# Short texts with fewer authors. Revisiting the boundaries of stylometry

**Rebora, Simone**
simone.rebora@uni-mainz.de
Johannes Gutenberg University Mainz, Germany

Extensive work has been dedicated to identifying the minimum text length for effective authorship attribution (Eder, 2013; Eder, 2017). However, reccommendations provided so far are based on artificial setups which do not always mirror the actual conditions of the research. In fact, the application of stylometric methods is often preceded by archival and philological research, which can help reduce the candidate authors to a few (Dimino et al., 2021), if not even just two (Rebora et al., 2019). This reduction simplifies the task for the stylometric analysis (Rybicki, 2015) and can help decreasing that threshold of 2,000 words—or, more cautiously, 5,000 words—, which was set by analyzing corpora composed by up to 21 authors.

This contribution therefore aims at integrating the guidelines provided so far by including also the number of candidate authors. Overall, it could be considered as another brick in the construction of an overarching set of guidelines, which can be completed only via the collaboration and integration of multiple proposals.

To give an interlinguistic scope to the research, analyses were performed on four different corpora of literary texts, composed by selections of novels in English (Computational Stylistics Group, 2022a), German (Computational Stylistics Group, 2022b), French (Schöch and Burnard, 2021, version "plain1"), and Italian (Ciotti et al., 2022, version "Orig"). All scripts have been shared through a dedicated *GitHub* repository ( https://github.com/SimoneRebora/stylometry_text_length) and are mainly built upon the pre-compiled functions of the *stylo* R package (Eder et al., 2016, version 0.7.4). Overall, the scripts take the texts from each corpus to randomly create experimental setups (by combining different lengths of text chunks and different numbers of candidates) and analyze them via different approaches (e.g., varying distance measures and most frequent words). To counter the effects of random selection, analyses are repeated multiple times for each configuration (20 times in this specific case). Produced score is a simple proportion of correct attributions. For more technical details, see full documentation in the *GitHub* repository.

A first overview of the results is provided by Figure 1, which shows how, for two candidate authors, a text of 1,000 words produces already about 75-80% correct attributions. It should also be noted how the use of character n-grams (4-grams in Figure 1) produces a better efficiency only for very short texts.

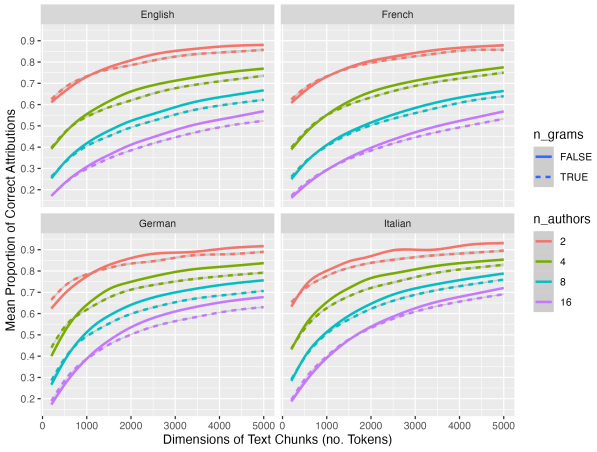


Figure 1. Results overview

Cosine Delta (Smith and Aldridge, 2011—called "dist.wurzburg" in the *stylo* implementation) is the best distance measure, independently from the language (Figure 2). When visualizing the results just for this measure, efficiency improves substantially, even moving beyond 90% when working with two authors and 2,000 words (Figure 3).

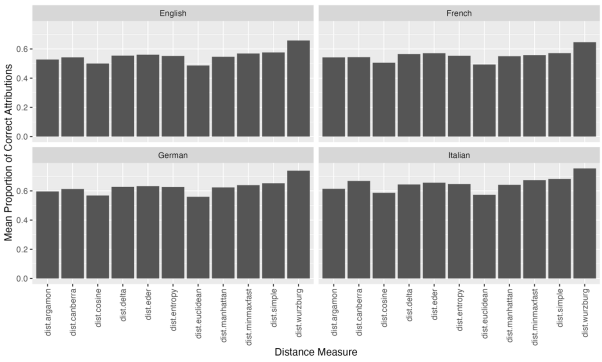


Figure 2. Results overview (focus on distance measure)

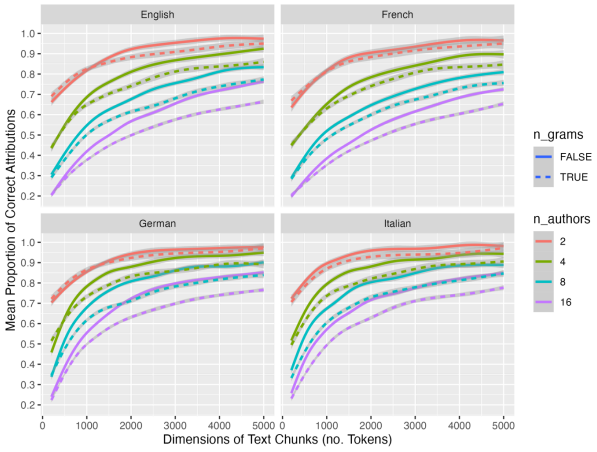


Figure 3. Results overview (limited to Cosine Delta distance)

Such results are in line with what already shown by Evert et al. (2017). They also confirm how efficiency does not just correlate positively with the number of most frequent words (MFW, see Figure 4). Closer inspection of the results for the German corpus shows in fact how positive correlation becomes evident only when considering longer texts (Figure 5) and higher numbers of candidates (Figure 6): a trend that is confirmed also for the other corpora. While this phenomenon calls for further investigation, a possible explanation can be in the fact that high selections of MFW include much more semantic information, which could become misleading when working with small experimental setups (Rebora, 2022).

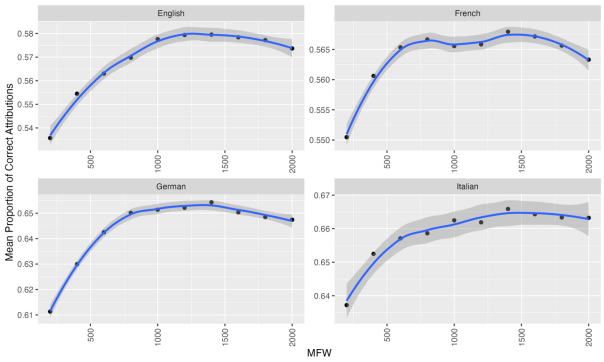


Figure 4. Results overview (limited to Word analysis and focus on MFW)

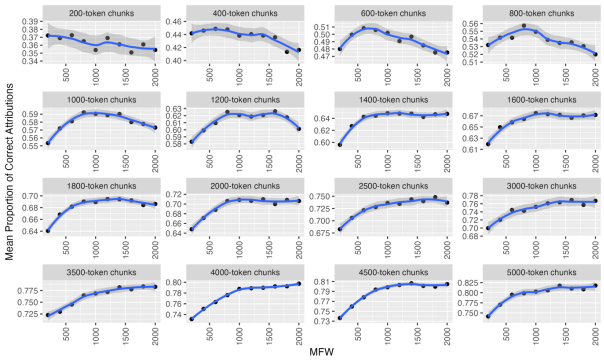


Figure 5. Results overview (limited to word analysis and German corpus, focus on MFW and text chunk dimensions)

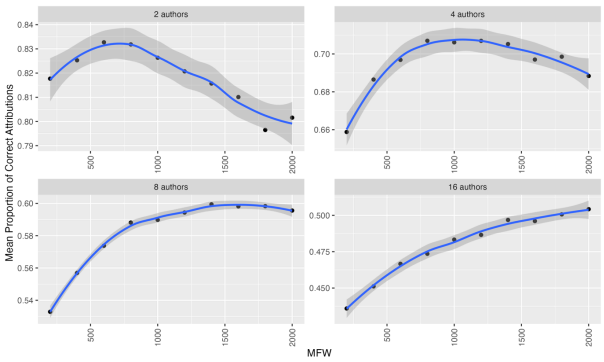


Figure 6. Results overview (limited to word analysis and German corpus, focus on MFW and number of authors)

Provisional outcome of this research is an overview of estimated efficiency scores (synthetically presented in Table 1 and fully available in the *GitHub* repository), which may be taken as a reference point by researchers when starting a new authorship attribution project. Still, the results presented here cannot be considered as final: in fact, outcomes might vary when considering different corpora (in different languages, but also including other genres and epochs) and using alternative methods. While further research is indeed needed, we hope to have shown how it should aim at better fitting the requirements of real case scenarios, by designing evaluation setups that imitate them as closely as possible.

| Text dimensions | Efficiency | Standard deviation |
|---|---|---|
| **200** | 0.655 | 0.271 |
| **400** | 0.698 | 0.229 |
| **600** | 0.769 | 0.205 |
| **800** | 0.781 | 0.217 |
| **1000** | 0.828 | 0.189 |
| **1200** | 0.833 | 0.194 |
| **1400** | 0.868 | 0.156 |
| **1600** | 0.901 | 0.147 |
| **1800** | 0.897 | 0.154 |
| **2000** | 0.938 | 0.115 |
| **2500** | 0.933 | 0.123 |
| **3000** | 0.955 | 0.103 |
| **3500** | 0.964 | 0.083 |
| **4000** | 0.974 | 0.067 |
| **4500** | 0.982 | 0.060 |
| **5000** | 0.971 | 0.088 |

Table 1. Overview of efficiency scores (sample for Cosine Delta distance, 200-2,000 MFW, two candidate authors, and English corpus)

# Bibliography

**Ciotti, F., Schöch, C. and Burnard, L.** (2022). ELTeC-ita European Literary Text Collection (ELTeC) https://github.com/COST-ELTeC/ELTeC-ita (accessed 31 October 2022).

**Computational Stylistics Group** (2022a). 100 English Novels ver. 1.4 https://github.com/computationalstylistics/100_english_novels (accessed 31 October 2022).

**Computational Stylistics Group** (2022b). 68 German Novels https://github.com/computationalstylistics/68_german_novels (accessed 31 October 2022).

**Dimino, M., Rebora, S. and Salgaro, M.** (2021). Between Austrian war propaganda and literary history. A stylometric analysis of Heimat. *EADH2021* https://eadh2021.culintec.de/REBORA_Simone_Between_Austrian_war_propaganda_and_literary_h.html.

**Eder, M.** (2013). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities* , 30 (2): 167–82 doi:10.1093/llc/fqt066.

**Eder, M.** (2017). Short samples in authorship attribution: a new approach. *DH2017 Book of Abstracts* . pp. 221–24.

**Eder, M., Rybicki, J. and Kestemont, M.** (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal* , 8 (1): 107–21.

**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): ii4–16 doi:10.1093/llc/fqx023.

**Rebora, S.** (2022). Stylometry and Reader Response. An Experiment with Harry Potter Fanfiction. In Ciracì, F., Miglietta, G. and Gatto, C. (eds), *AIUCD 2022 - Proceedings* . Bologna: AIUCD, pp. 30–34 http://amsacta.unibo.it/6848/ (accessed 19 September 2022).

**Rebora, S., Herrmann, J. B., Lauer, G. and Salgaro, M.** (2019). Robert Musil, a war journal, and stylometry: Tackling the issue of short texts in authorship attribution. *Digital Scholarship in the Humanities* , 34 (3): 582–605 doi:10.1093/llc/fqy055.

**Rybicki, J.** (2015). Success rates in most-frequent-word-based authorship attribution. a case study of 1000 polish novels from I gnacy K rasicki to J erzy P ilch. *Studies in Polish Linguistics* , 10 (2): 87104.

**Schöch, C. and Burnard, L.** (2021). French Novel Corpus (ELTeC-fra): April 2021 release Zenodo doi:10.5281/ZENODO.4662433. https://zenodo.org/record/4662433 (accessed 31 October 2022).

**Smith, P. W. H. and Aldridge, W.** (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method. *Journal of Quantitative Linguistics* , 18 (1): 63–88 doi:10.1080/09296174.2011.533591.