

MACIEJ EDER

PEDAGOGICAL UNIVERSITY OF KRAKÓW

POLISH ACADEMY OF SCIENCES, INSTITUTE OF POLISH LANGUAGE

Computational stylistics and Biblical translation: how reliable can a dendrogram be?

ABSTRACT. In the present study, two versions of the New Testament – Greek original and its Latin translation known as the Vulgate – are compared using stylometric methods. Although the study addresses some questions concerning stylistic differentiation between particular books, the main aim is to discuss the problem of reliability in stylometry. Last but not least, a simple way of improving reliability of cluster analysis plots using resampling of input data is introduced.

KEY WORDS. Computational stylistics, New Testament, reliability, cluster analysis, bootstrap

1. Introduction

Computational stylistics, also referred to as stylometry, has been traditionally focused on the problem of authorship attribution, i.e. the question whether hidden stylistic idiosyncrasies, traceable with advanced statistical procedures, might betray the person who wrote a disputed or anonymous literary text. Approaches to the problem of authorial “fingerprint” have quite a long tradition, dating back to studies by Augustus de Morgan, Conrad Mascol, and Thomas Mendenhall conducted as early as in the 1880s (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).

Introduced in the pre-computer era, stylometric methods gained their popularity rather slowly through the decades of the 20th century. The first to use them were mathematicians, quantitative linguists, computer scientists, i.e. scholars with scientific background rather than humanities-oriented researchers. However, it was only after Burrows published his seminal study on Jane Austen (Burrows 1987) when stylometry has become known to a broader circle of literary scholars. Indeed, the techniques used in authorship attribution can be easily generalized into a variety of issues in literary studies, such as diachronic investigations in style change, studies in genre recognition, literary inspirations, etc. Last but not least, the methods in question have been also applied in the area of translation studies. It was again Burrows who published a ground-breaking study on English translations of Juvenal (Burrows 2002a); since then, computational stylistics applied to translation studies has been thoroughly

examined and extended by Rybicki (2006, 2011, Heydel & Rybicki 2012), to name just a few studies.

The methods adopted or introduced by Burrows, Hoover, Craig, and others (Burrows 1987, 2002b, Hoover 2003a, Craig & Kinney 2009, etc.) were very intuitive and easily-applicable to literary studies. These are Principal Components Analysis, Cluster Analysis, Zeta and Iota. Despite their limitations (the lack of validation of the obtained results being the most obvious), they are still widely used. The awareness of their pitfalls is rarely demonstrated by the humanists, though.

On the other hand, the hard science has elaborated a number of well-performing, sophisticated machine-learning algorithms, suitable for classification tasks, derived mostly from the field of biometrics, nuclear physics, or software engineering. They include Naïve Bayes Classifier, Support Vector Machines, Nearest Shrunken Centroids, or Random Forests, to name but a few (Mosteller & Wallace 2007 [1964], Koppel et al. 2009, Jockers et al. 2008, Tabata 2012). Being surprisingly accurate, at the same time they are much too sophisticated (in terms of mathematical complexity) to be understood by the humanists, and thus they are usually ignored in literary-oriented studies. What is worse, they are sometimes claimed to be unsuitable for stylometric investigations due to their alleged unreliability. To exemplify, Love argues that interpreting groupings of samples on scatterplots or dendrograms – i.e. using graphical explanatory methods – should always be preferred to “black-box” approaches, as he refers to machine-learning classification methods (Love 2002: 142–147). His statement that supervised classification “offers none of the ways of assessing reliability offered by statistical methods” (ibid., 146) shows how different is the usage of the word “reliability” among literary scholars and computational scientists.

Since the gap between the two stylometric worlds is hopelessly getting wider, there seems to be a need for elaborating and promoting straightforward extensions of the existing methodology that could be used by literary scholars. If Decision Trees turn to be unavailable for a typical humanist, and nice-looking Cluster Analysis plots (dendrograms) are not reliable enough, the third way is to combine the two approaches. Using algorithms derived from the state-of-the-art classification methods, and visualization from the old-school techniques (e.g. dendrograms) might be a compromise. The promising examples include probabilistic and geometric extensions of classic Delta as introduced by Argamon (2009), and bootstrap consensus trees as a way of improving reliability of Cluster Analysis dendrograms. The latter method, inspired by the study of Papuan languages by Dunn et al. (2005, quoted in Baayen 2008: 143–147), will be discussed in greater detail below.

2. Reliability in computational stylistics

The question of reliability in non-traditional authorship attribution has been extensively discussed by Rudman (1998a, 1998b, 2003), who formulated a number of caveats concerning corpus preparation, sampling, style-markers' selection, interpreting the results, etc. Rudman's fundamental remarks, however, had not been preceded by an empirical investigation. Experimental approaches to the problem of reliability include an application of recall/precision rates as a way of assessing the level of (un)certainly (Koppel et al. 2009), a study on different scalability issues in stylometry (Luyckx 2010), a paper discussing the short sample effect and its impact on authorship attribution reliability (Eder 2010), an experiment using intensive corpus re-composition to test whether the attribution accuracy depends on particular constellation of texts used in the analysis (Eder & Rybicki 2012), a study aimed to examine the performance of untidily prepared corpora (Eder 2012), and so on.

Sophisticated machine-learning methods of classification routinely try to estimate the amount of potential error that may be due to inconsistencies in the analyzed corpus. A standard solution here is a 10-fold cross-validation, in terms of 10 random swaps between two parts of a corpus: a subset of reference texts and a subset of texts used in the testing procedure. Although it is rather disputable if bare 10 cross-checks are enough to ascertain the results of real-life linguistic data (Eder & Rybicki 2012), the general idea of reassessing the corpus with a number of random permutations of variables is a big step forward in stylometric investigations. So far, this is the only way to identify local anomalies in textual data, i.e. any texts that are not "representative" enough for their authors' idiolects.

Unsupervised methods used in stylometry, such as Principal Components Analysis or Cluster Analysis, lack this important feature. On the other hand, however, the results obtained using these techniques "speak for themselves", which gives a practitioner an opportunity to notice with the naked eye any peculiarities or unexpected behavior in the analyzed corpus. Also, given a tree-like graphical representation of similarities between particular samples, one can easily interpret the results in terms of finding out which group of texts a disputable sample belongs to.

Hierarchical cluster analysis – as applied in the present study – is a technique which seeks for the most similar samples (e.g. a literary text, etc.) and builds a hierarchy of clusters, using a "bottom up" approach. It means that the procedure starts with pairing the nearest neighboring samples into two-element groups, and then recursively joins these groups into larger clusters. What makes this method attractive is a very intuitive way of graphical representation of the obtained results (see Fig. 1–3). However, despite obvious advantages,

some problems still remain unresolved. The final shape of a dendrogram highly depends on many factors, the most important being (1) a distance measure applied to the data, (2) an algorithm of grouping the samples into clusters, and (3) the number of variables (e.g. the most frequent words) to be analyzed. These factors will be briefly discussed below.

In a study of multivariate text analysis using dendrograms, Burrows writes: “my many trials suggest that, for such data as we are examining, complete linkages, squared Euclidean distances, and standardized variables yield the most accurate results” (Burrows 2004: 326). The distance used by Burrows is a widely accepted solution in the field of computational stylistics; there are no studies, however, expalining the principles of using this particular measure. Presumably, “standardized variables” mean, in this context, relying on *z*-scores (i.e. scaled values) rather than on relative word frequencies. If this is true, the distance used here is in fact equivalent to the Linear Delta measure introduced by Argamon (2009: 134), a slightly modified version of the classic Delta measure as developed by Burrows (2002b). Since the distance measure embedded in Delta proved to be very effective – a fact confirmed by numerous attribution studies – it should be also, by extension, applicable to hierarchical cluster analysis procedure. The choice of this particular measure, however, was neither explained on theoretical grounds, nor confirmed by empirical comparisons with other distances. Should a chosen measure follow the inherent characteristics of linguistic data, such as the Zipf’s law? Should the same distance be used to analyze inflected (e.g. Latin) and non-inflected (e.g. English) languages? These and similar questions have not been answered yet.

Another factor affecting the final shape of a dendrogram is a method of linkage used. In the above-cited statement, Burrows favours the complete linkage algorithm as the most effective one. We do not know, however, which were the other algorithms considered by Burrows, and we do not know what method of comparison was used to test their effectiveness. In a similar study, Hoover argues that the best performance is provided by Ward’s linkage (Hoover 2003b); his claim is confirmed by a concise comparison of Ward’s, complete, and average linkages. The Ward’s method is quite often used in quantitative linguistics, corpus linguistics, and related fields. Although it seems to be accurate indeed, there is no awareness that this method has been designed for large-scale tests of more than 100 samples: for the sake of speed, the optimal clustering was not a priority (Ward 1963: 236). What is worse, the state-of-the-art linkage algorithms seem to be ignored by stylometrists, probably because they are not implemented in standard statistical software. One might want to ask a question: if, say, neighbor-joining methods for reconstructing phylogenetic trees (Saitou & Nei 1987) were

supported by out-of-the-box commercial software, would text analytics still promote complete or Ward's linkage?

"Blind borrowing of statistical techniques from other disciplines must stop", claims Rudman (1998b: 355). This is certainly true, and it applies, *inter alia*, to the choice of linkage method. The real problem is, however, that stylometry has not developed its own linkage algorithm, and the methods derived from other fields have not been systematically tested on linguistic data. So far, then, we are at the mercy of existing procedures, for better and for worse.

Last but not least, the results of cluster analysis depend on a number of features (e.g. frequent words) to be analyzed. This drawback is shared by all multivariate methods relying on distance measures. The question how many features should be used for stylometric tests has been approached in many studies, but no consensus has been achieved: some scholars suggest using a small number of carefully selected words (function words), others prefer long vectors of words, and so on. Although all these solutions are reasonable and theoretically justified, the choice of the number of features is usually arbitrary. This problem is sometimes referred to as cherry-picking (Rudman 2003); and it will be addressed in the present study.

One important thing needs to be stressed at this point: the endless discussions concerning the preferred linkage algorithm, choice of distance measure etc. all betray (implicitly) the real issue at stake. Namely, dendrograms produced by hierarchical cluster analysis are *unstable* and very *sensitive* to any changes in a number of features and/or methods of grouping the samples.

3. Data and research questions

To address the question of authorial uniqueness of a literary text translated into another language, and to assess the problem of reliability of cluster analysis, a particular case of textual tradition has been chosen, namely the New Testament. As a typical sacred text, it is believed to be written under the inspiration of God; this reason alone makes the question of authorship of some disputed books (e.g. Epistles) to be very interesting, to say the least. Also, as a sacred text, the New Testament requires special attention to be paid by its translators: the text has to be rendered with a rigid precision. This feature of Biblical translations gives us an opportunity to conduct a very interesting cross-language comparison, because different language versions are perfectly parallel.

The study will examine two versions of the New Testament: the Greek original and its Latin translation by St Jerome, commonly known as the Vulgate. Since the New Testament

consists of texts written by several authors, the study attempts to answer three different yet related questions:

- (1) Are the particular authors of the Gospels, Epistles, etc. recognizable in the Greek original?
- (2) Are the original authorial traces noticeable also in the Latin translation?
- (3) Are the differences (if there are any) between authors as strong in the translation as in the original?

The third question is based on the assumption – also known as the leveling-out hypothesis as formulated by Baker (2004) – that texts translated into a given language are generally more similar to each other than texts originally written in the language in question (in other words: translating usually weakens stylistic nuances noticeable in the original).

The aim of the present study, however, is to identify some pitfalls of multivariate analysis rather than to answer explicitly the above questions concerning similarities or dissimilarities between particular samples of the Holy Scripture. Modern scholarship has been approaching the problem of authorship of the New Testament for centuries (Helms 1997, Guthrie 1990, Brown 1997); there were also some stylometric studies addressing this issue (Kenny 1986, Greenwood 1995, Ledger 1995, etc.). It can be safely assumed, then, that the problems concerning the authorship of subsequent books of the New Testament have been thoroughly examined from linguistic, historical, theological, and rhetorical points of view. For this reason, the Scripture seems to be an ideal material for stylometric benchmarks, because the traditional scholarship can serve as a straightforward validation of the results obtained by using the computational approach.

The above remark applies also to the Latin version of the Bible. There is a strong agreement in biblical studies that St Jerome rendered the Old Testament from the Hebrew original, having previously translated some passages from the Septuagint. As to the New Testament, scholars are rather unanimous that St Jerome did not translate the whole text from scratch but rather revised and corrected existing translations, commonly referred to as *Vetus Latina* (Nautin 1986). In the following benchmarks, the facts determined by traditional scholarship will serve as a good point of reference.

Some books of the New Testament are rather too short for being approached with multivariate analysis; thus, a reasonable selection of the whole material has been collected: the Synoptic Gospels (Matthew, Mark, Luke), the Gospel of John, the Acts, a selection of Pauline Epistles (First Corinthians, Second Corinthians, Romans), James's Epistle, and the Revelation. All the tests have been performed twice: for Greek original, and for Latin

translation. The discussion presented below, however, focuses basically on the Greek version. The results for the Vulgate are briefly commented on in the final section of this paper.

4. The experiment

To approach the question of stylistic differentiation between particular books of the New Testament, a number of plots using different linkage algorithms and/or different distance measures have been generated. As expected, the obtained dendrograms were substantially heterogeneous – three examples (out of many) are shown on Fig. 1–3. Usually, even a little change in the settings affects the final results. Without deciding (yet) which dendrogram is more likely to be “true”, one has to admit that the particular groupings are quite unstable, to say the least.

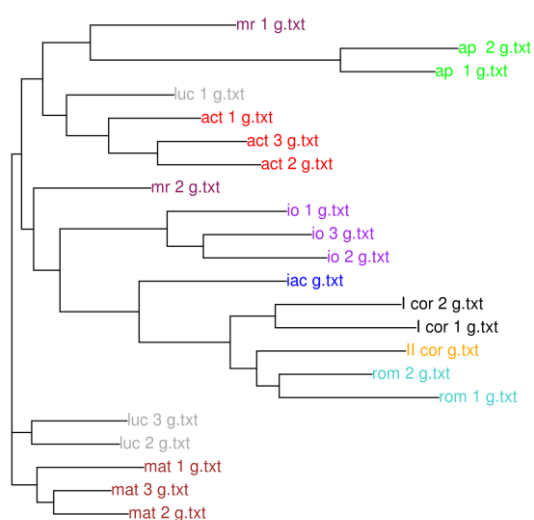


Figure 1. Greek New Testament, 30 MFW, Eder's simple distance

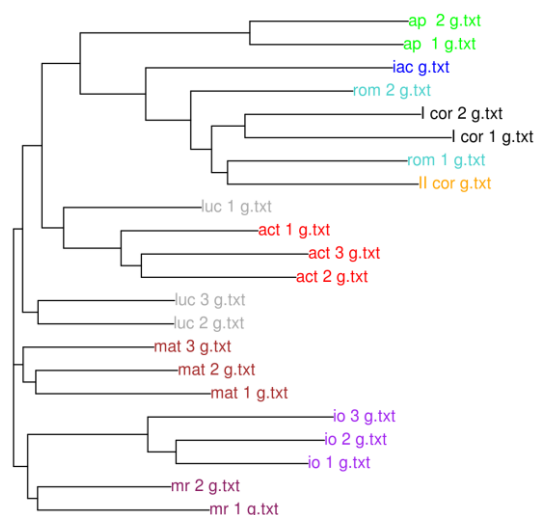


Figure 2. Greek New Testament, 300 MFW, Classic Delta distance

Fig. 1 shows the results for 30 the most frequent words (short ‘MFW’). One can clearly see that two parts of the Revelation are clustered together with the beginning of Mark; another discrete cluster stands for Matthew combined with the final passages of Luke. In the middle of the graph, there is a distinguishable cluster of Paul’s Epistles and James’s Epistle linked together. In Fig. 2 (300 MFW, classic Delta measure), the cluster of Epistles is even more distinct, but this time it unexpectedly absorbs the Revelation. An interesting thing is that a cluster containing the Acts attracts the first part of Luke – which might reflect some stylistic

similarities between the Acts and the Gospel of Luke (Greenwood 1995). The dendrogram for 1000 MFW and classic Delta distance measure (Fig. 3) combines, in a way, the information brought by the two previous graphs. Thus, the third plot seems to be the most convincing... Or does it?

At this point, a stylometrist inescapably faces the above-mentioned cherry-picking problem (Rudman 2003). When it comes to choosing the plot that is the most likely to be “true”, scholars more or less unconsciously pick the one that *looks* more reliable than others, or simply confirms their hypotheses. If common sense is used to evaluate the obtained plots, any counter-intuitive results will be probably dropped simply because they do not fit the scholars’ expectations. An interesting variant of cherry-picking is discussed by Vickers, who writes about the “visual rhetoric” of different lines, arrows, colors etc. added to a graph; being helpful, at the same time they suggest apparent separations of samples (Vickers 2011: 127).

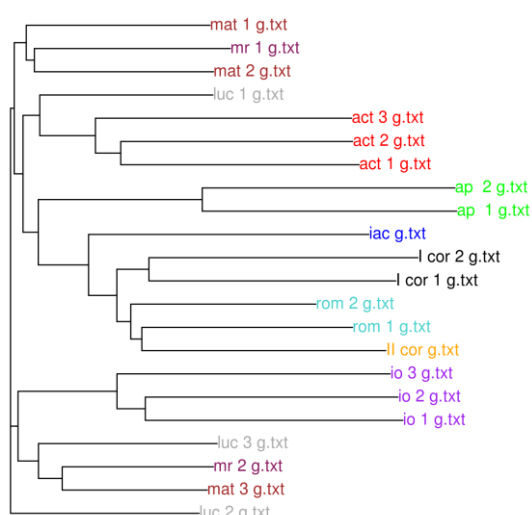


Figure 3. Greek New Testament, 1000 MFW, classic Delta distance

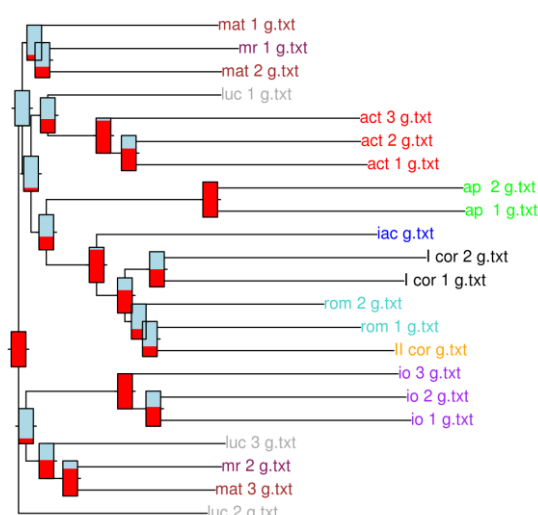


Figure 4. Greek New Testament, 1000 MFW, classic Delta, validated using the results of 5000 bootstrap turns

Is it possible to eschew the problem of cherry-picking? Yes, if one agrees to turn over the natural hierarchy in human-machine interaction, and accepts the superiority of automatic (i.e. machine-based) estimation of the most “reliable” picture. Even if it sounds like a post-human manifesto, it has been successfully used for decades in computer sciences, and also in computational stylistics. To exemplify, in a study aimed to identify the most “typical” works by analyzed authors (whatever a word “typical” means), an effective way to evaluate the

validity of particular samples turned out to be a procedure of intensive random permutation of the corpus to rule out the outliers (Eder & Rybicki 2012).

In the New Testament case, a similar approach might be used. An easiest way to get rid of cherry-picking is to apply a series of tests using different vectors of frequent words (e.g. 100, 110, 120, 130, ..., 1000), followed by an automatic evaluation of the dozens of pictures obtained throughout the analysis. One has to remember, however, that the arbitrary choice of 100, 110 etc. words might still lead to biased results. For this reason, a more advanced procedure, derived from a variety of bootstrap methods, is used instead.

The general idea of bootstrap is to perform a series of approaches to the input data: in a large number of trials, samples from the original population are chosen randomly (with replacement), and this chosen subset is analyzed in substitution of the original population (Good, 2006). Speaking of stylometric multivariate analyses, one can compute a list of the most frequent words from a corpus and use it as the “original population”, and then to produce a large number of virtual subsets containing randomly selected words. In the present approach, a list of 1000 MFW is used; a few dozen words occupying the top of this list are as follows (in descending order):

καὶ, ὁ, τοῦ, ἐν, δέ, τὸ, εἰς, τὸν, τὴν, ὅτι, τῆς, τῶν, τῷ, αὐτοῦ, οἱ, ἡ, ἐκ, οὐκ, μὴ, τῇ, τὰ, αὐτῷ, γὰρ, τοὺς, ἴνα, οὐ, οὖν, ἐπὶ, πρὸς, θεοῦ, αὐτὸν, αὐτῶν, ἐστίν, ὡς, μου, ὑμῖν, ἰησοῦς, ὑμῶν, τοῖς, αὐτοῖς, διὰ, εἶπεν, ἐγὼ, εἰ, ἀπὸ, λέγει, ἀλλὰ, ὑμᾶς, ἦν, τε, με, ἡμῶν, περὶ, τὰς, θεός, τοῦτο, ὑμεῖς, σου, τί, ἦ, αὐτόν, ἀλλ, ταῦτα, ἐάν, τις, κατὰ, μοι, μετὰ, κυρίου, πάντα, ἰησοῦ, ἐγένετο, αὐτοῦς, οὕτως, ὃν, αὐτῆς, πάντες, ταῖς, χριστοῦ, γῆς, μέν, ὃ, πνεῦμα, ἐξ, ἡμεῖς, σὺ, ὑπὸ, νῦν, ἀπεκρίθη, ἃ, δέ, ὅς, σὺν, καθὼς, ἡμῖν, ἡμᾶς, ἐπ, ἰδοὺ, μετ, ὑπὲρ, ...

Next, 100 words have been randomly harvested from this list (with replacement) in a very large number of iterations. Presumably, 5000 turns and 100 words in each turn is sufficient to cover the whole range of the approached fragment of the frequency list. In each turn, cluster analysis based on the selected 100 words were performed, and the results were recorded. Perhaps a straightforward way to assess the results is to produce 5000 subsequent dendrograms, one for each trial, but it is hardly feasible to inspect them all with the naked eye. Instead, the recorded information about particular clustering across the 5000 turns can be used to validate, say, a manually chosen dendrogram; it might be even the same plot that was cherry-picked at the earlier stage of the analysis (Fig. 3–4). The thermometers added to the plot (Fig. 4) represent the results of the bootstrap procedure. They show how reliable particular nodes on the graph are: the higher the “temperature”, the more robust a given

linkage, since the “temperature” reflects recurrence of the nodes across the 5000 bootstrap trials. It is evident in Fig. 4 that some of the clusters turned out to be rather accidental, while some other display a considerably high “temperature”: particularly John, the Acts, and the Revelation. Also, Paul’s Epistles and James’s Epistle are very robustly detached, even if they flock together in one common cluster.

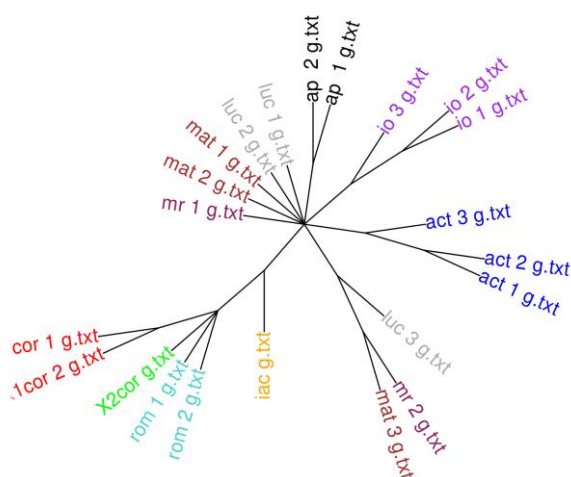


Figure 5. Greek New Testament, bootstrap consensus tree (consensus strength: 0.5)

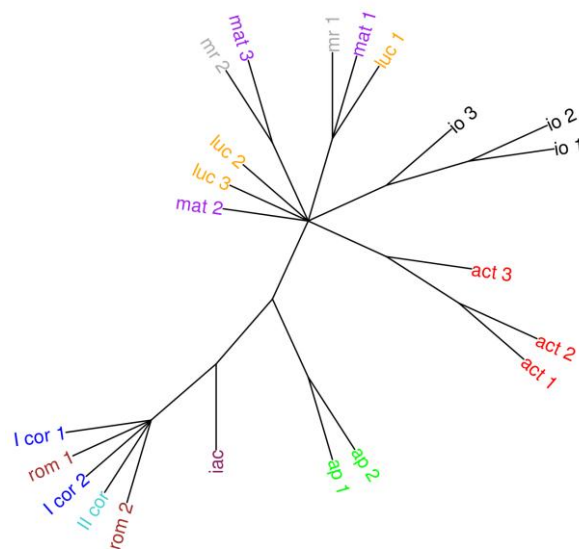


Figure 6. Latin New Testament, bootstrap consensus tree (consensus strength: 0.5)

The technique introduced above might serve as a comprehensive lie detector for testing particular plots’ reliability. For simple pictures, it might be a very convenient solution. However, interpreting a considerably complex dendrogram with numerous nodes can be a rather tough task. The last stage of the analysis, then, is to produce a compact plot (referred to as a “consensus tree”) that would summarize the information on clustering from the 5000 bootstrap iterations. The principle of building the plot is simple: if the “temperature” of a particular node is high enough, the node will appear on the consensus tree as well (Fig. 5).

At this point, we are really far away from the manual inspection of various dendrograms in search of the most “reliable” picture. The presented method of data verification using bootstrap seems to have solved the cherry-picking problem, but there is still a fly in the ointment. Namely, in the process of building the consensus tree one has to *decide* how high the “temperature” needs to be to establish a particular cluster. The decision is an arbitrary one.

The mechanism of hammering out the consensus can be compared to voting in an election: particular nodes appearing on different dendrograms “vote” for a certain cluster; the thermometers indicate the percentage of “votes” for and against the cluster in question. Like in real-life political systems, however, it has to be decided how many “votes” are needed to make the election valid. Usually, it is at least 50% of the votes, sometimes more, and some elections require unanimity; the same rules applies to consensus trees. Depending on the decided robustness threshold (or, the sufficient “temperature”), the final shape of the grown tree might differ significantly, as shown in Fig. 5–10.

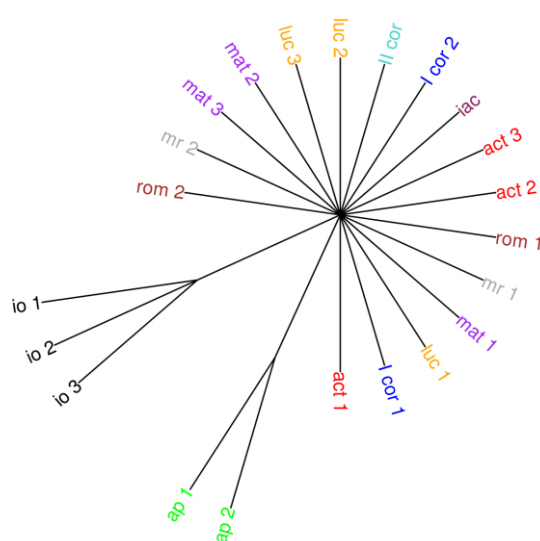
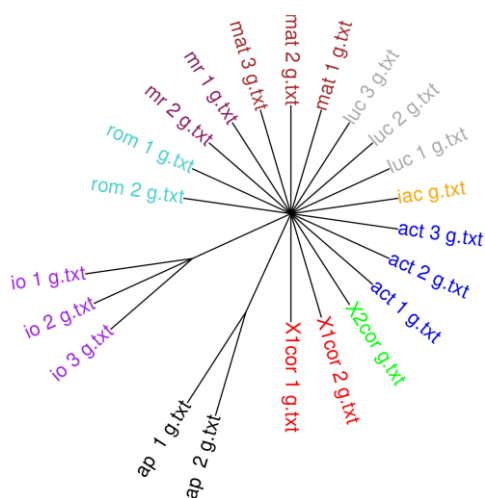


Figure 7. Greek New Testament, bootstrap consensus tree (consensus strength: 0.95) **Figure 8.** Latin New Testament, cbootstrap consensus tree (consensus strength: 0.95)

In Fig. 5, a very “democratic” type of consensus tree is shown: only those groupings that appeared in at least 50% bootstrap iterations were used to build a consensus tree (the “temperature” set to 0.5). One can easily identify a discrete branch for the Epistles (Jacob being put apart), branches for the Revelation, John, and the Acts. The remaining distinctive branch stands for three narrative variants of the crucifixion and resurrection of Christ described in three final parts of Matthew, Mark, and Luke – which is quite easy to explain, since the Synoptic Gospels share a great amount of textual material. The remaining samples are linked directly to the root of the tree, which means that they are ambiguous: in the subsequent 5000 bootstrap iterations, they are jumping from one cluster to another.

In Fig. 7, the robustness threshold was set to almost unanimous consensus (the “temperature” is decided to be as high as 0.95). This rigid version of consensus tree reveals an

interesting fact that in the Greek New Testament, only two books are soundly distinct in terms of stylistic differentiation: John and the Revelation.

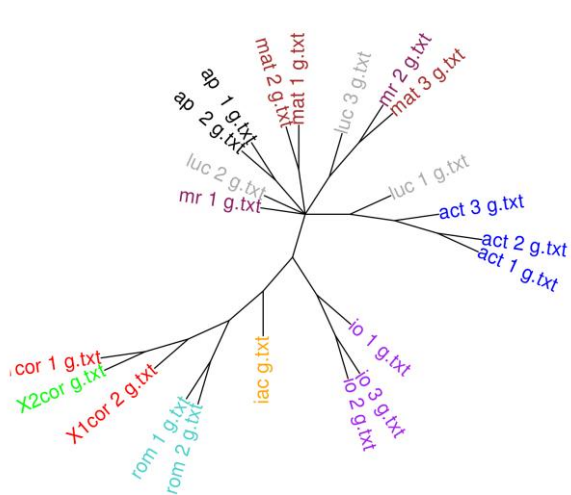


Figure 9. Greek New Testament, bootstrap consensus tree
(consensus strength: 0.3)

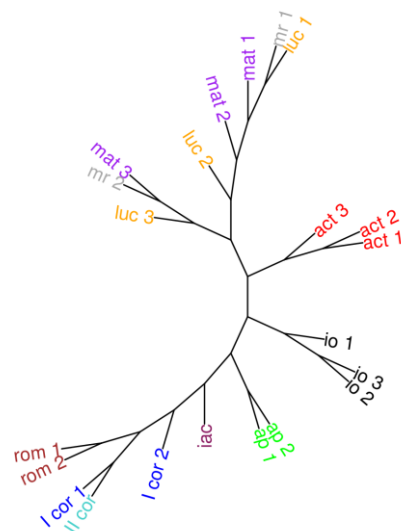


Figure 10. Latin New Testament, bootstrap consensus tree
(consensus strength: 0.3)

On the other pole is Fig. 9, where the consensus strength is set to an extremely low value of 30% (a voting system hardly imaginable in real-life democracies). Certainly, this plot is less reliable than the trees shown above, but at the same time it might betray some secondary regularities that are normally overwhelmed by strong authorial signals. Here, the cluster for the Acts seems to be interesting, because it absorbed the beginning part of Luke. Even if weak, this signal might to some degree confirm the hypothesis that St Luke was the author of the Acts (Guthrie 1990).

It is hard to decide which threshold of robustness should be chosen. Presumably, a reasonable approach is to generate a couple of consensus trees and to evaluate behavior of particular clusters. Now, if a given group of texts happens to be clustered on a unanimous consensus tree, it suggests that stylistic similarities between these texts are very strong indeed. On the other hand, if a tolerant consensus tree (the “temperature” around 0.5 or less) does not show any linkage between given samples, one has a convincing evidence of their actual significant differentiation.

5. Stylometry of translation

Finally, having discussed behavior of the Greek corpus, one can confront the results with its translated counterpart (Fig. 5–10). The general observation that can be made is that the twin versions of the New Testament – the Greek original and its Latin translation – display striking similarities. In the Vulgate, the original authorial signal is predominant and can be traced through the (almost) transparent layer of translatorial signal. The parallel trees representing the consensus of 95% are simply identical; on two remaining pairs of plots (consensus of 50% and 30%, respectively), most groupings in the Greek corpus are mirrored on the Latin side as well. The differences between the corpora are modest, yet interesting. First, a weak connection between the Acts and the Gospel of Luke, that could be seen on some plots on the Greek side, disappeared in the translation. In other words: the Latin translation differentiates stylistically the Acts and the Synoptic Gospels to a greater extent than the Greek original. Secondly, in the Latin translation the Synoptic Gospels tend to break into two discrete clusters: one for the opening parts, another for the closing sections of subsequent Gospels. In the Greek version, the Synoptic Gospels' clustering according to content (rather than to authorship) is not that clear.

6. Conclusions

In this paper, some reliability issues in computer-assisted translation studies have been discussed. The main methodological problem addressed in the study refers to the evaluation and validation of the results obtained using explanatory techniques of nearest neighbor classification. As presented above, hierarchical cluster analysis is vulnerable to a few factors, including the number of features, method of linkage, and distance measure used in the analysis. A dendrogram always represents a single precisely defined set of these variables (e.g. 100 frequent words + Ward's linkage + Euclidean distance), yet it might yield the correct results simply by chance. Even though, the results are unstable and their interpretation depends on arbitrary decisions made by a scholar: in evaluating the results, the risk of cherry-picking is obvious.

The procedure introduced above aims to help eschew the problem of arbitrariness. In a very large number of iterations, the variables needed to construct a dendrogram were chosen randomly, and a virtual dendrogram for each iteration was generated. Next, these numerous virtual dendrograms were combined into a single compact consensus tree. It is believed that this technique can provide an insight into *average* behavior of the analyzed corpus. However, there was no ideal consensus tree generated in the study, in terms of a single plot that would

tell the whole true about the input data. It seems that the Holy Grail of stylometric reliability is still beyond our capabilities.

References

- Argamon, Shlomo 2008: Interpreting Burrows's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing* 23, 131–147.
- Baayen, Harald 2008: *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*, Cambridge: Cambridge University Press.
- Baker, Mona 2004: A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics* 9, 167–193.
- Brown, Raymond E. 1997: *Introduction to the New Testament*. New York: Anchor Bible.
- Burrows, John 1987: *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.
- Burrows, John 2002a: The Englishing of Juvenal: computational stylistics and translated texts. *Style* 36, 677–699.
- Burrows, John 2002b: "Delta": A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17, 267–287.
- Burrows, John 2004: Textual analysis. In: Susan Schreibman, Ray Siemens, John Unsworth (eds.) 2004: *A Companion to Digital Humanities*. Oxford: Blackwell, 323–347.
- Craig, Hugh, Arthur F. Kinney (eds.) 2009: *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.
- Dunn, Michael, Angela Terrill, Geer Reesink, Robert Foley, Stephen Levinson 2005: Structural phylogenetics and the reconstruction of ancient language history, *Science* 309, 2072–2075.
- 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 2012: Mind your corpus: Systematic errors in authorship attribution. *Digital Humanities 2012: Conference Abstracts*. University of Hamburg, 181–185.
- Eder, Maciej, Jan Rybicki 2012: Do birds of a feather really flock together, or how to choose test samples for authorship attribution. *Literary and Linguistic Computing* 27, doi:10.1093/lc/fqs036 (published online 11 September 2012). ED 10/2012
- Good, Philip 2006: *Resampling Methods: A Practical Guide to Data Analysis*. Boston–Basel–Berlin: Birkhäuser.
- Greenwood, H. H. 1995: Common word frequencies and authorship in Luke's Gospel and Acts. *Literary and Linguistic Computing* 10, 183–187.
- Guthrie, Donald 1990: *New Testament: Introduction*. Leicester: Apollos.
- Helms, Randel 1997: *Who Wrote the Gospels?* Altadena, California: Millennium Press.
- Heydel, Magda, Jan Rybicki 2012: *Digital Humanities 2012: Conference Abstracts*, University of Hamburg, 212–215.
- Holmes, David 1998: The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing* 13, 111–117.

- Hoover, David 2003a: Multivariate Analysis and the Study of Style Variation. *Literary and Linguistic Computing* 18, 341–360.
- Hoover, David 2003b: Frequent collocations and authorial style. *Literary and Linguistic Computing* 18, 261–286.
- Jockers, Matthew, Daniela Witten, Craig Criddle 2008: Reassessing authorship of the “Book of Mormon” using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing* 23, 465–491.
- Kenny, Anthony 1986: *A Stylometric Study of the New Testament*. Oxford: Clarendon Press.
- Koppel, Moshe, Jonathan Schler, Shlomo Argamon 2009: Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology* 60, 9–26.
- Ledger, Gerard 1995: An exploration of differences in the Pauline Epistles using multivariate statistical analysis. *Literary and Linguistic Computing* 10, 85–97.
- 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.
- Luyckx, Kim 2010: *Scalability Issues in Authorship Attribution*. Diss. Univ. Antwerpen.
- Morton, Andrew 1978: *Literary Detection: How to prove authorship and fraud in literature and documents*. New York: Scribner.
- Mosteller, Frederick, David Wallace 1964: *Inference and Disputed Authorship: The Federalist*. Reprinted with a new introduction by John Nerbonne. Stanford: CSLI Publications, 2007.
- Nautin, Pierre 1986: Hieronymus. In: Gerhard Krause, Gerhard Müller (eds.) *Theologische Realenzyklopädie*. Vol. 15. Berlin–New York: Walter de Gruyter, 304–315.
- Rudman, Joseph 1998a: Non-traditional authorship attribution studies in the “Historia Augusta”: some caveats. *Literary and Linguistic Computing* 13, 151–157.
- Rudman, Joseph 1998b: The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities* 31, 351–365.
- Rudman, Joseph 2003: Cherry picking in nontraditional authorship attribution studies. *Chance* 16, 26–32.
- Rybicki, Jan 2006: Burrowing into translation: character idiolects in Henryk Sienkiewicz's “Trilogy” and its two English translations. *Literary and Linguistic Computing* 21, 91–103.
- Rybicki, Jan 2011: Alma Cardell Curtin and Jeremiah Curtin: the transtalar's wife's stylistic fingerprint. *Digital Humanities 2011: Conference Abstracts*, Stanford University, Stanford, CA, 219–222.
- Saitou, Naruya, Masatoshi Nei 1987: The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4, 406–425.
- Tabata, Tomoji 2012: Approaching Dickens' style through random forests. *Digital Humanities 2012: Conference Abstracts*, University of Hamburg, 388–391.
- Vickers, Brian 2011: Shakespeare and authorship studies in the twenty-first century. *Shakespeare Quarterly* 62, 106–142.
- Ward, Joe H. 1963: Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58, 246–244.

Streszczenie

Niniejszy artykuł omawia kilka kluczowych kwestii związanych z komputerową analizą stylu literackiego w badaniach nad przekładem. Tekstem stanowiącym podstawę porównań był Nowy Testament w dwóch wersjach językowych: oryginalnej greckiej i w łacińskim przekładzie św. Hieronima znanym pod nazwą Wulgata. Podstawowe pytanie badawcze stawiane w artykule było następujące: czy stosowane w stylometrii metody analizy wielowymiarowej (takie jak analiza skupień) dają wiarygodne wyniki? W stylometrii od kilkudziesięciu z górą lat stosuje się zaawansowane techniki probabilistyczne, w tym modelowanie, uczenie maszynowe etc., których cechą wspólną jest bardzo duży stopień sformalizowania matematycznego. Z drugiej strony stylometria największe swoje sukcesy odnosi od czasu, gdy badacze literatury zastosowali kilka podstawowych metod statystycznych do analizy stylistycznej dzieł literackich. Problem, w pewnym skrócie, przedstawia się następująco: literaturoznawcy stawiający istotne kwestie badawcze stronią od wyrafinowanych technik klasyfikacji; informatycy oferują metody dokładne, lecz nieatrakcyjne dla humanistów. Celem niniejszego artykułu było połączyć oba podejścia i wypracować metodę dokładną, lecz zarazem łatwą w interpretacji uzyskanych wyników. Efektem jest metoda wielokrotnego (automatycznego) przemieszania danych wejściowych i wykonania za każdym razem nowego testu analizy skupień. Uśrednienie wyników dla kilku tysięcy iteracji daje możliwość znalezienia powtarzalnych regularności i odrzucenia przypadkowych "podobieństw" między próbkami. Ostatnim etapem jest automatyczne wykreślenie graficznej reprezentacji uśrednionych wyników, tzw. drzewka konsensu. Polega to na tym, że najbardziej podobne próbki skupiają się na jednej "gałęzce" drzewa, próbki zaś, które nie mogły być wiarygodnie zaklasyfikowane, zostają przyłączone bezpośrednio do "korzenia" drzewa-grafu.

Address for correspondence:

Maciej Eder

Institute of Polish Studies

Pedagogical University of Kraków

ul. Podchorążych 2

30-084 Kraków, Poland

e-mail: maciejeder@gmail.com