Stylo visualisations of Middle English documents

Martti Mäkinen

Journal of Data Mining and Digital Humanities

Stylo visualisations of Middle English documents

Martti Mäkinen[1*]

1 Hanken School of Economics, Finland

*Corresponding author: Martti Mäkinen makinen@hanken.fi

Abstract

Automated approaches to identifying authorship of a text have become commonplace in the stylometric studies. The current article applies an unsupervised stylometric approach on Middle English documents using the script Stylo in R, in an attempt to distinguish between texts from different dialectal areas. The approach is based on the distribution of character 3-grams generated from the texts of the corpus of *Middle English Local Documents* (MELD). The article adopts the middle ground in the study of Middle English spelling variation, between the concept of relational linguistic space and the real linguistic continuum of medieval England. Stylo can distinguish between Middle English dialects by using the less frequent character 3-grams.

keywords

Middle English, historical dialectology, diatopical variation, non-standard spelling, stylometry, authorship attribution, R

1. Introduction

Until recently, in the study of Middle English spelling variation the concept of "linguistic space" has prevailed, which provides a relational map of attested spelling variants, but does not pin-point the text on a location on a geographical map ([LALME; Williamson 2000: 144-146]). Of late, also the real linguistic continuum of medieval England has become an object of interest in studies that answer questions related to the actual provenance of the extant texts ([Stenroos and Thengs, 2012]). The corpus of *Middle English Local Documents* ([MELD]) is being compiled at the University of Stavanger to answer questions of the latter type.

The current paper takes the middle road, using the material collected by the [MELD] project, but employing Stylo, an unattended script for R ([Eder, Rybicki and Kestemont, 2014]), originally designed for authorship attribution. Stylo creates a relative network of texts, relying on either word or character n-grams, and therefore disregarding the provenance of the texts. This paper argues for the usefulness of the method in historical dialectology and in the study of document texts, and shows how the script can be used to group and visualize the groupings and relations of [MELD] texts according to both the text category and diatopical variation. The images of textual networks are enhanced by the help of Gephi, another tool for visualizing networks ([Bastian, Heymann and Jacomy, 2009]). The method has earlier been tested on Middle English document categories ([Mäkinen, 2019]).

2. What is Stylo?

Stylo is a script for R (originally developed by [Eder, Kestemont, and Rybicki]; currently maintained by Eder), intended for authorship attribution. Stylo algorithm recurses the input data several times in order to establish links between texts, i.e. a link is not established unless it has been corroborated

several times by other similarities between the texts ([Eder, Rybicki and Kestemont, 2016, 112–117]). By way of visualizations, Stylo provides a selection of diagrams (MDS maps, dendrograms, cluster analysis diagrams, etc.) that illustrate the relative distance between texts in a 2-dimensional space. In addition to the diagrams, Stylo also provides results in word lists (organised according to frequency), logfiles that record the texts' potential for textual affinity, and tables of standard scores (z-scores) ([Eder, Rybicki and Kestemont, 2017, 4, 7]).

Stylo depends on word or character n-grams: the analysis is based on the extraction of character or word n-grams and the comparison of the n-gram standard scores. In the process of the analysis, the script first creates a list of unique words or n-grams (up to a desired frequency of n-grams, depending on the parameters set) in the data, with frequencies per individual text. The text-specific n-gram frequencies are turned into standard scores to normalize the results, which are then used in the various automated analyses (e.g. cluster analysis, multidimensional scaling, or principal components analysis, again according to the parameters chosen). Finally, the end product are diagrams that visualize the affinity or relative distance between the texts studied. ([Eder, Rybicki and Kestemont, 2017, 4, 7])

3. Stylo and Middle English

Stylometric analysis is, perhaps, not the obvious choice of method for analysing Middle English texts, which are characterized by diatopical or regional variation. Of course, the texts attest to authorial, idiosyncratic variation, temporal variation and genre variation as well; however, the strongest differentiating factor in Middle English spelling variation is usually regional variation.

This paper is not the first time a computational authorship attribution method has been applied on Middle English texts: the Paston letters have subjected to such an analysis ([Juola, 2008, 289–290]), and also Middle English *Pearl* poems have been studied through authorship attribution ([McColly and Weier, 1983]). The current data set has been analysed with Stylo before, in an attempt to differentiate automatically between Middle English document genres ([Mäkinen, 2019]). Other studies in which Stylo has been applied on historical texts before are a study on Hildegard of Bingen and Guibert of Gembloux's texts ([Kestemont, Moens and Deploige, 2015]), and another on the letters and prose of Queen Katherine Parr and Princess Elizabeth ([Evans, 2016]).

As previously stated, Stylo relies on word or character n-grams. As the dialectal variation in Middle English texts is reflected in spelling, any lemma of a lexical item becomes a lengthy list of items if texts from all dialect areas are observed for the lemma. Even a lemma of one lexical item per one text often contains a few spelling variants, i.e. the authors/scribes writing in Middle English often used alternative variants in one text ([Mäkinen, 2019, 154]).[1] This entails that using word n-grams in the Stylo analysis of unannotated Middle English texts would split the relevant information between so many spelling variants that the method would lose its analytical power. Therefore, character n-grams are the only viable alternative for an unattended Stylo analysis of this data.[2]

1 E.g. in the current material, a letter written in Durham (Durham, Prior's Kitchen, Locellus IX.67) contains the forms *hade writyn* and *had wryttyn* for the PDE *had written*.

2 Other approaches could be used if the intention of the analysis was to identify texts by the same scribe or author. For this purpose normalizing the spelling variants would be possible. Normalization has been tried on historical corpora before, e.g. the C*orpus of Early English Medical Texts* ([EMEMT]) provides a spelling-normalized flavour of the corpus, created with VARD2 ([Baron and Rayson, 2008]). Nevertheless, normalization would render the data unusable for the study of Middle English dialectology as it is the spelling variation that is regionally conditioned in Middle English texts.

In the earlier study on Middle English document categories and Stylo [Mäkinen, 2019] it became apparent that Stylo works on texts that are encumbered by spelling variation, at least in genres that have a lot of recurring, restricted vocabulary. This makes documents an ideal candidate for the approach, as the formulaicness of documents results in repetition of terminology from text to text. The restricted vocabulary has probably led to an early standardization of terminology in Middle English documents, and that has enriched the occurrence of certain spelling forms. This, obviously, benefits the analysis with Stylo, and the automatic detection of the different document categories.

4. MELD and analysis of data

The data for this study is from *A Corpus of Middle English Local Documents* (henceforth [MELD]), which is being compiled at the University of Stavanger. It contains documentary texts from 1400 to 1525. The version used in this study is 2017.1, consisting of 2,017 localizable scribal documents, and c. 850,000 words ([MELD]). Localizable documents are texts that either contain the information on the provenance of the document in the actual text, or provide circumstantial information about the provenance (through the use of personal and place names) so that localizing the origin of the document is, by some certainty, possible ([Stenroos and Thengs, 2012]).

The analysis of MELD data is based on two assumptions ([Mäkinen, 2019, 152]):

1) each MELD text attests to a unique set of character n-grams
2) such unique sets are more similar among texts that share a similar language variant

The use of words or characters as the basis of n-grams in authorship attribution has been discussed widely in earlier literature, (see e.g. [Hoover 2002, 2003, 2012], [Koppel et al., 2009], [Stamatatos, 2009], [Eder, 2011], and [Alexis et al., 2014]). The use of character n-grams means that the units of analysis have very little to do with the linguistic units of syllables, morphemes, and words ([Eder 2015]), i.e. the current approach is not intended for a study of style observable in these linguistic elements. The main object of the current analysis is the comparison of "spelling fingerprints" of texts as defined through the chosen character n-grams. Indeed, [Eder, 2013] has argued for using character n-grams with so-called "dirty" corpora, i.e. with text data that contain a lot of spelling variation, due to e.g. sub-standard optical character recognition in corpus compilation, or other reasons (see also [Juola, 2008, 285]).

3-grams not crossing word boundaries seem to provide the best discrimination of the material ([Mäkinen, 2019, 156]). The 3-grams used in the analysis are of the type

1) "Prefix" 3-grams: **BRO**KENn **CHA**LESSE
2) "Suffix" 3-grams: BROK**ENn** CHALE**SSE**
3) Word-internal 3-grams: B**ROK**ENn CH**ALE**SSE.

The words "Prefix" and "Suffix" are quoted as they do not refer to the morphological entities denoted by the terms.

5. Analysis

As has been mentioned, in earlier studies Stylo has been able to distinguish between different Middle English document sub-genres. Figure 1 (as presented in [Mäkinen, 2019, 162]) illustrates the created groupings in an MDS graph.

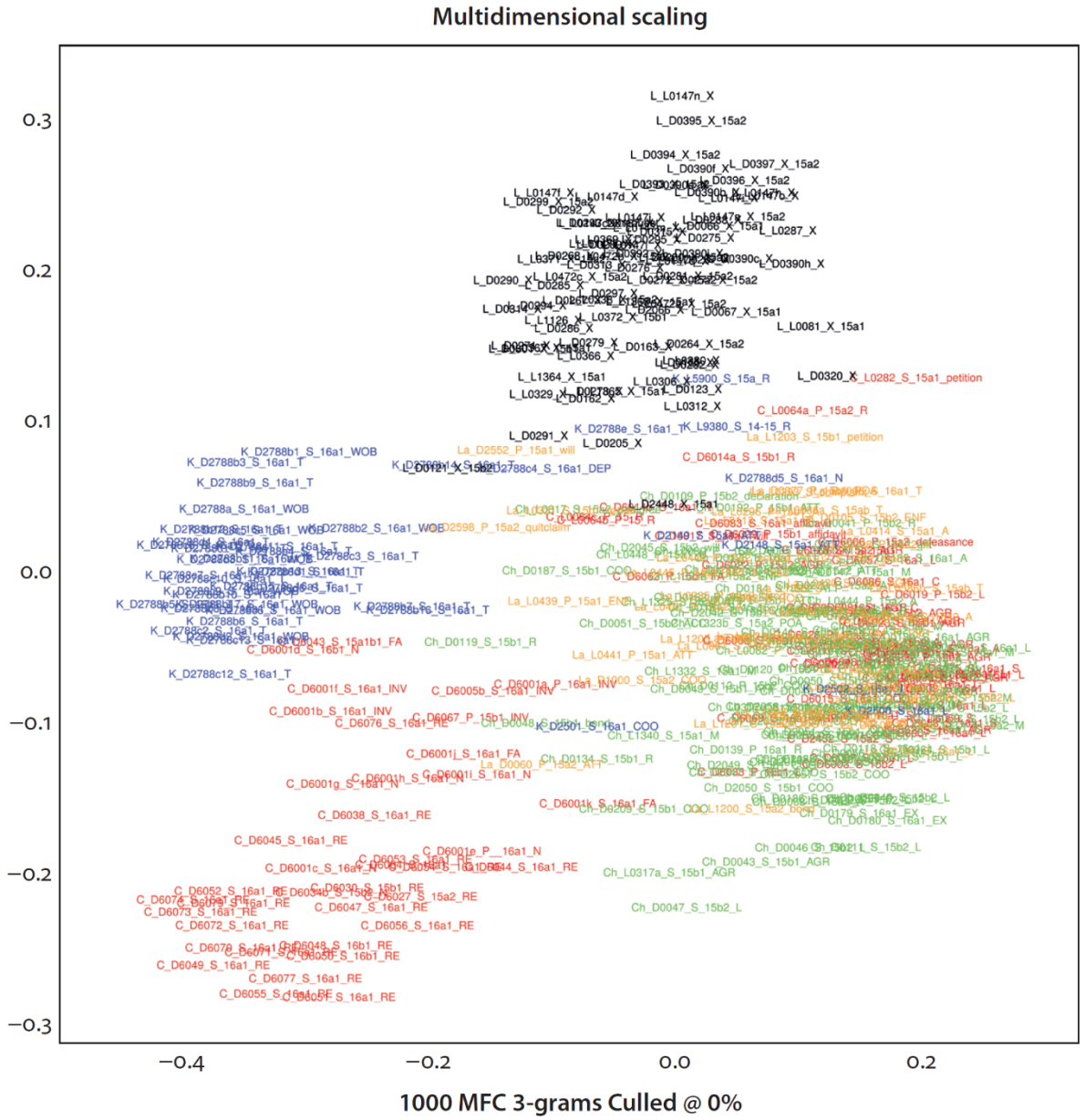

Figure 1. MELD subcorpus of letters, Cambridgeshire, Cheshire, Kent and Lancashire texts, 1000 most frequent 3-grams, Burrow's Delta (Source: [Mäkinen, 2019, 162])

Legend: Counties: C = Cambridgeshire, Ch = Cheshire, K = Kent, La = Lancashire / Functional labels: A = award, AGR = agreement, ATT = attestation, COO = condition of obligation, DEP = deposition, ENF = enfeoffement, EX = exchange, FA = financial account, INV = inventory, L = lease, M = marriage article, POA = power of attorney, R = rental, RE = receipt, T = testimony, WOB = wow of betrothal.

For Figure 1, a subset of MELD consisting of 389 texts was used (totalling c. 172,000 words), including texts from the counties of Cambridgeshire, Cheshire, Kent and Lancashire, and the category of letters in addition. In the figure, texts from the county of Kent are mostly on the left, in blue; the red-label Cambridgeshire texts are situated down on the left. On the right, in the middle, there is a mixed bag of orange Lancashire, green Chesire, red Cambridgeshire, and a few blue Kent text labels. At the top, there is the group of letters, in black. Despite the apparent grouping of texts by counties, the best explaining factor for these groupings are documentary sub-genres: texts from Kent are almost

exclusively testimonies, whereas Cambridgeshire texts lower left are almost invariably receipts. This means that we can see groupings defined by textual functions in Figure 1.

The divisions seen in Figure 1 are between left and right, and between up and down. The former division is between statements (testimonies, receipts, inventories, vows of bethrothal, and financial accounts) and conveyances (leases, conditions of obligation, agreements, marriage articles, bonds, enfeoffements etc.), i.e. documents that recorded transferral of rights and agreements between two or more parties. The up-down division in Figure 1 is between correspondence and other documents, reflecting (most likely) the difference between letter formulae and document formulae.

Figure 1 was created by analysing a selection of MELD texts according to 1,000 most frequent 3-grams, starting from rank 1. Thus, a Stylo analysis with an unrestricted n-gram list is able to distinguish between different textual categories, but not very well between the diatopically conditioned variants of Middle English: we do not see a grouping of texts according to dialect areas in Figure 1. The reasons for this are probably the frequent content words that recur in the document formulae. They may also have been more or less standardized already in the Late Middle English period, which would further explain the emerging sub-genre groupings in the figure. ([Mäkinen, 2019, 162])

The way the algorithm of Stylo works makes it challenging to distinguish between texts in different Middle English dialects without researcher intervention. In the analysis for Figure 1, we see different spaces of variation superimposed one another: the genre variation is mixed with dialectal variation, but as genre variation joins texts over dialect boundaries, it prevails in the figure, hiding the dialectal variation. Therefore, in order to be able to see the dialectal variation, one needs to manipulate the 3-gram list used. The guiding principle for the approach is the same as in [LALME], General Introduction: the 3-grams chosen need to be infrequent enough to be useful (i.e. to differentiate between dialects), but frequent enough not to be useless (they will still need to retain their analytical potential). This means that in search for suitable 3-grams one has to look beyond the top ranks of most frequent 3-grams, and focus on the levelling tail of the Zipf curve of rank/frequency ratio of 3-grams.

For Figure 2, a 3-gram list restricted to ranks from 50 to 200 were used, i.e. the list of 3-grams ordered according to frequency, the top 49 3-grams were left out, and the list was cut after the 200th 3-gram. This leaves the analysis with 802 different 3-grams (some of the low-frequency 3-grams share the same frequency, and thus also the same rank with other 3-grams). In addition to manipulating the 3-gram list, texts from one county were appended into one file. This further diminished the effect of genre variation, and boosted the method's ability to highlight the similarities and differences between different dialect areas. Also it provided a graph with 40 data points instead of the original 2,017 data points, thus enhancing the legibility of the graph.

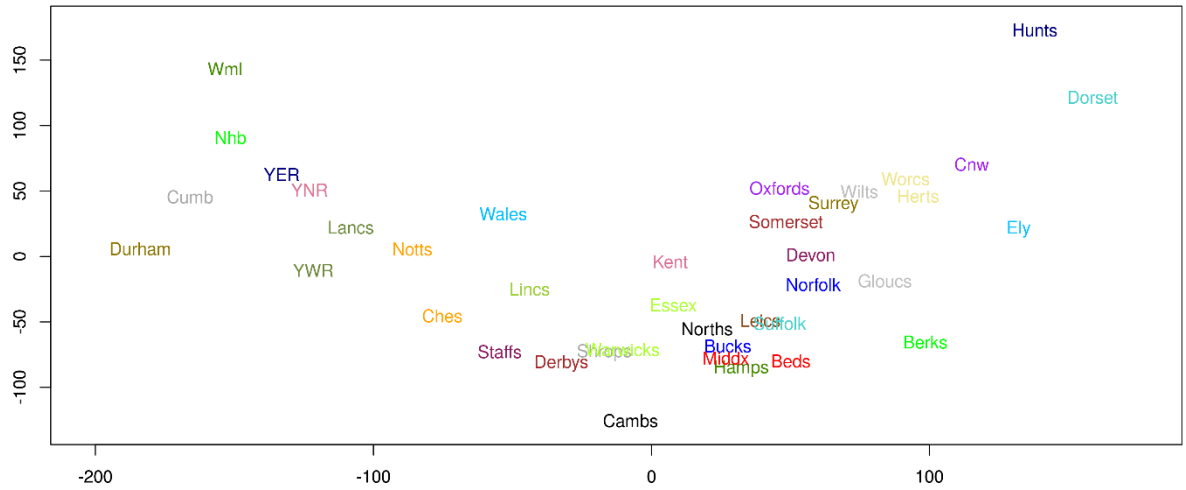

Figure 2. MELD corpus texts per counties MDS, 802 3-grams, ranks 50-200.

Legend: YER = Yorkshire, East Riding; YNR = Yorkshire, North Riding; YWR = Yorkshire, West Riding.

In Figure 2, the Stylo MDS creates the greatest distance between Northern and Southern counties, which was expected, and we see minimal variation between texts from Buckinghamshire, Middlesex, and Hampshire. Nevertheless, some of the counties are out of their expected places, and even if general interpretations of the textual network can be made based on the graph created, it does not tell anything about the strengths of the links between the texts used. Therefore, another tool, Gephi, was needed to visualize the information collected and provided, but not shown by Stylo.

Gephi is a tool designed for network manipulation and visualization ([Bastian, Heymann and Jacobi, 2009]). The advantage of Gephi is that it makes full use of the information produced by Stylo; in fact, Stylo also produces an EDGES file per each analysis, which can be fed as such into Gephi. In this way the visualizations can utilise and show the maximal amount of information accumulated in the analysis, e.g. how strong the links between the texts are. The algorithm used by Gephi, ForceAtlas2, recreates the textual network so that the text nodes with most links to other nodes gravitate to the centre of a graph, whereas the less linked nodes (the "lighter" nodes) are pushed to the fringes of the graph ([Eder, 2015, 10]).

Figure 3 is thus based on the information processed and accumulated by Stylo for Figure 2. In Figure 3, one can observe the way the ForceAtlas2 algorithm of Gephi works: On the whole, one can observe that the Northern text nodes are in the "North" of the graph, and the Southern texts are, more or less, situated in the "South". Futhermore, the Northern counties form a fairly close-knit network of their own. The graph reproduces the general, oblong shape of the map of England, even if no geographical data was fed into the system. Gephi also shows the strength of the links, providing ideas for the qualitative scrutiny of the data. In Figure 3, Warwickshire, Staffordshire, and Shropshire are the most linked text nodes, and therefore they can be found in the middle of the graph, leaving the less "weighty" text nodes towards the outer regions of the graph. This also means that the three counties commented on in Figure 2, Buckinghamshire, Middlesex, and Hampshire, are split around the fringes of the graph in Figure 3: the texts of these counties are less linked than the "link-heavy" counties, and therefore they will need to make way for Warwickshire, Warwickshire, Staffordshire, and Shropshire in the graph.

There are a few possible explanations for the out-of-place counties in relation to the real map of England (e.g. Oxfordshire, Cornwall, and Kent). In this approach it is fully possible that Figure 3 is influenced by 3-grams that are not conditioned by region nor genre, or they may be results of scribal mobility that can be seen as mixtures of regional varieties. However, the counties in question are all cases where the representativeness of the texts sampled for each of them seems to explain how they behave in the graph. Text nodes not in their expected places tend to be either for counties that are over-represented in the MELD corpus, or that are somewhat under-represented. This is a problem that almost any historical corpus faces, and it has been addressed also by the MELD compiler team. The situation will become better over time, as more texts will be added to MELD.

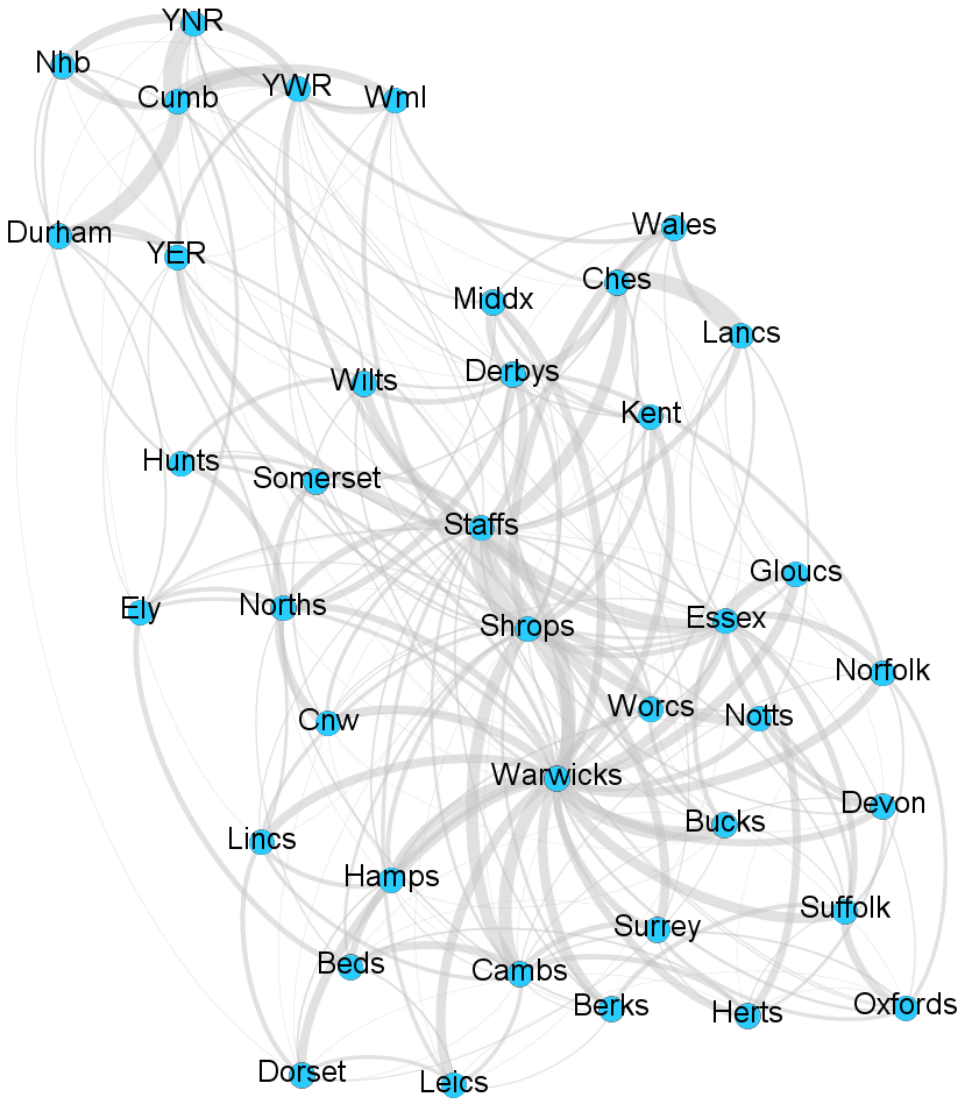

Figure 3. MELD corpus Gephi network, information drawn from Figure 2.

In reading Figure 3, one needs to bear in mind that with an automated, unattended approach on MELD texts one can see simultaneously the effects of diatopically, temporally, and idiosyncratically conditioned features, not forgetting the effect of sub-genre styles, i.e. the graph is not completely void

of interference from factors beyond the dialect variation. This happens, as the only selection criterion for the manipulated 3-gram list was rank. Nevertheless, even with the shortcomings of the approach the created graph conforms surprisingly well with our knowledge about diatopical variation in Middle English.

For the final figure, Figure 4, the sub-genre division illustrated in Figure 1 will be revisited: the documentary sub-genre graph in Figure 1 is re-rendered in Gephi to create Figure 4.

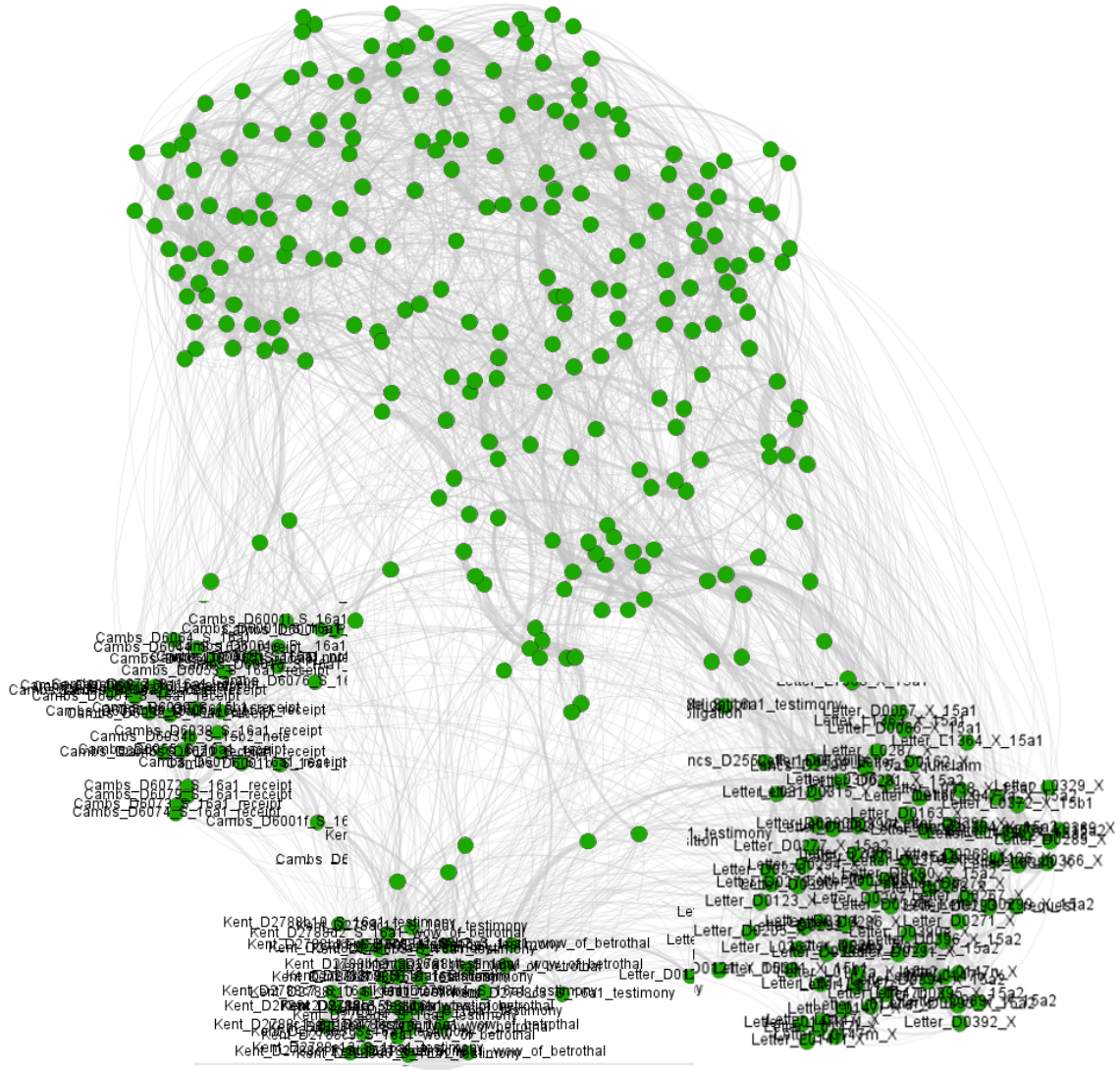

Figure 4. Gephi network based on Figure 1.

In Figure 4, the different sub-genres stand out as groups of their own: from left to right, receipts, testimonies, and letters. The first two belong to the umbrella category of statements, but they are represented as distinct groups of their own in the graph. What is noteworthy here is the fact that for Figures 1 and 4, the texts were not compiled into one text file per county, and yet we can perceive the affinity of similar texts in the graph. The benefit of the Gephi network graph is that the link strengths

can also be observed, and that is valuable information for a potential future qualitative analysis of the text nodes.

6. Discussion

As mentioned earlier, the Stylo algorithm compares between the standard scores of each item of analysis, in this study, each character 3-gram. Also, the algorithm is recursive, i.e. it requires several instances of similarity between two texts before an affinity is established. Therefore, the more frequent n-grams figure more prominently in the analysis. That is the reason why a completely unattended Stylo analysis would produce clusters of Middle English document text categories or subgenres as the more standardized and frequent content words in documentary formulae will prevail in the data. Therefore, texts in the analysis should be more or less equally long: comparatively short documents will give in to the pull of longer documents, as they necessarily attest to fewer n-grams, and also fewer n-grams outside the documentary formulae.

The diatopically conditioned spelling variants in Middle English found useful for a Stylo analysis can be found among the less frequent n-grams: in this analysis the 3-gram ranks from 50 to 200 were used, i.e. some manipulation of the automatically created 3-gram list was needed. This diminished the effect of genre variation, even if it did not eradicate it completely. In making full use of the information accumulated by Stylo, Gephi graphs were rendered from the Stylo output. They can be created directly from the EDGES files created by Stylo, and they provide a valuable resource for further qualitative work, beyond the computational analysis.

After this analysis, one needs to find ways to deal with the overlapping spaces of variation. The manipulated 3-gram list was a step in that direction; nevertheless, even more can be done to reach the desired outcome. One thing to explore is chunking the corpus into equally long time periods, in order to diminish the effect of temporal variation on Middle English spelling. Another approach could be creating sub-corpora according to text categories. Both of these approaches are at the mercy of the data: there may not be enough text mass per category or per time period for a reliable analysis. This, however, can be remedied by the texts that are being added to MELD corpus: the text selection for the "small" counties is gradually becoming more representative. Finally, random sampling is a means to counter-balance the various lengths of texts, and it is a function that is already available in Stylo.

In future, the observations of the exploratory studies (this and [Mäkinen, 2019]) should be tested in follow-up, qualitative studies. In that way the current observations can be triangulated, and their value for historical dialectology properly assessed.

7. Conclusion

The current study explored an automated, unattended approach to grouping Middle English documentary texts, and ways to control the feature sets that are active in the analysis. In an automated system without researcher intervention, it is impossible to study e.g. regional variation and genre variation at the same time. Stylo can make a distinction between Middle English document categories; also distinguishing between Middle English dialects is possible, by using the less frequent n-grams. The results of this paper may pave way to future exploratory uses of various computational methods in variationist studies of historical texts.

References

Alexis A., Craig H. and Elliot J. Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution. *Literary and Linguistic Computing*. 2014;29(2):147–63. https://doi.org/10.1093/llc/fqt028.

Baron A. and Rayson P. Automatic standardisation of texts containing spelling variation: How much training data do you need? *Proceedings of the Corpus Linguistics 2009 Conference*, Lancaster University (Lancaster). 2009. https://eprints.lancs.ac.uk/id/eprint/41666/.

Bastian M., Heymann S. and Jacomy M. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. 2009. https://gephi.org/publications/gephi-bastianfeb09.pdf.

Eder M. Visualization in Stylometry: Cluster Analysis Using Networks. *Digital Scholarship in the Humanities*, 2015;1–15. https://doi.org/10.1093/llc/fqv061.

Eder M. Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing*, 2013;28(4):603-614. https://doi.org/10.1093/llc/fqt039.

Eder M. Style-Markers in Authorship Attribution A Cross-Language Study of the Authorial Fingerprint. *Studies in Polish Linguistics*. 2011;6:99–114.

Eder M, Rybicki J and Kestemont M. 'Stylo': a package for stylometric analyses. *Computational Stylistics Group*. 2017;1-36. https://tinyurl.com/y449xxkk.

Eder M., Rybicki J. and Kestemont M. Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 2016;8(1):107-121. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Embleton S., Uritescu D. and Wheeler E. The Stability of Multidimensional Scaling over Large Data Sets: Evidence from the Digitized Atlas of Finnish. In Havu E., Helkkula M., and Tuomarla U. (eds.). *Du côté des langues romanes. Mélanges en l'honneur de Juhani Härmä*, (Mémoires de la Société Néophilologique de Helsinki, 77). Société Néophilologique (Helsinki). 2009;101-108.

EMEMT = *Corpus of Early Modern English Medical Texts*. Taavitsainen, Irma, Päivi Pahta, Turo Hiltunen, Ville Marttila, Maura Ratia, Carla Suhr, and Jukka Tyrkkö (eds), with the assistance of Anu Lehto and Alpo Honkapohja. John Benjamins (Amsterdam). 2010.

Evans M. Tudor women writing: Multimodal style and identity in the English letters and prose of Queen Katherine Parr and Princess Elizabeth. In Nevala M., Lutzky U., Mazzon G. and Suhr C. (eds.). *The Pragmatics and Stylistics of Identity Construction and Characterisation*. (Studies in Variation, Contacts and Change in English 17).VARIENG (Helsinki). 2016. http://www.helsinki.fi/varieng/series/volumes/17/evans/.

Hoover D.L. Frequent Word Sequences and Statistical Stylistics. *Literary and Linguistic Computing*. 2002;17(2):157–180. https://doi.org/10.1093/llc/17.2.157.

Hoover D.L. Another perspective on vocabulary richness. *Computers and the Humanities*. 2003;37(2):151–178.

Hoover D.L. (2012). The Tutor's story: A case study of mixed authorship. *English Studies*. 2012;93(3):324–339. https://doi.org/10.1080/0013838X.2012.668791.

Juola P. Authorship Attribution. *Foundations and Trends in Information Retrieval*. 2008;1(3):233–334. https://doi.org/10.1561/1500000005.

Kestemont M., Moens S. and Deploige J. Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux. *Digital Scholarship in the Humanities*. 2015;30(2):199–224. https://doi.org/10.1093/llc/fqt063.

LALME = McIntosh, Angus, Michael L. Samuels and Michael Benskin. *A Linguistic Atlas of Late Mediaeval English*. 4 vols. Aberdeen University Press (Aberdeen). 1986.

Mäkinen M. Testing a stylometric tool in the study of Middle English documentary texts. In: Bös B. and Claridge C. (eds.). *Norms and Conventions in the History of English*. John Benjamins (Amsterdam). 2019;149-166.

McColly W.B. and Weier D. Literary Attribution and Likelihood-Ratio Tests: The Case of the Middle English Pearl Poems. *Computers and the Humanities*. 1983;(17):65–75. https://doi.org/10.1007/BF02277126.

MELD = *The Middle English Local Documents Corpus*, version 2017.1. June 2017, University of Stavanger (Stavanger). https://www.uis.no/research/history-languages-and-literature/the-mest-programme/a-corpus-of-middle-english-local-documents-meld/.

Stamatatos E. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*. 2009;60(3):538–556. https://doi.org/10.1002/asi.21001.

Stenroos M. and Thengs K.V. Two Staffordshires: real and linguistic space in the study of Late Middle English dialects. In Tyrkkö J., Kilpiö M., Nevalainen T. and Rissanen M. (eds.). *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. (Studies in Variation, Contacts and Change in English 10). VARIENG (Helsinki). 2012. http://www.helsinki.fi/varieng/series/volumes/10/stenroos_thengs/.

Williamson K. Changing spaces: Linguistic relationships and the dialect continuum. In Taavitsainen I., Nevalainen T., Pahta P. and Rissanen M. (eds.). *Placing Middle English in Context*. (Topics in English Linguistics, 35.) De Gruyter Mouton (Berlin). 2000;141-180.