Curation Report

KEMPENSCH TAALEIGEN

BERGEIJKS DIALECTWOORDENBOEK

CLARIN-NL Data Curation Service

Version 1, 8 oktober 2013
Henk van den Heuvel
CLST, Radboud University Nijmegen

1. Introduction

There are various small local dialect dictionaries for the province of Noord-Brabant in the Netherlands. One of these dictionaries is: Panken, P.N. (1850) *Kempensch taaleigen.* Bergeijk: Johan Biemans [red. 2010]. This dictionary contains dialect entries for the village Bergeijk and surroundings in Noord-Brabant and is added to the curated version as a PDF file.

In this report we report upon the curation of this dictionary. This dictionary was offered for curation by prof dr Jos Swanenberg. The entries were manually enriched with a Dutch keyword, before they were provide to the DCS.

Each record contains the following information:

Field name	**English**
dialectwoord	dialectword
trefwoord	Dutch keyword
begrip	Sense
grammaticale informatie	Grammatical information
voorbeeldzinnen	Example sentences

Further information is known and added to the curated version:

Kloeke = K279p
Area = Noord-Brabant / Dutch Brabant
Place = Bergeijk
Sourcebook = Panken, P.N. (1850) *Kempensch taaleigen.* Bergeijk: Johan Biemans [red. 2010]

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.

3. Metadata

Parts of the Limburgian and Brabant 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 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.

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.

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.

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

Kempensch Taaleigen	**LMF**
trefwoord	Form Keyword=

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

dialectwoord	Form Representation Dialectform=
Begrip	Sense Meaning=
grammaticale informatie	Form Representation GrammaticalInfo=
voorbeeldzinnen	Context Example=
boek	Definition sourcebook=**Panken, P.N. (1850) *Kempensch taaleigen.* Bergeijk: Johan Biemans [red. 2010]**
bron	Location place=**Bergeijk**
	Location area=Noord-Brabant / Dutch Brabant
kloeke	Location kloeke=**K279p**

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

6. Documentation

Provided in this Curation Report.

Relevant information about the dictionary and its design can be found in the book:Panken, P.N. (1850) *Kempensch taaleigen.* Bergeijk: Johan Biemans [red. 2010] (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, the dictionary/book as PDF file, 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 .