Web as Corpus

Подходы к количественному и качественному анализу содержания Интернета

Serge Sharoff

Centre for Translation Studies
University of Leeds

s.sharoff@leeds.ac.uk

Outline

Яндексология

- НКРЯ: 90 миллионов слов за 1950-2003, русский

или Веб: триллионы слов, языки и жанры

(Turney, 2001) – синонимы (*levied* NEAR *imposed*)

(Keller, Lapata, 2003) – коллокации (*fulfill obligations*)

(Nakov, Hearst, 2005) – именные группы
(*female bus driver* vs *school bus driver*)

(Bergsma, 2005) – разрешение анафоры
(*Chengxiang himself* or *Chengxiang herself*)

- Googleology (Adam Kilgarriff), Яндексология

Ограничения Яндексологии

- нет частей речи, странная лемматизация
  frustrate, frustration; *Chang, change*, мятый→помни
- Limited metalanguage (NEAR operator)

  “fulfill the obligation”
  “fulfill the obligations”
  “fulfill his obligation”
  “fulfilling an obligation” etc
- Нет жанров и тем
- Непредсказуемые сервера, странности с частотными словами
  (*Chirac OR Sarkozy*)
- Изменения операторов и механизма поиска

Сделай сам

- Bootcat: использование интерфейса поисковых машин
- Нетематические слова для универсальных корпусов
  conditions, clearly, ground, much
- В результате можно получить корпуса 30000-80000 стр, 100-200MW:
  I-AR, I-DE, I-EN, I-EL, I-ES, I-FI, I-FR, I-IT, I-JP, I-PL, I-PT ...
- Тематические слова для тематических корпусов
  En *autosave, configuring, debugger, user-friendly*
  Ru автосохранение, настройка, отладчик, дружественный
  Zh 自动保存, 配置, 调试, 友好界面

Сбор корпуса по возобновляемой энергетике

- Ключевые слова из категории Википедии

fossil fuel	化石燃料	化石燃料	ископаемое топливо
power station	发电厂	發電廠	электростанция
hydroelectricity	水力发电	水力發電	гидроэнергетика
photovoltaics	太阳能光伏	太陽能光伏	фотоэлектричество

- Запросы: *“ископаемое топливо” “фотоэлектричество”*

	En	Ru	Zh(CN)	Zh(TR)
URLs:	5762	5991	2399	918
Words (MW):	6.5	5.8	8.4	4.2

Оценка содержимого по ключевым словам

7467 renewable energy	5629 источник энергия
4352 wind turbine	4550 окружающий среда
3973 fossil fuel	2754 электрический энергия
3127 greenhouse gas	2710 солнечный батарея
3049 natural gas	2274 солнечный энергия
2539 wind farm	2106 природный газ
2320 solar energy	1994 тепловой энергия
2265 energy efficiency	1870 возобновлять источник
1994 carbon dioxide	1561 производство электроэнергия
1920 solar cell	1508 возобновлять источник энергия
1782 wind energy	1439 изменение климат
1722 generate electricity	1401 парниковый газ
1559 solar patch	1315 альтернативный источник
1533 electricity generation	1289 экологически чистый
1529 fuel cell	1280 энергия ветер
1517 gas emission	1270 устойчивый развитие

Кролинг

- Heritrix, Nutch
- deWaC, itWaC, ukWaC, **ruWaC** – 2GW
- Отслеживание поискового спама, дубликатов
- Выделение текста (навигация, boilerplate)
- Лемматизация, синтаксический анализ

Что мы знаем о слонопотаме?
Сравнение корпусов и автоматическая разметка

Многомерное масштабирование

Многомерное масштабирование (Multi-dimensional scaling, MDS)

$$\Delta := \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & \vdots & & \vdots \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix}$$

Цель: найти I векторов $x_1, \ldots, x_I \in \mathbb{R}^N$ таких, что

$$\|x_i - x_j\| \approx \delta_{i,j} \forall i, j \in I$$

Геометрическая интерпретация

Поделим Ирландию

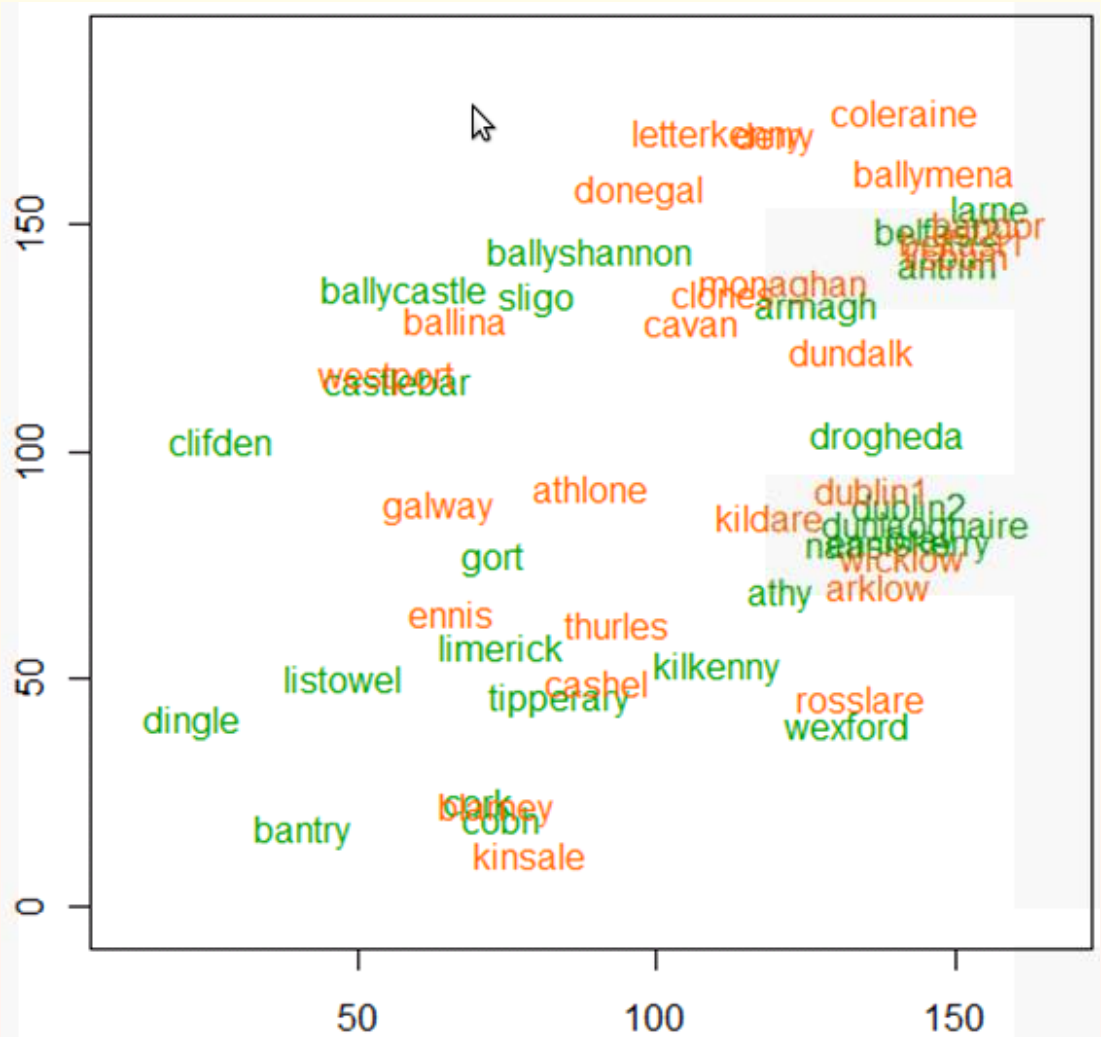
Town	antrim	armagh	athy
antrim	0	31	105
armagh	31	0	74
athy	105	74	0

...

Town	arklow	athlone	ballina
arklow	0	80	146
athlone	80	0	66
ballina	146	66	0

26 размерностей в каждом списке

Объединим Ирландию



Корреляция: 0.9941–0.9986

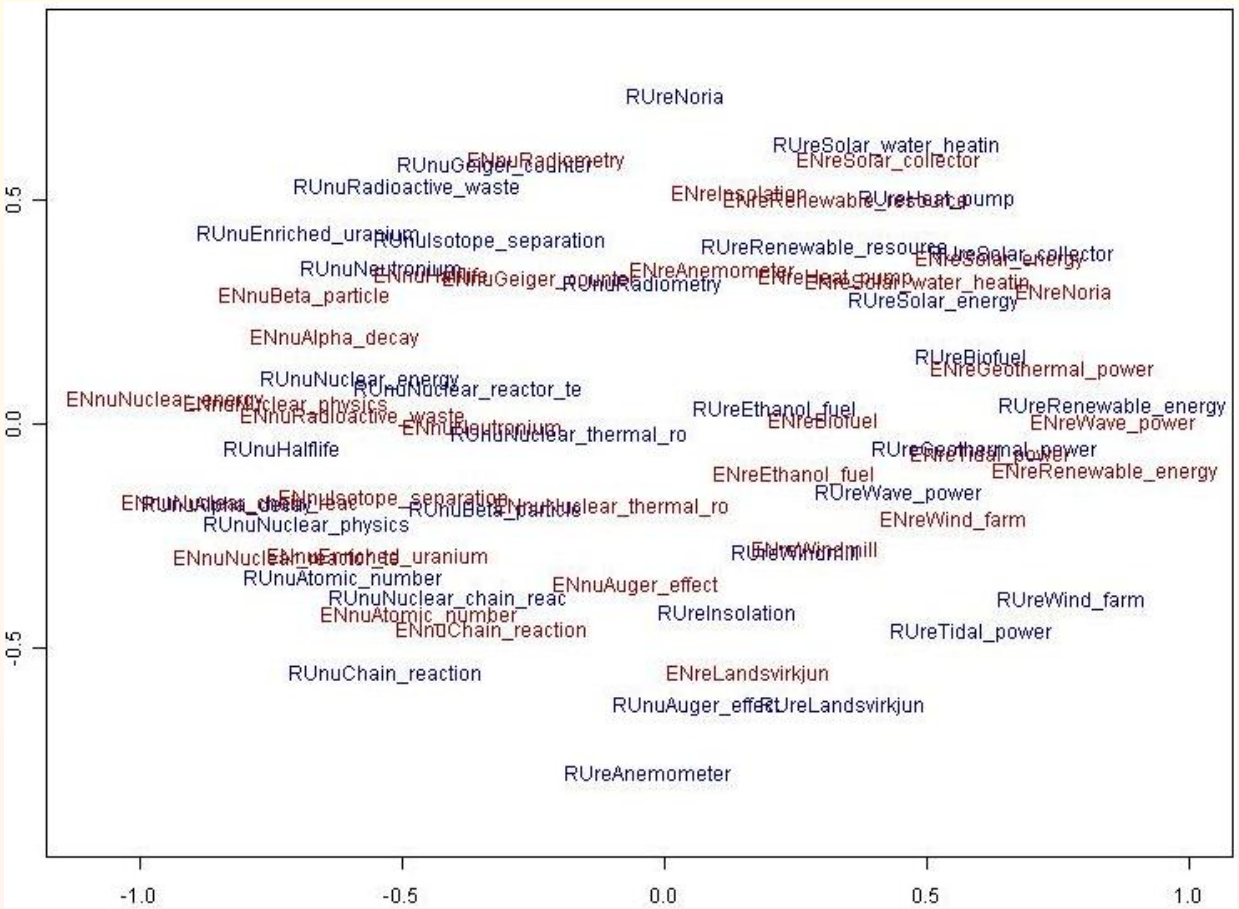
Применение масштабирования к сравнению корпусов

Тексты Оценка расстояний между текстами в каждом языке отдельно (ранговая корреляция частот)

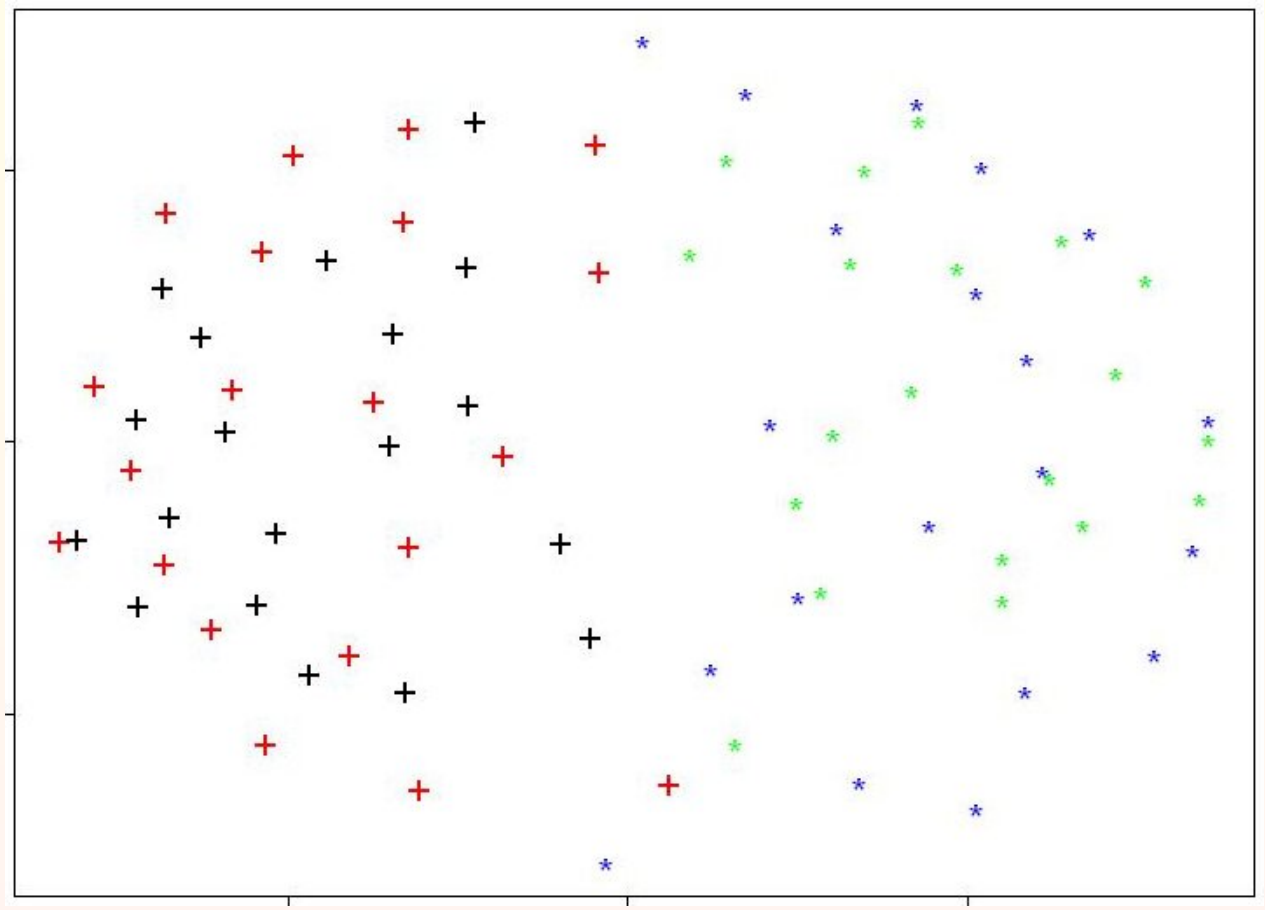
Масштабирование Применить MDS для каждого языка отдельно

Наложение Используем реперные точки (anchors), чтобы объединить заведомо похожие страницы (**NB** $N_a << N_c$, линейная регрессия)

Сравнимость между языками



Сравнимость между языками



+=RuNu, +=EnNu, *=RuRe, *=EnRe

Outline

Машинное обучение

- Признаки для описания документов
  предметные области: лексика (ключевые слова)
  жанры: слова, символьные коды, синтаксис
- Алгоритмы: с обучающей выборкой и кластеризация

Проблемы использования обучающей выборки

Необходимость большого количества данных (использование неразмеченных данных, Active learning, semi-supervised learning)

Проблемы использования кластеризации

- Выбор количества кластеров
- Выбор ключевых признаков
- Сравнение результатов

Выделение ключевых слов на корпусе

- TF-IDF, χ-statistics, Mann-Whitney test (Kilgarriff, 1997), mutual information (Baroni, 2004)
- log-likelihood (Rayson, Garside, 2000)

	Корпус A	Корпус B	Всего
Частота слова	a	b	a+b
Частота остальных слов	c-a	c-b	c+d-a-b
Общий размер	c	d	c+d

$$G^2 = a \ln \frac{a}{E1} + b \ln \frac{b}{E2}; E1 = c\frac{a+b}{c+d}; E2 = d\frac{a+b}{c+d}$$

- Статистически значимый порог

Кластеризация

- Оптимальное число кластеров (l_2, perlexity)
- Жесткая кластеризация (hard clustering), Repeated Bisections in Cluto
- Мягкая кластеризация (soft clustering), probabilistic topic models

Кластеризация БНК и Интернета

BNC	Words	Sim	**ukWaC**	Words	Sim
Fiction	21.0%	0.088	Life	12.4%	1.422
Politics	13.0%	0.151	Business	9.9%	9.334
Linguistics	6.6%	0.101	Travel (guides)	7.5%	-0.593
Museums/Art	5.8%	0.114	Politics	7.3%	0.577
Spoken (demo)	5.6%	0.244	Computing	6.3%	6.936
Sports	5.3%	0.211	DIY	5.8%	0.827
Food/Gardening	5.2%	0.078	Shopping	5.7%	0.832
Business	5.1%	0.125	Research	5.6%	-0.735
Music/Movies	5.1%	0.130	Computing	5.3%	0.555
Nat sciences	3.9%	0.115	Music	4.8%	-0.407
Travel/Housing	3.7%	0.114	Education	4.5%	1.184
Education	3.7%	0.157	Food/Gardening	3.5%	-0.630
Legal	3.2%	0.125	Travel (hotels)	3.3%	-0.537
Medicine	2.5%	0.180	Medicine	3.1%	0.141
Newsp reports	2.5%	0.334	Sports	3.1%	-0.147
Religion	2.0%	0.154	Healthcare	3.1%	0.210
Computing	1.8%	0.572	Sports	2.5%	-1.483
Spoken (context)	1.6%	0.138	Movies	2.4%	-0.418
Aviation/railways	1.3%	0.188	Religion	2.3%	-0.152
Environment	0.9%	0.301	House sale	1.4%	0.107

N	BNC codes
424	W.fict.prose
49	W.misc
43	S.oral.history
29	W.fict.poetry
26	W.biography
19	W.non.ac.soc
J.Dawson, *How do I look?*	
5	W.pop.lore
5	W.letters.personal
5	W.essay.school
5	S.brdcast.discussn
4	W.religion
4	W.non.ac.medicine
R.McCall, *Hearing loss?*	
4	S.speech.unscripted
4	S.classroom
3	W.non.ac.humanities
J.Jones, *Dostoevsky.*	
3	W.newsp.brdsht.social
3	S.consult
2	W.ac.polit.law
2	S.courtroom
1	W.newsp.other.social
1	W.newsp.brdsht.misc

BNC	LL-score	ukWac	LL-score
her	166623	I	18635
him	48459	her	7851
me	38013	he	6358
my	28576	she	6151
say	27237	my	5935
look	20199	his	5515
eye	19647	me	4444
smile	14535	him	2690
back	14432	i	2246
go	13128	it	2052
feel	12395	do	2000
door	12276	say	1922
tell	12165	PM	1713
hand	11973	post	1706
know	11822	man	1678
think	11606	go	1632
like	11058	have	1560
could	11020	love	1546
herself	10728	think	1540
face	10676	but	1507
man	10483	get	1338
down	9688	Wigton	1331
sit	9176	know	1324
come	8931	tell	1271

BNC	LL-score	ukWac,C1	LL-score	ukWac,C2	LL-score
Inc	31698	datum	5998	file	19892
Corp	26689	system	5819	search	7931
software	24243	software	5659	user	6585
user	24220	model	4489	text	4901
Unix	20408	use	3295	server	4878
system	19051	computer	2798	use	4491
lifespan	18968	network	2729	Windows	4196
IBM	18841	data	2509	web	4009
module	18417	technology	2183	page	3982
application	12362	user	2152	site	3446
version	12207	solution	1975	you	3432
package	11949	component	1952	library	3255
its	11928	algorithm	1586	directory	3228
file	10931	design	1511	Web	3170
disk	10784	method	1456	browser	2781
window	10006	WiMAX	1454	Internet	2756
database	9977	device	1450	HTML	2743
product	9912	nolvadex	1413	nodine	2671
will	9838	analysis	1393	password	2652
Microsoft	9282	interface	1339	click	2640
Sun	9089	Software	1330	database	2636
computer	8709	test	1270	command	2606
server	8608	wireless	1252	format	2546
Systems	8594	Systems	1247	default	2537

http://members.lycos.co.uk/clickforme/howtotakenolvadex.html, 1764 words

1796.61	90	nolvadex
487.64	30	Cached
264.53	17	tamoxifen
164.57	30	similar
92.53	7	estrogen
40.40	6	generic
32.64	4	steroid
24.42	10	woman
22.32	7	drug
20.81	4	prescribe
20.63	5	therapy

http://www.businessvision.co.uk/messages/28726.html, 2247 words

11410.43	509	nolvadex
1053.66	60	tamoxifen
245.37	16	10mg
226.01	24	tablet
199.83	12	doxycycline
151.78	18	generic
149.44	10	anticancer
130.57	28	treatment
125.87	8	uncoated
117.24	8	20mg
94.28	20	drug
81.45	10	hormone

itWac1	LL-score	itWac2	LL-score
vino	35572	mio	37072
olio	21983	mi	29263
cucchiaio	13327	che	27386
uovo	13066	suo	26494
burro	12502	non	26038
cuocere	12157	essere	24613
pasta	11880	dire	20620
piatto	11042	avere	20502
latta|latte	10573	io	18762
zucchero	10025	uomo	18355
formaggio	9698	tuo	17485
pepe	9428	fare	17153
farina	9011	lui	17022
cucina	8635	quello	16081
ingrediente	8562	ti	15871
oliva	8301	ma	15625
cipolla	8115	occhio	15328
tritare	7926	lo	15229
cottura	7849	cosa	15215
minuto	7326	amore	14900
acqua	7295	lei	14760
mescolare	7168	tu	14628
forno	7140	un	13431
fresco	6755	sapere	13045

ukWac	LL-score
plant	8343
fish	5455
bird	5160
garden	4051
flower	3669
tree	3633
food	3354
specie	2809
fruit	2606
wine	2348
seed	2245
vegetable	2050
cook	1832
sauce	1790
soil	1643
sugar	1613
eat	1538
meat	1537
flavour	1518
cheese	1382
herb	1372
egg	1363
recipe	1360
species	1354

не	2640349
я	2014056
он	1554719
она	739238
ты	502231
его	359128
сказать	298385
знать	247900
рука	156589
ее	154615
глаз	110351
голова	95409
спросить	61209
дверь	60287
голос	55064
подумать	39335
словно	28328
мистер	10080
кивнуть	9912

для	546727
система	99123
программа	61126
данные	36673
сеть	26129
компьютер	23942
пользователь	15954
диск	10663
приложение	8327
программный	7587
сервер	7320
код	7046
разработчик	5638
процессор	5307
linux	4239
интерфейс	3532
разъем	1148
гост	885
субд	829

use	4428090
system	1285828
design	899549
user	564370
datum	543017
model	469346
technology	422882
method	408956
network	393205
computer	390438
software	350971
solution	310791
analysis	296511
component	164674
data	160801
device	159826
interface	98791
algorithm	38926

Подходы к описанию жанров

Brown 500 отрывков, 2000 слов в каждом, по 15 жанров: A) Press: reportage, B) Press: editorial, C) Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, ...

BNC около 4,000 текстов по 70 жанров (ac.med, ac.tech, non-ac.tech, news...), публикации (book, periodical, ephemeral, ...), и т.п.

БЛ Adventure stories, Detective stories, Picaresque literature, Robinsonades, Sea stories, Spy stories, Thrillers, Allegories, Didactic fiction, Fables, Parables, Alternative histories, Dystopias, Bildungsromane, Arthurian romances, ...

Adamczik 4,500 Textsorten

Amazon http://www.amazon.co.uk/Books-Categories/

Примеры жанров на Интернете

MGC (20 genres)		KI-04 (8 genres)	Santinis (8 genres)
adults	official	article	blog
blog	personal	discussion	eShop
childrens	poetry	download	faq
commercial	prose	help	frontpage
community	scientific	linklists	hotlist
content delivery	shopping	portrait-nonpriv	homepage
entertainment	user input	portrait-priv	sitemap
error message		shop	Search page
FAQ			
gateway			
index			
informative			
journalistic			

349 жанров в корпусе Syracuse University

Жанры в Вавилонской библиотеке

By this art you may contemplate the variations of the 23 letters. R. Burton (1621) *The Anatomy of Melancholy*

Вселенная - некоторые называют ее Библиотекой - состоит из огромного, возможно, бесконечного числа шестигранных галерей, с широкими вентиляционными колодцами, огражденными невысокими перилами.
Как все люди Библиотеки, в юности я путешествовал. Это было паломничество в поисках книги, возможно каталога каталогов; теперь, когда глаза мои еле разбирают то, что я пишу, я готов окончить жизнь в нескольких милях от шестигранника, в котором появился на свет. Когда я умру, чьи-нибудь милосердные руки перебросят меня через перила, могилой мне станет бездонный воздух; мое тело будет медленно падать, разлагаясь и исчезая в ветре, который вызывает не имеющее конца падение.
Хорхе Луис Борхес (1941) Вавилонская библиотека

Требования к таксономии жанров

Michael Halliday's Introduction to functional grammar

to say sensible and useful things about any text in English

- Короткий список, подящий для большинства веб-страниц
- Обобщенные классы (FAQs vs. instructions)
- Возможность работы с новыми жанрами (Интернет-жалобы)
- Определение по внешней функции (genre vs. register)
- Look'n'feel vs. технические термины (блоги vs. argumentative, но разные виды блогов) $\rightarrow$ **инструкция**
- Надежность аннотирования: $x \in \Theta, \Phi$
- Внутренняя совместимость классов $x \in \Theta, y \in \Theta$ но $x \nsim y$

Sinclair's aims of text production

- Синклер: the EAGLES guidelines (1996)
  discussion – polemic, position statements
  recommendation – laws, reports and reviews
  recreation – fiction, non-fiction and popular lore
  religion – holy books, sermons, prophecies
  information – reference materials
  instruction – how-tos, textbooks and academic works
- Кром того теги для аудитории, вида публикации и автора

Проблемы

- Внешние признаки или интенции (e.g. opinions vs. report)
- Внутренняя неоднозначность (ср с традиционной риторикой)
- Совместимость классов (e.g. laws vs. reviews) $\rightarrow$
  Предсказуемость текста **исходя из** жанра

Functional Genre Classes (FGC)

1. **information** (catalogues, glossaries, home pages)
2. **instruction** (how-tos, FAQs, tutorials)
3. **promotion** (adverts, shops, political pamphlets[?])
4. **recreation** (fiction and popular lore)
5. **regulations** (laws, small print, contracts)
6. **reporting** (newswires, police reports)
7. **discussion**:
   academic (research papers and monographs)
   public (journalism and political debates)
   everyday communication (forums, emails, diary blogs)
8. **non-text** (pages with little running text):
   applications (Flash, Java, applets)
   online interfaces (query/login/purchase forms, download)
   linkerie (portals, link lists)

Аннотация в TTC

Жанры Наиболее типичные категории (22 labels)
academic article, news article, adverts, legal text, expert report, guide, FAQs entries, catalog, glossary entries, announcement, encyclopedic entries, no text, blog, threads, homepages, reviews, warning, editorial, abstract, other

Обобщенные жанры Functional Genre Classes
Information, discussion, instruction, listing, recreation, regulation, promotion, reporting, non-text

Регистр формальный/неформальный

Автор-аудитория Spec2Spec, Spec2Public, Public2Public

- 120 страниц, 8 аннотаторов
- Fleiss' κ score

Результаты разметки

Kappa score

Жанр	0,472
FGC	0,501
Регистр	0,345
Автор-аудитория	0,097

Fleiss' κ	Результат
< 0	Poor
0.0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Применение автоматических методов к Интернету

- символьные н-граммы имеют лучшую предсказательную силу

Labels	**Top ranking tetragrams**
Reporting	rday &_,_ week sday _wee rida nnou _ann noun e_19 quot;._ frid _th iday esda
Academic discussion	ms_t erab s_by ly_l eris ed_e ompa ly_u naly onta tiga ampl ch_w m_a_ n_ap y_lo alys

rday=yesterday,Saturday; &=marks abbreviations (Mr., Co.); *sday*=Tuesday, Wednesday, Thursday; _*wee*=(last, this) week; *rida*=Friday; *nnou*=announced; **Academic** *ms_t*=aims/claims/seems/problems/systems to/that; *erab*=considerably; *s_by*= presentations/lines/facilities/methods/beams by; *ly_l,ly_u,sly*=adverbs; *eris*=characteristic; *ompa*= compare /comparison; *onta*=contain; *tiga*=investigate/investigation; *ampl*=example/sample; *naly, alys*=analysis

- `http://sml c09.leeds.ac.uk/vincent/genre/`

Окончательное решение

Categories	BNC	I-EN	ukWac
discussion	37.42%	52.49%	38.21%
information	6.00%	4.03%	5.03%
instruction	26.66%	20.51%	18.77%
promotion	5.45%	11.24%	15.66%
recreation	21.43%	0.97%	1.03%
regulation	3.05%	2.21%	3.03%
reporting	6.00%	8.54%	18.27%

Частотные списки и статистика

- Изучающим язык важно знать более частотные слова (LDOCE, COBUILD)

maintain: Can they **maintain** it?
a. 维持
b. 扩大
c. 改善
d. 得到

- *he is a **soldier**; it has been **restored**; he couldn't stop **roving** <. . . <*

  *it was very **bawdy**; she used a **thesaurus**; he looked into her **limpid** eyes*

  *rampant, precinct, deform, **thesaurus**, heresy, skate, gamma, criminology, fondant*

Частотные списки и статистика

- Всплески частот (*whelk, nolvadex, nodine*)
- Неправильные корпуса:
  anchor, instrumental, sodium, ***banana****, tilt, hunter, armour*
  leer, enthrall, sheaf, ***toothbrush****, dungeon, stocky, lawsuit*
- Многословные выражения (*as well as, have to*);
- Трианулляция из нескольких языков в Kelly
  72 language pairs, *el* → *no* or *zh* → *ru*
  乐趣 (B1)→ удовольствие; 生活乐趣 →радости жизни

Поиск объективной оценки текста

- Преподаватели различаются в оценках одного текста
- Студенты различаются в использовании стратегий чтения
- Objective measures:
  - скорость чтения: количество понятого текста
  - отслеживание движений глаз (eye tracking) и сканирование мозга (brain imaging): мало информативно и нет связи с функцией
  - crowd sourcing: Simple Wiki

Признаки

1. покрытие наиболее частотными словами
   top 1000, top 2000, top 3000 words

   Источники General Service List, частотные списки из Интернета

   Покрытие не всегда релевантно: *Netanyahu gave in to demands to close up the deficit gap of 18 billion shekel.*

   Нетаньяху, 内塔尼亚胡
2. Частеречные тэги, словоформы, леммы

Feature set: text complexity

1. average sentence length (ASL)
2. average word length in syllables (ASW)
3. Flesch Reading Ease (FRE) – language-specific formulas
4. coverage by more frequent POS trigrams – language models
5. average number of lexical verbs per sentence
6. average number of passive verbs per sentence
7. average number of modal verbs per sentence
8. average number of prepositions per sentence
9. average number of punctuation marks per sentence

Признаки

1. покрытие наиболее частотными словами
   top 1000, top 2000, top 3000 words

   Источники General Service List, частотные списки из Интернета

   Покрытие не всегда релевантно: *Netanyahu gave in to demands to close up the deficit gap of 18 billion shekel.*

   Нетаньяху, 内塔尼亚胡
2. Частеречные тэги, словоформы, леммы

Feature set: text complexity

1. average sentence length (ASL)
2. average word length in syllables (ASW)
3. Flesch Reading Ease (FRE) – language-specific formulas
4. coverage by more frequent POS trigrams – language models
5. average number of lexical verbs per sentence
6. average number of passive verbs per sentence
7. average number of modal verbs per sentence
8. average number of prepositions per sentence
9. average number of punctuation marks per sentence

Признаки

1. покрытие наиболее частотными словами
   top 1000, top 2000, top 3000 words

   Источники General Service List, частотные списки из Интернета

   Покрытие не всегда релевантно: *Netanyahu gave in to demands to close up the deficit gap of 18 billion shekel.*

   Нетаньяху, 内塔尼亚胡
2. Частеречные тэги, словоформы, леммы

Feature set: text complexity

1. average sentence length (ASL)
2. average word length in syllables (ASW)
3. Flesch Reading Ease (FRE) – language-specific formulas
4. coverage by more frequent POS trigrams – language models
5. average number of lexical verbs per sentence
6. average number of passive verbs per sentence
7. average number of modal verbs per sentence
8. average number of prepositions per sentence
9. average number of punctuation marks per sentence

Coh-Metrix project (Uni Memphis)

- Word information: Familiarity, concreteness, imageability, Colorado meaningfulness, Paivio meaningfulness, . . . ;
- Mean logarigthm of frequency of content words;
- Incidence score: type/token ratio for POS tags;
- Density score for nouns and pronouns, for logical operators (*and, not, if*), connectives (*in other words*), density of NPs (number of modifiers per noun)
- Polysemy and hyperonymy scores: senses and levels

Text coherence measures

- Noun / argument / stem overlap (*apply heat/heated*)
- Cohesion matrix **R** with weighted distance between sentences
- Local cohesion $\frac{\sum R_{i,i+1}}{n-1}$ vs global cohesion $\frac{\sum_i \sum_j R_{ij|i<j}}{n \times \frac{n-1}{2}}$
- LSA cohesion (cosine distance between text elements)

PCA transform

Past (without coherence and log frequencies):

0.415prepositions+0.386verbs-0.352fre+0.334passiveverbs+0.32 top3000...
-0.416top2000-0.412top3000-0.41top1000+0.375punctuation+0.36conj

Using the new features:

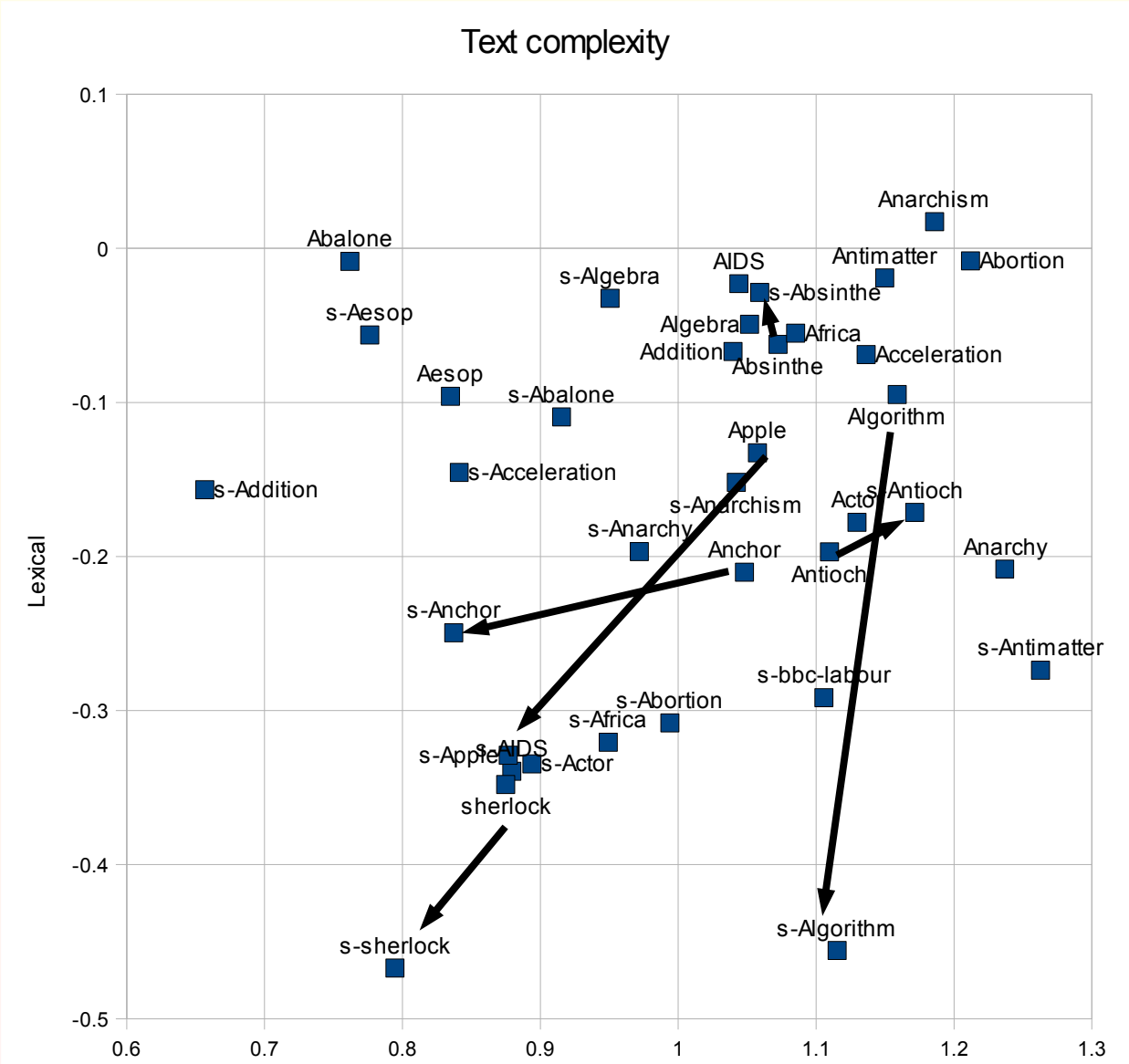
-0.385asl-0.369prepositions-0.318verbs-0.317noun-0.276passive-0.262adj-0.240postri...
0.316Ncoh2+0.316Ncoh3+0.315coh2+0.315coh3-0.314bigr-0.306trigr-0.273uni...
0.344coh3+0.343coh2-0.343pronoun+0.339Ncoh3+0.338Ncoh2...

Using the ratios between S and E:

-0.322prepositions-0.319asl-0.315uni-0.307bigr-0.301verbs-0.291top3000...
0.391fre-0.365asw-0.341noun-0.32punctuation-0.277conj+0.259trigr+0.255bigr...
0.472nouncoh3+0.469coh3+0.469coh2+0.466nouncoh2+0.132prepositions...

- no pos trigrams, modal verbs, proper nouns
- grammar/lexicon dimensions retained

Размерности для английского



Английские примеры

1.a `s-apple`: *People first grew apple trees in Central Asia.*

1.b `apple`: *The tree originated from Central Asia, where its wild ancestor is still found today.*

- Problems with simple wikis:
  `s-absinthe`: 'The English used in this article may not be easy for everybody to understand.'
  `s-antioch`: Сокращение вместо упрощения