Maciej Eder
Pedagogical University of Kraków
Polish Academy of Sciences, Institute of Polish Language

Style-Markers in Authorship Attribution
A Cross-Language Study of the Authorial Fingerprint

Abstract
The present study addresses one of the theoretical problems of computer-assisted authorship attribution, namely the question which traceable features of language can betray authorial uniqueness (a stylistic *fingerprint*) of literary texts. A number of recent approaches show that apart from lexical measures — especially those relying on the frequencies of the most frequent words — also some other features of written language are considerably effective as discriminators of authorial style. However, there have been no attempts to compare the attribution potential of these features. The aim of the present study, then, was to examine the effectiveness of several style-markers in authorship attribution. The style-markers chosen for the empirical investigation are those that can be retrieved from a non-lemmatized corpus of plain text files, such as the most frequent words, word bi-grams, different letter sequences, and markers of different nature, combined in one sample. Equally important, however, was to compare usefulness of the chosen style-markers across a few languages: English, Polish, German, and Latin. The results confirmed a high attribution effectiveness of word-based style-markers in the English corpus, but the alternative markers are shown to be usually more effective in the other languages.

Key words:
authorship attribution, stylometry, style-markers, multidimensional methods, Delta method, controlled attribution test

Streszczenie
Wskaźniki stylu w atrybucji autorskiej. Studium porównawcze autorskiego „odcisku palca" w kilku językach
Niniejszy artykuł poświęcony jest jednemu z teoretycznych problemów atrybucji autorskiej opartej o metody ilościowe, mianowicie kwestii, które kategorie językowe zdradzają indywidualny rys autorski (stylistyczny „odcisk palca") w tekstach literackich. Liczne prace powstające w ostatnich latach dowodzą, że oprócz miar leksykalnych — szczególnie tych, które oparte są na częstości wystąpień najczęstszych wyrazów — także inne cechy języka pisanego okazują się stosunkowo silnymi czynnikami różnicującymi styl autorski. Do tej pory nie pojawiły się jednak prace, które próbowałyby porównać atrybucyjne możliwości tych cech językowych z sobą. Celem niniejszego studium było zatem przetestowanie siły dyskryminacyjnej kilku wskaźników stylu w rozpoznawaniu autorów. Do empirycznej analizy wybrano te wskaźniki, które można wyłonić z nielematyzowanego korpusu, tj. ze zwykłych plików tekstowych, takie jak najczęstsze wyrazy, zestawienia dwóch słów, różne połączenia literowe, wreszcie wskaźniki niejednorodne, połączone w jednej próbce. Równie

ważne było jednak porównanie przydatności owych wybranych wskaźników stylu w kilku językach: angielskim, polskim, niemieckim i łacińskim. Wyniki potwierdziły wysoką wartość wskaźników leksykalnych w języku angielskim, podczas gdy w innych językach na ogół dokładniejsze okazywały się wskaźniki alternatywne.

Słowa klucze:
atrybucja autorska, stylometria, wskaźniki stylu, metody wielowymiarowe, metoda Delta, kontrolowany test atrybucyjny

1. Introduction

Studies in authorship attribution have their long and fascinating history, originating in Greek antiquity with the investigations on Homer's epic poems undertaken by scholars of the Alexandrian age (3rd century B.C.). Since then, numerous literary texts were ascribed to certain writers or excluded from their canons, attributed with certainty or judged spurious. To name but a few attribution cases, these include the canon of Plato's dialogues, some orations by Demosthenes, the anonymous *Batrachomyomachia*, poems entitled *Appendix Vergiliana*, or a collection of emperors' lives known as *Historia Augusta*. The fame of some brilliant attributions even reached a broader audience, like Lorenzo Valla's study of the *Donation of St Constantine* — a putative edict of emperor Constantine which turned to be a medieval forgery — or like Erasmus' edition of Seneca, where he rejected the widely-accepted Seneca's and St Paul's authorship of their alleged correspondence. Many of the attribution cases were solved since the beginnings of the discipline, but far more texts had to remain anonymous or disputed. Presumably, each national literature has its own famous unsolved attribution case, such as the Shakespearean canon, a collection of Polish erotic poems of the 16th century ascribed to Mikołaj Sęp Szarzyński, the Russian epic poem *The Tale of Igor's Campaign*, and many other. A concise survey of the attribution discussions of the past and present, yet limited to English and classical literature, is provided by Love (2002: 14–31).

Apart from traditional approaches to the problem of authorship, which rely on methods provided by philology, paleography, codicology, history, or forensic sciences, the discipline was substantially enriched by a number of non-traditional studies, i.e. those applying statistical tools to the analysis of style. Although referred to as non-traditional, the methods in question have their own and quite long tradition, begun as early as in the 1880s, with studies by Augustus de Morgan, Conrad Mascol, and Thomas Mendenhall (cf. Holmes 1998: 112; Rudman 1998: 354). Seminal founders of the discipline also include the inventor of the term *stylometry* and a scholar who proposed a new method of inferring the chronology of Plato's dialogues, Wincenty Lutosławski (1897; cf. Pawłowski and Pacewicz 2004). Even if his works today are known only to the most sophisticated experts in Plato, the impact of the term *stylometry* has never been questioned. It is usually used in a sense slightly broader than *non-traditional authorship attribution*: while the former denotes all statistical insights into style, the latter refers to a statistical approach to the question of authorship.

A part of quantitative linguistics, non-traditional authorship attribution is based on a number of linguistic and statistical assumptions, whether they are expressed explicitly by practitioners or not. We assume, then, that some linguistic features (e.g. lexical richness, different word collocations or particular syntax preferences) can be measured by means of statistical procedures. Next, we assume that in a language, there is a subset of (traceable) linguistic features dependent on an individual idiolect rather than shared by writers of the same epoch, genre, gender, etc. In a word, we believe that some features of a written text can betray the person who wrote it, despite his/her aesthetic, social, or historical conditions. Such a concept of an authorial *fingerprint* will be discussed below in detail. Again, there is also a silent — and somewhat naïve — assumption that texts in a corpus are purely "individual" in terms of being written solely by one author and not influenced by other writers — as if any text in the world could be created without references to the author's predecessors and to the whole literary tradition.

When it comes to the statistical background of stylometry, we silently assume that a sample text (e.g. a novel) written by a known author might serve as a representation of his/her "language" in general. Hence, given an anonymous text to be attributed and a comparison corpus of known texts, one believes that the degree of similarity between samples reflects real dependencies between authorial idiolects (while in fact what one deals with is only an approximation). Also, using statistics, we *ipso facto* accept the assumption that dealing with frequent phenomena, e.g. function words, leads to much more reliable conclusions than analyzing rare occurrences, such as hapax legomena. Next, most of the methods we use rely on the assumption that the units analyzed (e.g. words in a corpus) are distributed according to a Gaussian "normal" distribution, although word frequency distributions are not Gaussian at all (cf. Baayen 2001). Last but not least, there is a strong yet silent assumption that grammatical/lexical phenomena are independent of each other, as if words were not affected by their neighboring lexical context (cf. Argamon 2008: 140–144).

It becomes quite obvious that most of the assumptions mentioned above cannot be fulfilled in a quantitative analysis of natural language. At the same time, a large number of successful authorship experiments show that the techniques applied in stylometry indeed c a n d i s c r i m i n a t e analyzed texts according to authorial idiolects, and that the reasoning about the authorship on the basis of statistical analysis of style i s p o s s i b l e (with some caveats).

The most convincing proof that a computer-assisted stylometry really works is provided by several controlled attribution tests (Burrows 2002; Hoover 2004a, 2004b; Juola 2006; Juola and Baayen 2005; Jockers et al. 2008, 2010; Eder 2010; Rybicki and Eder 2011; Smith and Aldridge 2011, etc.). The general conception of such a controlled benchmark is to collect a corpus of texts written by known authors only, and then to perform a series of blind tests for authorship. Leaving the technical details aside, the way of testing is simple: the more samples are "guessed" correctly (in terms of being linked to their actual authors), the more accurate a given methodology. A very similar controlled benchmark will be applied below for examining different linguistic features as discriminators of authorial uniqueness.

There are dozens of statistical techniques applied to the study of language (cf. Baayen 2008; Gries 2010) but only a couple have been used in authorship attribution. Some of them focus on one phenomenon carefully retrieved from a corpus — as do different indexes of vocabulary richness or measures of the degree of rhythmicity — while others rely on a large number of features computed at once. The latter are called multidimensional and they are claimed to be much more sensitive to nuanced differences between samples. The reason for their attributive value is the fact that they aggregate the impact of many linguistic features of individually weak discriminating strength (cf. Nerbonne 2007: xvii).

Multidimensional methods include Principal Components Analysis, Factor Analysis, Cluster Analysis, Multidimensional Scaling, Discriminant Analysis, Support Vector Machines, Nearest Shrunken Centroids, and Burrows's tests for attribution: Delta, Zeta and Iota (cf. Burrows 1987, 2002; Hoover 2001, 2003, 2004a, 2004b; Argamon 2008; Juola 2006; Jockers et al. 2008, 2010, etc.). A detailed explanation of the mathematical foundations of these methods would be out of topic here; this problem will be addressed in another study, with an attempt at a new, made-to-measure method of attribution.

Some of the above multidimensional approaches produce attractive and comprehensive graphs (an example of such an approach is shown in Fig. 1 below) but they are not sufficiently accurate in real attribution experiments. Other techniques perform a classification of input samples, resulting in binary results: a given sample is always linked to one (and only one) candidate for authorship. For this reason, they are ideally suitable for controlled attribution tests.

A considerably high efficiency gained by some of the statistical techniques — even if some fundamental methodological assumptions are broken or not clearly obeyed — leads to at least two important conclusions. On the one hand, this may reflect a potentially strong attributive power of the methods in question. On the other hand, it suggests that we should not discard experiments which are counter-intuitive or have no convincing theory supporting the underlying hypotheses. Especially the latter statement will play some role in the present study.

2. The Stylistic Fingerprint

In non-traditional authorship attribution, the choice of an appropriate statistical method is one of the most important questions to be solved. This is a question of "how to measure style?" Perhaps even more important, however, is the question of "what to measure?", sometimes somewhat underestimated. While the first problem is of mathematical and/or computational nature, the latter is purely linguistic. In other words, the question arises which language features (if any) can betray a measurable difference in unique authorial style.

A search for authorial traces in a disputed text, however, should be preceded by a more general question whether the style is determined by an individual, as certainly is human DNA code, fingerprint, or patterns in one's iris. One of the founders of stylometry, Wincenty Lutosławski, has claimed the answer to be positive, and he compared the style to handwriting, assuming its uniqueness: "If handwriting can be

so exactly determined as to afford certainty as to its identity, so also with style, since style is more personal and characteristic than handwriting" (Lutosławski 1897: 66). Even if today's practitioners are far from such certainty, there is still strong belief that the process of writing is somehow affected by an unconscious personal stamp. The main problem, then, is to trace this stamp, or *stylistic fingerprint*, among the infinite number of linguistic (lexical, morphological, syntactic, etc.) features.

The correlation between particular language features and stylistic idiosyncrasy is quite a delicate and uneasy balance between what is common and what is unique in the language. Extracting as many unique elements of style as possible is the goal of authorship attribution. These identified and extracted stylistic features are referred to as *style-markers*. Usually, the most desired style-markers are those undiscoverable with the naked eye and thus beyond authorial control; this is because the real "uniqueness" in style should be resistant to imitation, plagiarism and parody.

In the long history of non-traditional attribution studies, a variety of style-markers have been examined, with better or worse results. In the first place, a vector of at least 100 most frequent words was used in a vast majority of recent approaches. Next go other style indicators, such as sentence length, word length, rhythmical patterns of stressed and unstressed syllables, vocabulary richness, frequencies of the most common function words, punctuation marks, common word collocations, frequencies of certain letter sequences, bi-grams of syntactic labels, and so on (for further discussion cf. Mosteller and Wallace 2007 [1964]; Holmes 1998; Baayen et al. 2002; Hoover 2002; Pawłowski 2003; Juola 2006; Grieve 2007; Hirst and Feiguina 2007).

The striking thing is, however, that there have been no attempts to compare the attribution effectiveness of these style-markers. A rare exception is a study by Hoover (2002), where the author compares the discriminating strength of (1) the most frequent words, and (2) frequent word collocations. Again, there are no cross-language benchmarks that would verify the silent assumption that style-markers are universal — i.e. that markers suitable for English would be as good (or as bad) for other languages. However, a prototype of such a benchmark has been proposed (cf. Juola 2009).

3. Motivation: Garbage In, Gospel Out?

Intuitively, if a stylistic fingerprint really exists, it should be somehow correlated with either lexical richness of a given text, or with different grammatical categories preferred by an author. In other words, one can safely assume that a hypothesis of authorship would be supported by a careful comparison of the usage of words within a corpus, or syntactic features, or other plausible grammatical measures rather than mere chunks of letters or punctuation marks.

However, a certain experiment, performed on a defective corpus, suggests that we are still far away from understanding where the authorial fingerprint might be hidden. The goal of the experiment in question was to test the effectiveness of a couple of nearest neighbor classification methods applied to ancient Greek epic poems. To avoid possible coding errors with non-Latin alphabet, the author converted the corpus into Beta-Code, i.e. a very popular way of representing Greek letters, accents and breathing

marks in a special markup language. The Beta-Code was developed in the 1970s and its main advantage is that it allows to write classical Greek on ASCII terminals (cf. Crane 2004: 51). This is done by transliterating the letter characters and replacing the accents and breathing marks with special control codes. The example below shows the opening line of the *Iliad* by Homer in Unicode and after conversion to Beta-Code:

(1) Μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος (Unicode)
(2) *mh=nin a)/eide qea\ *phlhi+a/dew *)axilh=os (Beta-code)

Even if all the samples were converted correctly, an ominous bug in the software affected the next stage of pre-processing. In consequence, the final text representation was heavily distorted, because all the control characters were treated as word-delimiters and thus the samples were split into quasi-random sequences of letters before and after the accents and/or breathing marks. The above Homeric line, then, turned into a sequence of following nonsense word-fragments:

(3) mh nin a eide qea phlhi a dew axilh os (Beta-code damaged in pre-processing)

The author was not aware of the problem, hence the experiment was not terminated and the corpus was analyzed by means of several methods used in stylometry. Certainly, the bug would have been discovered immediately if the results had been odd. Instead, all the methods gained very high accuracy: the samples were linked correctly to their actual authors, and no mistakes were noticed. The sample results computed with cluster analysis and plotted as a consensus tree (for more on this method, cf. Eder and Rybicki 2011) are shown on Fig. 1.

Fig. 1. Results of a controlled attribution experiment performed on damaged Greek samples. Despite a severe systematic error in the corpus, all the samples are clustered correctly on proper "branches" of a consensus tree

As one can easily notice, all the samples are clustered correctly and the Homeric texts are significantly detached from those by Nonnus, Apollonius, Hesiod, and Aratus. Hesiod's *Works and Days* were "guessed" to be similar to his *Theogony*, three different samples from Aratus' *Phainomena* moved close one to each other, and so on. In a word, authorial uniqueness was recognized across the 30 samples analyzed. Now, the success with k n o w n samples allows a reliable reasoning about the recognition of a n o n y m o u s ones. In this particular example, this is the case of *Batrachomyomachia*, an epic poem for centuries attributed to Homer. It is quite obvious in this graph that this poem has nothing in common with the assumed author of the *Iliad* and the *Odyssey*.

It is hard to believe that valid information about authorial uniqueness could be provided by a severely corrupted corpus. Even more counter-intuitive is the fact that when the bug was fixed, the obtained results became s l i g h t l y w o r s e. This effect is quite difficult to explain on theoretical grounds but it leads to the conclusion that further investigation of the effectiveness of style-markers not purely linguistic, or even markers without any linguistic relevance, might be promising. (On the other hand, the corpus in question was not so deeply damaged — despite a large amount of random noise obtained, many words were split more or less through their morphemic boundaries; if that was the reason of unexpected effectiveness of the attribution, morpheme-based markers deserve further detailed investigation).

4. Setting the Experiment

Since there is no premise — apart from the assumptions mentioned above — for favoring any kind of style-markers, the only feasible way of verifying their value seems to be an empirical investigation. The experiment presented below is an attempt at such a comparison of stylistic discriminators. Certainly, due to the infinite number of imaginable features of a written text, the investigation has to be limited to only a few. It will be focused, then, exclusively on those markers that can be retrieved from a non-lemmatized corpus of plain text files. Also, only those markers will be considered which are machine-countable and can be represented as relative frequencies, i.e. statistically normalized occurrences in a text. Other linguistic evidence (syntax, parts of speech, etc.), although potentially very promising in search of the authorial fingerprint (cf. Hirst and Feiguina 2007), will not be addressed here. The aim of this approach is, then, to test how much authorial information can be retrieved from a raw string of characters in a text sample, without any kind of annotation, parsing, information retrieval, or keyword extracting.

The markers chosen are as follows: (1) the most frequent words, a commonly-used and absolutely classical solution (cf. Mosteller and Wallace 2007 [1964]; Burrows 2002; Hoover 2001, 2003); (2) word bi-grams; (3) word tri-grams; (4) word tetra-grams; (5) letter bi-grams, or 2-character sequences of letters including spaces; (6) letter tri-grams; (7) letter tetra-grams; (8) letter penta-grams; (9) letter hexa-grams. Last but not least, a few combinations of the above features will be used: (10) a combination of words and word bi-grams (cf. Hoover 2002; Jockers and Witten 2010: 218); (11) a combination of words and letter penta-grams. Some preliminary tests

with other combined style-markers were abandoned due to unexpected methodological obstacles (this problem will be skipped here due to its mathematical rather than linguistic nature).

The procedure of splitting the input texts into word and/or letter n-grams was preceded with replacing all the punctuation marks with spaces; thus, the spaces became quite convenient word-delimiters. Consequently, the opening sentence of Jane Austin's *Pride and Prejudice* "It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife" would be split into different style-markers as follows (here, hyphens represent spaces):

(4) "it" "is" "a" "truth" "universally" "acknowledged" "that" "a" …
(single words)
(5) "it-is" "is-a" "a-truth" "truth-universally" "universally-acknowledged" …
(word bi-grams)
(6) "it" "t-" "-i" "is" "s-" "-a" "a-" "-t" "tr" "ru" "ut" "th" "h-" "-u" …
(letter bi-grams)
(7) "it-" "t-i" "-is" "is-" "s-a" "-a-" "a-t" "-tr" "tru" "rut" "uth" "th-" …
(letter tri-grams)
(8) "it-is-" "t-is-a" "-is-a-" "is-a-t" "s-a-tr" "-a-tru" "a-trut" "-truth"…
(letter hexa-grams)

Next, the retrieved character strings (i.e. assumed style-markers) were counted and the obtained numbers were converted to relative frequencies (i.e. percentages), due to possible differences in length between samples. For each experiment, current style-markers' frequencies were collected in a matrix similar to the table shown below (this particular table represents letter tri-grams' frequencies):

	ABronte_ Agnes	Austen_ Emma	CBronte_ Jane	Conrad_ Lord	Dickens_ Bleak	…
"-th"	1.4901	1.3523	1.5002	1.8030	1.5110	…
"the"	1.2484	1.0411	1.1717	1.4986	1.1899	…
"he-"	0.9835	1.0204	1.1500	1.4710	1.0677	…
"-an"	0.8949	0.7362	0.8166	0.6324	0.8086	…
"nd-"	0.8195	0.6858	0.8380	0.6497	0.8069	…
"er-"	0.6849	0.7428	0.6683	0.4968	0.6282	…
⋮	⋮	⋮	⋮	⋮	⋮	⋱

5. Corpora Used

To make the results more reliable, a series of parallel attribution experiments were performed on corpora in four languages. The corpora were roughly similar (ca. 70 prose texts by ca. 20 authors). The languages chosen were: (1) English, as an obvious reference point and a language thoroughly examined in many attribution tests;

(2) Latin, as a language highly inflected and an example of the Romance languages; (3) Polish, as one of the Slavonic languages; (4) German, as a language with many compound words, which makes German words longer than those in other languages — this might carry some import while comparing the effectiveness of attribution based on letter n-grams with that based on words.

Certainly, an experiment performed on such a limited number of languages will not allow to formulate universal linguistic conclusions. However, some main regularities can be traced, and the present study might serve as a point of reference in broader multilingual investigations. Perhaps the most evident absence here is that of ancient Greek, which was introduced above and indeed has motivated the whole experiment. However, due to a very complex set of preliminary problems that have to be solved in that case (dealing with accents, contractions, ellisions, punctuation, etc., which were partially introduced by modern scholars), attribution of ancient Greek texts will be addressed in a separate study.

6. Method of Testing

The testing of the particular style-markers' effectiveness was done with Burrows's Delta, an easily-applicable and reliable platform for benchmarking in stylometry (Burrows 2002; Hoover 2004; Eder 2010; Rybicki and Eder 2011). The obtained results, however, as they were not directly correlated with the method used, should be also valid for a vast majority of other multidimensional statistics, such as Principal Components Analysis, Multidimensional Scaling, or Cluster Analysis. For every language and every style-marker tested, the same series of controlled attribution experiments were performed.

Delta, like any other multidimensional method, is sensitive to the number of linguistic features (words, bi-grams, etc.) analyzed. Although this choice of the appropriate vector of units is essential, there is no consensus among scholars, even if they agree that in general the most frequent words are better than other style-markers. While some practitioners claim that a small number of function words provides best results (Mosteller and Wallace 2007 [1964]), other prefer longer vectors: 100 most frequent words (Burrows 2002), 300 (Smith and Aldrigde 2011), 500 (Craig and Kinney 2009), up to even 1000 or more (Hoover 2004). However, a multi-corpus and multi-language study (Rybicki and Eder 2011) shows that there is no universal vector of words and that the results are strongly dependent on the corpus analyzed.

Thus, to avoid fuzziness with unconvincing results, a bootstrap-like approach was applied in the present study (cf. Good 2006; Eder and Rybicki 2012). The general goal of such a procedure is to test a large number of randomly chosen permutations of the original data and to estimate an average score. Here, the number of units to by analyzed was chosen randomly in 1000 trials (e.g., 334, 138, 372, 201, 104, 145, 134, 462, …) and, for each trial, a nearest neighbor classification was performed, resulting in a percentage of correctly "guessed" authors. A mean of the obtained results was then recorded for each series of 1000 trials. The same was done for six ranges of the most frequent units: 30–100, 100–500, 500–1000, 1000–2000, 2000–5000, and 5000–10,000 words; and then, the whole procedure was repeated with the next language corpus. Rather important advantage of this approach is that even very small

discrepancies between final scores become statistically significant. The procedure described above results in 6000 single attribution tests for each style-marker, 66,000 tests for each language, and over 0.25 million tests in total.

Computing such a large number of attribution tests would not be possible in any imaginable amount of time with commercial statistics software. Instead, an adjustable open-source statistical environment R was used (http://cran.r-project.org) together with a tailored multi-task R script developed by the present author. The script, supplemented with a Tcl/Tk graphical user interface written by Jan Rybicki, performs myriads of iterated attribution tests in a fraction of the time needed by other tools (cf. Eder and Rybicki 2011). It was possible to compute the above number of 0.25 million iterations within a couple of hours rather than months. The script did all the work, including the pre-processing, splitting the input texts into style-markers, preparing tables of frequencies, applying the attribution tests, generating the graphical output (if applicable), and summarizing the final results of the classification.

7. Results

The analysis of the obtained results should begin with a comparison of different word constellations as style-makers. In every test in this part, then, the simplest and commonly known solution — frequencies of frequent words — turned to be the most effective (Fig. 2, black bars). For English prose, the accuracy reached 100% when a very long vector of words (some 7500 from the top of frequency list) was analyzed. Similarly satisfying results did not prove achievable either for the three remaining languages or for the remaining style-markers. Noticeably, the accuracy of "guessing" in the Polish corpus was significantly worse than for English, no matter which style-markers were considered, and slightly worse for both German and Latin. This kind of divergence between languages was observed — on a smaller scale — in recent studies (Rybicki and Eder 2011; Eder 2010) with a suggestion that a degree of inflection between the tested languages might be of some importance.

A comparison of word bi-grams and especially word tri-grams across languages was surprising to say the least. The differences were substantial if not fundamental (Fig. 2, dark grey and light grey bars). In English, the effectiveness increased with the number of the most frequent bi- or tri-grams tested, following the increasing "guessing" power of single words. Only for long vectors of word tetra-grams (those above 2000 units), the accuracy of attribution was dwindling. In other languages word collocations proved totally useless as style-markers. Especially in Latin, tri-grams and tetra-grams were completely "blind" in recognizing authorial uniqueness (Fig. 2d, light grey bars). Word bi-grams, even if slightly better, were still far from any acceptable level of accuracy. The results for Polish and German were not as bad as for Latin, but definitely the same phenomenon could be observed (Fig. 2b, c).

It is quite difficult to explain why word pairs do not discriminate Latin authors as compared to single words. It is even more difficult to find a plausible explanation of such a substantial divergence between English and Latin. Perhaps the reason is inherently connected with the nature of both languages, possibly with the considerably high number of

phrasal verbs (i.e. natural word bi-grams) in English? Or maybe this is about the rigorous word-order in English as compared to Latin? Or, could the degree of inflection, weak in English and essential in Latin, play some role here? Whatever it is, it deserves further investigation. Due to their very simple nature, style-markers applied in the present study are not suitable for testing the impact of syntax and/or inflection in authorship attribution. For such an experiment, parts-of-speech annotated corpora should be used.

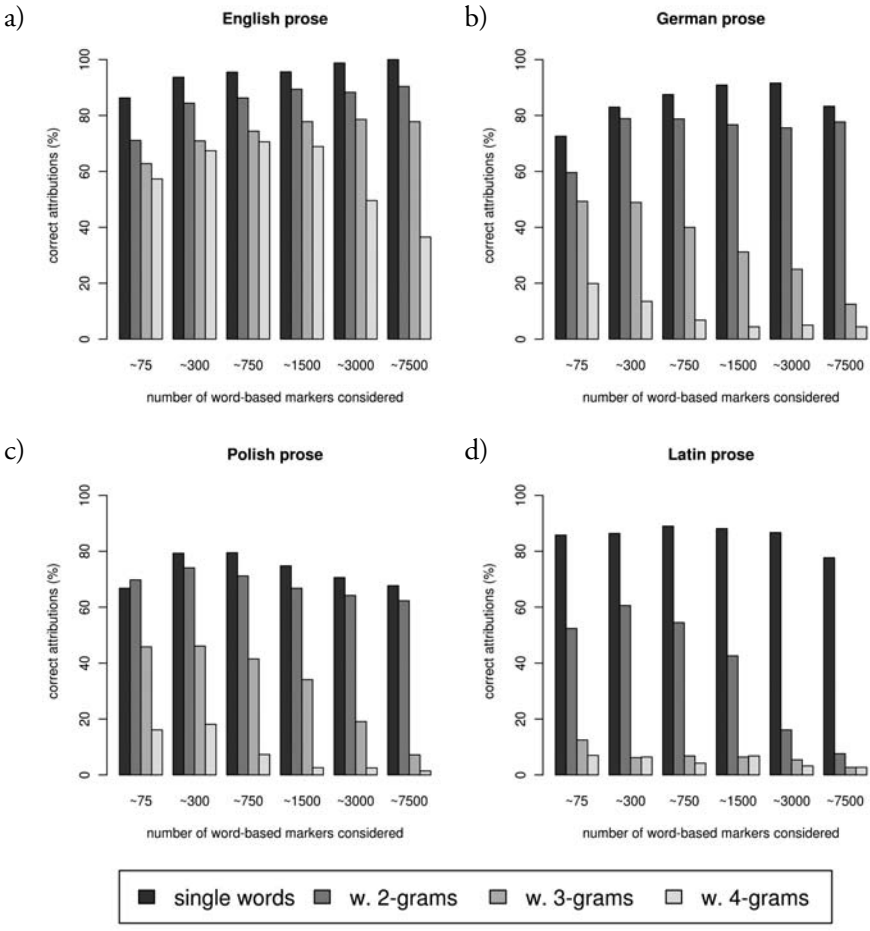

Fig. 2. Words and word n-grams as style-markers in four languages: (a) English, (b) German, (c) Polish, (d) Latin

The experiments with various letter n-grams were less exciting (Fig. 3). The results were rather similar for all letter-based style-markers in each language tested: the letter bi-grams proved slightly weaker discriminative strength than other features, then came letter tri-grams (the number of available bi-grams is limited, hence they are not represented above the range of 1000 units). The effectiveness of letter tetra-, penta- and hexa-grams was similar, and quite satisfying. In English, letter n-grams were slightly less accurate than

single words, but in Latin, German and Polish — letter tetra-grams or penta-grams occasionally reached better scores than words alone (cf. Fig. 2, black bars, and Fig. 3, grey bars). This was noticeable mainly for long vectors of units analyzed: i.e. in the Latin corpus, the accuracy achieved by the most effective word-based markers was as high as 85.8%, 86.4%, 89%, 88.1%, 86.7%, and 77.7% in consecutive ranges, while the same vectors of the letter penta-grams provided "guesses" at the levels of 78.1%, 85.3%, 89.7%, 90.7%, 90.4%, and 90.4%, respectively. In the Polish corpus, the best score ever obtained (81.3%) was produced indeed by the letter-based markers, namely hexa-grams (cf. Fig. 3c, light grey bars). This was also the case in the German corpus, where the highest efficiency (93.5%) was gained by a very long vector of letter penta-grams; then came letter tri-grams, which also turned to be considerably strong discriminators (cf. Fig. 3b). The correlation — if there is any — between the most effective letter n-grams and the average length of word in a given language should be examined further.

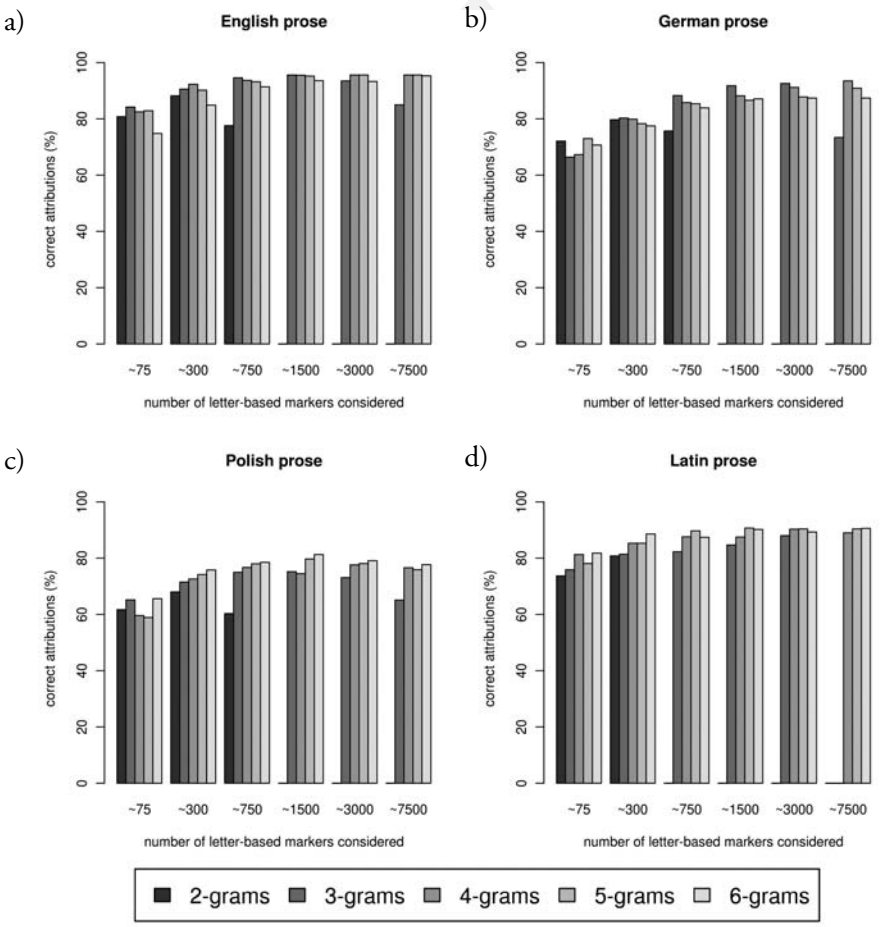

Fig. 3. Letter n-grams as style-markers in four languages: (a) English, (b) German, (c) Polish, (d) Latin

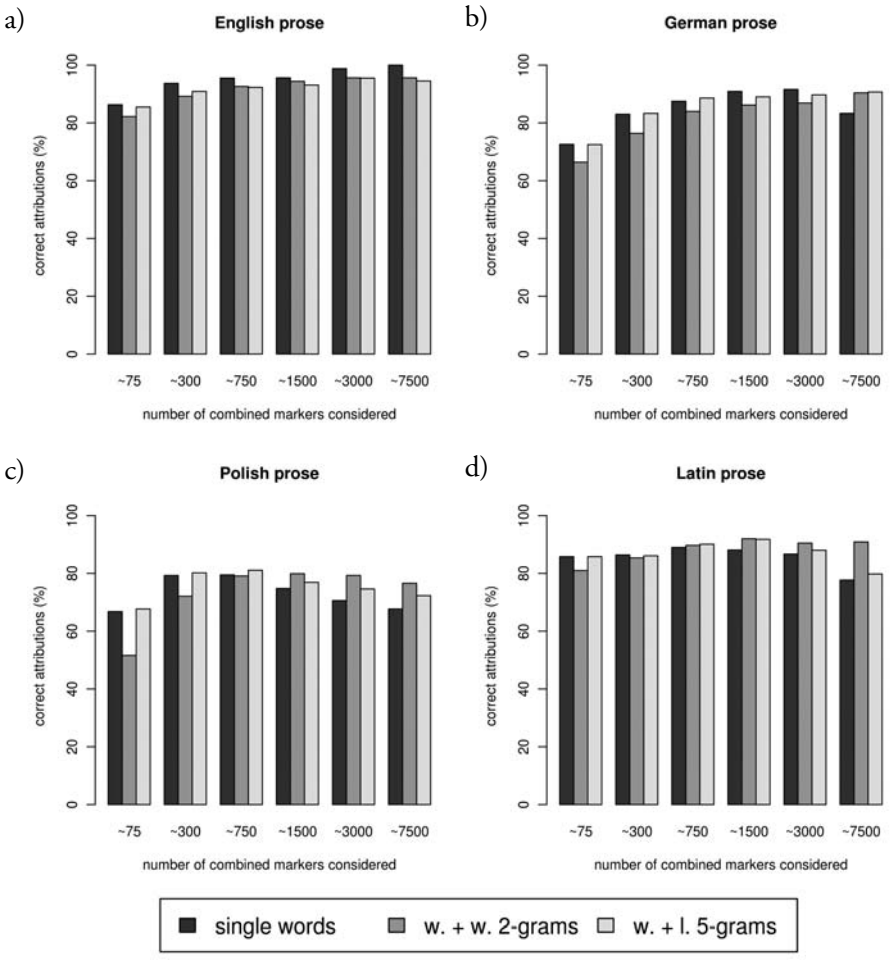

Fig. 4. Some combinations of style-markers in four languages: (a) English, (b) German, (c) Polish, (d) Latin

The last set of experiments focused on style-markers composed of units of different nature, combined in a single sample. Among a few combinations tested, best results were obtained for (1) a combination of words and word bi-grams, (2) a combination of words and letter penta-grams. For all the languages tested, their effectiveness was very good indeed (Fig. 4). The only exception was the English corpus, where both combined style-markers were just a little worse discriminators than single words. For German, the "guessing" strength of the combined markers was quite parallel to that displayed by single words, and slightly worse than the various letter-based markers could achieve. For Polish, the differences were even less significant: no matter if single words, letter n-grams, or combined markers were used, the results were similar yet not really impressive. Finally, in the Latin corpus, the two combined style-markers

reached their best scores for this language (92% and 91.8%), better than any word- or letter-based marker individually.

Last but not least, when carrying out a detailed analysis of the final results, one can easily miss a general observation of a great importance. Namely, the only really significant difference between style-markers' discriminative strength was that of the word n-grams as compared to single words. For all the other markers tested, the differences in their attributive strength rarely exceeded 5%. At the same time, the differences in "guessing" across languages could be as high as 40% (as between Polish and English). This is disappointing, and it means that elaborating the ideal set of style-markers may be just spoiled by choosing a poorly attributive language (certainly, since there are no differences in people's brains, the statement of attributive weakness of some languages sounds oddly; however, the above results clearly suggest such a conclusion). What is worse, some studies show that even the choice of particular samples in a comparison corpus might significantly affect the final results (Eder and Rybicki 2012). On the other hand, since we have no key to the attributive weakness of some languages, every little improvement of their efficiency is priceless.

8. Conclusion

The final results confirm the commonly-accepted intuition that lexical measures are generally a good choice for multidimensional approaches to authorship attribution — but this applies to English language only. Here, the best style-marker is a mass of single words. The situation is much more interesting (or confusing?) in other languages. The silent assumption that style-markers are language-independent turned to be unfounded: in languages other than English, alternative markers are shown to be usually more effective (cf. Fig. 3 and 4). The problem is, however, that we do not have prior knowledge of the preferred marker and we do not know why the best choice, for Latin, might be a combination of words and letter penta-grams, for Polish — words plus word bi-grams, and for German — simple letter tri-grams (on the other hand, some of the differences between the final scores are not statistically significant). Although there is no theoretical explanation of this dilemma, the markers alternative to simple words are doubtless worth considering. Certainly, the achievable improvements in attribution accuracy are not very high — e.g. in the Latin corpus, words revealed a maximum of 89% authors, while combination of words and word bi-grams shows the effectiveness of up to 91.8%. However, this is quite a lot in a real authorship attribution investigation.

The results also confirmed the well-known phenomenon: attribution effectiveness grows with the number of words analyzed, and at a certain point it tends to stabilize or slightly decreases (cf. Hoover 2004; Rybicki and Eder 2011; Smith and Aldridge 2011). In the present study, this observation can be extended to style-markers other than single words, too.

The main conclusion, however, could be formulated after a comparison of word bi-grams, and especially word tri-grams, in the English and the Latin corpora. The discrepancy evident between the low attribution accuracy for Latin and the considerably good effectiveness for English is very obvious and statistically significant. This leads to a claim concerning the importance of some syntax-based markers to be applied in later studies.

References

Argamon Shlomo (2008): Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. — *Literary and Linguistic Computing* 23, 131–147.

Baayen Harald (2001): *Word Frequency Distributions*. — Dordrecht: Kluwer.

Baayen Harald (2008): *Analyzing Liguistic Data: A Practical Introduction to Statistics Using R*. — Cambridge: Cambridge University Press.

Baayen Harald, Van Halteren Hans, Neijt Anneke, Tweedie Fiona (2002): An Experiment in Authorship Attribution. — *Proceedings of JADT 2002*, Université de Rennes, St. Malo, 29–37.

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, 267–287.

Craig Hugh, Kinney Arthur, eds. (2009): *Shakespeare, Computers, and the Mystery of Authorship*. — Cambridge: Cambridge University Press.

Crane Greg (2004): Classics and the Computer: An End of the History. — [In:] *A Companion to Digital Humanities*, Susan Schreibman, Ray Siemens and John Unsworth (eds.), 46–55. Oxford: Blackwell.

Eder Maciej (2010): Does Size Matter? Authorship Attribution, Small Samples, Big Problem. — *Digital Humanities 2010: Conference Abstracts*, King's College London, 132–135.

Eder Maciej, Rybicki Jan (2011): Stylometry with R. — *Digital Humanities 2011: Conference Abstracts*, Stanford University, Stanford, CA, 308–311.

Eder Maciej, Rybicki Jan (2012): Do Birds of a Feather Really Flock Together, or How to Choose Test Samples for Authorship Attribution. *Literary and Linguistic Computing* 27 (forthcoming).

Good Philip (2006): *Resampling Methods: A Practical Guide to Data Analysis*. — Boston–Basel–Berlin: Birkhäuser.

Gries Stefan (2010): *Statistics for Linguistics with R: A Practical Introduction*. — Berlin: De Gruyter Mouton.

Grieve Jack (2007): Quantitative Authorship Attribution: An Evaluation of Techniques. — *Literary and Linguistic Computing* 22, 251–270.

Hirst Graeme, Feiguina Ol'ga (2007): Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. — *Literary and Linguistic Computing* 22, 405–417.

Holmes David (1998): The Evolution of Stylometry in Humanities Scholarship. — *Literary and Linguistic Computing* 13, 111–117 .

Hoover David L. (2001): Statistical Stylistic and Authorship Attribution: an Empirical Investigation. — *Literary and Linguistic Computing* 16, 421–444.

Hoover David L. (2002): Frequent Word Sequences and Statistical Stylistics. — *Literary and Linguistic Computing* 17, 157–180.

Hoover David L. (2003): Multivariate Analysis and the Study of Style Variation. — *Literary and Linguistic Computing* 18, 341–360.

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

Jockers Matthew, Witten Daniela, Criddle Craig (2008): Reassessing Authorship in the Book of Mormon Using Delta and Nearest Shrunken Centroid Classification. — *Literary and Linguistic Computing* 23, 465–491.

Jockers Matthew, Witten Daniela (2010): A Comparative Study of Machine Learning Methods for Authorship Attribution. — *Literary and Linguistic Computing* 25, 215–223.

Juola Patrick (2006): Authorship Attribution. — *Foundations and Trends in Information Retrieval* 1, 233–334.

Juola Patrick (2009): Cross-linguistic Transference of Authorship Attribution, or Why English-only Prototypes Are Acceptable. — *Digital Humanities 2009: Conference Abstracts*, University of Maryland, College Park, MD, 162–163.

Juola Patrick, Baayen Harald (2005): A Controlled-corpus Experiment in Authorship Identification by Cross-entropy. — *Literary and Linguistic Computing* 20 (Suppl. Issue), 59–67.

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

Lutosławski Wincenty (1897): *The Origin and Growth of Plato's Logic: With an Account of Plato's Style and of the Chronology of his Writings*. — London: Longmans.

Mosteller Frederick, Wallace David (2007 [1964]): *Inference and Disputed Authorship: The Federalist*. Reprinted with a new introduction by John Nerbonne. — Stanford: CSLI Publications.

Nerbonne John (2007): The Exact Analysis of Text. — [Foreword in:] Mosteller and Wallace (2007 [1964]), xi–xx.

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

Pawłowski Adam, Pacewicz Artur (2004): Wincenty Lutosławski (1863–1954). Philosophe, helléniste ou fondateur sous-estimé de la stylométrie? — *Historiographia Linguistica* 21, 423–447.

Rudman Joseph (1998): The State of Authorship Attribution Studies: Some Problems and Solutions. — *Computers and the Humanities* 31, 351–365.

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

Smith Peter, Aldrigde W. (2011): Improving Authorship Attribution: Optimizing Burrows's Delta Method. — *Journal of Quantitative Linguistics* 18, 63–88.