Towards a new lexicographical Infrastructure for the Dutch Dialects: the Database of Southern Dutch Dialects project.

De Tier, Veronique; Depuydt, Katrien; De Does, Jesse (Dutch Language Institute)
Chambers, Sally; Vandenberghe, Roxane; Hellebaut, Lien; Van Keymeulen, Jacques (Ghent University)

The Southern Dutch dialect area is described in four separate dictionaries, available both in print and online. The Brabantic, Limburgian and Flemish dialect dictionaries were set up as parallel onomasiological dictionaries whereas the Zeelandic dictionary was ordered alphabetically. The idea behind a common data model was to provide a comprehensive overview of the entire dialect area covered by these dictionaries. Through the Dictionary of the Southern Dutch Dialects (DSDD) project, the three dictionaries were brought together in one portal, realising the first phase of a new infrastructure for Dutch dialects. This paper will discuss: how these three conceptually similar dictionaries were brought together, what the challenges to harmonise these three dialects datasets were and how this enabled the integrated dataset to be made accessible both via a user application with cartographic tools, and an API.

The data model
Even though the dictionaries used the same data model it became clear that an overarching concept layer was required to deal with the problem of similar, but not necessarily equivalent concepts (e.g. the choice for either a concept “frog”, or two concepts “frog” and “green frog”) and the corresponding heteronomy (i.e. all dialect words for a single concept).

Data format
The source data was received in a range of different formats, e.g. database extracts from Oracle and FileMaker or as OCR (XML), and then stored in a relational database (PostgreSQL).

Data quality
Some of the original material had been OCR’d and semi-automatically/manually corrected, which had resulted in poorly structured data. It was therefore unfortunately necessary to leave out some of the data. However, it was made sure that the data ingestion method allows for future updates.

Data curation and enrichment
Before aligning the concepts some curation and enrichment was necessary to avoid inconsistencies, for instance, the differences: a) between the ‘Dutchification’ of keywords and lexical variants, b) in the assignment of lexical variants to keywords, c) in spelling, etc. had to be resolved.

DSDD concept layer and linking
In the pilot phase, 1500 concepts from a number of thematic dictionary volumes were selected to explore different methods for aligning the data. For each theme a list of overarching concepts was compiled. Lex’it, a rapid database application development platform for linguistic data, developed at the Dutch Language Institute

(INT), was used to do the linking. When the names of the concepts were identical, concepts were linked semi-automatically. However, when they were not, strategies such as keyword overlap, searching in concept definitions etc. were used for linking. Later on other concepts were integrated and linked. Now, the database consists of 29.000 concepts.

Future work

The dataset is now accessible via a user application with cartographic tools and an API. In 2022/2023, the database will be extended with additional semasiological Dutch dialect dictionary data, such as the *Zeelandic Dictionary,* until ultimately, the dialect data in the lexicographical infrastructure covers the entire Dutch language area.

References

De Vriend, F. & L. Boves, H. van den Heuvel, R. van Hout, J. Kruijsen & J. Swanenberg, Jos. (2006). A unified structure for dutch dialect dictionary data. in: *Proceedings of The fifth international conference on Language Resources and Evaluation* (LREC 2006).

De Vriend, F. (2012), *Tools for Computational Analyses of Dialect Geography Data*. PhD Radboud University Nijmegen.

Van Keymeulen, J. (2004), Trefwoorden en lexicale varianten in de grote regionale woordenboeken van het zuidelijke Nederlands (WBD, WLD, WVD). In: De Caluwe J, G. De Schutter, M. Devos en J. Van Keymeulen*, Taeldeman, man van de taal, schatbewaarder van de taal.* Vakgroep Nederlandse Taalkunde UGent – Academia Press, Gent (2004); 897-908.

Van Keymeulen, J., V. De Tier, R. Vandenberghe & S. Chambers (2019), The dictionary of the Southern Dutch Dialects (DSDD): designing a virtual research environment for digital lexicological research. in: *Dialectologia. Special issue*, 8 (2019), 93-115.

Van den Heuvel, H, E. Sanders en N. van der Sijs (2016), Curation of Dutch Dialect Dictionaries In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 16).

Van Hout, R, N. van der Sijs, E. Komen en H. van den Heuvel (2018), A fast and flexible web interface for dialect research in the Low Countries. In: *Proceedings of The fifth international conference on Language Resources and Evaluation* (LREC 18).