A Unified Structure for Dutch Dialect Dictionary Data

Folkert de Vriend[1], Lou Boves[1,2], Henk van den Heuvel[1], Roeland van Hout[2], Joep Kruijsen[2], Jos Swanenberg[2]

[1] Centre for Language and Speech Technology (CLST)
[2] Center for Language Studies (CLS)
Radboud University Nijmegen
Erasmusplein 1, Nijmegen, The Netherlands
{f.devriend, l.boves, h.vandenheuvel, r.v.hout, j.kruijsen, j.swanenberg}@let.ru.nl

Abstract

The traditional dialect vocabulary of the Netherlands and Flanders is recorded and researched in several Dutch and Belgian research institutes and universities. Most of these distributed dictionary creation and research projects collaborate in the "Permanent Overlegorgaan Regionale Woordenboeken" (ReWo). In the project "digital databases and digital tools for WBD and WLD" (D-square) the dialect data published by two of these dictionary projects (*Woordenboek van de Brabantse Dialecten* and *Woordenboek van de Limburgse Dialecten*) is being digitised. One of the additional goals of the D-square project is the development of an infrastructure for electronic access to all dialect dictionaries collaborating in the ReWo. In this paper we will firstly reconsider the nature of the core data types - form, sense and location - present in the different dialect dictionaries and the ways these data types are further classified. Next we will focus on the problems encountered when trying to unify this dictionary data and their classifications and suggest solutions. Finally we will look at several implementation issues regarding a specific encoding for the dictionaries.

1. Introduction

The traditional dialect vocabulary of the Netherlands and Flanders is recorded and researched in several Dutch and Belgian research institutes and universities. Most of these distributed dictionary projects, which are in different phases of development and completion, collaborate in the "Permanent Overlegorgaan Regionale Woordenboeken" (ReWo). These are the *Woordenboek van de Brabantse Dialecten* (WBD), the *Woordenboek van de Drentse Dialecten* (WDD), the *Woordenboek van de Gelderse Dialecten* (WGD), the *Woordenboek van de Limburgse Dialecten* (WLD), the *Woordenboek van de Overijsselse Dialecten* (WOD), the *Woordenboek van de Vlaamse Dialecten* (WVD), the *Woordenboek der Zeeuwse Dialecten* (WZD), the *Stellingwarfs Woordeboek* (SW) and the *Woordenboek van de Achterhoekse en Liemerse Dialecten* (WALD). The focus of ReWo is on coordinating the efforts related to the digitisation of dialect data and the use of computer tools for interpreting data.

The dialect dictionaries WBD (for the provinces of Northern Brabant in the Netherlands and Antwerp and Flemish Brabant in Belgium) and WLD (for the provinces of Limburg, both in the Netherlands and Belgium) are the products of large conventional dialect geographic projects which were initiated around 1960. WBD has recently been completed with the publication of the last fascicle of the printed dictionary. WLD will be finished in 2007. In the final stage of these projects it became clear that there is a need for electronic access to the data in these dictionaries, for a large number of reasons. This gave rise to the project "digital databases and digital tools for WBD and WLD" ("D-square" in short).

The first goal of the project D-square is to digitise the enormous amounts of dialect data (about 3 million phonetic variants) as published in the many fascicles of the dictionaries WBD and WLD over the years. Some of the oldest fascicles had to be scanned, OCR'ed and then further converted to XML. Newer material could be converted to XML from Word, MacWrite and FileMaker formats. Access to the data will be provided through a web interface. Special attention will be given to cartography, since maps are widely used in interpreting dialect geographic data. The product of D-square is aimed at both the general public and linguists.

Additional goals of the D-square project include the development of user friendly tools for analyzing the large amounts of data and the development of an infrastructure for electronic access to all dialect dictionaries collaborating in the ReWo. The focus of this paper is on the latter goal of D-square. Eventually, this infrastructure will enable unified access to dialect geographic data for the complete Dutch language area through one interface and one set of research tools as if it were one homogeneous data collection.

In section 2 we reconsider the nature of the core data of the different dialect dictionaries and the ways this data is further classified. Section 3 will focus on how to unify this dictionary data and their classifications. Finally in section 4 we will look at several implementation issues.

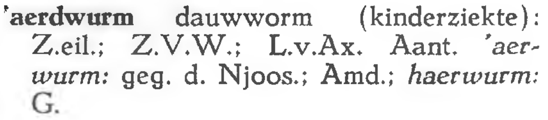
2. The Data Reconsidered

In order to realise a unified structure for the different Dutch dialect dictionaries, we did not take the printed dictionaries themselves as the starting point. Rather, we started from the questionnaires used to collect the data on which the dictionaries are based. The questionnaires more clearly show the essence of the data at hand. All dialect dictionary projects essentially go through the same process: documentation and classification of dialectal form variants that are used to talk about specific senses in specific locations (geographic coordinates). So the core data types they work with are *form*, *sense* and *location*. The most striking difference between the dictionary projects is the way they have organised their data for the purpose of publication. A choice for a certain organisation of the data was imposed upon the editors by the medium they had to use for presenting their findings in the twentieth century: i.e., the printed book. Books are one dimensional and linear and therefore the dialect dictionaries could not but present the data sequentially according to some ordering principle. In practice this meant that the editors had to choose one of the aforementioned types of core data - *form*, *sense* and *location* - as the primary type of core data as the most important organizing principle.

Traditionally, there have been two closely related fields of research that influenced dialect geography; lexicography and language geography (Kruijsen, 1996). The lexicographers used to take *form* as main core data type for presenting lexical data, ordered alphabetically. In the field of language geography it was common to use *sense* as main core data type, because it was felt that sense varied least in some geographic area.[1] Moreover, the questionnaires also tended to be organized on the basis of sense, rather than form or location. In a way, location has also been used as the primary criterion for making the data accessible, as testified by a number of dictionaries for a single city. However, the use of such local dictionaries is limited, and they incur a large amount of redundancy if they must be used in combination to cover a larger geographical area. We can see the two conventional approaches very clearly in the dictionary projects collaborating in the ReWo. Three of them follow the lexicographic form-based organisation of their data, while the other six follow the organisation most commonly used in language geography based upon senses. This raises the problem that uniform access and uniform research tools can only be provided if we can convert the data to a uniform internal structure.

2.1. Form-based organisation

The dictionaries WZD, WDD and SW have an alphabetical organisation based on forms. This kind of organisation has a very long tradition and is especially useful in situations where one encounters a certain dialectal form variant and wants to know its sense. Fig. 1 shows an example of a WZD entry:



'aerdwurm dauwworm (kinderziekte): Z.eil.; Z.V.W.; L.v.Ax. Aant. *'aer-wurm:* geg. d. Njoos.; Amd.; *haerwurm:* G.

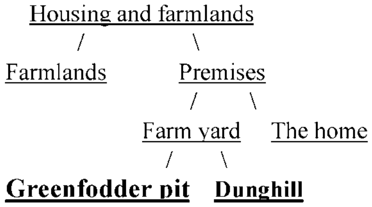
Figure 1: WZD entry

An entry can contain several related forms, as depicted in Fig. 1. In such a case one of the forms functions as reference form/headword. All forms are spelled in a phonetic alphabet and each form corresponds to a specific location. It is possible that one form refers to different senses in different locations.

An advantage of a form based dictionary over a sense based dictionary is the fact that it is completely based on observed data. It makes no tacit assumptions about the existence of specific forms in specific locations.

2.2. Sense-based organisation

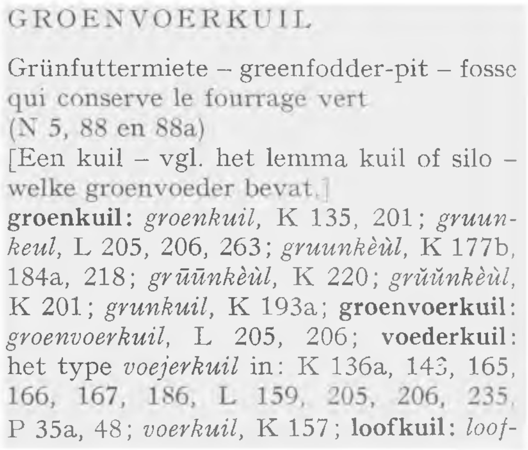
WBD and WLD are among the projects that based their data organisation on senses. Access to the data presented in the individual fascicles is provided firstly by traversing a taxonomy:[2]



Housing and farmlands
Farmlands
Premises
Farm yard
The home
Greenfodder pit
Dunghill

Figure 2: Partial taxonomy for the agricultural vocabulary

At the end leafs of the taxonomy the dictionary user ("reader") is presented with the entry belonging to a particular sense. These entries consist of a classification of the corresponding raw (i.e. "uninterpreted") dialectal forms. Part of the entry for *groenvoerkuil* ("greenfodder pit") in WBD is shown in Fig. 3.



GROENVOERKUIL

Grünfuttermiete – greenfodder-pit – fosse qui conserve le fourrage vert
(N 5, 88 en 88a)
[Een kuil – vgl. het lemma kuil of silo – welke groenvoeder bevat.]
groenkuil: *groenkuil*, K 135, 201; *gruun-keul*, L 205, 206, 263; *gruunkèùl*, K 177b, 184a, 218; *grūūnkèùl*, K 220; *grŭŭnkèùl*, K 201; *grunkuil*, K 193a; **groenvoerkuil**: *groenvoerkuil*, L 205, 206; **voederkuil**: het type *voejerkuil* in: K 136a, 143, 165, 166, 167, 186, L 159, 205, 206, 235, P 35a, 48; *voerkuil*, K 157; **loofkuil:** *loof-*

Figure 3: Partial WBD entry

The sense of the entry is in uppercase: g*roenvoerkuil.* The raw forms are in italics and in the phonetic alphabet Genoveva.[3] Each raw form matches with one or more locations where this form was recorded. The locations are specified with the geocoding system of Kloeke (Kloeke & Grootaers, 1934). All raw forms have been classified under so called heteronym categories. A heteronym is a synonym that has form variants that are geographically distinct (Weijnen 1961).[4] In Fig. 3 the heteronyms are in bold face.[5]

[1] Good examples are atlasses like Taalatlas van Noord - en Zuid-Nederland (G.G. Kloeke en L.J. Grootaers (1939-1972) Leiden) and Nouvel Atlas linguistique de la France (first fascicle published in 1957)

[2] The sense taxonomy used in WBD and WLD is based on *Begriffssystem als Grundlage für die Lexikographie: Versuch eines Ordnungsschemas* by R. Hallig and W. Von Wartburg (1952).

[3] Genoveva is a phonetic font especially designed for use in WBD and WLD.

[4] This is the definition as used in dialect geography. It differs somewhat from the more general defintion for heteronym.

[5] For the last fascicles of the WBD and WLD the editors have chosen to leave out the raw forms and in stead introduce a new intermediate level of classification in between raw forms and heteronyms: the lexical variant. Lexical variants group together

For WBD and WLD, there were several reasons for choosing the sense based organisation over the form based organisation which has a long tradition in lexicography.[6] The most important reason De Tollenaere and Weijnen (1963) give is that it makes it possible to present the form variation for the senses in dialect maps. These maps by their nature are linked to the senses. Entries with a substantial amount of form variation are accompanied with such a map.

A most practical reasons for an organisation based on senses is the self-contained nature of single fascicles covering a specific sense field instead of "all forms beginning with the letters a to g". This is of practical importance to both editors and buyers of the dictionaries.

2.3. Basic structure of the dialect data

It is interesting to note that all of the reasons given up till now for opting for one organisation and not the other are based not on fundamental differences in importance of one core data type over another, but purely on practical reasons. Different uses of the data are better catered for by one or the other organisation. But the nature of the data does not have an intrinsic "sense over form" or a "form over sense" hierarchy. The following UML class diagram illustrates how we model the relation between the core data types in a heterarchical manner:

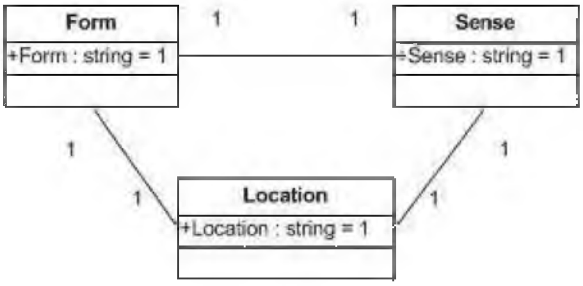


Figure 4: Core data types

A complete dictionary consists of a large number of such sets of three data points.

2.4. Higher order structure

Above these basic tripartite units different classifications can be created. The most natural way of classifying the senses is by using a taxonomy like the one depicted in Fig. 2. For the forms the natural way of classifying is to reduce raw phonetic variation to the standardised orthography of the meta-language used to describe the variation; in most cases this will be standard Dutch. Of course, also the locations can be further organised. The most natural organisation of the locations is a geopolitical taxonomy; villages are part of a municipality, a municipality is part of a province or region, etc. The diagram in Fig. 5 shows the relationship between the core data types and their classifications:

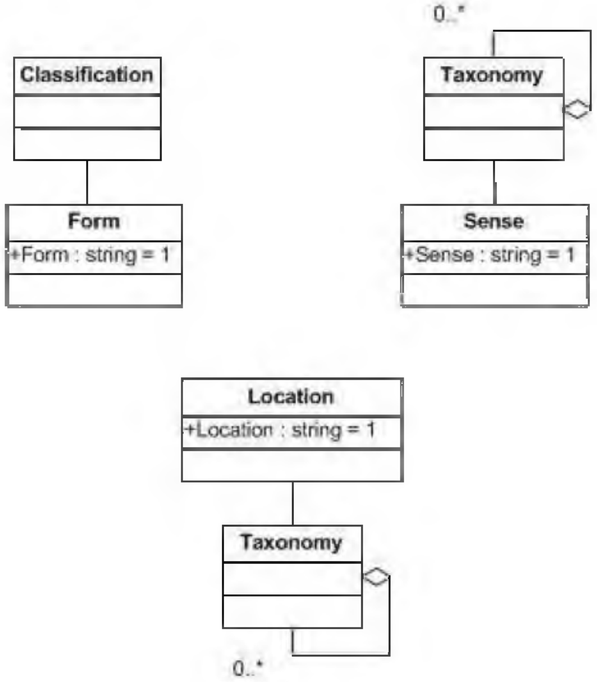


Figure 5: Core data types with classifications

2.5. Micro- and macrostructure

It is important to note that in the class diagram of Fig. 5 no one classification acts as main organisation for all core data types. This model also makes it possible to abandon the distinction between macro- en microstructure as it has traditionally been used for WBD and WLD. The macrostructure was seen as the basis on which to organise the data. WBD and WLD were sense based. The microstructure concerned the internal structure of the sense entries. In WBD and WLD the entry contained not only the form, sense en location relation depicted in Fig. 4, but also the classification of the form variation into higher level heteronyms (Kruijsen, 1996).

In our model we can redefine and simplify the distinction between micro- and macrostructure. The microstructure is reduced to the relation between the three core data types. The concept of macrostructure on the other hand is broadened. Every classification created above the basic tripartite units is a macrostructure in itself.

Adopting the model depicted in Fig. 5 has two advantages. First of all strictly separating the different classifications (macro structures) from the core data relations (microstructures) ensures optimal flexibility in working with the data. It will enable the user to choose the viewpoint most suitable to his needs. For instance, if he wants to know the form variation for the sense "plough", he will choose for the sense based view. If he wants to know what the sense distribution of the form "mus" (with the default meaning "sparrow") is, he will want to have the form based view on the data. And finally, if he wants to make a local dictionary (covering the town of Maastricht for instance) he will want to have a location based view on the data. The data can be used for more different purposes if it is possible to view them in multiple ways and from multiple perspectives. By offering this

raw variants that are distinct with regard to their consonant structure.

[6] WBD and WLD were the first comprehensive dictionaries in Europe that were based on sense.

possibility the dialect data are turned into a resource for eHumanities (Kircz, 2004).

The second advantage of this model becomes apparent when looking beyond the scope of any single dictionary. It also helps in realising the infrastructure that will enable unified access to the different dialect dictionaries in the ReWo.

3. A Unified Structure

By adopting the model presented in section 2 for every dictionary we can focus more clearly on where the fundamental inter dictionary differences exist. In this section we will try and analyse these differences more closely and go into more detail about strategies for merging the data and classifications belonging to different dictionaries.

3.1. Sense

Up till now we have used the term *sense* covering both the linguistic terms *concept* and *meaning.* From a language internal perspective forms can have *meanings.* From a language external perspective *concepts* can be referred to by forms that can be used in a language. A choice for one or the other perspective is directly related to choosing for a form based or a sense based methodology for data collection. In the form based dictionaries, what we have called sense so far, are more appropriately called 'meaning' from a linguistic point of view. The sense based dictionaries on the other hand are actually concept based. We propose to continue using sense as a linguistically and methodologically more neutral term covering both meaning and concept on the level of the core data.

The use of a published taxonomy does not solve all problems, because the concepts still must be referred to by words in a natural language. Two dictionaries can both use Dutch as language for describing senses in the taxonomy, but use slightly different wordings; one can use "hair of the dog", while the other uses "dog's hair". These are the well known and unavoidable problems of using a taxonomy in an environment that cannot enforce a fixed terminology. Also, one dictionary can use Dutch as language to describe its senses, while the other uses the dialect itself. The latter goes for WALD.

Another issue involves word forms that have no sense attached to them but only a grammatical function; the article "the" for instance. Such words have not been recorded for WBD and WLD because these dictionaries do not contain any closed word classes. Historically, the reason for this was that the need to record them was not felt as strongly as the need to record other word classes. The closed word classes are a rather stable core in every language. They are less subject to change than the open word classes, where words tend to disappear more easily. This was particularly relevant for the agricultural vocabulary that has been quickly disappearing since the start of the industrialisation. So, for WBD and WLD there are no word forms without a sense. But other dictionaries in the ReWo will introduce the problem of function words, if we want to collect all data in one unified environment. We can deal with this problem in several ways. We can be very strict and leave the sense field empty or extend the data model with a core data type "function". But we choose to be pragmatic here and follow the same strategy as most monolingual dictionaries do. That is to assign the grammatical function to the sense field and thereby make the notion of sense broader.

We propose to tackle the problems observed so far at the classification level; in a taxonomy. We could best deploy the WBD and WLD taxonomy (partly depicted in Fig. 2) as the basis for a sense classification that covers all dictionaries. These dictionaries have the most extensive sense taxonomy of the sense based dictionaries in the ReWo. The form based dictionaries have no sense taxonomy.

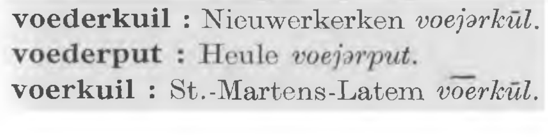
When deployed as the link between the senses used in the different dictionaries this sense taxonomy does not only describe hierarchical relations between the concepts themselves, but can also act as an interlingua between the senses used in the different dictionaries.

The senses of the other dictionaries need to be mapped onto this taxonomy. For form based dictionaries this means mapping the meanings to the more abstract concepts in the taxonomy. For the other sense based dictionaries the differences in wording and language of the senses can also be overcome by mapping those senses to standard concepts in the taxonomy. When dictionaries contain senses not already covered by the taxonomy, it should be possible to add new concepts to the taxonomy bottom up. The senseless forms have a grammatical function instead. This grammatical function will be mapped to a separate branch of the taxonomy that deals with a classification that adheres to that of the Dutch reference grammar *ANS.* (Haeseryn et al, 1997)

3.2. Form

In looking more closely at the differences in *form,* the first thing to note is that every dictionary uses its own kind of phonetic alphabet. For the dictionaries that have forms without a further classification into heteronyms, we have the problem how to relate them to forms in other dictionaries. All phonetic alphabets can be converted to IPA as a kind of objective reference, but this mapping is not necessarily trivial. Moreover, there is not yet a single standard for representing IPA symbols and diacritics in a computer readable and printable form.

On the level of the classifications of the forms there are a couple of typical problems when trying to unify them. The biggest problem with unifying the linguistically motivated heteronym classifications used in WBD, WLD and WVD can be illustrated by comparing the partial WBD entry in Fig. 3 with the following partial entry from WVD:



voederkuil : Nieuwerkerken *voejərkūl.*
voederput : Heule *voejərput.*
voerkuil : St.-Martens-Latem *vo͞erkūl.*

Figure 6: Partial WVD entry

Both entries show the variation for the sense *greenfodder pit.* However the form *voerkuil* has been classified under the heteronym "voederkuil" in WBD while the very similar form *vo͞erkūl* has been classified under a separate heteronym "voerkuil" in WVD. The problem is that the heteronym classification is based on a number of different linguistic criteria and that it is up to the intuition of the editor what criteria prevail (Van Keymeulen, 2004). This kind of inter-dictionary variation also exists between WBD and WLD.

Ideally, for a suitable unification consensus on the ways to classify forms into heteronyms should be reached. Because this is a very labour intensive undertaking, we suggest that users will get a choice between one of two possible automatically derived unifications. When comparing two dictionaries A and B, whenever a heteronym and a raw form are encountered that are identical, either the classification of dictionary A is adjusted to that of dictionary B, or the other way round. For the form variants *voerkuil* (WBD) and *vōērkūl* (WVD) this would result into one of the two classification mergers shown in tables 1 and 2:

	WVD forms	**WBD forms**
voederkuil	*voeɟərkūl*	*voejerkuil*
voerkuil	*vōērkūl*	*voerkuil*

Table 1: WBD classification adjusted to WVD classification

	WVD forms	**WBD forms**
voederkuil	*voeɟərkūl, vōērkūl*	*voejerkuil, voerkuil*

Table 2: WVD classification adjusted to WBD classification

Both strategies are lossy by nature. Either you loose the information that in WVD a variant had been classified as heteronym or you loose the information that in WBD this variant was classified as subordinate to a heteronym. We could let expert users choose one of the two mergers or no merger at all while presenting members of the general public with one kind of merger by default.

We are aware of the fact that the unification as presented does not automatically make it possible to draw methodologically sound conclusions from the dataset as a whole. However, since the data themselves cannot be altered anymore this is the best we can offer. When interpreting a unified data set one should always be kept aware of this.

3.3. Location

All dictionaries use either villages or cities as possible kinds of location. Some use the geocoding system of Kloeke, while others just use place names. In some dictionaries the place names have been abbreviated.

The dictionaries do not always cover mutually exclusive dialect areas. There are for instance locations that used to belong to WBD and later on became locations belonging to the dialect area of WLD. Just like one had to choose for either a form or a sense based organisation due to practical limitations imposed by the book medium, there have been practical reasons for deciding on the area any of the dictionaries would cover. The most important factors playing a role here were: who the funding organisation was, linguistic principles; isoglosses or the standard language of the area, or natural borders; the Nether Rhine acts as dialect border in WGD. None of these borders are strict natural dialect borders, however. By unifying all dictionaries again we see the advantage of being able to abstract away from enforced perspectives on the data: the original division into dialect areas. Ideally, users should be able to define the area in which they want to know the form variation. Information about the dialect area to which a variant was originally assigned should be of secondary importance.

Also place name ambiguity can be introduced when unifying the dictionaries. There might have been just one place *Berghem* in WBD, but when combining the data with other data sets all of a sudden three new *Berghems* might be introduced. There are two solutions for this. Either a geopolitical taxonomy covering all locations is introduced. Constructing such a taxonomy will not be very hard to do. Or all locations are converted to a geocoding system that can be used for uniquely encoding geographical locations world wide: longitude and latitude.

4. Implementation Issues

For implementation of our model and strategies for unifying the classifications and taxonomies we first need to decide what encoding to use. In our model the core dialect geography data is clearly data centred and heterarchical. For this type of data the relational data model is most appropriate. The taxonomies for the senses and locations have a natural and elaborate hierarchy; thus, the hierarchical data model of XML is most suitable (Wittenburg, 2004). The most suitable data model for the form classification is still under investigation.

4.1. Standardisation

For archival purposes and interoperability with projects outside the ReWo we also want to adhere to the Data Category Registry (Ide & Romary, 2004) and the Lexical Markup Framework (Francopoulo et al, 2006). The Lexical Markup Framework (LMF) is being developed in the ISO TC37/SC4 group and originated from the recognition of the troubles cross lexica search, merging, linking and comparison pose. The LMF core model is depicted in Fig. 7:

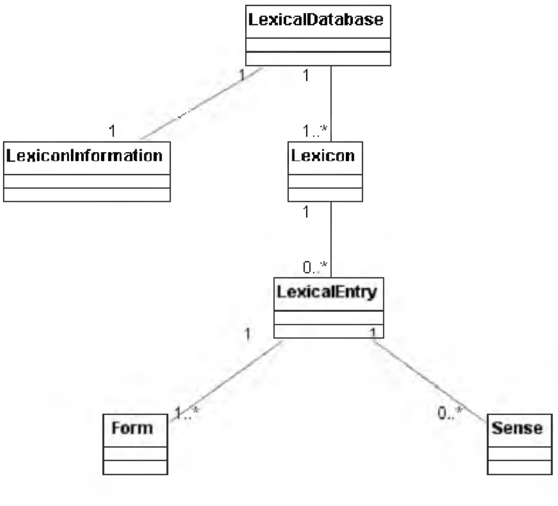


Figure 7: UML class diagram of LMF core model

The LMF core model has a *sense* and a *form* class but no *location* class. The LMF model being a flexible model, new components can be added to it. For our data we will attach a "location" extension to the LexicalEntry class. By doing so we do justice to the heterarchical nature of our data.

All core data will be imported into LMF with the use of the lexicon tool LEXUS (Kemps-Snijders, 2006).[7] Since LEXUS also provides support for the Data Category Registry (DCR) using predefined and accepted concepts and tag names is encouraged. However, also in the DCR there is no data category covering our data type location yet.

4.2. One interface for unified data

For the WBD data we have experimented with using Google Earth as cartographic interface, as shown in Fig. 8.[8] A unified structure for the Dutch dialect dictionaries as presented in this paper ultimately will make it possible to combine data from all dictionaries in such a location based interface.

Further functionality of the Google Earth interface is provided by the ability to combine data with overlays. This means the dialect data can be combined with all kinds of maps, for instance about historic geopolitical borders. Such combinations of different resources can shed new light upon the origin of patterns in dialect variation.

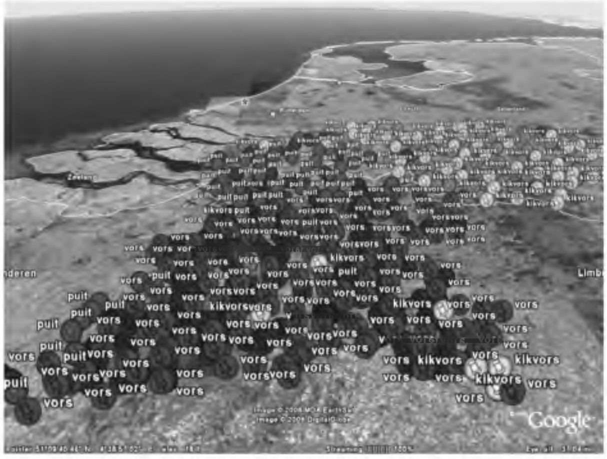


Figure 8: The three most frequent WBD heteronyms for "kikker" (frog) displayed in Google Earth

5. Conclusion

In the present paper we have reconsidered the data model of dialect geography and argued that the model helps to make new uses of the dialect resources more transparent. The focus has been on how adhering to this model helps in uniting the data and classifications from the different dictionaries in the ReWo, some of which are traditionally form based, most of which are sense based. We suggested to treat all data from the different dictionaries as one huge data set and let differences in the more precise nature of each of the data types be specified by the classifications. By doing so we shift all troubles in unifying the dictionaries to the classifications part of the model.

More information about D-square can be found on the project website: http://www.ru.nl/dialect/d2.

6. Acknowledgements

The project D-square is partly funded by Netherlands Organisation for Scientific Research (NWO).

7. References

Francopoulo, G., George, M.,Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC).* Genoa.

Ghijsen, H.C.M. (1964). Woordenboek der Zeeuwse Dialecten. Den Haag: Van Goor en Zonen.

Haeseryn, W., Romijn, K., Geerts, G., Rooij, J. de., Toorn, M.C. van den. (1997). Algemene Nederlandse Spraakkunst. Groningen/Deurne: Martinus Nijhoff uitgevers/Wolters Plantyn.

Ide, N., Romary, L. (2004). A Registry of Standard Data Categories for Linguistic Annotation. In *Proceedings of. the Fourth Language Resources and Evaluation Conference (LREC).* Lisbon, pp. 135-139.

Keymeulen, J. van. (2004). Trefwoorden en lexicale varianten in de grote regionale dialectwoordenboeken van het zuidelijke Nederlands (WBD, WLD, WVD). In J. De Caluwe, G. De Schutter, M. Devos, J. Van Keymeulen (Eds.), *Taeldeman, man van de taal, schatbewaarder van de taal.* Gent: Academia Press.

Kircz, J. (2004). E-based Humanities and E-humanities on a SURF platform. Utrecht: Stichting SURF.

Kloeke, G.G., Grootaers, L. (1934). Kloeke's systematisch en alfabetisch register van plaatsnamen voor Noord-Nederland, Zuid-Nederland en Fransch-Vlaanderen. 's-Gravenhage: Nijhoff.

Kruijsen, J. (1996). De Nijmeegse dialectlexicografische projecten. *Trefwoord,* 11. pp. 93-107.

Kemps-Snijders, M., Wittenburg, P. (2006). LEXUS - a web-based tool for manipulating lexical resources. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC).* Genoa.

Tollenaere, F. De., Weijnen, A. (1963). Het dialectwoordenbooek. In *Woordenboek en dialect. Lezingen gehouden voor de Dialectcommissie der Koninklijke Nederlandse Akademie van Wetenschappen op 4 november 1961 door Dr. F. de Tollenaere en Prof. Dr. A. Weijnen.* Amsterdam: N.V. Noord-Hollandsche Uitgevers Maatschappij.

WBD (1967-2005). Woordenboek van de Brabantse Dialecten. Assen: Van Gorcum.

Weijnen, A. (1961). De semantische en syntactische problematiek van het dialectwoordenboek. In *Tijdschrift voor Nederlandse Taal- en Letterkunde,* 78, 2.

Wittenburg, P., Broeder, D., Piepenbrock, R., Veer, K. van der. (2004). Databases for Linguistic Purposes: a case study of being always too early and too late. In *Proceedings of the E-MELD Workshop.* Detroit, pp. CD-ROM.

WLD (1983--). Woordenboek van de Limburgse Dialecten. Assen: Van Gorcum.

WVD (1979--) Woordenboek van de Vlaamse Dialecten. Gent/Tongeren: Michiels.

[7] This will be possible when there is an XML implementation of LMF. The XML implementation is expected in the spring of 2006.

[8] http://earth.google.com