# Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?

**Rybicki, Jan**
jkrybicki@gmail.com
Pedagogical University, Krakow, Poland

**Eder, Maciej**
maciej_eder@poczta.onet.pl
Pedagogical University, Krakow, Poland

In 2007, John Burrows identified three regions in word frequency lists of corpora in authorship attribution and stylometry. The first of these regions consists of the most frequent words, for which his Delta has become the best-known method of study. This is evidenced by a varied body of research with interesting modifications of the method (e.g. Argamon 2008; Hoover 2004, 2004a). At the other end of the frequency list, Iota deals with the lowest-frequency words, while "the large area between the extremes of ubiquity and rarity" (Burrows, 2007) is now the target of many studies employing Zeta (e.g. Craig, Kinney, 2009; Hoover, 2007).

Due to the popularity of the three methods it was only a matter of time before Delta (and, to a lesser extent, Zeta and Iota) were applied to texts in languages other than Modern English: Middle Dutch (Dalen-Oskam, Zundert, 2007), Old English (Garcìa, Martìn 2007) and Polish (Eder, Rybicki 2009). Delta has also been used in translation-oriented papers, including Burrowsís own work on Juvenal (Burrows, 2002) and Rybicki's attempts at translator attribution (2009).

It has been generally - and mainly empirically - assumed that the use of methods relying on the most frequent words in a corpus should work just as well in other languages as it did in English; this question was approached in any detail only very recently (Juola, 2009). We could not fail to observe that its success rates in Polish, although still high, fell somewhat short of its guessing rate in English (Rybicki 2009a). Also, the already-quoted study by Rybicki (2009) seemed to suggest that, in a corpus of translated literary texts, Delta was much better at recognising the author of the original than the translator. This justified a more in-depth look at the workings of Burrowsís method both in its "original" English and in a variety of other languages.

## 1. Methods

In this study, a single major modification has been applied to the usual Delta process. Each analysis was made for the top 50-5000 most frequent words in the corpus - but then the 50 most frequent words would be omitted and the next 50-5000 words taken for analysis; then the first 100 most frequent words would be omitted, and so on. This was done with a single R script written by Eder; the script produced word frequency tables, calculated Delta and produced "heatmap" graphs of Delta's success rate for each of the frequency list intervals, showing the best combinations of initial word position in wordlist and size of window, including variations of pronoun deletion and culling parameters. Thus, in the resulting heatmap graphs, the horizontal axis presents the size of each wordlist used for one set of Delta calculations; the vertical axis shows how many of the most frequent words were omitted. Each of the runs of the script produced an average of ca. 3000 Delta iterations.

## 2. Material

The project included the following corpora (used separately); each contained a similar number of texts to be attributed.

| Code | Language | Texts | Attribution |
|---|---|---|---|
| E1 | English | 65 novels from Swift to Conrad | Author |
| E2 | English | 32 epic poems from Milton to Tennyson | Author |
| E3 | English | 35 translations of Sienkiewicz's novels | Translator |
| P1 | Polish | 69 19th- and early 20th-century novels from Kraszewski to Øeromski | Author |
| P2 | Polish | 95 translations of 19th- and 20th-century novels from Balzac to Eco | Author |
| P3 | Polish | 95 translations of 19th- and 20th-century novels from Balzac to Eco | Translator |
| F1 | French | 71 19th- and 20th-century novels from Voltaire to Gide | Author |
| L1 | Latin | 94 prose texts from Cicero to Gellius | Author |
| L2 | Latin | 28 hexameter poems from Lucretius to Jacopo Sannazaro | Author |
| G1 | German | 66 literary texts from Goethe to Thomas Mann | Author |
| H1 | Hungarian | 64 novels from Kemèny to Bródy | Author |
| I1 | Italian | 77 novels from Manzoni to D'Annunzio | Author |

## 3. Results

The English novel corpus (E1, Fig. 1) was the one with the best attributions for all available sample sizes starting at the top of the reference corpus word frequency list; it was equally easy to attribute even if the first 2000 most frequent words were omitted in the analysis - or even the first 3000 for longer samples. The English epic poems (E2, Fig. 2) had their area of best attributive success removed away from the top of the word frequency list, into the 1000th-2000th most-frequent-word region. Some successful attributions could also be made with a variety of wordlists around the 2000 mark, starting at the 1st most frequent word.

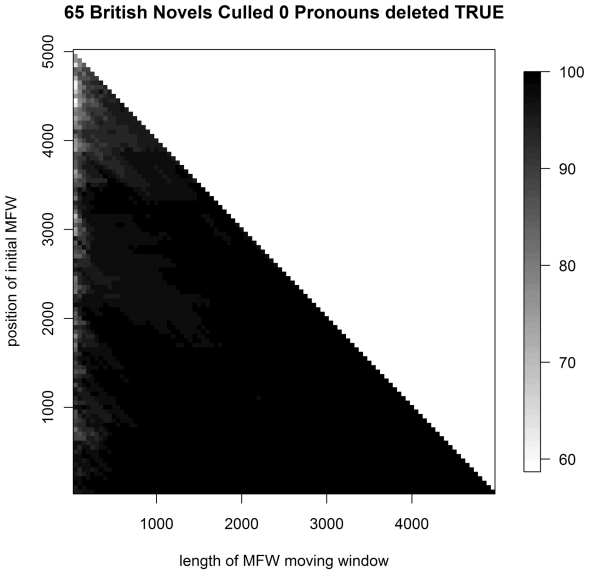


Figure 1. Heatmap of 65 English novels (percentage of correct attributions). Colour coding is from low (white) to high (black)

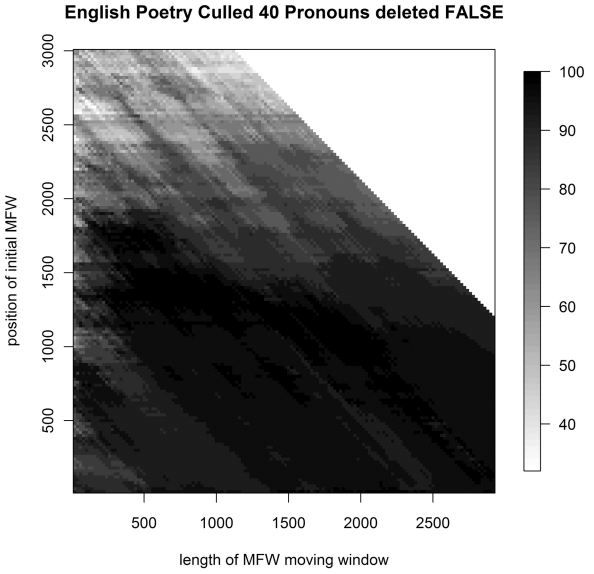


Figure 2. Heatmap of 32 English epic poems

The final "specialist" corpus in the English section of the project - 32 works by Polish novelist Henryk Sienkiewicz, translated into English by a number of translators (Fig. 3) - showed Delta's expected problems in translator attribution; however, for a variety of culling/ pronoun deletion parameters, a small yet fairly consistent hotspot would appear for small samples if the first 2000-3000 words were deleted from the frequency wordlist. The first Polish corpus, that of 69 19th- and early 20th-century classic Polish novels (P1, Fig. 4), showed marked improvement in Delta success rate when the wordlist taken for attribution started at some

450 words down the frequency list; the most successful sample sizes were relatively small: no more than 1200 words long.

When the corpus of Polish translations was studied for original authorship (P2, Fig. 5), the results were quite accurate for many sample sizes up to 1800 from the very top of the frequency list. Delta was equally successful for samples of up to 1400 words beyond the 800th most-frequent-word mark. The same corpus yielded lower attribution success when studied for translator recognition (P3, Fig. 6). In fact, it resembled somewhat the graph for Polish classics: a small range of passable attributions, usually for samples below 1000, and usually better when starting a hundred or so words down the frequency rank list.

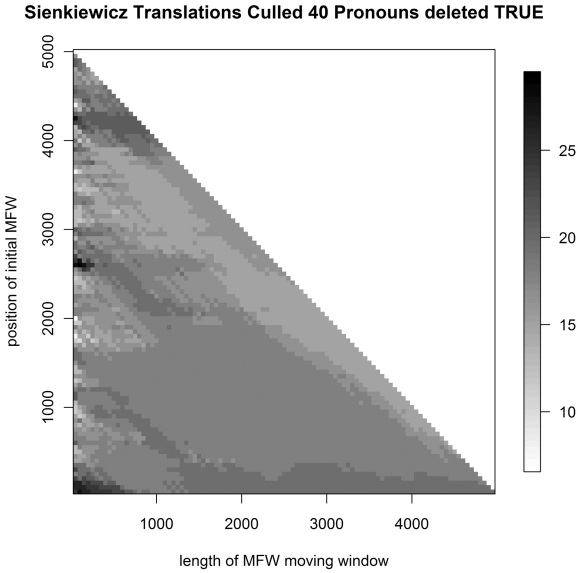


Figure 3. Heatmap of 35 English translations of Sienkiewicz's works

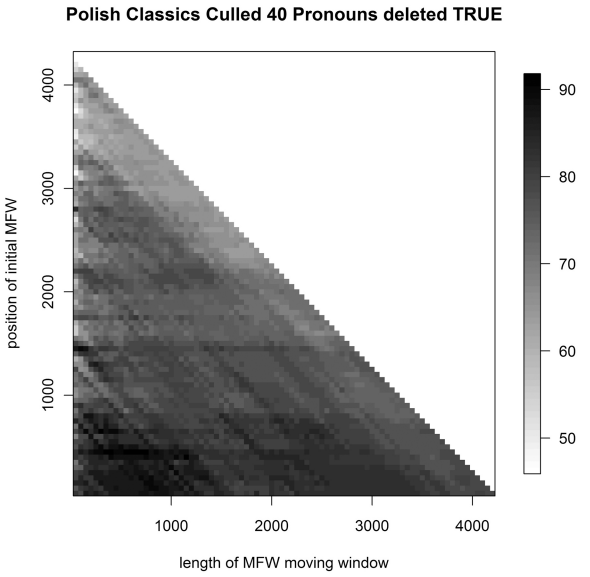


Figure 4. Heatmap of 69 Polish novel classics

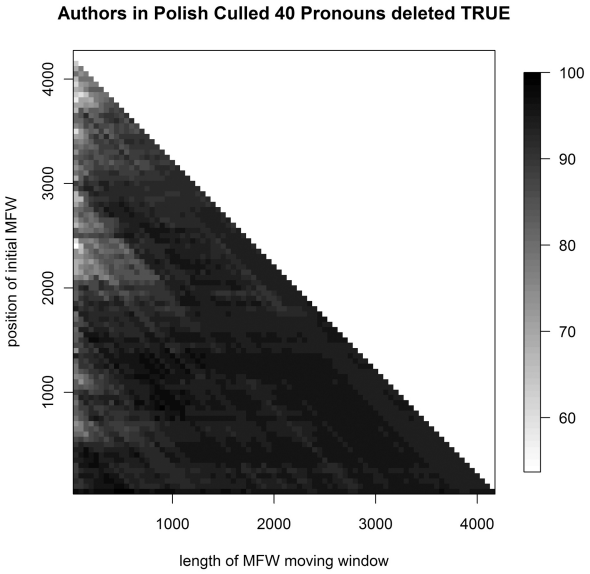


Figure 5. Heatmap of 95 Polish translations from Balzac to Eco (autorship attribution)

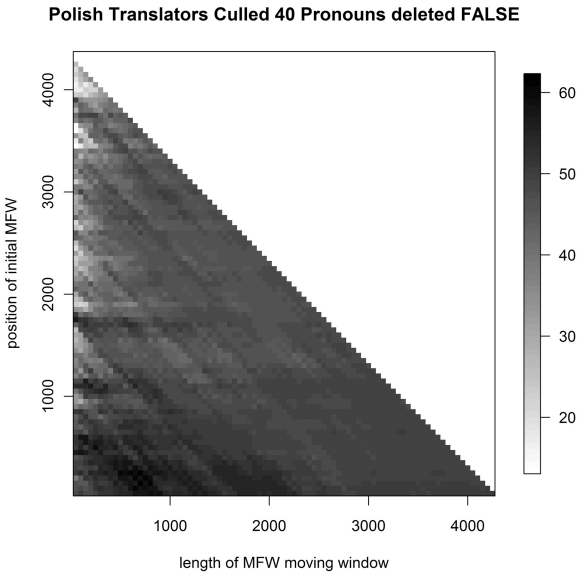


Figure 6. Heatmap of 95 Polish translations from Balzac to Eco (translator attribution)

The French corpus proved almost equally difficult (F1, Fig. 7): Delta was very successful mainly for small-sized samples from the top of the overall frequency wordlist. In contrast, the graph for the German corpus (G1, Fig. 8) presented a success rate akin to that for the English novels, with a consistently high correct attribution in most of the studied spectrum of sample size and for samples beginning anywhere between the 1st and the 1000th word in the corpus frequency list.

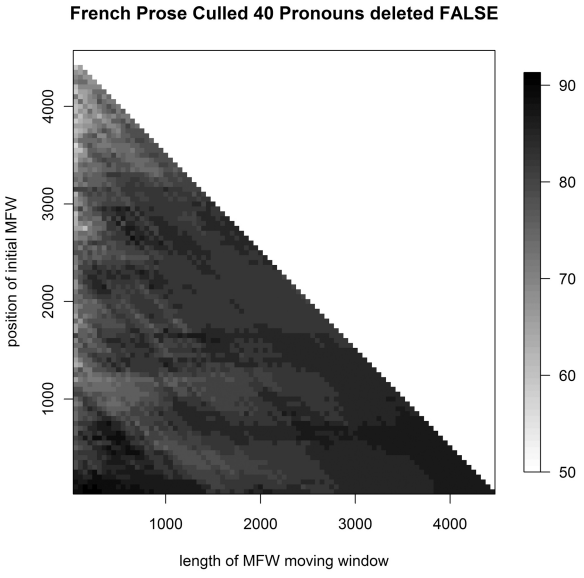


Figure 7. Heatmap of 71 French novels

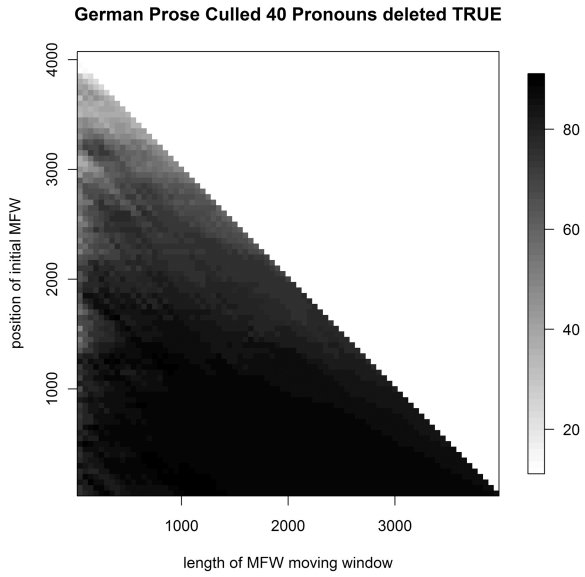


Figure 8. Heatmap of 66 German texts

Of the two Latin corpora, the prose texts (L1, Fig. 9) could serve as excellent evidence for a minimalist approach in authorship attribution based on most frequent words, as the best (if not perfect) results were obtained by staying close to the axis intersection point: no more than 750 words, taken no further than from the 50th place on the frequency rank list.

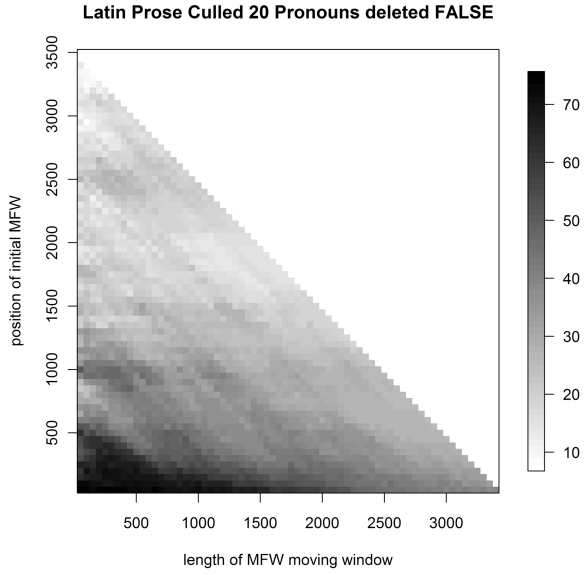


Figure 9. Heatmap of 94 Latin prose texts

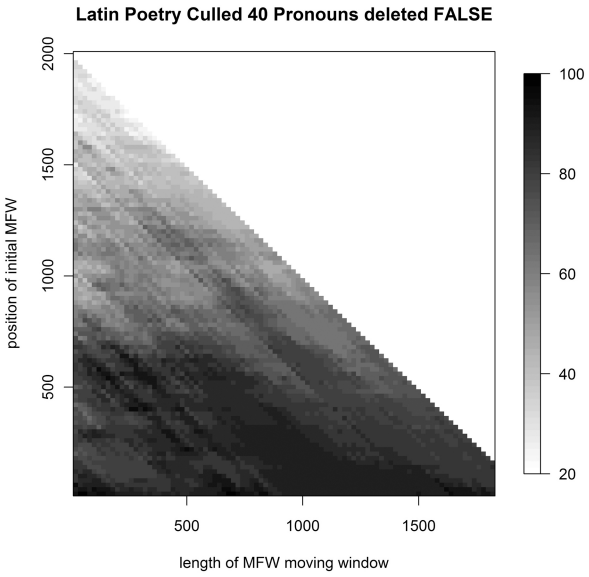


Figure 10. Heatmap of 28 Latin hexameter poems

The other Latin corpus, that of hexameter poetry (L2, Fig. 10), paints a much more heterogeneous picture: Delta was only successful for top words from the frequency list at rare small (150), medium (700) and large (1700) window sizes, and for a few isolated places around the 1000/1000 intersection point in the graph.

The corpus of 19th-century Hungarian novels (H1, Fig. 11) exhibited good success for much of the studied spectrum and an interesting hotspot of short samples at ca. 4000 words from the top of the word frequency list.

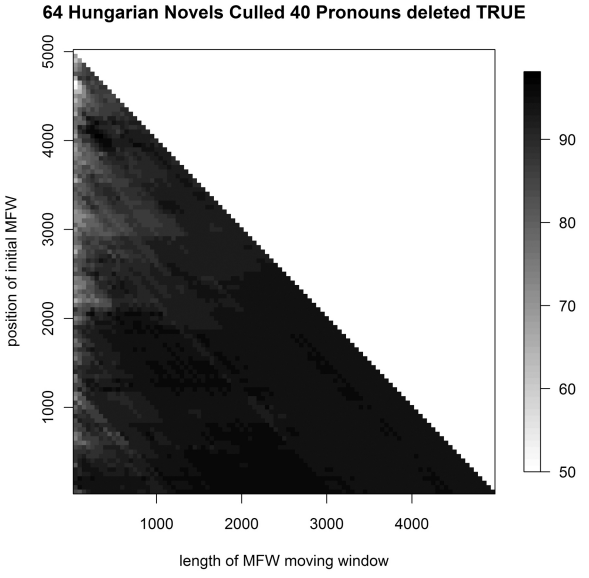


Figure 11. Heatmap of 64 Hungarian novels

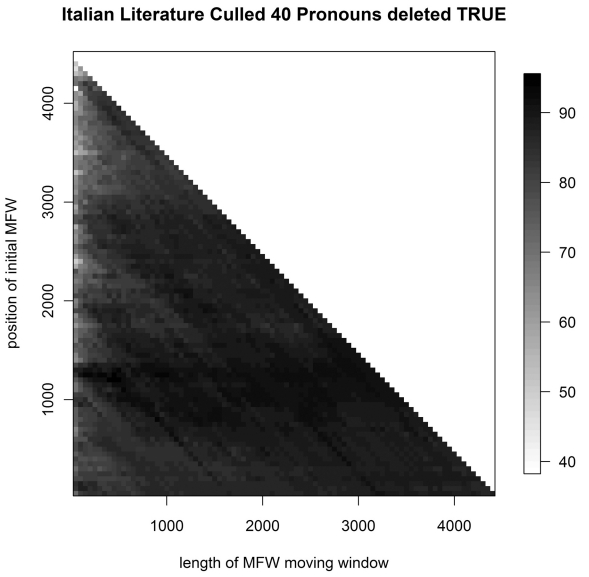


Figure 12. Heatmap of 77 Italian novels

With the Italian novels (I, Fig. 12), Delta was at its best for a broad variety of sample sizes, but only when some 1000 most frequent words were eliminated from the reference corpus.

## 4. Conclusions

1. Standard Delta (i.e. applied to the most frequent words) provides the best results for authorial attribution in English and German prose.
2. The same procedures still yield acceptable results in other languages and in translator attribution. The success here can be improved by manipulating the number of words taken for analysis and by selecting the reference wordlists at various distances from the top of the overall frequency rank list.
3. The differences in attributive success could be partially explained by the differences in the degree of inflection/agglutination of the languages studied, the strongest evidence of this being the relatively highest success rate in English and German.

## References

**Argamon, S.** (2008). 'Interpreting Burrows's Delta: Geometric and Probabilistic Foundations'. *Literary and Linguistic Computing*. **23(2)**: 131-147.

**Burrows, J.F.** (1987). *Computation into Criticism: a Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

**Burrows, J.F.** (2007). 'All the Way Through: Testing for Authorship in Different Frequency Strata'. *Literary and Linguistic Computing*. **22(1)**: 27-48.

**Burrows, J.F.** (2002). 'The Englishing of Juvenal: Computational Stylistics and Translated Texts'. *Style*. **36**: 677-99.

**Burrows, J.F.** (2002a). 'Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship'. *Literary and Linguistic Computing*. **17**: 267-287.

**Hoover, D.L.** (2003). 'Frequent Collocations and Authorial Style'. *Literary and Linguistic Computing*. **18**: 261-286.

**Hoover, D.L.** (2004). 'Testing Burrows's Delta'. *Literary and Linguistic Computing*. **19**: 453-475.

**Hoover, D.L.** (2004a). 'Delta Prime?'. *Literary and Linguistic Computing*. **19**: 477-495.

**Hoover, D.L.** (2007). 'Corpus Stylistics, Stylometry, and the Styles of Henry James'. *Style*. **41(2)**: 174-203.

**Rybicki, J.** (2009). 'Translation and Delta Revisited: When We Read Translations, Is It the Author or the Translator that We Really Read?'. *Digital Humanities*. College Park, 2009.

**Rybicki, J.** (2009a). 'Liczenie krasnoludków. Trochę inaczej o polskich przekładach trylogii Tolkiena'. *Po co ludziom krasnoludki?*. Warszawa, 2009.

**Craig, H., Kinney, A.F. (eds.)** (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

**Dalen-Oskam, K. van, Zundert, J. van** (2007). 'Delta for Middle Dutch–Author and Copyist Distinction in Walewein'. *Literary and Linguistic Computing*. **22**: 345-362.

**Eder, M., Rybicki, J.** (2009). 'PCA, Delta, JGAAP and Polish Poetry of the 16th and the 17th Centuries: Who Wrote the Dirty Stuff?'. *Digital Humanities*. College Park, 2009.

**Garcìa, A.M., Martìn, J.C.** (2007). 'Function Words in Authorship Attribution Studies'. *Literary and Linguistic Computing*. **22**: 49-66.

**Jockers, M.L., Witten, D.M., Criddle, C.S.** (2008). 'Reassessing Authorship in the Book of Mormon Using Delta and Nearest Shrunken Centroid Classification'. *Literary and Linguistic Computing*. **22**: 465-491.

**Mosteller, F., Wallace, D.L.** (1964) (2007). *Inference and Disputed Authorship: The Federalist*. CSLI Publications.

**Notes**

1. Reprinted with a new introduction by John Nerbonne