Abstract ICEHL 19
Martti Mäkinen
Hanken School of Economics

Testing a stylometric tool in the study of Middle English documentary texts

This paper investigates the usability of 'Stylo', a stylometric package written for _R_ (Eder, Rybicki and Kestemont, 2015) in the study of Middle English diatopical and other variation. The aim is to test the n-gram functionality in mapping potentially interesting groups of texts in advance of in-detail, traditional historical dialectal analysis of texts, based on the geographical variation in spelling (LALME).

The material used in this study is the Corpus of Middle English Local Documents (MELD), version 2015.1., containing English documentary texts (legal instruments, administrative writings, and personal letters) from the period between 1400-1525. The corpus contains currently 1,003 scribal texts that can be localised on non-linguistic grounds.

The analysis will be based on the extraction and comparison of character n-grams, the assumption being that each Middle English texts of the defined period will attest to a unique set of such n-grams. Such unique sets are also assumed to be more similar among texts that share a similar variant of Middle English, either diatopically conditioned or otherwise (Jensen, 2010; Stenroos and Thengs, 2012). The length of the n-gram will be of pivotal importance. In the visualisation of the data, the choice of the function has to be tested: 'Stylo' provides several functions to the analysis of n-grams, of which multidimensional scaling seems to be the most promising (Eder, Rybicki and Kestemont, 2015; cf. Embleton, Uritescu and Wheeler, 2009).

The expected results will show different groupings of texts, and some of them will be conditioned by genre or other extra-textual variables, not only by the location of composition. This paper contributes to the study of non-standardised historical texts and Middle English dialectology in particular.

Keywords: Middle English, historical dialectology, stylometry

References

Eder, Maciej, Jan Rybicki and Mike Kestemont. 2015. ‘Stylo’: a package for stylometric analyses. _Computational Stylistics Group._ [Online]. Available at <https://sites.google.com/site/computationalstylistics/>. Accessed Dec 10, 2015.

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: _Mélanges en l’honneur de Juhani Härmä_, (Mémoires de la Société Néophilologique de Helsinki,), ed. Eva Havu, Mervi Helkkula and Ulla Tuomarla. 101-108.

Jensen, Vibeke. 2010. Studies in the Medieval Dialect Materials of the West Riding of Yorkshire. PhD thesis, University of Stavanger.

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

MELD = The Middle English Local Documents Corpus, version 2015.1. September 2015, 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 10, 2015.