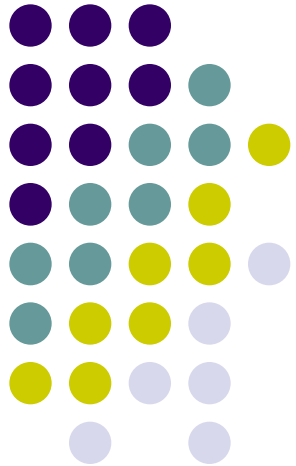
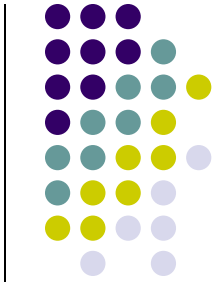
# Ю.Н. Орлов

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


# Методы статистического анализа литературных текстов



# Обсуждаемые вопросы

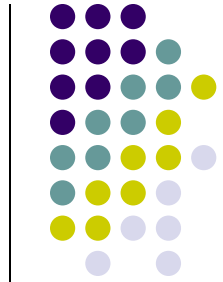


- Точность статистического анализа в зависимости от объема текста.

- Идентификация автора неизвестного текста в библиотеке эталонов, создание эталонов и кластеризация текстов.

- Оператор трансляций распределения текста по буквам и спектральные портреты. Эффект переводчика.

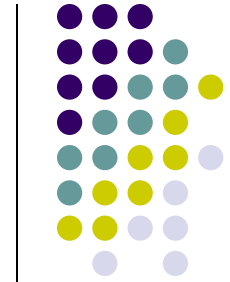
# Обсуждаемые вопросы



- Анализ авторских тандемов и проверка текста на однородность. Динамические системы, генерирующие ряд расстояний между одинаковыми буквами в тексте.

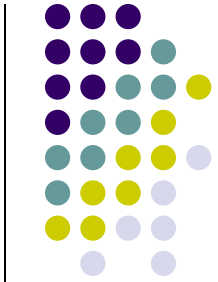
- Анализ европейских языков. Функция распределения букв по частоте встречаемости. Фонетический анализ алфавитов по избыточности или недостаточности символов.

# Цель и программа работы



- Сопоставление тексту структуры в фазовом пространстве (букв, слов и т.п.)
- Введение нормы как расстояния между структурами в фазовом пространстве
- Определение проекторов на подпространства с целью классификации: языка текста, эпохи написания, типа (проза или поэзия), формата (роман, очерк, эссе), жанра (детектив, триллер), автора

# Текстовый инвариант? – Нет!



- Текстовый инвариант – это функционал F(T) от текстовой структуры. Два текста близки в фазовом пространстве, если близки функционалы: $\left|F(T_1)-F(T_2)\right|<\varepsilon$.

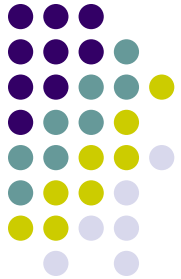
- Цель работы – ввести наилучшим образом расстояние в пространстве структур

$$\rho_{1,2}=\left\|T_1-T_2\right\|,$$

т.е. рассматривать не разность функционалов, а функционал разности.

# 1. Статистическая достоверность определения частот употребления букв в литературных произведениях

# Выборочное распределение текста по буквам

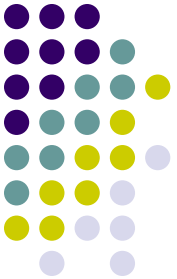


- Пусть $\xi$ – случайная величина (буква или буквосочетание), принимающая значения из конечного упорядоченного множества букв (пар букв, и т.д.) в алфавите.

- 1-ВПФР $f_{1N}(i)$ есть эмпирическая вероятность обнаружения данной ($i$-ой) буквы в тексте из N символов, 2-ВПФР $f_{2N}(i,j)$ – пары букв, и т.д.

- «Время» – это порядковый номер буквы в тексте. Пробелы и знаки игнорируются.

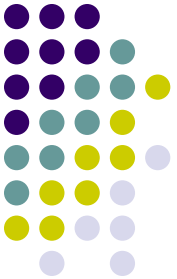
# Минимально достаточная длина текста



- Тексты должны быть таких длин, чтобы статистическая неопределенность в оценке вероятностей буквосочетаний была «много меньше» наблюденного расстояния между распределениями

- **Основная гипотеза: распределения буквосочетаний для каждого автора квазистационарны**

# Оценка достаточной длины текста



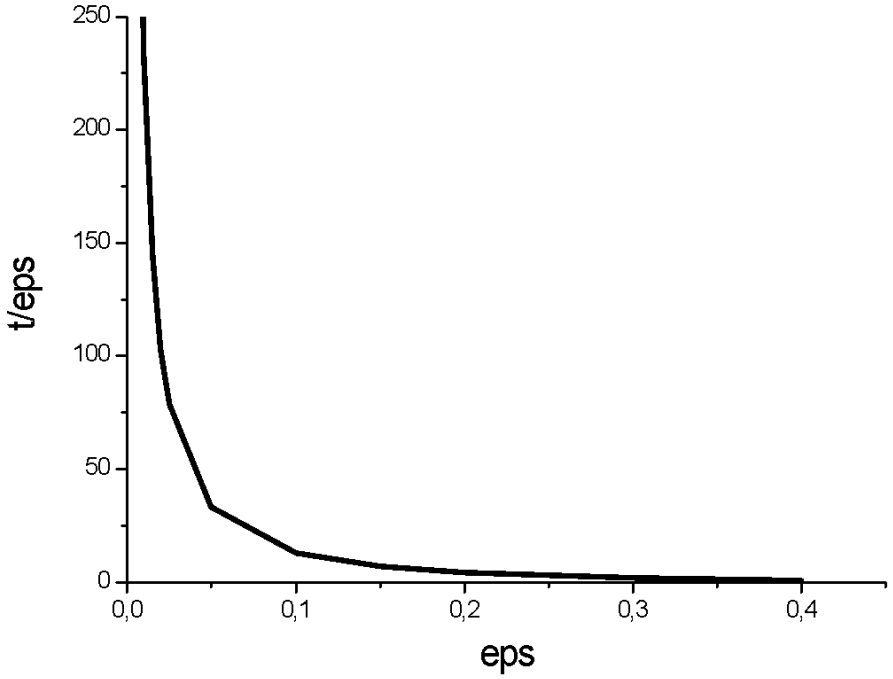
- Для стационарного процесса оценка среднего при неизвестной дисперсии дается статистикой Стьюдента

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

- Оценка длины текста для построения распределения с точностью ε:

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

# Достаточная длина текста




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

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

- При ε=0,05:

  для 1-ПФР $N$=40 тыс. знаков,

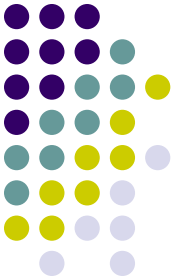
  для 2-ПФР $N$=400 тыс. знаков,

  для 3-ПФР $N$=4 млн знаков.

# Распределение расстояний между 2-ПФР в норме L1

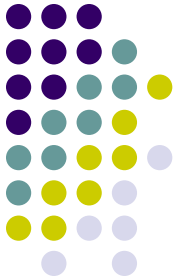
$$\rho_{a,b} = \|f_a - f_b\|_{L1} = \sum_j |f_a(j) - f_b(j)|$$

# Минимальная длина текста



| Тексты | 1-ПФР | 2-ПФР | 3-ПФР | 4-ПФР |
|---|---|---|---|---|
| Свои, $l_s$ | 0,05 | 0,14 | 0,26 | 0,38 |
| $\sigma_s$ | 0,02 | 0,05 | 0,10 | 0,10 |
| Чужие, $l_d$ | 0,12 | 0,27 | 0,45 | 0,55 |
| $\sigma_d$ | 0,05 | 0,08 | 0,09 | 0,09 |
| $\varepsilon = l_d - l_s$ | 0,07 | 0,13 | 0,19 | 0,17 |
| $N_{\min}$ , тыс. | 13 | 32 | 66 | 207 |

# Уровень нестационарности текстов



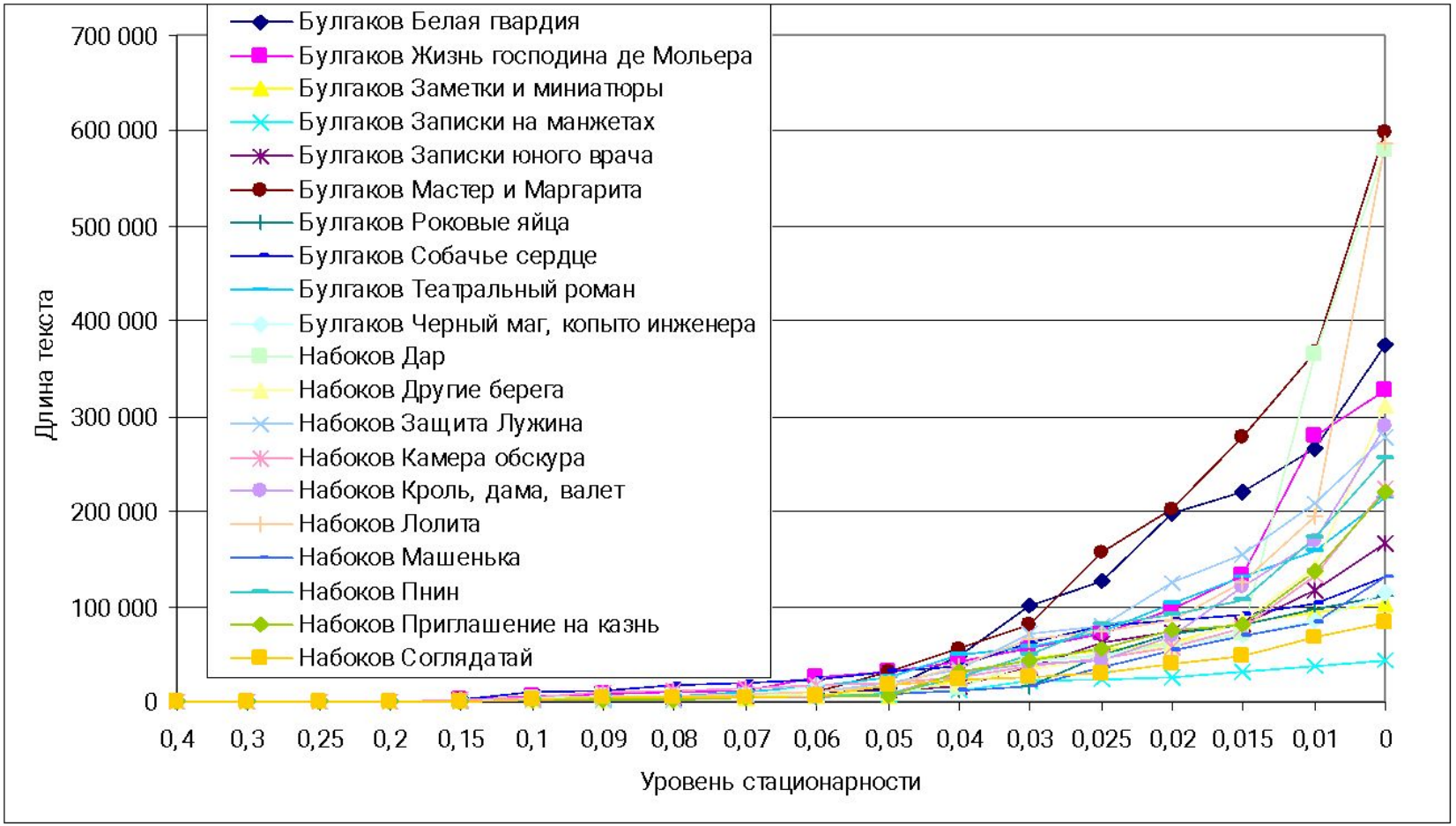
- Расстояние между ПФР текстов:

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

- Чтобы сравнивать распределения текстов разных объемов, следует убедиться в том, что каждый из них стабилизируется:

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

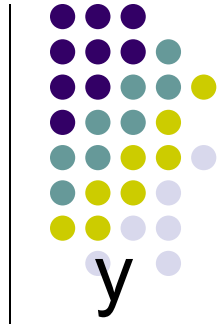
# Длина квазистационарности L(ε) для 1-ПФР




Для практических целей, где допустимы небольшие отклонения 1-ПФР отрывка от 1-ПФР всего текста, достаточно сравнительно небольших объемов текстов.

# 2. Кластеризация текстов, создание эталонных распределений и метод идентификации автора

# Идентификация автора текста



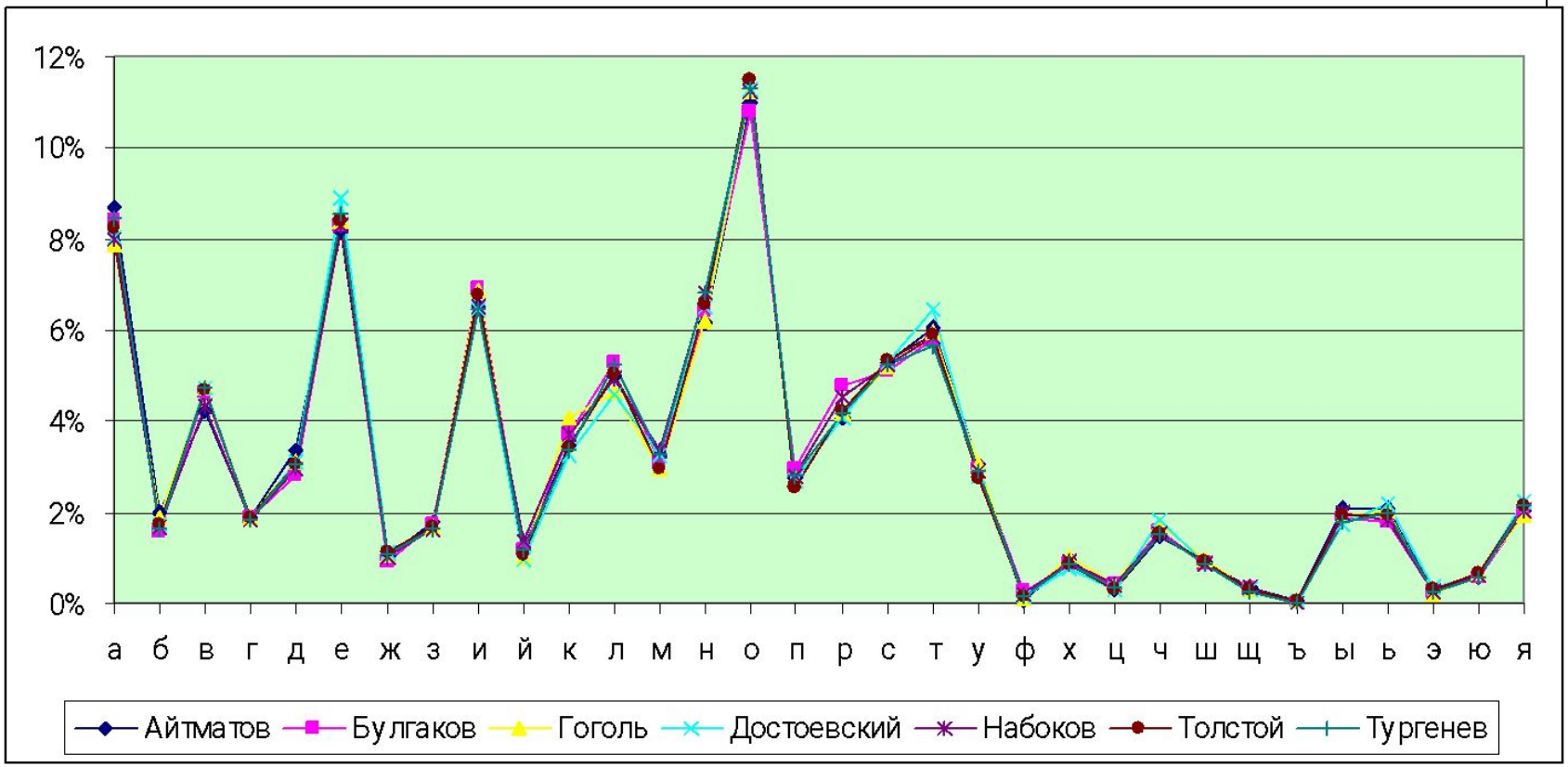
Пусть имеется библиотека из $A$ авторов, у $a$-го автора $K_a$ текстов, в $i$-ом тексте $N_{i,a}$ знаков, и $f_{i,a}(j)$ есть ПФР отдельного текста. Вводится эталонная ПФР автора:

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

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

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

# 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

# Ошибки 1-го и 2-го родов

$F_a^+(\rho)$ функция распределения расстояний текстов автора от его эталона;

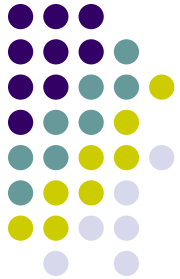
$F_a^-(\rho)$ чужих текстов от него же;

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

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

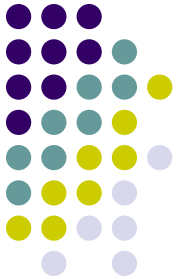
$1 - F_a^+\left(\rho_a^-\right)$ есть вероятность ошибочно признать чужой текст авторским (ошибка 2-го рода)

# Мощность статистических методов идентификации автора



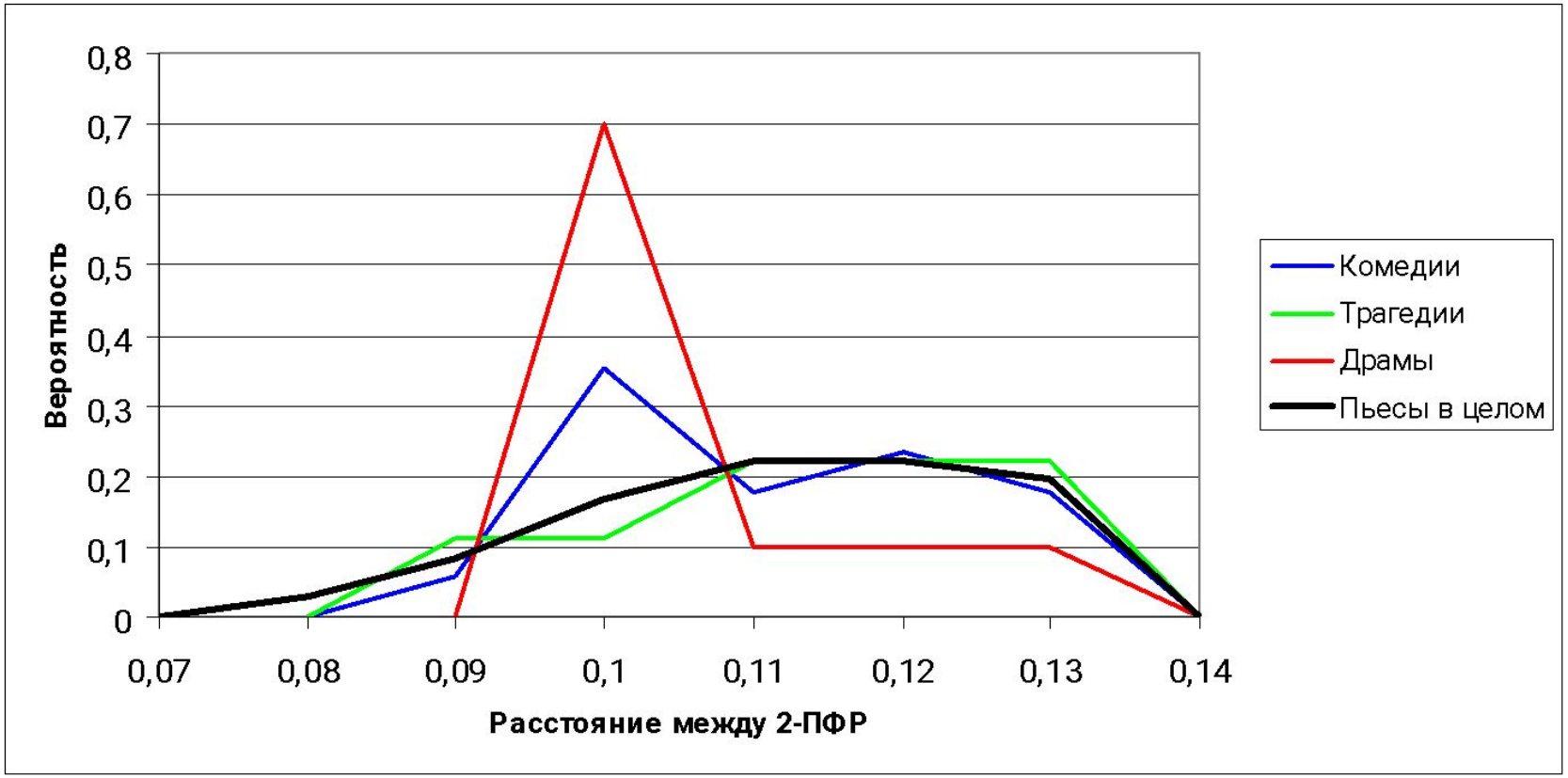
| 3000 текстов, 300 авторов | Ошибка, % |
|---|---|
| Близость 3-ПФР в норме L1 | 0 |
| Близость 2-ПФР в норме L1 | 4 |
| Близость вектора «подсознания» в норме L1 | 12 |
| Близость 1-ПФР в норме L1 | 15 |
| Доля служебных слов | 68 |
| Информационная энтропия 2-ПФР | 71 |
| Доля гласных | 81 |
| Среднее число слов в предложении | 87 |

# Шолохов – автор «Тихого Дона» по расстояниям между 2-ПФР текстов с вероятностью 96%

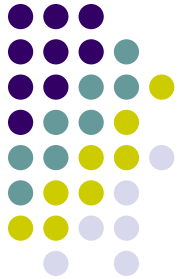


| Произведения Крюкова | | |
|---|---|---|
| | **До Крюкова** | **До Шолохова** |
| Булавинский бунт | 0,21 | 0,27 |
| В углу | 0,18 | 0,23 |
| Зыбь | 0,11 | 0,15 |
| Казачка | 0,14 | 0,19 |
| Шульгинская расправа | 0,16 | 0,19 |
| **Произведения Шолохова** | | |
| Они сражались за Родину | 0,16 | 0,13 |
| Повести | 0,21 | 0,15 |
| Поднятая целина | 0,17 | 0,11 |
| Путь дорога | 0,23 | 0,16 |
| Рассказы | 0,22 | 0,15 |
| Тихий Дон, части 1 и 2 | 0,16 | 0,08 |
| Тихий Дон, части 3 и 4 | 0,14 | 0,06 |

# Распределение расстояний между произведениями Шекспира



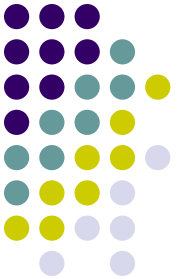
# Шекспир – автор всех своих пьес по расстояниям между 2-ПФР текстов с вероятностью 94%



| **Трагедии Шекспира и «спорный» текст** | Расстояние до средней 2-ПФР Шекспира | Расстояние до 2-ПФР Кида | Расстояние до средней 2-ПФР Марло |
|---|---|---|---|
| 1. Antony and Cleopatra | 0,116 | 0,201 | 0,170 |
| 2. The Tragedy of Coriolanus | 0,123 | 0,204 | 0,191 |
| 3. The Tragedy of Hamlet, Prince of Denmark | 0,086 | 0,153 | 0,152 |
| 4. The Life and Death of Julies Caesar | 0,138 | 0,224 | 0,204 |
| 5. King Lear | 0,096 | 0,162 | 0,149 |
| 6. The Tragedy of Macbeth | 0,110 | 0,165 | 0,150 |
| 7. Othello, the Moore of Venice | 0,101 | 0,174 | 0,169 |
| 8. Romeo and Juliet | 0,126 | 0,162 | 0,149 |
| 9. Timon of Athens | 0,118 | 0,166 | 0,164 |
| **Titus Andronicus** | **0,129** | **0,159** | **0,153** |

# 3. Спектральные портреты авторов и эффект переводчика

# Оператор трансляций

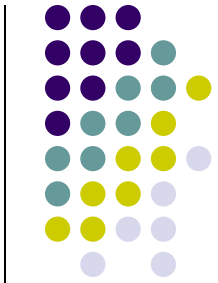


- Пусть $P_{ij}$ есть условная вероятность того, что буква $j$ следует за буквой $i$.

- Пусть также $K_i(t)$ есть $i$-ая компонента вектора вероятностей того, что буква $i$ реализуется в тексте в момент $t$.
- Тогда

$$\mathbf{K}(t+1) = P\mathbf{K}(t)$$

# Оператор трансляций на 1 шаг



- $P_{ij}$ выражается через 1-ПФР и 2-ПФР:

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

- По формуле полной вероятности

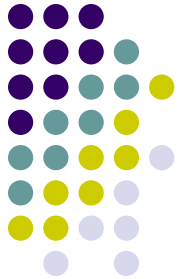
$$f(j) = \sum_k F(k, j) = \sum_k P_{jk} f(k)$$

- Следовательно, 1-ПФР

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

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

# ε-спектр оператора соседних трансляций



- Число λ называется принадлежащим ε-спектру $\Lambda_\varepsilon(P)$ матрицы P, если существует матрица Δ такая, что

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

- Резольвентой 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

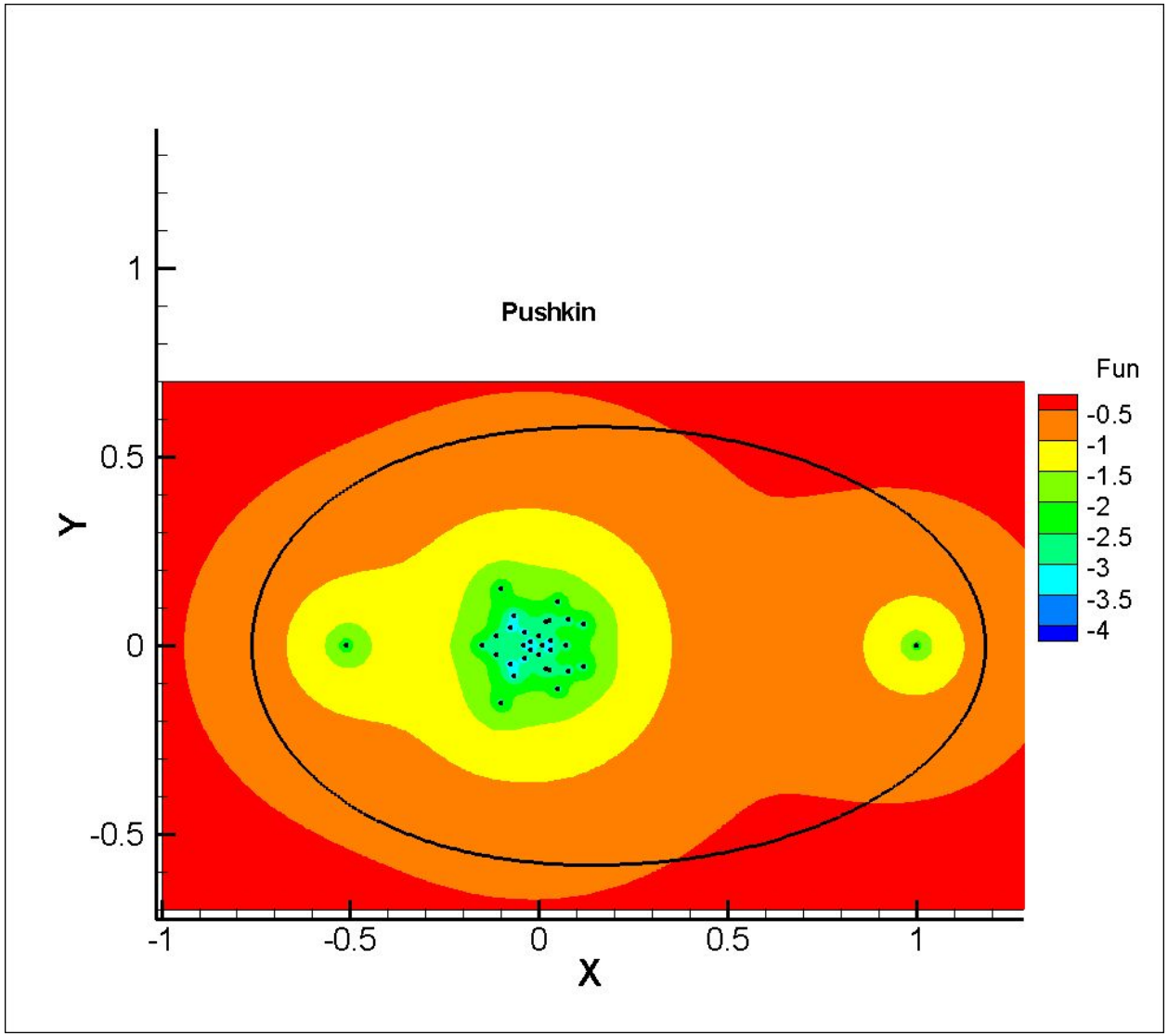
- Параметром дихотомии спектра относительно кривой $\gamma$ называется норма квадрата резольвенты на данной кривой:

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

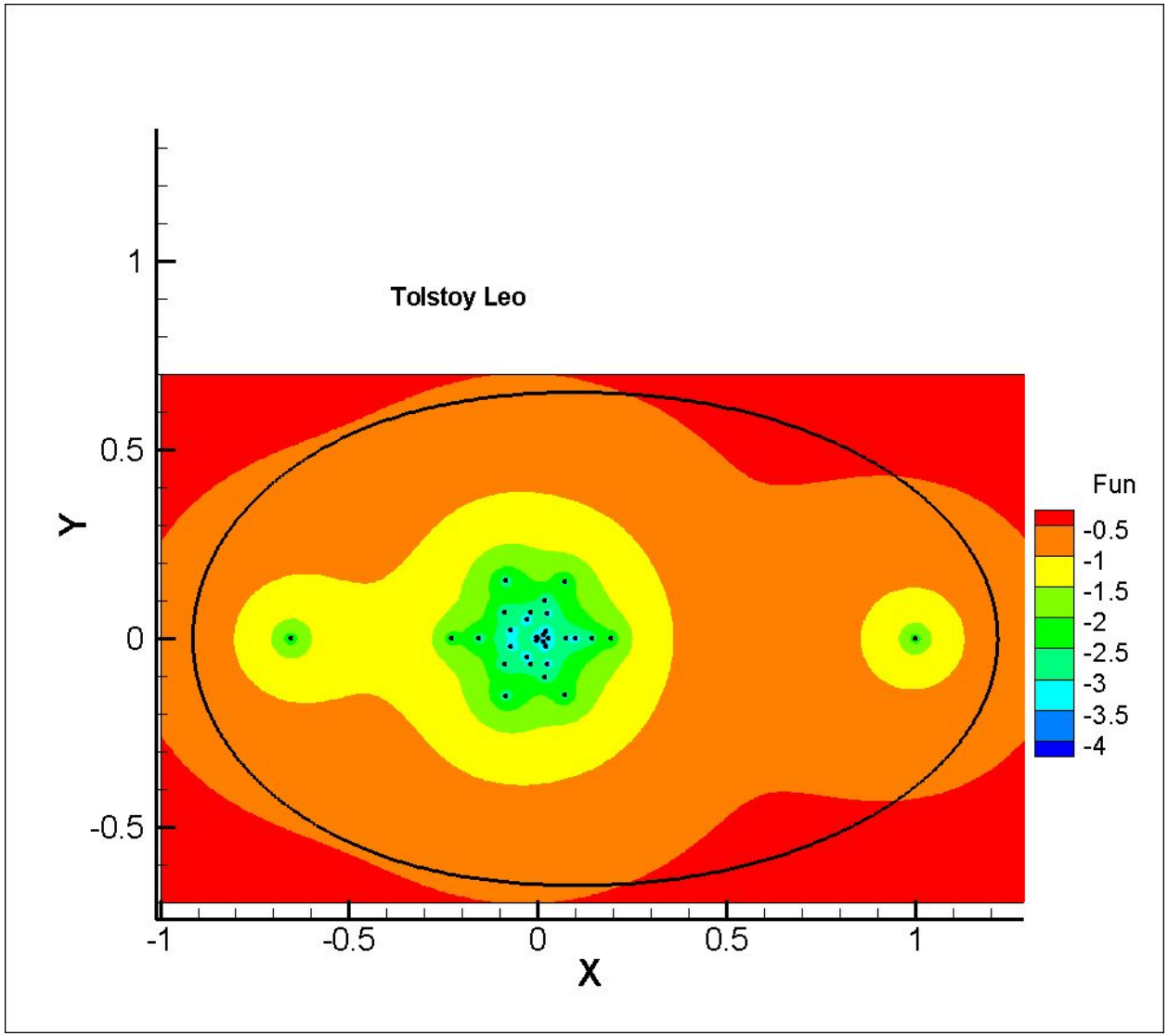
Если на кривой нет точек спектра, то норма резольвенты на этой кривой конечна.

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

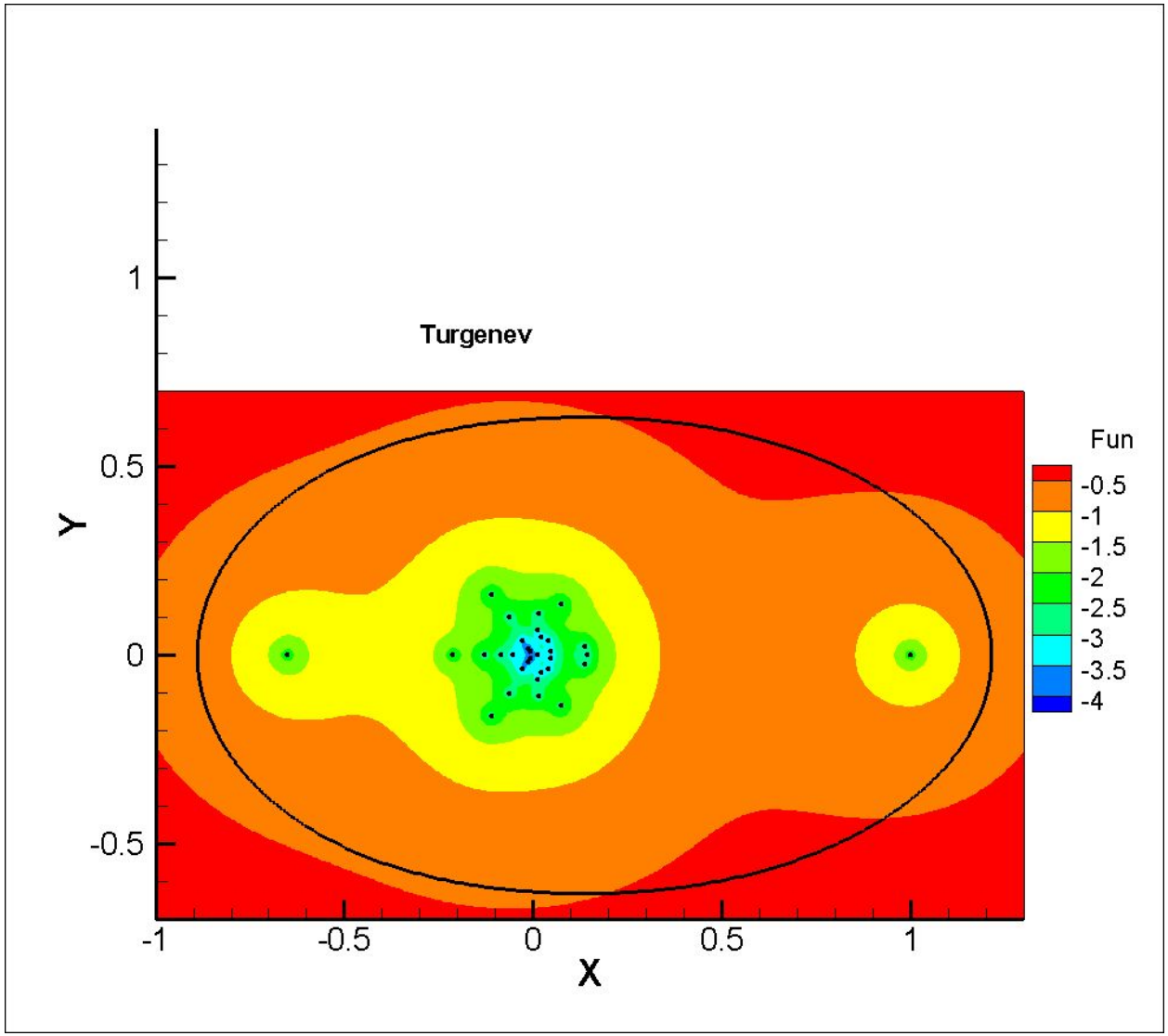
# Примеры спектральных портретов писателей



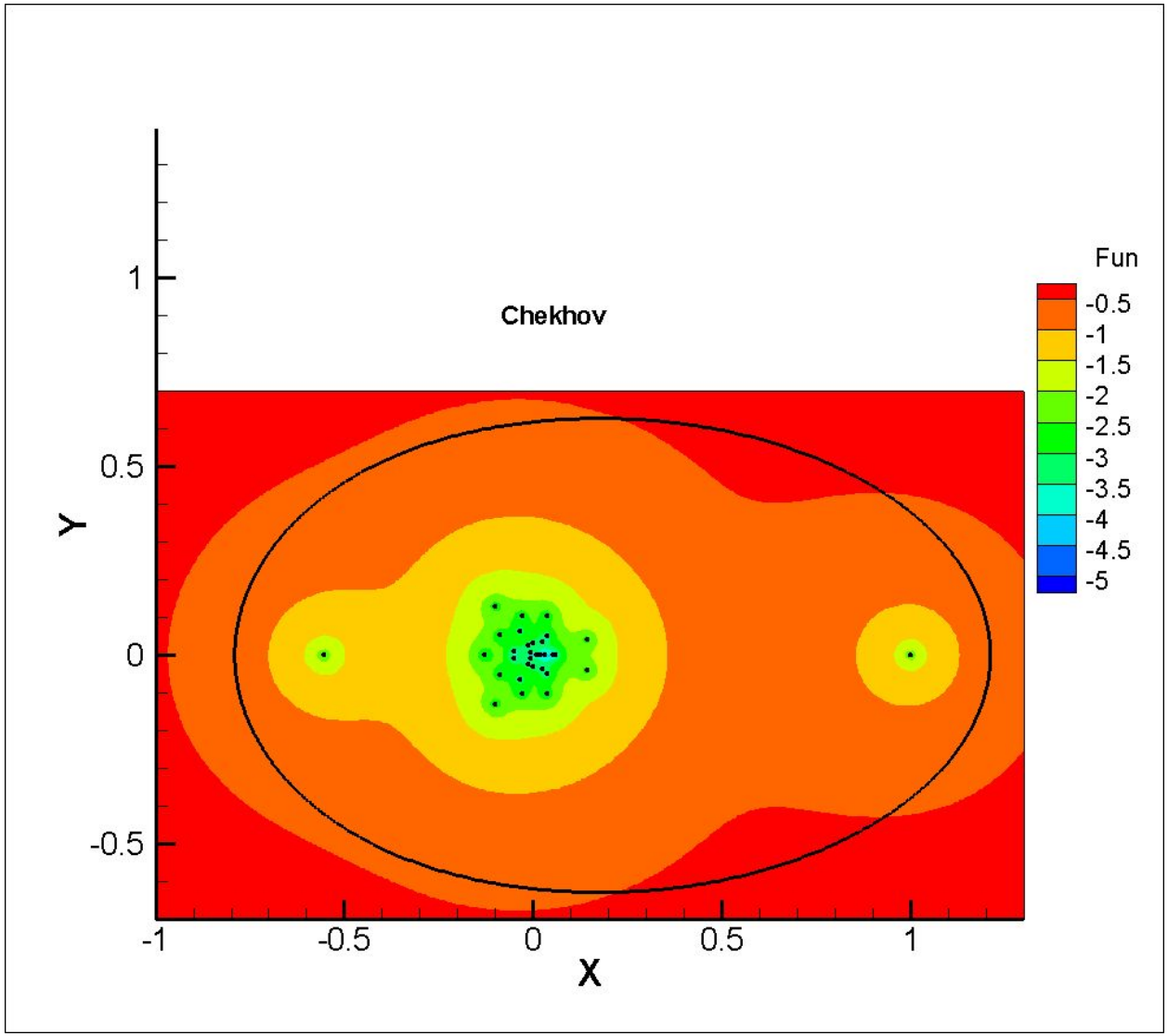
# Примеры спектральных портретов писателей



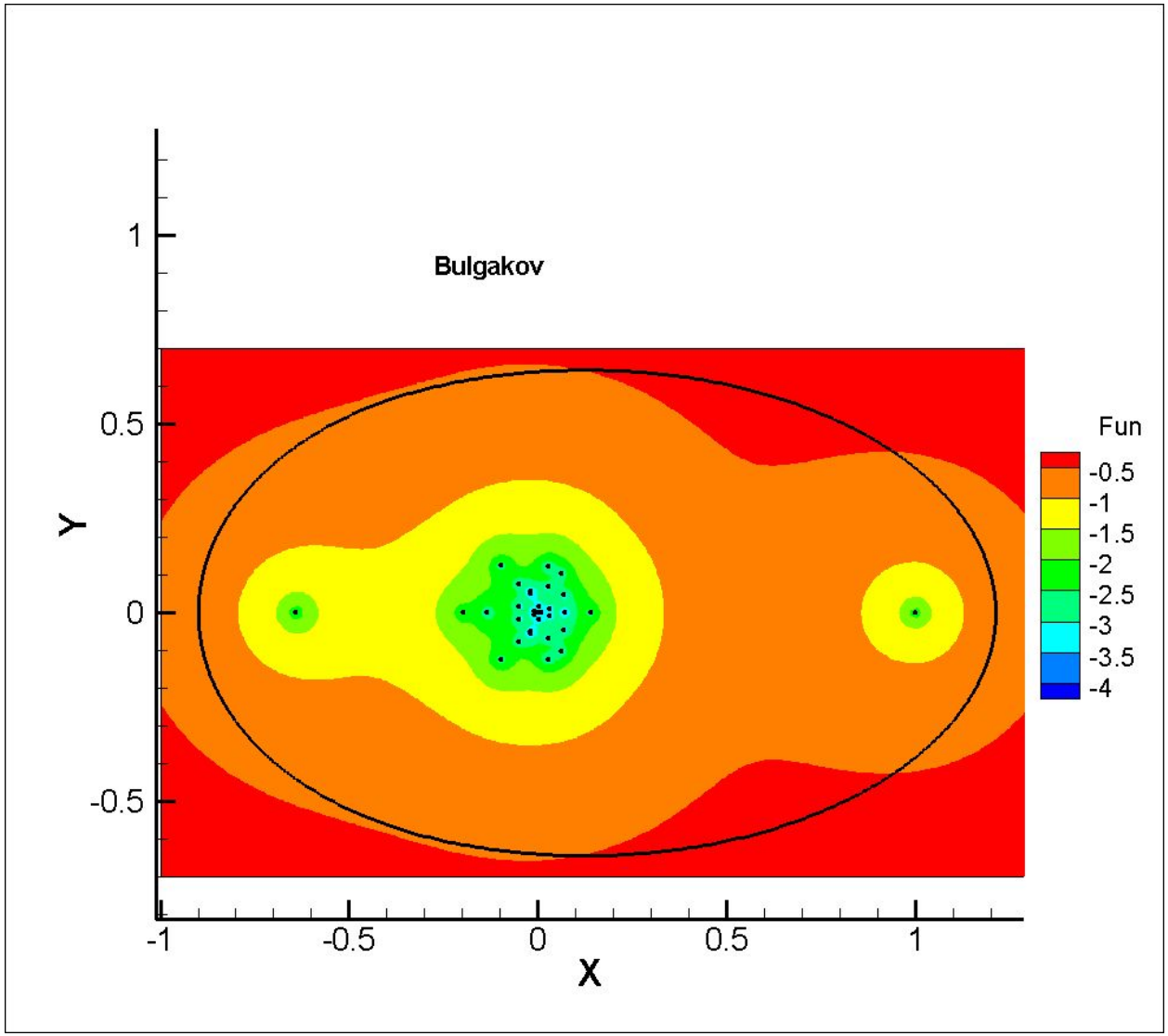
# Примеры спектральных портретов писателей



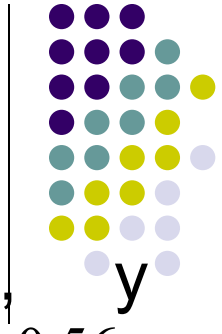
# Примеры спектральных портретов писателей



# Примеры спектральных портретов писателей



# Эффект переводчика и вектор «подсознания»



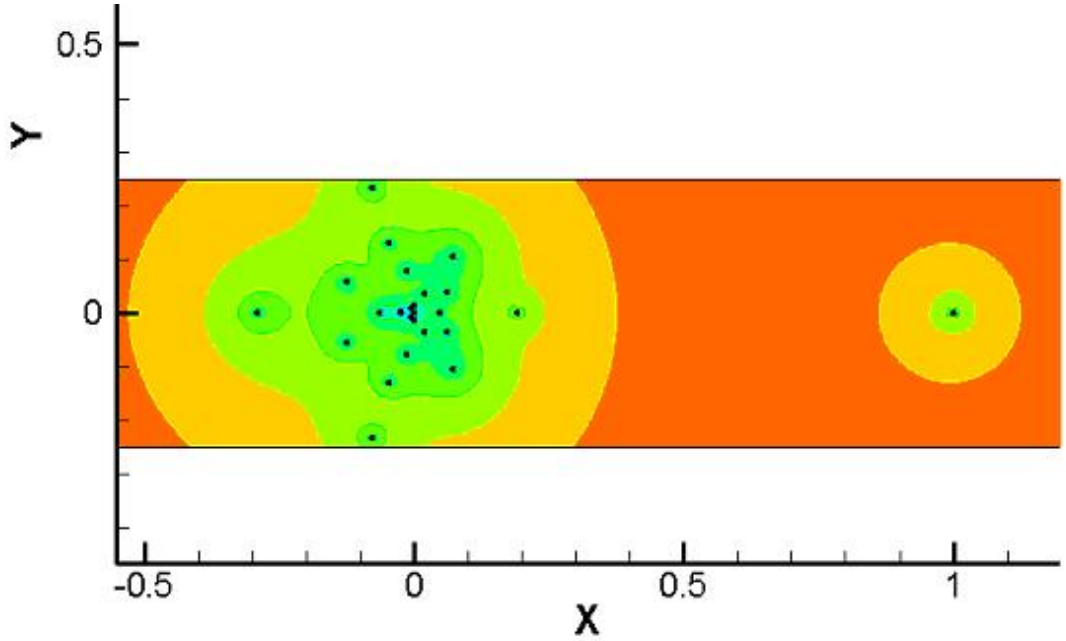
- Кроме с.з. $\lambda = 1$, которому отвечает с.в. 1-ПФР $f$, у оператора P(1) еще одно устойчивое с.з. $\mu \approx -0{,}56$
  Ему отвечает правый с.в. S и левый S*.

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

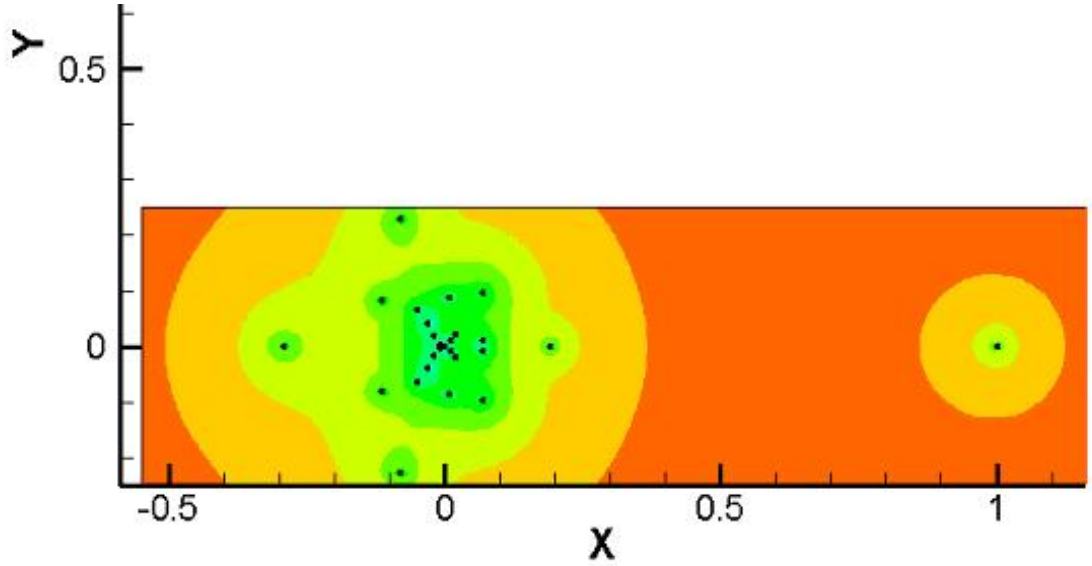
- Вектор S, как и вектор 1-ПФР $f$, весьма точно идентифицирует автора. Однако в переводах это идентификационное свойство теряется.

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

# Шекспир – оригинальный текст

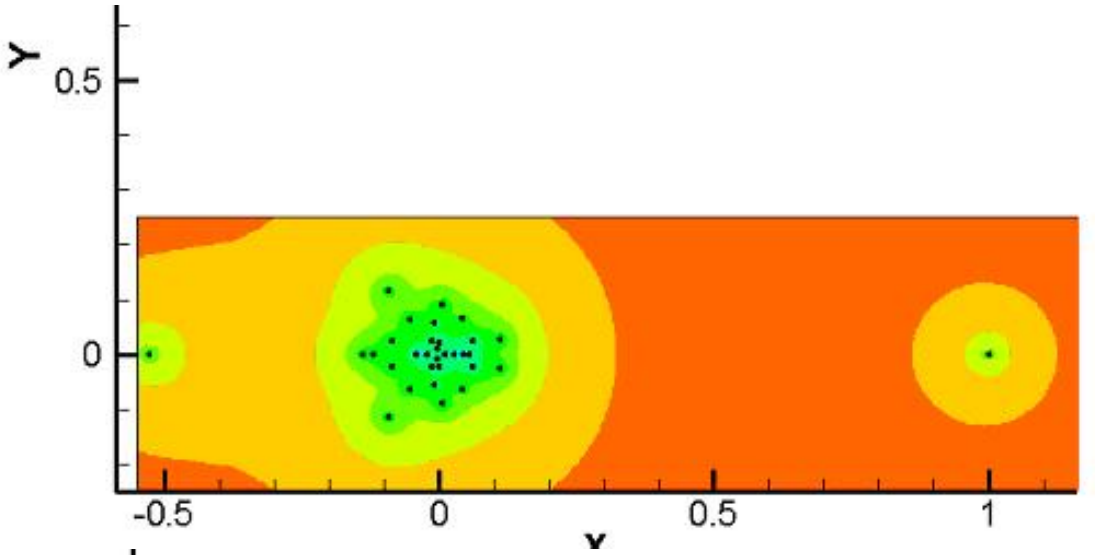



«Гамлет»

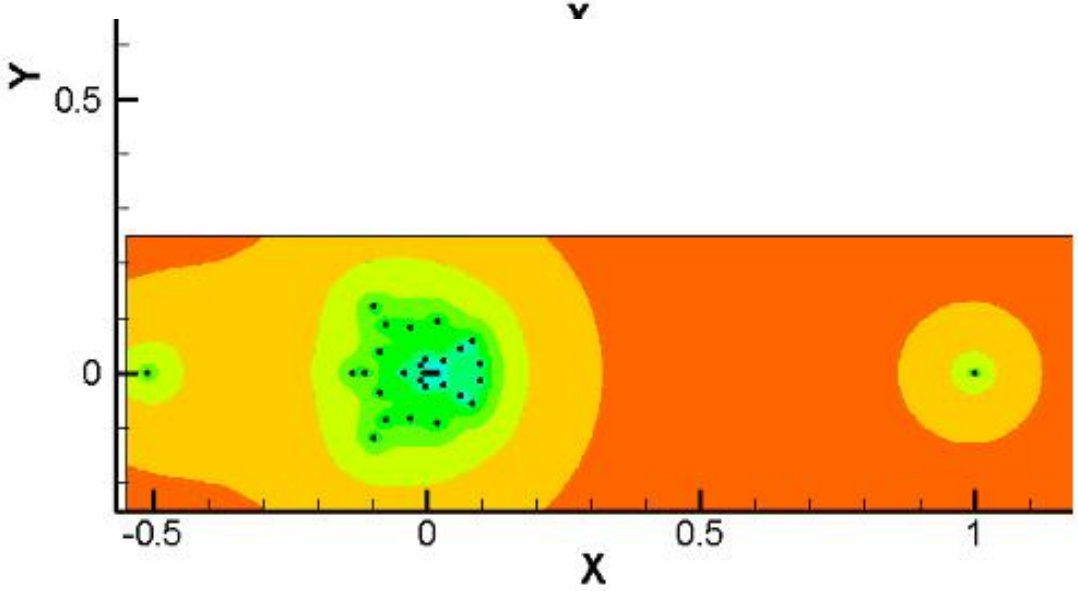



«Много шума из ничего»

# Шекспир – перевод



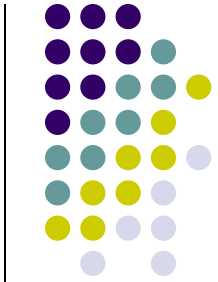

«Гамлет»
(Лозинский)




«Много шума из ничего»
(Щепкина-Куперник)

# 4. Анализ авторских тандемов и проверка текста на однородность

# Горизонтный ряд



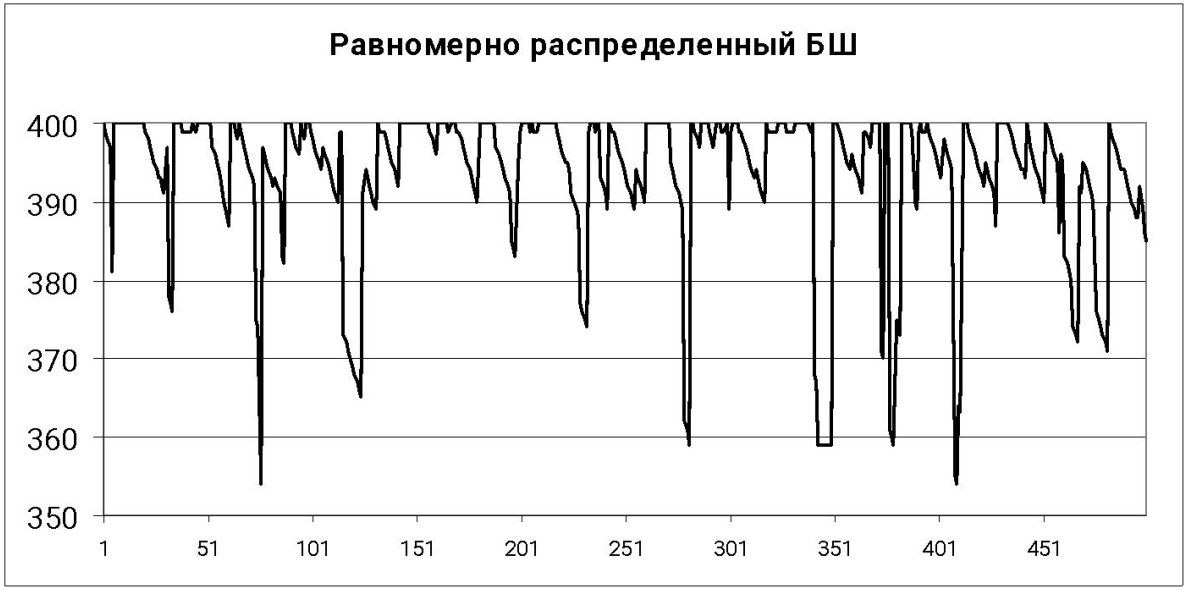
- Пусть $x(t)$ – эквидистантный временной ряд, $f(N,t)$ – его ВПФР, построенная к моменту $t$ по выборке объема $N$.

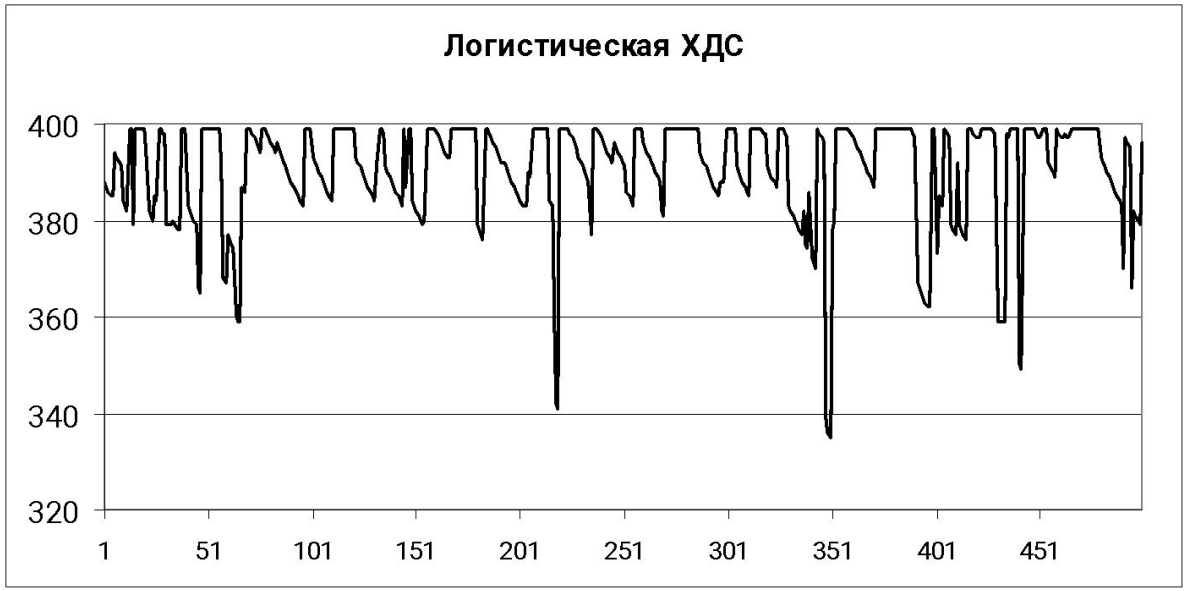
- Горизонтным рядом $h(t,\tau;\varepsilon)$ для ряда $x(t)$ называется минимальный объем выборки такой, что

$$\forall N \geq h(t,\tau;\varepsilon),\ \forall k \in [0;\tau]:\quad \left\| f(N,t+k) - f(N,t) \right\| \leq \varepsilon$$

$$h_{\max} = \min\left\{2; [2\tau/\varepsilon]\right\}$$

# Примеры горизонтных рядов




Сдвиг ВПФР на $\tau = 10$ при уровне $\varepsilon = 0{,}05$

Значения h, близкие к предельным, означают хаос, а уменьшение означает взаимосвязь элементов

# Горизонтный ряд как индикатор разладки ($\tau = 10$, $\varepsilon = 0{,}05$)

# Распределения горизонтных рядов для $\tau = 10$, $\varepsilon = 0{,}05$

# Распределение расстояний между одинаковыми буквами

$$\ln L(i) = 1{,}2 - 0{,}4 \ln f(i)$$

$$B_i(l) = \frac{(\lambda_i l)^{\nu_i}}{\Gamma(\nu_i + 1)} e^{-\lambda_i l}$$

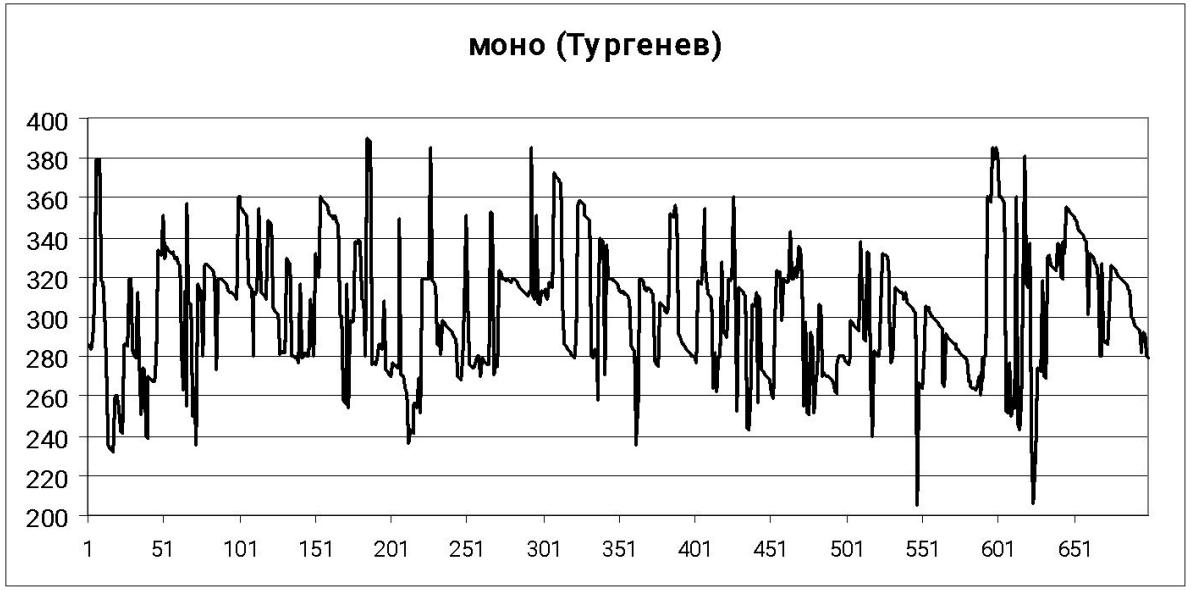
$$\lambda_i \approx \frac{1}{4 + \frac{1}{f(i)} + \frac{5}{2} \ln f(i)},$$

$$\nu_i = \frac{\lambda_i}{f(i)} - 1 \approx 1/4$$

# Распределение горизонтного ряда для расстояний «b-b» ($\tau$ = 10, $\varepsilon$ = 0,05)

Для всех букв распределение горизонтного ряда одинаково. Оно похоже на распределение для нелинейно коррелированных многомерных ХДС.

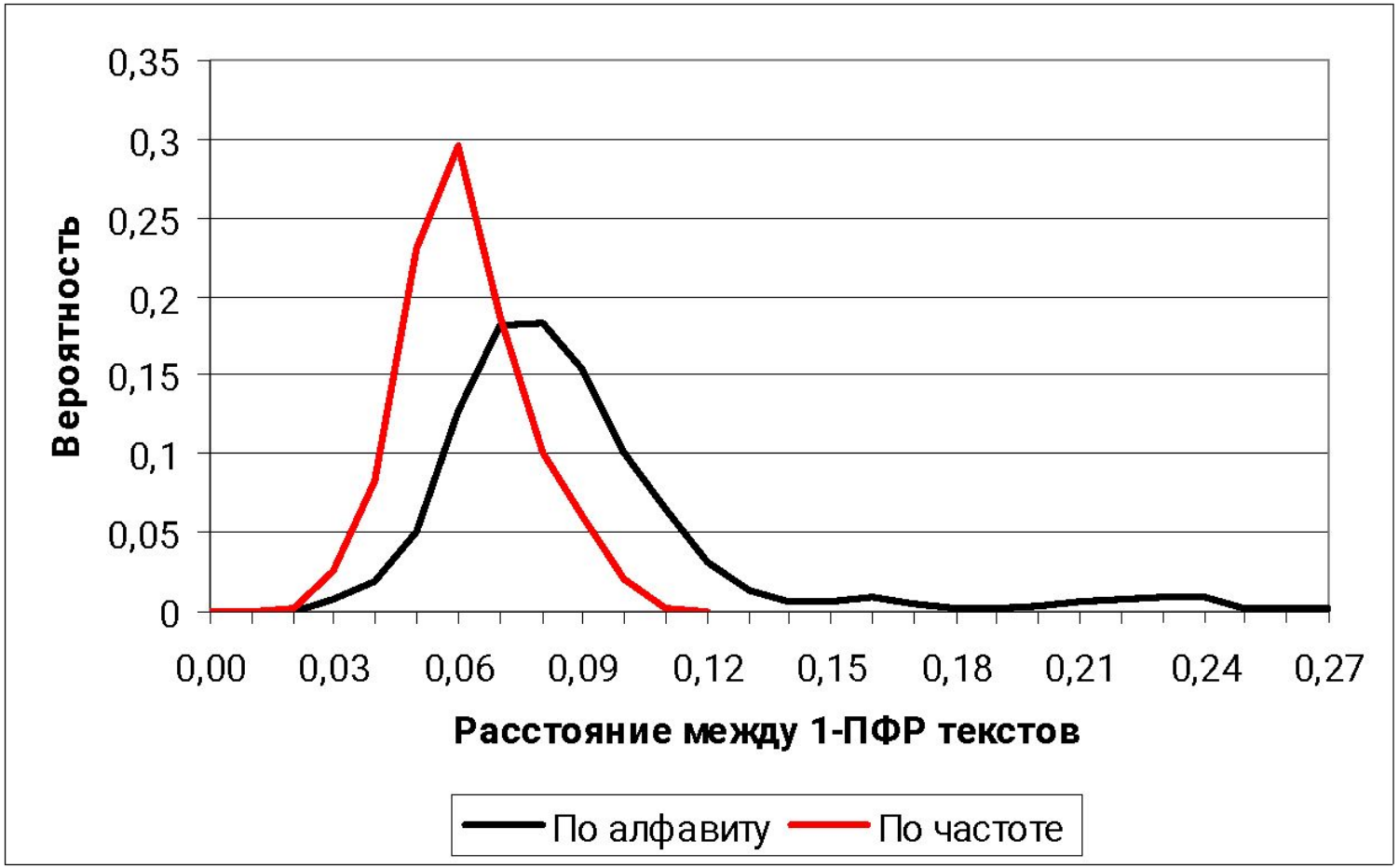
# Горизонтные ряды расстояний между гласными для моно и тандема




**У моно-писателей горизонтный ряд не достигает последней полосы шириной в горизонт, а у тандемов есть места максимальных рассогласований**

# 5. Упорядоченность букв по частоте встречаемости в европейских языках

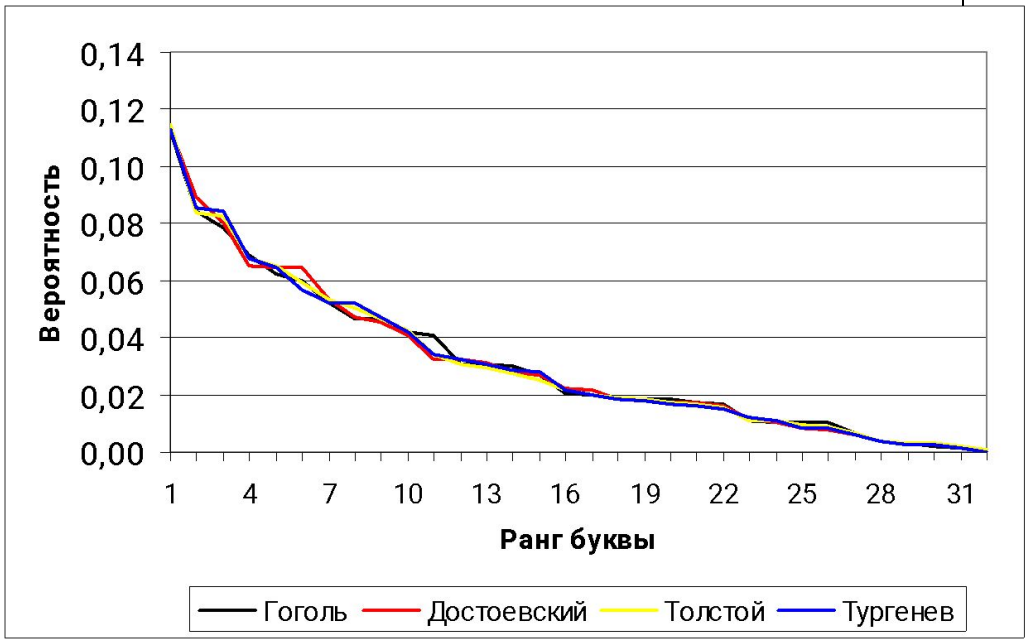
# Расстояния между текстами при различном упорядочении



# Распределение букв по частоте в алфавите из $n$=32 знаков

- С детерминацией 0,97

$$f(k) = a - b\ln(k)$$



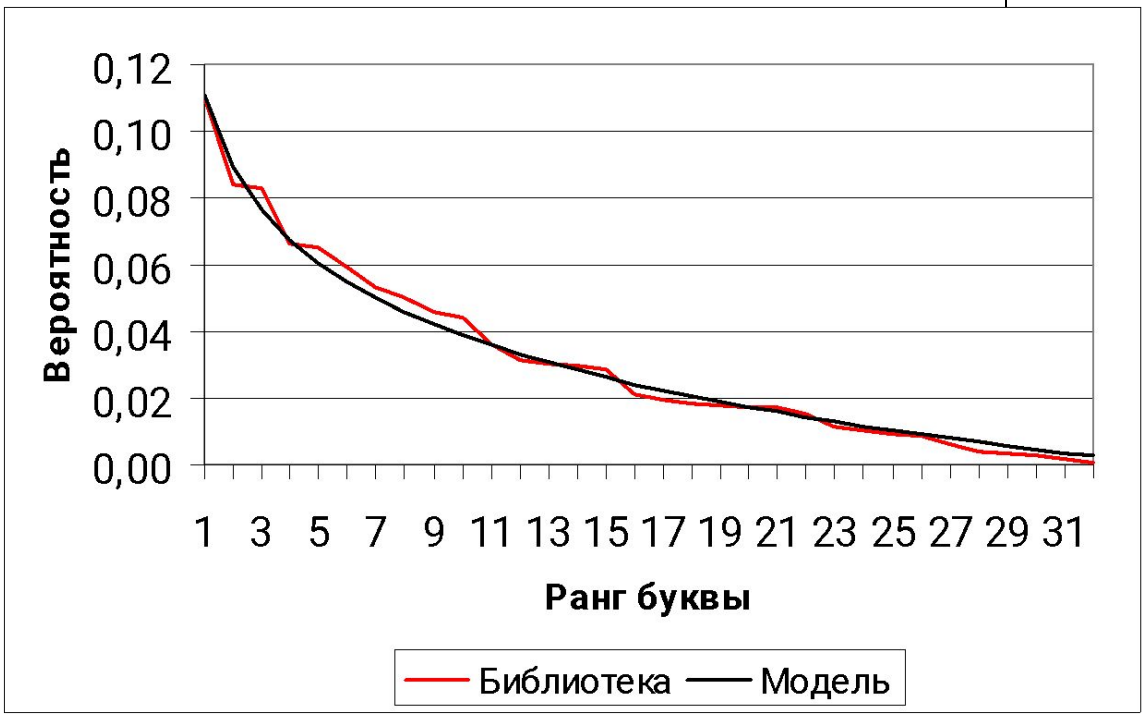

| Ранг буквы | Гоголь | Достоевский | Толстой | Тургенев |
|---|---|---|---|---|
| 1 | О | О | О | О |
| 2 | Е | Е | А | Е |
| 3 | А | А | Е | А |
| 4 | И | И | И | Н |
| 5 | Т | Н | Н | И |
| 6 | Н | Т | Т | Т |

# Распределение букв по частоте в текстах на русском языке

Минимальная интегральная ошибка приближения, равная 0,05, получается при $o$=0 в модели:

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

$$o = const = 0$$



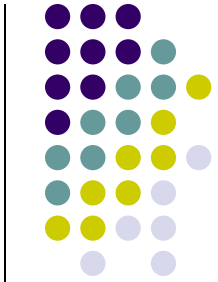

- Эта зависимость выполнена и для старославянских текстов (n=43), и для русской литературы XIX века (n=37). Для русских текстов в транслите (n=23 символа) $o$=+9.

# Избыточность и недостаточность алфавитов европейских языков

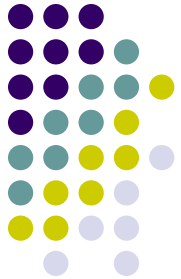
- Параметр $o$ трактуем как оценку избыточности ($o<0$) или недостаточности ($o>0$) алфавита по отношению к звуковому ряду. В текстах на всех языках без огласовки $n$ = 20, $o$ = 0.

| Языки | Число символов | Параметры оптимальной модели |
|---|---|---|
| Русский | $n$=33 | $n$=32, $o$= 0 |
| Болгарский | $n$=30 | $n$=30, $o$= –4 |
| Чешский | $n$=42 | $n$=30, $o$= +1 |
| Польский | $n$=32 | $n$=32, $o$= +3 |
| Шведский | $n$=29 | $n$=25, $o$= +1 |
| Датский | $n$=29 | $n$=28, $o$= –5 |
| Немецкий | $n$=30 | $n$=26, $o$= –4 |
| Английский | $n$=26 | $n$=26, $o$= 0 |
| Итальянский | $n$=26 | $n$=26, $o$= –4 |
| Испанский | $n$=27 | $n$=26, $o$= –4 |
| Французский | $n$=42 | $n$=26, $o$= –4 |

# Основные результаты



- 3-ПФР представляет ту текстовую структуру, расстояние в которой позволяет с высокой точностью опознавать автора
- Построен индикатор однородности текста (горизонтный ряд), позволяющий анализировать небольшие фрагменты на предмет количества возможных соавторов
- Изучен спектр оператора эволюции 1-ПФР и показана авторская устойчивость спектральных портретов. Пара главных направлений позволяет определить, собственный ли это текст автора, или изложение чужих мыслей
- Найдено универсальное полуэмпирическое распределение букв по частоте встречаемости в европейских языках, позволяющее оценить фонетическую адекватность алфавита



# СПАСИБО ЗА ВНИМАНИЕ