**A Parsed Corpus of Southern Dutch Dialects**

*Anne Breitbarth, Anne-Sophie Ghyselen & Jacques Van Keymeulen*
Ghent University

At Ghent University, 783 tape recordings (c. 700h) of spontaneous dialect speech from all Dutch-speaking provinces in Belgium, Zealand Flanders (Netherlands) and French Flanders (France) are available. They were recorded in the 1960s and 1970s, and most of the speakers were born around the turn of the 20th century. This collection is of immense value for (at least) (i) linguistic, (ii) historical, and (iii) cultural-historical reasons. **Linguistically**, the Southern Dutch dialects are known to have a number of striking typological characteristics compared to other Germanic languages, which still await systematic description and analysis. Besides, the tapes contain accounts of **oral history** that provide a wealth of information on e.g. the events around the World Wars. Finally, the recordings constitute a treasure trove of **cultural heritage**, such as lost professions and customs. However, while the tapes have recently been digitized, and short summaries have been created for most of them (http://www.dialectloket.be/geluid/stemmen-uit-het-verleden/), they have not yet been transcribed or annotated to facilitate systematic research in the 50 years of their existence. The digital exploration of this treasure is an urgent desideratum considering the rapid dialect loss in Flanders (cf. Ghyselen & Van Keymeulen 2014), which means that soon there will be no one able any longer to transcribe the recordings.

In the current short paper, we report on the first stages of an ongoing FWO-funded pilot project to transcribe and linguistically annotate 32 strategically selected recordings in order to make this unique collection of dialect data accessible for fundamental research, and to prepare a larger infrastructure project with the same aims. After having discussed the relevance of the current project for integrated Digital Humanities and (cultural-)historical research, we will focus on the developed project pipeline. This pipeline starts with a transcription stage, departing from a newly-developed unified transcription protocol. Using ELAN (https://tla.mpi.nl/tools/tla-tools/elan/), time-aligned transcriptions are produced in two layers, one closer to the dialect, and one closer to Standard Dutch, (cf. 1), in order to make the data more searchable. The time-alignment between audio and transcription facilitates (among others) phonetic research (as the transcription itself is not phonetic).

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (1) | en k#weten | ik | da#m#inder | evlucht | en | naa | Frankrijk | eni. |
| | en ik weet | ik | dat we wij | gevlucht | hebben | naar | Frankrijk | eni. |

In a second step, the data are tokenized, PoS-tagged, and lemmatized. In this step, we opt for an enrichment of ELAN-xml, as this allows maintaining the association with the time codes/the audio. Third, the data are syntactically annotated using a pipeline of scripts and revision queries, as well as the program Annotald (https://annotald.github.io). The parsing follows the Penn Treebank format. The syntactic annotations are stored in standoff format. Finally, the intention is to combine audio, aligned transcriptions and annotations in a sustainable and searchable online corpus. At a later stage, the digital transcriptions produced in the current project can be subject to topic modeling, as such creating infrastructure for historical research.