

ЦЕНТР ЦИФРОВЫХ  
ГУМАНИТАРНЫХ  
ИССЛЕДОВАНИЙ  
НИУ ВШЭ

# ГУМАНИТАРИИ В ПОИСКАХ ТОЧНОСТИ

Как мы пришли к анализу данных

Даниил Скоринкин

# План этой лекции

1. Далеко ли от традиционных humanities – до data science: «номотетические» гуманитарные науки
2. Случай стилометрии: как одна хорошая научная идея 100 лет ждала появления компьютеров
3. Идея Distant Reading: от распределенного литературоведения – к компьютерному

# 1. ТАК ЛИ УЖ ДАЛЕКИ ‘HUMANITIES’ – ОТ ‘SCIENCE’? А ОТ ‘DATA SCIENCE’?

О поиске точных законов и повторяющихся закономерностей в  
гуманитарных науках

# Что делают гуманитарии?

- Естественные науки — открывают закономерности (номотетический подход)
- Гуманитарные науки — описывают единичные уникальные объекты (идеографический подход)

# Что делают гуманитарии?

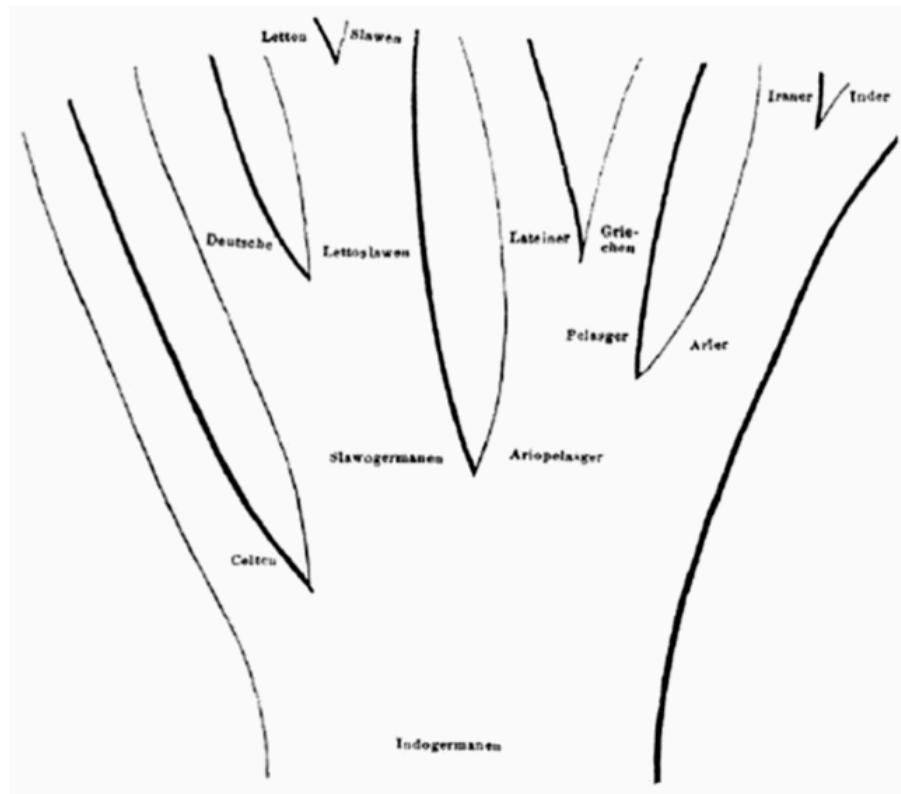
- Естественные науки – открывают закономерности (номотетический подход)
- Гуманитарные науки – описывают единичные уникальные объекты (идеографический подход)

# Гуманитарии давно ищут законы и закономерности

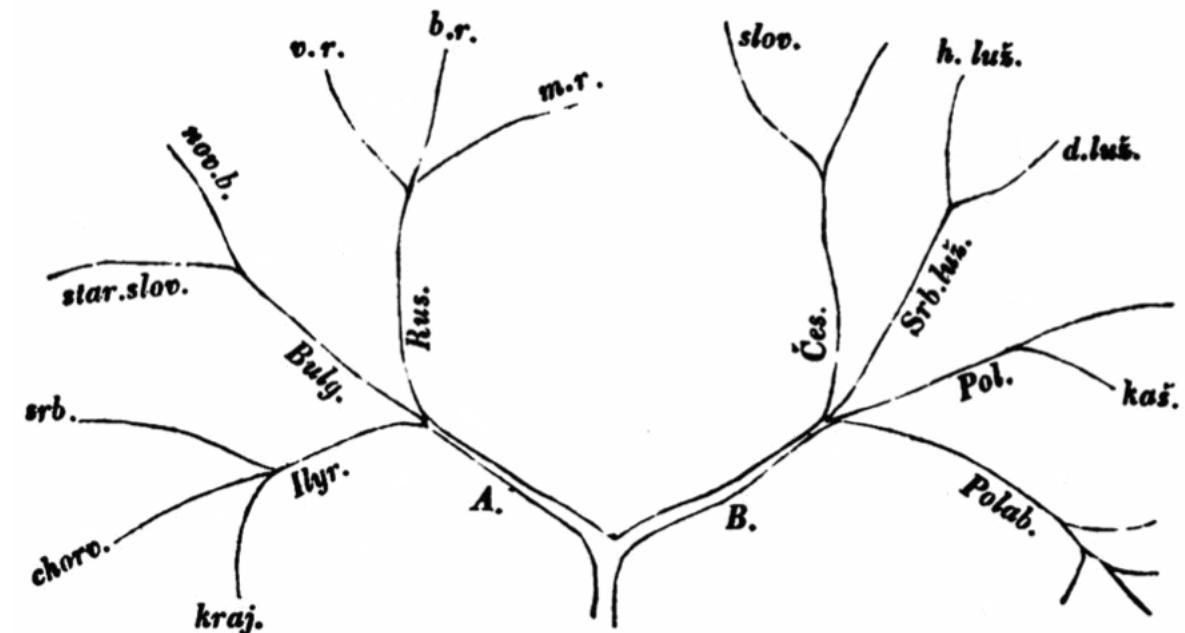
...there were both idiographic and nomothetic practices in every humanistic discipline, and the latter were often dominant.

Bod R. A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present. Oxford: Oxford University Press, 2013.

# Эволюция языков

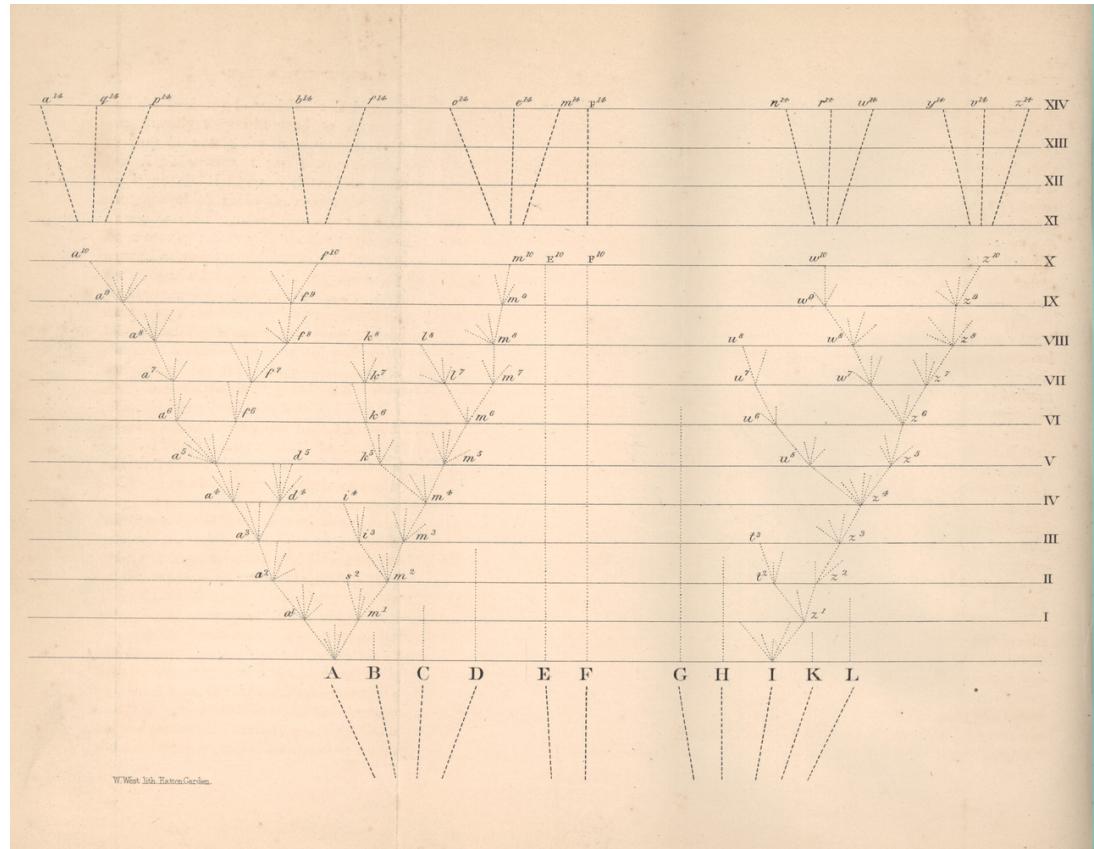


Schleicher, 1853



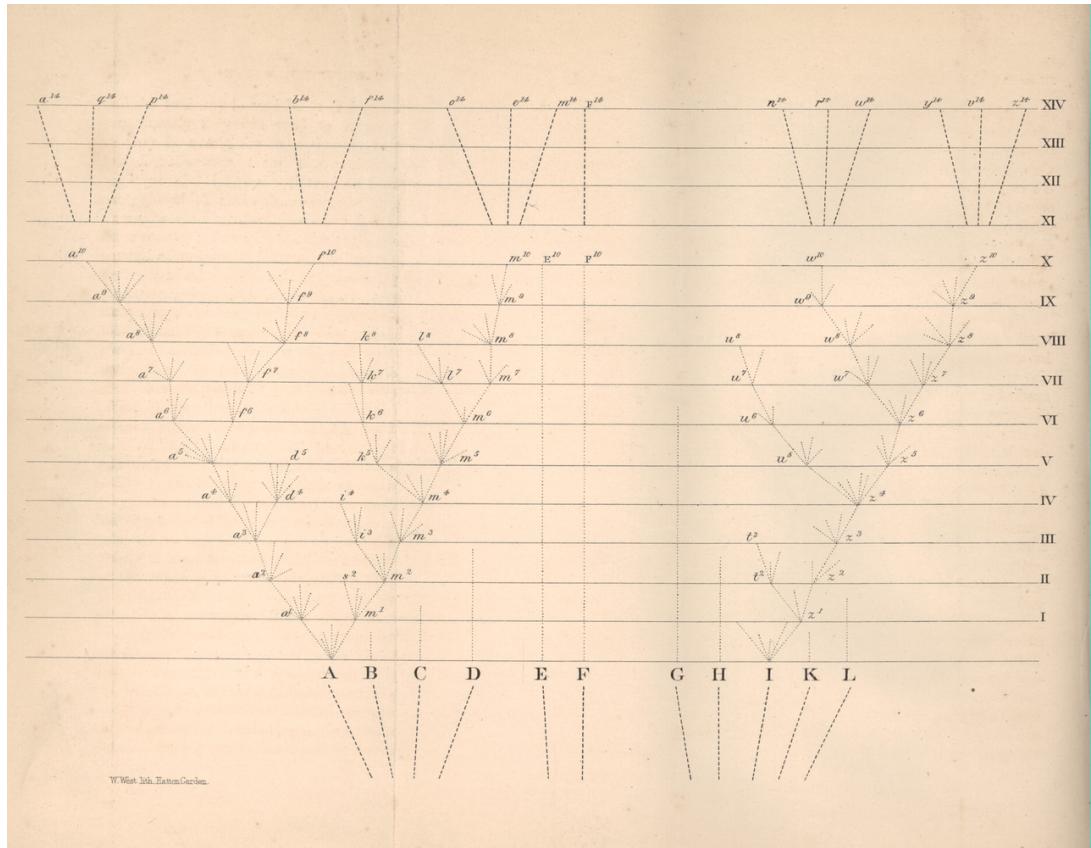
Celakovský, 1853

# Теория эволюции

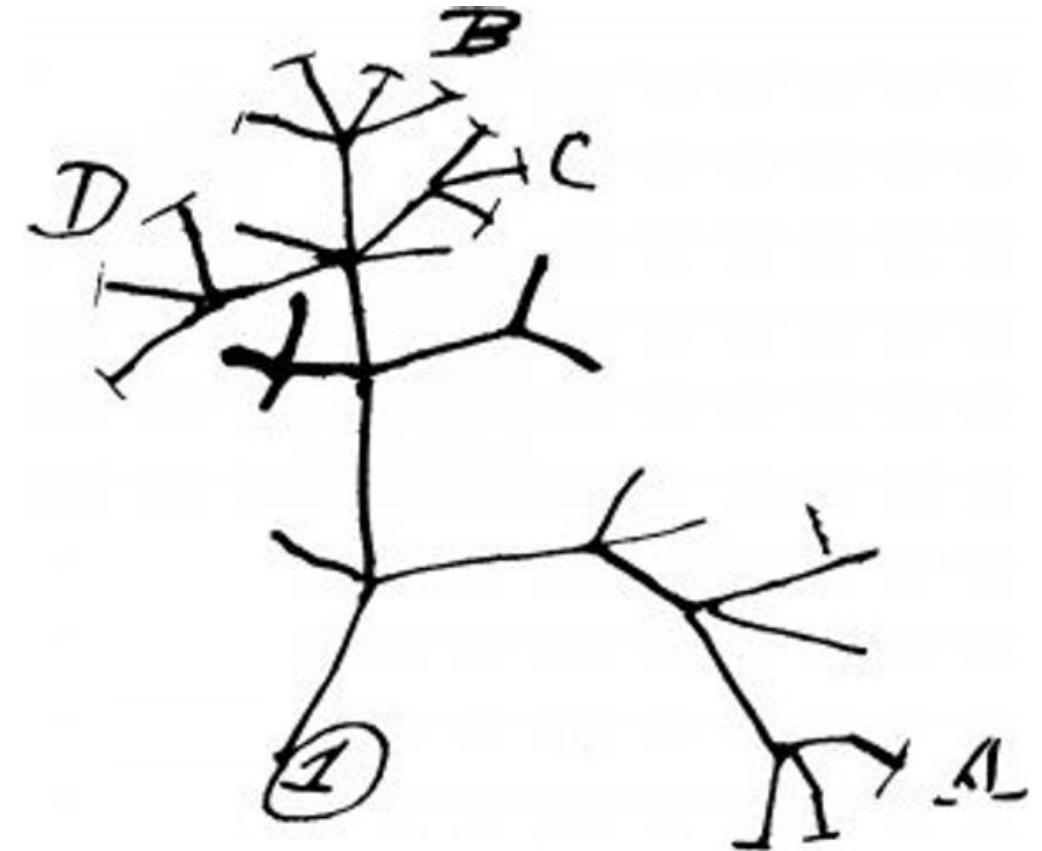


Darwin, On the Origin of Species, 1859  
(кстати, единственная иллюстрация в книге)

# Теория эволюции

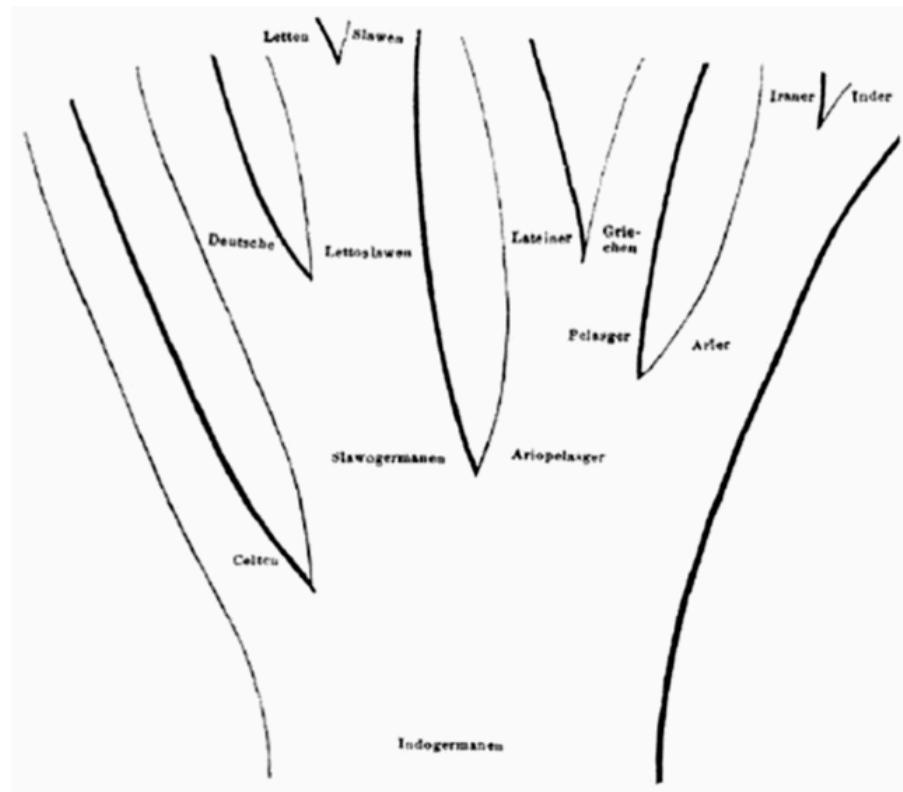


Darwin, On the Origin of Species, 1859  
(кстати, единственная иллюстрация в книге)

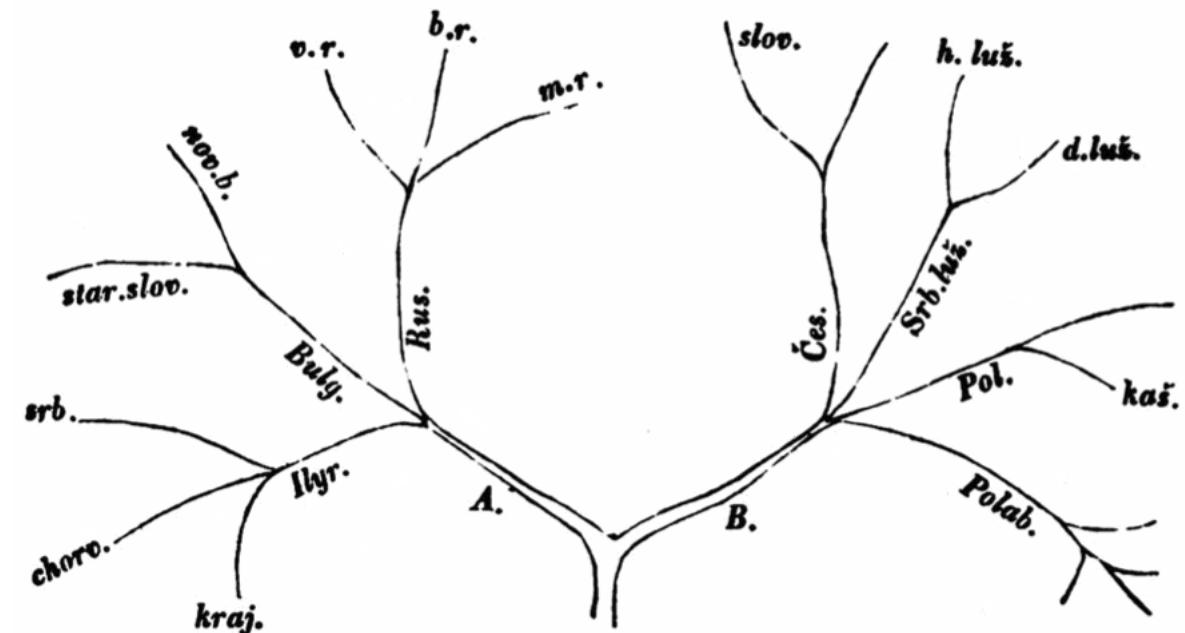


Darwin, 1837  
(набросок в блокноте)

# ЭВОЛЮЦИЯ ЯЗЫКОВ

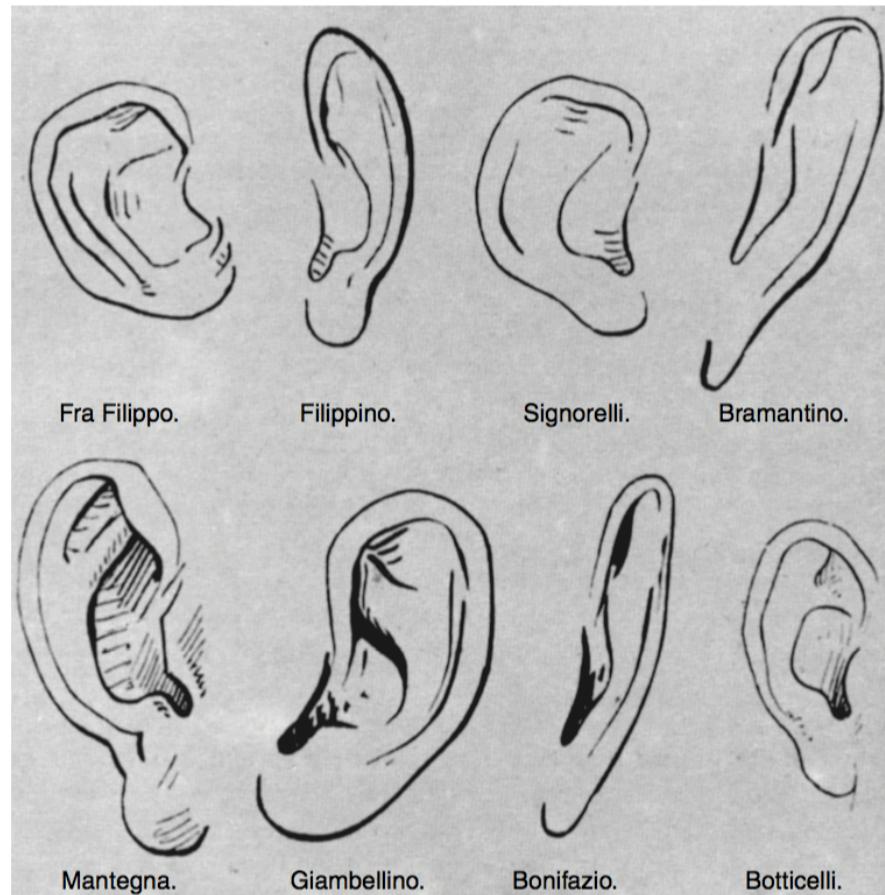


Schleicher, 1853



Celakovský, 1853

# Джованни Морелли и «стилеметрия» живописи



# Диттенбергер, Менденхол, Лютославский, Морозов: авторство и датировка в филологии

## SCIENCE.—SUPPLEMENT.

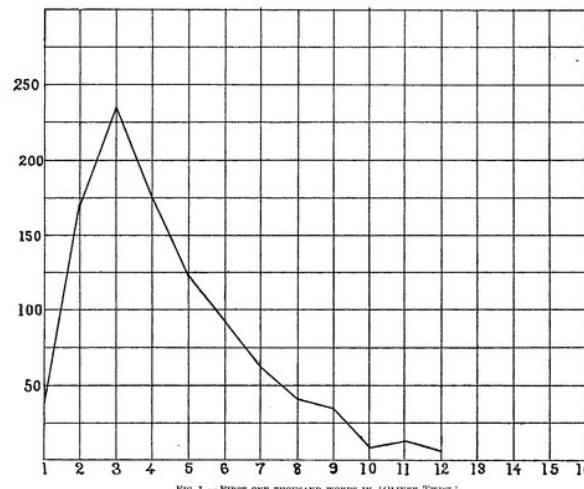
FRIDAY, MARCH 11, 1887.

### THE CHARACTERISTIC CURVES OF COMPOSITION.

AUGUSTUS DEMORGAN somewhere remarks (I think it is in his 'Budget of paradoxes') that some time somebody will institute a comparison among writers in regard to the average length of

mean word-length suggested itself. The new method, while scarcely more laborious than that proposed by DeMorgan, promised to yield results more quickly and of a definitely higher order. It also had the advantage of including, in its application, all that was necessary to the determination of mean word-length; so that, in reality, it furnished two distinct tests.

Preliminary trials of the method have furnished



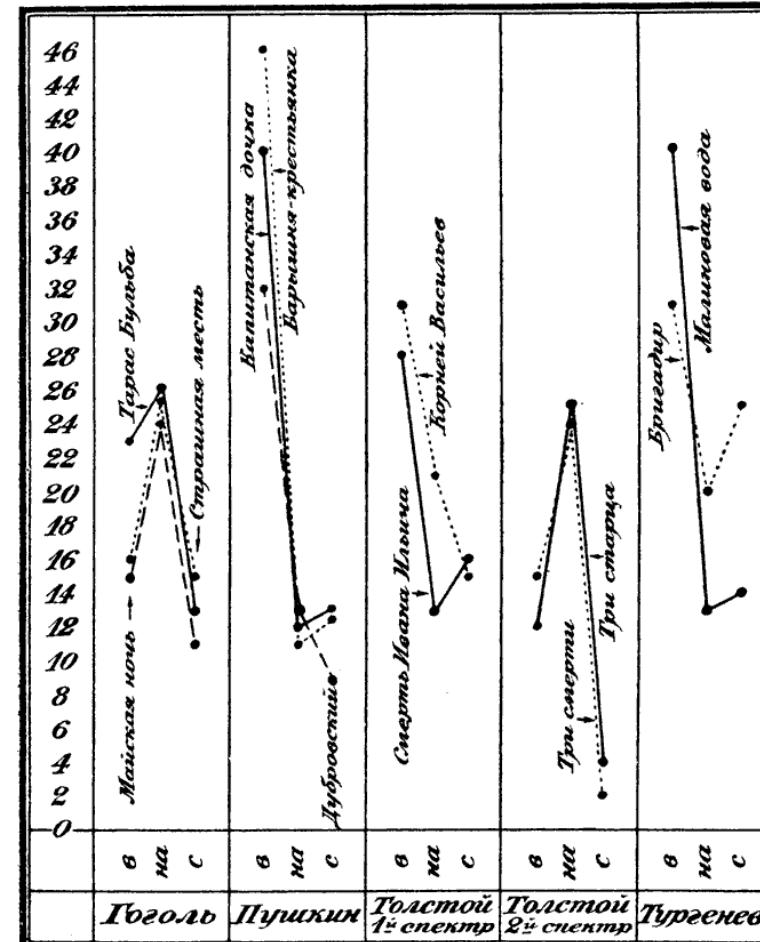
words used in composition, and that it may be found possible to identify the author of a book, a poem, or a play, in this way.

In reflecting upon this remark at various times within the past five or six years, always with the determination to test the value of the suggestion whenever time for the work seemed available, a more comprehensive and satisfactory method of analysis than that based simply upon

strong grounds for the belief that it may prove useful as a method of analysis leading to identification or discrimination of authorship, and it is therefore brought to the attention of the scientific and literary public in the hope that some one may be found who is at once able and willing to secure a satisfactory test of its validity.

The nature of the process is extremely simple, but it may be useful to point out its similarity to

Mendenhall T.C. The Characteristic Curves of Composition // Science. 1887. Vol. ns-9, № 214S. P. 237–246.



Морозов Н.А. Лингвистические спектры как средство для отличия плалиятов от истинных произведений известных авторов и для определения их эпохи. 1927 (1915)

# Первая волна количественного литературоведения

- Позитивизм середины XIX века (Конт и последователи)
- Статистика в стиховедении:
  - Чернышевский и его подсчеты метра (1850-е)
  - «Орфические гимны» Новосадского (1900)
  - «Символизм» Андрея Белого (1910)



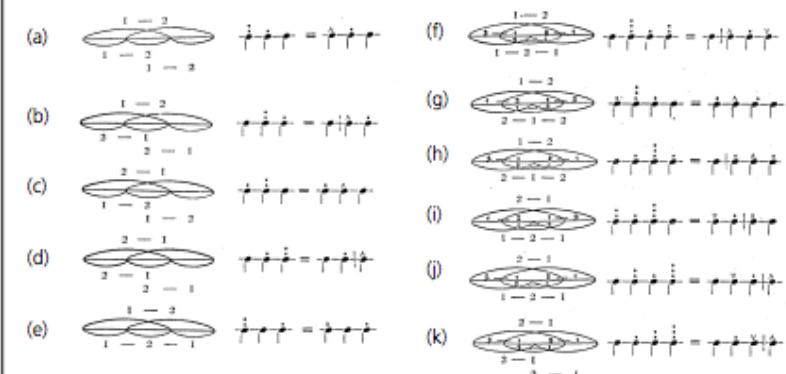
Сумма ускорений ямбического диметра (на 596 строках):	Ломоносов	Державин	Богданович	Дмитриев		
На первой стопе .....	13	46	24	25		
+ второй стопе .....	139	139	114	100		
+ третьей стопе .....	272	263	271	251		
+ первой и третьей .....	5	26	9	14		
+ второй и третьей .....	11	1	5	5		
+ первой и второй .....	*	*	*	*		
Капнист	Батюшков	Жуковский	Пушкин			
На первой .....	35	28	90	110		
+ второй .....	112	33	52	33		
+ третьей .....	230	313	280	341		
+ первой и третьей .....	7	7	44	60		
+ второй и третьей .....	9	5	2	1		
Лермонтов	Языков	Баратынский	Тютчев	Бенедиктов	Павлова	
На первой .....	101	126	164	115	59	107
+ второй .....	47	13	4	62	24	72
+ третьей .....	321	388	325	342	343	271
+ первой и третьей .....	58	85	62	76	30	44
+ второй и третьей .....	*	2	1	2	*	3
+ первой и второй .....	*	*	*	*	*	1
Полонский	Фет	Майков	Мей	Некрасов	А. Толстой	
На первой .....	96	139	77	123	81	83
+ второй .....	43	34	24	17	42	13
+ третьей .....	284	330	299	352	347	323
+ первой и третьей .....	40	62	36	65	44	41
+ второй и третьей .....	2	*	*	2	*	*
Мережковский	Сологуб	Брюсов	Блок	Городецкий		
На первой .....	86	146	73	13	77	
+ второй .....	16	27	48	173	11	
+ третьей .....	359	313	286	282	274	
+ первой и третьей .....	55	74	36	55	29	
+ второй и третьей .....	*	*	*	4	*	

# Музыка: формальные подходы к гармонии и размеру

- Мориц Хауптманн, Die Natur der Harmonik und der Metrik: zur Theorie der Musik (1853)
- Hugo Riemann, Musikalische Syntaxis: Grundriß einer harmonischen Satzbildungslehre (1877) – элементарные «мотивы» в музыке
- Fred Lerdahl and Ray Jackendoff, A Generative Theory of Tonal Music,

## Hauptmann's patterns of accentuation

Hauptmann holds that any basic (two-part) metrical formation may be “positive” by beginning with an accented element that progresses to an unaccented element (which he represents as “1—2”) or may be “negative” by beginning with an unaccented element progressing to an accented one (“2—1”). Since he conceives of triple and quadruple meters as originating out of two-part metrical formations, he can generate a variety of accentual patterns by allowing each component formation to be positive or negative. In the case of triple meter, eight possible patterns may result; for quadruple meter, the total increases to thirty-two patterns. A selection of patterns for each meter are shown here. By “adding up” the various accented elements within a pattern, Hauptmann generates beats that have differing accentual weights. Thus in the triple pattern (b), the first beat has no accentuation, the second beat has double accentuation, and the third beat has single accentuation. Hauptmann further interprets this pattern to represent a metrical group that begins with an upbeat leading to the metrical downbeat and concluding with the second beat (which also has some accentual strength). In pattern (c), the first beat has single accentuation, the second beat, double accentuation, and the third beat remains unaccented. Here, Hauptmann sees the first beat as the metrical downbeat, with the second beat having stronger accentuation. Although Hauptmann presents these many metrical patterns in the abstract, he means them to represent actual musical situations. Thus pattern (c) just discussed would represent the special weight accorded the second beat in a sarabande, for example. Or a four-beat motive that features a crescendo anacrusis to a downbeat would take the form shown in pattern (k).



Но вернемся к  
литературоведению

# Русский формализм: ОПОЯЗ и другие

- Борис Томашевский: инженер, математик-статистик — и выдающийся филолог; занимался статистикой стиха:
  - Ритмика четырехстопного ямба по наблюдениям над стихом «Евгения Онегина» (1917)
  - Пятистопный ямб Пушкина (1919)
- Виктор Шкловский: филология должна «выдерживать научную критику»
- Борис Эйхенбаум: «пропаганда объективно-научного отношения к фактам», «пафос научного позитивизма»



В настоящей работе практические, историко-литературные интересы стояли на втором плане. Тем не менее установленные факты имеют и историко-критическое применение. Вот пример такого применения. Много бумаги исписано — принадлежит ли Пушкину Зуевское «Окончание Русалки». Анализировался каждый стих этого «окончания», но в массе этого стиха не исследовали. Конечно — о каждом стихе могут быть бесконечные сомнения, как и о каждом слове: употребил ли его Пушкин или нет. Но ритм есть инерция, создаваемая цепью стихов. И эта инерция индивидуальна для поэта. Подделать слова — легко. Подделать ритм возможно лишь после тщательного изучения его, чего у Зуева понятно не было.

«Русалка» Зуева дает следующие цифры:

	Ударения на 2 сл.	на 4 сл.	на 6 сл.	на 8 сл.
Зуев . . . .	86,3	66,8	88,4	47,2
Пушкин				
В драматическом бесцезурн. стихе				
minim	80,3	69,1	82,5	50,6
maxim	87,0	76,8	87,3	59,8

Словоразделов

после 2 сл. 3 сл. 4 сл. 5 сл. 6 сл. 7 сл. 8 сл. 9 сл. 10 сл. 11 сл.  
Зуев . . . 24,1 34,7 77,4 10,0 31,7 45,2 40,7 25,1 41,2 58,8

Пушкин

Вдрам. бес-  
цезурн.стихе

minim	27,1	30,8	53,6	18,8	32,2	34,0	41,8	20,8	34,0	56,6
maxim	36,3	38,3	66,5	27,2	40,5	41,6	48,5	32,1	43,4	66,0

# «Анализ тональности прозы»: М. О. Лопатто

Элементы стиля у Пушкина.

	ВСЕГО.			Существит.			Глаголь.			Эпитет.		
	Существит.	Глаголь.	Эпитет.	Мыслим.	Эмоц.	Чувств.	Мыслим.	Эмоц.	Чувств.	Мыслим.	Эмоц.	Чувств.
Арапъ П. В. I .	250	208	116	307	21	22	149	36	9	84	20	12
Станціон Смотр., окончаніе . .	183	168	60	48	5	130	42	11	115	28	4	28
Пиковая Дама, 3	358	335	117	120	12	226	102	27	196	44	14	59

Лопатто М. О. Повести Пушкина: Опыт введения в теорию прозы. // Пушкинист. Историко-литературный сборник. Под ред. С. А. Венгерова, Вып. III. Пг., 1918. С. 3-50.

# В.Я. Пропп и морфология сказки

Подготовительная функция  
дарителя:

Просьбы

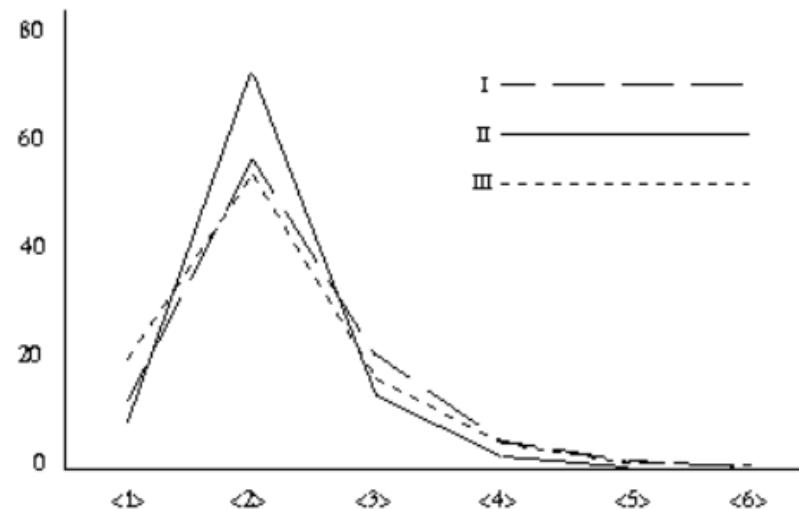
Испытание D<sup>1</sup>  
Выспрашивание D<sup>2</sup>  
загробного характера D<sup>3</sup>  
о пощаде и свободе D<sup>4,5</sup>  
о разделе D<sup>6</sup>  
другие D<sup>7</sup>  
Попытка уничтожить D<sup>8</sup>  
Схватка D<sup>9</sup>  
Предложение обмена D<sup>10</sup>

Формы передачи  
волшебного средства:

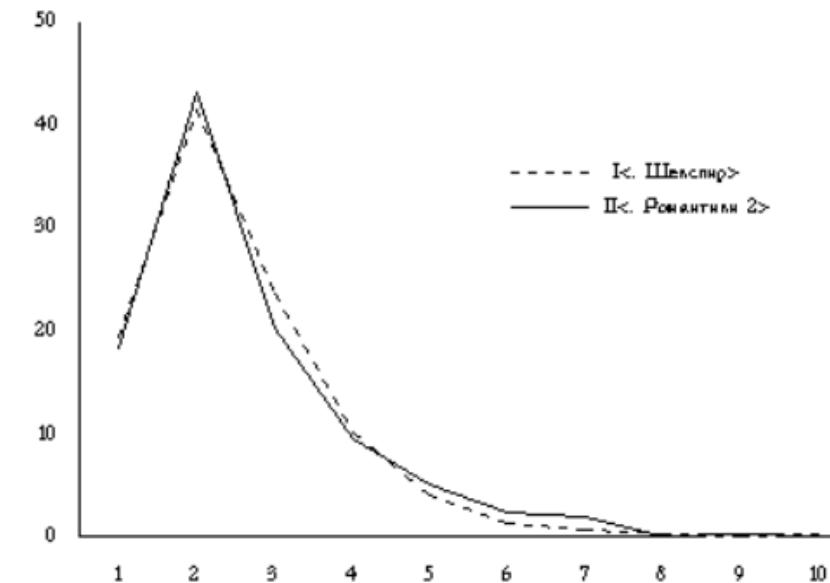
Z<sup>1</sup> Передача  
Z<sup>2</sup> Указание  
Z<sup>3</sup> Изготовление  
Z<sup>4</sup> Продажа  
Z<sup>5</sup> Находка  
Z<sup>6</sup> Появление  
Z<sup>7</sup> Поглощение  
Z<sup>8</sup> Похищение  
Z<sup>9</sup> Предложение услуг

# Б.И. Ярхо и статистические различия между жанрами драмы

<Число говорящих>	1	2	3	4	5	6
I. Ранний Корнелий (С I)	13,2	57,9	21,2	5,6	1,6	0,3
II. К<лассики> XVII <века>, 2(без С I)	9,1	73,9	13,8	2,6	0,6	
III. К<лассики> XVIII <века>	20,8	55,2	17,1	4,9	1,5	0,5



Число говорящих	1	2	3	4	5	6	7	8	9	10
I. Шекспир	19,3	41,3	23,6	10,0	4,0	1,2	0,7	0,3	—	0,1
II. Романтизм 2	18,2	43,0	20,2	9,4	5,1	2,2	1,8	—	0,1	—



Ярхо Б. И. Распределение речи в пятиактной трагедии: (К вопросу о классицизме и романтизме) // Philologica, 1997, т. 4, № 8/10, 201–284. (написано в 1930-е)

# Мечта Бориса Ярхо о «точном литературоведении»:

- Цитата: «Сходства между объектами литературоведения и естествознания настолько глубоки и существенны, что именно на них я считал возможным построить весь свой метод»
- Сборник «Методология точного литературоведения» (написан в 30-е, недописан, опубликован лишь в 2006 г. усилиями М. Шапира и М. Акимовой) – монументальный заход на научивание литературы.



Б. И. ЯРХО

**МЕТОДОЛОГИЯ  
ТОЧНОГО  
ЛИТЕРАТУРОВЕДЕНИЯ**

**Избранные труды  
по теории литературы**

Издание подготовили  
М. В. Акимова, И. А. Пильщиков, М. И. Шапир

Под общей редакцией М. И. Шапира



«Языки славянских культур»

Москва 2006

# «Заморозка» формализма

## Памятник научной ошибке

### I

Обостренное внимание, направленное сейчас на так называемый формальный метод, и неприязненность этого внимания легко обнаружимы.

Человек, который утверждает или утверждал, что классовая борьба не простирается на литературу, тем самымнейтрализует определенные участки фронта.

Говорить о лево-направленности сегодняшнего искусства невозможно. И, как будто, само собой интерес изучения в истории литературы переходит к наиболее направленным, так сказать, публицистическим эпохам.

В то же время оказывается, что в не-направленность искусства, его разгруженность там, где она существовала, преследовала свои весьма реальные и направленные цели.

В то же время так называемый формальный метод нельзя рассматривать, как реакцию против революции.

«Обличительный поэт», «Скорбный поэт», «Темный поэт» «Новый поэт», и даже «Новый поэт 2-й».

Б. Эзенбаум пытался провозгласить ревизию формального метода. Ревизия эта началась с правильной замены названия «формальный» метод названием — «морфологический» метод. Это избавляло от двусмысленности самого выражения «формальный» и в то же время точнее указывало на способ анализа.

Поворотным пунктом в эволюции метода явились чрезвычайно важные работы Тынянова, который ввел в литературоведение понятие о литературной функции—разнозначности литературных элементов в разное время.

От первоначального, уже панического, определения, что произведение различается сумме приемов, здесь осталось очень мало. Части литературного произведения не суммируются, а соотносятся. Сама литературная форма при кажущейся своей однозначности оказывается

ними резонаторами литературного произведения. Задание автора написать роман против разочаровщиков, так сказать, роман противореформенный, не удалось. Целевая установка автора не совпадала с обективной ролью его произведения.

### IV

Изучение литературной эволюции должно быть производимо при учете социального контекста, должно быть осложнено рассмотрением различных литературных течений, неравномерно просачивающихся в различные классовые прослойки и различноими вновь создаваемые.

Эти предпосылки определили собою мою последнюю работу о «Матвее Комарове, жителе города Москвы».

Мне казался невыясненным вопрос о визуальном появлении русской прозы в 30-х годах 19-го века.

Отыскивая ее истоки, я установил ее связь с прозой 18 века. От Вельтмана

«Вообще трагедию можно, подобно опере и балету, считать самой безвременной отраслью драматического искусства».

И в другом месте:

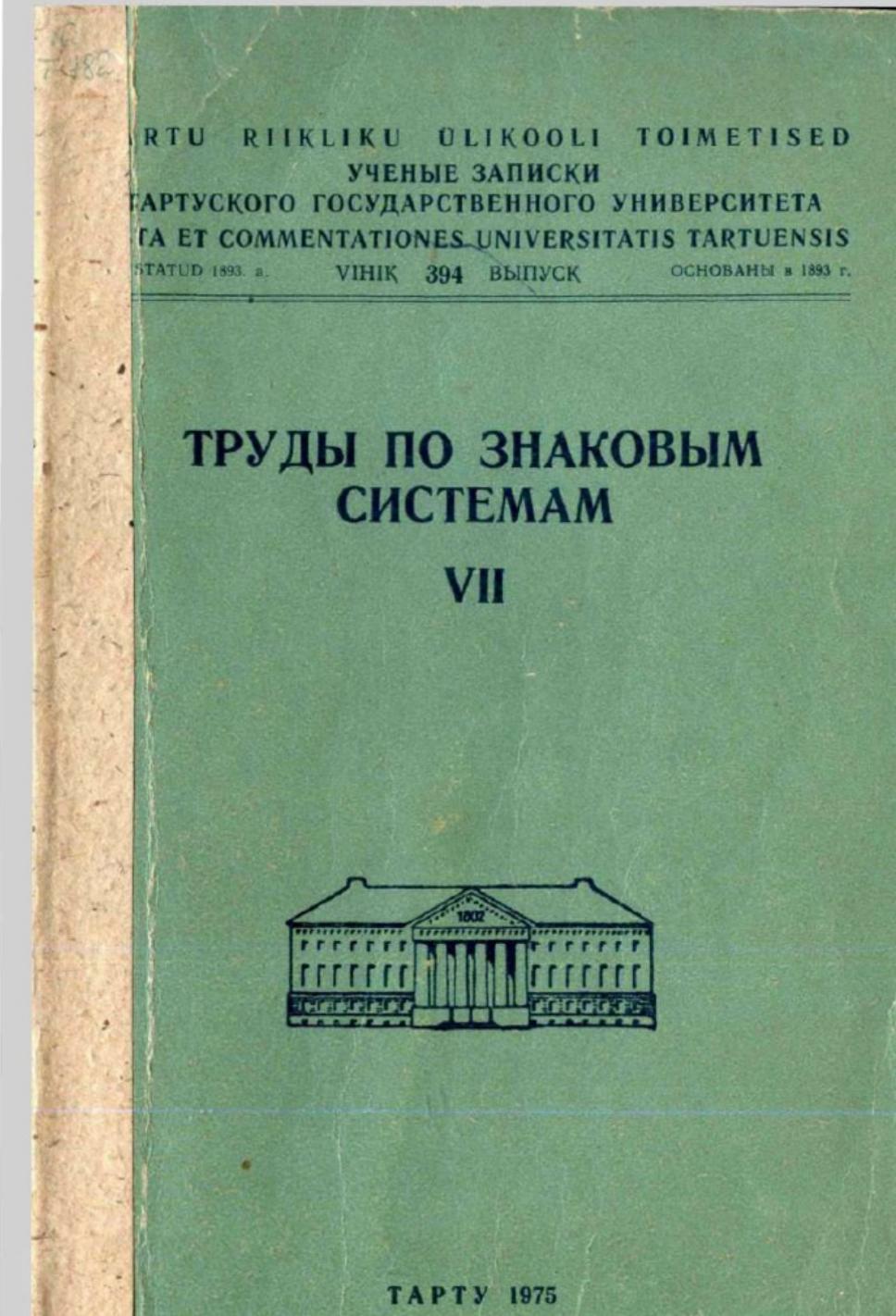
«Если трагедии предоставить более обширное поле, этим уменьшится влияние комедии. Публика, видевшая «Нормана Лири» с меньшим участием будет смотреть «Ревизора». Нашед в трагедии удовольствие чисто литературное художественное, она не будет с такой жаждой искать намека в комедии».

Совершенно ясно, что трагедия, в частности греческая трагедия и трагедия Шекспира, в свое время имела резко направленность. Но позже (ко времени Ольденкопа) трагедия сделалась «литературным удовольствием».

При учете значения учебы у классиков несомненно нужно учитывать этот характер «литературного удовольствия», связанным с самым понятием классики.

# Вторая волна «научного литературоведения»

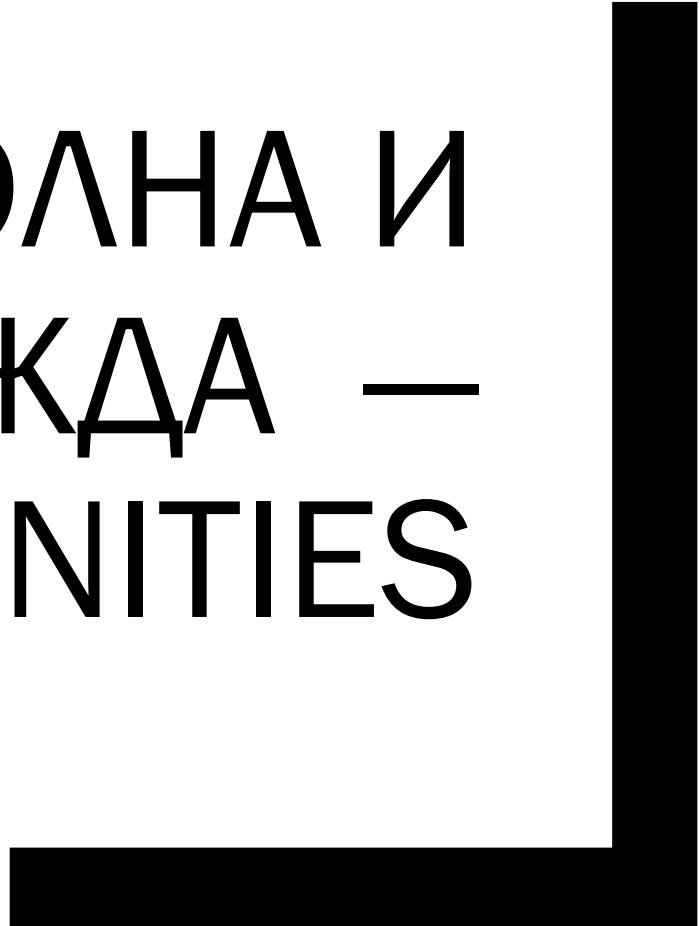
- Структурализм на западе (Леви-Стросс, Барт, Тодоров, Кристева, Греймас)
- Семиотика в СССР (Ю.М. Лотман, Б.А. Успенский, Ю.И. Левин, И. и О. Ревзины)
- Лотман (1967): «Литературоведение должно быть наукой»
- Роман Якобсон как передатчик эстафеты (Формализм -> Структурализм/Семиотика)

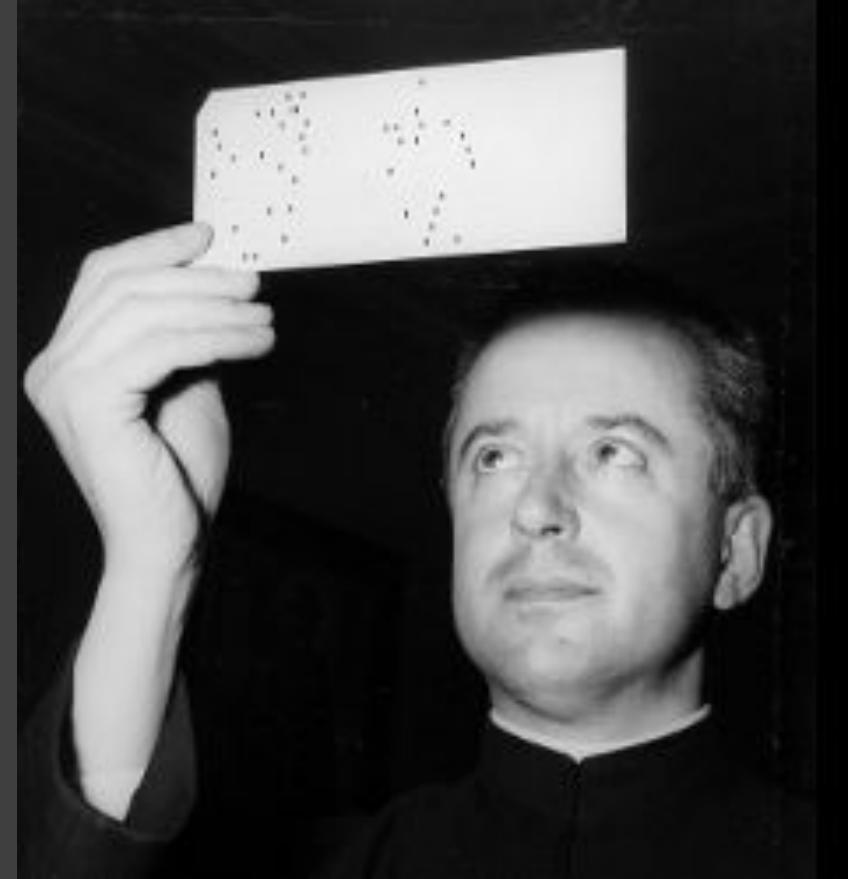


«Создается впечатление, что в каждом поколении литературоведов находятся те, кто более-менее самостоятельно осознает важность точных методов (формалисты и структуралисты — лишь два известных примера) и каждый раз терпит поражение. В этом есть что-то почти мистическое: необъяснимая невидимая сила постоянно разрушает недостроенное здание научного литературоведения»

Собчук О. Номотетическое литературоведение: пунктирный набросок // Новое литературное обозрение, 2 (132). 2015. с. 102-114

ТЕПЕРЬ ТРЕТЬЯ ВОЛНА И  
НОВАЯ НАДЕЖДА –  
DIGITAL HUMANITIES





ГУМАНИТАРИЙ И КОМПЬЮТЕРЫ, ГОД  
1949

# 2. SUCCESS STORY СТИЛОМЕТРИИ

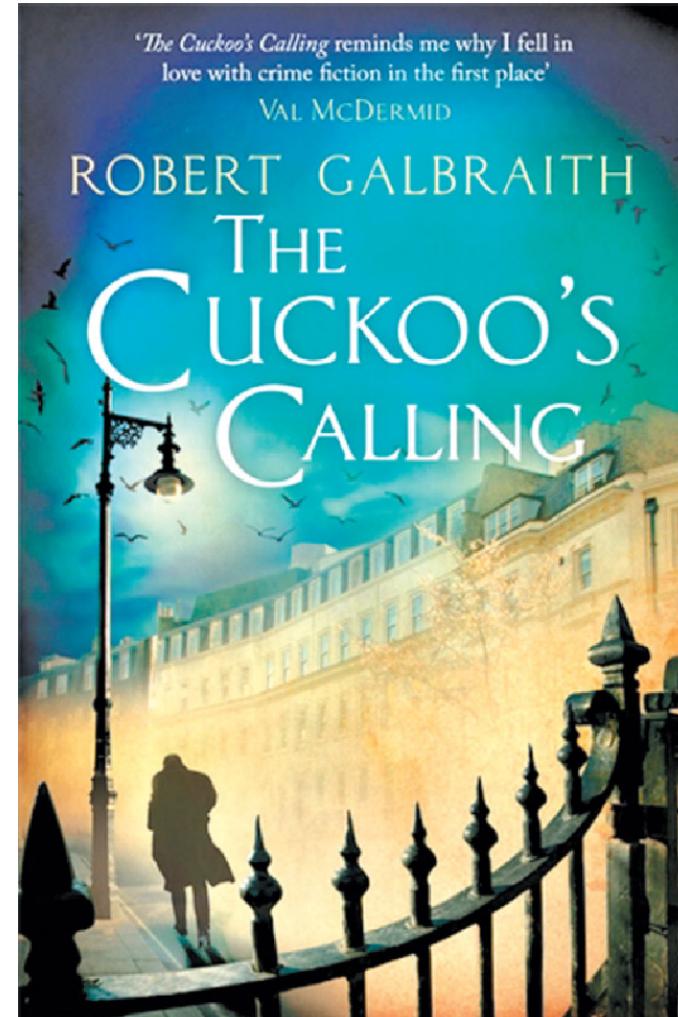
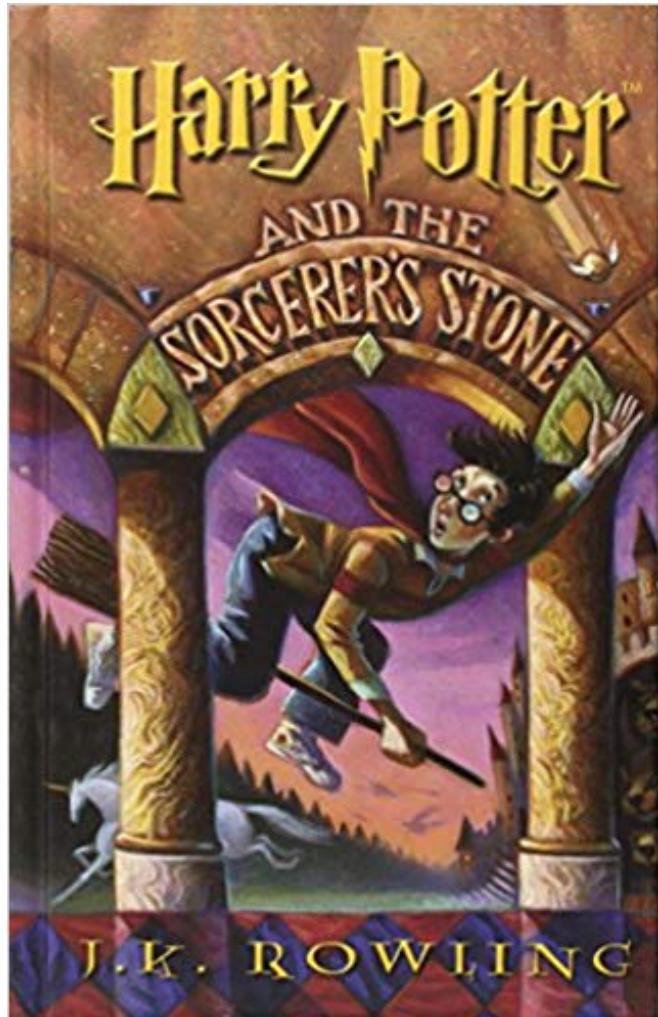
как одна хорошая научная идея  
100 лет ждала появления компьютеров





КТО  
НАПИСАЛ  
ЭТУ КНИГУ?

# Джоан Роулинг versus Роберт Гэлбрейт



TECHNOLOGY

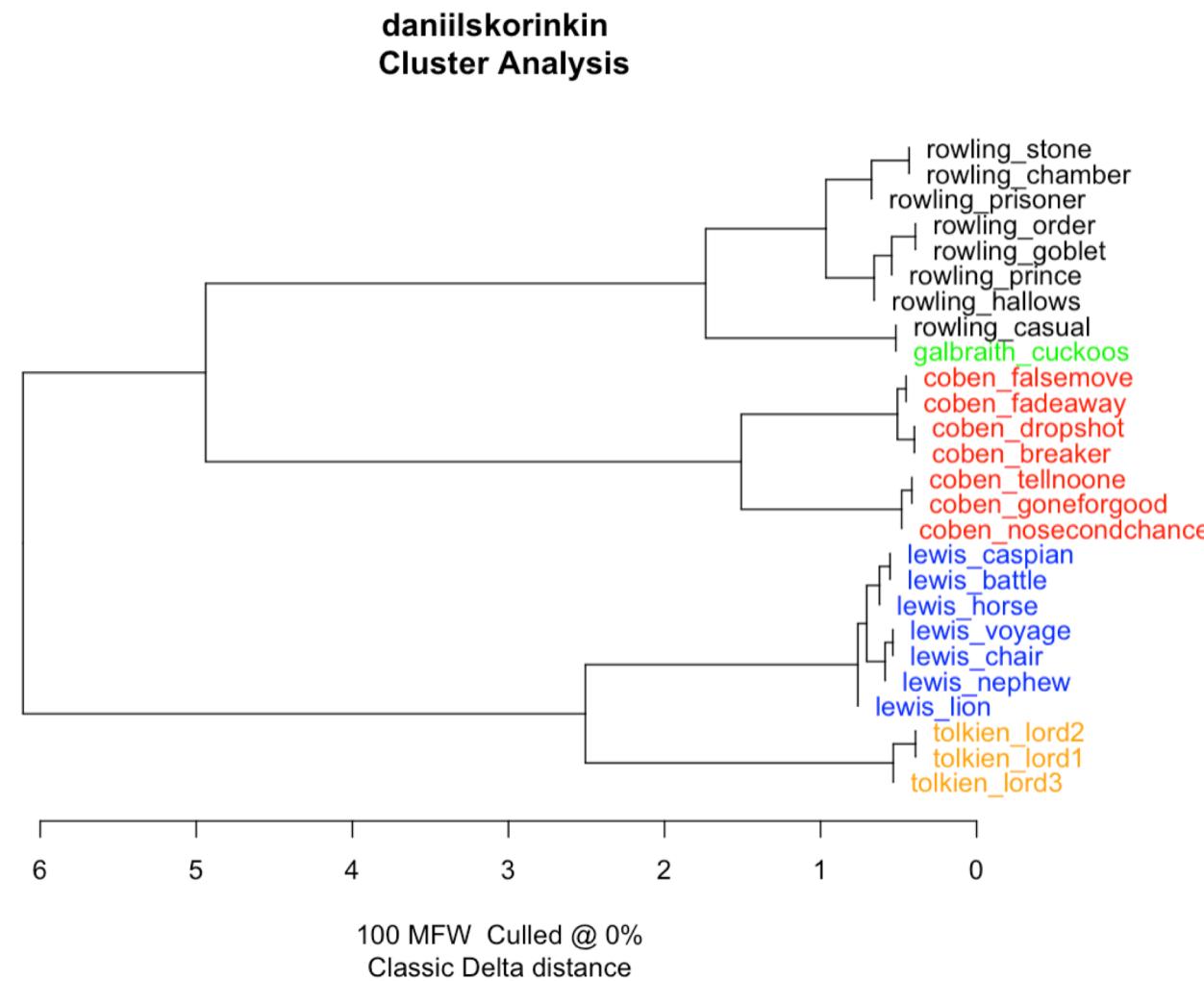
# HOW COMPUTER ANALYSIS UNCOVERED J. K. ROWLING'S SECRET NOVEL

OR, HOW YOUR FOUR-GRAMS MAY BE UNDERMINING YOUR ANONYMOUS EROTICA-WRITING CAREER

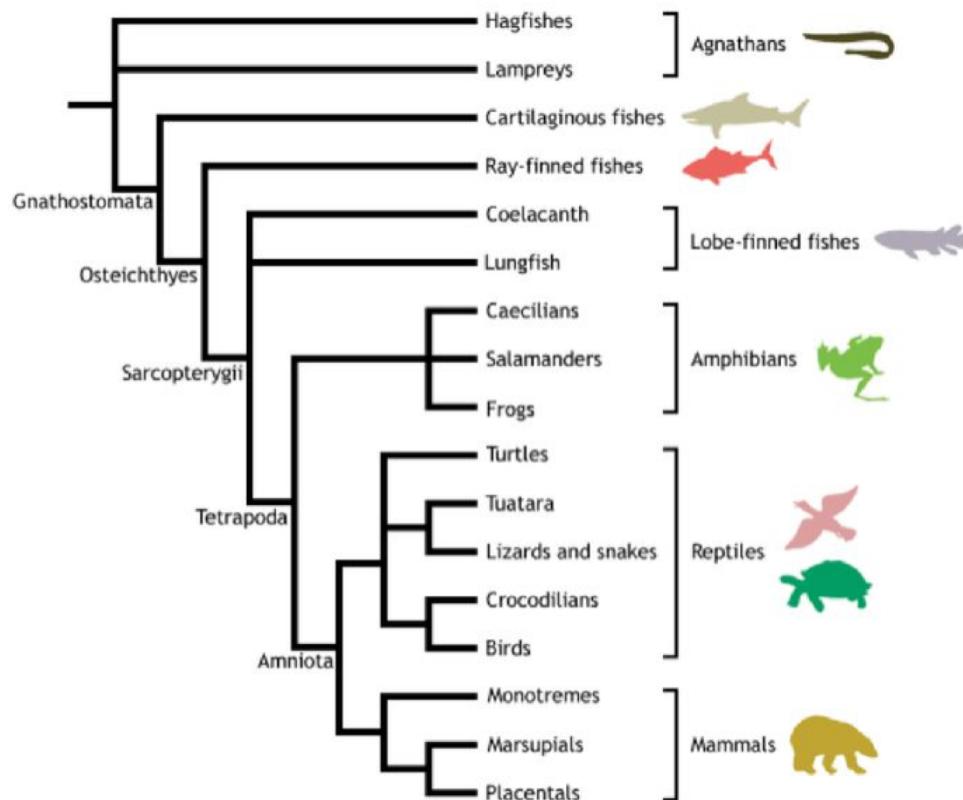
By Francie Diep July 18, 2013



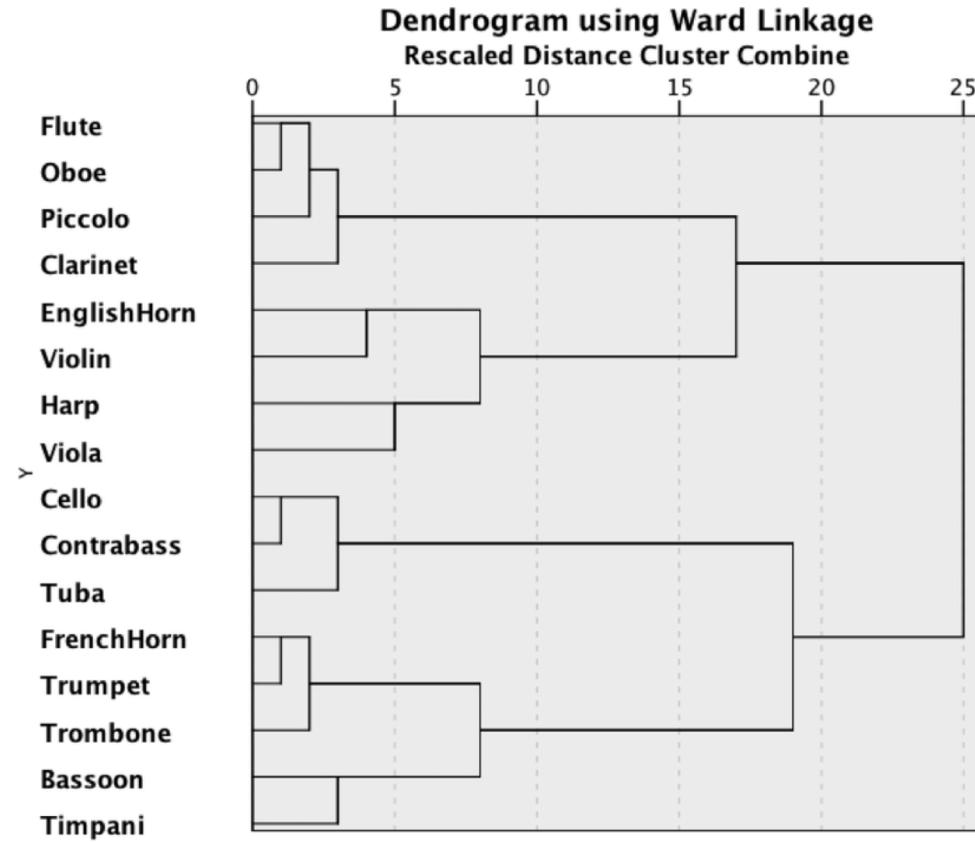
# Это дендрограмма стилометрической «близости»



Она отражает близость объектов по каким-то свойствам — как и вообще любая дендрограмма

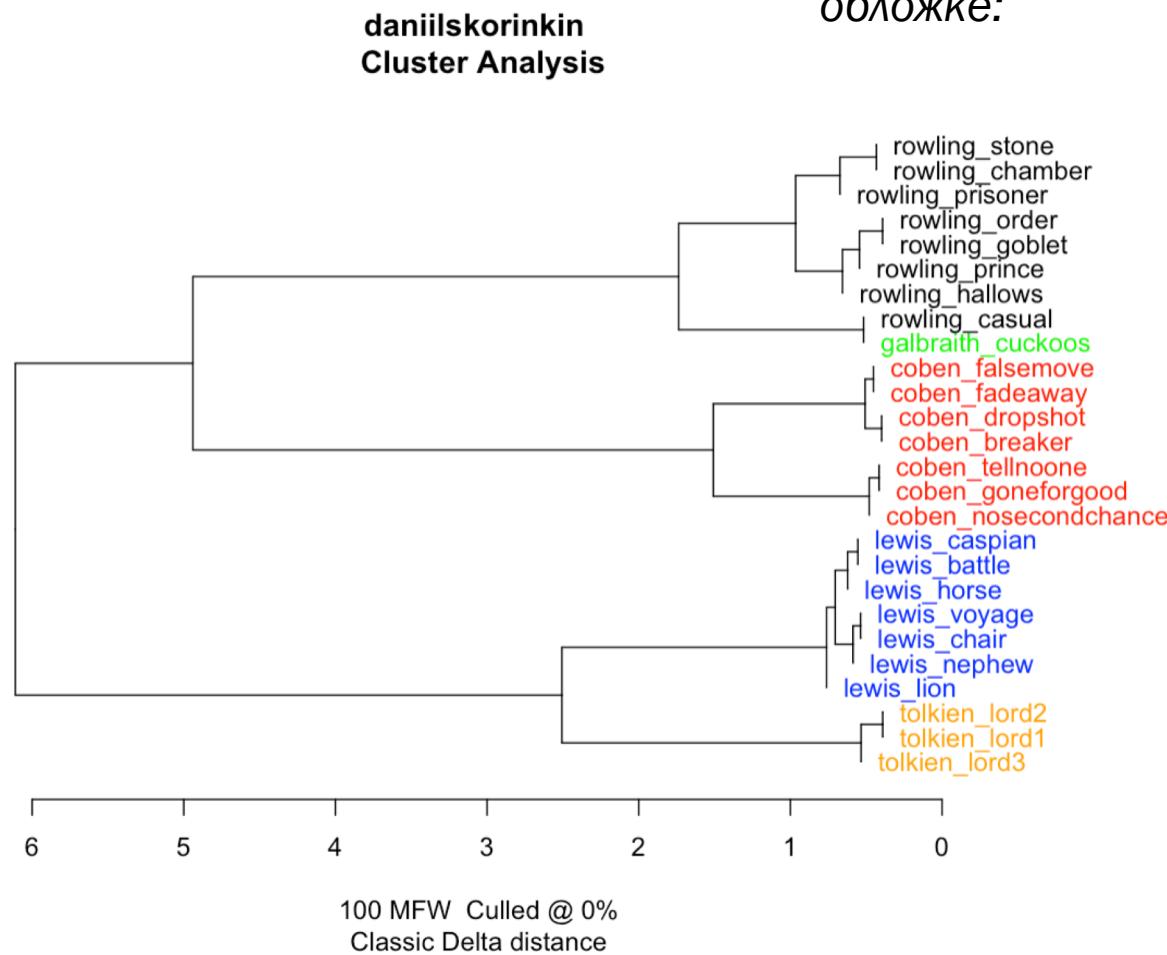


Она отражает близость объектов по каким-то свойствам — как и вообще любая дендрограмма



# Вернемся к дендрограмме стилометрической «близости»

Цвет подписи — по автору на  
обложке:



Как же это работает? 🤔

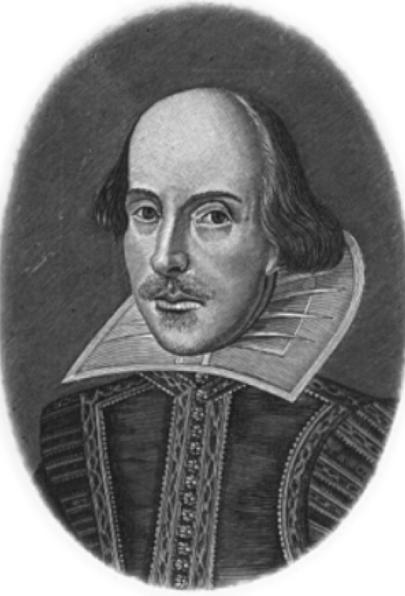
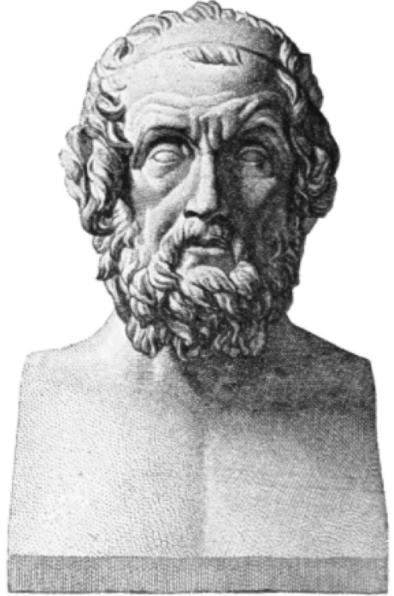
# Стилометрия (stylometry)

- «Стилометрия – это статистический анализ отклонений между литературными стилями разных авторов или жанров» (Oxford Dictionary)
- «В основе стилометрии лежит гипотеза о том, что у авторского стиля есть не осознаваемая автором составляющая» (Encyclopaedia of Statistical Sciences)

# Стилеметрия (stylometry)

Стилометрические исследования во всем их разнообразии имеют две общие черты: тексты должны быть каким-то образом преобразованы в числа, а числа — исследованы статистическими методами

M. Eder, M. Kestemont, J. Rybicki. ‘Stylo’: a package for stylometric analyses



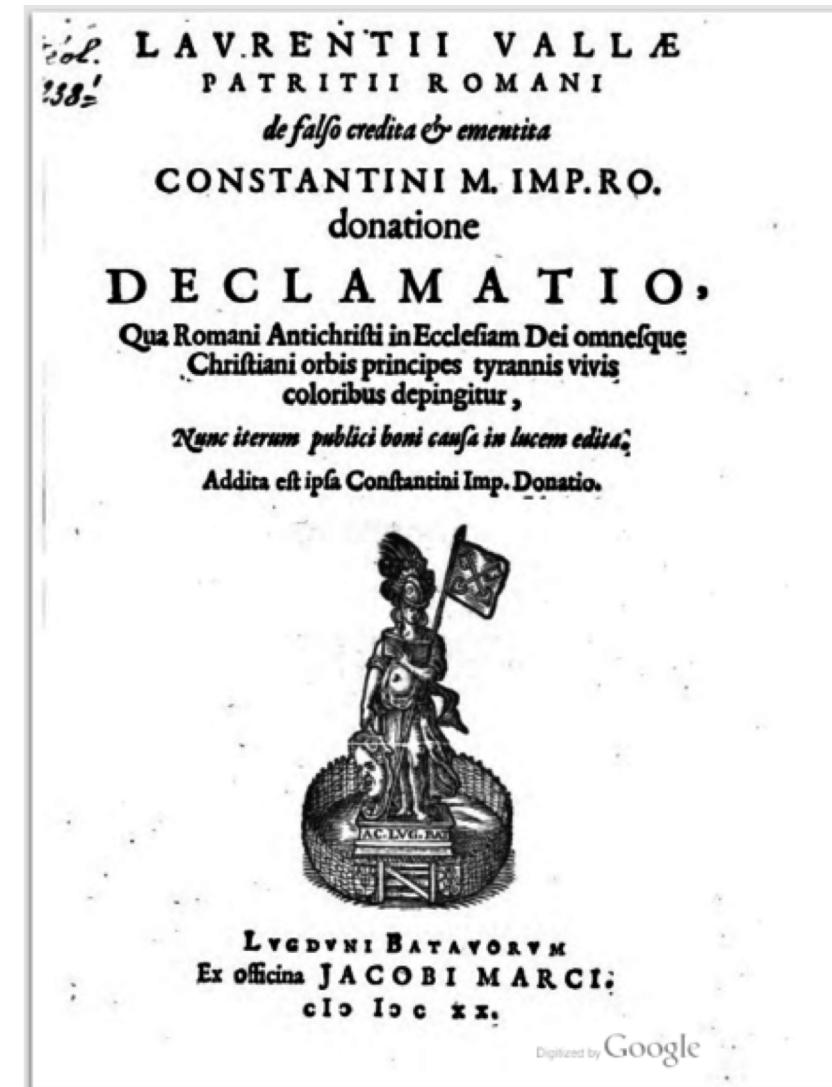
# Когда спор об авторстве решает судьбы государств:

- Лоренцо Валла (1407 – 1457) – итальянский гуманист, эпикуреец, знаток классической латыни, филолог
- В 1439-1440 году доказал, что «Константинов дар» (грамота, якобы написанная Константином Великим и передающая западную часть Римской империи римскому папе)
- Доказал, что текст «Константина дара» написан не той латынью, которая существовала в IV веке, а средневековой латынью VIII века
- Чуть не попал за это на 🔥



# Когда спор об авторстве решает судьбы государств:

- Лоренцо Валла (1407 – 1457) – итальянский гуманист, эпикуреец, знаток классической латыни, филолог
- В 1439-1440 году доказал, что «Константинов дар» (грамота, якобы написанная Константином Великим и передающая западную часть Римской империи римскому папе)
- Доказал, что текст «Константина дара» написан не той латынью, которая существовала в IV веке, а средневековой латынью VIII века
- Чуть не попал за это на 🔥





ЧТО-ТО ПОХОЖЕЕ (НО С  
ПРОТИВОПОЛОЖНЫМ  
РЕЗУЛЬТАТОМ В ИТОГЕ)  
СДЕЛАЛ И.А.  
ЗАЛИЗНЯК

# Слово о полку Игореве

Уже двести лет не прекращается дискуссия о том, что представляет собой «Слово о полку Игореве», — **подлинное древнерусское произведение или искусственную подделку под древность, созданную в XVIII веке.** <...> Гибель единственного списка этого произведения лишает исследователей возможности произвести анализ почерка, бумаги, чернил и прочих **материальных характеристик** первоисточника.

Наиболее прочным основанием для решения проблемы подлинности или поддельности «Слова о полку Игореве» оказывается в таких условиях **язык этого памятника.**

А.А. Зализняк. "Слово о полку Игореве": взгляд лингвиста.

Но что делать, когда у нас нет  
временного разрыва в столетия?

# Первая попытка: зайти через длину слова:

- 1851 – математик Августус де Морган предлагает длину слова как признак авторства
- 1887 – Томас Менденхолл (T. Mendenhall, на фото), The Characteristic Curves of Composition, первая известная работа по количественному определению авторства



Томас  
Менденхолл

# Первая попытка: зайти через длину слова:

- 1851 – математик Августус де Морган предлагает длину слова как признак авторства
- 1887 – Томас Менденхолл (T. Mendenhall), *The Characteristic Curves of Composition*, первая известная работа по количественному определению авторства

## SCIENCE.—SUPPLEMENT.

FRIDAY, MARCH 11, 1887.

### THE CHARACTERISTIC CURVES OF COMPOSITION.

AUGUSTUS DEMORGAN somewhere remarks (I think it is in his 'Budget of paradoxes') that some time somebody will institute a comparison among writers in regard to the average length of

mean word-length suggested itself. The new method, while scarcely more laborious than that proposed by DeMorgan, promised to yield results more quickly and of a definitely higher order. It also had the advantage of including, in its application, all that was necessary to the determination of mean word-length; so that, in reality, it furnished two distinct tests.

Preliminary trials of the method have furnished

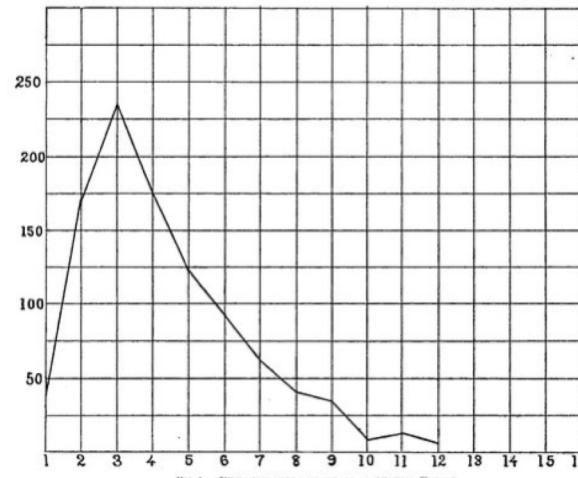


FIG. 1.—FIRST ONE THOUSAND WORDS IN 'OLIVER TWIST.'

words used in composition, and that it may be found possible to identify the author of a book, a poem, or a play, in this way.

In reflecting upon this remark at various times within the past five or six years, always with the determination to test the value of the suggestion whenever time for the work seemed available, a more comprehensive and satisfactory method of analysis than that based simply upon

strong grounds for the belief that it may prove useful as a method of analysis leading to identification or discrimination of authorship, and it is therefore brought to the attention of the scientific and literary public in the hope that some one may be found who is at once able and willing to secure a satisfactory test of its validity.

The nature of the process is extremely simple, but it may be useful to point out its similarity to

и его статья  
The Characteristic Curves of  
Composition

# Параллельно античники начинают применять статистику для решения вопросов о датировке

- 1880 – W. Dittenberger, Sprachliche Kriterien für die Chronologie der Platonischen Dialoge
- 1890 – W. Lutosławski, Principes de stylométrie
- 1897 – W. Lutosławski, The origin and growth of Plato's logic; with an account of Plato's style and of the chronology of his writings



Винцентий  
Лютославский

# Параллельно античники начинают применять статистику для решения вопросов о датировке

- 1880 — W. Dittenberger,  
Sprachliche Kriterien für die  
Chronologie der Platonischen  
Dialoge
- 1890 — W. Lutosławski, Principes  
de stylométrie
- 1897 — W. Lutosławski, The origin  
and growth of Plato's logic; with  
an account of Plato's style and of  
the chronology of his writings

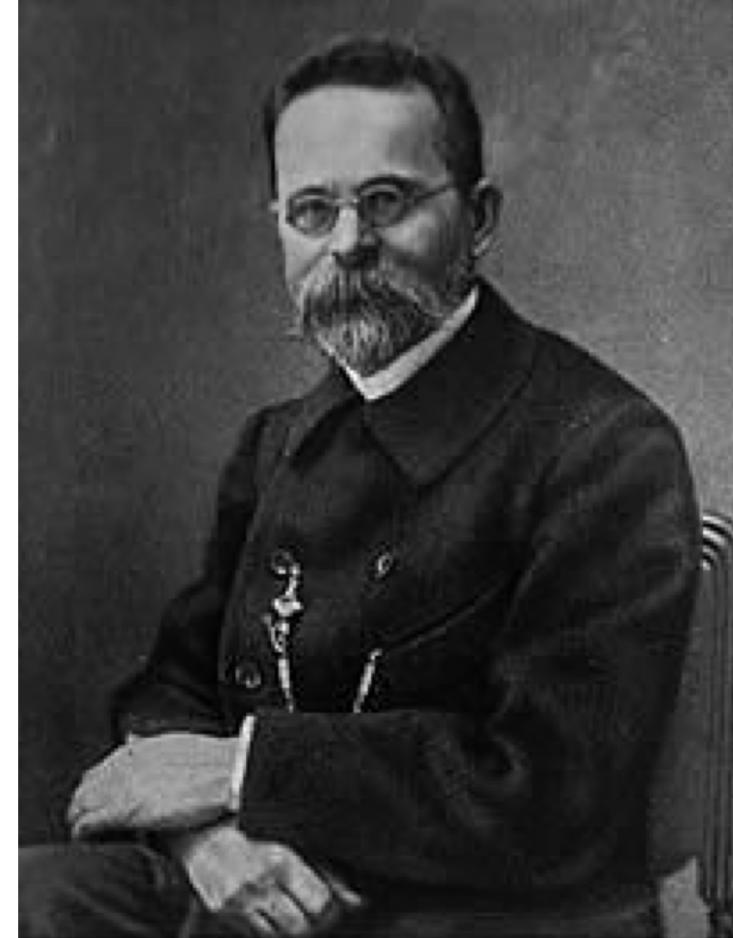
once in two pages (ed. Didot), when we shall call it very frequent. Besides such peculiarities we include here the following special observations whenever they refer to a dialogue :

12. Being the first member of a tetralogy projected later—this refers only to Republic and Theaetetus.
13. Partial prevalence of other teachers over Socrates. This refers only to Symposium and Parmenides. For in Sophist Politicus Timaeus Critias Laws Socrates is already completely supplanted by other teachers, and this constitutes a more important characteristic.
14. Periods less regular.
15. Natural order of words inverted, as generally observed by Campbell.
16. Recurrence of rhythmical cadence, as generally observed by Campbell.
17. Balancing of words to achieve harmony and symmetry.
18. Adjustment of longer and shorter syllables, idem.
19. Words common and peculiar to Timaeus, Critias, Laws more than once in two pages, but less than once in a page.
200. οὐσία less frequent than κούσια.
206. ἀλλὰ μή less frequent than καὶ μή.
306. νοίκευ more than four times oftener than μέντοι.
307. μέντοι less than once in two pages, but over once in five pages.
308. νοίκευ more than once in two pages.
317. ἔπειος prevailing over ἔλεγον.
318. Answers denoting subjective assent less than once in sixty answers.
325. Superlatives in affirmative answers more than half as frequent as positives, but not prevailing over positives.
334. πάντα μήσιν prevailing over πάντα γέ.
365. ξύμπαν prevailing over δρας.
366. ωδῆς and compounds between four and five times in one page.

...и его книга про  
Платона

# Русский последователь Лютославского и ко – Н.А. Морозов

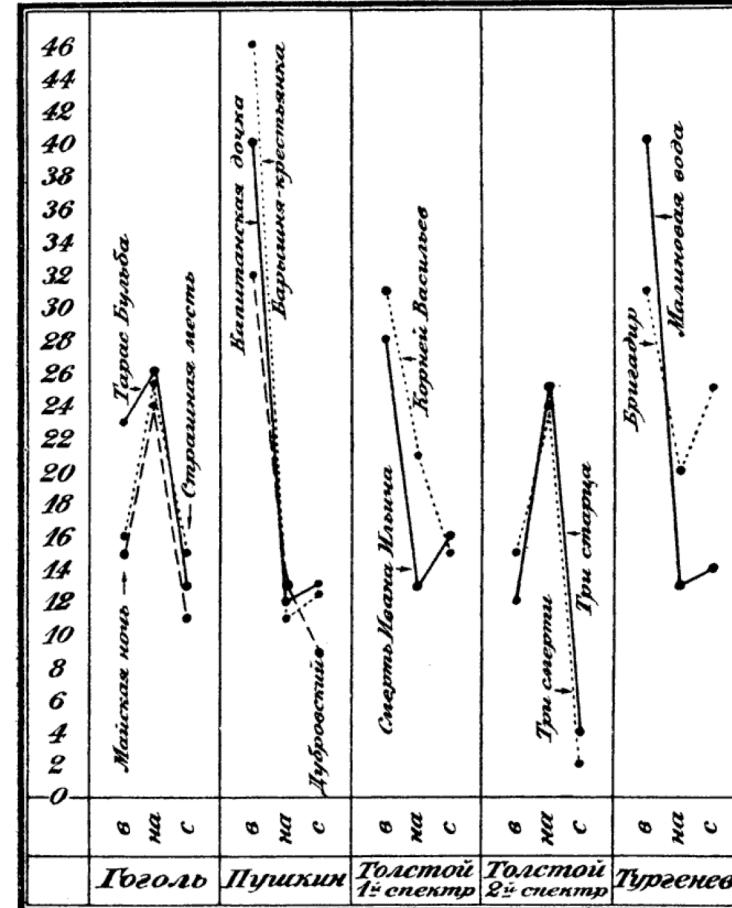
- 1915 – Морозов Н.А.  
Лингвистические спектры  
(вдохновлен Диттенбергером и  
Лютославским)
- Вслед за ними указал на важность  
служебных слов
- Но говорил уже в первую очередь  
об авторстве: «Лингвистические  
спектры, как средство для  
отличения плалиатов от истинных  
произведений того или другого  
известного автора и для  
определения их эпохи»



Николай Морозов

# Русский последователь Лютославского и ко – Н.А. Морозов

- 1915 – Морозов Н.А.  
Лингвистические спектры  
(вдохновлен Диттенбергером и  
Лютославским)
- Вслед за ними указал на важность  
служебных слов
- Но говорил уже в первую очередь  
об авторстве: «Лингвистические  
спектры, как средство для  
отличения плалиатов от истинных  
произведений того или другого  
известного автора и для  
определения их эпохи»



и его статья  
«Лингвистические

# Поиск универсального метода затормозился из-за трудозатратности:

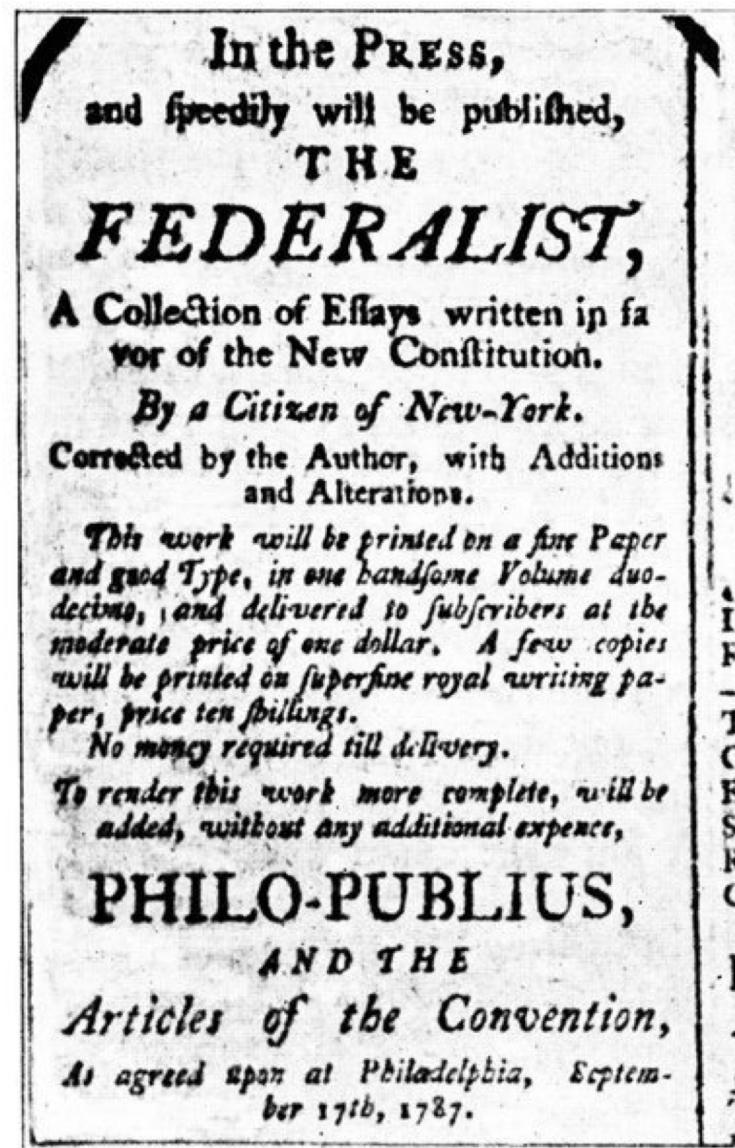
- «... results obtained through tiresome philological labour»  
(Lutosławski, The origin and growth of Plato's logic..., стр. 141)
- «Однако, мое время так было заполнено другими делами, что только летом 1915 г. я нашел несколько свободных дней, чтобы составить лингвистические спектры хотя бы нескольких писателей»  
(Морозов, Лингвистические спектры).



В 1960-Е У УЧЕНЫХ ПОЯВИЛИСЬ  
ЭЛЕКТРОННЫЕ ИНСТРУМЕНТЫ  
ВЫЧИСЛЕНИЯ – И ДЕЛО ПОШЛО

# Записки федералиста

- Серия знаковых статей эпохи Американской революции
- Написаны «отцами-основателями США» под псевдонимами
- 12 спорных (Хэмилтон или Мэдисон)
- Статья Ф.Мостелера и Д. Уоллеса Inference in an Authorship Problem (1963)
- '...определить авторство записок федералиста и предложить стандартный метод для решения проблем авторства'



# Mosteller, Wallace, 1963:

- The function words of the language appear to be a fertile source of discriminators, and luckily the high-frequency words are the strongest.
- <...> it is important to have a variety of sources of material <...>
- Madison is the principal author. These data make it possible to say far more than ever before that the odds are enormously high that Madison wrote the 12 disputed papers. <...>

TABLE 5.3. FINAL WORDS AND WORD GROUPS: ESTIMATED NEGATIVE BINOMIAL PARAMETERS BASED ON UNDERLYING CONSTANTS SET 31

Code No.	Word	$\mu_1$	$\mu_2$	$\sigma$	$\tau$	$\delta_1$	$\delta_2$
B3A 60	upon	3.24	.23	3.47	.932	.25	.39
B3B 3	also	.32	.67	.99	.327	.09	.10
4	an	5.95	4.58	10.53	.565	.02	.02
13	by	7.32	11.43	18.75	.390	.35	.40
39	of	64.51	57.89	122.40	.527	.24	.25
40	on	3.38	7.75	11.12	.304	.34	.42
55	there	3.20	1.33	4.53	.706	.23	.24
57	this	7.77	6.00	13.77	.564	.21	.21
58	to	40.79	35.21	76.00	.537	.39	.45
B3G							
73	although	.06	.17	.23	.267	.11	.11
78	both	.52	1.04	1.56	.334	.12	.14
90	enough	.25	.10	.35	.727	.47	.52
116	while	.21	.07	.28	.744	.23	.25
117	whilst	.08	.42	.50	.153	.15	.13
123	always	.58	.20	.78	.742	.07	.07
160	though	.91	.51	1.42	.639	.08	.08
B3E							
80	commonly	.17	.05	.23	.763	.05	.05
81	consequently	.10	.42	.52	.189	.16	.14
82	considerable(ly)	.37	.17	.54	.684	.07	.08
119	according	.17	.54	.71	.238	.30	.30
124	apt	.27	.08	.35	.770	.06	.07
B3Z							
87	direction	.17	.08	.25	.693	.31	.32
94	innovation(s)	.06	.15	.20	.278	.06	.06
96	language	.08	.18	.26	.316	.05	.05
110	vigor(ous)	.18	.08	.26	.680	.02	.02
143	kind	.69	.17	.86	.799	.25	.22
146	matter(s)	.36	.09	.45	.790	.05	.05
151	particularly	.15	.37	.51	.282	.14	.16
153	probability	.27	.09	.36	.757	.02	.02
165	work(s)	.13	.27	.40	.326	.46	.42

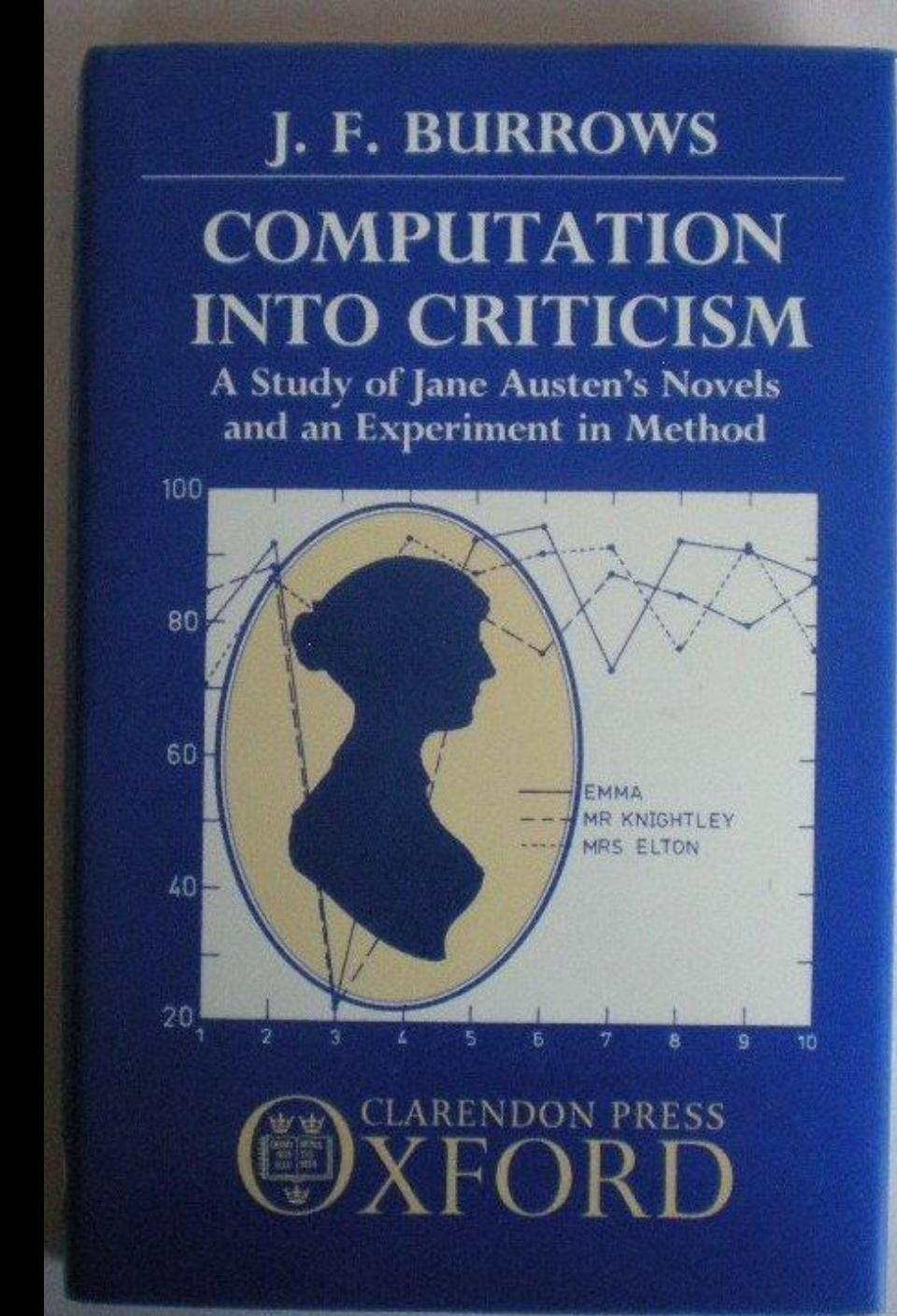


ДЖОН БАРРОУЗ,  
«ОТЕЦ»  
СОВРЕМЕННОЙ  
СТИЛЕМЕТРИИ

Барроуз изначально анализировал стили персонажей:

Most readers and critics behave as though common prepositions, conjunctions, personal pronouns, and articles — the parts of speech which make up at least a third of fictional works in English — do not really exist. But far from being a largely inert linguistic mass which has a simple but uninteresting function, these words and their frequency of use can tell us a great deal about the characters who speak them.

Preface to Computation into Criticism, 1987



Но в 90-е он доработал свой метод и догадался применить его к задаче определения авторства:

**‘Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship<sup>1</sup>**

---

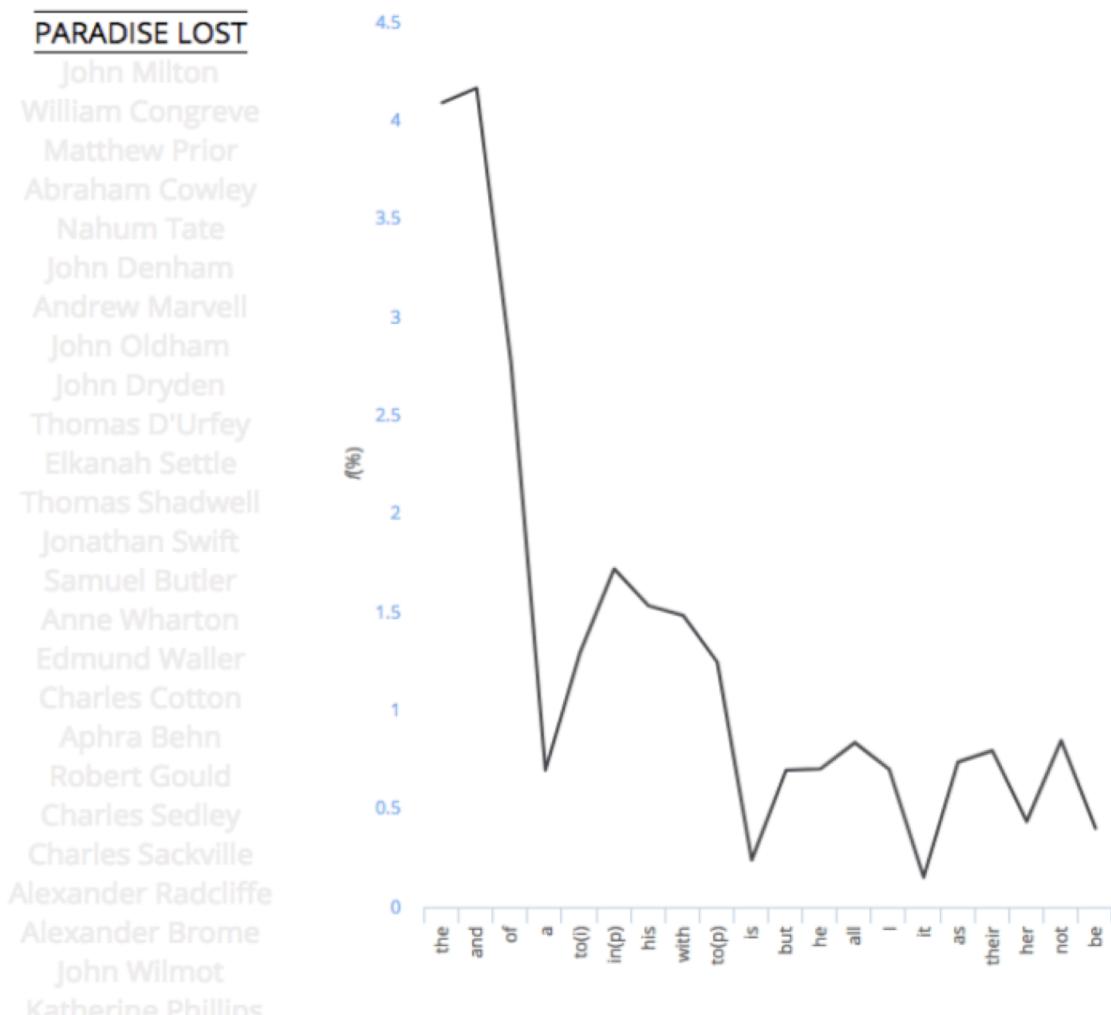
John Burrows  
University of Newcastle, Australia

---

**Abstract**

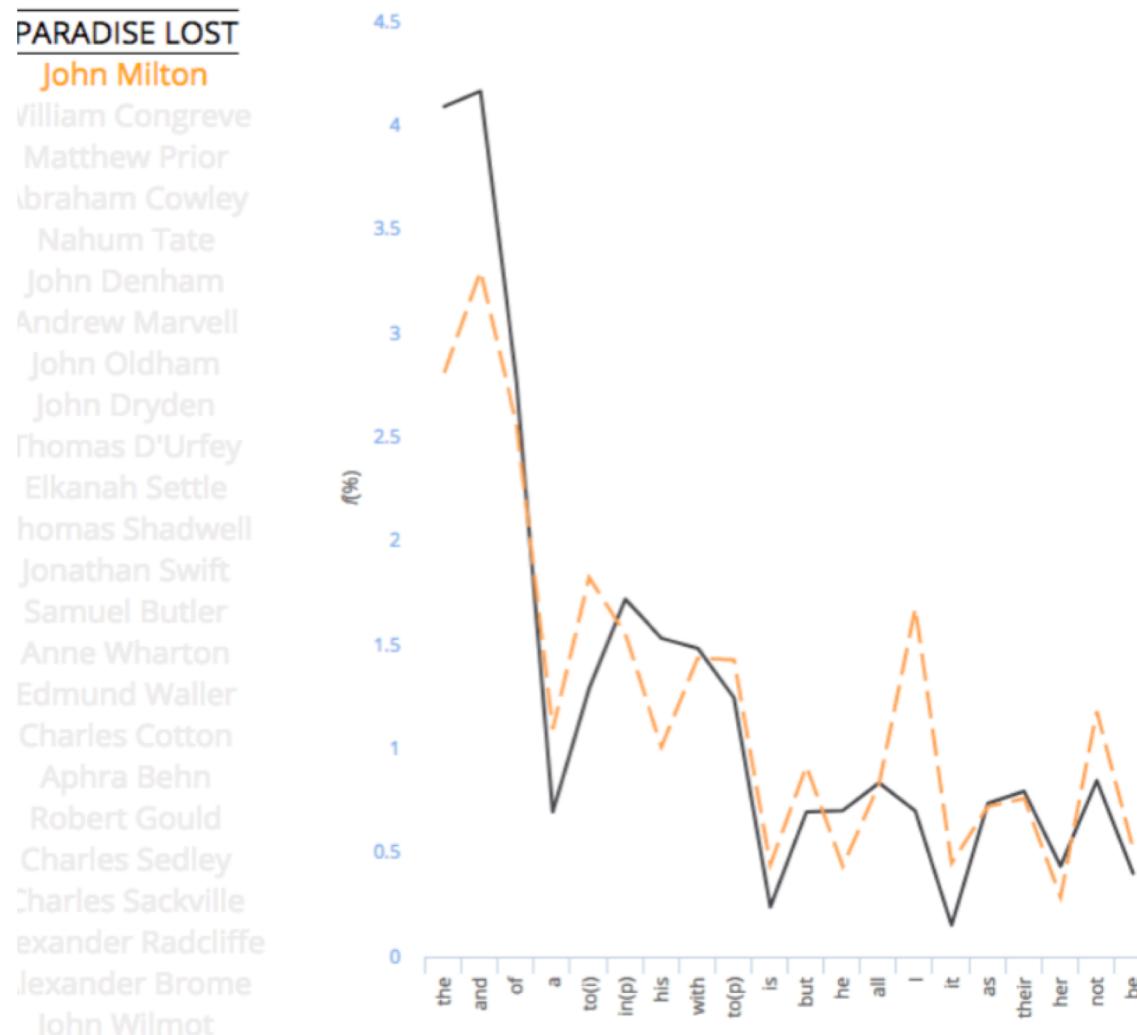
This paper is a companion to my ‘Questions of authorship: attribution and beyond’, in which I sketched a new way of using the relative frequencies of the very common words for comparing written texts and testing their likely authorship. The main emphasis of that paper was not on the new procedure but on the broader consequences of our increasing sophistication in making such comparisons and the increasing (although never absolute) reliability of our inferences about authorship. My present objects, accordingly, are to give a more complete account of the procedure itself; to report the outcome of an extensive set of trials; and to consider the strengths and limitations of the new procedure. The procedure offers a simple but comparatively accurate addition to our current methods of distinguishing the most likely author of texts exceeding about 1,500 words in length. It is of even greater value as a method of reducing the field of likely candidates for texts of as little as 100 words in length. Not unexpectedly, it works least well with texts of a genre uncharacteristic of their author and, in one case, with texts far separated in time across a long literary career. Its possible use for other classificatory tasks has not yet been investigated.

# Но Барроуз к этому шел по-другому:



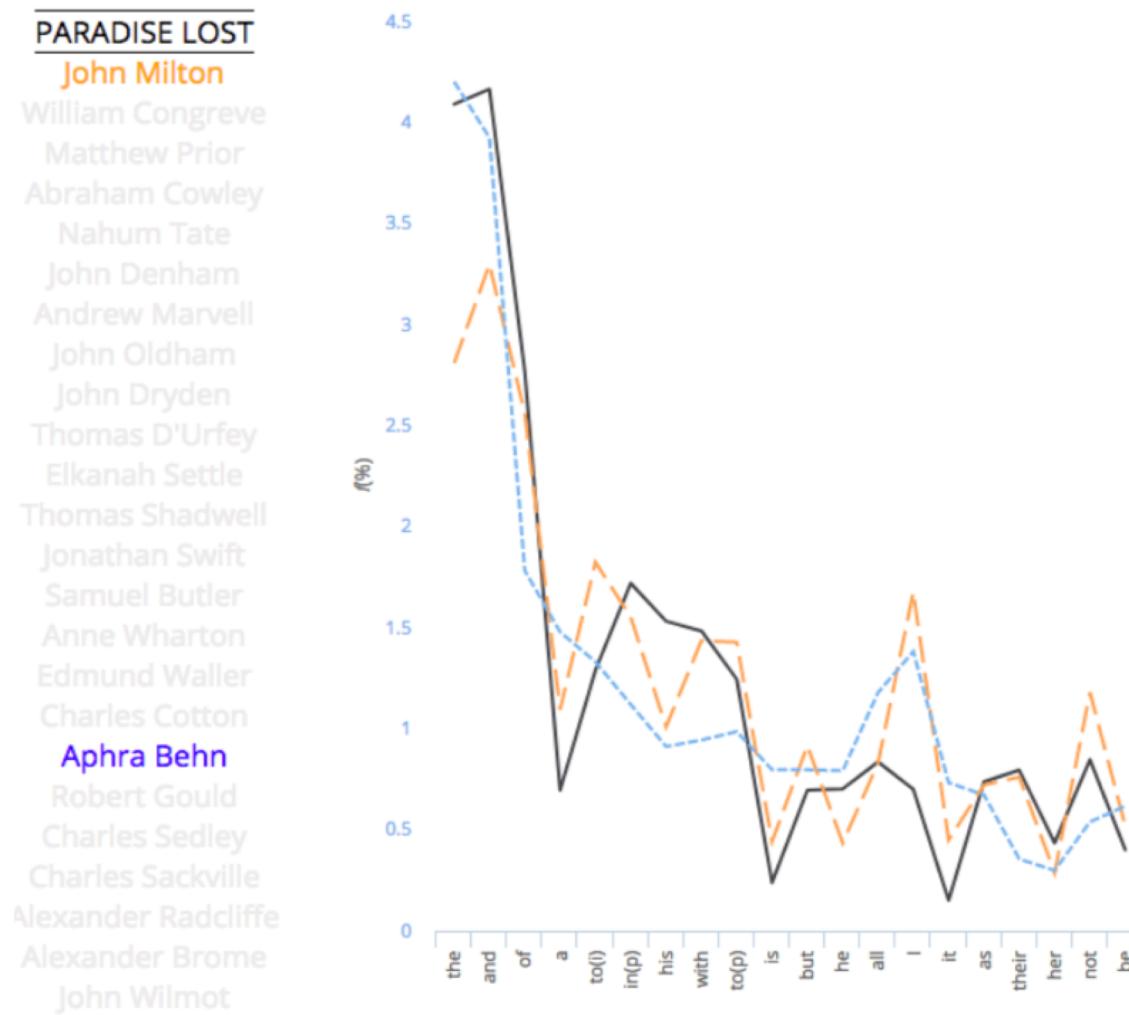
Картинка позаимствована из [выступления](#) Б.В. Орехова в ИГРАИ РАН

# Мильтон на фоне Мильтона же:



Картишка позаимствована из [выступления](#) Б.В. Орехова в ИГРАИ РАН

# Другой автор на фоне Мильтона:



Картина позаимствована из выступления Б.В. Орехова в ИРЛИ РАН

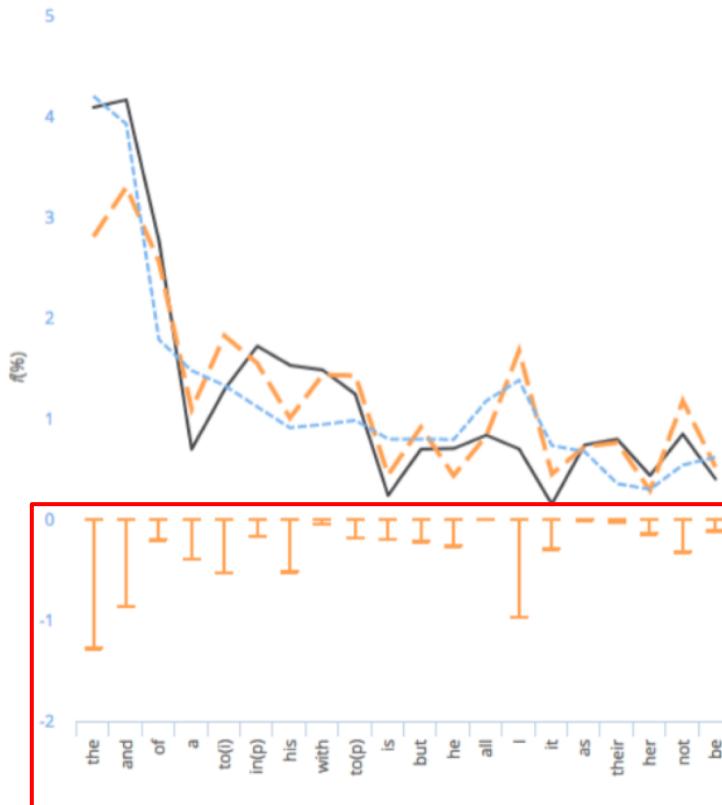
# Считаем расхождение для каждого слова в нашем списке:

## PARADISE LOST

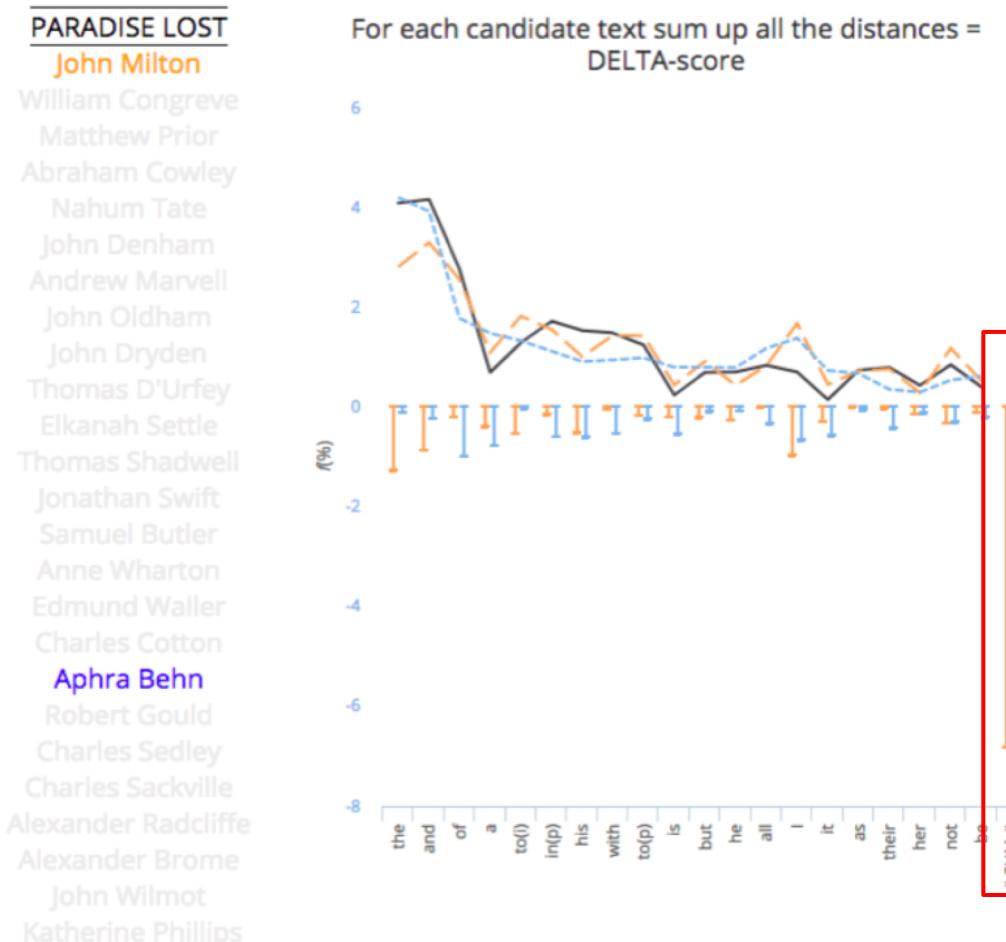
John Milton

William Congreve  
Matthew Prior  
Abraham Cowley  
Nahum Tate  
John Denham  
Andrew Marvell  
John Oldham  
John Dryden  
Thomas D'Urfey  
Elkanah Settle  
Thomas Shadwell  
Jonathan Swift  
Samuel Butler  
Anne Wharton  
Edmund Waller  
Charles Cotton  
Aphra Behn  
Robert Gould  
Charles Sedley  
Charles Sackville  
Alexander Radcliffe  
Alexander Brome  
John Wilmot  
Katherine Phillips

For each candidate text calculate the distances between its individual values and the ones of Paradise Lost



# Суммируем все расхождения – это и есть Delta:

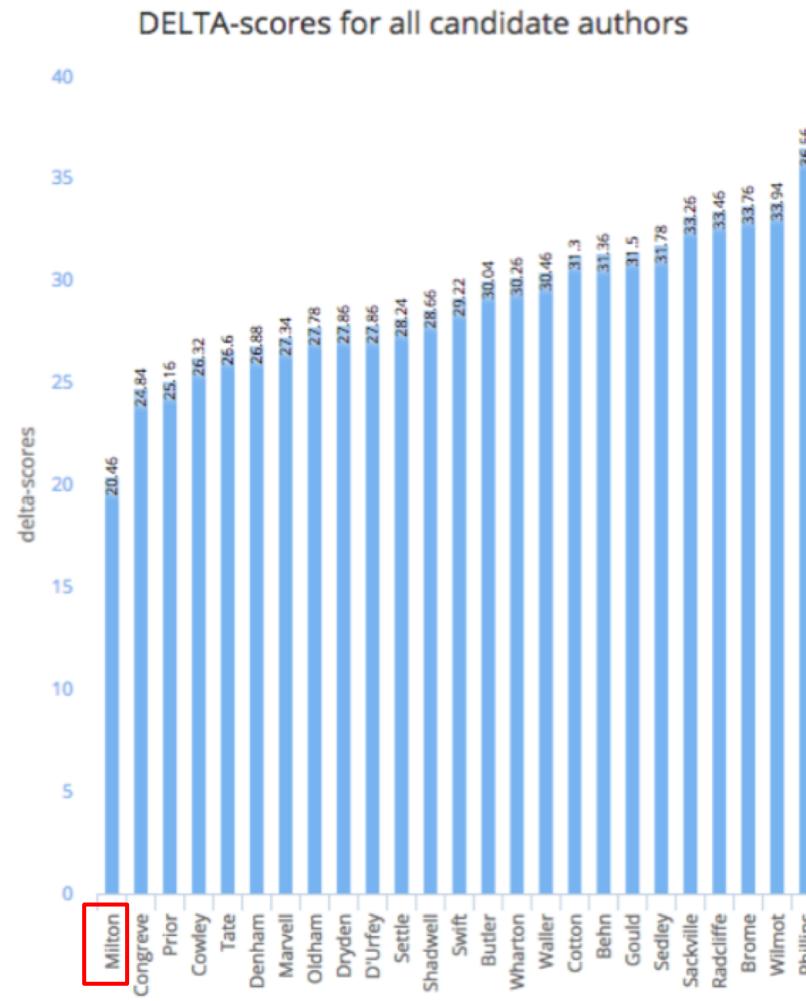


Картишка позаимствована из [выступления](#) Б.В. Орехова в ИРЛИ РАН

# Суммируем все расхождения – ЭТО И

## PARADISE LOST

John Milton  
William Congreve  
Matthew Prior  
Abraham Cowley  
Nahum Tate  
John Denham  
Andrew Marvell  
John Oldham  
John Dryden  
Thomas D'Urfey  
Elkanah Settle  
Thomas Shadwell  
Jonathan Swift  
Samuel Butler  
Anne Wharton  
Edmund Waller  
Charles Cotton  
Aphra Behn  
Robert Gould  
Charles Sedley  
Charles Sackville  
Alexander Radcliffe  
Alexander Brome  
John Wilmot  
Katherine Phillips

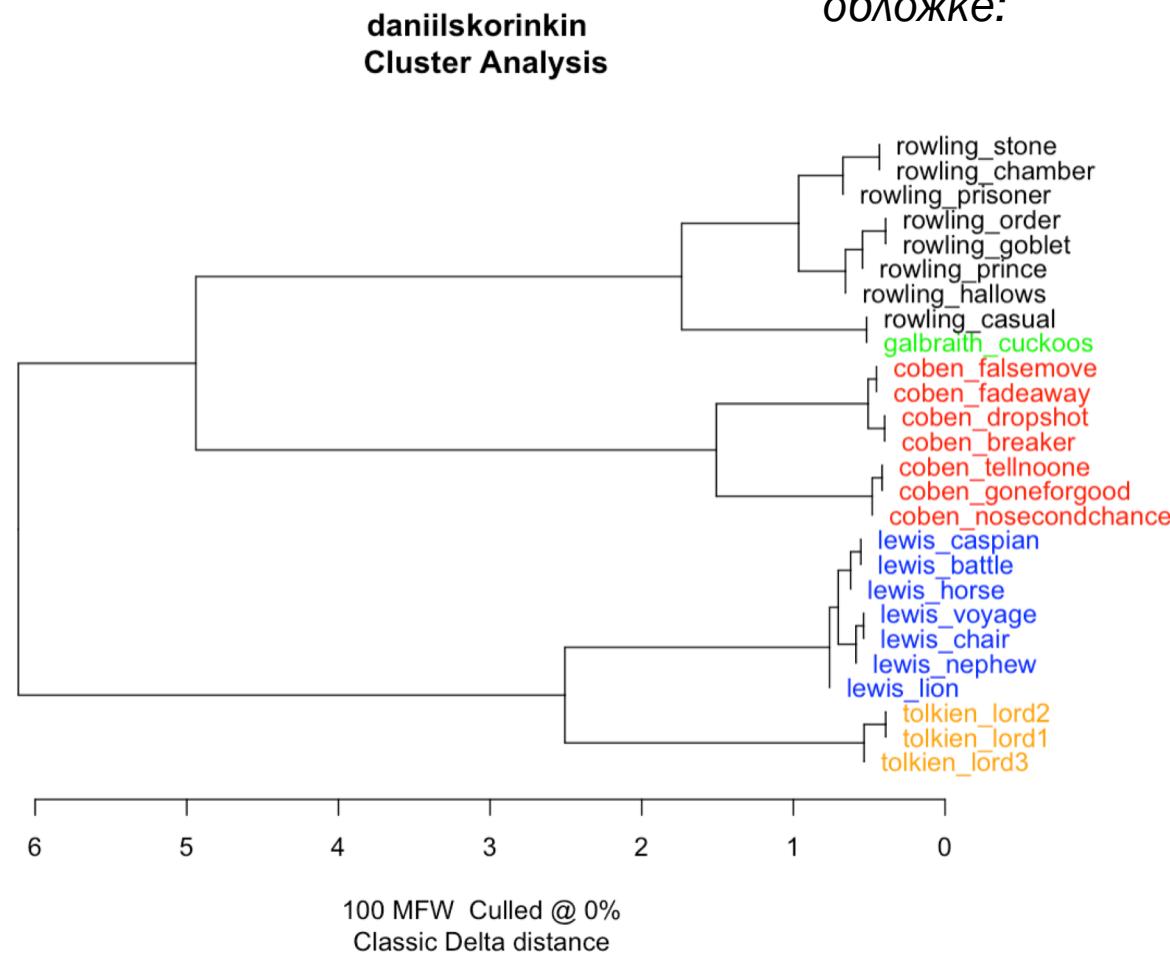


**Burrows's Delta <...> corresponds to the  
Manhattan distance of the word  
frequencies' z-scores.**

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, Thorsten Vitt,  
Understanding and explaining Delta measures for authorship attribution, *Digital Scholarship in the Humanities*, Volume 32, Issue suppl\_2, December 2017, Pages ii4–ii16, <https://doi.org/10.1093/llc/fqx023>

# Вернемся к дендрограмме стилометрической «близости»

Цвет подписи — по автору на  
обложке:



# Каждый текст – как набор чисел:

	A	B	C	D	E	F	G
1		galbraith_cuckoos	rowling_casual	rowling_chamber	rowling_goblet	rowling_hallows	rowling_order
2	the	4.52302745554585	4.7485038034729	4.41457832032396	4.48277661795407	4.6960087479497	4.2296641160
3	and	2.26740398874614	2.63908175930591	2.34403427431187	2.4258872651357	2.47293603061782	2.2525429458
4	to	2.49400579324491	2.62494930750986	2.35225071893891	2.48590814196242	2.24384909786769	2.5079899204
5	of	2.17939656008759	2.10819313531515	1.8768707083749	2.02244258872651	1.86987424822307	2.0301312190
6	a	2.14128310673153	1.76348420237671	2.00129115558425	1.79123173277662	1.44887916894478	1.7923542607
7	was	1.65550981941153	1.6461234070269	1.4813075884735	1.42327766179541	1.12575177692728	1.3870962785
8	I	1.12677227558106	0.560996890860605	0.881507130700159	0.848643006263048	0.525423728813559	0.8585323130
9	in	1.38039998336867	1.44335344647487	1.16790891484242	1.11691022964509	0.99507927829415	1.1677575981
10	he	1.32704114867019	1.18282476988682	1.16204002582311	1.36534446764092	1.37452159650082	1.3467625457
11	said	0.828101395645365	0.879898737910609	1.41440225365338	1.37212943632568	0.972662657189721	1.5150118312
12	you	1.07826424403698	0.597864156415519	0.962497799166618	0.894572025052192	0.963367960634226	1.0832488245
13	that	1.18567488531315	1.15886104727612	0.72656846059041	0.803757828810021	0.904319300164024	0.8366368581
14	it	0.856513242692612	0.907549187076795	0.991842244263161	0.934237995824635	1.0098414434117	0.9330536861
15	his	1.0907377378626	1.03228343553759	1.30524091789424	1.29488517745303	1.22799343903773	1.2188469930
16	had	1.31179576732776	2.01295269929829	0.788778684195082	0.970772442588727	0.950792782941498	0.8946406072
17	on	0.823943564370158	0.758236761579394	0.679617348435941	0.627870563674321	0.543466375068343	0.7344580682
18	at	0.555070475240115	0.666683052118024	0.758260461294677	0.85490605427975	0.631492618917441	0.8116683568
19	her	1.34783030504622	1.71187003059983	0.326310229473561	0.447286012526096	0.514488791689448	0.6722288804
20	with	0.760883123362854	0.740417583227852	0.550501790011151	0.592901878914405	0.537452159650082	0.6003964229
21	Harry	0.000692971879201142		0.1.76184048359646	1.52974947807933	1.4505194095134	1.4508619894
22	as	0.495474893628817	0.504467083676404	0.626797347262163	0.67223382045929	0.644067796610169	0.6645462647
23	for	0.567543969065735	0.592948521008197	0.470684899348553	0.484342379958246	0.441771459814106	0.4836206631
24	not	0.406081521211869	0.62735796885945	0.242972005399378	0.334029227557411	0.549480590486605	0.5677453059
25	him	0.480922484165593	0.505081538102319	0.56341334585363	0.708768267223382	0.644067796610169	0.6211394855
26	He	0.476071681011185	0.465142000417829	0.46833734374083	0.544885177453027	0.561509021323127	0.4444393227
27	The	0.388757224231841	0.354540203753088	0.359176007981689	0.367954070981211	0.468015308911974	0.3103776773
28	they	0.23075963577398	0.3348776621238	0.420212453782499	0.483820459290188	0.454346637506834	0.4294582219
29	she	1.14201765692348	1.2670050262372	0.201889782264217	0.319415448851775	0.353745215965008	0.4229279985
30	were	0.325003811345336	0.344094478512529	0.388520453078232	0.462421711899791	0.395844723892838	0.3841307888

# Каждый текст – как столбик частотностей:

	A	B	C	D	E	F	G
1		galbraith_cuckoos	rowling_casual	rowling_chamber	rowling_goblet	rowling_hallows	rowling_order
2	the	4.52302745554585	4.7485038034729	4.41457832032396	4.48277661795407	4.6960087479497	4.2296641160
3	and	2.26740398874614	2.63908175930591	2.34403427431187	2.4258872651357	2.47293603061782	2.2525429458
4	to	2.49400579324491	2.62494930750986	2.35225071893891	2.48590814196242	2.24384909786769	2.5079899204
5	of	2.17939656008759	2.10819313531515	1.8768707083749	2.02244258872651	1.86987424822307	2.0301312190
6	a	2.14128310673153	1.76348420237671	2.00129115558425	1.79123173277662	1.44887916894478	1.7923542607
7	was	1.65550981941153	1.6461234070269	1.4813075884735	1.42327766179541	1.12575177692728	1.3870962785
8	I	1.12677227558106	0.560996890860605	0.881507130700159	0.848643006263048	0.525423728813559	0.8585323130
9	in	1.38039998336867	1.44335344647487	1.16790891484242	1.11691022964509	0.99507927829415	1.1677575981
10	he	1.32704114867019	1.18282476988682	1.16204002582311	1.36534446764092	1.37452159650082	1.3467625457
11	said	0.828101395645365	0.879898737910609	1.41440225365338	1.37212943632568	0.972662657189721	1.5150118312
12	you	1.07826424403698	0.597864156415519	0.962497799166618	0.894572025052192	0.963367960634226	1.0832488245
13	that	1.18567488531315	1.15886104727612	0.72656846059041	0.803757828810021	0.904319300164024	0.8366368581
14	it	0.856513242692612	0.907549187076795	0.991842244263161	0.934237995824635	1.0098414434117	0.9330536861
15	his	1.0907377378626	1.03228343553759	1.30524091789424	1.29488517745303	1.22799343903773	1.2188469930
16	had	1.31179576732776	2.01295269929829	0.788778684195082	0.970772442588727	0.950792782941498	0.8946406072
17	on	0.823943564370158	0.758236761579394	0.679617348435941	0.627870563674321	0.543466375068343	0.7344580682
18	at	0.555070475240115	0.666683052118024	0.758260461294677	0.85490605427975	0.631492618917441	0.8116683568
19	her	1.34783030504622	1.71187003059983	0.326310229473561	0.447286012526096	0.514488791689448	0.6722288804
20	with	0.760883123362854	0.740417583227852	0.550501790011151	0.592901878914405	0.537452159650082	0.6003964229
21	Harry	0.000692971879201142		0.1.76184048359646	1.52974947807933	1.4505194095134	1.4508619894
22	as	0.495474893628817	0.504467083676404	0.626797347262163	0.67223382045929	0.644067796610169	0.6645462647
23	for	0.567543969065735	0.592948521008197	0.470684899348553	0.484342379958246	0.441771459814106	0.4836206631
24	not	0.406081521211869	0.62735796885945	0.242972005399378	0.334029227557411	0.549480590486605	0.5677453059
25	him	0.480922484165593	0.505081538102319	0.56341334585363	0.708768267223382	0.644067796610169	0.6211394855
26	He	0.476071681011185	0.465142000417829	0.46833734374083	0.544885177453027	0.561509021323127	0.4444393227
27	The	0.388757224231841	0.354540203753088	0.359176007981689	0.367954070981211	0.468015308911974	0.3103776773
28	they	0.23075963577398	0.3348776621238	0.420212453782499	0.483820459290188	0.454346637506834	0.4294582219
29	she	1.14201765692348	1.2670050262372	0.201889782264217	0.319415448851775	0.353745215965008	0.4229279985
30	were	0.325003811345336	0.344094478512529	0.388520453078232	0.462421711899791	0.395844723892838	0.3841307888

# Каждый текст – как $\overrightarrow{\text{вектор}}$ частотностей:

	A	B	C	D	E	F	G
1		galbraith_cuckoos	rowling_casual	rowling_chamber	rowling_goblet	rowling_hallows	rowling_order
2	the	4.52302745554585	4.7485038034729	4.41457832032396	4.48277661795407	4.6960087479497	4.2296641160
3	and	2.26740398874614	2.63908175930591	2.34403427431187	2.4258872651357	2.47293603061782	2.2525429458
4	to	2.49400579324491	2.62494930750986	2.35225071893891	2.48590814196242	2.24384909786769	2.5079899204
5	of	2.17939656008759	2.10819313531515	1.8768707083749	2.02244258872651	1.86987424822307	2.0301312190
6	a	2.14128310673153	1.76348420237671	2.00129115558425	1.79123173277662	1.44887916894478	1.7923542607
7	was	1.65550981941153	1.6461234070269	1.4813075884735	1.42327766179541	1.12575177692728	1.3870962785
8	I	1.12677227558106	0.560996890860605	0.881507130700159	0.848643006263048	0.525423728813559	0.8585323130
9	in	1.38039998336867	1.44335344647487	1.16790891484242	1.11691022964509	0.99507927829415	1.1677575981
10	he	1.32704114867019	1.18282476988682	1.16204002582311	1.36534446764092	1.37452159650082	1.3467625457
11	said	0.828101395645365	0.879898737910609	1.41440225365338	1.37212943632568	0.972662657189721	1.5150118312
12	you	1.07826424403698	0.597864156415519	0.962497799166618	0.894572025052192	0.963367960634226	1.0832488245
13	that	1.18567488531315	1.15886104727612	0.72656846059041	0.803757828810021	0.904319300164024	0.8366368581
14	it	0.856513242692612	0.907549187076795	0.991842244263161	0.934237995824635	1.0098414434117	0.9330536861
15	his	1.0907377378626	1.03228343553759	1.30524091789424	1.29488517745303	1.22799343903773	1.2188469930
16	had	1.31179576732776	2.01295269929829	0.788778684195082	0.970772442588727	0.950792782941498	0.8946406072
17	on	0.823943564370158	0.758236761579394	0.679617348435941	0.627870563674321	0.543466375068343	0.7344580682
18	at	0.555070475240115	0.666683052118024	0.758260461294677	0.85490605427975	0.631492618917441	0.8116683568
19	her	1.34783030504622	1.71187003059983	0.326310229473561	0.447286012526096	0.514488791689448	0.6722288804
20	with	0.760883123362854	0.740417583227852	0.550501790011151	0.592901878914405	0.537452159650082	0.6003964229
21	Harry	0.000692971879201142		0.1.76184048359646	1.52974947807933	1.4505194095134	1.4508619894
22	as	0.495474893628817	0.504467083676404	0.626797347262163	0.67223382045929	0.644067796610169	0.6645462647
23	for	0.567543969065735	0.592948521008197	0.470684899348553	0.484342379958246	0.441771459814106	0.4836206631
24	not	0.406081521211869	0.62735796885945	0.242972005399378	0.334029227557411	0.549480590486605	0.5677453059
25	him	0.480922484165593	0.505081538102319	0.56341334585363	0.708768267223382	0.644067796610169	0.6211394855
26	He	0.476071681011185	0.465142000417829	0.46833734374083	0.544885177453027	0.561509021323127	0.4444393227
27	The	0.388757224231841	0.354540203753088	0.359176007981689	0.367954070981211	0.468015308911974	0.3103776773
28	they	0.23075963577398	0.3348776621238	0.420212453782499	0.483820459290188	0.454346637506834	0.4294582219
29	she	1.14201765692348	1.2670050262372	0.201889782264217	0.319415448851775	0.353745215965008	0.4229279985
30	were	0.325003811345336	0.344094478512529	0.388520453078232	0.462421711899791	0.395844723892838	0.3841307888

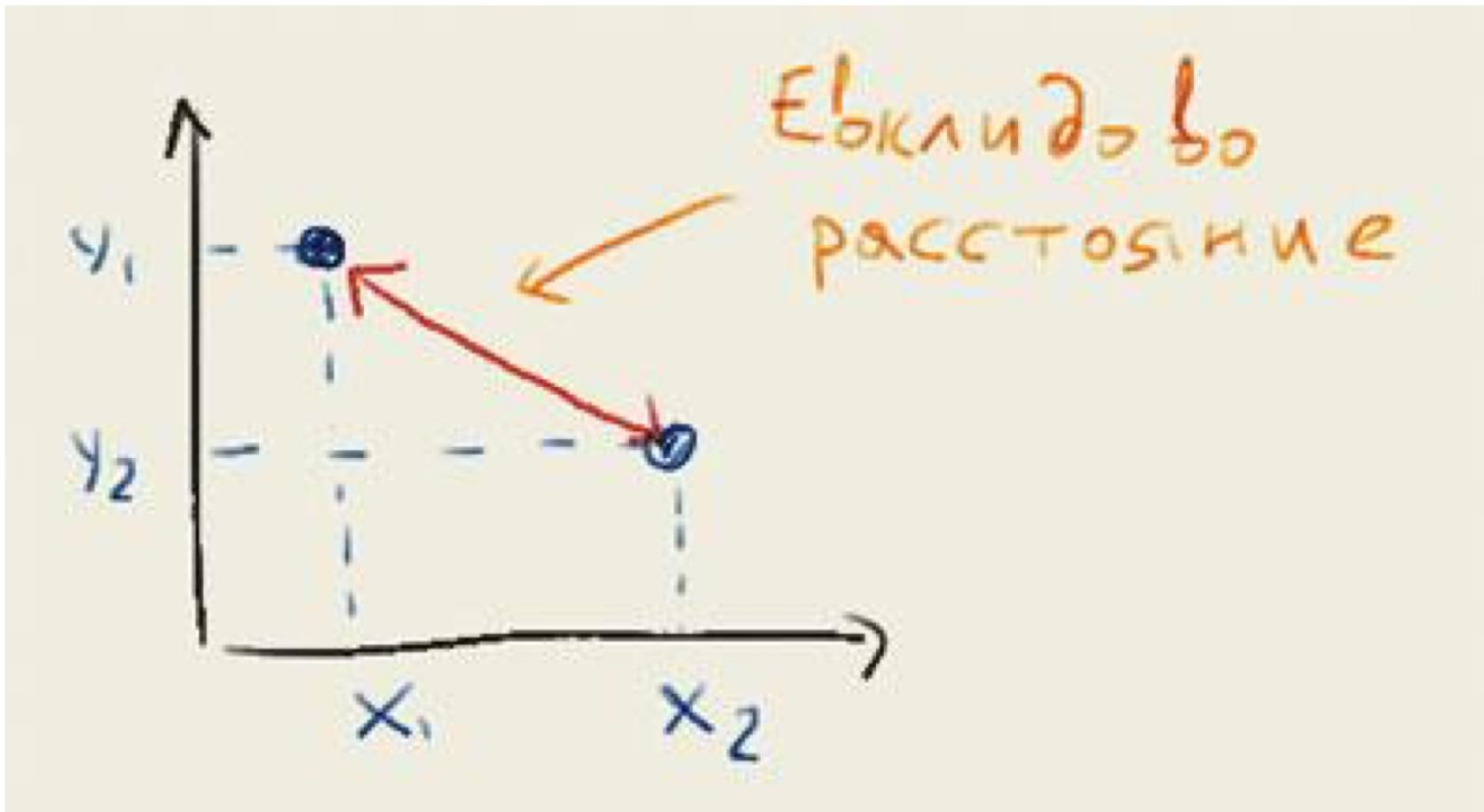
Давайте на время упростим модель до 2 измерений (вектора будут двумерные):

	A	B	C	D	E	F	G
1		galbraith_cuckoos	rowling_casual	rowling_chamber	rowling_goblet	rowling_hallows	rowling_order
2	the	4.52302745554585	4.7485038034729	4.41457832032396	4.48277661795407	4.6960087479497	4.2296641160
3	and	2.26740398874614	2.63908175930591	2.34403427431187	2.4258872651357	2.47293603061782	2.2525429458

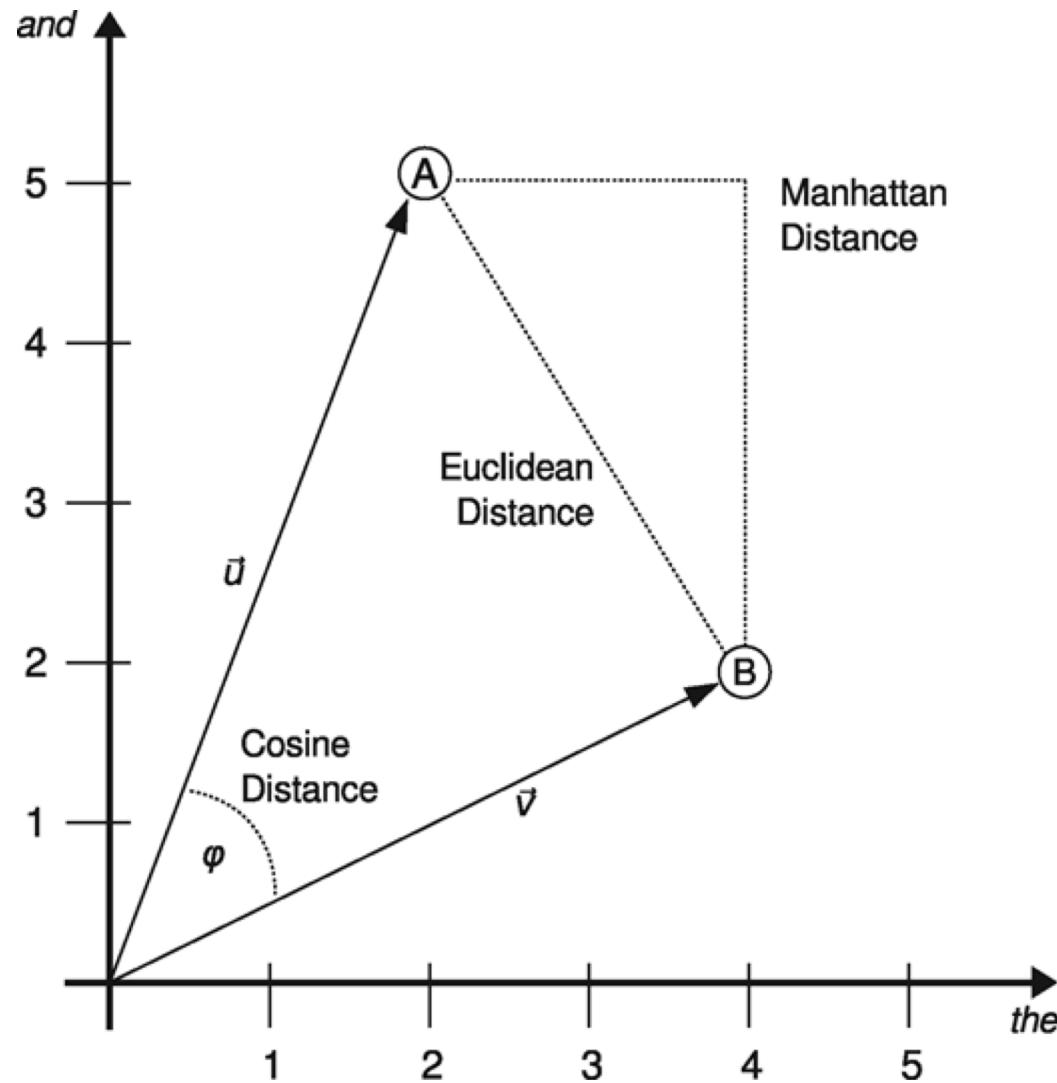




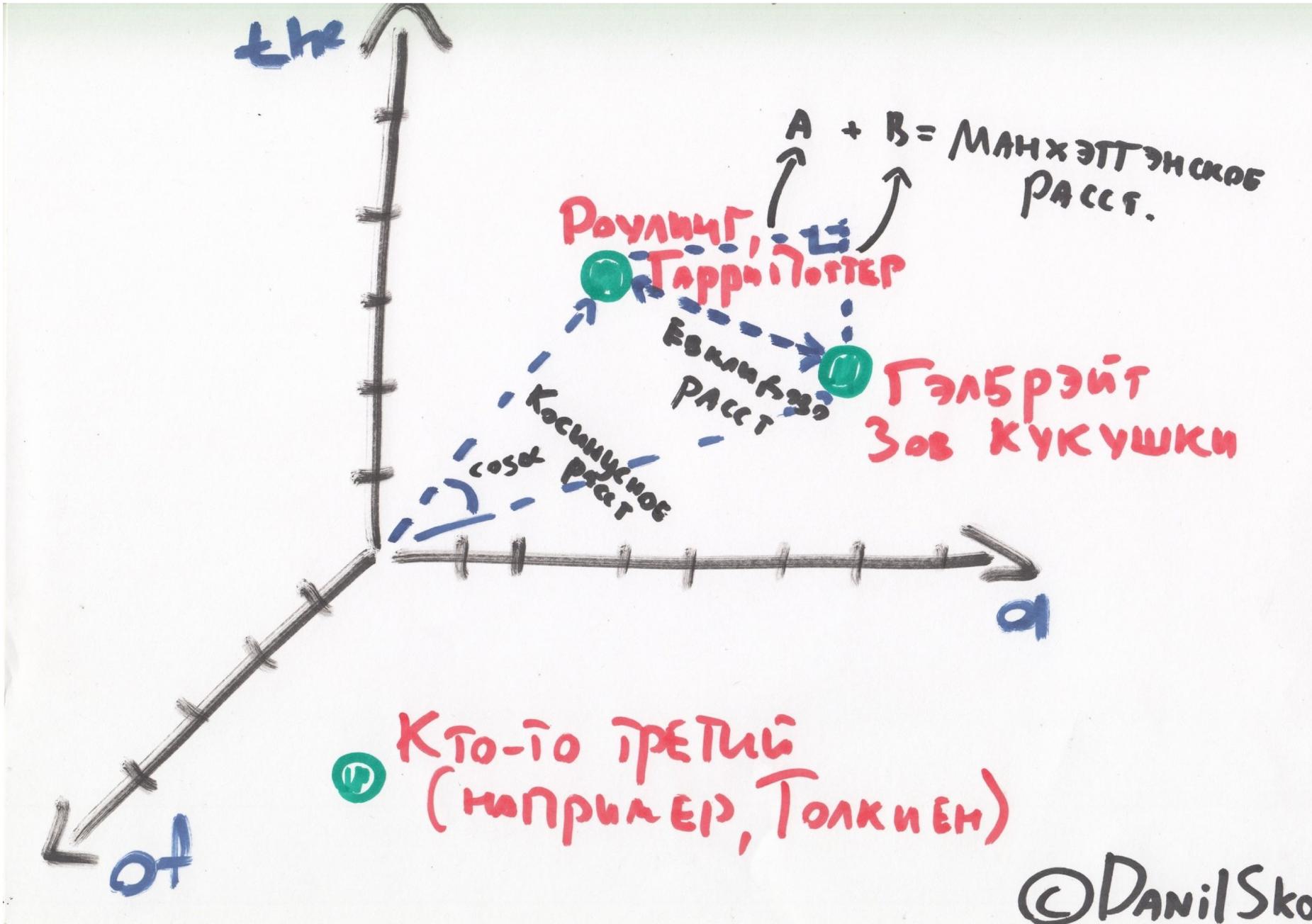
Между любыми такими векторами геометрия умеет измерить расстояние – например, так:



# Или по-другому:



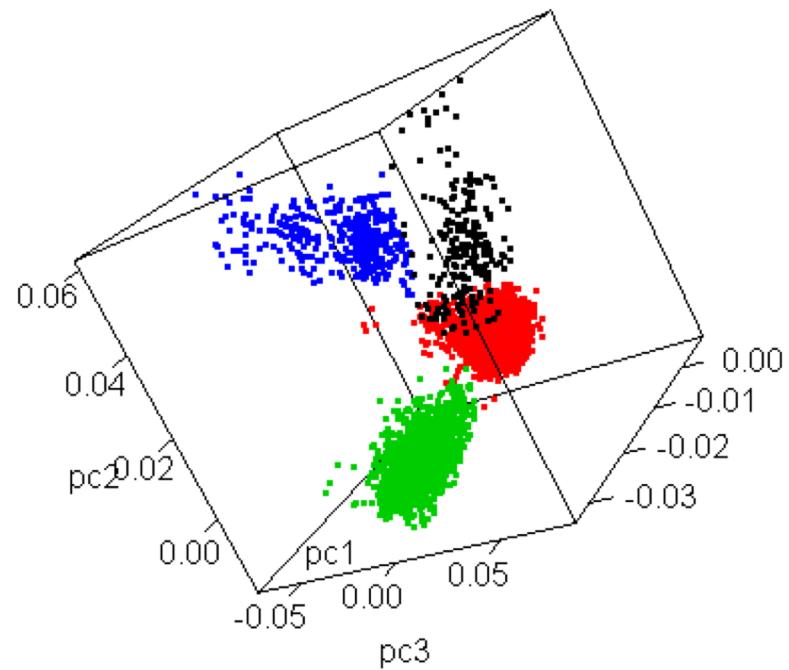




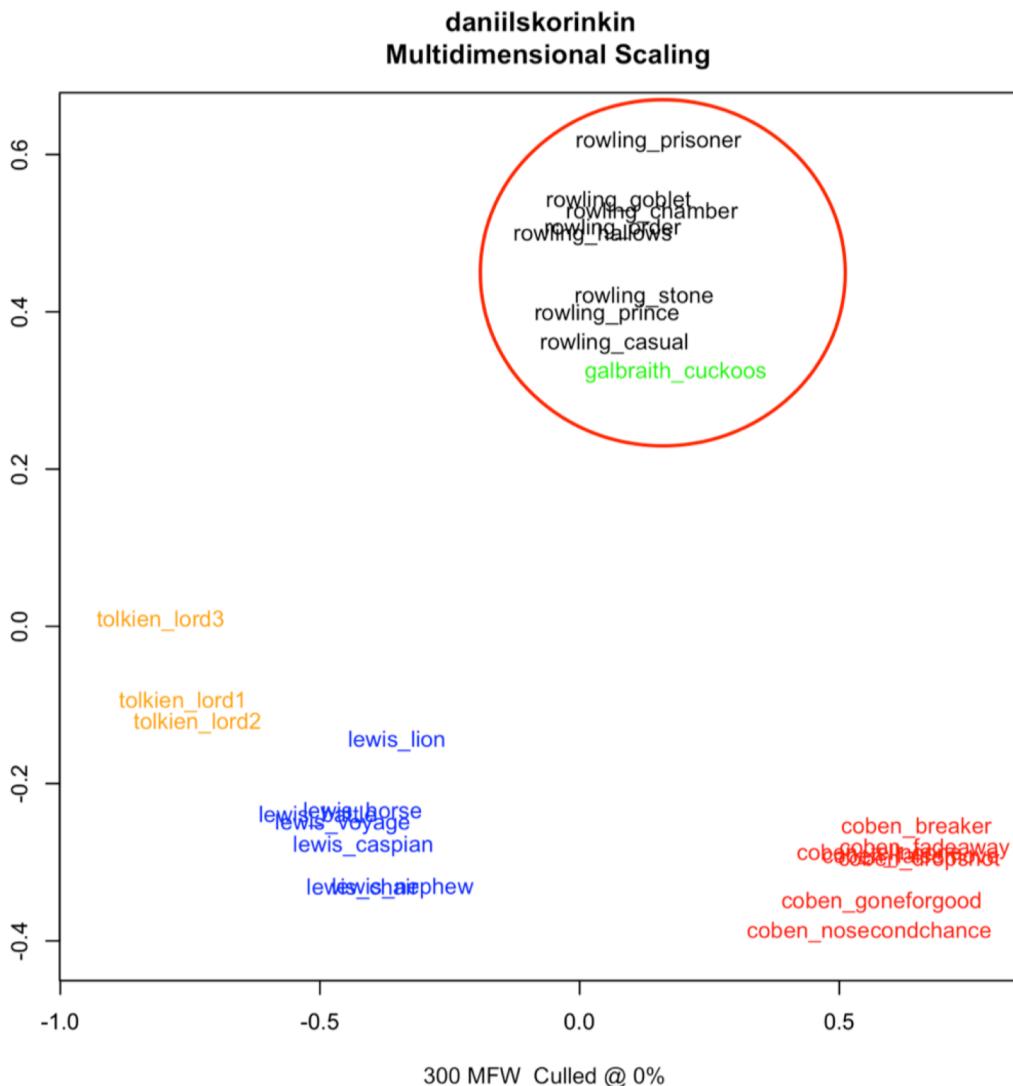
добавим третье измерение

@DanilSko

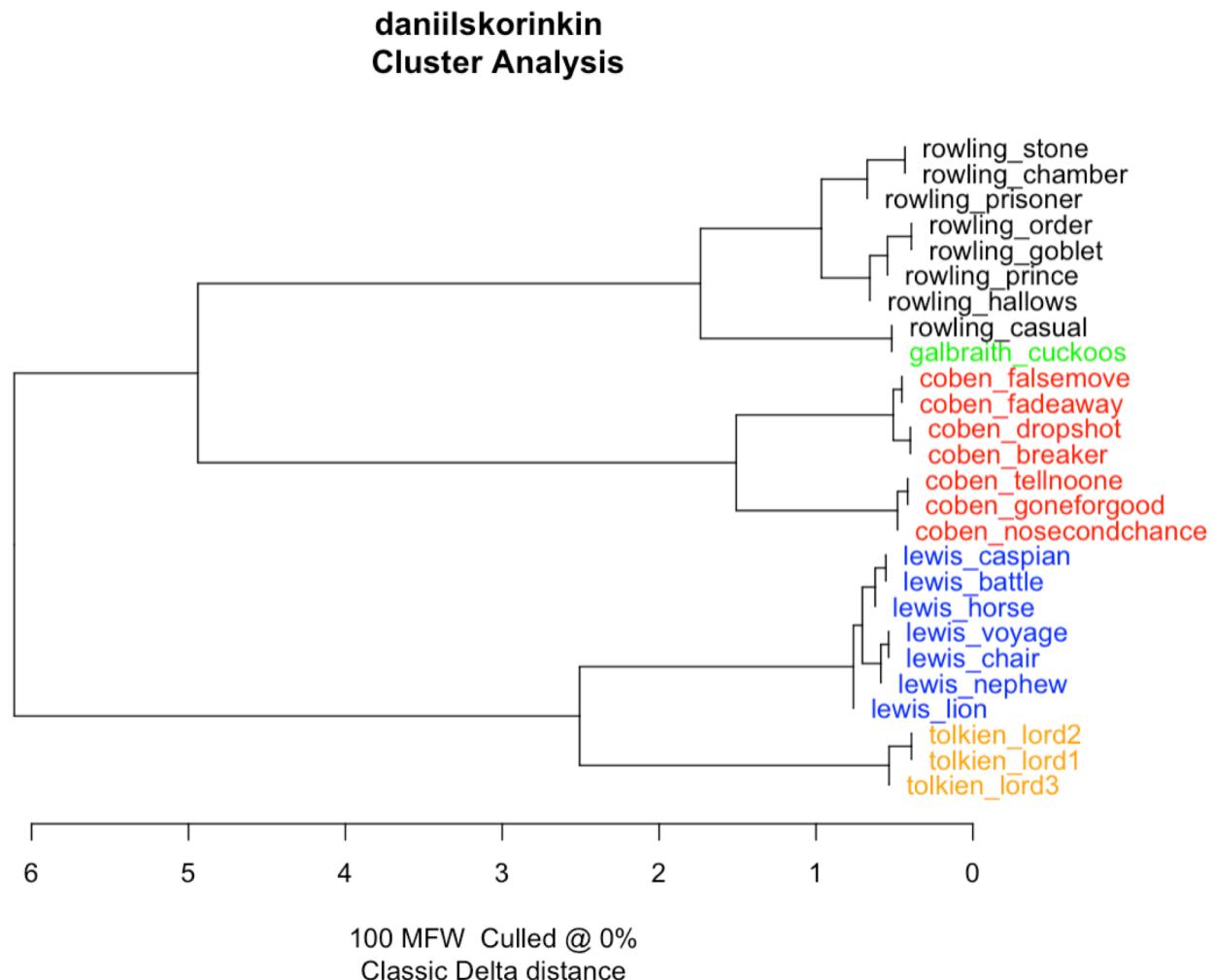
В стилометрии мы делаем это в  
100/300/500-мерном  
пространстве



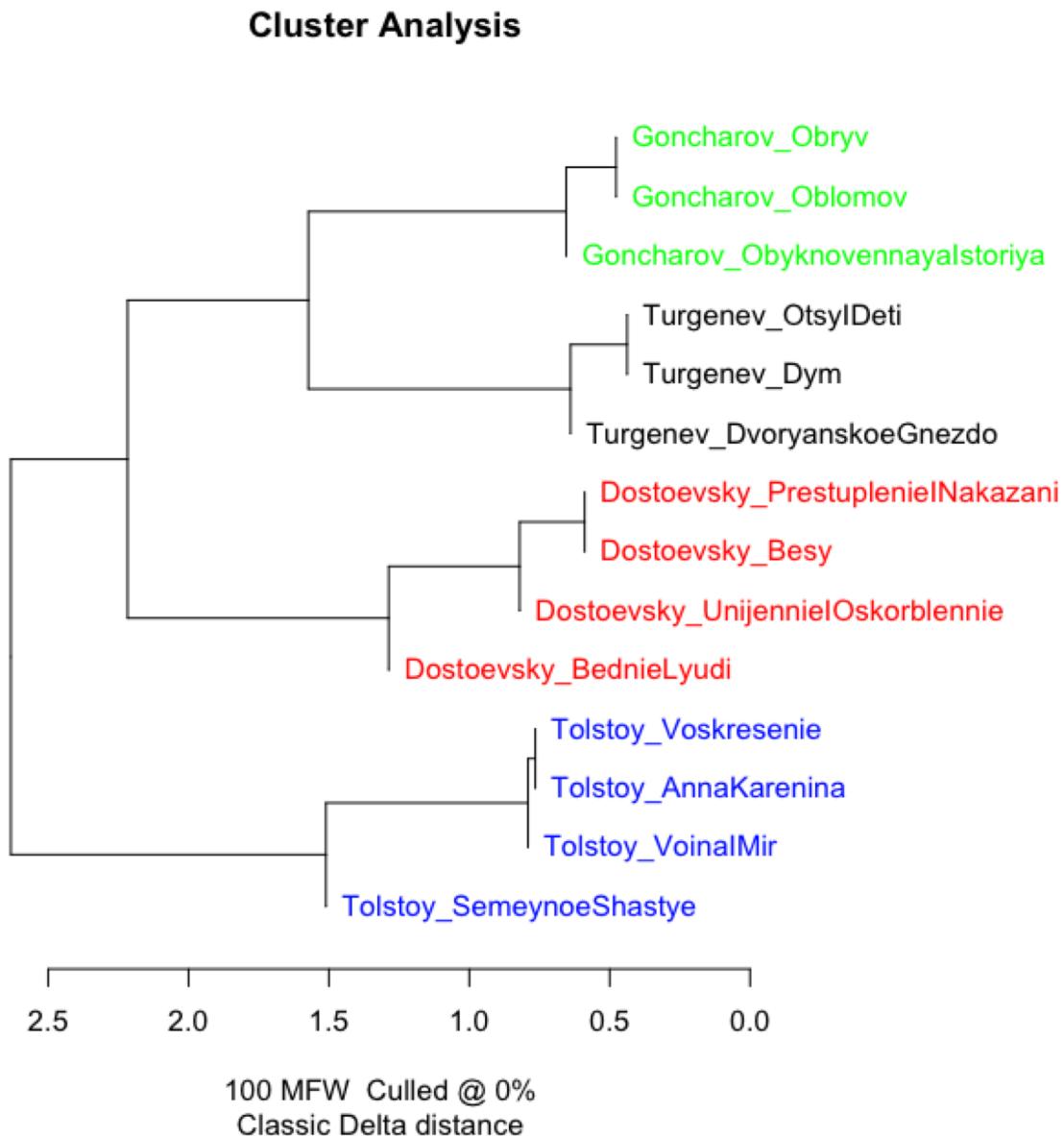
# Роулинг в 100-мерном пространстве (сжато до 2 измерений):



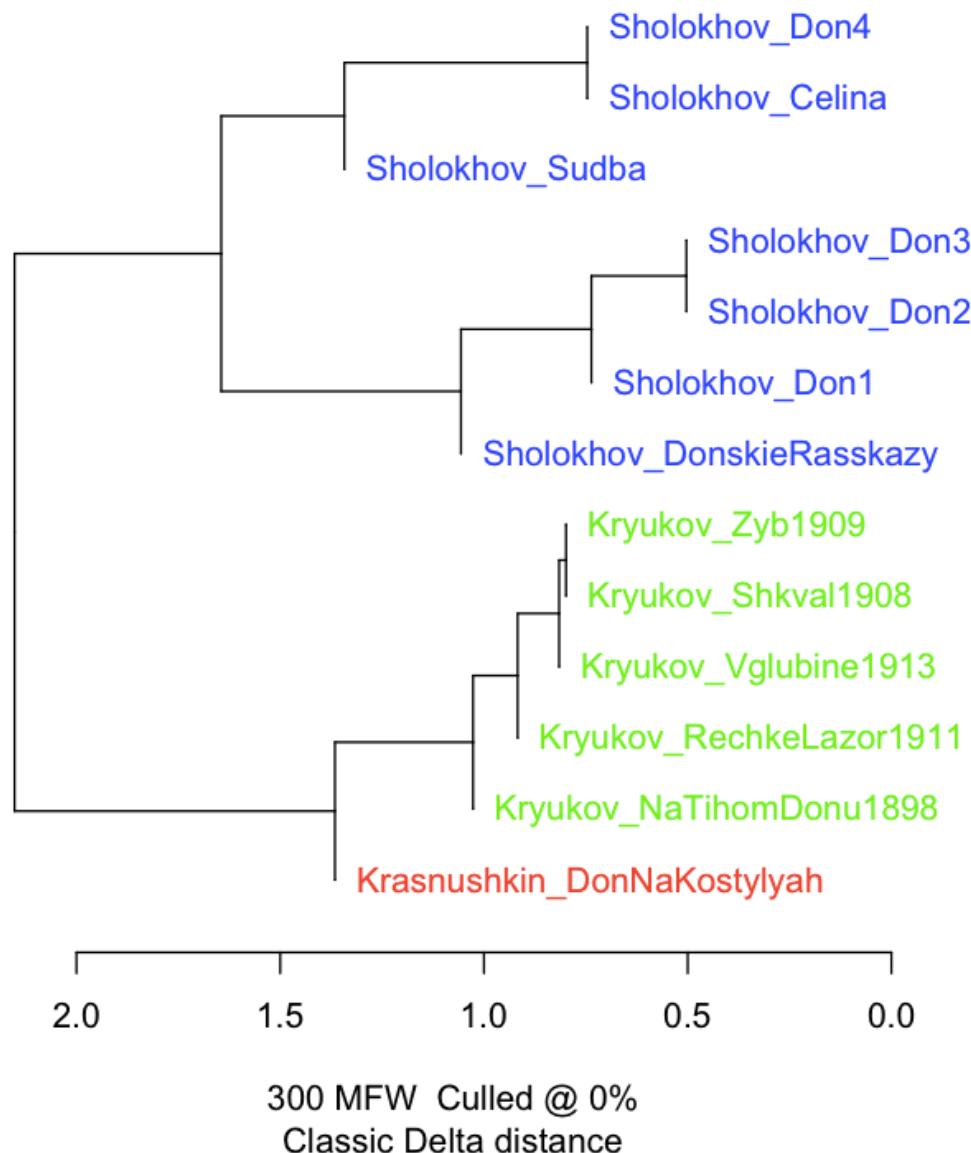
# То же самое на дендрограмме (вы уже видели)



# И для русского тоже работает:



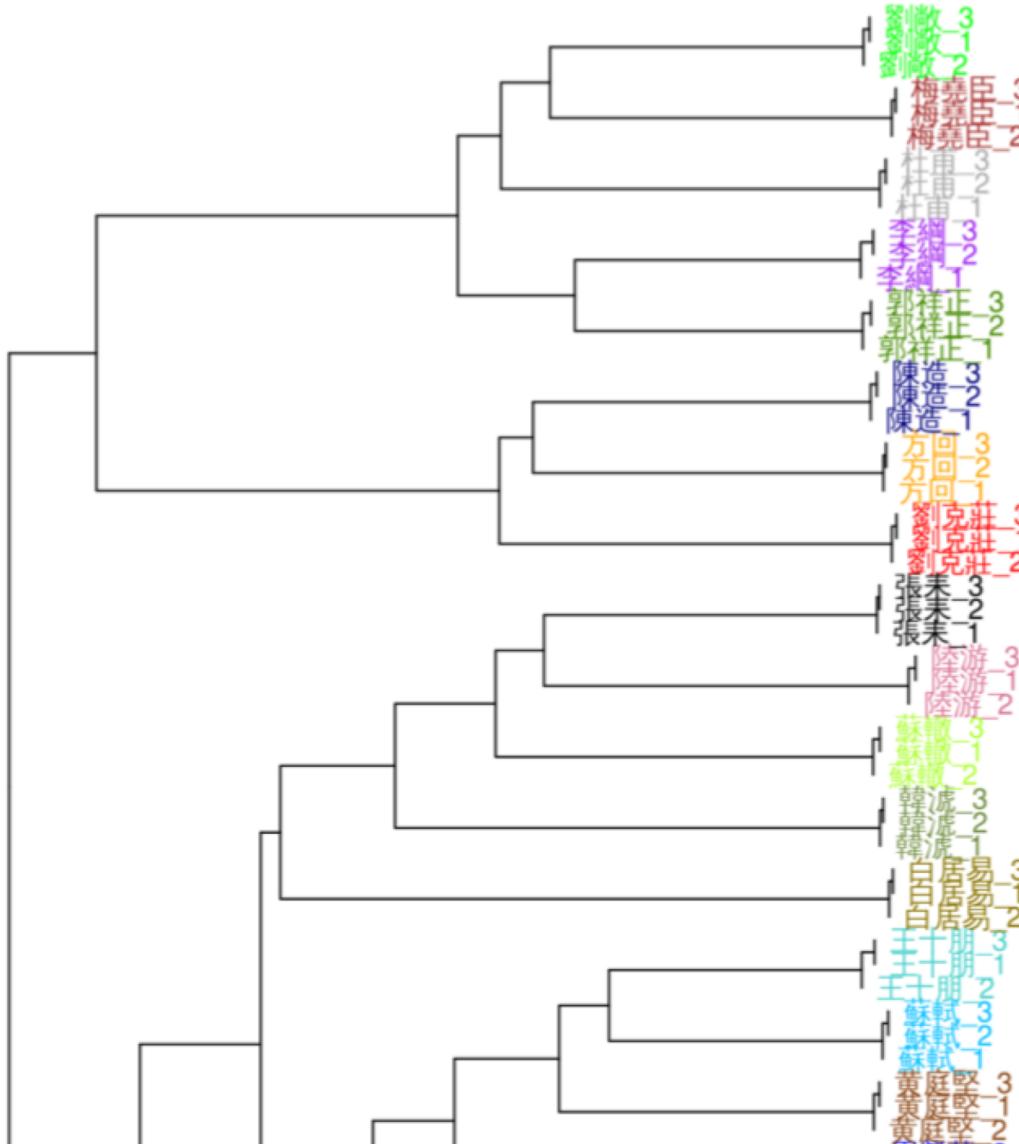
# ШОЛОХОВ И КОМПАНИЯ



# Delta сегодня – главный мировой инструмент для стилометрии

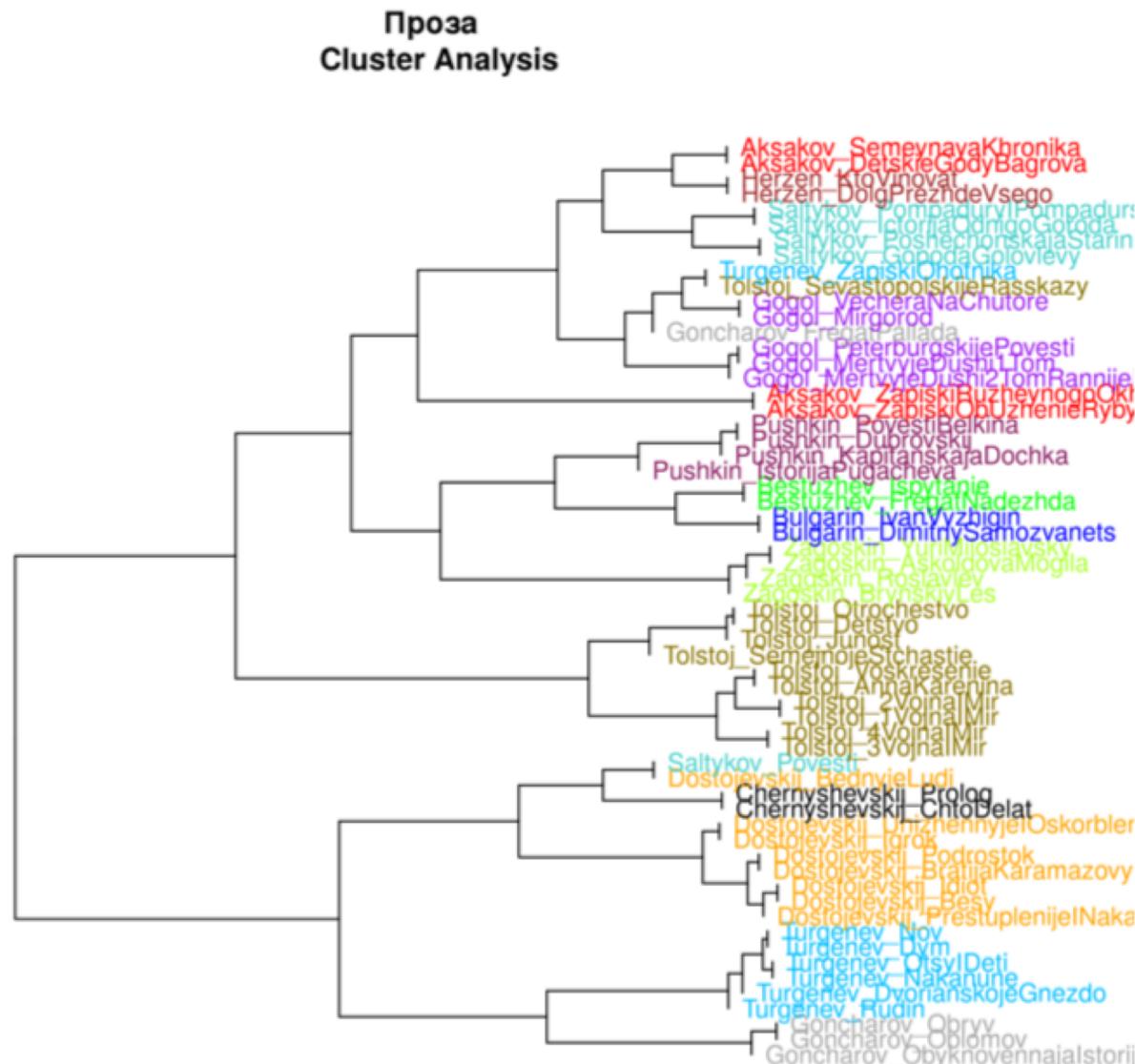
- Американский английский: Hoover D.L. Testing Burrows's Delta // Literary and Linguistic Computing. 2004. Vol. 19. No 4. P. 453–475.
- Древнеанглийский: García A. M., Martín J. C. Function Words in Authorship Attribution Studies // Literary and Linguistic Computing. 2007. Vol. 22. No 1. P. 49–66.
- Немецкий: Jannidis F., Lauer G. Burrows's Delta and Its Use in German Literary History // Distant Readings. Topologies of German Culture in the Long Nineteenth Century Studies in German Literature Linguistics and Culture. / под ред. M. Erlin, L. Tatlock. Rochester: Camden House, 2014. P. 29 – 54.
- Итальянский: Rybicki, J. (2018). Partners in life, partners in crime? In Tuzzi, A. and Cortelazzo, M. A. (eds), Drawing Elena Ferrante's Profile. Padova: Padova University Press, pp. 111–122,
- Польский: Rybicki, J. Heydel M., The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish. Literary and Linguistic Computing, 2013
- Хорошая статья в которой сравнивается качество атрибуции на разных языках (**латынь, польский, английский, немецкий**): Eder, M. Style-Markers in Authorship Attribution A Cross-Language Study of the Authorial Fingerprint. 99–114 (2011).

# Для китайского тоже работает:



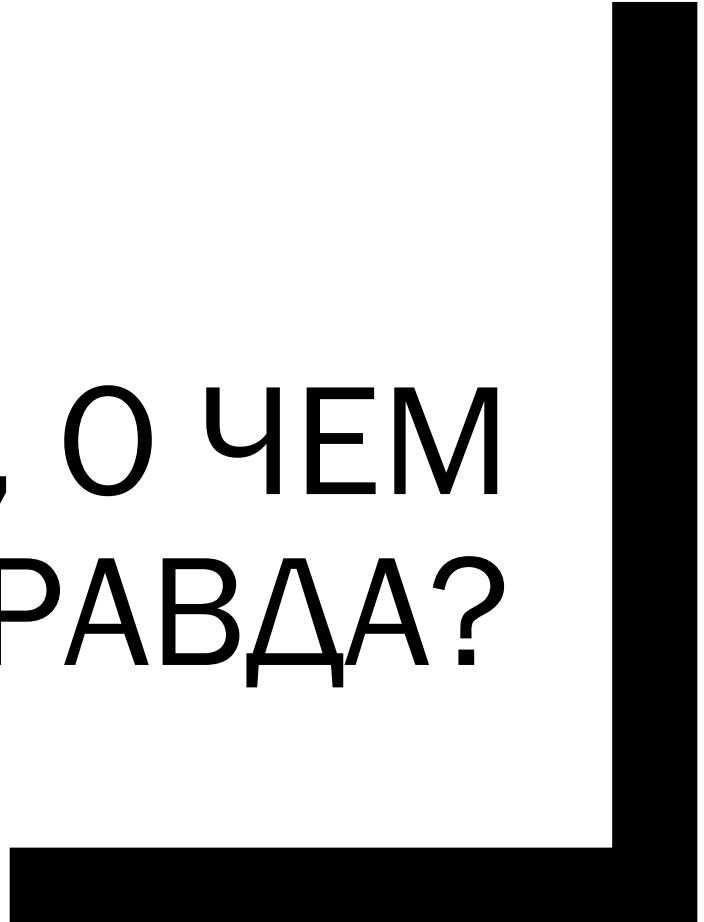
Эксперимент Б.В. Орехова и М.С. Зорькиной

# Delta для 60 русских текстов:



Эксперимент Б.В. Орехова

ЕЩЕ НЕ СОВСЕМ ТО, О ЧЕМ  
МЕЧТАЛ ЯРХО, ПРАВДА?



# Идея «Дальнего чтения» Франко Моретти (впервые – 2000 год):

«[...] if you want to look beyond the canon [...], close reading will not do it. It's not designed to do it, it's designed to do the opposite. [...] we know how to read texts, now let's learn how not to read them. Distant reading: where distance [...] is a condition of knowledge [...].»

Moretti F. Distant Reading. London: Verso, 2013. 244 p.

«Если же мы хотим выйти за пределы канона [...], то пристальное чтение нам не подходит. Оно не создано для таких задач, оно создано для решения задач противоположных. [...] мы умеем читать тексты, теперь нужно научиться не читать их. Дальнее чтение, для которого расстояние, повторюсь, является условием получения знаний [...].»

Моретти Ф. Гипотезы о мировой литературе (2000; пер. с англ. Олега Собчука) // В: Моретти Ф. Дальнее чтение. Москва, 2016, стр. 83

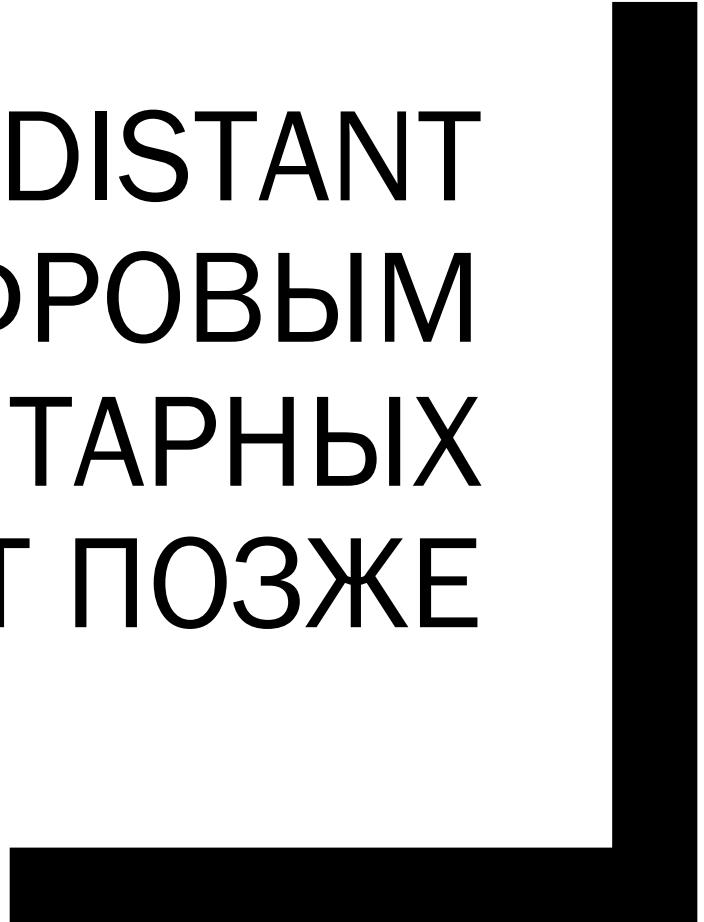
СНАЧАЛА МОРЕТТИ НЕ  
ДУМАЛ О «ЦИФРЕ»



**«это будет история литературы «из вторых рук» — мозаика, состоящая из исследований других людей, без какого-либо непосредственного прочтения текстов. Это не менее амбициозно, чем раньше (мировая литература!), но теперь амбиция прямо пропорциональна расстоянию от текста: чем более амбициозен замысел, тем большим должно быть расстояние».**

Моретти Ф. Гипотезы о мировой литературе (2000; пер. с англ. Олега Собчука) // В: Моретти Ф. Дальнее чтение. Москва, 2016, стр. 83

АДАПТАЦИЯ ИДЕИ DISTANT  
READING К ЦИФРОВЫМ  
МЕТОДАМ В ГУМАНИТАРНЫХ  
НАУКАХ ПРОИСХОДИТ ПОЗЖЕ



# Шокирующие признания от Моретти:

“Literature scholars should **stop**  
reading books, and start counting,  
graphing, and mapping them instead.”

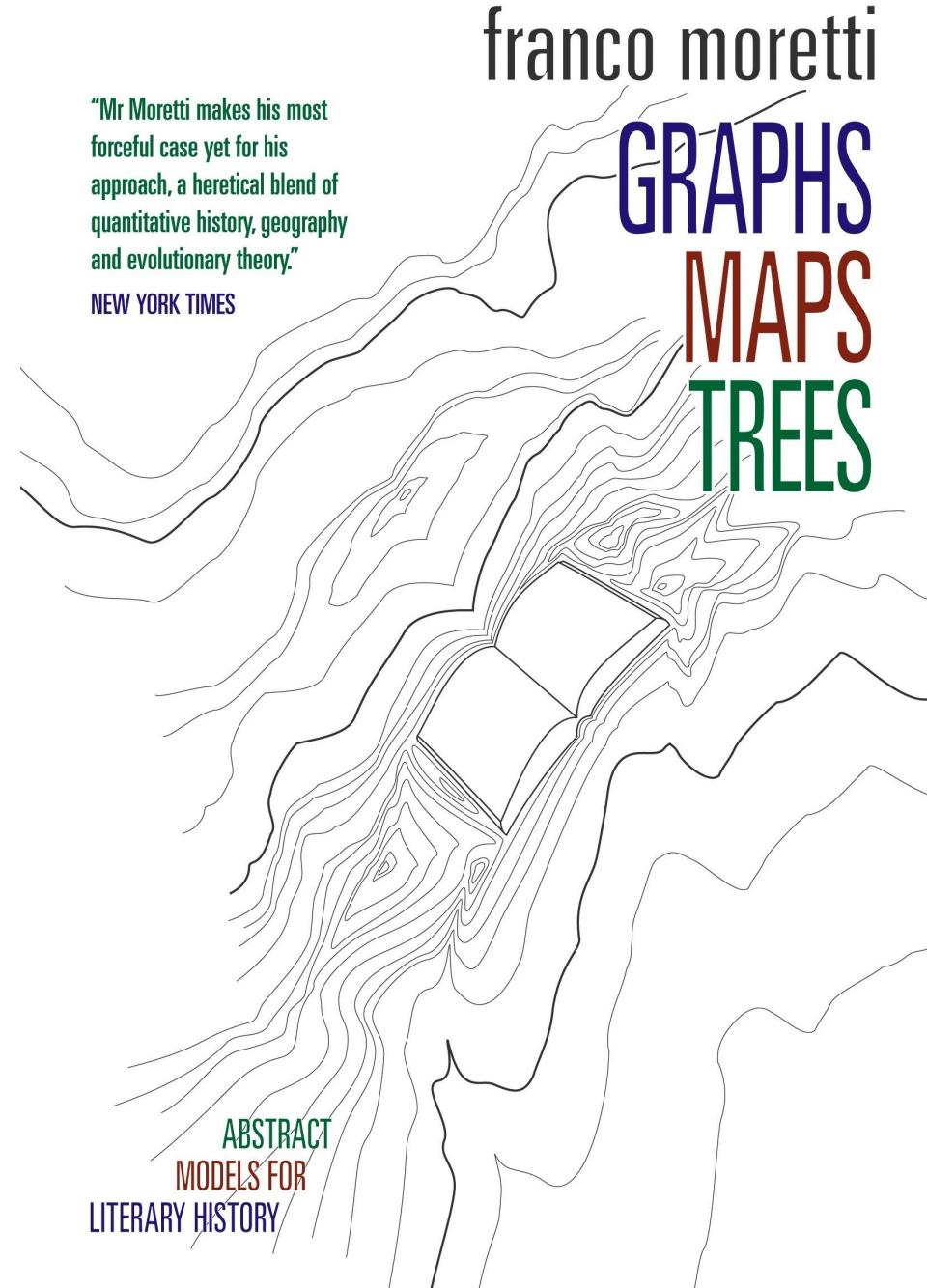
Moretti F. Graphs, Maps, Trees: Abstract Models  
for a Literary History. London: Verso, 2005.

Литературоведы должны прекратить  
читать книги и начать считать,  
визуализировать и картографировать  
их

Мой перевод

“Mr Moretti makes his most  
forceful case yet for his  
approach, a heretical blend of  
quantitative history, geography  
and evolutionary theory.”

NEW YORK TIMES



# Style Inc.

- Корпорация стиля: размышления о 7 тысячах заглавий (Ф. Моретти, 2009)
- «Британский роман 1740–1850 гг. Находящийся на периферии, часто презираемый в начале, к концу этого периода роман перемещается намного ближе к ядру национальной культуры. Поэтому это важный век в истории данной литературной формы». (Ф. Моретти, 2009)

# Что происходит?

«Основная метаморфоза названий XVIII в. проста: за время жизни двух поколений они становятся намного короче».

КОРПОРАЦИЯ СТИЛЯ



РИС. 1. Длина заглавий

# Очень короткие

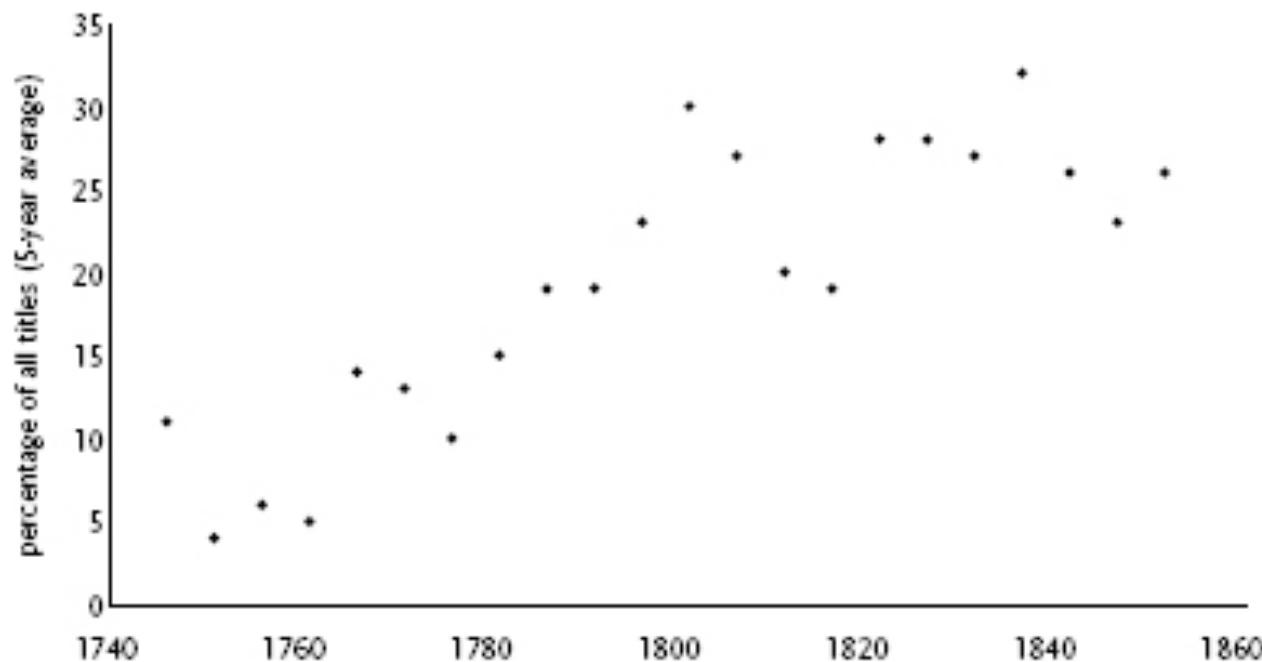


Figure 9: novels with very short titles

# Очень длинные

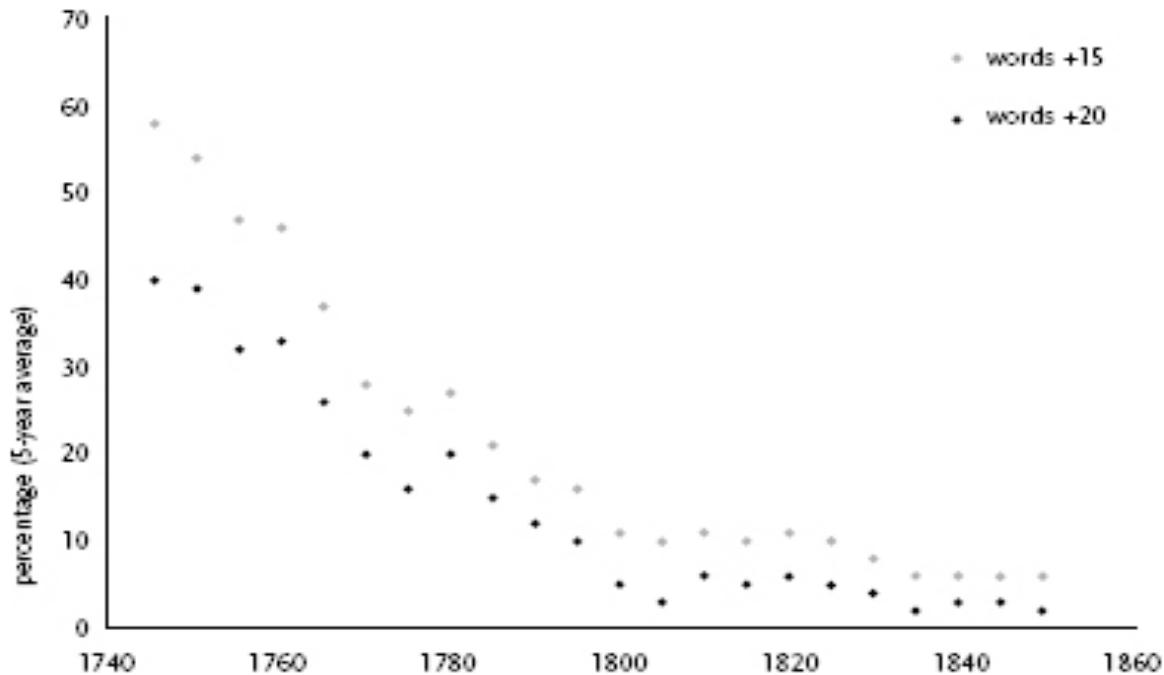


Figure 5: Novels with very long titles

# Длинные vs короткие

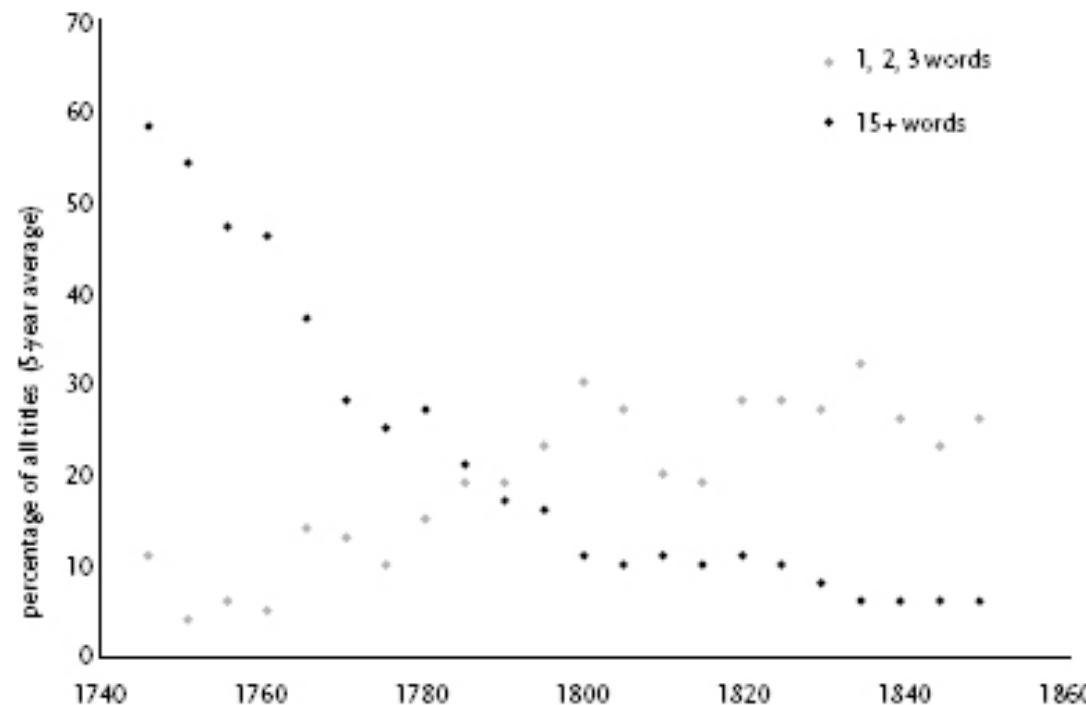


Figure 10: long titles, short titles

# А как выглядели длинные?

Письмо от Х—г—г, эсквайра, одного из лордов опочивальни, к молодому шевалье и единственному человеку из его свиты, сопровождавшему его во время долгого путешествия от Авиньона по Германии и другим местам; содержит множество замечательных и чувствительных происшествий, случившихся с П — в его таинственном странствии. Близкому другу.

(A letter from H—g—g, Esq; One of the Gentlemen of the Bedchamber to the Young Chevalier, And the Only Person of his Retinue that attended him from Avignon, in his late Journey through Germany, and elsewhere)

A

# LETTER

F R O M

H----G---g, Esq;

One of the Gentlemen of the Bed-chamber  
to the Young *Chevalier*, and the only Per-  
son of his own Retinue that attended him  
from *Avignon*, in his late Journey through  
*Germany*, and elsewhere:

CONTAINING

Many remarkable and affecting Occurrences, which  
happened to the P—, during the Course of  
his mysterious Progress.

T O

A PARTICULAR FRIEND.

---

*Vix irrix fortunæ sapientia.*

JUVENAL.

LONDON:

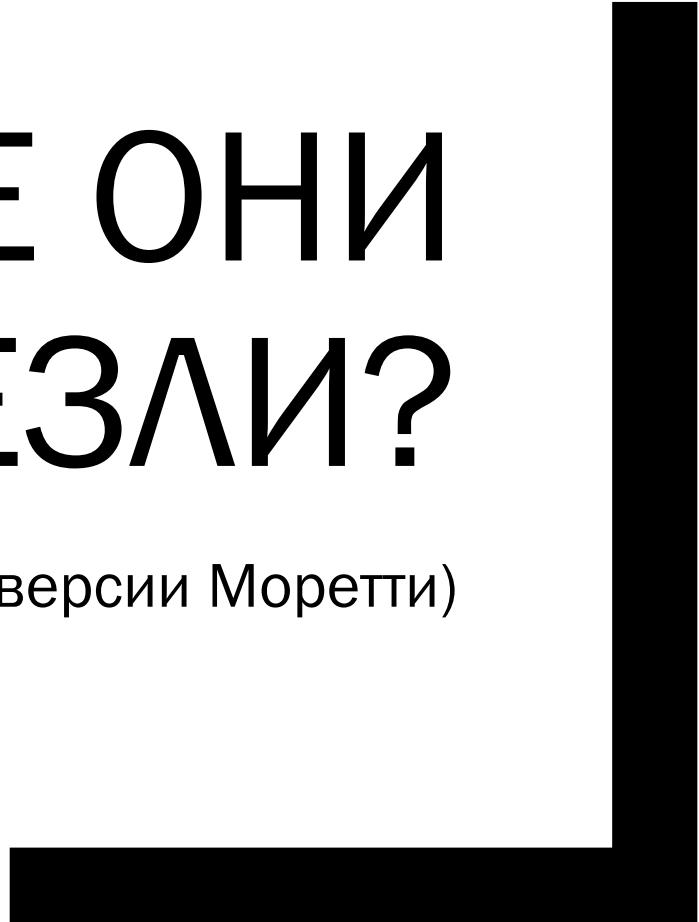
Printed, and Sold at the Royal Exchange, Temple-Bar,  
Charing-Cross, and all the Pamphlet-shops of London

«Сегодня это звучит странно, но на самом деле краткий пересказ в начале романа имеет смысл: роман — это повествование, а заглавие (в случае с титульным листом можно понять, зачем книге требовалась целая страница для титула) в качестве пересказа было укороченным повествованием — оно представляло основные события истории, персонажей, место действия, концовку. Это имело смысл».

(Ф. Моретти, 2009)

# ТАК ПОЧЕМУ ЖЕ ОНИ ИСЧЕЗЛИ?

(по версии Моретти)



# Общее число книг

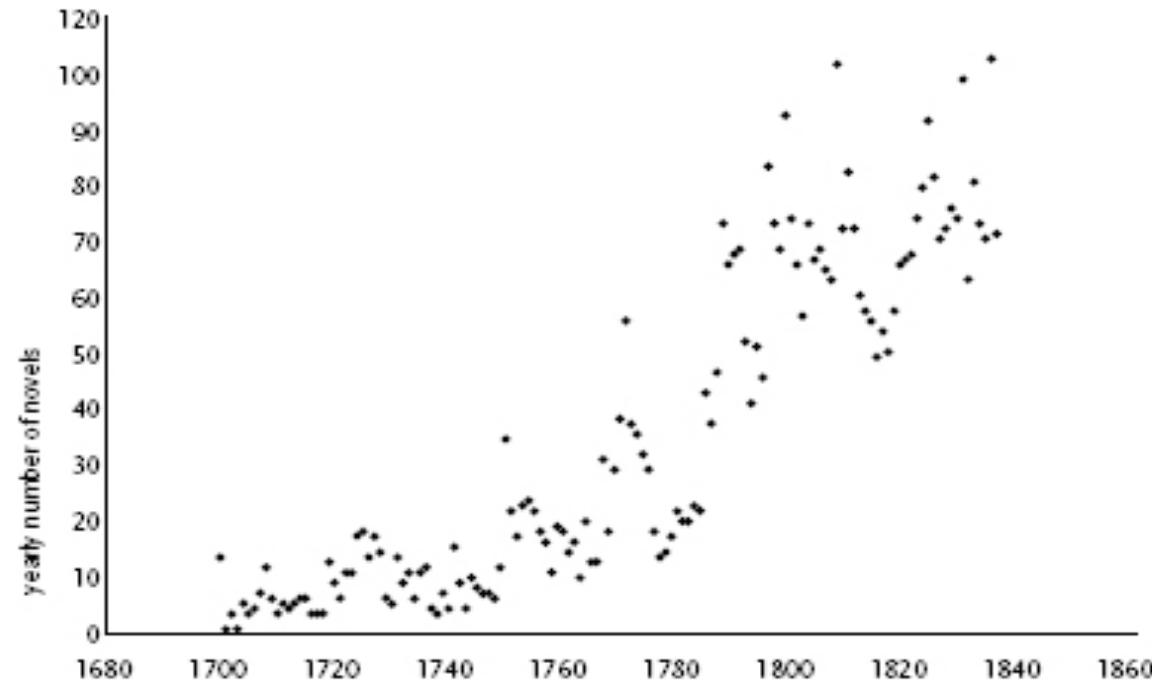


Figure 7: Publication of British novels, 1700-1836

<..> культурная экосистема изменялась таким образом, что становилась несовместимой с этими принципами: на протяжении XVIII в. количество опубликованных романов в Британии существенно выросло <..> И по мере циркулирования романов произошло две вещи.

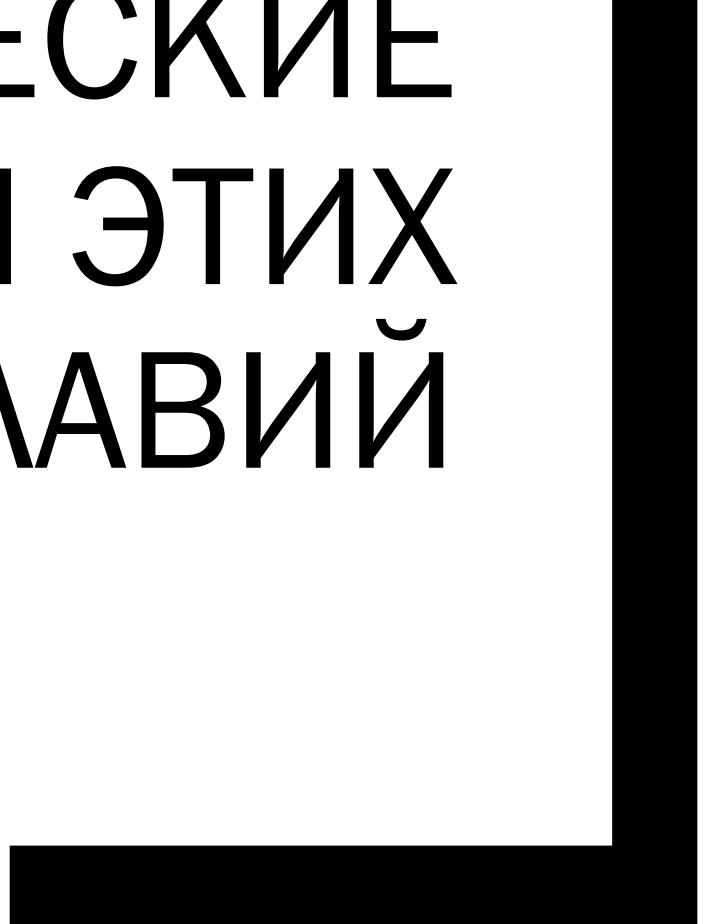
(Ф. Моретти, 2009)

# Что за две вещи?

1. В третьей и особенно в четвертой четверти XVIII в. Monthly и другие журналы стали печатать рецензии на многочисленные новые романы, что сделало заглавия-пересказы в некотором смысле избыточными. <...>
2. Кроме того, поскольку количество новых романов продолжало расти, временное окно для представления каждого из них на рынке сузилось, и для названия стало жизненно необходимым быстро и эффективно привлекать внимание публики. Пересказы не были для этого приспособлены. Они хорошо описывали книгу саму по себе, однако, когда дело касалось переполненного рынка, короткие заглавия справлялись лучше – хотя бы потому, что их было легче запомнить.

(Ф. Моретти, 2009)

ЛИНГВИСТИЧЕСКИЕ  
ОСОБЕННОСТИ ЭТИХ  
ЗАГЛАВИЙ



# 1. «Никакой» вампир против неприличной жены:

- «<...>оказалось, что прилагательное вовсе не уточняет семантическое поле, оно его трансформирует. В комбинации artikel-sуществительное половина заглавий, описывающих социальный тип, относится к экзотическому трансгрессивному полу — «Факир», «Вампир», «Пожиратель огня»
- «Однако стоит только появиться прилагательному, соотношение оказывается прямо противоположным (рис.11): факиры и распутники падают с 50 до 20%, а жены и дочери поднимаются с 16 до 40%: «Неприличная жена», «Брошенная дочь», «Неверный отец», «Братья-соперники», «Посмертно рожденная дочь», «Ложный друг», «Безумный отец»

(Ф. Моретти, 2009)

# 1. «Никакой» вампир против неприличной жены:

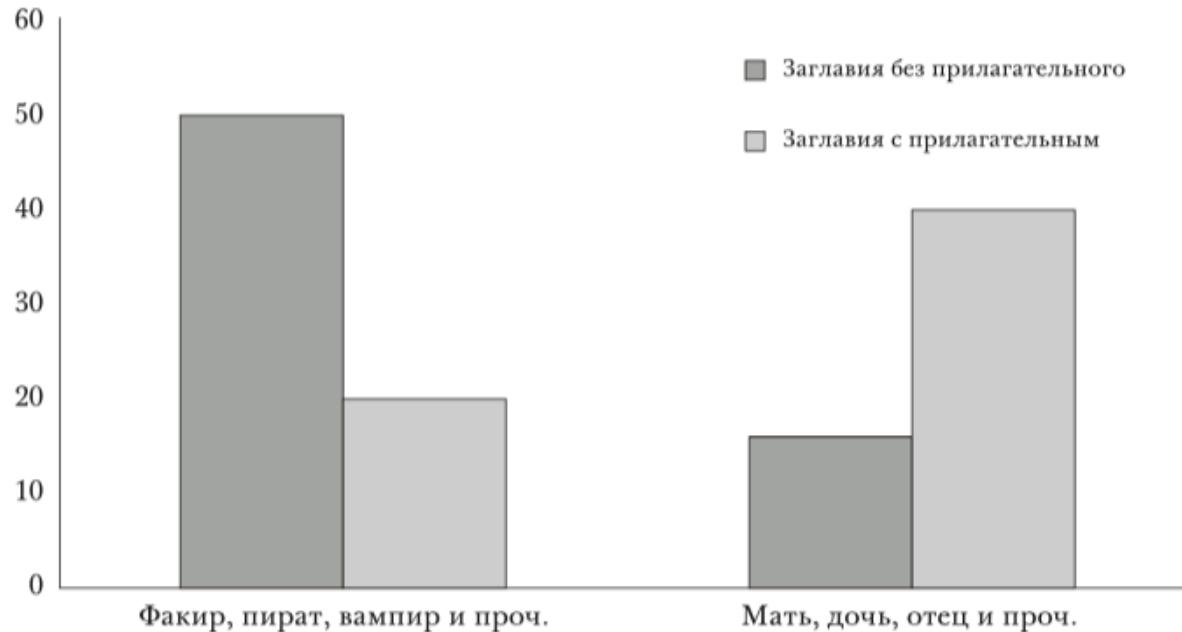


РИС. 11. Семантические поля в очень коротких заглавиях: роль прилагательных

# 1. «Никакой» вампир против неприличной жены:

«Без прилагательных мы находимся в мире приключений, а с прилагательными – в нарушенном домашнем укладе. <...> если в названии присутствует только существительное, то это существительное должно гарантировать интересную историю само по себе, и вампиры, и отцеубийцы в этом случае являются хорошим выбором. Однако когда появляется прилагательное, то даже хорошо знакомые фигуры могут стать чуждыми, превратившись в неверных отцов и посмертно рожденных дочерей».

# Женщины в заглавиях обретают фамилии:

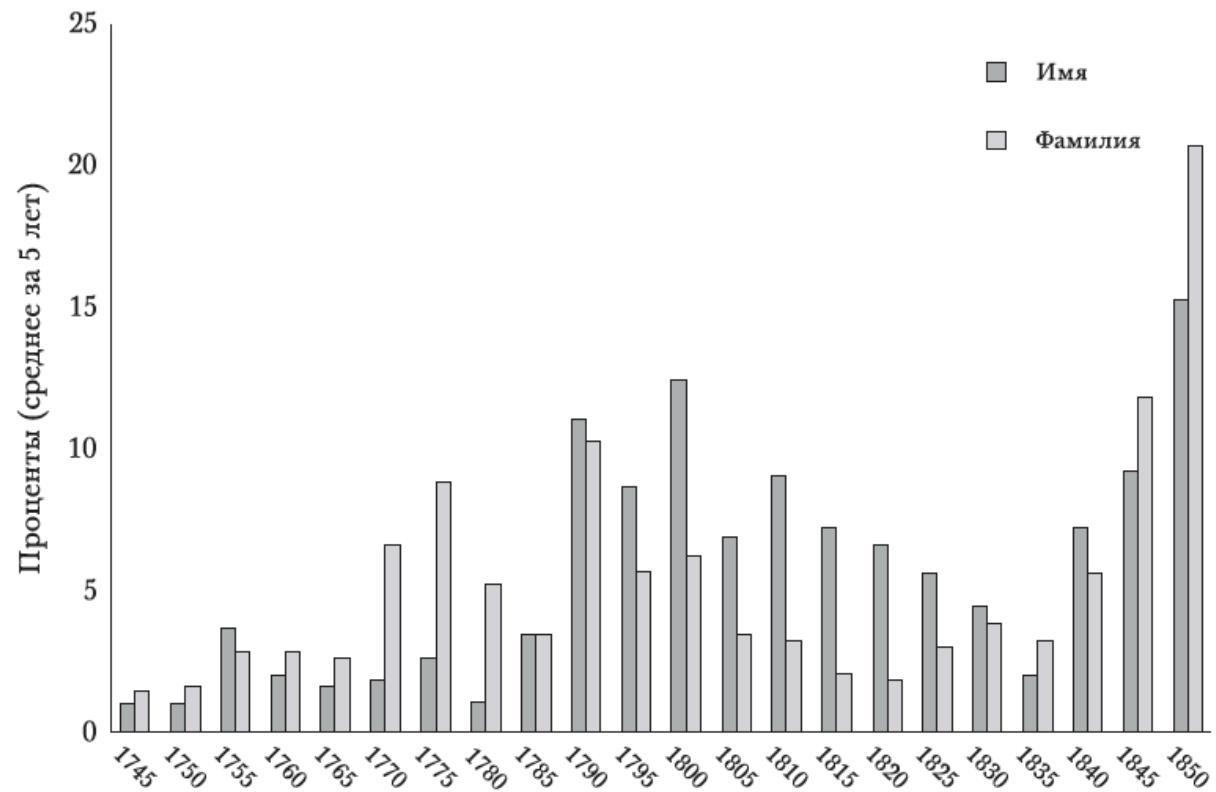


РИС. 16. Все заглавия, состоящие из женского имени

# Женщины в заглавиях обретают фамилии:

- «<...>героиня приобретала общественное положение, сразу отражающееся в названиях вроде «Джейн Эйр» или «Мэри Бартон». Обратите внимание, как много может быть сделано в коротких заглавиях благодаря небольшим вариациям: одно слово — и образ героини переворачивается на 180 градусов: от частного к публичному. Короткие названия были ограничением, навязанным рынком, да, но ограничение могло также стать замечательной возможностью для литературного воображения: искусство намека, сжатости — в конце концов заглавие становилось тропом».

# The X of Y

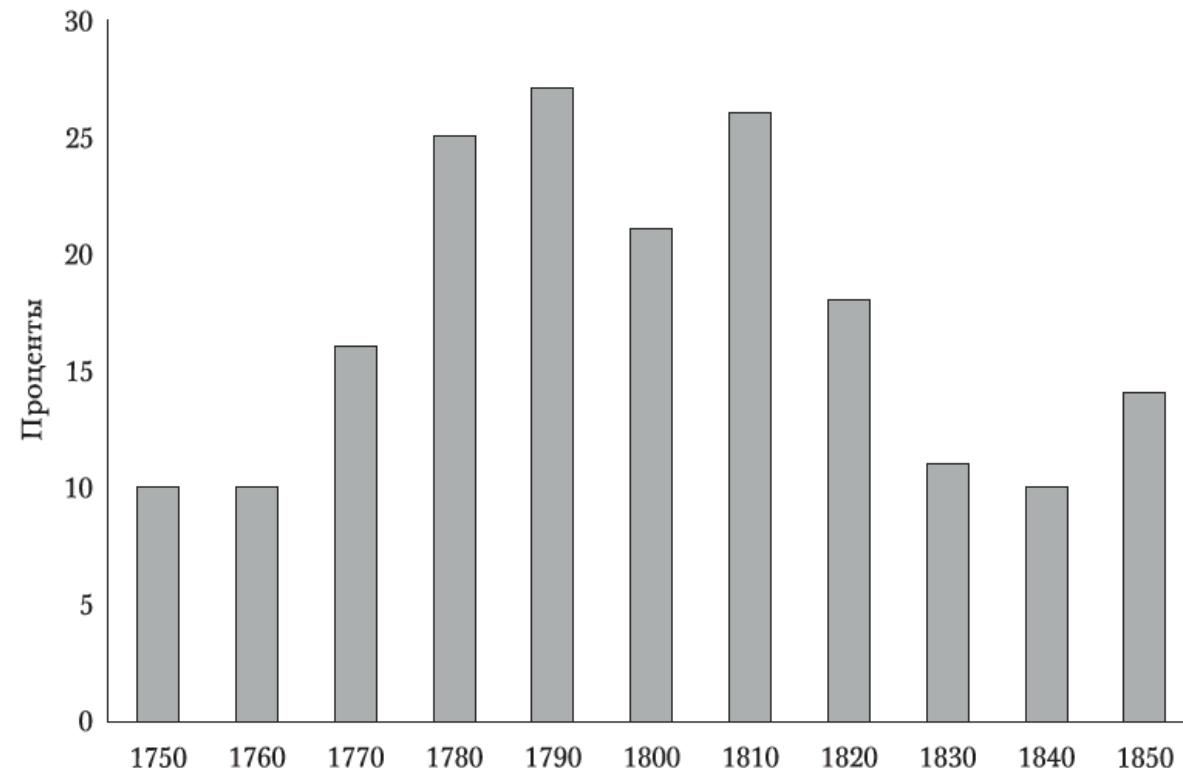


РИС. 19. Удачливая формула «the x of y»

# The X of Y

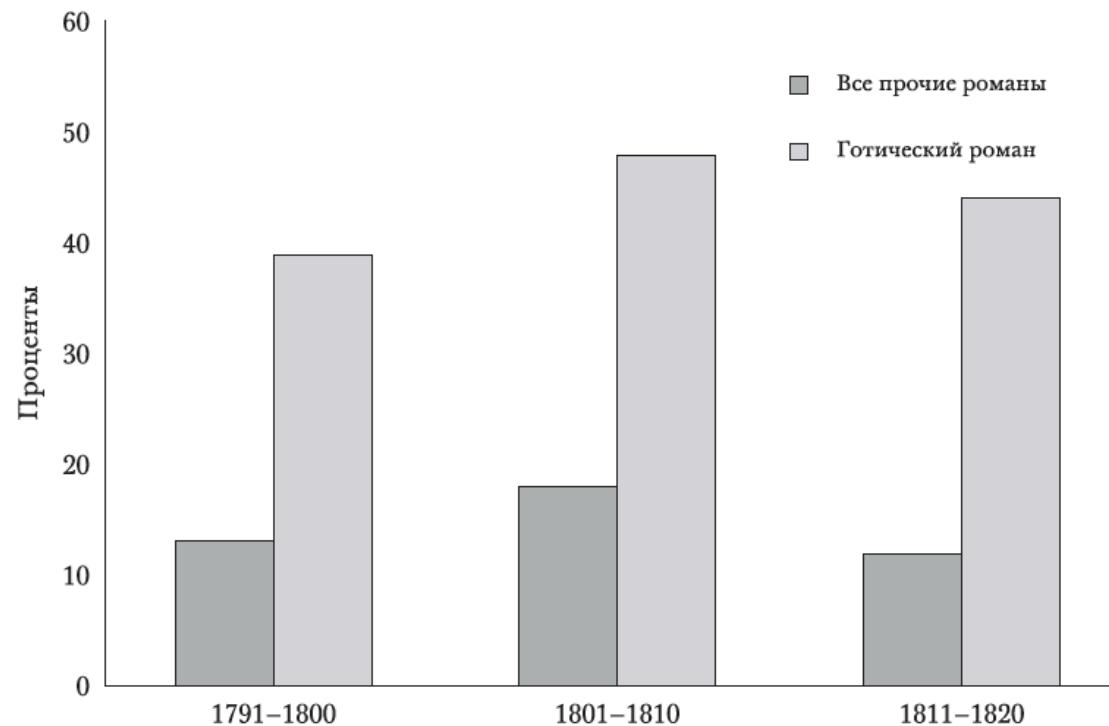


РИС. 20. Распределение формулы «the x of y»,  
1791–1820

# The X of Y

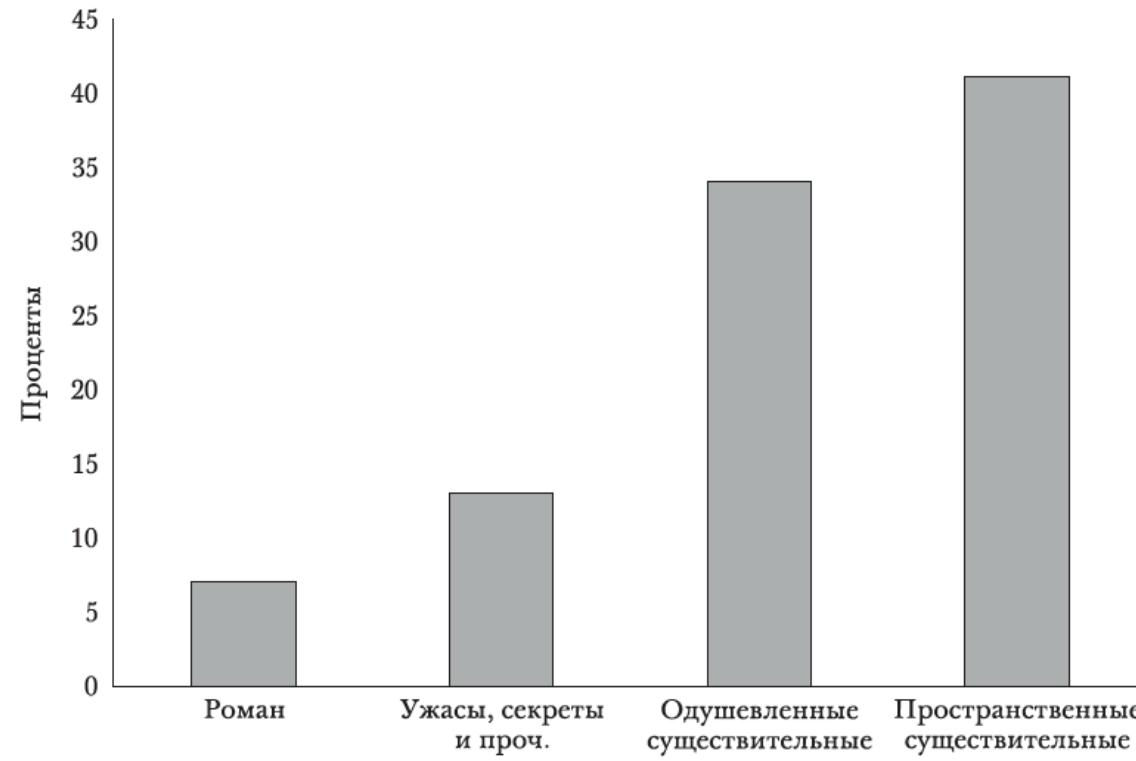


РИС. 21. «Х» в формуле «the X of Y»

# Дальнейшее дальнее чтение

- Многие цифровые гуманитарии подхватили идею Distant Reading
- Тем более что вокруг цифровой бум 2.0 середины 2000-х
- Вокруг самого Моретти появилась стэнфордская лаборатория с программистами



# Stanford Literary Lab

Director: Mark Algee-Hewitt

---

[About](#)   [People](#)   [Pamphlets](#)   [Projects](#)   [Techne](#)   [Events](#)

The Stanford Literary Lab is a research collective that applies computational criticism, in all its forms, to the study of literature. The Lab is open to students and faculty at Stanford, and, on a more ad hoc basis, to those from other institutions.

Our projects range from dissertation chapters to individual and group publications, lectures, courses, panels, and conferences. Typically, our research takes the form of an experiment that is then published in our [Pamphlet](#) series. Under [Projects](#) you will find the abstracts of our current activities; under [People](#), a list of those associated with our research. And under [Techne](#), our technical blog, we feature posts on our methodologies, both computational and critical.

# Измерения «литературной эволюции»

- Стэнфордская литературная лаборатория (Stanford Literary Lab)
- Метаморфозы британского романа в XVIII – XIX вв.  
(серия количественных исследований)
- Связь текстовой статистики – со
  - сменой жанров и направлений в литературе
  - социальными и экономическими процессами

# Измерения «литературной эволюции»

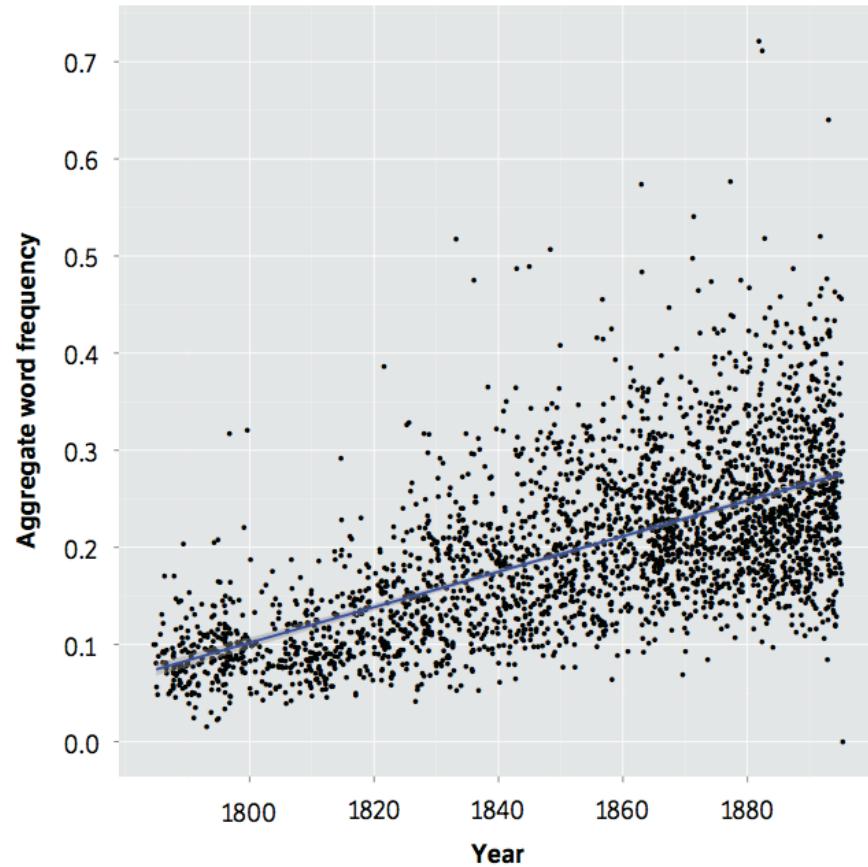


Figure 14: Aggregate term frequencies of the physical adjectives field in novels, 1785-1900.

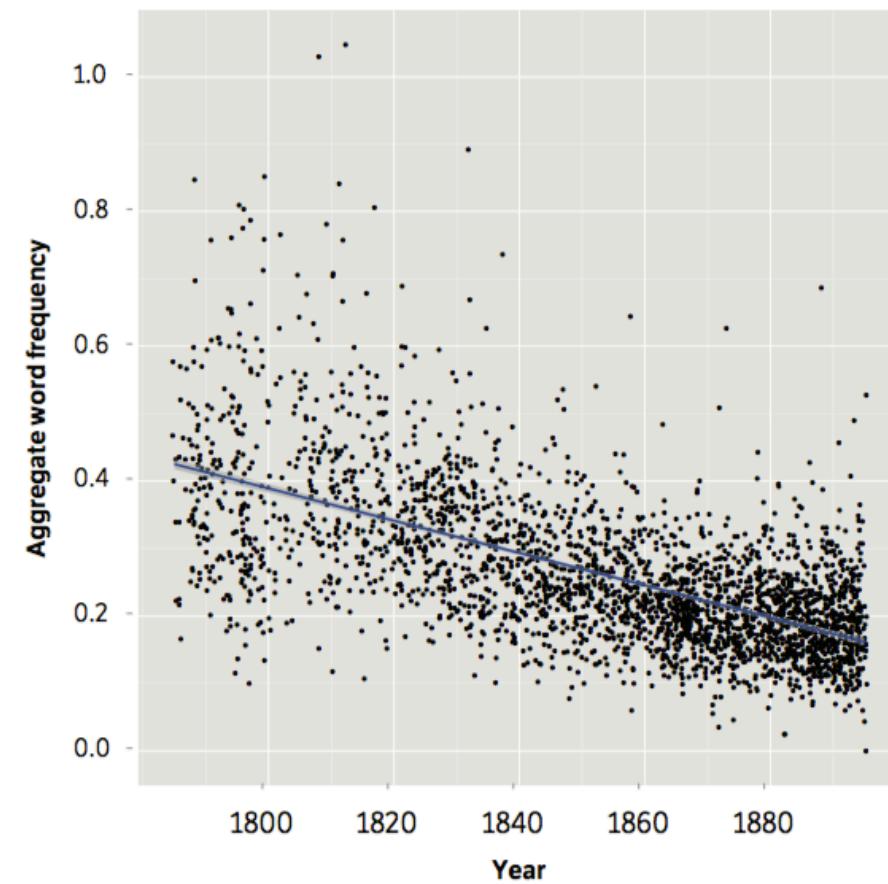
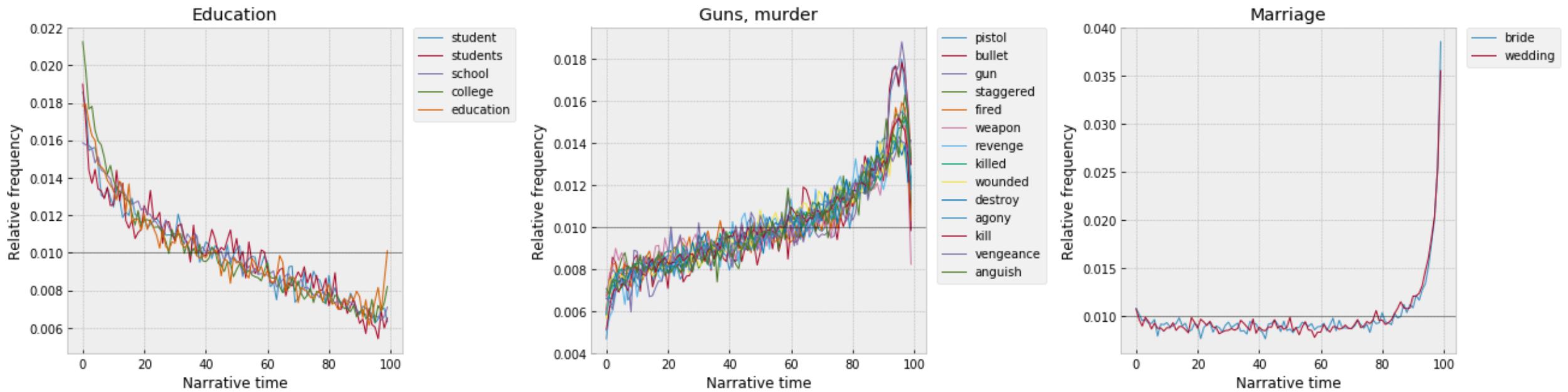


Figure 5: Aggregate term frequencies of the moral valuation field in novels, 1785-1900.

A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method  
(litlab.stanford.edu/LiteraryLabPamphlet4.pdf)

# Частотности групп слов в 50000 романов:



McClure D.W. Distributions Of Function Words Across Narrative Time In 50,000 Novels. // Digital Humanities 2018: Book of Abstracts / Libro de resúmenes. 2018.

Спасибо за внимание!