Delta vs. N-Gram Tracing: Evaluating the Robustness of Authorship Attribution Methods

Thomas Proisl*, Stefan Evert*, Fotis Jannidis†, Christof Schöch‡, Leonard Konle†, Steffen Pielström†

*Friedrich-Alexander-Universität Erlangen-Nürnberg, †Julius-Maximilians-Universität Würzburg, ‡Universität Trier

*Bismarckstr. 6, 91054 Erlangen, †Am Hubland, 97074 Würzburg, ‡Universitätsring 15, 54296 Trier

*{thomas.proisl, stefan.evert}@fau.de,

†{fotis.jannidis, christof.schoech, leonard.konle}@uni-wuerzburg.de, pielstroem@biozentrum.uni-wuerzburg.de,

‡schoech@uni-trier.de

Abstract

Delta measures are a well-established and popular family of authorship attribution methods, especially for literary texts. N-gram tracing is a novel method for authorship attribution designed for very short texts, which has its roots in forensic linguistics. We evaluate the performance of both methods in a series of experiments on English, French and German literary texts, in order to investigate the relationship between authorship attribution accuracy and text length as well as the composition of the comparison corpus. Our results show that, at least in our setting, both methods require relatively long texts and are furthermore highly sensitive to the choice of authors and texts in the comparison corpus.

Keywords: authorship attribution, stylometry, evaluation

1. Introduction

Authorship attribution, i. e. the identification of the true author of a text of unknown or disputed authorship based on quantitatively measured linguistic evidence (Juola, 2006; Koppel et al., 2009; Stamatatos, 2009), has applications in many fields, including literary studies, history, forensic linguistics and corpus stylistics. It is based on the assumption that individual writers have idiosyncratic habits of language use (which they are usually not aware of) that lead to stylistic similarities between texts written by the same author. A wide range of stylometric features has been proposed to capture these idiosyncrasies, ranging from relative frequencies of function words to measures of vocabulary richness and syntactic complexity. Based on such feature vectors, a disputed text can then be attributed to the most similar of a set of candidate authors.

If we want to apply authorship attribution methods in real-world settings, e. g. in forensic linguistics, the reliability and robustness of the methods are of utmost importance. Various factors can have an impact on the methods, raising questions such as: To what extent does authorship attribution accuracy depend on the length of the disputed text and the size of the corpus against which it is compared? Is there a minimum text length below which results become too unreliable? What impact does the composition of the comparison corpus have? Are the methods robust with respect to the selection of authors and texts for the comparison corpus?

In this paper, we try to answer those questions, at least partially, for Delta and N-gram tracing, two particularly simple but very successful authorship attribution methods that only rely on word- and character-level features.

2. Background and Related Work

2.1. Delta Measures

Delta measures (Burrows, 2002; Argamon, 2008) are a popular family of authorship attribution methods. They represent texts as simple bags-of-words, focusing on the n most frequent words (nMFW) in the corpus. Word frequencies are standardized to z-scores across the corpus, with a mean of 0 and standard deviation of 1 for each word form type. Text similarity is then quantified by some metric on the resulting vectors of z-scores. Popular choices are Manhattan distance, resulting in the original Burrows's Delta (2002), and angular ("cosine") distance, leading to Cosine Delta proposed by Smith and Aldridge (2011). Typically, (hierarchical) clustering is applied to the distance matrix of all text pairs and similarities between the texts are visualized in the form of a dendrogram. For the purpose of authorship attribution, a disputed text is assigned to the author of the majority of texts in its cluster. Alternatively, a nearest-neighbour classifier can be used or the MFW statistics can serve as features for a supervised machine learning algorithm.

Jannidis et al. (2015) showed that Cosine Delta is usually superior to other variants of Delta. Cosine Delta is also robust with respect to the choice of nMFW, which is why we focus on this particular variant in our experiments. Other key results on Delta measures were obtained by Rybicki and Eder (2011), who investigated the relationship between the number of MFW and authorship attribution success depending on the language of the materials and found notable differences between languages and even within genres. Eder (2013a) showed how text length interacts with attribution quality and found that depending on language and genre a minimum text length of 2,500 to 5,000 words is required for successful authorship attribution. Eder (2013b) investigated the influence of noise, e.g. from OCR errors, on attribution success rates and found that Delta is robust to a certain amount of noise. It should be noted, that – with the exception of Jannidis et al. (2015), who controlled for number of authors – none of the studies mentioned above controlled for text length or number of different authors.

Furthermore, the observations regarding attribution quality made in previous studies are mainly differences between

individual corpora. Variability caused by the selection of the actual texts for the corpora might contribute to the observed differences. None of the previous studies has systematically investigated the influence of sampling effects in corpus composition.

2.2. N-Gram Tracing

N-gram tracing (Grieve et al., submitted) is a novel authorship attribution method from the field of forensic linguistics. It has been designed for the comparison of a short disputed text with a much larger comparison corpus of plausible candidate authors. N-gram tracing extracts all distinct word or character n-grams of a certain length from the disputed text, then determines the percentage of overlap with each author in the comparison corpus. It is important to note that – in marked contrast to Delta – the frequency of the n-grams plays no role at all. The only thing that matters is how many n-grams types also occur in the author-specific parts of the comparison corpus.[1] Grieve et al. (submitted) also suggest a majority voting scheme to combine n-grams of different lengths in order to improve the robustness of the attribution. In their experiments they found that both word 1-to-3-grams and character 4-to-10-grams worked particularly well. This is also what we do in our experiments: We choose the author that is suggested by the majority of word 1-to-3-grams or by the majority of character 4-to-10-grams.

3. Methodology

To answer the questions raised in the introduction, we perform four experiments: Two shortening experiments that examine the performance of Cosine Delta and N-gram tracing depending on text length and two sampling exeriments that evaluate the robustness of both methods with respect to the composition of the comparison corpus.

To allow for a better comparison between Cosine Delta and N-gram tracing, we do not perform a clustering of the Delta distance matrix, but simply attribute the disputed text to the author of its nearest neighbor (i. e. we use a nearest-neighbour classifier).

For Delta, we use the 3,000 most frequent words, which has previously been found to be a robust choice for all languages (Evert et al., 2017). Fig. 1 shows the interaction between text length and nMFW: even for shorter texts, 3,000 MFW achieve better results than the much shorter word lists used by Burrows (2002) and other early work on Delta.

3.1. Shortening Experiments

The shortening experiments are based on three corpora of German, English and French novels (Jannidis et al., 2015; Evert et al., 2017).[2] Each corpus consists of 75 novels from 25 authors, with three texts from each author. We evaluate authorship attribution accuracy via a stratified three-fold cross-validation scheme: in each iteration 25 novels (one per author) are treated as disputed texts. The previous studies on these corpora found that, on the full texts, Cosine Delta showed excellent and robust results, with Adjusted Rand Index (ARI) scores greater than 0.9 for all three corpora (in general, attribution success was higher for the German novels than for the English and French data), but left open the question of the minimal amount of text required for robust authorship attribution results.

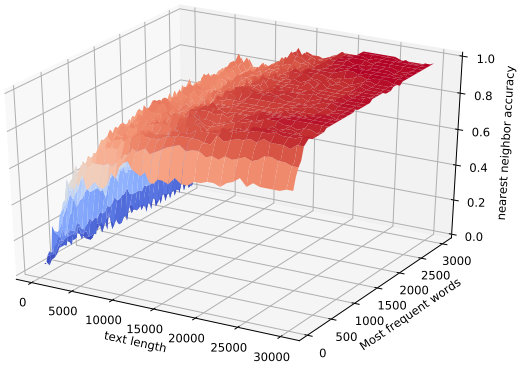


Figure 1: Interaction between text length and nMFW for Cosine Delta in experiment 1a (German corpus)

Experiment 1a: We shorten all texts in the corpus to the same number of tokens, skipping the first 10% of each text because we assume that beginnings and endings of literary texts differ in substantial ways from the rest.

Experiment 1b: We shorten only the disputed text and keep the size of the comparison corpus at a stable size of 30,000 tokens per text, again skipping the first 10% of each text.

3.2. Sampling Experiments

The sampling experiments are based on a collection of 973 German novels by 131 authors, with at least three novels from each author. All authors were native speakers, the collection contains no translations, and the novels were written between 1789 and 1914. We use this collection to draw a large number of samples similar in structure to the corpus of Jannidis et al. (2015), which was used for the shortening experiments.

Experiment 2a: We draw 5,000 random samples of 25 authors and randomly select three novels per author, i. e. each sample consists of 75 texts by 25 authors. Each text is shortened to 30,000 tokens, skipping the first 10%.

Experiment 2b: We select the 25 authors contributing the largest number of novels and draw 5,000 random samples of 75 texts (three per author). Each text is shortened to 30,000 tokens, skipping the first 10%.

The results are evaluated via a stratified three-fold cross-validation scheme as described in section 3.1.

4. Results

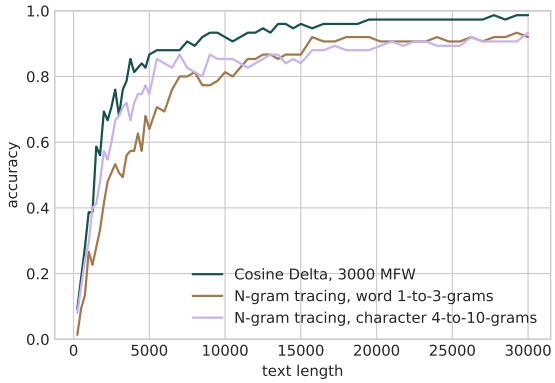
4.1. Shortening Experiments

4.1.1. Experiment 1a

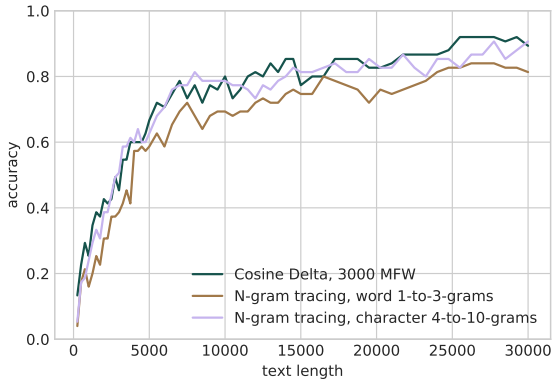
The results of experiment 1a, where we shorten all texts, are shown in Fig. 2. We display authorship attribution accuracy depending on text length for Cosine Delta and N-gram

[1] Wright (2017) describes a similar authorship attribution method based on word n-grams which uses the Jaccard coefficient instead. The Jaccard coefficient normalizes the number of n-grams that occur both in the disputed text and in the author-specific parts of the comparison corpus by the total number of distinct n-grams in those texts.

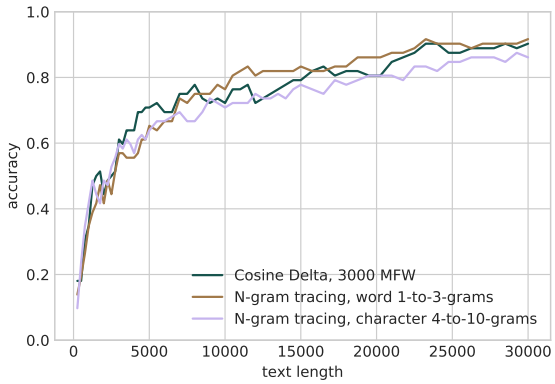
[2] https://github.com/cophi-wue/refcor



(a) German



(b) English



(c) French

Figure 2: Results of experiment 1a (shorten all texts)

tracing, using the majority votes of word 1-to-3-grams and character 4-to-10-grams as described in Section 2.2.

Unurprisingly, the accuracy of all three methods improves with larger text sizes. All methods perform rather poorly for very short texts, where they have to attribute, for example, a 250 word fragment to one of 25 possible authors with only 500 words of comparison text per author.

On the German corpus, shown in the top panel, Delta is consistently better than both N-gram tracing methods. Its performance is relatively stable for longer texts but drops for fewer than 5,000 words. The two N-gram tracing methods perform roughly identically on longer texts but for shorter text lengths character n-grams work better than word n-grams. The performance of both variants stabilizes for text lengths greater than 7,000 words.

On the English corpus, shown in the middle panel, Delta and character n-grams consistently outperform word n-grams. The performance of all methods drops for text lengths smaller than 7,000 words.

On the French corpus, shown in the bottom panel, Delta and word n-grams perform better than character n-grams. Performance drops at 5,000–7,000 words.

All in all, Delta usually performs better or at least approximately as well as N-gram tracing, and it is not entirely clear if word n-grams or character n-grams are better for the latter. Another observation that stands out is that in general the performance on the English and French corpora is notably worse than on the German corpus.

4.1.2. Experiment 1b

The results of experiment 1b, where we shorten only the disputed text, are shown in Fig. 3.

As was to be expected, the results for shorter text lengths are much better than in experiment 1a due to the much larger comparison corpus. In this scenario, N-gram tracing always outperforms Delta on very short texts by a large margin, achieving approximately 50% accuracy on 250-word fragments (while guessing would only achieve 4% accuracy).

On the German corpus, shown in the top panel, Delta is the best method for longer texts and N-gram tracing for shorter texts with less than 3,000 words. For both N-gram tracing methods, 1,000 tokens are sufficient for achieving more than 80% accuracy.

On the English corpus, shown in the middle panel, Delta and character n-grams are the best methods for longer texts, while both N-gram tracing methods are better than Delta for texts with less than 5,000 words. As in experiment 1a, performance is generally worse than on the German corpus and 3,000–4,000 words are needed for achieving 80% accuracy with N-gram tracing.

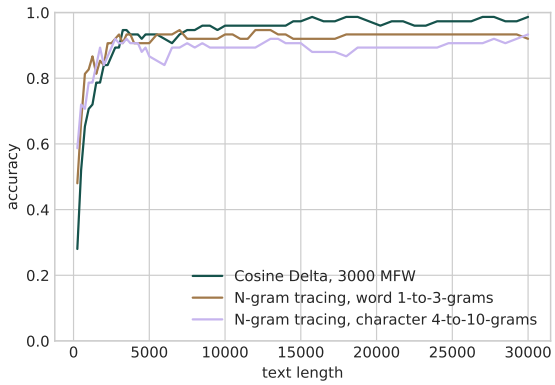
It is remarkable that on the French corpus, shown in the bottom panel, word n-grams are consistently the best method. Even though performance is not as good as on the German corpus, 1,000–2,000 words are sufficient for achieving more than 80% accuracy with word n-grams.

4.2. Sampling Experiments

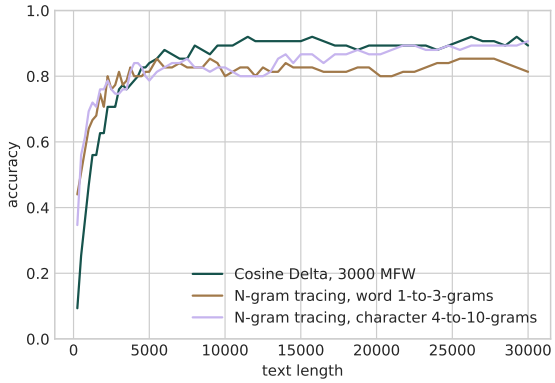
4.2.1. Experiment 2a

Fig. 4 shows the results of experiment 2a, where we draw 5,000 random samples of 25 German authors and shorten them to 30,000 words. For all methods, the central 50% of samples lie in a fairly narrow range of ± 5 percent points around the median (the colored boxes). However, for the remaining 50% there is considerable random variation: Classification accuracies lie between 80% and 100% for Cosine Delta and word n-grams and between 70% and 100% for character n-grams.

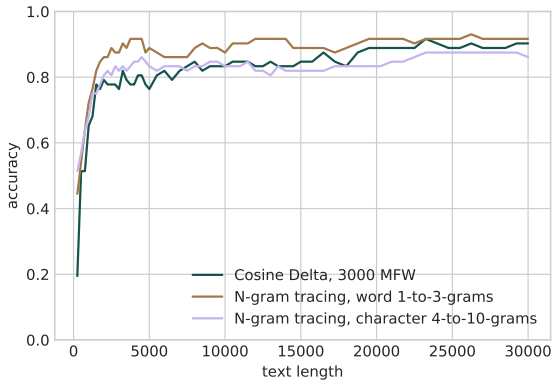
Cosine Delta seems to be a bit better than word n-grams and character n-grams seem to perform notably worse. But we cannot tell from Fig. 4 whether one method is usually better than the other on the same data because the differences between individual samples are much larger than the differences between methods. To this end, we computed pairwise accuracy differences between the three methods for



(a) German



(b) English



(c) French

Figure 3: Results of experiment 1b (shorten disputed text)

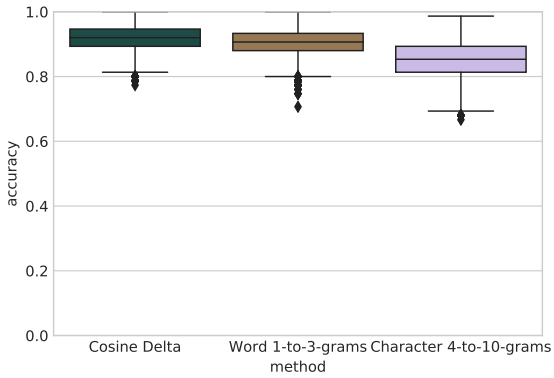


Figure 4: Boxplots with the results of experiment 2a (5,000 sets of 25 German authors)

each of the 5,000 samples. Their distribution is visualized in Fig. 5, showing that Cosine Delta in fact outperforms word n-grams for roughly 75% of the samples. Both Cosine Delta and word n-grams are almost always better than character n-grams.

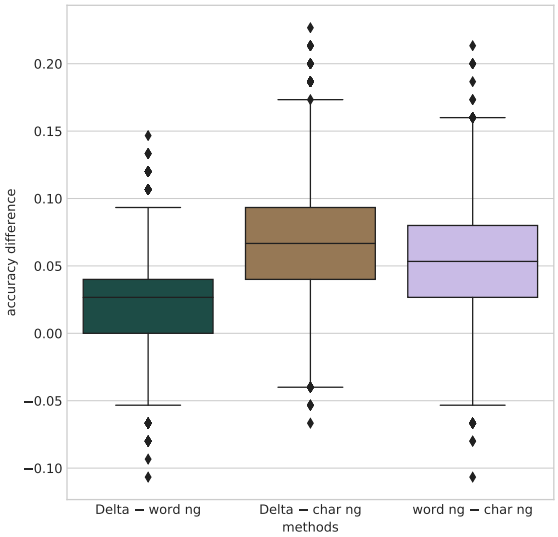


Figure 5: Boxplots with the accuracy differences between the methods in experiment 2a (larger values indicate that the first of the two methods is better)

4.2.2. Experiment 2b

The results of experiment 2b in Fig. 6 show that, as we would expect, sampling texts by the same set of authors results in somewhat less variability. However, the amount of variability is still surprising: Classification accuracies can easily fluctuate by 15 percent points. As before, Cosine Delta outperforms word n-grams and word n-grams outperform character n-grams.

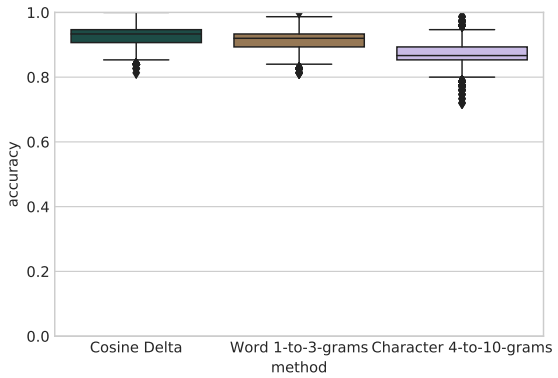


Figure 6: Boxplots with the results of experiment 2b (5,000 sets of texts from the same 25 German authors)

The pairwise differences between the methods are visualized in Fig. 7 and show the same pattern as for experiment 2a: Cosine Delta outperforms word n-grams for roughly 75% of the samples and both Cosine Delta and word n-grams are almost always better than character n-grams.

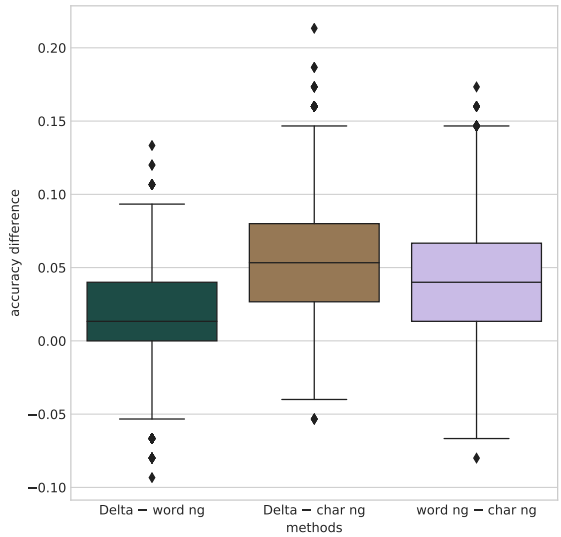


Figure 7: Boxplots with the accuracy differences between the methods in experiment 2b (larger values indicate that the first of the two methods is better)

4.2.3. Sampling vs. Shortening

The two sampling experiments showed that for both Cosine Delta and N-Gram tracing the accuracy of the authorship attribution depends to a certain extent on corpus composition. This raises the question of how meaningful the differences between the three corpora that we observed in the shortening experiments really are.

To address this question, we repeated experiment 1a on the 5,000 random samples of German authors drawn for experiment 2a. In Fig. 8, we show the results for Cosine Delta on the three corpora from experiment 1a. The grey area represents the range in which Cosine Delta lies in 95% of the 5,000 random samples.

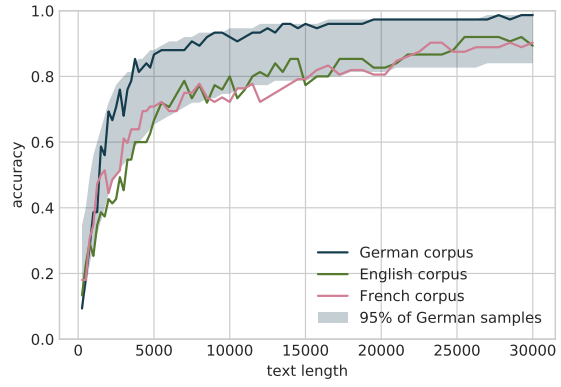


Figure 8: Accuracy of Cosine Delta in experiment 1a and on the 5,000 German samples from experiment 2a

As we can see, our German corpus happens to be a particularly easy sample. Fig. 8 also suggests that the differences observed in the shortening experiments may well be due to the selection of authors and texts.

5. Conclusion

For settings where the comparison corpus, i. e. the available amount of textual material written by the candidate authors, is large in comparison with the disputed text, we can confirm Grieve et al.'s (submitted) claim that N-gram tracing is more reliable than Delta for short texts up to 2,500–3,000. For longer texts, Delta is superior. In cases where both the disputed text and the comparison corpus are relatively small, neither Delta nor N-gram tracing yield reliable results. At least in our setting with a set of 25 possible authors, N-gram tracing requires text lengths of 1,000–3,000 words and a large enough comparison corpus to achieve an acceptable accuracy of 80%.

We also observed considerable performance differences of roughly ten percent points between English, French and German. It is tempting to blame those differences on typological differences between the languages and to speculate about the features that make the German language so well-suited for authorship attribution. However, as the sampling experiments show, the performance of the attribution methods varies considerably with corpus composition. Therefore, only future research comparing the spread of the measures based on many samples across languages will be able to answer the question whether the variance between languages is mainly a result of the corpus setup or whether there is also a factor in play related to language typology.

An interesting and somewhat worrying finding is that even with long texts the composition of the comparison corpus (i. e. the selection of authors and texts) has a large and unpredictable impact on the accuracy of the authorship attribution, which can easily fluctuate by as much as 20 percent points for all the methods tested. This aspect has so far been neglected and should both be kept in mind when interpreting previous results and be taken into account for future studies on authorship attribution.

The obvious next step would be to take the short analysis in Section 4.2.3. to the next level and to run shortening experiments on a large number of samples drawn from large collections of texts in many languages. Such experiments could in a reliable way shed light on the question whether the performance of authorship attribution methods varies between languages.

6. Bibliographical References

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

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

Eder, M. (2013a). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*.

Eder, M. (2013b). Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing*, 28(4):603–614.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., and Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl_2):ii4–ii16.

Grieve, J., Carmody, E., Clarke, I., Gideon, H., Heini, A., Nini, A., and Waibel, E. (submitted). Attributing the

Bixby Letter using n-gram tracing. *Digital Scholarship in the Humanities*. Submitted on May 26, 2017.

Jannidis, F., Pielström, S., Schöch, C., and Vitt, T. (2015). Improving Burrows' Delta – an empirical evaluation of text distance measures. In *Digital Humanities 2015: Conference Abstracts*.

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

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.

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):315–321.

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

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Wright, D. (2017). Using word n-grams to identify authors and idiolects. A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics*, 22(2):212–241.