Does size matter? Authorship attribution, small samples, big problem

Maciej Eder
Pedagogical University of Kraków, Poland and Polish Academy of Sciences, Institute of Polish Language, Kraków, Poland

Abstract

The aim of this study is to find such a minimal size of text samples for authorship attribution that would provide stable results independent of random noise. A few controlled tests for different sample lengths, languages, and genres are discussed and compared. Depending on the corpus used, the minimal sample length varied from 2,500 words (Latin prose) to 5,000 or so words (in most cases, including English, German, Polish, and Hungarian novels). Another observation is connected with the method of sampling: contrary to common sense, randomly excerpted 'bags of words' turned out to be much more effective than the classical solution, i.e. using original sequences of words ('passages') of desired size. Although the tests have been performed using the Delta method (**Burrows, J.F.** (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87) applied to the most frequent words, some additional experiments have been conducted for support vector machines and *k*-NN applied to most frequent words, character 3-grams, character 4-grams, and parts-of-speech-tag 3-grams. Despite significant differences in overall attributive success rate between particular methods and/or style markers, the minimal amount of textual data needed for reliable authorship attribution turned out to be method-independent.

Correspondence:
Maciej Eder, Institute of Polish Studies, Pedagogical University of Kraków, ul. Podchorążych 2, 30-084 Kraków, Poland.
Email:
maciejeder@gmail.com

1 Introduction

In the field of computational stylistics, and especially in authorship attribution, the reliability of the obtained results becomes even more essential than the results themselves: failed attribution is much better than false attribution (cf. Love, 2002). However, while dozens of outstanding articles deal with increasing the effectiveness of current stylometric methods, the problem of their reliability remains somehow underestimated. Especially, the simple yet fundamental question of the shortest acceptable sample length for reliable attribution has not been discussed convincingly.

It is true that the problem is not new. Its importance is stressed, although not directly, by Rudman in his seminal articles concerning reliability in authorship attribution inference (Rudman 1998a,b, 2003). In his investigation of style variation in Golding's *The Inheritors*, Hoover noticed that truncating all the samples to the size of the shortest chapter spoils the results, probably due to the short sample effect (Hoover, 2003, p. 439). In another instance, Rybicki discovered that his own

doi:10.1093/llc/fqt066 Advance Access published on 14 November 2013

results of remarkable similarities in the patterns of distance between idiolects in two different translations of the same trilogy of novels were due to the gap between talkative and nontalkative characters, the latter simply not saying enough to produce a reliable sample (Rybicki, 2006, 2008).

A few scholars have proposed an intuitive solution to this problem, e.g. that an analyzed text should be 'long' (Craig, 2004, p. 287), that 'for stylometric reliability the minimum sample size allowed is 1,000 words' (Holmes *et al.*, 2001, p. 406), that 'with texts of 1,500 words or more, the Delta procedure is effective enough to serve as a direct guide to likely authorship' (Burrows, 2002, p. 276), etc. Those statements, however, have not been followed by thorough empirical investigations. Additionally, many otherwise successful attribution studies do not obey even the assumed limit of 1,000 words per sample (Juola and Baayen, 2005; Burrows, 2002; Jockers *et al.*, 2008, etc.).

In those attribution studies based on short samples, despite their well-established hypotheses, good choice of style-markers, advanced statistics applied, and convincing results presented, one cannot avoid the simple yet nontrivial question whether those impressive results have not been obtained by chance, or at least have not been positively affected by randomness?

Recently, a few studies concerning different issues in scalability in authorship attribution have been published (Zhao and Zobel, 2005; Hirst and Feiguina, 2007; Stamatatos, 2008; Koppel *et al.*, 2009; Mikros, 2009; Luyckx, 2010; Luyckx and Daelemans, 2011). However, the problem addressed in the present study, that of an experimental estimation of the sample length that would provide a reliable attribution, has not been explored in a comprehensive manner. Also, there have been no cross-language studies addressing this problem, while this would seem an ideal way to validate the obtained results and to generalize patterns of behavior observed in particular case studies.

2 Hypothesis

Word frequencies in a corpus are not random variables in a strict sense; notably, an occurrence of a given word highly depends on its nearest context: the probability of finding, say, 'and' immediately followed by 'and' is extremely low. And yet words do display some characteristics of random variables: since the author is not obliged at any rate to distribute words regularly, particular word frequencies might vary considerably across different works, or even different passages (chapters, stanzas) written by the same person. Thus, similar to other probabilistic phenomena, word frequencies strongly depend on the size of the population (i.e. the size of the text used in the study). Now, if the observed frequency of a single word exhibits too much variation for establishing an index of vocabulary richness resistant to sample length (Tweedie and Baayen, 1998), a multidimensional approach—based on numerous probabilistic word frequencies computed at once—should be even more questionable.

On theoretical grounds, we can intuitively assume that the smallest acceptable sample length would be hundreds rather than dozens of words. Next, we can expect that, in a series of controlled authorship experiments with longer and longer samples tested, the probability of attribution success would at first increase quickly, indicating a strong correlation with the current text size; but then, above a certain value, further increase of input sample size would not significantly affect the effectiveness of the attribution significantly. However, in any attempt to find this critical point on a 'learning curve', one should be aware that its location might depend—to some extent—on the language, genre, or even the particular texts analyzed.

3 Data and Method of Testing

In any approach to the problem of scalability in authorship attribution, an appropriate choice of test data is of paramount importance. One possible solution is to perform a contrastive analysis of naturally long versus naturally short texts (e.g. novels versus short stories, essays versus blog posts, etc.), to estimate the possible correlation between sample length and attribution accuracy. The most apparent weakness of this kind of approach, however, is that the results might be biased by inherent cross-genre

differences between the two groups of texts. To remedy this limitation, in the present study, the same data set was used for all the comparisons: the goal was to extract shorter and longer virtual samples from the original corpus, using intensive resampling in a large number of iterations. The advantage of such a gradual increase of excerpted virtual samples is that it covers a wide range between 'very short' and 'very long' texts, and further enables one to capture a break point of the minimal sample size for a reliable attribution.

Several corpora of known authorship were prepared for different languages and genres (used separately): for English, Polish, German, and Hungarian novels, for Latin and Ancient Greek prose (nonfiction), and for English, Latin, and Ancient Greek epic poetry. Within a particular genre (novels, nonfiction, poetry), the collected corpora were roughly similar in size. As a matter of fact, keeping rigidly the same number of texts in each corpus seemed to be unrealistic; additionally, it should be stressed that any cross-corpus (or cross-language) comparison is never fully objective, even if the collected data sets are identical in size. The corpora used in the present study were as follows:

- Sixty-three English novels by 17 authors,
- Sixty-six German novels by 21 authors,
- Sixty-nine Polish novels by 13 authors,
- Sixty-four Hungarian novels by 9 authors,
- Ninety-four Latin prose texts by 20 authors,
- Seventy-two Ancient Greek prose texts by 8 authors,
- Thirty-two English epic poems by 6 authors,
- Thirty-two Latin epic poems by 6 authors,
- Thirty Ancient Greek epic poems by 8 authors.

The texts were gathered from a variety of public domain sources, including Perseus Project, The Latin Library, Bibliotheca Augustana, Project Gutenberg, Literature.org, Ebooks@Adelaide, and the author's private collection.[1] The acquired texts were edited in order to normalize for spelling errors (if applicable), to exclude footnotes, disclaimers, nonauthorial prefaces, and so forth.

For each corpus, three discrete controlled attribution experiments aimed to examine three different methods of sampling (discussed below in detail) were performed. To assess the textual data, a few attribution techniques have been applied. As the main methodological basis for all the experiments, however, the widely accepted Delta method (Burrows, 2002) was chosen, with the assumption that the results should be valid, by extension, for other distance-based methods as well. In all the tests, 200 most frequent words (MFWs) were analyzed. For the computation tasks, including text preprocessing, sampling, and classification, a tailored script for the open-source statistical environment R was used.

The reason for choosing Delta based on MFWs was that it combines high accuracy of supervised methods of classification with simplicity of multidimensional techniques using distance measures of similarity. Such a solution seems to be an acceptable compromise between two main approaches to stylometry: literary-oriented studies on stylistic similarities between texts (authors, genres, styles, and so forth) on the one hand, and information technology studies on authorship attribution 'in the wild' on the other. While the former approach usually involves explanatory distance-based techniques such as multidimensional scaling or cluster analysis, and is aimed to capture stylometric relationships between literary texts, the latter considers attribution as a particular case of a classification problem (where precision is of paramount importance), and usually relies on sophisticated machine-learning methods of supervised classification. In an attempt to find a balance between those two discrete stylometric worlds, Burrows's Delta seemed to be the best choice.

Obviously enough, Delta exhibits some methodological pitfalls, including the lack of validation of the obtained results, and the tacit assumption of variables' independence (Argamon, 2008). Also, Delta is sometimes claimed to be suboptimal in comparison with other classification algorithms: among computer scientists, there seems to be a consensus that support vector machines (SVM) using character *n*-grams is presently the single best, language-independent approach in the field of authorship attribution (Koppel *et al.*, 2009; Stamatatos, 2009). On the other hand, however, Delta was

Probleme von Delta & SVM als bester Ansatz für AA

found to perform almost equally well when compared with other classification techniques, including SVM (Jockers and Witten, 2010), and the claims about the robustness of character *n*-grams turned out to be unfounded for some languages (Eder, 2011).

The above arguments summarized, one has to admit that relying on one attribution method alone might lead to biased and/or unreliable results. Thus, apart from Delta based on MFWs, a number of additional tests have been conducted using other classifiers (SVM, *k*-NN) and other style markers [character 3-grams, character 4-grams, parts-of-speech (POS)-tags 3-grams].

The benchmarks were based on the standard procedure used in machine-learning methods of classification; namely, all texts from a given corpus were divided into two groups. The 'training' set consisted of one representative text per author, while the remaining samples were included into the 'test' set. Next, samples of a given length were extracted from the original texts, and each sample from the 'test' set was tested against the 'training' set to verify the quality of classification. The percentage of correctly 'guessed' samples, i.e. those linked to their actual authors, was regarded as a measure of attribution accuracy.

The procedure as described above, strictly following the original implementation of Delta (Burrows, 2002), provides an approximation to many real-case attribution studies that have been published in recent decades. Although this approach should reveal precisely the correlation between sample size and attribution performance, it is not reliable enough to test the attribution performance itself. This is because the 'training' texts might not have been 'representative' (whatever that means) for their authors, or there might have appeared other irregularities in stylometric profiles. One of the ways to avoid a possible bias is to swap randomly some texts between the 'training' and the 'test' set, and to replicate the classification stage. This solution is referred to as cross-validation and it will be addressed in the final sections of this article. However, since the point of gravity in this study was set toward literary studies (digital philology) rather than 'raw' authorship attribution, advanced topics of validation as well as rigorous statistical terminology were simplified as much as possible.

4 Experiment I: Bags of Words

The tests discussed in this article were aimed to excerpt, in many approaches, shorter and longer subsets from original literary texts. It can be argued, however, that in literary works, and particularly in novels, vocabulary is distributed slightly differently in narrative and dialogue parts (Burrows, 1987, pp. 163–75; Hoover, 2001, p. 426). Thus, to neutralize any possible impact of these natural inconsistencies in word distributions, in the first experiment a very intuitive way of sampling was chosen, which is sometimes referred to as 'bags of words': the goal was to pick the words randomly, one by one, from subsequent texts. This type of sampling provides a reasonable approximation to the original vocabulary distribution in text, but one has to remember that, at the same time, it deletes the original syntactic and lexical contexts of particular words.

The research procedure was as follows. For each text in a given corpus, 500 randomly chosen single words were concatenated into a new sample. These new samples were analyzed using the classical Delta method; the percentage of successful attributions was regarded as a measure of effectiveness of the current sample length. The same steps of excerpting new samples from the original texts, followed by the stage of 'guessing' the correct authors, were repeated for the length of 500, 600, 700, 800, . . . , 20,000 words per sample.

The results for the corpus of 63 English novels are shown in Fig. 1. The observed scores (black circles on the graph; gray circles are discussed below) clearly reveal a trend (solid line): the curve, climbing up very quickly, tends to stabilize at a certain point, which indicates the minimal sample size for the best attribution success rate. Although it is difficult to find the precise position of that point, it becomes quite obvious that samples shorter than 5,000 words provide a poor 'guessing' because they can be immensely affected by random noise. Below the size of 3,000 words, the obtained results are simply disastrous (more than 60% of false attributions for

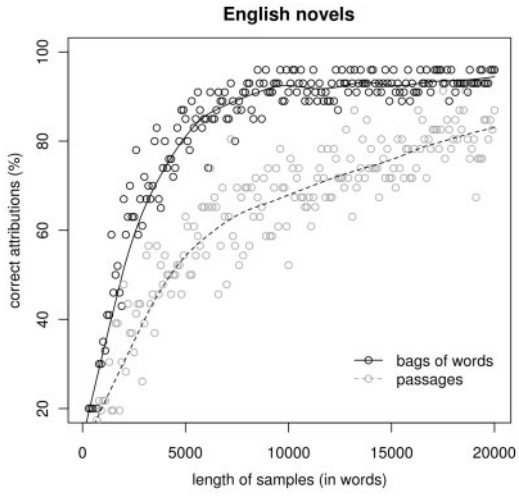

Fig. 1 Dependence of attribution accuracy and length of text samples: 63 English novels (200 MFWs tested). Black circles indicate the 'bags of words' type of sampling, gray circles indicate excerpted passages

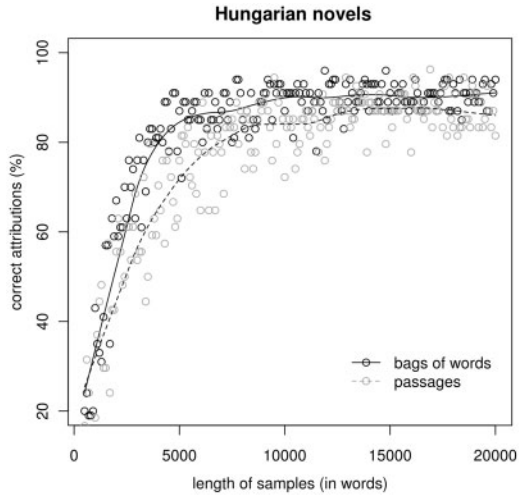

Fig. 2 Dependence of attribution accuracy and length of text samples: 64 Hungarian novels

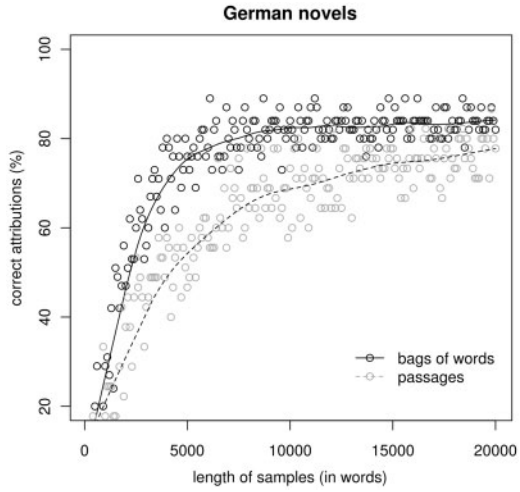

Fig. 3 Dependence of attribution accuracy and length of text samples: 66 German novels

1,000-word samples may serve as a convincing caveat). Other analyzed corpora showed quite similar shapes of the 'learning curves', although some interesting differences could also be noticed.

In particular, the overall attribution success rate varied considerably: among the corpora with modern prose, Hungarian (Fig. 2) gained lower scores than English, followed by German (Fig. 3), and Polish (Fig. 4). This phenomenon has already been observed in previous cross-language studies (Rybicki and Eder, 2011; Eder, 2011; Eder and Rybicki, 2013). The accuracy rates of both Ancient prose corpora (Figs 5 and 6) were fairly satisfying, Latin being slightly less attributable than Greek. Similar differences were observed in corpora with poetry: 'guessing' scores for English epic poems (Fig. 7) were substantially higher than for Greek poetry (Fig. 8), and very similar to Latin poetry (not shown).

The critical point of attributive saturation could be found at about 5,000 words per sample in most of the corpora analyzed in this study (and no significant differences between inflected and noninflected languages were observed). However, there were also some exceptions to this general rule. First of all, the corpus of Latin prose exhibited a significant improvement in resistance to short samples (its minimal effective sample size was of approximately 2,500 words; cf. Fig. 5, black circles). At the same time, the Latin corpus showed a very clear and distinctive trend of increasing attribution accuracy followed by fairly horizontal scores of statistical saturation. In the other corpora, especially in Polish novels, the 'learning curves' gained their saturation somewhat slowly and less distinctively.

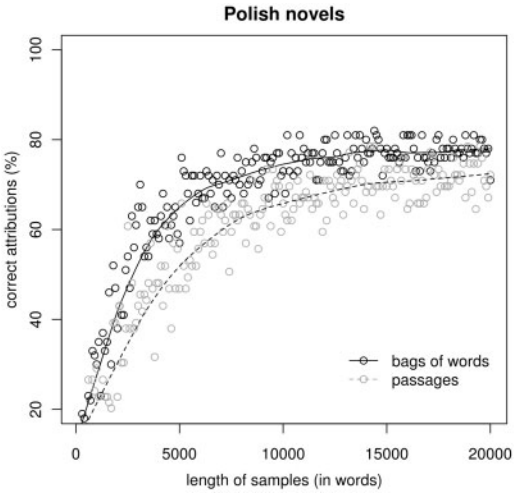

Fig. 4 Dependence of attribution accuracy and length of text samples: 69 Polish novels

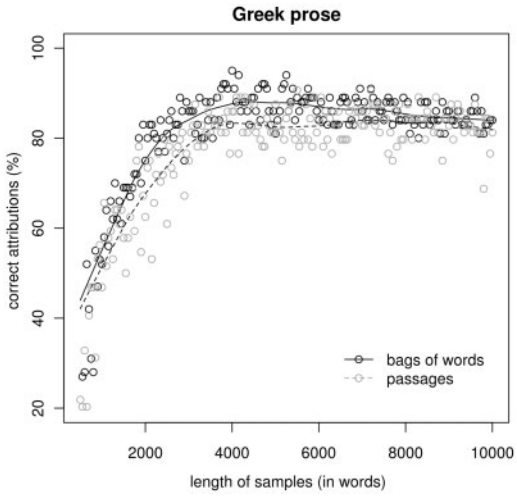

Fig. 6 Dependence of attribution accuracy and length of text samples: 72 Ancient Greek prose texts

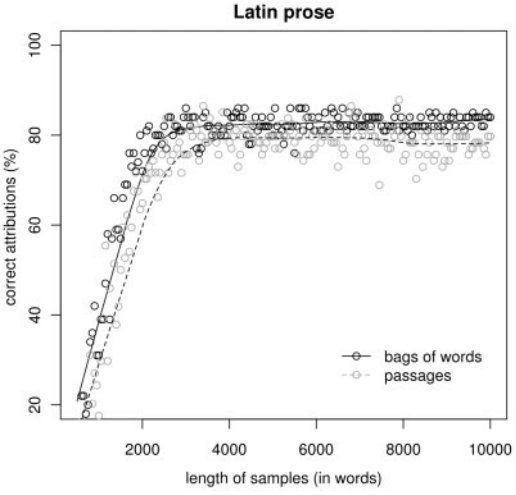

Fig. 5 Dependence of attribution accuracy and length of text samples: 94 Latin prose texts

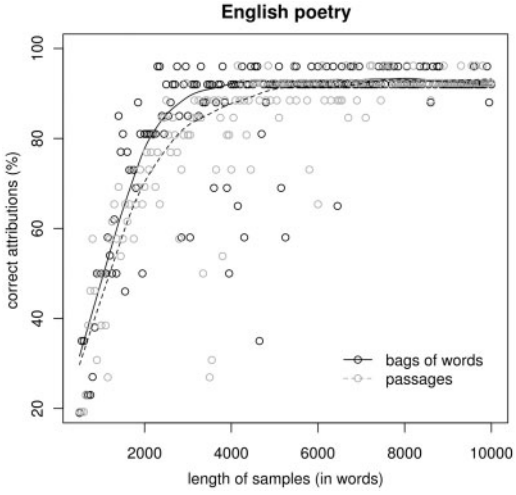

Fig. 7 Dependence of attribution accuracy and length of text samples: 32 English epic poems

The behavior of the poetic corpora of English and Latin should also be commented upon. At first glance, English epic poetry required a promising amount of 3,000 words per sample for reliable attribution (Fig. 7). However, the number of unpredictable and very poor scores scattered randomly on the plot (cf. the outliers much below the trend line) suggests that attributing English poems shorter than 5,000 words might involve a risk of severe misclassification. The picture of Latin epic poetry (not shown) was surprisingly similar.

Another peculiarity was observed in the case of two Greek corpora with prose and poetry, respectively (Figs 6 and 8). In both cases, after the usual rapid performance improvement due to increasing length of text samples, the attribution accuracy was

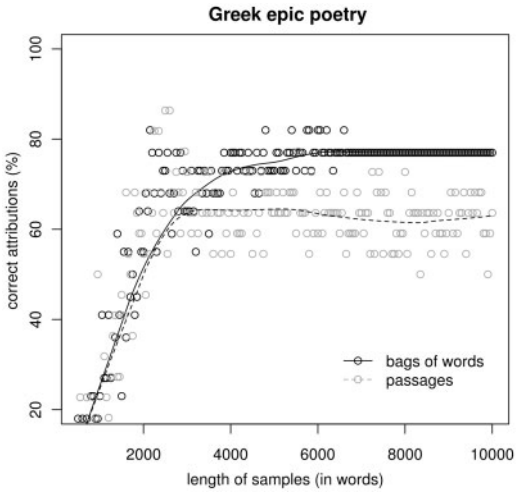

Fig. 8 Dependence of attribution accuracy and length of text samples: 30 Ancient Greek epic poems

found to decrease (!) for some time, and then to stabilize. It is difficult to explain this phenomenon; however, a similar behavior of Greek texts has been observed in a recent study (Eder, 2011, pp. 103–5), where a systematic error in corpus slightly improved the attribution accuracy rate.

As regard performance stabilization above a certain sample length, the 'guessing' scores for each corpus analyzed also seemed to show that effectiveness did not increase in samples exceeding 15,000 words. It represents another valuable observation, suggesting that there are limits to Hoover's statement that 'for statistical analysis, the longer the text the better' (Hoover, 2001).

5 Experiment II: Passages

The results of the first experiment were quite disappointing, to say the least. They might easily lead to the suspicion that the 'bags of words' type of sampling was a decisive factor here, since this way of combining samples breaks the original sequences of words with all their possible syntactic relations. A variant of the above experiment was prepared, then, to test the possible impact of sampling on attribution performance.

The way of preparing the samples by extracting a mass of single words from the original texts seems to be an obvious solution to the problem of statistical representativeness. In most attribution studies, however, shorter or longer passages, or blocks, of disputed works are usually analyzed, either randomly chosen from the entire text or simply truncated to the desired size (Hoover, 2003, p. 349; Reynolds *et al.*, 2012; etc.). The purpose of the second experiment was to test the performance of this typical sampling as opposed to extracted 'bags of words'. The whole procedure was repeated step by step as in the previous test, but now, yet instead of collecting individual words, sequences of 500 running words (then 600, 700, 800, . . . , 20,000) were chosen randomly from the original texts.

The excerpted virtual samples were analyzed using the same classification method, the same number of iterations, and the same number of frequent words as in the first experiment, but the results were found to be significantly different. The differences become particularly evident when the final scores of the two experiments are represented on shared graphs (Figs 1–8). The gray circles on the graphs and the dashed trend lines show the effectiveness of the 'passage' type of sampling, as opposed to the black circles followed by the solid trend lines of the 'bags of words'. Despite minor discrepancies, three main observations could be made at this stage that seem to be applicable to all the corpora examined:

(1) For each corpus analyzed, the attribution accuracy obtained in the second experiment (excerpted passages) was always worse than the scores described in the first experiment, relying on the 'bags of words' type of sampling. This is counterintuitive, and it means that the procedure of excerpting a mass of single words as described in the first experiment was not responsible for the considerably poor results. Quite the opposite, this way of sampling turned out to be a better solution. The results cannot be simply explained away by the fact that artificially excerpted samples ('bags of words') are no longer tied to a specific local context in texts. It is a well-known fact that topic strongly intervenes with

authorial style (Stamatatos, 2009; Luyckx and Daelemans, 2011), and the same can be said of narrative and dialogic parts of novels (Burrows, 1987; Hoover, 2001). The observed phenomenon, however, was noticeable irrespective of the assessed topics or genres.

(2) For 'passages', the dispersion of the observed scores was always wider than for 'bags of words', indicating a bigger impact of random noise for this kind of sampling. Certainly, as above, the effect might be due to the obvious differences in word distribution between narrative and dialogue parts in novels (Figs 1–4); however, the same effect was similarly strong for nonliterary prose in Latin and Greek (Figs 5 and 6) and in English poetry (Fig. 7), or even substantially stronger in Greek poetry (Fig. 8).

(3) The distance between both trend lines—for 'words' and for 'passages'—varied noticeably across different corpora. At the same time, however, there was also some regularity in this variance, quite evident in the corpora of novels (Figs 1–4) and probably related to the degree of inflection of the languages analyzed. More specifically, the more inflected the language, the smaller the difference in successful attribution between both types of sampling: the greatest in the English novels (Fig. 1 gray circles versus black), the smallest in the Hungarian corpus (Fig. 2). Both prose corpora of Latin and Greek—two highly inflected languages—also fitted the model (Figs 5 and 6). The only exception to it was the corpora of English and Greek poetry, where the results were ambiguous (Figs 7 and 8).

6 Experiment III: Chunks

At times we encounter an attribution problem where extant works by a disputed author are clearly too short to be analyzed in separate samples. The question is, then, if a concatenated collection of short poems, epigrams, sonnets, book reviews, notes, etc. in one sample (cf. Eder and Rybicki, 2009; Dixon and Mannion, 2012) would reach the effectiveness comparable with that presented above. Also, if concatenated samples are suitable for attribution tests, do we need to worry about the size of the original texts constituting the joint sample?

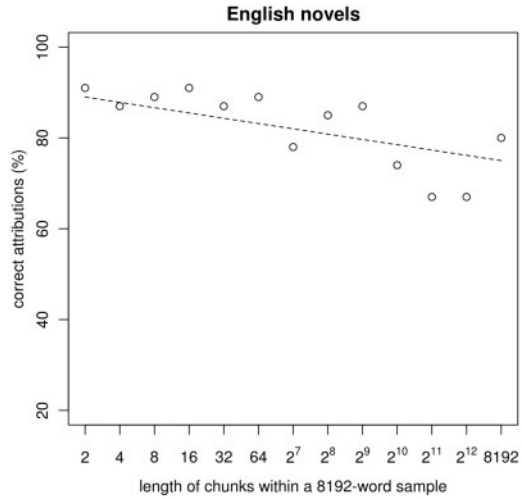

Fig. 9 Dependence of attribution accuracy and length of chunks within 8192-word samples: 63 English novels

To examine the usefulness of concatenated samples, an experiment slightly different from the previous two was prepared. Provided that 8,000 words or so seem to be sufficient to perform a reliable attribution (see above), in the present approach the size of 8,192 words was chosen to combine samples from shorter chunks. In 12 iterations, a number of word chunks, or *n*-grams, were randomly selected from each text and concatenated: 4,096 chunks of 2 words in length (2-grams), 2,048 word 4-grams, 1,024 chunks of 8 words, 512 chunks of 16 words, and so on, up to 2 chunks of 4,096 words. Thus, all the samples in question were 8,192 words long.

The obtained results were roughly similar for all the languages and genres analyzed, and somehow correlated with the results of the two previous experiments. As shown in Fig. 9 (for the corpus of English novels), the effectiveness of 'guessing' depends to some extent on the word-chunk size used. The attributive accuracy rates were worse for long chunks within a sample (4,096 words or so) than for 2-grams or 4-grams. This decrease in performance was linear, namely, the shorter a chunk,

the better the 'guessing' scores. Interestingly, the difference between the effectiveness of the shortest and the longest chunks followed the difference between 'bags of words' and 'passages' in the first two experiments (Fig. 1). Certainly, this is easy to explain, since single words are the extreme case of short chunks, and 'passages' are in fact very long chunks. The results of this experiment seem to fill the gap between the two trend lines for 'words' and 'passages' presented above. This observation applies to all the corpora tested.

The results seem to be fairly optimistic because there is no substantial difference in attribution accuracy between, say, a few chunks of 500 words combined in one sample and a dozen concatenated chunks of 100 words. It suggests that in real attribution studies, concatenated samples would display a very good performance.

One has to remember, however, that the above simulation of concatenated short texts was artificial. The chunks were excerpted randomly from long texts, regardless of sentence delimitation, etc. What is even more important is short literary forms, like epigrams or aphorisms, have their own stylistic flavor, substantially different from typical literary works (Mautner, 1976). Short literary forms are often masterpieces of concise language, with a domination of verbs over adjectives, particles, and so on, with a proverbial witty style, and with a strong tendency to compression of content. Thus, a collection of aphorisms will certainly have different word frequencies than a long essay written by the same author; similarly, a collection of short mails will differ from an extensive epistle, even if they are addressed to the same addressee. For that reason, further investigation is needed here.

7 Evaluation

This section aims to discuss the evaluation procedures used in the experiments presented above as well as to introduce a number of cross-check tests for other machine-learning techniques and alternative style markers. It should be stressed, however, that discussing all the tests that have been conducted in this section at the evaluation stage (dozens of tests involving some 2 million iterations and almost 100 million single 'guesses') is simply unrealistic. Thus, the corpus of English novels and the experiment with 'bags of words' will be used in this section as a case study.

Among the drawbacks of the Delta method, a particularly serious one is that the choice of texts to be included in the 'training' set is arbitrary or depends on subjective decisions which works by a given author are 'representative' of his/her stylometric profile. Even if this choice is fully automatic (e.g. when the samples for the 'training' set are chosen randomly by the machine), it is still very sensitive to local authorial idiosyncrasies. In consequence, the estimated classification accuracy might overfit the actual behavior of input data.

Advanced machine-learning methods of classification are routinely aimed at solving the problem of model overfitting due to possible inconsistencies in the selection of training samples. The general idea behind such 'cross-validation' tests is to replicate the original experiment multiple times with random changes to the composition of both the 'training' and 'test' sets: in a number of random swaps between the samples, followed by the stage of classification, one obtains an average behavior of the corpus. A commonly accepted solution, introduced to stylometry from exact sciences, is a 10-fold cross-validation (Zhao and Zobel, 2005; Juola and Baayen, 2005; Stamatatos, 2008; Koppel *et al.*, 2009; Jockers and Witten, 2010; Luyckx and Daelemans, 2011, and so forth). It has been shown, however, that mere 10 swaps might be far too little to betray potential model overfitting: a set of cross-language attribution tests with a large number of random recompositions of the 'training' and 'test' sets have shown substantial inconsistencies for some of the analyzed corpora (Eder and Rybicki, 2013).

To avoid homeopathic solutions, then, and to perform a robust cross-validation, the evaluation procedure applied in this study was as follows. For each type of sampling assessed, for each corpus, and for each sample size, the texts included into the 'training' set were chosen randomly (one text per author) in 100 independent iterations. In consequence, the classification test was applied 100 times, and the average attribution success rate was

recorded. The whole procedure was repeated for every single sample size tested: 500, 600, 700, . . . , 20,000 words (using 'bags of words' in the first experiment, 'passages' in the second one). This approach could be compared with an extreme version of a 100-fold cross-validation (extreme, because in each iteration the whole 'training' set was thoroughly recomposed). Certainly, the whole task was computationally very intensive.

The cross-validated attribution accuracy scores for the English corpus are shown in Fig. 10 (averaged performance rates represented by gray circles, a trend—by black solid line). The overall attributive rate is indeed lower than for the nonvalidated variant (Fig. 1), but model overfitting is not substantial.[2] Much more important for the scope of this study, however, is the shape of the 'learning curve' and the point where the curve becomes saturated—and they are almost identical (Fig. 1 versus Fig. 10). Thus, the conclusions concerning the minimal sample length seem to be validated, at least for Delta.

To test the possible impact of different classification algorithms on the sample size effect, another series of check tests, followed by 100-fold cross-validation, were performed. The methods tested were SVM (claimed to be the most accurate attribution method so far) and *k*-NN. The comparison of their performance is shown again in Fig. 10, which shows that SVM indeed outperforms other methods, Delta being an undisputed runner-up (Fig. 10, dashed line versus solid line). Significantly low performance of *k*-NN (Fig. 10, dotted line) can be explained by the fact that the 'training' set contained one sample per author only—and these settings are *a priori* suboptimal for *k*-NN. Despite the differences between particular classifiers, the shapes of the 'learning curves' remain stable. This shows that the short sample problem cannot be easily bypassed by switching to sophisticated machine-learning algorithms.

Another series of check tests was conducted to examine the performance of alternative style-markers as compared with short samples. Character *n*-grams seem to be a particularly promising proposition here. They are claimed to be robust and language-independent (Koppel *et al.*, 2009; Stamatatos, 2009), and resistant to untidily prepared corpora, e.g. containing a mass of misspelled characters

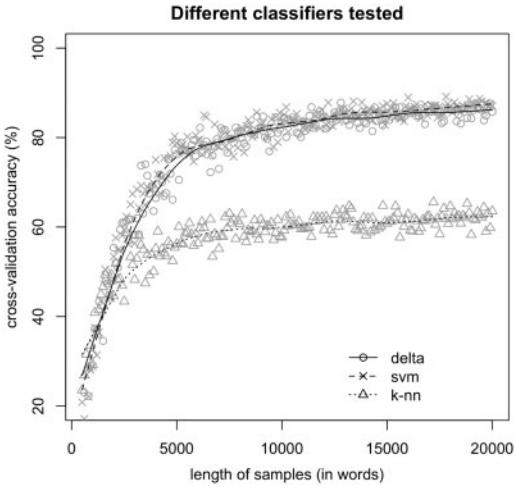

Fig. 10 Cross-validation scores for Delta, SVM, and k-NN with 200 MFWs analyzed: 63 English novels

(Eder, 2013). Besides character *n*-grams, potentially powerful style-markers also include syntax-based features, such as automatically recognized POS-tags,[3] analyzed separately or combined into *n*-grams (Hirst and Feiguina, 2007).

The results of 100-fold cross-validation for 12 possible combinations of four different style-markers (MFWs, character 3-grams, character 4-grams, POS-tag 3-grams) and three classifiers (Delta, SVM, *k*-NN) are presented in Fig. 11. No matter which classifier was used, MFWs proved to be the most accurate solution, followed by POS-tag 3-grams, and character-based markers. The attribution success rates for SVM based on MFWs clearly suggest that this particular combination provides the best performance in the corpus of English novels; Delta combined with MFWs was found to exhibit similar performance. The shapes of the 'learning curves' do not indicate any increase of resistance to short samples when alternative style-markers are used.

8 Discussion

8.1 Methodological remarks

The experiments described above were designed to be as reliable as possible; however, some

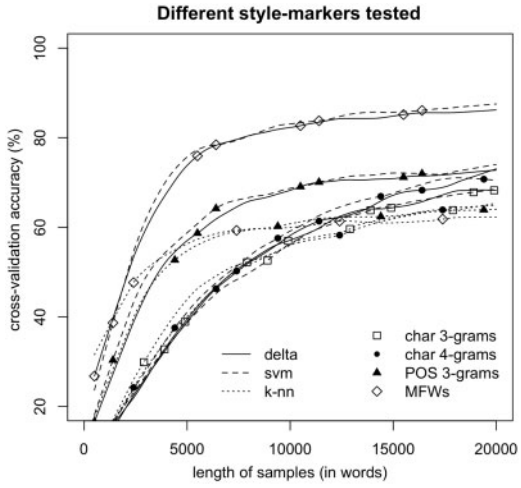

Fig. 11 Cross-validation scores for Delta, SVM, and k-NN combined with MFWs, POS-tag 3-grams, character 3-grams, and character 4-grams: 63 English novels (trend lines only, particular observations omitted for clarity)

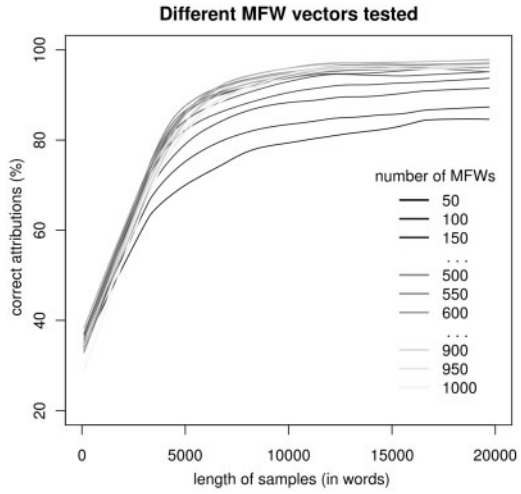

Fig. 12 Dependence of attribution accuracy, length of samples, and number of MFWs analyzed: 63 English novels

arbitrary choices were inescapable. These are as follows:

(1) Choice of style-markers. In nontraditional authorship attribution, many different features of style have been tested: indexes of vocabulary richness, measures of rhythmicity, frequencies of function words, frequencies of parts-of-speech tags, and so forth (an extensive list of possible style-markers is provided by Koppel *et al.*, 2009, pp. 11–13). In the present study, a few possible types of markers were tested, with preference given to the classical solution, i.e. vectors of frequencies of the MFWs. It is possible that in some languages, alternative style-markers might exhibit better performance.

(2) Number of style-markers to be analyzed. It is true that the effectiveness of nearest neighbor classifications, including Delta, is very sensitive to the number of features analyzed (Jockers and Witten, 2010). Unfortunately, as has been shown (Rybicki and Eder, 2011), there is no universal vector of MFWs suitable for different languages or genres. In the present study, 200 MFWs (200 character 3-grams, etc.) were used for each test; this arbitrary choice was a compromise between the small number of function words most effective for Latin, and the very long vectors of 1,000 or more words optimal for the corpus of English novels. To test the possible impact of this chosen number of 200 MFWs, an additional experiment was conducted using different MFW vectors. As shown in Fig. 12 (English novels), the overall attribution success rate indeed depends on the vector of MFWs analyzed, but—importantly—the shape of the 'learning curves' and the point of statistical saturation remain stable regardless of different MFWs settings.

(3) Number of texts tested and choice of 'training' and 'test' set members. It has been revealed that the effectiveness of attribution depends on corpus size and particularly on the number of authors tested (Luyckx, 2010; Luyckx and Dealemans, 2011). In short, a 2-class authorship attribution case needs less textual data than a 100-class case. Despite the significance of this problem, it was not addressed in the present study. Since in particular corpora the number of authorial classes was strictly constant for all the tests performed, the obtained results were not affected

by this factor (it is true, however, that any cross-language conclusions are limited, since the number of authorial classes was not fixed across the corpora).

(4) Choice of particular texts included in each corpus. One of the common beliefs demonstrated in technically oriented authorship studies is the slightly naive assumption that keeping the number of authorial classes constant and/or using strictly the same amount of training data guarantees the reliability of the experiment. Unfortunately, collecting textual data is probably the most unreliable stage of any corpus study; the uncontrollable factors are countless here. Unequal availability (representativeness) of texts in electronic version, stylistic differentiation in particular national literatures, possible existence of writers-outliers displaying either exceptional stylistic imagery or extreme dullness in style, inherent linguistic differences between national corpora, and so on—all these factors stand in the way of drawing any definite conclusions. These corpus-related problems might have had a significant impact on the results presented in the present study, with little hope to neutralize what is an inherent feature of textual data. This was also the reason why the corpora were not strictly of the same size: the cost of acquiring such a collection of allegedly 'comparable' corpora would be exorbitant, and the results would be still questionable, at least to some extent.

(5) Choice of a technique of attribution. The scores presented in this study, as obtained with classical Delta procedure, turned out to be slightly different when solved with other nearest neighbor classification techniques (Fig. 11). However, similarly to different vectors of MFWs (Fig. 12), the shapes of all the curves and the points where the attributive success rates remain stable are identical for each of these methods. The same refers to different combinations of style-marker settings—although different settings provide different 'guessing', they hardly affect the shape of the curves. Thus, since the obtained results are method-independent, this leads us to the conclusion about the smallest acceptable sample size for future attribution experiments and other investigations in the field of stylometry.

A few words should be added about explanatory multidimensional methods that are traditionally used in authorship attribution: principal components analysis (PCA), multidimensional scaling (MDS), and cluster analysis. In these methods, there is no direct way of examining the short sample effect and its impact on attribution. However, a very simple test might be performed to illustrate the significance of this problem. Using a method of excerpting 'bags of words' as introduced above, one can perform a series of, say, principal components analyses and compare the obtained plots. The results of two PCA tests of 1,000 randomly chosen words from the same corpus of 28 English novels are shown in Figs 13 and 14. In both cases, 100 MFWs were used. Without deciding which of the two pictures is more likely to be 'true', the substantial differences between them are quite obvious. In the first picture, one can distinguish discrete clusters for Eliot and Austen, a group of texts in the bottom part of the plot (including Fielding and Sterne), and a common cloud of the remaining samples. In the second picture, besides the central cloud, three distinguishable clusters are noticeable: for Charlotte Brontë, for Thackeray, and for Richardson/Fielding. The results of this fairly simple experiment show how misleading an explanatory interpretation of points scattered on a plot might be. What is worse, there are a number of real-life attribution cases based on samples of about 1,000 words (the *Federalists Papers* and *Scriptores historiae Augustae*, to name but a few); addressing them with PCA or MDS might involve a risk of being substantially mistaken.

8.2 Sample size

A characteristic paradox of nontraditional authorship attribution is that the most accurate state-of-the-art methods need very long samples to show their full potential, while in real life, there are not

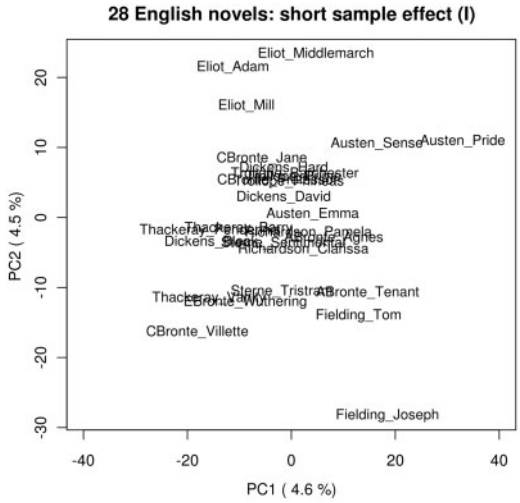

Fig. 13 Principal components analysis of 28 English novels: 1,000 randomly excerpted words from each novel, 100 MFWs analyzed

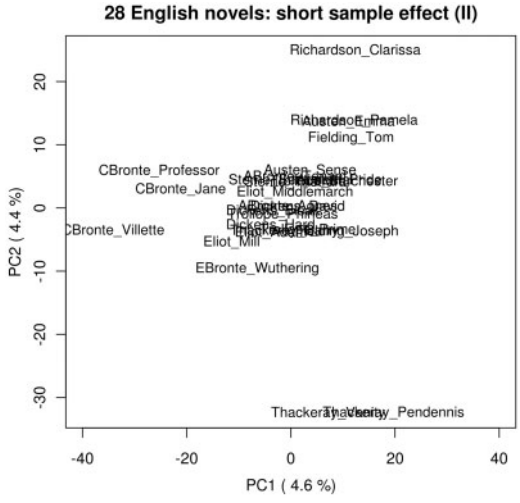

Fig. 14 Principal components analysis of 28 English novels: 1,000 randomly excerpted words from each novel, 100 MFWs analyzed

so many anonymous novels, as opposed to countless anonymous ballads, limericks, sonnets, or critical notes. Thus, paradoxically, the more helpful stylometry could be to supplement traditional literary criticism, the more unhelpful it seems to be in many real-life cases. For that reason, the results obtained in the present study—a few thousand words per sample, at least, to perform an attribution test—will not satisfy most literary scholars.

Certainly, this leads to the question of how to interpret the obvious discrepancy between these unsatisfactory results and several classic attribution studies where short samples have been successfully used. An extreme example is provided by Burrows, who observed that a poem of only 209 words by Aphra Behn was correctly attributed (Burrows, 2002, p. 278). The study by Koppel *et al.* (2009) goes even further, showing that a corpus of very short blog posts (of 217–745 words in length) displays an accuracy of 38.2–63.2%, depending on the classification method used.[4] These results are very impressive; they show how much authorial information can be retrieved from just a couple of paragraphs of running text. On the other hand, one should also remember that the remaining 61.8–36.8% of samples used in this study were misclassified, a crucial pitfall in real-life attribution cases. The commonly known thing is that natural language processing tools and techniques, such as parts-of-speech taggers or syntax parsers, easily achieve a few dozen percent of accuracy, but the actual problem is to gain every next following percent (and the difficulty increases exponentially rather than proportionally). The same can be said about the accuracy of stylometric methods.

A detailed inspection of final rankings of candidates obtained in the above experiments provides some further insights into the problems of attribution. Namely, for all the corpora, no matter which method of sampling is used, the rankings are considerably stable: if a given text is correctly recognized using an excerpt of 2,000 words, it will also be rightly 'guessed' in most of the remaining iterations; if a text is successfully assigned using 4,000 words, it will usually be correctly attributed above this word limit, and so on. At the point of statistical saturation, where increasing the length of sample does not improve the performance, only a few remaining texts are finally linked to their actual authors. For example, in the case of English novels, a fingerprint of both Charlotte and Anne Brontës was considerably well recognizable even for short samples, followed by Richardson, Dickens (*David*

Copperfield, *Pickwick Papers*), Eliot, Galsworthy, Austen, and again Dickens (*Hard Times*, *Oliver Twist*, *Great Expectations*). It was very hard to attribute Hardy, but particularly long samples were needed to distinguish novels written by James.

In other words, in some texts, the authorial fingerprint is quite easily traceable, even if very short samples are used. This was the case of the Latin prose corpus (Fig. 5), where particularly strong style-markers made it possible to distinguish authors using only 2,500 excerpted words per sample. However, there are also some other texts where the authorial signal is hidden, or covered by noise, and it needs to be carefully extracted using very long samples (as was the case of many texts in the Polish corpus, cf. Fig. 4). The only problem is, however, that, in real authorship attribution, we have no *a priori* knowledge of the category of an anonymous text. Thus, it seems that a minimal sample size for reliable attribution does not begin at the point where the first correct 'guesses' appear, but where the most problematic samples finally recognize their own authors.

9 Conclusions

The main research question approached in this study was how much data is sufficient to recognize authorial uniqueness. The results revealed no conclusive answer to this question, though. It seems that for corpora of modern novels, irrespective of the language tested, the minimal sample size is some 5,000 words (tokens). Latin prose required only 2,500 words, and Ancient Greek prose just a little more to display their optimal performance. The results for the three poetic corpora (Greek, Latin and English) proved to be ambiguous, suggesting that some 3,000 words or so would be usually enough, but significant misclassification would also occur occasionally. Thus, the results were found to depend on genre rather than on language examined.

Another conclusion is connected with the method of sampling. Contrary to common sense, randomly excerpted 'bags of words' turned out to be much more effective than the classical solution, i.e. using original sequences of words ('passages') of a desired size. This means that dealing with a text of 20,000 words in length, it is better to select 10,000 words randomly than to rely on the original string of 20,000 words. Again, it is better to excerpt 10 samples of 10,000 randomly chosen words from a whole novel rather than to rely on 10 subsequent chapters as samples. Certainly, the obtained results are partially dependent on the language tested (the level of inflection being one of the suspected factors), genre, literary epoch, number of assessed texts, number of authors, and probably also on particular selection of texts included in a corpus. Nonetheless, the results provide a convincing argument in favor of using randomly excerpted 'bags of words' rather than relying on arbitrarily chosen 'passages' (blocks of words).

This also means that some of the recent attribution studies should be at least reconsidered. Until we develop more precise style-markers than word frequencies, we should be aware of some limits in our current approaches, the most troublesome of these being the limits of sample length. As I have tried to show, using 1,000-word samples will hardly provide a reliable result, let alone examining shorter texts.

The results of the present study do not contradict the soundness of undertaking difficult attribution cases. Quite to the contrary, they attempt to show that stylometry has not said its last word, and there is an urgent need to develop more reliable techniques of attribution. Some promising methods include using part-of-speech tags instead of word frequencies (Hirst and Feiguina, 2007), sophisticated techniques of estimating the level of uncertainty of word counts (Hinneburg *et al.*, 2007), and, last but not least, the method of using recall/precision rates as described in the above-cited study by Koppel *et al.* (2009). Attributing short samples is difficult, but arguably possible, provided that it is approached with the awareness of the risk of misclassification.

Acknowledgements

I would like to thank Jan Rybicki, Georgios Mikros, Łukasz Grabowski, and the anonymous reviewers of this article for their helpful comments. Last but not least, special thanks go to the Alliance of Digital

Humanities Organizations for awarding this study the Paul Fortier Prize for the best young scholar paper at the 2010 Digital Humanities conference at King's College London.

References

Argamon, S. (2008). Interpreting Burrows's Delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, **23**(2): 131–47.

Burrows, J.F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Burrows, J.F. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Craig, H. (2004). Stylistic analysis and authorship studies. In Schreibman, S., Siemens, R., and Unsworth, J. (eds), *A Companion to Digital Humanities*. Oxford: Blackwell Publishing, pp. 273–88.

Dixon, P. and Mannion, D. (2012). Goldsmith's contribution to the 'Critical Review': a supplement. *Literary and Linguistic Computing*, **27**(4): 373–94.

Eder, M. (2011). Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, **6**: 99–114, http://www.wuj.pl/UserFiles/File/SPL%206/6-SPL-Vol-6.pdf (accessed 30 October 2013).

Eder, M. (2013). Mind your corpus: systematic errors and authorship attribution. *Literary and Linguistic Computing*, **28**(4): 603–14.

Eder, M. and Rybicki, J. (2009). PCA, Delta, JGAAP and Polish poetry of the 16th and the 17th centuries: who wrote the dirty stuff?. *Digital Humanities 2009: Conference Abstracts*. College Park, MA, pp. 242–4.

Eder, M. and Rybicki, J. (2013). Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing*, **28**(2): 229–36.

Hinneburg, A., Mannila, H., Kaislaniemi, S., Nevalainen, T., and Raumolin-Brunberg, H. (2007). How to handle small samples: bootstrap and Bayesian methods in the analysis of linguistic change. *Literary and Linguistic Computing*, **22**(2): 137–50.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, **22**(4): 405–17.

Holmes, D., Gordon, L.J., and Wilson, C.H. (2001). A widow and her soldier: stylometry and the American Civil War. *Literary and Linguistic Computing*, **16**(4): 403–20.

Hoover, D.L. (2001). Statistical stylistic and authorship attribution: an empirical investigation. *Literary and Linguistic Computing*, **16**(4): 421–44.

Hoover, D.L. (2003). Multivariate analysis and the study of style variation. *Literary and Linguistic Computing*, **18**(4): 341–60.

Jockers, M.L. and Witten, D.M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, **25**(2): 215–23.

Jockers, M.L., Witten, D.M., and Criddle, C.S. (2008). Reassessing authorship of the 'Book of Mormon' using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing*, **23**(4): 465–91.

Juola, P. and Baayen, R.H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, **20**: 59–67.

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

Love, H. (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.

Luyckx, K. (2010). *Scalability Issues in Authorship Attribution*. Dissertation. University of Antwerpen.

Luyckx, K. and Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, **26**(1): 35–55.

Mikros, G.K. (2009). Content words in authorship attribution: an evaluation of stylometric features in a literary corpus. In Köhler, R. (ed.), *Studies in Quantitative Linguistics 5*. Lüdenscheid: RAM, pp. 61–75.

Mautner, F.H. (1976). Maxim(e)s, sentences, Fragmente, Aphorismen. In Neumann, G. (ed.), *Der Aphorismus. Zur Geschichte zu den Formen und Möglichkeiten einer literarischen Gattung*. Darmstadt: Wissenschaftliche Buchgesellschaft, pp. 399–412.

Reynolds, N.B., Schaalje, G.B., and Hilton, J.L. (2012). Who wrote Bacon? Assessing the respective roles of Francis Bacon and his secretaries in the production of his English works. *Literary and Linguistic Computing*, **27**(4): 409–25.

Rudman, J. (1998a). Non-traditional authorship attribution studies in the 'Historia Augusta': Some caveats. *Literary and Linguistic Computing*, **13**(3): 151–57.

Rudman, J. (1998b). The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities*, **31**: 351–65.

Rudman, J. (2003). Cherry picking in nontraditional authorship attribution studies. *Chance*, **16**(2): 26–32.

Rybicki, J. (2006). Burrowing into translation: character idiolects in Henryk Sienkiewicz's 'Trilogy' and its two English translations. *Literary and Linguistic Computing*, **21**(1): 91–103.

Rybicki, J. (2008). Does size matter? A re-examination of a time-proven method. *Digital Humanities 2008: Book of Abstracts*. Oulu: University of Oulu, p. 184.

Rybicki, J. and Eder, M. (2011). Deeper delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, **26**(3): 315–21.

Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Information Processing and Management*, **44**(2): 790–99.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society of Information Science and Technology*, **60**(3): 538–56.

Tweedie, J.F. and Baayen, R.H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, **32**: 323–52.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. In Lee, G.G., Yamada, A., Meng, H., and Myaeng, S.-H. (eds), *Asia Information Retrieval Symposium 2005* (=Lecture Notes in Computer Science, vol. 3689). Berlin: Springer, pp. 174–89.

Notes

1 Many texts used in this study (especially those representing Polish literature) were extracted from corpora prepared for other projects by members of the Computational Stylistics Group (https://sites.google.com/site/computationalstylistics/) and from a variety of student MA projects conducted at the Pedagogical University of Kraków, Poland, and at the Jagiellonian University, Kraków, Poland. They can be made available by contacting the Computational Stylistics Group.

2 Since the classical Delta procedure counts *z*-scores based on the 'training' set alone, and then applies the variables' means and standard deviations to the 'test' set, the permutation of both sets somewhat impacts the *z*-scores and thus, possibly, the final results as well. In view of this, a parallel series of tests with *z*-scores calculated for both sets has been performed. The differences between those two approaches were noticeable yet not significant.

3 For a number of reasons, ranging from availability of well-trained taggers for particular languages (or particular historic periods), to substantially different grammars between the languages addressed in the present study, the check tests with POS-tag *n*-grams were limited to the corpus of English novels and the corpus of Latin poetry only. The software used for tagging were the Stanford NLP Tools (for English) and TreeTagger (for Latin).

4 The scores in question were obtained using 512 function words as style-markers. More sophisticated features, such as parts-of-speech tags or 1,000 character trigrams with 'highest information gain in the training corpus', displayed the accuracy of up to 86% for some classifiers (Koppel *et al.* 2009, p. 14).