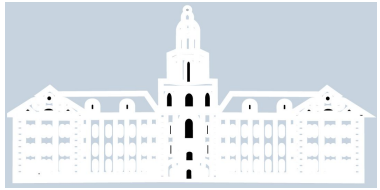


Лекция 4

Как интерпретировать цифры

Частотные списки, значимая лексика, коллокации

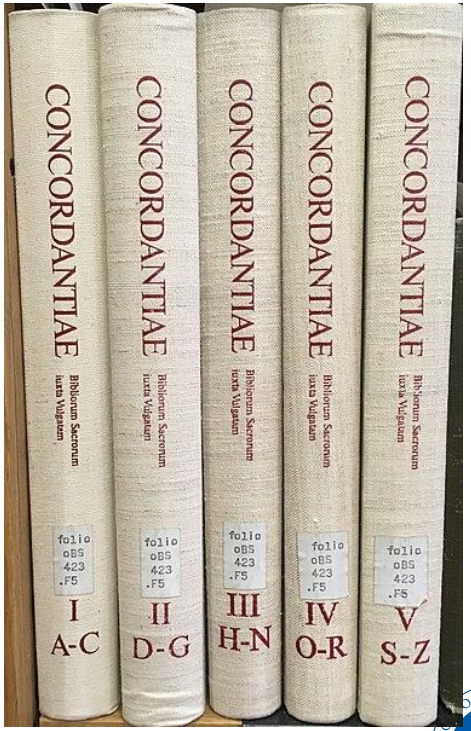
Ольга Ляшевская ** olesar@yandex.ru

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

Корпус - не только примеры

Что мы можем узнать из корпуса текстов:

- о словах?
- о корпусе (подкорпусах)?



Частотные списки

- Составляются для
  - всего корпуса
  - подкорпусов отдельных авторов, жанров, периодов и т. п.
  - для зоны заголовков, рифмовки в поэзии и т. п.

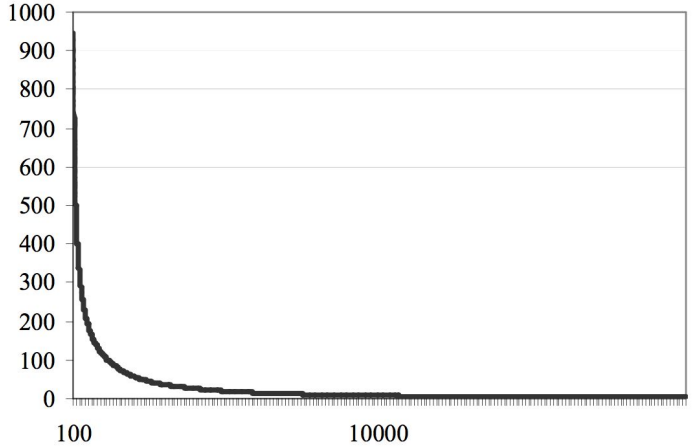
1	и	12851	99	лицо	275	999	можете	27	9999	вытянул	2
2	не	6474	100	сказать	275	1000	мои	27	10000	вытянуть	2
3	что	6070	101	этот	272	1001	Москвы	27	10001	выучить	2
4	в	5689	102	вас	271	1002	несомненно	27	10002	выучиться	2
5	он	5526	103	Левина	271	1003	новым	27	10003	выходившей	2
6	на	3584	104	раз	271	1004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

... второе по частоте слово встречается
~ в два раза реже, чем первое, третье -
~ в три раза реже, чем первое, и т. п.

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



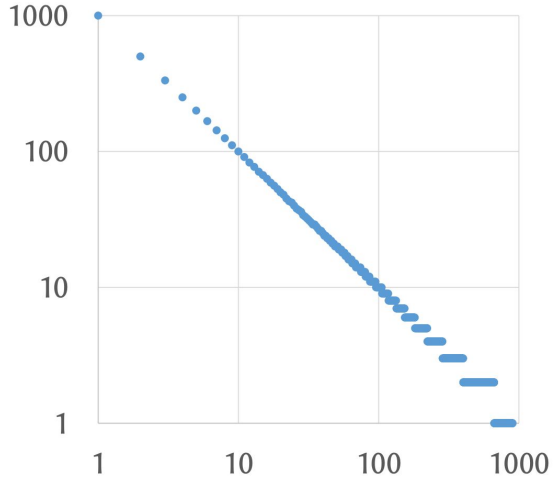
*Впервые сформулирован Жаном-Батистом Эсту как freq * rank = const (1908, Диапазон стенографии)

Закон Ципфа

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

... второе по частоте слово встречается
~ в два раза реже, чем первое, третье -
~ в три раза реже, чем первое, и т. п.

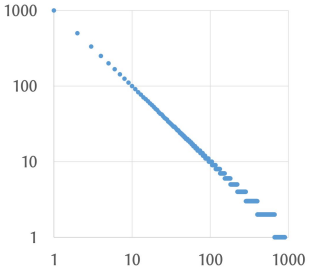
$$freq(w) * (rank(w) + \beta)^{\gamma} = Const$$



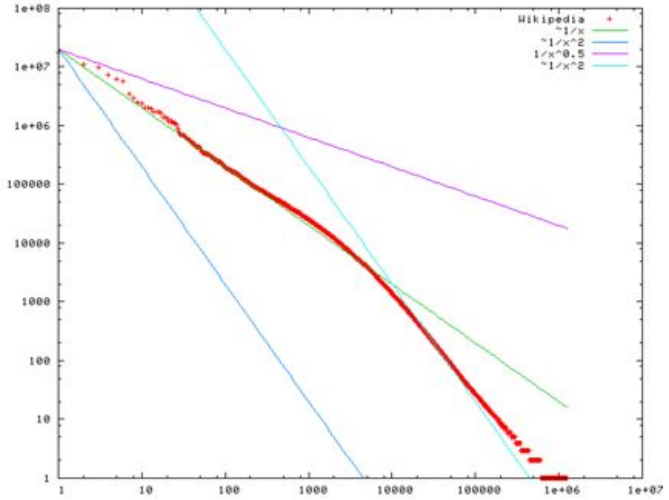
Кстати, на материале больших веб-корпусов этот закон выпол
Для морфологически богатых языков (ср. также словоформы - леммы) скорость убывает иначе.
γ - поправка Бенуа Мандельброта (1965) к закону Джорджа Кингсли Ципфа (1949).

Ломаная версия

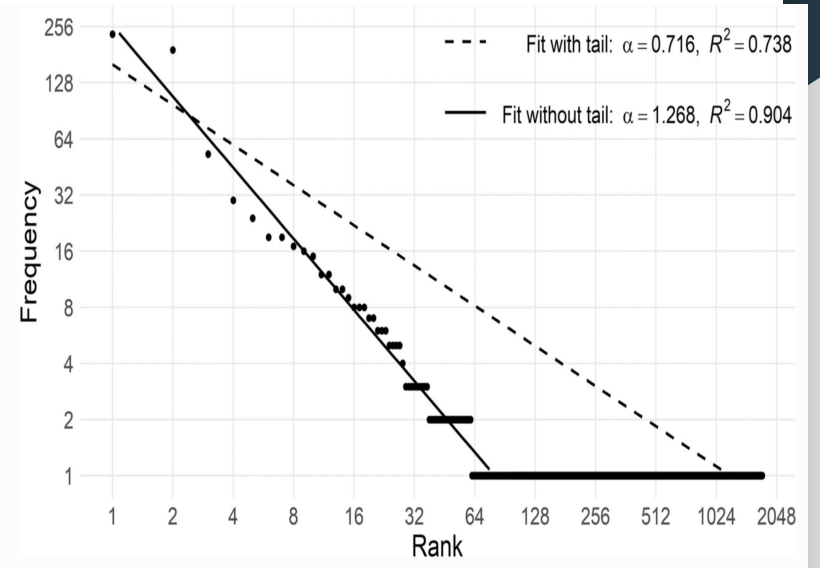
More examples:



RNC main corpus



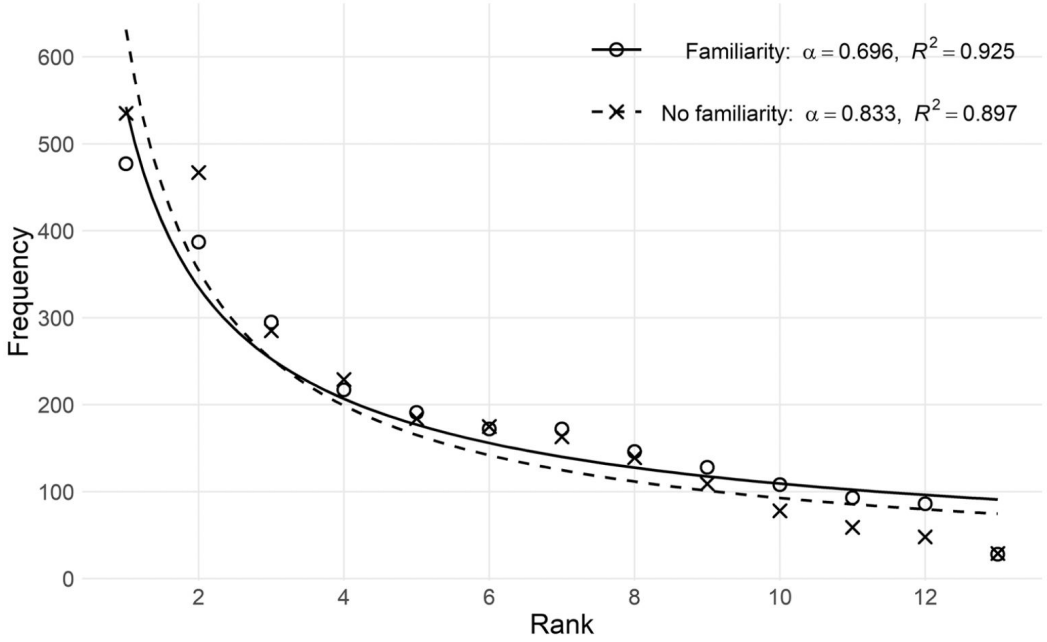
English Wikipedia (2009)



Switchboard Corpus, phone conversations between two strangers (Linders & Louwerse, 2023)

- broken power law: head - middle part - tail (hapax legomena)
  Б. Мандельброт выделил голову (стоп-слова), среднюю часть и хвост (гапаксы)

Закон Ципфа в сравнительных исследованиях



Закон Ципфа в двух корпусах диалогов:
разговаривающие знакомы друг с другом и нет
(Linders & Louwerse 2023)

Основные параметры частотных списков

- Частотный ранг - место в списке
- Абсолютная частота
- Относительная частота
  - ipm - items per million - доля употреблений на миллион слов (токенов) = *f*(item) / N(corpus size) * 1 000 000

сладкий - f_A=1472 в подкорпусе поэзии 1800-1850 гг. (N_A=2 млн.)
- f_B=2505 в подкорпусе поэзии 1900-1950 гг. (N_B=5 млн.)

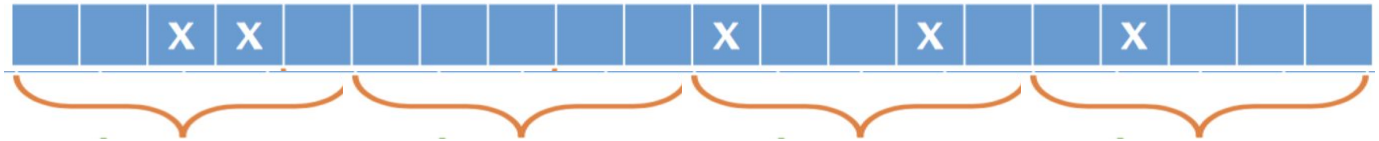
A: 677,7 *ipm* B: 458,4 *ipm**

.

N_A = 2 172 025, N_B = 5 464 590 словоупотреблений, включая вхождения наречия/предикатива *сладко*

Меры разброса появления единицы

- Документная частота
- Range (число сегментов корпуса, в которых встретилось слово, $k = 100$)



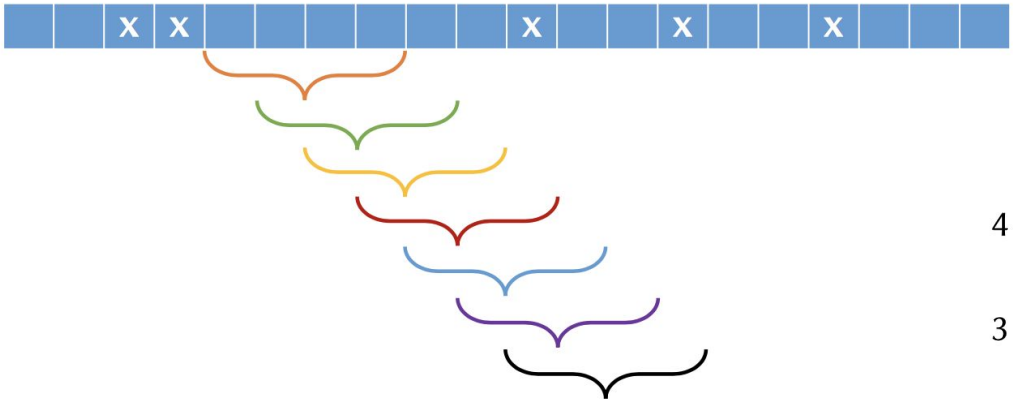
- Коэффициент *DP* Гриса $$DP = \frac{\sum_{i=1}^{n} |O_i - E_i|}{2}$$, где O_i, E_i - наблюдаемая и ожидаемая частота в каждом сегменте (могут быть разного размера)

- Коэффициент *D* Жуйяна

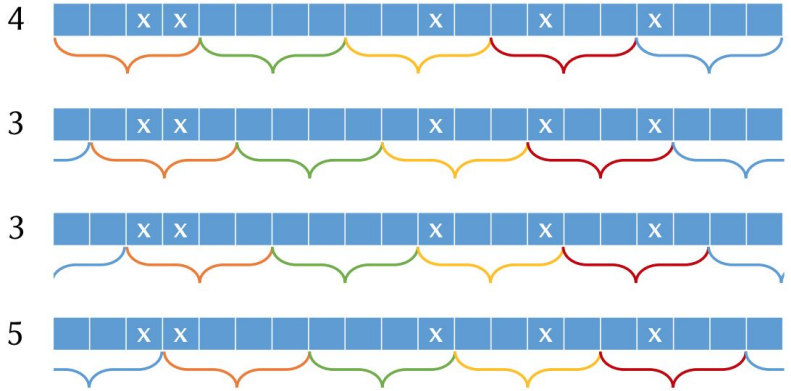
$$D = 100 \times \left(1 - \frac{\sigma}{\bar{v}\sqrt{n}}\right), \text{где} \quad \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(v_i - \bar{v})^2} \quad ; U = f\,D$$ (D модифиц., U Жуйяна)

Меры разброса появления единицы

- ARF (Averaged Reduced Frequency)

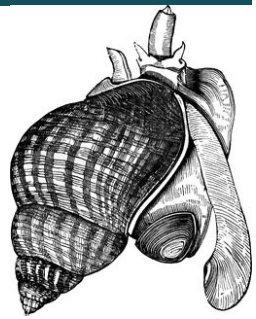


то же количество сегментов, что и в Range, но разбиение скользит по корпусу, начинаясь с каждого следующего слова



ARF = (4 + 3 + 3 + 5) / 4 = 3,75

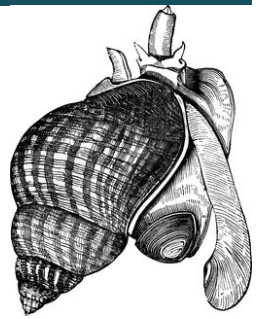
Ловушки частотных данных



Но груз, привязанный к веревке, — еще не **якорь** в современном смысле этого слова, ибо держащая сила современных **якорей** в десятки раз превосходит их собственный вес. А у камня, лежащего на дне, держащая сила даже меньше его веса. Поэтому назовем камень с привязанной к нему веревкой «якорным камнем». «Держащая сила **якоря** — это сила, которая приходится на единицу его веса и должна быть приложена для того, чтобы вырвать **якорь** из грунта в момент, когда веретено **якоря** расположено горизонтально. Удерживающая способность **якоря** — это сила, удерживающая корабль, который стоит на **якоре**, от перемещения под воздействием ветра и течения. Удерживающая способность **якоря** определяется произведением держащей силы **якоря** на его вес»… Л. Н. Скрягин. Книга о якорях (1973)

Все примеры — 1744

Ловушки частотных данных



- слова, **часто** встречающиеся в одном тексте (*веснянка, whelk*)
- слова, встретившиеся в одном-двух текстах (*страбизм*)
- стоп-слова: часто встречаются во всех текстах (*и, на, этот…*)
- не-слова: (*фывапролдж, кутруаз…*)

Меры разброса появления единицы

- Высокий Range, высокий D — лексическое “ядро” языка, активное слово
- Низкий R, высокий D — низкочастотное слово, периферия языка
- Низкий R, низкий D — тематически окрашенное слово
- Высокий R, низкий D — слово есть почти во всех сегментах, но с очень разными частотами ⇒ где-то наблюдается “whelk problem”

Приложения: wordandphrase

survive on. " Thomas Homer-Dixon has published extensively on population, environment and conflict. His theoretical framework first appeared in several seminal articles in the early to mid-1990s, 2 and research by Homer-Dixon and his colleagues on the relationship between population, environment and **violent** conflict has been influential in the field of political **science.** 3 In view of the widespread nature of human **conflict4** and the prevailing pessimism about population growth and environmental **destruction**, this linkage is clearly both important and policy-relevant. In addition to **outlining** what he believes to be the main causal mechanisms, Homer-Dixon has also examined a number of cases in detail. Homer-Dixon's work is far removed from the simplifications of some of the popular literature on the theme of population, environment and conflict. There is little if any of the **sensationalism** of Robert D. Kaplan5 or the **doomsday predictions** of Paul R. Ehrlich and Anne H. Ehrlich. 6 Homer-Dixon never asserts that population pressure and environmental degradation are the **sole** source of **violent**

SEE LISTS	FREQ RANK	1-500	501-3000	3000+		ACAD
	166 WORDS	54%	7%	6%		23%

https://www.english-corpora.org/coca/

Приложения: лексические минимумы

Table 12.1 EFL vocabulary size, formal EFL exams and the CEFR (from Meara and Milton, 2003, p. 5)

CEFR level	*Cambridge exam*	*XLex score (max. 5000)*
A1	Starters, Movers and Flyers	<1500
A2	Kernel English Test (KET)	1500–2500
B1	Preliminary English Test (PET)	2750–3250
B2	First Certificate in English (FCE)	3250–3750
C1	Cambridge Advanced English (CAE)	3750–4500
C2	Cambridge Proficiency in English (CPE)	4500–5000

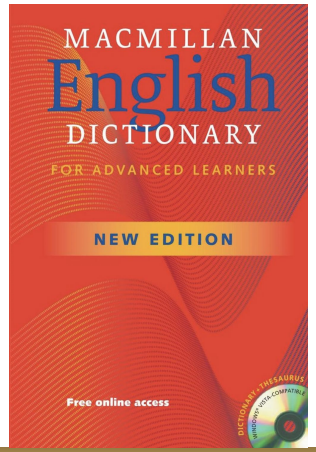
Приложения: словарь

★★★ — 1–2500 места в частотном списке (*the*, *animal*)

★★ — 2501–5000 места в частотном списке (*appropriate*, *tragedy*)

★ — 5001–7500 места в частотном списке (*restriction*, *allegedly*)

без звёдочек — остальные (*crescent*, *thatch*)



edge[1] /edʒ/ noun ★★★

1 part furthest out
2 sharp side of blade/tool
3 angry tone in voice
4 advantage
5 strange quality
+ PHRASES

1 [C] the part of something that is furthest from its centre: *Bring the two edges together and fasten them securely.* ♦ **+of** *The railway station was built on the edge of town.* ♦ *Victoria was sitting on the edge of the bed.*
2 [C] the sharp side of a blade or tool that is used for cutting things: *the knife's edge*
3 [singular] a quality in the way that someone speaks that shows they are becoming angry or upset: **+to/in** *Had she imagined the slight edge to his voice?*
4 [singular] an advantage that makes someone or something more successful than other people or things: **give sb/sth an/the edge over sb/sth** *Training can give you the edge over your competitors.*
5 [singular] a strange quality that something such as a piece of music or a book has that makes it interesting or exciting: *There is an edge to his new album that wasn't there in the last one.*
PHRASES **live on the edge** to have a life with many ...rs and risks, especially because you like to behave ...me and unusual way: *Despite the apparent*

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

Значимая лексика корпуса: примеры

Значимая лексика (лексические маркеры, ключевые слова): 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

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

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

Dataslov:
по какому корпусу создан этот список?

- QR-код
- Вакцина
- Вакцинация
- Инфоцыганство
- Киберспорт
- Ковикулы
- Пандемия
- Прививка
- Санкции
- Шмурдяк

Значимая лексика

- эти частотные меры должны оценить, насколько слово характерно для данного подкорпуса и насколько оно нехарактерно для контрастного подкорпуса

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

- частотная мера keyness $K = (a / c) / (b / d)$
  - версия Add-N keyness: добавим N (например, 10) к *относительным частотам*

$$K = ((a / c) + 10) / ((b / d) + 10)$$

  - помогает в случае, если b=0, сглаживает частотные эффекты: например, при N=1 выше будут ранжированы редкие слова, а при N=1000 выше будут ранжированы высокочастотные слова

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).

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 единица[9].

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)

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, характеризуют язык, текст, жанр

N-грамы 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

Можно также опросить носителей языка: характерные сочетания

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

Связанные (несвободные) сочетания слов, характеризуют язык, текст, жанр

N-грамы 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-грамы

редкие N-грамы

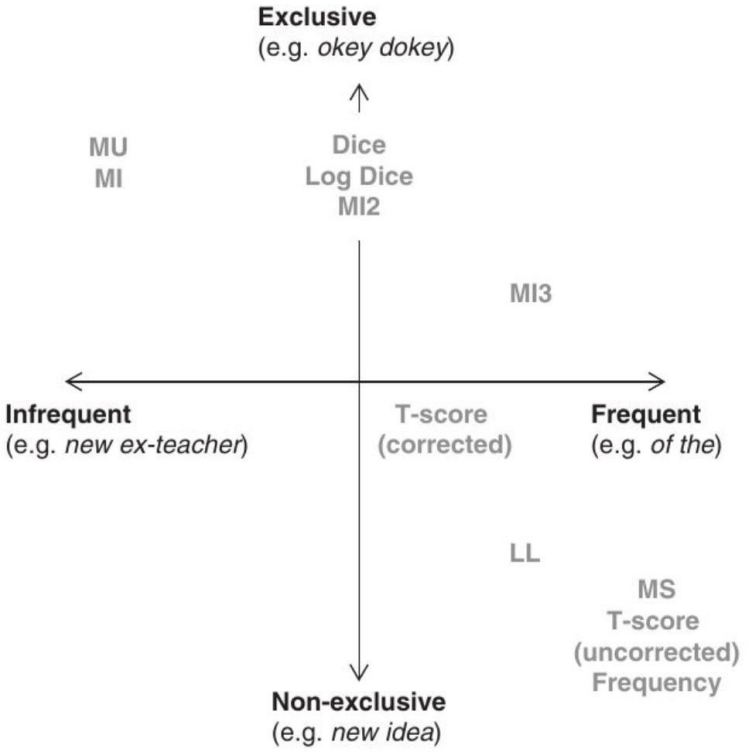
Сила связности коллокации

Ассоциативные меры: насколько коллокации не случайны?
Самые популярные статистические меры, позволяющие ранжировать выше редкие N-грамы:

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

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

Сила связности коллокации



Эксклюзивность – насколько обе единицы не употребляются друг без друга

Притяжение и **зависимость** – насколько один элемент употребляется только с данным словом, может быть, он больше любит компанию других слов?

Brezina V. (2018). Statistics in Corpus Linguistics. : A Practical Guide. Cambridge University Press.

Statistic	Equation
Freq. of co-occurrence	O_{11}
MU	$\frac{O_{11}}{E_{11}}$
MI (mutual information)	$\log_2 \frac{O_{11}}{E_{11}}$
MI2	$\log_2 \frac{{O_{11}}^2}{E_{11}}$
MI3	$\log_2 \frac{{O_{11}}^3}{E_{11}}$
LL (log likelihood)	$2 \times \left(\begin{array}{c} O_{11} \times log \frac{O_{11}}{E_{11}} + O_{21} \times log \frac{O_{21}}{E_{21}} + \\ O_{12} \times log \frac{O_{12}}{E_{12}} + O_{22} \times log \frac{O_{22}}{E_{22}} \end{array} \right)$
Z-score$_1$	$\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$

Statistic	Equation
T-score	$\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$
Dice	$\frac{2 \times O_{11}}{R_1 + C_1}$
Log Dice	$14 + \log_2 \frac{2 \times O_{11}}{R_1 + C_1}$
Log ratio	$\log_2 \frac{O_{11} \times R_2}{O_{21} \times R_1}$
MS (minimum sensitivity)	$min\left(\frac{O_{11}}{C_1}, \frac{O_{11}}{R_1}\right)$
Delta P	$\frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}; \frac{O_{11}}{C_1} - \frac{O_{12}}{C_2}$
Cohen's d	$\frac{Mean_{in\ window} - Mean_{outside\ wi}}{pooled\ SD}$

Разберем подробнее

My **love** is like a red, red rose that's newly sprung in June: My **love** is like the melody that's sweetly played in tune. As fair art thou, my bonnie lass, so deep in **love** am I: And I will **love** thee still, my dear, till a' the seas gang dry. Till a' the seas gang dry, my dear, and the rocks melt wi' the sun : And I will **love** thee still, my dear, while the sands o' life shall run. And fare thee weel, my only **love**, and fare thee weel a while! And I will come again, my **love**, thou'it were ten thousand mile. Robert Burns, A Red, Red Rose

N = 107 токенов
f (my, love) = 3
f (love) = 7
f (my) = 8
context window (1L,1R) = 2

	коллокат есть	*коллоката нет*	
ключ есть	$O_{11}=f(n,c)=3$	4	$R_1=f(n)=7$
нет ключа	5^*		R_2
всего	$C_1=f(c)=8$	C_2	$N = 107$

Разберем подробнее

My **love** is like a red, red rose that's newly sprung in June: My **love** is like the melody that's sweetly played in tune. As fair art thou, my bonnie lass, so deep in **love** am I: And I will **love** thee still, my dear, till a' the seas gang dry. Till a' the seas gang dry, my dear, and the rocks melt wi' the sun : And I will **love** thee still, my dear, while the sands o' life shall run. And fare thee weel, my only **love**, and fare thee weel a while! And I will come again, my **love**, thou'it were ten thousand mile. Robert Burns, A Red, Red Rose

N = 107 токенов
f (my, love) = 3
f (love) = 7
f (my) = 8
context window (1L,1R) = 2

	коллокат есть	*коллоката нет*	всего
ключ есть	$O_{11}=f(n,c)=3$	4	$R_1=f(n)=7$
нет ключа	5^*		R_2
всего	$C_1=f(c)=8$	C_2	$N = 107$

Математическое ожидание:
$E_{11}=R_1 \times C_1 / N = 7 \times 8 : 107 = 0{,}52$

Математическое ожидание (с коррекцией на окно):
$E_{11}=R_1 \times C_1 \times 2 / N = 7 \times 8 \times 2 : 107 = 1{,}05$

Разберем подробнее

My **love** is like a red, red rose that's newly sprung in June: My **love** is like the melody that's sweetly played in tune. As fair art thou, my bonnie lass, so deep in **love** am I: And I will **love** thee still, my dear, till a' the seas gang dry. Till a' the seas gang dry, my dear, and the rocks melt wi' the sun : And I will **love** thee still, my dear, while the sands o' life shall run. And fare thee weel, my only **love**, and fare thee weel a while! And I will come again, my **love**, thou'it were ten thousand mile. Robert Burns, A Red, Red Rose

N = 107 токенов
f (my, love) = 3
f (love) = 7
f (my) = 8
context window (1L,1R) = 2

	коллокат есть	*коллоката нет*	всего
ключ есть	$O_{11}=f(n,c)=3$	4	$R_1=f(n)=7$
нет ключа	5^*		$R_2=100$
всего	$C_1=f(c)=8$	$C_2=99$	$N = 107$

$E_{11}=0,52$	$E_{12}=6,48$	$R_1=7$
$E_{21}=7,48$	$E_{22}=92,52$	$R_2=100$
$C_1=8$	$C_2=99$	$N = 107$

Сила связности коллокации

2. Метрики, основанные на математическом ожидании:

MI = $\log_2 \frac{O_{11}}{E_{11}}$ если один из членов коллокации низкочастотный, то $E_{11}=f(n)\times f(c)/N$ становится очень маленьким, а значение MI - очень большим,
высоко ранжируются коллокаты во фразеологизмах типа *бить баклуши*,
где один из элементов редко употребляется вне коллокации, термины,
профессиональная лексика

t-score = $\frac{O_{11}-E_{11}}{\sqrt{O_{11}}}$ вес E_{11} в формуле незначительный, обычно чем больше O_{11}, тем выше t-score

несколько понижает в ранге стоп-слова (предлоги, союзы и т. п.),
хорошо выделяет неоднословные предлоги и др. служебные мультитокены

Сила связности коллокации

2. Метрики, основанные на математическом ожидании:

MI = $\log_2 \frac{O_{11}}{E_{11}}$ если один из членов коллокации низкочастотный, то $E_{11}=f(n)\times f(c)/N$ становится очень маленьким, а значение MI - очень большим,
высоко ранжируются коллокаты во фразеологизмах типа *бить баклуши*,
где один из элементов редко употребляется вне коллокации, термины,
профессиональная лексика

t-score = $\frac{O_{11}-E_{11}}{\sqrt{O_{11}}}$ вес E_{11} в формуле незначительный, обычно чем больше O_{11}, тем выше t-score

несколько понижает в ранге стоп-слова (предлоги, союзы и т. п.),
хорошо выделяет неоднословные предлоги и др. служебные мультитокены

LL-score: MI умножается на O_{11} (вклад E_{11} уменьшается), аналогично просчитывается отклонение во остальных 3х клетах таблицы и складывается;
затем все умножается на 2
несколько понижает в ранге стоп-слова, местоимения и числительные

! можно вначале обрезать список по частоте коллоката, потом ранжировать

Сила связности коллокации

2. Метрики, основанные на математическом ожидании:

MI = $\log_2 \frac{O_{11}}{E_{11}}$ если один из членов коллокации низкочастотный, то $E_{11}=f(n)\times f(c)/N$ становится очень маленьким, а значение MI - очень большим,
высоко ранжируются коллокаты во фразеологизмах типа *бить баклуши*, где один из элементов редко употребляется вне коллокации

t-score = $\frac{O_{11}-E_{11}}{\sqrt{O_{11}}}$ вес E_{11} в формуле незначительный, обычно чем больше O_{11}, тем выше t-score

несколько понижает в ранге стоп-слова (предлоги, союзы и т. п.)

3. Метрики н$\frac{O_{11}}{R_1}$имметричные:

Attraction = - “вес” f(n.c) в частоте ключа f(n), Reliance = $\frac{O_{11}}{C_1}$ - “вес” f(n,c) в частоте f(c)

$\Delta P = \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}; \frac{O_{11}}{C_1} - \frac{O_{12}}{C_2}$ две метрики, смотрят на соотношение пропорций по строкам столбцам: насколько ключ пропорционально чаще встречае в коллокации, чем не-ключ; аналогично для коллок

Сила связности коллокации

2. Метрики, основанные на математическом ожидании:

MI = $\log_2 \frac{O_{11}}{E_{11}}$ если один из членов коллокации низкочастотный, то $E_{11}=f(n)\times f(c)/N$ становится очень маленьким, а значение MI - очень большим,
высоко ранжируются коллокаты во фразеологизмах типа *бить баклуши*,
где один из элементов редко употребляется вне коллокации

t-score = $\frac{O_{11}-E_{11}}{\sqrt{O_{11}}}$ вес E_{11} в формуле незначительный, обычно чем больше O_{11}, тем выше t-score

несколько понижает в ранге стоп-слова (предлоги, союзы и т. п.)

3. Ме чные:

$\Delta P = \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}; \frac{O_{11}}{C_1} - \frac{O_{12}}{C_2}$ две метрики, смотрят на соотношение пропорций по строкам и столбцам: насколько ключ пропорционально чаще встречается в коллокации, чем не-ключ; аналогично для коллоката

4. Метрики дисперсии: Cohen's *d*

Примеры

	Cooccurrence count	Candidate count	T-score
,	523,108	1,435,343,425	609.532
.	403,526	939,675,323	550.464
и	258,726	563,822,183	445.127
"	221,153	312,665,167	432.167
в	154,718	550,416,368	313.149
говорить	90,855	13,510,062	298.853
не	89,909	203,518,343	260.951
это	83,795	90,162,067	271.624
быть	79,201	155,211,375	249.820
о	74,032	52,615,938	261.006
--	68,626	131,179,752	233.268
весь	63,693	60,666,300	238.599
что	61,511	145,980,475	214.282
по	59,288	115,135,561	216.393
сказать	53,383	12,413,704	227.968
на	53,127	267,904,726	163.883
-	50,652	91,782,970	201.689
комсомольский	49,459	170,517	222.349

коллокации существительного *правда*, Aranea Russicum Maximum

(ранжирование по t-score)

По какой метрике теперь?

	Cooccurrence count	Candidate count	T-score
,	523,108	1,435,343,425	609.532
.	403,526	939,675,323	550.464
и	258,726	563,822,183	445.127
"	221,153	312,665,167	432.167
в	154,718	550,416,368	313.149
говорить	90,855	13,510,062	298.853
не	89,909	203,518,343	260.951
это	83,795	90,162,067	271.624
быть	79,201	155,211,375	249.820
о	74,032	52,615,938	261.006
--	68,626	131,179,752	233.268
весь	63,693	60,666,300	238.599
что	61,511	145,980,475	214.282
по	59,288	115,135,561	216.393
сказать	53,383	12,413,704	227.968
на	53,127	267,904,726	163.883
-	50,652	91,782,970	201.689
комсомольский	49,459	170,517	222.349

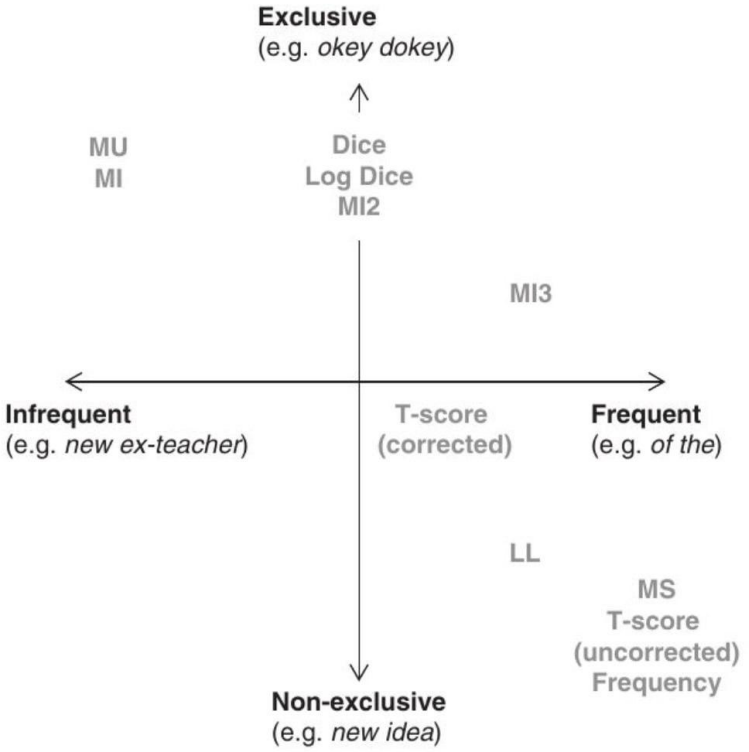
	Cooccurrence count	Candidate count
Алеманнская	16	11
Помезанская	10	10
комсомольсая	10	10
ukrpravda@gmail.com	20	20
Охотско-эвенская	13	13
Открыет	11	11
сырмяжная	12	12
Рипуарская	30	30
Салическая	344	362
Салическую	23	25
quotКомсомольская	11	12
серьмяжная	21	23
Помезанской	30	33
Бодхисаттвская	10	11
Салической	414	456
stapravda.ru	60	67
Вятско-Полянская	82	92

Примеры - LogDice

	Cooccurrence count	Candidate count	T-score
,	523,108	1,435,343,425	609.532
.	403,526	939,675,323	550.464
и	258,726	563,822,183	445.127
"	221,153	312,665,167	432.167
в	154,718	550,416,368	313.149
говорить	90,855	13,510,062	298.853
не	89,909	203,518,343	260.951
это	83,795	90,162,067	271.624
быть	79,201	155,211,375	249.820
о	74,032	52,615,938	261.006
--	68,626	131,179,752	233.268
весь	63,693	60,666,300	238.599
что	61,511	145,980,475	214.282
по	59,288	115,135,561	216.393
сказать	53,383	12,413,704	227.968
на	53,127	267,904,726	163.883
-	50,652	91,782,970	201.689
комсомольский	49,459	170,517	222.349

	Cooccurrence count	Candidate count	logDice
комсомольский	49,459	170,517	10.279
вера	29,533	1,341,490	8.611
газета	33,591	1,856,581	8.524
неправда	12,367	90,445	8.371
ложь	14,125	377,374	8.259
говорить	90,855	13,510,062	7.667
горький	7,539	214,264	7.518
служить	17,806	2,251,397	7.429
справедливость	7,742	474,269	7.301
божий	10,104	1,300,846	7.087
доля	12,759	1,998,788	7.060
сказать	53,383	12,413,704	7.012
узнать	20,211	4,085,879	6.987
правда	16,937	3,372,412	6.944
исторический	11,172	2,149,594	6.801
чистый	12,581	2,631,252	6.774
вымысел	3,850	43,200	6.744
истина	6,475	866,896	6.728
рассказать	16,434	4,026,452	6.705

Сила связности коллокации



Эксклюзивность – насколько обе единицы не употребляются друг без друга

Притяжение и **зависимость** – насколько один элемент употребляется только с данным словом, может быть, он больше любит компанию других слов?

Brezina V. (2018). Statistics in Corpus Linguistics. : A Practical Guide. Cambridge University Press.

Сила связности коллокации

Ассоциативные меры: насколько коллокации не случайны?
Самые популярные статистические меры, позволяющие ранжировать выше редкие N-грамы:

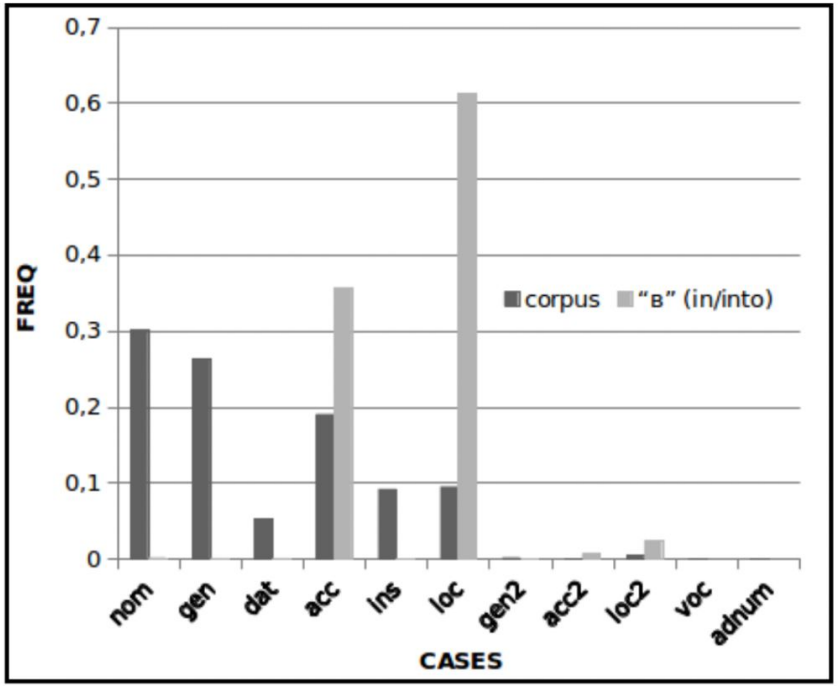
- взаимная информация (MI, PMI, MI^3): $MI(n,c)=log_2 \frac{f(n,c)\times N}{f(n)\times f(c)}$
- t-score: $$t-score=\frac{f(n,c)-\frac{f(n)\times f(c)}{N}}{\sqrt{f(n,c)}}$$

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

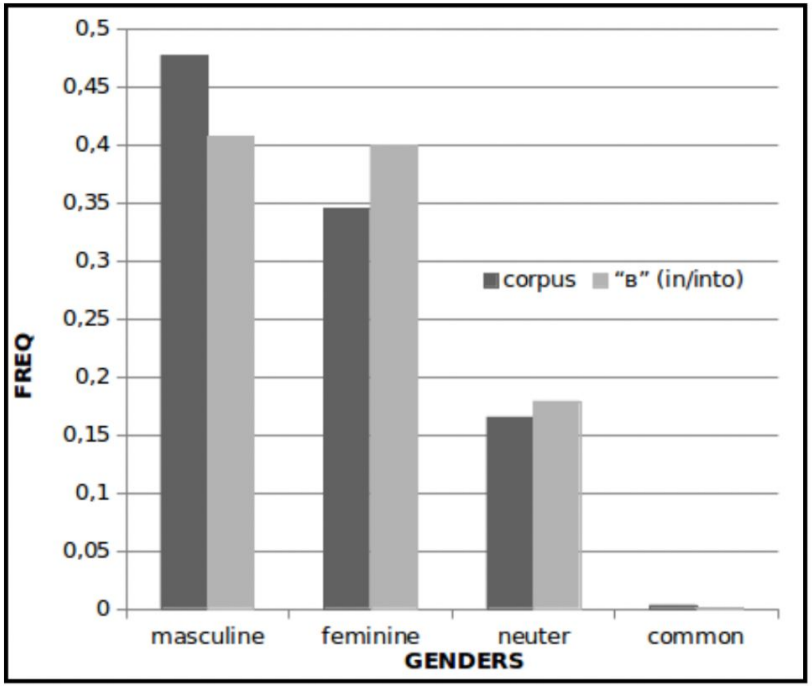
- логарифмическое правдоподобие: $$log\text{-}likelihood = 2\sum_{ij} Oij \times \log \frac{Oij}{Eij}$$
- logDice: $$\log Dice = 14 + \log_2 \frac{2f(n,c)}{f(n)+f(c)}$$

Профили поведения языковых единиц

Распределение падежей существительного в целом и после предлога “в”

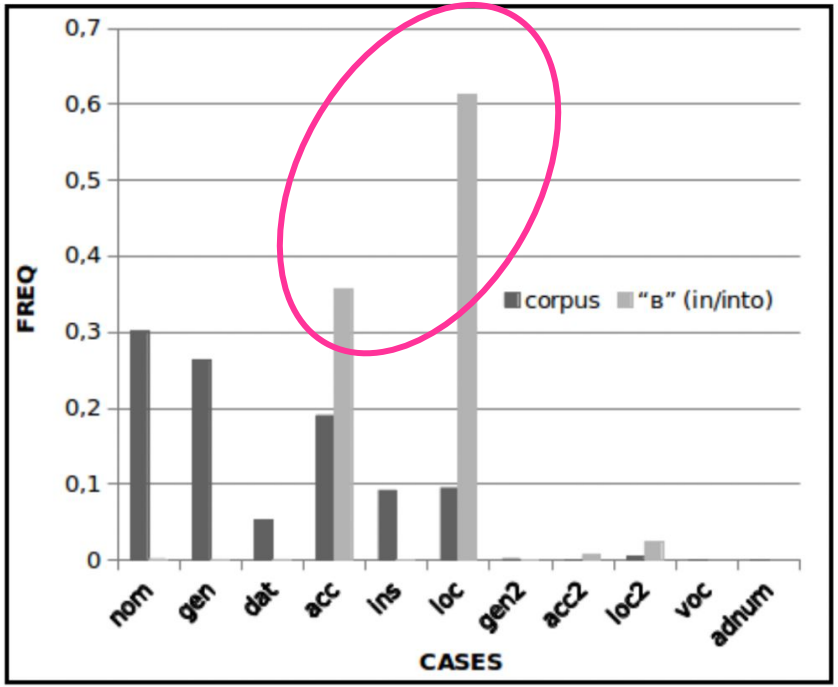


Распределение рода существительного в целом и после предлога “в” (Kopotev, Pivovarova 2013)

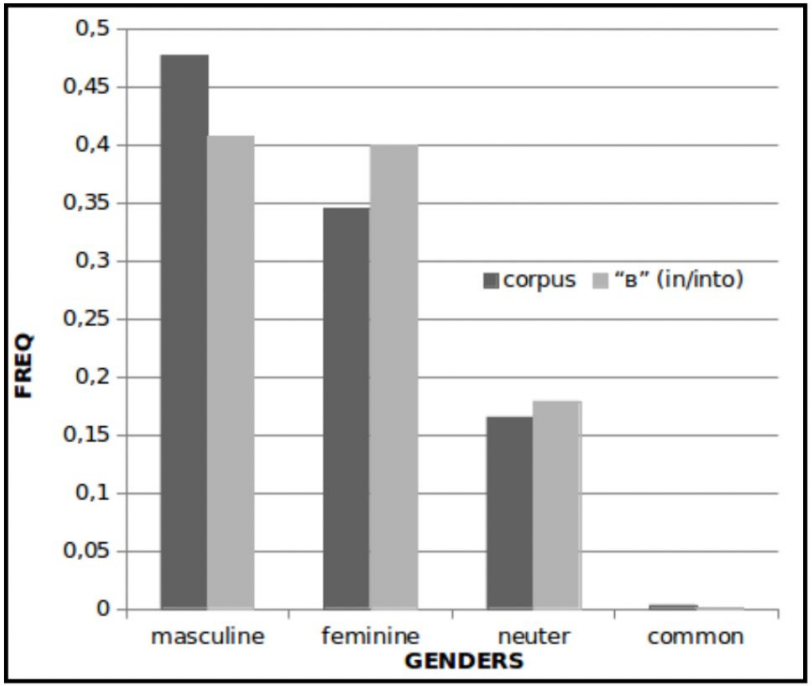


Профили поведения языковых единиц

Распределение падежей существительного в целом и после предлога “в”



Предлог контролирует падеж существительного, но не его род!



Профили поведения языковых единиц

Не до + S.Gen – **коллигация (сочетание с открытым списком лексем)**
Не до + $S.Gen_2$ (не до жиру, не до смеху) – **коллокация (сочетание с лексемой)**

Как определить характеристики слота в n-грамме не до ___ ?

Смотрим распределение всех лингвистических категорий (часть речи, одушевленность, падеж, время и т. д.)...

\+ Что устойчивее – распределение категорий или токенов или лемм?

Профили поведения языковых единиц

Не до + S.Gen – **коллигация (сочетание с открытым списком лексем)**
Не до + $S.Gen_2$ (не до жиру, не до смеху) – **коллокация (сочетание с лексемой)**

Как определить характеристики слота в n-грамме не до ___ ?

1) отношение частот значения категории: для Gen_2 weirdness = $\frac{f(Gen_2 \mid \text{не до } \_\_)}{f(Gen_2 \text{ в корпусе})}$

ср. Gen_2=(8/229) / (2344/5 988 177) = 89,25, Gen = (217/229) / (689 657/5 988 177) = 6,3

если weirdness < 1, то появление данного значения категории в n-грамме случайно

2) дивергенция (отклонение) для всей категории:

$$\text{Div(Case)} = \sum_{i=1}^{N} f(Case_i \mid \text{не до } \_\_) \cdot log \frac{f(Case_i \mid \text{не до } \_\_)}{f(Case_i \text{ в корпусе})}$$

* дивергенция Кульбака-Лейблера, подсчет по снятому корпусу НКРЯ

Профили поведения языковых единиц

Еще пример: слово в ___

В корпусе встречаются сочетания: случайные (*слово в театр*), частотные свободные сочетания (*слово в предложении*), устойчивые обороты (*слово в защиту*), идиомы (*слово в слово*)

Категория	**Нормализованная дивергенция**
Одушевленность	0,42
Падеж	0,39
Род	0,36
Число	0,20

Профили поведения языковых единиц

Еще пример: слово в ___

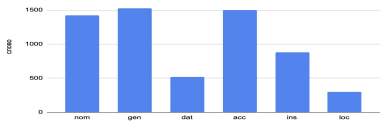
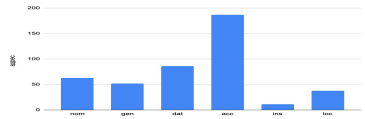
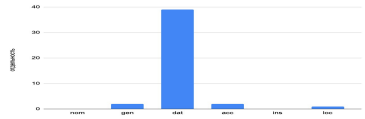
В корпусе встречаются сочетания: случайные (*слово в театр*), частотные свободные сочетания (*слово в предложении*), устойчивые обороты (*слово в защиту*), идиомы (*слово в слово*)

Категория	**Нормализованная дивергенция**	**Значение категории**	**Weirdness**
Одушевленность	0,42	inan	0,74
Падеж	0,39	acc	0,13
Род	0,36	n	0,12
Число	0,20	падеж: loc	0,028

Ни одно из грамматических значений не является достаточно стабильным, чтобы выделить коллигацию, например, *слово в + S.inan.acc*. Но одушевленность лучший кандидат, чем р

Профили поведения языковых единиц

Еще пример: слово в ___ – смотрим на лексику

Лемма	Weirdness	Грамматический профиль	Weirdness с учетом формы inan.acc
слово	151,55		96,16
адрес	232,44		76,47
отдельность	**1183,06**		не встретилось

https://cococo.cosyco.ru - Collocations, Colligations, Corpora

* ***отдельность*** побеждает, но за счет своей редкости в корпусе

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.

contest_NOUN

☑ High ☑ Medium ☐ Low

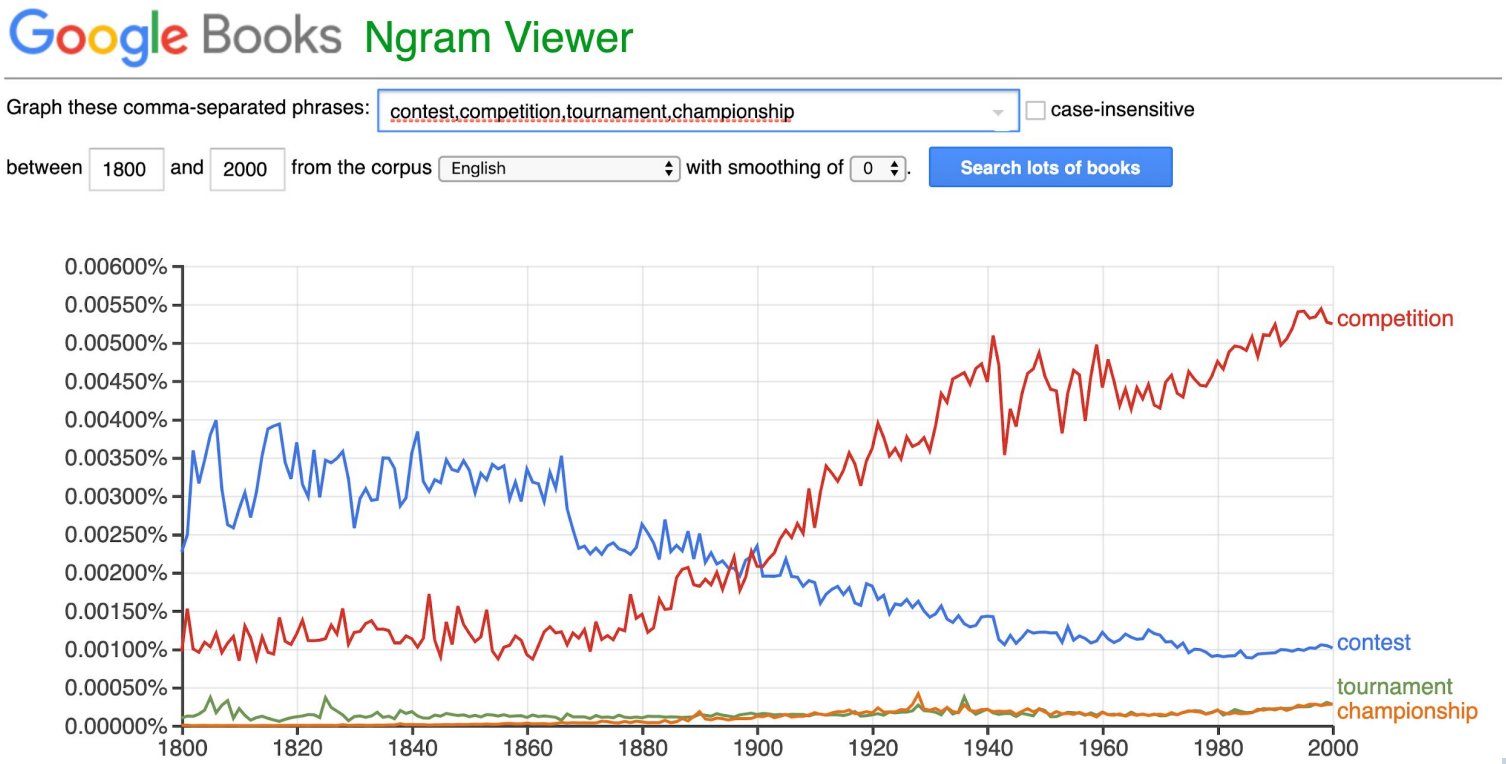
English Wikipedia

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

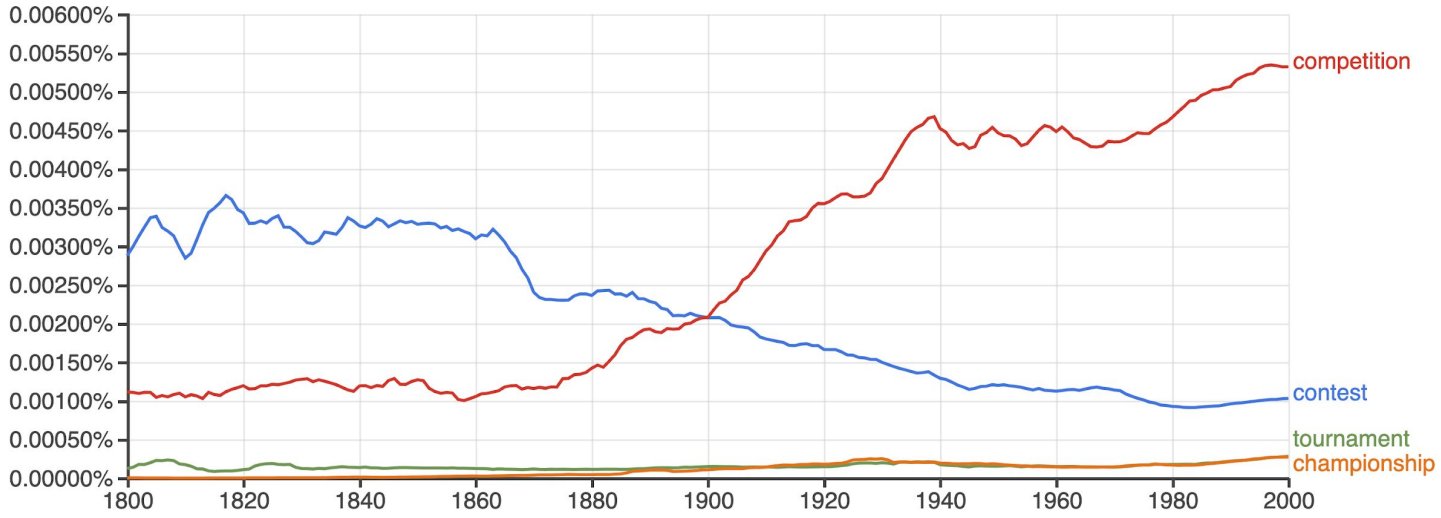


Распределение слов по времени

Google Books Ngram Viewer

Graph these comma-separated phrases: contest,competition,tournament,championship case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3. Search lots of books

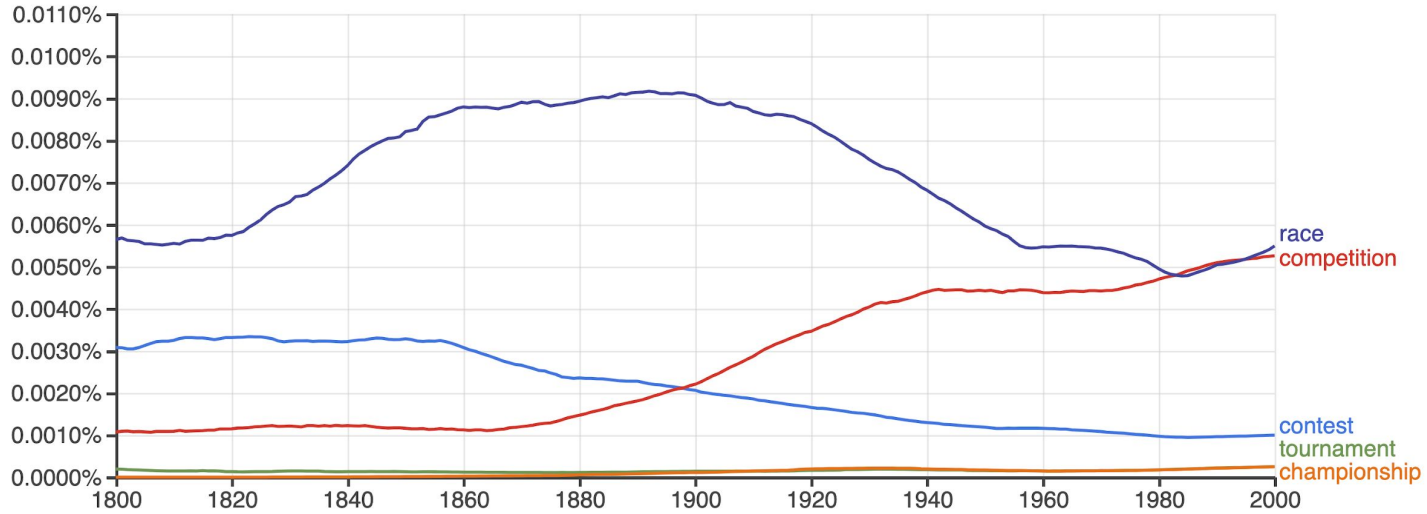


Распределение слов по времени

Google Books Ngram Viewer

Graph these comma-separated phrases: contest,competition,tournament,championship,race ☐ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 10. Search lots of books



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