Ю.Н. Орлов

**Институт прикладной математики им. М.В. Келдыша РАН,**
**кафедра высшей математики МФТИ**

# Статистическое распознавание образов в задачах с большими данными

(применительно к литературным текстам)

# План доклада

1. Использование функций распределения для задач с Большими Данными
2. Вычислительные аспекты байесовского распознавания образов
3. Кинетические методы анализа временных рядов на примере литературных текстов:

- распознавание авторских эталонов – распределений n-грамм
- способы кластеризации выборочных распределений
- спектральный портрет стохастической матрицы
- распределение букв по частоте встречаемости в разных языках
- показатель Херста ряда расстояний между символами

4. Примеры:

- спорные вопросы авторства: Шолохов-Крюков, Шекспир-Марло, Рерих-Блаватская, Булгаков-Ильф/Петров
- индикатор языковой группы индоевропейской семьи
- анализ алфавита, на котором написан Манускрипт Войнича

# 1. Точность оценки выборочной плотности функции распределения текстов по n-граммам

# Выборочные функции распределения в задачах с Большими Данными

Для временных рядов надо уметь строить совместные выборочные распределения приращений одного-двух-трех порядков по выборкам любого объема в любой момент времени в пределах заданной совокупности.

Если данных 10^9, то для анализа изменения ВПФР с точностью 1% по двум переменным, длине выборки и времени (с учетом сдвига по времени) требуется массив (10^2)^2*(10^9)^3 = 10^31 данных.

$$\rho(N;t;\tau) = \left\| f_N(x,t) - f_N(x,t+\tau) \right\|$$

# Оценка достаточной длины текста

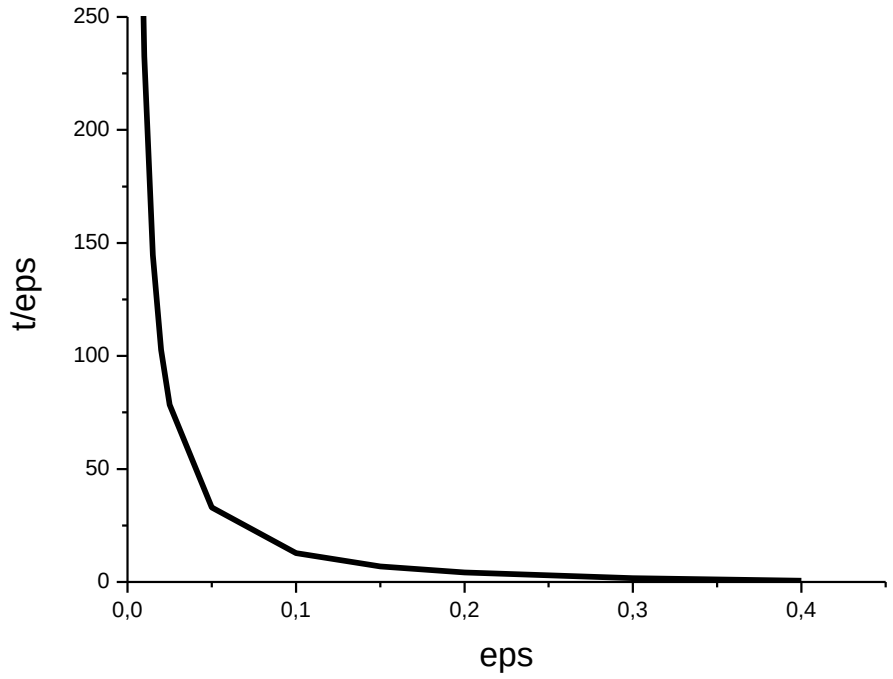
- Оценка средней частоты буквы в выборке при неизвестной дисперсии дается статистикой Стьюдента

$$t_\alpha(N-1)=\sqrt{N-1}\,\frac{\left|f_N(i)-f^*(i)\right|}{s_N(i)},\quad s_N(i)=\sqrt{f_N(i)\left(1-f_N(i)\right)}$$

- Оценка длины текста для построения распределения с точностью ε:

$$\frac{t_{1-\varepsilon/2}}{\varepsilon}\le\frac{\sqrt{N}}{\Sigma_N(n)},\quad \Sigma_N(n)=\sum_{i=1}^{n}\sqrt{f_N(i)\left(1-f_N(i)\right)},\quad \sum_{i=1}^{n}\left|f_N(i)-f^*(i)\right|\le\varepsilon$$

# Достаточная длина текста




$$\Sigma_1 \approx 5{,}1$$
$$\Sigma_2 \approx 19{,}5$$
$$\Sigma_3 \approx 61{,}1$$
$$\Sigma_4 \approx 108{,}3$$

$$N_{\min}(k) = \left( \frac{\Sigma_k \cdot t_{1-\varepsilon/2}}{\varepsilon} \right)^2$$

- При ε=0,05:

  для 1-ПФР $N$=40 тыс. знаков,

  для 2-ПФР $N$=400 тыс. знаков,

  для 3-ПФР $N$=4 млн знаков.

# Точность оценки n-ПФР

| Длина текста | 1-ПФР | 2-ПФР | 3-ПФР | 4-ПФР |
|---|---|---|---|---|
| $N$ = 10 тыс. | **0,08** | **0,22** | **0,40** | **0,60** |
| $N$ = 30 тыс. | **0,05** | **0,15** | **0,30** | **0,42** |
| $N$ = 50 тыс. | **0,04** | **0,12** | **0,25** | **0,39** |
| $N$ = 100 тыс. | **0,03** | **0,10** | **0,20** | **0,31** |
| $N$ = 500 тыс. | **0,02** | **0,05** | **0,15** | **0,20** |
| $N$ = 1 млн | **0,01** | **0,04** | **0,10** | **0,15** |

# Уровень нестационарности текстов

- Расстояние между ПФР текстов:

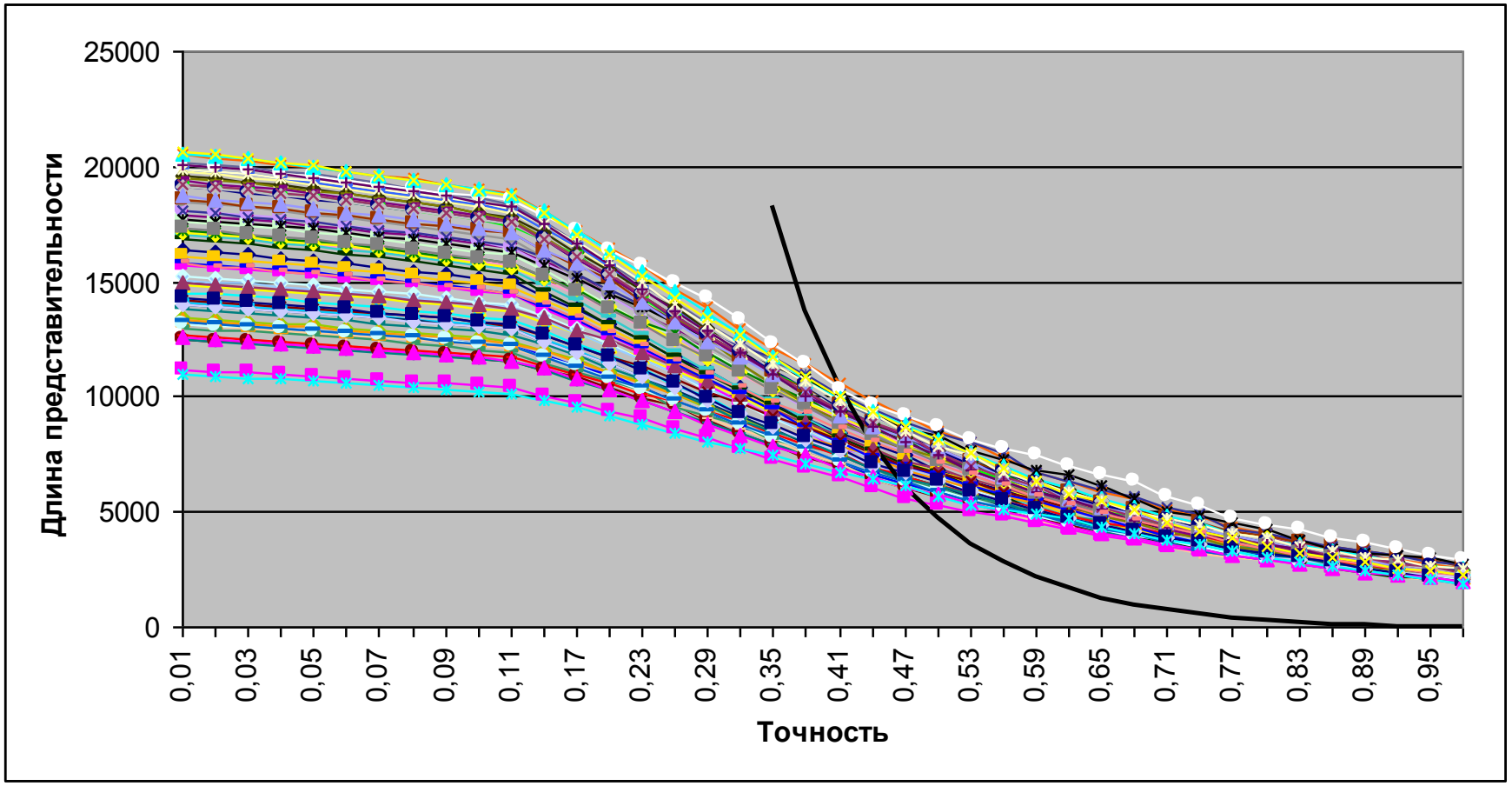
$$\rho_{12} = \left\| f^{(1)} - f^{(2)} \right\| = \sum_{i=1}^{n} \left| f_{N_1}^{(1)}(i) - f_{N_2}^{(2)}(i) \right|$$

- Чтобы сравнивать распределения текстов разных объемов, следует убедиться в том, что каждый из них стабилизируется на характерной «длине авторской представительности»:

$$\exists\, L(\varepsilon):\ \forall\, N_1, N_2 \geq L(\varepsilon) \quad \sum_{i=1}^{n} \left| f_{N_1}(i) - f_{N_2}(i) \right| \leq \varepsilon \quad \Rightarrow$$

$$\textit{для текста длины } N \quad L(\varepsilon) \leq \left[ (1-\varepsilon) N \right]$$

# Длина представительности L(ε) для 3-ПФР




Распределение символов в литературных текстах становится стационарным на длине порядка 10 тыс. знаков.

# 2. Составление эталонных распределений и метод идентификации автора текста

# Идентификация автора текста (Байесовское распознавание)

Пусть имеется библиотека из $A$ авторов, у $a$-го автора $K_a$ текстов, в $i$-ом тексте $N_{i,a}$ знаков, и $f_{i,a}(j)$ есть ПФР отдельного текста. Вводится эталонная ПФР автора:

$$f_a(j) = \frac{1}{N_a} \sum_{i=1}^{K_a} f_{i,a}(j) N_{i,a}, \quad N_a = \sum_{i=1}^{K_a} N_{i,a}.$$

Пусть $f_0(j)$- ПФР текста неизвестного автора. Автор определяется по правилу

$$\rho_a^0 = \| f_0 - f_a \|, \quad a^0 = \arg\min_a \rho_a^0$$

# Проектирование на пространство авторов

$$\mathbf{f} \in R^n \qquad \left\|\mathbf{f} - \Phi\mathbf{c}\right\|_{L2} \to \min \qquad \Phi = \left(\varphi_1,...,\varphi_p\right),\ \varphi_i \in R^n$$

Делается QR-разложение матрицы Ф эталонов:

$$\mathbf{f} - \Phi\mathbf{c} = I\mathbf{f} - QR\mathbf{c} = \left(I - QQ^T + QQ^T\right)\mathbf{f} - QR\mathbf{c} = Q\left(Q^T\mathbf{f} - R\mathbf{c}\right) + \left(I - QQ^T\right)\mathbf{f}$$

$$\left(R\mathbf{c} - Q^T\mathbf{f}\right)^T Q^T\left(I - QQ^T\right)\mathbf{f} = \left(R\mathbf{c} - Q^T\mathbf{f}\right)^T\left(Q^T{}_{p\times n} I_{n\times n} - I_{p\times p} Q^T{}_{p\times n}\right)\mathbf{f} =$$

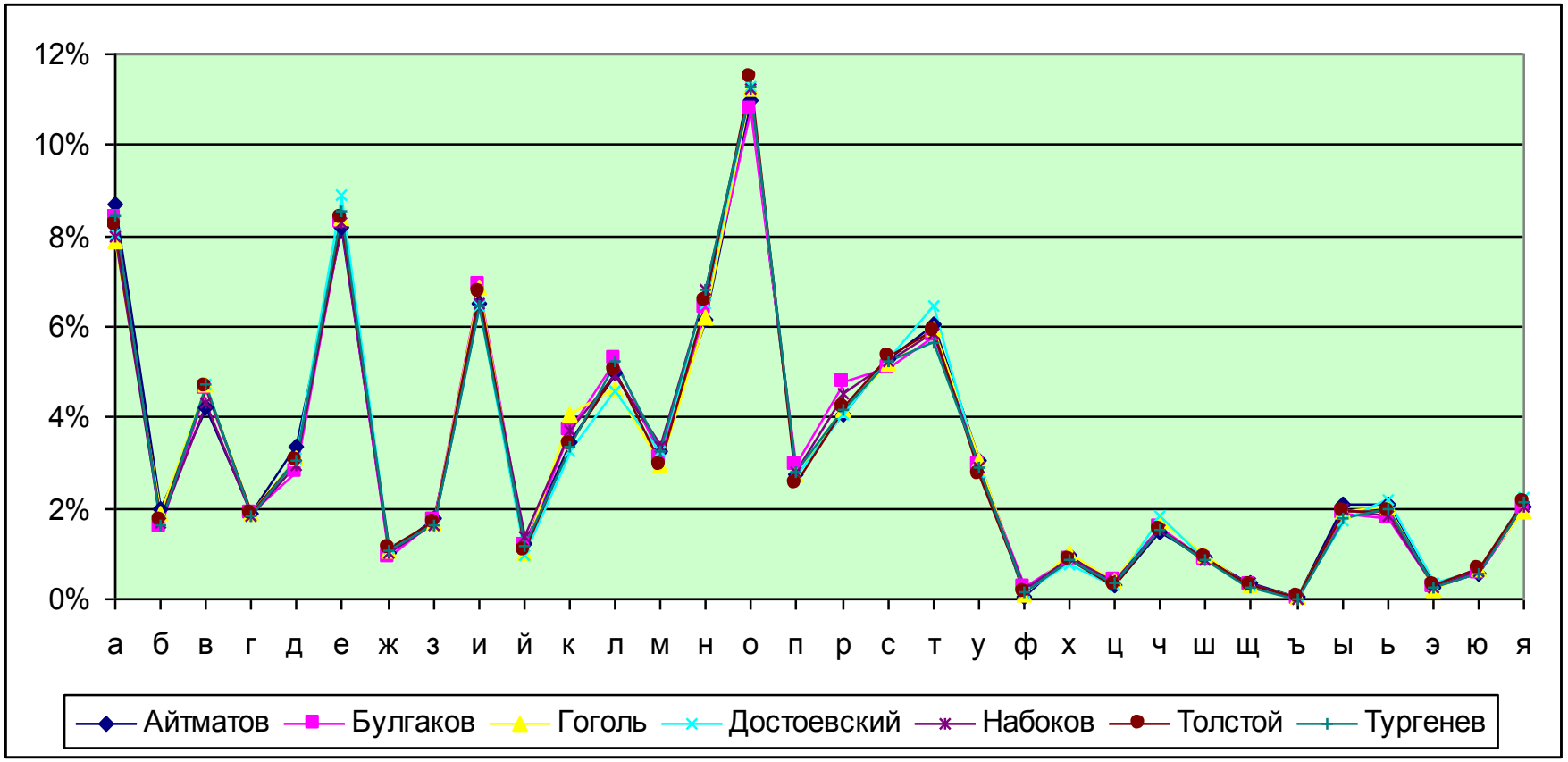
$$= \left(R\mathbf{c} - Q^T\mathbf{f}\right)^T O_{p\times n}\mathbf{f} = \mathbf{o}.$$

$$\mathbf{c}_{opt} = R^{-1}Q^T\mathbf{f} \qquad Q^TQ = I_{p\times p}$$

$$\sum_{i=1}^{n} f_i = 1, \quad \forall k \sum_{i=1}^{n}\varphi_{k,i} = 1, \quad f_i = \sum_{k=1}^{p} c_k\varphi_{k,i} \quad \Rightarrow \quad \sum_{k=1}^{p} c_k = 1$$

$$\mathbf{f} \equiv \varphi_j, \quad j = \arg\max c_k \quad \Rightarrow \quad j = \arg\min\left\|\mathbf{f} - \varphi_k\right\|$$

# Авторские 1-ПФР




- Вывод: авторские 1-ПФР очень близки, поэтому различие между ними должно выявляться на «тонкой структуре» их взаимных различий, а не функционала от них как таковых; вычислительная задача проектирования на «авторский базис» оказывается неустойчивой.

# Практические ограничения Байесовского распознавания

Решающее правило $\rho_a^0 = \| f_0 - f_a \|, \quad a^0 = \arg\min_a \rho_a^0$

корректно, если существует вероятностная интерпретация разложения неизвестного состояния по заданному базису:

$$\left\| \mathbf{f}_0 - \sum_{a=1}^{A} c_a f_a \right\| \le \varepsilon, \quad \sum_{a=1}^{A} c_a = 1, \quad c_a \ge 0; \quad a^0 = \arg\max c_a$$

При распознавании автора текста точное разложение по базису из авторских эталонов не имеет места, но решающее правило «ближайшего эталона» остается эффективным, хотя и не обоснованным, инструментом

Среднеквадратичное нарушение вероятностной интерпретации:

| **Число авторов** | **2** | **3** | **4** | **10** |
|---|---|---|---|---|
| **1-ПФР** | **0,15** | **0,26** | **0,54** | **2,50** |
| **2-ПФР** | **0,12** | **0,20** | **0,35** | **1,70** |
| **3-ПФР** | **0,09** | **0,18** | **0,29** | **1,25** |

# Вычислительная невязка

$$\left\| \mathbf{f} - \sum_{k=1}^{p} c_k \varphi_k \right\| \le \varepsilon, \quad \sum_{k=1}^{p} c_k = 1, \quad c_k \ge 0; \qquad \xi = \max\left( \frac{\|\Delta\Phi\|}{\|\Phi\|}, \frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} \right)$$

$$\frac{\|\Delta\mathbf{c}\|}{\|\mathbf{c}\|} \le \xi \cdot \left( \frac{2\kappa(\Phi)}{\cos\theta} + \kappa^2(\Phi) tg\,\theta \right) + O\left(\xi^2\right)$$

$$\kappa(\Phi) = \lambda_{\max} / \lambda_{\min}, \quad \sin\theta = \varepsilon$$

Число обусловленности

| Число авторов | 2 | 3 | 4 |
|---|---|---|---|
| 1-ПФР | 600 | 1800 | 3500 |
| 2-ПФР | 100 | 300 | 560 |
| 3-ПФР | 20 | 70 | 110 |

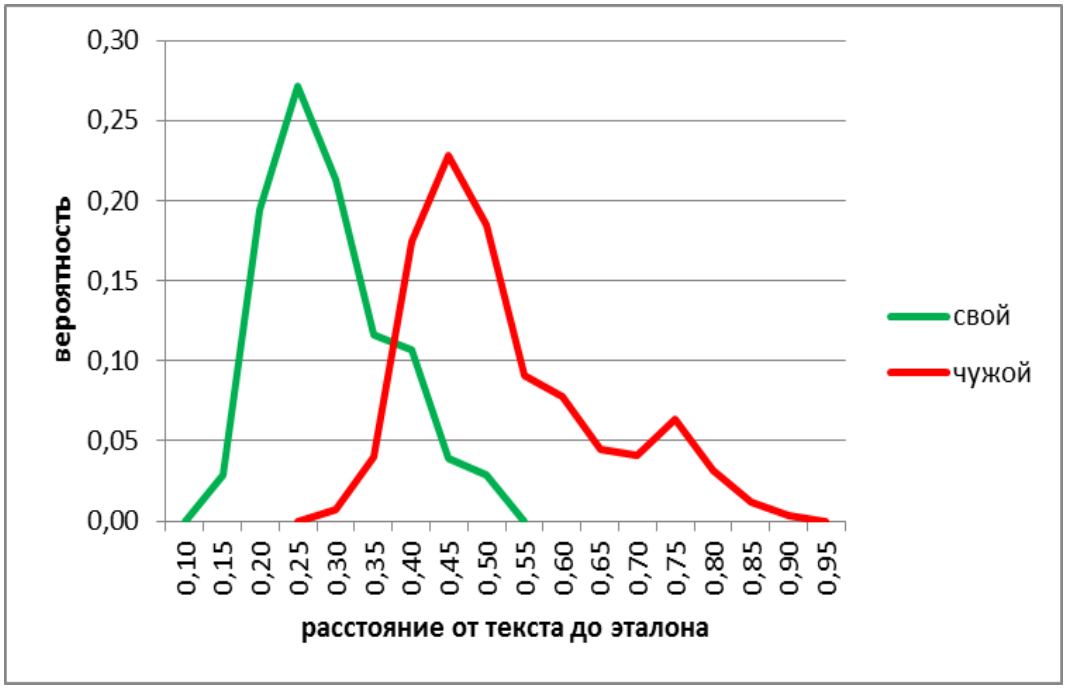
Среднеквадратичная невязка

| Число авторов | 2 | 3 | 4 |
|---|---|---|---|
| 1-ПФР | 0,05 | 0,05 | 0,05 |
| 2-ПФР | 0,16 | 0,16 | 0,16 |
| 3-ПФР | 0,33 | 0,33 | 0,33 |

# Задача Неймана-Пирсона

$$\rho_a^+ : \min \rho, F_a^+(\rho) = 1; \quad \rho_a^- : \max \rho, F_a^-(\rho) = 0$$

Существует уровень оптимального разделения: пропуск цели не превышает заданный уровень, а ложная тревога минимизируется




$F_a^+(\rho)$

функция распределения расстояний текстов автора от его эталона;

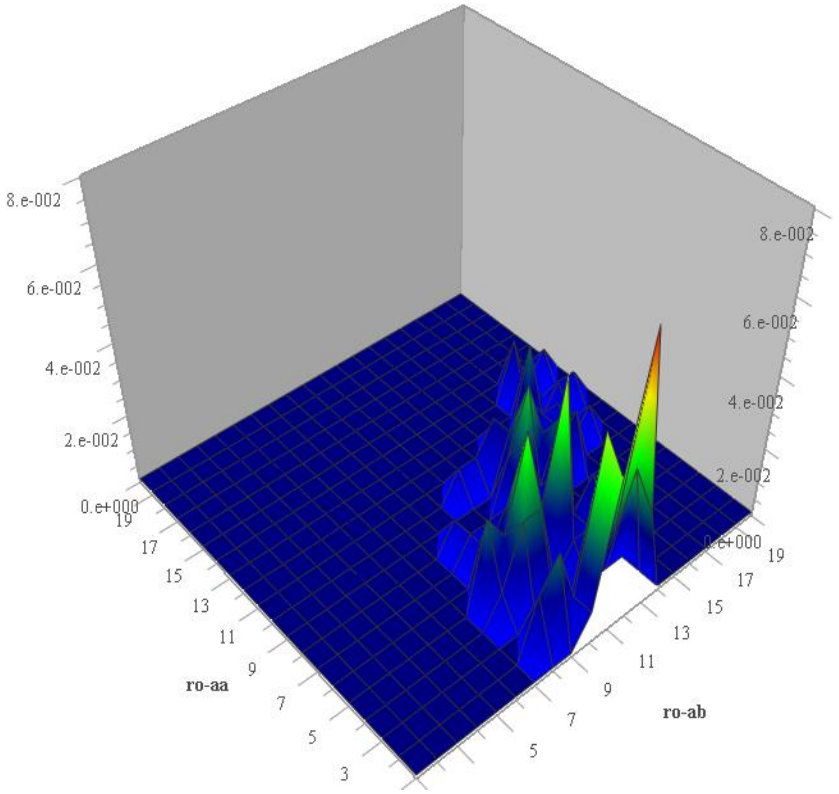
$F_a^-(\rho)$

чужих текстов от него же;

$F_a^-\left(\rho_a^+\right)$ есть вероятность ошибочно отклонить текст автора, посчитав его чужим (ошибка 1-го рода);

$1 - F_a^+\left(\rho_a^-\right)$ есть вероятность ошибочно признать чужой текст авторским (ошибка 2-го рода)

# Парное распределение расстояний «свой-чужой»



# Сравнение эффективности статистических методов идентификации автора

| 3000 текстов, 300 авторов | Ошибка, % |
|---|---|
| Близость 3-ПФР в норме L1 | 0 |
| Близость 4-ПФР в норме L1 | 2 |
| Близость 2-ПФР в норме L1 | 4 |
| Близость 1-ПФР в норме L1 | 15 |
| Доля служебных слов | 68 |
| Информационная энтропия 2-ПФР | 71 |
| Доля гласных | 81 |
| Среднее число слов в предложении | 87 |

# Кластеризация по жанрам путем создания эталона 1-ПФР

| Жанр | Боевик | Дамский детектив | Детектив | Ужас | Эротика | Классика | Соц. реализм |
|---|---|---|---|---|---|---|---|
| Б | **3,8** | 3,0 | 2,5 | 3,7 | 2,7 | 3,3 | 3,7 |
| ДД | | **4,1** | 2,9 | 3,7 | 3,0 | 4,1 | 3,3 |
| Д | | | **3,8** | 3,3 | 3,5 | 2,5 | 3,2 |
| У | | | | **4,2** | 3,3 | 4,1 | 3,5 |
| Э | | | | | **4,7** | 4,1 | 3,7 |
| К | | | | | | **4,4** | 3,1 |
| СР | | | | | | | **5,1** |

- Среднее расстояние 1-ПФР (%) между жанровым эталоном и текстом указанного жанра. Эталон для жанров не работает, ошибка идентификации 80 %.

# Кластеризация по жанрам методом попарной близости

| Жанр | Боевик | Дамский детектив | Детектив | Ужас | Эротика | Классика | Соц. реализм |
|---|---|---|---|---|---|---|---|
| Б | **5,5** | 6,2 | 5,8 | 6,5 | 7,1 | 6,4 | 6,7 |
| ДД | | **6,2** | 6,3 | 7,8 | 7,1 | 6,7 | 7,0 |
| Д | | | **5,7** | 7,0 | 7,4 | 6,5 | 7,0 |
| У | | | | **8,0** | 8,2 | 7,6 | 7,9 |
| Э | | | | | **7,4** | 7,7 | 7,6 |
| К | | | | | | **6,2** | 7,1 |
| СР | | | | | | | **6,4** |

- Среднее попарное расстояние между 1-ПФР текстов указанных жанров, %. Это хороший способ кластеризации, ошибка 20 %.

# 3. Спектральные портреты авторов, собственные векторы оператора трансляций, эффект переводчика

# Оператор трансляций на 1 шаг

- Рассмотрим стохастическую матрицу условной вероятности соседства пар символов, выраженную через 1-ПФР и 2-ПФР:

$$P_{jk} = F(k, j) / f(k)$$

- По формуле полной вероятности

$$f(j) = \sum_k F(k, j) = \sum_k P_{jk} f(k)$$

- Следовательно, 1-ПФР

$$f(j) = \sum_k F(k, j)$$

является с.в. оператора $P_{jk}$, отвечающим с.з. 1.

# ε-спектр оператора трансляций

- Число $\lambda$ называется принадлежащим $\varepsilon$-спектру $\Lambda_\varepsilon(P)$ матрицы P, если существует матрица $\Delta$ такая, что

$$\|\Delta\| \le \varepsilon \|P\| \qquad \det(\lambda E - P - \Delta) = 0$$

- Резольвентой R матрицы P называется матрица

$$R(\lambda) = (\lambda I - P)^{-1}$$

Тогда $\lambda \in \Lambda_\varepsilon(P)$ если $\|R(\lambda)\| \ge \dfrac{1}{\varepsilon \|P\|}$

# Вычисление ε-спектра

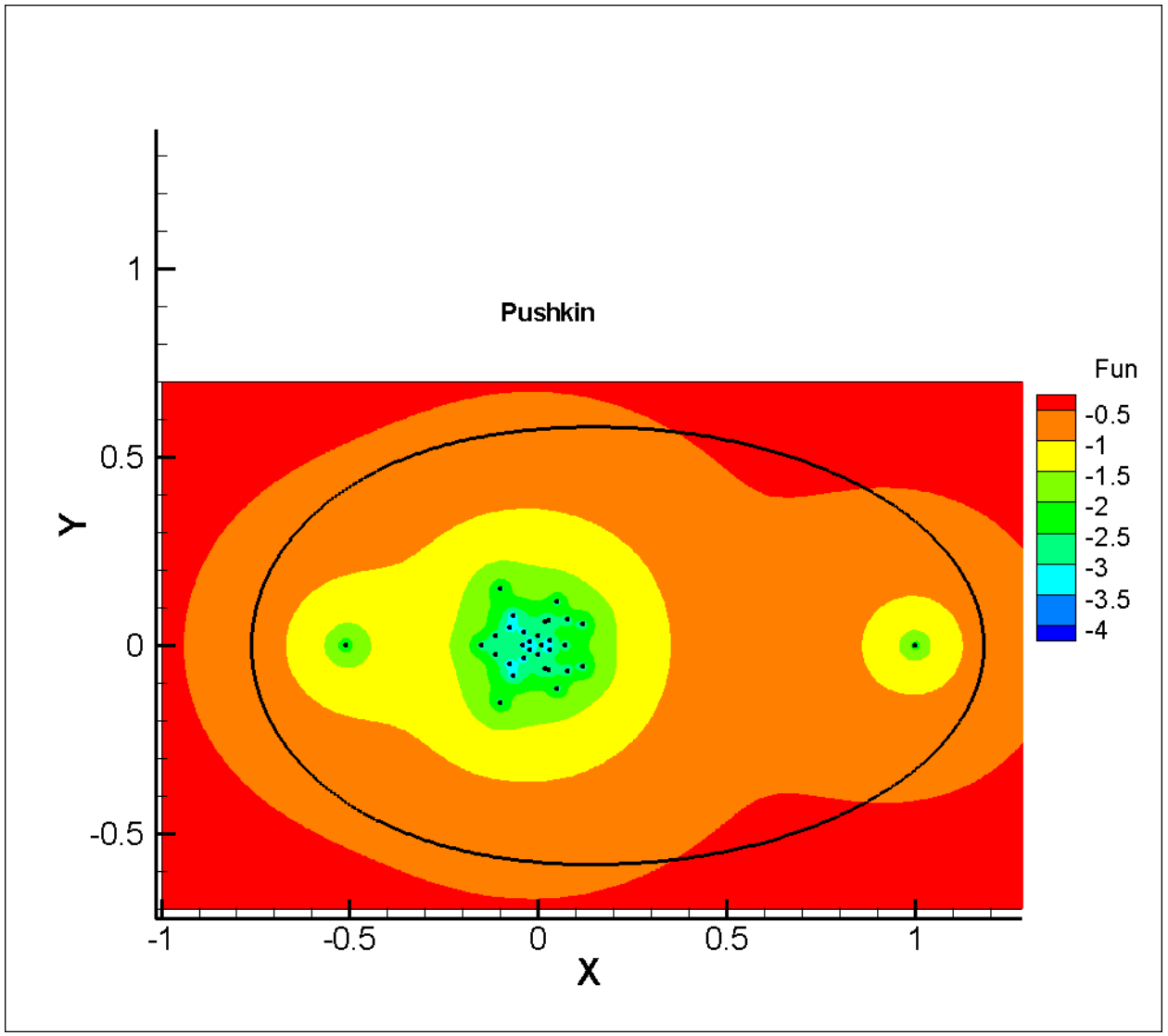
- Параметром дихотомии спектра относительно кривой $\gamma$ называется норма квадрата резольвенты на данной кривой:

$$\kappa_\gamma(P) = \frac{\|P\|^2}{2\pi r} \oint_\gamma \|R(\lambda)\|^2 d\lambda, \quad \gamma : \lambda = re^{i\varphi}$$

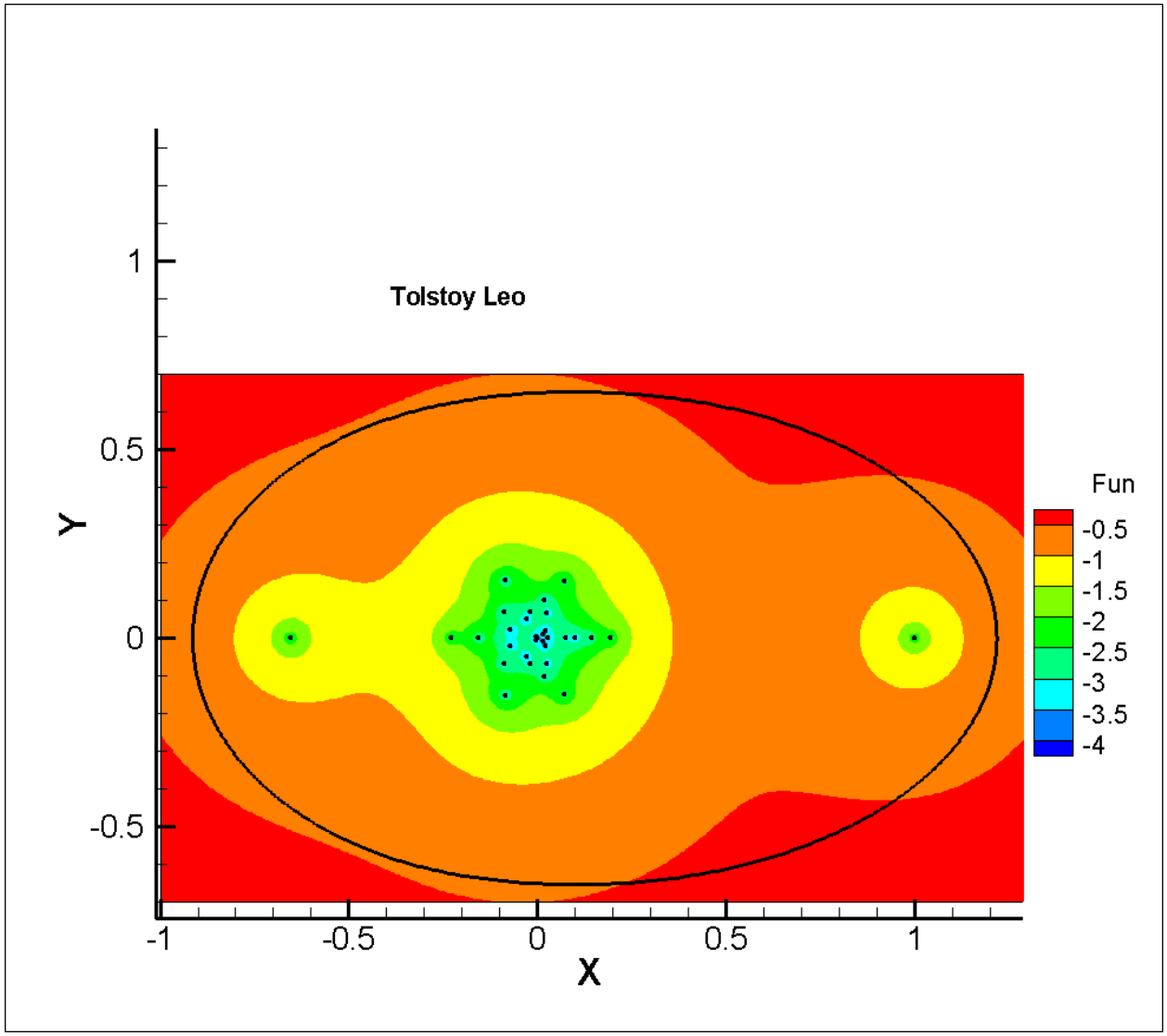
Если на кривой нет точек спектра, то норма резольвенты на этой кривой конечна.

- Спектральные портреты операторов P для разных авторов показывают устойчивость этой структуры для текстов одного автора и различающиеся картины для разных авторов.

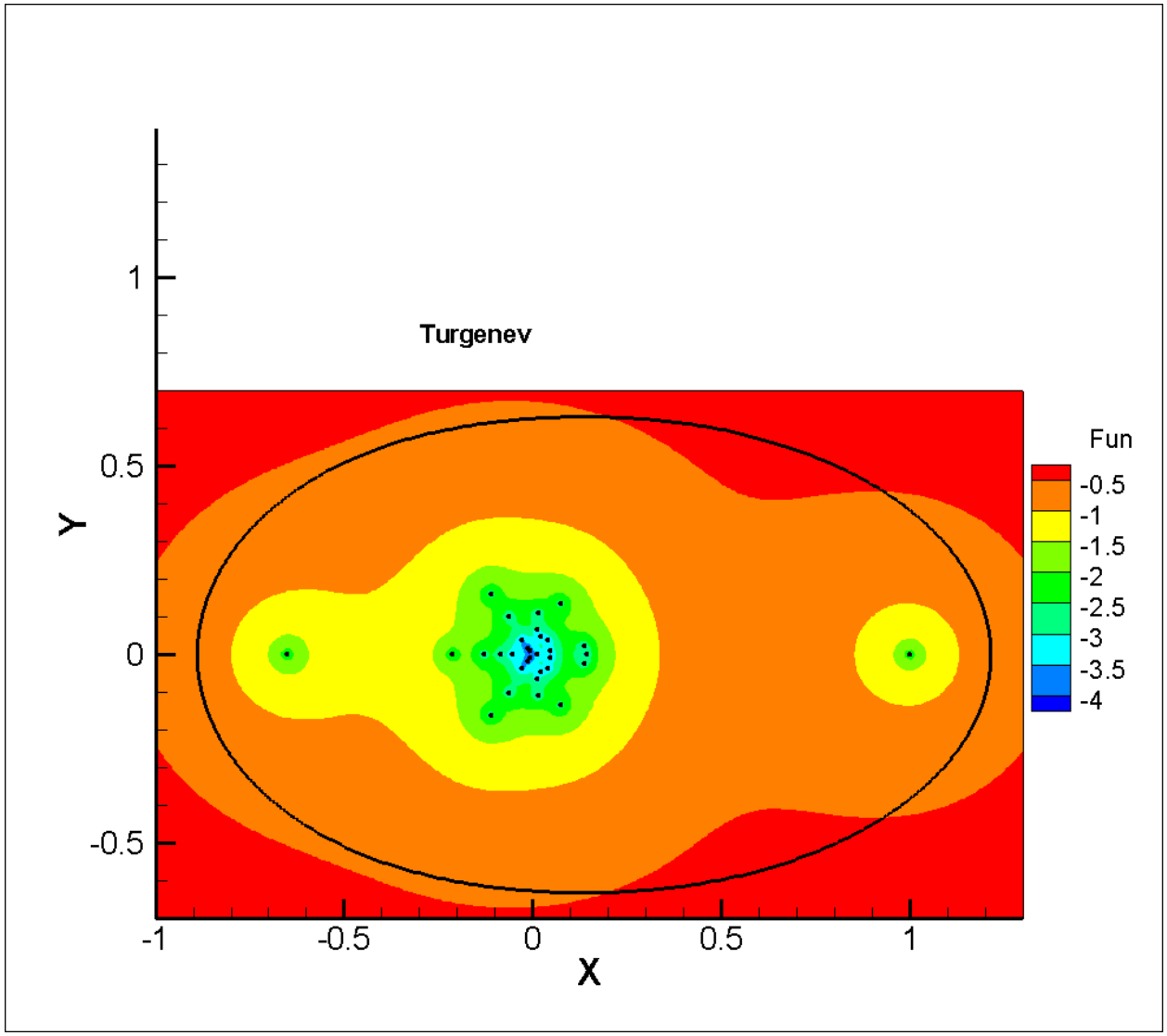
# Примеры спектральных портретов писателей



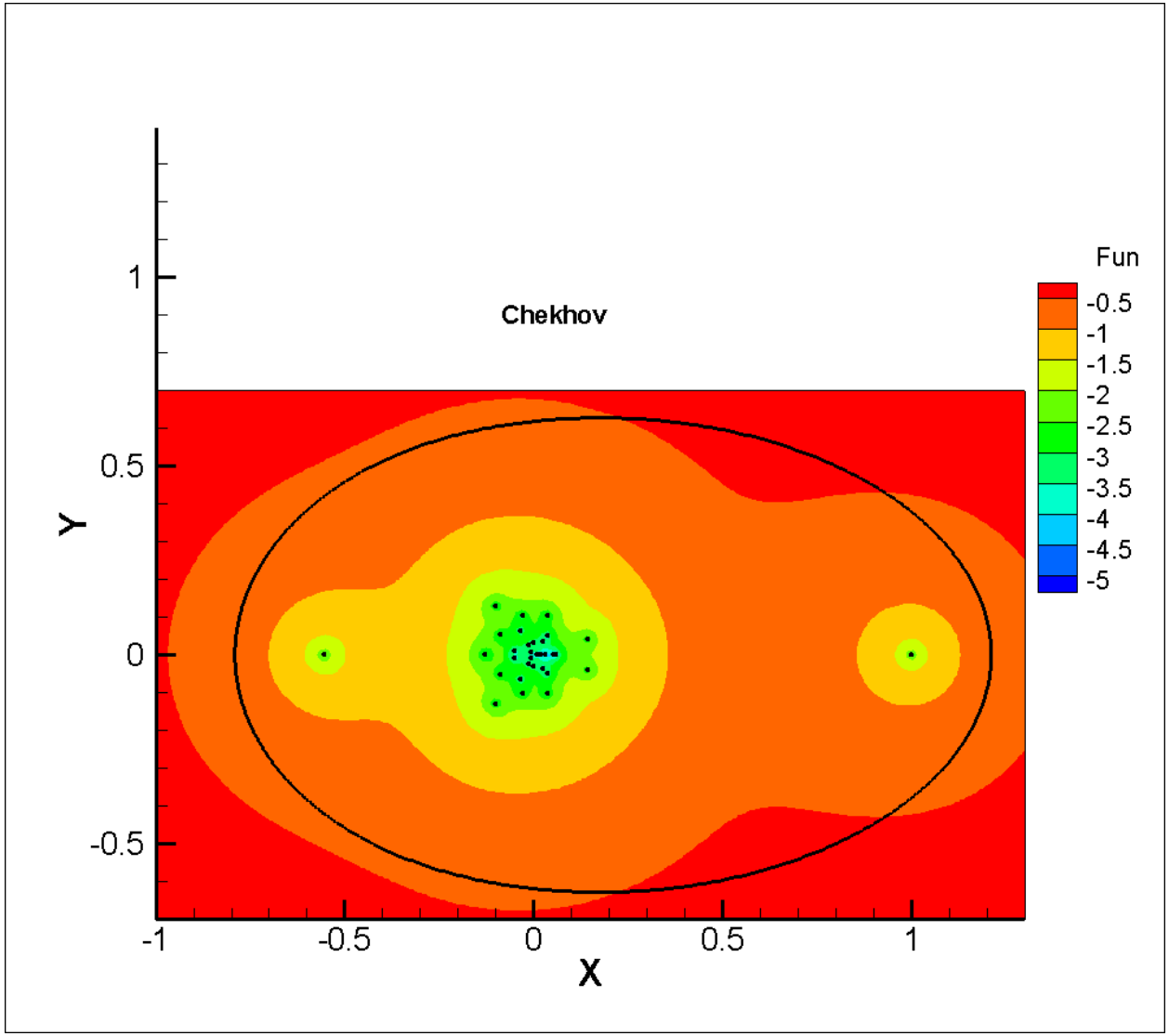
# Примеры спектральных портретов писателей



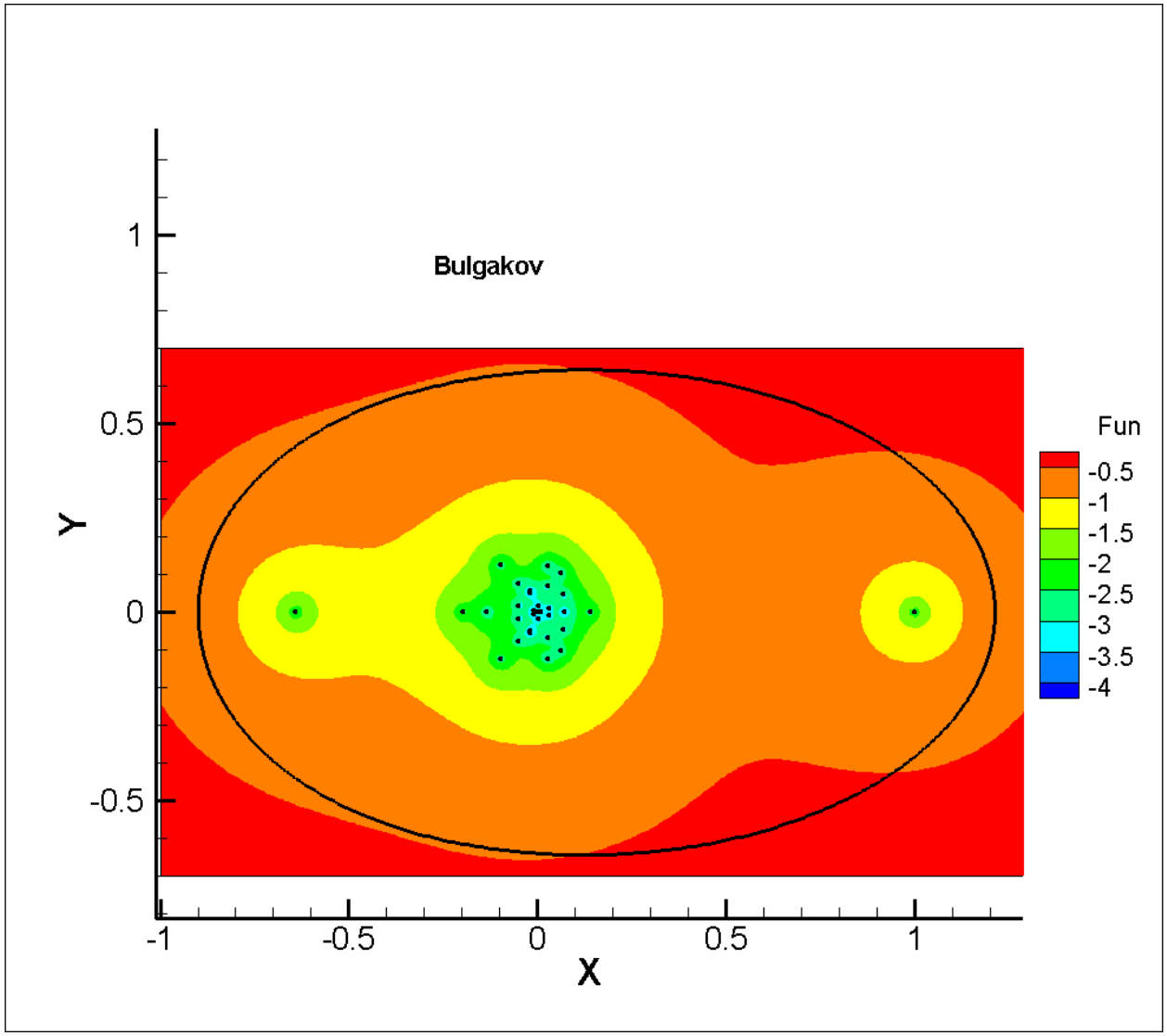
# Примеры спектральных портретов писателей



# Примеры спектральных портретов писателей



# Примеры спектральных портретов писателей



# Эффект переводчика

- Кроме с.з. $\lambda = 1$ , которому отвечает с.в. 1-ПФР $f$ , у оператора P(1) еще одно устойчивое с.з. $\mu \approx -0{,}56$ .

  Ему отвечает правый с.в. S и левый S*.

- Оказалось, что $\left(S^*, Pf\right) \approx 0$ , т.е. векторы S* и $f$ приближенно образуют главные направления оператора трансляций.

- Вектор S*, как и вектор 1-ПФР $f$ , весьма точно идентифицирует автора. Однако в переводах идентификационное свойство левого с.в. теряется. Переводчик как один и тот же писатель в переводах разных авторов не опознается. Автор же опознается без ошибок в разных переводах.

- Вывод: изложение можно отличить от сочинения, а переводчик не является соавтором.

# 4. Примеры решений спорных вопросов об авторстве

# Шолохов – автор «Тихого Дона» 3-ПФР

| Произведения Крюкова | | |
|---|---|---|
| | До Крюкова | До Шолохова |
| Булавинский бунт | 0,52 | 0,69 |
| В углу | 0,47 | 0,58 |
| Гулебщики | 0,49 | 0,56 |
| Зыбь | 0,27 | 0,38 |
| Казачка | 0,35 | 0,48 |
| Шульгинская расправа | 0,39 | 0,52 |
| **Произведения Шолохова** | | |
| Они сражались за Родину | 0,39 | 0,28 |
| Повести | 0,49 | 0,34 |
| Поднятая целина | 0,37 | 0,11 |
| Путь дорога | 0,51 | 0,40 |
| Рассказы | 0,48 | 0,34 |
| **Тихий Дон** | **0,34** | **0,26** |

# Шекспир – автор «Эдуарда III» 2-ПФР

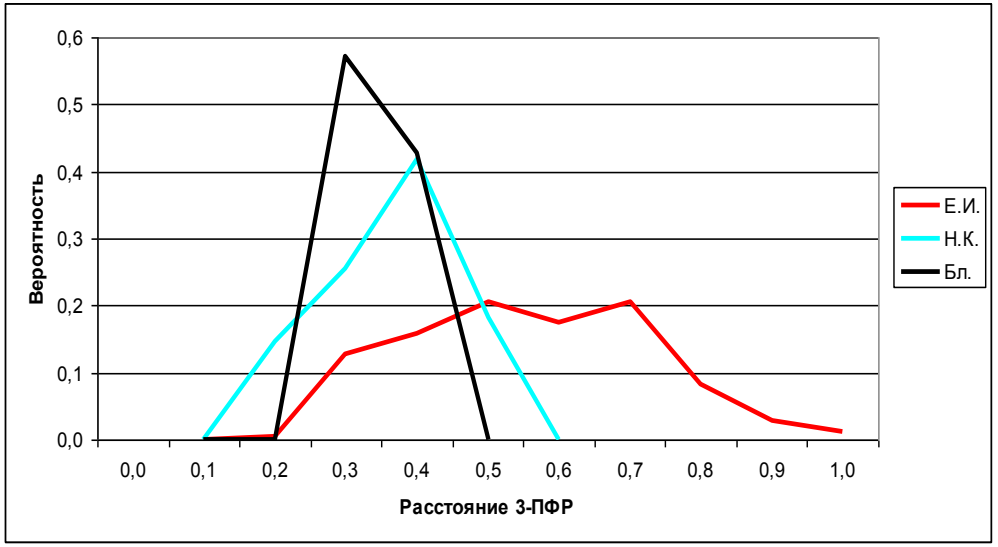
| Произведения Кристофера Марло | | |
|---|---|---|
| | До Марло | До Шекспира |
| Tamburlaine | 0,15 | 0,20 |
| The Jew of Malta | 0,16 | 0,21 |
| Edward II | 0,17 | 0,19 |
| Didona | 0,16 | 0,22 |
| The Massacre of Paris | 0,15 | 0,20 |
| **Произведения Вильяма Шекспира (Трагедии)** | | |
| Anthony and Cleopatra | 0,17 | 0,11 |
| Coriolanus | 0,19 | 0,12 |
| Hamlet | 0,15 | 0,09 |
| King Lear | 0,15 | 0,10 |
| Romeo and Juliet | 0,15 | 0,12 |
| **Edward III** | **0,15** | **0,13** |
| **Titus Andronicus** | **0,15** | **0,13** |

# Булгаков - возможный автор «Двенадцати стульев» и «Золотого теленка» (3-ПФР)

| | Булгаков | В. Катаев | Ильф/ Петров | Ильф | Петров |
|---|---|---|---|---|---|
| 12 стульев/ Золотой теленок | 0,30 | 0,31 | 0,32 | 0,40 | 0,46 |
| Мастер и Маргарита | 0,20 | 0,34 | 0,41 | 0,43 | 0,50 |
| Театральный роман | 0,25 | 0,39 | 0,43 | 0,47 | 0,51 |
| Белеет парус одинокий | 0,35 | 0,21 | 0,41 | 0,41 | 0,48 |
| Одноэтажная Америка | 0,37 | 0,39 | 0,21 | 0,40 | 0,46 |
| Светлая личность | 0,43 | 0,46 | 0,44 | 0,18 | 0,56 |
| Фронтовые записки | 0,43 | 0,41 | 0,39 | 0,51 | 0,28 |

# «Чаша Востока» Е.И. Рерих (3-ПФР): автор не «Махатма Мория», а Е.П. Блаватская

| | Эталон Е.И. | Эталон Н.К. | Эталон Бл. |
|---|---|---|---|
| **Тексты Е.И. (Агни-Йога)** | **0,43** | 0,53 | 0,57 |
| **Тексты Н.К.** | 0,55 | **0,41** | 0,52 |
| **Тексты Бл.** | 0,53 | 0,45 | **0,32** |
| **Чаша Востока** | **0,42** | **0,44** | **0,31** |




- 1. Е.И. не сама писала все тексты, ей приписываемые, но «Махатма Мория» в них не участвовал, если допустить, что он имел отношение и к текстам Блаватской;
- 2. Анонимное произведение «Чаша Востока», переведенное Е.И. с английского, наиболее вероятно, написано Блаватской и переведено на английский ее учениками;
- 3. Н.К. не был соавтором Е.И. и писал свои тексты самостоятельно.

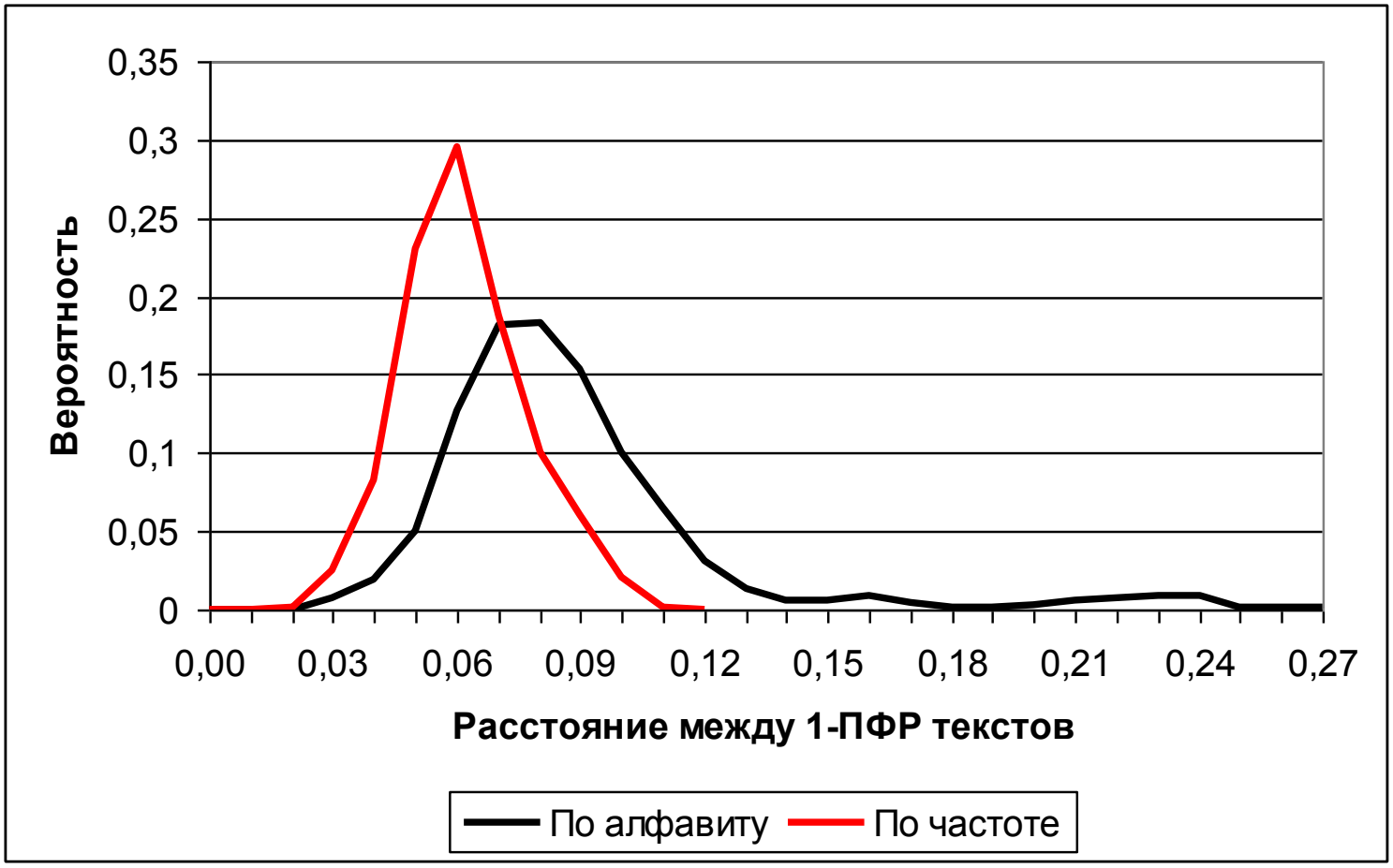
# 5. Упорядоченность букв в текстах на европейских языках и анализ Манускрипта Войнича

# Voynich Manuscript

Гипотезы:

- VM закодирован шифром прямого соответствия;
- VM содержит ложные пробелы между словами;
- существует рукопись-ключ, трафареты страниц которой дают расшифровку;
- одни и те же символы в разных местах текста отвечают разным буквам;
- VM написан с пропуском гласных букв;
- VM написан на латыни (немецком, финском, вьетнамском, индейском и т.д. языке);
- VM – это мистификация, т.е. просто бессмысленный набор знаков, проданный как магическая книга.

# Расстояния между русскими текстами при различном упорядочении



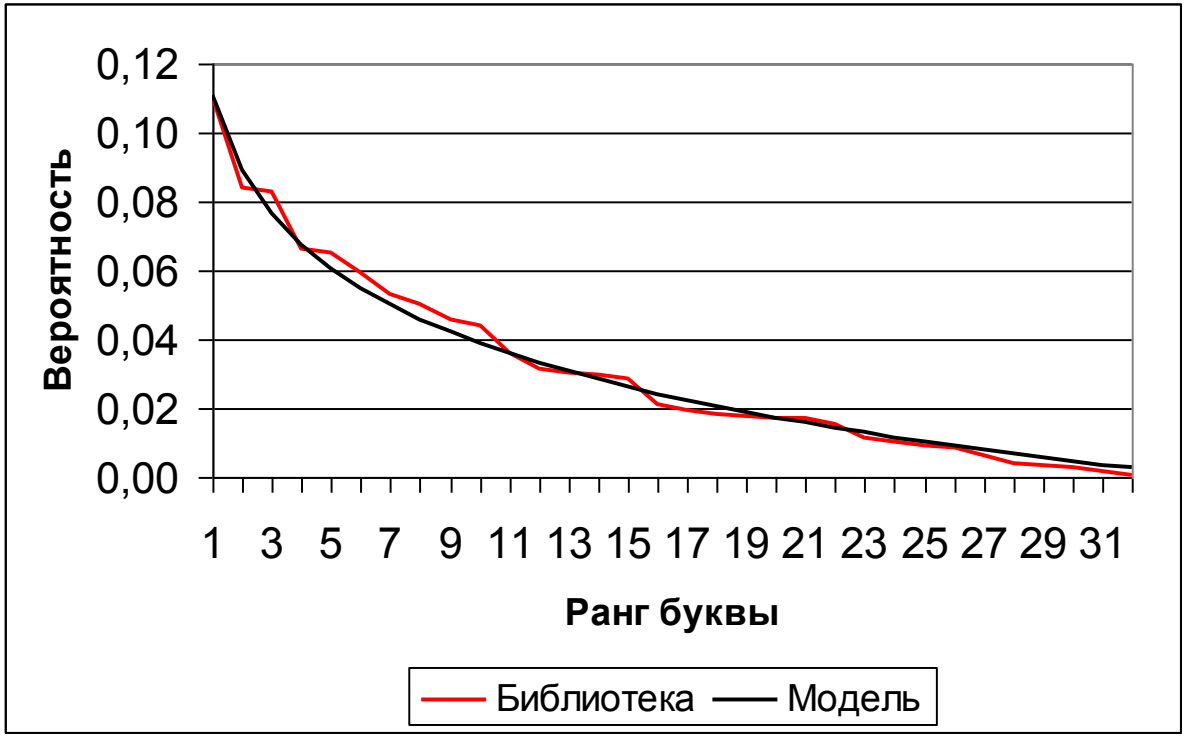

- Частотная упорядоченность стабильна и выражает свойство кодирования информации в том или ином языке

# Распределение букв по частоте в текстах на русском языке

Наилучшая детерминация 0,98 логарифмической аппроксимации получается при $o=0$ в модели:

$$f(k)=\frac{1}{n}\left(1+\frac{1}{n+o}\ln\frac{n!}{k^n}\right),$$

$$o=const$$

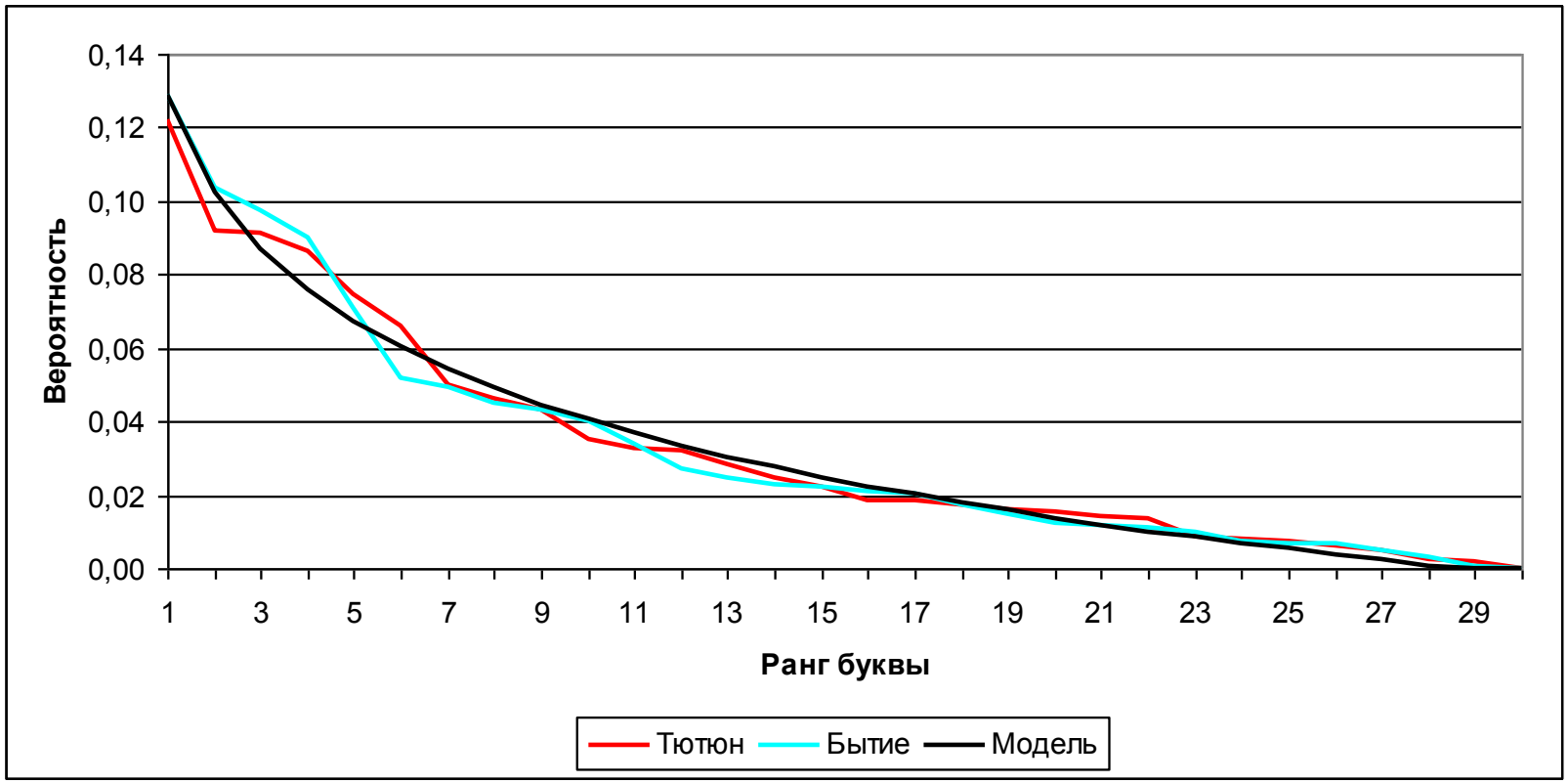



- Эта зависимость выполнена и для старославянских текстов (n=43), и для русской литературы XIX века (n=37). Для русских текстов в транслите (n=23 символа) $o=+9$.

# Распределение букв по частоте в текстах на болгарском языке

$$f(k) = \frac{1}{n}\left(1 + \frac{1}{n+o}\ln\frac{n!}{k^n}\right),$$

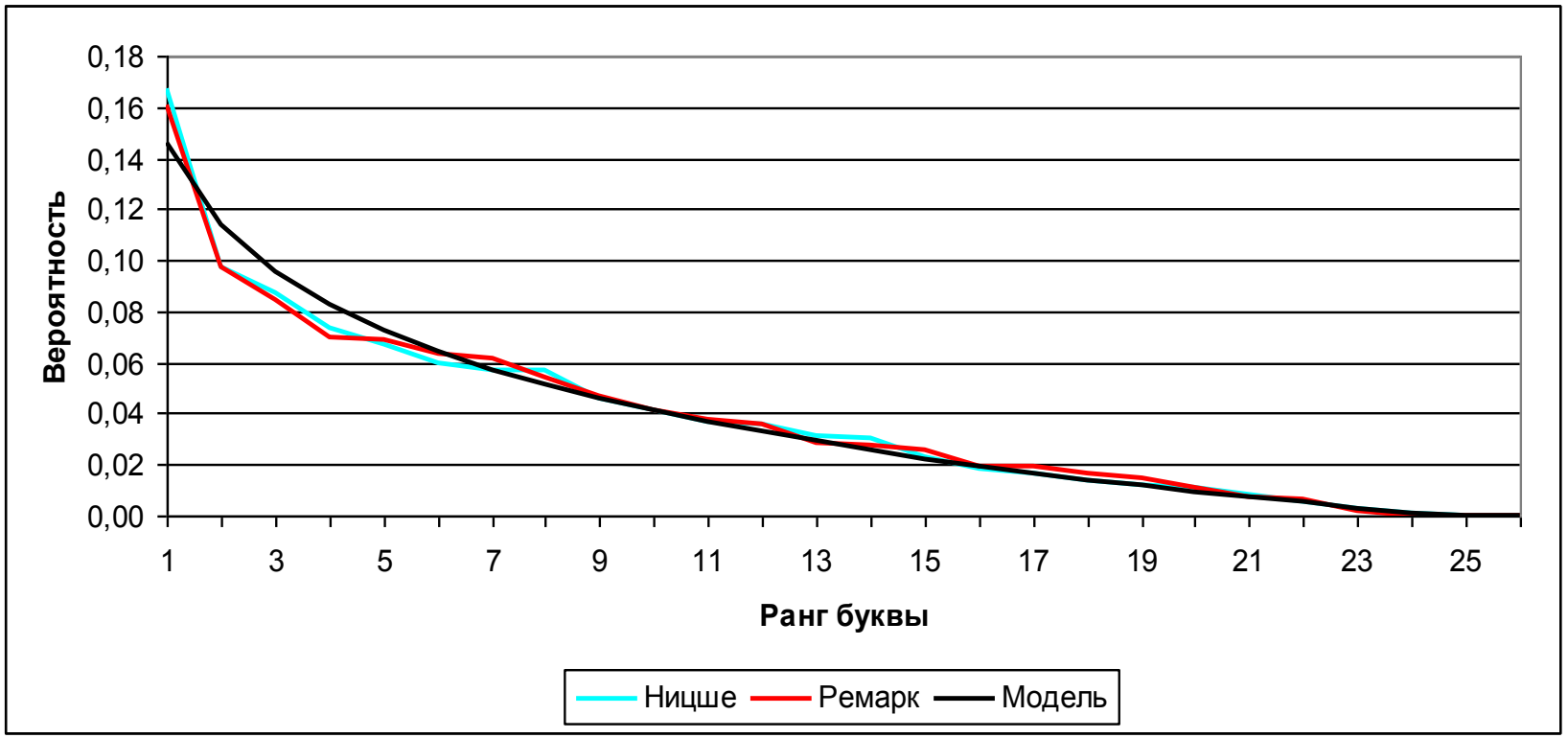
$$n = 30, \quad o = -4, \quad n_{fact} = 26$$



# Распределение букв по частоте в текстах на немецком языке

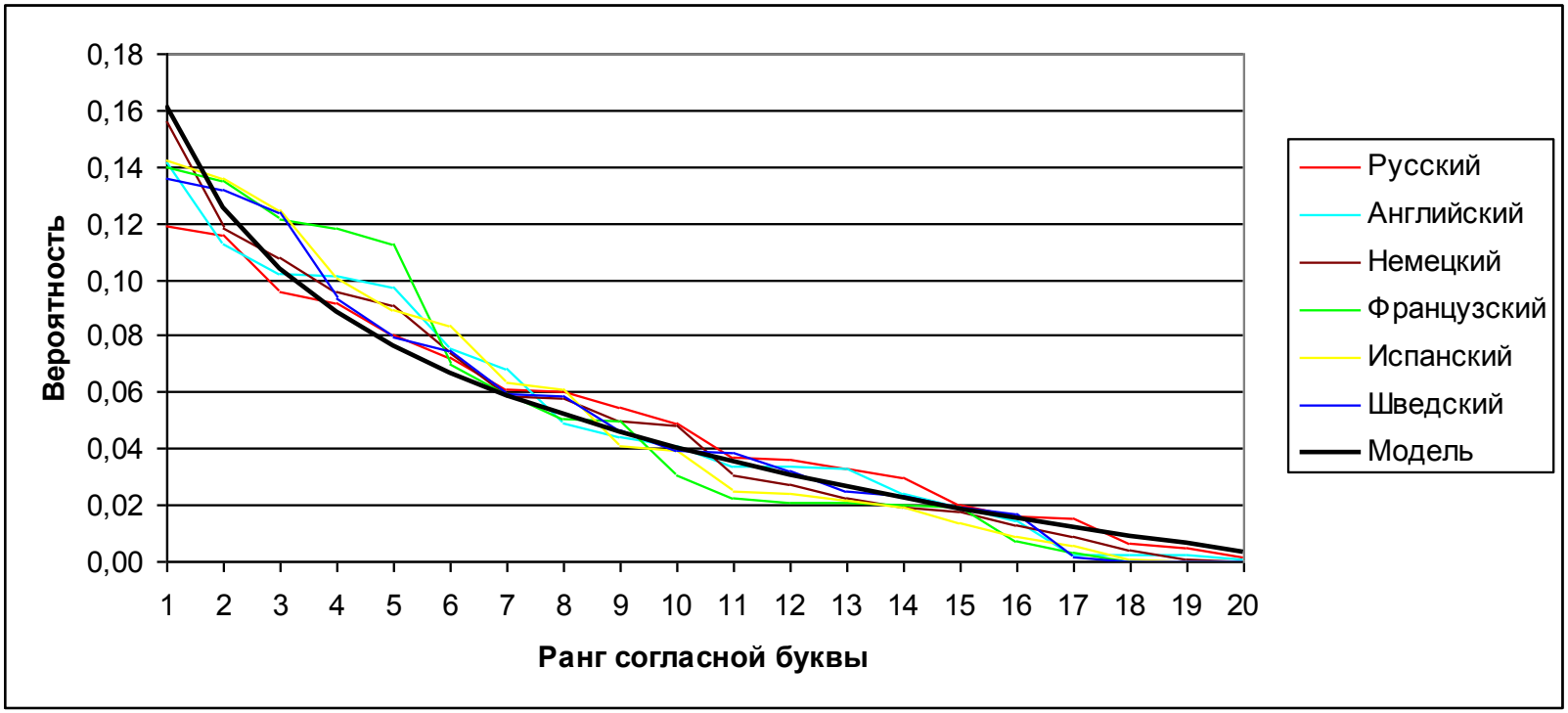
$$f(k) = \frac{1}{n}\left(1 + \frac{1}{n+o}\ln\frac{n!}{k^n}\right),$$

$$n = 26, \quad o = +4, \quad n_{fact} = 30$$

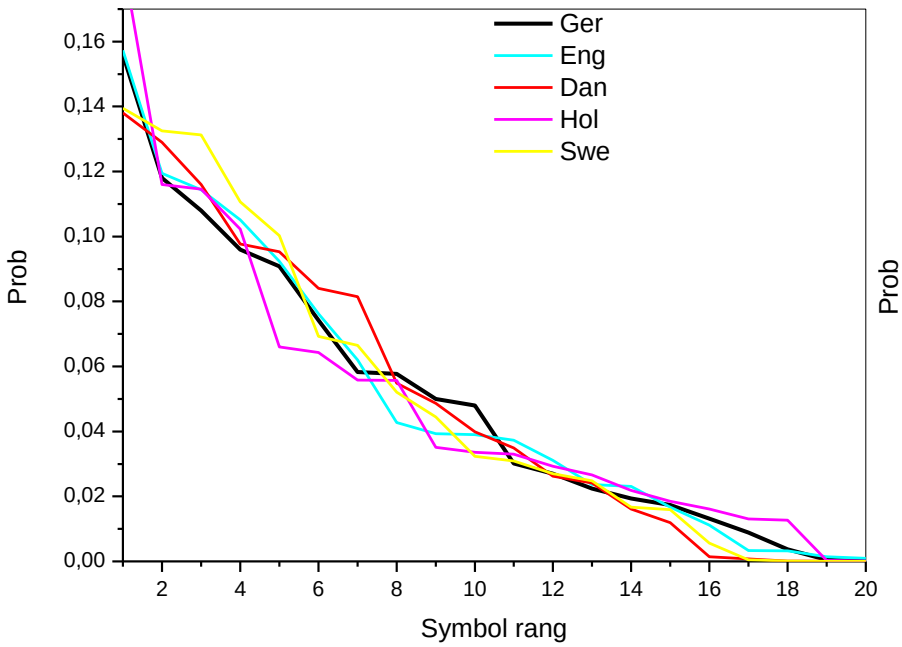


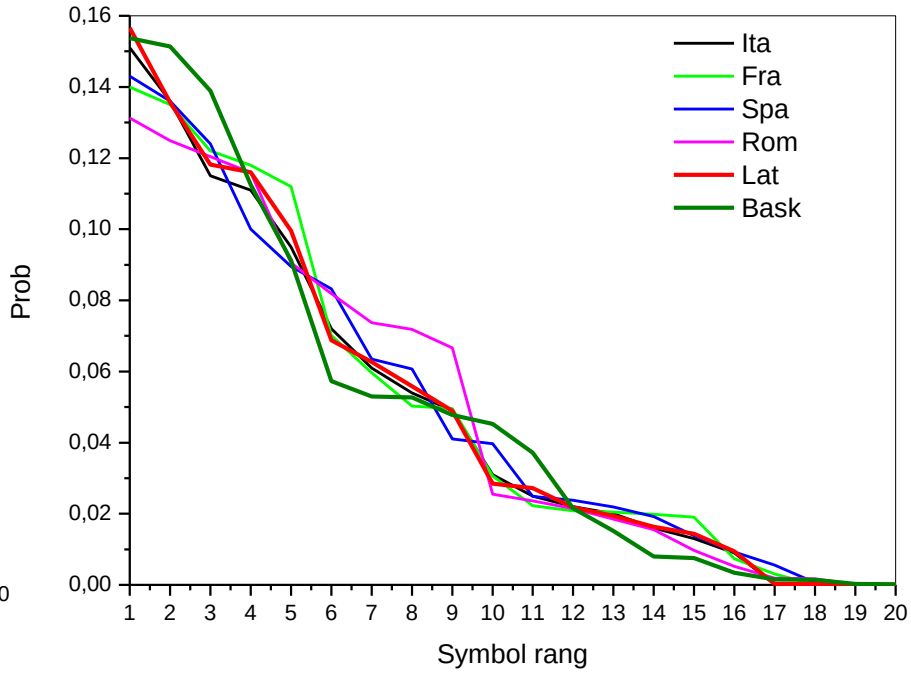
# Языки индоевропейской семьи имеют общий согласный остов

- Параметр $o$ трактуем как оценку избыточности ($o<0$) или недостаточности ($o>0$) алфавита по отношению к звуковому ряду. В текстах без огласовки на языках индоевропейской семьи $n$ = 20, $o$ = 0. Детерминация лежит в пределах 0,96-0,98.



# Распределение упорядоченных частот в текстах без огласовки



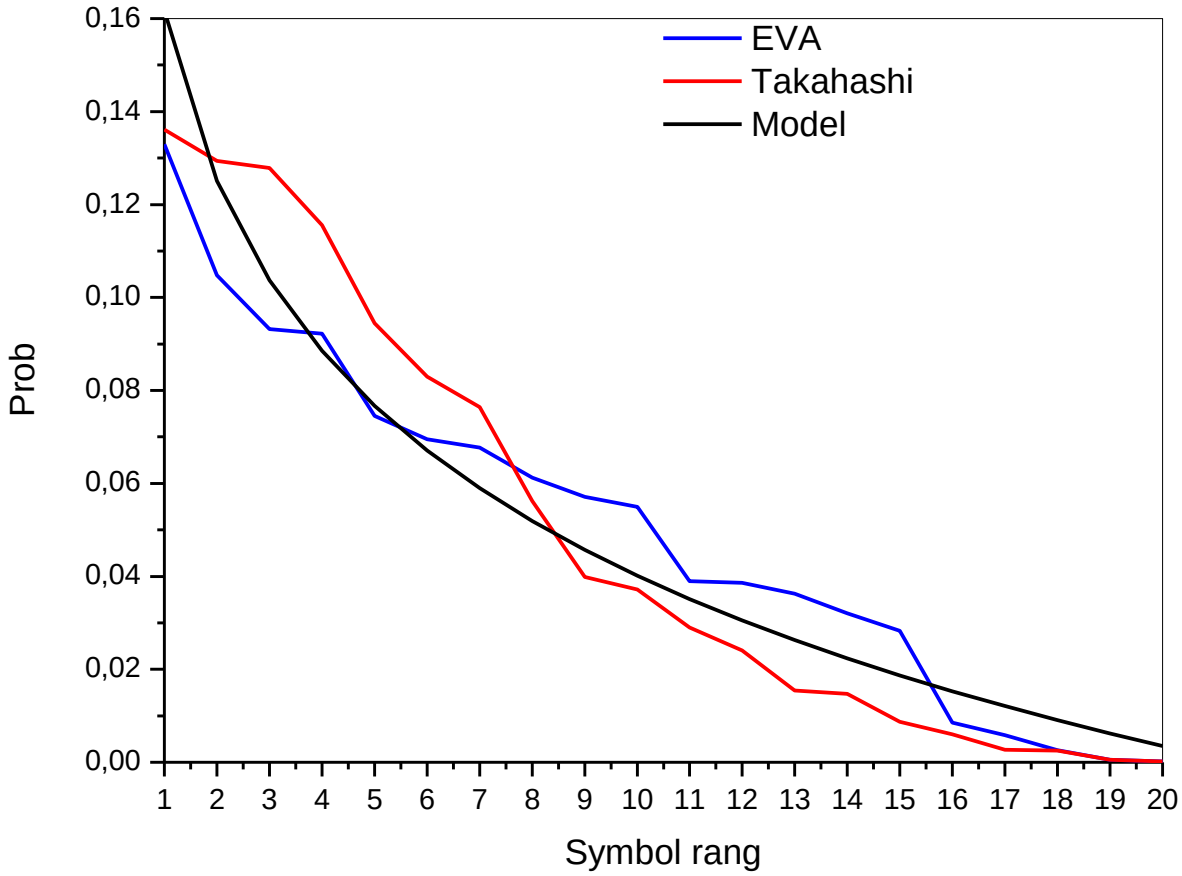

- Языки кластеризуются по близости распределений в норме L1 в соответствии с лингвистическими группами

# Группировка европейских языков

| | ger | eng | hol | dan | swe | nor | lat | ita | spa | fra | rom |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ger | | 8 | 11 | 13 | 11 | 12 | 12 | 13 | 11 | 15 | 19 |
| eng | | | 12 | 13 | 12 | 13 | 12 | 13 | 11 | 15 | 19 |
| hol | | | | 10 | 11 | 11 | 19 | 21 | 19 | 22 | 27 |
| dan | | | | | 11 | 10 | 13 | 13 | 9 | 14 | 13 |
| swe | | | | | | 11 | 15 | 15 | 10 | 14 | 18 |
| nor | | | | | | | 13 | 13 | 10 | 15 | 14 |
| lat | | | | | | | | 5 | 10 | 7 | 12 |
| ita | | | | | | | | | 10 | 7 | 12 |
| spa | | | | | | | | | | 11 | 13 |
| fra | | | | | | | | | | | 13 |
| rom | | | | | | | | | | | |

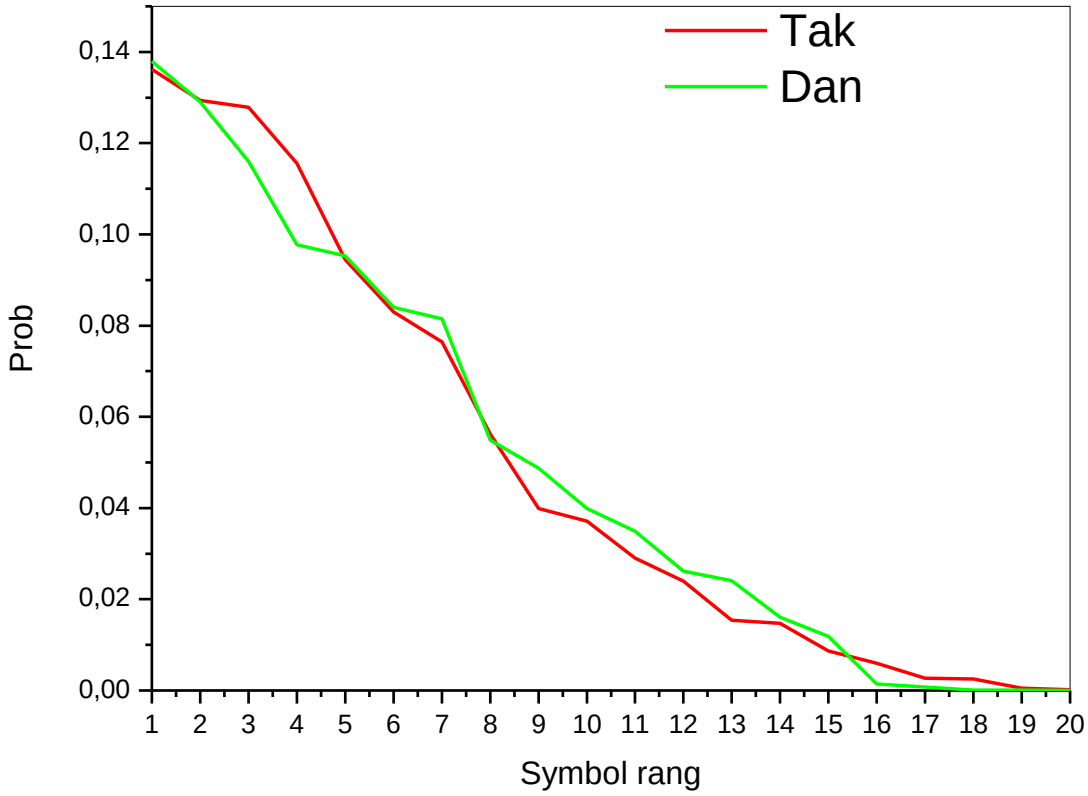
- Приведены расстояния между распределениями упорядоченных частот в текстах на разных языках в одном алфавите. Попарная кластеризация на уровне 13 (%).

# Распределение упорядоченных частот для двух транскрипций VM




- Отклонения текстов почти на всех языках от модели 0,08-0,11; только для датского, сербского и хорватского оно равно 0,17. Отклонение транскрипций VM от модели равно 0,17.

# VM написан на датском языке без огласовки шифром прямого соответствия?




- Отклонение распределений транскрипции VM и датского языка равно 0,10. Отклонение от латыни равно 0,11, но латынь отклоняется от модели на 0,13, а не на 0,17, как VM, т.е. менее подходит на роль языка рукописи.

# Показатель Херста как индикатор «качества» шума

$$\bar{x}(t,k) = \frac{1}{k}\sum_{i=t-k+1}^{t} x(i) \qquad \sigma_x^2(t,k) = \frac{1}{k}\sum_{i=t-k+1}^{t}\left(x(i)-\bar{x}(t,k)\right)^2$$

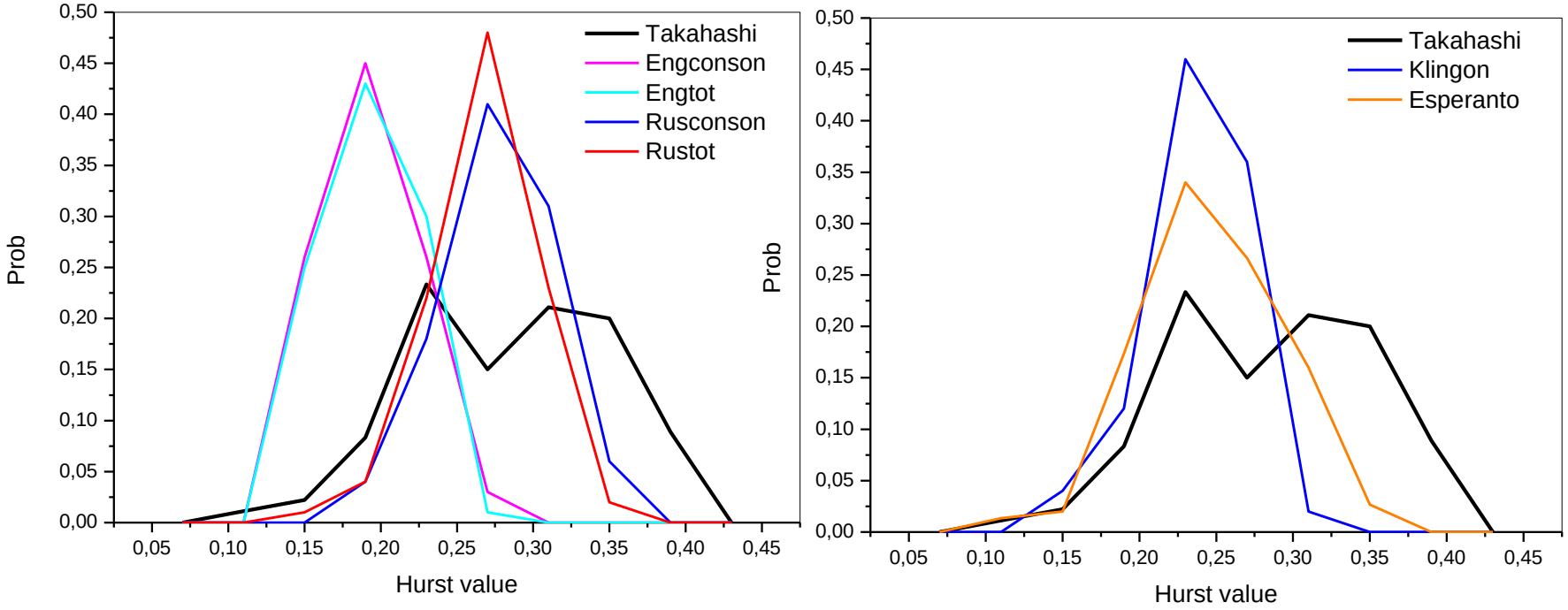
$$R(t,k) = \max_{j\le t}\left(\sum_{i=t-k+1}^{j}\left(x(i)-\bar{x}(t,k)\right)\right) - \min_{j\le t}\left(\sum_{i=t-k+1}^{j}\left(x(i)-\bar{x}(t,k)\right)\right)$$

$$\xi(t,k) = \ln\left(\frac{R(t,k)}{\sigma_x(t,k)}\right), \quad \bar{\xi}_N(t) = \frac{1}{N}\sum_{k=1}^{N}\xi(t,k)$$

$$H_N(t) = \frac{1}{N}\sum_{k=1}^{N}\left(\xi(t,k)-\bar{\xi}_N(t)\right)\left(1+\ln(k/N)\right)$$

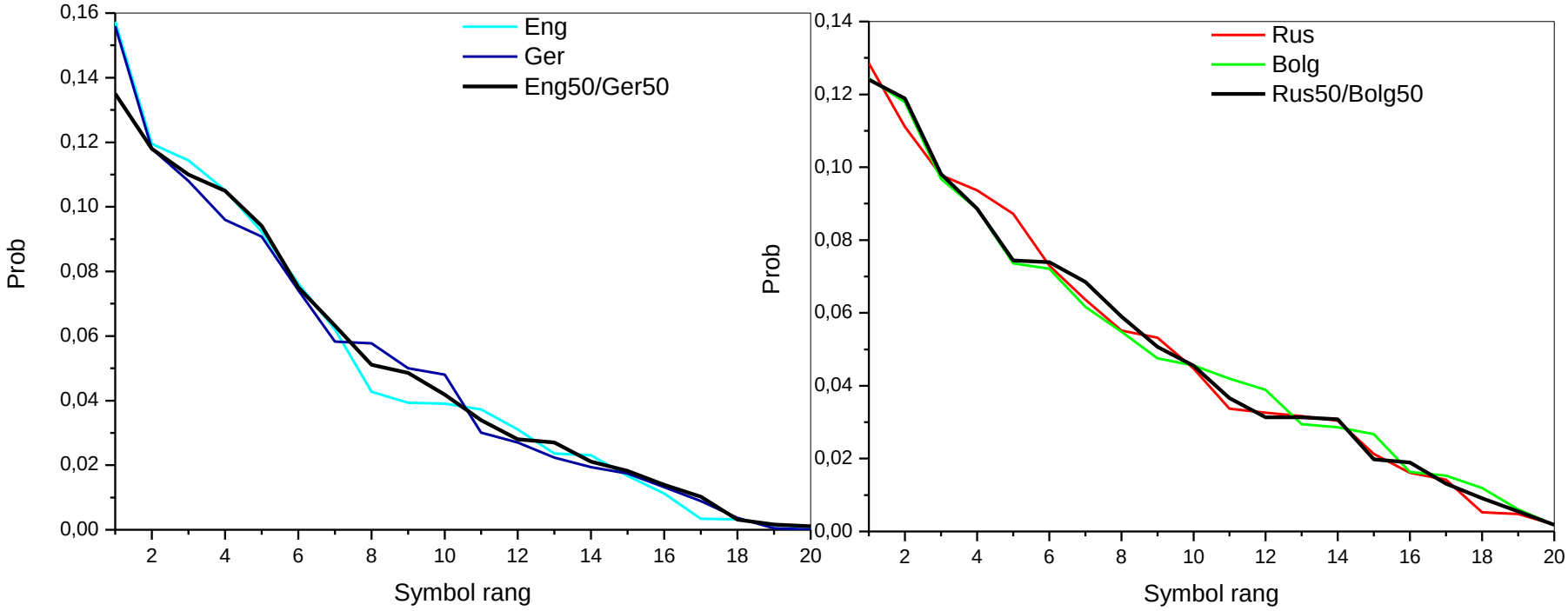
- Если H=1/2, то шум белый; если H<1/2, то ряд не сохраняет тренд; если H>1/2, тренд преимущественно сохраняется.

# Расстояния между одинаковыми символами образуют антиперсистентный ряд

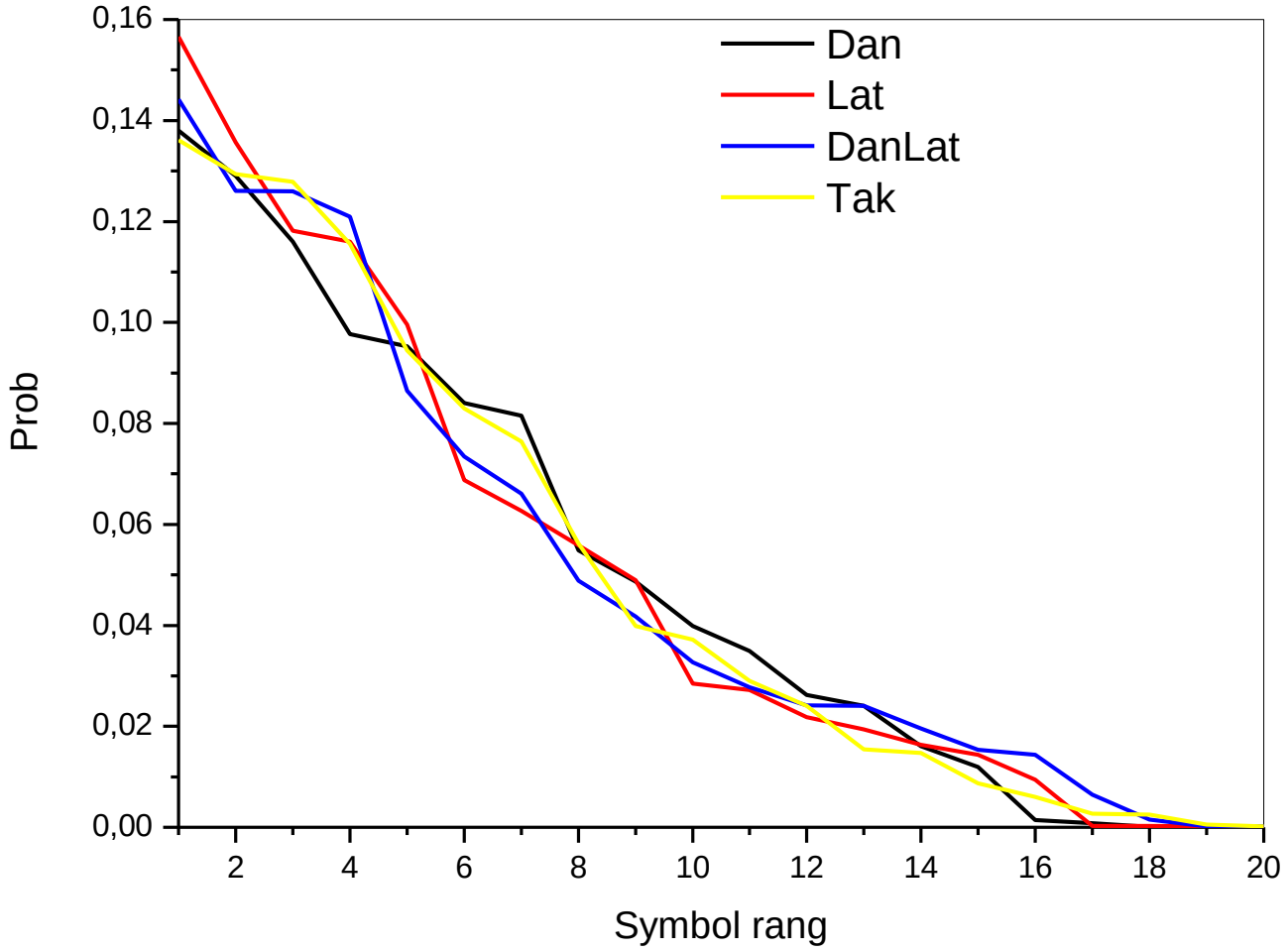



- Как естественные языки, так и искусственные (эльфийский quenya, инопланетный klingon, esperanto, volapuk и др.) имеют узко-унимодальные распределения показателя Херста по выборкам длин 10 тыс. знаков. Текст VM имеет более широкое распределение, что характерно для смешанных языков.

# Смесь языков из одной группы не меняет упорядоченного распределения частот



# Смесь языков из разных групп больше подходит на роль языка VM

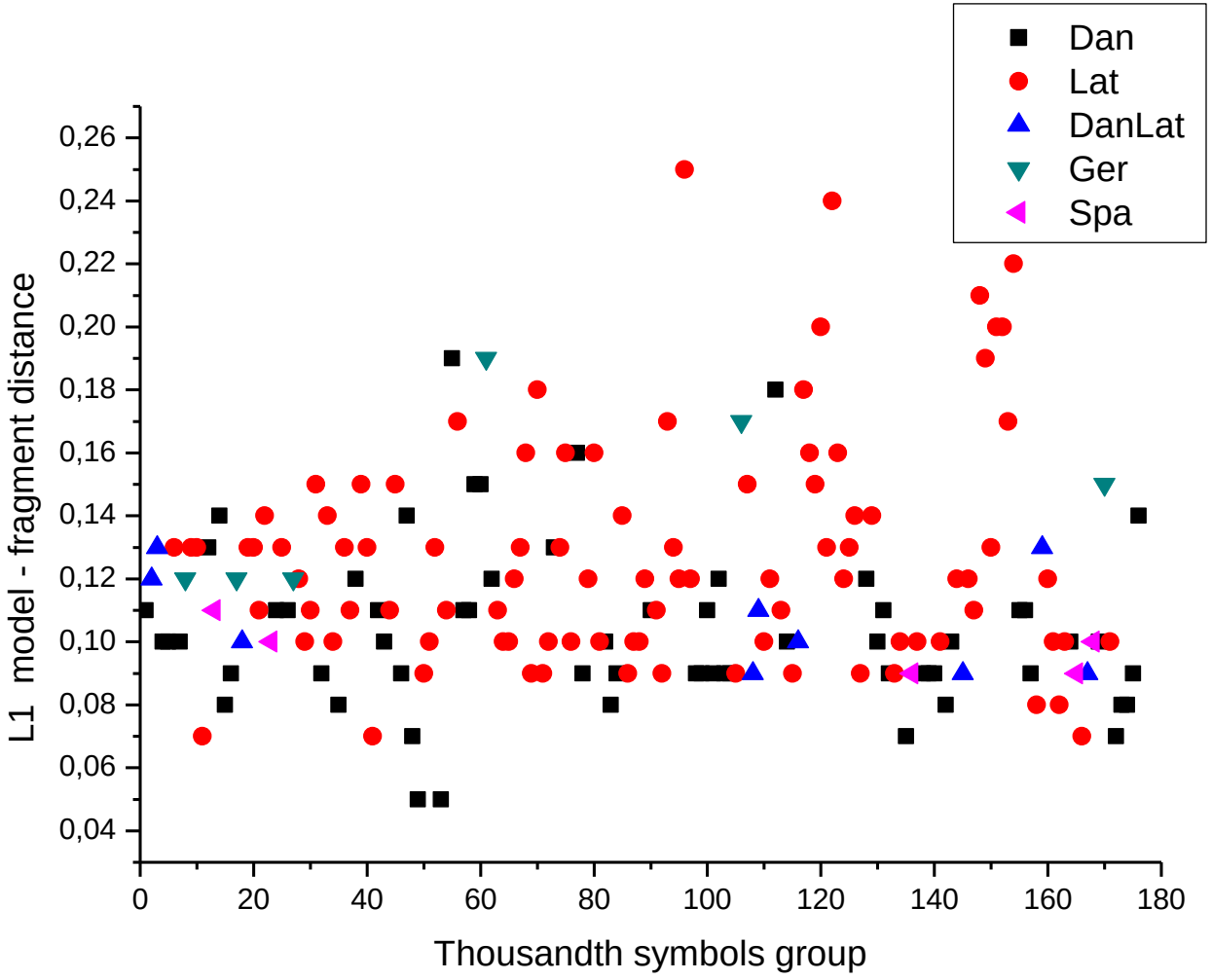



VM в целом написан на смеси латыни и датского языка в пропорции 70/30.

При написании были удалены гласные, слова были записаны без пробелов, но с ложными пробелами.

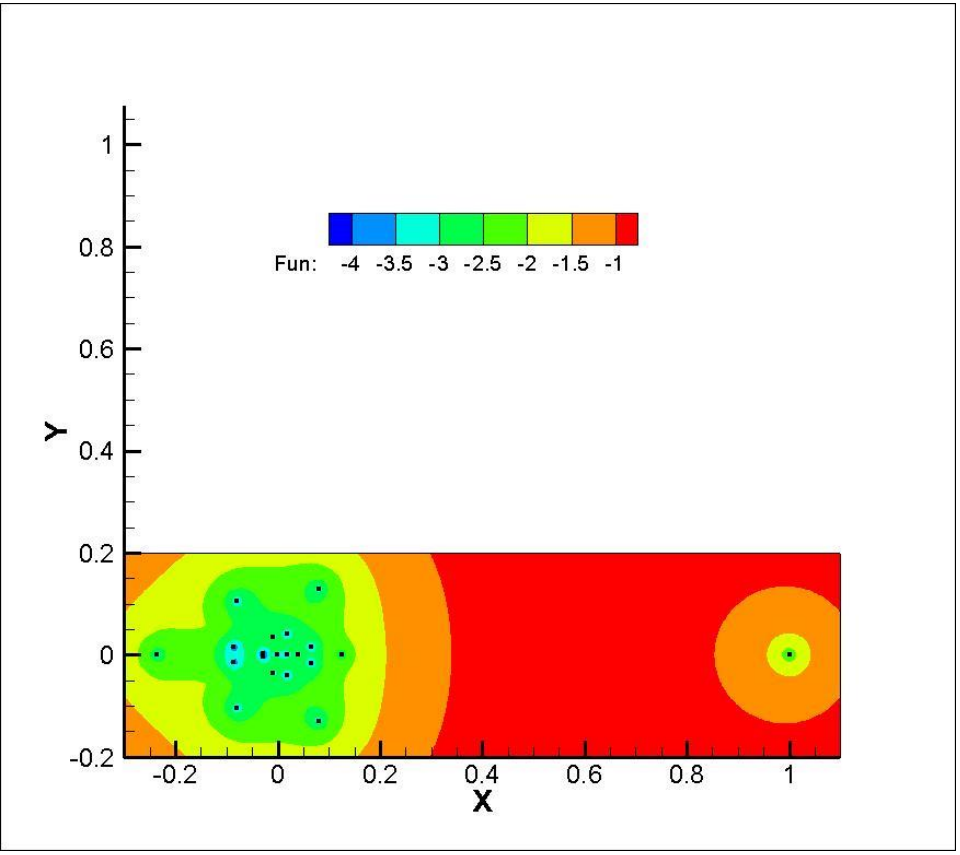
Затем применен прямой шифр.

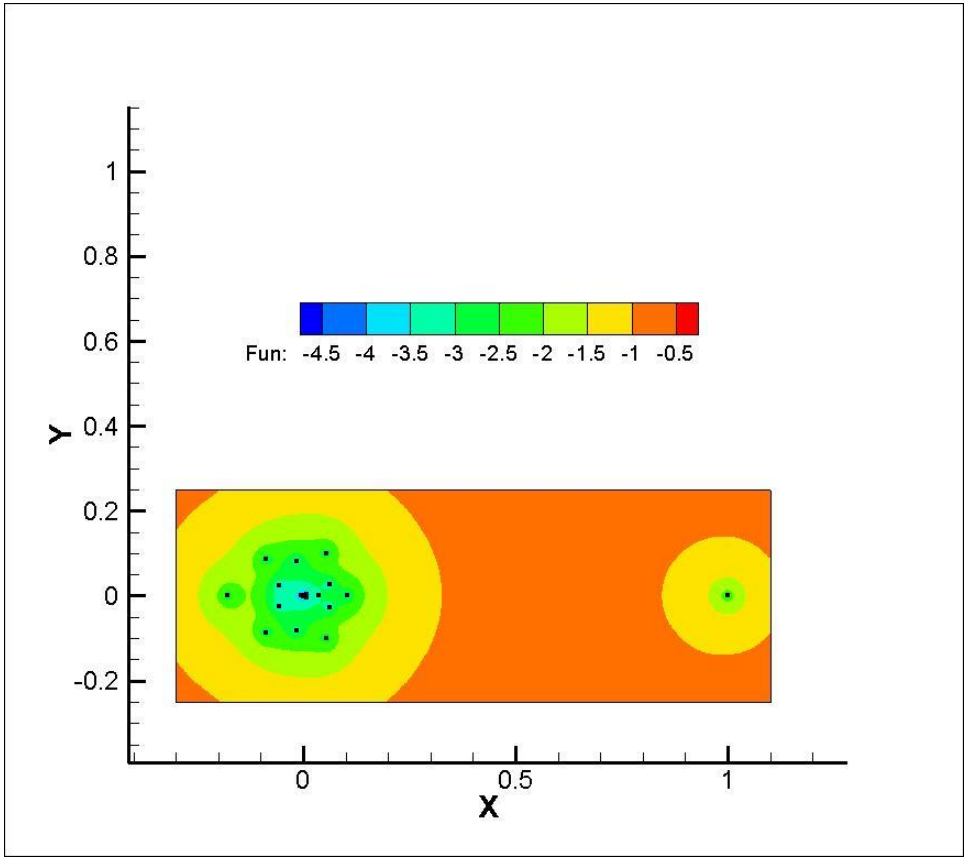
# Идентификация языка VM в окне 1000 символов (1 страница)




Примерно 85 % текста идентифицируется на уровне 15%-ой близости к эталону, что является характерной ошибкой для сравнения 1-ПФР. Листы VM, которые отклоняются от ближайшего языкового эталона более чем на 0,16, распознаны недостоверно.

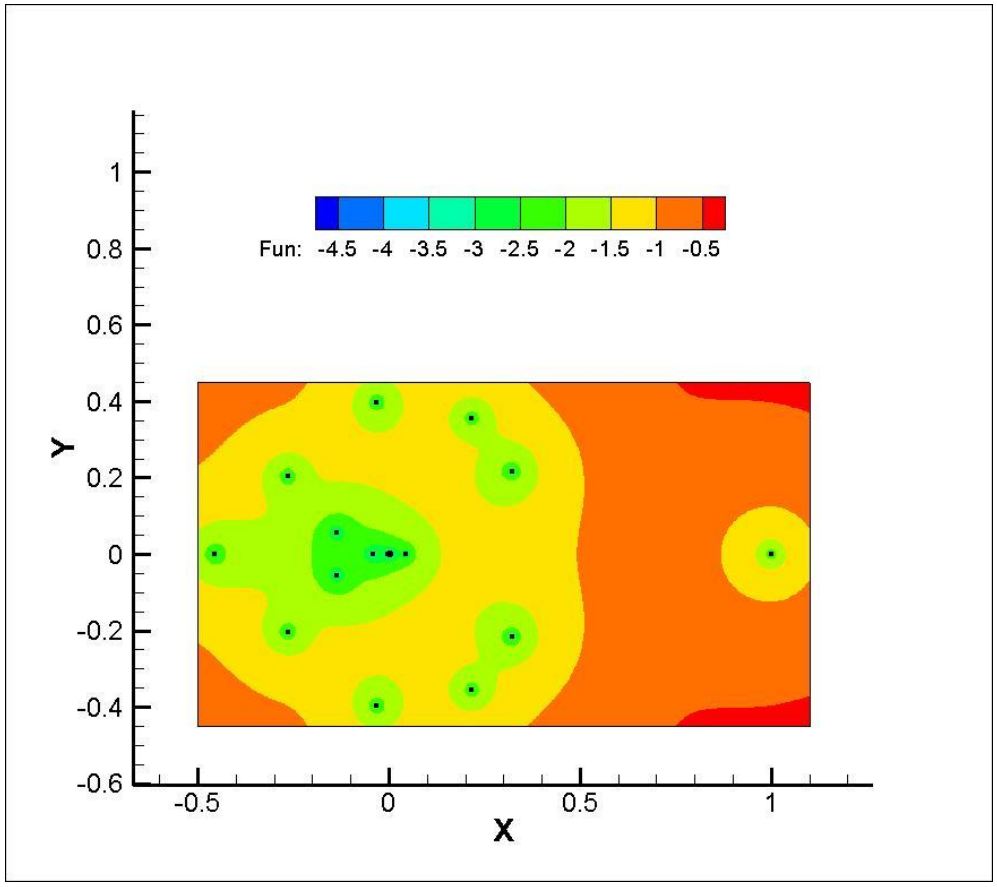
# Спектральные портреты текстов без огласовки




- Английский текст (слева) и латынь (справа) имеют круговую (зеленую) область расположения спектра радиусом примерно 0,2.

# Спектральный портрет VM отвечает смеси по крайней мере двух языков




- Круговая область расположения спектра имеет радиус примерно 0,4, что вдвое больше, чем для однородного текста.

# Выводы

- Использование функций распределения эффективно в задачах класса Big Data, поскольку позволяет сократить описание, разработать статистические индикаторы разладки высокой точности для машинного обучения распознавания образов, а также провести кластеризацию статистических объектов большой размерности в объекты (кластеры) малой размерности

- Применительно к литературным текстам кинетический метод позволил решить такие задачи, как определение атрибутов текста (автор, жанр, язык), и расширить область применения тонких методов численного анализа таких, как спектральные портреты

- Представленные методы анализа могут быть применимы к различным временным рядам, в том числе и нестационарным

# СПАСИБО ЗА ВНИМАНИЕ