

Цифровые гуманитарные технологии: ресурсы, инструменты, кейсы. Лекция 3. Компьютерная лингвистика и анализ данных

Борис Орехов

НИУ Высшая школа экономики

nevmenandr@gmail.com

12 марта 2018

Содержание

1 Компьютерная лингвистика в digital humanities

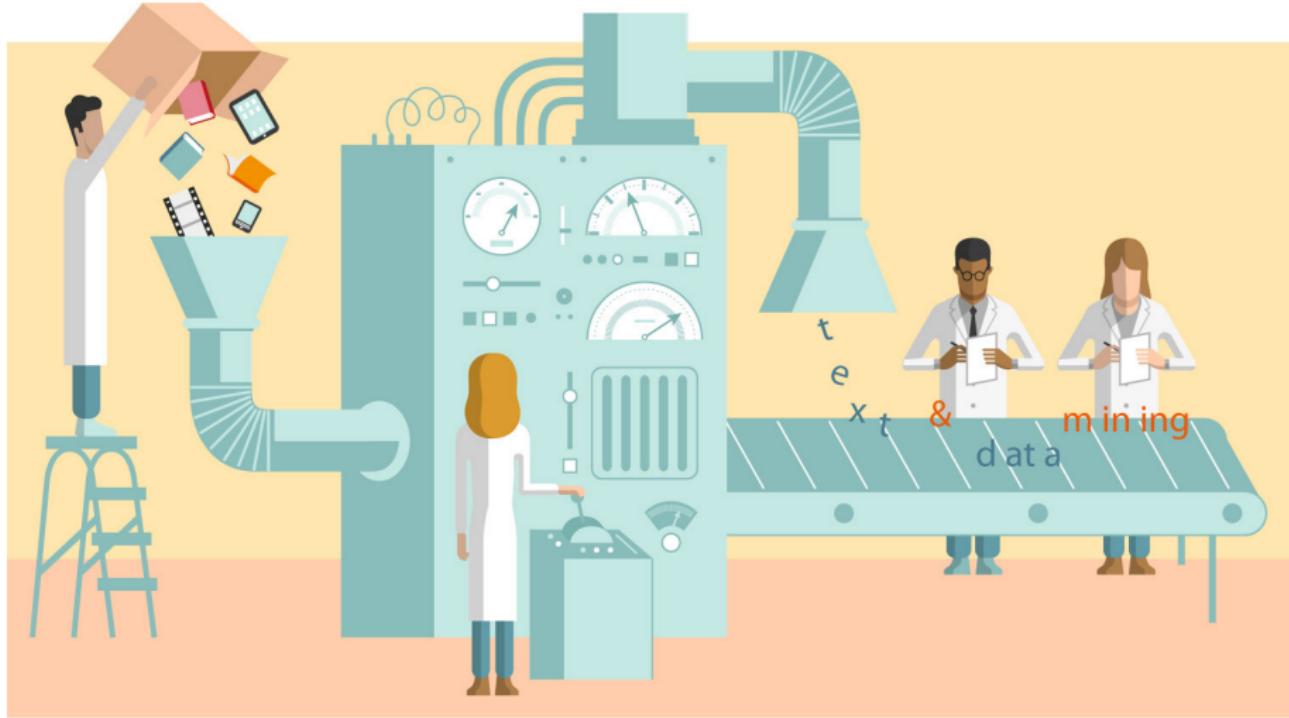
2 Анализ данных

Содержание

1 Компьютерная лингвистика в digital humanities

2 Анализ данных

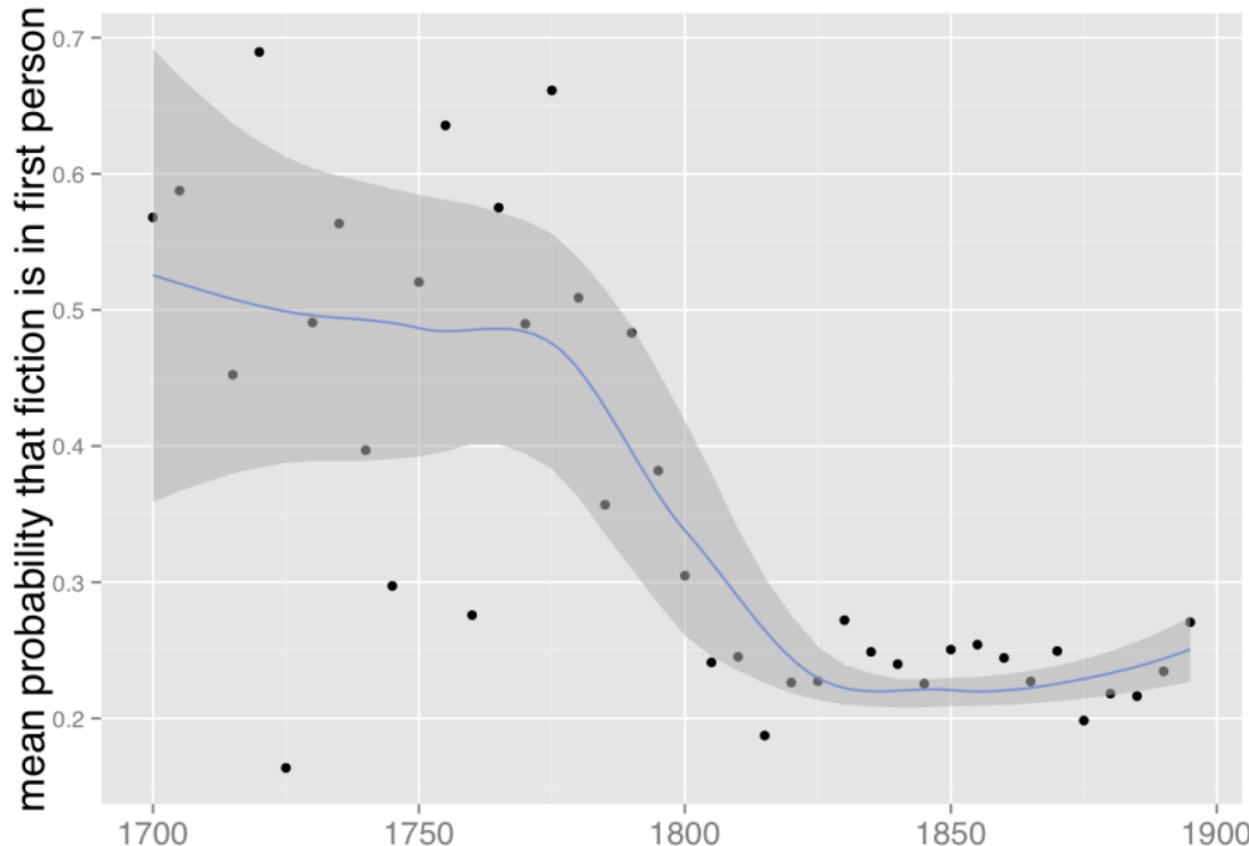
Когда текст оцифрован...



Что умеет компьютерная лингвистика?

- автоматическое извлечение информации
- выделение ключевых слов
- тематическое моделирование
- определение авторства
- анализ тональности
- выявление заимствований
 - токенизировать,
 - лемматизировать,
 - разметить синтаксически

Определение морфологической формы



Information Extraction

- Almost exclusively focused on explicit information

Person

Person

C0018795

"Bob Smith is a 61-year-old man referred by Dr. Davis for outpatient cardiac catheterization because of a positive exercise tolerance test. Recently, he started to have left shoulder twinges and tingling in his hands. A stress test done on C0015672-02 revealed that the patient exercised for 6 1/2 minutes, stopped due to fatigue. However, Mr. Smith is comfortably breathing in room air. He also showed accumulation of fluid in his extremities. He does not have any chest pain."

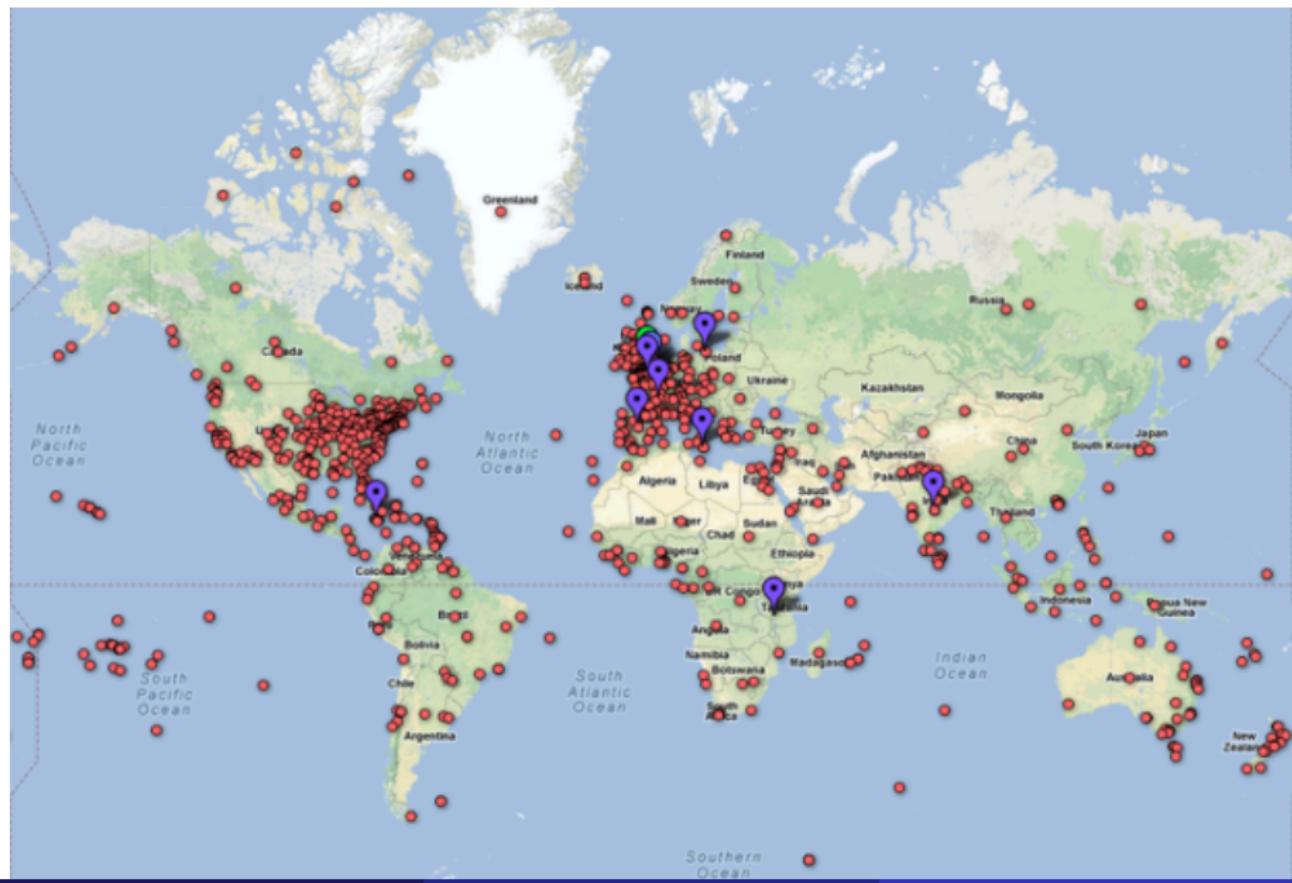
C0008031

Named Entity Recognition

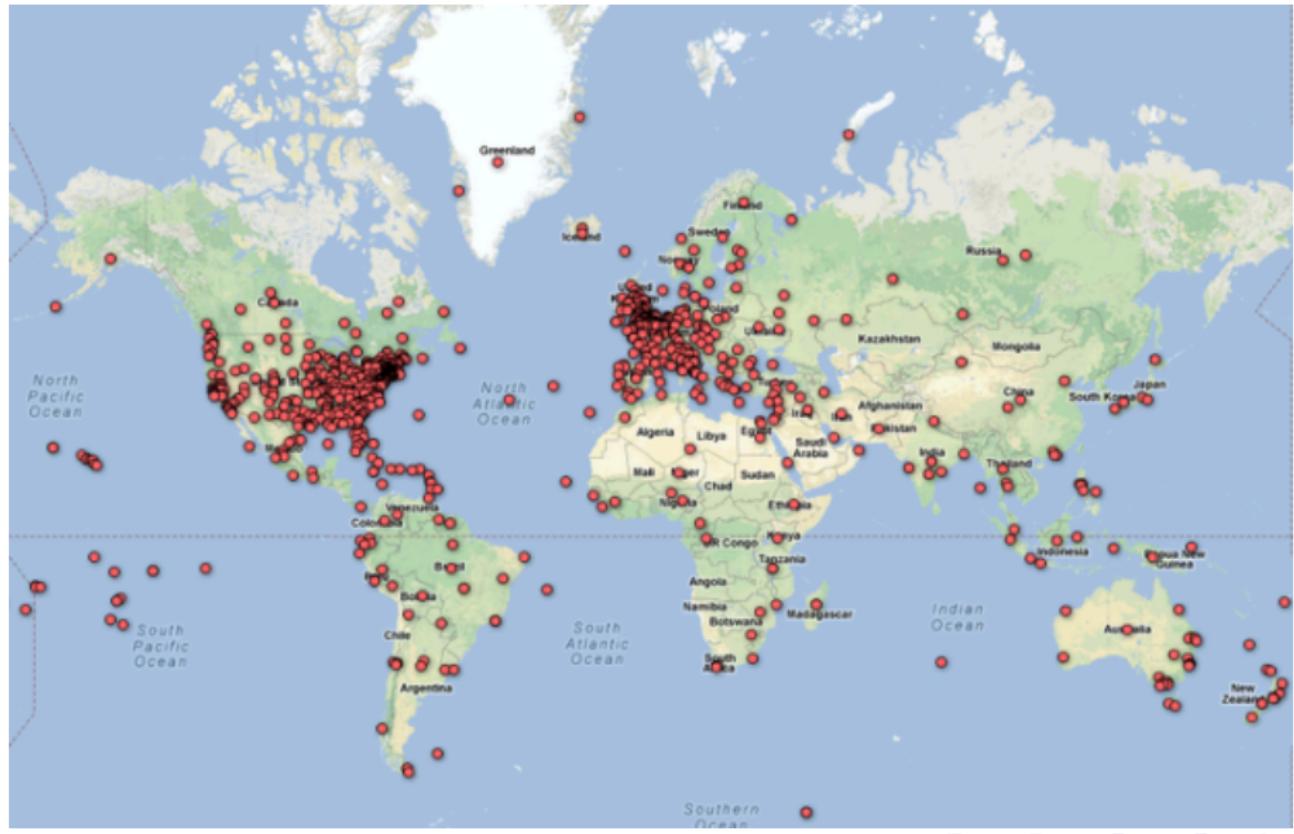
Relationship Extraction

Entity Linking

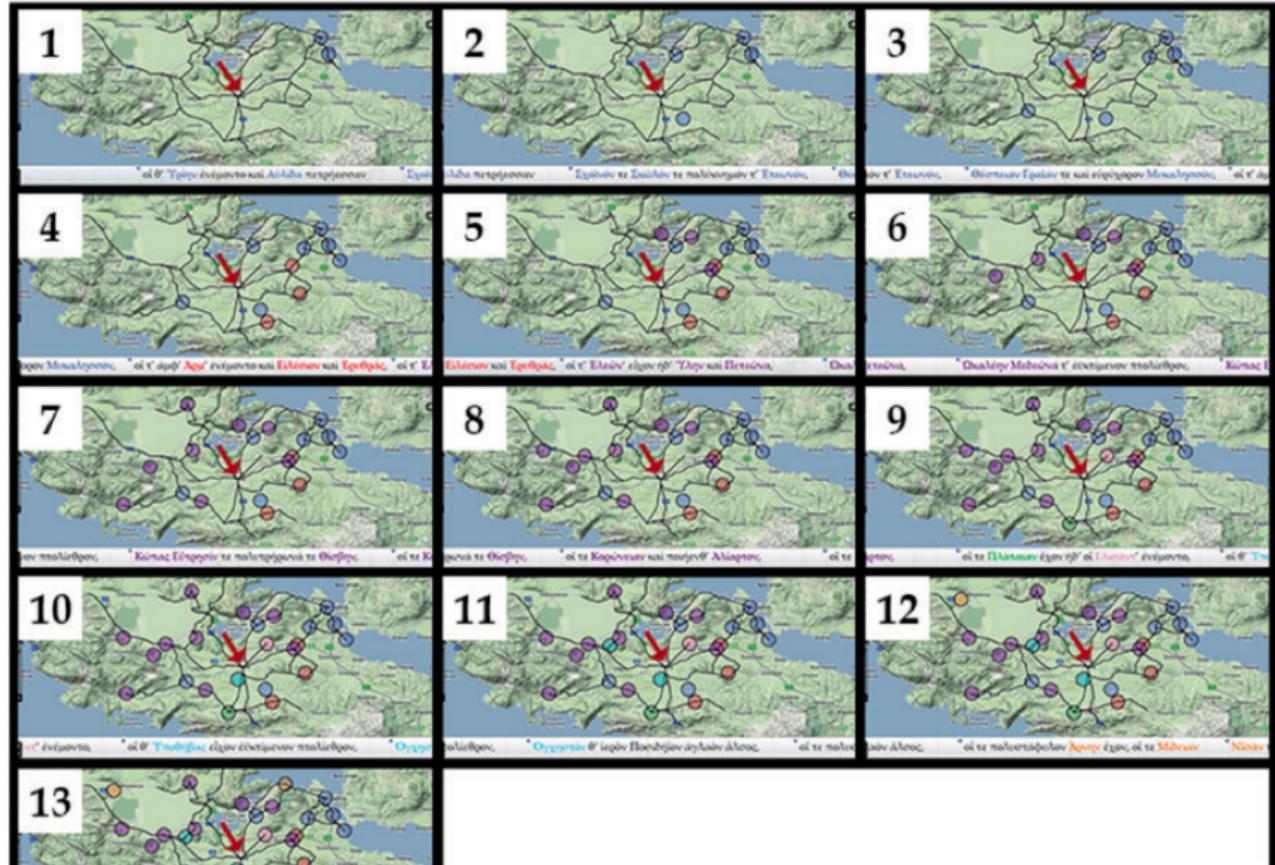
География английского романа XIX в.



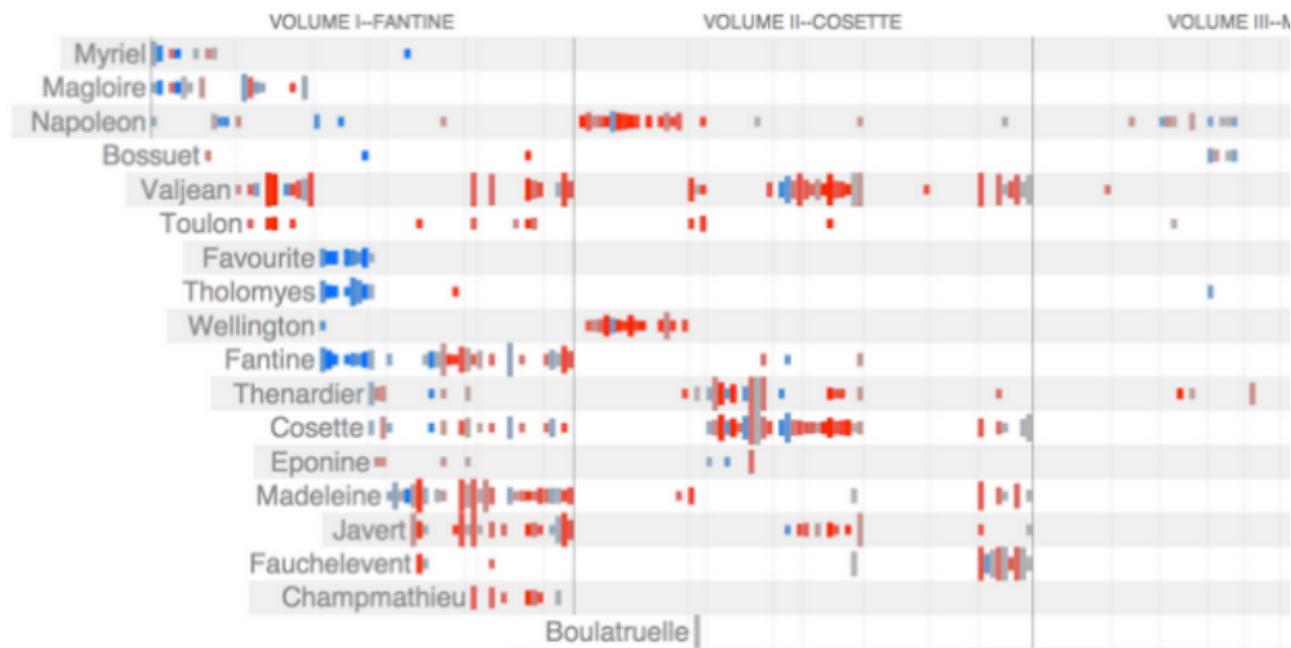
География американского романа XIX в.



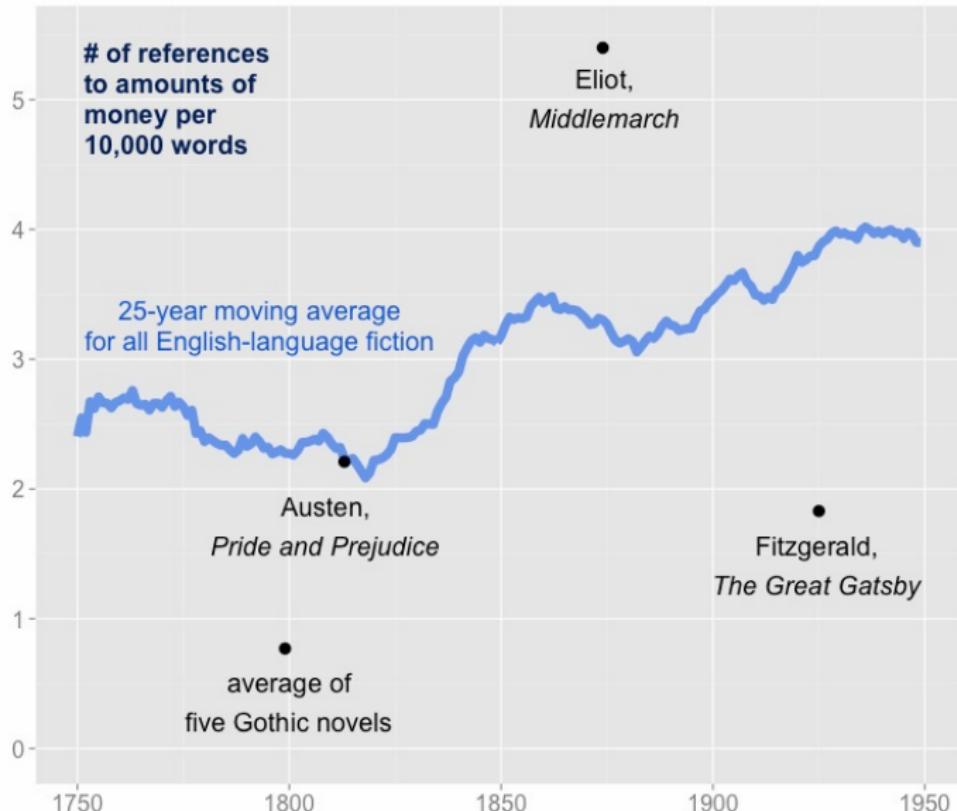
Каталог кораблей



Имена персонажей в «Отверженных»



Упоминание денег в романах



Сила голоса в «Идиоте»

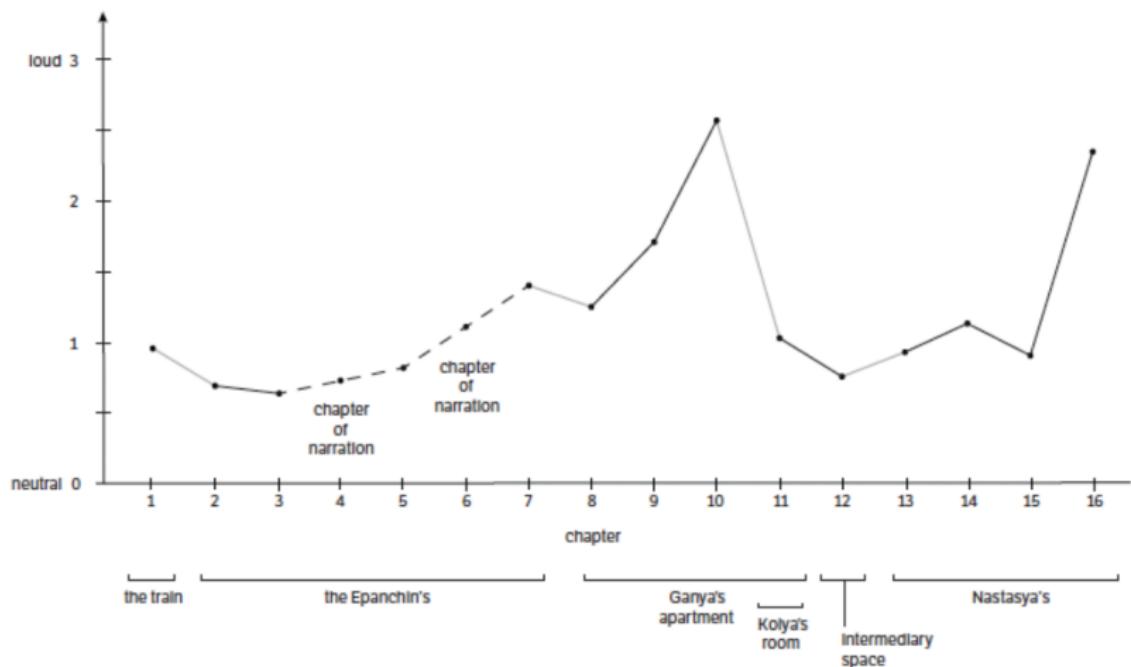


Figure 3: Loudness at the scale of the chapter: *The Idiot*, Book I.

Ещё сила голоса в «Идиоте»

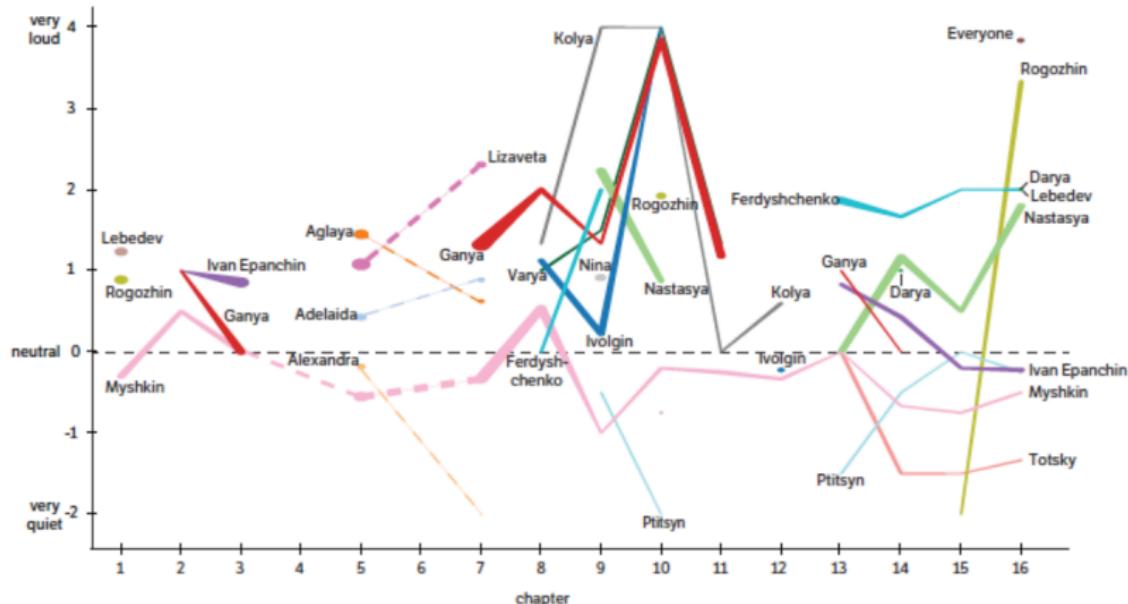
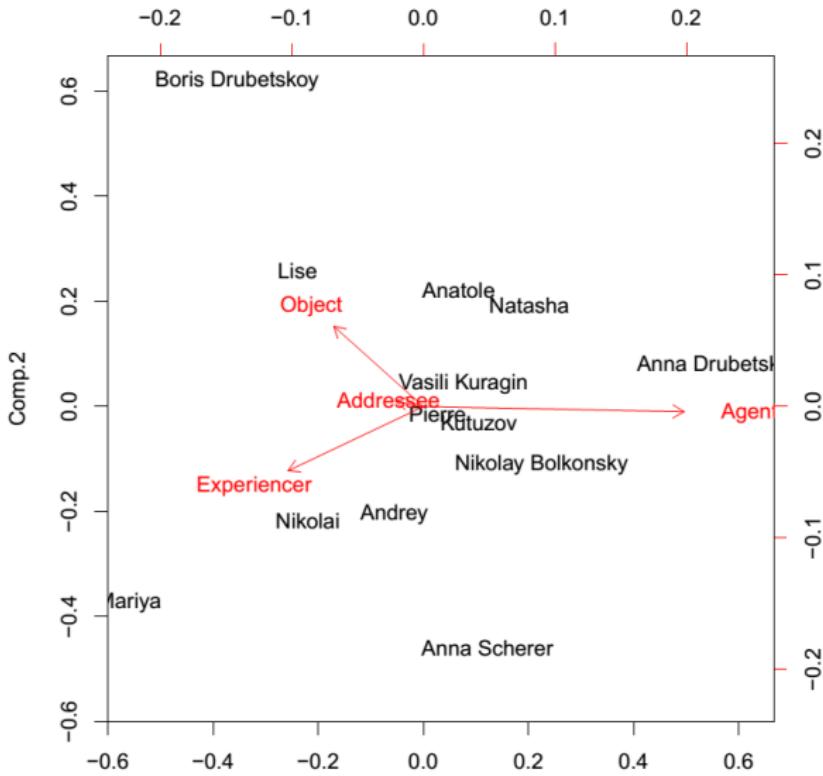
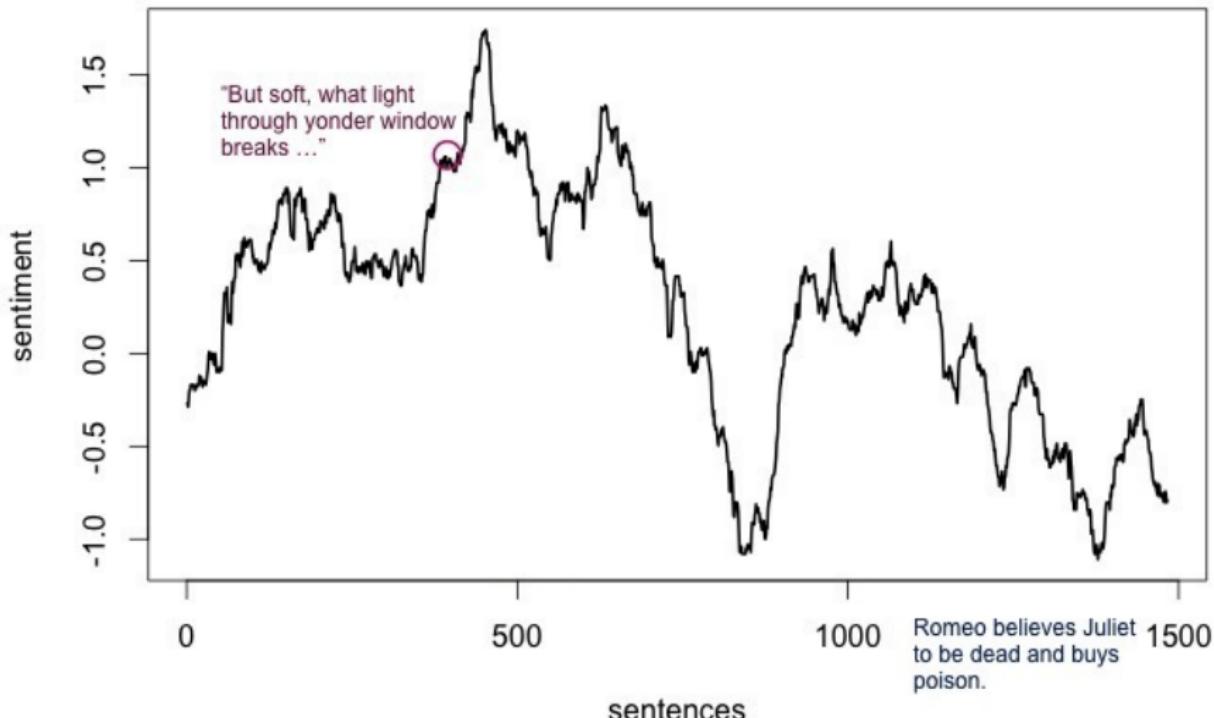


Figure 5: Loudness allocated between characters: *The Idiot*, Book I.

Семантические роли в «Войне и мире»



Анализ тональности в «Ромео и Джульетте»



Анализ тональности в книгах Толкиена



An analysis of
Tolkien's books

WORD COUNT AND DENSITY

CHARACTER MENTIONS

KEYWORD FREQUENCY

COMMON WORDS

SENTIMENT ANALYSIS

CHARACTER CO-OCCURRENCE

CHAPTER LENGTHS

WORD APPEARANCE

POSTERS

ABOUT

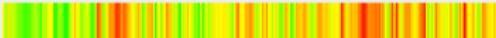
COMMENT

SENTIMENT ANALYSIS

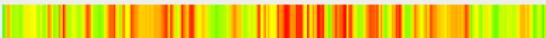
VIEW
Books
Graphs

These graphs show an analysis of the feeling for each page throughout Tolkien's works. The sentiment has been analysed for each sentence and then average over each page. Green, yellow and red indicate positive, neutral and negative sentiments respectively.

THE SILMARILLION



THE HOBBIT



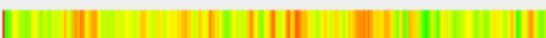
THE FELLOWSHIP OF THE RING



THE TWO TOWERS



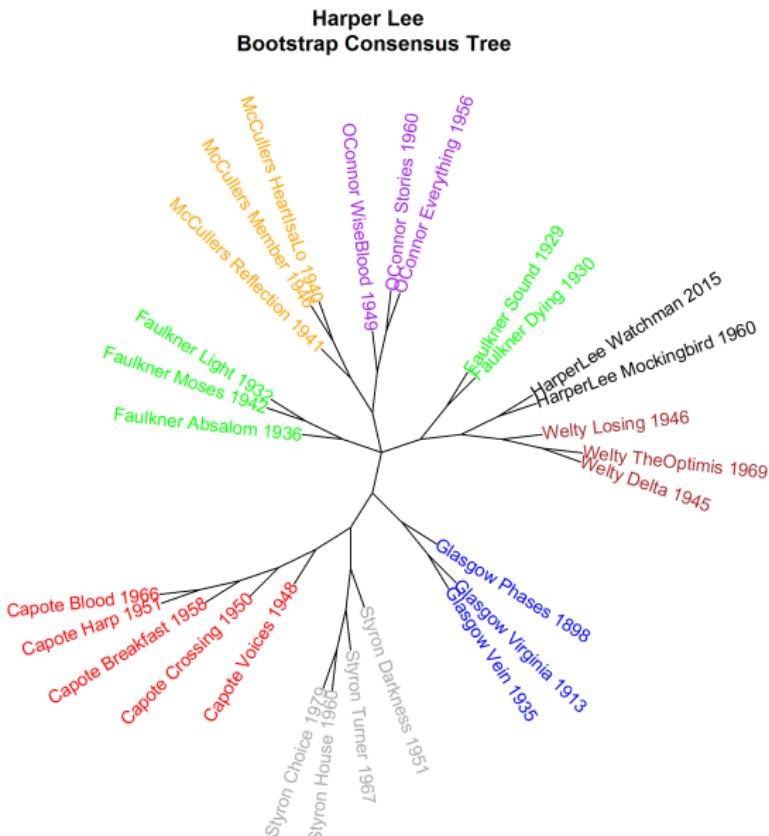
THE RETURN OF THE KING



About the graphs

Sentiment analysis is the science of assigning mood to pieces of text based on keywords and structure. On the web this kind of research is most commonly used in social media. In this project I decided to apply it to Tolkien's works to see if I could find patterns. Since sentiment analysis is rather difficult I used a free API called

Определение авторства (стилометрия)



Автоматический поиск цитат, формул и вариантов

- Можно сравнивать тексты, сопоставляя небольшие фрагменты (как последовательности) и находя цитаты, см. Tesserae Project
- Можно сравнивать тексты, сопоставляя их как наборы слов (не последовательности!). Т. н. «косинусное расстояние».

Поиск цитат

Tesserae

http://tesserae.caset.buffalo.edu/cgi-bin/transitional/get-data.pl?session=000001 Google

Tesserae

INTERTEXTUAL PHRASE MATCHING

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

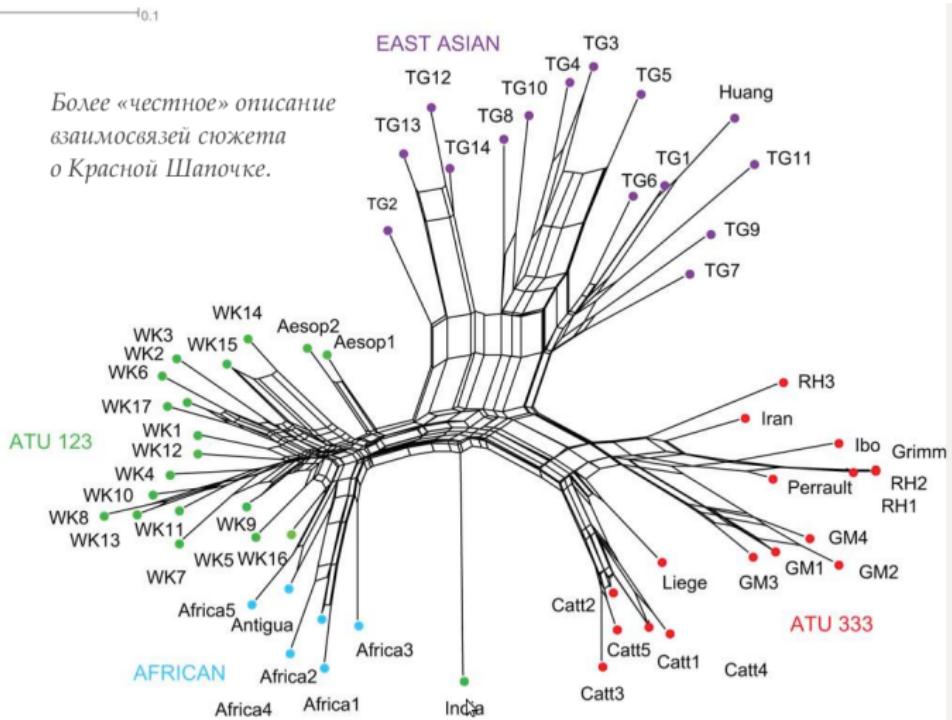
Sort by target phrase and format as html Change Display view session details

	target phrase	source matches	matched on	score
1.	<i>bella per emathios plus quam</i> civilia campos iusque datum sceleri canimus populumque potentem in sua vinctri 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	<i>verg.</i> <i>aen.</i> 1.5 <i>multa quoque et bello</i> passus dum conderet urbem inferretque deos latio genus unde latinum albanique patres atque altae moenia romae	multus, qui.2, belum, belius,	10
2.	<i>bella per emathios plus quam</i> civilia campos iusque datum sceleri canimus <i>populumque</i> potentem in sua vinctri 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	<i>verg.</i> <i>aen.</i> 1.21 <i>hinc populum late regem beloque</i> superbum venturum excidio libyae	<i>populus.1,</i> <i>populus.2,</i> belum, belius,	10
	<i>bella per emathios plus quam</i> civilia campos iusque datum sceleri canimus populumque potentem in sua vinctri conversum viscera	<i>verg.</i> <i>admetuens veterisque memor saturnia belli</i>	bellum,	

Косинусное расстояние

Тексты сравниваются как наборы слов, без учёта порядка их следования в текстах («мешок слов»). Каждый текст представляется в виде **вектора**, в котором количество употреблений каждого слова будет цифрой-координатой. Дальше мы считаем косинус угла между этими векторами. Чем расстояние меньше, тем тексты более похожи.

Теперь тексты можно сгруппировать по похожести

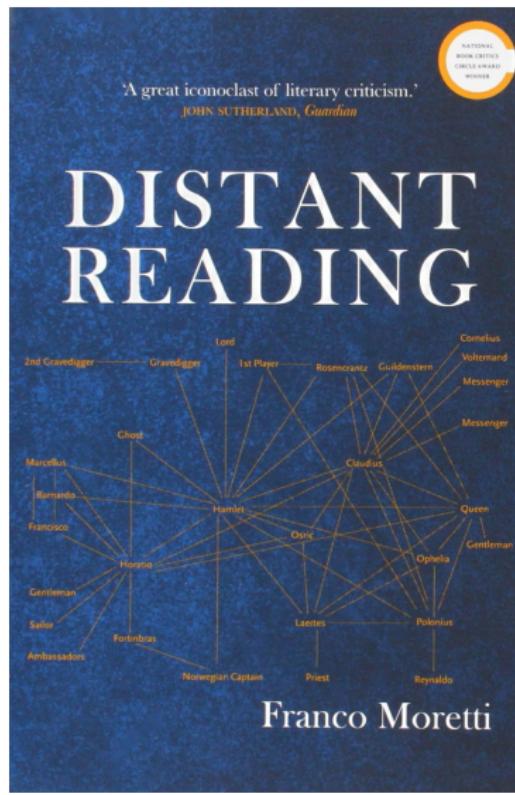


Содержание

1 Компьютерная лингвистика в digital humanities

2 Анализ данных

Франко Моретти «Дальнее чтение»

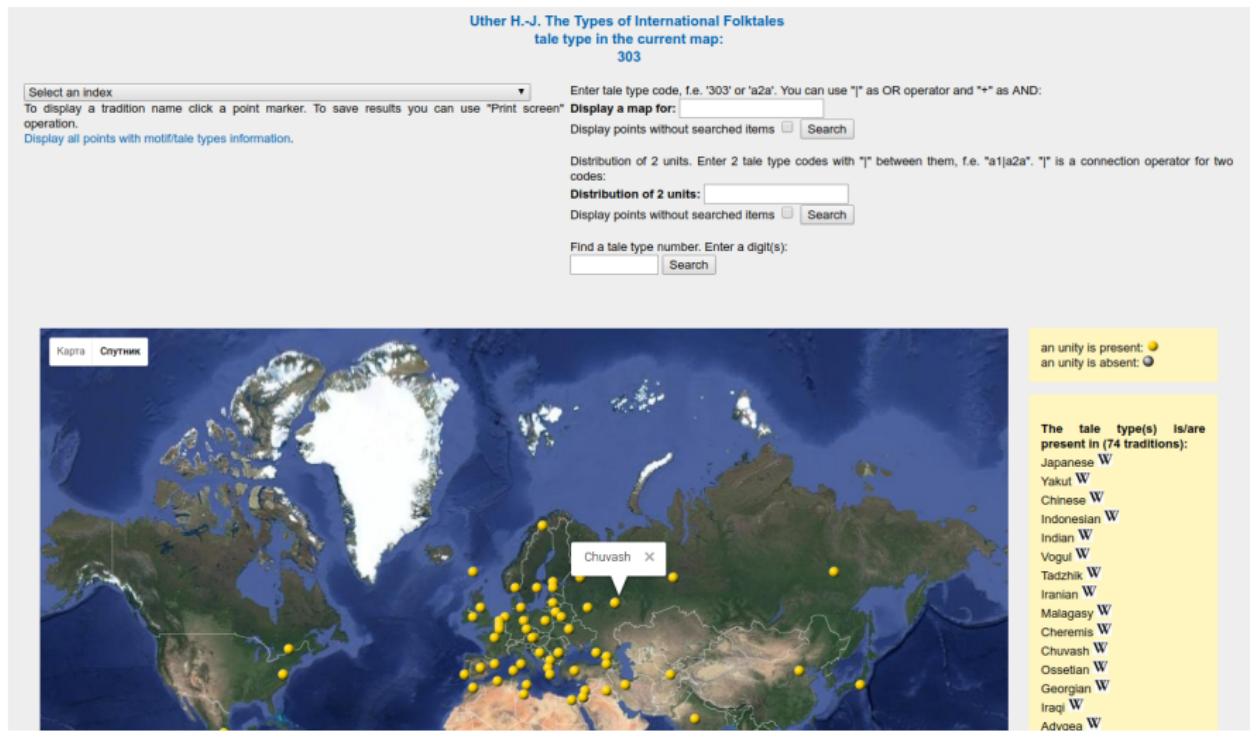


Анализ данных и литературоведение очень похожи

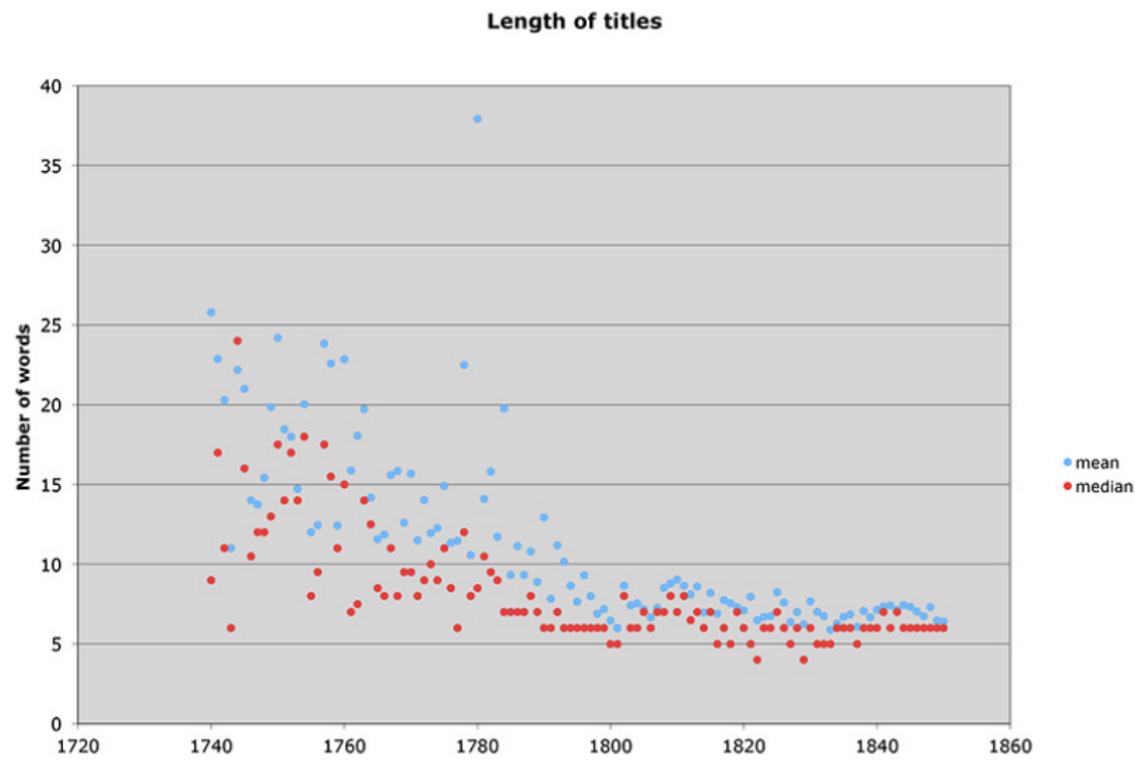
Они занимаются поиском неочевидных закономерностей.

Анализ данных в DH возможен благодаря компьютерной лингвистике.

Географическое распределение мифологических сюжетов в фольклоре

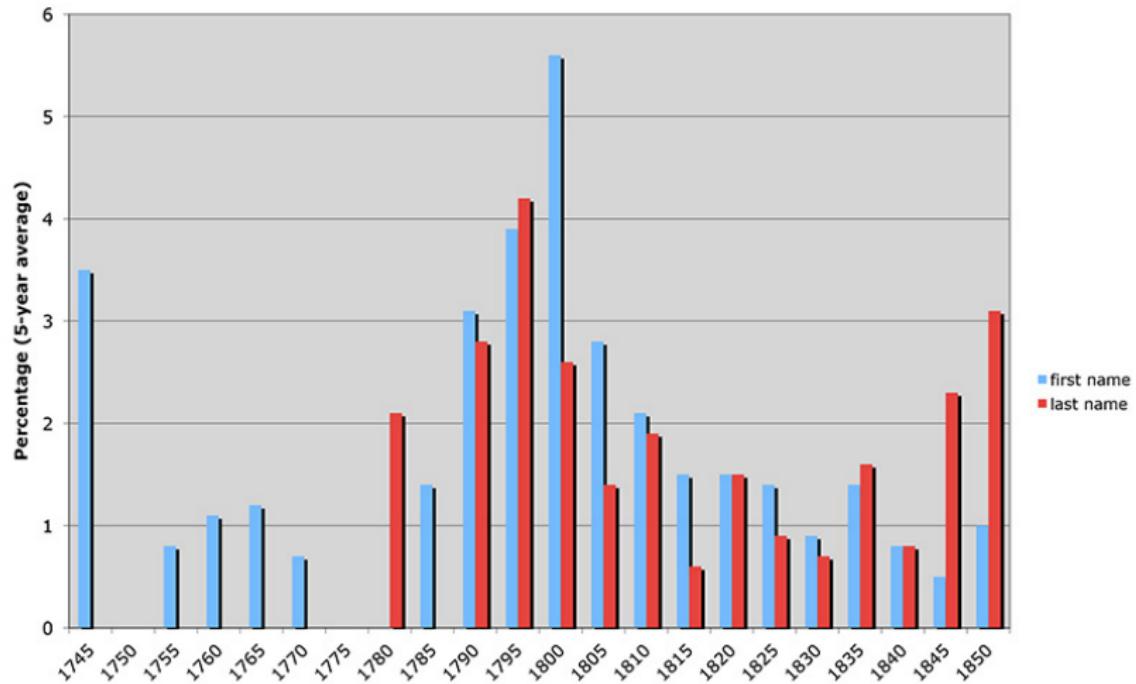


Длина заглавий



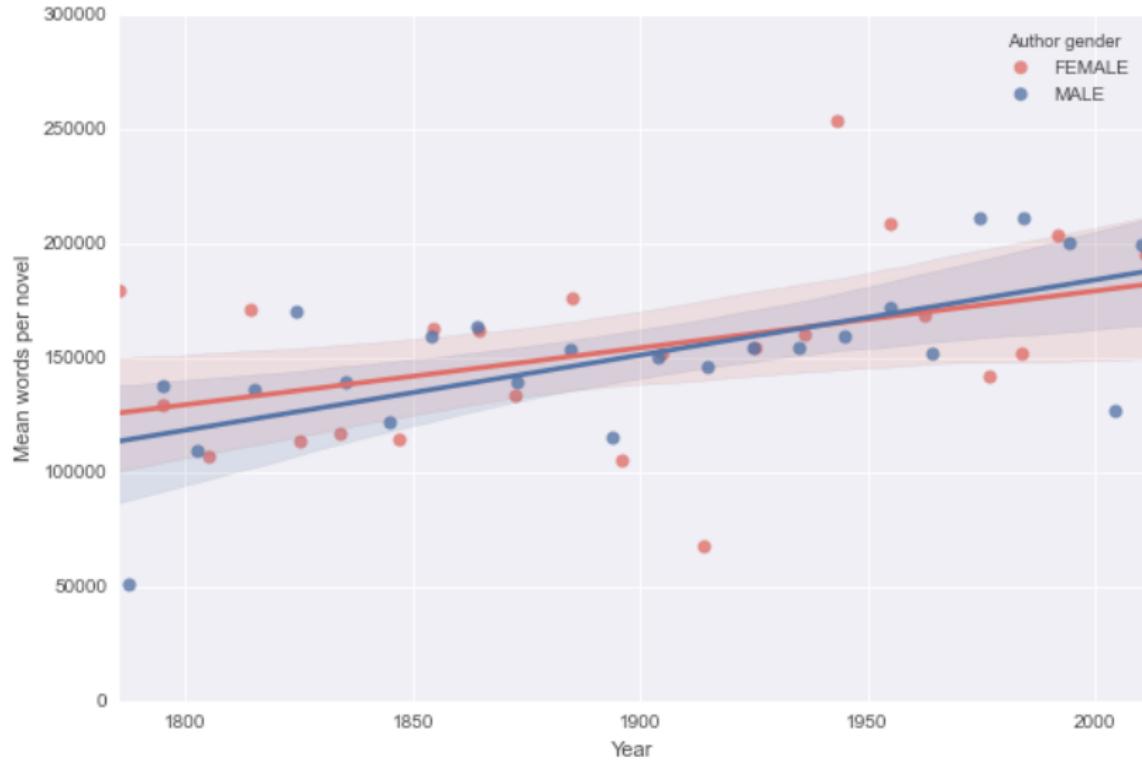
Женские имена в заглавиях

Short titles including only a woman's name



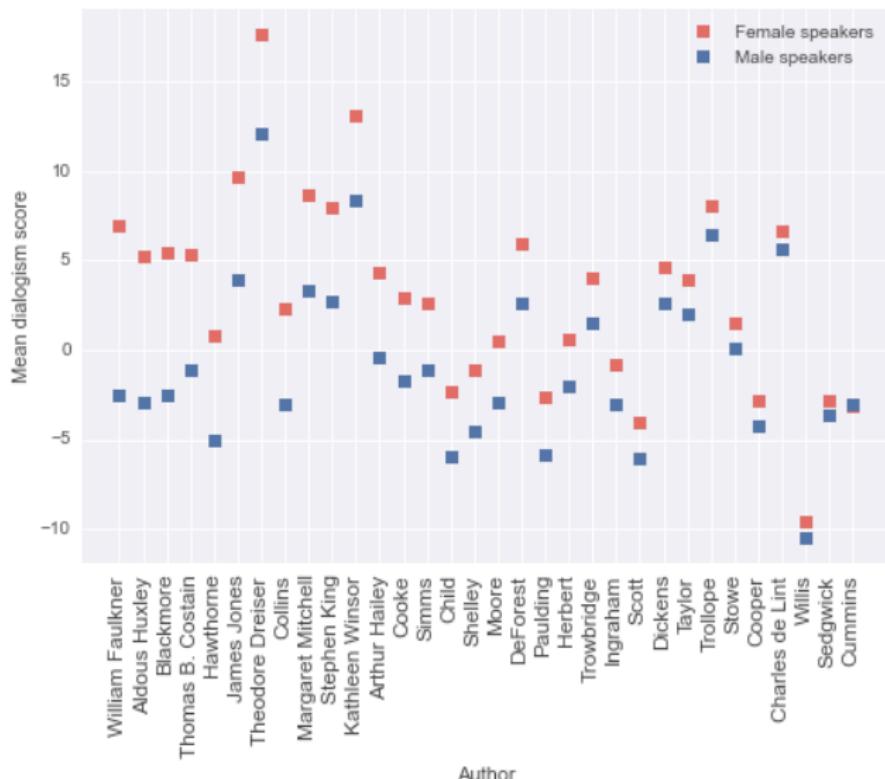
Ещё гендер

Правда ли, что авторы-женщины более многословны?



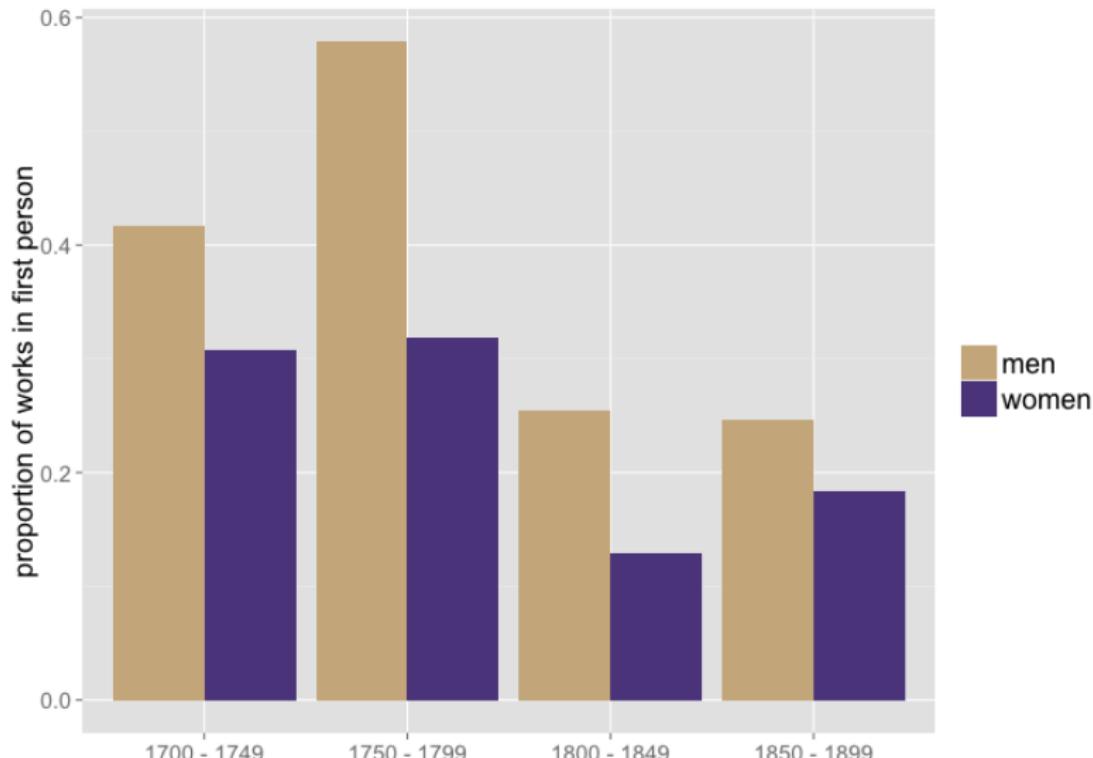
И опять гендер

Правда ли, что персонажи-женщины более многословны?

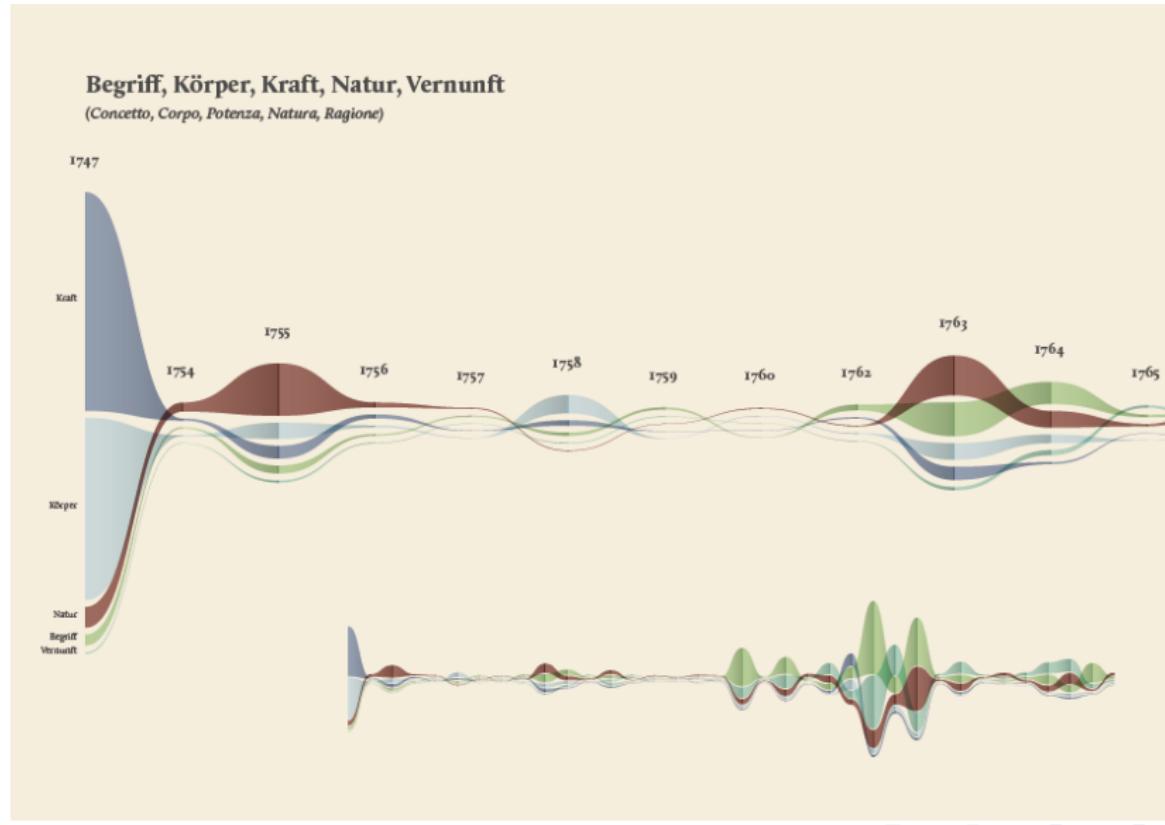


Ещё немного гендерра

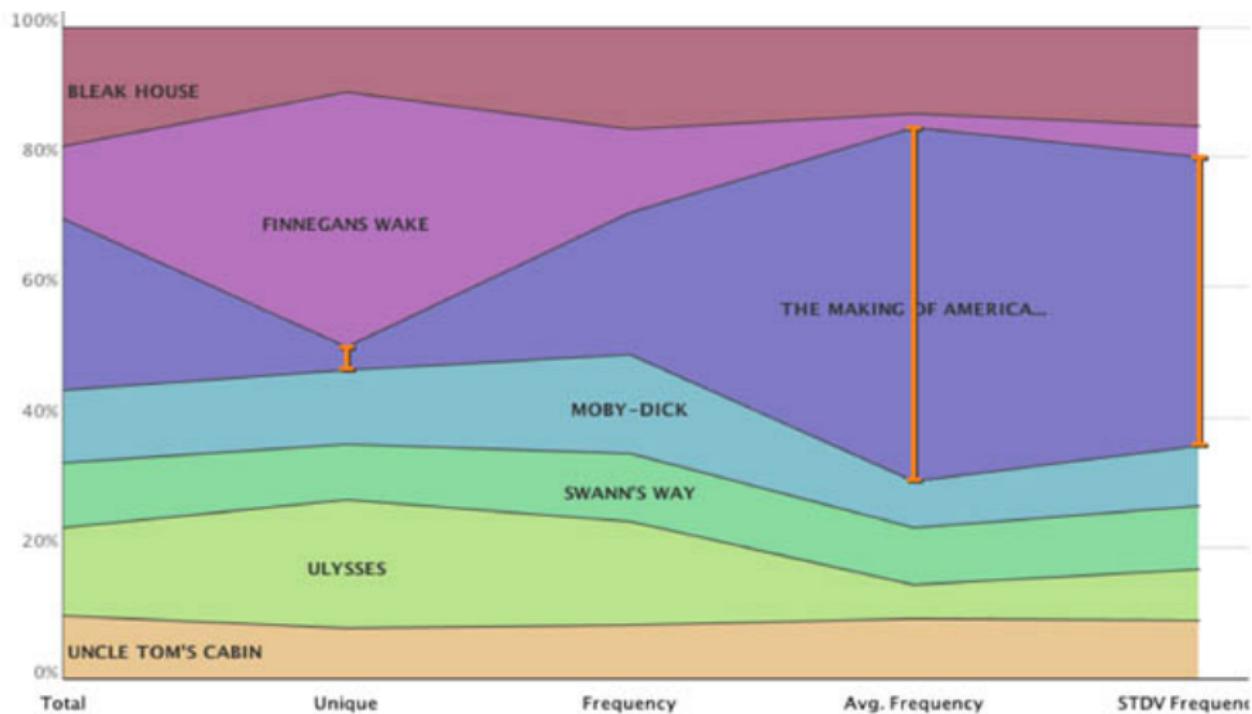
Правда ли, что авторы-женщины чаще используют Ichzählung?



Как распределяются понятия в трудах И. Канта?

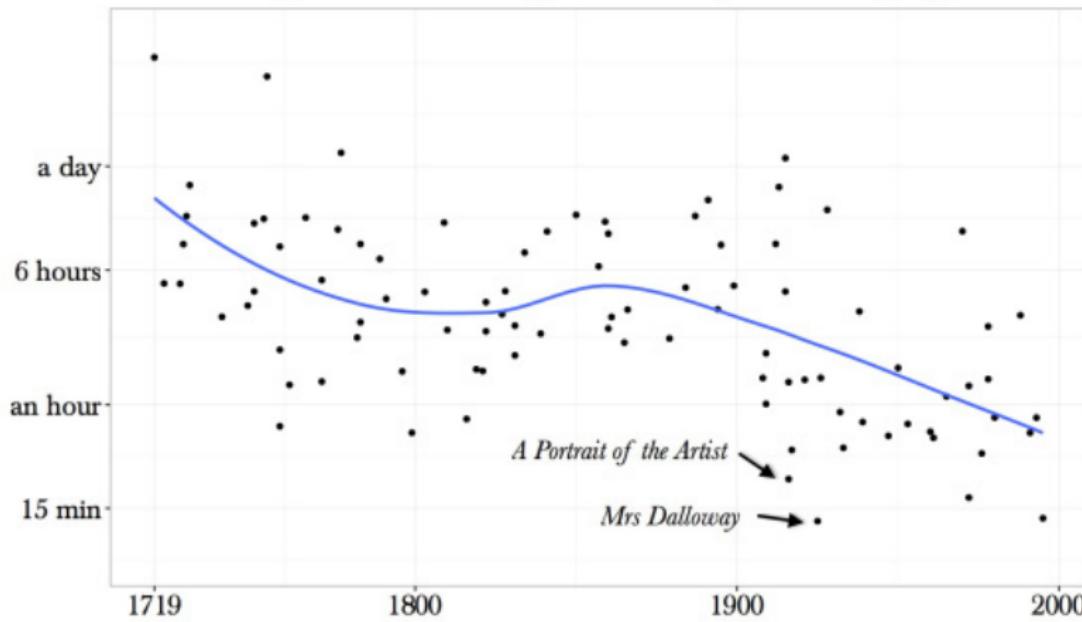


Частотные и уникальные слова в текстах разных стилей

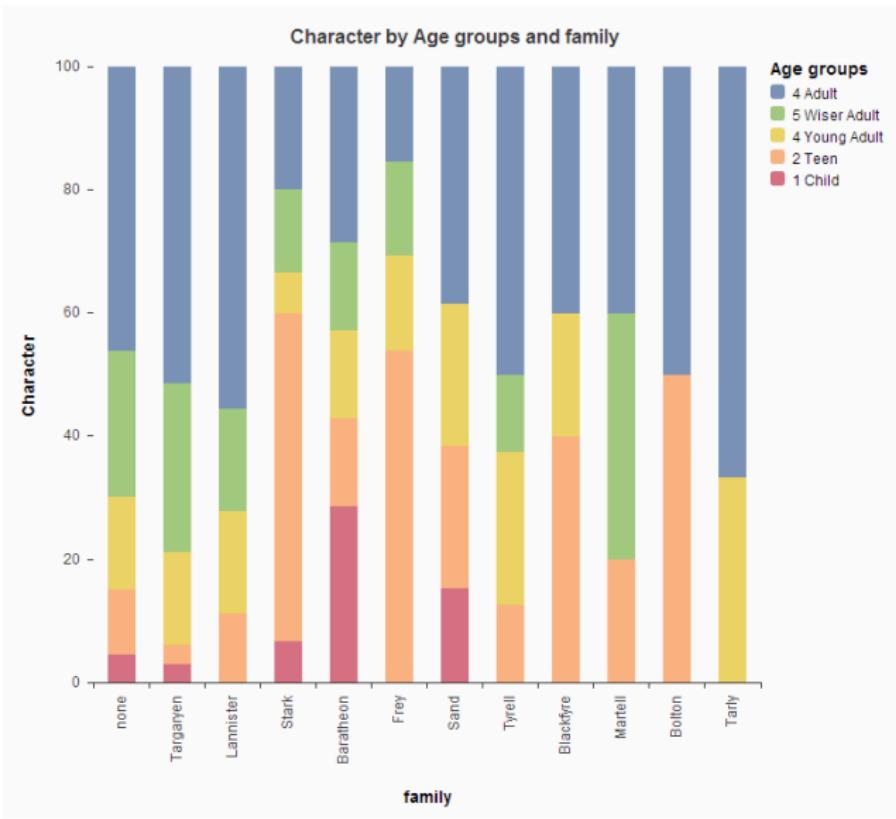


Time narrated, on average, in 250 words in a hypothetical world

(where modernism is responsible for changes in narrative pace)



Возраст персонажей



Список признаков

Example (С их помощью которых пытались предсказать смерть персонажей)

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* — осмыслиенные.

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