Anne Breitbarth
Ghent University

Parsing voices from the past:
The *Gesproken Corpus van de zuidelijk-Nederlandse Dialecten* (GCND)

The study of small languages and non-standard language varieties can contribute significantly to our understanding of the human language faculty. Additionally, dialect geography informs research on language contact and change over time, as it can help identify diffusion of innovations in space. The southern Dutch dialects — the dialects spoken in (i) Dutch-speaking Belgium, (ii) the three southern provinces of the Netherlands (Limburg, Noord-Brabant and Zeeland) and (iii) the northwestern part of Nord-Pas-de-Calais in France — are an invaluable testing ground, as they were affected by dialect levelling and loss only relatively late, and had for a long time little contact with the emerging Dutch standard language. This situation has meant that southern Dutch dialects have on the one hand preserved a number of archaic features, while in some cases also developing new ones.

While the focus of dialectology has traditionally been on the phonology, and to some extent the morphology of dialects, the study of syntactic variation in space has gained a lot of interest over the past twenty years. For the Dutch dialects, the monumental Syntactic Atlas of the Dutch Dialects (SAND) charts the syntactic variation in 267 places in the Netherlands, Flanders and France based on a written questionnaire and oral interviews. However, elicitation-based data collection tends to underreport phenomena that depend on specific discourse contexts, as is evident for instance from the SAND field notes, where fieldworkers report hearing speakers regularly use certain patterns they reject in direct elicitation. Nevertheless, spontaneous data as a source for dialectology have often been frowned upon (despite their regular use in neighbouring fields such as sociolinguistics) due to the difficulty of controlling the variables to be charted (sparsity). Clearly, if spontaneous speech data could be used, one could circumvent the problems posed by (in)direct elicitation, and if one had a large amount of such data, one could overcome the sparsity problem. Ghent University is in the fortunate position to host a large collection of recordings of spontaneous dialect speech: ca. 700 hours from 783 places in northern France, Belgium, and Zeeland in the Netherlands. These recordings were originally made in the 1960s and 1970s, with speakers born around 1900 (the oldest in 1871), before the introduction of general compulsory education in 1914. The speakers, often analphabets, were raised at a time when even bicycles were still rare, and hence mobility between places limited. This collection is therefore a historical corpus reflecting the state of the southern Dutch dialects before the influence of Standard Dutch through education, and before the onset of dialect levelling in Flanders from the 1960s onward. It furthermore preserves the already then threatened (and now moribund) Flemish dialects in northern France. The recordings were orginally made on reel-to-reel tape, but were digitised in 2014, and are now available online on `https://www.dialectloket.be/geluid/stemmen-uit-het-verleden/`.

In my presentation, I will report on the construction of the parsed Spoken Corpus of the southern Dutch dialects (Gesproken Corpus van de zuidelijk-Nederlandse Dialecten, GCND), currently (2020–2024) being built with funding by the Flemish Research Foundation FWO, which will finally make this wealth of data accessible (=searchable) for linguistic research by transcribing the recordings and annotating them with parts of speech and syntactic information. I will discuss the transcription protocol and the annotation, and demonstrate the advantages of this emerging resource for the study of so far overlooked or underreported syntactic phenomena of the southern Dutch dialects.