Studies in Polish Linguistics
vol. 10, 2015, issue 2, pp. 87–104
doi:10.4467/23005920SPL.15.004.3561
www.ejournals.eu/SPL

Jan Rybicki
Jagiellonian University in Kraków

Success Rates in Most-frequent-word-based Authorship Attribution. A Case Study of 1000 Polish Novels from Ignacy Krasicki to Jerzy Pilch

Abstract
The success rate of authorship attribution by multivariate analysis of most-frequent-word frequencies is studied in a 1000-novel corpus of Polish literary works from the late 18th to the early 21st century. The results are examined for possible influences of the number of authors and/or the number of texts to be attributed. Also, the success rates achieved in this study are compared to those obtained in earlier studies for smaller corpora, too small perhaps to produce regular patterns. This study shows that text sets of this size confirm the intuitive predictions as to those influences: 1) the more authors, the less successful attribution; 2) for the same number of authors, the number of texts to be attributed does not influence success rate.

Keywords
multivariate analysis, authorship attribution, Polish literature, stylometry

Streszczenie
W artykule zbadano skuteczność atrybucji autorskiej opartej na wielowymiarowej analizie najczęstszych słów w korpusie 1000 powieści polskich napisanych między końcem XVIII i początkiem XXI wieku. Oceniono wpływ liczby autorów i/lub tekstów na uzyskane wyniki. Porównano skuteczność atrybucji w niniejszej pracy z wynikami uzyskanymi we wcześniejszych opracowaniach wykorzystujących mniejsze korpusy – a więc te, które mogły nie wykazywać regularnych prawidłowości pod tym względem. Wykazano, że w dużych kolekcjach tekstów sprawdzają się intuicyjne przypuszczenia: 1) im więcej autorów, tym trudniej o skuteczną atrybucję; 2) przy tej samej liczbie autorów liczba tekstów nie ma wpływu na skuteczność atrybucji.

Słowa kluczowe
analiza wielowymiarowa, atrybucja autorska, literatura polska, stylometria

1. Introduction

For all the progress in information-from-language retrieval, it still largely relies on the "bag-of-words" assumption:

> Texts (e.g. queries and documents) are represented as unordered sets of terms. This means that any notion of term ordering is lost. For example, under this representation, the texts *the bear ate the human* and *the human ate the bear* are identical. However, these pieces of text clearly have different meanings. While this is an overly simplistic representation, very few have been able to develop non-bag of words retrieval models that are consistently and significantly better than the state-of-the-art bag of words models. (Metzler 2011: 3)

Not only term ordering is thus ignored; grammar, in a broad sense, is left out of the picture too, especially when information retrieval – and this is also true of reading novels, not just trawling the Internet – is focused on "meaningful" words, e.g. in various Web search engines. And while Google has been using a set of "stop-words" that are ignored in its queries: the most common words such as modals, prepositions or pronouns, it is exactly that manner of words that has been serving quite well in "non-traditional", or quantitative, authorial attribution in the case of material where meanings are usually supposed to reside in deeper linguistic structures than mere words: literature, *belles lettres*, artistic writing.

There is a visible epistemological gap between the "insignificant" feature that is most-frequent-word usage and the overall effect it seems to have in authorial style, if it is style that is reflected in that usage, as most stylometrists seem to assume. Explanations of this phenomenon have been metaphorical rather than satisfactory. Anthony Kenny speaks of a "stylistic fingerprint," which

> … would be a feature of an author's style – a combination perhaps of very humble features such as the frequency of such as – no less unique to him that a bodily fingerprint is. Being a trivial and humble feature of style would be no objection to its use for identification purposes: the whorls and loops at the ends of our fingers are not valuable or striking parts of our bodily appearance. (Kenny 1982: 12)

While a later paper co-authored by John Burrows, one of the crucial exponents of this new stylometric mode of literary research (Burrows 1987), goes a little further than that, the explanation is anything but explanatory:

> The possibility of using frequency patterns of very common words rests upon the fact that words do not function as discrete entities. Since they gain their full meaning through the different sorts of relationship they form with each other, they can be seen as markers of those relationships and. accordingly, of everything that those relationships entail. Thus, where the most common prepositions occur more often than usual,

> an abundance of prepositional phrases will usually mark a descriptive or reflective tendency in the writing. Where they are few, it is usually because the action is taking place upon a barer stage. … Other words than these have important stylistic implications of many other kinds. (McKenna et al. 1999: 152)

Although the above sources identify the question before the turn of the century, little has been done since to provide an definite answer. In a paper published quite recently in this journal, Maciej Eder has presented an extensive survey of the findings in stylometry and non-traditional authorship attribution in its evolutionary line from Mosteller and Wallace, through Baayen, Burrows, Hoover, Juola and Craig; major Polish contributions to the field include Pawłowski's work on the authorship of Romain Gary (Pawłowski 2003). In the above-mentioned paper that builds on that research, Eder presented numerous examples of how authorship attribution is best served by counting very simple lexical features of texts, such as most frequent single word or word n-gram frequencies (Eder 2011). What is more, even more recent studies seems to suggest that the authorial signal is more visible in frequencies of single words than in any syntax-based quantitative approaches – in word choice, then, rather than in sentence structure (Górski et al. 2014). All this may seem to some (literary rather than linguistic) scholars quite anti-intuitive: why should the difference between authorial voices be best discernible in the most frequent words – words that, according to all three Zipfian laws, are also shortest and least "meaningful"? The problem might seem even more controversial: now that the author has been apparently dead (Barthes 1967) to much of the literary studies community,[1] how is it possible that all it takes to identify him or her is frequencies of a few dozen of his or her (and everyone else's) articles, pronouns, prepositions and modals?

The stylometrist might strive to find less shaky theoretical grounds in two fields: in cognitive linguistics and in psychology. The former may help:

> … by integrating two, previously unrelated areas of literature, namely Burrows-style computational stylistics and Langacker's Cognitive Grammar. I suggest that this hybrid approach can help explain the contribution function words make to authorial style, and provide a theoretical connection between literary studies and computational stylistics. Cognitive Grammar provides insight into how function words operate in language, and their role in shaping style and the perception of style. Cognitive Grammar also points to an understanding of function words as part of a communicative strategy that is socially and culturally embedded. It thus goes some way towards closing the interpretive gap between raw counts represented in the data and the complexity of language as it is used in literary, social and textual contexts. (Connors 2013: 3)

[1] For a spirited stylometrist's response to this issue, cf. Love 2002: 3–7.

At the close of her text, Connors feels entitled to say:

> Cognitive Grammar supports the interpretation of stylistic results in an integrated, rhetorically motivated and non-reductive way and supports an engagement with the psychological, socio-cultural, and historical elements of the text. The key to this is the concept of embodiment and its insistence that language is a product of cognitive phenomena and is both produced and constrained by perceptual systems. Cognitive Grammar can explain the existence of computational results in a way that other accounts of language cannot. It offers a rich interpretive model that does not neglect the importance of author, reader, or context through its approach to language and literature as an expression of an innately constrained and embodied human mind. (Connors 2013: 219)

The other source of inspiration is grounded in psychology – and social studies. In his influential book *The Secret Life of Pronouns: What Our Words Say About Us*, James Pennebaker presumes to offer a helping hand, and he does that by echoing – apparently, quite independently – Burrows's above-quoted argument:

> Pronouns, articles, prepositions, and a handful of other small, stealthy words reveal parts of your personality, thinking style, emotional state, and connections with others. These words, typically called function words, account for less than one-tenth of 1 percent of your vocabulary but make up almost 60 percent of the words you use. Your brain is not wired to notice them but if you pay close attention, you will start to see their subtle power. Function words behave differently than you might think. For example, the most commonly used word in spoken English, I, is used at far higher rates by followers than by leaders, truth-tellers than liars. People who use high rates of articles – a, an, the – do better in college than low users. And if you want to find your true love, compare the ways you use function words with that of your prospective partners. (Pennebaker 2011: ix)

It is perhaps not a coincidence that Pennebaker, too, cites cognitive linguist George Lakoff as well as computational linguist Douglas Biber. And it is no wonder that his book has been quite favourably received by the more stylometrically-minded Digital Humanists, as evidenced by this review in their flagship journal, *Literary and Linguistic Computing*:

> The book deserves a review in LLC because it pays attention to linguistic and literary interests, and especially because it adds an interpretive dimension to stylometry (the study of style using exact techniques), which has been underdeveloped to-date …. As readers of this journal know, stylometry has come to focus increasingly on authorship attribution as an objective validation of its work, and has come to accept … that function word distributions are the most interesting indicators of authorship. I will criticize Pennebaker a bit later in the text for largely ignoring the stylometric literature, but I will focus on what he does contribute, and that is a great deal. (Nerbonne 2014: 140)

2. The problem

While the prospects for most-frequent-word stylometry's theoretical background seem somewhat brighter, there still always seems to be, in this field, at least one variable too many. One of these is the issue of comparing results between methods, or between results of the same methods for different corpora and/in different languages. It is true that stylometry seems to be working across all languages it has been so far tried on and, according to some scholars, methods prototyped in one language should work in any other (Juola 2009). But other results are less clear: Eder and Rybicki have conducted a cross-language study which showed that literary authorship attribution success rates vary from language to language and from corpus to corpus, and that there is seems to be no golden rule whether more texts or more authors make the guessing more or less difficult. And while they tried to blame the worst attribution accuracy rates they obtained in Polish corpora on the highly inflected language, the relatively better results for a set of 19th-century Hungarian realist novel seemed to disprove this elegant hypothesis (Eder and Rybicki 2013).

This in turn prompted an international group of researchers gathered at Stylometry@Kraków in 2013 to attempt establishing a series of corpora in a number of languages that would be based on similar rules of number of authors, number of texts per authors, genre and gender, etc. Since the realist novels functions as a discernible quality in most European national literatures, prototypes of 100 texts, with 3 texts per author on the average, were made. But then problems appeared: while it was quite easy to achieve the prescribed minimum of 33% of women among 19th-century English realist writers, and representative Polish literature of the same genre and era could go as high as 50% without respected levels of literary quality, or at least of literary celebrity, there simply were not enough French female writers to produce a similar balance. This has not been the only attempt to compile "comparative" corpora of different literatures. There is a report of Richard Forsyth's attempt to produce an English benchmark corpus for exactly the same purpose around 2003; Patrick Juola's Ad-hoc authorship attribution competition at the ALLC/ACH 2004 conference in Gothenburg was another one, and the participants competed in authorial recognition problems in several languages (2004); the idea of multilingual benchmarking was further developed by Forsyth and Sharoff (2014). Yet neither of these have been able to establish a series of sizeable and comparable corpora that would be adopted by other scholars to test their ever-emerging new methods of attribution.

Yet even if such benchmark sets of texts were feasible, there is absolutely no proof that the arbitrary choice of texts for such a benchmark corpus would not bring major individual differences that, in turn, would be constantly distorting any comparability between sets of texts derived from what has already been described here as a genre and a period common to most European literatures.

At a closer look, however, Polish 19th-century realism either goes towards the Auerbachian tendency novel or recedes into the post-Romantic historical romance; its worship of Balzac is Platonic rather than imitative, and it is very distant, in its mainstream, from Zola's naturalism for the simple reason that, according to Czesław Miłosz (1983: 290), "Polish theoreticians of prose were more fascinated by the artistry of Gustave Flaubert than by that of Zola".

No wonder, then, that comparability of such small corpora in different languages could not be achieved – at least at present. All that is presently feasible in terms of the main question of this paper is to expand the corpus of texts in a single language by at least one order of magnitude, and to try to obtain more cross-validated results for one national literature that would confirm or deny the intuitive and the mathematical expectation that, for instance, the less authors there are to guess in an attribution test, the easier it should be; after all, this would be the case if the attribution were decided by coin-toss. But then studies on smaller corpora sometimes produced results that refused to follow the path of the coin-toss, and that includes not only the already-cited Eder and Rybicki's (2013) paper, but also their earlier work (Rybicki and Eder 2011).

It should be stated at this point that this decision has also been influenced by the growing tendency in stylometry to expand its corpora. The temptation to use quantitative approaches is so great, partly because of stronger statistics, partly because electronic texts are simply *there*, and *en masse*. Matthew Jockers is perhaps not just deliberately provocative when he writes that "today's literary-historical scholar can no longer risk being just a close reader: the sheer quantity of available data makes the traditional practice of close reading untenable as an exhaustive or definitive method of evidence gathering" (Jockers 2013: 9). And then there is the ultimate reason behind Franco Moretti's distant reading, the need for which is stated simply by British writer Mark Mason:

> I turn 40 this year, so if you say I've got another 40 left, and I read average two books a month (work and parenthood mean it's well below that at the moment), this will bring my final total to 1760. For an activity that I've been doing all my life, and hopefully always will, that doesn't seem a very large number. It makes me feel inadequate, frustrated that there are so many books I'd like to read but never will. (Mason 2011)

Data from Poland's publishers paint a similar picture: recently, an average of ca. 4000 new novels are published in Polish every year; this includes both originals and translations from other languages (Dobrołęcki 2014). Less than three years of this production are enough to fill a human lifetime of reading, which becomes exhausted at 10,000 novels, counting half a book a day for sixty years. While there are legends about greater reading achievements the real question is about how many books can be assimilated and meaningfully analysed as a single set by a single traditional literary scholar. Now that literature has gone digital in a variety of e-book formats, traditional close reading can be – not

superseded, but complemented – by "distant reading:" computer-aided text- and data-mining of large quantities of cultural material to observe otherwise undiscernible relationships and, perhaps more importantly, to test the existing "models" of literature. Starting with the simplest and the most traditional ones: that, despite the already-mentioned death of the author, there are peculiarities of word usage that allow to see the difference between one cadaver and the next; that the works of such cadavers may be grouped into sets such as literary "periods" or "schools;" that there exist elements of individual cadavers' output that can be explained by their place in time and space; that "canonic" writers set the mood, the style, the qualities of their (or later) literary eras; or that translated literature shows traces of the original author, of the translator, or of both.

3. Material and method

In view of the above serious problems, the experiment described in this study might seem minimalist by contrast. It was conducted on a corpus of a mere 1000 Polish novels, novellas and short story collections by a mere 250 authors. The texts were at least 10,000 words long; this is usually considered the safe minimum for the results not to be influenced by such differences in size as that between Jacek Dukaj's *Lód* and Maria Dąbrowska's *Marcin Kozera* (Eder 2013). While the average number of texts per author was 4, in reality it could be anywhere between 2 and 12. The reason for this wide range of values was, quite obviously, that there were significant Polish writers who only published two books, Bruno Schulz being the case in point; the numbers at the other end of the range were reserved to important and prolific writers like Ignacy Kraszewski, who only recently lost his long-held world record of novels (232), or Henryk Sienkiewicz, the first Polish literary Nobel Prize winner. At the same time, single works by authors were eliminated for the obvious reason of there being nothing to guess; and the inclusion of Kraszewski's entire *oeuvre* would dangerously bias the composition of the master wordlist and move the whole study towards a differentiation between the author of *Stara Baśń* and the rest of Polish literature – an interesting subject for a separate study, perhaps, but something that should be avoided in the present context.

In chronological terms, the earliest work in the corpus was *Mikołaja Doświadczyńskiego przypadki*, the utopian novel by Ignacy Krasicki (1775), followed by his other venture into the genre, the physiocratic *Pan Podstoli* (1778–1803). Very typically for the European literature of the period, the next author on the list was Anna Mostowska with her four Gothic novels (1806–1807). The first half of the 19th century is represented by 40 novels and ends with Józef "Bolesta" Dzierkowski's satirical *Szpicrut honorowy* of 1848. The century's latter half includes as many as 151 novels and begins with one of

the multitude of Kraszewski's historical romances, *Kordecki* (1850); it includes the Great Three of Poland's realists: Eliza Orzeszkowa, Bolesław Prus, Henryk Sienkiewicz; and, at the end, it ushers in a new generation of writers with Stefan Żeromski's *Ludzie bezdomni* of 1899. The 20th century opens, somewhat inauspiciously, with the graphomaniac Wanda Grot-Bęczkowska-Korotyńska and her pretentious *Anima vagans* (1900). Two world wars, one restoration of the country's independence and one loss of same later, the mid-century arrives with the concentration camp novel by Igor Newerly, *Chłopiec z Salskich Stepów* (1948), which brings the number of novels published until 1950 to 211. The following pentadecade opens with perhaps the most troublesome item in terms of authorship: a reconstruction of scattered drafts to Maria Dąbrowska's *Przygody człowieka myślącego*, published by Ewa Korzeniewska in 1970. The 387 books of the 20th century's latter half end in as many as 13 published in 1999 – by authors as different as Maria Józefacka and Andrzej Stasiuk. The first thirteen years of the 21st are represented by more than 200 novels; the third millennium opens with Jerzy Pilch, Jacek Dukaj and Manuela Gretkowska.

This is a somewhat lengthy description of the corpus, albeit a necessary one, since it illustrates quite well the variety of theme, genre or trend that teems within those 239 years of Polish literature, with an average frequency of more than 4 books per year. The total size of the corpus was 68,129,241 word-tokens.

Since this paper aimed to confront the above-quoted earlier studies, Eder and Rybicki (2013) and Górski et al. (2014), which both used much smaller corpora, the method followed here was the one adopted in the earlier study and further modified for the latter one. All were based on machine-learning classification in which one text by each author became part of a reference set, and this was then used by the machine to "learn" the (hopefully) characteristic series of most-frequent-word frequencies for each author; all the other texts became the test corpus, and their authorship would be guessed by using the first 100, 200, 300 … 5000 most frequent words (based on a list established for the entire corpus). This would allow to measure the success rate achieved in various approaches to the corpus. For each wordlist length value, another series of 100 iterations were performed by randomly replacing the "reference" texts by a given author with one or another of his or her texts. This 100-fold cross-validation was necessary to minimize the risk of accidentally stumbling upon a particularly easy or particularly hard attribution case and, multiplied by the 10-fold repetition of the results at each of the above wordlist sizes, and then multiplied again by the 50 different values of the wordlist length parameter, resulted in 50,000 individual analyses performed on the corpus at each stage of the experiment.

The main difference from the Górski et al. (2014) study – that is, apart from using a much larger corpus – was that since it had established quite clearly that the Delta distance measure combined with single-word frequencies was the op-

timal one – it performed better than combinations of other distance measures and word- or part-of-speech n-grams – only unlemmatized words were used. It is true that lemmata slightly increased attribution success rate in the earlier experiment; yet the smaller corpus used in that study (just ca. 250 novels) made it possible to verify by hand the underlying automatic lemmatization; this was not an option here with 1000 texts, and the use of unverified automatic POS tagging threatened to introduce too many errors and make the results uncertain. As to the reliability of word n-grams in this particular corpus, Figure 1 speaks volumes on what happens whenever n is more than 1: single words out-attribute word n-grams by a wide margin, and success rate drops with every increase of n until, at n = 5 and more, it begins to approach coin-toss precision.

At the beginning of the first experiment of this study, the impact, on authorial attribution success rate, of the number of reference texts, or the number of authors to be guessed from the reference set, was measured. At first, this task was performed for the entire corpus, so that the reference set included a text by each of the 250 authors. Then the number of authors to be guessed was decreased by intervals, and only texts by this smaller number of authors

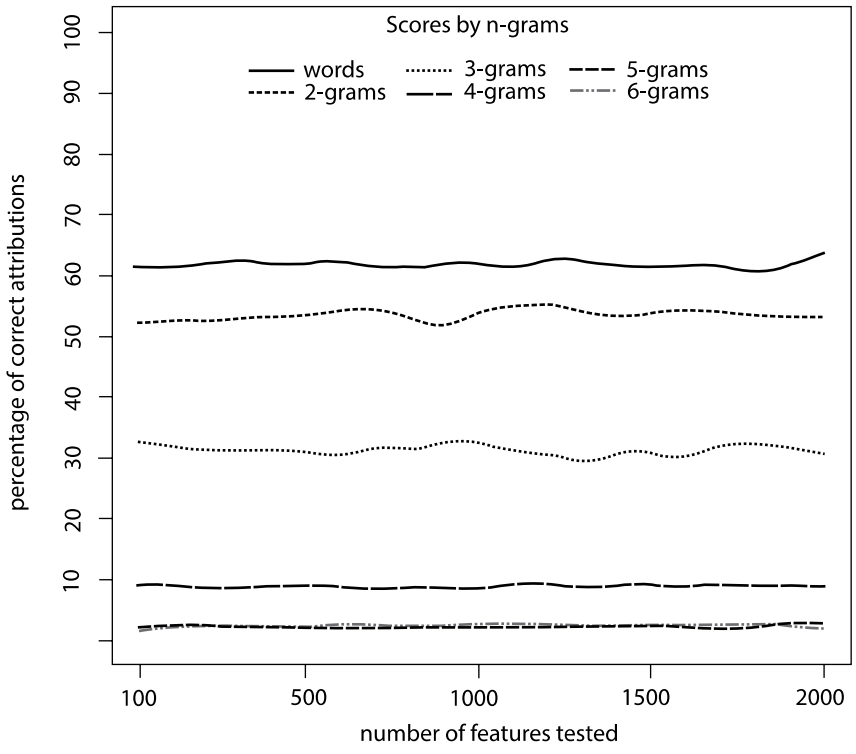

Figure 1. Attribution success rates for single words and word n-grams (n ranging from 2 to 6) in a corpus of 250 authors and 1000 texts

were attributed until the last one hundredfold iteration at just two authors. The results for each number of authors of the corpus were compared both directly and by plotting a density function; the latter is a way to show the distribution of the particular success rate values.

The second part of the experiment looked for a similar impact of the number of texts by a given author in the test set. This was assayed for 2, 3, 4 and 5 texts by the same author in the test corpus, slightly diminishing the number of test texts; as an obvious consequence, the reference became smaller and smaller, since there were less authors available with 3 texts in the test corpus than those with just 2, and even less authors with 4 texts each, etc. These results, too, were compared.

Finally, since the previous part was likely to introduce distortion, the number of authors in the reference test was limited to those who had at least four texts in the test set, and success rates were compared for attribution of one, two, three and then four books by each author. The whole procedure was conducted using R, the open-source statistical programming environment (R Core Team 2014), the *stylo* package written for this environment (Eder et al. 2013), especially its *classify* function, and several custom-made R scripts to cross-validate the results and produce graphs. The distance measure used was Burrows Classical Delta as implemented in the package: the value of Delta between two texts is obtained as the mean of the absolute differences between the z-scores for a set of relative word-frequencies in one text and in the other, according to the formula:

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

where

$$ü(\ _x(\ )) = \frac{f_x(T) - \mu_x}{\sigma_x},$$

$f_x(T)$ = raw frequency of word x in text T;
μ_x = mean frequency of word x in a collection of texts;
σ_x = standard deviation of frequency of word x.

In an attribution attempt within a corpus, an anonymous text is assumed to be authored by an author if the distance from that author's reference text (expressed by the Delta coefficient) is lowest among those from other reference texts (Burrows 2002). In this case, of course, the experiment used the same method on texts of known authorship to measure the rate of attributive success.

4. Results

Figure 2 presents the attribution success rates for the number of authors from 2 to 10, 15, 20, 50, 100, 150, 200 and 250. The plot for each of these individual values is represented by a smoothed curve (solid black for two authors and, with more authors, moving towards ever lighter grey). As can be seen, the obvious has been vindicated and it may be safe to say that, in a corpus of this size, a greater number of authors means lower attribution success. This is a consistent trend throughout the data, with the single exception of an almost axis-length overlap between the curves for five and six authors in the reference set. It is worth noting that most of the curves tend to peak somewhere between 500 and 1000 on the wordlist size scale – which might serve as yet another entry into the seemingly never-ending dispute on the preferred parameters in most-frequent-words-based authorship attribution.

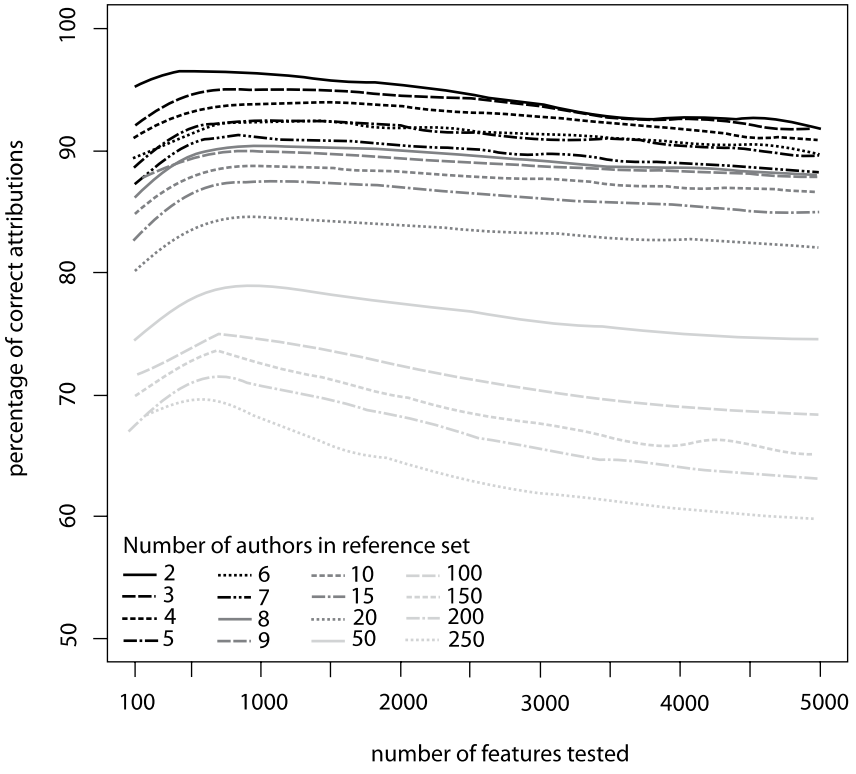

Figure 2. Attribution success rates for 2–250 authors. Colours become lighter with growing number of authors

The density plot performed for the same series of analyses shows the distribution of success rate values across the 1000 iterations for every number of reference set authors. Figure 3 shows the slow but sure drop in attributive success as the number of authors increase, and it slightly modifies the impression derived from the previous graph. Namely, the right-most curve for two authors is mainly distributed over the same section of the horizontal axis as is the second-form right curve for four authors. If anything, the latter is somewhat more reliable, since its distribution has a more consistent peak around 95%, while the former shows that it is just as probable to score 96% and 93%, and 94% and 95% are quite rare. For higher numbers of the reference set authors, the peaks extend further and further to the left and into the low attributive success territory; at the same time, they rise, showing that they might guess with less precision, but when they do, they do so more consistently. As the intervals between the authorial numbers increase (between 20 and 50 authors), the distribution curves move even more rapidly to the left, and the score for all the authors available for this study, while still much better than chance, can hardly be called respectable. The peak for the full corpus (first curve from the left) appears near the 60% success rate mark.

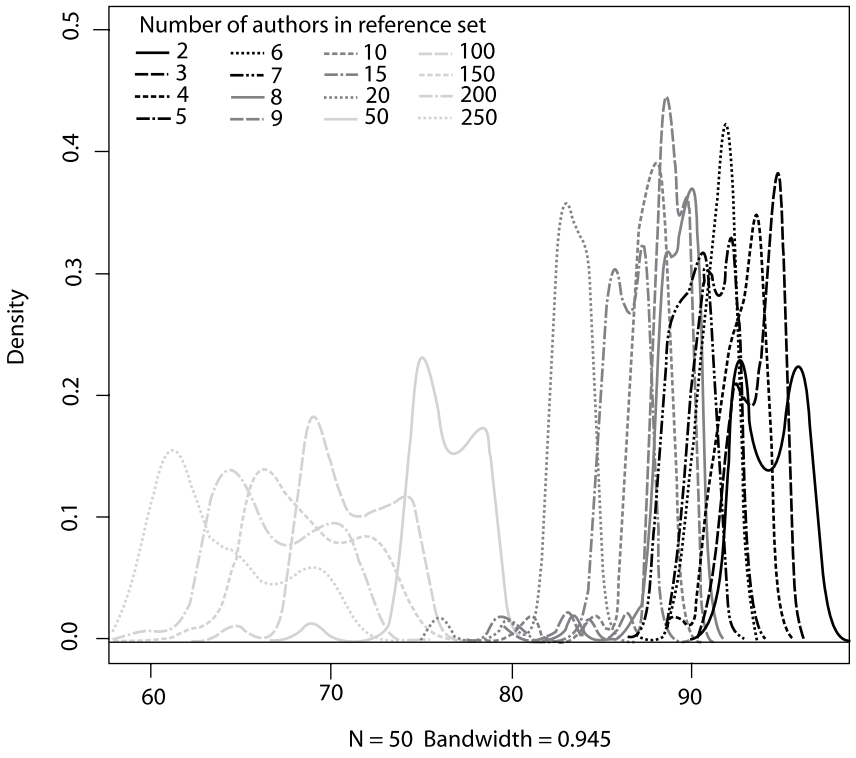

Figure 3. Attribution success rate density for 2–250 authors. Colours become lighter with growing number of authors

While the number of texts in the test set diminished proportionally to that for the reference set and maintained a fairly stable average 4:1 ratio of texts per author, it goes without saying that fluctuations in the number of texts by each individual author might be another influence on attribution rate success in any study of authorship. In order to see the impact of this factor, the following part of the experiment was conducted with two, three, four and five texts per author in the test set; possible distortions could be expected due the fact that there were naturally less and less authors available with more and more texts. In fact, there were 222 authors with just two books or more in the reference set; 122 with at least three; 80 with four or more; and just 24 with five. This is why the test sets consisted of, respectively, 444, 366, 320 and 120 texts. Attributive success was calculated for each of those four groups, and the results compared; but the only thing they showed was the expected bias introduced by differences in reference set size (Figure 4). Figure 5, in turn, shows not only a general drop in attributive success but also a gradual decrease in predictability of results, as the curves for the four cases become wider and wider with the rising number of author.

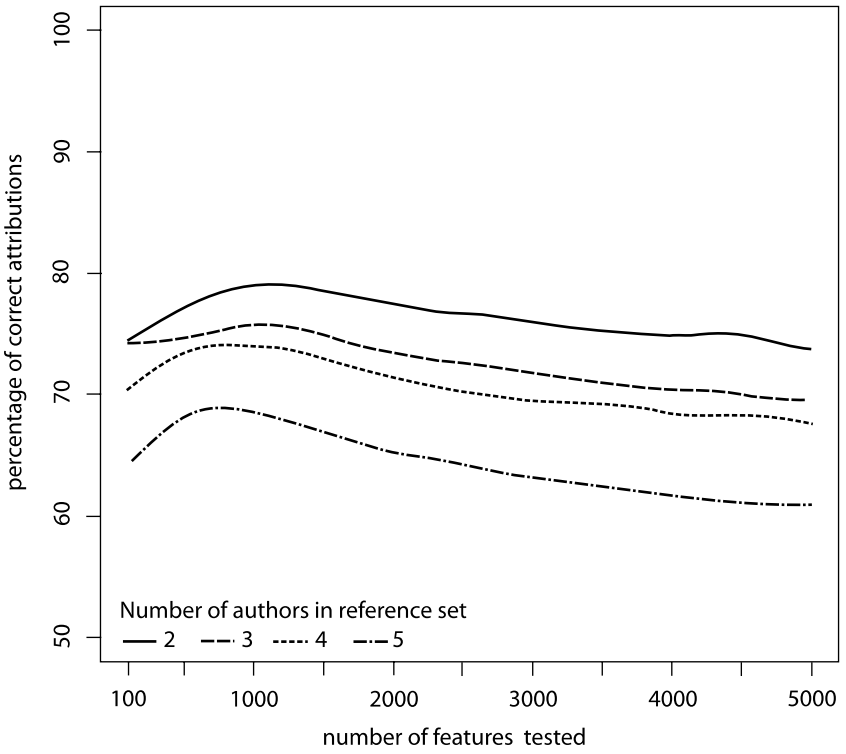

Figure 4. Attribution success rates for authors with 2, 3, 4 and 5 texts in the test set

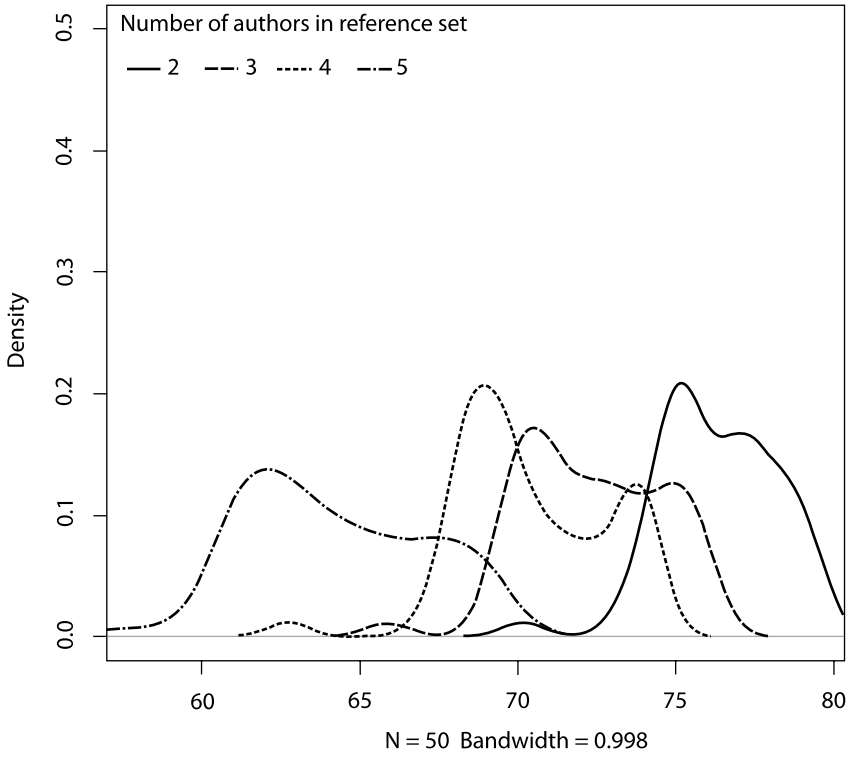

Figure 5. Attribution success rate density for authors with 2, 3, 4 and 5 texts in the test set

The final experiment in the series was one in which a relatively low number of reference set authors, 80, was tested against one text by each author in the test set; then against two texts, three, and four. The benefit was that there was no external distortion; the downside, that the test could only be performed on 80 rather than 250 authors, thus ignoring a part of the material available and significantly worsening the statistics. Uncharacteristically, Figure 6 allows some hope that even such a small sample is believable. Differences in attribution success are small between the four sets, but the bias for better attribution with more texts is there: these differences are statistically significant between the four series of results according to Student's t-test at a fairly low (but acceptable) confidence level of 90%; some, in fact, attain a confidence level of 95%. One might say that, at least in this context, every additional book in the test corpus is a chance to make another good guess rather than to err (Figures 6 and 7).

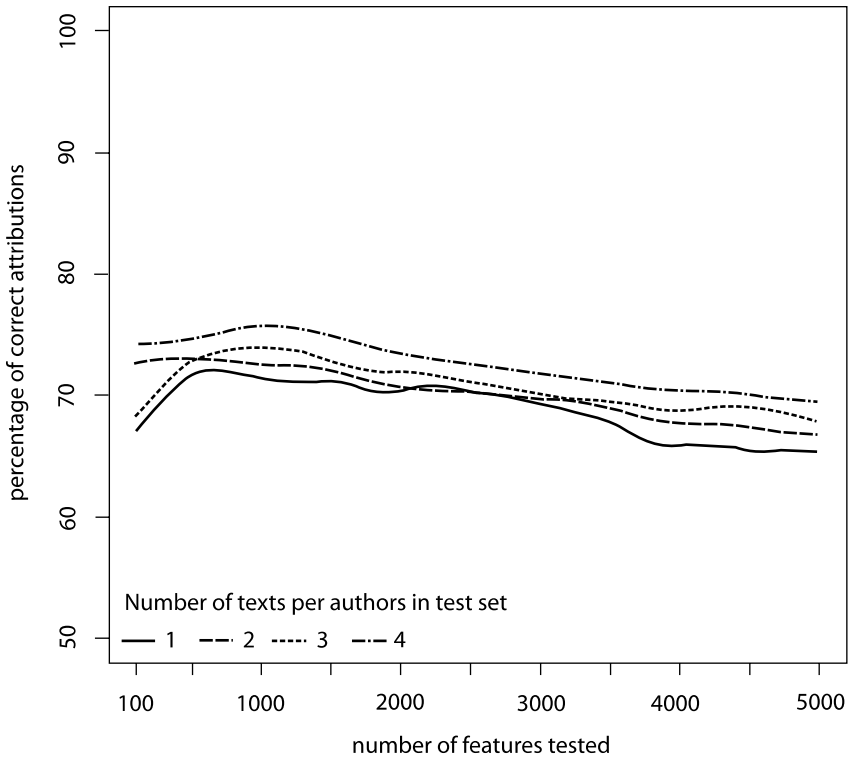

Figure 6. Attribution success rate for authors with 1, 2, 3 and 4 texts in the test set

5. Discussion

The results of this experiment are an addition and a certain modification of the results of the 2013 experiment by Eder and Rybicki, where some of the corpora refused to follow the simple rule: more authors, less successful attribution. In Eder and Rybicki, less authors made the results worse for German and Italian novels; the other corpora tested in that study, including Polish novels, behaved more predictably. And while this should now be tested on bigger corpora in other languages (especially those in German and Italian), it is to be expected that the greater number of texts simply should strengthen the statistics.

Interestingly, the results in this study for a comparable number of authors were significantly better than those obtained in 2013, and this might be attributed to improved algorithms used in the package. On that same note, one should also observe that there was an approximate agreement between the success rates for the comparable 57-author corpus in the Górski et al. (2014) experiment and for 50 authors in this experiment: both scored ca. 80%.

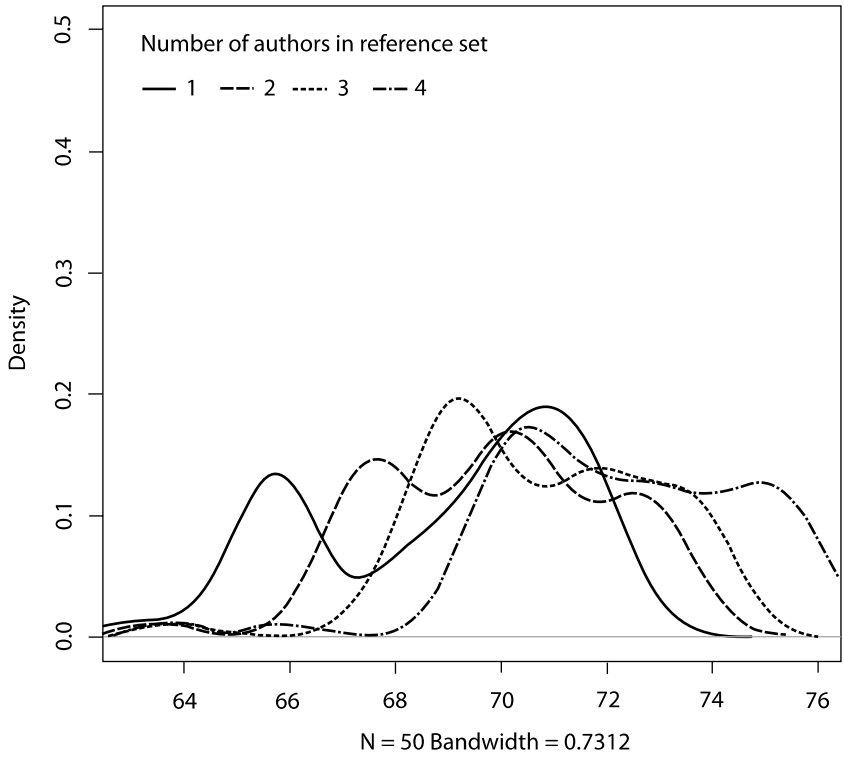

Figure 7. Attribution success rate density for authors with 1, 2, 3 and 4 texts in the test set

Once again: while this is still less than success rates in other languages, it is better than the result obtained for Polish by Eder and Rybicki (2013). One reason has already been given above, and another is that the 2013 study relied on a corpus dominated by 19th century novelists such as Orzeszkowa, Prus or Sienkiewicz; despite any personal animosities and rivalries that might have existed between the three, they all hailed from the same impoverished educated gentry class and they were all submerged in the same fate of a stateless nation – hence their similar interests and preoccupations, but also, perhaps, literary idiom. Such a corpus might simply have been more difficult to attribute than those of more fortunate, or more varied, literary cultures. It is true that once the corpus was made more representative, its authorship attribution success rates (almost) reached the "European" standard. This seems to be a direct vindication of another uncomfortable truth about literary authorship attribution: the additional variable of the literary and cultural history behind its components is a serious if very vague factor. The only consolation is that all that – the classifications, the periodizations and the intertextualities established by traditional literary scholarship through the ages – constitutes the sole

possible theoretical point of reference, the sole possible point of comparison for experimental and quantitative studies of the same material. Experimental stylometrists do not have their theoretical stylometrists as experimental physicists have their theorists, since, as has been stated above, the exact mechanism behind word-frequency-based authorship attribution remains unknown. And the mechanism is quite exact: adding more books to a corpus for attribution improves attribution – this is very good proof that the method works – somehow. So while there is still no hard theoretical basis for using a bag-of-words model and, more precisely, the frequencies of the most-frequent-words as the deciding feature – apart from the premises mentioned in the introduction to this paper – empirical evidence is plentiful and cannot be ignored.

References

Barthes Roland (1967). The death of the author. *Aspen Magazine* 5/6, 67–69.

Burrows John (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.

Connors Louisa (2013). Computational stylistics, Cognitive Grammar, and the Tragedy of Mariam: combining formal and contextual approaches in a computational study of early modern tragedy. Newcastle, NSW: University of Newcastle, Ph.D. thesis.

Dobrołęcki Piotr (2014). Rynek książki w Polsce 2014, *Instytut Książki*. [URL: http://www.instytutksiazki.pl/upload/Files/RYNEK_KSIKI_2014.pdf; accessed May 15, 2015].

Eder Maciej (2011). Style-markers in authorship attribution: A cross-language study of the authorial fingerprint, *Studies in Polish Linguistics* 6, 99–114.

Eder Maciej (2013). Does size matter? Authorship attribution, small samples, big problem, *Literary and Linguistic Computing*, first published online November 14, 2013. [URL: http://dsh.oxfordjournals.org/content/early/2014/12/02/llc.fqt066; accessed February 15, 2015].

Eder Maciej, Kestemont Mike, Rybicki Jan (2013). Stylometry with R: a suite of tools. *Digital Humanities 2013: Conference Abstracts*, 487–489. Lincoln, NE: University of Nebraska.

Eder Maciej, Rybicki Jan (2013). Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing* 28(2), 229–236.

Forsyth Richard, Sharoff Serge (2014). Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing* 29(1), 6–22.

Górski Rafał, Eder Maciej, Rybicki Jan (2014). Stylistic fingerprints, POS tags and inflected languages: a case study in Polish. Paper presented at *QUALICO 2014*, Olomouc, May 29–June 1, 2014.

Jockers Matthew (2013). *Macroanalysis. Digital Methods and Literary History*. Springfield: University of Illinois Press.

Juola Patrick (2004). Ad-hoc authorship attribution competition. Paper presented at *ALLC/ACH 2004*, Göteborg, June 11–16, 2004.

Juola Patrick (2009). Cross-linguistic transference of authorship attribution, or why English-only prototypes are acceptable. *Proceedings of Digital Humanities 2009*, 162–163. College Park, MD.

Kenny Anthony (1982). *The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities.* Oxford/New York: Pergamon Press.

Love Harold (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.

Mason Mark (2011). How many books will you read in your lifetime? *The Telegraph.* [URL: http://blogs.telegraph.co.uk/culture/markmason/100054373/how-many-books-will-you-read-in-your-lifetime/; accessed August 15, 2014].

McKenna Wayne, Burrows John, Antonia Alexis (1999). Beckett's Trilogy: Computational stylistics and the nature of translation. *Revue informatique et statistique dans les sciences humaines* 35(1–4), 151–171.

Metzler Donald (2011). *A Feature-Centric View of Information Retrieval*. Berlin/Heidelberg: Springer.

Miłosz Czesław (1983). *The History of Polish Literature*. Oakland: University of California Press.

Nerbonne John (2014). Review of Pennebaker (2011). *Literary and Linguistic Computing* 29(1), 149–151.

Pawłowski Adam (2003). O problemie atrybucji tekstu w lingwistyce kwantytatywnej (na przykładzie tekstów Romaina Gary. In *Prace językoznawcze dedykowane Profesor Jadwidze Sambor*, Jadwiga Linde-Usiekniewicz, Romuald Huszcza (eds.), 169–190. Warszawa: Wydział Polonistyki Uniwersytetu Warszawskiego.

Pennebaker James (2011). *The Secret Life of Pronouns: What Our Words Say about Us*. New York, Bloomsbury Press.

R Core Team (2014). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Wien. [URL: http://www.R-project.org/]

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.

Instytut Filologii Angielskiej
Uniwersytet Jagielloński
Al. Mickiewicza 9
31-120 Kraków
[jan.rybicki@gmail.com]