

РАЗДЕЛ II. ПРИКЛАДНАЯ ЛИНГВИСТИКА SECTION II. APPLIED LINGUISTICS

UDC 81'42

DOI: 10.18413/2313-8912-2026-12-3-0-3

Vladimir V. Bochkarev¹ 
Anton S. Afanasev² 
Rinat A. Bakirov³ 

Spearman's rank distance as a tool for stylometric attribution: a comparative analysis on heterogeneous corpora

¹ Kazan (Volga Region) Federal University,
18 Kremlyovskaya St., Kazan, 420008, Russia
E-mail: vbochkarev@mail.ru
ORCID: 0000-0001-8792-1491

² Kazan (Volga Region) Federal University,
18 Kremlyovskaya St., Kazan, 420008, Russia
E-mail: a.s.afanasyev@mail.ru
ORCID: 0000-0003-0571-833X

³ Kazan (Volga Region) Federal University,
18 Kremlyovskaya St., Kazan, 420008, Russia
Institute of Russian Literature (Pushkin House)
4 Makarov Emb., Saint Petersburg, 199034, Russia
E-mail: r1nt@ya.ru
ORCID: 0000-0001-6693-1076

*Статья поступила 17 июня 2026 г.; принята 15 сентября 2026 г.;
опубликована 30 сентября 2026 г.*

Abstract: This paper presents an empirical comparison of a metric based on Spearman's rank correlation coefficient with traditional and relatively new distance measures (Burrows's Delta, Cosine Delta, Labbé distance, Soergel distance, and Jensen-Shannon divergence, the latter receiving particular attention) used in stylometric text attribution. The relevance of the study lies in the need to search for and identify a stylometric tool (a distance measure and its "settings") that is robust to within-author variability and feature space sparsity. The research problem stems from methodological uncertainty in choosing an optimal distance measure for stylometric attribution. Despite the development of distance-based approaches, the effectiveness of metrics critically depends on corpus structure, the length of the analyzed fragments, and feature space parameters. In the absence of a single reliable metric for attributing complex text corpora, this study for the first time substantiates the use of Spearman's rank distance as a valid tool for stylometric analysis. The empirical basis of the study comprises four text corpora heterogeneous in terms of time, genre, and language (49 authors in total, 899 literary texts). The study included unified text preprocessing (tokenization, lowercasing, wordform-level analysis) and systematic parameter tuning for the six metrics. For each combination (author, metric, fragment length), the optimal configuration of parameters (number of frequent words; absence/presence of standardization) was determined,

providing the maximum true positive rate in attribution using the k-NN method, $k=1$. The results show that Spearman's rank distance demonstrates the greatest robustness and universality, maintaining high accuracy under genre heterogeneity, variability in text length, and the presence of stylistic anomalies. It is established that Spearman's rank distance systematically outperforms frequency-based and entropy-based metrics in robustness under conditions of genre heterogeneity and can serve as a reliable tool for stylometric attribution.

Keywords: Quantitative literary studies; Corpus linguistics; Stylometry; Distance measures; Spearman's rank distance; Jensen-Shannon divergence

Acknowledgments: The study was supported by the Russian Science Foundation grant No. 24-18-00570, <https://rscf.ru/project/24-18-00570/>.

How to cite: Bochkarev, V. V., Afanasev, A. S., Bakirov, R. A. (2026). Spearman's rank distance as a tool for stylometric attribution: a comparative analysis on heterogeneous corpora, *Research Result. Theoretical and Applied Linguistics*, 12 (3), 52–79. DOI: 10.18413/2313-8912-2026-12-3-0-3

UDC 81'42

DOI: 10.18413/2313-8912-2026-12-3-0-3

Бочкарев В. В.¹
Афанасьев А. С.²
Бакиров Р. А.³

Ранговое расстояние Спирмена как инструмент
стилометрической атрибуции: сравнительный анализ
на разнородных корпусах

¹ Казанский (Приволжский) федеральный университет
ул. Кремлевская, 18, Казань, 420008, Россия
E-mail: vbochkarev@mail.ru
ORCID: 0000-0001-8792-1491

² Казанский (Приволжский) федеральный университет
ул. Кремлевская, 18, Казань, 420008, Россия
E-mail: a.s.afanasyev@mail.ru
ORCID: 0000-0003-0571-833X

³ Казанский (Приволжский) федеральный университет
ул. Кремлевская, 18, Казань, 420008, Россия
Институт русской литературы (Пушкинский Дом)
наб. Макарова, 4, Санкт-Петербург, 199034, Россия
E-mail: r1nt@ya.ru
ORCID: 0000-0001-6693-1076

Received 17 June 2026; accepted 15 September 2026; published 30 September 2026

Аннотация: В статье представлено эмпирическое сравнение метрики, основанной на расчете коэффициента ранговой корреляции Спирмена, с традиционными и относительно новыми мерами расстояния (Дельта Берроуза, косинусное расстояние, расстояние Лаббе, расстояние Зергеля, дивергенция Йенсена-Шеннона, которой уделено особое внимание), используемых при стилометрической атрибуции текста. Актуальность исследования заключается в необходимости поиска и обнаружения стилометрического инструмента (меры

расстояния и ее «настроек»), устойчивого к внутриавторской вариативности и разреженности признакового пространства. Проблема исследования обусловлена методологической неопределенностью в выборе оптимальной меры расстояния для стилометрической атрибуции. Несмотря на развитие distance-based подходов эффективность метрик критически зависит от структуры корпуса, длины анализируемых фрагментов и параметров признакового пространства. В условиях отсутствия единой надежной метрики для атрибуции сложных текстовых массивов данное исследование впервые обосновывает использование расстояния Спирмена как валидного инструмента стилометрического анализа. Эмпирической базой исследования стали четыре разнородных по времени, жанрам и языку написания корпуса текстов (всего 49 авторов, 899 художественных текстов). Исследование включало унифицированную предобработку текстов (токенизация, приведение к нижнему регистру, анализ на уровне словоформ) и систематический перебор параметров для шести метрик. Для каждой комбинации (автор, метрика, длина фрагмента) определялась оптимальная конфигурация параметров (количество частотных слов; отсутствие/наличие стандартизации), обеспечивающая максимальную долю истинно положительных результатов при атрибуции методом k-NN, k=1. Результаты показали, что расстояние Спирмена демонстрирует наибольшую робастность и универсальность, сохраняя высокую точность при жанровой неоднородности, вариативности объема текстов и наличии стилистических аномалий. Установлено, что ранговое расстояние Спирмена систематически превосходит частотные и энтропийные метрики по робастности в условиях жанровой гетерогенности и может служить надежным инструментом стилометрической атрибуции.

Ключевые слова: Квантитативное литературоведение; Корпусная лингвистика; Стилметрия; Меры расстояния; Ранговое расстояние Спирмена; Дивергенция Йенсена-Шеннона

Благодарности: Исследование выполнено при поддержке гранта Российского научного фонда № 24-18-00570, <https://rscf.ru/project/24-18-00570/>.

Информация для цитирования: Бочкарев В. В., Афанасьев А. С., Бакиров Р. А. Ранговое расстояние Спирмена как инструмент стилометрической атрибуции: сравнительный анализ на разнородных корпусах // Вопросы теоретической и прикладной лингвистики. 2026. Т. 12. № 3. С. 52–79. DOI: 10.18413/2313-8912-2026-12-3-0-3

Introduction. The systematic use of stylometric text analysis dates back to the second half of the 19th century. Without delving into a detailed historical excursus (referring the reader on this matter to works by Grzybek (Grzybek, 2014), Martynenko (Martynenko, 2014; Martynenko, 2015), Kovalev (Kovalev, 2024)), we move directly to the beginning of the 21st century and recall that a turning point in the development of stylometric analysis was the publication of J. Burrows' now-classic work (Burrows, 2002), in which he proposed the Delta method for determining stylistic proximity between texts

and their attribution. Burrows' Delta (otherwise known as standardized Manhattan distance) is still one of the most popular distance measures in stylometric analyses (Skorinkin and Orekhov, 2023; Dviniatin and Kovalev, 2024; O'Sullivan, 2025; Mikros et al 2026), but is periodically (and, in our view, justifiably) subjected to constructive criticism and attempts at improvement (see, e.g., Hoover, 2004; Argamon, 2008; Smith and Aldridge, 2011; Rybicki and Eder, 2011), which stimulates the search for and development of alternative distance measures

that could overcome certain limitations of the classical Delta.

Currently, alongside standardized Manhattan distance, various metrics from the group of distance-based approaches are widely used in stylometry: unstandardized Manhattan distance, Euclidean distance, Cosine Delta, Canberra distance, Eder's Delta, Min-Max, and some others. Given this diversity of metrics, a natural question arises: which (or perhaps which one) of them demonstrates the greatest reliability and stability when solving stylometric problems?

Several research works compare distance measures used in stylometric studies.

A research team led by F. Jannidis compared 14 different distance measures (Jannidis et al., 2015) and concluded that Cosine Delta significantly surpasses the others in terms of accuracy and stability. A later article explains the advantage of cosine similarity over other distance measures: the process of normalising vector length mitigates the impact of deviation from the norm (Evert et al., 2017).

M. Kocher and J. Savoy (Kocher and Savoy, 2019), using French and English literature, compared Tanimoto, Matusita, Clark, and Cosine Delta for the authorship attribution task. It was found that the best results were achieved using 4-6-character letter n-grams as features and the Tanimoto, Matusita, and Clark distance metrics, which systematically outperformed Cosine Delta in accuracy.

D. Stanikunas, J. Mandravickaite, and T. Krilavicius (Stanikunas et al., 2017), analyzing Lithuanian texts (transcripts of parliamentary debates), compared 15 distance and similarity metrics, including Burrows's Delta, Eder's Delta, Eder's Simple Delta, Binomial Index, and Cosine Delta. The researchers concluded that Eder's Simple Delta demonstrates consistently high performance and can be used for general stylometric analysis.

O. Alieva (Alieva, 2024) conducts a comparative analysis of some of the metrics listed above and other distance measures that

have also been sporadically used for stylometric investigations and have proven their worth: Labbé distance (Manhattan distance between full vectors of relative frequencies), Tanimoto distance (which, in our interpretation, O. Alieva uses to describe Soergel distance), and Jeffrey's divergence (symmetrized Kullback-Leibler divergence). According to Alieva's experiment, "when classifying ancient Greek texts, the best results are achieved using cosine similarity on standardized frequency values, Tanimoto distance on relative frequency values, and Labbé distance on absolute frequency values" (Alieva, 2024: 23).

As we can see, a single universally successful metric does not exist: all studies confirm that the effectiveness of metrics strongly depends on language, genre, and frequency normalization method, and no universal solution has been proposed.

Conceptually significant for us is the study by D. Pronin and E. Kazartsev (Pronin and Kazartsev, 2026), who, using literary corpora in English, German, French, and Russian, compared classical metrics of the Delta family (Burrows's Delta, Cosine Delta, Eder's Delta) with two new measures – Rank-Turbulence Delta and Jensen-Shannon Delta. These metrics are based on the interpretation of standardized frequency vectors as probabilistic or rank distributions, making it possible to apply divergence measures from information theory and complex systems analysis. The article proves that Rank-Turbulence Delta demonstrates accuracy comparable to Cosine Delta, and Jensen-Shannon Delta systematically matches or surpasses Burrows's Delta in accuracy. One also cannot ignore an earlier work (Grieve, 2007), where the author considers rank methods as a robust alternative to Euclidean and Manhattan distances when working with sparse data.

Although the potential of Jensen-Shannon Delta (in our study, we will call the Jensen-Shannon Delta the Jensen-Shannon Divergence – JSD) has begun to be studied in stylometry (Pronin and Kazartsev, 2026), the

distance based on Spearman's rank correlation coefficient has not yet appeared in comparative experiments. Therefore, the aim of this study is to conduct an empirical comparison of distance metrics based on Spearman's rank correlation and JSD with metrics traditional for stylometric research (Manhattan distance, Cosine Delta, Labbé distance, and Soergel distance) under conditions of genre heterogeneity, variability in text size, and changes in feature space size.

Materials and Methods.

Corpus Materials.

For our experiment, we used four text corpora.

The "Greek" corpus includes 57 texts by 17¹ ancient Greek authors and was borrowed from the study by O. Alieva (Alieva, 2024). The composition of the corpus and source data are available in an open repository:

<https://github.com/locusclassicus/compareDis.t>.

The "Sholokhov" and "Rowling" corpora were formed to reproduce two classical stylometric cases: confirming M. Sholokhov as the author of the epic novel "And Quiet Flows the Don" and identifying J. Rowling under the pseudonym R. Galbraith.

The "Sholokhov" corpus is based on the corpus prepared by N. Velikanova and B. Orekhov (Velikanova and Orekhov, 2019: 79). We retained 11 authors but increased the number of works to 33 (M. Bulgakov: "The White Guard", "The Master and Margarita", "Heart of a Dog"; Vs. Ivanov: "Armoured Train 14-69", "Blue Sands", stories; F. Kryukov: "Group B", "Swell", "To the Source of Healings", "Mother", "Squall"; L. Leonov: "Badgers", "The Thief", "The End of a Small Man"; N. Ostrovsky: "How the Steel Was Tempered", "Born of the Storm"; A. Platonov: "The Foundation Pit",

"Chevengur", "The Juvenile Sea"; V. Seversky: "Don on Crutches"; A. Serafimovich: "The Iron Flood", stories; A. Fadeev: "The Rout", "The Spill", "The Birth of the Amgun Regiment"; D. Furmanov: "Red Landing", "Mutiny", "Chapaev"; M. Sholokhov: "Don Stories", "They Fought for Their Country", "Virgin Soil Uplifted", "The Fate of a Man", "And Quiet Flows the Don")².

The 11 authors included in the "Rowling" corpus were taken by us from the article by B. Orekhov (Orekhov, 2024: 2-4), but, as with "Sholokhov", the number of works was increased to enhance sample representativeness. Our corpus contains 54 literary texts: G. M. Burnett: "The Disappearance of Adèle Bedeau", "His Bloody Project", "Case Study", "Benbecula"; H. Coben: "Deal Breaker", "Drop Shot", "Fade Away", "One False Move", "Gone for Good", "No Second Chance", "Tell No One"; T. French: "In the Woods", "The Likeness", "Faithful Place", "Broken Harbour", "The Secret Place", "The Trespasser", "The Searcher", "The Hunter", "The Witch Elm"; R. Galbraith: "The Cuckoo's Calling", "The Silkworm", "Career of Evil", "Lethal White", "Troubled Blood", "The Ink Black Heart", "The Running Grave", "The Hallmarked Man"; C. Lewis: "The Lion, the Witch and the Wardrobe", "Prince Caspian: The Return to Narnia", "The Voyage of the Dawn Treader", "The Silver Chair", "The Horse and His Boy", "The Magician's Nephew", "The Last Battle"; J. Rowling: "Harry Potter and the Philosopher's Stone", "Harry Potter and the Chamber of Secrets", "Harry Potter and the Prisoner of Azkaban", "Harry Potter and the Goblet of Fire", "Harry Potter and the Order of the Phoenix", "Harry Potter and the Half-Blood Prince", "Harry Potter and the Deathly Hallows", "The Casual Vacancy"; B. Sanderson: "Mistborn: The Final Empire", "The Well of Ascension", "The Hero of

¹ The graphs represent 16 authors of the corpus: Antisthenes consistently showed NaN in our calculations.

² The curve for V. Seversky is not presented in the graphs because his oeuvre is represented by only one novel, "Don on Crutches".

Agas”; J. Stroud: “The Amulet of Samarkand”; “The Golem’s Eye”; S. Turton: “The Seven Deaths of Evelyn Hardcastle”, “The Devil and the Dark Water”, “The Last Murder at the End of the World”; J. Tolkien: “The Lord of the Rings”; F. Warrington: “A Blackbird in Silver”, “The Dark Arts of Blood”).

Regarding these three corpora, it was important for us to see how Spearman’s rank distance (hereinafter referred to as Spearman distance) and JSD perform on existing data.

The “Science fiction and rustic prose” (“SF&RP”) corpus is a hyper-corpus combining 755 works by 10 authors of Soviet science fiction and rustic prose: A. Belyaev – 46, I. Efremov – 27, A. and B. Strugatsky – 67, K. Bulychyev – 210, A. Kazantsev – 90; V. Astafyev – 97, V. Belov – 17, B. Mozhaev – 38, V. Rasputin – 25, V. Shukshin – 138. This corpus was used to assess the robustness of metrics to very large, cross-genre collections. There is an imbalance in the representation of texts by different authors (from 17 to 210 texts). This factor was considered when interpreting the results and is discussed in the limitations section.

The total vocabulary of the “Greek” corpus was 82955 words, “Rowling” – 60294, “Sholokhov” – 184686, and “SF&RP” – 454049.

Methodology.

Having analyzed previous work, we selected metrics that showed the best results in accuracy and stability. These are Labbé and Soergel distances, Cosine Delta, Burrows’ Delta (Manhattan distance with preliminary standardization of frequencies), and Jensen-Shannon divergence (hereinafter referred to as JSD), to which we will pay special attention in our study. We will compare these with Spearman distance.

These 6 metrics can be divided into three fundamentally different groups based on the type of information used. The first and most representative group consists of traditionally used metrics – Manhattan distance, Cosine Delta and Labbé distance. These approaches work directly with relative

frequencies of wordforms, measuring linear or angular discrepancies between frequency vectors. The second group comprises the probabilistic and information-theoretic measures of the Soergel distance and the JSD. They interpret frequency vectors as probability distributions and assess their divergence through entropy characteristics, making them sensitive to the shape of the distribution, not just the magnitude of frequencies. Finally, the third, separate group is represented by Spearman’s rank metric. Its fundamental difference lies in operating not with the frequencies themselves, but with word ranks – ordinal numbers in a sorted list. This fundamentally changes the stability of the metric: the rank approach is insensitive to extreme values and outliers and does not assume normality of distribution.

We will not discuss the traditional metrics that have been repeatedly used for stylometric analysis in detail, but will refer to existing research. The application of standardized Manhattan distance is described in detail by J. Burrows (Burrows, 2002) and critically evaluated in (Argamon, 2008). Cosine Delta (among other metrics) was the subject of analytical review in (Evert et al., 2017), and regarding Labbé distance, we refer to the work of the author himself (Labbé and Labbé, 2006) (the results of its application can be found in (Bortolato, 2016)). The Labbé distance is the Manhattan distance between the full vectors of relative frequencies (without employing standardization) of all words occurring in the compared texts.

We will discuss Spearman distance, JSD, and Soergel distance in more detail.

In statistics, Spearman’s rank correlation coefficient (ρ) is used to assess the strength of association between two variables (Spearman, 1904). Unlike the Pearson correlation coefficient, it is not the values of two quantities x_i and y_i that are compared directly, but their ranks ri and ri , that is, the numbers in a list sorted in ascending (or descending) order. The coefficient is then calculated by the formula:

$$\rho = 1 - \frac{6}{n(n-1)^2} \sum_{i=1}^n (r_i^x - r_i^y)^2$$

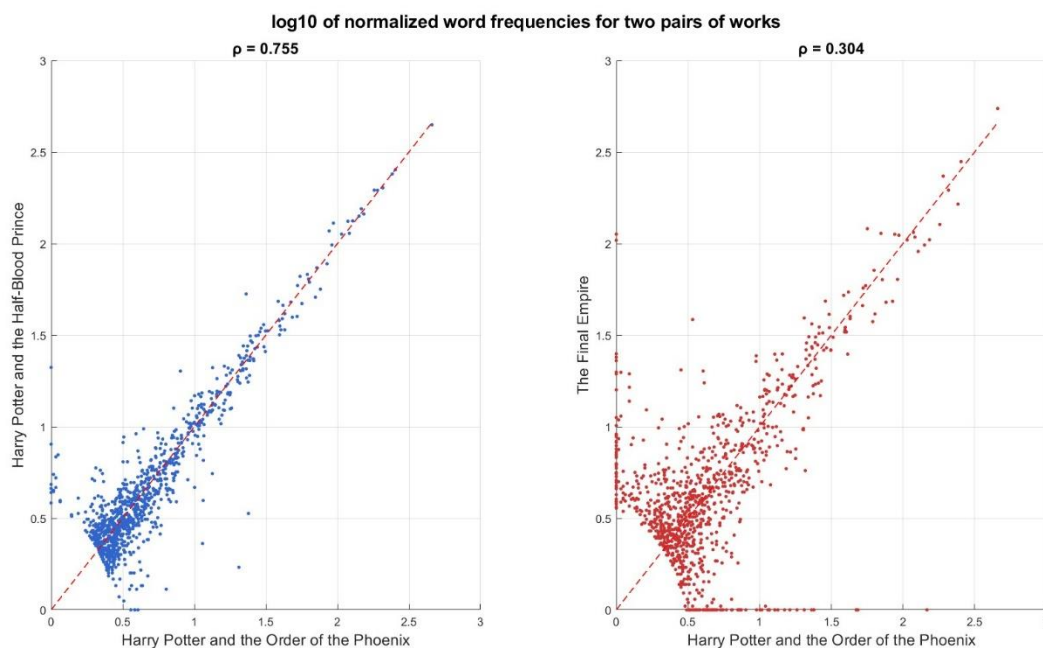
While the Pearson correlation coefficient and Cosine Delta provide an adequate estimate of the degree of similarity/difference only when the distribution of fluctuations in the data is close to normal, Spearman's rank correlation coefficient can be effectively applied to cases with a wide variety of probability distributions. This is extremely important for stylometric tasks, as the specifics of an author's word usage in their works is a complex and currently unexplained phenomenon arising from the intersection of the author's unique psychology, cognitive abilities, social experience, and conscious aesthetic intentions.

To illustrate the operation of Spearman's rank correlation, let us take two

pairs of texts. The first pair – J. Rowling's novels "Harry Potter and the Order of the Phoenix" and "Harry Potter and the Half-Blood Prince"; the second pair – J. Rowling's "Harry Potter and the Order of the Phoenix" and B. Sanderson's "The Final Empire". With 1000 features, Spearman's coefficient between word frequencies in the two novels by J. Rowling was $\rho = 0.755$. In the left graph (Figure 1), a clear linear trend is observed: the points are located near and along the line $y = x$. In the right graph, comparing texts by different authors, the coefficient drops to $\rho = 0.304$. The points form a sparse cloud, significantly deviating from the bisector, indicating weak agreement of lexical profiles. Such a contrasting visual divergence clearly demonstrates the discriminatory ability of the metric: it clearly records stylistic closeness within an author's corpus and just as confidently marks inter-author differences.

Figure 1. Spearman's rank correlation for Rowling/Rowling and Rowling / Sanderson novels

Рисунок 1. Ранговая корреляция Спирмена романов Дж. Роулинг / Дж. Роулинг и Дж. Роулинг / Б. Сандерсон



For stylometric analysis, it is more convenient to use not the Spearman coefficient ρ itself, but the value $1 - \rho$ which we will call Spearman distance. This

transformation is driven by the desire to uniformly describe all metrics used in stylometric analysis.

Jensen-Shannon divergence (JSD) (Lin, 1991; Endres and Schindelin, 2003) is used as a measure of dissimilarity between two probability distributions. Let two probability vectors $P = (p_1, p_2 \dots p_n)$ and

$$JSD(P, Q) = H\left(\frac{P + Q}{2}\right) - \frac{1}{2}\{H(P) + H(Q)\}$$

Here $H(P)$ is the entropy of distribution P , defined as

$$H(P) = - \sum_i p_i \log_2 p_i$$

JSD values range from 0 to 1. Values close to 0 indicate a high degree of similarity between the compared distributions, while values close to 1 indicate large differences. A successful example of using JSD within stylometric analysis can be found in (Arefin et al., 2014). Jensen-Shannon divergence has properties similar to symmetrized Kullback-Leibler divergence, also used in stylometric works (Savoy, 2020: 39-42). However, in the practical application of Kullback-Leibler divergence, one might encounter a problem if one of the frequencies in the compared distributions is 0 – in this case, the divergence value becomes infinite. Jensen-Shannon divergence is free from this drawback.

The Soergel distance is a generalization of the Jaccard distance to the case of comparing two probability distributions (Levy et al., 2025; Willett et al., 1998). The Soergel distance is calculated using the following formula:

$$1 - \frac{\sum_i \min(p_i, q_i)}{\sum_i \max(p_i, q_i)}$$

In the literature, this quantity can also be found under the names Jaccard distance, Ruzicka distance, and Tanimoto distance. When the two compared distributions completely coincide, the sums in the numerator and denominator are equal, and the Soergel distance becomes 0. The largest distance value will be observed when, in each pair p_i, q_i , one of the quantities is 0 (this situation, in relation to stylometric tasks, can

$Q = (q_1, q_2 \dots q_n)$ be given. The values of Jensen-Shannon divergence are calculated using the following formula:

occur if the lexicons of the compared texts do not intersect). In this case, the sum in the numerator is 0, and the distance value is 1.

For Manhattan, Cosine Delta, and Spearman distance, we tested both modes – with standardization of word frequencies (since it helps mitigate the impact of outliers and differences in frequency scale) and without it. For JSD, Soergel distance, and Labbé distance, standardization is not applied: after standardization, numbers can be both positive and negative, which precludes the calculation/application of these metrics; Labbé distance by definition operates on full vectors of relative frequencies without transformations.

The JSD and Soergel distance calculations were carried out using user scripts in the MATLAB language. The Manhattan distance, Cosine Delta, Labbé distance and Spearman distance were calculated using the MATLAB system's built-in functions.

Text Preprocessing and Experimental Protocol.

All texts underwent a unified preprocessing procedure implemented in the MATLAB environment using functions from the Text Analytics Toolbox.

At the first stage, tokenization was performed, preserving punctuation markers and removing numeric sequences. Next, case folding was applied. Stop-word filtering was not used to preserve the full stylistic signal. The analysis was conducted at the wordform level. For the ancient Greek corpus, it was critically important to retain polytonic diacritics. Since the built-in linguistic models of the Text Analytics Toolbox do not support Ancient Greek, lemmatization and

morphological tagging, as in O. Alieva's work, were not performed.

For each studied corpus, its lexicon was compiled, and the total occurrence of each wordform in the corpus was counted. The analysis was based on the most frequent words (MFW), so the list of wordforms was then ranked in descending order of word frequency across the entire corpus. Regarding the choice of the optimal number of features, we repeated all calculations with the number of features varying in the range of 100-3200 in steps of 50.

The texts presented in the corpora have different lengths. To enable comparisons for all authors and works under the same conditions, all texts were divided into fragments of fixed length (with 50% overlap). The range of fragment lengths varied depending on the characteristics of the corpus: for "Greek" – from 625 to 7071 words, for "Sholokhov" – from 1250 to 40000 words, for "Rowling" – from 2500 to 20000 words, for "SF&RP" – from 2500 to 40000 words. Fragment extraction was performed on a logarithmic grid with a multiplier of 2 (i.e., each subsequent length is obtained by multiplying the previous one by ≈ 1.414), which ensures uniform coverage of the length scale on a logarithmic scale and increased sensitivity to changes in accuracy on short segments. After that, for each fragment, the relative frequencies of each wordform present in the corpus were calculated, and then, for each compared metric and number of features, the distances from each fragment to all other fragments of the same length were computed.

Authorship attribution was carried out using the k-nearest neighbors method (k-NN) with $k=1$.

The choice of $k=1$ is standard for stylometric studies using the nearest neighbor method (see, for example, Evert et al., 2017; Jannidis et al., 2015). At $k=1$, the attribution result is maximally sensitive to the choice of metric. Increasing k "blurs" the contribution of a specific metric, averaging the decision across several neighbors. For each fragment of a given work, the most similar fragment

(i.e., the one with the minimum distance to it) belonging to other works in the corpus was found. Attribution was considered correct if the most similar fragment belonged to another work by the same author.

The primary measure of attribution accuracy in this study was the true positive rate (TPR). It reflects the absolute effectiveness of the metric – what proportion of fragments of a given author, on average, the metric attributes correctly. TPR was calculated using the formula:

$$TPR = \frac{N_{true\ positive}}{N_{total}}$$

where $N_{true\ positive}$ is the number of fragments correctly attributed by the algorithm as belonging to the given author, and N_{total} is the total number of fragments of this author subjected to the attribution procedure.

The choice of TPR as the primary metric is driven by the objective of stylometric text attribution: determining what proportion of fragments from a given author are correctly attributed to them. TPR directly measures this value. Accuracy and F1-score are more appropriate for binary/multiclass classification problems with a fixed set of classes, while in stylometry, the number of classes (in this case, authors) varies from study to study.

Results and Discussion

Dependence of Attribution Accuracy on MFW and Text Length.

Before performing evaluation of the metrics across the entire data, it is advisable to establish how each metric responds to changes in feature space size and the length of the analyzed segments. Such analysis allows us to identify parameter ranges where metrics yield the highest accuracy.

For this experiment, we selected three multilingual corpora: "Greek", "Rowling", and "SF&RP", and within each corpus – one author: Plato, J. Rowling, and the Strugatsky brothers. This choice is motivated by the following.

Plato can be considered a benchmark of stylistic uniqueness, since the corpus of his works is homogeneous (all written within a narrow chronological period and belonging to the same genre – philosophical dialogue), and the works themselves possess highly unique stylistic and lexical markers (dialogic form, philosophical terminology, specific particles).

J. Rowling was chosen due to the genre binary nature of her corpus (fantasy, realistic novel) and the associated classical stylometric case (see more on this in section 4.4).

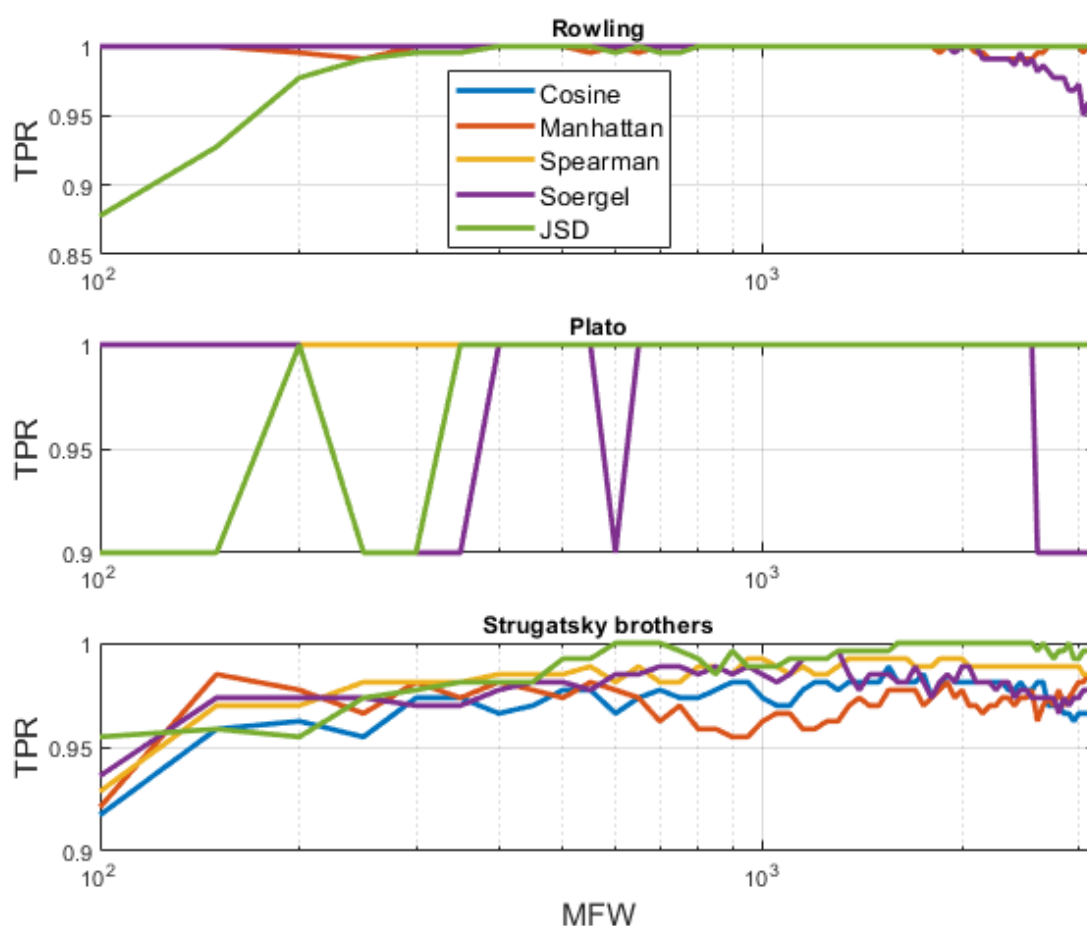
A. and B. Strugatsky represent great research interest, as they create certain difficulties for attribution. The complexity is explained by co-authorship (see stylometric investigations by V. Staroverova (Staroverova, 2025)) and chronological span,

broad genre diversity (novel, novella, short story, and their thematic modifications), as well as the inclusion of the Strugatsky brothers' works within the Russian-language hyper-corpus.

The experiment was conducted at a fixed fragment length of 10000 words, except for Plato, for whom the fragment length was set to 7071 words. The number of MFW features varied from 100 to 3200 in steps of 50. Three metrics – Spearman distance, Cosine Delta, and Manhattan distance – were used with preliminary frequency standardization, the others without standardization. The obtained TPR values for the 5 metrics (excluding Labbé distance) are presented in Figure 2.

Figure 2. Dependence of text attribution accuracy on MFW at a fixed text fragment length.

Рисунок 2. Зависимость точности атрибуции текста от MFW при фиксированной длине текстового фрагмента



First, we note that in many cases, TPR values reach 1.000. Any decision-making algorithm ensures decisions with some non-zero error probability (although under favorable conditions this probability can be extremely small). However, in a real experiment, we estimate TPR from finite sample sizes. The number of text fragments to be attributed may be insufficient for problematic cases to appear in the experiment (this especially applies to the region of long lengths, where the number of fragments becomes relatively small). Therefore, empirical TPR values in many cases are equal to 1.

For Rowling, all metrics achieve an accuracy of 1.000 already at MFW=400. Spearman distance, Cosine Delta, and JSD maintain this level across the entire MFW range. The Soergel distance shows a slight decrease in accuracy at MFW > 2500, likely due to noise introduced by rare words.

For Plato, three metrics – Spearman distance, Cosine Delta, and Manhattan distance – maintain an accuracy of 1.000 regardless of the size of the MFW list. A

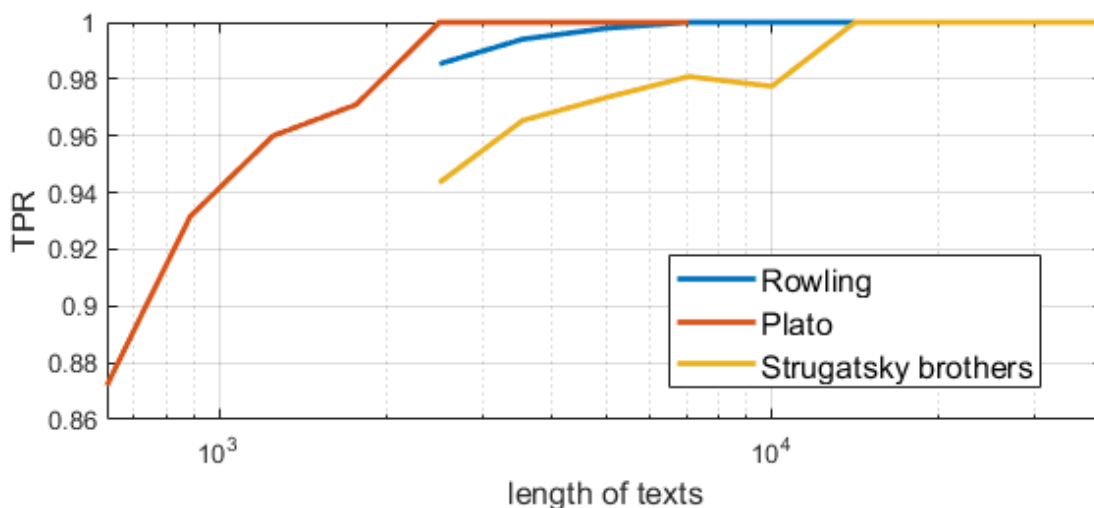
different picture emerges for the Soergel distance and the JSD: under certain feature configurations, their indicators drop to 0.900, which, in our opinion, reflects the increased sensitivity of these measures to the lexical composition of the sample.

The case of the Strugatsky brothers proved to be the most intriguing: JSD confidently reaches 1.000 at 600 and 1700-2200 features, surpassing its competitors. None of the other four metrics reached the maximum value on this case at the given fragment length (10000 words). This observation allows us to consider the entropy approach as a promising tool for texts with complex authorial structure and genre heterogeneity.

The Labbé distance works with the full vocabulary, and the MFW parameter is not set for it. However, its accuracy is sensitive to fragment length (see Figure 3). For the homogeneous Plato, 1.000 is recorded already at 2500 words, for the bi-genre Rowling – after 7000 words, and for the heterogeneous Strugatsky brothers, stabilization occurs only on segments exceeding 12000 words.

Figure 3. Dependence of text attribution accuracy on text fragment length for the Labbé distance

Рисунок 3. Зависимость точности атрибуции текста от длины текстового фрагмента для расстояния Лаббе



Comparison of Attribution Accuracy for Individual Authors.

The MFW and the application/non-application of frequency standardization are

among the tunable parameters, the choice of which is fixed by the researcher at the

experimental design stage. When comparing the effectiveness of different metrics, it seems natural that the optimal values of these two parameters will be selected for each metric independently. Within this study, we adhere to an approach where, for each metric, author, and fragment length, the configuration of parameters that maximizes attribution accuracy is determined. This strategy allows us to evaluate the potential of distance measures under optimal conditions for each of them.

Thus, for each corpus, author, metric, and fragment length, we choose whether to perform standardization or not, as well as the optimal MFW value, and then perform attribution of all fragments. Such calculations were performed for all authors of each of the four corpora with all 6 metrics. The attribution accuracies obtained during the experiment for a total of 46 authors (the figures represent the results for the authors with at least two works in the corpora) are shown in Figures 1-4 of the Appendix.

Spearman distance demonstrates the least dependence on authorial, corpus, and linguistic specifics. The key result of the experiment: when selecting the optimal value of MFW with a text length increase, the use of the given metric allows us obtaining a value of 1.000 for all authors in all four corpora. No exceptions in terms of failure to reach the accuracy ceiling were recorded. The difference between “simple” and “complicated” authors lies not in the fundamental possibility of reliable attribution but in the length of the text that allows one reaching it.

For stylistically homogeneous authors (Rasputin, Plotinus, French, Ivanov), the maximum value is recorded at small fragment lengths and maintained over a wide range of MFW. For authors with pronounced intra-author variability (Belov, Demosthenes, Kryukov, Warrington), the metric also reaches 1.000 but requires longer segments and more precise selection of MFW (most often 600-800). On short fragments, these authors show increased dispersion of indicators, but with

sufficient fragment size, the sample word frequencies stabilize, and the TPR reaches 1.000.

Manhattan distance and Cosine Delta behave almost identically. Manhattan distance reached 1.000 for all authors; Cosine Delta did not cope with Warrington texts (the highest TPR for this author was 0.913). For authors with a homogeneous set of texts in the corpus, the curve stabilizes already with a short length of a text fragment and maintains the maximum value across the entire subsequent range. For authors with pronounced intra-corpus variability, a TPR value of 1 is achieved in the region of longer segments. In case of Astafiev, Serafimovich, and Efremov, Manhattan distance started better than Spearman distance, but in most cases it acted as the one catching up (and always did catch up). The use of standardization noticeably accelerates the achievement of maximum accuracy on short fragments, while on ultra-long segments the difference between the standardized and non-standardized variants disappears.

The Labbé distance shows the sharpest polarisation. For stylistically balanced authors (Rasputin, Galen, Lewis, Fadeev), the curve rises and reaches 1.000 sufficiently quickly, making the metric highly effective on short, homogeneous texts. For the “complex” authors Plutarch, Platonov, Kryukov, the curve reaches a local maximum and then gradually decreases. For Turton and Belov, a monotonic increase is observed with maximum values of 0.586 and 0.833, respectively, even at the maximum fragment length. In several cases (4 observations in the “Sholokhov” corpus, 3 in “SF&RP”, 6 in “Greek”), we see a “rollback” effect: the curve first increases, reaches a local maximum at medium lengths, and then sharply drops. This effect is explained by the metric working with the full corpus vocabulary: on long fragments, rare wordforms and genre shifts accumulate, which disrupt the internal homogeneity of the segments and distort the stylistic signal,

leading to a sharp decrease in accuracy after reaching a local peak.

The Soergel distance curve on homogeneous authors (Mozhaev, Plotinus, Stroud) goes up and quickly reaches 1.000 at minimal lengths. Regarding “complex” authors (Belov, Turton, Platonov), accuracy gets stuck in the range of 0.550-0.850, with the curve not leveling off even when fragment length increases to the upper bounds of the scale.

JSD behaves differently from other metrics: either it quickly reaches 1.000, or it gets stuck at low values; there are almost no intermediate options.

On homogeneous corpora, JSD works almost instantly. For Lucian, Plutarch, Rasputin, and 5 authors of the “Rowling” corpus, the curve jumps to 1.000 already at the first points. Moreover, it often overtakes even Spearman distance and Manhattan. Apparently, for stable texts, the entropy signal is so clear that the metric does not require a long text fragment.

The situation with heterogeneous authors is different. For the same Belov and

Turton, JSD does not rise above 0.833 and 0.793, respectively: genre differences within the work of a single author create an entropy noise that the metric cannot filter out.

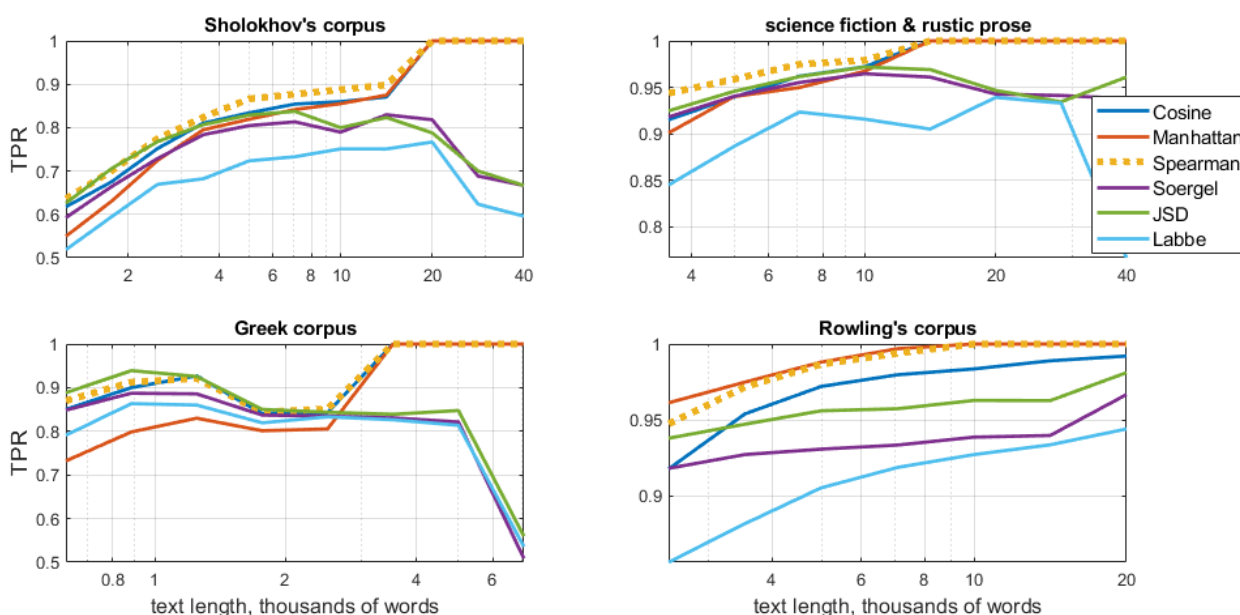
When compared with other metrics, JSD works more rigidly: Spearman distance and Manhattan distance tend to 1 smoothly, albeit at different speeds, while JSD either shoots immediately or does not shoot at all. For short, genre-pure texts, this is an excellent tool, but in heterogeneous corpora with internal style dynamics, it often fails.

Comparison of Attribution Accuracy for Corpora as a Whole.

In order to present the results in a more convenient and generalized form, for each of the four corpora we averaged the obtained TPR values across all authors included in that corpus. When calculating the average, all authors are weighted equally, regardless of how many of their works are represented in the corpus.

Analysis of the average attribution accuracy reveals a clear hierarchy of metrics (see Figure 4).

Figure 4. Average TPR for 4 corpora
Рисунок 4. Среднее TPR для 4 корпусов



Spearman distance demonstrates the greatest universality, systematically reaching an average accuracy value of 1.000 in all four corpora. Manhattan distance and Cosine Delta show similar dynamics: on homogeneous data, their curves almost merge with Spearman distance, while on heterogeneous materials, they lag only slightly on short segments, also reaching maximum accuracy with increasing text fragment length.

The Soergel distance and the JSD show ambiguous results. Both metrics on the “Rowling” and “SF&RP” corpora grow monotonically, confidently staying in the corridor > 0.900 , while on “Sholokhov” and “Greek”, after rising to 0.850-0.920 on small and medium fragments, they exhibit a noticeable decline to 0.550-0.700.

The Labbé distance shows the least stability growing smoothly only on the “Rowling” corpus and reaching the value of 0.944, while on the other three corpora, its average accuracy curve drops sharply.

For a summary assessment of metrics

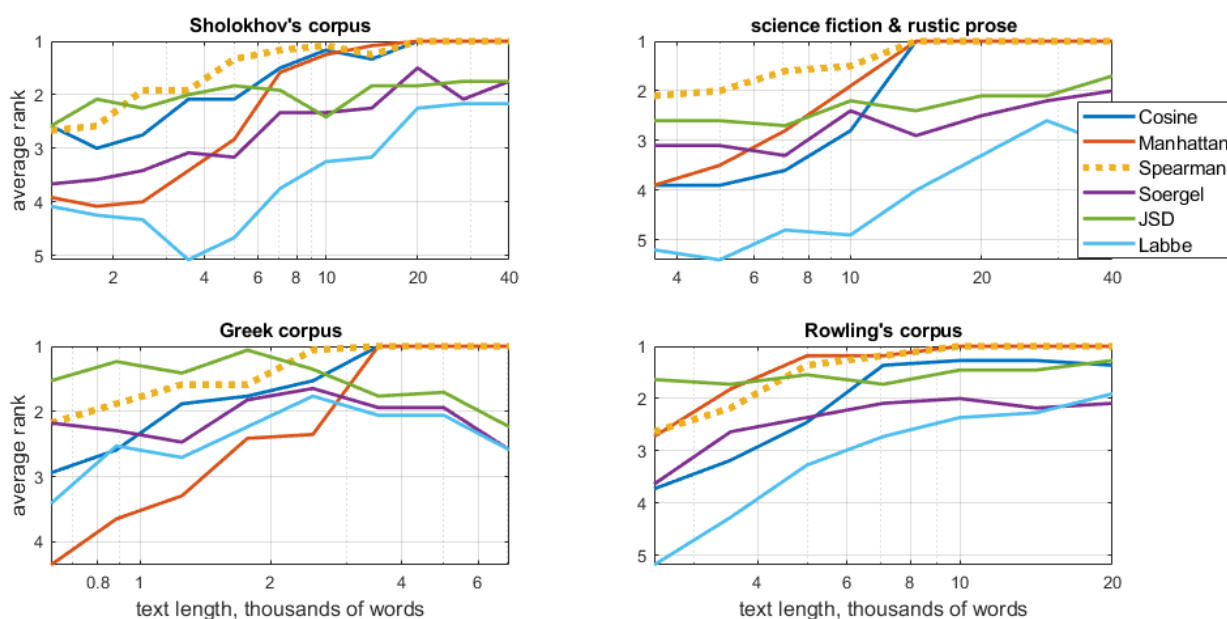
under inter-author variability, average ranks of the metrics were also calculated. They show the place the selected metric occupies among all six metrics in the list ordered by decreasing accuracy.

Ranking was carried out as follows: for each author represented in the corpus at a given text fragment length, the TPR values were ordered in descending order; the metric with the highest accuracy was assigned rank 1, the next – 2, and so on. If several metrics showed the same TPR, each was assigned the best of the numbers they occupy in the ordered list. For example, if three metrics showed the same TPR and occupied positions 2-4 in the ordered list, each was assigned rank 2. This rule was adopted to achieve a more transparent interpretation. In this case, for example, an average rank equal to 1 means that the metric for all authors either surpassed the other metrics or showed a result no worse than any other.

The average rank graphs (Figure 5) make the hierarchy of metrics even clearer.

Figure 5. Average rank for 6 metrics in different corpora

Рисунок 5. Средний ранг для 6 метрик в разных корпусах



As can be seen, Spearman distance is among the leaders on all corpora, regardless of text length; in most cases, it occupies the

first position. Manhattan distance in three out of four corpora noticeably lags behind Spearman distance on short text fragments but

catches up with the leader on all four corpora when the context increases. Cosine Delta, due to a drop on Warrington, does not share the top positions on the “Rowling” corpus. The Labbé distance systematically closes the ranking (ranks 4-5), especially in the corpus of Soviet prose, where the gap from the top metrics reaches 3-4 units.

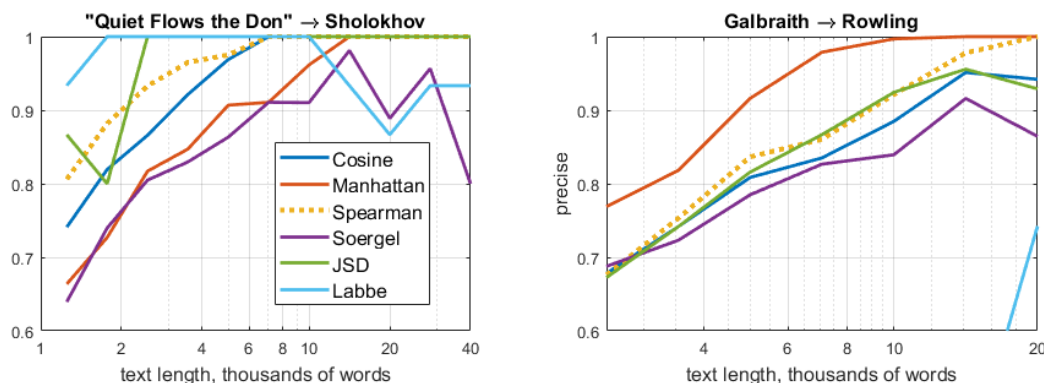
JSD exhibits consistently strong performance on short texts. It outperforms all competing metrics on the “Greek”, “Sholokhov”, and “Rowling” corpora at small sample sizes, and is surpassed only by Spearman distance on the “SF&RP” corpora. However, with increasing fragment length, it drops to the fourth rank. The Soergel distance behaves more reservedly: it rarely rises above 3rd or 4th position, demonstrating smooth but slow growth (the exception is “Greek”, where its curve falls).

The final part of our study is devoted to assessing the *accuracy of metrics in solving two well-known stylometric problems*: establishing the authorship of the novel “And Quiet Flows the Don” (see about this: Velikanova and Orekhov, 2019; Maslinsky, 2022) and identifying the real writer hiding behind the pseudonym R. Galbraith. For each case, we compared the effectiveness of 5 metrics on fragments of different lengths with the optimal MFW values selected for each case, and the Labbé distance on the full vocabulary.

The experiment on attributing disputed works in two heterogeneous cases revealed fundamental differences in the behavior of stylometric metrics depending on the quantitative and qualitative characteristics of the corpus (Figure 6).

Figure 6. Dependence of authorship attribution accuracy on text length for six stylometric metrics

Рисунок 6. Зависимость точности атрибуции авторства от длины текста для шести стилометрических метрик



Analysis of the graph for the “Sholokhov” corpus demonstrates unexpected behavior of JSD. Contrary to generalized expectations, JSD reaches the threshold value of 1.000 in the range of 2300 words, and, unlike the Labbé distance, which reaches 1.000 earlier than others, it does not show a subsequent decrease in accuracy on very long fragments. This behavior is explained by the specifics of the corpus material: with relative genre and size balance of the texts (in the corpus, the only exceptions in size and genre were the story “The Fate of a Man”; “Don

Stories” were not divided but went as a single text array), the entropy signal proves to be sufficiently stable, allowing the information-theoretic measure to instantly capture the authorial invariant. At the same time, rank and frequency metrics increase in a more gradual manner, only reaching their maximum by 7,000–12,000 words. In contrast, despite initially reaching 1.000, Labbé distance loses accuracy after 14,000 words due to the accumulation of stylistic noise and disruption to the internal homogeneity of long segments. This pattern confirms that, in a balanced

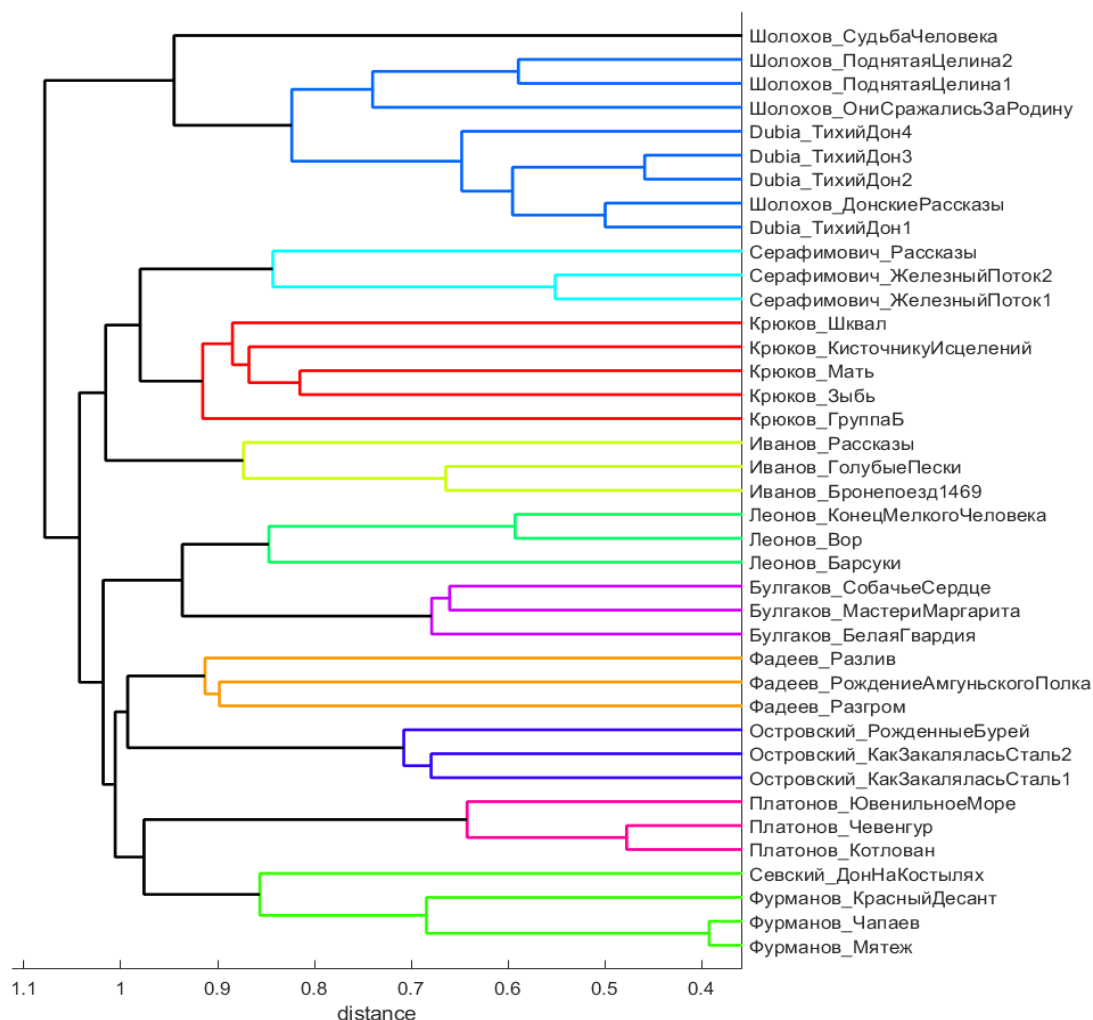
corpus, JSD can serve not only as an alternative, but also as an advanced attribution tool that combines high convergence speed with robustness to context expansion.

A different picture emerges in the “Rowling” case, where the pronounced asymmetry in text sizes (Rowling’s large-scale novels versus Galbraith’s more compact detective stories) creates specific conditions for attribution. The Labbé distance in these conditions is practically inapplicable: on fragments up to 10000 words, its accuracy approaches zero, and only at larger values (14142-20000 words) is moderate growth up to 0.750 recorded. This empirical result

confirms the theoretical proposition that comparable text lengths are necessary for the correct application of the Labbé distance. Absolute leadership in this corpus is demonstrated by Manhattan distance, reaching a value of 1.000 already at the 10000-word mark and maintaining it with further length increase. Spearman distance also shows high effectiveness, although it reaches the limit value only by 20000 words. The other metrics – Cosine Delta, Soergel distance, and Jensen-Shannon divergence – demonstrate significantly lower results under size imbalance conditions.

Figure 7. Dendrogram of the “Sholokhov” corpus, Spearman distance (with frequency standardization), MFW=2700

Рисунок 7. Дендрограмма корпуса “Шолохов”, расстояние Спирмена (со стандартизацией частот), MFW=2700



Based on the results obtained using Spearman distance, we decided to replicate the dendrogram for the “Sholokhov” corpus presented in the article “Digital Textology: Text Attribution on the Example of M. A. Sholokhov’s Novel “And Quiet Flows the Don”” by N. Velikanova and B. Orekhov. The researchers note: “A certain strangeness is observed only in the fact that “Don Stories” and Sholokhov’s later texts ended up in different parts of the tree. At the same time, this is quite natural: the novel was written over 16 years, the author’s style should have changed during this time” (Velikanova and Orekhov, 2019: 78). Spearman groups the block of “Don Stories”, “And Quiet Flows the Don”, and later works by M. Sholokhov (“Virgin Soil Uplifted”, “They Fought for Their Country”) into one cluster (see Figure 7). With increased distance, the only story in this corpus, “The Fate of a Man”, joins this cluster.

Discussion. The comparison of 6 metrics on four heterogeneous corpora reveals a stable pattern: Spearman distance ensures the highest accuracy of authorship attribution regardless of the genre structure of the sample, fragment length, and MFW. Unlike frequency and entropy approaches, the metric does not show critical failures when moving from stylistically aligned authorial corpora to hyper-corpora with pronounced intra-author variability. JSD behaves in two ways: on homogeneous texts, it quickly captures the authorial invariant; however, under conditions of genre or chronological heterogeneity, its discriminatory ability sharply decreases. Traditional measures (Manhattan distance, Cosine Delta) maintain high effectiveness, but their stability depends on preliminary standardization and precise selection of MFW, whereas Labbé and Soergel metrics prove sensitive to the accumulation of rare wordforms and imbalance in text size.

Such robustness of Spearman distance is associated with the fact that the metric operates not with relative or standardized frequencies of lexical units, but with their ranks, i.e., positions in a list ordered by

descending or ascending frequency. The transition to a rank scale mitigates the impact of extreme outliers and allows obtaining stable results for cases with different probability distributions of empirical word frequencies. Thanks to this, the stylistic signal is preserved even under significant intra-corpus shifts. Jensen-Shannon divergence, measuring the divergence of probability masses, interprets genre and thematic variations as statistical noise under the same conditions, which explains its instability on heterogeneous arrays.

Our results refine the findings of F. Jannidis and colleagues (Jannidis et al., 2015), who recorded the advantage of Cosine Delta and its robustness to the expansion of the MFW list due to vector length normalization. We have shown that the rank approach maintains comparable accuracy but demonstrates a broader area of parametric stability in heterogeneous corpora, where Cosine Delta requires more rigid “tuning”.

The conclusions obtained confirm the results of the study by D. Pronin and E. Kazartsev (Pronin and Kazartsev, 2026) on the effectiveness of entropy measures on controlled samples but point to their limitations when moving to heterogeneous arrays, where Spearman’s rank metric proves more robust.

The results of O. Alieva (Alieva, 2024), obtained on the homogeneous corpus of ancient Greek prose, demonstrate the high effectiveness of Cosine Delta, Tanimoto distance, and Labbé distance. The first metric in our studies also showed high results, but Tanimoto/Soergel and Labbé distance noticeably lost to all others: Tanimoto/Soergel loses discriminatory ability under imbalance in text size, and Labbé distance exhibits a “rollback” effect on long fragments. For “Sholokhov”, “Rowling”, and “SF&RP”, this decline can be explained by genre heterogeneity, accumulation of rare wordforms, and different text size, to which these measures are mathematically sensitive. The discrepancy between the results for the

ancient Greek corpus can be explained by differences in the experimental protocol.

Reviews of quantitative attribution methods (Koppel et al., 2009) and studies of the role of frequency features (Hoover, 2002) show that the effectiveness of a metric is determined by the structure of the corpus. Our results clarify these conclusions: under conditions of genre heterogeneity and volume imbalance, the rank approach proves more robust than frequency and entropy measures.

The totality of results allows us to formulate recommendations oriented towards the corpus structure and research task:

1. For heterogeneous genre collections with size imbalance, Spearman distance is preferred. Shift to the rank scale mitigates the influence of outliers in empirical frequency values and maintains informativeness even with a sparse feature space;

2. For homogeneous corpora, stylistically uniform texts, and short texts, it is advisable to use JSD. With optimal selection of MFW (600-2000 features), the metric provides fast convergence, outpacing frequency approaches;

3. The Labbé and Soergel distances require preliminary size-balance verification. Their application to heterogeneous samples without filtering is risky: the metrics are sensitive to zero frequencies and asymmetry of text lengths.

As we have seen, there is no single metric that yields the best results under all conditions. The question arises about an objective procedure, free from the researcher's personal preferences, for choosing a metric to solve specific tasks. In practice, a situation often occurs where the set of possible candidates for authorship of the text being attributed is known. Under such conditions, it is natural to start work by assembling a corpus of texts by the candidate authors, conducting a simulation similar to that described above, and evaluating the accuracy of different metrics in attributing texts specifically for this set of alternatives. Such a preliminary procedure will allow an

informed choice of the most suitable tool for solving a specific authorship attribution task.

Despite the results obtained, the study has a number of limitations:

The imbalance in author representation within the corpora (especially in "SF&RP") could have led to a bias in accuracy estimates in favor of authors with larger sample sizes (this concerns primarily Belov with his 17 texts in our hyper-corpus);

The focus on literary texts is due to the fact that stylometric attribution has historically developed precisely in this area and is the most well-developed methodologically. The applicability of the compared distance measures to non-literary genres – such as academic writing, journalistic texts, or social media posts – remains an open question and requires separate investigation.

The corpora we studied demonstrate similar patterns. Further expansion of validation across genres, languages, and historical periods will allow us to clarify the boundaries of applicability of rank and information-theoretic metrics;

The fixed set of MFW (100-3200) does not cover the entire spectrum of possible feature space configurations.

Conclusion. Our study has once again demonstrated that a universal metric does not exist in stylometry; under different conditions, different metrics may prove most useful. However, the distance based on Spearman's rank correlation coefficient more often than other metrics shows the best results, and thus represents the most robust tool for stylometric attribution of heterogeneous text collections. The transition to a rank scale ensures robustness to strong outliers in empirical frequencies and maintains informativeness even under conditions of genre heterogeneity, chronological dispersion, and size imbalance. Unlike frequency and probabilistic metrics, Spearman distance achieves maximum accuracy for all authors with adequate parameter tuning. Jensen-Shannon divergence

demonstrates the conditional nature of its effectiveness: on stylistically homogeneous texts and short fragments, it can surpass basic approaches; however, under conditions of pronounced heterogeneity, the metric's accuracy decreases. Traditionally employed measures remain reliable with sufficient text size whereas Labbé and Soergel metrics exhibit systematic sensitivity to the structural features of the corpus.

Statement on the Use of Generative

AI: The authors declare that no generative artificial intelligence or AI-assisted technologies requiring disclosure under the Journal's policy were used in conducting the research or preparing this manuscript.

Заявление об использовании генеративного ИИ: Авторы заявляют, что при проведении исследования и подготовке настоящей рукописи генеративный искусственный интеллект и ИИ-ассистированные технологии, подлежащие раскрытию в соответствии с политикой журнала, не использовались.

References

- Alieva, O. V. (2024). Distance measures for determining the authorship of ancient Greek texts, *Tsifrovye gumanitarnye issledovaniya* [Digital Humanities Research], 1 (1), 8–33. DOI: 10.31860/cgi-2024-1-8-33 (In Russian)
- Dvinyatin, F. N., Kovalev, B. V. (2024). “Sirin is not inferior to Leonov”: the Delta method for stylometric analysis of Russian novels of the interwar period, *Шаги / Steps* [Shagi/Steps], 10 (3), 207–229. DOI: 10.22394/2412-9410-2024-10-2-207-229. (In Russian)
- Martynenko, G. Ya. (2014). Stylometry: emergence and development in the context of interdisciplinary interaction. Part 1. First steps: 19th century, *Strukturnaya i prikladnaya lingvistika* [Structural and Applied Linguistics], 10, 3–23. (In Russian)
- Martynenko, G. Ya. (2015). Stylometry: emergence and development in the context of interdisciplinary interaction. Part 2. First half of the 20th century: expansion of interdisciplinary contacts of stylometry, *Strukturnaya i prikladnaya lingvistika* [Structural and Applied Linguistics], 11, 9–28. (In Russian)
- Maslinskiy, K. A. (2022). Refined digital textology: once again on the question of the authorship of the novel “And Quiet Flows the Don”, *Russkaya literatura* [Russian Literature], 1, 247–254. DOI: 10.31860/0131-6095-2022-1-247-254. (In Russian)
- Staroverova, P. V. (2025). The Strugatskys separated: Stylometric identification of individual authorial fragments in a collective text, *Tsifrovye gumanitarnye issledovaniya* [Digital Humanities Research], 1, 24–39. DOI: 10.31860/cgi-2025-1-24-39. (In Russian)
- Velikanova, N. P. and Orekhov, B. V. (2019). Digital textology: text attribution using the example of M. A. Sholokhov's novel “And Quiet Flows the Don”, *Mir Sholokhova* [Sholokhov World], 1 (11), 70–82. (In Russian)
- Arefin, A. S., Vimieiro, R., Riveros, C., Craig, H. and Moscato, P. (2014). An information theoretic clustering approach for unveiling authorship affinities in Shakespearean era plays and poems, *PLoS One*, 9 (10), e111445. DOI: 10.1371/journal.pone.0111445. (In English)
- Argamon, S. (2008). Interpreting Burrows' delta: Geometric and probabilistic foundations, *Literary and Linguistic Computing*, 23 (2), 131–147. DOI: 10.1093/lc/fqn003. (In English)
- Bortolato, C. (2016). Intertextual Distance of Function Words as a Tool to Detect an Author's Gender: A Corpus-Based Study on Contemporary Italian Literature, *Glottometrics*, 34, 28–43. (In English)
- Burrows, J. (2002). ‘Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship, *Literary and Linguistic Computing*, 17 (3), 267–287. DOI: 10.1093/lc/17.3.267. (In English)
- Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions, *IEEE Transactions on Information Theory*, 49 (7), 1858–1860. DOI: 10.1109/TIT.2003.813506. (In English)
- Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C. and Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution, *Digital Scholarship in the Humanities*, 32, 4–16. DOI: 10.1093/lc/fqx023. (In English)
- Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques, *Literary*

and *Linguistic Computing*, 22 (3), 251–270. DOI: 10.1093/lc/fqm020. (In English)

Grzybek, P. (2014). The Emergence of Stylometry: Prolegomena to the History of Term and Concept, available at: http://www.peter-grzybek.eu/science/publications/2014/grzybek_2014_stylometry.pdf (Accessed 05.04.2026). (In English)

Hoover, D. (2002). Frequent word sequences and statistical stylistics, *Literary and Linguistic Computing*, 17 (2), 157–180. DOI: 10.1093/lc/17.2.157. (In English)

Hoover, D. (2004). Testing Burrows's Delta, *Literary and Linguistic Computing*, 19 (4), 453–475. DOI: 10.1093/lc/19.4.453. (In English).

Jannidis, F., Pielström, S., Schöch, C. and Vitt, T. (2015). Improving Burrows' Delta – An empirical evaluation of text distance measures, available at: https://www.researchgate.net/publication/280086768_Improving_Burrows'_Delta-An_empirical_evaluation_of_text_distance_measures (Accessed 5 April 2026). (In English)

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution, *Journal of the American Society for Information Science and Technology*, 60 (1), 9–26. DOI: 10.1002/asi.20961. (In English)

Kocher, M. and Savoy, J. (2019). Evaluation of text representation schemes and distance measures for authorship linking, *Digital Scholarship in the Humanities*, 34 (1), 189–207. DOI: 10.1093/lc/fqy013. (In English)

Kovalev, B. V. (2024). From Classics to Digital Philology: On the Origin and Growth of Stylometry, *Philologia Classica*, 19 (2), 347–360. DOI: 10.21638/spbu20.2024.211. (In English)

Labbé, C. and Labbé, D. (2006). A Tool for Literary Studies: Intertextual Distance and Tree Classification, *Literary and Linguistic Computing*, 21 (3), 311–326. DOI: 10.1093/lc/fqi063. (In English)

Levy, A., Shalom, B. R. and Chalamish, M. (2025). A guide to similarity measures and their data science applications, *J Big Data*, 12, Art. 188. DOI: 10.1186/s40537-025-01227-1. (In English)

Lin, J. (1991). Divergence measures based on the Shannon entropy, *IEEE Transactions on Information Theory*, 37 (1), 145–151. DOI: 10.1109/18.61115. (In English)

Mikros, G., Sosoni, V., Manousakis, V., and Polychroniou, K. (2026). Burrows' Delta as a

Convergent Validator: Stylometric Analysis for Complementary Machine Translation Evaluation. *Journal of Quantitative Linguistics*, 1–29. DOI: 10.1080/09296174.2026.2612931. (In English)

Orekhov, B. V. (2024). Does Burrows' Delta really confirm that Rowling and Galbraith are the same author?, available at: <https://publications.hse.ru/pubs/share/direct/950068199.pdf> (Accessed 5 April 2026). DOI: 10.48550/arXiv.2407.10301. (In English)

O'Sullivan, J. (2025). Stylometric comparisons of human versus AI-generated creative writing. *Humanities and Social Sciences Communications*, 12, 1708. DOI: 10.1057/s41599-025-05986-3. (In English)

Pronin, D. and Kazartsev, E. (2026). Rank-Turbulence Delta and Interpretable Approaches to Stylometric Delta Metrics, available at: <https://arxiv.org/pdf/2604.19499> (Accessed 5 April 2026). (In English)

Rybacki, J. and Eder, M. (2011). Deeper Delta across genres and languages: do we really need the most frequent words?, *Literary and Linguistic Computing*, 26 (3), 315–321. DOI: 10.1093/lc/fqr031. (In English)

Savoy, J. (2020). Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling, Springer International Publishing, Cham, Switzerland. DOI: 10.1007/978-3-030-53360-1. (In English)

Skorinkin, D. and Orekhov, B. (2023). Hacking stylometry with multiple voices: Imaginary writers can override authorial signal in Delta. *Digital Scholarship in the Humanities*, 38 (3), 1247–1264. DOI: 10.1093/lc/fqad012. (In English)

Smith, P. and Aldridge, W. (2011). Improving Authorship Attribution. Optimizing Burrows' Delta Method, *Journal of Quantitative Linguistics*, 18 (1), 63–88. DOI: 10.1080/09296174.2011.533591. (In English)

Spearman, C. (1904). The proof and measurement of association between two things, *American Journal of Psychology*, 15 (1), 72–101. DOI: 10.2307/1412159. (In English)

Stanikunas, D., Mandravickaite, J. and Krilavicius, T. (2017). Comparison of distance and similarity measures for stylometric analysis of Lithuanian texts, available at: <https://portalcris.vdu.lt/server/api/core/bitstreams/423f881d-b231-4a02-b509-80fd53205028/content> (Accessed 28 April 2026). (In English)

Willett, P., Barnard, J. and Downs, G. (1998). Chemical similarity searching, *J Chem Inform Comput Sci*, 38, 983–996. DOI: 10.1021/CI9800211. (In English)

Литература

Алиева О. В. Меры расстояния для определения авторства древнегреческих текстов // Цифровые гуманитарные исследования. 2024. № 1 (1). С. 8–33. DOI: 10.31860/cgi-2024-1-8-33

Великанова Н. П., Орехов Б. В. Цифровая текстология: атрибуция текста на примере романа М. А. Шолохова «Тихий Дон» // Мир Шолохова. 2019. № 1 (11). С. 70–82.

Двнятин Ф. Н., Ковалев Б. В. «Сирин не уступает Леонову»: метод Delta для стилиметрического анализа русских романов межвоенного периода // Шаги/Steps. 2024. 10 (3). С. 207–229. DOI: 10.22394/2412-9410-2024-10-2-207-229

Мартыненко Г. Я. Стилиметрия: возникновение и становление в контексте междисциплинарного взаимодействия. Часть 1. Первые шаги: XIX век // Структурная и прикладная лингвистика. 2014. № 10. С. 3–23.

Мартыненко Г. Я. Стилиметрия: возникновение и становление в контексте междисциплинарного взаимодействия. Часть 2. Первая половина XX в.: расширение междисциплинарных контактов стилиметрии // Структурная и прикладная лингвистика. 2015. № 11. С. 9–28.

Маслинский К. А. Уточненная цифровая текстология: еще раз к вопросу об авторстве романа «Тихий Дон» // Русская литература. 2022. № 1. С. 247–254. DOI: 10.31860/0131-6095-2022-1-247-254

Староверова П. В. Стругацкие разделенные. Стилиметрическое выявление индивидуальных авторских фрагментов в коллективном тексте // Цифровые гуманитарные исследования. 2025. № 1. С. 24–39. DOI: 10.31860/cgi-2025-1-24-39

Arefin A. S. An information theoretic clustering approach for unveiling authorship affinities in Shakespearean era plays and poems / Arefin A. S., Vimieiro R., Riveros C., Craig H., Moscato P. // PLoS One. 2014. Vol. 9, № 10. e111445. <https://doi.org/10.1371/journal.pone.0111445>

Argamon S. Interpreting Burrows' delta: Geometric and probabilistic foundations //

Literary and Linguistic Computing. 2008. Vol. 23, № 2. Pp. 131–147. DOI:10.1093/lc/fqn003

Bortolato C. Intertextual Distance of Function Words as a Tool to Detect an Author's Gender: A Corpus-Based Study on Contemporary Italian Literature // Glottometrics. 2016. № 34. Pp. 28–43.

Burrows J. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship // Literary and Linguistic Computing. 2002. Vol. 17, № 3. Pp. 267–287. DOI:10.1093/lc/17.3.267

Endres D. M., Schindelin J. E. A new metric for probability distributions // IEEE Transactions on Information Theory. 2003. 49 (7). Pp. 1858–1860. DOI:10.1109/TIT.2003.813506

Evert S. Understanding and explaining Delta measures for authorship attribution / Evert S., Proisl T., Jannidis F., Reger I., Pielström S., Schöch C., Vitt T. // Digital Scholarship in the Humanities. 2017. Vol. 32. Pp. 4–16. DOI:10.1093/lc/fqx023

Grieve J. Quantitative authorship attribution: An evaluation of techniques // Literary and Linguistic Computing. 2007. 22 (3). P. 251–270. DOI: 10.1093/lc/fqm020

Grzybek P. The Emergence of Stylometry: Prolegomena to the History of Term and Concept. URL: http://www.peter-grzybek.eu/science/publications/2014/grzybek_2014_stylometry.pdf (дата обращения: 05.04.2026).

Hoover D. Frequent word sequences and statistical stylistics // Literary and Linguistic Computing. 2002. 17 (2). Pp. 157–180. DOI: 10.1093/lc/17.2.157

Hoover D. Testing Burrows's Delta // Literary and Linguistic Computing. 2004. Vol. 19, № 4. Pp. 453–475. DOI:10.1093/lc/19.4.453

Jannidis F. Improving Burrows' Delta – An empirical evaluation of text distance measures / Jannidis F., Pielström S., Schöch C., Vitt T. // Book of Abstracts of the Digital Humanities Conference 2015. URL: https://www.researchgate.net/publication/280086768_Improving_Burrows'_Delta-An_empirical_evaluation_of_text_distance_measures (дата обращения: 05.04.2026).

Koppel M., Schler J., Argamon S. Computational methods in authorship attribution // Journal of the American Society for Information Science and Technology. 2009. 60 (1). Pp. 9–26. DOI: DOI:10.1002/asi.20961

Kocher M. Evaluation of text representation schemes and distance measures for authorship

linking / Kocher M., Savoy J. // Digital Scholarship in the Humanities. 2019. Vol. 34. № 1. Pp. 189–207. DOI:10.1093/lc/fqy013

Kovalev B. V. From Classics to Digital Philology: On the Origin and Growth of Stylometry // Philologia Classica. 2024. Vol. 19. № 2. Pp. 347–360. DOI: 10.21638/spbu20.2024.211

Labbé C. A Tool for Literary Studies: Intertextual Distance and Tree Classification / Labbé C., Labbé D. // Literary and Linguistic Computing. 2006. Vol. 21. № 3. Pp. 311–326. DOI:10.1093/lc/fqi063

Levy A. A guide to similarity measures and their data science applications / Levy A., Shalom B. R., Chalamish M. // J Big Data. 2025. Vol. 12. Art. 188. DOI: 10.1186/s40537-025-01227-1

Lin J. Divergence measures based on the Shannon entropy // IEEE Transactions on Information Theory. 1991. 37 (1). Pp. 145–151. DOI: 10.1109/18.61115

Mikros G. Burrows' Delta as a Convergent Validator: Stylometric Analysis for Complementary Machine Translation Evaluation / Mikros G., Sosoni V., Manousakis V., Polychroniou K. // Journal of Quantitative Linguistics. 2026. Pp. 1–29. DOI: 10.1080/09296174.2026.2612931

Orekhov B. V. Does Burrows' Delta really confirm that Rowling and Galbraith are the same author? URL: <https://publications.hse.ru/pubs/share/direct/950068199.pdf> (дата обращения: 05.04.2026). DOI: 10.48550/arXiv.2407.10301

O'Sullivan J. Stylometric comparisons of human versus AI-generated creative writing // Humanities and Social Sciences Communications. 2025. Vol. 12. Art. 1708. DOI: 10.1057/s41599-025-05986-3

Pronin D. Rank-Turbulence Delta and Interpretable Approaches to Stylometric Delta Metrics / Pronin D., Kazartsev E. URL: <https://arxiv.org/pdf/2604.19499> (дата обращения: 24.04.2026).

Rybicki J. Deeper Delta across genres and languages: do we really need the most frequent words? / Rybicki J., Eder M. // Literary and Linguistic Computing. 2011. Vol. 26. № 3. Pp. 315–321. DOI: 10.1093/lc/fqr031

Savoy J. Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling. Cham: Springer International

Publishing, 2020. DOI: 10.1007/978-3-030-53360-1

Skorinkin D. Hacking stylometry with multiple voices: Imaginary writers can override authorial signal in Delta / Skorinkin D., Orekhov B. // Digital Scholarship in the Humanities. 2023. Vol. 38. № 3. Pp. 1247–1264. DOI: 10.1093/lc/fqad012

Smith P. Improving Authorship Attribution. Optimizing Burrows' Delta Method / Smith P., Aldridge W. // Journal of Quantitative Linguistics. 2011. Vol. 18. № 1. Pp. 63–88. DOI: 10.1080/09296174.2011.533591

Spearman C. The proof and measurement of association between two things // American Journal of Psychology. 1904. 15 (1). Pp. 72–101. DOI: 10.2307/1412159

Stanikunas D. Comparison of distance and similarity measures for stylometric analysis of Lithuanian texts / Stanikunas D., Mandravickaite J., Krilavicius T. 2017. URL: <https://portalcris.vdu.lt/server/api/core/bitstreams/423f881d-b231-4a02-b509-80fd53205028/content> (дата обращения: 28.04.2026).

Willett P. Chemical similarity searching / Willett P., Barnard J., Downs G. // J Chem Inform Comput Sci. 1998. Vol. 38. Pp. 983–996. DOI: 10.1021/C19800211

All authors have read and approved the final manuscript.

Все авторы прочитали и одобрили окончательный вариант рукописи.

Conflicts of interests: the authors have no conflicts of interest to declare.

Конфликты интересов: у авторов нет конфликтов интересов для декларации.

Vladimir V. Bochkarev, engineer, Research Laboratory “Multidisciplinary Textual Studies”, Kazan (Volga Region) Federal University, Kazan, Russia.

Владимир Владимирович Бочкарев, инженер, научно-исследовательская лаборатория «Мультидисциплинарные исследования текста», Казанский (Приволжский) федеральный университет, Казань, Россия.

Anton S. Afanasev, Doctor of Philology, Associate Professor, Dean of the Leo Tolstoy Higher School of Russian Philology and Culture,

Professor at the Department of Russian Literature and Methods of Teaching It, Kazan (Volga Region) Federal University, Kazan, Russia.

Антон Сергеевич Афанасьев, доктор филологических наук, доцент, декан Высшей школы русской филологии и культуры им. Льва Толстого, профессор кафедры русской литературы и методики ее преподавания, Казанский (Приволжский) федеральный университет, Казань, Россия.

Rinat A. Bakirov, Candidate of Philology, Associate Professor, Department of Russian Literature and Methods of Its Teaching, Kazan (Volga Region) Federal University, Kazan,

Russia; Research Fellow, Laboratory of Digital Literature and Folklore Studies, Institute of Russian Literature (Pushkin House), St. Petersburg, Russia.

Бакиров Ринат Альбертович, кандидат филологических наук, доцент кафедры русской литературы и методики ее преподавания, Казанский (Приволжский) федеральный университет, Казань, Россия; научный сотрудник лаборатории цифровых исследований литературы и фольклора Института русской литературы (Пушкинский Дом), Санкт-Петербург, Россия.

Appendix

Figure 8. Best probability of correct attribution (TPR) for the “SF&RP” corpus

Рисунок 8. Наилучшая вероятность правильной атрибуции (TPR) по корпусу «SF&PR»

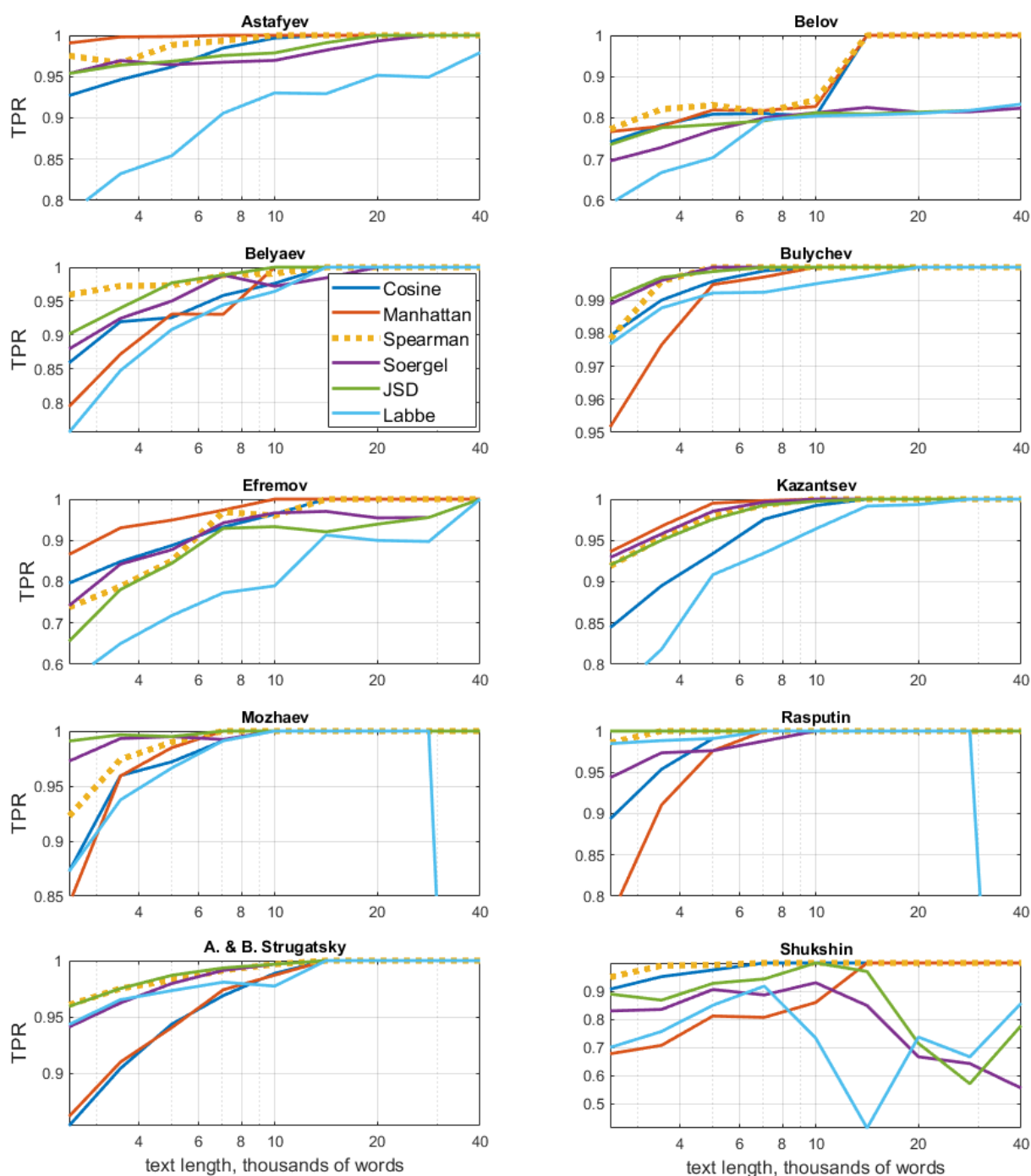


Figure 9a. Best TPR for the “Greek” corpus. Beginning
Рисунок 9а. Наилучший TPR по корпусу «Греки». Начало

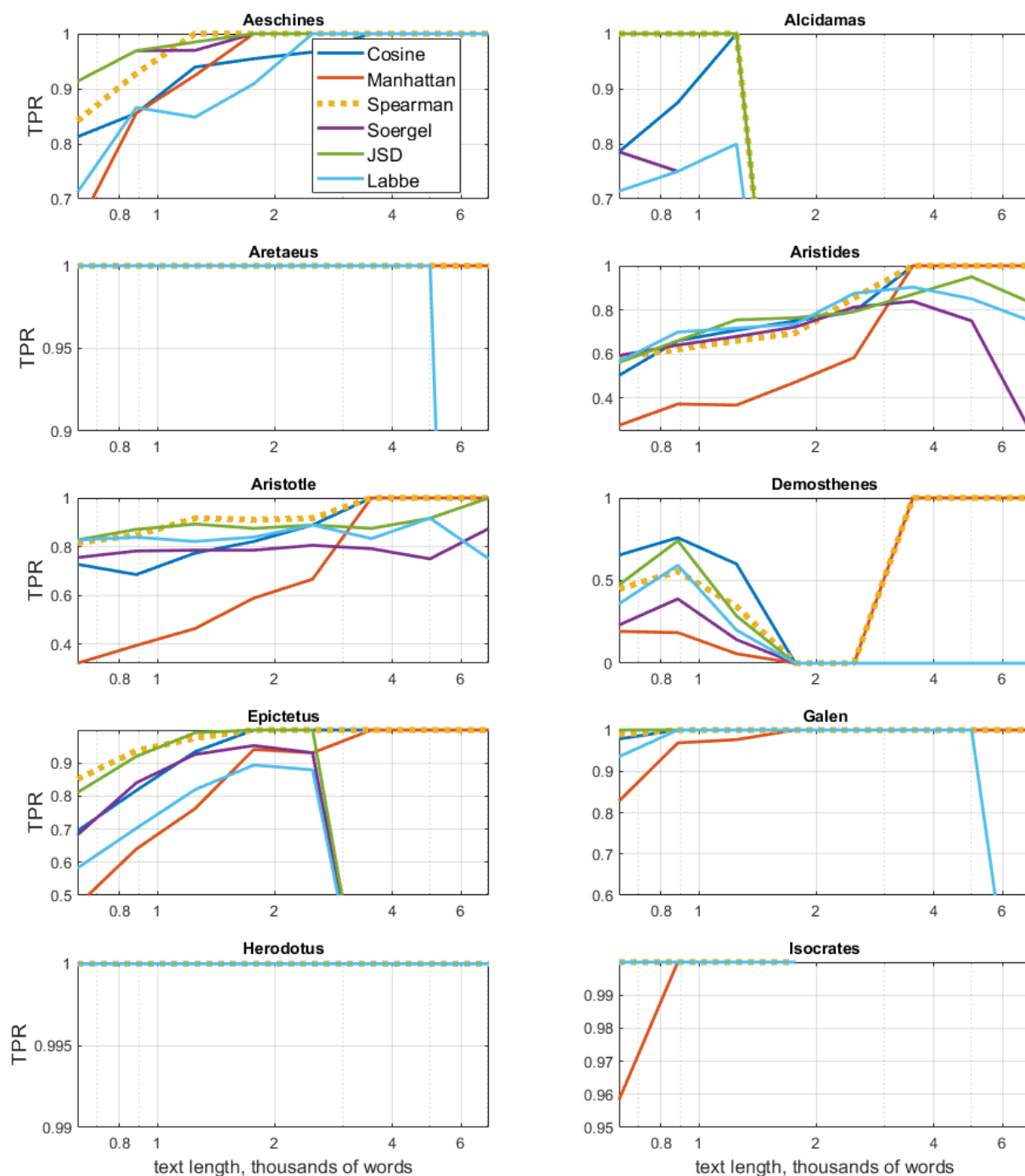


Figure 9b. Best TPR for the “Greek” corpus. End
Рисунок 9b. Наилучший TPR по корпусу «Греки». Окончание

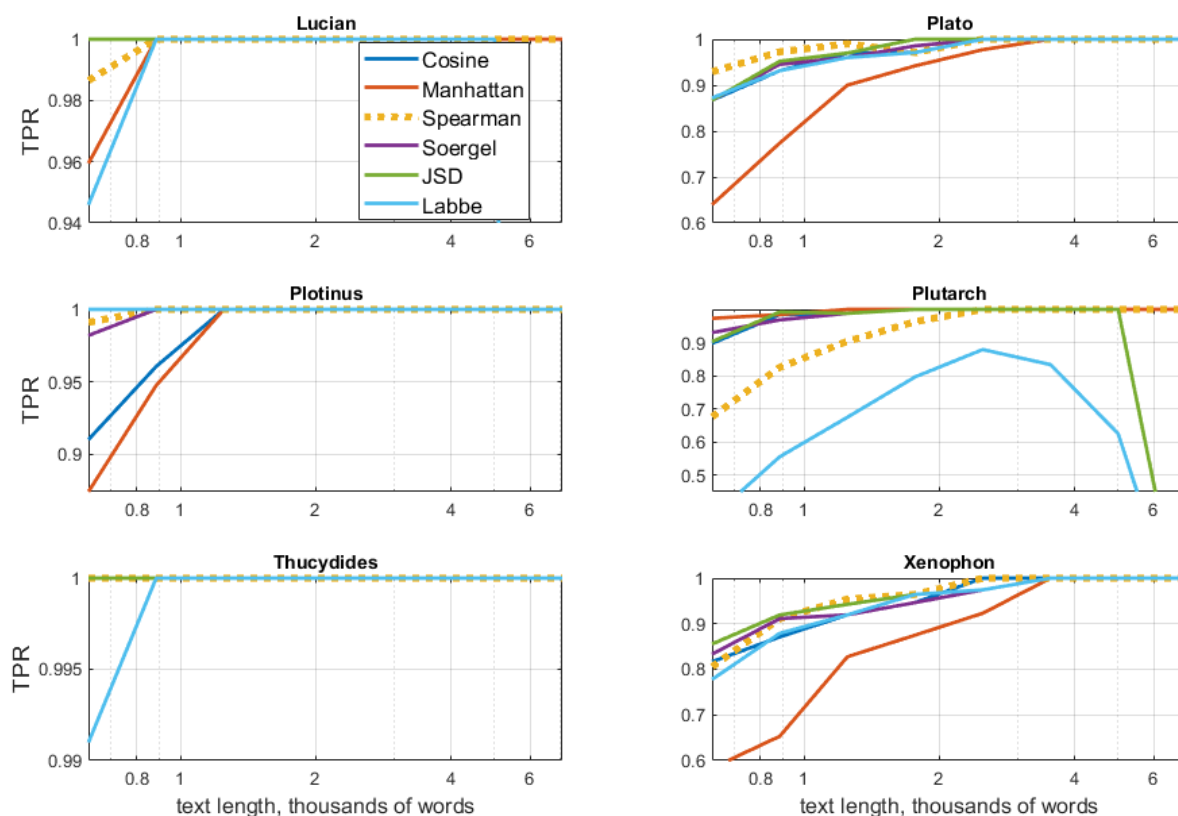


Figure 10. Best TPR for the “Rowling” corpus

Рисунок 10. Наилучший TPR по корпусу «Роулинг»

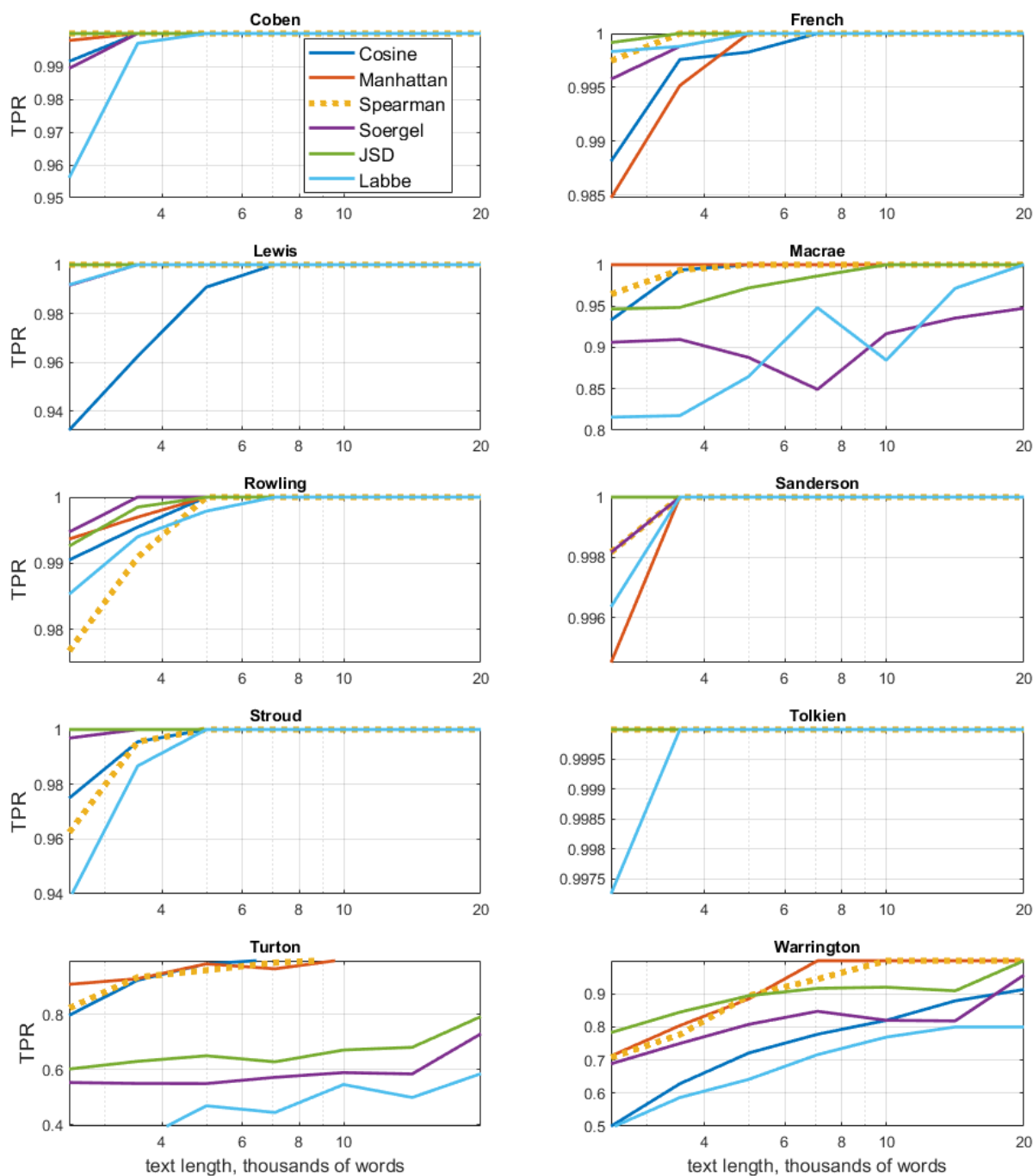


Figure 11. Best TPR for the “Sholokhov” corpus
Рисунок 11. Наилучший TPR по корпусу «Шолохов»

