МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ

Урок №3

Лингвистика

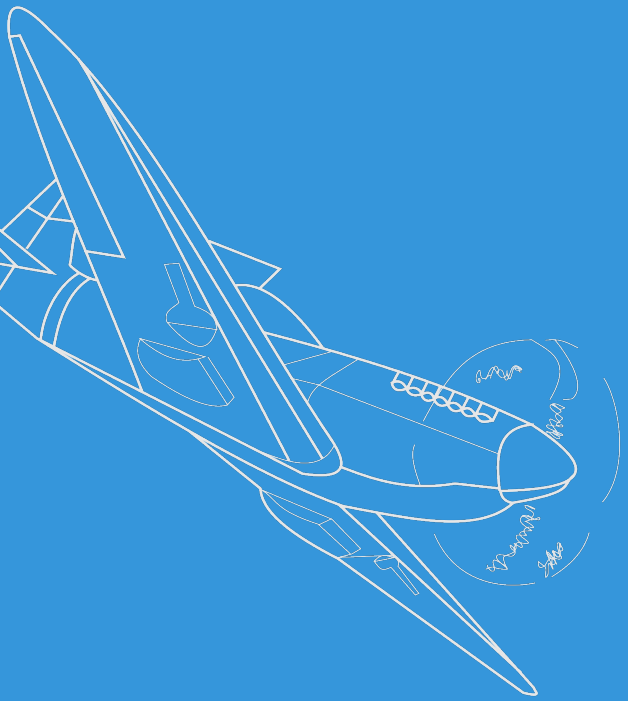
на которой расскажут про историю лингвистики, основные термины, закон Ципфа, основы обработки текста

16 октября 2023 года

Антон Кухтичев

Содержание занятия

1. Что такое лингвистика
2. История лингвистики
3. Разделы лингвистики
4. Подходы к языку
5. Морфология
6. Корпусная лингвистика
7. Закон Ципфа
8. Основы обработки текста



Что такое лингвистика?

Что такое лингвистика

Лингвистика (языкознание, языковедение; от лат. lingua — язык) — наука, изучающая языки. Это наука о естественном человеческом языке вообще и обо всех языках мира как индивидуальных его представителях.

Задачи лингвистики

- Как именно люди говорят?
- Как связаны человеческие фразы с окружающим миром?

История лингвистики

Пра-пра-лингвисты – кто они?

- **Индия**
  - Панини (IV в. до н. э.)
  - Яска (IV в. до н. э.)

История лингвистики

Пра-пра-лингвисты – кто они?

- **Китай**
  - Сю Шэнь (I в. н. э.)

История лингвистики

Пра-пра-лингвисты – кто они?

- **Греция**
  - Аристарх (II в. до н. э.)
  - Дионисий Фракийский (II в. до н. э.)
  - Аполлоний Дискол (II в. до н. э.)

История лингвистики

Пра-пра-лингвисты – кто они?

- **Рим**
  - Марк Теренций Варрон (I в. до н. э.)
  - Донат (III—IV в. н. э.)
  - Присциан (VI в. н. э.)

История лингвистики

Пра-пра-лингвисты – кто они?

- **Аравия**
  - Сибавейхи (VII в. н. э.)
  - Ибн Джинни (конец X — начало XI в.)

История лингвистики

Средние века и Новое время

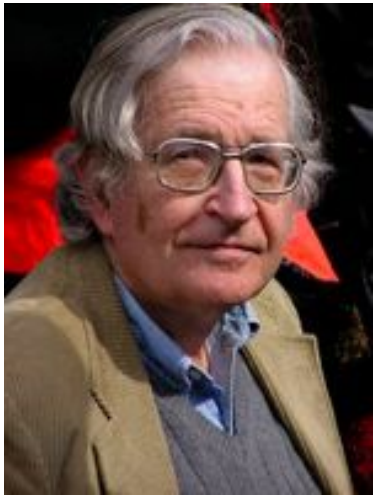
- Грамматика Пор-Рояля
- «Идеальный» язык
- Вильгельм фон Гумбольдт (1767—1835)

История лингвистики

- Фердинанд де Соссюр
  - «Курс общей лингвистики» (1916)
  - Язык vs речь
  - Структурная лингвистика

История лингвистики

- Ноам Хомский
  - «Синтаксические структуры» (1957)
  - Грамматические принципы, лежащие в основе языков, являются врождёнными и неизменными

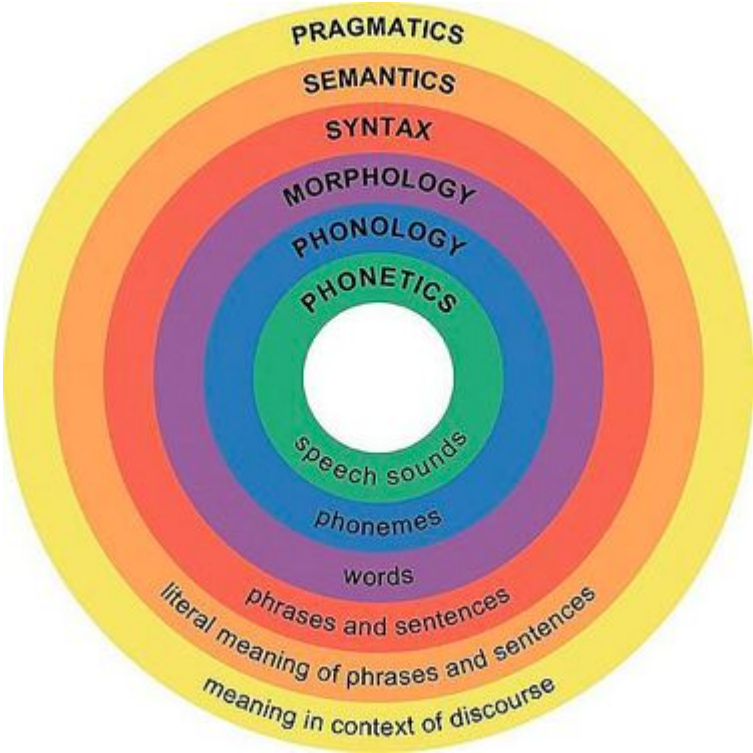


Чем занимаются лингвисты?

- Общая лингвистика
- Историческая лингвистика
- Лингвистическая типология
- Социолингвистика
- Диалектология
- Лексикография
- Психолингвистика
- Математическая лингвистика

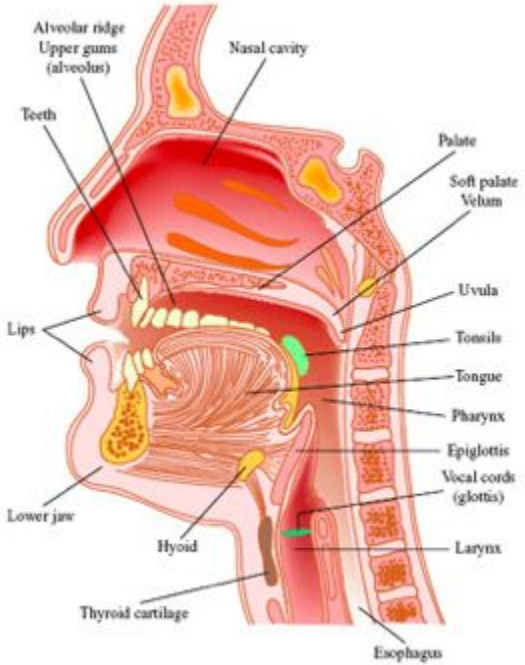
Чем занимаются лингвисты?

- Общая лингвистика
  - Фонетика
  - Фонология
  - Морфология
  - Синтаксис
  - Семантика
  - Прагматика



Общая лингвистика / Фонетика

- Звуки речи
  - с точки зрения его создания
  - как колебание воздуха
  - как объект восприятия



Общая лингвистика / Фонология

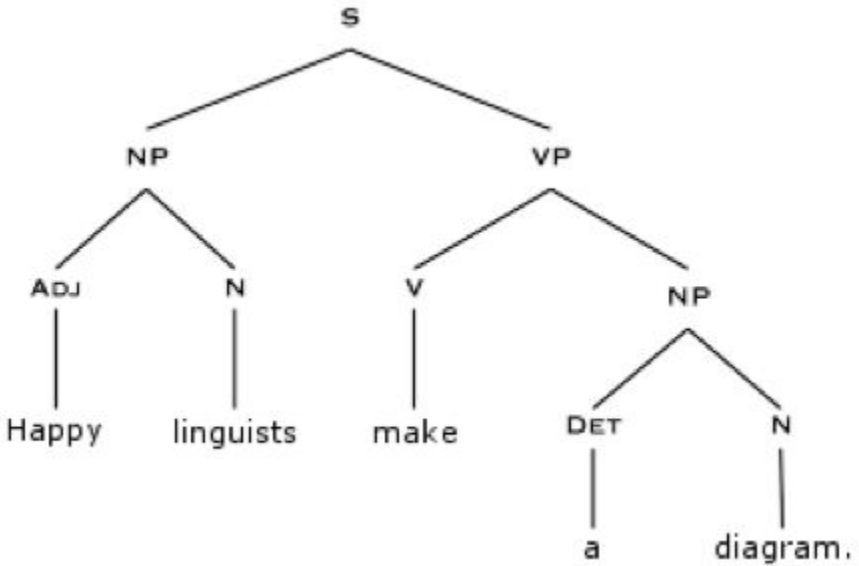
- Звуки, составляющие речь
- Фонема – минимальная смыслоразличительная единица языка
- Слог
- Транскрипция

Общая лингвистика / Морфология

- Внутренняя структура отдельных слов
  - Словообразование (лексическая морфология)
  - Словоизменение
- Морфемы

Общая лингвистика / Синтаксис

- Структура предложений и связи между словами



Неоднозначность синтаксического разбора

- Time flies like an arrow.
- Души прекрасные порывы.
- Эти типы стали есть в цехе.

Общая лингвистика / Семантика

- Значение отдельных слов и текстов

Общая лингвистика / Прагматика

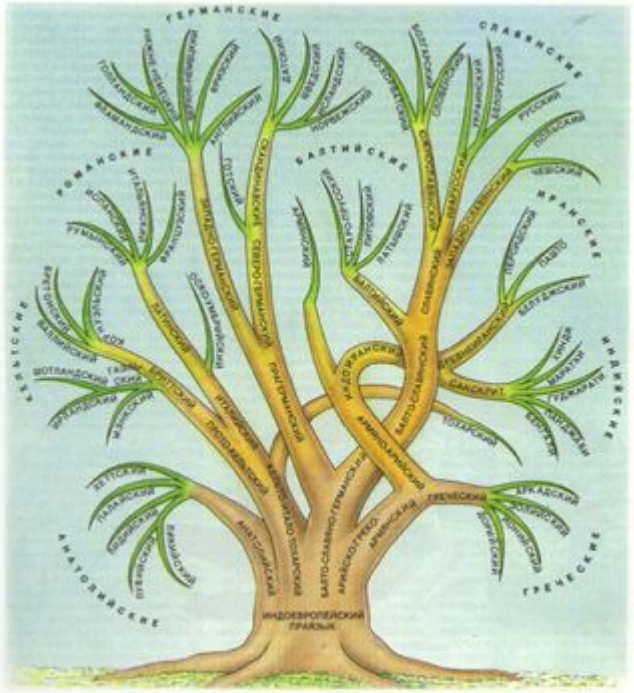
- Употребление языка в различных ситуациях

Историческая лингвистика

- История
- Сравнение

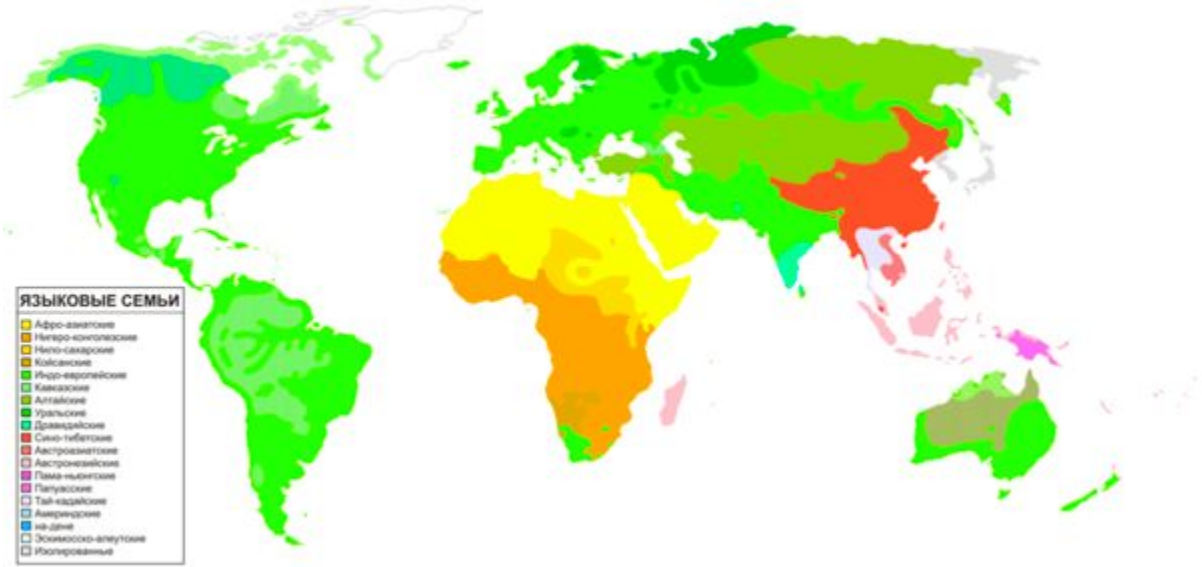
Spanish	English
constitución	constitution
revolución	revolution
investigación	?

- Диахрония/синхрония



Лингвистическая типология

- Сравнение языков, выделение характерных признаков



Социолингвистика

- Связь между языком и социальными условиями

Диалектология

Лексикография

- Составление словарей
- Есть еще лексикология, этимология…

ПОЛНЫЙ

ЦЕРКОВНО-СЛАВЯНСКІЙ

СЛОВАРЬ

(со внесеніемъ въ него важнѣйшихъ древне-русскихъ словъ и выраженій),

ПОСОБІЕ

1) для преподавателей русск. и ц.-слав. языка въ низшихъ и среднихъ учебныхъ заведеніяхъ; 2) для занимающихся изученіемъ русскихъ древностей, филологическими разысканіями въ области исторіи и этимологіи родного языка и т. п. работами; 3) для пастырей церкви, какъ совершителей богослуженія, законоучителей, проповѣдниковъ и миссіонеровъ и 4) для всѣхъ, желающихъ стать въ сознательно-разумныя отношенія какъ къ языку матери-церкви, такъ и къ родному слову въ его современномъ состояніи и историческихъ судьбахъ.

СОСТАВИЛЪ

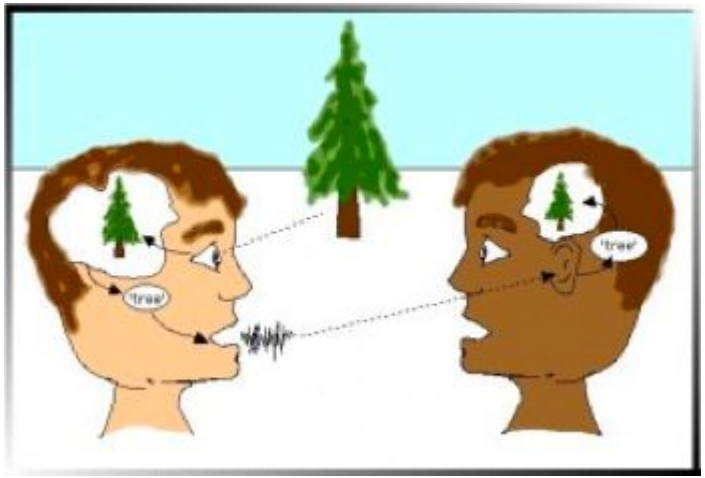
Священникъ магистръ Григорій Дьяченко

(бывшій преподаватель русскаго языка и словесности).

Всѣхъ словъ объяснено около 30,000.

Психолингвистика

- Связь языка, мышления и сознания



Математическая лингвистика

В узком смысле – теория порождающих грамматик

- Порождающая грамматика (generative grammar)

В широком смысле

- Количественная лингвистика
- Статистическая лингвистика
- Прикладная лингвистика

Статистическая лингвистика

- Анализ древних языков
- Снятие омонимии
- Машинный перевод

Подходы к языку

- Рационалистический
  - Значительная часть знаний заложена внутрь человека (например, наследуется) и не порождается из ощущений.
- Эмпирический
  - Человеческий мозг имеет некие способности к обобщению.
  - На основе этих общих способностей дети усваивают язык на примерах.

Рационализм

- Порождающая грамматика языка.
  - Любая корректная фраза порождается этой грамматикой, некорректная – нет.
- Больше интереса к «внутреннему языку», все его проявления снаружи – косвенные
- Важна интуиция носителей языка.
- Маргинальный аналог в ИИ: максимальное количество ручных правил, заложенных внутрь системы.

Эмпиризм

- Простая общая языковая модель.
- Параметры модели настраиваются на основе выбранного корпуса.
  - Статистика
  - Распознавание закономерностей
  - Машинное обучение
- Внимание сосредоточено именно на внешних проявлениях языка.

Здравый смысл

- Комбинация обоих подходов.
- До какой-то степени детализации идти рационально.
- Но когда деталей становится слишком много, применять эмпиризм.
- Пороги и подобные значения настраивать автоматически.

Морфология

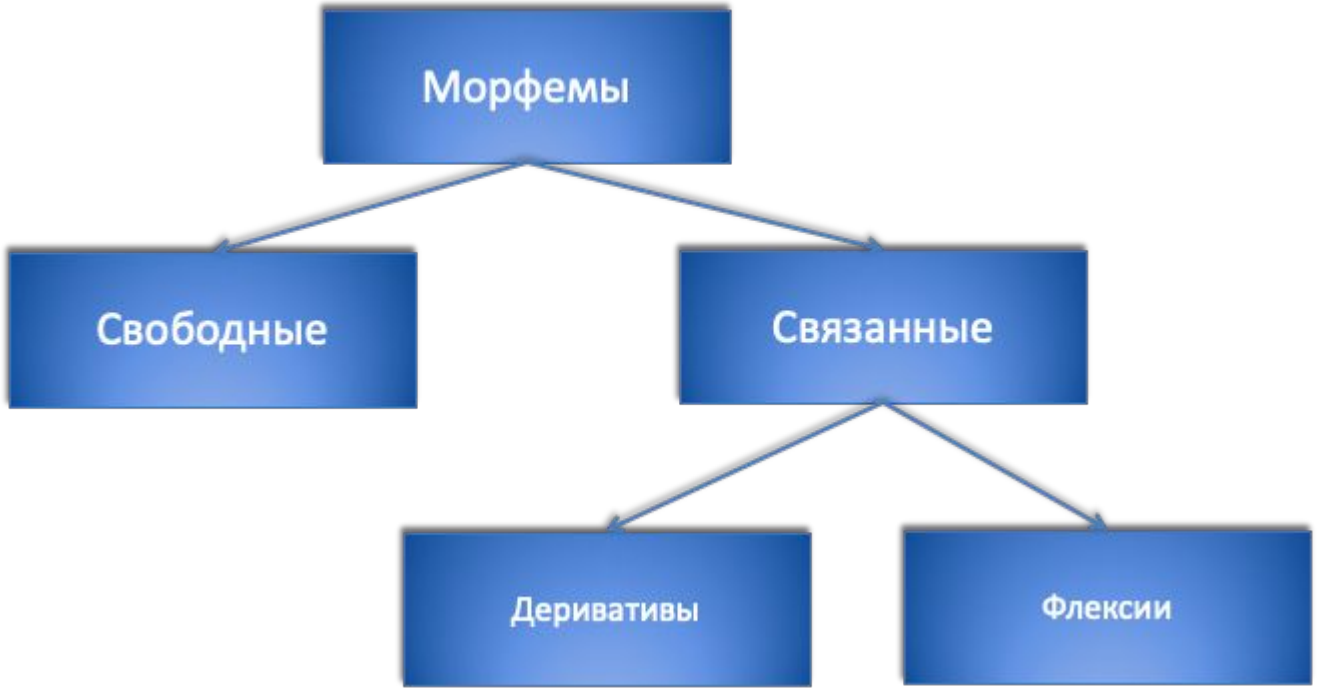
«Морфология есть часть лингвистики, занимающаяся словом во всех его релевантных аспектах» [Мельчук 1997]

Морфология

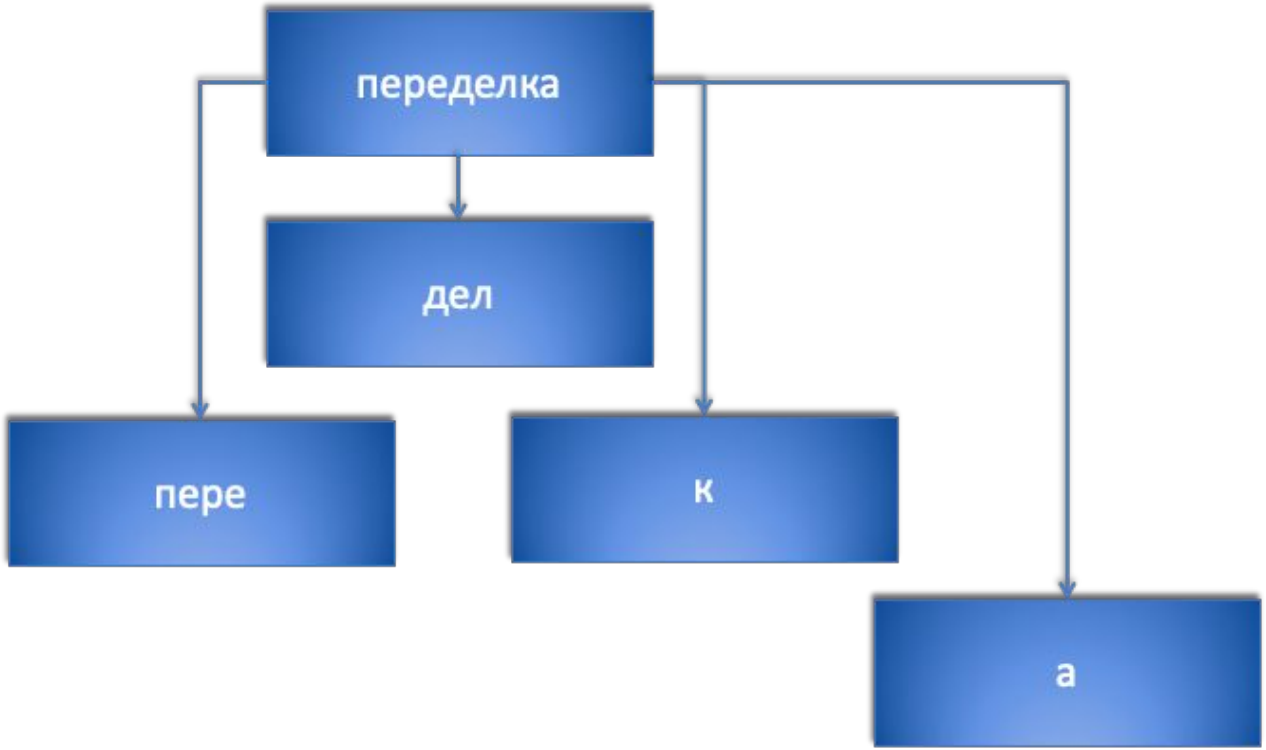
«Морфология есть часть лингвистики, занимающаяся словом во всех его релевантных аспектах» [Мельчук 1997]

Предмет морфологии - описание свойств слова и его (значащих) частей

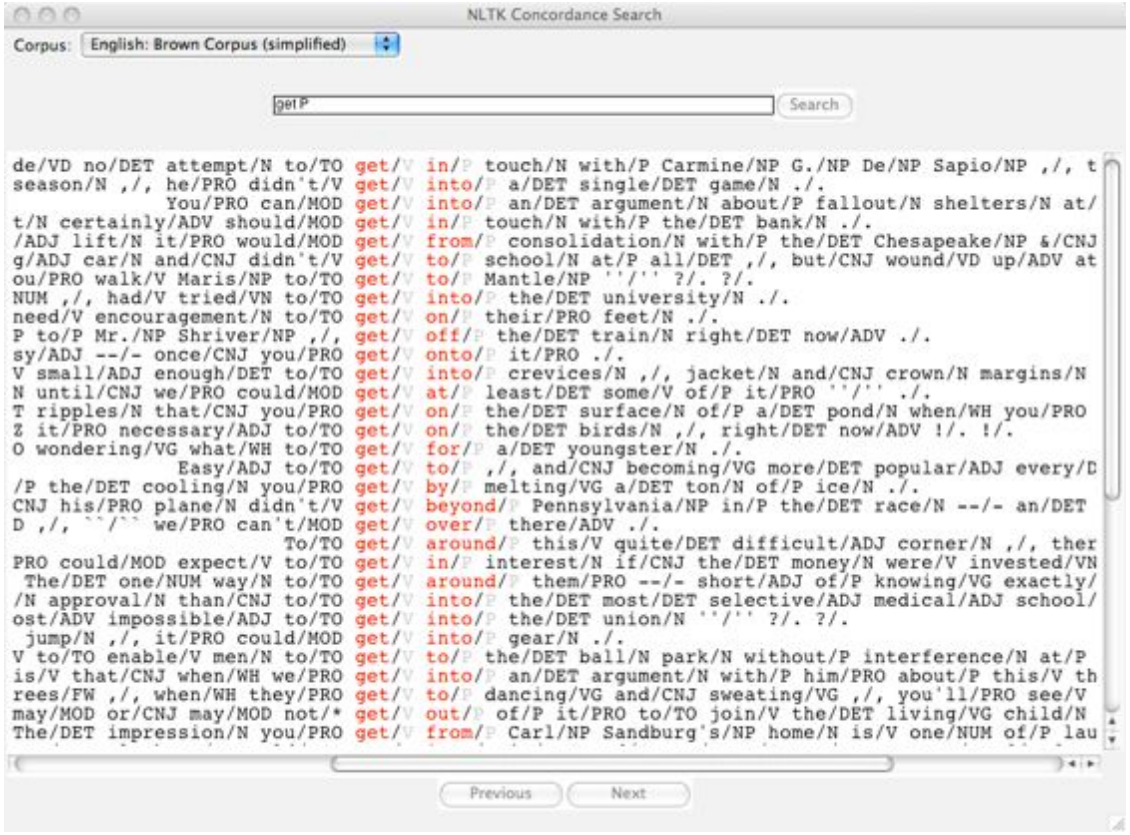
Морфология

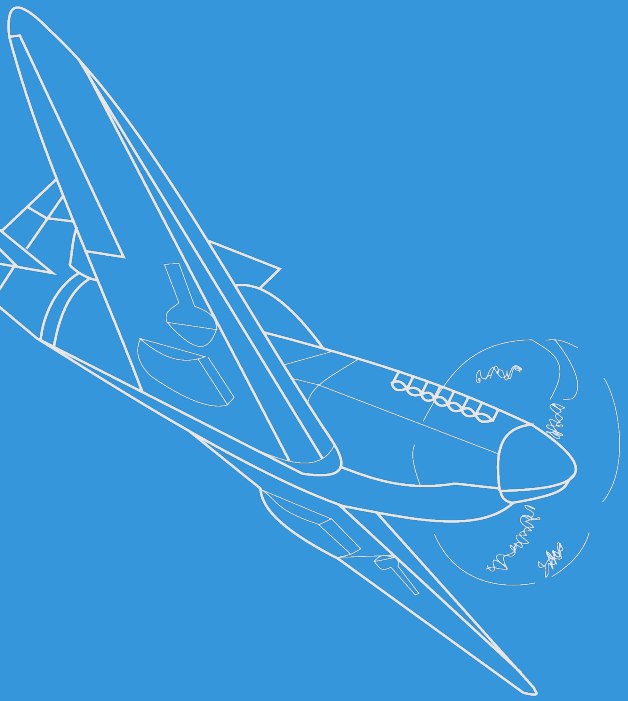


Морфология



Конкорданс


NLTK Concordance Search
Corpus: English: Brown Corpus (simplified)
get P
Search
de/VD no/DET attempt/N to/TO get/V in/P touch/N with/P Carmine/NP G./NP De/NP Sapio/NP ,/, t
season/N ,/, he/PRO didn't/V get/V into/P a/DET single/DET game/N ./.
You/PRO can/MOD get/V into/P an/DET argument/N about/P fallout/N shelters/N at/
t/N certainly/ADV should/MOD get/V in/P touch/N with/P the/DET bank/N ./.
/ADJ lift/N it/PRO would/MOD get/V from/P consolidation/N with/P the/DET Chesapeake/NP &/CNJ
g/ADJ car/N and/CNJ didn't/V get/V to/P school/N at/P all/DET ,/, but/CNJ wound/VD up/ADV at
ou/PRO walk/V Maris/NP to/TO get/V to/P Mantle/NP ''/'' ?/. ?/.
NUM ,/, had/V tried/VN to/TO get/V into/P the/DET university/N ./.
need/V encouragement/N to/TO get/V on/P their/PRO feet/N ./.
P to/P Mr./NP Shriver/NP ,/, get/V off/P the/DET train/N right/DET now/ADV ./.
sy/ADJ --/- once/CNJ you/PRO get/V onto/P it/PRO ./.
V small/ADJ enough/DET to/TO get/V into/P crevices/N ,/, jacket/N and/CNJ crown/N margins/N
N until/CNJ we/PRO could/MOD get/V at/P least/DET some/V of/P it/PRO ''/'' ./.
T ripples/N that/CNJ you/PRO get/V on/P the/DET surface/N of/P a/DET pond/N when/WH you/PRO
Z it/PRO necessary/ADJ to/TO get/V on/P the/DET birds/N ,/, right/DET now/ADV !/. !/.
O wondering/VG what/WH to/TO get/V for/P a/DET youngster/N ./.
Easy/ADJ to/TO get/V to/P ,/, and/CNJ becoming/VG more/DET popular/ADJ every/D
/P the/DET cooling/N you/PRO get/V by/P melting/VG a/DET ton/N of/P ice/N ./.
CNJ his/PRO plane/N didn't/V get/V beyond/P Pennsylvania/NP in/P the/DET race/N --/- an/DET
D ,/, ``/`` we/PRO can't/MOD get/V over/P there/ADV ./.
To/TO get/V around/P this/V quite/DET difficult/ADJ corner/N ,/, ther
PRO could/MOD expect/V to/TO get/V in/P interest/N if/CNJ the/DET money/N were/V invested/VN
The/DET one/NUM way/N to/TO get/V around/P them/PRO --/- short/ADJ of/P knowing/VG exactly/
/N approval/N than/CNJ to/TO get/V into/P the/DET most/DET selective/ADJ medical/ADJ school/
ost/ADV impossible/ADJ to/TO get/V into/P the/DET union/N ''/'' ?/. ?/.
jump/N ,/, it/PRO could/MOD get/V into/P gear/N ./.
V to/TO enable/V men/N to/TO get/V to/P the/DET ball/N park/N without/P interference/N at/P
is/V that/CNJ when/WH we/PRO get/V into/P an/DET argument/N with/P him/PRO about/P this/V th
rees/FW ,/, when/WH they/PRO get/V to/P dancing/VG and/CNJ sweating/VG ,/, you'll/PRO see/V
may/MOD or/CNJ may/MOD not/* get/V out/P of/P it/PRO to/TO join/V the/DET living/VG child/N
The/DET impression/N you/PRO get/V from/P Carl/NP Sandburg's/NP home/N is/V one/NUM of/P lau
Previous
Next



Закон Ципфа

Закон Ципфа

Частота слова (f) обратно пропорциональна его положению (r) в отсортированном по частотности списке слов

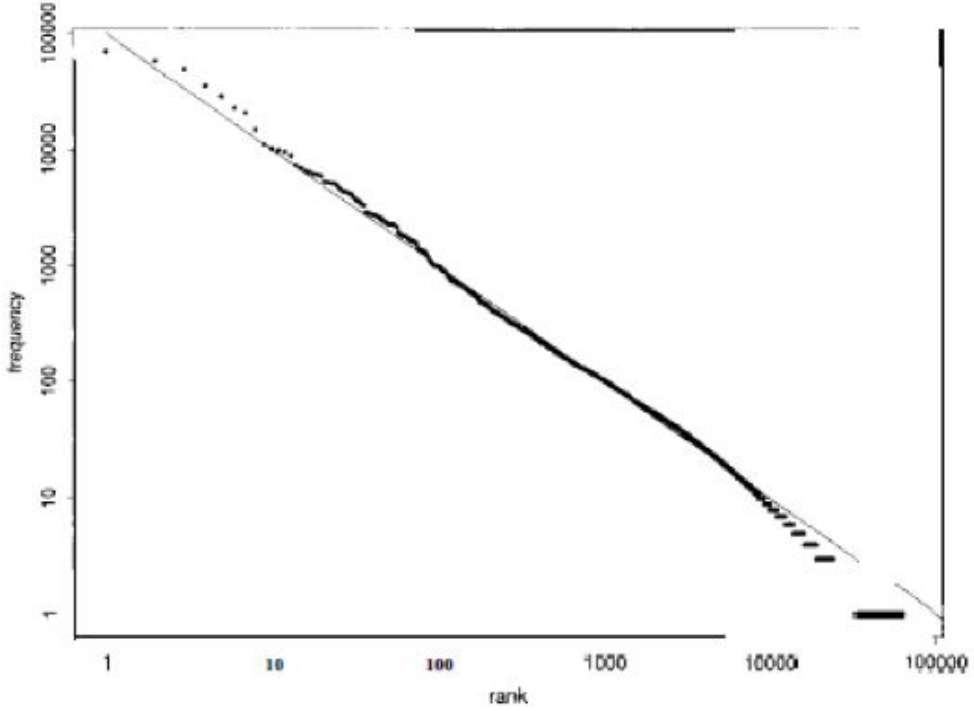
или существует такая константа k, что

$$f \cdot r = k$$

Закон Ципфа для Тома Сойера

Word	Freq. (f)	Rank (r)	$f \cdot r$	Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

Закон Ципфа, Brown Corpus



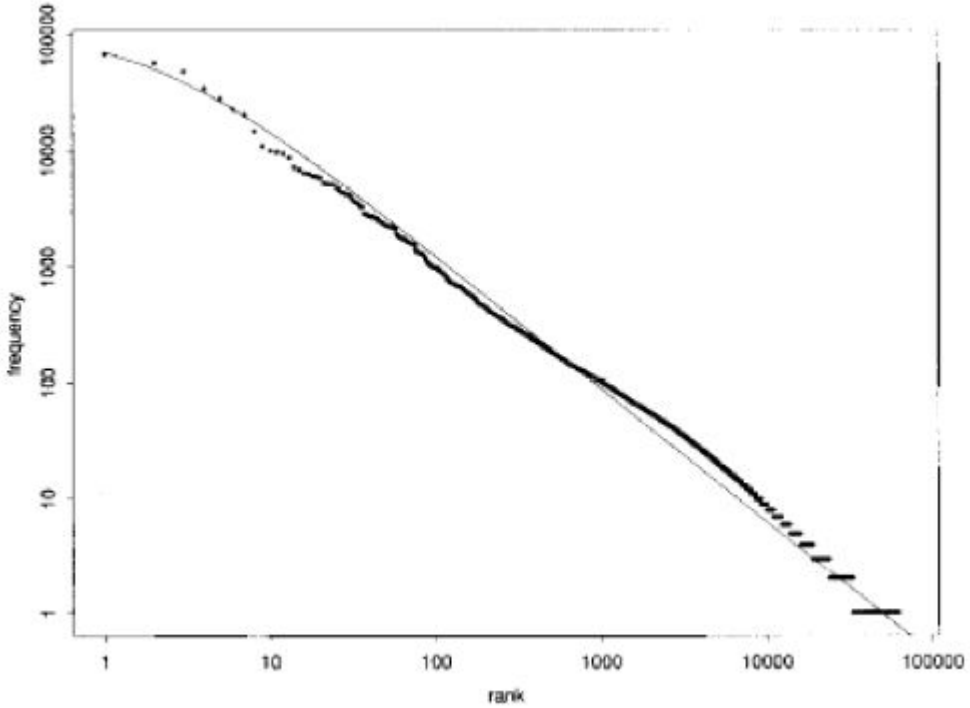
Поправки Мандельброта

- Закон Ципфа одинаков для всех слов, но корректно предсказывает только «средние» термины.
- Он ошибается в низкочастотной и высокочастотных частях словаря
- Более точная формула:

$$f = P \cdot (r + \rho)^{-B}$$

$$\log f = \log P - B \cdot \log (r + \rho)$$

Формула Мандельброта



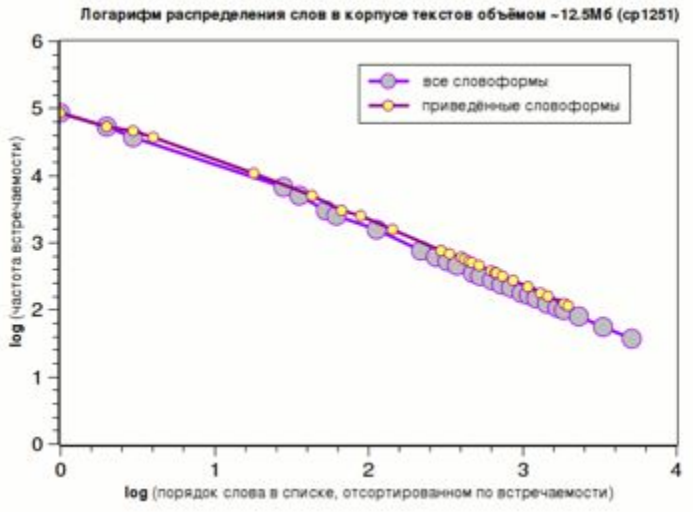
$P = 10^{5,4}$

$B = 1,15$

$\rho = 100$

Главный вывод

- Частотность слов не описывается нормальным распределением
- Лучше подходит гиперболическое распределение
    - Степенной закон



Другие законы Ципфа (1)

- Количество смыслов у слова:
    - Слушающий: один смысл – одно слово
    - Говорящий: больше смыслов у одного слова (уменьшает количество запоминаемых слов)
- Количество смыслов m подчиняется закону:

$$m \propto \sqrt{f}$$

Другие законы Ципфа (2)

- Количество строк или страниц текста между вхождениями одного и того же слова
- Если F – частота интервала и I – размер интервала, то:

$$F \propto I^{-p}$$

- p колеблется от 1 до 1,3.
- То есть, одинаковые слова чаще всего встречаются рядом.

Другие законы Ципфа (3)

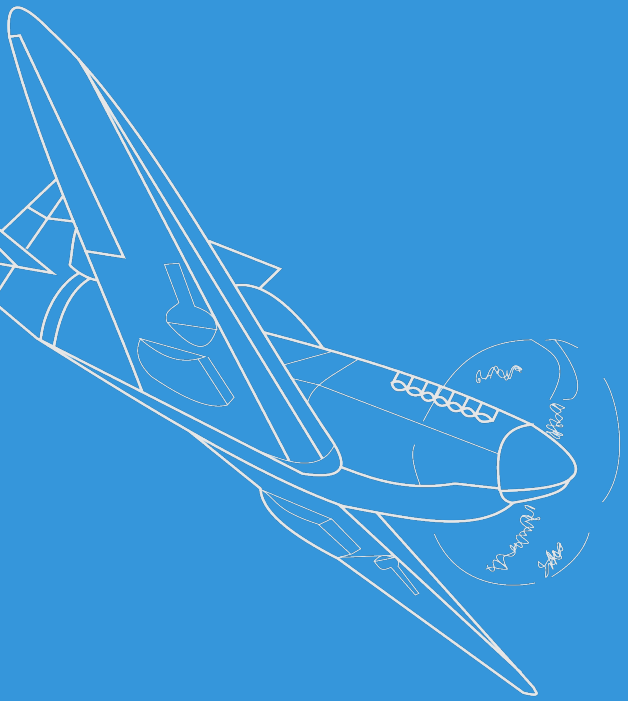
- Обратная зависимость между частотностью слова и его длиной

Важность законов Ципфа

- Если сгенерировать текст случайно, то этот текст будет удовлетворять закону Ципфа
- Вероятность слова длины *n* для английского алфавита:

$$\left(\frac{26}{27}\right)^n \cdot \frac{1}{27}$$

- С этой точки зрения их важность для описания естественного языка – невысока.
- Но! Как демонстрация степенного закона (подавляющее число слов редкие) – их роль очень важна.



Основы обработки текста

Терминология

Токен – экземпляр последовательности символов в документе, объединенных в семантическую единицу для обработки.

Термин – «нормализованный» токен (регистр, морфология, исправленные ошибки и т.п.)

Нормализация

- Необходимо «нормализовывать» термины как в индексируемом тексте, так и в запросе.
- Например: желательно считать одинаковыми термины U.S.S.R и USSR
- Обычно термины объединяются в классы эквивалентности.
- Можно поступать наоборот, расширять:
  - window ➔ window, windows
  - windows ➔ Windows, windows
  - Windows (нет расширения)
- Такой подход гибкий, но более ресурсоёмкий.

Документы

Мы неявно предполагаем:

- Мы знаем, что такое документ.
- Каждый документ доступен для автоматического разбора.

На самом деле, здесь может быть много проблем.

Лингвистика при обработке документов

- Определение формата документа (pdf, word, html и т.д.)
- Определение кодировки документа
- Определение языка документа
- Токенизация и сегментация
- Нормализация и лемматизация;
- Выделение объектов и зон
- Вычисление текстовых факторов

Нормализация

Нормализация зависит от языка документа.

- PETER WILL NICHT MIT. ➔ MIT = mit
- He got his PhD from MIT. ➔ MIT ≠ mit

Нормализация. Ударения и диакритика

- résumé vs. resume
- Умуляуты: Universität vs. Universitaet (заменяем на специальную последовательность «ae» или даже «æ»)

- Самый важный вопрос: как пользователи предпочитают писать запросы с этими словами?

Нормализация. Классы эквивалентности

- Soundex
  - фонетическая эквивалентность,

    Muller = Mueller

- Тезаурус
  - семантическая эквивалентность,

    car = automobile

Понижение регистра

- Понизить регистр всех букв.
- Возможны исключения, например, для
- капитализированных слов внутри предложения.
  - MIT и mit
  - Fed и fed
  - КОТ и кот (Калининградская областная таможня)
- NB: немецкий ➜ существительные с большой буквы
- Часто лучше понижать всё, потому что пользователи не заботятся о капитализации в запросах.

Проблемы токенизации: одно слово или два?

- Hewlett-Packard
- State-of-the-art
- co-education
- the hold-him-back-and-drag-him-away maneuver
- data base
- San Francisco
- Los Angeles-based company
- cheap San Francisco-Los Angeles fares
- York University vs. New York University

Числа

- 3/20/91
- 20/3/91
- Mar 20, 1991
- B-52
- 100.2.86.144
- (800) 234-2333
- 800.234.2333

Обработка запроса

- Запросы задают не по-русски
- Распознавание языка
- Исправление опечаток
- Токенизация
- Нормализация и лемматизация
- Кореференция (расширение запроса)
- Переформулировки запросов
- Сегментация запроса
- Извлечение объектов

Китайский: нет пробелов

李克强说，当前国际和地区形势发生复杂深刻变化，中越都处于发展的关键阶段，双方要从战略高度和长远角度出发，在发展中越关系十六字方针和“四好”精神指 引下，坚定不移推进中越友好。中方愿同越方保持高层战略沟通，加强治国理政经验交流，坚持经济优先、民生优先，深化务实合作，推动中越全面战略合作伙伴关 系迈上新台阶。

(c) news.xinhuanet.com

Китайский: нет пробелов

Или даже так

(c) Baidu.com

Китайский: нет пробелов

文字

Эти два иероглифа могут трактоваться как одно слово «письменность»

или как последовательность двух слов «культура» и «слово»

Другие случаи отсутствия пробелов

- Компаунды в датском, немецком, шведском, финском.
  - Computerlinguistik ➔ Computer + Linguistik
  - Lebensversicherungsgesellschaftsangestellter ➔ leben + versicherung + gesellschaft + angestellter - служащий компании страхования жизни.
  - Kallistuksenvaimennusjärjestelmä - система, предотвращающая крен (в погрузчиках).
- Льезоны в романских языках
  - em os ➔ nos
  - por a ➔ pela
- Эскимосы: tusaatsiarunnanngittualuujunga (Я не очень хорошо слышу)
- Таких языков довольно много.

Японский

4 разных алфавита.

ローマ字 Romaji

平仮名 Hiragana

片仮名 Katakana

漢字 Kanji

Запрос может быть сформулирован в любом из них.

Named Entity Recognition

Извлечение объектов (группа слов/токенов в запросе, которые обозначают одно понятие)

- ФИО
- Телефоны
- Адреса
- Даты
- Названия песен, фильмов, книг и т.д.

Кодировки

- ASCII (ISO 646) – 7-битовый стандарт
- ISO 8859
  - 8859-1, или ISO Latin-1
  - ISO 8859-5
- Русские кодировки
  - cp1251 aka Windows
  - 866 aka DOS
  - KOI-8
- Unicode
- UTF-8
- UTF-16

Кодировки. ASCII

- American standard code for information interchange (1967)

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Кодировки. koi8-r

- код обмена информацией, 8 бит

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
80	2500 ─	2502 │	250C ┌	2510 ┐	2514 └	2518 ┘	251C ├	2524 ┤	252C ┬	2534 ┴	253C ┼	2580 ▀	2584 ▄	2588 █	258C ▌	2590 ▐
90	2591 ░	2592 ▒	2593 ▓	2320 ⌠	25A0 ■	2219 ∙	221A √	2248 ≈	2264 ≤	2265 ≥	A0	2321 ⌡	B0 °	B2 ²	B7 ·	F7 ÷
A0	2550 ═	2551 ║	2552 ╒	451 ё	2553 ╓	2554 ╔	2555 ╕	2556 ╖	2557 ╗	2558 ╘	2559 ╙	255A ╚	255B ╛	255C ╜	255D ╝	255E ╞
B0	255F ╟	2560 ╠	2561 ╡	401 Ё	2562 ╢	2563 ╣	2564 ╤	2565 ╥	2566 ╦	2567 ╧	2568 ╨	2569 ╩	256A ╪	256B ╫	256C ╬	A9 ©
C0	44E ю	430 а	431 б	446 ц	434 д	435 е	444 ф	433 г	445 х	438 и	439 й	43A к	43B л	43C м	43D н	43E о
D0	43F п	44F я	440 р	441 с	442 т	443 у	436 ж	432 в	44C ь	44B ы	437 з	448 ш	44D э	449 щ	447 ч	44A ъ
E0	42E Ю	410 А	411 Б	426 Ц	414 Д	415 Е	424 Ф	413 Г	425 Х	418 И	419 Й	41A К	41B Л	41C М	41D Н	41E О
F0	41F П	42F Я	420 Р	421 С	422 Т	423 У	416 Ж	412 В	42C Ь	42B Ы	417 З	428 Ш	42D Э	429 Щ	427 Ч	42A Ъ

Кодировки. cp866

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
8.	А 410	Б 411	В 412	Г 413	Д 414	Е 415	Ж 416	З 417	И 418	Й 419	К 41A	Л 41B	М 41C	Н 41D	О 41E	П 41F
9.	Р 420	С 421	Т 422	У 423	Ф 424	Х 425	Ц 426	Ч 427	Ш 428	Щ 429	Ъ 42A	Ы 42B	Ь 42C	Э 42D	Ю 42E	Я 42F
A.	а 430	б 431	в 432	г 433	д 434	е 435	ж 436	з 437	и 438	й 439	к 43A	л 43B	м 43C	н 43D	о 43E	п 43F
B.	▒ 2591	▓ 2592	▩ 2593	┃ 2502	┤ 2524	╡ 2561	╢ 2562	╖ 2556	╕ 2555	╣ 2563	║ 2551	╗ 2557	╝ 255D	╜ 255C	╛ 255B	┐ 2510
C.	└ 2514	┴ 2534	┬ 252C	├ 251C	─ 2500	┼ 253C	╞ 255E	╟ 255F	╚ 255A	╔ 2554	╩ 2569	╦ 2566	╠ 2560	═ 2550	╬ 256C	╧ 2567
D.	╨ 2568	╤ 2564	╥ 2565	╙ 2559	╘ 2558	╒ 2552	╓ 2553	╫ 256B	╪ 256A	┘ 2518	┌ 250C	■ 2588	▬ 2584	▌ 258C	▐ 2590	▀ 2580
E.	р 440	с 441	т 442	у 443	ф 444	х 445	ц 446	ч 447	ш 448	щ 449	ъ 44A	ы 44B	ь 44C	э 44D	ю 44E	я 44F
F.	Ё 401	ё 451	Є 404	є 454	Ї 407	ї 457	Ў 40E	ў 45E	° B0	∙ 2219	· B7	√ 221A	№ 2116	¤ A4	■ 25A0	 A0

Кодовое пространство Unicode

- Обозначения: U+xxxx, U+xxxxx, U+xxxxxx
- Пространство разделено на 17 плоскостей по 2^{16} символов
- Первые 128 символов совпадают с ASCII
- Плоскость 0 (base multilingual plane) содержим основные символы
- Остальные плоскости содержат символы редких письменностей
- 2048 кодов U+DC00 - U+DFFF заняты под “суррогатные пары”

Всего символов в Unicode

$$17 * 2^{16} - 2048 = 1\,112\,064$$

Кодировки. UTF-8

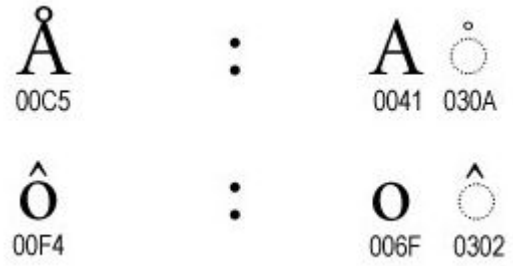
- Нужен для передачи Unicode по однобайтовым каналам связи
- Начало аналогично первой половине таблицы ASCII
- Обладает свойством самосинхронизации
- Мультибайтная кодировка

```
0xxxxxxx
110xxxxx 10xxxxxx
1110xxxx 10xxxxxx 10xxxxxx
11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
```

Кодировки. Объединение и дублирование символов

Для многих символов в Unicode есть отдельные коды. Это имеет место по историческим причинам, так как они присутствуют в национальных кодировках.

Новые составные символы в Unicode не добавляются, их нужно “конструировать” из нескольких кодов



Определение языка

- Подходы
  - Графематический
  - N-граммный
  - Лексический

Графематический

- Система письменности
  - Кириллица
  - Латиница
  - …
- Алфавит
  - Русский А … Я
  - Украинский - не используются Ё, Ъ, Ы, Э, но есть Ґ, Є, І и Ї
  - Казахский….

N-граммный

Russian		Ukrainian		English		French	
^п 1.91	^по 0.84	^п 1.97	^на 0.85	^t 3.17	^th 2.00	es 2.31	es$ 1.77
^с 1.71	^пр 0.68	^в 1.75	на$ 0.73	th 2.48	the 1.62	le 1.97	^de 0.98
^в 1.68	^на 0.66	^н 1.68	^по 0.72	^a 2.41	he$ 1.44	^d 1.84	le$ 0.82
^н 1.55	^и$ 0.61	на 1.45	^пр 0.63	he 2.24	ed$ 0.78	^l 1.74	de$ 0.76
ст 1.43	^в$ 0.60	^з 1.40	^за 0.59	in 1.94	nd$ 0.73	on 1.70	^le 0.72
то 1.29	^не 0.56	^с 1.25	^не 0.56	er 1.60	ing 0.73	re 1.48	re$ 0.68
но 1.23	ть$ 0.48	ро 1.13	ого 0.54	an 1.54	^an 0.72	^c 1.46	nt$ 0.58

- Ранговый
- Марковский

Пословный

- ???
  - án került vagy től majd új ami ő kategória ben szerint amikor hogy amerikai két ezt mint alatt magyar itt második már ….
- ???
  - cel cod său cu cea l după ro va județul această în către sunt pe toate astfel ani prin ca departamentul din timpul într ….
- ???
  - ayrıca iklimi gibi tarafından olu kültür birlikte ula yol tarihinde veya iyi sonra türk bulunan kar çalı göre oldu ….
- ???
  - Би биеэ үнэлэгчдэд шүлэг уншдаг Тэгээд би дээрэмчидтэй хамт архи хуурдаг…

Пословный

- Венгерский
  - án került vagy től majd új ami ő kategória ben szerint amikor hogy amerikai két ezt mint alatt magyar itt második már ….
- Румынский
  - cel cod său cu cea l după ro va județul această în către sunt pe toate astfel ani prin ca departamentul din timpul într ….
- Турецкий
  - ayrıca iklimi gibi tarafından olu kültür birlikte ula yol tarihinde veya iyi sonra türk bulunan kar çalı göre oldu ….
- Монгольский
  - Би биеэ үнэлэгчдэд шүлэг уншдаг Тэгээд би дээрэмчидтэй хамт архи хуурдаг…

Полезные инструменты

- CLD (Compact Language Detector) – C++, Python
  - http://code.google.com/p/chromium-compact-language-detector/
- LanguageDetection – Java
  - http://code.google.com/p/language-detection/
- Видеолекция (Яндекс, RuSSIR 2012)
  - http://videolectures.net/russir2012_grigoriev_language/

Кореференция: синонимы

Различные способы названия одного и того же объекта

- Синонимы: [“ШАВЕРМА”, “ШАУРМА”]
- Аббревиатуры: [“БМП”, “БОЕВАЯ МАШИНА ПЕХОТЫ”]
- Транслитерация: [“PLAZMA”, “ПЛАЗМА”]
- Грамматические замены: [“ПОЗДРАВЛЕНИЕ”, “ПОЗДРАВИТЬ”]
- Переводы: [“ВОЗДУШНАЯ ТЮРЬМА”, “CON AIR”]
- Джойны: [“АУДИО КОДЕКИ”, “АУДИОКОДЕКИ”]

Источники расширений

- Словари синонимов
- Энциклопедические сайты
  - Википедия
  - Тематические сайты (kinopoisk)
- Скобочные написания в документах
- Логи запросов с кликами
- Переформулировки запросов
- Грамматические преобразования
- Языковые модели и дистрибутивная семантика

Дистрибутивная семантика

Дистрибутивная гипотеза

- Значение лингвистической единицы складывается только из ее употребления, использования.
- В мозге хранится сумма всех тех контекстов, в рамках которых мы слышали или видели то или иное слово.
- Это и есть его смысл. Без знания типичных соседей никакой семантики нет.

Вывод:

- Слова с похожими типичными контекстами имеют схожее значение

Дистрибутивная семантика

Счетные модели

- Совместная встречаемость
- Косинусная близость

Predictive models

- word2vec

Усечение окончаний (стемминг)

- Отсекаем самое длинное возможное окончание от слова, надеемся, что это не очень ухудшит результат по сравнению с лемматизацией.
- Для каждого языка свои таблицы окончаний.
- Для некоторых языков другие аналогичные методы.
- Результат довольно смешной:
  - сочи и сочиться приводятся к одной форме.
  - По старому анекдоту про Жуковского - у глагола «ховать» есть форма повелительного наклонения?

Усечение окончаний (стемминг)

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Лучше ли становится от усечения окончаний?

- Усечение окончаний улучшает одни запросы и ухудшает другие.
  - Где хорошо: [tartan sweaters], [sightseeing tour san francisco]
- (классы эквивалентности: {sweater,sweaters}, {tour,tours})
- Алгоритм Портера определяет следующий класс эквивалентности
  - operate operating operates operation operative operatives operational.
- Тогда в этих запросах станет хуже:
  - [operational AND research]
  - [operating AND system]
  - [operative AND dentistry]

Лемматизация

- Привести все разные формы к одной начальной.
  - Пример: am, are, is ➔ be
  - Пример: car, cars, car’s, cars’ ➔ car
  - Пример: the boy’s cars are different colors ➔ the boy car be different color
- Лемматизация заключается в поиске правильной основной формы для леммы в словаре.

Лемматизация

- Слово = машинная основа + парадигма
- Парадигма

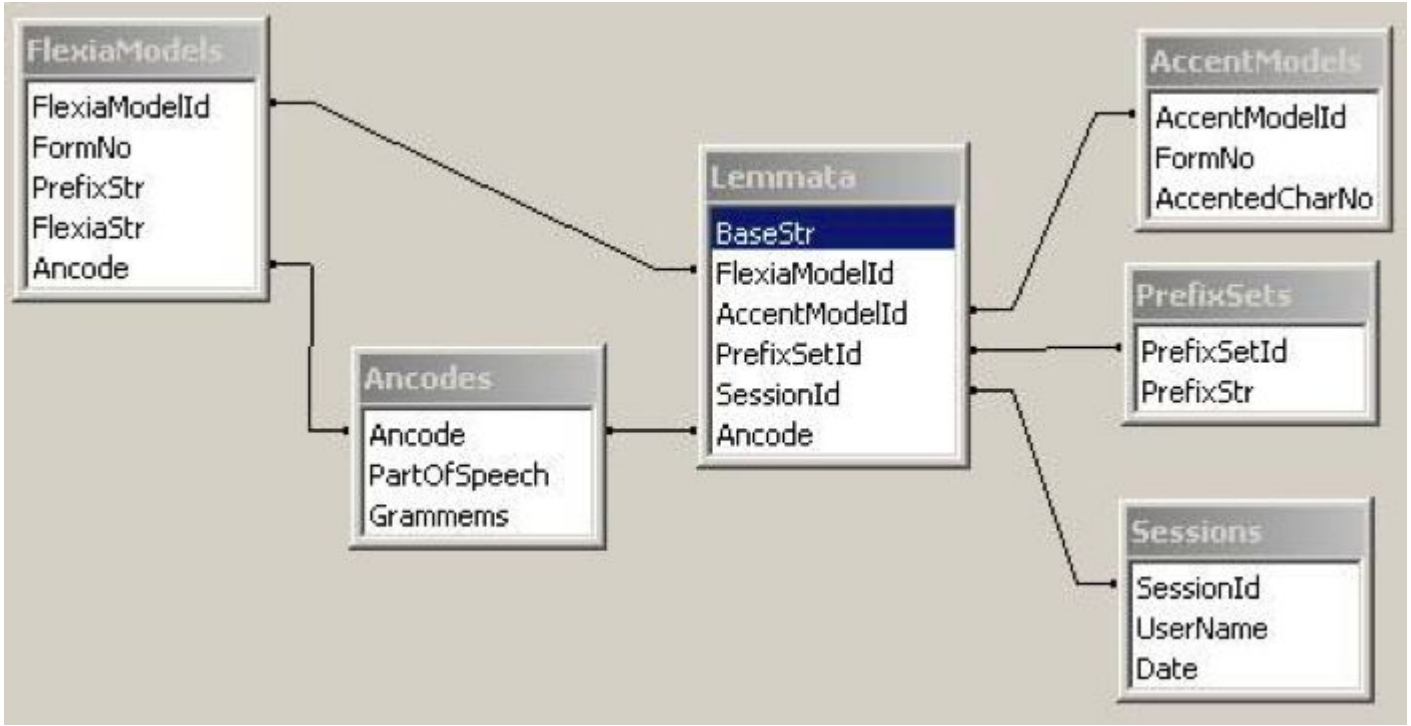
Лемматизация

- Слово = машинная основа + парадигма
- Парадигма
  - Парадигм а
  - Парадигм ы
  - Парадигм е
  - Парадигм у
  - Парадигм ой
  - Парадигм е

Лемматизация

- Слово = машинная основа + парадигма
- Парадигма
  - Парадигм а
  - Парадигм ы
  - Парадигм е
  - Парадигм у
  - Парадигм ой
  - Парадигм е

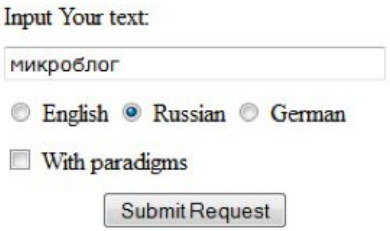
Лемматизация



(c) Сокирко А.В., http://www.aot.ru/docs/sokirko/Dialog2004.htm

Предиктор

- Что делать со словами, которых нет в нашем словаре?
- Ищем похожие!



Found	Dict ID	Lemma	Grammems	
-	но,	МИКРОБЛОГ	С мр,вн,им,ед,	АНАЛОГ
-		МИКРОБЛГИЙ	КР_ПРИЛ но,од,мр,ед,	НЕДОЛГИЙ

Откуда взять морфологию

- Для английского:
  - Стеммер Портера (Porter)
  - Стеммер Ловинса (Lovins)
- Для русского:
  - aot.ru
  - keva.ru (СтемКа)
  - MyStem (http://company.yandex.ru/technologies/mystem/)
  - pymystem
  - pymorphy
- Усекатель окончаний можно сделать самостоятельно.
  - Snowball - фреймворк для алгоритмов стемминга

Стоп-слова

- Стоп-слова очень часто встречающиеся слова, так что их появление в документе будет иметь мало ценности для выбора этого документа.
- В английском: a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with
- В русском: и, или, в, и, у, . . .
- Кроме того: 2010, 2011, корзина, . . .
- Ранее стоп-слова практически всегда удалялись.
- Но стоп-слова нужны для цитатного поиска: «King of Denmark», «Бахча У»
- Поисковые системы по вебу индексируют стоп-слова.

Где же проблемы?

- Языки разные
  - Изолирующие/Аналитические
  - Синтетические
    - Флективные
    - Агглютинативные
  - Полисинтетические

Изолирующие/Аналитические

- Изолирующие
  - Низкое отношение «морфема/слово»
- Аналитические
  - Грамматика – отдельными словами

- Китайский
- Английский

Флективные

- Индо-европейские
  - Русский
  - Испанский
  - ....

Флективные

- Финский

номинатив	talo	house
генитив	talon	of (a) house
эссив	talona	as a house
партитив	taloa	house (as an object)
транслатив	taloksi	to a house
инессив	talossa	in (a) house
элатив	talosta	from (a) house
иллатив	taloon	into (a) house
адессив	talolla	at (a) house
аблатив	talolta	from (a) house
аллатив	talolle	to (a) house
абессив	talotta	without (a) house
комитатив	taloineni	with (my) house(s)

Агглютинативные

- Тюркские языки

Turkish	English
ev	(the) house
evler	(the) houses
evin	your (sing.) house
eviniz	your (pl./formal) house
evim	my house
evimde	at my house
evlerinizin	of your houses
evlerinizden	from your houses

Полисинтетические

Чукотско-камчатские, эскимосско-алеутские и т.п.

Тымэйӈылевтпыгтыркын

(t-ə-mejŋ-ə-levt-pəγt-ə-rkən)

У меня сильно болит голова.

Омонимия

Разные по смыслу слова имеют одинаковое написание

Примеры:

белки бегали по лесу и ели орехи

(Лемма: белка, сущ. жен. род)

белки различаются по степени растворимости в воде

(Лемма: белок, сущ. муж. род)

- Словоформа - конкретная морфологическая разновидность слова

  белка, белку, белкой, белке

Неоднозначность

- Английский
  - Leg
  - Chair
- Русский
  - Лук
  - Очки
  - Лист

Снятие омонимии

- Rule-based
- Statistical

Rule-based

- X - verb or noun?
  - Preposition + X ➔ X – noun
  - Pronoun + X ➔ X – verb
  - You can do it by request
  - I request a book

Rule-based

```
//~ TN 26.1.a
//~ TC Если K - омоним с прилагательным или наречием
//~ TA И справа - enough
//~ TA И если омоним - переходный глагол, и справа от enough не
существительное
else if ((IsAdj(k) || IsAdverb(k))
        && CheckQuantitativeParticular(k + 1, QP_ENOUGH)
        && !CheckPrepParticular(k + 2, PP_OF)
        && !(IsTransitiveVerb(k) && IsNoun(k + 2)))
```

Статистическое снятие омонимии

- Размеченный корпус
- Собираем статистику
- Используем машинное обучение

Для интересующихся:

http://download.yandex.ru/company/Zelenkov_Segalovich.pdf

Рекомендуемая литература

[Введение в информационный поиск | Маннинг Кристофер Д., Шютце Хайнрих]

Для саморазвития (опционально)
[Чтобы не набирать двумя пальчиками]

Спасибо за внимание!

Антон Кухтичев

a.kukhtichev@mail.ru

@toshunster