Mind your corpus: systematic errors in authorship attribution

Maciej Eder

Abstract

In computational stylistics, any influence of unwanted noise – e.g. caused by an untidily-prepared corpus – might lead to biased or false results. Relying on contaminated data is quite similar to using dirty test tubes in a laboratory: it inescapably means falling into systematic error. An important question is what degree of nonchalance is acceptable to obtain sufficiently reliable results. The present study attempts to verify the impact of unwanted noise in a series of experiments conducted on several corpora of English, German, Polish, Ancient Greek and Latin prose texts. In 100 iterations, a given corpus was gradually damaged, and controlled tests for authorship were applied. The first experiment was designed to show the correlation between a dirty corpus and attribution accuracy. The second was aimed to test how disorder in word frequencies – produced by scribal and/or editorial modifications – affects the attribution abilities of particular corpora. The goal of the third experiment was to test how much 'authorial' data a given text needs to have in order to trace authorial fingerprint through a mass of external quotations.

1 Introduction

Non-traditional authorship attribution relies on advanced statistical procedures to distil significant markers of authorial style from a large pool of stylistic features that are not distinctive enough to provide reliable information about authorial uniqueness. The same general rule applies to other issues in computational stylistics – genre recognition, sentiment analysis, translation studies, assessing stylistic changes during an author's lifespan, etc. In all these and similar attempts to find distinctive stylistic features, the goal is to find as much order in 'randomness' as possible. The better the method applied, the more regularities can be extracted from a population that seems to contain nothing but noise. However, it does not mean that one can overcome the impact of randomness: noise is an inherent feature of all natural languages. In particular, word frequencies in a corpus behave as if they were randomly distributed. Although subjected to the rules of the language, words appear irregularly in texts, since the author has a freedom of choice between, say, different synonyms. Thus, word counts will always contain some randomness.

Although dealing with this unavoidable noise is *crème de la crème* of computational stylistics, any other influence of additional factors, either random or systematic – e.g. caused by an untidily-prepared corpus – might lead to biased or false results. Relying on contaminated data is quite similar to using dirty test tubes in a laboratory: it inescapably means falling into systematic error. Certainly, quite an important question is what degree of nonchalance is acceptable to obtain sufficiently reliable results. The importance of this issue should be emphasized especially in the age of easily acquirable 'web-scraped' corpora.

The problem of systematic errors in stylometry, and particularly in non-traditional authorship attribution, has already been discussed. Rudman (1998a, 1998b, 2003) has formulated a number of caveats concerning different issues in non-traditional authorship attribution, including possible pitfalls in corpus preparation. Noecker et al. (2008), in their attempt to test the impact of optical character recognition (OCR) errors on attribution accuracy, have observed that moderate damage of input texts does not affect the results significantly. Similarly, Eder (2011) has noticed that a faultily prepared corpus of Greek epic poems displayed an unexpectedly good performance. In other studies, a strong correlation between the amount of input data and attribution performance has been shown (Eder, 2010; Luyckx and Daelemans, 2011; Koppel at al., 2011). In these and many other studies, however, the problem of systematic errors has not been addressed systematically.

The present study attempts to test the extent to which different types of noise affect authorship attribution effectiveness, and to validate the obtained results using a number of corpora written in different languages. It should be stressed, however, that this is a pilot study rather than an exhaustive examination of every possible type of systematic error that can occur in a corpus. To keep things simple, one particular method has been used: Delta as developed by Burrows (2002), one type of style-markers: frequencies of the most frequent words (MFW), one corpus for each language, one set of texts selected to serve as training samples (i.e. no swapping between training and test sets), one type of damage added to the corpora (artificially generated systematic noise). Arguably, the obtained results are valid and reliable – but limited to the particular cases as discussed below. The author's ambition was, however, to propose a general framework of an empirical approach to the problem of systematic errors that could be thoroughly examined and extended in future studies.

2 Outline of the experiment

The nature of noise affecting the results is quite complex. On the one hand, a machine-readable text might be contaminated by poor OCR, mismatched codepages, improperly removed XML tags; by including non-authorial textual additions, such as prefaces, footnotes, commentaries, disclaimers, etc. On the other hand, there are some types of unwanted noise that can by no means be referred to as systematic errors; they include scribal textual variants (*variae lectiones*), omissions (*lacunae*), interpolations, hidden plagiarism, editorial decisions for uniform spelling, modernizing the punctuation, and so on. Both types of noise, however, share a very characteristic feature. Namely, the longer the distance between the time when a given text was written and the moment of its digitization, the more opportunities of potential error to occur, for different reasons.

Discussing possible factors that affect OCR accuracy, Holley (2009) argues that most OCR software brands claim 99% accuracy rates, but this applies either to clean, good quality images, or when manual intervention in text post-processing takes place. These accuracy rates are thus not applicable to historical documents: usually, the older the document in question, the lower the accuracy, also because old spelling variants make the character recognition more difficult (Robertson and Willett, 1993). Crane (2009) points out that commercial OCR systems are designed to use with the present standard English language. On his own words: 'While running English prose will be well served, books published before the mid-nineteenth century will produce much noisier output. Book printed in historical languages (e.g. classical Greek) would be essentially unsearchable. [...] The added noise from OCR can range from marginal (e.g., an early twentieth-century cleanly printed edition of Dickens) to catastrophic (e.g., a Greek source text)'. Another factor is a significant development in OCR software technology between 1993 and 2005 (Holley, 2009), making recently digitized texts much cleaner than documents OCR-ed some time ago.

The second type of noise is even more correlated with the time span between text creation and digitization. Before the era of print, the transmission of texts was based on copying by scribes. It is commonly known, for instance, that ancient classical texts were written on papyri (scrolls), and they needed to be rewritten roughly once in a century in order to avoid physical damage. In these copies of copies, textual variants were countless, some fragments were lost, some other were omitted in copying on purpose, e.g. due to obscene content (Reynolds and Wilson, 1978), and numerous interpolations were included. Another source of noise produced by scribes are innumerable spelling variants. Arguably, the longer the chain of copied copies in text transmission, the noisier the copied text, because of the aggregation effect. It is also worth remembering that sentence delimitation and punctuation marks were all introduced by modern scholars. Being the last element in text transmission, a critical edition is still subjected to editors' individual decisions and text variation (ultimately, the editors are the last generation of copyists). An illustrative example of editorial variance is provided by Wake (1957: 335–337), who shows interesting differences in sentence length distribution between two critical editions of Aristotle's *Categories* (Oxford Classical Text vs. Loeb Classical Library). The above general features of Greek and Latin classical texts apply also to medieval and early modern period, and to vernacular literatures as well.

Luckily enough, stylometric investigations are usually based on texts written in the same period, thus even if the impact of the above noise is substantial, it affects all the texts in question to a similar degree.

To verify the impact of different types of unwanted noise, a series of experiments has been conducted on several corpora of English, German, Polish, Ancient Greek and Latin prose texts, the corpora being roughly similar in text length and number of authors tested. They were as follows:

- sixty-three English novels (18th and 19th cent.) by seventeen authors;
- sixty-six German novels (19th cent.) by twenty-one authors;
- ninety-two Polish novels (19th and 20th cent.) by thirteen authors;
- ninety-four Latin prose texts (1st cent BC to 2nd cent. AD) by 20 authors;
- seventy-two Ancient Greek speeches and plays (6th to 4th cent. BC) by 8 authors.

The texts have been 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] In 100 iterations, a given corpus was gradually damaged, and controlled tests for authorship were applied (the procedure has been inspired by Noecker et al. 2008). It can be obviously assumed that heavy damage would spoil the results substantially. The aim of this study, however, is to test whether this decrease of performance is linear. On theoretical grounds, one can expect either a linear regression (the more errors, the worse the results), or some initial resistance to small errors, followed by a steep drop of performance.

In all the experiments, the Delta method (Burrows, 2002), and frequencies of the most frequent words (MFWs) have been used. There are several reasons of this decision. Firstly, it is a very intuitive procedure that performs considerably well in comparison with much more mathematically advanced techniques (Jockers and Witten, 2010). Secondly, authorship attribution studies using machine-learning techniques of classification, such as support vector machines (SVM) or nearest shrunken centroids (NSC), occupy only one mansion in the house of stylometry. Besides, there is a variety of popular explanatory methods used to assess stylistic differentiation between genres, authors, translators, and so on. They include cluster analysis, multidimensional scaling, principal components analysis and many other distance-based techniques, which are usually applied to frequencies of the most frequent words. Choosing Delta and MSWs in this study seemed to be a reasonable compromise, providing a reference point in the aforementioned machine-learning approaches to authorship on the one hand, and a good caveat in literary-oriented explanatory studies on the other.

The results obtained for Delta should be – by extension – valid for other methods that rely on multidimensional comparison of frequencies of MFWs.

[1] Many texts used in this study (especially those in Polish literature) come from corpora prepared for other projects by members of the Computational Stylistic 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 Stylistic Group.

On theoretical grounds, sophisticated machine-learning methods (e.g. SVM) and some robust style-markers (e.g. character n-grams or POS-tags[2]) might display significantly better accuracy than Delta applied to MFWs; their decrease of performance due to gradual damage in a corpus can also differ as compared to observed Delta behavior. Although verifying this assumption was not the principal aim of this study, some additional tests have been conducted using SVM and character n-grams as features. Needless to say that the number of all possible combinations of a few machine-learning techniques with a variety of possible style-markers is rather difficult to handle in one single study. Thus, further investigations are required here.

Like other multidimensional methods, Delta is very sensitive to the choice of number of features to be analyzed (Jockers and Witten, 2010). For that reason, each experiment has been approached in a series of 30 independent tests for attribution, increasing the number of MFWs analyzed: 100, 200, 300, and so on, all the way to 3,000. In each of these independent attribution trials, the percentage of correctly classified texts from the 'test' subcorpus to their actual authors represented in the 'training' subcorpus, was regarded as a measure of accuracy. The same procedure was applied to assess three different types of noise, as described below in detail. Since three experiments have been conducted, each of them performed 30 times in 100 iterations over five language corpora, the overall number of single attribution tests was as high as 45,000. Another 24,000 pilot tests have been performed for Delta applied to character 2-grams, Delta to character 3-grams, SVM to character 2-grams, SVM to character 3-grams, and SVM to MFWs.

For the sake of simplicity, the procedure of testing the attribution accuracy was not followed by a cross-validation step, in terms of 10 random swaps between training and test sets. It has been shown (Eder and Rybicki, 2013) that a standard 10-fold validation is by far not enough to estimate the degree of uncertainty in a real attribution case; it is more reliable to assess the problem with a very large number of random permutations of both sets. The present study is basically not devoted to the problem of authorship attribution itself, but to the question how any additional noise affects the results of attribution, hence evaluation of the best attributive scores obtained is not as important as capturing the decrease of performance in correlation to added noise. A straightforward validation of the results, however, is provided by repeating all the experiments 30 times with different MFW settings and by comparing the attribution performance across 5 language corpora.

There is yet one more important remark to be made. Namely, in any approach to the problem of measuring the impact of noise in a corpus (i.e. a miscellaneous collection of texts) one can deal with at least three general types of systematic errors. First, when all the samples are damaged to a similar degree (e.g. if the corpus has not been cleaned of markup tags, which are now counted as words of the actual text); second, when, in a carefully collected corpus, some of the samples (one or more) are of poor quality. Third, when texts included in a corpus are affected by different types of independent damage. This situation might take place when texts are harvested from different sources, and when poor OCR, inconsistent spelling, editorial normalization etc. aggregate in one collection. This type of a multi-layer systematic error is particularly difficult to identify and eliminate. It is also not straightforward to analyze its impact in controlled experiments. For this reason, the simulations presented below will focus on the first case only, i.e. one type of error spread throughout the whole corpus.

3 Misspelled characters

The first experiment addresses a very trivial yet severe type of damage – the situation where single

[2] Certainly, in any attempt to simulate attribution performance, using POS-tags extracted from dirty corpus, the core problem is the quality of POS-tagging rather than the final stage of data classification. Since this is the question of robustness of tagging algorithms, and not the question of attribution itself, it should be addressed in another study. The additional problem in a cross-language study is that while POS taggers do a fairly good job in English, their reliance is uneven for other languages (for reasons ranging from inflexion to finances) thus adding yet another variable to this already-complex equation.

characters are misspelled due to transmission disturbance, imperfect typing, poor quality of scanned documents, the use of untrained OCR software, etc. This general type of damage might be referred to as the 'dirty corpus' case.

There are many factors affecting the OCR accuracy; an extensive list is provided by Holley (2009), and some have already been discussed above. In an empirical approach to the impact of misspelled characters on authorship recognition, however, their possible sources are far less important than their distribution in texts to be analyzed. Ideally, developing an algorithm of artificial gradual damage in corpus should be preceded by a careful examination which characters are more likely to be misspelled in actual textual data. The most frequent confusions will certainly include the following pairs: *m / in*, *l / 1*, *h / li*, *c / e*, etc. A spectacular example of confusing *f* and the descending *s* (i.e. the letter ſ) can be found in the corpus provided by the Google Ngrams Viewer (http://books.google.com/ngrams): *beft* (instead of *best*), *fon* (instead of *son*), and numerous similar cases. Other misrecognized character pairs are also represented in this corpus, as evidenced by forms like *wc*, *liave*, *hly*, *aii*, *fony*, etc. (instead of *we*, *have*, *lily*, *all*, *sorry*, resp.).[3]

However, an attempt to measure the impact of real noise should also take into consideration that, in actual textual data, some misrecognized pairs of characters would occur more frequently than others. Next, confusions of particular characters would be distributed differently in different languages. Reynaert (2011) provides a list of the 20 most frequent character confusions in two Dutch corpora: since the most OCR-vulnerable characters from this list seem to be language-independent, their occurrences would vary from language to language. Next, the scripts using more diacritics (or, generally, more complicated glyphs) would be more likely to be misspelled; differences between English (pure Latin alphabet), German (some diacritics), and Ancient Greek (numerous accents, breathing marks, etc.) would be substantial. To conclude: a cross-language experiment in authorship attribution based on real-life noisy data seems unfeasible at this point. The problem requires further investigation and experimental approach to particular language corpora.

In the present study, a simple algorithm of random letter replacements was used in the belief that it would provide a good approximation to a real dirty corpus. Namely, it might be assumed that a particular spelling error is spread randomly in a string of characters; thus, multiple OCR confusions should be spread randomly as well. Next, some characters (letters) are more likely to be damaged than others. Especially, white spaces are rarely misspelled. An artificial damaging procedure might rely, then, on random replacement of non-space characters. A straightforward yet slightly naive solution might be as follows: damaging every 100th letter in the first iteration, every 99th letter in the second iteration, and so on. The easiest way to spoil a given letter would be replacing it with, say, 'x'. To give an example (the 20th iteration, with 20% damaged letters): 'Mrs Lo**x**g say**x** that **x**ethe**x**fiel**x** is ta**x**en by **x** youn**x** man o**x** larg**x** fort**x**ne'.

However, the procedure applied in a following experiment was somewhat different. In each of the 100 iterations, an increasing percentage of randomly chosen letters (excluding spaces) were replaced with other randomly chosen letters. In fact, the decision which letter would be spoiled was made using probabilistic rules: e.g. in the 20th iteration, every letter of the input text was intended to be damaged with a probability of 20%; in consequence, the corpus contained roughly 20% of independently damaged letters: 'Mrs Lon**r** sa**a**s **k**h**a**t **t**etherfi**i**ld is tak**s**n by a y**s**ung man of l**s**r**c**e fo**o**tune'.

The results were quite similar for most of the languages tested. As shown in Fig. 1 and 5–8 (color versions of the figures can be found in the online version of this article), short vectors of MFWs (up to 500 words) usually provide no significant decrease of performance despite a considerably large amount of noise added (the corpus of Polish novels being an exception). Even 20% of damaged letters would not affect the results in some cases! However, longer MFW vectors are *very sensitive* to misspelled characters: any additional noise means a steep decrease

[3] Recently, a new version of the Google Books corpus has appeared; OCR accuracy seems to have been substantially improved. However, the original dataset is still available, under the names 'English (2009)', 'American English (2009)', 'British English (2009)' etc.

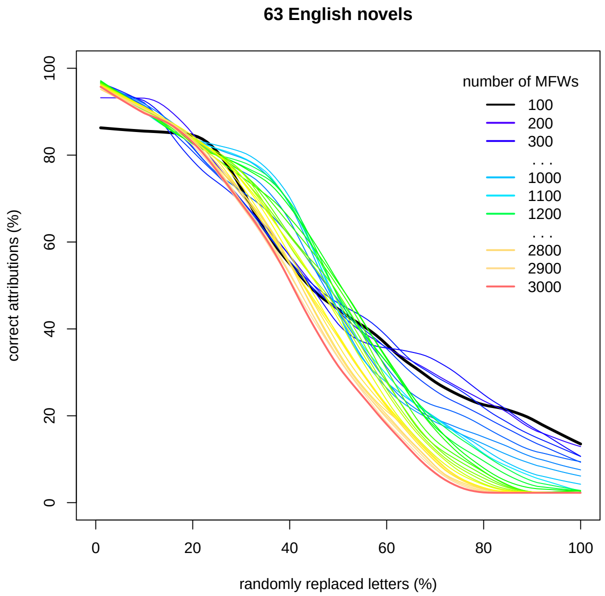

Figure 1: Simulation of poor OCR quality in the corpus of English novels: in 100 iterations, increasing percentage of intentionally misspelled characters has been tested for 30 different MFW vectors. Color coding is indicated by the legend.

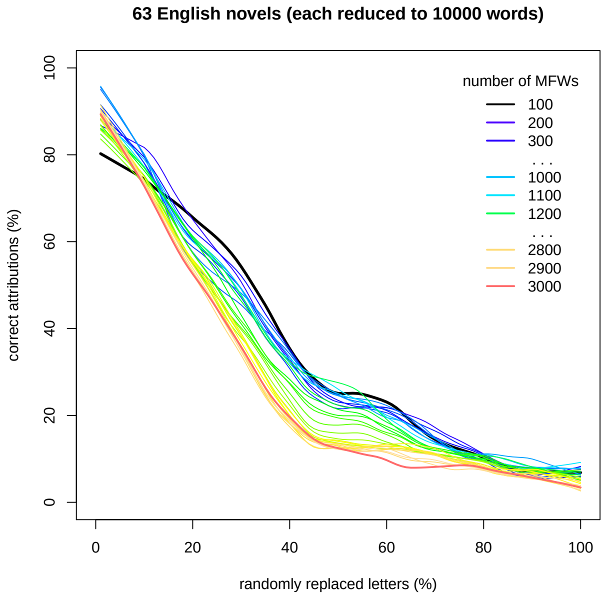

Figure 2: Simulation of poor OCR quality in the corpus of English novels: instead of entire texts, samples of 10,000 randomly excerpted words were tested.

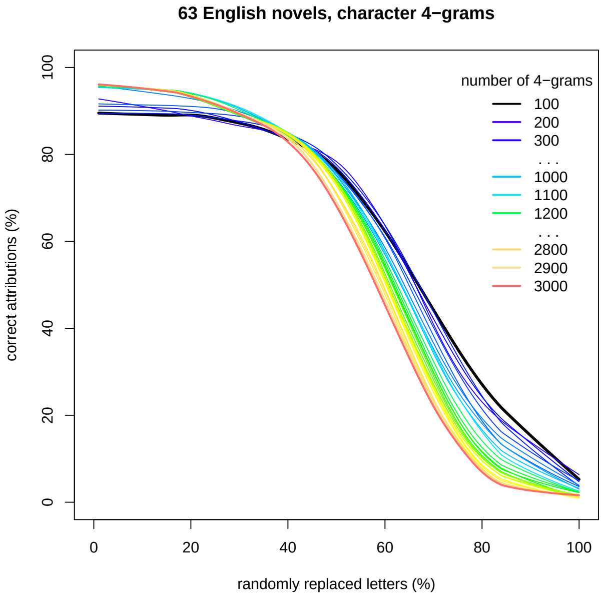

Figure 3: Simulation of poor OCR quality in the corpus of English novels: character 4-grams used as features, instead of MFWs.

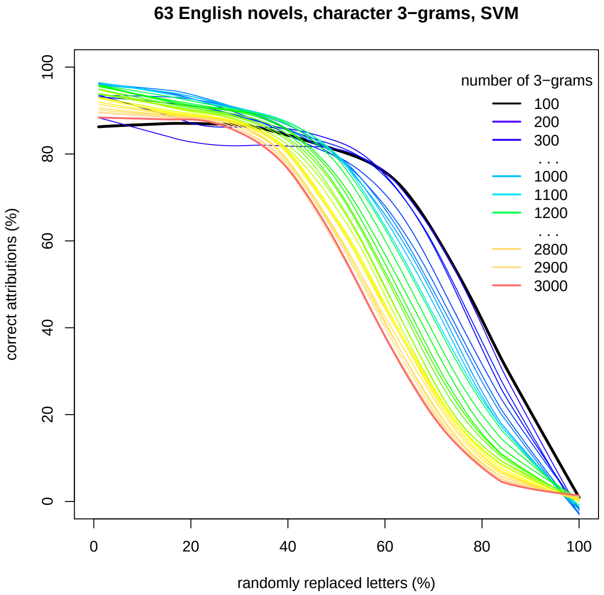

Figure 4: Simulation of poor OCR quality in the corpus of English novels: character 3-grams used as features, support vector machines (SVM) as a classification algorithm.

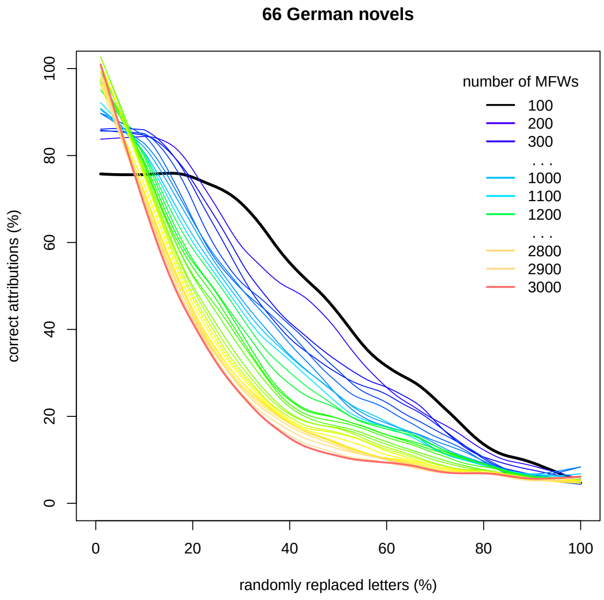

Figure 5: Simulation of poor OCR quality in the corpus of German novels.

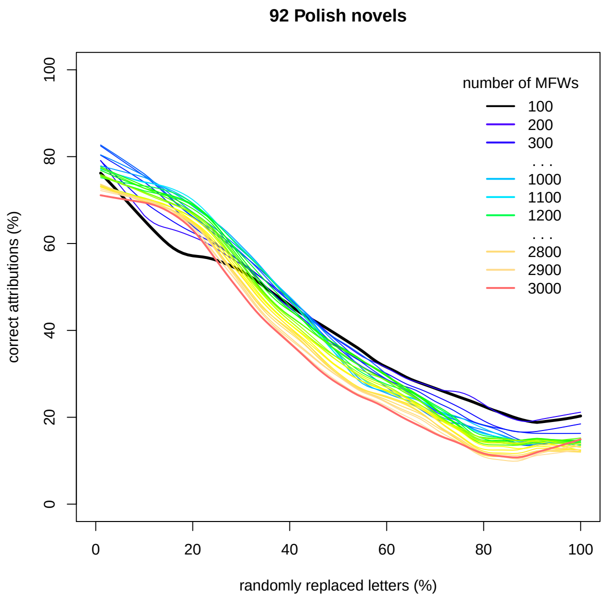

Figure 6: Simulation of poor OCR quality in the corpus of Polish novels.

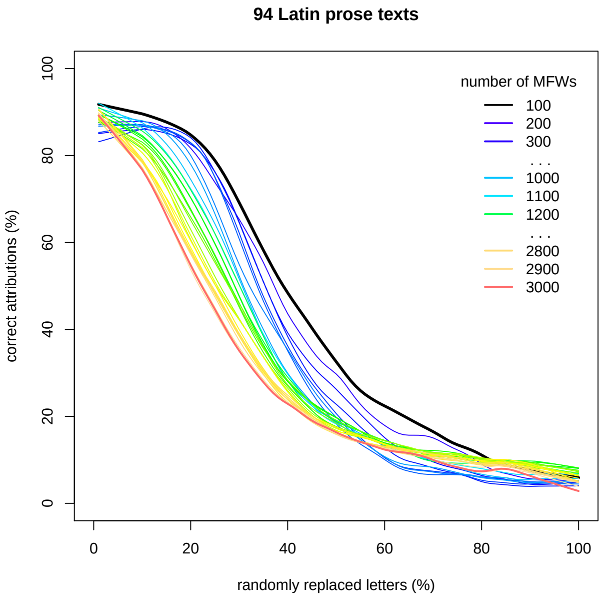

Figure 7: Simulation of poor OCR quality in the corpus of Latin prose texts.

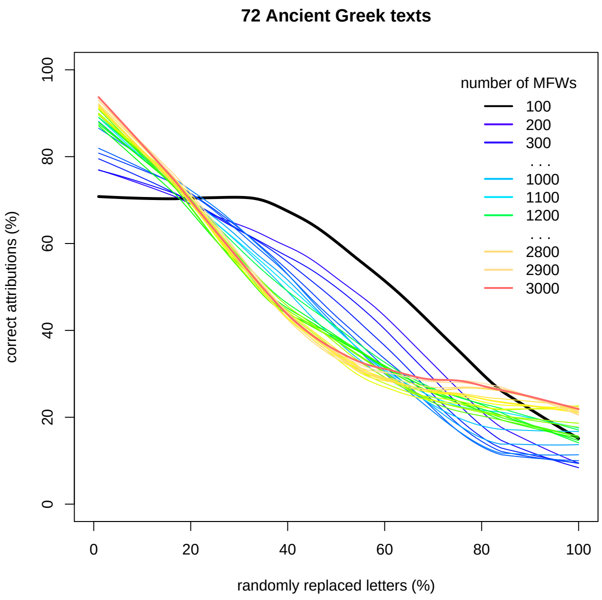

Figure 8: Simulation of poor OCR quality in the corpus of Ancient Greek prose texts.

of performance. Intuitively, this could be expected, as the top of frequency lists is usually occupied by short words, which are less likely to contain misspelled characters. This phenomenon is quite clearly evidenced in the German corpus (Fig. 5), i.e. in a language with words usually longer than those in other languages.

It is very important to stress, however, that using short MFW vectors – despite their considerable resistance to text damage – still provides worse performance than relying on a large number of MFWs. This means that the 'garbage in, gospel out' optimism is in fact illusory.

To validate the results, a few variants of the above experiment have been performed. First, instead of entire texts, samples of 10,000 randomly excerpted words have been analyzed. The results for the corpus of English are shown in Fig. 2; the other corpora exhibited very similar behavior. The observed effect of resistance to damage for shorter MFWs vectors is hardly noticeable now, but the performance of long vectors seems to be stable, i.e. independent of the modified sample size.

Next came a number of tests with a different classification algorithm (SVM) and/or other style-markers. Generally, no significant difference in attribution accuracy could be observed using SVM applied to MFWs. Character-based markers, however, revealed an impressive increase of performance, regardless of the classification method used. As evidenced in Fig. 3 (for Delta applied to character 4-grams), the threshold where the noise finally starts to overwhelm the attributive scores is settled somewhere around the point of 40% spoiled characters. The performance of SVM applied to character 4-grams was equally resistant to corpus damage.

The above impressive results deserve verification using other character-based markers: the performance of SVM applied to character 3-grams is shown in Fig. 4. It is true that the results are somewhat cluttered as compared to the scores obtained for 4-grams (Fig. 3), but it is the behavior of the shortest vector of 100 most frequent 3-grams (Fig. 4, thick black line) that is the reason to show this diagram. It is hard to believe how robust this type of style-marker is when confronted with a dirty corpus – to kill the authorial signal efficiently, one needs to distort more than 60% of original characters (!). On the other hand, however, one has to bear in mind that the most resistant sets of features – no matter if it is 100 character 3-grams, 100 character 4-grams, or 100 MFWs – are always the least effective in author discrimination as compared with longer vectors (with an exception of the Latin corpus). This means that to achieve better accuracy, one has to sacrifice resistance to noise, and *vice versa*. Nevertheless, the above tests with SVM and alternative markers show a big potential of character *n*-grams as a way to by-pass the problem of dirty corpora.

4 Noise in word frequencies

The aim of the second experiment is to explore the impact of scribal and editorial modifications of literary texts. These include orthographic variants, scribal interpolations, editorial textual adjustments, punctuation introduced by editors, etc. What is worse, scholars arguably claim that in preparing critical editions of early modern Latin texts, no consistent spelling principles can be proposed (Deneire, 2013). It means that even the future digital editions will continue to contain noise and editorial inconsistencies.

Difficulties in dealing with this type of noise in information retrieval are stressed by Gotcharek et al. (2011): 'it would be naive to expect that problems are solved as soon as OCR works in an acceptable way. [...] A non-expert user – who of course uses modern keywords in his queries – will miss a large amount of relevant documents in the answer set, being unaware of all the variants how words are written in historical texts'. A corpus that contains such texts is not merely *damaged*, i.e. it is clean of misspelled characters. However, occurrences of particular words and/or punctuation marks – as counted by a machine – will be more or less biased due to mismatching strings of characters. This bias might be used to find unique scribal idiolects (Kestemont and Dalen-Oskam, 2009; Thaisen, 2011); it can be also subjected to automatic disambiguation of spelling variants (Craig and Whipp, 2010, Gotscharek et al., 2011). In most approaches, how-

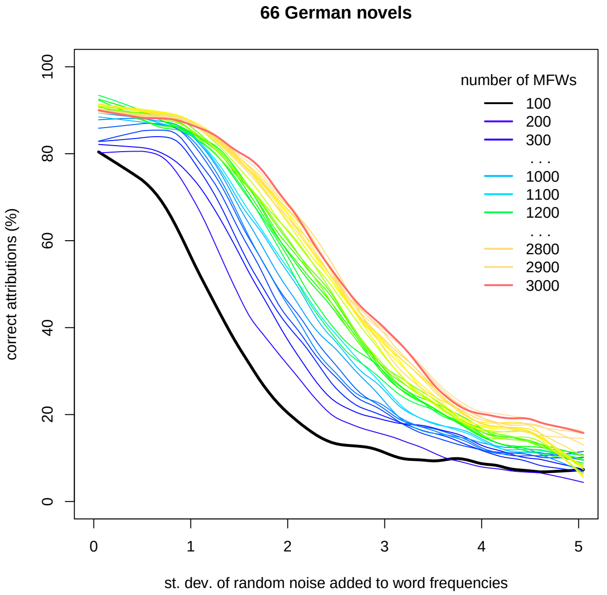

Figure 9: Simulation of editorial modifications in the corpus of German novels.

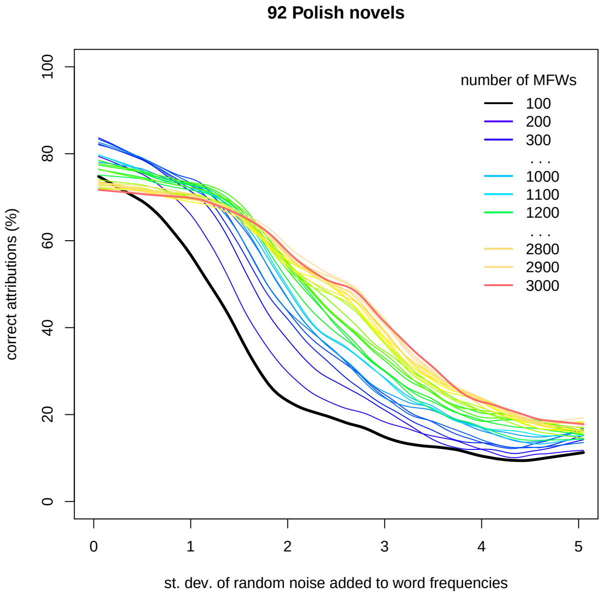

Figure 10: Simulation of editorial modifications in the corpus of Polish novels.

ever, there is no sufficient awareness of potential systematic error (Rudman, 1998a).

The potential bias in word counts can be simulated by adding random noise – gradually increasing its standard deviation – to the computed vectors of word frequencies. Thus, in the first of 100 iterations, the added noise would have as little variance as $0.05\sigma_i$ and the last iteration would include huge noise of $5\sigma_i$ variance (which means that the noise is 5 times stronger than the variance of given word frequencies it is added to).

The results seem to be quite similar to those obtained in the previous experiment – but it is worth to note that the pictures are in fact *symmetrical* (Fig. 1–8 vs. 9–10). Here, short MFW vectors are significantly sensitive to noise in word frequencies, while the longest vectors can survive a moderate earthquake: even very strong noise – its strength comparable with the variance of the words it infects – has a rather weak influence on attribution effectiveness. The results were roughly similar in each corpus tested.

5 Impact of literary tradition

The last type of noise can hardly be called systematic error. Namely, the aim of this experiment is to simulate the impact of literary inspirations (plagiarism, imitations, intertextuality, etc.) on attribution effectiveness. In authorship studies, there is always a tacit – and somewhat naïve – assumption that texts in a corpus are purely 'individual' in terms of being written solely by one author and not influenced by other writers – as if any text in the world could be created without references to the author's predecessors and to the whole literary tradition. It might be formulated using the words of Ecclesiastes 1, 10: 'Is there any thing whereof it may be said, See, this is new? it has been already of old time, which was before us'.

The problem of collaborative nature of early modern texts has been discussed by traditional literary criticism (Hirschfeld, 2001; Love, 2002), but it is hardly reported in computational stylistics – some general issues concerning plagiarism detection are discussed by Wilks (2004). In non-traditional authorship attribution, this inherent feature of written texts is certainly a pitfall; in the broad area of investigations concerning stylistic similarities between texts, however, the same feature makes it possible to use stylometric techniques to answer a question whether style is anyhow determined by chosen topic or genre, to explore diachronic aspects of style, such as development of style through literary periods, to trace stylistic similarities between different writers that could reflect their literary influences. These include conscious imitation of style, stylistic inspirations, parody, the phenomenon of prequels/sequels written by later authors, an impact of a distinguished writer on his contemporaries, collaborative authorship, collaborative translations, and other aspects of intertextuality (Hoover, 2009; Pennebaker and Ireland, 2011; Rybicki, 2011; Rybicki and Heydel, 2013; Jockers, 2013, etc.).

To test the role of 'intertextuality' in authorship attribution, the experiment was designed as follows. In each of 100 iterations, for each text, a consecutive percentage of original words were replaced with words randomly chosen *from the entire corpus*. Thus, a simulation of increasing intertextual dependence between the texts in a corpus was obtained. Certainly, quotations from external sources usually cover longer strings of words: sentences or even paragraphs rather than single words. On the other hand, however, subtle literary similarities and stylistic influences in most cases go beyond mere quotations; arguably, it is the usage of single words, word collocations, short phrases, idioms, and expressions – rather than the number of explicit quotations – that makes two texts similar. For this reason, replacing random individual words seemed to be a better way to assess intertextual similarities between literary works than swapping longer snippets between particular texts and the whole corpus.

The obtained results are quite interesting. The corpora of modern literatures, i.e. English (Fig. 11), German and Polish, displayed a gentle decrease of performance (despite the number of MFWs analyzed) in correlation with the amount of 'intertextuality' added. The Greek corpus exhibited a considerable stability despite a reasonable amount of 'intertextual' noise added. However, the Latin corpus (Fig. 12) behaved as if authorial uniqueness could still be traced through a mass of external quotations:

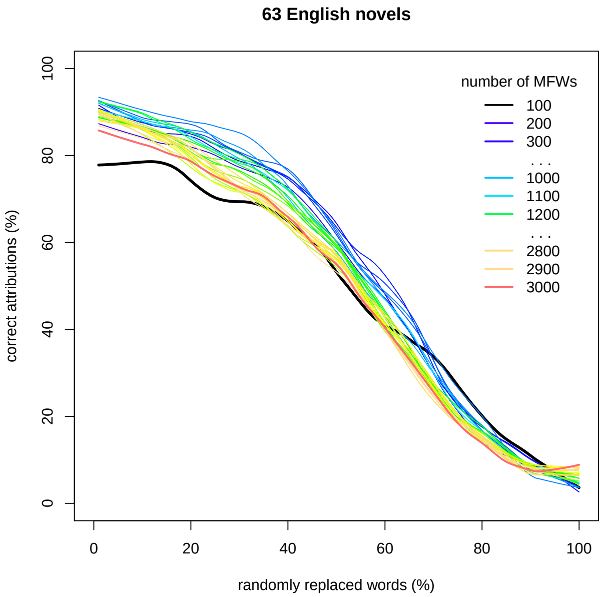

Figure 11: Simulation of extreme intertextuality in the corpus of English novels.

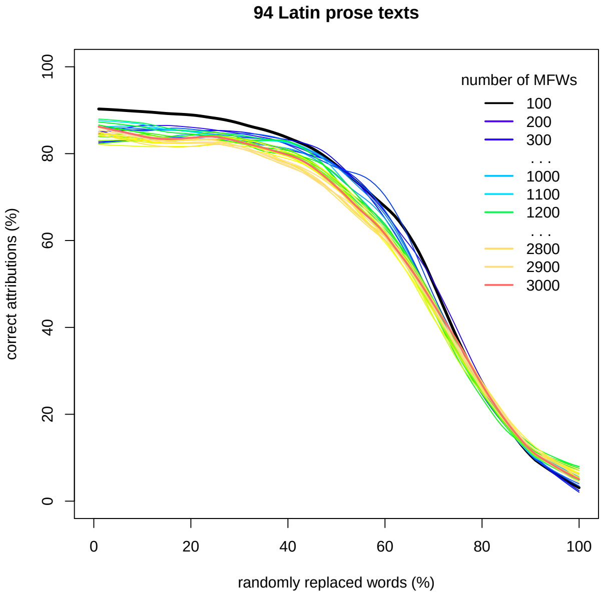

Figure 12: Simulation of extreme intertextuality in the corpus of Latin prose texts.

a considerably good performance was achieved despite 40% of original words replaced (these results were not further improved using SVM algorithm and/or character n-grams as features). This deserves further investigation. Intuitively, these results might be explained by very strong similarities between Latin classical texts. Since the educational system of Antiquity stressed the role of imitation in style development, we could expect Ancient texts to be very similar one to each other. This explanation, however, is too simple to be feasible. If the texts were really similar, the overall attribution scores would also be very low – and they were not.

6 Conclusions

In the present study, three different experiments to test the impact of three different types of noise have been conducted. The first test was designed to show the correlation between a dirty corpus and attribution accuracy. The results partially confirmed the claims that some tolerance in corpus preparation might be acceptable (the test for 100 MFWs did not prove significant decrease of performance even for 20% of damage); the general picture, however, showed the importance of tidily prepared corpora (a vast majority of tests turned out to be sensitive to corpus damage). In additional tests using SVM as classification technique and character n-grams as features, however, an impressive robustness of character-based markers has been observed.

The second experiment was aimed to test how disorder in word frequencies – produced by scribal and/or editorial modifications – affects the attribution abilities of particular corpora. The results were quite optimistic: very long vectors of words tested could overwhelm, to some extent, the effect of noise.

Last but not least, the goal of the third experiment was to test how much 'authorial' data a given texts needs to have in order to extract authorial fingerprint. Even if preliminary, the results seem interesting, and the behavior of the Latin corpus was difficult to explain. The approach to the problem of 'intertextuality' as presented below is but the top of an iceberg, though. Since actual and virtual links between literary texts are multi-layered and extremely complex, some more extensive research is guaranteed here. One of the aims of this study was to show the importance of this issue, usually underestimated in stylometric studies.

References

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. and Whipp, R. (2010). Old spellings, new methods: automated procedures for indeterminate linguistic data. *Literary and Linguistic Computing*, **25**(1): 37–52.

Crane, G. (2006). What do you do with a million books? *D-Lib Magazine*, **12**(3). http://www.dlib.org/dlib/march06/crane/03-crane.html (accessed 17 May 2013).

Deneire, T. (2013). Editing Neo-Latin texts: editorial principles; spelling and punctuation. In: Bloemendal J., Fantazzi C., Ford P. (eds.), *Encyclopaedia of Neo-Latin Studies*. Brill (in press).

Eder, M. (2010). Does size matter? Authorship attribution, small samples, big problem. *Digital Humanities 2010: Conference Abstracts*. King's College London, pp. 132–35.

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.unesco.uj.edu.pl/media/SPL-Vol.-65.pdf (accessed 17 May 2013).

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**, doi:10.1093/llc/fqs036 (published on-line 11 August 2012).

Gotscharek, A., Reffle, U., Ringlstetter, Ch., Schulz, K. U., Neumann, A. (2011). Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *International Journal of Document Analysis and Recognition*, **14**: 159–71.

Hirschfeld, H. (2001). Early modern collaboration and theories of authorship. *PMLA*, **116**(3): 609–22.

Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, **15**(3/4). http://www.dlib.org/dlib/-march09/holley/03holley.html (accessed 17 May 2013).

Hoover, D. L. (2009). Modes of composition in Henry James: dictation, style, and 'What Maisie Knew'. *Digital Humanities 2009: Conference Abstracts*. University of Maryland, College Park, MD, pp. 145–48.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.

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.

Kestemont, M. and Van Dalen-Oskam, K. (2009). Predicting the past: memory based copyist and author discrimination in Medieval epics. *Proceedings of the 21st Benelux Conference on Artificial Intelligence (BNAIC) 2009*. Eindhoven, pp. 121–28.

Koppel, M., Schler, J. and Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, **45**: 83–94.

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

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.

Noecker, J., Ryan, M., Juola, P., Sgroi, A., Levine, S. and Wells, B. (2009). Close only counts in horseshoes and... authorship attribution? *Digital Humanities 2009: Conference Abstracts*. University of Maryland, College Park, MD, pp. 380–81.

Pennebaker, J. W. and Ireland, M. E. (2011). Using literature to understand authors: the case for computerized text analysis. *Scientific Study of Literature*, **1**: 34–48.

Reynaert, M. W. C. (2011). Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal of Document Analysis and Recognition*, **14**: 173–87.

Reynolds, L. D. and Wilson N. G. (1978). *Scribes and Scholars: A Guide to the Transmission of Greek and Latin Literature*. Oxford: Clarendon Press.

Robertson, A. M. and Willett, P. (1993). A comparison of spelling-correction methods for the identification of word forms in historical text databases. *Literary and Linguistic Computing*, **8**(3): 143–52.

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. (2011). Alma Cardell Curtin and Jeremiah Curtin: the transtalor's wife's stylistic fingerprint. *Digital Humanities 2011: Conference Abstracts*. Stanford University, Stanford, CA, pp. 219–22.

Rybicki, J. and Heydel, M. (2013). The stylistics and stylometry of collaborative translation: Woolf's 'Night and Day' in Polish. *Literary and Linguistic Computing* **28** (in press).

Thaisen, J. (2011). Probabilistic analysis of Middle English orthography: the Auchinleck Manuscript. *Digital Humanities 2011: Conference Abstracts*. Stanford, CA, pp. 248–50.

Wake, W. C. (1957). Sentence-length distributions of Greek authors. *Journal of the Royal Statistical Society. Series A*, **120**(3): 331–46.

Wilks, Y. (2004). On the ownership of text. *Computers and the Humanites*, **38**: 115–27.