

Чем может и чем не может наука о данных помочь науке о литературе

Борис Орехов

НИУ Высшая школа экономики

nevmenandr@gmail.com

2 марта 2019

- 1 Литература и анализ данных
- 2 Как это работает?

1 Литература и анализ данных

2 Как это работает?

Они занимаются поиском неочевидных закономерностей.

Зачем вообще наука о данных науке о литературе?

- Человечество живет в ситуации перепроизводства художественных текстов.
- Формируется Великое Непрочтенное.
- Все это может прочитать компьютер.
- Дальнее чтение.



- Литературоведы работают со смыслами, уровнями высокого порядка.
- Компьютер эти уровни охватить пока не может и вынужден работать с атомарными фактами:
 - отдельными словами,
 - предложениями,
 - именованными сущностями

Б. И. Ярхо: «Ни один математический акт не должен совершаться, пока в него не будет вложен конкретный литературоведческий смысл».

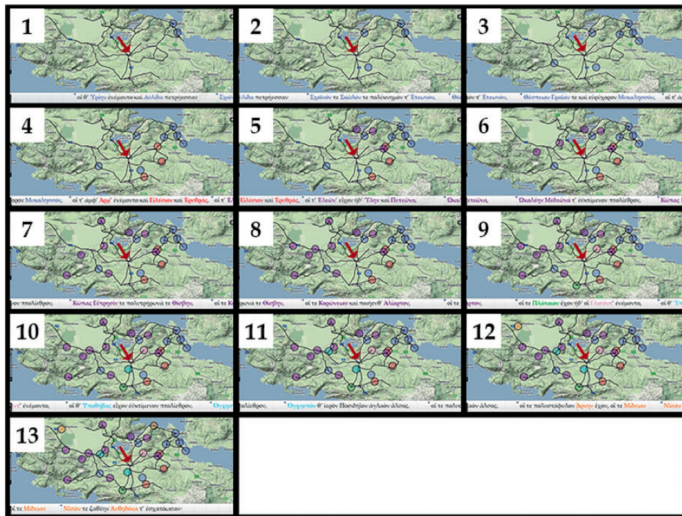
Лучше всего получается там, где задача и материал предельно просто формализуются

- Стиховедение.
- География (особенно фольклор).
- Язык художественной литературы.
- Текстология (сравнение вариантов, определение. заимствований, никакого выхода к содержанию).
- Определение авторства.
- Гендерные исследования.

1 Литература и анализ данных

2 Как это работает?

Каталог кораблей



Географическое распределение мифологических сюжетов в фольклоре

Uther H.-J. The Types of International Folktales
tale type in the current map:
303

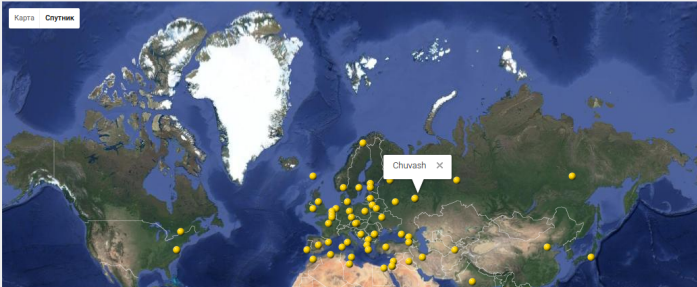
Select an index
To display a tradition name click a point marker. To save results you can use "Print screen" operation.
[Display all points with motif/tale types information.](#)

Enter tale type code, f.e. '303' or 'a2a'. You can use "|" as OR operator and "*" as AND:
Display a map for:
Display points without searched items ☐

Distribution of 2 units. Enter 2 tale type codes with "|" between them, f.e. "a1|a2a". "|" is a connection operator for two codes:
Distribution of 2 units:
Display points without searched items ☐

Find a tale type number. Enter a digit(s):

Карта Спутник

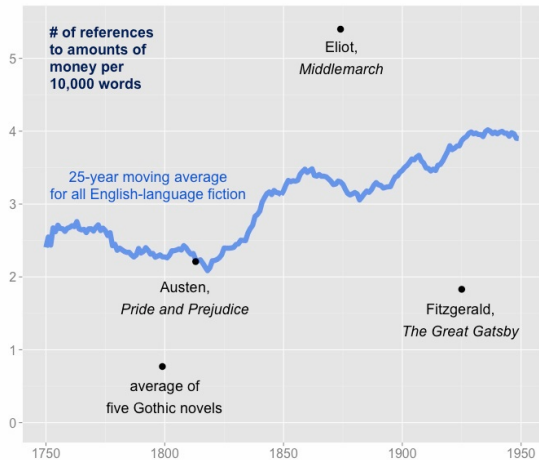


an unity is present: ●
an unity is absent: ○

The tale type(s) is/are present in (74 traditions):

- Japanese W
- Yakut W
- Chinese W
- Indonesian W
- Indian W
- Vogul W
- Tadzhik W
- Iranian W
- Malagasy W
- Cheremis W
- Chuvash W
- Osssetian W
- Georgian W
- Iraqi W
- Advoea W

Упоминание денег в романах



Сила голоса в «Идиоте»

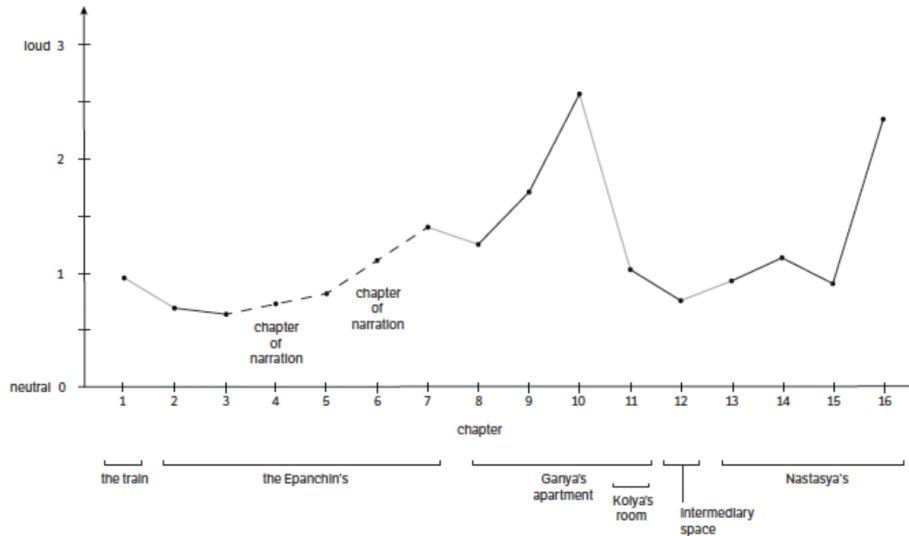
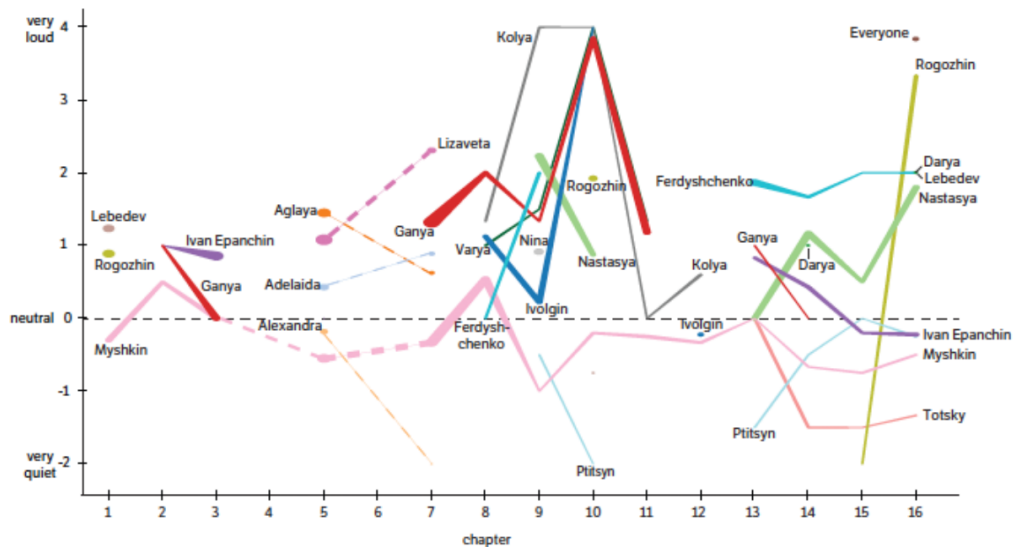
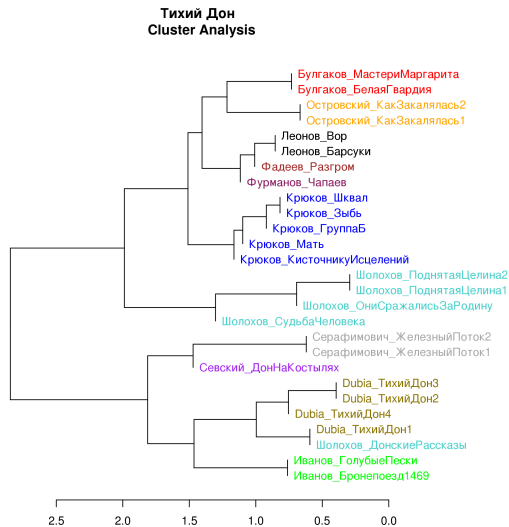


Figure 3: Loudness at the scale of the chapter: *The Idiot*, Book I.

Ещё сила голоса в «Идиоте»



Определение авторства (стилеметрия)



200 MEW, Cullid @ 0%

- Можно сравнивать тексты, сопоставляя небольшие фрагменты (как последовательности) и находя цитаты, см. Tesseract Project
- Можно сравнивать тексты, сопоставляя их как наборы слов (не последовательности!).

Tesserae

http://tesserae.caset.buffalo.edu/cgi-bin/transitional/get-data.pl?session=000000

Google

Tesserae

INTERTEXTUAL PHRASE MATCHING

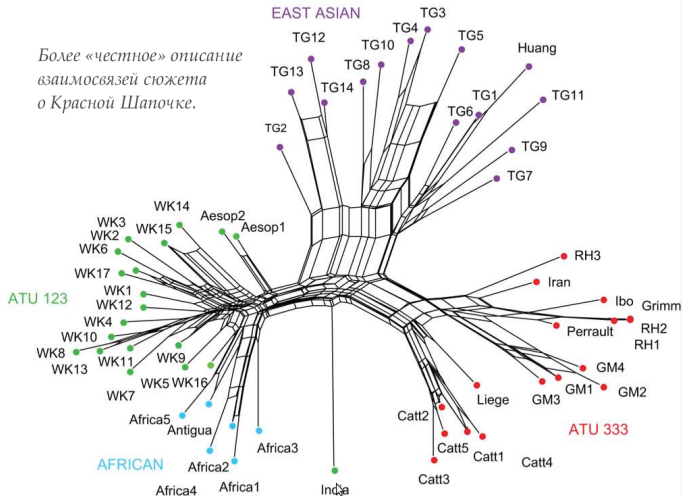
BASIC SEARCH | VERSION 2 | ABOUT TESSERAE | DEPT. OF CLASSICS | DEPT. OF LINGUISTICS

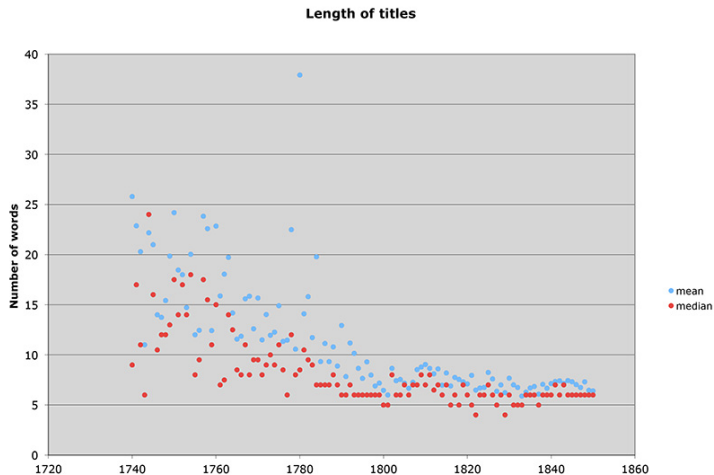
Sort by **target phrase** and format as **html** [Change Display](#) [view session details](#)

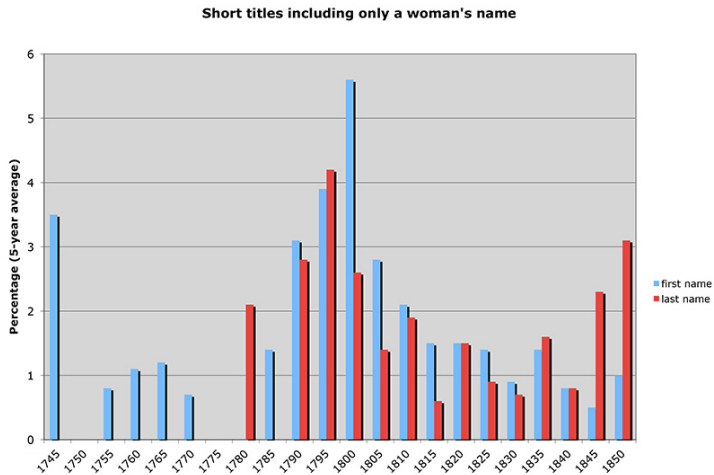
| | target phrase | source matches | matched on | score |
|----|---|--|---|-------|
| 1. | <p>bella per emathios plus quam civilia campos iusque datum sceleri canimus populumque potentem in sua vicitrici conversum viscera dextra cognatasque acies et rupto foedere regni certatum totis concussi viribus orbis in commune nefas infestisque obvia signis signa pares aquilas et pila minantia pilis</p> <p>luc. 1.1</p> | <p>multa quoque et bello passus dum conderet urbem inferretque deos latio genus unde latinum albanique patres atque altae moenia romae</p> <p>verg. aen. 1.5</p> | <p>multus, qui.2, bellum, bellus,</p> | 10 |
| 2. | <p>bella per emathios plus quam civilia campos iusque datum sceleri canimus populumque potentem in sua vicitrici conversum viscera dextra cognatasque acies et rupto foedere regni certatum totis concussi viribus orbis in commune nefas infestisque obvia signis signa pares aquilas et pila minantia pilis</p> <p>luc. 1.1</p> | <p>hinc populum late regem belloque superbum venturum excidio libyae</p> <p>verg. aen. 1.21</p> | <p>populus.1, populus.2, bellum, bellus,</p> | 10 |
| | <p>bella per emathios plus quam civilia campos iusque datum sceleri canimus populumque potentem in sua vicitrici conversum viscera</p> <p>luc. 1.1</p> | <p>id metuens veterisque memor saturnia belli edes, quod ed telum non sedit, non est, non</p> <p>verg.</p> | <p>bellum,</p> | |

Тексты можно сгруппировать по похожести: сказки

Более «честное» описание
взаимосвязей сюжета
о Красной Шапочке.

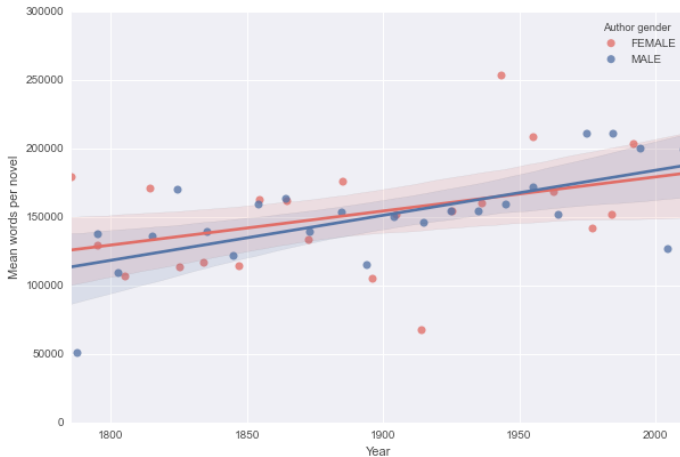






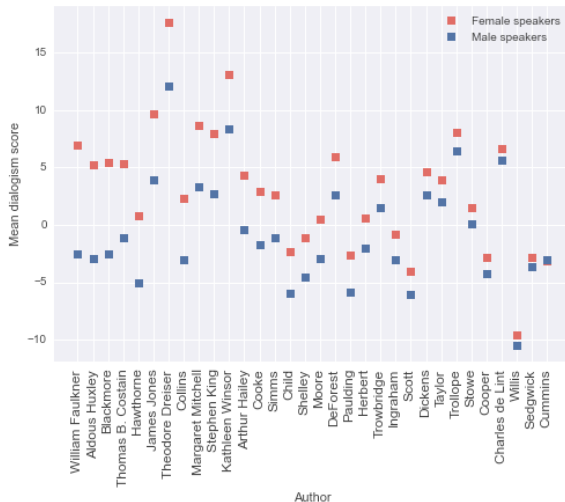
Ещё гендер

Правда ли, что авторы-женщины более многословны?



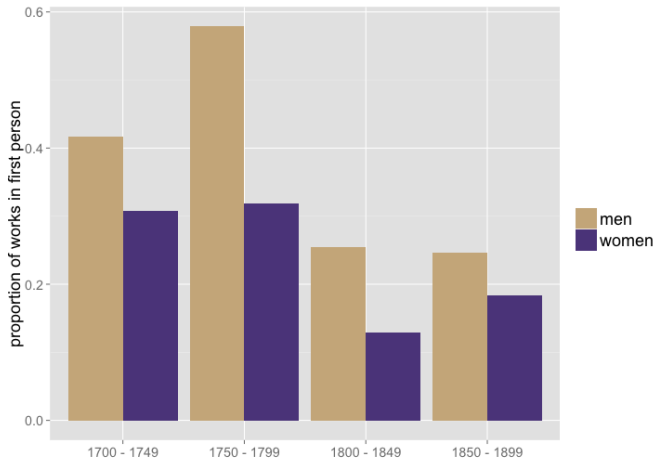
И опять гендер

Правда ли, что персонажи-женщины более многословны?



Ещё немного гендера

Правда ли, что авторы-женщины чаще используют речь от первого лица?



Example (С их помощью которых пытались предсказать смерть персонажей «Игры престолов»)

House to which a character belongs

Social group to which a character belongs

Male or female

Character's appearance in the book (все книги по отдельности)

Number of dead characters to whom a character is related

Whether the character is married

...

<https://got.show/machine-learning-algorithm-predicts-death-game-of-thrones>

Что не так с ЭТИМ СПИСКОМ?

Что же не так?

Эти признаки не отражают **поэтику** произведения.

Перечисленные признаки, разумеется, не случайны. Они взяты из практики применения анализа данных в жизни. Для **человека** с точки зрения статистики признаки принадлежности *семье, социальной группе, пол, состояние в браке* — осмысленны.

Для **персонажа** это не обязательно так.

Здесь мы видим отражение отношений **в вымышленном мире**, а не отражение поэтики. Из-за этого исследователь, который подошел к тексту с таких позиций, выглядит как вульгарный социолог, потому что не видит текста и не обращает внимание на то, как он построен.

Конечно, устройство вымышленного мира тоже отражает поэтику, но косвенно

- Лучше всего работать с фольклором, потому что в нем всё одинаковое.
- Пока компьютеры мало что могут сказать о важных для литературоведов характеристиках персонажей, сюжетах и мотивах.
- Но много могут сказать о том, что либо не относится к содержанию художественных произведений, либо мы и так знали.