Curation Report

Woordenboek van de Gelderse Dialecten Veluwe

CLARIN-NL Data Curation Service

Version 1, 7 June 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 *Veluwe* only. This dictionary was collected by Dr Harrie Scholtmeijer in the period 2005.

The dictionary consists of three parts: huis (Home), mens (man), wereld (world).

However for curation only the part Mens (“Man”) was available (75610 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)
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[1]. 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).

A corresponding metadatafile was populated for this part of the *Veluwe* area.

4. Restructuring the database

The TAB separated files 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[2]. 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.

[1] Refer to http://www.clarin.nl/page/about/projects/162#COAVA
[2] LMF: Lexical Markup Framework: http://www.lexicalmarkupframework.org/

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
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 source 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 source list and source list number.

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

WGD / Rivier Area	**LMF**
lemmatitel	Sense Lemma=
Lemmatitel (placeholder)	Form Keyword
Lemmatitel (as placeholder for NL-word)	Form Representation standardizedform=
standaardspelling	Form Representation Lexvariant=
dialectwoord	Form Representation Dialectform=
lijstnummer	Definition sourcelist=
vraagnummer	Definition sourcelistnumber=
herkomst	Location place=
	Location area=**Gelderland**
	Location subarea=**Veluwe**
opmerkingen	Context example=

Kloeke codes were not delivered.

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’. This was the case for 8813 records.

The Dutch version of the keyword (standardizedform) is not provided either. As placeholder Lemma was used again.

A corresponding LMF file was created for the domain provided (Man).

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 .