Software Demonstration of the Dictionary of the Flemish Dialects and the Pilot Project Dictionary of the Dutch Dialects

Veronique De Tier and Jacques Van Keymeulen

Ghent University (Belgium)

A. Dictionary of the Flemish Dialects

The relational database of the Dictionary of Flemish Dialects works under Oracle. The WVD input database (bronsoorten = sources) consists of subdatabases of one or more questionnaires. Once all the data have been put into the correct subdatabases, the lexicographer can start compiling the dictionary by selecting the concepts that are to be put in one particular fascicle, e.g. all the selected concepts for 'sheep'. This is done in a new database structure, called 'publications'. After selecting the data from the different sources, the lexicographer automatically can generate a dictionary article with all the results for one concept. From this point onwards, the dictionary article can be compiled and lexicological decisions have to be made. This database also forms the basis for drawing the word maps in MapInfo and for making a text file ('Wetenschappelijk Apparaat' (Scientific Database)) in which every entry with the different lexemes for that particular concept is followed by codes that indicate the location of the words.

1. Introduction

The Dictionary of the Flemish Dialects (= WVD) started at Ghent University in 1972 at a time when computers were not yet used for this kind of project. The first years of the computerless era were used to make questionnaires, and excerpt older questionnaires from other institutes and older (alphabetical) dialect dictionaries. All these lexemes were put on index cards, which were classified per concept.

The first fascicles were all made by hand and typewriter. All the maps were also drawn by hand. In the 80ies everything changed. In 1987 the first computer was acquired and the dictionaries, still made by hand and written on paper could now be typed by a secretary. The first digital text file in the series of the *Dictionary of the Flemish Dialects* was the fascicle that was published in 1990 (*De molenaar/the miller*), but the work itself was still done with the help of index cards and on paper.

A few years later the editorial board invested in computers and a database program for the input of the dialect data. From that time onwards dialect words were put into a database and a plotter was linked to the program to draw the word maps. The program was updated some years later, with the result that the mapping could be done by the computer itself (without using a plotter) with the program MapInfo, the cartographic tool still used today. The WVD computer program was not updated until about 2003. But thanks to a financial injection of the Ghent University, the editorial board was then able to invest in new software and hardware. Since 2004 a relational database working under Oracle has been in use. This software was developed by Infoservice Belgium (ISB) in Ghent.

2. The software of the Dialect of Flemish Dialects

In what follows we will try to explain the working of the software, used by the editorial board of the *Dictionary of Flemish Dialects*. As mentioned in the paper 'Towards the completion of the Dictionary of the Flemish Dialects[1]', this dictionary is based on questionnaires about one particular conceptual field, e.g. 'school', 'the pig', 'diseases' etc., filled in by volunteers.

---

[1] This paper is included in the Proceedings of the 14th Euralex International Congress 2010. For references to relevant literature please see this paper.

Depending on the subject, about 200 questionnaires are sent in. Most questionnaires have about 100 questions. When the questioning is finished, the lexicographer can start the input of all these questionnaires into the database.

The database of the *Dictionary of the Flemish Dialects* consists of smaller subdatabases. One of the main reasons to subdivide into different smaller databases is that the dictionary makes a distinction between younger (collected after 1950) and older (collected between 1880 and 1950) data, in the dictionary articles and on the maps. But selecting and separating the different sources is also interesting for other reasons, e.g. for drawing word maps per source.

2.1. Sources
The WVD input database (called *bronsoorten* = sources) thus consists of different subdatabases, such as

1) WVD-questionnaires (all new);
2) ZND-questionnaires (old: collected by the Catholic University of Louvain before 1950);
3) ZND-questionnaires (new: the two last questionnaires were collected after 1950);
4) DC-questionnaires (old: collected by the Meertens Institute before 1950);
5) DC-questionnaires (new: collected by the Meertens Institute after 1950);
6) RND-questionnaire (for the WVD-region the 140 sentences are collected before 1950);
7) Master theses (all new material);
8) Other monographic material;
9) ...

Not included in this database is material from published works such as published dialect dictionaries, books and articles with dialect words in it. The dialect words from these sources are added at a later stage of the dictionary making (see 2.4.).
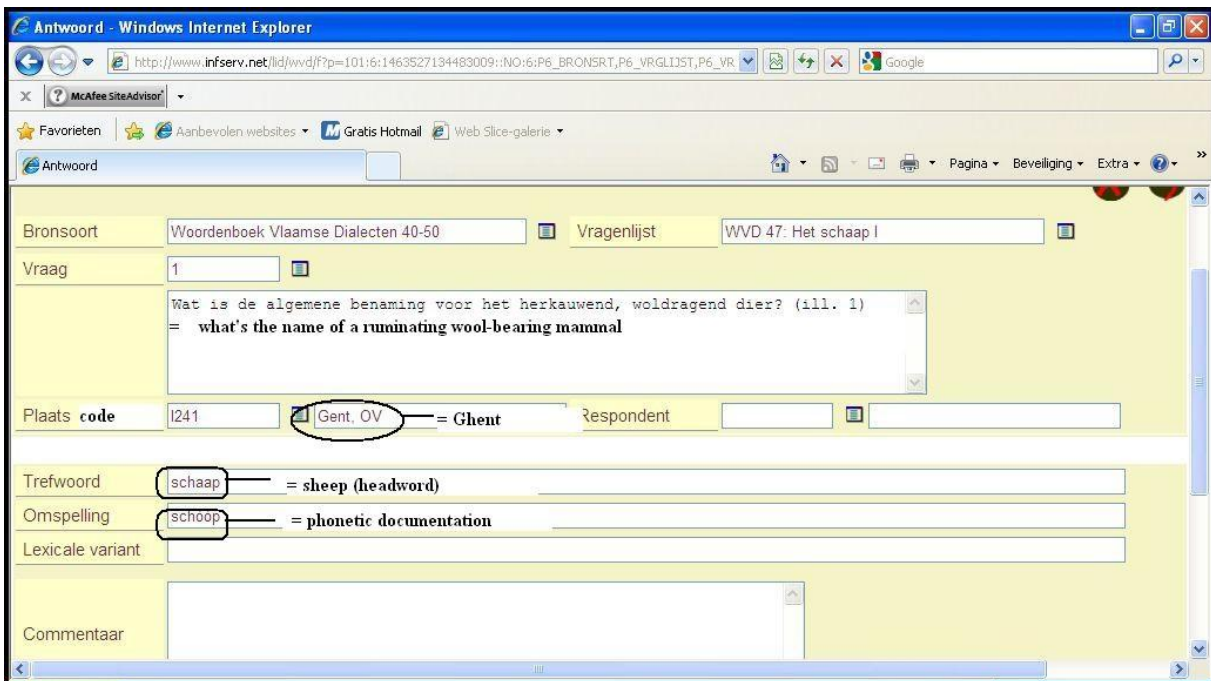
First of all, the editor has to choose the correct 'source' (= bronsoort). Each of these sources consists of one or more questionnaires (e.g. all WVD questionnaires numbered 40-50 = WVD 40-49). As an example we take sheep, the example that was given in the paper 'Towards the completion of the Dictionary of the Flemish Dialects'. The questionnaire sent by the editorial board of the *Dictionary of Flemish Dialects* about sheep is questionnaire 47. It consists of 132 questions. Before we start entering the data from this questionnaire, we have to put in all the names of the informants who answered the questionnaire. If this is done, the program can automatically give us the next locality of the same questionnaire, once we start giving the answers on for example question 1 (what's the name of a ruminating wool-bearing mammal) for all localities (see ill. 1). The editor puts the dialect word for sheep (*schaap* or *lemmen*) and the phoneographic realisation of this word for that particular locality in the correct fields, and the program moves to the next locality, until all the answers on question one are put into the database. The program then automatically goes on to question 2, and so on.

As for part I and II only headwords and phoneographic realisations are put into the database, but for part III also lexical variants can be added and the field phoneographic realisation can be left unfilled (see 'Towards the completion of the Dictionary of the Flemish Dialects 2.2.').
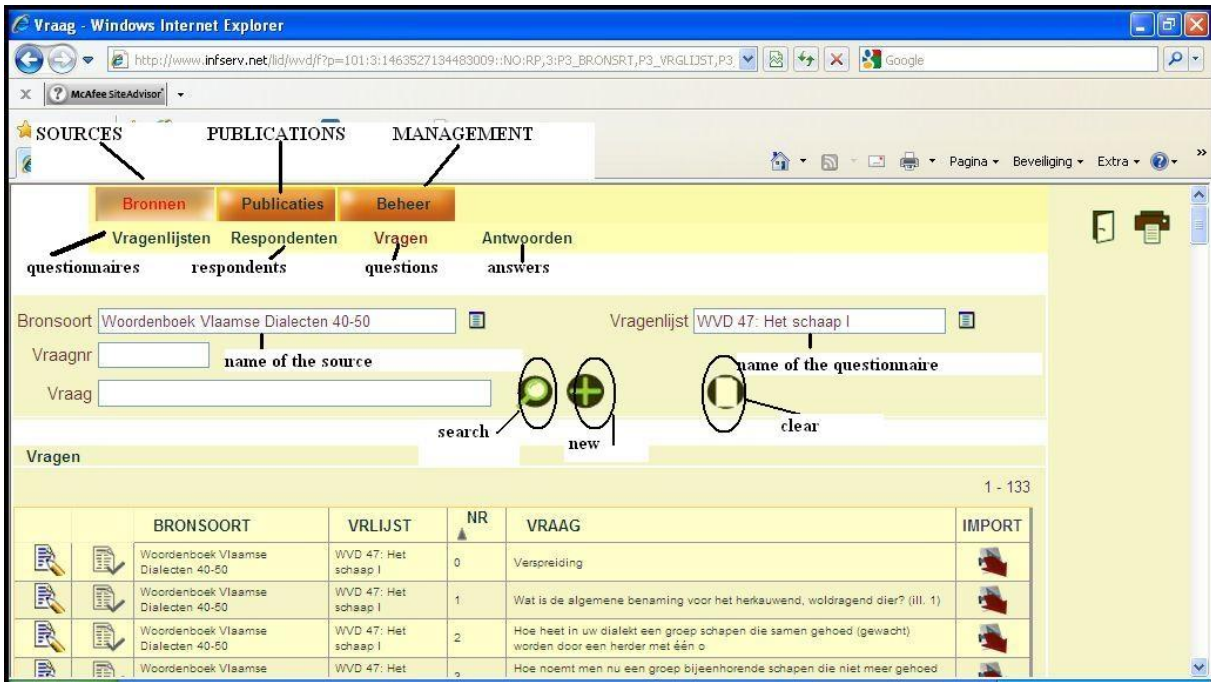
Ill. 1. Questionnaire 47 in the source Woordenboek Vlaamse Dialecten 40-50.

Ill. 2. Input of the data.

Veronique De Tier and Jacques Van Keymeulen

Ill. 3. Some answers on question 1 of questionnaire 47.

The questionnaire mentioned above (WVD 47), is one of the questionnaires that the editorial board distributed in the 80ies, but questions about sheep were found in older questionnaires of other institutes as well. So we choose another source (e.g. ZND) and also this subdatabase is filled with answers with names for sheep.

2.2. Publications

Once all the data have been put into the database (= answers to new and old questionnaires), the lexicographer can start compiling the dictionary (or a part of the dictionary) by selecting the concepts that are to be put in one particular fascicle, e.g. all the selected concepts for 'school', 'pig' or 'sheep'. This is done in a new database structure, called 'publications'. We create a new database, e.g. 'school', 'sheep' or 'food'. For each concept in a particular field (in the case of 'sheep' e.g.: sheep, ram, ewe, lamb, etc) questions with relevant data are selected from different sources/questionnaires. Sometimes only one question, sometimes a bulk of information from different master theses and different questionnaires are brought together for one dictionary article. For sheep we selected twelve different questions in ten sources. Seven of these questions were asked before 1950, five were asked after 1950. (see ill. 4)

SOURCES /BRONSOORT	QUESTIONNAIRE /VRGLIJST	QUESTION /VRAAG	Number of given answers
WVD 40-50	WVD 47: Het schaap I	1	94
WVD 151-160	WVD 161 : schaap en geit	1	69
ZND 1-10	ZND A2	22	1
ZND 1-10	ZND A2	313	50
ZND 1-10	ZND 6	25	30
ZND 31-40	ZND 38	40	461
DC voor 1950	DC 2	1	57
DC na 1950	DC 66	64	60
G V	Gemeenschappelijke Vragenlijst Meertens	3	69
Master theses E-T	Taquet	121a	10
Master theses U-Z	Van de Velde	1a	15
Willems	Willems	1	132

Ill. 4. The list of all the questions for sheep collected for the dictionary article sheep.

623

After selecting the data, the lexicographer can generate a dictionary article, which means that all the results (= words) for one concept are put together. From this point onwards, the dictionary article can be compiled and lexicological decisions have to be made. In WVD-speech, we call the collection of different dialect words denoting one concept a lemma.

Lemma-items

	PUBLICATION	NR	LEMMA	HEADWORD	NUMBER	DELETE	EXTRA INFORMATION	
	SCHAAP (127)	2	SCHAAP	Lemmen	14	✖	ℹ	
	SCHAAP (127)	2	SCHAAP	Schaap	1057	✖	ℹ	

1 - 2

Ill. 5. A dictionary article based on the selected questions is created in Publications
(14 x lemmen and 1057 X schaap).

As for the dictionary article 'sheep', there is no problem, because there are only two headwords, but mostly there are more headwords, sometimes even more than one hundred. Necessary alterations can be redirected to the original database (little typing or reading mistakes) by using the correction button. Some changes are not redirected to the original database, e.g. putting together diminutives and headwords, depending on the conventions made by the editorial board.
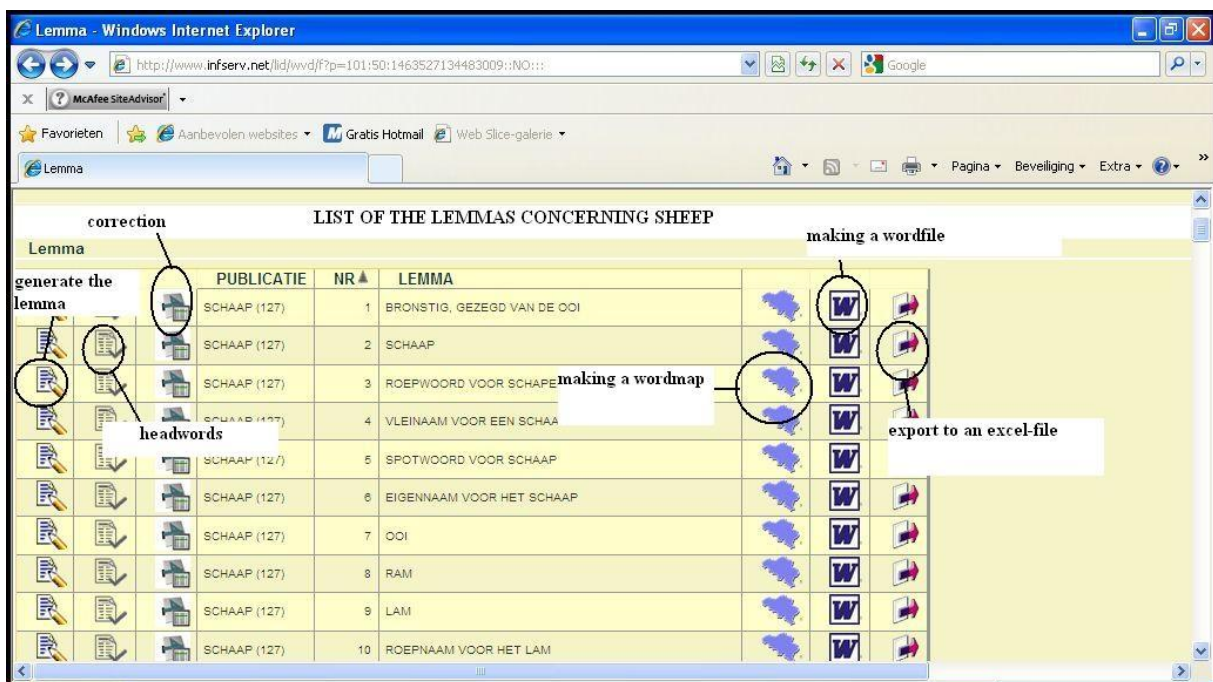
Ill. 6. List of dictionary articles concerning sheep.

This new database forms the basis for drawing the maps and making a text file. Therefore a Word macro has to be run to transfer all the results into a text file in word. This text we call the 'Wetenschappelijk Apparaat' (Scientific Database), a text in which every entry with the different lexemes for that particular concept is followed by codes that indicate the location of the words (see the example of sheep in 'Towards the completion of the Dictionary of the Flemish Dialects').

But because this 'scientific database' is not an easy text to read, the editorial board has chosen to publish a 'readable' text in which the code number row is replaced by generalized indications as to location, frequency and pronunciation. The codes are replaced by their locality names by running a macro in the Word file, but describing the regional dispersion of the words is still done by the editor himself. This is also the reason why we need maps of each dictionary article to see the geographical spreading of dialect words in a particular area .

2.3. Drawing maps

From each dictionary article, one or more maps are drawn (the data from each source can be put on a map separately: old material, new material, only ZND-material, etc.). In order to draw the maps, we use MapInfo, a cartographic tool. MapInfo automatically generates a map of the described region. Every word that occurs four times or more is put on the map, but an editor can decide to put some words together, or to show words that occur less than four times. On the maps, the lexicographer can see the different geographical patterns and describe where a certain dialect word is used (see as an example the word map of 'sheep' in 'Towards the completion of the Dictionary of the Flemish Dialects'). These maps are used to describe the geographical patterns. Interesting maps however are also published in the dictionary.

2.4. Adding published dialect words

Before publishing the dictionary article, the dialect words from published dialect dictionaries also have to be integrated in the text file. These words are excerpted from the different dialect dictionaries and put in an access database. The words are inserted manually in the text. The result is a dictionary article as you can see in the illustration of a page of the dictionary in 'Towards the completion of the Dictionary of Flemish Dialects'.

2.5. Published publications

Since 2009 the WVD-program has a third layer for 'published publications' in which all the 'titles' of the dictionary articles are put in the correct order (the same order as in the published fascicle), and in which the 'published material' is also added in the database itself. These publications will be made public later on and will replace the book version of the scientific database as it is published until now in part I an II. This procedure is now ready for most fascicles of the general vocabulary (part III) and will be put on the internet in the following months and years. Later on these databases can be integrated and combined with the databases of the two sister projects, and can then be the basis for drawing dialect maps of the whole region of the southern Dutch dialects.

3. Drawing maps of the three regional dictionaries

When the databases of the three regional dictionaries of the southern Dutch dialects will be connected, it will be possible to efficiently draw word maps for the whole southern Dutch area. The contents of the three dictionaries are not exactly the same, but thanks to the conceptual ordering, it is possible to combine most results of WVD, WBD and WLD, and even of the alphabetical WZD. It is too easy to say that it all happens 'by pushing one button', but it is relatively easy to make such word maps of the southern region.

At this moment, it is already possible to combine the databases of the three dictionaries, because a part of the databases of the two sister projects, is published on a website (www.ru.nl/dialect/wbd en www.ru.nl/dialect/wvd). These databases can be used in the database of the *Dictionary of Flemish Dialects* to draw maps. The procedure is to save the databases in an excel file as a csv-file. After importing this file in the WVD-database, it is

possible to combine the results with the dialect data of the *Dictionary of the Flemish Dialects*, and to make a combined dictionary article with the results of the three dictionaries. In the cartographic tool, only a different basic map has to be chosen to draw the correct map. Of course, this process will only result in a rough word map, with a lot of symbols on the map. It will probably look very chaotic. It is the task of the lexicographer to find and show some relationships by choosing corresponding symbols or colours. In the end, a readable word map for the whole southern Dutch area will be the result (see the word map 'daisy' in the paper 'Towards the completion of the Dictionary of the Flemish Dialects'). By combining the three databases the digital production of thousands of word maps will be possible. The immediate aim of the databases is the production of a word atlas of the southern Dutch dialects (for examples see: http://users.ugent.be/~jvkeymeu/cyberlemmata/).

B. Pilot project : the *Dictionary of the Dutch dialects*

The software of the Dictionary of Dutch Dialects works under the oracle platform as well. After digitizing dialect dictionaries by scanning and ocr-ing and after correcting the Word files of these dictionaries, the headwords are put into bold and two Hard Returns are inserted after each dictionary article. This Word document, converted into a standard XML file, can be imported into the database through the application built for this database. For each dictionary it is necessary to write a new custom script, which generates the XML-file by means of typographical conventions. Once the XML-files are uploaded, the database of the Dictionary of the Dutch Dialects can be made. The editors then may enrich the database with dutchification, translation and markers. The next step is to connect this database to a website with search facilities.

The software of the *Dictionary of Dutch Dialects* works under the oracle platform as well. The database of this pilot project is different from the database mentioned above, because in this project no questionnaires are used, but existing dialect dictionaries, made by amateur writers. As you can read in *Pilot project: The Dictionary of the Dutch Dialects*, the first step is to digitize dialect dictionaries by scanning and ocr-ing. The Word files of these dictionaries are corrected by volunteers. Once the Word file is corrected, and once the headword is put into bold and two Hard Returns are inserted after each dictionary article, this file is the input for this database.

---

*VETLAMPKE. o.
1. Aarden potje, met vet gevuld en voorzien van een wiek.
2. Gevuld met water en daarop een goede laag olie. Op die olie een kruisje in sterk papier, met in het midden een brandend wiekje. Destijds veel gebruikt voor vieringen.
Syn. LAMPIOENTJE, LANTJE.

VETLAP.
1. Iemand, die onzedige feiten bedrijft; die vuil en smerig is.
Syn. VETPOT.
2. Kleedingstuk, dat men gestadig draagt, waarmede men vuil werk verricht; waarop vette vlekken zijn : *ne — van nen hoed.*

*VETLEGGEN. ww. Alleen in den infinitief gebruikt. Hoornvee op de weiden vetten.

*VETLEGGER. m. Wie hoornvee vet op weiden (vetweiden) :
(1714) dat hij is gaen woonen ten huijse van de ouders van sijne huijsvrauwe vetlegghers. Crim. proc.

Ill. 7. Lievevrouw-Coopman, Gents woordenboek: Word file corrected by volunteers as input for the database.

The Word document is converted into a standard XML file by means of a script. The XML-file can be imported into the database through the application built for this database. For each dictionary it is necessary to write a new custom script, which generates the standard XML-file for the application. These scripts – as already mentioned – will generate the XML-file by means of typographical conventions. Examples already mentioned are the two hard returns, and the bold headwords.

Once the XML-files are uploaded, the database of the *Dictionary of the Dutch Dialects* can be made. Each bold headword followed by some information and ending with two hard returns will be transferred into the database of the *Dictionary of Dutch Dialects*.
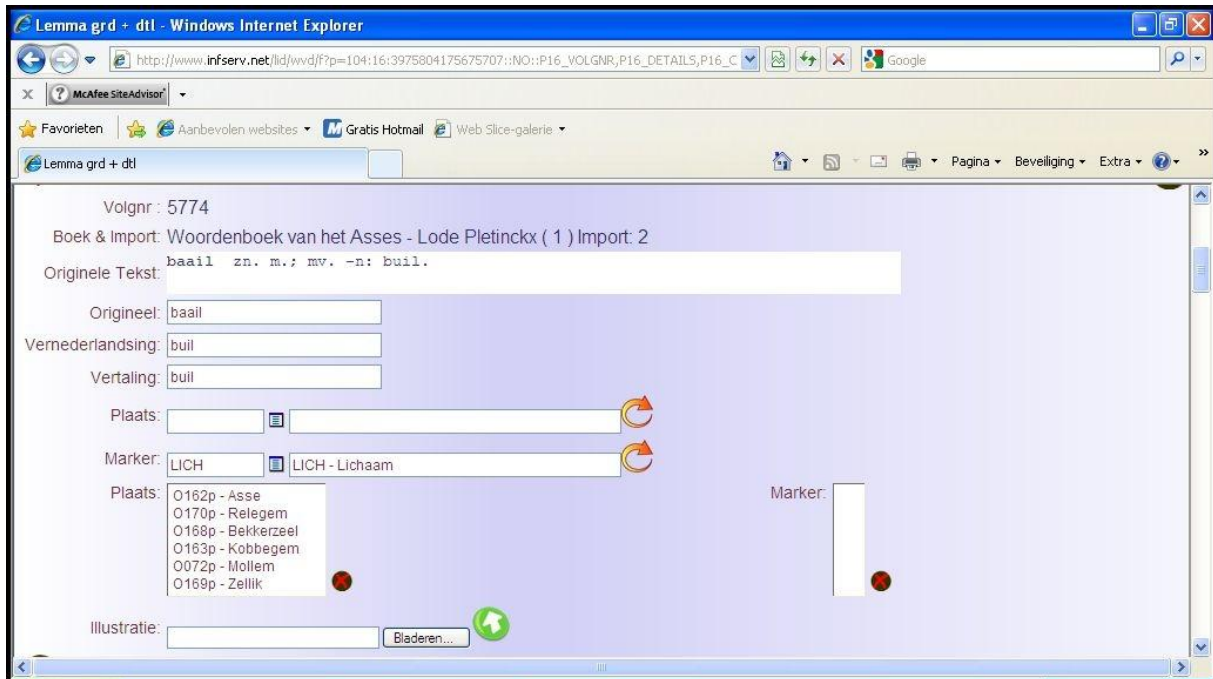
Ill. 8. Example of a lemma in the database of the Dictionary of Dutch Dialects.

The editors then may enrich the database with dutchification, translation and markers (see the example in ill. 8). Sometimes it is possible to add some of these enrichments automatically, but mostly the editor has to do this part manually.

The next step is to connect this database to a website with search facilities. So, if the database is filled with all the dialect dictionaries it will be possible to look e.g. for all the words for butterfly that are collected in all these dialect dictionaries. Or it will be possible to ask for a dialect word and find out where this word is used in the Dutch speaking area.