Санкт-Петербургский государственный
университет

Филологический факультет
Кафедра математической лингвистики

В.П. Захаров

КОРПУСНАЯ ЛИНГВИСТИКА

Учебно-методическое пособие

Санкт-Петербург
2005

ББК 81.1
З-38

Р е ц е н з е н т ы :
докт. филол. наук *Л.Н. Беляева* (Рос. гос. пед. ун-т им. А.И.Герцена)
канд. фил. наук *С.А. Коваль* (С.-Петерб. гос. ун-т)

Печатается по постановлению
Редакционно-издательского совета
С.-Петербургского государственного университета

Захаров В.П.
З-38 Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005. – 48 с.

Предлагаемое пособие содержит описание предмета и основного содержания корпусной лингвистики – нового направления в лингвистике. Оно включает также программу учебной дисциплины «Корпусная лингвистика», которая изучается студентами отделения структурной и прикладной лингвистики Санкт-Петербургского государственного университета. Пособие базируется на исследовательской и преподавательской деятельности автора.

Для студентов и аспирантов, специализирующихся в области прикладной лингвистики и автоматизированных систем обработки текста.

ББК 81.1

© В.П. Захаров, 2005
© Санкт-Петербургский государственный университет, 2005

1. Основные понятия

1.1. Введение: корпусы и корпусная лингвистика

Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий. Под названием ***лингвистический, или языковой, корпус текстов*** понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. В понятие «корпус текстов» входит также система управления текстовыми и лингвистическими данными, которую в последнее время чаще всего называют *корпусным менеджером* (или корпус-менеджером) (англ. corpus manager). Это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме.

Целесообразность создания и смысл использования корпусов определяется следующими предпосылками:

1) достаточно большой (репрезентативный) объем корпуса гарантирует типичность данных и обеспечивает полноту представления всего спектра языковых явлений;

2) данные разного типа находятся в корпусе в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения;

3) однажды созданный и подготовленный массив данных может использоваться многократно, многими исследователями и в различных целях.

Можно сказать, что все современные лингвистические исследования и работы по составлению словарей и грамматик так или иначе ориентированы на использование представительных корпусов текстов. Развитие современных интеллектуальных программных систем, предназначенных для обработки текстов на естественном языке, также требует

большой экспериментальной лингвистической базы. Спрос на корпусные данные совпал с появлением соответствующих технических возможностей.

Первые лингвистические корпусы текстов появились в 60-е гг. прошлого столетия. В 1963 г. в Брауновском университете (США) впервые был создан большой корпус текстов на машинном носителе (Brown Corpus). Авторы корпуса У. Френсис (W. Francis) и Г. Кучера (H. Kucera) спроектировали его как набор из пятисот двухтысячесловных прозаических печатных текстов американского варианта английского языка. Тексты принадлежали пятнадцати наиболее массовым жанрам англоязычной печатной прозы США и были напечатаны в 1961 г. Корпус сопровождался большим количеством материалов его первичной статистической обработки — частотный и алфавитно-частотный словарь, разнообразные статистические распределения. Появление Брауновского корпуса вызвало всеобщий интерес и оживленные дискуссии. Прежде всего они коснулись принципов отбора текстов и состава потенциально решаемых на таком корпусе задач. Затем последовали Ланкастерский корпус английского языка (Lancaster-Oslo-Bergen Corpus, LOB), Уппсальский корпус русского языка. Среди современных корпусов английского языка наиболее известны Британский национальный корпус (British National Corpus), Международный корпус английского языка (International Corpus of English), лингвистический Банк английского языка (Bank of English) и др. В настоящее время корпусы созданы для многих языков мира (см. Приложение 1). Ведется работа и над созданием Национального корпуса русского языка.

В первой половине 90-х гг. корпусная лингвистика окончательно сформировалась как отдельный раздел науки о языке. При этом она тесно взаимодействует с компьютерной лингвистикой, используя ее достижения и в свою очередь обогащая ее.

Поиск в корпусе данных позволяет по любому слову построить конкорданс – список всех употреблений данного слова в контексте со ссылками на источник. Корпусы могут использоваться для получения разнообразных справок и статистических данных о языковых и речевых единицах. В частности, на основе корпусов можно получить данные о частоте словоформ, лексем, грамматических категорий, проследить изменение частот и контекстов в различные периоды времени, получить данные о совместной встречаемости лексических единиц и т.д. Представительный массив языковых данных за определенный период позволяет изучать динамику процессов изменения лексического состава языка, проводить анализ лексико-грамматических характеристик в разных жанрах и у разных авторов, и т.д. Корпусы призваны служить также ис-

точником и инструментом многоаспектных лексикографических работ по подготовке разнообразных исторических и современных словарей. Данные корпусов могут быть использованы для построения и уточнения грамматик и в целях обучения языку.

Можно сказать, что корпусная лингвистика имеет своим предметом теоретические основы и практические механизмы создания и использования представительных массивов языковых данных, предназначенных для лингвистических исследований в интересах широкого круга пользователей.

1.2. Репрезентативность

Задача создателей корпуса – собрать как можно большее количество текстов, относящихся к тому подмножеству языка, для изучения которого корпус создается. Но главное не только и не столько в количестве языкового материала, сколько в его пропорциональности. Можно сказать, что корпус – это уменьшенная модель языка или подъязыка. Важнейшее понятие корпусной лингвистики – репрезентативность. Под *репрезентативностью* понимается необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т.п. Имеются разные подходы к определению репрезентативности, можно сказать, что применительно к общеязыковому (национальному) корпусу это понятие невозможно рассчитать и описать строго математически, однако к этому можно и нужно стремиться, как на этапе проектирования корпуса, так и на этапе его эксплуатации.

1.3. Размер корпуса

Термин «корпус» обычно обозначает собрание текстов конечного фиксированного размера. С течением времени объем и состав корпуса может меняться, однако эти изменения должны или не менять его репрезентативность, или менять обоснованно. Объем первых корпусов составлял 1 млн словоупотреблений (Брауновский корпус, Уппсальский корпус русского языка). В настоящее время считается, что объем общеязыкового корпуса должен быть не меньше 100 млн словоупотреблений.

1.4. Разметка

Для решения различных лингвистических задач мало лишь наличия массива текстов. Требуется также, чтобы тексты содержали в себе явным образом разного рода дополнительную лингвистическую и экстралингвистическую информацию. Так в корпусной лингвистике возникла идея размеченного корпуса. *Разметка* (tagging, annotation) заключается в приписывании текстам и их компонентам специальных меток (tag, tags): внешних, экстралингвистических (сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика; сведения об авторе могут включать не только его имя, но также возраст, пол, годы жизни и многое другое. Это кодирование информации имеет название *метаразметка*), структурных (глава, абзац, предложение, словоформа) и собственно лингвистических, описывающих лексические, грамматические и прочие характеристики элементов текста. Набор этих метаданных во многом определяет возможности, предоставляемые корпусами исследователям. При выборе этих данных необходимо руководствоваться целями исследования и потребностями лингвистов, а также возможностями по внесению в текст тех или иных дополнительных признаков. Среди лингвистических типов разметки выделяются:

- *морфологическая* разметка. В иностранной терминологии употребляется термин part-of-speech tagging (POS-tagging), дословно – частеречная разметка. В действительности морфологические метки включают не только признак части речи, но и признаки грамматических категорий, свойственных данной части речи. Это основной тип разметки: во-первых, большинство крупных корпусов являются как раз морфологически размеченными корпусами, во-вторых, морфологический анализ рассматривается как основа для дальнейших форм анализа – синтаксического и семантического, и, в-третьих, успехи в компьютерной морфологии позволяют автоматически размечать корпусы больших размеров;
- *синтаксическая* разметка, являющаяся результатом синтаксического анализа, или *парсинга* (англ. parsing), выполняемого на основе данных морфологического анализа. Этот вид разметки описывает синтаксические связи между лексическими единицами и различные синтаксические конструкции (например, придаточное предложение, глагольное словосочетание и т.п.);
- *семантическая* разметка. Хотя для семантики нет единой семантической теории, чаще всего семантические тэги обозначают семантические категории, к которым относится данное слово или словосочетание, и более узкие подкатегории, специфицирующие его значение;

- *анафорическая* разметка. Фиксирует референтные связи, например, местоименные;
- *просодическая* разметка. В просодических корпусах применяются метки, описывающие ударение и интонацию. В корпусах устной разговорной речи просодическая разметка часто сопровождается так называемой *дискурсной* разметкой, которая служит для обозначения пауз, повторов, оговорок, и т.д.

Существуют и другие типы разметки.

1.5. Технология создания корпусов

Технологический процесс создания корпуса можно представить в виде следующих шагов или этапов.

1. Определение перечня источников.

2. Оцифровка текстов (преобразование в компьютерную форму). Следует сказать, что насколько раньше задача ввода текстов в компьютер была тяжела и трудоемка, настолько сегодня эта проблема решается довольно легко, по крайней мере, что касается современных текстов и в современной орфографии. Эта легкость базируется на успехах в оптическом вводе (сканирование) и распознавании текстовой информации и на глобальной компьютеризации современной жизни, в том числе и в областях, связанных с обработкой текстовой информации. Тексты в электронном виде для создания корпусов могут быть получены самыми разными способами — ручной ввод, сканирование, авторские копии, дары и обмен, Интернет, оригинал-макеты, предоставляемые составителям корпусов издательствами и проч.

3. Предобработка текста. На этом этапе все тексты, полученные из разных источников, проходят филологическую выверку и корректировку. Также осуществляется подготовка библиографического и экстралингвистического описания текста.

4. Конвертирование и графематический анализ. Некоторые тексты проходят также через один или несколько этапов предварительной машинной обработки, в ходе которых осуществляются различного рода перекодировка (если требуется), удаление или преобразование нетекстовых элементов (рисунки, таблицы), удаление из текста переносов, «жёстких концов строк», обеспечение единообразного написания тире и проч. Как правило, эти операции выполняются в автоматическом режи-

ме. Обычно на этом же этапе осуществляется сегментирование текста на его структурные составляющие.

5. Разметка текста. Разметка текста заключается в приписывании текстам и их компонентам дополнительной информации (метаданных). Метаописание текстов корпуса включает как содержательные элементы данных (библиографические данные, признаки, характеризующие жанровые и стилевые особенности текста, сведения об авторе), так и формальные (имя файла, параметры кодирования, версия языка разметки, исполнители этапов работ). Эти данные обычно вводятся вручную. Структурная разметка документа (выделение абзацев, предложений, слов) и собственно лингвистическая разметка обычно осуществляются автоматически.

6. На следующем этапе осуществляется корректировка результатов автоматической разметки: исправление ошибок и снятие неоднозначности (вручную или полуавтоматически).

7. Заключительный этап – конвертирование размеченных текстов в структуру специализированной лингвистической информационно-поисковой системы (corpus manager), обеспечивающей быстрый многоаспектный поиск и статистическую обработку.

8. И, наконец, обеспечение доступа к корпусу. Корпус может быть доступен в пределах дисплейного класса, может распространяться на CD-ROM и может быть доступен в режиме глобальной сети. Различным категориям пользователей могут предоставляться разные права и разные возможности.

Конечно, в каждом конкретном случае состав и количество процедур могут отличаться от выше перечисленных, и реальная технология может оказаться гораздо сложнее.

1.6. Автоматическая разметка

Фактически, корпус в его современном понимании – это всегда компьютерная база данных, и в процессе его создания естественно использование специальных программ. Среди этих программ особое место занимают программы автоматической разметки. Разметка корпусов представляет собой трудоемкую операцию, особенно учитывая размеры современных корпусов. Если для некоторых видов разметки, в частности анафорической, просодической, создание автоматических систем пока представляется довольно сложным и основная часть работы проводится вручную, то для морфологического и

синтаксического анализа существуют различные программные средства, которые принято называть соответственно тэггеры (taggers) и парсеры (parsers). В результате работы программ автоматического морфологического анализа каждой лексической единице приписываются грамматические характеристики, включая часть речи, лемму (нормальную форму) и набор граммем (например, род, число, падеж, одушевленность/неодушевленность, переходность и т.п.). В результате работы программ автоматического синтаксического анализа фиксируются синтаксические связи между словами и словосочетаниями, а синтаксическим единицам приписываются соответствующие характеристики (тип предложения, синтаксическая функция словосочетания и т.п.).

1.7. Исправление ошибок и снятие неоднозначности

Однако автоматический анализ естественного языка небезошибочен и многозначен – он, как правило, дает несколько вариантов анализа для одной лексической единицы (слова, словосочетания, предложения). В этом случае говорят о грамматической омонимии. Снятие неоднозначности (морфологической, синтаксической) в целом является одной из важнейших и сложнейших задач компьютерной лингвистики. При создании корпусов для снятия неоднозначности используются автоматические и ручные способы. Корпусы нового поколения включают сотни миллионов слов, поэтому выдвигаются принципы разработки систем, которые бы минимизировали вмешательство человека. Автоматическое разрешение морфологической или синтаксической омонимии, как правило, основывается на использовании информации более высокого уровня (синтаксического, семантического) с применением статистических методов.

1.8. Форматы данных и стандартизация

Корпусы, как правило, предназначены для многократного использования многими пользователями, соответственно, и их разметка, и их программное обеспечение должны быть определенным образом унифицированы. Что касается разметки, то как лингвистическая, так и экстралингвистическая разметка должны базироваться на некоторых достаточно широко распространенных и принятых принципах описания текстов и языковых единиц. Параметры разметки и их значения должны

быть достаточно «естественными», т.е. должны соответствовать общепринятым научным классификациям. Что касается программного обеспечения, то оно должно поддерживать обработку типовых запросов и решение типовых задач. Большое значение имеет унификация форматов, как их наполнения, так и структуры. Единые форматы представления данных позволяют во многих случаях использовать единое программное обеспечение и обмениваться корпусными данными. Стандартизация в отношении корпусов, совместимость типов данных важны и с точки зрения сравнимости разных корпусов. Вопросы оценки корпусов, их пригодности к различным заданиям также требуют своих «стандартов оценки».

В настоящее время на основе международного опыта выработались де-факто стандарты представления метаданных, базирующиеся на описаниях текстов в рамках проекта Text Encoding Initiative (TEI) и на рекомендациях EAGLES (Expert Advisory Group on Language Engineering Standards). В качестве формального языка разметки широко применяются языки SGML и XML. В настоящее время стандарты EAGLES непосредственно включаются в технологическую среду языка XML, см., в частности, разработку стандарта Corpus Encoding Standard for XML (XCES).

1.9. Корпусные менеджеры

Работа пользователей с корпусом осуществляется с помощью специализированных программных средств – *корпусных менеджеров*, предоставляющих разнообразные возможности по получению из корпуса необходимой информации:

- поиск конкретных словоформ;
- поиск словоформ по леммам;
- поиск группы словоформ в виде разрывной или неразрывной синтагмы;
- поиск словоформ по набору морфологических признаков;
- отображение информации о происхождении, типе текста и т.п.;
- вывод результатов поиска с указанием контекста заданной длины;
- получение различных лексико-грамматических статистических данных;

– сохранение отобранных строк конкорданса в отдельном файле на компьютере пользователя и др.

Результаты поиска обычно выдаются в виде конкорданса (поэтому корпусные менеджеры еще называют *конкордансерами*), где искомая единица представлена в ее контекстном окружении и в виде статистических данных. Последние могут фиксировать частотные характеристики отдельных языковых единиц, или граммем, или могут характеризовать совместную встречаемость нескольких лексических единиц. Многие системы позволяют настраивать формат выдачи (менять длину левого и правого контекста, задавать объем выдачи и порядок сортировки данных, отображать или не отображать лингвистические и экстралингвистические характеристики, и т.д.).

Пример выдачи корпусных менеджеров см. в Приложении 1 (рис. 2–4).

1.10. Пользователи и способы использования корпусов

Пользователей корпусов, как правило, интересует не содержание конкретных текстов, а их метатекстовая информация и примеры употребления тех или иных языковых элементов и конструкций. Это, в первую очередь, лингвисты. Первоначальные лингвистические исследования, проводившиеся с помощью корпусов, сводились к подсчету частот встречаемости различных языковых элементов. Статистические методики используются в решении сложных лингвистических задач, таких как машинный перевод, распознавание и синтез речи, средства проверки орфографии и грамматики и т.д. Так, устойчивые словосочетания представляют собой с семантической точки зрения неделимую смысловую единицу, что очень важно учитывать в лексикографии, системах автоматической обработки текста. На материале корпуса статистическими методами можно определить, какие слова встречаются вместе регулярно и, таким образом, могут быть отнесены к устойчивым словосочетаниям. Корпусы являются богатым источником данных для исследований по лексикографии и грамматике. С исследованиями по лексикографии тесно связаны исследования в области семантики. Наблюдая окружения той или иной лингвистической единицы в корпусе, можно установить определенные семантические признаки, характеризующие данную единицу.

Лингвисты-теоретики используют корпусы в качестве экспериментальной базы для проверки гипотез и доказательства своих теорий. Прикладные лингвисты (преподаватели, переводчики и т.п.) используют

компьютерные корпусы при обучении языкам и для решения своих профессиональных задач. Особый класс пользователей представляют компьютерные лингвисты: они пытаются выявить и использовать статистические и лингвистические закономерности, присутствующие в текстах, для создания компьютерных моделей языка. Другие специалисты по языку (литературоведы, редакторы) также в ряде случаев могут получить ответы на интересующие их вопросы, обратившись к корпусу. Специалисты по общественным наукам (историки, социологи) также могут изучать свои объекты через язык, используя такие параметры текстов, как период, автор или жанр. Литературоведы используют корпусы для стилеметрических исследований. Наконец, корпусы используются для разработки и настройки различных автоматизированных систем (машинный перевод, распознавание речи, информационный поиск).

1.11. Типы корпусов

Несмотря на разнообразие корпусов, можно выделить два основных способа деления корпусов на классы: 1) это противопоставление корпусов, относящихся ко всему языку (часто к языку определенного периода), корпусам, относящимся к какому-либо подъязыку (жанр, стиль, язык определенной возрастной или социальной группы, язык писателя или ученого и т.п.); 2) разделение корпусов по типу лингвистической разметки. Несмотря на наличие множества типов разметки, большинство реально существующих корпусов относится к корпусам морфологического либо синтаксического типа (последние в англоязычной литературе называют treebanks, что можно перевести как «банки синтаксических структур»). При этом следует подчеркнуть, что корпус с синтаксической разметкой явно или неявно включает в себя и морфологические характеристики лексических единиц.

Вообще же существует большое число разных типов корпусов. Их разнообразие определяется многообразием исследовательских и прикладных задач, для решения которых они создаются, и различными основаниями для классификации. В зависимости от поставленных целей и классифицирующих признаков, можно выделить различные типы корпусов (см. таблицу).

Классификация корпусов

Признак	Типы корпусов
Тип данных	Письменные Речевые Смешанные
Язык текстов	Русский Английский и т.д.
«Параллельность»	Одноязычные Двуязычные Многоязычные
«Литературность», специфичность	Литературные Диалектные Разговорные Терминологические Смешанные
Жанр	Литературные Фольклорные Драматургические Публицистические
Доступность	Свободно доступные Коммерческие Закрытые
Назначение	Исследовательские Иллюстративные
Динамичность	Динамические (мониторные) Статические
Разметка	Размеченные Неразмеченные
Характер разметки	Морфологические Синтаксические Семантические Просодические и т.д.
Объем текстов	Полнотекстовые «Фрагментнотекстовые»
Хронологический аспект	Синхронические Диахронические
«Общность»	Общие Одного писателя
Структура	Центральные и архивные Ядерные и периферийные

1.12. Терминология

Терминология корпусной лингвистики еще не установилась. Во-первых, это естественно, учитывая ее недавнее происхождение. Во-вторых, корпусная лингвистика как отдельная ветвь лингвистики сложилась в США и в Великобритании. И соответственно, ее терминология складывалась и продолжает складываться в недрах английского языка. И, естественно, русская корпусная терминология строится на базе англоязычной. В качестве примера и образца приведем фрагмент будущего словаря-тезауруса по корпусной лингвистике (Приложение 2). Одновременно заметим, что методология корпусной лингвистики может быть применена и к ней самой. То есть необходимо составить корпус текстов по корпусной лингвистике и разрабатывать словарь непосредственно на живом текстовом материале. Некоторое число публикаций на русском языке, посвященных вопросам создания и использования корпусов, уже имеется. В приложениях 2 и 3 этот подход иллюстрируется на примере англоязычной терминологии. Что касается русского языка, то среди специалистов до сих пор нет единодушия в отношении главного термина: *корпус*. Каким должно быть множественное число от слова «корпус»? Как образуется соответствующее прилагательное? Словари допускают для разных значений этого существительного две формы множественного числа: ко́рпусы и корпуса́. Для значения «массив», которое имеет место в случае языковых корпусов, именительный падеж множественного числа должен быть «ко́рпусы» и, соответственно, прилагательное «ко́рпусный» (Большой толковый словарь русского языка, СПб., 1998). Однако анализ узуса специалистов пока свидетельствует в пользу форм «корпуса́», «корпусно́й», «корпусна́я», которые используются заметно чаще, так что можно, видимо, с осторожностью сказать, что в настоящее время этот вопрос остается открытым.

2. Программа учебной дисциплины «Корпусная лингвистика»

2.1. Организационно-методический раздел

Программа дисциплины составлена в соответствии с государственным образовательным стандартом высшего профессионального образования по направлению 021800 — Лингвистика.

Цель курса состоит в том, чтобы познакомить студентов с концепциями корпусной лингвистики, дать им возможность освоить основы корпусных технологий, приобрести навыки работы с корпусами.

Задачи курса:

- ознакомить студентов с новой парадигмой в лингвистических исследованиях;
- ознакомить студентов с историей корпусных исследований;
- изучить языковые и программные средства корпусной лингвистики;
- сформировать навыки работы с программными средствами и информационными ресурсами корпусной лингвистики;
- сформировать навыки исследовательской работы по анализу языка на базе корпусных данных.

Место курса в профессиональной подготовке выпускника: курс рассчитан на детальное ознакомление с новыми методами лингвистических исследований. Даются специальные знания для тех, кто хочет специализироваться в данном направлении лингвистической науки.

Требования к уровню освоения содержания курса.

В результате обучения студент должен подробно знать:

- основные понятия корпусных технологий,
- основные типы корпусов,
- понятие разметки,
- основные стандарты разметки,
- средства создания корпусов,
- основные имеющиеся корпусы,
- типы программных средств для работы с корпусами;

должен уметь:

- создавать языковые корпусы,
- работать с программами-менеджерами и конкордансерами,
- осуществлять поиск и исследования на базе корпусов.

2.2. Содержание курса

Курс состоит из трех частей, которые могут изучаться как последовательно, так и каждая в отдельности:

1) Часть 1. Введение в корпусную лингвистику.
2) Часть 2. Создание корпусов.
3) Часть 3. Использование корпусов.

Де-факто все три части между собою связаны, так, например, методы создания корпусов определяются их назначением и типологией, которые рассматриваются в первой части. Языки запросов и возможности корпусных менеджеров во многом определяются разметкой, которая рассматривается в третьей части. И так далее.

2.3. Часть 1. Введение в корпусную лингвистику

2.3.1. Разделы:

1) Основные понятия корпусной лингвистики.
2) История создания лингвистических корпусов.
3) Типология корпусов.

Краткое содержание разделов

Раздел 1. Основные понятия корпусной лингвистики
Тема 1. Основные понятия и определения.
Тема 2. Лингвистические (языковые) и нелингвистические корпусы.

Раздел 2. История создания лингвистических корпусов
Тема 3. История лингвистических корпусов: от картотеки к корпусу.
Тема 4. Корпусная лингвистика: современное состояние.
Тема 5. Корпусная лингвистика в России.

Раздел 3. Типология корпусов

Тема 6. Классификация (типология) корпусов по различным основаниям.

Тема 7. Типы корпусов по задачам.

Тема 8. Типы корпусов по формальным признакам.

2.3.2. Примерные вопросы для самоконтроля

Дать определения терминов:

Корпус
Разметка
Репрезентативность
Метаданные
Корпусный менеджер
Treebank
Лемматизация
Конкорданс
Параллельный корпус

Перечислить типы корпусов

Назвать и охарактеризовать наиболее известные корпусы.

2.3.3. Примерная тематика докладов, рефератов, курсовых работ

Способы использования корпусов в лингвистических исследованиях.

Исследование способов использования корпусов в лексикографии.

Изучение средств обработки корпусных данных, представленных на языке XML.

Создание электронной хрестоматии по корпусной лингвистике.

Исследование механизмов взаимодействия корпуса текстов и электронной картотеки (корпусы цитат).

Создание веб-сайта по корпусной лингвистике.

2.3.4. Примерный перечень вопросов к экзамену (зачету)

История лингвистических корпусов: от картотеки к корпусу.

Классификация (типология) корпусов.

Корпусная лингвистика: современное состояние.

Корпусная лингвистика в России.

Обзор существующих корпусов различных типов.

Корпус как поисковая система.

Корпусоподобные интерфейсы между лингвистом и поисковыми системами Интернета.

Лингвистические исследования, базирующиеся на корпусах.

2.3.5. Распределение часов курса по темам и видам работы

№ раз-дела	Наименование тем и разделов	ВСЕГО (ч)	Аудиторные занятия (ч)		Самостоя-тельная работа
			лекции	семинары	
1	Основные поня-тия корпусной лингвистики	40	8	2	30
2	История созда-ния лингвисти-ческих корпусов	40	10	–	30
3	Типология кор-пусов	52	10	2	40
ИТОГО:		132	28	4	100

2.3.6. Форма текущего, промежуточного и итогового контроля

В течение семестра слушатели выполняют лабораторные (практические) работы, готовят письменные работы (рефераты) по одной из выбранных тем, которые «защищаются» в конце курса в виде докладов. В конце курса — зачет.

2.3.7. Учебно-методическое обеспечение курса

Основная литература

Андрющенко В.М. Концепция и архитектура машинного фонда русского языка / Отв. ред. А.П. Ершов. М., 1989.

Баранов А.Н. Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику. М., 2001. С.112–137.

Вербицкая Л.А., Казанский Н.Н., Касевич В.Б. Некоторые проблемы создания национального корпуса русского языка // Научно-техническая информация. Сер. 2. 2003. № 6. С. 2–8.

Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» / Под ред. А.С. Герда. СПб., 2002.

Научно-техническая информация. Сер. 2. 2005. № 3.

Научно-техническая информация. Сер. 2. 2003. № 6.

Рыков В.В. Прагматически ориентированный корпус текстов // Тверской лингвистический меридиан. Вып. 3. Тверь, 1999. С. 89–96 // См. также http://rykov-cl.narod.ru/t.html.

Труды Международного семинара по компьютерной лингвистике и ее приложениям «Диалог-2000», «Диалог-2001», «Диалог-2002», «Диалог-2003», «Диалог-2004», «Диалог-2005».

Труды Международной научной конференции «Корпусная лингвистика 2004» / Под ред. А.С. Герда. СПб., 2004.

Чардин И.С. Лингвистические корпуса с синтаксической разметкой и их применение // Научно-техническая информация. Сер. 2. 2003. № 6. С. 18–24.

Дополнительная литература

English Corpus Linguistics: Studies in Honour of Jan Svartvik / Aijmer K., Altenberg B. (eds.). London, 1991.

Čermák F. Today's Corpus Linguistics: Some Open Questions // International Journal of Corpus Linguistics. 2002. Vol. 7, N 2. P. 265–282.

Fillmore C.J., Atkins B.T.S. Starting Where the Dictionaries Stop: the Challenge of Corpus Lexicography // Atkins B.T.S., Zampolli A. (eds.). Computational Approaches to the Lexicon. 1994.

Kennedy G. An Introduction to Corpus Linguistics. London, 1998.

Leech G. The State of Art in Corpus Linguistics // English Corpus Linguistics / Aijmer K., Altenberg B. (eds.). London, 1991. P. 8–29.

McEnery A., Wilson A. Corpus Linguistics. Edinburgh, 1996.

Francis N. W. Language Corpora B.C. // Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82. Stockholm, 4.–6. August 1991. / Svartvik J. (ed.). P. 17–32.

Proceedings of the LREC (Language Resource Evaluating Conference). 2002, 2003, 2004, 2005.

Quirck R. On Corpus Principles and Design // Directions in Corpus Linguistics. Berlin; New York, 1992. P. 461–462.

Sinclair J. M. The Automatic Analysis of Corpora // Directions in Corpus Linguistics. Berlin, 1992.

Svartvik, J. (ed.). Directions in Corpus Linguistics, Berlin. 1992.

Zakharov V. Russian Corpus of the 19th Century // Text, Speech and Dialogue: Proceedings of the 6th International Conference TSD 2003, České Budějovice, Czech Republic, September 2003 / Václav Matoušek, Pavel Mautner (eds.). Berlin; Heidelberg, 2003. P. 146–151. (Lecture Notes in Artificial Intelligence, 2807).

2.4. Часть 2. Создание корпусов

2.4.1. Разделы:

1) Предварительные работы по созданию корпуса.
2) Разметка. Средства создания и разметки корпусов.
3) Стандартизация в корпусной лингвистике.

Краткое содержание разделов

Раздел 1. Предварительные работы по созданию корпуса

Тема 1. Проблемы репрезентативности.

Тема 2. Отбор источников. Внешние и внутренние критерии отбора.

Тема 3. Нормализация файлов.

Тема 4. Графематический анализ.

Раздел 2. Разметка. Средства создания и разметки корпусов

Тема 5. Понятие разметки.

Тема 6. Типы разметки.

Тема 7. Автоматический морфологический и синтаксический анализ.

Тема 8. Металингвистическая разметка.

Тема 9. Параллельные корпусы. Проблема выравнивания.

Раздел 3. Стандартизация в корпусной лингвистике

Тема 10. Языковые средства представления размеченных текстов.

Тема 10. Международные стандарты и проекты (TEI, EAGLES, CDIF, XCES).

2.4.2. Примерные вопросы для самоконтроля

Дать определения терминов:

Разметка
Репрезентативность
Метаданные
Корпусный менеджер
Treebank
Лемматизация
Параллельный корпус

Перечислить типы корпусов

2.4.3. Примерная тематика докладов, рефератов, курсовых работ

Графематический анализ текстов.

Унификация текстов внутри корпуса 19 века.

Автоматическая морфологическая разметка текстов 19 века.

Исследование набора метаданных для корпуса 19 века.

База данных «Морфологический словарь языка 19 века».

Создание параллельного англо-русского корпуса.

Создание параллельного русско-чешского корпуса.

Создание параллельного русско-словацкого корпуса.

Методы снятия морфологической неоднозначности.

Исследование механизмов взаимодействия корпуса текстов и электронной картотеки (корпусы цитат).

Анализ функций сегментных внеалфавитных графем («межморфемный» дефис, «межслоговой» дефис, «межсловный» дефис, апостроф).

Проблема строчных и прописных букв в корпусах текстов (имена собственные и нарицательные, сплошная и начальная капитализация).

Проблема омографии – акцентно-ориентированный морфологический анализ.

Разработка модуля преобразования каллиграфем (жирность, курсивность, подчёркивание) в тэги языка XML.

Анализ функций точки (и других знаков препинания) с точки зрения структурной разметки текста.

Методы выделения структурных элементов текста (часть, глава, параграф, абзац).

Составные лексемы.

Методы снятия морфологической неоднозначности.

Методы выделения структурных элементов текста (часть, глава, параграф, абзац).

Составные лексемы.

Проект TEI (обзор).

Стандарты EAGLES (обзор).

Форматы CDIF и XCES.

2.4.4. Примерный перечень вопросов к экзамену (зачету)

Проблемы репрезентативности корпусов.

Проблемы хронологии в общеязыковых корпусах.

Отбор текстов для корпусов.

Графематический анализ.

Понятие разметки.

Типы разметки.

Морфологическая разметка.

Синтаксические корпуса (treebanks).

Семантическая разметка.

Технология создания корпусов. Стадии работы.

Понятие корпусоида.

Автоматическая морфоразметка.

Автоматический синтаксический анализ (parsing).

Языковые средства представления размеченных текстов (языки SGML, XML).

Международные стандарты (TEI, EAGLES, CDIF, XCES).

2.4.5. Распределение часов курса по темам и видам работы

№ раз-дела	Наименование тем и разделов	ВСЕГО (ч)	Аудиторные занятия (ч)		Самостоя-тельная работа
			лекции	семинары	
1	Предварительные работы по созданию корпуса	38	8	–	30
2	Разметка. Средства создания и разметки корпусов	60	16	4	40
3	Стандартизация в корпусной лингвистике	38	6	2	30
ИТОГО:		136	30	6	100

2.4.6. Форма текущего, промежуточного и итогового контроля

В течение семестра слушатели выполняют лабораторные (практические) работы, готовят письменные работы (рефераты) по одной из выбранных тем, которые «защищаются» в конце курса в виде докладов. В конце курса — зачет.

2.4.7. Учебно-методическое обеспечение курса

Основная литература

Богуславский И.М. и др. Аннотированный корпус русских текстов: Концепция, инструменты разметки, типы информации // Труды Международного семинара по компьютерной лингвистике и ее приложениям «Диалог-2000». Протвино, 2000.

Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» / Под ред. А.С. Герда. СПб., 2002.

Копотев М.В., Мустайоки А. Принципы создания Хельсинкского аннотированного корпуса русских текстов (ХАНКО) в сети Интернет // Научно-техническая информация. Сер. 2. 2003. № 6. С. 33–36.

Научно-техническая информация. Сер. 2. 2005. № 3, 6. 2003. № 6.

Труды Международного семинара по компьютерной лингвистике и ее приложениям «Диалог-2000», «Диалог-2001», «Диалог-2002», «Диалог-2003», «Диалог-2004», «Диалог-2005».

Труды Международной научной конференции «Корпусная лингвистика – 2004» / Под ред. А.С. Герда. СПб., 2004.

Шаров С.А. Параметры описания текстов корпуса. // http://bokrcorpora.narod.ru/header.html.

Шаров С.А. Формат выходного представления корпуса текстов. // http://bokrcorpora.narod.ru/format.html.

Дополнительная литература

Atkins S., Clear J., Ostler N. Corpus Design Criteria // Literary and Linguistic Computing. 1992. Vol. 7, N. 1. P. 1–16.

Biber D. Representativeness in Corpus Design // Literary and Linguistic Computing. 1993. Vol. 8, N. 4. P. 243–258.

Brill E. A Simple Rule-Based Part-of-Speech Tagger // Proceedings of the Third Conference on Applied Natural Language Processing (ANLF-92), Trento, Italy. 1992.

Burnard L. A Gentle Introduction to SGML. TEI P2. 1993.

Burnard L. A Gentle Introduction to XML. 1993 // http://www.tei-c.org/Guidelines2/gentleintro.html.

Burnard L. The Text Encoding Initiative: an Overview. // Spoken English on Computer / Leech G., Myers G., Thomas J. (eds.) New York, 1995. P. 223–235; См. также http://www-tei.uic.edu/orgs/tei/.

Lee D. Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle // Language Learning & Technology. September 2001. Vol. 5, N. 3, P. 37–72; См. также http://llt.msu.edu/vol5num3/pdf/lee.pdf

Leech G. Corpus Annotation Schemes // Literary and Linguistics Computing. 1993. Vol. 8. N. 4. P.275–281.

Proceedings of the LREC (Language Resource Evaluating Conference). 2002, 2003, 2004, 2005.

Sharoff S. Towards Basic Categories for Describing Properties of Texts in a Corpus. In Proc. of Language Resources and Evaluation Conference (LREC04). May, 2004, Lisbon, Portugal // http://www.comp.leeds.ac.uk/ssharoff/texts/lrec-04.pdf.

Sinclair J. Preliminary Recommendations on Text Typology. EAGLES Document EAG-TCWG-TTYP/P, 1996 // http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html.

TEI P4: Guidelines for Electronic Text Encoding and Interchange. 2001 / Sperberg-McQueen C. M., Burnard L. (eds.) // http://www.hcu.ox.ac.uk/TEI/P4X/index.html.

UCREL: Corpus Annotation // http://www.comp.lancs.ac.uk/ucrel/annotation.htm

XML Corpus Encoding Standard Document XCES 0.2. // http://www.cs.vassar.edu/XCES/

Zakharov V., Volkov S. Evaluating Morphological Tagging of Russian texts of the XIXth Century // Text, Speech and Dialogue: Proceedings of the 7th International Conference TSD 2004, Brno, Czech Republic, September 2004 / Petr Sojka, Ivan Kopeček, Karel Pala (eds.). Berlin; Heidelberg, 2004. P. 235–242. (Lecture Notes in Artificial Intelligence, 3206).

2.5. Часть 3. Использование корпусов

2.5.1. Разделы:

1) Обзор существующих корпусов различных типов.
2) Корпусные менеджеры.
3) Корпусные исследования.

Краткое содержание тем

Раздел 1. Обзор существующих корпусов различных типов

Тема 1. Зарубежные национальные корпусы.
Тема 2. Корпусы русского языка.
Тема 3. Специальные корпусы.

Раздел 2. Корпусные менеджеры

Тема 4. Корпус как поисковая система.
Тема 5. Языки запросов.
Тема 6. Выходные интерфейсы.
Тема 8. Сравнительный анализ.

Раздел 3. Корпусные исследования

Тема 9. Лексические исследования, базирующиеся на корпусах.
Тема 10. Грамматические исследования, базирующиеся на корпусах.
Тема 11. Семантические исследования, базирующиеся на корпусах.

Тема 12. Использование корпусов в социологии, исторической науке и др.

2.5.2. Примерные вопросы для самоконтроля

Когда был создан BNC?
Когда был создан CNK?
Как назывался первый корпус русского языка?
Каков был объем первого корпуса русского языка?
Корпусы каких писателей существуют?
Корпусы каких писателей доступны через Интернет?
Что такое язык регулярных выражений?
Что такое меры MI и T-score?

2.5.3. Примерная тематика докладов, рефератов, курсовых работ

Анализ и описание различных корпусов.
Анализ и описание корпусного менеджера Xaira.
Анализ и описание корпусного менеджера Bonito.
Анализ и описание корпусного менеджера QPL.
Анализ и описание интерфейса WebCorp.
Сравнительный анализ возможностей корпусов и поисковых систем Интернета.
Использование корпусов в социологии.
Использование корпусов в этнолингвистике.

2.5.4. Примерный перечень вопросов к экзамену (зачету)

Британский национальный корпус.
Чешский национальный корпус.
Польский национальный корпус.
Национальный корпус русского языка.
Мангеймский корпус немецкого языка.
Русско-английский корпус С. Шарова.
Корпус языка А.С. Грибоедова.
Корпус русского языка 19 века.
Языки запросов корпусных менеджеров: общая характеристика.
Языки запросов конкретных корпусных менеджеров.
Выходные интерфейсы корпусных менеджеров: общая характеристика.

Выходные интерфейсы конкретных корпусных менеджеров.
Типы лексических исследований, базирующихся на корпусах.
Типы грамматических исследований, базирующихся на корпусах.
Семантическое наполнение Национального корпуса русского языка.
Использование корпусов в других науках.
Статистические меры вычисления совместной встречаемости.
Веб как корпус.

2.5.5. Распределение часов курса по темам и видам работы

№ раз-дела	Наименование тем и разделов	ВСЕГО (ч)	Аудиторные занятия (ч)		Самостоя-тельная работа
			лекции	семинары	
1	Обзор существующих корпусов различных типов	24	4	—	20
2	Корпусные менеджеры	56	12	4	40
3	Корпусные исследования	56	12	4	40
ИТОГО:		136	28	8	100

2.5.6. Форма текущего, промежуточного и итогового контроля

В течение семестра слушатели выполняют лабораторные (практические) работы, готовят письменные работы (рефераты) по одной из выбранных тем, которые «защищаются» в конце курса в виде докладов. В конце курса – экзамен.

2.5.7. Учебно-методическое обеспечение курса

Основная литература

Венцов А.В., Касевич В.Б., Ягунова Е.В. Корпус русского языка и восприятие речи // Научно-техническая информация. Сер. 2. 2003. № 6. С. 25–32.

Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» / Под ред. А.С. Герда. СПб., 2002.

Захаров В.П. Чешский национальный корпус текстов: организация и способы использования // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» / Под ред. А.С. Герда. СПб., 2002. С. 72–79.

Копотев М.В. Корпусная лингвистика в Финляндии (обзор ресурсов) // Научно-техническая информация. Сер. 2. 2003. № 6. С. 37–41.

Научно-техническая информация. Сер. 2. 2003. № 6, 10. 2005. № 3.

Труды Международного семинара по компьютерной лингвистике и ее приложениям «Диалог-2000», «Диалог-2001», «Диалог-2002», «Диалог-2003», «Диалог-2004», «Диалог-2005».

Труды Международной научной конференции «Корпусная лингвистика – 2004» / Под ред. А.С. Герда. СПб., 2004.

Шаров, С.А. Представительный корпус русского языка в контексте мирового опыта // НТИ. Сер. 2. 2003. № 6. С. 9–17.

Дополнительная литература

Aarts Jan. Комментарий к статье «A New Corpus of English» (Sidney Greenbaum) // Directions in Corpus Linguistics. Berlin, 1992.

Adam.Kilgarriff. Web as Corpus// http://www.itri.bton.ac.uk/~Adam.Kilgarriff/wac_cfp.html.

Ball Catherine N. Tutorial: Concordances and Corpora // http://www.georgetown.edu/cball/corpora/tutorial.html.

BNC: The BNC Users Reference Guide, 2000. http://www.natcorp.ox.ac.uk/World/HTML/.

Český Národní Korpus – Úvod a Příručka Uživatele / Kocek J., Kopřivová M., Kučera K. (eds.). Praha, 2000.

Fillmore C.J., Atkins B.T.S. Starting Where the Dictionaries Stop: the Challenge of Corpus Lexicography // Computational Approaches to the Lexicon / Atkins B.T.S., Zampolli A. (eds.). 1994.

Gellerstam Martin. Modern Swedish Text Corpora // Directions in Corpus Linguistics. Berlin, 1992. P. 151–159.

Oakes M.P. Statistics for Corpus Linguistics. Edinburgh University Press, Edinburgh, 1998.

Proceedings of the LREC (Language Resource Evaluating Conference). 2002, 2003, 2004, 2005.

Sinclair J. Corpus, Concordance, Collocation, Oxford University Press, 1991.

Приложение 1

Корпусы в сети Интернет

Приведем сетевые адреса и краткие сведения о некоторых корпусах. В Интернете можно получить доступ и найти списки самых различных корпусов — см., например, D. Lee. Bookmarks for Corpus-based Linguists (http://devoted.to/corpora), веб-страницы М. Барбера (Manuel Barbera) (http://www.bmanuel.org/index.html) или М. Барлоу (Michael Barlow) (http://www.athel.com/corpus.html), сайт Language and Speech Resources (http://www.elsnet.org/resources.html) и др.

Национальный корпус русского языка **http://ruscorpora.ru**	70 млн слов[1] ***См. поисковые формы и образцы выдачи на рис. 1, 2, 5–7.***
Компьютерный корпус текстов русских газет конца XX-го века **http://www.philol.msu.ru/~lex/corpus**	200 тыс. слов Система поиска по корпусу временно недоступна
Корпус русского языка ХАНКО (Хельсинский университет) **http://www.ling.helsinki.fi/projects/hanco/**	100 тыс. слов Ручная морфологическая разметка
Корпуса русских текстов на сайте Университета в Лидсе, Великобритания **http://corpus.leeds.ac.uk**	
Русские корпуса Тюбингенского Университета **http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html**	
Словарь-корпус языка А.С. Грибоедова **http://www.inforeg.ru/electron/concord/concord.htm**	120 тыс. слов
Уппсальский корпус русских текстов Доступен для поиска на сайте http://www.sfb441.uni-**tuebingen.de/b1/en/korpora.html**	1 млн слов 600 текстов (публицистика 1985-1989; литературные произведения 1960-1988).
Банк английского языка (Bank of English) **http://www.collins.co.uk/books.aspx?group=153** Свободный доступ: **http://www.collins.co.uk/Corpus/CorpusSearch.aspx**	524 млн слов, 56 млн в свободном доступе (The Collins Wordbanks *Online* English corpus: 36 млн – брит. англ., 10 млн – амер. англ., 10 млн – брит. разговорн. англ.) ***См. образец выдачи на рис. 4.***

[1] Под словом здесь и далее имеется в виду словоупотребление (англ. token).

Британский национальный корпус **http://www.natcorp.ox.ac.uk/** или **http://sara.natcorp.ox.ac.uk/**	100 млн слов Корпусные менеджеры SARA и XAIRA (http://www.xaira.org)
Венгерский национальный корпус **http://corpus.nytud.hu/mnsz/**	100 млн слов
Корпус испанского языка (исторический) **http://www.corpusdelespanol.org/**	100 млн слов, тексты 13–20 вв. Создан в Иллинойском университете, США
Корпус латинских текстов «Персей» **http://www.perseus.tufts.edu**	
Корпус современного датского языка **http://www.korpus2000.dk/**	50 млн слов Тексты 1998–2002 гг.
Корпус современного итальянского языка CORIS/CODIS **http://www.cilta.unibo.it/ricerca.htm**	100 млн Слов
Корпус современного китайского языка (LIVAC Synchronous Corpus) **http://www.rcl.cityu.edu.hk/livac/**	720 млн слов (150 млн иероглифов)
Мангеймский корпус немецкого языка (Institut für Deutsche Sprache, Mannheim, Germany) **http://corpora.ids-mannheim.de/~cosmas/**	1610 млн слов Корпусный менеджер COSMAS
Национальный корпус словенского языка **http://www.fida.net/eng/**	Более 100 млн слов
Польский национальный корпус **http://korpus.ia.uni.lodz.pl/**	93 млн слов
Словацкий национальный корпус **http://korpus.juls.savba.sk**	180 млн слов Используется корпусный менеджер Manatee/Bonito.
Хорватский национальный корпус **http://www.hnk.ffzg.hr/**	53 млн слов Корпусный менеджер Manatee/Bonito.
Чешский национальный корпус **http://ucnk.ff.cuni.cz**	100 млн слов + 100 млн нового корпуса современной лексики Корпусный менеджер Manatee/Bonito. ***См. образец выдачи на рис. 3.***
Эстонский корпус **http://test.cl.ut.ee/korpused/baaskorpus/1980/index.html.en**	

Рис. 1. Запросная форма НКРЯ для поиска по морфологическим признакам.

Слово 1: **слово nom&pl**
расстояние между словами: **1**
Слово 2: **слово**

Область поиска: **основной корпус (со снятой и неснятой омонимией)**

Найдено документов: 59, контекстов: 69

1. ЕСЛИ ЧИНОВНИКИ НАЧАЛИ БОРЬБУ С КОРРУПЦИЕЙ — БЕРЕГИ КАРМАНЫ // «Красноярский рабочий», 2003.01.01 [омонимия не снята] Все контексты(1)

Слова, слова, слова... [ЕСЛИ ЧИНОВНИКИ НАЧАЛИ БОРЬБУ С КОРРУПЦИЕЙ — БЕРЕГИ КАРМАНЫ // «Красноярский рабочий», 2003.01.01] ☒

2. Юлия Рахаева. Две трети Аполлона Григорьева. Единственная профессиональная литературная премия назвала лауреатов // «Известия», 2003.01.26 [омонимия не снята] Все контексты(1)

Потом уже были **слова, слова, слова**... [Юлия Рахаева. Две трети Аполлона Григорьева. Единственная профессиональная литературная премия назвала лауреатов // «Известия», 2003.01.26] ☒

3. Евгений Ясин. ИНТЕРЕСНЫЙ ВОПРОС // «Известия», 2003.07.08 [омонимия не снята] Все контексты(1)

Что это: **слова словами**, а команда пока не дана? [Евгений Ясин. ИНТЕРЕСНЫЙ ВОПРОС // «Известия», 2003.07.08] ☒

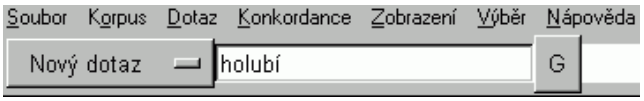
……..

Страницы: ← **1** **2** **3** **4** **5** **6** →

Поиск осуществлен системой Яndex.Server
При цитировании примеров просим ссылаться на Национальный корпус русского языка

Рис. 2. Образец выдачи в НКРЯ

1. Поиск словоупотреблений слова holubí (голубиный).



(. . .)

jej . Hedvábným slunečníkem **holubí** barvy procházelo slunce a
podobné zábrany rovněž slabé a **holubí** samec dokáže v těsné kleci
řeklo by se , že vybírala **holubí** vejce , to se o ní vědělo
odpověděl , " červené jako **holubí** nožky , červenější než ohromné
uniformu tvořila tunika v barvě **holubí** šedi , která se oblékala
znám jitra nadšená jak hejna **holubí** , a časem viděl jsem , co

(. . .)

2. Поиск словосочетания «holubí vejce» (голубиное яйцо) в любой форме и в любом написании (строчные и прописные)

[lemma="holubí"] [lemma="vejce"]

(. . .)

uzavřený v plachetce a připomíná **holubí vejce** . Později , po otevření ,
stromů , jen vzácně na zemi . **Holubí vejce** jsou bílá nebo nahnědlá .
spečenou žluč v kámen velikosti **holubího vejce** a břich na píď obalený tukem

(. . .)

3. Поиск всех прилагательных (A) в краткой форме (C), мужского рода (Y), единственного числа (S)

[tag="ACYS.*"]

(. . .)

to bylo včera , a ty jsi **schopen** jí všechno vyprdolit . Tak
společnost . Jeho vtip a šarm byl **znám** a poslední leč bez něj by
hladu nepoprali . Bořivoj byl **spokojen** . " Dobře jsme tu hospodu

(. . .)

Рис. 3. Образец выдачи в Чешском национальном корпусе.

Collocation Sampler

Type in your word:

Select a significance score to be calculated:

Mutual Information

T-score

To get collocations, press this button: Show Collocates

Note that output from this demo facility will be restricted to 100 collocates. These will be the statistically most significant ones according to the score you have selected.

Collocation for 'CORPUS'

Collocate	Corpus Freq	Joint Freq	Significance
the	2313407	189	5.540490
erm	84143	26	4.294184
million	15796	19	4.182154
christi	27	17	4.122786
spoken	1542	17	4.104865
er	98042	23	3.798765
a	973489	81	3.724491
habeas	12	12	3.463933
word	7972	11	3.199393
mm	73646	16	3.102023
software	1216	9	2.980231
based	7749	9	2.874020

Рис. 4. Интерфейс для вычисления коэффициента совместной встречаемости и образец выдачи в корпусе COBUILD

Приложение 2

Метаданные текстов в «Национальном корпусе русского языка» (НКРЯ)

Метаописание в НКРЯ состоит из двух блоков, первый из которых включает следующие признаки:

1) *Автор текста*: имя, пол, дата рождения (или примерный возраст);

2) *Название текста*;

3) *Время создания текста* (точно или приблизительно);

4) *Объем текста*: для художественных произведений принято, что обычная длина рассказа — менее 5 тыс. слов; обычная длина повести — от 5 до 15 тыс. слов; обычная длина романа — более 15 тыс. слов.

Второй блок содержит параметры метаописания трех основных *массивов* текстов корпуса: а) художественных текстов; б) нехудожественных текстов; в) драматургии.

Для художественных текстов предлагаются следующие параметры:

1) *Жанр текста:* нежанровая проза, автобиографическая проза, детектив, детская литература, историческая проза, криминальная литература, приключения, фантастика, юмор и сатира;

2) *Тип текста:* автобиографическая проза, ассоциативная проза, очерк, литературное письмо, повесть, пьеса, рассказ, роман, сказка, эссе;

3) *Хронотоп текста* (приблизительное указание на место и время описываемых в тексте событий; включается также помета «хронотоп не определен»). Реально предлагается следующее: древний Восток; Россия XVII в.; Россия XVIII в.; Россия XIX в.; Россия/СССР: советский период в целом; Россия, советский период – Германия 1920–1940-е; Россия/СССР – Европа 1960–1980-е; Россия/СССР: перестройка; Россия/СССР: советский и постсоветский период; Америка: современная жизнь; Америка: 1960–1980-е; Израиль: современная жизнь; Средняя Азия: современная жизнь; ирреальный мир и некоторые другие.

Для нехудожественных текстов установлены следующие параметры:

1) *Тип текста:* автобиография, дневник, договор, документ, закон, заметка, заявление, инструкция, информационное сообщение, кодекс, комментарий, объявление, отзыв, отчет, очерк, письмо, проповедь, резюме, рецензия, рецепт, сочинение, справочник, статья, учебник, характеристика, хроника, эссе, юридический документ (включается также помета «тип не определен») и пр. (всего 62 параметра).

2) *Тематика текста:* (открытый список в 5 подмножествах): бизнес, коммерция, экономика, финансы; война и вооруженные конфликты; дом; здоровье и медицина; досуг; искусство; криминал; наука (по разделам и отраслям); политика и общественная жизнь; право; производство; сельское хозяйство; спорт; природа; частная жизнь и т.п.

Помимо названной, в «Национальном корпусе» существует еще служебная или «имплицитная» метаразметка, которая не выносится на открытый доступ для широкого пользователя. К этой метаразметке относятся:

1) «текст-стиль», при этом выделяются академический, научно-популярный, официально-деловой, нейтральный, сниженный, сниженный с элементами грубого просторечия и жаргона, архаизованный, индивидуально-авторский, диалектный и пр. (всего 21);

2) аудитория-возраст;

3) аудитория-уровень образования;

4) аудитория-размер.

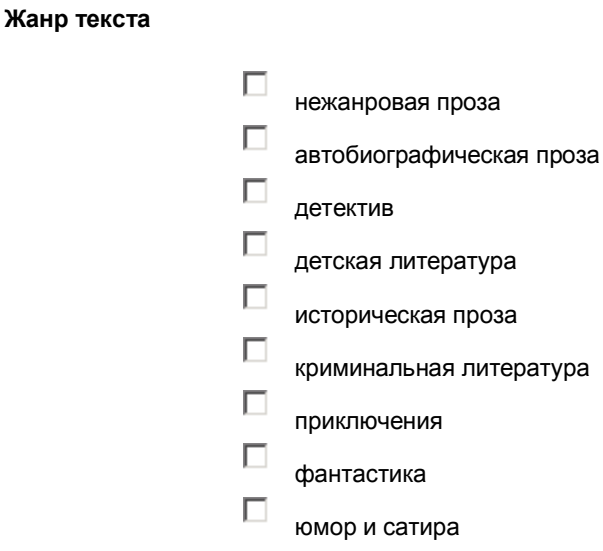


Рис. 5. Запросная форма НКРЯ для поиска по жанру.

Тип текста

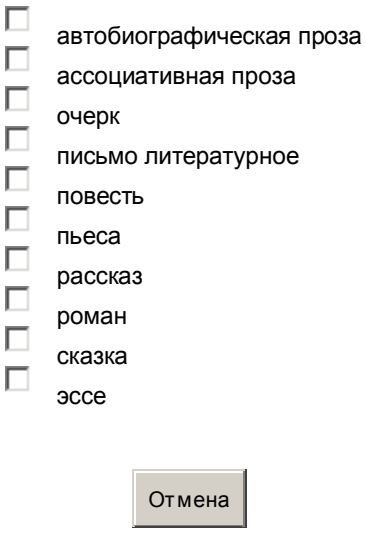


Рис. 6. Запросные формы НКРЯ для поиска по типу текста.

Автор текста

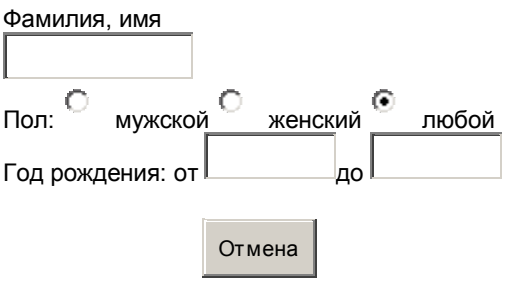


Рис. 7. Запросная форма НКРЯ для поиска по автору.

Приложение 3

Фрагмент словаря-тезауруса по корпусной лингвистике

В структуре словарных статей выделяются поля, которые помечены следующими метками: Term\ – англоязычный термин; Trans\ – русскоязычный термин; Def\ – определение; Syn\ – синоним; Ant\ – антоним; Up\ – вышестоящий термин; Down\ – нижестоящий термин, Cyt\ – цитата.

Term\ **aligned parallel corpus**

Trans\ выровненный параллельный корпус

Def\ Parallel corpus where texts in one language and their translations into other languages are aligned, sentence by sentence, phrase by phrase.

Up\ parallel corpus

Cyt\ A type of **multilingual corpus** where texts in one language and their translations into other languages are *aligned*, sentence by sentence, preferably phrase by phrase. Sometimes *reciprocate parallel **corpora*** are set up, **corpora** containing authentic texts as well as translations in each of the languages involved. This allows double-checking translation equivalents.

Cyt\ A parallel **corpus** is not immediately user-friendly. For the **corpus** to be useful it is necessary to identify which *sentences* in the sub-**corpora** are translations of each other, and which *words* are translations of each other. A **corpus** which shows these identifications is known as an *aligned **corpus*** as it makes an explicit link between the elements which are mutual translations of each other. For example, in a **corpus** the sentences "Das Buch ist auf dem Tisch" and "The book is on the table" might be aligned to one another. At a further level, specific words might be aligned, e.g. "Das" with "The". This is not always a simple process, however, as often one word in one language might be equal to two words in another language, e.g. the German word "raucht" would be equivalent to "is smoking" in English.

Term\ **aligned reciprocate parallel corpus**

Trans\ выровненный двусторонний параллельный корпус

Def\ Reciprocate parallel corpus where texts and their translations are aligned, sentence by sentence, phrase by phrase.

Up\ reciprocate parallel corpus

Cyt\ A type of **multilingual corpus** where texts in one language and their translations into other languages are *aligned*, sentence by sentence, preferably phrase by phrase. Sometimes *reciprocate parallel **corpora*** are set up, **corpora** containing authentic texts as well as translations in each of the languages involved. This allows double-checking translation equivalents.

Cyt\ A parallel **corpus** is not immediately user-friendly. For the **corpus** to be useful it is necessary to identify which *sentences* in the sub-**corpora** are translations of each other, and which *words* are translations of each other. A **corpus** which shows these identifications is known as an *aligned **corpus*** as it makes an explicit link between the elements which are mutual translations of each other. For example, in a **corpus** the sentences "Das Buch ist auf dem Tisch" and "The book is on the table" might be aligned to one another. At a further level, specific words might be aligned, e.g. "Das" with "The". This is not always a simple process, however, as often one word in one language might be equal to two words in another language, e.g. the German word "raucht" would be equivalent to "is smoking" in English.

Term\ **annotated corpus**

Trans\ размеченный корпус

Def\ Corpus enhanced with additional linguistic information.

Syn\ tagged corpus

Ant\ unannotated corpus

Up\ corpus

Down\ phonetically transcribed corpus

Down\ parsed corpus

Cyt\ A type of **corpus** enhanced with various types of linguistic information (or *tagged **corpus***). An **annotated corpus** may be considered to be a repository of linguistic information, because the information which was implicit in the plain text has been made explicit through concrete **annotation**.

Cyt\ More difficult is the question of *annotated **corpora***. It is proposed that this term is used for any **corpus** which includes codes that record extra information -- provenance, analytical marks, etc. Again the annotations should be separable from the plain text in a simple and agreed fashion. A set of conventions for removing, restoring and manipulating annotations is

necessary, especially as the next few years will see a large growth in the provision of annotated **corpora**. It is naive to expect that big **corpora** will remain easy to manage if they are full of various annotations; retrieval times are already critical.

Cyt\ For example, the form "gives" contains the implicit part-of-speech information "third person singular present tense verb" but it is only retrieved in normal reading by recourse to our pre-existing knowledge of the grammar of English. However, in an annotated **corpus** the form "gives" might appear as "gives_VVZ", with the code VVZ indicating that it is a third person singular present tense (Z) form of a lexical verb (VV). Such annotation makes it quicker and easier to retrieve and analyse information about the language contained in the **corpus**.

Cyt\ **Enriched data:** Many **corpora** have already been enriched with additional linguistic information such as part-of-speech annotation, parsing and prosodic transcription. Hence data retrieval from annotated **corpora** can be easier and more specific than with unannotated data.

Term\ **balanced corpus**

Trans\ исчерпывающий корпус

Look\ saturated corpus

Cyt\ A type of **corpus** composed according to parameters such as text type, genre or domain.

Cyt\ The Core **Corpus** of 2 million words is intended to be a representative subset of the whole **corpus**, in the sense that it contains samples from all the major subdivisions of the whole BNC, and in approximately the same proportions as those found in the BNC as a whole. There is one major exception to this statement however: whereas in the whole BNC, only c.10 million words (10% of the **corpus**) consist of spoken data, the Core **Corpus** is divided approximately equally between written and spoken material (c.1 million words each). It will generally be felt that in an ideal balanced **corpus** of the language, at least half of the material should be spoken English. It was only the impracticality of collecting and transcribing 50 million words of the spoken language which led to abandonment of this goal of "ideal balance" in the case of the whole BNC.

Term\ **corpus**

Trans\ корпус

Def\ Body of texts.

Down\ opportunistic corpus

Down\ saturated corpus

Down\ balanced corpus

Down\ Reference corpus

Down\ annotated corpus

Down\ tagged corpus

Down\ unannotated corpus

Down\ raw corpus

Down\ parsed corpus

Down\ treebank

Down\ phonetically transcribed corpus

Down\ monolingual corpus

Down\ multilingual corpus

Down\ monitor corpus

Down\ finite corpus

Down\ samples corpus

Down\ whole text corpus

Down\ whole text corpus

Down\ synchronic corpus

Down\ diachronic corpus

Cyt\ **Corpora** are sources of quantitative information beyond compare.

Cyt\ Leech (1992) argues that the **corpus** is a more powerful methodology from the point of view of the scientific method, as it is open to objective verification of results.

Cyt\ Whatever philosophical advantages we may eventually see in a **corpus**, it is the computer which allows us to exploit **corpora** on a large scale with speed and accuracy.

Cyt\ However, the notion of a **corpus** as the basis for a form of empirical linguistics is different from the examination of single texts in several fundamental ways.

Cyt\ In principle, any collection of more than one text can be called a **corpus**, (**corpus** being Latin for "body", hence a **corpus** is any body of text). But the term "**corpus**" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition.

Cyt\ We are therefore interested in creating a **corpus** which is maximally representative of the variety under examination, that is, which provides us with an as accurate a picture as possible of the tendencies of that variety, as well as their proportions.

Cyt\ Nowadays the term "**corpus**" nearly always implies the additional feature "machine-readable". This was not always the case as in the past the word "**corpus**" was only used in reference to printed text.

Cyt\ There is often a tacit understanding that a **corpus** constitutes a standard reference for the language variety that it represents. This presupposes that it will be widely available to other researchers, which is indeed the case with many **corpora** – e.g. the Brown **Corpus**, the LOB **corpus** and the London-Lund **corpus**.

Cyt\ Part-of-speech annotation is useful because it increases the specificity of data retrieval from **corpora**, and also forms an essential foundation for further forms of analysis (such as syntactic parsing and semantic field annotation).

Cyt\ Problem-oriented tagging (as described by de Haan (1984)) is the phenomenon whereby users will take a **corpus**, either already annotated, or unannotated, and add to it their own form of annotation, oriented particularly towards their own research goal.

Cyt\ In this session we will examine a few of the roles which **corpora** may play in the study of language. The importance of **corpora** to language study is aligned to the importance of empirical data. Empirical data enable the linguist to make objective statements, rather than those which are sub-

jective, or based upon the individual's own internalised cognitive perception of language.

Cyt\ It is important to note that although many linguists may use the term "**corpus**" to refer to any collection of texts, when it is used here it refers to a body of text which is carefully sampled to be maximally representative of the language or language variety.

Cyt\ A linguist who has access to a **corpus**, or other (non-representative) collection of machine readable text can call up all the examples of a word or phrase from many millions of words of text in a few seconds. Dictionaries can be produced and revised much more quickly than before, thus providing up-to-date information about language. Also, definitions can be more complete and precise since a larger number of natural examples are examined.

Cyt\ Because a **corpus** is sampled to maximally represent the population, any findings taken from the **corpus** can be generalised to the larger population. Hence quantification in **corpus** linguistics is more meaningful than other forms of linguistic quantification because it can tell us about a variety of language, not just that which is being analysed.

Cyt\ Most European languages (not to mention Chinese, Japanese, Korean etc.) now have some sort of **corpus** already and there is a growing awareness that a good **corpus** can be put to many uses; hence their importance grows. Despite initial disapprovals voiced by some linguists, doubts are dispelled by obvious and indisputable facts: nobody has ever been able to manually collect and subsequently process so much data in his or her lifetime as the computer can in a very short time.

Cyt\ It may still be premature to try to mark out exhaustively what **corpora** may do for language studies and linguists; undoubtedly, many new options are still to come while the appetite of linguists is gradually whetted and new ways of **corpus** exploitation are offered by **corpus** linguists. In fact, it is hard to see a linguistic discipline not being able to profit from a **corpus** one way or another, both written and oral. It is increasingly clearer that new ways and methods for retrieving information from **corpora** will have to be given more thought.

Приложение 4

Миникорпус корпусной терминологии (фрагмент)

Термин	Цитата
Corpus	**Corpora** are sources of quantitative information beyond compare.
Corpus	Leech (1992) argues that the **corpus** is a more powerful methodology from the point of view of the scientific method, as it is open to objective verification of results.
Corpus	Whatever philosophical advantages we may eventually see in a **corpus**, it is the computer which allows us to exploit **corpora** on a large scale with speed and accuracy.
Corpus	However, the notion of a **corpus** as the basis for a form of empirical linguistics is different from the examination of single texts in several fundamental ways.
Corpus	In principle, any collection of more than one text can be called a corpus, (**corpus** being Latin for "body", hence a **corpus** is any body of text). But the term "**corpus**" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition.
Corpus	We are therefore interested in creating a **corpus** which is maximally representative of the variety under examination, that is, which provides us with an as accurate a picture as possible of the tendencies of that variety, as well as their proportions.
Corpus	Nowadays the term "**corpus**" nearly always implies the additional feature "machine-readable". This was not always the case as in the past the word "**corpus**" was only used in reference to printed text.
Corpus	There is often a tacit understanding that a **corpus** constitutes a standard reference for the language variety that it represents. This presupposes that it will be widely available to other researchers, which is indeed the case with many **corpora** - e.g. the Brown **Corpus**, the LOB corpus and the London-Lund **corpus**.
Corpus	Part-of-speech annotation is useful because it increases the specificity of data retrieval from **corpora**, and also forms an essential foundation for further forms of analysis (such as syntactic parsing and semantic field annotation).
Corpus	Problem-oriented tagging (as described by de Haan (1984)) is the phenomenon whereby users will take a **corpus**, either already annotated, or unannotated, and add to it their own form of annotation, oriented particularly towards their own research goal.

Corpus	In this session we will examine a few of the roles which **corpora** may play in the study of language. The importance of **corpora** to language study is aligned to the importance of empirical data. Empirical data enable the linguist to make objective statements, rather than those which are subjective, or based upon the individual's own internalised cognitive perception of language.
Corpus	It is important to note that although many linguists may use the term "**corpus**" to refer to any collection of texts, when it is used here it refers to a body of text which is carefully sampled to be maximally representative of the language or language variety.
Corpus	A linguist who has access to a **corpus**, or other (non-representative) collection of machine readable text can call up all the examples of a word or phrase from many millions of words of text in a few seconds. Dictionaries can be produced and revised much more quickly than before, thus providing up-to-date information about language. Also, definitions can be more complete and precise since a larger number of natural examples are examined.
Corpus	Grammatical (or syntactic) studies have, along with lexical studies, been the most frequent types of research which have used **corpora**.
Corpus	Because a **corpus** is sampled to maximally represent the population, any findings taken from the **corpus** can be generalised to the larger population. Hence quantification in **corpus** linguistics is more meaningful than other forms of linguistic quantification because it can tell us about a variety of language, not just that which is being analysed.
Corpus	Most European languages (not to mention Chinese, Japanese, Korean etc.) now have some sort of **corpus** already and there is a growing awareness that a good **corpus** can be put to many uses; hence their importance grows. Despite initial disapprovals voiced by some linguists, doubts are dispelled by obvious and indisputable facts: nobody has ever been able to manually collect and subsequently process so much data in his or her lifetime as the computer can in a very short time.
Corpus	It may still be premature to try to mark out exhaustively what **corpora** may do for language studies and linguists; undoubtedly, many new options are still to come while the appetite of linguists is gradually whetted and new ways of **corpus** exploitation are offered by **corpus** linguists. In fact, it is hard to see a linguistic discipline not being able to profit from a **corpus** one way or another, both written and oral. It is increasingly clearer that new ways and methods for retrieving information from **corpora** will have to be given more thought.

Corpus	Since any language needs a consistent, perpetual and next-to-exhaustive coverage of its data, it should have a **corpus** of corresponding qualities, although in practice it is a gradual business of taking many minor decisions in the course of its construction and maintenance. This is particularly important in the case of small languages, which, unlike English and other languages, cannot afford the luxury of having a variety and multitude of **corpora** for specific purposes, at least not at the moment. What is really needed is a steady increase and perpetual growth of even, by present standards, very large **corpora** of billions of words, which should be as much representative as possible.
Corpus	Although the degree of the coverage of language by a large **corpus** is considerable, it is by no means true that today's **corpora** reflect language as a whole. Moreover, some **corpus** linguists are becoming more and more susceptible to another challenge here, namely the degree of representativeness of this coverage, which is very much an open issue and matter of much dispute.
Corpus	As information is to be found coming from all fields of human life and activity, it is hard to imagine that **corpora** can be based on a collection of, perhaps, newspapers only. On the other hand, this diversity of sources suggests that a mapping of proportions in which various kinds of information occur should take place and be reflected in the design and structure of the **corpus**, should this be a general type of **corpus**. This raises the problem of the **corpus** representativeness, mentioned above.
Corpus	More generally, one may wonder where this trend actually fits in, in an attempt to pursue purely practical and utilitarian goals, or in one aiming at an exhaustive, systematic and non-eclectic description of one's language. **Corpora** definitely offer the latter possibility.
Corpus	**Corpora** are cross-sections of a discourse universe comprising all communication acts. The texts they monitor are principally transient communication acts.
Corpus	It is the task of the linguist to define and delimit the scope of the discourse universe she or he is interested in in such a way that it can be reduced to a **corpus**. Parameters can be language, time segment, region, situation, external and internal properties of texts, and many others.
Corpus collection	**Corpus** collection continued and diversified after the diary studies period: large sample studies covered the period roughly from 1927 to 1957 - analysis was gathered from a large number of children with the express aim of establishing norms of development.
Early corpus linguistics	All the work of early **corpus** linguistics was underpinned by two fundamental, yet flawed assumptions: The sentences of a natural language are finite. The sentences of a natural language can be collected and enumerated.

Содержание

Учебное издание

Виктор Павлович Захаров

КОРПУСНАЯ ЛИНГВИСТИКА

Учебно-методическое пособие

Зав. редакцией *Г.И. Чередниченко*

Редактор *Н.Г. Михайлова*

Технический редактор *Л.Н. Иванова*

Обложка *А.В. Калининой*

Подписано в печать с оригинала-макета 28.11.2005.
Ф-т 60x84/16. Усл. печ. л. 2,79. Уч.-изд. л. 1,83.

Тираж 120 экз. Заказ № .

РОПИ С.-Петербургского государственного университета.
199034, С.-Петербург, Университетская наб., 7/9.

Типография Издательства СПбГУ.
199061, С.-Петербург, Средний пр., 41.