Введение в электронные лингвистические ресурсы

Учебное пособие для студентов,
обучающихся по специальностям «Филология» и «Фундаментальная и теоретическая лингвистика»

Пособие подготовлено на кафедре теории, истории языка и прикладной лингвистики Института филологии и журналистики Саратовского государственного университета им. Н.Г. Чернышевского

Составители – *проф. В.Е. Гольдин и проф. О.Ю. Крючкова*

Саратов
2011

Содержание

Лекция 1. Лингвистические ресурсы

1. Понятие «лингвистические ресурсы».
2. Первичные и вторичные лингвистические данные.
3. Средства обработки речевого материала.
4. Машинный фонд русского языка.
5. Рекомендации и задания.
6. Литература.

1. Слово «ресурс» (чаще – во множественном числе) широко используется в современной русской речи для обозначения источников чего-либо, каких-либо запасов, к которым можно обратиться в случае необходимости. Говорят о *природных ресурсах, информационных, финансовых, о ресурсах живого организма, о ресурсах исчерпанных, неиспользованных, возобновляемых* и под. Все чаще используется и словосочетание «лингвистические ресурсы», особенно в применении к компьютерным средствам поддержки работы лингвиста. Именно в этом последнем значении, то есть как **о компьютерных средствах поддержки работы лингвиста,** говорится о лингвистических ресурсах в данной лекции.

Сегодня еще далеко не все лингвисты, особенно получившие филологическое образование до начала XXI в., постоянно обращаются к компьютерным ресурсам, ориентируются в них и эффективно их используют. Между тем компьютерных средств поддержки работы лингвиста создано много, и они весьма разнообразны. Это неспециализированные и специализированные текстовые редакторы; это рассчитанные на специалистов звуковые анализаторы, компьютерные программы получения конкордансов, различные лингвистические базы данных с соответствующими средствами управления этими базами, например электронные словари; это специальным образом организованные собрания текстов, компьютерные программы для обучения иностранным языкам и многое другое.

Никакая лекция не может вместить даже одни только основные сведения, касающиеся современных лингвистических ресурсов. Не решают такой задачи и предлагаемые здесь материалы. Их цель значительно скромнее: **способствовать начальной ориентации** в лингвистических ресурсах, послужить отправным пунктом для тех лингвистов, кто впервые вступает на путь их использования.

Типология лингвистических ресурсов может опираться на различные основания. Так, возможно построение типологии лингвистических ресурсов на основе учета особенностей их представления и функционирования в компьютерных системах: представлены, например, в системе лингвистические данные на естественном языке или в формализованном виде, рассчитаны они на лингвиста как конечного пользователя или имеют внутренний характер, то есть обслуживают функционирование самой системы, поддерживают ее работу, предоставляя системе необходимые для работы лингвистические сведения, как, например, формализованные словари в системах машинного перевода. Но для лингвистов существенно рассмотрение ресурсов именно в лингвистических аспектах. С собственно лингвистической точки зрения важно различать, с одной стороны, ресурсы лингвистических данных и, с другой стороны, – средства обработки лингвистического материала, а среди ресурсов лингвистических данных – ресурсы первичных данных и ресурсы вторичных данных (см. схему 1).

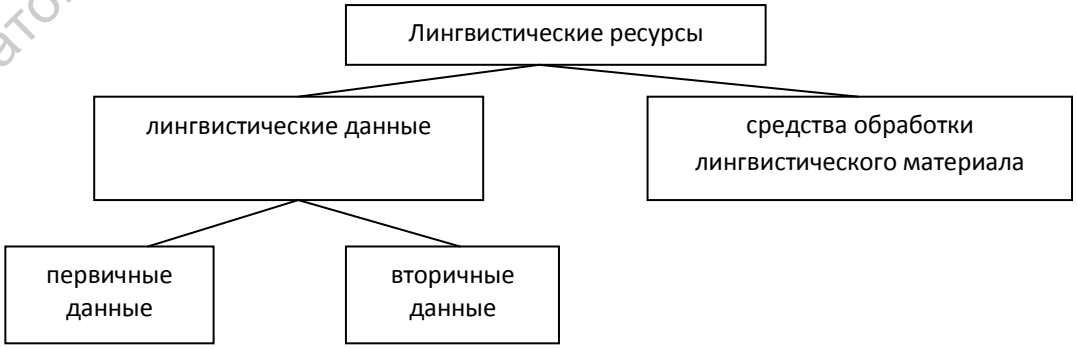


Схема 1. Основные типы лингвистических ресурсов

2. Первичные и вторичные лингвистические данные. Первичные лингвистические данные – это речь, в минимальной степени подвергнутая лингвистической интерпретации. Первичными лингвистическими данными можно, например, считать записи звучащей речи, представленные в аналоговой или в цифровой форме. Для речи письменной первичным является ее символьное представление, полностью сохраняющее состав и последовательность символов исходного текста (к письменной речи относим все тексты, которые создаются и функционируют в письменном виде: частные или официальные письма, тексты печатных СМИ, распорядительную документацию, государственные законы, тексты письменных художественных произведений и под.).

Ресурсами первичных лингвистических данных являются, например, собрания магнитофонных записей естественной диалектной речи, накапливаемые в научных центрах в составе фонотек, издаваемые в виде звуковых хрестоматий или публикуемые в Интернете. Для лингвиста, исследующего язык официальных документов, одним из ценных ресурсов первичных лингвистических данных может явиться, например, официальное периодическое издание "Собрание законодательства Российской Федерации", которое в электронном виде распространяется созданным в 1993 г. Научно-техническим центром правовой информации "Система" (www.systema.ru). «Собрание законодательства Российской Федерации" содержит федеральные конституционные законы, федеральные законы, акты палат Федерального Собрания, указы и распоряжения Президента Российской Федерации, постановления и распоряжения Правительства Российской Федерации, решения Конституционного Суда Российской Федерации о толковании Конституции Российской Федерации и др.

Образцом ресурса, представляющего в качестве первичных данных тексты прижизненных изданий пушкинских произведений, то есть без графических, символьных, орфографических и других замен, может служить соответствующий раздел Интернет-портала «Электронные публикации Института русской литературы (Пушкинского Дома) РАН» (http://lib.pushkinskiydom.ru) – см. Рис. 1.

Выберем на главной странице этого портала рубрику «Интернет-проекты. Пушкин. Прижизненные публикации» и в открывшемся окне – «Отдельные издания». Далее выделим, например, – «Руслан и Людмила (1820)». Читатель получает доступ к электронной копии

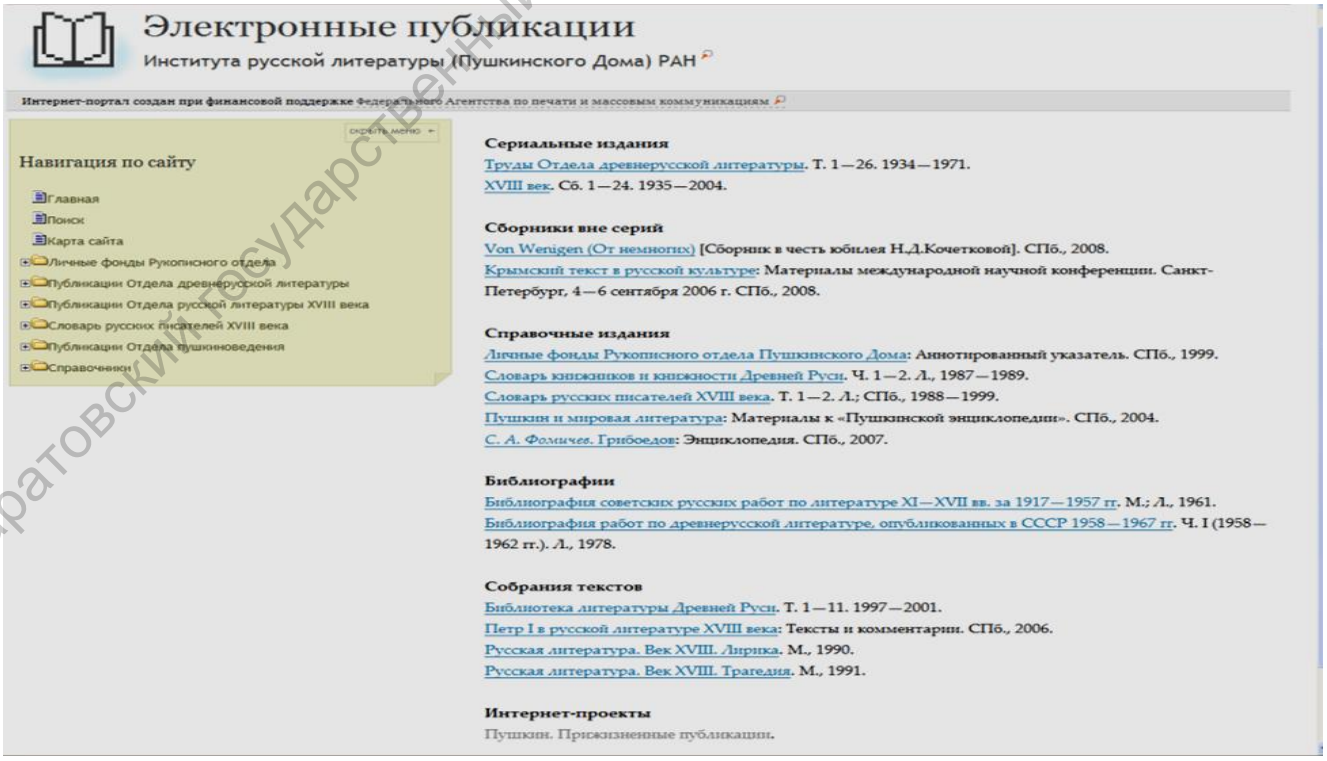
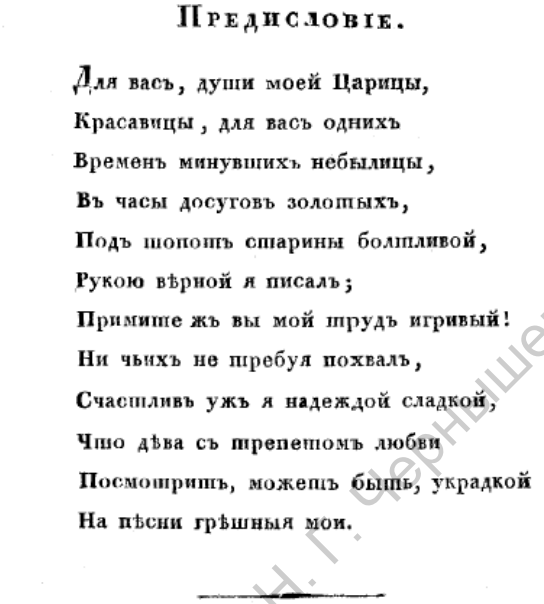


Рис. 1. Портал "Электронные публикации Института русской литературы (Пушкинского Дома) РАН"

издан
ия

1820 г. и может изучать все его страницы. Вот, например, как выглядела в публикации «Руслана и Людмилы» 1820 г. страница с «Предисловием» (см. Рис. 2).



Предисловіе.

Для васъ, души моей Царицы,
Красавицы, для васъ однихъ
Временъ минувшихъ небылицы,
Въ часы досуговъ золотыхъ,
Подъ шопотъ спарины болтливой,
Рукою вѣрной я писалъ;
Примите жъ вы мой трудъ игривый!
Ни чьихъ не требуя похвалъ,
Счастливъ ужъ я надеждой сладкой,
Что дѣва съ трепетомъ любви
Посмотритъ, можетъ быть, украдкой
На пѣсни грѣшныя мои.

Рисунок 2. Страница прижизненного издания поэмы "Руслан и Людмила" (1820)

Примером лингвистических данных вторичного типа могут служить символьные расшифровки звучащей речи, например тексты в фонетической транскрипции. Собрания таких транскрипций относятся к лингвистическим источникам вторичного типа, поскольку являются результатом оценивающей деятельности лингвиста: от его знаний, умений, от особенностей его слуха, от преследуемых целей, от избранной им системы знаков, от характера вспомогательных технических средств, применявшихся в процессе транскрибирования, зависят глубина транскрипции, ее точность, состав отражаемых в ней явлений. Известно, что не всякая фонетическая транскрипция передает, например, просодию, а применяемые исследователями системы записи просодических явлений весьма далеки от единства.

Существует немало оснований относить к ресурсам вторичных лингвистических данных и лингвистические словари (проблемы электронной лексикографии рассматриваются в лекции 4). Уже само выделение языковых единиц (морфем, например, или слов) совершается в процессе научной интерпретации, опирающейся на те или иные теоретические установки создателей словарей, и является результатом выбора определенных решений из ряда возможных. Если же обратиться к толкованиям значений слов, к стилистическим и нормативным квалификациям языковых единиц, то вторичный характер этих данных становится еще более очевидным. Среди словарей, по-видимому, только ассоциативные (в их современном виде) содержат первичные лингвистические данные: ответы испытуемых представлены в ассоциативных словарях в удобном для обозрения виде, но при этом в самих словарях никак не интерпретируются.

Имеется несколько основных разновидностей вторичных лингвистических данных. Это, во-первых, лингвистически обработанные тексты реальной коммуникации, во-вторых, – лингвистические базы данных (машиннообрабатываемые собрания лингвистических единиц, выделенных из речи), в-третьих, – научные тексты о языке и речи, к которым несколько условно можно отнести и тексты лингвистической метатеории, в том числе работы по истории языкознания.

Первичные данные имеют для лингвистов особую ценность, поскольку не навязывают пользователю решений, мнений, найденных другими специалистами, пусть даже наиболее квалифицированными, а предоставляют в распоряжение исследователя речь, не подвергнутую редукции и реструктуризации, – явлениям, неизбежным в процессе интерпретации речевого материала.

Однако, разграничивая первичные и вторичные данные, необходимо иметь в виду несколько обстоятельств.

Первое. Вторичный характер лингвистических данных ни в коей мере не означает их несущественности или «второсортности». Более того: результаты развития лингвистической теории всегда предстают в виде данных, вторичных по отношению к речи как объекту исследования. Результат собственных изысканий пользователя лингвистического ресурса и сам пополнит корпус вторичных данных об интересующих его речевых явлениях, о тех или иных лингвистических проблемах.

Второе. Первичные данные могут в том же источнике сочетаться с данными вторичного характера, дополняя друг друга. Кроме того, противопоставление первичных и вторичных данных и тем более ресурсов первичного и вторичного характера вообще не абсолютно. Оно имеет, скорее, градуальный характер. В самом деле, как бы тщательно в естественной материальной форме ни фиксировалась устная или письменная речь, уже сам процесс выделения текста из дискурса со всеми его сложными сторонами и отношениями оставляет исследователя без массы сведений, необходимых для его собственной работы с данным текстом, и, следовательно, полученный материал отчасти лишается свойства «первичности».

Из этого вытекает настоятельная необходимость включать в состав как первичных, так и тем более вторичных лингвистических ресурсов дополнительную информацию о дискурсах, из которых извлекались текстовые фрагменты (время и место протекания дискурса, адресант и его интенции, адресат и его коммуникативные возможности, язык общения, форма общения и т.д.), о конкретных текстах, послуживших источниками информации, и о тех электронных дискурсах (лингвистических ресурсах), в которые оказываются включенными подготовленные вторичные данные. Хорошо сконструированные лингвистические ресурсы обычно соответствуют этим условиям.

Так, обращаясь к собранию этимологических словарей лингвистического сайта «Вавилонская башня» (http://starling.rinet.ru), организованного в 1998 г. замечательным российским лингвистом Анатолием Сергеевичем Старостиным (1953-2005) и поддерживаемого его соратниками и учениками, пользователь получает информацию и о проекте в целом, и об А.С. Старостине, и об авторах каждого из собранных в базе этимологических словарей, и об источниках сформированных словарных данных, а представленные здесь же ссылки на рекомендуемые авторами специальные электронные ресурсы включают содержание сайта в широкий контекст мировой лингвистической науки.

Рисунок 4. Начало списка этимологических баз данных

Доступ ко всем этимологическим базам построен в «Вавилонской башне» одинаково и открывается со страницы сайта, на которой перечислены базы, указаны их создатели и время открытия на сайте (начало списка Этимологических баз данных см. на Рис. 4). Пользователю доступны описание каждой базы (ее составители, словарные источники, связи между базами) и последовательный просмотр содержащихся в базах материалов. Выбрав, например, базу «Индоевропейская этимология», можно последовательно просмотреть все имеющиеся в ней 3178 записей начиная с Proto-

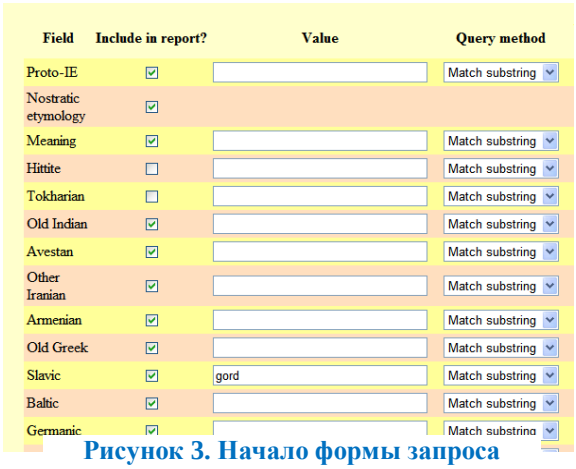


Рисунок 3. Начало формы запроса

IE *abel- и кончая Proto-IE *y^{e}nd- (Gr h-). Пользователь может сформировать конкретный запрос, заполнив соответствующую форму. Допустим, нас интересует этимология славянского корня *gord-. Включаем ее в форму запроса (фрагмент формы данного запроса см. на Рис. 3) и отмечаем интересующие нас этимологические связи. На Рис. 5 можно видеть выдаваемый базой ответ на данный запрос.

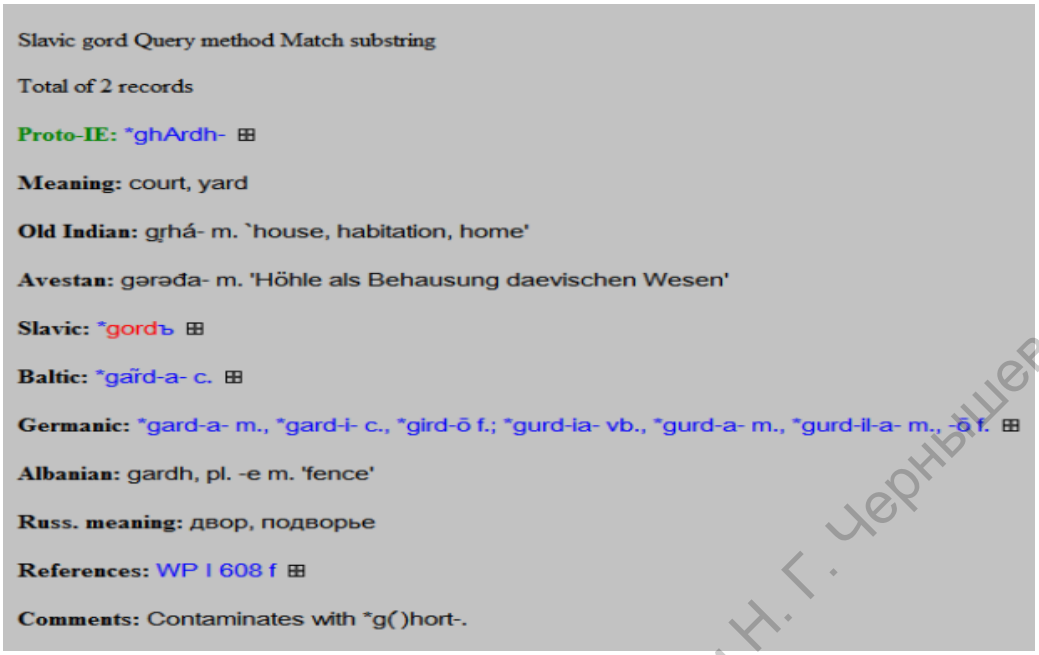


Рисунок 5. Ответ на запрос о корне *gord-

Еще один пример. Сайт «Лексикограф. Глагол» (http://lexicograf.ru) представляет основные идеи научного проекта «Лексикограф» – уникального русского семантического словаря, опирающегося на представление о том, что «в основе лексической системы языка лежат повторяющиеся смысловые компоненты (такие как ‘знать’, ‘видеть’, ‘двигаться’, ‘причина’, ‘предмет’) и параметры лексического значения – такие как категория, тематический класс, участник обозначаемой ситуации, таксономический класс участника».

Рисунок 6. Главная страница сайта "Лексикограф. Глагол"

В соответствии с задачей заявить основные идеи и результаты исследований, выполненных при реализации проекта, на сайте дана не только общая характеристика проекта, но представлены и его участники (Г.И. Кустова, Е.В. Падучева, Р.И. Розина), и списки их многочисленных публикаций, в том числе новейших, тезисы докладов и даже хендауты к некоторым лекциям и докладам.

Сайт развивается, и в сентябре 2009 г. на сайте уже выложена созданная участниками проекта база данных «Лексикограф», позволяющая не только знакомиться с содержащимися в

ней материалами, но и использовать ее как инструмент для собственных семантических разработок. На рис. 7 представлена в качестве примера закладка «Участники и толкование» с данными о лексеме ЛЕТЕТЬ1 (о птице).

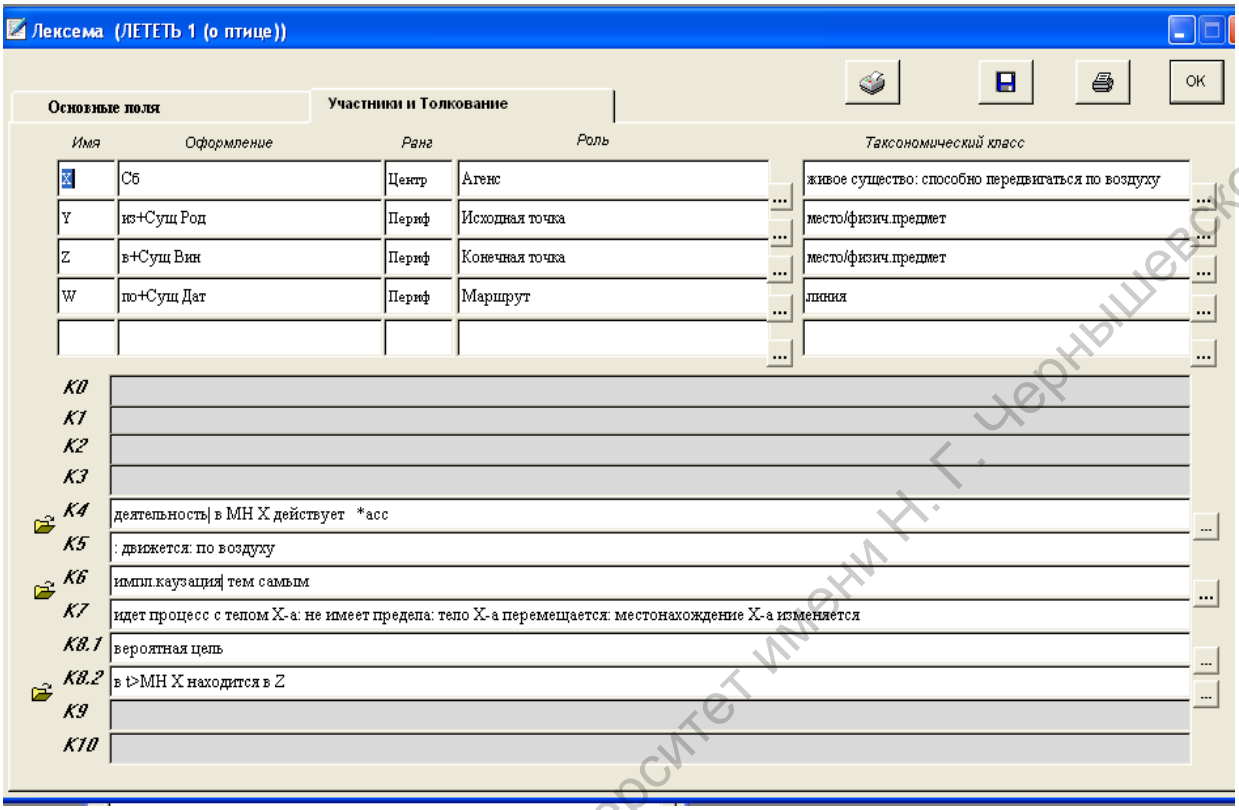


Рис. 7. Закладка «Участники и толкование»

Специально разработанные системы сопровождения речевых данных дополнительной информацией, необходимой для интерпретации этих данных, включаются в аннотированные текстовые корпуса (о них подробнее см. в лекции № 3).

Примером хорошо разработанного лингвистического ресурса с удобным пользовательским интерфейсом может служить один из продуктов SIL (см. ниже): база данных

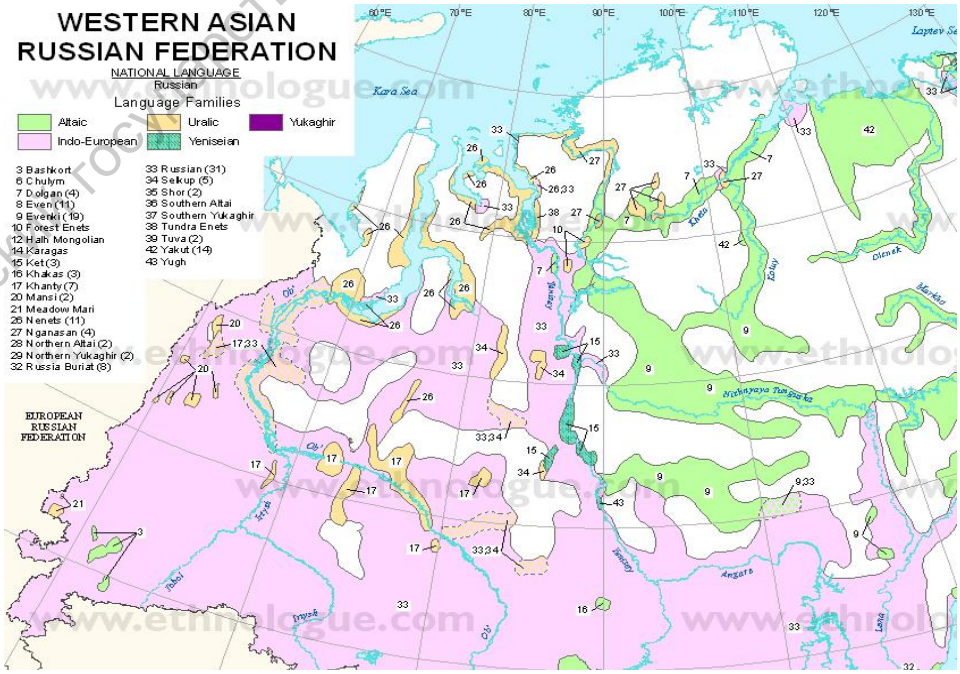


Рисунок 8. Фрагмент лингвистической карты России (по: www.ethnologue.com). Цветом обозначены языковые семьи

«Этнология» (http://www.ethnologue.com), содержащая краткие сведения о 6909 живых языках мира с демографическими указаниями, функциональными характеристиками, картами распространения. Материалы «Этнологии» могут быть полезными, но что касается малых языков и народов, то к приведенным сведениям нужно относиться критически: такие данные опираются на источники разной степени надежности, а соответствующие социальные и функциональные характеристики меняются достаточно быстро (ср., например, данные о кетах и кетском языке в «Этнологии» и в энциклопедическом словаре-справочнике «Языки народов России. Красная книга» (М., 2002).

Доступ к ряду on-line словарей организован на справочно-информационном портале «ГРАМОТА.РУ» (www.gramota.ru). Здесь же размещен постоянно пополняемый список ссылок на словари в Сети, в том числе на словари энциклопедические и терминологические. Коллекции ссылок на словари и энциклопедии находятся на страницах информационно поисковых систем (см., например: www.yandex.ru).

Разновидностью ресурсов вторичных лингвистических данных являются, как уже говорилось, электронные собрания работ по лингвистике. Собрания этого рода представлены на сайтах академических и образовательных учреждений, в составе научных электронных библиотек, в материалах крупных научных сообществ, организующих международные конференции периодического характера, в публикациях научных фондов, на сайтах ряда филологических журналов, на домашних страницах отдельных исследователей. Такие научные материалы предлагают, например, Институт лингвистических исследований РАН (http://iling.spb.ru/index.html), Институт языкознания РАН (см., в частности, статьи, монографии, тексты журнала «Вопросы психолингвистики» на сайте Института в разделе сектора психолингвистики) - http://www.iling-ran.ru/, Институт русского языка им. В.В. Виноградова РАН (www.ruslang.ru).

Лингвистические ресурсы, выложенные на сайте Института русского языка им. В.В. Виноградова особенно значительны. Это сайты Национального корпуса русского языка (см. о нем в лекции 3), Машинного фонда русского языка, собрание русских словарей и многое другое. Здесь, в частности, можно скачать оба тома академической «Русской грамматики» (1980) и воспользоваться поиском по ее тексту.

Здесь же в формате PDF размещен «Новый объяснительный словарь синонимов русского языка» (по второму, исправленному и дополненному, изданию – Москва; Вена: Языки славянской культуры: Венский славистический альманах, 2004 г. – 1488 с.) (http://www.ruslang.ru/agens.php?id=text_noss2_title). В Словаре подробно исследованы и описаны 354 синонимических ряда лексики русского языка (в основном антропоцентрического содержания). Словарь является новым не столько по времени создания, сколько по своему характеру. Как сформулировано в аннотации, «это словарь активного типа, согласованный с определенным грамматическим описанием русского языка, реализующий принципы системной лексикографии и ориентированный на отражение языковой, или «наивной», картины мира. Установка на детальное лингвистическое портретирование сочетается в нем с установкой на единообразное описание лексем, относящихся к одному лексикографическому типу. В Словаре последовательно отражаются семантические, референциальные, прагматические, коннотативные, коммуникативные, синтаксические, сочетаемостные, морфологические и просодические сходства и различия между синонимами, а также условия нейтрализации различий. Все словарные статьи содержат обширные справочные зоны, в которых перечисляются фразеологические синонимы, аналоги, точные и неточные конверсивы, конверсивы к аналогам, точные и неточные антонимы и дериваты (включая семантические) к элементам данного синонимического ряда. В некоторых случаях

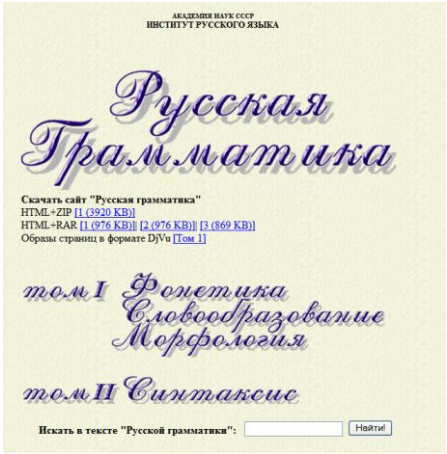


Рисунок 9. "Русская грамматика" на сайте ИРЯЗ им. В.В. Виноградова РАН

указываются специальные лингвистические работы, посвященные одной или нескольким лексемам, входящим в данный ряд».

Примером крупного ресурса научных работ по теории языка и компьютерной лингвистике являются материалы сайта международной конференции «Диалог» (www.dialog-21.ru). Конференция проводится уже более 30 лет; среди ее организаторов известные компании ABBYY и Яndex, Институт лингвистики РГГУ, Рос НИИ искусственного интеллекта, Институт проблем передачи информации РАН и другие научные центры. На сайте представлены полные тексты докладов, прочитанных на заседаниях конференции начиная с 2005 г., а также ряд более ранних материалов.

Важными данными металингвистического характера являются библиографические списки и базы, например библиографическая база работ по языкознанию, доступная на сайте ИНИОН (www.inion.ru). База ведется с 1986 г. и содержит почти 244 тыс. записей – аннотированных описаний поступивших в Фундаментальную библиотеку ИНИОН книг и статей на русском и других языках. Пользователям доступны поиск и сортировка данных по различным параметрам аннотаций, а также заказ электронных копий.

Примером современной электронной библиотеки, обеспечивающей не только поиск в ее ресурсах и доступ к полнотекстовым материалам, но и научную статистику, является Электронная научная библиотека «eLIBRARY.RU» (http://elibrary.ru).

3. Средства обработки лингвистического материала получили широкое распространение в современной лингвистике. В фонетических исследованиях они в значительной мере заменили использовавшиеся ранее осциллографы, спектрографы и подобную физическую аппаратуру для осуществления экспериментального изучения речи. Электронные информационные средства обработки речи оказывают существенную помощь и во многих других лингвистических исследованиях, а также в представлении их результатов.

Рисунок 10. Возможности шрифта Doulos SIL

Существует ряд Интернет-ресурсов, компилирующих различные средства обработки речи и предлагающих их пользователям. Один из таких ресурсов принадлежит SIL International (США) (www.sil.org/). Созданный еще в 1934 г. и известный как Летний лингвистический институт (SIL), сегодня SIL International – крупный международный центр, занимающийся изучением и описанием малых (в основном бесписьменных) языков, которым грозит исчезновение. Сотрудниками SIL International создан большой комплекс программных продуктов, бóльшая часть которых распространяется бесплатно. Среди них – достаточно мощный Speech Analyzer, осуществляющий акустический анализ речи, и большое количество компьютерных шрифтов, в том числе, например, шрифт Doulos SIL, дающий возможность соединять символы алфавитов с большим количеством различных диакритик, необходимых при детальной фонетической транскрипции речи (см. рис. 10).

Группа программистов и лингвистов, в основном выпускников РГГУ, работающая над созданием компьютерных средств обработки речи, с 2002 г. выкладывает свои продукты на широко известном сайте «АОТ. Автоматическая Обработка Текста» (www.aot.ru). Лингвисту, не занимающемуся программированием, может быть полезен раздел «ОНЛАЙН ДЕМО» (морфологический анализ словоформ, поиск по русскому размеченному массиву в 680 млн. словоупотреблений, поиск по биграммам) и «Система лингвистического поиска (DDC)» в разделе «СКАЧАТЬ». Последняя представляет собой вариант автоматического конкорданса.

Программы для создания автоматических конкордансов являются одним из наиболее активно используемых средств обработки текстов в их символьном представлении. Конкорданс к тексту (или корпусу текстов) – это список всех составляющих его слов (текстоформ или лемм) с приведением полного адресованного набора цитат, включающих данные слова. Во многих случаях пользователь автоматического конкорданса имеет возможность предварительно

размечать текст маркерами, чтобы раздельно анализировать разные части текста (отдельно, например, заголовки, отдельно сами тексты или комментарии к текстам, речь разных персонажей в драматургических произведениях и под.). Конкорданс создает основу для получения статистических данных о тексте, прежде всего – абсолютных и относительных частот составляющих его слов, предлагает способы различной фильтрации и сортировки материала, получения ответов на специальные запросы пользователя, предусматривает обычно и средства вывода полученных данных на печать.

Основы работы с автоматическим конкордансом хорошо изложены в книге Ж.Г. Аношкиной «Подготовка частотных словарей и конкордансов на компьютере» (М., 1995). Получил распространение созданный Ж.Г. Аношкиной текстоориентированный пакет программ UNILEX с удобной настройкой, морфологическим анализатором и значительным количеством различных опций конкорданса.

Интересную систему создания конкордансов, легко настраиваемую на обработку русского материала, предлагает R.J.C. Watt - www.concordancesoftware.co.uk.

Большая подборка ссылок на доступные в Сети компьютерные средства (морфологические и синтаксические парсеры, программы построения конкордансов, программы преобразования текстов, средства психолингвистического анализа, компьютерные словари и тезаурусы) содержится в Приложении к учебному пособию А.В. Всеволодовой для студентов, аспирантов, преподавателей-филологов «Компьютерная обработка лингвистических данных» (М.: Наука: Флинта, 2007). Из перечисленных в нем компьютерных ресурсов филологам, начинающим осваивать средства обработки лингвистического материала, можно рекомендовать самостоятельно разобраться в работе с **WordTabulator v2.2** (автор – С.В. Логичев): http://www.rvb.ru/soft/index.html и утилитой для Windows **Словогрыз** (автор И. Сагалаев) : http://www.softwaremaniacs.org/TR/.

Свободно распространяемая программа WordTabulator v.2.2.3 for Windows принимает на входе тексты в формате HTML или «обычный текст» и строит упорядоченные индексы искомых элементов, в качестве которых могут выступать словоформы, «словосочетания» или «синтагмы» (определения данных единиц см. в подробной справке к WordTabulator). Если выходной файл (индекс) задан в формате HTML-документа, то от каждого элемента индекса, например словоформы, возможен переход к обрабатываемому тексту, в котором подсвечиваются искомые единицы. При просмотре результата с помощью встроенного в программу Обозревателя фрагменты исходного текста с искомыми единицами могут копироваться обычными для Windows способами.

Существенной особенностью WordTabulator'а является возможность строить и параллельно обрабатывать два корпуса текстов, составы единиц которых могут по желанию пользователя либо объединяться в общем индексе, либо соотноситься по типу пересечения (на выходе показываются только элементы, встречающиеся в обоих корпусах одновременно), либо соотноситься по типу исключения (на выходе показываются только те элементы первого корпуса, которые не встречаются во втором корпусе.

В WordTabulator встроен разработанный Ж.Г. Аношкиной морфологический модуль, использование которого (его подключение нужно специально указывать в настройке) позволяет получать информацию обо всех словоформах указываемого в запросе слова. База морфологического модуля содержит 12000 фамилий, 2800 личных мужских и женских имен, 2100 отчеств, 40600 существительных, 18000 прилагательных, 20800 глаголов, 4000 других слов (наречий, междометий, предлогов и т.п.).

Словогрыз 3.1 - бесплатная утилита для Windows с функцией поиска и замен, допускает использование гибкой системы шаблонов с определяемыми пользователем переменными, удобна при необходимости осуществлять многошаговые замены, позволяет сохранять сценарии производимых замен и при необходимости запускать их вновь.

При выполнении самых различных исследований и прикладных работ оказывается необходимым ставить в соответствие некоторому количеству словоформ, образующих тексты или входящих в полученные каким-то образом списки, грамматические признаки этих словоформ и устанавливать начальные формы характеризуемых слов, то есть производить их

лемматизацию. Автоматический морфологический анализ и лемматизация используются в информационно-поисковых системах при обработке запросов, в больших текстовых корпусах при аннотировании материала и в других случаях. Ориентированные на русский язык специальные программы морфологического анализа (морфологические парсеры) опираются на данные уникального по полноте и строгости описания грамматических парадигм «Грамматического словаря русского языка» А.А. Зализняка (М., 1977), включающего около 100 000 слов. Однако в текстах кроме имен нарицательных функционируют и имена собственные, а также устаревшие слова, неологизмы, окказионализмы и другие лексические единицы, не представленные по тем или иным причинам в словаре А.А. Зализняка. Их морфологический анализ парсеры осуществляют по аналогии с представленными в их базе лексическими единицами.

Если автоматический морфологический анализ не сопровождается учетом контекста, то ему недоступно разграничение омонимичных единиц (омонимов, омоформ). Когда такие единицы встречаются в анализируемом материале, парсер показывает на выходе все варианты их морфологического разбора (варианты обычно отделяются один от другого знаком «|»). На сайте компании Яндекс (http://company.yandex.ru/technology/mystem/) можно скачать для некоммерческого использования созданный И.В. Сегаловичем консольный парсер Mystem, работающий с русским речевым материалом.

В качестве примера грамматического анализа, выполняемого парсером Mystem, покажем результат обработки парсером афоризма Б. Паскаля «*Только кончая задуманное сочинение, мы уясняем себе, с чего нам следовало его начать*»:

Только {только=ADV=|только=PART=|только=CONJ=|только=PART=}
кончая {кончать=V=непрош,деепр,несов}
задуманное {задумывать=V=прош,им,ед,прич,сред,сов,страд|задумывать=V=прош,вин,ед, прич,сред,сов,страд}
сочинение {сочинение=S,сред,неод=им,ед|сочинение=S,сред,неод=вин,ед}
мы {мы=SPRO,мн,од=им}
уясняем {уяснять=V=непрош,ед,прич,кр,муж,несов,страд|уяснять=V=непрош,мн,изъяв, 1-л,несов}
себе {себя=SPRO,ед,од=дат,жен|себя=SPRO,ед,од=дат,муж|себя=SPRO,ед,од=пр,жен|себя=SPRO,ед,од=пр,муж|себе=PART=}
с {с=PR=}
чего {чего=ADVPRO=}
нам {мы=SPRO,мн,од=дат}
следовало {следовать=V,несов=прош,ед,изъяв,сред}
его {его=APRO=им,ед,жен|его=APRO=им,ед,муж|его=APRO=им,ед,сред|его=APRO=им,мн|его=APRO=род,ед,жен|его=APRO=род,ед,муж|его=APRO=род,ед,сред|его=APRO=род,мн|его=APRO=дат,ед,жен|его=APRO=дат,ед,муж|его=APRO=дат,ед,сред|его=APRO=дат,мн|его=APRO=вин,ед,жен|его=APRO=вин,ед,муж,од|его=APRO=вин,ед,муж,неод|его=APRO=вин,ед,сред|его=APRO=вин,мн,од|его=APRO=вин,мн,неод|его=APRO=твор,ед,жен|его=APRO=твор,ед,муж|он=SPRO,ед,муж,од=род|он=SPRO,ед,муж,од=вин|оно=SPRO,ед,сред,од=род|оно=SPRO,ед,сред,од=вин}
начать {начинать=V=инф,сов}

В фигурные скобки заключена вся информация, относящаяся к одной текстоформе. После открывающей фигурной скобки записана начальная форма слова; знаком «=» разделены разные зоны грамматической характеристики: указания части речи, классифицирующих признаков, словоизменительных признаков. Знаком «|» разделяются, как уже сказано, грамматические варианты.

Обозначения частей речи, принятые в Mystem:

A - прилагательное,
ADV – наречие,

CONJ – союз,
INTJ - междометие,
NUM – числительное,
PART – частица,
PR - предлог ,
S – существительное,
V – глагол,
ANUM - порядковое числительное,
APRO - местоименное прилагательное,
ADVPRO - местоименное наречие,
SPRO - местоименное существительное.

Целый ряд относительно доступных программных продуктов поддерживает сопоставление и тематическую рубрикацию различных документов, в том числе, функционирующих в Сети. Для начального знакомства с этим типом ресурсов можно рекомендовать распространяемую с бесплатным испытательным сроком программу Rubryx (авторы – В. Поляков и В. Синицын) (www.sowsoft.com.rubryx/index.html). В соответствии со своими задачами пользователь (эксперт) указывает тематические классы (рубрики), по которым программа Rubryx должна распределять тексты, и подбирает для каждой рубрики несколько достаточно типичных, по его мнению, текстов в качестве образцов. На их основе программа автоматически строит тематические микрословари особого формата. Существенно, что в их состав входят однословные, двусловные и трехсловные единицы. Результат этой работы доступен контролю со стороны эксперта и может корректироваться им. При анализе новых текстов программа сравнивает состав их однословных, двусловных и трехсловных единиц с имеющимся в предварительно созданных микрословарях рубрик и устанавливает коэффициент родства между ними. Опытные прогоны позволяют поэтапно улучшать настройку программы на тематические рубрики.

Лингвистические принципы, положенные в основу Rubryx, обсуждаются в статье: *Поляков В.Н.* Использование технологий, ориентированных на лексическое значение, в задачах поиска и классификации // Scripta linguisticae applicatae. Проблемы прикладной лингвистики. Вып. 2. М., 2004.

Особый большой класс лингвистических ресурсов образуют лингводидактические программы, подробно рассматриваемые в учебном пособии *Бовтенко М.А.* Компьютерная лингводидактика. М.: Наука: Флинта, 2007.

4. В создании электронных ресурсов в области русистики важную роль сыграл проект «Машинный фонд русского языка» – программа комплексной информатизации исследований в русистике.

Идея создания Машинного фонда русского языка (МФРЯ) была высказана акад. А.П. Ершовым в 1978 г. на конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог». В 1980-х гг. был проведены специальные конференции, посвященные проблемам создания МФРЯ, сыгравшие важную роль в развитии отечественной компьютерной лингвистики. В это же время в Институте русского языка имени В.В. Виноградова РАН был создан Отдел машинного фонда русского языка, разработана «Комплексная программа научных исследований и прикладных разработок по созданию Машинного фонда русского языка на 1996-2000 гг. и информатизации исследований в Институте русского языка АН СССР». Руководителями Отдела были последовательно член-корреспондент АН СССР Ю.Н. Караулов (1985-1991 гг.), доктор филологических наук В.М. Андрющенко (1992-1998 гг.), профессор, доктор филологических наук А.Я. Шайкевич (1998-2006 гг.). В создании МФРЯ принимали участие более 40 организаций-соисполнителей, среди них Московский, Санкт-Петербургский, Харьковский, Гродненский, Сыктывкарский и Саратовский университеты. В 2006 г. Отдел был ликвидирован, работы по развитию МФРЯ продолжаются в рамках отдела корпусной лингвистики и лингвистической поэтики, которым руководит член-корр. РАН В.А. Плунгян.

В основе концепции, выработанной 1-ой Всесоюзной конференцией по созданию МФРЯ 1983 г., лежали две взаимосвязанные главные задачи: 1) создание компонентов лингвистического обеспечения задач информатики и 2) информатизация научных исследований

в русистике. В соответствии с этим в качестве базовых компонентов МФРЯ были определены: источники на машинных носителях и в базах данных; лингвистические программно-источниковые пакеты; компьютерные технологии подготовки научных трудов. На создание этих продуктов были направлены первоочередные усилия разработчиков МФРЯ.

Уже на первом этапе создания МФРЯ (1985—1992 гг.) были накоплены на машинных носителях и частично в базах данных текстовые источники русской литературы XIX-XX вв., главнейшие словари русского языка, Краткая академическая грамматика, некоторые другие материалы справочного характера, созданы текстовые корпуса поэзии, художественной прозы, общественно-политических и технических текстов. Были разработаны две подсистемы пакета UNILEX на персональных компьютерах (под MS DOS): текстоориентированная подсистема – UNILEX-T (разработчик Ж.Г. Аношкина), предназначенная для создания частотных словарей, словоуказателей (индексов слов к текстам) и конкордансов, и словарная подсистема – UNILEX-D (разработчик Л.И. Колодяжная) – для создания словарных баз данных и работы с ними. Было создано несколько программно-источниковых пакетов, таких как Автоматический Синтаксический словарь русского языка, Автоматический словарь синонимов русского языка, Автоматический вариант Словаря русского языка С.И. Ожегова, Автоматический словарь глагольного управления в русском языке и др.; разработаны технологии редакционно-издательской подготовки научных трудов и продуктов Машинного фонда русского языка.

Одним из удачных примеров реализации идей МФРЯ был созданный под руководством проф. Н.Н. Пшеничновой автоматический вариант Диалектологического атласа русского языка. База данных Диалектологического атласа представляет собой матрицу (характеристическую таблицу), в которой число строк равно числу обследованных пунктов, а число столбцов – числу языковых признаков, наличие или отсутствие которых характеризует соответствующие пункты. По наименованию пункта можно было получить все характеризующие его признаки, а по наименованию признака – список пунктов, в которых этот признак отмечен. Описание БД ДАРЯ приведено в издании [Андрющенко 1989].

Наряду с работой по наполнению источниковой базы МФРЯ, разработкой программных пакетов и компьютерных технологий для обработки лингвистических данных проводилось обучение филологов информатике, методам автоматизации филологических исследований, работе с компонентами Машинного фонда русского языка.

Еще в 1980-е гг., на начальном этапе разработки проекта МФРЯ, были выдвинуты важные теоретические идеи, сохраняющие свою актуальность для дальнейшего развития корпусной лингвистики (о корпусной лингвистике см. лекцию 3).

МФРЯ планировался и реализовывался прежде всего как словарная база данных. В соответствии с «лексикографической идеологией» вся информация МФ должна была группироваться вокруг слова, представлять собой базу данных, в которой содержатся всевозможные сведения о слове: его грамматические, стилистические, фонетические, контекстуальные и т.п. характеристики, сведения о разнообразном варьировании слова, в том числе поэтическом, диалектном, диахронном варьировании. Эти положения изложены в статье Ю.Н. Караулова «Методология лингвистического исследования и машинный фонд русского языка» в сборнике [Машинный фонд… 1986]. Для построения словарно-грамматической базы данных предлагалось перевести академические словари и грамматики в электронную форму; создать автоматически пополняемые словоуказатели и словари на базе текстов делового и разговорного стилей, научно-технической литературы и документации; соединить в единый языковой фонд данные об общеупотребительном языке и данные терминологических фондов; создать специальные программы машинной обработки лингвистического материала.

Лексикографическая ориентация МФРЯ не препятствовала параллельному формированию его «текстовой идеологии», согласно которой в МФРЯ в качестве его обязательной части должны были войти массивы (корпуса) текстов, представляющие подъязыки русского языка – его функционально-стилистические разновидности, различные сферы общения (Ю.Н. Караулов, Б.Ю. Городецкий, А.С. Герд, С.И. Гиндин, В.И. Перебейнос). Обсуждалась и возможность текстоориентированнной разработки МФРЯ. Ср.: «основной единицей хранения в фонде должна быть не лексическая единица, а текст (Ю.К. Орлов);

«тексты – то единственное первичное, достоверное, фактическое в языке, что вводится в машину (А.Я. Шайкевич); о диалектологическом подфонде МФРЯ: «машинный фонд может ориентироваться только на полные магнитофонные записи диалектных текстов» (А.С. Герд); текстовые диалектологические корпуса отдельных говоров должны составить основу диалектологического подфонда МФРЯ (В.Е. Гольдин).

И все-таки первоначально в структуре МФРЯ текстовым корпусам отводилась иллюстративная роль, текстовый подфонд был обозначен как «иллюстративно-текстовый» [Андрющенко 1989].

Актуальными для современной компьютерной лингвистики остаются сформулированные на конференциях по созданию МФРЯ общие принципы организации МФ:

– автоматизированная система, представляемая в МФРЯ, «должна быть адекватной и равнообъемной живому организму языка, но в то же время она должна быть анатомически отпрепарированной, разъятой, доступной для наблюдения, изучения и изменения» (А.П. Ершов);

– «фонд должен быть не просто банком данных, а творческой лабораторией. Это значит, что лингвист, пользующийся фондом, не ограничивается наведением справок, а находится в собственно исследовательской позиции», имея возможность перекомбинирования, переконцентрации материала, возможность получения новых классификаций. Все это, в свою очередь, обеспечивается «на основании максимально разветвленной и дробной параметризации языковых данных и явлений» (Ю.Н. Караулов);

– в МФРЯ должен быть осуществлен принцип всесторонности и междисциплинарности информации, связанной с русским языком, фонд должен включать не только собственно лингвистические данные, но и учитывать лингво-культурную и историко-этнографическую информацию (Ю.Н. Караулов, В.Е. Гольдин).

Уже при первоначальной постановке задачи создания МФ подчеркивалось, что МФ – это не только современная форма представления языкового материала, но и путь «к смене научной парадигмы в лингвистике», возможность перехода «от наблюдательного периода к измерительному» (А.П. Ершов). Обеспечивая полноту данных, позволяя исследователю «не ограничиваться образцами реализации того или иного правила, явления, иллюстрациями проявления какого-то признака, приблизительными обозначениями некоторой материальной области языковой структуры, а в каждом конкретном случае получать точное знание, доходить до исчерпывающего перечня всех элементов соответствующего множества» (Ю.Н. Караулов), использование МФРЯ должно, по мысли его инициаторов, иметь методологические последствия: раздвинуть границы понимания самого объекта (русского языка), увидеть его в новом ракурсе, выявить «белые пятна».

Известно, что ни одна научная парадигма не реализуется до конца, полностью. Обычно в ее недрах вызревают идеи сменяющей ее новой парадигмы, призванной разрешить трудности и противоречия предыдущей. Проект Машинного фонда русского языка был своеобразной научной парадигмой. Реализовать ее полностью не удалось. Единого комплекса, обеспечивающего информатизацию русистики, каким представлялся будущий Машинный фонд в начале работ по его созданию, сегодня не существует, но накопленные в нем материалы постоянно используются; работают и программные продукты Машинного фонда, а связанные с Машинным фондом научные идеи воплощаются в новых крупных проектах, в том числе – в Национальном корпусе русского языка.

5. Рекомендации и задания. Для приобретения хотя бы самой общей, первичной, ориентации в лингвистических ресурсах рекомендуется, во-первых, пройти по ссылкам на Интернет-ресурсы, приведенным в тексте лекции. Во-вторых, полезно попробовать самостоятельно отыскать в Сети ссылки на ресурсы, соответствующие той или иной лингвистической тематике. В поиске необходимых лингвистических ресурсов большую помощь оказывают информационно-поисковые системы: Яndex (www.yandex.ru), Rambler (www.rambler.ru), Апорт (www.aport.ru), Google (www.google.ru), Yahoo (www.yahoo.com), Copernic (www.copernic.com) и др. Они позволяют находить не только отдельные лингвистические источники, но и группы таких источников, в том числе списки «полезных ссылок»,

представленные на многих сайтах и лингвистических порталах. Хотя часть ссылок со временем неизбежно устаревает и перестает работать, а пополнение списков новыми данными не всегда происходит, собрания полезных ссылок являются важным средством, поддерживающим работу лингвистов. Их ценность состоит в том, что они обычно тематически организованы и помогают пользователю обнаруживать дополнительные ресурсы, относящиеся к актуальной для него проблеме.

Подробный и – что очень важно – хорошо ориентированный на нужды лингвистов список рекомендуемых ссылок представлен Институтом лингвистики РГГУ (http://il.rsuh.ru/links.html). В него входят Интернет-адреса сайтов образовательных учреждений, институтов РАН, международных лингвистических организаций и конференций, лингвистических рассылок и дайджестов, лингвистических корпусов по многим языкам мира, словарей и других баз данных. Завершается список адресами домашних страниц большого числа зарубежных и российских лингвистов, это позволяет знакомиться с перечнями их работ и скачивать отдельные научные и методические материалы.

Детально разработанный список ссылок на лингвистические Интернет-ресурсы предлагается на сайте Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета Московского государственного университета им. М.В. Ломоносова (http://www.philol.msu.ru/~lex/links.htm). Он включает ссылки на лингвистические сайты, ресурсы по русскому языку и иностранным языкам (словари, текстовые базы данных и корпуса), на крупные лингвистические центры и научные проекты, научные фонды, библиотеки, личные страницы лингвистов.

Добротный список ссылок на славистические издания, библиотеки, издательства можно найти на сайте «RUTHENIA» (http://www.ruthenia.ru/web/index.html) Тартуского университета. Многочисленные собрания ссылок на лингвистические ресурсы размещены на домашних страницах исследователей и разработчиков, в составе отдельных публикаций. Так, большой ссылочный материал и полезные советы по лингвистическому поиску в Сети содержит публикация профессионального переводчика Ю.В. Тиссена «Интернет в работе переводчика» (http://zhurnal.lib.ru/w/wagapow_a_s/tissen.shtml). Группой авторов (*Ю.А. Загорулько, И.С. Кононенко, Е.Г. Соколова*) создан Интернет-портал знаний по компьютерной лингвистике (http://uniserv.iis.nsk.su/cl).

Для закрепления полученных сведений выполните следующие задания:

1. Откройте в Интернете ресурс **«Рукописные памятники Древней Руси»** (http://www.lrc-lib.ru/?id=1) и ознакомьтесь с представленными на его главной странице сведениями о Некоммерческом Партнерстве «Рукописные памятники Древней Руси».

Выясните:

а) каковы цели Некоммерческого Партнерства «Рукописные памятники Древней Руси»;

б) чем обеспечивается научность информации, распространяемой Некоммерческим Партнерством «Рукописные памятники Древней Руси»;

в) каковы принципы взаимодействия Некоммерческого Партнерства «Рукописные памятники Древней Руси» с правообладателями памятников;

г) каким образом обеспечивается защита представленных на сайте изображений от их несанкционированной перепечатки?

д) какие типы рукописных памятников Древней Руси представлены на сайте?

2. Перейдите по ссылке на **сайт «Древнерусские берестяные грамоты»** и ознакомьтесь с представленной на нем информацией.

Выведите на экран и изучите грамоту № 590. Установите, что дает пользователю открытие фотографии грамоты в новом окне (1680x540 px). Удается ли при этом различить написанный на бересте текст? Предусмотрено ли дополнительное увеличение изображения? Что это дает?

Представлен ли в базе текст данной грамоты с разделением его на слова?

Доступен ли перевод данной грамоты?

Как документирована грамота (сведения о месте обнаружения, предположительной датировке грамоты, указание на ее тип и т.д.)?

3. Перейдите в **Базу данных** сайта «Древнерусские берестяные грамоты». Используя предложенный на сайте поисковый аппарат, получите список целых грамот жанра письмо, которые отнесены к 1050-1075 гг. и найдены в любом из новгородских раскопов. Выберите из списка грамоту «От Рожнета к Коснятину» и ознакомьтесь с представленной о ней информацией. По собственному выбору запросите в базе информацию о грамотах других жанров, другого типа сохранности, о грамотах, отнесенных к другим периодам времени, о грамотах, найденных не в Новгороде.

Ответьте на вопросы:

а) какие именно данные, относящиеся к древнерусским берестяным грамотам, сосредоточены в базе «Древнерусские берестяные грамоты»;

б) по каким параметрам организован поиск грамот в базе «Древнерусские берестяные грамоты»;

в) допускается ли указание в запросе сочетания нескольких параметров;

г) можно ли указывать в запросе одновременно несколько значений одного параметра;

д) является ли база завершенным электронным ресурсом или продолжает пополняться и уточняться;

е) каковы источники информации, представленной на сайте;

ж) какова роль прорисей в расшифровке текстов берестяных грамот;

з) являются ли прориси источником первичного типа (обоснуйте своё мнение)?

Выясните, какие научные тексты опубликованы на сайте и в каком формате, как построен библиографический указатель, что представляет собою «Кумулятивная библиография трудов о берестяных грамотах».

Дайте краткое описание ресурса «Древнерусские берестяные грамоты» в виде ответов на предложенные выше вопросы.

Литература

Основная:

Бовтенко М.А. Компьютерная лингводидактика. – М.: Наука: Флинта, 2007.

Всеволодова А.В. Компьютерная обработка лингвистических данных. – М.: Наука: Флинта, 2007.

Зубов А.В., Зубова И.И. Информационные технологии в лингвистике. – М.: Издательский центр «Академия», 2004.

Машинный фонд русского языка: идеи и суждения. – М., 1986.

Дополнительная:

Андрющенко В.М. Концепция и архитектура машинного фонда русского языка. М.: Наука, 1989.

Андрющенко В.М. Машинный фонд русского языка: Интеграционный подход (научно-методический материал). М.:, 1989.

Аношкина Ж.Г. Подготовка частотных словарей и конкордансов на компьютере. – М., 1995.

Гуманитарные исследования в Интернете / Под ред. А.Е. Войскунского – М., «Можайск–Терра», 2000.

Лукашевич Н.В., Добров Б.В. Тезаурус для автоматического концептуального индексирования как особый вид лингвистического ресурса // Труды Международной конференции ДИАЛОГ-2001 – М., 2001.

Поляков В.Н. Использование технологий, ориентированных на лексическое значение, в задачах поиска и классификации // Scripta linguisticae applicatae. Проблемы прикладной лингвистики. Вып. 2. М., 2004.

Лекция 2. Лингвистические аспекты гипертекстовой коммуникации

1. Введение.
2. Лингвосемиотические предпосылки гипертекстовой коммуникации.
3. Лингвистические средства системной организации структурированных гипертекстов.
4. Рекомендации и задание.
5. Литература.

1. Введение. Существуют различные определения гипертекста, но они в основном опираются на одну и ту же группу понятий: ‘текст’, ‘фрагмент’, ‘связь’, ‘ссылка’, и самое общее представление о гипертексте сводится к тому, что **это текст, относительно самостоятельные фрагменты которого связаны между собой явным образом с помощью специальных ссылок.**

При таком широком понимании гипертекста он не обязательно должен быть реализован на основе компьютерной технологии: гипертекстами можно считать, например, тексты кодексов законов в совокупности с множеством комментариев к ним, если каждый комментарий содержит явную ссылку на тот или иной комментируемый закон. В качестве гипертекста при таком его понимании можно рассматривать и любой текст, состоящий из ряда частей, глав, разделов и т.п., если при нем имеется оглавление с соответствующими номерами страниц, отмечающими пути перехода к каждой его части. Широкое понимание гипертекста не противоречит узкому, ограничивающему существование гипертекста компьютерной средой, но удобно тем, что позволяет обнаружить связи между общими тенденциями развития речевого общения и теми специфическими возможностями организации письменного общения, которые предоставляют компьютерные технологии.

Представление об относительности границ текста, предполагающее возможность подходить к тексту и как к целому, и одновременно как к части другого целого, далеко не ново: его проявление можно видеть уже в такой, например, сложившеюся в глубокой древности работе с текстами, как собирание и структурирование библиотек, архивов или составление сборников, включающих тексты сходной направленности. Во 2-ой пол. XX в. это представление актуализировалось, обогатилось новыми связями и развилось в понятие ‘гипертекст’ в связи с результатами разработки компьютерных технологий, использующих стандартные способы реализации текстовой относительности, обеспечивающие мгновенность переходов по ссылкам и возможность создавать гипертексты немыслимых ранее структур и размеров.

Актуализация этих представлений выразилась в том, что **теперь относительность границ текста стала осознаваться как одно из главных его свойств**. Одновременно изменились типы связываемых фрагментов, или гипотекстов; другими стали коммуникативные роли автора и получателя текстов; изменился характер создания, обработки, обслуживания и потребления текстов. Таким образом, внедрение гипертекстовых технологий затронуло саму структуру коммуникации, и сегодня под гипертекстом понимают и реализованный в компьютерной среде текст с относительно самостоятельными фрагментами, связанными с помощью специальных ссылок, и компьютерные технологии, поддерживающие создание и использование таких текстов, и новую сферу общения, складывающуюся и функционирующую на основе гипертекстовых технологий.

Историю гипертекста обычно начинают с 1945 г., когда в журнале «The Atlantic Monthly» появилась статья В. Буша (Vannevar Bush), бывшего советника по науке президента Ф. Рузвельта, «As we may think», в которой он описал настольную механическую систему Memex, использующую технологию микрофильмирования и позволяющую устанавливать и сохранять связи между отдельными текстами, делать пометки на полях «страниц», присоединять к текстам комментарии и использовать все это как единую систему хранения информации. Автору системы представлялось, что таким или сходным образом работает на

основе ассоциативных связей и человеческий мозг и может строиться объединенная информационная среда.

Так была заявлена идея гипертекста, но до 60-х годов ни соответствующей компьютерной технологии, ни самого термина «гипертекст» еще не существовало. Термин «гипертекст» был введен Т. Нельсоном (Theodor Holm Nelson) в 1965 г. в связи возникшей необходимостью определить специфику системы, предназначенной обслуживать на основе компьютерной технологии множество текстов со связывающими их переходами. Началась разработка специальных программных средств для создания и поддержки гипертекстовых систем и их стандартизация. В 1990 г. сотрудник Европейского центра ядерных исследований в Женеве Тим Бернерс-Ли (Tim Berners-Lee) разработал сетевой протокол передачи документов *HTTP* (*HyperText transfer protocol*) и язык гипертекстовой разметки *HTML* (*Hypertext Markup Language*). Их применение, а также создание и совершенствование специальных браузеров открыли эру «Всемирной паутины» (*World Wide Web*), как определил Бернерс-Ли новую сферу интеллектуального взаимодействия на базе гипертекстовых технологий.

Последовавшее бурное развитие гипертекстовых систем, с одной стороны, несомненно, обязано совершенствованию компьютерной техники и информационных технологий, без которых было бы невозможно создание и функционирование современных гипертекстов. С другой стороны, столь же несомненно, что развитие гипертекстовых систем обусловлено информационными потребностями общества и всей историей знаковой репрезентации речи, а это уже входит в сферу интересов лингвистики.

Ниже из всего комплекса проблем, связанных с гипертекстовой коммуникацией, мы сосредоточимся на двух: во-первых, на лингвосемиотических предпосылках развития гипертекстовой коммуникации и, во-вторых, на лингвистических средствах системной организации структурированных гипертекстов.

2. Лингвосемиотеческие предпосылки гипертекстовой коммуникации. Из числа особенностей коммуникации на гипертекстовой основе, соответствующих ее главным принципам, обратим внимание на следующие:

а) на нелинейность композиции гипертекста (включенные в текст ссылки позволяют «разветвлять» движение по тексту, выбирать ту или иную последовательность ознакомления с его частями),

б) на относительность границ гипертекста и текстов в составе гипертекста (текст, связанный ссылками с другими текстами, сохраняя некоторую автономность, вместе с тем структурно и содержательно становится частью большего образования, и его собственные границы теряют абсолютный характер),

в) на особую значимость роли читателя по отношению к гипертексту (именно читатель, выбирая одни связи и добавляя другие, формирует в конечном счете гипертекст).

Рисунок 6. Письмо индейца-чейена (по И.Е. Гельбу)

Являются ли эти коммуникативные особенности совершенно новыми? Чем обусловлена потребность в именно таком типе обмена письменными сообщениями?

Обратим прежде всего внимание на то, что для одной из самых древних систем передачи сообщений – пиктографии, – не существовало проблемы линейности текста. Простейшее «рисуночное» письмо не имело «начала» и «конца» и выражало передаваемое содержание связью всех своих компонентов в целом. Это свойство пиктографии прослеживается даже в тех пиктографических посланиях, которые создавались неграмотными представителями племен американских индейцев уже в новое время, в эпоху преобладающего использования других систем письма. Таково, например, письмо индейца из племени чейенов по имени Черепаха-

Следующая-За-Своей-Женой, которое он отправил по почте в обычном конверте сыну по имени Маленький Человек одновременно с переводом на его счет суммы в 53 доллара (см. Рис. 1).

И.Е. Гельб, воспроизводящий в своей монографии (Гельб 1982) данное письмо, раскрывает его содержание следующим образом: «Над фигурой человека, нарисованного слева, изображена черепаха, идущая вслед за своей женой и соединенная линией с головой этого человека, а над фигурой человека справа изображен маленький человек, от которого также проведена линия к голове этого второго человека. Над правой рукой этого второго человека нарисован еще один маленький человечек, поза которого изображает прыжок или движение в сторону Черепахи-Следующей-За-Своей-Женой, – человека, изображенного слева; от рта этого последнего отходят две линии, загнутые на концах как бы углом или крючком, словно он тянет маленькую фигурку к себе. Эта часть рисуночного послания, очевидно, и составляет содержание письма, то есть означает: ‘приезжай ко мне’. Более же крупные фигуры и их личные тотемы обозначают адресата и отправителя. Вверху между двумя крупными фигурами нарисованы 53 круглых предмета, под которыми подразумеваются 53 доллара. Оба индейца изображены в набедренном одеянии, указывающем на их принадлежность к племени чейенов…» (Гельб 1982: 39). Как сообщает исследователь, сын Черепахи-Следующей-За-Своей-Женой понял содержание письма и сразу отправился за отправленными на его счет долларами.

Благодаря изображенным в письме человеческим фигуркам оно получает для нас естественную ориентацию по вертикали. Следует ли, однако, читать его слева направо или справа налево (напомним о существовании систем письма, располагающих текст справа налево и иным образом), по вертикали или вращая, остается неизвестным и, по-видимому, несущественно. Неясен и порядок сообщений о двух ситуациях, отображенных в письме: ситуации смены собственности (присылка сыну 53 долларов) и ситуации приглашения к отцу: этого порядка просто нет.

Пиктография отображала не речь, а смысл сообщения. Однако, как известно, историческое развитие систем письма совершалось в значительной мере в сторону все более точной репрезентации ими внешней стороны речи, в том числе и такого ее свойства, как линейность. Линейность постепенно стала существенной характеристикой письменных и отдельных видов устных текстов (текстов ораторской речи, например) и закрепилась в виде представлений об их композиции. Ср. жесткий состав частей и порядок их следования в деловых документах и не менее важные, но по-другому реализуемые композиционные требования к литературно-художественным текстам.

Вторая характерная тенденция «пред-гипертекстовой» коммуникации – придание всё большего значения авторству текстов и развитие представления об авторской точке зрения как о факторе, организующем текст и выражающем его цельность. История культуры свидетельствует о том, что современные представления об авторстве текстов, как и текстовая линейность, складывались постепенно. Они явились следствием множества коммуникативных факторов, которыми были, с одной стороны, процессы усиления роли письменной коммуникации в сравнении с устной, укрупнения и усложнения текстов, развития возможностей их массового тиражирования и, с другой стороны, – обострение внимания к человеческой личности и повышение ценности индивидуального начала в общении. Всё это укрепляло идею авторства и придавало авторскому тексту свойства большей выделенности, самостоятельности, обособленности в континууме коммуникации.

Одновременно с формированием представления об авторе письменного текста как субъекте, создающем текст и имеющем на него особые права, формировалось и соответствующее ему представление о читателе как потребителе текста, лишенном права вносить в текст собственные изменения. Исторический процесс разделения коммуникантов на авторов и потребителей текстов нашел, в частности, яркое проявление в истории русских словарей,

Словари, отмечавшие и толковавшие слова иноязычного происхождения, развивались первоначально на основе заметок на полях, глосс, которые делались читателями, копировались переписчиками текстов и со временем стали сводиться в словарные списки, упорядоченные по

алфавиту первых букв слов. Вторая и последующие буквы первоначально не принимались во внимание, что делало списки открытыми и позволяло пользователям рукописей свободно дополнять словарные списки собственным материалом, выступая в роли читателя-соавтора. В эпоху печатных авторских словарей это становится уже невозможным: читателю предоставляется в словаре в крайнем случае пара чистых страниц «для заметок», но лишь строго за пределами авторского текста.

Дифференциация коммуникативных ролей «автор» и «читатель» представляет собой часть общего процесса социально-профессиональной дифференциации коммуникативной сферы. В России I пол. XIX в. социализацию и профессионализацию ролей «сочинитель», «издатель», «книгопродавец», «читатель», «критик» отмечал и остро переживал А.С. Пушкин. См., например, рассуждения поэта о «звании стихотворца» и о «публике», получившие выражение в тексте 1830 г.:

> «*Зло самое горькое, самое нестерпимое для стихотворца есть его звание, прозвище, коим он заклеймен и которое никогда его не покидает. Публика смотрит на него как на свою собственность, считает себя вправе требовать от него отчета в малейшем шаге. По ее мнению, он рожден для ее удовольствия и дышит для того только, чтобы подбирать рифмы. Требуют ли обстоятельства присутствия его в деревне, при возвращении его первый встречный спрашивает его: не привезли ли вы нам чего-нибудь нового? Явится ли он в армию, чтобы взглянуть на друзей и родственников, публика требует непременно от него поэмы на последнюю победу, и газетчики сердятся, почему долго заставляет себя ждать*…»

Сегодня ролевая дифференциация и профессионализация коммуникации достигли самого высокого уровня:

- тексты **профессионально создаются** (и отчасти ретранслируются) исполнителями социальных ролей «писатель», «журналист», «законодатель», «комментатор», «ученый», «экскурсовод», «пресс-секретарь» и под.,
- тексты **профессионально ретранслируются** (и отчасти создаются) исполнителями социально-профессиональных ролей «издатель», «диктор», «актер», «переводчик», «преподаватель» и под.,
- тексты **контролируются** «редакторами», «корректорами», «критиками», «судьями», «терминологами», «лексикографами» и под.,
- тексты **доставляются** «читателям», «слушателям», «зрителям» («публике»), «ученикам», «студентам» и подобным исполнителям социальных ролей группы «получатель»,
- **хранением, изучением, реставрацией, обработкой и распространением текстов профессионально занимаются** специальные службы и организации: канцелярии, архивы, библиотеки, музеи и др.

Названные тенденции (усиление линейности письменных текстов, бóльшая противопоставленность коммуникативных ролей автора и читателя, профессионализация общения) сами по себе не могли бы явиться предпосылкой развития гипертекстовой коммуникации, имеющей прямо противоположные свойства. **Однако важно иметь в виду, что развитие отмеченных коммуникативных тенденций сопровождалось в качестве следствия появлением и средств преодоления этих тенденций.**

Так, жесткость линейной композиции текста преодолевается его делением на части, разделы, главы, параграфы, присоединением к тексту оглавления с указанием номеров страниц каждой из частей, сопровождением текста специальными указателями, шрифтовым обозначением более и менее важных частей текста (ср. использование «мелкого шрифта» в учебной литературе), формированием комментариев, дополнений, приложений в качестве субструктур, соотносимых с основным текстом, выделением метаязыкового и метатекстового содержания в отдельные тексты (ср.: грамматики, словари, библиографии и под.) и др. Все эти

способы преодоления линейности текстов появляются и усиливаются, как известно, еще до становления компьютерных технологий, но находят в них поддержку и развитие.

Точно так же, чем более самостоятельной, структурно оформленной и с этой точки зрения относительно замкнутой единицей выступает текст, тем сильнее делается потребность указывать каким-то способом его связи с другими текстами, определять в явном виде его возможных и желательных адресатов, отмечать его место в коммуникации, в мире печатной продукции и/или в отдельных группах текстов (ср. с этой точки зрения функции аннотаций, индексов типа ББК или ISBN, индексов цитирования, списков использованной литературы; ср. также задачи библиографических указателей, обзоров, роль включения текстов в циклы, собрания сочинений, антологии, сборники и под.). Укреплению автономности текстов диалектически противостоит тенденция преодоления этой автономности, и развитие гипертекстовой коммуникации соответствует последней.

Связанное с укреплением авторского начала (прежде всего в письменной коммуникации) отлучение получателей от непосредственного участия в создании текстов компенсируется тем, что при создании текстов авторы вынуждены учитывать интересы предполагаемых получателей и их коммуникативные возможности, отбирать соответствующие темы, уровни изложения, писать или говорить на языке, понятном получателям текстов, демонстрировать различными способами признаки, по которым потенциальные получатели могли бы самостоятельно выделить в коммуникативном потоке необходимые им тексты и познакомиться с ними. Другими словами, творя тексты «единолично», авторы вследствие именно этого принципа своей работы вынуждены действовать и «за себя», и «за адресатов». Кроме того, в массовой коммуникации начинают развиваться и активно эксплуатироваться такие формы приобщения адресатов к творению текстов, как прямой эфир, когда телезрители или радиослушатели могут задавать выступающим вопросы или сообщать свое мнение, как приглашение в студию представителей адресатов передачи, как интерактивное голосование и т. п.

Следовательно, участие получателей текстов в их создании никогда не прекращается, но меняются формы этого участия: прямое участие адресатов в творении текстов, каким оно бывает, например, в непосредственном устном общении, заменяется во многих типах опосредованной коммуникации участием косвенным, прежде всего через отражение в тексте коммуникативных интересов и возможностей предполагаемых адресатов.

Таким образом, **в коммуникации действуют противоположно направленные тенденции**, и если сопоставить с ними принципиальные особенности гипертекстовой коммуникации, отмеченные в начале раздела, то можно сказать, что **успех гипертекстовой коммуникации достигнут не вопреки, а в полном соответствии с общим направлением развития речевых средств общения.**

Этот вывод не только не умаляет значимости явления гипертекста и компьютерных средств его реализации, но, напротив, позволяет оценить высокую актуальность и лингвосемиотическую обусловленность расцвета гипертекстовой коммуникации.

3. Лингвистические средства системной организации структурированных гипертекстов. Гипертекстовая коммуникация использует технические, программные и лингвистические средства.

Техническая и программная стороны дела обеспечиваются конструкторами и программистами. Создаются более совершенные версии языка разметки гипертекста HTML (*HyperText Markup Language*), функционально развиваются браузеры, предлагаются новые протоколы передачи данных, обеспечивающие бóльшие скорости обмена информацией, неуклонно расширяется круг участников Интернет-общения.

Разрабатываются и лингвистические проблемы организации гипертекстов (создание оптимальных гипертекстовых композиций, анализ многообразных связей между текстами и фрагментами текстов, выбор связей, наиболее соответствующих задачам гипертекста, определение опорных текстовых элементов, оценка семантической структуры создаваемой конструкции и др.). Обсуждению этих проблем посвящаются специальные конференции, журнальные и монографические публикации. Ориентированный на лингвистов подробный

анализ гипертекстовой коммуникации представлен в книге Р.К. Потаповой «Новые информационные технологии и лингвистика (М., 2002). Одно из наиболее полных и четко структурированных пособий на русском языке по теории и практике лингвистической работы с гипертекстами (Рязанцева Т.И. Теория и практика работы с гипертекстом: На материале английского языка. – М.) выпущено в 2008 г. Издательским центром «Академия». Положения работы Т.И. Рязанцевой учтены в последующем изложении.

Обращаясь к лингвистическим средствам организации гипертекстов, **необходимо различать гипертексты неструктурированные и структурированные**. Организация первых не подчиняется предварительно разработанным планам и возникает в самом процессе включения в гипертекст всё новых составляющих его фрагментов (гипотекстов), количество и характер которых определяются степенью погружения авторов в разрабатываемые ими темы. Когда говорят о беспредельности развертывания гипертекстов, о системной непредопределенности реализуемых в них связей, о наращивании гипертекстов тематически близкими ему гипотекстами на базе их автоматической индексации, о полной подчиненности гипертекста интересам пользователя, то имеют в виду прежде всего неструктурированные гипертексты.

Неструктурированные гипертексты постоянно создаются, но не во всех случаях эффективны. Неэффективны они прежде всего при решении задачи представлении таких справочных и обучающих текстов, в которых самим материалом предопределены системные отношения между блоками сообщаемых знаний и наложены ограничения на развертывание гипертекста.

Основную лингвистическую проблему при создании структурированных гипертекстов составляет выделение тех опорных элементов, через которые совершается расширение гипертекста дополнительными фрагментами, поскольку каждый элемент текста в принципе может быть связан с другими текстами и не одним типом отношений.

Возьмем, например, сообщение, относящееся к уже упоминавшимся индейцам оджибве (см. выше): «*Свои ритуальные песни индейцы оджибве в мнемонических целях записывали рисунками на бересте»*. В каких направлениях могло бы расширяться данное сообщение?

Во-первых, здесь возможны связи энциклопедического характера: сведения об индейцах вообще, об индейцах оджибве в частности (их всего около 30 тыс., живут в США и Канаде, в основном католики и т.д.), о мнемонических средствах (как происходит запоминание и воспроизведение запомненного, какие существуют приемы запоминания, какова степень их эффективности и т.д.), о возможности рисовать на коре березы, о песнях, ритуалах.

Во-вторых, расширение данного сообщения может идти по линии прагматической: откуда известно, что свои ритуальные песни индейцы оджибве в мнемонических целях записывали рисунками на бересте; где источник данной информации; насколько информация достоверна; является ли она новой, полезной, как можно ее использовать и т.д.

Рассматриваемая фраза содержит основания и для связей собственно лингвистического характера: каковы значения (лексические, грамматические) всех слов данного высказывания, каковы их стилистические характеристики, происхождение, какие особенности имеет произношение данных слов (так, нормативные словари разрешают произношение *берёста* и *береста* и отвергают вариант *берёсто*, который существовал в древнерусском языке и сохраняется в ряде современных говоров).

Все эти направления расширения исходного текста могли бы быть реализованы в неструктурированном гипертексте. Если же данная фраза включена в структурированный гипертекст, то ее **связи должны определяться характером планируемого целого**: темой и целями всего гипертекста, поэтому первый этап создания структурированного гипертекста – это определение его темы и цели, при этом одну из сторон целевой направленности текста составляет представление о его адресате.

Если, например, планируется собрать в гипертексте все, что известно о березе, то есть собрать гипертекстовую энциклопедию березы, то придется создавать гипотексты о березе как растении, о разновидностях березы и их распространении, о различном использовании древесины березы, ее коры, сока, ветвей и листьев, о березе в ритуалах, о березе как символе, об

образе березы в художественных произведениях, о научных исследованиях, посвященных березе, и т. п. (ср.: Мартинович 2008) При этом каждый гипотекст должен быть относительно цельным и законченным.

Заметим, что явления текстовой цельности и законченности получили в лингвистике и психолингвистике множество интерпретаций (см., например: Овчинникова 1998). Общим для большинства из них является представление о тесной связи и даже взаимообусловленности данных характеристик текста. Они отражаются в восприятии текста как подчиненного ведущей теме и имеющего иерархически организованный смысл, поддающийся компрессии до уровня одного высказывания, выражающего тему текста. Поэтому проверкой цельности и законченности текста может служить возможность озаглавить текст. Н.И. Жинкин утверждал: «Во всяком тексте, если он относительно закончен и последователен, высказана одна основная мысль, один тезис, одно положение. Все остальное подводит к этой мысли, развивает ее, аргументирует, разрабатывает» (Жинкин 1956).

Вне контекста рассматриваемое высказывание («*Свои ритуальные песни индейцы оджибве в мнемонических целях записывали рисунками на бересте»*) воспринимается как функционально неопределенное, и его содержание трудно определить кратким заголовком. Положение меняется, если включить данное высказывание в контекст хотя бы одного абзаца:

> «*Свои ритуальные песни индейцы оджибве в мнемонических целях записывали рисунками на бересте. Эти песни как правило исполняются при посвящении новообращенных в тайные культовые сообщества. Слова этих песнопений передаются из поколения в поколение без всяких изменений, так что многие из них уже давно устарели и не входят в лексику живого разговорного языка. Их не всегда понимают даже самые лучшие певцы-шаманы. Но ни одному индейцу не дозволено менять текст этих древних песен, так как это лишило бы их приписываемой им магической силы»* (По И.Е. Гельбу, с. 52)

Здесь сразу становится понятным, что речь идет не о березе, а об индейцах оджибве и их ритуальных песнях. Появляется возможность озаглавить этот относительно цельный фрагмент текста: «Ритуальные песни индейцев оджибве». Такой фрагмент уже может рассматриваться как кандидат на роль гипотекста в соответствующем гипертекстовом построении. Свое имя в качестве выражение цельности и законченности должен получить и весь структурированный гипертекст.

На объем гипотекстов обычно налагается ограничение: в каждый из них должно входить не более 7-9 высказываний, занимающих вместе с оформлением и различными иллюстрациями не более одного экрана (в разбираемом нами примере 7 предикативных единиц). Однако целесообразность данного правила не абсолютна, поэтому оно далеко не всегда выдерживается.

Когда гипотексты, отражающие различные стороны и блоки знания, для сообщения которого предназначается гипертекст, определены, наступает следующий (второй) этап работы: установление характера связей между гипотекстами в составе гипертекста. В неструктурированных гипертекстах они могут быть любыми, в том числе ассоциативными, отражающими структуру связей в памяти: связь между гипонимами и гиперонимами, между словами сходного значения, словами противоположного значения, между членами одной грамматической парадигмы, между словами, содержание которых относится к одной теме, между сходно звучащими словами и др.. В структурированных же гипертекстах это преимущественно логические и логико-композиционные связи. Типичные примеры связей между гипотекстами на логической основе:

положение и пример (образцом такой связи в рассматривавшемся примере могло бы быть присоединение к тексту о ритуальных песнях блока с записью конкретной песни),

тезис и доказательство (например, присоединение текста с историей неудачной попытки сделать ритуал понятнее),

тезис и его уточнение (добавление блока о том, в какой мере стали непонятными тексты ритуальных песен)

описание явления и сообщение о сходных, аналогичных явлениях,

описание события и раскрытие его причины,

описание события и конкретизация сведений о его участниках и соответствующих обстоятельствах (место, время, сопутствующие события, предсобытие, постсобытие),

термины и раскрытие их содержания

и др. (подробнее см.: Рязанцева 2008).

К логико-композиционным связям относятся связи основного текста и введения, основного текста и различных комментариев к нему, основного текста и заключения, списка использованной литературы, различных приложений.

Рис. 2. Фрагмент новостной ленты (Lenta.ru)

С вопросом о типах отношений между гипотекстами тесно связан вопрос о том, как именно должны размещаться в тексте ссылки на присоединяемые гипотексты. Специалисты по дизайну гипертекстов советуют не делать ссылок слишком много, чуть ли не от каждого слова или словосочетания: чем из большего количества путей чтения придется каждый раз выбирать свой путь пользователю, тем меньше останется у структурированного гипертекста его коммуникативных преимуществ. Безусловно, бессмысленны ссылки от служебных и (в большинстве случаев) от местоименных слов: гиперссылки этого типа (ср.: ***Свои*** *ритуальные песни индейцы оджибве* ***в*** *мнемонических целях записывали рисунками* ***на*** *бересте*) не подсказывают содержания скрывающихся за ними гипотекстов и, скорее всего, будут проигнорированы. С этой точки зрения предпочтительнее гиперссылки от знаменательных слов, словосочетаний или даже предикативных единиц.

Рассмотрим Рис 2. На нем представлен фрагмент новостной ленты сайта Lenta.ru. Дизайнер страницы явно предпочитает ссылки, хорошо выражающие содержание присоединяемого текста с изложением новости. Кроме гиперссылок от каждой картинки, мы находим здесь гиперссылки от предикативных единиц (Google анонсировал ускоритель Интернета, Microsoft пришлось оправдываться, Кремль подал первую заявку), от групп сказуемого, поскольку именно эти группы обычно содержат новое (оказался поклонником Интернета, создал новый язык программирования, решил увеличить свою долю в Rambler Media, ушел из Yahoo!, оказался в 15 раз меньше всей Сети, закончил тестировать новый поисковик), от именных групп, представляющих собой номинацию лица, дающего пресс-конференцию, и являющихся одновременно подписью к фотографии этого лица (Антон Волнухин, руководитель службы «Поиск по блогам» Яндекса, Николай Прянишников, президент Mikrosoft в России).

Окончательная сборка гипертекста осуществляется в соответствии с предварительно разработанной его моделью. Наиболее простой вид имеют иерархические модели с последовательным подчинением одних гипотекстов другим. При выборе модели важно предусмотреть, чтобы не только создатель структурированного гипертекста имел разработанную модель и руководствовался ею при сборке, но и чтобы для пользователя она оказалась достаточно прозрачной, чтобы пользователь мог составить себе общее представление о строении предлагаемого ему продукта и руководствоваться этим представлением в процессе навигации. Иерархическую структуру гипертекста нетрудно представить в текстовой форме как оглавление или в более абстрактном виде как граф, вершины которого соответствуют гипотекстам, а дуги – связям между гипотекстами.

Итак, рассматривая гипертекст с лингвистических позиций, мы, во-первых, установили историческую обусловленность развития гипертекстовых структур, их соответствие сложной, противоречивой природе письменной коммуникации и, во-вторых, – продемонстрировали, хотя и в самом общем виде, необходимость собственно лингвистического структурирования гипертекстов как одного из необходимых условий их эффективности.

Рекомендации и задание. Одним из важных практических применений гипертекстовых структур лингвистами является создание ими гипертекстовых учебных пособий. См., например, гипертекстовое пособие М.Л. Ремневой и О.В. Дедовой по курсу «Старославянский язык» (Ремнева, Дедова 2002)

Пособия, выполняемые в виде гипертекстов, обычно включают определенные типы материала: сообщения теоретического характера (научные утверждения), доказательства истинности сообщений, иллюстративные материалы различной модальности (текстовые, графические, звуковые и др.), терминологические справки и справочники, изложение истории отдельных разделов науки, задачи и упражнения, списки рекомендуемой литературы, учебные хрестоматии, персоналии, вопросы для самопроверки, тесты и под. Таким образом, можно говорить о существовании универсальных элементов структуры учебного текста, которые получают в гипертекстах свое наиболее явное автономное воплощение. Следствием их существования является то, что попытки построить гипертекст хотя бы для одной небольшой темы обычно ведут к более четкому осознанию различных ее сторон и связей.

Рекомендуем Вам проверить это утверждение, выполнив анализ текста статьи ОПРОЩЕНЬЕ «Лингвистического энциклопедического словаря» (ЛЭС). Ниже эта статья приведена полностью. Курсивом составители словаря выделяют отсылки к другим статьям данного издания):

ОПРО́ЩЕНИЕ (деэтимологизация) — лексико-морфологич. явление, состоящее в затемнении первонач. семантич. структуры слова вследствие стирания морфо-логич. границ между его компонентами, т. е. в результате превращения прежде членимой основы в нечленимый корень, ср. рус. «воздух», «запах» и т. п., нем. zurück 'назад', Vorrat 'запас'. Термин введен В. А. Богородицким. О. может вызываться фонетич. изменениями в процессе ист. развития языка и приводить к утрате прежней связанности однокоренных слов, напр. рус. «конец» и "на-чало», восходящие к одному и тому же индоевроп. корню *ken/*kon и имеющие этимологич. слав. суффиксы -ьць, -ло и префикс на-, воспринимаются как непроизводные и морфологически не связанные друг с другом. О. часто встречается при пиджинизации и креолизации (см. *Пиджины, Креольские языки)*. О. связано с фузионным соединением морфем (см. *Фузия)* и может сопровождаться предварительным *переразложением.* Характерно для флективных языков, но возможно и в языках иной структуры, затрагивая этимологически сложные слова.

Б о г о р о д и ц к и й В. А., Лекции по общему языковедению, 2 изд., Каз., 1915.

В.А. Виноградов

Рассмотрите текст и сформулируйте ответы на следующие группы вопросов:

1) Какие приемы преодоления текстовой автономности использованы в данной статье словаря? Можно ли выявить принцип выбора опорных единиц текста, от которых сделаны ссылки к другим статьям словаря? Влияет ли на характер ссылок наличие в

ЛЭС ряда приложений (Терминологического указателя, Указателя языков мира, Аннотированного именного указателя)?

2) Обладает ли текст данной статьи характеристиками, позволяющими использовать его в качестве одного из гипотекстов гипертекста «Исторические изменения морфемного состава слов»? Каковы эти характеристики?
3) Если бы текст данной статьи был включен в состав учебного гипертекста «Исторические изменения морфемного состава слов» (для студентов-филологов) в качестве одного из его гипотекстов, то какого рода его связи с другими гипотекстами целесообразно было бы предусмотреть? Какие другие гипотексты, связанные с рассматриваемым, потребовалось бы создать? Можно ли их оценить по шкале обязательности / факультативности?
4) Могли бы Вы теперь сконструировать достаточно эффективный учебный гипертекст «Исторические изменения морфемного состава слов» для студентов-филологов?

Литература

Основная:

Захаркина В.В. Язык структурной разметки гипертекста HTML. – СПб., 2007.

Потапова Р.К. Новые информационные технологии и лингвистика. – М., 2002.

Рязанцева Т.И. Теория и практика работы с гипертекстом: На материале английского языка. – М., 2008.

Дополнительная:

Гельб И.Е. Опыт изучения письма (основы грамматологии). – М., 1982.

Гольдин В.Е. Обращение: теоретические проблемы. – М., 2009.

Гуманитарные исследования в Интернете / Под ред. А.Е. Войскунского. – М., 2000.

Жинкин Н.И. Развитие письменной речи у учащихся III – VI классов // Изв. АПН РСФСР, 1956. № 78.

ЛЭС – Лингвистический энциклопедический словарь. – М., 1990

Мартинович Г.А. Текст и эксперимент: исследование коммуникативно-тематического поля в русском языке. – СПб., 2008

Овчинникова И.Г. О феномене цельности текста // Фатическое поле языка (памяти профессора Л.Н. Мурзина. – Пермь, 1998.

Ремнева М.Л. Старославянский язык. Учебное пособие. – М., 2004. Приложение: Электронный учебный курс: *Ремнева М.Л., Дедова О.В.* Старославянский язык. – CD.

Субботин М.М. Гипертекст как новая форма письменной коммуникации // Итоги науки и техники. Сер. Информатика. 1994. Т. 18.

Лекция 3. Текстовые корпуса русской речи

1. Структура текстовых корпусов. Аннотирование.
2. Текстовые корпуса русского языка.
3. Национальный корпус русского языка.
4. Лингвокультурологические корпуса как современное системное представление коммуникации. Саратовский диалектологический корпус.
5. Научные идеологические новации и эвристический потенциал современных текстовых корпусов.
6. Рекомендации и задания.
7. Литература.

1. Структура текстовых корпусов. Аннотирование. Под ***текстовым корпусом*** в корпусной лингвистике понимается структурированный, размеченный массив текстов или их значительных фрагментов, представленный в электронном виде и обеспеченный специализированной поисковой системой. Текстовые корпусы могут быть предназначены для решения различных лингвистических задач. Цель построения корпуса определяет его тип. Типы корпусов (фундаментальные корпусы текстов; динамические/мониторные vs. статические корпусы; исследовательские vs. иллюстративные корпусы; авторские корпусы) рассматриваются в [Баранов 2007; см. также: Баранов 2001; Захаров 2005].

Важнейшим общим принципом формирования текстовых корпусов является их ***репрезентативность***, которая определяется не только и не столько количеством языкового материала, но прежде всего его пропорциональностью. Например, в корпусе текстов одного писателя необходимо пропорциональное представление произведений различных периодов его творчества, жанров, в корпусе текстов русской художественной литературы – пропорциональность текстов различных периодов, жанров, стилей, авторов. Вот как, например, характеризует А.Н. Баранов стратегию формирования создаваемого в отделе экспериментальной лексикографии Института русского языка РАН динамического корпуса текстов по современной публицистике: «В плане репрезентативности основное внимание было обращено на выбор периодических изданий различной ориентации наиболее важных для общественного сознания в исследуемый период и на соблюдение пропорции, учитывающей значимость и популярность последних» [Баранов 2001: 131]. В пособии также подробно изложена принятая разработчиками корпуса методика выделения параметров для описания текстового массива современной публицистики (факторы автора текста, адресата, прагматических условий порождения текста и др.), позволяющая обеспечить репрезентативное представление данной предметной области в корпусе [см.: Баранов 2001: 132-133].

В идеале корпус – это уменьшенная модель языка или подъязыка: частота языкового явления в корпусе должна быть близка его частоте в отражаемой корпусом предметной области. Так, например, созданный на основе выбранных параметров корпус текстов современной публицистики «может рассматриваться как модель функционирования языка современной публицистики в дискурсе [Баранов 2001: 135]. Однако задача построения такой модели не всегда полностью выполнима. Например, невозможно с точностью определить пропорции различных видов (тем, жанров) устной – литературной и внелитературной – коммуникации. Поэтому при построении корпусов устной речи моделирование соответствующей предметной области принимает нежесткую, вероятностную форму. Тем не менее принцип репрезентативности не теряет своей актуальности и в этом случае как необходимый при построении любого текстового корпуса ориентир. «Если нет уверенности в

представительности корпуса, – пишет А.Н. Баранов, – его заведомо нельзя использовать для многих видов лингвистической деятельности, например, для оценки частоты употребления лексем в тех или иных значениях или для составления словников некоторой проблемной области» [Баранов 2001: 136]. Репрезентативность – важнейший принцип не только для проектирования корпуса, но и для использования корпусных данных, составления исследовательской выборки. Различные способы достижения репрезентативности корпуса обсуждаются в [Баранов 2007].

Другим отличительным свойством корпуса текстов является ***разметка (аннотирование)*** текстового массива. Разметка отличает корпус от других коллекций («библиотек») текстов. Разметка заключается в приписывании включенным в корпус текстам и их компонентам специальных меток (тэгов – от англ. tag ‘ярлык, метка’). Так, тэгами, кодирующими целый текст, могут быть имя автора, его возраст, пол, место жительства, год создания или записи текста (в зависимости от его первоначальной формы – письменной или устной), название, жанр, тематика и др. Все перечисленные параметры являются элементами метаразметки, передающими внешнюю, экстралингвистическую информацию о тексте.

При собственно лингвистической кодировке текстов в корпусах применяются различные виды разметки. В.П. Захаров приводит следующие виды лингвистической разметки текстовых корпусов:

- *морфологическая* разметка, в ходе которой обозначаются части речи всех словоформ размечаемого текста и признаки грамматических категорий, свойственных данной части речи;
- *синтаксическая* разметка является результатом синтаксического анализа (парсинга – от англ. parsing). Чаще всего в его основе лежит грамматика структур непосредственно составляющих. Этот вид разметки описывает синтаксические связи между лексическими единицами и различные синтаксические конструкции (например, придаточное предложение, глагольное словосочетание и т.п.);
- *семантическая* разметка. Семантические тэги обозначают обычно семантические категории, к которым относится данное слово или словосочетание, и более узкие подкатегории, специфицирующие его значение;
- *анафорическая* разметка. Фиксирует референтные связи, например, местоименные;
- *просодическая* разметка. В просодических корпусах применяются метки, описывающие ударение и интонацию. В корпусах устной разговорной речи просодическая разметка часто сопровождается так называемой *дискурсной* разметкой, которая служит для обозначения пауз, повторов, оговорок, и т.д.;
- *структурная* разметка, результатом которой является сегментирование текста на его структурные составляющие, например, абзац, реплики диалога, предложения, слова.

В настоящее время не существует общепризнанных стандартов представления лингвистической и других видов информации в текстах. Основным типом разметки в настоящее время является морфологическая разметка: большинство крупных корпусов являются морфологически размеченными корпусами. Неравномерный охват разных языковых уровней в разметке корпусов связан с состоянием теоретической разработки соответствующей области языкознания, со степенью ее формализации, наличием общепринятого стандарта описания. Морфологическая разметка опирается на принятый в качестве стандарта «Грамматический словарь русского языка» А.А. Зализняка. Для других видов разметки такого общепринятого стандарта нет.

Некоторые виды разметки (например, морфологическая, синтаксическая) проводятся с помощью специальных программ (тэггеров и парсеров). В результате работы программ автоматического морфологического анализа каждой словоформе приписываются грамматические характеристики, включая часть речи, лемму (начальную форму) и набор граммем (например, род, число, падеж, одушевленность/неодушевленность, переходность и т.п.). С помощью синтаксических парсеров (программных средств автоматического синтаксического анализа) фиксируются синтаксические связи между словами и словосочетаниями, а синтаксиче-

ским единицам приписываются соответствующие характеристики (тип предложения, синтаксическая функция словосочетания и т.п.). Компьютерным технологиям разметки корпусов посвящена статья А.Е. Полякова в сборнике [Национальный корпус… 2005].

Первичная автоматическая разметка не может быть абсолютно точной ввиду характерной для естественных языков грамматической омонимии. Ср. примеры автоматической разметки текстовых словоформ:

стоят{стоить=V,несов=изъяв,непрош,мн,3-л|стоять=V,несов=изъяв,непрош,мн,3-л};

людей{человек=S,муж,од=мн,род|человек=S,муж,од=мн,вин}.

Снятие грамматической омонимии (выбор релевантной для данного текста формы) проводится вручную или автоматически (с помощью специальных программ-редакторов).

Чем разнообразнее применяемая в корпусе разметка, тем выше его ценность как научного источника. По словам В.А. Плунгяна, корпусная лингвистика – «это прежде всего наука о том, как сделать хорошую разметку корпуса» [Национальный корпус… 2005: 6].

Электронная форма представления текстов, их разметка и обеспечение корпуса специализированной поисковой системой (типы программ, обрабатывающих текстовые массивы корпусов описаны в [Баранов 2001; Баранов 2007]) предоставляют пользователю разнообразные возможности быстрого и эффективного получения из корпуса необходимой информации. В.П. Захаровым выделены, например, следующие виды пользовательских запросов :

- поиск конкретных словоформ;
- поиск словоформ по леммам (по их словарной форме);
- поиск группы словоформ в виде разрывной или неразрывной синтагмы;
- поиск словоформ по набору морфологических признаков;
- отображение информации о происхождении, типе текста и т.п.;
- вывод результатов поиска с указанием контекста заданной длины;
- получение различных лексико-грамматических статистических данных;
- сохранение отобранных строк конкорданса в отдельном файле на компьютере пользователя и др.

Результаты поиска обычно выдаются в виде конкорданса – набора контекстов, включающих искомую единицу.

В настоящее время в разных странах созданы различные по своему предназначению и объему текстовые корпусы, часть из них функционирует в сети Интернет. В [Баранов 2001] и [Захаров 2005] содержится характеристика ряда корпусов текстов различных языков, приведены некоторые сетевые адреса текстовых корпусов.

2. **Текстовые корпуса русского языка.** Создание корпусов текстов русской речи началось сравнительно недавно. Некоторые из созданных и создающихся русских текстовых корпусов названы и охарактеризованы в [Баранов 2001; Захаров 2005]. Подробная сравнительная характеристика общедоступных[1] корпусов современного русского языка дается в статье Т.И. Резниковой и М.В. Копотева в [Национальный корпус… 2005].

Среди русских текстовых корпусов есть как корпусы, стремящиеся отразить состояние русского языка в целом на современном этапе его существования, так и корпусы, обращенные к отдельным его явлениям и подсистемам.

[1] Не все созданные или создающиеся корпуса доступны сегодня для широкого круга пользователей.

Наиболее представительным из русских текстовых корпусов первой группы (стремящихся отразить состояние русского языка в целом) является сегодня ***Национальный***[2] ***корпус русского языка***: http://ruscorpora.ru/ (подробнее о нем речь пойдет ниже)**.** К этой же группе корпусов можно отнести, также такие корпусы, как:

– Тюбингенские корпусы русских текстов: http://www.sfb441.uni-tuebingen.de/b1/rus/korpora.html;

– Корпус русского литературного языка: http://www.narusco.ru/resourses.htm **.**

Тюбингенские корпусы русских текстов (ТК), созданные в Тюбенгенском университете (Германия) под руководством проф. Т. Бергера, стали первыми корпусами русского языка открытого доступа в сети Интернет. В основу ТК лег первый электронный корпус русского языка – Уппсальский корпус современных русских текстов, составленный в Институте славистики Уппсальского университета (Швеция) под руководством проф. Леннарта Лёнгрена. Уппсальский корпус включает тексты художественной прозы и публицистики 1960-х – 1970-х гг. на русском языке объемом 1 млн. словоупотреблений. В составе ТК Уппсальский корпус образует отдельный подкорпус. Кроме этого, в ТК представлен корпус текстов интервью по темам "общество и политика", "экономика", "музыка", "литература", "молодежь" и "спорт", опубликованных в доступных в Интернете русских журналах и газетах с 1996 г. (объем корпуса постоянно увеличивается); подкорпус текстов статей из журнала «Огонек» за 1996–2002 гг. (объем более 9 млн. словоупотреблений); подкорпуса произведений целого ряда писателей XIX и XX вв. (общий объем подкорпусов художественной литературы свыше 14 млн. словоупотреблений). Поиск по словоформе или по произвольной ее части (по маске) в ТК может вестись по всему корпусу или по отдельному подкорпусу; контекст выдачи может расширяться от одного предложения до трех. В ТК реализована опция, позволяющая различать в поиске заглавные и строчные буквы; эта функция дает возможность оптимизировать выборку названий или словоформ, находящихся в начале предложения. Для корпуса текстов интервью предусмотрены дополнительные поисковые опции метатекстового характера, содержащие библиографические данные о тексте интервью.

Корпус русского литературного языка создается сотрудниками Санкт-Петербургского государственного университета при участии сотрудников Института лингвистических исследований РАН (СПб). Общее руководство проектом осуществляют Л.А. Вербицкая и В.Б. Касевич. В настоящее время в корпус входят только письменные тексты (как подчеркивают разработчики корпуса – «пока»), опубликованные признанными – официально зарегистрированными – издательствами. Текстовая база корпуса охватывает период с середины 20 в. и до настоящего времени. Размещенная в Интернете предварительная версия корпуса содержит ок. 1 млн словоупотреблений. В перспективе предполагается довести объем корпуса до 100 –150 млн словоупотреблений. Тексты корпуса примерно в равных объемах представляют художественную литературу (прозу), публицистику, драму, и научную (научно–популярную) литературу. На базе корпуса создан (частотный) словарь словоформ русского языка, насчитывающий около 125 тыс. единиц. Выдача производится с разбиением контекстов (при поиске по текстам корпуса) и данных о частотности словоформ (при поиске по словарю словоформ) по названным жанрам (См. Рис. 1.). В текстах и словаре все словоформы имеют акцентную разметку (отмечены основные и вторичные ударения); используется буква «ё».

[2] В европейской традиции корпусной лингвистики слово «национальный» в названиях корпусов используется как указание на самый большой и представительный корпус, характеризующий язык данной страны (напр.: Британский национальный корпус, Чешский национальный корпус). Представительность национального корпуса обеспечивается, помимо его значительного объема, включением в корпус всех типов текстов, порождаемых на данном языке в данный исторический период, и при этом содержать их в правильной пропорции [Национальный… 2005: 8].

Рис. 1. Поиск в словаре словоформ

Знак "+" — основное ударение;
Знак "^" — вторичное ударение

Допускается использование минуса (как дефиса).

Параметры поиска

- ☐ «Е» и «Ё» эквивалентны
- ☑ Искать и как часть составной словоформы
- ☑ Игнорировать ВЕРХНИЙ/нижний регистр
- ☑ Искать не только указанные ударения

Словоформа	**Число вхождений по жанрам** (в скобках приведено общее число словоформ в корпусе)				
	Драматургия (196107)	**Беллетристика** (354618)	**Публицистика** (303110)	**Научно-популярная литература** (198980)	**Всего** (1052815)
Интерне+т	0	0	15	0	15

Вторую группу корпусов (обращенных к отдельным явлениям и подсистемам русского языка) представляют такие корпуса, как:

– Компьютерный корпус газетных текстов русского языка конца XX века (КГТ): http://www.philol.msu.ru/~lex/corpus/;

– Корпус российских газет 90-х гг. XX в.: http://cfrl.ru/newspap.shtm;

– Хельсинкский аннотированный корпус (ХАНКО): http://www.ling.helsinki.fi/projects/hanco/;

– Словарь-конкорданс публицистики Ф.М. Достоевского: http://dostojevskij.karelia.ru/author.phtml;

– Конкордансы произведений Ф.М. Достоевского: http://petrsu.ru/~Dostoevsky/;

– Словарь языка Грибоедова: http://feb-web.ru/feb/concord/abc/;

– Корпус древнерусских берестяных грамот: http://gramoty.ru/index.php?key=bb;

– Параллельный корпус переводов «Слова о полку Игореве»: http://www.nevmenandr.net/slovo/.

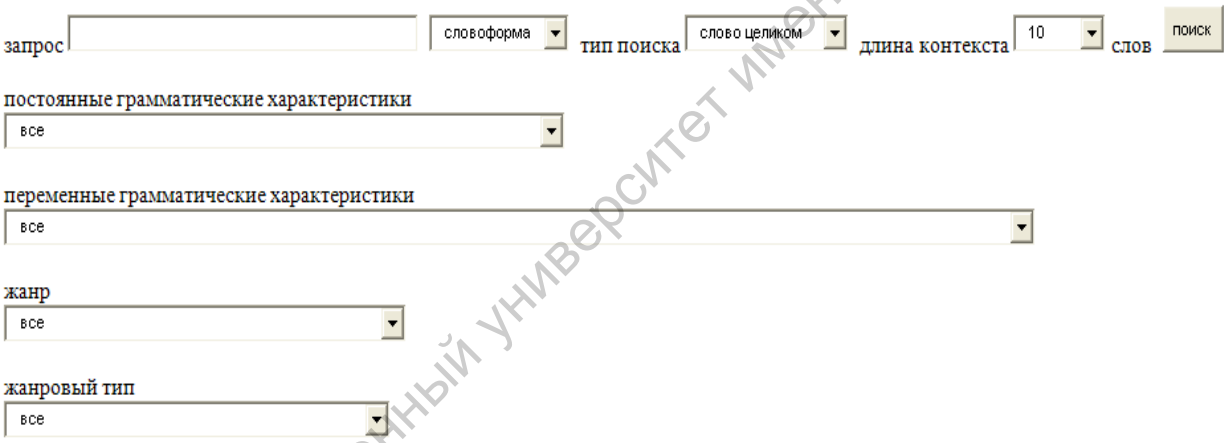
Компьютерный корпус газетных текстов русского языка конца XX века (КГТ) создан в 1999 г. и развивается в настоящее время в Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ под руководством заведующего Лабораторией А.А. Поликарпова. В настоящее время в Интернете доступен демонстрационный фрагмент корпуса объемом более 200 тыс. словоупотреблений (в нем представлены 446 текстов

российских газет). Полная Интернет-версия корпуса (более 1 млн. словоупотреблений) готовится к представлению. В КГТ включены полные тексты 13 известных российских газет различного типа (ежедневных и неежедневных, "левых" и "правых", центральных и местных, общих и профессионально ориентированных) за 1994–1997 гг.

Пользователь имеет возможность вести поиск по лемме, словоформе, элементу слова, по выбираемым из списка грамматическим характеристикам (постоянным и переменным). Длина контекста выдачи по запросу пользователя может варьироваться от 10 до 50 слов; контекст выдачи может быть также расширен до целого текста. Поиск в корпусе можно ограничить жанровым типом (информационно-публицистический, официально-деловой, рекламный и др. типы текстов) и конкретным жанром (репортаж, интервью, обзор и т.п.). См. Рис. 2.

Кроме того, благодаря осуществляемым в корпусе специфическим видам разметки (проводится морфемная сегментация словоформы, отмечается его принадлежность к определенной частотно-ранговой группе), КГТ позволяет пользователю отбирать материал по заданной морфемной модели, по частотным характеристикам лемм. Однако эти опции в демонстрационной версии недоступны.

Рис. 2. Поисковое окно КГТ



База данных **Корпуса российских газет 90-х гг. XX в**., созданного в Машинном фонде Института русского языка им. В.В. Виноградова РАН, содержит тексты 9 известных российских газет («Известия», «Литературная газета», «Московский комсомолец» и др.) общим объемом около 7,5 млн. словоформ. Поиск в текстовой базе ведется по слову, группе слов, а также по шаблону, описывающему задаваемую морфемную структуру (напр. шаблон #[base=][suff=ниц][flex=] задает поиск слов с суфф. -ниц при любой основе и любой флексии). См. Рис. 3.

Рис. 3. Страница подкорпуса «Литературной газеты»
в корпусе российских газет 90-х гг. XX в.

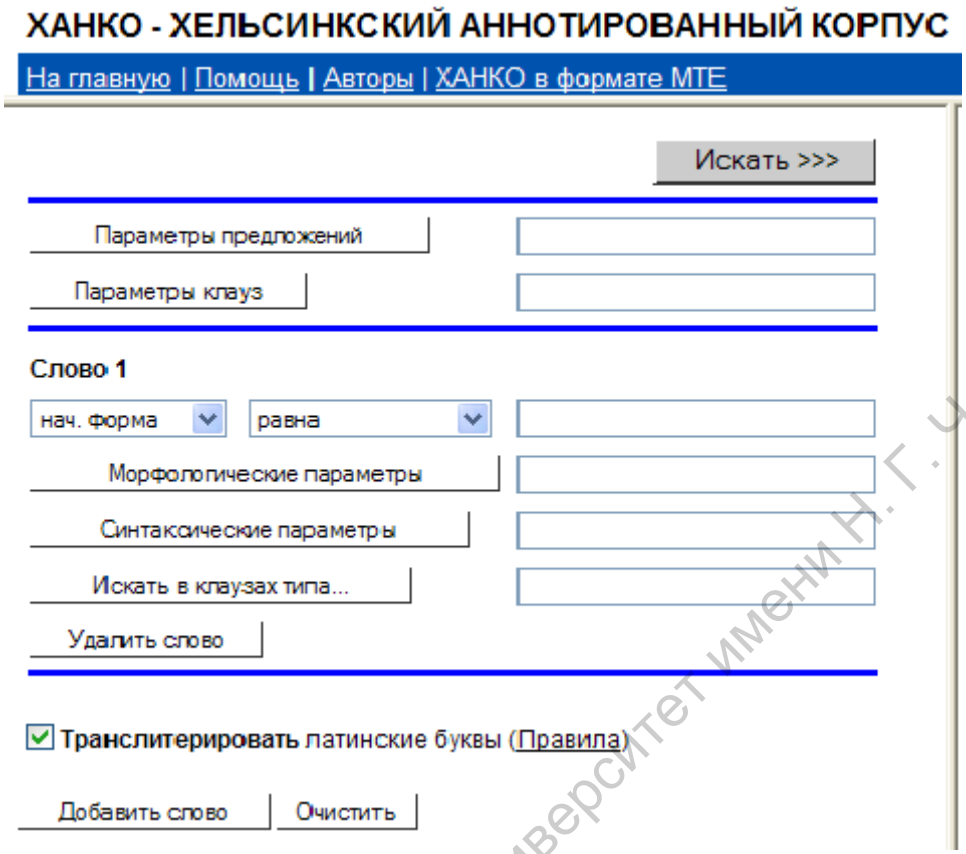
Литературная газета

Комплект	Архив	Текст	Поиск лексики	Комплект	Архив	Текст	Поиск лексики
Январь 1997 (274 Кб)	*Загрузить*	*Просмотреть*	*Искать лексику*	*Апрель 1997 (266Кб)*	*Загрузить*	*Просмотреть*	*Искать лексику*
Июль 1997 (497Кб)	*Загрузить*	*Просмотреть*	*Искать лексику*	*Август 1997 (432 Кб)*	*Загрузить*	*Просмотреть*	*Искать лексику*
Сентябрь 1997 (525 Кб)	*Загрузить*	*Просмотреть*	*Искать лексику*	*Октябрь 1997 (506 Кб)*	*Загрузить*	*Просмотреть*	*Искать лексику*
Ноябрь 1997 (512 Кб)	*Загрузить*	*Просмотреть*	*Искать лексику*	*Декабрь 1997 (393 Кб)*	*Загрузить*	*Просмотреть*	*Искать лексику*

Поиск лексики в комплекте «Литературной газеты», 1997 г.

Разработка ***Хельсинкского аннотированного корпуса*** (ХАНКО) ведется на отделении славянских и балтийских языков и литератур Хельсинкского университета под руководством проф. А. Мустайоки. ХАНКО – корпус текстов одного современного журнала «Итоги». Объем корпуса небольшой – 100 тыс. словоупотреблений; корпус отличает, по замыслу его создателей, «направленность на максимальный охват грамматической информации, а не на объем материала». В корпусе учтена возможность более чем одной интерпретации языковых фактов, отмечены все возможные варианты их интерпретации. В ХАНКО осуществлена морфологическая и синтаксическая разметка текстов; поиск ведется по начальной форме слова и конкретной текстоформе, по любым сочетаниям буквенных символов, входящих в состав начальной формы или текстоформы и занимающих в их составе различные позиции (ср. следующие параметры поиска: (текстофрма) «содержит», «начинается», «заканчивается» и т.д.), по морфологическим и синтаксическим признакам, выбираемым из числа предлагаемых в корпусе, в том числе и по сочетанию этих признаков. Для каждого контекста выдачи предусмотрена возможность его расширения и запрос синтаксического разбора. Все виды поиска могут быть ограничены типом предложения (простое, сложное и т.д.) и типом клаузы, характеризуемой по ее роли (самостоятельная и т.д.) и структуре (двусоставная и т.д.). См. Рис. 4.

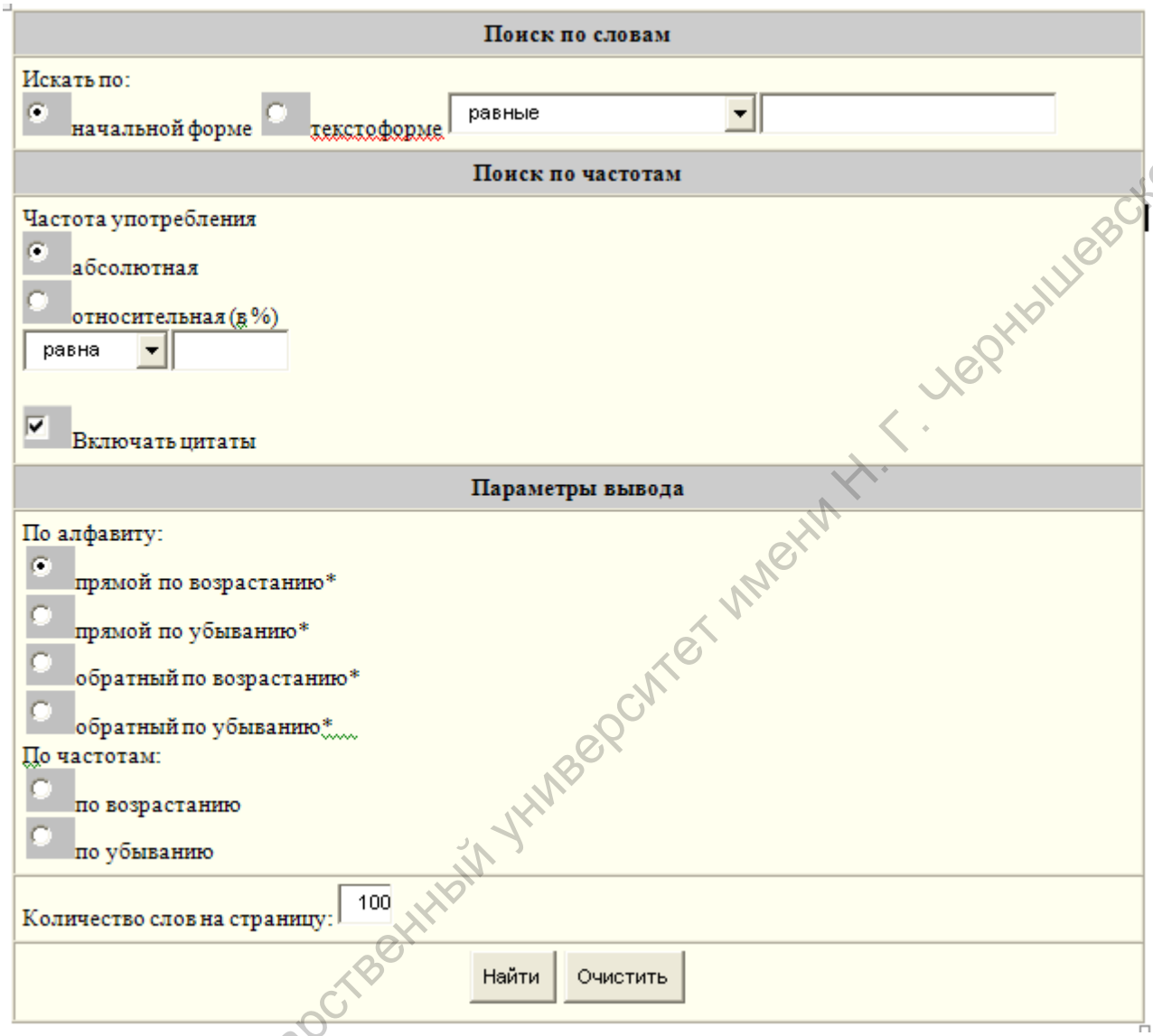
Рис. 4. Поисковое окно ХАНКО



Словарь-конкорданс публицистики Ф.М. Достоевского создан сотрудниками Петрозаводского государственного университета и Карельского государственного педагогического университета (руководитель коллектива разработчиков корпуса – М. В. Копотев) при участии сотрудников сектора экспериментальной лексикографии ИРЯ РАН (руководитель коллектива – Ю.Н. Караулов). Текстовой базой словаря-конкорданса являются тексты из раздела "Публицистика и письма" Полного собрания сочинений Ф.М. Достоевского в 30-ти томах. Словарь-конкорданс позволяет получить контекст употребления любого слова или словоформы (окружение в пределах 50-ти знаков до и после заданного слова), частотные характеристики леммы или словоформы, перейти к полному тексту произведения, а также узнать адрес этого произведения в полном собрании сочинений писателя (получить указание на том и страницу). В частотном словаре-конкордансе отдельно представлены частоты слов, принадлежащих собственно Ф. М. Достоевскому, и включенных автором в текст статьи как цитаты. Поиск может быть ограничен заданным пользователем кругом текстов или вестись по всему текстовому корпусу. См. Рис. 5.

Рис. 5. Поисковое окно

в Словаре-конкордансе публицистики Ф.М. Достоевского



Конкордансы произведений Ф.М. Достоевского размещены на сервере Петрозаводского государственного университета «Весь Достоевский». Поиск конкретной словоформы ведется в Конкордансах по выбранному из списка произведению. Результатом поиска являются все абзацы произведения, включающие данную словоформу, и указание на количество употреблений заданной словоформы в тексте. Задав выбранную из списка начальную двухбуквенную комбинацию пользователь может также получить полный список соответствующих словоформ с указанием их абсолютной частоты в данном тексте и, выбирая интересующие его словоформы переходить к их контекстам. См. Рис. 6.

Рис. 6. Окно выдачи на буквенную комбинацию «ап»
по роману «Преступление и наказание»

Ниже представлены слова, имеющие выбранное сочетание первых символов.
Выберите слово для просмотра его в контексте выбранного произведения

апатія (1)
апельсинничаешь (1)
апоплексію (1)
аппетита (2)
аппетитомъ (2)
аптеки (2)

Словарь языка Грибоедова содержит все слова, встретившиеся во всех опубликованных текстах Грибоедова, взятых из наиболее авторитетных источников.

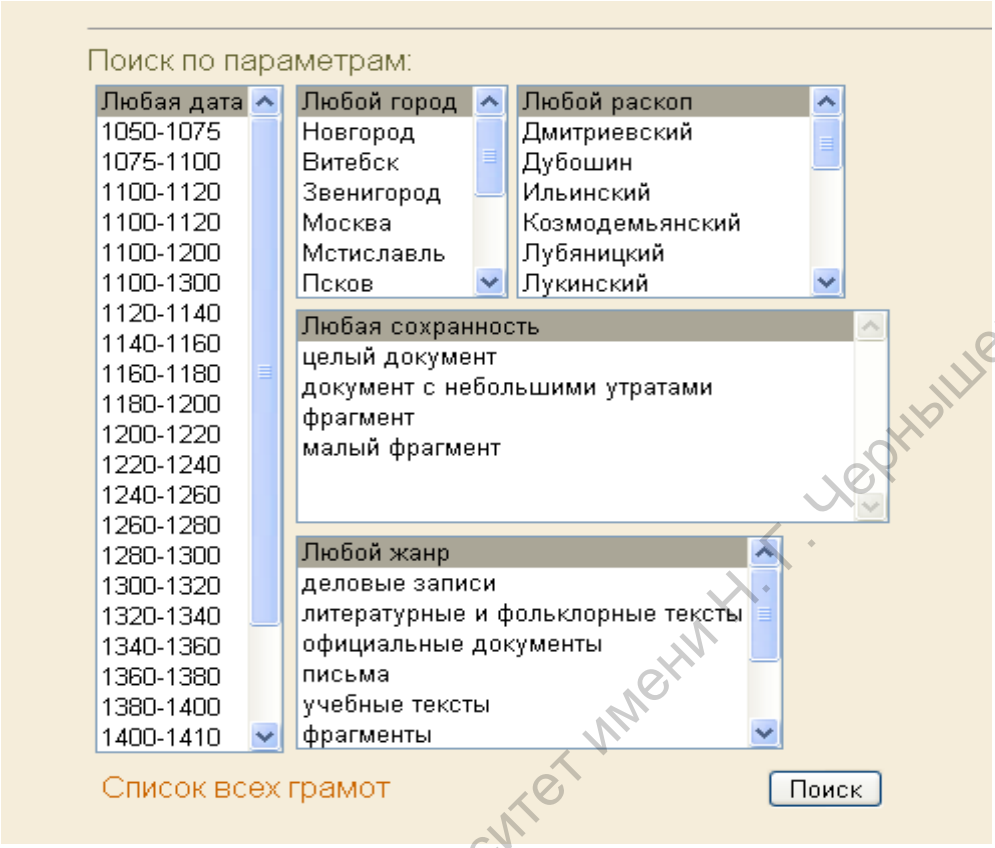
Основную часть словаря составляет алфавитно-частотный конкорданс к текстам А.С. Грибоедова, снабженный грамматической информацией. Конкорданс включает более 12 тыс. словарных статей, которые описывают в сумме более 150 тыс. словоупотреблений. В словарную статью входят лексема (заголовочное слово), ее грамматические признаки (часть речи, вид, род, одушевленность), суммарная частота по всем текстам, а при необходимости краткое толкование. Словоупотребления внутри статьи сгруппированы по грамматической форме (число, падеж, наклонение, время). Каждое словоупотребление содержит развернутый контекст и адрес, который включает код источника и его фрагмента (часть, глава, действие, явление и т.д.). См. Рис. 7.

Рис. 7. Окно выдачи по запросу на слово «басня»

басня *сущ.жен.неод.* (3)
ед.им.
и с тех пор ни одна сказка, ни **басня** не минует его рук Пс57.
мн.им.
ох! **басни** смерть моя! ГоУ 3.21.
мн.вин.
А если б, между нами, Был цензором назначен я, На **басни** бы налег ГоУ 3.21.

Корпус древнерусских берестяных грамот, размещенный на сайте «Рукописные памятники Древней Руси», представляет древнерусские грамоты на бересте XI–XV вв. Основой материалов корпуса стала книга: *А. А. Зализняк*. Древненовгородский диалект. 2-е изд., переработанное с учетом материала находок 1995—2003 гг. М., 2004, с добавлением данных о грамотах, найденных в 2004—2005 гг., а также серийное издание «Новгородские грамоты на бересте» (НГБ). База данных корпуса включает фотографии берестяных грамот, их прориси, оригинальные тексты грамот, их переводы на современный русский язык и основную историко-лингвистическую информацию о документах. Наполнение базы данных корпуса еще полностью не завершено. Сегодня корпус включает 1023 грамоты, обеспечивает поиск по параметрам «датировка грамот», «место нахождения грамот», «степень сохранности» (целый документ, фрагмент и т.д.), «жанр» (См. Рис. 8). Корпус содержит также библиотеку важнейших исследований берестяных грамот, карты мест их нахождения, библиографию трудов отечественных и зарубежных исследователей по проблематике берестяных грамот.

Рис. 8. Параметры поиска в Корпусе древнерусских берестяных грамот



Параллельный корпус переводов «Слова о полку Игореве» предназначен для сравнения переводов этого памятника древнерусской литературы. Весь текст «Слова…» разбит в корпусе на фрагменты в соответствии с делением, предложенным Р. О. Якобсоном. Выбрав нужный фрагмент и отметив интересующие его переводы из предложенного списка, пользователь получает заданный фрагмент «Слова…» и выбранные им переводы этого фрагмента. См. Рис. 9.

Рис. 9. Пример выдачи в Параллельном корпусе переводов «Слова о полку Игореве»

Параллельное представление фрагмента № 1

Источник	Текст
Древнерусский текст	Не лѣпо ли ны бяшеть, братіе, начяти старыми словесы трудныхъ повѣстій о пълку Игоревѣ, Игоря Святъславлича
Перевод Д. С. Лихачёва	Не пристало ли нам, братья, начать старыми словами печальные повести о походе Игоревом, Игоря Святослава?
Перевод В. А. Жуковского	Не прилично ли будет нам, братия, Начать древним складом Печальную повесть о битвах Игоря, Игоря Святослава!

3. Национальный корпус русского языка (НКРЯ). НКРЯ открыт в сети Интернет 29 апреля 2004 г. Текстовый массив корпуса охватывает период от начала XVIII до начала XXI в., при этом количественно преобладают в корпусе тексты современного периода – 2-й половины

XX – нач. XXI в. Этот период отражен в НКРЯ также и наиболее разнообразно по жанрам и типам речи. Объем НКРЯ в настоящее время – более 140 млн. словоупотреблений.

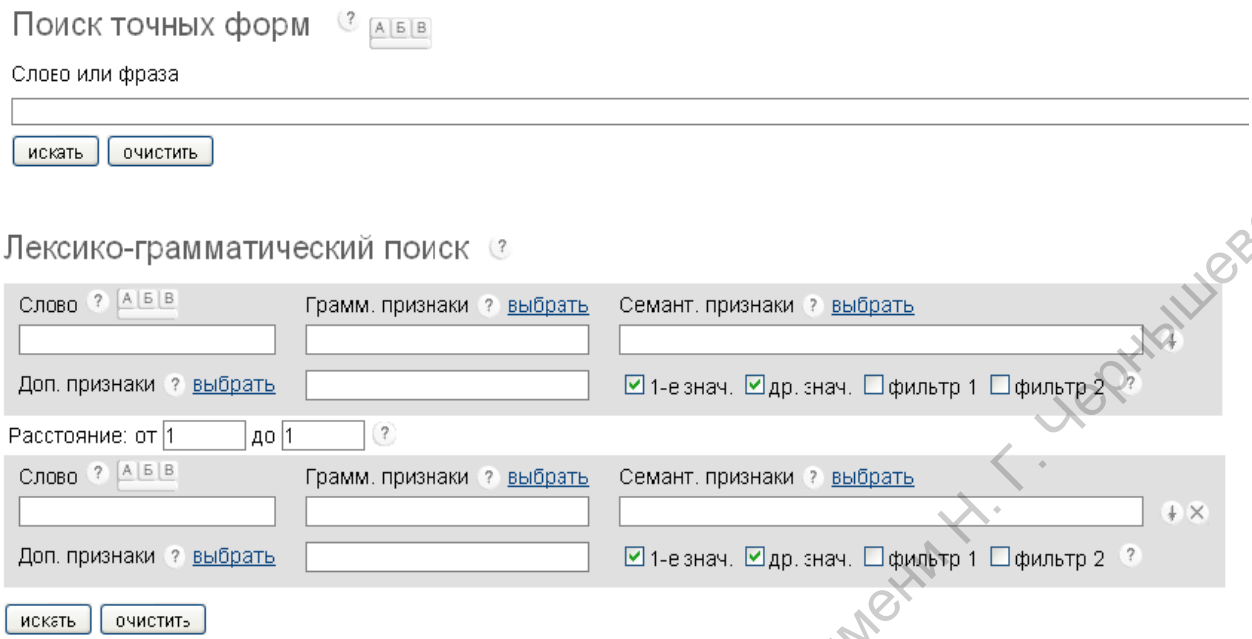
Русский язык представлен в НКРЯ в разных социальных формах его существования – литературной, разговорной, диалектной. Наиболее полно на сегодняшний день отражен в корпусе литературный вариант русского языка, который представлен значительным массивом художественных текстов разных жанров, другими видами письменной и (в меньшей мере) устной литературной речи: публицистика, научная и научно-популярная литература, частная переписка, дневники, документы, публичные выступления, газетные объявления и т.д. Корпус включает также небольшой пока подкорпус параллельных текстов – английских и русских, немецких и русских; планируется создание параллельных текстов и для других языков.

Специфика НКРЯ состоит в его принципиальной «нелитературоцентричности», хотя роль текстов классической и современной художественной литературы в корпусе достаточно велика. Такая установка продиктована представлением о том, что «учет именно этих текстов не является для многих задач приоритетным», стремлением представить в корпусе «образцы доминирующего в данном языковом коллективе дискурса». «На роль последнего, – пишет В.А. Плунгян, – может в современной ситуации претендовать скорее литература, относимая к жанру ‘non-fiction’, то есть литература с минимально декларируемой «художественностью», а также образцы устного городского фольклора: анекдоты, анонимные «истории из жизни», вербализующие стереотипы и мифы современного массового сознания» [Плунгян 2008: 13-14].

В настоящее время в НКРЯ используются метатекстовая, морфологическая, семантическая, акцентная разметки, разрабатывается синтаксическая разметка. Структура НКРЯ и система разметки в нем постоянно совершенствуются. Метатекстовой разметке в НКРЯ посвящена статья С.О. Савчук «Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции» [Национальный корпус… 2005]; проблемы морфологической разметки обсуждаются в статьях О.Н. Ляшевской, В.А. Плунгяна, Д.В. Сичинавы «О морфологическом стандарте Национального корпуса русского языка» [Там же], Д.В. Сичинавы «Обработка текстов с грамматической разметкой: инструкция разметчика» [Там же]; принципы семантической разметки изложены в статье Г.И. Кустовой, О.Н. Ляшевской, Е.В. Падучевой, Е.В. Рахилиной «Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы» [Там же]; перспективы синтаксического аннотирования корпуса представлены в статье коллектива авторов Ю.Д. Апресян и др. «Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы» [Там же].

Тексты, включенные в НКРЯ, не доступны для чтения и копирования как целые тексты. Они служат источниками примеров (контекстов), получаемых в результате поисковых запросов. Контекст выдачи равен одному предложению, однако по запросу пользователя может быть расширен до нескольких предложений. Каждая из текстоформ в контексте выдачи снабжена информацией о ее грамматических признаках (результат морфологической разметки), появляющейся в виде всплывающих окон. В НКРЯ предоставляется возможность поиска по слову, словоформе, словосочетанию, сочетанию слов, находящихся на определенном расстоянии друг от друга, по сегменту слова, по заданным грамматическим и семантическим характеристикам. В поисковый запрос могут быть включены также дополнительные признаки: повтор слова или грамматических характеристик, знаки препинания (находящиеся до или после запрашиваемого слова). См. Рис.10.

Рис.10. Поисковое окно в НКРЯ



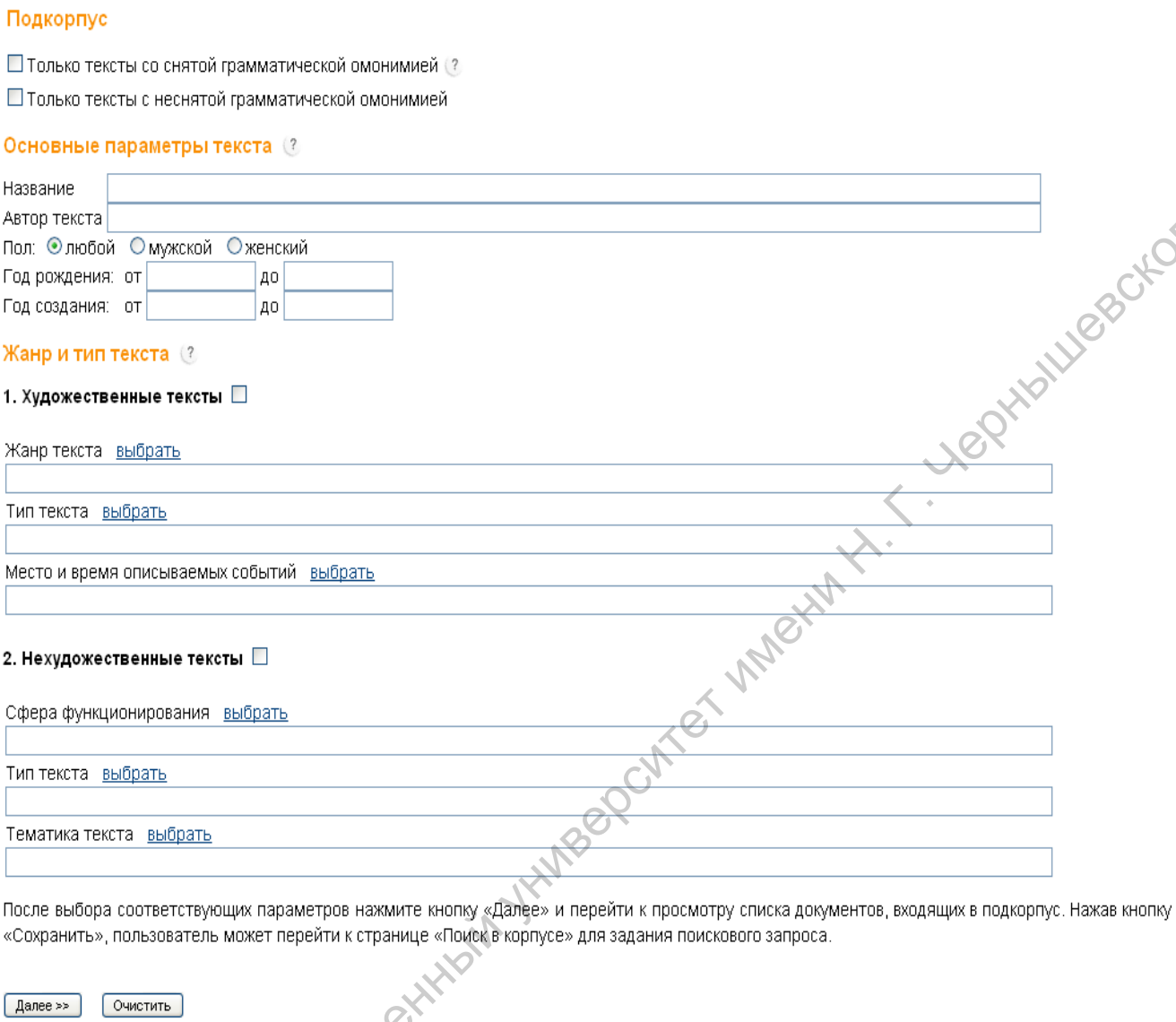
Поиск в НКРЯ может вестись как по целому корпусу, так и по определенному подмножеству текстов. Пользователь может, например, ограничить область поиска текстами определенного автора, определенного периода, определенного жанра, определенной тематики и т.п. Есть также возможность сформировать свой исследовательский подкорпус по принятым в НКРЯ параметрам его структурирования корпуса. См. Рис.11.

НКРЯ – динамично развивающийся корпус. В его составе активно разрабатываются новые подкорпуса, совершенствуются уже созданные. См., например, о перспективных проектах НКРЯ статью Е.А. Гришиной «Два новых проекта для Национального корпуса: мультимедийный подкорпус и подкорпус названий» и совместную статью Е.А. Гришиной и В.А. Плунгяна «Перспективы развития Национального корпуса русского языка» в [Национальный… 2005].

4. Лингвокультурологические корпуса как современное системное представление коммуникации. Саратовский диалектологический корпус. В настоящее время начата разработка особого, лингвокультурологического, типа корпусов, включающих, наряду с лингвистическими, также и нелингвистические данные. Корреляция таких данных в составе лингвокультурологического корпуса дает возможность системно представить определенный тип коммуникации, моделировать соответствующую разновидность общения как целостное культурно-коммуникативное образование.

Иллюстрацией лингвокультурологического корпуса является создаваемый в Саратовском государственном университете им. Н.Г. Чернышевского мультимедийный диалектологический корпус (СДК). В его основу положен принцип синтеза лингвистических и культурологических данных.

Рис.11.Окно выбора подкорпуса в НКРЯ



Центральное место в лингвокультурологической диалектном корпусе принадлежит репрезентативному массиву текстов на диалекте, отражающих важнейшие типы диалектной речи (речь бытовую, фольклорную, речь в условиях официального, обрядового общения); различные формы речи (диалог, полилог, монолог); разнообразную тематику сельского общения; социальную дифференциацию носителей говора (по полу, возрасту, профессии, уровню образования). Текстовая база корпуса содержит не только ценные лингвистические данные, но и уникальные сведения о судьбах людей, об истории родного села, края и страны в восприятии и оценках сельских жителей.

Тексты каждого отдельного говора образуют в составе СДК самостоятельный подкорпус. Текстовая часть каждого подкорпуса представлена в различных формах: в буквенной (близкой к орфографической) записи, в виде аудиозаписи или видеозаписи диалектной коммуникации. Наличие видеозаписи дает возможность наблюдать также и невербальные элементы общения, условия коммуникации.

Модель лингвокультурологического описания диалекта, наряду с собственно текстовой составляющей, должна включать также сведения о среде диалектного общения, его жанровом и тематическом репертуаре и коммуникативной релевантности конкретных тем и жанров. Поэтому база данных СДК содержит, помимо лингвистически аннотированной текстовой части, многоаспектную метатекстовую и нелингвистическую информацию.

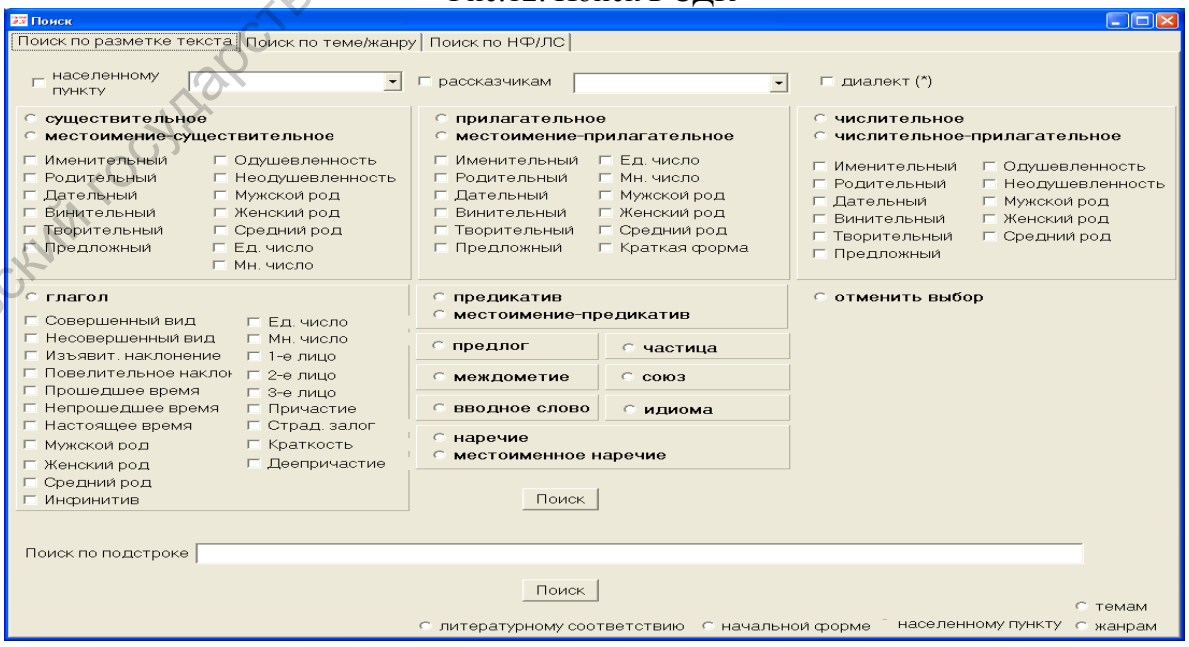
Задача лингвокультурологического описания говора в СДК обусловила включение в число параметров лингвистической разметки текстов жанровую и тематическую разметку, маркирующую все жанровые и тематические переходы в речевом континууме – единовременной записи диалектной коммуникации. С каждым введенным в базу корпуса текстом программно связаны модули с метатекстовой информацией: сведения об информантах (в том числе восстанавливаемая по текстовым данным биография диалектоносителя), о времени и месте записи текста, о конкретной ситуации общения, об адресатах речи, об упоминаемых в тексте лицах, о времени описываемых в тексте событий (до революции; революция и гражданская война; коллективизация; Великая Отечественная война; послевоенный советский период; постсоветский период), фотоиллюстрации к данному тексту. С каждым текстом в СДК связаны 3 модуля нелингвистического характера: 1) модуль с метаразметкой текста, 2) модуль, содержащий биографию информанта, 3) иллюстративный модуль.

Отдельный блок в электронной базе корпуса одного говора образует справочная нелингвистическая информация, обеспечивающая более полное понимание текстов пользователями-исследователями: краткая история населенного пункта, описание специфических природных особенностей места бытования говора, характеристика основных занятий местных жителей, план села с местными названиями его частей, актуальный топонимический и микротопонимический справочник (названия упоминаемых в текстах селений, кладбищ, полей, рек, ручьев, родников, колодцев, дорог, перекрестков и т.п.).

Реализуемая в СДК система поиска (см. Рис.12) позволяет оптимизировать выборку определенных жанровых и тематических фрагментов, проследить их соотношение в речи отдельного диалектоносителя и в коммуникативном пространстве данного говора определить место различных дискурсивных разновидностей и предметных областей в диалектной коммуникации. Тематический и пословный поиск в сочетании с данными о частотности словоформ позволяет выделить актуальные для диалектной коммуникации концептуальные переменные, а анализ конкордансов запрашиваемых словоформ предоставляет ценный материал для содержательного анализа соответствующих концептов и концептуальных оппозиций.

Лингвокультурологический диалектный корпус дает возможность получать комплексную информацию о говоре и условиях его бытования. Подробнее о принципах организации СДК и его эвристическом потенциале см. материалы, размещенные на сайте: www.sarteorlingv.narod.ru/projects.htm.

Рис.12. Поиск в СДК



5. Научные идеологические новации и эвристический потенциал современных текстовых корпусов. Наличие программно обеспеченных текстовых корпусов значительно

упрощает и ускоряет лингвистическую обработку больших массивов текстов, позволяет при минимальных затратах времени получить исследовательскую выборку, на составление которой у лингвистов уходили годы. Однако роль текстовых корпусов в лингвистических исследований – отнюдь не только техническая.

Использование текстовых корпусов обеспечивает новый уровень лингвистических исследований, опирающихся на репрезентативный, значительный по объему речевой континуум, в наиболее полной мере учитывающих функциональные и статистические характеристики языковых явлений. «Впервые в истории, – справедливо отмечают О.Н. Лагута и М.К. Тимофеева, – реально вырисовывается перспектива изучать «язык в действии», о чем некогда говорил Вильгельм фон Гумбольдт, а позже – И.А. Бодуэн де Куртенэ, А.А. Потебня, Л.В. Щерба и другие» [Лагута, Тимофеева 2007: 114].

В статье В.А. Плунгяна «Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики» обосновывается мысль о том, что появление НКРЯ «не просто дало в распоряжение лингвистов новый мощный инструмент анализа фактов языка – оно в определенной степени изменило теоретические приоритеты и отчасти даже взгляды на то, чем является язык и какие задачи изучения языка являются наиболее важными» [Плунгян 2008: 7-8]; «в настоящее время корпус – это не просто дань техническому прогрессу или более удобный инструмент для поиска примеров; это именно примета новой идеологии изучения языка, для которой язык – вообще говоря, и есть корпус…» [Там же: 12], а «текст… главный объект теоретической рефлексии» [Там же: 14].

Обращение лингвистов к корпусным данным, как подчеркивает В.А. Плунгян, отвечает современной научной парадигме, характеризующейся переходом от «системы» к «узусу» и от «языка» к «речи» и такими «идеологическими предпочтениями», как внимание к тексту (дискурсу), внимание к квантитативному компоненту языка, внимание к синхронной и диахронической вариативности языка, боле толерантное отношение к понятиям языковой нормы и языковой правильности [Там же: 8-10]. Не удивительно поэтому, что новая, ориентированная на узус лингвистика, становится лингвистикой корпусов. Рассуждая о связи корпусной лингвистики с новой теоретической моделью языка, В.А. Плунгян указывает на их взаимную обусловленность, «встречное движение»: «В каком-то смысле трудно сказать, что здесь причина, а что следствие: то ли идеологические установки «теории узуса»… привели к бурному расцвету корпусных исследований, то ли прогресс в такой первоначально сугубо прикладной области, как составление электронных корпусов языков, привел к резкой смене идеологических установок лингвистов (или ускорил эту смену). Скорее всего, имели место оба процесса, и правильнее говорить о встречном движении: в ходе теоретических поисков был обнаружен наиболее подходящий для этих поисков инструмент [Там же: 12].

Текстовый корпус позволяет получить данные, недоступные для наблюдения при обращении к другим, традиционным, источникам, ставить принципиально новые задачи, ранее практически невыполнимые из-за их трудоемкости. Так, объемный текстовый корпус может послужить надежным источником сведений о типичных для языковой единицы семантических и грамматических позициях, о частотности ее употребления, о характерных для нее прагматических и дискурсивных условиях функционирования, о динамике языковых явлений. Ср., например, приведенные В.А. Плунгяном некоторые конкретные вопросы, данные по которым могут быть добыты исследователем на материале НКРЯ «простым нажатием кнопки»: Какой из приставочных коррелятов – *прореагировать, отреагировать, среагировать* – употребляется в современном русском языке чаще? К каким контекстам тяготеет каждый из этих приставочных коррелятов (например, какой из них охотнее сочетается с наречием *быстро*)? В какой последовательности они появляются в современном языке – одновременно или по очереди? Различается ли частота их употребления в разные периоды? Сбор материала для ответа на подобные вопросы в отсутствии корпуса занял бы месяцы или годы [Национальный… 2005: 13-14]. Иначе говоря, текстовый корпус позволяет ставить и успешно

решать задачи, связанные с «микроэволюцией» языка на протяжении одного-двух столетий: наблюдать малозаметные изменения сочетаемости и значений слов, изменения частотности различных конструкций или частотности употребления лексических и грамматических вариантов, регистрировать появление или угасание отдельных явлений языка [Плунгян 2008: 14].

Наличие текстовых корпусов предъявляет новые требования к способам количественной и функциональной оценки исследованного языкового материала. «При наличии электронных коллекций уже не вполне приемлемым видится употребление неопределенных формулировок, которые характеризуют то или иное языковое явление просто как «редкое», «(не)типичное» и т.п., не подкрепляя это утверждение никакими количественными данными. …существование электронного языкового пространства в виде доступного и легко измеримого объекта сделает многие работы, проводимые в гуманитарных науках, более четкими и основательными» [Лагута, Тимофеева 2007: 128]. С момента открытия НКРЯ «русисты уже не могут игнорировать вопрос, насколько сведения, полученные в ходе изучения конкретного материала, собранного традиционными методами составления картотек (в большинстве случаев буквально «вручную»), соответствуют тенденциям, выявляемым при анализе аналогичных систем данных НКРЯ» [Там же: 115].

Опора на корпусные данные значительно повышает объективность, а следовательно, научный статус грамматических и словарных описаний языка. «Теперь, – справедливо отмечают А.Т. Хроленко и А.В. Денисов, – подлинно научные описания грамматического строя языков, а также авторитетные академические словари – практически все без исключений – должны составляться на основе корпусов этих языков» [Хроленко, Денисов 2007: 50].

Сотрудниками Института русского языка им. В.В. Виноградова РАН на основе НКРЯ уже создан ряд экспериментальных словарей. В настоящее время на сайте Института (**http://dict.ruslang.ru/**) размещены следующие словари этой серии: *Е.А. Гришина*. Грамматический словарь новых слов русского языка; *О.Н. Ляшевская*. Новый частотный словарь русской лексики; *Г.И. Кустова*. Словарь русской идиоматики. Сочетания слов со значением высокой степени; *Е.Ю. Калинина*. Словарь глагольной сочетаемости непредметных имен русского языка.

Все названные словари являются принципиально новыми лингвистическими ресурсами – и по своей источниковой базе, и по репрезентируемой предметной области, и по возможностям, предоставляемым пользователям. Словари составлены на основе значительного по объему (не менее 100 млн. словоупотреблений) текстового массива, включающего тексты самых разных жанров и типов, в том числе (в отличие от источниковой базы традиционных словарей) тексты современной художественной литературы, новостные и газетные, деловые и официальные тексты, тексты специальной научной и технической литературы, тексты электронной коммуникации (электронная переписка, чаты, форумы, «живой журнал» и др.) и тексты устной речи (интервью, дискуссии, бытовые разговоры, речь отечественного кино и др.). Словари, созданные на основе НКРЯ, являются аспектными словарями, наиболее полно и объективно репрезентирующими специальную предметную область. Пользователю предоставляется возможность формировать (с минимальной затратой усилий) разнообразные виды выборок, получать многообразную и репрезентативную информацию о языковых явлениях. См. Рис.13.

Рис.13. Главная страница Нового частотного словаря русской лексики

О. Н. Ляшевская
НОВЫЙ ЧАСТОТНЫЙ СЛОВАРЬ РУССКОЙ ЛЕКСИКИ

Как пользоваться словарем

Введение к Новому частотному словарю

I. Общая лексика

- Алфавитный список лемм
- Частотный список лемм
- Распределение лемм по функциональным стилям:

— Частотный словарь художественной литературы
— Словарь значимой лексики художественной литературы
— Частотный словарь публицистики
— Словарь значимой газетно-новостной лексики
— Частотный словарь другой нехудожественной литературы
— Частотный словарь значимой другой нехудожественной литературы
— Частотный словарь живой устной речи
— Словарь значимой лексики живой устной речи

- Алфавитный список словоформ

II. Общая лексика: части речи

- Частотный список имен существительных
- Частотный список глаголов
- Частотный список имен прилагательных
- Частотный список наречий и предикативов
- Частотный список местоимений (местоимения-существительные, прилагательные, наречия, предикативы)
- Частотный список числительных
- Частотный список лемм служебных частей речи

III. Вспомогательные таблицы

- Данные о частотности частеречных классов
- Частотность букв русского алфавита
- Частотность двубуквенных сочетаний

IV. Имена собственные и аббревиатуры

- Алфавитный список собственных имен и аббревиатур

Апелляция к текстовым корпусам национальных языков целесообразна в практике преподавания филологических дисциплин, в том числе в процессе изучения данного языка как иностранного. Об использовании НКРЯ в учебных целях см., например, статьи Н.Р. Добрушиной «Как использовать Национальный корпус русского языка в образовании?» и «Корпусные методики обучения русскому языку» в [Национальный корпус… 2005] и [Национальный корпус… 2009], методические разработки этого же автора [Добрушина 2008].

Для учебно-методических целей в НКРЯ специально выделен Обучающий подкорпус, грамматическая информация в котором соответствует современной школьной программе. Помимо общих для всего корпуса стандартных поисковых запросов, Обучающий подкорпус предоставляет возможность поиска по таким параметрам, как склонение существительных, спряжение глаголов, разряды существительных, прилагательных, местоимений, наречий. Перспективам использования Обучающего подкорпуса НКРЯ в образовательной деятельности посвящена статья С.О. Савчук, Д.В. Сичинавы «Обучающий корпус русского языка и его использование в преподавательской практике» [Национальный корпус… 2009].

Оптимизируя работу лингвиста в исследовании семантики языковых единиц, в оценке употребительности тех или иных семантических вариантов и языковых выражений, корпуса текстов эффективно используются при проведении лингвистических экспертиз. Опыт использования корпусных технологий в решении задач различных типов лингвистических экспертиз описан в [Баранов 2007].

Предпочтение, которое отдается современной лингвистикой корпусно-ориентированным исследованиям, объясняется в целом тем, что корпус «позволяет изучать действительно существующие в языке, а не мнимые явления» [Плунгян 2008: 17].

Рекомендации и задания. 1. Для общей ориентации в электронных текстовых корпусах найдите по ссылкам, приведенным в тексте лекции, размещенные в Интернете корпуса.. 2. Познакомьтесь с сайтами, включающим ссылки на текстовые корпуса. Такие ссылки есть, например, на сайтах НКРЯ (http://ruscorpora.ru/**)**, Фонда «Федеральная электронная библиотека» (ФЭБ – http://feb-web.ru/).

Для закрепления полученных сведений выполните следующие задания:

I.

1. Откройте в Интернете ***Национальный корпус русского языка*** (http://ruscorpora.ru/), прочтите общую информацию о НКРЯ (разделы «Что такое корпус?», «Подробнее о корпусе», «Состав и структура», «Новости проекта»), просмотрите перечень закладок на главной странице и ознакомьтесь с их содержанием.

2. Ответьте на вопросы:

а) какие характеристики НКРЯ обеспечивают его представительность и сбалансированность;

б) какие виды разметки используются в настоящее время в НКРЯ;

в) для каких целей предназначен НКРЯ;

г) какие подкорпуса входят в настоящее время в состав НКРЯ;

д) какую информацию предлагает пользователю Образовательный портал Национального корпуса русского языка?

3. Войдите в систему поиска НКРЯ; руководствуясь инструкцией (закладка в поисковом окне), проведите различные виды поиска: поиск по словоформе (*чиновники*), слову (*чиновник*), словосочетанию (*чиновники министерства*), сочетанию слов (*чиновник, власть*), находящихся на расстоянии друг от друга от 1 до 5, по заданным грамматическим характеристикам (*чиновник* + последовательно выбираемые значения *падежа* и *числа*, например, *чиновник* + *род.п. мн.ч.* и т.д.), по сочетанию признаков (1. сегмент слова **ник* + грамматические признаки «существительное, муж.р» + семантический признак «лица»; 2. слово *чиновник* + дополнит. признак «повтор лексемы»).

Задайте те же самые параметры поиска в подкорпусе Нехудожественных текстов официально-деловой сферы.

На материале полученных выборок опишите кратко количественные характеристики употребления слова «чиновник».

II.

1. Откройте в Интернете ***Тюбингенские корпусы русских текстов*** (http://www.sfb441.uni-tuebingen.de/b1/rus/korpora.html), прочтите информацию о ресурсе на главной странице.

2. Ответьте на вопросы:

а) какими кодировками можно пользоваться, работая с ТК;

б) какой вид разметки применяется в ТК;

в) тексты какой тематики включены в корпус текстов интервью;

г) какой словарь составлен на основе Уппсальского корпуса?

3. Войдите в систему поиска ТК, получите выборки по словоформе *молодость*, а также по маске *молод** отдельно из каждого включенного в ТК корпуса. Сопоставив выборки, кратко опишите полученные результаты.

III.

1. Откройте в Интернете Санкт-Петербургский ***Корпус русского литературного языка*** (http://www.narusco.ru/resourses.htm) и ознакомьтесь с представленными в разделе «О проекте» сведениями о принципах построения корпуса.

2. Ответьте на вопросы:

а) тексты какого типа и какого периода включены в Корпус;

б) является ли Корпус сбалансированным и чем это определяется;

в) для решения каких исследовательских и прикладных задач могут быть использованы материалы Корпуса?

3. Задайте одну и ту же словоформу, используя «поиск в словаре словоформ» и «поиск по текстам корпуса». Опишите, какие сведения и в какой форме получает пользователь, используя каждый из названных видов поиска.

IV.

1. Откройте в Интернете создаваемый в МГУ ***Компьютерный корпус газетных текстов русского языка конца XX века*** (http://www.philol.msu.ru/~lex/corpus/), ознакомьтесь с описанием корпуса по соответствующей ссылке на главной странице.

2. Ответьте на вопросы:

а) какие тексты образуют источниковую базу КГТ;

б) как обеспечивается в КГТ принцип пропорциональности данных;

в) какие параметры метразметки применяются в КГТ;

г) какая классификация жанровых типов использована в метаразметке КГТ;

д) какие сведения можно извлечь из частотно-распределительных словарей, построенных на основе КГТ?

3. Войдите в систему поиска КГТ, получите выборки на употребление леммы *Россия* в текстах разных жанров и жанровых типов. Сопоставьте выборки и кратко опишите полученные результаты.

V.

1. Откройте в Интернете созданный в МФРЯ ***Корпус российских газет 90-х гг. XX в.***: (http://cfrl.ru/newspap.shtm). Выбрав строку «О корпусе российских газет», ознакомьтесь с принятой в корпусе метакодировкой текстов.

2. Ответьте на вопросы:

а) тексты каких газет включены в Корпус;

б) какие данные можно получить в Корпусе по каждой газете?

3. Войдите в систему поиска по Корпусу, сделайте выборки на слово *культура* из Корпуса в целом и из каждой газеты в отдельности. Сопоставьте выборки и кратко опишите полученные результаты.

4. Оцените возможности поиска по шаблону.

VI.

1. Откройте в Интернете ***Хельсинкский аннотированный корпус*** (http://www.ling.helsinki.fi/projects/hanco/), ознакомьтесь с изложенными на главной странице принципами построения корпуса, а также с возможностями, предоставляемыми поисковыми механизмом корпуса (опция «Помощь»).

2. Ответьте на вопросы:

а) каков объем и источники ХАНКО;

б) какие виды аннотирования осуществляются в ХАНКО;

в) каким образом в ХАНКО представлены неоднозначно трактуемые в лингвистике языковые явления;

г) какие дополнительные сведения можно получить по каждому из контекстов выдачи?

3. Войдите в систему поиска ХАНКО, проведите поиск: а) по начальной форме выбранного Вами слова, используя дополнительно разные комбинации других параметров поиска; б) по выбранной текстоформе, используя дополнительно разные комбинации других параметров поиска. Опишите кратко полученные выборки и виды доступной в этих выборках информации.

VII.

1. Откройте в Интернете ***Словарь-конкорданс публицистики Ф.М. Достоевского*** (http://dostojevskij.karelia.ru/author.phtml), ознакомьтесь с принципами создания словаря, выбрав соответствующую опцию в меню; прочтите также комментарии к параметрам поисковых запросов (опция «Помощь»).

2. Ответьте на вопросы:

а) какую информацию о слове/ словоформе позволяет получить Словарь-конкорданс;

б) на каких принципах основывается формирование текстовой базы Словаря-конкорданса;

в) какие группы текстовых фрагментов выделены в качестве самостоятельных файлов, отдельно обрабатываемых для частотного словаря публицистики Ф.М. Достоевского;

г) по каким параметрам может быть отсортирован материал выдачи?

3. Войдите в систему поиска по корпусу, проведите поиск: а) по начальной форме выбранного Вами слова, используя дополнительно разные комбинации других параметров поиска; б) по выбранной текстоформе, используя дополнительно разные комбинации других параметров поиска. Опишите кратко полученные выборки и виды доступной в этих выборках информации.

VIII.

1. Откройте в Интернете ***Конкордансы произведений Ф.М. Достоевского*** (http://petrsu.ru/~Dostoevsky/) и ознакомьтесь с поисковым интерфейсом ресурса.

2. Ответьте на вопросы:

а) с чем связаны трудности поиска по конкретному слову в Конкордансах произведений Ф.М. Достоевского;

б) с помощью какой поисковой опции может быть устранено затруднение поиска по конкретному слову?

3. Войдите в систему поиска по корпусу, получите выборку на употребление слов с корнем *бог-/бож-* в дневниках писателя за разные годы. Опишите кратко полученные результаты, обратив внимание на характер информации, предоставляемой полученной выборкой.

IX.

1. Откройте в Интернете ***Словарь языка Грибоедова*** (http://feb-web.ru/feb/concord/abc/), прочтите Предисловие к Словарю.

2. Ответьте на вопросы:

а) каков объем и источники Словаря;

б) чем определяется принципиальная новизна Словаря, его отличие от традиционной писательской лексикографии

3. Войдите в систему поиска по Словарю, откройте словарную статью на слово *автор*, опишите структуру словарной статьи и виды получаемой пользователем информации.

X.

1. Откройте в Интернете ***Параллельный корпус переводов «Слова о полку Игореве»*** (http://www.nevmenandr.net/slovo/), прочтите статью «О проекте», выбрав в меню опцию «Проект → Описание».

2. Ответьте на вопросы:

а) какова цель Корпуса;

б) как размечен в Корпусе текст «Слова…»;

в) какие тексты, кроме переводов, включены в Корпус;

г) какую информацию содержат всплывающие окна?

3. Войдите в систему поиска по Корпусу, сформируйте подкорпус из 4-х текстов, выбрав по одному из каждой группы текстов, отметьте один из фрагментов «Слова…» и получите соответствующую запросу выборку. Используя опцию «Навигация», поработайте с

закладками «Следующий фрагмент», «Выбор переводов и фрагмента» («Расширенное», «Работа с блоками», Поиск»). Опишите кратко возможности, предоставляемые пользователю, поисковым механизмом Корпуса.

Литература

Основная:

Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. М.: Эдиториал УРСС, 2001.

Баранов А.Н. Лингвистическая экспертиза текста: теория и практика. Учеб. пособие. М.: Флинта : Наука, 2007 (Гл. 7. Технология экспертной деятельности: Корпусы текстов в лингвистической экспертизе текста).

Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – СПб.: СПб ГУ, 2005.

Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 2005.

Национальный корпус русского языка : 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009.

Плунгян В.А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. № 2 (16). 2008.

Дополнительная:

Гвишиани Н.Б., Герви О.Ю. Корпусная лингвистика и грамматика речи // Вестник МГУ. Сер. 9. Филология. 2001. № 2.

Добрушина Е.Р. Использование Национального корпуса русского языка в преподавании филологических дисциплин. Методические разработки в помощь преподавателям высшей и средней школы. М., 2008.

Лагута О.Н., Тимофеева М.К. Национальный корпус русского языка и Интегрум: итоги и перспективы // Русский язык в научном освещении. №2 (14). 2007.

Перцов Н.В. О роли корпусов в лингвистических исследованиях // Труды международной конференции «Корпусная лингвистика – 2006». СПб., 2006.

Перцов Н.В. К суждениям о фактах русского языка в свете корпусных данных // Русский язык в научном освещении. № 1 (11). 2006.

Хроленко А.Т., Денисов А.В. Современные технологии для гуманитария: практическое руководство. М.: Флинта : Наука, 2007.

Лекция 4. Русская электронная лексикография

1. Две ориентации электронных словарей.
2. Словари для пользователя-человека.
3. Рекомендации и задание.
4. Литература.

1. **Две ориентации электронных словарей.** Лексикографами разработано множество разновидностей лингвистических словарей: словари толковые, словари синонимов, антонимов, омонимов, паронимов, сокращений, личных имен, фамилий, образных средств и под., орфоэпические, морфологические, словообразовательные, фразеологические, грамматические, синтаксические; словари литературного языка, словари трудностей, словари разговорной речи, областные, терминологические, словари жаргонной лексики, словари личности, словари исторические, этимологические и многие другие. С развитием компьютерных технологий словари стали воплощаться не только в бумажной, но и в электронной форме, при этом возникло и более существенное различие: различие между словарями, предназначенными для использования их человеком, и словарями для компьютерных программ.

Главное различие между ними заключается в том, что первые представляют содержание слова и другие его свойства принципиально неполно и нестрого. Их задача – помочь владеющему языком человеку отличить одни единицы от других (по минимальному количеству признаков), активизировать его интуицию и подтолкнуть к узнаванию слова, а также частично дополнить и скорректировать имеющиеся у него языковые знания. Задача вторых – дать всю необходимую для работы программы информацию о внесенных в словарь единицах, притом – в формализованном виде.

Человек, владея языком и обладая знаниями о мире (фоновыми знаниями), понимает, что *сидеть на стуле* – не то же, что *сидеть на воздухе*, а *сидеть на диете* – это употребление глагола *сидеть* с еще одним особым значением. При этом в словаре, предназначенном для человека, не нужно перечислять или иным способом строго задавать абсолютно все сочетания глагола *сидеть* с другими словами: человек, зная категории вещей, понимает, что выражение *сидеть на веранде* подобно выражению *сидеть на воздухе*, а *сидеть на одном хлебе и воде* подобно выражению *сидеть на диете,* тогда как для успешной обработки компьютером высказываний с глаголом *сидеть* и вообще всех многозначных слов и омонимов необходимо исчерпывающее указание условий их употребления. Разрешение многозначности и снятие омонимии – одна из главных проблем компьютерной лексикографии.

Примером развитого словаря, ориентированного на использование компьютерными программами, является комбинаторный словарь в составе многофункционального лингвистического процессора ЭТАП-3, разрабатываемого Лабораторией компьютерной лингвистики Института проблем передачи информации РАН. По характеру представления информации о слове данный словарь является реализацией и развитием идей теории лингвистических моделей типа «Смысл ⇔ Текст» и использует формализмы «лексических функций» (см.: [Мельчук 1995; Апресян и др. 2007]) В работе [Цинман, Иомдин 1997] организация данных в комбинаторном словаре системы ЭТАП-3 иллюстрируется статьей АПЛОДИСМЕНТЫ. Вот этот пример:

15417 АПЛОДИСМЕНТЫ
POR:S

SYNT:мужск,мн!
DES:’факт’,’действие’,’абстракт’
D1.1:род
D2.1:дат
_SYN: рукоплескания/овация
_V0: аплодировать
_MULT: взрыв/буря/гром
_MAGN: бурный/продолжительный/восторженный/громкий
_ANTIMAGN: сдержанный/скупой/редкий/жидкий
_VER: заслуженный
_LABOR1-2: встречать/награждать
_OPER2: вызывать2
_INCEPFUNC0: раздаваться/грянуть/вспыхивать
_FINFUNC0: смолкать/стихать/прекращаться

Как можно видеть, в статье определена часть речи слова «аплодисменты» (POR:S – существительное), указаны синтаксические (SYNT) и семантические (DES) признаки слова (дескрипторы), а строки D1.1 и D2.1 описывают управляющие свойства слова; затем на языке лексических функций зафиксированы синонимы слова (SYN), глагольный дериват слова «аплодисменты» (V0), способ выражения составного характера явления аплодисментов (MULT), стандартные обозначения высокой (MAGN) и низкой (ANTIMAGN) степени явления, способ выразить, что явление таково, каким должно быть (VER), обозначение действий, соответствующих ситуации «АПЛОДИСМЕНТЫ» (LABOR1-2, OPER2), ее начальной и конечной фазам (INCEPFUNC0, FINFUNC0).

Структурированная подобным образом информация может быть использована на разных уровнях анализа и синтеза текстов, а также в осуществлении таких относительно самостоятельных функций ЭТАПа-3, как речевое перефразирование и обучение языку.

При решении многих прикладных задач (автоматическое индексирование и реферирование документов, информационный поиск, анализ и синтез текстов, машинный перевод и др.) необходимы электронные идеографические словари языков, репрезентирующие системные отношения между лексическими единицами. Словарь такого типа (WordNet) был создан в Принстонском университете (США) на базе английского языка под руководством Дж. Миллера и появился в свободном доступе в Интернете в 1995 г. (http://wordnet.princeton.edu/).

В словаре WordNet учтено, что системные связи слов разных частей речи имеют свою специфику, поэтому отношения между словами в системах существительных, глаголов, прилагательных и наречий представлены в нем раздельно.

Так, в WordNet при описании существительных определяются отношения синонимии, антонимии, родо-видовые отношения, отношение «часть – целое» (меронимия) и некоторые другие. В глагольной лексике слова, обозначающие действия и события, противопоставляются словам, обозначающим состояния; вводится специфическое отношение «следования» и в качестве его варианта – отношение «тропонимии». Оно связывает глаголы, являющиеся общим обозначением определенного действия, и глаголы, выражающие некоторую манеру, способ совершения действия (ср.: *говорить – мямлить, бормотать; идти – плестись, тащиться* и под.). Как особое отношение рассматривается причинная связь между событиями и состояниями.

Основная единица описания в WordNet – «синсет». Под синсетом понимаются слова, связанные между собой на основе широко понимаемых синонимических отношений. На исследованном материале английского языка в WordNet версии 2.1 выделяется 117 000 синсетов. Всем членам одного синсета, то есть синонимическому ряду в целом, приписывается в WordNet единое значение. В результате системные отношения в лексике предстают как отношения между синсетами.

Появление WordNet стимулировало разработку подобных компьютерных идеографических словарей для других языков в рамках организованного международного

проекта EuroWordNet . Частные идеографические словари в составе EuroWordNet («ворднеты» отдельных языков) соотносятся между собой не только на основе общих структурных принципов и разработанной для EuroWordNet системы основных предметно-логических понятий (онтологий), но и специальными индексами (ILI – Inter-Lingual-Index), позволяющими переходить от подсистем одного языка к соответствующим подсистемам других языков.

Для русского языка «ворднета» пока не существует. RusNet разрабатывается на кафедре математической лингвистики Санкт-петербургского государственного университета, но эта работа еще не завершена.

С машинными словарями мы сталкивается едва ли не при каждом обращении к компьютеру. Достаточно простые компьютерные словари являются, например, компонентами текстовых процессоров типа Microsoft Word, где ими обеспечиваются проверка орфографии, подбор слов со сходным или противоположным значением, обнаружение в текстах слов, которые не рекомендуется употреблять, с их помощью решается и ряд других задач; на основе встроенных словарей осуществляются многие привычные нам функции браузеров и других офисных программ.

2. **Словари для пользователя-человека**. За последние десятилетия русская лексикография пополнилась большим количеством электронных словарей, ориентированных на человека. Многие из них представлены в Интернете, в том числе с возможностью пользоваться ими on-line. О некоторых из них уже говорилось в предшествующих лекциях. Бóльшая часть этих словарей имеет бумажный прототип и выполнена в формате, обеспечивающем поиск нужной словарной статьи только по заглавному слову. Пользователю предлагается выбрать букву, с которой начинается интересующее его слово, а затем в открывшейся части словника указать искомое слово. В ответ на экран выводится соответствующая словарная статья.

Так, Фундаментальной научной библиотекой «Русская литература и фольклор» (http://feb-web.ru/feb/feb/dict.htm) представлено в свободном доступе несколько толковых словарей русского языка, в том числе «Словарь русского языка в 4-х томах» под ред. А.П. Евгеньевой (МАС) и вышедшие из печати выпуски «Словаря русского языка XVIII века» (гл. ред. Ю.С. Сорокин).

Выделив, например, в списке вокабул на букву *Ж* «Словаря русского языка XVIII века» строку *животное, животно*, мы получаем на экране полный текст словарной статьи, который при необходимости может быть скопирован в буфер:

> **ЖИВО́ТНОЕ**, *аго* и ◄ (*слав.*) **ЖИВО́ТНО**, а, *ср.* □ *мн. им. -ыя* и ◄ (*слав.*) *-ая, род.* ◄ (*слав.*) *-тен*. *Всякое живое существо.* Человѣк хотя и называется животным разумным, однако часто отступает от своего разума. Анчкв 1770 13. Земчуг. Малый шарик вещества бѣлаго, яснаго и твердаго, производимый животными, в раковинах живущими. Сл. нат. ист. I 169. Жизнь свойственна не одним животным, но и растѣниям. Рдщв Чел. II 165. | *О живых существах, исключая человека.* Демокрит вопрошен, какое разнство между людей и животных, отвѣща: разумно умствовати. Апофегм. 81. Что можеш разсуждати о четвероногих живо́тен. Мех. Штурма 19. Металлы служат нам в ловлении земных и морских животных. Лом. СС I 251. || *Перен. Простореч. О глупом, грубом человеке.* [Клара:] Да не только дочери, и племянница то ево от нево мучится .. [Пасквин:] Гнуснѣйшее животное! Сум. Лих. 96. [Размотаев:] Какое несносное животное всякой купец! Дв. купец 121.

Существенно, что при этом полностью сохраняются графика и орфография иллюстративного материала, а также всё оформление статьи, включая специальные знаки и шрифтовые выделения.

Словари этого типа не обогащены дополнительными функциями в области поиска и по сути дела являются электронными переизданиями осуществленных ранее словарных публикаций. Однако они имеют немалую ценность, поскольку чрезвычайно расширяют круг пользователей, облегчают и ускоряют работу со словарями, в том числе с труднодоступными. Из таких электронных изданий, отбирая и накапливая ссылки на словари, нетрудно сформировать достаточно большую словарную библиотеку, соответствующую конкретным потребностям специалиста. Один из вариантов такого словарного собрания предлагается на портале ГРАМОТА.РУ (www.gramota.ru). См. также «Словари» на www.yandex.ru.

Необходимо отметить и другую современную форму взаимодействия компьютерных технологий с «бумажной» лексикографией: крупные лексикографические проекты, результаты которых воплощаются в традиционной бумажной форме, нередко опираются сегодня на электронные текстовые корпусы как репрезентативный и сбалансированный источник речевого материала. Так, базой «Нового объяснительного словаря синонимов русского языка» (НОСС) явился сформированный авторами электронный текстовый корпус объемом выше 5 миллионов словоупотреблений; при создании «Словаря-тезауруса современной русской идиоматики» (СТИ) в качестве основных источников были использованы три текстовых корпуса («Современная русская публицистика», «Русская проза (60-90 гг.)», «Русский детектив»).

Поскольку словарные описания в основном хорошо структурированы и разные словари имеют совпадающие зоны (зона заглавного слова, зона грамматической информации, зона толкования, зона иллюстраций, зона стилистических помет, зона синонимов, зона антонимов, зона фразеологии и т.д.), то, во-первых, словарь достаточно просто отобразить в виде электронной базы данных и, во-вторых, различные электронные словари можно рассматривать как распределенную базу и ориентировать свои запросы сразу на совокупность словарей.

Так, в частности, организована проверка слов пользователями на портале ГРАМОТА.РУ. Например, поместив в окно запроса слово *огромный*, мы получаем в ответ справку, составленную из данных «Орфографического словаря» РАН под ред. В.В. Лопатина (**огромный**; *кр. ф.* -мен, -мна), полного текста соответствующей статьи «Большого толкового словаря» под ред. С.А. Кузнецова:

> **ОГРОМНЫЙ,** -ая, -ое; -мен, -мна, -мно. **1.** Очень большой по своим размерам, величине, объёму. *О-ая сумма. О. дом. Человек огромного роста. О. столетний дуб. Глазища у него о-ые!* **2.** Очень большой по количеству. *О-ая масса народу. О-ая армия. Получил о-ые деньги. Штат в учреждении был огромным.* **3.** Очень большой по силе, глубине и т.п. *О-ое влияние. О-ое впечатление. О. успех. У него о. талант. Брат обладает огромным самолюбием. Писатель огромного значения. Международный договор огромной значимости. Над человечеством нависла о-ая опасность. У него о-ые связи.* < Огромность, -и; *ж. О. зданий. О. армии. О. таланта.),*

из списков синонимов слова по материалам двух различных синонимических словарей, а также полный текст статьи ОГРОМНЫЙ из «Словаря антонимов» М.Р. Львова.

Организация лингвистического словаря в виде электронной базы данных существенно расширяет возможности работы с ним: увеличивается количество входов в словарь, допускаются разнообразные типы запросов, фильтрации, сортировки материала, получения различных количественных данных.

Разработаны различные способы организации электронных словарей. От избранной технологии зависят предоставляемые пользователям возможности поиска. Так, электронная версия «Толкового словаря русского языка» С.И. Ожегова, Н.Ю. Шведовой (по изданию Издательства «Азъ» 1992 г.) – http://ak.ak22.net/dict/ – позволяет строить запросы и по полным написаниям слов, и – что очень важно – по маске. Это дает возможность выбирать все статьи, заглавное слово которых включает заданную пользователем последовательность знаков. Например, на запрос **забот**, где символ «*» обозначает любое количество любых знаков, на

экран выводится полный текст словарных статей ЗАБОТА, ЗАБОТИТЬ, ЗАБОТИТЬСЯ, ЗАБОТЛИВЫЙ, ОЗАБОТИТЬ, ОЗАБОТИТЬСЯ, ПОЗАБОТИТЬСЯ.

Не менее важно то, что в данном словаре можно осуществлять прииск не только по заглавным словам, но и по любым компонентам толкующей части. Найдем, например, статьи словаря, в толкующей части которых встречается слово *вместилище*. В ответ на запрос получаем:

ЁМКОСТЬ -и, ж. 1. см. емкий. 2. **Вместилище** для жидких и сыпучих тел (спец.). Ёмкости для нефтепродуктов, для зерна. || прил. емкостный, -ая, -ое.

БАГАЖНИК -а, м. **Вместилище** в автомобиле, приспособление у велосипеда, мотоцикла для перевозки поклажи.

БОЛТАТЬСЯ -аюсь, -аешься; несов. 1. Свисая или вися свободно, двигаться из стороны в сторону, качаться (разг.). Конец пояса болтается. За спиной болтается пустой рюкзак. Пиджак болтается на плечах (слишком свободен). 2. О предметах, неплотно лежащих в каком-н. **вместилище**, или о жидкости в сосуде: двигаться из стороны в сторону, перемещаться, колыхаться (прост.). Во фляге болтаются остатки воды. Ноги болтаются в ботинках (ботинки слишком велики). 3. Ходить без дела, слоняться (разг.). Б. по улицам.

БОЧКА -и, ж. 1. Деревянное, обтянутое обручами, или металлическое цилиндрическое **вместилище** с двумя днищами и обычно с несколько выгнутыми боками. Дубовая б. Б. для воды. Б. с цементом. Набить обручи на бочку. 2. Старая русская мера жидкостей, равная сорока ведрам (около 490 л). 3. Фигура высшего пилотажа - полный оборот самолета вокруг его продольной оси. * На пороховой бочке сидеть - о положении, грозящем опасностью в любую минуту. Бочку катить на кого (прост.) - нападать, устраивать неприятности кому-н. Деньги на бочку! (разг.) - требование расплатиться сразу наличными деньгами. || уменьш. бочечка, -и, ж. (к I знач.). || прил. бочечный, -ая, -ое (к 1 знач.) и бочковый, -ая, -ое (к 1 знач.). Бочечная дощечка. Бочковое пиво (в бочках).

БУНКЕР -а, мн. -ы, -ов и -а, -ов, м. 1. Специально оборудованное **вместилище** для сыпучих и кусковых материалов. Б. для угля. Б. комбайна (для зерна). 2. Бетонированное подземное укрытие, убежище. || прил. бункерный, -ая, -ое.

ВАННА -ы, ж. 1. Большой, обычно продолговатый сосуд для купания, мытья. Сесть в ванну. 2. Мытье или лечебная процедура в таком сосуде. Принять ванну. Хвойная в. 3. перен. Лечение воздействием на тело солнца, воздуха, грязей. Солнечные, воздушные, грязевые ванны. 4. В технике: **вместилище** для растворов, жидкостей, используемое для их хранения или для различных технологических процессов. Красильная в. || уменьш. ванночка, -и, ж. (к 1 и 4 знач.). || прил. ванный, -ая, -ое. Ванная комната. В. павильон.

ВЛАГАЛИЩЕ -а, ср. 1. Конечный отдел половых проводящих путей у женщины. 2. Место, в к-ром укрепляется какой-н. орган, а также **вместилище** из кожных покровов (спец.). В. сухожилия. В. волоса, В. листа. н прил. влагалищный, -ая, -ое.

ДОЛБЛЁНКА -и, ж. (обл.). Долбленое **вместилище** (сосуд, колода) или лодка, улей.

ЗАРЯД -а, м. 1. Количество взрывчатого вещества, необходимое для взрыва, выстрела и содержащееся в соответствующем устройстве в специальном **вместилище**. 3. взрывчатки. Пороховой з. 3. энергии (также перен.: о скопившейся в ком-н. энергии). 2. Количество электричества, содержащееся в данном теле. Электрический з. (величина, определяющая интенсивность электромагнитного взаимодействия заряженных частиц). * Снежный заряд - внезапный и сильный снегопад. || прил. зарядный, -ая, -ое и зарядовый, -ая, -ое (ко 2 знач.; спец.). Зарядное устройство (в электротехнике). Зарядный ящик (повозка для снарядов; устар.).

ЗАРЯД -а, м. 1. Количество взрывчатого вещества, необходимое для взрыва, выстрела и содержащееся в соответствующем устройстве в специальном **вместилище**. 3. взрывчатки. Пороховой з. 3. энергии (также перен.: о скопившейся в ком-н. энергии). 2. Количество электричества, содержащееся в данном теле. Электрический з. (величина, определяющая интенсивность электромагнитного взаимодействия заряженных частиц). * Снежный заряд - внезапный и сильный снегопад. || прил. зарядный, -ая, -ое и зарядовый, -ая, -ое (ко 2 знач.; спец.). Зарядное устройство (в электротехнике). Зарядный ящик (повозка для снарядов; устар.).

КАДКА -и, ж. Цилиндрической формы **вместилище** со стенками из деревянных клепок, обтянутое обручами. Дубовая к, || прил. кадочный, -ая, -ое. Кадочное садоводство (выращивание растений в кадках).

КАПСУЛА -ы, ж. 1. Герметически закрытое **вместилище**. К. космического летательного аппарата. В фундамент здания заложена к. с запиской. 2. Желатиновая, крахмальная или иная легкая оболочка для нек-рых лекарств, облатка. Лекарство в капсулах. 3. Название соединительной оболочки у различных органов или их частей (спец.). || прил. капсульный, -ая, -ое.

КАРМАН -а, м. 1. Вшитая или нашивная деталь в одежде - небольшое обычно четырехугольное **вместилище** для платка, для мелких нужных под рукой вещиц. Вшивной, накладной к. Боковой, нагрудный к. К с отворотом, с клапаном. Положить, убрать, запихнуть в к. что-н. Торчит из кармана что-н. Платок, кошелек, билет, очешник в кармане. Залезть в чей-н. к. (также

перен.: украсть или ввести в расход, заставить потратиться). Спрятать в к. что-н. (также перен.: скрыть, не показывать виду. Спрятать самолюбие в к.). Широкий к. у кого-н. (также перен.: о том, у кого много денег). Пустой к. у кого-н. (также перен.: нет денег; разг.). Набить к. (также перен.: разбогатеть; разг. неодобр.). 2. Вделанное во что-н. особое отделение, л. рюкзака, сумки. 3. Углубление, выемка (спец.). л. в горной породе. К. раны. * Бить по карману (разг.) - вводить в расход, причинять убыток. Пены бьют по карману. Не по карману что коми (разг.) - слишком дорого для кого-н. Вещь дорогая, мм не по карману. В чужой карман смотреть (разг. неодобр.) - считать чужие деньги, чужое богатство. Тугой карман у кого (разг.) - о том, кто богат, обычно о скупом. В карман за словом не лезет кто (разг.) - о том, кто боек на язык, находчив в споре. Держи карман шире! (разг. ирон.) - возглас: напрасно ждешь, ничего не получишь. || уменьш. карманчик, -а, м. и кармашек, -шка, м. || прил. карманный, -ая, -ое. Карманные часы (для ношения в кармане). Карманные расходы (мелкие повседневные расходы). К. вор (ворующий из карманов). Книжка карманного формата (помещающаяся в кармане).

КОНТЕЙНЕР [тэ], -а, м. Стандартное **вместилище** для транспортировки в нем грузов без упаковки. Железнодорожный к. || прил. контейнерный, -ая, -ое. Контейнерные перевозки.

КОПИЛКА -и, ж. 1. **Вместилище** с узкой щелью для опускания монет с целью накопления. 2. перен., ед. Собрание чего-н. занимательного, ценного. К. курьезов (собрание занимательных фактов). В копилку знаний.

КОРЗИНА -ы, ж. 1. Плетеное изделие, служащее **вместилище**м для хранения вещей, для упаковки, переноски. Ивовая к. 2. В баскетболе: укрепленный на щите обруч с сеткой, в к-рую забрасывается мяч. || прил. корзинный, -ая, ое (к 1 знач.).

КОРМУШКА -и, ж. 1. Ящик, **вместилище**, в к-ром дается корм животным. Автоматизированная к. (автокормушка). 2. перен. Место, где можно, пользуясь бесконтрольностью, поживиться, приобрести что-н. для себя (разг. неодобр.). || прил. корму-шечный, -ая, -ое (к 1 знач.).

КОРОБКА -и, ж. 1. **Вместилище** для чего-н. в виде ящика, ящичка или другой формы. Деревянная, картонная, пластмассовая к. Круглая к. Швейная к. (со швейными принадлежностями). К. из-под сигарет, из-под печенья. К. конфет, спичек (с конфетами, со спичками). 2. Остов здания, а также вообще стандартное прямоугольное здание. Коробки заводских корпусов. * Черепная коробка (спец.) - костное **вместилище** головного мозга. Коробка скоростей (спец.) - механизм для изменения частоты вращения ведомого вала. || уменьш.

коробочка, -и, ж. (к 1 знач.). || прил. коробочный, -ая, -ое (к 1 знач.). Коробочная упаковка (в коробках).

И так далее.

Понятно, что поиск по словам категориальной семантики (*человек, животное, сооружение, здание, инструмент, вместилище, одежда* и под.), осуществленный в толкующей части словарных статей, оказывает существенную помощь в отборе слов с общими семантическими компонентами. Несомненное удобство представляют собой и гиперссылки от одних статей к другим (см. выше статью ЁМКОСТЬ).

С помощью данного электронного словаря также возможна фильтрация лексики по стилистическим пометам. Введя, например, в строку поиска помету *высок.*, мы получаем все словарные статьи (от АЛТАРЬ до ЭРА), в составе которых эта помета встречается.

Для оценки различий между бумажными и электронными словарями хороший материал дает русская ассоциативная лексикография. Сопоставим опубликованный в 2002 г. в Москве Издательством Астрель и Издательством АСТ «Русский ассоциативный словарь» (далее – РАС), материалы которого извлечены для печатного издания из электронной базы данных массовых ассоциативных экспериментов 1988 – 1997 гг., и построенную на материалах той же базы данных информационно-поисковую систему «Русский ассоциативный тезаурус» (далее – ИПС РАТ), открытую в 2008 г. в Интернете (http://www.philippovich.ru/Projects/ASIS/index.htm),

Печатное издание РАС – двухтомное. В первом томе помещены прямые статьи словаря (от стимула к полученным на данный стимул реакциям), во втором томе – обратные статьи (от реакции к стимулам, на которые данная реакция получена). В обоих случаях входом является только заголовочное слово (слово-стимул в первом томе, реакция – во втором).

Так, на стимул мешать в 1-ом томе РАС обнаруживается статья:

> **МЕШАТЬ:** работать *7;* думать *6;* заниматься, *ложкой,* спать *4;* жить, кому-то, тесто, учиться *3; кашу,* мне, помогать, *работе,* слушать, *соседу,* суп, товарищу *2; беспокоить, бурду,* бюрократ, варенье, *влияние, всем, встречному,* говорить, *движению, делать что-то, делу,* долго, *заниматься делом,* знать, играть, *идиоту,* идти, каша, коктейль, компот, *кому,* космос, лезть, ложка, *людям,* мысли, мыслить, *не давать делать,* очень, плохой, плыть, *победе, преподавателю, препятствовать,* работа, раствор, *слушать лекцию,* смотреть, *срать, стряпать,* трудиться, удар, *уединиться,* уйти, учить, *факты,* человеку, читать, *что-либо в чем-либо, что-либо делать,* шум *1;* 104+68+0+51

В статье приведены реакции, полученные от 104 испытуемых. Указаны абсолютная частота каждой реакции, общее количество различных реакций (68), количество единичных реакций (51), количество отказов записать реакцию (0). Курсивом выделены те реакции, которые совпадают со словами, вошедшими в общий список стимулов (другими словами: в 1-ом томе РАС имеются словарные статьи на данные слова). Реакции расположены в статье в порядке убывания их частот.

Как видим, словарь предоставляет пользователю большой круг данных о реакциях, связанных со стимулом, но, к сожалению, не все, на получение которых была направлена организация экспериментов. В экспериментах участвовали студенты различных вузов России; кроме реакций на стимулы, они фиксировали в анкетах свой пол и возраст, место проживания, специальность по которой обучаются. Учет этих параметров, безусловно, полезен при интерпретации полученных данных, однако отражение их в печатном издании увеличило бы во много раз и без того достаточно большой его объем (1-ый том – 784 стр., 2-ой том – 992 стр.) и сделало бы пользование словарем абсолютно неудобным.

Информационно-поисковая система «Русский ассоциативный тезаурус» (http://www.philippovich.ru/Projects/ASIS/index.htm), созданная, как уже сказано, на основе материалов той же электронной базы, что и печатный словарь, хотя и не увеличивает количество входов (остаются входы через слово-стимул и через реакцию), но значительно расширяет типы получаемых данных. Теперь пользователь может в интерактивном режиме

узнавать относительные частоты реакций в ассоциативных полях каждого из стимулов, фильтровать материал по полу, возрасту испытуемых и избранной ими специальности, сопоставлять между собой мужские и женские реакции на один и тот же стимул, сортировать рассматриваемые единицы в порядке убывания или возрастания их частот.

Вот, например, начало (первые 20 строк) таблицы, представляющей реакции на тот же стимул мешать, с делением ответов по полу испытуемых и нормализованными (относительными) частотами:

Статистика по запросу:
всего реакций на стимул: ***104****,*
различных реакций на стимул: ***68****,*
одиночных реакций на стимул: ***51****,*
отказов: ***0****.*

У мужчин:
всего реакций на стимул: ***54****,*
различных реакций на стимул: ***42****,*
одиночных реакций на стимул: ***36****,*
отказов: ***0****.*

У женщин:
всего реакций на стимул: ***50****,*
различных реакций на стимул: ***36****,*
одиночных реакций на стимул: ***26****,*
отказов: ***0****.*

Реакции	***Частота***	***Мужчины***	***Женщины***
Работать	**6.73**	**9.26**	4.00
Думать	**5.77**	**7.41**	4.00
заниматься	3.85	0.00	**8.00**
Ложкой	3.85	3.70	4.00
Спать	3.85	0.00	**8.00**
Жить	2.88	**5.56**	0.00
кому-то	2.88	3.70	2.00
Тесто	2.88	1.85	4.00
Учиться	2.88	3.70	2.00
Кашу	1.92	0.00	4.00
Мне	1.92	1.85	2.00
Помогать	1.92	1.85	2.00
Работе	1.92	1.85	2.00
Слушать	1.92	0.00	4.00
Соседу	1.92	1.85	2.00
Суп	1.92	0.00	4.00
Товарищу	1.92	0.00	4.00
беспокоить	0.96	1.85	0.00
Бурду	0.96	0.00	2.00
. . .	. . .	. . .	. . .

Обратим внимание на то, что общее количество женских ответов на стимул мешать несколько превышает количество мужских ответов (в данных по другим стимулам встречается и гораздо большее различие между количеством мужских и женских ответов). Это означает, что сопоставление мужских и женских реакций по их абсолютным частотам было бы некорректным и приведенные данные об относительных частотах реакций в этом случае совершенно необходимы.

В столбце «Частота» показаны относительные частоты реакций на стимул мешать без учета пола испытуемых. Реакции на стимул мешать, приведенные в первом столбце (*работать, думать, заниматься, ложкой* и т.д.), ранжированы по убыванию этих частот. В столбцах «Мужчины» и «Женщины» для каждой из представленных в первом столбце реакций указаны их относительные частоты отдельно в ответах мужчин и в ответах женщин.

Отметим наиболее частотные реакции, доли которых в ответах мужчин и в ответах женщин равны или превышают 5 %. У мужчин в эту зону попадают реакции *работать, думать, жить*, у женщин - *заниматься, спать*. При этом «женские» реакции *заниматься* и *спать* в ответах мужчин вообще не представлены. Реакции *работать* и *думать* давали не только мужчины, но и женщины, однако женщины заметно реже, а «мужская» реакция *жить* в ответах женщин на стимул мешать вообще не отмечена. Таким образом, в зоне наиболее частотных реакций ответы студентов и студенток не совпали, и это подтверждает важность сохранения в словаре информации о поле испытуемых.

Обращая внимание на различие мужских и женских реакций, необходимо вместе с тем отметить и то общее, что их объединяет: все перечисленные самые частотные реакции студентов и студенток (*работать, думать, жить, заниматься, спать*) связаны с одним и тем же значением слова «мешать»: «*Создавать препятствие в чем-нибудь, служить помехой*». Но в русском языке существуют, как известно омонимы мешать[1] и мешать[2]. Значение «*Создавать препятствие в чем-нибудь, служить помехой*» имеет мешать[1]. И у мужчин, и у женщин встретились также реакции, связанные со значением «*Переворачивать, взбалтывать круговым движением*» омонима мешать[2]. Это реакции *ложкой, тесто, кашу, суп, варенье, коктейль* и подобные. Однако доли каждой из этих реакций не достигают 5 %, и в этом также проявляется сходство ассоциативных полей у мужчин и женщин.

Теперь интересно сравнить данные РАС и ИПС РАТ с данными другого электронного ассоциативного словаря – «Ассоциативного словаря школьников Саратова и Саратовской области, материалы для которого собирались в экспериментах 1998-2008 гг. Словарь подготовлен в Саратовском государственном университете и имеет исследовательскую направленность (об АСШС см.: [Гольдин, Мартьянов, Сдобнова 2009]). Он поддерживает фильтрацию реакций по возрасту, полу испытуемых, месту их проживания, типу учебного заведения (школа или гимназия, лицей), дате проведения эксперимента и некоторым другим признакам. АСШС создает в табличной форме прямые статьи на заданные стимулы и обратные статьи по заданным реакциям, сопровождает статьи необходимыми количественными данными, предлагает стандартные перекрестные запросы, строит общие частотные перечни реакций и перечни реакций, ранжированные по убыванию так называемых «входящих» связей, выполняет и ряд других функций.

Выберем ответы учащихся 8-11-х классов на тот же стимул мешать и представим в таблице, подобной только что рассмотренной. Всего ответов - 297; в том числе ответов мальчиков – 154, девочек – 143. Начало таблицы (первые 20 строк) с наиболее частотными ответами школьников 8-11 классов на стимул мешать выглядит следующим образом:

Реакция	***Частота***	***Мальчики***	***Девочки***
кашу	**7,41%**	**6,49%**	**8,39%**
суп	**5,72%**	3,25%	**8,39%**
тесто	4,71%	4,55%	4,90%

Реакция	*Частота*	*Мальчики*	*Девочки*
ложка	3,37%	2,60%	4,20%
миксер	2,69%	1,95%	3,50%
помеха	2,36%	4,55%	0,00%
варить	1,68%	0,00%	3,50%
делать	1,68%	2,60%	0,70%
думать	1,68%	1,30%	2,10%
каша	1,68%	1,95%	1,40%
работать	1,68%	1,95%	1,40%
чай	1,68%	1,30%	2,10%
кому-то	1,35%	1,30%	1,40%
бить	1,01%	1,30%	0,70%
ложкой	1,01%	1,95%	0,00%
молоко	1,01%	0,00%	2,10%
помогать	1,01%	1,95%	0,00%
учиться	1,01%	0,65%	1,40%
в карты	0,67%	1,30%	0,00%
воду	0,67%	0,65%	0,70%
. . .	. . .	. . .	. . .

По данным АССШС, у мальчиков реакция с относительной частотой выше 5% всего одна: *кашу*. У девочек таких реакций две: *кашу* и *суп*. И у мальчиков, и у девочек самые частотные реакции оказались связанными не с мешать[1], а с «мешать[2]» в значении «*Переворачивать, взбалтывать круговым движением*». При этом реакции, вызванные восприятием стимула как «мешать[1]», имеющего значение «*Создавать препятствие в чем-нибудь, служить помехой*», в ответах детей также представлены (*помеха, думать, работать, помогать* и др.), но реже.

Могут быть выдвинуты различные предположения относительно того, чем именно обусловлены обнаруженные расхождения между реакциями школьников и студентов, однако для нашей темы важна другая сторона данного факта. Он свидетельствует о важности интеграции научных источников, которая могла бы обеспечить наиболее полный и объективный анализ лингвистических материалов. Такую интеграцию способна осуществить только компьютерная лексикография.

Отдельно следует сказать об электронных двуязычных и многоязычных словарях, поскольку они реализуют особый комплекс лексикографических функций. Одни из лучших многоязычных электронных словарей, созданных в России, – словари Lingvo фирмы ABBYY.

Словари Lingvo (англо-русский, русско-английский и многие другие) обеспечивают опознание формы слова и лемматизацию (приведение к начальной форме), перевод слов и словосочетаний с использованием тематически ориентированных групп словарей (используются словари общенаучной лексики, юридической, экономической, политехнической, медицинской, биологической и др.), поддерживают правильное написание слов и связанные с этим подсказки, позволяют получить полный перечень форм заданного слова, его транскрипцию и прослушать его произношение, предлагают специальную помощь в изучении

иностранной лексики. Пользователи словарей Lingvo имеют возможность активно настраивать их в соответствии со своими потребностями: создавать и редактировать карточки, удалять или добавлять словари и даже строить собственные словари, встраивая их в систему Lingvo. Таким образом, в системе Lingvo отчасти решается упомянутая выше актуальная проблема интеграции словарных данных.

Основная задача данной лекции, – показать, что идет активный процесс создания и совершенствования русских электронных словарей, что этими словарями обеспечивается комплекс функций, не реализуемых словарями в традиционной бумажной форме, и что сегодня современную русскую лексикографию невозможно представлять без учета ее электронной составляющей.

3. Рекомендации и задание. В лекции обсуждались в первую очередь русские электронные словари, представленные в свободном доступе в Интернете. Рекомендуем, если Вы еще этого не делали, создать собственный текстовый файл с адресами словарей, соответствующих Вашим интересам, и пополнять список адресов по мере знакомства с новыми словарями. При этом полезно присоединять к ссылкам краткие характеристики словарей с указанием их особенностей и возможности использования, а чтобы эти характеристики на самом деле оказались полезными, необходимо поработать с о словарем практически и таким путем протестировать его.

Задание. Исследуйте составленный Артемием Лебедевым «Словарь сокращений русского языка» (http://sokr.ru/). Выясните:

1. каковы источники словаря, достаточно ли они представительны и надежны;
2. кем и каким образом словарь пополняется;
3. каким образом достигается актуализация словаря (выводятся ли из него устаревшие сокращения, как маркируются новые);
4. какие типы сокращений представлены в словаре (*СГУ, СарГУ, дисс., стр., с., ун-т, мехмат...*?)
5. какой информацией сопровождаются в словаре представленные в нем сокращения;
6. возможно ли редактирование сведений, уже внесенных в словарь;
7. возможен ли поиск по маске;
8. в чем заключается разница между простым и расширенным поиском;
9. как можно уточнять запрос при использовании расширенного поиска.

Оцените ресурс Sokr.ru с точки зрения возможности опоры на него в сфере деловой, официальной коммуникации. Обладает ли данный ресурс свойствами научного источника, позволяющего использовать его данные в лингвистическом исследовании?

4. Литература

Основная:

Всеволодова А.В. Компьютерная обработка лингвистических данных. – М., 2007.

Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. М., 2006.

Мельчук И.А. Русский язык в модели «Смысл – Текст». – Москва – Вена, 1995. Ch. 1,2.

Дополнительная:

Апресян Ю.Д., Дьяченко П.В., Лазурский А.В., Цинман Л.Л. О компьютерном учебнике лексики русского языка // Русский язык в научном освещении. № 2 (14). 2007.

Гольдин В.Е., Мартьянов А.О., Сдобнова А.П. Электронный русский ассоциативный словарь школьников. // Компьютерная лингвистика и интеллектуальные технологии. «Диалог 2009». – М., 2009. Вып. 8 (15).

НОСС – Новый объяснительный словарь синонимов русского языка. М., 1997 – 2003. Т. 1-3.

СТИ – Словарь-тезаурус современной русской идиоматики. Под ред. А.Н. Баранова и Д.О. Добровольского. – М., 2007.

*Цинман Л.*Л., *Иомдин Л.Л.* Лексические функции и машинный перевод // Dialogue'97. Computational Linguistics and its Applications. Proceedings. Moscow, 1997.