Abstract ICEHL 20, AMC Workshop: Visualisations in Historical Linguistics

Martti Mäkinen
Hanken School of Economics

**Stylo visualisations of Middle English documents**

Middle English spelling variation has been the object of dialectal studies for a long time, and until recently the approach that employs 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 the map (LALME; Williamson 2000: 144-146). Of late, also the real, absolute linguistic continuum of medieval England has been raised as an object of interest, in particular to answer questions that are related to the actual provenance of the extant texts (Stenroos and Thengs 2012). The corpus of *Middle English Local Documents* (MELD), compiled at the University of Stavanger, is aimed to answer questions of the latter type.

This paper presents a follow-up study to a paper on Middle English diatopical and functional variation (Mäkinen 2017), as observed through the visualisations made with **stylo. Stylo** is a stylometric package written for R (Eder, Rybicki and Kestemont 2014). The paper tested character n-grams in mapping interesting groups of documents, drawn from MELD, in advance of in-detail, traditional analysis of texts. **Stylo** is able to discriminate between Middle English document genres in the visualisations it produces (e.g. MDS maps, dendrogrammes, and consensus trees) (Eder, Rybicki and Kestemont 2014; Embleton, Uritescu and Wheeler 2009); however, the way the algorithm of **stylo** works seems to make it incapable of discriminating between ME dialects. This paper will address the problem of diatopically interesting items invisible in **stylo** visualisations.

Originally, **stylo** is intended for authorship attribution (Juola 2008). Stylometry tools used for other purposes is less studied, even though the effect of dialect or genre can be detected by many of the current tools (e.g. Juola 2008, Eder 2015). In the analysis of ME documents, the infrequent items of personal style or of diatopical significance may be hidden behind the mass of purpose-of-writing-specific vocabulary that may have already been somewhat standardized in the late ME period. The fact that documents tend to be somewhat formulaic leads to the enrichment of text-category-specific features and their standardization to the extent that **stylo** relies on them in the analysis of documents. The forms that distinguish between diatopical variants of ME seem to be too infrequent to make a difference when using **stylo.**

The main aim of the paper is to identify the factors that make ME documents to cluster according to genres in a **stylo** analysis, and to discover ways to visualise the observable diatopical variation. This will require focusing on the long tail of n-gram inventories, where the infrequent items reside.

**References**

Eder, Maciej. (2015). Visualization in Stylometry: Cluster Analysis Using Networks. *Digital Scholarship in the Humanities*, 1–15. <https://doi.org/10.1093/llc/fqv061>. Accessed: Oct 10, 2017.

Eder, Maciej, Jan Rybicki and Mike Kestemont. 2014. ‘Stylo’: a package for stylometric analyses. *Computational Stylistics Group*. [Online]. Available at <https://sites.google.com/site/computationalstylistics/>. Accessed Dec 12, 2017.

Embleton, Sheila, Dorin Uritescu and Eric Wheeler. 2009. The Stability of Multidimensional Scaling over Large Data Sets: Evidence from the Digitized Atlas of Finnish. In: *Du côté des langues romanes. Mélanges en l’honneur de Juhani Härmä*, (Mémoires de la Société Néophilologique de Helsinki, 77). Eva Havu, Mervi Helkkula and Ulla Tuomarla (eds.). Helsinki, Société Néophilologique. 101-108.

Juola, Patrick. 2008. Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3), 233–334. <DOI: 10.1561/1500000005>. Accessed Dec 12, 2017.

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

Mäkinen, Martti, 2017. N-gram inventories in the study of Middle English variation. Paper presented at ICOME 10, University of Stavanger, May 31 – June 2, 2017.

MELD = *The Middle English Local Documents Corpus*, version 2017.1. June 2017, University of Stavanger.

Stenroos, Merja and Kjetil V. Thengs. 2012. Two Staffordshires: real and linguistic space in the study of Late Middle English dialects. _Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources_ (Studies in Variation, Contacts and Change in English 10), ed. Edited by Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen and Matti Rissanen. Helsinki: VARIENG. [Online]. Available at <http://www.helsinki.fi/varieng/series/volumes/10/stenroos_thengs/>. Accessed Dec 12, 2017.

Williamson, Keith. 2000. Changing spaces: Linguistic relationships and the dialect continuum. In: *Placing Middle English in Context*. Taavitsainen, Irma, Terttu Nevalainen, Päivi Pahta, Matti Rissanen (eds.) (Topics in English Linguistics, 35.) Berlin: De Gruyter Mouton. 141-180.