

A General Lexicographic Model for a Typological Variety of Dictionaries in African Languages

Gertrud Faaß, *Department of Information Science and Natural Language Processing, University of Hildesheim, Germany (gertrud.faaß@uni-hildesheim.de)* and *Department of African Languages, University of South Africa, Pretoria, South Africa*

Sonja E. Bosch, *Department of African Languages, University of South Africa, Pretoria, South Africa (boschse@unisa.ac.za)*

Rufus H. Gouws, *Department of Afrikaans and Dutch, Stellenbosch University, Stellenbosch, South Africa (rhg@sun.ac.za)*

Abstract: So far, there have been few descriptions on creating structures capable of storing lexicographic data, ISO 24613:2008 being one of the latest. Another one is by Spohr (2012), who designs a multifunctional lexical resource which is able to store data of different types of dictionaries in a user-oriented way. Technically, his design is based on the principle of a hierarchical XML/OWL (eXtensible Markup Language/Web Ontology Language) representation model. This article follows another route in describing a model based on entities and relations between them; MySQL (usually referred to as: Structured Query Language) describes a database system of tables containing data and definitions of relations between them. The model was developed in the context of the project "Scientific eLexicography for Africa" and the lexicographic database to be built thereof will be implemented with MySQL. The principles of the ISO model and of Spohr's model are adhered to with one major difference in the implementation strategy: we do not place the lemma in the centre of attention, but the sense description — all other elements, including the lemma, depend on the sense description. This article also describes the contained lexicographic data sets and how they have been collected from different sources. As our aim is to compile several prototypical internet dictionaries (a monolingual Northern Sotho dictionary, a bilingual learners' Xhosa–English dictionary and a bilingual Zulu–English dictionary), we describe the necessary microstructural elements for each of them and which principles we adhere to when designing different ways of accessing them. We plan to make the model and the (empty) database with all graphical user interfaces that have been developed, freely available by mid-2015.

Keywords: AFRICAN LANGUAGES DICTIONARIES, LEXICOGRAPHIC MODEL, MACRO-STRUCTURE, MICROSTRUCTURE, ACCESS STRUCTURE, ISO24613:2008, MYSQL, MULTIFUNCTIONAL LEXICAL RESOURCE

Opsomming: 'n Algemene leksikografiese model vir 'n tipologiese verskeidenheid woordeboeke in Afrikatale. Tot dusver bestaan daar min beskrywings oor hoe

om strukture te skep wat daartoe in staat is om leksikografiese data te berg. ISO 24613 2008 is een van die mees onlangse sodanige beskrywings. Nog een, naamlik dié van Spohr (2012) wat fokus op die ontwerp van 'n gebruikersgerigte multifunksionele leksikale bron, gebruik die voorstellingsmodel XML/OWL (eXtensible Markup Language/Web Ontology Language) wat in beginsel hiërargies is. In hierdie artikel word 'n ander roete gevolg om 'n model te beskryf wat gebaseer is op entiteite en hul onderlinge verhoudinge. MySQL (gewoonlik na verwys as Structured Query Language) beskryf 'n databasisstelsel van tabelle wat data bevat en definisies van hulle onderlinge verhoudinge. Die model is ontwikkel binne die konteks van die projek "Scientific e-Lexicography for Africa" en die databasis wat saamgestel word, sal met behulp van MySQL toegepas word. Die beginsels van die ISO-model asook dié van Spohr word gehandhaaf maar wel met een groot uitsondering in die implementeringstrategie: die lemma is naamlik nie sentraal in die bewerking nie, maar wel die beskrywing van betekenisonderskeidings — alle ander elemente, met insluiting van die lemma, is afhanklik van die beskrywing van die betekenisonderskeidings. Hierdie artikel bespreek ook die leksikografiese datastelle wat aangebied word en hoe hulle uit verskillende bronne versamel is. Aangesien dit ons doel is om verskillende prototipiese internetwoordeboeke saam te stel ('n eentalige Noord-Sotho woordeboek, 'n tweetalige aanleerderswoordeboek Xhosa-Engels en 'n tweetalige woordeboek Zoeloe-Engels) bespreek ons die noodsaaklike mikrostrukturele elemente vir elkeen van hierdie woordeboeke en watter beginsels gevolg word om verskillende maniere te ontwikkel om toegang tot hierdie woordeboeke te verkry. Die plan is om die model en die (leë) databasis met al die grafiese koppelvlakke wat ontwikkel is teen die middel van 2015 gratis beskikbaar te stel

Sleutelwoorde: AFRIKATAALWOORDEBOEKE, ISO24613:2008, LEKSIKOGRAFIESE MODEL, MAKROSTRUKTUUR, MIKROSTRUKTUUR, MULTIFUNKSIONELE LEKSIKALE BRON, MYSQL, TOEGANGSTRUKTUUR

1. Introduction

This article is concerned with the design of a lexicographic model, that is, a model of a data structure capable of storing lexicographic data, which will subsequently be used to compile several types of prototypical dictionaries for a selection of African languages¹. We keep in mind that there are no hard and fast rules for any typological model, but rather that different types of dictionaries may have certain features in common (Gouws and Prinsloo 2005: 45). In the last few years, several such lexicographic data collection models were published; the most general of all is the ISO standard for lexicography (ISO 24613:2008). This "Lexical Markup Framework²" (LMF) builds the background for several existing lexicographic data collections. A data collection model is not a database as such, but is defined as a standoff-XML-formatted framework of a number of files plus several external sources, each describing a different aspect of the dictionary that is compiled from them. For example, general data such as language or language coding is included, but also microstructural data related to lemma signs, such as information about its part of speech or its orthography. Concerning the possibilities to connect with other sources of

information, we agree with Spohr (2012: 23) who states that although LMF describes itself as interoperable, "it remains rather vague on its application in the various contexts, and in particular of its application in human usage situations".

Spohr's general graph-based formalism (Spohr 2012) can indeed be seen as an implementation of the LMF data model. His lexical resource, implemented in a graph based OWL model, is based on a typed formalism, similar to the adaptations the WWW is taking up to become the new Semantic Web (Spohr 2012: 38). Spohr places the lexeme in the focal point of the database, linking it for instance to its forms and senses (*ibid.* p. 68). He nevertheless states that "ideally, we would like senses to be the primary lexical entities, as all kinds of lexical relations seem to be defined between senses" (*ibid.* 67). Spohr, however, then argues against this concept saying that beginning with the item giving the sense (i.e. the item giving the paraphrase of meaning), it would not be possible to fill all other dependent fields, especially when acquiring lexicographic data from corpora (*ibid.* 68). This issue will receive further attention in section 6.3.

We want to mention two further publications here, which describe a lexicographic database or data collection model for generating online dictionaries in particular. A database that supplies several dictionaries for specific purposes with data is described by Bergenholtz and Bergenholtz (2013). In their article titled "One database, four monofunctional dictionaries", the kind of model that was utilized is unfortunately not mentioned. However, they do point out some items defined for the resulting database, as well as the fact that the compilation of several online dictionaries from one database, calls for a number of issues concerning its access features to be taken into account — see also our section 5 below.

Bosch, Pretorius and Jones (2007) propose a model for machine-readable lexicons, not only for the South African Bantu languages, but for the Bantu language family as a whole. The data model in the form of an XML DTD is intended to include all linguistic information of the languages in question and "provides flexibility and handles the various representations specifically applicable to Bantu languages, thereby making it applicable to diverse uses of machine-readable lexicons" as language resources for use in large-scale HLT/NLP applications. Only a fragment of the DTD is presented in the publication.

The majority of articles concerned with online dictionaries, however, refers to their visual representation (e.g. Prinsloo 2010 which is related to their implemented access strategies), others are concerned with the acquisition of data to populate lexicographic databases (e.g. L'Homme 2012 and Scholze-Stubenrecht 2013).

The research for this paper resides within a project entitled "Scientific e-Lexicography for Africa (SeLA)³" (i.a. described by Heid 2012), and it is carried out by the University of Hildesheim (Germany), the University of South Africa and the University of Pretoria, Stellenbosch University (South Africa), and the

University of Namibia in Windhoek (Namibia). The project intends to combine all of the above-mentioned issues: (1) designing a prototypical multifunctional database with the aim of compiling several monofunctional electronic dictionaries for the African languages; (2) solving the problem of data acquisition for resource-scarce languages; (3) defining "exactly which types of lexicographic data from the fact collection need to be selected in order to satisfy a given user need, as well as in deciding in which way such data have to be ordered and formatted (presented) for users with a given background and a given type of need" (Heid 2012: 438).

In the SeLA project, we are concerned with a multilingual African language data collection to be used for lexicographic purposes which we will store in a MySQL database. For the time being, the aim is not to compile comprehensive dictionaries from the database. Seeing the final implementation as a prototype, we plan to use this database for several other purposes, for instance, as part of intelligent Computer Assisted Language Learning (iCALL) software.

We consider it necessary to strictly differentiate between the database, which should be flexible, in other words, open to internal and external resources (so far unknown) to be added in the future, and the presentation of the (internal and external) data to the users, which depends on their requirements (see section 5). We also foresee access to a prototypical Natural Language Processing (NLP) machine performing morpho-syntactic analyses.

The database model is to be implemented with a MySQL database. Such a database may consist of (1) content tables containing the data itself, (2) relational tables linking data items with one another, and (3) tables generated from the data and their relations which are used for a faster access. One might wonder why we do not use XML/OWL, like the most up-to-date data collection models described above. Besides the fact that the SeLA team lacks the capacity to develop a full-scale Dictionary Writing System (DWS) or to make use of one to compile a full-scale dictionary, we consider a populated MySQL database implementation as equal to a standoff XML system. In both systems, all necessary data items can be described and a number of types of relations between those data items can be modelled. SQL, however, additionally allows for a fast and easy implementation without the need for DTDs, XML-editors or (commercial) Dictionary Writing Systems. Moreover, together with phpMyAdmin⁴, an online dictionary and the necessary maintainer facilities are speedily and simply implemented with a few PHP scripts. Another point of consideration is that most of the data will be imported from existing resources, which will populate the fields of the database only partially. The task of filling the gaps and generating full-scale dictionaries must be postponed to a later stage. To use MySQL for a start, does not imply that XML/OWL will not be used in the future. In such a case, the means will be found to fill the database with sufficient data to compile comprehensive dictionaries, and porting one system to the other will indeed be possible.

In summary, we describe a lexicographic model in this article which should

fulfil various requirements: (1) it should be open to a number of lexicographical functions as several different monofunctional online dictionaries will be compiled from it; (2) it should cover the specific linguistic phenomena of the languages belonging to the Bantu language family; and (3) concerning data acquisition — as we will need to populate the database with any relevant data that can be collected semi-automatically — the database should be tolerant of missing data items, even if they are considered essential for producing a dictionary. Furthermore, we will describe our current approach towards data acquisition and data accessibility.

2. Aims

Our aim as part of the SeLA project is to design and develop a lexicographic database that will contain multilingual data of three of the official African Languages of South Africa (i.e. Zulu, Northern Sotho and Xhosa). For some of these data sets, translation equivalents of South African English will be stored too. The data of other African languages, as well as Afrikaans, are foreseen to be added at a later stage. We begin by developing a database model, with the aim of fulfilling all the requirements to describe the language items thoroughly, while taking into account the languages in question and the external resources that are currently available. We take Spohr's (2012) data collection model into account too; however, as Spohr has suggested, we focus our attention on the polysemous senses of a word — the above-mentioned disadvantages (see section 95) only play a minor role for us, as is the case with the languages concerned, there are only few resources available which would allow for an automated filling of the database — most data will have to be added manually. The database will be utilised to compile a typologically diverse collection of prototypical monofunctional dictionaries (however, with few data sets), of which the majority are planned to be bilingual. Hence, we look at requirements of a good outer and inner access structure (see section 5), resulting in the design of different dynamic graphical user interfaces (GUIs) to be developed.

We will then examine ways and methods to import available external resources (the respective plans are described in section 6). Lastly, we plan to bind the resulting database into a language portal, a framework of lexicographic and other resources. We foresee linking it with other dictionaries, corpora, or other databases containing linguistic data, such as the ontology database of the part-of-speech items of Zulu and Northern Sotho described by Faaß, Bosch and Taljard (2012) or the e-learning tool "eZulu dictionary of possessives" assisting learners of the language in acquiring knowledge about producing possessives structures in Zulu, described by Bosch and Faaß (2014).

Setting the aims as described above, we need to examine aspects regarding macrostructures and microstructures of the foreseen dictionaries. On this basis, the data model can then be designed.

3. Aspects regarding the macrostructure and microstructure

3.1 Macrostructural elements for Bantu language dictionaries: a challenge of lemmatisation

The agglutinating nature of the Bantu languages that goes hand in hand with a complicated nominal and verbal derivation system, indeed poses challenges for lemmatisation (Gouws and Prinsloo 2005: 67). Different approaches to lemmatisation, the main one being word versus stem lemmatisation in the case of nouns and verbs, play an important role in dictionary compilation.

Because of the conjunctive writing system of Zulu, whereby parts of speech are written together, even full sentences may appear as one orthographic word. The sentence *bazokubona* "they will see it", for example, consists of several morphemes; *ba-* (subject concord of noun class 2) *-zo-* (future tense marker) *-ku-* (object concord of noun class 15) *-bon-* (verb root = "see") *-a* (verbal ending). We do not foresee to enable our system to analyse such input data, however, linguistic verbs consisting of several morphemes should, in principle, be analysed so that users can receive the data on the items related to their query. Users interested in stems on the other hand, should also be able to query those and get the data on all full forms containing a particular stem.

Concerning the disjunctively written Sotho languages, there are other challenges: The copulative of Northern Sotho, for example, consists of one or several, disjunctively written morphemes. These morphemes are highly ambiguous and the copulatives generated from them are homographous, too. The many forms cannot all be described in a printed dictionary due to space constraints. However, even in an electronic dictionary, the task of describing all forms might turn out to be too complex. An attempt has been made to extract these forms from corpora by using regular expressions (Faaß and Taljard 2013), however, due to the many homographs, no system to distinguish them could be found. Such rather morpho-syntactic challenges can be related to the issue of accessibility. We therefore do not see the electronic dictionary itself as the best solution, but rather develop connected systems that could, for instance, assist learners in producing the correct form, such as a decision tree-like device (described in Prinsloo, Bothma, Heid and Faaß 2012).

In an electronic dictionary, these analyses of input data, however, belong to access structure (see section 5), not to the data storage itself. One could, therefore, argue that in a lexicographic electronic data collection there is no macrostructure at all.

We place the sense element at the centre of our database, and since we link this sense with one (or more) orthographic forms and with a stem, we enable our system to allow for immediate access to stems of verbs and nouns, for instance, the Northern Sotho verb stem *bona* "[to] see", but also to full forms such as the Zulu address *sobonana* "see you (again)". Therefore, in terms of

orthographic forms, we foresee simplex and complex words which are both related to sense elements.

The change of focus is exemplified in the following two figures. Figure 1 illustrates a possible entry describing the English verb "[to] see" and its Zulu counterpart "[uku]bona" in a traditional lexicographic database where the lemma is the central element, and is linked to two senses, each extended with an example. The two translation equivalents are linked with each other.

In Figure 2, the same data is viewed from the perspective of our proposed model where English and Zulu data are entered independently, similar to Figure 1. The relational table "is_translation_of" informs that sense 1 and sense 3 are translation equivalents. Note that in Figure 1, the metaphorical sense of "[to] see"/"[uku]bona" was described in each language in the element "sense 2". In the new model, such a sense description does not appear as such. Instead, a literal sense description of "[to] understand"/"[uku]qonda" is included together with an example ("I understand what you mean."). The literal senses 1/2 ("[to] see"/"[to] understand") and senses 3/4 ("[uku]bona"/"[uku]qonda") are then linked with each other by items in the table "is_synonym_to" (see section 3.3). In this table, we learn that the synonymy is metaphorical and we also see the respective example sentences ("I see what you mean" | "Ngiyabona ukuthi uthini").

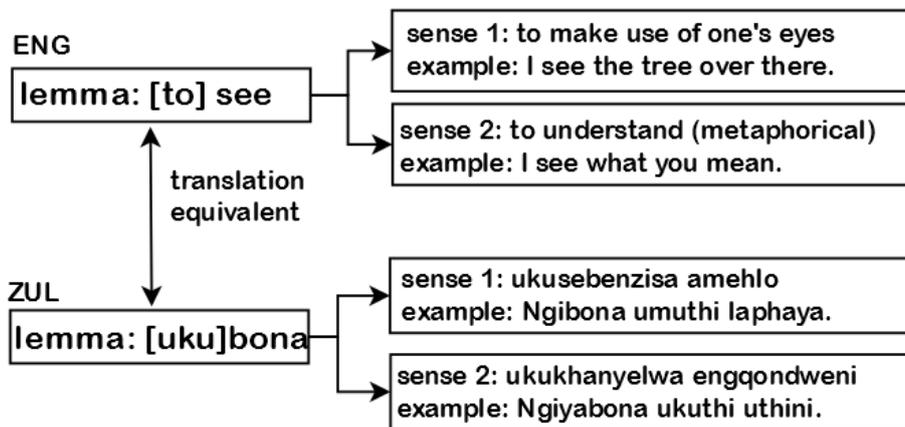


Figure 1: Illustration of the traditional data model: focus on the lemma

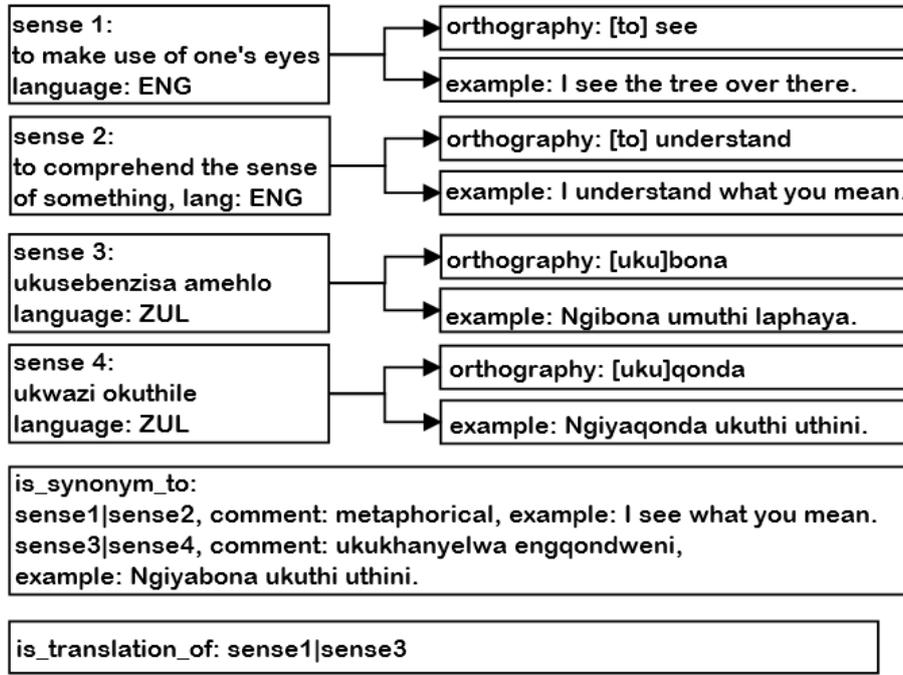


Figure 2: Illustration of the proposed data model: focus on the sense description

3.2 Microstructural items

We began with a general list of items which are usually part of the microstructural items in any dictionary, such as the lemma sign, its paraphrases of meaning, etcetera. For each of these items, we decided whether we require them for our database. Afterwards, we added all items that usually appear in the respective African language dictionaries that we are concerned with. We subsequently categorised the items, which we currently foresee: we generally differentiate between the categories "descriptions", "morpho-syntax", "phonetics", "etymology", "valency", "examples" and "idioms". Each of the tables representing these categories contains its microstructural items. As described above, we need to differentiate between data items to be filled for stems (the ones that are not identical with full forms) and data items to be filled for full forms. Table 1 shows the items foreseen, irrespective of the language they belong to. For the African languages, we add information on whether the item is described for full forms, for stems or for function words.

<i>Data category</i>	<i>Item giving the</i>	<i>Gloss</i>	<i>Full form y/n</i>	<i>Stem y/n</i>	<i>Function words y/n</i>
sense	short paraphrase of meaning	one or several brief semantic descriptions of what the item means (in comparison to other lemmas)	yes	optional	optional (for adv. prefixes/particles)
	paraphrase of meaning	one or several long semantic descriptions of what the item means (in comparison to other lemmas)	yes	optional	optional (for adv. prefixes/particles)
	source	source of brief or long semantic description	yes	optional	optional
	style marker	groups of humans who use the item (e.g. woman only, closed set)	yes	no	no
	subject area	what kind of subject does the item refer to (closed set)	yes	no	no
morpho-syntax	class	free unit (syntactically free: simple or complex, or idiom), or clitic (proclitic, mesoclitic or enclitic), or bound unit (stem or affix) see also Spohr (2012: 69)	yes	yes	yes
	abbreviation	an abbreviation of the item	yes	no	no
	degree of comparison	describes the degree of the item (the positive is not marked): comparative, superlative	yes	no	no
	gender	gender of the item (closed set)	optional (only persons)	no	no
	part-of-speech	morpho-syntactic classification (closed set)	yes	no	yes
orthography	grapheme lemma sign	one or several surface forms of the item in question	yes	no	yes
phonetics	pronunciation	pronunciation of the item (we have yet not decided on the format of this item, we however foresee using IPA)	yes	no	yes
	stressed syllable(s)	indicate which syllables of the item are stressed	yes	no	yes
	syllable division	result of syllabification	yes	no	yes
etymology	long description	long description of the etymology of the item (not a word formation issue)	yes	no	optional
	short description	brief description of the etymology of the item (not a word formation issue)	yes	no	optional
	source	source of long/short description	yes	no	optional
examples	example phrase	a phrase in which the item typically occurs in (to demonstrate the use of the item)	yes	no	yes

	example sentence	a sentence in which the item typically occurs in (to demonstrate the use of the item)	yes	no	yes
	source	source of example phrase or sentence	yes	no	yes
idioms	fixed expression	one or more example(s) of a (partially) idiomatic phrase (e.g. "kick the bucket" in English) which the item occurs	yes	no	yes
	idiom	one or more example(s) of an idiom ("der Krug geht so lange zum Brunnen bis er bricht") that the item occurs in	yes	no	no
	frequency of occurrence	how often does the idiomatic sense of the item occur in a corpus which — if possible — should be near representative; eventually we will use relative frequencies to abstract from the size of the corpus	yes	no	no
	source	source of fixed expression or idiom	yes	no	yes

Table 1: Microstructural items for all languages contained in the database

The items we need for the African languages only are listed in Table 2.

<i>Data category</i>	<i>Item giving the</i>	<i>Gloss</i>	<i>Full form y/n</i>	<i>Stem y/n</i>	<i>Function words y/n</i>
morpho-syntax	government	this item is in general used for items that govern the case in which another must appear (some German prepositions, for example, require their argument to appear in the dative). Concerning the African languages, we make use of this item to describe the influence of a conjunction on the verbal moods to follow it	no	no	yes
	noun class	the noun classes of the Bantu languages replace number/gender classes of other languages	yes	no	optional
phonetics	tone marker	high or low tone might lead to a difference in meaning	yes	yes	yes

Table 2: Microstructural items contained for the African languages only

Lastly, Table 3 contains the items only used for Afrikaans or English, respectively. We do not claim the tables to be comprehensive, other items might be added at a later stage.

Languages	Category	Item	Gloss
Afrikaans	morpho-syntax	attributive marker	marks the adjective taking a derivational "-e" when being used attributively
Afrikaans/ English	morpho-syntax	case	case that a verb refers to (closed set)
Afrikaans/ English	morpho-syntax	inflection	person, number and gender that a verb refers to (closed set)

Table 3: Microstructural items necessary for non-African languages only

3.3 Relational tables

While and after the data items are stored in the database with their respective descriptive items, additional tables describing the relations between them will be defined. In addition to the usual morpho-syntactic relations (e.g. "is-plural-of"), semantic relations are described too (e.g. "is-near-synonym-of"). So far, we do not foresee adding WordNet data. However, this is possible from a technical perspective, since the development of a prototype African Wordnet (AWN), which currently includes four languages, is an on-going project (Griesel and Bosch 2014). The resource has been developed by translating Common Base Concepts (CBC) from English and currently holds roughly 42 000 synsets.

To assign translation equivalents, we use the relation "is-translation-of". A rather general relation will be added as well: "is-linked-with" will contain relations between items not described in the others (i.e. miscellaneous kinds of relations that appear not frequent enough to give reason for an own relational table). This last table, however, will contain a data field where the type of relation is explained.

We relate senses of lemmas with the following tables:

- is-diminutive-of (for nominal items only)
- is-plural-of (for nominal items only)
- is-locative-of (for nominal items only)
- is-stem-of (see lemmatisation strategy above)
- is-homonym-of
- is-near-synonym-of
- is-antonym-of
- is-translation-of (relates items of different languages to each other)
- is-contained-in-example-sentence
- is-contained-in-fixed-expression

- is-contained-in-idiom
- has-morpho-syntax (relates a specific id of a type of morpho-syntactic item to one sense)
- has-phonetics
- has-valency (relates a specific id of a type of valency to one sense of an item taking arguments)
- is-linked-with

For space reasons, we describe only two of the tables in the following sections.

3.3.1 "has-morpho-syntax"

In any typical dictionary, the microstructure contains information on morphology and syntax of a lemma. Such information is repetitive not only for parts of speech appearing several times, but also for their morphological properties. Plural morphemes of English, for example the "-s" appearing in nouns like "type – types", "house – houses", must only be described once in our model. We foresee to fill a table called "morpho-syntax" with all the appearing categories (e.g. *noun*, -s). Each of the categories receives a unique id. In the relational table "has-morpho-syntax", we link the sense descriptions with one or several id(s) of morpho-syntactic categories that apply to them.

3.3.2 "has-valency"

Concerning the valency (or "valence", as described by Spohr 2012: 86f) of a lexicographic item, a similar situation occurs: one type of valency, for example "*verb, taking no object*" can be linked with several words ([to] sit⁵, [to] walk, etc.). We handle the situation in the same way as the "has-morpho-syntax"-table described above. A unique id is assigned to each valency type and sense descriptions are then related to the ids that apply to them.

Some relations between items will be added manually. For this purpose and the purpose of checking and correcting the data that will be inserted automatically (see section 6.4), the database will offer a maintainer interface.

4. Design and implementation method

In this section, we compile the items described above and define a basic lexicographic model where each category represents one table of the database (DB), see Figure 3 which, due to space constraints, does not show all of the items. In our model, we tentatively define relations between items, however, keeping them open for future changes by storing them into separate tables. In MySQL,

each item is identified via an "id"-data element (e.g. "sense-id" identifying one specific paraphrase of meaning). Such identifiers are marked as "primary key", which means that each may only appear once in the respective table. In the model shown in Figure 3, each of the items contained are to be pre-defined in respect of their type, "int" stands for integer, "varchar" for any kind of character. Lastly, "link" means that a URL will be entered.

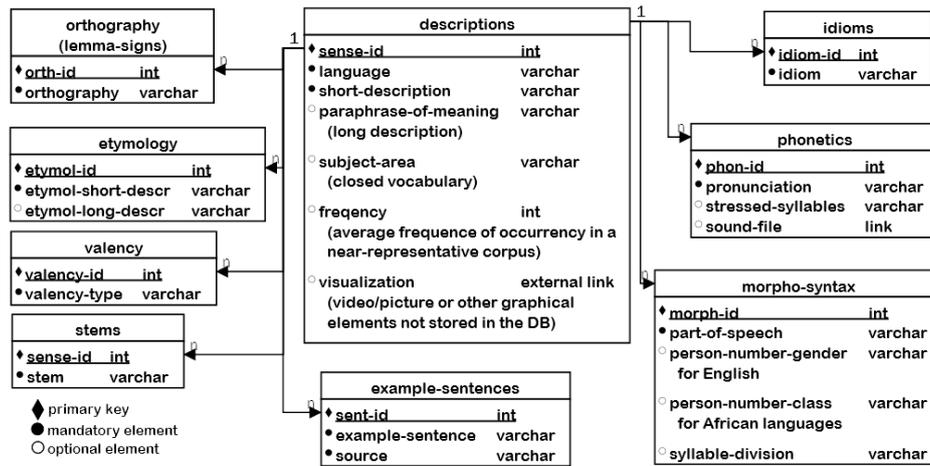


Figure 3: The basic database model showing tentative relations between data items

In an SQL database, relations between items of tables are to be described, which reflect dependencies between items (we can also define item as "hierarchies", as it is done in XML or in object-oriented database systems). A paraphrase of meaning, for example, should directly be related to one or several example sentences, similar to an integrated microstructure. The relation between those items is, therefore, 1:n where "n" stands for any integer number greater than zero. For example, the relation between the items "sense-id" of the table "descriptions" and "sent-id" of the table "example sentences" could be defined as "1:n". However, it could very well be the case that we could use one example sentence several times, by assigning several lemmas (or rather senses of those words) to it, therefore, we do not enforce the 1:n relation by directly linking items (e.g. foreign keys), but rather implement the word sense/example sentence relation by assigning a unique key to each of those items in the respective tables and by adding a separate table linking those ids to each other, see Figure 4.

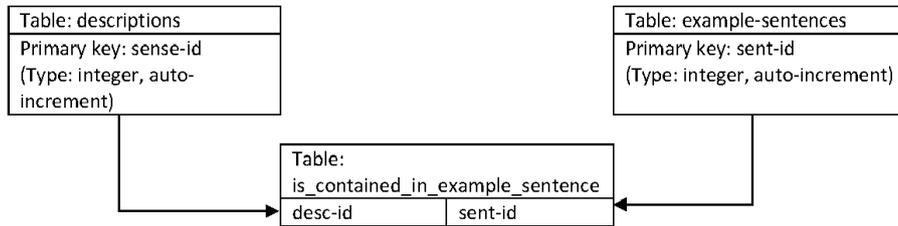


Figure 4: Adding relations between word-sense and example-sentence

The positive aspect of such an implementation is its openness towards a re-
definition of relations between items; a negative aspect might be that such
tables lead to a slow query processing of the database. Therefore, in our second
phase of implementation (i.e. after the available data will have been stored in
the database), we will automatically generate additional tables each containing
all relevant data for one of the dictionaries. Users will have access to each one
of these tables with one mouse click and one or several query words.

5. Data presentation: access structure

Bergenholtz and Gouws (2010: 103) maintain that "of critical importance in a
user-driven lexicographic approach is the need to ensure that the target users
of a specific dictionary gain unimpeded access to the data they need in order to
achieve an optimal retrieval of information". Such accessibility is typically
ensured by the access structure of any given dictionary. We adhere to the defi-
nition of Wiegand and Beer (2013: 111), who define accessibility as follows:
"The term 'data accessibility' refers to the access willingness and thereby to the
possibility to look up textual and illustrative lexicographical data; it is given
because the data are in the access domain of an access structure. A distinction is
made between the external and the internal data accessibility".

In printed dictionaries, the first step is determined by the knowledge a
user has of the specific dictionary. A user could embark on either the full or a
shortened outer access process, reaching the desired lemma via a rapid access
structure, for instance; thumb index markers or alphabet letters, or by merely
guessing where the relevant item will be and then following the running heads
until the desired page has been reached. Going down the lemmata, the desired
guiding item can then be found — the item, where the inner access route com-
mences. In e-dictionaries, a single word or multi-word string is typically typed
into the search box and this will immediately guide the user to the required
lemma sign without bringing any other outer access items into play. Other
systems offer a rapid access structure in the form of a list of clickable lemma

signs of which the user can select the required one. It is also possible to offer both, as described by Bothma and Gouws (2013).

The selection and the order of appearance of the data items both depend on several factors: (1) The type of dictionary; a bilingual dictionary will require a translation equivalent to appear, while a monolingual will not. (2) The part of speech of the lemma; some parts of speech need to be displayed with valency information, for others, valency plays no role. (3) The access route; the first resulting screen of a query will display only few items, from there, the user may click respective boxes on the screen to get more data (e.g. etymological information, idioms or example sentences). For each of the microstructural items above we need to define when it will appear on the screen (given that an orthographic form was entered as a query and this form was found in the database). Table 4 shows these decisions for several of the microstructural items above when a general monolingual dictionary is compiled; due to space constraints, not all assignments can be shown.

	<i>Only specific part-of-speech</i>	<i>Access route: first</i>	<i>Access route: more info</i>	<i>Solely on demand</i>
short paraphrase of meaning	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
paraphrase of meaning	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
source of paraphrase	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
style marker	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
subject area	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
class	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
abbreviation	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
degree of comparison	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
noun class	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
tone marker	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
etymological short/long description	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
example sentence	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
fixed expression	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Table 4: Examples of microstructural items being assigned to specific use situations

In section 5 above, we mentioned that for each of the foreseen dictionaries we will generate one table in the database containing all the necessary data. Table 4 above shows their elements for the planned monolingual dictionary of Northern Sotho.

5.1 External links

From a technical perspective, the database is planned to be connected, *inter alia*, with a morphological analyser. This is essential especially for the African languages that are written conjunctively; a user may enter, for instance, the orthographic word *abazukukhombisa* "they will not show it" — without knowing that this expression consists of a number of morphemes: *a-* (negative morpheme), *-ba-* (subject concord class 2), *-zu-* (future tense negative morpheme), *-ku-* (object concord class 15), *-khomb-* (verb root), *-is-* (causative extension), *-a* (verbal ending). Whenever such a query word is not found as a lemma by the database, this morphological analysis will be executed in order to deliver the linguistic units and their parts of speech which will be queried automatically by the system. The user will then see the results for each of the parts presented by the system and can select the items he or she is interested in to get further information displayed.

On the other hand, a user might enter a stem of a word; in this case, we will use the morphological analyser as generator and will generate full form words which could be queried in the database. It is foreseen to then suggest this list to the user, in order for the user to subsequently choose the ones he or she wants to know more about. Concerning productive purposes, we also foresee (user-activated) connections with the decision-tree system developed in the framework of the SeLA project (e.g. described in Prinsloo, Bothma, Heid and Faaß 2012). Another option will be to access corpus data, however, only maintainers will be allowed to see the whole of the data, as one cannot assume that all corpus data would be usable for exemplifying the meaning of a word (see section 6). The maintainers then will be able to choose example phrases or sentences to be added to the database.

6. Resources to be added to the database

It would be virtually impossible to fill such a database from scratch — corpora are scarce and the ones that do exist lack a description of their contents and are, therefore, not feasible for an automated retrieval of dictionary contents. However, there are some resources that we can indeed utilise for a start, as described below.

6.1 Available resources for the project

Language data for Northern Sotho is currently available in the form of a printed dictionary (Ziervogel and Mokgokong 1985), which was scanned⁶ into electronic format by means of Optical Character Recognition (OCR) and transformed at least partially to a structured data collection (Kebbe 2013). We also use a MySQL database containing about 600 full Zulu forms and their English

translation equivalents, generated in the SeLA sub-project on a Zulu dictionary of possessive constructions (Bosch and Faaß 2014). Lastly, we also have access to a file containing several thousand Xhosa nominal stems, information on the noun classes they appear in and their translations into English.

6.2 Other possible resources

In South Africa, the co-ordination of language resources is still in its infancy stages, however, the function of the newly established Language Resource Management Agency (RMA) is to develop and host reusable text and speech resources, and to manage and distribute these from one central point. Currently, relevant resources available are Annotated Text Corpora for all official languages of South Africa annotated with lemma, part of speech and morphological analyses. Initial versions of core technologies, namely lemmatisers, part of speech taggers and morphological decomposers are available as open source modules and could, therefore, be used for the annotation of text corpora of the various Bantu languages, although (Eiselen and Puttkammer 2014: 3702) point out that "there is still a lot of room for improvement, especially for lemmatisation and morphological decomposition".

6.3 Adding resources to the database

Despite several corpora for the African language that we are permitted to use for the purpose of, for instance, checking corpus frequencies of occurrences⁷ to be added (manually) to the database at a later stage, we were also able to get access to a scanned dictionary (Ziervogel and Mokgokong 1985). Unfortunately, the files we received were in word format, and all items were in the same font, so it was impossible to automatically identify item types by their format. Judith Kebbe, a student of information science at University of Hildesheim worked out an automated method to identify item types by their position in the dictionary article (Kebbe 2013) and wrote Perl scripts extracting those items, based on the descriptions of Faaß, Ramagoshi and Sebolela (2009). Her work resulted in structured, machine-readable data covering about half of the entries of the dictionary. As it turned out, however, the microstructure of this dictionary is not structured consistently; when trying to extract translation equivalents, especially, the automated method often failed. Another problem is described by Kosch (2013: 204) who points out the mixed lemmatisation approach of this dictionary, whereby a word approach is applied to nouns with irregular or non-overt class prefixes, although the overriding approach in the dictionary is stem based. The user is then given a cross-reference to the relevant stem. Examples are nouns such as *mmuši*, "ruler" and *pono* "vision", which are lemmatised as words and not as stems. According to the stem-based approach, the lemmatisation of the two nouns would have presented as *buši* and *bono*,

derived from the verb stems *-buša* "rule" and *-bona* "see" respectively.

Kebbe extracted several thousand links between dictionary entries, but only few dictionary entries describing translation equivalents. Hence, these data will be loaded into our database to cater for monolingual Northern Sotho only, mainly to test relational tables such as for instance "is-linked-with".

Bosch and Faaß (2014) populated a MySQL database with about 600 Zulu nouns and about 900 English translation equivalents, there is also information on their classes and numbers stored in this database. We will transfer these data to the SeLA database as well.

Bilingual Xhosa–English data was made available to us in .xls format. Here, not surface forms but several thousand noun stems, the classes they appear in, class prefixes, and English translation equivalents are contained. By way of shell scripts, we will generate full forms and fill the database with the respective data.

With the available resources, we cannot fill the sense descriptions in most cases; therefore, we will have to add them manually. During the import of the data, we plan to use English translations to have these mandatory fields filled, but these will have to be replaced manually with monolingual sense descriptions. As our team will not have the manpower to fill all of the foreseen database items, we plan to send out calls to the public, trying to find volunteers, as soon as the graphical user interfaces have been completed. For our aim to compile prototypical dictionaries, we consider the available data to be sufficient.

6.4 An example: monolingual Northern Sotho data

This article describes a lexicographic model which is still awaiting implementation. While implementing it, we might find errors or inconsistencies that will force us to change the model. Therefore, at this stage, we can only describe data that was examined during the development of the model. We chose the dictionary of Ziervogel and Mokgokong (1985) that contains several thousand noun stems with additional information. One of the dictionary entries contains data on the noun stem *mente*:

```
MENTE, -/di- (mêntê) munt (waar geld geslaan word) // mint (where money  
is coined)
```

Kebbe (2012:34) generated the following, machine-readable data from of this entry:

```
<entry>  
  lemma:MENTE  
  cppl:di  
  <translation>  
    <Afrikaans>munt (waar geld geslaan word)</Afrikaans>  
    <English>mint (where money is coined)</English>  
  </translation>  
</entry>
```

With these data, we cannot provide a Northern Sotho sense description to fill the mandatory item "short description" in the "descriptions"-table of our database. In a first attempt, we hence foresee to write scripts that make use of the English translation. The scripts however add the note "TO-BE-TRANSLATED-INTO-NSO" as an indication for the manual reworking which is foreseen at a later stage. The "language"-field can be filled automatically because we know that this is NSO data. Optional elements (as shown in Figure 3) are not filled:

1. Table "descriptions":
sense-id: 1,
language: NSO,
short-description: TO-BE-TRANSLATED-TO-NSO: mint (where money is coined),
paraphrase-of-meaning: empty field,
subject-area: empty field,
frequency: empty field,
visualization: empty field.

Next, we process the information on morpho-syntax: cppl stands for "class prefix plural" which is an indication that this new database entry describes a noun. As this prefix is *di* and as no singular prefix is given, the scripts can assume automatically that the noun is of class 9 which means that its orthographic form of the singular is identical to the stem (*mente*). Therefore the singular orthographic form is *mente* and the plural form *dimente*. Since this dictionary uses diacritics to indicate tone, we also learn about the high tone on the second vowel. The scripts can hence fill several tables:

2. "morpho-syntax":
morph-id: 1, part-of-speech: noun, person-number-class: 03-sg-09,
morph-id: 2, part-of-speech: noun, person-number-class: 03-pl-10.
3. "stems"
stem-id: 1,
stem: mente.
4. "orthography"
orth-id: 1, orthography: mente,
orth-id: 2; orthography: dimente.
5. "phonetics"
phone-id: 1,
pronunciation: mêtê,
stressed-syllables: empty field,
sound-file: empty field.

Lastly, the scripts will fill the necessary relational tables creating links between the items.

1. "is-plural-of":
orth-id:2 | orth_id:1.

2. "is-stem-of":
stem-id:1 | orth-id:1,
stem-id:1 | orth-id:2.
3. "has-morpho-syntax":
sense-id:1 | morph-id:1.
4. "has-phonetics":
orth-id:1 | phone-id:1.

7. Summary and future work

This article describes the design of a lexicographic data model which will be implemented with MySQL, resulting in a database capable of storing lexicographic data of several of the official languages of South Africa. We aim at compiling several prototypical dictionaries from there: a monolingual Northern Sotho dictionary, a bilingual Xhosa–English general language dictionary and a bilingual English–Zulu learners' dictionary. We have compiled lists of necessary microstructural elements and have decided to put the sense description at the centre, the "lemma" being just a realisation of the sense, in other words its surface form.

We have collected a number of resources, which will be loaded onto the database semi-automatically. At this stage, it is foreseen that all missing data items will require manual adding due to the lack of available resources. It is well known that the development of resources for African languages is often of a fragmented nature — the resources tend to be small, only usable for restricted purposes and, therefore, excluding connection with other resources. We, therefore, intend to investigate collaborative approaches and technologies for the accumulation and creation of data to ensure the continued filling of this lexicographic database (cf. Benjamin 2014).

8. Endnotes

1. The term "African languages" refers to languages belonging to the Bantu language family. Both terms are used in this article.
2. Francopoulo, G. (Ed.). 2013. *LMF Lexical Markup Framework*. London: Wiley-ISTE. ISBN: 978-1-84821-430-9.
3. SeLA is supported by the "Deutscher Akademischer AuslandsDienst", DAAD in their programme "Welcome to Africa", see also www.uni-hildesheim.de/iwist-cl/projects/sela/.
4. phpMyAdmin is available from www.phpmyadmin.net
5. For ease of understanding, we make use of the orthographic forms in our example. In the data model foreseen, we will however relate sense descriptions with the "has-valency" ids.
6. We gained permission from the publisher to use the scanned dictionary at least for our current research purposes and hope that we will be allowed to use the resulting data for the prototype of our dictionaries.

7. We are very grateful to Prof. Prinsloo for allowing us to use his African Languages corpora (see De Schryver and Prinsloo 2000).

9. Acknowledgement

This article has been written in the framework of the project *Scientific eLexicography for Africa* (SeLA) working towards the development of electronic dictionaries and terminology databases for the (South) African Bantu Languages. The project is funded by the German ministry for education and research (BMBF, Funding number 55516493) in the Framework of the DAAD programme "Welcome to Africa" (see also <http://www.uni-hildesheim.de/iwist-cl/projects/sela/>).

10. Bibliography

- Benjamin, M.** 2014. Collaboration in the Production of a Massively Multilingual Lexicon. *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era. LREC, 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, May 26-31, 2014*: 211-215.
- Bergenholtz, H. and R.H. Gouws.** 2010. A New Perspective on the Access Process. *Hermes* 44: 103-127.
- Bergenholtz, I. and H. Bergenholtz.** 2013. One Database, Four Monofunctional Dictionaries. *Hermes — Journal of Language and Communication in Business* 50: 119-125.
- Bosch, S.E., L. Pretorius and J. Jones.** 2007. Towards Machine-readable Lexicons for South African Bantu Languages. *Nordic Journal of African Studies* 16(2): 131-145.
- Bosch, S.E. and G. Faaß.** 2014. Towards an Integrated e-Dictionary Application: The Case of an English to Zulu Dictionary of Possessives. *Proceedings of the 16th Euralex Conference. Proceedings of the Sixteenth EURALEX International Congress, EURALEX 2014, Bolzano/Bozen, Italy, July 15-19, 2014*: 739-747. Bolzano/Bozen: Eurac.
- Bothma, T.J.D. and R.H. Gouws.** 2013. Mapping Indicators in Print Dictionaries to Structural Descriptions in e-Dictionaries. Paper read at the *eLex 2013 Conference, Tallin, Estonia, October 17-19, 2013*. Tallinn, Estonia: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.
- De Schryver, G.-M. and D.J. Prinsloo.** 2000. The Compilation of Electronic Corpora, with Special Reference to the African Languages. *Southern African Linguistics and Applied Language Studies* 18(1-4): 89-106.
- Eiselen, R. and M.J. Puttkammer.** 2014. Developing Text Resources for Ten South African Languages. *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era. LREC, 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, May 26-31, 2014*: 3698-3703.
- Faaß, G., R. Ramagoshi and F. Sebolela.** 2009. Updating the Setswana Monolingual School Dictionary, *Thanodi ya Setswana ya Dikole: An Experimental Study*. Otlogetswe, Thabelo J. (Ed.). 2009. *MLA Kgasa: A Pioneer Setswana Lexicographer*: 146-172. Cape Town: Centre for Advanced Studies of African Society (CASAS), book series 64.
- Faaß, G., S.E. Bosch and E. Taljard.** 2012. Towards a Part-of-Speech Ontology: Encoding Mor-

- phemic Units of Two South African Bantu Languages. *Nordic Journal of African Studies* 21(3): 118-140.
- Faaß, G. and E. Taljard.** 2013. Automatic Detection of Copulatives in Northern Sotho Corpora. (online) *Book of Abstracts on the 5th International Conference on BANTU Languages, Paris, September 2012*.
- Gouws, R.H. and D.J. Prinsloo.** 2005. *Principles and Practice of South African Lexicography*. Stellenbosch: SUN PReSS.
- Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds).** 2013. *Dictionaries. An International Encyclopedia of Lexicography*. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin/New York: De Gruyter.
- Griesel, M. and S.E. Bosch.** 2014. Taking Stock of the African Wordnet Project: 5 Years of Development. Orav, H., C. Fellbaum and P. Vossen (Eds). 2014. *Proceedings of the Seventh Global WordNet Conference (GWC 2014), Tartu, Estonia, 25–29 January 2014*: 148-153. Tartu: GWA. http://gwc2014.ut.ee/proceedings_of_GWC_2014.pdf.
- Heid, U.** 2012. SeLA — A New Project on Electronic Lexicography. *Lexicographica* 28(1): 437-440.
- ISO 24613:2008.** *Language Resource Management*. Edited by International Organization for Standardization, Geneva, 2008.
- Kebbe, J.** 2013. *Creating a Pre-dictionary Fact Base from a Printed Trilingual Dictionary*. Hildesheim: Bachelor project (BA International Information Science), University of Hildesheim, Institute for Information Science and Natural Language Processing (available on request).
- Kosch, I.** 2013. Expectation Levels in Dictionary Consultation and Compilation. *Lexikos* 23: 201-208.
- L'Homme, M.-C.** 2012. Adding Syntactico-semantic Information to Specialized Dictionaries: An Application of the FrameNet Methodology. *Lexicographica* 28: 233-252.
- Prinsloo, D.J., U. Heid, T. Bothma and G. Faaß.** 2012. Devices for Information Presentation in Electronic Dictionaries. *Lexikos* 22: 290-320.
- Prinsloo, D.J.** 2010. Internet Dictionaries for African Languages. *Lexicographica* 26: 183-194.
- RMA.** [accessed: 2014]. Language Resources Management Agency. <http://rma.nwu.ac.za>.
- Scholze-Stubenrecht, W.** 2013. The World Wide Web as a Resource for Lexicography. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds). 2013: 1365-1374.
- Spohr, D.** 2012. *Towards a Multifunctional Lexical Resource: Design and Implementation of a Graph-based Lexicon Model*. *Lexicographica*. Series Maior 141. Supplementary Volume. Berlin/Boston: Walter de Gruyter.
- Wiegand, H.E. and S. Beer.** 2013. Access Structures in Printed Dictionaries. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds). 2013: 110-149.
- Ziervogel, D. and P.C. Mokgokong.** 1985. *Groot Noord-Sotho-Woordeboek: Noord-Sotho, Afrikaans/Engels*. Pretoria: J.L. Van Schaik.