УДК 81:811.1/.2

Н. В. Козлова

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

E-mail: talja@ngs.ru

ЛИНГВИСТИЧЕСКИЕ КОРПУСА: ОПРЕДЕЛЕНИЕ ОСНОВНЫХ ПОНЯТИЙ И ТИПОЛОГИЯ

Рассматриваются основные понятия корпусной лингвистики, приводится типология корпусов. В приложении содержится краткий анализ основных существующих корпусов на материале русского, английского и немецкого языков.

Ключевые слова: корпус, корпусная лингвистика, разметка, репрезентативность, сбалансированность.

Определение понятия «корпус»

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 на завершающем 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.

Корпусная лингвистика появилась в 60-е гг. XX в., преимущественно на материале английского языка, но очень быстро начали возникать корпуса [1] на базе и других языков. В Брауновском университете 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. «Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten des Korpus sind typischerweise digitalisiert, d. h. auf Rechnern gespeichert und maschinenlesbar» [Lemnitzer, Zinsmeister, 2006. S. 7] – Корпус

[1] В корпусной лингвистике принято использование формы множественного числа «корпуса». См.: Толковый словарь русского языка под редакцией Д. Н. Ушакова: Корпус, *мн. ч.* корпуса. 10. Полное собрание, цельный свод каких-нибудь текстов.

ISSN 1818-7935
Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2013. Том 11, выпуск 1

представляет собой собрание письменных и устных высказываний. Данные корпуса, как правило, оцифровываются, т. е. хранятся на компьютерах и доступны в электронном виде (перевод наш. – *Н. К.*). 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 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: «A corpus is a collection of texts, designed for some purpose, usually teaching or research. […] A corpus is not something that a speaker does or knows, but something constructed by a researcher. It is a record of performance, usually of many different users, and designed to be studied, so that we can make inferences about typical language use. Because it provides methods of observing patterns of a type which have long been sensed by literary critics, but which have not been identified empirically, the computer-assisted study of large corpora can perhaps suggest a way out of the paradoxes of dualism» [Stubbs, 2001. P. 239–240].

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.

Таким образом, в каждом из представленных определений понятия «корпус» подчеркивается следующее:

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.).

Если рассматривать первый пункт, то здесь существенным критерием выступает *доступность* корпуса текстов в электронном виде. Все существующее множество корпусов текстов можно разделить на три обширные категории: 1) находящиеся в свободном доступе; 2) находящиеся в частичном доступе и 3) коммерческие. К первой категории относится довольно ограниченное количество из существующих на данный момент корпусов текстов (см. таблицу). 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.

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 частеречной) 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/corpora-structure.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.

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 прибавления 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 такой трактовке репрезентативности установить связи со статистикой, подчеркивается, что данное условие является необходимым, но все же недостаточным для 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 текстов 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).

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. На наш взгляд, 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.

Типология корпусов

Среди существующего многообразия исследовательских корпусов подчас очень сложно ориентироваться, ведь цели и задачи, стоящие перед лингвистом, очень часто совпадают в общем, но разнятся в отдельных отраслях и областях. Правильный выбор соответствующего корпуса – это первый шаг, который должен осуществить исследователь при анализе «изучаемых объектов». Все разнообразие существующих корпусов «определяется многообразием исследовательских и прикладных задач, для решения которых они создаются» [Захаров, 2005. С. 12], и может быть представлено следующей схемой (см. рисунок).

1. Устные – письменные – смешанные.

Под *устным* корпусом подразумевается структурированная совокупность речевых фрагментов, которая обеспечена программными средствами доступа к ним [Кривнова, 2006]. Первые устные корпуса появились в начале 80-х гг. прошлого века на материале американского варианта английского языка. Позже возникли специальные координационные центры по сбору, хранению, распространению и созданию устных корпусов. Например, LDC (Linguistic Data Consorcium, http://www.ldc.upenn.edu), CSLU (Center for Spoken Language Understanding, http://www.cslu.ogi.edu), ELRA (European Language Resources Association, http://www.elra.info).

Большинство из существующих корпусов относятся к *письменным* (например, находящаяся в свободном доступе часть Мангеймского корпуса немецкого языка, http://www.ids-mannheim.de) либо *смешанным*, однако доля лингвистически размеченных устных текстов даже в смешанных корпусах (чаще всего это национальные корпуса какого-либо языка, например: русского языка – http://www.ruscorpora.ru, американского варианта английского языка – http://corpus.byu.edu/coca, английского языка – http://corpus.byu.edu/bnc) ничтожно мала по отношению ко всему массиву корпуса.

2. Одноязычные – двуязычные / многоязычные.

Среди одноязычных корпусов можно выделить две группы: с одной стороны, корпуса, *охватывающие весь язык*, с другой – охватывающие только *язык для специальных целей*. Например, Corpus of Early English Medical Writing (CEEM) (подробнее см.: http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/index.html) – корпус медицинских текстов на английском языке 1375–1750 гг. общим объемом около 1,5 млн слов, в котором содержатся теоретические работы, справочники, стихотворные тексты на медицинские темы.

В двуязычных и многоязычных корпусах тексты могут быть представлены *соизмеримо* или *параллельно*. Так, в 1992 г. была создана Европейская корпусная инициатива (European Corpus Initiative (ECI)) – международная организация, занимающаяся созданием огромного многоязычного корпуса для научных целей (см.: http://www.elsnet.org/resources/eciCorpus.html). В данном соизмеримом корпусе содержатся в основном тексты европейских языков, а также тексты на таких языках, как турецкий, китайский, японский, русский и др., общим объемом более 98 млн слов, данный корпус носит коммерческий характер. Корпуса парал-

Лингвистические корпуса

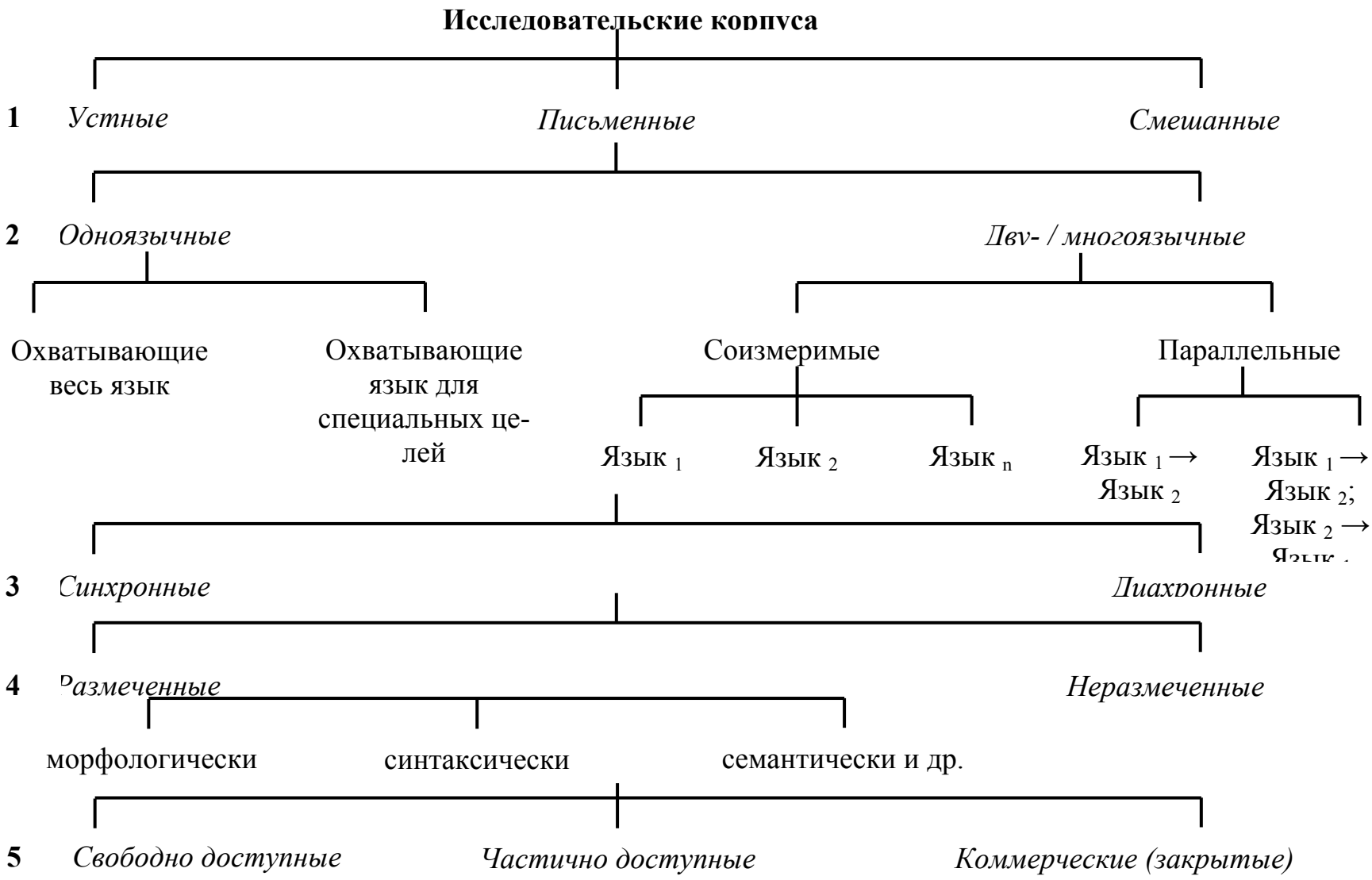
Название	Состав	Доступ	Разметка
Русский язык			
Национальный корпус русского языка, http://www.ruscorpora.ru	Более 500 млн слов. Кроме основного корпуса содержит газетный, параллельный, диалектный, поэтический, обучающий, устной речи, акцентологический и мультимедийный (пополняется)	Свободно доступный, 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://rhssl1.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)

Продолжение таблицы

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

Окончание таблицы

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	–


Исследовательские корпуса
1
Устные
Письменные
Смешанные
2
Одноязычные
Дву- / многоязычные
Охватывающие весь язык
Охватывающие язык для специальных целей
Соизмеримые
Параллельные
Язык 1
Язык 2
Язык n
Язык 1 → Язык 2
Язык 1 → Язык 2; Язык 2 →
3
Синхронные
Диахронные
4
Размеченные
Неразмеченные
морфологически
синтаксически
семантически и др.
5
Свободно доступные
Частично доступные
Коммерческие (закрытые)

лельных текстов предназначены, в первую очередь, для сопоставительного анализа текстов «оригинал – перевод» в целях обучения методам и приемам перевода. Удачный пример такого вида корпусов – European Parliament Proceedings Parallel Corpus 1996–2011 (http://www.statmt.org/europarl), где представлены параллельные тексты заседания Европейского парламента на всех европейских языках с переводом на английский.

3. Синхронный – диахронный.

Синхронные корпуса предполагают представление текстового материала для рассмотрения состояния языка как системы в определенный момент времени. В частности, в некоммерческой версии Британского национального корпуса (http://corpus.byu.edu/bnc) представлены лишь тексты конца XX в. – с 1980 по 1993 г.

Для рассмотрения исторического развития какого-либо языкового явления либо всей языковой системы в целом существуют *диахронные* корпуса. Например, Thesaurus Indogermanischer Text- und Sprachmaterialien (http://titus.uni-frankfurt.de), в котором представлены индогерманские тексты различных эпох.

4. Неразмеченные – размеченные.

Неразмеченный корпус – это массив текстов, которые содержат определенное количество упоминаний искомого элемента. При этом результаты поиска, предоставляемые в неразмеченных корпусах, могут быть использованы в лингвистических исследованиях, но только с чисто статистической точки зрения.

Размеченные (морфологически, синтаксически, просодически и др., см. выше) корпуса предоставляют намного больше возможностей для проведения лингвистического анализа.

Выводы

Таким образом, корпус – это представленный в электронном виде, как правило, размеченный для анализа в лингвистических целях, обеспеченный сравнительно простой в использовании поисковой системой репрезентативный массив неотредактированных текстов, представляющих как можно большее количество «вариантов» языка.

В период зарождения корпусной лингвистики вопросов компьютеризации данного направления не ставилось, и «исследователи указывали на возможность пренебречь языковой вариативностью, т. е. территориальной, социальной, профессиональной, возрастной, гендерной, индивидуальной и тому подобной дифференциацией языка» [Плунгян, 2006. С. 76–77]. Сегодня же, пренебрегая ею, мы сознательно ограничиваем себя различными рамками при изучении текстов определенного языка, что ставит под вопрос объективность подобного рода исследования. С появлением электронных корпусов многообразие форм существования языка стало более наглядным, возможности исследования языковых данных расширились. Современный лингвистический корпус содержит сотни миллионов словоупотреблений, а то, что с помощью электронного корпуса результаты примеров словоупотреблений можно получить за считанные доли секунд, существенно упрощает задачу лингвистам. Представленная типология корпусов, не претендуя на всеохватность, показывает нам существующее многообразие корпусов текстов и позволяет сориентироваться в нем для последующего проведения научного исследования.

Список литературы

Захаров В. П. Корпусная лингвистика. СПб., 2005.

Кибрик А. Е., Брыкина М. М., Леонтьев А. П., Хитров А. Н. Русские посессивные конструкции в свете корпусно-статистического исследования // Вопросы языкознания. 2006. Вып. 1. С. 16–45.

Кривнова О. Ф. Области применения речевых корпусов и опыт их разработки // Тр. XVIII Сессии Российского акустического общества РАО. Таганрог, 2006.

Мельников Г. П. Системная типология языков: Принципы, методы, модели / Отв. ред. Л. Г. Зубкова. М.: Наука, 2003.

Плунгян В. А. «Интегрум» и Национальный корпус русского языка в лингвистических исследованиях // Integrum: точные методы и гуманитарные науки. М., 2006. С. 76–84.

Рыков В. В. Прагматически ориентированный корпус текстов // Тверской лингвистический меридиан. Тверь, 1999. Вып. 3. С. 89–96.

Толковый словарь русского языка: В 4 т. / Под ред. Д. Н. Ушакова. М., 1935; Т. 1. М., 1938 Т. 2;. М., 1939 Т. 3; М., 1940 Т. 4. Репринтное издание: М., 2000. URL: http://ushakovdictionary.ru/

Lemnitzer L., Zinsmeister H. Korpuslinguistik: Eine Einführung. Tübingen, 2006.

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

Stubbs M. Words and phrases: corpus studies of lexical semantics. Oxford, 2001.

Материал поступил в редколлегию 01.02.2013

N. V. Kozlova

LINGUISTIC CORPUS: TYPOLOGY AND TERMS

The article deals with the theoretical aspects of corpus linguistics. The author presents an analysis of the declared problem and a systematization of existing theories, terms and concepts. It includes examples of Russian, English and German linguistic corpora.

Keywords: linguistic corpus, corpus linguistics, annotation, representativeness, balance.