Curation Report

Woordenboek van de Gelderse Dialecten Rivierengebied

CLARIN-NL Data Curation Service

Version 1, 31 May 2013
Henk van den Heuvel
CLST, Radboud University Nijmegen

1. Introduction

For the Dutch province Gelderland, various dialect dictionaries are available as part of the *Woordenboek van de Gelderse Dialecten* (Dictionaries of the dialects of Gelderland).

A website with more information about these dictionaries is: http://dialect.ruhosting.nl/wgd/ (in Dutch).

In this report we report upon the curation of the dictionary for the area *Rivierengebied* (River Area) only. This dictionary was collected by Dr Charlotte Giesbers in the period 2002-2008.

This dictionary consists of four parts: huis (Home), mens (man), wereld (world), woordenboeken en SGV (dictionaries and SGV[1] surveys)

Number of records for the individual parts:

1. Home: 27654
2. Man : 93716
3. World: 19230
4. dictionaries and SGV surveys: 34498

which totals to 175,098 records.

Each record contains the following information:

Field name	**English**
record_id	record ID
lemmatitel	Lemma title
tekst van de vraag	text of the question (from survey)
dialectwoord	dialectword in reponse to question
standaardspelling	standard orthography (=dialectword without diacritics)

[1] SGV refers to the dialectologists Schrijnen, Van Ginniken and Verbeeten

NL-woord	Dutch “translation” of dialectword (hardly used)
lijstnummer	List number (of survey)
vraagnummer	(Question number (of list)
kloeke	Kloeke code (A dutch code for dialect areas introduced by G.G. Kloeke, a Dutch dialectologist)
bron	Place (town) where dialect word was found
opmerkingen	Remarks

2. Data

The dictionary was provided in Filemaker Pro, version 5.
The tables were converted to version 11, and exported as CSV files (TAB separated), character codings in UTF-8.

3. Metadata

Parts of the Limburgs and Brabants dialect dictionaries (WLD and WBD) were digitized in the CLARIN-NL COAVA project[2]. In the COAVA project a CMDI profile was developed by Folkert de Vriend for WBD and WLD. This profile was extended by the to a more general profile for Dutch Dialect Dictionaries and published by the DCS in the http://catalog.clarin.eu/ds/ComponentRegistry/# as WND (Woordenboeken van de Nederlandse Dialecten).

4. Restructuring the database

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

[2] Refer to http://www.clarin.nl/page/about/projects/162#COAVA

5. Converting formats

The TAB separated files were converted to an LMF format[3]. 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.

The following classes and subclasses are defined and linked to ISOcat elements:

LMF feature	**Corresponding ISOcat element**
Sense lemma-id=	288 lemma identifier
Sense Lemma=	286 lemma
Sense Meaning=	464 sense
Form Keyword=	278 keyword
Form Representation Lexvariant=	5585 Lexical variant
Form Representation Morphologicalvariant=	5758 morphological variant (new, defined by DCS)
Form Representation Dialectform=	1851 geographical variant

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

Form Representation Phoneticform=	1837 phonetic form
Form Representation standardizedform=	1851 geographical variant
Definition Definition=	168 definition
Definition sourcelist= sourcebook=	5759 source list (new, defined by DCS) 471 source
Definition sourcelistnumber= sourcebookpage=	5760 souce list number (new, defined by DCS) 4126 pages
Context Timecoverage=	2502 time coverage OR 3664 Time coverage (Folkert)
Context Example=	3778 example
Location Place=	3759 source
Location Area=	3814 region
Location Subarea=	3814 region
Location informant-id=	3597 speaker id
location kloeke=	3651 Kloeke georeference

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

Furthermore, new ISOcat elements were introduced related to

Next, the data of the dictionary for the River Area were fitted into the model as shown below.

WGD / Rivier Area	**LMF**
record_id	Sense lemma-id=
lemmatitel	Sense Lemma=
tekst van de vraag	Definition Definition=
Lemmatitel (placeholder)	Form Keyword
NL-woord	Form Representation standardizedform=
standaardspelling	Form Representation Lexvariant=
dialectwoord	Form Representation Dialectform=
Vraagtekst	Definition Definition=
lijstnummer	Definition sourcelist= sourcebook=
vraagnummer	Definition sourcelistnumber= sourcebookpages=
bron	Location place=

	Location area=**Gelderland**
	Location subarea=**Rivierengebied**
opmerkingen	Context example=
kloeke	Location kloeke=

Kloeke codes were delivered as old Kloeke codes. They were converted to the new codes.

Since keywords were not provided for this dictionary, the position was filled by a placeholder (Lemma title). In case also Lemma Title was empty, the position was filled with the word 'EMPTY'.

The Dutch version of the keyword is contained in "Form Representation standardizedform=", but is hardly provided in the dataset.

Corresponding LMF files were created for the four domains provided:

1. Home
2. Man
3. World
4. dictionaries and SGV surveys

6. Documentation

Provided in this Curation Report.

Relevant information about the dictionaries and their design can be obtained via http://dialect.ruhosting.nl/wgd/inleiding_rivierengebied.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 four lmf files, this curation report and a cmdi metadata file are stored at the Meertens Institute as CLARIN data centre. Metadata harvesting and accessibility are taken care of of by Meertens .