Михаил Копотев

ВВЕДЕНИЕ В КОРПУСНУЮ ЛИНГВИСТИКУ

Учебное пособие для студентов филологических и лингвистических специальностей университетов

Прага
2014

Михаил Копотев

Введение в корпусную лингвистику: Учебное пособие для студентов филологических и лингвистических специальностей университетов

«Animedia»

2014

Копотев М. В.

Введение в корпусную лингвистику: Учебное пособие для студентов филологических и лингвистических специальностей университетов / М. В. Копотев — «Animedia», 2014

ISBN 978-80-7499-067-0

Пособие содержит основные сведения о корпусной лингвистике – одном из самых популярных разделов современного языкознания, целью которого является создание и использование языковых корпусов в лингвистических исследованиях. В учебнике на доступном уровне рассматриваются основы данной предметной области, перечисляются основные корпуса разных языков, показываются возможности использования методов корпусной лингвистики, а также описан вклад корпусной лингвистики в лингвистическую теорию. Учебник предназначен для студентов филологических и лингвистических факультетов высших учебных заведений. Может быть использовано аспирантами и преподавателями смежных дисциплин.Автор благодарит The Pygos Group. A HIT Entertainment company за разрешение использовать изображение Пингу и интернет-проект «ПостНаука» за разрешение использовать видеолекцию В. А. Плунгяна.

ISBN 978-80-7499-067-0

Содержание

Михаил Копотев
Введение в корпусную лингвистику (Учебное пособие)

От автора

Вы читаете электронный учебник, который, скорее всего, никогда не будет издан на бумаге. Вероятно, он никогда не будет издан на бумаге. У такого решения есть два преимущества. Во-первых, корпусная лингвистика тесно связана с компьютером, интернетом и электронной обработкой текстов, поэтому она идеально подходит и для электронного формата обучения. Во-вторых, я надеюсь, что купить электронное издание легче и дешевле, чем бумажное. Цена на учебник символическая, примерно столько же вы бы потратили на поездку в книжный магазин. Тем не менее, эта книга не бесплатна: она стоила определенного труда мне и моим помощникам, и ваша поддержка позволит периодически выпускать обновления. Спасибо за то, что купили!

Я бесконечно благодарен моим друзьям и коллегам, помогавшим мне советом и добрым словом: Э. Клышинскому, С. Крылову, А. Кутузову, О. Невзоровой, Л. Пивоваровой, Е. Маркасовой, А. Теснеру, А. Левиту, С. Шарову, Е. Ягуновой… – сожалею, что не могу перечислить всех! Этот учебник не вышел бы в свет без деятельного участия нескольких людей. Моя безграничная признательность – профессору Хельсинкского университета Арто Мустайоки за его содержательные комментарии, а также за финансовую поддержку издания в рамках гранта «Создание частотной грамматики русского языка». Мой смиренный поклон Ольге Митрениной, доценту кафедры математической лингвистики Санкт-Петербургского университета, согласившейся стать вторым рецензентом и нещадно критиковавшей меня как за незнание предмета, так и за незнание правил русского языка. Моя благодарность Дарье Кормачёвой, моей аспирантке, выпускнице той же петербургской кафедры, за то, что она собрала библиографию, подготовила словарь и убедилась, что все задания выполнимы. Наконец, я благодарен двум людям, превратившим текст в книгу: редактору, сотруднику Института русского языка им. В. В. Виноградова Наталии Занегиной, убравшей все неточности, повторы и ошибки, и художнице Марии Заборовской, лаконично и ясно визуализировавшей мои многословные объяснения. Спасибо вам!

Естественно, все не замеченные ими ошибки остаются на моей совести, с которой можно связаться по адресу: mihail.kopotev@helsinki.fi.

Предисловие

Корпусная лингвистика – это лингвистика корпусов, то есть собраний текстов. Для начала такого «определения» вполне достаточно. Такое направление лингвистики существует чуть более полувека, а в России это, по сути, наука XXI века: ее активное развитие пришлось на самое начало третьего тысячелетия.

> О «молодости» этой дисциплины говорит, в частности, неустойчивость ударения и морфологических форм самого термина корпус и его производных: *ко́рпусы – корпуса́, ко́рпусная – корпусна́я*. По моим наблюдениям, в устной речи специалисты по корпусной лингвистике предпочитают говорить корпуса́, корпусна́я. Письменная норма менее стабильна: в пяти русскоязычных сборниках по корпусной лингвистике встретилось 24 формы *корпуса* и 27 – *корпусы*.

Говоря о корпусной лингвистике, следует иметь в виду два ее направления:

- создание корпусов,
- корпусные исследования, то есть исследование языка с помощью корпусных методов.

Четкой границы между ними не существует, и практически все создатели корпусов проводят в то же время и собственно лингвистические исследования. В целом, корпусная лингвистика в первом значении более технологична и предполагает совместную работу лингвистов и специалистов по компьютерным технологиям. Это не столько теоретическое направление лингвистики, сколько технология. Корпусная лингвистика во втором значении – дело лингвистов, в том числе и специалистов по статистической обработке языка. Говоря о корпусной лингвистике, часто имеют в виду второе значение (‘корпусные исследования’), но необходимо помнить, что без первого в принципе не существовало бы и второго. В настоящем учебнике речь пойдет обо всех составляющих корпусной работы.

Главная задача учебника – введение в новую тему, многогранную и динамичную. Я старался построить его не как путеводитель по корпусам и программам (хотя ссылок в нем немало), а как рассказ об общих особенностях этого направления современной лингвистики. Идеальный читатель этого учебника – студент-филолог, который уже прослушал курсы по грамматике и еще не успел забыть школьную математику. Я строил этот учебник так, чтобы не перегрузить его сложным материалом, но совсем обойтись без сложностей (особенно математических) невозможно. В этой книге я рассмотрю следующие темы:

- определение и особенности языкового корпуса;
- история создания и классификация корпусов;
- различные виды корпусной разметки;
- одноязычные и многоязычные корпуса;
- интернет как корпус;
- создание собственного корпуса;
- количественные методы в корпусных исследованиях;
- вклад корпусной лингвистики в общую теорию языка.

Каждая глава сопровождается списком литературы и заданиями, позволяющими закрепить навыки или расширить представление о темах, обсуждаемых в соответствующей главе. Зведочкой (*) помечены задания повышенной сложности и задания для дискуссии.

Ниже приведен список англоязычных книг, которые я рекомендую для дополнительного чтения. Первые четыре – это современные учебники и словарь терминов; четыре последние – статьи и монографии, успевшие стать классическими за недолгую историю нашей дисциплины.

1. Biber D., Conrad S., Reppen R. Corpus linguistics: Investigating language structure and use. Cambridge University Press, 1998.

2. McEnery T., Hardie A. Corpus linguistics: method, theory and practice. Cambridge University Press, 2011.

3. Xiao R., Tono Y. Corpus-based language studies: An advanced resource book. Taylor & Francis, 2006.

4. Baker P., Hardie A., McEnery T. A glossary of corpus linguistics. Edinburgh: Edinburgh University Press, 2006.

5. Sampson G., McCarthy D. (ed.). Corpus linguistics: readings in a widening discipline. Continuum: International Publishing Group, 2005.

6. Sinclair J. Corpus, concordance, collocation. Oxford University Press, 1991.

7. Stubbs M. Text and corpus analysis: Computer-assisted studies of language and culture. Oxford: Blackwell, 1996.

8. Tognini-Bonelli E. Corpus linguistics at work. John Benjamins, 2001.

Глава 1. Что такое корпус?

Латинское слово *corpus* значит «тело, туловище, единое целое». Несколько лет назад я участвовал в конференции по корпусной лингвистике, которая проходила в помещении бывшего анатомического театра. Первый же докладчик, вспомнив картину Рембрандта «Урок анатомии доктора Тульпа», отметил символичность места: корпусной лингвист тоже работает с корпусом, препарируя его с помощью специальных инструментов. Добавлю, что традиция открытого для широкой публики доступа к *корпусу* исчезла из медицинской науки, но, как мы увидим, возродилась в лингвистике в виде общедоступного корпуса, позволяющего проверять и перепроверять утверждения лингвистов о языке.

Что же такое корпус в лингвистическом смысле? Ниже я привожу два определения, первое – из старого, но хорошего учебника, второе – из Википедии.

> (1) Корпус в современной лингвистике в отличие от любого набора текстов может быть более точно определен как ограниченный по объему набор электронных текстов, собранных с целью максимально точно представлять исследуемый вариант языка (McEnery & Wilson 1996: 24).
>
> (2) Лингвистическим корпусом называют собрание текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой. Иногда корпусом («корпус первого порядка») называют просто любое собрание текстов, объединённых каким-то общим признаком (языком, жанром, автором, периодом создания текстов) (Википедия, статья «Корпусная лингвистика», 2013).

Эти определения отличаются в частностях, которые связаны с развитием корпусной лингвистики за последние десятилетия – от коллекции текстов к аннотированному комплексу. В то же время обе формулировки позволяют определить минимальные требования к корпусу.

1. Тексты, входящие в корпус, должны быть собраны по определенным принципам, чтобы представлять определенный пласт языка или весь язык в определенный период времени. Это параметр называется **репрезентативность** (англ. *representativeness*).

> **Репрезентативность** – свойство корпуса, заключающееся в статистически достоверном представлении языка или его части и достигаемое за счет необходимого объема и жанрового разнообразия текстов.

Если сохранять латинскую этимологию, то языковой корпус – это тоже «тело», единое целое языка или подъязыка. В идеале таковыми являются все тексты, и такая ситуация вполне возможна, если мы изучаем язык конкретного автора и создаем, например, корпус произведений М. В. Ломоносова (www.lomonosov.pro), в который включены все произведения из всех одиннадцати томов полного собрания его сочинений. Этот корпус текстов уже не удастся существенно расширить, так что мы можем считать его хорошим примером полного корпуса. Но что делать, если речь идет о языке XVIII века в целом? Или о языке современного русского чата?

К счастью, лингвисты выяснили, что если тексты хорошо подобраны, то они могут представлять весь язык или его определенную часть. Для этого достаточно взять большой объем текстов, который будет представлять весь язык. Конечно, ключевой вопрос здесь – что значит «достаточно большой».

> Приведу пример. Если я, зайдя в аудиторию в восемь утра, начну спрашивать «Как дела?», – ответы, вероятнее всего, будут однотипными (и не очень позитивными). Если задавать тот же вопрос разным людям в разное время суток, то позитивные ответы все-таки появятся. Мы в какой-то момент заметим, что новых вариантов больше не слышно, а частотность каждого варианта ответа не меняется. С этого момента – условно говоря, после двух тысяч ответивших – мы можем прекратить опрос. Конечно, всегда есть вероятность получить оригинальный ответ от две тысячи первого человека, но обычно ученым для дальнейших исследований достаточно составить представление об общем распределении единиц.

Примерно так же поступают и корпусные лингвисты, которые собирают не все тексты всех носителей языка, а так называемую **представительную, или репрезентативную, выборку** (англ. *representative sampling*) – такой объем материала, увеличение которого уже почти никак не повлияет на распределение единиц. Невозможно раз и навсегда определить, какой объем достаточен. Во многих случаях, особенно для лексикографической работы, корпуса объемом в 100 миллионов слов недостаточно. С другой стороны, для решения множества задач (например, морфологических) достаточно текста объемом всего в 5 тысяч слов (три главы этого учебника), и дальнейшее увеличение объем не изменит лингвистический результат.

2. Второй важной характеристикой корпуса является его **сбалансированность** (англ. *balance*); этот параметр определяет, насколько равномерно представлены тексты разных типов.

> Согласно данным Частотного словаря русского языка, изданного в 1977 году, в сотню самых частых слов входят существительное «товарищ» и прилагательное «советский». Объем корпуса, на основе которого был создан словарь, достаточно большой даже по современным меркам – 1 млн слов. Но появление этих слов «на передовых рубежах» лексического состава языка того периода объясняется тем, что использовался несбалансированный корпус:

он включал в себя только письменные тексты советского периода. Если бы корпус состоял только из разговорных текстов, то в список самых частотных, вероятно, вошли бы совсем другие слова.

Надо сказать, что сбалансированность является ахиллесовой пятой многих существующих корпусов. Очевидно, что в реальной языковой практике объем произнесенного существенно превышает объем написанного (Подумайте сами, сколько слов вы сегодня написали, а сколько произнесли.). Но для создания корпуса оказывается удобнее и проще взять существующие письменные тексты, а не собирать устные записи. Эта проблема несбалансированности хоть и медленно, но решается.

Итак, репрезентативность и сбалансированность – свойства корпуса, позволяющие адекватно представлять всё разнообразие текстов в равных или неравных, но мотивированных реальным употреблением пропорциях. Не будем при этом идеализировать ситуацию: каким бы большим ни был корпус, он всего лишь отражение языковой стихии: в реальной живой речи всегда найдутся единицы, не вошедшие в корпус.

3. В зависимости от имеющихся задач корпус может состоять из нескольких тысяч или нескольких миллионов текстоформ, но в любом случае **объем корпуса** должен быть известен (англ. *finite-sized*). Информация и об общем объеме корпуса, и о количестве извлеченных из текста примеров должна быть доступна пользователю, чтобы он мог использовать «сырые» цифры или применять более сложные формулы лингвистической статистики. В главе 16 мы еще поговорим об этом, сейчас же – один пример.

Местоимение «аз» в корпусе XVIII века встретилось 355 раз, в корпусе XIX века – 603 раза, а в корпусе XX века – 887 раз. Значит ли это, что «аз» постепенно становится все более употребительным (см. график слева)? Совсем нет. Знание объема корпусов позволяет перевести сырые данные в относительные цифры и выяснить, что доля «аз» в корпусе XX века составляет всего 0,0007 процента (то есть слово очень редкое), а в корпусе XVIII века – 0,009 процента (в 10 раз чаще). Все встает на свои места (см. график справа).

4. В настоящее время корпуса существуют **в электронной форме**. Еще несколько лет назад значительная часть времени у многих студентов и исследователей уходила на то, чтобы собрать материал: найти и просмотреть бумажные издания, выписать примеры на карточки, все вручную пересчитать… Часто тот или иной диплом защищался с формулировкой «собран значительный языковой материал». Сейчас эта формулировка сохранилась, например, в полевой лингвистике или в тех областях, в которых еще не созданы корпуса. Электронная форма хранения корпуса обеспечивает быстрый поиск и извлечение материала, превращая исследовательскую работу в быструю проверку множества рабочих гипотез без утомительного этапа механического поиска примеров.

Важно понимать, что возможность поиска в современном корпусе ограничена поиском по буквам и другим знакам и сводится к точному составлению запросов в виде набора символов той или иной степени сложности. Даже когда мы ставим галочки и выбираем параметры из меню, мы по сути указываем, какие уже включенные в корпус символы или их комбинации нас интересуют (о некоторых исключениях я расскажу ниже).

> Например, поиск мужских или женских ролей в мультимедийном корпусе МУРКО (www.ruscorpora.ru/search-murco.html) возможен только потому, что корпус уже содержит заранее введенную информацию о том или ином актере. Поиск реплик актера по его изображению или тембру голоса невозможен и вряд ли необходим.

5. Из требования электронного формата следует возможность развития корпуса как в «ширину» (увеличение объема), так и в «глубину» (дополнительная информация о единицах корпуса). Последнее определяет требование к корпусу, которое сегодня все чаще становится обязательным. Я говорю о наличии специальной **разметки**, или **аннотации**. Именно она позволяет искать не только по текстоформам, но и по другим параметрам. Говоря по-простому, разметка представляет собой лингвистический разбор всех языковых единиц на выбранном языковом уровне, или, если говорить более формально:

> **разметка** (аннотация, англ. *annotation*) – это введенная автоматически или вручную лингвистическая или метатекстовая информация обо всех выбранных единицах корпуса: тексте, предложении, текстоформе, морфеме, звуке и т. д.

Этой важнейшей составляющей современного корпуса будет посвящено несколько глав учебника.

Дополнительная литература

1. Atkins S., Clear J., Ostler N. Corpus design criteria // Literary and linguistic computing. 1992. Vol. 7. № 1. P. 1–16.

2. Biber D. Representativeness in corpus design // Literary and linguistic computing. 1993. Vol. 8. № 4. P 243–257.

3. Integrum: точные методы и гуманитарные науки. М., 2006.

4. McEnery T., Wilson A. Corpus linguistics. Edinburgh: Edinburgh University Press, 1996.

5. O'Keeffe A., McCarthy M. (ed.). The Routledge handbook of corpus linguistics. Routledge, 2010. (Раздел 2: “Building and designing a corpus: what are the key considerations?”).

6. Материалы конференции «Диалог: Компьютерная лингвистика и интеллектуальные технологии». М.; Дубна, 1995-. Доступно по адресу: http://www.dialog-21.ru/.

7. Инструментарий русистики: корпусные подходы. Хельсинки, 2008.

8. Национальный корпус русского языка. 2003–2005: результаты и перспективы. М., 2003.

9. Национальный корпус русского языка. 2006–2008: новые результаты и перспективы. СПб., 2009.

10. Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. 2008. № 16 (2). С. 7–20.

11. Труды международной конференции «Корпусная лингвистика». СПб., 2004-. Доступно по адресу: http://www.corpora.phil.spbu.ru/.

12. Шимкова М. Репрезентативность корпуса как лингвистическая проблема // Сборник: Труды Международной конференции MegaLing-2005: Прикладная лингвистика в поиске новых путей. СПб.: Осипов. 2005. С. 130–139. Доступно по адресу: korpus.juls.savba.sk

Задания

1. Посмотрите видеолекцию Владимира Александровича Плунгяна (ВИДЕО, © ПостНаука; 13:30 мин.) и ответьте на следующие вопросы:

а) Какое определение дает В. А. Плунгян термину «корпус»?

б) Чем лингвист похож на ребенка?

в) Какую часть лингвистической работы сократили языковые корпуса?

г) Что такое Машинный фонд русского языка?

д) Подсчитайте, сколько раз В. А. Плунгян использует формы «корпусно́й / ко́рпусный» и «корпуса́ / ко́рпусы».

2. На сайте конференции «Диалог» (http://www.dialog-21.ru/) найдите программу последней конференции. Сколько раз в названиях докладов встречается слово «корпус» и его производные?

3*. Проведите описанный в главе эксперимент, задав окружающим вопрос «Как дела?». Какого количества опрошенных оказалось достаточно, чтобы ответы стали повторяться?

Глава 2. История корпусной лингвистики

Согласно данным корпусов английского языка, термин *corpus linguistics* был впервые использован в 1977 году. По меркам развития любой науки это не просто недавно, а прямо-таки вчера. Однако за это время корпусная лингвистика успела стать одним из ведущих направлений современной лингвистики. В России новый термин стал известен, по-видимому, в 1996 году благодаря лекциям одного из создателей знаменитого Международного корпуса английского языка (International Corpus of English, ice-corpora.net/ice) Сидни Гринбаума. Во всяком случае первый раз сочетание «корпусная лингвистика» встретилось в русском корпусе в связи с этим именем:

> «В декабре народ ломился на лекции по **корпусной лингвистике** профессора Гринбаума» (журнал «Карьера», № 2, 1999).
>
> Трудно сказать, кто из студентов написал эту заметку в 1999 году, но именно она войдет в историю корпусной лингвистики как первый случай письменной фиксации русского термина.

Конечно, корпусная лингвистика возникла не на пустом месте. Ей предшествовал многовековой период создания корпусов и применения, в сущности, корпусных методов. Однако ключевым отличием от современной корпусной лингвистики были неэлектронная форма хранения материала и, соответственно, неавтоматические способы извлечения данных. Этот период в истории корпусной лингвистики часто называют **доцифровым** (англ. *pre-electronic*).

> Знаменитая древнеиндийская грамматика, созданная великим Па́нини, была «антицифровой» по форме и корпусной по сути. Созданная приблизительно в V или IV веке до нашей эры, она передавалась буквально из уст в уста – в виде стихов. С другой стороны, она была основана на внушительном корпусе ведических текстов, представлявших уже мертвый на тот момент язык санскрит.

Многие другие доцифровые корпуса тоже были связаны со священными книгами разных религий. Среди них корпус библейских текстов стал самым популярным и наиболее исследованным. Основанные на Библии списки слов с указанием стихов получили название **симфоний**, или **конкорда́нций**. Первый конкорданс появился в начале XIII века и назывался «Concordantiae morales sacrae scripturae» («Нравственная конкорданция Священного Писания»).

Следующий этап в развитии доцифровых корпусов наступил в XVIII–XIX веках и был связан с созданием словарей и развитием лексикографии. Многие известные до сих пор словари были созданы авторами на основе многотысячных картотек, по сути – иллюстративных корпусов. Многие из этих корпусов до сих пор хранятся за крепкими дверями с надписью «Картотека» или «Словарный отдел». Однако результатами работы с такими картотекам стали, например, словарь американского английского Ноа Вебстера (Webster's dictionary) или Словарь живого великорусского языка В. И. Даля.

> В. И. Даль собирал материалы для словаря буквально до конца своей жизни: за несколько дней до смерти он добавил новые слова, услышанные от прислуги. Но несколько слов Владимир Иванович придумал сам (например, *живуля*), а ряд слов самого что ни на есть живого великорусского языка (например, русский мат), наоборот, исключил.

В конце XIX – начале XX века появляются корпуса, созданные для лингвистических исследований или – чаще – для решения практических задач. Одна из них – подсчет частотно-

сти языковых единиц. Первым словарем такого рода стал Частотный словарь немецкого языка (Häufigkeitswörterbuch der deutschen Sprache). Словарь был подготовлен для улучшения стенографической системы немецкого языка на основе корпуса в одиннадцать миллионов слов и издан под редакцией Фридриха Вильгельма Кэниннга в Берлине в 1897 году. С тех пор было создано множество частотных словарей и списков для разных языков, в том числе и для русского.

> В 1915 году в Известиях Отделения русского языка и литературы вышла работа, поставившая актуальный в те времена вопрос о «средстве для отличия плагиатов от истинных произведений». Н. А. Морозов составил «лингвистические спектры», или частотные графики, употребления служебных слов разными авторами. Это корпусное по методам исследование было выполнено на материале объемом в пять тысяч слов (большой по тем временам корпус!).

Примерно в то же время лингвисты нового поколения провозгласили отход от описания того, как нужно говорить: важно то, как носители языка говорят на самом деле. Этот принцип, сформулированный на рубеже XIX–XX веков, корпусная лингвистика услышала и сохранила как один из существенных для собственной методологии: корпусная лингвистика описывает прежде всего узус, а не норму.

> Датский ученый Отто Есперсен одним из первых объявил о переходе от прескрептивных (то есть нормативных) грамматик к дескриптивным (то есть описательным). Он отказался от искусственно сконструированных, «чистых» примеров в пользу реального языкового материала. Для своего главного труда «Modern English Grammar on Historical Principles» (1909–1949) он специально подбирал источники примеров. Список этих источников занимает 40 страниц и является прообразом современного репрезентативного и представительного корпуса.

Еще одним развитием этой же идеи ориентации на узус стал Словарь языка А. С. Пушкина, который, с одной стороны, входил в многовековую традицию составления словарей языка писателя, а с другой – ставил своей целью сплошное описание всего множества текстов (по сути, основу словаря составил доцифровой корпус всех текстов А. С. Пушкина).

Современные корпуса: от коллекции текстов к многоуровневой аннотации

С изобретением и широким распространением «электронно-счетных машин», «электронно-вычислительных машин» и «компьютеров» (что одно и то же) доцифровые корпуса никуда не ушли. В некоторых областях лингвистики работа с бумажными картотеками, с текстами на бересте или на глиняных дощечках была и остается существенной частью исследовательской работы. Вообще, для разных языков и разных текстов наблюдается большой разброс в типах и количестве корпусов. Локомотивом корпусной лингвистики является, безусловно, английский язык: никому уже не придет в голову просто собирать английские тексты, когда существуют очень большие и хорошо аннотированные корпуса для всех вариантов этого языка.

> В эру «до аннотирования» электронные корпуса представляли собой просто аккуратно собранную коллекцию текстов. Такими, например, были первые корпуса английского языка (Brown corpus, 1960-е годы) и русского языка (Упсальский корпус русских текстов, 1980-е годы).

В общем, первые электронные корпуса отличались от своих старших собратьев лишь форматом хранения, однако постепенно объем информации, заключенной в корпусе, суще-

ственно увеличился. В зависимости от количества и качества ресурсов для того или иного языка современным корпусом в одном случае назовут представительный, глубоко аннотированный ресурс, а в другом – простую электронную коллекцию текстов. Корпусная лингвистика – живое дело, и к моменту публикации этого учебника наверняка появится еще парочка новых ресурсов. О деталях мы поговорим в следующих главах, а здесь важно сказать, что каждый новый этап в развитии машинной обработки языкового материала открывал новые возможности сначала для создателей корпусов, а затем и для исследователей. По сути, это не покрытая пылью история, а современное состояние корпусной лингвистики: для части языков уже давно созданы морфологически и синтаксически размеченные корпуса, для других создаются первые, еще не аннотированные корпуса.

> Очень трудно создавать корпус древних текстов. Начнем с того, что сканировать древние рукописи очень сложно и даже опасно (для самих рукописей). Лингвистические сложности начинаются уже на первом этапе обработки: слово может писаться разными способами. Например: *фельдмаршалъ – фелд-маршалъ – фелтъ маршалъ* и т. д. Какой вариант считать правильным? И – главное – как искать лексему независимо от всех орфографических вариантов?

В любом случае современная лингвистическая работа часто невозможна без перевода текстов в электронную форму, что автоматически превращает их в, так сказать, «корпус первого порядка». И это прекрасно, что старые корпуса не умирают, а продолжают жить, наполняясь аннотациями, расширяясь и углубляясь. Как поется в одной старой песенке, «работа есть работа, работа есть всегда».

Задания

1. Прочитайте в Википедии статью про Панини на русском и на любом иностранном языке. Какая из статей оказалась более информативной?

2. Существуют ли конкордансы священных книг основных религий мира? С помощью Яндекса или Гугла попробуйте найти конкордансы Корана, Торы (Пятикнижия Моисея), Трипитака.

3*. По вашему мнению, кого из русских лингвистов «доцифровой» эпохи (условно говоря, до 1970-х годов) можно назвать «корпусным» лингвистом в докорпусную эру? Почему?

Глава 3. Самые известные корпуса

Два крупнейших специализированных каталога CLARIN (www.clarin.eu/) и ELRA (http://www.elra.info/) содержат информацию о более чем трех тысячах корпусов. Каждый год появляются новые корпуса, новые форматы и новые типы данных. Значительное число корпусов создается и уже создано для многих языков. Они активно используются как для лингвистических исследований, так и в прикладных целях. Вы можете сами посмотреть, сколько ресурсов создано для английского языка, сколько для русского или для любого другого. Ниже я подробно опишу самые известные и крупные корпуса (список основных корпусов для множества языков можно найти по адресу: www.aclweb.org/aclwiki).

Иноязычные корпуса

1. Британский национальный корпус (British National Corpus, BNC)
http://www.natcorp.ox.ac.uk/; corpus.byu.edu/bnc

100-миллионый корпус разговорных и письменных текстов британского варианта английского языка, охватывающий период конца XX – начала XXI века. Содержит морфологическую разметку.

2. Американский национальный корпус (American National Corpus, ANC)
http://www.anc.org/

22-миллионный корпус разговорных и письменных текстов американского варианта английского языка, охватывающий период конца XX – начала XXI века. Содержит морфологическую, частично синтаксическую разметку и разметку составных имен собственных.

3. Несколько корпусов испанского языка:

Корпус испанского языка (Corpus del español)
http://www.corpusdelespanol.org/

Содержит тексты XIII–XX веков объемом ок. 100 млн слов. Есть частеречная и металингвистическая разметки.

Корпус современного испанского языка (Corpus del español actual, CEA)
sfn.uab.es:8080/SFN/tools/cea/english

Содержит около 540 млн лемматизированных и морфологически аннотированных слов, извлеченных из Википедии и юридических документов (резолюции ООН и документы Европарламента).

4. Итальянский корпус (Corpus di Italiano Scritto)
corpora.dslo.unibo.it

Содержит современные письменные итальянские тексты объемом около 130 млн слов. Содержит частеречную разметку.

5. Корпус немецкого языка Cosmas II (das Projekt COSMAS II)
http://www.ids-mannheim.de/cosmas2/

Вторая версия немецкого национального корпуса, объединяющая свыше 100 разных подкорпусов общим объемом свыше 8,7 млрд слов. Содержит морфологическую и синтаксическую разметки.

6. Лексическая база данных французского языка FRANTEXT (le corpus Frantext)
artfl-project.uchicago.edu

К сожалению, хорошего национального корпуса французского языка не существует. Доступно только неразмеченное собрание текстов XVIII–XX веков общим объемом более 200 млн слов.

7. Греческий национальный корпус (Εθνικός Θησαυρός Ελληνικής Γλώσσας)

hnc.ilsp.gr/en

Корпус объемом более 47 млн слов разных жанров второй половины XX – начала XXI века. Разметка содержит леммы и части речи.

8. Ланкастерский корпус китайского языка (LCMC, Lancaster Corpus of Mandarin Chinese)

www.lancaster.ac.uk/fass/projects/corpus/LCMC

Корпус объемом около 1 млн единиц представляет тексты, написанные на современном мандаринском диалекте китайского языка. Тексты содержат метаразметку и указание на часть речи.

9. Корпус современного украинского языка (Корпус сучасної української мови)

www.mova.info/corpus.aspx

Корпус объемом 13 млн единиц состоит из четырех подкорпусов (художественные, официально-деловые, поэтические, фольклорные тексты). Существует возможность поиска по токенам, леммам и морфологической разметке.

10. Национальный корпус польского языка (Narodowy Korpus Języka Polskiego, NKJP).

nkjp.pl

Корпус объемом в миллиард слов разговорных и письменных текстов современного польского языка. Содержит неполную морфологическую разметку.

11. Чешский национальный корпус (Český národní korpus, ČNK)

ucnk.ff.cuni.cz

Содержит как современные, так и диахронические подкорпуса, устные и письменные тексты. Часть подкорпусов имеет морфологическую и синтаксическую разметки. Общий объем корпуса – более 500 млн единиц.

12. Словацкий национальный корпус (Slovenský národný korpus)

korpus.juls.savba.sk

Объем корпуса – более миллиарда употреблений, часть корпуса морфологический размечена.

13. Болгарский национальный корпус (Български национален корпус)

www.ibl.bas.bg/BGNC_bg.htm

Основной корпус объемом около 1 млн единиц и 14 параллельных подкорпусов объемом 4 млрд единиц. Корпус содержит частичную морфосинтаксическую разметку.

14. Корпуса древнерусского языка

1) Исторический корпус в составе Национального корпуса русского языка делится на несколько подкорпусов:

- церковнославянский: ruscorpora.ru/search-orthlib.html (объем – ок. 500 тыс. токенов);
- среднерусский: ruscorpora.ru/search-mid_rus.html (объем – ок. 3 млн токенов);
- древнерусский: ruscorpora.ru/search-old_rus.html (объем – ок. 500 тыс. токенов);
- берестяные грамоты: ruscorpora.ru/search-birchbark.html (объем – ок. 20 тыс. токенов).

Объем корпусов стремительно увеличивается, так что к тому моменту, когда вы читаете эти строки там наверняка появились новые тексты. Поиск в историческом корпусе с некоторыми ограничениями аналогичен поиску в основном корпусе: в нем есть богатая метаразметка, леммы, морфологические признаки.

Конец ознакомительного фрагмента.