Mind your corpus: systematic errors in authorship attribution

Maciej EDER

Pedagogical University, Kraków, Poland

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. In other words, 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 are *random variables*; the same can be said about any written authorial text, like a novel or poem.

Although dealing with this unavoidable noise is *crème de la crème* of computational stylistics, any other influence of additional randomness – 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 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 problem of systematic errors in stylometry 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 OCR errors on attribution accuracy, have observed that a 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 another study, a strong correlation between the length of input samples and attribution performance has been shown (Eder, 2010). In these and many other studies, however, the problem of systematic errors has not been addressed systematically.

The nature of noise affecting the results is quite complex. On the one hand, a machine-readable text might be contaminated by a 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.

To verify the impact of unwanted noise, a series of experiments has been conducted on several corpora of English, German, Polish and Latin prose texts, the corpora being roughly similar in length and number of authors tested. In 100 iterations, a given corpus was gradually damaged and controlled tests for authorship have been applied (the procedure has been inspired by the study of Noecker et al. 2008). It can be obviously assumed that heavy damage will 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 has been used (Burrows, 2002). It is a very intuitive procedure that performs considerably well as compared with much more sophisticated techniques. Like other machine-learning 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. The obtained results should be – by extension – valid for other methods that rely on multidimensional comparison of frequencies of MFWs.

It should be stressed that one can deal with two general types of systematic errors. First, when *all* the samples are damaged to a similar degree (e.g. if the corpus was not cleaned of markup tags); second, when, in a carefully collected corpus, *some* of the samples (e.g. one sample) are of poor quality. The latter case will not be discussed in the present study.

Misspelled characters

The first experiment addresses a very trivial, yet severe, type of damage – the situation where single characters are misspelled due to transmission disturbance, imperfect typing, poor quality of scanned documents, using untrained OCR software, etc. This type of error is distributed randomly in a string of characters; however, some letters are more likely to be damaged than others. Especially, white spaces are rarely misspelled.

The experiment, then, was designed as follows. In each of 100 iterations, an increasing percentage of letters (excluding spaces) were replaced with other randomly chosen letters. To give an example: in the 15th iteration, every letter of the input text was intended to be damaged with a 15% probability; in consequence, the corpus contained roughly 15% of independently damaged letters.

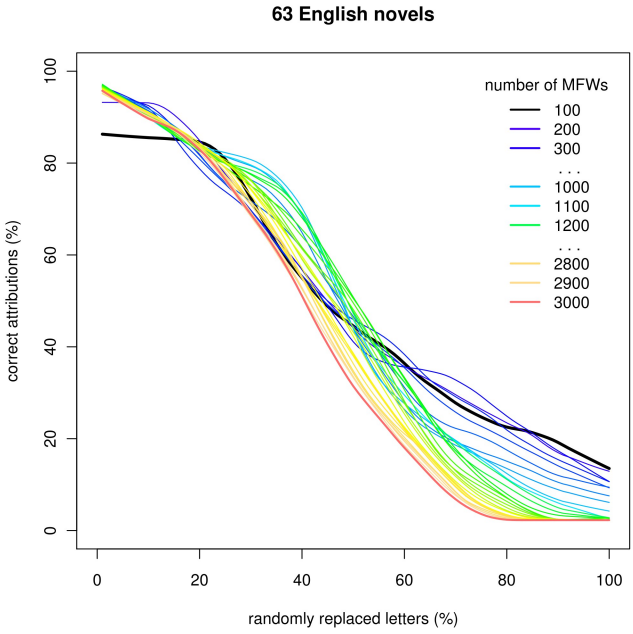

Figure 1. Simulation of poor OCR quality in the corpus of English novels: in 100 iterations, increasing percentage of intentionally misspelled characters has been tested for 30 different MFW vectors.

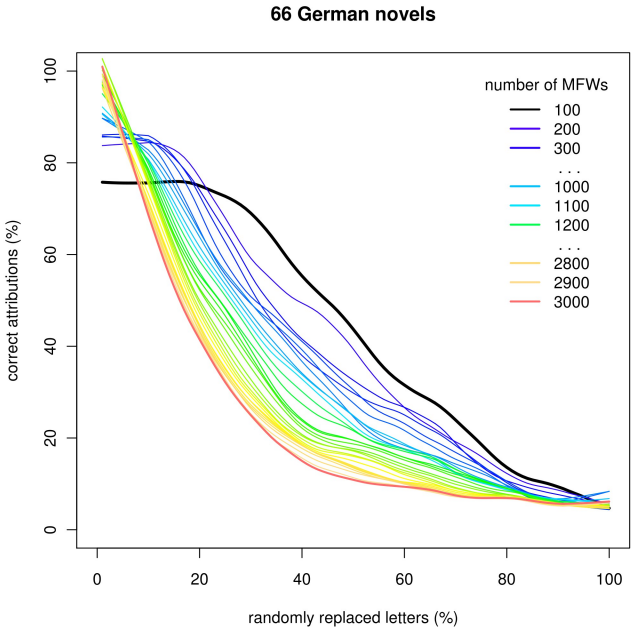

Figure 2. Simulation of poor OCR quality in the corpus of German novels.

The results were quite similar for most of the languages tested. As shown in Fig. 1 and 2, short vectors of MFWs (up to 500 words) usually provide no significant decrease of performance despite a considerably large amount of noise added. Even 20% of damaged letters will not affect the results in some cases! However, longer MFW vectors are *very sensitive* to misspelled characters: any additional noise means a steep decrease 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

randomly misspelled characters. This phenomenon is quite clearly evidenced in the German corpus (Fig. 2): i.e. in a language with words usually longer than those in other languages.

This is very important to stress, however, that using short MFW vectors – despite their considerable resistance to damaged texts – 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.

Noise in word frequencies

The aim of the second experiment is to explore the impact of scribal and editorial modifications of the literary texts. These include orthographic variants, scribal interpolations, editorial textual adjustments, punctuation introduced by modern scholars, etc. A corpus that contains such texts is not merely *damaged*, i.e. it is clean of misspelled characters. However, due to different spellings used, the obtained word frequencies are likely to be biased. This bias might be used to find unique scribal idiolects (Kestemont and Dalen-Oskam, 2009); it can be also subjected to automatic disambiguation of spelling variants (Craig and Whipp, 2010). In most approaches, however, there is no sufficient awareness of potential systematic error (Rudman, 1998a).

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

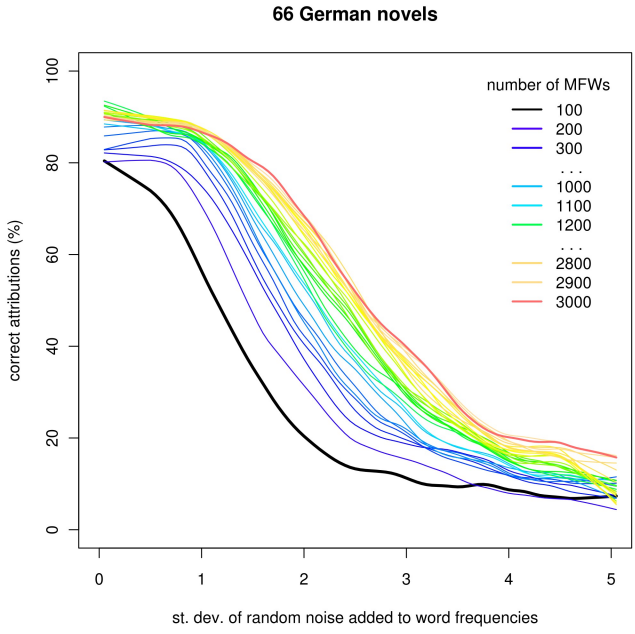

Figure 3. Simulation of editorial modifications in the corpus of German novels.

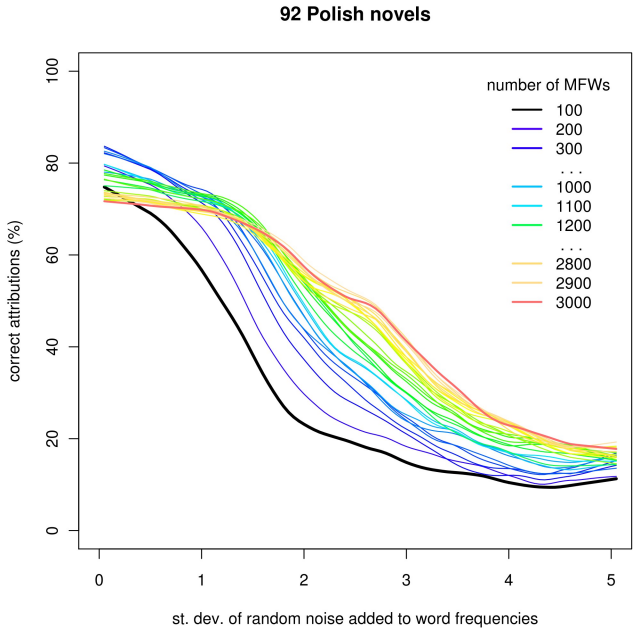

Figure 4. Simulation of editorial modifications in the corpus of Polish novels.

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. 1–2 vs. 3–4). Here, short MFW vectors are significantly sensitive to randomness in frequency tables, 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.

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. 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. In authorship attribution, this feature of written texts is certainly a pitfall; however, the same feature makes it possible to use stylometric techniques to trace stylistic imitations or unconscious inspirations between different authors.

The experiment is designed as follows. In each of 100 iterations, for each text, a consecutive percentage of original words are replaced with words randomly chosen *from the entire*

corpus. Thus, a simulation of increasing intertextual dependence between the texts in a corpus is obtained.

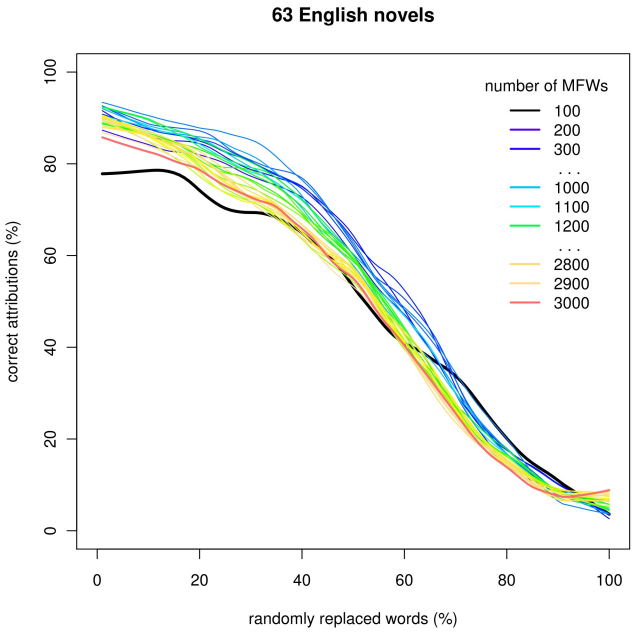

Figure 5. Simulation of extreme intertextuality in the corpus of English novels.

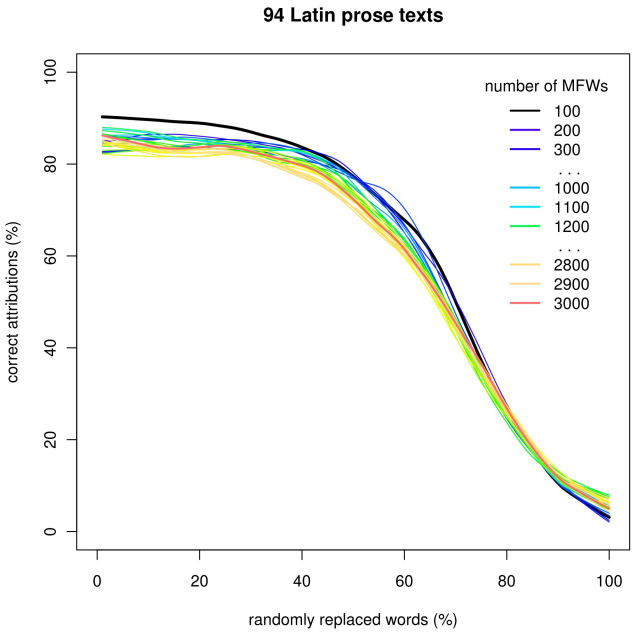

Figure 6. Simulation of extreme intertextuality, in the corpus of Latin prose texts.

The obtained results are quite interesting. The corpora of modern literatures, i.e. English (Fig. 5), German and Polish, displayed a gentle decrease of performance (despite the number of MFWs analyzed) in correlation with the amount of 'intertextuality' added. The Latin corpus (Fig. 6) behaved as if the authorial uniqueness could be traced through a mass of external quotations: a

considerably good performance was achieved despite 40% of original words replaced. This deserves further investigation.

References

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

Burrows, J. F. (2002). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship, *Literary and Linguistic Computing*, **17**(3): 267–87.

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** (in press).

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

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.

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

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.

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.