Deeper Delta across genres and languages: do we really need the most frequent words?

Jan Rybicki and Maciej Eder
Pedagogical University, Kraków, Poland

Abstract

This article examines the success of authorship attribution of Burrows's Delta in several corpora representing a variety of languages and genres. Contrary to the approaches of our predecessors, who only investigated the attributive effectiveness of the very top of the list of the most frequent words, hundreds of possible combinations of word vectors were tested in this study, not solely starting with the most frequent word in each corpus. The results show that Delta works best for prose in English and German and less well for agglutinative languages such as Polish or Latin.

Correspondence:
Jan Rybicki
Institute of Modern Languages,
Pedagogical University,
31-128 Kraków, Poland
Ul. Karmelicka 41.
Email:
jkrybicki@gmail.com

1 Introduction

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 (Burrows, 2002b). This is evidenced by a varied body of research with interesting modifications of the method (e.g. Hoover, 2004a, b; Argamon, 2008). 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 or its modifications, such as Craig's Zeta (e.g. Hoover, 2007; Craig and Kinney, 2009).

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 and Zundert, 2007), Old English (García and Martín, 2007), and Polish (Eder and Rybicki, 2009). Delta has also been used in translation-oriented articles, including Burrows's own work on Juvenal (Burrows, 2002a) and Rybicki's attempts at translator attribution (2009, 2011).

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 does in English; this question has not been approached in any detail until very recently (Juola, 2009). We cannot fail to observe that its success rates in Polish, although still high, fall somewhat short of its detection rate in English (Rybicki, 2009; Eder and Rybicki, 2011). Also, to further complicate the issue of multilingualism, the study by Rybicki mentioned above (2009) seems to suggest that, in a corpus of translated literary texts, Delta is much better at recognizing the author of the original than the translator. Or, to be more precise: with only two candidate translators of the same author, Delta fares well; however, at higher numbers of translators and of authors of the original, Delta's guessing favors the latter rather than the former. Additionally, genre differences between texts have often been blamed for worse (or better) results in authorship attribution by Delta. This was yet another good reason for a more in-depth look into the workings of Burrows's method not only in

Literary and Linguistic Computing, Vol. 26, No. 3, 2011.

doi:10.1093/llc/fqr031 Advance Access published on 14 July 2011

its 'original' English and in a variety of other languages, but also in a variety of genres.

2 Methods

The software we used provides several flavours of Delta (as well as other distance measures); however, the one consistently used in the final results of this study was Burrows's classic Delta, for the reason that it was the classic method and, perhaps more importantly, because tentative results obtained with the other Delta varieties were very similar.

In this study, a single major modification was applied to the usual Delta process. According to the standard Delta procedure, each corpus was divided into 'training' samples in a primary set (one representative sample per each author) and the remaining 'test' samples in a secondary set. The goal of such a procedure was to test how many samples of known authorship were 'guessed', or correctly classified to the proper 'training' sample.

Each analysis was first made with the top fifty most frequent words in the corpus; then the fifty most frequent words would be omitted and the next fifty words (i.e. words ranked 51–100 in the descending word frequency list) would be taken for analysis; then the next 50 most frequent words (those ranked 101–150), and so on until the required limit (usually the 5,000th most frequent word) would be reached. Then the procedure would restart with the first 100 words (1–100), the second 100 words (101–100), and so on. At every subsequent restart, the number of the words omitted from the top of the frequency list would be increased by fifty until the length of this 'moving window' descending down the word frequency list reached another limit (usually 5,000). This was done with a single 1,000-line script, written by Eder, for the statistical programming environment R.[1] The script produced word frequency tables, calculated the myriad Delta iterations 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. In fact, the heatmaps are probably the only feasible way of presenting such an amount of results in a comprehensive way.[2] In the resulting graphs below, the horizontal axis presents the size of each wordlist used for one set of Delta calculations (the 'moving window'); the vertical axis shows how many of the most frequent words were omitted (or where the 'moving window' began for each iteration). Each of the runs of the script produced an average of approximately 3,000 Delta iterations.

3 Material

The texts that constitute the corpora used in this study were taken from a variety of good-quality Internet sources (mostly, various national electronic libraries), cleaned of paraphernalia (such as extra titles or Project Gutenberg's legal disclaimers), and saved as Unicode text files; at this point, human editing ceased and the script took over to split the strings into words and perform the entire analysis.

The project included the following corpora (used separately).

Code	Language	Texts	Attribution
E1	English	Sixty-five novels from Swift to Conrad	Author
E2	English	Thrity-two epic poems from Milton to Tennyson	Author
P1	Polish	Sixty-nine 19th- and early 20th-century novels from Kraszewski to Żeromski	Author
F1	French	Seventy-one 19th- and 20th-century novels from Voltaire to Gide	Author
L1	Latin	Ninety-four prose texts from Cicero to Gellius	Author
L2	Latin	Twenty-eight hexameter poems from Lucretius to Jacopo Sannazaro	Author
G1	German	Sixty-six literary texts from Goethe to Thomas Mann	Author
H1	Hungarian	Sixty-four novels from Kemény to Bródy	Author
I1	Italian	Seventy-seven novels from Manzoni to D'Annunzio	Author
S1	English	Forty-two works by Shakespeare	Genre

4 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 2,000 most frequent words were omitted in the analysis—or even the first 3,000 for longer samples. This was also the only corpus where a perfect attributive score (100%) was achieved almost constantly, which is reflected, in the graph, by the widespread and smooth dark color in the heatmap. The English epic poems (E2, Fig. 2), on the other hand, while displaying a 100% accuracy in some 'pockets,' attributed in general significantly worse than the English novels. For less frequent words, i.e. below the 2,000th on the frequency list, the guessing effectiveness began decreasing very quickly; the area of best attributive success was removed from the top of the word frequency list, into the 1,000–2,000 most frequent word region.

The Polish corpus of Sixty-nine 19th- and early 20th-century classic Polish novels (P1, Fig. 3) showed marked improvement in Delta attribution rate when the wordlist started at some 450 words down the frequency list; the most successful sample sizes were relatively small: no more than 1,200 words long.

The French corpus proved difficult to interpret because there was no clear smooth area of good

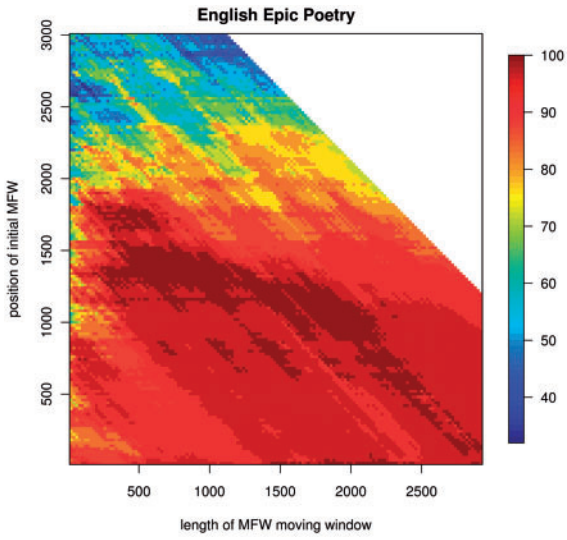

Fig. 2 Attribution accuracy for 32 English epic poems

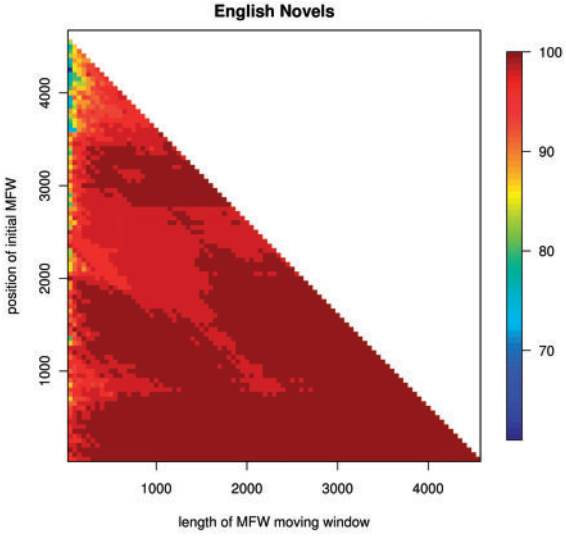

Fig. 1 Attribution accuracy for 65 English novels (percentage of correct attributions). Color coding is indicated by the sidebar scale

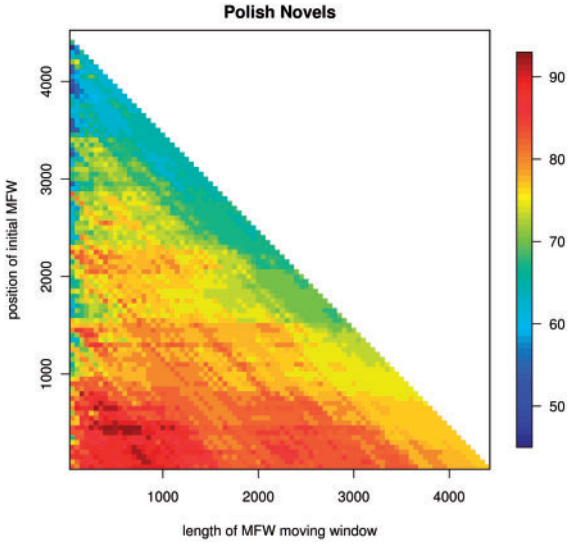

Fig. 3 Attribution accuracy for 69 Polish novel classics

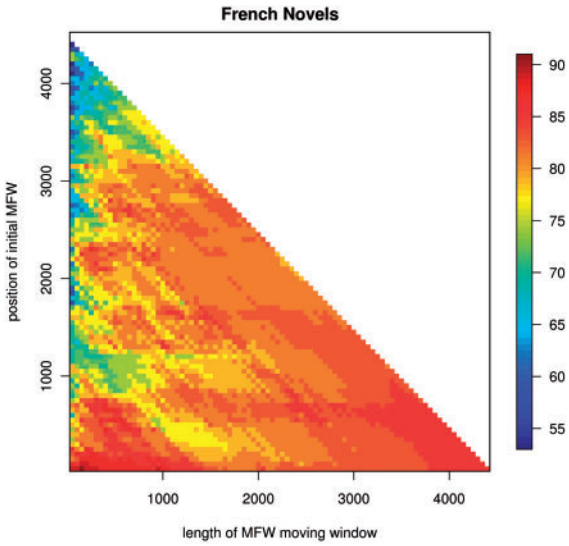

Fig. 4 Attribution accuracy for 71 French novels

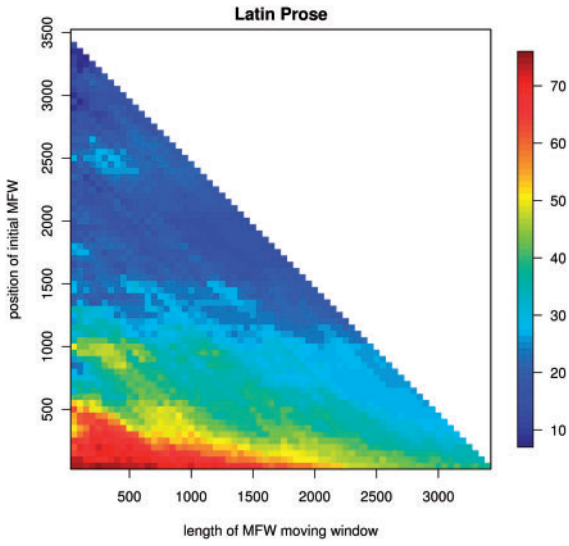

Fig. 6 Attribution accuracy for 94 Latin prose texts

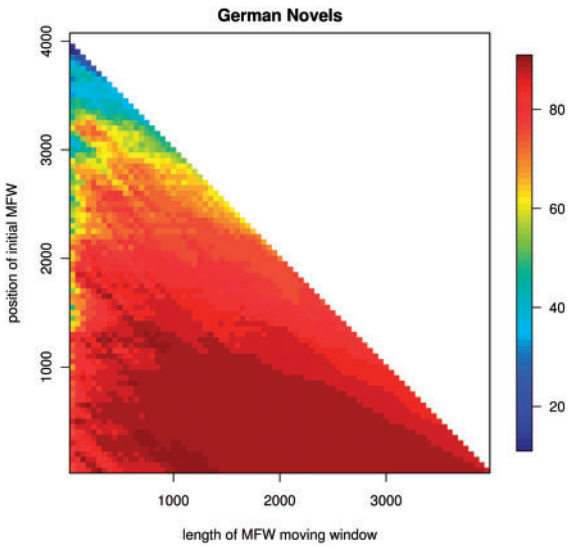

Fig. 5 Attribution accuracy for 66 German prose texts

accuracy (F1, Fig. 4): Delta was very successful mainly for small-sized pockets from the top of the overall frequency wordlist. In contrast, the graph for the German corpus (G1, Fig. 5) 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 1,000th word in the corpus frequency list. The best attribution was achieved in a narrow region around 1,000 MFWs from the top of the list.

Of the two Latin corpora, the prose texts (L1, Fig. 6) 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 the 50th place on the frequency rank list. The top score, 75%, was in fact achieved only once—at 250 MFWs from the top of the list.

The other Latin corpus, that of hexameter poetry (L2, Fig. 7), 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 (1,700) window sizes and for a few isolated places around the 500/500 intersection point in the graph. Again, the best score of 75% is represented by two pockets at 110 and 120 MFWs counting from the top of the list.

The corpus of 19th-century Hungarian novels (H1, Fig. 8) exhibited good success for much of

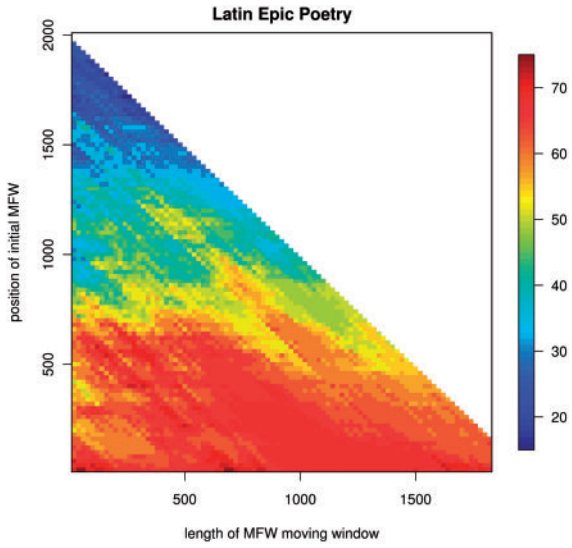

Fig. 7 Attribution accuracy for 28 Latin hexameter poems

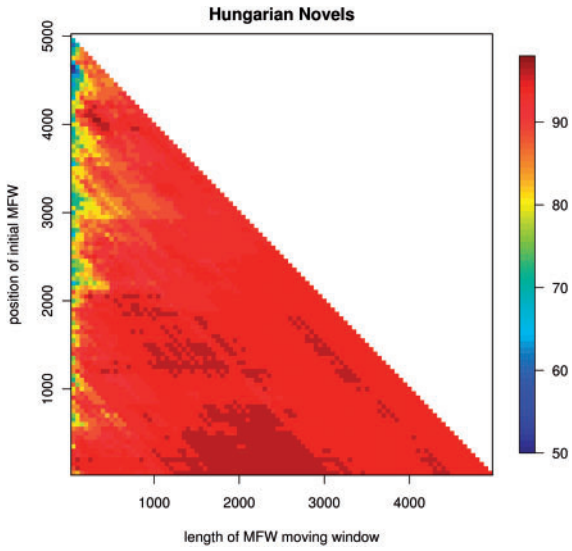

Fig. 8 Attribution accuracy for 64 Hungarian novels

the studied spectrum and an interesting hotspot of short samples at approximately 4,000 words from the top of the word frequency list. What was even more interesting, the hotspot was surrounded by an area of a very weak attributive success.

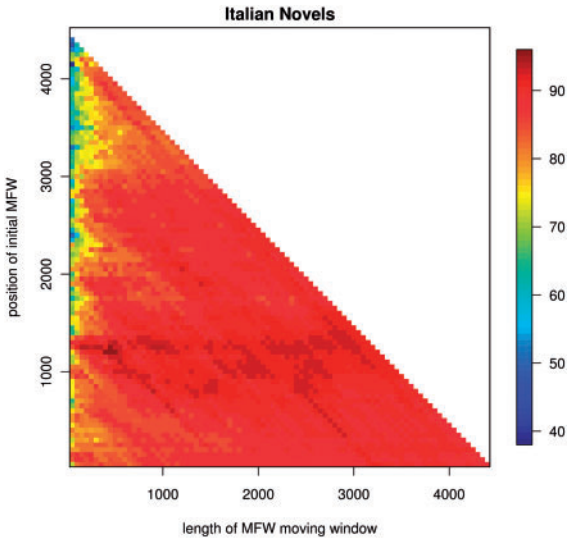

Fig. 9 Attribution accuracy for 77 Italian novels

With the Italian novels (I1, Fig. 9), 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. The top Italian score, 76%, appeared only a few times for wordlists of 400, 450, and 500 words starting at the 350th and the 400th most frequent word.

The final corpus used in this series of analyses was that of 42 works by Shakespeare (S1, Fig. 10). It was also the single case where Delta was tested for genre recognition—the works were categorized as poems, tragedies, comedies, romances, or histories. And while the overall reliability was poor, there is a smallish yet visible darker region in Fig. 10 for pockets of some 2,500 most frequent words starting at the top, or near the top, of the word frequency list.

5 Conclusions

The graphs presented above seem to confirm the suspicions that, while Delta is still the most successful method of authorship attribution based on word frequencies, its success is not independent of the language of the texts studied. This has not been noticed so far for the simple reason that Delta

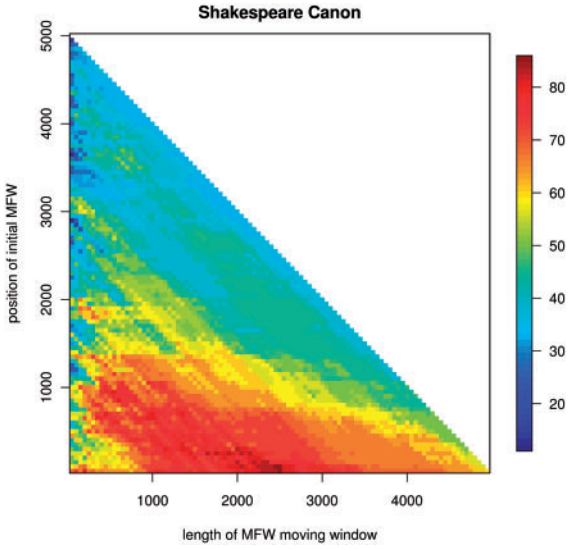

Fig. 10 Accuracy in genre recognition for 42 works by Shakespeare

studies have been done, in a great majority, on English-language prose. Yet even the switch from prose to poetry within the language of Dickens and Milton has consequences for the best-attribution region—perhaps for the simple reason that poetic texts (even those brought together in E2, a corpus of epic poetry, i.e. works of some length) provide less adequate statistics than material gathered from full novels.

Thus Delta's high success for prose texts, in general, is a positive and optimistic result of this series of experiments; less cause for optimism—and less uniformity—can be seen in Delta's behaviour in prose texts in other languages. Its high and consistent attributions throughout the frequency regions studied for the sixty-six German novels allows a hypothesis that Germanic languages might provide the best material for authorial attribution, and that their shared characteristics can be thanked for this. The relatively poorer results for Latin and Polish—both highly inflected in comparison with English and German—suggests the degree of inflection as a possible factor. This would make sense in that the top strata of word frequency lists for languages with low inflection contain more uniform words, especially function words; as a result, the most frequent words in languages such as English are relatively more frequent than the most frequent words in agglutinative languages such as Latin.

While diagrams for most other languages in this series of experiments seem, at the very least, not to disprove this working hypothesis, a severe blow to its simple elegance has been dealt by the Hungarian corpus, i.e. a collection of texts in a language generally deemed the most inflected one of those under study here. To make matters worse, Delta's success in this unlikely collection of texts was even more remarkable due to their relative similarity as representatives of the same trend in 19th-century Hungarian fiction. What seemed a difficult corpus in a difficult (i.e. highly agglutinative) language scored visibly better than the 'easier' corpora of Polish or Latin prose. At this point, it is worth mentioning that any statements on the relative ease and difficulty of corpora collected from various languages and literatures can be tentative at best and require further study.

The greatest methodological problem that this study shows as far as Delta is concerned is that, while 'pockets' of good attribution reliability can be found at a variety of parameters of culling, word-list length and/or number of the most frequent words omitted (or not) from the top of the frequency list, 'pockets' of similar size can be found nearby where attribution is anything but good. This study shows that obtaining near-perfect results for, say, the top 1,000 most frequent words does nothing to guarantee similar success for the top 2,000 words (with the possible exception of English or German corpora, where Delta's success has been shown to be more uniform than in the other languages studied). And that, while it might be a good idea to manipulate the above-mentioned parameters, it is not yet known *how* to manipulate them for a given corpus, language, genre, or attribution type. It seems so far that there is no 'best,' or 'most reliable,' or 'universal,' value for either the moving window or its initial position in the most frequent word lists. This is frustrating. And this calls for finding a way to even out the pockets of better and worse parameter combinations—to average them out and thus to eschew cherry-picking—possibly, with bootstrapping, as suggested by initial results of our recent

studies (Eder, 2011; Eder and Rybicki, 2011; Rybicki, 2011). But even more frustrating is the fact that we do not know why Delta in Hungarian performs oddly compared to English because, simply, no one knows why.

References

Argamon, S. (2008). Interpreting Burrows's Delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, **23**(2): 131–47.

Burrows, J. F. (2002a). The Englishing of Juvenal: computational stylistics and translated texts. *Style*, **36**(4): 677–99.

Burrows, J. F. (2002b). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Burrows, J. F. (2007). All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*, **22**(1): 27–48.

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

van Dalen-Oskam, K. and van Zundert, J. (2007). Delta for middle Dutch–author and copyist distinction in *Walewein*. *Literary and Linguistic Computing*, **22**(3): 345–62.

Eder, M. (2011). Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, **6** (in press).

Eder, M. and Rybicki, J. (2009). PCA, Delta, JGAAP and Polish poetry of the 16th and the 17th centuries: who wrote the dirty stuff? *Digital Humanities 2009: Conference Abstracts*. College Park: MD, pp. 242–44.

Eder, M. and Rybicki, J. (2011). Do birds of a feather really flock together, or how to choose test samples for authorship attribution. *Digital Humanities 2011: Conference Abstracts*. Stanford, CA, pp. 124–27.

García, A. M. and Martín, J. C. (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing*, **22**(1): 49–66.

Hoover, D. L. (2004a). Testing Burrows's Delta. *Literary and Linguistic Computing*, **19**(4): 453–75.

Hoover, D. L. (2004b). Delta prime?. *Literary and Linguistic Computing*, **19**(4): 477–95.

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 2009: Conference Abstracts*. College Park: MD, pp. 245–47.

Rybicki, J. (2011). Ślady żony tłumacza. Alma Cardell Curtin i Jeremiah Curtin. *Przekładaniec*, **24** (in press).

Notes

1 Since much of the testing of the script was done by one author's graduate students, the script included a simple Tcl/Tk GUI by Rybicki (for easier operation). Both authors wish to take this opportunity to thank the happy helpers: Barbara Bajak, Izabela Jakus, Magdalena Jamrych, Monika Jaworska, Agnieszka Jucha, Małgorzata Kozieł, Malwina Kuraś, Izabela Leoniak, Anna Mikulec, Monika Obrzut, Jakub Piasecki, Agnieszka Rybus, Alicja Usień, Katarzyna Szosta, Paulina Zegar and Agnieszka Zgoll.

2 Color versions of the heatmaps generated for this study can be found in the online version of this article.