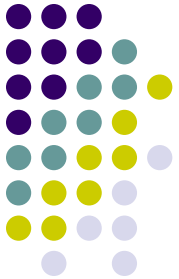
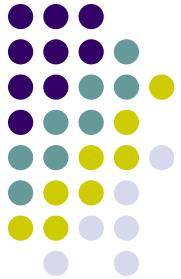


ТЕКСТОВЫЕ КОРПУСА И СТАТИСТИКА

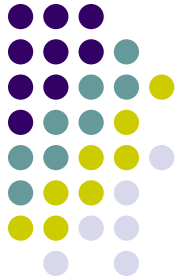
Большакова Елена Игоревна

СОДЕРЖАНИЕ



1. Текстовые корпуса
   - Корпуса и коллекции
   - Типы корпусов, состав, разметка, интерфейс
   - Проблемы корпусов, Применение в КЛ
2. Квантитативная лингвистика
   - Основные понятия
   - Статистика букв и буквосочетаний
   - Статистика слов, закон Ципфа
3. Статистические языковые модели
   - Особенности N-граммной модели
   - Сглаживание и перплексия
   - Применение языковых моделей в КЛ
4. Заключение, Домашнее задание № 1

КОРПУСНАЯ ЛИНГВИСТИКА



Зародилась в 1960-е годы в США

- Теория и практика создания и использования представительных массивов языковых данных для
  - лингвистических исследований ЕЯ
  - построения приложений КЛ и АТ
- Цели и задачи лингвистических исследований:
  - анализ современного состояния ЕЯ, а также его изменений во времени, в связи с географией и т.д.
  - проверка лингвистических теорий и моделей
  - исследование конкретных языковых явлений (типичные конструкции/выражения, жанры, стили и др.)
- ***Корпус*** – подобранная и обработанная по определенным правилам совокупность текстов (или образцов устной речи), используемых в качестве базы для исследования языка

КОРПУС ТЕКСТОВ

Представительный массив языковых данных

Принципы организации корпуса:

- предназначен для решения определённых задач (лингвистические/статистических исследования и т.п.)
- собран по принципу полноты охвата изучаемых языковых явлений: существенный объем данных
- сбалансирован по представительности (частоте, типичности) этих явлений
- представлен в электронном виде

Как правило, современный корпус:

- размечен по определенным правилам
- часто обеспечен специализированным ПО: *корпусным браузером* /*менеджером* – для многократного использования при решении различных задач

КОРПУС *Vs.* КОЛЛЕКЦИЯ ТЕКСТОВ

Коллекция текстов – собрание текстов, объединенных каким-то общим признаком

библиотека М. Мошкова, *Wikipedia*, *коллекция нормативно-правовых документов*

Основные отличия текстовой 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

Какие еще есть коллекции?

ОСОБЕННОСТИ КОРПУСОВ

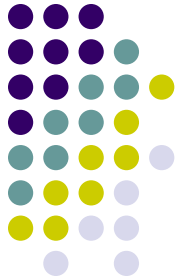
Признак	Значение
Язык текстов и параллельность	*русский*, *английский* и т.д.; *одно-*, *дву-*, *многоязычные*
Временные рамки текстов	*синхронистический*, *диахронистический*
Стиль и жанр	*литературная и разговорная речь*, *художественные, научно-технические, публицистические, фольклорные и т.п.*
Размер текстов	*полнотекстовые*, *фрагментотекстовые*
Вид данных	*письменные*, *речевые*, *мультимедийные*
Разметка и ее тип	*неразмеченные*, *размеченные*: *морфологический*, акцентологический и т.д.
Доступность	*общедоступные*, *коммерческие*
Назначение	*исследовательские*, *иллюстративные*
Динамичность	*открытые* (*пополняемые*), *закрытые*

ПРИМЕРЫ КОРПУСОВ

* Объем корпуса: в словоупотреблениях/токенах

Название	Язык	Год	Объем*	Особенности
Brown Corpus	амер.	1963	1 млн.	500 фрагментов за 1961 г.
BNC	брит.	1991	100 млн.	письмо и речь конца 20 в.
ČNK	чешск.	1994	9 млрд.	письмо и речь; параллельные подкорпусы
Уппсальский корпус	рус.	1980гг.	1 млн.	тексты за 1985-1989 гг. (язык 1960-1988 гг.)
НКРЯ	рус.	2004	600 млн.	письмо, речь, мультимедиа; от сер. 18 до нач. 20 в.; параллельные подкорпусы
Open Corpora	рус.	2009	1,5 млн.	создание сообществом; открытый
RuTenTen11	рус.	2011	14,5 млрд.	тексты из Интернет; открытый
ГИКРЯ	рус.	2013	20 млрд.	тексты из Интернет; открытый

СОСТАВ КОРПУСА



Массив данных с разметкой (собственно корпус)

и *Корпусный менеджер*, который обеспечивает:

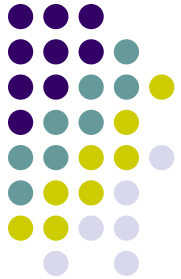
- поиск данных (слов и словосочетаний, их контекстов)
  - ✓ по признакам и шаблонам
  - ✓ с учетом различных типов разметки
- получение статистической информации
- отображение результатов в удобной форме

опционально:

- сохранение информации в различных форматах
- быструю работу с большими объемами данных

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

РАЗМЕТКА ТЕКСТОВ



Разметка – приписывание специальных меток тексту и лингвистическим единицам в нем

Виды разметки:

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

Способы разметки:

- автоматическая: с помощью соответствующих анализаторов
- ручная: более качественная, снята омонимия
- автоматизированная (наиболее часто сейчас): сначала автоматическая разметка, потом ручная коррекция

РАЗМЕТКА: ПРИМЕРЫ

Морфологическая разметка (BNC) для *, where is the body*:

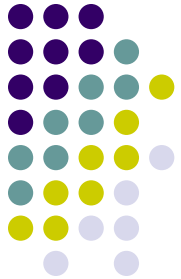
```
<c c5="PUN">,</c>
<w c5="AVQ" hw="where" pos="ADV">where</w>
<w c5="VBZ" hw="be" pos="VERB">is </w>
<w c5="AT0" hw="the" pos="ART">the </w>
<w c5="NN1" hw="body" pos="SUBST">body</w>
```

Морфологическая разметка (НКРЯ)

хорошо хороший ADJ Gender=Neut|Number=Sing|Variant=Short
бы бы PART _ _
иметь иметь VERB VerbForm=Inf|Voice=Act|Aspect=Imp
несколько несколько NUM Animacy=Inan|Case=Acc
жизней жизнь NOUN Case=Gen|Gender=Fem|Number=Plur

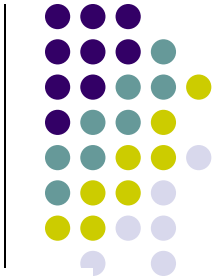
Синтаксическая разметка (НКРЯ):

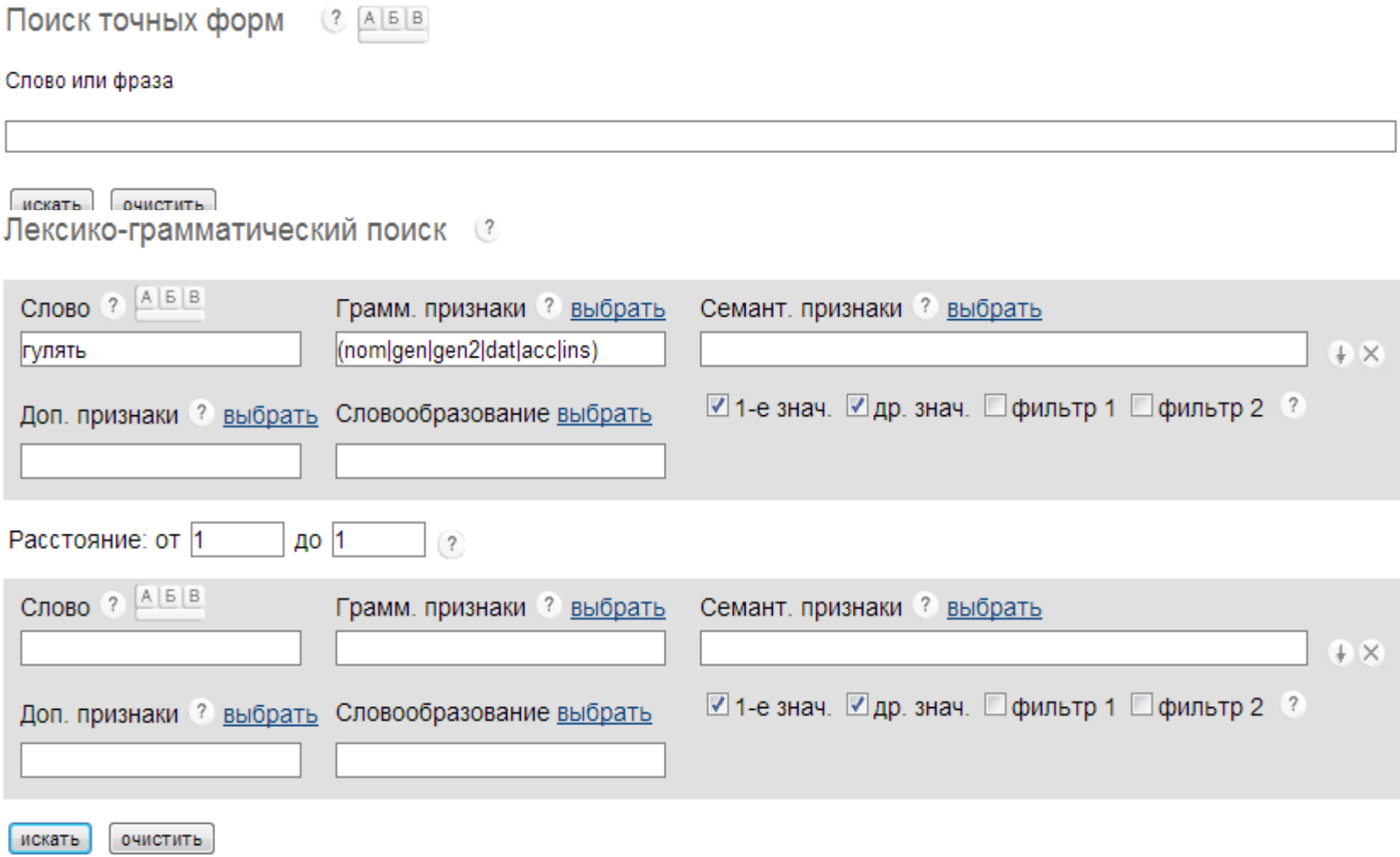
НКРЯ: СВОЙСТВА, ИНТЕРФЕЙС



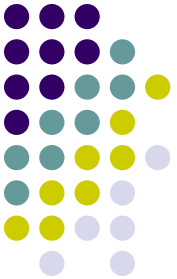
- *Национальный Корпус Русского Языка:* открытый, сбалансированный, представляет язык во всём его многообразии (письменная с XVIII в. и устная речь),
- Разметка: метатекстовая, морфологическая, синтаксическая, акцентная и семантическая
- Интерфейс (браузер): поиск осуществляется на основе существующей разметки, доступен по:
  - слову или фразе (точное совпадение), лемме
  - части речи и грамматическим признакам (падеж, род, число, время и пр.)
  - синтаксическим отношениям (предикативные, сочинительные и пр.)
  - семантическим группам (животные, время, еда и др.)
  - вокалической структуре (ударный гласный, количество слогов и др.) и т.д.

ИНТЕРФЕЙС ПОИСКА НКРЯ: ЗАПРОС





ИНТЕРФЕЙС ПОИСКА 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. Max Mara вывела vip-леди Самары на подиум (2002) // «Дело» (Самара), 2002.05.26** [омонимия снята] **Все примеры (1)**

   Уже в четыре часа дня в помещении Max Mara 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. Max Mara 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] ←…→

SKETCH ENGINE: ОСНОВНЫЕ ФУНКЦИИ

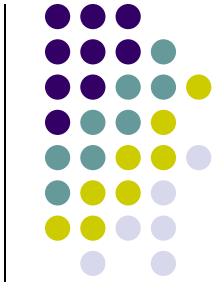
Развитый корпусный менеджер: https://www.sketchengine.co.uk

- создание (*краулинг*) и сравнение корпусов (для РЯ был создан корпус *RuTenTen11*)
- построение конкордансов для слова/словосочетания
- поиск биграмм с заданным словом; пример для *goal* :

goal *(noun)* **ukWaC freq = 168345** (107.5 per million)

object_of	58924	3.2	subject_of	25451	2.4	modifier	67879	1.6	modifies	11026	0.3
score	8390	11.28	score	903	8.59	ultimate	1911	9.27	scorer	389	9.39
achieve	9422	9.9	disallow	223	8.04	long-term	875	7.66	kick	634	8.86
concede	1421	9.39	concede	204	7.53	league	638	7.38	tally	129	7.9
accomplish	585	7.97	gape	76	6.5	winning	401	7.33	keeper	204	7.31
reach	1924	7.66	come	1316	5.44	primary	993	7.24	scramble	50	6.75
net	337	7.42	kick	76	5.44	second	2000	7.19	drought	78	6.65
pursue	648	7.41	rule	61	5.24	common	1529	7.17	difference	676	6.28
attain	400	7.35	orientate	34	5.06	strategic	645	7.1	cushion	53	6.26
grab	406	7.34	arrive	90	4.43	realistic	422	7.05	lead	267	6.24
set	2413	7.01	cap	20	4.38	achievable	290	6.97	setting	405	6.14
pull	501	6.88	beat	53	4.31	stated	259	6.8	kicker	25	6.04
disallow	190	6.67	direct	53	4.22	score	611	6.75	post	482	5.91

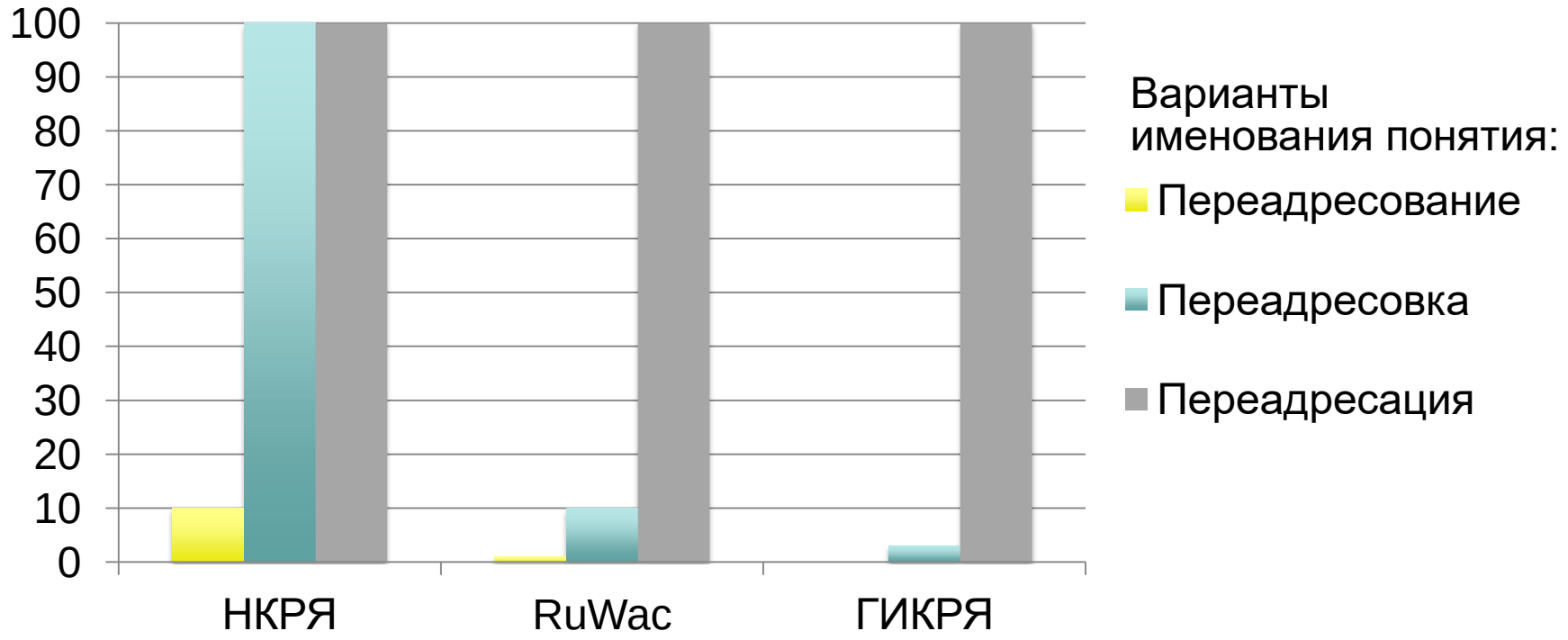
ПРОБЛЕМЫ ТЕКСТОВЫХ КОРПУСОВ



- Создание корпуса – трудоёмкий процесс:
  - выработка критериев отбора текстов
  - сбор тысяч текстов
  - решение проблем с авторскими правами
  - балансировка корпуса (нет единых критериев)
  - приведение текстов к единому формату
  - разработка/адаптация специализированного ПО
  - выбор формата разметки
  - разметка текстов
- Немасштабируемость данных (довольно часто)
- Ограниченность корпуса: проблема полноты покрытия слов языка (лексики),
  разреженность данных (*data sparseness*)

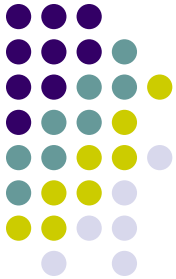
ПРОБЛЕМЫ КОРПУСОВ: НЕМАСШТАБИРУЕМОСТЬ

Получаемые по созданному корпусу данные могут быть не масштабируемы и не универсальны



А по данным словарей основной вариант – *переадресование*

ПРОБЛЕМА ОГРАНИЧЕННОСТИ



Часто корпуса плохо отображают:

- современную, новую лексику

  НКРЯ: *локдаун, майнить* – 0 употреблений ?
- региональные / профессиональные слова *челыш ?*
- редкие слова/явления: *ламбрекены ?*

Для преодоления этих проблем:

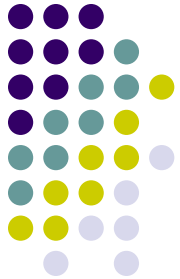
- для первых 5 000 частотных слов, необходимый размер корпуса – 10-20 млн. словоупотреблений
- для первых 20 000 частотных слов – более 100 млн. словоупотреблений

% слов, изменяющихся в языке за сто лет:

язык	% слов
американский	2,66
британский	4,11
русский	5,74

язык	% слов
немецкий	5,42
французский	3,43
испанский	2,97

СКОЛЬКО РАЗНЫХ СЛОВ В КОРПУСЕ? ЗАКОН ХИПСА



- Эмпирический закон, приписываемый лингвисту Х.С. Хипсу (реально:чешский лингвист Г. Хердан,1960г.?)
- С ростом объема текста (корпуса) количество различных слов в нем увеличивается

$$V = K * N^{b}$$

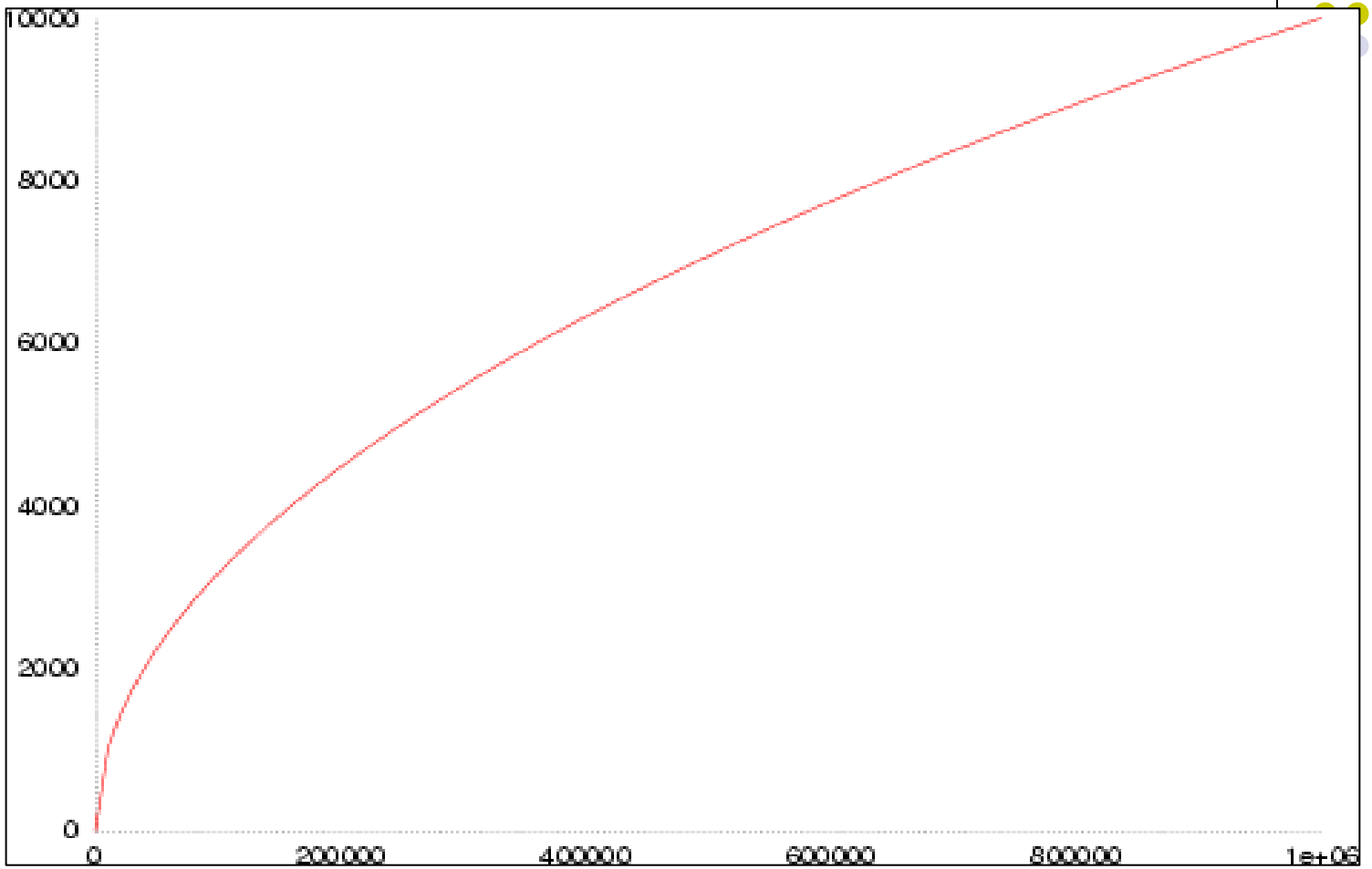
N – количество слов V – количество уникальных слов

β и K – константы, для европейских языков:

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

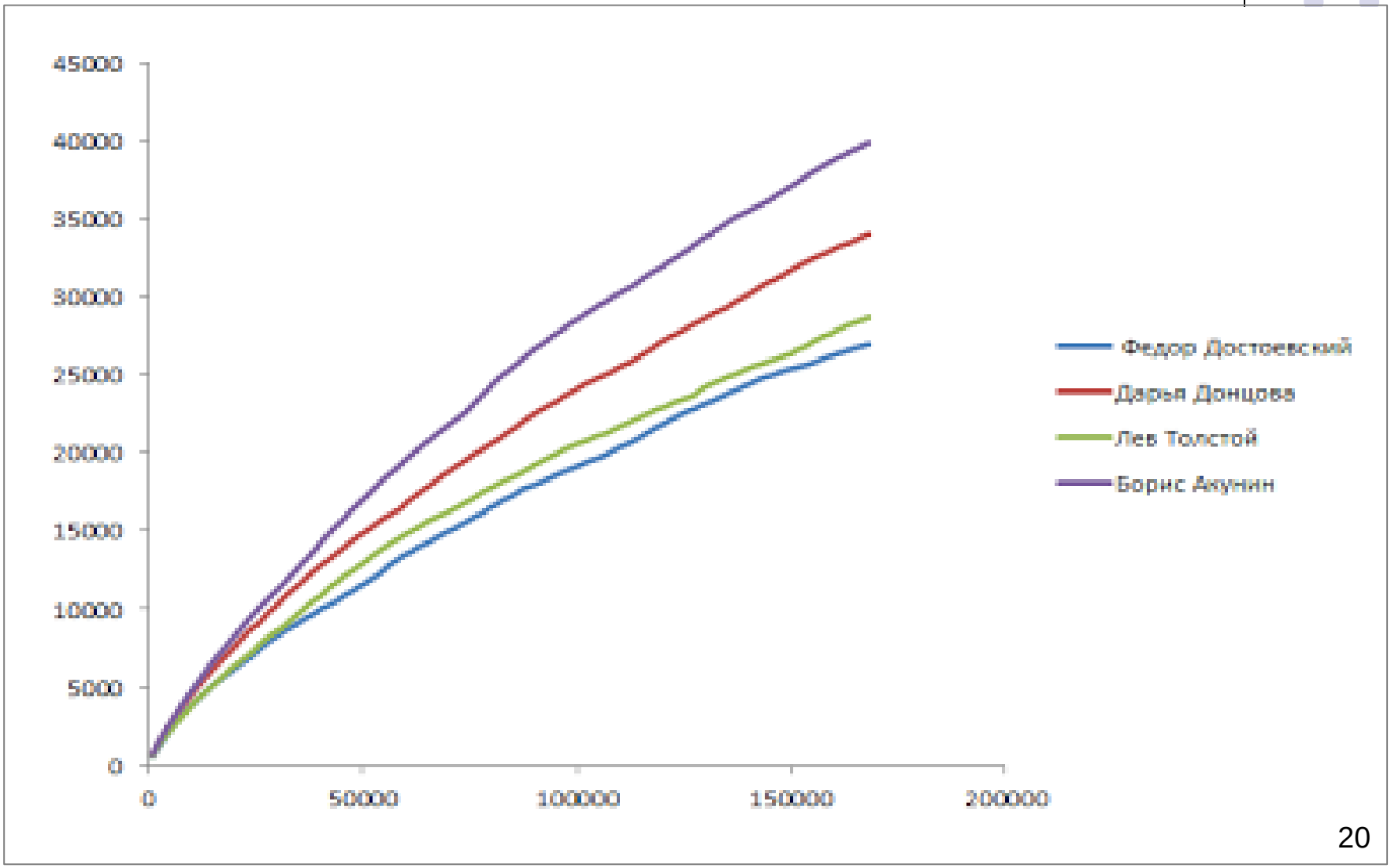
- Быстрый рост лексикона обычно обеспечивается редкими словами (названиями и собств. именами)
- Отклонения от закона при добавлении в корпус нового текста может означать наличие плагиата
- Корпус не может отобразить все богатство лексики

ГРАФИК ЗАКОНА ХИПСА

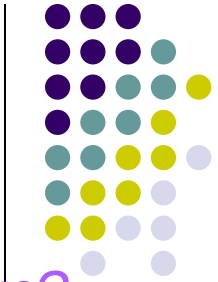


ГРАФИКИ ДЛЯ РАЗНЫХ АВТОРОВ

Как можно их интерпретировать?



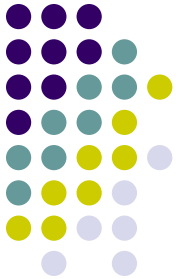
ИНТЕРНЕТ-КОРПУС: ПРОБЛЕМЫ



- Интернет: стихийно создаваемый массив текстов, разметка отсутствует/недостаточна – *это корпус?*
- Корпус м.б. построен автоматически из интернет-текстов (выкачивание страниц с заданными ключевые словами или с заданных адресов), так построены:
- *RuTenTen11* – с помощью SketchEngine + автоматическая морфологическая разметка

  ГИКРЯ – включает блоги, тексты Вконтакте и ЖЖ и т.п. автоматическая морфологическая разметка, доступен подкорпус 2 млн. словоупотреблений
- Проблемы:
  - ✓ Опечатки, обрывки слов, имена собственные, слова из других языков, экспрессивная лексика, новые слова и др. (*слу-шаю-с*, *красавчег*, *щаскакдам*)
  - ✓ Сбалансированность, представительность

ИСПОЛЬЗОВАНИЕ КОРПУСОВ В ЗАДАЧАХ КЛ



- Построение словарей, тезаурусов
- Обучение и тестирование лингвистических процессоров (морфологического, синтаксического и др.)
  - Применяются соответствующие корпуса/подкорпуса (например, *SinTagRus* из НКРЯ)
- Обучение и тестирование прикладных систем КЛ: (классификация текстов, извлечение информации и т.п.)
  - Применяются корпуса и *data sets* (наборы данных) – более широкое и менее конкретное понятие

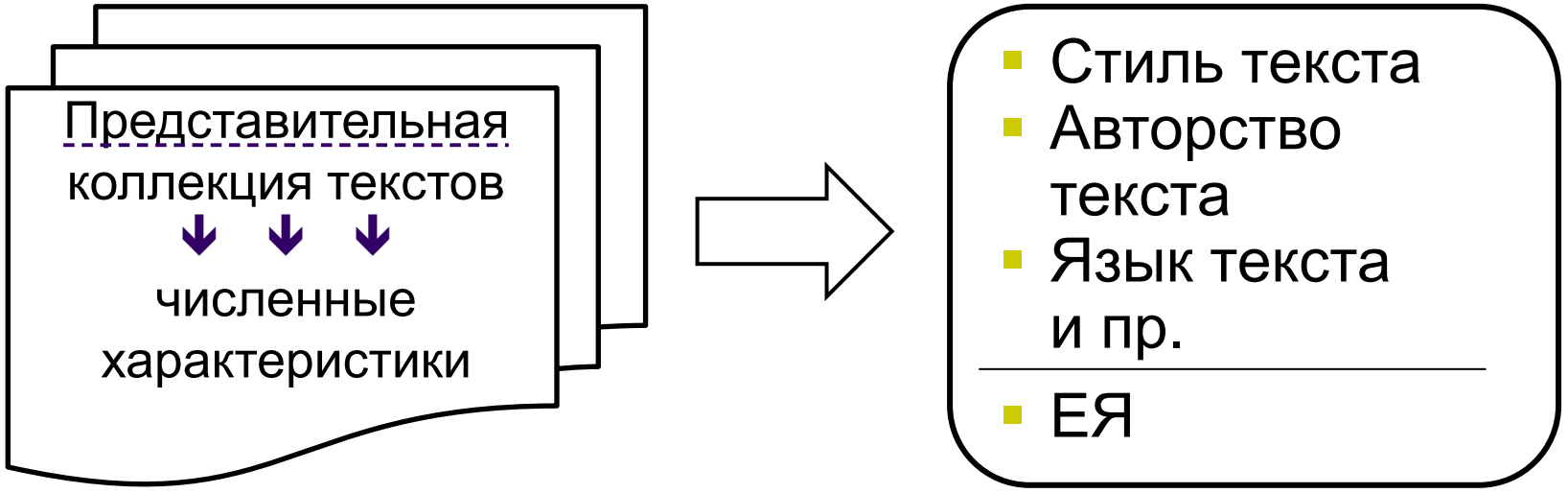
    В чем отличие датасета от корпуса/коллекции?
- Построение *статистических языковых моделей*: *Statistical Language Models*

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА

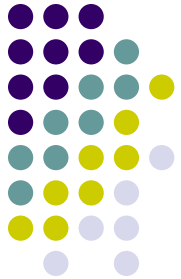
Дисциплина, изучающая количественные закономерности ЕЯ, проявляющиеся в текстах

Опирается на предположение:

что верно для одних текстов, верно для других или даже для ЕЯ в целом



СТАТИСТИЧЕСКИЕ ИССЛЕДОВАНИЯ ТЕКСТОВ



Объекты исследования:

- буквы, морфы, слова, словосочетания и т.д.
- их классы: гласные буквы, части речи
- последовательности: *N-граммы*

Например, для букв *лодка*

при $N=2$ *ло од дк ка* при $N=3$ *лод одк дка*

для текста *течет речка, печет печка*

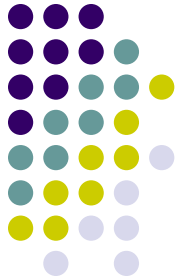
при $N=2$ *течет речка* / *речка печет* /*печет печка*

Численные характеристики:

- частота употребления
- вероятность (определяется статистически)

Вычисляются по представительной коллекции/корпусу

ЧАСТОТЫ УПОТРЕБЛЕНИЯ



- Для исследуемого объекта вычисляется количество его употреблений в наборе текстов T
- Полученные данные фиксируются в частотном списке/словаре

Абсолютная частота Fa_i – число употреблений i-ого объекта в совокупности текстов T (число вхождений в текст)

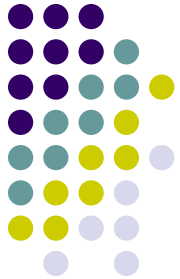
Относительная частота Fr_i: $Fr_i = Fa_i / N$

N – общее количество объектов в T (объекты могут быть разными)

❖ Частоты зависят от длины текста, его стиля/жанра/...

(в технических текстах буква *ф* становится более частой из-за слов *функция*, *коэффициент*, *диффузия* и др.)

СТАТИСТИКА БУКВ: ЧАСТОТЫ БУКВ ЛАТИНИЦЫ



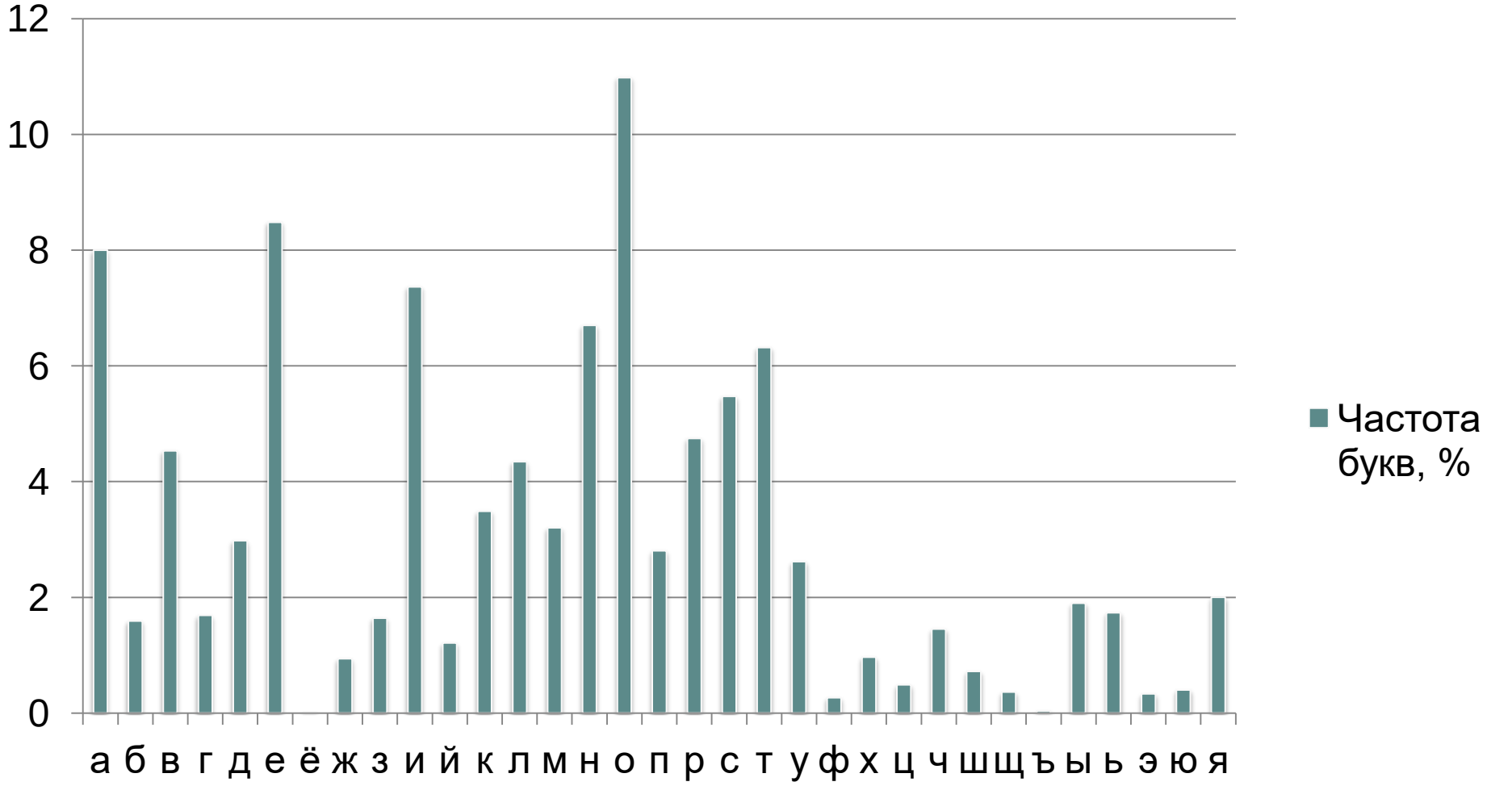
Буква	Франц.	Немец.	Англ.	Испан.	Итал.
A	7,68	5,52	7,96	12,90	11,12
B	0,80	1,56	1,60	1,03	1,07
H	0,64	5,02	5,39	0,91	0,83
J	0,19	0,16	0,16	0,24	0,00
K	0,01	1,13	0,41	0,00	0,00
O	5,34	2,14	6,62	8,84	8,92
U	6,05	4,22	2,48	4,00	3,09
W	0,00	1,38	1,80	0,00	0,00
Z	0,07	1,17	0,05	0,31	1,24

ОТНОСИТЕЛЬНЫЕ ЧАСТОТЫ (%) БУКВ КИРИЛЛИЦЫ

Буква	Русский	Белорусск.	Болгарский	Лакский
а	7,998	14,459	10,254	14,825
ё	0,013	0,711	–	–
ж	0,940	0,879	0,657	0,607
и	7,367	0,023	7,652	6,044
м	3,203	3,456	3,218	3,354
о	10,983	3,148	8,803	1,264
ц	0,486	2,563	0,381	1,572
ы	1,898	3,240	–	0,075
і	–	4,774	–	2,085

ЧАСТОТА БУКВ РЯ

10 наиболее частых букв русского языка – СЕНОВАЛИТР



ЧАСТОТЫ БУКВЕННЫХ БИГРАММ И ТРИГРАММ

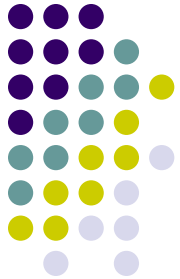
Биграмма	Частота в англ., %
th	2,5
ed	1,5
ti	1,0
ou	0,8

Биграмма	Частота в рус., %
ст	1,7
по	1,2
пр	1,0
он	0,7

Триграмма	№ в списке
the	1
ing	5
nce	11
men	16

Триграмма	Частота в %
что	0,6
ска	0,5
про	0,4
сно	0,3

СТАТИСТИКА БУКВ, *N-ГРАММ*: ПРИМЕНЕНИЕ В КЛ



Учитываются частоты букв и N-грамм букв

- Определение схожести текстов, дубликатов:
  чем больше одинаковых символьных N-грамм, тем более похожи тексты
- Определение кодировки текста:
  при верном определении кодировки количество недопустимых биграмм минимально
- Дешифровка текста (если известен язык):
  наиболее частым буквам текста ставятся в соответствие наиболее частые буквы языка
- Определение языка текста (русский/белорусский?):
  опора на частоты букв и допустимых N-грамм.

ЧАСТОТА УПОТРЕБЛЕНИЯ СЛОВ: ПРИМЕР

Эта страница красного цвета.
Красное солнце. Красное лето.
Красная площадь флаги полощет.

Количество *словоупотреблений* в тексте (вхождений словоформ, токенов) – 12

Количество различных *словоформ* – 11

Количество различных *лексем* (слов) – 9

$Fa_{площадь} = 1$ $Fr_{площадь} = 1 / 12 \approx 0{,}08$

$Fa_{красное} = 2$ $Fr_{красное} = 2 / 11 \approx 0{,}18$

$Fa_{красный} = 4$ $Fr_{красный} = 4 / 9 = 0{,}4(4)$

или: $Fr_{красное} = 2 /12$? $Fr_{красный} = 4 /12$?

СТАТИСТИКА СЛОВ: ЗАКОН ЦИПФА

Французский стенографист Ж.-Б. Эсту в 1916 г.,
Американский лингвист Д. Ципф в 1935 г.
выявили эмпирический закон:

Пусть *ранг* ***i*** – порядковый номер слова в списке, упорядоченном по убыванию частот.

Частота слова обратно пропорциональна ***i***

$$Fa_i = \frac{C}{i} \qquad Fr_i = \frac{C}{i} * N$$

C – константа, для английского языка ≈ 0,1
для русского языка ≈ 0,08 (0,06)

Эта зависимость присутствует и в других областях: распределение населения по городам, доходов среди людей, публикаций среди ученых и т.д.

ГРАФИК ЗАКОНА ЦИПФА

- Небольшое число очень частотных слов
- Среднее число раз используется среднее число слов
- Большинство слов используется редко и очень редко

ПРИМЕРЫ ЧАСТОТ

В Browns corpus (1 014 312 словоформы)

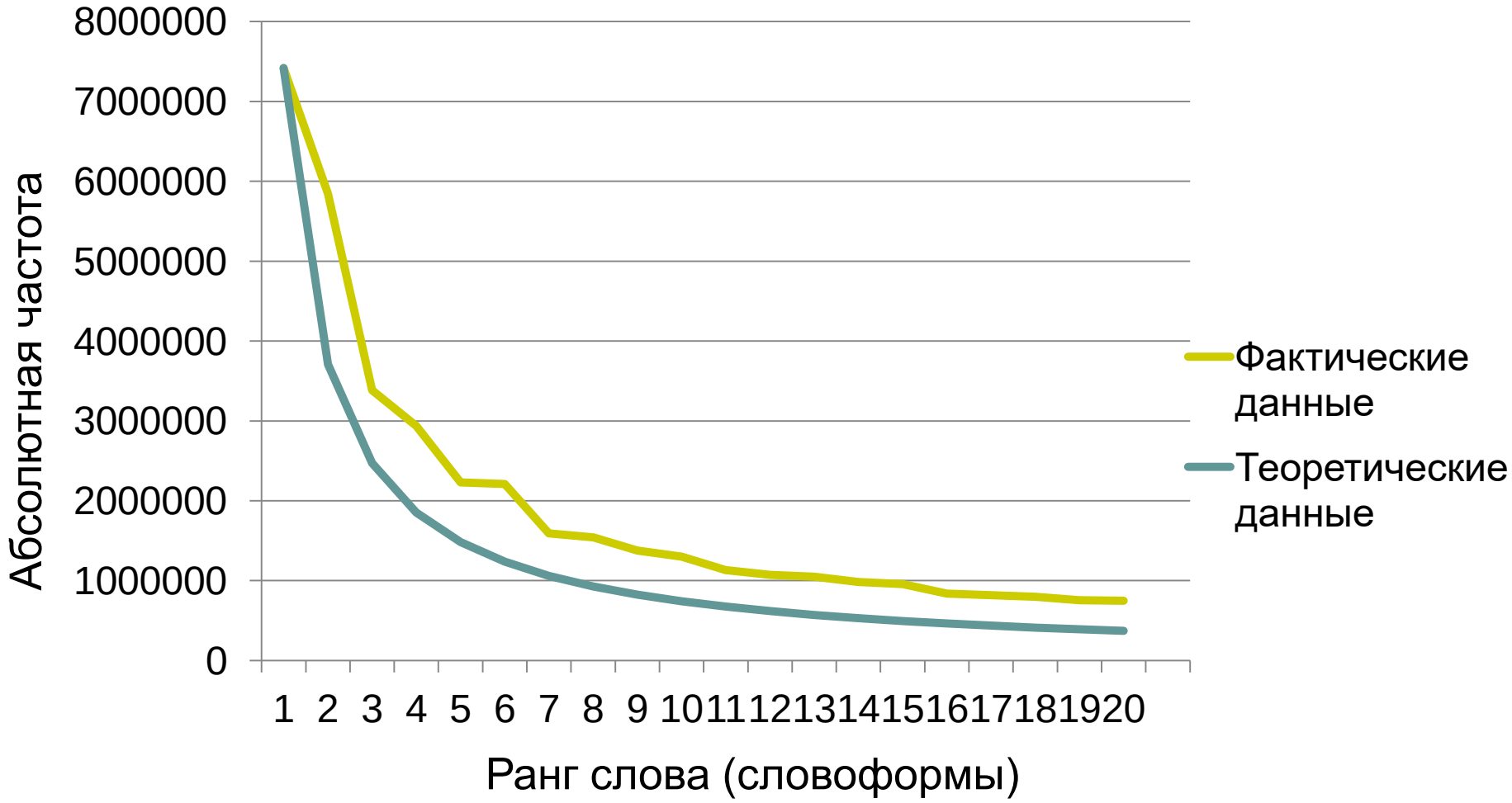
	the	of	and
фактическая Fa_i	69 971	36 411	28 852
фактическая Fr_i	0,069	0,036	0,028
теоретическая Fr_i ($C=0,1$)	0,100	0,050	0,033
теоретическая Fr_i ($C=Fr_1$)	0,069	0,035	0,023

В НКРЯ (192 689 044 словоформы)

	и	в	не
фактическая Fa_i	7 416 716	5 842 670	3 385 161
фактическая Fr_i	0,038	0,030	0,017
теоретическая Fr_i ($C=0,08$)	0,080	0,040	0,027
теоретическая Fr_i ($C=Fr_1$)	0,038	0,019	0,012

ГРАФИК ЗАКОНА ЦИПФА: НКРЯ

Для словоупотреблений закон работает лучше



ЗАКОН ЦИПФА-МАНДЕЛЬБРОТА

При $i < 15$ закон Ципфа выполняется плохо, и американский кибернетик Б. Мандельброт в 50-е гг. предложил ввести поправку ρ:

$$Fa_i = \frac{k * N}{(\rho + i)^{\gamma}}$$

N – общее количество словоупотреблений в тексте

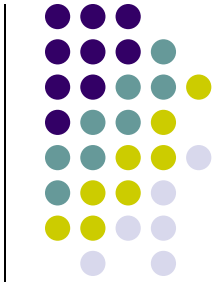
k – константа, зависит от количества слов в частотном словаре

γ – *коэффициент лексического богатства* текста (число разнообразных словоформ, число рангов)

ρ – поправочный коэффициент частых слов

Если $\rho = 0$ и $\gamma = 1$, получится закон Ципфа

КРИТИКА ЗАКОНА ЦИПФА-МАНДЕЛЬБРОТА



- Для текста определенной длины, языка, темы, жанра каждый раз нужно подбирать ρ и γ.
- Постоянство γ не сохраняется для маленьких и больших i
- Закон дает только грубое приближение к истинной статистической структуре текста.
- Закон удовлетворительно выполняется лишь для 2-3 тысяч наиболее частых словоформ.

  Для описания редких словоформ приходится оперировать другими зависимостями.
- Для художественных текстов закон выполняется точнее, чем для научно-технических.
- 1000 самых частых слов покрывают 85% текстов

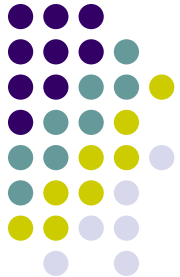
Закон ЦМ – закон не языка, а текстов

СТАТИСТИКА СЛОВ: 20 САМЫХ ЧАСТОТНЫХ

№	английский	русский	№	английский	русский
1	the	и	11	it	к
2	be	в	12	for	по
3	to	не	13	not	но
4	of	на	14	on	его
5	and	с	15	with	это
6	a	что	16	he	из
7	in	я	17	as	все
8	that	а	18	you	у
9	have	он	19	do	за
10	I	как	20	at	от

– Что у них общего ?

ЧАСТОТА И ДЛИНА СЛОВ



Теоретическое обоснование?

- Б. Мандельброт решал задачу:
  Необходимо сократить число букв в сообщении,
  не теряя его смысл.
  Как будут соотноситься частоты слов?
- Решение – Оптимальное кодирование:
  самые частые слова – самые короткие.
- Их частоты описываются законом Ципфа-Мандельброта.
- В процессе развития ЕЯ самооптимизировался:
  в нем в среднем самые частые слова (предлоги, местоимения, союзы, артикли) – самые короткие (древние, простые по структуре, но многозначные).
- Наиболее значимые слова лежат в средней части графика для закона Ципфа.

СРЕДНИЕ ДЛИНЫ СЛОВОФОРМ В ТЕКСТАХ НА ЕЯ

язык	длина, в буквах	язык	длина, в буквах
английский	3,042	немецкий	5,448
чеченский	3,424	латышский	5,786
ирландский	3,673	румынский	6,164
норвежский	4,222	таджикский	6,205
русский	4,701	узбекский	7,471
французский	4,855	эстонский	7,754
вьетнамский	4,979	эскимосский	8,296
украинский	5,156	тувинский	8,776

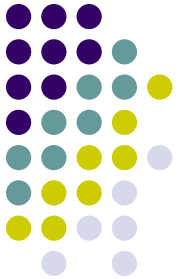
Наиболее развитые/используемые языки имеют небольшую длину словоформ.

РЯ все еще оптимизируется ?

ЧАСТОТЫ И ДЛИНЫ СЛОВ: ПРИМЕНЕНИЕ В КЛ

- Оценка естественности текста: выполнение закона ЦМ (но не для небольших фрагментов текстов)
- Выявление значимых слов: наиболее значимые слова имеют среднюю частоту
- Определение стиля/жанра/темы/автора текста: они характеризуются пропорциями частей речи, длинами слов, сложностью предложений и т.д.
  - В разговорной речи много коротких слов, в научных текстах много длинных/сложных терминов
  - В научных текстах глаголов в 2 раза меньше, а местоимений – в 1,5 раза, но существительных на 20% больше, чем в художественных
  - Определение путем сравнения с эталонными величинами? Проблема: подбор эталонных текстов

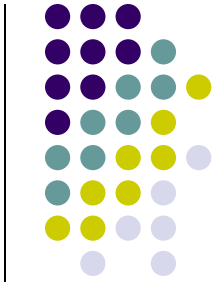
СТАТИСТИЧЕСКАЯ ЯЗЫКОВАЯ МОДЕЛЬ



Statistical Language Model

- Отвечает на вопрос, насколько данная фраза типична/правильна для языка
- Типичность/правильность можно задать с помощью вероятности путем приписывания любой фразе/предложению языка вероятности ее появления в тексте
- Позволяет предсказывать слова текста
- При создании статистической языковой модели:
  - Вычисляют вероятности (на основе статистики по корпусу/коллекции)
  - Устраняют ее недостатки – сглаживают
  - Оценивают качество построенной модели

ВАРИАНТЫ МОДЕЛИ



Вероятности могут быть приписаны:

- каждой фразе/предложению языка
- последовательностям из N слов –

N-граммная модель

- лексико-синтаксическим конструкциям

Пусть в языке ***1000*** слов, а длина предложения не более ***10***. Число возможных предложений:

$$1000^1 + \ldots + 1000^{10} =$$

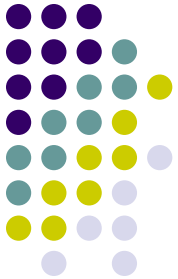
$$= 1001001001001001001001001001000$$

Для первого варианта модели необходимо хранить $\approx 10^{30}$ (2^{100}) чисел, большая часть из которых $= 0$,

т.к. многие предложения либо неверны, либо раньше не встречались.

Поэтому : второй вариант модели

N-ГРАММНАЯ МОДЕЛЬ



N-граммная модель – более компактная:
для *1000* слов: 10^6 биграмм, 10^9 триграмм

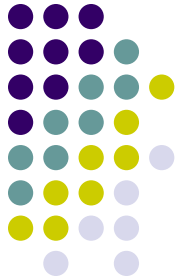
- Рассматривают *N*-граммы – последовательности из N единиц (слов). Обычно $N = 2, 3$ и реже 4
- Предположение (марковская модель): вероятность P появления очередного слова в предложении зависит только от предыдущих $N-1$ слов
- P вычисляют по корпусу с опорой на частоты фраз

$$P(w_N|w_1\ w_2\ \dots\ w_{N-1}) = Fa(w_1\ w_2\ \dots\ w_N) / Fa(w_1\ w_2\ \dots\ w_{N-1})$$

- P предложения вычисляется как произведение P входящих в него *N*-грамм. Например, для $N=2$:

$$P(w_1\ w_2\ \dots\ w_M) \approx P(w_1)*P(w_2|w_1)*P(w_3|w_2)*\dots*P(w_M|w_{M-1})$$

ПРИМЕР: ВЫЧИСЛЕНИЕ ВЕРОЯТНОСТИ ПРЕДЛОЖЕНИЯ



- Вычислим P для предложения *I want to eat* :

$$P(I\ want\ to\ eat) =$$
$$P(I) * P(want|I) * P(to|want) * P(eat|to)$$

- По *Berkeley Restaurant Corpus* получено:

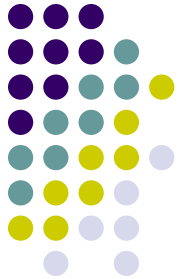
$$P(I) = 0{,}25 \qquad P(want|I) = 0{,}32$$
$$P(to|want) = 0{,}65 \qquad P(eat|to) = 0{,}26$$

где, в частности, $P(want|I) = Fa\ (I\ want) / Fa(I)$

- Тогда

$$P(I\ want\ to\ eat) = 0{,}25*0{,}32*0{,}65*0{,}26 \approx 0{,}014$$

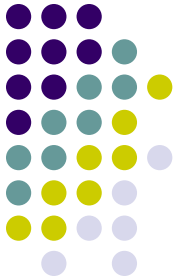
ОСОБЕННОСТИ N-ГРАММНОЙ МОДЕЛИ



+ Возможность построения языковой модели по текстовому корпусу достаточно большого размера
+ Относительная простота использования, скорость
+ Наиболее удачная модель этого класса – триграммная

− Предположение о независимости вероятности слова от более далеких слов (не отражены глубокие связи)
− Большие объемы обучающих и хранимых данных:
Если в языке 1000 слов, то получим
10^6 биграмм, 10^9 триграмм, 10^{12} тетраграмм
− Недостаточность данных для достоверных оценок (ограниченность корпуса: *data sparseness*)
$P=0$ *у неправильной N-граммы, но также*
$P=0$ *у N-граммы, которой нет в корпусе !*

СГЛАЖИВАНИЕ МОДЕЛИ



- Недостаточность данных для построения модели ведет к некорректным вероятностям:

  Пусть в корпусе есть фраза *I want to eat* ,
  но нет фразы *I want to eat British food*

  Получается, что $P(British \mid I\ want\ to\ eat) = 0$ и тогда
  $P(I\ want\ to\ eat\ British\ food) = 0$

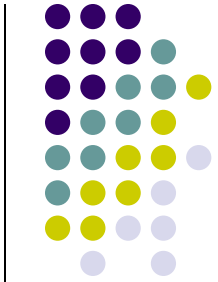
- Для устранения некорректности применяется ***сглаживание***: вероятность понижают для одних N-грамм и повышают для других

- Способ Лапласа: для биграмм – добавление $\alpha \leq 1$
  (если $\alpha=1$ – *"add-one"*)

$$P(w_2 \mid w_1) = \frac{Fa(w_1 w_2) + \alpha}{Fa(w_1) + V}$$

где V – количество слов (словоформ, лемм) в корпусе
а для $N > 2$: V – число возможных $N-1$ грамм

СГЛАЖИВАНИЕ: ПРИМЕР



- Пусть в корпусе нет биграммы *British food* и *eat British* (а также слова *British*) Тогда:

$$P(British|eat) = 0 \quad P(food|British) = 0$$

$$P(I\ want\ to\ eat\ British\ food) = P(I\ want\ to\ eat) *$$
$$* P(British|eat) * P(food|British) \approx 0{,}014 * 0 * 0 = 0$$

- Пусть размер корпуса 14180 словоупотреблений, 1616 слов (лемм) и
- По корпусу вычислено:

$Fa(I\ want) = 1134$ $Fa(I) = 3545$

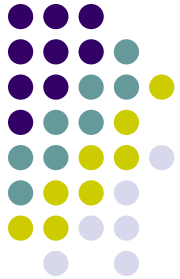
$Fa(want\ to) = 793$ $Fa(want) = 1220$

$Fa(to\ eat) = 942$ $Fa(to) = 3623$

$Fa(eat\ British) = 0$ $Fa(eat) = 3261$

$Fa(British\ food) = 0$ $Fa(British) = 0$

ПРИМЕР СГЛАЖИВАНИЯ: Лаплас



- Возьмем $\alpha = 1$, $V = 1616$. Тогда по формуле:

$$P(British \mid eat) = \frac{Fa(eatBritish) + 1}{Fa(eat) + V} = \frac{0 + 1}{3261 + 1616} \approx 0{,}0002$$

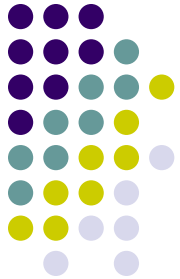
$$P(food \mid British) = \frac{0 + 1}{0 + 1616} \approx 0{,}0006$$

- При пересчете $P(I\ want\ to\ eat)$ получим $\approx 0{,}003$
- Теперь вычислим P для предложения

I want to eat British food

$$P(I\ want\ to\ eat\ British\ food) = P(I\ want\ to\ eat) *$$
$$* P(British|eat) * P(food|British) \approx$$
$$\approx 0{,}003 * 0{,}0002 * 0{,}0006 \approx 0{,}00000000036$$

УЧЕТ ПРЕДЛОЖЕНИЙ ПРИ ПОСТРОЕНИИ ЯЗЫКОВОЙ МОДЕЛИ



- В биграммной (и других моделях) вероятность первого слова рассчитывается не так, как остальных:

$$P(\text{Это лучший подарок}) = P(\text{это}) * P(\text{лучший}|\text{это}) * P(\text{подарок}|\text{лучший})$$

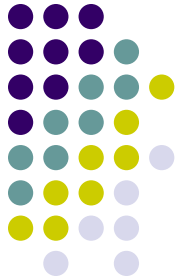
- Обычно текст разбит на предложения, и какие-то слова встречаются в начале предложений, а какие-то нет
- Логично это учитывать и при построении модели применяя специальное слово *BEGIN* или тег *< S >*

$$P(\text{Это лучший подарок}) = P(\text{это}|\ BEGIN) * P(\text{лучший}|\text{это}) * P(\text{подарок}|\text{лучший})$$

- Возможно также введение “слова” *END* в конец предложений : для согласованности модели и чтобы неправильные предложения имели меньшую вероятность:

He saw the. и *He saw the red house.*

ОЦЕНКА МОДЕЛИ: ПЕРПЛЕКСИЯ



- Пусть по корпусу построена N-граммная языковая модель. Как оценить ее качество?
- Предлагается оценить вероятность появления некоторого текста в рамках построенной модели: модель тем лучше, чем лучше она предсказывает случайный текст
- Используется *перплексия* (*perplexity*) – величина, обратная средней вероятности появления слов в тестируемом тексте T:

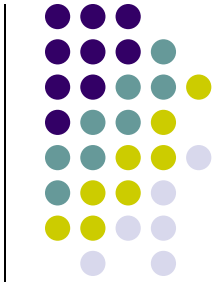
$$PP(T) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i \mid w_1 \ldots w_{i-1})}}$$

где N – объем (число слов) тестируемого текста

- Чем меньше показатель перплексии, тем лучше языковая модель. Важно:

чтобы в тексте T не было предложений из исходного корпуса

ПЕРПЛЕКСИЯ: ИНТЕРПРЕТАЦИЯ, ПРИМЕР



- Перплексию построенной языковой модели также можно трактовать как среднее количество слов, которые могут следовать за некоторым словом
- Таким образом, чем больше перплексия, тем выше неоднозначность, и следовательно, хуже модель
- Перплексия показывает предсказательную силу модели

Пример: Объем корпуса *Wall Street Journal* для построения моделей – 38 млн. словоупотреблений (V =19 979 слов),

Объем текста T – 1,5 млн. словоупотреблений

	униграммы	биграммы	триграммы
PP	*962*	*170*	*109*

ЯЗЫКОВЫЕ МОДЕЛИ: РАЗВИТИЕ

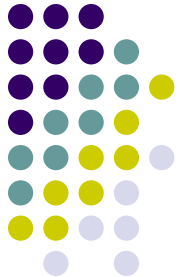
- Статистические языковые модели – модели языка в целом, а не текста (хотя построены по текстам)
- Кроме *Add-One Smoothing* применяются разные способы сглаживания: *Good-Turing Estimation* и др.
- Построена большая статистич. языковая модель: *GoogleNrams*: *https://books.google.com/ngrams*
- Разновидности статистических моделей: учет статистики синтаксических связей слов – *Syntactic Language Model*

❖ Другой подход/способ учета статистики: модель получают на основе обучения предсказанию (соседних) слов

❖ Машинное обучение на нейронных сетях – *Neural Language Model* , *предсказательные* модели

❖ Разные способы обучения дают разные модели, наиболее известные: *Word2Vec* и *BERT* – языковая модель для 104 разных ЕЯ, сложная архитектура сети

ЯЗЫКОВАЯ МОДЕЛЬ *Word2Vec*



Одна из первых нейронных моделей (Google, 2013)

- обучение на базе очень большой текстовой коллекции
- нейронная двухслойная сеть прямого распространения
- специальный алгоритм обучения на основе статистики совместной встречаемости слов
- после обучения скрытый слой сети образует пространство векторов, каждому уникальному слову соответствует свой числовой вектор – *word embedding*, отражает семантику

Пример: предсказание для слова *кофе*:

зернах 0.757 растворимый 0.709 чая 0.709 коффе 0.704 mellanrost 0.694 сублемированный 0.694 молотый 0.690 кофейные 0.680 чай 0.679 декофеинизированный 0.678 капучино 0.677 monoarabica 0.676 свежесваренный 0.676 декаф 0.674 расствоpимый 0.659

ЯЗЫКОВЫЕ МОДЕЛИ: ПРИМЕНЕНИЕ В ЗАДАЧАХ КЛ

- Распознавание речи: выбор правильного варианта из двух слов, произносимых одинаково

  P(у меня грипп) > P(у меня гриб)

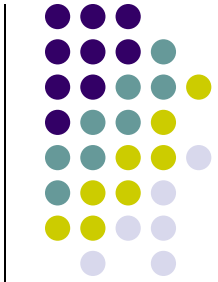
- Выявление ошибок:
  - опечатки: *P(на странице 25) > P(на страннице 25)*
  - лексические ошибки (сочетаемость):

    P(small house) > P(small home)

- Машинный перевод: выбор правильного слова при переводе: *a herd of horses* – *табун* или *стадо лошадей*
- Генерация текста: составление рефератов, аннотаций и др. *Яндекс-рефераты:* https://yandex.ru/referats/

Как сгенерировать случайное предложение в N-граммной модели?

ЗАКЛЮЧЕНИЕ



- Коллекции и корпуса текстов, датасеты – неотъемлемые компоненты решения задач КЛ
- Для разных задач нужны разные корпуса, и их много en.wikipedia.org/wiki/List_of_text_corpora
- Статистика используется в КЛ повсеместно, но разными способами,
  важно: статистические данные существенно зависят от текстов, на которых они получены
- Часто полезно учитывать эмпирические закономерности (закон Ципфа и др.) и следствия из них
- Статистическая языковая модель – исторически первая из ряда языковых моделей, но до сих пор используется

СПАСИБО ЗА ВНИМАНИЕ!

ДОМАШНЕЕ ЗАДАНИЕ № 1

Тема: морфология, статистика, На выбор 5 вариантов:

A. Статистическое исследование разных видов морфологической омонимии в текстах РЯ
B. Программирование и тестирование собственного графематического анализатора для РЯ
C. Сравнение возможностей двух морфопроцессоров для РЯ, и оценить точность их морфологического анализа.
D. Лексико-статистический анализ двух текстов на русском языке с применением морфоанализатора
E. Анализ качества разрешения морфологической омонимии для одного из морфологических процессоров

Срок выполнения – до **21 февраля** включительно

Предполагаемый объем отчета: 1-5 страниц

Оценивается общий объем работы, полнота тестирования

ПОДКОРПУС НКРЯ
СинТагРус (SynTagRus)

Год создания	с 1998 года
Язык	русский
Размер	более 1,1 млн слов (около 77 тыс. предложений)
Назначение	лингвистические исследования, источник лингвистических данных
Вид данных	тексты разных авторов и разных жанров
Разметка	полная морфологическая и синтаксическая разметка со снятой омонимией