**EDER — Taking**

<2 Figures>

# 

# **Taking Stylometry to the Limits: Benchmark Study on 5,281 Texts from *Patrologia Latina***

## Eder, M.

In the era of large-scale stylometry—which is about to come—some basic but difficult methodological questions have not been answered yet. Certainly, the most important one is whether a given method, reasonably effective for a collection of, say, 100 texts, can be scaled up to assess thousands of texts without any significant side effects. When one deals with historical corpora, however, this question becomes much more complex, since several additional factors have to be taken into consideration. Spelling variation, insufficiently trained NLP models, corpora a priori unbalanced—these are the obvious issues. However, one should also take into account less obvious yet equally important factors that make any stylometric investigations nontrivial, to say the least. These include editorial corrections, punctuation introduced by modern scholars, hidden plagiarism and/or text reuse problems, innumerable authorship issues, and many other sources of potential stylometric ‘noise’ (Eder, 2013).

In the present study, some of the above issues will be undertaken. The benchmark, however, will be focused on one aspect of text classification only, namely authorship attribution. Other stylistic layers—e.g., genre recognition and topic distinction—will be addressed in a follow-up study.

### Dataset

The complete collection of *Patrologia Latina* that was used in this study, and that will be soon made available in the form of a carefully prepared corpus with morphosyntactic annotation (A. Guerreau, 2014, pers. comm.), gives us a great opportunity to test some of the above assumptions and possible drawbacks of the state-of-the-art stylometric methods. The aforementioned collection consists of 5,821 texts by over 700 authors—the Latin Church Fathers. It covers a time span of over 1,000 years (2nd–13th centuries). Even if the texts represent a few genres, the collection is thematically very consistent: for obvious reasons, theology overwhelms other topics. At the same time, however, *Patrologia Latina* is a pre-Internet example of a big-yet-dirty text collection, published in the years 1844–1855; the goal was to publish all the available material in a relatively short period of time, with the assumption that particular volumes would be gradually replaced by carefully prepared critical editions.

The above factors make the electronic collection of *Patrologia Latina* a relatively difficult benchmark corpus, close to real-life attribution cases. Since the number of texts to be analyzed is very high, one cannot inspect them manually and/or emend the transcription. Also, one cannot reliably exclude all external quotations, passages copied from other sources, and similar intertextual links; the same applies to any instances of (hidden) mixed authorship. In this case big data means big noise. On the other hand, however, being large and noisy, the collection is a perfect case for massive authorship attribution test, and particularly for exploring the ‘needle in a haystack’ attribution scenario (Koppel et al., 2009). Moreover, since the texts are supplemented with grammatical codes (lemmata and POS-tags), it gives us a unique opportunity to test the attributive performance of lemmatized vs. nonlemmatized corpus, POS-tag *n*-grams vs. MFWs, and so forth. Last but not least, the corpus is supplemented with consistent metadata—the authors’ names are easily retrievable, as well as genre, chronology, and the length of particular texts.

### Preselection of Texts

The collection of *Patrologia Latina* contains a few dozen texts known to be anonymous, as well as several works of uncertain authorship. For obvious reasons, these texts were excluded from the analysis. What should be emphasized, however, is the fact that the authorship identification of the remaining works is as reliable as the 19th-century scholarship on the Church Fathers. To give an example: the treatise *On the Spectacles* (*De spectaculis*), for centuries attributed to St. Cyprianus and published under the name Pseudo-Cyprianus in *Patrologia Latina*, is nowadays assumed to have been written by Novatianus. In the whole collection of over 5,000 texts, there must be a good share of similar, yet still undiscovered, mismatches.

The corpus was further reduced to contain only those texts that had at least 3,000 tokens, since sample length is known to be a major issue in attribution (Eder, 2014). For the sake of supervised text classification, which requires at least two texts per author to carry out the attribution test, further culling was performed—the authors of fewer than three works have been filtered out. After all the reductions, the corpus consisted of 1,665 works by 197 authors.

Since the number of authors is one of the most important factors affecting the performance of attribution methods (Luyckx and Dealemans, 2011), two smaller corpora have been prepared as well. The first consisted of a randomly chosen subset from the entire *Patrologia Latina* (10% of texts) and further abridged to meet the criteria of 3,000 tokens and known authorship (resulting in 81 texts by 15 authors); the second corpus was based on a 5% subset from the original collection, further reduced according to the same rules (22 texts by five authors after culling). The experiments discussed in this paper have been performed on the big corpus, and then replicated using the two smaller subcorpora.

### Method of Benchmarking

A standard procedure of supervised classification has been applied. Namely, the input dataset was divided (randomly) into two groups: a ‘training set’ containing two texts per each author, and a ‘test set’ for all the remaining samples. Then the stage of ‘guessing’ was applied, which was aimed to examine if the samples from the second group were correctly linked to their actual authors as represented in the training corpus.

A great number of independent controlled experiments have been performed for every combination of different parameters: three types of style markers (most frequent words, POS-tags, lemmata) in two variants (with and without punctuation), three levels of *n*-grams (1-grams, 2-grams, 3-grams), at five levels of culling (0%, 20%, 40%, 60%, 80%), for twenty different vectors of the most frequent features (100, 150, 200, 250, . . . , 1,000), and for six classifiers. The classifiers were as follows: support vector machines, nearest shrunken centroids (Jockers et al., 2008), *k*-Nearest Neighbor classifier, naïve Bayes, classic Delta (Burrows, 2002), and a home-brew classifier available in the stylometric package ‘stylo’ (Eder et al., 2013) under the name Eder’s Delta. For each test, 20-fold cross-validation was performed.

To compute such a big number of stylometric tests—the total number of experiments exceeded 30,000—a tailored version of the R package ‘stylo’ was used, supplemented by a few high-performance packages, e.g., ‘multicore’ or ‘ff’. Even still, it took several weeks to complete the task on a 6-processor Xeon machine equipped with 16 Gb of RAM.

### Results

The outcome of a vast majority of tests was disappointing: attribution accuracy was unexpectedly poor. First suspicion that the number of almost 200 authorial classes was responsible for this effect turned out to be unfounded when smaller subsets of *Patrologia Latina* were scrutinized. The results were obviously better, but still unsatisfactory. For the subset of 81 texts by 15 authors, the optimal attributive success achieved was as high as 50% for the most effective set of input parameters (Figures 1 and 2). This surprisingly weak authorial signal in the corpus of medieval Latin deserves further investigation.

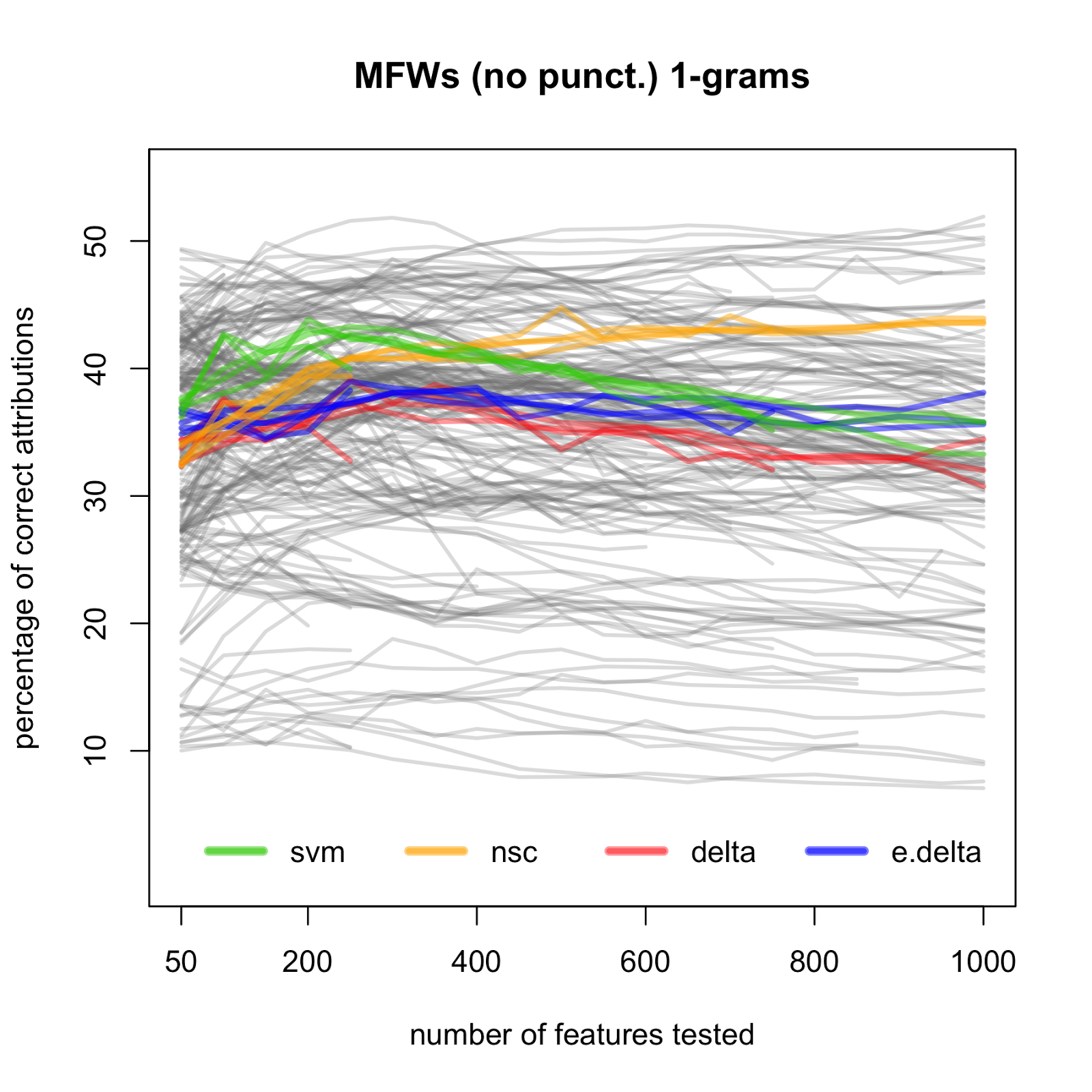


Figure 1. Performance of four methods of classification tested on the Most Frequent Words (grey lines in the background represent ca. 600 remaining combinations of style-markers, n-grams, classifiers, etc.).

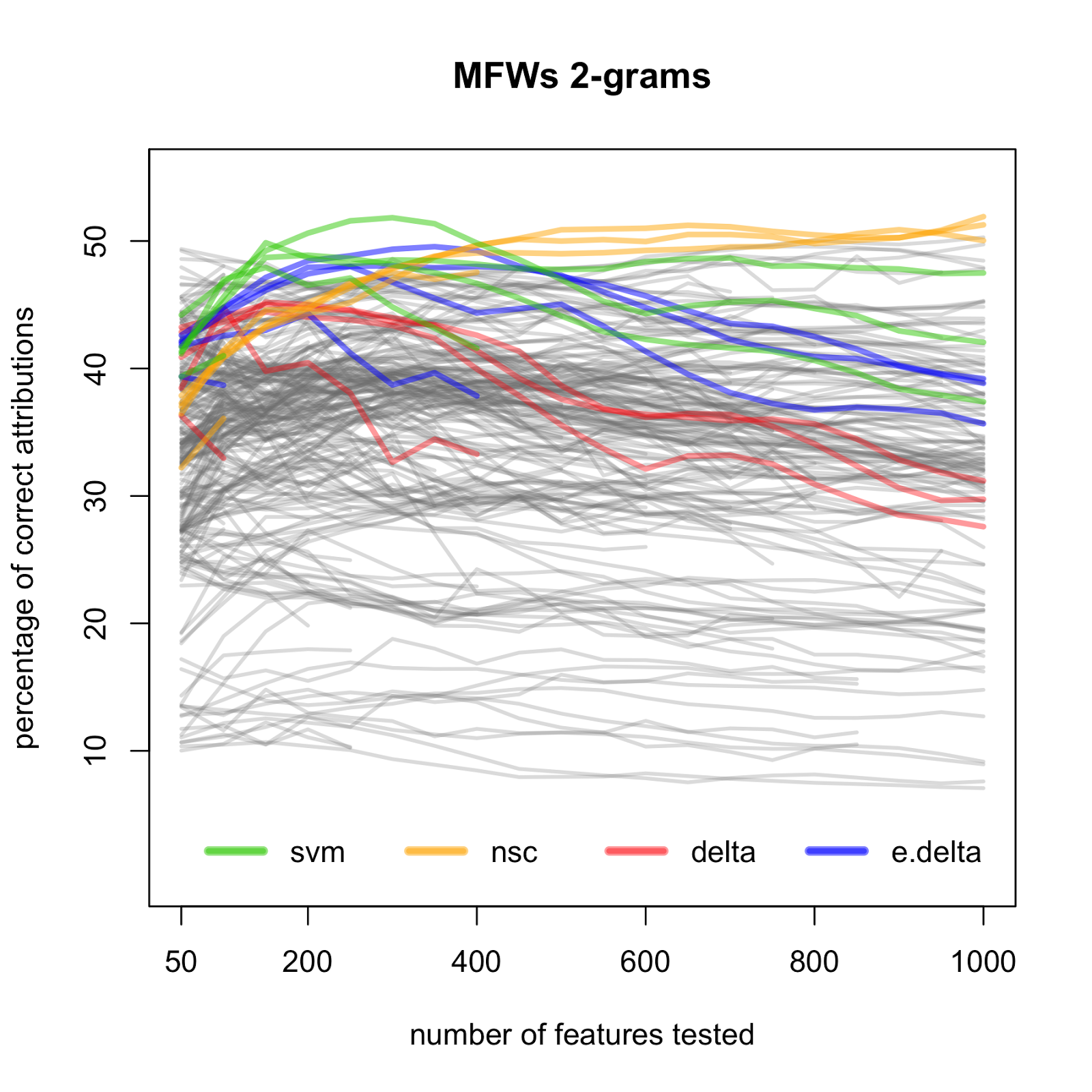


Figure 2. Performance of four methods of classification tested on bi-grams of frequent words including punctuation marks.

Massive stylometric experiments allowed a systematic comparison of particular parameters’ performance. It turned out that the most accurate classifier was nearest shrunken centroids (especially when applied to long vectors of features), then came support vector machines. Eder’s Delta outperformed the classic Burrowsian measure in almost every instance; both *k*-nearest neighbor classifier and naïve Bayes occupied the tail of the ranking. When it comes to the style-markers (features) examined, some counterintuitive results could be observed as well. Frequencies of Most Frequent Words (MFWs)—that is, the type of style markers routinely used for authorship attribution—proved to be suboptimal (Figure 1), while the best performance was achieved using frequent word bi-grams, or combinations of two adjacent words including punctuation (Figure 2). Lemmata and POS-tags followed a similar pattern: they were more accurate when combined into bi-grams including punctuation.

Although it has been shown that punctuation generally increases attribution accuracy (Baayen et al., 2002), this statement must be considered suspicious in the context of ancient/medieval Latin. It is a fact commonly accepted that Latin originally had no punctuation—it was introduced by early modern scholars (Reynolds and Wilson, 2013 [1978]). Apparently, then, it should always be avoided as an authorial style discriminator. On the other hand, however, Latin punctuation is not a randomly scattered artifact, but it inescapably follows the rules of language. Even if artificial, it reveals some linguistic characteristics of analyzed texts, like in the Platonic cave, in which vague shadows reflect some actual phenomena. Interestingly, punctuation was a stronger style discriminator than a more explicit indicator of syntax, such as *n*-grams of POS-tags.

### References

**Baayen, H., van Halteren, H., Neijt, A. and Tweedie, F.** (2002). An Experiment in Authorship Attribution. *JADT 2002: 6es Journées internationales d'Analyse statistique des Données Textuelles*. St. Malo, pp. 29–37.

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

**Eder, M.** (2013). Mind Your Corpus: Systematic Errors and Authorship Attribution. *Literary and Linguistic Computing,* **28**(4): 603–14.

**Eder, M.** (2014). Does Size Matter? Authorship Attribution, Small Samples, Big Problem. *Literary and Linguistic Computing,* **29**, doi:10.1093/llc/fqt066.

**Eder, M., Kestemont, M. and Rybicki, J.** (2013). Stylometry with R: A Suite of Tools. *Digital Humanities 2013: Conference Abstracts*. University of Nebraska–Lincoln, pp. 487–89.

**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.

**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.

**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.

**Reynolds, L. D. and Wilson, N. G.** (2013 [1978]). *Scribes and Scholars: A Guide to the Transmission of Greek and Latin Literature*. Clarendon Press, Oxford.