Русские Google-биграммы при исследовании культурных трендов: подходы и наблюдения

Борис Орехов (НИУ ВШЭ, ИРЛИ РАН)

Культуромика

Культуромика

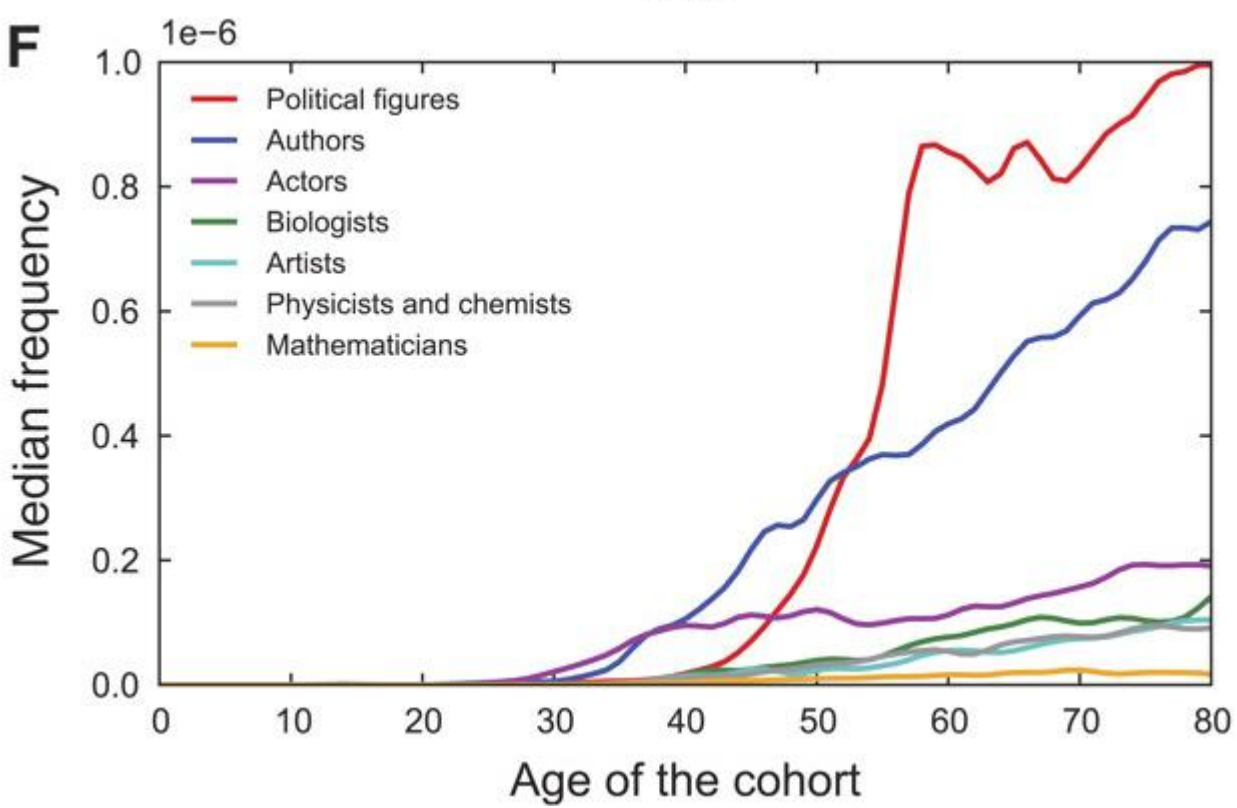
Ключевая идея культуромики состоит в том, что изменения частотностей слов рассматриваются как сигналы (или «следы») культурных и социальных изменений.

- *Бонч-Осмоловская А. А.* Культуромика: исследование культуры и языка с помощью больших текстовых данных // Цифровые гуманитарные исследования. Красноярск: СФУ, 2023, 57-99.
- *Michel, J.-B. et al*. Quantitative Analysis of Culture Using Millions of Digitized Books // Science. 2011. 331 (6014), 176–182.

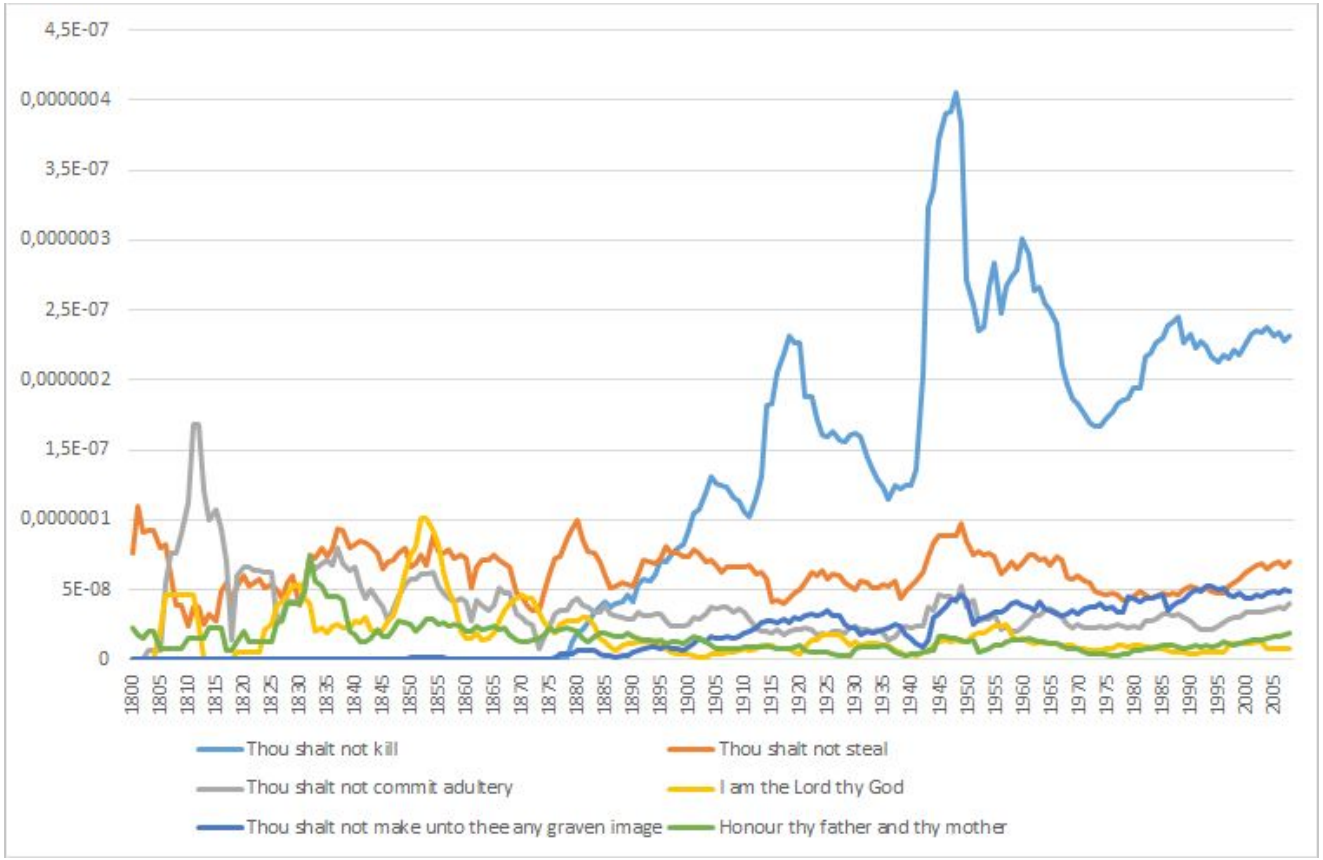
Ключевые идеи

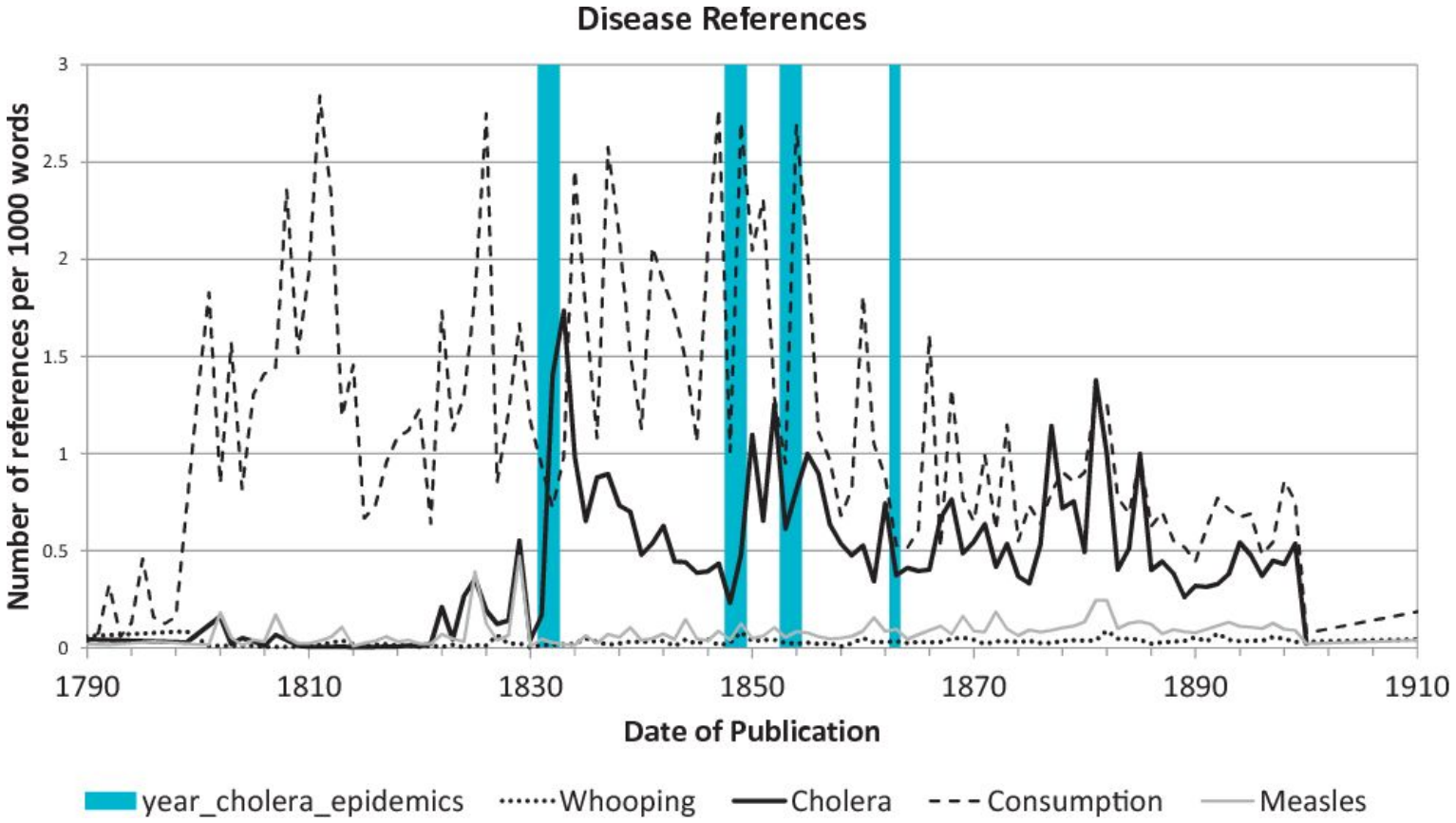
Так же как геномика изучает совокупность генов в живых организмов, культуромика должна изучать некую совокупность уникальных единиц, из которых складывается культура. Такими единицами являются словоупотребления в гигантском массиве данных, состоящем из оцифрованных книг коллекции Google Books.

Данные о том, как в течение времени меняется частотность определенных слов и словосочетаний, дают нам новые знания о социально-культурных трендах, общественных изменениях и процессах, открывает возможности сравнить и измерить, казалось бы, недоступные для квантитативных методов социальные концепты.



Michel, J.-B., Liberman A., Aiden A. P., Veres A., Gray M. K., Pickett J. P., Hoiberg D., Clancy D., Norvig P., Orwan J., Nowak M., Pinker St.Quantitative Analysis of Culture Using Millions of Digitized Books // Science. 16 December 2010. 331 (6014): 176 – 82.


4,5E-07
0,0000004
3,5E-07
0,0000003
2,5E-07
0,0000002
1,5E-07
0,0000001
5E-08
0
1800
1805
1810
1815
1820
1825
1830
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
Thou shalt not kill
Thou shalt not steal
Thou shalt not commit adultery
I am the Lord thy God
Thou shalt not make unto thee any graven image
Honour thy father and thy mother


Disease References
Number of references per 1000 words
3
2.5
2
1.5
1
0.5
0
1790
1810
1830
1850
1870
1890
1910
Date of Publication
year_cholera_epidemics
Whooping
Cholera
Consumption
Measles

Русскоязычные биграммы XX века

Особенности представления данных

В оригинальном наборе данных Google Ngrams записи выглядят следующим образом:

```
( отно.  1924,1,1     1928,1,1     1936,2,2     1953,2,2     1956,2,21957,2,2 1959,1,1     1962,2,2     1963,1,1
( недоразвитие_NOUN  1924,2,1     1925,5,5     1926,1,1     1927,1,11929,3,3 1930,2,2     1931,1,1     1932,2,2
( растворим_VERB 1934,1,1     1939,1,1     1945,1,1     1950,2,21951,1,1 1953,2,2     1955,1,1     1957,1,1
( Kohayashi_NOUN 1932,3,1     1937,1,1     1949,2,2     1953,1,11955,1,1 1958,8,3     1959,2,1     1960,4,2
) приходился_VERB    1868,2,2     1894,1,1     1896,1,1     1899,4,41902,2,2 1903,1,1     1904,3,3     1905,2,2
)_. указываться_VERB 1931,1,1     1970,1,1     1971,1,1     1972,1,11976,2,2 1988,1,1     1992,1,1     1997,1,1
```

Алгоритм обработки

1. Получаем файл с биграммами с сайта Google;
2. Распаковываем;
3. Берем только записи в диапазоне 1918—2010;
4. Данные о частотности биграммы за каждый год были нормируем на число словоупотреблений;
5. Числа меньше 0.0000000001 отбрасываем;
6. Лемматизируем (mystem);
7. Сливаем дубли;
8. Убираем некириллические биграммы.

Особенности представления данных

В оригинальном наборе данных Google Ngrams записи выглядят следующим образом:

```
( отно.  1924,1,1    1928,1,1    1936,2,2    1953,2,2    1956,2,21957,2,2 1959,1,1    1962,2,2    1963,1,1
( недоразвитие_NOUN  1924,2,1    1925,5,5    1926,1,1    1927,1,11929,3,3 1930,2,2    1931,1,1    1932,2,2
( растворим_VERB 1934,1,1    1939,1,1    1945,1,1    1950,2,21951,1,1 1953,2,2    1955,1,1    1957,1,1
( Kohayashi_NOUN 1932,3,1    1937,1,1    1949,2,2    1953,1,11955,1,1 1958,8,3    1959,2,1    1960,4,2
) приходился_VERB    1868,2,2    1894,1,1    1896,1,1    1899,4,41902,2,2 1903,1,1    1904,3,3    1905,2,2
)_. указываться_VERB 1931,1,1    1970,1,1    1971,1,1    1972,1,11976,2,2 1988,1,1    1992,1,1    1997,1,1
```

nevmenandr/russian-20th-century-bigrams

Объемы данных

Исходно: 385 547 647 записей

После лемматизации: 270 061 709

Только кириллица: 87 537 955

DOI: 10.57967/hf/5987

6 файлов по 15 млн строк каждый

```
@misc{boris_orekhov_2025,
	author = { Boris Orekhov },
	title = {
russian-20th-century-bigrams
(Revision 8ef57d3) },
	year = 2025,
	url = {
https://huggingface.co/datasets/nevm
enandr/russian-20th-century-bigrams
},
	doi = { 10.57967/hf/5987 },
	publisher = { Hugging Face }
}
```

как

анализировать?

Находим в частотностях выбросы

```
def detect_outliers(data):
    """Функция для определения выбросов на основе IQR."""
    q1, q3 = np.percentile(data, [25, 75])
    iqr = q3 - q1
    upper_bound = q3 + (1.5 * iqr)
    return [(head[i], x) for i, x in enumerate(data) if x > upper_bound]
```

Выбросы

ь_ADV в 1918,9.04e-08 1922,1.146e-07 1930,9.09e-08 1932,1.034e-07 1943,9.35e-08
ь_VERB о_ADP 1918,1.809e-07 1929,1.832e-07 1930,1.958e-07 1933,2.009e-07
ь_NOUN не_PRT 1918,4.071e-07 1941,2.053e-07

Обращаемся к Википедии

Выбросы содержат 1176779 записей;
Берем из биграмм слова длиннее 4 букв, убираем частеречную отметку, получается 21574 леммы;
Ищем все эти слова в Википедии;
Для 4337 поиск неуспешен;
Для 1739 найдены совпадения годов-выбросов и годов в статьях

Код поиска в Википедии

```
w = wikipedia.page("Москва")
w.content[:500]
```

'Москва́ (МФА: [mɐˈskvä]) — столица России, город федерального значения, административный центр Центрального федерального округа и центр Московской области, в состав которой не входит. Мегаполис; крупнейший по численности населения город России и её субъект — 13 274 285 человек (2025), что делает Москву 22-й среди городов мира по численности населения. Центр Московской городской агломерации.

Визуальный анализ

"начинаться": {
"page": "Понедельник начинается в субботу",
"years": [
"**1965**"
]
} # «Понеде́льник начина́ется в суббо́ту» (**1965**) — фантастическая юмористическая повесть братьев Стругацких
цветение начинаться 1936,3.577e-07 1965,5.27e-07

Природа совпадений

"капитал": {
"page": "Капитал",
"years": [
"1920"
]
} Работы итальянского экономиста Пьеро Сраффа в середине 1920-х годов заложили теоретические основы неорикардианства

централизация капитал 1920,1.7779e-06

Ищем биграммы

Совпало

174948 биграмм из 1176779

148143 не получилось найти

Царская фамилия

"царский фамилия_NOUN": {
"page": "Романовы",
"years": [
"**1926**"
]
} Последние представители мужского пола, Дмитрий Романович (**1926**—2016) и Николай Романович (1922—2014), умерли

Артефакты поиска


```
"o_ADP проводить_VERB": {
 "page": "Провод",
 "years": [
  "2009"
 ]
}
```

```
"обследовать и": {
 "page": "Чуваши",
 "years": [
  "1925",
  "1936"
 ]
}
```

Артефакты поиска


```
 "у под": {
   "page": "Катастрофа Ил-62
под Дмитровом",
   "years": [
    "1928"
   ]
 }
```

```
  "так надоедать_VERB": {
   "page": "Толстой, Лев
Николаевич",
   "years": [
    "1928"
   ]
```

Библиография


```
"с договор_NOUN": {
  "page": "Договор",
  "years": [
    "2001",
    "2006"
  ]
}
```


- Брагинский М. И., Витрянский В. В. Договорное право. М., 2001. Кн. 1: Общие положения.
- Гражданское право: В 4 т. / Под ред. Е. А. Суханова. 3-е изд. М., 2006.

Библиография

```
"с польский": {
  "page": "Польский язык",
  "years": [
    "1920",
    "1954"
  ]
}
```

Лер-Сплавинский Т. Польский язык. — М.: Издательство иностранной литературы, **1954**.

Другие промахи


```
 "свой необыкновенный_ADJ": {
  "page": "Необыкновенный
концерт",
  "years": [
   "1944",
   "1945"
  ]
```


Работа над новым спектаклем началась уже весной 1944 года. Почти год продумывалась пьеса, осенью 1945 года были готовы первые эскизы кукол, затем начались репетиции.

Обнадеживающие примеры


```
"трест_NOUN и_CONJ": {
 "page": "Трест",
 "years": [
  "1929"
 ]
}
```


К середине 1-й пятилетки (1929—1934) тресты превратились в промежуточное звено административного управления.

Обнадеживающие примеры

```
"в_ADP гибралтар": {
 "page": "Гибралтар",
 "years": [
  "1940"
 ]
}
```

В 1940 году в начале Второй мировой войны каудильо Франсиско Франко, соблюдая вооружённый нейтралитет, отверг предложение Гитлера захватить британский Гибралтар.

Обнадеживающие примеры


```
 "валюта для": {
  "page": "Валюта",
  "years": [
   "1944",
   "1993"
  ]
}
```


На смену ей в 1944 году пришла Бреттон-Вудская валютная система, представляющая собой систему золотовалютного стандарта.

Обнадеживающие примеры


```
"олений_ADJ рог_NOUN": {
 "page": "Северный олень",
 "years": [
  "1940",
  "1953"
 ]
```


Охота была разрешена снова в 1943 году. Поголовье неуклонно растёт, начиная с 1940-х годов.

Стадо на острове Св. Георгия вымерло к 1953 году.

Обнадеживающие примеры


```
"он военный": {
  "page": "Военный округ",
  "years": [
   "1942",
   "1943",
   "1944",
   "1945"
```

Обнадеживающие примеры

```
"от паровоз": {
 "page": "Паровоз",
 "years": [
  "1933"
 ]
}
```

В 1933 году на советских магистральных железных дорогах появился новый вид тяги — электровозная.

Обнадеживающие примеры


```
 "наш_ADJ
автомобиль_NOUN": {
  "page": "Автомобиль",
  "years": [
   "1923"
  ]
 }
```


В 1923 году фирма Бенца изготовила первый грузовой автомобиль с двигателем Дизеля

Обнадеживающие примеры


```
"на_ADP имущество": {
 "page": "Имущество",
 "years": [
  "2008"
 ]
}
```


Сама идея передвижных домов не нова, но она вновь стала актуальной в связи с ипотечным кризисом 2008 года в США: за год их производство выросло на 19 %

Обнадеживающие примеры


```
"c_ADP тихий_ADJ": {
 "page": "Тихий океан",
 "years": [
  "1928",
  "1945",
  "1946",
  "2000"
```


6 и 9 августа 1945 года Вооружёнными силами США были осуществлены атомные бомбардировки японских городов Хиросима и Нагасаки

С 1946 по 1958 года на атоллах Бикини…