Studia Metrica et Poetica 5.2, 2018, 29–54

Versification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry

Petr Plecháč, Klemens Bobenhausen, Benjamin Hammerich*

Abstract: This article describes pilot experiments performed as one part of a long-term project examining the possibilities for using versification analysis to determine the authorships of poetic texts. Since we are addressing this article to both stylometry experts and experts in the study of verse, we first introduce in detail the common classifiers used in contemporary stylometry (Burrows' Delta, Argamon's Quadratic Delta, Smith-Aldridge's Cosine Delta, and the Support Vector Machine) and explain how they work via graphic examples. We then provide an evaluation of these classifiers' performance when used with the versification features found in Czech, German, Spanish, and English poetry. We conclude that versification is a reasonable stylometric marker, the strength of which is comparable to the other markers traditionally used in stylometry (such as the frequencies of the most frequent words and the frequencies of the most frequent character *n*-grams).

Keywords: authorship attribution; stylometry; versification; Czech verse; German verse; Spanish verse; English verse

1. Introduction

The most frequent task type that we encounter in authorship attribution (AA) begins with the following situation: we have a text of unknown or doubted authorship (the target text) and a set of candidate authors. Contemporary stylometry has developed extremely accurate and sophisticated methods for handling this task type. Their underlying logic is that one can determine the author by measuring the degree of stylistic similarity between the target text and specific texts written by candidate authors. Various style markers are taken into account for this purpose: frequencies of words, frequencies of parts of

* Authors' addresses: Petr Plecháč, Czech Academy of Sciences, Institute of Czech Literature, Na Florenci 1420/3, 11000 Praha 1, Czech Republic. E-mail: plechac@ucl.cas.cz; Klemens Bobenhausen, Metricalizer, Erwinstr. 76, 79112 Freiburg im Brsg., Germany. E-mail: bobenhausen@gmail.com; Benjamin Hammerich, Metricalizer, Erwinstr. 76, 79112 Freiburg im Brsg., Germany. E-mail: benjamin.hammerich@gmail.com.

https://doi.org/10.12697/smp.2018.5.2.02

speech, frequencies of character *n*-grams, frequencies of collocations, etc. One important aspect of style (of one important form of literature) is however so far completely disregarded by stylometry, and that aspect is versification.

In this article, we first argue why versification can and should be employed within the AA process (section 2), and then touch upon the history of quantitative AA methods and provide a detailed description of those we employ in our own work (section 3).[1] Finally, we report the results of our experiments comparing ordinary AA analysis and versification-based analysis of Czech, German, Spanish, and English poetry texts (sections 4 and 5).

2. Motivations

There are several reasons for assuming that versification analysis may be useful in AA; to name a few of the most important:

- Most of the features used in stylometry (such as words and *n*-grams) amount to what are known in statistics as "rare events". Therefore, rather large text samples are required.[2] However, in practice these are rarely available in practice for AA of poetry texts – usually a single poem or just a few are in question, not an entire collection. On the other hand, versification features are usually Boolean (e. g. stressed/unstressed position), or can take on only a limited number of values (e. g. rhythmic types), and thus may be analyzed even with significantly smaller samples.
- Versification is much more topic-independent than the usual stylometric features (words, word and character *n*-grams, etc.) – vocabulary may change considerably across poems of different genres written by the same author, but we may assume that their rhythm and rhyming technique will remain more or less stable.
- Some verse experts have stated that rhythm is more complicated to forge than lexicon.[3]

---

[1] A more detailed history of the field may be found in Juola 2006; Koppel, Schler, Argamon 2009; Stamatatos 2009.

[2] Eder 2013 states that 5000 words are a minimum sample, but Eder revises this estimation in Eder 2017 to make it highly dependent on the composition of the corpus.

[3] Compare: "Rhythm is inertia created by the chain of verses. And this inertia is individual for every poet. It is easy to forge a word. But in order to forge a verse rhythm the forger has to study the imitated rhythm very hard, and [forgers are usually] not prepared for this" (Tomashevsky 1923/2008: 238; English translation from Lotman 2015: 145).

- Some stylometrists propose combining different features within a single analysis, e. g. the most frequent words + character *n*-grams + word *n*-grams (cf. Mikros, Perifanos 2013; Eder 2011), but the frequencies of these features are strongly correlated. Versification, on the other hand, should be almost entirely independent of these. We thus may expect the combined analysis of lexicon and versification to be more powerful than the analysis of lexicon alone.

3. History and Related Works

Many scholars trace the origins of quantitative approaches to AA to the works of T. C. Mendenhall (1841–1924), namely his papers "The Characteristic Curves of Composition" (1887) and "A Mechanical Solution of a Literary Problem" (1901). In the first work, he suggests the capturing of the peculiarities of an authorial style via the distribution of the relative frequencies of word-lengths measured by number of characters. According to Mendenhall, if the samples are large enough (he recommends 100,000 words), the curve defined by these values should be more or less stable in the works of one author, but should have different shapes in works by different authors. In the second work, he employed this method within a real-world question of authorship – the work attributed to William Shakespeare. Mendenhall compares the shape of the curves extracted from Shakespeare's works with those of Francis Bacon and Christopher Marlowe (see Figures 1 and 2) and cautiously concludes that Bacon could not have written Shakespeare's work, but that there is a strong evidence that Marlowe actually did so (1901: 104–105). The differences in the curves of Shakespeare and Bacon were, however, later found to have been caused by the comparing of the versified texts of the former with the non-versified texts of the latter (see Williams 1975), and Mendenhall's method was swept off the table.

It is worth noting that long before Mendenhall's lexical analysis, there were attempts to shed some light upon the authorship of Shakespeare's works based on the quantification of verse rhythm and rhyme. These included Malone 1787; Weber 1812: 166; Spedding 1850; and especially the works of *New Shakespeare Society* members such as Ingram 1874, and Fleay 1874, 1876. But these studies seem not to have had a real impact on the later development of stylometry (cf. Grieve 2005; Grzybek 2014).

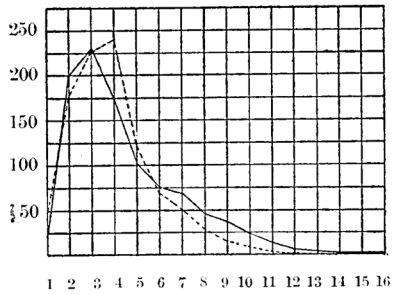

Figure 1. Relative frequencies (per mille) of word-lengths measured by number of characters in works of W. Shakespeare (dashed) and F. Bacon (full line). Source: Mendenhall 1901: 104 (facsimile).

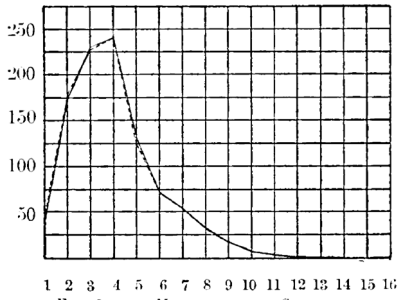

Figure 2. Relative frequencies (per mille) of word-lengths measured by number of characters in works of W. Shakespeare (dashed) and C. Marlowe (full line). Source: Mendenhall 1901: 105 (facsimile).

During the first half of the 20th century, many textual characteristics – such as sentence length (Yule 1938) and the measurement of vocabulary richness (Simpson 1949) – were proposed for the purposes of authorship attribution (see Juola 2006: 240–241). What all these methods share in common is a basis in *univariate statistical analysis*. Although they may characterize text by a set of numerical values (as in Mendenhall's case above) rather than by a single one, the attribution itself is based on the comparison of single values (e. g.: Shakespeare uses 4-letter words more often than Bacon), not on the information that may be extracted from the whole. None of these early methods are considered reliable today, as has been shown in many comparative studies (e. g. Grieve 2007).

Since the groundbreaking study by Mosteller and Wallace (1964) on the authorship of the *Federalist Papers*, modern stylometry has turned to the much more reliable multivariate approach. Various approaches have been proposed that employ various methods of multidimensional statistics and machine learning and rely on complex textual characteristics such as the frequencies of the most frequent words, character *n*-grams, and POS-tags (to name just a few). Versification, however, has not been taken into account.

However, despite the lack of interest by mainstream stylometry, versification was still widely used as a discriminant for authorship during the 20th century in the studies performed by certain verse experts – namely those associated with the so-called "Russian School" of metrics. In the early 1920s, for example, Boris Tomashevsky used versification as evidence proving that the end of Pushkin's unfinished poem *Mermaid,* which Dmitry Zuev had claimed to have found in 1889, was a forgery (Tomashevsky 1923/2008: 238–239). Other instances of the

use of verse rhythm and rhyme as evidence of authorship include the rejection of the authenticity of alleged fragments of *Eugen Onegin's* tenth chapter (Lotman, Lotman 1986), the questioning of the authenticity of works newly added to Alexander Ilyushin's edition of Gavriil Batenkov's poems (Shapir 1997, 1998), and especially the extensive work by Marina Tarlinskaja on Shakespeare and his contemporaries (1987, 2014 in particular).[4]

Due to these versification-based approaches' isolation from the main branch of stylometry, there has been a large gap arising between stylometry's more and more advanced methods and these approaches, which have continued in the use of rather simple methods of univariate statistics. In the sections below, we would like to illustrate this through several examples. First we will explore the example of the univariate versification-based approach provided by the above-mentioned verse expert Marina Tarlinskaja (section 3.1). Then we will focus on one multivariate lexically based model, which has been the most widely used in the field of literary studies in recent years, the "Burrows' Delta" measure (section 3.2) and its later modifications (section 3.3). We will also briefly mention one popular machine-learning method, called the Support Vector Machine (section 3.4). Finally, we will attempt to combine the advantages of versification analysis presented in section 3.1 with those of the advanced multivariate models presented in sections 3.2, 3.3 and 3.4, which – to the best of our knowledge – has so far been reported in one article only, written by programmers, and covering old Arabic poetry, which differs greatly from modern European versifications (Al-Falahi, Ramdani, Bellafkih 2017).

3.1. Marina Tarlinskaja: The Attribution of *Henry VIII*

In her book *Shakespeare and the Versification of English Drama, 1561–1642* (2014), Marina Tarlinskaja provides many examples of versification-based AA. Let us focus here upon one such example: the attribution of the 17th-century play *Henry VIII*.

Most scholars agree that *Henry VIII* was a collaborative text, wherein certain recognizable parts were written by John Fletcher (the "A" parts) and the remainder by William Shakespeare (the "B" parts). Tarlinskaja (2014: 140–149) brings in evidence for this hypothesis from the domain of versification, namely

[4] In part due to the influence of Tarlinskaja, versification has been used as an argument in 20th century Shakespearean studies even within the Western tradition, namely in the long discussion on authorship of *Funeral Elegy* between Don Foster on one side and Ward Elliot and Robert J. Valenza on the other (see Grieve 2005: 6–8).

the distribution of what she calls "strong syntactic breaks" after particular metrical positions. Her argument is that the A and B parts differ in whether the peak of the distribution takes place after the 6th or 7th syllable, and that the entire distributions found in the A parts and B parts are similar to those found in the two authors' other texts written in the very same period: Fletcher's *Bonduca* and Shakespeare's *Tempest* respectively (see Figure 3).

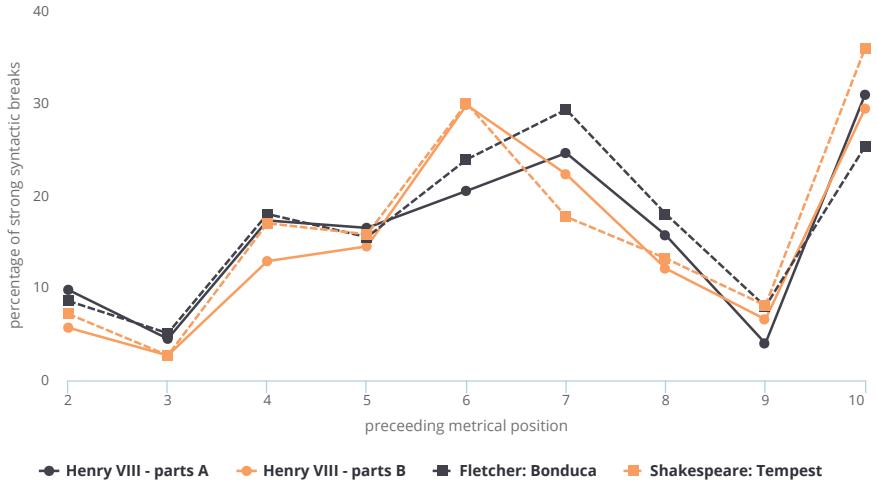

Figure 3. The frequencies of strong syntactic breaks after particular metrical positions in the A and B parts of *Henry VIII*, and in Fletcher's *Bonduca* and Shakespeare's *Tempest*. Source: Tarlinskaja 2014: Table B.3.

While this is a strong and valid argument, we may assume that even more reliable evidence may be gathered by moving from a univariate approach (the comparison of individual values) to a multivariate one (the comparison of entire sets of values). Let us also add here that a solely visual comparison may be misleading (especially when dealing with large data sets). In the sections below, we will present several ways in which this problem is being handled in modern stylometry.

3.2. Burrows' Delta measure

Burrows (2002, 2003) has proposed Delta as a measure of the stylistic difference between texts. It is based on a comparison of the frequencies of the *n* most frequent words (MFW), and it deals with the above-mentioned question of

how to funnel multiple differences into one value. Burrows' solution is rather simple and intuitive. Let us illustrate it with an example based on data provided by Burrows himself (2002: 270–272).

We will use a model situation – there is a manuscript entitled *Paradise Lost* and strong evidence that it was written by either John Milton or Aphra Behn. The question, of course, is: which of these two is the real author? To find an answer, we collect one set of texts that were provably written by Milton and one set of texts that were provably written by Behn. Next we reach for the n words that are the most frequent across all the texts collected. For the sake of clarity, let us work with $n = 20$, even though much larger numbers (from hundreds to thousands) are usually used in Burrows' Delta. What we are attempting to do here is to find out which of the two authors has works that are more similar to *Paradise Lost* in terms of the relative frequencies of these words.

The most straightforward method would be to plot the frequencies (Figure 4) and compare the curves, just as Mendenhall (section 3) and Tarlinskaja (section 3.1) did. But such a visual judgment is rather vague and unreliable in this case. What Burrows suggests instead is to express the degree of dissimilarity between the texts as the mean value of the differences between the frequencies of specific words. But – as we know and as we may also observe in Figure 4 – word frequencies generally tend to decrease rapidly after the top ranks (Zipf's law). Thus the difference between the frequencies of the most frequent word will be generally much larger than the difference between the frequencies of the 50th or 100th most frequent word in any given body of texts. So as to be able to consider each word as a marker of equal weight, Burrows transforms the frequencies of individual words into z-scores. Very roughly speaking, such a transformation shrinks or extends the frequency ranges so that the ranges are approximately the same for each word (see Figure 5).[5]

[5] More precisely, it transforms the distribution into another one with mean = 0 and standard deviation = 1. For a particular word in text T having a frequency of f_T, z-score is calculated as follows: $z_T = (f_T - \mu) / \sigma$, where μ stands for the mean frequency of the word across all texts, and σ stands for its standard deviation.

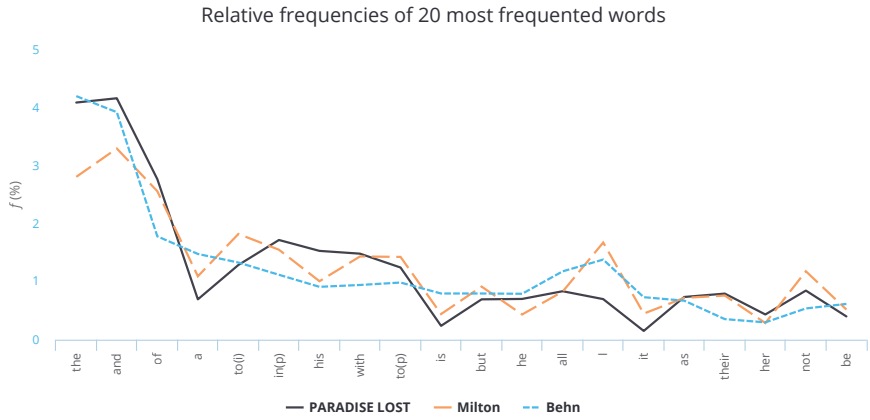

Figure 4. The relative frequencies of the 20 most frequent words in Milton's *Paradise Lost*, other works by Milton, and the work of Aphra Behn (1640–1689). The characters in parentheses disambiguate homographic forms: i = infinitive and p = preposition. Source: Burrows 2002: 270–272.

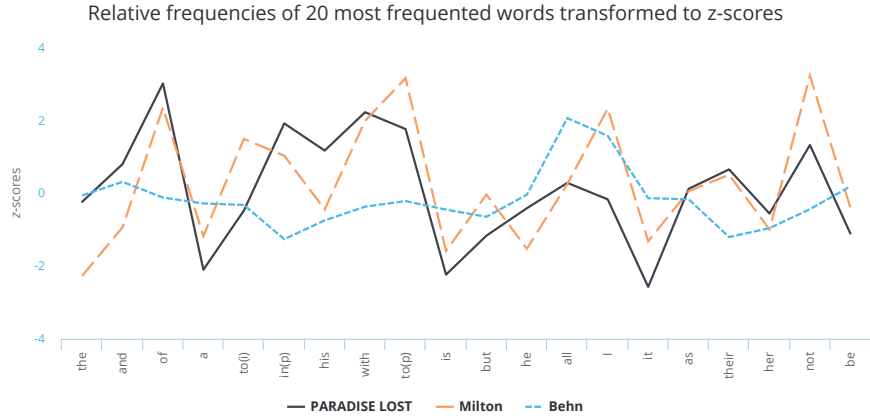

Figure 5. The relative frequencies of the 20 most frequent words in Milton's *Paradise Lost*, other works by Milton, and the work of Aphra Behn (1640–1689). Transformation into *z*-scores based on all of the samples in the corpus. The characters in parentheses disambiguate homographic forms: i = infinitive and p = preposition. Source: Burrows 2002: 270–272.

At this point, each of the three samples (*Paradise Lost*, Milton, and Behn) is represented by a set of 20 values corresponding to the frequencies of the 20 words transformed into *z*-scores. In order to find which of the two candidates is closer to the values of *Paradise Lost*, Burrows chose the most intuitive method:

(1) For each word in each candidate sample, count the difference between its own frequency transformed to *z*-score and the one found in *Paradise Lost*. As we are only interested in the size of the differences, not their directions, we work with their absolute values.

(2) The value of the Delta measure between each of the candidate samples and *Paradise Lost* is then calculated as the arithmetic mean of the particular differences. Figure 6 shows the entire procedure and the expected result, wherein John Milton is found to be more similar to his own work than Aphra Behn.

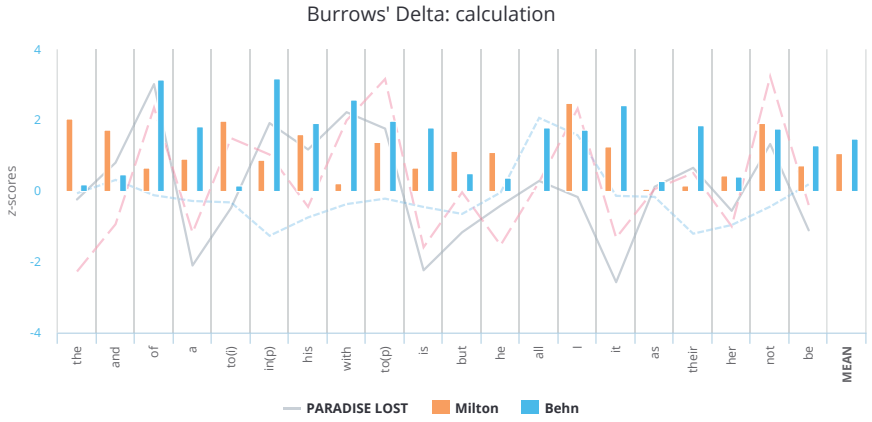

Figure 6. The calculation of Burrows' Delta between Milton's work and *Paradise Lost* and Behn's work and *Paradise Lost*.

Through a simple generalization of the procedure above, we may arrive at the general formula for Burrows' Delta: let $z_i(T_1)$ be the *z*-score for the relative frequency of a word in the text (or set of texts) T_1 that is the *i*-th most frequent in the entire corpus, and $z_i(T_2)$ be the *z*-score for the relative frequency of the same word in the text (or set of texts) T_2. The Delta score for T_1 and T_2 based on the *n* most frequent words ($\Delta_n(T_1,T_2)$) is then calculated as follows:

[1]

$$\Delta_n(T_1, T_2) = \frac{\sum_{i=1}^{n} |z_i(T_1) - z_i(T_2)|}{n}$$

3.3. The Mathematical Basis for and Modifications to Burrows' Delta

Soon after its publication, Delta became a popular and widely used method within AA. Several modifications have been proposed (e. g. Hoover 2004a, 2004b), but what seems to be the most important step forward from today's perspective is Shlomo Argamon's interpretation of Delta's key principle. Argamon (2006) has pointed out that the method intuitively proposed by Burrows is in fact equivalent to what is known in mathematics as a measurement of *Manhattan distance*, and that Delta may thus be considered as an instance of the *nearest neighbor classifier*. In the following section, we will clarify this relationship using a "real-world" example and will mention some other alternatives used in contemporary AA, namely the Euclidean-distance-based Quadratic Delta (Δ^{Q}) proposed by Argamon himself (2008) and the cosine-similarity-based Cosine Delta ($\Delta^{\angle}$) proposed by Smith and Aldridge (2011).

Since Manhattan distance actually takes its name from the grid of east-west Streets and north-south Avenues in New York City's borough of Manhattan, let us locate our example there. Imagine a pedestrian seeking the shortest path from the Chelsea Hotel to the Empire State Building (Figure 7). It doesn't matter if he chooses to walk 10 blocks north via 7th Avenue and then east via 33rd Street (as indicated by the red line in the map below), or to instead walk e. g. one block east via 23rd Street, 10 blocks north via 6th Avenue, and the rest via 33rd Street; the distance is always the same – it is the simple sum of the individual distances walked along the streets (d_1) and avenues (d_2). This corresponds to the Manhattan distance between these two buildings on a two-dimensional map: $D_{\mathrm{MAN(2)}} = d_1 + d_2$. The Euclidean distance, on the other hand, corresponds to the path "as the crow flies". Notice that as long as we know d_1 and d_2, this distance may easily be calculated using the Pythagorean theorem: $D_{\mathrm{EUC(2)}} = \sqrt{d_1^2 + d_2^2}$. Finally, the cosine similarity may be roughly depicted as the size of the angle under which these two buildings are seen by some distant observer (we will reach the precise formula later).

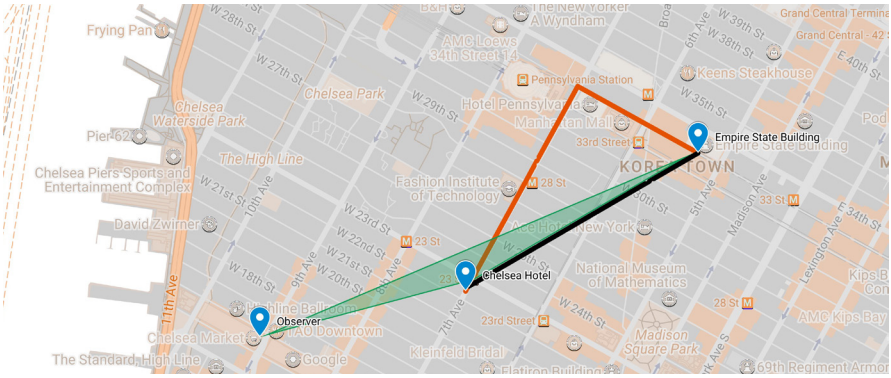

Figure 7. Manhattan distance (red), Euclidean distance (black), and cosine similarity (green). "Real-world" examples in 2D.

Now consider a similar task in 3-dimensional space – here we are interested in the distance from the *entrance* of the Chelsea Hotel to the *very top floor* of the Empire State Building (Figure 8). To get the Manhattan (pedestrian) distance, we simply sum up the distances walked along streets (d_1) and avenues (d_2), along with the elevator ride (d_3): $D_{MAN(3)} = d_1 + d_2 + d_3$. To get the Euclidean distance (a phantasmal crow's flight), we start by taking the direct connecting line on the ground, which – as has been shown above – is equal to the root of the sum of the squares of the distances along streets and avenues ($\sqrt{d_1^2 + d_2^2}$). We may then complete our calculation of the Euclidean distance by using the Pythagorean theorem once again – with this connecting line being one leg of the right triangle, and the vertical distance (the elevator ride) being the other (see Figure 8), thus in this case:

[2]

$$D_{EUC(3)} = \sqrt{\sqrt{d_1^2 + d_2^2}^2 + d_3^2} = \sqrt{d_1^2 + d_2^2 + d_3^2}$$

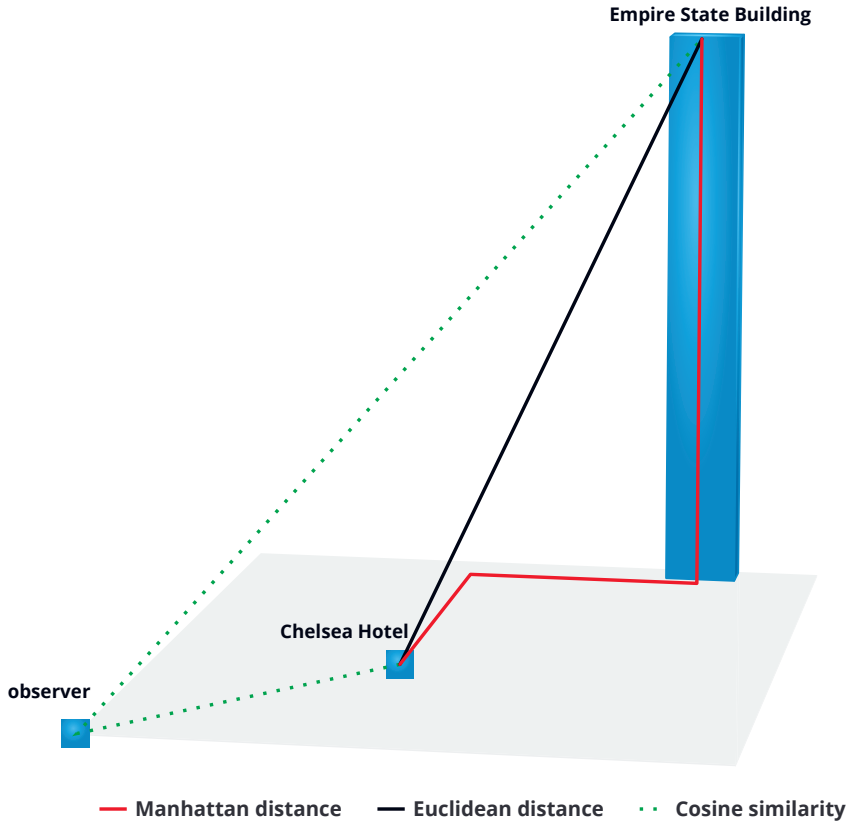

Figure 8: Manhattan distance, Euclidean distance, and cosine similarity. "Real-world" examples in 3D.

How does all this relate to Burrows' Delta? Let us return to our model situation that deals with *Paradise Lost* (T_1), the other works of John Milton (T_2), and the works of Aphra Behn (T_3). For the sake of clarity, we will focus only on the three most frequent words (1: *the*; 2: *and*; 3: *of*) and the *z*-scores of their frequencies in these three samples: $z_1(T_1)$; $z_2(T_1)$; $z_3(T_1) \mid z_1(T_2)$; $z_2(T_2)$ … As has been shown above, Burrows' Delta between T_1 and T_2 is in this case equal to:

[3]

$$\Delta_3(T_1, T_2) = \frac{\sum_{i=1}^{3} |z_i(T_1) - z_i(T_2)|}{3} = \frac{|z_1(T_1) - z_1(T_2)| + |z_2(T_1) - z_2(T_2)| + |z_3(T_1) - z_3(T_2)|}{3}$$

When we plot the sets of *z*-scores representing these two samples as data points (vectors) in a 3-dimensional diagram (Figure 9), it becomes clear that this is equal to their Manhattan distance divided by 3. And as long as we use Delta purely to rank the candidate samples, it makes no difference if we divide all distances by a constant (the number of words analyzed) or not. We may thus simplify the calculation to be fully equal to the Manhattan distance:

[4]

$$\Delta_3(T_1, T_2) = D_{MAN(3)}(T_1, T_2) = |z_1(T_1) - z_1(T_2)| + |z_2(T_1) - z_2(T_2)| + |z_3(T_1) - z_3(T_2)|$$

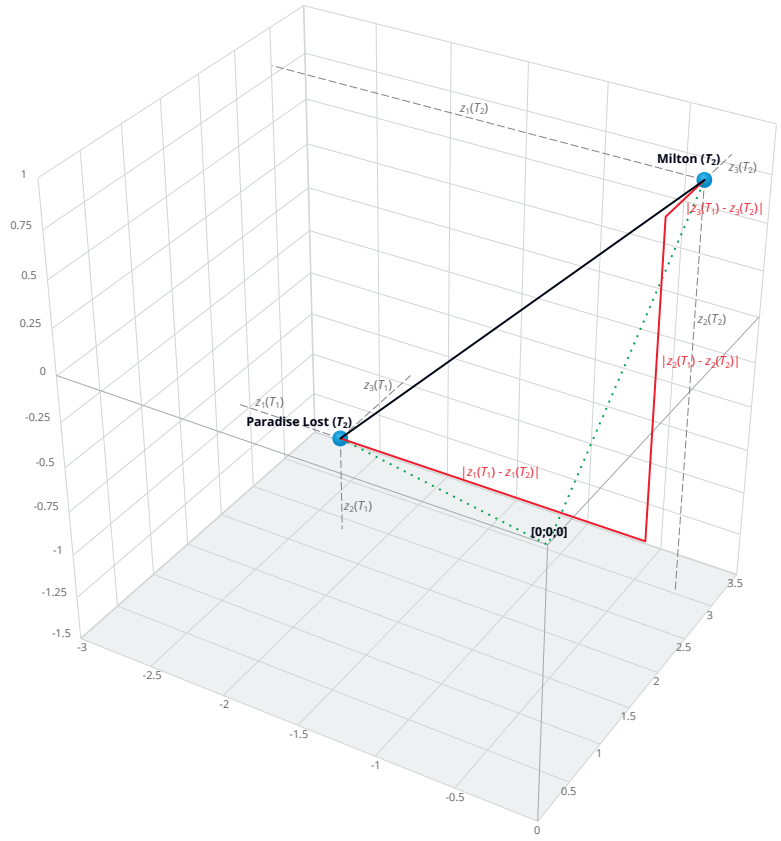

Figure 9. The Manhattan distance, Euclidean distance, and cosine similarity between ***Paradise Lost*** (T_1) and other works by Milton (T_2), as represented by the ***z***-scores of the three most frequent words (***the; and; of***).

Burrows' Delta thus turns out to be an instance of a *nearest neighbor classifier* – it simply picks the author of the candidate sample that is the *nearest* one to the target text in terms of its Manhattan distance. And even though the human imagination is limited to the 3-dimensional space, mathematics is not – no matter whether we choose to work with the 25 or the 500 most frequent words, the principle remains the same: we are still seeking the nearest neighbor in a 25-dimensional or 500-dimensional vector space.

The above-mentioned alternatives – the Quadratic Delta (Δ^{Q}) and the Cosine Delta ($\Delta^{\angle}$) – follow the very same principle of nearest-neighbor classifier, just with a different distance function. The Quadratic Delta simply replaces Manhattan with a straightforward Euclidean distance, which in 3-dimensional space – as has been shown above – equals the square root of the sum of the squares of the distances along each dimension:

[5]

$$D_{EUC(3)}(T_1, T2) = \sqrt{(z_1(T_1) - z_1(T_2))^2 + (z_2(T_1) - z_2(T_2))^2 + (z_3(T_1) - z_3(T_2))^2}$$

Via a simple generalization of the procedure above, we can arrive at the formula that applies for n-dimensional vector space in general:

[6]

$$D_{EUC(3)}(T_1, T2) = \sqrt{\sum_{i=1}^{3} (z_i(T_1) - z_i(T_2))^2}$$

And just as the division of each distance by a constant in Burrows' Delta does not affect the final ranking, the same holds true for extracting the root in the Euclidean distance. The formula for Argamon's Quadratic Delta is thus defined as the square of the Euclidean distance:

[7]

$$\Delta_n^Q(T_1, T_2) = (D_{EUC(n)}(T_1, T_2))^2 = \sum_{i=1}^{n} (z_i(T_1) - z_i(T_2))^2$$

Finally, the Cosine Delta takes the cosine similarity of vectors as the ranking principle, or in other words, it takes the cosine of the angle between the lines connecting the data points with the origin of the chart (the data point where all coordinates equal zero); see Figure 9. The logic behind the formula for the calculation of the cosine similarity (cos α):

[8]

$$\cos\alpha = \frac{\sum_{i=1}^{n} z_i(T_1)z_i(T_2)}{\sqrt{\sum_{i=1}^{n} z_i(T_1)^2}\ \sqrt{\sum_{i=1}^{n} z_i(T_2)^2}}$$

is rather complicated, and we are not willing to bother the reader with its detailed derivation. What is, however, worth noting is that the cosine of the angle may range from 1 (maximum similarity) to -1 (maximum dissimilarity). In order to reflect the greater degree of vectors' similarity for lower values and vice versa (as in the case of the original Burrows' Delta and Argamon's quadratic Delta), Smith and Aldridge's cosine Delta is defined as:

[9]

$$\Delta_n^{\angle}(T_1, T_2) = 1 - \cos\alpha$$

3.4. The Support Vector Machine

Metrics from the Delta family have been widely used and successfully tested with various configurations for the number of words analyzed as well as with other features, such as the most frequent character *n*-grams or the most frequent word *n*-grams (see e. g. Eder 2011; Jannidis et al. 2015), but recently it seems that machine-learning methods are poised to overtake them due to the latter methods' even-better performance. Let us thus briefly mention one machine-learning method that is usually judged to be the most powerful one in authorship attribution – the *Support Vector Machine* (SVM).

The SVM is a supervised learning model, which means that the algorithm uses labeled training data to infer a classification function. Let us clarify how it functions with a very simple example based on artificial data: assume that we have a text of unknown authorship (the target text) and a set of 20 text samples from each of two candidate authors (author 1 and author 2). All of the texts are characterized by the *z*-scores of the frequencies of the two most frequent words ("the"; "and").

During the first phase (learning), the SVM is fed with data from author 1 and author 2 (training data), labeled according to who wrote which sample, and attempts to find a function that correctly separates them with respect to their labels. This is done using a *hyperplane* – a subspace of one dimension less than the original vector space. In our example with its 2-dimensional

data, this means a 1-dimensional space, i. e. a line. During the second phase (classification), the hyperplane inferred from the training data is used for classifying the target text.

As the first chart of Figure 10 indicates, there is an infinite number of hyperplanes that can correctly separate our training data, with some of them attributing the target text to author 1, and some of them to author 2. From all of these possible lines, the SVM chooses the one that maximizes the distance to the nearest vector on each side (with these being called *support vectors*), as shown in the second chart of Figure 10. The SVM thus classifies the target text as a text of author 1.

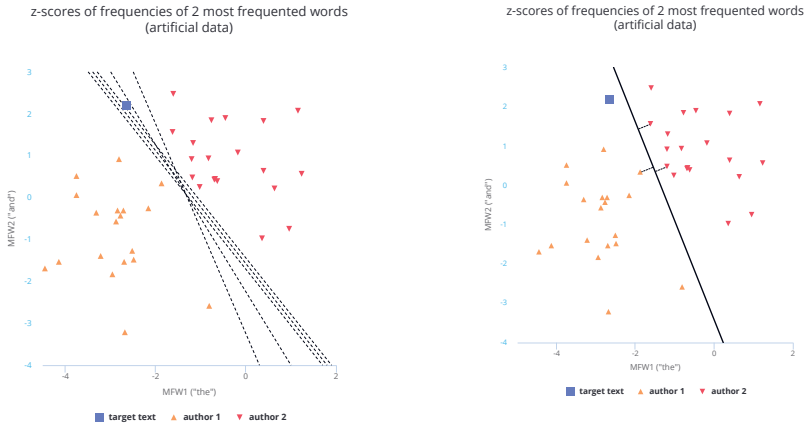

Figure 10. The Support Vector Machine. Left: various possible hyperplanes separating training data from author 1 and author 2. Right: a maximum-margin hyperplane; the dashed lines indicate the distances to the support vectors.

This example is – of course – a very elementary one. Not only do we need to deal with data of much higher dimensions in real-world attribution tasks, but in the vast majority of cases, one is faced with data that is not linearly separable. Furthermore, in most attribution tasks, one needs to perform multiclass classification, i. e. to decide among more than two candidates. The ways in which the SVM deals with these two issues are too complex to be discussed here. We have illustrated at least some of the SVM's very basic principles, and we will note that the SVM (in addition to the distance functions discussed above) is implemented in the scientific libraries of many programming languages (e. g. scikit-learn in Python), where it is ready to be used even without deep knowledge of its workings.

4. Method

Let us now turn to our experiments, wherein we aimed to test whether or not versification features may be considered as authorship markers with effectiveness equal to that of the traditional features used in AA. We worked with four corpora of poetic texts: Czech (CS), German (DE), Spanish (ES), and English (EN), with data taken from the following sources:

- CS: *Corpus of Czech Verse* (Plecháč, Kolár 2015; Plecháč 2016; <http://versologie.cz>)
- DE: *Metricalizer* (Bobenhausen 2011; Bobenhausen, Hammerich 2015; <http://metricalizer.de>)
- ES: *Corpus de Sonetos del Siglo de Oro* (Navarro-Colorado 2015; Navarro-Colorado, Ribes-Lafoz, Sánchez 2016; <https://github.com/bncolorado/CorpusSonetosSigloDeOro>)
- EN: *Chicago Rhyming Poetry Corpus* (Reddy, Knight 2011; <https://github.com/sravanareddy/rhymedata>)

All of the corpora were tokenized, phonetically transcribed, and annotated in terms of their meters and rhymes.[6]

In order to test the above-mentioned hypothesis, we extracted samples of essentially the same size from each corpus (100 lines in CS, DE, and EN; 98 lines – i. e. 7 sonnets – in ES). The samples consisted of lines written in specified meters: masculine and feminine trochaic tetrameters in CS, feminine trochaic tetrameters in DE, hendecasyllabic lines in ES, and masculine iambic pentameters in EN. Each sample was written by a single author, and none of the poems were divided into two or more samples. The samples' basic characteristics are given in Table 1 and Table 2.

[6] All of the annotations were provided by the authors of the corpora, with the following exceptions: the Spanish corpus was phonetically transcribed using eSpeak Speech Synthetizer (<http://espeak.sourceforge.net), and rhymes were annotated by rhymeTagger (Plecháč 2018; <http://github.com/versotym/rhymeTagger>). The phonetic transcription and metrical annotation of the English corpus was performed using the Prosodic parser (<http://github.com/quadrimegistus/prosodic>).

Table 1. The numbers of samples extracted from each corpus, and the numbers of their authors

	# of samples	# of authors
CS	275	21
DE	142	8
ES	323	10
EN	142	12

Each sample was represented by the frequencies of the stressed syllables at particular metrical positions (its stress profile) and the frequencies of particular sounds (a simple way to "caption" the vague notion of "euphony"). In DE, we also worked with various characteristics of rhyme: the vowel length match frequency, the frequency of closed rhymes (i. e. ending with a consonant), and the frequencies of vowel and consonant pairs. All of the values were transformed into *z*-scores. This ensured that each sample was represented by a vector consisting of a few dozen values.

Using this data, we tested all of the above-discussed classifiers: Burrows' Delta (Δ), Argamon's Quadratic Delta (Δ^{Q}), Smith–Aldridge's Cosine Delta ($\Delta^{\angle}$), and the Support Vector Machine (SVM). This evaluation was performed using the "leave one out" cross-validation method. In this method, in order to estimate the accuracy of the classifier, each sample is iteratively picked out to be treated as the target text, with the rest of the samples being treated in that iteration as candidates. The accuracy is than calculated as the percentage of cases in which the real author was recognized successfully.

Apart from versification itself, we also tested the classifiers with: (1) the 100 most frequent words, (2) the 100 most frequent character trigrams, and (3) combined vectors consisting of versification features, the 100 most frequent words, and the 100 most frequent character trigrams.

For each corpus, we also report the value of the random baseline – an estimation of what portion of the samples would be attributed correctly if assignment of authorship were to be completely random. This value is calculated as:

[10]

$$\text{random baseline} = \sum_{i=1}^{N} \left(\frac{a_i}{X}\right)^2$$

where N is the number of authors, X is the number of samples and a_i is the number of samples written by author i.

Table 2. The authors and the numbers of samples from each of them

CS	DE	ES	EN
Čelakovský, František Ladislav (11) Chmelenský, Josef Krasoslav (7) Furch, Vincenc (28) Hajniš, František (6) Havelka, Matěj (17) Hněvkovský, Šebestián (7) Klumpar, Jan Květoslav (9) Kulda, Beneš Metod (32) Macháček, Simeon Karel (5) Nejedlý, Vojtěch (15) Picek, Václav Jaromír (27) Pohan, Václav Alexander (11) Puchmajer, Antonín Jaroslav (6) Rubeš, František Jaromír (9) Ryba, Jakub Jan (10) Šnajdr, Karel Sudimír (8) Šrámek, Jan (6) Tablic, Bohuslav (14) Uhlíř, Josef (10) Villani, Karel Maria Drahotín (20) Vinařický, Karel Alois (17)	Brentano, Clemens (16) Eichendorff, Joseph von (20) Fleming, Paul (13) Gleim, Johann Wilhelm Ludwig (12) Goethe, Johann Wolfgang (23) Heine, Heinrich (26) Lenau, Nikolaus (22) Schlegel, August Wilhelm (10)	Argensola, Bartolome (22) Cetina, Gutierre de (35) Gongora (16) Herrera, Fernando de (45) Litala y Castelvi, Joseph de (21) Quevedo (73) Rojas, Pedro Soto de (17) Tassis y Peralta, Juan de (29) Ulloa y Pereira, Luis de (15) Vega, Lope de (50)	Constable, Henry (6) Crosland, T.W.H. (6) Drayton, Michael (9) Dryden, John (6) Finch, Anne (9) Jonson, Ben (36) Lovelace, Richard (7) Pope, Alexander (7) Smith, Charlotte Turner (31) Spenser, Edmund (11) Swift, Jonathan (5) Wordsworth, William (9)

5. Results

Figure 11 presents the results of our experiments. Because the values of the random baselines differ greatly, it does not make sense to compare the results across specific corpora. On the other hand, a comparison of features and classifiers within particular corpora shows some interesting trends:

1. In each corpus and with each classifier, the success rate for versification features is significantly higher than the value for the random baseline. Versification thus may be considered to be a reasonable stylometric marker.

2. The success rate for versification relative to words and trigrams varies strongly; it is significantly higher in CS, slightly higher with Delta measures and slightly lower with the SVM in ES, slightly lower in DE, and significantly lower in EN.[7]

3. When comparing classifiers among the same feature set, in each corpus the SVM *consistently* gives the best performance out of all the classifiers.[8] Out of the Delta family, Cosine Delta usually evinces the best performance.

4. The combining of versification features with words and trigrams always displays better performance than these features alone.

---

[7] Here let us note that the English corpus comprises several extensive works (such as Spenser's *Faerie Queene*), the chapters of which are treated as separate poems. The extremely high accuracy with words here may thus simply be due to overfitting – the classifiers may actually not be recognizing the author's style, but rather the specific vocabulary of a given work.

[8] Two other machine learning methods were tested, namely random forest and naive Bayes classifier, but these were outperformed in all cases by the SVM.

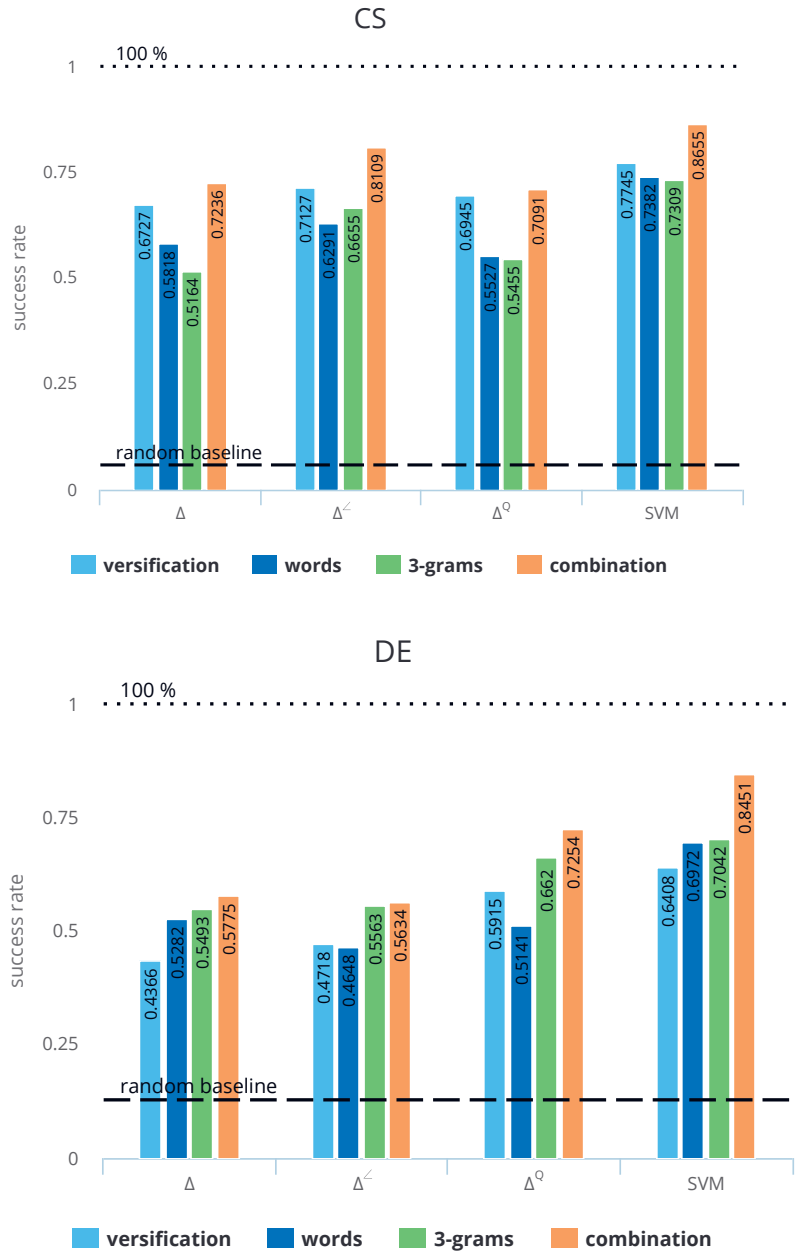
CS
100 %
1
0.75
0.5
0.25
0
success rate
random baseline
0.6727
0.5818
0.5164
0.7236
0.7127
0.6291
0.6655
0.8109
0.6945
0.5527
0.5455
0.7091
0.7745
0.7382
0.7309
0.8655
Δ
Δ'
Δ^Q
SVM
versification
words
3-grams
combination
DE
100 %
1
0.75
0.5
0.25
0
success rate
random baseline
0.4366
0.5282
0.5493
0.5775
0.4718
0.4648
0.5563
0.5634
0.5915
0.5141
0.662
0.7254
0.6408
0.6972
0.7042
0.8451
Δ
Δ'
Δ^Q
SVM
versification
words
3-grams
combination

Figure 11.

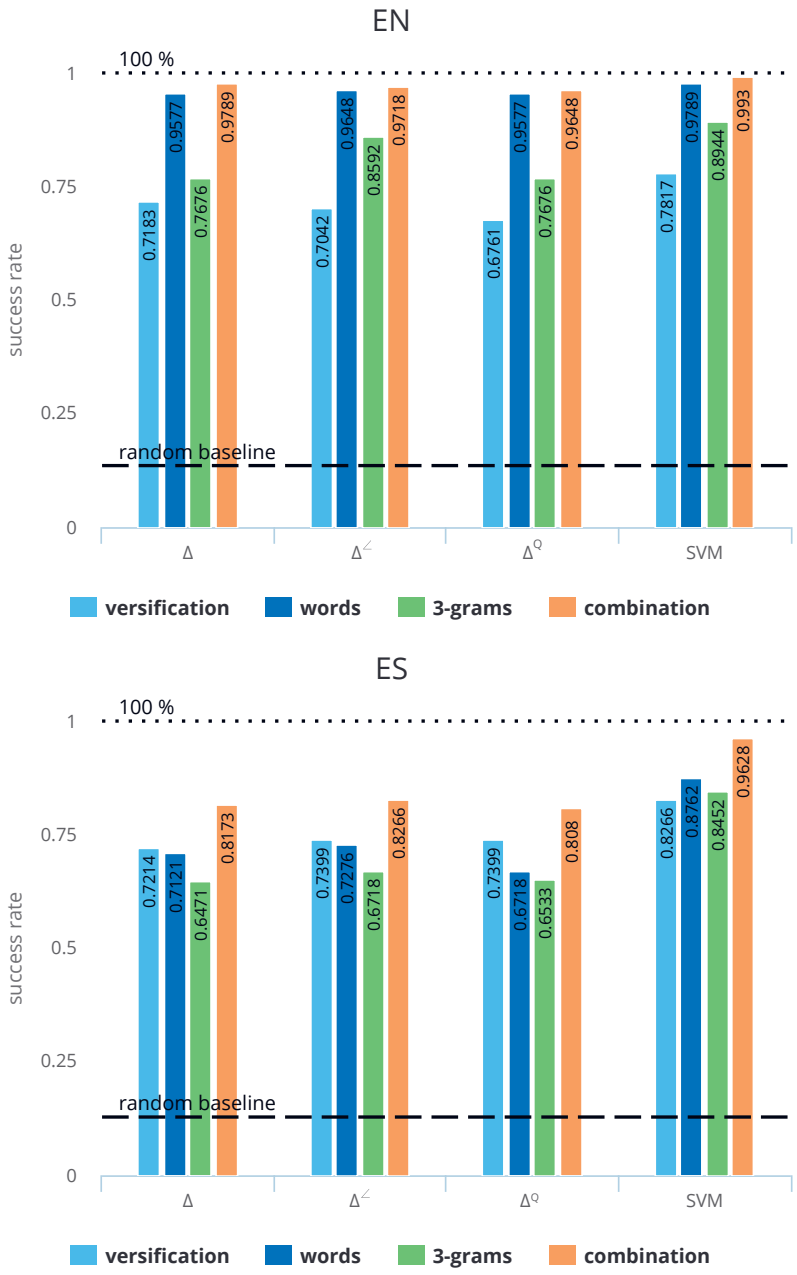

Figure 11. The accuracy of Burrows' Delta (Δ), Cosine Delta (Δ^{Q}), Quadratic Delta (Δ^{A}), and the Support Vector Machine (SVM) with versification features, the 100 most frequent words, the 100 most frequent character trigrams, and the combination of these three.

Conclusions and Future Work

We have shown that versification is a reasonable stylometric indicator, of comparable strength to those that are used in stylometry traditionally. So far we have worked with only limited sets of versification features (the stress profile and sounds' frequencies and rhyme characteristics), which were chosen rather *ad hoc*, as were the poetic meters with which we were working (the Czech and German trochaic tetrameter; the English iambic pentameter). In our future work, we would like to systematically test the methods proposed with other features added (e. g. the frequencies of word boundaries after particular metrical positions), as well with other meters and even other languages. Apart from merely testing the method, we would also like to employ it for real attribution tasks. We do believe that this article has provided evidence that versification can and should be used for such purposes.[9]

References

Al-Falahi, Ahmed; Ramdani, Mohamed; Bellafkih, Mostafa 2017. Machine Learning for Authorship Attribution in Arabic Poetry. In: *International Journal of Future Computer and Communication* 6(2), 42–46.

Argamon, Shlomo 2008. Interpreting Burrows's Delta: geometric and probabilistic foundations. In: *Literary and Linguistic Computing* 23(2), 131–147.

Bobenhausen, Klemens 2011. The Metricalizer: Automated Metrical Markup for German Poetry. In: Küper, Christoph (ed.), *Current Trends in Metrical Analysis*. Frankfurt am Main [etc.]: Peter Lang, 119–131.

Bobenhausen, Klemens; Hammerich, Benjamin 2015. Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme Metricalizer². In: *Langages* 199, 67–87.

Burrows, John F. 2002 "Delta": a measure of stylistic difference and a guide to likely authorship. In: *Literary and Linguistic Computing* 17(3), 267–287.

[9] The authors would like to thank Ryan Heuser (Stanford University) for his kind help with setting up the Prosodic package, and Borja Navarro-Colorado (University of Alicante) for his helpful comments on Spanish verse. The study was supported by the Czech Science Foundation, project GA17-01723S (Stylometric Analysis of Poetic Texts).

Burrows, John F. 2003. Questions of authorship: attribution and beyond. In: *Computers and the Humanities* 37(1), 5–32.

Eder, Maciej 2011. Style markers in authorship attribution. A cross-language study of the authorial fingerprint. In: *Studies in Polish Linguistics* 6, 99–114.

Eder, Maciej 2013. Does size matter? Authorship attribution, short samples, big problem. In: *Digital Scholarship in the Humanities* 30(2), 167–182.

Eder, Maciej 2017. Short samples in authorship attribution: A new approach. In: *Digital Humanities 2017: Conference abstracts*. Montreal: McGill University, 221–224. https://dh2017.adho.org/abstracts/341/341.pdf.

Fleay, Frederick Gard 1874. On the authorship of The Taming of the Shrew. In: *Transactions of the New Shakespeare Society* 1, 85–129.

Fleay, Frederick Gard 1876. *Shakespeare Manual*. London: Macmillan.

Grieve, Jack 2005. *Quantitative Authorship Attribution. A History and an Evaluation of Techniques*. Master thesis. Burnaby, BC: Simon Fraser University.

Grieve, Jack 2007. Quantitative authorship attribution: an evaluation of techniques. In: *Literary and Linguistic Computing* 22(3), 251–270.

Grzybek, Peter 2014. The emergence of stylometry: prolegomena to the history of term and concept. In: Kroó, Katalin; Torop, Peeter (eds.), *Text within Text – Culture within Culture*. Budapest, Tartu: L'Harmattan, 58–75.

Hoover, David L. 2004a. Testing Burrows's Delta. In: *Literary and Linguistic Computing* 19(4), 453–475.

Hoover, David L. 2004b. Delta Prime? In: *Literary and Linguistic Computing* 19(4), 477–495.

Ingram, John Kells. 1874. On the "weak endings" of Shakespeare, with some account of the history of the verse tests in general. In: *Transactions of the New Shakespeare Society* 1, 442–456.

Jannidis, Fotis; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten 2015. Improving Burrows' Delta. An empirical evaluation of text distance measures. In: *Digital Humanities 2015: Conference abstracts*, Sydney: University of Western Sydney. http://dh2015.org/abstracts/

Juola, Patrick 2006. Authorship attribution. In: *Foundations and Trends in Informational Retrieval* 1(3), 233–334.

Koppel, Moshe; Schler, Jonathan; Argamon, Shlomo 2009. Computational methods in authorship attribution. In: *Journal of the Association for Information Science and Technology* 60(1), 9–26.

Lotman, Mihhail 2015. A study on Shakespeare's verse in its historical context (Marina Tarlinskaja, Shakespeare and the Versification of English Drama, 1561–1642, Ashgate, 2014) [Review article]. *Studia Metrica et Poetica* 2(1), 140–153.

Lotman, Yuri; Lotman, Mihhail 1986. Vokrug desjatoj glavy "Evgenija Onegina". In: Petrunina, Nina Nikolaevna (ed.), *Pushkin: Issledovanija i materialy* XII. Leningrad: Nauka, 124–151.

Malone, Edmond 1787. *A dissertation on parts one, two and three of Henry the Sixth tending to shew that those plays were not written originally by Shakespeare*. London: Henry Baldwin.

Mendenhall, Thomas Corwin 1887. The characteristic curves of composition. In: *Science* 9, 237–249.

Mendenhall, Thomas Corwin 1901. A mechanical solution to a literary problem. In: *Popular Science Monthly* 9, 97–110.

Mikros, George K.; Perifanos, Kostas A. 2013. Authorship attribution in Greek tweets using author's multilevel *n*-gram profile. In: *Papers from the 2013 AAAI Spring Symposium*. "*Analyzing Microtext*", *25–27 March 2013*, *Stanford*, *California*. Palo Alto, CA: *AAAI* Press, 17–23.

Mosteller, Frederick; Wallace, David L. 1964. *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.

Navarro-Colorado, Borja 2015. A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects. In: *Computational Linguistics for Literature NAACL 2015*, Denver, CO, 105–113. http://www.aclweb.org/anthology/W/W15/W15-0712.pdf

Navarro-Colorado, Borja; Ribes-Lafoz, María; Sánchez, Noelia 2016. Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Portorož.

Plecháč, Petr 2016. Czech verse processing system KVĚTA: Phonetic and metrical components. In: *Glottotheory* 7, 159–174.

Plecháč, Petr (2018). A collocation-driven method of discovering rhymes (in Czech, English, and French poetry). In: Fidler, Masako; Cvrček, Václav (eds.), T*aming the Corpus. From Inflection and Lexis to Interpretation*. Cham: Springer, 79–95.

Plecháč, Petr; Kolár, Robert 2015. The Corpus of Czech Verse. In: *Studia Metrica et Poetica* 2(1), 107–118.

Reddy, Sravana; Knight, Kevin 2011. Unsupervised discovery of rhyme schemes. In: *Proceedings of the 49th Annual Meeting on the Association for Computational Linguistics*. Portland, OR: ACL, 77–82.

Shapir, Maxim 1997. Fenomen Batenkova i problema mistifikatsii (lingvostikhovedcheskij aspekt 1–2). In: *Philologica* 4, 85–144.

Shapir, Maxim 1998. Fenomen Batenkova i problema mistifikatsii (lingvostikhovedcheskij aspekt 3–5). In: *Philologica* 5, 49–132.

Simpson, Edward H. 1949. Measurement of diversity. In: *Nature* 163, 688.

Smith, Peter W. H.; Aldridge, W. 2011. Improving authorship attribution: Optimizing Burrows' Delta method. In: *Journal of Quantitative Linguistics* 18(1), 63–88.

Spedding, James 1850. Who wrote Shakespeare's *Henry VIII*. In: *The Gentleman's Magazine*, 115–123.

Stamatatos, Efstathios 2009. A Survey of modern authorship attribution methods. In: *Journal of the Association for Information Science and Technology* 60(3), 538–556.

Tarlinskaja, Marina 1987. *Shakespeare's Verse: Iambic Pentameter and the Poet's Idiosyncrasies*. New York: Peter Lang.

Tarlinskaja, Marina 2014. *Shakespeare and the versification of English Drama, 1561–1642*. Farnham [etc.]: Ashgate.

Tomashevsky, Boris Viktorovich 1923/2008. Pjatistopnyj jamb Pushkina. In his: *Izbrannye raboty o stikhe*. Moskva, Sankt-Peterburg: Akademija, 140–242.

Weber, Henry (ed.) 1812. *The Works of Beaumont and Fletcher in Fourteen Volumes*. Vol. 13. Edinburgh: J. Ballantyne & Co.

Williams, Carrington Bonsor 1975. Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. In: *Biometrika* 62(1), 207–212.

Yule, George Udny 1938. On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. In: *Biometrika* 30, 363–390.