Abstract DH Benelux 2020

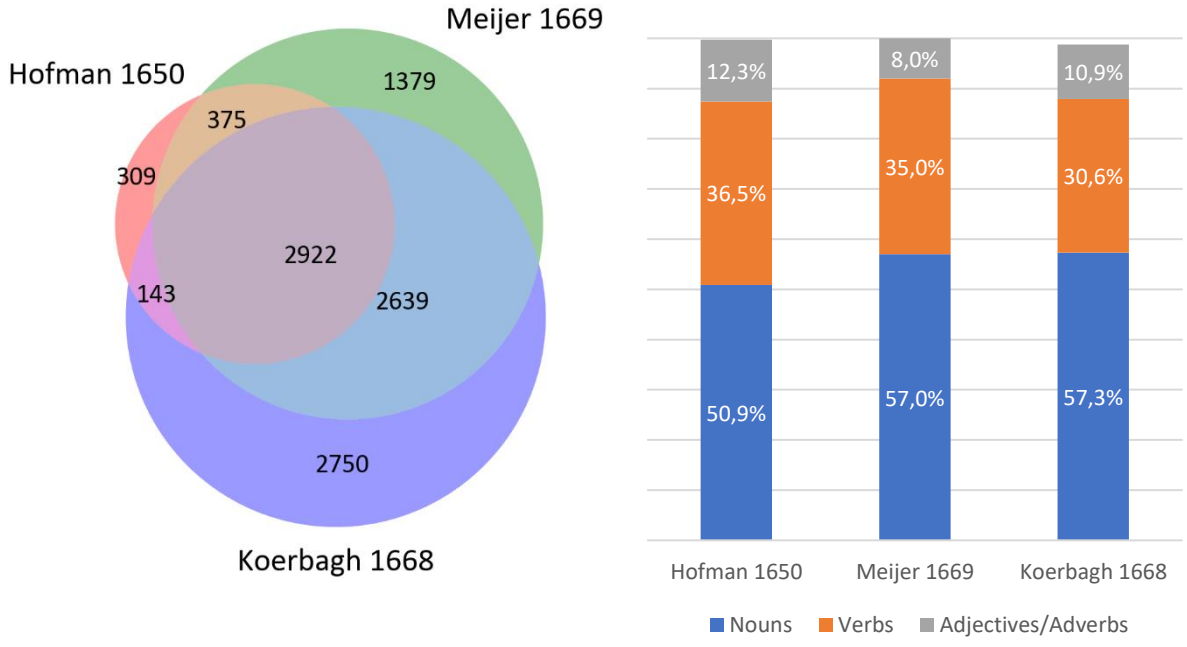
Name: Lucas van der Deijl
Institution: University of Amsterdam
Presentation category: short paper
Theme: 2. Replication, evaluation and quantitative analysis in the DH era

Automatic loan word extraction for early modern Dutch

Introduction
The Dutch language has never been exclusively 'Dutch': throughout its history, Dutch grammar and vocabulary have been enriched by elements from other languages. These foreign influences provide a meaningful feature of historical texts because they signal socio-linguistic norms among language users (e.g. Van der Sijs 2005). Contributing to the historical study of those norms, this paper presents a replicable lexicon-based method for automatic extraction of Latin and French loan words for early modern Dutch.

Lexicon
The lexicon consists of 10,457 distinct loan words obtained from three digitised seventeenth-century loan word dictionaries: Johan Hofman's *Nederlandtsche woorden-schat* (1650), the fifth edition of Lodewijk Meijer's *L. Meijers Woordenschat* (1669), and Adriaan Koerbagh's *Een bloemhof van allerley lieflijkheyd sonder verdriet* (1668). There is a strong similarity between these dictionaries in terms of overlapping lemmas and grammatical categories (see Figure 1; cf. Salverda de Grave 1906; Van Hardeveld-Kooi 2000). Because of their general agreement on the definition and categorisation of loan words, the selected dictionaries offer a valid representation of contemporary vocabularies considered to be 'foreign'.

Figure 1. Overlapping lemmas and grammatical categorisation in the three dictionaries



Method
The method involved two pre-processing steps to improve its recall: spelling normalisation and lemmatisation. Spelling normalisation was done using the spelling normalisation software VARD2 (Baron & Rayson 2008), which was trained on early modern Dutch by Wijckmans & Kisjes (2018). Lemmatisation relied upon FROG (Van den Bosch et al. 2007) and CLARIN's PICCL pipeline (Reynaert et al. 2015). These pre-processing steps increased the procedure's mean recall from 0.69 to 0.83. Recall was calculated based on manual annotation of 15,245 unique lemmas from 6 different texts. Note that precision is not a relevant evaluation metric since any word type will be classified as loan word if (and only if) it is included in the lexicon, thus leaving no false positives.

The formal procedure, documentation and lexicon are accessible through a non-programmer-friendly Jupyter Notebook and made available for further improvement and applications to other corpora.[1]

Results

This paper finally demonstrates the value of automatic loan word extraction through a case study of 33 texts by four seventeenth-century Dutch translators of Early Enlightenment philosophers René Descartes and Baruch Spinoza: Jan Hendrik Glazemaker, Pieter Balling, Stephan Blankaart, and Jacob Copper. Their loan word use was compared to a reference corpus of 207 manually transcribed texts from various genres published between 1650 and 1699 (downloaded from the DBNL). This reference corpus offered a norm to identify significant deviations from the average loan word use (M=1.59% of distinct lemmas, SD=0.81%) during the second half of the seventeenth century. The aim of this comparison was to better understand the factors explaining the variance in loan word use by translators who held explicit ideas about language philosophy, linguistic purism, and the status of the Dutch language. As automatic loan word extraction reveals high variances of loan word frequencies within the oeuvre of individual authors, this paper argues that loan word use depended on intellectual discourse and intended readership rather than individual lexical preferences. Moreover, the presented method tailored for historical Dutch complements similar tools developed for modern Dutch (Van der Sijs & Van der Meulen (forthcoming)[2]), thus enabling comparative analysis of foreign influences in trans-historical Dutch corpora.

References

Baron, A. & P. Rayson. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham.

Van den Bosch, A., Busser, G.J., Daelemans, W., and Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch, In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, Leuven, Belgium, pp. 99-114.

Hardeveld–Kooi, I. (2000). *Lodewijk Meijer (1629-1681) als lexicograaf*. Leiden: Dissertation Leiden University.

Leeuwenburgh, B. (2013). *Het noodlot van een ketter. Adriaan Koerbagh 1633-1669*. Nijmegen: Vantilt.

Reynaert, M., M. van Gompel, K. van der Sloot and A. van den Bosch (2015). PICCL: Philosophical Integrator of Computational and Corpus Libraries. *Proceedings of CLARIN Annual Conference 2015*, pp. 75-79. Wrocław, Poland.

Salverda de Grave, J.J. (1906), *De Franse woorden in het Nederlands*. Verhandelingen der Koninklijke akademie van wetenschappen te Amsterdam, afd. Letterkunde. Nieuwe reeks deel VII. Amsterdam.

Sijs, N. van der (2005). *Van Dale Groot Leenwoordenboek. De invloed van andere talen op het Nederlands*. Utrecht/ Antwerp: Van Dale.

Sijs, N. van der & M. van der Meulen. Tokenaanwezigheid van leenwoorden in Nederlandse kranten 1951-2002. Forthcoming.

Wijckmans, T. & I. Kisjes (2018), Adapting a Spelling Normalization Tool Designed for English to 17th Century Dutch. Paper presented at *DH 2018*.

[1] https://github.com/lucasvanderdeijl/automatic-loan-word-extraction

[2] https://github.com/INL/leenwoordenzoeker