What linguistic resources reveal about rusted bicycles and hoop making

In the 60s and 70s of the last century, a collection of 783 tape recordings (ca. 700h) of spontaneous dialect speech from all Dutch-speaking provinces in Belgium, Zeelandic Flanders (Netherlands) and French Flanders (France) was assembled. The speakers were all born around the turn of the 20th century (the oldest in 1871). The collection was meant to capture the unique linguistic features of the dialects before they would be lost. However, the recordings also form a unique window into the rural culture of the region around 100 years ago, as they form the biggest collection of life-stories of common, uneducated people living in the first half of the 20th century. They contain invaluable information about topics such as disappeared professions (e.g. rope and hoop making), the first and second World War, the introduction of electricity, bikes and cars and the typical coastal Flemish fishing expeditions to Iceland, the latter of which has barely been documented in writing.

The tapes have been digitised (www.dialectloket.be), but not yet digitally transcribed or linguistically annotated. With an eye on fast advancing dialect loss across Flanders (De Caluwe & Van Rentergem 2011; Ghyselen & Van Keymeulen 2014), it is an urgent desideratum that this wealth of data be transcribed, annotated and made available for linguistic and historical research, as there are now only very few people who are able to understand the recordings, let alone transcribe them. The current paper reports on an on-going project developing and testing a transcription and annotation standard for this important collection on a strategic sample of 40 recordings.

The recordings are transcribed in two tiers using the free ELAN tool (https://tla.mpi.nl/tools/tla-tools/elan/). One tier is closer to the dialect (cf. 1a, for instance capturing clitic clusters, marked with #), and one closer to Standard Dutch (2a), in order to facilitate further (semi-)automatic annotation, as well as searching through the data with minimal knowledge of the dialect.

(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	worked	with	a	plough	
		‘No (that is not the case), I have worked with a plough.’							

The transcribed data are then tokenized, lemmatized, PoS-tagged, and parsed. We opt for enrichment of ELAN-xml, as this allows maintaining the association with the time codes/the audio. The data are PoS-tagged using FROG (https://languagemachines.github.io/frog/) and corrected manually, in order to ensure interoperability with other corpora of spoken Dutch, esp. the CGN (https://ivdnt.org/downloads/tstc-corpus-gesproken-nederlands).

Given the great potential of the collection for research into cultural and oral history, the next step is the addition of keywords to facilitate searching the corpus for topics, e.g. all passages relating to hoop making. Using crowdsourcing, we have set up a network of dialect-speaking volunteers across Flanders, who have created content summaries of the tapes, accompanied by time indications per topic. These are used to add an extra time-aligned tier with these topics within ELAN-xml. The topics will be standardised and hierarchically classified using a controlled vocabulary.

Ultimately, 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).

References

De Caluwe, J. & E. Van Renterghem. 2011. Regiolectisering en de opkomst van tussentaal in Vlaanderen. *Taal en Tongval* 63: 61-77.

Ghyselen, A.S. & J. Van Keymeulen 2014. Dialectcompetentie en functionaliteit van het dialect in Vlaanderen anno 2013. *Tijdschrift voor Nederlandse Taal- en Letterkunde* 130, 117-139.