Nordic Journal of African Studies 24(2): 146–168 (2015)

Implementation of a Part-of-Speech Ontology: Morphemic Units of Bantu languages

Elsabé TALJARD
University of Pretoria, South Africa,
Gertrud FAAß
University of South Africa, South Aftica
and
Sonja BOSCH
University of South Africa, South Aftica

ABSTRACT

In a previous article (Faaß et al., 2012), a first attempt was made at documenting and encoding morphemic units of two South African Bantu languages, i.e. Northern Sotho and Zulu, with the aim of describing and storing the morphemic units of these two languages in a single relational database, structured as a hierarchical ontology. As a follow-up, the current article describes the implementation of our part-of-speech ontology. We give a detailed description of the morphemes and categories contained in the database, highlighting the need and reasons for a flexible ontology which will provide for both language specific and general linguistic information. By giving a detailed account of the methodology for the population of the database, we provide linguists from other Bantu languages with a road map for extending the database to also include their languages of specialization.

***Keywords**: part-of-speech ontology, morphemic categories, Bantu morphology, Northern Sotho, Zulu.*

1. INTRODUCTION AND CONTEXTUALIZATION

Computational lexicons containing lists of words for computational processing have been generated since the first days of Natural Language Processing (NLP). These lists are usually labelled with parts of speech (POS) which categorize words according to their distribution in texts. Well-known inventories such as these are the Penn Treebank tagset for English or the Stuttgart Tübingen TagSet (STTS) for German, often used as input for software that annotates words in texts automatically with POS labels (POS-taggers). Usually, words in these two languages are described as (lexical) products of morphological processes and they are used as the starting point for syntactic analyses. Following this perspective or the principle of morphology free syntax “PCFM”, as described for instance by Zwicky (1992:354), morphology and syntax can be viewed as two different levels of linguistic description. For the above mentioned languages, morphological software processors have been developed that formally analyse and/or generate lexical words labelled with their parts-of-speech categorisation. An example is SMOR for German (Schmid et al., 2004) which is fed by lexicons containing thousands of morphemes and their allomorphs together with their categories and (usually sequential) rules on how to merge them into words.

From the perspective of NLP, as mentioned above, the POS categorization of a word is highly dependent on its distribution in the text. The reason is that the software that annotates POS (the "tagger") firstly never has all words possibly occurring in a text in its lexicon, and secondly often finds several POS possible for one word in its lexicon. In order to enable such a tool to identify the correct POS for a specific word, its surrounding words (and/or POS) are taken into account. The more POS are contained in a tagset, the more difficult this decision will be, thus tagsets should ideally only contain an absolute minimum number of POS. Such a constraint, however, contradicts the necessities of syntactic analysis (parsing): A parser needs as much information as possible to correctly assign structural information to a given text. To solve this issue for German, where a parser needs case, number and gender information to correctly assign words to phrases, Schmid and Laws (2008) developed a POS-tagger that works with a hierarchical tagset, where features of tags are described on several levels that are (1) coarse category, (2) finer category, (3) case, (4) number, and (5) gender. A neuter singular definite article ("das") appearing in the nominative hence receives the tag: ART.Def.Nom.Sg.Neut. Schmid and Laws (2008) proved that such a hierarchical tagset together with a tagger calculating the probability of each level independently, allows for a finely grained tagset while achieving high correctness rates. In the same year, Khoury et al. developed a hierarchical tagset for English succeeding in high precision and recall rates for keyword extraction (Khoury et al., 2008).

In traditional Bantu linguistics, a finely grained description of categories was developed for several of the languages of which most are frankly, simply unusable for an automated tagging process (see Taljard et al. (2008)), because distribution issues or NLP in general were not taken into account. Moreover, as in these languages morphological and syntactic processes cannot be separated easily when looking at their orthographies (see the following sections), existing tagsets contain both lexical words and (sub-word) morphemes. NLP processing for languages like Zulu and Northern Sotho is currently well under way as suggested by a number of publications on morpho-syntactic analyses (inter alia Anderson and Kotzé (2006), and Faaß and Prinsloo (2011)), and tagging (inter alia de Schryver and De Pauw (2007)). It was therefore deemed necessary to reconsider the given traditional morpho-syntactic categories from the perspective of NLP to a certain extent. Starting with the assumption that most of the Southern African Bantu languages share a pool of similar if not identical morphemic categories (see Taljard and Bosch, (2006) for a comparison of Zulu and Northern Sotho), a first step towards NLP processing is to create a machine-readable inventory of morphemes.

Taking the works of Khoury et al. (2008) and Schmid et al. (2004) into account, this inventory is to be designed as a hierarchy of categories in order to make a finely grained description of items possible while still aiming at high precision tagging (as well as parsing). We however soon realized that generating this ontology of morphemic items is a project of its own: So far, no full inventories are available to build upon as existing publications rather focus on selected phenomena. The overall project thus consists of four phases: (1) the design of a database that contains all morphemic categories of Zulu and Northern Sotho (Faaß et al., 2012); (2) the population of this database with all known morphemes; (3) the population of the database with data from other, related languages; and finally, (4) the development of morpho-syntactic analyzers/generators utilizing the data collected. Additionally, as this project is a part of the "Scientific e-Lexicography for Africa (SeLA)" collaboration, we aim at connecting the database with another lexicographic database thus enabling interested users of e-dictionaries based on the SeLA database to query detailed information on single morphemes and their categories.

This article describes phase 2, i.e. the database and the manual and semi-automatic procedures utilized to populate it with sufficient data. It also shows the results of some structural design adjustments that were deemed necessary while populating it. Currently, the ontology of morphemic items contains 593 different morphemes with 1,049 assignments to 60 categories of which two are considered productive (noun and verb roots). We will focus on the challenges of a cross-linguistic view on morpheme categories and will describe how the interested user can access and browse our data. Lastly, we contextualize our work within the larger framework of e-lexicography.

2. Categories and Morphemes of the Two Languages

In our ontology, no distinction is made between open and closed classes because such categorizations are more applicable to surface word forms than to morpheme categories. The ontology however contains all morphemes belonging to closed class parts of speech (for example, all of the morphemes and their allomorphs forming subject concords of Northern Sotho). Additionally, a number of roots and stems are contained which are open class parts of speech (e.g. ideophones) or which form the basis of open class parts of speech (e.g. common noun roots). We aim at filling this part of the inventory by using corpus data as part of an ongoing process.

Both languages treated so far distinguish noun class dependent and noun class independent categories. Both of these contain free and bound roots and word formation affixes. Morphological processes usually distinguish compounding (which is not relevant for morpheme categorization and thus not contained in the database), derivation and inflection. The latter two processes are accomplished by adding prefixes and/or suffixes. Figure 1 shows the hierarchy levels 1 to 4 of the current ontology: the highest level is that of morpheme, containing class dependent and class independent items, both of which contain free and bound morphemes. Free as well as bound morphemes can be further categorised as roots, or as derivational or inflectional prefixes and suffixes.

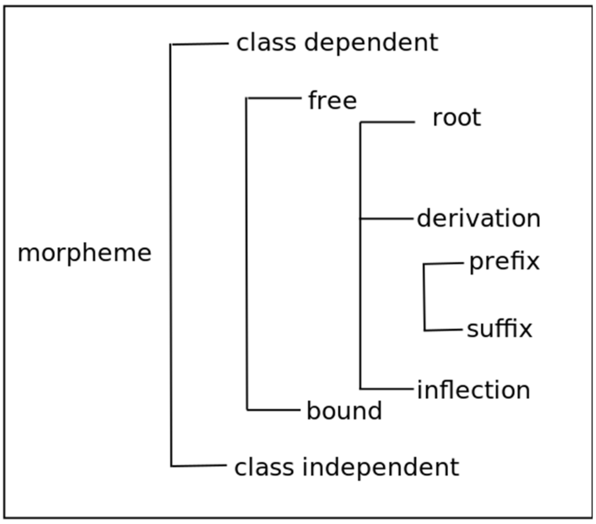


Figure 1. *Levels 1 to 4 of the ontology.*

It should be noted that we do not recognize infixes as a morphemic category in either of the two languages. As Kosch (2006:9) points out, true infixes are hard to find in the Bantu languages and are largely a matter of definition. Possible infixes, such as the passive -*w*- in Northern Sotho which is said to be inserted into the past tense suffix –*ile* (forming –*ilwe*), can easily also be subsumed under the term suffix, as the morpheme "verbal ending" -*e* can be assumed to have been added after the inflectional passive had been formed. We therefore assume that derivational or inflectional affixes are added to incomplete word forms, which do not contain a final morpheme – the final verbal or nominal morpheme is only added after the derivation process has been completed.

For each of the morphological items added to the database, we assign the attributes "person" and "number" wherever appropriate. The attribute "person" distinguishes "first", "second" and "third" (the latter which is represented by the noun classes).

Concerning a naming convention for our morpheme categories, we need to follow the traditional naming conventions of Bantu linguistics so that others will not find it difficult to make use of our system; on the other hand, we deem it necessary to follow the recommendations of the former Expert Advisory Group on Linguistic Engineering Standards (EAGLES[1]), proposing *non-ambiguity*, *compactness*, *readability* and *processability* (Leech and Wilson, 1999:59) of the names we choose.

Levels 5 to 7 of the ontology already describe some language specific morphemic categories; for Northern Sotho and Zulu nominal, verbal, qualificative

[1] http://www.ilc.cnr.it/EAGLES96/browse.html

and a number of other roots, e.g. ideophones are contained. Here, we follow the tradition of not using the term “root” or “stem” for ideophones (IDEO), adverbs (ADV), conjunctions (CONJ), interrogatives (INT), as well as the three categories solely distinguished for Northern Sotho, i.e. hortative and interrogative particles (PART_Hort and PART_Int), and class independent question words (QUE_nil). The bound roots, however, all begin with the name “ROOT” in order to distinguish them from the affixes. Figure 2 shows a complete tree for a common noun root as an exemplification.

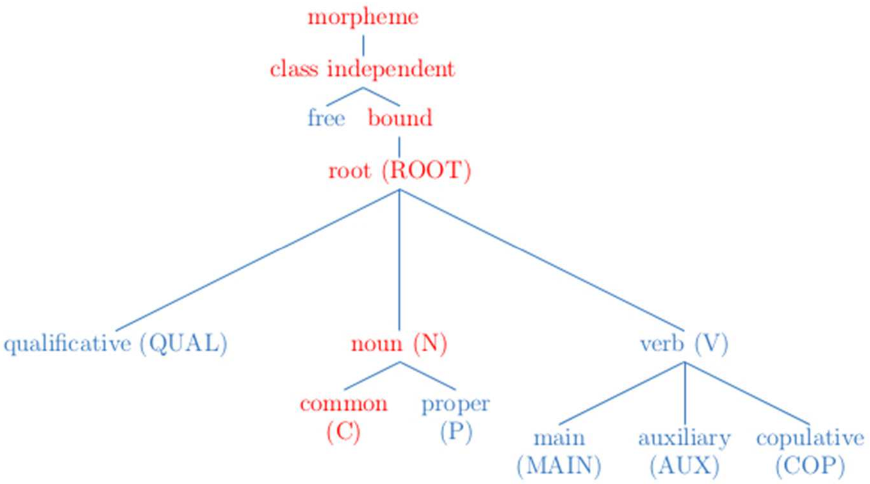


Figure 2. *The path through the ontology to the common noun root.*

2.1 DEVIATION FROM TRADITIONAL GRAMMARS

Any endeavour to describe morphological categories of Northern Sotho and Zulu (or of any language, for that matter) needs to take into consideration that existing grammatical descriptions are embedded in specific linguistic traditions and conventions. These descriptions are often very finely granulated and a lesser degree of granularity is sometimes necessary for a computational application. Some grammatical distinctions, though linguistically sound, are often not relevant for the purpose of our ontology. One such a case is the distinction which is made between copulative particles (respectively prefixes) and subject concords in Northern Sotho, cf. Louwrens (1991). In essence, copulative particles are subject concords which function as fully-fledged copulative verbs, although they appear without any copulative verb stem. Subject concords typically appear as verbal prefixes of verb stems (cf. (2) below), and based on these distributional and syntactic differences, a distinction is drawn between these two categories – subject concords with a copulative function are categorised as particles in grammatical descriptions of Northern Sotho and thus regarded as linguistic words, whereas subject concords are regarded as agreement morphemes. Compare the following examples (for part of speech annotation, the tags described in Taljard et al. (2008) are used):

(1)	*Basadi*	*ba*	*bohlale*	
	N02	PartCop	N014	
	'Women are clever'			
(2)	*Basadi*	*ba*	*bala*	*kuranta*
	N02	CS02	VS	N09
	'Women read the paper'			

Despite these differences, in our ontology we categorize both the so-called copulative particles and the subject concords as class dependent agreement morphemes, partly because the category 'particle' is a rather problematic one when viewed against the background of generally accepted linguistic principles for part-of-speech distinction, and secondly, because what is currently categorized as copulative particles developed from fully-fledged copulative verbs from which the copulative verb stem has been dropped.

Another example of deviation from traditional grammars is the case of the locative noun classes. In Northern Sotho for example, 5 different locative noun classes are distinguished, typically classes 16 (*fa*-), 17 (*go*-), 18 (*mo*-), and two unnumbered classes with prefixes *N*-[2] and *ga*- respectively; in Zulu, the locative classes are 16 (*pha*-), 17 (*ku*-) and 18 (*mu*-). This distinction between the classes is mostly a morphological one, based on the different class prefixes; however, these prefixes do not carry any specific semantic implications in these two languages and furthermore, are no longer productively used to change the meaning of a non-locative noun to a locative one. As a result of this process of semantic bleaching, all classes mainly use one set of concords, namely that of class 17. In our ontology therefore, we do not make provision for separate locative classes, but assign all locative nouns and morphemes related to these classes to the category LOC, as was already suggested by Taljard et al. (2008).

We are furthermore aware of the fact that the so-called absolute or emphatic pronouns in both Northern Sotho and Zulu are not monomorphemic; in most cases they consist of a concordial element, plus a root *-o-* , followed by a suffixal morpheme *–na*, e.g. *s-o-na* (class 7), *b-o-na* (class 2). However, there is no consistency with regard to the morphological decomposition of pronouns in either of these two languages, especially with regard to the first person singular (NSO *nna*, ZUL *mina*) and the second person singular *wena*. It was therefore decided to enter all pronouns as class-dependent, but free roots. A similar approach is followed by Hendrikse and Poulos (2012:260) when they categorise pronouns in Bantu languages as simple, i.e. atomic schematic structures which appear in the lexicon as free morphemes.

In conclusion to the discussion on deviation from traditional grammars, we pay attention to the naming conventions *root* vs *stem* in descriptions of Bantu

[2] We make use of the placeholder *N*- to describe nasals such as *m*- or *n*- appearing as class 9 prefixes.

languages, with particular reference to Zulu and Northern Sotho. Kosch (2006:10–11) summarises the naming convention succinctly by explaining that the possible reason for the indiscriminate use of the term *stem* when *root* is actually meant, is the perpetuation of a descriptive tradition in Bantu language grammars over many years. Doke (1980:286–287), who played a leading role in the earlier linguistic development of Bantu languages in South Africa, states that: “The distinction between roots and stems is more or less arbitrary, and one employed for convenience.”

For the purposes of our ontology we follow a root morphology approach whereby we distinguish between two types of morphemes, namely roots (as free or bound morphemes) and affixes (as prefixes and suffixes). The root is regarded to be the constant core element of a non-compound word in the Bantu languages, carrying the basic meaning, while the remainder of the word or word form represents inflection and derivation (Hurskainen, 1997:631). Therefore, in the case of nouns we may distinguish between noun roots and verb roots (in the case of deverbative nouns), as shown in (3).

Zulu

(3)	a.	*umuntu*			
		umu-	*-ntu*		
		PrefClass_01	ROOT_NC		
		‘person’			
	b.	*isinkwana*			
		isi-	*-nkwa-*	*-ana*	
		PrefClass_07	ROOT_NC	SuffDim	
		‘small bread’			
	c.	*imibono*			
		imi-	*-bon-*	*-o*	
		PrefClass_04	ROOT_VMAIN	SuffNEnd	
			‘see’		
		‘idea’			
	d.	*umfundisi*			
		um-	*-fund-*	*-is*	*-i*
		PrefClass_01	ROOT_VMAIN	SuffExtCause	SuffNEnd
			‘learn, study’		
		‘teacher’			

2.2 DESCRIPTIVE DIFFERENCES BETWEEN LANGUAGES

First, descriptive differences between languages are often the result of adherence to different descriptive frameworks, rather than of fundamental linguistic differences. This is particularly evident when a comparison between Northern Sotho and Zulu is made. Northern Sotho grammatical descriptions mostly adhere to a structuralist framework as advocated by Van Wyk, whereas those of Zulu are mainly done within the functional approach of Doke, cf. Kosch (1993:32). These different theoretical frameworks lead to similar grammatical phenomena being categorised differently in the two languages, although they are fundamentally the same. One such an example is categorization of locative nouns, that is nouns belonging to the so-called locative classes 16, 17 and 18. Within the structuralist approach, these are categorized as belonging to the word class 'noun' in Northern Sotho, based on their morphological structure, i.e. class prefix + root, e.g. NSO *fa-se*, but in terms of their function, Doke classifies them as adverbs in Zulu. Critical consideration is therefore required when items such as these are assigned to the categories in the database: should they be assigned to different categories based on the way in which they are described in the literature, or should some compromise be reached in assigning them to the same category?

Secondly, cases are found of morphemes which appear in both languages, have the same function and are subsumed under the same term, the hortative being a case in point. Hortatives describe a request or a wish, but are morpho-syntactically different in the two languages, possibly as a result of different historical origins. In Zulu, the hortative is clearly an inflectional morpheme, appearing as a verbal prefix; in Northern Sotho however, it is a particle, which is a class independent free morpheme.

For instance, in the two ZUL examples in (4) a polite request and a stronger request or command are expressed by prefixing a hortative morpheme such as *ma-* to the subjunctive form of the verb; or by inserting *-bo-* after the subject concord in the indicative form of the verb:

(4) a. ***Ma**sibuye*

ma-	*-si-*	*-buy-*	*-e*
PrefHort	CS_1p_pl	ROOT_V MAIN	SuffVEnd

'Let us return'

b. *U**bo**hamba kusasa*

u-	*-bo-*	*-hamb-*	*-a*
CS_2p_sg	PrefHort	ROOT_V MAIN	SuffVEnd

'You must go tomorrow'

In Northern Sotho, the hortative appears in constructions similar to the first example, cf. (5 a). On the other hand, the hortative in Northern Sotho can also appear outside of the verbal structure; in fact, separated from the verb by the subject NP, cf. (5 b).

(5) a. ***A** re boe*

A	*re*	*bo-*	*-e*
PrefHort	CS_1p_pl	Root_VMAIN	-SuffVEnd

‘Let us return’

b. ***A** basadi ba tsene*

A	*ba-*	*- sadi*	*ba*	*tsen-*	*-e*
Part_Hort	PrefClass_02	Root_NC	CS_02	Root_VMAIN	-SuffVEnd

‘Let the women enter’

In examples as given in (5), the hortative *a* can clearly not be categorized as an inflectional verbal prefix, and based inter alia on the fact that it can be separated from the verb by means of another linguistic word, it is categorized as a particle – a category which is admittedly a rather idiosyncratic one. Rather than assigning hortatives to two different categories we decided to categorize all Northern Sotho hortatives to the category ‘hortative particle’ since this category would also provide for examples such as those in (5).

Thirdly, there are cases with inherent differences between the two languages. In Zulu, two types of quantitative pronouns are distinguished, i.e. inclusive and exclusive quantitative pronouns. The inclusive quantitative pronoun is characterised by a suffixal morpheme *–nke* which carries the meaning ‘the whole of’ in the singular form and ‘all’ in the plural form (cf. Poulos and Msimang, 1998:124–125), as illustrated in (6).

(6) a. *bo**nke** abantu*

bonke	*aba-*	*-ntu*
PRO_QUANT_02	PrefClass_02	ROOT_NC

‘all the people’

b. *wo**nke** umzimba*

wonke	*um-*	*-zimba*
PRO_QUANT_03	PrefClass_03	ROOT_NC

‘the whole body’

The exclusive quantifier in Zulu is used to express ‘only’ or ‘alone’ and is marked by means of a suffixal morpheme -*dwa*, as in (7).

(7) a. *amadoda o**dwa***

ama-	*-doda*	***odwa***
PrefClass_06	ROOT_NC	PRO_QUANT_06

‘only the men’

b. *umfana uhambe ye**dwa***

um-	*fana*	*u-*	*-hamb-*	*-e*	*ye**dwa***
PrefClass_01	ROOT_NC	CS_01	ROOT_VMAIN	SuffVEnd	PRO_QUANT_01

‘the boy went alone’

In Northern Sotho, only the inclusive category exists and is identified by the quantitative stem *–ôhlê* 'the whole of; all' (Poulos and Louwrens, 1994:79). The category exclusive quantitative pronoun would then be a language specific one which needs to be distinguished to make provision for Zulu, and for that matter for all the languages belonging to the Nguni group.

Moving further afield, although some noun classes (no longer) exist in either Northern Sotho or Zulu, the ontology needs to be flexible enough to provide for other languages in which these classes do appear. Approximately 27 distinct noun classes have been identified in the various languages belonging to the Bantu language family (cf. Poulos and Msimang, 1998:28). Although there is no Bantu language that boasts the whole spectrum of these classes, they need to be provided for in an ontology of Bantu. For instance, classes 20 and 21 do not occur in Zulu and Northern Sotho, but they do occur in at least one of the other South African Bantu languages, namely Venda. Another example is the singular class 11 which is quite common in Zulu, but in the case of Northern Sotho, nouns which originally belonged to class 11, have been reassigned to class 5 (class prefix *le*-). In Tswana, a language closely related to Northern Sotho, class 11 is still distinguished with class prefix *lo*-, cf. (8).

(8)	a.	*loleme*	vs.	*leleme*
		(Tswana)		(Northern Sotho)
		PrefClass_11 + ROOT_NC		PrefClass_05 + ROOT_NC
		'tongue'		
	b.	*losea*	vs.	*lesea*
		(Tswana)		(Northern Sotho)
		PrefClass_11 + ROOT_NC		PrefClass_05 + ROOT_NC
		'baby'		

Another inherent language difference is the so-called relative construction which is language specific in our ontology, since it occurs in Zulu but not in Northern Sotho. The relative construction in Zulu is characterised by a relative morpheme which carries the meaning of 'who/which/that'. Therefore the semantic function of the relative construction is that of qualifying the noun to which it refers, cf. (9).

(9)	a.	*amanzi* ***a****bandayo*					
		ama-	*-nzi*	***a-***	*-band-*	*-a-*	*-yo*
		PrefClass_06	ROOT_NC	CRel_06	ROOT_VMAIN	SuffVEnd	SuffRel
		'water **that** is cold'					
	b.	*umfundi* ***o****hlakaniphile*					
		um-	*- fundi*	***o-***	*-hlakaniph-*	*-ile*	
		PrefClass_01	ROOT_NC	CRel_01	ROOT_VMAIN	SuffVEnd	
		'a scholar **who** is intelligent'					

It is significant that relative constructions may be formed from a variety of word categories, e.g. the copula construction, the verb, adverbial forms, possessive constructions as well as pronouns.

2.3 LANGUAGE INTERNAL DIFFERENCES IN GRAMMATICAL DESCRIPTIONS

In 2.1 it was pointed out that differences in the descriptive frameworks can in some cases account for perceived differences between languages. Differences in grammatical descriptions within the same language can also be the result of different theoretical stances taken by linguists. We will only refer to two examples of such differences in descriptions, the first being the categorisation of radical pronouns vs enumeratives in Northern Sotho. Lombard (1985), following a structuralist approach, categorizes the items *-tee* 'one', *-šele* 'strange', *-fe* 'which' and *-šoro* 'cruel' as radical pronouns, the term 'radical' in this case referring to its etymology, i.e. radix, meaning root. However, in the more functional approach taken by Poulos and Louwrens (1994) enumeratives are, together with adjectives and nominal relatives regarded as qualificatives, since their primary function is to qualify or describe a nominal antecedent. Their pronominal function is a secondary one, which is only fulfilled when the nominal antecedents of these enumeratives are deleted. In keeping with the functional approach, in our ontology we categorize enumeratives as qualificative roots.

A second example is the numbering system of the noun classes in Bantu. For ease of reference it has become customary in the description of the Bantu languages to assign numbers to the different noun class prefixes. In our ontology, we follow Meinhof's (1932:48ff) numbering system of the noun class prefixes. In general, the uneven class prefixes represent singular prefixes with the following even number representing the corresponding plural prefix. There are, however, some exceptions such as the prefixes of class 15, as well as those of the locative classes that are not associated with any grammatical number. It should be noted that there are other approaches to noun class categorisation as well, for example Doke's numbering system that divides classes into singular and plural pairs, sometimes referred to as noun class genders with a single number for each pair.

2.4 MORPHOSYNTACTIC PHENOMENA NOT DESCRIBED PREVIOUSLY

Although both Northern Sotho and Zulu are to a large extent standardized in the sense that their grammars have been extensively described and investigated, some linguistic phenomena seem to have slipped through the cracks of grammatical description. In Northern Sotho for example, a whole paradigm of concordially based forms are not mentioned in any grammar, nor have they been the object of any form of linguistic investigation, despite their high frequency of use. They consist of a prefixal element *na-* which is affixed to a shortened form of the so-called emphatic pronoun and carries the meaning '(together) with him/her/it/them', e.g. *nabo* 'with them' (class 2), *naye* 'with him/her/it' (class 1) and *natšo* 'with them' (class 8/10). These forms do not readily fit into any of the existing word classes of Northern Sotho, although it could be argued that function-wise, they could be categorized as adverbs. Adverbs however, normally do not contain a concordial element. In cases such as these, a decision needs to be taken by the expert linguists as to the correct categorization of these forms – in this case, it was decided to add the whole paradigm to the category adverb, based on the function of these forms.

3. Design of the Data Model, the Contents and Front End

When designing a database (DB), especially a relational one like the MySQL-DB shown in Figure 3, one should distinguish between data items and relations between those items. Morphological categories like part of speech, number or noun class are viewed as data items, as are the morphemes themselves. Each of those data items is stored only once in the database. By assigning one morpheme to a language and to a part of speech (and possibly, to a noun class and a number), we create a relation which is then stored in a relational table (see "morph_assign" in Figure 3). The lines connecting the tables in Figure 3 describe the numeral relation between entries in the tables: 1:n means that each morpheme appearing once for instance in the table "morphemes" may appear several times in the table "morph_assign". Diamonds stand for primary (unique) keys, filled circles mean that for each entry, this field has to be filled, and empty circles mean that filling the field is optional.

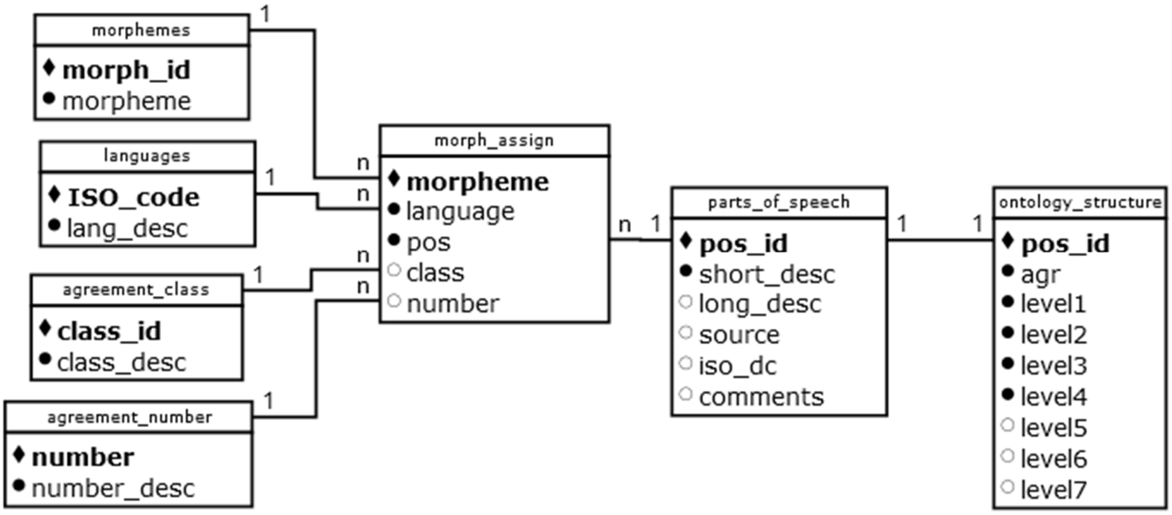


Figure 3. *Database design.*

In our database, one table, namely the table "morphemes" contains all of the morphemes and allomorphs[3]; the field "morpheme" is defined as unique which means that each morpheme may only appear once in the table. The table "languages" is generated to contain all languages to be described. There is also a table "parts-of-speech" describing categories with additional information like short and long description and a respective ISOCAT[4] data category, if available (cf. Pretorius and Bosch, 2014). The table "ontology-structure" represents the levels described above for each of the morpheme categories and is connected 1:1 with the table "parts_of_speech" to make sure that no morpheme category will be added to the database that is not sorted into the hierarchy first. There are tables containing all known data items of "class" (including 1st and 2nd person), and "number", too. The relational table morph_assign lastly contains assignments of the morphemes each to a part-of-speech category, a language and, if appropriate, a class and number information. A class dependent item can thus be identified via a POS and a class and/or a number, e.g. *ba* being a subject concord (CS) of class 2 for the language Northern Sotho (no number assigned) or *ke* being a subject concord (CS) of the 1st person singular for the same language, see Figure 4a.

[3] We distinguish allomorphs (like *tlo* and *tla* both being future tense morphemes of Northern Sotho; and *umu-* and *um-* both being class 1 noun prefixes of Zulu) and the results of morpho-phonological changes at word boundaries when merging morphemes. Results of the latter are not contained in the database.

[4] http://www.isocat.org/rest/dcs/119#index

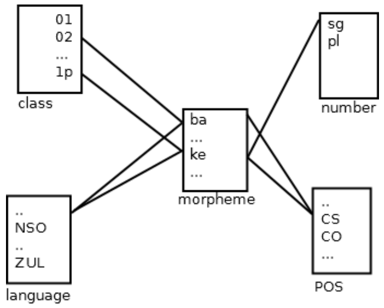


Figure 4a. *Graphical view of assignments done for **ke** and **ba** in the table morph_assign.*

3.1 Populating the Database

There are several ways to populate the database; we opted for adding most of the data from excel-sheets that had been filled in by language experts. These experts also had the option of using the editing function of the database to correct, delete or add items. For adding further languages, such as Tswana, it is possible to generate an excerpt of the database containing all of the Northern Sotho assignments, e.g. in excel sheets, in .csv or similar format. A language expert can then change the data with an external program before these are added collectively to the database by the database administrator. New single assignments can be done by language experts with editing rights via the interface.

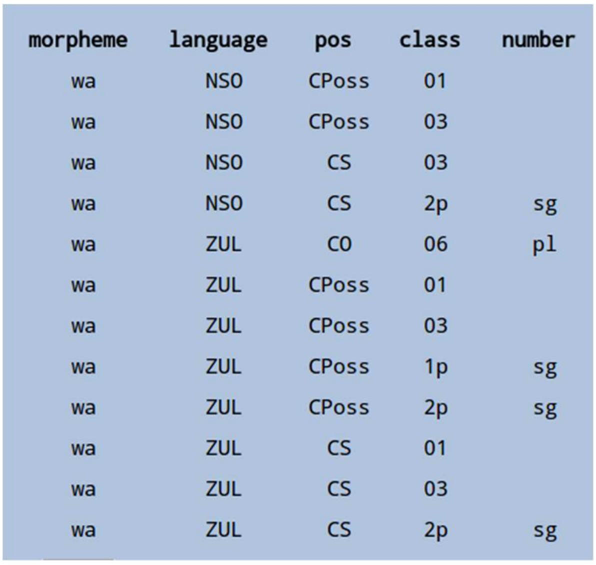
Concerning the front end, we distinguish three kinds of users: (1) the browsing user may examine the data contained, but may not make any changes; (2) the maintainer can add, delete or edit single morphemes and/or morpheme assignments, and (3) the database administrator, who has all the rights above and may do all else, such as adding external data from lists (or excel sheets) provided by language experts or deleting data that exhibit certain properties, e.g. morphemes that are not assigned to any category.

3.2 Contents of the Current Database

So far, 60 different morpheme categories are fully described and sorted into the hierarchy. A total of 1,049 assignments (595 for Zulu and 454 for Northern Sotho) for 593 different morphemes ("types") have been stored. In Northern Sotho, *go* shows the highest number of assignments, namely 13, followed by *a* and *ba*, each with 12. In Zulu, *ku* and *ba* each are assigned to 11 morpheme categories, followed by *zi* with 9 assignments. We find altogether 34 morphemes which are assigned to both languages, most of them to several categories, see Table 1 for an excerpt. Figure 4b demonstrates the current assignments for the morpheme *wa*.

Table 1. *An excerpt of morphemes assigned to both Zulu and Northern Sotho.*

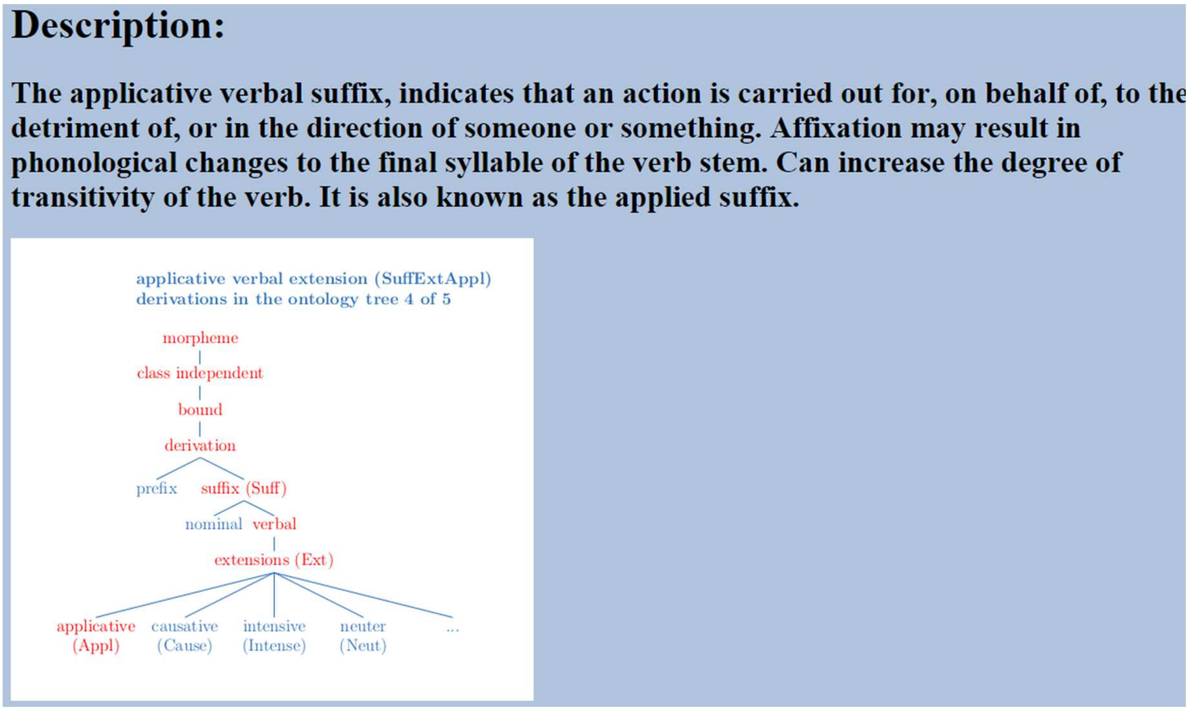
Morpheme	***Assignments to ZUL***	***Assignments to NSO***	***Total no. of assignments***
a	9	12	21
ba	11	12	23
be	3	2	5
bona	2	2	4
ma	4	1	5
lona	2	2	4
wa	8	4	12
wena	1	1	2
wona	4	2	6
yena	2	1	3
yona	2	4	6



morpheme	language	pos	class	number
wa	NSO	CPoss	01	
wa	NSO	CPoss	03	
wa	NSO	CS	03	
wa	NSO	CS	2p	sg
wa	ZUL	CO	06	pl
wa	ZUL	CPoss	01	
wa	ZUL	CPoss	03	
wa	ZUL	CPoss	1p	sg
wa	ZUL	CPoss	2p	sg
wa	ZUL	CS	01	
wa	ZUL	CS	03	
wa	ZUL	CS	2p	sg

Figure 4b. *Table view of current assignments for the morpheme* ***wa*** *as shown by the database.*

All morpheme categories contained in the database have been provided with a description in the form of a terminological definition. The aim of these definitions is to enable linguists who would want to enter the data of any other language to identify the proper category where their data should be entered. This is necessary since as we have pointed out in section 2.2 above, differences in the linguistic descriptions of languages may obscure similarities between languages. Although not all definitions have been done within the classical genus-differentiae convention of terminological definitions, we have attempted to refer to the relevant superordinate concepts of the term being defined whenever possible and / or applicable. By doing so, we have tried to give some indication of the conceptual relationship between the concept being defined and other related concepts. In this way, we give some indication of the position of the morpheme in the diagram. The definition of the applicative verbal extension (SuffExtAppl) and its position in the relevant diagram are given as an example in Figure 5.



Description:

The applicative verbal suffix, indicates that an action is carried out for, on behalf of, to the detriment of, or in the direction of someone or something. Affixation may result in phonological changes to the final syllable of the verb stem. Can increase the degree of transitivity of the verb. It is also known as the applied suffix.

Figure 5. *Long description and ontology tree for SuffExtAppl*.

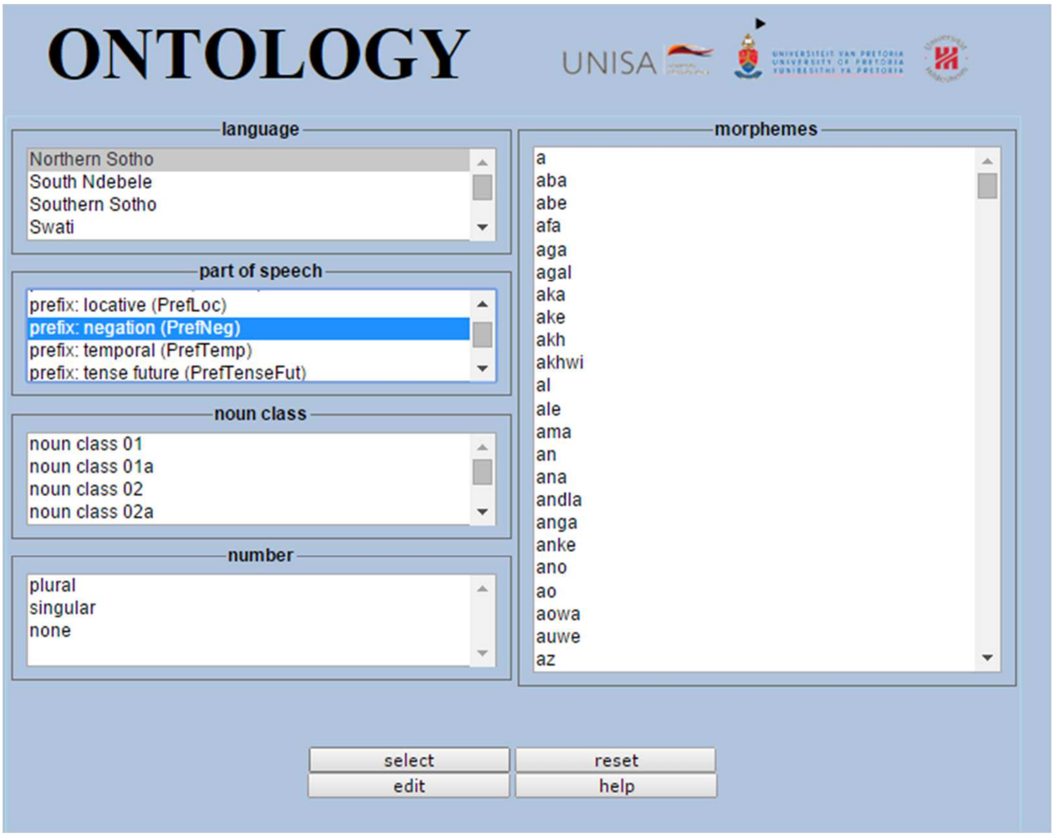
Our long descriptions are furthermore an attempt to provide for both language specific and general linguistic information. In cases where we are aware of different interpretations of specific linguistic phenomena, a reference to such viewpoints is already made in the description. Compare in this regard our definition of the potential morpheme, which we regard as an aspectual morpheme, but which is described as a marker of a potential mood in some grammars:

> "The potential morpheme expresses aspectual notions such as possibility, permission and condition. In some grammatical descriptions, the potential is regarded as a separate mood or verb form."

3.3 Browsing Data (Front End)

The current front end of our database is still a beta version waiting for its evaluation by language experts who will extend it to other languages. Figure 6 shows the starting screen[5] of the browsing mode (we will describe the edit mode after it has been evaluated). The user may click on the elements that (s)he is interested in. It is possible to double click on a morpheme to obtain information about the assignments for this morpheme (the result of such an action for the morpheme *wa* is displayed in Figure 4b) or to generate a selection by clicking on a language, a morpheme category, and/or on class and/or a number, followed by clicking on the "select" button. As Figure 7 exemplifies, only the morphemes of the selected categories are then displayed.

[5] https://www.uni-hildesheim.de/iwist-cl/projects/ontology/



ONTOLOGY

language
Northern Sotho
South Ndebele
Southern Sotho
Swati

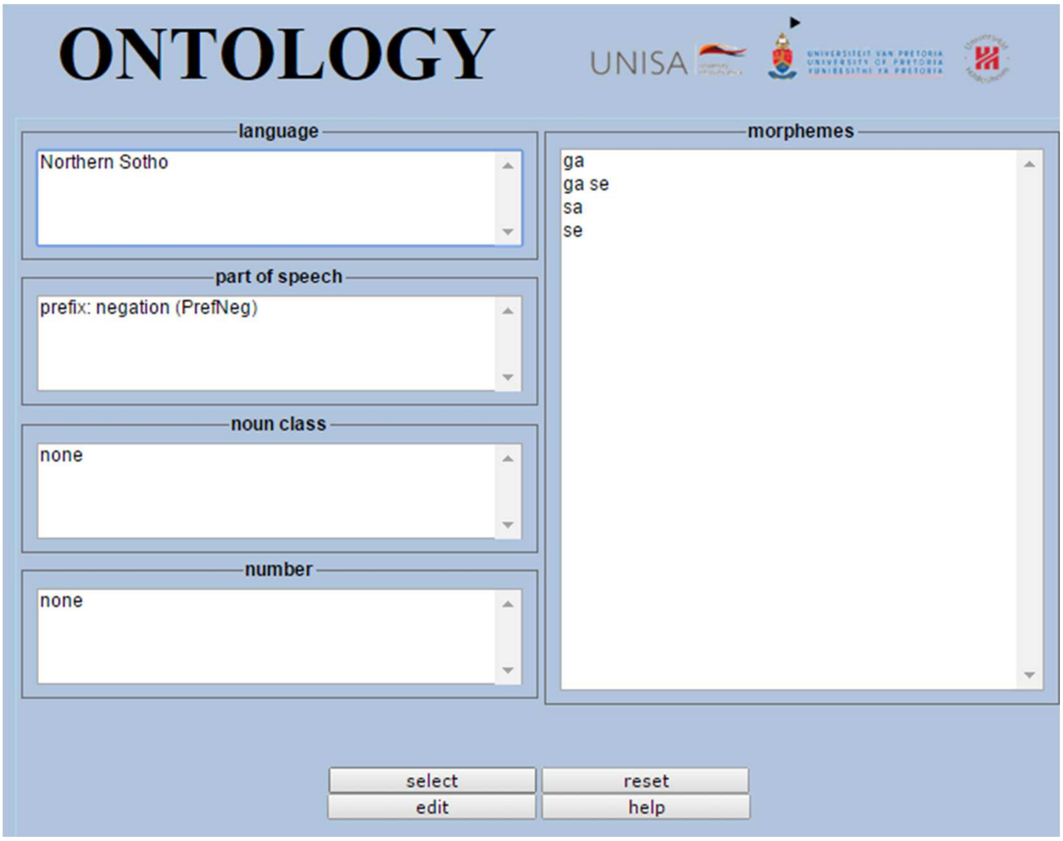
part of speech
prefix: locative (PrefLoc)
prefix: negation (PrefNeg)
prefix: temporal (PrefTemp)
prefix: tense future (PrefTenseFut)

noun class
noun class 01
noun class 01a
noun class 02
noun class 02a

number
plural
singular
none

morphemes
a
aba
abe
afa
aga
agal
aka
ake
akh
akhwi
al
ale
ama
an
ana
andla
anga
anke
ano
ao
aowa
auwe
az

Figure 6. *Selecting "Northern Sotho" and "prefix: negation" on the start screen.*



ONTOLOGY

language
Northern Sotho

part of speech
prefix: negation (PrefNeg)

noun class
none

number
none

morphemes
ga
ga se
sa
se

Figure 7. *Resulting screen after clicking on "select".*

If a user needs more information about a part of speech category, (s)he may double click on the respective part of speech. This action will lead to a pop-up window giving the long description of the data item and the path to it through the ontology, as shown in Figure 5 above.

After login (clicking on the "edit" button leads to a login screen), a language expert may add or change morpheme assignments to the parts of speech described in the database. In case a new morpheme is assigned, it is added to the database automatically; if the last assignment for a morpheme is deleted, the morpheme itself is deleted in the database too. Language experts may not change the existing hierarchy. This is to ensure that all categories are discussed with all language experts involved before they are changed, added or deleted. Such activity may only be performed by the administrator who makes sure that all experts agree.

3.4 Extension to Other Languages (Roadmap)

The preceding detailed account of the methodology for the population of the database serves as roadmap for linguists specializing in other Bantu languages for extension of the database. It was shown that the database is extendable via the web so that data from Bantu languages in various degrees of relatedness can be added.

A case in point is the relative suffix. In Zulu as well as in Northern Sotho, the relative suffix is an invariable, class independent morpheme as shown in (10) and (11).

Zulu

(10) a. *umfana ofunda**yo***

um-	*-fana*	*o-*	*-fund-*	*-a-*	***-yo***
PrefClass_01	ROOT_NC	CRel_01	ROOT_VMAIN	SuffVEnd	SuffRel

'the boy who studies'

b. *abafana abafunda**yo***

aba-	*-fana*	*aba-*	*-fund-*	*-a-*	***-yo***
PrefClass_02	ROOT_NC	CRel_02	ROOT_VMAIN	SuffVEnd	SuffRel

'the boys who study'

Northern Sotho

(11) a. *mošemane yo a kitima**go***

mo-	*-šemane*	*yo*	*a-*	*-kitim-*	*-a-*	***-go***
PrefClass_01	ROOT_NC	CDem_01	CS_01	ROOT_VMAIN	SuffVEnd	SuffRel

'the boy who is running'

b. *bašemane ba ba kitima**go***

ba-	*-šemane*	*ba*	*ba-*	*-kitim-*	*-a-*	***-go***
PrefClass_02	ROOT_NC	CDem_02	CS_02	ROOT_VMAIN	SuffVEnd	SuffRel

'the boys who are running'

However, in some Bantu languages such as Swahili, the relative suffix morpheme is actually class dependent, which implies that additional provision would have to be made for such a morpheme within the ontology as shown in (12). This further motivates our decision to implement a flexible ontology.

Swahili

(12) a. *mntu asoma**ye***

m-	*-ntu*	*a-*	*-som-*	*-a-*	***-ye***
PrefClass_01	ROOT_NC	CS_01	ROOT_VMAIN	SuffVEnd	SuffRel_01

‘a man who reads’

b. *kengele ilia**yo***

Ø	*kengele*	*i-*	*-li-*	*-a-*	***-yo***
PrefClass_09	ROOT_NC	CS_09	ROOT_VMAIN	SuffVEnd	SuffRel_09

‘a bell which rings’

c. *kitu kiangukа**cho***

ki-	*-tu*	*ki-*	*-anguk-*	*-a-*	***-cho***
PrefClass_07	ROOT_NC	CS_07	ROOT_VMAIN	SuffVEnd	SuffRel_07

'a thing which falls down'

In these examples it is illustrated that co-variance or agreement exists between the relative suffix morpheme appearing in bold and the antecedent. Although in Swahili this bold-printed form is regarded as an absolute pronoun, it occurs in the same position and has the same function as the described relative suffix morpheme in the other languages that have been exemplified, the difference being that it agrees in class with the antecedent.

In Tsonga, the vowel of the relative suffix has become phonologically identical to the final vowel of the verb stem (cf. Poulos, 1986:287–288), and is therefore no longer an invariable morpheme, see (13).

Tsonga

(13) a. *vanhu lava va vulavula**ka***

va-	*-nhu*	*lava*	*va*	*vulavul-*	*-a-*	***-ka***
PrefClass_02	ROOT_NC	PRO_DEM_02	CS_02	ROOT_VMAIN	SuffVEnd	SuffRel

‘people who are talking’

b. *vanhu lava va nga vulavuli**ki***

va-	*-nhu*	*lava*	*va*	*nga*	*vulavul-*	*-i-*	***-ki***
PrefClass_02	ROOT_NC	PRO_DEM_02	CS_02	PrefNeg	ROOT_VMAIN	SuffVEnd	SuffRel

‘people who are not talking”

For cases described in (11) to (13), new categories will therefore have to be added to the ontology. The language expert should hence contact the database administrator suggesting the category and its position in the ontology. The administrator will then contact the language experts involved to find a consensual solution.

4. Conclusion and Future Work

The ontology of morphological items can be used by linguists as a separate knowledge base. However, we plan to also connect it with the lexicographic database currently developed in the SeLA project (Faaß et al., 2014) since we assume that users not familiar with the Bantu languages, might enter fully fledged orthographic words when seeking lexicographic information. A morphological analyser is planned which will split the given orthographic word (if necessary) into lexicographic units (e.g. stems) for which lexicographic information can subsequently be found in the SeLA database. Should a user enter a grammatical morpheme which is not contained in the SeLA database, the ontology database will be able to provide a description. The morphological analyser will hence make use of the ontology database so that information on morphological units not provided with lexical information can also be displayed to the user.

As described above, we are now aiming at an extension of the ontology database towards other Bantu languages and collaboration with relevant language experts will be sought.

Acknowledgement

This paper 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, see also http://www.uni-hildesheim.de/iwist-cl/projects/sela/).

The authors would like to acknowledge the assistance rendered by Ronny Jauch with regard to the implementation of the front end of the part-of-speech ontology, as well as Dr. Beata Wójtowicz for the Swahili examples.

REFERENCES

Anderson, W.N. and Kotzé, P.M. 2006.
Finite state tokenisation of an orthographical disjunctive agglutinative language: The verbal segment of Northern Sotho. In: *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 22–28 May, 2006, pp.1906–1911. Available: http://www.lrec-conf.org/proceedings/lrec2006/

De Schryver, G-M and De Pauw, G. 2007.
Dictionary Writing System (DWS) + Corpus Query Package (CQP): The case of TshwaneLex. **Lexikos** 17 (AFRILEX-reeks/series 17): 226–246.

Doke, C.M. 1980.
Textbook of Zulu grammar. **Cape Town: Maskew Miller Longman**.

Faaß, G. and Prinsloo, D.J. 2011.
A computational implementation of the Northern Sotho Infinitive. **South African Journal of African Languages** 31(2): 281–301.

Faaß, G., Bosch, S. and Taljard, E. 2012.
Towards a Part-of-Speech Ontology: Encoding Morphemic Units of Two South African Bantu languages. **Nordic Journal of African Studies** 21(3):118–140. http://www.njas.helsinki.fi/

Faaß, G., Bosch, S. and Gouws, R.H. 2014.
A General Lexicographic Model for a Typological Variety of Dictionaries in African Languages. **Lexikos** 24 (AFRILEX-reeks/series 24:2014): 94–115.
http://lexikos.journals.ac.za/pub/article/view/1254/766

Hendrikse, A.P. and Poulos, G. 2012.
Tagging an agglutinating language: A new look at word categories in the Southern African indigenous languages. **Language Matters: Studies in the Languages of Africa** 37(2): 246–266.

Hurskainen, A. 1997.
A language sensitive approach to information management and retrieval: the case of Swahili. In: Herbert, RK (ed), *African Linguistics at the Crossroads: Papers from Kwaluseni*, pp. 629–642. Cologne: Rüdiger Köppe.

Khoury, R., Karray, F. and Kamel, M. 2008.
Keyword extraction rules based on a part-of-speech hierarchy. **International Journal of Advanced Media and Communication** 2(2): 138–153.

Kosch, I.M. 1993.
A historical perspective on Northern Sotho linguistics. Via Afrika Monograph series 5. Pretoria:Via Afrika.

2006. Topics in Morphology in the African Language Context. Pretoria: Unisa Press.

Leech, G. and Wilson, D. 1999.
Standards for Tagsets. In: van Halteren, H (ed.), *Syntactic Wordclass Tagging*, pp. 55 – 80. Dordrecht: Kluwer Academic.

Lombard, D. P. 1985.
Introduction to the Grammar of Northern Sotho. Pretoria: J. L. van Schaik.

Meinhof, C. 1932.
Introduction to the phonology of the Bantu languages. Berlin: Dietrich Reimer/Ernst Vohsen.

Poulos, G. 1986.
Instances of semantic bleaching in South-Eastern Bantu. In: Dimmendaal, G.J. (ed.), *Current approaches to African linguistics.* Vol. 3. Dordrecht: Foris Publications.

Poulos, G. and Louwrens, L. 1994.
A linguistic analysis of Northern Sotho. Pretoria: Via Afrika.

Poulos, G. and Msimang, C.T. 1998.
A linguistic analysis of Zulu. Pretoria: Via Afrika.

Pretorius, L. and Bosch, S. 2014.
Towards Extending the ISOcat Data Category Registry with Zulu Morphosyntax. In: *Proceedings of the 9th edition of the International Conference on Language Resources and Evaluation (LREC)*: 39–43. Reykjavik, Iceland. Available:
http://www.lrecconf.org/proceedings/lrec2014/index.html

Schmid, H. and Laws, F. 2008.
Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging, COLING 2008, Manchester, Great Britain.

Schmid, H., Fitschen, A. and Heid, U. 2004.
SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In: *Proceedings of the 4th edition of the International Conference on Language Resources and Evaluation (LREC 2004)*, pp.1263–1266, Lisbon, Portugal.

Taljard, E., Faaß, G., Heid, U. and Prinsloo, D.J. 2008.
On the development of a tagset for Northern Sotho with special reference to the issue of standardization. **Literator** – *special edition on Human Language Technologies*, 29(1): 111–137.

Taljard, E. and Bosch, S. E. 2006.
A Comparison of Approaches to Word Class Tagging: Disjunctively vs. Conjunctively Written Bantu Languages. **Nordic Journal of African Studies** 15(4): 428–442.
http://www.njas.helsinki.fi/

Zwicky, A. M. 1992.
Some choices in the theory of morphology. In: Levine, Robert (ed.), *Formal grammar: Theory and Implementation*, pp. 327–371. Oxford: Oxford University Press.

About the authors: *Elsabé Taljard* is professor in the Department of African Languages at the University of Pretoria, Republic of South Africa. Her language of specialization is Northern Sotho (also known as Sepedi or Sesotho sa Leboa). Her fields of interest include corpus linguistics, terminology and computational linguistics.

Gertrud Faaß is a visiting researcher in the Department of African Languages University of South Africa (UNISA). Her research concerns inter alia morphology, syntax, corpus linguistics, and computer assisted language learning.

Sonja Bosch is professor in the Department of African Languages at the University of South Africa (UNISA). Her main field of interest is natural language processing of the Nguni language family, with specialization in morphological analysis.