извлекаемым объектам и фактам:

(1) ***Директору Института экспериментальной биологии и медицины СО АН СССР докт. мед. наук Е. Н. Мешалкину*** присуждена Ленинская премия за разработку новых операций на сердце и крупных кровеносных сосудах.

(2) ***В составе Сибирского отделения АН СССР*** организована самостоятельная ***Лаборатория измерительной и вычислительной электроники, которую возглавил чл.-корр. АН СССР В. Н. Авдеев***.

(3) Существование в плазме бесстолкновительных ударных волн теоретически предсказано ***чл.-корр. АН СССР Р.З. Сагдеевым (Институт ядерной физики СО АН СССР)***.

Тексты архива, являясь отрывками, извлеченными из документов различных жанров (официальных постановлений, деловых писем, газетных статей и т.п.), характеризуются небольшим размером (от одного до 4–5 предложений) и утрачивают жанровые особенности, присущие концептуальной структурной организации первоисточников. Однако тексты сохраняют лексические и синтаксические особенности деловых документов данной предметной области:

- номенклатурная лексика — наименования организаций, должностей и званий;
- шаблонная, унифицированная лексико-грамматическая структура словосочетаний, представляющих составные наименования организаций (*Институт экспериментальной биологии и медицины СО АН СССР*);
- аппозитивные конструкции, представляющие дескрипции персон, включая наименования ученых степеней, званий и составные имена (*академик Михаил Алексеевич Лаврентьев*);
- сокращения (*докт. техн. наук, доктор геол.-мин. наук, чл.-корр.);*
- однородные и скобочные конструкции, используемые для увеличения семантической емкости предложения;
- высокочастотное употребление двучленной (безагенсной) страдательной конструкции

с инвертированным порядком слов (*в состав Сибирского отделения включен Институт леса АН СССР*).

4. Представление лингвистических знаний

Лингвистическая база знаний содержит всю совокупность лингвистических знаний, необходимых для анализа текста, и включает словари предметной лексики, модель сегментации документов и схемы извлечения фактов.

4.1. Словарный компонент

При формировании словарей использовалась словарная технология, описанная в [7]. Она позволяет создавать предметные словари, которые включают однословные и многословные термины (фиксирующие частотные в анализируемом подъязыке словосочетания), а также шаблонные лексические конструкции (позволяющие определять произвольные символьные выражения, в том числе выражения, маркирующие границы сегментов). Создаваемые словари могут содержать семантические характеристики, а также накапливать статистику встречаемости терминов в текстах.

Для задачи анализа сообщений хроник были разработаны следующие словари.

1. Словарь предметной лексики, включающий имена, фамилии известных ученых, локативную лексику и термины, связанные с деятельностью научных организаций (наименования научных должностей, степеней, званий, типов мероприятий и организаций, релевантные предикаты и т.п.), — около 3 тыс. терминов.
2. Словарь лексических конструкций, включающий шаблоны дат и наименований организаций, сокращения и служебные конструкции, — около 700 шаблонов.

С целью минимизации объема ручной работы было осуществлено автоматическое начальное наполнение словарей. Для первого словаря применялись методы обучения, использующие универсальный морфологический словарь (www.aot.ru). Для второго был разработан модуль, который по набору опорных слов-классификаторов (*институт, филиал, президиум* и т.п.) и списку аббревиатур (*АН, РАН, СО РАН* и т.п.) формирует шаблоны наименований организаций вида:

[ИСИ СО РАН] =
институт**…**_систем_информатики**(**_
[ершова])_[СО РАН]
иси_**[СО РАН]**

Для определения основ слов из левого контекста опорного слова (когда в наименовании организации опорному слову предшествует цепочка согласованных с ним прилагательных) используется морфологический словарь. В дальнейшем эксперт вручную исправляет ошибки в аббревиатурах, устанавливает эквивалентность наименований, отмечает необязательные фрагменты, формирует иерархию шаблонов и т.п. При пополнении словаря в качестве маркеров правых границ шаблонов служат шаблоны уже известных (вышестоящих) организаций.

4.2. Модель извлечения фактов

Факт, представляя собой зафиксированное в высказывании (языковом выражении) эмпирическое знание об объектах, их свойствах и ситуациях, может быть формализован в виде когнитивной схемы, соотносящей его с понятиями и отношениями онтологии. Каждый факт имеет свой тип — название отношения и список его аргументов (например, отношение *Сотрудник*). Модель извлечения факта из текста должна учитывать множество языковых способов репрезентации данного отношения носителями подъязыка и обеспечивать их трансформацию в формальную структуру факта. Такую модель мы будем называть схемой извлечения факта (СИФ).

Формально, СИФ — это тройка вида < A, Res, C >, где A — множество дескрипторов аргументов факта, где дескриптором может быть тип словарной единицы, класс информационного объекта (понятие или отношение онтологии) или тип служебного факта.

Res = < t, op(t), P > — результат применения СИФ, где

1. t — задает тип элемента (класс нового объекта или один из аргументов);
2. op(t) — тип операции (создание и/или редактирование аргумента), применяемой, если выполнены ограничения C;
3. P — множество правил для формирования/редактирования объекта. Каждое правило ставит в соответствие атрибуту результирующего объекта либо точное значение, либо значение атрибута одного из аргументов.
4. C — множество ограничений, накладываемых на характеристики аргументов факта. Выделяются следующие ограничения:
5. условия на морфологические и семантические характеристики аргументов схемы (например, arg1.Падеж = рд, arg1.SemClass=Лок)
6. ограничение синтаксической сочетаемости вершин синтаксических групп, реализующих аргументы схемы (например, Synt = Согл(число, падеж), см. схему 2),
7. структурно-текстовые ограничения на взаиморасположение аргументов в тексте: позиция

аргументов относительно друг друга, тип контактности (например, Contact=Common — аргументы могут разделяться в тексте знаками препинания и/или незначимыми словами), тип сегмента.

Приведем пример типичной схемы:

Scheme Персона_с_инициалами **(1)**
arg1: Term*::ФИО*(фио-тип: *фам*)
arg2: Term_lex*::инициалы*()

Condition Position = preposition_priority, Contact = absolute

⇨ Object::*Персона*(Фамилия: arg1.Name, Инициалы: arg2.Value)

Данная схема имеет два аргумента, которые описывают элементы из словарей разного типа (словарь предметной лексики и словарь шаблонов). Позиционное ограничение определяет контактность терминов в тексте, а также указывает на тот факт, что расположение фамилии справа от инициалов является приоритетным в случае, когда есть альтернатива (это позволяет корректно обрабатывать ситуации вида *Г. Петров И.И.*). В результате будут формироваться объекты класса Персона с двумя определенными атрибутами (*Фамилия* и *Инициалы*). Эта схема применима к любому контексту типа *ФИО*, в котором фамилия сопровождается инициалами.

5. Извлечение информации из текста

Процесс обработки текста включает следующие этапы: графематический анализ, лексический анализ, сегментация, морфологический анализ, сборка фактов и формирование контента документа.

5.1. Сегментация

Предусмотрена возможность осуществлять два вида сегментации текста — первичную (логическую) и жанровую [8]. В процессе первичной сегментации производится разбиение линейного текста на строковые объекты, оформленные как сегменты и упорядоченные в соответствии с их встречаемостью в тексте. В рассматриваемом примере документы не имеют выраженной жанровой структуры, поэтому процесс сегментации порождает только логические сегменты: абзац, предложение, клауза и т.п.

Разбиение на сегменты используется при сборке фактов, где, при наличии соответствующего структурного ограничения, на вход алгоритму подается не весь текст целиком, а только фрагмент текста. В этом случае алгоритм сборки фактов запускается столько раз, сколько найдено требуемых сегментов.

5.2. Сборка фактов

Процесс извлечения фактов из текста хроник базируется на схемах извлечения фактов, при формировании которых максимально полно учитываются различные способы выражения в текстах объектов и отношений предметной онтологии.

5.2.1. Извлечение объектов.

Извлечение из текста объекта класса Персона, представленного именной группой типа ФИО, демонстрируется схемой 1 в разд. 4.2.

При вводе в текст объекта класса Персона могут быть указаны ученые степень и звание. В этом случае дескрипция объекта, как правило, представляет собой аппозитивную именную группу, в которой к группе типа ФИО примыкают, чаще в препозиции, фрагменты (сокращения или согласованные с ФИО по числу и падежу именные группы), реализующие Степень и Звание. Схема 2 демонстрирует определение значения атрибута Звание, характеризующего объект класса Персона. С помощью аналогичной схемы извлекается значение атрибута Степень.

Scheme Персона_звание : segment *Предложение* **(2)**
arg1: Object::*Персона*()
arg2: Term::*Звание*()

Condition Position = postposition_priority, Contact = common, Synt = Согл (число,падеж)

⇨ arg1(Звание: arg2.Name)

В текстах хроник отмечаются ситуации нереферентного употребления имен собственных, когда идентифицированный в тексте фрагмент ФИО не вводит конкретного объекта класса Персона: *А. П. Виноградова* в контексте *Институт геохимии* ***им. А. П. Виноградова***; упоминание персон в позиции актанта (С рд) предиката *в честь, памяти,* а также *имя* (в контексте *присвоить, получить, носить*), как *акад. А. П. Виноградова* в примере (4). Это случаи, в которых может иметь место тот или иной вариант локальной неоднозначности (наименование персоны vs. фрагмент наименования организации).

(4) Институту геохимии СО АН СССР присвоено имя выдающегося советского ученого акад. А. П. Виноградова.

В первом случае омонимия снимается уже на уровне сборки лексических шаблонов объектов: подстрока *А. П. Виноградова* входит в лексическую конструкцию, реализующую шаблон наименова-

ния объекта класса Организация. Таким образом, шаблон, охватывающий подстроку, соответствующую группе ФИО, имеет отрицательную видимость и объекта не создает.

В остальных случаях снятие неоднозначности требует не только лексического анализа, но и обработки на этапе сборки фактов. Идентификация объекта Персона в указанной актантной позиции позволяет изменить статус найденного объекта на нереферентный (схема 3), означающий, что в БД ему не сопоставляется никакой конкретный объект, а если такой объект уже присутствует в БД, то ему не добавляется никакой новой информации. Одновременно инициируется формирование служебного факта Именование: отношение Именование не представлено в онтологии хроник, но данный факт позволяет извлечь связанную с (пере)именованием дату.

Scheme Имя_Персоны: segment *Клауза* **(3)**
arg1: Term::*Предикат_Имя*()
arg2: Object::*Персона*(Падеж: рд)

Condition Position = preposition, Synt = Упр(arg1, рд)

⇨ arg2(Visibility: false), Fact::*Именование*(second: arg2)

5.2.2. Извлечение отношений

Рассмотрим пример извлечения из текста отношения Сотрудник, аргументами которого являются объекты классов Персона и Организация. Это отношение в текстах хроник представляется как факт сотрудничества персоны в организации в некоторой должности, которая может быть определенной (*директор, научный сотрудник*), неопределенной (*сотрудник, ученый*) либо недоопределенной должностной ролью ‘первого лица организации’ (*глава, руководитель*).

Различные способы репрезентации отношения Сотрудник сводятся к двум основным вариантам (ср. [5]).

1. В первом варианте для репрезентации отношения Сотрудник используется непредикативная конструкция — связь объектов Организация и Персона реализуется в синтаксических рамках именной группы.

В примере (1) реализована наиболее типичная схема: именная группа <Должность + Организация> (построенная по схеме С+Срд) и примыкающая согласованная (по числу и падежу) именная группа, реализующая объект Персона. Вся конструкция неразрывна, хотя возможны различные варианты взаимного расположения групп и сегментации: все компоненты представлены аппозитивной конструкцией в одном сегменте, как в примере (1); все или некоторые компоненты принадлежат разным сегментам, в том числе возможны сегменты скобочного типа, как в примере (3). Заметим, что скобочная структура не предполагает согласованности группы в скобках с остальными компонентами.

В другой типичной схеме связь групп Организация и Персона реализуется через предложно-падежное примыкание (*во главе с, под руководством*).

2. Во втором варианте связь объектов Организация и Персона реализуется предикативно, с помощью эксплицитных глагольных предикатов, включая лексемы, непосредственно репрезентирующие отношение сотрудничества (*принять, уволить, утвердить, назначить, избрать, работать*), предикаты, вводящие должностную роль ‘первого лица’ (*возглавлять, руководить*), а также связочные глаголы (*быть, становиться, являться, занимать, находиться*). В реализации этой связи в актантной позиции регулярно используется группа Должность (*принять в институт техником, назначен на пост директора института*).

В качестве подкласса предикатов, репрезентирующих отношение Сотрудник, в словаре имеются глаголы *руководить, возглавлять*, в семантике которых синкретично выражено значение должностной роли первого лица организации. В процессе анализа примера (2) данная информация извлекается посредством схемы 4, которая применима без каких-либо ограничений на морфологический класс (часть речи) предиката.

Scheme Предикат_ Сотрудник_ первое_лицо **(4)**
arg1: Term::*СотрудникПЛ*()

⇨ Relation::*Сотрудник*(status1:«missing», status2:«missing», Должность_роль: «первое лицо»)

Далее используются схемы, применимые к произвольному предикату класса Сотрудник, представленному в любой глагольной форме, возможной в позиции вершины клаузы (личный глагол, причастие, инфинитив и т.п.). Ограничение синтаксической сочетаемости проверяет согласованность грамматических признаков вершин синтаксических групп, реализующих аргументы схемы, в соответствии со стандартными правилами согласования (Согл) и управления (Упр i). Конкретный вид ограничения определяется значениями морфологических характеристик аргументов. В примере (2) отношение (arg1 схемы) представлено личным глаголом (*возглавил*), а 1-му семантическому актанту отношения (Персона) соответствует подлежащее. Это означает применение ограничения (arg2.Падеж=им и Согл (Род, Число)). В ситуации со страдательным причастием (*возглавляемый*) применяется ограничение (arg1.Упр2=arg2.Падеж), в данном случае это arg2. Падеж=тв. Соответственно, действительное причастие в позиции arg1 схемы означало бы ограничение вида Согл (Род, Число, Падеж).

При отсутствии одного из актантов в пределах клаузы (в примере (2) это Организация) наличие в ней местоименного заместителя (*который*) означает применимость схем разрешения анафоры. Восстановление антецедента происходит в два этапа. Сначала путем проверки падежной формы местоимения уточняется факт анафорической замены второго семантического актанта (arg2.Упр2 = arg1.Падеж). Затем происходит собственно установление антецедента (схема 5, требующая согласования местоименного заместителя *который* с объектом, претендующим на роль антецедента), что и завершает процесс извлечения факта на основе отношения Сотрудник из текста (2).

Scheme Сотрудник_Антецедент_Орг: segment *Предложение* **(5)**
arg1: Relation::*Сотрудник*(status2: «antecedent — left segment»)
arg2: Object::*Организация*()

Condition Position = postposition, Synt = Согл(Род, Число)

⇨ arg1(second: arg2, status2 = «complete»)

5.3. Генерация информационных объектов

После извлечения фактов из текста осуществляется генерация информационных объектов, соответствующих найденным фактам. На данном этапе происходит взаимодействие с БД системы с целью уточнения объектов и их характеристик:

- Слияние референтных объектов на основе результата процедуры поиска соответствующих объектов в БД (если двум объектам сопоставился один объект БД, то делается вывод о тождестве данных объектов);
- Уточнение неявно выраженных характеристик, например, если в отношении Сотрудник атрибут Должность_роль = *«первое лицо»*, то определяется значение атрибута Должность на основании информации о типе Организации и названии руководящей должности для данного типа:

Сотрудник.Должность_роль = «*первое лицо*» +
Организация.Тип = «*лаборатория*»
⇒ Сотрудник.Должность = «*заведующий лабораторией*».

Для того чтобы сформировать результирующий контент документа, необходимо:

1. обеспечить контроль корректности значений атрибутов информационных объектов, полученных в результате анализа;
2. идентифицировать полученные объекты, т.е. заполнить ключевые атрибуты и сопоставить с объектами базы данных ИнС (если такие существуют);
3. добавить объекты в информационное пространство системы и связать их с документом.

При редактировании объекта в БД могут возникать противоречия между старыми и новыми значениями его характеристик. Для решения данной проблемы была выбрана стратегия сохранения всех данных с указанием даты.

6. Заключение

Проводимое исследование по извлечению фактической информации из текстов хроник ограничено отношениями между персонами и организациями, а также между организациями, с учетом локативных и временных характеристик этих отношений. Схемы описывают реализации фактов в рамках именных и предикативных конструкций, которые могут осложняться анафорой и однородными группами (*члены-корреспонденты АН СССР А. А. Трофимук, М. М. Шемякин*).

В ближайшее время планируется расширение лингвистической базы знаний средствами репрезентации объектов и отношений онтологии и совершенствование собственно лингвистического анализа в аспекте нерешенных проблем. В частности, одним из источников ошибок являются составные наименования объектов и атрибутов при установлении референции (*Председателем РИСО утвержден акад. С. Л. Соболев, его заместителем — акад. А. Л. Яншин*. — антецедент не находится) и обработке сочинительного сокращения (*с институтами: геологии, горного дела, биологическим, радиофизики и электроники* — 4 или 5 организаций?). Требуют внимания и ситуации со снятым или нереальным отношением (*бывший директор*, *не переизбран на пост ректора,* предполагается принять на должность).

В дальнейшем предполагается проведение эксперимента на широкой тестовой базе (новости и постановления Президиума СО РАН).

Литература

1. *Национальный* Корпус Русского Языка www.ruscorpora.ru.
2. *Nedellec C. and Nazarenko C.* Ontology and Information Extraction: A Necessary Symbiosis // Ontology Learning from Text: Methods, Evaluation and Applications. Buitelaar P., Cimiano P. and Magnini B.(eds .), IOS Press Publication:2005.
3. *Нариньяни А. С.* Автоматическое понимание текста — новая перспектива // Труды международного семинара Диалог'97 по компьютерной лингвистике и ее приложениям. М.: 1997. С. 203–208.
4. *Ермаков А.Е.* Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей // Труды международной конференции Диалог' 2007 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2007.
5. *Гершензон Л. М., Ножов И. М., Панкратов Д. В.* Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности // Труды международной конференции Диалог'2005 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2005. С. 97–101.
6. *Азарова И. В., Гребеньков А. С., Ландо Т. М.* Использование маркеров актантных позиций при анализе деловых текстов для расширения логической схемы предметной области // Труды международной конференции Диалог' 2008 «Компьютерная лингвистика и интеллектуальные технологии». М.: РГГУ, 2008. Вып. 7 (14). С. 11–16.
7. *Сидорова Е. А.* Многоцелевая словарная подсистема извлечения предметной лексики // Труды международной конференции Диалог' 2008 «Компьютерная лингвистика и интеллектуальные технологии». М.: РГГУ, 2008. Вып. 7 (14). С. 475–481.
8. *Кононенко И. С., Сидорова Е. А.* Обработка делового письма в системе документооборота // Труды международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. Протвино: 2002. Т. 2. С.299–310.

Модели и методы анализа иерархически структурированных текстов

Models and methods for the analysis of hierarchically structured texts

Скатов Д. С. (ds@dictum.ru),
Ерехинская Т. Н. (te@dictum.ru), **Окатьев В. В.** (oka@dictum.ru)
ООО «Диктум», Нижний Новгород, Россия

В статье обсуждается задача анализа иерархически структурированных текстов (законы, кодексы, стандарты). Дано описание задачи, способы применения результатов анализа и обзор разработок в области. Описаны разработанные модели и методы анализа иерархически структурированных текстов.

1. Введение

Большинство исследователей под обработкой естественно-языкового текста традиционно понимают обработку текста, представляющего собой набор предложений без выраженной структуры. В настоящее время становится актуальной задача обработки документов, обладающих высокой степенью формализованности и, как следствие, внутренней иерархической структурой. К ним относятся, в первую очередь, юридические тексты, нормативно-техническая документация, описания стандартов. Интерес к анализу подобных текстов обусловлен тем, что учет информации об иерархии в системах поиска и анализа документов позволяет предоставить их пользователям новые инструменты, повышающие эффективность работы ([7, 8, 9, 10]).

При поиске информации в коллекции сложно структурированных текстов пользователю недостаточно одного лишь списка релевантных документов в качестве поисковой выдачи по причине больших объемов и высокой сложности документов. Повышение эффективности поиска в таких документах может быть достигнуто, если пользователь будет получать в качестве поисковой выдачи не только документы, но и цитаты из них — точные дословные выдержки из текста, обладающие смысловой законченностью. Цитаты могут быть получены с помощью анализа иерархической структуры текстов, и далее могут быть уточнены с применением синтаксического анализа. В результате пользователь получает компактную поисковую выдачу, в которой отсечен значительный объем информации, нерелевантный запросу. Способ поиска информации, основанный на извлечении цитат из текстов, описан в [14].

Задача извлечения цитат тесно связана с задачами автоматического реферирования текстов [9, 10], поэтому предложенные в статье модели и методы могут быть эффективно использованы в этой области.

Сложность задачи анализа иерархически структурированных текстов обусловлена следующими их свойствами:

1) как правило, разметка заголовков и маркеров (с помощью стилей, тэгов и т.д.) в документе присутствует лишь частично или отсутствует;
2) заголовки из разных уровней иерархии могут быть неотличимы по виду;
3) заголовок и ссылка на него в тексте могут иметь одинаковый вид;
4) богатство конфигураций непрерывных текстовых фрагментов: предложение может состоять из нескольких таких фрагментов, один фрагмент может включать несколько предложений, группа предложений может быть вложена в предложение в виде комментария.

Эти сложности преодолены в рамках подхода, рассматриваемого в статье.

2. Обзор

До последнего времени не было разработано математических моделей, пригодных для построения систем анализа иерархически организованных текстов ([1, 2, 3]). В работах [4, 5, 6] рассматривается задача деления на предложения линейного текста, при этом возможное наличие в тексте иерархической структуры не учитывается. В публикации [7]

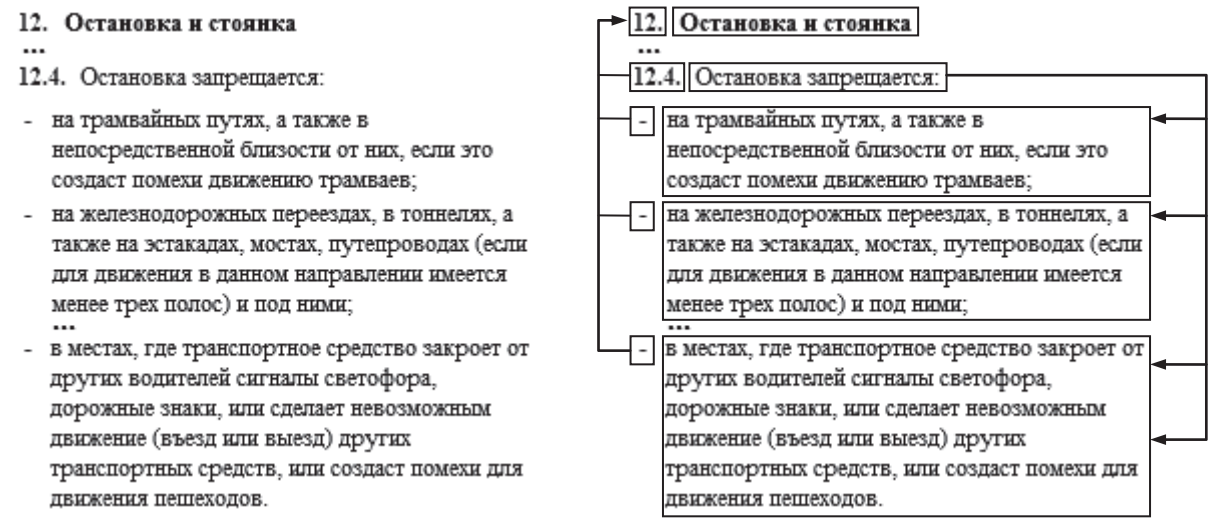


Рис. 1. Фрагмент текста ПДД (слева) и результат анализа его иерархической структуры (справа)

признается важность учета иерархии текстов при поиске, описывается модель поведения пользователей при работе со структурированными документами, анализируются примеры возможных текстовых иерархий. В работе [8] приведены краткие сведения о поиске в иерархически структурированных документах. В публикациях [9, 10] приводится информация об использовании иерархической структуры в задаче автореферирования текстов. Однако во всех упомянутых работах не описаны модели и методы анализа сложно структурированных текстов, которые были бы применимы для решения широкого спектра задач поиска информации.

В HTML-документах информация об иерархии содержится в гипертекстовой разметке. Ее дерево легко можно получить с помощью одной из множества существующих библиотек [15, 16]. В plain-текстах стандартизированная и тривиально определяемая разметка (подобная HTML) отсутствует: перед анализом иерархии определяющую ее разметку (прежде всего — заголовки и маркеры) фактически необходимо выявить, при этом учитывая, что различные тексты, как правило, обладают различной разметкой. В сравнении с методами очистки веб-страниц [17, 18], по сути удаляющих из документа информацию об иерархии, в представленном далее подходе эта информация используется для поиска и получения цитат.

Предложенный в статье подход к анализу иерархической структуры связан с анализом нумерации (маркировки) разделов документа. Определение нумерованных (маркированных) фрагментов позволяет более точно определить границы предложений и сформировать их в виде, удобном для последующей обработки. Полученная в результате разметка текста может быть использована в поисковых системах, в частности, она используется в системе извлечения цитат «Dictum» [12, 13].

Конечно, зная характерную структуру *конкретного документа*, легко разработать утилиту, которая строит дерево этого документа. На практике требуется анализировать сотни документов с *различной внутренней структурой*, что возможно в рамках предлагаемого в данной статье подхода.

Поясним суть задачи анализа иерархии текста на примере текста Правил Дорожного Движения [11]. Рассмотрим фрагмент, содержащийся в п. 12.4 этого текста. На Рис. 1 слева показан исходный текст.

Под иерархическим анализом текста понимается выполнение следующих действий:

1) определение в исходном тексте элементов маркировки — заголовков,
2) определение вложенности групп заголовков,
3) определение отдельных предложений (цитат) из фрагментов текста, расположенных на разных уровнях иерархии.

На Рис. 1 справа показана разметка, полученная в результате анализа иерархии: заголовки и маркеры связаны отношением подчинения (в левой части рисунка), текстовые фрагменты — отношением следования (в правой части рисунка). Из текстовых фрагментов по отношению следования могут быть выделены цитаты.

3. Линейная структура текста

3.1. Текстовый фрагмент

Рассмотрим фрагмент *txt* входного текста. *Текстовым фрагментом* назовем набор (l, r, txt, pl, pr, P).

- *txt* будем называть *значащей частью фрагмента*.
- $P \in D$ представляет собой *разделитель* — символ, который следует после значащей части фрагмента. Между фрагментом и его разделителем допускается наличие пробельных символов. Разделитель является свойством фрагмен-

```
1 Сигналы регулировщика имеют следующее значение :
2 руки вытянуты в стороны или опущены :
  со стороны левого и правого бока разрешено движение трамваю прямо,
3 безрельсовым транспортным средствам прямо и направо, пешеходам разрешено
  переходить проезжую часть ;
```

Рис. 2. Пример разделителей в тексте

та, но не входит в значащую часть *txt*. Пример изображен на Рис. 2: в строках 1–3 части текста, обведенные в сплошную рамку, представляют собой значащие части (*txt*) фрагментов, а символы, заключенные в пунктирную рамку — их разделители.

- Параметры *pl, pr* принимают булевские значения и означают соответственно наличие слева и справа от фрагмента символа переноса строки. В примере на Рис. 2 для трех представленных фрагментов $pl = pr = true$.

В процессе анализа иерархии строится разбиение входного текста α на непересекающиеся фрагменты, в совокупности образующие упорядоченное множество.

3.2. Ячейка

Ячейка представляет собой запись некоторого порядкового значения (величины, номера) в фиксированной системе нумерации. Определим ячейку как набор (T, L, C, os, N). Первые четыре компонента определяют *систему нумерации ячейки*, последний компонент — *числовой эквивалент ячейки* в ее системе нумерации.

- $T \in$ {*числовая, буквенная*} — определяет тип системы нумерации ячейки;
- $L \in$ {*латинская, русская, арабская, римская*} — представляет собой локализацию системы нумерации. Напр., (1) *ячейка «б» записана в системе с русской локализацией, ячейка «XI» — с римской*.
- Для латинской и русской локализаций предусматриваются многопозиционные системы нумераций. В них допускаются порядковые значения, составленные из нескольких символов алфавита. Напр., при такой нумерации допустимы ячейки (2) *«аа» и «аб»*.
- $C \in$ *{верхний, нижний}* — регистр системы нумерации, которой принадлежит данная ячейка. Напр., (3) *ячейка «A» записана в системе с верхним регистром*. Свойство регистра не задано для ячеек с арабской локализацией.
- $os \in D$ представляет собой открывающий символ данной ячейки. Напр., (4) *ячейка «.а» имеет открывающий символ «.»*. Если ячейка не имеет открывающего символа, то $os = null$.
- N — числовой эквивалент ячейки в ее системе нумерации. Строится функция φ: если $\alpha \in D^*$ представляет собой запись какого-либо порядкового значения в некоторой системе нумерации, то $\varphi(\alpha) = N$ представляет собой число, однозначно определяющее эту запись в данной системе нумерации; иначе $\varphi(\alpha) = \infty$. Полагается $N = \varphi(\alpha)$. Напр.: (5) *α = «VII» — это запись порядкового значения в системе нумерации с римской локализацией, верхним регистром, а число $\alpha = 7$ однозначно определяет эту запись в данной системе нумерации. α = «остановка» не является записью порядкового значения в какой-либо системе нумерации, поэтому $\varphi(\alpha) = \infty$.*

На множестве всех ячеек вводятся отношения *равенства* «=» и *непосредственного следования* «$\prec$». Отношение равенства «=» определяется как равенство векторов:

$Cell_1 = (T_1, L_1, C_1, os_1, N_1) = (T_2, L_2, C_2, os_2, N_2) = Cell_2 \Leftrightarrow$
$\Leftrightarrow os_1 = os_2,\ T_1 = T_2,\ L_1 = L_2,\ C_1 = C_2,\ N_1 = N_2.$

Отношение непосредственного следования «$\prec$» определим так: $Cell_1 \prec Cell_2 \Leftrightarrow (os_1 = os_2,\ T_1 = T_2,\ L_1 = L_2,\ C_1 = C_2)$ & $N_2 - N_1 = 1$. Напр., (6) *«а» $\prec$ «b», но «а» $\not\prec$ «c»*.

Множество всевозможных ячеек, распознаваемых системой, описывается композицией конечных автоматов, реализация которых и выполняет определение ячеек в тексте.

3.3. Заголовок

Заголовок представляет собой конечный набор ячеек с дополнительными свойствами. Определим его набором: (T, *Cells*, *txt*, *pos*, *os*, *cs*, *pr*, *t*).

- $T \in$ {*пустой, маркированный, нумерованный, корневой*} — определяет тип заголовка.
  - *Маркированный заголовок* используют для обозначения серии однотипных пунктов. В неразмеченных текстах такие заголовки представляются некоторым декоративным символом — маркеро м, напр. (7) *«*» или «–»*.
  - *Пустой заголовок* часто встречается в неформатированных текстах. В начале каждой строки, которая логически представляет собой новый пункт, можно разместить виртуальный маркер ∅.

- *Корневой заголовок* вводится в рассмотрение искусственно как будущий корень иерархии.
- *Нумерованный* заголовок определяется свойством *Cells*. Они различаются согласно порядку нумерации. Если T = *нумерованный*, то $Cells = (Cell_1, \ldots, Cell_k)$ представляет собой набор ячеек, образующих данный заголовок, иначе *Cells* = *null*. Напр., (8) *заголовок «1.б» имеет следующий набор ячеек*:

$((\text{числовая}, \text{арабская}, \varnothing, \text{null}, 1), (\text{буквенная}, \text{русская}, \text{нижний}, ".", 2))$.

- *txt* представляет собой часть исходного текста, определяющую заголовок (если T = *пустой*, то *txt* = *null*).
- *pos* задает расположение *txt* в исходном тексте, если $T \neq$ *пустой*, и ту позицию, куда может быть вставлен виртуальный маркер, если T = *пустой*.
- Параметры $os, cs \in D$ представляют собой соответственно открывающий и закрывающий символы заголовка. Напр., (9) *заголовок «(1.а)» имеет открывающий символ* os = «(», *закрывающий символ* cs = «)».
- Для каждого текста можно указать набор префиксов — слов, которые могут открывать заголовок. Множество префиксов P упорядочивается лексикографически, и его элементы нумеруются согласно этому порядку, начиная с нуля. Тогда *pr* представляет собой номер префикса данного заголовка. Напр., (10) *если* P = {*раздел*, *статья*}, *то заголовок «Статья 1.1» имеет префикс pr* = 1.
- В процессе анализа некоторые фрагменты могут быть определены как названия соответствующих заголовков — тогда фрагмент изымается из множества F и становится атрибутом t заголовка. Напр., (11) *в тексте «Статья 6.1. Порядок исчисления сроков, установленных законодательством о налогах и сборах» имеется заголовок с префиксом и фрагмент после него, причем фрагмент является названием данного заголовка*.

Множество всех заголовков H строится так, чтобы части текстов, соответствующие различным заголовкам из H, не пересекались. Упорядочим элементы H в порядке их следования в тексте, и установим отношения *непосредственного следования* «$\prec$» и *непосредственного подчинения* «$\lhd$».

Пусть $H_1, H_2 \in H$,
$H_j = (T_j, Cells_j, txt_j, pos_j, os_j, cs_j, pr_j, t_j),\ j = 1,2$.
Отношение непосредственного следования «$\prec$» определяется следующей цепочкой проверок:

- $T_1 \neq T_2 \Rightarrow H_1 \not\prec H_2$;
- $T_1 \neq$ *нумерованный* & T_2 = *нумерованный* $\Rightarrow$ $H_1 \prec H_2$;
- Пусть $T_1 = T_2$ = *нумерованный*, тогда:
- $cs_1 \neq cs_2 \Rightarrow H_1 \not\prec H_2$;
- $os_1 \neq os_2 \Rightarrow H_1 \not\prec H_2$;
- $pr_1 \neq pr_2 \Rightarrow H_1 \not\prec H_2$;
- $length(Cells_1) \neq length(Cells_2) \Rightarrow H_1 \not\prec H_2$;
- $length(Cells_1) = 1$ & $cs_1 = "."$ & $pr_1 = \text{null}$ & $(t_1 = \text{null}\ \&\ t_2 \neq \text{null} \mid t_2 = \text{null}\ \&\ t_1 \neq \text{null}) \Rightarrow H_1 \not\prec H_2$;
- Если $length(Cells_1) > 1$, то:
  - $\exists\, j \in \{1, length(Cells_1) - 1\}: Cell_{1j} \neq Cell_{2j} \Rightarrow H_1 \not\prec H_2$;
  - $k = length(Cells_1),\ \forall\, j \in \{1, k-1\}$ $Cell_{1j} = Cell_{2j} \Rightarrow H_1 \prec H_2 \Leftrightarrow Cell_{1,k}.T = Cell_{2,k}.T$ & $Cell_{1,k} \prec Cell_{2,k}$;
- $T_1 = T_2 =$ *маркированный* $\Rightarrow$ $H_1 \prec H_2 \Leftrightarrow \text{ord}\, H_2 - \text{ord}\, H_1 = 1$;
- $T_1 = T_2 =$ *пустой* $\Rightarrow$ $H_1 \prec H_2 \Leftrightarrow \text{ord}\, H_2 - \text{ord}\, H_1 = 1$;
- в остальных случаях $H_1 \not\prec H_2$.

Отношение непосредственного подчинения «$\lhd$» определяется цепочкой проверок:

- ! ($T_1 = T_2$ = *нумерованный*) $\Rightarrow H_1 \not\lhd H_2$;
- ! ($length(Cells_2) > 1$ & $length(Cells_2) - length(Cells_1) = 1) \Rightarrow H_1 \not\lhd H_2$;
- $pr_1 \neq pr_2 \Rightarrow H_1 \not\lhd H_2$;
- $\exists\, j \in \{1, length(Cells_1)\}: Cell_{1j} \neq Cell_{2j} \Rightarrow H_1 \not\lhd H_2$;
- в остальных случаях $H_1 \lhd H_2$.

4. Иерархическая структура текста

4.1. Древовидная модель иерархии

Иерархию в тексте представим деревом, в котором узлы соответствуют объектам, а ветви задают отношение подчинения (включения). Если у некоторого объекта-вершины имеются непосредственные объекты-потомки, то они находятся с родительской вершиной в отношении непосредственного подчинения.

Будем рассматривать *линейные компоненты текста — заголовки* и *фрагменты* — как вершины дерева. Расширим множество заголовков и фрагментов дополнительным объектом — *параграфом*. *Параграф* — это вершина, которой могут быть подчинено некоторое кол-во фрагментов.

Т.о., моделью иерархической структуры текста является дерево (V, E) (V — множество его вершин, E — множество ребер), заданное как бинарное отношение непосредственного подчинения на V. Множество вершин состоит из элементов вида (T, *Obj*), где тип элемента задан параметром $T \in$ {*заголовок, фрагмент, параграф*}, а параметр *Obj* обозначает объект типа T, представляющий собой данную вершину.

Задача анализа иерархии состоит в построении дерева (V, E), моделирующего иерархическую структуру этого текста. Это построение выполняется в три прохода:

1) на первом проходе строится дерево T_1, состоящее только из заголовков;
2) второй проход выявляет фрагменты и присоединяет их к дереву T_1 в линейном порядке, результатом является дерево T_2; также выполняется определение некоторых фрагментов как названий соответствующих заголовков;
3) третий проход выполняет коррекцию дерева T_2 с целью определения иерархии фрагментов; результатом является дерево T_3.

4.2. Определение иерархии заголовков

Введенные отношения непосредственного следования и подчинения на множестве H используются в методе определения иерархии для этого множества.

Полагаем, что множество H упорядочено по линейному расположению заголовков в тексте. В начало множества H добавим заголовок H_0 с T = *корневой*.

Построим процедуру формирования дерева. В ней параметры *from, to* представляют собой некоторый диапазон элементов множества H, *parent* — элемент H, который является родительским на данном уровне рекурсии, *Father* — узел в дереве T_1, который является предком для других узлов на данном уровне рекурсии.

```
procedure Recurrent (from, to, parent, Father) {
    prev := from; next := from + 1;
    пока (prev != to) {
        если (next = to) {
            V.T := заголовок;
            V.Obj := H_prev ∈ H;
            добавить_потомка (T_1, Father, V);
            если (next — prev != 1)
                Recurrent (prev+1, next, prev, V);
            прервать_цикл;
        }
        если (H_prev ≺ H_next) {
            V.T := заголовок;
            V.Obj := H_prev ∈ H;
            добавить_потомка (T_1, Father, V);
            если (next — prev != 1)
                Recurrent (prev+1, next, prev, V);
            prev := next;
            next:=next+1;
            продолжить_цикл;
        }
        если (parent != null) {
            если (H_next ◁ H_parent) {
                V.T := заголовок;
                V.Obj := H_prev ∈ H;
                добавить_потомка (T_1, Father, V);
                если (next — prev != 1)
                    Recurrent (prev+1, next, prev, V);
                prev := next;
                next:=next+1;
                продолжить_цикл;
            }
        }
        next := next+1;
    }
}
```

Теперь построение дерева можно выполнить вызовом рекуррентной процедуры: **Recurrent** (0, *lenght*(H), null, *root*(T_1)).

4.3. Определение и добавление фрагментов

Для добавления фрагментов к дереву будем выполнять проход по дереву T_1 в ширину (breadth-first). При построении T_2 используется тот факт, что фрагменты заключены между заголовками текста. В процессе прохода по дереву фиксируются участки текста, заключенные между смежными заголовками, в этих участках текста выявляются фрагменты, которые добавляются как потомки соответствующих заголовков.

4.4. Коррекция вершин и финальная обработка

Чтобы T_2 приобрело требуемый вид T_3, необходимо применить к нему процедуру коррекции вершин. Она выполняется в три этапа.

На первом этапе осуществляется вынесение вершин-фрагментов из внутренних уровней во внешние. Решение о перемещении некоторой группы вершин принимается на основании свойств вершин, окружающих эту группу.

Появление иерархии в структуре текста наблюдается в двух ситуациях:

1) Разбиение множества предложений на иерархические группы: совокупность предложений разделяется на множества, каждому из которых присваивается определенный уровень иерархии и позиция в этой иерархии;
2) Разбиение одного предложения на иерархические части: компоненты, на которые таким образом разбивается предложение, могут быть разделены текстовыми вставками, независимыми от данного предложения — например, комментариями.

Возможна также композиция этих ситуаций. Обрабатывать ситуацию (а) позволяют вершины-заголовки, (б) — вершины-фрагменты, подчиняющие себе другие вершины. Вершины-фрагменты выявляются на *втором этапе*.

На третьем этапе выполняется введение в дерево вершин-параграфов и подчинение последовательно идущих фрагментов этим вершинам так, чтобы вершины-параграфы и их потомки отражали структуру абзацев исходного текста.

Детально эти этапы рассмотрены в работе [12].

5. Заключение

Полученные результаты позволяют сделать вывод, что разработанные модели и методы анализа иерархической структуры текста могут быть эффективно использованы в ряде приложений поиска и анализа текстов. Решенная задача является актуальной, т.к. возможность учета иерархии в текстах в настоящее время востребована, однако моделей и методов, которые могут быть использованы для решения широкого класса задач, ранее представлено не было. В рамках проведенных исследований разработана математическая модель линейной и иерархической структур текста, а также методы выявления структуры текста. На основе разработанных моделей и методов построен прототип системы извлечения цитат Dictum, который доступен в режимах браузера и ICQ-клиента на сайте [13].

Исследование проводилось малым инновационным предприятием ООО «Диктум» при поддержке Фонда содействия развитию малых форм предприятий в научно-технической сфере, проект № 6466 «Разработка компьютерной системы извлечения цитат из текстов на естественном языке». Подробные результаты исследований опубликованы в работе [12].

Литература

1. *Кормалев Д. А., Куршев Е. П.* Приложение технологии извлечения информации из текста: теория и практика. // Переяславль-Залесский: ИПС РАН, 2003.
2. *Шереметьева С. О.* Теоретические и методологические проблемы инженерной лингвистики. // М.: ВИНИТИ, 1998.
3. *Лахути Д. Г.* Автоматический анализ естественно-языковых текстов. // М.: ВИНИТИ, 2003.
4. *http://www.aot.ru/docs/fragman.html.*
5. *Липатов А. А., Мальцев А. А.* Методы автоматизации построения и пополнения двуязычных словарей с использованием корпусов параллельных текстов. // Труды международной конференции «Диалог 2006». М.: Изд-во РГГУ, 2006.
6. *Palmer D.* SATZ — An Adaptive Sentence Segmentation System // Report No. UCB/CSD-94–846, Computer Science Division (EECS). University of California: December 2004.
7. *Hertzum M., Lalmas M. and Frokjer E.* How Are Searching and Reading Intertwined during Retrieval from Hierarchically Structured Documents? // INTERACT 2001. Japan, July 2001.
8. *Lalmas M., Reid J. and Hertzum M.* Information-seeking Behaviour in the Context of Structured Documents // ECIR 2003: European conference on IR research No. 25. Pisa, ITALIE: 2002. Vol. 2633, pp. 104–119.
9. *Браславский П., Колычев И.* Автоматическое реферирование веб-документов с учетом запроса. // Интернет-математика 2005. Автоматическая обработка веб-данных. М.: 2005. С. 485–501.
10. *Yang Ch. C., Wang, F. L.* Fractal Summarization for Mobile Devices to Access Large Documents on the Web // Proceedings of the WWW2003. Budapest, Hungary: May 20–24, 2003.
11. *Правила* Дорожного Движения Российской Федерации (в ред. Постановлений Правительства РФ от 08.01.1996 N 3, от 31.10.1998

N 1272, от 21.04.2000 N 370, от 24.01.2001 N 67, от 28.06.2002 N 472, от 07.05.2003 N 265, от 25.09.2003 N 595, от 14.12.2005 N 767, от 28.02.2006 N 109).

12. *Окатьев В. В., Гергель В. П., Алексеев В. Е., Таланов В. А., Баркалов К. А., Скатов Д. С., Ерехинская Т. Н., Котов А. Е., Титова А. С.* Отчет о выполнении НИОКР по теме: «Разработка пилотной версии системы синтаксического анализа русского языка» (инвентарный номер ВНТИЦ 02200803750) // М.: ВНИТЦ, 2008.
13. *http://www.dictum.ru*
14. *Окатьев В. В., Баркалов К. А.* Патент на изобретение «Способ поиска информации», RU 2320005, // М.: Федеральная служба по интеллектуальной собственности, патентам и товарным знакам, 2008. Бюллетень № 8.
15. *http://htmlparser.sourceforge.net/*
16. *http://homepage.mac.com/pauljlucas/software/html_tree/*
17. *http://www.jafsoft.com/detagger/*
18. *http://www.prcyonline.info/totext.html*

Опыт систематизации знаний и интернет-ресурсов для портала знаний по компьютерной лингвистике

Experience of systematizing knowledge and internet resources for a knowledge portal on computational linguistics

Соколова Е. Г. (minegot@rambler.ru)
Российский государственный гуманитарный университет, Москва

Загорулько Ю. А. (zagor@iis.nsk.su), **Кононенко И. С.** (irina_k@cn.ru)
Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск

В статье описывается опыт систематизации и интеграции знаний и интернет-ресурсов по компьютерной лингвистике в интернет-портал знаний. Рассматривается состав и структура объектов портала, место портала среди других каталогов по компьютерной лингвистике, опыт создания двуязычного словаря терминов по компьютерной лингвистике с использованием процедур автоматического извлечения терминов из текстов.

1. Введение

В ходе двухгодичного исследовательского проекта (2007–2008 годы), поддержанного РГНФ, создан интернет-портал знаний по компьютерной лингвистике (КЛ). Общая информация и проблемы, возникающие при описании в виде онтологии портала такой области как КЛ, рассмотрены нами в [1]. В данной статье мы обсуждаем различные аспекты проведенного исследования — принципы отбора информации, особенности портала и его место среди других каталогов по КЛ, классификацию и представление знаний и ресурсов, опыт создания двуязычного (англо-русского) словаря терминов по компьютерной лингвистике с использованием процедур автоматического извлечения терминов из текстов.

Принципы, которых мы придерживались, выбрав парадигму создания порталов знаний, заложенную работами [2, 3], состоят в следующем. 1. Главная цель проекта — представить российским исследователям всестороннюю картину КЛ, не претендуя на полноту в деталях. Тем самым мы стремимся отразить на портале достижения мировой КЛ, но при этом уделить особое внимание российской КЛ. 2. Российские исследования примерно на 15 лет позже, чем западные, начали переходить в стадию технологий и, как правило, не имели финансирования от организаций, требующих общедоступности результатов исследований, в связи с чем, в отличие от западных, редко завершались созданием ресурсов и функциональных систем. При этом отечественные исследования часто основывались на глубоких подходах и содержали интересные идеи, которые могут быть полезны сейчас или в будущем, поэтому нам хотелось включить и кратко охарактеризовать и такие отечественные публикации и материалы. 3. Из имеющихся в Интернете каталогов хотелось взять, прежде всего, «живые» системы и ресурсы, а также связанные с русским языком. 4. В начальной реализации портала язык описания его информационных объектов (проектов, методов и т. д.) — русский. Английский используется для указания исходных английских названий проектов, методов и моделей, например, «Расширенные сети переходов (Augmented Transition Networks (ATN))»; 5. Объекты онтологии сопровождаются только краткими описаниями, содержащими самую общую информацию об объекте. 6. Принципиальным требованием к таким проектам, как портал знаний, является наличие обратной связи с пользователями.

2. Границы описываемой порталом области исследований

В [1] мы показываем, что из нескольких терминов, соответствующих описываемой порталом области исследований, таких как прикладная

лингвистика, автоматическая обработка естественного языка, автоматическая обработка текстов, наиболее адекватен термин «компьютерная лингвистика», подчеркивающий тот факт, что компьютер является необходимой составляющей этой науки. Но недостаточной. Относится ли к КЛ оцифрованный и показанный на экране компьютера традиционный словарь или Грамматический словарь А. А. Зализняка, изданный в виде книги? Судя по тематике докладов на конференциях Диалога, на оба вопроса дается положительный ответ. В первом случае отнесение к КЛ оправдывается тем, что при формировании оцифрованного словаря используются компьютерные методы представления словарей и некоторые элементы формализации содержания словаря, хотя в целом оно остается не формализованным и ориентированным на человека. Во втором случае мы имеем дело с печатным изданием, но оно содержит формализованное описание русской морфологии, которое после введения в компьютер используется автоматизированными системами[1] морфологического анализа и синтеза русского языка. Если использование компьютера не является необходимым, то где граница между лингвистикой и КЛ? Граница есть, так как КЛ имеет свои собственные объект и предмет — преобразования текстов и звучащей речи[2], — которыми традиционная лингвистика не занималась и не занимается. Цели, которые ставятся в задачах преобразования текстов и речи, тем более выходят за рамки традиционной, преимущественно описательной, лингвистики, поэтому раньше КЛ называлась «прикладной». Современная КЛ имеет обширные пересечения с теоретической лингвистикой, искусственным интеллектом и математикой. КЛ использует теоретические достижения лингвистики для построения моделей языка в действии и преобразования текстов и речи, а сама дает методическую базу и ресурсы для проверки гипотез теоретической лингвистики. КЛ пересекается с ИИ в области обработки знаний и информации, выраженной на ЕЯ, использует методы математики для обработки текстов и речи в прикладных системах.

3. КЛ в зеркале онтологии портала знаний по КЛ

Понятия онтологии предметной области «Компьютерная лингвистика» организованы в 5 иерархий «общее-частное»: Объекты исследования, Предметы исследования, Разделы науки, Методы исследования, Научные результаты. Понятия в иерархиях связываются между собой посредством ассоциативных отношений.

Объекты исследования: Речевое произведение (РП) как объективная форма существования и использования естественного языка в виде Текста или Устной речи и Языковые единицы в составе РП, соответствующие различным языковым уровням: Синтаксические, Лексические, Морфологические и Фонетико-фонологические единицы. Для представления связи между целостными РП и их структурными единицами используется отношение «Включение».

Общеметодологический термин «объект исследования» ориентирован на традиционную науку и не совсем точен для КЛ, точнее было бы «объект моделирования». Особенность КЛ состоит в том, что собственно исследованием определенных лингвистических единиц она не занимается, но занимается созданием ресурсов лингвистических единиц — формализованных баз и корпусов, представляющих совокупности таких единиц. Организуя базы и корпуса, КЛ использует достижения традиционной лингвистики для систематизации и разметки единиц. Размеченные корпуса и компьютерные базы, с одной стороны, служат исходным материалом для настройки систем КЛ (например, при машинном обучении), а с другой стороны, позволяют верифицировать результаты лингвистических исследований. Объекты КЛ включают тексты и звучащие отрезки речи, которые обычно не являлись объектами систематического описания в лингвистике. В частности, структура текста как объект такого описания возникает в рамках КЛ для моделирования структуры текстов в системах генерации текстов.

Предмет исследования — аспект в исследуемом/моделируемом материале, на который направлена научная деятельность. Предметом исследования в КЛ являются 1. Процессы, связанные с функционированием языковых единиц в коммуникации. Среди них выделены процессы Анализа речи, Синтеза Речи, Анализа текста и Синтеза текста. В модели процессов выделены подпроцессы, которые относятся к уровням языка. Так, например, класс понятий Анализ текста представлен в иерархии подклассами: Сегментация текста, Морфологический анализ, Синтаксический анализ, Семантическая интерпретация, Анализ дискурса. Процессы анализа и синтеза имеют разный состав и не рассматриваются как обратимые. 2. Прикладные процессы, имеющие практическую ценность, отвечающие

---

[1] При подготовке Грамматического словаря русского языка А.А. Зализняк инициировал создание «Обратного словаря русского языка» с использованием методов машинной обработки материала, без которого работа над Грамматическим словарем не могла бы быть эффективно закончена.

[2] Ср. определение Автоматической Обработки Текстов, которое мы цитировали в [1]: «преобразование текста на искусственном или естественном языке с помощью ЭВМ» (с. 14, В.М. Андрющенко). Другими науками, порожденными компьютером, являются Искусственный Интеллект (Artifitial Intelligence), Вычислительная математика и Информатика (Computer science).

определенному социальному запросу, к которым относятся машинный перевод, автоматическое реферирование, идентификация говорящего по голосу и многое другое.

Методы так же, как и объекты, относятся скорее к способам моделирования, а не к способам исследования, хотя такие традиционные лингвистические методы, как, например, компонентный анализ, также включены в иерархию. Иерархия методов составляет центральную иерархию портала, так как КЛ является по сути методологической наукой. К методам отнесены Средства представления знаний, Грамматические формализмы, Методы теоретической лингвистики, Формальные механизмы и методы обработки ЕЯ, Методы оценки работы алгоритмов и систем. Теории языка в КЛ также носят методологический характер, т. е. объясняют не устройство систем конкретных единиц языка, а способы моделирования языковых средств для использования автоматическими системами. Такими теориями — моделями, породившими компьютерные системы, — являются структурная модель «Смысл-Текст» И. А. Мельчука, Ю. Д. Апресяна (компьютерные системы ЭТАП-2, RealPro и др.) и функциональная модель М. А. К. Хэллидея (компьютерные системы Penman, KPML, AGILE и др.). Модели Н.Хомского также представлены в иерархии методов как структурные модели, однако, если стандартную синтаксическую теорию Н. Хомского естественно рассматривать в рамках КЛ, где она повлияла на создание других уровневых моделей, в частности, российских (И. А. Мельчука, Ю. С. Мартемьянова), то теория GB перешла в ранг лингвистических теорий, позволяющих объяснять синтаксические явления.

В основе иерархии Разделов КЛ лежит классификация базовых теоретических и прикладных направлений КЛ: Моделирование языка и языковой деятельности (с разделами Автоматическая обработка текста (АОТ), Речевые технологии (РТ), Формализация описаний языковых средств и свойств речевых произведений) и Создание прикладных систем. В зависимости от направления моделирования (анализ или синтез) в первых двух разделах Моделирования языка и языковой деятельности выделены, соответственно, подклассы Понимание текста и Генерация текста, Распознавание речи и Синтез речи. В зависимости от объекта обработки (текст или звучащая речь), Прикладные системы разделены на Прикладные системы АОТ и Прикладные системы РТ.

Научные результаты представлены следующими классами: Технологии и программные продукты, Прикладные системы, Лингвистические ресурсы. Последний класс делится на такие классы, как Словари, Корпуса и Лингвистические БД. Класс Лингвистические БД, в свою очередь, разделен на Грамматические, Лексико-семантические, Семантико-синтаксические и Синтаксические ресурсы, а также Морфологические БД и Речевые БД. Корпуса разделяются на Корпуса текстов и Речевые корпуса.

4. Место портала по КЛ среди других интернет-каталогов

В Интернете имеются каталоги разработок и публикаций по КЛ, кратко рассмотренные нами в [4]. Наиболее крупный зарубежный каталог LINGUIST List [5] послужил прототипом для сайта «Российская лингвистика (RUSLING)» [6], созданного в Отделении лингвистических исследований ВИНИТИ РАН около 20 лет назад. Сайт «Лингвистика в России: ресурсы для исследователей», создан в феврале 2006 года по инициативе НИВЦ МГУ им. М. В. Ломоносова и ГОУВПО «Казанский государственный университет им. В. И. Ульянова-Ленина» [7]. Особенность Портала по КЛ по сравнению с этими каталогами состоит в том, что на нем представлена более узкая предметная область — КЛ, а информация структурирована в соответствии с онтологией предметной области. Полезными для КЛ являются такие источники, как Кругосвет, где имеются статьи, отражающие современное состояние лингвистики, например, статья, описывающая понятие дискурса, и интернет-энциклопедия Википедия, в которой можно найти полезную информацию о моделях, методах, интернет-ресурсах, персонах и организациях современной КЛ.

В Интернете представлена и более узко специализированная информация по отдельным направлениям КЛ. В качестве примера можно привести российский сайт «Речевые технологии» [8], всесторонне охватывающий теоретические и прикладные аспекты развития данного направления (технологии, программные средства, коллективы разработчиков, конкретные системы и т. п.).

Наиболее полными, точными и долговечными являются узкоспециализированные каталоги, поддерживающиеся западными исследователями, например, каталог систем генерации текстов [9]. В нем содержится информация обо всех западных системах и проектах по созданию таких систем, всего на момент написания данной статьи 383 системы. Например, для проекта AGILE приводится следующая информация:

Имя системы: AGILE
Разработчики: Krujiff, Korbayová, Teich, Hartley, Bateman, Sharoff, Scott, Staykova, Sokolova
Даты разработки: 1999–2001
Языки: Bulgarian, Czech, Russian
URL (if available) http://www.itri.bton.ac.uk/projects/agile who's who
Построен на основе: KPML
Описание: AGILE is a tool which allows a technical author to specify, in a non-linguistic representation, the 'content' of different tasks that can be performed by users of CAD-CAM software. The AGILE system can then automatically express these content specifications in styles appropriate to different sections of a CAD-CAM manual (procedures, ready reference ...) in Bulgarian, Czech and Russian. The generated texts are displayed in a browser as hyperlinked documents. No expertise in knowledge representation is required, although some training with the interface is needed. The system has been evaluated and the results are described in the relevant project deliverables.
Ссылки на публикации по проекту (3 ссылки).

Задача создателей этого каталога облегчается тем, что описываемые системы создавались и создаются преимущественно в рамках целенаправленно финансируемых научных проектов, имеющих конкретный срок разработки от 2 до 5 лет и направленных на создание одной конкретной системы или среды.

В отличие от узкоспециализированных каталогов, портал по КЛ охватывает все типы метапонятий, определенных прототипом портала, в соответствии с чем на портале представлена как информация собственно по КЛ (теории, методы и т.д.), так и информация по автоматическим системам КЛ, которая разнесена по онтологическим классам, а соответствующие им объекты связаны в единую сеть содержательными ассоциативными отношениями, обеспечивающими переходы по сети от одного элемента информации к другому.

Так, информация о проекте AGILE распределена по онтологическим классам следующим образом: Деятельность → Проект → проект AGILE; Объект исследования → Речевое произведение → Текст → Инструкция; Раздел науки → Моделирование языка и языковой деятельности → Автоматическая обработка текста → Генерация текста; Научные результаты и продукты → Прикладная система → система AGILE; Персона → <список исполнителей>; Организация → <список организаций-участников>.

Объекты, соответствующие этим классам, связаны отношениями: «Исследует объект», «Использует результат», «Результат деятельности», «Персона-Участник деятельности», «Публикация о деятельности» и т. д. Например, отношение «Использует результат» связывает проект AGILE со средой для разработки многоязыковых генераторов KPML, который, в свою очередь, позволяет выйти на персону-автора данного ресурса и т. д.

Портал знаний не только обеспечивает гибкое представление информации по КЛ, но и предоставляет пользователям удобные средства поиска и навигации по ней.

Для навигации по контенту портала используется дерево понятий онтологии. При выборе в этом дереве определенного узла пользователь получает список соответствующих ему информационных объектов. Если на объектах выбранного понятия задано отношение включения, по желанию пользователя этот список может быть представлен в виде дерева. Информация о конкретном объекте и его связях отображается в виде html-страницы, при этом объекты и интернет-ресурсы, связанные с данным объектом, представляются на его странице в виде гиперссылок, по которым можно перейти к их детальному описанию.

Портал предоставляет пользователю два вида поиска: простой поиск и расширенный поиск. Простой поиск позволяет находить объекты, в значениях атрибутов которых содержится строка поиска. Расширенный поиск предоставляет пользователю возможность задания запроса в терминах предметной области портала. При этом пользователь может указать не только объекты какого понятия и с какими свойствами он хочет найти, но и задать ограничения на значения атрибутов объектов, связанных с искомым объектом.

Таким образом, портал может служить основой для конечных пользователей при поиске проектов, интернет-ресурсов и методов исследований по КЛ. С другой стороны, на нем могут базироваться библиотечные классификации данной области, которые в настоящее время остаются неразработанными.

В связи с масштабностью поставленной задачи на портале представлены не все проекты, лингвистические группы и исследователи, не полностью охвачены модели и методы, используемые в КЛ. Поэтому работа по развитию онтологии КЛ и наполнению контента портала новыми данными и ресурсами будет продолжена.

5. Словарь терминов КЛ

Для поддержки автоматизации сбора интернет-ресурсов и новостей, относящихся к области КЛ, а также обеспечения визуализации контента портала и поиска в нем информации на разных языках было разработано несколько словарей. Основой этих словарей стали английский, русский и англо-русский словники, построенные в ходе выполнения проекта в 2007–2008 годах.

В качестве материала для создания этих словников были сформированы два корпуса текстов — английский, включающий учебник под ред. Р. Митковa [10] и обзор под ред. Р. Коула [11], и русский, включающий труды конференций «Диалог» за последние три года (2006, 2007, 2008), взятые с сайта конференции Диалог, ввиду отсутствия фронтальных обзоров[3] (о чем мы писали в статье [1]). Оба корпуса были обработаны с помощью технологии автоматического создания терминологической базы по текстам предметной области, авторами которой являются Н. В. Лукашевич и Б. В. Добров [12], в результате чего были получены списки русских и английских терминов, в которые вошли слова и правильные словосочетания (в основном, пары слов —

[3] Учебники по КЛ и близким областям, изданные в России за последние два десятилетия: Шемакин Ю. И. Начала компьютерной лингвистики. М.: Изд-во МГОУ А/О «Росвузнаука», 1992. — 114 с.; Прикладное языкознание. Учебник. (ред. А.С.Гердт). СПб., 1996.; Баранов А. Н. Введение в прикладную лингвистику. Серия «Новый лингвистический учебник». М.: Эдиториал УРРС. 2001; Леонтьева Н.Н. Автоматическое понимание текстов. Системы. Модели. Ресурсы. М.: Academia, 2006, — не дают стройной картины направления, которую мы видим в западных учебниках и обзорах этого же периода.

согласованные Прил+Сущ и Сущ+Сущ в родительном падеже, а также порожденные на их основе трехкомпонентные словосочетания). Русский список был пополнен с помощью технологии [13], примененной к текстам, содержащим определения понятий, атрибутов, доменных значений и объектов, составленных экспертом при описании онтологии КЛ. В результате объем русской терминологической базы достиг 13 тыс. слов, английской — 15,5 тыс. слов. После просмотра и редактирования экспертом количество терминов существенно сократилось и составило, соответственно, 4801 русских и 4972 английских термина. На основе этих списков с учетом статистических данных был создан двуязычный (англо-русский и русско-английский) словарь по КЛ.

Сравнение полученных русских и английских списков показало, что многие единицы в них, с точки зрения эксперта, не являются терминами, кроме того, они содержали не вполне коррелированные наборы понятий. В результате в словарь вошла часть терминов, для которых было установлено межъязыковое соответствие путем сопоставления английского и русского терминов из списков, а другая часть получена путем перевода английских терминов на русский язык.

Полученный словарь включает около 1900 англоязычных и русскоязычных терминов и их переводов (английских — 1866, русских — 1875, общее количество связей — 2199: связей больше ввиду синонимии; все непереведенные слова были автоматически удалены из словаря). Этот результат свидетельствует об ограниченности проведенного эксперимента (нужны значительно более объемные корпуса текстов) и о неполной сопоставимости исследований по КЛ в России (как они представлены в Диалоге за последние три года) и за рубежом.

Кроме того, на основе английской и русской терминологии было разработано два предметных словаря, содержащих морфологическую, статистическую и семантическую информацию. Эти словари использовались для настройки модулей портала, отвечающих за автоматизацию наполнения его контента и сбора новостных сообщений по тематике КЛ.

6. Заключение

В настоящее время портал знаний доступен по адресу http://uniserv.iis.nsk.su/cl. Его контент включает более 600 интернет-ресуров, около 2000 информационных объектов, связанных примерно 4000 отношениями. Пользователь может видеть не только иерархию «общее-частное», заданную на понятиях онтологии, но и иерархии «часть-целое», заданные на информационных объектах.

Работа по развитию онтологии КЛ и наполнению контента портала новыми данными и ресурсами будет продолжена. Разработанный англо-русский словарь будет дополнен и станет базой для визуализации контента портала и поиска в нем информации на двух языках — русском и английском.

Обратная связь с пользователем на данный момент осуществляется через адрес электронной почты, указанный на сайте в разделе «О портале». Планируется организация специального форума для обсуждения онтологии и контента портала.

Авторы благодарят О. Ф. Кривнову за помощь в систематизации и работе с терминологией для разделов по обработке речи, а также Н. В. Лукашевич и Б. В. Доброва за проведение эксперимента по извлечению терминологии.

Литература

1. *Соколова Е. Г., Кононенко И. С., Загорулько Ю. А.* Проблемы описания компьютерной лингвистики в виде онтологии для портала знаний // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2008» (Бекасово, 4-8 июня 2008 г.). М.: РГГУ, 2008. Вып. 7 (14), С. 482–487.
2. *Боровикова О. И., Загорулько Ю. А., Загорулько Г. Б., Кононенко И. С., Соколова Е. Г.* Разработка портала знаний по компьютерной лингвистике // Труды 11-ой национальной конференции по искусственному интеллекту с международным участием КИИ-2008 (г. Дубна, Россия). М.: ЛЕНАНД, 2008. Т. 3, С. 380–388.
3. *Боровикова О. И., Загорулько Ю. А.* Организация порталов знаний на основе онтологий // Компьютерная лингвистика и интеллектуальные технологии: Труды международного семинара «Диалог 2002» (Протвино, 6–11 июня 2002 г.). М.: Наука, 2002. Т. 2, С.76–82.
4. *Загорулько Ю. А., Боровикова О. И., Загорулько Г. Б.* Портал знаний по компьютерной лингвистике: содержательный доступ к лингвистическим информационным ресурсам // Компьютерная лингвистика и интеллектуальные технологии. Электронные публикации Международной конференции «Диалог-2008» (http://www.dialog-21.ru/dialog2008/materials/html/Zagorulko.htm)
5. *LINGUIST List* (http://linguistlist.org/)
6. *«Российская* лингвистика (RUSLING)» (http://rusling.narod.ru)
7. *«Лингвистика* в России: ресурсы для исследователей» (http://uisrussia.msu.ru/linguist/index.jsp)
8. *Портал* «Речевые технологии» (http://speech-soft.ru/)
9. *«John Bateman and Michael Zock's* List of NLG systems»http://www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/NLG-table-root.htm
10. *Mitkov Ruslan* (ed.) The Oxford handbook of computational linguistics // N.Y.: Oxford university press, 2003.
11. *Cole Ronald* (ed.) Survey of the state of the Art in Human Language Technology // 1996. (http://cslu.cse.ogi.edu/HLTsurvey/).
12. *Добров Б. В., Лукашевич Н. В., Сыромятников С. В.* Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции», 2003. С. 201–210.
13. *Сидорова Е. А.* Многоцелевая словарная подсистема извлечения предметной лексики // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2008» (Бекасово, 4–8 июня 2008 г.). М.: РГГУ, 2008. Вып. 7 (14), С. 475–481.

Использование лексико-грамматических баз данных в русской диалектной лексикографии[1]

The use of lexico-grammatical databases in the Russian dialectal lexicography

Тер-Аванесова А. В. (teravan@mail-ru)
Институт русского языка им. В. В. Виноградова РАН

Крылов С. А. (krylov-58@mail.ru)
Институт востоковедения РАН, Институт системного анализа РАН

С помощью СУБД STARLING обогащена построенная ранее лексико-грамматическая база данных (ЛГБД) по русским народным говорам с различением двух фонем «типа *о*». К созданной ранее базе данных по среднерусскому говору с. Пустоша Шатурского р-на Московской обл. добавлена ЛГБД по вологодскому слободскому говору, включающая ок. 30 тыс. словоформ, представляющих ок. 4500 лексем. Ядерный диалектный корпус (ЯДК) содержит тексты с частичной лексико-грамматической разметкой. В сентенциальной базе единицами являются предложения ЯДК в фонологической транскрипции, пронумерованные в порядке вхождения в ЯДК. На ее основе создан прямой алфавитный лексико-грамматический конкорданс и обратный алфавитный лексико-грамматический указатель словоформ. ЛГБД содержит информацию об условной фонологической транскрипции данной единицы, о словоизменительных и акцентных типах лексем, смысловые пометы о лексических значениях семантических диалектизмов, а также метаязыковые социолингвистические пометы о возрастных и территориальных особенностях употребления словоформы.

1. Предмет исследования и материал: русские народные говоры с различением двух фонем «типа *о*»

В рамках проекта РГНФ в 2006 г. было продолжено создание ЛГБД по русским народным говорам с различением двух фонем «типа *о*». К созданной ранее базе данных среднерусского (владимирско-поволжского) говора с. Пустоша Шатурского р-на Московской обл. добавилась построенная формате STARLING лексико-грамматическая база данных по севернорусскому слободскому говору деревень Арзубиха, Захариха и Злобиха Харовского р-на Вологодской области.

Предметом исследования являются русские говоры, в системе вокализма которых представлены две фонемы «типа *о*», распределенные в соответствии с правилом Л.Л. Васильева — А.А. Шахматова: «*о* закрытое» (фонема /уо/) выступает на месте **о* под праславянским «восходящим» ударением, «*о* открытое» (фонема /о/) — на месте **о* под «нисходящим» ударением, на месте **ѣ, *е, *ь*. В говорах с различением двух фонем «типа *о*» обычно также различаются две фонемы «типа *е*», наряду с фонемами /а/, /у/, /и/, в связи с чем их системы вокализма получили название семифонемных. В настоящее время такие говоры достаточно редки, не образуют сплошных ареалов, сохраняются главным образом в восточной части Европейской территории России и лишь отдельными вкраплениями — к югу и юго-западу от Москвы. Данные русских говоров с семифонемным вокализмом имеют особое значение для истории русского языка и славянской акцентологии, поскольку тембр ударного *о* < **о* является (по крайней мере, в части случаев) отражением праславянских слоговых тонов.

Некоторые косвенные данные свидетельствуют о том, что в прошлом системы вокализма рассматриваемого типа были распространены в русских говорах гораздо шире. Ареалы таких систем должны были быть не меньше современных ареалов нескольких типов диссимилятивного яканья, пред-

[1] Данная работа выполнена при финансовой поддержке Программы ОИФН РАН «Генезис и взаимодействие социальных, культурных и языковых общностей» (проекты «Восточнославянский диалектный корпус: праславянское наследие и лингвогеография» и «Генезис балто-славянской языковой общности: акцентологический аспект»), а также гранта РГНФ № 08-04-12132в.

полагающих семифонемный и шестифонемный ударный вокализм (обоянский, задонский, дмитриевский, новосёлковские, ореховские типы яканья). Карты и материалы Диалектологического атласа русского языка показывают, что небольшие кружевные ареалы юго-восточных семифонемных систем вокализма «вписаны» в несравненно более обширные ареалы перечисленных типов диссимилятивного яканья. Следовательно, эти типы яканья являются косвенным свидетельством наличия в прошлом различения под ударением двух фонем «типа *о*» на гораздо большей территории, чем сегодня. Локализация памятников письменности XIV–XVII вв., графико-орфографические системы которых отражают противопоставление двух фонем «типа *о»,* показывает, что говоры с различением двух *о* в старорусский период были распространены почти на всей территории русского языка.

В 2006 г. были проведены три диалектологические экспедиции для сбора материала по говорам изучаемого типа: в с. Новосёлки Рыбновского р-на Рязанской обл., с. Пустоша Шатурского р-на Московской обл. и в дд. Арзубиха, Захариха и Злобиха Харовского р-на Вологодской обл. Расшифровки магнитофонных записей речи уроженцев названных вологодских деревень стали материалом для ЛГБД.

Несколько говоров рассматриваемого типа были предварительно сопоставлены, в том числе с помощью построенных ЛГБД, как в отношении акцентных систем и распределения двух фонем «типа *о*», так и в отношении их лексического состава: западно-вологодские слободской и тотемский (последний — по описанию О. Брока); владимирско-поволжский говор с. Пустоша; восточный среднерусский акающий говор с. Лека Шатурского р-на Московской обл. (по описанию А.А. Шахматова); задонский говор (по материалам В. Тростянского); рязанский говора с. Новосёлки.

Говоры с противопоставлением двух *о* обнаруживают сильные различия по общим наборам признаков (они относятся к разным наречиям и группам говоров). Одновременно, относясь к восточной диалектной зоне, все говоры с двумя *о* имеют ряд важных общих черт: «моновариантное» склонение типа рус. лит., маргинальную подвижность ударения в прош. времени глаголов с корнями на нешумные, в целом схожее распределение непроизводных существительных по акцентным типам и ряд других сходств. Так, все семифонемные говоры обнаруживают нетривиальное сходство: сохранение у небольшого числа существительных муж. рода (**u-*, **i-* и консонантные основы а. п. *d*) рефлекса смешанной акцентной парадигмы (с формой-энклиноменом в И.ед. и окситонезой прочих форм). Обнаружены признаки, противопоставляющие друг другу отдельные группы говоров с различением двух фонем *о,* например, 1) /уо/ из **о* в формах мн. числа слов ж. и ср. рода а. п. *b* (*вдуóвы, вдуóвами; долуóты, долуóтами*) в средне- и южнорусских говорах рассматриваемого типа; в севернорусских в тех же случаях — /о/; 2) накоренное ударение в наст. времени и пов. наклонении *i*-глаголов а. п. *b* при насуффиксальном — в инфинитиве и прош. времени, характерное для говоров Рязанской группы и «рязанского ареала» говоров с различением двух *о* (Пустоша *хóжу, хуóдишь,* Новосёлки *хуóжу, хуóдишь*). Списки глаголов с указанной инновацией, однако, сильно различаются в говорах Пустошей и Новосёлок: если в Пустошах этот список ограничен итеративами а. п. *b1* (*ходить, носить, возить, водить, молотить, просить* и т. д.), то в Новосёлках в него входят каузативы и деноминативы а. п. *b2* и даже *c.* Последнее различие должно указывать на гетерогенный характер говоров с различением двух фонем «типа *о*» в «рязанском ареале».

Построение лексико-грамматической базы данных слободского говора (Харовский р-н Вологодской обл.). Аудиозаписи речи носителей говора старшего поколения были расшифрованы и записаны в аллофонемной транскрипции. На основе получившихся текстов с помощью интегрированной информационной среды STARLING (автор — чл.-корр. РАН С. А. Старостин) построена лексико-грамматическая база данных говора — так называемый ядерный диалектный корпус (ЯДК). ЯДК представляет собой исчерпывающее описание говора в рамках определенного корпуса текстов и охватывает тексты общей длиной около 30 тыс. речевых словоформ. Они репрезентируют 7047 языковых словоформ без учёта пунктуации, 9591 пунктуационно-грамматическую словоформу (пунктуационный вариант языковой словоформы).

2. Структура базы данных слободского говора

Структура базы данных слободского говора идентична структуре созданной ранее базы данных говора с. Пустоша Шатурского р-на Московской обл. В качестве исходной базы данных выступает ЯДК. Лингвистическая информация в ЯДК организована по многоступенчатому принципу. Выделяется 7 уровней членения письменного текста; на каждом из них выделяется своя основная (базовая) единица членения; каждой единице членения каждого уровня в ЯДК приписан уникальный номер, способный служить адресом отсылки к этой единице.

1. Уровень целого текста. На этом уровне вводятся параметры, характеризующие личность информанта: фамилия, имя, отчество, год и место рождения, образование и т.п.

2. Уровень абзаца (сверхфразового единства). Сверхфразовое единство — это отрезок текста, пун-

ктуационно выделенный особым абзацным делимитатором («красной строкой», «отступом»). У сверхфразового единства есть некоторая единая общая смысловая тема.

3. Уровень предложения (сентенциальный уровень). Границы предложений помечены сентенциальными делимитаторами. В начале предложения стоит инициальный делимитатор — суперсегментная пунктограмма «заглавности»; в конце предложения стоит финальный делимитатор — пунктограммы «.», «?», «!», «…». Содержательно предложение соответствует законченной мысли, а фонетически — интонационно законченному отрезку.

4. Уровень клаузы // предикации (клаузальный уровень). Границы клауз помечались так: предложение состоит из клауз, а между клаузами внутри предложения стоит один из клаузальных делимитаторов. К ним относятся пунктограммы «;», «:» и «–». Содержательно и интонационно клаузы примерно соответствуют простым предложениям и отдельным предикациям в составе сложных предложений.

5. Уровень синтагмы (синтагматический уровень). Границы синтагм внутри клаузы помечены пунктуационным синтагматическим делимитатором — пунктограммой «запятая». Содержательно и интонационно синтагмы примерно соответствуют словосочетаниям.

6. Уровень такта (тактовый уровень). Границы такта в ходе расшифровки помечались стандартной орфографической пунктограммой «пробел», но после создания ЯДК они были размечены вручную так: был использован пунктуационный тактовый делимитатор — пунктограмма «знаменательный (паузальный) пробел». Такты примерно соответствуют фонетическим словам, членам предложения, «синтаксическим молекулам», формам слова (как аналитическим, так и синтетическим). Важнейшее фонетическое свойство такта: внутри него невозможна (или по меньшей мере нетипична) пауза.

7. Уровень глосса (глоссовый уровень). Границы глоссов в ходе расшифровки помечались либо стандартной орфографической пунктограммой «пробел», либо стандартной орфографической пунктограммой «дефис», но после создания ЯДК их границы были размечены вручную. Каждый такт состоит из одного или нескольких глоссов. Глоссы, входящие в состав одного такта, обладают признаком потенциальной подвижности в предложении. Для обозначения границ глоссов при разметке был использован специальный набор нескольких метаязыковых глоссовых делимитаторов — пунктограммы «служебных пробелов». Выделены служебные пробелы шести типов: «{» между проклитикой и её правой опорой; «}» между энклитикой и её левой опорой; «<» между проклитикоидом и его правой опорой; «>{» между энклитикоидом и его левой опорой; «<>» между членами квази-композита с неустойчивым просодическим центром; «&» между компонентами «фразеологического штампа» с множеством просодических центров. Глоссы примерно соответствуют по длине морфологическим словам (в т. ч. служебным словам, синтетическим формам слов и подвижным компонентам аналитических форм). Внутри глосса (так же как внутри такта) невозможна пауза. Фактически наиболее близкий аналог глоссов в русском письменном тексте, записанном по правилам русской орфографии — это графические слова.

Ценность предложенной многоуровневой схемы ЯДК состоит в том, что при необходимости вывести на обозрение список отрезков текста, обладающих некоторым общим свойством, STARLING позволит пользователю по выбору вывести (на экран, на принтер или в файл) отрезок не только одного формата, но разных форматов — графическую словоформу (глосс), минимальный контекст этой словоформы (аналитическую форму, например, предложно-падежную, сочетание клитики с акцентно автономной словоформой и т. п. — такт), словосочетание (синтагму), предикацию (клаузу), предложение, абзац.

Лингвистическая информация о единицах текста на данном этапе в ЯДК такова: 1. Условная фонологическая транскрипция данной единицы (в сочетании с её пунктуационной разметкой). 2. Словоизменительный и акцентный тип данной единицы. 3. Смысловые пометы (при лексических диалектизмах). 4. Метаязыковые социолингвистические пометы о возрастных и территориальных особенностях употребления словоформы.

3. Приложение. Образцы словарных статей (иллюстрирующих семантическое поле «позвоночник» в харовском говоре)

Лён 1 ‘шейный отдел позвоночника’ <а.т. В>.

Вот <nom. sg.> *лён етот са́мой у ч’ело-вие́ка. Голова́ с позвоно́ч’ником свя́зана, между ни́м* <nom. sg.> *лён. А здие́с го́рло. А ше́йя ето фсё вми́сте ше́йа и йёс. А у шшу́ките ние́ту* <gen. sg.> *лну́-то е́тово. Како́й у шшу́ки* <nom. sg.> *лён. Ние́ту* <gen. sg.> *лну́ у шшу́ки.* <Егоров Виктор Никол. 1940 г. р. Род. в д. Злобиха Харовск. р-н, Волог. обл. 7 кл. Зап. Белова, Тер-Аванесова в д. Злобиха, Харовск. р-н, Волог. обл., 2003 г.>.

Осёл 1 ‘шестой шейный позвонок’ <а.т. А>.

<nom.-acc. sg.> *осёw,* <gen. sg.> *осёла* <Клешнина Нина Васил. 1936 г. р. Род. в д. Арзубиха Харовск. р-н, Волог. обл. 7 кл. Зап. Тер-Аванесова в д. Арзубиха, Харовск. р-н, Волог. обл., 2002 г.>.

<nom. sg.> *Осёw ето хря́шш о́коло хрепта́, о́коло позвоно́ч’ника, пе́рва-та шы́шка. Ешшо́ до позвоно́ч’ника не дошлẃ, и ше́йа конц’я́ец’ц’е —*

ето <nom. sg.> *осёw.* <nom. sg.> *Осёw болúт, горúт, как ц'úрей рвёт, как пересúлиш себя́, етот* <nom. sg.>*осёw.* <Егорова (Фокина) Зинаида Никол. 1941 г. р. Род. в д. Полутиха Харовск. р-н, Волог. обл. 8 кл. Зап. Белова, Тер-Аванесова в д. Злобиха, Харовск. р-н, Волог. обл., 2003 г.>.

Хребёт 'хребет, позвоночник без шейного отдела' <а.т. В>.

<nom.-acc. sg.>*хребёт,* <gen. sg.> *хрептá,* <instr. sg.>*хрептóм,* <loc. sg.>*на хрептué,* <nom. pl.> *хрепты́* <Клешнина Нина Васил. …>; <gen. sg.> *óколо хрептá* <Егорова (Фокина) Зинаида Никол. …>.

Хрип 1 'соединение позвоночника и черепа у рыб' <а.т. А sg.>.

Йа сломúw шшýки <acc. sg.> *хрúп, ет тóлко схрýпало. Фсё, онá уш готóва. Ф сиéтку попадёт, ну кáк йейó задавúт? гóлову рáс — фсё. В éтом миéсте у нейó слáбойо éто миéсто-то. А болшýйу кáк, никáк немóжно, мнué прóшлой гóт попáла, повезлó — йедвá* <acc. sg.> *хрúп сломúw. Шшýчина болшáйа, ак* <acc. sg.> *хрúп йедвá сломúл. Ниéту лнý у шшýки. Йёсли бы бы́w у шшýки лён, шшýки тáг бы* <acc. sg.> *хрúп не сломúт. Уш* <gen. sg.> *хрú па не сломúт.* <Егоров Виктор Никол. 1940 г. р. Род. в д. Злобиха Харовск. р-н, Волог. обл. 7 кл. Зап. Белова, Тер-Аванесова в д. Злобиха, Харовск. р-н, Волог. обл., 2003 г.>

4. Приложение. Полные синонимы в слободском говоре (на материале существительных)

При помощи ЛГБД легко показать, что говор деревень Арзубиха, Захариха и Злобиха, а также других деревень бывш. Слободского с/с Харовского р-на Вологодской обл., является единым. Наблюдаемые различия отчасти объясняются, по свидетельству информантов, как относящиеся к «младшей» или «старшей» разновидностям говора, причем соответствующие единицы «младшей» разновидности, как правило, заимствованы из литературного языка. Имеются и такие различия внутри говора, которые представляют собой внутрисистемные колебания. Крайне редки различия, которые могут претендовать на принадлежность к разным диалектным системам (например, название шеста, вокруг которого укладывают сено в стог: Арзубиха, Захариха *стожаир* или *стогаир,* Злобиха *островиина;* название помещения над избой, чердака: Арзубиха *иизбиця,* Митиха *потолwика*).

Ниже приводятся пары существительных — полных синонимов, выявленных в словаре слободского говора при помощи ЛГБД. Первый из пары синонимов является элементом традиционного лексического состава говора (иногда даже это — устаревшее слово); второй, как правило, представляет собой заимствование из литературного языка. Этот список выделен из словаря существительных, включающего около 2000 лексем; тем самым, примерно десятая часть словаря существительных говора представляет собой пары полных синонимов.

Пары полных синонимов (в большинстве своих случаев обязанных факту диалектно-литературного двуязычия) выделялись на основе явных показаний информантов («можно так сказать, а можно и так сказать»). Хронологические различия между членами пар («старое» / «новое») отмечались также на основании показаний информантов («сейчас говорят так-то, а раньше говорили так-то»).

бáба — жонá, бáба — жéншына, батýок — пáwка, берéмё — охáпка, блю́до — мúска, бог — икóна, божáт — хрёсной, божáтка — хрёсна мáт, брусни́ця — брусни́ка, брюшы́на, брю́хо — жывуóт, вар — пиéна, вéред — нарыw, ви́ця — виéтка, вуóлиха ~ óлха 'ольха', *вówненця — вownýшка, вóтен — лентя́й, грýда — кýчя* (предметов), *губá — подборуóдок, губá — чя́га, губá — погáнка, двойни́к — близнéць, долуóн — ток* (в гумне), *кáтаник — вáленок, колобáшка — лепёшка, колуóда — корыто, ком — ломóт* (хлеба), *косúця — висóк, кощýля — шýба* 'шуба, крытая сукном', *кут — кýхня, куфтыр — жывуóт, избá — куóмната, úзбиця — чердáк, йиéжа — йедá, зáгорода — úзгород, нéпогодь — бýря, лабазá — лесá* (строительные), *лáва — мóстик, лáwка — магазúн, лёжен — лентя́й, лён — шéя, лóшат — куóн, ля́га — топ, ледúна — болуóто, мётлá — вúник, миéw — заквáска* (из пивного сусла), *мизгúр, мызгыр — паýк, мост — пow, мостúна — половúця, наберёг — заквáска* (из остатков ржаного теста), *назём — навуóз, настáwниця* устар. *— учúтелниця, оболочúна — тýчя, оболочúнка — óблако, обýтка — óбуw, острежнúк — стрехá, отерёбок — замухрышка, отченáш — молúтва, плёнка — пузыр* 'околоплодный пузырь', *поскуóтина, поскуóтка — пáзбишшо, поганéць — погáнка, постéлкя — послиéд, потолуóка — чердáк, простокúша — простоквáша, рúзен — кусóк* (хлеба), *роднúк — колуóдець, рыло — нуóсик* (чайника), *слúзен — улúтка, собáка — пёс, стекля́шка, скля́нка — бáнка, бутыwка, солодя́ха — сыройиéшка, соромотá — стыд, сиéра — смолá, стáя — хлеw, тоскá — бол* 'боль', *úлик — úлей, фатá — платóк, хребёт — позвонуóчник* (позвоночник без шейного отдела, ср. *лён*), *хрестóк — хря́шшык, черёд — óчеред, чéрен — рýчка* (лопаты, вил), *чилúк — поберёзоватик, чиря́к — чúрей, шáм — мýсор, сóр, шыморá — прохиндéй, шубня́к — шýба* 'шуба мехом вверх'.

Литература

1. *Брок 1907* — О. Брок. Описание одного говора из юго-западной части Тотемского уезда // Сборник ОРЯС, 1907. — Т. 83.
2. *Васильев 1929* — Л. Л. Васильев. О значении каморы в некоторых древнерусских памятниках XVI–XVII вв. К вопросу о произношении звука о в великорусском наречии. Л., 1929.
3. *Зализняк 1985* — А. А. Зализняк. От праславянской акцентуации к русской. М., 1985.
4. *Крылов С. А.* Измерение частотности синтаксических молекул (на материале Генерального корпуса русского языка) // Кибрик А. Е. (ред.), Компьютерная лингвистика и интеллектуальные технологии. Вып. 7 (14). По материалам международной конференции «Диалог'2008» (Бекасово, 4–8 июня 2008 г.), М.: РГГУ, 2008а. С. 254–261.
5. *Крылов С. А.* О частотном словаре фонетических слов (на материале Генерального корпуса русского языка) // Архипов А. В. и др. (ред.). Фонетика и нефонетика. К 70-летию Сандро В. Кодзасова. М.: Языки славянских культур, 2008б, с. 387–399.
6. *Крылов С. А., Тер-Аванесова А. В.* Лексико-грамматические базы данных как инструмент диалектологического описания // Труды международной конференции «Диалог 2006». М., 2006. С. 493–497.
7. *Ter-Avanesova A.* Russian dialects with the distinction of two o-phonemes and their contribution to Slavonic accentology (Русские говоры с различением двух о-фонем и их значение для славянской акцентологии) // Second International Workshop on Balto-Slavic Accentology. Copenhagen, 2006. P. 20–24.
8. *Тер-Аванесова А. В.* Акцентуационные особенности русских говоров с различением двух фонем «типа о» // Тезисы докладов Международной конференции «Актуальные проблемы русской диалектологии» 23–25 октября 2006 г. М., 2006. С. 177–180.

Лексические функции и возможности оптимизации поиска информации в интернете (на материале параметрических слов)

Lexical functions and search engine optimization (based on words with numeric values).

Тимошенко С. П. (timoshenko@iitp.ru), **Цинман Л. Л.** (cinman@iitp.ru)

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

На базе лингвистического процессора ЭТАП-3 была разработана опция экспериментального перифразирования. Она дополняет двух- или трехсловный поисковый запрос о числовом значении параметра до неполного предложения. Эксперимент доказал, что показатель точности поиска повышается в среднем на 24 %.

1. Вводные замечания

Задача повышения точности поиска в Интернете не всегда хорошо решается чисто математическими методами. Для лингвиста это предсказуемо, поскольку в таких ситуациях мы имеем дело с асимметрией между означающим и означаемым. Применительно к поиску можно сказать, что означаемым является искомый смысл, а означающим — вся совокупность выражающих этот смысл предложений. Предложения языка L, выражающие один и тот же смысл, могут очень сильно отличаться друг от друга. Задача поисковой машины в таком случае — распознать смысл, игнорируя формальные различия. Соблазнительно попробовать решить эту задачу с помощью лингвистических инструментов, ориентированных на описание разнообразных способов выражения смысла — с помощью средств лексической семантики. Одним из таких инструментов является аппарат лексических функций, разработанный И. А. Мельчуком и А. К. Жолковским и значительно дополненный и усовершенствованный Ю. Д. Апресяном. См. [4–9; 1, 3].

2. Роль лексических функций в перифразировании

В самом общем виде можно сказать, что лексические функции — это тривиальные смыслы, словесное выражение которых в тексте зависит от того, при каком конкретном слове этот смысл выражается. Для некоторых фрагментов лексической системы языка разработанные лексической семантикой правила вида: «При слове X смысл f_1 выражается словом X′, при слове Y смысл f_1 выражается словом Y′» обладают большой предсказательной силой. Одним из таких фрагментов является класс параметрических слов. Под параметрическими словами мы понимаем имена существительные со значением параметра, допускающего числовое значение, например: *высота, вместимость, объем; продолжительность, возраст; мощность, сила, масса, давление, магнитуда; рождаемость, смертность; цена, стоимость, зарплата, выручка; энтропия, уровень, коэффициент, индекс* и т. д.

Правила лексической семантики, описывающие функционирование слов этого класса, в русском языке носят особенно строгий характер. Это связано с характерной чертой языка: хотя параметры являются универсальным типом предикатов, прототипическими представителями данного класса в русском языке являются не глаголы, а существительные ([2]:С. 74). Глаголов с соответствующими значениями, таких как *стоить, весить, длиться, вмещать*, очень мало. Даже для выражения такого тривиального параметра, как 'высота', специализированного глагола нет (в отличие, например, от английского: *The Pisa tower*

rises 56 meters).[1] Поэтому естественным способом приписывания какому-либо объекту определенного параметра является конструкция из параметрического существительного и глагола (*Пизанская башня* ***достигает в высоту*** *55 метров*).

Отношения между существительным и глаголом в огрубленом виде предстают как отношения названия ситуации и вспомогательного глагола, служащего для выражения категориальных значений — вида, времени и т.д. При более внимательном рассмотрении оказывается, что несвободные сочетания вида «глагол-связка + существительное» делятся на группы, объединенные общим элементом смысла, который и есть значение лексической функции.

Аппарат лексических функций лежит в основе опции перифразирования, реализованной в системе ЭТАП-3. Перифразирование позволяет, опираясь на записанные в словарных статьях слов значения лексических функций, построить некоторое количество предложений, синонимичных или квазисинонимичных заданному, то есть предложений, имеющих с ним одинаковую структуру на глубинном семантическом уровне. В системе ЭТАП для описания несвободных словосочетаний, вообще говоря, используется несколько десятков лексических функций. Они позволяют из предложения *Чистка матрицы стоит тысячу рублей.* получить следующий набор предложений (1):

Чистка матрицы имеет стоимость тысяча рублей.

Чистка матрицы достигает стоимости тысяча рублей.

Чистка матрицы имеет цену на тысячу рублей.

Стоимость чистки матрицы составляет тысячу рублей.

Стоимость чистки матрицы достигает тысячи рублей.

Стоимость чистки матрицы равняется тысяче рублей.

Цена чистки матрицы составляет тысячу рублей.

С помощью тех же лексических функций можно, подав на вход любое из получившихся предложений, снова получить весь набор перифраз.

Опция перифразирования была приспособлена для решения поисковых задач следующим образом: из именного словосочетания, в вершине которого находится параметрическое слово, можно получить ряд неполных предложений, содержащих все элементы, кроме численного значения параметра.

Ввод: *глубина Марианской впадины*

Перифразы, генерируемые системой перифразирования ЭТАП-3 в экспериментальном режиме (2):

Глубина Марианской впадины равна.

Глубина Марианской впадины составляет.

Глубина Марианской впадины достигает.

Глубина Марианской впадины равняется.

Марианская впадина имеет в глубину.

Марианская впадина достигает в глубину.

Марианская впадина имеет глубину.

Марианская впадина достигает глубины.

3. Коротко об алгоритмической организации экспериментального перифразирования

Коснемся лишь того фрагмента перифразирования, который относится к нашей задаче. Подробно перифразирование в системе ЭТАП описано в[3]. Приведенный выше куст перифраз (2) строится на основе информации о лексических функциях, содержащейся в словарной статье параметрического слова *глубина*, и некоторых универсальных правил перифразирования.

Для нашей задачи в слове *глубина* интерес представляют записи, относящиеся только к трем функциям:

OPER1:ИМЕТЬ/ДОСТИГАТЬ
FUNC2:СОСТАВЛЯТЬ1/ДОСТИГАТЬ/РАВНЯТЬСЯ1
LABOR1–2:ИМЕТЬ<В1>/ДОСТИГАТЬ<В1>

Из всего многообразия лексических функций, обслуживающих слово *глубина* эти три соединяют имя параметра с глаголом таким образом, чтобы получилось описание ситуации, когда что-либо имеет определенную глубину. Функция OPER1 позволяет обозначить ситуацию так, что параметр выступает

---

[1] Русские глаголы *выситься* и *возвышаться*, хотя и допускают конструкции со значением параметра (*Скала возвышается на 140 метров над уровнем Атлантического океана.*), в подавляющем числе случаев имеют не значение ***'X имеет высоту Y'***, а значение ***'X, имеющий большую высоту, располагается в Z'*** (или 'имея большую высоту'?)(Для глагола возвышаться это значение еще уточняется. ***'X, имеет большую высоту, и эта высота существенно больше высоты Z, рядом с которым располагается X'*** Типичные предложения с этими глаголами: *На ее могиле у Заонежского села Кузаранда высится стела. …над нами возвышается гигантский пресс.* (Примеры взяты из Национального корпуса русского языка и из СинТагРуса).

при соответствующем глаголе первым дополнением, например: *Гора Котопакси имеет **высоту** почти 6 км.* Функция FUNC2 позволяет обозначить ситуацию так, чтобы параметр был подлежащим при функциональном глаголе: ***Высота** горы Котопакси составляет почти 6 км.* Функция LABOR1–2 позволяет так обозначить ситуацию, что параметр занимает место второго дополнения при функциональном глаголе, а места подлежащего и сказуемого занимают первый и второй актанты соответственно: *Гора Котопакси достигает 5 870 м **в высоту**.*

Из универсальных правил перифразирования (их в системе ЭТАП несколько десятков) задействованы лишь три двусторонних правила

OPER1 + X <--> FUNC2 + X (*иметь глубину <--> глубина составляет*)

OPER1 + X <--> LABOR1–2 + X (*иметь глубину <--> иметь в глубину*)

FUNC2 + X <--> LABOR1–2 + X (*глубина составляет <--> иметь в глубину*)[2]

Если какая-либо лексическая функция имеет несколько значений (в нашем примере все три функции представлены альтернативными значениями), то система перифразирования строит предложения поочередно со всеми значениями. Один и тот же глагол может быть значением различных лексических функций (в нашем примере: *достигать высоты, высота достигает, достигать в высоту*). У различных параметрических слов значения этих трех лексических функций часто совпадают, но встречаются и заметные различия. Например, для слова *мощность* функция OPER1 имеет три значения

OPER1:ИМЕТЬ/ДОСТИГАТЬ/РАЗВИВАТЬ,

а функции LABOR1–2 для этого слова не существует.

Приведенный выше куст перифраз состоит из 8 предложений. Система перифразирования построит этот куст целиком, если ей на вход подадут либо именную группу (как в приведенном выше примере), либо любое предложение из этого куста.

Отметим, что для решения нашей задачи потребовалось установление отдельного режима работы системы ЭТАП. Дело в том, что все правила перифразирования в ЭТАПе настроены на работу с полными фразами, а фразы приведенного выше куста полными не являются. Поэтому при этом режиме работы системы после получения синтаксической структуры входного (неполного) предложения производится достройка этого предложения до полного. Например, если на входе была именная группа *глубина Марианской впадины*, то достраиваем это предложение до *глубина Марианской впадины равняется чему-то.* Глагол *равняется* — дежурное значение лексической функции FUNC2 для абсолютного большинства параметрических слов, а существительное *что-то* выполняет роль временного дополнения, необходимого при перефразировании. После синтеза очередной перифразы это временное существительное стирается. Если на вход ЭТАПа подается одно из предложений куста, где глагол присутствует, то синтаксическая структура пополняется только временным дополнением.

Упомянем еще одну проблему, которая возникает, когда на вход подается именная группа. Например, для запроса *водоизмещение 'Титаника'* наша система в настоящий момент построит запросы

водоизмещение 'Титаника' составляет,
'Титаник' имеет водоизмещение,
…,

в которых глаголы стоят в настоящем времени. Такие запросы не могут улучшить поиск, поскольку форма точного запроса не предполагает в качестве результата словоформы, отличающиеся от заданных наборами грамматических характеристик, а в текстах о 'Титанике' эти глаголы, скорее всего, употреблены в прошедшем времени. Подобная ситуация может возникнуть и с характеристикой вида глагола (сов/несов). Решение этой проблемы могло бы заключаться в порождении всех перифраз, где глаголы имеют различные характеристики времени и вида. Алгоритмически это сделать не сложно, но число перифраз при этом заметно возрастет.

В то же время упомянутая проблема исчезает, если на вход подается запрос, содержащий глагол в требуемой для данного запроса форме

водоизмещение 'Титаника' составляло
смертность в России к 2010 г. достигнет
…

В этом случае система перифразирования сохранит для всех перифраз характеристики глагола, поступившего на вход.

4. Эксперимент, определяющий эффективность поиска по перифразам

На основе исходного списка из 100 параметрических слов было составлено около 120 осмысленных коротких запросов (параметр + носитель пара-

[2] Особняком стоит правило OPER1 + X <--> X + *(быть) равным* (*иметь глубину <--> глубина равна*), поскольку, с одной стороны, это правило не универсальное (оно справедливо только для параметрических слов, у которых есть OPER1), а, с другой стороны, выражение *(быть) равным*, строго говоря, не является значением функции FUNC2.

метра). Именно к таким запросам, насколько можно судить по статистике Яндекса, чаще всего прибегают пользователи. С помощью системы ЭТАП-3 каждый запрос преобразовывался в блок перифраз. В ходе эксперимента перифразы вводились по одной в поисковые системы «Яндекс» и «Google». Поскольку нас интересовала принципиальная возможность улучшить поиск, автоматически расширяя запрос, время выполнения запроса и время обработки запроса системой ЭТАП-3 не учитывалось, равно как и дата проведения эксперимента и загрузка поисковых серверов. Поскольку система перифразирования выдает целостные структуры, не выходящие за пределы одного предложения и не предусматривающие разрывов и пропусков, для тестирования была выбрана форма точного запроса. Применение логического оператора «ИЛИ» внутри точного запроса языками поисковых запросов не поддерживается. Употребление дизъюнкции при неточном запросе приводит к тому, что находятся не сайты с целостной структурой, а сайты, просто содержащие заданные слова, возможно, не связанные синтаксически. Поэтому группа слов, выражающая численное значение и зависящая от глагола, может относиться к своершенно другому объекту или к другому параметру. Частично снимает эту проблему опция поиска с пропуском заданного количества слов. Можно себе представить неточный запрос вида *водоизмещение Титаника /+1 составлять*. Однако и он не дает стопроцентной точности. На первую же страницу почему-то попадает такой фрагмент:

Через два часа тридцать пять минут после катастрофы крен «**Титаника**» **составлял** почти 90 градусов. … Длина парома "Геральд оф Фри Энтерпрай» **составляла** 132 метра, **водоизмещение** — 7951 тонна. Он являлся составной частью флота, управляемого компанией…

Очевидно, эти опции пока еще не всегда корректно обрабатываются поисковыми машинами. Если усложнить запрос, насытив его «лингвистическими» маркерами (например «Гора Котопакси» /+1 достигает /+3 «в высоту»), доля неподходящих ответов еще возрастет.

Употребление точной формы запроса привело к тому, что различия в работе двух упомянутых поисковых систем практически нивелировались, за исключением различий в составе баз документов (например, «Яндекс» индексирует больше личных блогов, чем Google). Эти различия оказались настолько незначительными, что мы сочли возможным ими пренебречь.

Цель эксперимента заключалась в том, чтобы определить, насколько использование системы экспериментального перифразирования повышает эффективноость поиска. Для этого был разработан следующий протокол оценки эффективности. Релевантным признается такой результат, когда искомая численная информация содержится непосредственно в сниппете, предлагаемом поисковой машиной. Показателем эффективности поиска считается количество релевантных результатов на первой странице (т. е. количество релевантных результатов среди первых десяти), выраженное в процентах. Если количество найденных документов меньше 10, то оно и принимается за 100%. В качестве контрольного уровня эффективности поиска был принят уровень эффективности поиска по исходным запросам — именным группам, подаваемым на вход системы перифразирования ЭТАПа-3. Эти запросы также оформлялись как точные — это ограничение дает возможность оценить чистый эффект использования перифраз.

В результате этих операций получаем данные по каждому запросу:

Запрос	**Найденные страницы**	**Найденные сайты**	**Кол-во релевантных документов на 1 странице**
«твердость алмаза»	2044	791	3/10 (30%)
«твердость алмаза равна»	42	18	10/10 (100%)
«твердость алмаза составляет»	18	6	3/6 (50%)
«твердость алмаза достигает»	0	0	0
«твердость алмаза равняется»	0	0	0
«алмаз имеет твердость»	79	35	9/10 (90%)
«алмаз достигает твердости»	0	0	0

Первая строка в данных по каждому запросу будет представлять эффективность поиска по необработанным словосочетаниям, а все остальные — по обработанным перифразам. Чтобы оценить среднее изменение результата, можно взять среднее значение эффективности необработанных запросов и среднее значение по всем перифразам вместе. Однако такой подход кажется нам неверным. Представим себе, как могла бы работать поисковая машина, умеющая генерировать перифразы. Она получает на входе запрос, создает перифразы, а затем обрабатывает каждую из них как точный запрос. Какие-то перифразы приносят результат, какие-то — нет. Очевидно, что на экран в таком случае выводится сумма найденных по всем перифразам документов, а несработавшие перифразы, конечно, ничего не добавляют, но и не ухуждшают картины. Поэтому мы рассчитывали повышение эффективности поиска для каждого запроса отдельно, и уже для этих результатов затем рассчитывался средний показатель, характеризующий общее изменение эффективно-

сти. Так, для приведенного выше запроса средний показатель увеличения эффективности составит

$$((100\% - 30\%) + (50\% - 30\%) + (90\% - 30\%)) / 3 = \\ = 50\% \text{ или } 0{,}5.$$

Бывают случаи, когда необработанный запрос приносит какое-то количество релевантных ответов, а перифразы не приносят ничего. Например, *«абсолютный минимум температуры на Земле»*. Это связано с тем, что информация по этому вопросу очень часто представлена в Рунете перепечаткой статьи из Большой советской энциклопедии, где использована конструкция с нулевой связкой вместо глагола. В этом и других похожих случаях мы признавали результат использования системы перифразирования отрицательным: если результаты поиска по необработанному запросу были релевантны в 10 случаях из 10, то эффективность поиска с перифразированием равна −100% или −1, если в 6 случаях из 10 — −60% или −0,6 и т.д.

Распределение показателей изменения эффективности поиска при использовании системы перифразирования представлено на гистограмме (рис. 1).

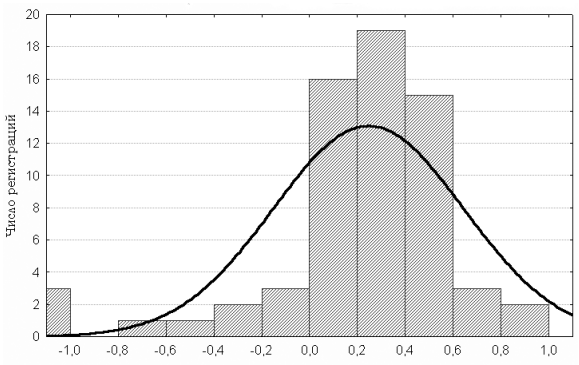


Рис. 1. Изменение эффективности поиска при использовании запросов на основе перифраз аппроксимация: 65 × 0,2 × normal (x; 0,2395; 0,3966)

На графике видно, что в среднем точность поиска повышается на 24%. Любопытно, что если разделить параметрические слова по тематическим группам, то этот средний показатель будет варьироваться. Если смотреть данные по «географическим» запросам (параметры географических объектов, как природных, так и культурных, например, *ширина реки Амазонки в среднем течении* или *численность населения Киева*), то повышение эффективности поиска составит в среднем 18,5%, а в группе данных по физическими запросам (примеры запросов: *молярная масса меди, сила тяжести на Марсе*) этот показатель будет равен 27%. Результаты поиска по геометрическим объектам (примеры запросов: *площадь поверхности сферы, объем конуса*) оказались неожиданными. Реализующий значение лексической функции FUNC2 глагол *равняться,* а также добавленная в поисковое перифразирование конструкция с кратким прилагательным *равен,* отличаются от других средств тем, что регулярно присоединяют в качестве первого дополнения формулу или её словесное описание: *Объём конуса равен одной трети произведения основания на высоту; Объем конуса равняется одной трети объема цилиндра с теми же основанием и высотой.* Показатель эффективности использования перифраз для поиска информации такого рода составляет 13%. Однако ЭТАП-3 можно настроить на поиск подобной информации, добавив в систему перифразирования глаголы *характеризовать*, *измеряться* и, возможно, некоторые другие. Отдельно подчеркнем, что эти глаголы не являются значениями лексических функций параметрических слов. Точно так же не описываются лексическими функциями конструкции типа *площадь восьмиугольника может быть вычислена как…*. Информацию о том, что формулы могут вводиться подобными предложениями мы черпаем из экстралингвистических фактов. Соответственно, в рамках системы ЭТАП можно построить ряд правил, позволяющих получать подобные предложения, однако эти правила, скорее всего, будут крайне неполными: при ничем не ограниченной сочетаемости слов практически невозможно описать все возможные предложения и тем боле угадать предложения, реально присутствующие в интернет-документах.

5. Переводы запросов на английский язык

Поскольку одной из опций процессора ЭТАП является автоматический перевод текста, то для системы не составляет труда перевести все полученные путем перифразирования словосочетания на английский. Из запроса *высота Пизанской башни* с помощью ЭТАПа легко получить набор словосочетаний (3):

The height of the Pisa tower equals

The height of the Pisa tower reaches

The height of the Pisa tower is reaching

The height of the Pisa tower amounts to

The height of the Pisa tower attains

The height of the Pisa tower is attaining

Однако эффективность поиска в данном случае практически не повышается. И это вновь связано с особенностями строя языка, в частности, с тем хорошо известным фактом, что в английском язы-

ке параметры задаются прилагательными со значением верхнего полюса шкалы (30 feet high, 6 feet tall, 25 years old), что русскому языку практически не свойственно (за исключением известных редких примеров типа *как велика вероятность n*). В случае параметров для носителей языка естественнее употреблять прилагательные: *The Cupola is 55 meters high and 16 meters wide.* Поэтому перифразы, построенные ЭТАПом с помощью существующих на сегодняшний день правил, несмотря на то, что грамматически они абсолютно правильны, могут вообще не встречаться среди источников. Именно так дело обстоит с Пизанской башней.

Кроме того, играет роль общеизвестный факт обязательности глагола-связки в английском языке. В тех случаях, когда параметр все-таки обозначается существительным, носителям английского языка не нужно подбирать глагол, который мог бы выразить грамматические категории — достаточно просто использовать глагол *be: The speed of light is 300 million metres per second*. На данном уровне экспериментальное перифразирование не может быть использовано для перевода запросов. Однако именно в силу того, что в английском разница между формой запроса и формой ответа столь заметна, это поле деятельности представляется довольно интригующим. Чтобы двигаться в этом направлении, необходимо расширять как правила перифразирования, так и инструменты, позволяющие переводить глубинные английские структуры предложений типа *The Cupola is 55 meters high в глубинные русские структуры предложений типа Высота купола составляет 55 метров. Это может стать шагом к построению полноценного глубинно-синтаксического представления в той действующей модели языка, которая лежит в основе лингвистического процессора ЭТАП-3.*

6. Неточные запросы

Данные эксперимента показывают, что использование перифраз повышает точность поиска за счет выбора только тех документов, которые содержат искомую информацию, наличие которой однозначно предсказывается глаголом, реализующим ту или иную лексическую функцию. Однако требование точного запроса приводит к тому, что отсеиваются и релевантные документы, в которых мысль выражена немного по-другому: не будет найден документ, содержащий предложение *Гора Котопакси, высота которой составляет почти 6 км…* Неточный поиск по перифразе приводит к тому, что конструкция разрывается и по запросу *продолжительность жизни в Голландии составляет,* находятся страницы новостей, где есть информация о продолжительности жизни в Китае и сообщение о политической ситуации в Голландии. Однако некоторые попытки такого поиска показывают, что и в этом случае точность поиска по перифразам выше точности поиска без них. Это происходит из-за того, что наличие глагола, пусть даже и оторванного от именного словосочетания, меняет общую направленность страницы. Например, по запросу *высота юкки* находятся страницы, содержащие объявления о продаже пальмы какой-либо высоты, а по запросу *высота юкки достигает* находятся статьи из разнообразных справочников по цветоводству, содержащих общую информацию о растении. Можно сказать, что в действие вступает стилистический фактор, так как глаголы, реализующие лексические функции OPER1, FUNC2, LABOR1–2 часто используются в научно-публицистических и научных текстах. Это открывает для системы перифразирования определенные перспективы, тем более что в системе ЭТАП-3 предусматривалось применение стилистических фильтров.

7. Выводы

Эксперимент показал, что точность результатов поиска по запросам, предполагающим численные ответы, может быть увеличена за счет использования системы экспериментального перифразирования. В случае точного запроса количество релевантных результатов увеличивается на 24%. Точный запрос означает, что к искомым документам предъявляются самые жесткие требования. В случае менее жестких требований, то есть при использовании нестрогого запроса, результаты непредсказуемы. Помимо варьирования форм запросов, зависящего от поисковой машины, пути улучшения работы экспериментального перифразирования лежат в расширении числа используемых лексических функций и более точной их настройки на разные типы информации. В частности, эксперимент показал, что, изменяя участвующие в перифразировании глаголы, можно получать либо конкретные числовые значения, либо формулы и их словесные описания.

Литература

1. *Апресян Ю. Д.* Избранные труды. Лексическая семантика. Синонимические средства языка. // М.: 1995. Т. 1, 2-е издание.
2. *Апресян Ю. Д.* Основания системной лексикографии // Языковая картина мира и системная лексикография. М.: Школа «Языки русской культуры», 2006.
3. *Апресян Ю. Д., Цинман Л. Л.* Формальная модель перифразирования предложений для систем переработки текстов на естественных языках // Русский язык в научном освещении. 2002. № 4. С. 102–146.
4. *Жолковский А. К., Мельчук И. А.* О возможном методе и инструментах семантического синтеза // Научно-техническая информация. 1965. № 6.
5. *Жолковский А. К., Мельчук И. А.* О семантическом синтезе // Проблемы кибернетики. — 1967. № 19. С. 177–238.
6. *Жолковский А. К., Мельчук И. А.* К построению действующей модели языка «Смысл <=> Текст» // Машинный перевод и прикладная лингвистика. 1969. №11. С. 5–35.
7. *Мельчук И. А.* Опыт лингвистических моделей «Смысл <=> Текст». Семантика, синтаксис // М.: Школа «Языки русской культуры», 1999.
8. *Мельчук И. А., Жолковский А. К.* Толково-комбинаторный словарь современного русского языка. Опыт семантико-синтаксического описания русской лексики // Wien: Wiener Slawistischer Almanach, 1984.
9. *Mel'čuk I.* Lexical Functions: A Tool for the Discription of Lexical Relations in a Lexicon // Lexical Functions in Lexicography and Natural Language Processing. Ed. By L. Wanner. Amsterdam (Philadelphia). 1996. P. 37–102.

Применение методов лингвистической семантики и машинного обучения для повышения точности и полноты поиска в поисковой машине «Exactus»

Applying linguistic semantics and machine learning methods to search precision improvement in search engine «Exactus»

Тихомиров И. А. (matandra@isa.ru), **Смирнов И. В.** (ivs@isa.ru)

Институт системного анализа РАН

Доклад посвящен проблемам применения методов лингвистической семантики и машинного обучения в задачах поиска и анализа текстов. Приведена экспериментальная оценка алгоритма поисковой машины Exactus в рамках семинара РОМИП. Сделаны выводы о перспективах использования логических методов машинного обучения в задачах анализа текстов.

1. Введение

В 2008 году поисковый алгоритм Exactus претерпел значительные изменения по сравнению с 2007 годом. Основной новинкой явилось включение в алгоритм анализа текстов контекстных правил установления значений минимальных синтактико-семантических единиц текста (синтаксем) [1]. Эти правила были получены с помощью методов машинного обучения на основе электронной версии синтаксического словаря Золотовой Г. А.[2]. Это позволило улучшить качество семантического анализа и снизить шум при поиске, что подтверждено результатами экспериментов в рамках российского семинара по оценке методов информационного поиска (РОМИП) [3]. О примененных методах машинного обучения и лингвистической семантики пойдет речь в данной работе.

2. Особенности поискового алгоритма Exactus

В различных публикациях неоднократно отмечалось, что поисковый алгоритм Exactus объединяет статистические и лингвистические методы поиска [4]. Лингвистическая составляющая поискового алгоритма заключается в учете семантики — смысловых значений слов, которые определяются на основании теории коммуникативной грамматики русского языка [5], [6].

Семантический анализ текста имеет своей целью извлечение смысла из текста и отображение его в формальную модель, которая позволяет находить смысловую близость двух текстов (применительно к задаче поиска — близость запроса и документа). При компьютерном семантическом анализе текста множество синтаксем каждого предложения отображается в неоднородную семантическую сеть, предложенную Г. С. Осиповым [7], с синтаксемами в вершинах и семантическими связями на множестве синтаксем в качестве ребер [8].

Семантический анализ текста оперирует в основном именными синтаксемами, которые выделяются в результате морфологического и синтаксического анализа. Именная синтаксема представляется в тексте именной или предложной группой — словосочетанием с существительным или предлогом в качестве управляющего слова. Именная синтаксема характеризуется морфологической формой — предлогом, падежом, и категориально-семантическим классом существительного, от которого она образована. Морфологическая форма синтаксемы и категориально-семантический класс определяются с помощью лингвистического анализатора текста. Синтаксема характеризуется также синтаксической функцией, которую она может выполнять в предложении, и синтаксическим значением. В ходе семантического анализа текста необ-

ходимо установить **значения** именных синтаксем, которые являются обозначениями смыслов, передаваемых текстом.

Морфологическая форма и категориально-семантический класс именной синтаксемы не однозначно задают её значение, а синтаксическую функцию, в которой выступает конкретная синтаксема, встречаемая в тексте в ходе анализа, автоматически определить невозможно. Таким образом существует проблема семантической многозначности синтаксем. Обычно для разрешения этой многозначности в анализ вовлекается контекст — глагол или отглагольное существительное, т.е. предикатное слово, при котором именная синтаксема встречается в предложении. Учет такого рода контекста требует создания специального словаря, описывающего наиболее частые сочетания определенного глагола с возможными синтаксемами при нем, и такой словарь был создан для глаголов и отглагольных существительных, наиболее часто встречаемых в текстах определенной тематики.

Словарь предикатных слов не может охватить все глаголы и отглагольные существительные, т.к. перечисление возможных синтаксем при глаголе вручную является весьма трудоёмкой задачей для лингвистов. Поэтому часто при семантическом анализе невозможно опираться на предикатное слово, так как его нет в словаре предикатных слов, следовательно, для точного установления значения синтаксемы в таких случаях необходимо учитывать другой контекст синтаксемы.

В безглагольных предложениях или предложениях, для которых предикатное слово не найдено в словаре, синтаксемы присутствуют рядом с другими элементами предложения, и несут своё значение только в данном контексте. Зависимость значения синтаксемы от собственных морфологических характеристик и характеристик соседних элементов предложения (не глаголов) является языковой закономерностью, которую необходимо обнаружить и зафиксировать для выполнения семантического анализа безглагольных предложений в дальнейшем. Такую закономерность для значения синтаксемы можно записать в виде **правила**, где в посылке правила находятся характеристики самой синтаксемы и контекста — окружающих её синтаксем и других элементов предложения, а заключение правила содержит значение, которое необходимо приписать целевой, рассматриваемой синтаксеме. Контекстные правила позволяют однозначно устанавливать значения синтаксем, т. е. снимать семантическую многозначность синтаксем.

Построение правил установления значений синтаксем экспертом-лингвистом требует больших трудозатрат на просмотр текстов, где встречаются анализируемые синтаксемы, анализ контекста синтаксем, обобщение признаков, влияющих на значение синтаксемы в разных текстах. Поэтому встала задача автоматического построения таких правил с помощью методов машинного обучения.

ДСМ-метод порождения гипотез, предложенный В. К. Финном [9], применяется для выявления скрытых причинно-следственных закономерностей в некоторой предметной области. Его задачей является обнаружение причин возникновения некоторого явления, или наличия свойств у объектов из некоторого множества. Решение этой задачи основывается на фактах или обучающем множестве объектов. Найденные причины используются для прогнозирования наблюдения явлений в дальнейшем.

Индуктивный вывод в ДСМ-методе основывается на принципе единственного сходства, сформулированном Д. С. Миллем: е*сли какое-то обстоятельство постоянно предшествует наступлению исследуемого явления, в то время как иные обстоятельства изменяются, то это обстоятельство есть, вероятно, причина данного явления*.

На практике построение гипотез о причинах того, что некоторые объекты обладают определенным свойством, заключается в нахождении характеристики сходства этих объектов — максимального множества признаков, которое принадлежит двум или более объектам с данным свойством. Эта характеристика сходства будет тем самым обстоятельством, которое не меняется от случая к случаю при наблюдении явления.

Для применения ДСМ-метода к решению задачи автоматического получения правил установления значений синтаксем было сделано следующее:

- введено отношение выводимости на морфологических признаках синтаксем (предлог, падеж, категориальный класс) и задана операция вычисления сходства морфологических признаков синтаксем, что позволяет вычислять характеристики сходства синтаксем;
- введено понятие составного морфологического признака — совокупности других морфологических признаков, рассматриваемых как один целый признак, задана операция вычисления сходства составных морфологических признаков;
- введено понятие синтаксем в контекстах, задана операция вычисления сходства синтаксем в контекстах.

Всё это позволило оперировать сложными лингвистическими объектами «синтаксема» или «синтаксема в контексте» без нарушения их внутренней логической структуры.

На описанных выше принципах была разработана программная реализация метода порождения правил установления значений синтаксем.

Материалом для построения обучающих примеров (синтакскем в контекстах) послужила электронная версия синтаксического словаря Г. А. Золотовой [2], предоставленная сотрудниками Машинного фонда русского языка Института русского языка РАН. В словаре приводятся синтаксемы с при-

мерами их вхождений в тексты литературы и периодики. В электронной версии словаря границы синтаксем выделены с помощью знаков подчеркивания «_». Это дает возможность автоматически выделить фрагменты текста, содержащие примеры синтаксем, и построить синтаксемы в контекстах. Следует заметить, что в синтаксическом словаре для каждой синтаксемы приводится очень мало примеров встречаемости в текстах, что делает неэффективным применение статистических методов машинного обучения. Преимущество логических методов машинного обучения состоит в том, что они срабатывают на обучающих выборках малого объёма (например, для индуктивного вывода в ДСМ-методе достаточно двух обучающих примеров).

В результате выполнения программной реализации было порождено более тысячи правил установления значений синтаксем. Для каждого правила сохранялись примеры, из которых оно было получено. Каждый пример содержит тексты, из которых были созданы целевая и соседняя синтаксемы, а также обрабатываемое предложение целиком. Таким образом, каждое правило хранит своё обоснование, которое может быть полезным как для оценивания адекватности реализации метода, так и при анализе лингвистом правильности установления значений синтаксем в дальнейшем.

Приведем пример правила установления значения «дестинатив» (назначение предмета или действия) для синтаксемы родительного падежа с предлогом «для»:

Правило: *Если встречается синтаксема в падеже <родительный> с предлогом <для>, имеющая категориальный класс <личное>, а до неё встречается синтаксема в падеже <именительный>, имеющая категориальный класс <предметное>, то полагается, что первая синтаксема имеет значение <дестинатив — назначение предмета или действия >*

Обоснование:

(40) ЗНАЧЕНИЕ = дестинатив
ЦЕЛЕВАЯ СИНТАКСЕМА = для тебя; КСК: личное
СОСЕДНЯЯ СИНТАКСЕМА = Все; ПРЕДЛОГ: ;ПАДЕЖ: им.вин.; КСК: предметное; ПОЗИЦИЯ: до
===КОНТЕКСТ: и песни, и силы — Все для тебя.

(41) ЗНАЧЕНИЕ = дестинатив
ЦЕЛЕВАЯ СИНТАКСЕМА = для различных рачков; КСК: личное
СОСЕДНЯЯ СИНТАКСЕМА = пища; ПРЕДЛОГ: ;ПАДЕЖ: им.; КСК: предметное; ПОЗИЦИЯ: до
===КОНТЕКСТ: Эти растения — пища для различных рачков

В примерах поле «КОНТЕКСТ» содержит предложение, из которого был построен пример.

Предложен алгоритм снятия смысловой многозначности синтаксем на основе полученных контекстных правил, который позволяет выбрать одно значение для синтаксемы из всех возможных, что уменьшает число ошибок семантического анализа текста в среднем в 4 раза (в случае срабатывания правила).

Реализованный алгоритм снятия смысловой многозначности синтаксем на основе полученных правил был внедрён в поисковый алгоритм Exactus, что позволило повысить точность семантического анализа и, соответственно, поиска документов.

3. Результаты экспериментов

Экспериментальная оценка поискового алгоритма Exactus, включающего методы лингвистической семантики и машинного обучения, проводилась в рамках российского семинара по оценке методов информационного поиска в 2008 году [10].

Общепринятым критерием оценки работы поисковых алгоритмов является 11-точечный график TREC, который отображает совмещенные показатели точности и полноты при разных показателях точности. На рисунках 1 и 2 ниже приведены 11-точечные графики TREC оценок AND и OR для системы Exactus и других участников семинара для коллекции «Белорусский WEB».

Заметим, что около 37% запросов, оцениваемых асессорами РОМИП, содержат синтаксемы с ролями, и только 16% из этих запросов содержат предикатные слова, т. е. в остальных запросах для синтаксем происходило снятие многозначности. Например, для двух оцениваемых запросов: «работа для студентов», «магазины для беременных» выполняется приведенное выше правило.

Разработчиками Exactus были сданы два прогона, которые отличались друг от друга настроечными параметрами поискового алгоритма. По графикам видно, что экспериментальный алгоритм Exactus получил ощутимо лучшие оценки по всем точкам TREC-графика для OR-оценки и большинству точек TREC-графика для AND-оценки.

4. Заключение

Применение методов интеллектуального анализа данных к обработке текстов на естественном языке может быть полезно для решения многих задач компьютерной лингвистики, но накладывает определённые требования на используемые методы

и получаемые результаты. Такие требования состоят, например, в способности оперировать сложными лингвистическими объектами и в интерпретируемости результатов. Систематическое применение в прикладной лингвистике методов машинного обучения, реализующих индукцию и аналогию, подтверждает возможность автоматического получения корректных результатов, которые наглядно объясняют некоторые языковые закономерности, а также помогают решать прикладные задачи, в частности снимать семантическую многозначность.

Массовость эксперимента, проведенного в рамках РОМИП, не позволяет оценить вклад отдельно каждого из используемых подходов к поиску. Анализ результатов участия в РОМИП показывает, что нельзя выделить какой-либо один фактор, существенно влияющий на показанное преимущество поискового алгоритма Exactus по сравнению с аналогами. Хороших результатов позволила достичь совокупность методов машинного обучения, лингвистической семантики, статистики, а также хороший уровень технического обеспечения и программирования.

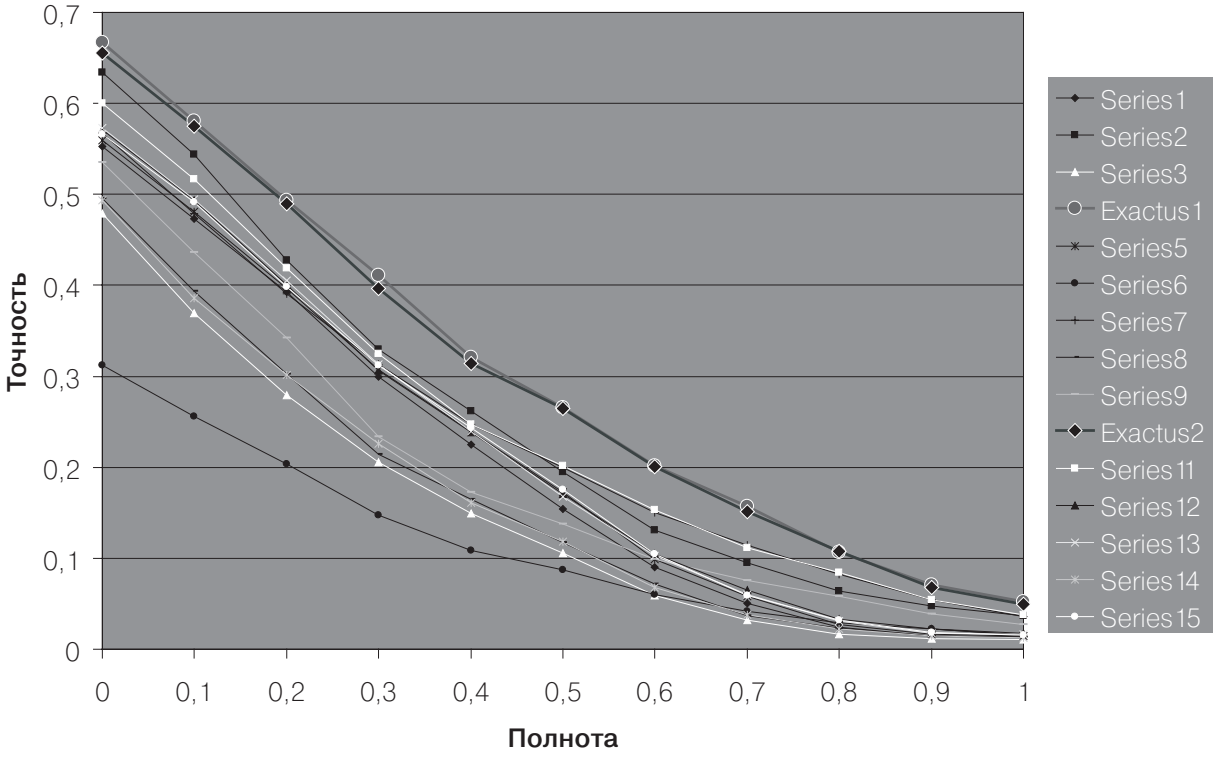


Рис. 1. Белорусский WEB, график TREC: OR-оценка.

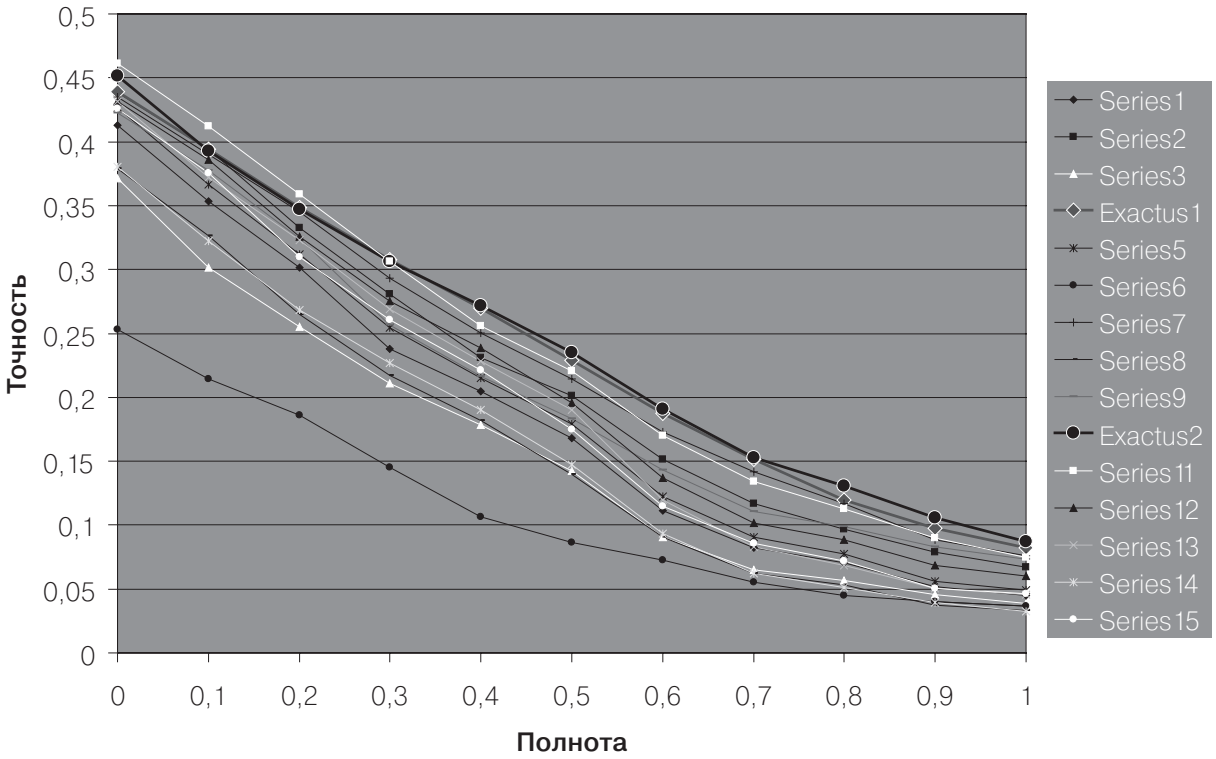


Рис. 2. Белорусский WEB, график TREC: AND-оценка.

Литература

1. *Смирнов И. В.* Метод автоматического установления значений минимальных синтаксических единиц текста // Информационные технологии и вычислительные системы. — 2008. — №3. — С. 30–45.
2. *Золотова Г. А.* Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса // М.: Эдиториал УРСС, 2001.
3. *Смирнов И. В., Соченков И. В., Муравьев В. В., Тихомиров И. А.* Результаты и перспективы поискового алгоритма Exactus. // Труды российского семинара по оценке методов информационного поиска РОМИП'2007–2008. Санкт-Петербург: НУ ЦСИ. — 2008. — С. 66–76.
4. *Золотова Г. А., Онипенко Н. К., Сидорова М. Ю.* Коммуникативная грамматика русского языка // М.: Институт русского языка РАН им. В. В. Виноградова, 2004.
5. *Тихомиров И. А., Смирнов И. В.* Интеграция лингвистических и статистических методов поиска в поисковой машине Exactus. // Труды мезденародной конференции Диалог'2008. — С. 485–491.
6. *Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Olga Zavjalova.* Application of Linguistic Knowledge to Search Precision Improvement.//Proceedings of 4th International IEEE conference on Intelligent Systems 2008. Volume 2. — P. 17-2–17-5.
7. *Осипов Г. С.* Приобретение знаний интеллектуальными системами // М.: Наука. Физматлит, 1997.
8. *Осипов Г. С., Смирнов И. В., Тихомиров И. А.* Реляционно-ситуационный метод поиска и анализа текстов и его приложения. // Искусственный интеллект и принятие решений. — 2008 — №2 — С. 3–10.
9. *Финн В. К.* ДСМ-метод как средство анализа каузальных зависимостей в интеллектуальных системах. // Научно-техническая информация, Сер. 2, Информ. процессы и системы. — №11. — 2000 — С. 1–5.
10. *Russian* Information Retrieval Evaluation Seminar // http://romip.ru

К проблеме вариативности видовых форм императива

On the problem of variability of imperative aspectual forms

Труб В. М. (trub44@ukr.net)

Институт украинского языка НАН Украины, Киев, Украина

Статья посвящена рассмотрению соотношения разных видовых форм императивных глаголов и исходит из предпосылки, что семантическая интерпретация побуждений, оформленных разными видовыми императивными формами, должна фиксировать смысловые различия между ними или же указывать причины аномальности, возникающей при употреблении одной из противоположных видовых форм

В данной работе рассматривается вопрос о соотношении видовых форм императивных глаголов. По этой проблеме существует обширная литература, одно из важных обобщений которой содержится, в частности, в работах [5, 6]. Как отмечает Розанна Бенаккьо, «…присущая императиву темпоральность проявляется настолько ограниченно, что традиционные видовые значения…, хотя, несомненно, и существуют, не являются столь прозрачными и часто приобретают совершенно иное значение…» [6]. Задача семантического анализа разных видовых форм императива заключается в объяснении смысловых различий между ними или же в объяснении причины аномальности употребления той или иной видовой формы.

Здесь нам хотелось бы коснуться некоторых непростых аспектов этой обширной и сложной проблематики, в частности — обратить внимание на связь между видовой формой глагольного императива и спецификой его аспектуальных значений, таких как комплетивность / инкомплетивность — ср. [2] (в частности деятельность / достижение результата (цели) действия — ср. соответственно соотношение *вытекать — вытечь* и, с другой стороны, *откручивать — открутить*), контролируемость / неконтролируемость, моментальность / процессуальность, однократность / неоднократность и т. д.

Видовое противопоставление НСВ и СВ, как хорошо известно (ср. [3]), далеко не всегда связано с противопоставлением типа «процесс — достижение предела процесса». Одним из многочисленных примеров «регулярного» видового соотношения типа «деятельность — достижение результата (цели) действия» служит однокорневая видовая пара *разыскивать — разыскать*, маркирующая соотношение типа ‘поиск референтного объекта’ / ‘его обнаружение’. Между тем результативным коррелятом для семантически очень близкого *разыскивать* глагола *искать* является СВ *найти*. В то же время противоположной видовой парой для *искать* будет его словообразовательный дериват — СВ-коррелят с делимитативным значением *поискать* (ср. также [7]), а для глагола *найти* — НСВ *находить*.

Употребление императивных форм нормативно выполняет функцию побуждения к достижению результата, который говорящий считает посильным для слушающего — адресата побуждения. При этом степень проблематичности достижения нужного результата может отражаться в варьировании видовых (инкомплетивных / комплетивных) форм одного и того же или разных глаголов:

(1) *Ищи слесаря* ≈ ‘действуй (приложи усилия) с целью обнаружения слесаря, чтобы привлечь его к решению текущей проблемы’.

(2) *Поищи слесаря* ≈ ‘попробуй найти слесаря, чтобы…’ ≈ ‘осуществи квант деятельности с целью обнаружения слесаря, чтобы…’

(3) *Найди слесаря* ≈ ‘сделай так, чтобы обнаружить слесаря (обнаружь слесаря), чтобы…’

В (1), (2) и (3) речь идёт о побуждении слушающего к достижению определённой цели — обнаружения слесаря. При этом (1–2) побуждают адресата к началу деятельности для достижения нужного результата. Следует подчеркнуть, что в (1) степень проблематичности реализации первичной цели (обнаружения слесаря) предполагается говорящим меньшей, чем в (2), т. е. употребление императивной формы СОВ *поищи* предполагает большую, чем в (1) труд-

ность достижения желаемого результата. Ведь в (2) содержится призыв лишь к кванту («порции») поисковой деятельности независимо от того, будет ли она иметь успех. А в (1) имеется в виду, что адресат должен прилагать свои усилия до тех пор, пока не обнаружит слесаря. В (3) же главный акцент падает на сам результат, а необходимость приложения ведущих к нему усилий считается само собой разумеющимся, причём какая-либо вероятность их безуспешности даже не допускается или остаётся «за кадром».

Отметим, что видовое противопоставление *ищи/поищи, найди* хорошо согласуется с проводимым в [5] тезисом о маркируемой формами СВ **первичности** идеи побуждения. В отличие от (2–3), в (1) форма НСВ *ищи* (вместе с прямым дополнением) выполняет кооперативную функцию практического рассуждения — совета в ранее актуализированной проблемной ситуации, в которой оказался слушающий (или слушающий вместе с говорящим). А (2) и (3) выступают в функции первичной просьбы, распоряжения, которые преследуют интересы прежде всего говорящего.

Дальнейшая дифференциация степени достижимости побуждаемого действия может отражаться с помощью поверхностного употребления императивных форм предикатов со значением попытки и приложения усилий — ср. *стараться, пытаться, пробовать* и т. д. В подобных ситуациях обе видовые формы данных глаголов (если они возможны) указывают на уже актуализованную проблемную ситуацию, «рецепт» преодоления которой выражается соответствующей императивной структурой. Полезно рассмотреть особенности сочетаемости этих глаголов с разными видовыми формами подчинённых предикатов, обозначающих побуждаемую акцию. Эти особенности во многом обусловлены спецификой значения каждого из внешних предикатов, что в свою очередь накладывает определённые ограничения на значения предикатов внутренних. Как видно из приводимых ниже примеров, далеко не все из теоретически возможных сочетаний видовых форм внешних и внутренних предикатов являются допустимыми. Правила согласования (непротиворечивости) сочетаний этих форм могут варьироваться для предикатов разных типов. Разумеется, приводимые ниже группы примеров и их комментариев очень далеки от того, чтобы охватить все возможные типы взаимодействия разных видовых форм императивных глаголов и подчинённых им предикатов. В качестве обозначений побуждаемых действий рассмотрены только три глагола:

1) *ложиться (спать)* как обозначение простого (контролируемого) действия;
2) *искать* как средство выражения слабо контролируемого процессуального действия, успех которого не может быть заранее гарантирован;
3) *находить*, указывающее на достижение цели действия ‘искать’.

(4.1.а) Постарайся лечь во-время
(4.1.б) Постарайся ложиться во-время
(4.1.в) ?Старайся лечь во-время
(4.1.г) Старайся ложиться во-время

(4.2.а) Постарайся найти слесаря
*(4.2.б) *Постарайся находить слесаря*
(4.2.в) ?Старайся найти слесаря
*(4.2.г) *Старайся находить слесаря*

(4.3.а) ?Постарайся поискать слесаря
*(4.3.б) *Постарайся искать слесаря*
(4.3.в) ?Старайся поискать слесаря
*(4.3.г) * Старайся искать слесаря*

Значение предиката *стараться* предусматривает, что его разные императивные формы указывают на исходящее от говорящего побуждение к целевому действию, результат которого требует от адресата приложения максимального количества усилий. При этом субъект-адресат должен быть нацелен на то, чтобы прилагать усилия до тех пор, пока нужный результат не будет достигнут. А предикат, подчинённый *стараться*, может обозначать как вполне контролируемое действие, так и действие, достижение цели которого нормативно не может быть заранее гарантировано.

Что касается видовых характеристик рассматриваемых императивных комплексов, то более естественным является совпадение (согласованность) видовых форм внешнего и внутреннего предикатов. Это в первую очередь относится к сочетанию типа СВ + СВ, указывающему на побуждение к выполнению одного действия — ср. примеры 4.1.а, 4.2.а и другие. Случаи, когда такое сочетание видовых форм оказывается присущим аномальным фразам, будут особо оговорены.

Как видим, в каждой из подгрупп рассматриваемой группы примеров (как и в последующих группах) последовательно просматривается аномальность сочетания НСВ + СВ — ср. (4.1.в, 4.2.в, 4.3.в). Действительно, побуждение к одному конкретному действию (*лечь во-время, найти, поискать*) не может выражаться формой НЕСОВ, которая в данном случае навязывает представление о побуждении к неоднократному выполнению этого действия.

Глагол *ложиться* в сочетании с сирконстантом *во-время* приобретает значение ‘укладываться спать’ и воспринимается как обозначение «моментального» (не процессуального) действия. Поэтому в форме НСВ *ложиться во-время* нормативно воспринимается в итеративном значении. Таким образом, императивные комплексы (4.1.б, 4.1.г) обозначают побуждение к выполнению многократных действий. Следует отметить, что они имеют практически одинаковые значения, которые, однако, формируются разными способами. Значение (4.1.г) может быть условно представлено как ‘Всякий раз если ложишься

спать, то делай это во-время'. А (4.1.б), где внешний предикат представлен формой СВ, указывает на то, каким образом следует выполнять целый комплекс однотипных действий типа 'ложиться спать'.

В примерах 4.2.б, 4.2.г НСВ от результатива успешно завершённого слабо контролируемого процессуального действия (*находить*) в сочетании с единственным числом прямого дополнения не допускает итеративной интерпретации, что и объясняет аномальность этих примеров. В то же время реализация объектного дополнения этого глагола во множественном числе делает соответствующие фразы приемлемыми: (4.2.б')*Постарайся находить слесарей;* (4.2.г')*Старайся находить слесарей*. Ведь в этом случае внутренний предикат приобретает дистрибутивное значение — множественное число, указывая на необходимость обнаружения не одного, а многих слесарей, тем самым указывает на результаты неоднократных поисков. Поэтому данные фразы допускают итеративную интерпретацию, подобную примерам 4.1.б и 4.1.г.

А все видовые варианты сочетаний, образующих подгруппу 4.3, оказываются аномальными. Это обусловлено значением внешнего предиката *стараться*, значение которого предусматривает обязательное достижение цели действия, обозначенного внутренним предикатом. Вместе с тем, значение глагола *искать* указывает лишь на цель, но не на её достижение. Данное противоречие и объясняет аномальность примеров 4.3.а — 4.3.г.

(5.1.а) ?Попытайся лечь во-время
(5.1.б) ?Попытайся ложиться во-время
(5.1.в) ?Пытайся лечь во-время
(5.1.г) ?Пытайся ложиться во-время

(5.2.а) Попытайся найти слесаря
*(5.2.б) *Попытайся находить слесаря*

(5.2.в) ?Пытайся найти слесаря
*(5.2.г) *Пытайся находить слесаря*

(5.3.а) ? Попытайся поискать слесаря
*(5.3.б) *Попытайся искать слесаря*

(5.3.в) ?Пытайся поискать слесаря
*(5.3.г) *Пытайся искать слесаря*

Императивная структура с внешним предикатом *пытаться* является средством побуждения к приложению усилий для выполнения заведомо трудной деятельности, ведущей к достижению нужного результата. В данном случае имеется в виду, что проблематичность достижения результата обусловлена именно трудоёмкостью способа, ведущего к цели, его неполной контролируемостью или же отсутствием изначального представления о способе, «пути», ведущему к нужному результату. В то же время не исключается, что деятельность к которой побуждается адресат, может оказаться и безрезультатной. Данное свойство объясняет частое употребление форм прошедшего времени *пытаться* (как и *пробовать*) в конструкции с элементом *было*, обозначающей неудачную попытку: *Он попытался (попробовал) было выехать, но…* Ср. в этой связи невозможность употребления в такой структуре формы СОВ *постараться*, предписывающей обязательное достижение преследуемой цели: **Он постарался было выехать, но…* Поэтому одно из важных требований, предъявляемых к значению внутреннего предиката при *пытаться*, состоит в том, что, указывая на конечный результат побуждаемой деятельности, он не должен быть обозначением действия, которое с общепринятой точки зрения считается заведомо контролируемым. Именно нарушением этого требования объясняется странность, если не аномальность примеров (5.1.а–5.1.г).

Внутри подгруппы 5.2. по естественным причинам (см. выше) нормально интерпретируется лишь сочетание СВ + СВ, где внутренний предикат представлен формой СВ (*найти*) от результатива слабо контролируемого действия *искать*.

Особый интерес представляют примеры (5.2.б и 5.2.г). Разумеется, они аномальны по тем же причинам, что и рассмотренные выше примеры (4.1.б) и (4.1.г) (невозможность итеративной интерпретации *находить* в сочетании с единственным числом его прямого дополнения). Однако если в данных примерах заменить единственное число этого дополнения на множественное, то результат такой замены будет несколько иной, чем тот, который наблюдался в случаях (4.2.г') и (4.2.б').

Сравним для начала фразы (4.2.а) *Постарайся найти слесаря* и (5.2.а) *Попытайся найти слесаря*. В (5.2.а) имеется в виду, что обнаружение слесаря представляет собой более трудоёмкую задачу, чем в (4.2.а). Ведь, в отличие от (4.2.а), (5.2.а) предусматривает, что поиск слесаря может окончиться и неудачей. То же отличие характеризует и фразы (4.2.г') *Старайся находить слесарей* и (5.2.г') *Пытайся находить слесарей*. Ведь в (5.2.г') подразумевается, что не каждая попытка обнаружения слесаря (или слесарей) будет успешной. В то же время, в отличие от (4.2.б') *Постарайся находить слесарей*, пример (5.2.б') ?*Попытайся находить слесарей* воспринимается странно. В нём речь идёт о побуждении к выполнению комплекса (множества) однотипных слабо контролируемых действий. В то же время одинаковый успешный исход **каждого** из действий, входящих в соответствующий комплекс, в силу самого значения *пытаться* не может быть гарантирован. Это противоречие и объясняет странность (5.2.б').

Значение *пытаться*, как и *стараться*, предусматривает, что ситуация, которая является конечной целью прилагаемых усилий, должна быть обозначена подчинённым предикатом. Как уже отмеча-

лось, глагол *искать*, выступающий в качестве подчинённого предиката и в подгруппе 5.3, указывает лишь на преследуемую цель (обнаружение нужного объекта), но не на её достижение. В принципе в изъявительном наклонении подобные глагольные сочетания возможны — ср. *Он не раз пытался её искать, но безрезультатно*. Однако императивная форма предиката *пытаться* требует, чтобы подчинённый ему глагол обозначал именно **достижение** конечной цели усилий, к которым говорящий побуждает слушающего. Невыполнение этого требования в подгруппе 5.3 объясняет аномальность примеров (5.3.а — 5.3.г).

(6.1.а) Попробуй лечь во-время
(6.1.б) Попробуй ложиться во-время
(6.1.в) ?Пробуй лечь во-время
(6.1.г) Пробуй ложиться во-время

(6.2.а) Попробуй найти слесаря
*(6.2.б) *Попробуй находить слесаря*
(6.2.в) ?Пробуй найти слесаря
*(6.2.г) *Пробуй находить слесаря*

(6.3.а) Попробуй поискать слесаря
*(6.3.б) *Попробуй искать слесаря*
(6.3.в) ?Пробуй поискать слесаря
*(6.3.г) * Пробуй искать слесаря*

Императивный комплекс, образуемый внешним предикатом *пробовать*, также является побуждением к выполнению не вполне контролируемого действия, достижение результата которого не может быть заведомо гарантировано. Но, в отличие от глагола *пытаться*, *пробовать* не может указывать на повторное, неоднократное приложение для достижения нужного результата **одинаковых** усилий или способов. *Пытаться* может быть использовано для обозначения множества одинаковых, настойчивых попыток — ср. (7) *Он долго пытался сдвинуть шкаф (ситуацию с мёртвой точки)*. В то же время для *пробовать* это неверно: (8) ?*Он долго (настойчиво) пробовал сдвинуть шкаф (ситуацию с мёртвой точки)*. Каждая новая «проба» требует обращения к новым средствам разрешения проблемной ситуации. Ср. (9) *Он уже (пере)пробовал много лекарств, но ни одно не помогло*. Рассматриваемое значение *пробовать* указывает на кратковременное приложение физических и / или интеллектуальных усилий, достаточное для того чтобы то ли добиться успеха, то ли убедиться в неэффективности или бесполезности этих усилий. Поэтому, в частности, сомнительны и фразы типа (6.2.б') и (6.2.г') с множественным числом объекта, в которых речь идёт о множестве одинаковых проб поисковой деятельности: (6.2.б') ?*Попробуй находить слесарей*; (6.2.г') ?*Пробуй находить слесарей*.

В то же время различается по крайней мере два типа интерпретации таких императивных комплексов, обусловленные разными вариантами данного значения *пробовать*. Дело в том, что в отличие от предиката *пытаться*, *пробовать* может сочетаться с глаголами, обозначающими как слабо контролируемые, так и вполне контролируемые действия.

В первом случае речь идёт о побуждении к действию, которое, с точки зрения говорящего, является объективно нужным или даже необходимым для решения уже актуализированной проблемы. Так, в (6.2.а) имеется в виду, что обнаружение слесаря — это объективно правильный путь для разрешения проблемы, которой озабочены оба коммуниканта. Отмеченная выше особенность рассматриваемого значения *пробовать*, по-видимому, объясняет приемлемость примера (6.3.а) (в отличие от 5.3.а). Неконтролируемость «пробы» в сочетании с её кратковременностью хорошо согласуется со значением *поискать*, которое, указывая на цель слабо контролируемой поисковой деятельности, не содержит указания на её достижение (т.е. обнаружение искомого объекта) и, кроме того, обозначает лишь «квант» этой деятельности.

Во втором случае (сочетание *пробовать* с обозначением контролируемого действия) последнее указывает на предполагаемый способ решения актуальной проблемы — способ, за однозначную эффективность которого говорящий не может полностью поручиться (ср. [1, 4]). Например, в (6.1.а, 6.1.б и 6.1.г) имеется в виду, что у говорящего нет полной уверенности в том, что однократный или периодический своевременный отход ко сну, не представляющий, по общепринятому мнению, каких-либо трудностей (в отличие, скажем, от самого слабо контролируемого процесса засыпания или своевременного пробуждения), поможет решить обсуждаемую проблему.

Отмеченная вариативность значения *пробовать* проявляется, например, в его сочетаемости с близкими по значению глаголами *помогать* и *поддерживать* — ср. (10) *Попробуй ей помочь* и (11) *Попробуй её поддержать*. В отличие от *помогать*, предикат *поддерживать* в наиболее распространённом непрямом значении — ср. *X поддерживает У-ка в Р* (где речь не идёт о поддержке морально-психологической) обозначает заведомо контролируемое действие. Это связано с тем, что эффективность поддержки зависит не от возможностей, а от воли Х-а. Данное значение *поддерживать* предусматривает, в частности, что Х обладает более высоким авторитетом (полномочиями, социальным, материальным статусом), чем У, что позволяет ему легко распоряжаться ресурсом, необходимым У-ку для осуществления его цели Р. Например, высокий социальный статус Х-а является весомым аргументом для того, чтобы склонить другие полномочные инстанции в пользу принятия положительного решения относительно У-ка. Таким образом, всё зависит только от воли Х-а — захочет ли он употре-

бить тот или иной подвластный ему ресурс в пользу интересов У-ка. Поэтому в (11) проблематичность усматривается не в том, удастся ли адресату побуждения её поддержать (в этом никто не сомневается), а в том, насколько полезной окажется эта поддержка для разрешения проблемной ситуации, которой озабочены участники коммуникации.

В то же время в значении *X помогает У-ку в P* о статусных различиях между Х-ом и У-ком ничего не сообщается. Поэтому в сочетании с *попробовать* в качестве внешнего предиката глагол *помогать* воспринимается как обозначение слабо контролируемого действия. Так, в побудительной фразе (10) не вызывает сомнения объективная необходимость оказания помощи, но остаётся неясным, удастся ли адресату побуждения предоставить ресурс, достаточный для реализации её цели. Ср. также естественность сочетания внешнего предиката *пытаться* (который, как мы помним, требует, чтобы подчинённый ему глагол не был обозначением заведомо контролируемого действия) с *помочь* и проблематичность его сочетания с *поддержать*: (12) *Попытайся ей помочь*, но (13) ?*Попытайся её поддержать* (не в значении морально-психологической поддержки).

Подобно тому, как рассмотренное модальное значение *пробовать* не может указывать на повторное приложение **одинаковых** усилий или приёмов, так и «дегустационное» *пробовать* предусматривает однократность снятия пробы. Ведь для того чтобы оценить качество блюда или напитка достаточно отведать его один раз. Поэтому в данном значении *пробовать* не может обозначать многократную дегустацию одного и того же: (14) ?*Он много раз пробовал пиццу Божоле)*. Указание на многократную дегустацию требует «родового» обозначения её объектов во множественном числе, которое подразумевает их вариативность: (15) *Он (пере)пробовал много блюд (напитков)*.

Императивное употребление «дегустационного» *пробовать* выполняет функцию побуждения не только оценить вкусовые качества продукта, но и поглотить его в объёме, достаточном для удовлетворения аппетита адресата. Нормативно императивное *попробуй(те)* и особенно *пробуй(те)* предполагает множественность предлагаемых блюд. При этом форма СВ может, в полном соответствии с принципом, отстаиваемым в работе [5], обозначать и первичность побуждения: (16) *Попробуйте мой салат*. Форма же НСВ должна опираться на прагматическую презумпцию о том, что адресат говорящего уже намерен отведать предлагаемое угощение и говорящий «подсказывает» ему, на чём остановить свой выбор в данный момент. В других случаях форма НСВ может выполнять и иллокутивную функцию разрешения-подбадривания адресата продегустировать уже выбранное им блюдо: (17) *Пробуйте мой салат*.

Те же особенности характеризуют употребление в императиве *пробовать* в другом варианте его «дегустационного» значения — ‘установить степень соответствия приготовленного (приготовляемого) блюда известной (субъекту) норме его вкусовых качеств’, т. е. определить, нет ли каких-либо отклонений от нормы (недосола, пересола, избытка или дефицита специй, подгорания и т. д.).

Ещё раз отметим, что рассмотренные здесь семантические эффекты, вызванные вариативностью глагольного вида в императиве, не могут претендовать на универсальность. Поэтому представляет несомненный интерес изучение поведения в императиве глаголов с другими особенностями лексического значения.

Литература

1. *Жолковский А. К.* Лексика целесообразной деятельности // Машинный перевод и прикладная лингвистика. — М.: 1964. — Вып. 8. С. 67–103.
2. *Князев Ю. П.* Грамматическая семантика. Русский язык в типологической перспективе. ЯЗЫКИ СЛАВЯНСКИХ КУЛЬТУР. МОСКВА, 2007. — 706 с.
3. *Падучева Е. В.* Семантические исследования: Семантика времени и вида в русском языке. Семантика нарратива. — М.: Языки русской культуры, 1996. — 464 с.
4. *Труб В. М.* Лексика целесообразной деятельности (опыт описания) // Логический анализ языка. Ментальное действие. Институт языкознания РАН, М. 1993. С. 58–66.
5. *Шатуновский И. Б.* Несовершенный vs. совершенный вид в императиве (к проблеме начала) // Логический анализ языка. Семантика начала и конца. Издательство «ИНДРИК», Москва 2002. — С. 267–309.
6. *Benacchio Rosanna.* Конкуренция видов, вежливость и этикет в русском императиве // Russian Linguistics 26. 2002. P. 149–178.
7. *Bogusławski Andrzej.* Pary czy wieloczłony aspectowe? // PRACE FILOLOGICZNE. TOM XLVI. Warszawa 2001, s. 69–77.

Разговорные словечки *как бы* и *конкретно*[1]

Russian colloquial words *kak by* (lit. ‘as if, like’) and *konkretno* (lit. ‘specifically’)

Урысон Е. В. (x-uryson@mtu-net.ru)

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

Описывается семантика частиц как бы и конкретно, ср. Я как бы отчет написал, Яблоки конкретно сладкие. Слово как бы в своем основном значении указывает на сходство и закономерно развивает значение принадлежности к классу. Структура многозначности слова конкретно развивается по аналогии со структурой многозначности слова вообще; причина в том, что слова вообще и конкретно в некоторых значениях антонимичны.

0. Объект предлагаемого исследования — разговорные слова *как бы* и *конкретно*, представленные в контекстах типа:

(1) Студентка обращается к преподавателю с просьбой перенести защиту/сдачу курсовой работы. Мотивируя свою просьбу, она произносит: *Я как бы беременна* [сообщено М. А. Реформатской].

(2) [аспирант, мнс, обращается к руководителю сектора] *Юрий Трофимович, я тут как бы смету составил по гранту, Вы не посмотрите*.

(3) *Я тут как бы на операцию ложусь.* [*Ты не могла бы гулять с моей собакой на следующей неделе?*]

(4) [диалог в магазине] — *Вот эти яблоки кислые или сладкие? — Яблоки конкретно сладкие*.

(5) [диалог в магазине] — *Эти ботинки все-таки на осень. А у вас нет зимних ботинок, на меху? — Вот на этой полке конкретно зимняя обувь*.

Подобные высказывания вызывают улыбку у людей, следящих за своей речью: *как бы* и *конкретно* воспринимаются как смешные словечки-паразиты. При этом они имеют разную стилистическую маркированность. *Как бы* употребляется очень широко, причем и в интеллигентной среде. *Конкретно* воспринимается, скорее, как слово вульгарное. Стилистический анализ данных слов не входит в нашу задачу. Наша цель — показать, что слова *как бы* и *конкретно* закономерно развивают данную семантику. Иными словами, употребление слов *как бы* и *конкретно* в подобных контекстах закономерно с точки зрения лексической системы языка.[2]

1. Начнем с *как бы*. Прежде всего обратим внимание на то, что высказывания типа (1)–(3) весьма часто сопровождаются специфической просодией — говорящий как будто немножко извиняется. Он сообщает соответствующую информацию вскользь. Эта информация — не главная, обычно она предваряет более важную информацию. Говорящий старается быть максимально вежливым, в частности, он старается не акцентировать эту предваряющую информацию. Таковы примеры (1)–(3). Впрочем, такая просодия необязательна. Ср.

(6) *Мы как бы унифицировали модельный ряд (автомашин)* [представитель пресс-центра сообщает о том, что чиновники отныне будут ездить на машинах отечественного производства; тональность высказывания деловая и жесткая].

Очевидно, что слово *как бы* в данных контекстах выступает в каком-то особом значении — в термино-

[1] Работа выполнена при финансовой поддержке Программы фундаментальных исследований ОИФН РАН «Русская культура в мировой истории» и гранта Президента РФ № НШ-5611.2006.6.

[2] Слово *конкретно* в интересующем нас значении как будто еще никем не рассматривалось. Слову *как бы* посвящено исследование А. А. Летучего [Летучий…], в котором сравниваются союзы *как бы* и *как будто* и при этом дается характеристика дискурсивных употреблений *как бы*. Наши результаты не противоречат этой работе.

логии московской семантической школы перед нами особая лексема слова *как бы*. С точки зрения грамматики это частица. Она выражает какую-то специфическую модальную рамку. Заметим, что если Р — в фокусе внимания, например, если говорящий подчеркивает, что Р, или эмфатически выделяет Р, то словечко *как бы* не употребляется. Мы не услышим:

(7) **Представляешь, она как бы беременна!* [нормально: *Представляешь, она беременна!*]

(8) *Ура! *Наконец-то как бы смету составил!* [нормально: *Ура! Наконец-то смету составил!*]

(9) **Я как бы на операцию ложусь, предупредили, что исход — пятьдесят на пятьдесят. Нужно родным сказать* [в таком контексте нормально: *Я на операцию ложусь*].

Подобные высказывания в лучшем случае пародийны.

Тонкие семантические нюансы, которые выражаются в подобных контекстах с *как бы*, описаны в статьях А. А. Летучего [Летучий 2008 а,б]. Мы ставим следующий вопрос: каким образом данная лексема *как бы* семантически связана с другими значениями этого служебного слова?

В идеале для ответа на этот вопрос требуется дать лексикографический портрет слова *как бы* и указать мосты (семантические переходы) между его лексемами. Мы сейчас не будем решать эту задачу в полном объеме. Обрисуем лишь нужный нам фрагмент полисемии слова *как бы* и покажем, что он вполне системен: в русском языке есть и другие слова с аналогичным центральным значением и с таким же фрагментом структуры полисемии.

Союз *как бы* в своем центральном значении выражает сравнение. Ср.

(10) [*Кот*] *потерся о ноги Кузьмина, промурлыкал и ушел обратно в ночные комнаты, как бы приглашая Кузьмина за собой* (К. Паустовский, МАС).

В [МАС] данное сравнение квалифицируется как условно-предположительное, [Грамматика-80] называет такое сравнение недостоверным. Есть еще ряд союзов, выражающих такое недостоверное, или условно-предположительное, сравнение: *как будто, словно, точно* и т.п.

(11) *Конь спотыкается под ним, Храпит, как будто гибель чует* (М.Ю.Лермонтов).

Для полноты приведем пример на достоверное сравнение:

(12) *Зачем Арапа своего Младая любит Дездемона, Как месяц любит ночи мглу?* (А. С. Пушкин).

(13) *Наташа стала, как была, Опять румяна, весела* (А. С. Пушкин).

(14) *Наш сад как проходной двор* (А. П. Чехов).

Мы не будем сейчас разбирать различия между достоверным и недостоверным сравнением. Нам важно, что *как бы* входит в класс сравнительных союзов.

Сравнительные слова, в частности сравнительные союзы, указывают на сходство двух объектов или ситуаций. Так, в (10) речь идет о том, что действия кота похожи на приглашение. Грубо говоря, в (12) речь идет о том, что любовь Дездемоны к Отелло похожа на то, как месяц любит ночи мглу. В (14) говорится о сходстве данного сада с проходным двором.

Но в примерах (1)–(3) *как бы* не указывает на сходство. Требуется понять, как устроен переход от сравнительного слова *как бы* к тому слову-паразиту, которое представлено в данных примерах.

Обратим внимание на следующий факт.

Идея сходства некоторых объектов лежит в основе объединения объектов в класс. Действительно, человек объединяет в один и тот же класс однотипные, сходные объекты. Класс — это множество сходных объектов. Тем самым, ‘класс’ и ‘сходство’ — это связанные друг с другом смыслы: ‘сходство’ — это компонент смысла ‘класс’. Мы вправе ожидать в системе языка полисемию следующего типа:

(15) ‘класс Х-ов’ — ‘принадлежащий классу Х-ов’ — ‘похожий на Х’.

Более точно: мы вправе ожидать сосуществование в многозначном слове лексем с такими значениями. Подчеркнем, что мы не занимаемся направлением развития многозначности, т.е. не ставим вопрос, какое значение послужило источником для развития другого значения. Для решения нашей задачи достаточно статичного подхода, констатирующего сосуществование таких-то значений в структуре многозначности.

Проверим нашу гипотезу.

Слово *тип*. Одна из лексем этого слова обозначает совокупность однотипных, похожих в некотором отношении объектов, т.е. класс объектов. Такой лексемой является биологический термин *тип*. Ср. *тип членистоногих*.

(16) *тип членистоногих* — ‘класс живых организмов, который составляют Х-ы’.

Очень близкая лексема слова *тип* представлена в контексте *новый тип двигателя*. В [МАС] это значение слова *тип* толкуется так:

(17) *новый тип двигателя* [*тип Х-а*] — ‘модель Х-а, которой соответствует класс Х-ов’ [МАС].

Предлог *типа* + РОД:

(18) *Построил что-то типа сарая*; *Схватил что-то типа гайки*.

Словосочетание *что-то типа X* указывает на сходство данного объекта (‘что-то’) с Х-ом. При этом существительное Х всегда выступает в родовом денотативном статусе. Как известно, существительное в родовом денотативном статусе обозначает класс объектов. Поэтому в каком-то смысле можно говорить, что семантика класса сохраняется и у предлога *типа*+РОД. Ясно, однако, что у этого предлога на первом плане — идея сходства, а идея класса уходит на задний план: она отражается лишь в денотативном статусе управляемой лексемы.

В следующем типе контекстов наблюдаем аналогичную перестройку семантики:

(19) *Пробормотал какие-то слова, типа того что очень сожалеет*.

Слово *типа* здесь тоже указывает на сходство: конкретная ситуация ‘он пробормотал какие-то слова’ интерпретируется говорящим как похожая на ситуацию ‘он выразил сожаление’.

Несколько иначе устроен следующий пример:

(20) *Петя типа того что заболел* [не утверждается прямо, что заболел; заболел или находится в состоянии, имеющих много общих признаков с болезнью].

Данное высказывание обычно в определенном прагматическом контексте. Например, говорящий не уверен в том, что Петя заболел, и поэтому не хочет говорить ‘Петя заболел’. Вместо этого он сообщает следующее: ‘Петя находится в состоянии, имеющем много общих признаков с болезнью’. Или говорящий знает, что Петя не заболел, и не хочет врать, но вынужден говорить неправду (‘Петя заболел’); «смягчая» свою ложь, он подает информацию так, как будто она не вполне точная. Или же говорящий не хочет сообщать эту информацию, но ему приходится это делать, и он «смягчает» ее, как в предыдущем случае. Ср. аналогичный пример:

(21) *Она типа того что беременна*.

По-видимому, высказывания типа (20)–(21) имеют следующую семантическую структуру: *X типа того что P* = ‘X P; или ситуация P(X) имеет много общих признаков с ситуацией P’. Как кажется, идея класса в подобных высказываниях практически стерта.

Примеры (20)–(21) интересны с синтаксической точки зрения: в них лексически выражен только «эталон сравнения», а само состояние субъекта не обозначено никаким словом — оно подразумевается. Между тем в остальных случаях лексически обозначены и сравниваемый объект и эталон сравнения. Ср. *что-то* [сравниваемый объект] *типа сарая* [‘сарай’ — эталон сравнения]; *пробормотал…* [сравниваемая ситуация] *типа того что извиняется* (‘извиняется’ — эталон сравнения].

Еще один пример интересующей нас многозначности — слово *род*.

В следующем примере слово *род* обозначает класс. Ср.

(22) *Мужчины были двух родов: одни тоненькие, которые увивались около дам <…>. Другой род мужчин составляли толстые* (Н. В. Гоголь, МАС).

Биологический и логический термин *род*, представленный в случаях типа *род воробьиных*, *этот род явлений*, также обозначает класс объектов.

Родственный слову *род* предлог *вроде* (из *в* + *роде*) сближается с предлогом *типа*. Ср.

(23) *Что-то вроде самолета*.

(24) *Он вроде тебя — такой же упрямый*.

(25) *Он вроде того что болен* [говорится, что состояние субъекта похоже на состояние ‘быть больным’].

О подобных употреблениях слов *типа* и *вроде* писала Н. Д. Арутюнова.

Для нас сейчас важен семантический переход ‘класс’ — ‘сходство’, который мы наблюдаем в гнезде *род — вроде*.

Аналогичный случай: существительное *вид* — предлог *в виде (звездочки)*. Существительное указывает на класс, предлог — на сходство формы. Ср. *(этот) вид грибов* VS. *коржик в виде звездочки*. С точки зрения динамического описания направление семантического перехода здесь такое: от указания на внешний вид к указанию на совокупность. Между тем, в случаях выше переход прямо противоположен. Но для статического описания важно не направление перехода, а факт сосуществования в пределах одного многозначного слова лексем с данной семантикой.

Интересно в рассматриваемом отношении слово *такой* [Богуславская 1991] (некоторые примеры ниже заимствованы из этой работы).

(26) *Маша носила старинное серебряное кольцо. Такое кольцо было у его прабабки*.

Упрощенно:

(27) *такой X* = ‘X, имеющий то же свойство, что и объект, упомянутый ранее, и на этом основании объединяемый говорящим в один класс с этим упомянутым объектом’.

Такое кольцо … — это кольцо, которое имеет то же свойство ‘старинное и серебряное’, что и кольцо, которое носит Маша (это объект, упомянутый ранее); на этом основании данное кольцо (т.е. кольцо прабабки) объединяется в один класс с кольцом Маши.

Ср. аналогичный пример:

(28) *В самом углу у двери примостился маленький столик. Такие [столики] стоят в уличных кафе* (Ю.Домбровский).

Такой указывает на тождество каких-то свойств. На основе тождества свойств объекты объединяются в класс. Тождество и сходство — это разные вещи, хотя между ними есть определенная общность. Замечательно, что и у слова *такой* есть значение, в котором есть только указание на принадлежность к классу, а указание на тождество уже отсутствует. Ср.

(29) *Карл Маркс — это такой экономист* (из анекдота).

(30) *К вечеру вышли к Рождественскому — село такое старинное на берегу Волги*.

Здесь *такой* указывает только на принадлежность к данному классу:

(31) *Y — это такой X* = ‘Y принадлежит к классу X-ов’.

В одном значении слова *такой* сосуществуют компоненты ‘тождество свойств’ и ‘класс’. А в последнем значении компонента ‘тождество’ нет, есть только компонет ‘класс’.

Подчеркнем, что когда в подобных случаях говорят о семантическом переходе, то это условность «динамического» описания. В действительности мы можем говорить лишь о сосуществовании в структуре полисемии слова или в гнезде родственных слов двух лексем. Одна указывает на класс сходных объектов, а другая — только на сходство. (Реальное направление развития полисемии — объект отдельного описания).

Вернемся к слову *как бы*. Мы считаем, что в наших разговорных высказываниях это словечко полностью утратило компонент ‘быть похожим’ и указывает на класс ситуаций.

Как бы P = ‘говорящий сообщает, что данная ситуация входит в класс ситуаций P’.

Иными словами, говорящий не напрямую сообщает адресату: *P*, а говорит, что имеющая место ситуация относится к классу P. Этот «непрямой» способ сообщения и создает почву для развития специфических модальностей, присущих контекстам с данным словечком и часто выражаемых просодически.

Как бы в рассматриваемом значении выступает как своего рода артикль: указание на принадлежность к классу — это одна из семантических функций неопределенного артикля, но *как бы* — это «артикль» не перед обозначением предмета, а перед обозначением ситуации.

Итак, слово *как бы*, подобно другим лексическим единицам, указывающим на сходство, имеет значение, указывающее на класс. Перед нами явление, которое мы называем или полисемической аналогией [Урысон 2003].

2. Слово *конкретно*. Мы начнем наше рассуждение со слов, как будто не имеющих ничего общего с описываемыми словами. Это *вообще* и *вообще-то*. Слово *вообще* описано Ю.Д. Апресяном, в частности, в связи с его нетривиальными просодическими свойствами [Апресян 1995б: 188–189; 2004] (ниже используются примеры из этих работ); см. также статью Н.В.Гатинской [Гатинская 2002].

Начнем со следующих примеров:

(32) *Он вообще <вообще-то> добрый, только деньги не любит в долг давать*.

(33) *Она вообще-то ничего, только нос длинноват*.

Вообще и *вообще-то* указывают здесь на то, что описываемая ситуация обладает очень многими, но не всеми признаками ситуации P (‘быть добрым’, ‘иметь неплохую внешность’). Эта семантика очень близка семантике, описанной выше. Действительно, ‘входить в класс A’ ≈ ‘обладать всеми признаками объектов из класса A’. В класс объединяются однотипные объекты, а однотипные объекты — это объекты, обладающие одинаковыми (или сходными) признаками.

Словечки *вообще* и *вообще-то* очень характерны для разговорной речи. Подобно *как бы*, они указывают на класс, но в данном случае это класс признаков ситуации P. Но *вообще* и *вообще-то* указывают не просто на класс признаков, а еще и на то, что в данном случае речь идет почти о всех, но все-таки не о всех признаках P. Заметим, что может возникнуть необходимость в обратном — подчеркнуть, что речь идет о всех без исключения признаках P. Но прежде чем переходить к этой теме, рассмотрим другое значение слова *вообще*.

(34) *Пили за дам вообще и за Людмилу в частности <в особенности, конкретно>*.

Здесь *вообще* указывает на совокупность, на класс. Отдельный объект класса — это конкретный представитель класса. Общее (класс) противопоставляется частному, единичному, конкретному. Прилагательное *конкретный* оказывается антонимом прилагательного *общий* и развивает ту же структуру полисемии, что и *общий*. Причем *конкретный*

начинает выступать как антоним не только слова *общий*, но вообще слов с корнем *общ-*. Ср.

(35) *Пили за дам вообще и за Людмилу конкретно*.

В таком контексте слово *конкретно* противопоставляется словам *в частности* и *в особенности*. Слово *в частности* указывает на то, что данный представитель класса — один из многих рядовых его представителей и, естественно, не главный из них. Слов *в особенности*, напротив, указывает на то, что данный объект выделен по признаку P ('пили за дам') — за Людмилу пили больше, чем за остальных дам. Слово *конкретно*, в отличие от *в частности* и *в особенности*, указывает только на то, что объект является представителем данного класса, и не выражает никаких «лишних» смыслов.

Теперь обратимся к примеру *Яблоки конкретно сладкие*. Данная лексема *конкретно* противопоставлена словам *вообще* и *вообще-то*. Ср. *Яблоки вообще-то сладкие (но попадаются с кислинкой)*. Лексема *конкретно* указывает на то, речь идет о всех без исключения признаках P. Аналогами этой лексемы являются слово *безоговорочно* и выражение *И никаких гвоздей!*

Возьмем теперь еще одну лексему слова *вообще*. Ср. *Я вообще слышал эту версию* [пример Ю.Д.Апресяна]; *Я вообще всегда хотел стать врачом*; *Мы тогда вообще в Париже жили*. Как отметил Ю.Д.Апресян, эта лексема *вообще* никогда не несет никакого фразового ударения. Ю.Д.Апресян толкует эту лексему так: 'говорящий оставляет за собой право на оговорки и уточнение своего чересчур общего утверждения'. Данная лексема слова *вообще* чрезвычайно характерна для разговорной речи, она превращается в своего рода «слово-заполнитель», почти слово-паразит. Класс подобных словечек постоянно обновляется. Естественно предположить, что обновление этого класса словечек подчиняется своим закономерностям. В частности, в этом классе может появиться (квази)антоним уже имеющемуся слову. Какой может быть антоним у данной лексемы *вообще*? Какая-то лексема слова *конкретно*. Действительно, если слово *конкретно* в каком-то своем значении уже выступает как антоним некоторой лексемы слова *вообще*, то естественно, что слово *конкретно* разовьет ту структуру многозначности, которой обладает слово *вообще*. Это предположение поддерживается следующим системным фактом.

Слова, относящиеся к одному и тому же полю, развивают схожие структуры многозначности. Этот факт, по-видимому, был впервые описан в конце 19 века проф. М.М. Покровским на материале древних языков [Покровский 1959/1895]. Хорошо известно развитие по аналогии структуры полисемии у синонимов, см. об этом Апресян 1995а:225 и ссылки там. (Точнее, если какие-то две лексемы синонимичны, то многозначные слова, к которым относятся данные лексемы, развивают аналогичные структуры многозначности.) Другие случаи развития многозначночти по аналогии описаны в работе [Урысон 2003] и в трудах Е.В.Падучевой, в частности в книге [Падучева 2004]. Естественно предположить, что структура многозначности антонимов тоже может развиваться по аналогии.

Тогда у слова *конкретно* должна быть лексема, антонимичная лексеме *вообще*, представленной в случаях типа *Мы тогда вообще в Париже жили*. И действительно, в разговорной речи мы можем услышать высказывания типа: *Мы тогда жили конкретно в Париже*, где фигурирует «вульгарная» лексема *конкретно*. (Она несколько отличается от той лексемы, которая выступает в случаях типа *Яблоки конкретно сладкие* — здесь лексема *конкретно* противопоставлена другой лексеме *вообще*, ср.: *Яблоки вообще(-то) сладкие*.)

Мы попытались показать, что наличие в языке смешных разговорных словечек-паразитов *как бы* и *конкретно* вполне закономерно: они заполняют «пустые клетки» в лексической системе языка. Структура многозначности слова *как бы* закономерно развивается от указания на сходство к обозначению принадлежности к классу. Слово *конкретно* в одном из своих значений выступает как антоним слова *вообще*, и в результате структура многозначности слова *конкретно* развивается по аналогии со структурой многозначности слова *вообще*. Рассматриваемые словечки возникли в разговорной речи как результат действия аналогии в системе языка.

Литература

1. *Апресян Ю. Д.* (1995а) Лексическая семантика // Ю. Д. Апресян. Избранные труды. Т. 1. М., 1995. [Первое издание: М., 1974].
2. *Апресян Ю. Д.* (1995б) Типы лексикографической информации об означающем лексемы // Ю. Д. Апресян. Избранные труды. Т. 2. М., 1995. [Впервые в сб. «Типология и грамматика», М., 1990].
3. *Апресян Ю. Д.* «Фразовое ударение» // Лингвистическая терминология словаря / Новый объяснительный словарь синонимов русского языка. Изд. 2-е, исправл. и дополн. М. — Вена, 2004.
4. *Богуславская О. Ю.* Словарная статья местоимения ТАКОЙ // Семиотика и информатика. Вып. 32. М., 1991. С. 42–54.
5. *Гатинская Н. В.* Портрет лексемы ВООБЩЕ (опыт функционально-семантического описания) // Вестник Московского университета. Серия 9. Филология. 2002. № 5. С. 122–130.
6. *Летучий А. А.* Конструкции сравнения ситуаций с показателями как бы и как будто // НТИ.. Сер. 2. 2008. № 2. С. 19–27.
7. *Летучий А. А.* Сравнительные конструкции, ирреализ и эвиденциалис // Wiener Slawistischer Almanach. Bd. 72. 2008.
8. *МАС* — Толковый словарь русского языка в 4-х томах. М., 1985–1990.
9. *Падучева Е. В.* Динамические модели в семантике лексики. М., 2004.
10. *Покровский М. М.* Семасиологические исследования в области русских языков [первое издание: 1895] // М.М. Покровский. Избранные работы по языкознанию. М., 1959.
11. *Урысон Е. В.* — Проблемы исследования языковой картины мира: Аналогия в семантике. М., 2003.

Значения предлогов «по» и «к» в русском языке: кодирование сирконстантов и семантических ролей

Meanings of the prepositions «po» and «k» in Russian: encoding of adjuncts and semantic roles

Усачёва М. Н. (mashastroeva@gmail.com)

Московский государственный университет им. М. В. Ломоносова, Россия

Работа посвящена применению аппарата описания семантики пространственных показателей, разработанного преимущественно на материале дагестанских языков, но претендующего на типологическую достоверность ([Ганенков 2002, 2005], [Мазурова 2007]), к значениям двух русских предлогов — «по» и «к».

В последние несколько десятилетий языковые единицы, входящие в семантическое поле пространства, находятся в центре внимания представителей различных научных дисциплин — лингвистов, литературоведов, психологов и др. (краткий обзор работ, посвященных данной теме, содержится в диссертации [Мазурова 2007], с. 5–7). Пространственные концепты подвергаются всестороннему изучению. Рассматриваются как отдельные пространственные лексемы, так и лексико-семантические группы и семантические поля, привлекаются достижения культурологи и антропологии. В последнее время особое внимание уделяется составлению полных и систематизированных описаний слов с пространственной семантикой, основанных на анализе широкого типологического материала. Такой подход представлен, в частности, в работах [Ганенков 2002] и [Мазурова 2006]. Д. С. Ганенков делает акцент на непространственных значениях пространственных показателей языков мира, в то время как в центре внимания Ю.В. Мазуровой находятся средства выражения пространственной локализации на вертикальной оси. Концепты[1], выделяемые авторами в процессе анализа значений пространственных лексем, предлагается включать в Универсальный грамматический набор. Таким образом, данные концепты претендуют на статус «грамматических атомов» — неделимых значений, формально представленных хотя бы в одном языке мира.

Набор концептов, выделенных Д. С. Ганенковым и Ю. В. Мазуровой, уже успешно использовался нами для анализа пространственных падежей ряда уральских языков (см. работы [Строева 2008a,b]). В данной статье излагается опыт применения такого аппарата описания к русскому языку, который по ряду параметров значительно отличается от уральских. Во-первых, в русском языке пространственные значения кодируются преимущественно предлогами и пространственными наречиями. В случае же с уральскими языками нами рассматривались падежные значения. Во-вторых, материалы по уральским языкам были получены нами методами анкетирования информантов и изучения доступных текстов, представляющих собой записи устной спонтанной и квазиспонтанной речи. В ходе работы с русским языком использовался Национальный корпус русского языка (http://www.ruscorpora.ru), содержащий письменные тексты. Первое различие, с нашей точки зрения, особого значения не имеет, поскольку при выделении прототипических концептов авторы названных работ рассматривали различные языковые единицы — не только падежи и адлоги[2], но и превербы, наречия, глагольные суффиксы и локативно-директивные глагольные основы. Второе же различие оказывается существенным, поскольку аппарат Д. С. Ганенкова и Ю. В. Мазуровой разрабатывался с опорой на бесписьменные языки. Смена модуса влечет частичную смену лексики и грамматики, что, как будет показано ниже, снижает описательную силу аппарата; кроме того,

[1] К сожалению, набор концептов, выделяемых данными авторами, слишком велик, чтобы приводить его здесь: он включает 11 прототипических концептов локализации и 60 базовых непространственных значений. Мы надеемся, что в ходе изложения читатель сможет составить примерное представление об используемом нами аппарате.

[2] Адлогами мы, следуя за Ю.В. Мазуровой, называем предлоги и послелоги ([Мазурова 2007, с. 67]).

встает вопрос о необходимости приведения его к более строгому виду.

В качестве объекта исследования нами были взяты русские предлоги «к» и «по» в сочетаниях с существительными, обозначающими пространство и место. Эта семантическая группа была выбрана не случайно. По нашему предположению, в большинстве сочетаний с такими существительными предлоги «к» и «по» должны выражать базовые, «прототипические» пространственные значения, от которых достаточно легко отделить значения предложных групп, не поддающиеся описанию в терминах локализаций и ориентаций. Впоследствии так и оказалось. Анализу подвергались только предложные группы, зависимые от предикативов. Группы, зависимые от отглагольных имен (типа «приезд к родственникам»), из рассмотрения исключались.

Материал был получен из Национального корпуса русского языка (http://www.ruscorpora.ru) с использованием семантических фильтров «пространство и место», «здания и сооружения», «вместилища» и «горизонтальные поверхности». Рассматривались только свободные словосочетания; фразеологизированные обороты и коллокации (например, «по горячим следам») из рассмотрения исключались[3]. Было проанализировано около 13.000 предложений из текстов второй половины XIX — начала XXI в.в., содержащих глагольные актанты и сирконстанты вида «предлог к/по + существительное из вышеуказанных семантических групп». Таких существительных было найдено 84[4]: *аквариум, апартаменты, аэровокзал, балка, банк, банка, бездна, берлога, водоём, водохранилище, выбоина, гавань, галерея, гнездо, гнездышко, дебаркадер, закуток, залив, источник*[5]*, канал, карьер, колдобина, колея, ледник, ложа, ложе, лоток, лужа, луна, лунка, манеж, могила, море, муравейник, нора, нутро, овражек, озеро, океан, окоп, омут, отпечаток, отсек, очаг, пазуха, планета, платформа, площадка, подземелье, пол, поле, пристань, притон, пролив, пропасть, проток, протока, пруд, резиденция, река, речка, родник, родничок, русло, ручей, рытвина, сеновал, скважина, склон, след, солнце, солнышко, станция, сцена, тайник, термитник, терраса, трещина, ухаб, ущелье, щелка, щель, эшафот, юрта.*

Большинство существительных из списка, по данным корпуса, сочетается с обоими рассматриваемыми предлогами. Четыре существительных (*карьер, отпечаток, площадка, ухаб*) не встретились в сочетаниях с предлогом «к», одиннадцать (*источник, родник, родничок, солнышко, очаг, аквариум, гнездышко, пазуха, скважина, термитник, щелка*) — с предлогом «по». С нашей точки зрения, этот факт связан с неполнотой корпуса. Так, например, предложения типа «Вася шел по солнышку» представляются столь же естественными, как и предложения типа «Вася шел по солнцу» (последние в корпусе встречаются). Контексты «ударить по термитнику палкой» и «подойти к термитнику», как кажется, также не противоречат языковой интуиции.

Большая часть полученных нами сочетаний распадается на три группы, семантика которых хорошо описывается в терминах базовых локализаций (In, Super, Sub, Ad, Post, Ante, Apud, Inter) и ориентаций[6] (*эссив, латив, элатив, пролатив, директив*):

A. «Глагол движения или перемещения + предлог к/по + существительное»

В данной группе предлог «по» выражает сочетание пролативной ориентации ‘перемещаясь через’ с концептами In ‘внутри’ и Super ‘на’. С помощью первого концепта описываются контексты со словами из подгруппы зданий и сооружений (аэровокзал, банк, дебаркадер, юрта и т.п., см. *пример (1)*) — а также с существительными типа «подземелье», обозначающими пространственные вместилища (*пример (2)*):

(1) *Вскоре после того, как я уселся на орон (нары), на подложенные под меня в несколько рядов оленьи шкуры, мое внимание было привлечено*

---

[3] Заметим сразу, что основная цель данной статьи состоит в изложении результатов применения некоторой методики к описанию семантики пространственных показателей. Мы стремимся скорее к демонстрации достоинств и недостатков избранного нами метода, чем к полному и подробному описанию языковых единиц,. Примером такого описания является работа [Иомдин 1990], посвященная рассмотрению значений предлога «по».

[4] Разумеется, семантическая группа существительных со значением пространства и места далеко не исчерпывается восьмьюдесятью четырьмя словами. Во-первых, не каждое существительное, включенное разработчиками корпуса в данную группу, встречается в сочетании с предлогами «к» и «по». Примером может служить слово «депозитарий» из подгруппы зданий и сооружений. Во-вторых, семантическая разметка корпуса, видимо, не является законченной. Так, слово «дом», очевидно, являющееся одним из центральных в группе зданий и сооружений и в то же время высокочастотным, в эту группу не попало. Среди обозначений пространственных объектов наличествует слово «овражек», а «овраг» отсутствует. Однако представляется, что данный факт не имеет большого значения для нашего исследования, поскольку описание семантической группы пространственных существительных в его задачи не входит, а для выделения основных групп значений двух предлогов восьмидесяти четырех слов, на наш взгляд, вполне достаточно.

[5] Некоторые из представленных в таблице слов имеют несколько значений (например, «телевизионный канал» vs «Суэцкий канал »). В рассмотрение включались только пространственные значения этих слов.

[6] О категориях и граммемах локализации и ориентации см., например, [Мельчук 1998, с. 47–60, 324–350].

к серенькому шарику, который безостановочно ***шмыгал по юрте*** *под ногами у присутствовавших. [В. М. Зензинов. Нена (1925)]*

(2) *Ведь однородная масса из тысячелетнего льда, камней и грязи* ***пронеслась по ущелью****, сметая все на своем пути. [Александр Богомолов. В Кармадонском тоннеле могут найти тела людей. Тоннель. Что знала девушка Нана (2003) // «Известия», 2003.02.13]*

Концепт Super выделяется в сочетаниях с прочими существительными:

(3) А со мной был годовалый сын Филипп, который ползал по полу между безумно дорогими вазами и уникальным антикварiатом… [Артем Тарасов. Миллионер (2004)]

Предлог «к» в таких контекстах выражает директивную ориентацию:

(4) *Во все стороны было одно и то же — гора, хвоя;* ***спуститься к морю****, подумал Ять, тамто уж вдоль берега вернемся, — но и вниз пути не было, ибо склон вдруг обрывался. [Дмитрий Быков. Орфография (2002)]*

Б. «Глагол местонахождения + предлог по + существительное»

В данную группу входят значения предложной группы вида ‘концепт In/Super+ эссивная ориентация’. In возникает в сочетаниях предлога «по» с названиями вместилищ, зданий и сооружений (примеры (5), (6)), которые при этом выступают, как правило, во множественном числе, что придает предложениям дистрибутивный оттенок:

(5) *Вот-вот новая пугачевщина сотрясет отчизну нашу, а мы будем* ***сидеть по щелям*** *своим. [Борис Васильев. Дом, который построил Дед (1990–2000)]*

(6) *Живет месяц, другой,, а потом опять исчезает,* ***ютится по притонам****, рисуя в трактирах, по заказам буфетчиков, за водку и еду. [В. А. Гиляровский. Москва и москвичи (1934)]*

Super выделяется в сочетаниях с существительными, обозначающими горизонтальные поверхности:

(7) *Я ринулся на манеж. Рита* ***распласталась по полу****. Она дышала тяжело, с присвистом. [Вальтер Запашный. Риск. Борьба. Любовь (1998–2004)]*

Предлог «к» с глаголами местонахождения не сочетается, что понятно, поскольку основное его пространственное значение — директивная ориентация.

В. «Глагол физического воздействия + предлог к/по + существительное»

В данном случае предлог «к» также выражает директивную ориентацию, однако в этом случае добавляется семантический компонент ‘наличие контакта с ориентиром’:

(8) *Подтянуть ягодицы. Стопы ног плотно* ***прижать к полу****. МАТИГА. [Алексей Яшкин. Стойки в каратэ (2004) // «Боевое искусство планеты», 2004.09.09]*

Значение предлога «по» в таких контекстах описывается через взаимодействие пролативной ориентации и концепта Super:

(9) *Подала им знак и* ***прочертила по морю*** *золотую линию к берегу в стороне от « погибелья »… [Вадим Бурлак. Хранители древних тайн (2001)]*

В данных трех группах представлены основные, пространственные значения предлогов «к» и «по». С описанием таких значений проблем не возникает. Однако сочетания рассматриваемых предлогов с пространственными существительными могут иметь и непространственные значения, см. примеры *(10)* и *(11)*:

(10) *И, догоняя караван, я всегда* ***вижу по следам****: вот про шли другие связки, а здесь, стороной, шел Афанасий. [С. В. Обручев. В неизведанные края. Путешествия на Север 1917–1930 г.г. (1954)]*

(11) *Он рад был отдыху, остывал перед тем, как вновь* ***идти по солнцу****, и выпитая вода выходила из него потом. [Г. Я. Бакланов. Навеки девятнадцатилетние (1979)]*

Некоторые такие значения адекватно описываются в терминах семантических ролей, предложенных в работе [Ганенков 2002]. Так, в примере *(10)* предложная группа «по следам» обозначает ИСТОЧНИК ИНФОРМАЦИИ. Для значения же «идти по солнцу» (‘идти по пространству, освещенному солнцем’) нам не удалось найти удовлетворительной трактовки ни в инвентаре ролей Ю. С. Ганенкова, ни в наборе базовых локализаций и ориентаций, ни в расширенном списке локализаций и ориентаций, предлагаемом Ю. В. Мазуровой.

Приведем список непространственных значений предлогов «по» и «к», поддающихся описанию в терминах семантических ролей Д.С. Ганенкова. В скобках после каждого названия роли указывается базовый пример, приводимый этим автором в качестве ее иллюстрации. Знаком «?» отмечаются лексемы, отнесение которых к данной роли вызывает сомнение:

1. Предлог «по»

СТИМУЛ ЭМОЦИЙ (*Дети смеются **над стариком**.*): соскучиться по, тосковать по, затосковать по ;

МИШЕНЬ (*Не ори **на детей***!): стрелять по (?);

ИСТОЧНИК ИНФОРМАЦИИ (*Мальчик спросил **у мамы** …*): ориентироваться по (солнцу), определяться по (следам), находить по, найти по, выделить по, узнать по, определять по, определить по, найти по, распознавать по, разыскивать по (отпечатку), восстановить по, видеть по, догадаться по;

КООРДИНАТИВ (*Я все сделал **по закону***.): ложиться по (солнцу), креститься по (солнцу), жить по (солнцу);

ДИСТРИБУТИВ[7] (*разлить варенье **по банкам***): разноситься по, разнести по, разнестись по, разносить по, разметать по, рассыпаться по, рассыпать по, разбрасывать по, разбросать по, разбросаться по, раскидать по, рассеять по, рассеиваться по, рассеяться по, растягиваться по, разливать по, разлить по, рассаживать по, разлететься по, расползтись по, расшвыривать по, расточать по, разгонять по, разогнать по, размыкать по, рассылать по, растаскивать по, рассесться по, распадаться по (руслам), распустить по.

2. Предлог «к»

СТИМУЛ ЭМОЦИЙ (*Дети смеются **над стариком***): относиться к чему-л. как-л., захотеться к, хотеться к, хотеть к, привыкнуть к (?), привыкать к (?), притерпеться к (?), приспособиться к (?), ревновать к, испытывать чувства к;

МИШЕНЬ (*Не ори **на детей***!): подлизываться к (?), влечь к (?);

ЦЕЛЬ ДВИЖЕНИЯ (*Мать пошла **за водой***.): пробивая путь к (?), попроситься к (?), командировать к (?), влечь к (?), пробираться к (?), отмахнуть к (?);

АДРЕСАТ РЕЧЕВОГО АКТА (*Сын сказал **отцу** …*): обращаться к;

СТИМУЛ (*Я смотрю **на этого человека***.): присматриваться к, приглядеться к, принюхиваться к, присматриваться к;

ТЕРМИНАТИВ (*Я дошел **до магазина** и повернул обратно*.): расти от станции **к станции.**

Возможность отнесения ряда представленных выше контекстов к семантическим ролям вызывает вопросы. С нашей точки зрения, источником вопросов является способ определения ролей. Дело в том, что Д.С. Ганенков не дает строгих дефиниций. Для каждой роли он приводит несколько глаголов, в семантическую структуру которых она входит, а также несколько типичных примеров. Это объясняется задачей работы [Ганенков 2002], которая состоит в выявлении элементов универсального грамматического набора на материале большого числа языков различных семей и ареалов. Однако при попытках использовать данный набор ролей для описания значений языковых единиц и грамматических категорий вопрос о «сходстве» рассматриваемого примера и «прототипического» примера (достаточно ли глагол «обращаться к» похож на глагол «сказать»; возможно ли распространить семантическую роль мишени с глаголов речевого воздействия на ряд других глаголов), приводимого автором, приходится решать на основе интуиции исследователя. В связи с этим представляется перспективным построить для ролей Д.С. Ганенкова толкования на семантическом метаязыке, что, с нашей точки зрения, позволило бы ввести более четкие критерии трактовки языковых единиц как представителей ролей.

В ходе работы с Национальным корпусом русского языка нами также был получен ряд лексем, не поддающихся описанию в терминах избранного нами аппарата. К ним относятся: тяготеть к, приурочиваться к, переходить к (чему-л.), пригласить к, причислить к, уставиться к (солнцу), подходить к (по критерию), приучить к, приучать к, подпустить к, увеличиться по (скважине), отмечаться по (скважине), сводиться к, получить название по, называть по, приготовить к (сцене), подготовить к (сцене), привести к (чему-л). Заметим, что абсолютное большинство этих слов относятся к модусу письменной речи.

Кроме этого, была выделена группа контекстов типа «селиться по рекам» и «залечь по берлогам», плохо поддающаяся описанию в терминах локализаций и ориентаций. Однако, с нашей точки зрения, эту проблему можно решить путем привлечения понятий метафорического и метонимического переноса.

---

[7] Отметим, что Д. С. Ганенков дистрибутивного значения пространственных показателей не выделяет. С его точки зрения, такой тип употреблений «связан, скорее, не с указанием на семантическую роль участника, а с выражением количественных характеристик ситуации» ([Ганенков 2002, с. 124]). Представляется, что такое решение связано с нежеланием подменять количественное аспектуальное значение глагола значением пространственных единиц. Так, например, дистрибутивная интерпретация предложения «*Люди разбрелись по вагонам*» возникает в связи с семантикой глагола. Однако в русском языке, с нашей точки зрения, существует ряд случаев, в которых средством выражения количественных характеристик ситуации служит не глагол, а предлог. Речь идет о контекстах типа «дать X-ам Y» и «дать X-ам по Y-у» (*Дал лоцманам рубль* vs *Дал лоцманам по рублю*), представляющих собой минимальные пары с разным заполнением одной валентности. Учитывая это обстоятельство, мы выделяем дистрибутивное значение предлога «по». Сочетания глаголов, имеющих дистрибутивное значение, с группами вида «существительное+предлог «по»» мы трактуем как случаи двойного маркирования этого значения.

Подведём итоги. В данной работе представлено применение аппарата описания семантики пространственных показателей, разработанного Д. С. Ганенковым и Ю. В. Мазуровой преимущественно на основе бесписьменных языков и языков, получивших письменность в XX веке, к языку с давней письменной традицией — русскому. С помощью этого аппарата описываются основные значения предлогов «к» и «по». Отметим, что, с нашей точки зрения, несомненным плюсом избранной методики является единообразие описания семантики разных показателей, что позволяет легко приводить сравнение. Однако для создания полного описания необходимо доработать аппарат в направлении описания лексем, принадлежащих письменной речи, и ввести четкие толкования семантических ролей. Необходимо также использовать понятия метафорического и метонимического переноса.

Представленное в данной статье описание значений русских предлогов «по» и «к» служит как типологическим целям, так и целям создания полного описания семантики русского языка. Кроме того, списки глаголов и сочетающихся с ними предложно-субстантивных сочетаний, приведенные в работе, могут быть использованы в ходе морфологической и синтаксической разметки Национального корпуса русского языка.

Литература

1. *Ганенков Д. С.* Модели полисемии пространственных показателей. Дипломная работа. М.: 2002.
2. *Ганенков Д. С.* Контактные локализации в нахско-дагестанских языках и их типологические параллели. Диссертация на соискание ученой степени кандидата филологических наук. М.: 2005.
3. *Иомдин Л. Л.* Русский предлог ПО: этюд к лексикографическому портрету // Metody formalne w opisie jezykow slowianskich. Bialystok: 1990. Pp. 241–260.
4. *Мазурова Ю. В.* Типология средств выражения пространственной локализации (вертикальная ось). Диссертация на соискание ученой степени кандидата филологических наук. М.: 2007.
5. *Мельчук И. А.* Курс общей морфологии. Том II. Ч. 2. Вена: 1998. С. 47–60, 324–350.
6. *Строева М. Н.* Семантика «конкретных» падежей в уральских языках // Acta linguistica Petropolitana. С-Пб.: 2008. Том IV, часть 2. С. 174–180.
7. *Строева М. Н.* Семантика локативных падежей в уральских языках. Дипломная работа. М.: 2008.

Влияние места словесного ударения на распознавание слов в русской устной речи

Processing initial-stress and non-initial-stress words in spoken-word recognition in Russian

Фёдорова О. В. (olga.fedorova@msu.ru),
Шаврыгина А. С. (shavrygina@gmail.com)

Московский государственный университет им. М. В. Ломоносова

Рассматривается вопрос о принципиальной возможности использования в русском языке (как языке с сильной редукцией) нового экспериментального приема, предложенного в работе [Mattys, Samuel 2000] для проверки метрической стратегии сегментации устной речи. Результаты двух экспериментов дают положительный ответ на поставленный вопрос.

1. Введение. Проблема сегментации устной речи

Одним из важных вопросов в области восприятия и распознавания звучащей речи является вопрос о ее сегментации. Почему при восприятии речи на родном языке эта процедура деления речевого потока на отдельные слова в большинстве случаев происходит быстро и как бы сама собой, а при восприятии незнакомого нам языка превращается в большую проблему? Почему родная речь порой воспринимается нами как слишком медленная, с длинными и четкими паузами между словами, а иностранная часто сливается в один непрерывный поток, ведь в обоих случаях обязательные для письменной речи пробелы между словами при этом отсутствуют? Какие механизмы стоят за этой почти автоматической способностью сегментировать родную речь?

Подобные вопросы, связанные с сегментацией устной речи, исследуются как в области восприятия устной речи, так и в области ее распознавания[2]. Существующее разделение этих двух научных областей хотя и имеет под собой серьезные основания (традиционно, при изучении восприятия устной речи исследуется, как человек воспринимает и идентифицирует отдельные звуки языка, а при изучении распознавания устной речи нас в первую очередь интересует вопрос идентификации целых слов), но в некоторой степени все же искусственно: нельзя утверждать, что сначала мы распознаем все звуки, а потом складываем из них слова. Наоборот, знание конкретного слова помогает нам правильно распознать звуки, из которого это слово состоит; кроме того, часто нам удается распознать слово еще до того, как оно было произнесено до конца.

Рассмотрим теперь, как вопрос о сегментации решается разными авторами при разных подходах к восприятию и распознаванию устной речи, которые могут быть условно разделены на подходы с опорой на акустические характеристики звуков или на их статистические свойства (так называемые **до-лексические** подходы) и на **лексические** подходы, авторы которых предполагают, что при установлении границ между словами уже активно используется информация о самих этих словах.

[1] Работа выполнена при частичной финансовой поддержке Российского гуманитарного научного фонда в рамках проекта № 08-04-00165а.

[2] В отличие от английского термина **speech perception**, который в большинстве случаев переводится на русский язык как восприятие устной (или звучащей) речи, перевод на русский язык термина **spoken word recognition** имеет множество вариантов: устное опознание слова, опознавание слова со слуха, узнавание слова, идентификация слова, распознавание слова на слух, устное распознавание слова, распознавание звучащей речи, распознавание устной речи, распознавание слов устной речи и под. В настоящей работе мы в дальнейшем будем использовать русскоязычные термины **восприятие устной речи** и **распознавание (слов) устной речи**. В свою очередь, термины **word recognition** и **lexical access** не представляют особой трудности для перевода (**распознавание слов** и **доступ к слову**, соответственно), однако, их точное определение и взаимозаменяемость представляют определенные проблемы. Так, по мнению некоторых авторов доступ к слову происходит раньше распознавания слова, по мнениюдругих — наоборот, позже; во многих работах эти термины используются как синонимы. Мы в дальнейшей работе будем использовать только термин распознавание слова.

В свою очередь, лексические подходы представлены, в частности, когортной моделью Марслен-Вильсона [Marslen-Wilson 1973] (ее последняя версия [Gaskell, Marslen-Wilson 2002], однако, содержит некоторые элементы коннекционизма), а также несколькими коннекционистскими моделями, среди которых стоит выделить TRACE [McClelland, Elman 1986] и Shortlist [Norris 1994].

Остановимся немного более подробно на долексических теориях, связанных с дистрибутивными (или статистическими) характеристиками слов. Одна из основополагающих идей при этом подходе состоит в том, что при определении границ между словами слушающий активно опирается на просодические свойства слов, а именно, что каждый сильный слог является потенциальным началом нового слова. Согласно данной **стратегии метрической сегментации** речи [Cutler & Norris 1988] слова с ударением на первом слоге обрабатываются быстрее и легче, чем слова с ударением не на первом слоге. Этот факт был подтвержден уже достаточно большим количеством экспериментальных исследований, однако в работе [Mattys, Samuel 2000] авторам впервые удалось подготовить стимулы таким образом, чтобы использовать один и тот же акустический материал в различных экспериментальных условиях.

2. Экспериментальная методика из работы [Mattys, Samuel 2000]

Эксперимент, описанный в работе [Mattys, Samuel 2000], проводился по методике **обнаружения фонемы** (phoneme monitoring или phoneme detection [Connine, Titione 1996]): испытуемому называют определенный целевой звук (или показывают на экране соответствующую букву), который ему нужно будет распознать, после чего он слушает предложения или список несвязанных между собой слов и должен нажать на определенную клавишу, как только услышит целевой звук. Считается, что испытуемые выполняют подобные задания не на до-лексическом уровне обработки, а уже на лексическом. В качестве контроля выполнения задания именно на лексическом уровне эксперимент часто дополняется вспомогательным чисто лексическим заданием (в частности, в эксперименте, описанном в работе [Mattys, Samuel 2000], испытуемый должен был нажимать на специальную клавишу всякий раз, когда слышал слово, относящееся к категории орудий труда).

В каждой экспериментальной попытке испытуемый слышал последовательность из семи слов, произнесенных практически без пауз, и должен был нажать на клавишу, как только услышит целевой звук. Этот целевой звук всегда был начальным согласным слога и мог находиться в (i) начальном безударном слоге, (ii) начальном ударном слоге, (iii) втором безударном слоге и (iv) втором ударном слоге. Особенностью данного эксперимента было то, что слова с разным местом ударения в действительности представляли собой один и тот же акустический материал, то есть одни слова были вырезаны из других. Так, исследователи записывали пару слов-близнецов (test twins) *saga-zealous*, а потом получали из нее четыре комбинации: /g/ во втором безударном слоге слова *saga* (вариант iii), /g/ в начальном безударном слоге слова ***g**azelle* (вариант i), /z/ в начальном ударном слоге слова ***z**ealous* (вариант ii) и /z/ во втором ударном слоге слова *ga**z**elle* (вариант iv). Таким образом авторам работы [Mattys, Samuel 2000] удалось избежать проблем, связанных с вариативностью произносимых слов.

3. Эксперименты на русском материале

Вышеописанная стратегия метрической сегментации (а также более строгий её вариант, который гласит, что только слова с главным ударением на первом слоге воспринимаются как начала слов и благодаря этому требуют меньше времени и когнитивных усилий для своей обработки) проверялась по большей части на материале английского языка, где 90% всех слов имеют ударение на первом слоге (кроме английского тестировался также голландский язык [Vroomen, de Gelder 1995]). Таким образом, проверка данной гипотезы на русском материале не только добавит в исследовательскую копилку еще один язык, но и даст возможность верифицировать эту теорию на принципиально ином метрическом материале.

Однако проведение аналогичных экспериментов на русском материале[3] может осложниться сильной редукцией, характерной для русского языка по сравнению с английским. Слова, вырезанные подобным образом из русских слов-близнецов, могут вызвать у испытуемых серьезные проблемы с их распознаванием. Поэтому прежде чем приступать к проведению эксперимента, аналогичного описанному в [Mattys, Samuel 2000], нам предстояло провести большую подготовительную работу по проверке потенциальной возможности распознавания слов, вырезанных из слов-близнецов. Описанию двух таких экспериментов и будет посвящена следующая часть нашей работы.

[3] Фонетические, фонологические и просодические особенности русского языка, несомненно, очень хорошо изучены (свежий обзор психолингвистических работ см., например, в [Ягунова 2008]), однако работ по аналогичной [Mattys, Samuel 2000] методике «вырезания» слов, насколько нам известно, на материале русского языка еще не проводилось.

3.1. Эксперимент с использованием методики регистрации движений глаз

В качестве стимульного материала нами было составлено 24 пары слов-близнецов (например, *доми**нота**релка*, полный список см. в Приложении 1), 12 из вырезанных слов были хореями, 12 остальных — ямбами. Затем слова-близнецы были записаны и обработаны в звуковом редакторе Cool Edit Pro таким образом, чтобы между словами практически не было пауз, после чего были составлены и записаны (тоже без пауз на стыках слов) «семерки» слов (24 собственно экспериментальных, 12 отвлекающих (филлеров) и 4 тренировочные). Все слова, входящие в экспериментальные семерки, были среднечастотными, имели привычное звучание на стыках; семерки были записаны по такому принципу, что позиция целевого слова в них была сбалансирована от 2-го до 6-го места; две пары слов в семерке были записаны вместе (слова в скобках), а две пары слов отдельно, например: *картон — **нота** — (бамбук — карандаш) — школьник — (ковер — радость)*; на каждом из двух экспериментальных листов семерки с хореическим целевым словом чередовались с семерками с ямбическим целевым словом. В экспериментальных листах филлеры и экспериментальные семерки стояли на тех же самых позициях, но если на одном листе целевое слово было «вырезанным», то на другом листе на том же самом месте стояло целиком записанное целевое слово.

В процессе тестирования экспериментального материала мы провели несколько пилотных экспериментов, в которых задавали испытуемым вопрос: «Есть ли в данной последовательности слов что-то не очень естественно звучащее?». Большинство испытуемых указывали в этом случае на записанные вместе пары слов и на наши экспериментальные вырезанные слова. Таким образом, гипотеза, которую мы проверяем в ходе этого эксперимента звучит так: «Будет ли тот факт, что слова, вырезанные из двух слов-близнецов, звучат не совсем естественно, мешать их быстрому и правильному распознаванию по сравнению со словами, записанными целиком?».

Эксперимент был выполнен в виде презентаций в программе Microsoft PowerPoint. Инструкция испытуемым была следующей: «В каждой экспериментальной попытке Вы будете слышать последовательность из семи слов, начитанных с уменьшенными паузами между словами (что затрудняет их понимание); одновременно на экране Вы будете видеть восемь слов, только одно из которых будет совпадать с услышанными Вами словами. Ваша задача как можно быстрее кликнуть мышью по совпадающему слову. Через секунду после этого вокруг правильного слова (независимо от того, правильным или нет был Ваш ответ) будет появляться красная рамка».

Данный эксперимент был проведен с шестью испытуемыми по методике записи свободных движений глаз (free-viewing eye-tracking, оборудование ETL-500 фирмы ISCAN Inc., подробнее об этой методике см. [Федорова 2008]), при этом фиксировались момент воспроизведения ключевого слова, момент первого после конца воспроизведения ключевого слова взгляда на ключевое слово, момент нажатия на мышку, правильность выбора, а также высчитывались время с момента воспроизведения ключевого слова до первого взгляда на ключевое слово (наиболее важный параметр) и время с момента воспроизведения ключевого слова до нажатия мышки. Результаты (см. Приложение 2) свидетельствуют о том, что статистически значимое различие есть только во времени ответов на вырезанные и записанные целиком ямбические слова.

3.2. Эксперимент на воспроизведение слов

В нашем втором эксперименте были использованы те же 40 экспериментальных семерок, но каждая из них воспроизводилась два раза подряд, чтобы исключить пропуск целевого слова в том случае, если с первого раза испытуемые не смогут запомнить все слова. В инструкции говорилось, что эксперимент имеет целью определение объёма рабочей памяти в условиях плохой перцептивной различимости слов. После прослушивания стимульной семерки испытуемые должны были воспроизвести все слова, которые они смогли запомнить; сразу после первой попытки слова данной семерки воспроизводились второй раз, что помогало некоторым испытуемым повторить во второй попытке большее количество слов. Если в первом эксперименте мы смотрели, будет ли в случае вырезанных слов увеличиваться время, необходимое на поиск целевого слова, и количество неправильных ответов, то в этом эксперименте мы проверяли гипотезу «будет ли тот факт, что слова, вырезанные из двух слов-близнецов, звучат не совсем естественно, мешать их правильному воспроизведению по сравнению со словами, записанными целиком?».

Эксперимент, представленный в виде презентаций в программе Microsoft PowerPoint, был проведен с 40 испытуемыми. Результаты (см. Приложение 3) говорят о том, что (i) повторение в семерках-филлерах (в среднем 5 из 7) статистически значимо лучше, чем в экспериментальных словах (4,15 из 7); (ii) повторение в **целиком записанных словах** (4,3 из 7) статистически незначимо лучше, чем в вырезанных (4 из 7); (iii) повторение целевого слова в **целиком записанных хореях** статистически незначимо лучше, чем в вырезанных хореях; (iv) повторение целевого слова в **целиком записанных ямбах** статистически значимо лучше, чем в вырезанных ямбах; (v) повторение целевого слова в **хореях** статистически значимо лучше, чем в ямбах.

В последнем разделе настоящей работы мы сопоставим результаты двух экспериментов, подведем итоги и кратко наметим перспективы дальнейших исследований.

4. Обсуждение результатов

Итак, по результатам двух проведенных экспериментов можно сделать общий вывод, что в целом не совсем естественное звучание вырезанных слов не мешает ни их правильному распознаванию (эксперимент 1), ни их правильному воспроизведению (эксперимент 2). Таким образом, русская редукция не препятствует принципиальной возможности проведению эксперимента, аналогичного [Mattys, Samuel 2000], на русском материале.

Обратим, однако, внимание на тот факт, что и в первом и во втором экспериментах вырезанные ямбические слова и распознавались хуже (время реакции в первом эксперименте статистически значимо больше), и воспроизводились хуже (количество правильно воспроизведенных слов для вырезанных ямбов значимо меньше как по сравнению с вырезанными хореями, так и по сравнению с записанными целиком ямбами, в то время как для хореев это различие статистически незначимо). Более того, в целом все (т.е. и вырезанные, и записанные целиком) хореические слова во втором эксперименте воспроизводились значимо лучше, чем все ямбические слова. Данный факт, по нашему мнению, может свидетельствовать как о несовершенстве нашего ямбического стимульного материала, так и о том, что хореические слова русского языка понимаются (то есть распознаются и воспроизводятся) лучше, чем ямбические. Последняя гипотеза может сама по себе уже служить некоторым подтверждением использования не только англоязычными, но и русскоязычными носителями стратегии метрической сегментации.

Для того, чтобы развести эти две гипотезы, нам нужно будет провести еще один подготовительный эксперимент, аналогичный эксперименту 2, заменив в стимульном материале все слова, особенно ямбические, которые плохо распознавались и воспроизводились по результатам уже проведенных экспериментов (см. Приложение 4). Если при замене наиболее неудачных слов (как то: ямбические слова *топор*, *коза*, *комар*, *роса*, *нога* и *коса*; хореические *палец* и *ветка*), мы получим аналогичные результаты, это будет уже с большей вероятностью свидетельствовать в пользу использования в русском языке стратегии метрической сегментации. Однако более строго доказать это утверждение можно будет только после проведения основного эксперимента, подобного описанному в работе [Mattys, Samuel 2000], на русском материале.

Приложение 1. Список экспериментальных слов-близнецов

№	хорей	целевое слово	звуки	№	ямб	целевое слово	звуки
1	домино-тарелка	нота	н-т	1	слово-дама	вода	в-д
2	письмо-река	море	м-р	2	мясо-ваза	сова	с-в
3	такси-лопата	сила	с-л	3	полка-маршал	комар	к-м
4	молоко-жара	кожа	к-ж	4	конфета-порция	топор	т-п
5	тропа -лицо	палец	п-л	5	палка-занавес	коза	к-з
6	ответ-карман	ветка	в-к	6	ворона-галстук	нога	н-г
7	слеза-поход	запах	з-п	7	долина-радуга	нора	н-р
8	попугай-конверт	гайка	г-к	8	телега-радио	гора	г-р
9	метро-залив	роза	р-з	9	утро-сабля	роса	р-с
10	толпа-русалка	парус	п-р	10	берёза-водка	завод	з-в
11	борода-частушка	дача	д-ч	11	колено-сокол	носок	н-с
12	кошмар-лягушка	марля	м-л	12	точка-сахар	коса	к-с

Приложение 2. Первый эксперимент

Первая цифра — целиком записанные слова, вторая через запятую — вырезанные.

	правильные ответы (из 36)	время
хорей	33, 29	1.302, 1.447
ямб	32, 26	1.327, 2.191

Приложение 3. Второй эксперимент

Первая цифра — целиком записанные слова, вторая через запятую — вырезанные.

	кол-во слов (из 1680)		кол-во целевых (из 240)	
	1 попытка	2 попытка	1 попытка	2 попытка
хорей	1067, 1028	1294, 1257	132, 114	180, 159
ямб	1007, 936	1266, 1139	101, 42	155, 72
филлер	2419 (из 3360)	2814 (из 3360)		

Приложение 4. Стимульные слова

Первая цифра — целиком записанные слова, вторая через запятую — вырезанные).

Слова, выделенные жирным шрифтом, распознавались плохо; слова, выделенные курсивом, распознавались на среднем уровне.

№	слово	тип ударения	первый эксперимент, ошибки	первый эксперимент, время	второй эксперимент, целевое слово
1	нота	хорей	0, 1	долго, долго	9, 12
2	сова	ямб	0, 1	быстро, быстро	17, 12
3	море	хорей	0, 0	быстро, быстро	13, 12
4	топор	ямб	1, 1	быстро, быстро	6, 3
5	сила	хорей	0, 1	быстро, быстро	4, 11
6	коза	ямб	1, 0	быстро, быстро	6, 2
7	кожа	хорей	0, 1	быстро, быстро	9, 5
8	вода	ямб	0, 0	быстро, долго	8, 1
9	палец	хорей	1, 0	быстро, долго	1, 5
10	комар	**ямб**	0, 2	быстро, быстро	5, 1
11	**ветка**	**хорей**	0, 2	быстро, быстро	17, 0
12	*нора*	*ямб*	0, 1	быстро, быстро	10, 2
13	запах	хорей	0, 0	быстро, быстро	9, 11
14	**роса**	**ямб**	1, 1	быстро, быстро	5, 0
15	гайка	хорей	0, 1	быстро, быстро	15, 11
16	*завод*	*ямб*	0, 0	быстро, быстро	10, 2
17	роза	хорей	1, 0	долго, быстро	13, 14
18	**нога**	**ямб**	0, 1	быстро, быстро	7, 1
19	парус	хорей	0, 0	быстро, быстро	13, 20
20	*гора*	*ямб*	0, 1	долго, быстро	8, 3
21	дача	хорей	0, 0	быстро, быстро	17, 8
22	**коса**	**ямб**	1, 1	быстро, быстро	8, 1
23	*марля*	*хорей*	0, 1	быстро, долго	12, 5
24	носок	ямб	0, 1	быстро, быстро	11, 14

Литература

1. *Фёдорова О. В.* Методика регистрации движений глаз «Визуальный мир»: шанс для сближения психолингвистических традиций // Вопросы Языкознания, 6. 2008. С. 98–120.
2. *Ягунова Е. В.* Вариативность стратегий восприятия звучащего текста. Пермь, 2008.
3. *Connine C. M., & Titone, D.* Phoneme monitoring // Language and Cognitive Processes, 11(6), 1996. P. 635–646.
4. *Cutler A. & Norris D.* The role of strong syllables in segmentation for lexical access // Journal of Experimental Psychology: Human Perception and Performance, 14, 1988. P. 113–121.
5. *Gaskell M. G. & W. D. Marslen-Wilson*. Representation and competition in the perception of spoken words // Cognitive Psychology, 45, 2002.
6. *Marslen-Wilson W. D.* Linguistic structure and speech shadowing at very short latencies // Nature, 244, 1973.
7. *Mattys S. & Samuel A.* Implications of stress pattern differences in spoken word recognition // Journal of Memory & Language, 42, 2000. P. 571–596.
8. *McClelland J. L. & Elman J. L.* The TRACE model of speech perception // Cognitive Psychology, 18, 1986.
9. *Norris D.* Shortlist: a connectionist model of continuous speech recognition // Cognition, 52, 1994. P. 189–234.
10. *Vroomen J. & de Gelder B.* Metrical segmentation and lexical inhibition in spoken-word recognition // Journal of Experimental Psychology: Human Perception and Performance, 21, 1995. P. 98–108.

Восточноармянский национальный корпус www.eanc.net

Eastern Armenian national corpus www.eanc.net

Хуршудян В. Г. (vk@corpustechnologies.com),
Даниэль М. А. (misha.daniel@gmail.com),
Левонян Д. В. (dl@renovacapital.com),
Плунгян В. А. (plungian@gmail.com),
Поляков А. Е. (pollex@mail.ru), **Рубаков С. В.** (rubakov@gmail.com)

Corpus Technologies, Москва

Восточноармянский национальный корпус (ВАНК) — это лингвистическая информационно-поисковая система, основанная на обширной коллекции текстов (около 110 млн.) на восточноармянском языке, покрывающая период с середины 19-го века до наших дней и снабженная мощной и гибкой поисковой функциональностью. ВАНК находится в открытом доступе в интернете (www.eanc.net).

ВАНК является репрезентативным, сбалансированным и полным электронным корпусом современного восточноармянского языка. Проект ВАНК был запущен в январе 2006 г. по инициативе группы московских исследователей и компании CorpusTechnologies. Летом 2007 г. был открыт портал www.eanc.net, на котором был размещен первый релиз корпуса. Второй релиз был размещен на том же портале весной 2008 г., а третий релиз — в марте 2009 г.

Третий релиз отличается от предыдущих объемом поискового корпуса (около 110 млн. словоупотреблений вместо 90 млн. во втором и 60 млн. в первом релизах соответственно), некоторыми функциональными расширениями, а также добавлением возможности просмотра статистических данных, связанных с употреблением и распределением определенной словоформы в корпусе.

Интернет-корпуса (www.sd-editions.com/LALT/home.html, Лейден и http://titus.uni-frankfurt.de/indexe.htm, Франкфурт-на-Майне) и электронные библиотеки (www.digilib.am, Ереван) существуют для древнеармянского языка; Анаид Донабедян (INALCO, Париж) занимается разработкой корпуса западноармянского языка [Donabédian, Boyacioglu 2007]. Кроме того, имеются электронные библиотеки на современном восточноармянском языке (например, www.armenianhouse.org, www.hayeren.hayastan.com, www.artgrak.am, www.people.cornell.edu, www.arlis.am, www.brusov.am и т.д.). Все эти ресурсы использовались при создании ВАНК; они составляют около 5% художественных и нехудожественных текстов корпуса. Значительный объем восточноармянских публицистических текстов содержится в архивах интернет-изданий www.azg.am, www.aravot.am, www.yerkir.am, www.iravunk.com и т. д.; эти ресурсы легли в основу подкорпуса современной прессы. Попыток создания электронных корпусов восточноармянского языка ранее не предпринималось.

1. Состав корпуса ВАНК

В ВАНК вошли тексты разных жанров, в том числе проза, поэзия, официальные, научные, религиозные, публицистические тексты, а также устная речь современного Еревана. Функциональность выбора подкорпуса (см. ниже) позволяет ограничивать поиск отдельными жанрами и группами жанров, которые интересуют пользователя в данный момент. В целом, ВАНК проектировался таким образом, чтобы максимально полно отразить лингвистическое разнообразие современного восточноармянского языка (см. Приложение 1. Состав ВАНК (на март 2009 г.)).

Важнейшим аспектом корпусной лингвистики являются микроисторические исследования, ориентированные на «быстрые» языковые изменения. Для таких исследований корпус должен иметь временную координату, по которой можно отслеживать изменения значений лексемы или граммемы, отмирание старых и появление новых конструкций. ВАНК покрывает весь новый период истории армянского письменного языка с самого начала аш-

харабара (Хачатур Абовян «Раны Армении» 1841 г.). Временные характеристики текстов также можно использовать при выборе подкорпуса — например, работать только с текстами двадцатого или второй половины двадцатого века.

В идеале каждый жанр должен быть относительно равномерно распределен по годам — или, если это не так, существующая неравномерность должна отражать культурную ситуацию данного периода. В качестве примера препятствующих такой временной сбалансированности технических ограничений можно привести распределение прессы ВАНК по годам. Значительная часть прессы относится к постсоветскому периоду за счет широкой представленности текстов открытых периодических интернет-изданий (около 35 млн. словоупотреблений), доступный объем которых практически неограничен. Эта проблема была отчасти преодолена во втором релизе после осуществления совместного с Национальной библиотекой Армении проекта, в рамках которого были отсканированы, распознаны и включены в корпус избранные выпуски 60 периодических изданий общим объемом более 12 млн. словоупотреблений. Таким образом, архив периодики ВАНК покрывает всю историю существования армянской прессы, начиная от 70-х годов 19-го века по поздний советский период. Тем не менее, временной баланс прессы остается неидеальным.

Важнейшим жанром письменных текстов является электронная коммуникация: электронная почта, смс, instant messengers, блоги. Тексты этого типа в ВАНК только начинают подключаться: в третьем релизе будет содержаться небольшой корпус блогов. Здесь основное затруднение заключается в том, что до самого последнего времени в текстах такого типа использовались в основном разного рода косвенные и мало стандартизованные способы передачи армянской письменности.

Специального обсуждения заслуживает проблема устного корпуса. Иногда возникает вопрос, зачем устные тексты вообще были включены в ВАНК. Претензии, которые предъявляются к устному корпусу (не только в ВАНК, но и, например, Национальному корпусу русского языка) — это отклонение от языковой нормы, массовое использование английских и русских слов, нарушения корректных синтаксических структур. Те же доводы часто приводятся против включения в корпус текстов электронной коммуникации. Все эти претензии, на самом деле, апеллируют не к недостаткам, а к языковым особенностям разговорной речи, которая имеет собственную, иногда значительно отличную от литературной норму, широко использует переключение кода, обладает собственным синтаксисом и т.п. Именно для исследования этих особенностей устной речи и формируются подобные корпуса. Эти исследования относятся не только к лингвистике, но и к смежным областям — социолингвистике (переключение кода), психолингвистике (особенности построения высказывания). Нарушения литературной нормы, особенно если они носят системный характер, должны использоваться при языковом планировании и разработке языковых реформ: такие реформы, которые противонаправлены вектору развития устной речи, обречены на провал. Иными словами, устная речь не является неправильной, не нормативной письменной — она просто другая, иная языковая субстанция.

Отсюда вытекает ответ и на другой вопрос, связанный с сбалансированностью корпуса — как определить правильное соотношение между количеством письменных и устных текстов? Теперь ясно, что это вопрос бессодержательный. Письменный и устный подкорпуса могут находится в произвольном количественном отношении между собой, так как по сути это два разных способа существования языка, два разных корпуса, одновременный поиск по которым имеет ограниченную научную ценность.

С весны 2008 г. на сайте открыт раздел электронной библиотеки, содержащий полные тексты более 100 произведений классической армянской литературы. От других электронных библиотек библиотека ВАНК отличается наличием лексико-морфологического анализа словоформы для всех разбираемых словоформ (более 90% словоупотреблений), для большей части из которых даются также английские переводные эквиваленты (более 85% словоупотреблений).

2. Грамматический словарь

В основу грамматического словаря ВАНК положен словник объемом около 80 тыс. слов. Этот словник является компиляцией многих источников — в первую очередь словаря Е. Г. Галстян [Галстян 1985] и части словаря Э. Б. Агаяна [Агаян 1976], но также словаря аббревиатур Д. С. Гюрджиняна и Н. А. Экекян [Гюрджинян, Экекян 2007], словаря географических названий А. Гргеаряна и Н. Арутюнян [Гргеарян, Арутюнян 1987–1989], различных имен собственных и пр.

Составление грамматического словаря — трудо- и времяемкая работа, которая ранее на армянском материале, насколько нам известно, не проводилась; проект, который решал небольшую часть этой задачи — словарь форм множественного числа [Гюрджинян 2005]. Для сравнения скажем, что первый и достаточно полный грамматический словарь русского языка, содержащий свыше ста тысяч лексем, появился уже более тридцати лет назад [Зализняк 1977].

Несмотря на богатую традицию грамматического описания армянского языка, в готовом виде получить из какого-либо источника классификацию

именных или глагольных парадигм, пригодную для автоматического анализа текстов, невозможно. Для построения алгоритма лемматизации армянских именных словоформ мы выделили более 50 формальных типов именного словоизменения. Полный список *всех* различных парадигматических типов приведен на сайте проекта в разделе «Разметка».

3. Поисковая функциональность ВАНК

ВАНК представляет гибкую функциональность для лингвистического поиска, ориентированную в первую очередь на лексические и грамматические запросы. Синтаксические запросы возможны лишь опосредовано, так как корпус не имеет синтаксической разметки. По поисковой функциональности ВАНК очень близок Национальному корпусу русского языка, который до определенной степени служил его прототипом.

1. ***Поиск словоформы или лексемы.*** ВАНК позволяет искать как вхождения конкретной словоформы (например, մարդու *mardu человек.gen*), так и вхождения всех словоформ определенной лексемы (например, словоформ մարդ *mard человек.nom*, մարդու *mardu человек.gen*, մարդիկ *mardik человек.pl.nom* и т.д. от лексемы մարդ *mard человек.nom*).
2. Вхождения лексем можно искать по их ***английским переводным эквивалентам.***
3. ***Поиск по грамматическим признакам***. ВАНК позволяет искать все словоформы, обладающие определенной грамматической характеристикой или набором грамматических характеристик (например, имперфективный конверб в пассиве). Грамматические признаки можно искать как вне зависимости от того, в какой лексеме они встретились, так и вместе с лексемой. При поиске можно учитывать словоизменительный тип словоформы. Грамматический запрос может:
    a. быть определен как логическая конъюнкция или дизъюнкция нескольких категорий, или
    d. совмещать конъюнкцию и дизъюнкцию в одной логической формуле; собственно, именно последний тип грамматического запроса является наиболее частотным и естественным.

Кроме этих, собственно лингвистических параметров поиска, можно использовать дополнительные графематические и иные параметры, иногда позволяющие эффективно сузить запрос. Так, можно искать только словоупотребления в начале или в конце предложения, накладывать определенные ограничения на регистр (написание с первой прописной или со всеми прописными), указывать знаки препинания слева и справа от вхождения и т.п. ВАНК позволяет искать контексты, в которые одновременно входит несколько поисковых элементов. Расстояние между вхождениями можно изменять, меняя интервал допустимых расстояний.

Сравнивая поисковую функциональность ВАНК с поисковой функциональностью его ближайшего аналога, Русского национального корпуса, можно отметить следующие отличия. ВАНК менее гибок в смысле отрицания словоформ и лексем, но зато представляет возможности отрицания граммем. Он также представляет более гибкие механизмы поиска с учетом регистра и пунктуации, позволяет искать вхождения, находящиеся в начале, в конце или не в начале и не в конце предложения, а также только такие вхождения, которые не имеют омонимичных разборов, причем в ВАНК 3.0 различается внутрилексемная (грамматическая) и межлексемная (лексическая) омонимии. Отсев вхождений с омонимичными разборами в некоторых случаях позволяет сократить количество поискового шума.

Любой запрос, который может быть применен к ВАНК, может быть также применен и определенному пользователем подкорпусу ВАНК. Окно подкорпуса состоит из следующих трех основных зон: авторы и произведения, период, жанр, и трех дополнительных: проза/поэзия, оригинальные / переводные тексты, детская / общая литература.

4. Отображение результатов ВАНК

ВАНК позволяет осуществлять сортировку контекстов по целому ряду параметров: начальная форма словоформы-вхождения (лексема), словоформа-вхождение, словоформа слева от словоформы-вхождения, автор, название, год создания (как по возрастанию, так и по убыванию), жанр. При этом ВАНК поддерживает четыре формата отображения найденной информации:

1. *полный* (по умолчанию): каждый контекст сопровождается базовыми библиографическими сведениями (автор, название, год создания);
2. *краткий*: библиографические сведения приводятся только в окне расширенного контекста;
3. *KWIC (Key Words In Context)*: принятый в корпусных интернет-ресурсах способ отображения контекстов таким образом, чтобы они были визуально выровнены друг относительно друга по вхождению. Формат KWIC используется обычно вместе с сортировкой по словоформе или левой словоформе (см. Приложение 2).
4. *глоссированный*: этот формат предназначен в первую очередь для лингвистов-типологов и изучающих армянский язык. Отображение текста близко к так называемому морфологическому глоссированию (interlinear morphological

glossing), используемому в типологических публикациях и описаниях малых языков, но без разбиения на морфемы и поморфемного перевода. Для всех словоформ, за исключением словоформ, которые не разбираются парсером ВАНК, на экран в виде столбца, расположенного непосредственно под лексемой, выводится лексико-грамматический анализ, который в других типах выдачи доступен только при наведении мыши. В первой строчке столбца содержатся исходная форма и лексические признаки (например, частеречная характеристика). Во второй строке в фигурных скобках приводятся грамматические (словоизменительные) признаки словоформы (за исключением неизменяемых лексем). Если лексеме приписан перевод, он дается в третьей строчке. Если у словоформы существует несколько разборов, они отделяются друг от друга светло-серой чертой (см. Приложение 3).

Отображение результатов возможно как в армянском алфавите, так и в транслитерации (см. Приложение 4). Используемая в ВАНК транслитерация в основном следует международной арменоведческой традиции Хюбшманна-Мейе, адаптированной под Unicode. Транслитерация используется в том числе при отображении имен авторов и названий произведений.

Каждый контекст представлен в окне выдачи одним предложением (исключением является поиск, при котором областью поиска является документ); слова-вхождения при этом выделены оранжевым цветом. При каждом контексте приводятся базовые библиографические характеристики (если они известны) — автор, название, год создания, для прессы также номер или дата выпуска. ВАНК позволяет расширить контекст найденного предложения. По умолчанию на экран выводятся три предложения — то предложение, в котором обнаружены искомые вхождения, а также одно предложение до него и одно предложение после него. Расширяя контекст, можно увеличивать размер контекста вплоть до девяти предложений (четыре предложения до и четыре предложения после того предложения, в котором обнаружено вхождение).

При каждом запросе в верхней части экрана отображается общая информация о запросе и полученных результатах (см. Приложение 2):

- число вхождений (в случае контекстного запроса с числом контекстов более 10,000 — примерная оценка их общего числа в корпусе),
- число документов (в случае контекстного запроса с числом контекстов более 10,000 — примерная оценка числа документов, в которых они могут встретиться),
- критерии сортировки (если они выбраны пользователем),
- размер подкорпуса, по которому осуществлялся поиск (в процентах от общего числа словоупотреблений в корпусе).

Кроме того, в третьем релизе ВАНК добавлена новая функциональность — интерфейс, с помощью которого пользователь может получить не только общую информацию о количестве вхождений словоформы в корпусе, но и подробную картину ее распределения по основным жанрам и декадам (см. Приложение 6. Статистическое употребление словоформы աստծո *astco* (*бог.gen)* по данным ВАНК на март 2009 г.). Кроме абсолютного числа употреблений словоформы, на сайте приводятся еще две характеристики: WPM (число вхождений на миллион), которая позволяет получить представление о частотности словоформы с учетом объема корпуса, а также ее ранг (логарифм отношения частотности самой частотной словоформы к частотности данной словоформы), которая показывает, насколько данная словоформа менее частотна, чем самая частая словоформа данного сегмента. Для краткости приводится только таблица значений показателя WPM, значения которого легче всего интерпретировать.

5. Замечания о программном обеспечении ВАНК

Программное обеспечение для проекта ВАНК разрабатывается и поддерживается компанией Corpus Technologies. Оно создавалось с учетом перспективы масштабирования корпуса, а в конечном итоге — с целью создания языково-независимой программной платформы для корпусных исследований.

Система спроектирована таким образом, чтобы сделать возможным индексирование корпусов разноструктурных языков; проиндексированный системой корпус обеспечивает эффективную обработку запросов разной степени сложности. Единственным, но необходимым требованием является следование разработанному Corpus Technologies стандарту разметки текстов. Только парсер ВАНК и пользовательский интерфейс жестко ориентированы на армянскую грамматику, все остальные структурные элементы системы могут работать практически с любым морфологическим типом языка и алфавитом и разметками разной степени детальности и глубины.

Отметим также алгоритм рандомизации, реализованный далеко не во всех современных крупных корпусах. На настоящий момент пользовательская выдача имеет ограничение 10 тыс. контекстов на запрос. Если число удовлетворяющих запросу контекстов превышает этот лимит (например, при поиске частотной лексемы или отдельного грамматического признака), поисковая система ВАНК использует специальную процедуру, позволяющую избежать нежелательной «конденсации» найденных контекстов в определенной части корпуса и работать с квазирепрезентативной выборкой примеров, более или менее равномерно покрывающей весь корпус.

6. Целевая аудитория ВАНК

Аудиторией проекта является в первую очередь сообщество арменистов, работающих с лексикой и грамматикой ашхарабара, а также специалисты по западноармянскому языку или грабару, исследования которых носит сравнительный характер. С точки зрения представленности различных форм и временных срезов, ВАНК покрывает значительную часть языкового материала и ограничен почти только рамками логически невозможных (несовременные устные тексты) или крайне труднодоступных (жанр частной переписки) типов текстов. Важно подчеркнуть, что корпусом могут пользоваться исследователи, не владеющие или не вполне владеющие армянским языком — арменисты, которые только начинают изучать армянский язык, а также лингвисты-типологи, вообще не специализирующиеся в армянской филологии. Кроме возможности ввода запросов в латинской транслитерации (виртуальная клавиатура) и переключения в латинскую транслитерацию отображения армянской графики, во втором релизе в подсветку грамматического разбора (появляющуюся при наведении мыши на словоформу) включен краткий список английских переводных эквивалентов и псевдоглоссированная выдача (см. Приложение 3).

Благодаря включению в разметку английских переводов пользование корпуса значительно упростилось и для изучающих армянский язык, как лингвистов, так и нелингвистов. Студент-лингвист может проводить собственные микроисследования, пользуясь корпусом точно так же, как и опытный исследователь-арменист, но при необходимости обращаясь к грамматическим разборам и переводам незнакомых форм.

Важной частью целевой аудитории для корпуса, как мы надеемся, могут стать преподаватели армянского языка как иностранного и школьные и вузовские преподаватели армянского языка как родного. Использование корпусов в преподавании — вполне активная, а количественно — чуть ли не доминирующая сфера использования языковых корпусов (см. [Добрушина 2005, 2008]), поскольку корпус позволяет работать с живым языковым материалом и отойти от традиционных методов обучения, опирающихся на закрытый и ограниченный объем признанной литературной классики.

Здесь естественно также упомянуть о той сфере использования корпусов, которая лежит в "серой» зоне между филологами и преподавателями языка — нормативной лингвистике. Как таковая эта отрасль не принадлежит ни к какому направлению академического лингвистического исследования и относится скорее к общественно-политической, чем научной сфере. Еще раз подчеркивая, что корпус ни в коем случае не является образцом нормы, следует отметить, что именно корпус, представляя действительный языковой узус, может и должен становиться основой для работы над нормой. Именно в корпусе видны тенденции языкового развития, изменения узуса, на которые должно ориентироваться языковое планирование. Языковые реформы, которые оторваны от живых языковых процессов, обречены на фиаско, а если таковые реформы принимаются, то в конечном итоге они будут сметены стихией живого языка. В здоровом социуме лингвистический произвол и "вкусовщина» при языковом реформировании невозможны, так как языковой процесс не поддается законодательному регулированию. И здесь важную роль может сыграть как сам ВАНК, фиксирующий языковые сдвиги на протяжении более чем полутора веков, так и устный корпус ВАНК, демонстрирующий живые языковые процессы и обладающий значительным (по сравнению с устными корпусами многих других языков мира) объемом около 3,5 млн. словоупотреблений.

Для представителей других специальностей — историков, социологов, культурологов и др. — корпус может представлять интерес лишь постольку, поскольку они в своих исследованиях обращаются к языковому материалу (что происходит относительно редко). Речь идет о том, как социальные факторы или исторические процессы отражаются в языке, то есть о своего рода исторической социолингвистике. Собственно социолингвисты в основном работают с современным состоянием языка и составляют собственные микрокорпуса, ориентированные на частные задачи. А вот на частные вопросы об узусе того или иного социального значимого концепта, о том, когда он впервые упоминается в письменных текстах, как распространяется, как отмирает, как меняется его наполнение, ВАНК сможет ответить достаточно однозначно. Здесь гибкая грамматическая и контекстная функциональность поиска оказывается излишней (достаточно поиска по лексемам), зато на первый план выступает репрезентативность корпуса и особенно большой объем прессы, для ВАНК — включение во второй релиз значительного архива армянской периодики (около 47 млн. словоупотреблений), покрывающий весь период ее существования.

Наконец, часть потенциальной аудитории корпусов составляют люди, для которых обращение к корпусу вызвано не профессиональной потребностью, а личным интересом к языковому материалу. Языковая рефлексия, рассуждения об узусе, о значении тех или иных слов, как нам кажется, характерны для интеллигентного человека вообще. Поэтому при разработке интерфейса ВАНК мы пытались сделать его функциональность по возможности интуитивной и прозрачной, а разъяснения того, как искать лингвистическую информацию, максимально неспециальными и свободными от лингвистической терминологии. Пользователь корпуса — нелингвист может искать редкие слова, в значении кото-

рых он сомневается, формы, которые кажутся ему неправильными, но которые он встретил в живой речи или в тексте или наоборот, запрещаемые нормой формы, которые кажутся ему естественными или допустимыми.

7. Перспективы проекта

По полноте и репрезентативности литературного языка ВАНК приблизился к некоторому качественному порогу, преодолеть который не только трудно, но и не необходимо. Возможное осмысленное развитие проекта — включение принципиально новых текстов на армянском языке в широком смысле этого слова — литературных западноармянских, диалектных, средне- и древнеармянских текстов.

Технически было бы важно оптимизировать некоторые типы запросов: например, запросы с отрицанием, обработка которых на настоящий момент занимает значительное время. Высокий уровень грамматической омонимии (17% словоупотреблений), как внутри-, так и межлексемной, характерная для армянской грамматики, позволяет говорить о полезности, если не необходимости, работы по снятию омонимии или хотя бы создания подкорпуса со снятой омонимией (ср. корпус с вручную снятой омонимией в НКРЯ).

Создание синтаксической модели и внесение в корпус синтаксической разметки могло бы резко увеличить число академических областей применимости корпуса. Однако такая работа, в первую очередь автоматический синтаксический парсинг, требует огромной теоретической разработки и не может быть осуществлена в обозримом будущем.

На настоящем этапе исследовательская группа ВАНК приступает к корпусно ориентированным исследованиям армянской грамматики, первым из которых стала разработка словаря глаголов, снабженных полной словоизменительной и словообразовательной информацией. В рамках этого же направления компания CorpusTechnologies, технически и финансово поддерживавшая разработку ВАНК, открыла программу корпусных исследований в арменистике, подробное описание которой размещено на сайте корпуса.

Литература

1. *Агаян Э. Б.* Արդի հայերենի բացատրական բառարան [Толковый словарь современного армянского языка]. Т. 1–2. Ереван: 1976.
2. *Галстян Е. Г.* (ред.). Հայ-ռուսերեն բառարան [Армяно-русский словарь]. Ереван: 1985.
3. *Гргеарян А. К., Арутюнян Н. М.* Աշխարհագրական անունների բառարան [Словарь географических названий]. Ереван: 1987–1989.
4. *Гюрджинян Д. С.* Անուն խոսքի մասերի թվի կարգը արդի հայերենում. Քերականական բառարան-տեղեկատու [Категория числа имен в современном армянском. Словарь-справочник]. Ереван: 2005.
5. *Гюрджинян Д. С., Экекян Н. А.* Հայերենում գործածվող տառային հապավումների բառարան [Словарь. Инициальные аббревиатуры в армянском языке]. Ереван: 2007.
6. *Добрушина Н. Р.* Как использовать Национальный корпус русского языка в образовании? // Национальный корпус русского языка: 2003 — 2005. Результаты и перспективы. М.: 2005. С. 308–330.
7. *Добрушина Н. Р.* (ред.) Национальный корпус русского языка и проблемы гуманитарного образования. Теис: 2007.
8. *Зализняк А. А.* Грамматический словарь русского языка. Словоизменение. М. 1977.
9. *Хуршудян В. Г., Подлесская В. И.* Армянское ban как дискурсивный маркер речевого сбоя // Армянский гуманитарный вестник № 1. Ереван: 2006. С. 21–42.
10. *Хуршудян В. Г.* Средства выражения хезитации в устном армянском дискурсе в типологической перспективе. Дис. … канд. филол. наук. М.: РГГУ. 2006.
11. *Donabédian, A., Boyacioglu, N.* «La lemmatisation de l'arménien occidental avec Nooj» // S. Koeva, D. Maurel, M. Silberztein (eds.) Formaliser les langues avec l'ordinateur, de INTEX à NooJ. Presses Universitaires de Franche-Comté. 2007. P. 55–75.
12. *www.aravot.am* — ежедневная газета, Ереван.
13. *www.arlis.am* — Информационная система армянского законодательства (Armenian Legal Information System (ARLIS).
14. *www.armenianhouse.org* — электронная библиотека.
15. *www.artgrak.am* — иностранная литература на армянском языке.
16. *www.azg.am* — ежедневная газета, Ереван.
17. *www.brusov.am* — Ереванский гос. лингвистический университет им В. Брюсова.
18. *www.digilib.am* — электронная библиотека древнеармянских текстов.
19. *www.eanc.net* — Восточноармянский национальный корпус.

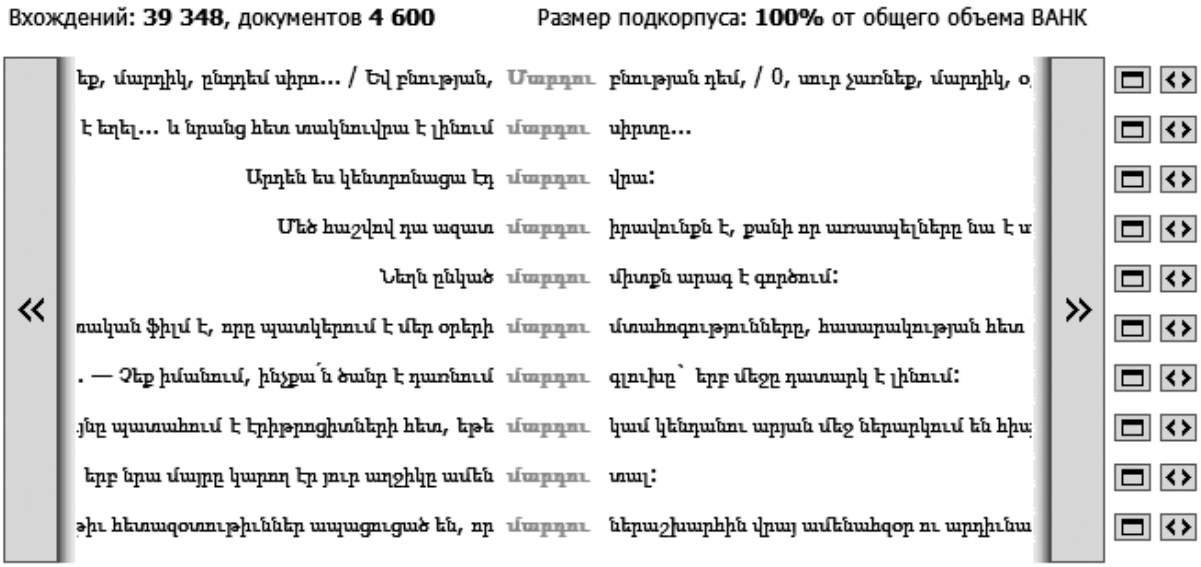
20. *http://forum.am-kayq.com* — армянский форум.
21. *www.hayeren.hayastan.com* — армянский образовательный портал.
22. *www.iravunk.com* — ежедневная газета, Ереван.
23. *www.people.cornell.edu* — текст Библии на восточноармянском.
24. *www.ruscorpora.ru* — Национальный корпус русского языка.
25. *www.sd-editions.com/LALT/home.html* — Leiden Armenian Lexical Textbase.
26. *http://titus.uni-frankfurt.de/indexe.htm* — Thesaurus Indogermanischer Text- und Sprachmaterialien.
27. *www.yerkir.am* — еженедельная газета, Ереван.

Приложение 1. Состав ВАНК (на март 2009 г.)

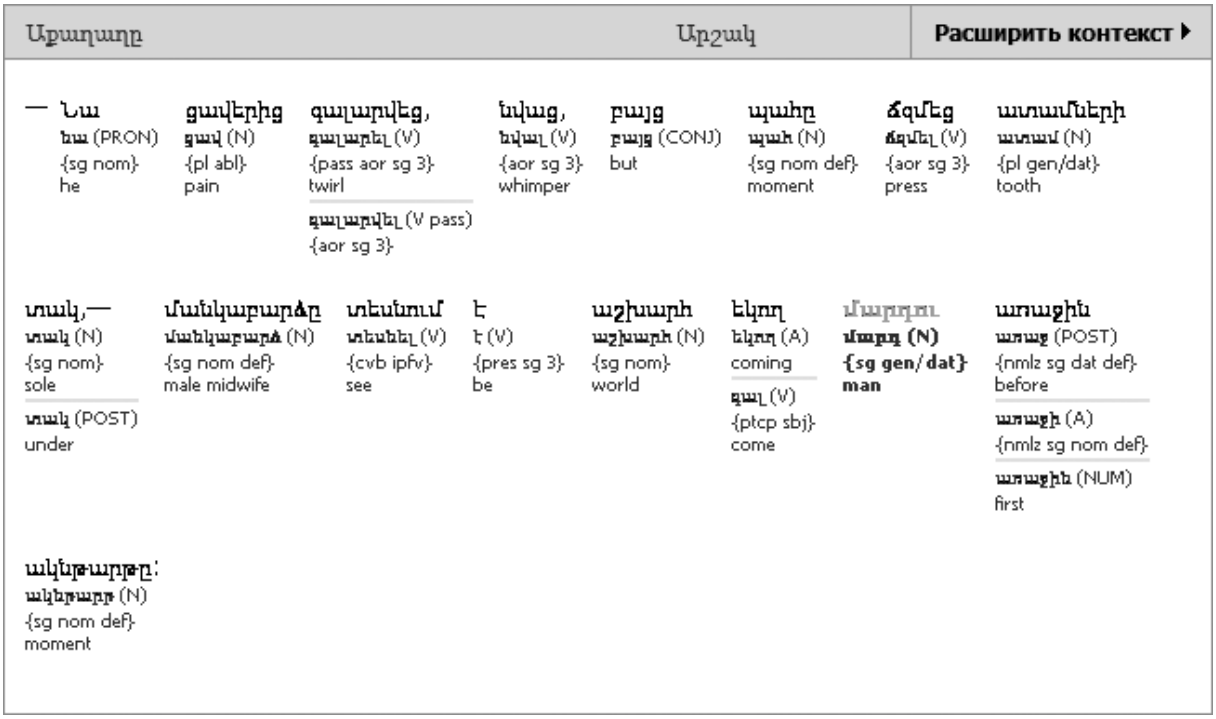
	Письменные тексты	словоупотребления	доля в ВАНК	документы	
	Художественная литература				
	проза: романы	29,909,172	27.1%	371	вкл. 99 переводных
	проза: рассказы	5,959,142	5.4%	183	вкл. 56 переводных
	проза: драматургия	1,411,030	1.3%	55	вкл. 8 переводных
	итого прозы	**37,279,344**	**33.8%**	**609**	
	поэзия	3,648,160	3.3%	227	вкл. 43 переводных
	Пресса	47,264,735	43.0%	7858	
	Нехудожественные тексты				
	научные тексты	13,875,930	12.6%	113	вкл. 22 переводных
	эссе, мемуары, официальные и религиозные тексты	4,735,997	4.3%	379	вкл. 8 переводных
	Итого письменных текстов	**106,804,166**	**96.8%**	**9,186**	

	Устная речь	словоупотребления	доля в ВАНК	документы
	Спонтанная устная речь	1,029,646	0.94%	208
	Публичная устная речь	1,933,899	1.76%	543
	Стимулированные нарративы	70,010	0.06%	22
+	Электронная коммуникация	442,399	0.40%	1
	Итого устной речи	**3,475,954**	**3.2%**	**774**
	Итого в ВАНК	**110,280,120**	**100%**	**9,960**

Приложение 2. KWIC выдача



Приложение 3. Глоссированная выдача



Приложение 4. Отображение результатов в транслитерации

Gorc, 1990.08 #21	1990	Расширить контекст ▸
«Tariner aŕaǰ,– asel ē na,– Erewanum Ēdvard Harutʻyunyani glxavorutʻyamb, orn ayžm kendani čʻē, himnel enkʻ mardu iravunkʻneri paštpanutʻyan hanjnaxumb:		

Mǰnašen	Xalapʻyan Zorayr	Расширить контекст ▸
— Miangamayn aŕoġǰ mardu , očʻ mi hogekan šeġum:		

Azg, 12.20	2006	Расширить контекст ▸
Manuk Gasparyan-Xozi misə hamov ē, bayc‘ nra arark‘nerə, gorcoġut‘yunnerə, ōrinak‘ məštapes c‘exi meǰ t‘avalvelə, thač en mardu hamar:		

Приложение 5. Статистика запроса

Вхождений: **46 477**, документов **5 700** Размер подкорпуса: **100%** от общего объема ВАНК

Մետաքսի ճանապարհը	Շեկոյան Արմեն	Расширить контекст ▸
Արդեն չորս տարի է՝ Կիրակոսը Կոմայգի է գալիս ամեն օր, ռեժիմով, ինչպես, ասենք, մարդիկ աշխատանքի են գնում:		

Ջերմանց միսիթարություն	Խալափյան Զորայր	Расширить контекст ▸
Սկսել եմ ձիու քայլերով ման գալ, մարդիկ էլ աչքիս շախմատի քարեր են երևում:		

Նրա ճանապարհը, մաս 4	Թաթիկյան Շահեն	Расширить контекст ▸
Վատ մարդիկ չեն, կծանոթացնեմ:		

Приложение 6. Статистическое употребление словоформы աստծո **astco (бог.gen) по данным ВАНК на март 2009 г.**

Словоформа: աստծո Число вхождений: **9,229** Ранг: **386.3** WPM: **550.**

	Худ.	Нехуд.	Пресса	Устные	Всего по декаде
(1800–1859)	2	n/a	0	n/a	2
(1860–1869)	17	n/a	0	n/a	17
(1870–1879)	61	10	0	n/a	71
(1880–1889)	129	6	0	n/a	135
(1890–1899)	57	n/a	n/a	n/a	57
(1900–1909)	46	10	0	n/a	56
(1910–1919)	92	n/a	0	n/a	92
(1920–1929)	15	0	2	n/a	17
(1930–1939)	167	2	6	n/a	175
(1940–1949)	82	12	0	n/a	94
(1950–1959)	205	18	11	4	238
(1960–1969)	1160	19	30	11	1220
(1970–1979)	759	198	11	2	970
(1980–1989)	906	143	21	6	1076
(1990–1999)	237	77	59	1	374
(2000–2009)	166	528	1673	201	2568
недатированные	2038	15	1	13	2067
Всего по жанру	**6139**	**1038**	**1814**	**238**	9229

Синтаксические нули в теории грамматики[1]

Zero Categories in Universal Grammar

Циммерлинг А. В. (meinmat@yahoo.com)

Московский государственный гуманитарный университет
им. М. А. Шолохова, Россия

В статье обсуждается статус нулевых элементов в теоретическом синтаксисе. В некоторых случаях введение синтаксических нулей является оптимальным инструментом описания трансформационных отношений между разными конструкциями. Сведение всех гипотетических нулевых местоимений финитного предложения к единому таксону ‘pro’ непродуктивно. Нулевые местоимения 1–2 л. регулярно обнаруживают свойства отличные от нулей 3 л. Нулевые подлежащие с ролевой семантикой Агенса могут сосуществовать с нулевыми подлежащими без выраженной ролевой семантики.

Правильно построенные структуры иногда могут порождаться и распознаваться носителями языка и в том случае, когда не все их звенья выражены эксплицитно. Это побуждает добавить в грамматику раздел, изучающий условия полной и неполной реализации синтаксических структур и свойства т.н. *нулевых категорий*, которые необходимы для порождения и распознавания синтаксических объектов, но могут быть лишены звуковой оболочки. По современным представлениям, общая грамматика включает операции трех типов — Объединения [поддеревьев] (Merge), Перемещения [поддерева] (Move)[2] и Озвучивания (Spell-Out). В [Циммерлинг 2002: 21, 561] совокупность условий полной и неполной реализации синтаксических структур названа «теорией неполноты», что соответствует термину Spell-out. К компетенции теории неполноты относится выявление факторов, в силу которых эксплицитная реализация компонентов необязательна, затруднена или невозможна. Ключевую роль играют понятия эллипсиса (ellipsis) или стирания составляющей (deletion) и понятие нулевой категории (zero lexeme у И. А. Мельчука, empty category у Н. Хомского).

В прикладном синтаксисе нулевые элементы всегда соотнесены с заданным набором конструкций. Введение в их представление «нулей» показывает, что данные конструкции порождаются при помощи таких правил, которые позволяют интерпретировать отсутствие некоторой внешне выраженной категории (например, «подлежащего», или «Именной Группы», или «личного местоимения в имен.п.», или «связки») как результат операции по ее устранению из наличного состава высказывания[3]. Благодаря нулям соответствующие конструкции получают разметку, показывающую, как они порождены. Нулевые элементы повышают распознаваемость синтаксических структур и помогают решать одну из главных задач прикладной лингвистики — *парсинг* естественного языка. Чем больше нулей мы введем, тем больше конструкций мы потенциально сможем распознать. Но набор нулей нельзя расширять ad libitum — иначе обращаясь к каждому новому корпусу текстов и к каждому новому языку, мы будем получать все новые и новые разновидности нулевых категорий. Общая грамматика решает эту задачу, предлагая строго ограниченный (2–4 типа) набор нулевых таксонов и догматические критерии их определения. Ни одна из версий общей грамматики не основана на полном переборе множества языков мира. В связи с этим встают две проблемы. Во-первых, подгонять новый материал под заранее принятые таксоны нужно так, чтобы различия между нулевыми сущностями в пределах одного и того же таксона не превысили различия между нулями предположительно разного типа. Во-вторых, сами наборы нулевых таксонов в разных

[1] Статья написана при поддержке гранта РГНФ 09-04-00297 «Типология синтаксических ограничений».

[2] Мы не касаемся попыток сторонников Н.Хомского объявить операции Move и Merge реализациями одного общего механизма.

[3] В [Holmberg 2005] предлагается считать, что часть нулей (PRO) соответствуют критерию «запрета на ненулевой план выражения», а другая часть нулей (pro) соответствуют критерию «операцию по устранению элемента из некоторой синтаксической позиции».

версиях общей грамматики не совпадают, поэтому надо проверить, можно ли добавлять те или иные типы нулей в версии общей грамматики, отличные от той, где их изначально выдвинули. Наиболее известными теориями, прибегающими к понятию синтаксических нулей, остаются учение Ш.Балли об эллипсисе и нулевом знаке [Балли 1955: 175–181] и концепция четырех типов нулевых категорий, выдвинутая в рамках Теории Управления и Связывания Н.Хомского [Chomsky, Lasnik 1993: 518–523]. Необходимо рассмотреть также теорию И.А.Мельчука [Mel'čuk 1979] и ревизии хомскианской доктрины, выдвинутые сторонниками Минималистской Программы (Minimalist Program) в 2000-е гг., ср. [Sigurðsson 2008], [Holmberg 2005].

В естественном языке есть непустое пересечение между явлениями, относимыми к компетенции теории неполноты, и явлениями, относимыми к базовому компоненту синтаксиса, так как не все правила построения верифицируемы. A priori неясно, есть ли в итальянских предложениях типа ит. Questo decisione non rende [e] felice, букв. «Это решение не делает [кого-либо] счастливым», нулевое дополнение с генерическим значением, как утверждает Л. Рицци. Чем больше постулатов о свойствах правильно построенных структур содержит теория синтаксиса, тем сильнее она нуждается в критериях неполноты. Самая подробная теория нулевых категорий возникла в рамках доктрины Хомского, где принят постулат об изоморфности всех составляющих.

Понятие эллипсиса сугубо синтаксично: оно применимо лишь к тем объектам, которые состоят из синтаксических элементов низшего уровня, т.е. к предложениям и словосочетаниям. Понятие нулевого знака разработано структуралистами начала XX в. в связи с анализом грамматического значения: оно применимо как к морфологическим, так и к синтаксическим объектам. Р.О. Якобсон в статье 1939 г. показал, что нулевые словоформы и нулевые лексемы, т. е. элементарные знаки, имеющие означаемое, но лишенные означающего, могут быть выведены на основе тех же отношений между планом выражения и планом содержания оппозиций сложных синтаксических знаков, что и нулевые морфы и нулевые морфемы. Программа Якобсона была развита И. А. Мельчуком. Постулировав для предложений типа рус. *Улицу засыпали песком* нулевую лексему 3 л. мн.ч. со значением «люди», И. А. Мельчук признал ее нетождественной реальной лексеме *люди*, так как отсутствие материального подлежащего с генерическим значением дает в русском языке другой эффект, нежели употребление лексем *люди* или *кто-нибудь* [Мельчук 1995: 180, 184, 187]. Но если синтаксический нуль нигде не синонимичен материальному элементу, то критерий выделения нулевых лексем не соответствует определению нулевого знака по Балли как такого знака, который имеет «определенное значение и определенную позицию» в синтагме, которую «можно заменить одной или несколькими синтагмами того же вида, где этот [знак] имеет эксплицитную форму». Поэтому попытка игнорировать разницу между нулевыми знаками в синтаксисе и в других уровнях репрезентации обречена на неудачу.

Понятие нулевых категорий в Теории Управления и Связывания держится не на постулате о знаковых свойствах элементов, но на несовпадении уровней репрезентации [Chomsky, Lasnik 1993: 518–523]. Четыре типа пустых категорий различаются значением признаков матрицы из двух столбцов [± анафоричность, ± местоименность]. Один из четырех типов — нулевые местоимения (*pro*) [– анафоричность, + местоименность] — обобщает случаи опущения неанафорического подлежащего и частично соответствует нулевым лексемам Мельчука. Синтактика всех четырех типов — нулевых местоимений (*pro*), нулевых подлежащих вставленной предикации (*PRO*) [+ анафоричность, + местоименность], следов ИГ, устраняемых в позиции повторной номинации [+ анафоричность, – местоименность], следов линейного перемещения [– анафоричность, – местоименность] вытекает из их места в структуре.

[– анафоричность, + местоименность]	Нулевые местоимения (*pro*)
[+ анафоричность, + местоименность]	Нулевые подлежащие вставленной предикации (*PRO*)
[+ анафоричность, – местоименность]	Следы ИГ, устраняемые в позиции повторной номинации
[– анафоричность, – местоименность]	Следы линейного перемещения

Рис. 1

Различия между нулевыми лексемами по Хомскому и по Мельчуку не столь велики, чтобы отрицать сходство их теорий. В обоих случаях нули объясняют неполное соответствие разных уровней репрезентации, и это, по-видимому, единственный способ обоснования нулевых сущностей в синтаксисе. Неверно, будто теория Мельчука отказывается от уровневого подхода, при котором более глубокий (*n*-й) уровень интерпретирует отношения на *n+1*-м уровне организации языка. В теории Хомского нулевые категории возникают в силу наложения двух формально-синтаксических репрезентаций предложения, а у Мельчука интерпретирующим уровнем для поверхностного синтаксиса оказывается семантика предложения и его ролевая структура. По-существу, нулевые лексемы 3 л. ед. ч. и 3 л. мн. ч. в предложениях рус. *Улицу засыпало песком* и *Улицу засыпали песком* в теориях Якобсона и Мельчука вводятся потому, что Якобсон и Мельчук признают в них предикат *засыпать песком* агентивным, а актант *улицу* — Пациенсом. Это побуждает

их рассматривать пары *Рабочие засыпали улицу песком → Улицу засыпали песком* и *Буря засыпала улицу песком → Улицу засыпало песком* как диатезные преобразования и постулировать для вторых членов пар «нулевой Агенс». Тем самым, 'нули Мельчука' есть ролевые сущности с приписываемыми им референтными свойствами, но этот факт замаскирован семиотическими соображениями. Нулевые категории Хомского — референциально-дейктические сущности без выраженной ролевой семантики, что признается открыто.

При обращении к материалу встают следующие вопросы:

- В каких синтаксических позициях возможны нулевые категории?
- Обязательна ли гипотеза о нулевых категориях?
- Можно ли вводить в описание одного и того же языка нулевые категории, выдвинутые с позиций разных теорий синтаксиса?
- Как дифференцировать разные типы нулей в одном и том же языке?

Первый вопрос должен решаться по-разному в прикладном и теоретическом синтаксисе. При создании синтаксического процессора и разметке анафорических связей может понадобиться введение нулей самых разных типов. С инженерной точки зрения полезно вводить разные нули для случаев эллипсиса, сочинительного сокращения, для предполагаемых подлежащих инфинитивных, причастных, деепричастных оборотов, для зависимых финитных: отсутствие точных помет только замедлит работу процессора и затруднит пользователю доступ к информации. Но удобство описания не является достаточным основанием для введения нулевых категорий в теоретическом синтаксисе: здесь надо доказать, что нулевые лексемы имеют ряд общих свойств с некоторыми ненулевыми синтаксическими элементами. Пока что такие доказательства были предложены для выражений двух типов — нулевых подлежащих и нулевых связок (ср. нулевую связку «быть» в наст.вр. изъявит. накл. в русском языке). Нулевые дополнения более эфемерны, так как их признание зависит от наличия структур, где отсутствие материально выраженного дополнения однозначно свидетельствует о том, что позиция дополнения сохранена, но заполняется нулевой лексемой. Гипотезы о нулевых подлежащих и нулевых связках тоже зависят от формализма (are theory-internal), но наличие позиций подлежащего и связки в языках мира в куда меньшей степени зависит от выбора предиката (переходный/непереходный и т.п.) и его синтактики.

Вопрос об альтернативах нулям поднят в работе Л. Бэбби [Babby 2002], которая посвящена проблеме введения нулевых подлежащих в структуру безличных предложений. Бэбби отрицает, что позиция подлежащего входит в валентную рамку глагола и что глагол приписывает подлежащему какую-либо семантическую роль (theta-role). Бэбби отрицает также, что все предложения изоморфны и открывают позицию подлежащего; поэтому он не хочет вводить нулевые подлежащие для таких случаев, как рус. *лодку качает* или укр. *Вазу було розбито (вітром)*. Минусом такого подхода является отказ от вывода односоставных предложений из двусоставных (или наоборот). Если не прибегать к предположению о том, что продуктивные модели предложений сосуществуют в одном языке случайно, нужно обосновать механизм, который связывает употребления рус. *качать* и укр. *розбить* при материально выраженном подлежащим в имен.п., и при отсутствии такового. Именно это и предлагают теории нулевых подлежащих — от Пауля до Хомского. Это не бесспорное и не единственное решение — можно брать односоставные предложения за исходную точку и добавлять позицию подлежащего за счет особых правил. Но в любом случае отказ выводить одну конструкцию предложения из другой не является подлинной альтернативой гипотезе о нулевом подлежащем. Симптоматично, что Дж. Лавин, исследовавший те же факты, что и Бэбби, вновь ввел синтаксические нули [Lavine 2005]. Альтернативой введению нулей в представление предложений вроде *Вазу було розбито (вітром)* было бы признание ИГ *вазу* подлежащим в косв. п. (oblique subject), но эту возможность Лавин отклонил.[4] Доказательство реальности нулевых подлежащих состоит в том, что они исключают некоторые другие типы выражений, которые способны занимать позицию подлежащего в том же языке. Этот подход апробирован на материале финского языка в [Holmberg 2005] и на материале русского языка в [Zimmerling 2008], где выделено несколько типов неканонических подлежащих, включая несогласуемое слово *это*.

Третий вопрос решается в зависимости от того, как моделируется теория. Поскольку нулевые лексемы Мельчука реально обосновываются несовпадением разных уровней репрезентации синтаксиса, непреодолимых препятствий для введения их в другие концепции, где используется представление об уровнях (этапах, фазах и т.п.) порождения предложения, в том числе — в Минималистскую Программу Хомского, нет, что мы пытались показать в [Zimmerling 2007]. Добавить новые виды нулевых категорий (следы и *pro*), можно и в модель «Смысл–Текст». То, что ни то, ни другое до сих пор не сделано, объясняется человеческим фактором.

[4] Если синтаксические нули и некоторые аргументы в косв.п. имеют те же свойства, что и материально выраженные подлежащие, маркированные стандартным для данного языка способом, их нельзя произвольно соединять в пределах одного финитного предложения: надо либо постулировать для рус. *мне было холодно* нуль в позиции подлежащего, либо объявить подлежащим элемент мне, но не применять эти гипотезы одновременно.

Четвертый вопрос — догматический. Исходная матрица из двух столбцов [± анафоричность, ± местоименность], дающая четыре типа нулевых категорий (*pro*, PRO, след линейного перемещения, след R-выражения), недостаточна для описания материала. Таксон *pro* используется для обозначения двух заведомо разных явлений — нулевых личных местоимений и гипотетических аналогов формального подлежащего, ср. дат. *Det* er gott «это хорошо» и исл. ∅-er gott “то же”. Сама комбинация черт [+anaphoric, +pronominal], в 1980-е гг. постулированная Хомским для PRO, т.е. подлежащего контролируемого инфинитивного оборота, с точки зрения лингвиста выглядит странно, о чем его последователи решились сказать сейчас. Х. Сигюрдссон пишет о том, что конструкты ‘anaphoric’ и ‘pronominal’ не есть собственно языковые признаки, так как за ними не стоят никакие синтаксические операции (they are not accessible or visible to syntax as objects or units) и они не поддаются семантической интерпретации (get no interpretation at the semantic interface). Сигюрдссон предлагает считать, что PRO — референтная категория местоименного типа, имеющая морфологический падеж и согласовательные признаки (a *phi*-feature variable with morphological case), см. [Sigurðsson 2008a: 2]. Такая ревизия PRO превращает ее в аналог нулей, опирающихся на понятие означаемого. Другие теоретики стремятся свести PRO к рутинной процедуре вычеркивания кореферентных составляющих в матричной и вставленной клаузе: такой подход называется термином ‘movement theory of control’, ср. [Boeckx & Hornstein 2004]. Но movement theory of control, утверждающая, что PRO — просто стертый след (deleted copy), не объясняет свойства тех конструкций, для которых в свое время был предложен таксон PRO. Кроме того, вызывает вопросы само понятие «копирования», на которое опирается movement theory of control. Речь идет о попытке выдать инженерное решение за теоретический постулат и свести базовое понятие синтаксического перемещения (movement) к предположительно более простым операциям — присоединению поддеревьев (merge) и копированию (copying) с возможностью/невозможностью восстановления копии (reconstruction).

Утверждения, будто прогресс доктрины Хомского отменил понятия Перемещения и следа (trace), не стоит принимать буквально: перемещение поддерева (Internal Merge), моделируется как векторное преобразование ровно за счет введения следов, указывающих на первоначальное положение элемента, что признается в грамматиках Э. Стэблера, являющихся строгими моделями Минималистской Программы [Stabler 1997]. Но верно, что у сторонников Хомского отпала нужда приписывать следам разных типов разные категориальные свойства, в отличие от того, что Теория Управления и Связывания предписывала делать в 1980-е гг. А. Алексяду и А. Анагнастополу изучали связь между базовым порядком слов (SVO ~ SOV ~ VSO), параметром нулевого подлежащего по Л.Рицци и реализацией *pro*. Они возражают против практики объяснять порядок VSO как обращение порядка SVO с добавлением *pro* в предглагольную позицию (SVO → *pro* VSO) и принимают тезис, близкий к идеям упомянутого Э. Стэблера: они утверждают, что синтаксическая позиция подлежащего открывается перед глаголом лишь в том случае, когда в предикативном ядре предложения (узел AgrSP или IP, в позднейшей записи) есть сильная черта (strong nominal feature), требующая ее открытия [Alexiadou & Anagnastopoulou 1998: 501]. Языки с «нулевым подлежащим», по их мнению, порождаются в результате передвижения глагола (verb raising), который пересекает позицию подлежащего: такая операция возможна потому, что согласовательная морфология «сильна» и позволяет идентифицировать лицо и число подлежащего без вставки местоименного подлежащего (нулевого или ненулевого). Поэтому Алексяду и Анагнастополу не постулируют начальное *pro* для языков с «нулевым подлежащим», хотя признают в них *pro* в других случаях. Рассмотренные ими строгие языки VSO, такие как арабский и ирландский, а также романские и балканские языки SVO → VSO с передвижением глагола и объектными клитиками, являются языками *pro-drop* в том классическом варианте, который принято называть «аргументными языками *pro-drop*» (argument pro-drop). Их особенностью является то, что опущенное местоимение 3 л. однозначно идентифицируется и в том случае, когда оно не является анафорическим и не восстанавливается с опорой на дискурс. Русский не является стандартным языком pro-drop. Фраза вроде *pro пошел на лекцию* легко интерпретируется в зн. 1 л. = «Я (ГОВОРЯЩИЙ) пошел на лекцию», с некоторой натяжкой — в зн. 2 л. = «Ты (СЛУШАЮЩИЙ) пошел на лекцию», а интерпретация в зн. 3 л. = «NN пошел на лекцию» исключена за пределами контекстов, где опущенное местоимение анафорично. Ср. диалог «<А. Где Ваня$_i$?> В. *pro*$_i$ пошел на лекцию». Напротив, для таких языков как испанский или новогреческий интерпретация в зн. 3 л. = «NN пошел на лекцию» является дефолтной, а поиск антецедента нулевого местоимения в ближайшем линейном окружении не требуется.

В последовательных языках pro-drop сложились запреты на одновременное употребление личных местоимений и показателей предикатного согласования. Так, в русинском языке (диалект Воеводины), по данным У.Брауна, возможны фразы вроде *Добри=є*, *Вон добри*, но не **Вон=є добри*, где одновременно представлены клитика 3 л. ед.ч. *=є* и личное местоимение 3л. ед.ч. м.р. *вон*. Значение «Я член какой-л. организации», можно выразить по-русински и как *pro член=сом*, и как *я член*, но плеоназм **я=сом член* запрещен. Аналогично, значение «я не читал нечто»,

может быть выражено как при помощи фразы с ненулевым личным местоимением — *я нє читал*, так и при помощи фразы с клитикой — *нє читал=сом*, но комбинация обоих маркеров согласования исключена: **я нє читал=сом*, **я=сом нє читал, *я нє=сом читал* [Browne 2008].[5] Сходным образом устроен древненовгородский диалект XI-XV вв. В этих языках связка-клитика согласуется с подлежащим только в том случае, когда подлежащее внешне не выражено! Более эффектное доказательство существования *pro* трудно себе представить.

А. Алексяду и Е. Анагностопулу верно переводят наблюдения типологов над особенностями языков VSO и SVO → VSO на метаязык Минималистской Программы, хотя некоторые их доктринарные решения скорее экстравагантны, чем доказательны: так, исследовательницы утверждают, что порядок SVO в балканских и южно-романских языках возникает за счет «клитизации подлежащего слева к глаголу» (clitic-left-dislocation). Этот тезис нужен, чтобы развести строгие VSO языки (-EPP/XP; +SpecTP в терминологии авторов) и языки SVO → VSO типа новогреческого (-EPP/XP; -SpecTP в их терминологии). В строгих VSO языках вроде арабского препозитивное подлежащее при порядке S || VO «топикализовано», т.е. занимает позицию SpecTP, а в SVO → VSO языках типа новогреческого порядок SVO не трактуется как топикализация (подлежащее не занимает SpecTP). Для теории нулевых категорий важен вывод о том, что постпозитивное подлежащее в строгих VSO языках остается за пределами ГГ (is VP-external), в то время как в SVO → VSO постпозитивное подлежащее остается в пределах ГГ и, предположительно, занимает узел SpecIP. Отсюда следует, что если нулевым местоимениям приписывается синтаксический статус подлежащего, их тоже следует помещать в узел SpecIP, по крайней мере, для SVO → VSO языков вроде новогреческого.

А. Хольмберг, который критикует Алексяду и Анагностополу за стремление ограничить применение *pro*, разделяет одно из допущений их анализа: он устанавливает позицию *pro* на основе позиций эксплетивных подлежащих вроде англ. it, there, которые он признает материализовавшимися нулями. По Хольмбергу, эксплетивы и коррелятивные им нули могут стоять либо в SpecTP, (SpecCP, в записи Хольмберга), либо в SpecIP. «Топикализованные» эксплетивы/нулевые формы, по его мнению, свойственны языкам типа исландского. Стоящая в SpecCP форма возможна в начале предложения, но заменяется нулем, если предфинитная позиция уже заполнена другим элементом. Проиллюстрируем этот тезис Хольмберга собственными примерами, взятыми из фарёрских текстов: в (1) эксплетив *tað* сохраняется в начале предложения, в (2) он опускается в постпозиции глаголу.

(1) фар. *Vit byrjuðu hesa greinina við at siga,*
[$_{CP}$ *at* ***tað****-Expl* *gongur-3Sg upp og niður* *í fiskivinnuni-DatPrepSg*].
«Мы начали эту статью с утверждения о том, [$_{CP}$ что в рыболовстве дела идут то лучше, то хуже»], букв. «…[что **это** идет вверх и вниз»].

(2) фар. *Í fiskivinnuni-DatPrepSg ∅-gongur-3Sg altíð upp og niður*,
«В рыболовстве ∅-идет все время (то) вверх, (то) вниз».

В качестве образца *pro* и эксплетива, занимающих узел SpecIP, Хольмберг ссылается на ситуацию в финском языке. Здесь гипотетические нулевые местоимения и эксплетивы, как и в исландском и фарёрском языках, могут употребляться при одних и тех же предикатах, ср. грамматичные примера(3 а-б) при неграмматичном (3в)

(3) фин. а. *pro* Meni hullusti.
went wrong
«(дело) пошло плохо»
б. **Sitä** meni hullusti.
EXPL went wrong
«то же»
в. *Meni nyt hullusti.
Went now wrong

Неграмматичность порядка V + Adv + AdvPred в (3 в) объясняется не тем, что финский язык запрещает предложения с начальным глаголом — это не так, ср. (3а), а тем, что элемент, для которого типична роль темы, должен выноситься в предглагольную позицию: *Meni nyt hullusti → nyt$_i$ meni t$_i$ hullusti. Если же такого перемещения не происходит, в предглагольную позицию вставляется эксплетив sitä, ср. (3б) или нулевое местоимение *pro* (3а). В отличие от фарёрского, sitä и другой эксплетив, se[6] не устраняются в постпозиции глаголу, например, в общих вопросах, см. (4).

(4) фин. Meni-kö **sitä** taas hullusti?
Went-Q EXPL again awry
‘Дела что ли снова пошли плохо?’.

Как и эксплетивы в исландском и фарёрском, финское sitä возможно в двусоставных конструкциях, в том числе, при подлежащем в 1 и 2 л. и согласо-

[5] В русинском языке допускается однократное употребления маркеров лица и числа во фразах типа *Барз=є красни ‘(Это) очень красиво/-ый’* или *Авто барз красни ‘машина очень красивая’*, но запрещается плеоназм типа **Авто=є барз красни «Машина очень красивая»*.

[6] Se и sitä являются разными падежными формами одного и того же местоимения 3 л.

ванной форме глагола. В этом случае sitä допустимо как в предфинитной, так и в постфинитной позиции, что сближает его с употреблением несогласуемого слова *это*, ср. рус. *это мы с матерью пришли*. Однако фин. sitä, в отличие от рус. *это*, не может предшествовать тематическому подлежащему, если оба они вынесены перед глаголом: (5б) грамматично потому, что местоименное подлежащее minä «я» является ремой или фокусом контраста, в то время как для (5в) найти соответствующую коммуникативную интерпретацию, видимо, нельзя:

(5) фин. а. **Sitä** olen minä -kin käynit Pariisissa.
EXPL have-1Sg I-too visited Paris-INE
'Я действительно тоже бывал в Париже»'
б. Minä **sitä** olen käynit Pariisissa.
'И я бывал в Париже (представь себе)' /'Я-то бывал в Париже'.
в. *__sitä__ minä olen käynit Pariisissa.

Подобные факты дают основания трактовать фин. sitä не столько как подлежащее, сколько как «формальную тему», но Хольмберг эту возможность отвергает, настаивая на том, что тема или «топик» имеет строго определенные линейные позиции в финском предложении. Последний тезис требует проверки, но ясно, что часть запретов на употребление sitä стоит объяснять синтаксически. Так, финский допускает опущение личных местоимений 1 и 2 л. при согласуемой форме глагола, что оправдывает гипотезу о *pro*, но при этом запрещает вставку эксплетива как при внешне выраженном, так и при опущенном местоимении 1 и 2 л., ср. (6а–б) и (7а–б)

(6) фин. а. *__Sitä__ puhun englantia.
EXPL speak-1Sg English
Букв. *"**это** говорю по-английски".
б. *__Sitä__ *minä* puhun englantia.
EXPL I speak-1Sg English
Букв. *«**это** *я* говорю по-английски».

(7) фин. а. Oletteko (*__sitä__) käyneet Pariisissa?
Have-2Pl-Q EXPL visited Paris-INE
Букв. "Бывали-ли-*вы* (*__это__) в Париже?"
б. Oletteko te (*sitä) käyneet Pariisissa?
Have-2Pl.-Q you EXPL visited Paris-INE
Букв. «Бывали-ли-вы *вы* (*__это__) в Париже?»

Эти данные свидетельствуют о том, что вставка фин. sitä приурочена к определенным позициям, но они не дают ответа, каков в действительности статус sitä — подлежащее или маркер темы, так как все формальные запреты одновременно влекут за собой некоторые коммуникативные ограничения. Если сопоставить финские предложения с русскими, где, как и в финском, опущение тематических местоименных подлежащих возможно лишь в 1–2 л., но не в 3 л., мы увидим, что русские аналоги (6а), (7а–б) исключены полностью — ср. рус. **это говорю по английски, *бывали ли это в Париже*? и *бывали ли вы это в Париже*?[7], а русские аналоги (6б) возможны лишь при определенном чтении, ср. *это* ↘↘*Я говорю по английски* (местоимение «я» является фокусом контраста) ∨ *это* ↘↘*я говорю по* ↘↘*английски* (коммуникативно нерасчлененное предложение), но неуместны при других. Грамотные носители русского языка назовут предложения рус. *говорю по-английски* и *бывали ли в Париже?* 'двусоставными неполными'; многие согласятся, что в иной системе терминов этот результат можно записать, как *pro говорю по-английски*. Но отсюда вовсе не следует, что пример **это говорю по-английски* аномален потому, что слово *это* (чей статус в качестве грамматического подлежащего еще нужно доказать) стоит в той же позиции, что *pro* (которое гипотетически отличается от результата эллипсиса русских личных местоимений *я/ты/мы/вы*, что также еще нужно доказать). Предпосылкой для употребления несогласуемого безударного ситуативного слова *это* является его референция к ситуации, названной в предтексте или существующем в экстралингвистической реальностей и имплицирующей некоторую логическую связь между текущей и внешней ситуацией, ср. *<Тише! > Это* ↘↘*Анатолий Сергеевич звонит*. Нет причин думать, что в языках, где употребление ситуативных слов более грамматикализовано, чем в русском, их вставка производится по автономным от коммуникативного синтаксиса причинам. Примеры Хольмберга на вставку sitä в финские предложения с глаголом в 3 л. и эксплицитно выраженным подлежащим и дополнением не рассеивают этого ощущения. Сам Хольмберг приводит (8а-г) в доказательство того, что финский язык удовлетворяет ограничению «глагол на втором месте от начала предложения» (Verb-Second-Constraint, V2), в силу чего якобы запрещается порядок VSO с начальным глаголом (8а), но разрешается порядок Expl+V+S+O, где инверсия глагола и подлежащего компенсируется. Больше оснований заключить, что финский язык не допускает коммуникативно-расчлененные на тему и рему предложения с начальным глаголом и двумя именными аргументами[8]: если ни один из актантов, способных играть роль темы, не перемещается в предглагольную позицию, ср. (8б) и (8в), вставляется эксплетив sitä (8г):

(8) фин. а. *Leikki lapsia kadulla.
Играют-3Pl дети-PAR улица-INE
б. Kadulla leikki lapsia.
улица-INE играют-3Pl дети-PAR 'Дети играют на улице'

[7] Мы исключаем чтения, где *это* является междометием.

[8] Возможно, также и коммуникативно-нерасчлененные предложения с начальным глаголом. Русский язык соответствующие предложения разрешает.

с. Lapsia leikki kadulla.
дети-PAR играют-3Pl улица-INE ‘то же’
г. **Sitä** leikki lapsia kadulla.
EXP играть-3Pl дети- PAR улица-INE
‘то же’

Финский язык разрешает предложения с начальным глаголом, ср. (3а): если кто-либо скажет, что примеры вроде (3а) *pro* meni hullusti удовлетворяют параметру V2, поскольку перед глаголом meni стоит нулевая категория *pro*, возникнет порочный круг. Как бы ни обозначать позицию *pro* (SpecIP vs SpecCP), ограничения типа германского V2 устанавливаются с опорой на эксплицитно выраженные категории: иначе бы не было никакой разницы между аномальным примером (8а) **pro* leikki lapsia kadulla и грамматичным примером (8г) **Sitä** leikki lapsia kadulla — и теоретики имели бы повод заявить, что ИГ lapsia «дети», стоящая в постпозиции глаголу leikki, в обоих случаях не является подлинным подлежащим.

Фин. sitä можно определить как эксплетивное слово со статусом темы и ситуативной референцией: оно может занимать аргументные позиции, хотя его статус в предложениях с согласованной формой предиката требует уточнения. Нет оснований считать его материализацией *pro* или какой-либо иного нуля; рискованно также определять место *pro* в дереве предложения с опорой на дистрибуцию *pro*, и наоборот. Анализ А. Хольмберга пробуждает скепсис в плане того, что все рассмотренные случаи употребления гипотетических нулевых местоимений подводимы под единый таксон *pro*. В его изложении сюда попадают как случаи эллипсиса референтных местоимений 1 и 2 л., так и употребления ситуативных местоимений 3 л., предположительно коррелятивных эксплетивам se и sitä (которые при определенных условиях сочетаются с ненулевыми местоимениями 1 и 2 л., что дополнительно запутывает ситуацию). Можно усомниться и в том, что финский (а также русский) вообще являются «языками с нулевым подлежащим». Как отмечалось выше, в 3 л. опущение неанафорического местоименного подлежащего в этих языках невозможно, что заставляет отнести оба языка к классу языков с «дискурсивным опущением *pro*», а не к классическим языкам с «аргументным опущением *pro*» (ср. испанский, греческий, сербохорватский, русинский, древненовгородский и т. д.).

Попытка А. Хольмберга демонстрирует те трудности, которые возникают, если за основу взять «дискурсивное опущение местоимений», а полученные таким путем критерии *pro* распространить на конструкции, где нулевые элементы имеют нетривиальные синтаксические свойства (связанные с ролевой семантикой, референцией и выбором определенной граммемы лица и т.д.). Если такой подход чреват трудностями даже в языке с богатым согласованием, есть опасения, что понятие «дискурсивное опущение местоименных подлежащих» окажется фикцией для многочисленных языков, где согласование между подлежащим и сказуемым отсутствует.

И. А. Мельчук в 1979 г. предложил программу, близкую к тому, что сейчас называется «argument pro-drop». Он настаивает на том, что 1) нулевые подлежащие следует вводить лишь в том случае, если глагольная форма имеет внешне выраженное согласование; 2) нулевую лексему можно постулировать

	Структуры с субъектно-предикатной связью и согласованием	
Pro	I. Фин. (*Minä*) puhun englantia. Рус. (*Я*) говорю по-английски.	I. Фин. (*Me*) puhumme englantia. Рус. (*мы*) говорим по-английски.
Pro	II. Фин. (*Sinä)* puhut englantia. (*Ты*) говоришь по-английски	II. Фин. (*Te*) puhutte englantia. Рус. (*Вы*) говорите по-английски.
*	III. Фин. (***Han**) puhuu englantia. Рус. (***Он/она**) говорит по-английски.	III. Фин. (***Не**) puhuvat englantia. Рус. (***Они**) говорят по-английски.[9]
Кореферентное сокращение во вставленном предложении NP1 = NP2	I-III. Фин. Pekka$_i$ väittää [että Ø$_i$ /hän$_i$ puhuu englantia]. ‘П. считает, [$_{CP}$что (он) говорит по-английски].’ Рус. Я считаю/ты считаешь/Петр$_i$ считает [$_{CP}$что ?(я) говорю/, (ты) говоришь/ (он$_i$) говорит по-английски].	
	Структуры с нулевым/эксплетивным элементом, без согласования с предикатом	
Ø^{3Sg} ~ EXPL	Маркер безличного предиката: фин. Ø^{3Sg} meni hullusti; Nyt Ø^{3Sg} / *se* taas sataaa ’Теперь снова идет дождь’. Рус. Ø^{3Sg} плохо подвигается; Ø^{3Sg} (*оно) морозит, светает.	
Ø^{3S} ~ EXPL	Коррелятивный маркер CP: Фин. Ø^{3Sg} /**Se**$_i$ oli hauskas [$_{CP}$että tulit käymään]$_i$ ‘(**Это**$_i$) Приятно, [$_{CP}$что вы пришли навестить]$_i$’ . Рус. Ø^{3Sg} /**Это**$_i$ странно, что [[$_{CP}$что он опоздал]]$_i$. (*Это$_i$) мне странно, что [[$_{CP}$что он опоздал]]$_i$.	

Рис. 2. «Дискурсивное опущение местоимения» и нулевые формы

[9] В таблице не учтены неопределенно-личные предложения вроде так говорят по-английски, для которых в теории И. А. Мельчука предусмотрена особое нулевое подлежащее ØPeople (нулевой нереферентный Агенс).

только в том случае, если она вносит в представление высказывания информацию, которая не сигнализируется какой-либо ненулевой формой; 3) каждая постулированная нулевая лексема должна иметь специфические ролевые и референтные свойства, которые не дублируются в полном объеме какой-либо другой лексемой, ненулевой или нулевой. Концепция Мельчука в полном объеме апробирована лишь на материале современного русского языка; альтернативы ей и границы ее применения в типологии обсуждаются в [Zimmerling 2007], [Zimmerling 2008], [Циммерлинг 2008]. Здесь нас интересуют следствия принятия анализа Мельчука для русского языка и языков с близкими морфосинтаксическими параметрами. Напомним, что Мельчук ввел два разных нуля для 3 л. ед. ч. и мн. ч. в позицию подлежащего в те конструкции, которые в русистике называют неопределенно-личными, ср. (9) и безличными, ср. (10). Оба нуля имеют генерическое значение и роль Агенса, но нуль 3 л. мн.ч. ($\varnothing^{\text{people}}$ в записи Мельчука) сигнализирует контролируемое и умышленное действие, производимое одушевленным Агенсом, а нуль 3 л. ед.ч. ($\varnothing^{\text{elements}}$ в записи Мельчука) — неконтролируемое действие, производимое природной силой:

$\varnothing^{\text{people}}$ [[Vf.3.Pl]; + умышленное действие]	$\varnothing^{\text{elements}}$ [[Vf.3.Sg]; — умышленное действие]
(9) *Улиц-у засыпал-**и** песк-ом.*	(10) *Улиц-у засыпал-**о** песк-ом.*

Рис. 3. Нулевые лексемы с ролевой семантикой в модели «Смысл — Текст»:

Если пойти дальше и поставить вопрос о морфологии $\varnothing^{\text{people}}$ и $\varnothing^{\text{elements}}$, их можно признать нулевыми неопределенными местоимениями 3 л. в им. п. [Zimmerling 2008]. $\varnothing^{\text{people}}$ and $\varnothing^{\text{elements}}$ обладают рядом синтаксических свойств, общих с ненулевыми подлежащими русского языка: они контролируют и определяют по крайней мере три черты — согласование глагола в числе и роде (3Sg.Neut. для $\varnothing^{\text{elements}}$ и 3Pl. для $\varnothing^{\text{people}}$), рефлексивизацию и присоединение деепричастного оборота, ср. (11) и (12).

Контроль за рефлексивизацией	**Контроль деепричастного оборота**
(11) *В **своем**$_i$ доме $\varnothing^{\text{people}}{}_i$ обычно не гад-**ят***	(12) *Пиратскими дисками $\varnothing^{\text{people}}{}_i$ торгу-**ют** в Лужниках,* [$_{\text{IP}}$ PRO$_i$ ***обеспечивая**$_i$ всех москвичей*][10].

Рис. 4. Нулевые лексемы как контролеры подлежащных свойств

[10] Мы добавили PRO в позицию подлежащего вставленного деепричастного оборота и отметили его кореферентность лексеме $\varnothing^{\text{people}}$ в главном предложении. В рамках самой модели «Смысл — Текст» такая запись не нужна.

В древнерусском языке то значение, которое в русском выражается неопределенно-личными конструкциями в глаголом в 3 л. мн. ч. (ср. лексему $\varnothing^{\text{people}}$ Мельчука), выражалось безличной конструкцией с глаголом в 3 л. ед.ч., как показал Й. Зубаты [Zubáty 1954]. Б. Гавранек указал, что в древних славянских языках не было четкого деления глагольного словаря на «личные» и «безличные» лексемы, так как позицию подлежащего можно было устранить вообще при любом предикате [Havránek, 1962]. Такое состояние обнаруживается непосредственно, либо восстанавливается из сопоставления разных временных срезов. Так, др.рус. пример (13а) из «Слова о Полку Игореве» был уже непонятен переписчику, который исправил его, вставив возвратную частицу *ся:* тем самым, он создал двусоставное предложение с ИГ *гласъ* в позиции внешне выраженного подлежащего в имен. п., ср. (13б).

Конструкция с нулевым подлежащим	**Грамматическое подлежащее в имен. п.**
(13а) *На Дунаеви Ярославнынъ **гласъ**=*Acc Sg $\varnothing^{\text{3Sg}}$ *слыш**ить***	(13б) *На Дунаеви Ярославнынъ **гласъ**=*Nom Sg *слышить **ся**.*
“[**One can hear**] Yaroslavna’s **voice** at Danube”	“Yaroslavna’s **voice** can be heard at Danube”

Рис. 5

В этих условиях подход Мельчука необязателен — каждую глагольную вокабулу легко представить в виде пары лексем, ср. др. рус. *слышить* 3Sg {a)“(someone) hears”, b) “they hear”, “one can hear”}. Но если вводить нулевые подлежащие с ролевыми свойствами, достаточно единственной нулевой лексемы $\varnothing^{\text{3Sg}}$ со значением «УСТРАНЕННЫЙ СЕМАНТИЧЕСКИЙ СУБЪЕКТ». В современном русском $\varnothing^{\text{people}}$ и $\varnothing^{\text{elements}}$ имеют роль Агенса, но в других языках, как в древнерусском, возможны и нулевые нереферентные местоимения с ролью Экспериенцера и прочими неагентивными ролями.

Нулевые и формальные подлежащие могут иметь разные свойства даже там, где они связаны отношением генетической преемственности. В ряде языков конструкции с нулевыми подлежащими были перестроены за счет добавления формальных слов, (ср. Eng. *It, there*, Fr. *Il*, Ge. *es*, Da. *det, der*), причем последние унаследовали синтактику нулевых подлежащих [Циммерлинг 2002, 571–600; 634–664]. Заманчиво трактовать эксплетивы как ‘материализовавшиеся нули’. Такой анализ годится для финитных схем с глаголом в 3 л. акт. залога, но конструкции с причастием и эксплетивом нередко запрещают ‘пассивизацию’ непереходных глаголов действия (ср. значения “COME”, “GO”, “DANCE”), но разрешают ‘пассивизацию’ стативных глаголов (ср. значение “SEE”). Те же самые глаголы ранее допускали безличный пассив с нулевым подлежащим:

	+ Нулевое подлежащее: − Нулевое подлежащее.	− Нулевое подлежащее, + эксплетивное подлежащее.
Языки	Древнеисландский	Датский, шведский, норвежский
Стативы: (−активность)	*Var*=3Sg ***sofit***=Sup *lengi* «спали долго», букв. «было **спавши** долго»	Норв. ***Det***=Expl *blev*=3Sg ***sovet***=Sup lenge hver morgon, букв. «**Это** было **спавши** долго каждое утро»
Глаголы движения: (+активность)	*Var*=3Sg *til dura* ***gengit***=Sup. Букв. «было **пойдено** к дверям» *Var*=3Sg ***komit***=Sup *snemma* букв. «пришедши рано»	Норв. **Det*=Expl *blev*=3Sg ***kommet***=Sup, букв. * «Это было **пришедши**»; **Det*=Expl *blev*=3Sg ***fart***=Sup букв. «**Это** было **поехано**»

Рис. 5 Нулевое подлежащее и эксплетивное подлежащее в скандинавских языках

Может ли нулевое подлежащее иметь морфологический падеж? Для русских $\varnothing^{elements}$ и $\varnothing^{people}$, вводимых в модели «Смысл — Текст», ответ положительный, для *pro* — отрицательный. Мы уже отмечали, что существуют теории (большого) PRO, где данной категории приписывается падеж, причем не абстрактный, а морфологический. Иногда уместно постулировать два разных нуля в двух связанных между собой схемах с ядерным актантом в одном и том же падеже, если их ролевая семантика различна. Ровно это имеет место в современном исландском языке, где, руководствуясь критериями Мельчука, надо вводить два разных нулевых местоимения 3 л. ед. ч. в двух сходных дативных схемах:

Неодушевленный Агенс, {−CONTROL}	Одушевленный агенс {± CONTROL}
Глагольно-безличная конструкция:	Безличный пассив:
Bátnum=DatSg $\varnothing^{1}$-*hvolfdi*=3SgPret «лодку опрокинули», букв. «**лодке** опрокинул(о)».	***Bátnum***=DatSg $\varnothing^{2}$-var=3SgPret *hvolft*=Sup букв. «**лодке** опрокинул(о)».
****Bátnum***=DatSg $\varnothing^{1}$-*hvolfdi*=3SgPret ***viljandi*** букв. *«**лодке** опрокинул(о) **умышленно**».	***Bátnum***=DatSg $\varnothing^{2}$-*var* *hvolft*=Sup ***viljandi***, букв. «**лодку** опрокинуто **умышленно**».

Рис. 6. Два типа нулевых агентивных лексем в исландском языке

То различие, которое исландский выражает противопоставлением глагольных и причастных схем, выражается в русском двумя глагольными схемами с разными нулевыми местоимениями, ср. аномальный пример **лодку* $\varnothing^{elements}$ *опрокинуло умышленно* с грамматичным примером *лодку* $\varnothing^{people}$ *опрокинули умышленно*. Различение финитных/нефинитных схем, схем с глаголом в 3 л. мн. ч. /3 л. ед. ч. оказывается лишь аспектом синтаксической техники, кодирующей одно и то же семантическое противопоставление, что косвенно свидетельствует о реальности нулевых подлежащих с ролевой семантикой.

Можно ли одновременно выделять нескольких местоименных нулей, часть которых имеет ролевые свойства, а часть нет? Современный исландский и фарёрский дают шанс это проверить. А. Хольмберг и К. Платцак относят эти языки к языкам с нулевым подлежащим, где Согласование и Падеж «сильные», в терминах Минималистской Программы [Holmberg & Platzack 1995]. Исландский и фарёрский имеют эксплетивные местоимения 3 л. ср. р. (исл. *það,* фар. *tað*), но они факультативны: предикатов, которые требовали бы эксплетивов при любом порядке слов, нет. При этом вставка *það/tað* возможна лишь в одном классе исландских и фарерских предикатов. Вставка исл. *það* запрещена при таких глаголах как исл. *lægia* «сходить на нет» или *slota* «успокаиваться». Ср. исл. *vindinn-AccSg* ∅*-lægir-3Sg* «ветер стихает», но не ****það****-Expl lægir vindinn*; исл. *veðrinu-DatSg* ∅*-slotaði-3Sg*, но не ****það****-Expl slotaði veðrinu*. Значение ‘Пациенту хуже’ можно выразить по-исландски как в схеме с *það*, ср ***það***-Expl dregur-3Sg stöðugt af sjúklingunum-Dat.Prep. букв. ‘***Это*** тащит постоянно из больного’, так и в схеме, запрещающей вставку *það*, ср. sjúklingunum–Dat. Sg. ∅-hrakar-3Sg, букв. «больному худшает». Нельзя лишь произвольно варьировать эти схемы при одном и том же предикате.

Применив общий таксон для случаев опущения *það/ tað* в классе глаголов типа *draga,* и для случаев, где *það/tað* вообще нельзя вставить (ср. глаголы типа *hraka* type), мы проигнорируем разницу между двумя совершенно разными классами предикатов и синтаксическими схемами. Исландский и фарерский в разных классах предикатов используют оба вида нулевых категорий — и *pro,* и ‘нули Мельчука’. Эти языки лексикализовали употребление нулевого подлежащего $\varnothing^{3Sg}$ в классе *hraka*: класс *hraka* семантически задан и включает только глаголы действия и процесса, но не глаголы других групп. Тем самым, нулевое местоимение $\varnothing^{3Sg}$ отлично от *pro* и имеет следующие ролевые и референтные свойства: {+Агенс; −Референтность; −Одушевленность}. Класс *draga* является открытым, поэтому употребление *pro* и *það/tað* в исландском и фарерском не обусловлено выражением какой-либо ролевой семантики.

В свете выявленных различий между *pro* и ‘нулями Мельчука’ стоит оценить гипотезу о наличии в русском нулевой лексемы 2 л. ед. ч. $\varnothing^{2Sg}$, нетождественной эллипсису местоимения *ты*: напомним,

что эллипсис личного местоимения, по крайней мере, в нотации Хольмберга, квалифицируется как *pro*. Если ограничиться материалом, который привел сам автор, ср. рус. *Его не переспоришь, Идешь, бывало, по Бродвею* [Мельчук 1995: 187], то ни интуитивных, ни формальных оснований признавать $Ø^{2Sg}$ нет, так как эксплицитно выраженное местоимение *ты* тоже часто употребляется в генерическом значении. Но в русском есть две конструкции, которые побуждают ввести нуль (нули) 2 л.: это т. н. коррелятивные предложения типа рус. *Как посеешь, так и пожнешь* и структуры с вторичной предикацией типа рус. *Такое обучение не сделает* [$Ø^{2Generic}$] *умным/умными/умной*, *Пластический хирург не сделает* [$Ø^{2Generic}$] *красавцем*. Русский запрещает коррелятивные предложения в 3 л. с *pro,* ср. (14а) и требует эксплицитно выраженного ненулевого местоимения в левой части коррелятивного комплекса, ср. (14б); в правой части *pro* допустимо (14в), хотя оно и необязательно.

(14) а. * [pro_i *ищет*] $_i$, pro_i *всегда найдет*; *[pro_i *что болит*] $_i$, pro_i *о том и говорит*.
б. [***Кто***$_i$ *ищет*] $_i$, ***тот***$_i$ *всегда найдет*. [***У кого***$_i$ *что*$_j$ *болит*] $_i$, ***тот***$_i$ *о том*$_j$ *и говорит*.
в. [***Кто***$_i$ *ищет*] $_i$, ***pro***$_i$ *всегда найдет*. [*У кого*$_i$ *что*$_j$ *болит*] $_i$, pro_i *о том*$_j$ *и говорит*.

Это является серьезным основанием вводить лексему $Ø^{2Generic}$ в представление примеров вроде (15а), (16а) тем более что в правой части $Ø^{2Generic}$ не может заменяться эксплицитно выраженным генерическим местоимением *ты*, ср. аномальные примеры (15б) и (16б):

(15) а. [$Ø^{2Generic}_i$ Ищешь в чужом государстве],
[$Ø^{2Generic}_i$ / pro_i находишь в своем болоте].
б. *[$Ø^{2Generic}_i$ Ищешь в чужом государстве],
[*ты*$_i$ находишь в своем болоте].

(16) а. [$Ø^{2Generic}_i$ как посеешь]$_i$,
[$Ø^{2Generic}_i$/ pro_i так и пожнешь].
б. *[$Ø^{2Generic}_i$ как посеешь]$_i$,
[так *ты*$_i$ и пожнешь].

Неясно, следует ли признать для русского языка одно или два нулевых местоимений 2 л. С одной стороны, $Ø^{2Generic}$ в коррелятивных предложениях указывает на форму имен. п. ед. ч., а в конструкциях со вторичной предикацией генерическое местоимение 2 л. стоит в вин. п. и может принимать формы всех родов и чисел. С другой стороны, введение единого нулевого генерического местоимения 2 л., способного принимать формы имен. п./вин. п. обеих чисел, но не другие падежные формы, кажется меньшим злом для грамматической теории, чем признание синтаксического нуля, специализированного в функции дополнения[11]. Доказательство того, что в конструкциях со вторичной предикацией реализуется именно $Ø^{2Generic}$, а не эллипсис (*pro*-форма) генерического *ты*, представляет дистрибуция обсуждаемых местоимений, так как эксплицитно выраженное генерическое *ты* согласуется по роду и полу с полом реального референта: пример (17) с нулевым местоимением 2 л. уместен и в том случае, если референт является мужчиной, а пример (18) в этой ситуации звучит аномально.

(17) *Пластический хирург не сделает* $Ø^{2Generic}{}_i$ {M/F; Sg} *красавицей* $_i${F; Sg}.

(18) *Пластический хирург не сделает* ***тебя***$_i$ {F; Sg} *красавицей*$_i$ {F; Sg}.

Подведем промежуточные итоги. Лучший способ проверить реальность синтаксических нулей состоит в обращении к типологии, но типология нулевых элементов еще не создана. Сведение всех нулевых местоимений финитного предложения к общему таксону *pro* непродуктивно; дискурсивное опущение личных местоимений, кореферентное сокращение местоимений и запреты на эксплицитное выражение местоименных подлежащих — три разные механизма. Возможности дальнейшей классификации нулевых местоимений связаны с различением нулевых подлежащих, имеющих ролевую семантику (ср. ‘нули Мельчука’), и лишенных ее, а также местоимений, привязанных и непривязанных к определенным граммемам лица: часть нулевых местоимений 3 л. имеет признаки, не свойственные местоимениям 1–2 л. и нулевым местоимениям, принимающим формы разных лиц. Возможно семантическое противопоставление нулевых местоимений 3 л. ед. ч. и 3 л. мн. ч., но те же значения могут быть выражены и с помощью другой комбинации нулевых категорий. Часть нулевых местоимений охарактеризована в плане падежа и рода, что связано не с семантикой, а с согласовательными ресурсами конкретного языка. Нулевые местоимения с ролевой семантикой Агенса и без нее могут сосуществовать при разных классах предикатов.

---

[11] Структурный вин. п. и имен. п. коррелятивны, а актант в вин. п. в структурах со вторичной предикации естественно квалифицировать как подлежащее вставленного предложения.

Литература

1. *Alexiadou A. & Anagnastopoulou E.* 1998. Parametrizing AGR: Word Order, V-movement and EPP-checking // Natural Language and Linguistic Theory 16: 491–539.
2. *Babby L.* 2002. Subjectlessness, External Subcaterization, and the Projection Principle // Journal of Slavic Linguistics. 10: 341–88.
3. *Boeckx C. & N. Hornstein.* 2004. Movement under Control // Linguistic Inquiry 34, 269–280.
4. *Browne* 2008. Clitic Ordering in Vojvodina Rusinski // Slavic Linguistic Society 3, Ohio June 10–12, 2008.
5. *Chomsky N., Lasnik H.* 1993. Syntax in generativen Grammatik // Syntax: An International Handbook of Contemporary Research / J. Jacobs, A. von Stechow, W. Stermfeld, T. Venneman, eds. Berlin & New York, Walter de Gryuter, 506- 569.
6. *Gilligan, Gary M.* 1987. A cross-linguistic approach to the pro-drop parameter. Univ. of Southern California PhD.
7. *Havránek B.* 1962. K historickosrovnávacímu poznání syntaxe slovanských jazyků // Otázky slovanských syntaxe. Praha.
8. *Holmberg, Anders.* 2005. Is there a little pro? Evidence from Finnish. Linguistic Inquiry 36, 533–564.
9. *Holmberg A. & C. Platzack.* 1995. The Role of Inflection in Scandinavian Syntax. N.Y: Oxford UP.
10. *Jaeggli O., Safir K.* (eds.) 1989. The null subject parameter. 1989. Dordrecht, Foris.
11. *Lavine J.* 2005. The morphosyntax of Polish and Ukrainian –no/-to // Journal of Slavic Linguistics 13: 75–117.
12. *Mel'čuk I.* 1979. Syntactic, or Lexical Zero in Natural Language //Proceedings of the Berkeley Linguistic Society.Berkeley: UCB, 224–260.
13. *Rizzi L.* 1986. Null subjects in Italian and the Theory of pro. L. 501–557.
14. *Sigurðsson Halldór Ármann.* 2008. Conditions on argument drop — in press.
15. *Sigurðsson Halldór Ármann.* 2008a. The Case of PRO // Natural language and linguistic Theory — in press.
16. *Stabler, E. P.* 1997. Derivational minimalism. In Christian Retore, ed., Logical Aspects of Computational Linguistics. Springer, p. 68–95.
17. *Zimmerling A.* 2007. Zero Lexemes and Derived Sentence Patterns. Wiener Slawistischer Almanach, Sondetrband 69.
18. *Zimmerling A.* 2008. Dative Subjects and Semi-Expletive Pronouns in Russian // Formal Description of the Slavic Languages, FdSL 7 / U.Junghanns, L. Szucsic, G. Zybatow (eds) — in press.
19. *Zubáty J.* 1954. Studie a články, II. Praha.
20. *Мельчук И. А.* 1995. Русский язык в модели Смысл ↔ Текст. Москва-Вена: Языки русской культуры. (Wiener Slawistischer Almanach, Sonderband 39).
21. *Циммерлинг А. В.* 2002.Типологический синтаксис скандинавских языков. Москва.
22. *Циммерлинг А. В.* 2008. Нулевые лексемы в синтаксисе: догматика и типология. Acta linguistica Petropolitana IV, part 2. Санкт-Петербург, 2008, 226–244.

Статистический анализ и контекстуальные правила разрешения графической омонимии при синтезе речи по тексту

Statistical analysis and contextual rules of homograph disambiguation on text-to-speech synthesis

Цирульник Л. И. (L.tsirulnik@newman.bas-net.by)
Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

Барбук С. Г. (sviatos@tut.by)
Минский государственный лингвистический университет, Минск, Беларусь

Лобанов Б. М. (Lobanov@newman.bas-net.by)
Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

Описываются правила определения позиции ударения в омографах, основанные на результатах контекстуального и статистического анализа текстовых корпусов. Разработанные правила используются в системе русскоязычного синтеза речи по тексту «Мультифон» и позволяют повысить степень адекватности смыслового восприятия синтезированной речи.

Введение

При разработке систем синтеза речи по тексту одной из актуальных проблем является разрешение графической омонимии. Как показывает опыт создания системы русскоязычного синтеза речи по тексту, при некорректном определении позиции ударения в слове-омографе и последующем синтезе такого слова затрудняется восприятие смысла всего предложения. В качестве иллюстрации таких ситуаций можно привести отрывки из сказки П.П.Ершова «Конёк-горбунок», в которых неверно установлено ударение в омографах[1]:

«В долгом времени аль вскоре
Приключилося им горе+…»;

«Он и усом не ведёт,
На пе+чи в углу поёт,
Изо всей дурацкой мочи+:
“Распрекрасные вы очи!”»;

«Грива в землю золотая,
В мелки+ кольца+ завитая.»

«По исходе же трёх дней
Двух ро+жу тебе коней…»

Читателю предлагается самостоятельно определить, насколько сложно понять смысл такой фразы, если она произнесена с некорректным ударением в слове-омографе.

Анализу частоты встречаемости омографов в текстах русского языка и выявлению правил определения позиции ударения в омографах посвящена данная работа.

1. Классификация омографов

В основу классификации омографов, используемой для вычисления статистических характеристик, положено разбиение, предложенное в словаре [1]. Согласно этой группировке выделяются следующие классы: разные лексемы одной части речи; раз-

[1] Здесь и далее позиция ударения обозначается знаком «+» после ударного гласного

ные формы одной лексемы; разные лексемы разных частей речи; разные варианты одной лексемы.

В процессе анализа омографов было принято решение на рассматривать класс разных вариантов одной лексемы, который включает в себя следующие подклассы: один из вариантов — профессионализм, один из членов пары — допустимый вариант, один из вариантов — с пометой «в народнопоэтической речи». Очевидно, что для корректной расстановки омографов этого класса недостаточно ни статистических, ни контекстуальных правил, а требуется, в общем случае, глубокий семантико-синтаксический анализ текста.

Кроме того, из перечня омографов были исключены слова — так называемые «ё-омографы». Исследованию их статистических характеристик посвящена статья [2].

Число омографов различных категорий, используемых для проведения эксперимента, представлено в таблице 1.

Таблица 1. Количество омографов, используемых для проведения статистического анализа.

Класс омографов	Количество
Разные формы одной лексемы	1689
Разные лексемы одной части речи	1918
Разные лексемы разных частей речи	323
Общее количество омографов	3930

2. Исходные данные для проведения статистического эксперимента

Для статистического анализа были выбраны текстовые корпуса научного и художественного стилей. В качестве корпуса научного стиля текста использовались доклады конференции «Диалог-2008», содержащие 87 текстов, включающих 237543 слова.

В качестве корпуса художественного стиля текста использовались произведения современных авторов: Б. Акунина, Л. Петрушевской, Д. Рубиной. Корпус включал 8 произведений Б. Акунина общим объёмом 239 460 слов, 56 произведений Л. Петрушевской общим объёмом 87 712 слов и 5 произведений Д. Рубиной общим объёмом 52 105 слов. Общий объём корпуса составил 379277 слов.

3. Проведение статистического эксперимента

Для проведения статистического эксперимента были разработаны специальные программные средства, которые принимают на вход список омографов и текстовый корпус и позволяют вычислять следующие числовые характеристики:

1. n — количество слов в корпусе;
2. m — количество различных омографов в корпусе;
3. m_{10} — количество омографов, встретившееся в текстах более 10 раз;
4. 4. m_1 — количество омографов, встретившееся в текстах только один раз;
5. m_z — общее количество омографов, вычисляемое в соответствии с формулой

$$m_z = \sum_{i=1}^{m} h_i q_i \qquad (1)$$

где h_i — i-тый омограф,
q_i — частота встречаемости i-того омографа в корпусе.

6. p — процентное содержание омографов в текстах, вычисляемое в соответствии с формулой:

$$p_z = \frac{m_z}{n} \times 100\% \qquad (2)$$

7. d — процентное содержание в текстах омографов из входных списков, вычисляемое в соответствии с формулой:

$$d = \frac{m}{k} \times 100\% \qquad (3)$$

где k — количество омографов во входном списке.

Числовые значения k для различных классов омографов приведены выше, в таблице 1.

4. Результаты статистического эксперимента

4.1. Результаты эксперимента для общего перечня омографов

Результаты статистического анализа омографов приведены в таблице 2.

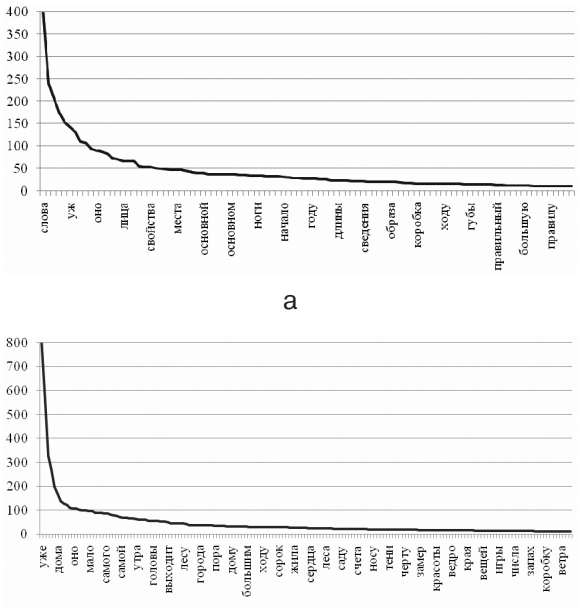
Таблица 2. Результаты статистического анализа общего перечня омографов

Стиль корпуса / Числовые характеристики	Научный	Художественный
n	237543	379277
m	537	1093
m_{10}	104	195
m_1	186	419
m_z	6089	13682
p	2,56 %	3,61 %
d	13,7 %	27,8 %

Как видно из таблицы, количество *различных* омографов, встретившихся в корпусе художественного стиля, более чем в два раза превышает количество *различных* омографов в корпусе научного стиля. В то же время около 50% омографов в корпусе художественного стиля встретилось только 1 раз[2]. В корпусе научного стиля встретилось по одному разу 34% омографов.

Интересно отметить, что, как и ожидалось, процентное содержание слов-омографов среди всех слов корпуса (величина *p*), а также процентное содержание омографов из входных списков (величина *d*) в художественных текстах выше. Такое явление можно объяснить более широким, разнообразным лексиконом текстов художественного стиля.

На рис. 1 иллюстрируется дифференциальное распределение частоты встречаемости омографов. Двадцать наиболее часто встретившихся омографов в каждом из корпусов и количество их появлений в текстах представлены в таблице 3.



а

б

Рис. 1. Дифференциальное распределение частоты встречаемости омографов:
а) в текстах научного стиля;
б) в текстах художественного стиля

Приведённые диаграммы демонстрируют более резкое «падение» частоты встречаемости омографов в корпусе художественного стиля, при этом два самых частотных омографа в художественных текстах — «уже» и «потом» — в большинстве случаев являются служебными частями речи (уже+, пото+м.).

То, что наиболее частотные омографы в корпусе научного стиля — это «слова» и «корпуса», объясняется, пожалуй, лексиконом предметной области исследуемых текстов. Шесть омографов встретились в списках двадцати наиболее частотных как в научных, так и в художественных текстах, а именно, «уже», «второй», «мало», «руки», «оно», «слова».

Таблица 3. Наиболее частотные омографы в исследуемых корпусах

Корпус научного стиля		**Корпус художественного стиля**	
Омограф	**Количество появлений в корпусе**	**Омограф**	**Количество появлений в корпусе**
слова	399	уже	811
корпуса	241	потом	555
уже	209	глаза	328
мало	176	руки	270
связи	153	голову	198
уж	143	дома	168
части	132	самом	137
правило	110	слова	127
правила	107	моя	123
стороны	95	двери	109
Оно	91	оно	108
Тона	87	деньги	108
Методы	83	должно	102
Года	74	ноги	98
Второй	71	второй	98
Лица	67	мало	97
Руки	66	окна	96
Рода	66	стороны	90
Тела	55	тому	88
Числа	54	кому	88

4.2. Результаты эксперимента для различных классов омографов

Результаты статистического анализа для разных форм одной лексемы, разных лексем одной части речи и разных лексем разных частей речи приведены, соответственно, в таблицах 4, 5, 6.

Таблица 4. Результаты статистического анализа разных форм одной лексемы

Стиль корпуса / **Числовые характеристики**	**Научный**	**Художественный**
n	237543	379277
m	155	327
m_{10}	35	73
m_1	45	100
m_z	2275	4059
p	0,96 %	1,07 %
d	9,18 %	19,36 %

[2] Эта величина вычисляется как $(m_1/m)*100\%$

Таблица 5. Результаты статистического анализа разных лексем одной части речи

Стиль корпуса / Числовые характеристики	Научный	Художественный
n	237543	379277
m	251	525
m_{10}	40	51
m_1	100	231
m_z	1746	2690
p	0,74 %	0,71 %
d	13,09 %	27,37 %

Таблица 6. Результаты статистического анализа разных лексем разных частей речи

Стиль корпуса / Числовые характеристики	Научный	Художественный
n	237543	379277
m	100	172
m_{10}	24	48
m_1	27	50
m_z	1387	3879
p	0,58 %	1,02 %
d	30,96 %	53,25 %

Как видно из таблиц, среди всех классов омографов наиболее часто (с учётом частоты появления каждого омографа) в текстах и научного, и художественного стилей встречаются разные лексемы одной части речи: 1746 и 2690 раз соответственно. При сравнении количества *различных* омографов в текстах научного и художественного стилей видно, что наибольшее их количество среди всех классов омографов принадлежит разным формам одной лексемы: 155 и 327 единиц соответственно.

Важно отметить, что процентное содержание омографов, являющихся разными формами одной лексемы и разными лексемами одной части речи, в текстах и научного, и художественного стилей практически одинаково, в то время, как процентное содержание омографов — разных лексем разных частей речи в художественных текстах в два раза больше, чем в научных.

Процентное содержание омографов из входных списков (величина p) для всех классов омографов в художественных текстах практически в два раза больше, чем в научных.

5. Контекстуальные правила

Проведённый статистический анализ позволил выявить омографы, наиболее часто встречающиеся в текстах как научного, так и художественного стилей. Затем был осуществлён экспертный анализ места ударения в таких омографах в конкретном текстовом окружении. Для этих целей использовалась специально разработанная программа, позволяющая извлекать из текстов предложения, содержащие указанные омографы, а также сайт «Национальный корпус русского литературного языка» [3], позволяющий осуществлять поиск по акцентуированному корпусу.

В результате экспертного анализа были сформулированы контекстуальные правила разрешения графической омонимии, приводимые в данном разделе.

5.1. Правила для омографов, которые являются разными формами одной лексемы

Правило 1. Для существительных женского рода третьего склонения в родительном, дательном падежах единственного числа, которые являются омографами форм слова в предложном падеже единственного числа, а также для существительных женского рода третьего склонения в именительном падеже множественного числа, являющихся омографами форм предложного падежа единственного числа, а именно: *бро+ви — брови+*, *гря+зи — грязи+*, *да+ли — дали+*, *кро+ви — крови+*, *ма+зи — мази+*, *но+чи — ночи+*, *свя+зи — связи+*, *те+ни — тени+*, *че+сти — чести+*, ударение в предложном падеже смещается с основы на окончание, если:

а. Существительное входит в несвободное словосочетание.

(1) *Примеры. Дело на мази+, всё на мази+, быть в чести+.*

б. Перед существительным находится предлог «на» (в слове «брови»).

(2) *Пример. На правой брови+.*

в. Перед существительным находится предлог «в» (в словах «дали», «ночи», «связи», «тени»).

(3) *Примеры. В дали+, в глубокой тени+.*

г. Перед существительным находится предлог «на» либо предлог «в» (в словах «грязи», «крови»).

(4) *Примеры. В грязи+, на грязи+; в тёмной крови+, на крови+.*

Исключениями из этого правила являются следующие несвободные словосочетания: «надвинуть на бро+ви», «в неразрывной свя+зи».

Правило 2. Для существительных мужского рода в форме родительного падежа единственного числа, которые являются омографами форм слова в родительном падеже единственного числа, а именно: *ря+да — ряда+, сле+да — следа+, хре+на — хрена+, ча+са — часа+, ша+га — шага+, шу+та — шута+*, ударение смещается с основы на окончание, если:

а. Это существительное второго склонения и непосредственно перед ним находятся количественные числительные «два», «три» или «четыре».

(5) *Примеры. Два ряда+, три следа+, четыре часа+, три шага+.*

б. Непосредственно перед существительным находится предлог «без» (в слове «следа»).

(6) *Пример. Без следа+.*

в. Непосредственно перед существительным или после него находится словосочетание «не осталось».

(7) *Пример. Следа+ не осталось.*

г. Существительное входит в несвободное словосочетание.

(8) *Пример. Ни хрена+.*

д. В паре омографов «шу+та — шута+» ударение падает на основу только в несвободном словосочетании «какого шу+та».

Правило 3. Для существительных мужского рода в форме дательного падежа единственного числа, родительного падежа единственного числа, которые являются омографами форм слова в предложном падеже единственного числа, а именно: *бо+ку — боку+, бы+ту — быту+, ве+тру — ветру+, ви+ду — виду+, гла+зу — глазу+, го+ду — году+, до+лгу — долгу+, до+му — дому+, ду+ху — духу+, за+ду — заду+, кра+ю — краю+, кру+гу — кругу+, ле+су — лесу+, лу+гу — лугу+, мо+згу — мозгу+, ни+зу — низу+, но+су — носу+, пле+ну — плену+, ро+ду — роду+, ря+ду — ряду+, са+ду — саду+, све+ту — свету+, сле+ду — следу+, сне+гу — снегу+, со+ку — соку+, ты+лу — тылу+, хле+ву — хлеву+, хо+ду — ходу+, цве+ту — цвету+, ча+су — часу+, ша+гу — шагу+, шка+фу — шкафу+, я+ру — яру+*, ударение в предложном падеже смещается с основы на окончание, если:

а. Перед ним находится предлог «на» либо предлог «в» (в словах «боку», «заду», «краю», «кругу», «носу», «роду», «следу», «снегу», «ходу», «часу», «шагу», «шкафу», «глазу»).

(9) *Примеры. На левом боку+, в носу+, на роду+, в глубоком снегу+, во втором часу+.*

б. Перед ним находится предлог «в» или «во» (в словах «быту», «долгу», «году», «мозгу», «плену», «ряду», «саду», «соку», «тылу», «хлеву», «цвету», «лесу», «яру»).

(10) *Примеры. В быту+, во вражьем плену+, в яру+.*

в. Перед ним находится предлог «на» (в словах «ветру», «дому», «лугу», «низу», «свету», «духу»).

(11) *Примеры. На сильном ветру+, на дому+, на духу+.*

г. Слово входит в несвободное словосочетание.

(12) *Примеры. Быть на виду+, иметь в виду+, ни в одном глазу+.*

Правило 4. Для существительных мужского и среднего рода в форме родительного падежа единственного числа, которые являются омографами форм слова в именительном падеже множественного числа (всего более 150 пар) ударение смещается на окончание, если перед ним или после него стоит глагол, прилагательное, краткое прилагательное, существительное, притяжательное или возвратное местоимение в форме множественного числа.

(13) *Примеры. Новые слова+, глаза+ смотрели.*

Правило 5. Для существительных мужского и среднего рода *антите+ла — антитела+, во+йска — войска+, дела+ — де+ла, де+ревца — деревца+, зе+ркала — зеркала+, кру+жева — кружева+, кру+жевца — кружевца+, ма+сла — масла+, ме+ста — места+, мо+ря — моря+, мы+ла — мыла+, мя+са — мяса+, о+блака — облака+, о+блачка — облачка+, по+ля — поля+, пра+ва — права+, радиозе+ркала — радиозеркала+, се+рдца — сердца+, сло+ва — слова+, ста+да — стада+, те+ла — тела+, те+льца — тельца+* ударение смещается на основу, если:

а. Перед существительным стоит количественное числительное.

(14) *Пример. Три ста+да.*

б. перед существительным стоит слово «нет» или глагол с отрицательной частицей «не».

(15) *Пример. Если для этого нет сло+ва; не было бы сло+ва; не имеете сло+ва; не знать сло+ва.*

в. перед существительным стоит местоимение в единственном числе.

(16) Пример. Этого во+йска.

Правило 6. Для существительных женского и среднего рода *ви+на — вина+, воло+кна — волокна+, гу+мна — гумна+, доло+та — долота+, ду+пла — дупла+, же+рла — жерла+, ка+йла — кайла+, ко+льца — кольца+, ли+ца — лица+, льноволо+кна — льноволокна+, о+кна — окна+, пи+сьма — письма+, поло+тна — полотна+, полуко+льца — полукольца+, полусу+кна — полусукна+, пя+тна — пятна+, ру+жья — ружья+, ря+дна — рядна+, со+пла — сопла+, стекловоло+кна — стекловолокна+, су+кна — сукна+, тя+бла — тябла+, ха+йла — хайла+, чи+сла — числа+, ядра+ — я+дра, яйца+ — я+йца, я+рма — ярма+* ударение смещается на основу, если перед существительным или после него стоит глагол, прилагательное, краткое прилагательное, существительное, притяжательное или возвратное местоимение в форме множественного числа.

(17) *Примеры. Красные ви+на, разбили о+кна, ли+ца тусклы.*

Правило 7. Для пары омографов *у+тра — утра+* ударение падает на окончание, если перед словом находится предлог «до», «от», «с» или количественное числительное.

(18) *Примеры. Девять утра+, до утра+.*

Правило 8. Для пары омографов *у+тру — утру+* ударение падает на окончание, если перед словом стоит предлог «к».

(19) *Пример. К самому утру+.*

5.2. Правила для омографов, которые являются разными лексемами одной части речи

Правило 1. Для пары омографов *сто+ит — стои+т* ударение падает на основу, если:

а. слово входит в несвободные словосочетания.

(20) *Примеры. Не сто+ит того. Этот парень коечего сто+ит!*

б. слово употреблено в безличном предложении.

(21) *Пример. Мне не сто+ит лезть в штаб-квартиру.*

Правило 2. Для пары омографов *ме+тоды — мето+ды* ударение падает на первый слог, если перед словом или после него стоит прилагательное, притяжательное или возвратное местоимение в форме множественного числа.

(22) *Пример. Математические ме+тоды.*

Правило 3. Для пары омографов *го+лову — голову+* ударение падает на окончание только в словосочетании «*городского голову+*».

5.3. Правила для омографов, которые являются разными лексемами разных частей речи

Правило 1. Для пары омографов *по+том — пото+м* ударение падает на первый слог:

а. В несвободных словосочетаниях «по+том и кровью», «умываться по+том», «умыться по+том».

(23) *Примеры. Доставалось по+том и кровью, добывать по+том и кровью.*

б. Перед или после данного слова находится глагол «покрыться», «покрываться», «облиться», «обливаться», «пахнуть», «осыпать», «осыпаться», «провонять», «залить».

(24) *Пример. После тяжёлой тренировки он обливался по+том.*

Правило 2. Для пары омографов *то+му — тому+* ударение падает на окончание, если:

а. Перед словом стоит предлог «к», а после него частица «же».

(25) *Пример. К тому+ же.*

б. Перед словом стоят предлоги «к», «по», а после них существительное в форме дательного падежа.

(26) *Примеры. К тому+ времени, к тому+ моменту.*

в. После слова стоит запятая, а после запятой одно из слов «что», «чем», «кто», «как», «какое», «с каким», «чтобы», «который», «без чего», «чему», «зачем», «о чём».

(27) *Примеры. Она не привыкла к тому+, чтобы ею командовали.*

г. Перед словом или после него находятся слова «подтверждение», «пример», «свидетель», «объяснение», «причина», «поражаясь», «поверить», «положить», «повинуясь», «накладывая», «содействовать», «верить», «готовясь к», «благодаря».

(28) *Примеры. Объяснением тому+ была плохая погода.*

Правило 3. Для пары омографов *дру+гом — друго+м* :

а. Ударение падает на основу, если слово входит в словосочетания «*друг с дру+гом*», «*друг над дру+гом*», «*друг за дру+гом*»; перед словом находится предлог «с» («со»); перед словом или после него находится местоимение или прилагательное в форме творительного падежа.

(29) Примеры. Со своим дру+гом, хорошим дру+гом был мой сосед.

б. Ударение падает на окончание, если перед словом стоят предлоги «в», «на».

(30) *Примеры. В друго+м помещении, на друго+м берегу.*

Правило 4. Для пары омографов *ми+нут — мину+т* ударение падает на окончание, если перед словом или после него находится количественное числительное, или местоимение, которое его заменяет («несколько», «столько-то», «сколько», «сколько-нибудь»), или существительное «пара».

(31) *Примеры. После двадцати мину+т, на пару мину+т.*

Правило 5. Для пары омографов *са+мом — само+м* ударение падает на основу, если слово входит в несвободные словосочетания «*в са+мом деле*», «*на са+мом деле*».

Заключение

Приведённые правила программно реализованы и применяются в системе русскоязычного синтеза речи по тексту «Мультифон» [4]. Использование описанных правил позволило повысить степень адекватного смыслового восприятия синтезированной речи.

Необходимо отметить, что описанные правила базируются только на лексико-грамматической информации о словоформах и охватывают далеко не все ситуации, встречающиеся в текстах.

Дальнейшие усилия авторов будут направлены на разработку и использование правил разрешения графической омонимии, основанных на результатах синтаксического анализа текстов.

Авторы выражают искреннюю благодарность Елене Ягуновой за предоставление «Словаря омографов русского языка» [1], а также разработчикам «Национального корпуса русского литературного языка» [3].

Литература

1. *Венцов А. В., Грудева Е. В., Касевич В. Б., Корешкова Е. И., Сведенцова Е. А., Ягунова Е. В.* Словарь омографов русского языка // СПб.: Филологический факультет СПбГУ, 2004.
2. *Лобанов Б. М.* Проблема разрешения «Ё»-омографов при синтезе речи по тексту. В данном сборнике докладов.
3. *Электронный ресурс: http://www.narusco.ru/.*
4. *Лобанов, Б. М., Цирульник Л. И.* Компьютерный синтез и клонирование речи // Минск: Белорусская Наука, 2008. — 342 с.

Алгоритм обнаружения ссылочного спама

An algorithm of link spam detection

Шарапов Р. В. (info@vanta.ru), **Шарапова Е. В.** (mivlgu@mail.ru)

Муромский институт (филиал)
Владимирского государственного университета

В статье рассматриваются подходы к выявлению ссылочного спама на основе анализа содержания страницы. Основное место в работе посвящено выявлению рекламных (платных) ссылок. Анализируются признаки, характерные для рекламных ссылок. Дается алгоритм выявления спам-ссылок и приводятся результаты его работы.

1. Введение

Количество информации, доступной пользователям глобальной сети Интернет, с каждым годом становится все больше и больше. Постоянно растет число сайтов, увеличивается число их страниц. Например, поисковая система Яндекс, на момент написания статьи, осуществляла поиск по 17 миллионам сайтов и 4,5 миллиардам веб-страниц [15]. Интернет становится не только средством получения информации и общения, но и средством ведения бизнеса. Естественно, нахождение сайта по наиболее популярным поисковым запросам и положение сайта в списке результатов поиска является актуальной проблемой для большинства владельцев сайтов. В связи с тем, что сайтов одной тематики иногда бывает слишком много, и каждый хочет быть на вершине списка результатов поиска по ключевым запросам, владельцы начинают прибегать к различным ухищрениям, чтобы поднять свои сайты как можно выше к началу списка. Простым выходом кажется применение различных технологий манипулирования поисковыми системами, например с помощью поискового спама. Существует большое количество методик, используемых для того, чтобы ввести в заблуждение поисковые системы [8]. Рассмотрим одну из них — ссылочный спам.

Увеличение числа ссылок на сайты стало одним из основных методов манипулирования поисковыми системами в последнее время. Масштабы манипуляции постоянно растут. Если несколько лет назад основным способом являлся так называемый обмен ссылками, который проводился вручную, то теперь ему на смену пришли различные способы автоматического размещения ссылок. Можно выделить несколько вариантов такого размещения [17]:

1. Использование специализированных программ для автоматического добавления ссылок в каталоги, гостевые книги, форумы и т.д.
2. Покупки ссылок у рекламных брокеров.

С первым вариантом поисковые системы научились бороться, выявляя ресурсы, где есть возможность простого, немодерируемого добавления ссылок. Вес ссылок с таких ресурсов сильно снижается. Размещение же ссылок с использованием рекламных брокеров представляет для поисковых систем много большую проблему.

В настоящее время в русском сегменте интернет действует около десятка крупных рекламных брокеров, занимающихся продажей текстовых ссылок. Только один из них, Sape.ru, имеет возможность размещать ссылки на более чем 55 миллионах страниц (за прошедшие 6 месяцев число страниц увеличилось на 20 миллионов) [11]. Несмотря на то, что ссылки в таких системах называют «рекламными», их основная цель — не реклама с целью привлечения посетителей (ссылки часто размещаются в самых неприметных местах страницы и пользователь их просто не замечает), а улучшение своего положение в поисковых системах. Стоимость такой «рекламы» также часто бывает номинальной, иногда всего 0.01$ за месяц размещения. Ссылки, размещенные с помощью рекламных брокеров, будем называть «платными», подчеркивая что ссылки имеют искусственное происхождение (т. е. не имеют никакого отношения к содержанию страницы). Они размещаются владельцем страницы за деньги, а не из «уважения» к сайту, на который ссылаются.

В чем же основная опасность крупномасштабного ссылочного спама, наблюдаемого последние несколько лет? Опасность заключается в том, что ссылки активно используются современными поисковыми системами для ранжирования результатов поиска. Со ссылками связано и понятия Индекса цитируемости в Яндекс и определение PageRank в Google. Массовое увеличение ссылок неестественного происхождения (ссылочного спама) может сильно «испортить» эффективность их работы. Ситуация осложняется тем, что «платные» ссылки могут размещаться на любых сайтах, в том числе и на очень уважаемых и популярных ресурсах. Таким образом, становится невозможным простое деление страниц на «хорошие» и страницы для ссылочного спама [17].

2. Текущее состояние проблемы

В настоящее время существует несколько подходов к определению поискового спама. Множество работ посвящено анализу ссылочной информации — в первую очередь взаимосвязях страниц, объединяемых ссылками и текстам самих ссылок.

Ряд разработчиков предлагают алгоритмы, построенных на основе PageRank. Например, в работе [7] описывается алгоритм TrustRank для борьбы со спамом. Принцип TrustRank строится на том, что «хорошие» страницы обычно ссылаются на «хорошие» страницы и редко используют ссылки для спама. Сначала выбирается набор «хороших» страниц и им назначается высокий вес. Далее используется подход, аналогичный PageRank: вес разделяется на исходящие ссылки к другим страницам. Наконец, после конвергенции, страницы с высоким весом принимаются за хорошие страницы. Авторы считают, что использование алгоритма TrustRank дает более качественные результаты, чем PageRank.

В работе [4] предлагается алгоритм HostRank (PageRank, вычисленный по графу хостов), который более гибок по отношению к ссылочному спаму. Алгоритм позволяет сократить число сомнительных сайтов в результатах поиска, что достигается уменьшением веса, получаемого сайтами от ссылочного спама.

В работе [1] извлекаются особенности, основанные на связанных образцах сайтов. Кластеризация пространства особенностей позволяет выделить кластеры, сайты которых принадлежат одной и той же группе спам-сайтов.

Для выявления ссылочного спама с помощью Truncated PageRank [2] предлагается анализировать топологию сети ссылок. На основе вычисляемых алгоритмом атрибутов производится классификация ссылок на предмет спама.

В работе [12] предлагается идентифицировать страницы с «ферм ссылок», основываясь на наблюдении, что входящее и исходящее их окружение имеет тенденцию пересекаться. Набор «плохих» страниц многократно расширяется и ссылки между ними отбрасываются.

Другая группа работ основана на анализе содержания страниц.

В [10] рассматриваются различные характеристики страницы (число слов на странице и в заголовке, длина слов, процент видимого текста и т.д.). Проводя сравнение выявленных характеристик с их распределением на «обычных» страницах можно выявить страницы, содержащие спам.

В работе [6] предлагается статистический анализ для выявления автоматически сгенерированных страниц со спамом. Отклонения от нормального распределения различных свойств страниц, включая имена и IP-адреса, входящие и исходящие ссылки, содержание страницы и норму изменения, — все это может свидетельствовать о спаме.

В работе [3] предлагается применять дерево решений для отделения спам-ссылок от обычных.

Таким образом, существующие алгоритмы базируются на анализе структуры сети ссылок, выявлении спамерских страниц и сайтов и т.д. Но существующие алгоритмы практически не предназначены для обнаружения «хороших» и «спамерских» ссылок на каждой отдельной странице [17].

Цель нашего исследования — определение спам-ссылок на любых веб-сайтах, в том числе авторитетных. На каждой отдельной странице могут присутствовать и обычные, и спам-ссылки.

3. Выявление ссылочного спама

Рассмотрим признаки определения рекламных/платных ссылок [5, 9, 17]:

3.1. Ссылки, отмеченные как рекламные объявления

Для этого необходимо просмотреть окрестность ссылки (текст, соседствующий ссылке). Признаки платной ссылки — слова: «Реклама», «Спонсоры», «Наши Партнеры», и т.д.

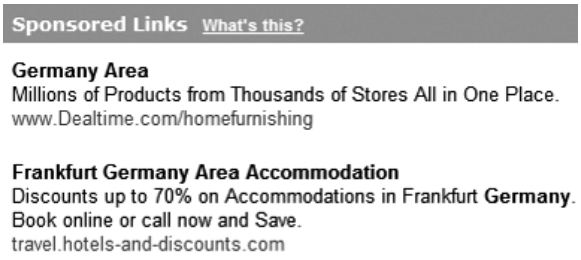


Рис. 1. Пример рекламных объявлений

3.2. Большой блок ссылок

Повышенная плотность ссылок на небольшом участке страницы (блок ссылок) может свидетельствовать об их неестественном происхождении.

3.3. Ссылки на агентства по продаже ссылок/рекламы

Часто вблизи рекламных блоков можно увидеть ссылки на рекламных брокеров.

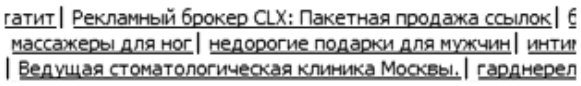


Рис. 2. Пример ссылки на рекламного брокера в блоке ссылок

3.4. На сайте есть информация о том, как можно купить ссылки

Если на сайте или около блока ссылок содержится такая информация, это является фактом, заставляющим усомниться в том, что ссылки являются естественными, а не рекламными.

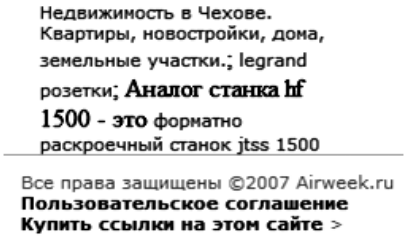


Рис. 3. Пример предложения о покупке ссылок

3.5. Тематическая близость ссылки

Если текст ссылки или тематика сайта, на который ведет ссылка сильно отличается от тематики страницы, на которой ссылка расположена, то ссылку можно считать спамом.

Однако определение тематики ссылки не всегда является тривиальной задачей. Ссылка может располагаться внутри предложения (хотя и не быть частью основного текста страницы). Поэтому ориентироваться на текст в непосредственной близи к ссылке не всегда оправдано.

Агентство дает ипотечный кредит под 9 процентов.

Агентство дает ипотечный кредит под 9 процентов.

Рис. 4. Пример ссылки с текстом по краям

Часто ссылки указывают на ресурс с достаточно общей тематикой (например, при ссылках на источник новостей или сайт автора какой-либо статьи).

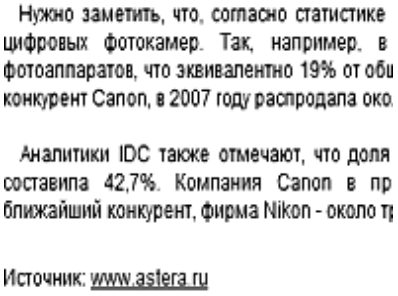


Рис. 5. Пример ссылки на источник

Для правильного определения тематики ссылки может помочь глубокий анализ тематика сайта, на который ведет ссылка. Понятно, что задача эта трудоемка и требует большого количества времени.

3.6. Тематическая близость соседних ссылок

Для этого необходимо проанализировать тематику группы ссылок, размещенных на странице. Если ссылки не являются тематическими и имеют явно выраженный разброс тематики, то они — рекламные.

3.7. Место расположения ссылок

Для этого необходимо проанализировать расположение ссылок на странице. Чем дальше ссылки от основного содержания страницы, тем более вероятно, что они являются рекламными. Например, часто такие ссылки размещаются внизу страницы или в правом столбце, когда основной текст располагается посередине.

3.8. Код ссылок

Многие автоматизированные системы установки ссылок (биржи, обменники, брокеры) устанавливают код автоматически по шаблону. Наличие блока идентичных по коду ссылок может указывать на их спамерское происхождение.

3.9. Динамичность/Время жизни ссылок

Частое изменение ссылок на страницах без изменения остального содержания может свидетельствовать о неестественном их происхождении. Ссылки могут либо просто на время исчезать со страниц (в случае неполадок систем по автоматическому размещению ссылок), либо часть их может заменяться на новые.

3.10. Сообщение о платных ссылках

О платных ссылках могут сообщить конкуренты, бывшие покупатели ссылок, бывшие сотрудники и т.д.

3.11. Просмотр страницы человеком

Просмотр страниц модератором и выявление ссылочного спама вручную.

4. Алгоритм обнаружения ссылочного спама

Теперь рассмотрим алгоритм, способный выявить спамерские ссылки. Он состоит из нескольких этапов.

Этап 1: Формирование предварительного набора спам-ссылок *S*. Набор формируется из следующих ссылок:

- выбранных вручную;
- определенных алгоритмом раннее, как спам;
- определенных на основе анализа кода рекламных брокеров для автоматического размещения ссылок.

Наибольший интерес представляет именно последний способ. У некоторых рекламных брокеров можно обнаружить отличительные особенности в размещении кодов, которые могли бы помочь их идентифицировать. Например, при автоматическом размещении кода Prospero.ru можно заметить следующее:

```
<a class=prospero href="http://www.logipark.
ru">таможенное оформление Япония</a>
<a class=prospero href="http://www.svadbaexclusive.
com/">ЗАГСы Москвы, организация свадьбы
в Москве</a>
```

При кэшировании рекламных ссылок иногда можно увидеть такой код:

```
<!--from cache 14:18:25 13.04.2008-->
<a href="http://www.clinicsex.ru/" target=_
blank>цитомегаловирус затем гарднереллез
анализы мочи</a>
<a href="http://zemnovosti.ru" target=_
blank>Статьи земельная тематика</a>
<!--/from cache-->
```

Оригинальный способ предлагается на [14]. Суть его заключается в том, что ссылки от рекламных брокеров устанавливаются для определенных страниц, и чаще всего с помощью одного и того же кода. Соответственно, рекламный брокер узнает о том, какой код разместить на странице, анализируя строку адреса страницы, например, http://www.site.ru/index.php?cat=1&page=11. Тогда, передав дополнительный параметр (например, http://www.site.ru/index.php?cat=1&page=11&**aa=bb**) можно ввести в заблуждение рекламного брокера, и он не установит рекламные ссылки на страницу. Сравнив содержание страницы в первом и втором случае, появляется возможность выявить платные ссылки.

Еще один метод заключается в отслеживании динамики изменения содержания страницы. Если в течение времени на странице изменяется только группа ссылок, то эта группа может являться платными ссылками. Аналогичные выводы можно сделать, увидев на странице в какой-то момент времени следующее сообщение:

```
<b>Warning</b>: mysql_connect(): Too many
connections in <b>/home/clx/inc/conf.inc</b>
on line <b>56</b><br />
```

Надо заметить, что не все ссылки, определенные алгоритмом как спам стоит заносить в набор *S*, а только те, чьи признаки спама носят явно выраженный характер (чтобы исключить случайного попадания ссылок в разряд спама).

На этапе 1 можно использовать различные алгоритмы классификация и машинного обучения.

Этап 2: Выявление спам-ссылок на основе содержания страницы. Основная идея состоит в анализе содержания страницы и выявлении признаков спама. За каждый признак спама на ссылку налагается штраф q_i. Если суммарный штраф превышает определенный порог, ссылка признается спамом.

Шаг 1. Страница сканируется на наличие ссылок S_b, занесенных в список *S*, сформированный на Этапе 1. При обнаружении таких ссылок сканируется область вокруг них. Если ссылки обнаружены, то им назначается штраф q_1, величина которого снижается по мере удаления от ссылки S_b.

Шаг 2. Страница сканируется на наличие признаков рекламного блока. Признаком могут служить слова «Реклама», «Спонсоры», «Наши Партнеры» и т.д. При обнаружении признаков рекламного блока, ссылкам в его окрестностях назначается штраф q_2.

Шаг 3. Страница сканируется на наличие ссылок на рекламного брокера. При обнаружении таких признаков рекламного блока, ссылкам в его окрестностях назначается штраф q_3.

Шаг 4. Страница сканируется на наличие информации о продаже ссылок (и о том, каких можно купить). При обнаружении таких признаков, ссылкам в их окрестностях назначается штраф q_4.

Шаг 5. Страница сканируются на наличие большого блока ссылок. Если количество ссылок в блоке больше определенного порога, им назначается штраф q_5.

Шаг 6. Ссылки сканируются на признаки кода рекламного брокера, в случае обнаружения которого ссылкам назначается штраф q_6.

Шаг 7. Проверяется соответствие тематики ссылки и общей тематики страницы. В случае несоответствия, ссылке назначается штраф q_7. Для проверки тематики часто бывает достаточно просто просканировать текст страницы на совпадение слов с текстом ссылки.

Шаг 8. Проверяется соответствие тематики ссылки и тематики ссылок в ее окрестностях. В случае несоответствия, ссылке назначается штраф q_8.

Шаг 9. Проверяется место размещения ссылки. Если ссылка находится в самом конце страницы, ей назначается штраф q_9.

Этап 3. Анализ структуры сайта с целью выявления спама. Этот этап является самым сложным. Его цель — выявить особенности структуры сайта и места на страницах, где встречаются «платные» ссылки.

Для этого из страниц сайта удаляется весь изменяющийся контент (кроме ссылок). Далее производится объединение страниц с одинаковым шаблоном в кластеры. Следующий этап: для каждого кластера удаляются повторяющиеся ссылки и идентифицируются области, где ссылки меняются на каждой страницы кластера. Для ссылок, входящих в такие области назначается штраф q_r.

Этап 4. Для каждой ссылки все начисленные штрафы суммируются. Если сумма превышает определенный порог, делается вывод, что ссылка — спам. В этом случае ссылка заносится в список S.

5. Результаты исследований

Для формирования начального набора спам-ссылок S были просканированы 20 сайтов, размещающих платные ссылки (информация о местах размещения платных ссылок были предоставлены нам владельцами сайтов). Число страниц на каждом сайте — от 100 до 5000. После удаления дубликатов был получен набор из 15000 платных ссылок.

В предыдущих исследованиях [17] были реализованы этапы 1,2 и 4. В текущей работе мы продолжили реализацию этапа 3.

При слабо выраженных иных факторах, тематическая близость становится наиболее значимым мерилом спама. Особо это актуально для одиночных ссылок. В наших предыдущих исследованиях [17] мы использовали «наивный» подход для определения тематической близости, основанный на совпадении набора слов. Конечно, такой подход являлся достаточно приближенным и не точным. Кроме того, в рамках наших исследований мы остановились лишь на анализе страницы, на которую ведет ссылка, а не всего сайта. Это позволило ускорить задачу анализа страниц. Недостатком такого подхода явилось возникновение неточностей в определении тематической близости (так как часто ссылки ведут на главную страницу, а не на подраздел сайта). В частности такой подход вызвал ошибочное отнесение ряда ссылок в разряд спама.

Теперь мы усовершенствовали методику определения тематической близости. Некоторые идеи были почерпнуты их [16].

Для оценки качества работы алгоритма использовалась методика, описанная в [2].

$$\text{Precision} = \frac{\text{Число спам-ссылок, отмеченных как спам}}{\text{Число ссылок, отмененных как спам}}$$

$$\text{Recall} = \frac{\text{Число спам-ссылок, отмеченных как спам}}{\text{Общее число спам-ссылок}}$$

$$\text{FalseSpam} = \frac{\text{Число обычных ссылок, отмеченных как спам}}{\text{Общее число обычных ссылок}}$$

$$\text{FalseNotSpam} = \frac{\text{Число спам-ссылок, отмеченных как не спам}}{\text{Общее число спам-ссылок}}$$

Для тестирования были вручную отобраны 100 страниц с числом внешних ссылок от 1 до 30 на каждой. Общее количество ссылок составило 783. Для каждой страницы были вручную отмечены спам-ссылки, которых оказалось 519. В результате работы алгоритма 488 ссылок были отмечены как спам, их которых 461 действительно были спам-ссылками (совпали с отобранными вручную). Результаты оценки качества работы алгоритма приведены в таблице 1.

Таблица 1. Результаты тестирования алгоритма

Precision	0,94
Recall	0,89
FalseSpam	0,102
FalseNotSmap	0,112

Ряд ошибок [17] в выявлении спам-ссылок возникало из-за неглубокого анализа тематической близости. Усовершенствование алгоритма определения тематического подобия позволило повысить Precision на 2%, Recall на 3%, снизить FalseSpam на 5% и FalseNotSmap 3%.

Одиночные рекламные ссылки (в основном, размещенные вручную) близкой к странице тематики не были выявлены как спам. Это связано с тем, что у таких ссылок признаки спама часто отсутствуют. Решением может служить только анализ структуры страниц сайта и выявление мест размещения рекламы (планируется осуществить в будущем), а также анализ времени жизни ссылок, для чего необходим длительный мониторинг страниц.

Таким образом, предложенный алгоритм демонстрирует достаточно неплохие результаты в определении спам-ссылок.

Литература

1. *Amitay E., Carmel D., Darlow A., Lempel R., Soffer A.* The connectivity sonar: Detecting site functionally by structural patterns. In Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia, Aug. 2003, pages 38–47.
2. *Becchetti L., Castillio C., Donato D., Leonardi S., Baeza-Yates R.* Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection. Technical report, DELIS — Dynamically Evolving, Large-Scale Information Systems, 2006.
3. *Davison B. D.* Recognizing nepotistic links on the web. In AAAI-2000 Workshop on Artificial Intelligence for Web Search, Austin, TX, July 30 2000, pages 23–28.
4. *Eiron N., McCurley K. S., Tomlin J. A.* Ranking the web frontier. In Proceedings of the 13th International World Wide Web Conference (WWW), New York, NY, USA, 2004. ACM Press, pages 309–318.
5. *Enge Eric.* 15 Methods for Paid Link Detection http://www.stonetemple.com/blog/?p=167
6. *Fetterly D., Manasse M., Najork M.* Spam, damn spam, and statistics — Using statistical analysis to locate spam web pages. In Proceedings of the 7th International Workshop on the Web and Databases (WebDB), Paris, France, 2004.
7. *Gyöngyi Z., Garcia-Molina H., Pedersen J.* Combating web spam with TrustRank. In Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), Toronto, Canada, 2004.
8. *Gyöngyi Z., Garcia-Molina H.* Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, 2005.
9. *Nash Tim.* How to find a paid link? http://paymentblogger.com/2007/10/07/how-to-find-a-paid-link/
10. *Ntoulas A., Najork M., Manasse M., Fetterly D.* Detecting spam web pages through content analysis. In Proceedings of the World Wide Web conference, Edinburgh, Scotland, May 2006, pages 83–92.
11. *Sape.ru* — главная страница, 2009. http://www.sape.ru/
12. *Wu B., Davison B. D.* Identifying link farm pages. In Proceedings of the 14th International World Wide Web Conference (WWW), 2005.
13. *Кравцов Алексей.* Ссылочный спам: найти и обезвредить http://www.kravcov.ru/2007/03/11/nnueiiue-niai-e-eae-n-iei-aidhiouny/
14. *Детектор* продажных ссылок, 2008. http://venality.name/
15. *Компания* Яндекс, 2009. http://company.yandex.ru/
16. *Некрестьянов И. С.* Тематико-ориентированные методы информационного поиска: Диссертационная работа к.т.н.: 05.13.11 // Санкт-Петербургский государственный университет. СПб., 2000. 80 с.
17. *Шарапов Р. В., Шарапова Е. В.* Обнаружение ссылочного спама // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Десятой Всероссийской научной конференции «RCDL'2008» (Дубна, Россия, 7–11 октября 2008 г.). Дубна: ОИЯИ, 2008. С. 191–196.

Коммуникативы и методы их описания

Communicatives and methods of its description

Шаронов И. А. (Igor_sharonov@mail.ru)

Российский государственный гуманитарный университет

В статье рассматриваются единицы особого рода — короткие аграмматичные стереотипные реплики диалога, или коммуникативы. Перечисляются категориальные и формальные признаки коммуникативов и принципы их конверсационного анализа. Для полноценного описания рассматриваемых единиц необходима по крайней мере следующая информация:

- выражаемая коммуникативом речевая интенция или эмоция;
- способ ее передачи (прямой или косвенный);
- форма высказывания-источника и способы ее преобразования в коммуникатив;
- границы дискурсивной сочетаемости коммуникатива;
- интонационный контур коммуникатива и другие фонетические характеристики.

Некоторые области языка до сих пор слабо поддаются лингвистическому анализу и полноценному лексикографическому описанию. К таковым, например, относятся короткие аграмматичные стереотипные реплики диалога, или коммуникативы. Это единицы типа *ни-ни-ни, пустое, ну уж, окстись, еще чего, дудки, отпад, ни боже мой, ладушки, так и быть* и т.п. Не обладая обычными характеристиками языковых единиц, такими как словоизменение и синтаксические связи, эти единицы за счет устойчивой формы, интонационного оформления, неявных прагматических и дискурсивных свойств активно участвуют в речевом взаимодействии.

Категориальными характеристиками единиц класса коммуникативов являются:

- синтаксическая независимость (это, главным образом, ответные реплики диалога);
- интенциональное значение, которое слабо и неочевидным образом связано с семантикой составляющих коммуникатив компонентов, поскольку такие единицы состоят из служебных и (частично) десемантизированных слов и их сочетаний;
- относительная фиксированность формы;
- устойчивый интонационный контур.

Интенциональное значение, передаваемое коммуникативами, позволяет объединять их в большие группы единиц, выражающих подтверждение, согласие, возражение, отказ, благодарность, клятву, угрозу и т.п. Принципы выявления и сбора единиц, а также идеографическое описание некоторых рядов коммуникативов см. в Шаронов 1996, 1997. Группы интенционально маркированных фразеологических единиц (речевых формул) можно найти также, например, в словаре Баранов, Добровольский 2008.

В традиционной лингвистике понятие *интенции*, или *речевого намерения* в самом общем виде было представлено через классификацию высказываний по цели: утвердительные, вопросительные, побудительные и восклицательные высказывания. Более подробная разработка этого понятия проводится в современных прагматических теориях и теориях речевых актов. Перформативы, побудительные речевые акты, речевые акты извинения, обещания и т. д. описывают интенцию высказывания в социальном контексте коммуникации и межличностного взаимодействия. Важной составляющей прагматического описания высказываний как речевых актов считаются условия их употребления в дискурсе, способствующие, с одной стороны, адекватности понимания речевого акта, а с другой — приемлемости его употребления в дискурсе.

Коммуникативы — особая, с позиций языкового анализа, группа речевых актов. Денотативная опустошенность и аграмматичность обусловливают их зависимость от структуры дискурса и интонационного оформления. Описание коммуникатива требует:

1) выявить речевую интенцию как функцию, которую единица выполняет в структуре диалога;
2) учитывать при описании, как — прямо или косвенно — выражается речевая интенция;
3) выявить значение эталонной формы коммуникатива, для чего восстановить начальную форму высказывания, возвращая утерянные компоненты, и «вынимая» ее из модифицирующих высказывание риторических приемов, таких как ирония, риторический вопрос и т. п.;

4) 4) выявить границы дискурсивной сочетаемости коммуникатива со смежными репликами;
5) 5) выявить устойчивые интонационные контуры и другие фонетические характеристики коммуникатива.

Ниже мы на конкретных примерах проиллюстрируем каждый из перечисленных факторов описания коммуникатива.

1) Выявление речевой интенции, или функции коммуникатива в диалоге

Формальным тестом для выяснения речевой интенции является перевод используемого в диалоге коммуникатива в косвенную речь. При такой операции коммуникатив целиком, без остатка заменяется на предикаты со значением интенции, ментального или эмоционального состояния говорящего. Например, *то-то и оно* — заменяется на: подтвердить что-л.*; еще чего* — на отказать в чем-л.; *ни боже мой* — на *отрицать что-л.*; *так уж и быть* — на разрешить что-л.; *креста на тебе нет!* — на осудить кого-л. за что-л, или на обвинить кого-л. в чем-л. и т. д.

Пересказывая коммуникатив: *И то хлеб!* мы используем глагол удовлетвориться чем-л. На этом основании коммуникатив попадает в группу единиц, выражающих **удовлетворение** (синонимический ряд Слабое удовлетворение): *ну, хоть так; и на том спасибо!; ну и ладно; сойдет (для сельской местности)*. Таким образом осуществляется сбор и первичная классификация коммуникативов в интенциональные ряды, такие как подтверждение гипотезы или оценки; согласие с мнением; согласие, готовность выполнить требуемое действие; отрицание факта; несогласие с мнением или оценкой; отказ выполнить требуемое действие; клятва; удивление; недоумение; удовлетворение результатом и др.

2) Прямая или косвенная форма выражения речевой интенции

Группа собранных по интенциональному признаку единиц может доходить до сотни. При этом единицы различаются по своей структуре и не всегда свободно заменяют друг друга в тексте. Кроме того, мы заметили интересный факт: коммуникативы одной группы могут выстраиваться в реплике последовательной цепочкой. Возьмем в качестве примера несколько коммуникативов подтверждения, согласия с мнением собеседника: *да, конечно, а как же, в том-то и дело*.

Каждая из репликовых единиц способна выражать данный речевой акт самостоятельно.

Он был там? — *Да.*
Он был там? — *Конечно.*
Он был там? — *А как же.*
Он был там? — *В том-то и дело.*

Все перечисленные единицы могут выстроиться при подтверждении друг за другом:

Он был там? — *Да! Конечно! А как же! В том-то и дело*.

Несмотря на то, что единицы синонимичны при выражении подтверждения, не все из них легко поменять местами. В первую очередь это касается позиции коммуникатива *да*. Ср. некоторую неестественность ответных реплик:

Он был там? — ?*Конечно! Да!*
Он был там? — ?*А как же! Да!*
Он был там? — ?*В том-то и дело. Да!*

Кроме того, заметна тенденция коммуникатива *В том-то и дело* занимать в такой цепочке конечную позицию:

Он был там? — ?*В том-то и дело. Конечно!*
Он был там? — ?*В том-то и дело. А как же!*

Причина затруднений при перемещении единиц внутри цепочки может быть объяснена прагматическим принципом построения ответной реплики. Собеседник ждет от говорящего непосредственного подтверждения своей гипотезы. Прямым выражением этого речевого акта является, безусловно, коммуникатив *да*. Поэтому для него естественно занимать первую позицию в списке. Две следующих коммуникатива являются косвенными речевыми актами. Они каждый по-своему аргументируют подтверждение. В. Ю. Апресян (НОСС 2004, с. 476) пишет: «<…>*конечно* сближается по смыслу и функциям с частицей *да,* отличаясь от нее указанием на большую убежденность говорящего и эмоциональностью. *Да* выражает согласие; *конечно* выражает горячее согласие». Наш анализ собранного материала показывает, что ответная реплика *конечно* в нескольких своих дискурсивных разновидностях выражает не только подтверждение, согласие с мнением, но и апеллирует к абсолютной уверенности говорящего. Ее отличие от коммуникатива *Да* — одновременно с подтверждением успокоить, поддержать собеседника, снять его сомнения.

(1) *[Галина:] Мы здесь заживем дружно, верно?*
[Зилов:] ***Конечно.***
[Галина:] Как в самом начале (А. Вампилов. Утиная охота).

(2) *Значит, двадцать первый век, может быть, станет корить нас за то, что мы кого-то проглядели… —* ***Ну конечно.*** *Сплошь и рядом так и бывает.*
(Репетиция воскресенья. Беседа с А. Синявским. Новая газета № 19, 1997).

(3) *[Медведь.] Я уезжаю!*
[Трактирщик.] Это невозможно.
[Медведь.] Я не боюсь урагана.
[Трактирщик.] ***Конечно, конечно!*** *Но вы разве не слышите, как стало тихо?*
(Е. Шварц. Обыкновенное чудо).

Коммуникатив *а как же!* предлагает иную мотивацию для подтверждения: ‘подтверждающий не знает, не может представить альтернативы предложенной гипотезе’.

Вероятно, при классификации косвенных речевых актов необходима определенная иерархия степеней косвенности: косвенность 1-ой степени, косвенность второй степени и т.д. Коммуникатив *в том-то и дело*, кроме значения подтверждения, несет дополнительный смысловой компонент: ‘указание на важность или актуальность высказанной идеи, вывода’, что и заставляет занимать единице в цепочке коммуникативов позицию после прямого акта подтверждения и после речевых актов, передающих подтверждение с «усилительным» компонентом.

Синтагматика коммуникативов служит индикатором при выявлении различий в употреблении единиц одного интенционального ряда. Для толкования коммуникативов необходимым становится сопоставление форм прямой и косвенной передачи интенции и выявление дополнительных смысловых компонентов в косвенных речевых актах.

3) Определение эталонной формы коммуникатива

Достаточно много коммуникативов являются результатом опрощения, приращения, эллипсиса, ассоциативных замен и некоторых риторических преобразований как обычных, так и фразеологизированных высказываний. Для описания особенностей функционирования коммуникатива в дискурсе часто бывает необходимо определить исходное значение начальной формы высказывания-источника.

Перечислим и проиллюстрируем разные виды преобразования языковых высказываний, обращающие их в коммуникативы, а также модифицирующие коммуникативы:

Опрощение

Окстись!	← окрестись!
Спасибо!	← спаси (тебя) Бог!
Во дела!	← вот (так) дела!
Не(е), не-а	← нет.

Приращение

Нетушки	← нет
Фигушки!	← фиг
Спасибочки	← Спасибо.
Ясненько-понятненько	← ясно, понятно.

Эллипсис

Еще чего!	← Еще чего *тебе / ему / ей / вам / им не хватало*!
	← Ещё чего *скажешь / придумаешь / выдумаешь / попросишь*?
Иди ты	← *Да* иди ты *к чёрту*!
А как же	← А как же *может / могло быть иначе*?

Ассоциативные замены

Иди *к чёрту*	→ Иди *к дьяволу / к лешему / на фиг / в болото / в баню* и т. д.
Черта с два!	→ *Фига / хрена* с два!

Ироническое перевертывание

Очень нужно!	← не нужно;
велика радость!	← я не рад;
только попробуй	← даже не пытайся, а то будет плохо.

Риторический вопрос

Это еще зачем?! / Зачем это надо?!	← Незачем это делать / Не надо это делать
Как же иначе / Как же (может быть) иначе?	← Иначе быть не может.

Иногда оказывается, что в результате преобразований в одном коммуникативе воплощаются абсолютно разные высказывания. В качестве примера возьмем коммуникатив из молодежного жаргона *Фигас,* используемый

1) в качестве выражения удивления:

(4) Фигас, опять ничего не успел сегодня сделать. (pashitos.livejournal.com/6382.html).

2) в качестве возражения:

(5) 15.00 — думаете домой пошла? фигас! на консультацию по материаловедению. а она две пары длится (sergeydobrin.ru/verdad/2parte2.html).

Процесс восстановления начальной формы с учетом эллипсиса и ассоциативных замен идет по двум разным направлениям. Ср.:

Удивление: *Фигас* ← *фигасе,* ← *Ни фигас-се* ← *ни фига себе* ← ничего себе.

Возражение: *Фигас* ← *фигас два* ← *фига с два* ← черта с два[1].

4) Выявление границ дискурсивной сочетаемости со смежными репликами

Ориентированность реплик друг на друга в диалоге, взаимообусловленность диалогической пары как стимула и реакции время от времени становятся предметом специального анализа (см. Арутюнова 1970, Падучева 1982, Баранов, Крейдлин 1982. и др.). В работах по конверсационному анализу и анализу диалога традиционно отмечается взаимосвязь диалогических пар ВОПРОС — ОТВЕТ, ПОБУЖДЕНИЕ —

[1] В целях экономии иллюстративные примеры для каждой из перечисленных форм не приводятся. Желающий может обнаружить их в большом количестве по поиску в Yandex.

СОГЛАСИЕ / ОТКАЗ, СОВЕТ — БЛАГОДАРНОСТЬ и т.д. Для адекватного использования коммуникативов требуется более конкретное рассмотрение их дискурсивной сочетаемости и ее границ. Рассмотрим два конкретных случая ограничений, накладываемых на употребление коммуникатива в зависимости от характера и свойств предшествующей реплики.

А. Ограничение на использование коммуникатива: *Обязательно*

Наречие *обязательно* толкуется в МАС (Т. 2) как ‘во что бы то ни стало, непременно’ и иллюстрируется двумя примерами как в плане прошедшего, так и в плане будущего времени:

(6) *В другое время Владик обязательно заспорил бы…* (А. Гайдар. Военная тайна).

(7) *Сейчас же обязательно еще поговорю с главным инженером* (С. Антонов. Первая должность).

В толковых словарях описания репликового употребления *обязательно* обнаружить не удалось, а между тем оно достаточно прихотливо. В позиции ответной реплики *обязательно* используется только для подтверждения полувопросительного высказывания о какой-либо гипотетической ситуации, которая находится в зоне контроля субъекта подтверждения. Чаще всего реплика, стимулирующая использование *обязательно*, описывает ситуацию в будущем времени. Ср.:

(8) *[Галина:] Мы здесь заживем дружно, верно? <…> По вечерам будем читать, разговаривать… Будем?*
[Зилов:] ***Обязательно****.* (А. Вампилов. Утиная охота).

(9) *— Значит, несмотря на некоторые пессимистические прогнозы, «Калина» будет? —* ***Обязательно!*** [П. Меньших, В. Каданников. В Тольятти придут инвесторы // «За рулем», 2003.05.15]

(10) *Боитесь / что он отомстит? № 4. Конечно /* ***обязательно****.* [Беседа в Новосибирске (2003)]

В обыденной нелитературной речи можно встретить использование коммуникатива *обязательно* в контексте прошедшего времени: *Ты ходил туда? — Обязательно!*, что представляется определенным нарушением узуса**.** Также не совсем удачными кажутся высказывания о будущей ситуации, независимой от воли говорящего:

(11) *— Солнце и через сто лет будет гореть так же ярко?*
— Конечно / Разумеется / ***?Обязательно****.*

Б. Ограничение на использование коммуникативов подтверждения со значением мнения

Коммуникативы *наверняка; не иначе; как пить дать*, в отличие от близких им *конечно, естественно, разумеется, ну так* и т.д., являются путативными, и это обстоятельство ограничивает дискурсивную сочетаемость этих единиц. Они используются только в ответ на предположение или гипотетическую, нефактивную оценку собеседника. Ср.:

(12) *— Там леса, наверно, бесконечные!*
— ***Наверняка / Не иначе/ Как пить дать****. Сибирь всё-таки.*

(13) *— Там леса, знаешь, ну просто бесконечные! — Ну так / Естественно /* ****Наверняка / *Не иначе / *Как пить дать****. Сибирь всё-таки.*

5) Выявление устойчивых типов интонационного контура коммуникатива.

Отсутствие предметных значений сближает коммуникативы с междометиями в той роли, которую выполняет в них интонационный контур и некоторые другие фонетические характеристики. При помощи фонетических средств может быть выражена как собственно речевая интенция, так и эмоциональное состояние говорящего лица. Влияние фонетики на формирование смысловых вариаций рассмотрим на примере коммуникатива *Ну да.* Носитель акцента в коммуникативе — частица *да*. Утвердительная интонация произнесения коммуникатива выражает значение *подтверждения*, вопросительная интонация несет значение *удивления*, *неуверенности*, просьбы подтвердить сказанное. Экспрессивная интонация, сопровождаемая насмешливой или пренебрежительной мимикой говорящего, передает ироническое *возражение, несогласие* с мнением собеседника.

Дадим примеры для каждого значения коммуникатива.

- Подтверждение

(14) *— Тех, которых люди хотели читать, не печатали, а тех, которых печатали, никто не читал. Правильно? —* ***Ну да****, — сказал я неуверенно. — Это было, конечно, не совсем так, но в общих чертах* (В. Войнович, Москва 2042).

- Удивление

(15) *— Что она сказала? — Черт те что, — сказала Таиса, вытаращив глаза. — Замуж, говорит, выходит.*
— ***Ну да?*** *— Сказала: замуж* (В. Панова, Конспект романа).

- Ироническое возражение

(16) — *Записался, брат, из попов в купцы […]*
*— Зато грошей будет много! — **Ну да!** Дулю мне под нос, а не гроши* (А. Чехов. Степь).

Частица *ну* модифицирует значение подтверждения, вводя в него компоненты ‘задумчиво’, ‘недоуменно’, ‘озаренно’ в зависимости от формы интонационного контура. В целом же, выполняя хезитационную функцию, эта частица на мгновение оттягивает произнесение *да*, то есть подтверждение, согласие с мнением собеседника, дает говорящему временную паузу для обдумывания услышанного. Причин для паузы, неготовности мгновенно подтвердить может быть несколько:

а) необходимость времени для оценки аргументов собеседника, которыми он обосновывает утверждение;
б) необходимость времени для припоминания чего-либо;
в) необходимость времени для поиска причин переспроса собеседника.

То или иное ментальное состояние говорящего находит отражение в интонации, с которой произносится коммуникатив. Изучение таких интонаций — необходимая составляющая полноценного описания дискурсивных единиц (см. Кобозева 2006).

Дадим примеры для каждого из перечисленных вариаций согласия, подтверждения, выражаемых коммуникативом *ну да*.

А) Подумав, я соглашаюсь с утверждением, оценкой.

(17) — *Так кому же все-таки говорили, что скоты? — А как же их еще называть! Скоты и есть. Совсем бедная женщина, в очереди, а у нее деньги украли, двадцать рублей. <…> — **Ну да**, — сказал он, — понимаю. Действительно* (А. Битов. Пенелопа).

Б) Припомнив, я подтверждаю сказанное.

(18) — *А не снился ли вам приезд государя императора в город Кострому? — В Кострому? Было такое сновидение. Позвольте, когда же это? **Ну да**, третьего февраля сего года…* (И. Ильф и Е. Петров. Золотой теленок).

В) Не понимая причину переспроса собеседника, я подтверждаю то, что уже сказал.

(19) — *Но вообще я биолог и у меня есть свой институт, который занимается созданием нового человека. — Нового человека? — удивился я. — **Ну да,** — подтвердил он. Нового человека.* (В. Войнович. Москва 2042).

Разумеется, перечисленными характеристиками анализ коммуникативов не ограничивается. Для их лексикографического описания могут потребоваться жестово-мимическое описание коммуникатива, стилистические пометы и другие более частные характеристики для каждой рассматриваемой единицы.

В качестве заключения выскажем мысль, освещенную лучом осторожного оптимизма: описание семантики и функционирования коммуникативов — задача крайне трудоемкая, но все же не безнадежная.

Литература

1. *Арутюнова* 1970 — Арутюнова Н.Д. Некоторые типы диалогических реакций и почему-реплики в русском языке // НДВШ, ФН, №3, 1970.
2. *Баранов, Крейдлин* 1992 — Баранов А.Н., Крейдлин Г. Е. Иллокутивное вынуждение в структуре диалога // ВЯ, № 2, 1992. С. 84–100.
3. *Баранов, Добровольский* 2008 — Словарь-тезаурус современной русской идиоматики. Под. ред. А. Н. Баранова и Д. О. Добровольского. М., 2007.
4. *Кобозева* 2006 — Кобозева И.М. Описание означающего дискурсивных слов в словаре: нереализованные возможности // Вестник Московского университета. Сер. 9. Филология. 2006. № 2. с. 37–56.
5. *МАС* — Словарь русского языка. АН. СССР. Т. 1–4.
6. *НОСС* 2004 — Новый объяснительный словарь синонимов русского языка. М., Вена, 2004.
7. *Падучева* 1982 — Падучева Е.В. Прагматические аспекты связности диалога // Изв. АН СССР, СЛЯ, 1982, № 4, с. 305–313.
8. *Шаронов* 1996 — Шаронов И.А. Коммуникативы как функциональный класс и как объект лексикографического описания.// Русистика сегодня № 2, 1996, с. 89–112.
9. *Шаронов* 1997 — Шаронов И. А. Глаголы речевых актов и коммуникативы. // Московский лингвистический журнал, вып. 4. М. 1997, с. 54–72.

Вариативность, продолжение и серийность анекдотов: проблемы построения базы данных

Variation, continuation, and seriality of jokes: problems of database construction

Шмелёва Е. Я. (eshkind@mail.ru),
Шмелёв А. Д. (shmelev.alexei@gmail.com)

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

В статье рассматриваются различные виды вариативности анекдотов, а также интертекстуальные связи между анекдотами. Обсуждаются такие явления, как реализация анекдота, вариант анекдота, продолжение анекдота, модификация исходного анекдота, добавление к анекдоту, серия анекдотов, цикл анекдотов.

1. Вступительные замечания

В ряде докладов на конференции «Диалог», начиная с 1998 года, мы обсуждали различные проблемы, возникающие при создании и индексации базы данных русских (советских и постсоветских) анекдотов. В настоящей статье речь пойдет о вариантах анекдотов, сериях и циклах анекдотов и анекдотах, в которых происходит постепенное увеличение числа персонажей или событий.

Поскольку анекдот — это устный фольклорный жанр, отдельной задачей является выявление «канонического» текста анекдота, который и должен быть представлен в базе данных. «Канонический» текст анекдота практически всегда несколько отличается от реализации этого текста в конкретном речевом действии (при рассказывании анекдота). Но во многих случаях эти отличия не осознаются носителями языка. Например, если один рассказчик рассказывает анекдот, начиная его предложением с глаголом в настоящем времени (*приходит муж с работы, а у жены любовник*), а другой рассказчик рассказывает этот же анекдот, начиная его предложением с глаголом в прошедшем времени (*пришел муж с работы, а у жены любовник*), это, как правило, не ощущается слушателями как варьирование. Такое незначащее (свободное) варьирование анекдотов в базе данных не отражается. В этом случае в базе данных дается максимально нейтральный текст, очищенный от индивидуальных особенностей реализации текста анекдота в том или ином речевом действии[1].

Лишь варианты анекдотов, которые осознаются носителями языка как таковые, подаются в базе данных как варианты, (напр., анекдот поручике Ржевском, который меняет носки «только на водку», и об украинце, который меняет носки «тильки на сало»). Однако здесь встает еще один вопрос, который мы здесь рассматривать не будем: о разграничении вариантов одного анекдота и нескольких разных, хотя и похожих анекдотов.

Мы кратко коснулись этого вопроса в уже упомянутой книге [Шмелева, Шмелев 2002: 115], где отметили, что «решающую роль играют восприятие участников коммуникативной ситуации, отражаемое, в частности, в метатекстовом вводе» (*А был еще похожий анекдот* или *А я знаю этот анекдот в другом варианте*). Разумеется, восприятие разных людей может не совпадать. Так, следующая пара примеров иногда воспринимается как варианты одного анекдота; однако скорее это все же два разных анекдота (поскольку только во второй версии появляется мотив попытки абсурдного отождествления обнаруженного убитого с собеседником и установление невозможности такого отождествления на основе совершенно случайного признака):

[1] Этой тактики мы старались придерживаться в своей книге [Шмелева, Шмелев 2002: 47–63]. Единственное существенное отклонение от него, когда мы на стр. 110 воспроизвели анекдот о раввине, у которого пропал велосипед, вместе со всеми словами-паразитами и другими индивидуальными речевыми особенностями, относящимися к конкретному акту его реализации, вызвало справедливое замечание О.В. Смолицкой в ее рецензии, опубликованной в журнале «Новое литературное обозрение» [Смолицкая 2003]. Впрочем, на стр. 114 нашей книги мы отметили, что данный анекдот мог бы быть рассказан в иной манере (в частности, без «паразитических» словечек *вот*, *типа*).

(1) Проходит в тундре обмен паспортов. Милиционерам неохота объезжать все стойбища, собирать фотографии. Они думают: «Все чукчи на одно лицо. Можно одного сфотографировать, и эту фотографию во все паспорта наклеить». Так и сделали. И вот раздают милиционеры паспорта, один чукча взял, другой, третий. И вдруг один говорит: «Начальник, однако паспорта не моя». Ну, паспортисты испугались, что сейчас их обман обнаружится, давай чукчу уговаривать: «Как это не твой паспорт? Посмотри на фотографию, разве это не твое лицо?» — «Лицо-то моя, начальник, — говорит чукча, — кухлянка, однако, не моя!»

(2) Два охотника постучали в дверь дома чукчи. «Привет, чукча!» — «Привет, однако». — «Мы только что обнаружили в лесу тело убитого и подумали, что, может быть, это ты…» — «А как он выглядит?» — «Да примерно такого же роста». — «В красной фланелевой рубашке?» — «Нет, в коричневой». — «Тогда это не я, однако».

2. Интертекстуальные связи между анекдотами

От варьирования анекдота следует отличать интертекстуальные связи между анекдотами, а также случаи, когда уже существующий анекдот получает продолжение. Интертекстуальные связи между анекдотами были рассмотрены нами ранее (см. [Шмелева, Шмелев 2002: 125]), и суть их состоит в том, что некоторый анекдот не понятен без отсылки к другому анекдоту, который предположительно входит в общий фонд знаний слушателей. В случае если рассказчик сомневается, что исходный анекдот известен данной аудитории, он либо отказывается от рассказывания «вторичного» анекдота, либо предваряет рассказывание вопросом: «Помните, был такой анекдот?» В базе данных исходные анекдоты выступают как самостоятельные единицы, хотя после них имеется отсылка к вторичному анекдоту или вторичным анекдотам. А для вторичных анекдотов с самого начала сообщается, что это вторичный анекдот, и он предваряется отсылкой к исходному анекдоту.

Несколько иной тип интертекстуальных связей обнаруживается, когда исходный анекдот просто получает продолжение. Если в первом случае, рассказчик в норме не рассказывает исходный анекдот, то в случае продолжающихся анекдотов рассказчик рассказывает весь анекдот целиком, добавляя в него новую концовку. Тем самым в этом случае от аудитории не требуется знания исходного анекдота. Правда, в процессе рассказывания может оказаться, что слушатели помнят исходный анекдот и начинают останавливать рассказчика, подозревая, что он пытается рассказать им анекдот «с бородой». В этом случае рассказчик может редуцировать рассказывание исходного анекдота и быстро перейти к концовке[2].

Продолжение анекдота весьма характерно для анекдотов о политике разных правителей или о том, как разные правители ведут себя в сходных ситуациях (при этом последовательность, в которой фигурируют правители, в норме соответствует исторической последовательности их правления). Появляется новый правитель, его правление характеризуется теми или иными особенностями, и анекдот получает продолжение.

В анекдотах с продолжением также может иметь место варьирование. Так, в начале 1960-х гг. рассказывался (с незначительным варьированием) следующий анекдот о том, на каком транспорте советский народ двигался к коммунизму:

(3) При Ленине — как на поезде через туннель: вокруг темно, а впереди свет. При Сталине — как в трамвае: один ведет, а остальные — кто сидит, кто трясется. При Хрущеве — как в аэроплане: один ведет, всех остальных тошнит, а выйти некуда.

Позднее у анекдота появилось продолжение:

(4) При Брежневе — как в такси: чем дальше, тем дороже.

Однако в то же самое время существовал и другой вариант продолжения этого анекдота, зафиксированный, напр., в известной книге [Штурман, Тиктин 1987: 165]:

(5) Теперь — как в космосе. Неизвестно зачем, неизвестно куда и опереться не на что.

Упомянем также анекдот о том, как поезд, идущий в коммунизм стал, потому что кончились рельсы, и как на это реагировали разные руководители (варианты этого анекдота были приведены в книге [Шмелева, Шмелев 2002: 129] и в статье [Шмелева, Шмелев 2007: 239–240]). Первоначально анекдот заканчивался реакцией Хрущева, приказавшего

[2] В некоторых случаях исходный анекдот всегда цитируется полностью. Во время «перестройки» возник анекдот: *Сначала будет перестройка, затем перестрелка, затем перепись оставшегося населения*. Когда в 2002 г. проводилась перепись населения России, появилось продолжение анекдота, которое рассказывалось в следующем виде: *Помнишь, во время перестройки был такой анекдот, что сначала будет перестройка, затем перестрелка, затем перепись оставшегося населения. Ура, прорвались!*

разбирать рельсы сзади поезда и класть их спереди, а затем с каждым новым руководителем (в некоторых вариантах Андропов и Черненко пропускались) анекдот получал продолжение.

В базе данных анекдоты «с продолжением» подаются как одно вхождение, но с соответствующей разметкой; приводится исходный анекдот (разумеется, указываются варианты «канонического» текста), а затем указываются его продолжения. Если в анекдоте есть факультативные части, то они особым образом помечаются.

3. Модификация исходного анекдота

Иногда появление продолжения приводит к тому, что «исходный» вариант анекдота несколько видоизменяется. Проиллюстрируем это на примере анекдота о том, как разные руководители комментировали сказку о мухе-цокотухе. Ряд вариантов этого анекдота приведен со ссылкой на сборник [Krikmann 2004] и проанализирован в статье [Неклюдов 2007]. Приведем его в том виде, в каком мы слышали его в начале 1984 г. (отвлекаясь от индивидуальных особенностей конкретных актов рассказывания, т. е. реконструируя «канонический вариант):

(6) Приходит Корней Чуковский к Ленину. «Владимир Ильич! Я стихотворение написал. Хотел бы его опубликовать». — «Ну, читайте». — «Муха, Муха, цокотуха, / Позолоченное брюхо, / Муха по полю пошла, / Муха денежку нашла. / Пошла Муха на базар / И купила самовар…» — «Стоп, стоп. Това'ищ Чуковский! Почему на база', а не в коопе'атив? Это политическая ошибка. Пе'епишите стихотво'ение!» Приходит Чуковский к Сталину. «Товарищ Сталин! Я стихотворение написал, хотел бы его опубликовать». — «Ну, читайте». — «Муха, Муха, цокотуха, / Позолоченное брюхо, / Муха по полю пошла, / Муха денежку нашла…» — «Стоп, стоп. Таварыш Чуковский! У нас дэнги на полэ нэ валяются. Перепишите стихотворение». Приходит Чуковский к Хрущеву с той же просьбой. Начинает читать: «Муха, Муха, цокотуха, / Позолоченное брюхо, / Муха по полю пошла…» — «Стоп, стоп. Товарищ Чуковский! Если каждый будет шастать по колхозному полю, у нас кукуруза не уродится. Исправьте». Приходит Чуковский к Брежневу с новой редакцией стихотворения. «Муха, Муха, цокотуха, / Позолоченное брюхо…» — «Стоп, стоп. Товарищ Чуковский! У нас в стране каждый грамм золота на счету, а у вас какая-то муха с позолоченным брюхом. И что это значит — позолоченное брюхо?! На кого вы намекаете? Чем это вам наши Герои Социалистического Труда мух напоминают?! Перепишите». Приходит Чуковский к Андропову. «Юрий Владимирович! Никак не могу опубликовать стихотворение, помогите». — «Ну, читайте». — «Муха, Муха, цокотуха…» — «Какая такая ЦэКатуха? Что вы там про ЦК сказали?!»

Разумеется, для рассказывания анекдота совершенно безразлично то, что на самом деле Корней Чуковский умер раньше Брежнева, так что к Андропову он приходить никак не мог: для анекдотов смешение временных планов является нормой [Шмелева, Шмелев 2002: 14].

Краткое правление Черненко, насколько нам известно, не привело к появлению у этого анекдота каких-либо продолжений, а при Горбачеве продолжение появилось:

(7) Наступила гласность. Идет Чуковский к Горбачеву и говорит: «Я стихотворение написал, а мне все не давали опубликовать». — «Читай». — «Муха, муха…» — «Это кто там под мухой? Мы тут с пьянством боремся! Нет, не пойдет».

Мы видим, что текст, с которым Чуковский приходит к каждому следующему вождю, все более сокращается, и кажется, что на Горбачеве уже достигнут предел. Тем не менее в 1990-е гг. появилось дальнейшее продолжение этого анекдота:

(8) Развалился Советский Союз. Идет Чуковский к Ельцину и говорит: «Я стихотворение написал, а коммунисты запрещали публиковать». — «А нам вообще поэты не нужны!»

И, наконец, уже в годы президентства Путина, после принятия «нового старого гимна» у анекдота появилось несколько неожиданное продолжение:

(9) Приходит Михалков к Путину и говорит: «Я стихотворение написал, никак опубликовать не могу». — «Читайте!» — «Муха, муха, цокотуха, / Позолоченное брюхо. / Муха по полю пошла, / Муха денежку нашла. / Пошла муха на базар / И купила самовар!» — «Замечательные стихи!» И дал Путин Михалкову медаль за прославление рыночной экономики и недр России, которые богаты золотом, нефтью и газом.

Появление Михалкова должно быть как-то мотивировано; поэтому, как правило, при рассказывании анекдота в этой, самой последней версии с самого начала говорится: *Написали Чуковский и Михалков стихотворение, приходят к Ленину…* Далее в какой-то момент (чаще всего — начиная с Брежнева) Михалков начинает приходить к вождям один, но все же говорит: *Мы тут с Чуковским стихотворение написали*; но в заключительном эпизоде, в разговоре с Путиным вместо местоимения *мы* он уже

говорит: *я*. Здесь можно видеть намек на историю советского (resp. российского) гимна: в качестве автора «сталинской» и «брежневской» версий указывались Михалков и Эль-Регистан, а автором «путинской» версии является уже один Михалков. Примечательно, что при таком продолжении новыми красками играет и фраза анекдотического Ельцина «нам вообще поэты не нужны»: все помнят, что при Ельцине гимн России был без слов (ср. [Неклюдов 2007: 39–40]).

4. Добавления к исходному анекдоту

Несколько иной тип продолжений анекдота имеет место в тех случаях, когда к анекдоту добавляется новая концовка. В этом случае «соль» анекдота может состоять в обманутых ожиданиях слушателей, которые считают, что они знакомы с анекдотом — но выясняется, что они не знают самого главного (а слушатели, незнакомые с исходным анекдотом, могут считать, что он закончен, но выясняется, что самое главное еще впереди). Некоторые примеры таких добавлений были приведены в нашей книге [Шмелева, Шмелев 2002: 124–125, 128–129]. Приведем еще пример. Во второй половине 1920-х гг. рассказывали следующий анекдот:

(10) В какие игры играют кремлевские вожди? — Рыков — в пьяницу, Крупская — в Акульку, Сталин — в короли, а Калинин — в дурака.

Ср. иную версию этого анекдота, с добавлением:

(11) В какие игры играют кремлевские вожди? — Рыков — в пьяницу, Крупская — в Акульку, Сталин — в короли, Калинин — в дурака, а Ленин уже сыграл в ящик.

Иногда добавление может быть значительным, равным по продолжительности исходному анекдоту. Так, известен (и часто цитируется) анекдот о трех путешественниках, попавших в плен к людоедам (в другом варианте — к инопланетянам):

(12) Поймали дикари трех путешественников: американца, индуса и русского — и говорят: «Кто сможет удивить и развеселить нашего вождя, того отпустим, а остальных — съедим!» Дают они каждому по два титановых шарика и сажают на ночь в бетонный бункер. Утром вождь делает обход. Заходит к американцу, а у него шарики по полу катаются, сталкиваются, выделывают разные фигуры. Вождь посмотрел: не интересно! Ну и сожрали его. Заходит к индусу — а тот сидит медитирует, шарики у него вокруг головы летают. Не интересно! Тоже пошел на суп. Заходит к русскому: через две минуту вываливается из камеры от хохота. У вождя спрашивают: что ж он такое придумал? Вождь: «Да этот придурок один шарик потерял, а второй — сломал!!!»

В какой-то момент к этому анекдоту было сделано добавление (впрочем, по общему мнению, оно уступает исходному анекдоту):

(13) Через какое-то время снова поймали те людоеды трех путешественников: на этот раз итальянца, англичанина и опять русского. Ну и дали им титановые шарики на тех же условиях, то есть того, кто сделает с ними что-то для них удивительное, отпустят. Только для разнообразия не два, как в прошлый раз, а три. Итальянец шариками жонглировать начал, чем дикарей удивил: тремя шариками жонглировать гораздо труднее, чем двумя. Англичанин игру настольную изобрел типа бильярда: оказалось, что тремя шариками играть гораздо интереснее. Но отпустили все же русского. То, что он один шарик потерял, а другой сломал, дикарей не особо удивило: в прошлый раз то же самое было. Но вот как он, находясь в бетонном бункере, умудрился третий шарик пропить — этот вопрос мучит их до сих пор.

Часто добавление делаются к мультинациональным анекдотам — как разные народы ведут себя в тех или иных ситуациях, как невероятных, так и вполне реальных: на курорте или на необитаемом острове, при встрече с женой или с джинном. Число народов может варьироваться рассказчиком в зависимости от аудитории или увеличиваться, если по тем или иным причинам актуализуются отношения с каким-то народом, но последним в норме все равно оказывается родной народ рассказчика, реакция и поведение которого и составляют пуанту анекдота. Однако в русских анекдотах после русских иногда появляются представители еще какого-то народа (чаще всего евреи). Так, анекдот о том, что говорят представительницы разных народов после первой брачной ночи, в исходном варианте заканчивался фразой русской: *А душу мою ты так и не понял!* Добавление к анекдоту — фраза еврейки: *Тут болит, там болит!* [Шмелева, Шмелев 2002: 82, 129]. Ср. аналогичный пример:

(14) У англичанина есть жена и любовница, но любит он только жену. У немца есть жена и любовница, но любит он любовницу. У француза есть жена и любовница, а любит он обеих. У русского есть жена и любовница, а любит он водку.

Добавление к анекдоту:

(15) У еврея есть жена и любовница, а любит он маму.

Сходным образом получил добавление и следующий анекдот советского времени:

(16) Один англичанин — джентльмен, два англичанина — пари, три англичанина — парламент. Один француз — любовник, два — дуэль, три — революция. Один немец — солдат, два немца — два солдата, три немца — три солдата. Один русский — пьяница, два — драка, три — первичная парторганизация.

Добавление:

(17) Один еврей — завмаг, два — матч на первенство мира по шахматам, три — русский симфонический оркестр.

Реже добавление бывает связано с каким-то другим народом, как в анекдоте о книгах, которые представители разных народов публиковали о слонах. Исходный анекдот завершался сообщением о советском трехтомнике (название первого тома стало крылатой фразой — «Россия — родина слонов»); однако затем анекдот получил добавление, в соответствии с которым в Болгарии был перепечатан советский трехтомник и к нему был добавлен четвертый том — «Болгарский слон — младший брат советского слона» [Шмелева, Шмелев 2002: 124]. Ср. также изредка встречавшиеся в советское время добавления к вышеприведенному анекдоту об одном, двух и трех представителях разных народов:

(18) Один украинец — разбойник, два — партизанский отряд, три — партизанский отряд с предателем.

(19) Один грузин — Сталин, другой грузин — Берия, третьего пока не было.

Иногда добавление представителей некоторого народа бывает связано с тем, что анекдот имеет хождение в соответствующей среде. Приведем версии анекдота, в которых фигурируют один, два и три представителя разных народов, имевшие хождение в среде русских, живущих в Молдавии и Венгрии (но получившие некоторое распространение и в российской городской среде):

(20) Один русский — пьяница, два — драка, три — первичная партячейка. Один еврей — завмаг, два — матч на первенство мира по шахматам, три — оркестр молдавских народных инструментов. Один молдаванин — дурак, два молдаванина — два дурака, три молдаванина — Молдавская академия наук.

(21) Один француз — бабник, два — дуэль, три — революция. Один немец — педант, два — завод, три — война. Один русский — пьяница, два — драка, три — администрация. Один еврей — завмаг, два — матч на первенство мира по шахматам, три — Русский национальный банк. Один венгр — просто венгр, два — политическая партия, три — не бывает, один из них либо еврей, либо немец.

В базе данных анекдоты с добавлениями подаются так же, как анекдоты с продолжением, т. е. как одно вхождение, но с соответствующей разметкой.

5. Серийные анекдоты

Длинный анекдот, возникший в результате соединения ряда однотипных добавлений, может превращаться в серию анекдотов. Так, приведенный выше анекдот состоящий из однотипных частей, построенных по схеме «Один X — это …, два X-а — это …, три X-а — это …» может рассматриваться как серия однотипных анекдотов. С другой стороны, однотипные серийные анекдоты могут соединяться в один длинный анекдот. Так, серия анекдотов, в которых армянскому радио задается вопрос: «Можно ли построить коммунизм в Армении, Израиле, Швейцарии и т. п.?», иногда объединяется в один анекдот, который завершается вопросом: «А можно ли вообще построить коммунизм?» (ответ: «Построить-то можно, но выжить при нем нельзя!»). Вообще говоря, не так важно, имеет ли место соединение однотипных анекдотов в один или разбиение длинных анекдотов на серию однотипных анекдотов. Поскольку держать в памяти и рассказывать длинные анекдоты бывает «неудобно», в некоторых случаях объединение анекдотов происходит преимущественно на письме (например, в Интернете), а люди рассказывают отдельные короткие анекдоты. Но поскольку во многих компаниях принято рассказывать анекдоты один за другим (есть даже особое название для такого рассказывания анекдотов «травить анекдоты»), то часто один анекдот серии «цепляет» за собой другие.

Главное отличие серии однотипных анекдотов и длинного анекдота, состоящего из однотипных частей, заключается в следующем. Серийные анекдоты «равноправны» и в ситуации, когда «травят» анекдоты, они могут следовать друг за другом в каком угодно порядке (причем разные анекдоты серии могут рассказываться разными людьми), а анекдот, состоящий из однотипных частей, предполагает «ударную» концовку и тем самым порядок следования его частей не является произвольным. Поэтому, если нет анекдота, претендующего на роль «ударного», завершающего, то мы имеем дело с серией

(как в случае серии «Сколько надо Х-ов, чтобы вкрутить лампочку?»), если есть — то это скорее длинный анекдот с ударной концовкой. Однако в сериях обычно есть некоторое число исходных анекдотов, а затем могут появиться вторичные анекдоты, предполагающие знание исходных.

Серию анекдотов следует отличать от цикла. Анекдоты объединяются в цикл по персонажам. Каждый анекдот цикла автономен, хотя предполагает, что слушатели имеют какое-то представление о свойствах основного персонажа. Поэтому анекдоты одного цикла, как правило, не соединяются в один длинный анекдот. Анекдоты одной серии имеют схожее формальное строение и, как правило, рассказываются друг за другом; поэтому серия анекдотов часто воспринимается как один длинный анекдот и именно таким образом передается при письменной фиксации.

Иными словами, анекдоты одного цикла — это анекдоты с постоянными персонажами, но имеющие совсем разную структуру («анекдоты о Штирлице», «анекдоты о Вовочке», «анекдоты о чукче», «анекдоты о новых русских» и т. д.), а анекдоты одной серии — это анекдоты тождественной структуры, но с переменными персонажами. Этим и определяется их различное представление в базе данных.

6. Заключительные замечания

Разграничения и понятия, обсуждавшиеся в данной статье (разные реализации одного анекдота, варианты одного анекдота, похожие анекдоты, интертекстуальные связи между анекдотами, продолжение исходного анекдота, модификация исходного анекдота, добавление к исходному анекдоту, серия анекдотов, цикл анекдотов) могу показаться схоластическими. Однако они необходимы для принятия решений при создании и индексации базы данных русских анекдотов.

Литература

1. *Неклюдов С. Ю.* Происхождение анекдота: «Муха-цокотуха» под судом советских вождей // Post-Socialist Jokelore — Постсоциалистический анекдот. International symposion — Международный симпозиум. January 15th–16th 2007 — 15–16января 2007. Tartu, 2007. С. 37–42 (препринт).
2. *Смолицкая О. [рец.:] Шмелева Е. Я., Шмелев А. Д.* Русский анекдот: текст и речевой жанр. М., 2002 // Новое литературное обозрение, 2003, № 64. С. 377–380.
3. *Шмелева Е. Я., Шмелев А. Д.* Русский анекдот. Текст и речевой жанр. // М.: Языки славянской культуры, 2002.
4. *Шмелева Е. Я., Шмелев А. Д.* Анекдот в современной русской речи: интертекстуальные связи // Вопросы культуры речи. Вып. 9. М.: 2007. С. 226–242
5. *Штурман Д., Тиктин С.* (сост.) Советский Союз в зеркале политического анекдота. // Иерусалим: Экспресс, 1987.
6. *Krikmann A.* (ред. и сост.) Netinalji Stalinist — Интернет-анекдоты о Сталине — Internet humor about Stalin. Арво Крикманн. Тарту, 2004.

Разрешение синтаксической неоднозначности: эффекты прайминга и самопрайминга

Syntactic ambiguity resolution: priming and self-priming effects

Юдина М. В. (dietiefe@yandex.ru)
ABBYY Software, Московский государственный университет им. М. В. Ломоносова

Фёдорова О. В. (olga.fedorova@msu.ru)
Московский государственный университет им. М. В. Ломоносова

В докладе представлен первый опыт экспериментального исследования синтаксического прайминга на конструкциях с относительным придаточным в русском языке. В рамках синтаксического прайминга выделяются два эффекта: собственно прайминг (настройка на заданную конструкцию) и самопрайминг (настройка на собственную стратегию).

Синтаксический прайминг представляет собой одно из проявлений более общего феномена прайминга (или преднастройки, подробнее про терминологию см. [Величковский 1982], [Фаликман, Койфман 2005]). Праймингом называется «изменение скорости или точности решения задачи после предъявления информации, связанной с содержанием или контекстом этой задачи, но не соотносящейся прямо с ее целью, а также повышение вероятности спонтанного воспроизведения этой информации в подходящих условиях.» ([Фаликман, Койфман 2005]). Явление прайминга относят к имплицитной памяти, его использование дает возможность ответить на вопрос о том, насколько глубоко анализируется информация об объектах.

Имплицитная память находится вне контроля сознания. Её существование подтверждается нейрофизиологическими исследованиями на больных с амнезией: испытуемые значительно лучше выполняли задачи распознавания, если объект находился в их имплицитной памяти, в то время как вследствие болезни объект не мог при этом находиться в эксплицитной памяти испытуемых.

Выделяется много типов языкового прайминга: семантический, лексический, морфологический, синтаксический. В рамках данной статьи мы рассмотрим явление синтаксического прайминга и эффекты, которые он может оказывать при разрешении синтаксической неоднозначности.

Явление синтаксического прайминга заключается в следующем: при ответной реакции на какой-либо стимул говорящий склонен использовать те синтаксические конструкции, которые он в недавнем прошлом каким-либо образом обработал (услышал, прочитал, сказал). Одним из типичных проявлений синтаксического прайминга является синтаксическая координация участников диалога. Высказывание, осуществляющее преднастройку, называется «праймом», а высказывание, на порождение или понимание которого, как предполагается, окажет влияние прайм, называют «целью».

Среди используемых методик в первые годы изучения СП преобладали методики, в которых исследовалась речевая продукция одного человека (within-subjects design): например, в работе [Bock 1986] испытуемые сначала повторяли предложения-праймы, а потом описывали картинки-цели. Такой синтаксический прайминг называется «самопраймингом».

В 2000 г. авторами работы [Branigan et al. 2000] была разработана специальная процедура для изучения синтаксического прайминга в диалоге (between-subjects design). Согласно последним работам, в диалоге прайминг-эффекты статистически значимо сильнее, чем при автопрайминге (например, [Фёдорова 2002]).

В работе [Scheepers 2003] описаны первые эксперименты на синтаксический прайминг неодно-

значных синтаксических конструкций, а именно, конструкций с относительным придаточным, известных также как конструкции с «ранним-поздним закрытием». Феномен раннего-позднего закрытия заключается в следующем: это неоднозначность сложноподчиненных предложений с относительными придаточными, например, Преступник застрелил служанку актрисы, которая стояла на балконе. Данное предложение можно понять двояко: при одном понимании придаточное предложение относится к вершине именной группы, далее ИГ ('служанка стояла на балконе'), это мы называем «ранним закрытием» (далее РЗ), а при другом — к зависимому члену ИГ ('актриса стояла на балконе'), это мы называем «поздним закрытием» (далее ПЗ). Раннее-позднее закрытие широко изучалось на материале различных языков ([Frazier, Fodor 1978], [Frazier, Clifton 1997], [Fodor 1998], [Sekerina, Fedorova 2004], [Юдина 2006] и др.), такой интерес был вызван тем, что первоначальная гипотеза об одинаковых механизмах разрешения неоднозначности в разных языках не подтвердилась проведенными экспериментами: например, в работе [Cuetos, Mitchell 1988] показано, что англоговорящие предпочитают ПЗ, а испаноговорящие — РЗ. Подобные различия стимулировали дальнейший поиск факторов, влияющих на разрешение данной синтаксической неоднозначности.

Эксперимент [Scheepers 2003] проводился на немецком материале по методике завершения предложений, экспериментальный блок состоял из четырех предложений: прайм с ранним закрытием (далее РЗ-прайм), прайм с поздним закрытием (далее ПЗ-прайм), базовый прайм (блокировал присоединение относительного придаточного) и целевое предложение (далее цель). Суть эксперимента заключалась в следующем: РЗ- и ПЗ-праймы могли быть продолжены испытуемыми только одним способом, базовый прайм не предполагал конструкции с относительным придаточным, а цель была составлена так, что она могла быть закончена испытуемыми двояко в силу омонимии относительного местоимения. Таким образом, предполагалось, что вынужденное использование испытуемым РЗ или ПЗ в прайме вызовет использование соответствующей структуры в синтаксически неоднозначном предложении — цели. Базовые праймы включались в экспериментальные наборы для того, чтобы проверить, каково будет предпочтение закрытия в цели в отсутствие прайминга. Результаты эксперимента показали значительный прайминг-эффект (испытуемые присоединили придаточное предложение к первому существительному ИГ в 44% после РЗ-прайма, в 29% после ПЗ-прайма и в 33% случаев после базового прайма).

Мы провели серию аналогичных экспериментов на русском материале. Первый эксперимент ([Юдина 2007]) был сконструирован на основе экспериментального дизайна работы [Scheepers 2003], стимульный материал в значительной степени состоял из стимульного материала уже проведенных экспериментов на раннее-позднее закрытие ([Юдина 2006]).

Приведем пример экспериментального блока:

РЗ-прайм: *Толпа с удивлением рассматривала переводчицу министра, которая…*

ПЗ-прайм: *Толпа с удивлением рассматривала переводчицу министра, который…*

Базовый прайм: *Толпа с удивлением рассматривала переводчицу министра, потому что…*

Цель: *Пассажир попытался позвать напарника стюардессы, …*

Из-за особенностей русской морфологии нам пришлось убрать из цели местоимение «который», чтобы испытуемые могли ее продолжить, исходя из типа закрытия прайма или исходя из собственной стратегии закрытия.

Эксперимент I, вопреки ожиданиям, не показал значимого эффекта синтаксического прайминга в силу того, что из-за особенностей экспериментального дизайна испытуемые подвергались не только действию синтаксического, но и морфологического прайминга ([VanWagenen, Pertsova 2005]) формой местоимения «который»: «который» (мужской род) или «которая» (женский род), поскольку выбранная для закрытия цели форма местоимения морфологически совпадала с одним праймом, а синтаксическая структура, образованная местоимением в такой форме, соответствовала другому прайму. Таким образом, в ряде случаев оказалось невозможно указать причину выбранного испытуемым типа закрытия: это мог быть и морфологический, и синтаксический прайминг. Кроме того, Эксперимент I показал очень большой процент предпочтения РЗ, выше, чем в других экспериментах на раннее-позднее закрытие на русском материале ([Юдина 2006], [Sekerina 2003], [Sekerina, Fedorova 2004]) (см. табл. 1)

Для того, чтобы убрать конкурирующий эффект морфологического прайминга, мы сконструировали новый экспериментальный дизайн. В отличие от Эксперимента I, в целевых предложениях Экспериментов II и III данного исследования было эксплицитно введено местоимение «который» в форме «которым», равновероятно относящееся либо к творительному падежу вершины ИГ, стоящей в единственном числе, либо к дательному падежу зависимого члена ИГ, стоящего во множественном числе. Поскольку структура праймов осталась прежней (то есть РЗ-прайм содержал местоимение который в форме, согласующейся с вершиной ИГ по роду и в именительном падеже, а ПЗ-прайм — местоимение который в форме, согла-

сующейся с зависимым членом ИГ по роду и в именительном падеже), эффект морфологического прайминга был полностью исключен.

Таким образом, экспериментальный блок Экспериментов I и II выглядел следующим образом:

РЗ-прайм: *На заседании утвердили бюджет организаций, который …*

ПЗ-прайм: *На заседании утвердили бюджет организаций, которые …*

Базовый прайм: *На заседании утвердили бюджет организаций, и оказалось …*

Цель: *Власти решили учредить конкурс газонокосилок, которым …*

Эксперименты II и III показали значительный эффект синтаксического прайминга:

Таблица 1. Сравнительные результаты Экспериментов I, II, III

Эффект синт. прайминга	**Эксперимент I**	**Эксперимент II**	**Эксперимент III**
РЗ после РЗ-прайма — РЗ после ПЗ-прайма, %	78–75,7	57–46,8	82–36
РЗ после базового прайма, %	83,2	61	62,6

Из таблицы видно, что результаты Эксперимента II близки к результатам, полученным в работе [Scheepers 2003], однако, результаты Эксперимента III показывают еще более высокий процент синтаксического прайминга.

Разница между экспериментами Экспериментом II и Экспериментом III заключалась в расстоянии между экспериментальными парами «прайм — цель». В Эксперименте II между двумя парами «прайм — цель» находились 2 отвлекающих предложения (не содержащих неоднозначной конструкции с относительным придаточным). Однако при обработке данных по испытуемым оказалось, что большинство испытуемых склонно заканчивать цель не так, как того требует прайм, непосредственно находящийся перед данной целью, а таким типом закрытия, каким была закончена предыдущая цель.

Таким образом, цепочку экспериментальных предложений можно представить следующим образом (рис. 1):

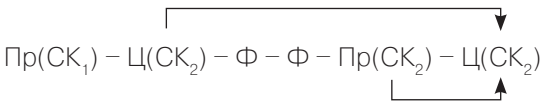


Рис. 1

Как видно на рис. 1, выбор структуры $СК_2$ для цели может быть обусловлен как «манипулируемым праймингом» (нижняя стрелка), так и самопраймингом (верхняя стрелка). Эффект самопрайминга заключается в том, что испытуемый ориентируется не на ту синтаксическую конструкцию, на которую его настраивает прайм, а на свою собственную стратегию. В свою очередь, эта ориентированность на собственную стратегию может быть двух типов: с одной стороны, испытуемый может настроиться на какую-либо синтаксическую конструкцию в ходе прохождения конкретного эксперимента (локальный синтаксический прайминг), а с другой стороны, с возрастом носитель языка настраивается на определенную синтаксическую конструкцию, обычно являющуюся в языке более частотной (глобальный синтаксический прайминг). Явление глобального прайминга описывается Гипотезой языковой настройки на предыдущий опыт ([Mitchell, Cuetos, 1991], [Mitchell et al., 1995], [Драгой 2006]), согласно которой разрешение синтаксической неоднозначности испытуемым тесно связано с тем, как именно он разрешал данную неоднозначность в предыдущем опыте (что, в свою очередь, зависит от предпочтительного типа закрытия в данном языке).

В Эксперименте III расстояние между парами «прайм — цель» было увеличено до 6 предложений, что позволило получить значительно больший прайминг-эффект. Интересно отметить, что в работе [Boyland, Anderson 1998] авторы обнаружили эффект СП после 20ти минутной задержки, а в работе [Bock, Griffin 2000] была выявлена некоторая функция, в соответствии с которой эффекты СП распределяются при наличии между праймом и целью от 1 до 10 стимулов-филлеров. Наши же результаты показывают, что значимое уменьшение эффекта самопрайминга наблюдается уже при увеличении «расстояния» между целями до 6 единиц (отвлекающих предложений).

Таким образом, наши результаты показывают, что эффект самопрайминга можно уменьшать с помощью экспериментального дизайна. Этот вывод имеет большое практическое значение. В случаях, когда направления прайминга и самопрайминга совпадают, мы не можем утверждать, какой именно эффект обусловил выбор испытуемым того или иного типа закрытия, поэтому все полученные эффекты прайминга в половине случаев могут быть обусловлены на самом деле самопраймингом. Обладая возможностью манипулировать уменьшением эффекта самопрайминга с помощью экспериментального дизайна, можно получить более точные данные про эффекты прайминга.

Однако целевые предложения, сконструированные нами для Эксперимента II и III, были в большой степени искусственными в силу их структуры (напомним, в данных экспериментах цели содержало местоимение «который» в форме «которым», равно-

вероятно относящееся либо к творительному падежу вершины ИГ, стоящей в единственном числе, либо к дательному падежу зависимого члена ИГ, стоящего во множественном числе): *В Москве недавно открылся салон автомобилей, которым…* Испытуемые отмечали сложность завершения данных предложений, несмотря на то, что при конструировании данных предложений возможность их завершения в пользу любого члена ИГ была продумана, например, данное предложение можно закончить либо *…которым были довольны покупатели* (РЗ), либо *…которым было уже довольно много лет* (ПЗ). При проведении эксперимента, однако, экспериментатор настаивал на том, что все предложения анкеты должны быть завершены без изменения их структуры, но любым содержанием; поэтому испытуемые были вынуждены придумывать окончания данных предложений. Возможно, в этом и заключается причина того, что завершение одной цели становилось источником эффекта синтаксического прайминга для другой цели: ведь одним из свойств человеческой памяти является то, что решение задачи, вызвавшей определенные трудности, запоминается лучше, чем задача, подобных трудностей не вызвавшая (данное явление носит название эффекта Зейгарник, [Величковский 2006]) и запоминается вплоть до окончания действий, релевантных по отношению к данной задаче. Этим также объясняется то, что в эксперименте [Scheepers 2003] подобного эффекта выявлено не было: цели в этих экспериментах являлись совершенно нормальными предложениями языка и не могли быть восприняты испытуемыми как неестественные. Мы предполагаем, что некая «неестественность» наших экспериментальных предложений способствовала таким образом увеличению эффекта самопрайминга.

Попробуем объяснить полученные нами результаты с точки зрения двух гипотез о механизме прайминга. Одна гипотеза — гипотеза остаточной активации — предполагает, что механизм прайминга основан на активации ментальных репрезентаций синтаксических структур ([Pickering, Branigan 1998]). Прайм, согласно данной гипотезе, активирует в ментальном языковом аппарате испытуемого определенную синтаксическую структуру (включающую в себя также порядок применения правил реализации данной структуры, [Scheepers 2003]), благодаря чему в момент, когда испытуемый должен сделать собственный выбор синтаксической структуры из двух равновероятных, одна уже оказывается выделенной, активированной. Выбор же более выделенной структуры обусловлен общим устройством когнитивного аппарата человека.

Другая гипотеза (гипотеза имплицитного научения) объясняет эффекты прайминга тем, что механизм прайминга основан на имплицитном научении, то есть связан не столько с активацией определенной структуры, сколько с закреплением ее в ментальном аппарате говорящего ([Bock, Griffin 2000]). В пользу данной гипотезы свидетельствует, к примеру, исследование [Hartsuiker, Westenberg 2000]: эффект прайминга может быть кумулятивным: структуры, которые испытуемый склонен повторять в рамках одного эксперимента, оказывается также более частотны и в следующем эксперименте, проводимом с данным испытуемым.

Полученные нами результаты подтверждают обе гипотезы, но только гипотеза остаточной активации актуальна скорее для локального прайминга, а гипотеза имплицитного научения — для глобального. Данное утверждение нуждается в дальнейшей проверке, кроме того, необходимо выяснить, сохранится ли полученный эффект в экспериментах, основанных на диалоге.

Литература

1. *Величковский Б. М.* Современная когнитивная психология.// М.: 1982.
2. *Драгой О. В.* Разрешение синтаксической неоднозначности: правила и вероятности. // Вопросы языкознания, № 6, 2006
3. *Фаликман М. В., Койфман А. Я.* Виды прайминга в исследованиях восприятия и перцептивного внимания // Вестник Московского Университета. М.: МГУ, 2005, серия 14, Психология, № 3, 4.
4. *Фёдорова О. В.* Синтаксическая координация в диалоге: миф или реальность? // Диалог. М.: 2002.
5. *Юдина М. В.* Понимание и порождение высказываний с синтаксической неоднозначностью (на примере относительных придаточных в русском языке) // Диалог. М.: 2006.
6. *Юдина М. В.* Разрешение синтаксической неоднозначности: возможна ли преднастройка? // Диалог. М.:2007
7. *Bock J. K.* Syntactic persistence in language production. // Cognitive Psychology. 1986, № 18, С. 355–387.
8. *Bock K., Griffin Z.* The persistence of structural priming: Transient activation or implicit learning? // Journal of Experimental Psychology: General 1292. 2000
9. *Boyland J. T., & Anderson J. R.* Evidence that syntactic priming is long-lasting. Proceedings of the Twentieth Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.
10. *Branigan H. P., Pickering M. J., Cleland A. A.* Syntactic priming in written production: Evidence for rapid decay. // Psychonomic Bulletin and Review 6. 1999
11. *Branigan H. P., Pickering M. J., Cleland A. A.* Syntactic co-ordination in dialogue. // Cognition. 2000, № 75, B13–B25.
12. *Cuetos, F. & Mitchell, D. C.* Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. // Cognition, 30, 1988
13. *Desmet T., Declercq M.* Cross-linguistic priming of syntactic hierarchical configuration information // Journal of Memory and Language. 2006, № 54
14. *Fodor, J. D.* Learning to parse? // Journal of Psycholinguistic Research, 27, 2, 1998.
15. *Frazier, L. & Clifton, C. J.* Construal: Overview, Motivation, and Some New Evidence. // Journal of Psycholinguistic Research, 26, 3. 1997
16. *Hartsuiker, R. J., Westenberg C.* Persistence of word order in written and spoken sentence production. // Cognition 75. B27–B39. 2000
17. *Mitchell, D. C., & Cuetos, F.* The origins of parsing strategies. // C. Smith (Ed.), Current Issues in Natural Language Processing. Center for Cognitive Science, University of Austin, TX, 1–12. 1991.
18. *Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M.* Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. // Journal of Psycholinguistic Research, 24, 6, 1995.
19. *Pickering M. J., Branigan H. P.* The representation of verbs: evidence from syntactic priming in language production.// Journal of Memory and Language. 1998, №39
20. *Pickering, Martin J., Holly P. Branigan, and Janet F. McLean.* Constituent structure is formulated in one stage. // Journal of Memory and Language 46, 2002
21. *Scheepers C.* Syntactic priming of relative clause attachments: persistence of structural configuration in sentence production. // Dundee: University of Dundee, 2003.
22. *Sekerina, I.* The Late Closure Principle in Processing of Ambiguous Russian Sentences. // The Proceedings of the Second European Conference on Formal Description of Slavic Languages. Universität Potsdam, Germany. 2003
23. *Sekerina I. A., Fedorova O. V.* Questionnaire Studies of Relative Clause Attachment Ambiguity in Russian. // Unpublished manuscript. The City University of New York, 2004.
24. *VanWagenen S., Pertsova K.* Asymmetries in the Russian lexicon: Evidence from inflectional priming. // A presentation at the University of California, Los Angeles, Psychobabble, 2005

Структура атрибутивных значений в тезаурусе RussNet (на материале перцептивных прилагательных)

Structuring of attributive word meanings in RussNet thesaurus (in the group of Russian perceptional adjectives)

Яворская М. В. (yav.mas@gmail.com),
Азарова И. В. (ivazarova@gmail.com)

Санкт-Петербургский государственный университет, Россия

В докладе рассматриваются прилагательные, выражающие перцептивные значения в русском языке и особенности их представления в компьютерном тезаурусе RussNet. Обсуждаются основные значения поля Восприятие: воспринимаемый зрением, слухом, осязанием, обонянием, на вкус; их взаимодействие для частотных и низкочастотных прилагательных.

1. Введение

На кафедре математической лингвистики Санкт-Петербургского государственного университета создается компьютерный тезаурус RussNet для современного русского языка [9, 10], который следует основным принципам, разработанным при построении Принстонского WordNet [8] и других wordnet-словарей [6]. Словари такого типа дают широкие возможности для отображения семантических отношений между значениями, получившими лексикализованное выражение в рамках некоторого языка. Среди отношений есть традиционные для тезаурусной организации лексических значений, например, родовидовые, или гипонимические, а также менее традиционные: меронимические (часть-целое), каузативные, различные семантико-деривационные отношения [2].

Структуризация атрибутивных значений представляется сложной задачей. В первую очередь, это отмечается для прилагательных. Например, в традиции wordnet-словарей неоднократно отмечалось [8], что значения прилагательных образуют своеобразные кластеры: пары антонимичных значений, к которым присоединяются стилистически окрашенные или специализированные каким-либо образом значения. Таким образом, получаются двухуровневые структуры. Построение же многоярусной родовидовой структуры считается невозможным, хотя в некоторых wordnet-словарях такие структуры были созданы [3].

В настоящем докладе мы попытаемся обозначить основные проблемы структуризации атрибутивных значений перцептивных прилагательных. Эта группа выбрана ввиду ее многочисленности и разнообразности: по нашим подсчетам они составляют около 3000–3500 ipm. Ядерные прилагательные этой группы входят в число наиболее частотных (*большой* —1631 ipm*; белый* —493 ipm*; тяжелый* — 247 ipm*; тихий* — 138 ipm и т. д.). Поэтому группа атрибутивных значений перцептивных прилагательных представляется весьма перспективной: если реально построить родовидовую структуру и установить другие характерные семантические отношения, то как раз в ней станут очевидными все проблемы, которые необходимо разрешить.

Еще одним вопросом, является то, какова природа родовидовых отношений: совпадает ли она с гипонимией существительных или тропонимией [8] глаголов, или же представляет совершенно иной тип.

2. Перцептивные прилагательные и различные подходы к представлению значений

В докладе мы будем оперировать термином «перцептивные прилагательные», подразумевая качественные прилагательные, «в семантике которых репрезентированы перцептивные образы признаков

предметов, сформированные в сознании человека на основе работы его органов чувств» [16; с. 72].

Есть ли у класса перцептивных прилагательных свои особенности? Т. И. Клименко отмечает, что наиболее упорядоченными и многочисленными по составу членов являются «родо-видовые и видовидовые группы прилагательных со значением «цвет», «форма», «размер», «расположение в пространстве», «вкус»… и некоторые другие» [15; с. 8.], то есть те качества, значение которых мы воспринимаем непосредственно органами чувств. Для перцептивных признаков характерна как полимодальность (один и тот же атрибут объективно может относится к нескольким модусам перцепции, ср. *тошнотворный* (вкус, запах, вид), *пронзительный* (звук, тактильные ощущения, запах), так и интермодальность, основанная на явлении синестезии, то есть «употреблении слова, значение которого связано с одним органом чувств, в значении, относящемся к другому органу чувств. При интермодальности всегда можно выделить первичный модус, по аналогии с которым рассматриваются другие» [17; с. 20]. При разработке представления значений перцептивных прилагательных, мы в дальнейшем постараемся, по возможности, проследить переход значений между доменами.

Современная прикладная лингвистика предлагает целый ряд возможных представлений значений имен прилагательных. Одним из самых распространенных подходов является рассмотрение конструкций типа «прилагательное + существительное», когда на основе сочетаемости имен делается вывод об их семантической структуре: «сочетаемость лексических единиц имеет прямую зависимость от совместимости понятий и находит свое воплощение в комбинаторике концептов, выраженных этими лексическими единицами. В результате оказывается очевидным тот факт, что совместимость или несовместимость слов выходит далеко за пределы собственно лингвистической сферы в логическую, лингвокультурную, когнитивную и даже онтологическую плоскость» [22; с. 45].

Другой распространенный подход — это компонентный анализ имен прилагательных с опорой на словарные дефиниции (в том числе, интерпретация данных ассоциативных экспериментов, отраженных в ассоциативных словарях). Данный метод также позволяет зафиксировать способы концептуализации мира и их отражение во фрагментах языковой картины мира. Например, одно из положений в диссертационном исследовании Н.С. Павловой говорит о «наличии типовых русских запахов, определяющих специфику русской обонятельной картины мира» [21; с. 6]

И, наконец, существует необходимость разработки представлений значений имен прилагательных под конкретные компьютерные приложения, в нашем случае это электронный тезаурус типа wordnet, отражающий отношения парадигматического типа между словами. Очевидно, что традиционные для имен прилагательных отношения синонимии и антонимии не будут представлять сложности. Поэтому нас в первую очередь будет интересовать возможность существования и представления дополнительных отношений между прилагательными, а также прилагательными и существительными.

3. Иерархия атрибутивных значений

Т. И. Клименко, говоря о глобальном характере родовидовых отношений, выявляющих процессы категоризации явлений внешнего мира в семантической структуре слов, подчеркивает, что «отличительной чертой родовидовых связей имен прилагательных является то, что их парадигмы в целом беднее по количеству членов и семантических оттенков, чем парадигмы других частей речи, например, существительных» [14; с. 21].

Что же касается гипонимии с точки зрения прилагательных в целом, то, как известно, «в парадигматике гипонимы характеризуются, с одной стороны, наличием существенного общего компонента значения, содержательно равного их гипероними и составляющего гиперсему их интенсионалов. Этот общий компонент значения может быть выражен отдельным словом, а может быть представлен описательно (что более характерно для имени прилагательного)» [14; с. 21]. Более того, гиперонимом прилагательного может выступать имя существительное: в поле цвета «самым общим родовым наименованием является гипероним-существительное “цвет”» [14. С. 28.].

4. ЛСГ перцептивных прилагательных в русском языке

Для анализа структуры значений группы перцептивных прилагательных были отобраны 82 прилагательных. Значения этих слов, представленные в Словаре русского языка в четырех томах под редакцией А.П. Евгеньевой (М., 1981) – далее МАС, используются для первоначальной разметки значений контекстов в корпусе современных текстов кафедры математической лингвистики объемом в 21 млн. словоупотреблений [8]. Это стандартная методика определения структуры значений для некоторой лексической группы в RussNet [8], которая позволяет экстраполировать результаты разметки случайной выборочной совокупности контекстов (100–300) на корпус в целом, выявляя таким образом реальную структуру распределения значений в лексико-семантической группе. Выявленное по корпусу распределение значений меняет структуру и даже состав статьи. Так, например, *тяжелый2* по МАС (‘*прост*. Беременная’) не встретилось в выборке ни разу, поэтому это значение не включа-

ется в RussNet. Зато анализ контекстов выявил перцептивное значение ‘характеристика звука с преобладающими низкими частотами’ 1,08 ipm, которого нет в MAC, но оно будет введено в RussNet. По корпусу было выделено 178 атрибутивных значений и проанализировано 31080 контекстов.

Среди выявленных атрибутивных значений можно выделить группу высоко- и среднечастотных с частотой более 10 ipm: *большой1, белый1, огромный1, маленький1, небольшой1, крупный1, белый2, мелкий1, широкий1, длинный1, высокий2, теплый1, малый2, круглый1, длинный2, тяжелый2, цветной1, мягкий1, длинный5, соленый1, короткий2, высокий6, глубокий2, теплый2, гигантский1, значительный3, сладкий1, легкий2, маленький3, громкий1, острый2*; периферию составляют низкочастотные атрибутивные значения: *острый2, легкий3, короткий3, гладкий1, крошечный1, громадный1, острый3, низкий2, обширный1, теплый3, теплый4, кислый2, горький1, невысокий1, крохотный1, мягкий2, мутный1, глубокий5, миниатюрный1, теплый6, вонючий1, квадратный1, душистый1, легкий6, бескрайний1, рослый1, ароматный1, необъятный1, низкий5, пресный1, жирный2, короткий6, пряный1, круглый2, пронзительный1, тяжелый6, мягкий6, глубокий7, крупный5, мягкий8, невысокий3, увесистый1, мелкий6, хриплый1, жирный3, здоровый4, высокий9, бездонный1, мягкий9, мягкий10, здоровый5, грузный1, протяжный1, тяжелый10, студеный1, невесомый1, объемный1, едкий2, исполинский1, квадратный3, пахучий1, тяжелый11, объемный2, колоссальный4, внушительный2, острый8, глубокий8, легкий11, легкий12, пряный2, изрядный2, тяжелый12, острый10, глубокий9, вонючий3, сочный1, теплый8, высоченный1, долговязый1, мутный2, соленый4, сладкий7, сладкий6, большущий1, зловонный1, терпкий1, ломкий1, кислый5, мутный4, жгучий2, жирный4, гладкий3, крошечный2, аппетитный1, легкий15, крохотный2, сочный3, теплый9, теплый10, аппетитный2, жгучий4, гладкий5, пронзительный4, беспредельный4, жирный6, пахучий2, душистый3, жгучий5, безграничный5, терпкий2, объемный3, сочный4, прелый1, высоченный2, горький5, горьковатый1, жирный7, долговязый2, круглый3, ломкий2, тошнотворный4, тошнотворный3, зловонный2, аппетитный4, едкий4, лакомый1, миниатюрный2, пронзительный7, трескучий2, горьковатый2, пресный3, мутный7, жирный10, жирный8, трескучий3, студеный2, мягкий16, ароматный2, вяжущий2, горький8, пронзительный8, едкий6, невесомый3.*

4.1. Ядро и периферия группы значений прилагательных, определяющих визуальные параметры

В данной группе представлены характеристики таких параметров, как размер, цвет, форма, свойство поверхности. Самой представительной является группа атрибутивных значений, задающих параметры размера. В этой группе центральное место занимает WM[1] *большой1*[1] ‘значительный по одному или нескольким измерениям’. На рис. 1 приведена схема семантических связей в данной группе. Каждый из элементов нашей схемы содержит синсет, члены которого располагаются в порядке убывания частотности их значений (в ipm). Стрелками обозначены отношения между синсетами.

Корневой синсет с доминантой *большой1 (большой стол, крупная кошка, здоровый камень)* присоединяет гипонимы разного типа.

Синсет с доминантой *огромный1* (*огромный зонт, исполинский зверь, громадная звезда*) указывает на интенсивность проявления признака, при этом типичная сочетаемость с существительными сохраняется. Данное отношение сходно с традиционной для существительных гипонимией: интенсионал гиперонима включен в интенсионал гипонима, к которому добавлен компонент ‘высокой степени’ (Magn).

Другая группа гипонимов (*широкий1 — длинный1 — высокий2 — глубокий2*) уточняет значительность размеров в одном из измерений, поэтому сочетаемость существенно у́же, чем у гиперонима (*широкая грудь, широкое поле, *широкая кошка; длинные ноги, длинное шоссе, *длинная звезда; высокая скала, высокий стакан, *высокий зонт; глубокий колодец, глубокая долина, *глубокий стол*). При этом общая схема сочетаемости кажется сходной (конкретные видимые объекты реальности), однако включения интенсионала нет. В этом случае семантическое соотношение между вышестоящим и нижестоящим синсетами скорее напоминает тропонимию глаголов (ср. *идти* и *семенить, маршировать, хромать* и проч.). Это характеристика значительного размера некоторым специфическим способом. Сложение значений гипонимов дает значение гиперонима *большой1*.

Третья группа гипонимов на схеме указывает еще на один вариант гипонимии: сужения сферы сочетаемости. Применительно к характеристике людей и животных появляются специализированные синсеты, например, *долговязый1, грузный1*. Этот тип гипонимии имеет общие черты и с гипонимией существительных, и тропонимией глаголов.

Очевидной периферией в данной подгруппе являются атрибуты ‘производящий впечатление такого-то веса’:

- *Тяжелый6* ‘Производящий впечатление большого веса, тяжести’, 2,4 ipm
- *Легкий15* ‘Производящий впечатление невесомого, изящный’, 0,53 ipm
- *Невесомый3* ‘Производящий впечатление отсутствия веса, легкости’, 0,05 ipm

[1] WM — word meaning, т.е. значение, выраженное определенным словом (иногда устойчивым словосочетанием). WM является элементом синсета (синонимического ряда), основной единицы wordnet-словаря.

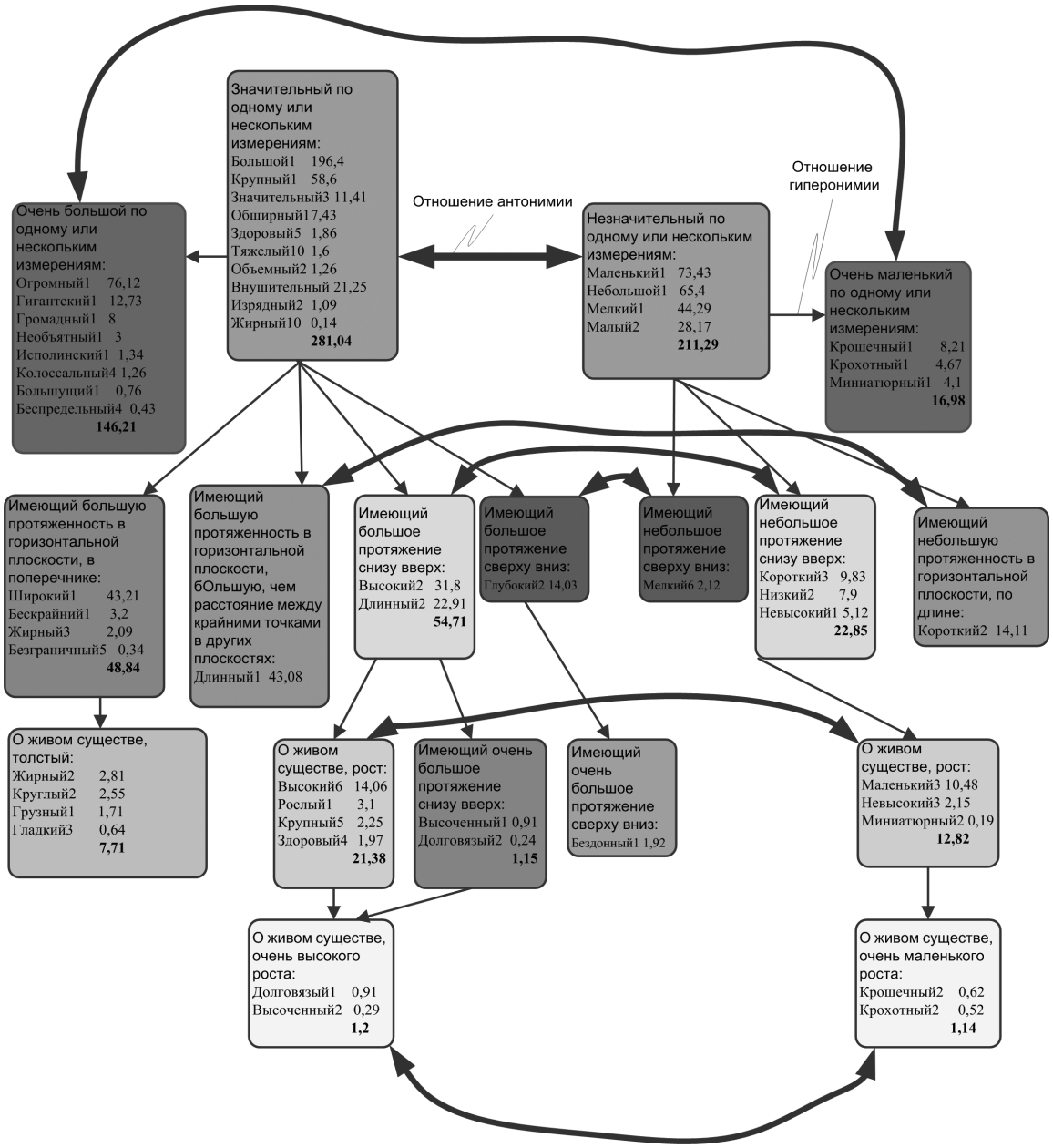


Рис. 1. Схема семантических отношений в группе атрибутивных характеристик размера.

Эта характеристика может не отражать реального веса объекта: *тяжелая, по-питерски мрачная колоннада; легкая архитектура; утренняя и невесомая дорога.*

4.2. Ядро и периферия группы прилагательных, определяющих слуховые параметры

В этой группе отражены такие характеристики звука, как громкость, высота, длительность, чистота и место образования. Довольно неожиданно наиболее частотным является значения из синсета с доминантой *длинный5* ‘продолжительный [по времени]’ (*длинная тирада, длинная песня*). Схема значений этой группы приведена на рис. 2.

Рис. 2. Схема семантических отношений в группе атрибутивных характеристик звука.

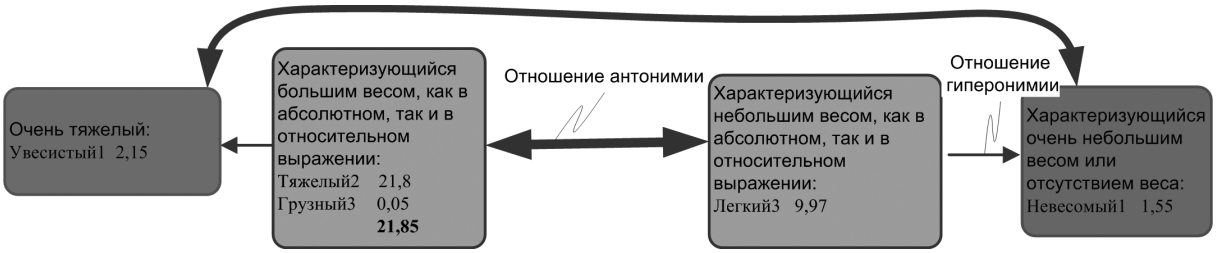


Рис. 3. Схема семантических отношений в группе атрибутивных характеристик веса.

На периферии группы находятся значения, в которых указывается эмоциональная реакция на звук: *пронзительный1* ‘сильно, резко действующего на органы чувств’ и *мягкий10* ‘приятный для слуха’.

4.3. Ядро и периферия группы прилагательных, определяющих тактильные параметры

Данная группа атрибутивных значений характеризует вес, температуру, влажность, свойства всего тела (*ломкий, упругий*) и свойства поверхности объектов действительности. Наиболее частотной является группа температурных характеристик. На рис. 3 приведена схема для отношений в группе характеристик веса с ядром *тяжелый2* ‘характеризующийся большим весом, как в абсолютном, так и в относительном выражении’.

На периферии характеристики с гедонистическим значением ‘приятный на ощупь’.

гладкий1	**приятный на ощупь**, без выступов, впадин и шероховатостей, ровный.	**9,43**
мягкий2	**приятный на ощупь**, тонкий, шелковистый (о шерсти, волосах, тканях).	**4,62**
мягкий8	нежный, **приятный на ощупь**, пухлый (о теле, частях тела).	**2,24**
гладкий5	**приятный на ощупь**, однородный, без комочков.	**0,48**
		16,77

4.4. Ядро и периферия группы прилагательных, определяющих обонятельные параметры

Согласно исследованию Н. С. Павловой, характеристики запаха в русском языке малочисленны, что свидетельствует о «недостаточной разработанности когнитивного представления ольфакторных признаков» [21; с. 8]. Даже частотные характеристики имеют свойства периферии: дают (субъективную) оценку запаха.

Вонючий1, едкий2, зловонный1, кислый5, горький5, тошнотворный3, пронзительный8 ‘характеризующийся очень неприятным запахом’ (7,42 ipm). *Вонючий лак, едкая вонь, зловонная жижа, несло кислым табачищем, легкий дымок, горький как отзвук выстрела, тошнотворная вонь, пронзительный запах.*

Душистый1, ароматный1, сладкий6, теплый10, аппетитный4 ‘характеризующийся приятным запахом’ (8,06 ipm). *Душистый мед, ароматная мазь, сладкий аромат, теплый запах сена, аппетитный запах.* В этом синсете отсутствует интенсификация признака, поэтому даже отношение антонимии выглядит сомнительным.

Пахучий1, тяжелый11, пряный2, острый10, терпкий2, жирный12 ‘обладающий сильным запахом’ (5,22 ipm). *Пахучее мыло, тяжелый цветочный запах, пряные благовония, острое зловоние рыбы, терпкая нота, жирный дух.*

Качество запаха регулярно передается в тексте через сочетание с родительным падежом существительного *запах чего-л,* которое указывает источник запаха (*запах прения*) или запах-ориентир (*запах горького миндаля*).

Подгруппа значений указывает на несобственный запах: *вонючий3 ‘разг.* пропахший очень неприятным запахом’ (*вонючий барак*), *зловонный2* ‘пропитавшийся очень неприятным запахом, зловонием’ (*зловонное белье*), *душистый3* ‘пропитавшийся или имеющий на себе вещество, обладающее сильным, приятным запахом’ (*душистый пеньюар*).

4.5. Ядро и периферия группы прилагательных, определяющих вкусовые параметры

В этой группе отсутствует явная доминанта. Выделяются названия прототипов соответствующих вкусов: *острый, горький, сладкий, соленый, кислый и т.д.,* между которыми практически невозможно установить семантические связи. В редких случаях реализована антонимия по наличию/отсутствию

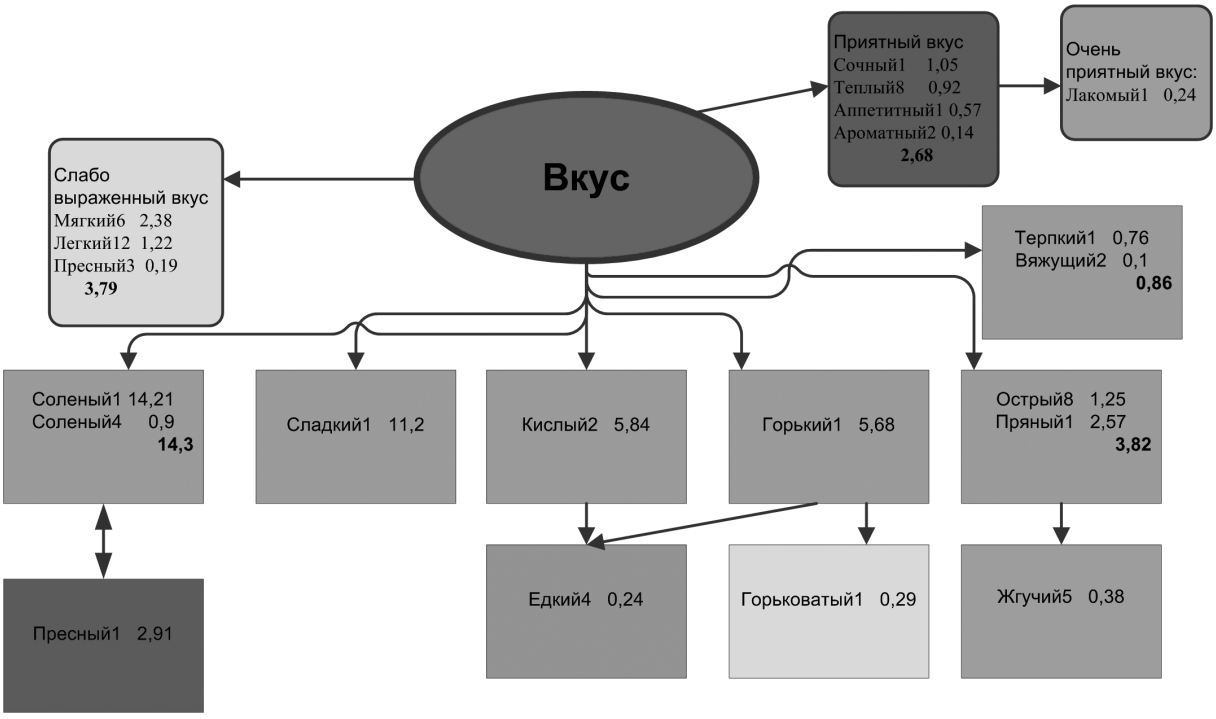


Рис. 4. Схема семантических отношений в группе атрибутивных характеристик вкуса.

заданного вкуса (*соленый1 — пресный1*) или гипонимия по степени выраженности (*горький1 — горьковатый1*). Часто характеристика вкуса задается конструкцией с существительным в родительном падеже (*вкус клубники*).

Однако, Т. И. Клименко утверждает, что в данной группе перцептивных прилагательных, наряду с размером, цветом и формой, наиболее ярко проявляются родовидовые отношения. Основные эталоны вкуса похожи на тропонимы глаголов, могут соединяться: *сладко-соленый, терпко-горький, остро-кисло-сладкий, едкий* ‘кисло-горький вкус’ и т. д.

5. Выводы и перспективы исследования

Исследованные группы значений перцептивных прилагательных довольно неоднородны как в плане разной «проработанности» целой области перцептивного восприятия, так и во внутренней структуре отдельных областей. В тех случаях когда возможно иерархическое упорядочение значений, реализуются 3 разные схемы родовидовых связей: по типу существительных, по типу глаголов и смешанная схема. Оценочно-экспрессивный характер некоторых перцептивных атрибутов затрудняет выявление иерархии, чаще всего это периферийные зоны перцептивных областей.

Расширение круга атрибутивных значений перцептивных прилагательных позволит уточнить схему отношений характеристик в перечисленных группах.

Литература

1. *Alonge A.* & others. Encoding information on adjectives in a lexical-semantic net for computational applications. P. 42–49.
2. *Azarova I., Mitrofanova O., Sinopalnikova A., Yavorskaya M., Oparin I.* RussNet: Building a Lexical Database for the Russian Language // Workshop on WordNet Structures and Standardisation, and how these affect Wordnet Application and Evaluation. 28th May 2002. Las Palmas de Gran Canaria, 2002. P. 60–64.
3. *http://www.sfs.uni-tuebingen.de/GermaNet/*
4. *Kanzaki Kyoko* & others. Construction of an Objective Hierarchy of Abstract Concepts via Directional Similarity. // International Conference On Computational Linguistics. Proceedings of the 20th international conference on Computational Linguistics. Geneva, Switzerland. 2004. Article No. 1147.
5. *Mendes Sara*. Adjectives in WordNet.PT // GWC 2006, Proceedings, P. 225–230.
6. *The Global* WordNet Association // URL: http:// www.globalwordnet.org/
7. *Wierzbicka A.* Ethno-syntax and the Philosophy of Grammar // Studies in Language. 1979. № 3.
8. *WordNet:* An Electronic Lexical Database / Fellbaum Ch. (ed.). MIT Press. 1998.
9. *Азарова И. В., Синопальникова А. А.* Использование статистико-комбинаторных свойств корпуса современных текстов для формирования структуры компьютерного тезауруса RussNet // Труды международной конференции «Корпусная лингвистика 2004». СПб.: 2004. С. 5—15.
10. *Азарова И. В., Синопальникова А. А., Яворская М. В.* Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004 М.: 2004. С. 542—547.
11. *Арутюнова Н. Д.* Аксиология в механизмах жизни и языка // Проблемы структурной лингвистики, 1982. М.: 1984. С. 5—23.
12. *Вольф Е. М.* Функциональная семантика оценки. М.: 2002.
13. *Клименко А. П.* Лексическая системность и ее психолингвистическое изучение. Минск: 1974.
14. *Клименко Т. И.* Виды семантических отношений в системе имен прилагательных (на материале «Русского семантического словаря» и словарей эпитетов). Диссертация на соискание ученой степени к.ф.н. Брянск: 1994.
15. *Клименко Т. И.* Виды семантических отношений в системе имен прилагательных (на материале «Русского семантического словаря» и словарей эпитетов). Автореферат диссертации на соискание ученой степени к.ф.н. Санкт-Петербург: 1995.
16. *Лаенко Л. В.* Перцептивный признак как объект номинации // Вестник ВГУ. Серия: Филология. Журналистика. 2004. №2. С. 71–77.
17. *Лаенко Л. В.* Перцептивный признак как объект номинации. Воронеж: 2005.
18. *Лингвистический* энциклопедический словарь. Ред. Ярцев В. Н. М.: 2002.
19. *Матвеева Т. М.* Перцептивная категория вкуса и лингвистические средства ее реализации. Автореферат диссертации на соискание ученой степени кфн. М.: 2005
20. *Никитин М. В.* Основы лингвистической теории значения. М.: 1990.
21. *Павлова Н. С.* Лексика с семой «запах» в языке, речи, тексте. Автореферат диссертации на соискание ученой степени к.ф.н. Екатеринбург: 2006.
22. *Юдина Н. В.* Сочетания «прилагательное + существительное» в лингвокогнитивном аспекте. Москва-Владимир: 2006.

Формирование наборов опорных слов в разных типах восприятия речи[1]

Best recognizable words under different experimental settings

Ягунова Е. В. (iagounova_elena@mail.ru)

Санкт-Петербургский государственный университет

В докладе рассматриваются основные характеристики и предполагаемые пути формирования наборов опорных слов (НОС). Под опорными понимаются наиболее распознаваемые слова при восприятии в разных экспериментальных режимах. Данные подтверждают гипотезу о существенном значении функционального стиля и степени динамичности текста.

1. Вводные замечания

Настоящий доклад является продолжением исследования опорных слов (Ягунова 2008б). Под опорными понимаются наиболее распознаваемые слова при восприятии осмысленного текста в условиях искажения.

В основе работы лежат следующие положения:

- речевая ткань звучащего текста неоднородна по своей природе, т.е. лишь некоторые ее фрагменты могут распознаваться за счет анализа фонетических характеристик (ср. Бондарко и др. 1974);
- функциональный стиль текста и экспериментальные условия восприятия задают уровень и тип неоднородности текста (Ягунова 2008а);
- распределение опорных слов в ткани текста (неоднородность) является видом структурированности текста (Ягунова 2008а).

При восприятии осмысленного текста разборчивость слов определяется как фонетическими характеристиками текста, так и процедурами контекстной предсказуемости, позволяющей восстанавливать фрагменты текста на основании смысловой (внефонетической) информации. Можно предположить, что эти слова несут особую смысловую нагрузку, которой и обусловлена высокая распознаваемость именно этих слов. Набор опорных слов (НОС) представляет собой свертку текста: упорядоченный набор слов, сопоставленный исходному тексту и в определенных условиях выступающий в качестве его «представителя» (Ягунова 2008б).

С другой стороны — функциональная нагрузка слов, распознающихся на основании собственно фонетической информации, и слов, восстанавливаемых на основании контекстной предсказуемости, различны. Для исследования вопроса о возможных путях формирования НОС мы прибегли к сопоставлению характеристик НОС, полученных при восприятии осмысленного текста в шуме (при соотношении сигнал/шум 0 дБ) (Ягунова 2008б) и при восприятии псевдотекстов, лишенных лексико-семантической информации.

Подобное сопоставление позволит проанализировать то, как влияют фонетические и/или внефонетические характеристики стимула на формирование НОС (отдельно рассмотреть роль контекстной предсказуемости).

В качестве исследуемого материала было выбрано два переводных текста на русском языке:

1. Отрывок из официальной публикации «Закон об иностранных инвестициях во Вьетнаме и нормативные акты, изданные на его основе» (в дальнейшем «деловой текст»);
2. Отрывок сюжетной художественной прозы Нам Као «Ти Фео» с элементами диалога (в дальнейшем «художественный текст»)[2].

Художественный текст является динамическим, т. е. характеризуется сменой ситуаций. Композицион-

[1] Работа выполнена при частичном финансировании РГНФ (проект №07-04-00161а)

[2] Существенным ограничением использования эксперимента по восприятию псевдотекста явилось ограничение объема предъявляемого материала — начальный фрагмент текстов, обладающий некоторой законченностью (смысловой и интонационной) и составляющий около трети от исходного текста, анализирующегося в (Ягунова 2008б).

ная структура художественного текста включает преамбулу (ориентацию) и завязку (по Чейфу). Деловой текст (текст-предписание) является статическим.

Исходные псевдотексты были созданы на основании двух исходных текстов, в графической записи которых ***все*** согласные были заменены на парадигматические аналоги (сонорные на сонорные, глухие взрывные на глухие взрывные и т. д.). Диктор читал псевдотексты, ***имитируя*** просодические структуры исходных текстов. Т.о., можно говорить о максимально возможном фонетическом подобии исходного текста и псевдотекста (далее псевдотекст-I или псевдо-I).

Пример преобразования фрагментов исходного текста в псевдотекст-I:

*Плановые показатели вьетнамскими государственными органами не устанавливаются. ⇒ Тр**б**мозые топав**б**кеми льем**б**ршкири дохуб**б**цвемыни **у**лбарани ре успар**б**мнизаюча. — Опять притащился! Пора бы знать своё место, здесь тебе не банк! ⇒ Отя**й**к снипаф**ъ**нчя! Ков**б** лы жр**б**пь хмоё р**й**фто, гв**й**х пег**й** ле д**б**рт!*

Под опорными в настоящем докладе понимаются словоупотребления (фонетические слова), распознающиеся не менее чем 30% испытуемых при восприятии текста в следующих условиях:

- при восприятии осмысленных исходных текстов в шуме при соотношении сигнал/шум 0 дБ;
- при восприятии псевдотекстов двух типов:
  - псевдо-I — исходные псевдотексты,
  - псевдо-М — монотонизированные (с помощью программы WinPitch) по основному тону псевдо-I[3].

В эксперименте по восприятию текста в шуме приняло участие 40 испытуемых. В эксперименте по восприятию псевдо-I приняло участие более 60 испытуемых и псевдо-М — 40 испытуемых. В инструкции испытуемым предлагалось прослушать текст удобными порциями, останавливая прослушивание с помощью клавиши «пауза» и записывая в орфографии каждый услышанный фрагмент. Текст можно было слушать один раз, не возвращаясь назад.

Каждый испытуемый мог участвовать лишь в одном из экспериментов[4].

В дальнейшем НОС, полученные при восприятии в этих трех условиях, будем называть *НОС-шум*, *НОС-псевдо-I* и *НОС-псевдо-М*. В качестве опорных слов в НОС-псевдо-I и НОС-псевдо-М определяются псевдослова. Однако если принцип подобия преобразовании текста в псевдотекст оказывается работающим, то логично предположить возможность сопоставления опорных псевдослов и их исходных проекций. Напр., опорность фонетических псевдослов *гв**й**х, пег**й**,* и *ле д**б**рт* (в *НОС-псевдо-I* и *НОС-псевдо-М*) мы можем интерпретировать как то, что анализ лишь фонетических характеристик позволяет распознать фонетические слова (ФС) *здесь*, *тебе*, *не банк* в исходном тексте, более того, это возможно даже после монотонизации по основному тону. При анализе внефонетических характеристик НОС-псевдо (в НОС-псевдо-I и НОС-псевдо-М), таким образом, рассматриваются наборы соответствующих слов исходного текста.

На материале НОС-шум нами были получены следующие результаты (Ягунова 2008б):

- В отличие от набора ключевых слов (КС) — традиционного вида свертки текста характеризующего текст как целостный объект[5] — НОС соответствует «следам» восприятия и понимания текста ***в текущем режиме времени***, когда смысловая структура извлекается не только из текста целиком, но и из его структурных составляющих (Ягунова 2008б);
- Для смыслового структурирования текста и формирования НОС существенное значение имеют признаки «функциональный стиль текста» и «степень динамичности текста».
- Статичность делового и динамичность художественного текстов могут быть показаны
  - с помощью минимального статистического анализа НОС (доли в них существительных и КС);
  - через восстановление внутритекстовых ассоциативных связей между словами (в том числе связей в разных составляющих: синтагмах, фразах, ситуациях и т.д.), т.е. через эксперимент по восстановлению текстов на основе НОС[6].

---

[3] Монотонизация как искажение фонетических характеристик псевдо-I использовалась для того, чтобы проследить вклад этого признака в формирование НОС, т.к. мелодические характеристики в существенной степени влияют на опорность слов и на выделение наиболее важных компонентов в смысловых структурах текста (Ягунова 2006; Ягунова 2007; Ягунова 2008а).

[4] В настоящем исследовании ставилась задача изучения восприятия в условиях ограничений на «базу знаний» адресата (предметная область делового текста и реалии вьетнамской жизни в художественном тексте были незнакомы испытуемым). В этом случае восприятие текста в максимальной степени опирается на структуры самого текста.

[5] Ср. (Мурзин, Штерн 1991). КС нами определялись — по традиционно используемой методике (Мурзин, Штерн 1991) — в ходе вспомогательного эксперимента, в котором испытуемым предлагалось следовать следующей инструкции: «Прослушайте текст. Подумайте над его содержанием. Выпишите из текста 10–15 слов, наиболее важных с точки зрения его содержания». Испытуемые заполняли анкеты ***после*** окончания прослушивания текста. Следовательно, НКС, по-видимому, отражает смысловую структуру текста как целостного объекта.

[6] Впрочем, состав художественного НОС-шум не позволяет восстановить динамичность художественного текста (включающего только два смысловых компонента: преамбулу и завязку).

Задачей настоящей работы является исследование формирования НОС на материале сопоставления *НОС-шум*, *НОС-псевдо-I* и *НОС-псевдо-М*. В основе сопоставления лежат следующие гипотезы:

- Распределение в тексте опорных слов, распознающихся за счет анализа лишь фонетических характеристик, несет существенную информацию о смысловой структуре текста.
- Для фонетического и смыслового структурирования текста и формирования НОС-псевдо-I и НОС-псевдо-М существенное значение имеют признаки «функциональный стиль текста» и «степень динамичности текста».
- Роль мелодических признаков можно проследить через анализ *НОС-псевдо-М*, тех компонентов, которые остались разборчивыми в условиях монотонизации.
- Развертывание НОС-псевдо-I и НОС-псевдо-М в текст — в отличие от НОС-шум — позволяет определить позиции слов, обладающих максимальной предсказуемостью

2. Сопоставление наборов опорных слов

2.1. Наборы разборчивых слов (разных уровней разборчивости)

Что из себя представляет наборы разборчивых слов в разных экспериментальных условиях? Какую часть текста эти слова покрывают? Как изменяется доля распознаваемых слов с изменением уровня разборчивости? Почему опорные слова были определены как слова 30%-го уровня разборчивости? Проиллюстрируем (рис. 1) основные закономерности на примере сопоставления двух кривых, характеризующих соотношения числа слов 8ми уровней разборчивости для осмысленного текста в шуме, когда разделить влияние фонетических и внефонетических признаков невозможно, и для псевдотекста (на примере псевдо-I), когда мы можем говорить об изолированном влиянии собственно фонетических признаков.

Кривые, описывающие данные по естественным текстам, оказываются между кривыми по данным псевдотекстов; в среднем наборы разборчивых слов для делового и художественного псевдотекстов значимо различаются, аналогичные наборы разборчивых слов по данным делового и художественного естественных текстов — различаются незначимо Полученную картину можно интерпретировать как признание того, что взаимодействие фонетических и внефонетических признаков обеспечивает, условно говоря, большее единообразие, чем действие исключительно фонетических признаков.

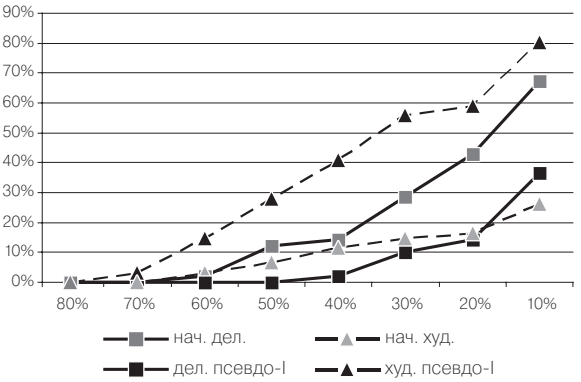


Рис. 1. Отношение доли фонетических слов 8ми уровней разборчивости (от общего числа ФС в текстах) для восприятия текста в шуме и псевдо-I

С другой стороны, полученные данные дают наглядное представление о существенном влиянии функционального стиля на распределение в тексте наборов разных уровней разборчивости. Для художественного псевдотекста разборчивые слова (распознающиеся на основании собственно фонетической информации) покрывают бóльшую часть «карты» текста, чем для делового текста. Длина слов художественного текста оказывается оптимальной для распознавания, а слова делового текста — слишком длинными. Для осмысленного текста, при распознавании которого невозможно разделить роль фонетических и внефонетических признаков, данные по опорным словам художественного и делового текстов оказываются зеркальным образом перевернутыми: доля опорных слов для делового текста существенно выше, чем для художественного. По-видимому, подобное «перевертывание» обусловлено механизмами контекстной предсказуемости, в результате которых могут реконструироваться слова осмысленного текста; для делового текста результаты работы контекстной предсказуемости выше, чем для художественного. Одна из причин высокой контекстной предсказуемости лежит в более высокой синтаксической структурированности делового текста, чем художественного текста (подробнее см. Ягунова 2007б).

2.2. Внефонетические характеристики наборов опорных слов

В табл. 1а и 1б приведены наборы опорных слов по результатам 4 экспериментов. Эти наборы характеризуют особенности, определяющиеся структурой текстов разных функциональных стилей, фонетическими характеристиками самих опорных слов и условиями искажения.

Набор КС характеризует текст как целостный объект, в КС в незначительной степени заложена динамичность смены ситуаций (Ягунова 2008б).

Среди КС наиболее значимое место занимают существительные (см., напр., Мурзин, Штерн 1991), чем больше доля существительных, тем выше степень статичности текста. Эти положения позволяют нам ограничить анализ внефонетических характеристик двумя — часть речи (соотносимость с существительными) и соотносимость с КС (смысловыми вехами текста, определяемыми в ходе дополнительного эксперимента с носителями языка).

Таблица 1а. Наборы опорных слов для трех экспериментальных режимов восприятия деловых псевдотекстов

осмысленный текст в шуме	псевдо-I	псевдо-M
сто́роны	сто́роны	
участвующие		
с иностранным		
капиталом	капиталом	
	в технико-	в технико-
		к заявке
имеют		
право		
самостоятельно		
устанавливать		
программы		
сто́роны		
договора		
сотрудничестве		
с иностранным		
	капиталом	
		включать
	показатели	

Условные обозначения: словоупотребления упорядочены в соответствием с порядком введения в исходный текст, подчеркиванием выделены слова, соотносимые с ключевыми для данного фрагмента

Таблица 1б. Наборы опорных слов для трех экспериментальных режимов восприятия художественных псевдотекстов

осмысленный текст в шуме	псевдо-I	псевдо-M
когда	когда	когда
подобные		
мысли	мысли	
трудно		
настроение		
	и выдержку	
	а тут	а тут
	еще	еще
	Ти Фео	Ти
	клянчить	
	на выпивку	
	ему	ему
	пять хао	пять хао
	опять	опять
пора бы	пора бы	пора бы
знать	знать	знать
свое		
место	место	место
	здесь	здесь
	тебе	тебе
	не банк	не банк
	с этими	с этими
	словами	словами
		он
	с размаху	
	монету	
	на землю	
	забирай	забирай
	и катись	и катись
	чтобы	чтобы
	духу	
	живо	
	сколько	сколько
	раз	раз
	говорил	
	жить	жить
	честно	честно

Анализа НОС-шум показывает, что для статичного делового текста (по сравнению с более динамичным художественным) подавляющее число опорных слов являются ключевыми (54% vs. 33%). Бóльшая доля существительных рассматривается как признак бóльшей статичности текста (набора). НОС-шум статичного делового текста (по сравнению с художественным) характеризуется бóльшей долей существительных (57% vs. 33%, см. Ягунова 2008).

НОС-псевдо-I (распознаваемость на основании сохранной фонетической информации) демонстрирует статичность делового: бóльшую долю КС (80%

vs. 34%) и существительных (80% vs. 41%) (по сравнению с художественным), т. е. результаты аналогичны вышеописанным (см. табл. 2 и 3). Эти данные можно интерпретировать как то, что наборы опорных слова, распознающихся на основании лишь сохранных фонетических характеристик, несут в себе смысловую информацию о статичности/динамизме текста.

Для художественного текста во всех экспериментальных режимах доля существительных сравнительно мала (см. табл. 2). Монотонизация (отсутствие изменения основного тона, псевдо-М) как для делового, так и для художественного текстов дает минимальную (не более трети) долю существительных. Вероятно, движение частоты основного тона дает основания для определения статичности/динамичности.

Таблица 2. Доля слов в рассматриваемых НОС, соотносимых с КС

эксп. режим / текст	**НОС-шум**	**НОС-псевдо-I**	**НОС-псевдо-М**
деловой	0,54	0,80	0
художественный	0,33	0,34	0.30

Таблица 3. Доля слов в рассматриваемых НОС, соотносимых с существительными

эксп. режим / текст	**НОС-шум**	**НОС-псевдо-I**	**НОС-псевдо-М**
деловой	0,57	0,80	0,33
художественный	0,33	0,41	0,26

Анализ НОС при монотонизации текста (искажении фонетической информации) лишает нас мелодической составляющей: искажаются характеристики НОС делового псевдо-М (он содержит одно существительное и ни одного КС). Доля существительных и КС в составе НОС художественного псевдо-М мало изменилась (см. табл. 2 и 3).

Таким образом, монотонизация художественного псевдотекста уменьшает объем НОС, но не изменяет его внефонетические характеристики.

3. Восстановление текста на основе наборов опорных слов

Признаки «функциональный стиль» и «степень динамичности», заложенные в НОС-псевдо, необходимо проследить также через восстановление внутритекстовых ассоциативных связей между словами (в том числе связей в разных составляющих: синтагмах, фразах, ситуациях и т.д.), т.е. через эксперимент по восстановлению текстов[7] на основе НОС-псевдо. Ранее нами был проведен соответствующий эксперимент на материале НОС-шум (Ягунова 2008б), его результаты использовались нами для сопоставления.

Эксперимент сможет пролить свет на вопросы о том, в какой степени два типа НОС-псевдо отражают смысловую структуру текста, а также о том, как работает контекстная предсказуемость при восстановлении смысловой структуры.

В основу описываемого далее эксперимента легли следующие ***гипотезы***:

- опора на НОС позволяет осуществить построение целостного связного текста;
- НОС задают функциональный стиль развертываемого текста;
- НОС отражают предметную (тематическую) область развертываемого текста;
- развертывание НОС позволяет определить позиции слов, обладающих максимальной предсказуемостью.

Эксперимент проводился в письменно-письменной форме. Опорные слова представляли собой фонетические слова, т.е. знаменательные словоформы с клитиками (предлогами, союзами, частицами). Анкета для испытуемых включала два НОС, записанных в столбцы под названием «текст 1» и «текст 2», и письменной инструкции. Столбцы соответствовали наборам для делового («текст 1») и художественного («текст 2») текстов одного экспериментального режима. В докладе рассматриваются некоторые результаты двух серий эксперимента: для начального фрагмента текста и всего текста. В письменной инструкции было указано: «*Перед Вами* ***последовательность*** *слов, извлеченных из текста. Попробуйте на их основе восстановить текст*», письменная инструкция дополнялась устным требованием сохранять грамматическую форму слов, вводить слова в текст в указанной последовательности и пожеланием минимизирования литературных и «философских» фантазий. Время на выполнение задания не ограничивалось.

Анкеты, не удовлетворяющие требованиям инструкции, далее не рассматривались. В результате было получено для НОС-псевдо-I и НОС псевдо-М — более чем по 20 анкет. Каждый испытуемый участвовал только в одной из серий эксперимента.

Определение того, принадлежат ли восстановленные тексты тому же функциональному стилю, производилось на основании двух критериев:

- заключения эксперта о принадлежности восстановленного текста к данному функциональному стилю (деловому или художественному),
- степени статичности vs. динамичности смены описываемых ситуаций.

[7] В экспериментах по восстановлению текстов участвовали испытуемые незнакомые с исходными текстами.

Ранее в соответствующем эксперименте для серии НОС-шум (более 30 анкет) основные гипотезы были подтверждены. Единственным исключением была невозможность для испытуемых восстановить динамичность художественного текста (состав и объем художественного НОС-шум).

4. Результаты эксперимента. Выводы

4.1. Деловой текст

- Все развернутые тексты — как и исходный текст — относятся к деловому функциональному стилю. Большинство развернутых текстов можно отнести к жанру нормативного акта.
- Все развернутые тексты — как и исходный текст — относятся к статическому варианту: как правило, предписание, регулирующее положение дел, или — в ряде случаев — описание некоторого положения дел.
- Развертывание НОС в текст не позволило определить позиции слов (конструкций), предсказывающихся не менее чем в 30% случаев.

4.2. Художественный текст

- Все развернутые тексты — как и исходный текст — относятся к художественному функциональному стилю.
- Большинство развернутых на основании НОС-шум художественных текстов характеризуется статичностью (отсутствием смены ситуаций). Как правило, при разворачивании НОС не восстанавливается сюжет с двумя действующими лицами, в тексте отсутствуют диалоговые фрагменты.
- Для НОС-псевдо-I и НОС-псевдо-М общий смысл и композиция художественного текста в подавляющем числе случаев восстановлены правильно. Восстанавливается наличие двух действующих лиц (в наборах присутствует имя одного из них, второе обозначается, чаще всего, с помощью личного местоимения я, он, она), чередование «нарративных» и диалоговых фрагментов, общая тема «выпрашивание денег».
- Развертывание НОС-шум в текст не позволило определить позиции слов, точно предсказывающихся не менее чем в 30% случаев.
- Развертывание НОС-псевдо-I и НОС-псевдо-М в текст позволило определить те позиции, в которых слова предсказываются не менее чем в 30% случаев (см. табл. 4).

Рассматривалось два вида контекстной предсказуемости: полная, т. е. точное восстановление единицы исходного текста, и смысловая, т.е. восстановление эквивалента единицы исходного текста с точностью до контекстного синонима. В таблицах 1 и 2 полужирным шрифтом выделены слова, предсказуемые более чем в 30% случаев.

Примеры полной восстановимости:

«*свое*» восстанавливается (см. контекст — выделенный квадрат в табл. 1б) в «*Пора бы знать* _____ *место!*» в 55% и 31% случаев (для НОС псевдо-I и псевдо-М, соответственно),

«*швырнул*» восстанавливается (см. контекст — выделенный квадрат в табл. 1б)

в «*С этими словами* _____ *с размаху монету на землю*» в 35% случаев (НОС псевдо-I)

в «*С этими словами он* _____» только в 8% случаев (НОС псевдо-М).

Под контекстным синонимом понимаются традиционные лексические синонимы или замены, достаточно точно отражающие общий смысл предложения (*одолевают* (о мыслях) → *пришли в голову*, *путаются*, *разбегаются*; *научись* (жить честно) → *пора, надо*).

Примеры смысловой восстановимости:

в указанных выше контекстах «*швырнул*» восстанавливается с точностью до контекстного синонима в 100% (*швырнул* 35%, *бросил* 35%, кинул 26% и пр.) и в 45% (*швырнул* 8%, *бросил* 8%, *кинул*, *протянул*, *дал*, *вынул*, и пр.) случаев (для псевдо-I и псевдо-М, соответственно).

Два типа НОС — псевдо-I и псевдо-М — были получены в результате экспериментов по восприятию псевдотекстов, исключающих действие механизмов контекстной предсказуемости. Если данные эксперимента по развертыванию наборов в текст интерпретировать как обнаружение слов, максимально предсказуемых контекстом, то эти НОС (см. табл. 1а и 1б) можно расширить за счет этих предсказуемых слов. В таблице 4 приведены наборы опорных слов и наборы предсказуемых слов для экспериментальных режимов восприятия художественных псевдотекстов; полужирным шрифтом выделены точно восстанавливаемые словоформы, курсивом с точностью до контекстного синонима.

Таблица 4. Расширенные наборы (НОС-псевдо и множество предсказуемых слов) для художественных псевдотекстов

псевдо-I		**псевдо-М**	
опорные слова	*предсказуемые*	*опорные слова*	*предсказуемые*
когда		когда	
	одолевают		

псевдо-I		**псевдо-М**	
опорные слова	*пред-сказуемые*	*опорные слова*	*пред-сказуемые*
мысли			
	сохранить		
и выдерж-ку			
а тут		а тут	
еще		еще	
Ти Фео		Ти	
клянчить			
на выпивку			
ему		ему	
пять хао		пять хао	
опять		опять	
пора бы		пора бы	
знать		знать	
	свое		***свое***
место		место	
здесь		здесь	
тебе		тебе	
не банк		не банк	
с этими		с этими	
словами		словами	
		он	
с размаху			
	швырнул		*швырнул*
монету			
на землю			
забирай		забирай	
и катись		и катись	
чтобы		чтобы	
духу			духу
	твоего		твоего
	не было		*не было*
живо			
сколько		сколько	
раз		раз	
говорил			Говорил
	научись		*Научись*
жить		жить	
честно		честно	

Предсказуемые слова включают глаголы и компоненты фразеологизмов (см. табл. 4); доля множество предсказуемых слов по отношению к набору опорных слов составляет 22% для НОС-псевдо-I и 30% для НОС-псевдо-М. Появление предсказуемых слов может являться следствием высокой предсказуемости компонентов фразеологизмов и устойчивых сочетаний: *чтобы духу твоего не было* и *пора бы знать свое место*. С другой стороны, появление предсказуемых слов — особенно предсказуемых глаголов — может характеризовать процедуры развертывания свертки в текст: процесс восстановления пропозиций как отражения описываемых ситуаций. Доля же существительных в расширенных НОС-псевдо уменьшается (как формальный показатель увеличения степени динамичности художественного текста): 0,41 → 0,33 для НОС-псевдо-I и 0,26 → 0,23 для НОС-псевдо-М.

Возможность восстановления искомых пропозиций на основании рассматриваемых наборов опорных слов, в свою очередь имеет и «обратную» трактовку. Так, одним из результатов эксперимента по восстанавливанию текста из НОС-псевдо является то, что общий смысл и композиция динамичного художественного текста в подавляющем числе случаев восстановлена правильно, т.е. наиболее распознаваемые на основании только фонетических характеристик (формально выделяемые) слова оказываются словами, важными для понимания текста.

Сама по себе монотонизация художественного псевдотекста существенно снижает разборчивость псевдослов (Ягунова 2008а). Однако «включение» процедур контекстной предсказуемости в эксперименте по восстановлению текста может существенно улучшать результаты восприятия (расширенный НОС-М и восстановленные тексты).

5. Вместо заключения

Выдвинутые гипотезы подтвердились. На основании проведенного исследования можно сделать следующие выводы:

- Существенное значение для смыслового структурирования текста имеет функциональный стиль и степень динамичности текста.
- НОС задают функциональный стиль тех текстов, что восстанавливаются испытуемыми в эксперименте.
- НОС-псевдо (распознаваемые на основании сохранных фонетических характеристик) позволяют адекватно восстановить тексты: «функциональный стиль», «степень динамичности», структура, предметная (тематическая) область) соответствуют исходному тексту.
- Распределение в тексте опорных слов, распознающихся за счет анализа лишь фонетических характеристик, несет существенную информацию о смысловой структуре текста.
- Расширение НОС-псевдо за счет предсказуемых слов художественного текста демонстрирует динамичность текста, проявляющуюся в закономерностях развертывания свертки в текст.

Исследование различных путей формирования НОС позволяет решить многие вопросы восприятия и понимания текста человеком. Бо-

лее того, хочется надеяться, что такое экспериментальное исследование сокращает существующую оторванность теории речевой деятельности от прикладных задач извлечения знаний из текста и распознавания речи. В реальных ситуациях анализа текста человек (или автомат как «искусственный носитель языка») способен распознать лишь ограниченное число фрагментов, остальное реконструируется на основе контекстной предсказуемости. В то же время даже фрагментарное распознавание позволяет восстановить смысловую структуру предъявляемого текста.

Литература

1. *Бондарко Л. В., Вербицкая Л. А., Гордина Л. А., Зиндер Л. Р., Касевич В. Б.* Стили произношения и типы произнесения // Вопросы языкознания, 1974, №2. — С. 64–70.
2. *Мурзин Л. Н., Штерн А. С.* Текст и его восприятие. Свердловск, 1991
3. *Ягунова Е. В.* Мелодические признаки и опорные элементы при восприятии текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2006» (Бекасово, 31 мая — 4 июня 2006 г.) / Под ред. И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея.– М.: Наука, 2006
4. *Ягунова Е. В.* Коммуникативная и смысловая структуры текста и его восприятие // Вопросы языкознания. 2007а, № 6
5. *Ягунова Е. В.* Фонетические признаки опорных сегментов и восприятие русского текста // Русский язык в научном освещении 2(14) М., 2007б.
6. *Ягунова Е. В.* Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). Пермь : Издательство Перм. ун-та, 2008а.
7. *Ягунова Е. В.* Набор опорных слов как вид свертки текста (в сопоставлении с набором ключевых слов) // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Периодическое издание. Выпуск 7 (14). М., 2008б.

Русские обращения: словарная информация и вокативные конструкции[1]

Russian vocatives: lexicon and constructions

Янко Т. Е. (tanya_yanko@list.ru)

Институт языкознания РАН

Рассматриваются предпочтения и запреты, связанные с формированием обращений. Выделяются лексические единицы и словоформы с близкой семантикой, но различным вокативным потенциалом. Анализируются специальные вокативные конструкции русского языка.

В работах Э. Щеглова и А. Цвикки показано, что английские слова с близкими значениями *doctor* ‘доктор’ и *physician* ‘врач’ в роли обращений ведут себя по-разному. *Doctor!* — это «хорошее» обращение, а *Physician!* — «плохое» [Schegloff 1968; 1972; Zwicky 1974]. Это наблюдение полностью приложимо и к русским словам *доктор* и *врач*. Мы говорим *Доктор, я буду жить?*, но не *?Врач, я буду жить?*. Совершенно аналогично, слова *водитель*, *шофер* и *таксист* в обращении к человеку, сидящему за рулем, обнаруживают различный вокативный потенциал. *Водитель!* — это «хорошее» обращение, а *Шофер!* и *Таксист!* — «плохие»: *Водитель! Откройте заднюю дверь!* vs. *?Шофер*! *Остановите у светофора!*.

Если в аудитории более одного слушающего, мы обращаемся к ней со словами *Друзья!*, *Коллеги!*, *Ребята!*, *Девочки!*, *Мальчишки!*, *Господа!*, *Офицеры!*, но не *?Гости!*, *?Соседи*!, *?Родственники*!, *?Студенты*!?. Однако чем друзья лучше родственников, а коллеги — соседей?

Почему слова, значение которых специально обозначает слушающего, то есть того, к кому мы обращаемся, например *радиослушатель* или *телезритель*, оказываются «плохими» обращениями? Дикторы радио и телевидения никогда не говорят *?Радиослушатели!*, а только *Дорогие радислушатели!* или *Уважаемые телезрители!*. От слов *радиослушатели*, *гости*, *соседи*, *телезрители* и других форма обращения обычно образуется с помощью прилагательных *дорогой* и *уважаемый*, даже если слушающий не всегда действительно дорог говорящему. Нам говорят *Уважаемые абоненты радиотрансляционной сети*!, *Дорогие гости!*, *Уважаемые пассажиры*!, а формы *Телезрители!, Зрители!* и *Гости!* как обращения не используются. Таким образом, обращения как иллокутивный акт вызывают много вопросов. Данная статья посвящена некоторым словарным свойствам слов в связи с их способностью использоваться в функции обращений, а также способам получения «хороших» обращений из «плохих» и анализу специализированных вокативных конструкций русского языка.

1. Информация о вокативном потенциале слова в словаре

Способность использоваться в качестве обращений — это словарное свойство слов. Отчасти это свойство может быть объяснено социолингвистическими традициями использования слов *доктор*, *врач*, *таксист*, *водитель*, *гость*, *друг*, *сосед*, *офицер*, *зритель*, *пассажир* и других. Анонимный рецензент настоящей работы обратил наше внимание на то, что способность слов использоваться в роли обращений может быть связана с категорией вежливости, которая выражается в большинстве языков, но наиболее разработана в языках Юго-Восточной Азии. Категория вежливости влияет на использование слов не только в качестве обращений, но и на именования людей в их присутствии. Соответственно, лексические единицы делятся на «вежливые» и «невежливые». Русское слово *шофер* и подобные ему в русском и в других языках не используются в качестве обращений и описаний присутствующих при акте речи. Заметим, однако, что некоторые слова, например *студенты*, *ученики*, *зрители* и другие, могут использоваться в качестве описаний, и не только заглазных, но не в качестве обращений. Можно предположить, что существуют градации вежливости, которые ранжируют ситуации, связанные с исполь-

[1] Работа над темой финансируется Российским Государственным гуманитарным фондом, проект 09-04-00106а.

зованием имен. Кроме того, имеются не только социолингвистические, но собственно семантические признаки «хороших» обращений. Один из шагов в анализе семантики слов, которые используются как обращения, будет сделан ниже в разделе 2.2. Там рассматриваются именования слушающего не по его постоянному или долговременному признаку, который выражен, скажем, существительным *таксист*, а по его актуальной роли в ситуации, ср.: *истец, дежурный, старший, встречающие мистера Питкина из Нью-Йорка*. Имена из второго класса имеют больший вокативный потенциал, чем имена из первого класса. Однако это не решает задачи классификации слов, потому что многие слова могут выражать как постоянный признак, так и актуальный. Так, *командующий* концептуализуется как выражение актуального признака, а *командир* и *главнокомандующий* могут выражать и постоянный (виртуальный) признак, и актуальный. Кроме того, некоторые имена, которые с семантической точки зрения выражают, казалось бы, актуальный признак, например *радиослушатели* или *телезрители*, получают в языке своего рода «виртуальную» концептуализацию, т. е. понимаются как постоянные признаки или индифферентны по этому параметру. Далее, слова могут начать в узусе свою собственную социолингвистическую жизнь, которая смещает семантические акценты, имеющиеся в толковании. Так, обозначения высокого места в социальной иерархии вследствие того, что говорящий по политесным соображениям склонен позиционировать слушающего как старшего, могут пополнить класс этикетных обращений, ср. такие имена, как *господин*, *батюшка*, *государь*, *шеф*, *начальник*, *командир*. Однако это не значит, что все имена со значением высокого социального статуса окажутся «хорошими» обращениями. Аналогично, по социолингвистическим причинам в качестве обращений закрепляются и многие слова со значением социального равенства между говорящим и слушающим, ср. *товарищ*, *коллега*, *земляк*, *подруга*. Таким образом, полностью решить задачу семантической классификации именования слушающего и использования имени в функции обращения не удается. Надежные семанические признаки «хорошего» обращения отсутствуют, поэтому информацию о вокативном потенциале конкретных слов следует приводить в словаре.

Кроме того, необходимо знать, какие обращения и в каких ситуациях следует использовать. Так, обращаясь с поздравлением, объявлением или тостом, мы можем сказать *Друзья!*, *Дорогие друзья!* и *Господа!*, но обратиться к аудитории со словами ?*Дорогие господа!* мы не можем. Это словарная особенность слова *господин*. Про некоторые обращения следует знать, что они используются только в начале официальной речи, например: *Офицеры, разрешите поздравить вас…!*; *Дамы и господа! Сегодня у нас замечательный день!*. Если же цель говорящего состоит в том, чтобы только поддержать уже начатый контакт, обращения *офицеры* и *дамы и господа* будут выглядеть неуместно: ?*Позвольте, дамы и господа, на этом закончить*; ?*Об этом мы поговорим на следующем заседании, офицеры!*. Между тем те же предложения, но с обращениями *господа* или *коллеги* звучат совершенно естественно: *Позвольте, господа, на этом закончить*; *Об этом мы поговорим на следующем заседании, коллеги!*

Пометы в словаре должны иметь характер рекомендаций по использованию лексем в функции обращений и в определенном типе дискурса. Так, в современном городском общении женщины, собравшиеся в кружок во дворе, как правило, не называют друг друга соседками — скорее, здесь будут использованы обращения *Женщины!* или *Девочки!*, но в другом обществе или при стилизации под другой способ общения, например в переводе на русский язык рассказа о жизни грузинской деревни, обращение *Соседки!* окажется совершенно нормативным.

В словаре следует указывать и грамматическое число, в котором слова используются в качестве обращений, потому что в роли обращения слово может функционировать только в какой-нибудь одной форме, например только во множественном или только в единственном числе. *Друзья!* и *Господа!* — это обращения, которые можно использовать в очень широком круге различных ситуаций общения, а *Друг!* и *Господин!* — в единственном числе — могут использоваться только очень ограниченно. Мы не можем обратиться к незнакомцу со словами ?*Господин! Не подскажете ли, где тут метро?*, при том, что обращение *Господа! Посторонитесь, пожалуйста, у меня тут коляска с ребенком*, хоть и несколько старомодное, но все-таки не выходит за рамки нормы. И наоборот, *Профессор!* — это «хорошее», хоть и несколько устаревшее, обращение в академических кругах, а *Профессора́!* — нет. Между тем *Коллега!* и *Коллеги!* прекрасно звучат при обращении и к одному слушающему, и ко многим. Тем самым, обращаясь по-русски к профессорам, мы называем их *Коллеги!* и *Господа!*, но не *Профессора́!*. *Доктор!* — это тоже «хорошее» обращение, а *Докторá!* как обращение не используется.

2. Вокативные средства

Возникает вопрос: существуют ли способы получить «хорошие» обращения из «плохих»? И какие выражения функционируют только в функции обращений? К специальным «вокативизирующим» контекстам можно отнести следующие средства русского языка.

- Слова *уважаемый*, *многоуважаемый* и *дорогой*, как было показано выше;
- Междометие *эй!* (*Эй, пассажиры, вы куда?*);

- Местоимение *мой* (*Моя радость!*; *Рыбка моя!*; *Моя звездочка!*; *Горе мое!*);
- Использование уменьшительно-ласкательных и других суффиксов (*Дружок!*; *Дружочек!*; *Дружище!*[2]);
- Использование вокативной конструкции *Эй* плюс имя места, где находится слушающий (*Эй, на барже!*);
- Развернутая характеризация слушающего по его актуальной роли в конкретной ситуации (*Таксисты, обслуживающие центр города, зайдите в диспетчерскую!*; *Студенты третьего курса, экзамен по математике переносится на понедельник*; *Пассажиры, находящиеся в первом и последнем вагоне! Будьте осторожны при выходе из поезда!*).

В связи с вокативизирующими лексемами *уважаемый*, *дорогой*, *мой* и суффиксами мы уже высказывались в работе [Янко 2008], поэтому ниже мы сосредоточимся на двух последних в этом списке типах вокативных контекстов, связанных с именованием адресата по месту (подраздел 2.1 ниже) и по его актуальной роли в ситуации (подраздел 2.2).

2.1. Эй, на барже!: место слушающего как обращение к нему

Интересным средством образования обращений служит русская конструкция с междометием *Эй* плюс имя места, в котором находится слушающий, например, *Эй, на барже!*. Имя места может показаться неожиданным способом обращения к слушающему, но такой тип зова слушающего довольно часто используется в русской речи, например, когда говорящий видит, что на капитанском мостике кто-то есть, но он не знает или не видит кто именно. И тогда говорящий кричит: *Эй, на мостике!*. На первый взгляд можно предположить, что такие обращения принадлежат исключительно морскому языку. Действительно, около 70 процентов обращений с *Эй* плюс имя места, которые встретились нам в интернете, касались водной тематики. Вот некоторые примеры: *Эй, на барже! Кидай якорь!*; *Эй, на том берегу!*; *Эй, на Акуле! — крикнул капитан*; *Эй, на линкоре!*; *Эй, на острове!*; *Эй, на корабле!*; *Эй, на судне!*; *Эй, на шхуне!*; *Эй, на мостике!*; *Эй, на ледоколе!*; *Эй, на лодке!*; *Эй, на румбе!*; *Эй, на берегу!*; *Эй, на борту!*; *Эй, на палубе!*; *Эй, на шлюпке!*; *Эй, на буксире!*; *Эй, на катамаране!*; *Эй, на руле! Круче к ветру!*; *Эй, на «Беде», счастливого плаванья!*; *Эй, на буксире, прими конец!*; *Эй, на «Малыгине»!*; *Эй, на баркасе!*; *Эй, на желтой лодке!*; *Эй, на камбузе!*; *Эй, на мостике!*; *Эй, на плотике!*. Однако оказывается, что «сухопутные» обращения такого типа тоже встречаются: *Эй, на задней площадке, все оплатили проезд?*; *Эй, на верблюде!*; *Эй, на трибуне!*; *Эй, на уазике!*. Значит, объяснять функции и особенности обращений по месту следует не принадлежностью к общению между собой морских волков, а особыми условиями, при которых используется обращение по месту.

Когда можно использовать обращение по месту? Почему *Эй, на трибуне!* — это «хорошее» обращение, а *Эй, в лесу!* и *Эй, в поле!* — «плохие»? Почему даже морская тематика не спасает обращение *Эй, на барже!*, если говорящий находится на другой барже? Почему, когда мы говорим *Иди вниз!*, низ понимается с точки зрения слушающего — здесь слушающий находится высоко и должен спуститься ниже (где находится говорящий, в данном случае неважно: он может быть и выше, и ниже, и на одном уровне со слушающим), а когда мы говорим *Эй, внизу!*, низ рассчитывается, наоборот, с точки зрения говорящего? И наконец, почему, когда мы слышим *Эй, на берегу!*, это значит, что говорящий находится на воде, а не с другой стороны, скажем, на шоссе, идущем вдоль берега?

Ответы на эти вопросы должен дать анализ обращения к слущающему в соответствии с тем местом, где он находится. Мы покажем, что обращения к слушающему по месту имеет более жесткие ограничения, чем другие способы локализации объекта в языке, т. е. вне иллокутивной силы обращения. Таким образом, перед нами встает задача показать, что такая периферийная конструкция русского языка, как обращение говорящего к адресату по месту, где он находится, представляет собой уникальный тип концептуализации пространства, воплощенный в языке.

Прежде всего, слушающий должен находиться в месте, которое легко выделяется в пространстве или в некоей системе ценностей. Параметр системы ценностей мы обсудим ниже, а для начала обратимся канализупространства. Идеальное место для слушающего, к которому можно обратиться со словами *Эй, на X-е!* — это лодка на водной глади, остров в океане, капитанский мостик, верхушка дерева на фоне неба, трибуна, сцена, вышка, при том что говорящий находится на известном удалении от этих точек, ср. понятия фигуры и фона, введенные в работе Л. Талми [Talmy 1978]. Фигура говорящего должна легко выделяться на общем фоне. Обращения ?*Эй, в поле!* и ?*Эй, в лесу!* «нехороши» потому, что в них отсутствует выделенная точка, противопоставленная местоположению говорящего. Поле и лес — это не точки в пространстве, а обширные просторы, на лоне которых адресат, скорее, затеряется, чем откроется взору говорящего. Заметим попутно, что другой — менее удобный и менее этикетный способ обращения — по шапке или по одежде — тоже основан на принципе легкого узнавания на общем однообразном фоне: *Эй, в шляпе!*; *Эй, в красной куртке!*.

[2] Существительное *дружище* используется только как обращение. Другие аналогичные примеры читатель найдет в статье Даниэль (в печати).

Далее. Говорящий и слушающий должны находиться на объектах, противопоставленных друг другу в некоторой системе ценностей. Что здесь понимается под системой ценностей? Обратимся к примерам. Пусть говорящий едет в машине по шоссе, а слушающий стоит у дороги. Тогда говорящий может обратиться к слушающему со словами *Эй, на обочине!*, потому что дорога и обочина дороги принадлежат одной системе ценностей и представляют собой разные полюса одной системы. Адресат находится на точке пересечения края дороги с линией, соединяющей направление взора говорящего, устремленного на слушающего. Это хорошо видная точка. Если же говорящий вышел из леса и увидел слушающего, стоящего также на обочине дороги, он, скорее всего, обратится к нему со словами *Эй, у дороги!*, потому что кромке леса противопоставлена дорога, а не ее обочина. Обочина дороги противопоставлена не лесу, а прохожей или проезжей части дороги.

Приведем другие примеры. Мы говорим *Эй, на чердаке!*, если сами мы находимся в том же доме, скажем, на лестнице, ведущей на чердак. Если же мы сами находимся в другом доме или на улице, обращение к сидящим на чердаке другого дома *Эй, на чердаке!* оказывается неуместным. Здесь мы можем крикнуть (например, в громкоговоритель, как это бывает в детективных сериалах): *Эй, на вилле! Выходи по одному!*. Обращение *Эй, в доме!* здесь тоже не подойдет, потому что вокруг дома могут быть другие дома, следовательно, место слушающего не будет считаться выделенной точкой. Наша гипотеза состоит в том, что концептуализация дома в русском языке предусматривает нахождение дома в окружении других домов. Не будет удачным и обращение *Эй, на даче!*, потому что дача тоже концептуализуется как летний дом в кругу других таких же домов. Вилла же и хутор понимаются как центры достаточно большого незастроенного пространства, поэтому они больше подходят для обращения по месту, чем дома и дачи: *Эй, на хуторе!*. Хорошо подходят для обращений по месту строения, имеющие уникальные имена: *Эй, в «Приюте орла»!*; *Эй, в «Белом олене»!*; *Эй, в «Сороке и пне»!*. Однако это не могут быть названия слишком больших зданий, несоразмерных фигуре говорящего, который надеется докричаться до слушающего: обращения *Эй, в Кемпински!*, *Эй, в Хилтоне!* в прямом — не-метафорическом — значении невозможны.

Таким образом, именование говорящим слушающего зависит не только от того, где находится слушающий, но также и от того, где находится сам говорящий, а также от того, образуют ли место расположения говорящего и место расположения слушающего естественную систему, такую как дорога — обочина, край леса — дорога, идущая вдоль леса, трибуна — зал, чердак — пространство внутри дома под чердаком, водоем — берег. Именование слушающего без учета системы ценностей, в которую встраивается его локализация, по-видимому, тоже возможно, но предпочтительны обращения, которые учитывают такие противопоставления. Очевидно, что учет взаимного расположения говорящего и слушающего накладывает на способ вербализации места слушающего больше ограничений, чем при выражении в языке места объекта безотносительно к расположению других объектов в пространстве.

Покажем теперь, что в обращении по месту точкой отсчета (или центром ориентации в пространстве) служит фигура говорящего и что при других типах вербализации места точкой отсчета могут быть другие объекты. Какой объект станет точкой отсчета — говорящий, слушающий или объект, названный в тексте? Когда мы говорим слушающему *Спускайся вниз!*, низ здесь понимается с точки зрения слушающего, а когда мы говорим *Пусть Вася идет домой*, здесь имеется в виду дом Васи, названного в тексте. В обращении же по месту точкой отсчета служит именно говорящий. Об этом говорят обращения *Эй, внизу*!; *Эй, наверху*!; *Эй, на том берегу!*, где низ, верх и противоположный берег отсчитываются от того места, на котором расположен говорящий.

И наконец, если говорящий и слушающий находятся на одноименных объектах, обращение по месту оказывается неуместным. Чисто теоретически говорящий, находящийся на барже, мог бы обратиться к слушающему со словами *Эй, на другой барже!* по аналогии с обращением *Эй, на том берегу!*, но реально такие обращения не встречаются.

Мы показали, что в обращениях по месту есть немало ограничений. Говорящий и слушающий должны располагаться на разноименных объектах, которые принадлежат одной системе ценностей, слушающий должен находиться в выделенной точке пространства, и точкой ориентации в этом пространстве должна служить фигура говорящего.

Последнее замечание, которое тут следует сделать, состоит в том, что если какое-либо из условий называния слушающего по месту не выполняется, добавление в обращение местоимений *ты* и *там* (*Эй ты, на дороге!*; *Эй там, в поле!*; *Эй ты, там на улице!*) восстанавливает соблюдение указанных условий, потому что превращает место расположения слушающего в противопоставленную и соразмерную фигуре говорящего выделенную точку, которая называется *там* (т.е. в том месте, которое противопоставлено месту говорящего), или понимается как место слушающего, который противопоставлен говорящему и называется *ты*.

2.2. Именование слушающего по его актуальной роли в ситуации как вокативное средство

И. И. Ковтунова в специальном разделе, посвященном поэтическим обращениям, книги по по-

этическому синтаксису выделила особый класс обращений-характеризаций, которые сочетают в себе одновременно номинативную и предикативную функции. В поэзии такие обращения утрачивают собственно вокативную функцию и используются для характеристики адресата речи [Ковтунова 1986: 108]: *Дуб богоборческий! В бои Всем корнем шествующий! Ивы-провидицы мои! Березы-девственницы!* (Цветаева); *Я посетил тебя, пленительная осень, не в дни веселые…* (Баратынский); *Испепеляющие годы! Безумья ль в вас, надежды ль весть?* (Блок). Можно заметить, что имена с характеризациями всегда служат «хорошими» обращениями.

2.2.1. Обращения-характеризации

Нетрудно также заметить, что обращения-характеризации, основанные на предицировании имени слушающего характеризующих признаков, широко используются не только в поэтической речи, но и в сугубо деловом жанре объявлений, как правило, письменных. Здесь развернутые определения в структуре обращений ограничивают и актуализируют круг лиц, которым адресованы объявления: *Студенты третьего курса, просьба после окончания сессии сдать зачетки в учебную часть*; *Ученики старших классов, вас приглашает на собеседование электромеханический колледж г. Москвы*; *Владельцы пластиковых карт Виза Аэрофлот, для вас объявляется акция «Весна в Юго-Восточной Азии»*; *Мастер транспортного цеха судостроительного участка, пройдите к пульту*; *Кто забыл в туалете бриллиантовое кольцо, обратитесь к бортпроводнику*. В тексте *Студенты третьего курса, просьба после окончания сессии сдать зачетки в учебную часть* содержится не только обращение к студентам, но им также приписывается принадлежность к некоторому более узкому классу, чем класс студентов вообще, и обладание некоторым актуальным признаком.

Если сравнить обращения без определений и обращения с определениями, а также другие обращения со сложной лексико-синтаксической структурой, можно заметить, что определения-характеризации служат мощным вокативизирующим контекстом: *?Студенты, после третьей пары зайдите ко мне в кабинет* vs. *Студенты из тех, кто работает сейчас с информантами, зайдите ко мне в кабинет*. Аналитические именования лиц (*учащиеся старших классов*; *тот, кто забыл кольцо*; *третьекурсники*; *таксист Иванов*; *кассир Сидоров*) преобразуют имена, характеризующиеся некоторым виртуальным специализированным признаком (‘быть по профессии таксистом’, ‘быть студентом’, ‘быть врачом’) в имена, которые характеризуют адресата по его актуальной роли в реальной ситуации (*быть студентом третьего курса*; *быть таксистом таксопарка, который обслуживает центр города*, *быть врачом-ординатором, проходящим практику на базе Первой Градской больницы*). Если обращения *врач!*, *таксист!* и *студент!*, как было показано выше при постановке задачи, принадлежат к словарному классу «плохих» обращений, то определения-характеризации переводят их в «хорошие» обращения, именующие адресата по его актуальной роли в текущей ситуации. Имя роли адресата служит для обращения к нему.

2.2.2. Лексические единицы, служащие именами ролей

Развернутые характеризации это не единственное средство актуализации имени лица как исполнителя конкретной роли в ситуации. Существуют лексемы, которые концептуализуются именно как роли и вне контекста характеризующих определений. К ним относятся существительные *председатель*, *истец*, *ответчик*, *судья*, *свидетель*, *покупатель*, *начальник*, *абитуриент*, *соискатель*, *победитель*, *охрана*. Как актуальные роли в ситуации концептуализуются и существительные, которые по происхождению являются субстантивированными прилагательными и причастиями. Соответствующие концепты сохранили семантическую связь с категорией предикативности и отражением актуальной роли в ситуации, ср.: *верующие*, *больной* и *беременная* (в условиях лечебного учреждения), *рядовой*, *старший*, *дежурный*, *связной*, *вахтенный*, *заведующий*, *заключенный*, *потерпевший*, *осужденный*, *уполномоченный*, *раненый*, *нарочный*, *новобрачные*, *новорожденный*, *сопровождающий*, *встречающие*, *провожающие*, *отъезжающие, побежденный*. Такие существительные также служат источником «хороших» обращений: *Покупательница, возьмите сдачу!*; *Не сердись, начальник!*; *Больной, вам в другой кабинет!*; *Беременная, вы куда?* (Этот пример нам сообщил И.А. Шаронов); *Истец, передайте заявление адвокату*; *Соискатель, вам присуждается кандидатская степень*; *Дежурный, пройдите к пульту!*; *Охрана! На выход!*; *Провожающие, проверьте, не остались ли у вас билеты отъезжающих пассажиров*.

Итак, в связи с семантическим противопоставлением ‘человек-носитель признака’ vs. ‘роль человека в ситуации’ выделяется три класса имен со значением лица. 1) Большинство существительных могут обозначать как человека, обладающего (постоянным или долговременным) признаком, так и роль человека в конкретной ситуации. Это существительные *отец*, *командир*, *профессор*, *экзаменатор*, *солдат*, *генерал*. 2) Некоторые — немногочисленные — существительные могут обозначать только роль: *истец*, *заявитель*, *председательствующий*. 3) И наконец, к третьему типу относятся относятся имена, преимущественно обозначающие человека-носителя постоянного (виртуального) признака: *врач*, *таксист*, *интеллигент*. Имена второго класса со значением актуальной роли, которую может играть некий анонимный адресат в некоторой конкретной ситуации, свободно используются в вокативной функции.

Границы классов нечетки: имена могут переходить из одного класса в другой. Регулярным средством перевода слова из первого и третьего класса во второй служат характеризующие определения.

Приведенная классификация может объяснить некоторые запреты и предпочтения, существующие при формировании обращений в русском языке. Так, можно предположить, что слово *водитель* принадлежит к классу «хороших» обращений, потому что обозначает водителя конкретной машины в определенной ситуации на дороге: водитель этот тот, кто сидит за рулем. Когда мы говорим *Водитель, откройте заднюю дверь*!, мы имеем в виду человека, который ведет конкретный трамвай. Между тем аналогичные обращения, использующие лексические единицы *шофер* и *таксист*, например *Шофер, остановите у светофора!*, не запрещены, но статистически маловероятны, потому что *шофер* и *таксист* выражают постоянный и узко специальный признак человека, а именно — его профессию, а не его роль в ситуации.

Объяснительная сила этой классификации невелика. Она не охватывает всех запретов и предпочтений, потому что, как уже говорилось, под влиянием контекста слова могут переходить из одного класса в другой. Кроме того, классификация не учитывает социолингвистических факторов, которые в формировании обращений могут иметь решающее значение. И наконец, парадоксальным образом лексические единицы, онтологически выражающие временную функцию, такие как *телезритель* и *радиослушатель*, могут концептуализоваться в языке не как имена актуальных ролей, а как имена виртуальных признаков. По сочетаемостному поведению слов *телезритель* и *радиослушатель* можно предположить, что зрители, телезрители и радиослушатели понимаются не как те, кто в текущий момент времени смотрит на сцену, сидит у телеэкрана или вслушивается в звуки радиоприемника, а как люди, имеющие постоянные занятия и пристрастия, такие как смотреть телевизор, ходить в театр или слушать радио. Мы предполагаем, что зрители концептуализуются в языке как театралы, т. е. как люди, которые любят ходить в театр. Актеры говорят: *Наши зрители! Они любят наше искусство. Мы работаем для них*. Здесь зритель понимается как константа, а не как меняющаяся раз от разу аудитория. В этом сказывается общая концептуализация имени существительного как обозначающего постоянный признак или множество признаков. Этой проблеме посвящена большая литература, см. например работу [Рахилина 2000: 28] и цитированную там литературу. Исключениями из этого класса концептуализаций служат субстантивированные прилагательные и причастия, такие как *старший*, *дежурный* и *встречающие*, которые сохранили способность обозначать актуальный признак лица, а также некоторые исконные существительные, служащие только именами ролей, такие как *истец* и *соискатель* (ср. неудачные сочетания **краснощекий истец*, **седоватый ответчик*, *?худощавый соискатель*, которые говорят о том, что эти существительные обозначают не человека, а только его роль, ср., с другой стороны, нормативные сочетания *краснощекий / седоватый / худощавый генерал*).

Итак, для объяснительных целей достоверно можно использовать только одну полученную здесь импликацию: если имя лица концептуализуется как роль, оно может функционировать в качестве обращения.

***

В данной работе были рассмотрены предпочтения и запреты, связанные с формированием обращений. Были выделены лексические единицы с близкой семантикой, но различным вокативным потенциалом, и рассмотрены специальные вокативные конструкции русского языка, которые используются при именовании адресата речи по месту, в котором он находится, и по его актуальной роли в конкретной ситуации.

Литература

1. *Даниэль М. А.* (в печати) Звательность как дискурсивная категория. Несколько гипотез.
2. *Ковтунова И. И. Поэтический синтаксис. М.: Наука, 1986.*
3. *Рахилина Е. В.* Когнитивный анализ предметных имен: семантика и сочетаемость. М. Русские словари. 2000.
4. *Янко Т. Е.* Интонационные стратегии русской речи в сопоставительном аспекте. М. ЯСК. 2008. С. 97–106.
5. *Янко Т. Е.* Просодия в толковом словаре и словарь уникальных просодий // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2008 по компьютерной лингвистике и ее приложениям. 2008. С. 595–601.
6. *Schegloff, E. A.* Notes on a conversational practice: formulating place// P. P. Giglioli (ed.) Language and social context, Harmondsworth: Penguin, 1972.
7. *Schegloff, E. A.* Sequencing in conversational openings // American anthropologist 70.6. 1968.
8. *Talmy L.* Figure and ground in complex sentences // Greenberg et al. (eds)/ universals of human language. Stanford/ Vol. 4. P. 625–649/
9. *Zwicky A.* Hey, whatsyourname! // Papers from the Tenth Regional Meeting of the Chicago Linguistics Society. Chicago: Chicago Linguistics Society, 1974.

Создание и использование многоязычного корпуса объектно-ориентированных топонимических текстов для оптимизации задачи автоматического генерирования описания изображений

Development and implementation of multilingual object type toponym-referenced text corpora for optimizing automatic image description generation

Gornostay T. (tatjana.gornostaja@tilde.lv)
Tilde, Riga (www.tilde.com)

Aker A. (a.aker@dcs.shef.ac.uk)
Department of Computer Science, University of Sheffield

С точки зрения обработки стремительного роста объема графической информации в сети Интернет целесообразна разработка автоматических методов генерирования ее описаний. В последнее время используется метод автоматического реферирования набора документов определенной тематики. В настоящей статье описывается подход к созданию и использованию многоязычного корпуса объектно-ориентированных топонимических текстов для четырех языков (английского, немецкого, итальянского и латышского) в контексте оптимизации задачи автоматического реферирования для генерирования описаний графических изображений топонимов.

1. Introduction

In recent years the number of images on the Web has experienced an immense growth facilitated by the development of affordable digital hardware and the availability of online image sharing social sites. For successful indexing and retrieving the available images, their content has to be correctly identified and annotated. One way of annotating images is by tagging them with image descriptions.

Image description, or *summary*, is a more general term for *image index* (Hollink et al., 2004) which is used broadly in both image indexing and retrieval disciplines. Image descriptions can contain miscellaneous information about an image, including generic, specific, and interpretative explanation of what is shown in the image (Shatford, 1986). Apart from being essential for automatic indexing and retrieval of images, image descriptions are also useful for human users to get information about the content of the image. For example, descriptions of images showing locations could help a user who seeks information about a certain place, or a journalist, who writes an article about a location, or a tourist who looks for interesting places to visit (Aker and Gaizauskas, 2008).

Image descriptions can be written manually by individuals who are specially hired for this purpose or describe images for their private usage. However, manual generation of image descriptions is a tedious, expensive and inaccurate task (Pan et al., 2004; Jamieson et al., 2007). Therefore, methods of automatizing image description generation have been developed recently.

There are different approaches to automatic generation of image descriptions (Mori et al., 2000; Barnard and Forsyth, 2001; Duygulu et al., 2002; Barnard et al., 2003; Pan et al., 2004; Deschacht and Moens, 2007; Feng and Lapata, 2008). All these approaches generate image captions based on texts associated with the image in combination with or without analysis of image features (color, shape, texture). The resulting image descriptions contain named entities (e.g. person names), and/or a set of open class words (nouns, verbs, adjectives and adverbs) describing the image.

Another approach of image captioning is that of Aker and Gaizauskas (2008) who apply generic and

query-based multi-document summarization techniques to generate image descriptions for toponym-referenced images. *Toponyms* are terms describing places, such as *Westminster Abbey*, *University of Sheffield*, etc. *Toponym-referenced images* are images tagged with toponyms. Image descriptions generated automatically for toponym-referenced images can be called a *toponym-referenced description*, or *summary*.

In contrast to other researchers, who assume that the image has an associated text with it, Aker and Gaizauskas (2008) generate toponym-referenced descriptions from multiple web documents containing toponym-referenced texts which describe the places reflected in the image. The web documents are retrieved using the toponyms of the image.

Aker and Gaizauskas (2008) have shown that query-based toponym-referenced descriptions outperform generic descriptions. Query-based summaries are generated by biasing the summarizer towards the query which is the set of toponyms associated with the image. This makes sure that sentences which contain the query (toponyms) are more highly scored than the ones which do not contain any query term. In contrast, for generation of generic summaries the query (toponym) is not used to bias the summarizer. Therefore, the sentences containing the toponym are not necessarily scored more highly than those which do not contain any query term.

Although toponym-referenced descriptions generated by query-based summarizer were better than those generated by the generic summarizer, the authors noted that the best image descriptions were still far from perfect. The evaluation showed that the agreement between query-based descriptions and human generated summaries was not satisfactory. In contrast, the agreement between descriptions generated by humans for each image was high. This observation allowed the authors to hypothesize that humans have some conceptual model of what is salient regarding a certain *object type* of a toponym (churches, bridges, etc.). The authors suggest collecting existing textual resources about object types as one way of capturing these conceptual models. More specifically, they propose to compile a corpus of toponym-referenced texts for each object type and use the information commonly associated with each object type in the corpus to bias the summarizer. For example, a collection of texts describing object type *church* would contain information about the age of the church, dates of construction, the architectural style, its height, etc. This object type specific information can be used by the summarizer to give higher score to sentences which also contain this information.

In this paper we describe a possible way to derive such object type text corpora from the Web for different languages. We use Wikipedia as a resource and demonstrate how each Wikipedia article can be categorized automatically by object type for English. We also show how such a corpus can be extended to further languages. We are in particular interested in four languages: English, German, Italian and Latvian, for which we collect object type corpora. Finally, we exemplify how such object type corpora can be used for automatic multi-document summarization.

In Section 2 we define our requirements, give an overview of existing text resources on the Web and argue for our choice of Wikipedia as the most suitable resource to derive an object type corpus. Wikipedia is described in Section 3. Section 4 focuses on the procedure of automatic categorization of Wikipedia articles by object types and reports on evaluation of our categorization procedure. We also describe how object type corpora of German, Italian and Latvian can be obtained from the English object type corpus. In Section 5 we explain how object type corpora can be used for automatic generation of image descriptions using multi-document summarization. Section 6 concludes the paper and outlines future directions.

2. Corpus development scenarios

Object type text corpora that we aim to build for our target languages are collections of toponym-referenced texts for single object types (churches, bridges, etc.) compiled from texts available on the Web. The Web is generally accepted to be the most suitable and helpful resource for corpus development (Kilgarriff and Grefenstette, 2003; Cheng et al., 2004; Liu and Curran, 2006; Bernardini et al., 2006; Kilgarriff, 2007). Its main advantage is that a large amount of textual data about locations is available in electronic form and multiple languages and can be efficiently searched and processed. A particularly important aspect behind our decision to use the Web is that one of our target languages, Latvian, is considered to be an under-resourced language with few corpus resources available, e.g. the JRC-Acquis corpus (Steinberger et al., 2006) and Contemporary Latvian language text corpus[1]. However, none of the currently existing resources is rich in toponym-referenced descriptions of places. Using the Web as a resource for corpus development is therefore a way to avoid data scarcity problems that would arise from lack of toponym-referenced texts in existing Latvian corpora.

Collecting a corpus from the Web is a challenging task since the Web is unstructured, not systematically organized, and does not have any definite directory (Kuo and Yang, 2004; Liu and Curran, 2006: 234). Therefore, searching domain-specific text resources on the Web often appears to be a low precision activity (much invaluable and unnecessary information can be obtained) (Kuo and Yang, 2004; Hughes, 2006). In addition, the following observation is relevant in our case with respect to Latvian: "*finding relevant*

[1] www.korpuss.lv

materials for low density <…> languages on the Web is in general an increasingly inefficient exercise even for experienced searchers" (Hughes, 2006). To alleviate these disadvantages, we use Wikipedia as a single resource for derivation of toponym-referenced text corpora from the Web. Wikipedia is a well structured Web resource, rich in location descriptions and available in multiple languages, including our four target languages. We therefore consider it particularly suitable for our purposes.

3. Wikipedia as a corpus

Wikipedia is a free multilingual encyclopedia project by the non-profit Wikimedia Foundation[2]. With 11 million articles written in different languages it is currently the largest and most popular general reference work on the Web. Altogether 265 languages are represented in Wikipedia. Different languages have different growth rates. In 2006 there were 2 languages with no article other than the main page, 5 languages have about 100 articles, and 25 languages less than 100 articles (Adafre and Rijke, 2006). The remaining languages are now represented by more than 100 articles. For example:

- English 2.7 Million
- German, Spanish, French, Italian, Polish: 300,000
- Esperanto, Catalan, Ukrainian: 100,000
- Bulgarian, Estonian, Lithuanian: 50,000;
- Latin, Macedonian: 20,000.

As these numbers show, the overwhelming part of Wikipedia is more or less developed and constitutes a rich resource for study of language. It has recently been successfully used for a number of natural language processing tasks like deriving a large scale taxonomy (Ponzetto and Strube, 2007), named entity recognition and translation in question answering (Bouma et al., 2006), named entity disambiguation, translation and transliteration (Wentland et al., 2008), domain specific query translation in multilingual information access (Jones et al., 2008) among others.

Wikipedia has at least three features that can be of use in corpus development: redirection pages, disambiguation pages, and internal links.

Redirection pages are used to normalize different variants or synonyms of a given concept (Bouma et al., 2006; Wentland et al., 2008), e.g. *Rīga* and *Rīgas pilsēta* — the capital of Latvia. Figure 1 shows an example of a redirection page. The redirection to the main page *Rīga* is given as: #REDIRECT [[Rīga]].

Disambiguation pages illustrated in Figure 2 are used to disambiguate homonyms (Wentland et al., 2008), e.g. *Rīga* as Latvia's capital, sport team, airport, cinema, etc.

[2] http://en.wikipedia.org/wiki/Wikipedia

<page>
<title>**Rīgas pilsēta**</title>
<id>**66353**</id>
<revision>
<id>**416001**</id>
<timestamp>**2008-10-07T09:21:13Z**</timestamp>
<contributor>
<username>**Juzeris**</username>
<id>**23**</id>
</contributor>
<comment>**Pāradresē uz [[Rīga]]**</comment>
<text xml:space="**preserve**">**#REDIRECT [[Rīga]]**</text>

Figure 1. Wikipedia redirection page from the article about Rīgas pilsēta to the article about Rīga in XML format

Rīga (nozīmju atdalīšana)

Vikipēdijas raksts

Rīga var būt:

- Rīga, Latvijas galvaspilsēta
- sporta komandas:
  - ASK Rīga, dažādas komandas
  - FK Rīga, futbola komanda
  - HK Rīga 2000, hokeja komanda
- Arēna Rīga
- Starptautiskā lidosta "Rīga"
- Kinoteātris "Rīga"
- Rīga, mazā planēta, atklāta 1966
- Rīga, mopēds

Skatīt arī

apdzīvotas vietas ASV:

- Riga, pilsēta Ņujorkas štatā, ASV
- Riga Township, ciemats Mičiganā, ASV

personvārds:

- *Riga Mustapha*, Ganā dzimis nīderlandiešu futbolists

Figure 2. Wikipedia disambiguation page about Rīga

Wikipedia is a hypertext document and has an internal link structure. Wikipedia's internal links can be of two types: *cross-article* links and *cross-language* links. Usually the text of a Wikipedia article contains links to other articles, cross-article links. For example, the article about the river *Abava* has 17 links to other Wikipedia articles. These articles are topically associated with the main article and refer to more detailed or more general information related to it (Adafre and Rijke, 2005). Cross-language links are links from any page describing an entity in one Wikipedia language to a page describing the same entity in another language (Wentland et al., 2008). For the article describing *Riga*, there are 97 links to Wikipedia pages in other languages with the same entry. The format of cross-language links is ***[[language code: Title]]***, e.g. ***[[ru: Рига]]*** the link to the article about *Riga* in Russian.

Since Wikipedia articles in different languages on the same topic are closely related, this relation can be of use in deriving comparable corpora for different languages. Several research studies use this feature of Wikipedia, demonstrating that Wikipedia has reached a level where it can support multilingual research (Adafre and Rijke, 2005, 2006; Bouma et al., 2006; Declerck et al., 2006; Jones et al., 2008). In our work cross-language links were used to develop object type corpora for languages other than English (cf. Section 4.2).

4. Wikipedia's content categorization procedure

To build the object type corpus for English we categorized each article in the entire Wikipedia dump from 24/07/2008 by *object type* by applying *Is-A patterns*. Our patterns are similar to the ones described in Hearst (1992) and Mann (2002) who used manually written patterns to extract hyponyms from large text corpora. Is-A patterns are described in the automaton shown in Figure 3.

We took a Wikipedia article, split it into sentences and POS tagged each sentence using shallow text analysis tools (OpenNLP tools[3]). Then each sentence was checked for the occurrence of an Is-A pattern. If multiple Is-A patterns were found, the first one was taken. In case it matched, the part of the sentence before the pattern was deleted and only the subsequent text was kept. Example 1 demonstrates the first sentence of the article about *Westminster Abbey*:

(22) *The Collegiate Church of St Peter at Westminster, which is almost always referred to by its original name of Westminster Abbey, is a large, mainly Gothic church, in Westminster, London, just to the west of the Palace of Westminster.*

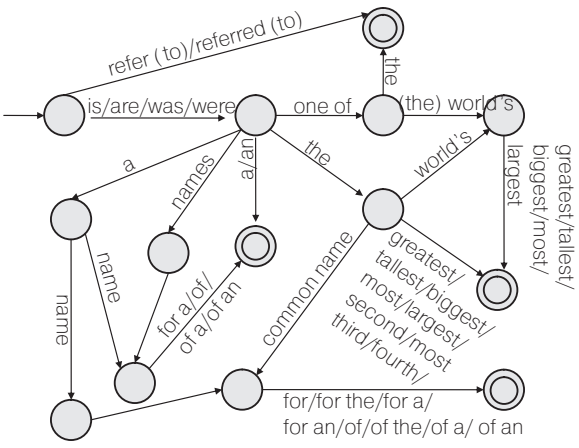


Figure 3. Is-A Patterns

[3] http://opennlp.sourceforge.net/

Is-A pattern was matched in the sentence and everything until the first occurrence of "*large*" was deleted, so that the following short version of the sentence was kept (Example 2).

(23) large/ADJ mainly/ADV Gothic/ADJ church/N ,/, in/P Westminster/PN ,/, London/PN ,/, just/ADV to/P the/DET west/N of/P the/DET Palace/PN of/P Westminster/NP

In order to find the category of the image, noun phrase (NP) search was applied to the short version of the sentence. For NP search, an approach similar to Bennett et al. (1999) rule-based NP identification for medical texts was implemented. The texts are pre-processed by tokenization and POS tagging and NPs are identified on basis of rules composed of different sequences of POS tags as shown in the automaton in Figure 4.

The noun input in the automaton is the starting POS tag for any sequence of rules. Therefore, the first occurrence of a noun in the shortened version of the sentence is searched in the beginning. Then, further POS tags which can be combined with a noun, e.g. Noun + Noun, Noun + Possessive, Noun + Possessive + Adjective + Noun, etc. are checked for. There is a possibility in each state of the automaton to terminate if the new input (*something else*) is not allowed to follow the term or POS tag found before. After an NP has been found, the nouns which occur at the end of the NP as the object type of the image are extracted. In the example with the above shortened sentence for *Westminster Abbey*: ***church/N*** was found as the first noun, and nothing described by the automaton was found after it. Therefore the automaton terminates and ***church*** is returned as the object type for the article about *Westminster Abbey*.

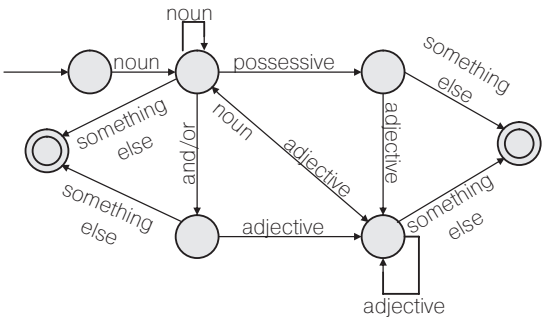


Figure 4. Noun Phrases

In this way about a half of Wikipedia articles (2.1 of the 5.4 millions) were automatically categorized. All together 40648 categories have been identified by our procedure. However, not all of these categories are about places, e.g. there are categories such as politician, leader, person, etc. which don't describe places and are not useful for our purposes. We manually filtered all identified categories in order to retain the ones describing places. That resulted in a set of 734 categories. From that set we retained 175 categories which are associated

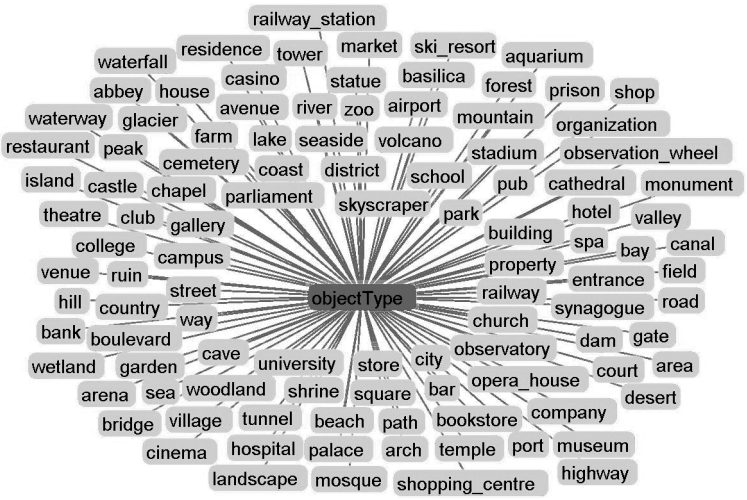


Figure 5. Object Types

with at least 50 Wikipedia articles. Finally, we manually assigned specific categories such as ***catholic church*** to a more general category ***church***. In this way we collected 107 categories containing articles about places around the world (Figure 5).

4.1. *Is-A pattern* evaluation

To assess the accuracy of the object type categorization we randomly selected 35 object type corpora and 50 articles from each corpus. Then we checked for each of these articles whether it is correctly or wrongly assigned to its object type. Finally, we calculate an accuracy value for each object type by dividing the number of correctly assigned articles by 50 (cf. Table 1). We observed an average accuracy of 80% for all 35 object types.

Table 1. Object types and the accuracy of the categorization

shopping center	0.9	bank	0.74
ski resort	1.0	monument	0.62
mountain	0.92	university	0.98
highway	0.82	building	0.52
railway station	1.0	park	0.96
mosque	0.66	gallery	0.725
waterfall	0.88	museum	0.7
street	0.58	canal	0.82
landscape	0.5	temple	0.74
restaurant	0.86	tower	0.52
island	0.92	prison	0.83
airport	1.0	residence	0.8
area	0.64	aquarium	0.62
volcano	0.92	castle	0.86
village	0.96	bridge	0.72
zoo	0.96	waterway	0.83
arena	0.96	river	0.94
wetland	0.79	average accuracy	0.80

A similar categorization has already been conducted for Wikipedia articles. It is stored in an ontology DBpedia[4]. However, the categorization of places by object types in the current version is not precise enough because it does not cover all object types identified by our procedure, e.g. churches, volcano, valley, pub, etc.

4.2. Multi-lingual object-type corpora collection

To obtain multi-lingual object type corpora we used cross-language links (cf. Section 3) to find Wikipedia articles in other languages for each article in the English corpus. For instance, the article about *Eiffel Tower* from the English Wikipedia was added to the object type ***tower***. Using cross-language links from this article we were able to obtain articles in those languages for which an article about the *Eiffel Tower* exists. Following this strategy we have collected object type text corpora for German, Italian and Latvian. Table 2 below shows the top-20 object types and the number of articles found for each object type for the four languages: English (EN), German (DE), Italian (IT), and Latvian (LV).

Table 2. Top-20 object types with the number of articles for English, German, Italian, and Latvian

Language / Object type	EN	DE	IT	LV
Airport	6493	560	201	9
Area	6934	731	382	31
Church	3005	392	371	7
City	14233	3788	2815	316
Company	5734	625	250	7
Country	3186	349	267	100

[4] http://www4.wiwiss.fu-berlin.de/dbpedia/dev/ontology.htm

Table 2. *Ending*

Language / Object type	EN	DE	IT	LV
District	6565	858	860	26
Island	6400	1292	689	73
Lake	3649	500	190	151
Mountain	5290	952	534	11
Museum	2320	277	121	0
Organization	9393	673	332	17
Park	3754	425	154	11
River	5851	1102	604	64
Road	3421	980	455	14
School	15794	292	128	5
Stadium	3665	509	274	11
University	7101	901	211	14
Village	39970	3550	3042	150
Way	2508	284	172	20

5. Example use of the object type corpus: Deriving language models for automatic summarization

Object type corpora for different languages described in the previous section can be used to derive language models to improve the results of multi-lingual image summary generation.

In multi-document summarization *language models* are used to improve the sentence selection procedure. Jagadeesh et al. (2005), for instance, generate language models using the query terms and a large text corpus to approximate the probability of a word occurring in a sentence relevant to the query, i.e., a sentence is generated on basis of the terms occurring in the language model. The more language model terms a sentence contains, the higher is the probability that the output sentence can be created on its basis. This probability is calculated and all the sentences are scored according to this calculation. As a result, the summary consists of sentences with highest ranks.

We implemented a similar approach for the task of toponym-referenced image description generation. Object type corpora are used to derive language models which in turn are used to bias the sentence selection during the summarization process. For instance, if an image summary is to be generated for the object *Westminster Abbey* then its object type ***church*** is automatically identified first with the help of Is-A patterns. Then uni-gram and bi-gram language models are derived from the corpus for the object type ***church***. Finally, while generating the summary for the object *Westminster Abbey* the ***church*** language model is used to rank the sentences of the documents to be summarized. More precisely, sentence generation process described in Jagadeesh et al. (2005) is applied to calculate the probability that the sentence is generated based on the ***church*** language model. Sentences with high probabilities are used to generate a summary to an image. Example 4 demonstrates the description automatically generated for the object *Westminster Abbey* in English.

(24) The Westminster abbey museum is located in the 11th century vaulted undercroft of St Peter beneath the former monks' dormitory in Westminster Abbey. The church is one of the most famous in Britain and is one of London's most visited tourist attractions. Westminster Abbey's long history can be traced back to the community of Benedictine monks established here c. 960 by Dunstan, bishop of London. It was most probably designed for the High Altar of the Abbey, although it has been damaged in past centuries. The Westminster Abbey's a magnifical monument, full of history and meaning. Westminster Abbey was originally a Benedictine monastery, refounded as the Collegiate Church of St. Peter in Westminster (today one of the boroughs constituting Greater London) by Queen Elizabeth I in 1560. It is the traditional place of coronation and burial site for English monarchs. The Westminster Abbey is certain that in about AD 785 there was a small community of monks on the island and that the monastery was enlarged and remodelled by St. Dunstan in about AD 960.

6. Conclusions and future work

In this paper we proposed a method for developing object type text corpora for four languages: English, German, Italian and Latvian with the aim to optimize multi-document summarization for generating image descriptions. We argued that Wikipedia, being a well structured web resource, suits our goals best since it is a rich source of location descriptions and offers features that facilitate multi-lingual cross-referencing. As such it is a valuable source for such languages as Latvian which is considered to have an under-resourced status. We described and evaluated categorization of Wikipedia's content according to object types using Is-A patterns . The evaluation results have shown that Is-A patterns are one simple way to identify the object type corpora which nevertheless renders satisfactory results. Finally, we illustrated how such object type corpora can be used to enhance multi-document summarization on the example of building automatic image captions by summarizing multiple web documents.

In spite of the fact that object type corpora were developed for all the four languages, most experiments

on automatic generating toponym-referenced descriptions of images have been performed for the English. The work on our remaining target languages is ongoing. We plan to carry out evaluation experiments for all the four languages using our object type corpora. It will allow us to investigate whether image summaries that incorporate conceptual models about objects derived from object type corpora are better in quality than those generated without such models and whether the contribution of conceptual models can be observed across languages.

Acknowledgment

The research reported was partly funded by the TRIPOD project (TRI-Partite multimedia Object Description)[5] supported by the European Commission under the contract No. 045335. We would like to thank Lars Borin, Robert Gaizauskas, Emina Kurtic, Andrew Salway, Inguna Skadiņa and Raivis Skadiņš for discussions and comments.

[5] http://tripod.shef.ac.uk/

References

1. *Adafre S. F. and Rijke M.* (2005) Discovering Missing Links in Wikipedia // Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications — LinkKDD, 2005: 90–97.
2. *Adafre S. F. and Rijke M.* (2006) Finding Similar Sentences across Multiple Languages in Wikipedia // Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006: 62–69.
3. *Aker A. and Gaizauskas R.* (2008) Evaluating automatically generated user-focused multi-document summaries for geo-referenced images // Proceedings of the 22nd International Conference on Computational Linguistics — COLING 2008, Manchester.
4. *Barnard K. and Forsyth D.* (2001) Learning the semantics of words and pictures // Proceedings of International Conference on Computer Vision, Vol. 2, Vancouver: IEEE, pp. 408–415.
5. *Barnard K., Duygulu P., Forsyth D., de Freitas N., Blei D., Jordan M.* (2003) Matching words and pictures // The Journal of Machine Learning Research, 3: 1107–1135.
6. *Bennett N., He Q., Powell K., Schatz B.* (1999) Extracting noun phrases for all of MEDLINE // Proceedings of American Medical Informatics Association.
7. *Bernardini S., Baroni M., Evert S.* (2006) A WaCky introduction // Wacky! Working papers on the Web as Corpus, Bologna: GEDIT, 2006: 9–40.
8. *Bouma G., Fahmi I., Mur J., Noord G., Plas L., Tiedermann J.* (2006) Using Syntactic Knowledge for QA // Proceedings of Cross Language Evaluation Forum Workshop, Aarhus, Denmark, 2008.
9. *Cheng P.-J., Pan Y.-C., Lu W.-H., Chein L.-F.* (2004) Creating Multilingual Translation Lexicons with Regional Variations // Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain, Article No. 534, 2004.
10. *Declerck T., Gómez-Pérez A., Vela O., Gantner Z., Manzano-Macho D.* (2006) Multilingual Lexical Semantic Resources for Ontology Translation // Proceedings of the 5th International Conference on Language Resources and Evaluation: LREC 2006, Genoa, Italy, 2006: 1492–1495.
11. *Deschacht K. and Moens M.* (2007) Text Analysis for Automatic Image Annotation // Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. East Stroudsburg: ACL.
12. *Duygulu P., Barnard K., de Freitas J., Forsyth D.* (2002) Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary // Proceedings of the 7th European Conference on Computer Vision (ECCV), 4: 97–112.
13. *Feng Y. and Lapata M.* (2008) Automatic Image Annotation Using Auxiliary Text Information // Proceedings of Association for Computational Linguistics (ACL) 2008, Columbus, Ohio, USA.
14. *Hearst M.* (1992) Automatic acquisition of hyponyms from large text corpora // Proceedings of the 14th conference on Computational linguistics, Vol. 2: 539–545.
15. *Hollink L., Schreiber A., Wielinga B., Worring M.* (2004) Classification of user image descriptions // International Journal of Human-Computer Studies, 61(5): 601–626.
16. *Hughes B.* (2006) A web search service for minority language communities // Proceedings of Open Road 2006 Conference: Challenges and Possibilities, 2006.
17. *Jagadeesh J., Pingali P., Varma V.* (2005) A relevance-based language modeling approach to DUC 2005 // Proceedings of Document Understanding Conferences (along with HLT-EMNLP 2005), Vancouver, Canada.
18. *Jamieson M., Fazly A., Dickinson S., Stevenson S., Wachsmuth S.* (2007) Learning Structured Appearance Models from Captioned Images of Cluttered Scenes // Computer Vision, 2007, ICCV 2007, IEEE 11th International Conference: 1–8.

19. *Jones G., Fantino F., Newman E., Zhang Y.* (2008) Domain-Specific Query Translation for Multilingual Information Access using Machine Translation Augmented With Dictionaries Mined from Wikipedia // Proceedings of the 2nd International Workshop on "Cross Lingual Information Access" Addressing the Information Need of Multilingual Societies, 2008: 34–41.
20. *Kilgarriff A.* (2007) Googleology is bad science // Computational Linguistics, Vol. 33(1): 147–151.
21. *Kilgarriff A. and Grefenstette G.* (2003) Introduction to the special issue on the web as corpus // Computational Linguistics, Vol. 29: 333–347.
22. *Kuo J.-S. and Yang Y.-K.* (2004) Constructing Transliteration Lexicons from Web Corpora // Proceedings of the Association for Computational Linguistics 2004 on Interactive poster and demonstration sessions, Barcelona, Spain, Article No. 3, 2004.
23. *Liu V. and Curran J. R.* (2006) Web Text Corpus for Natural Language Processing // Proceedings of the European Chapter of the Association for Computational Linguistics, 2006: 233–240.
24. *Mann G.* (2002) Fine-grained proper noun ontologies for question answering // Proceedings of International Conference On Computational Linguistics, 2002: 1–7.
25. *Mori Y., Takahashi H., Oka R.* (2000) Automatic word assignment to images based on image division and vector quantization // Proceedings of RIAO 2000: Content-Based Multimedia Information Access.
26. *Pan J., Yang H., Duygulu P., Faloutsos C.* (2004), Automatic image captioning // Multimedia and Expo, 2004. ICME'04, Vol. 3.
27. *Ponzetto S. and Strube M.* (2007) Deriving a Large Scale Taxonomy from Wikipedia // Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence, Vancouver, B.C., Canada, 2007: 1440–1445.
28. *Shatford S.* (1986) Analyzing the Subject of a Picture: A Theoretical Approach // Cataloging and Classification Quarterly, 6(3): 39–61.
29. *Steinberger R., Pouliquen B., Widiger A., Ignat C., Erjavec T., Tufiş D., Varga D.* (2006) The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages // Proceedings of the 5th International Conference on Language Resources and Evaluation: LREC 2006, Genoa, Italy, 2006.
30. *Wentland W., Knopp J., Silberer C., Hartung M.* (2008) Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration // The sixth Language Resources and Evaluation Conference: LREC'08, Marrakech, Morocco, 2008.

Транскрибирование, структурирование и временной анализ речевого корпуса эстонского языка при выборе единиц в системе синтеза (текст-речь)

Transcrbing, structuring and temporal analysis of fluent speech corpus for a unit selection tts system for Estonian

Meelis Mihkla (meelis@eki.ee), **Indrek Kiissel** (indrek@eki.ee),
Tõnis Nurk (tonis@eki.ee), **Liisi Piits** (liisi@eki.ee)

Institute of the Estonian Language, Tallinn, Estonia

В статье рассматриваются проблемы создания системы синтеза, основанной на выборе единиц из корпуса эстонского языка (текст-речь). Авторы предлагают правила транскрибирования и принципы фонологического структурирования, облегчающие выбор языковых единиц. Исследуется также интенсивность коллокации (сочетаемости) в зависимости от темпа речи и разрабатываются соответствующие модели длительности.

1. Background

Around the turn of the millennium (1997–2002) the first generation of Estonian corpus-based TTS synthesizer was developed (Mihkla et al. 1999). This time the basic speech units were diphones, chosen as sources of natural phone transitions. With this undertaking we also became part of the international MBROLA project. Although our first-generation synthesis was also corpus-based, its database provided no more than one diphone for each transition. It has been argued — and proved in practice — that the large number of concatenation points make the synthetic speech sound unnatural, even if the spectral discontinuities have been minimized by carefully smoothing the concatenation points, considering phonetic criteria (Donovan, Woodland 1999).

Two years ago a new project of corpus-based synthesis of Estonian was launched in the framework of the national programme "Language technological support of Estonian". The aim is to develop a high-quality second-generation speech synthesizer, based on unit selection. Our ambition is to create on the bases of moderate speech corpus (up to 60 minutes) high quality Estonian speech synthesizer for male and female voices. The new synthesizer, however, draws its acoustic material from the whole speech corpus. The idea of corpus-based or unit-selection synthesis is that the corpus is searched for maximally long phonetic strings to match the sounds to be synthesized. As compared to diphone or triphone synthesis, corpus-based speech tends to elicit considerably higher ratings of naturalness in auditory tests (Nagy et al., 2003). As the corpus in its entirety provides the acoustic basis for such synthesis, the development of an optimal corpus represents an essential task of corpus-based synthesis.

Our speech corpus (Piits et al. 2007) contains phonetically rich sentences and various phonological structures of Estonian. The corpus includes words which contain all Estonian diphones and many numerals, alongside with frequent Estonian words and expressions. Development of a unit selection based TTS system for Estonian take place in two directions: on the one hand the system is developed in the Festival environment with Cluster unit selection. On the other hand we would like to test just how far we can go by using a high-quality corpus and good algorithms of unit selection without any synthesis engine, just applying some very simple methods of signal processing. Both applications demand high quality speech corpus, thus the development of an optimal corpus represents an essential task of corpus-based synthesis. A system with a good selection module and a high-quality speech corpus may yield output speech of extremely high quality, even if the signal processing module is rather simple (Bozkurt et al., 2002).

In order to convert an Estonian written text into synthesized speech we have to solve the following

tasks: to convert an orthographic text into a phonetic-phonological one and to compile rules or models for the control of segment durations and F0 contours. The most difficult problem is: how to find the third quantity degree Q3 and palatalization automatically in the written text. We also investigate how collocation strength might correlate with speech rate and what role it might play in duration models. In order to optimize the unit selection algorithm and to guarantee the necessary quality of the synthetic speech the whole speech database as well as the utterance to be synthesized is represented as a phonological tree.

2. Transcription rules for fluent speech corpus

When transforming an orthographic Estonian text into a pronounced text one should take into account that not every phonological opposition is spelt out in Estonian. This applies, for example, palatalized vs. unpalatalized consonants and, in a general case, to the 2nd and 3rd quantity degrees. In Estonian long stressed syllables can be pronounced with (third quantity degree: Q3) or without (second quantity degree: Q2) prosodical quantity. In principle, both palatalization and word quantity are marked in the lexicon, after morphological analysis of the sentence. Still, there are cases where the phonetic quantity and palatalization defined on the basis of the lexicon need additional adaptation to meet the rules of fluent speech.

In Estonian, which is considered a stress-timing language, the foot (1–3 syllables) carries such prosodic phenomena as stress and quantity. Two- or three-syllable words may be classified into first-, second- and third-quantity words, whereas the mono-syllabics have, theoretically, all been marked out for the third quantity. Actually fluent speech is characterized by a considerable number of monosyllabic words (pronouns, conjunctions, adverbs) capable of either adhering to other words in longer feet or occurring in an unstressed position as clitics in the speech flow (Hint 1998:145–146; Lehiste 1997:11). Therefore the more frequent pronouns and conjunctions are not subjected to the general rules of transcription, but treated as special cases, e.g the conjunctions *et* 'so as to' [*ett*, not *et:t*], *kui* 'if; than' [*kui*, not *kui:*]; the pronoun *neid* 'he/she Part. Pl.' [*neit*, not *nei:t*]. At the same time, the third quantity is retained in content words of a similar structure, e.g. *vett* 'water Part. Sg.' [*vet:t*].

The Estonian language has four palatalized speech sounds: the palatal lateral approximant [l'], the palatal nasal [n'], the alveolar ejective fricative [s'], and the voiceless palatal plosive [t']. Our orthographic spelling makes no difference between the palatalized and unpalatalized sounds, even though there are several cases where palatalization does have a distinctive function as, for example, in *palk* 'salary' [*palk:k*] and *palk* 'log' [*pal'k:k*]. In addition to the lexicon-based palatalization, palatalization is a feature added automatically to such *l, n, s, t* that, preceding *i* or *j,* immediatedly follow either the vowel of a main stress syllable or a palatalised consonant, e.g. *hiilima* 'to sneak' [*hii:lima -> hii:l'ima*], *kasti* 'box Gen. Sg.' [*kas'ti -> kas't'i*], but *kunsti* 'art Gen. Sg.' [*kun'sti*].

Foreign letters are transcribed according to the Estonian tradition: the fricatives *z* and *ž* (voiced retroflex fricative), for example, sound voiced in many languages, but not in Estonian, where their pronunciation does not differ from that of *s* and *š* (voiceless retroflex fricative).

The long *üü* (close front round vowel) is diphthongized if followed by a vowel or *j,* like, e.g., in *müüjad* 'salesperson Nom. Pl.' [*myi:jat*], *hüüe* 'shout' [*hyie*]. If a vowel follows a long *ii* or an *i*-final diphthong, it is pronounced as if preceded by the glide *j*, e.g. *saiu* 'white bread Part. Pl.' [*sai:ju*]. A similar rule works for a long *uu* and an *u*-final diphthong, where the *u* is pronounced as if followed by the voiced labiovelar approximant *w*, e.g. *suue* '(river) mouth' [*suuwe*].

If three stops happen to meet on a word boundary, the single plosive phonemes are usually reduced to such an extent that, as a rule, a dissimilatory loss can be diagnosed, e.g. the sequence *kõlblik kokk* 'a fit cook' is transcribed as *kõl:plik kok:k*, not as *kõl:plikk kok:k*. Dissimilatory loss may also be found in vowels, e.g. the phrase *ma ei tea* 'I don't know' is pronounced as [*maitea*].

The letters *k, p, t, t', f* and *š* occurring between voiced sounds on syllable boundary or at the end of a word are transcribed doubly, whereas in the immediate neighbourhood of a voiceless sound as well as at the beginning of a word they remain single, e.g. *kokpit* 'cockpit' [*kokpitt*], *Aafrika* 'Africa' [*aaffrikka*], *part* 'duck' [*part:t*], *kašelott* 'cachalot' [*kaššelot':t'*].

3. Structuring the speech corpus

In order to optimize the unit selection algorithm and to guarantee the necessary quality of the synthetic speech the whole speech database as well as the utterance to be synthesized is represented as a phonological tree. Figure 1 represents a fragment of the database tree. Our phonological tree has the following levels: phoneme, syllable structure, syllable, foot, word, phrase, and sentence (cf. Breen et al. 1998, Taylor et al. 1999). Search for appropriate speech units involves all those levels, beginning from the higher ones, e. g. preferring longer units.

The free word order and close intertwinement of the Estonian syntactic phrases frustrates, as a rule, the attempts to build a binary tree between the word and sentence levels. For the present project a phrase is defined as a clause that is separated by an intrasentence punctuation mark or conjunction and that includes a predicate.

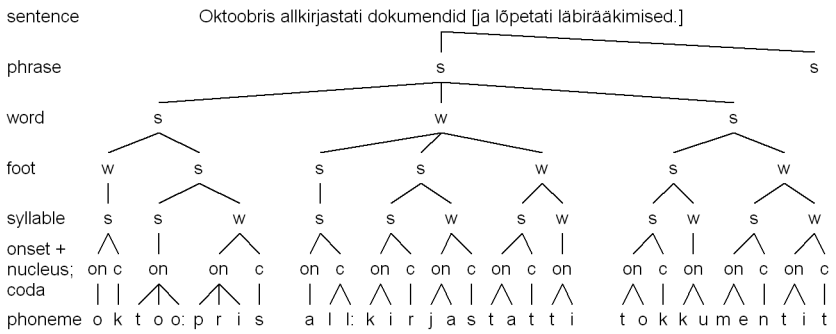


Figure 1. Fragment of the phonological tree

Black and Taylor's phonological structure has been criticized for fixed representation of word boundaries (Möbius 2000). True, there are some Estonian (compound) words (e.g. *raudtee* 'railway', *õunapuu* 'apple-tree') with co-articulation occurring on word boundary, yet the basic unit in Estonian speech is the word and thus an emphasis on the marking of word boundaries presents no problem for speech synthesis. But there is an extremely rich morphology to consider. Thus we first attempted a joint use of the phonological tree and a morphological one consisting of word forms divided into stem and grammatical morphemes. Unfortunately the representation of phonological and morphological information in one and the same tree turned out too complicated a task, because morpheme and syllable boundaries do not coincide. So we had to keep to the phonological tree after all. Still, our consideration for morphological information had been not quite in vain, providing great help in finding phonologically suitable units.

In some ways the branching of the phonological tree on sub-syllable levels is also morphologically bound. Syllable division starts from separating the coda. This enables the use of the onset and nucleus in word form synthesis even in case the coda is wrong. Thus, after separating the coda we can use the word form *ka-la-l* 'fish Adess. Sg.' to form, e.g. *ka-la-s* 'fish Iness. Sg.' *ka-la-st* 'fish Elat. Sg.' or *ka-la-lt* 'fish Ablat. Sg.'

4. Effect of collocational strength on speech rate

While recording an Estonian corpus for corpus-based synthesis some fluctuations of the speech rate were observed, even though the text was read out by a professional radio announcer. The slowings down could be due to difficult clusters (the corpus was required to contain all diphones possible in Estonian, however rare (see Piits et al. 2007), which could, in turn, occur in rare words. A quickening rate, however, could have to do with frequent words as well as collocational phrases. It has been argued before that the high frequency of a word and the predictability of its context may have a reducing effect on the pronunciation of the word (Pluymaekers et al. 2005, Bell et al. 2003). In some cases the effects of word frequency and contextual predictability on word duration have been studied in combination (M. L. Gregory et al.1999).

In Estonian, the word has a very important role both in grammar and phonetics, while the morphology is extremely rich. The aim of the present study is to find out if, apart from word frequency, Estonian word length could in any way depend on the collocational strength between the words.

We set up the following hypothesis: collocational strength between words in Estonian has an effect on word duration, i.e. words occurring side by side more often tend to be pronounced more rapidly. Our scrutiny is focused on the verb *olema* 'be' as the most frequent word in Estonian.

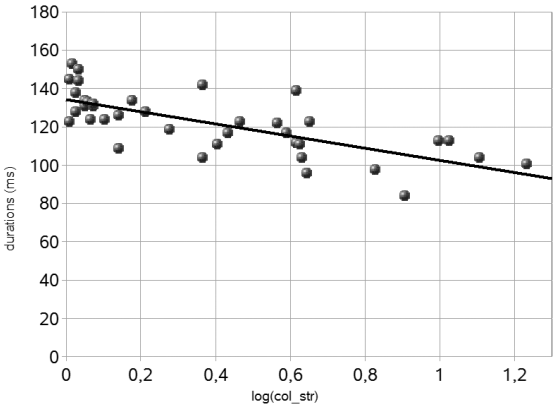


Figure 2. Relation between duration of stem sounds and collocational strength.

We investigated how collocation strength might correlate with the durations of *olema* forms and what role it might play in predictive models. Figure 2 shows the relation between duration of stem sounds of verb *olema* 'be' and collocational strength (in logarithmic scale). The hypothesis was tested by means of different statistical methods (linear regression, CART trees), enabling to disclose small, hidden, but possibly significant effects between input and output (Sagisaka 2003).

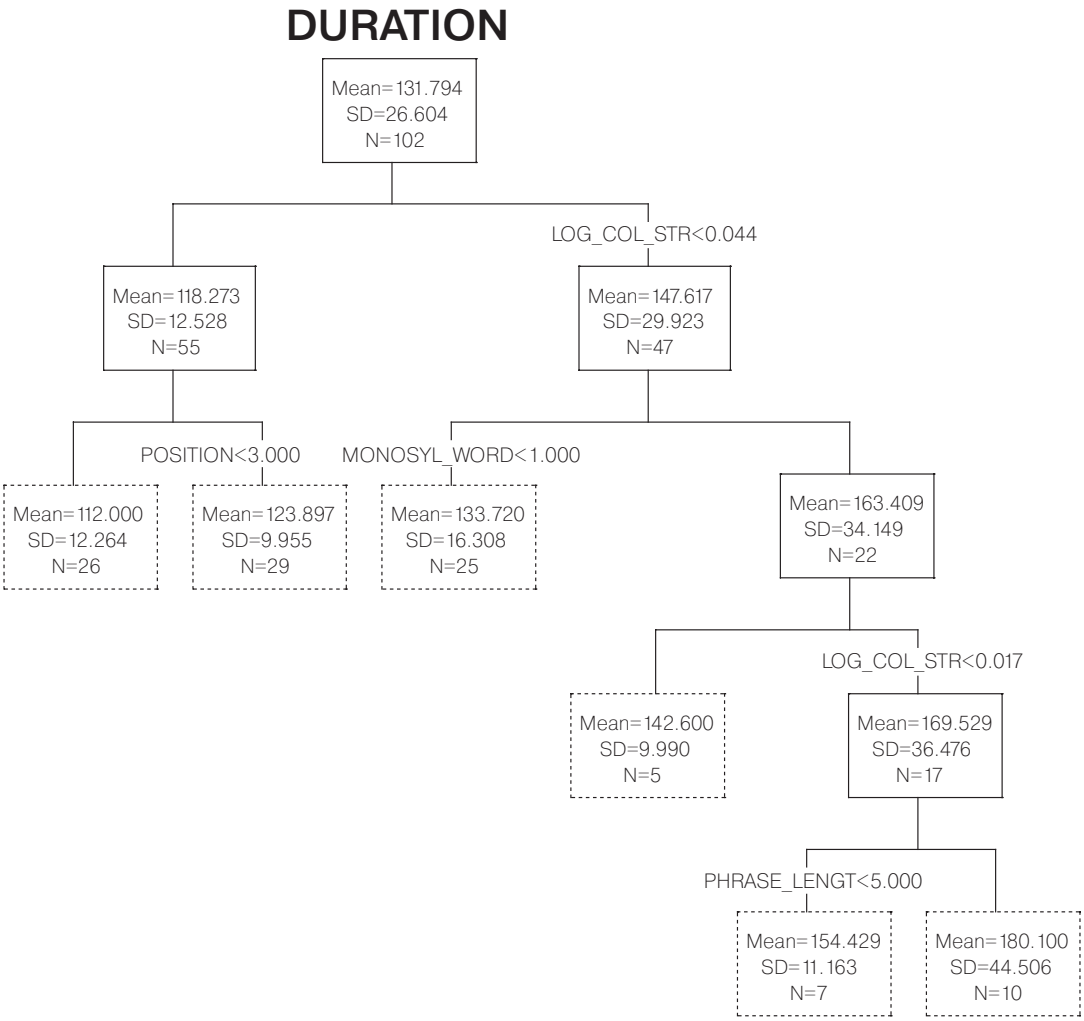


Figure 3. CART model of durations of the verb ***olema*** 'be'.

A simple durational model was compiled to predict the duration of the verb *olema* 'be' from collocation strength, length of phrase, position of the verb in the phrase, and a binary characteristic indicating whether a concrete verb form had just one syllable or more (Fig. 3). According to the resulting models there were two features — the binary one and collocational strength — that were significant in all models. Consequently, in the material studied collocational strength does have an effect on the durations of the verb *olema* 'be'.

5. Conclusion

The aim of the speech corpus described was to develop an acoustic basis for a relatively naturally sounding synthetic speech. To reduce the number of concatenation points in the synthetic utterance it was necessary to create a speech corpus enabling searching for units larger than diphones. The phonological structure describes different levels from where the units were found. The rules for transforming an orthographic Estonian text into a pronounced text should certainly provide for the perception of phonologically essential distinctions. Yet apart from that the transcription rules should not overlook some additional features, which, without being distinctive, still play an important role in making fluent speech sound natural. The study also demonstrated that the strength of collocation between words shortens the duration of the Estonian verb *olema* 'be' and that contextual predictability is a significant feature to be considered in developing models of word duration. Whether this indicates a stable relation between input and output or an occasional hidden one is a question pending further research involving measurement of collocation strength and durations of other words on more copious speech material.

Acknowledgements

This work is supported by National Programme for Estonian Language Technology, grant ETF7998 and project SF0050023s09.

References

1. *Bell A., Jurafsky D., Fosler-Lussier E., Girand C., Gregory M. and Gildea D.* 2003. Effects of disfluencies, predictability and utterance position on word form variation in English conversation. // Journal of the Acoustical Society of America, 113(2), pp. 1001–1024.
2. *Bozkurt B., Dutoit, T., Prudon, R. C., D'Alessandro, C., Pagel, V.* 2004. Reducing discontinuities at synthesis time for corpus-based speech synthesis. // In: Narayanan, S.; Alwan A. (eds). Text To Speech Synthesis: New Paradigms and Advances, pp. 1–17.
3. *Breen, A. P., Jackson, P.* 1998. Non-Uniform unit selection and the similarity metric within BT's Laureate TTS system. // In: Proc. Third ESCA Workshop on Speech Synthesis, pp. 373–376.
4. *Donovan, R. E., Woodland P. C.* 1999. A hidden Markov-model-based trainable speech synthesizer. // Computer Speech and Language 13, pp. 223–241.
5. *Gregory M. L., Raymond W. D., Bell A., Fosler-Lussier E. and Jurafsky D.* 1999. The effects of collocational strength and contextual predictability in lexical production. // CLS-99, pp. 151–166. Chicago: University of Chicago.
6. *Hint M.* 1998. Häälikutest sõnadeni. // Eesti Keele Sihtasutus, Tallinn, pp. 145–146.
7. *Lehiste, I.* 1997. Search in phonetic correlates in Estonian prosody. // In: Lehiste, I., Ross, J. (eds.) Estonian Prosody: Papers from Symposium, Institute of the Estonian Language, Tallinn, pp. 11–35.
8. *Mihkla, M., Eek, A., Meister, E.* 1999. Diphone synthesis of Estonian. // In: Dialogue'99 : Computational Linguistics and its Applications: International Workshop: Proceedings. Vol. 2. Applications. (Toim.) Narin'yani,A.S.. Tarusa: 1999, pp. 351–353.
9. *Möbius, B.* 2000. Corpus-based speech synthesis: methods and challenges. // Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart) 6(4), pp. 87–116.
10. *Nagy, A., Pesti, P., Németh, G., Bőhm, T.* 2005. Design Issues of a Corpus-Based Speech Synthesizer. // Hungarian Journal on Communications 6, pp. 18–24.
11. *Piits L., Mihkla M., Nurk T. and Kiissel I.* 2007. Designing a speech corpus for Estonian unit selection synthesis. // Nodalida 2007 Proceedings: The 16th Nordic Conference of Computational Linguistics, pp. 367–371.
12. *Pluymaekers M., Ernestus M. and Baayen H. R.* 2005. Articulatory planning is continuous and sensitive to informational redundancy. // Phonetica, 62, pp. 146–159.
13. *Sagisaka Y.* 2003. Modelling and perception of temporal characteristics in speech. // Proceedings of 15th International Congress of Phonetic Sciences, Barcelona, pp. 1–6.
14. *Taylor, P., Black, A. W.* 1999. Speech synthesis by phonological structure matching. // Proc. Eurospeech'99 Budapest, pp. 623–626.

Динамический характер значения прилагательных

The dynamics of adjective meaning[1]

Partee Barbara H. (partee@linguist.umass.edu)
University of Massachusetts, Amherst, MA, USA; Moscow State University

Значение находится в динамическом взаимодействии с контекстом. Наша задача — описать зависимость значения от контекста, не отказываясь от принципа композициональности. Мы проиллюстрируем эту проблему на примере семантики некоторых сортов прилагательных. Будет показано, как взаимодействуют друг с другом композициональная семантика, лексическая семантика и контекст.

Introduction

A central concern for the study of meaning is how the meanings of expressions are composed from the meanings of their constituent parts. What are "parts"? The Principle of Compositionality requires a notion of part-whole structure that is based on syntactic structure.

> **Principle of Compositionality:** The meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined.

This is a good if informal statement of the basic principle. There have been many challenges to it of different sorts, but it makes a good working principle: apparent counterexamples are invitations to work hard to uncover new descriptive accounts or to make revisions somewhere in the theoretical framework.

The focus of the paper will be some aspects of the dynamic interaction of meaning and context. One important challenge faced by compositional approaches is how to account for context-dependent meaning shifts without abandoning compositionality. The semantics of different kinds of adjectives presents challenges of this sort. The interplay of context-dependence and intensionality is crucial for showing why *skillful* is intensional but *large* is not, even though we may consider a large house not to be a large building. I will also take up the puzzles of "privative" adjectives like *fake* and *counterfeit* and "redundant" adjectives like *real*. I will try to show how attention to the semantics of syntactic structure (compositional semantics) sheds light on word meaning, and how compositional semantics, lexical semantics, and the context of the utterance all interact.

1. Introduction to adjective semantics

Montague [3] presented a semantic treatment of adjectives which he credited to unpublished work by Hans Kamp and by Terence Parsons; that work, and similar work of Romaine Clark, was subsequently published [4–6]. The central claim in that work was that adjective meanings should be analyzed as functions from properties to properties. Among adjective meanings, some might satisfy further constraints such as intersectivity or subsectivity, but no such constraint can be imposed on the class as a whole, the argument goes, because of the existence of adjectives like *false, ostensible, alleged*.

Since Montague insisted on having a uniform semantic type for each syntactic category, he gave all adjectives the type of functions from properties to properties. More restricted subclasses of adjectives, such as the subsective (*skillful, good*) and intersective (*purple, carnivorous*) adjectives, could then be indicated by the use of meaning postulates. Most current theories allow type multiplicity and type-shifting, and assign to the intersective adjectives the simpler type of one-place predicates.

Kamp and Partee [7] review the standard "hierarchy" of classes of adjectives as a preliminary to a discussion of the possible appropriateness of prototype theory for

[1] I am grateful to Anita Nowak for showing me the Polish Split-NP facts, and to Meredith Landman for initial discussion. For valuable comments I thank Lisa Matthewson and her UMass Fall 2000 Pro-Seminar, several classes of students at RGGU and MGU in Moscow, participants of a colloquium in honor of Terry Parsons at Notre Dame in 2003, and audiences in Leipzig, Copenhagen, Amsterdam, Arizona, Prague, and Göteborg. I thank Maria Gouskova, Bozena Cetnarowska, and Bozena Rozwadowska for helpful discussions of the Polish data, and an anonymous reviewer for suggestions for improvement. This talk is based on [1] and [2]. This work was supported in part by NSF Grants BCS-9905748 and BCS-0418311 to Partee and Vladimir Borschev.

some adjective-noun combinations. The hierarchy ranges from intersective adjectives like *carnivorous* to privative adjectives like *counterfeit, fake,* and *fictitious.* The same article makes proposals for coercion of adjective meanings in context, driven by certain general constraints, which help to explain a number of kinds of meaning shifts. Some problem cases remained, especially the case of *stone lion*, where it seems that the noun rather than the adjective shifts its meaning, contrary to our predictions.

But now I want to argue that in fact adjective meanings are more constrained than was appreciated in the work of Montague, Kamp, Parsons and Clark or in the work of Kamp and Partee. I will argue that some facts about the possibility of "NP-splitting" in Polish and Russian cast serious doubt on the standard hierarchy, and that the data become much more orderly if privative adjectives like *counterfeit, fake,* and *fictitious* are reanalyzed as subsective adjectives. Further evidence for that move comes from long-standing puzzles about what to say about sentences like *Is that gun real or fake?* The revised account requires the possibility of coerced expansion of the denotation of the noun to which such an adjective (as well as adjectives like *real*) is applied. Such coercion can be motivated by treating the constraints on possible adjective meanings as presuppositions that must be satisfied by any use of an adjective; the corresponding coercion is then a form of presupposition accommodation.

The rest of the paper is structured as follows. Section 2 briefly reviews the adjective classification familiar since the work of the 1970's as summarized in [7] and [8]. The Polish NP-splitting data [9] and the problem they pose for the familiar hierarchy are presented in Section 3. In Section 4 I review some of the constraints on possible adjective meanings proposed in [7] and propose further constraints that exclude privative adjectives and account for the coercion of the noun meaning in cases that would otherwise come out as privative.

A qualification must be made at the beginning. There are many Adj-Noun combinations that are idioms, compounds, or otherwise lexicalized, non-compositional units: they must be learned as wholes, although their parts may give a clue to their meaning. No proposals in this paper are claimed to apply to phrases like "blackbird", "black death" (= the plague), etc.

2. Adjective classification

2.1. Meaning postulates for classes of adjectives

An adjective like carnivorous is **intersective**, in that the informally stated meaning postulate (1) holds for any N.

(1) $||\textit{carnivorous}\ \text{N}|| = ||\textit{carnivorous}|| \cap ||\text{N}||$

But *skillful* is not, as shown by the invalid inference pattern in (2), familiar from the work of Kamp, Parsons, Clark, and Montague.

(2) *Premise: Francis is a skillful surgeon.*
Premise: Francis is a violinist.

Conclusion: Francis is a skillful violinist. INVALID

Skillful is not intersective, but it is **subsective:** meaning postulate (3) holds for any N.

(3) *Subsectivity*: $||\textit{skillful}\ \text{N}|| \subseteq ||\text{N}||$

The adjectives *former, alleged, counterfeit* are neither intersective nor subsective.

(4) (a) $||\textit{former senator}|| \neq ||\textit{former}|| \cap ||\textit{senator}||$
(b) $||\textit{former senator}|| \not\subset ||\textit{senator}||$

Nonsubsective adjectives may either be "plain" nonsubsective (no entailments at all, no meaning postulate needed), or **privative**, entailing the negation of the noun property. The meaning postulate for privative adjectives is stated informally in (5).

(5) $||\textit{counterfeit}\ \text{N}|| \cap ||\text{N}|| = \varnothing$

Additional examples of each type are given below.

(6) (i) intersective: *sick, carnivorous, blond, rectangular, French*.
(ii) non-intersective but subsective: *typical, recent, good, perfect, legendary*.
(iiia) non-subsective and privative: *would-be, past, spurious, imaginary, fictitious, fabricated* (in one sense), *mythical* (maybe debatable); there are prefixes with this property too, like *ex-, pseudo-, non-*.
(iiib) plain non-subsective: *potential, alleged, arguable, likely, predicted, putative, questionable, disputed*.

The conclusion drawn by Parsons, Kamp, Clark and Montague was that the simplest general rule for interpretation of the combination of an adjective with a noun (or common noun phrase: CNP) is the following: Adjectives are functions that map the (intensional) semantic value of the CNP they combine with onto the semantic value of the ADJ + CNP combination. That is, "The denotation of an adjective phrase is always a function from properties to properties. (This was one of the proposals advanced by Kamp and Parsons.)" [3, p.211 in Montague 1974]

Meaning postulates specify various restrictions on these functions, characterizing various subclasses of adjectives. "Semantic features" may be seen as labels for meaning postulates which give them determinate content. Thus a lexical entry for an intersective adjective like *green* might contain the "feature" +Intersec-

TIVE, which can be taken as labelling a semantic property of the adjective, spelled out by a meaning postulate.

Alternatively, and more commonly, intersective adjectives (and only those) can be interpreted in type <e,t>. This alternative treatment automatically guarantees their intersectivity and eliminates the need for a meaning postulate. Type-shifting rules of the sort described in Partee (1995) will give them homonyms of a functional type when needed.

The "plain" nonsubsective adjectives (*alleged, possible*) have no meaning postulate; this class is "noncommittal": an *alleged murderer* may or may not be a *murderer*.

The **subsective** adjectives (*skillful, good*) have a meaning postulate specifying that the ADJ + CNP combination denotes a subset of the CNP denotation: *a skillful surgeon* is a *surgeon*. The **privative** adjectives (*fake, counterfeit*) have a "negative" meaning postulate; a *fake gun* is not a *gun*.

On this familiar classification, adjectives are seen as forming a hierarchy from intersective to subsective to nonsubsective, with the privative adjectives an extreme case of the nonsubsective adjectives.

Among many other debated points, one which has always been troubling, and to which we will return in Section 4, is the question of whether an adjective or adjectivally used noun like *fake* or *toy* is really privative. One problem is the tension between the possible truth of (7a) and the undeniable well-formedness and interpretability of (7b).

(7) (a) *A fake gun is not a gun.*
(b) *Is that gun real or fake?*

2.2. Is tall intersective or subsective?

There are many questions and disputes when it comes to assigning particular adjectives to particular classes. Kamp [5] added an important insight in arguing that adjectives like *tall*, which seem to be non-intersective, are actually intersective but context-dependent. Kamp's analysis found linguistic support in Siegel's analysis of long-form and short-form adjectives in Russian [10, 11].

In Section 2.1 we said that the inference pattern (2) was a test of whether an adjective was intersective. By this test, vague adjectives like *tall* would appear to be non-intersective:

(2') *Premise: Tom is a tall 14-year-old.*
Premise: Tom is a basketball player.

Conclusion: Tom is a tall basketball player. INVALID??

Does this mean that *tall* is not intersective? No; perhaps it is intersective but vague and context-dependent. How can we tell the difference?

First argument. Keep the ADJ-N sequence constant but change other aspects of the context. That can help to show whether it is the intension of the noun that is crucial.

(2") (a) *My two-year-old son built a really tall snowman yesterday*.
(b) *The linguistics students built a really tall snowman last weekend.*

Further evidence that there is a difference between truly nonintersective subsective adjectives like *skillful* and intersective but vague and context- dependent adjectives like *tall* was noted by Siegel (1976b): the former occur with *as*-phrases, as in *skillful as a surgeon*, whereas the latter take *for*- phrases to indicate comparison class: *tall for an East coast mountain*. (An adjective can be nonintersective and **also** vague, and then one can use both an *as*-phrase and a *for*-phrase: *very good as a diagnostician for someone with so little experience*.)

There has been much further work on the semantics of adjectives in the intervening years, and the context-dependence of interpretation of adjectives is central in the work of Klein [12] and more recently of Kennedy [13].

3. Privative adjectives and Polish NP-split phenomena

Nowak [9] studied the phenomenon of "split PPs" and "split NPs" in Polish, a construction that is conditioned primarily by topic-focus articulation [14, 15]. An NP consisting of Adj and N in Polish may be "split", with either Adj or N sentence-initial and the other sentence-final. Sequences of Adj's can be sentence-initial; only a single element can be sentence-final. Examples of NP-splits are given in (8 — 9) below, with the relevant constituents underlined.

Sentences (8b) and (9b) are 'split' versions of sentences (8a) and (9a), which represent the unmarked word order. All examples are from Nowak [9][2].

(8) (a) *Kelnerki rozmawiały* *o przystojnym chłopcu*.
Waitresses talked about handsome-LOC boy-LOC
'The waitresses talked about a handsome boy.'
(b) *O przystojnym* *kelnerki rozmawiały* *chłopcu*.
about handsome-LOC waitresses talked boy-LOC
'The waitresses talked about a handsome BOY'

(9) (a) *Włamano* *się do nowego sklepu*.
broke-in (one) reflex. to new-GEN store-GEN
'Someone broke into the new store.'
(b) *Do sklepu* *włamano* *się* *nowego*.
to store-GEN broke-in (one) reflex. new-GEN
'Someone broke into the NEW store.'

What is of particular relevance for this paper is that some adjectives can participate in the splitting construction and some cannot.

---

[2] Bożena Cetnarowska (p.c.) has informed me that the data are less black-and-white than they appear here; I will not discuss the complexities here, but only note that the generalizations made in the text still seem to hold.

(10) *Do rozległej weszliśmy doliny.*
to large-GEN (we)entered valley-GEN
'We entered a large VALLEY.'

(11) **Z potencjalnym widzieli się kandydatem.*
with potential-INSTR (they)saw reflex. candidate-INSTR
'They met with a potential CANDIDATE'

Those that CAN split include:

(12) (a) *rozległy* 'large, vast'
(b) *biedny* 'poor' in the sense of 'not rich', not in the sense of 'pitiful'
(c) *zdrowy* 'healthy', *amerykański* 'American', *gadatliwy* 'talkative' (intersective)
(d) *dobry* 'good', *sławny* 'famous', *wprawny* 'skillful' (subsective)
(e) *fikcyjny* 'fictitious', *wymyślony* 'imaginary', *fałszywy* 'fraudulent' (privative [!])

Those that CANNOT split include:

(13) (a) *biedny* 'poor' in the sense of 'pitiful'
(b) *potencjalny* 'potential', *rzekomy* 'alleged', *sporny* 'disputed', *oczekiwany* 'expected, due, anticipated' (non-subsective, non-privative ('modal')

Another important fact is that the ones that cannot split also cannot occur predicatively.

On the traditional classification outlined in Section 2, the adjectives which can participate in the NP-split phenomenon are not a "natural class". It is unexpected for the intersective, subsective, and privative adjectives to pattern together, while the non-subsective adjectives that are "noncommittal" (and generally "modal"), cannot participate in the NP-split.

4. Principles of interpretation and the "no privative adjectives" hypothesis

The hypothesis I propose is that Nowak's data provide a clue that the adjectives *fake* and *imaginary* aren't actually privative, but subsective, and that no adjectives are actually privative. In interpreting a question like (7b) above or sentences like (14a) and (14b) below, I propose that we *expand* the denotation of *fur* to include both fake and real fur.

(14) (a) *I don't care whether that fur is fake fur or real fur.*
(b) *I don't care whether that fur is fake or real.*

In fact, even in (7a), it is reasonable to suppose that the first occurrence of *gun*, modified by *fake*, is similarly coerced, whereas the second, unmodified, occurrence is not. Normally, in the absence of a modifier like *fake* or *real*, all guns are understood to be real guns, as is evident when one asks how many guns the law permits each person to own, for instance. Without the coerced expansion of the denotation of the noun, not only would *fake* be privative, but the adjective *real* would always be redundant[3].

Kamp and Partee [7], in discussing the "recalibration" of adjective interpretations in context, introduced a number of principles, including the following "Non-Vacuity Principle".

(15) **Non-vacuity principle (NVP):**
In any given context, try to interpret any predicate so that both its positive and negative extension are non-empty. [7, p.161]

The Non-Vacuity Principle applies not only to simple predicates but to predicates formed, for instance, by combination of an adjective and a noun: these should be interpreted in such a way that the ADJ + N combination is a non-vacuous predicate.

Kamp and Partee [7] also argued, using example (16), that in ADJ + N constructions, one first interprets the noun in the larger context (ignoring the adjective), and then "recalibrates" the adjective as necessary. This is expressed as the "Head Primacy Principle" in (17).

(16) (a) *giant midget*
(*a midget, but an exceptionally large one*)
(b) *midget giant*
(*a giant, but an exceptionally small one*)

(17) **The Head primacy principle (HPP):** In a modifier-head structure, the head is interpreted relative to the context of the whole constituent, and the modifier is interpreted relative to the local context created from the former context by the interpretation of the head.[4] [7, p.161]

In many cases, the Non-Vacuity Principle and the Head Primacy Principle cooperate to account for the observed results, including not only the examples in (16), but also the fact that the truth of (18b) below is compatible with a non-redundant use of the modifier in (18a).

(18) (a) *This is a sharp knife.*
(b) *Knives are sharp.* [7, p.162]

If the Head Primacy Principle is absolute, the proposed shift in the interpretation of the head noun under coercion by a privative adjective like *fake* or a "tautologous" adjective like *real* would be impossible. But there are other examples that suggest that the Head Primacy Principle is non-absolute. In particular, there is a large

[3] This property of *real* is noticed in passing by Lakoff in [16].

[4] "In the simplest cases, the effect of the interpretation of a head noun on a given context will be to restrict the local domain to the positive extension of the head in the given context." [7].

and productive class of "constitutive material" modifiers that occur in examples like *stone lion*, *wooden horse, velveteen rabbit, rubber duck*. It is evidently so easy to shift nouns from their literal meaning to a meaning "representation/model of …" that we hardly notice the shift.

The perspective of Optimality Theory suggests that we can account for this situation by saying that the Non-Vacuity Principle outranks the Head Primacy Principle. We normally try to obey both. But if there is no reasonable way to obey the Non-Vacuity Principle without shifting the noun outside its normal bounds (as in the case of *fake* and *real*), then it may be shifted in such a way as to make the compound predicate obey the Non-Vacuity Principle.

So I suggest that **no adjectives are privative** [2]. "Normal" adjectives are always subsective, and there should be some ways to identify "modal" adjectives as a special subclass, such that only they are not necessarily subsective.

If the "no privatives" hypothesis can be maintained, then the classification of adjectives is much more neatly constrained. Adjectives are still functions from properties to properties in the general case, but in harmony with the traditional notion of *modifiers*, they are normally constrained to be subsective. We still need to allow for the 'modal' adjectives, which are not so constrained; the Polish data provide fuel for a proposal to consider them syntactically as well as semantically distinct. Of course more work also needs to be done on the detailed lexical semantics of each of the putatively privative adjectives, since they are far from identical; but that is beyond the scope of this paper.

5. Conclusions

The adjective puzzles that I have been discussing were designed to illustrate several issues. One is the need to study lexical semantics and principles of semantic composition together; decisions about either may have major repercussions for the other. More importantly for this context, I have tried to show that while contextually influenced meaning shifts pose challenges for compositionality, we can see that compositionality plays an essential role in constraining the kinds of meaning shifts that take place. We hold the principle of compositionality constant in working out (unconsciously) what shifts our interlocutors may be signaling. In the extreme case we (like children) depend on compositionality to figure out the meanings of novel words: if we can use contextual clues to guess what a whole sentence or phrase means, we can then "solve" for the meaning of the unknown word. Compositionality thus appears to be one of the most cognitively basic principles in the realm of semantics.

References

1. *Partee B. H.* Compositionality and coercion in semantics: The dynamics of adjective meaning // Cognitive Foundations of Interpretation. Amsterdam: Royal Netherlands Academy of Arts and Sciences, 2007. P. 145–161.
2. *Partee B. H.* Privative adjectives: subsective plus coercion // Presuppositions and Discourse. Amsterdam: Elsevier, In press.
3. *Montague R.* English as a formal language // Linguaggi nella Società e nella Tecnica. Milan: Edizioni di Comunità, 1970. P. 189–224.
4. *Clark R. L.* Concerning the logic of predicate modifiers // Noûs, 1970. 4. P. 311–335.
5. *Kamp H.* Two theories about adjectives // Formal Semantics of Natural Language. Cambridge: Cambridge University Press, 1975. P. 123–155.
6. *Parsons T. Some problems concerning the logic of grammatical modifiers // Synthèse, 1970. 21. P. 320–340.*
7. *Kamp H., Partee B.* Prototype theory and compositionality // Cognition, 1995. 57. P. 129–191.
8. *Partee B.* Lexical semantics and compositionality // An Invitation to Cognitive Science (Second Edition). Vol. 1: Language. Cambridge: MIT Press, 1995,. P. 311–360.
9. *Nowak A.* On split PPs in Polish // Amherst: unpublished manuscript, 2000.
10. *Siegel M.* Capturing the Russian adjective // Montague Grammar. New York: Academic Press, 1976. P. 293–309.
11. *Siegel M. E. A.* Capturing the Adjective. Amherst: University of Massachusetts, 1976.
12. *Klein E.* A semantics for positive and comparative adjectives // Linguistics and Philosophy, 1980. 4. P. 1–45.
13. *Kennedy C.* Projecting the Adjective. // Santa Cruz: University of California Santa Cruz, 1997.
14. *Melhorn G.* Information structure of discontinuous constituents in Russian // Current Issues in Formal Slavic Linguistics (FDSL-3). Frankfurt/Main: Lang, 2001. P. 344–352.
15. *Siewierska A.* Phrasal discontinuity in Polish // Australian Journal of Linguistics, 1984. 4. P. 57–71.
16. *Lakoff G.* Women, Fire, and Dangerous Things // Chicago and London: The University of Chicago Press, 1987.

Онтологическая семантика и абдукция: обработка эллипсиса

Ontological semantics and abduction: parsing ellipsis

Petrenko M. (mpetrenk@gmail.com)

Московский Гуманитарный Институт имени Е. Р. Дашковой, Москва, Россия

В работе рассматриваются возможости абдуктивного (инференционного) анализа естественных текстов с эллиптическими сегментами в рамках Отнологической Семантики на основе инференционных правил установления зависимости между семантическими ролями, а также правил зависимости классов событий и значений скалярных атрибутов.

1. Paper goals

The paper explores a promising yet currently understudied area of application of Ontological Semantics — ellipsis processing. After a brief outline of the Ontological Semantics framework and the mechanism of abduction, i.e. inference-based form of reasoning, it will be demonstrated how an Ontosem-informed NLP application handles elliptic input abductively, i.e. in a two-step fashion similarly to an abducing human agent. Two directions for developing abductive NLP module are explored. Pertinent examples are provided.

2. Direct Meaning Access: theory, structure and applications

In the current range of largely non-semantic, method-driven and domain-restricted computational NLP systems, Ontological Semantics, or Direct Meaning Access (its current incarnation), offers a semantics-informed, formalism-independent, problem-driven and cross-domain toolbox for describing and modeling human language competence in its complexity and dynamicity. A rapidly growing list of publications provides a detailed description of the methodology [14, 28], structure [9] and application domains [24, 25, 26, 28]. Below follows a brief and generalized overview of the system.

Stemming from the fundamental tenet — the unavoidability of semantics in designing any computational NLP system — DMA incorporates a large and richly structured hierarchy of ontological concepts — the Ontology. In addition to the basic ALL → (EVENT, OBJECT, ATTRIBUTE) branching (see Figure 1), each of ca. 8,000 concepts is also defined through a large set of properties (both unique and inherited) of a slot-filler structure, whose fillers are other concepts (see Figure 2). This results in:

- a highly complex nature of the ontology, which is constrained, on the one hand, by the general principle of parsimony of its acquisition, and on the other, by the natural organization of objects, events and properties in the world, which the ontology models;
- a highly versatile nature of the ontology; the highly entangled (hypero-hyponymic, mereological, causal, etc.) conceptual network enables the ontology to emulate human semantic competence;

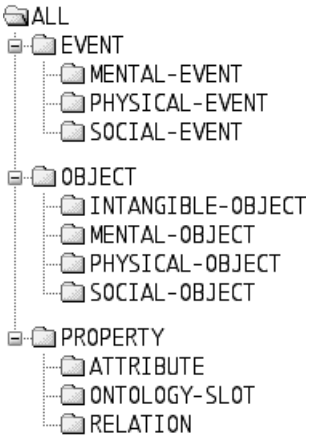


Figure 1. Basic Ontology branching: the root branch ALL breaks into OBJECT, EVENT, and PROPERTY

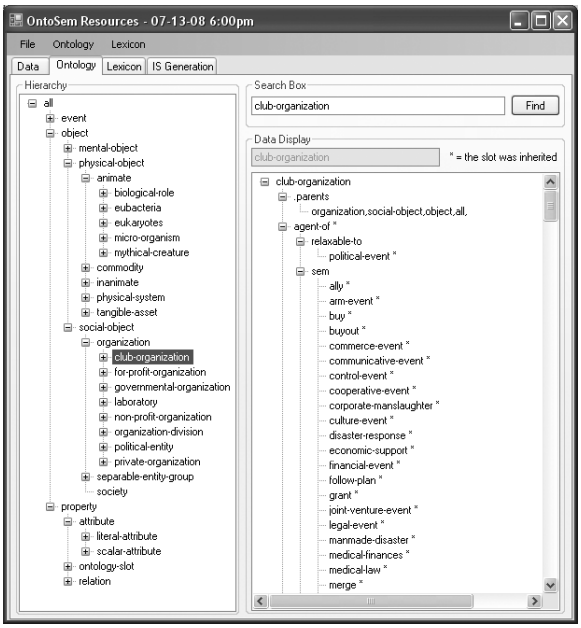


Figure 2. A snapshot of the ontology browser with the concept hierarchy and a detailed description for the CLUB-ORGANIZATION concept

A language-dependent Lexicon (120,000 entries for English, fewer for Spanish, Russian and Turkish) constitutes another static knowledge resource within DMA. The Lexicon features semantic (linking to a concept and its properties) and syntactic (case roles, selection restrictions) information for each entry (see Figure 3). Proper names are stored in a 25,000 entries-large Onomasticon.

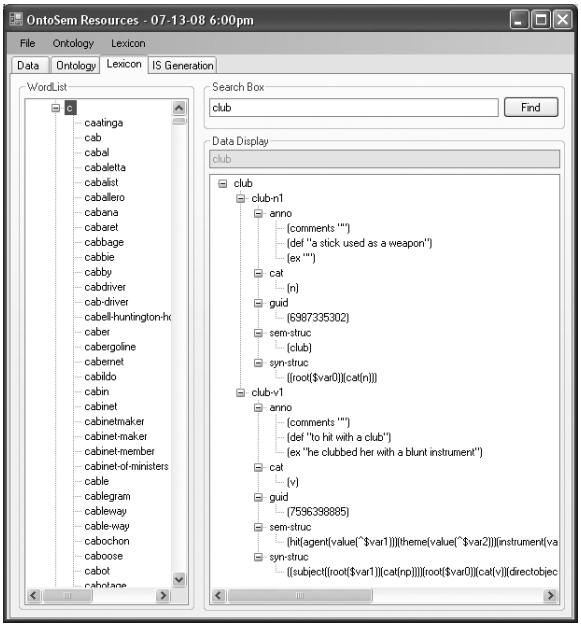


Figure 3. Lexical entries of club-n1, and club-v1 for the "club" super-entry in the Lexicon

The Fact Repository stores instances of concepts (head concepts, constraining properties and case-role fillers) from the immediate input. This allows the parsing module to operate across clauses and reconstruct elliptic and contradictory segments from recent slot fillers (see [25] for details on the contradiction-detection application of DMA).

The OntoParser, a dynamic processing module, utilizes static knowledge resources and proceeds in a step-by-step fashion from clause-breaking to lexical instantiations of concepts, syntactic constituents (e.g. multiple NP resolution), events, their case-role fillers and, ultimately clause merging, temporal, modality features, resulting in the text-meaning representation (TMR). Figure 4 illustrates the working of the OntoParser:

TMR is the final output of a DMA-informed NLP application. It constitutes lock, stock and barrel of any NLP enterprise and serves as a foothold for further machine-based applications: (information assurance, security, search, retrieval, etc.). Within DMA, a typical TMR features an event-driven description of a clause (from sentential level up) with the head event(s) and its case-role fillers. For example, the processing of the input

(1) The outlaws ran cocaine into the U.S.

would yield the following TMR:

```
(smuggle
(agent(sem(criminal)))
(theme(sem(cocaine)))
(destination(sem(country(has-name(value("united-states"))))))
)
```

, where the concept SMUGGLE in the Ontology is disambiguated through the lexical verbal entry run-v6 in the Lexicon, whose theme is the concept CONTROLLED-DRUG, which, in turn, is the ontological parent for the concept COCAINE (a more in-depth analysis of the example can be found in [28]).

Table 1 illustrates a wide range of application of Ontological Semantics and DMA with various degree of implementation (adapted from Raskin et al. 2004).

2.1. Abductive reasoning and Ontological Semantics: theory

The research on abductive reasoning, first defined in [17], is developing rapidly [1, 4, 5, 10, 11, 13, 15, 18, 29]. Generally, abduction is resorted to when an explanation of a fact is required, and no ultimately definitive theories are available. Most sources agree on a two-phased structure of abduction. It involves delineating a set of hypotheses, at which point a "leap of faith" is done by an abducing human when from the set of generated hypotheses the most plausible yet potentially defeasible candidate is selected. Selection criteria and algorithms are subject to debates in the literature (see [11] for a detailed overview).

Interesting parallels between abduction and the OntoParser can be drawn. Similar to abducing human

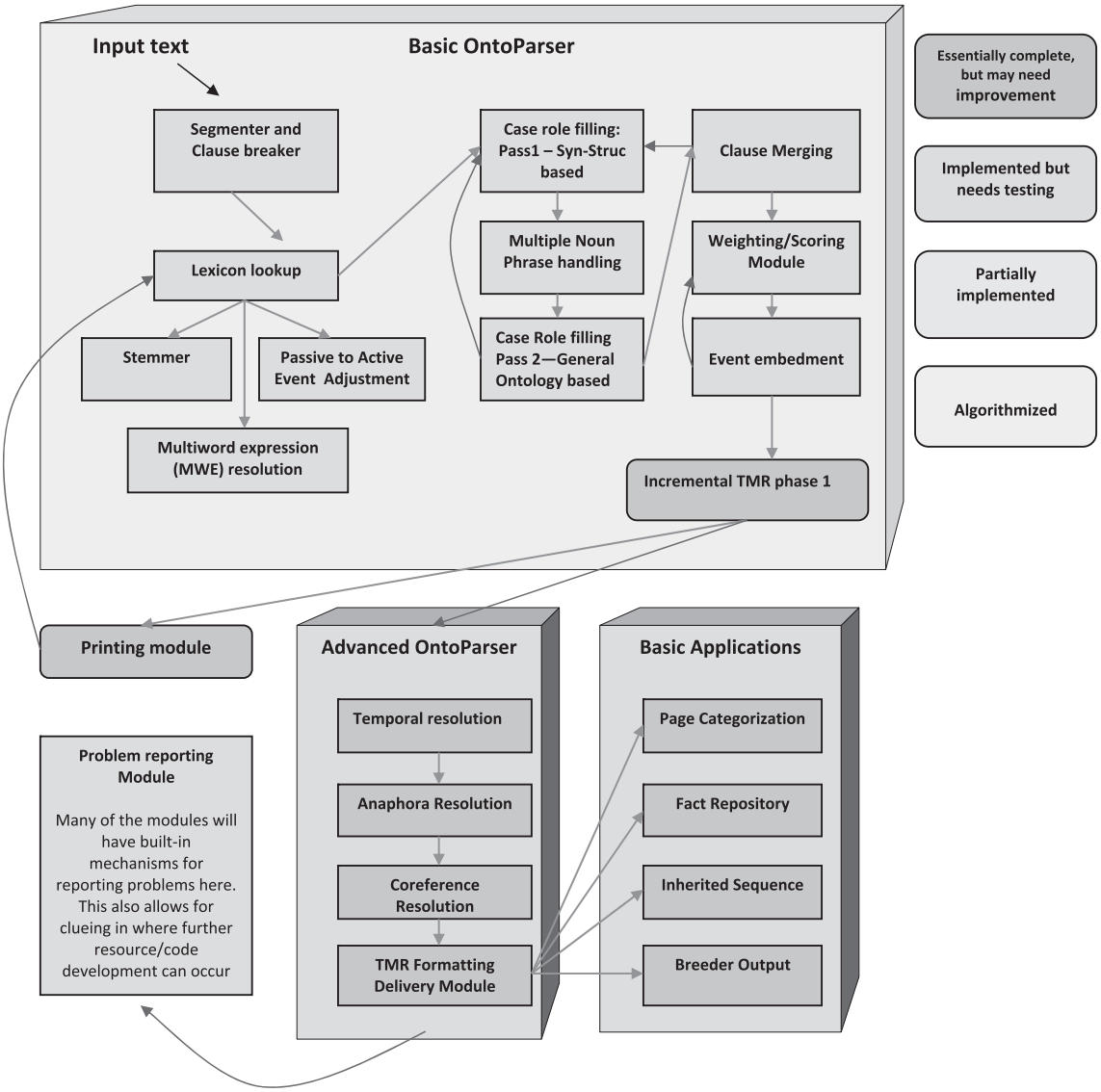


Figure 4. The flowchart of the OntoParser with sub-modules and steps of processing

Table 1. Areas of DMA application

Application	Function	Implementation	Reference
Syntactic NL Watermarking	Embeds the watermark in the syntactic tree of a sentence	Pilot/demo	[2]
Semantic NL Watermarking	Embeds the watermark in the TMR tree of a sentence	Pilot	[3]
NL Tamperproofing	Embeds a brittle watermark to detect any changes to the text	Pilot	[3]
NL Sanitization	Seamlessly removes and replaces sensitive information	Proof of concept	[12]
Automatic Terminology Standardizer	Translates different terminological dialects in IAS into TMRs	Proof of concept	[23]
Perimeter Protection	Sanitizes outgoing e-mail online	Proof of concept	[22]
NL Streaming Processor	Interprets incoming information before it is complete	Research	[23]
Ontosem-based humor research	Processing, modeling and generation of humorous texts based on Ontosem and General Theory if Verbal Humor frameworks	Research	[6, 8, 19]
Ontosem-driven Internet search engine	Semantic search, classification, QA applications	Pilot/demo	[9]
Abductive reasoning modeling	Processing elliptic input segments	Research	[20]
Ontosem-based (de)classification	Controls information sharing across security levels	Research	[27]

agents handling ambiguity, the OntoParser processes elliptic input by:

- projecting its static knowledge resources (Ontology, Lexicon, Onomasticon) onto immediate input;
- constraining the resources by selecting a pool of possible slot-fillers, and, ultimately,
- detecting the most plausible candidate (or a minimal set of them) within that pool.

Abduction thus lies in the core of OntoParser's processing of ambiguity. Both human agent and OntoParser operate on same principles. What the "pool of possible explanations" and the "best explanation" are to the human competence is what the "multiplicity of slot-fillers (mainly case-role fillers)" and the "constrained filler" are to the Ontoparser (Table 2).

Table 2. Similarity of two-staged abductive reasoning by human agent and OntoParser

Phase	Human competence	OntoParser
Hypotheses set	Pool of plausible hypotheses selected based on general knowledge	Set of case-role fillers delineated based on ontological properties and data from lexical entries in the input
Candidate selection	Best plausible explanation generated based on immediate goals	Most plausible case-role filler identified based on Fact Repository data and other explicit slot fillers

2.2. Abductive reasoning and Ontological Semantics: practice

The section will discuss two cases of ellipsis and strategies for their processing by the OntoParser.

Modeling abduction within Ontological Semantics is an ongoing study. Currently, two routes are being explored for enabling the parser to effectively process ambiguous (in particular, elliptic) input. Each route is defined by the initial input conditions, i.e. what segments of the text are missing and need restoring.

- For input with elliptic case roles (and explicit clause-forming events), inference rules can be designed that capture dependencies between case role fillers across clauses; if the dependency is there, the elliptic case role filler can be reconstructed based on the rule and the already present fillers from the recent input (stored in the Fact Repository);
- For input with elliptic clause-forming events (and explicit case roles), inference rules can be designed that capture correlations between scalar attributes and the epistemic modality values; when applied, the rule allows to narrow down the pool of potential elliptic events;

Examples will illustrate each case.

The implementation of an inference rule establishing case-role dependencies in clauses with specific event classes was described in detail in [20]. The rule allowed reconstructing non-verbalized fillers for the case role of destination in events whose EFFECT property slot was filled by a MOVEMENT-EVENT with an explicit destination case role:

(2) A bomb was thrown at a building. No serious damage reported. (Roughly adapted from [20])

```
(throw                                  (damage
(theme(sem(bomb)))              →       (theme(sem(building —
                                        reconstructed)))
(destination(sem(building))))           )
```

Based on the rule and employing the immediate data from Fact Repository and proceeding algorithmically, the parser reconstructs missing theme case role for the event DAMAGE with BUILDING from the preceding THROW (which IS-A MOVEMENT-EVENT) event, where BUILDING fills in the destination slot.

The inference rule below stipulates identical fillers for agent and theme case roles in CHANGE-IN-QUANTITY events related through the PRECONDITION and EFFECT properties. Example (3),

(3) If the US plans to increase its troops in Afghanistan, the reduction in Iraq needs to be accelerated.

features two events, INCREASE and DECREASE (represented by increase-v1 and reduce-v1), the latter having a non-verbalized theme filler. Before the inference rule is applied, the following TMR would be produced:

```
(increase
 (agent(sem(country(has-name(value("united-states"))))))
 (theme(sem(military-unit)))
 (location(sem(country(has-name(value("afghanistan")))))))
 (volitive(value(>0.5)))
 (precondition(sem
  (decrease
   (agent(sem(nothing))))
   (theme(sem(nothing)))
   (location(sem(country(has-name(value("iraq"))))))
   (saliency(value(>0.5)))
   (velocity(value(>0.5)))
  )
 )
)
```

The italicized "nothing's" indicate that no explicit or default fillers for the slots have been found the parser. This being said, the following rule can be formulated:
For events E1 and E2,
If E1 and E2 are CHANGE-IN-QUANTITY events, which stand in the PRECONDITION or EFFECT relation, then E1 and E1 have identical fillers for agent and theme case roles, if no other explicit or default fillers are available.
A formal definition of the rule is provided below:

IF (E1(pre-condition/effect(E2)));
E1 = CHANGE-IN-QUANTITY
E2 = CHANGE-IN-QUANTITY
No explicit case role fillers for E2 available;
THEN (E1(agent)) = (E2(agent))
(E1(theme)) = (E2(theme))

As the result of the rule application, the concepts COUNTRY(has-name(value("united-states"))) and MILITARY-UNIT will fill in the slots for agent and theme case roles in the DECREASE event.

Reconstructing elliptic events presents a challenge. In the ongoing study in this area, an interesting avenue is currently being explored, in which for cases with explicit scalar attribute values, explicit case roles and implicit events, inference rules can be designed based on the correlation between the value of epistemic modality and the value of large class of SCALAR-ATTRIBUTE's for the non-verbalized event. In the current Ontology, it appears, some scalar attributes can determine the value of the epistemic modality of the clause-forming event. To illustrate, the zero value of INTELLIGENCE attribute seems to precondition the zero value of epistemic modality for the branch of COGNITIVE-EVENT's:

IF (intelligence(value(0)))
THEN (cognitive-event(epistemic(value(0))))

Example (4),

(4) I tried to explain him the theory, but he turned out to be completely dumb.

contains an explicit zero value of the INTELLIGENCE attribute and a non-verbalized EVENT:

(expressive-act
(agent(sem(human)))
(theme(sem(theory)))
(beneficiary(sem(human intelligence(value(0)))))
(domain(value(*event missing*)))
(epiteuctic(value(<0.5)))
)

Naturally, in the given situation, successful understanding presupposes a higher-than-zero value of intelligence property attributed to the recipient. In the TMR, the italicized *event missing* can be reconstructed through an inference rule which would constrain the domain[1] for the INTELLIGENCE attribute to a number of event classes, COGNITIVE-EVENT being one of them. The case role filler match would then narrow the search down to a specific event. A rule can thus be formulated:

For a clause C that contains a verbalized case role CR, a verbalized scalar attribute SA preconditioning a set of event classes EC, and non-verbalized event E, the value of SA will be identical with the value of the epistemic modality of E, and CR will fill in one of the slots for E;

For (E(SCALAR-ATTRIBUTE(value(x)))(case-role(y)))
IF E is elliptic;
x is verbalized;
y is verbalized;
THEN
(E(case-role(y));
(epistemic(value(x)))),
where E ∈ {EC}; y ∈ {EC(case-roles)};

The tentative list of preconditioning scalar attributes includes, but not restricted to:

difficulty-attribute, feasibility-attribute, intensity, orderliness, precision-attribute, rapidity, safety, secrecy-attribute, success-attribute, survivability, treatability, stability, age, endurance, flexibility, resistance, roundness, almost all SCALAR-HUMAN-ATTRIBUTE, etc.

For each of these attributes, a limited number of events can be defined as their domain. Based on this dependency, the OntoParser would:

1) Look up the set of events preconditioned by the attribute in question;
2) From this set, select events whose case-role fillers can be found in the given input;
3) Identify the non-verbalized event;
4) Based on the given attribute value, determine the probability of the event by assigning a specific value of epistemic modality;

A tentative algorithm can thus be formulated:

1. Input contains an ATTRIBUTE?
   Yes — go to 2. No — terminate.
2. Non-verbalized EVENT in the clause?
   Yes — go to 3. No — terminate.
3. Identify events preconditioned by the attribute.
   Yes — go to 4. No — terminate.
4. Identify case-role fillers in the clause.
   Yes — go to 5. No — terminate.

[1] The notions of "attribute domain" and "attribute range" need clarification here. Within Ontological Semantics, along with OBJECTS and EVENTS, the ontology features a class of PROPERTIES, of which ATTRIBUTE is a subclass. Every PROPERTY is defined through domain and range: property (domain, range). The set of events or objects to which PROPERTY pertains is defined in its domain, whereas the set of values of PROPERTY defined in its range (see [14], but also [16: 31] for similarities with the notions of domain and range of a mathematical function). To illustrate, in the ontology, AGE is a literal attribute, whose domain is EVENT and OBJECT, and whose range is ANY-NUMBER (i.e. a numeric value of age of a particular object or event). The author is grateful to the anonymous reviewers for emphasizing the need to elaborate on the notions.

5. Match the case-role fillers with those for the events preconditioned by the attribute.
   Declare the selected set of events elliptic.
6. Identify the value of the ATTRIBUTE.
   Assign identical value of EPISTEMIC modality to the event or events declared elliptic.

Consider the examples:

(5) He saw the shore on the horizon but was too tired.

(6) He found the spade but was too tired.

For both examples, the parser would identify the following set of events preconditioned by the STRENGTH-ATTRIBUTE: dig, enclose, shift, shift-material, entwine, unwrap, wrap, fasten-together, operate-device.

In the set, the events would be selected which could have COASTAL-GEOLOGICAL-ENTITY (lexical item "shore") and SHOVEL (lexical item "spade") as case-role fillers:

```
(5) (coastal-geological-entity
      (destination-of(sem(horizontal-liquid-motion)))
    )

(6) (shovel
      (instrument-of(sem(operate device dig shift-material
      enclose fasten-together operate-device unwrap wrap)))
    )
```

Having narrowed down the search, the parser would then proceed with the STRENGTH-ATTRIBUTE and EPISTEMIC value assignment, at which point the ultimate TMR's for the two examples will be generated:

```
(5) He saw the shore on the horizon but was too tired.
(visual-event
   (agent(sem(human(gender(value(male))))))
   (theme(sem(coastal-geological-entity)))
(horizontal-liquid-motion
   (agent(sem(human
         (gender(value(male))))))
         (strength-attribute(value(<0.2)))
   (destination(sem(coastal-geological-entity)))
   (epistemic(value(<0.2)))
   )
)

(6) He found the spade but was too tired.
(find
   (agent(sem(human(gender(value(male))))))
   (theme(sem(shovel)))
(operate device dig shift-material enclose fasten-together
operate-device unwrap wrap
   (agent(sem(human
         (gender(value(male))))))
         (strength-attribute(value(<0.2)))
   (instrument(sem(shovel)))
   (epistemic(value(<0.2)))
   )
)
```

In (6), the natural ambiguity of the input would be registered even by a human agent since it is unclear for what purposes the spade would be used. This is interpreted by the OntoParser as the multiplicity (significantly reduced, after to the inference processing) of possible events.

References

1. *Aliseda A.* Abductive Reasoning: Logical Investigations into Discovery and Explanation // Mexico: National Autonomous University of Mexico, 2006.
2. *Atallah M. J., Raskin V., Crogan M., Hempelmann C. F., Kerschbaum F., Mohamed D., Naik S.* Natural Language Watermarking: Design, Analysis, and a Proof-of-Concept Implementation // I. S. Moskowitz (ed.), Information Hiding: 4th International Workshop, IH 2001, Pittsburgh, PA, USA, April 2001 Proceedings. Berlin: Springer, 2001. P. 185–199.
3. *Atallah M. J., Raskin V., Hempelmann C. F., Karahan M., Sion R., Topkara U., Triezenberg K. E.* Natural Language Watermarking and Tamperproofing. // F. A. P. Petitcolas (ed.), Information Hiding: 5th International Workshop, IH 2002, Proceedings. Berlin: Springer, 2002. P.196–210.
4. *Attardo S.* On the Nature of Rationality in (Neo-Gricean) Pragmatics // International Journal of Pragmatics. (Special issue on Neo-Gricean pragmatics, edited by Ken Turner). 2003. V. 14, P. 3–20.
5. *Flach P. A., Kakas A. C.* Abduction and Induction: Essays on their Relation and Integration // Kluwer Academic Publishers, 2002.
6. *Hempelmann C. F.* Paronomasic Puns: Target Recoverability towards Automatic Generation. Unpublished Ph.D. Dissertation // Interdisciplinary Program in Linguistics. West Lafayette, IN: Purdue University, 2003.
7. *Hempelmann C. F., Raskin V., Triezenberg K. E.* Semantic Forensics: NLP Systems for Deception Detection // Proceedings of the First Annual MidwestColloquium in Computational Linguistics. Damir C., Rodriguez P.(Eds.) Bloomington, IN: Indiana University. 2004.

8. *Hempelmann C. F.* Computational Humor: Beyond the Pun // The Primer of Humor Research. Raskin V. (Ed.). Berlin, New York: Mouton de Gruyter, 2008. P. 335–363.
9. *Hempelmann C. F., Raskin V.* Semantic Search: Content Vs. formalism // Rome: Proceedings of Langtech 2008. http://www.langtech.it/en/technical_program/technical_program.htm (full paper).
10. *Hobbs J. R., Stickel M. E., Appelt D. E., Martin P.* Interpretation as Abduction. // Pereira, F.C.N. and Barbara J. Crosz (eds.) Natural Language Processing. MIT Press, 1994. P.69–142.
11. *Gabbay D. M., Woods J.* The Reach of Abduction. Insight and Trial // A Practical Logic of Cognitive Systems. Amsterdam: Elsevier Science, 2005. V. 2.
12. *Mohamed D.* Ontological Semantics Methods for Automatic Downgrading. An unpublished Masters' thesis // Interdisciplinary Program in Linguistics and CERIAS. West Lafayette, IN: Purdue University, 2001.
13. *Niiniluoto I.* 1998. Defending Abduction // Philosophy of Science, Supplement. Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association, 1999. Part I: Contributed Papers. Vol. 66.
14. *Nirenburg S., Raskin V.* Ontological Semantics // Cambridge, MA: MIT Press, 2004. A prepublication draft, chapter by chapter, can be found on www.ontologicalsemantics.com
15. *Paavola S.* Abduction as a Logic and Methodology of Discovery: The importance of strategies // Foundations of Science. 2004. V. 8. P. 267–283. http://www.helsinki.fi/science/commens/papers/abductionstrategies.html
16. *Partee B. H., Meulen A. T., Wall R. E.* Mathematical Methods in Linguistics. Dordrecht, Boston, London: Kluwer Academic Publishers, 1990.
17. *Peirce Ch. S.* Philosophical Writings of Peirce (1955) // Collected Papers. Charles Hartshorne and Paul Weiss (eds.), New York: Dover: Harvard University Press, 1960–1966. V. 8.
18. *Peng Y., Reggia J. A.* Abductive Inference Models for Diagnostic Problem Solving // New York, Berlin: Springer-Verlag, 1990.
19. *Petrenko M.* The Narrative Joke: Conceptual Structure and Linguistic Manifestation. An unpublished Ph.D. dissertation // Interdisciplinary Program in Linguistics, West Lafayette, IN: Purdue University, 2008.
20. *Petrenko M., Raskin V.* Modeling Abduction within Ontological Semantics // Proceedings of Midwestern Computational Linguistics Colloquium 5. East Lansing: Michigan State University, May 2008.
21. *Raskin V., Atallah M. J., McDonough C. J., Nirenburg S.* Natural Language Processing for Information Assurance and Security: An Overview and Implementations // M. Schaefer (ed.), New Security Paradigm Workshop. Proceedings. New York: ACM Press. Ballycotton, County Cork Ireland: 2001. P. 51–65.
22. *Raskin V., Atallah M. J., Hempelmann C. H., Mohamed D.* Hybrid Data and Text System for Downgrading Sensitive Documents // Technical report. West Lafayette, IN: Center for Education, Research in Information Assurance and Security, 2001.
23. *Raskin V., Nirenburg S., Atallah M. J., Hempelmann C. F., Triezenberg K. E.* Why NLP should move into IAS // Steven Krauwer (ed.), Proceedings of the Workshop on a Roadmap for Computational Linguistics. Taipei, Taiwan: Academima Sinica, 2002. P. 1–7.
24. *Raskin V., Nirenburg S., Nirenburg I., Hempelmann C. F., Triezenberg K. E.* The genesis of a script for bankruptcy in ontological semantics // ACL: Edmonton, Alberta, Canada: G. Hirst and S. Nirenburg (eds.), Proceedings of the Text Meaning Workshop, HLT/NAACL 2003: Human Language Technology and North American Chapter of the Association of Computational Linguistics Conference., 2003
25. *Raskin V., Hempelmann C. F., Triezenberg K. E.* 2004. Semantic Forensics: An Application of Ontological Semantics to Information Assurance // ACL Publication. G. Hirst and S. Nirenburg (eds.), Proceedings of the Second Workshop on Text Meaning and Interpretation, ACL, Barcelona, Spain, Edmonton, Alberta, Canada: 2004.
26. *Raskin V., Hempelmann C. F., Triezenberg K. E.* Ontological Semantic Forensics: Meaning-based Deception Detection // Meaning computation, 2008. V. 1.
27. *Raskin V., Hempelmann C. F., Triezenberg K. E., Buck B., Keen A.* Accessing and Manipulating Meaning of Textual and Data Information for Information Assurance and Security and
28. *Intelligence* Information // Proceedings of the Cyber Security and Information Intelligence Research Workshop, Oak Ridge National Laboratory, Oak Ridge, Tennessee: May 12–14, 2008.
29. *Raskin V., Hempelman C. F., Taylor J. M., Petrenko M. S., Trienzenberg K. E., Buck B.* The Why's, How's, and What-of's of Natural Language Ontology // 2nd Interdisciplinary Ontology Conference. Proceedings. Tokyo: 2009 (forthcoming)
30. *Walton D.* Abductive Reasoning // University of Alabama Press, 2004.

Abstracts

KEYWORD SEARCH USING SYLLABLE LATTICE

Aliyev R. M. (RomanAliyev@gmail.com), **Kheidorov I. E.** (igorhmm@mail.ru), **Yan Jinbing** (yjbemail@163.com),
Belarussian State University, Minsk, Belarus

Syllable lattice-based keyword search methods may help to overcome the problem of Out of Vocabulary (OOV) words and compensate the loss of search performance caused by recognition error. While there has been no effective search model in lattice-based search approaches, a syllable posterior probability-based search model is proposed. The model takes account of the lattice structure and syllable posterior probability. A search method based on the model is proposed. A series of experiments shows that our method is suitable for keyword search.

SEMANTIC SOURCES OF CONCESSION

Apresjan V. Ju. (valentina.apresjan@gmail.com), Russian Language Institute

The paper addresses the issue of concession as a complex derived meaning and analyzes its semantic origins. It also considers polysemy of concessive words and proposes semantic tools to distinguish among closely synonymous concessives derived from words with a non-concessive primary meaning. In particular, the following lexical items are analyzed: concessive conjunction "tol'ko" derived from a restrictive particle, and concessive conjunction/parenthetical word "pravda" derived from a factual noun. Their similarities and differences are analyzed in the light of the primary meanings of "tol'ko" and "pravda".

STATISTICAL DISTRIBUTIONS OF WORDS IN A COLLECTION OF RUSSIAN TEXTS

Baglei S. G. (baglei@galaktika.ru), **Antonov A. V.** (alexa@galaktika.ru), **Meshkov V. S.** (meshkov@galaktika.ru),
Sukhanov A. V. (sukhanov@galaktika.ru), Galaktika Corporation, Moscow, Russia

Statistical properties of texts have been widely studied in the fields of applied mathematics and linguistics. We explored statistical distribution of words in documents of a large collection of Russian texts using a probabilistic Bernoulli text generation process in our model. Unlike the traditional Bernoulli process, each document in the collection is considered as a finite text. We explored distributions of word frequencies in texts within a model representing a set of "bags-of-words". We plan to use the obtained results to elaborate a more realistic estimated probability of word generation in arbitrary Russian text with regard to word correspondence to the text collection.

SEMANTIC CORRELATES OF FORMAL VARIATION IN THE FIELD OF IDIOMATICS (THE OPERATION OF SUBSTITUTION)

Baranov A. N. (baranov_anatoly@hotmail.com), Institute of Russian Language, Moscow, Russia

The issue of formal variation of idioms is discussed. The paper focuses on the operation of substitution of different components on an idiom. A classification of different types of substitution operation is elaborated. It is hypothesized that formal variations of different kinds have specific semantic and discursive functions. Linguistic description of variation in the field of idiomatic presupposes an analysis of correlation between formal variation and meaning changes in an idiom. In is shown that substitution of the components of an idiom in most cases results in a generation of alternative semantic levels and, consequently, a linguistic play.

WWW STATISTICAL ESTIMATION OF THE FUNCTIONAL PROPERTIES OF LEXICAL ITEMS

Belikov V. I., Russian Language Institute n. a. V. V. Vinogradov, Russian Academy of Sciences
Akhmetova M. V., Journal "Zhivaia Starina" (Moscow, Russia)

The paper deals with the possibilities of using web-cite statistics for objective estimation of functional properties of vocabulary items: their stylistic status, territorial distribution, obsolescence of an item and its replacement by a new one, etc. The functional properties of particular words and phraseological units reveal themselves in their frequencies in different text arrays (classical vs. web-literature, official texts, weblogs, etc).

CREATING CONCEPTUAL GRAPHS AS ELEMENTS OF SEMANTIC TEXT LABELING

Bogatyrev M. Y.(okkambo@mail.ru), **Tuhtin V. V.**, Tula StateUniversity

Prospects of applying conceptual graphs as elements of semantic text labeling are considered. This kind of labeling is metadata that can be used to effectively solve some of the Text Mining problems. An algorithm for creating conceptual graphs is proposed and some results of its applications to modeling abstracts of scientific papers are presented.

A SPEECH CORPUS AS A TOOL FOR MONITORING AND FIXATION OF VARIOUS FORMS OF NATURAL LANGUAGE

Bogdanova N. V. (nvbogdanova_2005@mail.ru), **Asinovsky A. S.** (a.s.asinovsky@gmail.com),
Rusakova M. V. (mvrusakova@gmail.com), **Ryko A. I.** (aryko@yandex.ru),
Stepanova S. B. (stsvet_2002@mail.ru), **Sherstinova T. Yu.** (sherstinova@gmail.com),
Saint Petersburg State University, Russia

The paper concerns methodological principles and describes the technology of creation of the Corpus of Spontaneous Russian Speech and the structure of the database. Preliminary investigations based on Corpus material are briefly presented.

CROSSLEXICA: A LARGE ELECTRONIC DICTIONARY OF COLLOCATIONS AND SEMANTIC LINKS BETWEEN RUSSIAN WORDS

Bolshakov I. A. (bolshakov34@mail.ru), National Polytechnic Institute, Mexico

A large Russian electronic dictionary contains a vocabulary of 185,000 entries, 1.75 million collocations, 2 million semantic links, English translations of entry titles, and their morphoparadigms. It functions dialogically (for text editing or language learning) and is also accessible from external software for parsing, word sense disambiguation, detection & correction of malapropisms, steganography, etc.

CREATING A SEMANTIC DICTIONARY OF PREPOSITIONAL CONSTRUCTIONS ON THE BASIS OF THE UKRAINIAN NATIONAL LINGUISTIC CORPUS

Bugakov O. V. (ovbugakov@gmail.com), Ukrainian Lingua-Information Fund, NAS of Ukraine, Kiev, Ukraine

Search capabilities of the Ukrainian national linguistic corpus and linguistic databases built on its basis are examined. The structure of the semantic dictionary of prepositional constructions built in accordance with the theory of lexicographic systems is described. Key words: preposition, main word, dependent word, semantic state, electronic semantic dictionary of prepositional constructions.

MARKUP OF TEXT FRAGMENTS DURING CLASSIFICATION

Vasilyev V. G. (vvg_2000@mail.ru), Institute of Informatics Problems of the Russian Academy of Sciences (IPI RAN)

A comparative analysis of approaches to the selection of meaningful fragments of texts by using statistical methods of classification is presented. We consider new algorithms based on hidden Markov models covering the text by special hierarchical multiple fragments, as well as based on pre-segmenting the text into fragments without taking account of the information about the structure of classes.

RUSLED DICTIONARY AS TOOL FOR SEMANTIC STUDY

Voskresenskiy A. L. (avosj@yandex.ru), **Gulenko I. E.** (gig@yandex.ru), **Khakhalin G. K.** (gkhakhalin@yandex.ru)

The use of Russian sign language dictionary as an indicator of various Russian words meanings is described. This approach is enabled to more purposefully carry out analysis of context for word disambiguation.

THE DIGITAL RUSSIAN ASSOCIATIVE DICTIONARY OF SCHOOLCHILDREN

Goldin Valentin (goldinve@yandex.ru), **Martianov A. O.** (comrad-mao@mail.ru),
Sdobnova Alevtina (sdobnovaap@yandex.ru), Saratov State University

The paper deals with some ways of solving different issues of psycholinguistics, sociolinguistics and culturology based on the materials of the digital “Associative Dictionary of Schoolchildren of Saratov city and Saratov region”.

ON THE NATURE OF SYNTACTIC POLISEMY

Grigorian E. L. (elena_grigorian@yahoo.co.uk), South Federal University, Rostov-on-Don, Russia

The analysis of variations of actant structures reveal the fact that most syntactic structures represent a set of semantic features which are not necessarily realized in every context. In many cases semantic distinctions are neutralized and the constructions differ only in communicative structure or style.

ON GESTURE–WORD CORRELATION (VOCAL GESTURE OH IN SPOKEN RUSSIAN)

Grishina Elena (rudi2007@yandex.ru), Institute of Russian Language, RAS, Moscow, Russia

The paper analyzes the usage of the vocal gesture Oh according to the data of the future Multimodal Russian Corpus (MURCO). The investigation is based on the analysis of the body and face movements that accompany this vocal gesture in the process of oral speech. As a result three meanings were detected 1) Oh as a deixis, 2) Oh as an interjection, and 3) Oh as a physiological exclamation.

UNIVERSAL DICTIONARY OF CONCEPTS

Dikonov V. G. (dikonov@iitp.ru), **Boguslavsky Igor M.** (bogus@iitp.ru),
Institute for Information Transmission Problems, Russian Academy of Sciences

A universal dictionary of concepts, developed as a part of the ongoing effort to create a semantic intermediary language for global information exchange, is presented. The article describes basic principles and contents of the dictionary and outlines the current state of the project. The dictionary can evolve into an open and freely available language-neutral resource with many potential applications. For example, the extensible dictionary of concepts can serve as a pivot to uniformly record and link meanings of words of different languages and facilitate creation of bi- and multilingual dictionaries. Another possible use is word sense markup of corpora. The dictionary of concepts is going to be linked at the word sense level with lexicons of major world languages including Russian, English, Spanish, French, Arabic, Hindi, etc.

RUSSIAN NET, GERMAN NEIN , ENGLISH NO: CONTRASTIVE SEMANTIC ANALYSIS WITH PARALLEL CORPORA

Dobrovol’skij D.O. (dm-dbrv@yandex.ru), **Levontina I. B.** (irina.levontina@mail.ru),
Russian Language Institute, Russian Academy of Sciences

‘No’ seems to be a very simple and universal idea. However, surprisingly enough, the German word nein or the English no are not always good equivalents for the Russian word net, and vice versa. Parallel corpora show that in many cases net is translated differently, even though the respective phrase with nein/no is acceptable. And we often see net in Russian translation instead of some other units. We assume that such lack of coincidence must have certain semantic reasons. They are probably rooted in semantic differences between net and nein/no. In our paper we try to reveal these reasons.

NATURAL LANGUAGE QUERY PROCESSING FOR SEARCH ENGINE BASED ON LINGUISTIC ANALYSIS

Ermakov A. E. (ermakov@metric.ru), **Pleshko V. V.** (vp@rco.ru), RCO Ltd (www.rco.ru), Moscow

A new method of transforming natural language queries into search engine language queries is described, which is based on the automatic analysis of syntactic relations between words and their representation as relevant search engine language operators saving the meaning of an original query to the extent possible.

ON THE NOTION OF SEMANTIC SHIFT

Zalizniak Anna A. (anna-zalizniak@mtu-net.ru), Institute of linguistics, Russian Academy of Sciences

The paper deals with the notion of "semantic shift" as a category of semantic typology and the unit of the "Catalogue of semantic shifts in the languages of the world"; it reflects some results of the work on a project, realized in the Institute of Linguistics, Russian Academy of Sciences, by a group of linguists (Anna A. Zalizniak, Maria Bulakh, Dmitriy Ganenkov, Ilya Gruntov, Timur Maisak and Maxim Russo). The problem of identification of semantic shifts in cases of syncretism (semantic generality) is discussed in more detail.

<STRIKE>I'VE NEVER TOLD THAT</STRIKE>: ABOUT LITURATIVES, STRIKEOUTS OR IMAGINARY TEXTS

Zanegina N. N. (zanegina@list.ru), The Institute of the Russian Language

This paper deals with linguistic peculiarities of strikeout texts — their semantics and syntax. These texts are very often used in Internet communication.

AN APPROACH TO AUTOMATED ONTOLOGY BUILDING IN TEXT ANALYSIS PROBLEMS

Zakharova I. V. (iren@csu.ru), **Gorodechnyj P. P.** (petr.gorodechnyj@edu.csu.ru),
Chelyabinsk State University, Dept. of mathematics

An approach to how to automatically build an ontology for complex tasks of full-text document classification using UDC is discussed.

UNIVERSAL SYNTAX ANNOTATION SYSTEM OBJECTATE

Zobnin A. I. (Alexey.Zobnin@gmail.com), Lomonosov State University,
Russian Language Institute n.a. V. V. Vinogradov, Russian Academy of Sciences
Sakharova A. V. (nenen@mail.ru), Russian Language Institute n.a. V. V. Vinogradov, Russian Academy of Sciences

The object model and the features of the Universal text annotation system ObjectATE are described. This system is used in Vinogradov Institute of Russian language of RAS for semimanual morphological and syntactical annotation of ancient manuscripts. It allows the user to define his own annotation models by describing classes, add-ins, fields and relations in the metadata layer (for example, for syntax markup).

EVERYDAY TERMINOLOGY. IN PURSUIT OF STANDARDS

Iomdin B. L. (iomdin@ruslang.ru), Russian Language Institute n.a. V. V. Vinogradov, Russian Academy of Sciences

The paper is devoted to the vocabulary describing everyday life artifacts. This vocabulary is shown to be treated very differently in dictionaries, production standards, and usage; henceforce, unified lexicographic definitions of words belonging to this vocabulary are hardly possible at all. A draft project of an explanatory and encyclopedic thesaurus of everyday life terminology is presented.

SYNTACTIC CORRELATES OF PROSODICALLY MARKED ELEMENTS OF THE SENTENCE AND THEIR ROLE IN THE TASKS OF TEXT-TO-SPEECH SYNTHESIS

Iomdin L. L. (iomdin@iitp.ru), Institute for Information Transmission Problems, Russian Academy of Sciences
Lobanov B. M. (lobanov@newman.bas-net.by), United Institute of Informatics Problems,
National Academy of Science of the Republic of Belarus

The paper describes a feasibility study of using syntactic parsing of written text at an initial stage of text-to-speech synthesis algorithm. An attempt has been made to establish correlations between the elements of an automatically created dependency tree structure of a sentence, on the one hand, and prosodically strong elements of this sentence, on the other hand. First experimental results show that the approach may be effective.

PROSODIC TRANSCRIPTION: LEVELS OF DETAIL

Kibrik A. A. (kibrik@comtv.ru), Institute of Linguistics RAS, **Khudyakova M. V.** (mariya.kh@gmail.com), Lomonosov Moscow State University, **Kodzasov S. V.** (sankod@yandex.ru), Lomonosov Moscow State University

In the book Kibrik and Podlesskaya (eds.) 2009, a prosodically oriented system of discourse transcription for spoken Russian was proposed. In this paper a number of extensions for that system are suggested, such as the distinction between expiratory and pitch accents, a more detailed account of pitch accents, interval of tone in an accent, dynamic vowel doubling, etc.

TOWARDS THE PROBLEM OF LINGUISTIC VARIABILITY: A MULTIFACTOR SECOND-ORDER CALCULUS METHOD

Kibrik A. E., Lomonosov Moscow State University, Russia

Clausal coordination is studied in 23 related Daghestanian idioms. Clausal coordination is extremely variable across this language sample: there is not a pair of idioms with identical coordinate clausal constructions. At first sight, the choice of formal coding technique used by specific idioms appears random and chaotic. Such situation creates irresolvable theoretical difficulties. Neither the traditional method of classification nor the structural calculus method are helpful.

In the paper an alternative method is employed. It can be called the multifactor second-order calculus method. A calculus of coordinate constructions is implemented at the level of parameterized principles and strategies predetermining specific coordinate constructions, rather than at the level of coordinate constructions themselves.

THE METHOD OF AUTOMATED SYNTAX SEGMENTATION RULES GENERATION

Klyshinsky E. S. (klyshinsky@mail.ru), Keldysh Institute of Applied Mathematics RAS
Manushkin E. S. (EugeneLebowsky@mail.ru), Moscow State Institute of Electronics and Mathematics

The paper proposes a method of automated generation of syntax segmentation rules. The method is based on FIRST, LAST, FIRST2 and LAST2 sets calculated for existing BNF grammars describing the rules for syntax analysis of natural languages texts.

SYNTACTIC INCOMPATIBILITY AS A PROPERTY OF THE LINEAR ORGANIZATION OF A RUSSIAN SENTENCE

Kobzareva T. Yu. (stamstam@mtu-net.ru), Russian State University for Humanities, Russia

The paper considers a property of the linear organization of sentence in Russian, the so-called syntactic incompatibility, or impossibility of simultaneous appearance of some components in its fragments set by punctuation marks or coordinative conjunctions. The property can be taken into account at different stages of automatic analysis.

SEMANTICS OF THE VERB PONIMAT': FROM PROPOSITIONAL TOWARDS INTERPERSONAL ATTITUDE

Kobozeva I. M. (kobozeva@list.ru), Lomonosov Moscow State University, Moscow, Russia

The Russian verb ponimat' 'understand' in constructions with a personal direct object is studied. 6 of its readings, corresponding to different intentional states (rational, emotional, interpersonal) are explicitly defined. The emergence of non-rational readings is explained on the cognitive basis.

THE DATABASE ON INTONATION OF RUSSIAN NARRATIVE TEXTS

Kodzasov S. V. (sankod@philol.msu.ru), **Arkhipov A. V.** (arxipov@philol.msu.ru), **Zakharov L. M.** (leon@philol.msu.ru),
Krivnova O. F. (okri@philol.msu.ru), Lomonosov Moscow State University, Moscow, Russia

The paper represents the results obtained at the 2nd stage of development of the DB "Intonation of the Russian informative and narrative texts". This stage opened the 2nd triennial cycle of inquiry into Russian intonation.

DETECTION OF NOMINALIZED STRUCTURES IN PARALLEL PATENT TEXTS IN RUSSIAN AND IN GERMAN

Kozhunova O. S. (kozhunovka@mail.ru), Institute for Informatics Problems of the Russian Academy of Sciences

In the paper nominalization in bilingual situation (Russian-German) involving comparative study results for three languages (Russian, English, and German), approach of parallel texts identification for patent sphere and transformation types have been analyzed.

PARENTHESES IN RUSSIAN IDIOMS

Kozerenko A. D. (akozerenko@mail.ru), V. V. Vinogradov Russian Language Institute, Russian Academy of Sciences

The paper offers a semantic analysis of Russian idioms containing the word parentheses. A paradoxical fact is observed that two idioms that in Russian sound as enclose in parentheses and put outside the parentheses have the same meaning. The clue is given and other Russian idioms containing the word parentheses are examined.

PAUSES AFTER POSTPOSITIONS AND TOPICAL PARTICLE WA IN JAPANESE: A CORPUS STUDY

Komarova A. D. (komarovochka@gmail.com), Russian State University for Humanities, Russia

This research discusses pauses after postpositions and topical particle wa and before them in Japanese. It aims to find out how frequent and probable these pauses are, their usual length and if they differ depending on the syntactic position.

A CORPUS STUDY OF PAUSATION AT SYNTACTIC BOUNDARIES: WHY PAUSES DO NOT ALWAYS APPEAR WHERE WE EXPECT THEM

Korotaev N. A. (n_korotaev@hotmail.com), Russian State University for Humanities (RSUH)

The so-called ideal delivery presupposes a pause at every elementary discourse unit boundary. Natural discourse, however, provides numerous examples when such pauses are missing. The paper reports a corpus-based study of these cases in spoken Russian. It is argued that they are characterized by a high degree of semantic integration, which correlates with syntactic and prosodic properties of the examined sequences. For instance, absence of pauses is intrinsic to most complex clauses. Analyzing a corpus of night dream stories, it has also been found that the ratio between boundaries with and without pauses varies appreciably from one story to another.

PATTERNS OF EMOTIONAL REACTIONS IN COMMUNICATION: PROBLEMS OF CORPORA STUDIES AND APPLICATION TO COMPUTER AGENTS

Kotov A. A. (kotov@harpia.ru), Russian State University for the Humanities

We study cognitive architecture of computer agents, simulating emotional speech behavior, and changing their mood in time. Basing on a multimodal corpus (records of university exams) we study sequences of contrastive emotional reactions and the possibility to apply the sequences to computer agents.

CITIZEN-INSTITUTION NON-MEDIATED DIALOGUE: THE RUSSIAN DIRECT LINE CASE

Cotta Ramusino P. (paola.cottaramusino@unimi.it), University of Milan, Italy

This paper analyses a specific kind of institutional discourse: Russian Direct Line. It aims to give account of interactional strategies used by subordinate participants of the given interaction. It tries to investigate how "naïf" interviewers, who are not familiar with strategies regulating a neutral or "neutralistic" position, manage to avoid possible consequences of their own speech acts, by using pragmatic and metapragmatic acts, basically aimed at downgrading.

THE NONVERBAL BEHAVIOR OF PEOPLE OF DIFFERENT CULTURES IN A DIALOG I: FINNISH AND RUSSIAN GESTURE SYSTEMS

Kreydlin G. E. (gekr@iitp.ru), Russian State University for Humanities

The paper presents some reflections of the so-called exterior observer about Finnish nonverbal semiotic culture, some nonverbal signs and models of Finnish dialog behavior. Corresponding Russian nonverbal data are given for comparison.

ON THE SEMANTIC CLASSIFICATION PROGRAM ProSeCa: THEORETICAL AND PRACTICAL ASPECTS

Kretov A. A. (a_a_kretov@rambler.ru), Voronezh State University, Russia
Rafaeva A. V. (anna_raf@rambler.ru), Lomonosov Moscow State University, Russia

A modified version of E. Kuznetsova's definition-based semantic identification method is proposed. The main point of it is that lexical semantics is concentrated in the most common nouns. A computer program of semantic classification is described. Perspectives of using and developing the program are outlined.

"QUASI-CORPUS" INVESTIGATION OF LEXICAL PRODUCTIVITY OF NON-TRIVIAL BASIC DIATHESES OF RUSSIAN WITH SPECIAL REGARD TO S. I. OZHEGOV'S DICTIONARY OF RUSSIAN

Krylov Sergey A. (krylov-58@mail.ru), Institute of Oriental Studies of Russian Academy of Sciences, Moscow;
Institute of System Analysis of Russian Academy of Sciences, Moscow

"Quasi-corpus" linguistics allows the investigation of both primary and secondary information sources (like grammars and dictionaries). The paper studies the statistics of grammatical data (on government patterns, transitivity, impersonality etc.) in the text of S. I. Ozhegov's "Dictionary of Russian" (1989).

THE ADJECTIVES WITH MEANING OF HIGH AND LOW TEMPERATURE AND LINGUISTIC ESTIMATION OF TEMPERATURE

Krylova T. V., V. V. Vinogradov Russian Language Institute, Russian Academy of Sciences

In this article the adjectives холодный, прохладный, горячий, жаркий, теплый are considered. In the first part we analyze their division to groups. In the second part we consider their combinations with adverbs of degree. We advance the hypothesis that many differences in using of temperature adjectives are caused by difference in linguistic estimation of high and low temperature. In conclusion the same idea is illustrated by the material of verb with meaning of temperature.

SEMIAUTOMATIC MARKING TOOLS FOR THE RUSSIAN MULTIMEDIA CORPUS (MURCO)

Kudinov M. S. (peshka1@mail.ru), MSU, **Grishina E. A.** (rudi2007@yandex.ru), Institute of Russian Language, Moscow

The paper describes two workbenches for corpus markers: a speech act marker's workbench (Marker) and a gesture marker's workbench (GesturesMarker). These programs allow the annotator to describe in quick and uniform manner Russian gesticulation and speech acts used in Russian spoken language.

MEANS FOR TUNING OF THE "SEMANTIX" LINGUISTIC PROCESSOR TO SUBJECT FIELDS

Kuznetsov I. P. (igor-kuz@mtu-net.ru), Institute of Informatics problems, Moscow, Russia
Efimov D. A. (d.efimov@synsys.ru), ZAO Synergetic Systems, Moscow, Russia

The linguistic processor "Semantix" for automatic formalization of natural language texts is presented It extracts data on user objects, their links and actions from texts. The processor uses special tools and methods for tuning to new subject fields. As an example the process of tuning for the text corpus about monuments is considered.

THE SEMANTIC DATABASE OF VERBAL ADJECTIVES: STRUCTURE AND TYPES OF INFORMATION

Kustova G. I. (galinak03@gmail.com), Moscow State Pedagogical University

The paper discusses the issues of elaboration of an electronic semantic dictionary (database) of Russian verbal adjectives (like vkhodnoj, lechebnyj, osvetitel'nyj etc). The topics considered include: a) the correlation between the verbal adjective and the verbal situation and the possibilities of expressing verbal arguments, e.g. stiral'naja mashina ('washing machine', instrument), vs. stiral'nyj poroshok ('washing powder', means); b) the correlation between the semantic class and the functional predicate of a noun and the semantic model of combinations like «verbal adjective + noun»; c) information types in the database; d) specification of semantic marking in the dictionary of the National Corpus of Russian language.

THE APPROACH TO CREATION OF MULTILINGUAL PARALLEL CORPORA OF WEB PUBLICATIONS

Lande D. V. (dwl@visti.net), **Zhigalo V. V.** (vladlen@visti.net), ElVist iInformation centre, Kiev, Ukraine

An algorithm of creating bilingual parallel corpora of documents from web publications is described. The algorithm uses frequency morphological dictionaries and empirical statistical properties of texts. An approach of homonymy resolution by means of statistical approach is presented, which allows choosing the most frequent normal forms. The algorithm has been developed as a software complex and integrated into the InfoStream system of content monitoring. As a result of algorithm operation aimed to determine basic word forms, a bilingual parallel corpus of electronic texts from web publications that contains more than 450 000 pairs of documents.

AN EDITOR OF AUGMENTED TRANSITION NETWORKS WITH A GRAPHICAL USER INTERFACE

Lebedev A. S. (andremoniy@gmail.com), Moscow State Institute of Electronics and Mathematics

The problem of semantic search is considered on an example of search for abstracts. An approach to the creation of a linguistic processor using augmented transition networks, inserted graphs, and arrangement of objects based on their descriptive part is proposed.

THE PROBLEM OF THE «Ë»-HOMOGRAPHS RESOLUTION IN TEXT-TO-SPEECH SYNTHESIS

Lobanov B. M. (lobanov@newman.bas-net.by),
United Institute of Informatics Problems, National Academy of Science of the Republic of Belarus

The problem of adequate ambiguity resolution in text-to-speech synthesis, for a special case of graphic homonymy related to the letter Ë is considered. Statistical characteristics of homographic pairs including Ë homographs and distributions among the frequent pairs of such homographs are investigated. The methods of resolution for the highly frequent homographic pair «ВСЁ» and «ВСЕ» are discussed.

SUMMARIZATION OF NEWS CLUSTERS BASED ON THEMATIC REPRESENTATION

Loukachevitch N. V. (louk@mail.cir.ru), **Dobrov B. V.** (dobroff@mail.cir.ru), Research Computer Center of M. V. Lomonosov Moscow State University (MSU NIVC); NCO Center for Information Research

The paper describes a technology of multi-document summarization, based on news cluster topical structure, lexical cohesion modelling and thesaurus descriptions of lexical senses. Lexical knowledge helps to improve cohesion and recall of a summary and reduce repetitions.

RUSSIAN FRAMENET: TOWARDS A CORPUS-BASED DICTIONARY OF CONSTRUCTIONS

Lashevskaya O. (olesar@mail.ru), **Kuznetsova Ju.** (julia.kuznetsova@uit.no), University of Tromsø (Norway)

The paper presents our basic approach to creating a FrameNet-oriented resource for Russian language, which involves extracting sampling from the Russian National Corpus and adding a layer of semantic and syntactic annotation. We discuss aims and methods of the project and give several examples of argument labeling in the dictionary and in the companion corpus.

NAMES OF BODY PARTS FROM THE VIEWPOINT OF TOPOLOGY

Makhova A. A. (discourse@yandex.ru), **Lyashevskaya O. N.** (olesar@mail.ru), **Desyatova A. V.** (patine@gmail.com)

The paper describes Russian names of body parts through the notion of topological type as introduced by L. Talmy. The corpus analysis of collocation with adjectives of shape and dimension makes it possible to define a number of topological types of body parts, such as juts, rods etc. and identify some peculiarities of their spatial perception.

AUTOMATIC ANALYSIS OF TERMINOLOGY IN THE RUSSIAN TEXT CORPUS ON CORPUS LINGUISTICS

Mitrofanova O. A. (alkonost-om@yandex.ru),Saint Petersburg State University, Russia
Zakharov V. P. (vz1311@yandex.ru), Saint Petersburg State University; Institute of Linguistic Studies, Russian Academy of Science

The paper presents the results of semi-automatic analysis of terminology in the Russian text corpus on Corpus Linguistics. Special attention is given to extraction of one-word and multi-word terms as well as to the use of lexical-grammatical patterns in the description of term structure and contexts of use.

AN EXPERIENCE OF CREATION OF THE NATIONAL CORPUS OF DAGESTAN LANGUAGES

Mutalov R. O. (mutalovr@mail.ru), Dagestan State University, Makhachkala

Problems and prospects of national corpora of six literary languages of Dagestan, created in the Dagestan State University, are considered. Special attention is given to the creation of a system of automatic markup of texts and digitization of printed texts.

COREFERENCE ANNOTATION IN PRAGUE DEPENDENCY TREEBANK

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz), Charles University, Prague, Czech Republic

The paper presents the pattern for annotating coreferential relations on the PTD corpus. Three levels of annotation are discussed: annotating grammatical coreference (the antecedent is calculated according to the grammar rules of a given language); annotating textual pronominal coreference; an extended pattern for annotating nominal textual coreference and associative anaphora. The first two (grammatical coreference and pronominal coreference) have been annotated on the whole PDT corpus, whereas the nominal coreference and assosiative anaphora are currently in the focus of the author's research. Certain complicated cases are going to be discussed and first results of the research presented.

SEGMENTATION OF ORAL NARRATIVE DISCOURSE AND ILLUSTRATIVE GESTURES: VISUAL CLUES AS SEGMENT MARKERS

Nikolaeva Y. V. (lis_julia@list.ru), Lomonosov Moscow State University, Russia

The paper is devoted to the interrelations between speech accompanying gestures and the discourse structure. The main aim was to find out how different characteristics of illustrative gestures mark discourse segment boundaries.

MODELS AND METHODS OF PUNCTUATION USE IN RUSSIAN LANGUAGE SYNTAX PARSING

Okatiev V. V. (oka@dictum.ru), **Erekhinskaya T. N.** (te@dictum.ru), **Skatov D. S.** (ds@dictum.ru),
DICTUM Ltd., Nizhny Novgorod, Russia

The paper describes functional ambiguity of punctuation marks in the Russian language. A formal model of isolations and series of coordination members is presented. Mathematical target setting for punctuation use in syntax parsing and the algorithm for this task are suggested.

TRANSLATION OF GERMAN PARTICLE *DOCH* USED IN STATEMENTS INTO RUSSIAN (IN STATEMENTS): *VED'*, *ŽE*, *VSE ŽE* OR *VSE-TAKI*?

Orlova S. V. (svetlachok-star@yandex.ru), Lomonosov Moscow State University, Russia

The paper is devoted to the comparative analysis of the semantics of the German particle DOCH in statements and its translation equivalents taken from German-Russian dictionaries — the Russian particles VED', ŽE, VSE ŽE and VSE-TAKI.

ETYMOLOGICAL DICTIONARY: LEXICOGRAPHIC STRUCTURE AND REPRESENTATION IN DIGITAL ENVIRONMENT

Ostapova I. V. (iros@zeos.net), Ukrainian Linguo-Information Fond, National Academy of Sciences of Ukraine

A technology for building an instrumental system for supporting the dictionary in digital environment was developed. The technology is based on a formal model of lexicographic system of etymological dictionaries. The main focus is given to mechanisms of language indexation.

POSSESSIVES AND MANNER OF ACTION NOUNS: CORPUS BASED EXPLORATION

Paducheva E. V. (elena708@gmail.com), Institute of Scientific and Technical Information (VINITI), Russian Academy of Sciences

Possessives (i.e. possessive pronouns and adjectives) resemble the genitive, but possessive Subject co-occurs with a genitive Object in the context of a verbal noun (мейерхольдовская постановка Ревизора), while genitive Subjects are not compatible with genitive Objects. Possessive-genitive diathesis serves as a diagnostics for NOUNS OF MANNER.

DERIVATIONAL PATTERNS AND SYNTACTIC POSITIONS OF DEVERBAL NOMINALS (ON CORPUS DATA)

Pazelskaya A. G.(anna_pz@abbyy.com), ABBYY Software

This paper is a part of general study of differences in behaviour in Russian deverbal nominals derived via various patterns. The investigation is done on the basis of corpus data, mostly obtained from the Russian National Corpus. We study preferences of nominals ascending to the three most productive derivational patterns with respect to the syntactic position of the resulting nominal in a sentence.

PROSODY OF THE GERMAN VOCATIVE NPs IN CONTRAST TO THE RUSSIAN ONES

Palko M. L. (m_palko@mail.ru) Institute of Linguistics (Russian Academy of Sciences)

The prosody of the German vocative NPs is discussed as contrasted to the prosody of the Russian vocatives. The analysis shows that the German vocatives do not allow for prestressed articulations that are highly characteristic of the Russian vocatives used in unofficial and close contacts between the hearer and the listener, cf. MOLODOJ chelovek! with a wordform molodoj to be accented. The non-vocative NPs also demonstrate more restrictions in prestressed patterns formation, which seems to be the typological parameter of German and of most West European languages.

NONVERBAL COMMUNICATIVE ACT OF CONSOLATION: MATERIALS FOR A DICTIONARY OF NONVERBAL COMMUNICATIVE ACTS

Pereverzeva S. I. (P_Sveta@hotmail.com), Russian State University for the Humanities, Russia

The paper discusses some issues regarding the dictionaries of Russian speech acts and Russian nonverbal acts. I provide a preliminary draft of a dictionary entry "consolation" as an example of lexicographical description of nonverbal acts.

THE ROLE OF DISCOURSE MARKERS IN LOCAL DISCOURSE STRUCTURE: A CORPUS STUDY

Podlesskaya V. I. (podlesskaya@ocrus.ru), Russian State University for the Humanities
Kibrik A. A. (aakibrik@gmail.com), Institute of Linguistics RAS

Based on a corpus of spoken narratives, the study shows how discourse markers can be differently integrated into local discourse structure: some can be used as a separate "minimal discourse unit", while others are always integrated into a bigger unit with a propositional meaning. The two discourse markers most frequent in the corpus, VOT and NU, are compared and VOT is shown to be less integrated into prosodic, linear and hierarchic structure than NU.

LOMONOSOV CONCORDANCE — CONCEPT AND IMPLEMENTATION

Polyakov A. E. (pollex@mail.ru), NTC «Informregistr»
Bergelson M. B. (mirabergelson@gmail.com), Lomonosov Moscow State University, Russia
Pilshckov I. A. (pilshch@yandex.ru), IMK of Lomonosov Moscow State University

This paper qualifies the concepts and terminology relevant to the development of comprehensive digital Concordance to the texts of Lomonosov, and discusses the practical decisions which are necessary for the implementation of this lexicographical product. The concordance is based on the corpus of author's texts supplied with structural, philological and grammatical markup. We describe the technology we use to build the corpus and the concordance, the principles of corpus markup, and the structure of concordance vocabulary entries, as well as its application to linguistic research.

SYNTACTICALLY THE INVARIANT METHOD OF IDENTIFICATION OF SEMANTICS OF THE INFORMATION

Potapov M. V. (potapov_mv@rgrta.ryazan.ru), Ryazan State Radio Engineering University

In the report the description of practically approved method of an estimation of the semantic maintenance of the information streams based on statistic — a linguistic way of primary processing of the bit information and approaches of the theory of recognition of images contains at the analysis of multivariate attributes

UNSUPERVISED PARSING

Potemkin, S. (potemkin@philol.msu.ru), Philological Faculty, Moscow State University, Russia

A statistical approach to parsing of raw text is described. The parsing algorithm builds a projective dependency tree in quadratic time after training on an unannotated corpus.

MULTI-TIER MARKUP OF SPEECH CORPUS FOR HYBRID RUSSIAN TTS SYSTEM "VITALVOICE"

Prodan A. I. (prodan@speechpro.com), **Korolkov E. A.** (korolkov@speechpro.com), **Oparin I. V.** (ilya@speechpro.com),
Talanov A. O. (andre@speechpro.com), Speech Technology Center, Russia

The paper deals with the features of a system for multi-level markup of speech corpora. These corpora are used for the hybrid Russian TTS system "VitalVoice" developed at Speech Technology Center (STC). VitalVoice is basically a Unit Selection TTS system complemented with triphone inventory. The basic advantage of this approach is that it allows getting speech units from the speech corpus in a quick and efficient way. The database consists of interrelated levels of markup (phrases, intonation models, words, syllables, etc.). The levels of markup, their use in the TTS system and automatic markup checking are described in detail.

SEMANTIC-DERIVATIONAL MODELS OF POLYSEMOUS ADJECTIVES: METAPHOR, METONYMY AND THEIR INTERACTION

Rakhilina E. V. (rakhilina@gmail.com), Institute Of Russian Language, RAS
Karpova O. S. (o_k_inbox.ru), Russian State University for Humanities
Reznikova T. I. (tanja.reznikova@gmail.com), All-Russian Institute of Scientific and Technical Information, RAS

The paper reports on a project intended to provide a corpus-based description of semantic-derivational models for Russian adjectives. The research deals with high-frequency adjectives in the attributive use denoting the quality of a person or thing. We discuss basic metonymical and metaphorical patterns and analyze several non-regular shifts.

THE SO-CALLED: SEMANTIC ANALYSIS OF PARENTHETICAL METALINGUISTIC PHRASES

Rozina R. I. (rarozina@yandex.ru), Russian Language Institute n.a. V. V. Vinogradov, Russian Academy of Sciences

The paper is concerned with meaning and textual functions of a group of parenthetical phrases expressing the speaker's attitude to the manner of speech. It is argued that their function is to ensure the transition in the text between different styles, the relation between which changes in the course of time, and that the meaning of these phrases is extended in the way that might be regular.

AUTHORSHIP IDENTIFICATION WITH SUPPORT VECTOR MACHINE IN CASE OF TWO POSSIBLE ALTERNATIVES

Romanov Aleksandr S. (ras@ms.tusur.ru), **Mescheriakov Roman V.** (mrv@keva.tusur.ru),
Tomsk state university of control system and radioelectronics

Authorship identification problem is viewed as a classification task. The importance of resolving the binary authorship classification problem for authorship identification is justified. Description and results of authorship identification experiment with support vector machine in the case of two possible alternatives are given.

STRATEGIES OF DELIMITATION OF SYNTACTIC UNITS IN SPONTANEOUS SPEECH

Ryko A. I. (aryko@mail.ru), **Stepanova S. B.** (stsvet_2002@mail.ru), Saint Petersburg State University, Russia

The paper discusses methods of dividing spontaneous speech into syntactic units using the Corpus of Spoken Russian. We analyze individual strategies of experts who took part in the experiment, and examine connections between the boundaries of sentences and their final intonation.

ON ENCYCLOPAEDIC DATA IN AN APPLIED SEMANTIC DICTIONARY

Semenova S. Yu. (Sonya_sem@mail.ru), INION RAS, Russia

Inclusion of information on ontological realities into a semantic dictionary, which is a trend in modern lexicography, corresponds to ideas of cognitive science with its focus on the wholeness of the information perception process. The paper is concerned with the encyclopaedic data within the NLP-aimed semantic dictionary that has the rigid formats for lexical data representation. Encyclopaedic functions in the RUSLAN machine semantic dictionary are considered. Some ways of loading and enhancement of the functions are discussed. A number of words and lexical classes relevant to certain types of encyclopaedic data are considered.

AN ONTOLOGY-BASED APPROACH TO FACT EXTRACTION

Sidorova E. A. (lena@iis.nsk.su), **Kononenko I. S.**(irina_k@cn.ru), A. P. Ershov Institute of Informatics Systems, Russia

An approach is proposed to develop fact extraction technology applicable in information systems of various kinds. The approach makes use of the knowledge base including domain ontology, domain vocabulary, model for text segmentation, and fact extraction schemes that relate vocabulary items and lexical-syntactic constructions to ontology entities.

MODELS AND METHODS FOR THE ANALYSIS OF HIERARCHICALLY STRUCTURED TEXTS

Skatov D. S. (ds@dictum.ru), **Erekhinskaya T. N.** (te@dictum.ru), **Okatiev V. V.** (oka@dictum.ru),
DICTUM Ltd., Nizhny Novgorod, Russia

The analysis of hierarchically structured texts (laws, standards etc.) is discussed. An overview of developments in the domain are given. The developed models and methods for the analysis of hierarchically structured texts are described.

EXPERIENCE OF SYSTEMATIZING KNOWLEDGE AND INTERNET RESOURCES FOR A KNOWLEDGE PORTAL ON COMPUTATIONAL LINGUISTICS

Sokolova E. G. (minegot@rambler.ru), Russian State University for Humanities, Moscow
Kononenko I. S. (irina_k@cn.ru), **Zagorulko Yu. A.** (zagor@iis.nsk.su), A. P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk

The paper describes an experience of systematizing knowledge and internet resources for a knowledge portal on computational linguistics. A composition and structure of objects of the portal, place of the portal among other catalogues on computational linguistics, an experience of development of bilingual vocabulary of terms on computational linguistics with using procedures of automatic extraction of terms from text are considered.

THE USE OF LEXICO-GRAMMATICAL DATABASES IN THE RUSSIAN DIALECTAL LEXICOGRAPHY

Aleksandra V. Ter-Avanesova (teravan@mail.ru), Institute of Russian Language of Russian Academy of Sciences, Moscow
Sergej A. Krylov (krylov-58@mail.ru), Institute of Oriental Studies of Russian Academy of Sciences, Moscow;
Institute of System Analysis of Russian Academy of Sciences, Moscow

The lexico-grammatical database (LGDB) for Russian folk dialects with two [o]-like phonemes that was built with the help of StarLing informational system is significantly enriched. It includes now the data on a Middle Russian dialect of the village Pustosha (Shatura district, Moscow region, and a LGDB for Vologda suburban dialects, including about 30 thousand word-forms that represent about 4500 lexemes. The kernel dialectal corpus (KDC) contains texts with partial lexico-grammatical tagging.

LEXICAL FUNCTIONS AND SEARCH ENGINE OPTIMIZATION (BASED ON WORDS WITH NUMERIC VALUES)

Timoshenko Svetlana (timoshenko@iitp.ru), **Leonid Cinman** (cinman@iitp.ru), Institute for Information Transmission Problems, Russian Academy of Sciences

To provide more precise web search we have developed a special option in the ETAP-3 multifunctional NLP environment. The search query consisting of two or three words has been supplemented with the values of certain lexical functions to generate an incomplete sentence which lacks only the numeral information. We expect that it may help in searching numeral data like “The height of the Pisa tower”. The results of the experiment show that the search precision index in this domain of knowledge increases by 24 % on the average.

APPLYING LINGUISTIC SEMANTICS AND MACHINE LEARNING METHODS TO SEARCH PRECISION IMPROVEMENT IN SEARCH ENGINE “EXACTUS”

Tikhomirov I. A. (matandra@isa.ru), **Smirnov I. V.** (ivs@isa.ru) Institute for Systems Analysis of RAS, Moscow

The paper considers problems of using linguistic semantics and machine learning methods in the Exactus search engine. An experimental evaluation of search quality showed that these methods improve search precision and recall. Prospects of applying linguistic semantics and machine learning methods in search engines are discussed.

ON THE PROBLEM OF VARIABILITY OF IMPERATIVE ASPECTUAL FORMS

Trub V. M. (trub44@ukr.net)

The paper deals with the correlation between different aspectual forms of imperative verbs. We believe that one of the aims of semantic interpretation of inducements conveyed by different aspectual forms consists in the explication of semantic differences between them and the explanation of causes of irregularities reflected in the use of a form opposed to the default one.

KAK BY (lit. ‘as if, like’) AND KONKRETNO (lit. ‘specifically’)

Uryson E. V. (x-uryson@mtu-net.ru) Institute of Russian Language, Moscow

The semantics of Russian colloquial “parasitic” particles KAK BY (lit. ‘as if, like’) and KONKRETNO (lit. ‘specifically’) is described. The goal is to show that their emergence in the language is due to the lexical system of the language. KAK BY in its first meaning denotes similarity, and the words denoting similarity usually have a meaning denoting a set (a class). This is the way of “desemantization” of the conjunction KAK BY. The particle KONKRETNO develops its parasitic meaning by analogy with the word VOOBSHCHE (‘in general’); the cause is that some meanings of KONKRETNO are antonyms to some meanings of VOOBSHCHE.

MEANINGS OF THE PREPOSITIONS “PO” AND “K” IN RUSSIAN: ENCODING OF ADJUNCTS AND SEMANTIC ROLES

Usacheva M. N. (mashastroeva@gmail.com), Lomonosov Moscow State University, Russia

This work is devoted to the application of the spatial meaning description method (developed primarily for Dagestani languages but claimed to be typologically universal: see [Ganenkov 2002, 2005[, [Mazurova 2007]) to Russian prepositions “po” and “k”.

PROCESSING INITIAL-STRESS AND NON-INITIAL-STRESS WORDS IN SPOKEN-WORD RECOGNITION IN RUSSIAN

Fedorova O. V. (olga.fedorova@msu.ru), **Shavrygina A. S.** (shavrygina@gmail.com), Lomonosov Moscow State University, Russia

The data of an experimental investigation of spoken-word recognition in Russian are presented. Two experiments showed that word recognition and word recall are faster and better in initial-stress word than in non-initial-stress words. The results support the metrical segmentation theory.

EASTERN ARMENIAN NATIONAL CORPUS www.eanc.net

Khurshudian V. G. (vk@corpustechnologies.com), **Daniel M. A.** (misha.daniel@gmail.com),
Levonian D. V. (dl@renovacapital.com), **Plungian V. A.** (plungian@gmail.com), **Polyakov A. E.** (pollex@mail.ru),
Rubakov S. V. (rubakov@gmail.com), Corpus Technologies

Eastern Armenian National Corpus (EANC) is a comprehensive linguistic database of annotated texts in Eastern Armenian from the mid 19th century to the present. The EANC contains about 110 million tokens and is enhanced with a powerful search engine. EANC is available at www.eanc.net.

ZERO CATEGORIES IN UNIVERSAL GRAMMAR

Zimmerling A. V. (meinmat@yahoo.com) Moscow State University for the Humanities, MGGU

The paper discusses the status of zero categories in general syntax. The taxon 'pro' is not sufficient for tagging all covert pronouns in finite clauses. Moreover, the notion of 'discourse pro-drop languages' is not a valid tool in syntactic typology. Discourse-linked dropping of anaphoric pronouns, coreferent deletion and constraint on overt realization of pro-forms are different syntactic operations. More specifically, I am challenging some points in Holmberg's analysis of Finnish pro and claiming that 1-2 person pro-forms regularly display features different from 3rd person pronominal zeros. Finally, I am discussing the status of 'Mel'čuk's zeros', e.g. theta-role sensitive zero lexemes and proving for that theta-role sensitive zero pronouns with an Agentive value and theta-role neutral pro-forms may coexist in one and the same language.

STATISTICAL ANALYSIS AND CONTEXTUAL RULES OF HOMOGRAPH DISAMBIGUATION ON TEXT-TO-SPEECH SYNTHESIS

Tsirulnik Liliya I. (L.tsirulnik@newman.bas-net.by), United Institute of Informatics Problems, National Academy of Sciences of Belarus
Svetlana G. Barbuk (sviatos@tut.by), Minsk State Linguistic University, Belarus
Boris M. Lobanov (Lobanov@newman.bas-net.by), United Institute of Informatics Problems, National Academy of Sciences of Belarus

The rules of accent position location in the homographs based on the results of contextual and statistical analysis of scientific and artistic text corpora are described. The implementation of the developed rules in Russian TTS synthesis system "MultiPhone" increase the degree of adequacy of sense understanding of synthesized speech.

AN ALGORITHM OF LINK SPAM DETECTION

Sharapov R. V. (info@vanta.ru), **Sharapova E. V.** (goldenstuff@mail.ru), Murom Institute of Vladimir State University

Approaches to detecting spam links on the basis of the analysis of page content are considered. We focus on the detection of advertisement (paid) links. Features of paid links are analyzed. The algorithm of detecting a spam link is given.

COMMUNICATIVES AND METHODS OF ITS DESCRIPTION

Sharonov I. A. (Igor_sharonov.@mail.ru), Russian State University for the Humanities

Short dialogical utterances with fixed and vague grammatical structure are analyzed. We call these utterances "communicatives" and focus on the main principles underlying the classification of such language forms and ways of their pragmatic and conversational analysis. To describe the functioning of a communicative in conversation we need to clarify their semantic, formal and discursive characteristics, which include: — communicative intention or emotional state; — what kind of speech act — direct or indirect — a communicative represent; — the source form of the communicative and the mode of its transposition into communicative; — the discursive boundaries with adjacent utterances; — standard intonation patterns and other phonetic characteristics of the communicative in speech.

VARIATION, CONTINUATION, AND SERIALITY OF JOKES: PROBLEMS OF DATABASE CONSTRUCTION

Shmeleva E. Y. (eshkind@mail.ru), **Shmelev A. D.** (smelev.alexei@gmail.com), Institute of Russian Language, Moscow

The paper deals with different kinds of joke variation and intertextual relations between jokes. We discuss such phenomena as realization of a joke, versions of a joke, continuation of a joke, modification of the original joke, addition to the original joke, series of jokes, joke cycle.

SYNTACTIC AMBIGUITY RESOLUTION: PRIMING AND SELF-PRIMING EFFECTS

Yudina M. (dietiefe@yandex.ru) (ABBYY, Moscow State University)
Fedorova O. (olga.fedorova@msu.ru) (Moscow State University)

The report is devoted to the first experimental research on the influence of syntactic priming on syntactic ambiguity resolution of relative clauses in Russian. Within the frame of syntactic priming we can see two effects: the syntactic priming itself and self-priming (persistent preference of subject's own syntactic strategy).

BEST RECOGNIZABLE WORDS UNDER DIFFERENT EXPERIMENTAL SETTINGS

Iagounova E. V. (iagounova_elena@mail.ru), St.Petersburg State University

Basic features of the sets formed by the words, best recognizable under white-noise masking and within meaningless text fragments have been analyzed. It is observed that the sets are crucially dependent on such broad text parameters as professional text vs. fiction and dynamic vs. static text.

STRUCTURING OF ATTRIBUTIVE WORD MEANINGS IN RUSSNET THESAURUS (IN RUSSIAN ADJECTIVES OF PERCEPTION)

Yavorskaya M. V. (yav.mas@gmail.com), **Azarova I. V.** (ivazarova@gmail.com), Saint Petersburg State University, Russia

Adjectives with perceptional meanings are described. We focus on the problem of attributive meanings structuring for computer thesaurus RussNet. 178 attributive word-meaning pairs are marked up in the random samples of corpus contexts. Attributes for different spheres of perception are compared.

RUSSIAN VOCATIVES: LEXICON AND CONSTRUCTIONS

Yanko T. E. (tanya_yanko@list.ru), Institute of Linguistics (Russian Academy of sciences)

According to Zwicky, semantically parallel NPs often have distinct vocative properties. Whether a given NP can be used as a call or an address is a dictionary information. In this paper a variety of specific vocative strategies and vocative constructions that change a vocative potential of lexical items is analyzed.

DEVELOPMENT AND IMPLEMENTATION OF MULTILINGUAL OBJECT TYPE TOPONYM-REFERENCED TEXT CORPORA FOR OPTIMIZING AUTOMATIC IMAGE DESCRIPTION GENERATION

Gornostay T. (tatjana.gornostaja@tilde.lv), Tilde, Riga (www.tilde.com),
Aker A. (a.aker@dcs.shef.ac.uk), Department of Computer Science, University of Sheffield

The fast growing amount of images available on the web has motivated development of automatic approaches for image description generation. Using multi-document summarization for this task has been proposed recently. This paper describes a method for developing and implementing object type toponym-referenced text corpora in the context of optimizing the multi-document summarization for generating toponym-referenced descriptions of images. Object type corpora are developed for four different languages: English, German, Italian and Latvian.

TRANSCRBING, STRUCTURING AND TEMPORAL ANALYSIS OF FLUENT SPEECH CORPUS FOR A UNIT SELECTION TTS SYSTEM FOR ESTONIAN

Mihkla M. (meelis@eki.ee), **Kiissel I.** (indrek@eki.ee), **Nurk T.** (tonis@eki.ee), **Piits L.** (liisi@eki.ee), Institute of the Estonian Language

The paper reports the development of a speech corpus for Estonian text-to-speech synthesis based on unit selection. The process of transforming an orthographic Estonian text into a pronounced text, requiring the consideration of quantity, palatalization and other essential features of an Estonian pronounced text, is described. In order to optimize the unit selection algorithm and to guarantee the necessary quality of the synthetic speech the whole speech database is represented as a phonological tree. We present the evidence that the collocational strength shortens the duration of words and that contextual predictability is a significant feature to be considered in developing models of word duration.

THE DYNAMICS OF ADJECTIVE MEANING

Partee Barbara H. (partee@linguist.umass.edu), University of Massachusetts, Amherst, MA, USA; Moscow State University

Meaning and context interact dynamically; how can one account for context-dependence without abandoning compositionality? We illustrate with the semantics of different kinds of adjectives. We show how compositional semantics sheds light on word meaning, and how compositional semantics, lexical semantics, and context all interact.

ONTOLOGICAL SEMANTICS AND ABDUCTION: PARSING ELLIPSIS

Petrenko M. (mpetrenk@gmail.com), Princess Dashkova Moscow Humanities Institute, Russia

New avenues for modeling abductive reasoning within the framework of Ontological Semantics are explored. Specifically, the rich knowledge resources and dynamic parsing module of Ontological Semantics allow processing elliptic input with a set of inference rules, which establish on the one hand, dependencies between verbalized and non-verbalized case-roles across clauses, and on the other hand, dependencies between scalar attribute values and specific event classes. Examples are provided to illustrate each case.

Авторский указатель

Азарова И. В. 559
Алиев Р. М. 1
Антонов А. В. 13
Апресян В. Ю. 6
Архипов А. В. 181
Асиновский А. С. 38
Ахметова М. В. 25
Баглей С. Г. 13
Баранов А. Н. 19
Барбук С. Г. 530
Беликов В. И. 25
Бергельсон М. Б. 396
Богатырёв М. Ю. 31
Богданова Н. В. 38
Богуславский И. М. 91
Большаков И. А. 45
Бугаков О. В. 51
Васильев В. Г. 57
Воскресенский А. Л. 64
Гольдин В. Е. 69
Городечный П. П. 116
Григорьян Е. Л. 75
Гришина Е. А. 80, 248
Гуленко И. Е. 64
Даниэль М. А. 509
Десятова А. В. 313
Диконов В. Г. 91
Добров Б. В. 299
Добровольский Д. О. 97
Ерехинская Т. Н. 346, 458
Ермаков А. Е. 102
Ефимов Д. А. 262
Жигало В. В. 278
Загорулько Ю. А. 465
Зализняк Анна А. 107
Занегина Н. Н. 112
Захарова И. В. 116
Захаров В. П. 321
Захаров Л. М. 181
Зобнин А. И. 120
Иомдин Б. Л. 127
Иомдин Л. Л. 136
Карпова О. С. 420
Кибрик А. А. 143, 390
Кибрик А. Е. 149
Клышинский Э. С. 165
Кобзарева Т. Ю. 170
Кобозева И. М. 176
Кодзасов С. В. 143, 181
Кожунова О. С. 185
Козеренко А. Д. 192
Комарова А. Д. 197
Кононенко И. С. 451, 465
Корольков Е. А. 415
Коротаев Н. А. 204
Котов А. А. 211
Котта Рамузино П. 219
Крейдлин Г. Е. 224
Кретов А. А. 230
Кривнова О. Ф. 181
Крылова Т. В. 243
Крылов С. А. 236, 471
Кудинов М. С. 248
Кузнецова Ю. Л. 306
Кузнецов И. П. 262
Кустова Г. И. 271
Ландэ Д. В. 278
Лебедев А. С. 284
Левонтина И. Б. 97
Левонян Д. В. 509
Лобанов Б. М. 136, 291, 530
Лукашевич Н. В. 299
Ляшевская О. Н. 306, 313
Мартьянов А. О. 69
Махова А. А. 313
Мешков В. С. 13
Мещеряков Р. В. 432
Митрофанова О. А. 321
Муталов Р. О. 329
Недолужко А. 332
Николаева Ю. В. 340
Окатьев В. В. 346, 458
Опарин И. В. 415
Орлова С. В. 352
Остапова И. В. 359
Падучева Е. В. 365
Пазельская А. Г. 373
Палько М. Л. 379
Переверзева С. И. 384
Пильщиков И. А. 396
Плешко В. В. 102
Плунгян В. А. 509
Подлесская В. И. 390
Поляков А. Е. 396, 509
Потапов М. В. 405
Потемкин С. Б. 409
Продан А. И. 415
Рафаева А. В. 230
Рахилина Е. В. 420
Резникова Т. И. 420
Розина Р. И. 426
Романов А. С 432
Рубаков С. В. 509
Русакова М. В. 38
Рыко А. И. 38, 438
Сахарова А. В. 120
Сдобнова А. П. 69
Семенова С. Ю. 444

Сидорова Е. А. 451
Скатов Д. С. 346, 458
Смирнов И. В. 483
Соколова Е. Г. 465
Степанова С. Б. 38, 438
Суханов А. В. 13
Таланов А. О. 415
Тер-Аванесова А. В. 471
Тимошенко С. П. 476
Тихомиров И. А. 483
Труб В. М. 488
Тюхтин В. В. 31
Урысон Е. В. 493
Усачёва М. Н. 499
Фёдорова О. В. 504, 554
Хахалин Г. К. 64
Хейдоров И. Э. 1
Худякова М. В. 143
Хуршудян В. Г. 509
Циммерлинг А. В. 519
Цинман Л. Л. 476
Цирульник Л. И. 530
Шаврыгина А. С. 504
Шарапова Е. В. 537
Шарапов Р. В. 537
Шаронов И. А. 543
Шерстинова Т. Ю. 38
Шмелёв А. Д. 548
Шмелёва Е. Я. 548
Юдина М. В. 554
Яворская М. В. 559
Ягунова Е. В. 566
Янко Т. Е. 574
Янь Цзинбинь 1
Aker A. 580
Gornostay T. 580
Indrek Kiissel 588
Liisi Piits 588
Meelis Mihkla 588
Partee Barbara H. 593
Petrenko M. 598
Tõnis Nurk 588

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной Международной конференции «Диалог 2009»

Выпуск 8 (15)

Ответственный за выпуск **Левченкова И. А.**
Компьютерная вёрстка **Климентовский К. А.**

Издательский центр РГГУ
125993, Москва, Миусская пл., д. 6
Тел.: +7 (499) 973 42 00

Отпечатано с готового оригинал-макета
в типографии ООО «Издательско-полиграфический центр Маска»
117246, Москва, Научный пр-д, д. 20, стр. 9

Подписано в печать 12.05.2009 г.
Формат 60×84/8
Тираж 200 экз. Заказ № 213