A parsed corpus of Southern Dutch dialects

Anne Breitbarth, Melissa Farasyn, Anne-Sophie Ghyselen & Jacques Van Keumeulen
Ghent University

This paper reports on the creation of an annotated corpus of spontaneous Southern Dutch dialect speech. It is based on a collection of 783 recordings (about 700 hours) from 617 locations in the Dutch-speaking provinces in Belgium, Zealand Flanders (Netherlands) and French Flanders (France). The tapes were recorded in the 1960s and 1970s, and contain the speech of dialect speakers born around the turn of the 20th century, the oldest informant being born in 1871. The dialect is hardly influenced by the standard language, as the speakers are almost exclusively monolingual and received only minimal formal education. The collection is therefore of immense value for linguistic research. Furthermore, it is a unique historical and cultural-historical resource, as it reports on topics such as both World Wars, the rise of electricity, bikes and cars and offers a unique perspective on lost professions and leisure activities. Although the recordings are already digitized and available online, they have not been transcribed or annotated systematically, which makes them hardly searchable for word forms or content, let alone for structures. The digital markup and exploration of this valuable treasure is an urgent desideratum considering the rapid dialect loss in Flanders, which means that soon there will be no one able any longer to transcribe the recordings.

The paper reports more specifically on an ongoing pilot project to transcribe and linguistically annotate about 40 recordings in the preparation of a larger infrastructure project to eventually make all the recordings accessible for fundamental research. We will first outline the transcription protocol developed specifically for this project, and then focus on the annotation pipeline, which starts with time-aligned transcriptions in ELAN (https://tla.mpi.nl/tools/tla-tools/elan/). The transcriptions consist of two layers, one closer to the dialect (cf. 1a, for instance capturing clitic constructions – marked with # – and ablaut phenomena), and one closer to Standard Dutch (2a), in order to make the data more searchable.

(1) a neen#t… k#en ik ewrocht met een ploef
b neen het... ik heb ik gewrocht met een ploeg
no it… I have I workedwith a plough
'No (that is not the case), I have worked with a plough.'

The time-alignment between audio and transcription facilitates (among others) phonetic research (as the transcription itself is not phonetic). In order to improve the quality of the transcriptions, we make use of crowdsourcing by setting up a network of dialect-speaking volunteers to check the transcriptions and to resolve ambiguities or doubts. After the transcription phase, the data is tokenized, lemmatized, PoS-tagged and parsed. We opt for an enrichment of ELAN-xml, as this allows maintaining the association with the time codes/the audio. The PoS-tags are awarded automatically and corrected manually. The tagset is still under development, but will connect as much as possible with existing tagsets for Dutch such as the CGN tagset in order to facilitate large-scale comparative research. The same considerations of interoperability guide the syntactic annotation, which follows the format of the Penn parsed corpora of historical English. The parsing is partly automated as well, using a pipeline of scripts for revision and shallow parsing, amongst others with CorpusSearch revision queries and the graphical user interface Annotald (https://annotald.github.io). Finally, it is the intention to combine audio, aligned transcriptions and annotations in a sustainable and searchable online corpus, made available via CLARIN in collaboration with the *Instituut voor Nederlandse Taal* (INT). At a later stage, the digital transcriptions can be subject to topic modeling, as such creating infrastructure for historical research.