Visualization in stylometry: Cluster analysis using networks

Maciej Eder
Pedagogical University of Kraków, Poland
Institute of Polish Language, PAS

Abstract

The aim of this article is to discuss reliability issues of a few visual techniques used in stylometry, and to introduce a new method that enhances the explanatory power of visualization with a procedure of validation inspired by advanced statistical methods. A promising way of extending cluster analysis dendrograms with a self-validating procedure involves producing numerous particular 'snapshots', or dendrograms produced using different input parameters, and combining them all into the form of a consensus tree. Significantly better results, however, can be obtained using a new visualization technique, which combines the idea of nearest neighborhood derived from cluster analysis, the idea of hammering out a clustering consensus from bootstrap consensus trees, with the idea of mapping textual similarities onto a form of a network. Additionally, network analysis seems to be a good solution for large data sets.

Correspondence:
Maciej Eder, Institute of Polish Studies, Pedagogical University of Kraków, ul. Podchorążych 2, 30-084 Kraków, Poland.
E-mail:
maciejeder@gmail.com

1 Introduction

Most of the computational methods used in stylometry have been originally introduced to solve authorship attribution problems. This fact had an immense influence on the further development of the whole discipline. The seminal study by Mosteller and Wallace (2007 [1964]) showed in a very convincing way that authorship attribution based on statistical analysis of style is ultimately the problem of classification. In its standard form, attribution is aimed at extracting a unique authorial profile from a disputed text and from texts written by possible 'candidates'; the goal is to compare the profiles and to single out the matching 'candidate'. Even if one deals with an open-set attribution case—where the list of possible candidates cannot be reliably established—the general idea does not differ substantially from other classification problems.

Exact science has developed 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, that could be easily adopted to authorship attribution. They include naïve Bayes classification, support vector machines, nearest shrunken centroids, or random forests, to name but a few (Mosteller and Wallace, 2007 [1964]; Jockers *et al.*, 2008; Koppel *et al.*, 2009, Tabata, 2012).

Independently, a ground-breaking monograph on Jane Austen published by Burrows (1987) ushered stylometry into literary criticism. It turned out that from a literary perspective, matching profiles of 'candidates' is not as important as obtaining a broader picture of relations between different novels, types of narration, main characters' voices, and so forth. The methods adopted or introduced by Burrows, Hoover, Craig, and others (Burrows, 1987, 2002, 2007; Hoover, 2003a, b; Craig and Kinney, 2009) were very intuitive and easily-applicable to literary studies. These include principal components

Digital Scholarship in the Humanities, Vol. 32, No. 1, 2017.
doi:10.1093/llc/fqv061 Advance Access published on 2 December 2015

analysis, multidimensional scaling, cluster analysis, Delta, Zeta, and Iota. Despite their limitations (the lack of validation of the obtained results being the most obvious), they are still widely used.

The reason of their popularity is that they meet the needs of literary scholars, also because they offer convincing visualizations.

Needless to say, visualization has an undeniable explanatory power. Scatterplots, maps, trees, and diagrams provide an insight into the whole corpus at one glance. Moreover, they allow to draw conclusions about literature from a distant-reading perspective, through a visual interpretation of groupings and separations of several samples. Certainly, this is particularly desired in stylometry beyond authorship attribution. The attractiveness of visualization in computational literary criticism is confirmed not only by the aforementioned studies by Burrows or Hoover, but also by immense popularity of beautiful yet relatively simple plots presented by Moretti, Jockers, Posavec, and others (Morretti, 2005; Posavec, 2007; Jockers, 2013; Sinclair and Rockwell, 2014). The aim of this article is to discuss reliability issues of a few visual techniques, and to enhance the explanatory power of visualization with a procedure of validation inspired by advanced statistical methods.

2 Reliability in Computational Stylistics[1]

The question of reliability in non-traditional authorship attribution has been extensively discussed by Rudman (1998a,b, 2003), who formulated a number of caveats concerning corpus preparation, sampling, selection of style-markers, interpreting the results, etc. Rudman's fundamental remarks, however, have not been preceded by 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)certainty (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, 2015), an experiment using intensive corpus re-composition to test whether the attribution accuracy depends on particular constellation of texts used in the analysis (Eder and Rybicki, 2013), a study aimed to examine the performance of untidily prepared corpora (Eder, 2013a), and so forth.

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, or 10 random swaps between two parts of a corpus: a subset of training texts and a subset of texts used in the testing procedure.

Most unsupervised methods used in stylometry, such as principal components analysis, multidimensional scaling, 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 the group of texts to which a disputed sample belongs.

Hierarchical cluster analysis—as discussed in the present study—is a technique which tries to find the most similar samples (e.g. literary texts) and builds a hierarchy of clusters, using a 'bottom-up' approach. What makes this method attractive is the very intuitive way of graphical representation of the obtained results: contrarily to the scatterplots as produced by multidimensional scaling or principal components analysis, where the goal is to interpret relative positions of several points settled on a rectangular plot, cluster analysis produces explicit links between neighboring items (see Figs 1–4). 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) the particular distance measure applied to the data, (2) the 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.

(1) In a study of multivariate text analysis using dendrograms, Burrows concludes, '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, p. 326). The distance used by Burrows is a widely accepted solution in the field of computational stylistics; there are no studies, however, that would satisfactorily explain 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, p. 134), a slightly modified version of the classic Delta measure as developed by Burrows (2002). There is no denying that Delta, and *ipso facto* the distance measure embedded in it, proved to be very effective—a fact confirmed by numerous stylometric studies; thus, it should be also applicable to hierarchical cluster analysis procedure. Even if convincing at first glance, however, the choice of this particular measure needs to be theoretically justified and confirmed by empirical comparisons with other distances.

(2) Another factor affecting the final shape of a dendrogram is the method of linkage used. In the above-cited statement, Burrows favors 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. Good performance of Ward's method has been also proven in many other applications within the field of quantitative linguistics, corpus linguistics, and related disciplines. Although it seems to be accurate indeed, there is no awareness, however, 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, p. 236).

(3) Even if some issues still remain unresolved, scholars roughly agree that Euclidean (normalized) distance and Ward's linking algorithm provide acceptable results. However, the same cannot be said about the third factor cluster analysis depends on, which is the number of features (e.g. frequent words) to be analyzed, and the type of countable features (e.g. words, word *n*-grams).

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 (often, function words), others prefer long vectors of words, and so on. Although all these solutions are reasonable and theoretically justified, the final choice of the number of features to analyze is *a priori* arbitrary. This problem is sometimes referred to as 'cherry-picking' (Rudman, 2003). Awareness of this issue, followed by partial solution, can be observed in the studies by Hoover (2003a, b), who assesses a given corpus with a few discrete cluster analyses for different most frequent word (MFW) values. Even if still subject to arbitrary choices, this approach gives a fairly good insight into variability of the input data. This way of dealing with uncertainty will be discussed below in detail, with its possible extension to other visualization techniques.

3 Multilayer Model of Written Text

As will shortly be demonstrated, even the slightest change in the experiment setup might cause a severe reshaping of the final dendrogram. Without deciding which of the three factors discussed in the previous section—linkage algorithm, distance measure, and the number of words analyzed—is more likely to affect the final shape of a dendrogram, one must admit that the first two are related to the method of clustering, while the third factor is inherently linked to certain linguistic features of analyzed texts.

Endless discussions of how many frequent words or *n*-grams should be taken into account (e.g. Mosteller and Wallace, 1964; Hoover, 2003a; Burrows, 2007; Koppel *et al.*, 2009; Eder, 2013b; Schöch, 2013) show rather clearly that there is no universal frequency strata where the authorial fingerprint is hidden. Just the opposite, it seems that the authorial signal is spread throughout the whole frequent and not-so-frequent words spectrum, but at the same time it may become obscured by additional and unpredictable signals, which are considered noise in classical approaches to attribution. In stylometry beyond attribution, however, this 'noise' is worth a closer look. Why are some authors misclassified? Which texts are wrongly attributed to a given author, and why are they linked to this very author and not to others? These and similar questions are probably much more interesting than the never-ending fine-tuning of the parameters of this or that classification algorithm in order to neutralize the impact of the 'noise'.

Obviously, the problem is not new. Cross-genre authorship attribution, for one, has always been a major challenge (Kestemont *et al.*, 2012; Schöch, 2013). Also, there have been a few attempts to extract particular signals hidden in texts: author's nationality (Jockers, 2013), psychologic profile (Noecker *et al.*, 2013), gender (Pennebaker, 2011), genre (Koppel *et al.*, 2009), and translator's fingerprint (Rybicki and Heydel, 2013). On theoretical grounds, function words should be responsible for authorial recognition, while content words should be more topic- and genre-related. The abovementioned empirical studies, however, do not really confirm this assumption. There is no clear rule here, and the same words are sometimes claimed to reveal different signals. For instance, the definite article 'the' is considered to discriminate British versus American flavors of English in one study (Jockers, 2013, p. 105), and female versus male language in another (Pennebaker, 2011, p. 42).

The difficulties with separating one specific signal suggest that a text (written or spoken) is a multilayer phenomenon, in which particular layers are correlated. These layers include authorship, chronology, personality, gender, topic, education, literary quality, translation (if applicable), intertextuality, literary tradition (e.g. sources of inspiration), and probably many more. Arguably, literary quality somehow depends on education, genre depends on topic, authorial voice is affected by chronology, gender affects personality, and so on. Some layers might be barely noticeable, and some others might become surprisingly strong. In authorship attribution, this complex system of uncontrollable layers is a problem of unwanted noise, and in literary-oriented computational stylistics, an opportunity to see more.

4 Dendrogram, or One Snapshot at a Time

Since particular frequency strata are responsible, to some extent, for different signals hidden in a literary text, the dendrograms generated using longer or shorter MFW vectors presumably will also be heterogenous. And they actually are (Figs 1–4); the only problem is that their variability is much bigger than one could expect and—what is worse—the changes in dendrograms' shapes are unpredictable. Different combinations of linkage algorithms, number of MFWs, and distance measures applied, one obtains a convincing example of how unstable the final results might be.

Worth noticing, however, that the authorial 'leaves' on the dendrograms are usually correctly clustered regardless of the parameters used. In Fig. 1 (Ward's linkage, 100 MFWs), most of the authors are recognized to be stylistically homogenous; the exceptions include Charles Dickens and Henry James. When the number of features increases to 300 MFWs, the 'leaves' of the dendrogram are matched with no misattributions (Fig. 2). In any attempts to visualize larger groupings of texts, however, one needs to admit that the 'branches' are significantly less predictable than the 'leaves': is Galsworthy stylometrically similar to George Eliot or to Joseph Conrad? Is Thackeray linked to Walter Scott or to Charles Dickens? What does the main division into two large clusters mean? Figures 1–4 might support many contradictory hypotheses.

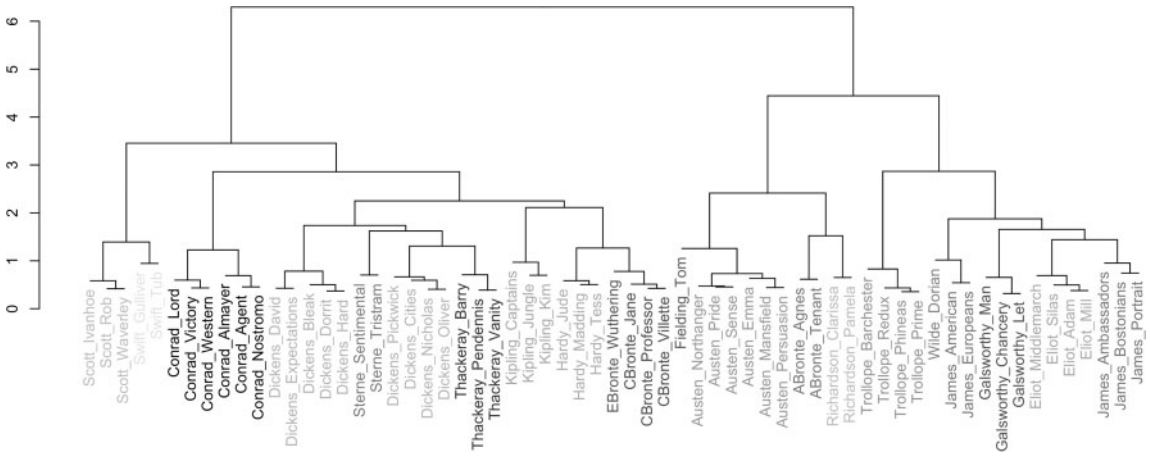

Fig. 1. Cluster analysis of 66 English novels, 100 MFWs, classic Delta distance, Ward's linkage. Color versions of all figures are available online.

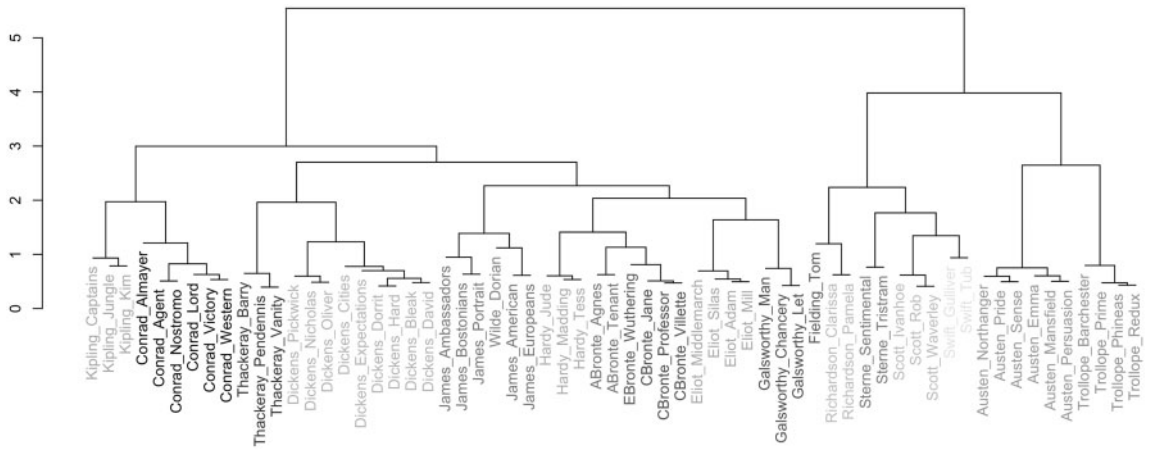

Fig. 2. Cluster analysis of 66 English novels, 300 MFWs, classic Delta distance, Ward's linkage.

The problems do not end here: a detailed inspection of multiple dendrograms generated for gradually increasing number of features (MFWs) shows that substantial rearrangements might occur quite suddenly. An example of this behavior is shown in Figs 3 and 4. Cluster analysis using McQuitty's linkage and 136 MFWs (not shown) reveals a perfect authorial recognition, but when 137 MFWs are used, the cluster for Joseph Conrad is split into two parts and remains detached (along many other substantial rearrangements of the corpus) until the same corpus is assessed at 969 MFWs (Fig. 3). *Almayer's Folly* jumps back from Kipling's branch to Conrad's cluster exactly between the words 969 and 970 on the frequency list (Fig. 4). The knowledge that this 970th word is 'wine' does not help much, however, since multivariate analyses take into consideration a great number of features at a time. The word 'wine', not very discriminative itself, was the factor to tip the scale in favor of Conrad. What is more important here is the side-effect: apart from the local Kipling/Conrad change, the whole dendrogram has been severely affected and, in consequence, significantly reshaped. Such

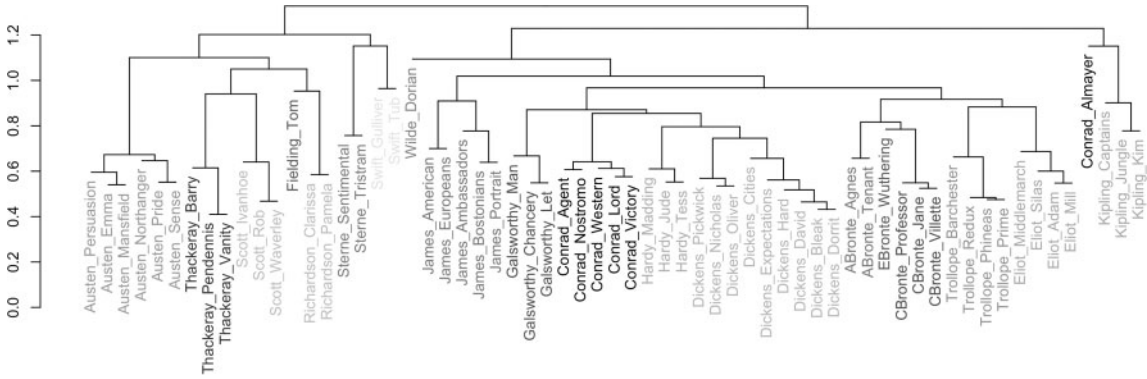

Fig. 3. Cluster analysis of 66 English novels, 969 MFWs, classic Delta distance, McQuitty's linkage.

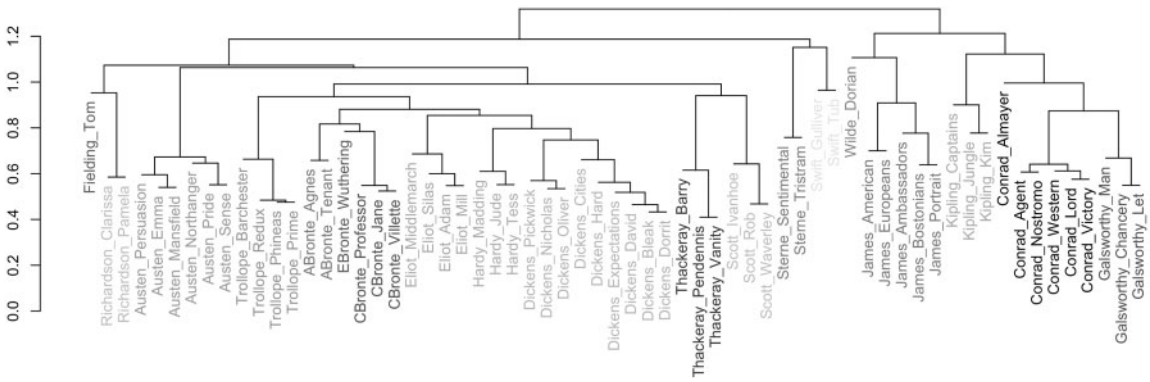

Fig. 4. Cluster analysis of 66 English novels, 970 MFWs, classic Delta distance, McQuitty's linkage.

abrupt changes seem to be a rule rather than the exception, at least for textual data sets.

The decision which of the dendrograms presented above reveal the actual separation of the samples and which show fake similarities is not trivial at all. Generating hundreds of dendrograms covering the whole spectrum of MFWs, a variety of linkage algorithms, and a number of distance measures, would make this choice even more difficult. At this point, a stylometrist inescapably faces the abovementioned cherry-picking problem (Rudman, 2003). When it comes to choosing the plot that is the most likely to be 'true', scholars are often in danger of more or less unconsciously picking the one that looks more reliable than others, or that 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, and so forth added to a graph; while helpful, at the same time they suggest apparent separations of samples (Vickers, 2011, p. 127).

4 Consensus Tree, or Many Dendrograms Combined

A partial solution of the cherry-picking problem involves combining the information revealed by numerous dendrograms into a single consensus plot. This technique has been developed in phylogenetics (Paradis *et al.*, 2004) and later used to assess

differences between Papuan languages (Dunn *et al.*, 2005). It has been also introduced into stylometry (Eder, 2013b) and applied in a number of stylometric studies (Rybicki, 2012; Rybicki and Heydel, 2013; van Dalen-Oskam, 2014). This approach assumes that, in a large number of 'snapshots' (e.g. for 100, 200, 300, 400,. . ., 1,000 MFWs), actual groupings tend to reappear, and apparent similarities are likely to remain accidental. The goal, then, is to capture the robust patterns across a set of generated snapshots. The procedure is aimed at producing a number of virtual dendrograms, and then at evaluating robustness of groupings across these dendrograms. If a given link—say, between Richardson's *Pamela* and Fielding's *Tom Jones*—turns out to appear frequently enough, it is reproduced on a consensus plot. In other words, several regular (yet virtual) dendrograms 'vote' for the most robust links—the procedure summarizes the information on clustering from particular plots.

In Fig. 5, a consensus tree of the corpus of 66 English novels has been shown (the 'snapshots' were computed for 100, 200, 300, etc. up to 1,000 MFWs). Some text groupings can be easily identified, including, among others, an expected cluster of the three Brontë sisters, and a branch of Kipling/Conrad—clearly subdivided into two distinct authorial voices. Unlike typical dendrograms, however, the established links do not represent stylometric distances between samples. Instead, they indicate the strength of the consensus, or the repetitiveness across a number of virtual 'snapshot' dendrograms.

Upgrading the procedure from a cherry-picked cluster analysis into a consensus tree is a significant step toward reliable stylometry. Such a tree captures the average behavior of a corpus for a given frequency strata (in this case, 100–1,000 MFW). More importantly, it filters out local disturbances (artifacts) that could otherwise be considered as valid results. Some arbitrary decisions cannot be avoided, though. They include the number of features to be assessed, the number of iterations ('snapshots') to produce a consensus tree, and—last but not least—the linkage algorithm embedded in the whole procedure. A considerably simple way to neutralize these issues is to reproduce a given experiment using different settings. Sooner or later, however, other limits of consensus tree approaches become painful, especially when the number of analyzed texts increases. The technique introduced below is aimed at overcoming these limits.

5 Consensus Network, or Importance of Runners-Up

Although the problem of unstable results can be partially by-passed using consensus techniques, two other issues remain unresolved. Firstly, when the number of analyzed samples exceeds a few dozen, the plot becomes cluttered and thus illegible. Secondly, the procedure of hammering out the consensus is aimed at identifying nearest neighbors only, which means extracting the strongest patterns (usually, the authorial signal) and filtering out weaker textual similarities. Consequently, samples on a consensus tree are very likely to be grouped into many discrete authorial clusters rather than into a few larger branches. When the number of analyzed texts is considerably small, the granulation of clusters is barely noticeable (Fig. 5); in large corpora, however, numerous little branches are linked directly to the root of the dendrogram. Useful in explanatory authorship attribution, such a plot will not support stylometric interpretations of similarities between texts, authors, genres, styles or literary epochs. Arguably, large-scale stylometry will be interested in deeper textual relations rather than in mere nearest neighborhood.

To overcome the two aforementioned issues, it seems reasonable to leverage the idea of consensus, in terms of embedding it into a flexible way of visualization. Techniques of network analysis seem to be particularly promising.

The concept of network has already been used to assess linguistic data: the applications included an analysis of syntactic structures in English (Cancho i Ferrer, 2005), syntactic structures in Czech, German and Romanian (Cancho i Ferrer *et al.*, 2004), commonly occurring English adjectives and nouns (Newman, 2006, p. 14), word associations (Lai *et al.*, 2004; Lancichinetti, 2011, p. 17). Network analysis has been also used to compare differences between several texts in a corpus, namely, to

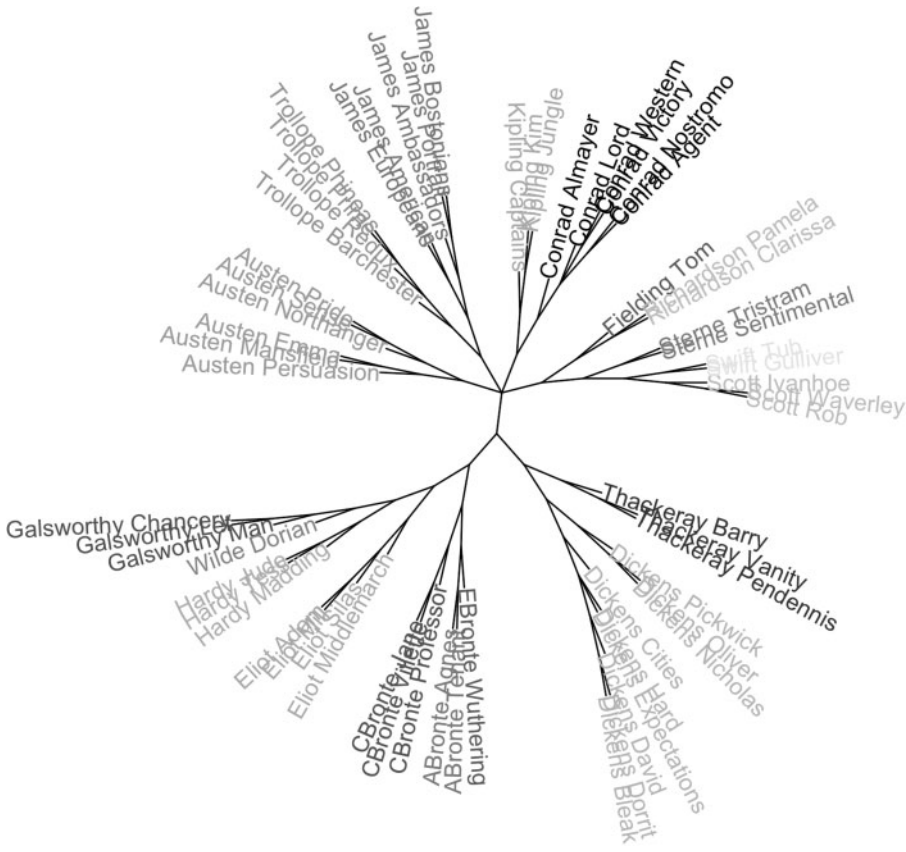

Fig. 5. Consensus tree of 66 English novels, 100–1,000 MFWs, classic Delta distance, Ward's linkage.

investigate the process of word network growth given a number of *n* sequences (Caldeira *et al.*, 2006), and recently to visualize relations in a corpus of a few hundred English novels (Jockers, 2013). The method introduced below is somewhat inspired by these studies. It relies on the assumption that particular texts can be represented as nodes of a network, and their explicit relations as links between these nodes. The most significant difference, however, between the approaches applied so far and the present study is the way in which the nodes are linked. This new procedure of linking is two-fold: one of the involved algorithms computes the distances between analyzed texts, the other is responsible for establishing a consensus of links.

A typical approach to authorship attribution involves a comparison of a disputed (anonymous) sample against a reference corpus, in order to identify the nearest neighbor of the disputed sample. To do this, stylometric distance between each pair of samples is estimated, and then the texts are ordered from the most to the least similar. To give an example: in the case of *The Jungle Book* by Kipling, the ranking begins with *Kim* (the nearest neighbor), the next is *Captains Courageous*, then *Lord Jim* by Conrad, and so on, and the last place in this procession is given to *Gulliver's Travels* by Swift. Each text in the corpus is associated with its own ranking of neighbors, from the nearest to the farthest one.

Now, these rankings can be reused to produce a stylometric network. In a simple variant, the links would be established between nearest neighbors only: Kipling's *The Jungle Book* connected to *Kim*, Hardy's *Far from the Madding Crowd* connected to *Jude the Obscure*, and so forth. However, since in literature-oriented studies, weaker or hidden textual

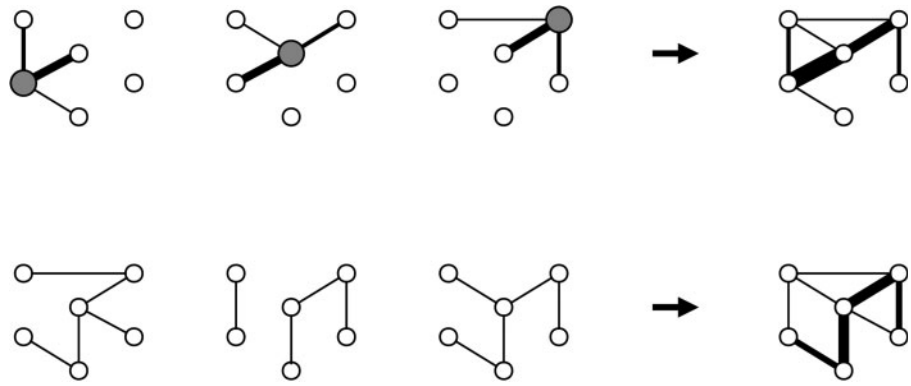

Fig. 6. Two algorithms of mapping textual relations: establishing weighted links to a nearest neighbor and two runners-up (top); producing a consensus network (bottom).

relations are potentially more interesting than explicit similarities, it makes sense to use the rankings more extensively. In stylometric terms, it means that runners-up (i.e. a few texts that have been ranked immediately after the nearest neighbor) should not be excluded from the analysis, even if, in typical approaches to classification, these runners-up are considered as unwanted noise and routinely filtered out.

Let the algorithm establish, then, for every single node, a strong connection to its nearest neighbor (i.e. the most similar text), and two weaker connections to the 1st and the 2nd runner-up. The outline of the algorithm is represented in Fig. 6 (top). Consequently, the final network will contain a number of weighted links, some of them being thicker (close similarities), some other revealing weaker connections between samples. Arguably, in most literary analyses, the thick connections will betray authorial similarities (usually the strongest stylometric signal), while thin links will reflect hidden layers of subtle intertextual correlations. In this article, it is assumed that three neighbors—a nearest one and its two runners-up—provide enough information about weaker similarities. However, one can set any number of neighbors to be connected. An empirical comparison of different ways of connecting the nodes will be discussed in a separate study.

The second algorithm (Fig. 6, bottom) is aimed at overcoming the problem of unstable results. It is an implementation of the idea of consensus dendrograms as discussed above into network analysis. The goal is to perform a large number of tests for similarity with different number of features analyzed (e.g. 100, 200, 300, . . . , 1,000 MFWs). Finally, all the connections produced in particular 'snapshots' are added, resulting in a consensus network. Weights of these final connections tend to differ significantly: the strongest ones mean robust nearest neighbors, while weak links stand for secondary and/or accidental similarities. Validation of the results—or rather self-validation—is provided by the fact that consensus of many single approaches to the same corpus sanitizes robust textual similarities and filters out apparent clusterings.

The two algorithms combined, one is presented with a robust picture of actual (strong) clusterings, emerging from an ethereal web of weaker stylistic similarities in the background. The above two-fold procedure of linking is implemented in the package 'stylo', an open-source stylometric library written in the R programing language (R Core Team, 2013) and available at CRAN repository (http://cran.r-project.org).[2]

The next crucial step in network analysis is to arrange the nodes on a plane in such a way that they reveal as much information about linkage as possible. Apart from very small networks that can be arranged manually, usually an algorithmic layout is applied. In the present study, one of the force-directed layouts was chosen, namely the algorithm

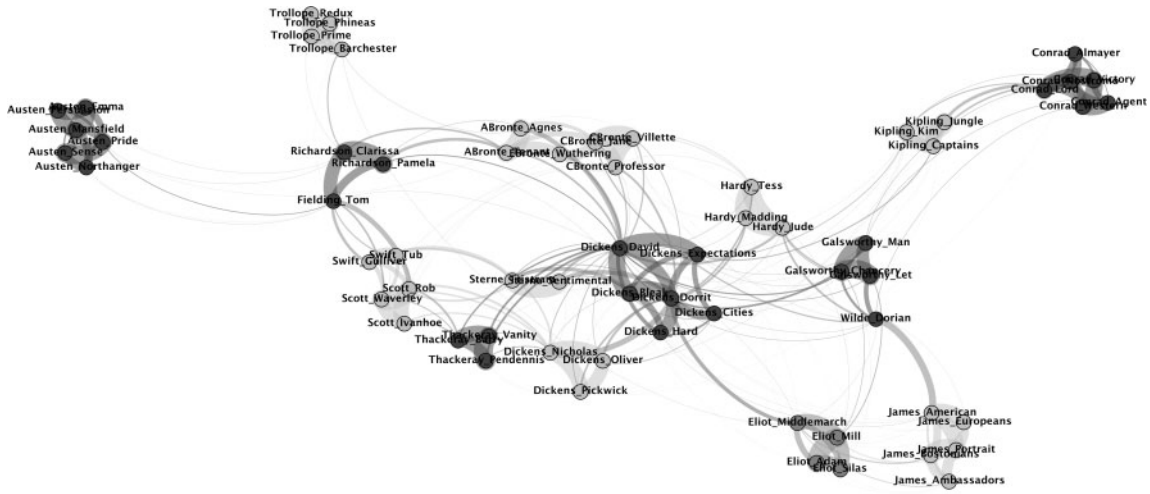

Fig. 7. Consensus network of 66 English novels: classic Delta distance, 100–1,000 MFWs, modularity 0.5.

ForceAtlas2 embedded in GEPHI, an open-source tool for network manipulation and visualization (Bastian *et al.*, 2009). Force-directed layouts perform gravity-like simulation and pull the most-connected nodes (i.e. the ones that have several links and/or their links are very strong) to the center of the network, while the least connected nodes are pushed outside.

A network produced using the above procedure is fairly informative *per se*: it usually reveals some clusterings discoverable with the naked eye, some centrally located nodes as well as peripheries, some denser and sparser areas, and so forth. At the same time, however, such a network can be subjected to a variety of standard measures used in networks analysis, which make the interpretation of the results more complete. These include measures of network size, its density, centrality of the nodes (closeness, betweenness, degree), and others. The measure of modularity, used as a community detection tool, might be particularly helpful to interpret clusters of stylistically similar texts.

In Fig. 7, a network of 66 English novels produced using the above procedure is shown. Spatial arrangement of the nodes was established by the said force-directed layout, and the nodes' colors were assigned according to the modularity measure. The network is clearly split into a few groups that obviously confirm the predominance of authorial signal in the data set. What is more interesting, however, is the relations between particular authorial clusters—and this is one notable advantage of networks over consensus trees. The outliers include Austen, Trollope, James, and Conrad, while the central parts are occupied mostly by the works of Dickens and Sterne. A circle of immediate satellites formed by Hardy, Galsworthy, the Brontës, Richardson, Fielding, and Thackeray is also noteworthy. Moreover, modularity-based color assignment sheds new light onto the already-interesting picture: while different works of a given author are usually recognized to form a distinct group, notable exceptions include a common cluster for Richardson, Fielding, Swift, and Scott; another common cluster is formed by the Brontë sisters, and the Dickensian oeuvre is split into two discrete groups (quite well connected with each other, though). Last but definitely not least, the network clearly shows a chronological pattern undiscoverable using consensus trees: a diagonal timeline beginning at the left side of the network, i.e. the late 18th-century area occupied by Fielding, Richardson, and Swift, through the Victorians (roughly in the middle), all the way to the early modernist Joseph Conrad.

Modularity is not the only way in which stylistic properties of particular texts/nodes can be assessed.

Another useful yet extremely simple measure is the degree or the number of connections that a particular node has. The real potential of this measure, however, comes on stage when the nodes are re-linked to form a directed consensus network.

6 Directed Network, or Seeking Stylistic Hubs

In the variant of a network discussed so far, all the connections of particular 'snapshots' were simply added, regardless of their direction. It means that any two nodes are connected no matter if the node *A* points to *B* as its neighbor, or if is pointed to by *B*. It is true that in most cases the relation between the nodes is mutual. However, since the rankings of candidates are calculated independently for every single text in a corpus, some non-symmetrical relations might occur as well. This is particularly the case when untypical texts are analyzed: such a text will point to its nearest neighbors anyway, but it would hardly be pointed to by other texts. Arguably, a directed network will discover such situations.

The procedure of establishing the connections does not differ from the undirected variant as introduced above, except that the direction of the links is recorded. Also, any mutual relations are not summed into one connection, but kept as two independent links: $A \rightarrow B$ and $A \leftarrow B$. Consequently, every single node will have, by definition, at least three outcoming links pointing to the nearest neighbor and to two runners-up. It is possible, however, that a minority of well-defined nodes might send numerous links in different directions, while others would constantly point to but three neighbors. And the other way around: it is possible that some nodes receive a vast majority of links from the entire network, while other nodes remain unpointed. In other words, measuring the number of connections of particular nodes should lead to identifying 'hubs', or texts that are stylistically followed (high incoming degree), and the stylistic followers (high outcoming degree).

In Fig. 8, a directed consensus network with node coloring according to outdegree is shown. One can easily identify a few hotspots—they represent the 'radiating' hubs, or the texts from which the number of outcoming links is the highest. These are: *Dorian Gray* by Wilde (12 links), *Sentimental Journey* by Sterne (10), *Kim* by Kipling (10), *Tom Jones* by Fielding (9), and *Agnes Grey* by Anne Brontë (9).

It is easy to explain the behavior of *Dorian Gray* and *Tom Jones*, one might say, since these are the only novels by Wilde and Fielding, respectively, included into the corpus. In the absence of natural nearest neighbors—i.e. other texts written by the same author—the analyzed novels blindly seek any similarities around. On the other hand, however, this does not apply to *Wuthering Heights*, the only novel of Emily Brontë: she turns out to be surprisingly introvert, with her mere five outcoming links, while her elder sister sends links to nine novels by Austen, Eliot, Trollope, Dickens, and Charlotte Brontë. It is also surprising to see the extroversion of Sterne's *Sentimental Journey*, especially when compared with a very modest behavior of *Tristram Shandy*.

Since the procedure of linking the nodes is based on classification principles, the existence of radiating hubs betrays the texts likely to be misclassified in a real-case authorship attribution study. A provisional interpretation of this phenomenon is that a given text turns into a radiating hub whenever it lacks in strong authorial signal, or when its authorial voice is overshadowed by other signals: genre, gender, chronology, and so forth. Needless to say that the ability of detecting radiating hubs makes this technique a potentially useful addition to authorship attribution toolbox—as a straightforward way to identify unstable samples.

From a literary point of view, however, the incoming links are potentially much more interesting, especially when they happen to form any 'absorbing' hubs. Such a hub represents a text pointed out as the nearest neighbor by several other texts from the corpus. Measure of incoming links, or indegree, applied to the corpus of 66 English novels is represented in Fig. 9. Two major absorbing hubs can immediately be spotted; they focus on two novels by Dickens, *David Copperfield,* and *Little Dorrit.* Two other hotspots are also fairly noticeable, namely

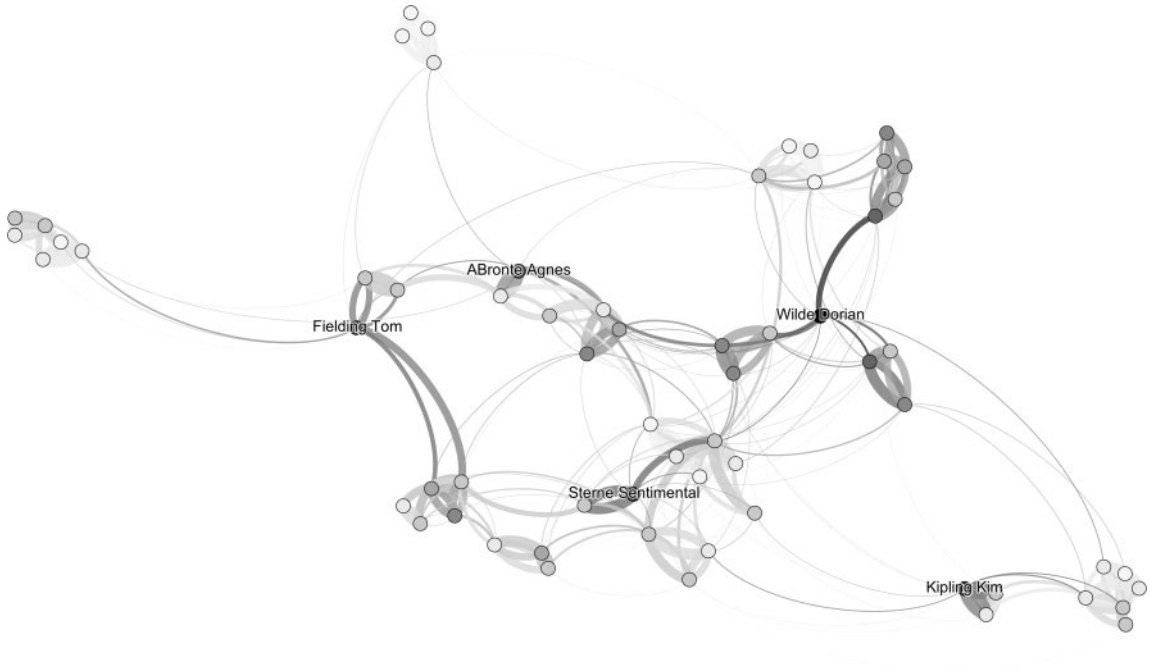

Fig. 8. Consensus network of 66 English novels (directed): the degree of outcoming links marked in color.

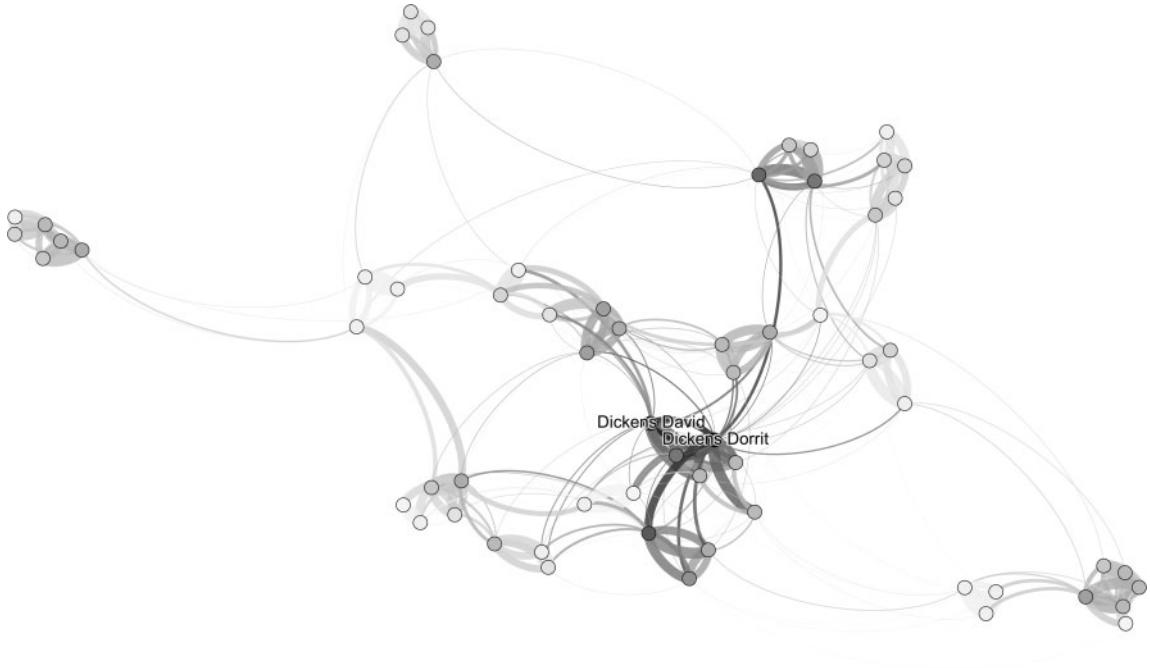

Fig. 9. Consensus network of 66 English novels (directed): the degree of incoming links marked in color.

Middlemarch by Eliot and *Nicolas Nickleby*, again by Dickens. Poorly connected novels found their place on the other pole of the indegree measure: *Dorian Gray* by Wilde (no incoming links at all), Sterne's *Sentimental Journey* (a single yet very strong incoming link from *Tristram Shandy*), and Swift's *Gulliver* (a single strong link from *A Tale of a Tub*).

Unlike radiating hubs, the absorbing ones are harder to interpret. In social sciences, physics etc., the hubs are usually considered to betray the most important events/agents/phenomena. In stylistics, however, what they really mean remains largely open to dispute. Jockers's approach to the question of literary influence seems to assume that the hubs indicate the most influential works (Jockers, 2013, pp. 154–168). Arguably, however, the picture is far more complex here.

The most striking observation is that according to the incoming links, Dickens would have had to live much earlier to have influenced Richardson, Sterne, or Swift. Is it the method, then, that is wrong, or the interpretation? In the aforementioned study on literary imitation, Jockers filters out all textual similarities that could not have happened due to chronological reasons, before undertaking actual analysis (Jockers, 2013, p. 163). However, discarding the backward time links cannot deny the fact that they do appear in the corpus.

It seems reasonable to assume that the absorbing hubs should be interpreted as sources of stylistic influence in a very broad sense, for instance as witnesses of stylistic mood of an entire literary epoch. It is true that these hubs might indeed indicate the most influential texts (copied, paraphrased, sequelled, consciously/unconsciously imitated, and so forth). At the same time, however, they might also reflect texts stylistically 'average', typical for their times rather than exceptional. In any case, the absorbing hubs betray texts lacking in a single, distinct stylistic signal.

A slightly oversimplified interpretation of both types of hubs might be as follows. The absorbing hubs stand for receivers of stylistic appreciation (regardless of their actual stylistic quality), radiating hubs represent emitters of stylistic appreciation (not mere followers, though, since they do not follow a single author).

7 Conclusions

In the present study, a few reliability issues of explanatory methods used in stylometry were discussed. They include unstable output—because final results highly depend on the setup of the experiment—as well as lack of validation. A promising way of extending cluster analysis dendrograms with a self-validating procedure involved producing numerous particular 'snapshots', or dendrograms produced using different input parameters, and combining them all into the form of a consensus tree. This approach, however, inherits some drawbacks of cluster analysis—dependence on a chosen linkage algorithm being the most painful—and introduces a few new pitfalls: granulation of clusters, and cluttered visualization when a corpus becomes large.

Significantly better results were obtained using a new visualization technique, which combines the idea of nearest neighborhood derived from cluster analysis, the idea of hammering out a clustering consensus from bootstrap consensus trees, with the idea of mapping textual similarities onto a network. Additionally, network analysis seems to be a good solution for large data sets.

The added value of consensus trees over standard dendrograms is the reliability of the results represented in a plot, and the added value of stylometric consensus networks is at least three-fold: the reliability inherited from consensus trees, insight into a more complete picture of textual relations beyond mere nearest neighborhood, and, last but not least, the capability of handling dozens, or even hundreds, of text samples in a single plot. The only limitation here seems to be the paper size one wants to use for drawing a literary network. Regardless of the printing issues, however, the aim of this study was to encourage stylometrists to produce a reliable map of literature in its entirety, and to propose a methodological background for such a map.

Acknowledgements

The idea of enhancing clustering procedures with network analysis has been developed during my visiting fellowship at the eHumanities Group (Royal Netherlands Academy of Arts and Sciences, The

Netherlands): I am grateful to Sally Wyatt, Andrea Scharnhorst, and Karina van Dalen-Oskam for the many inspiring discussions we had during my stay in Amsterdam. I am also grateful to the anonymous reviewers of this article for their valuable suggestions.

References

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

Bastian, M., Heymann, S. and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In: *International AAAI Conference on Weblogs and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154 (accessed 30 June 2014).

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

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

Burrows, J. (2004). Textual analysis. In: Schreibman, S., Siemens, R. and Unsworth, J. (eds), *A Companion to Digital Humanities*. Oxford: Blackwell, pp. 323–47.

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

Caldeira, S. M. G., Petit Lobão, T. C., Andrade, R. F. S., Neme, A. and Miranda, J. G. V. (2006). The network of concepts in written texts. *European Physical Journal B*, 49: 523–529.

Cancho i Ferrer, R. (2005). The structure of syntactic dependency networks: Insights from recent advances in network theory. In: Levickij, V. and Altman, G. (eds), *Problems of Quantitative Linguistics*. Chernivtsi: Ruta, pp. 60–75.

Cancho i Ferrer, R., Solé, R. V. and Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E*, 69: 1–8.

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

Dalen-Oskam, K. Van (2014). Epistolary voices: The case of Elisabeth Wolff and Agatha Dekken. *Literary and Linguistic Computing*, 29(3): 443–51.

Dunn, M., Terrill, A., Reesink, G., Foley, R. and Levinson, S. (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309: 2072–75.

Eder, M. (2013a). Mind your corpus: Systematic errors in authorship attribution. *Literary and Linguistic Computing*, 28(4): 604–14.

Eder, M. (2013b). Computational stylistics and Biblical translation: How reliable can a dendrogram be? In: Piotrowski, T. and Grabowski, Ł. (eds), *The Translator and the Computer*. Wrocław: WSF Press, pp. 155–70.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2): 167–82.

Eder, M. and Rybicki, J. (2013). Do birds of a feather really flock together, or how to choose test samples for authorship attribution. *Literary and Linguistic Computing*, 28(2): 229–36.

Hoover, D. (2003a). Multivariate analysis and the study of style variation. *Literary and Linguistic Computing*, 18(4): 341–60.

Hoover, D. (2003b). Frequent collocations and authorial style. *Literary and Linguistic Computing*, 18(3): 261–86.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.

Jockers, M., Witten, D. and Criddle, C. (2008). Reassessing authorship of the 'Book of Mormon' using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing*, 23(4): 465–91.

Kestemont, M., Luyckx, K., Daelemans, W. and Crombez, T. (2012). Cross-genre authorship verification using unmasking. *English Studies*, 93(3): 340–56.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60: 9–26.

Lai, Y.-C., Motter, A. E. and Nishikawa, T. (2004). Attacks and cascades in complex networks. In: Ben-Naim, E., Frauenfelder, H. and Toroczkai, Z. (eds), *Complex Networks*. Berlin–Heidelberg: Springer, pp. 299–310.

Lancichinetti, A., Radicchi, R., Ramasco, J. J. and Fortunato, S. (2011). Finding statistically significant communities in networks. *PLoS One*, 6(4): 1–18.

Luyckx, K. (2010). *Scalability Issues in Authorship Attribution*. Diss. University. Antwerpen.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London–New York: Verso.

Mosteller, F. and Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist*. Reprinted with a new introduction by John Nerbonne. Stanford: CSLI Publications, 2007.

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74: 1–19.

Noecker, J., Ryan, M. and Juola, P. (2013). Psychological profiling through textual analysis. *Literary and Linguistic Computing*, 28(3): 382–7.

Paradis, E., Claude, J. and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20: 289–90.

Pennebaker, J. W. (2011). *The Secret Life of Pronouns: What our Words Say About Us*. New York: Bloomsbury Press.

Posavec, S. (2007). Literary organism. http://www.stefanieposavec.co.uk (accessed 30 June 2014).

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org/ (accessed 30 June 2014).

Rudman, J. (1998a). Non-traditional authorship attribution studies in the 'Historia Augusta': Some caveats. *Literary and Linguistic Computing*, 13(3): 151–57.

Rudman, J. (1998b). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31: 351–65.

Rudman, J. (2003). Cherry picking in nontraditional authorship attribution studies. *Chance*, 16: 26–32.

Rybicki, J. (2012). The great mystery of the (almost) invisible translator: Stylometry in translation. In: Oakes, M. and Ji, M. (eds), *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins, pp. 231–50.

Rybicki, J. and Heydel, M. (2013). The stylistics and stylometry of collaborative translation: Woolf's 'Night and Day'. *Literary and Linguistic Computing*, 28(4): 708–17.

Schöch, C. (2013). Fine-tuning our stylometric tools: Investigating authorship, genre, and form in French classical theater. In: *Digital Humanities 2013: Conference Abstracts*. University of Nebraska–Lincoln, pp. 383–86.

Sinclair, S. and Rockwell, G. (2014). Voyant tools. http://voyant-tools.org (accessed 30 June 2014).

Tabata, T. (2012). Approaching Dickens' style through random forests. In: *Digital Humanities 2012: Conference Abstracts*. Hamburg: University of Hamburg, pp. 388–91.

Vickers, B. (2011). Shakespeare and authorship studies in the twenty-first century. *Shakespeare Quarterly*, 62: 106–42.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58: 236–44.

Notes

1 An earlier version of the Section 2 has been published in a paper discussing relations between the Greek New Testament and its Latin translation (Eder, 2013c).

2 The newest versions of the package 'stylo' are posted at the Computational Stylistics Group webpage (https://sites.google.com/site/computationalstylistics/), with a concise manual, installation instructions, and other supplementary materials.