

Лекция 4

Работа с собственным корпусом

Частотные списки, коллокации, ключевые слова

Ольга Ляшевская ** olesar@yandex.ru

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

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!



Обработка данных для корпуса

Стандартная

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

Любой этап может быть пропущен, в зависимости от нужд исследования



Need more corpus please!



Частотные списки

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

Л.Н.Толстой, Анна Каренина

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 выходявшей	2
6 на	3584	104 раз	271	1004 ног	27	10004 выходу	2
...		

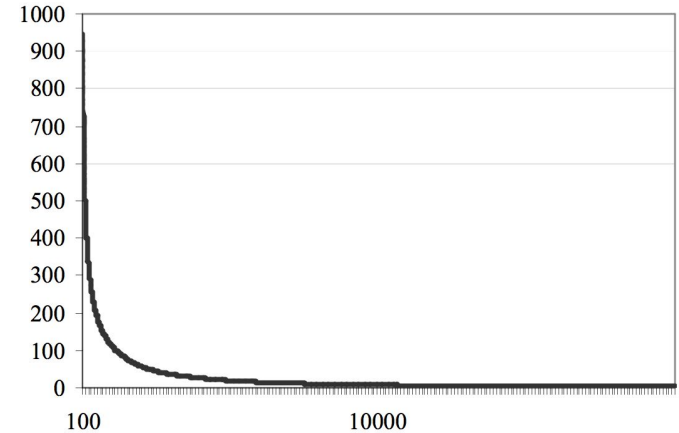


Закон Ципфа

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.

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

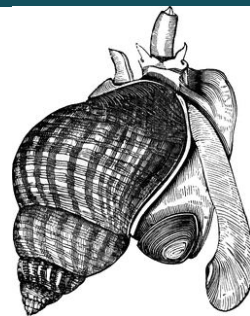


Кстати, на материале больших веб-корпусов этот закон выполняется примерно для половины слов. Для морфологически богатых языков (ср. также словоформы - леммы) скорость убывает иначе. γ - поправка Бенуа Мандельброта (1965) к закону Джорджа Кингсли Ципфа (1949). Он выделил голову (стоп-слова), среднюю часть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



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



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

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



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,

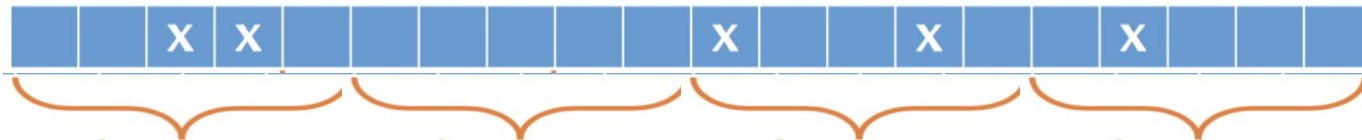
$f_D - f_d$ – количество анализируемой словоформы в контрастном подкорпусе, кроме количества словоформы в документе, в которую входит анализируемая единица⁹.

Меры дистрибуции появления единицы

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

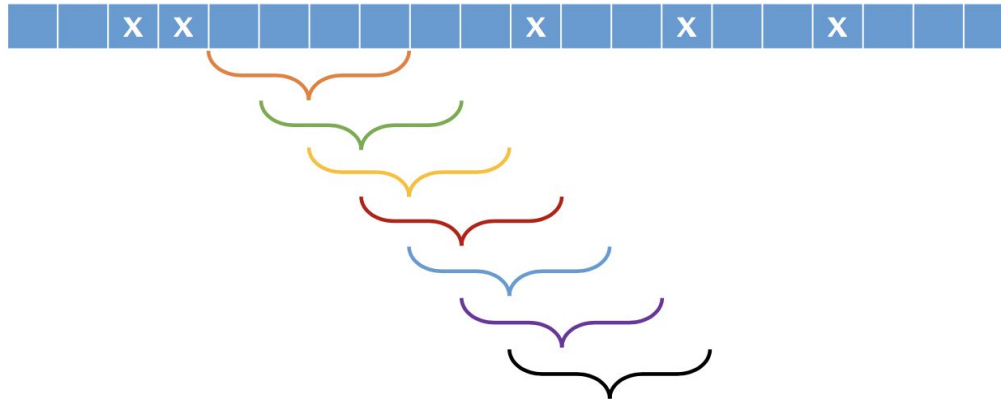
$$D = 100 \times \left(1 - \frac{\sigma}{\bar{v}\sqrt{n}}\right), \text{ где } \sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (v_i - \bar{v})^2} ; U = fD \text{ (D модифиц.)}$$

- 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



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-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

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-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 коллокации

Сила связности коллокаций: насколько коллокации не случайны?

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

- взаимная информация (MI, PMI, MI³): $MI(n,c) = \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}$

- 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:

$$\log\text{-likelihood} = 2 \sum_{ij} O_{ij} \times \log \frac{O_{ij}}{E_{ij}}$$

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



Ресурсы и литература

- **AntConc** и его производные (для [Windows, MacOS, Linux](#))
- 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](#)

