Curation Report

Woordenboek van de Brabantse Dialecten (WBD)

Part III

CLARIN-NL Data Curation Service

Version 1, 20 December 2013
Henk van den Heuvel
CLST, Radboud University Nijmegen

1. Introduction

The Dictionary of Brabant Dialects (WBD) covers together with the Dictionary of the Limburg Dialects (WLD) and the Dictionary of the Flemish dialects (WVD) by a same type of descriptive dialect lexicography the entire Southern Dutch-speaking region below the major rivers. This area stretches over three countries: the Netherlands, Belgium and France. To the study area include Flemish Brabant and Antwerp in Flanders and Brabant in the Netherlands for WBD, both Limburg and the northeast of Liege for WLD, and the east, west, French and Zeeland part of Flanders for WVD.

The WBD and WLD are compiled at the University of Nijmegen and the University of Leuven; the WVD is made by the University of Ghent.

This curation report deals with part III of the WBD. In Part III, the known vocabulary of the dialect-speaking community is dealt with that is not tied to the exercise of a profession. The third part of the WBD, the *General Vocabulary*, consists of four main sections, namely: "Man, as an individual," "The domestic life", "Community life" and "the world facing man."

The curation of Part III entails the material collected in the following books:

1.1 P. Vos
Het menselijk lichaam
2005, 431 blz. – ISBN 90 5179 223 9

1.2 P. Vos
Beweging en gezondheid
2005, 691 blz. - ISBN 90 5179 226 3

1.3 G. Coupé
Kleding en lichamelijke verzorging
2005, 469 blz. - ISBN 90 5179 235 2

1.4 M. Ooms
Karakter en gevoelens
2005, 795 blz. - ISBN 90 5179 207 7

2.1 lic. M. Ooms
De woning
2003. 640 blz. - ISBN 90 232 3916 4

2.2 M. Ooms, I. Blomme, J. Kruijsen, J. Swanenberg
Familie en seksualiteit
2005. 209 blz. - ISBN 90 5179 262 X

2.3 I. Blomme
Eten en drinken
2004, 453 blz. - ISBN 90 5179 206 9

3.1 lic. B. Decroos
Maatschappelijk gedrag, school en onderwijs
2002. 568 blz. - ISBN 90 232 3782 X

3.2 lic. G. Coupé
Feest en vermaak
2004. 434 blz. ISBN 90 232 4013 8

3.3 lic. G. Coupé
Kerk en geloof
2004. 478 blz. ISBN 90 232 4014 6

4.1 dr. J. Swanenberg
Fauna; vogels
2001. 272 blz. - ISBN 90 232 3755 2

4.2 dr. J. Swanenberg
Fauna; overige dieren
2001. 293 blz. - ISBN 90 232 3756 0

4.3 dr. J. Swanenberg
Flora
2002. 526 blz. - ISBN 90 232 3829 X

4.4 dr. J. Swanenberg
De stoffelijke en abstracte wereld
2005, 535 blz. – ISBN 90 5179 227 1

These books are delivered in PDF-format as part of the curation.

In this report we report upon the curation of Part III of the WBD. This dictionary was offered for

curation by prof dr Jos Swanenberg. The dictionary was available in the form of MySQL databases and are stored at the Meertens Instute as a result of CLARIN-NL's COAVA project (http://www.meertens.knaw.nl/coavasite/?page_id=8).

Each record contains the following information:

WBD/WLD	**English**
lemma_id	lemma-id
lemmatitel	Lemma title
vraagtekst	Text of question OR definition
trefwoord	Keyword
lexicale_variant	Lexical variant
opgave_genoveva	Dialectform in genoveva letter type
opgave_SIL	Dialectform in SIL phonetic form
bron	source
pag_of_vraagnummer_bron	page OR sourcelist number
plaatsnaam	place name
gebiedscode	code of region
subgebiedscode	code of subregion
volgnummer_informant	informant-id
toelichting	Comment
kloeke_nieuw	kloeke code (new version)

2. Data

The dictionary was provided in as text dump of SQL. The fields mentioned above were split and

converted into a CSV file (tab separated) in UTF8 encoding.

It appeared that the information in opgave_genoveva was corrupted since it was partly coded as Windows West-European Mac. Since the same information is in a more complete form available in opgave_SIL, it was decided to discard the opgave_genoveva from the curation.

opgave_SIL is coded as IPA Unicode, using partly hexadecimal character representations. The codes can be retrived in:
http://symbolcodes.tlt.psu.edu/bylanguage/ipavowels.html and
http://en.wikipedia.org/wiki/Phonetic_symbols_in_Unicode#From_IPA_to_Unicode)

3. Metadata

In the COAVA project a CMDI profile was developed by Folkert de Vriend for WBD and WLD. This profile was extended by the DCS to a more general profile for Dutch Dialect Dictionaries and published in the http://catalog.clarin.eu/ds/ComponentRegistry/# as WND (Woordenbank van de Nederlandse Dialecten). This profile was used to generate the CMDI metadatafile for this dictionary. We created one cmdi-file for this resource covering all fascicles for all LMF files and the corresponding PDF book files.

4. Restructuring the database

The MySQL files were converted into TAB separated files and these were used as starting point for converting the data into LMF format.

5. Converting formats

The TAB separated files were converted to an LMF format[1]. The LMF model for dialect dictionary data was developed by the DCS in close cooperation with Menzo Windhouwer. During this process dialectologists were consulted as to the proper inclusion and naming of lexical features in the model.

The model consists of three main classes for a Lexical Entry : Sense, Form, Location.
Location is a new class in the model.

Keyword (*trefwoord* in Dutch) is the only mandatory feature for a lexical entry in the model.

[1] LMF: Lexical Markup Framework: http://www.lexicalmarkupframework.org/

Next, the data of the dictionary were fitted into the model as shown in the table below.

WBD/WLD	**English**	**LMF**
lemma_id	lemma-id	Sense lemma-id=
lemmatitel	Lemma title	Sense Lemma=
vraagtekst	Text of question OR definition	Definition Definition=
trefwoord	Keyword	Form Keyword
lexicale_variant	Lexical variant	Form Representation Lexvariant=
opgave_genoveva	Dialectform in genoveva letter type	Form Representation Dialectform=
opgave_SIL	Dialectform in SIL phonetic form	Form Representation Phoneticform=
bron	source	Definition sourcelist=
pag_of_vraagnummer_bron	page OR sourcelist number	Definition sourcelistnumber=
plaatsnaam	place name	Location place=
gebiedscode	code of region	Location area=
subgebiedscode	code of subregion	Location subarea=
volgnummer_informant	informant-id=	Location

		informant-id=
toelichting	Comment=	Context Comment=
kloeke_nieuw	kloeke code (new version)	Location kloeke=

A corresponding LMF file was created including the LMF categories in the table above.

In order to validate the IPA Unicode characters in the Phonetic Form field the header of the LMF files ahd to be extended by:

```
<!DOCTYPE lmf:LexicalResource [
<!ENTITY % xhtml-lat1
   PUBLIC "-//W3C//ENTITIES Latin 1 for XHTML//EN"
   "xhtml-lat1.ent" >
<!ENTITY % xhtml-special
   PUBLIC "-//W3C//ENTITIES Special for XHTML//EN"
   "xhtml-special.ent" >
<!ENTITY % xhtml-symbol
   PUBLIC "-//W3C//ENTITIES Symbol for XHTML//EN"
   "xhtml-symbol.ent" >
   %xhtml-lat1;
         %xhtml-special;
         %xhtml-symbol;
]>
```

This invokes the required extra character sets needed to parse the phonetic symbols. These definitions can be found in:

http://www.w3.org/TR/xhtml-modularization/dtd_module_defs.html#a_dtd_xhtml_character_entities

The validity of the resulting XML files was tested using xmllint.

6. Documentation

Provided in this Curation Report.

Relevant information about the WBD and all its material can be found on the website http://dialect.ruhosting.nl/wbd/index.htm (in Dutch)

7. Persistent identifiers

Persistent identifiers were attributed by the CLARIN Data Centre (Meertens Institute).

8. Transfer data to CLARIN data centre

The curated dictionary consisting of the lmf file, this curation report and the cmdi metadata files are stored at the Meertens Institute as CLARIN data centre. Metadata harvesting and accessibility are taken care of of by Meertens .