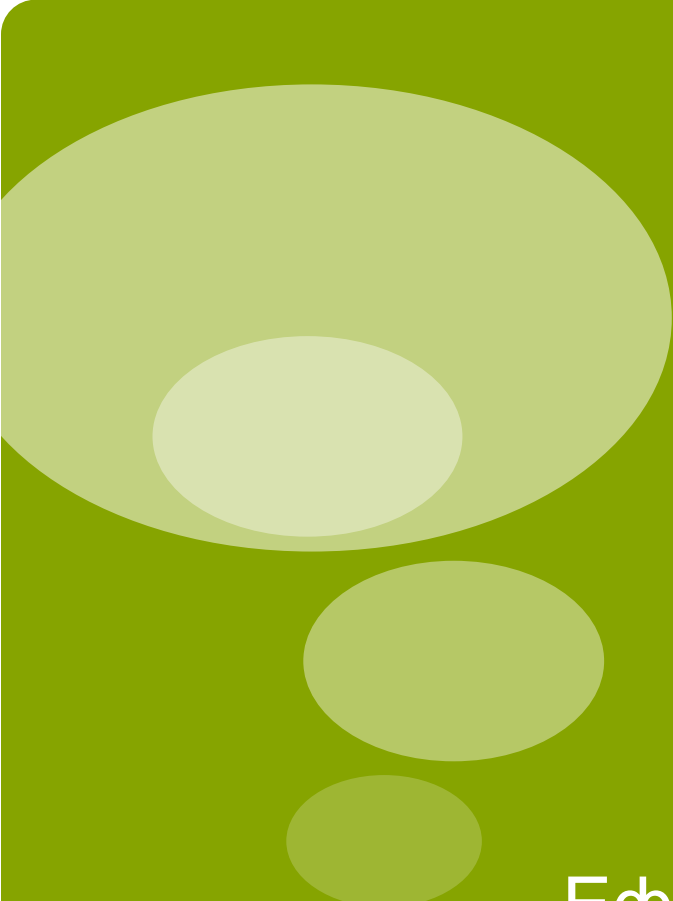




Корпусная ЛИНГВИСТИКА

Ефремова Наталья Эрнестовна
Грацианова Татьяна Юрьевна

Содержание



- Задачи корпусной лингвистики
- Типы корпусов
- Состав типичного корпуса
- Виды разметки текстов
- Примеры корпусов
- Параллельные и псевдопараллельные корпуса
- Интернет и корпусы
- Проблемы корпусной лингвистики

Корпусная лингвистика



- *Корпусная лингвистика* – теория и практика создания и использования *корпусов текстов*
- *Лингвистический, или языковой корпус* – представительный массив языковых данных
- Корпуса используются для
 - ✓ решения лингвистических задач
 - ✓ построения приложений компьютерной лингвистики (КЛ)
- Корпусная лингвистика зародилась в 1960-е годы в Америке

Первые корпуса



- Брауновский корпус (*Brown Corpus*):
 - ✧ 1960-е гг., Университет Брауна
 - ✧ состояние английского языка в США за 1961 гг.
 - ✧ 500 фрагментов текстов ≥ 2 тысячи слов, всего – 1 014 312 слов
 - ✧ для слов указаны части речи
 - ✧ были проведены исследования, созданы частотные словари
 - polite* – 7 раз, *polite letter* – 1 раз, *smile* – 0 раз
- В 1977 г. на основе корпуса текстов в 1 млн. слов был создан частотный словарь русского языка Л.Н. Засориной (около 40 000 слов)

Лингвистические задачи



Основная задача – проведение лингвистических и/или статистических исследований

- **объект:** язык, вид речи, текст, стиль/жанр и т.д.
- **цель исследования:**
 - ◆ анализ современного состояния
 - ◆ анализ изменений и различий (связанных со временем, географией, автором и т.д.)
 - ◆ проверка лингвистических теорий
 - ◆ исследование языковых и речевых явлений (например, типичные контексты употребления)
 - ◆ построение моделей этих явлений

Построение приложений КЛ



Корпуса используются при:

- получении эталонных статистических данных
- создании статистических моделей
- построении лингвистических ресурсов (словарей, тезаурусов, онтологий, ...)
- обучении и тестировании лингвистических процессоров (графематического, ...)
- обучении и тестировании прикладных систем (машинный перевод, оценка тональности, классификация текстов, снятие омонимии, ...)

Принципы создания корпуса



- **Соответствие решаемой задаче** (их может быть несколько)
- **Представительный объем** – типичность данных и полнота охвата изучаемых явлений
- **Естественность контекста данных** – возможность их всестороннего и объективного изучения
- **Разметка по определенным правилам**
- **Электронный вид и наличие корпусного менеджера** – многократное использование данных при решении различных задач
 - ➔ Разные задачи – разные типы корпусов

Корпуса текстов vs. коллекции текстов



Коллекция текстов – собрание текстов,
объединенных каким-то общим признаком
библиотека М. Мошкова, Wikipedia,
коллекция нормативно-правовых документов

Основные отличия коллекции от корпуса:

- ☐ коллекция решает нелингвистические задачи
- ☐ отбор текстов – на усмотрение составителей
- ☐ тексты рассматриваются не как образцы языковых явлений, а сами по себе
- ☐ тексты не имеют лингвистической разметки



Типы корпусов (1)

Определяются значениями признаков:

- ❑ Язык текстов и параллельность
русский, английский и т.д.
одноязычные, двуязычные, многоязычные
- ❑ Даты выхода текстов
синхронистический, диахронистический
- ❑ Стилъ и жанр
литературная и разговорная речь,
художественные, фольклорные и т.д.
- ❑ Размер текстов
полнотекстовые, фрагментотекстовые

Типы корпусов (2)



- ❑ Вид данных
письменные, речевые, мультимедийные
- ❑ Разметка и ее тип
*неразмеченные, размеченные:
морфологический, синтаксический и т.д.*
- ❑ Доступность
общедоступные, коммерческие
- ❑ Назначение
исследовательские, иллюстративные
- ❑ Динамичность
открытые (пополняемые), закрытые

Состав корпуса



Массив данных с разметкой (собственно корпус) и *корпусный менеджер*, который обеспечивает:

- поиск данных (слов и словосочетаний, контекстов употреблений):
 - ✓ по шаблонам
 - ✓ с учетом различных типов разметки
- получение статистической информации
- отображение результатов в удобной форме
- сохранение информации в различных форматах
- быструю работу с большими объемами данных

Примеры: *интерфейс поиска НКРЯ*,
Sketch Engine (RuTenTen11)

Разметка текстов



Разметка – приписывание специальных меток

Виды разметки:

- *экстралингвистическая* (сведения об авторе и тексте: *автор*, *название*, *дата создания*, ...)
- *структурная* (*глава*, *предложение*, *токен*, ...)
- собственно *лингвистическая* (*морфологическая*, *синтаксическая*, *семантическая*, ...)

Способы разметки:

- *автоматическая*: с помощью соответствующих анализаторов
- *ручная*: более качественная, снята омонимия
- ✓ сейчас в основном полуавтоматическая: сначала автоматическая разметка, потом ручная правка

Разметка: примеры



Морфологическая разметка (BNC)

<c c5="PUN">,</c>

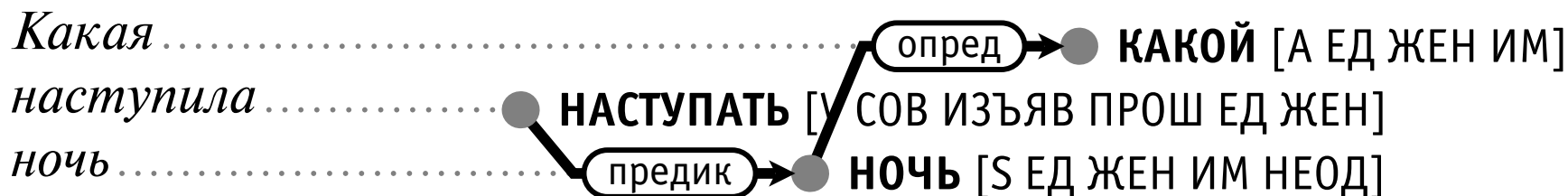
<w c5="AVQ" hw="where" pos="ADV">**where**</w>

<w c5="VBZ" hw="be" pos="VERB">**is** </w>

<w c5="AT0" hw="the" pos="ART">**the** </w>

<w c5="NN1" hw="body" pos="SUBST">**body**</w>

Синтаксическая разметка (НКРЯ)



Примеры корпусов. Британский национальный корпус (BNC)



Год создания	1991, закрытый, сбалансированный
Язык	британский английский
Размер	100 млн словоупотреблений
Назначение	образец типичного разговорного и письменного языка конца XX в.
Вид данных	письменная (художественная литература, газеты, журналы, ...) и устная речь
Разметка	частеречная
Доступность	поиск через корпусный менеджер Xaira; возможно скачивание (нужно принять условия)
Формат	XML

Чешский национальный корпус (ČNK)



Год создания	с 1994, открытый
Язык	чешский, английский, немецкий, ...
Размер	9 млрд словоупотреблений, из них 6,248 – в корпусе иностранных языков
Назначение	сохранение норм языка, международное сотрудничество
Вид данных	письменная (газеты, словари, книги. ...) и устная речь, параллельный корпус, ...
Разметка	морфологическая
Доступность	поиск через корпусный менеджер Manatee, полный доступ – после регистрации

Мангеймский корпус немецкого языка



Год создания	с 1964, открытый, несбалансированный
Язык	немецкий
Размер	> 4 млрд словоупотреблений
Назначение	основа для научного исследования современной немецкой письменной речи
Вид данных	тексты с 1956 года (научная литература, газеты, ...), устной речи
Разметка	морфологическая, метаданные
Доступность	для зарегистрированных пользователей поиск и анализ данных через систему COSMAS II

Уппсальский корпус РЯ



Год создания	Швеция, 1980 г., закрытый
Язык	русский
Размер	1 млн словоупотреблений
Назначение	отражение языка 1960-1988 гг.
Вид данных	600 текстов (художественная литература, информативные тексты) за 1985-1989 гг.; записаны латиницей
Разметка	морфологическая (при помощи TnT)
Доступность	поиск при помощи программы CQP
Сейчас входит в состав түбингенских корпусов русских текстов	

Открытый корпус РЯ (OpenCorpora)



Год создания	с 2009, открытый, создается и редактируется сообществом
Язык	русский
Размер	1,536 млн словоупотреблений
Назначение	создать морфологически, синтаксически и семантически размеченный корпус текстов на РЯ, в полном объеме доступный для исследователей
Вид данных	тексты (блоги, новости, художественная литература, ...)
Разметка	морфологическая
Доступность	поиск по словарю, доступен для скачивания

Национальный корпус русского языка (НКРЯ)



Год создания	с 2004 (в Интернет), открытый, сбалансированный
Язык	русский, параллельные корпуса
Размер	> 600 млн слов
Назначение	представляет язык во всём его многообразии, обеспечение научных исследований
Вид данных	письменная (с XVIII в.) и устная речь, мультимедиа
Разметка	метатекстовая, морфологическая, синтаксическая, акцентная и семантическая
Доступность	поиск с помощью корпусного менеджера, при подписании соглашения – получение подкорпуса ≈ 1 млн словоупотреблений

Корпусный менеджер НКРЯ: запрос



Поиск точных форм ?

Слово или фраза

Лексико-грамматический поиск ?

Слово ?

Грамм. признаки ? [выбрать](#)

Семант. признаки ? [выбрать](#)



Доп. признаки ? [выбрать](#)

Словообразование [выбрать](#)

☒ 1-е знач. ☒ др. знач. ☐ фильтр 1 ☐ фильтр 2 ?

Расстояние: от до ?

Слово ?

Грамм. признаки ? [выбрать](#)

Семант. признаки ? [выбрать](#)



Доп. признаки ? [выбрать](#)

Словообразование [выбрать](#)

☒ 1-е знач. ☒ др. знач. ☐ фильтр 1 ☐ фильтр 2 ?

Корпусный менеджер НКРЯ: результаты



1. Чтой-то случилось (2003) // «Марийская правда» (Йошкар-Ола), 2003.01.10 [омонимия снята] [Все примеры \(1\)](#)

В отличие от прошлых лет, даже в Рождественскую ночь **гуляющие** горожане смогли полюбоваться ледяными скульптурами (раньше от них, как правило, оставались жалкие глыбы льда). [Чтой-то случилось (2003) // «Марийская правда» (Йошкар-Ола), 2003.01.10] [омонимия снята] [←...→](#)

2. Эльвира Савкина. Мах Мага вывела vip-леди Самары на подиум (2002) // «Дело» (Самара), 2002.05.26 [омонимия снята] [Все примеры \(1\)](#)

Уже в четыре часа дня в помещении Мах Мага можно было наблюдать беспорядочно снующих моделей и грустно **гуляющих** по залу журналистов в ожидании почётного гостя показа — президента фонда "Русский силуэт" Татьяны Михалковой. [Эльвира Савкина. Мах Мага вывела vip-леди Самары на подиум (2002) // «Дело» (Самара), 2002.05.26] [омонимия снята] [←...→](#)

3. Вера Белоусова. Второй выстрел (2000) [омонимия снята] [Все примеры \(1\)](#)

Чтобы согреться, я немного попрыгал на месте под удивлёнными взглядами редких **гуляющих** и пустился трусцой наугад по какой-то аллее. [Вера Белоусова. Второй выстрел (2000)] [омонимия снята] [←...→](#)

4. Фазиль Искандер. Курортная идиллия (1999) [омонимия снята] [Все примеры \(1\)](#)

Я выбрал себе извилистый путь в этом маленьком крымском городке и очутился в незнакомом месте, хотя густая толпа **гуляющих** казалась той же самой, что и на нашей улице. [Фазиль Искандер. Курортная идиллия (1999)] [омонимия снята] [←...→](#)

НКРЯ: состав



- **Основной корпус:** прозаические письменные тексты с XVIII до начала XXI века
- **Синтаксический (глубоко аннотированный):** с морфологической и синтаксической разметкой
- **Газетный:** статьи 1990-2000-х годов
- **Параллельные корпуса**
- **Диалектных текстов:** запись диалектной речи различных регионов России
- **Поэтических текстов**
- **Обучающий:** со снятой омонимией
- **Устной речи**
- **Акцентологический:** история русского ударения
- **Мультимедийный:** снабжённые видео- и аудиорядом фрагменты кинофильмов 1930-2000-х годов

НКРЯ: параллельные корпуса



- ❖ Тексту на русском языке сопоставлен его перевод на другой язык и/или наоборот
- ❖ Тексты выравнены: между предложениями оригинального и переводного текста установлено соответствие
- ❖ Используется для научных исследований (теория и практика перевода), при обучении ЕЯ
- ❖ Доступны тексты на английском, немецком, испанском, итальянском, польском, украинском, белорусском, китайском и пр. языках
- ❖ Объем – 72,2 млн слов
- ❖ Имеется многоязычный подкорпус

Параллельные корпуса НКРЯ: пример поиска



Для слова *солнце* и *китайского* языка получим:

1. Н. А. Островский. Как закалялась сталь (1932) [омонимия не снята] [Все примеры \(1\)](#)

ru	Утреннее солнце лениво подымалось из-за громады лесопильного завода. [Н. А. Островский. Как закалялась сталь (1932)] [омонимия не снята] ←...→
zh	早晨的太阳从锯木厂高大的厂房后面懒洋洋地升起来。[尼·奥斯特洛夫斯基 / Ni. Aositeluofusiji. 钢铁是怎样炼成的 / Gangtie shi zenyang lianchengde] [омонимия не снята] ←...→
zh_2	zǎochén de/dī/dí/dì tàiyang cōng/cóng/zòng jùnmùchǎng gāodà de/dī/dí/dì chǎngfáng hòumian lǎnyāngyāng de/dì shēngqǐ lái. [尼·奥斯特洛夫斯基 / Ni. Aositeluofusiji. 钢铁是怎样炼成的 / Gangtie shi zenyang lianchengde] [омонимия не снята] ←...→

2. 鲁迅 / Lu Xun. 狂人日记 / Kuangren riji (1918) [омонимия не снята] [Все примеры \(1\)](#)

zh	太阳也不出，门也不开，日日是两饭。[鲁迅 / Lu Xun. 狂人日记 / Kuangren riji (1918)] [омонимия не снята] ←...→
zh_2	tàiyang yě bù chū, mén yě bù kāi, rìrì shì liǎng fàn. [Лу Синь. Записки сумасшедшего (С. Тихвинский,))] [омонимия не снята] ←...→
ru	И солнце не всходит, и двери не открываются: кормят два раза в день. [Лу Синь. Записки сумасшедшего (С. Тихвинский,))] [омонимия не снята] ←...→

Псевдопараллельные (сопоставимые) корпуса



- ❖ «Близкие» тексты на более чем одном языке.
Тексты оригинальные, не являются переводами
- ❖ Виды псевдопараллельных корпусов (С. Шаров):
 - «шумные» параллельные: допускается небольшая адаптация исходного текста к языку перевода
 - строго сопоставимые: не являются переводами, но посвящены одной теме (*статьи Wikipedia*)
 - слабо сопоставимые: посвященные одной теме, но созданные независимо друг от друга (*учебники*)
- ❖ Терминологическое выравнивание: сопоставление по терминологическим единицам и устойчивым словосочетаниям с ними

Aranea – семейство псевдопараллельных корпусов для 18 языков, созданы в одно время по одному принципу

Интернет как корпус



- Доступ через поисковые системы (*Googleology*)
- Интерпретация результатов – по числу страниц или по первым ссылкам
- Проблемы
 - ✓ статистика примерна и нестабильная: дублирование текстов, скрытое цитирование, географическая привязка и др.
 - ✓ алгоритмы работы поисковой системы меняются
 - ✓ недостаточная разметка
 - ✓ неоднородность интернет-страниц: содержание vs. оформление, реклама
- Последняя тенденция: автоматическое получение корпуса из Интернета



Интернет-корпуса РЯ

Сбор и разметка текстов из Рунета автоматически

RuTenTen11

- 2011 год, 14,5 млрд слов,
- без дубликатов
- морфологическая разметка
- поиск с помощью SketchEngine

ГИКРЯ

- с 2013, > 20 млрд слов, открытый: ВКонтакте, Живой Журнал, блоги Мейл.ру, ...
- метаразметка, токены, морфологическая разметка (TnT)
- для скачивания доступен «Серебряный стандарт» в 2 млн. словоупотреблений

Разметка ГИКРЯ: пример



TEXTID=9110_1*****77536943_1049.dat

536943	0001	УХ_ты	[УХ_ты]	I
536943	Р	!		
536943	0002	красивая	[красивый]	Apfsnf
536943	0003	очень	[очень]	Rp
536943	Р	!		

TEXTID=9112_1*****77536945_1049.dat

536945	0001	ага	[ага]	Ncfsn-y
536945	Р	,		
536945	0002	работа	[работа]	Ncfsn-n
536945	0003	с	[с]	Sps-i
536945	0004	настроением	[настроение]	Ncnsi-n
536945	Р	.		
536945	0005	у	[у]	Sps-g
536945	0006	Ани	[Аня]	Npfsg-y
536945	0007	еще	[ещё]	Rp
536945	0008	много	[много]	Rp
536945	0009	красоты	[красота]	Ncfsg-n
536945	0010	в	[в]	Sps-l
536945	0011	ЖЖ	[ЖЖ]	Npmi--n
536945	Р	.		

Проблемы Интернет-корпусов



- Опечатки, ошибки капитализации, обрывки слов, имена собственные, слова из других языков, экспрессивная лексика, новые слова
слу-шаю-с, хрюкмены, щаскакдам, приве-е-ет, иоаннутый, советобоязнь, нью-вэйв
- Частотного словаря *Araneum*, 5 млн слов

Янукович	240
Януковичем	142
я-то	129
ЯНАО	83
яже	80
Яннушка	47
явл-ся	29

які	28
Яффо	19
якудза	11
языкъ	11
яться	10
языкоблудием	1
языкк Ncmsan	1

Ngram Viewer



- На основе Google Books, 9 языков, есть разметка
- Для РЯ > 67 млрд словоупотреблений
- Поиск по словам, биграммам, частям речи, ...

Google Books Ngram Viewer

Graph these comma-separated phrases: бурные аплодисменты ☐ case-insensitive

between 1900 and 2000 from the corpus Russian with smoothing of 0 [Search lots of books](#)



Проблемы корпусной лингвистики: создание корпуса



Трудоёмкий процесс:

- ☐ выработка критерия отбора текстов
- ☐ сбор тысяч текстов
- ☐ решение проблем с авторскими правами
- ☐ *балансировка* корпуса
- ☐ приведение текстов к единому формату
- ☐ разработка специализированного ПО
- ☐ выбор типов и формата разметки
- ☐ разметка текстов

Сбалансированность и представительность



- ❑ *Сбалансированность* – соответствие типов текстов их доле в языке
- ❑ *Представительность* – всестороннее отражение явлений языка (исходя из задачи). Добавление новых данных практически ничего не меняет
- ❑ Нет четких алгоритмов их обеспечения
- ❑ Корпус конечен, поэтому неизбежны:
 - ✓ случайность подбора текстов
 - ✓ неадекватность отражения всего языка

Сейчас корпус считается сбалансированным и
представительным, если имеется негласный
договор об этом между
его создателями и пользователями

Масштабируемость



Пусть идеальный корпус создан, но:

- ❑ частота слова в одном тексте может влиять на позицию слова в общем частотном списке
- ❑ сложно определить позиции малочастотных слов
- ❑ есть омонимичные словоформы
- ❑ есть неизвестные слова (при автоматическом анализе)
Барклай – Барклаивать
- ❑ будут ли результаты идентичны на другом аналогичном корпусе?
- ❑ с ростом объема корпуса количество различных слов в нем увеличивается (Закон Хипса)

**Получаемые по корпусу данные не всегда
масштабируемы на весь язык**

Закон Хипса



- В 1960 году чешский лингвист Г. Хердан описал закон, который впоследствии был ошибочно приписан другому лингвисту – Х.С. Хипсу:

с ростом объема текста (корпуса) количество различных слов в нем увеличивается

$$V = K * N^{\beta}$$

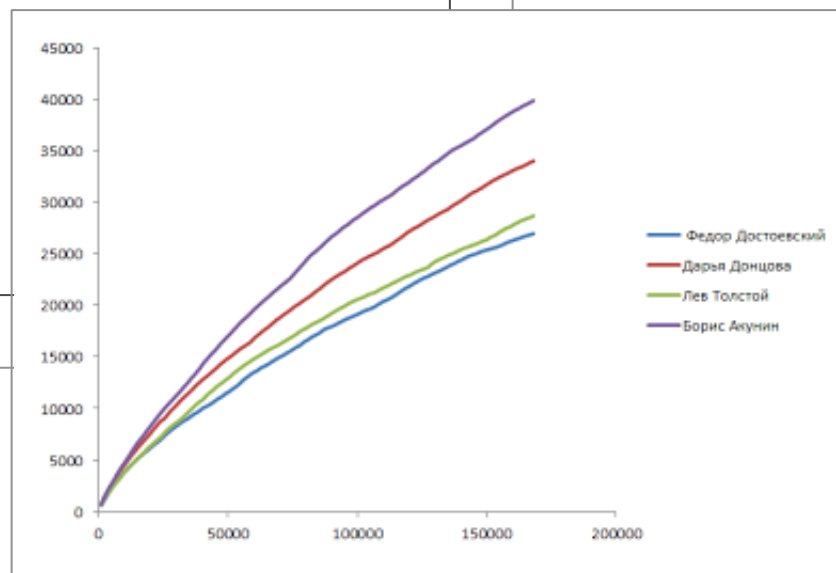
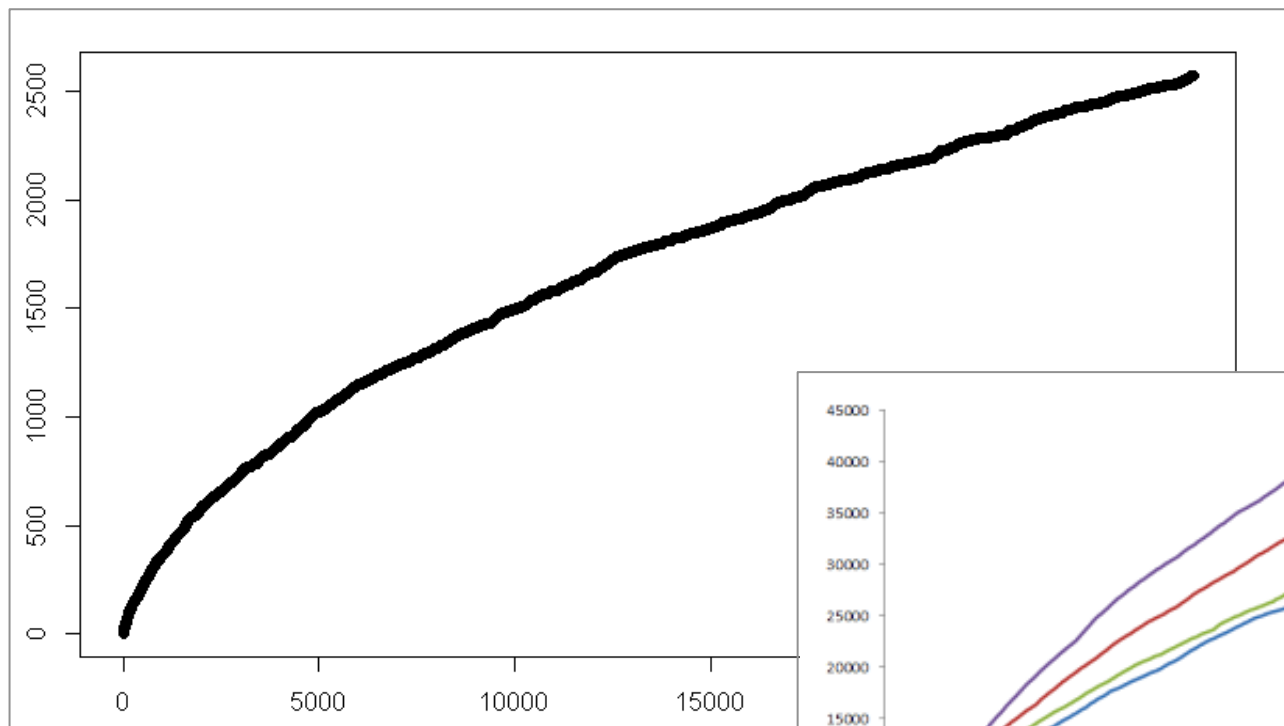
V – количество уникальных словоформ

β и K – константы, для европейских языков:

$$10 \leq K \leq 100 \quad 0,4 \leq \beta \leq 0,6$$

- Быстрый рост лексикона обеспечивается редкими словами (названиями и собственными именами)
- Отклонения от закона при добавлении в корпус нового текста может означать наличие плагиата

График закона Хипса





Выводы

- Для разных задач нужны разные корпуса
- Нельзя создать один универсальный корпус
- Создание корпуса сложно и трудоемко
- Необходимо критично относиться к получаемым результатам (особенно из Интернет). Например, НКРЯ
 - ✓ не отображает современные явления
спиннер, лагать – по 1 употреблению
 - ✓ не отображает региональные/социальные/профессиональные явления
корпусная лингвистика – 1 употребление
 - ✓ плохо отражает редкие явления:
белиберда встречается чаще, чем *хрюкнуть*

Несмотря на все проблемы пока корпуса – наиболее эффективный способ изучения языковых явлений



Спасибо за внимание!

Проект Татоэба



- ✓ > 2 млн фраз и предложений
- ✓ > 130 языков
- ✓ позволяет добавлять новые и изменять существующие предложения
- ✓ содержимое можно скачать бесплатно



#		Желаю счастья!	
→		お幸せに！	
→		I hope you'll be happy forever.	
→		Semoga bahagia!	
→		Ti auguro tanta felicità.	
→		Jag önskar dig all lycka i framtiden.	
→		Jag önskar dig all lycka i livet.	

Собрание текстов Достоевского



- Все художественные произведения, публицистика и эпистолярное наследие писателя
- Единицей хранения является отдельное произведение
- Источник для словаря языка Достоевского: готов частотный словарь, сформирована база данных по идиоматике Достоевского