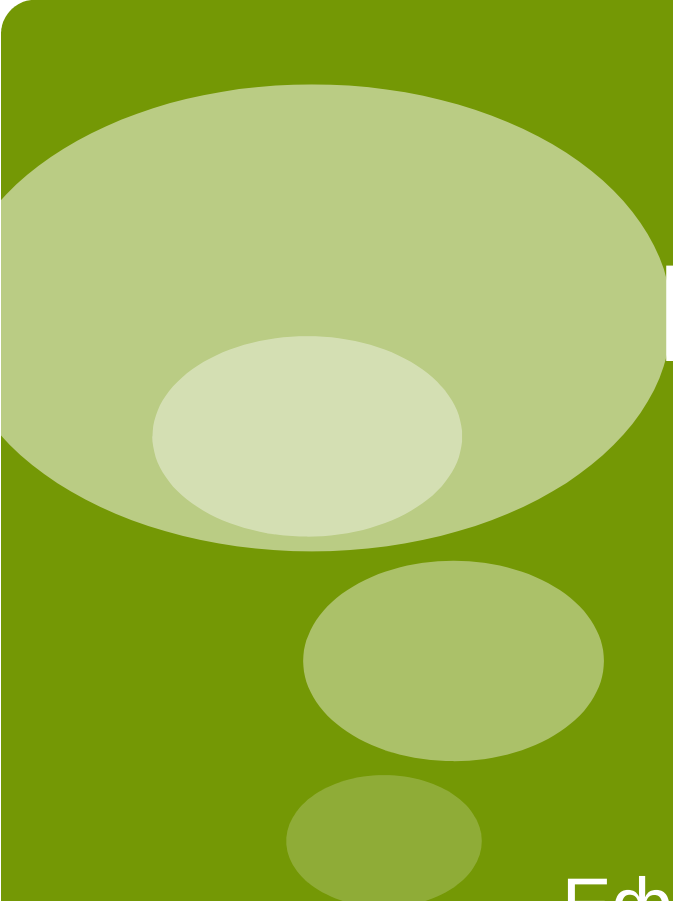




# Квантитативная ЛИНГВИСТИКА

Ефремова Наталья Эрнестовна  
Грацианова Татьяна Юрьевна

# Содержание



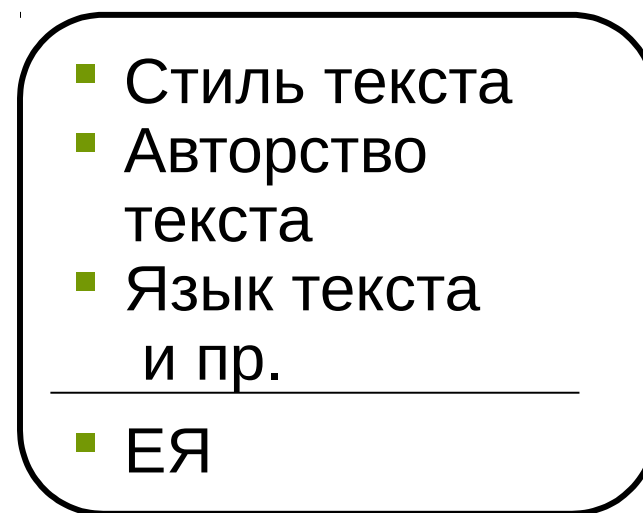
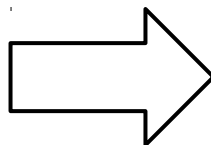
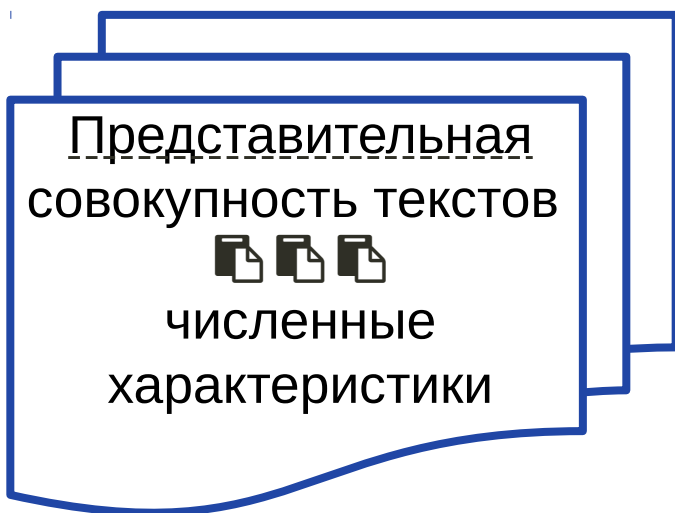
- Квантитативная лингвистика
- Статистические характеристики текстов и их применение:
  - статистика букв
  - статистика слов
- Закон Ципфа-Мандельброта и его интерпретация
- Применение статистики слов в КЛ

# 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**)

Для букв слова *успех*

при  $N=2$  *ус сп не ех*      при  $N=3$  *усп спе пех*

## Численные характеристики:

- частота употребления
- вероятность

Работаем, как правило, не со случайными, а со специально собранными текстами – с корпусом

# Частота употребления



*Абсолютная частота  $Fa_i$  – число употреблений  $i$ -ого объекта в совокупности текстов  $T$*

*Относительная частота  $Fr_i$ :  $Fr_i = Fa_i / N$ , где  $N$  – общее количество объектов в  $T$*

- ❑ Для исследуемого объекта подсчитывается число его употреблений в текстах  $T$
- ❑ Полученные данные фиксируются в частотном списке/словаре
- ❑ В разных источниках – разные значения частот, т.к. они зависят от длины текста, его стиля/жанра/...

В технических текстах буква **Ф** становится частой из-за слов **функция**, **коэффициент**, **диффузия** и пр.

# Частоты букв латиницы и кириллицы



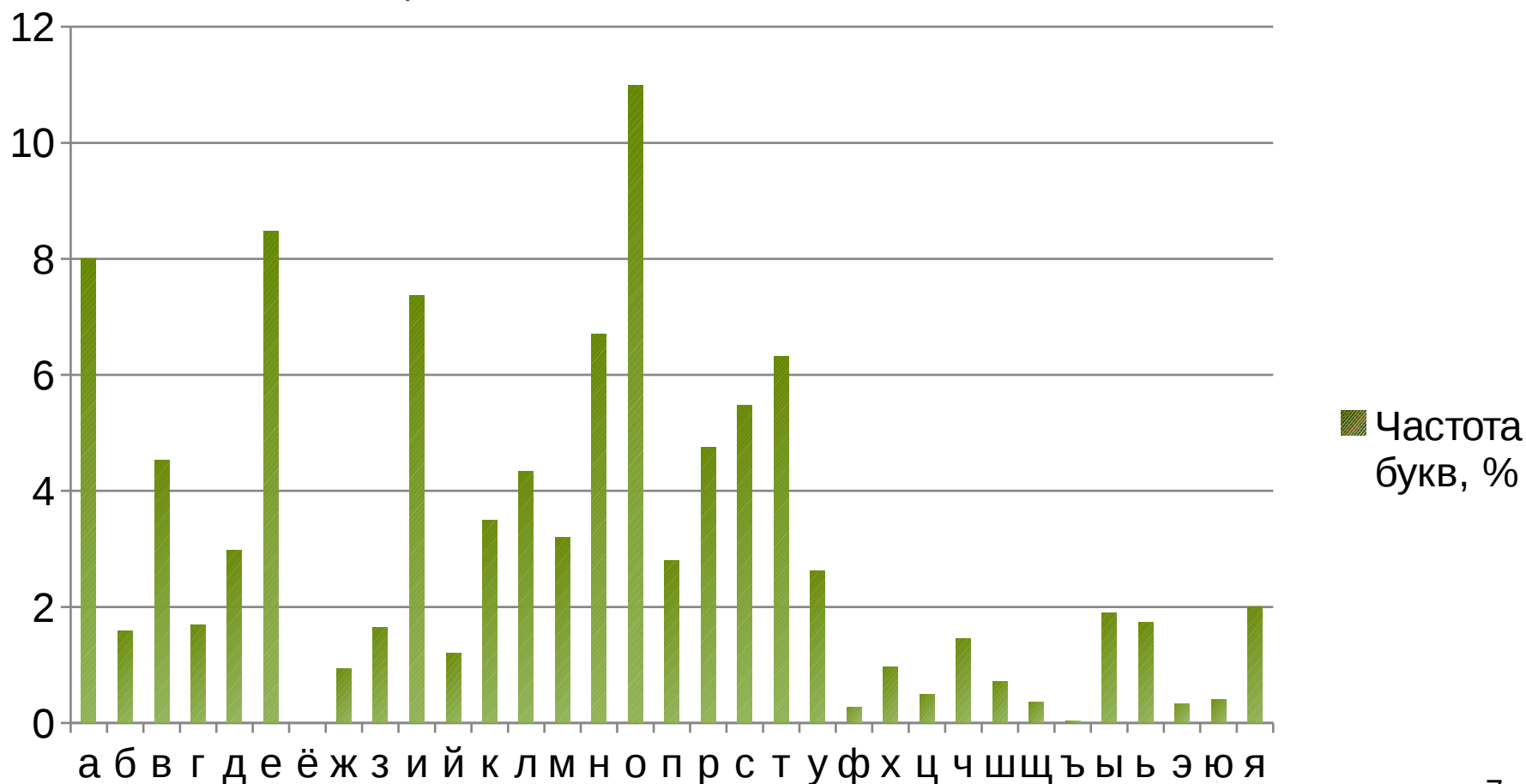
| Буква | Франц.      | Немец.      | Англ.       | Испан.       | Итал.       |
|-------|-------------|-------------|-------------|--------------|-------------|
| А     | 7,68        | <b>5,52</b> | 7,96        | <b>12,90</b> | 11,12       |
| Н     | <b>0,64</b> | <b>5,02</b> | 5,39        | 0,91         | 0,83        |
| Ј     | <b>0,19</b> | <b>0,16</b> | <b>0,16</b> | <b>0,24</b>  | <b>0,00</b> |
| К     | 0,01        | 1,13        | 0,41        | 0,00         | 0,00        |
| О     | 5,34        | 2,14        | 6,62        | 8,84         | 8,92        |

| Буква | Русский      | Белорусск.    | Болгарский   | Лакский      |
|-------|--------------|---------------|--------------|--------------|
| а     | <b>7,998</b> | <b>14,459</b> | 10,254       | 14,825       |
| ё     | 0,013        | 0,711         | —            | —            |
| и     | <b>7,367</b> | <b>0,023</b>  | 7,652        | 6,044        |
| м     | <b>3,203</b> | <b>3,456</b>  | <b>3,218</b> | <b>3,354</b> |
| о     | 10,983       | 3,148         | 8,803        | 1,264        |

# Частота букв РЯ



10 наиболее частых букв русского языка –  
СЕНОВАЛИТР, английского – TETRIS-HONDA



# Частоты гласных и согласных букв



| Язык        | Количество гласных, % |
|-------------|-----------------------|
| Испанский   | 47.95%                |
| Итальянский | 46.80%                |
| Французский | 44.27%                |
| Русский     | 42.21%                |
| Немецкий    | 39.27%                |
| Английский  | 39.21%                |

Частота биграмм из гласных и согласных для РЯ:

|           | Гласная | Согласная | Всего  |
|-----------|---------|-----------|--------|
| Гласная   | 6 588   | 38 310    | 44 898 |
| Согласная | 38 296  | 16 806    | 55 102 |



# Частоты буквенных биграмм и триграмм



| Биграмма  | Частота в АЯ, % | Биграмма  | Частота в РЯ, % |
|-----------|-----------------|-----------|-----------------|
| th        | 2,5             | ст        | 1,7             |
| ed        | 1,5             | по        | 1,2             |
| ti        | 1,0             | пр        | 1,0             |
| ou        | 0,8             | он        | 0,7             |
| Триграмма | № в списке      | Триграмма | Частота в РЯ, % |
| the       | 1               | что       | 0,6             |
| ing       | 5               | ска       | 0,5             |
| nce       | 11              | про       | 0,4             |
| men       | 16              | сно       | 0,3             |

# Применение статистики букв в КЛ



- ❑ **Определение языка текста:** опора на частоты букв и допустимых N-грамм, т.к. разные ЕЯ имеют разные частоты для букв
- ❑ **Дешифровка текста** (если известен язык): наиболее частым буквам текста ставятся в соответствие наиболее частые буквы языка
- ❑ **Определение кодировки текста:** при верном определении кодировки количество недопустимых биграмм минимально
- ❑ **Определение схожести текстов, дубликатов:** чем больше одинаковых N-грамм, тем более похожи тексты

# Частота употребления слов: пример (1)



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

Словоупотребление – единица текста: *Красное, солнце, Красное, лето, Красная, ...*

Словоформа – единица частотного списка: *Красное, солнце, лето, Красная, ...*

Слово – единица словаря: *красный, солнце, лето, ...*

Лемма – начальная форма слова

# Частота употребления слов: пример (2)



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

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

Количество различных **словоформ** – 11

Количество различных слов – 9

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

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

$$Fa_{\text{красный}} = 4 \quad Fr_{\text{красный}} = 4 / 9 \approx 0,44$$

# Статистика слов: закон Ципфа



Немецкий физик Ф. Ауэрбах в 1913 г.,  
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**  $i$

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

$C$  – константа, для английского языка  $\approx 0,1$

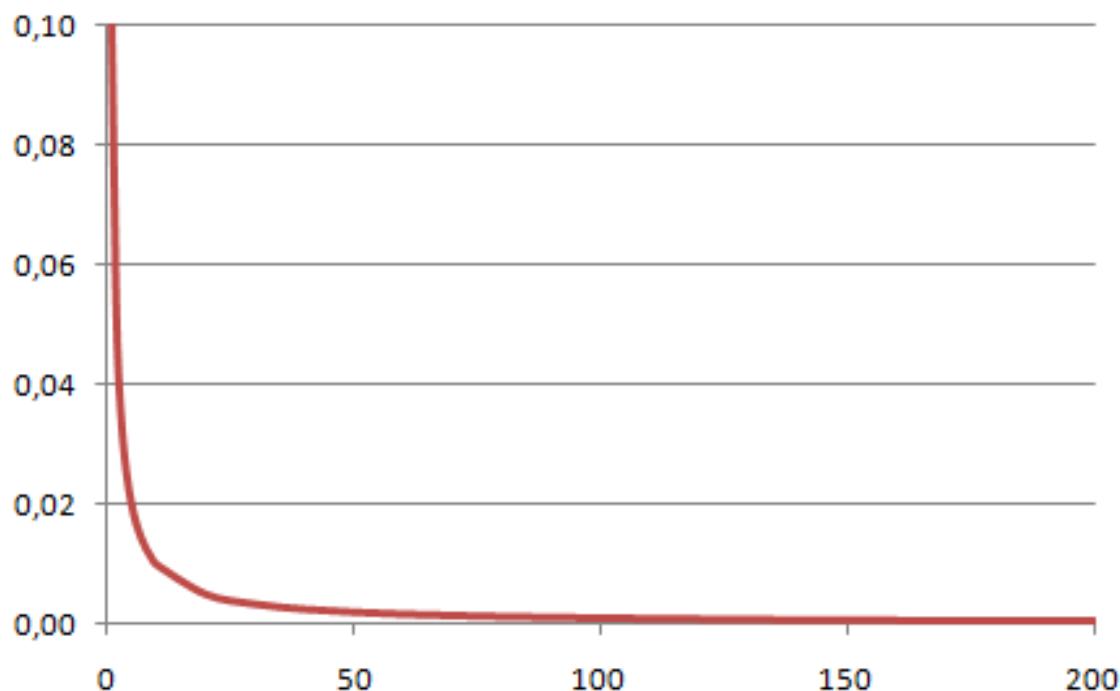
Закон аналогичен принципу Парето (1906 г.):

80% земель в Италии принадлежит 20% ее жителей

# График закон Ципфа



- Чаще всего используется малое количество слов
- Среднее число раз используется среднее число слов
- Большинство слов используется редко и очень редко



# Примеры частот



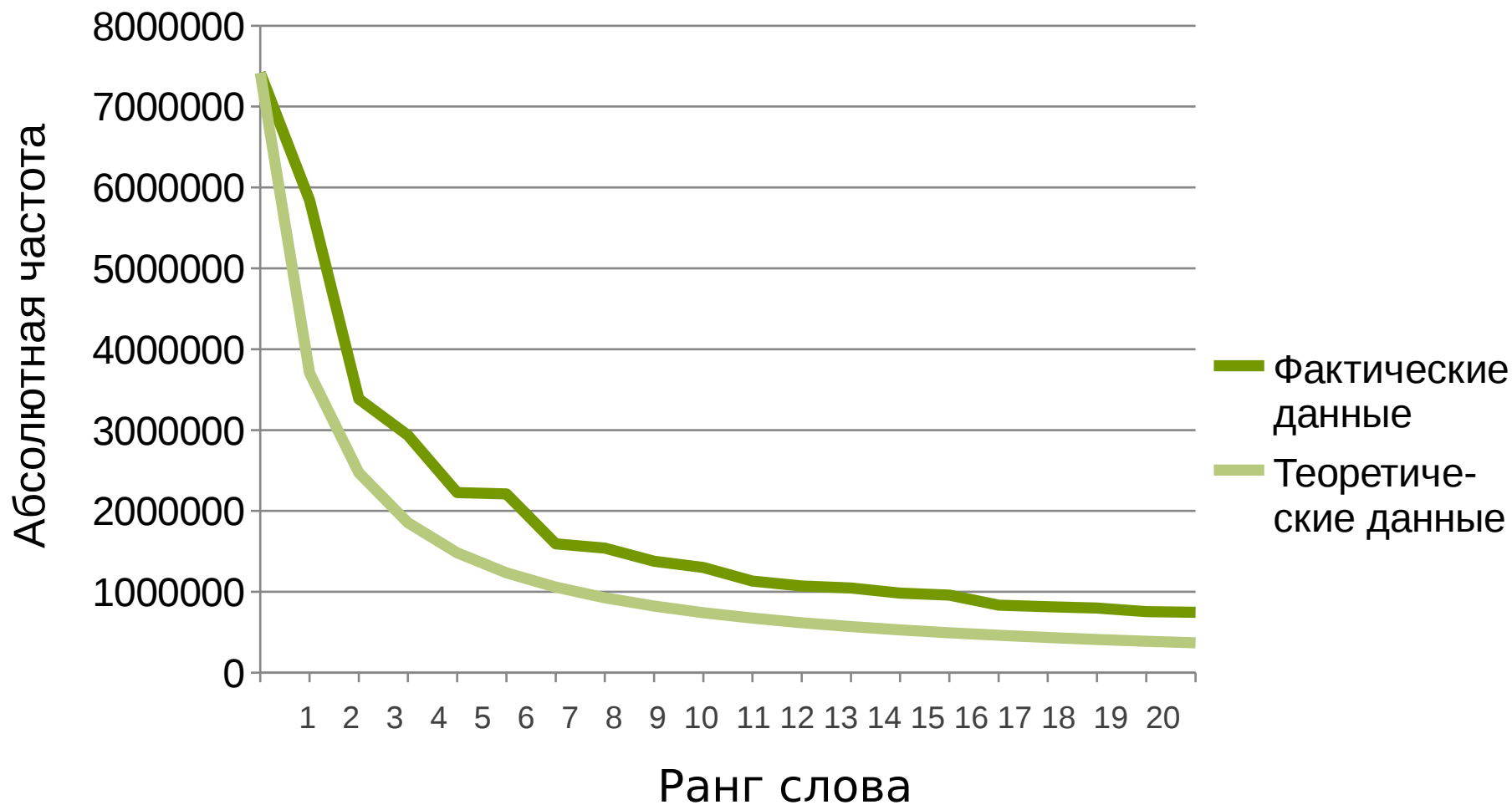
В Browns corpus (1 014 312 словоформы)

|                                 | 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 для НКРЯ





# Эксперимент Кондона



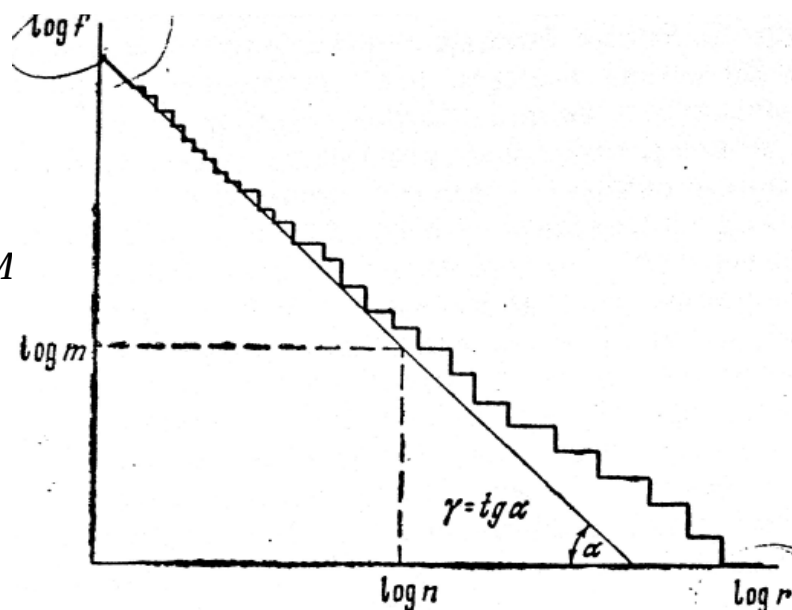
В 1928 году американец Э. Кондон использовал логарифмическую систему координат:

- на оси  $OX$  – логарифм  $i$
- на оси  $OY$  – логарифм  $Fa_i$

Получилось, что:

- пересечение с  $OY$  –  $\log Fa_1$
- пересечение с  $OX$  –  $\log Fa_M$   
( $M$  – количество слов в частотном списке)

⚙ Зависимость между  $\log Fa_i$  и  $\log i$  **близка к линейной**

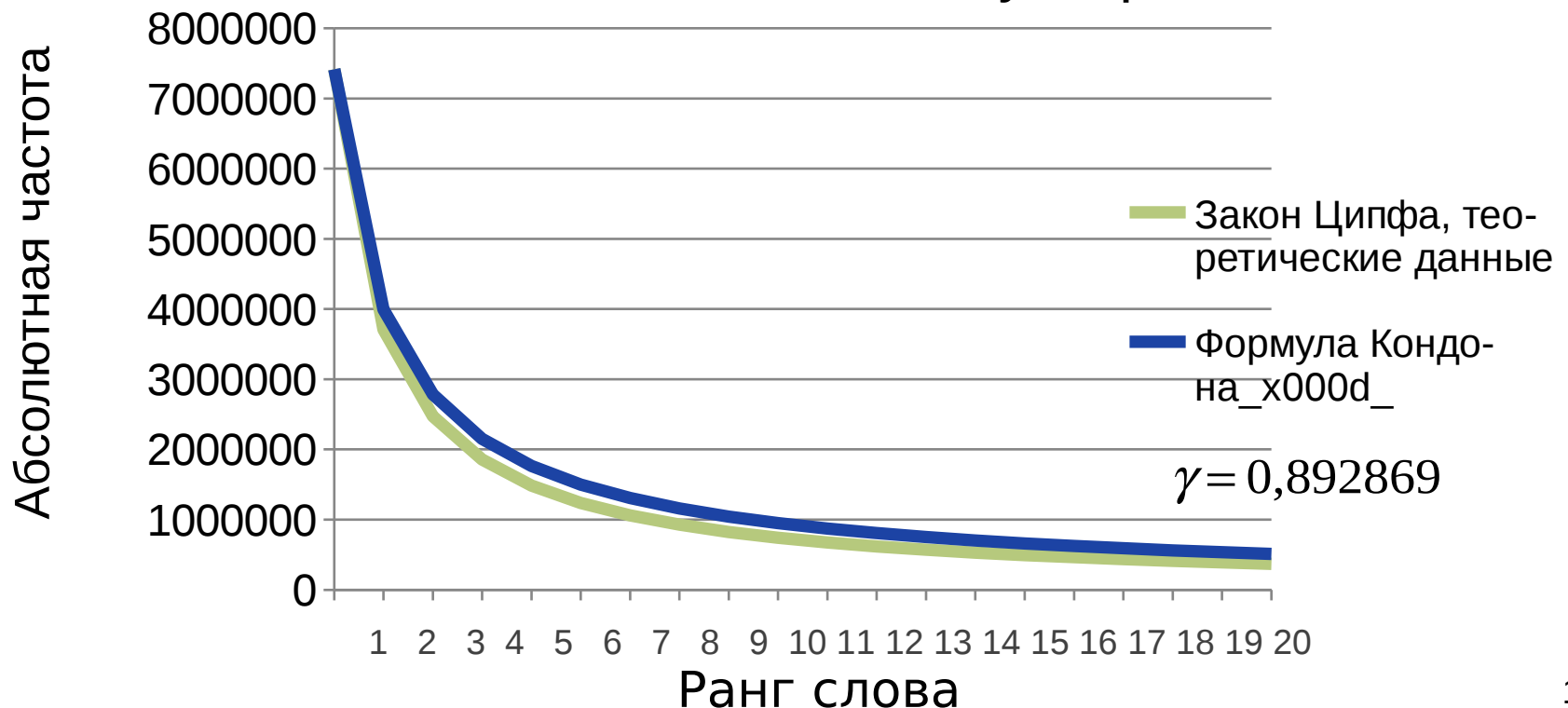


# Формула Кондона



$$Fr_i = \frac{C}{N * i^\gamma}$$

$C$  – частота слова с рангом 1 ( $Fa_1$ )  
 $\gamma$  – константа, задает угол наклона  
(зависит от языка/стиля/жанра и пр.)  
 $N$  – число словоупотреблений в  $T$



# Исследование Мандельброта



- Все понимали, что закон Ципфа работает, но не понимали, почему
- Американский кибернетик Б. Мандельброт в 50-е гг. решал **задачу**:  
Надо 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

# 20 самых частотных слов



Самые частотные слова, как правило, являются наиболее краткими, древними, многозначными и простыми (по морфоструктуре)

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

# Закон Ципфа-Мандельброта



Т.к. при  $i < 15$  закон Ципфа выполняется плохо, Мандельброт предложил ввести поправку  $\rho$ :

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

$N$  – общее количество словоупотреблений в  $T$

$k$  – константа, зависит от количества слов  
в частотном словаре

$\gamma$  – константа, указывающая на угол наклона

$\rho$  – поправочный коэффициент частых слов

Если положить  $\rho = 0$  получится формула Кондона

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

# Применение и критика закона Ципфа-Мандельброта



- Закон не выполняется для фрагментов текстов  выполнение закона является признаком естественности текста
- Для художественных текстов закон выполняется точнее, чем для научно-технических
- Для текста определенной длины, стиля, языка и т.д. каждый раз нужно подбирать  $\rho$  и  $\gamma$
- Постоянство  $\gamma$  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

# Применение статистики слов в КЛ



- ◆ Исследование текстов – стилеметрия
- ◆ Исследование ЕЯ:
  - ✓ глоттохронология
  - ✓ составление частотных словарей,  
квантитативная морфология
  - ✓ оценка развития лексики языка

*% слов. изменяющихся в языке за сто лет*

| язык         | % слов | язык        | % слов |
|--------------|--------|-------------|--------|
| американский | 2,66   | немецкий    | 5,42   |
| британский   | 4,11   | французский | 3,43   |
| русский      | 5,74   | испанский   | 2,97   |

**Для получения эталонных частот  
необходимо создание корпуса текстов**

# Стилеметрия



- ◆ Количественное исследование стилей/жанров/тем/авторства текстов
- ◆ Стилль/жанр/тема/автор характеризуются пропорциями частей речи, длинами слов, сложностью предложений и т.д.

*В научных текстах гл. в 2 раза <, мест. в 1,5 раз <, сущ. на 20% > чем в художественных*

*В разговорной речи много коротких слов (мест.), в научных текстах много терминов  средняя длина слов возрастает*

- ◆ Распределение частот анализируемого текста сравнивается с эталонным распределением



# Частоты частей речи в РЯ (1)



«Частотный словарь русского языка»  
под редакцией Л. Н. Засориной, 1977 год

| Часть речи      | Кол-во,<br>% | Часть речи   | Кол-во,<br>% |
|-----------------|--------------|--------------|--------------|
| Существительное | 24,9%        | Наречие      | 7,9%         |
| Глагол          | 15,9%        | Союз         | 7,6%         |
| Местоимение     | 14,1%        | Числительное | 1,1%         |
| Предлоги        | 11,5%        | Частица      | 1%           |
| Прилагательное  | 8,3%         | Причастие    | 0,4%         |

# Частоты частей речи в РЯ (2)

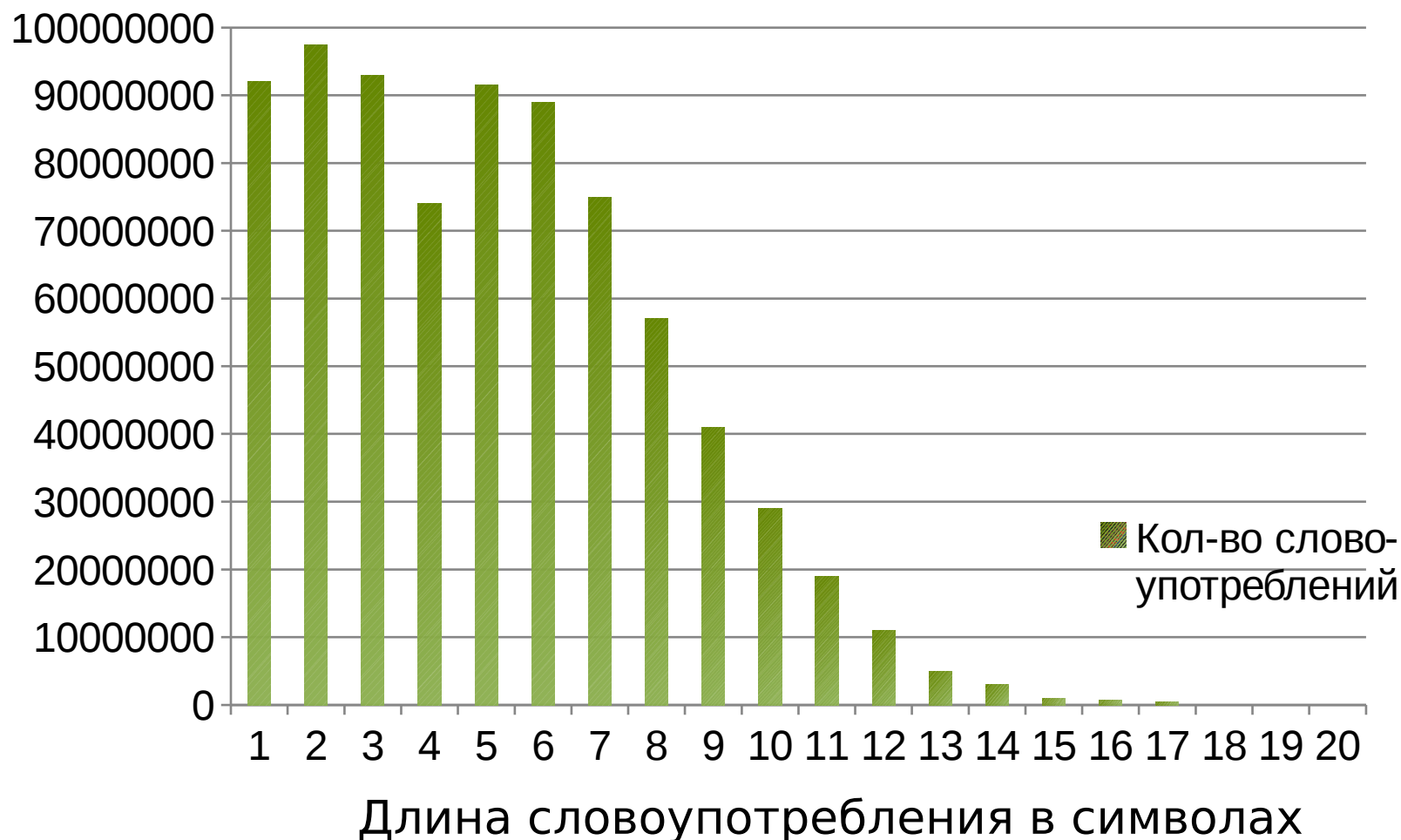


«Новый частотный словарь русской лексики»

С.А. Шаров, О.Н. Ляшевская, 2009 год

|    | Часть речи                                   | Помета                | М: Абс. Частота | % М   |
|----|----------------------------------------------|-----------------------|-----------------|-------|
| 1  | существительное                              | s                     | 1431979         | 28.9% |
| 2  | глагол                                       | v                     | 826995          | 16.7% |
| 3  | предлог                                      | pr                    | 515112          | 10.4% |
| 4  | прилагательное                               | a                     | 426549          | 8.6%  |
| 5  | союз                                         | conj                  | 383623          | 7.7%  |
| 6  | местоимение-существительное                  | spro                  | 388792          | 7.9%  |
| 7  | наречие, предикатив, вводное слово           | adv, praedic, parenth | 253234          | 5.1%  |
| 8  | местоимение-прилагательное                   | apro                  | 223267          | 4.5%  |
| 9  | частица                                      | part                  | 229557          | 4.6%  |
| 10 | местоименное наречие, местоимение-предикатив | advpro, praedicpro    | 112111          | 2.3%  |
| 11 | числительное                                 | num                   | 97193           | 2.0%  |
| 12 | числительное-прилагательное                  | anum                  | 20261           | 0.4%  |
| 13 | инициал                                      | init                  | 10682           | 0.2%  |
| 14 | междометие                                   | intj                  | 6466            | 0.1%  |
| 15 | прочее                                       | nonlex, com           | 26248           | 0.5%  |

# Распределение длин словоупотреблений



# Словоупотребления и словоформы



*На<sup>1</sup> дворе<sup>2</sup> – трава<sup>3</sup>, на<sup>4</sup> траве<sup>5</sup> – дрова<sup>6</sup>.  
Не<sup>7</sup> руби<sup>8</sup> дрова<sup>9</sup> на<sup>10</sup> траве<sup>11</sup> двора<sup>12</sup>.*

Всего словоупотреблений – 12

Словоупотреблений длины 2 – 4 (0,33%)

Словоупотреблений длины 4 – 1 (0,08%)

Словоупотреблений длины 5 – 7 (0,58%)

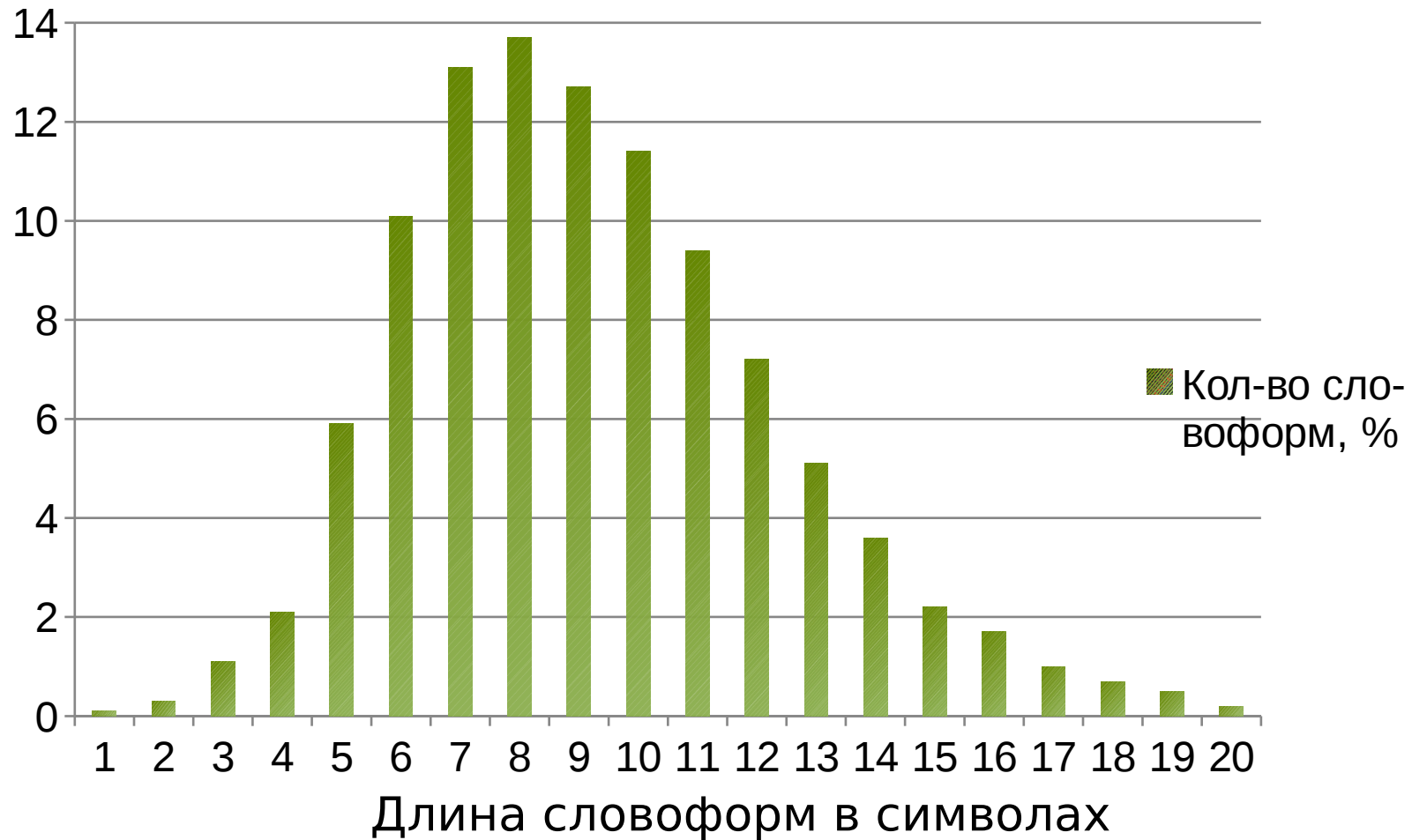
Всего различных словоформ – 8

Различных словоформ длины 2 – 2 (0,25%)

Различных словоформ длины 4 – 1 (0,13%)

Различных словоформ длины 5 – 5 (0,63%)

# Распределение длин словоформ



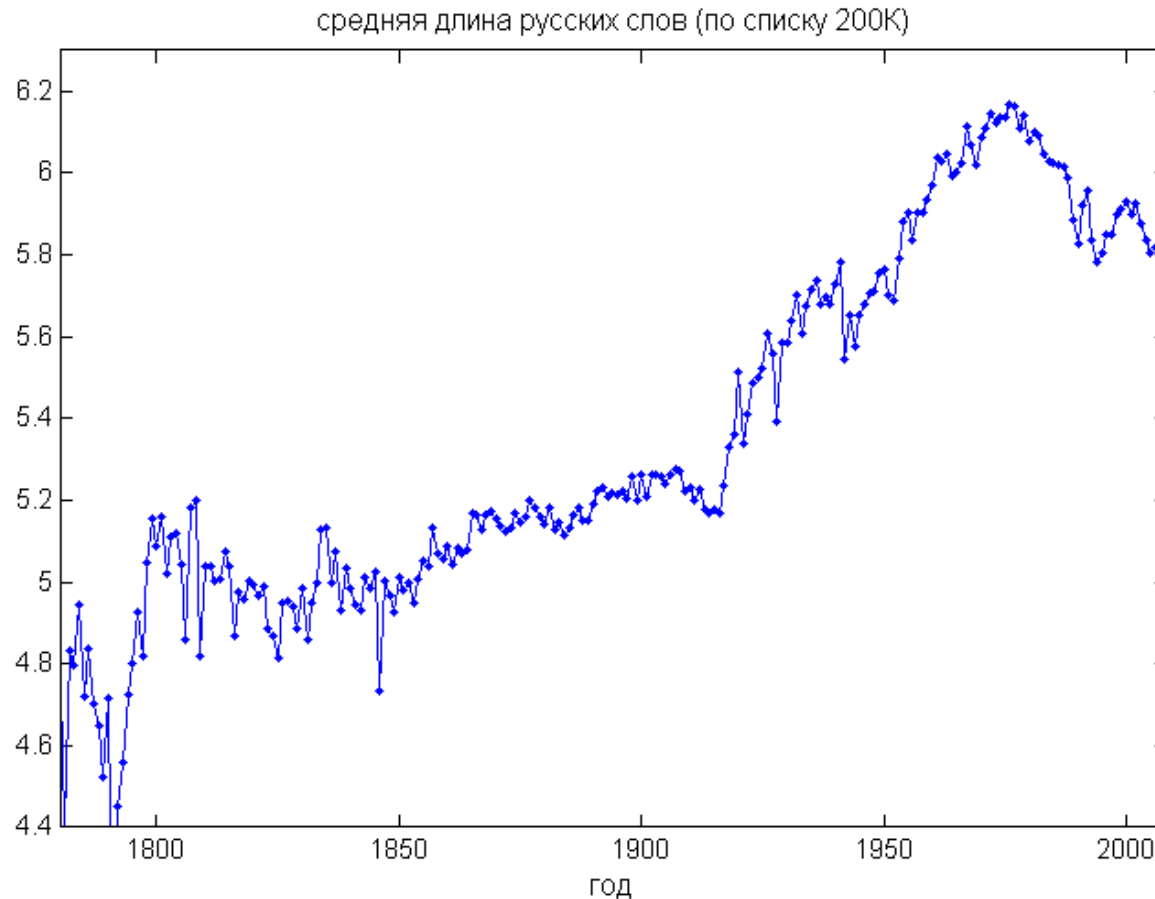
Наиболее значимые слова имеют среднюю частоту

# Средние длины словоформ в текстах



| язык        | длина, в<br>буквах | язык        | длина, в<br>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              |

# Изменение со времени средней длины слов



Аналогично для АЯ   
средняя длина зависит от исторических событий

# Прикладные задачи стилеметрии



- ◆ Классификации текстов, определения жанров, авторства
- ◆ Определение схожести текстов (в том числе и определение плагиата)
- ◆ Предсказание популярности новых книг:
  - ✓ учитываем слова, словосочетания, части речи, сложность конструкций, тональность
  - ✓ используются методы машинного обучения
  - ✓ точность – 74%
  - ✓ слова из популярных книг: *not*, *said*, *whom*, *I*, *in*  
слова из непопулярных книг: *arm*, *risk*, *boat*, *took*

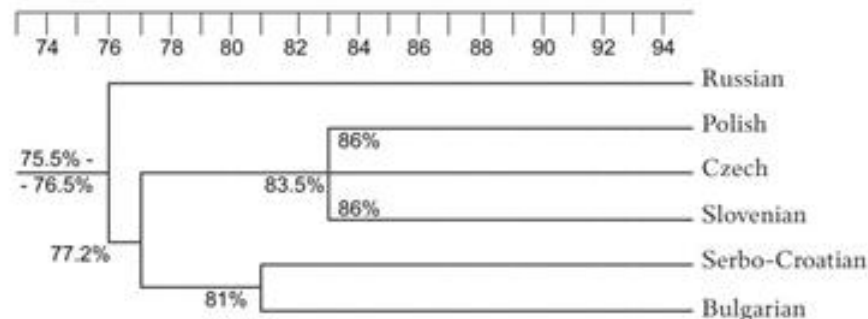
!? можно написать диплом, приятный для чтения ?!



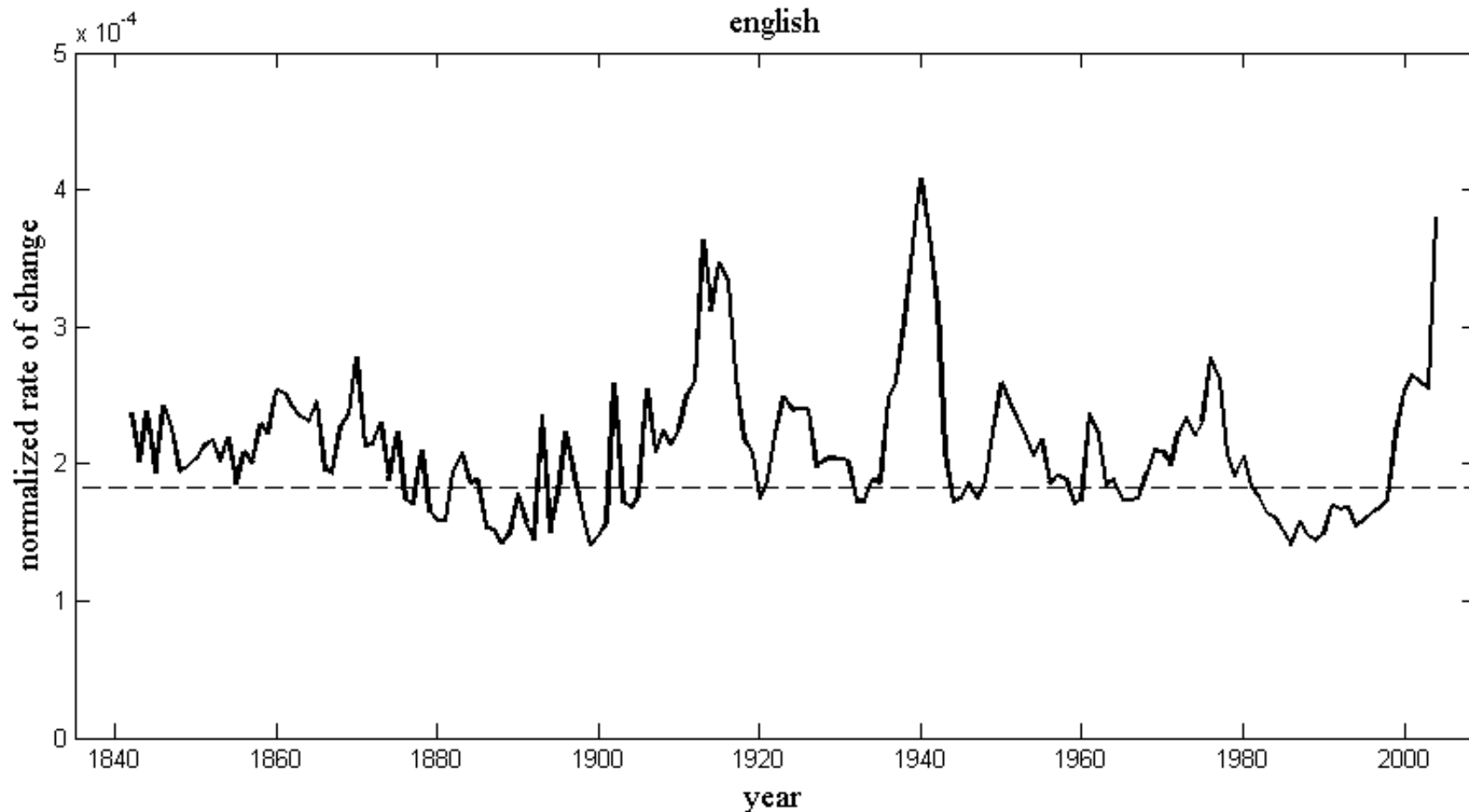
# Глоттохронология



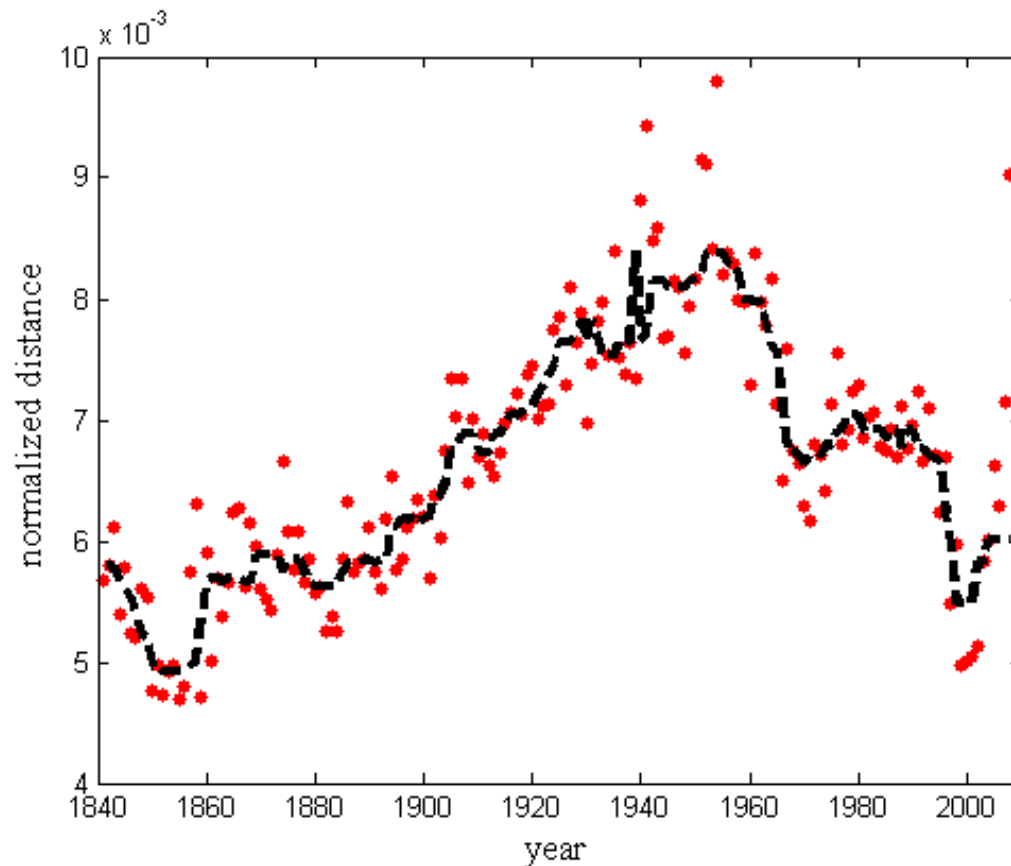
- ◆ Метод определения времени разделения родственных языков. Основан на гипотезе, что **скорость изменения базового словаря языка остается примерно одинаковой**
- ◆ Базовый словарь – список слов предположительно устойчивых к заимствованию (*личные местоимения, части тела, основные глаголы, числительные*)
- ◆ Для базовых словарей родственных языков вычисляется процент родственных слов
- ◆ Чем больше процент, тем позже разделились языки



# Динамика скорости изменения лексики английского языка



# Расстояние между частотными распределениями слов



Британский и американский диалекты АЯ



Спасибо за внимание!

# Частотные словари



- ❑ Чтобы понимать 80-90% текста достаточно знать около 1000 слов (закон Ципфа-Мандельброта)
- ❑ Из частотных словарей можно узнать, какие слова нужно знать в первую очередь
- ❑ Частотность слов зависит от специфики текстов
- ❑ Использование  $F_a$  и  $F_r$  затрудняет сравнение частот из разных словарей
- ❑ Поэтому используется общая частота ( $imp$ ) – *instances per million words*
- ❑ Применение частотных словарей:
  - ✓ дешифровка текста, определение стиля писателя
  - ✓ машинный перевод, распознавание речи, снятие омонимии

# Новый частотный словарь русской лексики



## I. Общая лексика

- Алфавитный список лемм
- Частотный список лемм
- Распределение лемм по функциональным стилям:
  - Частотный словарь художественной литературы  
Словарь значимой лексики художественной литературы
  - Частотный словарь публицистики  
Словарь значимой газетно-новостной лексики
  - Частотный словарь другой нехудожественной литературы  
Частотный словарь значимой другой нехудожественной литературы
  - Частотный словарь живой устной речи  
Словарь значимой лексики живой устной речи
- Алфавитный список словоформ

## II. Общая лексика: части речи

- Частотный список имен существительных
- Частотный список глаголов
- Частотный список имен прилагательных
- Частотный список наречий и предикативов
- Частотный список местоимений (местоимения-существительные, прилагательные, наречия, предикативы)
- Частотный список числительных
- Частотный список лемм служебных частей речи

## III. Вспомогательные таблицы

- Данные о частотности частеречных классов
- Частотность букв русского алфавита
- Частотность двубуквенных сочетаний

## IV. Имена собственные и аббревиатуры

- Алфавитный список собственных имен и аббревиатур

# Фрагменты частотного списка



|       | Лемма            | Лемма            | Часть речи | Частота (ipm) | Худ.лит.1950-60-е | 1970-80-е | 1990-2000-е | Публ. 1950-60-е | 1970-80-е | 1990-2000-е |
|-------|------------------|------------------|------------|---------------|-------------------|-----------|-------------|-----------------|-----------|-------------|
| 1     | и                | и                | conj       | 35801.8       | 37546.0           | 39187.7   | 36514.7     | 36703.6         | 38352.1   | 35539.5     |
| 2     | в                | в                | pr         | 31374.2       | 23943.5           | 25005.7   | 26185.7     | 35295.3         | 33264.3   | 36372.5     |
| 3     | не               | не               | part       | 18028.0       | 21154.3           | 22921.5   | 22349.1     | 16127.4         | 18796.9   | 16853.2     |
| 4     | на               | на               | pr         | 15867.3       | 17169.9           | 16411.7   | 16663.0     | 15249.8         | 15961.9   | 16675.2     |
| 5     | я                | я                | spro       | 12684.4       | 17022.7           | 17090.5   | 18257.0     | 13881.5         | 16386.7   | 10447.1     |
| 6     | быть             | быть             | v          | 12160.7       | 12467.9           | 13492.8   | 12778.1     | 14458.2         | 15091.8   | 12646.0     |
| 7     | он               | он               | spro       | 11791.1       | 19708.0           | 18587.1   | 16882.0     | 11721.6         | 12799.3   | 8811.7      |
| 8     | с                | с                | pr         | 11311.9       | 11186.1           | 10974.4   | 11655.3     | 11060.5         | 11662.5   | 11462.7     |
| 9     | что              | что              | conj       | 8354.0        | 8155.2            | 8313.9    | 8938.8      | 7634.5          | 9702.4    | 8743.1      |
| 10    | а                | а                | conj       | 8198.0        | 10298.0           | 10772.2   | 9897.5      | 5748.9          | 7643.3    | 7298.8      |
| 19995 | пытливый         | пытливый         | a          | 2.6           | 3.7               | 3.8       | 2.2         | 7.4             | 2.2       | 3.1         |
| 19996 | свастика         | свастика         | s          | 2.6           | 6.4               | 2.0       | 1.7         | 5.9             | 2.6       | 3.6         |
| 19997 | смертоносный     | смертоносный     | a          | 2.6           | 3.0               | 1.7       | 2.6         | 4.5             | 1.8       | 3.2         |
| 19998 | спекулятивный    | спекулятивный    | a          | 2.6           | 0.0               | 1.4       | 0.7         | 1.5             | 1.1       | 4.2         |
| 19999 | счастливец       | счастливец       | s          | 2.6           | 3.5               | 3.8       | 3.1         | 8.9             | 4.8       | 2.6         |
| 20000 | тюлень           | тюлень           | s          | 2.6           | 3.5               | 4.0       | 4.6         | 20.8            | 1.1       | 1.3         |
| 20001 | умножать         | умножать         | v          | 2.6           | 1.6               | 1.8       | 2.3         | 5.9             | 4.4       | 2.6         |
| 20002 | ходуном          | ходуном          | adv        | 2.6           | 2.7               | 6.3       | 5.0         | 5.9             | 2.9       | 1.3         |
| 20003 | хохолок          | хохолок          | s          | 2.6           | 28.9              | 1.9       | 1.2         | 0.0             | 0.7       | 0.0         |
| 20004 | целеустремленный | целеустремленный | a          | 2.6           | 1.9               | 2.7       | 1.8         | 5.9             | 4.0       | 3.3         |

# Значимая лексика живой устной речи



|    | Лемма    | Часть речи | Частота в корпусе (ipm) | Частота в подкорпусе (ipm) |
|----|----------|------------|-------------------------|----------------------------|
| 1  | ну       | part       | 1114.6                  | 17208.0                    |
| 2  | да       | part       | 787.5                   | 11847.0                    |
| 3  | вот      | part       | 1785.1                  | 15698.6                    |
| 4  | там      | advpro     | 1128.1                  | 10531.7                    |
| 5  | ты       | spro       | 3171.2                  | 13503.8                    |
| 6  | угу      | intj       | 24.6                    | 2068.8                     |
| 7  | я        | spro       | 12684.4                 | 33686.4                    |
| 8  | нет      | part       | 589.2                   | 4618.9                     |
| 9  | а        | conj       | 8198.0                  | 20593.3                    |
| 10 | вообще   | adv        | 417.6                   | 2989.8                     |
| 11 | у        | pr         | 4306.1                  | 11807.9                    |
| 12 | знать    | v          | 1713.8                  | 6269.4                     |
| 13 | говорить | v          | 1755.0                  | 6275.7                     |
| 14 | ой       | intj       | 64.5                    | 1234.7                     |
| 15 | э        | intj       | 19.4                    | 815.2                      |
| 16 | э-э      | intj       | 11.5                    | 711.9                      |
| 17 | ага      | intj       | 40.2                    | 975.2                      |
| 18 | да       | conj       | 801.0                   | 3445.9                     |
| 19 | давай    | part       | 100.3                   | 1179.3                     |
| 20 | ладно    | part       | 110.3                   | 1202.0                     |