# Метод Дельты Бёрроуза для определения авторства анонимных и псевдонимных литературных произведений на русском языке

# Burrows's Delta for Authorship Attribution of Russian Literary Texts

**Н. К. Мамаев**[4]
**Nikita Mamaev**[4]
nikita.mamaev1@gmail.com

**М. А. Марусенко**[2]
**Mikhail Marusenko**[2]
mamikhail@yandex.ru

**К. Р. Пиотровская**[1]
**Xenia Piotrowska**[1]
krp62@mail.ru

**А. Л. Ронжин** [3]
**Andrey Ronzhin** [3]
ronzhin@iias.spb.su

[1]Российский государственный педагогический университет им. А. И. Герцена
[2]Санкт-Петербургский государственный университет
[3]Санкт-Петербургский институт информатизации и автоматизации РАН
[4]Университет ИТМО
Санкт-Петербург, Российская Федерация

[1]Herzen State Pedagogical University of Russia
[2]Saint Petersburg State University
[3]The Saint Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
[4] ITMO University
Saint Petersburg, Russian Federation

CEUR-WS.org/Vol-2233/Paper_9.pdf

**Abstract**

The intertextual distance measure proposed by J. Burrows (Burrows's Delta and its variants) and its application to authorship attribution of some novels in Russian are under discussion. The research aim is to test classic (Burrows's) Delta and Eder's Delta, which

implemented in stylo package for computational text analysis developed for the R software environment. The novels “Twelve Chairs” and “Golden Calf”, officially attributed to Soviet writers I. Ilf and E. Petrov, are the research material.

**Keywords:** *authorship attribution, quantitative analysis, intertextual distance, Burrows's Delta, Eder's Delta, Cosine Delta.*

**Аннотация**

В работе обсуждается эффективность применения меры межтекстового расстояния, предложенной Дж. Бёрроузом, при уточнении авторства литературных произведений на русском языке. Производится тестирование классической Дельты и Дельты Эдера с помощью пакета для стилеметрического анализа stylo, разработанного для программной среды R. В качестве материала исследования выступают романы “Двенадцать стульев” и “Золотой телёнок”, официально приписываемые советским писателям И. Ильфу и Е. Петрову.

**Ключевые слова:** *атрибуция, квантитативная анализ, исследование авторства, межтекстовое расстояние, Дельта Бёрроуза, Дельта Эдера, косинус-Дельта.*

# 1 Введение

Наравне с историко-филологическим анализом материала, установление авторства анонимных и псевдонимных произведений традиционно опирается на постоянно расширяемый аппарат квантитативного анализа, который предоставляется точными науками: математической статистикой, информатикой, а также сравнительно новым направлением – текст-майнингом. Для решения задач атрибуции используются традиционные математические инструменты, широко применяемые различными областями знания: манхэттенское и евклидово расстояния; мера Жаккара; критерии Стьюдента, Сёренсена-Чекановского, $\chi$-квадрат и др., а также некоторые средства, предложенные специально для квантитативного анализа текстов на естественном языке: Дельта Бёрроуза и межтекстовое расстояние Лаббе [Jannidis et al. 2015; Labbe et al. 2001]. Межтекстовое расстояние, предложенное Дж. Ф. Бёрроузом [Burrows 2002] в 2001 г. и называемое Дельтой Бёрроуза (Burrows's Delta) или классической Дельтой (Classic Delta), получило в стилеметрии довольно широкое распространение, и за последние 15 лет были получены следующие результаты:

- было предложено несколько модификаций названной методики: Eder's Delta, Hoover's Delta, Rotated Delta, Cosine Delta и др. [Hoover 2004; Argamon 2008; Rybicki et al. 2011; Jannidis et al. 2015], эффективность которых была проверена в сравнении с манхеттеновским и евклидовым расстояниями на текстах на различных языках Западной Европы [Evert et al. 2017];

- были разработаны открытые программные пакеты для стилеметрического анализа и атрибуции (stylo для программной среды R и JGAAP, написанный на языке Java) [Rybicki et al. 2011; Juola 2009];

- применение методов кластеризации, классификации и понижения размерности данных предоставило возможность повысить наглядность представления результатов исследований.

Эти шаги привели к популяризации Дельты Бёрроуза, но, несмотря на активное применение, потенциал её использования для анализа текстов на русском языке до сих пор не был изучен в полной мере. В связи с этим мы предприняли попытку тестирования заявленной методики на текстах некоторых классических литературных произведений советской эпохи 20-30-х гг., официальное авторство которых за последнее десятилетие было подвергнуто сомнению.

## 2 О проблеме авторства романов “Двенадцать стульев” и “Золотой телёнок”

Согласно официальным литературоведческим данным дилогия, включающая романы “Двенадцать стульев” и “Золотой телёнок”, была написана советскими писателями-соавторами Ильёй Ильфом и Евгением Петровым. Впервые романы были опубликованы в 1928 и 1931 году соответственно.

Авторы романов, осмеивая тёмные стороны реалий молодого советского государства, предлагают новый сатирический стиль, отличный от стиля других произведений этого периода: остроумный и содержательный, насыщенный оптимизмом и тонким юмором. В этом же ключе работали такие мэтры русской литературы, как М. А. Булгаков и С. С. Заяицкий, произведения которого пользовались популярностью в 20-х гг. М. А. Булгаков – автор “трагических буффонад”: ср. “Роковые яйца” (1924). “Собачье сердце” (1925), “Зойкина квартира” (1925), “Адам и Ева” (1931), “Иван Васильевич” (1936), “Мастер и Маргарита” (1929-1940), а С. С. Заяицкий – автор ряда опубликованных сатирических произведений: “Земля без солнца” (ср. альманах “Рол”, 1925, № 4), “Красавица с острова Люлю” (1926, под псевдонимом: Пьер Дюмьель), “Баклажаны” (1927), “Жизнеописание Степана Александровича Лососинова” (1928). Оба автора любили литературные мистификации, закодированные послания читателям и иногда писали под псевдонимами (ср. Г. П. Ухов - Булгакова или Пьер Дюмьель - Заяицкого).

Первым изданием, высказавшим сомнения по поводу достоверности авторства романов, была монография И. Амлински “12 стульев от Михаила Булгакова” [Amlinski 2013]. В своей книге автор проанализировала данные, касающиеся истории возникновения замысла, работы над произведениями и их публикации – как общепринятой, так и альтернативных версий. Также Амлински рассмотрела подготовительные материалы, варианты и редакции всех произведений М. Булгакова, включая черновики, все произведения Ильфа и Петрова, а также воспоминания современников об этих писателях. Опираясь на собранный материал, автор провела глубокий анализ лексического и образного строя произведений. Выявив схожести в структуре описания сцен, основных образах и мелких деталях повествования, Амлински заключила, что стиль романов очень близок к авторскому стилю Булгакова, и является слишком личным, чтобы допустить, что над ними работало два человека.

Рассмотрев правдоподобность этой мистификации, литературный критик В. Козаровецкий [Kozaroveckii 2013] подкрепил выводы Амлински биографическими сведениями: на время написания романов Булгаков, Ильф и Петров работали в редакции московской газеты “Гудок”, и вполне могли организовать добровольно-принудительный творческий союз. Также Козаровецкий отметил, что против Булгакова к середине 20-х гг. была настроена как советская литературная критика, так и спецслужбы, и по этой причине он не мог надеяться на возможность публикации своих новых сочинений.

Наконец, альтернативную гипотезу предложил В. Тришин, ссылаясь на архивные материалы, утверждающие, что писатель С. Заяицкий перед смертью закончил неизвестный большой, не найденный впоследствии роман, который он мог успеть передать или продать [Trishin 2017]. Проведенный Тришиным сравнительный историко-филологический анализ также показал близость авторского стиля Заяицкого к стилю атрибутируемых произведений.

Наконец, высказанные критиками сомнения поддерживаются общеизвестной историей публикации романа "Двенадцать стульев". В 1926 г. В. Катаев предложил Е. Петрову и И. Ильфу "поработать литературными неграми над романом про стулья, в которых спрятаны деньги, и которые нужно найти" [Petrov 2001: 146-147]. И хотя известно, что работа в соавторстве замедляет процесс создания произведения, в 1928 г. роман был уже опубликован, стремительно пройдя все этапы публикации и цензуры.

# 3 Метод исследования

Мера межтекстового расстояния Дельта, предложенная Дж. Бёрроузом в 2001 [Burrows 2002], была испытана многими западными исследователями на больших объёмах разнородных текстовых данных:

- английская проза начала XX в. [Hoover 2004],
- современная английская поэзия [Hoover 2005], а также поэтические произведения на латыни [Rybicki et al. 2011]
- прозаические произведения крупных форм на различных западноевропейских языках: английском, французском, итальянском, немецком, польском, венгерском языках, а также на латыни и арабском [Rybicki et al. 2011; Evert et al. 2015; Jannidis et al. 2015],
- политические тексты на английском языке, в том числе атрибуция речей американских президентов [Savoy 2015].

Математическое обоснование межтекстового расстояния Дельта формулируется следующим образом.

Пусть есть множество из $n$ слов, представляющих интерес для исследования, относительно которых будет вычисляться мера Дельта. Назовём это множество слов $\{w_i\}$, определяя $f_i(D)$ как частоту слова $w_i$ в тексте $D$, а $\mu_i$ – как среднюю частоту слова по выборке, и $\sigma_i$ – как стандартное отклонение этой частоты. Тогда стандартизованная оценка, или $z$-оценка частоты употребления слова $w_i$ в тексте $D$ вычисляется по формуле

$$z(f_i(D)) = \frac{f_i(D) - \mu_i}{\sigma_i}. \tag{1}$$

Таким образом, имеем следующее математическое выражение, соответствующее среднему абсолютных значений разностей стандартизованных оценок частот слов из $\{w_i\}$ между текстами $D$ и $D'$, называемое мерой Дельта:

$$\Delta(D, D') = \frac{1}{n} \sum_{i=1}^{n} |z(f_i(D)) - z(f_i(D'))|. \tag{2}$$

Эту формулу можно преобразовать следующим образом:

$$\Delta(D, D') = \frac{1}{n}\sum_{i=1}^{n} |z(f_i(D)) - z(f_i(D'))| = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{f_i(D) - \mu_i}{\sigma_i} - \frac{f_i(D') - \mu_i}{\sigma_i} \right| =$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left| \frac{f_i(D) - f_i(D')}{\sigma_i} \right|. \quad (3)$$

Это преобразование показывает, что Дельта на самом деле не зависит от средних частот слов $\mu_i$ по выборке, и её можно рассматривать как нормированную меру разницы между частотами каждого из слов в текстах $D$ и $D'$. Также заметим, что поскольку взятие среднего включает деление суммы на постоянную, равную числу рассматриваемых слов $n$, то при сравнении вычисленных результатов этим действием можно пренебречь. Поэтому формула преобразуется до вида:

$$\Delta(D, D') = \sum_{i=1}^{n} \frac{1}{\sigma_i} |f_i(D) - f_i(D')|, \quad (4)$$

т.е. меру Дельта относительно пары текстов $(D, D')$ принимаем равной сумме по множеству слов $w_i$ абсолютных значений разностей частот слов между текстами $D$ и $D'$, делённых на стандартное отклонение $\sigma_i$.

Используя Дельту в задаче атрибуции, мы делаем попытку сравнить кандидатов на авторство текста $D'$ путём оценивания многомерного расстояния до текста $D$, при этом каждое измерение (частота употребления слова) масштабируется множителем $\frac{1}{\sigma_i}$ (т. о., небольшие отклонения могут повлиять на результат, если принадлежат измерению с малым разбросом частот) [Argamon 2008].

Одной из модификаций Дельты является "Дельта Эдера" (Eder's Delta), предполагающая увеличенный вес частоупотребимых слов и предложенная учёным М. Эдером для исследования текстов на языках синтетического типа [Eder et al 2016]:

$$\Delta_E^{(n)}(D, D') = \frac{1}{n}\sum_{i=1}^{n} \left( \frac{|f_i(D) - f_i(D')|}{\sigma_i} \cdot \frac{n - i + 1}{n} \right). \quad (5)$$

Вкупе с Дельтой часто применяются алгоритмы кластеризации, позволяющие получить результат в виде дендрограммы. В настоящем исследовании был применён алгоритм кластеризации, формирующий кластеры согласно методу Варда. Была проведена атрибуция эталонной выборки на произведениях И. Бунина и А. Куприна, которая показала, что произведения одного автора не объединялись с произведениями другого до тех пор, пока все его произведения не объединились друг с другом, что послужило основанием к применению методики к основной атрибуционной проблеме данной работы.

# 4 Материалы исследования

Априорные корпусы настоящего исследования были составлены из следующих произведений:

- “1001 день, или Новая Шахерезада” (1927), “Светлая личность” (1928), “Необыкновенные истории из жизни города Колоколамска” (1928), “Одноэтажная Америка” (1937) И. Ильфа и Е. Петрова;
- “Земля без солнца” (1925), “Красавица с острова Люлю” (1926), “Баклажаны” (1927), “Человек без площади” (1927), “Женитьба Мечтателева” (1927), “Рука бога Му-га-ша” (1928), “Внук золотого короля” (1928), “Жизнеописание Степана Александровича Лососинова” (1928), “Шестьдесят братьев” (1929) С. Заяицкого;
- “Собачье сердце” (1925), “Копыто инженера” (1929), “Тайному другу” (1929), “Мастер и Маргарита” (1940) М. Булгакова,
- главы атрибутируемых романов “Двенадцать стульев” и “Золотой теленок”,
- записные книжки И. Ильфа (1925–1937).

Построенные атрибуционные гипотезы формулируются следующим образом.

1. Нулевая гипотеза $H_0$ соответствует официальным данным и гласит, что оба романа были написаны Ильфом и Петровым.
2. Альтернативная гипотеза $H_1$ гласит, что романы были написаны М. Булгаковым.
3. Альтернативная гипотеза $H_2$ гласит, что романы были написаны С. Заяицким.

Предобработка данных для процедуры анализа включила: деление атрибутируемых произведений на главы и исключение из частотного словаря имён собственных. Было проведено два набора испытаний: первый набор – без использования лемматизации, второй – с использованием автоматического лемматизатора MyStem 3.1 (tech.yandex.ru/mystem).

В обоих случаях анализ проводился с объёмами частотного словаря в 250, 500, 750 и 1000 слов с использованием мер межтекстового расстояния классическая Дельта, Дельта Эдера и косинус-Дельта. Для вычислений и построения дендрограмм использовался пакет для вычислительного анализа текста **stylo**, разработанный для программной среды R в 2013 г. международной группой учёных М. Эдером, Я. Рубицким и М. Кестемонтом [Eder et al. 2016].

# 5 Обсуждение результатов

Результаты, полученные в ходе первого набора испытаний, мы качественно поделили на группы, а затем выделили превалирующие (рис. 1, 2 и 3). Вторая и третья группы поддерживают нулевую гипотезу (третья группа дендрограмм подчёркивает связь атрибутируемых романов с неоконченной повестью Булгакова “Тайному другу”), остальные шесть результатов (первая группа и выбросы) не позволяют соотнести дилогию ни с одним из авторов.[1]

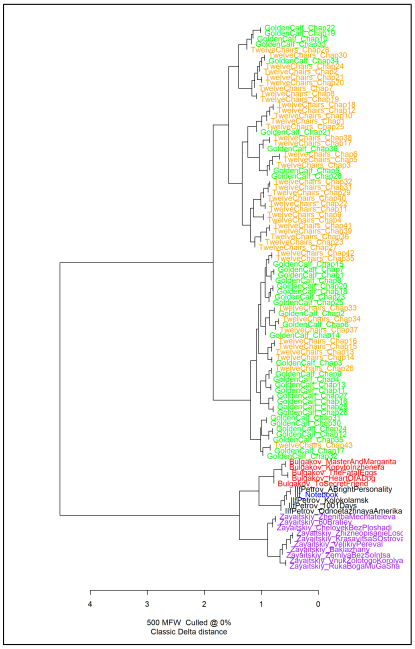


Рис. 1: Первая группа: 4 из 12 результатов.

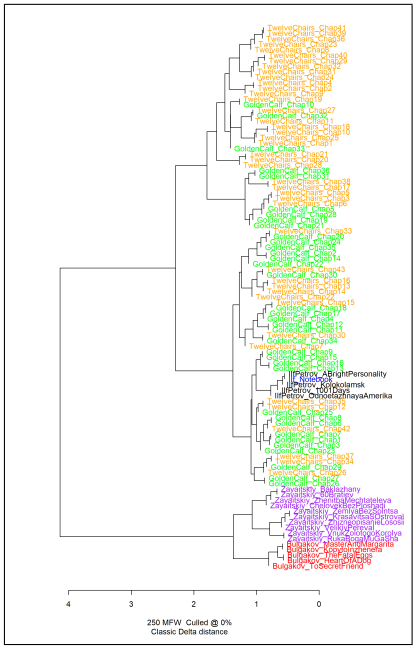


Рис. 2: Вторая группа: 3 из 12 результатов.


Рис. 3: Третья группа: 3 из 12 результатов.

Результаты, полученные после применения лемматизатора, также образовали три качественные группы. В первую группу вошло шесть дендрограмм, построенных при анализе с мерой классическая Дельта, объём словаря – 250 и 500 слов; с мерой Дельта Эдера, объём словаря – 250, 500 и 750 слов, а также мерой косинус-Дельта, объём словаря – 1000 слов (рис. 4). В данном случае произведения Ильфа и Петрова объединяются в кластер с главами романов. Произведения С. Заяицкого формируют кластер, также как и произведения М. Булгакова, затем эти кластеры совмещаются.

[1]В рисунках чёрным цветом обозначены произведения Ильфа и Петрова, красным – произведения М. Булгакова, фиолетовым – произведения С. Заяицкого, жёлтым – главы романа “Двенадцать стульев”, зелёным – главы романа “Золотой телёнок”, синим – записные книжки И. Ильфа.

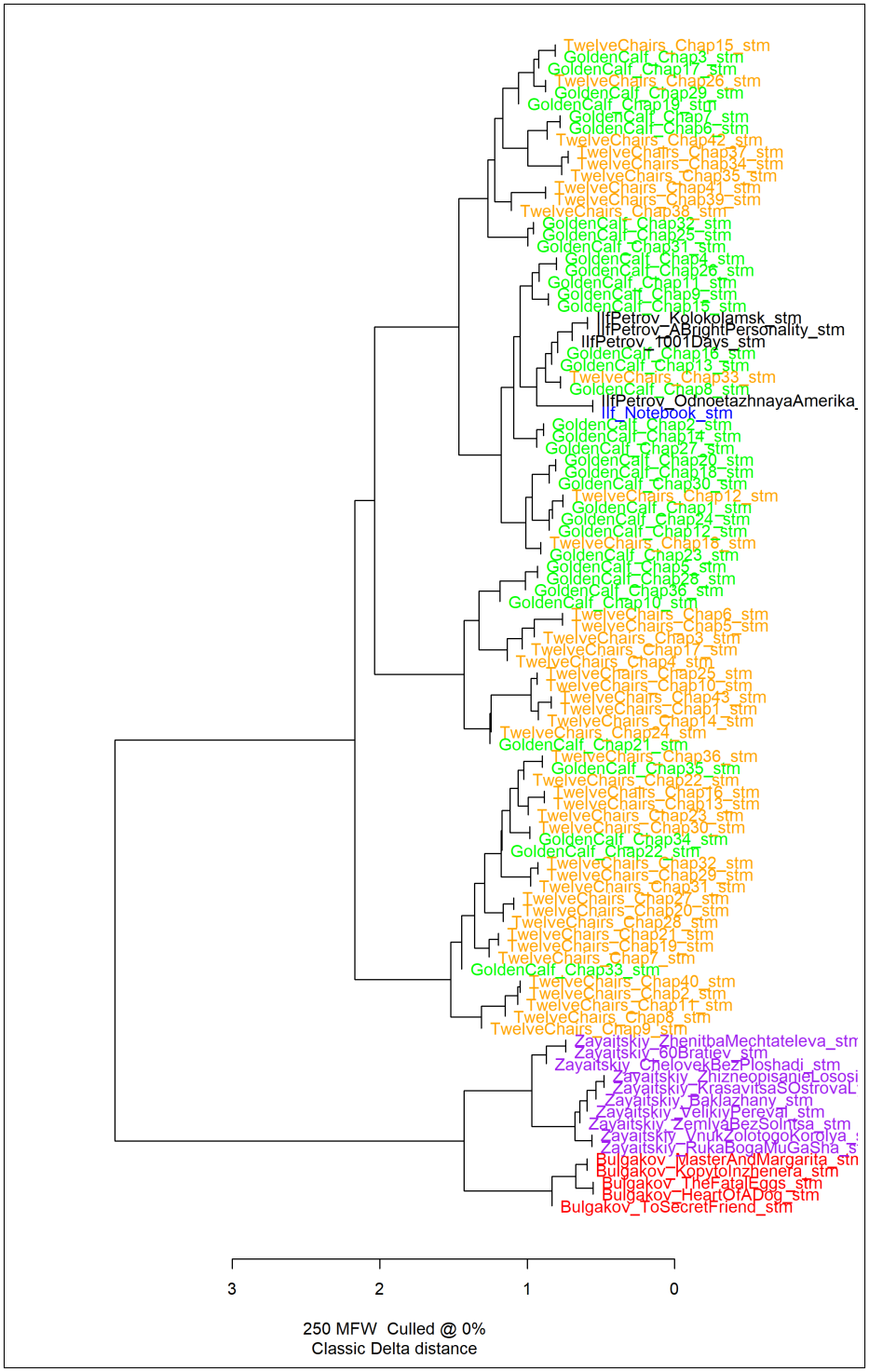


Рис. 4: Результат анализа с мерой классическая Дельта, объём словаря – 250 слов.

Кластер, содержащий произведения Ильфа и Петрова и главы атрибутируемых романов, объединяется с кластером, содержащим произведения других писателей, только на последнем шаге. Данный результат поддерживает нулевую гипотезу $H_0$.

Во вторую группу вошло три дендрограммы, построенные при анализе с мерой косинус-Дельта, объём словаря – 250, 500 и 750 слов (рис. 5).

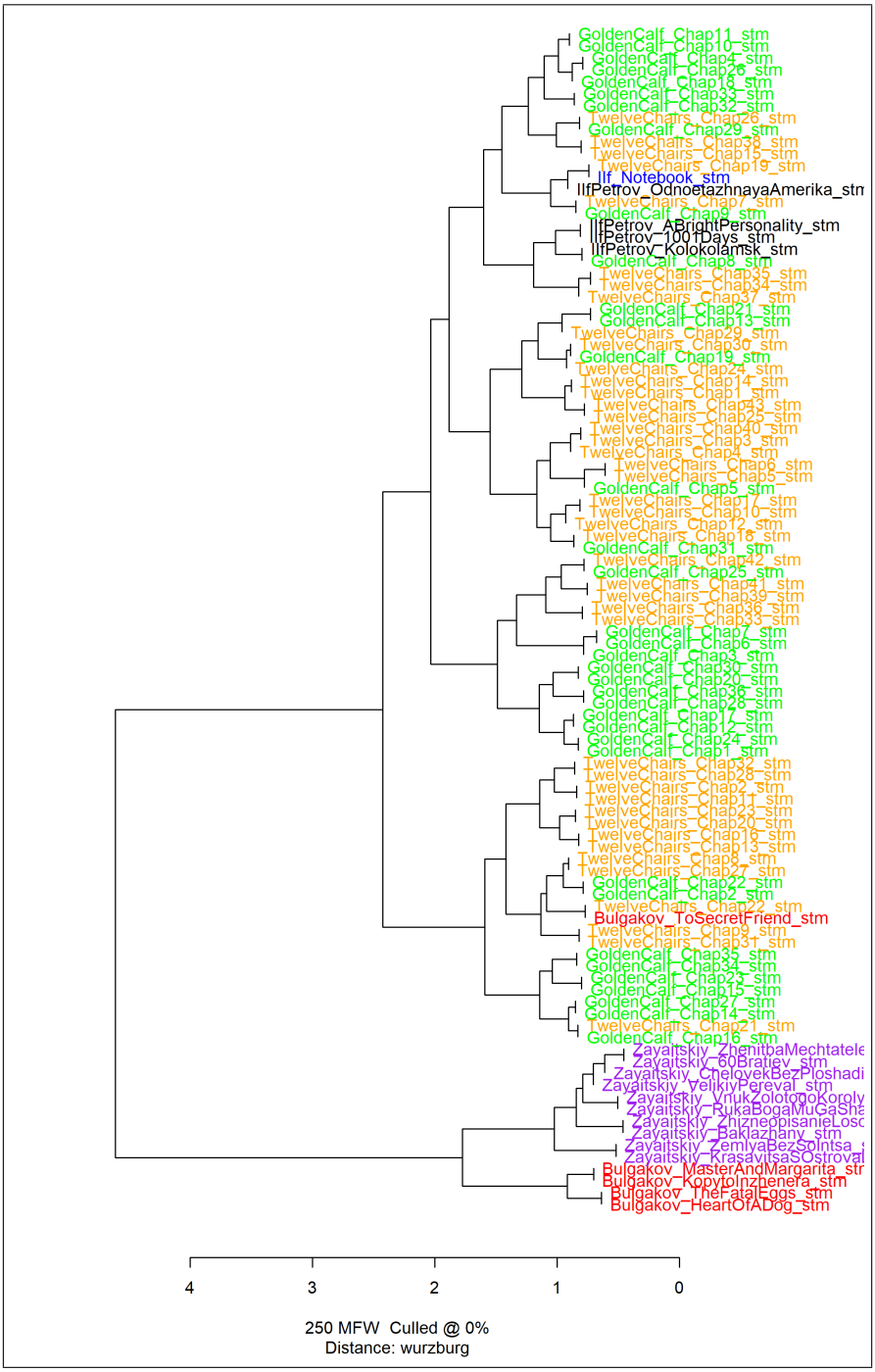


Рис. 5: Результат анализа с мерой косинус-Дельта, объём словаря – 250 слов.

Этот результат работы алгоритма свидетельствует, что повесть “Тайному другу” стилистически демонстрирует близость к романам “Двенадцать стульев” и “Золотой телёнок”, а также к произведениям Ильфа и Петрова. Заметим, что это сходство определяется только с помощью меры косинус-Дельта. В целом, тем не менее, данный результат также поддерживает нулевую гипотезу $H_0$.

В третью группу вошло три дендрограммы, построенные при анализе с мерой классическая Дельта, объём словаря – 750 и 1000 слов, а также мерой Дельта Эдера, объём словаря – 1000 слов (рис. 6).

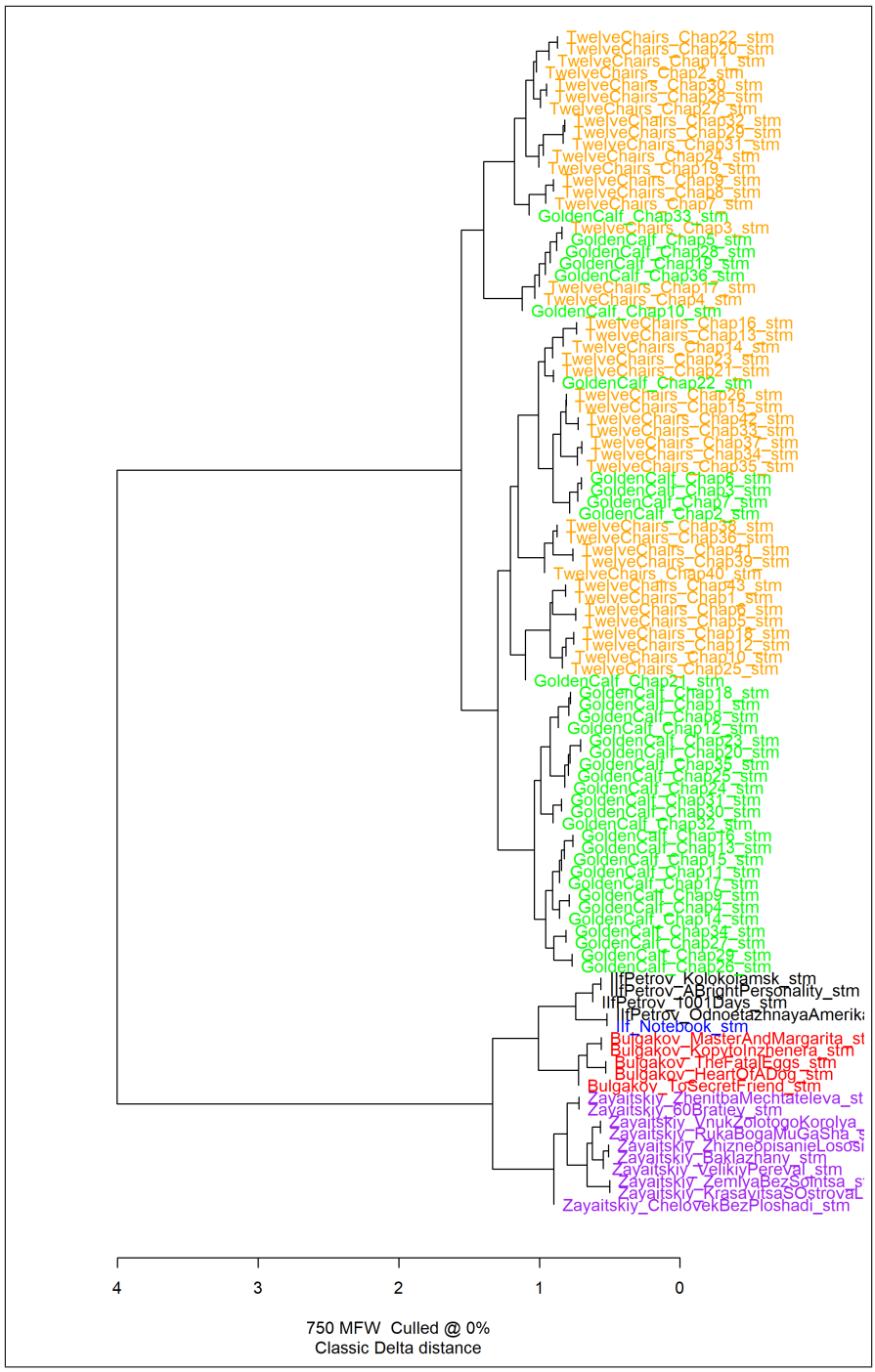


Рис. 6: Результат анализа с мерой классическая Дельта, объём словаря – 750 слов.

Данные результаты подчёркивают стилистическую однородность дилогии, однако не позволяют сделать выбор в пользу одной из атрибуционных гипотез.

Итак, в девяти из двенадцати попыток результат поддерживает нулевую гипотезу, в трёх попытках из двенадцати сформировать вывод не получилось. По результатам провер-

ки гипотез $H_0$ и $H_1$ гипотеза $H_0$ представляется наиболее вероятной, а гипотезы $H_1, H_2$ отвергаются.

# 6 Выводы

Изучение Дельты Бёрроуза и её модификаций позволяет заключить, что это – принципиально простой и в некоторых случаях эффективный набор инструментов, позволяющий исследователю подкреплять или опровергать атрибуционные гипотезы, рождающиеся в процессе стилистического и историко-филологического этапов исследования.

По итогам исследования можно сделать некоторые выводы о предполагаемом авторстве романов и обстоятельствах их написания. Полученные нами результаты свидетельствуют, что с точки зрения частот употребления слов дилогия близка к произведениям И. Ильфа и Е. Петрова, а также намечают связь между дилогией и произведениями М. Булгакова — скорее всего, можно говорить о возможном вкладе писателя в работу над романами. Предстоит дальнейшее тестирование Дельты Бёрроуза и сравнение эффективности предлагаемого инструментария с традиционными стилеметрическими квантитативными методами.

# Список литературы


[Amlinski 2013] Amlinski I. (2013) 12 stul'ev ot Mihaila Bulgakova.["12 Chairs" from M. Bulgakov] M.: Izd-vo "Kirschner Verlag". – 328 s. (In Russian) = Амлински И. 12 стульев от Михаила Булгакова. М.: Изд-во "Kirschner Verlag", 2013. – 328 с.

[Argamon 2008] Argamon S. Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. // Literary and Linguistic Computing Vol. 23, No. 2. Pp. 131–147.

[Burrows 2002] Burrows J.(2002) 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship //Literary and Linguistic Computing, Vol. 17, No. 3. Pp. 267-287

[Eder et al. 2016] Eder M., Rybicki J., Kestemont M. (2016) Stylometry with R: A package for computational text analysis //The R Journal, Vol. 8, No. 1. Pp. 107-121

[Evert et al. 2017] Evert St., Proisl Th., Jannidis F., Reger Is., Pielström St., Schöch Chr., Vitt Th. (2017) Understanding and explaining Delta measures for authorship attribution//Digital Scholarship in the Humanities, Vol. 32, Issue 2, 2017. Pp. 114–116

[Grieve 2007] Grieve, J. (2007) Quantitative Authorship Attribution:An Evaluation of Techniques//Literary and Linguistic Computing, Vol. 22, No. 3, Pp. 251-270

[Jannidis et al. 2015] Jannidis F., Pielstrom St., Schoch C., Vitt Th. (2015) Improving Burrows' Delta – An empirical evaluation of text distance measures.//Digital Humanities Conference, 2015, Sydney.

[Juola 2009] Juola P. (2009) JGAAP: A System for Comparative Evaluation of Authorship Attribution//JDHCS 2009, Vol. 1 No. 1

[Hoover 2005] Hoover D. (2005) Delta, Delta Prime, and Modern American Poetry: Authorship Attribution Theory and Method //Proceedings of the 2005 ALLC/ACH Conference

[Hoover 2004] Hoover D. L. (2004) Testing Burrows's Delta //Literary and Linguistic Computing, Vol. 19, No. 4. Pp. 453–475

[Kozaroveckii 2013] Kozaroveckii V. (2013) Moskovskie baranki i odesskie bubliki. //Literaturnaya Rossiya, № 41, 2013. (In Russian) = Козаровецкий В. Московские баранки и одесские бублики.[Moscow bagels and bagels from Odessa] //Литературная Россия, № 41, 2013. – С. 48-50

[Petrov 2001] Petrov E. P. (2001) Moi drug Il'f / Sostavlenie i kommentarii A. I. Il'f.[My friend Il'f ] M.: Tekst. – 351 s. (In Russian) = Петров Е. П. Мой друг Ильф / Составление и комментарии А. И. Ильф. М.: Текст, 2001. – 351 с.

[Rybicki et al. 2011] Rybicki J., Eder M. (2011) Deeper Delta across genres and languages: do we really need the most frequent words? //Literary and Linguistic Computing, Vol. 26, No 3. Pp. 315-321

[Savoy 2010] Savoy, J. (2010) Lexical analysis of US political speeches // Journal of Quantitative Linguistics, Vol. 17, No. 2. Pp. 123–141

[Savoy 2015] Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages //Quantitative Linguistics Vol. 19, Issue 2. Pp. 132-161

[Trishin 2017] Trishin V. N. (2017) Kto avtor "Dvenadcati stul'ev" i "Zolotogo telenka"? [Who is the author of "12 Chairs" and "Golden golden Calf"]//Simvol nauki [Science Symbol], № 10, 2017. S. 48-50. (In Russian) = Тришин В. Н. Кто автор "Двенадцати стульев" и "Золотого телёнка"? //Символ науки, № 10, 2017. С. 48-50