The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation

Jan Rybicki

Institute of Modern Languages, Pedagogical University of Krakow, Poland

Introduction.

Stylometry, or the study of measurable features of (literary) style, such as sentence length, vocabulary richness and various frequencies (of words, word lengths, word forms, etc.), has been around at least since the middle of the 19th century, and has found numerous practical applications in authorship attribution research. They are usually based on the belief that there exist such conscious or unconscious elements of personal style that can help detect the true author of an anonymous text; that there exist stylistic fingerprints that can betray the plagiarist; that the oldest authorship disputes (St. Paul's epistles; Shakespeare's plays; Sęp Szarzyński's erotic poems) can be settled with more or less sophisticated statistical methods.

While specific issues remain largely unresolved (or, if closed once, they are sooner or later reopened), a variety of statistical approaches have been developed that allow, often with spectacular precision, to identify texts written by several authors based on a single example of each author's writing. In this machine-learning procedure, the traceable differences between texts in a corpus are first used to produce a set of rules – a classifier – for discriminating authorial "uniqueness." The second step is to use the trained classifier to assign other texts samples to the authorial classes established by the classifier; any disputed or anonymous sample will be assigned to one of the classes as well.

The texts for this procedure are divided into two groups: the primary (training) set and the secondary (test) set. The first set, a collection of single texts written by known authors, serves as a sub-corpus for finding the best classifier. The second set contains texts of the known authors, works by other authors, and anonymous (or disputed) texts. The better the classifier, the more samples from the test set are attributed correctly and the more reliable the attribution of the disputed texts (Eder & Rybicki 2011).

Some of the most successful attributive applications involve the use of frequencies of the most frequent words (MFWs) in the entire corpus as the classifier; since the most frequent words are often function words, this approach can be traced back to at least 1964, when Mosteller and Wallace performed their attribution of the Federalist papers. Multivariate analysis (such as Principal Components Analysis, Cluster Analysis, Multidimensional Scaling) is used to evaluate the distances, or differences, between the frequency data for each text; these data are usually normalized in some way, either as correlations of relative frequencies (i.e. relative to the size of the text in which they occur) or, as in the case of Burrows's Delta, as z-scores of the word frequencies. In most studies, the analysis produces graphs that simplify the multidimensional matrix of frequencies of each word in each text in the corpus to a two-dimensional map of distances between the text or (as in Cluster Analysis)

presents them in a tree diagram, where texts most similar to each other are placed on neighboring branches.

Burrows's Delta has been established in the last decade as perhaps the most widely-used of the above methods. As has been mentioned above, it normalizes the frequency of the most frequent words by using z-scores:

$$z(f_i(T)) = \frac{f_i(T) - \mu_i}{\sigma_i}$$

where $f_i(T)$ is the raw frequency of word i; μ_i is that word's mean frequency in the corpus; σ_i is its standard deviation. The z-scores for all words studied in all the texts considered are then compared; Burrows's Delta is "the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text" (Burrows, 2002), or, for two texts, T and T_1, and a set of n words,

$$\Delta(T, T_1) = \frac{1}{n}\sum_{i=1}^{n}\left|z\left(f(T)\right) - z(f(T_1))\right|$$

Based on the results of Delta, authorship is claimed for the author of that text in the primary set for which the Delta distance is the smallest from the disputed text.

Since its first application by Burrows, Delta has accumulated a number of modifications; the most notable include Argamon's simplification of the formula and its "geometric interpretation" (2008), and Hoover's variants (2004). Hoover has also noticed that certain manipulations of the most-frequent-word list can improve attribution. The omission of personal pronouns, for instance, often helps attribution in English corpora if they contain texts written in both first- and third-person narration; culling too-characteristic words for single texts can also improve precision in bigger corpora (Hoover 2004a).

The size of the most-frequent-word list itself is a matter of some controversy. Some scholars use a relatively small number of the most frequent words, usually the 30 to 150 from the top of the rank list (e.g. Burrows 2002a); others study how the effectiveness of attribution methods could be increased (or decreased) when the number of words analyzed is extended to hundreds or even thousands of words down the frequency list (Hoover 2004, 2004a, 2007; Eder & Rybicki 2009; Smith & Aldridge 2011). Further attempts are made by omitting the top of the frequency rank list, which might improve attribution in a variety of languages and genres (Rybicki & Eder 2011). The optimal size of the attributed texts themselves has been discussed, and better precision for texts of lengths exceeding 10,000 words has been shown (Eder 2010).

Criticism of the Delta method (recently summarized by Vickers 2011) is usually based on the fact – acknowledged, it is true, by one of its most enthusiastic users – that, while "simple and intuitively reasonable, like previous statistical authorship attribution techniques," it "lacks any compelling theoretical justification" (Hoover 2005). Indeed, it dangerously assumes mutual independence of word frequencies (Argamon 2008) and is helpless in cases when the real

author is not present among the suspected writers (Smith & Aldridge 2011); to quote Burrows himself, what Delta really shows is the "least unlikely" author rather than the most likely one (Burrows 2002). Most recently, the machine-learning procedure employed not only by Delta but also by many other statistical approaches to authorship attribution has been shown to be strongly dependent on the choice of the "exemplary" texts of each author that make up the primary set (Eder & Rybicki 2011).

It has also been noticed that Delta's precision in the recognition of English texts is not matched by that in other languages (Rybicki & Eder 2011). As the creation of Delta and many of the initial studies made with this method happened in an English-language environment, it has been intuitively and understandably assumed that Delta – and other word-frequency-based authorship attribution methods – should work in all languages alike. This optimism has been shared by most researchers in the field; the reasoning presented in a rare attempt at discussing this issue (Juola 2009) is sound and persuasive. This comes in some contrast to the fact that, as has been mentioned above, experience gathered over extensive corpora in a variety of languages shows that, while still reliable, results for some languages (Polish and Hungarian 19th-century realistic prose was tested extensively from this point of view) do not match the accuracy achieved for English and German texts of the same genre and literary period. At this point yet another caveat must be added: this type of data "prevents direct comparisons of accuracy" and, furthermore, "it is hard to imagine ways to establish that two authorship attribution tasks are 'comparably difficult' to enable such direct comparisons" (Juola 2009: 163).

And yet despite the above shortcomings and uncertainties, Delta (and similar measures) is more often right than wrong. In fact, its precision combined with its assumption of independence of word frequencies seems to raise an interesting linguistic question that goes well beyond the practicalities of authorship attribution: why mere word frequencies are very often enough to differentiate between authors? While this question cannot be resolved satisfactorily here, it is interesting to see if this non-traditional method of authorship attribution is equally successful in recognizing authors in translation – or if translators' traces obliterate authors' individual use of the most frequent words – or if multivariate analysis of MFWs can tell translator from translator. This will be presented over a variety of translational corpora.

Method.

The version of authorship-attribution-oriented multivariate analysis used in this study employs z-scores according to the original Delta formula; these are then submitted to Cluster Analysis to produce tree diagrams for a given set of parameters, such as: number of MFWs studied; pronoun deletion; culling rate. The latter, expressed in percentages, specifies the number of texts in a corpus in which a given word must be found in order to be included in the analysis. Thus, a 100% culling rate limits the analysis to words that appear at least once in every text in the corpus; at a 50% culling rate, a word is included into the analysis when it appears in at least half of the texts in the corpus; a 0% culling rate (or no culling) means that no words are omitted. Then, these results, produced for a great variety of parameter

combinations, are used as input for a bootstrap procedure, similar to that employed by Dunn et al. in a study of Papuan languages (2005, quoted in Baayen 2008: 143-147):

> The basic idea of the bootstrap (…) is that we sample (with replacement) from the columns of our data matrix. For each sample, we construct the distance matrix and grow the corresponding unrooted tree with the node-joining algorithm. Finally, we compare our original dendrogram with the dendrograms for the bootstrap samples and calculate the proportions of bootstrap dendrograms that support the groupings in the original tree (Baayen 2008: 148).

In other words, a host of individual Cluster Analysis tree diagrams (or dendrograms) conduct a vote on the final configuration; the resulting bootstrap tree is a consensus between possibly different findings. It has been shown recently that while single Cluster Analysis diagrams can be misleading, a combination of a great many of them yields a much more reliable result. This approach is in fact an attempt at cashing in from the empirical fact stated above: that Delta is more often right than wrong (Eder & Rybicki 2011b).

The whole procedure was performed with a single script for the R statistical programming environment: the script processed the electronic texts to create a list of all the words used in all texts studied, with their frequencies in the individual texts, to create an initial input matrix of words (rows) by individual texts (columns), each cell containing a given word's frequency in a given text. The script then normalized the frequencies (using the R command "scale"); selected words from stated frequency ranges for analysis; performed the additional procedures (automatic deletion of personal pronouns and culling); compared the results for individual texts; performed the Delta calculations for each set of parameters; clustered the Delta similarities/distances obtained; finally, produced the above-mentioned bootstrap consensus trees (the entire procedure is presented in detail in Eder & Rybicki 2011b).

The validity of the method can be best evaluated in test runs, in which all authors are known. Figure 1 presents a corpus of 27 English novels by 11 authors from Sterne to Thackeray. As can be seen, works by the same authors have been correctly placed on the same "branches" of the dendrogram; what is more, some of the immediate-neighbor groups make sense in terms of traditional literary studies: above all, the common branch of the Brontë sisters, but also those of Dickens and Eliot, and of Richardson and Fielding. The range of the parameters used in this study – from 100 MFWs all the way to 5000, at culling values from 0% (no words are removed from the MFWs list) to 100% (frequencies are analyzed only for MFWs that appeared in all the texts) – shows that this bootstrap tree is a consensus between as many as 250 Cluster Analysis diagrams. Even if the number of authors and books is increased to, respectively, 19 and 65, the method still produces an attributively acceptable – if somewhat cluttered – diagram (Fig. 2).

Fig. 1. 27 novels by 11 authors in the English original

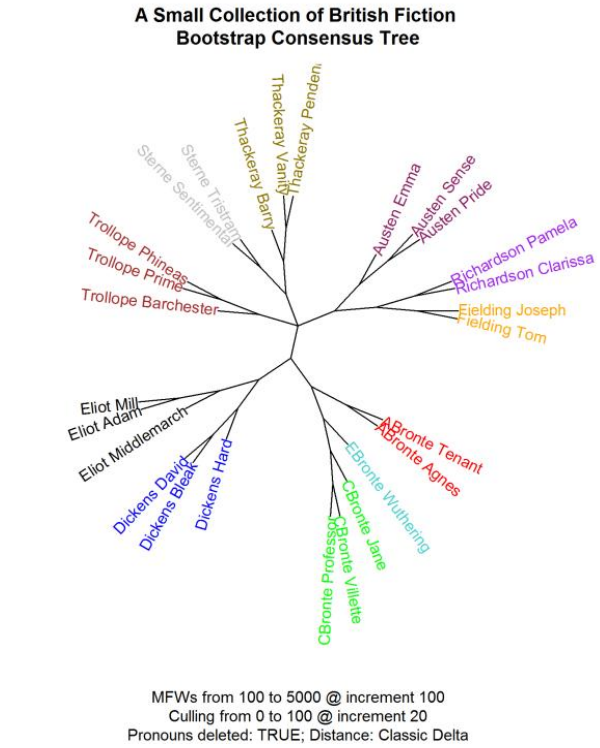
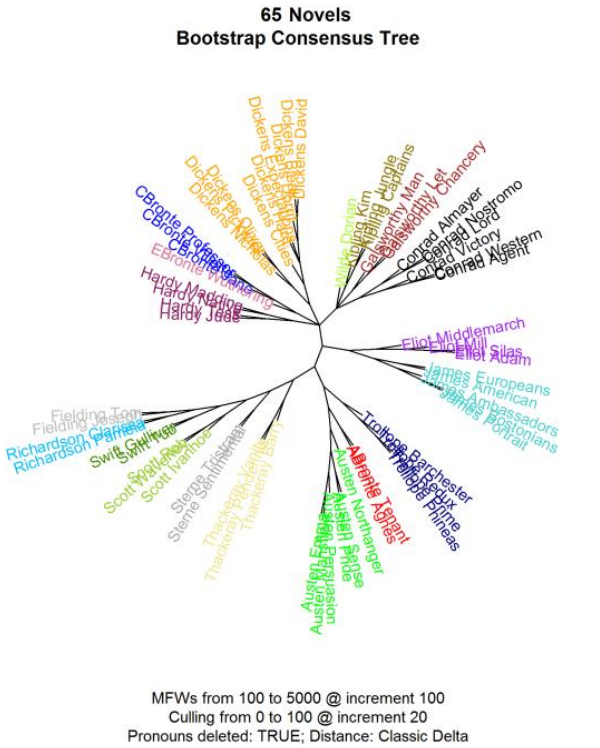


Fig. 2. 65 novels by 19 authors in the English original

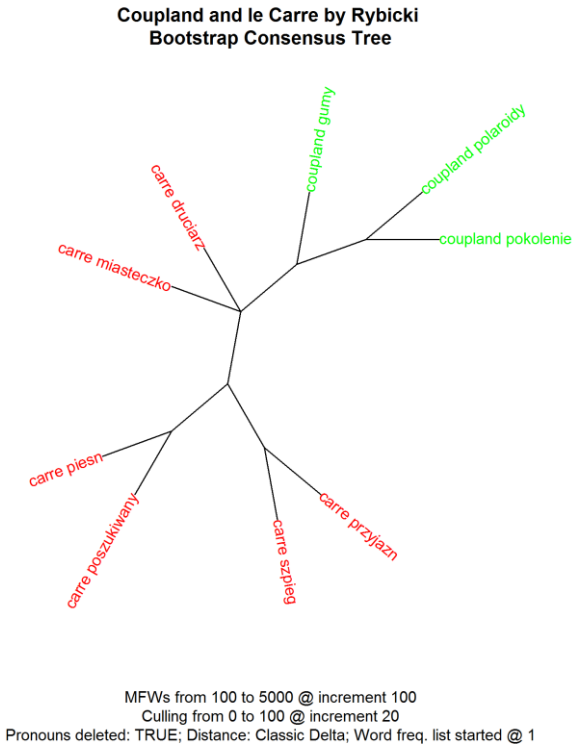


Results

Interpres, recense te ipsum: it is only fitting that my own English-to-Polish translations come first. Of the almost thirty novels I have translated since 1990, only two authors featured more

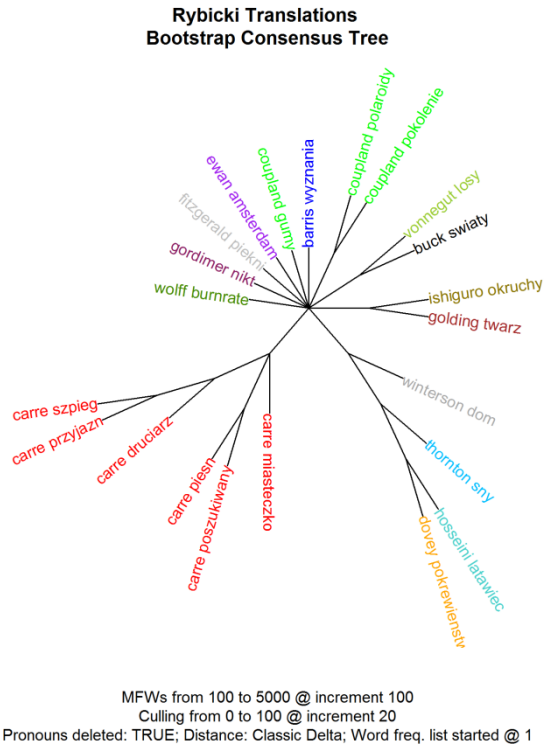
than once: the Canadian Douglas Coupland, represented in this corpus by my translations of three of his novels: *Generation X*, *Polaroids from the Dead* and *The Gum Thief*; and the Englishman John le Carré, with six novels: *A Perfect Spy*, *Absolute Friends*, *The Missions Song*, *Tinker Tailor Soldier Spy*, *A Most Wanted Man* and *A Small Town in Germany*. When these nine translations are subjected to testing with Delta, the result is quite representative of studies with two authors translated by a single translator: the diagram (Fig. 3) shows two fairly distinct groups for each.

Fig. 3. Rybicki's Polish translations of two authors



When my translations of novels by other authors are added to the corpus, the authorship attribution becomes both better and worse (Fig. 4): better, because all novels by le Carré are placed on their own cluster of branches, and worse, as the Coupland novels are not (although two of these, *Generation X* and *Polaroids from the Dead*, remain immediate neighbors).

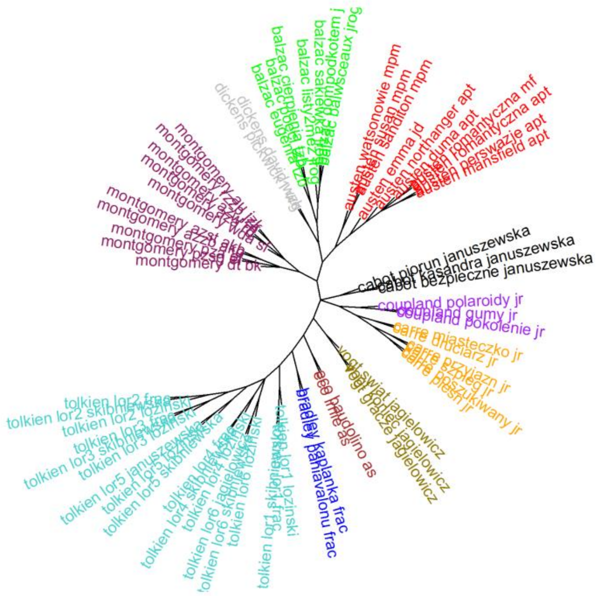
Fig. 4. Rybicki's Polish translations



When the corpus of Polish translations is expanded even further into works by other authors and translators, the pattern of correct authorial attributions becomes even more evident. In Fig. 5, the Polish translations of 65 novels by 11 authors (English, French and Italian) are distributed on neighboring branch clusters in a clear dependence on the author of the original. It is interesting to observe that works by individual authors cluster together whether or not each has been translated by the same translator; that, within some authorial clusters, some translator clusters can be observed (as in the Austen translations); that separate clusters of authors translated by the same translator occupy adjacent positions on the graph (the Coupland and le Carré translations by Rybicki); finally, that the three translations of the individual volumes of the same book series cluster by volume rather than by translator (for Tolkien in Polish).

Fig. 5. Polish translations by 20 translators of 65 novels by 11 authors

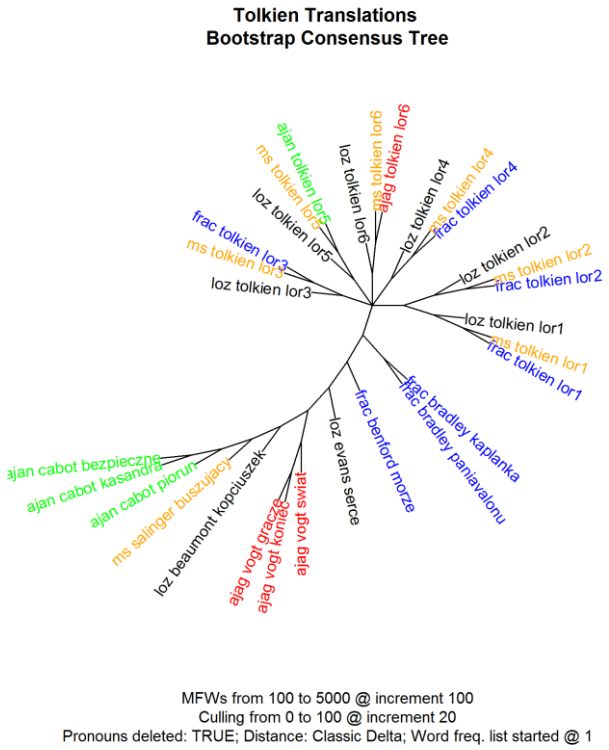
Polish Translations
Bootstrap Consensus Tree



MFWs from 100 to 5000 @ increment 100
Culling from 0 to 100 @ increment 20
Pronouns deleted: TRUE; Distance: Classic Delta; Word freq. list started @ 1

This last phenomenon is just as evident when the above corpus is limited to the three Polish translations of Tolkien and to other novels translated by translators involved therein (Fig. 6). The situation is quite complex here: the earliest translation by Skibniewska, made in the 1960s, has acquired two rivals in the 1990s: the controversial work by Łoziński and the joint effort by Cezary Frąc and Maria Frąc, responsible for the trilogy's first two books; Cezary Frąc also thoroughly edited the final one, initially translated by Aleksandra Jagiełowicz (first half) and Aleksandra Januszewska (second half).

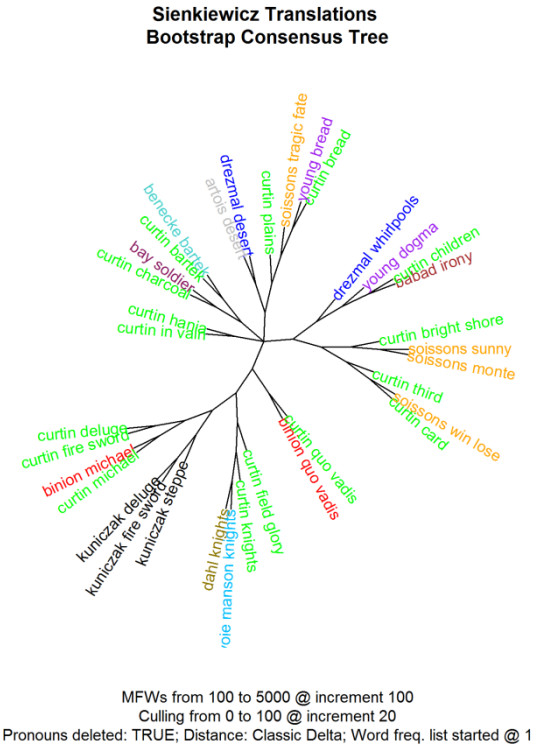
Fig. 6. 5 Polish translators of Tolkien novels and the same translators' other translations



A similar corpus of Polish-to-English translations of a single author is presented in Fig. 7; in fact, it also contains two different translations of a trilogy (with yet another of its books translated by a third translator). Here, however – as can be seen in the bottom left of the diagram – clustering by volume is limited to the Curtin and Binion translations, while the work of Kuniczak constitutes a separate cluster. This has not been entirely unexpected: the Kuniczak trilogy is the most extreme example – possibly of all translations in all corpora presented here – of an adaptative, modernized and explicative translation; in fact, it has been received by some critics as an adaptation rather than a translation (c. f. Segel 1991). Its length in tokens has been expanded at a ratio of 150-170% (the usual rate for Polish to English translation is 120-130%, Rybicki, 2010) by the translator's additions of explanatory passages (at times, much more than mere footnotes incorporated into the text), and that despite deleting extensive final chapters in two out of the three novels in the series. Other multiple translations of individual novels cluster together, even in the case of the three translations of *Krzyżacy*: Curtin's complete *Knights of the Cross* is similar to abridgements by Dahl, and by Savoie and Manson. The Sienkiewicz translations exhibit a fairly visible division into their author's specialty, historical romances (the bottom half of the diagram), as opposed to novels set in the

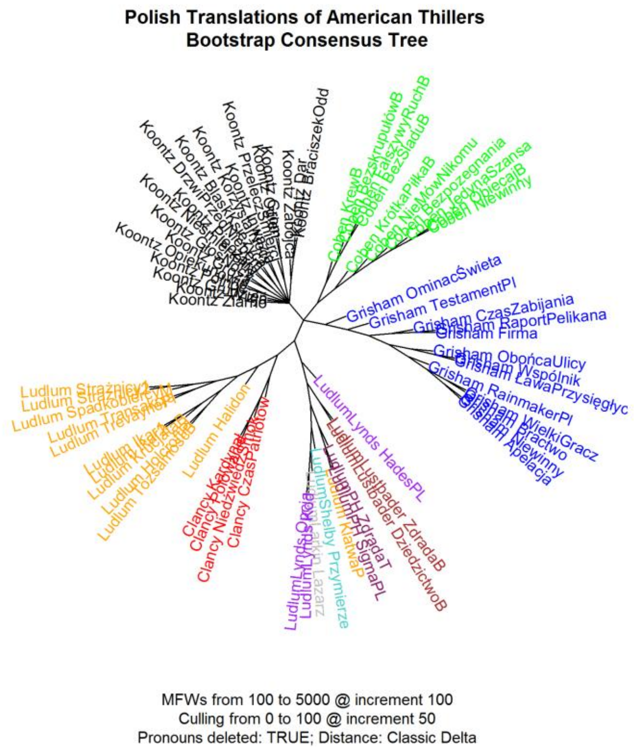
writer's own latter half of the 19th century (top). This is a reflection of a similar layout for Sienkiewicz's originals; in fact, the two genres usually refuse to appear separate in Delta diagrams for any Polish literary corpus that includes Poland's first Nobel Prize winner – and this is one of the symptoms of Delta's lesser accuracy in Polish (Rybicki & Eder 2011).

Fig. 7. English translations of Sienkiewicz



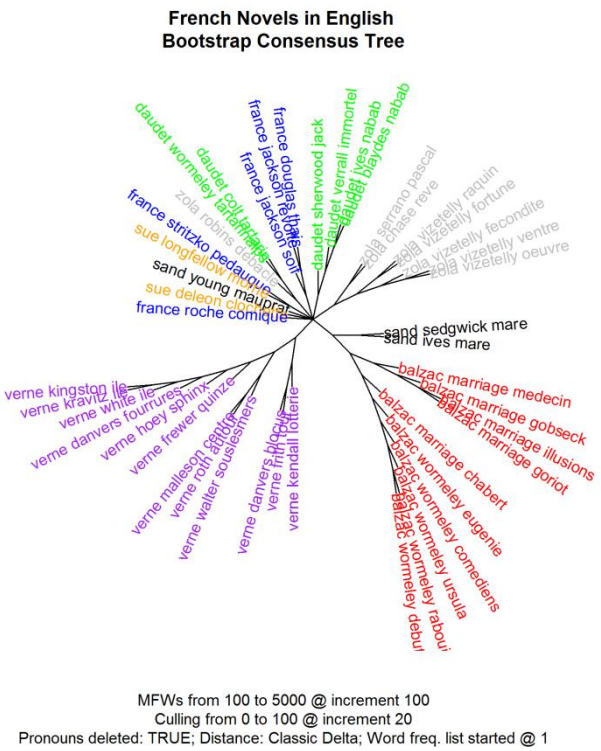
However, less complex cases seem to follow the rule of seeing through translation in authorial attribution. The large Polish translational corpus presented in Fig. 5 is all but mirrored in another corpus of comparable size, that of Polish translations of 70 thrillers by Clancy, Coben, Grisham, Koontz and Ludlum, including the latter's collaborations (Fig. 8). Not only do translations of works of the same author appear in separate clusters; Delta even seems to distinguish (with a single exception) Ludlum's collaborations from his individual efforts, despite, often, shared translators (Jamrych 2011: 18).

Fig. 8. Polish translations of 70 thrillers



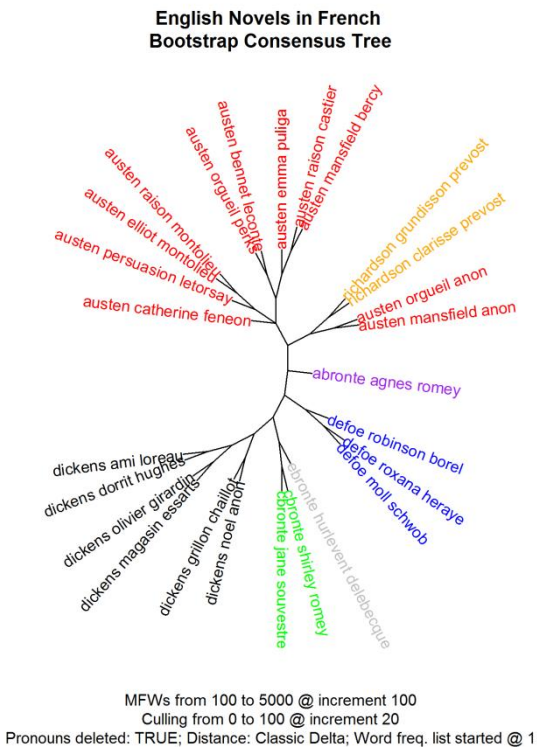
To provide yet another of many similar examples, Fig. 9 presents a Delta bootstrap tree for a corpus of English translations of French novels. Although most original authors are contained within individual branch clusters, some (including Daudet, France, Sand, Sue, Zola) are scattered; in none of these cases is this caused by translations of a single translator grouping together. In fact, works by the two translators who translated more than one author in this corpus, Ives (Daudet, Sand) and Wormeley (Balzac, Daudet), share their branch with at least one other novel by the author of the original.

Fig. 9. 46 English translations by 30 translators of 42 novels by 7 French authors



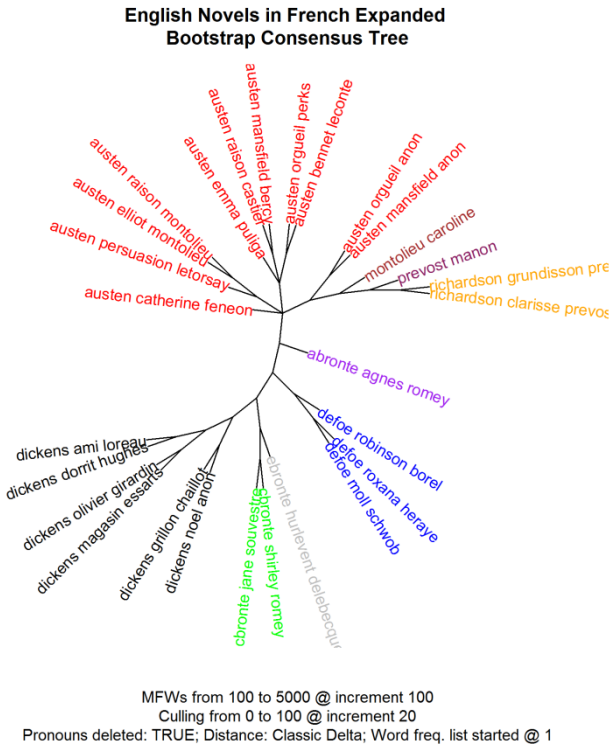
In a reciprocal corpus, that of English novels translated into French, ordering by original author is also very visible (Fig. 10). In fact, the few exceptions to this rule can be easily explained. Namely, the two translations of Austen that do not belong to the Austen cluster are in fact early and abridged anonymous French versions; the preface in one of them suggests the same publisher for both and hence, probably, the same translator/abbreviator. Equally interesting is the cluster of Charlotte and Emily Brontë, which reflects quite well the proximity between the two (and sometimes the three) sisters in diagrams for the original English novels (Figs 1 and 2).

Fig. 10. 26 French translations by 22 translators of 21 novels by 5 English authors



Interestingly, when this corpus receives two additional texts: original French works by two of its translators, the effect is not uniform (Fig. 11). Abbé Prévost's *Manon Lescaux* seems quite similar to his two translations of Richardson; Baroness de Montolieu's *Caroline de Lichtfield*, by contrast, does not place itself any closer to her translations of Austen. This variation has already been described from other material. Namely, in one of the earliest applications of Delta, Burrows's own study of English translations of Juvenal's *Tenth Satire*, Dryden is shown to be "able to conceal his hand" as a translator, while Johnson "strikes his own note and holds it" (Burrows 2002a: 688)

Fig. 11. As above, with two French originals added



Conclusions

The above collection of exemplary diagrams seems an unexpected corroboration of Venuti's observation on translator's invisibility. Indeed, it is adding insult to injury. Not only do "translators receive minimal recognition for their work" (Venuti 1995: 8) in fame and fortune and law; not only is their work usually best praised when it is not mentioned at all – as I know from my own experience as a literary translator. Now this study seems to be adding an additional dimension to "the translator's shadowy existence": statistics – what is more, simple statistics of word usage – make them invisible too. It has been one of the original tenets of this variety of authorship attribution that it deals with frequent words – after all, the 5K most frequent words that are the basis of such conclusions are a fairly thin layer on top of the 50K words in adult vocabulary (Miller 1996) – and thus, in a great part, it remains outside the realm of conscious choice, and free of manipulation (c.f. Burrows 1987). In other words, multivariate analysis of most-frequent-word usage further – and in a novel way – condemns translators to stylometric invisibility. In the context of this study, they only emerge from it when they do something wrong, or at least controversial, like deleting fragments of a novel or adding their own two pence to the original writer's guinea. Of course, those who believe that invisibility is the translator's main task will be gratified with the results of translation attribution analysis by means of Delta.

Stylometric translator invisibility goes somewhat against the grain of one of the main preoccupations of translation studies; after all, the entire field deals, among other things, with how translators distort the original – as evidenced by terms such as "translator's traces" or Berman's "deforming tendencies." This, in turn, seems to quarrel with the fact, famously

remarked a decade ago by Mona Baker, that translational style "has been somewhat neglected in translation studies" (Baker 2000: 245). Now the said style is coming under the scrutiny of translation scholars, perhaps most visibly so of the corpus-linguistic variety. Stylometry is certainly mentioned in a crucial text in the field, Olohan's *Introducing Corpora in Translation Studies* (Olohan 2004), although with little reference to stylometry by most frequent words. In spite of appearances, the results obtained in this study do not negate this intuition: they establish a crucial fact that "traditional" non-traditional authorship attribution methods such as Delta might not necessarily be adequate for differentiating between individual translators' styles, possibly for the simple reason that word usage – even the most unconscious usage of the most content-less function words – is *not* style, or not *solely* style. The consistent tendency of the various translations in the diagrams presented in this paper to cluster by author and by volume rather than by translator might in fact indicate that Delta has its content-conscious side, which becomes more influential in studies of translations because two translations of the same text into the same language share much more than any other two literary texts written in the same language.

Stylometry does not end with Delta. Even without leaving the sphere of Burrows, two of his other attributive methods, Zeta and Iota, might be the answer for translator attribution, as they search for authorial evidence in, respectively, the middle and the lowest word frequency strata and – contrarily to the avowedly multidimensional Delta – identify the individual words responsible for the differences between texts by two (or more) translators (Burrows 2007).

References

Baker, Mona. 2000. Towards a methodology for investigating the style of a literary translator. *Target* 12(2): 241-266.

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

Burrows, John. 2002. 'Delta': A Measure of Stylistic Difference and A Guide to Likely Authorship. *Literary and Linguistic Computing* 17(3): 267-287.

Burrows, John. 2002a. The Englishing of Juvenal: computational stylistics and translated texts. *Style* 36: 677-99.

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

Eder, Maciej. 2010. Does Size Matter? Authorship Attribution, Small Samples, Big Problem. In *Proceedings of the 2010 Digital Humanities conference*, 132-135. London: King's College.

Eder, Maciej & Rybicki, Jan. 2011. Do Birds of a Feather Really Flock Together, or How to Choose Test Samples for Authorship Attribution. In *Proceedings of the 2011 Digital Humanities conference,* 124-127. Stanford: Stanford University.

Eder, Maciej & Rybicki, Jan. 2011a. PCA, Delta, JGAAP and Polish Poetry of the 16th and the 17th Centuries: Who Wrote the Dirty Stuff? *Literary and Linguistic Computing*, (forthcoming).

Eder, Maciej & Rybicki, Jan. 2011b. Stylometry with R. *Proceedings of the 2011 Digital Humanities conference*, 308-311. Stanford: Stanford University.

Hoover, David. 2004a. Testing Burrows's Delta. *Literary and Linguistic Computing* 19(4): 453-475.

Hoover, David. 2004b. Delta Prime? *Literary and Linguistic Computing*, 19(4): 477-495.

Hoover, David. 2005. Delta, Delta Prime, and Modern American poetry: Authorship Attribution Theory and Method. *Proceedings of the 2005 ALLC/ACH conference*, 79-80. Victoria: University of Victoria.

Hoover, David. 2007. Corpus Stylistics, Stylometry, and the Styles of Henry James. *Style* 41: 174-203.

Jamrych, Magdalena. 2011. The Digital Mystery of the Thriller Genre: A Multifaceted Stylometric Analysis of English Originals and Polish Translations. M.A. thesis, Uniwersytet Pedagogiczny w Krakowie.

Juola, Patrick. 2009. Cross-linguistic Transference of Authorship Attribution, or Why English-Only Prototypes Are Acceptable. *Proceedings of the Digital Humanities 2009 Conference*, 162-163. College Park: University of Maryland.

Miller, G.A. 1996. *The Science of Words*. New York: Freeman.

Mosteller, F. and Wallace, D. 1964. *Inference and Disputed Authorship: The Federalist Papers.* New York: Springer.

Olohan, M. 2004. *Introducing Corpora in Translation Studies*. London: Routledge.

Rybicki, Jan. 2010. Original, Translation, Inflation. Are All Translations Longer than Their Originals? *Proceedings of the 2010 Digital Humanities conference*, 363-364. London: King's College.

Rybicki, Jan. & Eder, Maciej. 2011. Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? *Literary and Linguistic Computing* 26(3): 315-321.

Segel, Harold Bernard. (1991), book review in *The Polish Review* 36(4): 486-495.

Smith, Peter & Aldridge, W. 2011. "Improving Authorship Attribution: Optimizing Burrows's Delta Method." *Journal of Quantitative Linguistics* 18(1): 63-88.

Venuti, L. 1995. *The Translator's Invisibility: A History of Translation*. London: Routledge.

Vickers, Brian. 2011. Shakespeare and Authorship Studies in the 21st Century. *Shakespeare Quarterly* 62(1): 106-142.