Лингвистический корпус как дидактический ресурс

Актуальность темы

- Любое исследование, осуществляемое лингвистом, должно быть ориентировано, по меньшей мере, на следующие этапы деятельности:

  1) выбор принципов и оснований («эталонов») классификации изучаемых объектов;

  2) процесс распределения объектов по классам в соответствии с этими основаниями («эталонами»);

  3) осмысление, интерпретация, истолкование результатов распределения объектов по классам, *объяснение причин такого распределения* [Мельников, 2003. С. 29].
- При этом первый этап данной деятельности подразумевает наличие «изучаемых объектов», т. е. сбор эмпирического материала для построения назавершающем этапе исследования теории.
- В настоящее время все большую популярность при сборе и анализе практического материала приобретает корпусная лингвистика. И это естественный шаг в лингвистике вслед за стремительным развитием информационных технологий.

Что такое лингвистический корпус

- Корпусная лингвистика появилась в 60-е гг. XX в., преимущественно на материале английского языка, но очень быстро начали возникать корпуса на базе и других языков. В Брауновском Бниверситете США в 1963 г. учеными У. Н. Френсисом и Г. Кучерой был создан первый корпус текстов на электронном носителе (Брауновский корпус, свободный доступ с сайта университета Лидс: http://corpus.leeds.ac.uk/ protected/).
- В нем содержалось 500 текстов 15 самых популярных жанров англоязычной прозы США по 2 000 слов в каждом. К корпусу прилагались указатель частотности и алфавитно-частотный указатель, а также некоторые статистические распределения.
- Корпусом считается собрание текстов одного или нескольких языков, связанных между собой определенными параметрами.
- Корпус представляет собой собрание письменных и устных высказываний. Данные корпуса, как правило, оцифровываются, т. е. хранятся на компьютерах и доступны в электронном виде. При этом составные части корпуса, тексты, состоят из данных, а также, возможно, из метаданных, описывающих эти данные, и из лингвистических аннотаций, которые эти данные упорядочивают.

Корпусная лингвистика как раздел языкознания

- Корпусная лингвистика как отдельный раздел языкознания окончательно сформировалась в первой половине 90-х гг. XX в. В это же время начал оформляться и понятийный аппарат. Так, Дж. Синклер описывает корпус как «a collection of naturally-occurring language text, chosen to characterize a state of variety of a language» [Sinclair, 1991. P. 171].
- В данном определении подчеркивается один из основополагающих принципов при выборе текстов для построения корпуса – речь идет о *неотредактированных текстах, т. е.* язык представлен в том виде, в котором он проявил себя в речи (будь то речь устная или письменная). Кроме того, в корпусе представлены не существующие «образцы» и «предписания» для правильного построения сообщения, а как можно большее количество *«вариантов» языка, пусть некоторые* из них и находятся на периферии языковой системы.
- В последующие годы понятие «корпус» все больше конкретизируется: На наш взгляд, наиболее полное определение понятия «корпус» можно найти у В. П. Захарова. Исследователь говорит о корпусе как о большом, представленном в электронном виде, структурированном и размеченном, филологически представительном массиве языковых данных, предназначенных для решения определенных лингвистических задач (см.: [Захаров, 2005. С. 3]). Данное определение можно охарактеризовать как «функциональное», в общих чертах описывающее лингвистическую направленность упорядоченных массивов текстов.

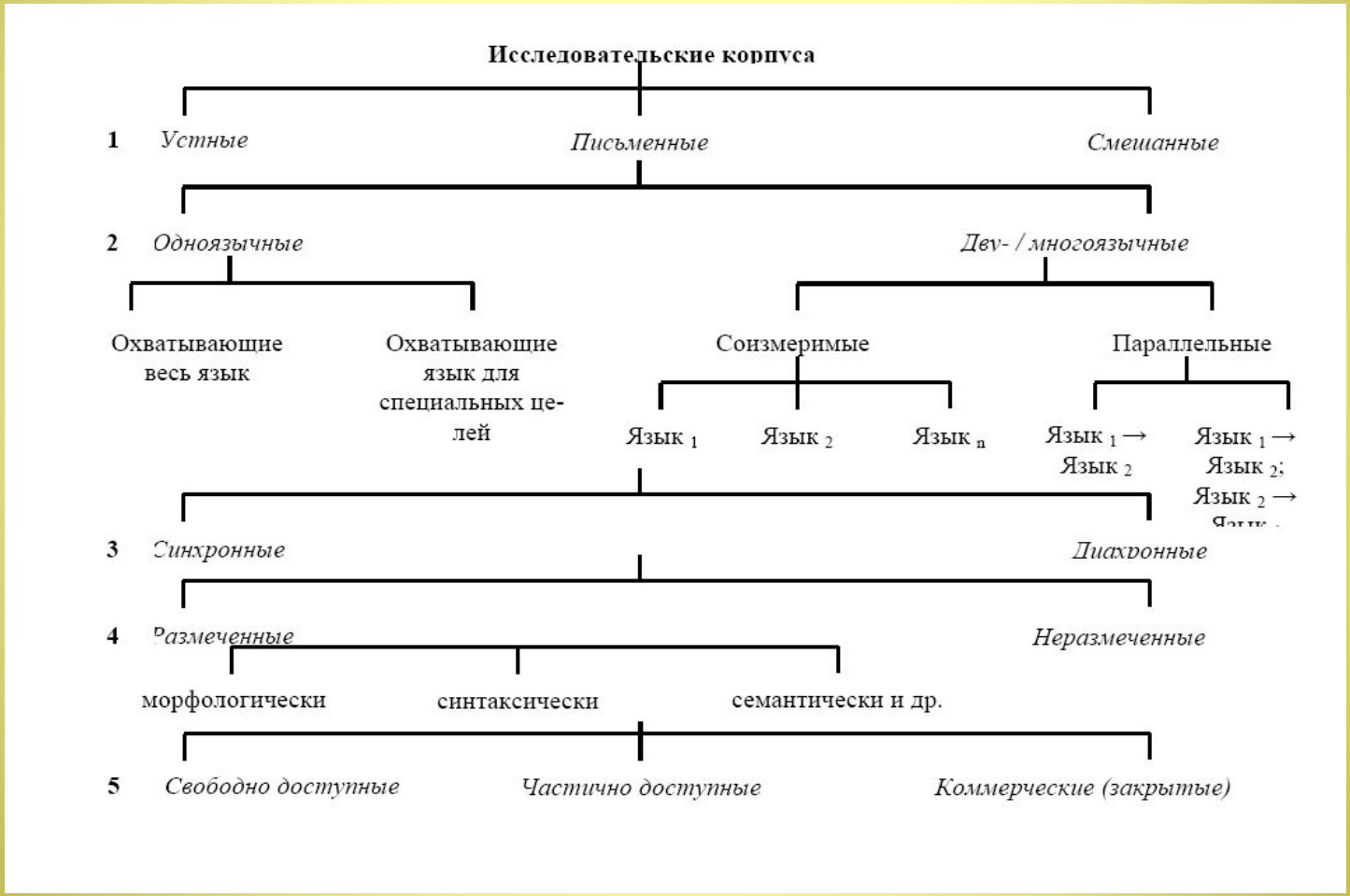
Основные свойства корпуса

1) множество текстов должно быть представлено в электронном виде (в сети Интернет или на диске);

2) языковые данные должны быть размечены для анализа в лингвистических целях;

3) в результате проведенного анализа должна существовать возможность различного распределения полученного языкового материала (по жанровой принадлежности, году создания текста, тематике и т. п.).

Классификация корпусов



Лингвистические корпуса русского языка

Лингвистические корпуса

Название	Состав	Доступ	Разметка
Русский язык			
Национальный корпус русского языка, http://www.ruscorpora.ru	Более 500 млн слов. Кроме основного корпуса содержит газетный, параллельный, диалектный, поэтический, обучающий, устной речи, акцентологический и мультимедийный (пополняется)	Свободно доступный, оффлайновая версия недоступна, однако для свободного пользования предоставляется случайная выборка предложений из корпуса со снятой омонимией объемом 180 тыс. словоупотреблений	Морфологическая (для 6 млн слов со снятой морфологической омонимией), морфо-синтаксическая со снятой омонимией
Хельсинкский аннотированный корпус русских текстов ХАНКО, http://www.ling.helsinki.fi/projects/hanco/	Содержит тексты журнала «Итоги» (пополняется)	Свободно доступный	Морфологическая и синтаксическая
Машинный фонд русского языка, http://cfrl.ru/	Содержит тексты русской прозы, поэзии и драматургии XIX–XX вв., подкорпус текстов российских газет 90 гг. XX в., произведения русских историков XIX–XX вв., а также подкорпус по фольклору (русские народные сказки А. Н. Афанасьева)	Свободно доступный	Морфологическая (частично)
Regensburg Russian Diachronic Corpus (RRuDi), http://rhss11.uni-regensburg.de/SlavKo/korpus/rrudi-new/	Содержит тексты на церковнославянском и древнерусском языках (пополняется)	Свободный доступ для выполнения исследовательских задач предоставляется после подписания лицензионного соглашения	Морфосинтаксическая (большинство текстов проверено вручную)

Лингвистические корпуса английского языка

Название	Состав	Доступ	Разметка
Английский язык			
BYU-BNC: British National Corpus, созданный Марком Дэвисом, http://corpus.byu.edu/bnc/	100 млн слов британского варианта английского языка (1980–1993 гг.)	Свободный доступ для выполнения исследовательских задач после несложной процедуры регистрации на сайте	Морфологическая (можно искать конкретную словоформу, все формы одной лексемы по возможным начальным формам, словосочетания, выбранные грамматические формы лексемы)
Corpus of Contemporary American English (COCA), созданный Марком Дэвисом, http://corpus.byu.edu/coca/	Более 450 млн слов американского варианта английского языка (1990–2012 гг.). Содержит в одинаковых пропорциях тексты разговорной речи (скрипты более чем 150 ТВ- и радиопередач), художественной литературы, публицистики (популярные журналы и газеты), а также тексты академических журналов (пополняется)	Свободный доступ для выполнения исследовательских задач после несложной процедуры регистрации на сайте	Морфологическая (можно искать конкретную словоформу, все формы одной лексемы по возможным начальным формам, словосочетания, выбранные грамматические формы лексемы)
Corpus of Historical American English (COHA), созданный Марком Дэвисом, http://corpus.byu.edu/coha/	Более 400 млн слов американского варианта английского языка (1810–2009 гг.). Содержит тесты художественной литературы и публицистики	Свободный доступ для выполнения исследовательских задач после несложной процедуры регистрации на сайте	Морфологическая (можно искать конкретную словоформу, все формы одной лексемы по возможным начальным формам, словосочетания, выбранные грамматические формы лексемы)
Bank of English, http://www.collinslanguage.com/content-solutions/wordbanks	Более 553 млн слов различных вариантов английского языка, сбалансировано по разным жанрам (пополняется)	Коммерческий, пробная версия предоставляется бесплатно на один месяц после процедуры регистрации	Частеречная с элементами морфологической

Лингвистические корпуса немецкого языка

Название	Состав	Доступ	Разметка
Brown Corpus, http://corpus.leeds.ac.uk/protected/	Первый представительный корпус. Состоит из 500 прозаических фрагментов в 2 000 слов, взятых из текстов, опубликованных в США в 1961 г.	Свободно доступный с сайта университета Лидс (100 примеров использования)	Морфологическая и синтаксическая
Немецкий язык			
Мангеймский корпус немецкого языка, DeReCo, http://www.ids-mannheim.de/kl/projekte/korpora/	Самый представительный корпус немецкого языка, поддерживаемый Институтом немецкого языка (Мангейм). Более 5,4 млрд слов. Содержит тексты художественной, научной и научно-популярной литературы, периодики, а также подкорпус устной речи	Свободно доступный после регистрации на сайте и подписания лицензионного соглашения. Требуется установка специальной программы – оболочки COSMAS II	Частичная морфологическая. Можно искать конкретную словоформу, все формы одной лексемы по возможным начальным формам, словосочетания
LIMAS, http://korpora.zim.uni-duisburg-essen.de/Limas/	Более 1 млн словоупотреблений. Состоит из 500 текстов 33 различных рубриках	Свободно доступный	Поиск по слову, контексту, фразе
Корпус Берлинско-Бранденбургской Академии наук DWDS, http://www.dwds.de	Около 1,8 млрд слов. Содержит тексты художественной литературы XX–XXI вв., периодики (Berliner Zeitung, Bild, Süddeutsche Zeitung, Tagesspiegel, WELT, ZEIT), устной речи и др. В разработке корпус текстов 1650–1900 гг.	Свободно доступный после регистрации на сайте	Можно искать конкретную словоформу, все формы одной лексемы по возможным начальным формам, словосочетания
Многоязычные корпуса			
TITUS, http://titus.uni-frankfurt.de/indexe.htm	Тезаурус материалов по индоевропейским языкам (древнее, среднее и для ограниченного количества языков современное состояние)	Свободно доступный. Тексты доступны для поиска, просмотра и скачивания	Возможен поиск по грамматическим формам слова
European Parliament Proceedings Parallel Corpus, http://www.statmt.org/europarl	Корпус слушаний парламента (1996–2011 гг.). Тексты на всех языках европейского парламента	Свободно доступный для скачивания	–

Доступность корпусов

- Существенным критерием выступает *доступность корпуса текстов в электрон*ном виде. Все существующее множество корпусов текстов можно разделить на три обширные категории:
  1) находящиеся в свободном доступе;
  2) находящиеся в частичном доступе
  3) коммерческие.
- К первой категории относится довольно ограниченное количество из существующих на данный момент корпусов текстов. Наиболее обширным (общим объемом более 500 млн слов) является Национальный корпус русского языка (www.ruscorpora.ru).
- Большинство из существующих корпусов относится ко второй категории, однако для решения конкретных лингвистических задач такой частичный доступ является чаще всего достаточным. Так, в Британском национальном корпусе (http://www.natcorp.ox.ac.uk/) выдача результата ограничена 50 случайными примерами, кроме того, отсутствуют многие возможности поискового интерфейса, поставляемого вместе с полной (платной) версией корпуса.
- Наряду с этим существует некоммерческая версия данного корпуса (http://corpus.byu.edu/bnc/), доступная после несложной процедуры регистрации, в которой для поиска представлено 100 млн слов в текстах 1980–1993 гг. Довольно представительная подборка из Мангеймского корпуса немецкого языка (http://www.ids-mannheim.de/kl/projekte/korpora/) доступна также после процедуры регистрации
- и установки специальной программы (оболочки COSMAS II). К третьей группе можно отнести, например, Банк английского языка (Bank of English) с возможностью пробной бесплатной подписки на один месяц для получения доступа в Collins Wordbanks Online (553 млн слов) (http://www.collinslanguage.com/content-solutions/wordbanks), после чего необходимо приобрести платную версию корпуса.

Разметка корпуса

- Следующим существенным признаком лингвистического корпуса текстов является наличие или отсутствие *разметки, так как* для решения лингвистических задач наличия простого массива текстов недостаточно.
- Под разметкой понимается приписывание текстам и их компонентам специальных меток: внешних, экстралингвистических, структурных и собственно лингвистических, описывающих лексические, грамматические и прочие характеристики элементов текста [Захаров, 2005. С. 6]. Метаразметка включает в себя сведения об авторе и о самом тексте. Рассмотрим собственно лингвистические виды разметки на примере некоторых из существующих корпусов. Остановимся, прежде всего, на *морфологической (или час*теречной) разметке. Данный вид разметки является наиболее распространенным в существующих корпусах, при этом учитывается не только признак части речи, но и признаки грамматических категорий.
- Морфологическая разметка осуществляется с помощью специальных программ автоматического морфологического анализа. Например, в небольшой части Национального корпуса русского языка (объемом 6 млн словоупотреблений) произведено ручное снятие морфологической омонимии и дополнительная коррекция результатов работы программы автоматического морфологического анализа. «Эта часть образует так называемый корпус со снятой омонимией, который может служить удобным полигоном для тестирования различных программ поиска, морфологического анализа и автоматической обработки текстов, а также для исследований современной русской морфологии, требующих повышенной точности поиска» (см.: [http://ruscorpora.ru/corporastructure. html]).
- В Британском национальном корпусе, как и в Банке английского языка, также представлены метатекстовая и морфологическая разметки. В Мангеймском корпусе немецкого языка морфологическая разметка присутствует в основном в подкорпусах публицистических текстов. Среди других видов разметки особо следует выделить *синтаксическую, которая представлена* не во всем массиве корпуса (Национального корпуса русского языка, Мангеймского корпуса немецкого языка), а только в его небольшой части, так как данный вид разметки, подразумевающий указание синтаксической структуры для каждого предложения, осуществляется фактически вручную и требует огромных временных затрат.
- Кроме того, в корпусе могут присутствовать и другие виды разметки, такие как семантическая, просодическая, анафорическая, графематическая и др. – все это во многом позволяет облегчить процесс непосредственного сбора материала исследователем при условии правильно заданных критериев поиска.
- Однако, чтобы созданный корпус текстов удовлетворял различным лингвистическим задачам, стоящим перед исследователем языка, он должен также обладать еще по меньшей мере двумя признаками.

Репрезентативность корпуса

- Прежде всего, речь идет о так называемой *репрезентативности корпуса текстов.* По мнению А. Е. Кибрика, М. М. Брыкиной, А. П. Леонтьева и А. Н. Хитрова, репрезентативность можно оценить «по изменению относительной частоты рассматриваемого явления при увеличении выборки. Если относительная частота явления от рибавления каждого последующего фрагмента текста будет изменяться все меньше и меньше, то это означает, что корпус в целом репрезентативен» [Кибрик и др., 2006. С. 21].
- При этом хоть и отмечается невозможность при определения репрезентативности корпуса текстов. В целом, вопрос определения репрезентативности того или иного корпуса текстов является по сей день актуальным, однако, к сожалению, недостаточно разработанным.
- Именно репрезентативность превращает обычный *набор разнообразных тек*стов непосредственно в *корпус текстов,* пригодный для проведения лингвистического исследования. Однако языковая деятельность человека настолько разнообразна, что чрезвычайно трудно объективно отразить все существующие «*варианты» языка, о* которых мы уже упоминали выше.
- Вследствие этого вопрос репрезентативности корпуса текстов является скорее вопросом из области объективности любого научного исследования. Здесь следует опираться на здравый смысл самого исследователя, если речь идет о пользовательском корпусе (создается самим исследователем в зависимости от целей его исследования), либо группы исследователей, если речь идет о создании корпуса, претендующего на всеохватность языковых явлений, стилей, жанров и т. п. (например, национального корпуса определенного языка).

Простота корпуса

- Немаловажным критерием при определении корпуса выступает также и *простота* его использования, другими словами, корпус должен быть обеспечен специализированной поисковой системой, которая должна быть (в идеальном случае) довольно понятна и проста в использовании.
- Так, предлагаемая поисковая система в Мангеймском корпусе немецкого языка довольно сложна в использовании, в то время как при использовании Национального корпуса русского языка, Британского национального корпуса и Банка английского языка особых трудностей не возникает.
- Корпус должен сокращать количество времени, необходимое на поиск конкретного явления, а не предлагать сложный алгоритм этого поиска, ознакомление с основными пунктами которого требует от исследователя-лингвиста подчас чисто технических и математических знаний.

- **Благодарю за внимание!**