# Корпусная лингвистика

Лекция №1

## Оглавление



### *Технические детали*

Курс «Корпусная лингвистика» читается студентам специальности «Перевод и переводоведение» с двоякой целью. Во-первых, знание основ этой ветви науки о языке входит в требования к кругозору любого лингвиста с высшим образованием. Эту цель можно назвать «информирующей». Во-вторых, методы корпусной лингвистики могут быть чрезвычайно полезны и для перевода текстов. В этом аспекте данный курс даёт студентам возможность овладеть инструментами, которые облегчают труд переводчика.

Кроме того, изучение корпусной лингвистики предоставляет студентам большую свободу в выборе интересной темы дипломной работы.

Данный курс предусматривает 12 лекций. Практические занятия планом не предусмотрены, но могут быть проведены за счёт других предметов (необходим компьютерный класс). Рабочую программу курса (в том числе и список рекомендованной литературы) можно посмотреть на сайте кафедры перевода и переводоведения.

Вот план лекций по курсу:

1. Вводная лекция. Понятие лингвистического корпуса. История корпусной лингвистики. Корпусный (эмпирический) подход в сравнении с хомскианской лингвистикой.
2. Задачи и основные направления корпусной лингвистики. Корпусная лингвистика и компьютерная лингвистика.
3. Предмет исследования. Развитие лингвистических корпусов в мире. Первое и второе поколение корпусов.
4. Типы корпусов: устные и письменные, одноязычные и многоязычные.
5. Типы корпусов: аннотированные и неаннотированные. Лингвистическая аннотация и метаданные.
6. Лингвистические исследования на базе корпуса: изучение лексики.
7. Лингвистические исследования на базе корпуса: изучение других уровней языка.
8. Методы извлечения информации из корпуса. Типы извлекаемой информации. Конкорданс. Программы для работы с корпусами.
9. Создание своего корпуса. Планирование. Сбор и оцифровка данных. Кодировка

текста.

10. Аннотирование корпуса. Хранение, публикация и обновление корпусов.
11. Заключительная лекция. Основные проблемы и направления развития современной корпусной лингвистики.
12. Защита проектов.

Для зачёта необходимо будет пройти компьютерное тестирование и защитить проект. Проект будет состоять из небольшого тренировочного корпусного исследования. В принципе, это исследование вполне может затем вырасти в дипломную работу.

Лекции будут доступны с компьютера в этой аудитории и с сайта кафедры. Кроме лекций желательно почитать литературу из списка.

## *Понятие лингвистического корпуса*

Прежде, чем говорить о корпусной лингвистике, необходимо определить само понятие лингвистического корпуса. По-английски это будет **linguistic corpus** или **text corpus**, множественное число **linguistic corpora** (corpuses употребляется реже). Существует довольно много определений, которые сходятся в одном: корпус есть «некоторый филологический объект». Вот несколько дефиниций:

- корпус — это организованное определённым образом словесное единство, элементами которого являются тексты или специальным образом отобранные отрывки из текстов;
- корпус – это набор лингвистических данных из определённого языка в форме записанных высказываний или письменных текстов, доступный для анализа;
- корпус — это набор естественных текстов на любом языке, устных или письменных, который хранится в электронном виде и позволяет организовать компьютеризированный поиск;
- пожалуй, наиболее полное определение: **корпус есть собрание отрывков текстов в электронной форме, отобранных в соответствии с внешними критериями, чтобы наиболее полно представлять язык или вариацию языка. Функционирует как источник данных для лингвистических исследований.** (John Sinclair)

Вот примеры корпусов:

- **тексты конкретного писателя или писателей;**
- **тексты за конкретное десятилетие или столетие;**
- **современные тексты определённой тематики;**
- **современные тексты, адекватно представляющие язык или общество.**

В одном из определений было сказано, что корпус может быть как устным, так и письменным. Вообще, существует мнение, что лингвистические корпусы не являются ни устными, ни письменными, ни печатными, а представляют собой четвёртую фактуру речи — тексты на машинном носителе — тот самый digital text. Впрочем, с этим взглядом можно спорить.

Понятно, что корпус — это набор текстов, с которыми можно что-то делать. Но что же может делать корпус? Ответ может показаться неожиданным: сам корпус не может делать ничего. Но мы можем использовать специальное программное обеспечение, чтобы искать в корпусе что-либо и производить некоторые вычисления. Что же мы можем искать? В первую очередь, это слова и фразы, которые имеют культурную или лингвистическую значимость.

Кроме того, предметом поиска могут являться какие-либо пометки, которые вы добавили к корпусу, например, пометка «существительное».

А вот примеры того, что может нам выдать поиск по корпусу:

- **все употребления выбранного слова в непосредственном контексте;**
- **вариации и последовательность в использовании лексики;**
- **слова, которые чаще всего стоят рядом с выбранным словом;**
- **наиболее важные различия между двумя наборами текстов;**
- **как тот или иной писатель использует слова и фразы;**
- **интертекстуальность: значение слова как сумма его употреблений;**
- **скрытые (потенциальные) модели использования лексики;**
- **развитие концептов во времени;**
- **сравнение языков.**

В частности, нам, как переводчикам, наиболее актуальны возможности поиска **контекстов** слов, имеющих несколько переводных **эквивалентов**, а также подбор эквивалентов терминологических и фразеологических словосочетаний в **параллельных корпусах**, о которых мы будем говорить в следующих лекциях.

Важнейшее свойство корпуса – **репрезентативность**, то есть, способность отражать все свойства проблемной области. Репрезентативность определяется фонетическими, морфологическими, синтаксическими и стилевыми параметрами корпуса. Именно репрезентативность отличает корпус от простого набора текстов. Не в последнюю очередь репрезентативность зависит от **размера** корпуса.

## *Эмпирический подход в сравнении с хомскианской лингвистикой*

Некоторые русскоязычные источники указывают, что впервые идея о том, что достоверные лингвистические данные могут быть получены лишь из большого массива текстов, была высказана Р.Г. Пиотровским в 60-х годах. На самом деле, осмысленные исследования в области корпусов начались ещё в сороковые годы (Блумфилд, Фрайс и Бонджерс). Но в 50-60-е годы возобладала концепция Ноама Хомского[1] (хомскианская лингвистика, chomskyan linguistics). Она заключалась в том, что нужно изучать лишь ***competence*** (языковое знание, «язык» по Соссюру), а не ***performance*** (языковое употребление, «речь» по Соссюру). Ведь число высказываний естественного языка бесконечно, поэтому исследовать их бессмысленно. С другой стороны, количество языковых правил, которые и составляют *competence*, конечно. Поэтому их можно исследовать. Таким образом, произошёл уход от эмпирики в сторону рационализма и интроспекции (использования интуиции носителей языка). Тем не менее, некоторые учёные продолжали использовать корпусные методики и в период безраздельного господства генеративной лингвистики.

Причина повышения интереса к корпусным исследованиям в последнее время — появление компьютеров, которые сделали возможной обработку огромных массивов текстов. Кроме того, всё больше учёных склоняется к тому, что интроспекция как метод изучения языка не всегда адекватна, и более научно опираться на естественные данные. Известные корпусные лингвисты Тони Мак-Эннери и Эндрю Уилсон пишут, что нужно использовать и эмпирику, и интроспекцию, и искусственные данные, и естественные. Корпусная лингвистика ни в коем случае не отрицает ценности и необходимости речевых данных, не представленных в корпусной форме. Кроме того, из корпуса текстов невозможно извлечь все возможные лингвистические выводы, то есть, корпус текстов не является самодостаточным[2].

1 Основатель генеративной лингвистики

2 Например, корпус в принципе не может дать ответ на вопрос, какие конструкции в данном языке

Так, Чейф считает, что корпусный лингвист должен не только **описывать** явления языка, но и стараться **объяснить** их. Вообще, в центре внимания корпусной лингвистики оказалась **языковая личность**, то есть, её речевая деятельность, массовая коммуникация, проблема её описания.

В этой таблице (её автор — Владимир В. Рыков) показаны основные отличия корпусной лингвистики от традиционной (хомскианской):

| *Корпусная лингвистика* | *Традиционная лингвистика* |
|---|---|
| Основное внимание – изучение **речи** | Основное внимание – изучение **языка** |
| Цель – описание языка в том виде, как он проявил себя в речи, представленной в виде специально подобранного корпуса текстов | Цель – описание и объяснение языка |
| В своих исследованиях опирается на данные корпуса текста | В своих исследованиях идёт от теории к её объяснению и подтверждению в фактах речи |
| Предпочитает **квантитативные (количественные)** методы | Предпочитает **квалитативные (качественные)** методы |
| Видит себя частью традиций, базирующихся на **эмпирических** методах | Видит себя частью традиций, базирующихся на **рационалистических** методах |
| Текст рассматривается как некоторая **физическая сущность** | Текст рассматривается как некоторая **абстракция** |
| Составление грамматики **конкретных** языков | Изучает языковые **универсалии** |
| Основное внимание уделяется **форме** | Основное внимание – не только форме, но и **содержанию** |
| Рассматривает тексты в **глобальной** перспективе | Рассматривает тексты в **локальной** перспективе |
| Фокусирует своё внимание на как можно более **широком** взгляде на текст, неограниченном ни какими догмами | Анализирует некоторую **конкретную**, искусственно ограниченную, проблемную область |
| В своих выводах опирается на наблюдение речевой деятельности, проявленной в виде текстов | Опирается на интуицию в отборе речевого материала, в отборе эмпирических материалов своих исследований |
| Часто пользуется **вероятностными методами** и **статистикой** для первичной обработки речевого материала | Предпочитает **логические рассуждения** |
| Проводится работа с лингвистическими данными (словоупотреблениями) в том виде, в каком они встречались в контексте | Предпочитаются искусственные примеры, из изолированных от текста словоупотреблений |
| Предпочитает **индуктивные** методы обработки эмпирического словесного материала, считает их сутью научного метода | Предпочитает **дедуктивные** методы обработки эмпирического словесного материала |

**невозможны**

| | |
|---|---|
| Верит в научные открытия, основанные на **обработке эмпирических данных** | Верит в открытия, основанные на **процедурах, оценках, сравнениях** и т.д. |

## *История корпусной лингвистики*

Собственно, корпусы люди составляли и изучали ещё до появления корпусной лингвистики, начиная с XVIII века. Примеры: исследования Библии (Cruden и многие другие), составление словарей (Johnson, Oxford English Dictionary, Webster Dictionary), преподавание языков (частотный корпус Thorndike'a, 1921), дескриптивная грамматика (Fries, 1940, Quirk, 1968). Корпус Квирка (Survey of English Usage) включал один миллион словоупотреблений и изначально представлял собой один миллион карточек размером 6 на 4 дюйма, 17 строк текста на каждой. Этот корпус стал последним **не электронным**. Его составление заняло 25 лет, и к 1989 году, когда он был закончен, технология ушла далеко вперёд. Пришлось срочно переводить корпус в цифровую форму. Теперь этот корпус доступен в Юниверсити Колледж в Лондоне.

Основные вехи создания компьютерных корпусов:

1. 1960-е: Брауновский корпус, (США), 1 млн. слов
2. 1970-е: LOB корпус (Великобритания, Норвегия), 1 млн. слов
3. 1980-е: Машинный Фонд русского языка
4. Уппсальский корпус русского языка (Швеция), 1 млн. слов
5. 1990-е: British National Corpus, 100 млн. слов, национальные корпуса (венгерский, итальянский, хорватский, чешский, японский) объёмом 100 млн. слов
6. The Bank of English, Birmingham (Collins Cobuild), 600 млн. слов
7. 2000-е: American National Corpus, 100 млн. слов
8. Corpus of Contemporary American English, 400 млн. слов.
9. Национальный корпус русского языка, 140 млн. слов
10. Gigaword corpora: английский, арабский, китайский, 2 млрд. слов
11. Oxford English corpus, 2 млрд. слов.

Таковы основные продукты деятельности корпусной лингвистики на сегодняшний день. В.В. Рыков даже пишет, что корпусная лингвистика – спорный термин, так как непонятно, имеется ли в виду наука о том, как **создавать корпусы** или же лингвистика, основанная на **данных из корпусов**. На практике, обычно под корпусной лингвистикой понимают и то, и другое. То есть, корпус для корпусной лингвистики, с одной стороны, **исходный речевой материал**, с другой – **результат деятельности**.

Подытоживая:
Корпусная лингвистика сделала возможным:

1. **Уточнить результаты и выводы проведённых ранее исследований речи.**
2. **Произвести новые, более широкие и системные** (по охвату эмпирического речевого материала) **лингвистические исследования.**

## *Рекомендуемая литература*

1. *Список основной литературы по теме: http://scholar.google.com/scholar?q=corpus+linguistics&hl=en&lr=&btnG=Search*
2. *Список последних статей по теме: http://scholar.google.com/scholar?*

*q=corpus+linguistics&hl=en&lr=&scoring=r&as_ylo=2003*

3. **Гальперин И.Р.** Текст как объект лингвистического исследования. - М.: Едиториал УРСС, 2005. - 144 с.
4. **Коваль С.А.** Роль корпуса в создании реалистичных моделей словоизменительной морфологии. URL: http://skowal.narod.ru/research/corpora2006/Koval_Corpora.2006.htm
5. **Марчук Ю.Н.** Основы компьютерной лингвистики. - М.: Изд-во МПУ, 2000
6. **Плунгян В.А.** Почему современная лингвистика должна быть лингвистикой корпусов. 2009
7. **Рыков В.В.** Курс лекций по корпусной лингвистике. URL: http://rykov-cl.narod.ru/c.html
8. **Kennedy, Graeme.** An Introduction to Corpus Linguistics / Graeme Kennedy. - London: Longman, 1998. - 315 p
9. **Tony McEnery, Andrew Wilson.** Corpus Linguistics. - Edinburgh University Press, 2001. URL: http://books.google.com/books?id=nwmgdvN_akAC
10. Developing linguistic corpora: a guide to good practice. Edited by Martin Wynne. URL: http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm
11. **D Biber, S Conrad, R Reppen.** Corpus Linguistics: Investigating Language Structure and Use. - Cambridge University Press, 1998. URL: http://books.google.com/books?id=2h5F7TXa6psC
12. ICAME (International Computer Archive of Modern and Medieval English) Journal. URL: http://icame.uib.no/journal.html

# Корпусная лингвистика. Лекция 2

## *Основные задачи и направления корпусной лингвистики. Взаимодействие корпусной лингвистики и компьютерной (computational) лингвистики.*

# Оглавление



## Основные задачи

Как уже говорилось в предыдущей лекции, деятельность в рамках корпусной лингвистики может быть сведена к **созданию корпусов** и к **лингвистическим исследованиям на их базе (**все задачи по изучению больших массивов текстов). В каком-то смысле, корпусная лингвистика сама создаёт свой материал, точнее, самостоятельно структурирует его. Именно это делает её самостоятельной лингвистической дисциплиной – у неё специфический характер используемого словесного материала (корпусы) и свой собственный инструментарий (программы анализа корпусов). А самостоятельность науки как раз и определяется наличием у неё собственного материала, либо собственных методов его исследования. Корпусная лингвистика обладает как тем, так и другим.

В качестве своей главной цели изучаемая нами наука видит **объективное лингвистическое описание языковой системы, причём к этому описанию корпусная лингвистика подходит от изучения конкретной человеческой коммуникации,** от реальных текстов, которые ранее рассматривались лишь как досадная помеха. В качестве вторичной задачи рассматривается выработка особого способа отражения речевого материала в корпусе текстов. Этот способ, в свою очередь, может использоваться другими лингвистическими дисциплинами.

Ещё одно отличие в подходах между традиционной лингвистикой и корпусной заключается в том, что традиционно языкознание изучало **возможность (possibility)** или невозможность какого-либо лингвистического явления. Например, традиционный учебник английского языка скажет вам, что конструкция *I'm not* в литературном английском возможна, а конструкция *I ain't* – нет. Корпусная лингвистика дополнительно изучает и **вероятность (probability)** лингвистических явлений. То есть, с точки зрения корпусной лингвистики, мы не можем сказать, что употребление *I ain't* в литературном языке совершенно невозможно. Оно всего лишь маловероятно.

## Основные направления

Кратко и неполно расскажем об основных направлениях современной корпусной лингвистики.

Во-первых, это **лексикографические исследования**, создание словарей. Практически все современные словари английского языка (Collins, Webster, MacMillan и т.д.) издаются на основе огромных корпусов, которые позволяют сделать словарь репрезентативным. То есть, словарь может быть верным или не верным относительно данного корпуса.

Во-вторых, изучение корпусов позволяет получать точные данные о **лексическом составе** языков, об относительных частотах употребления тех или иных слов. В частности, при помощи корпусной лингвистики был окончательно доказан так называемый **закон Ципфа**, утверждающий, что если в любом естественном языке все слова упорядочить по убыванию частоты их использования, то частота любого слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру (так называемому *рангу* этого слова). Например второе по частоте слово встречается примерно в два раза реже, чем первое, третье – в три раза реже, чем первое, и так далее.

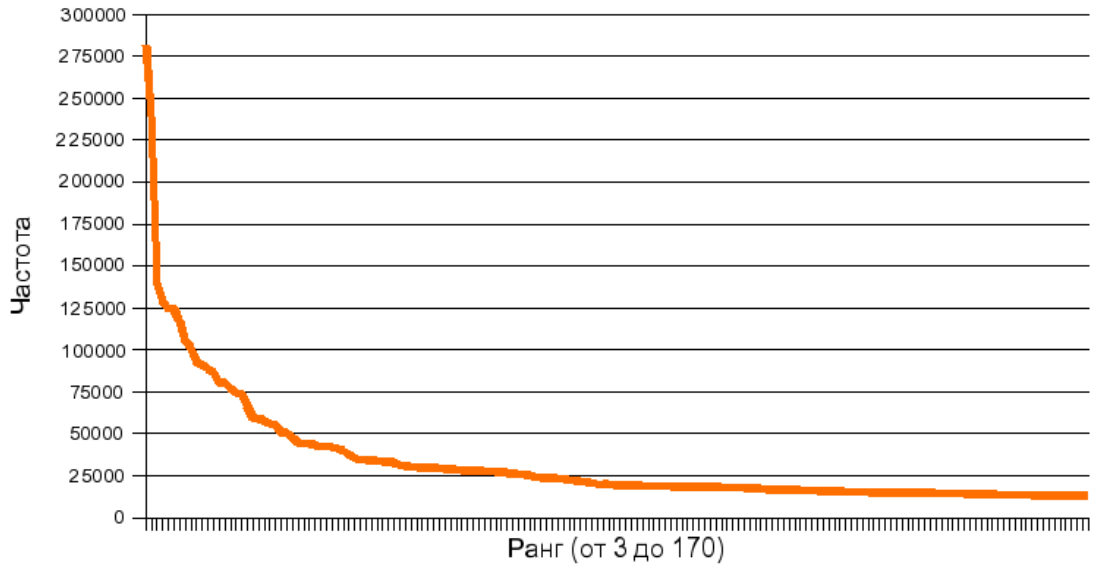


*Рисунок 1: Закон Ципфа*

Выводом из закона Ципфа является утверждение о том, что **язык – это большой набор редких событий**. То есть, редких слов в языке значительно больше, чем частых.

В-третьих, корпусная лингвистика изучает и изменения в лексическом составе языков, различные его вариации (например, появление и исчезновение **неологизмов**).

Четвёртое направление корпусной лингвистики – изучение **грамматики** естественных языков, в частности – сочетаемости тех или иных грамматических явлений

друг с другом. Естественно, что данные, полученные из живой речи, гораздо более актуальны, чем умозрительные грамматики традиционной лингвистики. Кроме того, получается более объективное исследование: грамматика верна лишь относительно того или иного корпуса текстов.

В-пятых, не оставлено без внимания и изучение **текстов**. Например, используя корпусы, мы можем научиться определять **функциональный стиль** через статистические характеристики текста – среднюю длину слова и предложения, характерные сочетания слов и т.д. Такие методы уже существуют и используются в автоматическом реферировании и тематическом поиске. Причём, изучать таким образом можно не только письменный, но и **устный дискурс**.

В-шестых, корпусная лингвистика активно используется в **лингводидактике**, то есть, в обучении иностранным языкам. Чтобы знать, чему, собственно, учить, необходимы точные количественные данные о преподаваемом языке – состав наиболее частотной лексики, вероятности употребления тех или иных грамматических конструкций и т. д. Что немаловажно, корпусная лингвистика даёт возможность обновить набор примеров, которые используются в преподавании языка.

И наконец, особый интерес для нас, как переводоведов, представляют, конечно, **многоязычные корпусы**, особенно **«выровненные» или «сопоставленные» (aligned)**. В «выровненном корпусе» каждой фразе на одном языке соответствует её эквивалент на другом языке или языках. Такие корпусы используются при подготовке переводчиков или при создании двуязычных словарей. Очень важны они для создания систем автоматического машинного перевода (если такая система опирается на корпус переводов, сделанных переводчиками-людьми, её качество будет гораздо выше). Кроме того, такой корпус можно использовать для исследований, связанных со сравнением оригинальных и переводных текстов.

## Корпусная лингвистика и компьютерная лингвистика

Довольно часто звучит вопрос о соотношении корпусной и так называемой **«компьютерной лингвистики»**. Эти ветви науки о языке, действительно, близки друг другу, но всё же не совпадают.

Что такое «компьютерная лингвистика»? Вообще, термин довольно расплывчат, тем более, что существует ещё некая **«математическая лингвистика»**. В англоязычном языкознании проще – там есть один общий термин **computational linguistics**, то есть, «вычислительная лингвистика». Мы для простоты будем говорить «компьютерная лингвистика», поскольку сейчас без компьютеров всё равно никто уже ничего не вычисляет. Так вот, обычно говорят, что компьютерная лингвистика – это такая

**междисциплинарная ветвь лингвистики, занимающаяся либо статистическим либо rule-based[1] моделированием языка с использованием компьютеров**. Моделирование – это приблизительный эквивалент английского термина **sampling**. То есть, компьютерная лингвистика строит модели языка. Кстати, корпусная занимается примерно тем же, поэтому они друг другу помогают.

Вот некоторые точки приложения компьютерной лингвистики:

- автоматический перевод;
- автоматизированное извлечение информации из естественных текстов;
- конструирование удобных интерфейсов между человеком и машиной;
- количественное описание общения на естественных языках;

Немаловажно, что компьютерная лингвистика создаёт **инструменты** (то есть, программы) **для корпусной лингвистики**. В этом смысле они тоже дополняют друг друга. Например, корпусным лингвистам необходимы средства для **автоматической разметки классов слов** в корпусах. Если у вас есть корпус на 100 миллионов словоупотреблений и вам нужно отметить часть речи у каждого слова, то вручную это сделать совершенно нереально. Тут и понадобится специализированное программное обеспечение. Обычно сначала его нужно «обучить», то есть разметить вручную какое-то небольшое количество слов, чтобы система «натренировалась». После этого разметка по классам слов[2] будет происходить в автоматическом режиме.

Очень активно в современном мире используются программы **морфологического и синтаксического анализа**[3]. Именно они лежат в основе автоматической проверки орфографии и грамматики, которая в текстовых процессорах подчёркивает вам красным неправильные слова и фразы. Для создания таких программ равно необходимы как программисты, так и лингвисты.

Для исследования корпуса бывает важно сначала **снять лексическую неоднозначность**, то есть, выделить слова-омонимы. Например, в корпусе русских текстов нужно отделить слово «лук» в значении «овощ» от слова «лук» в значении «оружие». В большом корпусе сделать это вручную затруднительно. Поэтому компьютерная лингвистика создаёт программы **семантического анализа текстов**, которые могут в более или менее автоматическом режиме определять, в каком значении употреблено то или иное слово.

И, наконец, компьютерная лингвистика активно занимается вопросами создания **параллельных корпусов**, о которых говорилось выше. Ведь это очень интересная

1 На основе правил.
2 Англ. *POS (part-of-speech) tagging.*
3 По английски синтаксический анализ – parsing.

лингвистическая задача – как в автоматическом режиме «сопоставить»[1] два текста, один из которых является переводом другого? Как «соотнести» друг с другом отдельные предложения на языке оригинала и на языке перевода? Здесь достаточно проблем и трудностей, но решения уже есть и уже существуют автоматические системы сопоставления текстов. Некоторые из таких программ мы будем изучать в рамках курса «компьютерные технологии в переводе».

Итак, как можно видеть, компьютерная лингвистика выступает для корпусной в качестве «поставщика» инструментов анализа и обработки корпусов. Поскольку большой корпус можно обрабатывать только при помощи компьютера, необходимы программы. А написанием лингвистически ориентированных программ как раз и занимается компьютерная лингвистика. С другой стороны, в современной науке порой сложно отделить корпусного лингвистика от компьютерного, поскольку чаще всего учёные занимаются и тем и другим.

1 Англ. Text alignment.

# Корпусная лингвистика

Лекция 3

# Предмет исследования корпусной лингвистики. Развитие лингвистических корпусов в мире: первое и второе поколение

## 1 Предмет исследования

Корпусная лингвистика рассматривает текстовые массивы как поле изучения и как источник фактов для лингвистического описания и аргументации. Как уже говорилось, она сосредотачивается на «речи» (*performance*), а не на «языке» (*competence*).

Как и вся наука о языке, корпусная лингвистика занимается в основном описанием и объяснением сущности, структуры и использования языка, а так же более частными вопросами: изучение языков, их изменение и т.п. Однако корпусная лингвистика стоит в языкознании несколько особняком

Можно отметить, что часто она ограничивается изучением скорее лексики и лексической грамматики, нежели синтаксиса. В чём-то это результат использования методики конкордансов *(списков слов в контекстах, в последующих лекциях будет более подробно)* – ширины экрана или печатного листа (обычно 130 символов) просто не хватает на то, чтобы анализировать синтаксис или дискурс.

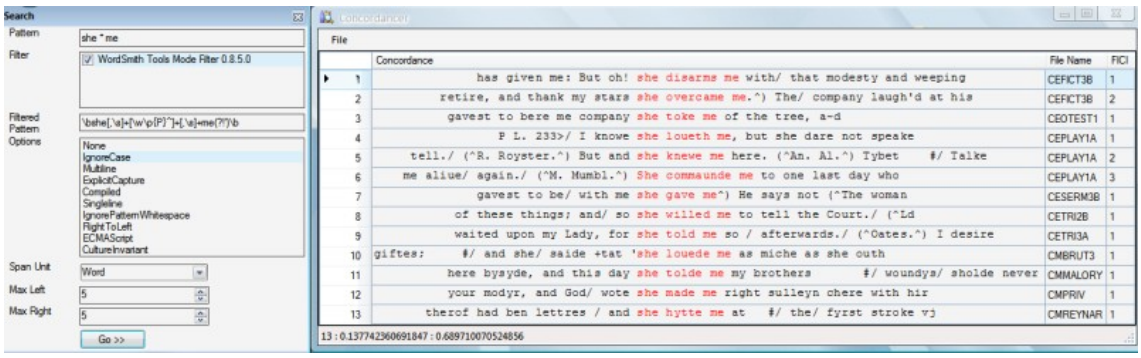


*Рисунок 1: Пример конкорданса в программе Corsis*

Существует четыре группы корпусных лингвистов:

1. Создатели корпусов (corpus compilers).
2. Разработчики программ для анализа корпусов (corpus software developers)
3. Дескриптивные лингвисты, которые используют существующие корпусы для адекватного описания лексики и грамматики языка. В основном используется вероятностный подход.[1]
4. Те, кто занимается использованием корпусов в новых прогрессивных приложениях –

1Например, выяснилось, что в английском языке *quite* чаще сочетается со словами типа *obviously*, а *absolutely* – с отрицательной лексикой (*absolutely not*).

изучение и преподавание языков, машинная обработка естественного языка (например, для распознавания речи и для автоматического перевода).

# 2 История электронных лингвистических корпусов

## 2.1 Первое поколение корпусов

### 2.1.1 The Brown Corpus

Точное название: Brown University Standard Corpus of Present-Day American English. Составлялся с 1961 по 1964 год. Язык корпуса: американский английский, письменные тексты, 1 миллион словоупотреблений (это количество стало фактическим стандартом для всего первого поколения корпусов). В то время в лингвистике доминировала концепция Хомского, так что Nelson Francis и Henry Kucera (создатели Брауновского корпуса) делали свою работу в очень неблагоприятной атмосфере. Корпус состоит из 500 текстов по 2000 слов каждый.

Фактически, он задал стандарт для корпусных исследований, поскольку была очень хорошо продумана структура и выбор категорий текстов. Этот же проект установил традицию свободного доступа к корпусам для исследовательских нужд. На этом корпусе уже в 1969 году был основан словарь American Heritage Dictionary.

### 2.1.2 Lancaster-Oslo/Bergen (LOB) Corpus

1970-78 год, проект университетов Ланкастера и Осло и научного центра в Бергене. Британский английский, 1 миллион словоупотреблений, структура похожа на Брауновский корпус. Учёные уже начали понимать, однако, что одного миллиона словоупотреблений недостаточно для анализа низкочастотных элементов языка (а их большинство). Тем не менее, на Брауновском и LOB корпусах основаны многие сотни качественных и интересных исследований. Сайт проекта - http://khnt.hit.uib.no/icame/manuals/lobman/

### 2.1.3 London-Lund Corpus (LLC)

В 1975 году было завершено создание корпуса **устной** английской речи. Он содержал около 500 тысяч словоупотреблений с орфографической транскрипцией, фонетической и просодической разметкой. Эта грандиозная работа сначала была выполнена в бумажном варианте сотрудниками University College London, а затем переведена в компьютерную форму лингвистами из шведского города Лунд. Сайт проекта - http://www.ucl.ac.uk/english-usage

Помимо упомянутых, составлялись корпусы для лексикографических исследований (**American Heritage Intermediate**), для изучения разговорного английского (**Lancaster/IBM Spoken English Corpus, Corpus of Spoken American English, etc**), диахронические корпусы (**Helsinki Corpus of English Texts: Diachronic Part**, 1,5 миллиона словоупотреблений), корпусы для лингводидактических исследований (**International Corpus of Learner's English**) и другие.

### 2.1.4 Машинный Фонд русского языка

Создание первого советского лингвистического корпуса началось в 1985 году в Институте русского языка Академии Наук СССР. Успели только разработать концепцию и архитектуру корпуса и несколько программ, а также собрать какое-то количество текстов. В районе 1991 года финансирование прекратилось и работы заглохли.

### *2.1.5 Уппсальский корпус русского языка*

Одновременно (в 1980-е годы) в институте славистики университета Уппсалы (Швеция) был создан Уппсальский корпус современных русских текстов. 1 миллион словоупотреблений, около 600 текстов. Сайт проекта - http://www.slaviska.uu.se/korpus.htm

## 2.2 Второе поколение корпусов

К 90-м годам технологии извлечения и хранения текстов позволили создавать корпусы из ста миллионов словоупотреблений и более.

### *2.2.1 The Cobuild Project / The Bank of English*

Началось всё в 1980 году, когда издательство Collins принялось за составление корпуса для создания нового словаря.

В 1990 году было объявлено об объединении усилий Collins и факультета английского языка университета Бирмингема в инициативу под названием *The Bank of English*. *The Bank of English* — это так называемый мониторинговый корпус. Вот слова руководителя проекта Джона Синклера: ***мониторинговый корпус*** *это огромный, вечно изменяющийся поток языка, не имеющий чётко определённого размера. Этот поток проходит через фильтры, которые извлекают из него лингвистические данные.* Около 300 миллионов словоупотреблений в 1997 году, а в 2005 уже 525 миллионов. Каждый месяц в корпус поступает два с половиной миллиона новых словоупотреблений.

25 процентов корпуса составляет устная речь, 75 процентов — письменная.

По адресу http://www.collins.co.uk/Corpus/CorpusSearch.aspx можно использовать тестовую версию корпуса (56 миллионов словоупотреблений).

### *2.2.2 The Longman Corpus Network*

Коммерческая база данных нескольких корпусов, созданных компанией Longman и университетом Ланкастера. 50-100 миллионов словоупотреблений. Сайт в Интернете - http://www.pearsonlongman.com/dictionaries/corpus/index.html

### *2.2.3 British National Corpus (BNC)*

100 миллионов словоупотреблений, представляет английский язык в целом, а не один жанр. Этот корпус имеет конечный размер, в отличие от Cobuild Project. 90 процентов письменных текстов, 10 устных.

В создании принимали участие многие организации, включая Британское правительство. Процесс завершился около 1995 года. Корпус состоит из 4124 текстов, из которых 863 транскрибированы из устных бесед или монологов. Каждый текст сегментирован на орфографические предложения, а внутри них каждому слову автоматически назначен код класса слова (части речи). Во всём корпусе 6,4 миллионов орфографических предложений. Сегментирование и классификация слов были выполнены программой стохастической разметки CLAWS, разработанной в университете Ланкастера. Классификационная схема предусматривает 65 частей речи, которые описаны в прилагающейся документации. Все тексты размечены с использованием наиболее стандартных способов — языка SGML и системы TEI.

При создании корпуса были использованы новые подходы к отбору текстов, многоуровневая система контроля.

Корпус доступен по адресу http://www.natcorp.ox.ac.uk/

#### *2.2.4 The International Corpus of English (ICE)*

Совместный проект нескольких десятков университетов. 20 параллельных подкорпусов, по миллиону словоупотреблений каждый, вместе 20 миллионов словоупотреблений. Можно изучать специфику стран, где английский – второй или официальный язык (Австралия, Канада, Новая Зеландия и т.п.). Разработано сложное программное обеспечение специально для анализа этого корпуса. Веб-сайт: http://www.ucl.ac.uk/english-usage/ice/

#### *2.2.5 American National Corpus*

Первый выпуск состоялся в 2003 году. Планируется 100 миллионов словоупотреблений, но пока только 11. Доступ исключительно на платной основе. Корпус в XML-формате. Веб-сайт проекта: http://www.americannationalcorpus.org/

#### *2.2.6 Gigaword corpora*

Мониторинговые корпусы английского, арабского, китайского и других языков. Спонсируются Европейским Союзом, создаёт их компания Linguistic Data Consortium. Уже 1 миллиард словоупотреблений. В основном тексты взяты из публицистики и новостей. Корпусы довольно дорогие. Посмотреть на список можно на сайте http://www.ldc.upenn.edu

### 2.3 Современные российские лингвистические корпусы

#### *2.3.1 Национальный корпус русского языка*

Общедоступный для поиска корпус русских текстов (сокращённо НКРЯ). Открыт 29 апреля 2004 в Интернете по адресу http://ruscorpora.ru. Работы по созданию Корпуса были начаты в 2001 году группой лингвистов из Москвы, Петербурга, Воронежа и других городов. Основные участники – Институт русского языка РАН, Институт языкознания РАН и компания «Яндекс».

Письменные, устные, поэтические диалектные тексты. 140 миллионов словоупотреблений в 2007 году. Корпус морфологически и семантически размечен и полностью свободен для использования при помощи веб-сайта.

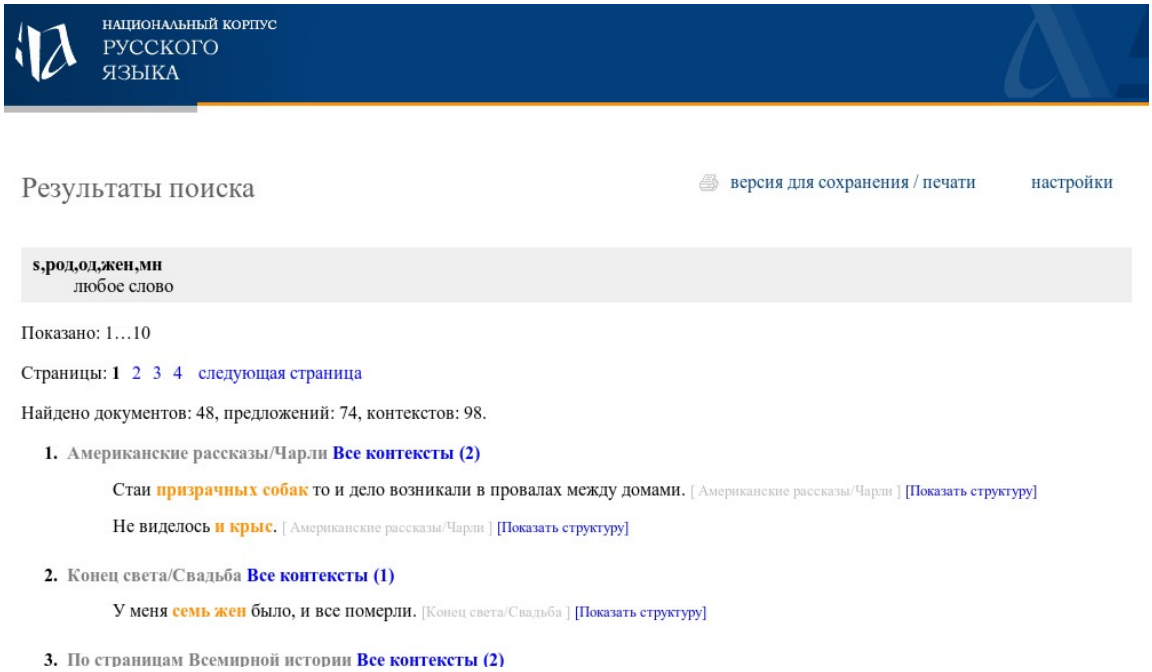


*Рисунок 2: Пример поиска одушевлённых существительных женского рода множественного числа в родительном падеже (НКРЯ)*

# Корпусная лингвистика

Лекция 4

## Корпусы: устные и письменные, одноязычные и многоязычные

Если бы нам пришла в голову идея исследовать корпус текстов по корпусной лингвистике (например, книг и научных статей) методами самой корпусной лингвистики, то оказалось бы, что чаще всего к слову «*корпус*» примыкает глагол «*составлять*»[1]. Какими же бывают корпусы по методу составления?

## *Устные – письменные*

Большая часть корпусов 1 поколения были исключительно письменными. Письменные тексты гораздо легче собирать. Существуют три метода ввода письменных текстов в компьютер:

- заново набирать тексты (это лучше, чем пробивать перфокарты, как было с Брауновским корпусом);
- использовать тексты, которые уже существуют в электронной форме;
- сканировать напечатанные тексты (но при этом нужно исправлять много ошибок).

Большие современные корпусы обычно комбинированные, с преобладанием письменных текстов. Даже в **BNC** лишь 10% текстов устные. Выделяется **ICE**, в котором 60% текстов устные.

Между тем, язык в основном существует именно в устной форме, письменная его форма вторична. Поэтому так важны устные корпусы, либо смешанные.

Среди специфически устных корпусов нужно назвать **London Lund Corpus** (LLC, 1975 г.) и **Lancaster/IBM Spoken English Corpus** (1992), сокращённо SEC. Этот последний состоит из 52600 словоупотреблений. Он поставляется на CD-ROMe вместе с аудиозаписями, полностью размечен на предмет ударений, интонации, пауз и т.п. Однако, он не содержит информации о социальном статусе и образовании респондентов, что ограничивает его использование в социолингвистике.

**Corpus of Spoken American English** (1991), миллион словоупотреблений, 80 часов звучания.

**Map Task Corpus** (1991, университет Глазго, Шотландия), 147 тысяч словоупотреблений, 16 часов звучания.

Устные корпусы включают меньше словоупотреблений, чем письменные, не только из-за трудоёмкости сбора данных, но и потому, что для просодических исследований обычно достаточно меньшего количества слов. Так, для изучения интонации достаточно корпуса в сто тысяч словоупотреблений.

Устные корпусы могут включать как монологическую, так и диалогическую речь. Для сбора материала используются записи с радио и телевидения или опрос по выборочным

1*Хотя вообще-то лингвистические корпусы предназначены для того, чтобы быть основой для анализа и описания языков.*

методикам социологии и социолингвистики. Отметим, что скрытая запись сейчас считается неэтичной (в отличие от 70-х годов).

Обычно собирают довольно подробную информацию о респондентах:

- место записи
- что респондент делает
- время
- дата
- количество участников
- степень спонтанности беседы
- тема
- пол участников
- возраст участников
- этническая принадлежность участников
- основной язык участников
- профессия
- образование
- социальный статус
- отношение к записывающему
- диалект

Самая трудоёмкая стадия — transcription. Орфографическая транскрипция одного часа записи с минимальной интонационной разметкой может занять около 10 часов. Если же размечать текст по всем правилам TEI (Text Encoding Initiative), то на это может уйти 25 часов и более. А без разметки корпус устных текстов не имеет смысла — как минимум, должна быть указана продолжительность пауз, размечена одновременная речь, ударение, интонация. Иногда включают контекстные комментарии типа «*ест печенье*». Именно благодаря подробной разметке корпус LLC стал стандартном для корпусов устной речи.

## *Статические – динамические*

Первые корпусы были статичными снимками языка. Наиболее значимый современный корпус (BNC) тоже статичен. Но начали появляться и динамические мониторинговые корпуса, которые пополняются постоянно. Пример — Cobuild Project. Такие корпусы ещё называются «открытые». Их проблема в том, что они часто не совсем адекватно представляют язык, поскольку не подчиняются чётким критериям отбора, тексты не сбалансированы.

## *Одноязычные — многоязычные*

Корпусных лингвистов (особенно связанных с переводом) всегда интересовала задача составления корпусов на нескольких языках. Уже в первом поколении начали появляться двуязычные корпусы для таких языков, как английский, финский, французский, немецкий, греческий, норвежский, испанский, шведский, валлийский. Такие корпусы ещё называются

bitexts.

Естественно, нет никаких технических препятствий к тому, чтобы делать корпусы не дву- а трёх-, четырёх- и более язычными. Вообще говоря, само появление многоязычных корпусов спровоцировало всплеск научных исследований, поскольку для их анализа требуются другие инструменты и даже другие концепции, нежели чем для анализа корпусов одноязычных. Вполне естественно, что можно представить себе два типа двуязычных корпусов:

- корпус, в котором тексты являются переводами друг друга
- корпус, в котором просто присутствуют тексты на разных языках (возможно, одной и той же тематики).

Корпусы второго типа иногда называют «переводными» (translation corpora) и используются для изучения различий в выражении схожих мыслей на разных языках. Корпусы первого типа называют «параллельными» (parallel corpora) и используются для исследования различных аспектов собственно перевода. Например, существует параллельный корпус текстов заседаний канадского парламента (английский/французский).

Параллельные корпусы также могут быть двух типов — выровненные (aligned) и не выровненные (not aligned). «Выровненность» означает, что в корпусе существует чёткая связь между единицами перевода, которые соответствуют друг другу. То есть, мы можем быстро найти, как то или иное слово или предложение переводилось на другой язык. Обычно такими единицами перевода служат всё-таки предложения, поскольку часто сложно выровнять слова (ведь обычно переводят не дословно). Такой корпус наиболее полезен для переводчика, поскольку представляет собой ту самую «память переводов» (translation memory) — бесценный ресурс, позволяющий использовать предыдущие переводы. Невыровненные корпусы ещё называют «сравнительными».

*«Выровнять текст с его переводом на другой язык означает показать какие части текста переведены какими частями второго текста»* (Kay & Röscheisen 1993: 121)

Выравнивание (alignment) можно делать автоматически, а можно вручную. Первый способ быстрее, но чреват ошибками. Например, если при переводе произошло членение или объединение предложений, то не всегда можно легко определить, какое из предложений перевода соответствует какому предложению оригинала.

Одним из примеров выровненного многоязычного корпуса может послужить база данных *Acquis Communautaire* Европейского Союза (DGT-TM). Это память переводов европейского законодательства на 22 языка[1], которую выложили в открытый доступ в ноябре 2007 года. Всего в ней около миллиарда слов, она выровнена по предложениям (sentence-aligned). Вот пример предложения из этой базы данных:

***EN**: Articles 5 to 7 of this Directive do not apply to containers for gases which are compressed, liquefied or dissolved under pressure.*

***BG**: Членове 5 - 7 на настоящата директива не се отнасят за контейнери с газове, които са сгъстени, втечнени или разтворени под налягане.*

***CS**: Články 5 až 7 této směrnice se nevztahují na kontejnery pro plyny, které jsou stlačené, zkapalněné nebo rozpuštěné pod tlakem.*

1 Поскольку все новые члены ЕС обязаны принимать все его законодательные акты, ЕС вынужден переводить весь Acquis на все языки.

***DA**: Artiklerne 5-7 i dette direktiv finder ikke anvendelse på beholdere , der indeholder komprimerede , flydende eller under tryk opløste gasser.*

***DE**: Die Artikel 5 bis 7 gelten nicht für Behälter , in denen sich verdichtete , verflüssigte und unter Druck gelöste Gase befinden.*

***EL**: Τα άρθρα 5 έως 7 της παρούσης οδηγίας δεν έχουν εφαρμογή επί δοχείων που περιέχουν αέρια συμπιεσμένα , υγροποιημένα ή διαλελυμένα υπό πίεση.*

***ES**: Los artículos 5 a 7 de la presente Directiva no se aplicarán a los recipientes que contengan gases comprimidos, licuados y disueltos a presión.*

***ET**: Käesoleva direktiivi artikleid 5-7 ei kohaldata mahutitele, mis on täidetud suru-, vedel- või rõhu all lahustatud gaasiga.*

***FI**: Tämän direktiivin 5-7 artiklaa ei sovelleta säiliöihin, jotka sisältävät puristettua, nesteytettyä tai paineen alla liuotettua kaasua*

***FR**: Les articles 5 à 7 de la présente directive ne sont pas applicables aux récipients qui contiennent des gaz comprimés, liquéfiés et dissous sous pression.*

***HU**: Ezen irányelv 5-7. cikke nem alkalmazható sűrített, cseppfolyósított vagy nyomás alatt oldott gázok tárolóira.*

***IT**: Gli articoli 5, 6 e 7 della presente direttiva non sono applicabili ai recipienti contenenti gas compressi liquidi e disciolti sotto pressione.*

***LT**: Šios direktyvos 5-7 straipsniai netaikomi suslėgtų, suskystintų ar aukštame slėgyje ištirpintų dujų konteineriams.*

***LV**: Šīs direktīvas 5. līdz 7. pantu nepiemēro tvertnēm saspiestai, sašķidrinātai vai zem spiediena izšķīdinātai gāzei.*

***MT**: L-Artikoli 5 sa 7 ta'din id-Direttiva ma japplikawx għall-kontenituri għal gassijiet li huma kompressati, likwifikati jew maħlula taħt pressa.*

***NL**: De artikelen 5 tot en met 7 zijn niet van toepassing op houders die samengeperste, vloeibaar gemaakte en onder druk opgeloste gassen bevatten.*

***PL**: Artykuł 5-7 niniejszej dyrektywy nie stosuje się do zbiorników lub pojemników zawierających gazy sprężone, skroplone lub rozpuszczone pod ciśnieniem.*

***PT**: Os artigos 5º., 6º. e 7º. da presente directiva não são aplicáveis aos recipientes que contêm gases comprimidos, liquefeitos e dissolvidos sob pressão.*

***RO**: Articolele 5 - 7 din prezenta directivă nu se aplică recipientelor care conțin gaze comprimate, lichefiate sau dizolvate sub presiune.*

***SK**: Články 5 až 7 tejto smernice neplatia pre nádrže na plyn, ktorý je stlačený, skvapalnený alebo rozpustený pod tlakom.*

***SL**: Člena 5 in 7 te direktive se ne uporabljata za posode za plin, ki je stisnjen, utekočinjen ali raztopljen pod pritiskom.*

***SV**: Artikel 5 7 i detta direktiv gäller inte behållare för gas som är komprimerad eller kondenserad eller löst under tryck.*

Ценность параллельного корпуса, как и других корпусов, возрастает с его размером и количеством языков. В этой связи трудно переоценить важность *Acquis Communautaire*, который является самым большим параллельным корпусом в мире. Ещё два его преимущества — бесплатность и наличие редких пар языков, типа «мальтийский-эстонский», «словенский-

финский». Сам корпус представлен в стандартном открытом формате памяти переводов **TMX**, про который я ещё расскажу на лекции по компьютерным технологиям в переводе.

Этот и подобные корпусы можно использовать для многих целей. Например:

- выявление типичных переводческих приёмов и трансформаций
- обучение статистических систем автоматического перевода
- создание одноязычных и многоязычных словарей
- обучение и тестирование программ извлечения информации
- автоматическая проверка правильности перевода
- и конечно, облегчение труда переводчика через подбор возможных эквивалентов

Двуязычные корпусы — ещё одно благодатное поле для студентов-лингвистов, которые могут использовать их для выполнения своих квалификационных работ. Корпус в данном случае может пониматься не как самостоятельная цель, а как инструмент для получения некоторых языковых данных. Соответственно, здесь возможны либо исследования процесса и результата перевода (берём оригинал и перевод), либо контрастивные исследования (берём схожие тексты на языке 1 и языке 2).

# Корпусная лингвистика

Лекция 5

# Корпусы: аннотированные и неаннотированные. Лингвистическая аннотация (разметка) и метаданные.

## Оглавление



В соответствии с классификацией по признаку наличия какой-либо индексации, корпусы бывают *raw* и *annotated*, или, говоря по-русски, *простые* и *аннотированные*. Вообще, для широко известных современных корпусов эта классификация уже утратила актуальность, поскольку все они являются аннотированными. Что, впрочем, не исключает возможности сделать какой-то небольшой корпус безо всякой аннотации.

Итак, чем же корпусных лингвистов так привлекают аннотированные корпусы?

### *Что такое разметка?*

*Знаки пунктуации — это разметка.*

*Маргиналии на полях средневековых манускриптов — это разметка.*

Под **лингвистической аннотацией** или **разметкой** корпуса (по-английски **linguistic markup**) подразумевается наличие в корпусе неких данных, не являющихся частью текста, но несущих какую-то информацию о нём (так называемые **метаданные**). Простейший пример таких данных — отметки частей речи. Выглядеть это может так:

*I will use Google before asking dumb questions.*

Размечаем:

*I* (pronoun) *will* (verb) *use* (verb) *Google* (noun) *before* (preposition) *asking* (verb) *dumb* (adjective) *questions* (noun).

В основном это нужно для облегчения автоматического анализа корпуса. Один раз отметив в тексте все части речи, затем можно производить любые исследования, связанные с ними без необходимости заново выявлять, например, все прилагательные в корпусе. Понятно, что если такой разметки нет, то, к примеру, поиск по слову «*will*» выдавал бы все случаи его появления в корпусе, вне зависимости от того, существительное это или вспомогательный глагол. Но ведь обычно исследователя интересует лишь какой-то один из этих случаев! И это далеко не единственный тип разметки, который бывает нужен корпусному лингвисту.

## *История систем разметки*

В 80-х годах был принят стандарт разметки электронных текстов под названием **SGML**[1] (Standard Generalized Markup Language).Он был разработан внутри типографской индустрии, но быстро распространился на другие отрасли. Смысл SGML был в том, чтобы документы, набранные в разных текстовых процессорах, можно было редактировать, анализировать и изменять в любом из них.

## Тэги как лингвистический инструмент

SGML ввёл концепцию **тэгов**. Тэги (англ. tags) — это служебные пометки в тексте, содержащие информацию о самом тексте. Для каждого случая можно определять собственные тэги и таким образом создавать диалекты языка SGML.

Традиционно тэги заключаются в угловые скобки и бывают парными: открывающими и закрывающими. Например, *<a>* - это открывающий тэг, а *</a>* - закрывающий. Закрывающий тэг сигнализирует, что то, о чём сообщал открывающий тэг, закончилось. Приведём пример тэга ** (выделение важного в тексте, emphasis):

*Это относится в первую очередь к вам!*

В данном случае слова «в первую очередь» помечены как важные. Тэги могут быть вложенными друг в друга:

<ds>*Это относится в первую очередь к вам!</ds> - сказал он.*

Текст «это относится в первую очередь к вам» заключён в тэги *<ds>*, означающие прямую речь (direct speech), а внутри него слова «в первую очередь» дополнительно заключены в тэги **. Количество уровней вложенности не ограничено.

Тэги могут быть и не парными, то есть, не иметь «открывающей» и «закрывающей» части. Например, при разметке устных корпусов употребляется тэг *<pause>*, означающий, что в этом месте произошла задержка речи. Он одиночный.

Сами тэги в обычных обстоятельствах пользователю не показываются. Программа, отображающая размеченный текст, интерпретирует тэги в соответствии с заложенными в неё правилами и показывает пользователю текст, оформленный согласно им.

## Текст с разных сторон: alternative views

Одно из наиболее значительных преимуществ разметок семейства SGML — возможность нескольких **представлений текста** (alternative views). Это означает, что один и тот же размеченный текст легко представить в нескольких видах, в зависимости от нашей текущей задачи. Например, мы хотим выделить из корпуса только текст, **не** являющийся прямой речью. Тогда та программа, в которой мы просматриваем текст, просто скроет все символы, заключённые в тэги *<ds>* и наш пример будет выглядеть уже так:

*- сказал он.*

Или мы можем указать, чтобы текст, помеченный, как важный, был зелёного цвета, а прямая речь выделялась полужирным шрифтом:

***Это относится в первую очередь к вам!*** *- сказал он.*

1 Стандарт ISO 8879:1986

Можно представить и гораздо более сложные alternative views. Например, тэгами можно разметить слова, которые произносят разные персонажи пьесы, а затем представлять их диалог либо в виде последовательных строчек (слова одного персонажа под словами другого), либо дать каждому персонажу отдельную колонку — так, чтобы, высказывания и ответы на них находились на одной строке[1].

## Потомки SGML

Язык разметки SGML — это как бы «конструктор» языков. Сам по себе, в своём первозданном виде, он очень сложен и используется довольно редко. Но на его базе были созданы такие широко известные языки разметки, как **HTML** и **XML**.

Язык **HTML** (Hyper-Text Markup Language), на котором написано подавляющее большинство страничек интернет-сайтов, создали из SGML путём выделения чётко определённого ограниченного набора тэгов, в основном относящихся к оформлению, а не к содержанию документа. В результате мы получили WWW (Всемирную Паутину).

Второе широко известное подмножество SGML — расширяемый язык разметки **XML** (eXtensible Markup Language), который применяется для хранения любых структурированных данных — в том числе и текстов в корпусах. Фактически, это свод синтаксических правил для описания структуры данных. Например, формат офисных документов Open Document построен именно на XML.

Специально для разметки текстовых данных (корпусов) несколько университетов[2] разработали систему, описывающую, какие именно параметры текстов нужно размечать. Эта система использует XML и называется Text Encoding Initiative Guidelines (TEI Guidelines). Это список различных особенностей текстов, которые вообще можно кодировать, размечать и индексировать. Например, система перечисляет различные типы исправлений в тексте, помарок, цитат, иностранных слов и т.д. и т.п. В настоящее время практически все проекты по созданию корпусов (в том числе British National Corpus) стараются в той или иной мере следовать рекомендациям TEI. Подробнее почитать о них можно на http://www.tei-c.org/Guidelines/index.xml.

Естественно, каждый, кто создаёт корпус, может сам выбирать, что именно ему размечать и насколько подробно. Но считается, что в письменном корпусе нужно размечать части речи, границы высказываний, цитаты, списки, заголовки, аббревиатуры, имена собственные, инициалы и акронимы, главы книг. В устных текстах важно разметить обмен репликами, прерывания, перекрывающуюся речь, диалектные формы, паузы и неразличимую речь.

В приложении к этой лекции приведён пример текста, размеченного в соответствии с рекомендациями TEI.

## *Автоматическая разметка текстов*

Понятно, что размечать большие корпусы вручную — занятие очень долгое и дорогое. Поэтому уже в 70-х годах появляются первые проекты по поручению этой задачи компьютеру. Тогда программа TAGGIT смогла корректно назначить тэги частей речи 77% слов в Брауновском корпусе. Остальные пришлось размечать вручную в течение 10 лет. Но прогресс

1 Пример Graeme Kennedy
2 Оксфордский, Брауновский, университет Вирджинии и некоторые другие

не стоял на месте. В 80-е годы система CLAWS (Constituent Likelihood Automatic Word-tagging System) правильно разметила уже около 95% Брауновского корпуса. В ней использовался вероятностный подход. В настоящее время для основных европейских языков уже реализованы как автоматическая разметка **частей речи** (морфологический анализ, word-class tagging), так и автоматическая разметка **членов предложения** (синтаксический анализ, parsing). Эти достижения используются, в том числе, и в системах автоматического перевода и интернет-поиска.

В этой связи нужно отметить немалый вклад рабочей группы учёных под названием «Автоматическая обработка текста» (сайт http://www.aot.ru). В основном они занимаются русским языком. Выросла эта группа из факультета лингвистики РГГУ и занимается приложением теоретической лингвистики к современным компьютерным технологиям. Они разработали модули графематического (определение границ слов), морфологического (определение частей речи), синтаксического (определение членов предложения) и семантического (выявление семантических связей между словами) анализа текстов на русском, немецком и английском языках.

### *Что ещё почитать про разметку?*

1) **James H. Coombs, Allen H. Renear, Steven J. DeRose**. Markup Systems and the Future of Scholarly Text Processing, 1987
2) **Darrel R. Raymond, Frank Wm. Tompa, Derick Wood**. Markup Reconsidered, 1992
3) **Stuart A. Yeates**. Text Augmentation: Inserting mark-up into natural language text with PPM Models, 2006

   Все эти статьи легко найти в Интернете.

## *Приложение*

Пример текста, размеченного в соответствии с рекомендациями TEI, из работы Cetin Sert «Keywords of Protagonists in Shakespeare's Tragedies». Это пьеса Шекспира «Венецианский купец». Вначале идёт заголовок корпуса со сведениями о нём, а затем сам текст:

```
2 <!-
3
4 TENKA SOLUTIONS
ETEXT ARCHIVE
6 SHAKESPEARE MARLOWE CORPUS PROJECT
7
8 FOR EXAMPLE PURPOSES ONLY
9 SUBJECT TO FURTHER CHANGE
11
12 GLOSSARY OF TEMPORARY SYMBOLS & NOTATIONS
13
14 ... missing section:
will be added later after thoroughly
16 studying TEI guidelines
17
18 xxx TEI public release version number:
19 probably P5 will be used
21
22 2006-06-26
23 CETIN SERT
24
```

```
26 -->
27 <TEI.XXX>
28 <teiHeader>
29 <fileDesc>
<titleStmt>
31 <title>The Merchant of Venice (1623 First Folio Edition)</title>
32 <author>Shakespeare, William, 1564-1616</author>
33 <respStmt>
34 <resp>Retagged by</resp>
<name>Cetin Sert</name>
36 </respStmt>
37 </titleStmt>
38 <sourceDesc>
39 <bibl>
The first folio of Shakespeare, prepared by Charlton Hinman
41 (The Norton Facsimile, 1968)
42 </bibl>
43 <!-- ... -->
44 </sourceDesc>
</fileDesc>
46 <encodingDesc>
47 <projectDesc>
48 <p>Prepared for use in the production of a series of old-spelling
49 concordances for the corpus linguistics term-paper of Cetin Sert</p>
</projectDesc>
51 <editorialDecl>
52 <correction>
53 <p>Line numbers are omitted</p>
54 <p>Line number pattern: "^[0-9]+: ?"</p>
</correction>
56 </editorialDecl>
57 </encodingDesc>
58 </teiHeader>
59 <text>
60 <body>
61 <div1 type="Act" n="1" part="N">
62 <l><head rend="italic">Actus primus</head></l>
63 <div2 type="Scene" n="1" part="N">
64 <l><stage rend="italic">
<move type="enter" who="Antonio; Salario; Solanio">[ Enter Anthonio, Salarino, and
Salanio.]</move></stage></l>
66 <l />
67 <l />
68 <sp who="Antonio">
69 <l><speaker>Anthonio.</speaker></l>
<l />
71 <l>In sooth I know not why I am so sad,</l>
72 <l>It wearies me: you say it wearies you;</l>
73 <l>But how I caught it, found it, or came by it,</l>
74 <l>What stuffe 'tis made of, whereof it is borne,</l>
<l>I am to learne: and such a Want-wit sadnesse makes of</l>
76 <l>mee,</l>
77 <l>That I haue much ado to know my selfe.</l>
78 </sp>
79 <l />
```

# Корпусная лингвистика

Лекция 6

# Лингвистические исследования на базе корпусов

## Оглавление



Лингвистические корпусы составляют, чтобы предоставить основу для более точного и адекватного описания структурных и функциональных параметров языка. Сегодня мы поговорим о результатах некоторых корпусных исследований и опишем как вообще использование корпусов может помочь лингвистике.

## *Описания лексики*

Конечно, чаще всего корпусные описания лексики применяются в лексикографии. Практически все современные словари английского языка построены на базе корпусов. Корпусы помогают достоверно определить набор словоформ (types) в языке, показывают появление новых словоформ, используются для уточнения разных значений одного слова и их относительных частот. Самый известный пример — словари издательства Collins, построенные на базе корпуса Cobuild Project.

Корпусы могут давать очень интересные лексикографические сведения о языке. Так, даже в относительно небольшом London-Lund Corpus слово good встречается 800 раз. Оно выступает в 20 значениях как прилагательное, а кроме того, может являться междометием в различных функциях. У всех этих значений разная частота употребления. Кроме того, корпусы показывают появление неологизмов. Так, в 1994 году в английских газетах появились следующие интересные слова: *complainy, dial-a-video, bespokesman, cleavage-wielding, event-driven, fruitcakeland, infotainers, over-housed, unbusy, anarchitecture, bimboisation, bonkable, crashworthiness.*

Статистические исследования лексики на материале корпусов начались ещё в докомпьютерную эпоху. Основным их результатом стал известный закон Ципфа (30-е годы). Напомним, что суть его в том, что в любом массиве текстов небольшое число словоформ[1] (types) образует большую часть реальных словоупотреблений[2] (tokens). Соответственно, например, 90-95 процентов словоупотреблений в английских текстах составлено из 2-5 тысяч наиболее употребительных словоформ. Более того, около половины текста — это словоупотребления 50-100 самых актуальных словоформ (хотя конкретный их набор может

1 Словоформа есть единица речи, повторяющаяся одинаковая последовательность звуков или букв [Щерба , Гируцкий]

2 Словоупотребление есть единица речевой деятельности, любая цепочка букв или звуков между двумя пробелами [Щерба, Гируцкий]

быть разным для разных стилей и подъязыков). Это открытие имело большое значение для преподавания английского языка, поскольку позволило сосредоточиться на предъявлении учащимся самой частотной лексики.

Существуют исследования, которые описывают, какая лексика специфична для определённых типов текстов и вряд ли появится в других. Так, для научных текстов одним из таких слов является глагол *to measure*, а для художественных — *to kiss*.

Появление электронных корпусов дало возможность уточнить частотные параметры лексики. Вот, например, список 50 наиболее частых словоформ в Birmingham Corpus.

| | | | |
|---|---|---|---|
| *1) the* | *14) you* | *27) are* | *40) so* |
| *2) of* | *15) on* | *28) or* | *41) what* |
| *3) and* | *16) with* | *29) by* | *42) their* |
| *4) to* | *17) as* | *30) we* | *43) if* |
| *5) a* | *18) be* | *31) she* | *44) would* |
| *6) in* | *19) had* | *32) from* | *45) about* |
| *7) that* | *20) but* | *33) one* | *46) no* |
| *8) I* | *21) they* | *34) all* | *47) said* |
| *9) it* | *22) at* | *35) there* | *48) up* |
| *10) was* | *23) his* | *36) her* | *49) when* |
| *11) is* | *24) have* | *37) were* | *50) been* |
| *12) he* | *25) not* | *38) which* | |
| *13) for* | *26) this* | *39) an* | |

С другой стороны, многие слова встречаются в данном корпусе только один раз. Такие слова называют hapax legomena (от греческого - «*нечто, сказанное один раз*»). В коротком тексте из 200 слов около 150 слов обычно являются hapax legomena. И даже в больших корпусах с 5 миллионами словоупотреблений почти 40% слов встречаются лишь однажды.

Другое применение корпусов в лексических исследованиях — описание сочетаемости слов (collocation). Есть два подхода — подход **Скиннера**, который рассматривал предложения как словесные цепи, в которых вероятность следующего слова определяется предыдущими, и подход генеративной грамматики (**Хомский**), которая утверждает, что язык всегда порождает новые неожиданные словосочетания по определённым синтаксическим правилам. «Идиоматический принцип» и «принцип открытого выбора», как назвал их Дж. Синклер. Корпус может помочь найти лексикализованные словосочетания, образованные по первому принципу (типа *at least*). Выяснилось, что, например, около 70% слов в корпусе LLC входят в повторяющиеся словосочетания. Впрочем, иногда случаются парадоксальные вещи. Так, некоторые словосочетания появляются в корпусе LOB лишь один раз, но носители языка однозначно воспринимают их как повторяющиеся и устойчивые: *at a first glance, at his mercy*.

Электронные корпусы позволяют исследовать проблему сочетаемости более объективно. Например, мы можем посчитать статистическую значимость разницы между тем, как фактически сочетаются слова в корпусе и тем, как они теоретически должны были бы сочетаться, если исходить из их индивидуальных частот. Например, слово Christmas часто сочетается с такими редкими словами, как day, eve, tree, cards и present.

Дальнейшие разработки в анализе корпусов открывают огромное поле для исследований

в области описания языка. И тут корпусная лингвистика может помочь определить, наконец, грань между лексикой и грамматикой.

## *Исследования частей речи*

Здесь можно говорить о работах, посвящённых различным аспектам использования глагольных форм, предлогов, союзов, наречий и т.п. Особое место занимают исследования глагольных форм, поскольку уже по результатам изучения Брауновского корпуса и корпуса LOB был сделан вывод о том, что глаголы составляют почти 20% всех слов корпуса, причём 23 наиболее часто употребляемых глагола составляют 95% всех глагольных словоупотреблений.

Существуют работы, связанные с распределением различных временных форм глагола. Так, известно, что в устном английском всегда преобладает время simple present (в частности, глагол *to be*). В письменном языке (особенно в нарративах) учащается использование simple past, но и настоящее время продолжает оставаться частотным.

Также изучаются модальные глаголы. Эти исследования показывают, как распределение их грамматических шаблонов и семантических ролей зависит от жанра и на какие формы следует обратить внимание при преподавании языка. Например, выяснилось, что в письменном английском самый употребительный модальный глагол — *would*, а вот в устной речи это *will*.

Корпусные лингвисты обращают внимание и на залог глагола. Так, Svartvik показал, что пассив гораздо чаще употребляется в информативной прозе, чем в художественной. Например, в научных текстах пассивный залог встречается почти в 8 раз чаще, чем в рекламных объявлениях.

Другие интересные корпусные исследования, связанные с частями речи:

- Постглагольные частицы. Самые частотные — *up* и *out*, за ними идут *off*, *back*, *down, on* и *in*.
- Предлоги. Каждое восьмое слово (12%) в любом английском тексте является предлогом, причём *of* составляет 30% всех предложных словоформ. Выяснился интересный факт — данный предлог (несмотря на его огромную частотность) функционирует не совсем так, как остальные предлоги. Большая часть предлогов входит в коллокации с последующими словами (*at least, for a while*), в то время как of тяготеет к сочетанию с предшествующими (*descriptions of, the basis of*). После выявления различий в сочетаемости предлогов, зародились обоснованные сомнения в том, можно ли вообще объединять эти слова в один класс — ведь они даже сочетаются с разными частями речи.
- Союзы. Исследовались различия в употреблении и многочисленных значениях *since*, *when* и *once*, а так же *more* и *less*. Кстати, было показано что очень большое количество слов тем или иным образом отражают количественные отношения (являются квантификаторами). Причём большая их часть выражает **неточное** количество — *few*, *substantial* и т.д. Это открытие имеет большое значение для лингводидактики.

## *Исследования синтаксических процессов на уровне предложения*

Изучать синтактику в корпусе сложнее, чем лексику. Основная причина — сложность автоматического синтаксического анализа языка (парсинга).

Тем не менее, при переходе от уровня слов на уровень фраз и предложений, корпусные методы не теряют своей значимости. Корпусная лингвистика убедительно показала ложность

бытующего мнения, что синтаксические конструкции сочетаются и варьируются абсолютно свободно. Напротив — они так же зависят от жанра, как лексика. Так, ещё в 70-х годах была доказана зависимость средней длины предложения от жанра текста: в информативной прозе предложения длиннее, чем в художественной (так называемый «флэш-тест»). Причём в основном это происходит не за счёт увеличения количества предикатных групп, а за счёт увеличения количества слов в этих группах. Например, средняя длина предложения в американских новостных текстах составляет 23,7 слова (от 3 до 70 слов).

Распространённые темы синтаксических корпусных исследований включают в себя:

- Типы придаточных предложений, в частности, лексикализованные сущности, типа *that's right* или *wait a minute*.
- Noun Phrases, например, изучение присоединённых конструкций (apposition) с выходом на их зависимость от того, разделяют ли продуцент и реципиент текста одну и ту же картину мира.
- Отношения условия (conditionality). Эти исследования, например, показали частые ошибки в тактике преподавания английской условности. Выяснилось, что наиболее частым способом передачи условности в английском языке является фактуальный, через simple present + simple present (*If you don't return, I die*). Между тем, во многих курсах грамматики этот тип условности вообще не рассматривается, зато большое внимание уделяется конструкциям с *would* и т.д., которые вовсе не так частотны.
- Сослагательное наклонение. Выяснилось, что американский английский тяготеет к конструкциям типа *It is important that he visits the lecture*, а британский — *We insist that he should visit it*.
- Отношения причинности (causation). Чаще всего выражаются словами *because, so* и *for*. Однако, последнее слово релевантно только для письменного языка — в устном оно фактически не встречается в значении причинности (также, как *since, thus* и *therefore*). Кстати, корпусная лингвистика позволяет существенно расширить представления о способах выражения причинности в английском. Даже лучшие учебники грамматики (Quirk) дают около 40 таких способов, в то время, как корпусные лингвисты Fang и Kennedy идентифицировали на основе корпуса LOB более 130.
- Отрицание (negation). Tottie показала, что в устной речи отрицание употребляется в два раза чаще, чем в письменной, а так же, что формы *no* и *not* далеко не всегда являются взаимозаменяемыми (*no point, there is not lot of money available*).
- Усиление (clefting). *It was him who did it*. Эта конструкция более частотна в письменных текстах, конкретно — в информативных.

## *Исследования прагматики и устной речи*

Развитие корпусной лингвистики в 60-х совпало с возросшим интересом к социальным и прагматическим функциям устной речи. Одно и то же высказывание может иметь разное значение в зависимости от обстоятельств. Интерес представляет и сегментирование устной речи. Ведь письменный текст достаточно легко членится на слова, фразы и предложения, а устный дискурс обычно делят на интонационные сегменты, что значительно труднее.

До середины 90х все корпусные исследования устной речи были основаны на корпусе London-Lund (LLC), который единственный имел полную разметку. Соответственно, его изучение можно было автоматизировать. Как пример исследования, можно привести изучение слова *well* в работах Svartvik. Он показал, что дискурсивная частица *well* используется в устной речи в значении идентификатора тематического сдвига. Но в письменной речи в таком значении это слово не используется практически никогда. Интересно также употребление так

называемых **hedge words** типа *sort of, kind of* и т.д. Исследовались усилительные наречия (*absolutely, completely; very, terribly*). В устной речи используются лишь немногие из них, но очень интенсивно (например, *quite*).

## *Исследования вариаций языка*

Мы уже упоминали, что результаты многих корпусных исследований показывали различия между вариациями английского языка. Остановимся на этом подробнее.

Огромные различия прослеживаются между письменным и устным английским. Их можно проследить на основе устного и письменного корпуса. Например, в письменном корпусе LOB слово *pretty* в 44% случаев играет роль прилагательного, в устном корпусе LLC эта цифра составляет лишь 4%. В устном английском это слово чаще употребляется как усилитель. Различны и частоты употребления таких слов, как *really, right* и *just*.

Открытия различий в частоте и функциях лексики и грамматики для разных текстов могут привести к масштабным изменениям в классификации жанров и типов текстов (см., например, D. Biber, «Variations across Speech and Writing»). Часто устный и письменный тексты могут быть ближе друг к другу, чем два письменных текста разных типов. Другие лингвисты считают, что устная речь всё же кардинально отличается от письменной, поскольку является не продуктом, а процессом.

Много работ посвящено сравнению региональных разновидностей английского языка (в частности, на материале корпуса ICE). В основном различия проявляются на уровне лексики: хотя в целом 50 наиболее частотных слов совпадают (за исключением *so*, которое частотно в британском английском, и *more*, которое частотно в американском). Грамматически языки скорее совпадают, разве что в американском варианте больше номинативных конструкций.

Корпусные исследования могут помочь разрешить сложные грамматические коллизии по поводу нормы. Так после слова *different* могут идти слова *than*, *to* или *from*. Неясно, какое из них более правильное. Исследования на материале Брауновского корпуса и LLC показали, что *different from* используется гораздо чаще остальных, а *different than* скорее характерно для американского английского.

Проводились исследования и других вариантов английского языка. Например, Collins показал, что в австралийском английском глагол *must* гораздо чаще употребляется в значении «*You must be joking*», чем в прямом. Фильм англичане и новозеландцы назовут *film*, а американцы и австралийцы — *movie*.

Естественно, что существует обширная литература по стилистическим различиям в языке (жанры и регистры). Вероятность употребления слова или грамматического явления зависит от регистра. В то же время, регистры могут «перетекать» друг в друга даже в пределах одного текста.

На основе диахронических корпусов (например Хельсинкского) проводились исследования изменений языка. Например, выяснилось, что начиная со средних веков, в английском увеличивается частота употребления глагольной формы progressive (be + ing).

Для переводчиков большой интерес представляют исследования бинарной оппозиции **«оригинальный текст — переводной текст»**. Имеется в виду выяснение, существуют ли объективные отличия переводных (вторичных) текстов от текстов первичных. Здесь можно назвать имена M. Baker и M. Olohan.

Итак, мы увидели, что описания таких аспектов языка, как лексика, морфология, синтактика и устная речь, на базе анализа корпусов дали нам очень много новых знаний о языке. Корпусная лингвистика пока не осуществила системное описание всех аспектов английского языка (тем более, русского), но сделано уже достаточно, чтобы предположить, что корпусные описания языка будут давать всё больше информации об относительных частотах явлений языка в разных его вариациях. В свою очередь, это повлияет на преподавание иностранных языков на всех уровнях системы образования.

# Корпусная лингвистика

Лекция 7

# Извлечение информации из корпуса

## *Методы извлечения информации из корпуса*

Лингвистическая информация из корпуса извлекается при помощи специальных компьютерных программ. Есть два основных источника разработки подобных программ. Во-первых, это лингвистические отделы больших коммерческих проектов, в основном, связанных с публикацией словарей. Например, Cobuild Project. Часто это закрытое программное обеспечение, стоящее больших денег. Второй источник разработки — компьютерная лингвистика и учёные, которые ей занимаются. В её рамках было создано немало программ, осуществляющих автоматический анализ грамматики и семантики, анализ и синтез текста, автоматический перевод и другие приложения для компьютерной обработки естественного языка. Конечно, не был обойдён стороной и анализ корпусов, в том числе, средства автоматической грамматической и синтаксической разметки — вероятностные (probabilistic), либо на основе правил (rule-based). Такие программы, разработанные самостоятельными исследовательскими группами (или даже отдельными учёными) часто бесплатны или вообще открыты для изучения (open source).

## *Типы извлекаемой информации*

Для поиска и извлечения информации из корпуса используется некоторое количество довольно стандартных процедур. Самый простой формат отображения информации о корпусе — это простые списки. Эти списки могут быть разных типов — от простых глоссариев до конкордансов. Давайте посмотрим на то, как всё это может быть представлено.

## Списки слов и конкордансы

Часто нужно разобраться со словами, которые употребляются в тексте. Список слов (word list) в самой простой своей форме – это попросту список всех слов, содержащихся в исследуемом тексте. Многие программы создают лемматизированные (lemmatized) списки, в которых разные грамматические формы слова показаны, как одно слово. Например, *goes* и *will go* будут показаны в одной строке с *go*. Иногда программа позволяет создать список не только по словам, но и по словосочетаниям из двух или трёх слов.

Часто этот список отсортирован по частоте встречаемости слов или по алфавиту. Такой список даёт базу для терминологических исследований и позволяет составить глоссарий. Например, возьмём такой текст:

*There are two possible approaches to automating the translation process:*

*Machine translation:*

*Machine translation has been a Holy Grail of the IT industry for more than 40 years. There have been significant advances in language technology over this period and we all benefit from these on a day to day basis when we use spelling and grammar checkers and ever more sophisticated search engines.*

*One of the fundamental reasons why machine translation has not so far produced convincing results is that language is more than mere words and grammar. Language conveys meaning and until you can clearly define and understand what is being conveyed you cannot hope to translate it. A good test of a Machine Translation system is to translate the text into the target language and then back again - the results can be quite comical.*

*Translation Memory:*

*Translation memory works by aligning previously translated text in a target language with the source language. This is accomplished either by the use of a manual tool, or automatically by using a controlled environment for the translation process. Alignment is usually done at a sentence level. This affords the best level of usable granularity. The aligned source and target text is held in a repository. The next time the document is updated the repository is searched in order to locate any text that has not changed. Where such a sentence is identified the source language text can be replaced with the target language text.*

*This relatively low tech method can nevertheless provide benefits in terms of translation consistency and reduced costs.*

Вот список всех словоформ или типов, которые встречаются в данном тексте, отсортированных по частоте:

| *Ранг* | *Частота* | *Слово* |
|---|---|---|
| *1* | *14* | *the* |
| *2* | *10* | *is* |
| *3* | *9* | *a* |
| *4* | *9* | *and* |
| *5* | *7* | *language* |
| *6* | *6* | *of* |
| *7* | *6* | *text* |
| *8* | *6* | *translation* |
| *9* | *5* | *in* |
| *10* | *5* | *to* |
| *11* | *4* | *can* |
| *12* | *4* | *target* |
| *13* | *3* | *by* |
| *14* | *3* | *has* |
| *15* | *3* | *Machine* |
| *16* | *3* | *more* |
| *17* | *3* | *source* |
| *18* | *3* | *This* |
| *19* | *3* | *Translation* |

| | | |
|---|---|---|
| *20* | *2* | *be* |
| *21* | *2* | *been* |
| *22* | *2* | *day* |
| *23* | *2* | *for* |
| *24* | *2* | *grammar* |
| *25* | *2* | *level* |
| *26* | *2* | *not* |
| *27* | *2* | *process* |
| *28* | *2* | *repository* |
| *29* | *2* | *results* |
| *30* | *2* | *sentence* |
| *31* | *2* | *than* |
| *32* | *2* | *that* |
| *33* | *2* | *The* |
| *34* | *2* | *There* |
| *35* | *2* | *translate* |
| *36* | *2* | *use* |
| *37* | *2* | *we* |
| *38* | *2* | *with* |
| *39* | *2* | *you* |
| *40* | *1* | *A* |
| *41* | *1* | *accomplished* |
| *42* | *1* | *advances* |
| *43* | *1* | *affords* |
| *44* | *1* | *again* |
| *45* | *1* | *aligned* |
| *46* | *1* | *aligning* |
| *47* | *1* | *Alignment* |
| *48* | *1* | *all* |

| | | |
|---|---|---|
| *49* | *1* | *any* |
| *50* | *1* | *approaches* |
| *51* | *1* | *are* |
| *52* | *1* | *at* |
| *53* | *1* | *automatically* |
| *54* | *1* | *automating* |
| *55* | *1* | *back* |
| *56* | *1* | *basis* |
| *57* | *1* | *being* |
| *58* | *1* | *benefit* |
| *59* | *1* | *benefits* |
| *60* | *1* | *best* |
| *61* | *1* | *cannot* |
| *62* | *1* | *changed* |
| *63* | *1* | *checkers* |
| *64* | *1* | *clearly* |
| *65* | *1* | *comical* |
| *66* | *1* | *consistency* |
| *67* | *1* | *controlled* |
| *68* | *1* | *conveyed* |
| *69* | *1* | *conveys* |
| *70* | *1* | *convincing* |
| *71* | *1* | *costs* |
| *72* | *1* | *define* |
| *73* | *1* | *document* |
| *74* | *1* | *done* |
| *75* | *1* | *either* |
| *76* | *1* | *engines* |
| *77* | *1* | *environment* |

| | | |
|---|---|---|
| *78* | *1* | *ever* |
| *79* | *1* | *far* |
| *80* | *1* | *from* |
| *81* | *1* | *fundamental* |
| *82* | *1* | *good* |
| *83* | *1* | *grail* |
| *84* | *1* | *granularity* |
| *85* | *1* | *have* |
| *86* | *1* | *held* |
| *87* | *1* | *holy* |
| *88* | *1* | *hope* |
| *89* | *1* | *identified* |
| *90* | *1* | *industry* |
| *91* | *1* | *into* |
| *92* | *1* | *IT* |
| *93* | *1* | *it* |
| *94* | *1* | *Language* |
| *95* | *1* | *locate* |
| *96* | *1* | *low* |
| *97* | *1* | *machine* |
| *98* | *1* | *manual* |
| *99* | *1* | *meaning* |
| *100* | *1* | *memory* |
| *101* | *1* | *Memory* |
| *102* | *1* | *mere* |
| *103* | *1* | *method* |
| *104* | *1* | *nevertheless* |
| *105* | *1* | *next* |
| *106* | *1* | *on* |

| | | |
|---|---|---|
| *107* | *1* | *One* |
| *108* | *1* | *or* |
| *109* | *1* | *order* |
| *110* | *1* | *over* |
| *111* | *1* | *period* |
| *112* | *1* | *possible* |
| *113* | *1* | *previously* |
| *114* | *1* | *produced* |
| *115* | *1* | *provide* |
| *116* | *1* | *quite* |
| *117* | *1* | *reasons* |
| *118* | *1* | *reduced* |
| *119* | *1* | *relatively* |
| *120* | *1* | *replaced* |
| *121* | *1* | *search* |
| *122* | *1* | *searched* |
| *123* | *1* | *significant* |
| *124* | *1* | *so* |
| *125* | *1* | *sophisticated* |
| *126* | *1* | *spelling* |
| *127* | *1* | *such* |
| *128* | *1* | *system* |
| *129* | *1* | *tech* |
| *130* | *1* | *technology* |
| *131* | *1* | *terms* |
| *132* | *1* | *test* |
| *133* | *1* | *then* |
| *134* | *1* | *these* |
| *135* | *1* | *this* |

| | | |
|---|---|---|
| *136* | *1* | *time* |
| *137* | *1* | *tool* |
| *138* | *1* | *translated* |
| *139* | *1* | *two* |
| *140* | *1* | *understand* |
| *141* | *1* | *until* |
| *142* | *1* | *updated* |
| *143* | *1* | *usable* |
| *144* | *1* | *using* |
| *145* | *1* | *usually* |
| *146* | *1* | *what* |
| *147* | *1* | *when* |
| *148* | *1* | *Where* |
| *149* | *1* | *why* |
| *150* | *1* | *words* |
| *151* | *1* | *works* |
| *152* | *1* | *years* |

В целом, в этом тексте 152 словоформы (types) и 259 словоупотреблений (tokens).

Уже по такому простому списку можно получить большое количество информации об употреблении слов в тексте. Например, можно видеть, что 10 самых частотных слов (*the, is, a, and, language, of, text, translation, in, to*) в целом соответствуют средним значениям для английского языка, за исключением появления трёх слов — language, text и translation. Здесь уже можно говорить о ключевых словах текста.

Однако, формат простого списка не даёт возможности снять полисемию и неоднозначность грамматического класса слова, поскольку это невозможно сделать без контекста. Чтобы разобраться с этим вопросом, нам нужно будет перейти к понятию «конкорданс» (concordance).

Конкорданс - это не просто список слов или словосочетаний. Его ценность в том, что он даёт контекст слова. То есть, мы можем запустить поиск и получить все появления данного конкретного слова в тексте. Результаты поиска показываются в формате, который называется KWIC (key word in context). Обычно при щелчке на строку программа-конкордансер выдаёт полный контекст.

Результаты поиска можно сортировать по-разному. Вы можете настроить программу на показ того или иного количества слов справа и слева от искомого термина. Также возможно изменять порядок строк конкорданса: например, если вы искали существительное, то можете

попросить конкордансер, чтобы он отсортировал в алфавитном порядке слова, непосредственно предшествующие слову поиска. Это поможет вам найти подходящие прилагательные, которые можно употреблять со словом поиска. Таким образом можно, например, обнаружить, что справа от слова computer очень часто стоят слова hardware, software и problem.

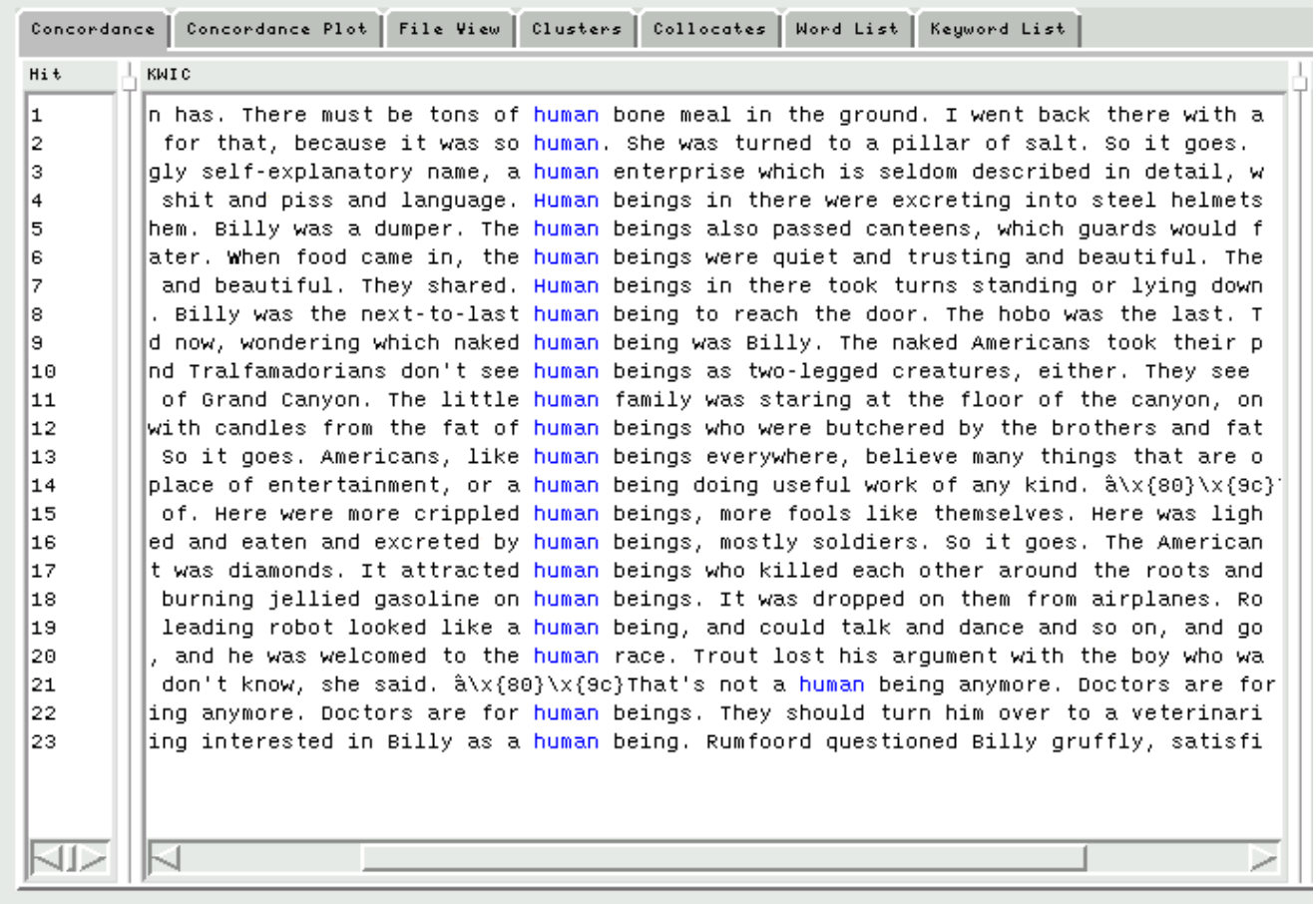


*Иллюстрация 1: Конкорданс, сгенерированный программой AntConc по слову "human" (поиск по роману Курта Воннегута Slaughterhouse-Five)*

Можно видеть, что конкордансы чрезвычайно полезны для изучения устойчивых словосочетаний (коллокаций). Мы можем искать типичные случаи употребления слов в одной коллокации.

Одной из наиболее распространённых программ-конкордансеров является WordSmith Tools[1] Майка Скотта из Оксфордского университета, но она платная. Учитывая, что автор живёт в Великобритании, купить её в России затруднительно. Впрочем, можно скачать демонстрационную версию с ограниченными возможностями. Практически ничем WordSmith не уступает бесплатный AntConc, разработанный японскими учёными[2]. В нём реализованы все необходимые функции — список слов, конкорданс, поиск коллокаций. Отечественная лингвистика может гордиться разработками группы «Автоматическая обработка текста»[3], среди которых есть и доступный для свободного скачивания конкордансер Dialing Concordance

1 http://www.lexically.net/wordsmith/
2 http://www.antlab.sci.waseda.ac.jp/software.html
3 http://www.aot.ru

(DDC). По возможностям он пока значительно уступает AntConc, но зато обладает встроенным морфологическим анализатором и способен понимать русское словоизменение, например, по запросу «студент», находить так же слова «студентов» и «студенткой». Недавно появился полностью свободный конкордансер Corsis[1] (ранее назывался Tenka Text), который стремится стать полнофункциональной заменой для WordSmith Tools. Он разрабатывается в Германии.

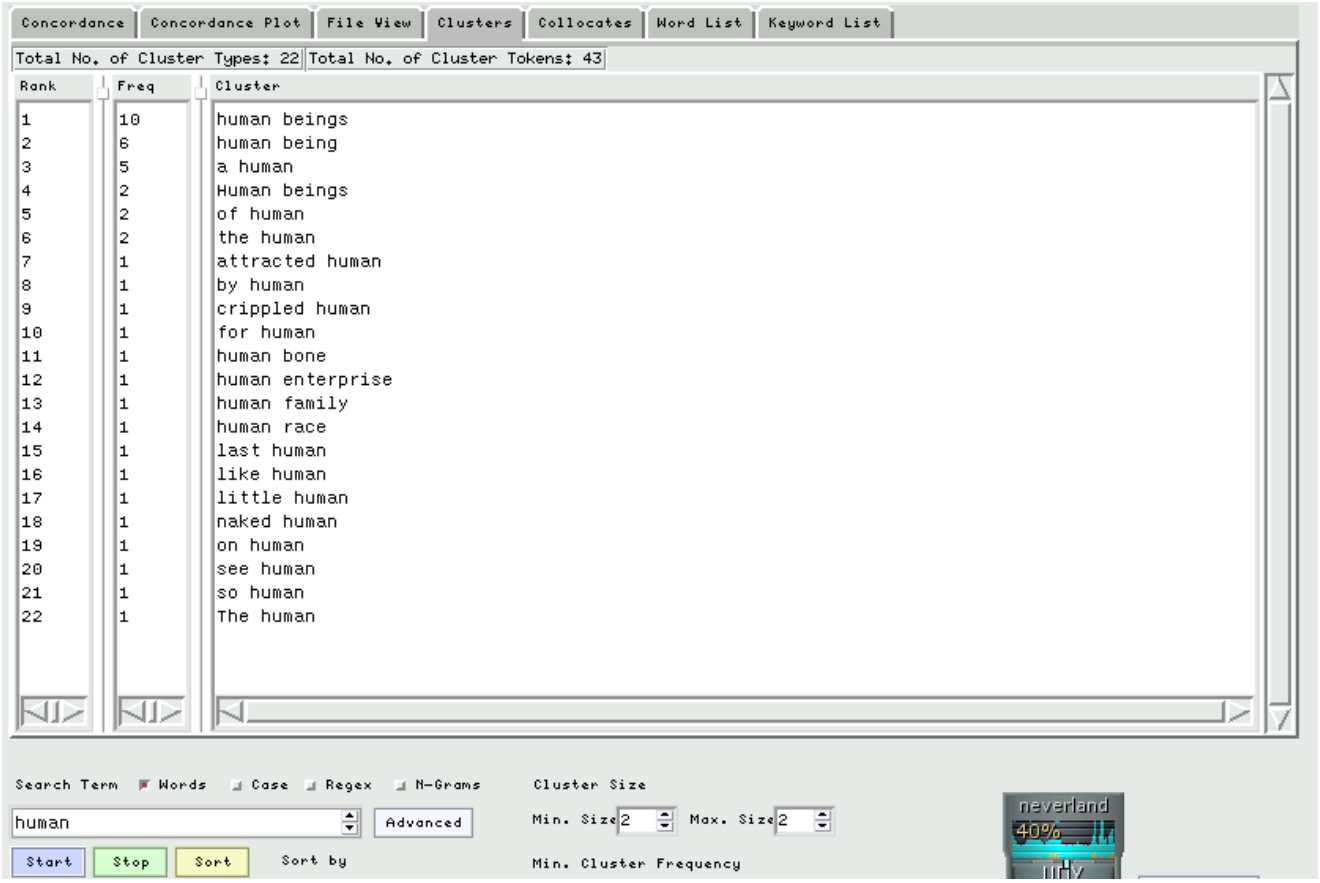


*Иллюстрация 2: Поиск коллокаций в программе AntConc*

Помимо конкордансов программы анализа корпусов обычно отображают и базовую статистическую информацию о корпусе: соотношение числа словоформ и словоупотреблений, среднюю длину предложения, количество предложений и их распределение по длине, индекс исключительности (каков процент слов, употреблявшихся лишь один раз), индекс постоянства (каков процент частых слов) и так далее.

## *Анализ корпуса на зачёт*

Для зачёта по лекционному курсу «корпусная лингвистика» (помимо успешного прохождения электронного теста) вам нужно будет защитить проект пробного корпусного исследования. Проект заключается в создании и анализе корпуса размером не менее 10 тысяч словоупотреблений. Корпус может быть двуязычным или одноязычным.

В проекте должны присутствовать следующие обязательные модули:

1. Обоснование корпуса, описание источников материала. Основная задача корпуса.
2. Выявление базовых параметров корпуса (*соотношение количества словоформ и*

1 http://corsis.sourceforge.net/

*словоупотреблений, индекс исключительности, индекс постоянства,наличие hapax legomena, относительная частота и коллокаты частых слов, составление полного конкорданса для одного-двух ключевых слов*) при помощи как минимум двух разных программ. Рекомендуется использование программ WordSmith, AntConc, Dialing Concordance, Corsis.

3. Краткое сравнение функциональности и удобства использованных программ с точки зрения корпусного лингвиста.
4. Описание какой-либо лингвистической особенности (особенностей) корпуса, которая вскрылась в процессе выполнения проекта. Выводы.

   Проект сдаётся преподавателю в электронном виде — корпус и сопутствующий текст.

Перед тем, как создавать свой корпус, будет небесполезно прочитать текст Джона Синклера «How to build a corpus», расположенный по адресу *http://www.ahds.ac.uk/creating/guides/linguistic-corpora/appendix.htm*

# Корпусная лингвистика

Лекция 9

# Создание собственного корпуса

Многие лингвисты работают с уже существующими масштабными корпусами — например, с British National Corpus или с корпусом Cobuild Project. Тем не менее, часто возникает необходимость изучить какие-то тексты, до сих пор не вошедшие в известные корпусы. Кроме того, не всегда есть возможность использовать эти «гиганты». В этом случае лингвист может составить свой собственный корпус из своих источников и исследовать уже его. В этой лекции мы рассмотрим некоторые вопросы, связанные с созданием своего корпуса — в основном, те, которые могут вызвать сложности при работе над проектом на зачёт.

## Планирование

Лингвистический корпус — это некое собрание текстов, в основе которого лежит логический замысел, логическая идея, объединяющая тексты. Тип корпуса и его структура зависят от его предназначения. Это означает, что прежде чем начать планирование корпуса, необходимо определиться, что же именно мы хотим исследовать. Например, нам интересны лингвистические особенности рекламных текстов в журнале *Cosmopolitan* в 2007 году. Здесь уже определена тематика текстов, а так же место и время их размещения. Это вполне конкретная лингвистическая задача. В данном случае корпус будет синхроническим, но вполне можно себе представить соответствующее диахроническое исследование — например, сравнение рекламных текстов в этом же журнале в 80-х годах и в 2000-ых. Отметим, что изучать просто «тексты из *Cosmopolitan*» было бы, наверное, некорректно, поскольку они принадлежат к разным категориям с разными коммуникативными интенциями. Для такого исследования нужно было бы сначала категоризировать все тексты этого журнала и затем собирать корпус с учётом этой классификации.

Принципиальным так же является решение об устном или письменном наполнении корпуса. Устные тексты сложнее разбить на категории, поэтому даже планирование устного корпуса связано с немалыми трудностями. Но, скорее всего, большинство студентов будут сдавать на зачёт письменный корпус, так что им это не грозит.

Важное значение имеет и размер корпуса. Ранее в лекциях говорилось об устоявшемся стандарте в 1 миллион словоупотреблений (word tokens, running words). На зачёт вам необходимо собрать корпус всего лишь не менее 10 тысяч словоупотреблений. Но это не значит, что он будет «неполноценным». Такой размер корпуса вполне достаточен для многих лингвистических исследований — разве что кроме изучения масштабных дискурсивных связей в длинных текстах. Есть хорошая фраза: «Не пытайтесь сочинить идеальный корпус — лучше подробно опишите имеющийся.»

При наборе текстов в корпус всегда желательно учитывать такие экстралингвистические факторы, как источники текстов, их авторы (их пол, возраст, профессия, национальность), носитель текста, место действия, тематика, дата публикации, возраст и размер предполагаемой аудитории и т.д.

Продолжим наш пример с *Cosmopolitan.* После того, как мы определились с тематикой и временем (рекламные тексты, 2007 год), нам нужно выбрать собственно сами тексты. И тут мы встаём перед выбором: либо взять в корпус **все** рекламные тексты журнала за этот год, либо

провести выборку (sampling) нескольких текстов, на базе анализа которых можно будет делать выводы о всех остальных. Выборку применяют очень часто, поскольку редко когда можно внести в корпус все интересующие нас тексты. В данном случае выборка играет роль модели явления. Выборка должна быть строго случайной, не зависящей от субъективных моментов. Например, мы можем выбирать 5 рекламных текстов из каждого номера за этот год, причём эти 5 текстов должны быть равномерно распределены по журналу (к примеру, быть расположенными на 10, 30, 50, 70 и 90 страницах). Аналогичным образом выборка осуществляется и в других случаях. Ещё раз отметим, что если явление, которое мы анализируем, является сложным, состоящим из нескольких классов, то и наша выборка (модель) должна отражать это деление.

Итак, перед тем, как начать составлять корпус, нам нужно знать следующее:

1. Какова логическая идея которая положена в основу корпуса?
2. С каким объёмом данных мы будем работать при составлении корпуса? Насколько это необходимо и реалистично?
3. Используем отрывки из текстов, полные тексты или то и другое?
4. Какова процедура отбора текстов в корпус?

## Сбор и оцифровка данных

В качестве источников текстов для корпуса можно использовать как цифровые, так и не цифровые носители. Естественно, в последнем случае понадобится каким-то образом ввести текст в компьютер: заново набрать его, либо отсканировать и распознать (конечно, с последующим редактированием). Например, в нашем случае с *Cosmopolitan* у нас нет электронных версий рекламных текстов, поэтому нам придётся приложить усилия для их оцифровки (приведения в computer-readable вид).

Однако в настоящее время большинство корпусов составляются из текстов, которые уже находятся в цифровой электронной форме (благодаря нарастающей компьютеризации). Для проекта на зачёт вам так же, скорее всего, будет логично использовать уже существующие электронные версии необходимых документов. Это резко снижает сложность составления корпуса.

Один из очевидных источников уже оцифрованных текстов — Интернет, который сам по себе является титаническим текстовым корпусом. В первую очередь, это, конечно, веб-страницы, но не нужно забывать и про другие интернет-каналы, по которым циркулируют огромные объёмы текстов: электронная почта, общение в ICQ и других мессенджерах, в социальных сетях, чаты, IRC и т.п. Можно использовать и другие источники текстов в электронном виде, если составитель корпуса может обосновать их привлечение.

Ввод в компьютер звуковых данных (в случае с устным корпусом) ещё более затруднён, но и результаты, которые может дать такой корпус более интересны.

## Формат и кодировка текста

Храните тексты для корпуса в простом текстовом формате (plain text, *.txt). Во-первых, он занимает меньше места, чем сложные форматы типа MS Word. Во-вторых, хотя современные программы анализа корпусов обычно могут работать с документами в формате HTML (XML), но всё-таки это менее надёжно, чем простой текст. Plain text — это простоя

последовательность букв, пробелов и знаков пунктуации. Такие файлы будет понимать любая программа везде и всегда, а при необходимости вы в любой момент сможете сконвертировать их в любой другой формат по своему выбору. Не храните ваши корпусы в MS Word — это не имеет никакого смысла! Кстати, не забывайте про резервные копии.

Ещё один тонкий момент — кодировка ваших файлов. Дело в том, что компьютеры создавались для работы на английском языке (точнее, с латинскими алфавитами). Отсюда многочисленные проблемы, связанные с тем, что нет чёткого договора, как именно компьютер должен обрабатывать и отображать символы других алфавитов (например, кириллицу, которой пользуется русский язык). Многие наверняка сталкивались с этим, когда видели в Интернете страницы, на которых текст представлен нечитаемыми «иероглифами» или вместо текста на экране оказывается бессмысленная последовательность кириллических букв. Это происходит из-за того, что существует несколько так называемых «кодировок» (англ. encodings), которые описывают русский алфавит — среди них **koi8-r** или **cp1251**. Ни одну из них нельзя назвать стандартом. Кроме того, не так давно появилась кодировка Unicode, которая поддерживает символы всех алфавитов всех языков мира, включая даже египетские иероглифы. Но пока не все программы готовы с ней работать.

Любой текстовый файл сохранён в одной из этих кодировок. Соответственно, если программа анализа корпуса считает, что кодировка одна, а на самом деле она другая — то файл будет прочитан неверно и вместо слов вы получите те самые бессмысленные наборы символов. Что тут можно посоветовать? Мы рекомендуем пользоваться либо Unicode (предпочтительнее), либо CP-1251. CP-1251 является стандартной кодировкой для MS Windows, а Unicode удобнее, поскольку может использоваться для любого языка. Когда вы сохраняете файл как «кодированный текст» в MS Word или в OpenOffice.org, то вам будет предложено выбрать кодировку.

Если вы анализируете текст в AntConc, то там вы можете указать кодировку для файлов, которые загружаете в него (в меню Global Settings — Encodings). Corsis воспринимает кириллические тексты только если они сохранены в кодировке Unicode. Dialing, напротив, считает, что кириллические тексты должны быть только в CP-1251. Но у вас всегда есть выход — вы можете сохранять один и тот же текст сколько угодно раз в различных кодировках.

С английскими текстами таких проблем нет, они будут нормально читаться и анализироваться вне зависимости от кодировки.

## Разметка (аннотирование) корпуса

Вы можете разметить свой корпус, то есть, добавить в тексты какие-то служебные пометки (например, части речи). В этом случае внимательно перечитайте лекцию номер пять, в которой говорилось о лингвистической разметке корпусов. Разметка поможет вам искать какие-то специфические места в текстах, но, учитывая небольшой размер корпусов, вряд ли имеет смысл разрабатывать масштабную систему разметки. Если же она всё-таки понадобится, то, скорее всего, нужно будет использовать при её создании стандарты XML и TEI.

## Хранение и презентация корпуса

Окончательный корпус должен соответствовать отраслевым стандартам и быть представлен, как продукт, готовый к отправке заказчику. То есть, он должен быть адекватно оформлен.