

Лекция 4

Как интерпретировать частотные данные

Частотные списки, коллокации, ключевые слова

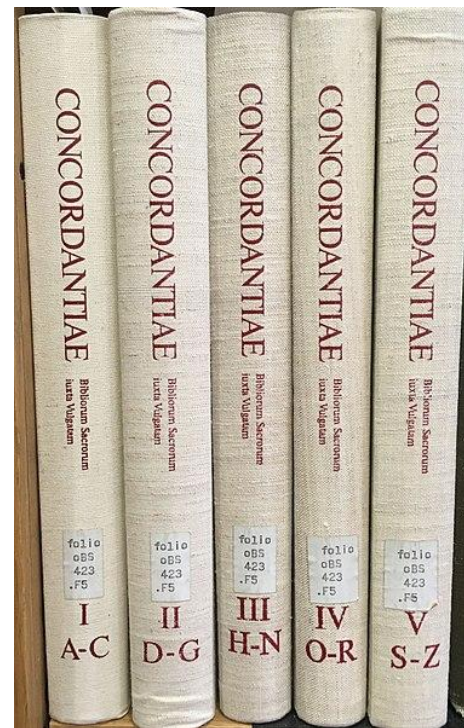
Ольга Ляшевская ** olesar@yandex.ru

Курс “Лингвистические данные”, 1 курс ФикЛ(б) НИУ ВШЭ

Корпус - не только примеры

Что мы можем узнать из корпуса текстов:

- в каких контекстах используется слово
- **лексикографический портрет** слова - самые характерные для него формы, словосочетания и конструкции
- как используются **словосочетания и конструкции**
- как два (под)корпуса различаются по частотному распределению слов, словосочетаний и конструкций



Частотные списки

- Составляются для
 - всего корпуса
 - подкорпусов отдельных авторов, жанров, периодов и т. п.
 - для зоны заголовков, рифмовки в поэзии и т. п.
- Языковые единицы характеризоваться абсолютной и относительной частотой, а также рангом

1 и	12851	99 лицо	275	999 можете	27	9999 вытянул	2
2 не	6474	100 сказать	275	1000 мои	27	10000 вытянуть	2
3 что	6070	101 этот	272	1001 Москвы	27	10001 выучить	2
4 в	5689	102 вас	271	1002 несомненно	27	10002 выучиться	2
5 он	5526	103 Левина	271	1003 новым	27	10003 выходившей	2
6 на	3584	104 раз	271	1004 ног	27	10004 выходу	2
...		

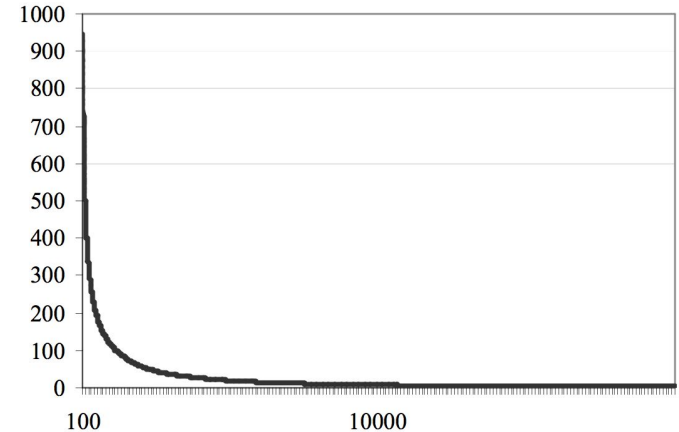


Закон Ципфа

Если все слова упорядочить по убыванию частоты, то частота n -ного слова окажется примерно обратно пропорциональна его рангу (порядковому номеру).

Например, второе по частоте слово встречается примерно в два раза реже, чем первое, третьи три раза реже, чем первое, и т. п.

$$\text{freq}(w) * \text{rank}(w)^\gamma = \text{Const}$$



Кстати, на материале больших веб-корпусов этот закон выполняется примерно для половины слов. Для морфологически богатых языков (ср. также словоформы - леммы) скорость убывает иначе. γ - поправка Бенуа Мандельброта (1965) к закону Джорджа Кингсли Ципфа (1949). Он выделил голову (стоп-слова), среднюю часть и хвост (гапаксы) - broken power law.



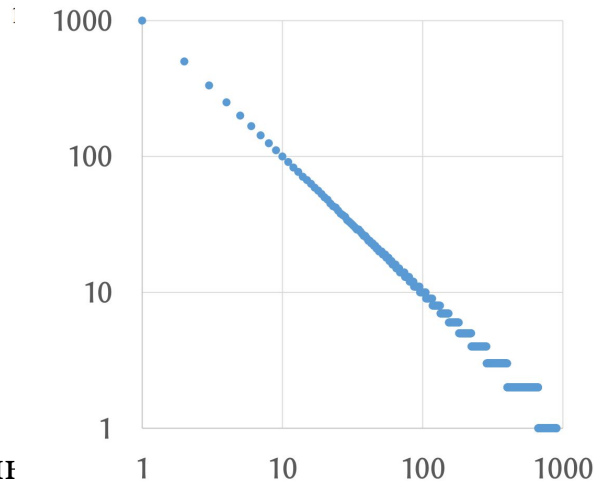
Закон Ципфа

Если все слова упорядочить по убыванию частоты, то частота n -ного слова окажется примерно обратно пропорциональна его рангу (порядковому номеру).

Например, второе по частоте слово встречается примерно в два раза реже, чем первое, третьи три раза реже, чем первое, и т. п.

$$\text{freq}(w) * \text{rank}(w)^\gamma = \text{Const}$$

Кстати, на материале больших веб-корпусов этот закон выполняется. Для морфологически богатых языков (ср. также словоформы - леммы) скорость убывания γ - поправка Бенуа Мандельброта (1965) к закону Джорджа Кингсли Ципфа (1949). Он выделил голову (стоп-слова), среднюю часть и хвост (гапаксы) - broken power law.



Частотные списки

- Могут представлять
 - словоформы, лексемы (леммы)
 - части речи, пунктуацию
 - буквы, сочетания букв
 - сочетания слов (биграммы, триграммы - для форм и лемм)
 - (синтаксические) конструкции - более сложные запросы
 - пары синтаксически связанных слов (синтаксические биграммы)

1 <i>ост</i>	1723246	6 <i>ста</i>	1281976
2 <i>ени</i>	1652167	7 <i>что</i>	1249647
3 <i>про</i>	1591587	8 <i>при</i>	1223950
4 <i>ого</i>	1564559	9 <i>енн</i>	1206823
5 <i>ств</i>	1331722	10 <i>льн</i>	1185312

(По данным Частотного словаря
русского языка, 2011)



N-граммы

И долго буду тем любезен я народу



- биграмма: сочетание словоформ, не всегда информативна



И долго буду тем любезен я народу

- синтаксическая биграмма: сочетание связанных синтаксическим отношением словоформ или лемм
- могут отличаться в зависимости от выбранного способа анализа:



И долго буду тем любезен я народу



Коллокации

Связанные (несвободные) сочетания слов, которые характеризуют язык в целом, конкретный текст или жанр/регистр

N-граммы корпуса на шкале:

случайные сочетания (*и в, красный же...*)

свободные сочетания (вы были)

коллокации (ставить условие, резкий рост)

неоднословные номинации и термины

(Иван Грозный, транспортное средство)

фраземы (идиомы) (ничего себе,
всего доброго)



Коллокации

Композициональность

- значение сложного выражения есть функция от значения его частей и правил их комбинации
- часто нарушается:
дом - красный дом, но уголлок - красный уголлок
куртка - красная куртка книга - красная книга
- носители языка легко справляются с некомпозициональностью, свободно используют и понимают единицы:
потупить взгляд, нанести урон, бурные аплодисменты,
– лексические функции



Коллокации

Можно также попросить носителей языка заполнить пропуски. Много ли вариантов получится?

между молотом и _____

тогда _____ вопрос, когда же закончится конфликт?

красный как _____

_____ к числу сторонников оппозиции

Интересный лингвистический материал:

- лексическая сочетаемость
- лексическая избирательность конструкций
- “легкие” (семантически почти пустые) глаголы и другие слова-функции
- идиоматика



Коллокации

Основные свойства коллокаций:

- некомпозициональность (*отбросить коньки != отбросить & коньки*)
- нерегулярность - коллокации не порождаются в соответствии со стандартными правилами языка
- устойчивость - нельзя заменить один элемент на другой (синоним):
крепкий чай != сильный чай
- частотность



Коллокации

Вычислительный подход:

N-граммы корпуса на шкале:

случайные сочетания (*и в, красный же...*)

свободные сочетания (вы были)

коллокации (ставить условие, резкий рост)

неоднословные номинации и термины
(Иван Грозный, транспортное средство)

фраземы (идиомы) (ничего себе,
всего доброго)

частые N-граммы

редкие N-граммы



Сила связности коллокации

Сила связности коллокаций: насколько коллокации не случайны?

Самые популярные статистические меры, позволяющие ранжировать выше редкие N-граммы:

- взаимная информация (MI): $MI(n,c) = \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}$

$f(n,c)$	$f(n)$
$f(c)$	N

- t-score: $t - score = \frac{f(n,c) - \frac{f(n) \times f(c)}{N}}{\sqrt{f(n,c)}}$

- логарифмическое правдоподобие: $log - likelihood = 2 \sum_{ij} O_{ij} \times \log \frac{O_{ij}}{E_{ij}}$

- logDice: $\log Dice = 14 + \log_2 \frac{2f(n,c)}{f(n) + f(c)}$



Сила связности коллокации

- MI - чувствительная к редким коллокатам, находит тематические коллокации
- t-мера - чувствительна к частоте самой коллокации (выводит в топ списка такие частотные пары), находит устойчивые сочетания
- (log)Dice - находит симметричные устойчивые сочетания (w_1 и w_2 встречаются только вместе), лучше работает в корпусах большого размера (миллиарды словоупотреблений)

выбор оптимальной меры зависит от задачи



Основные параметры частотных списков

- Ранг - место в списке
- Абсолютная частота
- Относительная частота
- ipm - items per million - доля употреблений на миллион слов/токенов
 - $= f(\text{item}) / N(\text{corpus size}) * 1\,000\,000$
- Корпус для сравнения (reference corpus) - позволяет сравнить **наблюдаемую** в данном корпусе частоту и **ожидаемую** частоту ("теоретическое" ("математическое") ожидание)



Значимая лексика

- частотная мера keyness

- Add-N version:

$$K = \frac{f_{foc} / T_{foc} + N}{f_{ref} / T_{ref} + N}$$

f_{foc} -- количество вхождений слова в фокусном подкорпусе

T_{foc} -- объем фокусного корпуса

f_{ref} -- количество вхождений слова в референсном подкорпусе

T_{ref} -- объем референсного корпуса

- мера логарифмического правдоподобия LL

	Подкорпус	Другие тексты	Весь корпус
Частота	a	b	a+b
Размер	c	d	c+d

На основе этой матрицы значение отношения правдоподобия G^2 (LL-score) можно вычислить как:

$$= 2(a \ln(\frac{a}{E1}) + b \ln(\frac{b}{E2})); \text{ где } E1 = c \frac{a+b}{c+d}; E2 = d \frac{a+b}{c+d}$$

Здесь a, b, c, d – наблюдаемые величины, а $E1$ и $E2$ – ожидаемый показатель в сравниваемых подкорпусах (см. Rayson & Garside 2000).



Значимая лексика корпуса: примеры

Значимая лексика (лексические маркеры): ремарки у Достоевского (Шайкевич и др. 2003)

- *ввернуть, вставить, ввязаться, включить, подсказать*
- *заторопиться, протянуть, поспешить, скороговоркой, впопыхах*
- *проворчать, промямлить, промычать, прошамкать*



Значимая лексика корпуса: примеры

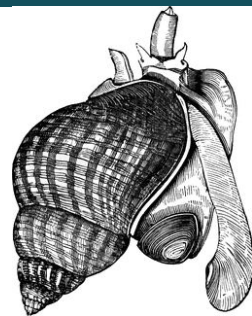
- Угадай корпус

1	ну	part	1114.6
2	да	part	787.5
3	вот	part	1785.1
4	там	advpro	1128.1
5	ты	spro	3171.2
6	угу	intj	24.6
7	я	spro	12684.4
8	нет	part	589.2
9	а	conj	8198.0
10	вообще	adv	417.6

11	у	pr	4306.1
12	знать	v	1713.8
13	говорить	v	1755.0
14	ой	intj	64.5
15	э	intj	19.4
16	э-э	intj	11.5
17	ага	intj	40.2
18	да	conj	801.0
19	давай	part	100.3
20	ладно	part	110.3

Ловушки частотных данных

- слова, часто встречающиеся в одном тексте (*веснянка, whelk*)
- стоп-слова: часто встречаются во всех текстах (*и, на, этот...*)
- все частотные меры пытаются оценить, насколько слово характерно для данного подкорпуса и насколько оно нехарактерно для контрастного подкорпуса

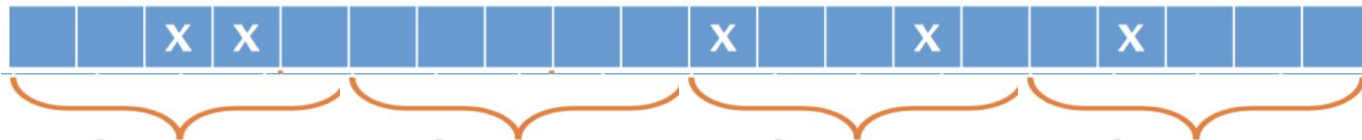


Меры дистрибуции появления единицы

- Документная частота
- Range (число секций корпуса, в которых встретилось слово, нп. $k = 100$)
- Коэффициент D Жуйяна

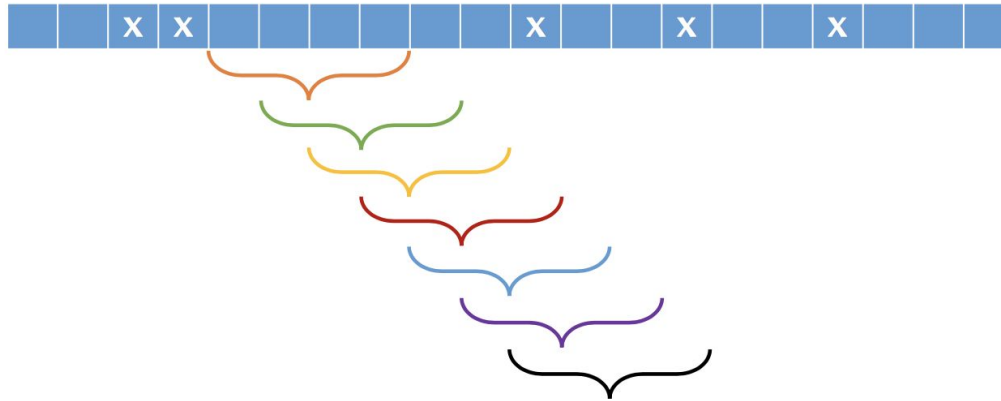
$$D = 100 \times \left(1 - \frac{\sigma}{\bar{v}\sqrt{n}}\right), \text{ где } \sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (v_i - \bar{v})^2} ; U = fD \text{ (D модифиц.)}$$

- Коэффициент DP Гриса $DP = \frac{\sum_{i=1}^n |O_i - E_i|}{2}$, где O_i, E_i - наблюдаемая и ожидаемая частота в каждом сегменте (могут быть разного размера)



Меры дистрибуции появления единицы

- ARF (Averaged Reduced Frequency)



то же количество секций,
что и в Range, но разбиение
скользит по корпусу,
начинаясь с каждого
следующего слова

Не больше v сегментов начиная с $(n_{i-1} + 1)$ -го по n_i -й содержат слово x



Ресурсы и литература

- Конкордансер **AntConc** и его производные (для [Windows, MacOS, Linux](#))
- **Voyant [tools](#)** -- для работы с собственными корпусами
- **Google N-grams [viewer](#)**
- **RusVectōrēs** и его аналоги для вычисления контекстной [близости](#) слов
- Ляшевская О. Н., Шаров С. А. Введение к частотному словарю современного русского языка (2011) [PDF](#)
- Шайкевич А. Я., В. М. Андрющенко, Н. А. Ребецкая. Статистический словарь языка Достоевского (2003). Введение. [PDF](#)
- Захаров В. П., Хохлова М. В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке (2010) [PDF](#)
- Statistics used in Sketch Engine (2015) [PDF](#), [logDice](#), [MI-score](#), [t-score](#)
- OpenCorpora.org Wiki: Коллокации [сайт](#)



Какие слова имеют близкие контексты?

[WebVectors](#)[Similar words](#)[Visualizations](#)[Calculator](#)[Miscellaneous](#)[Models](#)[About](#)

Computing associates

Enter a word to produce a list of its 10 nearest semantic tag («tea_NOUN»). Otherwise, *WebVectors* will detect it.

☒ High ☒ Medium ☐ Low

English Wikipedia

1. **competition** NOUN 0.61
2. **eurovision** NOUN 0.56
3. **entrant** NOUN 0.56
4. **winner** NOUN 0.53
5. **pageant** NOUN 0.51
6. **finalist** NOUN 0.51
7. **semi-finalist** NOUN 0.50
8. **preselection** NOUN 0.49
9. **runner-up** NOUN 0.49

English Gigaword



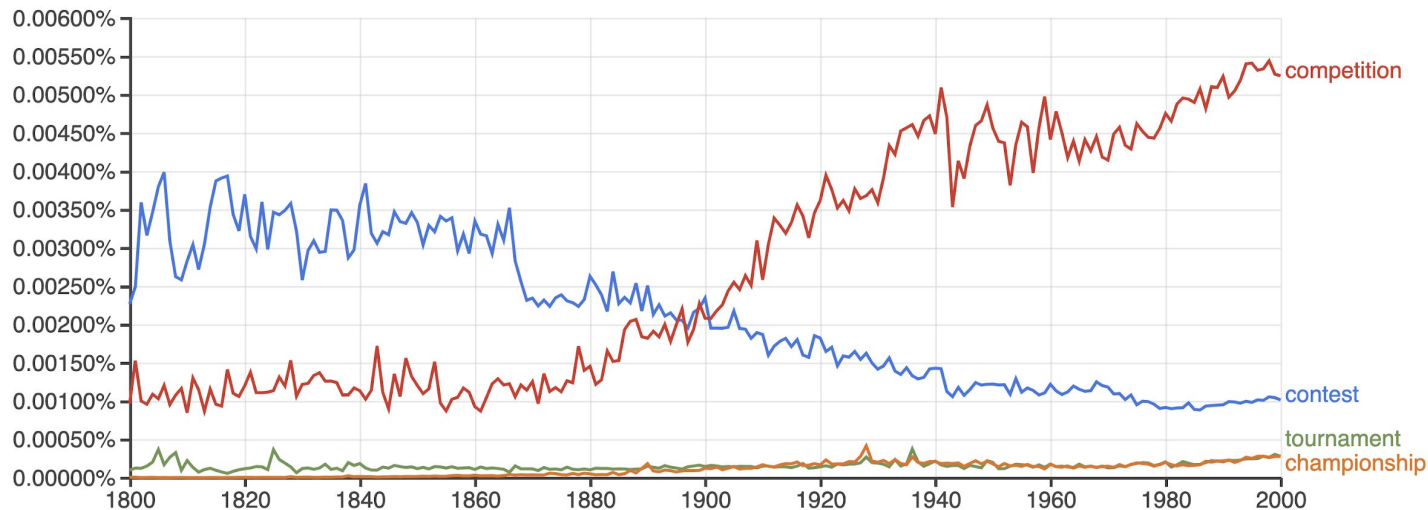
1. **race** NOUN 0.61
2. **primary** NOUN 0.58
3. **competition** NOUN 0.56
4. **contestant** NOUN 0.49
5. **match-up** NOUN 0.49
6. **duel** NOUN 0.48
7. **challenger** NOUN 0.45
8. **matchup** NOUN 0.44
9. **rematch** NOUN 0.44
10. **contender** NOUN 0.44

Распределение слов по времени

Google Books Ngram Viewer

Graph these comma-separated phrases: ☐ case-insensitive

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



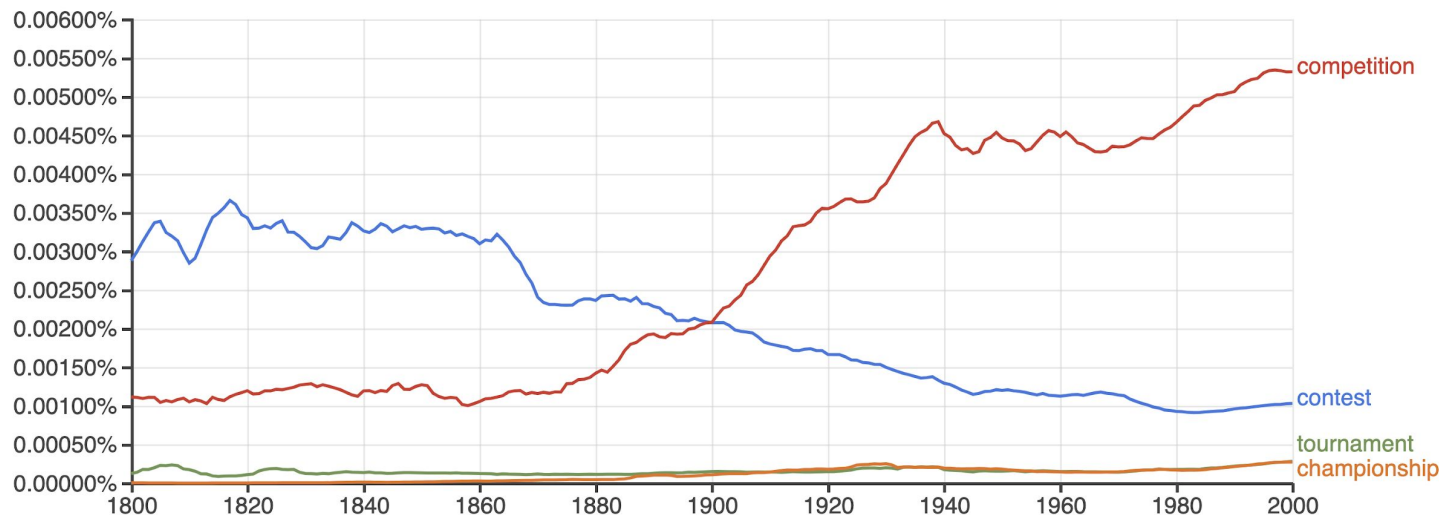
Распределение слов по времени

Google Books Ngram Viewer

Graph these comma-separated phrases: ☐ case-insensitive

between and from the corpus with smoothing of

[Search lots of books](#)

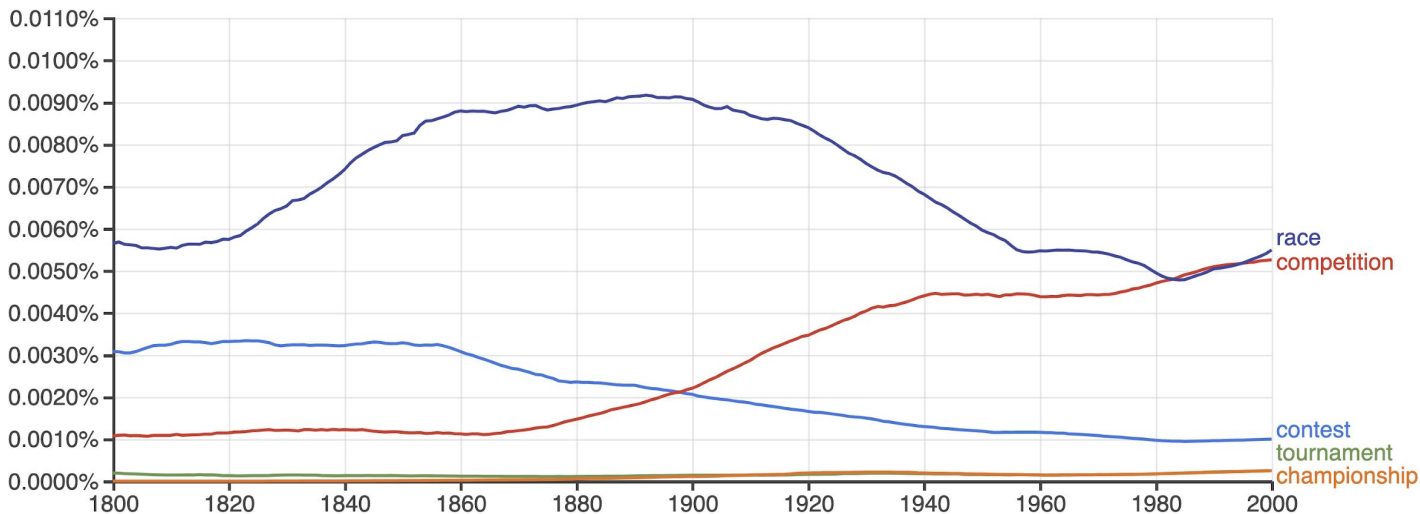


Распределение слов по времени

Google Books Ngram Viewer

Graph these comma-separated phrases: ☐ case-insensitive

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



Пользовательские корпуса

Несколько примеров

- корпус твиттера / отзывов booking.com
- корпус Михаила Шолохова
- корпус школьных сочинений
- корпус речей президентов США

Обычно отличаются

- размером (сильно больше или сильно меньше, чем BNC)
- доступностью (для себя)
- разметкой (под свои исследовательские задачи)



Need more corpus please!



NATIONAL RESEARCH
UNIVERSITY

Обработка данных для корпуса

Стандартная

- препроцессинг текстов (дубликаты, опечатки, служебная информация)
- метаразметка
- разбиение на предложения, токены
- лемматизация
-

Этап могут быть пропущены или добавлены, в зависимости от задач корпуса и нужд исследования



Need more corpus please!



NATIONAL RESEARCH
UNIVERSITY

Частотные меры

- TF*IDF

- TF*ICTF (term frequency – inverse collection term-frequency)

$$\text{TF*ICTF} = \frac{f_d}{F_d} * \log \frac{F_D}{f_D}, \text{ где}$$

f_d – количество анализируемых словоформ/лемм (term) в документе,

F_d – количество всех словоформ/лемм в анализируемом документе,

F_D – общее количество словоформ/лемм контрастном подкорпусе,

f_D – количество анализируемых слов/лемм контрастном подкорпусе.

- модифицированная

$$\text{TF*ICTF}' = (0,5 + 0,5 \frac{f_d}{F_d}) * \log \frac{F_D - F_d}{f_D - f_d}, \text{ где}$$

$F_D - F_d$ – объем контрастного подкорпуса без объема документа, в которую входит единица, для которой вычисляется вес,

$f_D - f_d$ – количество анализируемой словоформы в контрастном подкорпусе, кроме количества словоформы в документе, в которую входит анализируемая единица⁹.