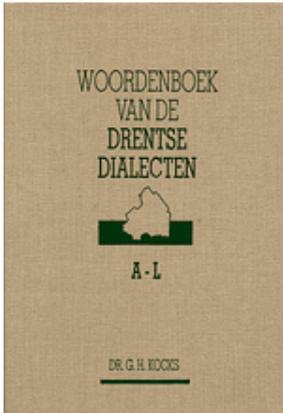


Improving a Dialect Database

Jelke Bloem and Erik Tjong Kim Sang

University of Groningen

Introduction



Woordenboek van de Drentse dialecten is an online version of a paper dialect dictionary for Drents developed between 1969 and 1996. The recent digital version was derived from the original material by hand-written rules which covered the most frequent structures in the paper dictionary.

Research problem

3.4% of the material in the paper dictionary was not included in the online version because it could not be categorized by the conversion rules and did not fit in the available database structure. In this study we improved the previously developed conversion rules and created a more elaborate database structure in order to allow for exceptional lemma structures.

Rule set adaption

Several adaptations were made to the conversion step from HTML to XML, for example:

- new trigger words, like *onpers*, were added
- references (*zie*) were allowed at more positions
- rules were applied repeatedly rather than once

The adaptations caused the coverage of the rules to go up from 96.6% of the lemmas to 97.6%.

Database structure modifications

An analysis of the online dictionary showed that most of the problems were caused by the database structure. Therefore several adaptations have been made here, for example:

- extra fields were added, like lemma numbers
- new combinations of data were allowed, like example sentences with sources
- field cardinalities were increased, like allowing nouns to have both *de* and *het* as article

The effect of the changes have not been evaluated.

Concluding remark

We have improved a digital dialect dictionary by expanding the conversion process and using a more elaborate database structure.

References

- Jelke Bloem, *Een digitale structuur voor het Woordenboek van de Drentse dialecten*. BSc thesis, University of Groningen, 2009 (in Dutch).
- Jelke Bloem, Lars Buitinck, Erik Joling, Gerwin Kramer en Harmen Kreulen, *De Dikke Kocks online*. Project report for the course Capita Selecta Natuurlijke Taalverwerking, University of Groningen, 2009 (in Dutch).

Woordenboek van de Drentse dialecten

[Home](#) [Uitgebreid zoeken](#) [Zoektips](#) [Suggesties](#) [Over](#)