# Curation of Dutch Regional Dictionaries

**Henk van den Heuvel**[1], **Eric Sanders**[1], **Nicoline van der Sijs**[2,3]

[1]CLS/CLST, Radboud University Nijmegen, the Netherlands
[2]Meertens Instituut, Amsterdam, the Netherlands
[3]NCI, Radboud University Nijmegen, the Netherlands
{h.vdheuvel, e.sanders, n.vandersijs}@let.ru.nl

**Abstract**

This paper describes the process of semi-automatically converting dictionaries from paper to structured text (database) and the integration of these into the CLARIN infrastructure in order to make the dictionaries accessible and retrievable for the research community. The case study at hand is that of the curation of 42 fascicles of the Dictionaries of the Brabantic and Limburgian dialects, and 6 fascicles of the Dictionary of dialects in Gelderland.

**Keywords:** data curation, dialect, dictionary

## 1. Introduction

Between 1967 and 2008 the Dictionaries of the Brabantic and Limburgian dialects (*Woordenboek van de Brabantse Dialecten*, WBD, and *Woordenboek van de Limburgse Dialecten*, WLD) have appeared in press. They consist of three parts, published in 69 fascicles. The first part concerns the agricultural terminology of the Southern Dutch dialects, the second the technical terminology (industries, trades), and the third the general lexicon. WBD and WLD are prepared at the Radboud University in Nijmegen and set up by the famous Dutch dialectologist A.A.Weijnen. At Ghent the Flemish counterpart (*Woordenboek van de Vlaamse Dialecten*, WVD) is compiled. As a follow-up, in 2002 the Dictionary of the Guelders Dialects (*Woordenboek van de Gelderse Dialecten*, WGD) was set up at Radboud University. It consists of three fascicles (House 2005, Man 2006, World 2008) for two regions: the Veluwe and the Rivierenland (river area). The dictionaries were elaborated with a database progam. All these dictionaries are onomasiologically organized, i.e. as an entry the (Standard Dutch) concept is given, followed by the various dialect forms, and the places where these forms are used. These places are given in codes consisting of a capital letter plus three digits, the so-called Kloeke codes, named after the dialectologist G. Kloeke who invented the system (Kruijsen and van der Sijs, 2010). Figure 1 shows the location of the provinces covered by the dictionaries. The dictionaries associated with the provinces are in red capitals.

Part III of the WBD and WLD dictionaries has from the start been elaborated with a database program. Within the NWO project D-Square the data have been made digitally available on a website, to which is added a cartographic tool (van den Heuvel et al., 2015; de Vriend et al., 2006). However, the first and second part of the dictionaries exist in print only. The 42 fascicles of these two parts, together 9706 pages, contain an enormous amount of dialect data that are on the verge of disappearing. These data are invaluable for scientific research into the Dutch dialects and for research on lexical semantics in general[1].

In the past, efforts have been made to digitize the data of the WBD and WLD. The printed fascicles have been scanned and the Omnipage program has been trained to read the phonetic representation of the dialect forms. The result was poor, since especially in the Limburgian parts an enormous amount of subtle phonetic differences were noted, which seriously thwarted the optical character recognition (OCR). The project has halfway been abandoned. It resulted in computer readable MS Word documents without any internal structure. In principle Dutch entries are in bold and dialect forms in italic, but unfortunately in a number of fascicles all typographical information was lost during the process.

In 2015 the CLARIN-NL program has granted a project called CARE: CurAtion and integration of REgional dictionaries. The goal of CARE is to semi-automatically convert the Word documents to structured text (database), and to combine the data with those of part III. As output format the Lexical Markup Framework LMF has been chosen as an accepted standard with CLARIN for lexical data. Within LMF a generic hierarchical data model (feature system definition, FSD) was set up, into which other dialect dictionaries could fit as well. The objectives of CARE were:

- Define a generic database structure for dialect dictionaries (LMF);
- Link the structure to *Woordenboek van de Vlaamse Dialecten* (WVD) and other regional dictionaries;
- Curate the *Woordenboek van de Brabantse dialecten* (WBD) and the *Woordenboek van de Limburgse Dialecten* (WLD) parts I and II;
- Update the curation of WBD and WLD Part III;
- Include the *Woordenboek van de Gelderse Dialecten* (WGD).

[1]Extensive information about the dictionaries (in Dutch) is available via http://dialect.ruhosting.nl


Figure 1: The locations of WBD, WGD and WLD shown in a map with the Dutch provinces.

## 2. Material and Method

### 2.1. Dictionary Layout

Figure 2 shows an example from the Brabantic dictionary. The dictionary is divided in Lemmas, each with a Bronnenlijst (list of sources) and Toelichting (explanations). Each Lemma is divided in Trefwoorden (key words), which in turn is divided in Dialectopgaven (dialect entries). Each Dialectopgave has one or more Kloekecodes (codes that indicate the city or village where this Dialectopgave is used). Trefwoorden, Dialectopgaven and Kloekecodes can also have explanations, but these are not in the example.
The layout of the dictionaries is in principle uniform:

- A Lemma is in all capitals.
- A Bronnenlijst (List of sources of a dialect entry) is between parentheses.
- A Toelichting (Comment) to a Lemma is in square brackets or curly brackets.
- A Trefwoord (Keyword) is in bold, followed by a colon.
- A Dialectopgave (Dialect form) is in italic.
- A Kloekecode is of the form capital letter + 3 digits, but with omission of repeating letters and leading zeros. They are separated by commas and the last one is followed by a semicolon.

### 2.2. Macro and Script

The typographical information (layout) is used as anchor points for the conversion of text to database. For dictionaries this type of conversion based on typography can be deployed generically, since all dictionaries use typography essentially in the same way, even if there are small individual differences (for instance, dialect word can be typeset in bold or bold italics). First, the Word document is converted to Unicode text by a macro that also adds tags (<I></I> and <B></B> respectively) around text that is *italic* or **bold**.
A python script was created that reads the text and parses the text on the basis of typographical information to extract the fields in the dictionary. The script repairs common OCR errors and regular deviations in typographical information, such as instances in which a closing tag is found in the middle of a word (paardentuig: instead of paardentuig:).
The script was developed in Python especially for this project, but with generalisability in mind. The script works hierarchically: it tries to detect the dictionary elements and treats them in top-down order (Lemma, Trefwoord, Dialectopgave, Kloekecode). On each level, first OCR errors and other errors are traced and repaired and in a second step separate elements are extracted and stored. Both error de-

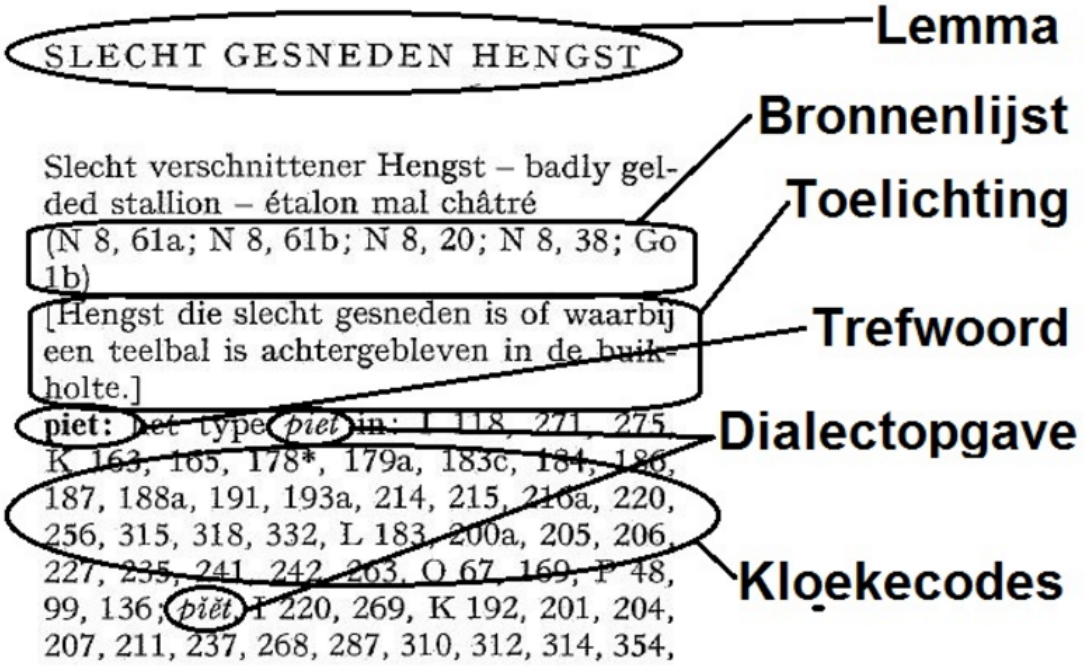


Figure 2: Example from the Brabantic dialect dictionary with explanation of how it is structured.

tection and element extraction are done by extensive pattern matching.
The output of the script is in CSV (comma separated value) format with a line for each Kloekecode. The output CSV file contains the following information in the consecutive columns:

- Lemma
- Lemma comment
- Source list
- Keyword
- Keyword comment
- Dialect form
- Comment Dialect form
- Kloeke code
- Place name
- Comment Kloeke code

This CSV output is the basis for further processing, such as the conversion into an LMF representation as in our case. (see section 2.4).

### 2.3. Manual Intervention

If the script detects patterns it does not recognise it will output an error message. Trained assistants inspect the output of the script. They repair the errors in an iterative process with running the script, until the script does not indicate errors anymore. They also check whether the script parses the text without errors (omitting elements or extracting elements in the wrong category). A problem the script cannot solve, is that there are irregularities in the dictionary text: at any point commentaries or explanations can be added in roman characters. These commentaries can refer to the Dutch entry, the dialect entry or the Kloeke code, and the commentaries can be placed before or after the element they refer to. In all these cases the script returns an error message. At first the assistants manually corrected the text where an error message occurred and then run the script again. This turned out to be a very elaborate and time-consuming procedure, and to speed up the work, it was decided to mark the commentaries before the script is applied. This was done by two volunteers, and led to a considerable saving of time. The volunteers marked a commentary before the element it is referring to with the symbols %% and for a commentary after the element with the symbols $$. The script was adapted so that it recognized these symbols and on encountering them, did not return an error message. Since the number of commentaries greatly varies in the various fascicles of the dictionaries, it is hard to calculate the time needed for preprocessing, but a trained volunteer could do this quite quickly. Running preprocessed text through the script appeared an easy task and the assistants managed to process an average of 110 pages in a days work.

### 2.4. LMF Data Representation

A second script converts a CSV file into a corresponding LMF file. LMF is an XML standard which is typically suited to capture hierarchical lexicon structures. Within the CLARIN project COAVA (Cornips et al., 2011) we developed a first LMF model for part III of the WBD and WLD. We then extended this model so that it could be used to fit in

other dialect dictionaries as well. Our LMF model is based on three head features associated with Lexical Entry, viz.

- Form
- Sense
- Location

Two further head features are Definition and Context (both positioned under Sense). Each individual feature is linked to an ISOcat data category (cf. (Windhouwer and Wright, 2013)). The model described in (van den Heuvel et al., 2015) was augmented to accommodate the commentaries at the various data categories referred to above. The model is of generic nature and is able to include a wide range of dialect dictionaries of Dutch and other Germanic languages. In the appendix is a full overview of the LMF model and its implementation for our dictionaries.
The script is built such that the columns in the CSV files may contain arbitrary information. Allocation of each column to the appropriate LMF feature is provided in the header of the script.

### 2.5. CMDI files

Each part of a dictionary obtained a metadata file. To this end we used a CMDI[2] metadata profile as developed for the COAVA project (Cornips et al., 2011) and slightly adapted it. For each part of a dictionary (I, II, II) the resulting CMDI profile is named *WND* and can be found at https://catalog.clarin.eu/ds/ComponentRegistry. It contains information about owners and collectors of the material, the field of research, time and space dimensions, the created LMF files and the associated PDF versions (books) of the dictionaries .

### 2.6. Including the WGD

Another regional dialect dictionary is the *Woordenboek van de Gelderse Dialecten* (WGD). It is briefly described in the full context of other regional dialect dictionaries by (van Keymeulen and de Tier, 2010)[3]. Previously (van den Heuvel et al., 2015) the part Rivierengebied (River area) was curated. In the CARE project the entries were checked and corrected once more and the collection was extended with the Veluwe area covering the topics House, Man and World. These were already available in digital form but had to be transformed into CSV files yet. The information contained in the subsequent columns in the CSV files is:

- Standardized spelling
- Place name
- Number of question list
- Number of question
- Question
- Kloeke code
- Comment
- Dialect form
- Lemma

The CSV files were converted into LMF using our second script.

## 3. Results and deliveries

For each part of the dictionaries a series of files was delivered as resulting output of the project.

1. The PDF files of the fascicles
2. The corresponding CSV files
3. The corresponding LMF files
4. The CMDI metadata file
5. The Curation Report

Table 1 shows the number of records per Part of each dialect dictionary

| | I | II | III |
|---|---|---|---|
| WBD | 314,001 | 110,305 | 1,245,314 |
| | (8 of 8) | (6 of 9) | (13 of 14) |
| WLD | 325,493 | 116,938 | 1,277,246 |
| | (9 of 13) | (10 of 12) | (14 of 14) |
| WGD | 175,098 | 70,578 | N/A |
| | (3 of 3) | (3 of 3) | – |

Table 1: Number of records per dictionary part. For WGD Part I corresponds to Rivierengebied and Part II to Veluwe. Between brackets are the number of curated fascicles.

## 4. Discussion and Conclusions

In this paper the CARE project was described in which dictionaries of Dutch dialects were curated. The method was intended to be reusable for future dialect dictionary curation projects. The method itself is general and can be used in dictionaries that have a similar (hierarchical) structure as the WBD, WGD and WLD. The implementation, however, is only partly reusable: The layout of the dictionaries is specific and the (OCR) errors differ from book to book. Even within this project the variety of book layouts and the typecasting of information was enormous. and the resulting script that generates the corresponding CSV files is therefore typically suited for those dictionaries but fails when it encounters other dictionary lay outs. However, we learned an important lesson which can be used in similar projects, namely that it is time-saving to preprocess the texts by manually marking all irregular commentaries. Also, the updated LMF model is quite generic and thus suited for a large variety of dialect dictionaries. Finally, the script which converts CSV files into LMF is very generic. By setting a couple of header switches it can deal with arbitrarily structured CSV files. In conclusion, the experience we gained from the CARE project has allowed us to consider new, more generic solutions to convert text to structured data.

[2] See http://www.clarin.eu/CMDI
[3] See also http://dialect.ruhosting.nl/wgd/index.htm

## 5. Acknowledgements

We greatly thank research assistant Aukje Borkent, student assistants Jorik van Engeland and Inge Otto, interns Maaike Borst and Eline Dimmendaal for their meticulous and diligent work on the manual text processing. We also thank volunteers Jantien Kettenes-Van den Bosch and Herman Wiltink for their work on the preprocessing of the text documents. Finally we thank Menzo Windhouwer for his advice and support with the LMF model, and Charlotte Giesbers and Hugo de Vos for cleaning, updating and manually correcting the entries of the WGD.

## 6. Bibliographical References

Cornips, L., Snijders, M. K., Snijders, M., Swanenberg, J., and de Vriend, F. (2011). Bridging the gap between first language acquisition and historical dialectology with the help of digital humanities. In *Proceedings Supporting Digital Humanities. Copenhagen, 17-18 November 2011*. http://www.meertens.knaw.nl/coavasite/wp-content/uploads/2011/11/Paper-SDH.pdf.

de Vriend, F., Boves, L., van den Heuvel, H., van Hout, R., Kruijsen, J., and Swanenberg, J. (2006). A unified structure for dutch dialect dictionary data. In *LREC 2006, Language Resources and Evaluation Conference, Genova, Italy 2006*, pages 1660–1665.

Kruijsen, J. and van der Sijs, N. (2010). Mapping dutch and flemish. In *Alfred Lameli and Roland Kehrein and Stefan Rabanus (eds), Language and Space: An International Handbook of Linguistic Variation: Language Mapping*, Handbooks of linguistics and communication science ; 30.2, pages 180–202. De Gruyter Mouton.

van den Heuvel, H., Oostdijk, N., Sanders, E., and de Lint, V. (2015). Data curations by the dutch data curation service. overview and future perspective. In *Odijk, J. (2015): Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands.*, pages 54–62. Linkoping Electronic Conference Proceedings, 116. http://www.ep.liu.se/ecp_article/index.en.aspx?issue=116;article=005.

van Keymeulen, J. and de Tier, V. (2010). Pilot project: a dictionary of the dutch dialects. In Anne Dykstra et al., editors, *Proceedings of the XIV Euralex International Congress*, pages 754–763. Fryske Akademy. https://biblio.ugent.be/publication/1008579/file/1008580.

Windhouwer, M. and Wright, S. E. (2013). Lmf and the data category registration: Principles and application. In *Gil Francopoulo (ed.), LMF Lexical Markup Framework. Chapter 3*. Wiley-ISTE.

## Annex: LMF model for regional dialect dictionaries

| LMF feature | WBD/WLD I / II | WBD/WLD III | WGD |
|---|---|---|---|
| **Form**<br>**keyword**<br>(= dutchification; mandatory)<br>**comment**<br>(=Comment at trefwoord) | trefwoord<br><br>Toelichting bij trefwoord | trefwoord<br><br>— | lemma<br><br>– |
| **FormRepresentation**<br>**aggregatedkeyword**<br>(= category or theme) | | | |
| **FormRepresentation**<br>**dialectform**<br>(=dialect form; mandatory)<br>**comment**<br>(=comment at dialect form) | dialectopgave<br><br>opmerking | fonetische variant | dialectwoord |
| **FormRepresentation**<br>**standardizedform**<br>(= respelling of dialect form) | — | — | standaardspelling |
| **FormRepresentation**<br>**lexvariant**<br>(= lexical variant) | — | Lexicale variant | |
| **FormRepresentation**<br>**phoneticform**<br>(= symbolic phonetic form) | | | |
| **Sense**<br>**lemma-id**<br>(= unique id) | lemma-id | lemma-nummer | record-ID |
| **Sense**<br>**lemma**<br>(= lemma; mandatory)<br>**comment**<br>(=comment at lemma) | lemmatitel<br><br>opmerking | lemmatitel | lemmatitel<br><br>opmerkingen |
| **Definition**<br>**definition**<br>(=definition or explanation of lemma) | toelichting op lemmatitel | vraagtekst | vraag |
| **Definition**<br>**sourcebook**<br>(=reference to source book) | bronnen | bronnen | – |
| **Definition**<br>**sourcebookpages**<br>(= pages in source book) | — | pag-bron | – |
| **Definition**<br>**sourcelist**<br>(= name question list) | – | vragenlijst | lijstnummer |
| **Definition**<br>**sourcelistquestion**<br>(= question number) | – | vraagnummer | vraagnummer |
| **Context**<br>**timecoverage**<br>(= time interval covered by source) | | | |
| **Context**<br>**publicationyear**<br>(= year of first publication) | | | |

| LMF feature | WBD/WLD I / II | WBD/WLD III | WGD |
|---|---|---|---|
| **Context**<br>**example**<br>(=example sentence) | | | |
| **Context**<br>**Comment**<br>(=general comment) | | commentaar | |
| **Location**<br>**kloeke**<br>(= Kloeke code for places)<br>**comment**<br>(Comment at Kloeke code) | Kloekecode<br><br>Toelichting Kloekecode | Kloeke-nieuw<br><br>— | Location Kloeke<br><br>— |
| **Location**<br>**place**<br>(= place name) | – | plaatsnaam | herkomst of bron |
| **Location**<br>**area** | – | gebiedscode | – |
| **Location**<br>**subarea** | – | subgebiedscode | – |
| **Location**<br>**informant-id**<br>(= informant code) | – | Informantencode | – |
| **Location**<br>**informant-birthyear** | | | |
| **Context**<br>**linktopublicationscan**<br>(=link to file - scan or PDF) | | | |