Classical Hebrew Stylometry: Initial challenges

Pierre Van Hecke (KU Leuven, Belgium)

Johan de Joode (KU Leuven, Belgium)

Our current research project aims at developing stylometric studies for Classical Hebrew and in particular the Hebrew Bible and the Dead Sea Scrolls. The present contribution describes the first research results of this project, highlighting both its promises and its challenges. Bible and Scrolls research have used computational methods extensively since the 1970s, but have often done so for information retrieval purposes only, and not for the computational study of textual similarity. Our project focuses on applying state-of-the-art stylometric methods to contribute to scholarly debates on the classification and provenance of texts in the aforementioned corpora (Burrows 1987; Koppel a.o. 2012; Daelemans 2013), given that these methods yield excellent results in other, also historical, corpora (Kestemont a.o. 2015; Stover & Kestemont 2016; Stover a.o. 2016).

The corpora of both Hebrew Bible and the Dead Sea Scrolls have been digitized and tagged lexically, morphologically and syntactically, and are available in a number of formats, both in open source and commercially. For the initial stages of our research, we have extracted the digitized text and the tagging from the databases used in the commercial Accordance® software. The reason for doing so is that this software package is the only one to have a consistent tagging of both corpora, which allows for an integrated and comparative study. Copyright issues make it difficult, however, to make public the enriched data and additional taggings that we are developing for the needs of our research. A major challenge recognized by many in the field, is the development of open source texts and consistent open source annotations of the Hebrew corpora. The BHSA database developed by the VU Amsterdam provides exactly this for the Hebrew Bible, a comparable open source database for the Dead Sea Scrolls still lacking.

The first stylometric analyses on both corpora have yielded promising results (Van Hecke 2018). The analysis of the datasets with the help of the Stylo library for R (Eder a.o. 2016) showed that both corpora (Hebrew Bible and [non-biblical] Dead Sea Scrolls) could adequately be differentiated stylometrically, at least for texts with a minimum length of 500 non-reconstructed word tokens. Also the clustering of artificially chunked portions of major Dead Sea compositions and the clustering of different manuscripts of single Dead Sea compositions was successfully executed. In our analyses, we selected both most frequent words (MFW) and most frequent character trigrams (MFC) as clustering feature types, calculating pair-wise distances between documents (with Burrow's Delta Measure [Burrows 2002; Argamon 2008]) on the basis of 500 most frequent features, visualizing our results both in dendrograms and in PCA graphs.

Successful though these preliminary analyses are, they leave many questions open for further research. Firstly, stylometric analysis detects stylistic similarity or distance between documents, but does not answer the questions by what this proximity or distance is caused. One possibility is that stylistic similarity between documents is caused by their common provenance or milieu. This insight can potentially have a great influence on classical scholarship in the field, which is highly interested in historical-critical questions on provenance. Many other factors may affect the observed stylistic similarity between documents, however. Three of these we have already analyzed in greater details (Van Hecke & de Joode, forthcoming). Firstly, we have studied the effect of textual fragmentation: since many of the Dead Sea Scrolls are only extant in heavily fragmented manuscripts, whereas the text of the Hebrew Bible is well established, we had to ascertain that our methodology was not unintentionally measuring the degree of textual fragmentation of the documents, by the way the features were selected. A second possible factor causing observed similarity between documents is orthography. Since Dead Sea Scrolls tend to have a different orthographic profile (frequently using consonants *y* and *w* to indicate vowels), we had to exclude that our stylometric analysis were not simply measuring differences in spelling. In order to do so, we harmonized the text orthographically by changing the orthography of lemma-morphology combinations with more than one orthographic variant to the most frequent one. Both factors do not seem to have a decisive effect on the clustering, so that the latter should be explained in a different way. Finally, we researched whether the feature selection (e.g. 300 most frequent character trigrams) affects our clustering. By building consensus trees of many analyses with slightly different feature selections, we could conclude that the observed stylistic differences between the Biblical and the Dead Sea Scrolls corpora does not depend our feature selection, and that a number of inner-corpus document clusterings are particularly strong and require more detailed analysis.

Four more factors possibly affecting document clustering need to be researched, before any conclusive arguments about document provenance or background on the basis of stylistic similarity can be made. Two of them are easy to study: document size and textual overlap. Since the corpus consists of documents with very different lengths, it should be analyzed to what extent these different lengths have an effect on our clustering, for example by favoring features occurring more frequently in longer documents. This challenge can be dealt with by using same length samples of the different documents. Secondly, some literary compositions are transmitted in several manuscripts with partial textual overlap, which will obviously cause documents displaying overlap to be clustered more closely together. Eliminating the overlap by ascribing the common text to one of the documents will solve this bias. Two other factors demand more complex research, namely genre effects and content bias. Our surveys have indicated that documents of similar genres tend to cluster together. It is therefore necessary to identify genre characteristics and to study their effect on document clustering, in order to determine which effects should be ascribed to genre, and which can be taken to be indicative of provenance or milieu. Finally, the stylometric methodology with its selection of *most frequent* features is relatively immune to content bias, by which documents with similar content tend to be clustered together. Nonetheless, the precise effect of similar content on document clustering should be determined.

References:

S. Argamon. 2008. Interpreting Burrow’s Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing* 23/2: 131-7.

J. F. Burrows. 1987. *Computation into Criticism : A Study of Jane Austen’s Novels and an Experiment in Method.* Oxford : Clarendon.

J. F. Burrows. 2002. ‘Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17/3: 267–87

W. Daelemans. 2013. Explanation in Computational Stylometry. Pages 451-62 in *Computational Linguistics and Intelligent Text Processing*. Edited by A. Gelbukh. Berlin/Heidelberg: Springer.

M. Eder, J. Rybicki and M. Kestemont. 2016. Stylometry with R: A Package for Computational Text Analysis. *The R Journal* 8/1: 107–21.

M. Kestemont, S. Moens, and J. Deploige. 2015. Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux. *Digital Scholarship in the Humanities* 30/2: 199-224.

M. Koppel, J. Schler, S. Argamon and Y. Winter. 2012. The ‘Fundamental Problem’ of Authorship Attribution. *English Studies* 93: 284–91.

J. A. Stover and M. Kestemont. 2016. The Authorship of the Historia Augusta: Two New Computational Studies. *Bulletin of the Institute of Classical Studies* 59/2: 140-157.

J. A. Stover et al.. 2016. Computational Authorship Verification Method Attributes a New Work to a Major 2nd Century African Author. *Journal of the Association for Information Science and Technology* 67/1: 239-242.

P. Van Hecke. 2018. Computational Stylometric Approach to the Dead Sea Scrolls. Towards a New Research Agenda. *Dead Sea Discoveries* 25: 62–87.

P. Van Hecke – J. de Joode. Forthcoming. Promises and Challenges in Designing Stylometric Analyses for Classical Hebrew. *Proceedings of the International Symposium on the Hebrew of the Dead Sea Scrolls and Ben Sira 2016*. Studies in the Texts of the Desert of Judah. Edited by Steve Fassberg. Leiden: Brill.