Измерение расстояний между текстами

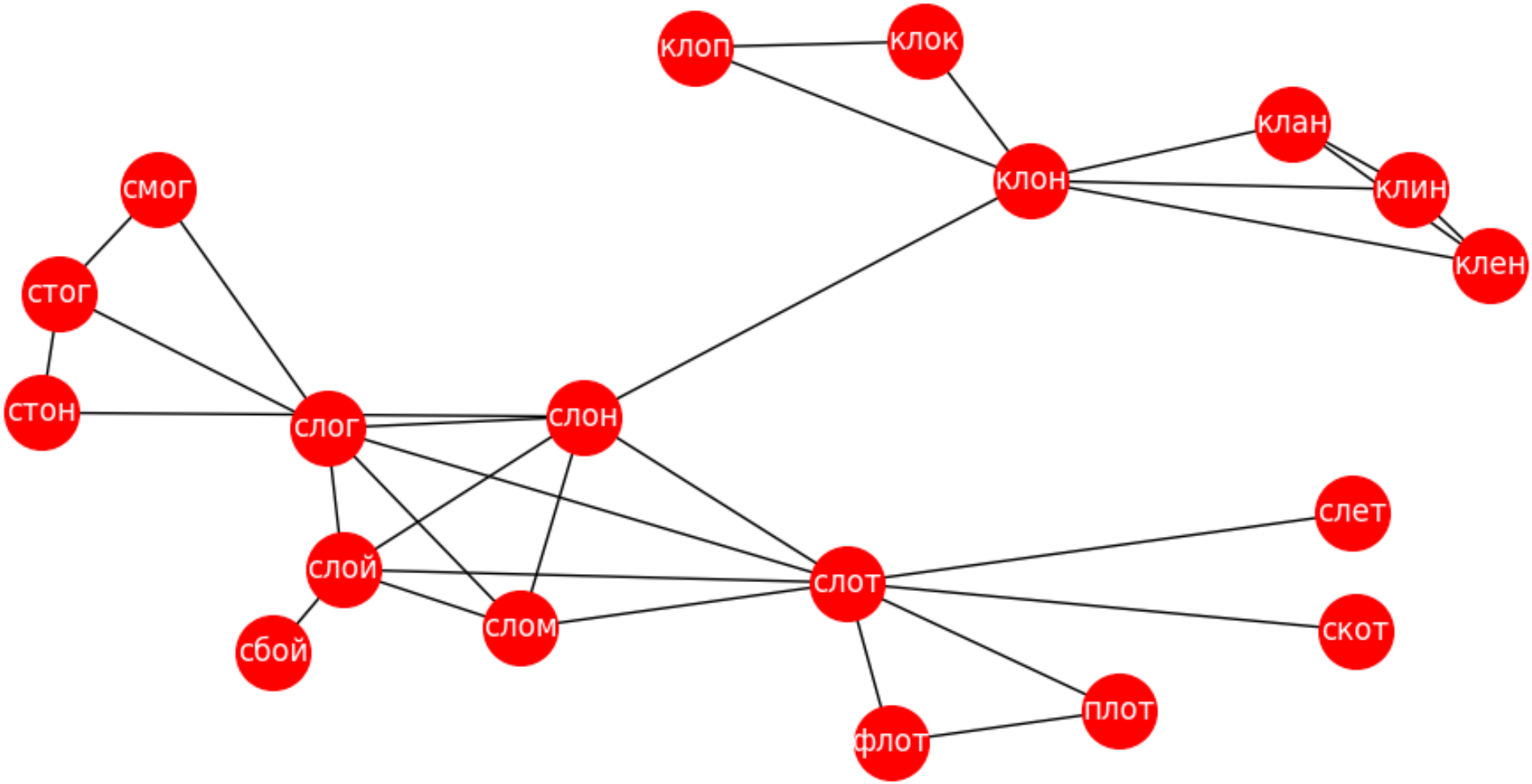
Александр Пиперски

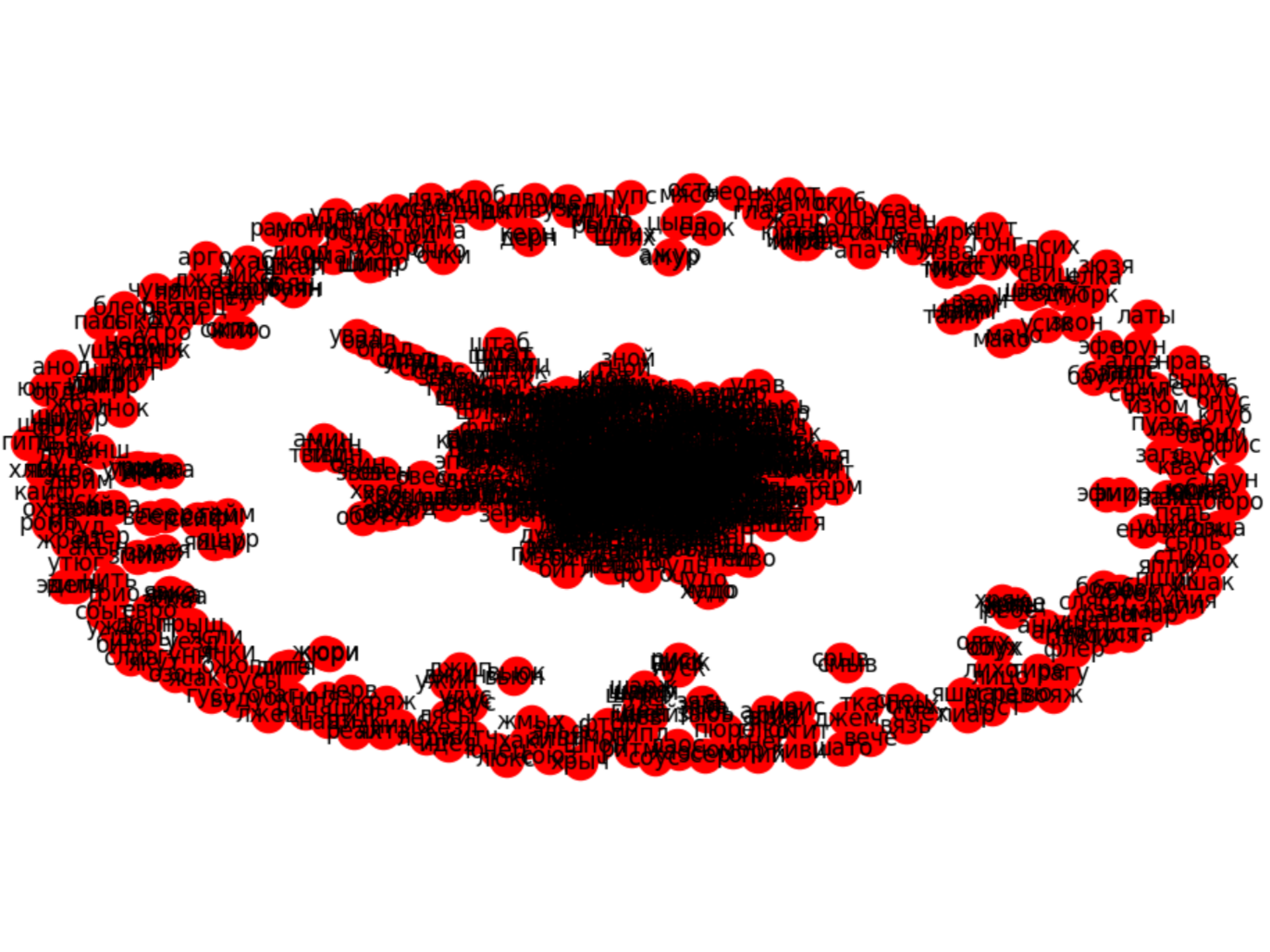
Малый мехмат, 10.02.2018

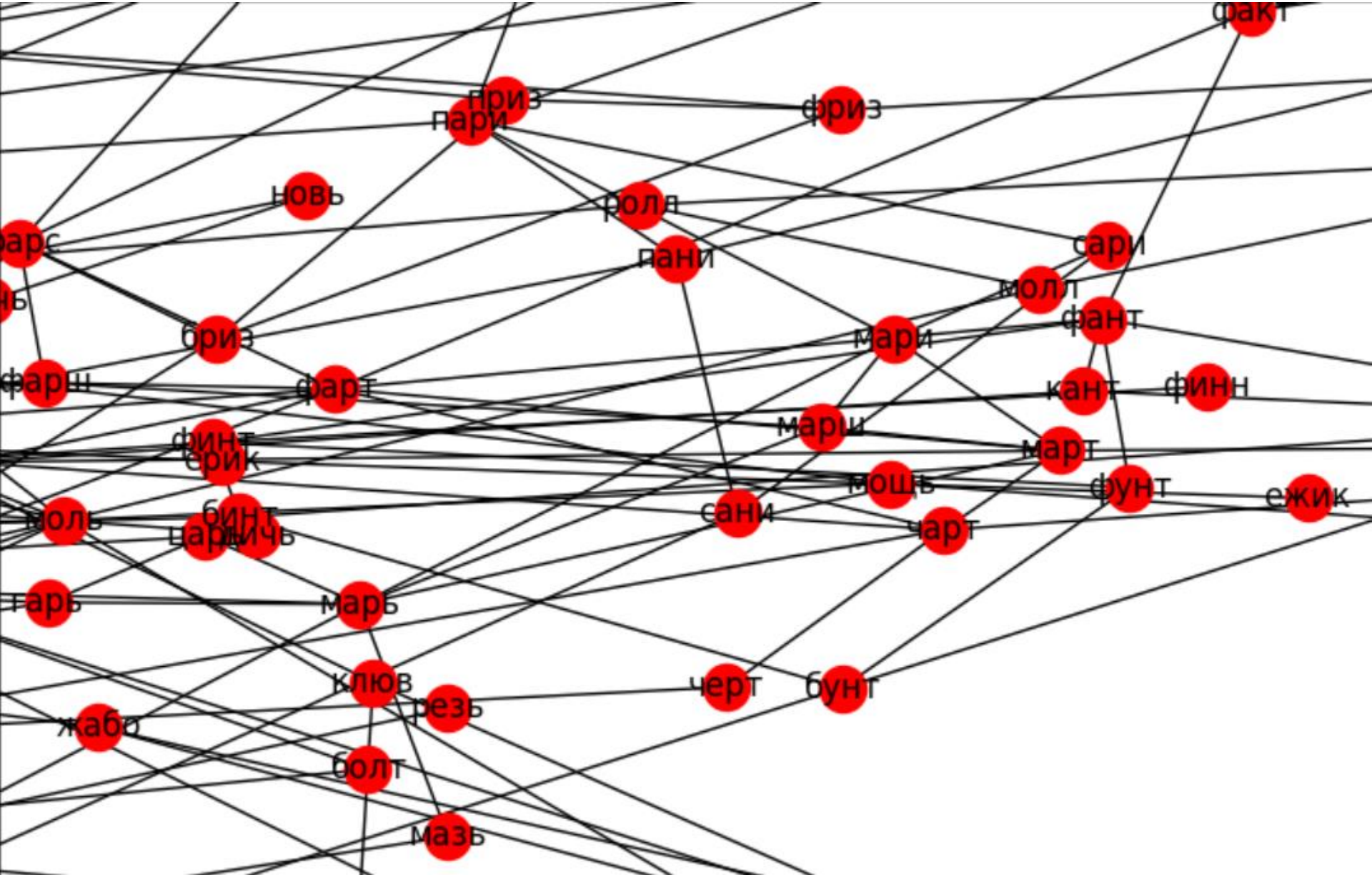
Как сделать из мухи слона

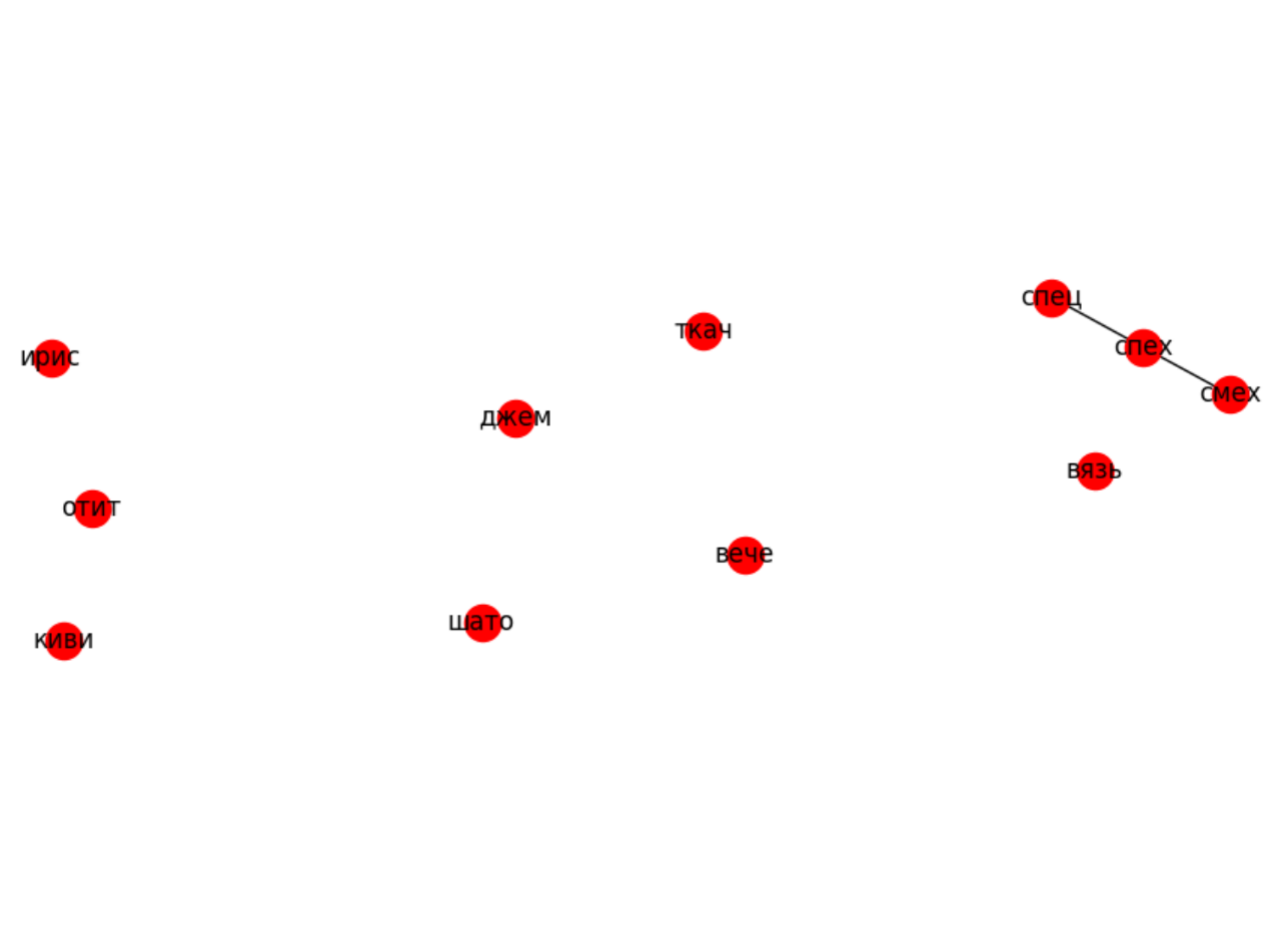
- Задача: превратить одно русское нарицательное существительное в другое (напр., *стол* ⟶ *стул*, *фарш* ⟶ *поза*, *флот* ⟶ *клоп*, *муха* ⟶ *слон* и т. п.)
- На каждом шаге можно заменять одну букву на другую так, чтобы получалось существующее русское нарицательное существительное
- Назовём допустимыми слова, вошедшие в частотный словарь О. Ляшевской и С. Шарова (2009)

Подграф вокруг слова *слон*






факт
приз
фриз
пари
новь
ролл
сари
пани
молл
фант
бриз
мари
фарш
фарт
кант
финн
финт
марш
марк
мощь
фунт
ежик
моль
сани
чарт
марь
клюв
черт
бунт
резь
жабо
болт
мазь


спец
ткач
спех
ирис
смех
джем
вязь
отит
вече
шато
киви

Некоторые свойства графа

- 1221 вершина
- Компоненты связности: 551 вершина, 309 вершин, 6 вершин, 5 вершин, ...
- 218 вершин имеют степень 0 (= слова не имеют соседей: *атом*, *бюро*, *дядя*, *евро*)
- 28 пар слов, связанных только друг с другом (*глаз – глас*, *цирк – цинк*)

Расстояния между словами

- На этом графе можно измерять длины путей из одной вершины в другую
- Длина самого короткого пути ≈ сложность игры для данной пары слов

Расстояния между словами

- d_{min}(*кора*, *панк*) = 4
  кора, *кара*, *пара*, *парк*, *панк*
- d_{min}(*тмин*, *штаб*) = 22
  тмин, *твин*, *овин*, *овен*, *овес*, *свес*, *свет*, *слет*, *слот*, *слон*, *клон*, *клан*, *кран*, *крап*, *храп*, *храм*, *хлам*, *шлам*, *шлак*, *шпак*, *шпат*, *штат*, *штаб*
- d_{min}(*муха*, *слон*) = ∞
  (*муха* и *слон* принадлежат двум разным компонентам связности)

муха и *слон*

- По другому словарю («Грамматический словарь русского языка» А. Зализняка, 1713 слов, самая большая компонента связности — 1361 слово):
- *муха*, *мура*, *мара*, *пара*, *парк*, *паек*, *саек*, *стек*, *стен*, *стон*, *слон*

Расстояние Хэмминга

- **Расстояние Хэмминга** — число позиций, в которых символы слов одинаковой длины различны
- `флот` `муха`
  `клоп` `слон`
- $d_{Hamming}(\textit{флот}, \textit{клоп}) = 2$
- $d_{Hamming}(\textit{муха}, \textit{слон}) = 4$

Расстояние Хэмминга: проблемы

- `муха`
  `ухаб`
- $d_{Hamming}(муха, ухаб) = 4$

- $d_{Hamming}(муха, мушка) = ???$

Расстояние Хэмминга: усовершенствования

- `му␣ха` `мух␣а` `␣муха`
  `мушка` `мушка` `мушка`
- $d_{Hamming}(му␣ха, мушка) = 2$
- $d_{Hamming}(мух␣а, мушка) = 2$
- $d_{Hamming}(␣муха, мушка) = 5$
- Разрешаем добавлять символ ␣ оптимальным (минимизирующим расстояние) образом

Расстояние Левенштейна

- `муха␣`
  `␣ухаб`
- $d_{Hamming}(муха␣, ␣ухаб) = 2$
- **Расстояние Левенштейна** — минимальное количество редакционных операций (вставок, удалений и замен символов), необходимых, чтобы преобразовать одну строку в другую

Владимир Иосифович Левенштейн (1935–2017)

- «Двоичные коды с исправлением выпадений, вставок и замещений символов» (1965)
- 9317 цитирований в Google Scholar на 07.02.2018 — самая цитируемая работа, написанная советским / российским математиком!

Расстояние Левенштейна

- $d_{Levenshtein}$(*лакрица*, *планетарий*) = ?
- ␣ла␣␣␣крица
  планетарий␣
- $d_{Levenshtein}$(*лакрица*, *планетарий*) =
  $d_{Hamming}$(␣*ла*␣␣␣*крица*, *планетарий*␣) = 7

Вычисление расстояния Левенштейна

	ε	п	л	а	н	…	…	…	…	…	…
ε	0	1	2	3	4	…	…	…	…	…	…
л	1	1	1	2	3	…	…	…	…	…	…
а	2	2	2	1	2	…	…	…	…	…	…
к	3	3	3	2	2	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…

Расстояние Левенштейна между текстами

- *У попа была собака*
- *У Алевтины Ивановны была кошка*
- Можно считать единичными символами слова, а словами — тексты
- Не очень удобно на практике, поскольку сходство на самом деле так не работает

- *У попа была собака, он её любил. Она съела кусок мяса, он её убил.*
- *Я вас любил: любовь еще, быть может, в душе моей угасла не совсем; но пусть она вас больше не тревожит; я не хочу печалить вас ничем.*
- *Я больше люблю собак, чем кошек, но у меня нет собаки, потому что мне не хватает денег, чтобы кормить её мясом.*

Расстояние между корпусами

- Задача: найти меру $d(A, B)$, которая позволит оценить, насколько похожи между собой тексты / 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?
  - На что больше похожи стихи Мандельштама: на стихи Ахматовой, Блока или Гумилёва?

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 (ср. обучение с противником)

Частотные списки

- Проще подступиться к задаче сравнения корпусов между собой, если сравнивать не корпуса, а их частотные списки

она	3		вы	4		я	3
он	2		не	3		нет	2
быть	1		я	2		собака	2
кусок	1		больше	1		больше	1
любить	1		быть	1		деньги	1
мясо	1		душа	1		кормить	1
поп	1		еще	1		кошка	1
собака	1		любить	1		любить	1
съесть	1		любовь	1		мясо	1
у	1		мой	1		но	1
убить	1		мочь	1		она	1
ВСЕГО	14		ничто	1		потому	1
			но	1		у	1
			она	1		хватать	1
			печалить	1		чем	1
			пусть	1		что	1
			совсем	1		чтобы	1
			тревожить	1		ВСЕГО	21
			угаснуть	1			
			хотеть	1			
			ВСЕГО	26			

Нормированные частотные списки

- Полученные списки плохи тем, что числа в них зависят от размера текста
- Нормируем их — например, приведя к процентам (в корпусной лингвистике обычно приводят к единицам на миллион, ipm)

она	3	21		вы	4	15		я	3	14
он	2	14		не	3	12		нет	2	10
быть	1	7		я	2	8		собака	2	10
кусок	1	7		больше	1	4		больше	1	5
любить	1	7		быть	1	4		деньги	1	5
мясо	1	7		душа	1	4		кормить	1	5
поп	1	7		еще	1	4		кошка	1	5
собака	1	7		любить	1	4		любить	1	5
съесть	1	7		любовь	1	4		мясо	1	5
у	1	7		мочь	1	4		но	1	5
убить	1	7		мой	1	4		она	1	5
ВСЕГО	14	98		ничто	1	4		потому	1	5
				но	1	4		у	1	5
				она	1	4		хватать	1	5
				печалить	1	4		чем	1	5
				пусть	1	4		что	1	5
				совсем	1	4		чтобы	1	5
				тревожить	1	4		ВСЕГО	21	104
				угаснуть	1	4				
				хотеть	1	4				
				ВСЕГО	26	103				

Для простоты пренебрежём погрешностью округления

	C_1	C_2	C_3			C_1	C_2	C_3
больше	0	4	5		она	21	4	5
быть	7	4	0		печалить	0	4	0
вы	0	15	0		поп	7	0	0
деньги	0	0	5		потому	0	0	5
душа	0	4	0		пусть	0	4	0
еще	0	4	0		собака	7	0	10
кормить	0	0	5		совсем	0	4	0
кошка	0	0	5		съесть	7	0	0
кусок	7	0	0		тревожить	0	4	0
любить	7	4	5		у	7	0	5
любовь	0	4	0		убить	7	0	0
мой	0	4	0		угаснуть	0	4	0
мочь	0	4	0		хватать	0	0	5
мясо	7	0	5		хотеть	0	4	0
не	0	12	0		чем	0	0	5
нет	0	0	10		что	0	0	5
ничто	0	4	0		чтобы	0	0	5
но	0	4	5		я	0	8	14
он	14	0	0					

Частотные списки как векторы

- Каждому корпусу теперь соответствует упорядоченная последовательность чисел — **вектор**
- C_2: (4; 4; 15; 0; 4; 4; 0; ...; 0; 8)
- Как можно измерять расстояния между такими векторами?

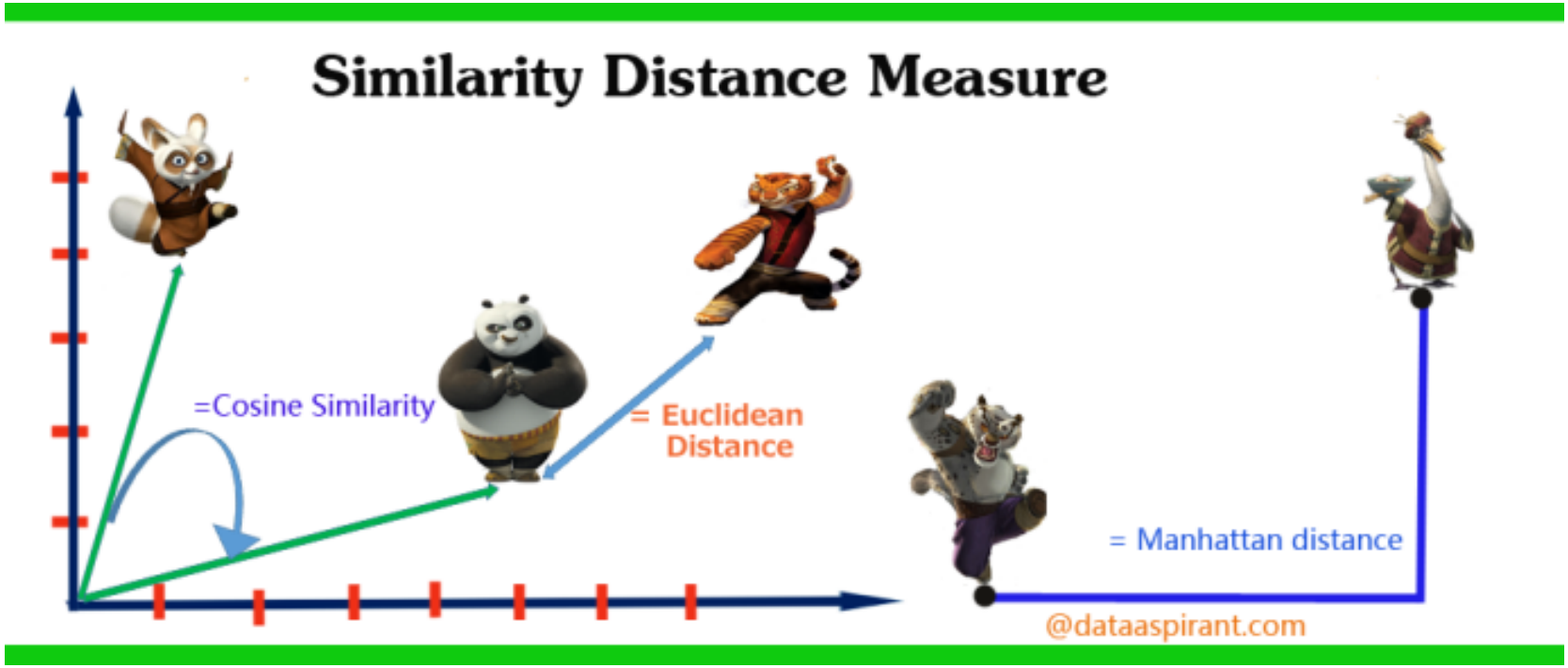
Геометрические меры

1. Евклидово расстояние
2. Манхэттенское расстояние

(1–2. Расстояние Минковского)

3. Косинусное расстояние

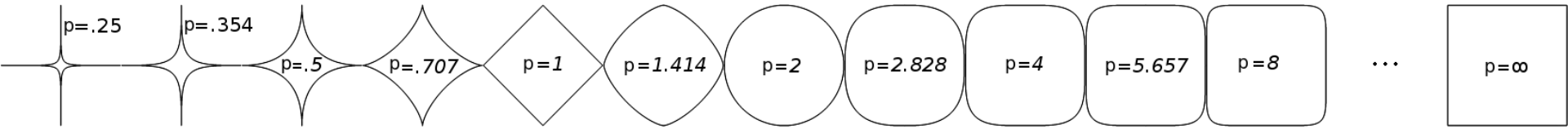
Геометрические меры



Евклид, Манхэттен, Минковский

- Евклидово расстояние:
  $$D(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$
- Манхэттенское расстояние:
  $$D(x,y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$$
- Расстояние Минковского:
  $$D(x,y) = \sqrt[p]{|x_1 - y_1|^p + |x_2 - y_2|^p + \cdots + |x_n - y_n|^p}$$

Единичная окружность при раличных *p* Минковского



Косинусное расстояние

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

distance = 1 – similarity

Статистические меры

4. χ^2
5. Коэффициент Спирмена

Мера 4: 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) × 2 (добавив строку для всех остальных слов) и вычисляем χ^2 обычным образом

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

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

	Худ50-60	Публ50-60	Σ
и	0.038	0.037	0.075
в	0.024	0.035	0.059
не	0.021	0.016	0.037
на	0.017	0.015	0.032
я	0.017	0.014	0.031
быть	0.012	0.014	0.026
он	0.020	0.012	0.032
с	0.011	0.011	0.022
что	0.008	0.008	0.016
а	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	
и	7×10^{-6}	7×10^{-6}	
в	0.0010	0.0010	
не	0.0003	0.0003	
на	6×10^{-5}	6×10^{-5}	
я	0.0001	0.0001	
быть	7×10^{-5}	7×10^{-5}	
он	0.0010	0.0010	
с	0	0	
что	0	0	
а	0.0005	0.0005	
ПРОЧЕЕ	3×10^{-5}	3×10^{-5}	
			0.0064

Мера 4: 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 2 (абсолютно разные частоты; *кстати, какие?*)

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 5: Спирмен

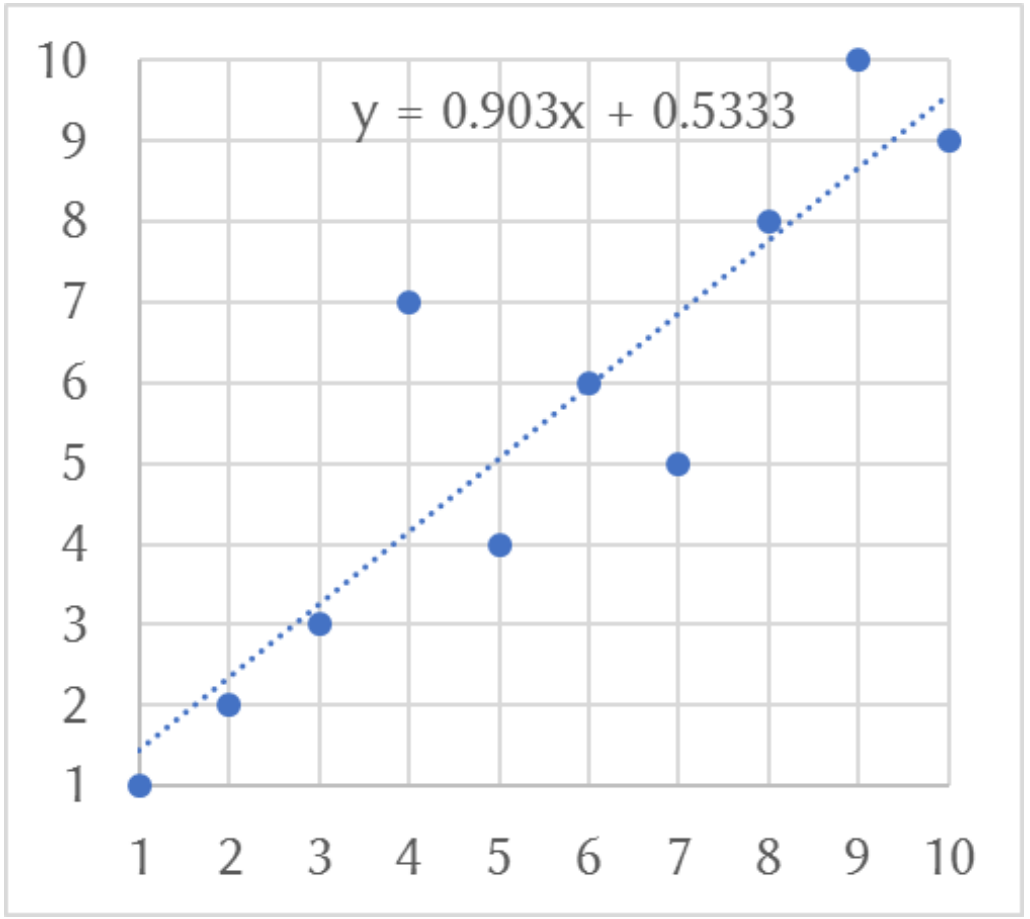
- Строим частотные списки корпусов A и B
- Выбираем топ-N слов из объединённого частотного списка для двух корпусов
- Определяем ранги слов в частотном списке, включающем в себя эти N слов, для A и в таком же списке для B и вычисляем коэффициент корреляции Спирмена r_s обычным образом исходя из разностей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

$$r_s = 1 - \frac{6 \sum D_i^2}{N(N^2 - 1)}$$

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

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

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



Мера 5: 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

Мера 6: 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
- $k(w) = \max\left(\frac{f_A(w)+n}{f_B(w)+n}, \frac{f_B(w)+n}{f_A(w)+n}\right)$
- Берём среднее значение k по N наиболее характерным словам

Reference corpus: **Early English Books Online (EEBO)**
Switch focus and reference (sub)corpus

Page 1 Go Next >

	British Academic Spoken English Corpus (BASE)		*Early English Books Online (EEBO)*		
word	**frequency**	**frequency/mill**	**frequency**	**frequency/mill**	**Score**
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

КС: сравнение 4 английских корпусов в SketchEngine (n = 100 / 1000000 = 0.0001)

	(BASE)	(BAWE)	Brown	(EEBO)
British Academic Spoken English Corpus (BASE)	1.00	3.28	3.12	4.20
British Academic Written English Corpus (BAWE)	3.28	1.00	2.21	3.45
Brown	3.12	2.21	1.00	2.80
Early English Books Online (EEBO)	4.20	3.45	2.80	1.00

Мера 6: сходство по ключевым словам (Kilgarriff 2009; SketchEngine)

- Диапазон: от 1 (абсолютно идентичные частотные списки) и более

Пожелания к мерам расстояния между корпусами (1)

- Мера должна являться метрикой (в математическом смысле)
  - Тождество: $d(A, A) = 0$
    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**

Оценка мер расстояния: Known-Similarity Corpora (KSC)

- Kilgarriff 2001
- 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_{54}
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 $a \neq$ с 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

Kilgarriff 2001: сравнение мер

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

- Победитель — χ^2

Сравнение мер

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

Материал исследования

- 20 подкорпусов Британского национального корпуса (BNC) объёмом ≥ 550 000 токенов
- 10 случайно выбранных подкорпусов используются на этапе подбора параметров, 10 подкорпусов — на этапе тестирования мер

ID	Source	Token count	BNC file ID's
acc	Accountancy	598,379	CBT–CBY
art	The Art Newspaper	696,904	CKT–CKY, EBS–EBX
bel	The Belfast Telegraph	847,445	HJ3–HJ4, K29–K35
bio	The Dictionary of National Biography: Missing persons	773,750	GSX–GTH
eco	The Economist	929,886	ABD–ABK, CR7–CRC
gua	The Guardian	984,553	A7S–AAX
gut	Gut: Journal of Gastroenterology and Hepatology	808,792	HU2–HU4, HWS–HWT
han	Hansard Extracts	1,234,630	HHV–HHX
ind	The Independent	1,131,300	A1D–A5X
kee	Keesings Contemporary Archives	2,875,490	HKP–HLT
law	The Weekly Law Reports	721,892	FBS–FE3
liv	Liverpool Daily Post and Echo	921,162	K36–K4M, K97
mir	The Daily Mirror	833,598	CH1–CH3, CH5–CH7
new	Central television news scripts	1,332,831	K1B–K28
nor	Northern Echo	1,272,118	K4N–K55
nsc	New Scientist	980,897	ANX, B71–B7N
sco	The Scotsman	1,394,751	K56–K5M
tel	Daily Telegraph	1,319,936	AH9–AL6
tod	Today	1,419,457	CBC–CBG, CEK–CEP
uni	Unigram X	609,537	CMW–CN0, CS8–CTV

Подбор параметров

- Для Евклидова, манхэттенского и косинусного расстояния, χ^2, Спирмена — количество слов, участвующих в сравнении
- Для ключевых слов — количество отбираемых ключевых слов и константа, добавляемая к числителю и знаменателю

Подбор параметров: χ^2

N	Медиана	Ср. арифм.		N	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.
Евклид	N = 1000	99,17	96,36
Манхэттен	N = 215	99,47	96,95
Косинус	N = все слова	98,86	96,54
χ^2	N = 215	99,47	97,19
Спирмен	N = 215	99,09	96,05
Ключевые	N = 68, n = 0.01	99,62	97,12

Тестирование

...	acc	bio	gua	gut	law	liv	mir	sco	tod	uni
acc		EMCχK	K	χ	χK	K	M	χ	MC	χK
bio			M	EMCK	EMC	Mχ	EMCχSK	EMCK	EMCK	EMCχSK
gua				EC	M	K	E	S	EC	K
gut					M	M	EMC	M	M	E
law						M	M	M	M	M
liv							C	E	S	C
mir								C	E	χK
sco									E	M
tod										M
uni										

E 15
M 24
C 14
χ 9
S 4
K 13

Тестирование
(20 прогонов × 45 пар корпусов)

• Манхэттен	429 побед	48%
• Евклид	415 побед	46%
• Косинус	386 побед	43%
• χ^2	293 победы	33%
• Ключевые	260 побед	29%
• Спирмен	127 побед	14%

Что лучше сравнивать?

- Во всех экспериментах по оценке мер сравнивались словоформы
- Может быть, надо сравнивать что-то другое?
  - не униграммы, а биграммы?
  - не слова, а символы?
  - ...

Что лучше сравнивать: эксперимент

- Материал: 20 подкорпусов BNC
- 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

Частотные списки разных уровней (BNC: British Medical Journal)

word1			**char2**			**tag3**	
,	0.046		*e*␣	0.023		S S S	0.026
the	0.041		*s*␣	0.020		S Prep S	0.025
.	0.028		␣*t*	0.019		S S Pun	0.021
to	0.027		␣*a*	0.015		Adj S S	0.020
and	0.025		*th*	0.014		Prep Art S	0.019
of	0.022		*in*	0.014		S Pun S	0.018
a	0.015		*t*␣	0.013		Art Adj S	0.017
in	0.014		*er*	0.012		Adj S Prep	0.017
...	...		...	...		...	...

Худшие результаты

		char1	char2	char3	word1	word2	word3	lemma1	lemma2	lemma3	tag1	tag2	tag3	claws1	claws2	claws3
184	mir tod	93.64	92.42	69.39	84.55	68.33	75.91	86.36	85.61	51.97	66.67	56.97	60.76	66.97	58.33	69.39
185	gua sco	62.58	74.85	74.09	76.82	86.21	56.82	68.18	71.82	79.09	56.82	66.97	81.52	53.94	76.21	68.94
186	gua tel	78.03	74.7	77.12	64.09	71.82	81.97	71.67	75	82.27	61.36	47.58	46.82	64.24	69.85	72.73
187	ind sco	71.06	89.09	67.42	62.88	76.82	73.48	69.7	71.06	56.97	58.03	57.12	79.24	52.27	81.06	59.85
188	sco tel	79.55	75.91	75.45	77.58	66.67	75.61	66.36	79.09	74.09	60.61	51.97	66.21	53.33	61.52	50.3
189	ind tel	70.45	69.24	73.79	80.3	75.15	59.55	57.73	77.42	84.7	63.79	66.67	56.97	54.7	58.03	64.85
190	gua ind	85.61	67.58	45	53.48	64.85	65.91	55.3	56.06	57.12	36.67	40.3	67.12	54.24	53.48	70.91

Лучшие результаты

		char1	char2	char3	word1	word2	word3	lemma1	lemma2	lemma3	tag1	tag2	tag3	claws1	claws2	claws3
1	bio new	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2	new uni	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
3	bio uni	100	100	100	100	100	99.85	100	100	100	100	100	100	100	100	100
4	bel bio	100	100	100	100	100	100	100	100	100	100	100	99.7	100	100	100
5	bio mir	100	100	100	99.55	100	100	100	100	99.85	100	100	100	100	99.85	100
6	bio eco	100	100	100	100	100	99.85	99.39	100	99.85	100	100	100	100	100	100
7	bio han	100	100	100	99.85	100	99.09	100	100	99.85	100	100	100	100	100	100

Сравнение уровней анализа

Unit of analysis	Ngrams	Median	Mean	Standard deviation	Highest score (% cases)
Characters	2	99.39	97.16	5.31	43%
Characters	3	96.62	96.62	7.01	39%
Lemmata	1	99.09	96.03	7.12	16%
Characters	1	98.94	96.43	5.93	36%
Words	1	98.94	96.10	6.71	22%
CLAWS5	2	98.86	94.61	9.02	21%
PoS	3	98.79	95.10	8.44	22%
Lemmata	2	98.33	95.79	6.59	12%

НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА

главная

основной
синтаксический
газетный
параллельный
обучающий
диалектный
поэтический
устный
акцентологический
мультимедийный
мультипарк
исторический

использование корпуса

Поэтический корпус

терминологический указатель | инструкция | список авторов | задать подкорпус

Поиск точных форм

Слово или фраза

искать | очистить

Лексико-грамматический поиск

Слово | Грамм. признаки выбрать

Доп. признаки выбрать | Семант. признаки выбрать

Расстояние: от 1 до 1

Слово | Грамм. признаки выбрать

Доп. признаки выбрать | Семант. признаки выбрать

искать | очистить

Работы в 2015—2016 г. выполнены при поддержке РГНФ, проект № 15-04-12018 «Развитие специализированных модулей НКРЯ».

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	1823	1866

Источник текстов

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

Частотные словари отдельных авторов

- Что может сказать о творчестве отдельного автора его частотный словарь?
- Будет ли он отличаться от общего частотного словаря и от частотных словарей других авторов?

Частотный словарь Пушкина

и	8137	41535
в	4976	25400
я	4384	22378
он	3590	18325
не	3335	17024
ты	2542	12976
на	2347	11980
с	2217	11317
мой	1566	7994
быть	1419	7243

Частотный словарь Лермонтова

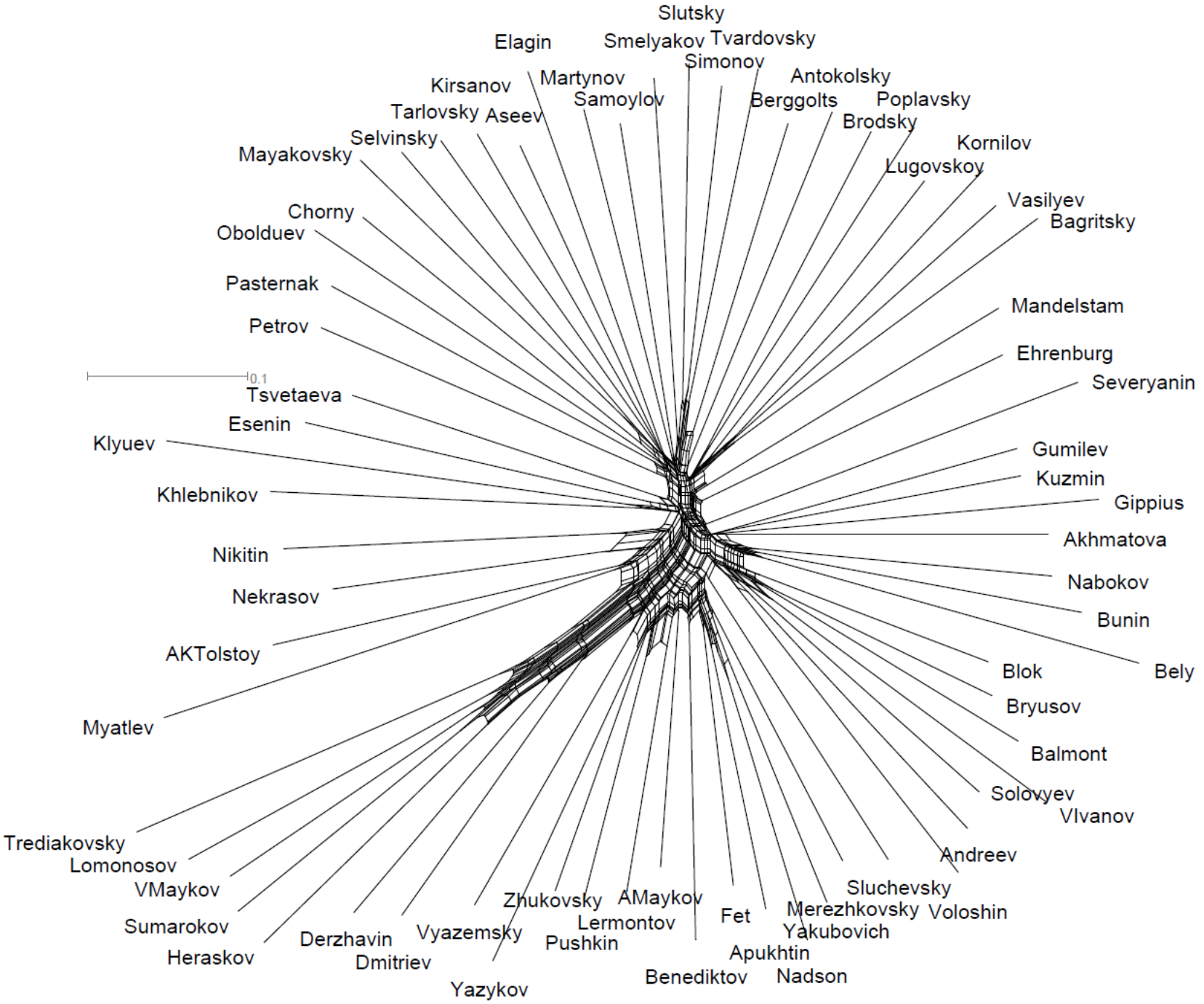
и	6183	46862
я	3522	26694
в	3211	24337
он	2928	22192
не	2820	21373
на	1621	12286
как	1507	11422
быть	1447	10967
ты	1431	10846
с	1321	10012

Матрица расстояний (Манхэттенское расстояние)

	Ахматова	Блок	Фет	А. К. Толстой	…
Ахматова	0	0,36	0,39	0,54	…
Блок	0,36	0	0,30	0,58	…
Фет	0,39	0,30	0	0,54	…
А. К. Толстой	0,54	0,58	0,54	0	…
…	…	…	…	…	…

Материал исследования

- 69 поэтов за всю историю русской литературы, для которых в Национальном корпусе русского языка представлено более чем 50 тыс. словоформ


Slutsky
Elagin
Smelyakov
Tvardovsky
Simonov
Kirsanov
Martynov
Antokolsky
Samoylov
Berggolts
Poplavsky
Tarlovsky
Aseev
Brodsky
Selvinsky
Kornilov
Mayakovsky
Lugovskoy
Vasilyev
Chorny
Bagritsky
Obolduev
Pasternak
Mandelstam
Petrov
Ehrenburg
0.1
Severyanin
Tsvetaeva
Esenin
Gumilev
Klyuev
Kuzmin
Khlebnikov
Gippius
Akhmatova
Nikitin
Nabokov
Nekrasov
Bunin
AKTolstoy
Blok
Bely
Myatlev
Bryusov
Balmont
Solovyev
VIvanov
Trediakovsky
Andreev
Lomonosov
Sluchevsky
VMaykov
Zhukovsky
AMaykov
Merezhkovsky
Voloshin
Sumarokov
Lermontov
Fet
Yakubovich
Derzhavin
Vyazemsky
Pushkin
Heraskov
Apukhtin
Dmitriev
Benediktov
Nadson
Yazykov

Вопрос для размышления

- Как использовать эти меры для атрибуции авторства?