

Разные методы сравнения корпусов на основе частотных словарей

Александр Пиперски

Семинар по теоретической семантике
ИППИ РАН, 17.03.2017

Постановка задачи

- Найти меру $d(A, B)$, которая позволит оценивать, насколько похожи между собой корпуса текстов A и B , и сравнивать такие оценки между собой.
- Пример:
 - На что больше похожи английские любовные романы — на научную фантастику или на приключенческие романы?

Сходные задачи

- Информационный поиск
- Детекция плагиата

Информационный поиск: отличия

- Оценивается сходство текстов существенно различного размера (длинных документов с коротким поисковым запросом).
- Существенно только ранжирование документов, наиболее сходных с запросом.

Детекция плагиата: отличия

- Ищутся точные текстуальные совпадения.
- Подразумевается злая воля противоположной стороны.

Сходство vs. различие

- Взгляд на сходство с другой стороны — различие.
- NB: сегодня я употребляю попеременно слова «близость» / «сходство» и «расстояние» / «различие».

Пожелания к мерам расстояния между корпусами (1)

- Мера должна являться метрикой (в математическом смысле)
 - Тожество: $d(A, A) = 0$
Корпус находится от на расстоянии 0 от самого себя.
 - Симметрия: $d(A, B) = d(B, A)$
Расстояние от A до B равно расстоянию от B до A .
 - Неравенство треугольника:
 $d(A, C) \leq d(A, B) + d(B, C)$

Пожелания к мерам расстояния между корпусами (2)

- Ограниченность:
мера должна иметь фиксированный диапазон значений (тогда расстояние легко преобразуется в близость и наоборот).
- Устойчивость к окружению:
 $d(A, B)$ не должно меняться в зависимости от того, какие корпуса есть в общем наборе для исследования.
- **Соответствие реальности**
(об этом позже)

Пожелания к мерам расстояния: пример нарушения

- Стандартная поисковая мера tf-idf нарушает пожелание устойчивости к окружению, так как idf (inverse document frequency) зависит от общего количества сравниваемых документов и от количества документов, содержащих искомое слово.

Методы сравнения корпусов

- Kilgarriff, Adam. 2001. Comparing Corpora. *International Journal of Corpus Linguistics* 6.1: 97–133.



Два типа мер расстояния

- Меры на основе перплексивности (perplexity)
 - Строим n -граммную языковую модель на основе корпуса A и смотрим, насколько хорошо она позволяет предсказать данные корпуса B .
- Меры на основе частотности слов
- Частотность слов лучше соответствует реальности (об этом опять-таки позже).

Классические меры расстояния на основе частотных списков

- χ^2
- Коэффициент ранговой корреляции
Спирмена
- Описания: Kilgarriff (2001), Schäfer &
Bildhauer (2013)

Мера 1: χ^2

- Строим частотные списки корпусов A и B .
- Выбираем топ- N слов из объединённого частотного списка для двух корпусов.
- Представляем данные в виде таблицы сопряжённости $(N+1) \times 2$ (добавив строку для всех остальных слов) и вычисляем χ^2 обычным образом:

$$\chi^2 = \sum_{i,j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

χ^2 : сравнение двух корпусов по Ляшевская & Шаров (2009)

	Худ50-60	Публ50-60	Σ
<i>и</i>	0.038	0.037	0.075
<i>в</i>	0.024	0.035	0.059
<i>не</i>	0.021	0.016	0.037
<i>на</i>	0.017	0.015	0.032
<i>я</i>	0.017	0.014	0.031
<i>быть</i>	0.012	0.014	0.026
<i>он</i>	0.020	0.012	0.032
<i>с</i>	0.011	0.011	0.022
<i>что</i>	0.008	0.008	0.016
<i>а</i>	0.010	0.006	0.016
ПРОЧЕЕ	0.822	0.832	1.654
Σ	1.000	1.000	2.000

χ^2 : сравнение двух корпусов по Ляшевская & Шаров (2009)

	Худ50-60	Публ50-60	
<i>и</i>	7×10^{-6}	7×10^{-6}	
<i>в</i>	0.0010	0.0010	
<i>не</i>	0.0003	0.0003	
<i>на</i>	6×10^{-5}	6×10^{-5}	
<i>я</i>	0.0001	0.0001	
<i>быть</i>	7×10^{-5}	7×10^{-5}	
<i>он</i>	0.0010	0.0010	
<i>с</i>	0	0	
<i>что</i>	0	0	
<i>а</i>	0.0005	0.0005	
ПРОЧЕЕ	3×10^{-5}	3×10^{-5}	
			0.0064

Мера 1: χ^2

- NB: в таблицу внесены относительные частоты, а не абсолютные, потому что нас интересует только размер эффекта, а не статистическая значимость (не нужно, чтобы пропорциональное увеличение выборок меняло расстояние между корпусами).
- Диапазон: от 0 (абсолютно одинаковые частоты) до 1 (абсолютно разные частоты).

χ^2 : расстояния между 4 корпусами
по Ляшевская & Шаров (2009)

	Худ50–60	Худ70–80	Публ50–60	Публ70–80
Худ50–60	0	0.0003	0.0064	0.0039
Худ70–80	0.0003	0	0.0065	0.0035
Публ50–60	0.0064	0.0065	0	0.0012
Публ70–80	0.0039	0.0035	0.0012	0

Мера 2: Спирмен

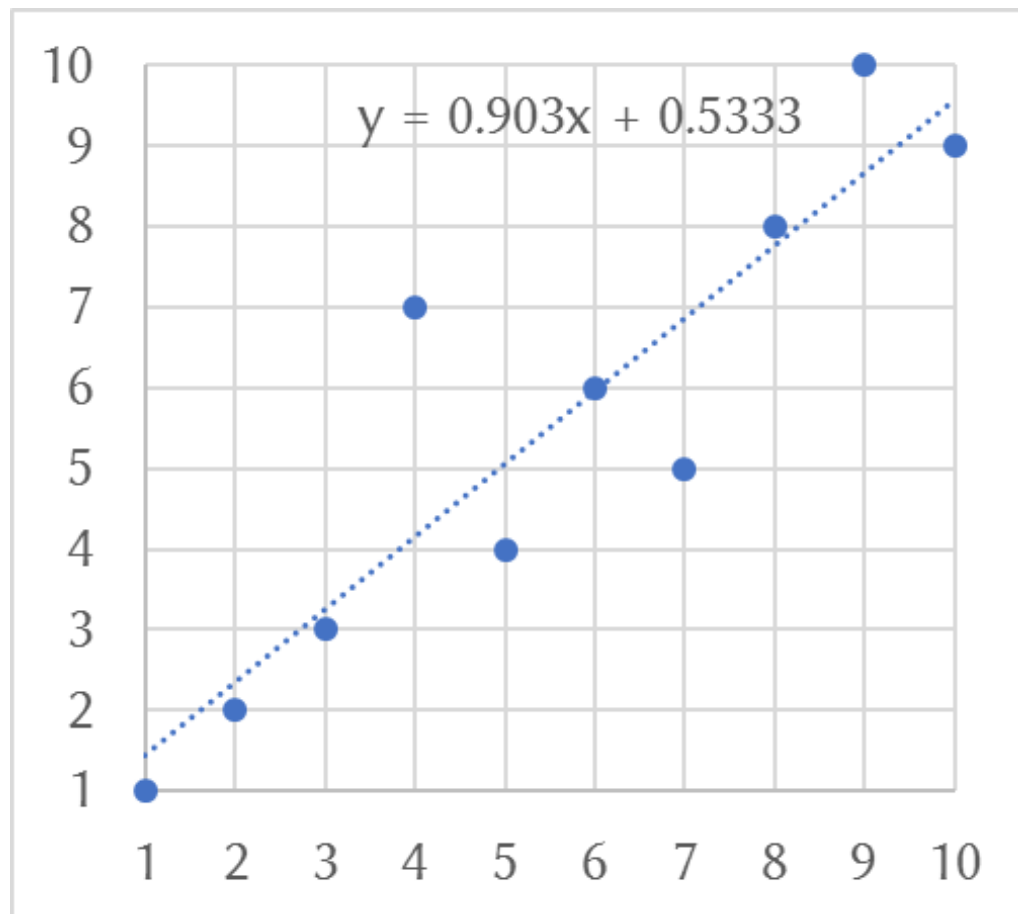
- Строим частотные списки корпусов A и B .
- Выбираем топ- N слов из объединённого частотного списка для двух корпусов.
- Определяем ранги слов в частотном списке, включающем в себя эти N слов, для A и в таком же списке для B и вычисляем коэффициент корреляции Спирмена r_s обычным образом:

$$r_s = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

Спирмен: сравнение двух корпусов по Ляшевская & Шаров (2009)

	Худ50-60		Публ50-60	
<i>и</i>	0.038	1	0.037	1
<i>в</i>	0.024	2	0.035	2
<i>не</i>	0.021	3	0.016	3
<i>на</i>	0.0172	5	0.015	4
<i>я</i>	0.0170	6	0.0139	6
<i>быть</i>	0.012	7	0.0145	5
<i>он</i>	0.020	4	0.012	7
<i>с</i>	0.011	8	0.011	8
<i>что</i>	0.008	10	0.008	9
<i>а</i>	0.010	9	0.006	10

Спирмен: сравнение двух корпусов по Ляшевская & Шаров (2009)



Мера 2: Спирмен

- Мера слишком чувствительна к частотам внизу списка (если переставить местами 1-е и 2-е слово, эффект тот же, как если переставить 100-е и 101-е слово)
- Диапазон: от -1 (противоположный порядок) до 1 (идентичный порядок)
- Мера расстояния: $1 - r_s$

Спирмен: расстояния между 4 корпусами по Ляшевская & Шаров (2009)

	Худ50–60	Худ70–80	Публ50–60	Публ70–80
Худ50–60	0	0.012	0.097	0.097
Худ70–80	0.012	0	0.121	0.085
Публ50–60	0.097	0.121	0	0.036
Публ70–80	0.097	0.085	0.036	0

Оценка мер расстояния: Known-Similarity Corpora

- Kilgarriff (2001): Known-Similarity Corpora
- Берутся два исходных корпуса A и B , которые априорно считаются существенно различными.
- Эти корпуса нарезаются на куски одинакового размера, из которых составляются корпуса известной степени сходства (Known-Similarity Corpora)

Оценка мер расстояния: Known-Similarity Corpora



C_0	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}
C_1	A_1	B_{11}	B_{12}	B_{13}	B_{14}	B_{15}	B_{16}	B_{17}	B_{18}	B_{19}
C_2	A_2	A_3	B_{20}	B_{21}	B_{22}	B_{23}	B_{24}	B_{25}	B_{26}	B_{27}
C_3	A_4	A_5	A_6	B_{28}	B_{29}	B_{30}	B_{31}	B_{32}	B_{33}	B_{34}
C_4	A_7	A_8	A_9	A_{10}	B_{35}	B_{36}	B_{37}	B_{38}	B_{39}	B_{40}
C_5	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	B_{41}	B_{42}	B_{43}	B_{44}	B_{45}
C_6	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}	A_{21}	B_{46}	B_{47}	B_{48}	B_{49}
C_7	A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	A_{27}	A_{28}	B_{50}	B_{51}	B_{52}
C_8	A_{29}	A_{30}	A_{31}	A_{32}	A_{33}	A_{34}	A_{35}	A_{36}	B_{53}	B_{54B}
C_9	A_{37}	A_{38}	A_{39}	A_{40}	A_{41}	A_{42}	A_{43}	A_{44}	A_{45}	B_{55}
C_{10}	A_{46}	A_{47}	A_{48}	A_{49}	A_{50}	A_{51}	A_{52}	A_{53}	A_{54}	A_{55}

Оценка мер расстояния: Known-Similarity Corpora

- На наборе из 11 KSC:
660 сравнений $d(C_a, C_b) ? d(C_c, C_d)$, где $c \leq a < b \leq d$, при этом $a \neq c$ или $b \neq d$
- Мера расстояния хорошо отражает реальность, если даёт правильный знак в большой доле из этих 660 случаев.
- Доля правильных знаков и есть оценка меры.

Оценка мер расстояния: Known-Similarity Corpora

- Kilgarriff (2001): сравниваются несколько мер на основе перплексивности, χ^2 и Спирмен.
- Материал исследования: British National Corpus (BNC).
- Берутся подкорпуса из обильно представленных источников; из них составляются наборы из 11 KSC объемом по 100 000 слов.

Подкорпуса BNC (Kilgarriff 2001)

- Accountancy (acc)
- The Art Newspaper (art)
- British Medical Journal (bmj)
- Environment Digest (env)
- The Guardian (gua)
- The Scotsman (sco)
- Today (tod)

Kilgarriff 2001: результаты

Table 7. Comparison of four measures

	spear	χ^2	closed	type 1	type 2
KSC-set					
acc_gua	93.33	91.33	82.22	81.11	80.44
art_gua	95.60	93.03	84.00	83.77	84.00
bmj_gua	95.57	97.27	88.77	89.11	88.77
env_gua	99.65	99.31	87.07	84.35	86.73

Table 8. Spearman/ χ^2 comparison on all KSCs

	spear	χ^2	tie	total
Highest score	5	13	3	21

- closed, type1 и type2 — различные варианты меры на основе перплексивности.
- Победитель — χ^2 .

Более новые меры

- Сходство по ключевым словам
(Keyword Similarity; КС; Kilgarrriff 2009;
SketchEngine)
- Сумма минимальных частот
(СМЧ; Шайкевич 2015)

Мера 3: сходство по ключевым словам (Kilgarriff 2009; SketchEngine)

- Берём по топ-5000 частотных слов из каждого из корпусов
- Для каждого из слов в объединении этих множеств вычисляем, насколько оно характерно для одного или другого корпуса.
- $$k(w) = \max \left(\frac{f_A(w_i)+n}{f_B(w_i)+n}, \frac{f_B(w_i)+n}{f_A(w_i)+n} \right)$$
- Берём среднее значение k по N наиболее характерным словам.

КС: сравнение 4 английских корпусов в SketchEngine ($n = 0.0001$)

	(B A S E)	(B A W E)	B r o w n	(E E B O)
British Academic Spoken English Corpus (BASE)	1.00	<u>3.28</u>	<u>3.12</u>	<u>4.20</u>
British Academic Written English Corpus (BAWE)	<u>3.28</u>	1.00	<u>2.21</u>	<u>3.45</u>
Brown	<u>3.12</u>	<u>2.21</u>	1.00	<u>2.80</u>
Early English Books Online (EEBO)	<u>4.20</u>	<u>3.45</u>	<u>2.80</u>	1.00

Reference corpus: **Early English Books Online (EEBO)**

[Switch focus and reference \(sub\)corpus](#)

Page [Next >](#)

<i>British Academic Spoken English Corpus (BASE)</i>			<i>Early English Books Online (EEBO)</i>		Score
word	frequency	frequency/mill ?	frequency	frequency/mill	
i	15,450	12337.7	145,195	147.1	50.3
okay	3,552	2836.5	0	0.0	29.4
n't	6,681	5335.2	89,115	90.3	28.6
just	4,066	3246.9	84,060	85.1	18.1
going	3,965	3166.3	94,406	95.6	16.7
actually	2,407	1922.1	26,344	26.7	16.0
got	3,037	2425.2	87,386	88.5	13.4
yeah	1,414	1129.2	0	0.0	12.3
sort	2,560	2044.3	74,752	75.7	12.2
get	2,783	2222.4	121,150	122.7	10.4
really	1,806	1442.2	50,276	50.9	10.2
something	1,687	1347.2	41,626	42.2	10.2

Мера 3: сходство по ключевым словам (Kilgarriff 2009; SketchEngine)

- Диапазон: от 1 (абсолютно идентичные частотные списки) и более.
- NB: плохо удовлетворяет пожеланиям к мерам расстояния, в частности ограниченности диапазона и тождеству.

Мера 4:

сумма минимальных частот

- Строим частотные списки корпусов A и B .
- Выбираем топ- N слов из объединённого частотного списка для двух корпусов.
- Для каждого слова находим минимальную из двух частот и среднюю частоту
- Делим сумму минимальных частот на сумму средних частот
- $$SMF = \frac{\sum \min(f_A(w_i), f_B(w_i))}{\sum 0,5(f_A(w_i) + f_B(w_i))}$$

Художественная литература vs. публицистика (Шайкевич 2015)

- Источник данных:
Ляшевская & Шаров (2009)

	ХЛ	Пуб.	min		ХЛ	Пуб.	min
и	.0365	.0355	.0355	на	.0167	.0167	.0167
в	.0262	.0364	.0262	я	.0183	.0104	.0104
не	.0223	.0169	.0169	Σ	.1200	.1159	.1057
			C = .1057/.1180				= .896

Лексическое сходство жанров НКРЯ (Шайкевич 2015)

	Устная речь	Худож. лит-ра	Публи- цистика	Прочее
Устная речь		.751	.667	.670
Худож. лит-ра	.661		.892	.852
Публи- цистика	.561	.716		.933
Прочее	.527	.700	.859	

Справа сверху: слова с 1-го по 100-е

Слева снизу: слова со 101-го по 500-е

Мера 4:

сумма минимальных частот

- Если взять все слова из обоих корпусов, есть простая вероятностная интерпретация: ожидаемая доля пересечений в достаточно больших выборках из двух корпусов.
- Диапазон: от 0 (непересекающиеся списки слов) до 1 (идентичные частотные списки).
- Мера расстояния: $1 - SMF$

Сравнение мер

- Воспроизводим эксперимент Килгарриффа, но с бóльшим количеством мер и на большем количестве корпусов.
- Два этапа:
 - подбор параметров
 - собственно тестирование мер

Материал исследования

- 19 подкорпусов BNC объемом $\geq 550\,000$ токенов
- 9 случайно выбранных подкорпусов используются на этапе подбора параметров, 10 подкорпусов — на этапе тестирования мер.

acc	Accountancy	598,379
art	The Art Newspaper	696,904
bel	The Belfast Telegraph	847,445
bio	The Dictionary of National Biography: Missing persons	773,750
gua	The Guardian	984,553
gut	Gut: Journal of Gastroenterology and Hepatology	808,792
han	Hansard Extracts	1,234,630
ind	The Independent	1,131,300
kee	Keesings Contemporary Archives	2,875,490
law	The Weekly Law Reports	721,892
liv	Liverpool Daily Post and Echo	921,162
mir	The Daily Mirror	833,598
new	Central television news scripts	1,332,831
nor	Northern Echo	1,272,118
nsc	New Scientist	980,897
sco	The Scotsman	1,394,751
tel	Daily Telegraph	1,319,936
tod	Today	1,419,457
uni	Unigram X	609,537

Подбор параметров

- Для χ^2 , Спирмена и СМЧ — количество слов, участвующих в сравнении.
- Для КС — количество отбираемых ключевых слов и константа, добавляемая к числителю и знаменателю.

Подбор параметров: χ^2

<i>N</i>	Медиана	Ср. арифм.		<i>N</i>	Медиана	Ср. арифм.
10	97,27	94,65		316	99,39	97,12
15	95,98	94,63		464	99,09	97,09
22	97,50	95,13		681	99,02	96,98
32	98,26	95,37		1000	98,94	96,74
46	99,24	96,01		1468	98,94	96,73
68	99,32	96,38		2154	98,71	96,74
100	99,47	96,68		3162	98,64	96,72
147	99,47	96,99		Все	97,73	96,01
215	99,47	97,19				

Подбор параметров: результаты

	Параметры	Медиана	Ср. арифм.
χ^2	$N = 215$	99,47	97,19
Спирмен	$N = 215$	99,09	96,05
КС	$N = 68, n = 0.01$	99,62	97,12
СМЧ	$N = 215$	99,47	96,95

Тестирование

[illegible]

M	31
X	14
K	13
S	3

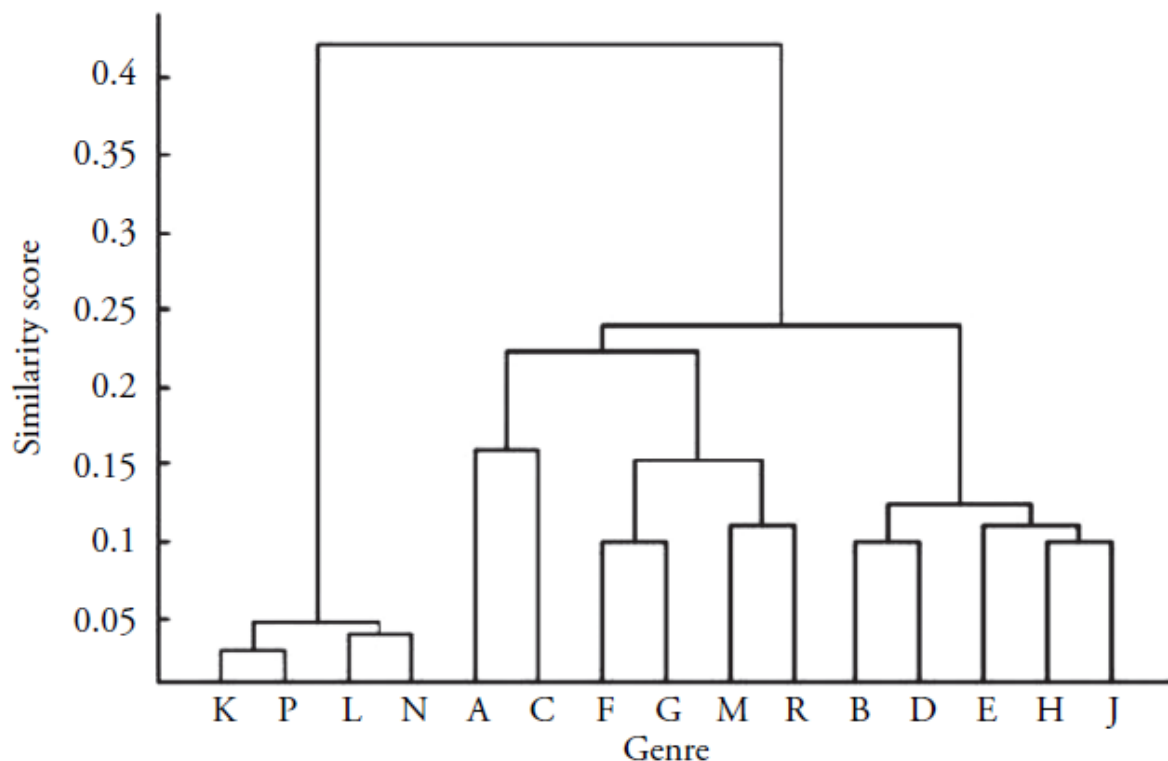
Тестирование (20 прогонов × 45 пар корпусов)

• СМЧ	622 победы	69%
• χ^2	320 побед	36%
• КС	310 побед	34%
• Спирмен	112 побед	12%

Что дальше делать с расстояниями?

- Расстояния имеют смысл, только если у нас имеется ≥ 3 корпуса, которые мы сравниваем.
- На основе расстояний можно сделать кластеризацию.

Жанры в корпусе британского английского LOB (Oakes 2009)



A: Press: Reportage
B: Press: Editorial
C: Press: Reviews
D: Religion
E: Skills & Hobbies
F: Popular Lore
G: Belles Lettres, etc.
H: Miscellaneous
J: Learned
K: General Fiction
L: Mystery and
Detective Fiction
M: Science Fiction
N: Adventure and
Western Fiction
P: Romance and Love
R: Humor

Жанры в Брауновском корпусе американского английского (Шайкевич 2015)

[illegible]

Жанры в американском английском и в китайском (Шайкевич 2015)

- Брауновский корпус:
 - «деловой кластер» — ABCDEFGJ
 - «литературный кластер» — KLMNP
- Ланкастерский корпус китайского языка:
 - «деловой кластер» — BCDEFG
 - «литературный кластер» — KLMNP + R: Humor
 - между ними: A: Press. Reportage и G. Belles Lettres, Biography, Memoirs, etc.

Case study: лексическое сходство русских поэтов

- Поэтический корпус в составе Национального корпуса русского языка (www.ruscorpora.ru)
- 872 автора
- Общий объём текстов: 11 млн словоформ



[главная](#)

[архив новостей](#)

[поиск в корпусе](#)

[что такое корпус?](#)

[состав и структура](#)

[статистика](#)

[графики](#)

[частоты](#)

[морфология](#)

[обороты](#)

[синтаксис](#)

[семантика](#)

[параметры текстов](#)

[о проекте](#)

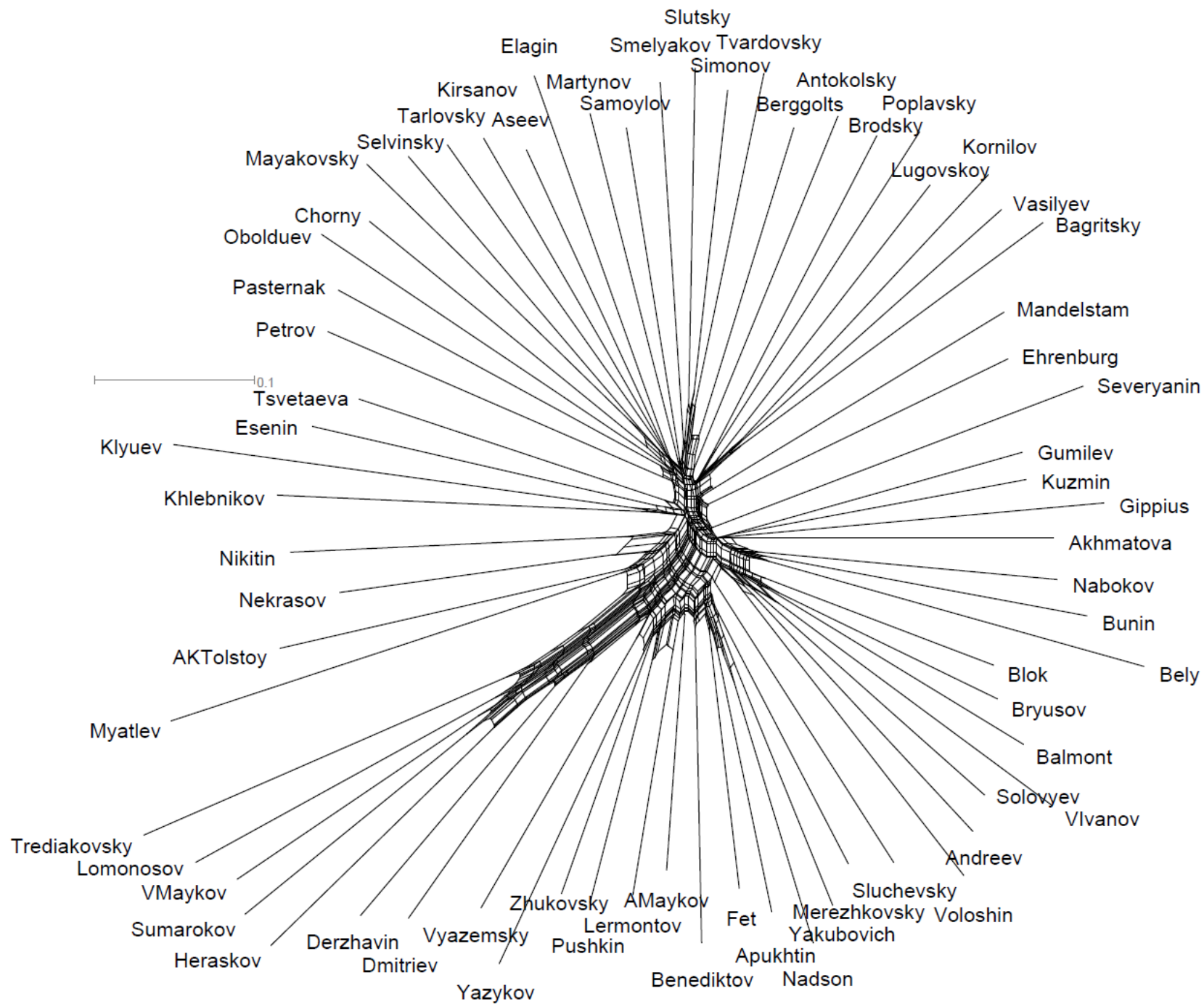
[участники проекта](#)

Список авторов поэтического корпуса

Автор ▲	Год рождения	Год смерти
А. А. Аблесимов	1740	1783
Н. Я. Агнивцев	1888	1932
А. Е. Адалис	1900	1969
Г. В. Адамович	1892	1972
М. Н. Айзенберг	1948	
И. А. Аксенов	1884	1935
В. Д. Александровский	1887	1934
Л. А. Алексеева	1909	1989
Н. Н. Алл	1890-е	после 1966
Б. Н. Алмазов	1827	1876
Дж. Алтаузен	1907	1942
С. Я. Алымов	1892	1948
А. Альвинг	1885	1942
С. А. Алякринский	1889	1938
А. Н. Аммосов	1822	1866

Материал и метод исследования

- 69 поэтов за всю историю русской литературы, для которых в Национальном корпусе русского языка представлено более чем 50 тыс. словоформ.
- Лемматизация с помощью `mystem` с автоматическим снятием омонимии
- Сравнение с помощью СМЧ без ограничения по количеству слов
- Визуализация: NeighborNet (Bryant & Moulton 2004)



Наблюдения

- Сеть хорошо отражает представления, привычные нам по традиционной истории литературы.
- Вероятно, этот метод можно применить и там, где у нас нет таких сильных априорных представлений.

Материал исследования (2)

- 64 поэтов, которые родились в 1890 году или позже и для которых в Национальном корпусе русского языка представлено более чем 20 тыс. словоформ.



Наблюдения

- Выделяются некоторые известные группы авторов:
 - обэриуты: Хармс, Введенский, Заболоцкий
 - имажинисты: Шершеневич, Мариенгоф, Есенин
 - советская романтика: Тихонов, Багрицкий (но не Светлов)
 - эмигранты: Г. Иванов, Набоков, Адамович, Чиннов
 - Маяковский + Кирсанов
 - Липкин + Тарковский
 - ...



Что лучше сравнивать?

- Во всех экспериментах, описанных в первой части доклада, сравнивались словоформы.
- В поэтическом эксперименте сравнивались леммы.
- Может быть, надо сравнивать что-то другое?
 - не униграммы, а биграммы?
 - не слова, а символы?
 - ...

Что лучше сравнивать: (предварительный) эксперимент

- Материал: 6 подкорпусов BNC, взятых в Kilgarriff (2001)
- Мера: χ^2 на топ-400 элементах
- Уровни анализа:
 - символы
 - словоформы
 - леммы
 - части речи (глагол, существительное, etc.)
 - части речи CLAWS5 (более дробные пометы)
- От 1-грамм до 3-грамм

Что лучше сравнивать: (предварительные) результаты

Место	Единицы	<i>n</i> -граммы	Медиана	Ср. арифм.
1	символы	2	98,64	97,56
2	леммы	1	98,48	97,29
3	символы	3	98,03	97,52
4	CLAWS5	3	98,03	95,99
5	слова	1	97,58	96,61

Заключение (1)

- Меры сравнения корпусов на основе частотных словарей позволяют получать адекватные реальности оценки расстояния между корпусами.
- Рейтинг 4 проанализированных мер:
 1. СМЧ Шайкевича;
 - 2–3. χ^2 , КС Килгарриффа;
 4. Спирмен.

Заключение (2)

- На основе вычисленных расстояний между корпусами можно делать содержательные выводы об их кластеризации.
- Наилучшим по качеству уровнем для анализа, хотя и не самым лингвистически содержательным, могут оказаться символные n -граммы.

Вопросы на будущее

- Математические свойства мер?
- Устойчивость мер к размеру корпуса?
- Наилучшие уровни для анализа для разных языков?

Спасибо
за внимание!