МИНОБРНАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГБОУ ВПО «Воронежский государственный университет»

Факультет романо-германской филологии

Кафедра теоретической и прикладной лингвистики

О.М. Воевудская, И.А. Терентьева

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ЛИНГВИСТИКЕ

Учебное пособие для вузов

Издательско-полиграфический центр

Воронежского государственного университета

2012

Утверждено научно-методическим советом факультета романо-германской филологии, протокол № 5 от 16 мая 2012 г.

Составитель: Воевудская О.М., Терентьева И.А.

Рецензент: к.ф.н., доц. Шилихина К.М.

Учебное пособие подготовлено на кафедре теоретической и прикладной лингвистики факультета РГФ Воронежского государственного университета

Рекомендуется для студентов 1 курса дневной и вечерней форм обучения по специальностям 035700 – Лингвистика и 035800 – Фундаментальная и прикладная лингвистика

ВВЕДЕНИЕ

Предлагаемое учебное пособие «Информационные технологии в лингвистике» предназначено для студентов 1 курса всех форм обучения факультета романо-германской филологии.

Его цель – приобщить студентов к информационным технологиям в учебно-образовательной и научно-исследовательской работе. Эта цель достижима в результате решения нескольких задач:

- показать студенту возможности информационных технологий в обучении и самообразовании;
- научить студента пользоваться программными продуктами, ориентированными на задачи избранной специальности, начиная с простейших – обработки текста, проверки правописания, использования электронных словарей и ресурсов библиотек – до более сложных, таких как программы автореферирования, квантитативные методы обработки лингвистических данных, поисковые системы, программы машинного перевода, программы распознавания и синтеза звучащей речи и др.;
- выработать практические навыки обращения к интернет-ресурсам по гуманитарным дисциплинам;
- использовать полученные знания и навыки при выполнении курсовых, а затем и выпускной квалификационной работы.

Пособие включает лекции, задания на закрепление пройденного материала, а также список литературы, которая может быть полезна студентам при подготовке к экзамену (зачету) по данному курсу.

Особое внимание авторов было направлено на активизацию мыслительной деятельности студентов, на сознательное и самостоятельное добывание ими знаний.

Содержание

Тема 1: ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ЛИНГВИСТИКА

1. Информационные технологии и причины, способствовавшие их появлению
2. Прикладная лингвистика: направления и методы
3. Связь функций языка и направлений прикладной лингвистики
4. Методы прикладной лингвистики
5. Компьютерная лингвистика и теория знаний
6. Структуры представления знаний
7. Задачи прикладной лингвистики с использованием информационных технологий, будущее информационных технологий

§ 1.1. Информационные технологии и причины, способствовавшие их появлению

До последнего времени мощь любого государства определялась уровнем развития промышленности, новизной и эффективностью ее технической базы. Именно с опорой на новейшее техническое оборудование совершенствовалась технология материального производства.

Под словом ***технология*** (от греч. *techne* ‘искусство, мастерство, умение’ и *logos* ‘слово, учение’) понимается совокупность методов обработки, изготовления, изменения свойств, формы и т.д. сырья или материалов, используемых в процессе производства продукции. Такая технология была основой индустриального общества.

70-е гг. XX века стали периодом создания персональных компьютеров и началом развития информационного общества. Основное отличие этого общества от индустриального заключается в том, что наряду с технологией материального производства все большую роль начинают играть информационные технологии (далее – ИТ) как совокупность методов и средств получения, хранения, преобразования и передачи и информации с помощью компьютеров.

Появлению и бурному развитию ИТ способствовали следующие достижения научно-технического прогресса:

1) создание ПК с большой памятью и большой скоростью выполнения операций;

2) разработка звуковых плат, дающих возможность воспроизводить и записывать речь, звуки и музыку в большом диапазоне частот;

3) изобретение видеоплат, позволяющих выводить на экран компьютеров изображение с телеэкранов и видеомагнитофонов;

4) разработка мультимедийных компьютеров, позволяющих воспроизводить на экране дисплеев цвет, звук, музыку, движение;

5) создание специальных процессоров и устройств, способных передавать информацию в сети от одного компьютера к другому;

6) разработка устройств электронной связи (модемов), позволяющих передавать информацию на далекие расстояния (по телефонным линиям, кабелям, радиоканалам и т.п.);
7) создание электронной оргтехники, связанной с ПК и позволяющей осуществлять высокоскоростную печать документов, их копирование и размножение.

Конкретизируя определение понятия ИТ по отношению к лингвистике, можно сказать, что ИТ в лингвистике – это совокупность методов и средств получения, хранения, преобразования и передачи с помощью компьютеров информации о языке и законах его функционирования.

«…все существующие и потенциальные ИТ, интересующие гуманитария, можно разделить на три группы: образовательные инструментальные и поддерживающие.

Образовательные информационные технологии обеспечивают гуманитарное образование – полное или частичное. Выпускник школы может с помощью системы дистанционного образования получить специальность гуманитария. Практикующий гуманитарий может приобрести дополнительную специальность или сумму новых знаний, повысить квалификацию путём освоения дистанционного курса по избранной дисциплине или научной проблеме.

Инструментальные информационные технологии – это набор компьютерных программ, используемых для решения конкретной научной проблемы, над которой специалист работает.

Поддерживающие информационные технологии – это совокупность информационных ресурсов, к которым специалист постоянно обращается за поддержкой и советом: найти нужную информацию, отыскать требуемый текст, комментарий к тексту, толкование слова (термина), перевод информации с чужого языка на свой или наоборот и т. д. Это то, что зовется порталами, электронными библиотеками, поисковыми машинами, электронными переводчиками, электронными энциклопедиями и т. д. и т. п.» [9, с.4].

§ 1.2. Прикладная лингвистика: направления и методы

Термин «прикладная лингвистика» многозначен. В российской и западной лингвистике он имеет совершенно разные интерпретации. В западной лингвистике он связывается, прежде всего, с преподаванием иностранных языков, включая методику преподавания, особенности описания грамматики для учебных целей, преподавание как родного языка, так и иностранных, и пр. Это «узкий» подход к пониманию лингвистики, так называемая *applied linguistics*.

Другая трактовка данного термина связана с применением компьютеров для решения практических лингвистических задач. В англоязычной традиции данному пониманию соответствует термин *computational linguistics*. В отечественной литературе в качестве синонимов используются термины

«компьютерная лингвистика», «вычислительная лингвистика», «автоматическая лингвистика», «инженерная лингвистика.

Группа ученых Ленинградского (ныне Санкт-Петербургского) университета, опираясь на три основных аспекта любой области знания – теорию, эксперимент, практику, – выделили в языкознании три взаимосвязанных направления: теоретическую лингвистику, экспериментальную лингвистику и прикладную лингвистику, в рамках которой функционирует компьютерная лингвистика. Всё это «узкие» подходы к интерпретации трактовки термина «прикладная лингвистика».

В широком понимании «зонтиковый» термин «прикладная лингвистика» объединяет все те направления лингвистических исследований, результаты которых используются для решения практических задач, связанных с передачей, обработкой и хранением информации, обучением языку, коммуникативным воздействием и т.д. Таким образом, прикладная лингвистика (в широком понимании) – раздел языкознания, в котором разрабатываются методы и способы решения практических задач, связанных с использованием естественного языка в различных сферах человеческой деятельности.

§ 1.3. Связь функций языка и направлений прикладной лингвистики

Классифицировать разнообразные направления прикладной лингвистической деятельности удобно, опираясь на те функции, которые язык выполняет в обществе. Такая классификация была предложена А.Н. Барановым [1, с. 16].

Основная функция языка – *коммуникативная*, то есть в обществе язык служит средством общения и передачи информации. Именно на оптимальную организацию общения нацелены разработки таких лингвистических направлений, как теория и практика перевода, машинный перевод, теория и практика преподавания языков, теория и практика информационно-поисковых систем. Результатами практических разработок названных направлений пользуется сегодня каждый человек.

Еще одна функция языка – *социальная*, которая является важной составной частью коммуникативной функции: язык функционирует в обществе, следовательно, неизбежно взаимовлияние общества и языка. Изучением взаимоотношений языка и общества занимается в первую очередь социолингвистика. Кроме нее к этой группе относятся такие направления исследований, как политическая лингвистика, теория воздействия, теория рекламы. Эти дисциплины изучают влияние языка на поведение человека. В эту же группу прикладных исследований войдут орфография и орфоэпия, поскольку их основная задача – нормирование языка, необходимое для успешной коммуникации.

Следующая функция языка – *эпистемическая*. Это означает, что язык является средством хранения и переработки информации. Здесь прикладные лингвистические задачи решаются в рамках таких направлений

исследований, как терминология и терминография. Эти направления сосредоточены на создании и стандартизации терминологических систем. Оптимизацией эпистемической функции языка также занимается лексикография, а также корпусная и полевая лингвистика.

Последняя группа направлений прикладных лингвистических исследований – это дисциплины, занимающиеся оптимизацией *когнитивной* функции языка. Когнитивная функция связана с возможностями человека познавать окружающий мир, а язык является одним из средств познания. В эту группу входят компьютерная лингвистика, психолингвистика, лингвистической криминологии, афазиологии, квантитативная лингвистика.

Компьютерная лингвистика решает задачи оптимизации естественного языка для эффективного взаимодействия человека и компьютера. Основные задачи психолингвистики – изучение механизмов порождения и восприятия речи человеком. Квантитативная лингвистика изучает язык статистическими методами.

Таким образом, круг вопросов, которые сегодня решает прикладная лингвистика, весьма широк. Поэтому далеко не всегда можно четко отделить прикладное языкознание от теоретического. Правильнее говорить о том, что практически любая область исследований в современной лингвистике обладает как прикладным, так и теоретическим аспектом. К тому же, очень часто прикладные задачи не могут быть решены без предварительных исследований теоретического характера.

§ 1.4. Методы прикладной лингвистики

Методы, которыми пользуются лингвисты для решения практических задач, чрезвычайно разнообразны. Выбор метода исследования определятся, во-первых, особенностями исследуемого объекта, а во-вторых, – целями и задачами конкретного исследования.

Поэтому мы можем говорить о том, что в прикладной лингвистике применяются как общенаучные методы исследования, так и частные методы, которые используются в рамках какой-либо конкретного научного направления.

К общенаучным методам можно отнести классификацию – установление набора тех параметров, которые позволяют описать объект изучения с достаточной степенью полноты.

Выбор метода исследования в лингвистике осложняется тем, что язык – это специфический объект изучения, который недоступен в прямом наблюдении. Поэтому одним из основных методов его изучения является моделирование.

Необходимость в моделировании возникает в тех научных областях, где объект изучения недоступен непосредственному наблюдению. В таких случаях он уподобляется некоему «черному ящику», о котором известно только то, какие начальные материалы он получает «на входе» и какие конечные продукты он выдает «на выходе». Задача исследователя состоит в

том, чтобы изучить процесс переработки исходных материалов в конечные продукты. Однако напрямую этого сделать нельзя, поскольку, нарушив целостность «черного ящика», мы одновременно нарушим и механизм его функционирования. Можно сказать, что лингвисты тоже имеют дело с «черным ящиком»: единственной реальностью, с которой они непосредственно имеют дело, является текст (устный или письменный), а интересующие их механизмы языка, лежащие в основе речевой деятельности человека, не даны в прямом наблюдении.

По определению американского математика К.Э. Шеннона, модель является представлением объекта, системы или понятия (идеи) в некоторой форме, отличной от формы их реального существования.

Методом моделирования пользуются и в теоретической, и в прикладной лингвистике. Модели, создаваемые в рамках теоретического языкознания, призваны описать языковую систему во всей ее полноте. Например, в теоретической лингвистике часто используются следующие типы моделей:

• компонентные модели или модели структуры (из чего сделан Х);
• предсказывающие модели (предсказать поведение Х в тех или иных обстоятельствах);
• имитирующие модели (внешне вести себя как Х);
• диахронические модели (как и почему меняется Х с течением времени).

В прикладном же языкознании основной целью создания модели является потребность в решении определенной задачи, а не познание того, «как все обстоит на самом деле». Таким образом, различия между теоретическими и прикладными моделями можно свести к следующему: во-первых, прикладные модели требуют большей степени формализации. Во-вторых, прикладные модели используют знания о языке выборочно, позволяя значительно упрощать объект моделирования.

§ 1.5. Компьютерная лингвистика и теория я знаний

Компьютерная, лингвистика – самостоятельное направление в прикладной лингвистике, ориентированное на использование компьютеров для решения задач, связанных с использованием естественного языка. Иногда компьютерную лингвистику считают разделом информатики, поскольку это направление применяет методы информатики к особому объекту – естественному языку. Такое неоднозначное отнесение компьютерной лингвистики и к сфере языкознания, и к информатике говорит о том, что мы имеем дело с научным направлением, образовавшимся на стыке двух наук.

Две причины обусловили появление новой науки. Во-первых, исследователи-лингвисты надеялись, что современные точные науки (и прежде всего математика) помогут лингвистике обрести большую точность. Появление компьютеров укрепило эти надежды, так как многим языковедам с самого начала было ясно, что компьютер – это инструмент, который позволяет автоматизировать работу с текстами, тем самым упрощая и ускоряя исследования. Компьютеру можно поручить рутинную работу,

которая отнимает много времени: статистический анализ различных языковых единиц, ведение словарных и лексических картотек, создание лексикографических баз данных и корпусов текстов.

Во-вторых, с появлением компьютеров почти сразу же возникла проблема общения с ними неподготовленных пользователей. Наиболее удобной формой общения человека и компьютера является естественный язык. Но для организации человеко-машинного взаимодействия на естественном языке надо было прежде понять законы и особенности использования языка в процессе общения людей между собой.

У лингвистов появилась возможность проверять теоретические гипотезы о языке с помощью компьютеров, и это стимулировало появление новых идей.

В современной компьютерной лингвистике широко используется терминология когнитивной науки и эпистемологии (теории знаний). Это связано с тем, что идеи, разработанные в рамках теории знаний, лежат в основе компьютерного моделирования естественного языка.

В теории знаний выделяют два основных вида знаний: декларативные знания («знание ЧТО») и процедурные знания («знание КАК»).

Декларативные знания – это утверждения о чем-либо. Типичный пример декларативного знания – толкование слова в толковом словаре.

Процедурные знания – это алгоритм, то есть последовательность действий, которые следует выполнить в определенной ситуации. Пример процедурных знаний – инструкция к любому бытовому прибору. Процедурным знанием также является функциональная способность человека к ходьбе, бегу, использованию языка.

Основное различие между декларативными и процедурными знаниями состоит в следующем: если декларативные знания можно верифицировать, то есть определить их истинность или ложность, то процедурные знания верификации не подлежат. Оценить качество процедурных знаний можно только с точки зрения успешности или не успешности работы алгоритма.

§ 1.6. Структуры представления знаний

В теории знаний также изучаются различные структуры представления знаний: фреймы, сценарии, планы и т. д.

Фрейм – это структура для представления декларативного знания о какой-либо стереотипной ситуации. Формально фрейм представляется как совокупность узлов и отношений между ними. Каждый узел задает определенный параметр, который может заполняться конкретной информацией. Ядро фрейма составляют узлы, соответствующие постоянным для данной ситуации понятиям. Периферия фрейма – узлы, заполнение которых информацией каждый раз не является обязательным.

Помимо фрейма существуют и другие структуры представления знаний, например, сценарий. *Сценарий* – это тоже представление о стереотипной ситуации или стереотипном поведении, только элементами сценария

являются шаги алгоритма или инструкции. Можно, например, говорить о «сценарии посещения ресторана» или «сценарии покупки в магазине».

Другой термин, который используется в когнитивных исследованиях – план. Под *планом* понимают представление знаний о возможных действиях, которые необходимы для достижения определенной цели. План возникает на основе одного или нескольких сценариев, необходимых для решения данной проблемной ситуации. Выполнимость плана – обязательное условие его порождения, а к понятию «сценарий» критерий выполнимости неприменим.

Еще одно важное понятие теории знаний – *модель мира*. Под моделью мира понимается совокупность знаний о мире. В модели мира эти знания не хаотичны, а организованы определенным образом.

§ 1.7. Задачи прикладной лингвистики с использованием информационных технологий, будущее информационных технологий

Среди задач прикладной лингвистики с использованием возможностей ИТ можно назвать следующие:
- создание систем искусственного интеллекта;
- создание систем автоматического перевода;
- создание систем автоматического аннотирования и реферирования текстов;
- создание систем порождения текстов;
- создание систем обучения языку;
- создание систем понимания устной речи;
- создание систем генерации речи;
- создание автоматизированных информационно-поисковых систем;
- создание систем дешифровки анонимных текстов;
- разработка различных баз данных (словарей, каталогов, реестров и т.п.) для гуманитарных наук;
- разработка различного типа автоматических словарей;
- разработка систем передачи информации в сети Интернет и т.д.

Философы, психологи и другие специалисты отмечают, что в будущем социально защищенным может считаться лишь тот человек, который способен гибко перестраивать направление и содержание своей деятельности в связи со сменой технологий или требований рынка. Любой информационный ресурс представляет реальную ценность лишь в том случае, когда к нему организован соответствующий доступ. Интеллектуальная собственность, представленная в цифровом формате, станет главной «валютой» XXI века.

Задание: Перечислите те проблемы Вашей специальности, решением которых занимается прикладная лингвистика. Существуют уже готовые решения или проблемы находятся в стадии разработки?

Тема 2: ГИПЕРТЕКСТОВЫЕ ТЕХНОЛОГИИ

1. Из истории возникновения метода «гипертекст»
2. Отличие гипертекста от традиционного текста
3. Компоненты гипертекста
4. Типология гипертекста
5. Гипертекстовые системы
6. Роль лингвистов в создании гипертекста

§ 2.1. Из истории возникновения метода «гипертекст»

Идея гипертекста связывается с именем Ванневара Буша – советника президента Ф. Д. Рузвельта по науке, который в своей статье ‘*As we may think*’, опубликованной в журнале *The Atlantic Monthly* в 1945 г., описал настольный аппарат, названный им Мемекс (англ. *Memex* от ***memory extender*** ‘расширитель памяти’). Он задумывался как прибор, в котором человек хранит все свои книги, записи, сообщения, быстродействующий и удобный в обращении. Управляемый с помощью ручек, кнопок и клавиатуры и основанный на технологии микрофильмирования, Мемекс представляет собой механическую модель компьютера как средства воспроизведения и отображения информации. Более того, пользователь мог делать пометки и комментарии на полях так, словно перед ним была страница книги или журнала. Суть замысла заключалась в возможности Мемекс устанавливать ассоциативные связи между текстами. Автор представлял его как систему, которая работает так же, как работает человеческий мозг.

Однако отсутствие компьютерной техники сделало проект трудно реализуемым, поскольку механическая система оказалась чрезмерно сложной для практического воплощения. Идея В. Буша в 60-е гг. получила второе рождение в системе «Ксанаду» (англ. *Xanadu*) Теда Нельсона, которая уже предполагала использование компьютерной техники. «Ксанаду» позволял пользователю прочитывать совокупность введенных в систему текстов различными способами, в различной последовательности, программное обеспечение давало возможность как запоминать последовательность просмотренных текстов, так и выбирать из них практически любой в произвольный момент времени. Множество текстов со связывающими их отношениями (системой переходов) было названо Т. Нельсоном гипертекстом.

§ 2.2. Отличие гипертекста от традиционного текста

Существует различные определения гипертекста: «Гипертекст – это текст, связанный ссылками с другими текстами. <…> Гипертекст – текст, устроенный таким образом, что он превращается в систему, иерархию текстов, одновременно составляя единство и множество текстов. Простейший пример гипертекста – это любой словарь или энциклопедия, где каждая статья имеет отсылки к другим статьям этого же словаря» [4, с. 182].

Самый яркий пример современного гипертекста – всемирная паутина WWW (WorldWideWeb), состоящая из связанных воедино веб-сайтов. Там почти невозможно найти страницу, которая не была бы связана ссылками со всеми остальными веб-страницами Интернета. Именно поэтому она и получила название паутины.

Многие исследователи рассматривают создание гипертекста как начало новой информационной эпохи, противопоставленной эре книгопечатания. Линейность письма, внешне отражающая линейность речи, оказывается фундаментальной категорией, ограничивающей мышление человека и понимание текста.

Гипертекстовые же технологии позволяют легко сочетать различные виды информации – обычный текст, рисунок, график, таблицу, схему, звук и движущееся изображение.

Как традиционный текст, так и гипертекст – феномены, порожденные новыми технологиями. В первом случае технология позволила легко тиражировать и распространять знания самых различных типов, а во втором – компьютерные технологии дали возможность изменить сам внешний вид текста и его структуру.

Разнородность гипертекста – это первое технологическое свойство гипертекста, технологическое в том смысле, что оно непосредственно следует из используемой компьютерной технологии. Второе технологическое свойство гипертекста – его нелинейность. Гипертекст не имеет стандартной, обычной последовательности чтения. Прочие свойства гипертекста в той или иной степени являются следствиями из этих двух технологических свойств.

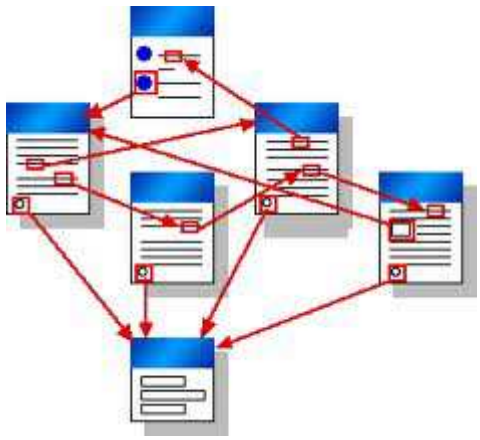
Суммировать различия текста и гипертекста можно следующим образом:
1) конечность, законченность традиционного текста vs. бесконечность, незаконченность, открытость гипертекста;
2) линейность текста vs. нелинейность гипертекста;
3) точное авторство текста vs. отсутствие авторства (в традиционном понимании) у гипертекста, снятие противопоставления между автором и читателем;
4) субъективность, односторонность обычного текста vs. объективность, многосторонность гипертекста;
5) однородность обычного текста vs. неоднородность гипертекста.

Гипертекст	Обычный текст
1) является открытым, незаконченным, бесконечным	1) является конечным, законченным
2) принципиально не линеен	2) линеен
3) в традиционном понимании не имеет автора	3) имеет автора
4) читатель сам определяет маршрут передвижения по тексту, может стать соавтором гипертекста, внося в него что-то новое	4) читатель, как правило, не может определять маршрут передвижения по тексту, не может стать его соавтором

5) сочетает в себе самые разнообразные способы передачи информации: текст, графику, видео, аудиофрагменты	5) однороден в плане способа представления информации

§ 2.3. Компоненты гипертекста

Структурно гипертекст может быть представлен как граф, в узлах которого находятся традиционные тексты или их фрагменты, изображения, таблицы, видеоролики и т.д. Узлы связаны разнообразными отношениями, типы которых задаются разработчиками программного обеспечения гипертекста или самим читателем.



Отношения задают потенциальные возможности передвижения или навигации по гипертексту. Отношения могут быть однонаправленными или двунаправленными. Соответственно, двунаправленные стрелки позволяют двигаться пользователю в обе стороны, а однонаправленные – только в одну. Цепочка узлов, через которые проходит читатель при просмотре компонентов текста, образует путь или маршрут.

Тип чтения гипертекста определяется не только маршрутом, но и качественными характеристиками, связанными с пониманием информации, содержащейся в узлах. Медленное чтение предполагает внимательное знакомство с информацией каждого узла. Часто медленное чтение сопровождается заметками, которые читатель может привязывать к узлам гипертекста.

Быстрое чтение – браузинг – наиболее часто используется в информационных системах, основанных на гипертекстовой технологии. При поиске конкретной информации пользователь быстро передвигается по узлам сети, маркируя нужные фрагменты. Для браузинга создается специальная программная поддержка.

Совокупность смежных узлов образует окрестность данного узла. Понятно, что окрестность узла образуют те узлы, в которых содержится информация, близкая по семантике к содержанию данного узла. Узлы сети, в которые входит и выходит много стрелок-отношений, образуют центральную

часть гипертекста, а те, которые почти изолированы от других узлов – его периферию.

§ 2.4. Типология гипертекста

Первое противопоставление относится к структуре гипертекста. Гипертекст может быть иерархическим или сетевым. Иерархическое – древовидное – строение гипертекста существенно ограничивает возможности перехода между его компонентами. В таком гипертексте отношения между компонентами напоминают структуру тезауруса, основанного на родо-видовых связях. Иерархический гипертекст не реализует всех возможностей технологии гипертекста. Сетевой гипертекст позволяет использовать различные типы отношений между компонентами, не ограничиваясь отношениями «род–вид».

Второе противопоставление характеризует не саму структуру гипертекста, а возможности программного обеспечения. Здесь различаются простые и сложные гипертексты. Примером простого программного обеспечения гипертекста может служить электронное оглавление документа, которое позволяет перейти к любой части оглавления, минуя этап просмотра всего текста. К простому гипертексту относится и система, которая дает возможность просматривать отсылки к литературе, содержащиеся в тексте, не обращаясь непосредственно к списку литературы. Сложные гипертексты обладают богатой системой переходов между компонентами гипертекста, в них отсутствует представление о базовом тексте, с которым связаны второстепенные по значимости тексты.

По способу существования гипертекста выделяются статические и динамические гипертексты. Статический гипертекст не меняется в процессе эксплуатации; в нем пользователь может фиксировать свои комментарии, однако они не меняют существо дела. Для динамического гипертекста изменение является нормальной формой существования. Обычно динамические гипертексты функционируют там, где необходимо постоянно анализировать поток информации, то есть в информационных службах различного рода. Гипертекстовой является, например, Аризонская информационная система (*AAIS*), которая ежемесячно пополняется на 300-500 рефератов в месяц.

Отношения между элементами гипертекста могут изначально фиксироваться создателями, а могут порождаться всякий раз, когда происходит обращение пользователя к гипертексту. В первом случае речь идет о гипертекстах жесткой структуры, а во втором – о гипертекстах мягкой структуры. Жесткая структура технологически вполне понятна. Технология организации мягкой структуры должна основываться на семантическом анализе близости документов (или других источников информации) друг к другу. В настоящее время широко распространено использование технологий мягкой структуры на ключевых словах. Переход от одного узла к другому в сети гипертекста осуществляется в результате поиска ключевых слов.

Поскольку набор ключевых слов каждый раз может различаться, каждый раз меняется и структура гипертекста. Структура Интернета часто функционирует как гипертекст мягкой архитектуры.

Технология построения гипертекстовых систем не делает различий между текстовой и нетекстовой информацией. Между тем включение визуальной и звуковой информации (видеороликов, картин, фотографий, звукозаписей и т. п.) требует существенного изменения интерфейса с пользователем и более мощной программной и компьютерной поддержки. Такие системы получили название гипермедиа или мультимедиа. Наглядность мультимедийных систем предопределила их широкое использование в обучении, в создании компьютерных вариантов энциклопедий.

§ 2.5. Обзор некоторых гипертекстовых систем

Технологически в основе гипертекста лежат компьютерные программы, которые поддерживают следующие базовые функции:

• обеспечение быстрого просмотра информационного массива (браузинг);

• обработка ссылочных отношений (обращение и вызов фрагмента текста или другой информации, на которую производится отсылка);

• навигация по гипертексту, запоминание маршрута движения; представление пути движения в легко воспринимаемой форме;

• возможность формирования обычного линейного текста как результата движения по гипертексту;

• дополнение гипертекста новой информацией;

• введение новых отношений в структуру гипертекста (для систем с жесткой структурой).

Программные оболочки гипертекста, как правило, универсальны. Они могут использоваться в различных областях для создания тематически разных гипертекстов. Таковы, например, оболочка *ZOG* и разработанная на ее основе промышленная гипертекстовая система *KMS* (университет Карнеги-Меллон, США). Сферы применения этих гипертекстовых систем необычайно разнообразны – от работы с документацией и поддержки электронной почты до гипертекстов, предназначенных для экспертов, работающих над бюджетом.

Имеются и специализированные системы. Так, система *NoteCards* (продукт компании *Xerox PARC*) предназначена для аналитической работы, а система *WE*, моделирующая особенности получения нового знания – для помощи в авторской работе.

Наиболее популярны в настоящее время программные пакеты *HyperCard* для ЭВМ *Macintosh* компании *Apple*. Они относительно просты в использовании. Гипертекст в оболочке *HyperCard* представляется в виде каталожных карточек. Пользователь с помощью довольно простого интерфейса организует структуру карточки и устанавливает связи между

карточками. Пакеты *HyperCard* позволяют сочетать различные типы информации, в частности карточки могут включать графическую, звуковую и другую информацию.

Следует отметить, что современные базы данных также включают поля для визуальной и звуковой формы данных (например, базы данных *ACCESS* 7, работающая в среде *Windows, Toolbook* для ЭВМ *PC/Windows*). Близка к *HyperCard* по своим свойствам и программа *SuperCard* фирмы *Silicon Beach*, а также классические системы гипертекста *Hyperties, KMS, NoteCards, SEPIA.*

Некоторые системы гипертекста содержат специальные средства ориентации пользователя в гиперпространстве – карты или закладки, отмечающие наиболее посещаемые узлы гипертекста. Комплексом средств ориентации обладает система *Hypergate Writer* фирмы *Eastgate Systems Inc*.

Метод гипертекста с успехом применяется в различных учебных курсах. Примером может служить известный курс английской литературы в Брауновском университете США. Этот гипертекст предназначен как для преподавателей (он помогает им организовывать и представить учебный материал), так и для студентов (он помогает им изучать учебный материал и добавлять к нему свои аннотации и доклады). Студенты, интересующиеся биографией какого-либо писателя, могут проследить в хронологическом порядке политические события, имевшие место в период его жизни, или подобрать материал, в котором сопоставляются особенности творчества современников этого писателя.

Курсы по истории искусств строятся на основе средств гипермедиа. Например, в гипертекстовом курсе истории музыки биографические данные о композиторе соединяются связями с его портретом, фотографией дома, где он родился, а текстовый рассказ об отдельных произведениях (письменный или устный) – с исполнением этих произведений или их фрагментов.

§ 2.6. Роль лингвистов в создании гипертекста

Для гипертекстов с мягкой структурой требуется разработка семантических процессоров, устанавливающих отношения семантической близости между документами в автоматическом режиме. Гипертексты с жесткой структурой требуют установления системы смысловых отношений между компонентами гипертекста, что является одной из важнейших задач лингвистической семантики и лингвистики текста. Еще одна задача, входящая в сферу интересов прикладной лингвистики – отбор информации в узлы сети гипертекста. Лингвиста должен определять, какие смысловые связи должны войти в гипертекст, а какие – нет. При этом принципы отбора смысловых отношений определяются практической ориентацией гипертекстовой системы (работа с документацией, работа над бюджетом, получение нового знания и т.д.).

Задание: Найдите не менее 10 различных ссылок в Интернете для следующих групп языков:
- романские;
- германские;
- балтийские;
- славянские;
- кельтские;
- финно-угорские;
- самодийские;
- тюркские;
- монгольские;
- индийские;
- семито-хамитские;
- нигеро-конголезские;
- дравидийские;
- австралийские.

Тема 3: АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТА
1. Распознавание текста
2. Анализ текста
3. Синтез текста
4.Текстовые редакторы
5. Текстовые процессоры
6. Возможности автоматического аннотирования и реферирования

Разработки в области автоматической обработки письменных текстов ведутся в трех направлениях:
• распознавание текста
• анализа текста
• синтез текста

§ 3.1. Распознавание текста

В наши дни для быстрого и качественного ввода текстовой информации в компьютер широко используются сканеры. Сканер работает по принципу фотоаппарата, позволяя ПК «увидеть» текст. Для того чтобы «понять» его содержание, т.е. перевести графическое (точечное) изображение символов в пригодную для дальнейшей обработки (редактирования, реферирования, перевода и т.д.) текстовую форму, необходима система автоматического чтения текста или оптического распознавания символов (OCR-система – *Optical Character Recognition*).

В классическом понимании система автоматического чтения текста – это компьютерная программа, позволяющая преобразовать текст с бумажного носителя в электронный текстовый файл, который может быть прочитан средствами обработки текстов. Исходный текст должен быть вначале введен в ПК с помощью сканера или получен на факс-модем.

История появления современных программ в области распознавания начинается с конца 40-х годов XX века, когда ученые многих стран стали работать над идеей обучения компьютера умению решать разные интеллектуальные задачи. Автоматическое чтение текста, распознавание речи, решение шахматных задач и головоломок и даже сочинение музыки и стихотворений – вот далеко не полный перечень идей, которые выдвигались и разрабатывались в то время. К концу 50-х годов эти идеи оформились в отдельную область знания – искусственный интеллект. Одной из задач, которая вскоре выделилась в отдельное направление, и была задача распознавания образов. Идеальная компьютерная система распознавания должна уметь формировать, анализировать и интерпретировать любое изображение, в том числе и символьное.

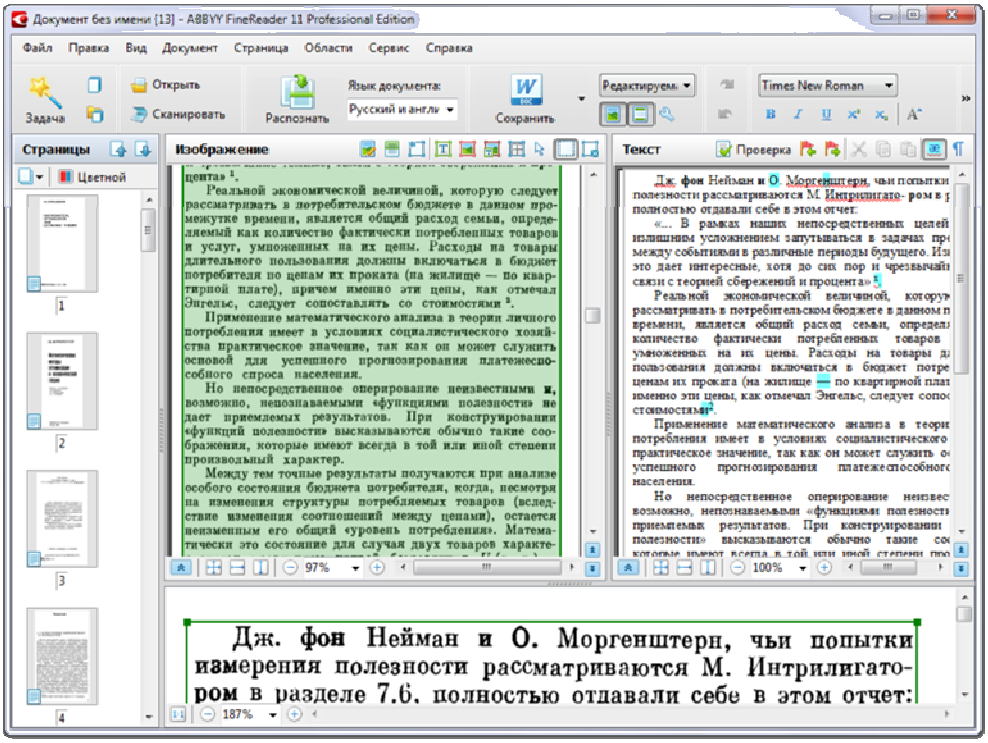
Несмотря на определенные трудности распознавания текстов (на одной странице может встретиться до трех и более шрифтов разного размера, стиля начертания и гарнитуры; тексты часто бывают многоязычными, многоколоночными, включают в себя таблицы, графические изображения и т.д.), возможности систем автоматического чтения текста огромны, а

точность распознавания OCR-систем на текстах хорошего и среднего качества достигает в настоящее время достигает 97–99 %.

Начальной характеристикой текста на естественном языке, введенного в память компьютера, является его буквенный состав. Сюда входят буквы алфавита, знаки препинания, другие графемы (скобки, кавычки, тире). Технологии распознавания текстов базируются на статистических данных о частоте употребления графем, а также на данных о возможных правилах сочетаемости графем в конкретном языке и о частотности определенных буквосочетаний. Данные о графематическом составе языка используются в программах оптического распознавания символов (в настоящее время во всем мире широко известны две OCR-системы, созданные российскими разработчиками. Это *FineReader* компании *ABBYY Software House* и *CuneiForm* фирмы *Congitive Technologies*), которые позволяют при помощи сканера перенести в электронный вид текст с бумаги.

ABBYY FineReader позволяет извлекать текстовые данные из цифровых изображений (фотографий, результатов сканирования, PDF-файлов). Полученное в результате распознавания может быть сохранено в различных форматах файлов: *Microsoft Word, Microsoft Excel, Microsoft Powerpoint, Rich Text Format, HTML, PDF/A, searchable PDF, CSV* и текстовые *(plain text)* файлы.

Поддерживает распознавание текста на 188 языках и имеет встроенную проверку орфографии для 45 из них.



CuneiForm Cognitive OpenOCR свободно распространяемая открытая система оптического распознавания текстов российской компании *Cognitive Technologies*. *CuneiForm* позиционируется как система преобразования

электронных копий бумажных документов и графических файлов в редактируемый вид с возможностью сохранения структуры и гарнитуры шрифтов оригинального документа в автоматическом или полуавтоматическом режиме.

§ 3.2. Анализ текста

Анализ текстов на естественном языке является необходимым этапом работы систем машинного перевода, а также информационно-поисковых систем. Чтобы проводить анализ текста автоматически, необходимо ответить на вопрос, существуют ли строгие формальные правила, по которым строится структура предложения и структура текста.

В результате проведенных исследований стало ясно, что за каждым текстом (в том числе и за отдельным предложением, являющимся своего рода мини-текстом) скрывается не одна, а несколько формальных структур, которые можно разделить на три уровня.

Первый уровень – это поверхностная синтаксическая структура. В этой структуре каждое предложение текста рассматривается изолированно от других и для каждого проводится что-то вроде разбора предложения по составу. Выделяются подлежащее и сказуемое, определения, дополнения и обстоятельства разного вида. Но этой структуры для анализа оказывается мало.

Следующий шаг – построение глубинной синтаксической структуры. Идея существования глубинной синтаксической структуры связана с пониманием того, что различные естественные языки, отличаясь друг от друга многими внешними синтаксическими особенностями, передают весь спектр взаимосвязей между объектами, явлениями, их свойствами и протекающими с их участием процессами, характерными для окружающего мира. И этот мир един, каким бы языком мы его ни описывали. Следовательно, в каждом тексте существуют не зависящие от особенностей языка некие глубинные структуры, которые определяют адекватное отображение той или иной ситуации в окружающем мире.

С этой идеей тесно связано использование так называемых глубинных падежей, или падежей Чарльза Филлмора, названных по имени американского исследователя, впервые введшего их в научный оборот.

Рассмотрим две фразы: «Мальчик сорвал цветок» и «Цветок был сорван мальчиком». В первом предложении субъект действия «сорвал» – это «мальчик». И это слово играет здесь роль подлежащего, о чем свидетельствует именительный падеж. Во втором же предложении роль подлежащего играет слово «цветок», а слово «мальчик» стоит в творительном падеже. Но субъектом действия «сорвал» и здесь остается все тот же «мальчик». А цветок в любом из двух приведенных предложений играет роль объекта действия. Понимание ситуации, описываемой любым из этих предложений, заключается, в частности, в том, что мы выделяем в тексте некоторое действие, а также его субъект и объект. Позиции субъекта и

объекта служат примером глубинных падежей, которые ввел Ч. Филлмор. Эти два падежа (субъектный и объектный) не единственные. Разные исследователи выделяют разное количество таких падежей (инструментальный, временной, пространственный и т.д.).

Синтаксическая структура, построенная на основе глубинных падежей, позволяет перейти от синтаксического уровня предложения к его семантическому уровню. На этом уровне для анализа привлекаются дополнительные данные, связанные с наличием у лексических единиц языка (в частности, слов) определенных значений. Сами значения известны носителю языка и хранятся в его памяти. Обращение к памяти позволяет приписать элементам предложения соответствующие им значения и использовать их для понимания текста на семантическом уровне.

В семантических структурах (третий уровень формальных структур) также можно выделить поверхностный и глубинный уровни, в чем-то похожие на соответствующие уровни в синтаксических структурах. Структуры наиболее «глубокого» уровня, возникающие при анализе предложений, могут быть названы прагматическими. Из них следует понимание того, какова коммуникативная цель данного текста. Прагматические структуры связывают текст в единое целое и оказывают определенное влияние на адресата текста.

В процессе анализа текстов, содержащих более одного предложения, возникают новые структуры, обеспечивающие сцепление этих предложений в рамках некоторой описываемой ситуации или последовательности ситуаций. Возникают межфразовые связи, позволяющие понять текст как единое целое. Эти структуры пока изучены значительно хуже, чем структуры, лежащие в основе одного предложения.

Таким образом, наличие нескольких уровней формальных структур в тексте приводит к тому, что для их выделения при автоматическом анализе надо пройти несколько последовательных этапов:

1) исходный текст →
2) морфологический анализ →
3) поверхностный синтаксический анализ →
4) глубинный синтаксический анализ →
5) поверхностный семантический анализ →
6) глубинный семантический анализ →
7) прагматический анализ →
8) выявление текстовых структур.

Указанные этапы охватывают всю задачу анализа текстов на естественном языке. Необходимость в исполнении тех или иных этапов при анализе конкретного текста зависит от тех целей, для которых тот анализ осуществляется.

§ 3.3. Синтез текста

Синтез текстов на естественном языке также является необходимым этапом работы систем автоматического перевода, систем автоматического реферирования, экспертных систем. Синтез текста – это задача, которая может рассматриваться как обратная по отношению к анализу.

Если заданы некоторая тема и цель будущего текста, то можно считать заданной прагматическую структуру текста. Ее надо разложить на прагматические структуры отдельных предложений и для каждого предложения пройти все этапы анализа в обратном направлении. На сегодняшний день здесь еще масса нерешенных проблем. Особенно много вопросов вызывает возможность синтеза прагматической структуры текста из тех целей, которые стимулируют создание текста. Например, неясно, как эту структуру разбить на прагматические структуры предложений и как от этих частных прагматических структур перейти к глубинным семантическим структурам.

Один из возможных путей состоит в использовании актантов действий. С каждым действием связан некоторый набор сопутствующих ему объектов и характеристик. Они, как правило, совпадают с глубинными падежами Ч. Филлмора. Наличие актантных структур действий позволяет представить процесс синтеза текстов в виде ряда следующих друг за другом шагов.

На первом шаге генерируется нужная последовательность глаголов-действий. На следующем шаге заполняются их актантные структуры, что приводит к появлению глубинной семантической структуры отдельных предложений. Затем эти структуры связываются с учётом общих действующих субъектов и используемых объектов, а также иных параметров в единый текст. Последний шаг – образование синтаксически правильных конструкций в предложениях.

§ 3.4. Текстовые редакторы

Текстовый редактор – компьютерная программа, предназначенная для обработки текстовых файлов (создание и внесение изменений). Условно выделяют два типа редакторов: потоковые и интерактивные.

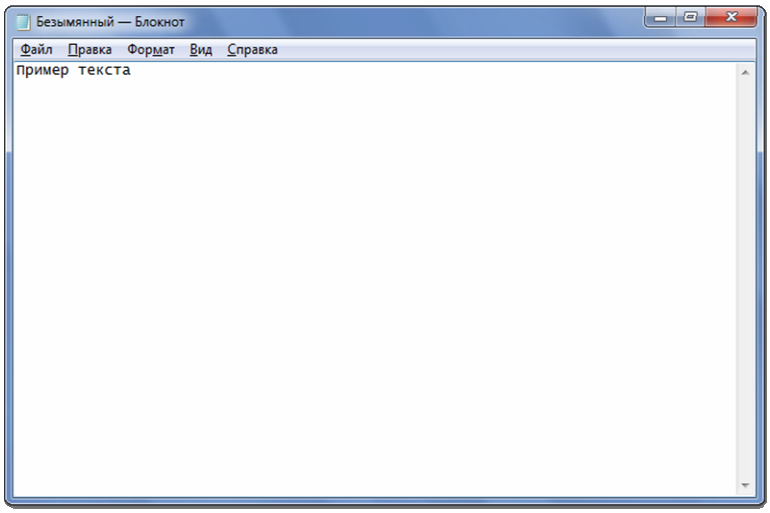
Потоковые текстовые редакторы представляют собой компьютерные программы, которые предназначены для автоматизированной обработки входных текстовых данных в соответствии с заранее заданными пользователями правилами. Примером такого текстового редактора может служить редактор *Sed (Stream Editor*). Sed вначале загружает в себя набор команд, а затем применяет их к каждой строчке текста.

Интерактивные текстовые редакторы – это семейство компьютерных программ предназначенных для внесения изменений в текстовый файл в интерактивном режиме, т.е. они сначала загружают текст, а затем используются команды. Такие программы позволяют отображать текущее состояние текстовых данных в файле и производить над ними различные действия.

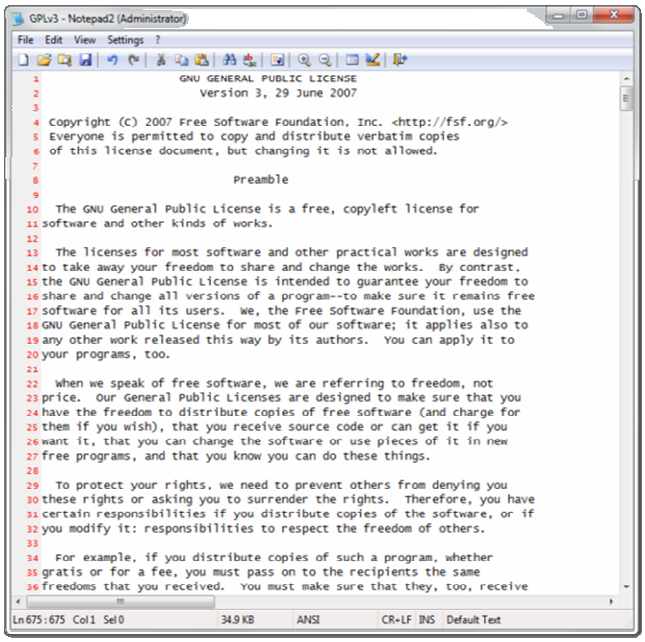
Часто интерактивные текстовые редакторы содержат значительную дополнительную функциональность, н-р, подсветка синтаксиса.

Примеры текстовых редакторов: Блокнот (*Notepad*), *WordPerfect (Corel WordPerfect)*, *Notepad++* (все для *Windows)*, *Лексикон* (для *DOS*), *WordPad* (среднее между текстовым редактором и процессором) и др. Рассмотрим некоторые из них.

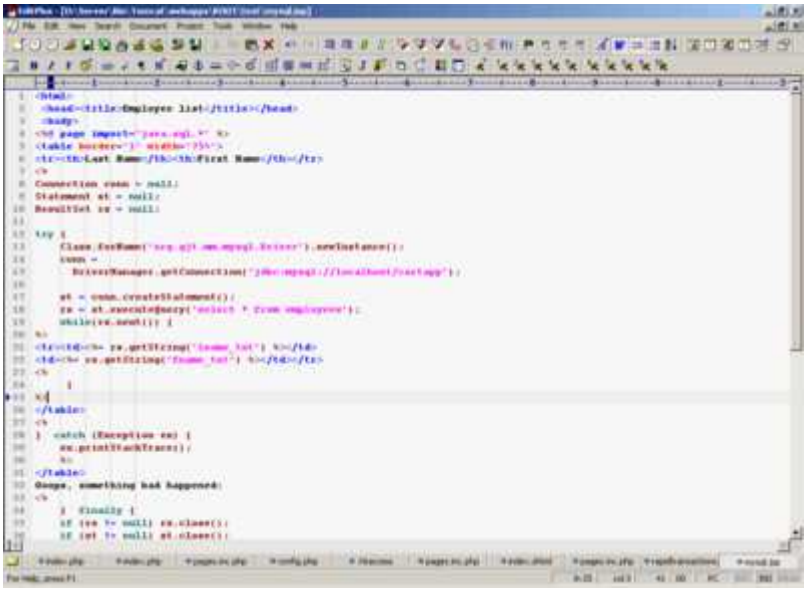
Блокнот *(Notepad)* – простой текстовый редактор, являющийся частью операционных систем Microsoft Windows, начиная с вышедшей в 1985 году *Windows 1.0*, и *Windows NT*.



Notepad2 – свободный текстовый редактор с открытым исходным кодом для *Microsoft Windows*. Программа написана Флорианом Балмером с помощью компонента *Scintilla* в апреле 2004 года. Текстовый редактор построен на принципах *Microsoft Notepad*, является маленьким, быстрым и полезным.



EditPlus – текстовый редактор, редактор веб-страниц и редактор программиста для *Windows*, ориентированный на Интернет. Может служить отличной заменой стандартному *Блокноту Windows*, и в то же время имеет множество мощных и удобных возможностей для создателей веб-страниц, переводчиков программ, программистов.

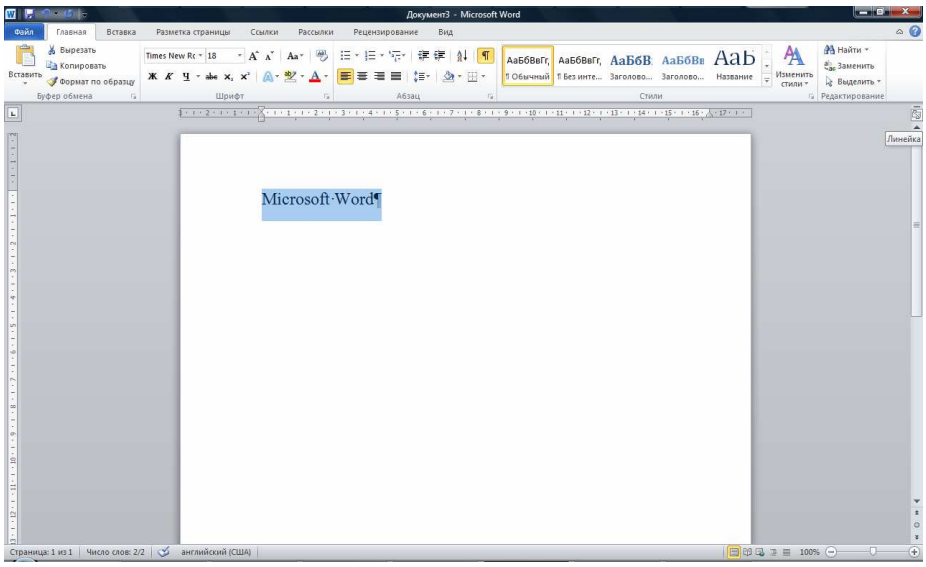


§ 3.5. Текстовые процессоры

Строго говоря, текстовый процессор может быть причислен к интерактивным текстовым редакторам, однако для данного класса компьютерных программ их возможность применения в качестве интерактивного текстового редактора не является главной целевой.

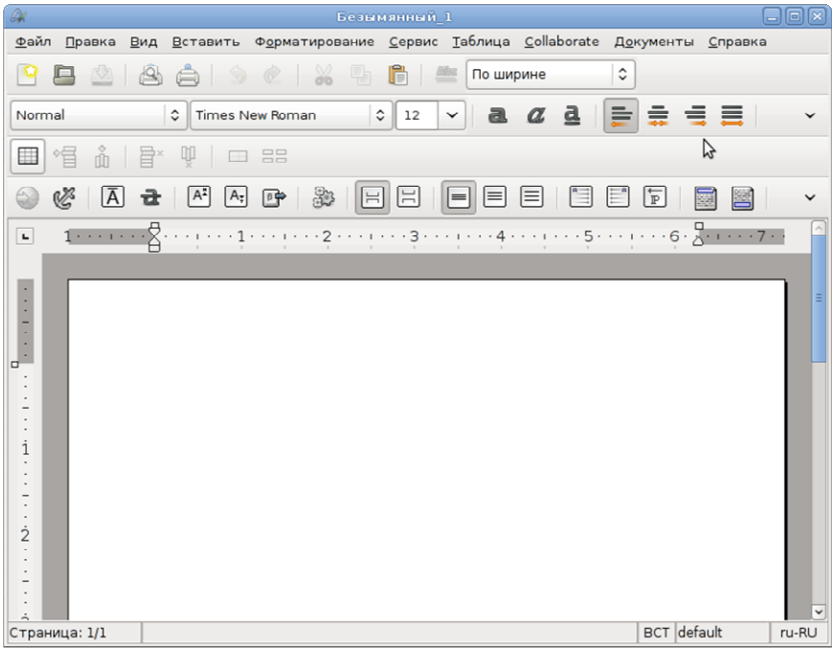
Текстовыми процессорами в 1970-80-е гг. называли предназначенные для набора и печати текстов машины индивидуального и офисного использования, состоящие из клавиатуры, встроенного компьютера для простейшего редактирования текста, и электрического печатного устройства. Позднее название «текстовый процессор» стало использоваться для компьютерных программ, предназначенных для аналогичного использования.

Текстовые процессоры, в отличие от текстовых редакторов, имеют больше возможностей для форматирования текста, внедрения в него графики, формул, таблиц и других объектов. Поэтому они могут быть использованы не только для набора текстов, но и для создания различного рода документов, в том числе официальных. Наиболее известным примером текстового процессора является *Microsoft Word*.

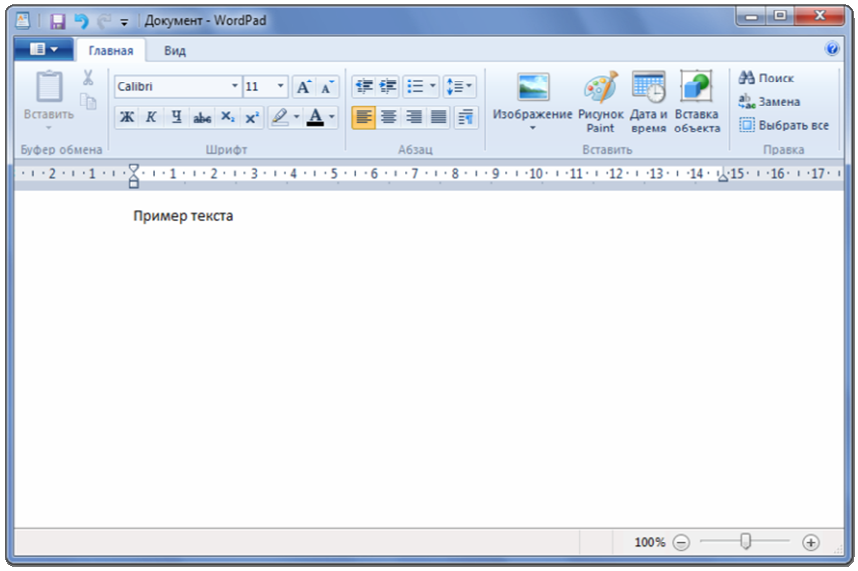


Кроме него, к наиболее популярным текстовым редакторам относятся следующие – *OpenOffice.org Writer*, *WordPerfect*, *AbiWord*, *WordPad* и др. Рассмотрим некоторые из них.

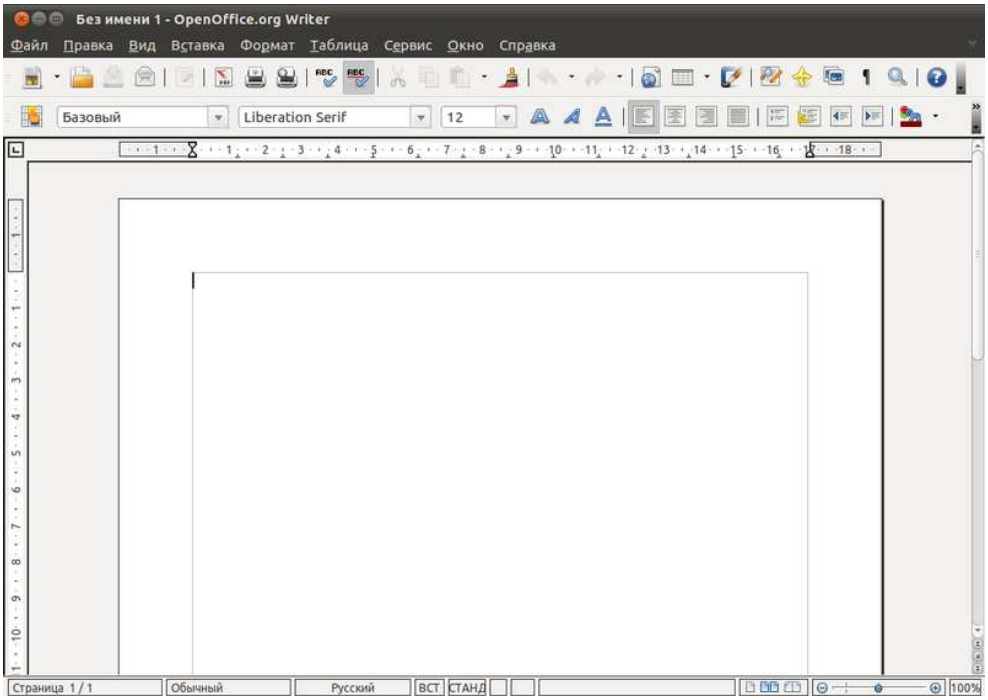
AbiWord – свободный текстовый процессор, распространяемый согласно *GNU General Public License*. Поддерживается на платформах *Linux, Mac OS X (PowerPC), Microsoft Windows, ReactOS, SkyOS, BeOS* и других.



WordPad – текстовый редактор, входящий в состав Microsoft Windows, начиная с *Windows 95*. Обладает бóльшим набором инструментов, чем *Блокнот*, но не дотягивает до уровня полноценного текстового процессора типа *Microsoft Word* или *OpenOffice.org Writer*. *WordPad* представляет собой эволюционировавшую версию программы *Windows Write* из *Windows 1.0*. Поддерживает форматирование и печать текста, но не имеет ряда таких важных инструментов, как таблицы и средств проверки орфографии.



OpenOffice.org Writer – текстовый процессор и визуальный редактор *HTML*, входит в состав *OpenOffice.org* и является свободным программным обеспечением (выпускается под лицензией *LGPL*). Также имеет некоторые возможности, отсутствующие в *Word*, например, поддержку стилей страниц. *Writer* позволяет сохранять документы в различных форматах, включая *Microsoft Word, RTF, XHTML и OASIS Open Document Format*.



WordPerfect – компьютерная программа для электронной обработки текстов. Пик её популярности приходится на конец 1980-х – начало 1990-х годов XX века. В это время программа фактически являлась стандартным текстовым редактором, который затем был вытеснен с рынка редактором *Microsoft Word*.

§ 3.6. Возможности автоматического аннотирования и реферирования

Рефератом называется связный текст, который кратко выражает центральную тему или предмет какого-либо документа, цель, применяемые методы, основные результаты описанного исследования или разработки.

Рефераты обычно составляют к научно-техническим документам, статьям, патентам на изобретение и т.п. Реферат помогает человеку ориентироваться в информационных потоках, оперативно отбирать для себя наиболее ценную и полезную информацию. Процесс составления реферата называется реферированием.

Аннотацией называют краткое изложение содержания документа, дающее общее представление о его теме.

Таким образом, если реферат в краткой форме знакомит читателя с сутью излагаемого в документе содержания (фактами, методикой, экспериментами и т.п.), то аннотация выполняет лишь сигнальную функцию, сообщая о том, что опубликована статья или книга на определенную тему. Процесс составления аннотации называется аннотированием.

Рефераты и аннотации представляют собой вторичные документы. Первичные, или исходные, документы – это книги, статьи, патенты и т.п.

В каждом вторичном документе можно выделить два компонента информации: 1) содержательный, 2) документографический.

Первый компонент содержит информацию первоисточника (о чем книга, статья). Второй компонент – это сведения о самом первичном документе (тип документа: книга, статья и т.п.; вид: печатный, рукописный; год издания; место издания и т.д.). В дальнейшем речь пойдет только о первом компоненте вторичного документа.

Научно-технический прогресс привел к появлению большого числа публикаций (книг, статей и т.п.) по самым разным проблемам науки, техники, образования, и специалисты не успевают следить за новейшей литературой по своей области знания. Для этого, как установлено, человек должен был бы прочитывать ежедневно 1500 страниц текста на разных

языках, что явно превышает его физические возможности. Поэтому для оперативного «поверхностного» знакомства с новейшими публикациями используются рефераты и аннотации книг и статей, которые составляются в специальных организациях и публикуются в реферативных журналах и реферативных сборниках.

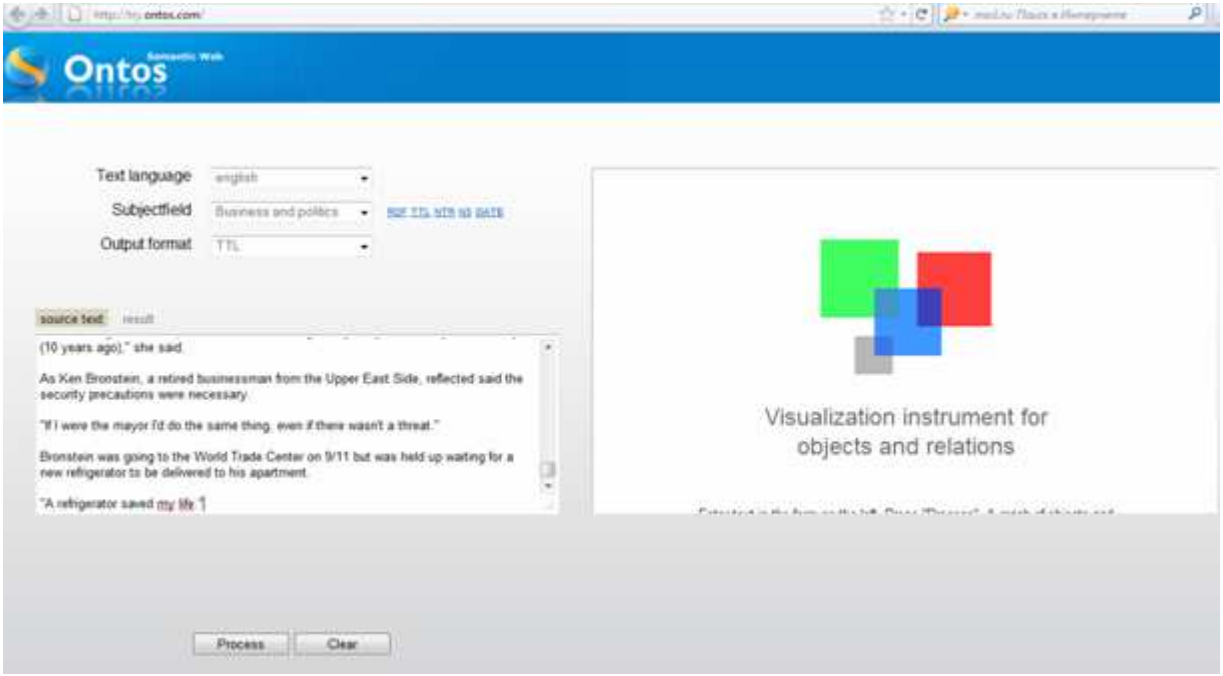
Реферирование и аннотирование текста являются довольно сложными и трудными видами интеллектуальной деятельности. Составление человеком рефератов или аннотаций занимает много времени. Это приводит к тому, что до ученых, педагогов, инженеров и других специалистов новейшая информация (особенно зарубежная) доходит очень медленно, что, в свою очередь, ведет к повторению в разных странах и в пределах одной страны одних и тех же исследований, более позднему применению новейших методик, технологий, процессов. Чтобы как-то избежать этого, для составления рефератов и аннотаций применяют современные компьютеры.

Составление реферата или аннотации текста с помощью компьютера называется автоматическим реферированием или аннотированием. Например, система автоматического реферирования «Либретто» была разработанная по технологии компании «МедиаЛингва». Она осуществляет автоматическое реферирование русских и английских текстов любого объема и любого уровня сложности. Исходный текст может сжиматься с необходимым пользователю коэффициентом сжатия, а реферат выдается в виде цепочки ключевых предложений или ключевых слов (аннотация).

На западном рынке к системам подобного типа относятся системы автоматического реферирования *Inxight Summarizer*, *Prosum*, *Researcher*. Несколько упрощенный вариант реферата в виде последовательности именных групп выдают системы *Extractor* и *TextAnalyst*, разработанные в инновационном центе «Микросистемы» в Москве.

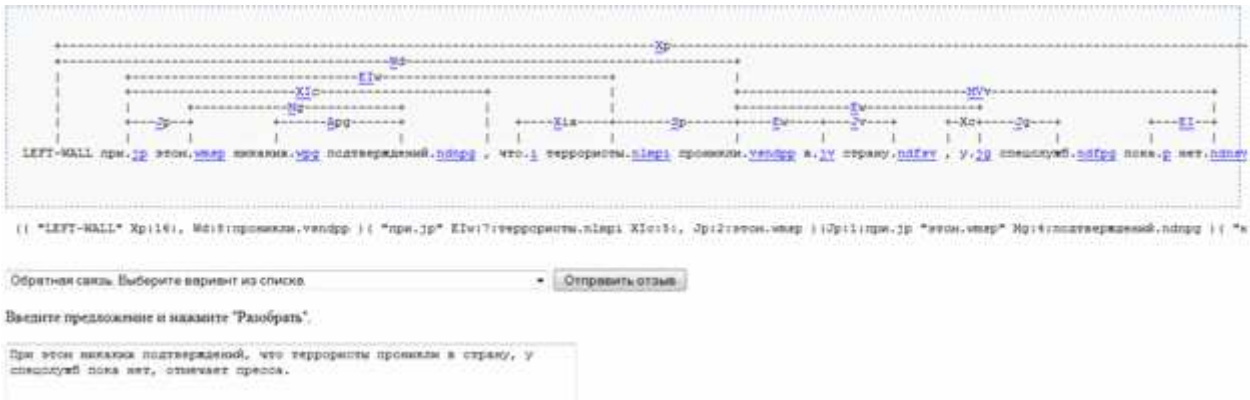
Рассмотрим некоторые программы, используемые при составления аннотаций.

Ontos – программные продукты предназначены для анализа текстовых документов, составления аннотаций, обработки данных *(OntosMiner, LightOntos for Workgroups, Ontos SOA, TAIS Ontos)*. Алгоритмы функционирования основаны на графематическом, морфологическом и семантическом анализе текстовой информации. Системы используют морфологические словари для основные языков (английского, немецкого, французского, русского). Обеспечивается выявление фактографической информации и представление ее в форме различного вида отчетов, в том числе в виде графа связей объектов.



Программа ***Galaktika-ZOOM*** позволяет выявлять значимые слова и словосочетания документа, проводить поиск документов по вводимым пользователем ключевым словам с учетом их синонимов, а также формировать отчеты по частоте встречаемости слов в документах. Программа обеспечивает обработку русскоязычных текстов. Алгоритмы основаны на использовании морфологического и статистического анализа.

Link Grammar Parser for Russian *(С. Протасов)* – on-line программа синтаксического анализа предложений русского языка. Алгоритм работы синтаксического анализатора основан на использовании грамматики связей.



Задание: Ознакомьтесь с указанными в данном разделе программами. Найдите другие программы по автоматической обработке текстов естественного языка. Постарайтесь найти программы, которые могут помочь вам в Вашей специальности.

Тема 4: АВТОМАТИЧЕСКАЯ ОБРАБОТКА ЗВУЧАЩЕЙ РЕЧИ

1. Особенности автоматической обработки звучащей речи
2. Практическое применение систем автоматической обработки звучащей речи
3. Проблемы синтеза звучащей речи
4. Структура программ распознавания и синтеза звучащей речи
5. Основные задачи систем обработки и распознавания звучащей речи
6. Обзор некоторых программы распознавания и синтеза звучащей речи

§ 4.1. Особенности автоматической обработки звучащей речи

Потребность в создании систем автоматической обработки естественного языка возникла постольку, поскольку невозможно обучить всех пользователей программированию. Оптимальной формой диалога человека и компьютера является диалог на естественном языке. А так как естественный язык существует в двух формах – письменной и устной, то и создание систем автоматической обработки естественного языка ведется в двух направлениях: обработка устной речи и обработка письменного текста.

Под обработкой устной речи понимается разработка методов, технологий и конкретных систем, которые обеспечивают общение человека с компьютером на естественном или ограниченно естественном языке. Речевой диалог обладает рядом преимуществ по сравнению с традиционным вводом информации с помощью клавиатуры:

1) устное общение не требует специальной предварительной подготовки пользователя;

2) диалог освобождает руки и зрение;

3) за счет системы распознавания голоса возможна защита от недозволенного доступа к объекту;

4) диалоговое взаимодействие дает возможность пользоваться компьютером людям с ограниченными возможностями.

Однако связь с помощью голоса имеет и свои недостатки: подверженность шумовым помехам, невозможность неограниченного ввода данных в компьютер в течение длительного времени.

§ 4.2. Практическое применение систем автоматической обработки звучащей речи

Системы автоматической обработки устной речи находят практическое применение в информационно-справочных службах, где можно получать информацию из базы данных в режиме диалога (например, в медицине или на транспорте). Кроме того, такие системы необходимы и для организации приема и озвучивания сообщений (например, получение электронной почты по телефону), а также для перевода звучащей речи в привычный текст в электронной форме. Компьютеры могут оказывать помощь и при обучении иностранному языку с помощью автоматических фонетических тренажеров.

История практического применения систем автоматической обработки звучащей речи началась еще в XVIII в., когда появились первые механические синтезаторы речи. Их создатели ставили целью воспроизвести процессы произнесения звуков с помощью механического устройства, имитируя строение голосового аппарата человека.

В начале XX века механические устройства сменились электрическими вокодерами. Первое устройство для распознавания речи появилось в 1952 г., оно могло распознавать произнесённые человеком цифры. В 1964 г. на ярмарке компьютерных технологий в Нью-Йорке было представлено устройство *IBM Shoebox*.

Коммерческие программы по распознаванию речи появились в начале 90-х годов. Обычно их используют люди, которые из-за травмы руки не в состоянии набирать большое количество текста. Эти программы (н-р, *Dragon NaturallySpeaking, VoiceNavigator*) переводят голос пользователя в текст, таким образом, разгружая его руки. Надёжность перевода у таких программ не очень высока, но с годами она постепенно улучшается.

Увеличение вычислительных мощностей мобильных устройств позволило и для них создать программы с функцией распознавания речи. Среди таких программ стоит отметить приложение *Microsoft Voice Command*, которое позволяет работать со многими приложениями при помощи голоса. Например, можно включить воспроизведение музыки в плеере или создать новый документ.

Прогресс, однако, не стоит на месте и в последнее время в телефонных интерактивных приложениях все чаще стали использоваться системы автоматического распознавания и синтеза речи. В этом случае общение с голосовым порталом становится более естественным, так как выбор в нем может быть осуществлен не только с помощью тонового набора, но и с помощью голосовых команд. При этом системы распознавания являются независимыми от дикторов, то есть распознают голос любого человека.

Следующим шагом технологий распознавания речи можно считать развитие так называемых *Silent Speech Interfaces (SSI*) (Интерфейсов Безмолвного Доступа). Эти системы обработки речи базируются на получении и обработке речевых сигналов на ранней стадии артикулирования. Данный этап развития распознавания речи вызван двумя существенными недостатками современных систем распознавания: чрезмерная чувствительность к шумам, а также необходимость четкой и ясной речи при обращении к системе распознавания. Подход, основанный на *SSI*, заключается в том, чтобы использовать новые сенсоры, не подверженные влиянию шумов в качестве дополнения к обработанным акустическим сигналам.

На сегодняшний день существует два типа систем распознавания речи – 1) работающие по принципу «клиент-сервер» (*client-server*), 2) «на клиенте» (*client-based*). При использовании клиент-серверной технологии речевая команда вводится на устройстве пользователя и через Интернет передается

на удаленный сервер, где обрабатывается и возвращается на устройство в виде команды (*Google Voice, Vlingo*); ввиду большого количества пользователей сервера система распознавания получает большую базу для обучения.

Первый вариант работает на иных математических алгоритмах и встречается редко (*Speereo Software*) – команда вводится на устройстве пользователя и обрабатывается в нем же. Плюс обработки «на клиенте» в мобильности, независимости от наличия связи и работы удаленного оборудования. Так, система, работающая «на клиенте» кажется надежнее, но иногда ограничивается мощностью устройства на стороне пользователя.

§ 4.3. Проблемы синтеза звучащей речи

На основе анализа существующих методов, можно сделать вывод о наличии следующих проблем в области синтеза речи:

1) искусственность речи;
2) отсутствие эмоциональной нагрузки;
3) низкая помехоустойчивость синтезированной речи.

Проблема искусственности речи заключается в том, что, несмотря на кажущееся качество произношения текста речевыми синтезаторами, такая речь тяжела для восприятия и понимания человеком. В основу технологии речевого синтеза положено использование заранее записанной фонетической базы и слова формируются с помощью статистического расчёта по принципу максимального правдоподобия фонетической сочетаемости, а пробелы и недочеты фильтруются человеческим мозгом. То есть качественный синтезатор с хорошо подобранной фонетической базой может восприниматься на слух в течение 10-15 минут, после чего синтезируемая речь перестает быть понятной. Это связано с тем, что для прослушивания синтезируемой речи человек использует дополнительные центры обработки головного мозга, и мозг просто устает. Таким образом, головной мозг не воспринимает синтезированную речь как естественную, которая сразу обрабатывается в речевом центре. Подобный эффект сравним с изучением иностранного языка.

Второй проблемой в области синтеза речи является отсутствие эмоциональной нагрузки, то есть личного восприятия произносимого текста читателем. При чтении текста человеком, он, поневоле, пропускает смысл воспроизводимого через себя, и в интонациях и нюансах чувствуется его отношение к прочитанному. Современные программы этого не могут, однако самые передовые из них пытаются имитировать интонацию путем модуляции тембра, длительности фонем и пауз. Но данная реализация является лишь подражанием, в связи, с чем мозг быстро устает исправлять огрехи воспроизведения, и слушатель теряет нить повествования.

Очевидно, что для решения этой задачи требуются методы из области теории искусственного интеллекта для «извлечения смысла» из воспроизводимого текста. Поэтому синтезаторы, учитывающие смысл

воспроизводимого текста должны строиться с учётом результатов междисциплинарных исследований.

Третья проблема – низкая помехоустойчивость синтезированной речи. Как показали и показывают эксперименты, достаточно наличие лишь слабого источника шума, чтобы слушатель перестал воспринимать смысл текста, воспроизводимого речевым синтезатором. Объяснение этому также находится в области нейрофизиологии. Так как для обработки синтезированной речи головной мозг использует дополнительные центры, то при наличии постороннего шума, разговора или необходимости выполнения слушателем какой-то работы, мозг просто не справляется, и человек перестает понимать смысл произносимого. Эффект помех существенно ограничивает возможности применения синтезатора в реальных условиях техногенных и природных шумов.

§ 4.4. Структура программ распознавания и синтеза звучащей речи

Появление компьютеров оказало мощное влияние на технологию синтеза речи, т.к. от аналоговых устройств исследователи перешли к цифровым. Однако на начальном этапе использования компьютеров для синтеза речи исследователи использовали ограниченное количество речевых образцов, которые хранились в памяти компьютера, поэтому результатом работы первых систем был не собственно синтез звучащей речи, а восстановление этих образцов.

С 60-х годов XX века перед исследователями встала задача озвучивания любого сообщения, подобно тому, как человек читает тексты вслух. В результате получили развитие синтезаторы типа «Текст – Речь». В них впервые появился этап предварительной лингвистической обработки текста.

Современные синтезаторы речи включают два блока: блок лингвистической обработки текста, с помощью которого строится полная фонетическая транскрипция синтезируемого текста, а также блок акустического синтеза, который генерирует речевой сигнал.

Блок лингвистической обработки текста имеет достаточно сложную структуру, поскольку создание транскрипции включает несколько этапов: определение языка входного текста, устранение возможных орфографических ошибок, проведение морфологического анализа словоформ для постановки ударения. Самая трудная задача этапа лингвистической подготовки текста – формирование интонации и просодических характеристик фразы. Во многих случаях для этого необходим значительно более сложный семантический и синтаксический анализ фразы. Последний этап работы блока лингвистической подготовки текста – создание фонетической транскрипции. На этом этапе применяются стандартные правила чтения, при этом сложность и трудоемкость этого этапа определяется соотношением между орфографией и произношением каждого конкретного языка.

После создания фонетической транскрипции начинает работу второй блок синтезатора блок акустического синтеза. Его задача – перевод транскрипции в цифровой сигнал, который, в свою очередь, преобразуется в звуковые колебания при помощи обычного цифро-аналогового преобразователя.

§ 4.5. Основные задачи современных систем распознавания и обработки звучащей речи

Для создания систем автоматического распознавания речи необходимо решить чрезвычайно трудную задачу – формализовать естественный диалог. Трудность этой задачи не только практическая, но и теоретическая. Достаточно сказать, что до сих пор не существует единой теории диалога, в которой были бы учтены лингвистические, социологические и психологические данные исследований.

Задача систем автоматического распознавания речи состоит в установлении того, что было сказано, и выдаче результата, например, в виде фонетической транскрипции или записи другого вида. Для таких систем важно, чтобы не было никакой посторонней информации, кроме акустической. Иными словами, данные системы не ориентированы на распознавание смысла высказывания.

Имеющиеся лингвистические и акустические знания недостаточны для создания эффективной системы по автоматическому распознаванию речи, поэтому ученые обратились к спектральному анализу речевого сигнала.

Спектральный анализ предполагает установление того, какие частоты участвуют в образовании данного звука и какова их интенсивность по отношению друг к другу. В результате спектрального анализа ученые получают амплитудно-частотные спектры. Спектральный анализ стал методом анализа звуков, поскольку известно, что человеческое внутреннее ухо осуществляет предварительный спектральный анализ речевого сигнала непосредственно перед поступлением его в мозг.

Для проблемы автоматического распознавания существенны следующие параметры распознающей системы: количество распознаваемых единиц; ограничения, связанные с голосом диктора; распознавание интонации, акцентов, особенностей произношения, а также время и условия распознавания.

Сегодня наиболее успешно с распознаванием речи справляются те системы, которые используют статистические и вероятностные модели звучащей речи.

§ 4.6. Обзор некоторых программ распознавания и синтеза звучащей речи

Dragon Naturally Speaking – это мировой лидер в программном обеспечении по распознаванию человеческой речи. Программа дает большие возможности при использовании компьютера. Пользователь может диктовать

тексты в микрофон, и программа будет писать их сама, например, в текстовом процессоре.

Программные решения синтеза русской и английской речи, а также программные комплексы распознавания английской речи предлагаются следующими компаниями:

Sakrament TTS (Text-to-Speech) Engine – система нового поколения, осуществляющая качественный речевой синтез. Она может использоваться как отдельное приложение для озвучивания электронных текстов, в качестве речевого движка для других приложений, а также для интеграции с различными информационными системами.

Sakrament ASR Engine – разработка компании «Сакрамент», осуществляющая высокоточное распознавание речи на различных платформах. Технология распознавания речи используется при создании средств речевого управления – программ, управляющих действиями компьютера или другого электронного устройства с помощью голосовых команд, а также при организации телефонных справочных и информационных служб. Программа рассчитана на применение в различных аппаратных системах и программных приложениях, использующих технологии распознавания речи, таких как: *IVR*-системы, мобильные электронные устройства, бытовая техника и т.д. *Sakrament ASR Engine* может быть легко перенесена на любую существующую программную или аппаратную платформу, а также настроена под конфигурацию любого приложения.

Задание: Ознакомьтесь с программами распознавания и синтеза звучащей речи.

Тема 5: ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ И БАЗЫ ДАННЫХ

1. Причины появления информационно-поисковых систем
2. Виды ИПС
3. Основные понятия ИПС
4. Лингвистический компонент ИПС
5. Базы данных: основные понятия; способы организации; системы управления; способы доступа к информации

§ 5.1. Причины появления информационно-поисковых систем (ИПС)

Поиск информации является часто возникающей задачей, причем, если раньше потребность в отборе конкретной информации из всего информационного множества возникала в основном у специалистов, то с появлением Интернета проблему поиска и отбора необходимой информации приходится решать и рядовым пользователям. Помимо самой информационной потребности при больших объемах информации появляется необходимость в систематизации этой информации и облегчению ее поиска. Эту проблему призваны решать информационно-поисковые системы.

§ 5.2. Виды информационно-поисковых систем

Принято различать ручные, механизированные и автоматизированные ИПС. В качестве примера системы с *ручным* поиском можно привести ситуацию поиска литературы по определенной теме в библиотеке, когда необходимо сначала обратиться к тематическому каталогу, а затем читать полностью или выборочно отобранную литературу. Первые *механизированные* информационно-поисковые устройства представляли собой технические средства, которые обеспечивали отбор нужных документов путем механического сопоставления поисковых образцов документов с запросами. В этих устройствах применялись различного рода перфокарты. С использованием компьютеров для поиска информации стали говорить о создании *автоматизированных* ИПС. Таким образом, современные ИПС – это программные системы для хранения, поиска и выдачи интересующей пользователя информации.

Кроме того, по характеру поискового массива и выдаваемой информации ИПС подразделяют на документальные и фактографические.

Документальная ИПС предназначена для поиска документов (статей, книг, отчетов, описаний к авторским свидетельствам и патентам), содержащих необходимую информацию. Поисковый массив такой ИПС состоит из поисковых образов документов или из самих документов. В ответ на предъявляемый информационный запрос ИПС выдает некоторое множество документов (или адреса их хранения), содержащих необходимую пользователю информацию.

Фактографическая ИПС обеспечивает выдачу непосредственно фактических сведений, затребованных потребителем в информационном запросе. Поисковый массив состоит из фактографических записей, т.е. из

описаний фактов, извлеченных из документов и представленных на некотором формальном языке.

§ 5.3. Основные понятия ИПС

В теории и практике создания ИПС разработана своя терминология. Ее важнейшими понятиями являются следующие:

Запрос – вербально выраженная потребность пользователя в определенной информации.

Документ – любой осмысленный текст, который обладает определенной логической завершенностью и содержит сведения о его источнике и создателе. Документы хранятся в базе данных ИПС.

Информационно-поисковый язык – формальный язык, предназначенный для описания отдельных аспектов содержания документов, а также формулировки запроса для ИПС.

Дескрипторы – лексические единицы информационно-поискового языка. Дескриптор ставится в однозначное соответствие группе ключевых слов естественного языка, отобранных из текста определенной предметной области. Например, в качестве дескриптора может быть выбрано любое (предпочтительно наиболее часто используемое или короткое) ключевое слово или словосочетание или же цифровой код. Многозначному слову естественного языка соответствует несколько дескрипторов, а нескольким синонимичным словам и выражениям – один дескриптор.

Тезаурус – специально организованный нормативный словарь лексических единиц информационно-поискового и естественного языка. Назначение тезауруса – помочь пользователю сформулировать информационный запрос так, чтобы он был правильно понят системой. Тезаурус учитывает семантические связи между словами: антонимию, синонимию, родовидовые отношения, ассоциации.

Индексирование – выражение центральной темы какого-либо текста или описание какого-либо объекта на информационно-поисковом языке.

Поисковый образ документа – определенный информационный код, который в результате индексирования присваивается каждому документу, входящему в базу данных конкретной ИПС.

Поисковое предписание – текст на информационно-поисковом языке, содержащий признаки документов, затребованных пользователем в запросе.

Формальная релевантность – соответствие поискового образа документа поисковому предписанию.

Смысловая релевантность – действительное соответствие содержания выданного документа содержанию запроса

Точность поиска – отношение между количеством выданных релевантных текстов к общему количеству выданных системой текстов.

Полнота поиска – соотношение между количеством выданных релевантных текстов или документов к общему количеству релевантных документов,

имеющихся в данной информационной системе. В идеальном случае количественное выражение полного и точного поиска равно единице.

§ 5.4. Лингвистический компонент ИПС

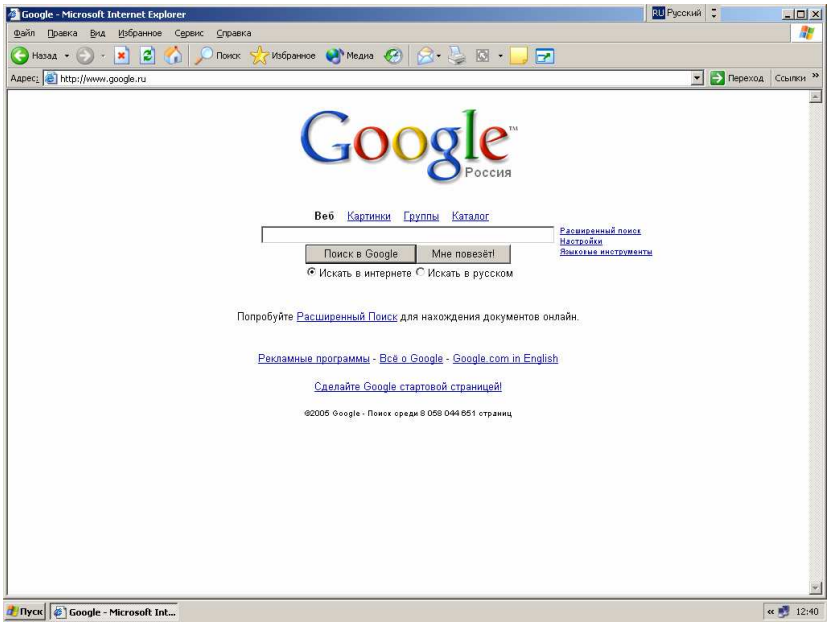
При работе ИПС могут возникать ошибки двух типов. Первый тип связан с ситуацией, когда текст является релевантным по смыслу, но не является релевантным с формальной точки зрения. В результате информационно-поисковая система не выдает этот текст пользователю. Второй тип ошибок связан с тем, что текст обладает формальной релевантность, не обладая при этом смысловой. В результате возникает так называемый информационный шум, когда пользователь на выходе получает множество текстов, не являющихся релевантными по смыслу.

Увеличить эффективность работы ИПС можно за счет детальной обработки текста документа. Существуют системы, которые для простоты в качестве поискового образа документа принимают его название, однако оно в силу разных обстоятельств не всегда формально отражает содержание текста. Поэтому применяют программы, производящие лингвистическую обработку текстов на естественном языке с учетом морфологии и синтаксиса. Только с их помощью можно установить, являются ли слова с похожим написанием формами одного слова или же это совершенно разные слова, в соответствие которым поставлены разные семантические единицы.

§ 5.5. Поисковые системы

Одним из достоинств системы Интернет считается наличие в ней нескольких поисковых систем, или машин.

Всемирная поисковая система *Google (http://www.google.com)* отличается великолепными результатами поиска, отличающимися высокой степенью релевантности. Место сайта в списке напрямую связано с количеством ссылок на него с других серверов аналогичной тематики.

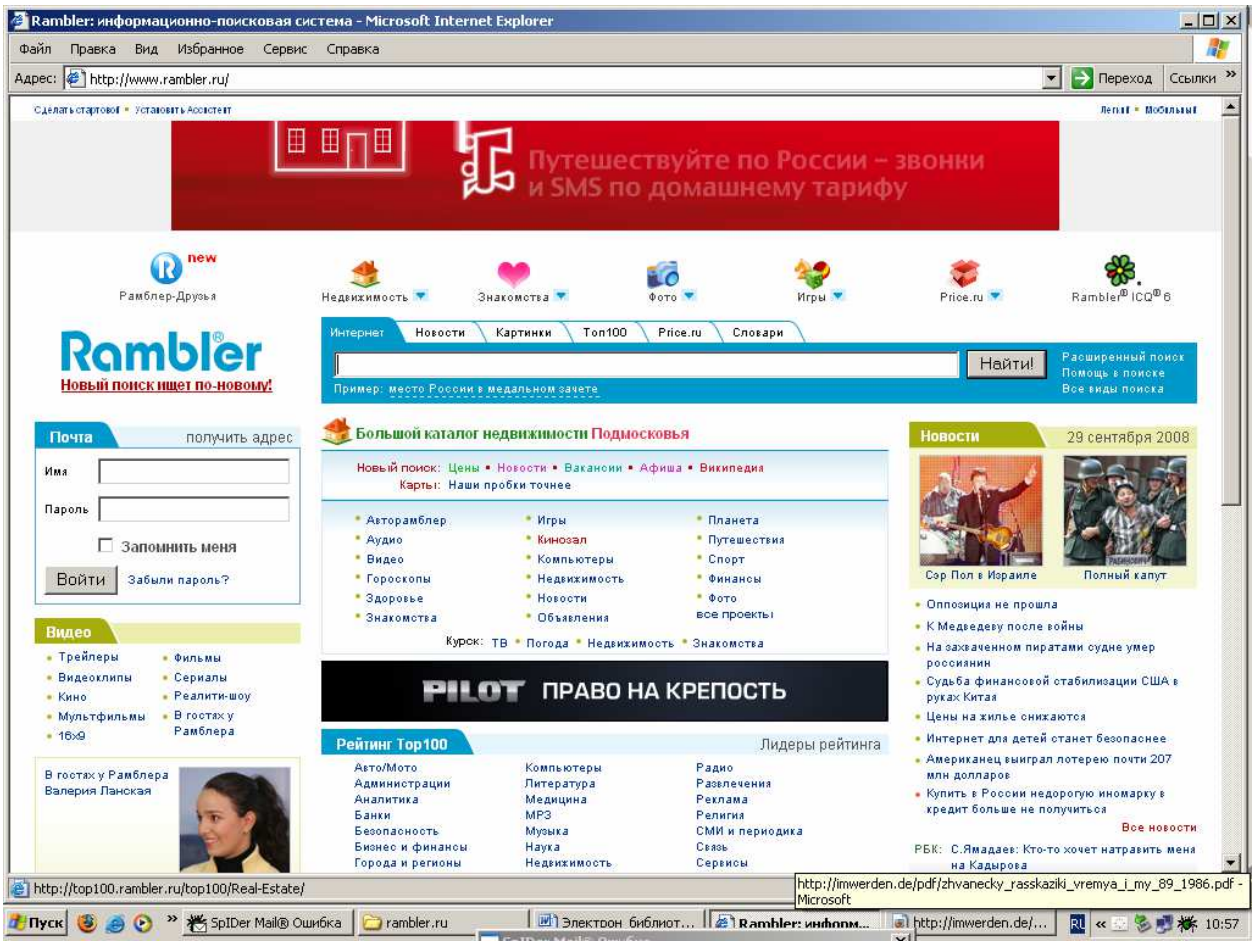


Компания Google была создана в 1998 г. Название, которое она получила, представляет собой слегка искаженный математический термин *googol*, обозначающий число из единицы и ста нулей. За прошедшие годы Google стала популярнейшим поисковым механизмом в Сети, которым регулярно пользуются десятки миллионов человек. Кроме того, Google предлагает своим поклонникам множество сервисов, которые также высоко ценят многие пользователи.

Русскоязычные поисковые машины

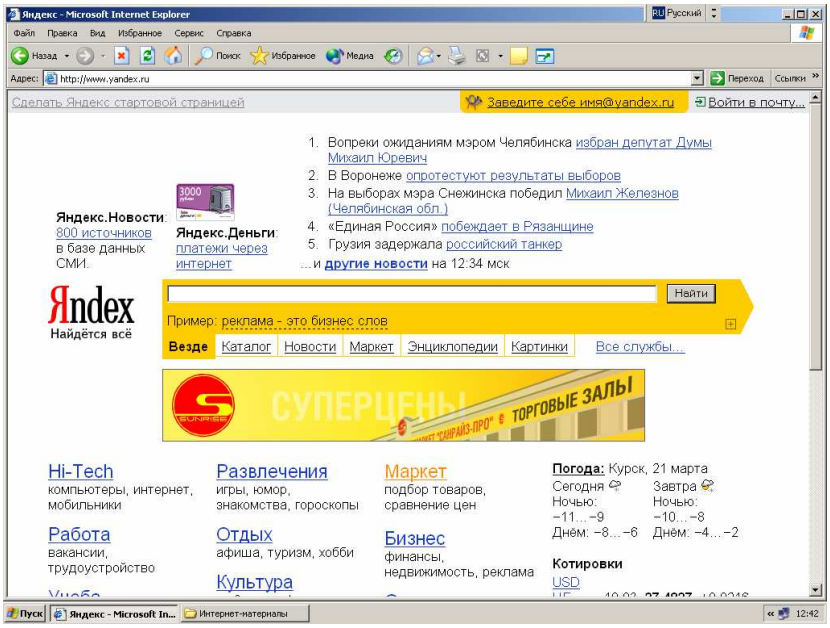
«Рамблер» (http://rambler.ru) – первая поисковая система русскоязычного сегмента сети Интернет, созданная в 1996 году разработчиками из подмосковного Пущино.

Это открытый портал, уникальное сочетание поисковой системы, соединяющей в себе традиционный горизонтальный поиск по Интернету со специализированными вертикалями (Цены, Новости, Вакансии, Афиша, Википедия, Авто и др.), постоянно обновляемой ленты новостей и огромного выбора сервисов – аудио и видео, фото и открытки, знакомства и объявления, покупки и финансы, путешествия и спорт, а также возможности для открытого общения – почта и ICQ. Многозначительно имя поисковика – *rambler* ‘скиталец, странник, бродяга’.



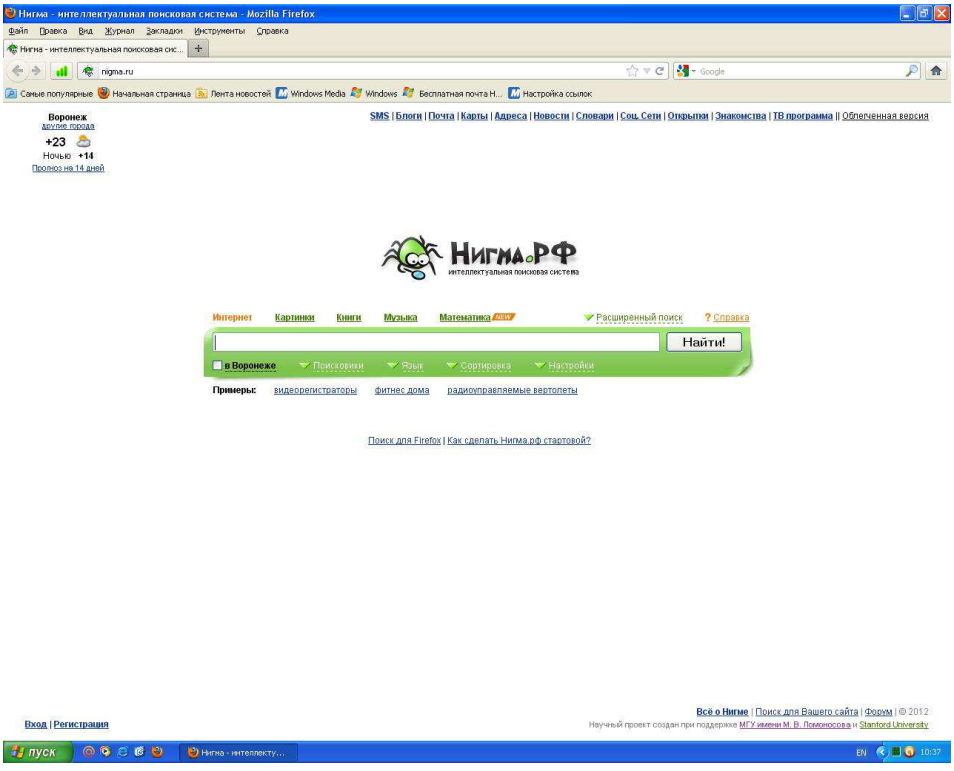
Яndex (http://www.yandex.ru)

Основным достоинством Яndex'а является способность находить заданные слова независимо от формы, в которой они употребляются в документах.

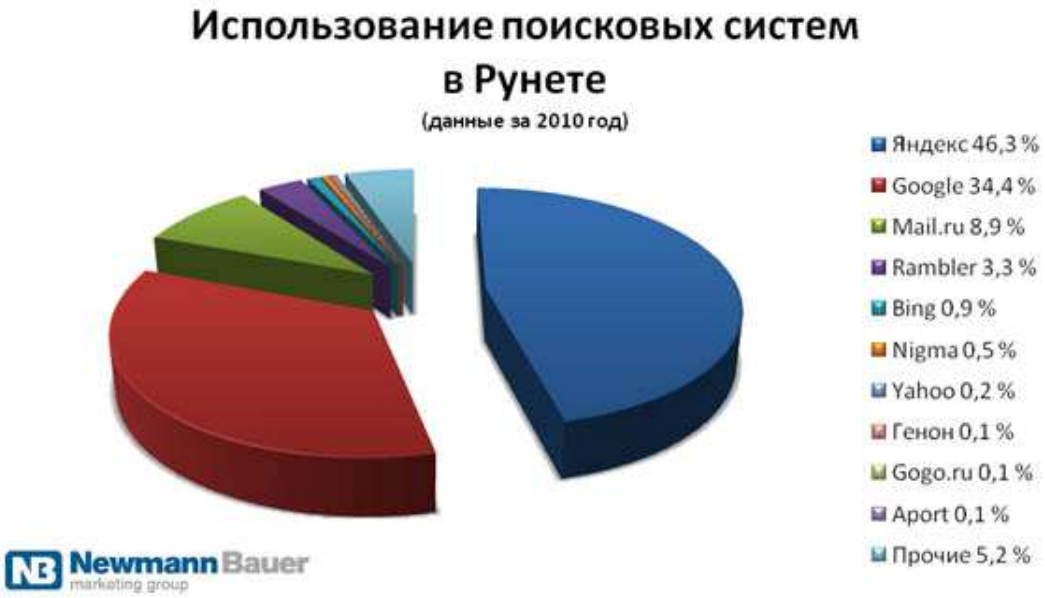


Нигма (http://nigma.ru)

Нигма – российская интеллектуальная метапоисковая система, первая кластеризующая поисковая система в Рунете. Проект создан при поддержке факультетов ВМК и психологии МГУ, а также Стэнфордского университета. Название «Nigma» (один из родов пауков семейства Dictynidae) было выбрано по ассоциации со Всемирной паутиной.



Использование ИПС в Рунете:



§ 5.6. Базы данных: основные понятия; способы организации; системы управления; способы доступа к информации

База данных – это совокупность определенным образом упорядоченных сведений о некоторых объектах (например, список студентов университета с указание Ф.И.О., пола, года рождения, факультета, группы, № зачетки, адреса, телефона, размера стипендии и т.д.).

В реальном мире каждый объект обладает определенными свойствами (вес, длина, цвет и т.д.). Эти свойства называются данными. Группу данных, образующих в таблице одну строку, называют записью об объекте. Набор записей одного столбца называют полем или доменом (например, пол, дата рождения и т.д.).

Различают следующие типы данных, используемых в БД: 1) текстовые, 2) числовые, 3) дата/время, 4) денежные (для хранения денежных сумм), 5) счетчики (для хранения автоматически наращиваемых чисел), 6) логические (да/нет; true/false) и некоторые другие.

Существует несколько способов организации БД: 1) иерархический, 2) сетевой, 3) реляционный.

Иерархическая модель представляет взаимосвязь данных в виде иерархического дерева, состоящего из узлов. На самом верхнем уровне иерархии имеется только один узел – корень. Каждый узел, кроме корня, связан с одним из узлов на более высоком уровне, называемым исходным узлом для данного узла. Ни один элемент иерархической модели не имеет более одного исходного. Каждый элемент может быть связан с одним или несколькими элементами на более низком уровне. Они называются порожденными. Элементы, расположенные в самом конце ветви и не имеющие порожденных, называются листьями.

В *сетевой* модели любое данное может быть связано с любым другим данным. Однако в отличие от иерархической модели у порожденного узла может быть несколько исходных узлов.

Данных в *реляционной* базе представляются в виде таблицы. Каждая таблица выражает отношения (relations) между включенными в нее данными (например, в домене «пол» выделяются как студенты женского, так и мужского пола).

Система управления базами данных (СУБД) – это совокупность программных средств, позволяющих осуществлять ведение баз данных (создание, обновление, удаление элементов баз данных) и поиск в них информации.

Доступ к БД осуществляется или через специальный программный интерфейс *API (Application Programming Interface)*, или через универсальные механизмы доступа (драйверы и провайдеры).

В практической работе с БД применяются три основных режима выдачи информации: 1) *on-line*, 2) *off-line*, 3) *ИРИ* (избирательное распространение информации.

В режиме on-line вся информация выдается мгновенно на экран ПК, в режиме off-line – на экран выдаются только количественные данные о результатах поиска (н-р, сколько книг по грамматике английского языка найдено в базе данных, но сами названия книг не выдаются; в режиме ИРИ запрос пользователя помещается в специальных каталог базы данных и автоматически обрабатывается при каждом обновлении информации, при этом информация передается пользователю в режиме off-line.

Задание: Используйте ИПС для сбора данных по следующим темам:

1) *Учебные заведения, где осуществляют подготовку по выбранной Вами специальности в России и за рубежом.*
2) *О грантах, программах обмена и стажировках по выбранной Вами специальности/ тематике в России и за рубежом.*
3) *Материалы о жизни и творчестве выдающихся общественных деятелей, ученых, художников, композиторов, музыкантов и т.д. прошлого и современности.*
4) *Необходимую информацию о стране, в которой Вы мечтаете побывать (оформление документов, транспорт, бронирование отелей, достопримечательности, особенности национальной кухни и т.п.).*
5) *О Вашем хобби.*
6) *О той профессии, которой бы Вы хотели овладеть и месте, где бы Вы хотели работать.*
7) *Назовите известные Вам ИПС. Какими Вы пользуетесь и почему? В чем их достоинства по сравнению с другими ИПС?*

Тема 6: ЛИНГВИСТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ РЕСУРСЫ
1. Проблемы создания лингвистических информационных ресурсов
2. Электронные библиотеки
3. Проект ‘Linguist List’

§ 6.1. Проблемы создания лингвистических информационных ресурсов

Лингвистические информационные ресурсы – это множество определенным образом организованных речевых и языковых данных, находящихся на машинных носителях информации и используемых в различных сферах практической деятельности (образовании, промышленности, экономике, культуре, искусстве и т.д.).

Выбор этого термина выражает идею о том, что большие массивы лингвистических данных, используемых для создания и развития эффективных систем обработки текста и речи, играют такую же существенную, фундаментальную роль, как и железные дороги, электросети, средства коммуникации для промышленного и экономического развития страны.

Выделяют активные и пассивные лингвистические ресурсы. К *пассивным* формам относят письменный лексикон, терминологические словари, письменные текстовые массивы (корпуса текстов), фонетические ресурсы, электронные библиотеки и т.д., к *активным* – алгоритмы, модели, программы, базы знаний.

Проблемам создания лингвистических ресурсов ежегодно посвящается большое количество научных конференций во всем мире. Создан ряд организаций, занимающихся разработкой лингвистических ресурсов: *LDC (Linguistic Data Consortium, USA), ELRA (European Language Resources Association), TELRI (TransEuropean Language Resources Infrastructure)*. Перед ними стоят следующие задачи:

1) разработка единых стандартов создания ресурсов,
2) разработка способов защиты от несанкционированного доступа,
3) создание единых экспертных требований,
4) планирование единой стратегии разработки лингвистических ресурсов,
5) создание многофункциональных лингвистических ресурсов большого объема для использования в разных странах.

§ 6.2. Электронные библиотеки

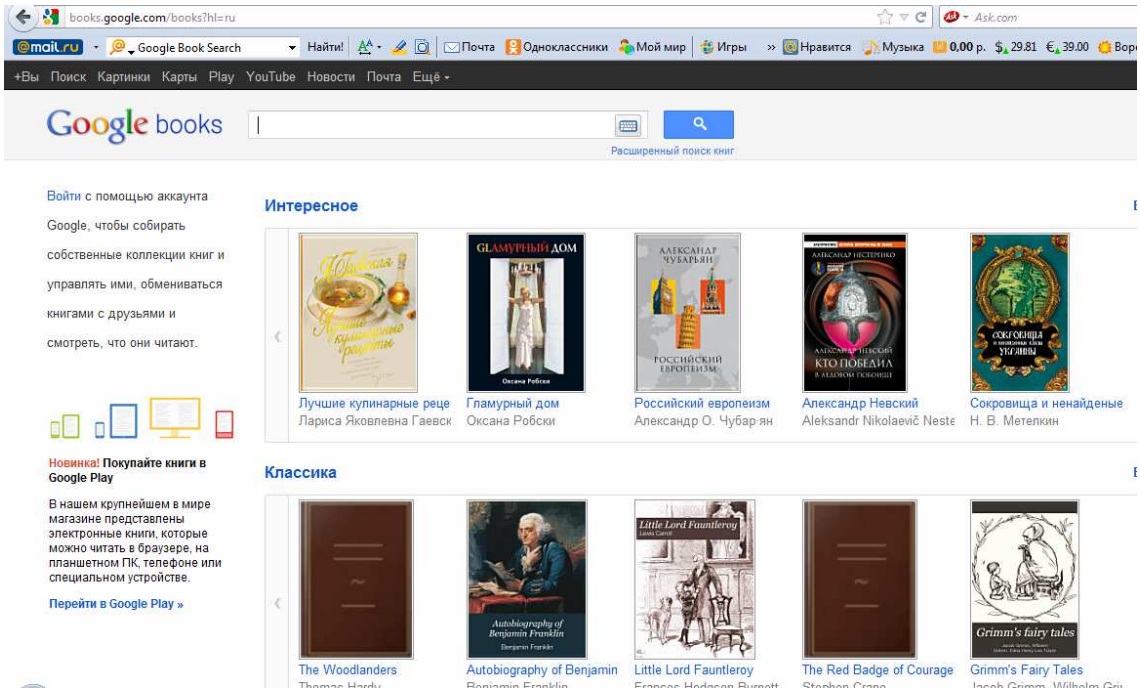
Электронная библиотека – упорядоченная коллекция разнородных электронных документов (в том числе книг), снабженных средствами навигации и поиска. Электронные библиотеки могут быть универсальными, стремящимися к наиболее широкому выбору материала (как Библиотека Максима Мошкова или Либрусек), и более специализированными, как Фундаментальная электронная библиотека или проект Сетевая Словесность,

нацеленный на собирание авторов и типов текста, наиболее ярко заявляющих о себе именно в Интернете.

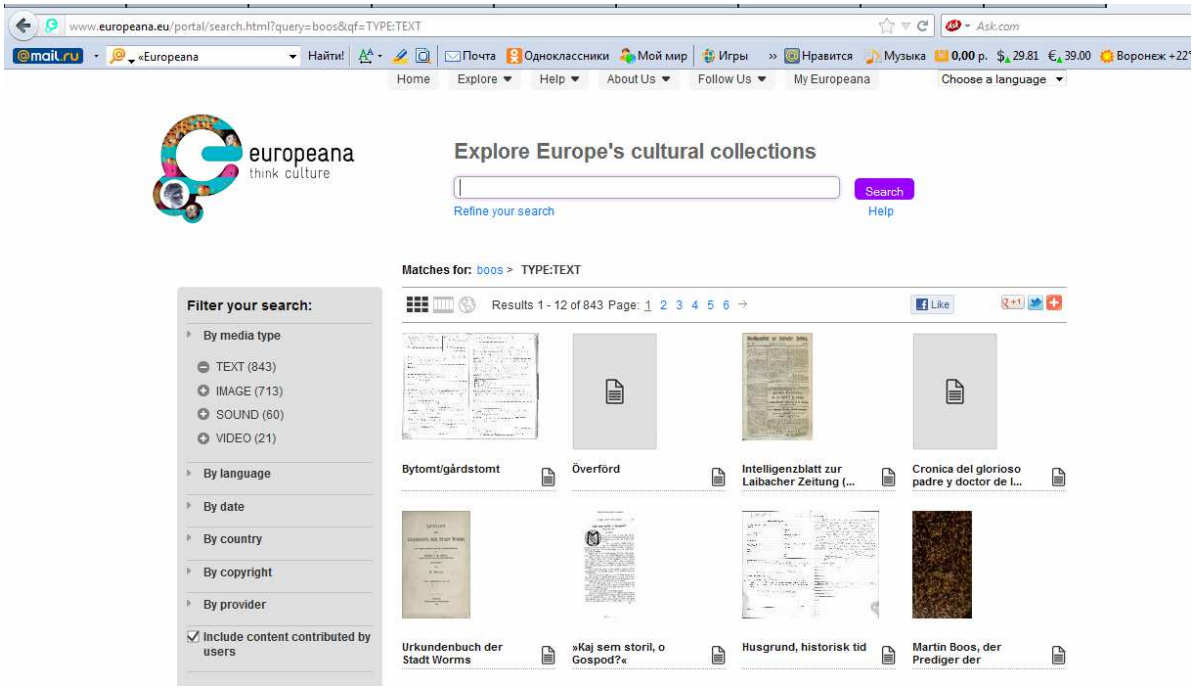
Первым проектом по созданию ЭБ стал Проект «Гутенберг» (1971 г). В Рунете первой ЭБ стала библиотека Максима Мошкова. С ростом числа пользователей компьютеров и Интернета всё большее количество людей начинает пользоваться электронными книгами. В связи с этим многие библиотеки начали создавать электронные версии хранящихся в их фондах книг. Особое место в ряду электронных библиотек занимают библиотеки научно-образовательной тематики, в которых собраны издания, необходимые для осуществления образовательного процесса. Среди электронных библиотек можно выделить следующие:

- *Универсальные международные*

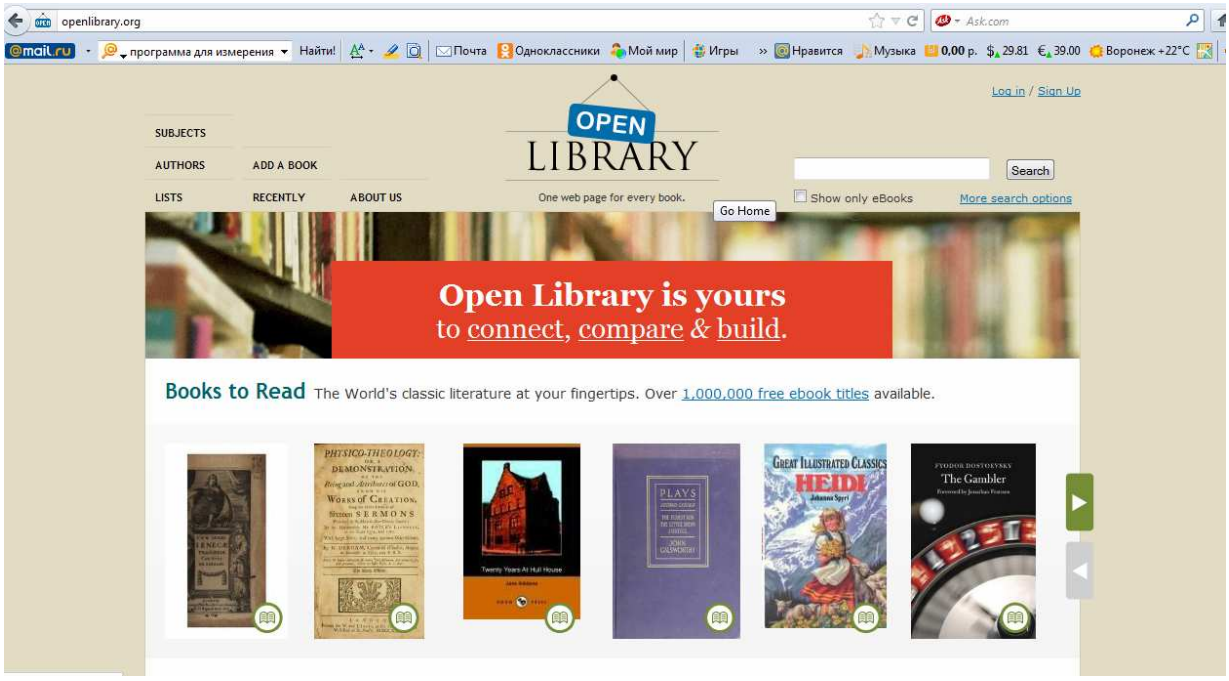
1) *Google Book Search (Google Books)* – проект компании *Google*, содержит значительное количество полных текстов книг, в том числе и на русском языке.



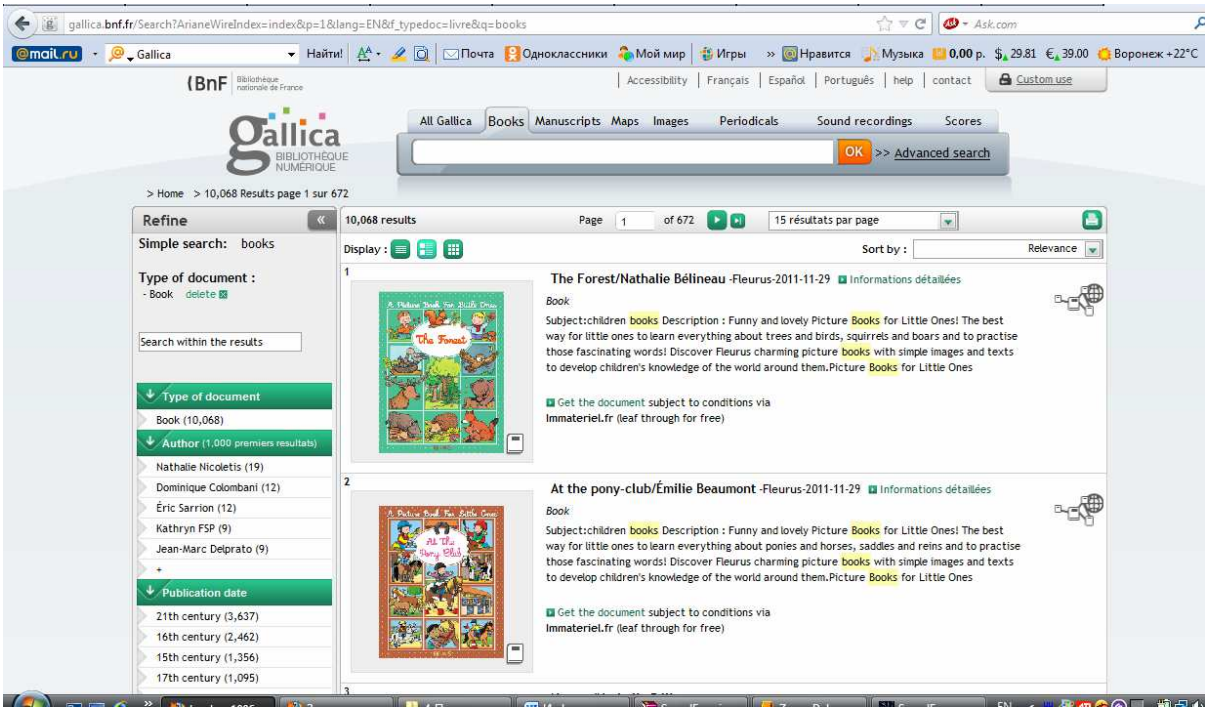
2) *Европейская электронная библиотека Europeana* – содержит оцифрованные объекты культурного наследия Европы: книги, картины, фотографии, аудиозаписи.



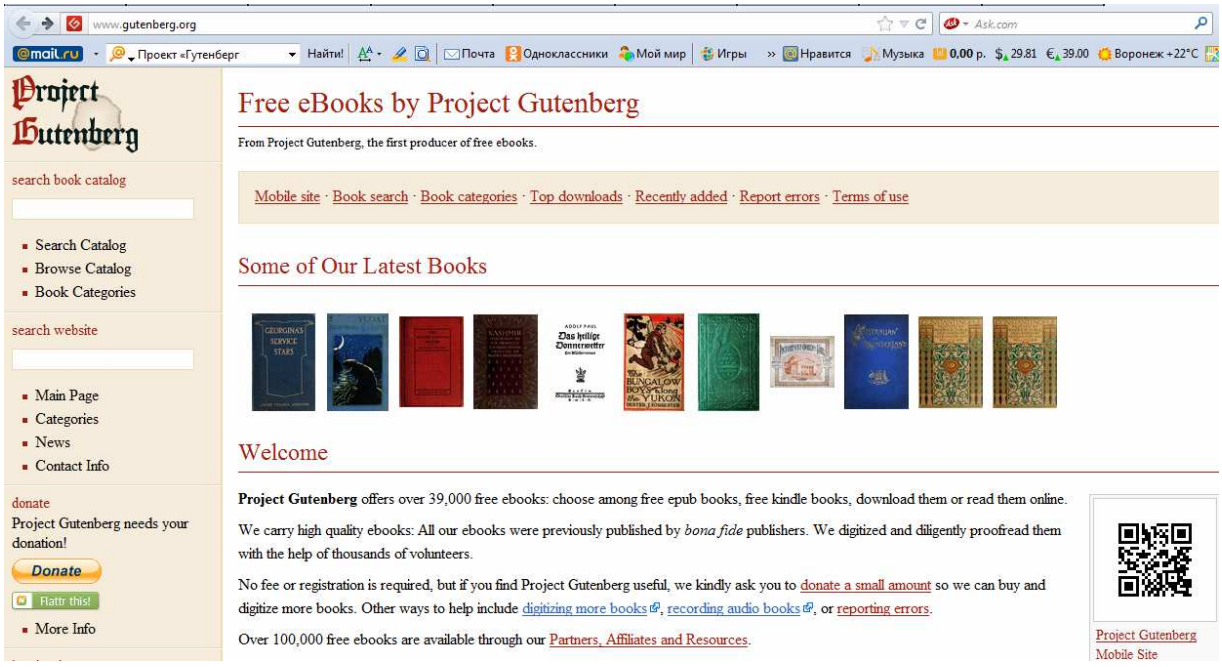
3) *openlibrary.org* – проект некоммерческой организации *Internet Archive* по оцифровке книг, находящихся в общественном достоянии; посетителям сайта из фондов Бостонской общественной библиотеки предлагается выбрать желаемую книгу для бесплатной оцифровки (*Scan-On-Demand*). Содержит большое количество книг на русском языке XIX–XX вв.



4) *Gallica* – французский архив, ставит своей целью оцифровать все содержимое Национальной библиотеки Франции.

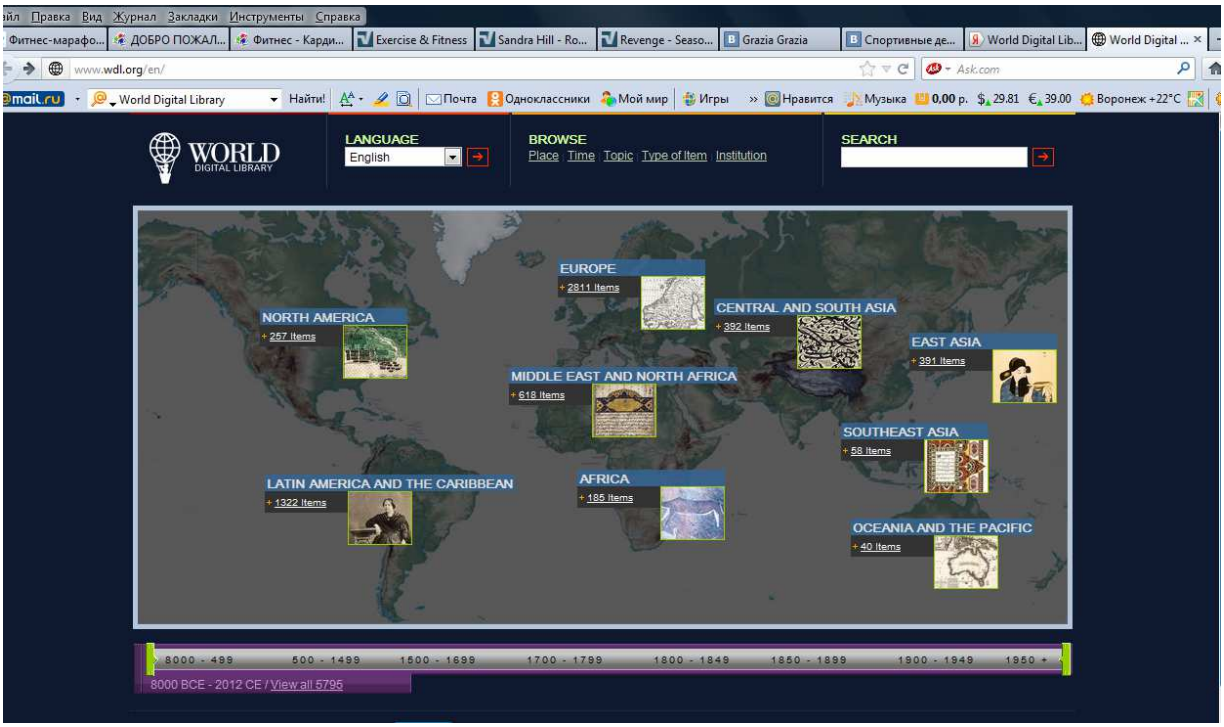


5) *Проект «Гутенберг»* – первая в мире электронная библиотека.



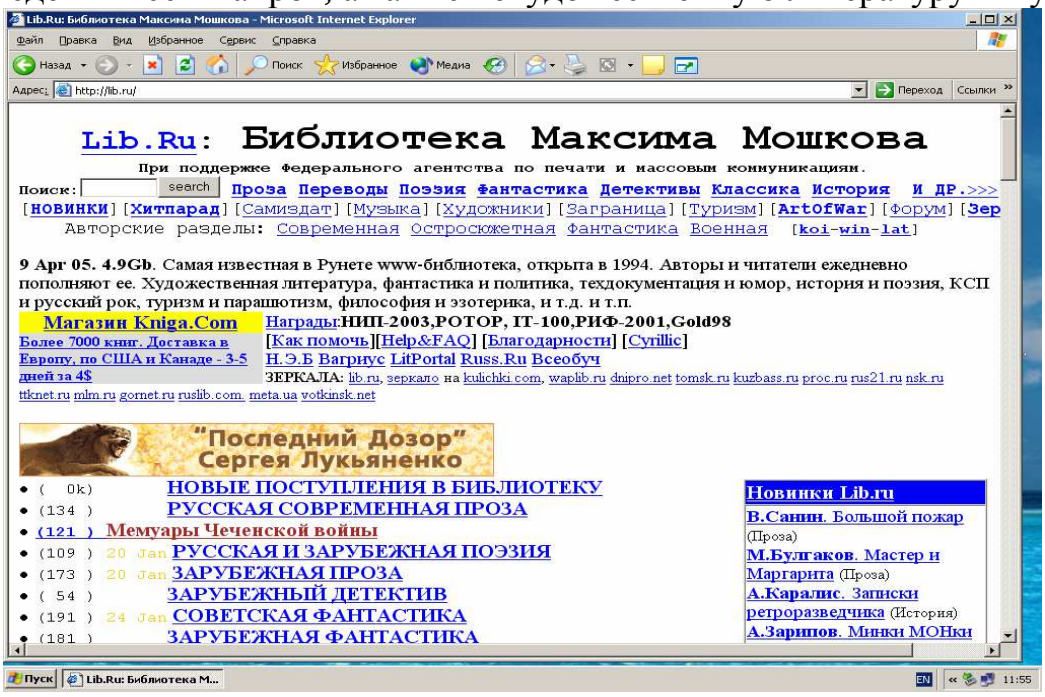
6) *LibriVox* – проект по созданию аудиокниг в общественном достоянии, на основе текстов, из проекта «Гутенберг».

7) *Всемирная цифровая библиотека (World Digital Library)* – проект Библиотеки Конгресса. В 2007 году к проекту присоединилась Российская национальная библиотека.



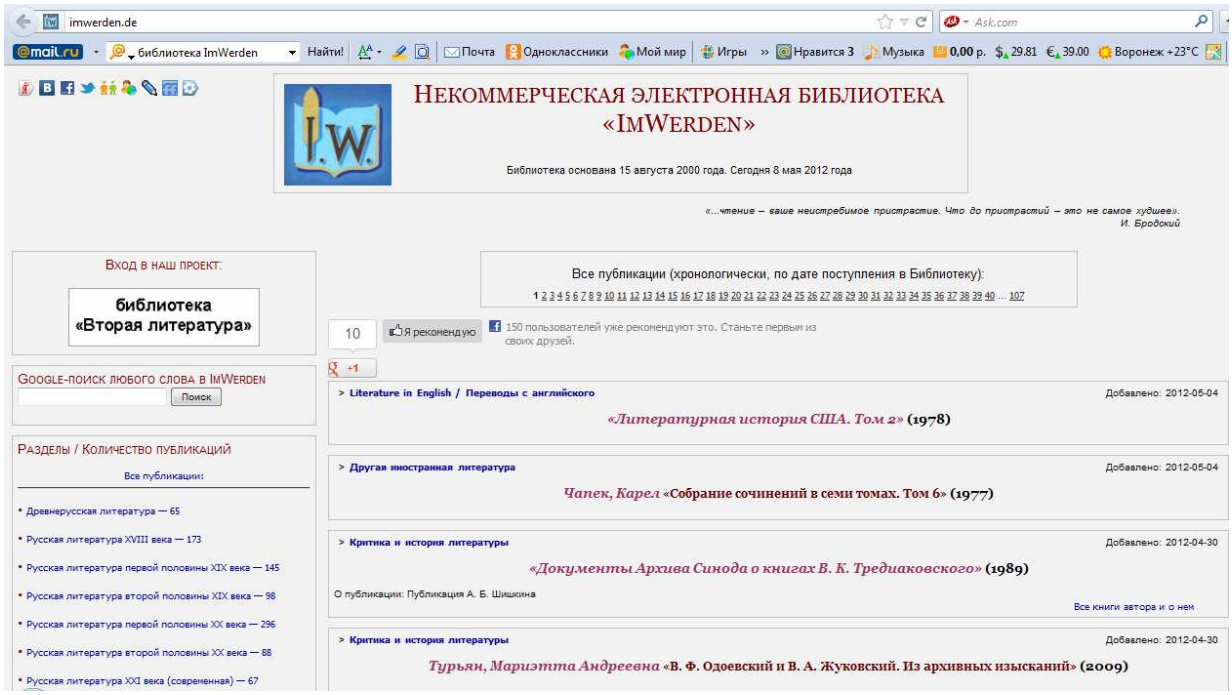
- *Русскоязычные*:

1) *библиотека Максима Мошкова* – старейшая, наиболее известная и одна из крупнейших русскоязычных сетевых библиотек. Содержит литературные произведения всех жанров, а также нехудожественную литературу и музыку.



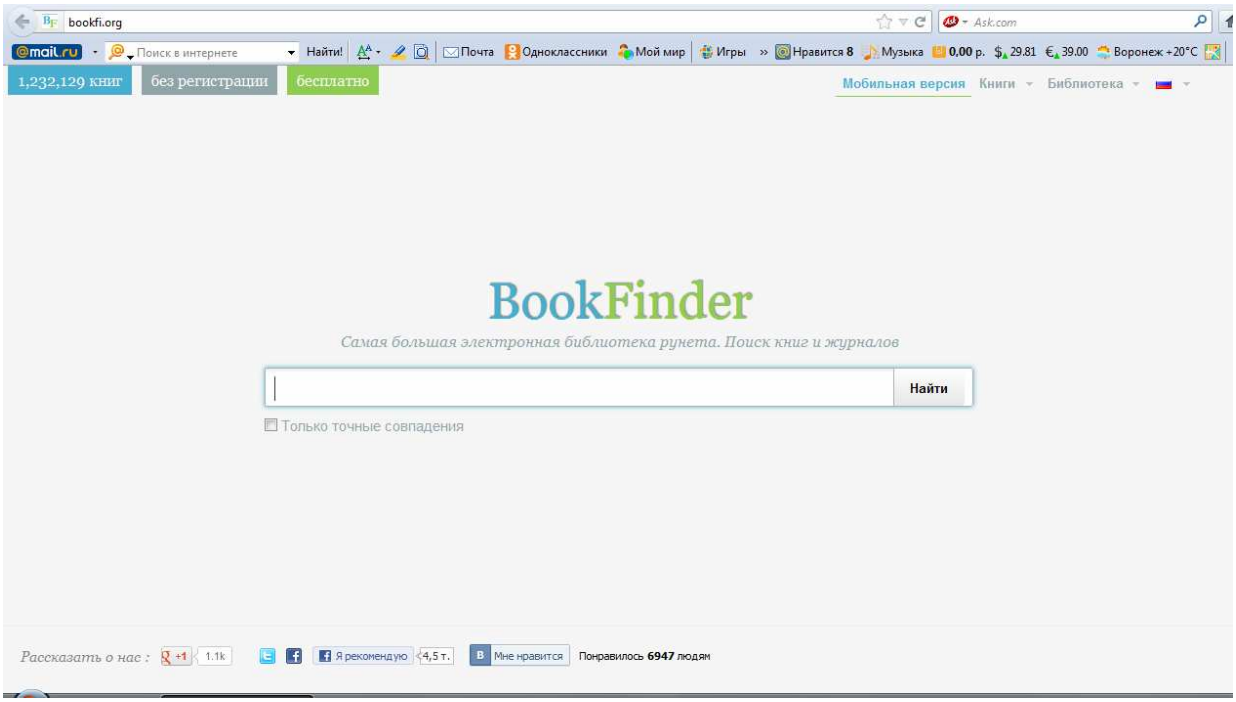
2) *библиотека «Альдебаран»*

3) *библиотека ImWerden*

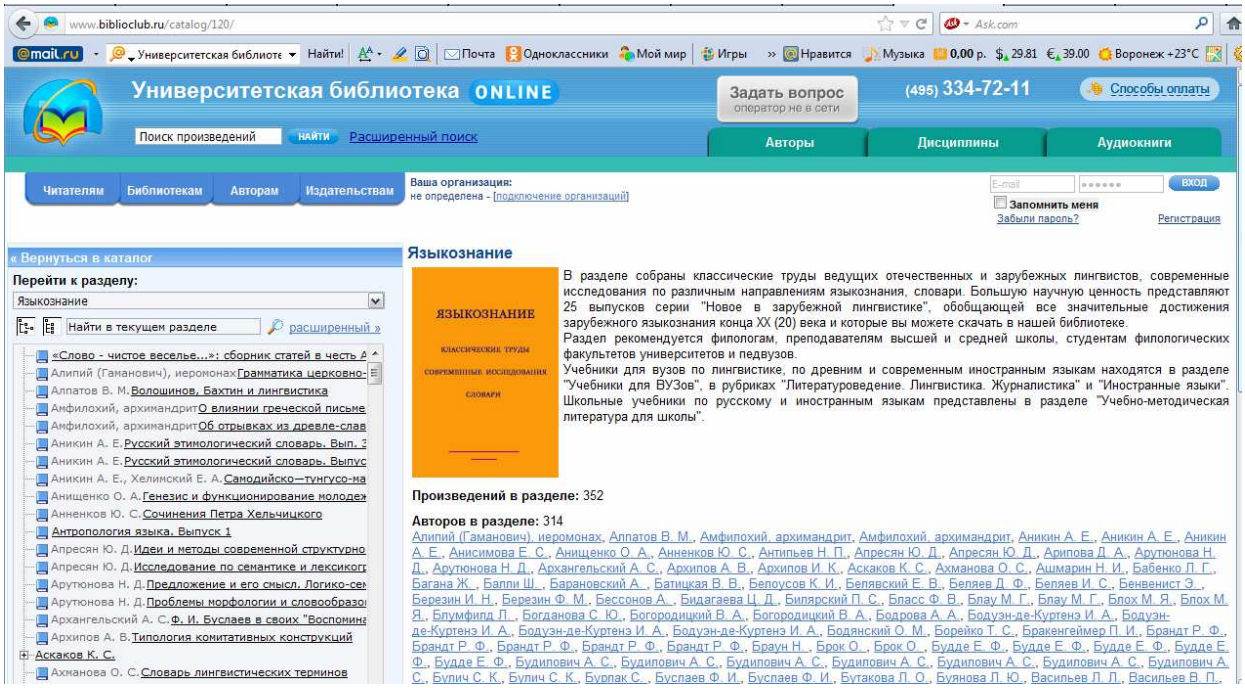


4) *Либрусек* – известная библиотека, позиционирующая себя как «сообщество пиратов».

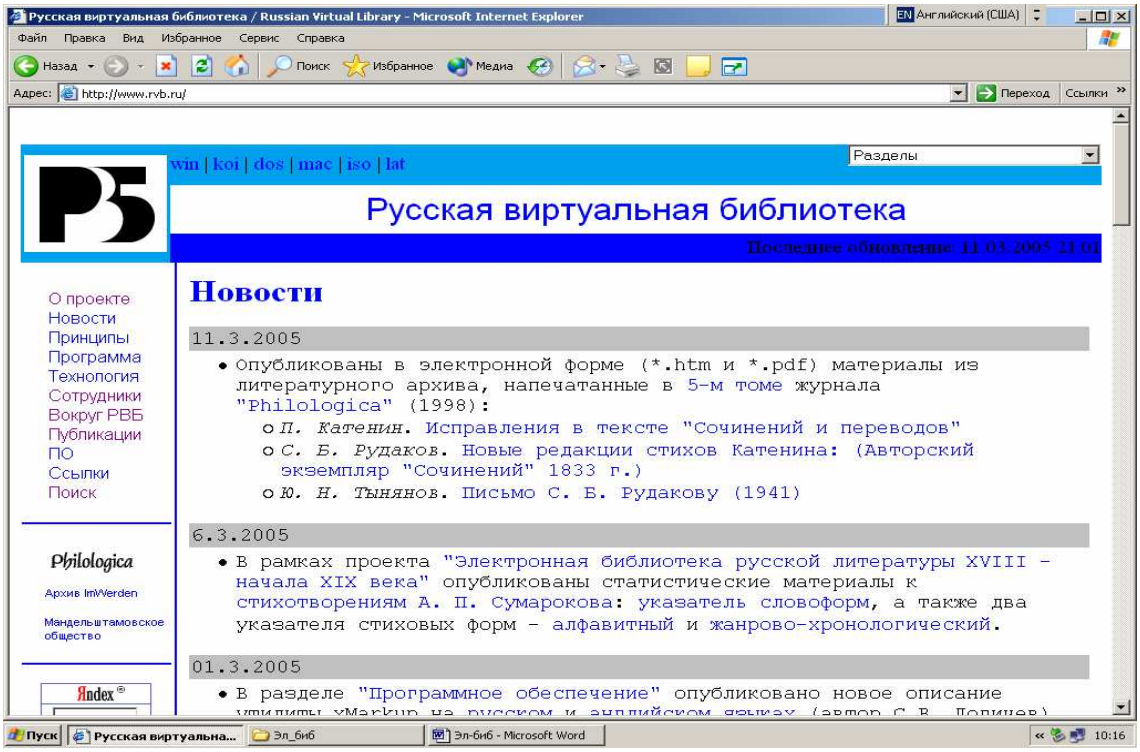
5) *BookFi.org* – относительно новый проект магистров СПбГУ, организована слиянием всех крупнейших художественных и научных коллекций. На май 2011 года содержит более 900 тысяч книг и журналов.



6) *Университетская библиотека онлайн* – крупнейшая легальная онлайн библиотека, основу, которой составляют учебные, образовательные и научные книги в электронном формате.



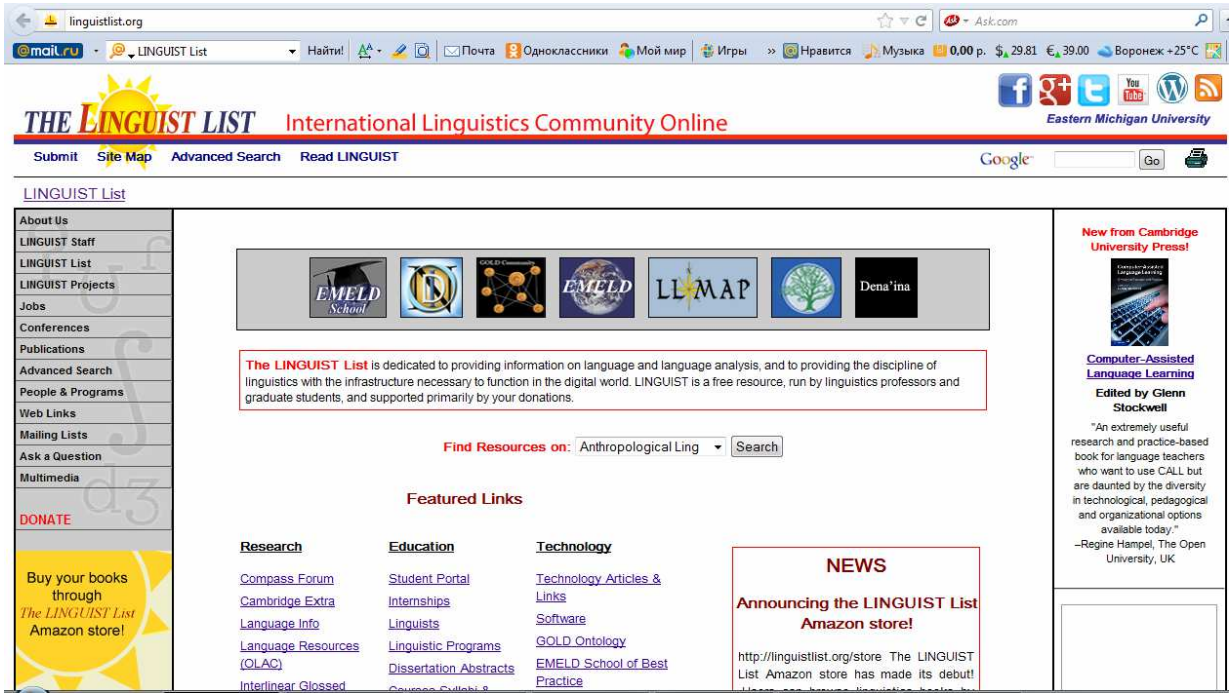
7) *Русская виртуальная библиотека (www. rvb. ru)*



Цель проекта – электронная публикация классических и современных произведений русской литературы по авторитетным источникам с приложением необходимого справочно-комментаторского аппарата. Программа публикаций РВБ предполагает максимально широкий охват художественных и литературно-критических произведений, созданных на русском языке с XVIII в. до наших дней. В корпус публикаций будут включены произведения древнерусской литературы и литературы XVII века.

§ 6.3. Проект ‘LINGUIST List’

LINGUIST List (ЛингвистЛист) – крупнейший Интернет-ресурс международного лингвистического сообщества. Основан в 1990 году Энтони Аристаром.



Сайт поддерживается на английском языке и содержит следующие разделы:
1) общая информация о сайте и его разработчиках;
2) информация об основной электронной рассылке сайта, а также архив рассылки;
3) база данных других лингвистических рассылок;
4) информация о вакансиях для лингвистов;
5) информация о конференциях;
6) информация о публикациях (книги, периодика, диссертации и пр.);
7) информация о лингвистических ресурсах (корпуса, словари) и инструментарии (шрифты, программы);
8) информация о лингвистах (более 23 тыс. персоналий) и организациях;
9) раздел «Задай вопрос лингвисту»;
10) обучающие аудио- и видеоматериалы.

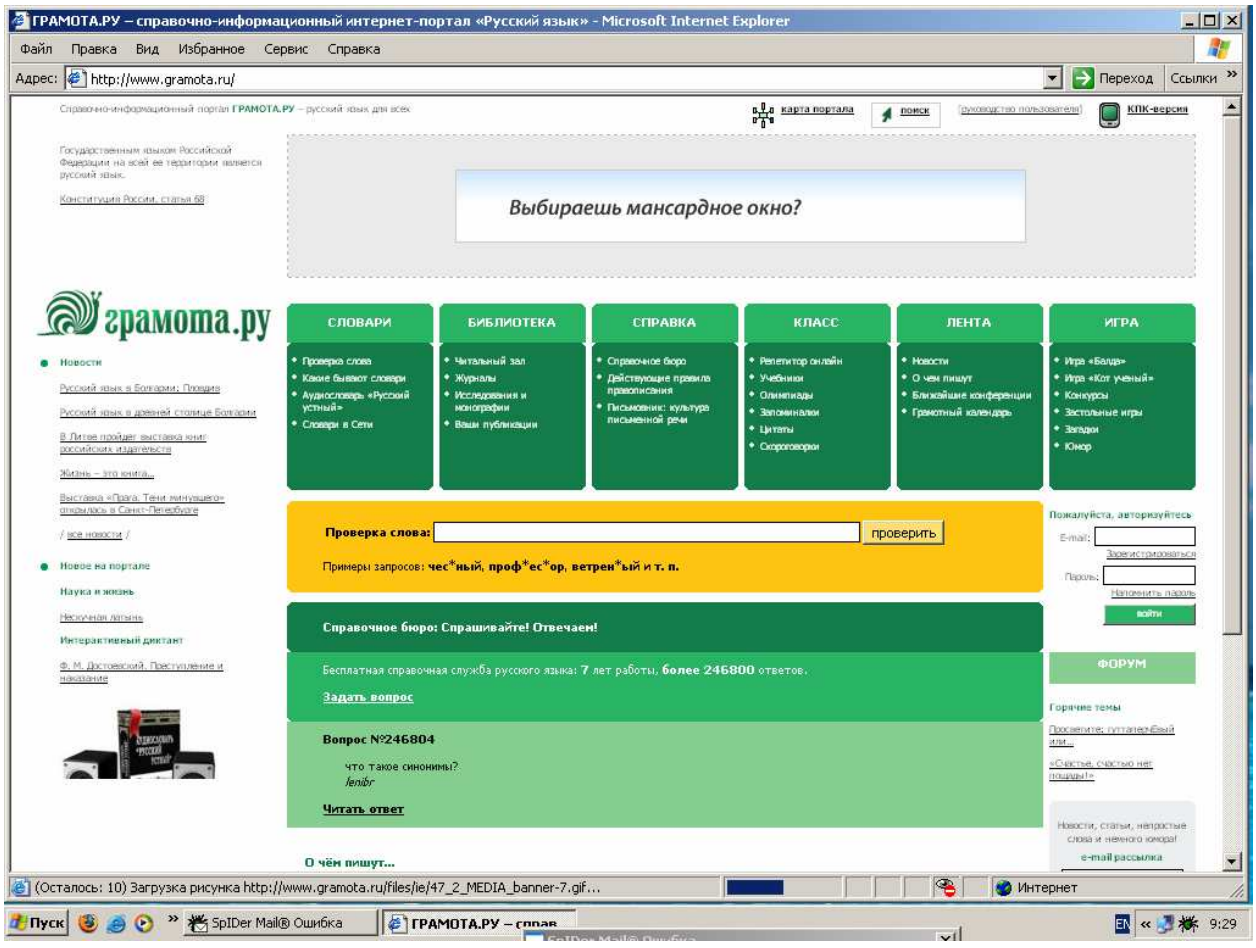
Работу по поддержанию сайта вели сотрудники Университета Восточного Мичигана и Университета Уэйна, однако в 2006 г. при Университете Восточного Мичигана для этой цели был основан специальный центр – Институт лингвистической информации и технологий. Работа над сайтом в разные годы поддерживалась грантами Национального научного фонда США *(National Science Foundation)*, а также частными пожертвованиями.

§ 6.4. Образовательные порталы

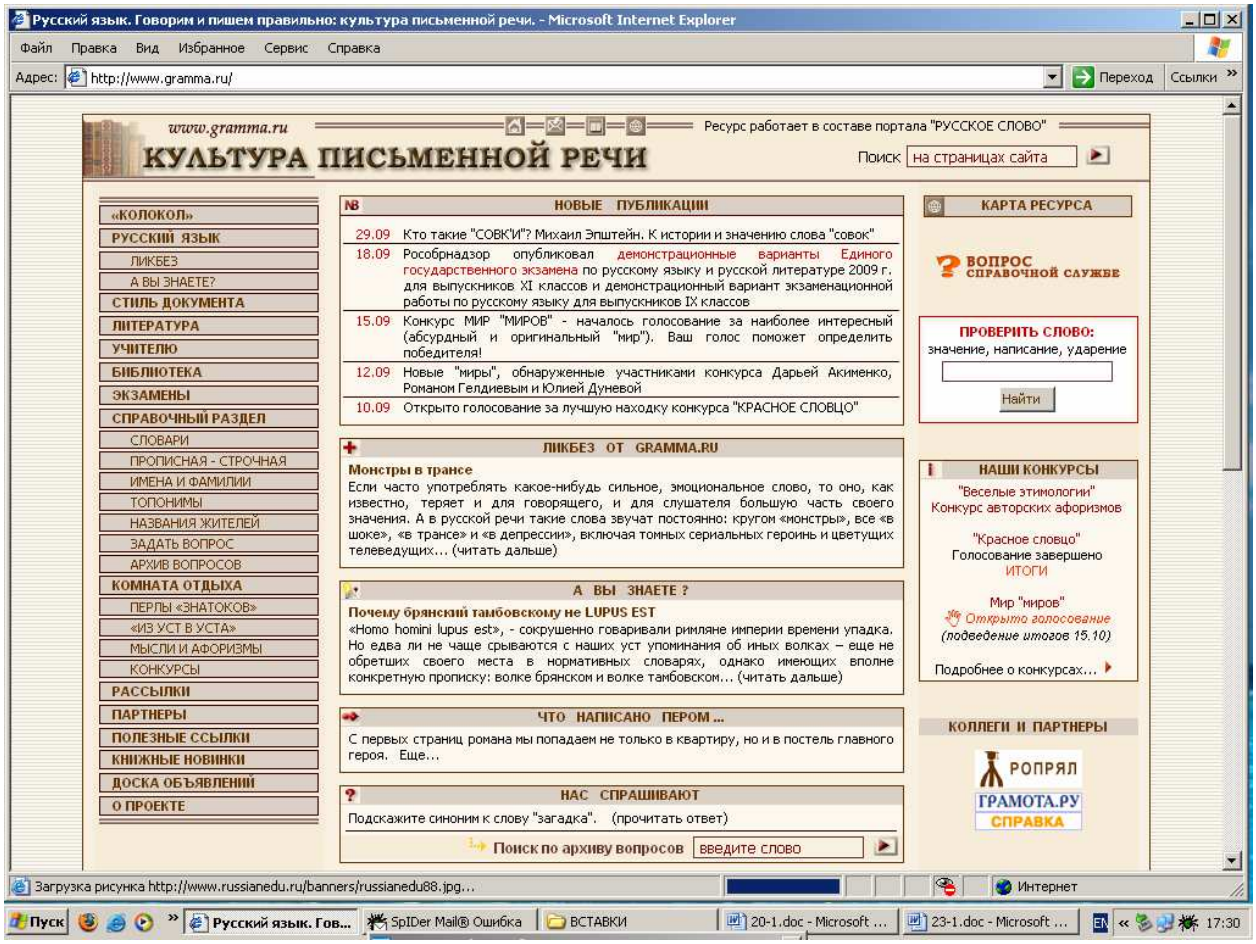
Существенную роль в самообразовании гуманитария могут играть образовательные порталы. Портал – это сайт Интернет, предоставляющий пользователю доступ сразу к нескольким сервисам Сети – справочнику, новостям, поисковой системе и т. д.

Портал ***«Русский язык»*** (gramota. ru) предлагает следующую информацию: «Орфографический словарь русского языка», «Словарь трудностей произношения и ударения в современном русском языке», «Новый толково-словообразовательный словарь», «Словарь ударений русского языка. Информация организована следующим образом: СЛОВАРЬ: Проверка слова; Какие бывают слова; Аудиословарь «русского устного»; Словарь в Сети. БИБЛИОТЕКА: Читальный зал; Журналы; Исследования и монографии; Ваши публикации. СПРАВКА: Справочное бюро; Действующие правила правописания; Письмовник: культура письменной речи. КЛАСС: Репетитор онлайн; Учебники; Олимпиады; Запоминалки; Цитаты; Скороговорки. ЛЕНТА: Новости; О чём пишут; Ближайшие конференции; Грамотный календарь. ИГРЫ.

Систематически проводимый интерактивный диктант – это прекрасная возможность поддерживать хорошую орфографическую и пунктуационную форму.



Портал ***«Культура письменной речи»*** (gramma. ru).

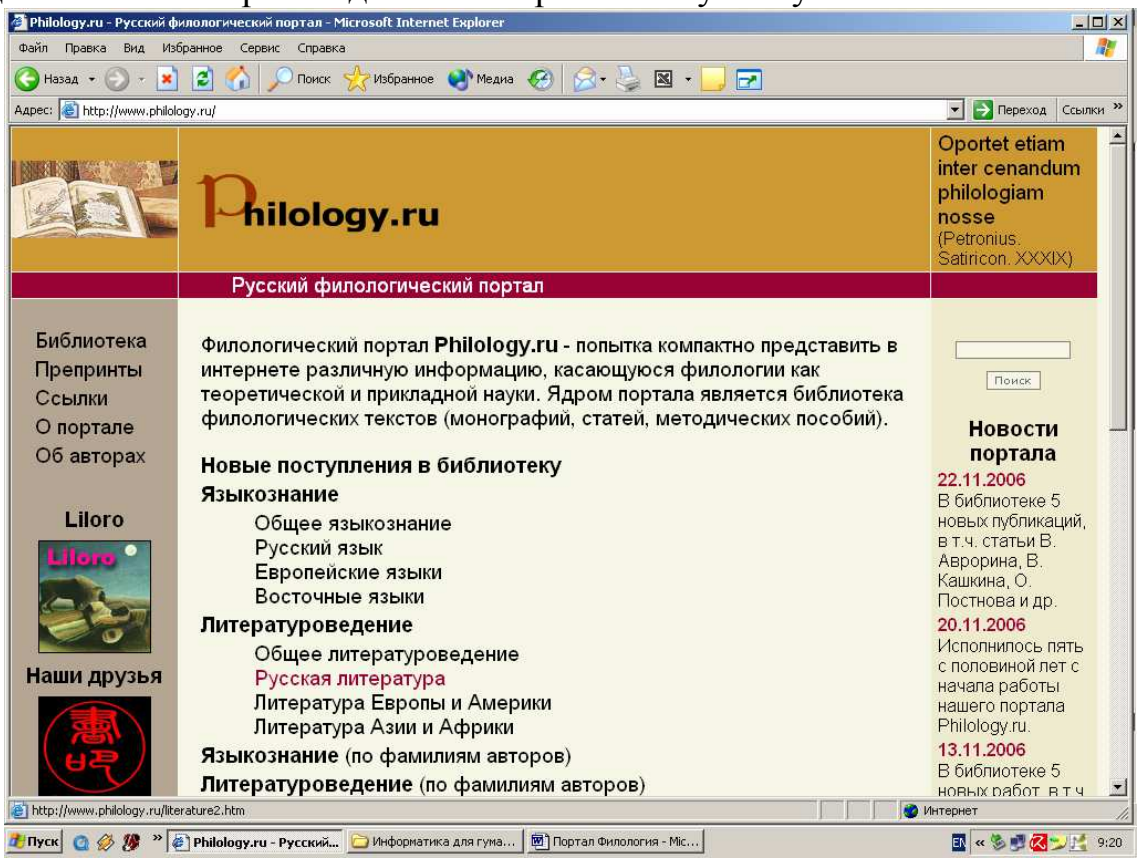


Темы раздела:

- Современный русский язык
- Морфология и словообразование
- Принципы орфографии
- Правила орфографии
- Синтаксис и пунктуация
- Лексикология
- Фразеология
- Стилистика
- Орфоэпия
- Типичные ошибки
- Экзамен
- Тесты, задания
- Ликбез от «Gramma. ru»
- А вы знаете…

Портал ***Philology.ru*** даёт информацию о важнейших русскоязычных филологических ресурсах. Публикуются только *научные работы на русском языке*, изданные ранее в виде книг, брошюр и статей; работы, не

издававшиеся ранее в бумажном виде, публикуются в разделе «Препринты»; художественные произведения на портале не публикуются.



Библиотека портала имеет следующую структуру:

Новые поступления в библиотеку

Языкознание

- Общее языкознание
- Русский язык
- Европейские языки
- Восточные языки

Литературоведение

- Общее литературоведение
- Русская литература
- Литература Европы и Америки
- Литература Азии и Африки

Языкознание (по фамилиям авторов)

Литературоведение (по фамилиям авторов)

Рецензии

Препринты

Распределение сайтов по рубрикам достаточно условно, на некоторых из них содержится информация не только по основной теме, но и по другим разделам. Рассмотрим сайты с лингвистической тематикой.

ЯЗЫКОЗНАНИЕ

Общелингвистические ресурсы

homepages.tversu.ru/~ips
Сайт И.П. Сусова (Тверской университет). Один из наиболее богатых русскоязычных ресурсов по общей лингвистике. Многочисленные ссылки на другие лингвистические сайты.

starling.rinet.ru
"The Tower Of Babel" – уникальный лингвистический веб-проект, созданный С.А. Старостиным – база данных по этимологиям некоторых языковых семей (алтайской, семитской, сино-тибетской и др.). Имеется библиотека работ по сравнительно-историческому языкознанию.

iling.nw.ru.org
Сайт Института лингвистических исследований РАН (Санкт-Петербург) – одного из ведущих научных заведений России. В разделах "Публикации" и "Материалы" – много работ по различным разделам языкознания.

www.lingvisto.org
"Lingvisto – языковая энциклопедия" – активно развивающийся сайт, содержащий информацию о ряде языков (чешский, словацкий, арабский и др.).

www.languages-study.com
"Изучение языков в интернете: лучшие методики и пособия". Сайт содержит много интересной информации и ссылок по различным языкам мира.

www.peoples.org.ru
"Языки России в Интернете". Сайт, специально посвященный языкам народов России. Содержит информацию о веб-ресурсах по этой исключительно плохо разработанной теме.

Русский язык

www.gramota.ru
Справочно-информационный портал "Русский язык" – поддерживается Министерством по делам печати, телерадиовещания и средств массовых коммуникаций. Разделы: официальные документы, новости, журнал, мониторинг культуры речи и др.

www.rusword.org
"Мир слова русского" – сайт по русской филологии. Содержит популярную информацию по русскому языку и литературе.

www.russian.iztok.net
"Balkan Rusistics" – болгарский проект, посвященный русской филологии. Публикация статей, описание новых проектов, информация о новых книгах.

www.gramma.ru
"Грамма" – портал, посвященный культуре письменной речи. Содержит

литературу по этой теме и словари.

rusjaz.da.ru
"Русскій язык" – "ресурс для лингвистов-филологов, семиологов, учителей русского языка и литературы". Веб-проект Д. Яцутко. Имеется библиотека литературы по русскому языку.

ksana-k.narod.ru
"Библиотека Фронтистеса". В разделе "Эл. книги" – ряд ценных публикаций по русскому и старославянскому языку, а также общему языкознанию и некоторым другим языкам.

www.trishin.ru
В разделе "Некоммерческая деятельность" представлена электронная версия "Толкового словаря русского языка с синонимами", разработанного В.Н. Тришиным. Словарь содержит около 200 тыс. слов и выражений.

Английский язык

artefact.cns.ru/english
"English for Everyone" – сайт, посвященный изучению английского языка. Имеется несколько работ по грамматике английского языка, а также богатая библиотека книг на английском и некоторых других языках..

www.english4u.dp.ua
"Английский язык от ::English4U:: – интернет-портал изучающих английский язык" – грамматика, сленг, литература, тексты. Проект Д. Хозина. В разделе "Статьи" – много интересных работ по методике изучения языков и грамматике английского языка.

Немецкий язык

apuzik.deutschesprache.ru
"Немецкий язык – история языка, общая лингвистика, перевод, история Германии, этимология, грамматика" – веб-проект А.А. Пузика. Имеется богатый материал по изучению немецкого языка, а также библиотека лингвистической литературы.

frank.deutschesprache.ru
Сайт И.М. Франка. Много разнообразного материала, причем не только по немецкому языку, но и по другим языкам мира: от латинского до индонезийского.

Скандинавские языки

norse.narod.ru
"Norroen Dyrd" (Северная Слава) – активно развивающийся проект, посвященный истории и культуре Древней Скандинавии и древнеисландскому языку. Имеется библиотека работ по этим темам.

Латинский язык

linguaeterna.com
"Lingua Latina aeterna" – веб-проект М.П. Поляшева. Совместный российско-украинско-голландский ресурс по латинскому языку. В числе публикаций – большой русско-латинский словарь, составленный автором проекта.

Испанский язык

moscu.cervantes.es
Интернет-страница Института Сервантеса, испанского культурного центра под патронатом посольства Испании. Содержит информацию о языках и культуре Испании и стран Латинской Америки.

Алтайские языки (тюркские, монгольские и т.д.)

www.kyrgyz.ru
"Центральноазиатский исторический сервер" – веб-портал Р. Абдуманапова, содержащий ценный материал по истории, культуре, языкам и литературе тюркских народов Центральной Азии.

www.tataroved.ru
Сайт, посвященный истории и культуре татарского народа. Содержит ряд статей, затрагивающих проблемы татарского языка и тюрских языков в целом.

www.altaica.narod.ru
"Monumenta Altaica" – проект И. Грунтова, посвященный алтайской языковой семье. Имеется библиотека и богатая коллекция ссылок.

turkolog.narod.ru
"Тюркологические публикации" – сайт, посвященный истории, этнографии и языкам тюркских народов. В разделе "Библиотека" – много статей о тюркских языках и народах.

Баскский язык

euscara.narod.ru
"Euscara. Exotica On Line" – сайт К. Панфилова, содержащий информацию о баскском языке. Планируется размещение материалов по различным экзотическим языкам.

Прикладная лингвистика и лингвистическая экспертиза

www.philologia.ru
"Филология в задачах" – веб-проект С.А. Шаповал. Сайт посвящен анализу текста. Имеется раздел, включающий филологические задачи.

lexis-asu.narod.ru
"Юрислингвистика" – сайт учебно-исследовательской лаборатории

юрислингвистики и развития речи, созданной при Алтайском госуниверситете. Имеется библиотека лингвистической литературы.

Задание:

Найдите в электронных библиотеках книги по следующим темам:

1) *Учебные материалы по различным аспектам изучаемого Вами иностранного языка*

- *фонетика,*
- *лексика,*
- *грамматика,*
- *страноведение,*
- *история и т.п.*

2) *Существующие лингвистические ресурсы по классическим и современным авторам, пишущим на изучаемом Вами языке.*
3) *Книги по информационным технологиям и прикладной лингвистике.*

Тема 7: ОРГАНИЗАЦИЯ И КОМПЬЮТЕРНАЯ ОБРАБОТКА ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

1. Квантитативная лингвистика. Сферы применения количественных методов анализа
2. Дешифровка
3. Экспертиза авторства текста
4. Синтаксический парсинг
5. Контент-анализ
6. Организация данных в программе Excel (сортировка, статистическая обработка языковых данных)

§ 7.1. Квантитативная лингвистика. Сферы применения количественных методов анализа

Квантитативная лингвистика – это междисциплинарное прикладное направление, в котором объектом изучения является язык или речь, а инструментом анализа – количественные или статистические методы. Когда говорят, что в исследовании были использованы статистические методы, имеют в виду, что в ходе исследования были собраны некоторые данные, затем с помощью статистических приемов их обработали, а затем на основании полученных числовых данных сделали определенные выводы о свойствах изучаемого объекта.

Квантитативная лингвистика бурно развивается благодаря тому, что современные компьютеры позволяют хранить и автоматически обрабатывать большие массивы текстов.

Количественные данные проливают свет на наши представления о возможностях функционирования языковой системы. С помощью статистических методов могут быть проанализированы единицы любого уровня языка.

Количественные методы применяются чаще всего в лексикологии при изучении количественного состава словаря, процессов словообразования. Информация о частотности употребления того или иного слова либо словосочетания может оказаться полезной, например, при изучении иностранного языка, когда встает вопрос, какую именно лексику должен знать коммуникант для успешного общения. Данные о частоте использования слов могут оказать влияние на выбор говорящего в ситуации, когда из ряда синонимов необходимо выбрать одно слово.

Статистические методы анализа текста также используются для решения других задач: дешифровки текста, авторизации текста, для синтаксического парсинга, при проведении контент-анализа, в системах автоматического перевода, в информационно-поисковых системах.

Кратко рассмотрим некоторые возможности применения статистических методов в лингвистических исследованиях.

§ 7.2. Дешифровка

Дешифровка – это исследование сообщений или текстов для обнаружения информации, причем эта информация представлена способом, не известным исследователю. При дешифровке исследователь может столкнуться со следующими ситуациями:

1) Неизвестна только письменность, но язык известен. Решение этой задачи – это установление правил чтения забытых знаков.

2) Неизвестен только язык, но письменность известна. В качестве иллюстрации этой ситуации может выступать код. Решение такой задачи предполагает установление значения единиц языка, звучание которых известно. Эта ситуация называется интерпретацией.

3) Неизвестный язык записан неизвестным письмом. Аналогом здесь выступает зашифрованный код. Решение такой задачи, т.е. установление и звучания, и значения единиц называется раскрытием.

При дешифровке используются структурные методы, которые позволяют исследовать тексты на основе их формы, без привлечения значения. В основе структурного анализа лежит убеждение в том, основную информацию о языке можно получить непосредственно из текста (как письменного, так и устного), если изучить все встречающиеся в нем сочетания единиц. Разрабатываются алгоритмы, в основе которых заложены статистические данные о сочетаемости и частотности графем.

§ 7.3. Экспертиза авторства текста

Экспертиза авторства текста может быть рассмотрена с точки зрения трех возможных ситуаций:

1) Имеется множество текстов или их фрагментов. Необходимо установить, скольким авторам принадлежат эти тексты, и определить конкретное авторство каждого текста. Этот случай анализа называют множественной неопределенностью.

2) Вторая ситуация – это случай, когда имеется несколько образцов текстов определенного автора. Задача исследователя – определить, является ли он и автором некоторого другого текста. Такая ситуация называется сравнением по образцу.

3) В третьей возможной ситуации имеются образцы текстов нескольких авторов. Необходимо установить, кто из них является автором спорного текста. Это так называемая конкуренция образцов.

Поскольку в современной лингвистике авторский стиль понимается как категория структурно-синтаксическая, то использование количественных методов анализа оказывается неизбежным. Одно из перспективных направлений в этой области – теория распознавания образов. В рамках этой теории стиль описывается как пространство количественно выразимых параметров. Например, количественное описание получают средняя длина предложения, количество слов в предложении, количество предложений в абзаце и т.д. Далее анализируемый текст выражается через вектор,

координаты которого задаются значениями выбранных параметров. Сходство векторов является основанием для заключения о сходстве стилей.

В качестве примера можно привести многолетний спор по поводу того, кто является истинным автором романа «Тихий Дон». В свое время в него включились норвежские ученые под руководством Г. Хьетсо. Они взяли тексты, бесспорно принадлежащие М. Шолохову, и тексты донского писателя Ф. Крюкова, которому приписывалось авторство великого романа, и проанализировали их, выявляя особенности писательской манеры каждого. Учёные сравнили длину предложений, распределения длины предложений по количеству слов, распределение частей речи, сочетание частей речи в начале и в конце предложения и т.д. Сделать это оказалось возможным только с помощью мощной вычислительной техники. Вывод однозначен: из двух претендентов на авторство «Тихого Дона» Крюков явно обладает наименьшим правом. Недавно найденная рукопись великого романа (885 рукописных страниц, 605 из которых написаны рукой самого Шолохова, а 285 страниц – женой писателя и её сестрой) окончательно утвердила авторство М. А. Шолохова и правоту скандинавских ученых [9, сс.33-34].

Лингвоанализатор *(Д. Хмелев)* – on-line версия программы математического анализа структуры текста. Целью анализа является определение близости любого из предлагаемых пользователем текстов к одному из авторских эталонов, определенных заранее и взятых из ресурсов Русской Фантастики. Программа анализирует входной текст и выдает имена трех писателей, которые могли бы быть его наиболее вероятными авторами. Кроме этого, программа находит три произведения каждого из авторов, которые наиболее близки данному тексту.

§ 7.4. Синтаксический парсинг

Синтаксический парсинг – это процесс сопоставления линейной последовательности лексем языка с его формальной грамматикой. Результатом обычно является дерево разбора (синтаксическое дерево). При парсинге исходный текст преобразуется в структуру данных, обычно – в дерево, которое отражает синтаксическую структуру входной последовательности и хорошо подходит для дальнейшей обработки. Как правило, результатом синтаксического анализа является синтаксическая структура предложения, представленная либо в виде дерева зависимостей, либо в виде дерева составляющих, либо в виде некоторой комбинации первого и второго способов представления.

§ 7.5. Контент-анализ

Контент-анализ – количественный анализ текстов и текстовых массивов с целью последующей содержательной интерпретации выявленных числовых закономерностей. Смысл КА как исследовательского метода состоит в восхождении от многообразия текстового материала к абстрактной модели содержания текста.

Первые примеры использования КА датированы XVIII веком, когда в Швеции частота появления в тексте книги определенных тем служила критерием её еретичности. Однако всерьёз говорить о применении КА можно лишь начиная с 30-х годов XX века в США. Использовался он преимущественно в социологических исследованиях, в том числе при изучении рекламных и пропагандистских материалов. Применение компьютерной техники значительно активизировало методики контент-анализа (сейчас широко применяется в антропологии, управлении персоналом, психологии, литературоведении, истории, в связях с общественностью и т.д.).

С помощью контент-анализа можно анализировать такие различные типы текстов, как сообщения СМИ, заявления политических деятелей, программы партий, правовые акты, рекламные и пропагандистские материалы, исторические источники, литературные произведения.

Наименьшей единицей КА может являться слова (как правило) или тема, под которой в данном случае понимается отдельное высказывание об отдельном предмете. Единицу КА называют концептуальной переменной. Например, К-переменной могут быть такие категории, как «свой-чужой», «демократия», «права человека», «терроризм» и др. К конкретном тексте выбранная К-переменная будет реализовываться конкретными значениями. Например, А.Н. Баранов приводит следующие варианты для категории «свой-чужой»: мой, наш, мы, я, привычный, знакомый, близкий, их, его, он, она, непривычный, дальний, незнакомый и др. [1, с. 217]

Для правильности КА очень важно определить весь список значений или языковых репрезентантов К-переменной, чтобы результаты исследования стали максимально достоверными.

Проведение КА состоит из нескольких этапов:

1) отбор материала, формирование корпуса языковых данных (ими могут быть газетные публикации за определенный период, литературные произведения определенного жанра, написанные в определенную эпоху);

2) определение К-переменной и ее языковых реализаций в тексте;

3) выбор единицы кодирования (значения К-переменной могут приписываться текстам, их фрагментам (н-р, абзацам), предложениям, словосочетаниям, отдельным словам);

4) кодировка данных по заранее определенным единым критериям;

5) подсчет данных (составление специальных таблиц, применение компьютерных программ, специальных формул, статистических расчетов);

6) интерпретация результатов в соответствии с целями и задачами конкретного исследования (обычно на этом этапе выявляются и оцениваются такие характеристики текстового материала, которые позволяют делать заключения о том, что хотел подчеркнуть или скрыть его автор).

Такой вид КА называется содержательным (или качественным), т.к. в его процессе выделяется определенная концептуальная переменная и исследуется ее значение в тексте.

Если же исследователя интересует не столько что говорится (содержание), сколько как говорится, то в этом случае мы имеем дело со структурным (количественным) КА. При таком методе анализа изучаются формальные реализации К-переменной, например, фотографии, сколько места посвящалось определенной проблеме, на каких полосах газет и т.д.

§ 7.6. Квантитативные методики в гуманитарных науках

Еще в X веке ученый и философ эпохи Возрождения Николай Кузанский в трактате «Об ученом познании» утверждал, что все познания о природе необходимо записывать в цифрах, а все опыты над нею производить с весами в руках. Философ И. Кант был убеждён, что точное естествознание простирается до тех границ, в пределах которых возможно применение математического метода.

Если науки естественного цикла сравнительно давно заговорили на языке математики, то гуманитарные науки обратились к нему только в XX в. Первой среди них была лингвистика.

Связь языкознания с математикой наметилась уже давно. Так, известный русский лингвист И. А. Бодуэн де Куртене, набрасывая контуры будущего языкознания, непременным условием его считал тесную и органическую связь с математикой. «Нужно чаще применять в языкознании количественное, математическое мышление и таким образом приблизить его всё более к наукам точным» [цит. по 9, сс. 29-31].

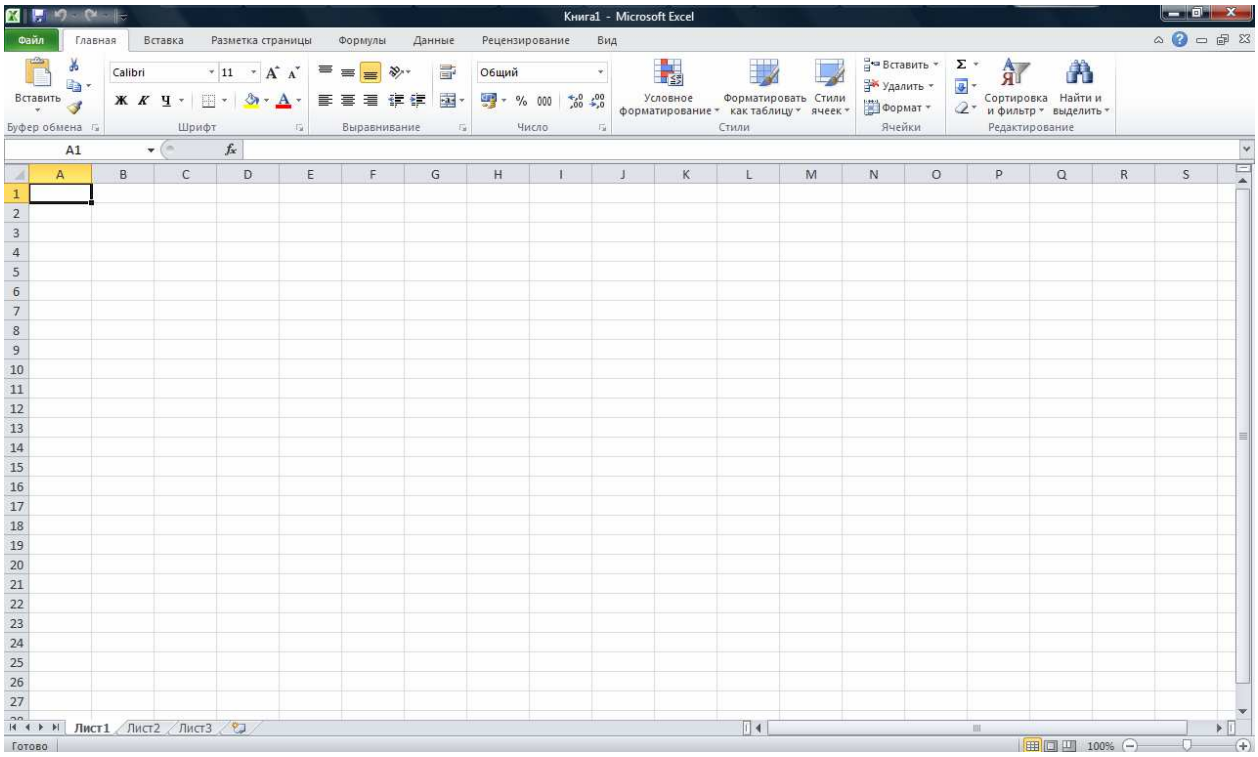
Для описания и исследования лингвистических фактов привлекаются различные разделы математики: алгебра, теория множеств, математическая логика, теория информации, теория вероятностей и математическая статистика.

Квантитативная лингвистика отличается от математической лингвистики бóльшим вниманием к языковой специфике, которая стоит за количественными отношениями. Главная её задача – поиск связи между количественными и качественными сторонами языка: между длиной слова и его употребительностью, полисемией и употребительностью и т.д.

§ 7.7. Организация данных в программе *EXCEL* (сортировка, статистическая обработка языковых данных)

Microsoft Excel – программа для работы с электронными таблицами, созданная корпорацией *Microsoft* для *Windows*. Она предоставляет возможности статистических расчетов и графические инструменты; входит в состав Microsoft Office и на сегодняшний день *Excel* является одним из наиболее популярных приложений в мире.

Первая версия *Excel* для *Windows* была выпущена в ноябре 1987 года. Текущая версия для платформы Windows – *Microsoft Office Excel 2010*.



Excel был первым табличным процессором, позволявшим пользователю менять внешний вид таблицы (шрифты, символы и внешний вид ячеек). Он также первым представил метод умного пересчета ячеек, когда обновляются только те ячейки, которые зависят от изменённых ячеек (раньше табличные процессоры либо постоянно пересчитывали все ячейки или ждали команды пользователя).

Электронная таблица – компьютерная программа, позволяющая проводить вычисления с данными, представленными в виде двухмерных массивов, имитирующих бумажные таблицы. Использование математических формул в ЭТ позволяет представить взаимосвязь между различными

параметрами некоторой реальной системы. Решения многих вычислительных задач (в том числе и лингвистических), которые раньше можно было осуществить только с помощью программирования, стало возможно реализовать через математическое моделирование в электронной таблице.

На сегодняшний день метод работы с электронными таблицами широко используется квантитативной лексикологией. Общепризнанно, что словарь языка можно стратифицировать на ядро и периферию. В ядре будут находиться наиболее частотные единицы языка, на периферии – наименее употребительные.

Частотность лексики можно определять по 4-м параметрам: 1) функциональная активность (употребительность; здесь важна длина слов – чем слово короче, тем чаще оно употребляется), 2) синтагматическая активность (широкая сочетаемость, способность сочетаться с большим количеством лексических единиц, в том числе ФС), 3) парадигматическая активность (вхождение в многочисленные синонимические ряды), 4) эпидигматическая активность (многозначность, полисемия).

Лексика любого языка помещается в соответствующие ячейки таблицы, потом производится подсчет по каждому из параметров, ранжирование лексики и выделение «ядерной» и «периферийной» лексики.

Результаты подобной «обработки» языкового материала находят практическое применение при составлении кратких двуязычных словарей, разговорников, на начальных этапах обучения ИЯ и др.

Задания:
1) Сделайте сообщение о применении метода дешифровки в лингвистике.
2) Напишите доклад о наиболее известных экспертизах авторства текста.

Тема 8: КОРПУСНАЯ ЛИНГВИСТИКА: ПОИСКОВЫЕ И АНАЛИТИЧЕСКИЕ ВОЗМОЖНОСТИ

1. Лингвистические корпуса как источник информации о языке, их практическое использование.
2. Из истории лингвистических корпусов.
3. Принципы отбора и обработки материала в языковых корпусах.
4. Типы корпусов.
5. Современные корпуса текстов: национальный корпус русского языка; Британский национальный корпус; другие иноязычные лингвистические корпуса.
6. Параллельные корпуса.

§ 8.1. Лингвистические корпуса как источник информации о языке, их практическое использование

Корпусная лингвистика – раздел языкознания, занимающийся разработкой, созданием и использованием текстовых (лингвистических) корпусов.

Лингвистическим корпусом называют совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой. Иногда корпусом («корпус первого порядка») называют просто любое собрание текстов, объединённых каким-то общим признаком (языком, жанром, автором, периодом создания текстов).

Целесообразность создания текстовых корпусов объясняется:
1) представлением лингвистических данных в реальном контексте;
2) достаточно большой репрезентативностью данных (при большом объёме корпуса);
3) возможностью многократного использования единожды созданного корпуса для решения различных лингвистических задач. Среди них можно выделить следующие:
1) в лексикографии и лексикологии – для составления различных словарей, определения значений многозначных слов, выявления ассоциативных связей слов в тексте и т.д.,
2) в грамматике – для определения частоты употребления грамматических морфем в Различных текстах, выявления наиболее употребляемых типов словосочетаний и предложений, частоты употребления классов слов;
3) в лингвистике текста – для дифференциации типов текста, выявления связей между предложениями в абзацах, между абзацами и т.д.,
4) при автоматическом переводе текстов – для поиска контекстов слов, имеющих несколько переводных эквивалентов, поиска переводных эквивалентов в параллельных текстах и т.д.,
5) в учебных целях – для выбора цитат, отдельных фрагментов произведений, примеров, при создании учебников и учебных пособий и т.д.;

6) к корпусам текстов также обращаются программисты, занимающиеся разработкой систем автоматической обработки текстов. Для них корпус служит своеобразным «полигоном», на котором проверяется эффективность работы компьютерных программ.

§ 8.2. Из истории лингвистических корпусов

Первым большим компьютерным корпусом считается Брауновский корпус американского варианта английского языка, который был создан в 1962-63 гг. под руководством У. Фрэнсиса в Университете Брауна и содержал 500 фрагментов текстов по 2 тысячи слов в каждом. В результате он задал стандарт в 1 млн словоупотреблений для создания представительных корпусов на других языках. По модели близкой к БК в 1970-е годы был создан частотный словарь русского языка Л.Н. Засориной, построенный на основе корпуса текстов объемом также в 1 миллион слов и включавший примерно в равной пропорции общественно-политические тексты, художественную литературу, научные и научно-популярные тексты из разных областей и драматургию. По аналогичной модели был построен и русский корпус, созданный в 1980-е годы в Университете Уппсалы, Швеция.

Размер в один миллион слов достаточен для лексикографического описания только самых частотных слов, поскольку слова и грамматические конструкции средней частоты встречаются по несколько раз на миллион слов (со статистической точки зрения язык является большим набором редких событий). Так, каждое из таких обыденных слов, как, н-р, англ. *polite* ‘вежливый’ или англ. *sunshine* ‘солнечный свет’ встречается в БК всего 7 раз, выражение англ. *polite letter* лишь один раз, а такие устойчивые выражения как англ. *polite conversation, smile, request* ни разу.

По этим причинам, а также в связи с ростом компьютерных мощностей, способных работать с большими объемами текстов, в 1980-е годы в мире было предпринято несколько попыток создать корпуса большего размера. В Великобритании такими проектами были Банк английского языка *(Bank of English)* и Британский Национальный Корпус *(British National Corpus, BNC)*. В СССР таким проектом был Машинный Фонд русского языка, создававшийся по инициативе А. П. Ершова.

В настоящее время представительные корпуса существуют (или разрабатываются) для немецкого, польского, чешского, словенского, финского, новогреческого, армянского, китайского, японского и других языков. Национальный корпус русского языка, создаваемый при РАН, содержит на сегодняшний день более 149 млн словоупотреблений.

§ 8.3. Принципы отбора и обработки материала в языковых корпусах

Создание корпусов включает в себя отбор текстов, разработку средств кодирования и средств поиска внутри базы данных. Подбор текстов осуществляется на основе четко сформулированных критериев (жанровая принадлежность текста, время его создания и др.).

Поскольку собрать все тексты языка практически невозможно (исключение составляют лишь мертвые языки), при отборе текстов необходимо следить за тем, чтобы были равномерно представлены все стили языка, чтобы в корпус были включены примеры употребления низкочастотной лексики. Считается, что для национального корпуса, который достоверно описывает некоторый язык, размер базы данных должен быть не менее 100 миллионов словоупотреблений. О таком корпусе принято говорить как о достаточно репрезентативном. Репрезентативность корпуса – одно из важнейших условий его использования. Поскольку от того, насколько корпус является репрезентативным, зависит достоверность результатом исследований, которые проводятся на материале этого корпуса.

В корпус могут включаться тексты самых разных жанров: произведения художественной литературы, публикации СМИ, деловые документы, записи диалогов, телевизионных ток-шоу, переписка по электронной почте и т.д. Такой подбор обеспечивает репрезентативность, т.е. показывает, как на самом деле функционирует язык в обществе. Иногда оказывается, что языковое употребление значительно расходится с нормой, представленной в грамматиках и словарях. Необходимо также отметить, что включение в корпус текстов, отражающих реальный процесс использования языка в определенных коммуникативных контекстах, контрастирует с подходом генеративной лингвистики, где порождение высказываний – это результат размышлений лингвиста об использовании языка, т.е. высказывания порождаются вне контекста на основе строгих правил.

Включенные в корпус тексты получают морфологическую и синтаксическую разметку (или аннотацию), которая необходима для того, чтобы пользователь мог осуществлять поиск необходимых фрагментов по заданным параметрам. Разметка – это приписывание грамматической информации о входящих в тексты словоформах. Наличие такой информации значительно обогащает корпус и облегчает процедуру поиска. Другие преимущества разметки заключаются в ее эксплицитности (т.е. информация о грамматических свойствах словоформ дается в явном виде), а также многофункциональности (аннотированный корпус может быть использован в различных исследовательских целях).

Разметка может осуществляться как вручную, так и в автоматическом режиме. Для того чтобы аннотирование происходило автоматически, специалисты по корпусной лингвистике используют специальные программы: лемматизаторы (т.е. программы, функция которых – приведение словоформы к начальной форме), программы, расставляющие указатели частей речи (*part-of-speech taggers*) и др.

Кроме того, в больших корпусах возникает проблема, которая ранее была неактуальной: поиск по запросу может выдавать сотни и даже тысячи результатов (контекстов употребления), которые просто физически невозможно просмотреть в ограниченное время. Для решения этой проблемы разрабатываются системы, позволяющие группировать результаты поиска и

автоматически разбивать их на подмножества (кластеризация результатов поиска), либо выдающие наиболее устойчивые словосочетания (коллокации) со статистической оценкой их значимости.

§ 8.4.Типы корпусов

Существуют различные подходы к классификации корпусов текстов в зависимости от типа текстов, способов их организации, языка и т.д.

С точки зрения их использования лингвистами наиболее значимы следующие виды корпусов:

1) исследовательские – создаются с целью изучения различных аспектов функционирования языка;

2) иллюстративные – служат для выделения в них лингвистических примеров, подтверждающих те или иные языковые факты, обнаруженные иными лингвистическими приемами;

3) статические – содержат тексты какого-то небольшого временного промежутка;

4) в динамические корпуса включают письменные источники большого временного периода, они предназначены для проведения различных диахронических исследований.

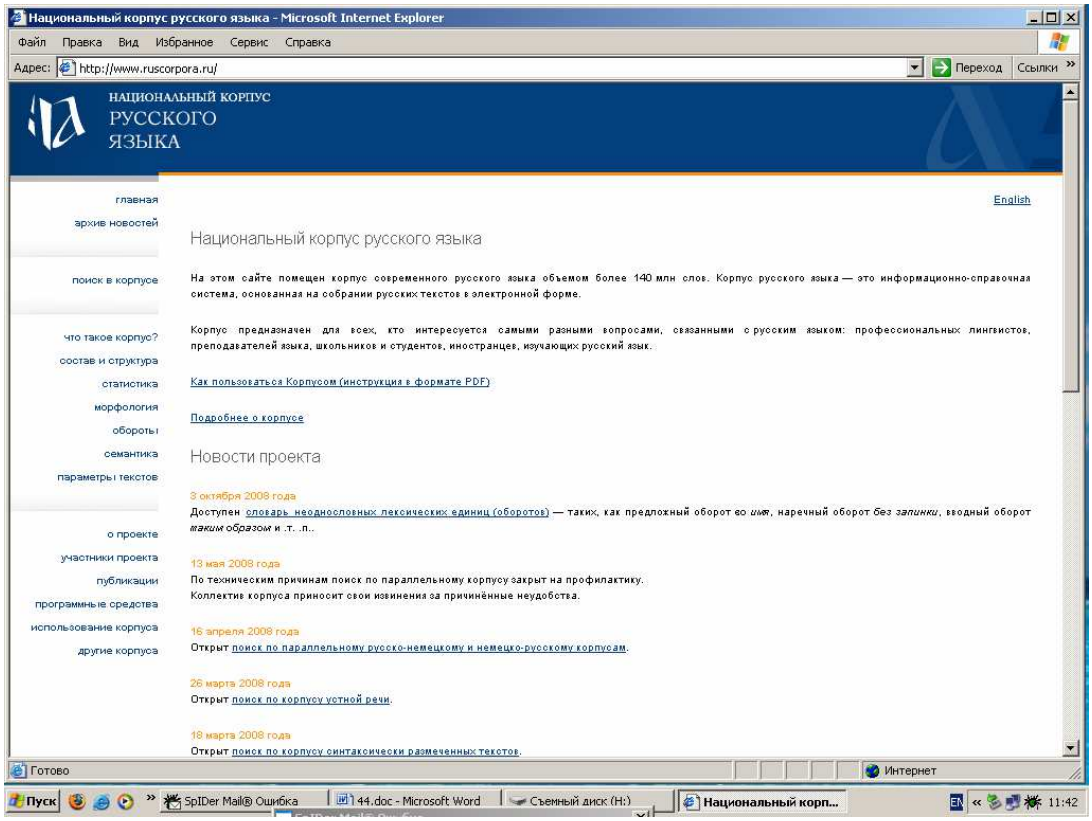
Если в корпус включены тексты только на одном языке, то это одноязычный корпус. Существуют также многоязычные корпуса, которые объединяют несколько одноязычных корпусов с приблизительно одинаковой выборкой текстов и репрезентативностью. Также разрабатываются корпуса параллельных текстов: в них включаются тексты с их переводами на другой язык (или языки).

§ 8.5. Современные корпуса текстов

Национальный корпус русского языка

Корпусная лингвистика в России развивается с некоторым отставанием. Первые электронные корпуса РЯ начали появляться не в России, а в Европе. Самым известным из таких корпусов является Упсальский корпус русского языка, созданный в Швеции. Сегодня этот корпус хранится на сервере Тюбингенского университета в Германии.

Национальный корпус русского языка (*http://www.ruscorpora.ru/*) – общедоступный для поиска электронный онлайновый корпус русских текстов – был создан недавно (2004 г.) и находится в стации разработки.



В Корпус входят как письменные тексты (художественные, мемуары, публицистика, научная, религиозная литература, повседневная печатная продукция), так и записи устных текстов (публичной речи и частных бесед). В корпус также входят подкорпуса поэтических и диалектных текстов, русско-английский, англо-русский и немецко-русский корпуса параллельных текстов, синтаксический, акцентологический и обучающий подкорпуса. Объём Национального корпуса русского языка составляет свыше 70 тыс. текстов общим объемом свыше 150 млн словоупотреблений. На сегодняшний день в корпусе используется четыре типа разметки: метатекстовая, морфологическая, акцентная и семантическая. Поиск можно осуществлять как во всем массиве текстов, так и в текстах, отобранных по определенному критерию (жанр, автор, время написания и др.).

Британский национальный корпус

Одним из самых авторитетных корпусов сегодня считается *British National Corpus (http://www.natcorp.ox.ac.uk)*. Корпус был создан в 1990-х гг. Правила разметки, которые использовались при его создании, принялиза образец еще более ста появившихся позднее систем. Появление корпуса стимулировало развитие англоязычной лексикографии: данными корпуса пользуются при составлении наиболее авторитетных англоязычных словарей.

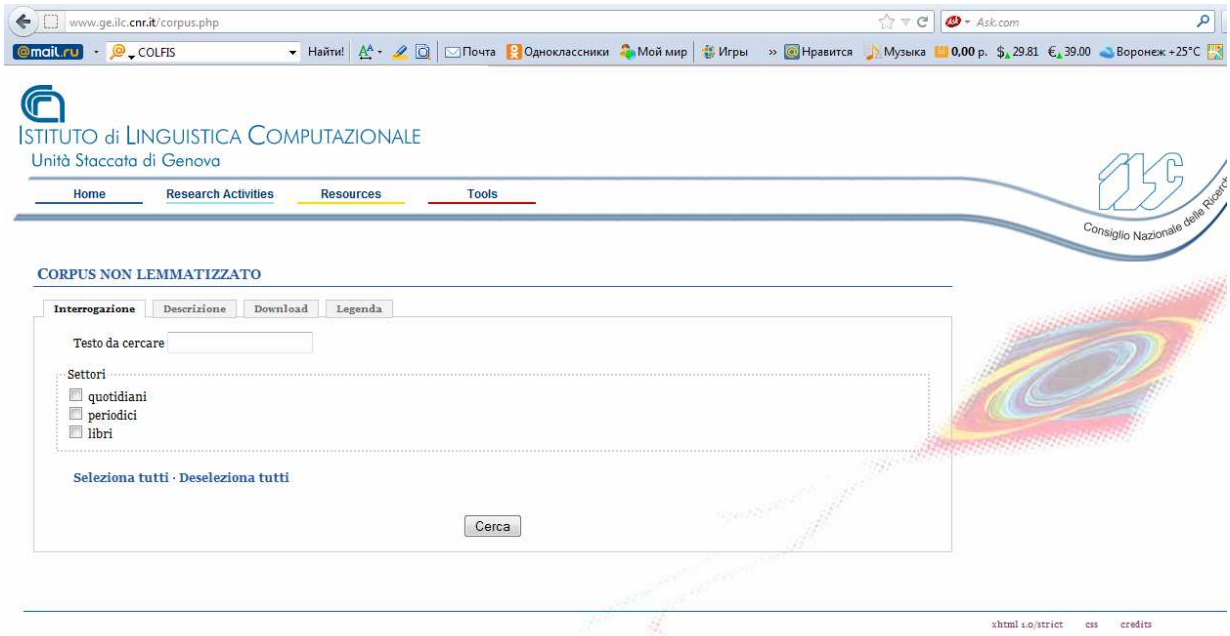
Другие иноязычные лингвистические корпуса

Корпус современного американского варианта английского языка *(Corpus of Contemporary American English – COCA)* – создан проф. Марком Дэвисом из Университета Бригама Янга *(Brigham Young University)*. В настоящее время насчитывает более 425 млн. слов и более 160 млн. текстов за период с 1990 по 2011 г.

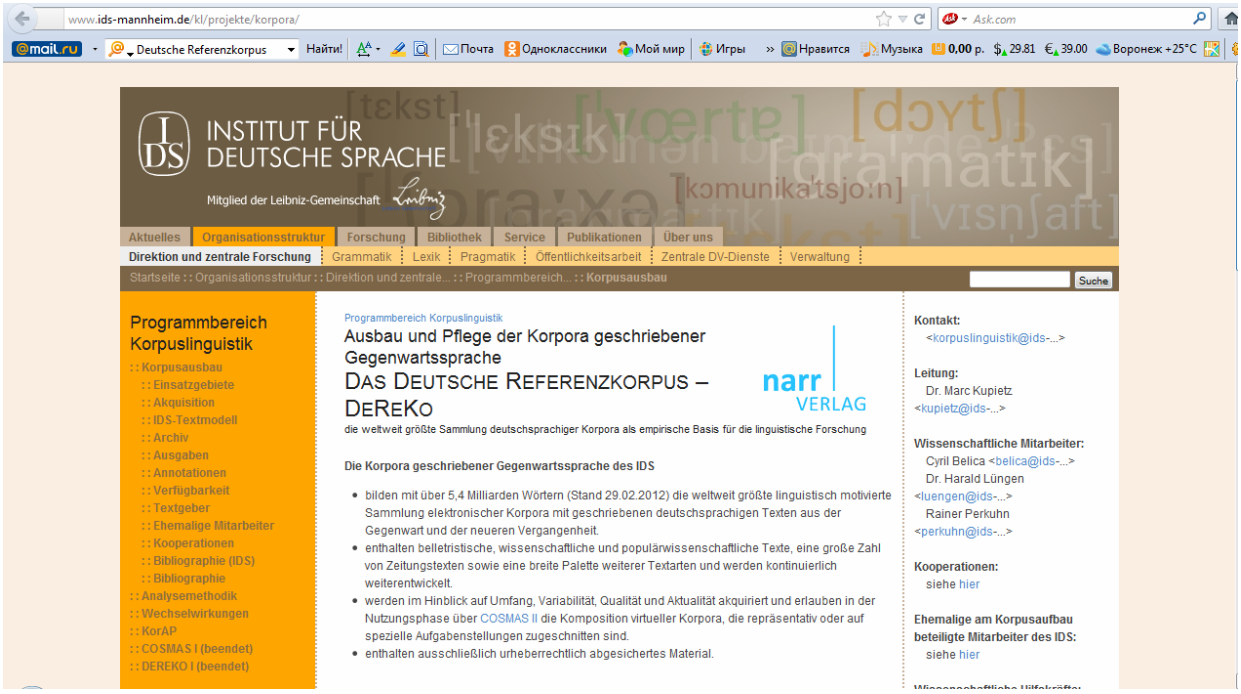
Корпусная лингвистика во Франции: один из наиболее грандиозных французских проектов – создание «Сокровищницы французского языка» (Tresor de la langue francaise informatise), включающей корпус текстов в 90 млн словоупотреблений. Разработка проекта началась в 1963 г., ввод основного материала был закончен к 1968 г. Введены тексты 19-20 вв., всего 80 тыс. вводов.

Корпус испанского языка – *Corpus del Espanol*.

Корпуса итальянского языка – *COLFIS, CORIS, CODIS.*



Корпус немецкого языка – *Deutsche Referenzkorpus (DeReKo)* (в Институте немецкого языка в Мангейме).



§ 8.6. Параллельные корпуса

Параллельный текст (битекст) – текст на одном языке вместе с его переводом на другой язык. Большие собрания параллельных текстов называются «параллельным корпусом» (англ. *parallel corpora*).

«Выравнивание параллельного текста» – это идентификация соответствующих друг другу предложений в обеих половинах параллельного текста. Эта процедура является необходимой предпосылкой для различных

аспектов лингвистических исследований. В процессе перевода предложения могут разделяться, сливаться, удаляться, вставляться или менять последовательность. В связи с этим выравнивание часто становится сложной задачей.

В области перевода «битекст» – это совмещенный документ, состоящий из версий соответствующего текста на исходном и целевом языках.

Битексты создаются с помощью специальных компьютерных программ, которые называются «инструментами для выравнивания» (*alignment tool*) или «инструментами для битекста» (*bitext tool*), которые позволяют автоматически выравнивать оригинальную версию текста и его перевод. Подобные программы, как правило, приводят в соответствие два текста (оригинал и перевод) по каждому предложению.

Идея битекста принадлежит Брайану Хэррису (Brian Harris), который первым написал исследование по данной концепции в 1988 году, и была впоследствии развита группой ученых при Университете Монреаля.

Задания:

1) Используйте параллельный корпус НКРЯ для поиска и анализа вариантов перевода следующих слов:

- watch (глагол);
- like (глагол);
- light (прилагательное);
- cool (прилагательное);
- force (существительное);
- mind (существительное).

2) Используя НКРЯ, проанализируйте левостороннюю сочетаемость слова «стол» *с прилагательными.*

3) Проанализируйте правостороннюю сочетаемость слова «вода» *с глаголами.*

4) Проанализируйте правостороннюю сочетаемость слова «окно» *с существительными в родительном падеже.*

Тема 9: КОМПЬЮТЕРНАЯ ЛЕКСИКОГРАФИЯ

1. Лексикография: направления исследования и задачи.
2. Типы словарей
3. Основные структурные компоненты словаря
4. Основные структурные компоненты словарной статьи
5. Компьютерная лексикография.
6. Принципы создания электронного словаря
7. Электронные словари в Интернете

§ 9.1. Лексикография: направления исследования и задачи

Лексикография – это теория и практика составления словарей. Поэтому говорят о двух направлениях исследований: практической лексикографии и теоретической лексикографии.

Практическая лексикография выполняет несколько общественно важных функций:

1) обеспечивает обучение языку, как родному, так и неродному;
2) словари различных типов описывают и нормализуют родной язык (т.е. в них разрабатывается языковая норма);
3) словари обеспечивают межъязыковое общение;
4) на основе словарей проводятся исследования в рамках теоретической лингвистики.

Теоретическая лексикография охватывает следующий комплекс проблем:

1) разработка общей типологии словарей и словарей новых типов:
2) разработка макроструктуры словаря (отбор лексики, установление принципов расположения слов и словарных статей, выделение омонимов, вопросы включения в словарь иллюстраций или грамматических статей);
3) разработка микроструктуры словаря, т.е. отдельной словарной статьи (например, решение о включении в словарь фонетической информации и грамматического комментария к слову; о выделении и классификации значений у многозначных слов; о системе помет и специальных знаков и др.).

§ 9.2. Типы словарей

Словарь – это определенным образом организованное собрание слов с комментариями к ним. В комментариях разъясняются смысловая или формальная структура данного слова и особенности его функционирования. Помимо слов объектами словарного описания могут стать компоненты слов (например, существуют словари морфем) или словосочетания различных типов (пословицы, поговорки, фразеологизмы, крылатые выражения, цитаты и т. д.). Все словари делятся на две категории: энциклопедические и лингвистические.

Энциклопедические словари – это научные или научно-популярные издания, которые представляют собой систематизированный свод знаний в каких-либо областях.

Объектом описания энциклопедических словарей являются понятия, термины, исторические события, персоналии, географические реалии.

Словники энциклопедических словарей в основном включают существительные и сочетания с ними.

Словарная статья энциклопедического словаря содержит в основном экстралингвистическую информацию и сопровождаться иллюстрациями, схемами или картами.

Энциклопедические словари в свою очередь подразделяются на универсальные («Большой энциклопедический словарь»), отраслевые (например, «Философский энциклопедический словарь»), иногда выделяют региональные словари (типа «Африка»). Особо выделяются биографические словари, объектом которых является жизнь и деятельность ученых, политиков, деятелей искусства и т.д.

Объектом описания *лингвистических* словарей являются языковые единицы: слова, устойчивые словосочетания, морфемы и др.

Словники лингвистических словарей включают все части речи.

В лингвистических словарях слово описывается с точки зрения его языковых и речевых характеристик (в словарную статью будет включена, например, грамматическая информация, помечена стилистическая окраска слова, данные о происхождении лексемы и т.д.).

Что касается лингвистических словарей, то по количеству используемых языков их можно разделить на одноязычные, двуязычные или многоязычные, или переводные.

В зависимости от целей и способов лексикографического описания выделяют толковые лингвистические словари. Их основная задача – объяснение значений слов и иллюстрация их употребления в речи.

По функциям и целям толковые словари делятся на дескриптивные и нормативные.

Цель дескриптивного толкового словаря – дать наиболее полное описание лексики и все имеющиеся релевантные случаи употребления. Дескриптивными являются словари сленгов и жаргонов, диалектные словари.

Цель нормативного словаря – показать норму употребления слова. При этом должны исключаться не только употребления, считающиеся неправильными вследствие незнания значения слова, но и те употребления, которые так или иначе не соответствуют коммуникативной ситуации (например, жаргонные). Таким образом, нормативные толковые словари задают литературную норму. Словари этого типа являются действенным инструментом языковой политики.

По характеру словника толковые словари делятся на общие и частные. БАС (17-ти томный словарь современного русского литературного языка), МАС (4-х томный словарь русского языка), словарь Ожегова, Ушакова, Даля

– примеры общих словарей. К частным относятся фразеологические словари, словари сленгов, жаргонов, диалектные словари, словари иностранных слов. Словники этих словарей ограничены сферой использования языкового материала.

По способу описания основных единиц словаря и отношений между ними выделяются словари синонимов, антонимов, омонимов и паронимов. В словарях синонимов в одну словарную статью помещаются близкие по значению слова, образуя синонимические ряды. Если в словарях синонимов толкование значения необязательно, иногда значение становится ясно из самого синонимического ряда, то в словарях омонимов толкование значений обязательно. Паронимические словари содержат слова с частичным звуковым сходством при их семантическом различии (главный – заглавный, гуманный – гуманитарный и др.)

Словари, отражающие некоторые тематические и стилистические пласты лексики делятся на: терминологические, диалектные, словари просторечий, словари арго (табуированная лексика), словари языков писателей.

По расположению материала в словарях – идеографические (содержат не слова, а рисунки со смыслов, например, глаз с капающей слезой – горе), аналогические (в них слова располагаются не по алфавиту, а по смысловым ассоциациям – мебель: стул, стол, кровать, диван, кресло и т.д.), обратные (слова располагаются по алфавиту конечных букв).

По назначению (адресату) – словари трудностей какого-либо языка, словари ошибок, учебные словари, «ложные друзья» переводчика.

Тезаурусы (от греч. θησαυρός ‘сокровище’) – особая разновидность словарей общей или специальной лексики, в которых указаны семантические отношения (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т. п.) между лексическими единицами. В отличие от толкового словаря, тезаурус позволяет выявить смысл не только с помощью определения, но и посредством соотнесения слова с другими понятиями и их группами.

В прошлом термином тезаурус обозначались по преимуществу словари, с максимальной полнотой представлявшие лексику языка с примерами её употребления в текстах.

Первый тезаурус (тезаурус Роже) был создан в 1852 г. в Англии как средства «для облегчения выражения мыслей и помощи при написании сочинений». Особенно широко тезаурусы использовались в 70-х гг. XX в., когда с целью поиска фактических данных об оборудовании (станках, автомобилях и т.п.) или технической литературы начали активно разрабатываться информационно-поисковые системы.

§ 9.3. Основные структурные компоненты словаря

Каждый словарь состоит из ряда компонентов, благодаря которым читатель или пользователь, в том числе и неопытный, может достаточно легко найти в словаре необходимую информацию. Разработанность структуры влияет на удобство пользования словарем.

Важнейшим компонентом любого словаря является словник. В словник включаются все единицы, которые входят в область описания словаря и являются входами словарных статей.

Элементарной единицей любого словаря является словарная статья. Словарной статьей называется каждый отдельно взятый объект описания словаря и относящиеся к этому объекту словарные характеристики. Множество словарных статей формирует основной текст словаря.

Еще одним важным структурным компонентом словаря являются указатели или индексы. В обычных толковых словарях индексы встречаются нечасто, однако в словаре идиом указатели необходимы для того, чтобы идиому можно было бы легко найти по любому из ее компонентов. Также без указателей не обходятся тезаурусы, поскольку необходимо определить, в какие таксоны, т.е. тематические группы, входит то или иное слово.

Профессионально сделанные словари включают в себя также вводную статью, где авторы объясняют принципы пользования словарем, структуру словарной статьи, указывают на объем словника и т.д.

Отдельным компонентом крупного словаря обычно является список сокращений. Часто для удобства пользователя отдельно печатается алфавит. Он необходим, как правило, для иноязычных пользователей.

§ 9.4. Основные структурные компоненты словарной статьи

Базовая единица словаря – это словарная статья. Обычно словарная статья состоит из нескольких зон описания, причем каждая зона содержит особый тип словарной информации. Количество зон и характер информации зависит от типа словаря.

Самой первой зоной словарной статьи является лексический вход, который также называют вокабула или лемма. Часто в вокабуле указывается ударение. Графически лексический вход отмечается полужирным шрифтом.

Вслед за лексическим входом чаще всего идет зона грамматической информации и зона стилистических помет. В качестве грамматической дается информация о принадлежности слова к части речи, указываются особые грамматические формы. Стилистические пометы указывают на сферу употребления слова (например, является ли слово термином, или принадлежит к слою разговорной или просторечной лексики).

Далее следует зона значения, внутри которой выделяют следующие подзоны: номер значения, дополнительные грамматические и стилистические пометы, толкование, примеры и иллюстрации употребления, зона оттенков значения.

В толковых словарях словарная статья, как правило, завершается зоной фразеологизмов. В некоторых случаях в словарной статье дается информация о происхождении слова, тогда говорят об особой этимологической зоне.

Словарная статья тезауруса отличается тем, что в ней представлена иерархия семантических отношений внутри лексики. Поэтому, применительно к структуре тезауруса, принято использовать термин таксон.

Таксон – это любая совокупность слов, словосочетаний, объединенных общей темой. Иными словами, в основании таксона лежит единство семантики входящих в этот таксон единиц.

Для маркировки различных зон словарной статьи используются различные виды графического выделения. Это позволяет читателю легко находить необходимую информацию. Например, лексический вход всегда выделяется полужирным шрифтом, грамматическая информация может даваться либо более мелким шрифтом, либо выделяться курсивом. Знак ромба обычно отделяет фразеологическую зону, и т.д.

§ 9.5. Компьютерная лексикография

С развитием компьютерных технологий в лексикографии появилась новая отрасль – компьютерная лексикография, занимающаяся созданием электронных словарей. Сегодня она является особым направлением в практической лексикографии со своими собственными подходами не только к отображению, но и к содержанию словаря.

Компьютерная лексикография развивается в двух направлениях, во-первых, компьютеры используются для создания обычных, бумажных словарей, поскольку это значительно упрощает работу с языковым материалом и ускоряет процесс создания словаря. Во-вторых, создаются собственно электронные словари.

§ 9.6. Принципы создания электронного словаря

Электронный словарь – это особый лексикографический объект, в котором могут быть реализованы и введены в обращение многие продуктивные идеи, облегчающие и ускоряющие поиск информации. Главная особенность электронного словаря – это возможность быстрого получения информации и ее сортировки по различным критериям.

Всякий электронный словарь состоит из двух частей – собственно словаря, т.е. текста или базы данных особого формата, и программы, позволяющей задать вопрос и быстро получить на него ответ. Чтобы пользователь мог получить наиболее полную информацию за короткий промежуток времени, в электронных словарях применяются различные лингвистические технологии: морфологический и синтаксический анализ, полнотекстовый поиск, распознавание и синтез звука и др.

Процесс создания обычного словаря с помощью компьютера проходит несколько этапов. Вначале создаются лексикографические базы данных, на основе которых затем строятся словарные статьи. Еще один важный этап создания электронного словаря – это подбор примеров, которые должны показать, в каких контекстах употребляется данное слово. Подбор примеров осуществляется не вручную, а с использованием корпусов текстов, которые хранятся в памяти компьютера. Выбор происходит в автоматическом режиме. Поиск примеров на употребление слова, или контекстов его употребления, называется построением конкордансов.

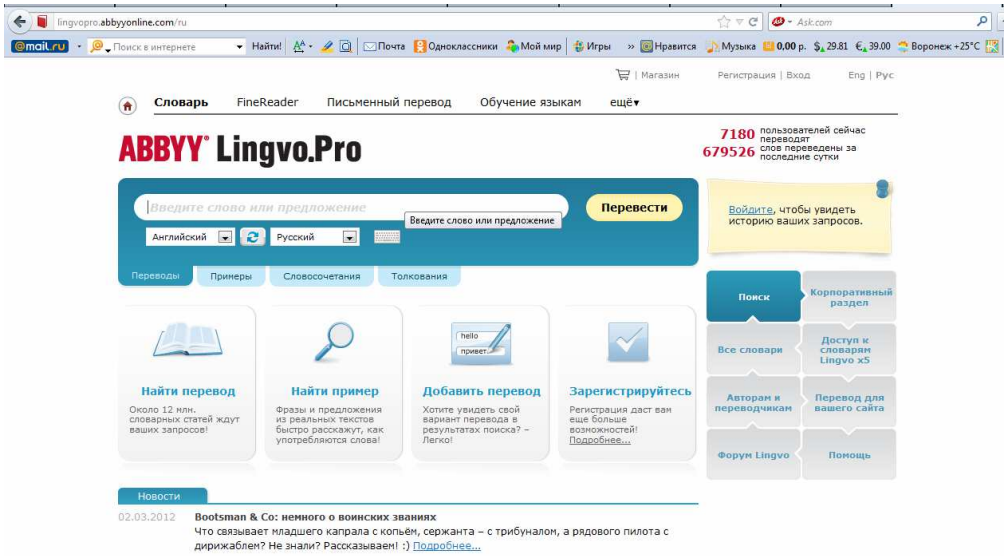
Как правило, этот контекст состоит из трех предложений: 1) предложения, в котором встретилось данное слово (словоформа), 2) предложения, стоящего перед основным предложением, 3) предложения, стоящего после него. Предполагается, что контекст такого объема является достаточно полным и содержит внутри себя отрезок, законченный в смысловом отношении.

После подбора материала начинается составление словарных статей, но делается это не в текстовом редакторе, а в базе данных. Такой режим работы существенно упрощает создание системы отсылок и указателей, сортировку данных.

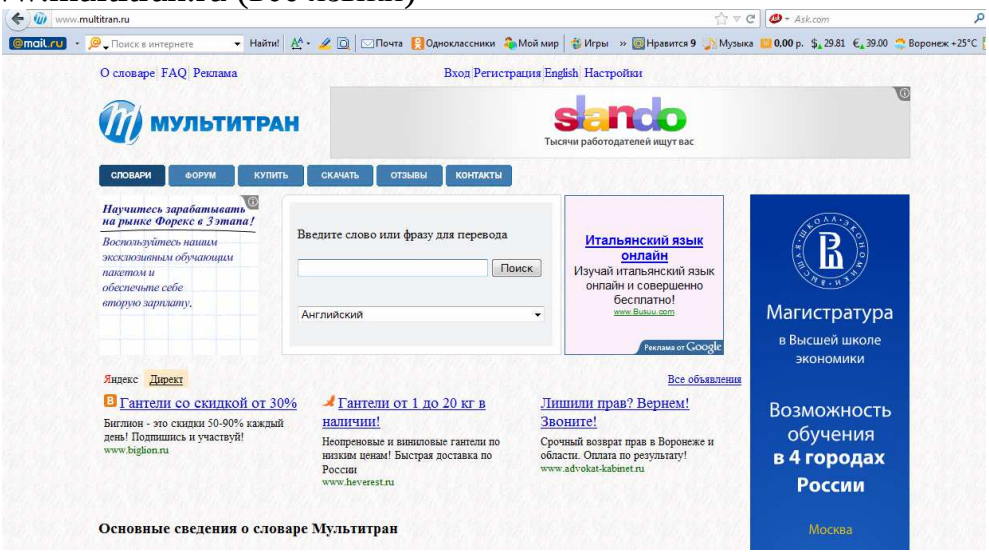
Заключительные этапы создания словаря – формирование текста словаря, его редактирование и корректура, создание оригинал-макета книги. При этом поля записи базы данных автоматически преобразуются в привычные для нас зоны словарной статьи.

§ 9.7. Электронные словари в Интернете

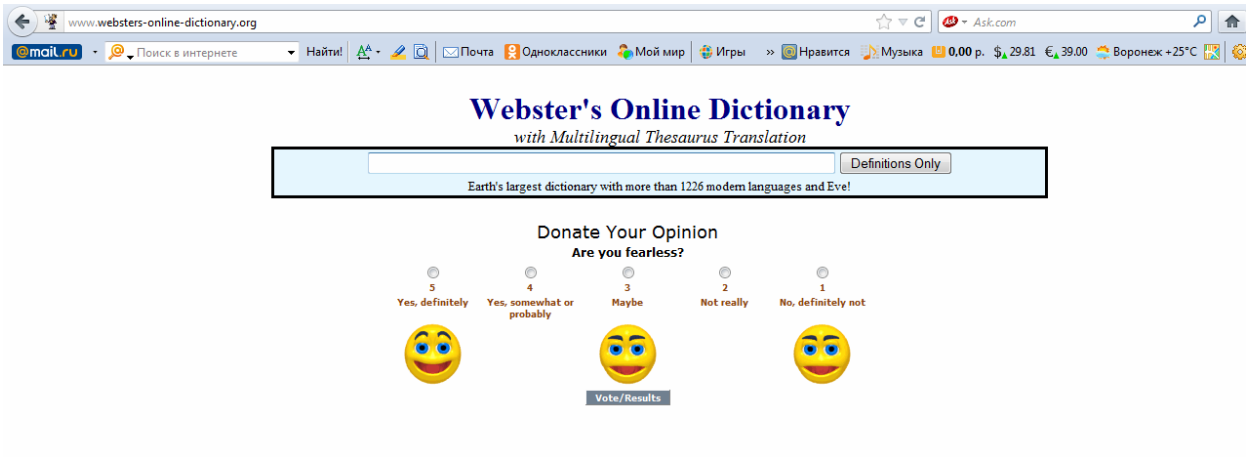
http://lingvopro.abbyyonline.com/ru (все языки)



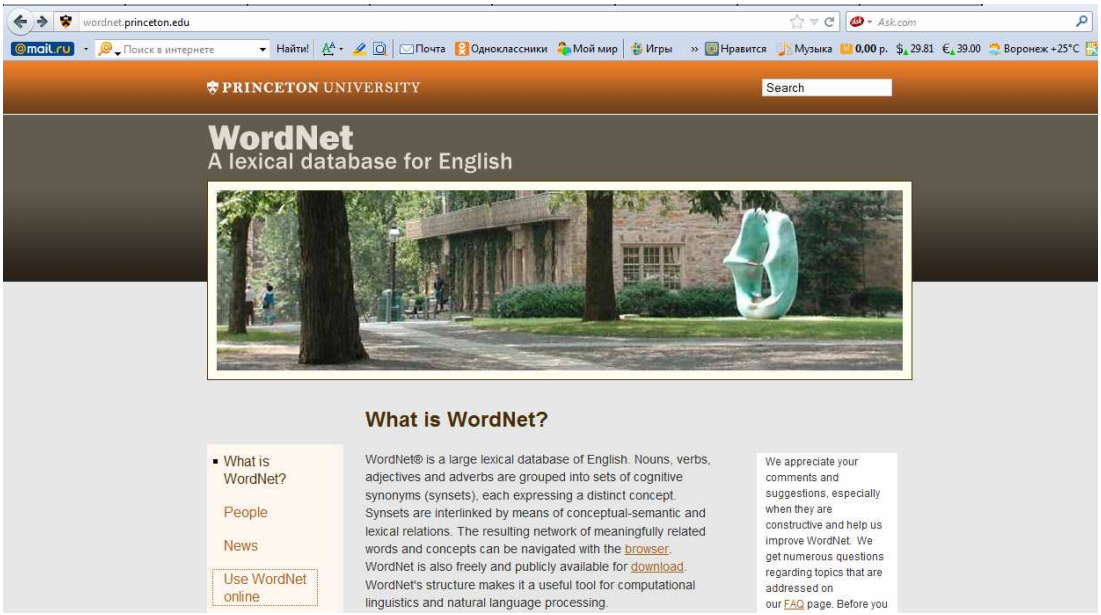
http://www.multitran.ru (все языки)



http://www.websters-online-dictionary.org/ (англо-английский)



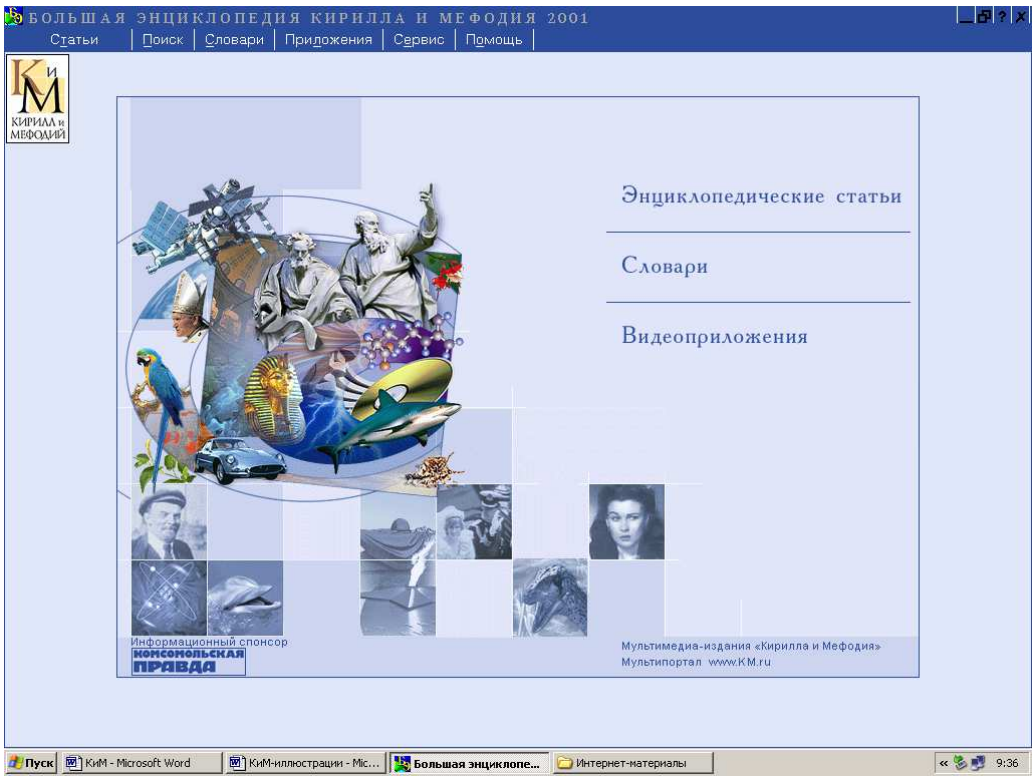
http://wordnet.princeton.edu/ (семантическая сеть английского языка)



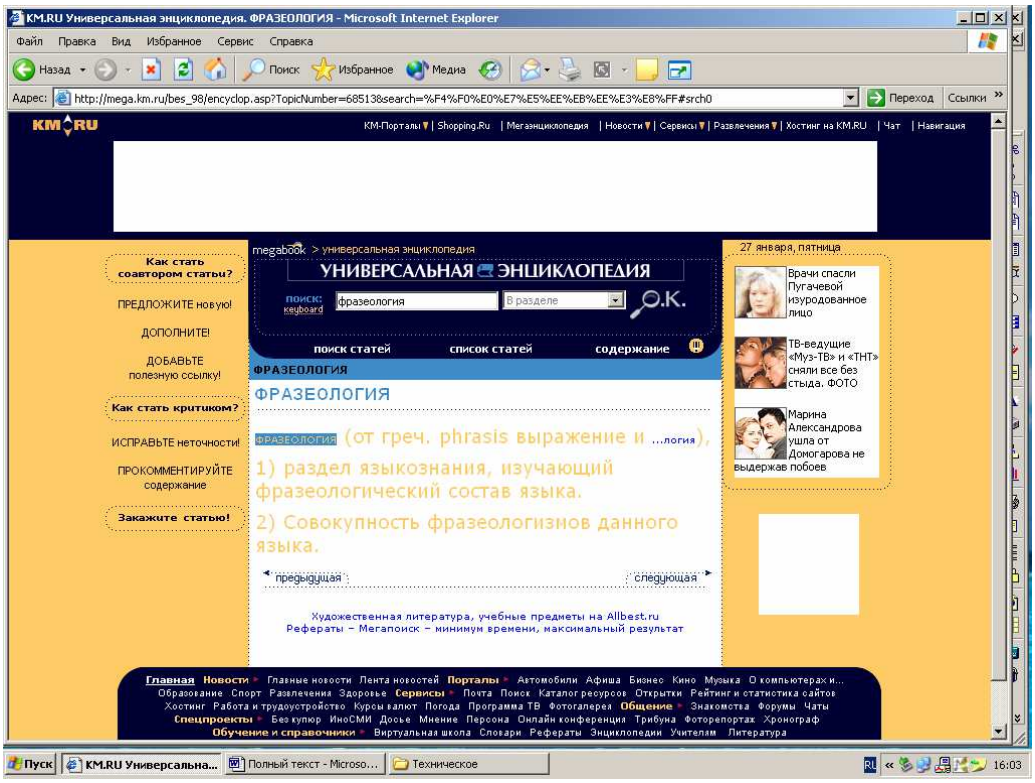
§ 9.8. Электронные энциклопедии

Энциклопедия – это научное или научно-популярное издание, содержащее систематизированный свод знаний. Энциклопедия – это веками отработанное средство информационной поддержки образования и самообразования. Современная электронная энциклопедия помимо фотографий содержит звукозаписи, музыкальное сопровождение и видеофрагменты.

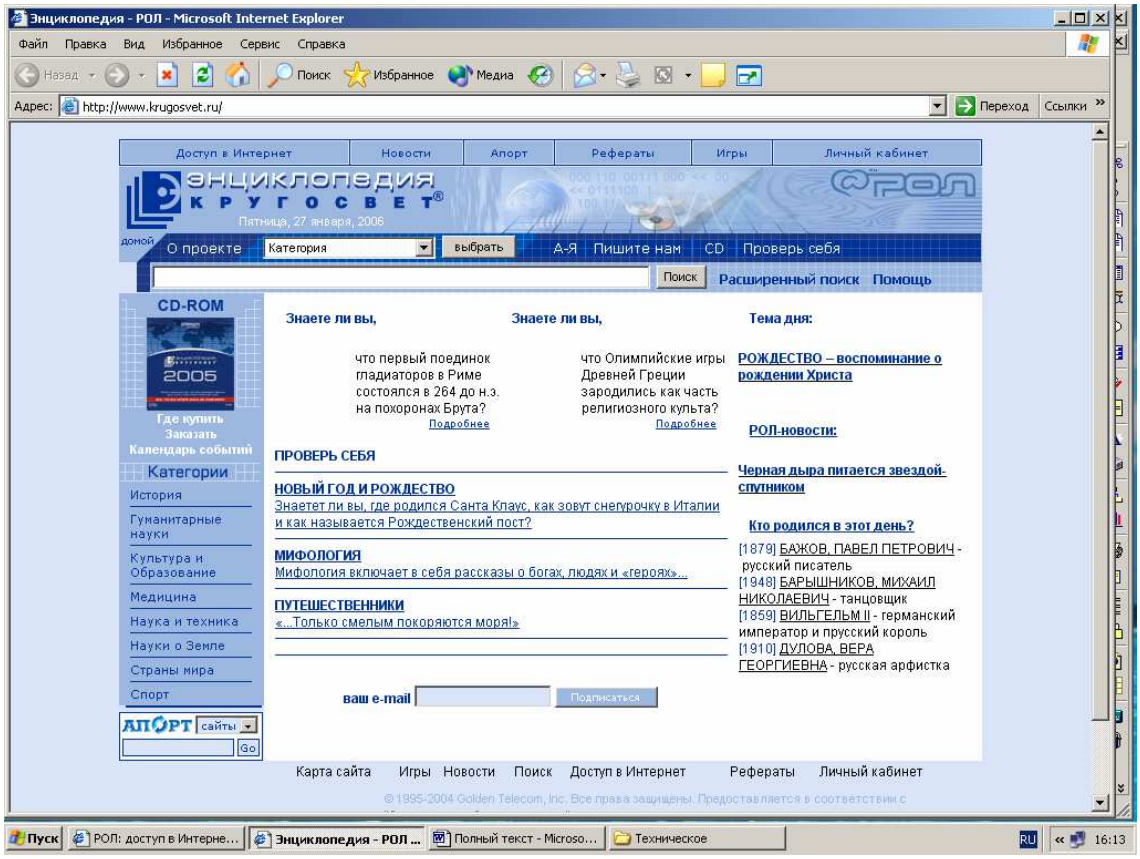
«***Большая энциклопедия Кирилла и Мефодия***» – современная универсальная российская энциклопедия – основана на Большом Энциклопедическом Словаре в дух томах издательства «Российская энциклопедия» (изд. 1996 г.). В ней 85000 энциклопедических статей (около 5 млн. слов), 7000 иллюстраций и более 20 больших приложений. В ней также 80 мин. видео (150 видеофрагментов), 129 мин. звука (222 звуковых фрагмента), более 200 географических карт.



БЭКМ – это универсальная энциклопедия, в ней содержатся сведения по всем областям науки, техники, литературе и искусству; вся важнейшая историческая, экономическая, географическая и социально-политическая информация по всем странам мира; все крупнейшие персоналии всех времен и народов; все значительные события общественной и культурной жизни России и мира. Онлайновый вариант энциклопедии по адресу *http: // mega.km.ru*



Энциклопедия «***Кругосвет***» является дополненным и исправленным изданием в переводе на русский язык «Энциклопедии Кольера» (*Collier's Encyclopedia)*, выходившей в США в 1952-1998 гг. Авторы проекта сообщают следующее: «Говорят, что энциклопедия – это свод вчерашних знаний, составленный сегодняшним днем для завтрашнего. Что ж, от «вчерашних знаний» никуда не уйдешь – это опыт, накопленный человечеством. В таком опыте есть и вечные истины, а их надо напоминать, иначе они вечно забываются. Так что «Кругосвет» никак не чурается «вчерашних знаний». Однако составители «Кругосвета» стремятся собирать и самые свежие, наиновейшие знания.

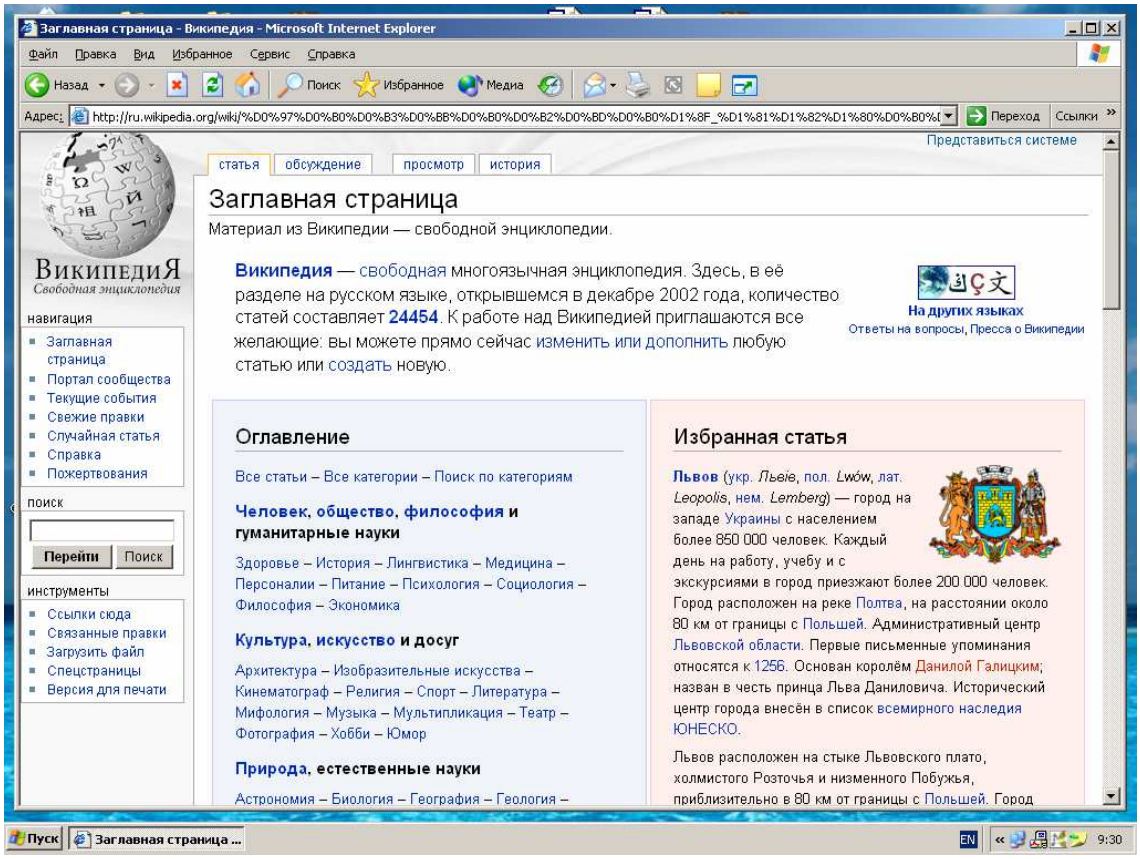


Содержание Энциклопедии распадается на несколько разделов: История, Гуманитарные науки, Культура и образование, Медицина, Наука и техника, Наука о земле, Страны мира, Спорт. Каждая категория в свою очередь дифференцируется. Например, Гуманитарные науки: Лингвистика, Психология и педагогика, Экономика и право, Философия.

Википе́дия (wikipedia.org) – общедоступная, свободно распространяемая многоязычная энциклопедия, издаваемая в Интернете.

Была создана 15 января 2001 года как проект по созданию англоязычной онлайн–энциклопедии, где любой посетитель может вносить изменения и дополнения. Цель проекта – создание полной, беспристрастной, свободной от авторских ограничений энциклопедии на всех языках Земли. Википедия приобрела популярность среди пользователей Сети. Позже появились разделы Википедии на других языках, включая русский.

Ежедневно энциклопедия пополняется в среднем 150-200 новыми статьями, а каждый час в содержание энциклопедии вносится около сотни правок. 30 тыс. статей – это объем толстого толкового словаря, но в отличие от книг онлайновая энциклопедия содержит куда больше информации в каждой статье, которые к тому же связаны между собой ссылками.



Задание: Проанализируйте любой словарь по следующему плану:

1. *Авторы, год выпуска, издательство, качество издания.*
2. *Объем словника.*
3. *Потенциальный потребитель.*
4. *Макроструктура словаря: наличие предисловия, какая информация включена, порядок расположения материала (алфавитный, обратный, гнездовой).*
5. *Микроструктура: структура словарной статьи, характер представленной информации, последовательность отражения информации, Как представлена парадигма слова? Есть ли этимологическая информация? Какие стили представлены? (разговорная, сниженная лексика и т.д.)*

Найдите как минимум 10 ресурсов с электронными словарями.

Тема 10: ПРИМЕНЕНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ В ПРЕПОДАВАНИИ ИНОСТРАННЫХ ЯЗЫКОВ

1. Применение информационных технологий в преподавании иностранных языков
2. Методы обучения с применением персонального компьютера
3. Способы использования персонального компьютера при обучении иностранным языкам
4. Содержание компьютерных программ индивидуализированного обучения иностранным языкам
5. Виды обучающих программ.
6. Дистанционное обучение, его особенности, применение информационных технологий в дистанционном обучении

§ 10.1. Применение информационных технологий в преподавании иностранных языков

Широкое использование компьютеров в различных аспектах деятельности человека не обошло стороной проблему обучения человека языкам. Уже сейчас создано большое число компьютерных программ обучения ИЯ. Сложность задачи обучения языкам объясняется тем, что любое обучение – задача комплексная, требующая учета данных психологии, педагогики, методики, особых свойств изучаемого предмета. По существу, каждая обучающая программа – это сложная система искусственного интеллекта.

За последние годы значительно изменился принцип применения компьютерных программ для обучения ИЯ. Если ранее утверждалось, что наиболее эффективно они могут быть использованы в рамках автоматизированных обучающих систем, то с появлением компьютеров в доме обучаемых такие программы все чаще используются индивидуально и подбираются в зависимости от цели обучения.

§ 10.2. Методы обучения с применением персонального компьютера

С точки зрения принципов восприятия информации при обучении с помощью компьютера выделяют, как правило, два теоретических подхода: бихевиористский и когнитивно-интеллектуальный.

Бихевиористский подход связан с постулатом «чем чаще употреблено слово, тем лучше оно запоминается» и основан на жесткой формуле «стимул – реакция». В рамках данного подхода различают следующие методы автоматизированного обучения:1) программирование учебной деятельности обучаемого; 2) тестирование; 3) информирование.

Первый из этих методов обучения характерен тем, что управляющие воздействия на обучаемого полностью определяются обучающей программой. В такой программе каждому обучаемому в зависимости от его

уровня знаний полностью задается последовательность учебных или контрольных заданий.

При тестировании компьютер по специальным программам выявляет индивидуальные профессиональные и психологические характеристики обучаемых и достигнутые ими уровни знаний. При этом обучаемый лишь отвечает на вопросы, но оценку за знания не получает. Этот метод достаточно часто используется при оценке различных аспектов знания иностранных языков (тестирование словарного запаса, способности к изучению иностранных языков и т.п.).

Суть метода информирования заключается в том, что в память компьютера помещаются некоторые справочно-информационные данные (грамматический справочник, орфографический словарь, двуязычный словарь и т.п.), которые обучаемый может использовать при подготовке к занятиям или непосредственно в процессе занятий.

К сожалению, бихевиористский подход не может преодолеть механистичность обучения и отсутствие развития когнитивных (мыслительных) способностей обучаемых, которым отводилась пассивная роль объекта, а не творческая роль субъекта обучения.

При *когнитивно-интеллектуальном* подходе у обучаемого активизируются познавательные функции. Для успешной реализации такого подхода в памяти компьютера создается универсальная учебная среда, включающая различные грамматические справочники, словари, спеллеры, другие вспомогательные материалы. При таком подходе в принципе возможны следующие методы автоматизированного обучения: 1) моделирование учебной среды; 2) свободное обучение.

Суть процесса моделирования учебной среды сводится, как и при бихевиористском подходе к тому, что обучаемый может выбирать учебные задания лишь из некоторого конечного множества заданий, включенного в универсальную учебную среду. Однако метод свободного обучения дает обучаемому возможность самому выбрать тематику обучения и способ работы с компьютером.

Так, например, при обучении ИЯ можно дать задание компьютеру выступить в роли продавца в магазине, где обучаемый, ведя диалог с продавцом, хочет купить некоторую вещь. При этом обучающая программа может корректировать действия обучаемого или после каждого его шага принятия решения, либо она может комментировать его действия после окончательного завершения всей программы.

§ 10.3. Способы использования персонального компьютера при обучении иностранным языкам

Если рассматривать современные персональные компьютеры не просто как средство технической поддержки учебного процесса, а как устройство, способное выполнять педагогические функции, несущее в себе конкретные

знания и передающее эти знания в процессе диалога с обучаемым, то можно выделить три способа использования компьютеров в обучении:

1) Компьютер – помощник преподавателя. В этом случае процесс обучения строится в соответствии с традиционным содержанием образования и методами передачи знаний от преподавателя к обучаемым. Используемые при этом обучающие программы лишь моделируют некоторые задачи, темы, разделы изучаемого курса и общаются с обучаемым по достаточно жесткому сценарию. Здесь преобладает групповой метод обучения в традиционных группах, классах и т. п.

2) Компьютер – преподаватель. При таком способе также моделируется традиционная методика обучения и строится жесткий сценарий обучения. Однако соответствующие обучающие программы направлены на обучение целому курсу (информатике, английскому языку и т.д.). Как правило, они предназначены для индивидуализированного обучения (чаще всего в домашних условиях).

3) Компьютер – источник знаний и «оцениватель» знаний обучаемого. Здесь используется так называемая альтернативная педагогика, когда обучаемый, исходя из целей обучения и своих возможностей, опираясь на собственный опыт и знания, обращается к компьютеру как к носителю необходимых для него знаний или «оценивателю» полученных обучаемым знаний. Такой подход возможен как при групповом, так и индивидуализированном обучении в рамках дистанционного обучения.

§ 10.4. Содержание компьютерных программ индивидуализированного обучения иностранным языкам

Компьютерные программы индивидуализированного обучения языкам обычно представляют собой некоторые законченные курсы, предназначенные для начального обучения, совершенствования языка и т.д. Среди большого числа требований, которым должны удовлетворять компьютерные программы индивидуализированного обучения, наиболее важными с педагогической точки зрения являются следующие:

1) совмещать в себе обучающую, контрольную и поисковую функции;

2) опираться на сценарии, приближенные к обычному традиционному обучению;

3) максимально использовать принцип наглядности и доступности, т.е. выводить на экран компьютера не только текст, но и звук, иллюстрации, видео и т.п.;

4) иметь средства быстрой и объективной оценки знаний обучаемых даже в тех случаях, когда ответ обучаемого далек от наиболее ожидаемого;

5) содержать возможность настройки на конкретного обучаемого (выбор способа подачи нового материала, типа упражнений, скорости ответа и т.п.).

Как правило, обучающие программы, используемые для индивидуализированного обучения, реализуются в виде так называемых мультимедийных обучающих программ. Слово мультимедийный появилось

вне связи с компьютерами в англо-русском словаре 1969 года издания. В то время урок, проводимый преподавателем, назывался мультимедийным, если в нем присутствовали и рассказ учителя, и магнитофонная запись, и кино, и слайды, и любые средства технического обучения. Сегодня под мультимедийной обучающей программой понимается компьютерная программа, использующая текст, звук, цвет, графику и движение.

В понятие звук входят речь, музыка, а также различные звуковые эффекты. Графика в таких программах может быть представлена различными рисунками, геометрическими фигурами, символами, фотографиями и сканированными изображениями. Движение в мультимедийных программах представляется в виде последовательности статических элементов (кадров) и может быть в виде анимации (последовательность рисованных изображений) и видео (последовательность черно-белых или цветных фотографий, пропускаемая на экране компьютера со скоростью около 24 фотографий в секунду). Такие программы дают возможность активизировать различные каналы восприятия информации и повышают степень запоминания и усвоение учебного материала.

В целом процесс создания мультимедийных обучающих программ включает следующие этапы:

1) разделение всего курса, который будет предложен для обучения, на определенное число тем и подтем;
2) отбор для каждой темы или подтемы определенного лексического и грамматического материала;
3) создание для каждой темы или подтемы набора сценариев, в рамках которых будут закрепляться лексический материал и грамматические правила;
4) подбор в соответствии со сценариями необходимых текстов, аудио- и видеоматериалов;
5) программирование сценариев.

Такая работа может выполняться высококвалифицированными программистами вместе с опытными преподавателями ИЯ, методистами, психологами, электронщиками. К числу наиболее известных фирм такого рода относятся американские фирмы: *Seracuse Language System, Compulink, Foreign Language Software Company*, российские фирмы: «Мультимедиа Технологии», *Istrasoft, New Media Generation* и др.

§ 10.5. Виды обучающих программ

По типу пользователей различают программы: 1) для детей; 2) для молодежи и взрослых; 3) для бизнес-применений; 4) специализированные программы.

По назначению: 1) для игр; 2) для начального обучения языку; 3) для совершенствования знаний языка; 4) для сдачи различных экзаменов; 5) для работы с деловыми текстами.

Так, для начального обучения английскому языку молодежи и взрослых можно использовать такие мультимедийные программы, как *Bridge to English*, Репетитор *English*, Профессор Хиггинс, *Learn to speak English, Everyday English in Communication, Talk to Me*.

Для совершенствования знания английского языка молодыми людьми и взрослыми полезны такие программы, как: *Complete English, English for Communication, English Cold, English Platinum*.

Совершенствование знаний английского языка взрослыми в области бизнеса возможно путем использования мультимедийных программ *Business English и EBC (English Business Contracts).*

Примером специализированных обучающих программ, ориентированных на молодежь и взрослых и предназначенных для сдачи международного теста на владение английским языком как иностранным (TOEFL), является программа *The Heinemann TOEFL*.

Для того чтобы представить возможности мультимедийной обучающей программы, рассмотрим программу *English Gold*. Она предназначена для совершенствования знаний английского языка молодежью и взрослыми. Программа включает пять разделов: «Фонетика», «Грамматика», «Словарь», «Диалоги», «Фильм» – и 144 урока. Комментарии и подсказки оформлены на русском языке. Словарь включает 12000 слов. Каждая словарная единица представлена в письменном и звуковом видах, а также в виде изображения предмета. Логический объем звука программы составляет более 100 часов. Программа содержит 2096 иллюстраций.

§ 10.6. Дистанционное обучение, его особенности, применение информационных технологий в дистанционном обучении

В настоящее время в России возрастает интерес к дистанционному обучению на базе новых информационных технологий.

В самом общем плане дистанционное обучение – это обучение на расстоянии, т. е. в ситуации, когда обучаемый отделен от преподавателя в пространстве или во времени. В сознании большинства преподавателей это понятие идентично заочному обучению.

Однако содержание этого термина постепенно меняется: дистанционное обучение – это новая форма организации учебного процесса, соединяющая в себе традиционные и новые информационные технологии обучения, основывающаяся на принципе самостоятельного получения знаний, предполагающая в основном телекоммуникационный принцип доставки обучаемому основного учебного материала и интерактивное взаимодействие обучаемых и преподавателей как непосредственно в процессе обучения, так и при оценке полученных ими в процессе обучения знаний и навыков. Дистанционное обучение (далее ДО) – основная составляющая дистанционного образования.

Существуют различные подходы к классификации моделей ДО. Наиболее оптимальной представляется классификация, зависящая от

способов доставки обучаемым учебного материала и принципов общения с преподавателем:

1) Интерактивное телевизионное обучение. В данном случае преподаватель ведет занятие в одном из классов, где установлена видеокамера, и весь урок по телевизионным кабелям передается в другое здание, другой город, другую область или другое государство. По телевизионным каналам обучаемым передается также учебный материал и задания. Проверка знаний с выдачей соответствующих дипломов и аттестатов осуществляется при личных контактах преподавателя со студентами. Такой способ дистанционного обучения широко используется в настоящее время в США. Наиболее известным образовательным учреждением в этом плане является Национальный технологический университет (*NTU – National Technological University*). Он является центром дистанционного образования штата Колорадо (г. Форт-Коллин) и объединяет 40 дистанционных школ, расположенных на его территории. Для такого обучения в США активно используется и Система публичного телевещания (*Public Broadcasting System, PBS–TV*). Учебные курсы по бизнесу, управлению, различным областям науки передаются по четырем образовательным каналам и доступны людям всей страны.

2) Дистанционное обучение с использованием носителей учебной информации на компакт-дисках (CD). Для получения различных консультаций у преподавателей широко используется глобальная сеть Интернет. Проверка знаний и навыков в этом случае, как и в первом виде дистанционного обучения, осуществляется при личных контактах обучаемых с преподавателями.

3) Дистанционное обучение с широким использованием телекоммуникационных сетей. В этом случае как передача знаний обучаемым, так и проверка их знаний и навыков осуществляются в интерактивном режиме работы с глобальной сетью Интернет. Занятия проводятся в следующих формах:

Чат-занятия – учебные занятия, осуществляемые с использованием чат-технологий. Чат-занятия проводятся синхронно, то есть все участники имеют одновременный доступ к чату. В рамках многих дистанционных учебных заведений действует чат-школа, в которой с помощью чат-кабинетов организуется деятельность дистанционных педагогов и учеников.

Веб-занятия – дистанционные уроки, конференции, семинары, деловые игры, лабораторные работы, практикумы и другие формы учебных занятий, проводимых с помощью средств телекоммуникаций и других возможностей Всемирной паутины. Для веб-занятий используются специализированные образовательные веб-форумы – форма работы пользователей по определённой теме или проблеме с помощью записей, оставляемых на одном из сайтов с установленной на нем соответствующей программой.

Телеконференции – проводятся, как правило, на основе списков рассылки с использованием электронной почты.

Использование технологий ДО позволяет:

1) снизить затраты на проведение обучения (не требуется затрат на аренду помещений, поездок к месту учебы, как учащихся, так и преподавателей и т. п.);

2) проводить обучение большого количества человек;

3) повысить качество обучения за счет применения современных средств, объемных электронных библиотек и т.д.

Несмотря на то, что в последние годы идея ДО нашло широкую поддержку во всем мире, в ее практической реализации имеется ряд существенных трудностей. На первый план здесь выходит самостоятельная познавательная деятельность обучаемого (учение, а не преподавание); кроме того, обучаемый должен обладать определенным уровнем владения компьютером.

В России у истоков ДО стояла Евгения Семёновна По́лат (1937-2007) – доктор педагогических наук, профессор, заведующая лабораторией дистанционного обучения Института содержания и методов обучения РАО (с 1996 г.)

В настоящее время в нашей стране, помимо университетов классического типа, сочетающих традиционные формы учебной работы с элементами ДО, возникают специальные вузы. Например, Современная гуманитарная академия. Она вошла в книгу рекордов Гиннеса по наибольшему числу филиалов. Костяк академии – 350 докторов наук, 5 членкоров, 400 кандидатов наук. Работают в ней и профессора из Кембриджа. Академия известна эффективными виртуальными семинарами. Информация дозирована и распределена по модулям. Каждый обучающийся движется по своей учебной траектории. Первокурсник сам формирует планы программ. Темп – индивидуальный и пролонгированный. Каждые пять дней – тестирование. В случае успеха – переход на другой уровень. СГА – единственный вуз в России, получивший официальное подтверждение Минобразования РФ о готовности к полнообъёмному дистанционному обучению, поскольку академия имеет спутниковое образовательное телевидение с собственный телепортом для круглосуточного вещания по двум каналам.

Задание: Разделитесь на две команды. Составьте план-урока по английскому языка, используя как можно больше средств современных информационных технологий.

Тема 11: МАШИННЫЙ ПЕРЕВОД

1. Перевод текстов: общие понятия
2. Виды перевода
3. Причины создания систем машинного перевода
4. Преимущества и недостатки машинного перевода
5. Совершенствование систем машинного перевода
6. Классификация систем машинного перевода. Рабочее место переводчика
7. Обзор некоторых системы машинного перевода

§ 11.1. Перевод текстов: общие понятия

Существуют различные определения понятия «перевод текстов». В качестве рабочего определения примем следующее: ***перевод*** есть вид человеческой языковой деятельности, в результате которой некоторый текст на одном языке ставится в соответствие тексту на другом языке, при этом обеспечивается их смысловая эквивалентность.

Слово «перевод» понимают двояко: 1) как сам процесс перехода от текста на одном языке к этому же тексту на другом языке, 2) как и результат этого перехода, т.е. тот текст, который получается в результате перевода.

Переводом текстов человек начал заниматься еще в античном мире – более 20 веков назад. Одним из первых «переводчиков» был Цицерон, древнеримский политический деятель, оратор и писатель. Он переводил произведения древних греков на латинский язык и считал, что переводить следует не слова, а мысли, не букву, а смысл, в соответствии с условиями и духом своего языка.

Однако такой взгляд на перевод не являлся общепринятым как в древние, так и в последующие средние века. Вред научной теории перевода принесли переводы древнееврейских религиозных текстов на другие языки. При этом любое отступление от оригинала рассматривалось как ересь. Такой подход привел к возникновению и развитию *пословного перевода*, буквализму, искажавшему смысл и стиль текста исходного языка.

В эпоху Возрождения (XIV–XVI вв.) появились шедевры мировой литературы: произведения Ф. Рабле, У. Шекспира, С. Сервантеса, Ф. Петрарки, Дж. Боккаччо и целого ряда других писателей и поэтов, что привело к резкому возрастанию количества переводов. В это время особенно отчетливо стали осознаваться трудности перевода художественных текстов. Как протест против пословного, буквального перевода начинает возникать *вольный перевод*, при котором на другом языке передается лишь общая идея текста исходного языка. Такой перевод не учитывает стилистические особенности автора исходного текста, реалии места и времени описываемых в нем событий и многие другие особенности переводимого произведения. В процессе вольного перевода тексты исходного языка порой искажались до неузнаваемости.

Идеи перевода интересовали И. В. Гёте, Н.В. Гоголя, А.С. Пушкина и других писателей, поэтов, переводчиков. Их труды постепенно помогли подойти к созданию истинно научных теорий перевода, утверждающих возможность хорошего перевода текста с любого языка на другой язык. Такие теории начали создаваться в 50–60-е годы XX века, они опираются на метод моделирования процесса перевода текстов человеком.

§ 11.2. Виды перевода

Насчитывается большое количество типов и видов переводов, существуют различные подходы к их классификации.

В зависимости от *переводного материала* различают:

1) перевод художественной литературы;

2) научно-технический перевод (перевод научно-технических, военных, юридических текстов и т.д.);

3) общественно-политический перевод (перевод газет, политических журналов и т.д.);

4) бытовой перевод (перевод текстов разговорно-бытового характера).

По *форме презентации* текста перевода и текста оригинала выделяют:

1) письменный перевод;

2) устный перевод.

На основании того, с какой *скоростью* осуществляется устный перевод, выделяют:

1) синхронный (перевод производится практически одновременно с произнесением текста на исходном языке, при этом максимальное отставание не должно превышать 10 секунд),

2) последовательный (переводчик прослушивает значительный фрагмент текста, фиксирует его в своей памяти или записывает с помощью переводческой скорописи, а потом переводит его для слушателей).

По признаку *основной прагматической функции* перевода (с какой целью делается перевод) различают:

1) практический перевод (в целях получения новой технической, экономической, политической, эстетической и другой информации);

2) учебный перевод (для обучения основам перевода);

3) экспериментальный перевод (например, для оценки умения переводчика и качества работы);

4) адаптивный (направлен на то, чтобы приспособить текст к потребностям пользователя; чаще всего это достигается путем сокращения текста, сжатия информации, отсюда его второе название – реферативный перевод; такой перевод обычно используется в сфере технической документации);

4) эталонный перевод (как образцовый перевод, с которым сравниваются другие переводы).

По *степени механизации* процесса перевода выделяют:

1) традиционный («ручной») перевод, выполняемый человеком;

2) перевод, выполняемый человеком с помощью компьютера (когда компьютер по запросу пользователя ищет переводные эквиваленты иностранных слов);

3) перевод, выполняемый компьютером с помощью человека (например, когда компьютер по специальной программе делает перевод, а за справками – неизвестным переводным эквивалентом, неизвестной синтаксической структурой и т. д. – обращается к человеку-интерредактору);

4) машинный или автоматический перевод (выполняется компьютером без вмешательства человека).

В процессе перевода, как устного, так и письменного, переводчик может столкнуться с целым рядом трудностей, которые объясняются как различиями между языками, так и некоторыми экстралингвистическими явлениями.

Среди лингвистических проблем принято выделять семантические трудности (перевод лакун, идиом, метафор), синтаксические трудности (порядок слов, перевод личных местоимений, отрицательных конструкций и т.п.). В качестве примера экстралингвистических трудностей можно назвать культурные различия.

§ 11.3. Причины создания систем машинного перевода

Начало работ по машинному переводу (МП) или автоматическому переводу (АП) относят к 50-м гг. XX в. Идея МП обязана своим происхождением чисто практическим нуждам. В начале 50-х гг. происходит информационный взрыв – существенно возрастают объемы научно-технической информации. Дополнительный импульс исследованиям в области МП дала «холодная война»: противостоящие общественно-политические системы внимательно следили за развитием научно-технического потенциала друг друга. Именно по этой причине многие первые зарубежные системы МП работают с русским языком.

Формальная дата начала эры машинного перевода – 1949 г. В этом году известный американский специалист по дешифровке Уоррен Уивер составил меморандум, в котором теоретически обосновал принципиальную возможность создания систем МП. Меморандум был разослан двумстам специалистам в области лингвистики, дешифровки и теории программирования. С этого времени в США появляются коллективы разработчиков МП (в Массачусетском технологическом институте, в Калифорнийском университете, в Национальном бюро стандартов в Лос-Анджелесе, в Техасском университете).

Наконец, в 1954 г. проводится известный Джорджтаунский эксперимент, в процессе которого осуществляется перевод с русского языка на английский текста по физике. Хотя программа работала со словарем всего лишь в 250 слов, успех этого эксперимента стимулировал дальнейшие исследования в области МП.

В СССР первый эксперимент по МП прошел в 1955 г.: был осуществлен перевод на русский язык текстов по прикладной математике.

К этому времени относится начало работ по МП в Институте прикладной математики АН СССР под руководством О. С. Кулагиной и И. А. Мельчука. Коллектив разработчиков создал экспериментальные системы МП – с французского языка на русский и с английского на русский.

В 1959 г. открывается Лаборатория машинного перевода в МГПИ-ИЯ им. М.Тореза (ныне Московский государственный лингвистический университет).

В 1974 г. в Москве создается Всесоюзный центр переводов, в котором несколько научных коллективов работали над системами МП – АМПАР (англо-русский перевод), НЕРПА (немецко-русский перевод), ФРАП (французско-русский перевод).

§ 11.4. Преимущества и недостатки машинного перевода

В настоящее время Европейское экономическое сообщество имеет свою службу перевода, включающую около 2 тыс. переводчиков. Они переводят в год примерно 600 тыс. страниц текстов с пяти языков и не справляются со все возрастающими потоками заказов на переводы. Это приводит к тому, что до специалистов различных стран зарубежная информация доходит с большим опозданием (порой через 5–10 лет). Единственным способом увеличения скорости перевода является использование в переводческой деятельности современных компьютеров, которые в миллиарды раз быстрее человека могут выполнять необходимые для перевода логические действия.

Человек-переводчик тратит 20 % своего времени на перевод, 40 % – на поиск по словарю незнакомых слов и 40 % – на перепечатку и оформление перевода. Компьютер же в процессе перевода тратит 95 % времени непосредственно на перевод и 5 % на пополнение словаря.

Если максимальная производительность труда переводчика составляет 4–5 авторских листов в месяц, то такая, например, система машинного перевода, как SYSTRAN, переводит, в час до 1 млн. словоупотреблений (около 120 авторских листов).

Эти количественные характеристики свидетельствуют о преимуществе компьютерных систем. Однако качество такого перевода значительно уступает переводу, сделанному человеком.

И тем не менее проблемами машинного перевода сейчас активно занимаются во всех развитых странах: США, Франции, Японии, Германии, Китае и т.д. Ежегодно по машинному переводу проводится несколько крупных международных конференций, в разных странах постоянно издаются журналы и книги по этой проблеме.

§ 11.5. Совершенствование систем машинного перевода

Разработка систем МП прошла несколько этапов. Первые модели машинного перевода базировались на принципе перекодирования текста на

одном языке в текст на другом: грамматика в традиционном понимании в них отсутствовала полностью.

Изначально разработчики систем машинного перевода планировали разработать универсальные системы перевода, не ограничивать сферу их применения какой-либо проблемной сферой или тематикой. Однако довольно быстро обнаружилась неприменимость машинного перевода, например, к художественным текстам. Кроме того, одна система машинного перевода не могла переводить тексты из различных отраслей науки, т.е. пришлось ограничивать сферу применения созданных систем.

Причины, по которым пришлось это сделать, можно условно разделить на лингвистические и экстралингвистические.

К *лингвистическим* причинам относят недостаточность знаний о том, как на самом деле функционирует язык и как происходит процесс перевода в голове человека. К *экстралингвистическим* причинам относят тот факт, что процесс понимания, который является важнейшей составляющей механизма перевода, не обеспечивается исключительно лингвистической информацией. Мы понимаем текст или сообщение, опираясь при этом на ситуацию общения, или контекст, на накопленный опыт. Иными словами, человек при переводе опирается на имеющуюся у него картину мира. Именно это помогает человеку-переводчику воздержаться от абсурдного перевода. Компьютер этого сделать пока не может.

Разработка систем машинного перевода прошла несколько этапов, поэтому говорят о существовании систем нескольких поколений.

К системам машинного перевода первого поколения относятся простейшие модели, использующие стратегию прямого перевода, в которых выбор эквивалентов не производится. На выходе выдаются все переводные эквиваленты, имеющиеся в словаре. Преобразования текста в таких системах сводятся к последовательной замене слов и словосочетаний исходного текста на словарные эквиваленты текста перевода. Возможности таких систем определялись доступными размерами словарей, прямо зависящими от объема памяти компьютера. Важной особенностью ранних систем было то, что в них не производилось различий между пониманием (т.е. анализом) и порождением текста (т.е. синтезом). Качество перевода было очень низким.

Более поздние разработки привели к созданию систем с «трансфером». Учеными была высказана идея о том, что создать запись для прямой связи текстов разных языков невозможно. Эта связь может быть установлена только путем последовательных преобразований на различных языковых уровнях: лексическом, морфологическом, синтаксическом. *Трансфер* – это такой этап межъязыковых операций, на котором производится анализ синтаксиса и семантики входного текста и уточняется его структура.

Еще одна стратегия перевода – перевод через язык-посредник. Преимущество данной стратегии состояло в том, что одна система создавалась не для одной пары языков, а для нескольких языков, т.к. язык-посредник мыслился как универсальный код, с помощью которого можно

единым образом выразить грамматическую и семантическую информацию, содержащуюся в тексте на любом языке. Язык-посредник можно описать как формальный язык, понятный компьютеру и записанный в виде символов, не имеющих воплощения в звуках. Однако и эти системы не обеспечили адекватность перевода, поскольку с введением языка-посредника неизбежно терялись существенные характеристики переводимых текстов, например, коммуникативные и прагматические установки автора текста, актуальное членение и т.д. Кроме того, усложненность этапов преобразований вместе с низкой скоростью действия компьютеров того времени также понижали эффективность работы систем.

Эффективность работы системы автоматического перевода может быть повышена за счет сужения проблемной области. Системы машинного перевода ориентируются на работу с текстами определенной сферы. Качество перевода улучшается благодаря тому, что язык, обслуживающий конкретную сферу деятельности, легче поддается формализации. Иногда компании, чья продукция экспортируется, и, соответственно, требуется перевод большого объема технической документации, идут на то, чтобы с помощью лингвистов были разработаны строгие стандарты для внутреннего языка, чтобы упростить процедуру машинного перевода. Н-р, в Европейской ассоциации авиапроизводителей разработан такой стандарт, поскольку документация к самолетам достигает более 100 тыс. страниц.

§ 11.6. Классификация систем машинного перевода

В 1990 г. Ларри Чайлдс, специалист по машинному переводу, предложил следующую классификацию современных систем машинного перевода:

1) *МАНТ (Machine-assisted Human Translation)* – перевод, осуществляемый человеком с использованием компьютера;

2) *HAMT (Human-assisted Machine Translation)* – машинный перевод при участии человека;

3) *FAMT (Fully-automated Machine Translation)* – полностью автоматизированный машинный перевод;

При первом способе текст переводит человек, а за переводными эквивалентами слов, которые ему неизвестны, он обращается к автоматическим (электронным) словарям. Данный способ перевода используется сейчас достаточно широко. Для этого разработано много больших электронных словарей, например *Lingvo*, «Контекст», «Мультитран».

При втором способе человек на определенных этапах подключается к процессу перевода текста компьютером. Возможны три варианта подключения человека к такому процессу перевода:

1) предварительная подготовка текста;

2) определенные преобразования текста непосредственно в процессе перевода;

3) редакторская правка текста после перевода.

При *предварительной подготовке* текста перед его вводом в компьютер человек проводит целый ряд операций, например: сложные предложения сводит к нескольким простым, в тексте специальным образом отмечает фразеологические и идиоматические выражения, фиксирует одно переводное значение для многозначных слов ИЯ и т. п. Лингвист, выполняющий такую работу, называется *предредактором*.

Определенные преобразования текста непосредственно в процессе перевода осуществляет *интерредактор***.** Он реагирует на поступающие запросы компьютера: упрощает структуру выведенного на экран неясного для системы МП предложения исходного языка, указывает синтаксическую структуру этого предложения, отмечает не переведенные системой терминологические или фразеологические сочетания, указывает однозначный перевод многозначного слова и т.п.

Полную редакторскую правку переведенного автоматической системой текста осуществляет опытный переводчик – *постредактор.*

Компьютерные системы, соединяющие работу человека и машины, иногда называют автоматизированным рабочим местом переводчика.

При третьем способе для полностью *автоматического перевода* текстов с одного языка на другой создана серия программ. Наиболее известными из них являются *Stylus, Socrat, Retrans, Ertrans, Multitran, PROMT* и др.

§ 11.7. Обзор некоторых систем машинного перевода

Количество реально работающих и проектируемых систем МП к настоящему времени перевалило за сотню. Рассмотрим некоторые из них.

Система GAT (Georgetown Automatic Translation) – одна из первых систем МП, разрабатывавшаяся с 1952 г. в Джорджтаунском университете США. Проблемная область – перевод русскоязычных текстов по физике на английский язык. Стратегия создания – прямой перевод, сопровождавшийся некоторыми синтаксическими перестановками, приближавшими русский порядок слов к порядку слов английской фразы. Программа не имела под собой никакой серьезной лингвистической базы, эксплуатировалась до 1976 г.

Системы CETA и GETA русско-французского машинного перевода – разрабатывались во Франции в Гренобльском университете с 1961 по 1971 гг. Стратегия построения – использование языка-посредника. Опыт разработки оказался не вполне удачным, поскольку сконструированный язык-посредник приводил к потере релевантной информации.

Система TAUM предназначена для перевода английских текстов на французский язык, разрабатывалась в Монреальском университете с 1965 г. Строилась как система с трансфером. Изначально проект не имел направленности на конкретную проблемную область. Позднее система была переориентирована на перевод прогнозов погоды и инструкций по эксплуатации авиационной техники.

Системы семейства ЭТАП (ЭлектроТехнический Автоматический перевод) (французско-русский и английско-русский перевод) – предназначены для перевода связных текстов и заголовков патентов. Относятся к системам МП с трансфером.

Переводческий комплекс АНРАП состоит из двух больших систем – АМПАР (англо-русский перевод) и НЕРПА (немецко-русский перевод). Предназначен для использования в крупных информационных службах и переводческих организациях для перевода текстов различных тематических областей. Для обеспечения тематической привязки предусматривается возможность подключения дополнительных терминологических словарей, описывающих конкретные тематические сферы. Скорость перевода довольно высока (3-5 авторских листов в час), что является необходимым условием функционирования промышленных систем МП, однако качество перевода невысоко. Постредактирование переводов оказывается необходимым.

Система CULT (Chinese University Language Translator) представляет собой типичный пример системы человеко-машинного перевода. Разработка системы, предназначенной для перевода китайских математических и физических текстов на английский язык, началась в Китайском университете Гонконга в 1968 г. Работа программы CULP требует активного участия человека не только на этапе предредактирования, но и в процессе самого перевода.

Системы семейства ALPS – типичный пример компьютерного инструментария, образующего рабочее место переводчика. Системы фирмы ALPS позволяют проводить экранное редактирование текста в многооконном текстовом редакторе, осуществлять оперативный поиск слова в словарных базах данных, переносить информацию из баз данных в текстовый файл, а также делать пословный перевод текста, опирающийся на введенные в систему словарные источники. В настоящее время системы поддержки перевода, распространяемые на рынке фирмой ALPS, обеспечивают перевод на английский, немецкий, французский, португальский и испанский языки.

Задание: Сделайте сообщение об одной из систем машинного перевода.

Тема 12: ПРЕДСТАВЛЕНИЕ РЕЗУЛЬТАТОВ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

1. Представление информации в виде диаграмм, гистограмм, таблиц.
2. Создание презентаций в среде PowerPoint.

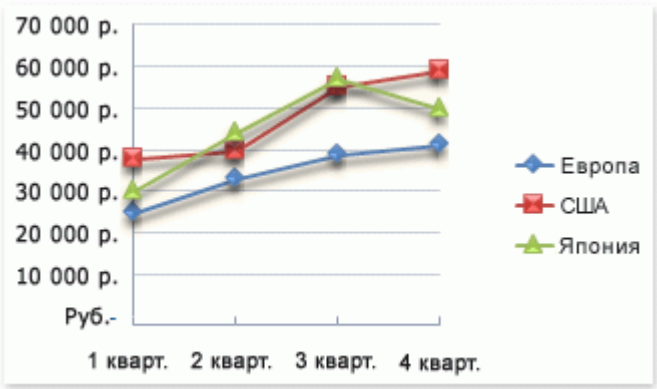
§ 12.1. Представление информации в виде диаграмм, гистограмм, таблиц

Диаграмма (греч. Διάγραμμα (diagramma) – изображение, рисунок, чертёж) – графическое представление данных, позволяющее быстро оценить соотношение нескольких величин. Представляет собой геометрическое символьное изображение информации с применением различных приёмов техники визуализации.

Иногда для оформления диаграмм используется трёхмерная визуализация, спроецированная на плоскость, что придаёт диаграмме отличительные черты или позволяет иметь общее представление об области, в которой она применяется.

Диаграммы в основном состоят из геометрических объектов (точек, линий, фигур различной формы и цвета) и вспомогательных элементов (осей координат, условных обозначений, заголовков и т. п.). Также диаграммы делятся на плоскостные (двумерные) и пространственные (трёхмерные или объёмные).

Диаграммы-линии или графики – это тип диаграмм, на которых полученные данные изображаются в виде точек, соединённых прямыми линиями. Точки могут быть как видимыми, так и невидимыми (ломаные линии). Также могут изображаться точки без линий (точечные диаграммы). Для построения диаграмм-линий применяют прямоугольную систему координат. Обычно по оси абсцисс откладывается время (годы, месяцы и т. д.), а по оси ординат – размеры изображаемых явлений или процессов. На осях наносят масштабы.



Диаграммы-линии целесообразно применять тогда, когда число размеров (уровней) в ряду велико. Кроме того, такие диаграммы удобно использовать, если требуется изобразить характер или общую тенденцию развития явления или явлений. Линии удобны и при изображении нескольких динамических рядов для их сравнения, когда требуется сравнение темпов роста. На одной

диаграмме такого типа не рекомендуется помещать более трёх-четырёх кривых. Их большое количество может усложнить чертёж, и линейная диаграмма может потерять наглядность.

Столбчатые и линейные диаграммы (гистограммы)

Классическими диаграммами являются столбчатые и линейные (полосовые) диаграммы. Также они называются гистограммами. Столбчатые диаграммы в основном используются для наглядного сравнения полученных статистических данных или для анализа их изменения за определённый промежуток времени. Построение столбчатой диаграммы заключается в изображении статистических данных в виде вертикальных прямоугольников или трёхмерных прямоугольных столбиков. Каждый столбик изображает величину уровня данного статистического ряда. Все сравниваемые показатели выражены одной единицей измерения, поэтому удаётся сравнить статистические показатели данного процесса.

Разновидностями столбчатых диаграмм являются линейные (полосовые) диаграммы. Они отличаются горизонтальным расположением столбиков. Столбчатые и линейные диаграммы взаимозаменяемы, рассматриваемые в них статистические показатели могут быть представлены как вертикальными, так и горизонтальными столбиками.

Круговые (секторные) диаграммы

Достаточно распространённым способом графического изображения структуры статистических совокупностей является секторная диаграмма, так как идея целого очень наглядно выражается кругом, который представляет всю совокупность. Относительная величина каждого значения изображается в виде сектора круга, площадь которого соответствует вкладу этого значения в сумму значений. Этот вид графиков удобно использовать, когда нужно показать долю каждой величины в общем объёме. Сектора могут изображаться как в общем круге, так и отдельно, расположенными на небольшом удалении друг от друга.

Круговая диаграмма сохраняет наглядность только в том случае, если количество частей совокупности диаграммы небольшое. Если частей диаграммы слишком много, её применение неэффективно по причине несущественного различия сравниваемых структур.

Радиальные (сетчатые) диаграммы

В отличие от линейных диаграмм, в радиальных или сетчатых диаграммах более двух осей. По каждой из них производится отсчёт от начала координат, находящегося в центре. Для каждого типа полученных значений создаётся своя собственная ось, которая исходит из центра диаграммы. Радиальные диаграммы напоминают сетку или паутину, поэтому иногда их называют сетчатыми. Преимущество радиальных диаграмм в том, что они позволяют отображать одновременно несколько независимых величин, которые характеризуют общее состояние структуры статистических совокупностей.

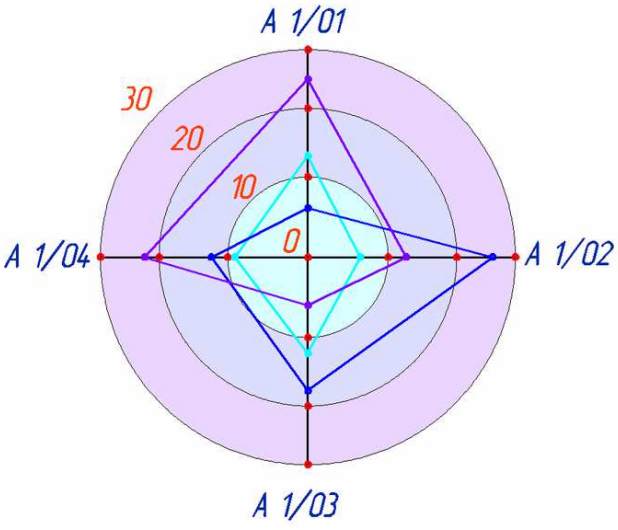


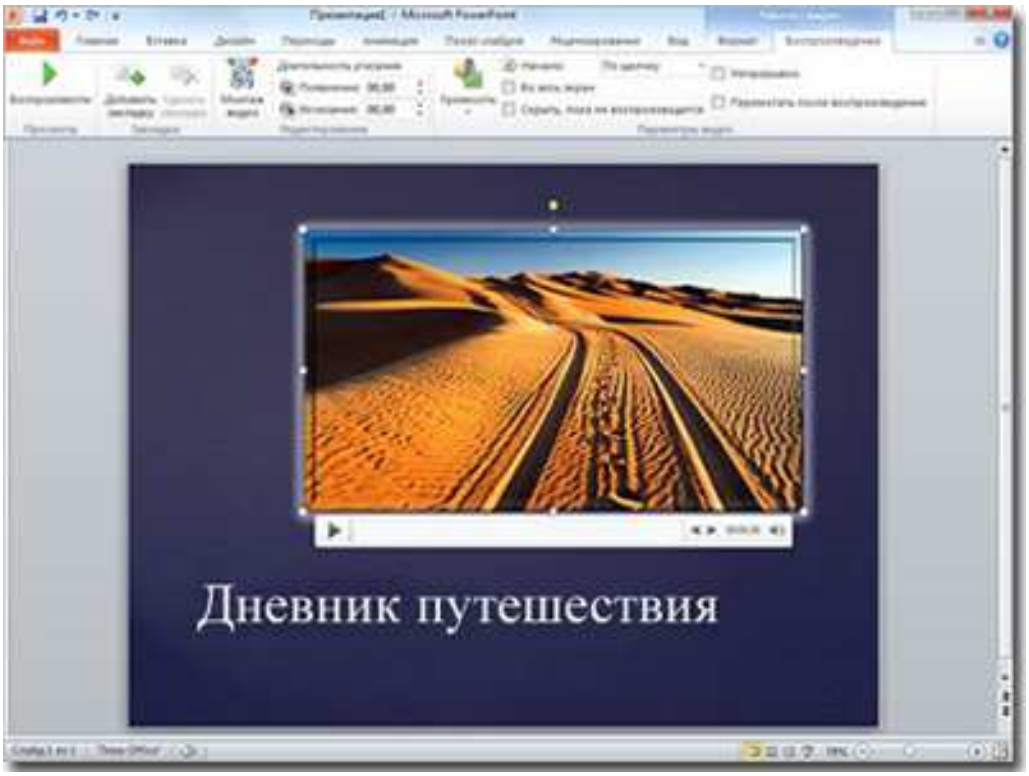
Таблица (из лат. tabula «доска») – способ передачи содержания, заключающийся в организации структуры данных, в которой отдельные элементы помещены в ячейки, каждой из которых сопоставлена пара значений – номер строки и номер колонки. Таким образом, устанавливается смысловая связь между элементами, принадлежащими одному столбцу или одной строке.

§ 12.2. Создание презентаций в среде PowerPoint

Microsoft PowerPoint – программа для создания и проведения презентаций, являющаяся частью Microsoft Office и доступная в редакциях для операционных систем Microsoft Windows и Mac OS.

Идея PowerPoint появилась у Боба Гаскинса, студента университета Беркли, который решил, что наступает век графических интерактивных материалов. В 1984 году Гаскинс нанял разработчика Денниса Остина и, объединив свои усилия, они создали программу Presenter. Позже было решено сменить имя на PowerPoint, которое и стало названием конечного продукта.

В 1987 году вышел PowerPoint 1.0 для Apple Macintosh. Он работал в чёрно-белом цвете. Вскоре появились цветные Macintosh, и новая версия PowerPoint не заставила себя ждать. В 1990 году вышла версия для Windows, и с этого года PowerPoint стал стандартом в наборе программ Microsoft Office.



Задания:

1) Постройте все указанные типы диаграмм в программе Microsoft Excel.

2) Сделайте презентацию на одну из указанных в списке тем:

1. *Интернет-ресурсы в помощь преподавателю иностранного языка (использование ПК в обучении лексике, грамматике, фонетике и другим аспектам языка – на выбор).*
2. *Использование информационных технологий в дистанционном обучении.*
3. *Компьютерная лексикография (электронные словари в Интернете).*
4. *Корпусная лингвистика (корпуса языков, параллельные корпуса).*
5. *Электронные библиотеки: ресурсы и возможности.*

6. *Автоматическая обработка текстов.*
7. *Автоматическая обработка звучащей речи.*
8. *Информационно-поисковые системы: история, преимущества, недостатки.*
9. *Лингвистические информационные ресурсы: настоящее и будущее.*
10. *Квантитативная лингвистика: сферы применения.*
11. *Системы машинного перевода (причины создания, обзор, преимущества и недостатки).*

Литература

1. Баранов А. Н. Введение в прикладную лингвистику / А.Н. Баранов; Моск. гос. ун-т им. М.В. Ломоносова. Фил. фак. – М.: Эдиториал УРСС, 2001. – 358 с.
2. Бовтенко М.А. Компьютерная лингводидактика: учебное пособие / М.А. Бовтенко. – М.: Флинта: Наука, 2005. – 215 с.
3. Всеволодова А.В. Компьютерная обработка лингвистических данных: учебное пособие: для студентов, аспирантов, преподавателей-филологов / А.В. Всеволодова. – 2-е изд., испр. – М. : Флинта : Наука, 2007. – 92 с.
4. Зубов А. В. Информационные технологии в лингвистике / А.В. Зубов, И.И. Зубова. – М.: Academia, 2004. – 205 с.
5. Калмыков А.А., Коханова Л.А. Интернет-журналистика. – М.: ЮНИТИ-ДАНА, 2005. – 383 с.
6. Леонтьева Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы: учебное пособие для студентов лингвистических факультетов вузов / Н.Н. Леонтьева. – М.: Академия, 2006. – 302 с.
7. Марчук Ю. Н. Компьютерная лингвистика: учебное пособие для студентов вузов, специализирующихся по направлению и специальности «Филология» / Ю.Н. Марчук. – М.: Восток–Запад, 2007. – 317 с.
8. Потапова Р. К. Новые информационные технологии и лингвистика: учебное пособие для студ. вузов / Р.К. Потапова; Моск. гос. лингв. ун-т.– Изд. 2-е. – М.: Едиториал УРСС, 2004. – 317 с.
9. Хроленко А.Т. Современные информационные технологии для гуманитария: практическое руководство / А.Т. Хроленко, А.В. Денисов. – М.: Флинта: Наука, 2007. – 127 с.
10. Цатурова И.А. Компьютерные технологии в обучении иностранным языкам: учебно-методическое пособие / И.А. Цатурова, А.А. Петухова. – М.: Высш. шк., 2004. – 94 с.
11. Шилихина К.М. Основы прикладной лингвистики: учебное пособие по специальности 021800 (031301) – Теоретическая и прикладная лингвистика / К.М. Шилихина; Воронеж. гос. ун-т. – Воронеж: ЛОП ВГУ, 2006. – 51 с.