Historical Corpus of Dutch

A new multi-genre corpus of Early and Late Modern Dutch

Abstract

This presentation has a two-fold aim. On the one hand, we intend to present the Historical Corpus of Dutch, discussing its structure and composition, but also reflecting on the use of citizen science in its compilation process. On the other hand, we will illustrate the usefulness of the corpus with a number of smaller case studies across linguistic domains, leading to a plea concerning data and corpus compilation for historical language studies.

The Historical Corpus of Dutch will be presented as a new multi-genre, diachronic corpus of Early and Late Modern Dutch (ca. 1550-1850). It consists of a digitized collection of printed administrative texts (e.g. town council meeting reports), handwritten ego-documents (e.g. diaries and travelogues), and printed pamphlets (e.g. of a political or religious nature). In addition, the corpus is also built up around different time periods, gathering data from around the middle of the sixteenth, seventeenth, eighteenth and nineteenth centuries. The corpus is also balanced between Northern and Southern material, with data from the provinces of Holland and Zeeland for the North, and from Flanders and Brabant for the South. All in all, the corpus consists of 209 different texts representing over 460,000 tokens, excluding tags and metadata. It has been manually annotated for named entities (personal and place names), and will be made available for research purposes in both TEI-compliant XML and plain text formats. As the corpus has in part been compiled using volunteers transcribing data in the citizen science project Wikiscripta Neerlandica II, we will also briefly touch upon the advantages and limitations of crowd sourcing data digitization for historical corpus building.

To illustrate the value of the new corpus, our presentation will then proceed to discuss a number of smaller case studies focused on linguistic changes spreading through time and space. These will range from orthographical and pronunciation-based variables to morphosyntactic variables, and serve to show how the available material can help us chart out changes in real time (e.g. visualizing S-curve patterns of language change), across regions (e.g. investigating North-South tensions from a pluricentricity perspective), and across genres (e.g. showing how ego-documents often display older or dialectal variants no longer found in more formal genres). Based on these exploratory experiences with the corpus, we will conclude by launching a plea for historical corpus building not to focus too much on ‘big’ data, but rather shift attention to ‘good’ data. Rather than bringing together large amounts of unstructured and poorly documented data with significant issues of representativeness, we will plea for the construction of smaller but balanced and well-structured corpora of different text genres, across the literacy/orality continuum, and enriched with sufficient metadata to allow for sociolinguistically informed analyses of language variation and change.