Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond Delta: Introducing an Angle Metric for Stylometric Similarity

Olga G. Gorina

gorina@bk.ru

National Research University Higher School of Economics

Natalya S. Tsarakova

National Research University Higher School of Economics

Research Article

Keywords: stylometry metrics, authorship attribution, Burrow's Delta, Cosine Delta, Delta, Angle metrics

Posted Date: April 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8457236/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Beyond Delta: Introducing an Angle Metric for Stylometric Similarity

Olga G.Gorina, |HSE University (0000-0002-9184-326X), gorina@bk.ru,

Natalia S. Tsarakova, HSE University (0000-0003-3276-6484) n. carakova@gmail.com

ABSTRACT

This article evaluates stylometric distance metrics, including established (Burrows's Delta) and lesser-known methods (Argamon's Delta, Cosine Delta, Minkowski metrics). It introduces a novel angle-based method that compares word vectors in n-dimensional space using original relative frequencies, eliminating the need for a reference corpus or z-score transformation. The proposed approach shows promise for assessing stylistic similarity, even in short texts.

Key words: stylometry metrics, authorship attribution, Burrow's Delta, Cosine Delta, Delta, Angle metrics

1 Introduction

The focus throughout is on exploring the effectiveness of various methods for measuring stylistic distance in texts that vary in length, particularly how different stylometric methods perform in terms of reliability when analyzing relatively short texts. Another important parameter to investigate is the set of the Most Frequent Words (MFW), which is integral to Burrows's Delta and other Delta methodologies. Through systematic experimentation, this study aims to provide a comprehensive understanding of how stylistic distance can inform authorship attribution.

Of note that established metrics typically rely on a selected corpus of works by known authors to ascertain the authorship of texts with unknown origins; these texts must be of significant length. In contrast, angle metrics, as a relatively new approach developed in this study, allow for conducting pairwise comparisons among texts without relying on a corpus of known authorship works. Thus, the objectives of angle metrics research are twofold: first, to measure its effectiveness in general and second, to assess the possibility of pairwise comparisons of shorter texts. The results underscore the significance of selecting appropriate methods and parameters while recognizing the evolving nature of authorship in the digital age.

The main distinction of the angle metric studied by the authors lies in the fact that we operate with relative frequencies for the words constituting the MFW set, rather than with z-scores of these frequencies. All relative frequencies, consequently, obviously lie within the range [0..1]. Accordingly, if we consider the representation of a text as a vector directed from the origin of an n-dimensional space to a point defined by the values of the relative frequencies of the words constituting the MWF-set, then the angle between two compared texts will obviously lie within the range of 0-90 degrees. Unlike the cosine distance method, we will reason in terms of the value of the angle itself, not its cosine. For convenience and to emphasize the distinction between our method of assessing angular distance and the cosine delta.

In our study, we define Most Frequent Words (MFW) strictly as the highest-frequency lexical items without excluding function words (conjunctions, prepositions, articles, auxiliaries). We maintain an inclusive approach towards function words, which are far from being neutral systematically co-vary with content words to create measurable stylistic patterns. Consequently, as far as MFW is concerned authorial style emerges from the entire lexical ecosystem, not just content words or function words. In linguistic classifications grammar or function words normally include pronouns, prepositions, auxiliary and copular verbs, determiners, etc., which have no independent meaning (O'Keeffe et al., 2007; Scott and Tribble 2006).

1.1 Related Work: General Overview

Stylometry is known to be a field that focuses on the statistical investigation of one's style features, such as sentence length, lexical richness, syntactic complexity, and hapax legomena (Gómez-Adorno et al. 2018). One of the prominent tasks that falls within the scope of this field is authorship attribution, which aims to identify the author of a given text through quantitative techniques (Holmes 1994). In this sense, researchers strive to uncover an authorial fingerprint by analyzing various text characteristics, which ultimately sheds light on the unique writing style of an individual (Holmes 1994; Juola 2008). A prime example of this can be found in the linguistic investigation of "The Cuckoo's Calling" where word frequencies were used to prove that the novel, published under the pen name of Robert Galbraith, was actually written by J.K. Rowling (Zimmer 2013).

Apart from literary studies, stylometric techniques are also used in historical analysis (Baum 2017), forensic investigations (Rocha et al. 2017; Yang and Chow 2014), and philosophy, positioning attribution as an important branch of interdisciplinary research. In fact, academic literature demonstrates that the aforementioned techniques can be used to detect plagiarism (Suman et al. 2022), aid in the cybercrime investigation process (Abbasi and Chen 2005), and provide historical insights into political writings (Ivanov 2016). These and other related tasks require varying methodological approaches that, based on current research, are associated with distance-based techniques (Stamatatos 2009). In the subsequent section, we will discuss this approach separately to provide an overview of how researchers address the statistical notion of authorship in their work.

1.2 Distance-based measures

John Burrows is known to have proposed a new Delta measure to operationalize the stylistic resemblance of two texts. The major idea behind his approach is that documents in question can be presented in vector space, where each word corresponds to a different dimension (Burrows 2002). This allows researchers to utilize geometric measures such as the Euclidean or the Manhattan distance to assess the similarity of two documents (ibid.).

In "A Vector Space Model for Automatic Indexing," Salton, Wong, and Yang (1975) explain how texts can be converted into vectors within a multi-dimensional space, facilitating similarity calculations. This approach, introduced in author's work forms the foundation of the vector space model (Salton et al. 1975).

Relying on this approach, calculating the Burrows's Delta can be broken down into six steps (Evert et al. 2017; Stamatatos 2009). First, text data is preprocessed to extract words and their frequencies. This is followed by feature selection, in which the most common words across all the researched texts are chosen (typically, these include conjunctions, articles, and other function words). During the third stage, word frequencies are normalized to account for varying document lengths. Concurrently, mean and standard deviation values are calculated for each word to employ z-score transformation. Finally, the Delta score is computed by summing the absolute differences between individual z-scores for each text pair being compared (i.e. the Manhattan distance) — smaller Delta scores suggest a higher probability that the documents share the same author. This approach was successfully evaluated on literary works, making the proposed method a staple in the field of authorship attribution (Burrows 2002; Hoover 2004).

Despite being positively received by the academic community, the Delta statistic has been revisited by various scholars with the single purpose of improving its performance (Argamon 2007; Eder 2015; Smith and Aldridge 2011). Argamon was among the first to propose changes to the initial formula, asserting that the use of z-score transformation lacks mathematical justification as "distance ranking by Burrows's Delta is equivalent to probability ranking using what is known as the Laplace distribution" (Argamon 2007, p. 7). In this sense, the author states that the classical standardization is not suitable due to the fact that it relies on the assumption of a normal distribution. To address this issue, he suggests two modifications to the original formula: first, he introduces the Linear Delta that employs parameters of the Laplace distribution instead of z-scores; second, he proposes the Quadratic Delta that retains z-scores but substitutes the Manhattan distance with the Euclidean one. Additionally, Argamon points out that delta-like measures are based on the assumption that word frequencies are independent, which is not always true (Argamon 2007). To solve this problem, he introduces the Rotated Delta, which makes use of the whitening technique. Evert and his colleagues summarize this technique in the following way: "the raw frequency matrix is rotated by a 'covariance matrix' determined from the whole corpus or from a reference corpus, thus eliminating the dependencies between word frequencies before calculating the Euclidean

distances" (Evert et al. 2017, p. ii7). Thus, the Rotated Delta can be seen as a non-axis parallel Quadratic Delta. While Argamon's theoretical claims may appear valid, empirical studies reveal that the suggested modifications do not yield any significant improvements over the initial approach (Evert et al. 2017).

Another crucial alteration of the Delta can be seen in the work of Smith and Aldridge who make use of the Cosine similarity measure (Smith and Aldridge 2011). In this sense, researchers suggest considering the angle between two vectors in a high-dimensional space, which in their experiment shows much accurate results compared to the Burrows's Delta. Specifically, the paper highlights the statistically significant difference in accuracy between the two metrics at a significance level of 0.05 and suggests adopting the classical formula to a trigonometric similarity function (Smith and Aldridge 2011). Following this publication, several scholars investigated the proposed claims, ultimately concluding that employing Cosine similarity measure indeed enhances the overall accuracy. For example, Evert and his colleagues performed a number of experiments to test multiple Delta modifications using a set of novels written in three languages (Evert et al. 2017). They discovered that the alteration proposed by Smith and Aldridge consistently outperformed other variations of the metric, which is important to note for further research.

2 Methodology

Stylometric analysis in this study treats a text as a frequency-ordered word list. We focus solely on word frequency methods, representing each text as a list of words paired with their relative frequencies, sorted descending. A subset of the most frequent words (MFW) words act as coordinate axes, transforming the text into a point or vector in an n-dimensional space, where n is the size of the MFW set. To facilitate this analysis, various denotations and measurements should be introduced. One such denotation is T_i, which represents each text that is being analyzed. The size of the text vocabulary is denoted as V_i**.** The aggregate of word frequencies in the text, referred to as N_i, equals the total number of running words in the text.

As the analysis we conducted involves comparisons across a corpus of $\boldsymbol{k}$ selected texts. Each text is indexed by $\boldsymbol{i}$, where $\boldsymbol{i}$ varies from $\boldsymbol{1}$ to $\boldsymbol{k}$. As was mentioned, the sum of the frequencies of all words in a given text's wordlist corresponds to the overall length or size of that text.

Of note is that in order to compare frequency characteristics accurately, it is necessary to use the relative frequencies of words in the text. If this approach is not employed, then the comparison would involve values that are not comparable. Thus, this paper operates with the text, which is represented as a table of ordered pairs $(w_i ; f_i)$, where w_i represents the word and f_i represents the relative frequency of its occurrence in the text, arranged in descending order. Throughout the article, the term "frequency" refers to relative frequency unless specified otherwise.

At the core of the methods being discussed lies the hypothesis formulated in the late 19th century and the early 20 century [1] , which suggests that an author's style can be discerned by examining their usage of most frequent words (MFW), often classified as grammar words or function words. Thus, the methods being considered utilize a set, or, precisely, a set of these commonly used words, which is generated for a given corpus of texts. In our understanding, it is crucial for this set of words to be the same for all texts under comparison when determining authorship. In order to compare frequency characteristics, the relative frequencies of the set words are used in relation to the each corpus text in question. Essentially, the relative frequencies of the set words serve as coordinates in an *n*-dimensional space, where *n* represents the size of the set. Consequently, each text is assigned a point in this n-dimensional space.

Thus, the comparison of two texts can be represented by their respective points in an ***n***-dimensional space. If the texts are identical, their points will coincide, while different texts have distinct points. To assess the stylometric distance between the texts, a mathematical quantity that measures the distance between these points is used; this measurement should determine whether the distance is large or small. A small distance indicates that the texts are stylistically similar, whereas a large distance suggests dissimilarity. It is crucial to emphasize that all comparisons are carried out using a specific comparison model such as Burrows's, Argamon's, Cosine Delta, employing a specific distance measure, and focusing on a specific set of function words. This last point is particularly significant.

Consequently, it is evident that using the same calculation models relying on various sets yields different results. Therefore, it is essential to ensure that the same measures and sets are employed when comparing the results of stylometric differences across different texts.

To introduce our methodology let us first turn to Burrows's Delta. The method involves a corpus of ***k*** texts for comparison (T_1, T_2,...., T_k) a MFW set of size ***n*** (w_1, w_2,...., w_n) . For each text T_j in the corpus, an ordered representation is compiled as follows:

(1) (f_{1j}, f_{2j},...., f_{nj})

[1] The Polish-German school of Antiquaries and, in particular, Vincetius Lutoslawski proposed to count grammar (function) words in texts, on the basis of which to make cautious statements about the time of their creation. At the beginning of the XX century, these studies inspired Morozov N. A. to count grammar words to determine the authorship of the text.

Wincenty Lutosławski (1897) *The Origin and Growth of Plato's Logic. With an Account of Plato's Style and of the Chronology of His Writings.* — London. —547 p.

Morozov, N. A. (1915) Linguistic spectra: A means to distinguish plagiarism from the true works of one or another famous author: A Stylometric etude // Izvestiya of the Department of Russian Language and Literature of the Imperial Academy of Sciences. — St. Petersburg,— Vol. XX. Book 4. — pp. 93-134.

Where f_{ij} represents the frequency of the ***i*th** word from the set in the ***j*th** text from the corpus of texts. Subsequently, a linear transformation (2)

(2) $$z_{ij} = \frac{f_{ij} - \mu_i}{\sigma_i}$$

is applied to the frequency representation of the text (1) to obtain a **z**-score representation (3) of the ***j*th** text from the corpus:

(3) $$(z_{1j}, z_{2j},...., z_{nj})$$

In formula (2) $\boldsymbol{\mu_i}$ denotes the average value of the frequencies of the ***i* th** word across all the texts in the corpus, and $\boldsymbol{\sigma_i}$ represents the corresponding standard deviation.

Once we have obtained the z-scores for each text, which can be depicted as points in ***n***-dimensional space, we can proceed with comparison operations using a specific measure.

For the purpose of comparison (4), Burrows's approach employs the Minkowski measure with a parameter value of ***p*=*1***, which is normally referred to as L_1 :

(4) $$\|T_k T_m\|_p = (\sum_{i=1}^{n} |z_{ik} - z_{im}|^p)^{1/p}$$

This measure is commonly referred to as the Manhattan distance, which the linear transformation (2) expresses as (5):

(5) $$\|T_k\ T_m\| = \sum_{i=1}^{n} \frac{|f_{ik} - f_{im}|}{\sigma_i}$$

It is often recommended to use the averaged distance value, which is normalized by the size of the set. This can be expressed as follows:

(6) $$\|T_k\ T_m\| = \frac{1}{n} \sum_{i=1}^{n} \frac{|f_{ik} - f_{im}|}{\sigma_i}$$

When comparing texts from different corpora with different MFW sets, using the average value is advisable. This average value, calculated using equation (5), is commonly known as the Burrows's Delta. As was mentioned, there are several variations of the Burrows's Delta, including the Eder Delta proposed by Eder et al. (2016), which incorporate the rank of a word in the frequency-ordered set MFW:

(7) $$\|T_k\ T_m\| = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{|f_{ik} - f_{im}|}{\sigma_i} \cdot \frac{n-i+1}{n}\right)$$

This is done so that more frequent words from the MFW contribute more to the distance calculation. From our perspective, to achieve this effect, it is sufficient to use relative frequencies as coordinates rather than their z-scores.

By setting the value *p*=2, which corresponds to Minkowski metrics L_2 , the Euclidean distance between the z-scores (8) can be obtained (Argamon 2007); this measure is often referred to as the Argamon'sDelta.

(8) $$\|T_k\, T_m\|_2 = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(z_{ik} - z_{im})^2}$$

In their study Smith and Aldridge (2001) make use of the Cosine Delta, an analytical tool in the form of the Cosine of the angle α between the z-estimate vectors of pairwise taken texts, which is expressed as the ratio between the scalar product of vectors representing the z-scores and the product of the lengths of these vectors based on the Euclidean distance.
In their study, Smith and Aldridge (2001) used Cosine distance as their metric, as shown in formula (9):

(9) $$\|T_k\, T_m\|_2 = 1 - \frac{\sum_{i=1}^{n} z_{ik} \cdot z_{im}}{\sqrt{\sum_{i=1}^{n}(z_{ik})^2} \cdot \sqrt{\sum_{i=1}^{n}(z_{im})^2}}$$

It is important to discuss the approaches described above. Applying the Z-transformation ensures that each word in the set contributes equally to the construction of the text vector. Each of ***n*** dimensions' values
varies within approximately the same range. If, instead, relative word frequencies were used as values, the values along coordinate axes corresponding to more frequent words would differ by many times from those of less frequent words, which is a direct consequence of the rank-frequency distribution nature.
The methods described above works under the assumption that the words included in the MFW set contribute equally to shaping an author's style; however, this assumption might not be self-evident. Moreover, the assumption that all MWF should contribute equally to the formation of the text vector lacks any theoretical justification. It is worth explaining that if we were to employ the Manhattan distance, Euclidean distance, or any Minkowski measure with $\boldsymbol{p > 2}$, the resulting distances between texts would range from **0** to virtually unlimited values, depending on both the dimension of the MFW set and the specific value of ***p***. Furthermore, even when comparing the same pair of texts within the same collection, using the same distance metric but varying the size

of the MFW set, we would again observe significantly different distance values. As a result, interpreting the values of these stylometric distances becomes problematic, since it is impossible to determine whether a particular distance should be considered 'large' or 'small'. Alternatively, the Cosine distance avoids this shortcoming, as its values lie within the range **[0..2]** when Z-transformation is employed. However, the use of z-transformation still implicitly assumes that each MWF word contributes equally to the formation of the text vector, which is not a given. Furthermore, we argue that the more frequent an MWF word is, the greater its contribution to the vector formation should be.

The Curse of Dimensionality. Another challenge is the so-called *curse of dimensionality*, a term coined by Richard E. Bellman (Bellman, 1961). This concept refers to the difficulties in measuring distances using different metrics in high-dimensional linear spaces. The core issue lies in the exponential growth of the volume of a "unit cube" as dimensionality increases. In high-dimensional spaces, data become extremely sparse, distances expand, and the distinction between "near" and "far" neighbors diminishes or blurs.

Extensive research has characterized the challenges of distance metrics in high-dimensional spaces, where the dimensionality ***n*** fundamentally alters their behavior. Aggarwal et al. (2001) proved that when measuring distances in linear spaces of large dimension using the Minkowski measure metric (L_p), the accuracy of the obtained values decreases with increasing value of p.

It has been shown that for large values of ***n***, using an indicator with a coefficient of ***p*<1** gives more accurate results (Weber et al., 1998; Berchtold et al., 1998). However, as is well known, the Minkowski measure with a coefficient ***p*<1** is not exactly a metric since for such a measure, the axiom of the triangle inequality does not hold. To circumvent this problem, as well as to make it possible to compare the distances obtained from (i) different collections of texts, (ii) various dimensional MFW sets and (iii) different metrics choices, we linearly normalize all distance values to the interval **[1..2]**:

(10) $$r_i = \frac{(R_i - R_{min})}{(R_{max} - R_{min})} + 1$$

Where R_i is the computed value of the distance, R_{max} and R_{min} are respectively the maximum and minimum computed values of distances among all compared text pairs in the entire table. Therefore, r_i is the normalized value for the distance R_i.

It has to be clarified that, strictly speaking, expression (10) might not be considered to act as a metric because it does not satisfy the identity axiom. However, there are reasons for using (10) as a metric under our

circumstances. Firstly, this "metric" is only used on a restricted set of texts, none of which are identical. Secondly, identical texts have not been compared with each other, since such a comparison would be meaningless. Other metric axioms, as can be easily verified, will hold.

3 Study

3.1 Experiment #1

The first experiment uses the selected corpus that comprises 16 works from four distinguished authors, consisting of four Harry Potter novels by J.K. Rowling, the detective novel "The Cuckoo's Calling" published under a pseudonym, four novels by F. Scott Fitzgerald, four novels by Raymond Chandler, and three novels by Charlotte Brontë.

The texts selection is a deliberate mix that combines temporal and stylistic diversity to test the robustness of stylometric analysis. The corpus is presented in Table #7 of Appendix; for each select novel, the date of publication, the size and the text genre are provided. While the authors (Brontë, Chandler, Fitzgerald, Rowling) represent different eras and genres, each author's chosen works are stylistically consistent within their own corpus (e.g., Chandler's detective novels share genre-specific patterns; Rowling's Harry Potter and detective novel contrast genres but retain core stylistic traits). This allows the study to isolate stable authorial features from era- or genre-related variations, ensuring the model identifies authorship even across intentional shifts in style or time period.

Thus, the essence of the experiment lies in representing all 16 works as points in an n-dimensional space, with varying MFW-set's dimensions ranging from the maximum (1,096 words) to the minimum (100 most frequent words). The maximum MFW- set consists of words appearing at least once in every compared text. We then sort this MFW-set by word frequency across the selected texts and progressively reduce it—first to 500 words, then to 100.

At each stage, our custom software computes pairwise distances between texts using:

1. Argamon's Delta (equivalent to Euclidean/L_2 metric),
2. Burrows's Delta (equivalent to Manhattan/L_1 metric),
3. Cosine Delta with Z-score normalization,
4. Minkowski metric with fractional exponent $p = 0.5$.

To enable comparison, results are normalized to a unified scale using transformation (10).

Next, we perform bottom-up hierarchical clustering via the complete-linkage method. Initially, each text is its own cluster. At each step, the two closest clusters merge. Distance between clusters is defined as the maximum pairwise distance between their elements. Theoretically, we would expect four clusters, each corresponding to one author. The final output is a symmetric distance matrix, where the cell at row «*i*» and column «*j*» gives the distance between clusters «*i*» and «*j*». Thus, the diagonal entries report: number of texts in the cluster, mean inter-cluster distance, and variance of inter-cluster distances. These results are tabulated in Appendix Tables 1–3, with distances visualized via grayscale gradients reflecting lighter gray colour as smaller distance.

Our key findings for full MFW-set sized 1,096 words suggest that all metrics successfully separated texts by authorship. Charlotte Brontë's works exhibited the strongest stylistic similarity with lowest mean and variance. Additionally, intra-cluster distances were significantly smaller than inter-cluster distances. For reduced MFW sets of 500 and 100 words the results are as follows. All the metrics *except Cosine Delta* merged *The Cuckoo's Calling* with Scott Fitzgerald's works, despite high intra-cluster distances that indicates stylistic divergence. Meanwhile, Cosine Delta maintained correct authorship clusters for Brontë and Chandler, whereas Rowling's and Fitzgerald's texts clusters showed high variance.

3.2 Experiment # 2

It differs in that we added K. Fox's genre-divergent work to the corpus, adjusting the maximum MFW set to 993 words; each word in this set now appears in every text. For this set of 993 words, all the metrics except Cosine Delta correctly clustered texts by authorship, isolating K. Fox as a distant outlier. Cosine Delta also separated The Cuckoo's Calling into its own cluster. Notably, intra-cluster distances shrank significantly vs. the 16-text experiment, suggesting that adding a highly divergent text made the remaining works stylometrically closer. For the reduced MFW set of 500 and 100 words, Cosine Delta performed nearly identically to the full-set case of 993 words.

Of note that Argamon's and Burrows's Deltas again merged The Cuckoo's Calling with Fitzgerald's works. The important result is that only the fractional Minkowski metric (p=0.5) preserved correct authorship groupings. Results are detailed in Appendix Tables 4–6.

3.3 Experiment #3

The purpose of this experiment was to develop an accurate method for determining the stylometric similarity between any pair of texts, specifically shorter texts of up to 5,000 words. This method aims to attribute authorship to student essays, preventing the misuse of AI-enabled ghostwriting tools for academic assignments. It is important that this method deals with pair –wise texts comparisons.

The research methodology used a size ***n*** MFW- set for both texts, represented as (w_1, w_2,...., w_n) , along with frequency representations (1) for each text. Every text was depicted as a point in an n-dimensional space, where coordinates were derived from their respective frequency representations. The degree of similarity between the pair of texts was ascertained by calculating the angle separating the vectors originating from the coordinate system's inception point and reaching the corresponding points within the ***n***-dimensional space.

A salient departure from the Delta family methods lies in the coordinate assignment for each text's point; rather than employing z-scores, frequency representations dictate these coordinates. This approach is characterized by a significant attribute: all n coordinates consistently reside within the range **[0..1]**. Consequently, in the designated space, the angles formed between the vectors corresponding to any two texts persistently span between 0 and 90 degrees. For the sake of convenience, the angle between the vectors, originating from the n-dimensional space's origin and leading to the points representing two texts under comparison, is henceforth designated as the angle between the texts.

To avoid confusion with the cosine similarity measure, results are presented as the angle between vectors in degrees, not its cosine. This analysis is based on several hundred pairwise comparisons of texts. The first computational step identifies the maximum possible MFW-set, which is MFWmax; it defined as all the words present in both text. This set is sorted in descending order by the average pooled relative frequency of each word across both texts, and alphabetically in cases of equal frequency.

As a result, four important observations are noted:

1. For large texts (over 100,000 words), the size of MFWmax constitutes about 3–4% of the combined vocabulary. For small texts (under 5,000 words), the size constitutes 10–15%.

2. For large texts, words from the MFWmax set account for over 90% of all word tokens. For small texts, they account for 60–70%.

3. Larger texts produce a larger MFWmax.

4. When the compared texts differ significantly in size (by an order of magnitude or more), the key parameters—text coverage by the MFW-set and its proportion of the total vocabulary—diverge greatly for each text. Consequently, the angular distance between them increases substantially. For illustration, the angle calculated between a full text and 1/16 of that same text corresponded to a value indicating a significant difference.

Therefore, for more accurate results, it is best to compare texts of roughly the same order of magnitude in size.Using the MFWmax, the angular distance between the texts was calculated. We then systematically reduced this MFW-set by removing the least frequent tenth of words from its end and recalculated the angle. This process of reduction and recalculation was repeated until the MFW-set contained only 10 words.

From a large number of such calculations, the following consistent patterns emerged for all comparisons:

1. As the MFW-set is reduced, the angle between texts remains stable until the set contains approximately 100 words.

2. With further reduction below 100 words, the angular distance begins to decrease.

3. An MFW-set of 100 words consistently covers about half of any given text, regardless of its total size.

These calculations were performed using selected texts from Experiments 1 and 2. To clearly demonstrate the influence of text size, one of the compared samples was a 1/16 segment of text vs a whole text.

In the experiment, the angles were compared for the following pairs of texts:

1. J. K. Rowling.Harry Potter and the Chamber of Secrets -- Harry Potter and the Prisoner of Azkaban
2. J. K. Rowling. The Cuckoo's Call -- Kate Fox Watching the English
3. J. K. Rowling Harry Potter and the Prisoner of Azkaban -- 1/16 Kate Fox Watching the English
4. F. Scott Fitzgerald. The Great Gatsby -- J. K. Rowling.Harry Potter and the Chamber of Secrets
5. J. K. Rowling. The Cuckoo's Call -- Harry Potter and the Chamber of Secrets
6. J. K. Rowling. The Cuckoo's Call -- Harry Potter and the Prisoner of Azkaban
7. F. Scott Fitzgerald. The Great Gatsby -- Kate Fox Watching the English
8. Charlotte Bronte Jane Eyre -- Charlotte Bronte Villette
9. Charlotte_Bronte_Jane Eyre -- The Professor
10. Charlotte Bronte Villette --The Professor
11. Kate Fox Watching the English -- 1/16 Kate Fox Watching the English

Figure 1 shows how the angular distance between texts changes as the size of the MFW-set is reduced.

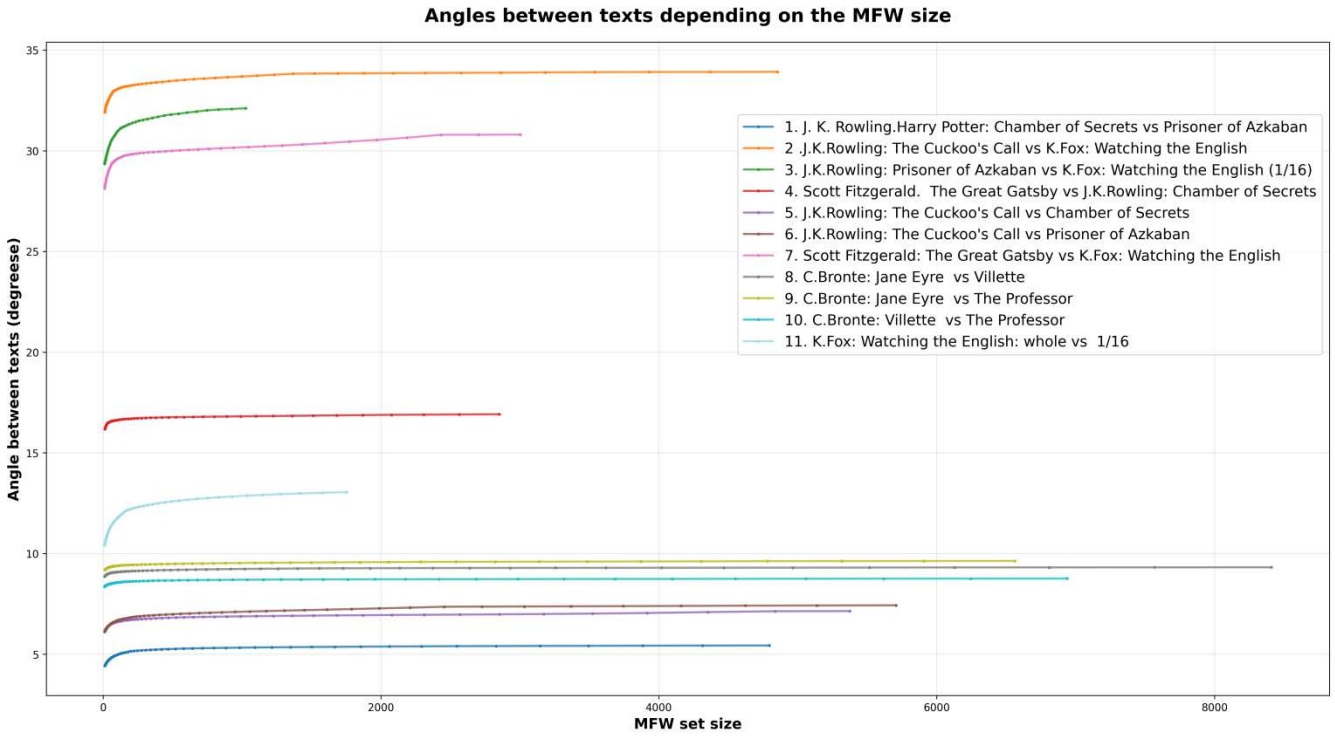


Fig. 1. Angle between texts for different MFW sizes, for the text pairs numbered in the list above.

The graphs clearly demonstrate that for most text pairs, the angle remains constant while the **MFW-set** is reduced from its maximum size down to about 100 words. The angle only starts to decrease when the set is reduced further.

This pattern changes when comparing texts of very different lengths, as seen for pairs 3 and 11. For these, the angle begins to decrease immediately as the **MFW-set** is reduced from its maximum size. This effect is especially pronounced for pair 11, which compares a full text with one-sixteenth of itself, resulting in a significant measured angle.

Key Observations from Figure 1:

- Kate Fox's Work: Consistent with Experiments 1 and 2, the work by Kate Fox remains stylometrically distinct. The angle between it and other texts (pairs 2,7) stays within the 30–35 degree range, whether comparing full texts or a full text with a fragment.
- J.K. Rowling's Harry Potter Series: These works are stylometrically close to each other (pair 1). Notably, *The Cuckoo's Calling* shows a consistently small angle when compared with two other Rowling novels (pairs 5,6) a finding that differs from the results of Experiments 1 and 2.

- Charlotte Brontë's Novels: These three works again show strong stylometric proximity, with a stable pairwise angle of about 9 degrees.(pairs 8,9,10)
- Less Closely Related Works: The comparison between F. Scott Fitzgerald and J.K. Rowling (pair 4) consistently shows an angle greater than 15 degrees.

Conclusions from the Analysis:

Based on a large number of pairwise text comparisons, we can draw the following conclusions:

1. For more accurate results, compare texts of roughly the same length. A two-fold size difference has little impact, but a difference of an order of magnitude (10x) distorts results significantly.
2. For similarly sized texts, an angular distance of 5–15 degrees indicates stylometric closeness.
3. An angle of 15–25 degrees signifies a clear stylometric difference.
4. An angle greater than 25 degrees means the texts are stylometrically very distant. At this level, differences in text length become irrelevant.
5. If the compared texts are very different in length, or if both are short (under 10,000 words), use the maximum MFW-set (MFWmax) for the most accurate angle calculation.
6. For significantly large texts (over 50,000 words), a MFW-set of the top 100 most frequent words is sufficient for a reliable angle calculation.

3.4 The Novelty of the Suggested Metrics

Our method is novel primarily in its approach. While all Delta methods require a pre-existing corpus of texts with known authorship to assess an unknown text, our method directly addresses the core task of comparing two texts. In today's context, there is a growing need to evaluate just two texts for stylistic similarity—often irrespective of authorship—including texts of different lengths and even short texts under 5,000 words (for example, a tutor checking student essays for plagiarism or AI-generated content).

After calculating angles between several hundred texts, we observed the following:

- The measured angles typically range from 3 to 40 degrees.
- The angle remains stable when the MFW-set is reduced in size, as long as its words still cover over half of each text. A gradual decrease in the angle only begins when the set is reduced beyond this point.
- The top 100 most frequent words nearly always cover about half of a text. If these words are present in both texts, they can effectively serve as the MFW-set.

The key innovation is that pairwise comparison using Burrows's Delta is not practical without a reference corpus, as results depend heavily on MFW-set size and text length. Typically, such a corpus of stylistically similar texts is needed to calibrate distances. In contrast, our method directly calculates the angle between two texts of any length and classifies them as stylometrically close or distant based on this value (e.g., using the thresholds defined in our conclusions: the 5–15°, 15–25°, 25°+ scale).

5 Results and Conclusions

The L1 and L2 metrics perform reliably only with large MFW-sets; their accuracy diminishes as the set size is reduced. The fractional Minkowski metric demonstrates greater stability, with smaller coefficient values yielding better results. However, like the others, any Minkowski metric depends on vector length and is highly sensitive to the dimensionality of the set. Therefore, when comparing large texts, MFW-sets smaller than 100 words should be avoided.

The Cosine metric, however, produces stable results even with a reduced MFW-set. Moreover, Cosine Delta yields significantly larger distances both within and between clusters, indicating it is a more stringent measure. Its stability with smaller sets also enables reliable comparisons between texts of substantially different lengths.

In Experiment #3, we applied a pairwise comparison method based on the angular distance between texts. This approach offers a major advantage over Delta family methods, as it does not require a pre-existing corpus of related texts. The results show little variation in angular distance even when the number of MFWs is decreased. The calculated angles, reflecting stylistic dissimilarity, range from a few degrees for highly similar texts to over 30 degrees for texts with considerable stylistic differences. Notably, the method performs exceptionally well on short texts of 5,000 words or less.

Consequently, we believe that ready-made statistical packages and authorship attribution software are best suited for standard investigations where their established procedures apply.

However, for more complex research tasks, it may be necessary to engage with the underlying theory and adapt the methodology to the specific problem. This could involve selecting custom features, adjusting methodological parameters, or implementing advanced machine learning techniques. Ultimately, the choice of method depends on the task's nature, complexity, and the research objectives.

Another promising direction for future research involves analyzing GPT-generated texts. Assessing their "authorship" raises fascinating questions. For instance, measuring the stylometric distance between human-

authored texts and GPT outputs could reveal how effectively AI mimics human style and could aid in detecting AI-generated content in academic settings. The widespread acceptance of such texts would also challenge and potentially expand our definitions of authorship, with implications for ethics, literary analysis, forensics, and intellectual property law.

Acknowledgements

The publication was prepared within the framework of the Academic Fund Program at HSE University in 2022–2023 [grant № 23-00-005]

References

Abbasi, A., Chen, H. (2005). Applying authorship analysis to extremist-group Web forum messages. IEEE Intelligent Systems, 20(5), pp. 67–75. https://doi.org/10.1109/MIS.2005.81

Aggarwal C., Hinneburg A., Keim D. (2001) On the Surprising Behavior of Distance Metrics in High Dimensional Space // Lecture Notes in Computer Science. V 1973pp 420-434. doi:10.1007/3-540-44503-x_27.

Argamon, S. (2007). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. Literary and Linguistic Computing, 23, pp. 131–147. https://doi.org/10.1093/llc/fqn003

Baum, A. D. (2017). Content and Form: Authorship Attribution and Pseudonymity in Ancient Speeches, Letters, Lectures, and Translations—A Rejoinder to Bart Ehrman. Journal of Biblical Literature, 136(2), pp. 381–403. https://doi.org/10.15699/jbl.1362.2017.200369

Bellman, R.E. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

Burrows, J. (2002). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing, 17(3), pp. 267–287. https://doi.org/10.1093/llc/17.3.267

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

Eder, M. (2015). Taking Stylometry to the Limits: Benchmark Study on 5,281 Texts from "Patrologia Latina." ADHO 2015 - Sydney. https://dh-abstracts.library.cmu.edu/works/2364

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 (suppl_2), ii4–ii16. https://doi.org/10.1093/llc/fqx023

Gómez-Adorno, H., Posadas-Duran, J.-P., Ríos-Toledo, G., Sidorov, G., Sierra, G., Gómez-Adorno, H., Posadas-Duran, J.-P., Ríos-Toledo, G., Sidorov, G., Sierra, G. (2018). Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts. Computación y Sistemas, 22(1), pp. 47–53. https://doi.org/10.13053/cys-22-1-2882

Holmes, D. I. (1994). Authorship Attribution. Computers and the Humanities, 28(2), pp. 87–106.

Hoover, D. L. (2004). Testing Burrows's Delta. Literary and Linguistic Computing, 19(4), pp. 453–475. https://doi.org/10.1093/llc/19.4.453

Ivanov, L. (2016). Using Alliteration in Authorship Attribution of Historical Texts. In P. Sojka, A. Horák, I. Kopeček, and K. Pala (Eds.), Text, Speech, and Dialogue. pp. 239–248. Springer International Publishing. https://doi.org/10.1007/978-3-319-45510-5_28

Juola, P. (2008). Authorship Attribution. Foundations and Trends® in Information Retrieval, 1(3), pp. 233–334. https://doi.org/10.1561/1500000005

Kreuz, R. (2023). Linguistics fingerprints. How language creates and reveals identity. Prometheus Books, Guilford, (CT). — 283 p.

Lutosławski, W. (1897) *The Origin and Growth of Plato's Logic. With an Account of Plato's Style and of the Chronology of His Writings.* — London. — 547 p.

Morozov, N. A. (1915) Linguistic spectra: A means to distinguish plagiarism from the true works of one or another famous author: A Stylometric etude // Izvestiya of the Department of Russian Language and Literature of the Imperial Academy of Sciences. — St. Petersburg,— Vol. XX. Book 4. pp. 93-134.

O'Keeffe A., McCarthy M., Carter R. (2007) From Corpus to Classroom: language use and language teaching. — Cambridge: Cambridge Univ. Press. — 315 p.

Peter, W. H., Smith Aldridge, W. (2011) Improving Authorship Attribution: Optimizing Burrows' Delta Method*. Journal of Quantitative Linguistics, 18(1), pp. 63–88, February.

Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A. R. B., and Stamatatos, E. (2017). Authorship Attribution for Social Media Forensics. IEEE Transactions on Information Forensics and Security, 12(1), 5–33. https://doi.org/10.1109/TIFS.2016.2603960

Salton, G., Wong, A., Yang, C.S. (1975). A vector space model for automatic indexing, CACM 18(11), 613-616.

Savoy, J. (2020). Machine learning methods for stylometry. Authorship attribution and author profiling. Cham: Springer.

Scott, M., Tribble, C. (2006) Textual Patterns: key words and corpus analysis in language education: Studies in Corpus Linguistics. — Amsterdam/Philadelphia: John Benjamins. — 200 p.

Smith, P., Aldridge, W. (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method*. Journal of Quantitative Linguistics, 18, 63–88. https://doi.org/10.1080/09296174.2011.533591

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556. https://doi.org/10.1002/asi.21001

Suman, C., Saha, S., and Bhattacharyya, P. (2022). Authorship Attribution of Scientific Abstracts. 2022 26th International Conference on Pattern Recognition (ICPR), 1522–1528. https://doi.org/10.1109/ICPR56361.2022.9956343

Yang, M., Chow, K.-P. (2014). Authorship Attribution for Forensic Investigation with Thousands of Authors. In N. Cuppens-Boulahia, F. Cuppens, S. Jajodia, A. Abou El Kalam, and T. Sans (Eds.), ICT Systems Security and Privacy Protection (pp. 339–350). Springer. https://doi.org/10.1007/978-3-642-55415-5_28

Zimmer, B. (2013, July 16). "The Cuckoo's Calling": The Science That Uncovered J.K. Rowling's Literary Hocus-Pocus (Update). Wall Street Journal. http://blogs.wsj.com/speakeasy/2013/07/16/the-science-that-uncovered-j-k-rowlings-literary-hocus-pocus/

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendixnew.docx