Nordic Journal of African Studies 21(3): 118–140 (2012)

Towards a Part-of-Speech Ontology: Encoding Morphemic Units of Two South African Bantu Languages

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

ABSTRACT

This article describes the design of an electronic knowledge base, namely a *morpho-syntactic database* structured as an ontology of linguistic categories, containing linguistic units of two related languages of the South African Bantu group: Northern Sotho and Zulu. These languages differ significantly in their surface orthographies, but are very similar on the lexical and sub-lexical levels. It is therefore our goal to describe the morphemes of these languages in a single common database in order to outline and interpret commonalities and differences in more detail. Moreover, the relational database which is developed defines the underlying morphemic units (morphs) for both languages. It will be shown that the electronic part-of-speech ontology goes hand in hand with part-of-speech tagsets that label morphemic units. This database is designed as part of a forthcoming system providing lexicographic and linguistic knowledge on the official South African Bantu languages.

Keywords: *part-of-speech ontology, morpho-syntactic database, tagging, Northern Sotho, Zulu.*

1. INTRODUCTION

The aim of this article is to describe the design of an electronic knowledge base, namely a *morpho-syntactic database* structured as an ontology of linguistic categories, containing linguistic units of two Bantu languages. It will be argued that the electronic part-of-speech (POS) ontology goes hand in hand with POS tagsets. For the purpose of this discussion, the term POS tagging will be extended to also include tagging of morphemic units, the reason of which will become clear in section 3. This database is designed as part of a forthcoming system providing lexicographic and linguistic knowledge on the official Bantu languages of South Africa. It is also to be used as a knowledge base for tools that annotate POS in written text.

Northern Sotho and Zulu both belong to the Bantu language family, more specifically the South-Eastern zone of Bantu languages. These two languages are also part of a larger grouping of languages namely the Sotho and Nguni language groups respectively. In general, Bantu languages are characterised by two basic morphological systems, namely the noun class system, and the resulting system of concordial agreement: “The noun class system classifies nouns into a number of noun classes, as signalled by prefixal morphemes also known as noun prefixes. These noun prefixes have, for ease of analysis, been divided into classes with numbers by historical Bantu linguists and represent an internationally accepted numbering system.” (Taljard and Bosch, 2006: 429). Noun prefixes contribute significantly to the morphological structure of the Bantu languages in that they link the noun to other words in the sentence. This linking is expressed by a system of concordial agreement, which governs grammatical agreement in verbs, adjectives, possessives, pronouns and so forth.

POS tagging is usually performed on electronic texts with the aim of facilitating linguistic research and for preparing such texts for further computational linguistic processing, e.g. parsing. For disjunctively written languages such as Northern Sotho, POS tagging has hitherto been done on the orthographic (graphemic) level, but as a result of the conjunctive writing system of Zulu, such tagging has not been possible. Therefore, tagging texts of both languages with one knowledge source (the designed database) requires working on a deeper level: the level of underlying morphemes. Surface text will therefore be split into morphemes first by making use of morphological analyzers similar to the prototypes developed by Pretorius and Bosch (2003) for Zulu, and Anderson and Kotzé (2008) for Northern Sotho, after which POS tagging will follow.

In aiming at high precision tagging of morphemes contained in texts of both languages, we however face two major challenges:

- The richness of the morphology of both languages constitutes the first challenge. When utilizing automated ways of annotating or adding linguistic labels to morphemes by means of heuristic POS taggers, the number of types used should be kept small to avoid sparse data problems. Distinguishing more types implies having fewer occurrences of each type. Since heuristic taggers usually assume the most frequent type for an unknown item given a specific environment, sparse data hence leads to poor tagging quality.
- Furthermore, a high degree of morphemic homography, especially with regard to the disjunctively written Sotho languages, leads in turn to high levels of type ambiguity resulting in a second challenge, namely that of disambiguation of closed class items as described in Faaß et al. (2009).

In order to address both of these challenges, the EAGLES (1996) 'Recommendations for the Morphosyntactic Annotation of Corpora'[1] offer a possible solution. EAGLES differentiate between mandatory, optional and recommended labels. These recommendations suggest different levels of annotation and each level can be tagged separately, either with different tools or with taggers that calculate the probability of each label for each of the levels separately. Heid (2000: 679, referring to Leech) also describes such a hierarchical system as being better suited for natural language processing (NLP) and for mnemonical purposes. Taggers making use of this strategy are available, for instance the RF Tagger developed by Schmid and Laws (2008) which has been used successfully on Northern Sotho texts (Faaß et al., 2009).

Any solution, however, first requires a proper description of each category on every level, and the levels themselves must be defined on a sound linguistic basis. Our work aims at describing morphemic categories on such levels of annotation for two of the Bantu languages, Northern Sotho and Zulu, and at storing them electronically in a knowledge base accessible for the development of taggers and other applications, e.g. electronic dictionaries. At a later stage, we plan to add the data for more of the official Bantu languages of South Africa.

As we would not want to focus merely on POS tagging of two languages with two different writing systems, the database we aim at is intended to be a valuable opportunity for cross-linguistic research as well. In the near future, we would like to be able to answer the following questions: Can we assume that the same morphemic categories exist in these languages, and that the apparent differences are merely the result of different traditions of linguistic description? Are there categories that are language specific and thus only appear in one of the languages?

The following section describes related work and past and present developments; in section 3 we develop the ontology of morphemic units for Northern Sotho and Zulu as representatives of the Sotho and Nguni language groups, and describe the current state of implementation, while section 4 concludes the article and points to future work. An appendix lists a combined tagset for Northern Sotho and Zulu.

2. CONTEXTUALIZATION

Kahrel et al. (1997) aim at tagging English texts and, referring to the EAGLES (1996) Guidelines for Annotation, illustrate that POS tags can indeed be organized on different levels, namely the obligatory and the recommended level. In summary, each level describes a whole set of parts of speech for lexical items, but in different degrees of granularity. Following this approach, different levels of linguistic representation can be implemented, as shown in Table 1: the first,

[1] Available online http://www.ilc.cnr.it/EAGLES/annotate/annotate.html.

coarse level of POS labels (obligatory) may be sufficient for certain research tasks, while the second, more finely grained level (recommended), describes subtypes carrying detailed information about the word forms it refers to or represents.

Level	***1***	***2***
Category	**noun:** *tree, trees*	**noun.sg:***tree* **noun.pl:***trees*
	verb: *grow, grows, grew, grown*	**verb.sg.1st pers.pres tense:***grow* **verb.sg.2nd pers.pres tense:***grow* **verb.sg.3rd pers.pres tense:***grows* **verb.sg.past tense :***grew* **verb.pastparticiple :***grown*

Table 1. *A simple example of different levels of parts of speech definition.*

Kahrel et al. (1997) did not primarily aim at designing or compiling a database describing the necessary items; the work rather focussed on the tags to be split into several levels aiming at a high level of precision. Actually, to our knowledge, most of the computational linguistic research done on the Bantu languages has been focussing on tagset design and/or the development and evaluation of automatic POS tagging tools of linguistic or graphemic words, and their morphological analysis. For instance, a general tagset for use in an automatic word-class tagger was developed for Tswana (Van Rooy and Pretorius, 2003), a language which is closely related to Northern Sotho. The design of this tagset is done in accordance with the EAGLES (1996) standards, and functions strictly at the level of the linguistic word. It therefore makes no provision for morphological information contained within linguistic words. Although the authors indicate that the tagset is to be used for automatic word-class tagging, an implementation has thus far not been done. Allwood et al. (2003) propose a tagset to be used on a corpus of spoken Xhosa, a conjunctively written Bantu language. The latter draft morpho-syntactic tagset for Xhosa is revised and amended by Hendrikse and Mfusi (2008) within a Construction Grammar approach. However, in this paper, we argue for a generic categorization system of parts of speech from which words can be built at a later stage, while Hendrikse and Mfusi (2008) go a step further and describe a grammar that builds words from such categories. In other words, our ontology does not structure words, but morpheme categories. We will come back to their work in our forthcoming article on the implementation of a morpho-syntactic generation process based on the ontology described here.

Taljard et al. (2008) present the development of a multilevel tagset for Northern Sotho in order to account for the morphological complexity of the language. The first level of annotation contains obligatory information such as class membership and the specification of the feature person (1st and 2nd), while the second level contains optional and recommended information such as further specification of the first level features, e.g. MORPH_fut and MORPH_prog for future tense and progressive morphemes respectively, where the underscore serves to distinguish between the different levels. We view this tagset as an organic one, which can be adapted and extended according to specific needs and further applications.

Concerning electronic databases containing morphological items, we are only aware of works in lexicography. Bosch et al. (2007) initiated the development of a machine-readable (multilingual) lexicon for the South African Bantu languages, representing fully inflected forms on the basis of word stems. The aim of the data model is to ensure maximum inclusiveness of linguistic information while providing flexibility and dealing with the diverse representations relevant specifically to Bantu languages, thereby making it applicable to wide-ranging uses of machine-readable lexicons (Bosch et al., 2007: 143).

In this project phase we describe the design of the future database: Beginning with the POS inventory developed by Taljard et al. (2008) for Northern Sotho graphemes which already contains two levels of description, we extend it to provide for Zulu items as well (see Appendix). We then sort these categories according to the general principles of morphological categorization: bound and free items, items that generate concordial agreement with others and items that don’t. We also differentiate between the basic POS categories like nominal, verbal and others. By filling the ontology with all the morphemes we know of and which we find in written texts of both of the languages, we attempt a complete description of all of the existing categories[2].

3. AN ONTOLOGY OF MORPHEMIC UNITS OF NORTHERN SOTHO AND ZULU

Most lexical knowledge bases currently used in NLP are designed as ontologies, i.e. a formal representation of knowledge as a set of concepts, also representing the relationships between these concepts. Examples of such ontologies are the Princeton WordNet (Fellbaum, 1998) and the African languages Wordnet that is under construction for four South African Bantu languages as described by Le Roux et al. (2007). Making use of an ontology to describe morphemic units or parts of speech is a less known application, but foreseen in NLP, as discussed in section 2 above.

The POS hierarchies described by Kahrel et al. (1997), Khoury et al. (2008) and Van Rooy and Pretorius (2003), can also be regarded as ontologies, as shown in Figure 1(a), (b) and (c) respectively. Instead of hyperonyms and hyponyms, we find supertypes and subtypes, where the subtype inherits all the properties assigned to the supertype.

[2] We do not attempt to fill the database with all of the stems and roots of nouns and verbs, we will however collect as many as we can find.

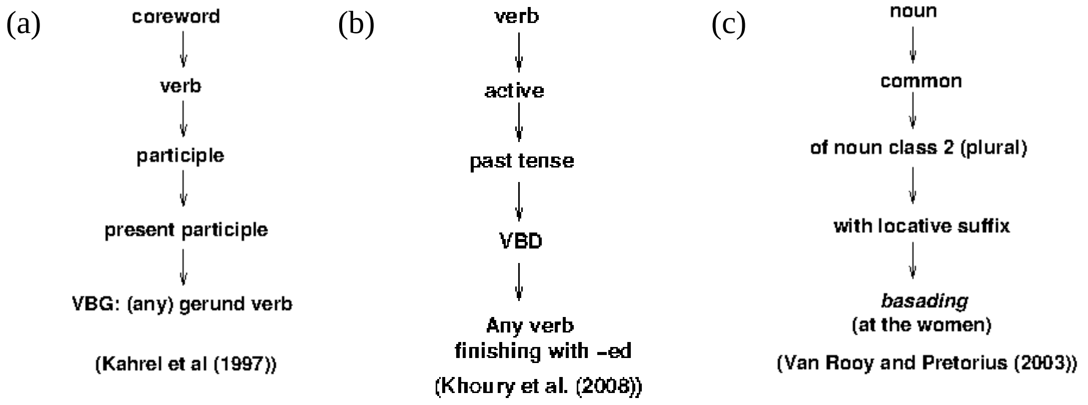

Figure 1. *POS hierarchies as ontologies.*

Ontologies can be implemented either as hierarchical databases, implementable with for example XML or as relational databases. The latter are implementable with e.g. SQL. Since converting between the formats is no longer considered a problem, we chose the relational layout and SQL for ease of the first implementation.

3.1 MORPHOLOGICAL ITEMS AS A BASIS OF NATURAL LANGUAGE PROCESSING

In many languages, tokens/word forms are described on the level of POS, leading to a more abstract linguistic representation level of these forms. For generating such a POS level, a morphological analysis often has to be done first. Syntactic structures can then be defined on the basis of this representation. These form another, more abstract representation.

This is clearly not the case for the Bantu languages. In Table 2, the Northern Sotho phrase *monna yo e lego morutiši* (the man who is a teacher) is shown with all levels of representation, i.e. morphological, POS, and the syntactic structure levels 1 and 2. For ease of reference, translations are added.

Levels:					
Surface	*monna*	*yo*	*e*	*lego*	*morutiši*
Morphological Representation	*mo-nna* 01-noun, stem is *-nna*	*yo* 01-demonstrative concord	*e* NEUT-subject concord	*le-go* copulative verb with relative suffix	*mo-rutiši* 01-noun, stem is *-rutiši*
Translation	man	the/this		who is	teacher
POS Representation	N.01	CDEM.01	CSNEUT	VCOP	N.01
Syntactic Structure level 1	noun phrase		verbal phrase		
Syntactic level 2	noun phrase				

Table 2. *Levels of linguistic representation in Northern Sotho.*

Table 2 demonstrates that specifically in the Bantu languages, morphological and syntactical analyses often have to go hand in hand, as official orthographies sometimes obscure the underlying morphological structure of linguistic texts. In these languages, a graphical token can either represent a morpheme, a lexeme, a phrase or even a sentence[3]. If words of these languages are thus described on the level of morphemic units instead of parts of speech, these can form a linguistically sound basis for e.g. an electronic grammar doing both morphological and syntactic generation or analysis. However, using morphological items as a basis for natural language processing in our particular context poses certain challenges, specifically with regard to the two languages selected, as was pointed out by Hendrikse and Mfusi (2008: 189). The relevant issues are briefly addressed below.

As indicated in the introduction above, two different writing systems are utilized for the Sotho and Nguni groups of languages; the former using a partial disjunctive system, the latter a conjunctive one. Northern Sotho, a member of the Sotho languages, is a disjunctively written language, where some morphs forming a word are written as separate orthographic entities. Of these, the majority can be considered grammatical morphs (surface forms of morphemes), i.e. elements with no referential meaning, but a grammatical function (e.g. the subject concord *e* in Table 2 above). Zulu, a member of the Nguni languages, is a conjunctively written language in which morphemes are often merged on the surface level resulting in a one to one correspondence of orthographic words (graphemes) and linguistic words, where the term linguistic word is understood to refer to any unit or units constituting a specific word category, e.g. verb, noun, ideophone, etc. (Kosch, 2006: 4). At first glance, these phenomena seem to require very different methodologies for describing linguistic units when automated processing is one of the aims of the description. However, when texts of these languages are segmented into their morphemic units, the underlying structural similarities are clearly revealed (Taljard and Bosch, 2006). The close relationship between the South African Bantu languages, which are lesser resourced languages, makes re-use of any application a promising and practical proposition. However, different theoretical approaches to word class categorization in Northern Sotho and Zulu can be a potential stumbling block in this regard. This being said, it does not fall within the ambit of this article to re-evaluate existing word class categorizations. Therefore, existing categorizations as described in standard grammars of these two languages are used as point of departure for our analysis. Secondly, often as a result of different types of word class categorization, terminological differences arise, which need to be resolved before the actual design of an ontology can be attempted. These two aspects will briefly be discussed.

[3] See for instance Faaß (2010), for a morpho-syntactic description of Northern Sotho for encoding an electronic grammar on a token basis.

Traditionally, Zulu grammarians favour the word class categorization proposed by C M Doke (Kosch, 1993: 32), which is often called a functionalist approach, since according to Doke, the (syntactic) function of a word takes precedence over its morphological features when a categorization is made. Northern Sotho grammarians on the other hand, usually opt for the more structuralist approach as formulated by Van Wyk (1961). In his classification, syntactic and morphological principles take precedence over phonological and semantic principles. As a result of the utilization of different classificatory principles, different categorizations emerge. To illustrate: in Northern Sotho, words such as *fase* 'below' and *godimo* 'on top of' are categorized as (locative) nouns, inter alia based on their ability to potentially function as the subjects of sentences, and the fact that their morphology is typically that of nouns, i.e. consisting of a class prefix (*fa*- and *go*- respectively) and a stem (*-se* and *-dimo* respectively). Within a Dokean framework, the Zulu equivalents *phansi* 'below' and *phakathi* 'on top of', although morphologically similar to the Northern Sotho examples, are classified as adverbs, since this is the function attributed to these words. Since there are no inherent differences between the Northern Sotho and Zulu items, the ideal would be to find some compromise, in order to maximally utilize the similarities existing between these two languages. An example of such a compromise concerns the word class 'particle', which is distinguished for Northern Sotho but not for Zulu. This category includes the instrumental particle *ka* 'with', the associative particle *le* 'together with' and the agentive particle *ke* 'by'. Similar items in Zulu are categorized as prefixes. Since it was found that the definition of prefix can be interpreted to also subsume that of particle (Louwrens, 1994: 133), it was decided to categorize these Northern Sotho items that are traditionally classified as particles for the purposes of the ontology design as prefixes. As a result of the deviation from traditionally distinguished categories of grammatical description, the descriptive part of the ontology will be very important: here, we can document language specific descriptions, provide examples and also list and/or describe the respective terms that have been used by other linguists.

3.2 A MULTI-LEVEL APPROACH

Linguists usually differentiate between different levels of representation of the units of a language. The smallest such unit is the morpheme representing the morphemic level. One or several morphemes form words that are represented at word level. Morphemes are sorted into different categories: bound morphemes never appear alone, but be attached to other morphemes, while free morphemes may appear alone. Inflectional morphemes are always bound morphemes, as they add morpho-syntactic information to other morphemes, e.g. the English morpheme 's' added to a verb in order to indicate that the subject of this verb is of the 1st person singular. Morphemes put together (some merging to one grapheme, some not) form words, of which the ones with a specific sense are called 'lexemes', representing the word level. Hence it might happen that one morpheme also represents a word in itself (like most English verbs).

In most languages, only the word level is usually labelled with parts of speech; however, in a number of descriptions of the Bantu languages, such as Van Rooy and Pretorius (2003), graphemic units are all described as if they are fully-fledged parts of speech, i.e. on the same linguistic level, and a text is seen as a sequence of these, though there are indeed categories from different linguistic levels to be found when taking a closer look (as already indicated in Table 2). Table 3 shows some morphemic units that form a Northern Sotho text. The labelling of these units follows the proposal put forward by Taljard et al. (2008).

No.	***Surface***	***Category (translation)***	***Tag***	***Morpheme***	***Description***
1	*basadi*	noun	N02	*ba-*	class prefix class 2
2		('women')		*-sadi*	noun stem
3	*ba*	verb	CS02	*ba*	subject concord class 2
4	*a*	('sell them')	MORPH_pres	*a*	present tense morpheme
5	*di*		CO08/10	*di*	object concord class 8 or 10
6	*rekiša*		V	*rek-*	verb root
7				*-iš-*	causative extension
8				*-a*	verbal ending

Table 3. *Current POS annotation versus morphemic units for Northern Sotho (NSO).*

In this example, the only graphical token that represents a POS is *basadi* 'women', which is a noun in class 2. Morphologically, this noun consists of a class prefix of class 2 and a noun stem. Units three and five are agreement morphemes, the fourth belongs to the group of inflectional morphemes, the sixth is a lexeme, followed by another inflectional morpheme indicating causality, whereas the last one constitutes the verbal ending. Even though it is therefore clear that these elements belong to different structural levels, they are often described as if belonging to the same level (annotations done as described by Taljard et al. (2008)).

In Table 3 above, every unit represents a morpheme. Within the category morpheme, two types are distinguished, namely class dependent and class independent morphemes; the latter in this case reflecting tense. We need to differentiate between class dependent and independent morphemes as this helps us to distinguish between morphemes that require class information when forming words and those that do not. Zulu utilizes a conjunctive writing system, and to our knowledge only Spiegler et al. (2010a/b) have developed a POS tagset for this language. However, they first apply a morphological analysis and annotate morphemic units making use – similar to Hendrikse and Mfusi (2008) – of a (Definite clause) grammar. The grammar creates one or several hypothetic morphological analyses for each of the graphemic units. After manual disambiguation (many graphemic units get several morphological analyses), their tagger annotates POS, making use of a simplified tagset containing 15 tags

(Verb, Noun, Relative, Prepositional, Possessive, Copulative, Locative, Adverb, Modal, Demonstrative, Pronoun, Interjection, Presentative, Adjective, and Conjunction).

Pretorius and Bosch (2003: 210) apply morphological analysis only, as illustrated in Table 4.

No	*Surface*	*Category*	*MORPH*	*Tag*	*Description*
1	*abazulithengisa*	*verb*	*a-*	[NegPre]	negative morpheme
2			*-ba-*	[SC2]	subject concord class 2
3			*-zu-*	[FutNeg]	future tense negative morpheme
4			*-li-*	[OC15]	object concord class 5
5			*-theng-*	[VRoot]	verb root
6			*-is-*	[CausExt]	causative extension
7			*-a*	[VerbTerm]	verbal ending

Table 4. *Labelling of morphemic units for Zulu (ZUL).*

The orthographic word *abazulithengisa* ‘they will not sell it’ is a linguistic word of Zulu, belonging to the verb category. This correspondence between orthographic and linguistic words is a characteristic feature of Zulu, which distinguishes it from Northern Sotho (Taljard and Bosch, 2006). Morphologically this verb consists of a negative morpheme *a-*, two class dependent agreement morphemes in the second and fourth units, and an inflectional morpheme indicating future tense in the third unit. The verb root *-theng-* ‘buy’ is followed by another inflectional morpheme in unit 6, namely the causative extension, followed by the verbal ending *-a*. In the Bantu languages, so-called verbal extensions may be suffixed to verb roots in order to modify the basic meaning of the verb root. In certain cases, more than one extension may be added. Examples of verbal extensions are: applied, causative, neuter, passive and reciprocal. See the Zulu examples in (1a-c) and the Northern Sotho examples in (1d-f):

(1) a) *-fund-a* > *-funda* ‘learn’

b) *-fund-is-a* > *-fundisa* ‘teach’

c) *-fund-is-an-a* > *-fundisana* ‘teach each other’

d) *-seny-a* > *-senya* ‘destroy’

e) *-seny-el-a* > *-senyela* ‘destroy for’

f) *-seny-el-iš-a* > *-senyetša* ‘cause to destroy for’

The development of our ontology consists of three components:

– Identification and classification of the morphemic units of both languages. We need to ascertain which morphemic units are common to both languages, and which are language specific. In the longer term, the project will be extended to include other (South African) Bantu languages as well. The model could then be adapted, if necessary, to make provision for any language specific elements.

- Categorization of all known morphemic units according to their linguistic properties. Elements that share properties need to be identified and listed in order to combine these elements in a sensible manner. The ontology of categories is then built based on these properties.
- Identification of closed classes and an inventory for open classes. For every language, the members of closed classes need to be identified, and for open classes, an inventory must be compiled. Such an inventory would be organic in the sense that, as the work progresses, new items can continually be added to it.

This work is therefore a database under construction and forms the basis for the development of tools concerned with the morphological and syntactic analysis and generation of surface words and text. In the following section the design of such an ontology is described.

3.3 DESIGN OF THE ONTOLOGY

The main part of the ontology contains two major categories: class independent morphemes shown in Figures 2 to 4, and class dependent (bound) morphemes as shown in Figure 5.

Class independent morphemes consist of either free morphemes or bound morphemes. According to Kosch (2006: 7) a free morpheme is a self-contained word that appears in isolation while carrying a complete meaning on its own. Although free morphemes are rather scarce in the Bantu languages they do occur, as listed in Fig. 2. Bound morphemes on the other hand, occur more commonly and always rely on at least one other morpheme to complete the meaning of a word. In the Bantu languages class independent bound morphemes are classified as verbal, adverbial, adjectival and nominal.

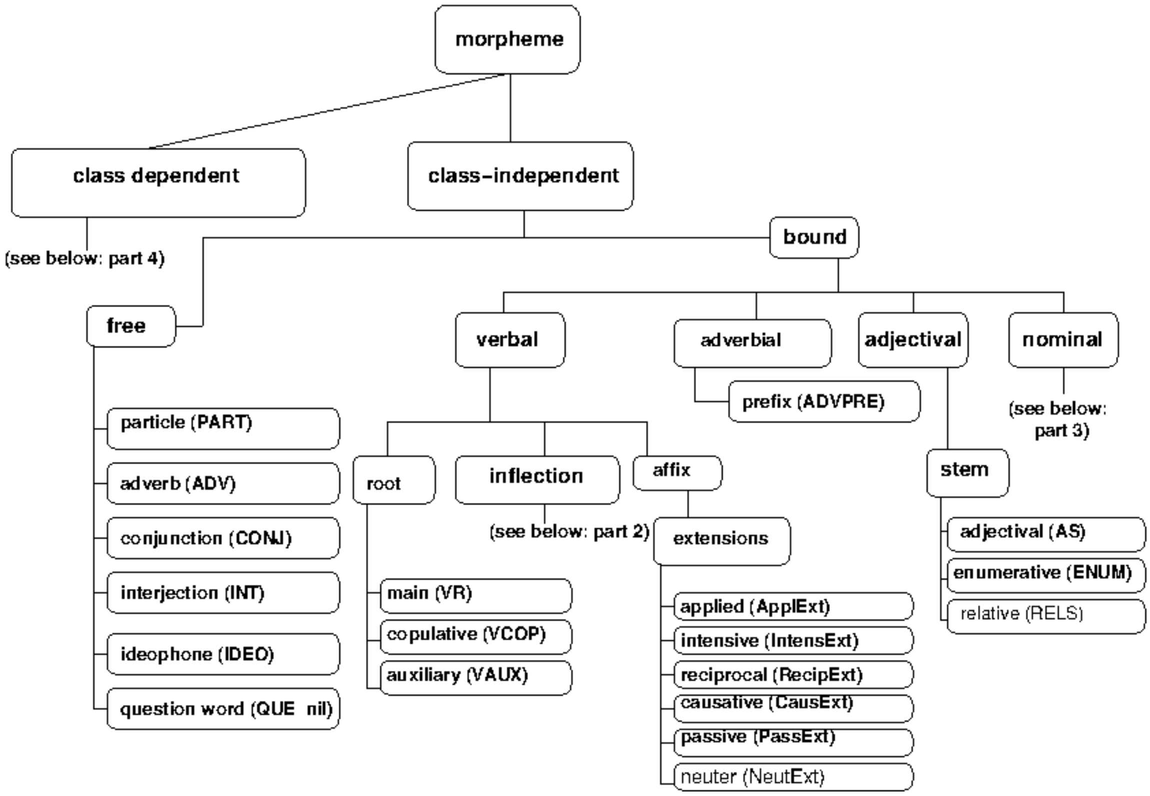

Figure 2. *The class-independent morphemes part 1/4.*

Bound verbal morphemes are in turn divided into three categories, i.e. roots, inflectional morphemes and affixes. We use the term ‘verbal root’ to signify “single morphemes that carry the principal semantic load of a word” (Kosch, 2006: 7). Inflectional morphemes (cf. Fig. 3) typically have grammatical meaning, and do not change the lexical meaning of a word to which they are affixed. If a verb is, for example, inflected to refer to an action in the past, in the present moment or to a future event, the lexical meaning of the verb is not affected. The relative suffix could perhaps be regarded as the prototypical inflectional morpheme – it has no lexical content whatsoever, and simply serves to complete a syntactic construction, as is illustrated in (2) below:

(2) a) *Siya esibhedlela nengane ekhalayo* > *e*$_{\text{REL09}}$*khal*$_{\text{VR}}$*a*$_{\text{Vend}}$*yo*$_{\text{RelSuff}}$
‘We go to the hospital with the child that is crying’

b) *Siya esibhedlela nengane ekhala njalo* > *e*$_{\text{REL09}}$*khal*$_{\text{VR}}$*a*$_{\text{Vend}}$
‘We go to the hospital with the child that is always crying’

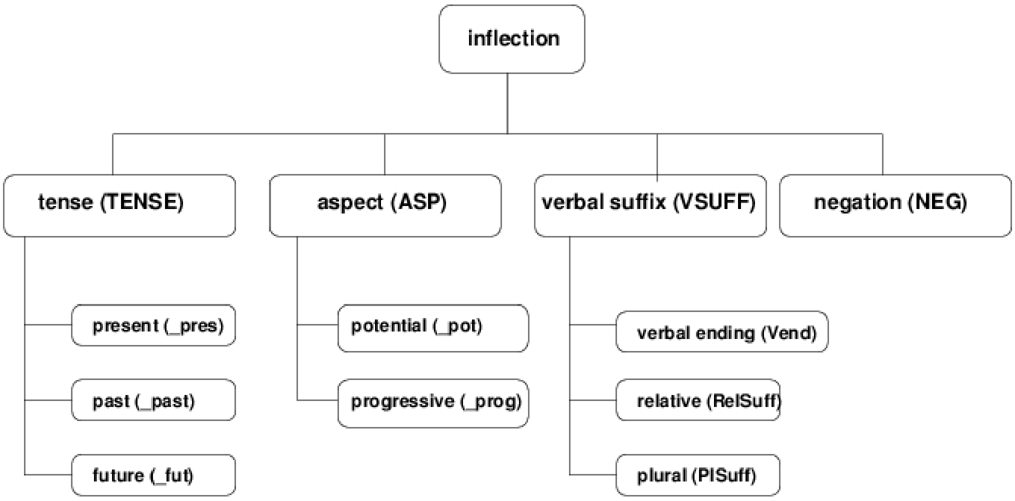

Figure 3. *The class-independent inflectional morphemes part 2/4.*

Nominal affixes (cf. Fig. 4) are optional morphemes and are generally not dependent on noun classes. According to Zulu grammars, there are some exceptions such as the locative prefix *ku-* in Zulu (cf. 3a) that occurs only in the case of nouns in classes 1, 1a, 2 and 2a [+human] nouns in class 6, whereas other noun classes use the locative prefix *e-* followed by a locative suffix *-ini* (cf. 3b); and in exceptional cases by prefixation of the prefix *e-* only (cf. 3c).

(3) a) *ubaba*$_{\text{N01a}}$ ‘father’
ku-u-baba$_{\text{N01a\_loc}}$ > *kubaba* ‘to/at father’

b) *indlela* $_{\text{N09}}$ ‘road’
e-in-ndlela-ini$_{\text{N09\_loc}}$ > *endleleni* ‘on the road’

c) *ikhaya*$_{\text{N05}}$ ‘home’
e-i(li)-khaya $_{\text{N05\_loc}}$ > *ekhaya* ‘at home’

However, De Schryver and Gauton’s (2002: 209) corpus-based research reveals that the locative prefix *ku-* is used widely with nouns other than those mentioned above, and can therefore no longer be restricted to specific noun classes.

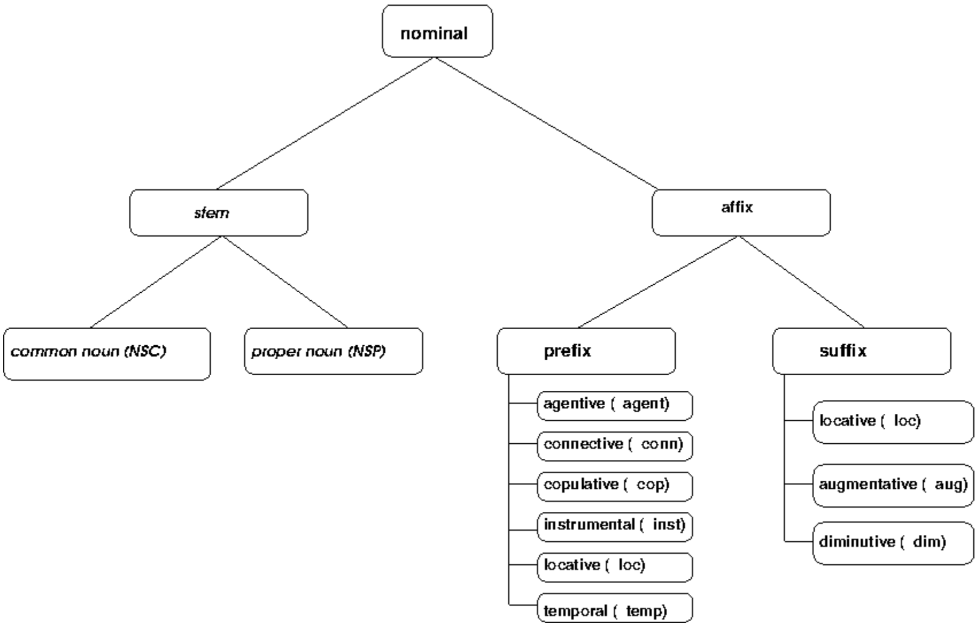

Figure 4. *The class-independent nominal morphemes part 3/4.*

Figure 5 reflects the distinction that is made between the so-called 'adjective' and 'relative' constructions in the grammars of the Zulu language. The adjective consists of an agreement morpheme or adjective concord plus an adjective stem. A limited number of adjective stems occur in Zulu. The relative on the other hand, consists of an agreement morpheme (or relative concord) that differs slightly in form from the adjective concord, particularly in the classes containing a nasal in the class prefix morpheme. The relative stem that follows on the relative concord is either a primitive relative stem or is based on verb or copulative stems, cf. (4).

(4) a) *Umfana omude* 'The tall boy' > $\textit{omu}_{\text{ADJ01}}$ - $\textit{de}_{\text{ADJ\_stem}}$

b) *Umfana o-qotho* 'The honest boy' > $\textit{o}_{\text{REL01}}$ - $\textit{qotho}_{\text{REL\_stem}}$

Furthermore, for the purpose of the design of our ontology, it was decided to distinguish two categories for the demonstrative of Northern Sotho (cf. Fig. 5), although in traditional grammars, it is mostly classified as a (deictic) pronoun – a categorization which is not always uncontested. The demonstrative also forms part of qualificative constructions, such as the adjective and verbal relative constructions, where it has neither a deictic, nor a pronominal function. In cases such as these, the demonstrative functions as a grammatical marker of a particular syntactic structure, and is therefore regarded as a demonstrative concord. Compare the following examples by way of illustration:

(5) a) *Morutiši yo o ruta Sepedi* 'This teacher teaches Northern Sotho'
(*yo* = demonstrative with deictic function)

b) *Morutiši yo a rutago Sepedi* … 'The teacher who teaches Northern Sotho …'
(*yo* = grammatical marker of the relative construction)

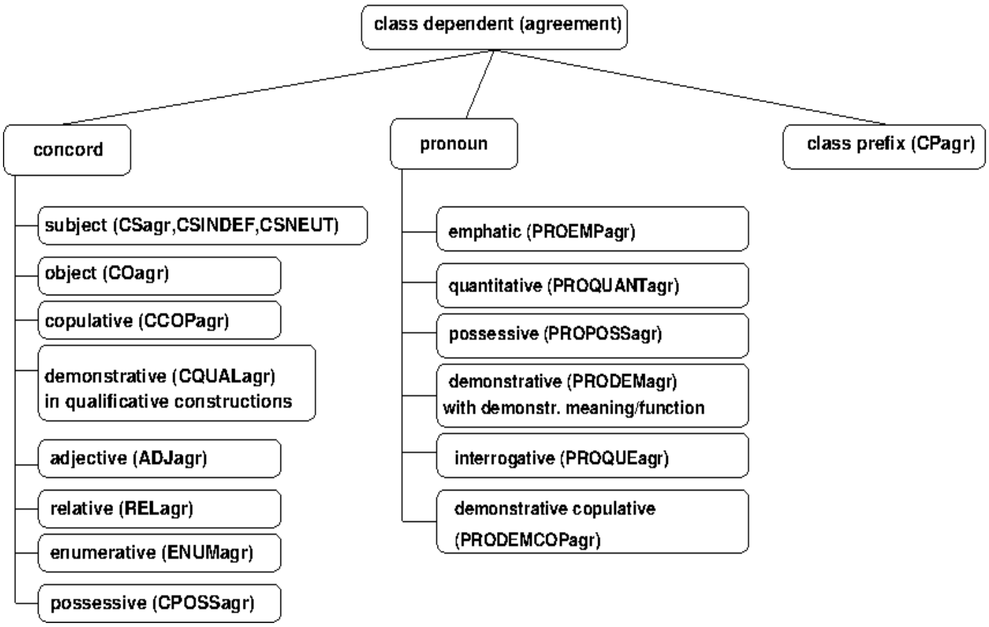
class dependent (agreement)
concord
pronoun
class prefix (CPagr)
subject (CSagr,CSINDEF,CSNEUT)
object (COagr)
copulative (CCOPagr)
demonstrative (CQUALagr) in qualificative constructions
adjective (ADJagr)
relative (RELagr)
enumerative (ENUMagr)
possessive (CPOSSagr)
emphatic (PROEMPagr)
quantitative (PROQUANTagr)
possessive (PROPOSSagr)
demonstrative (PRODEMagr) with demonstr. meaning/function
interrogative (PROQUEagr)
demonstrative copulative (PRODEMCOPagr)

Figure 5. *The class-dependent agreement morphemes 4/4.*

In the ontology parts described above, agreement has only been marked with a place holder 'agr'. We however aim at assigning information on person, class and number to all of the items that play a role in syntactic agreement, therefore a table is to be added to the database describing all possible combinations. When entering e.g. a subject concord to the database, it will be linked with one specific entry of this table to provide information on its person, number, and/or class.

Figure 6 demonstrates our encoding approach: we follow the idea of storing the same (type of) information for a particular category, we suggest designing the supertype agreement (*AGR*) in a way that it not only uses the noun class (*CLASS*) as a sub-type, but also person (*PERS*) and number (*NUM*). *CLASS* is a subtype of *PERSON*, as all units of the third person are described there. As nominal stems do in general not appear in all possible classes, class information for all nominal stems will need to be stored in the proposed database. Hence, lexical and grammatical information about an item will be linkable and – from a technical perspective – only correct forms will be generated by the future morphological analyzer/generator. Note that the element *NUM* is defined as being optional because this information may not be available or not applicable (e.g. for abstract nouns).

The last table of the database will contain the languages of which items are described in the ontology. All of the entries in the database will be linked with one or several of the languages.

$$\mathrm{AGR}\begin{bmatrix}\mathrm{PERS} & \begin{bmatrix}\mathit{1^{st}}\\ \mathit{2^{nd}}\\ \mathrm{CLASS}\begin{bmatrix}\mathit{01,\ 02,\ 03,\ ..\ 10,\ 14,\ LOC}\end{bmatrix}\end{bmatrix}\\ (\ \mathrm{NUM} & \begin{bmatrix}\mathit{sg}\\ \mathit{pl}\\ \mathit{none}\end{bmatrix})\end{bmatrix}$$

Figure 6. *An attribute-value matrix describing the type hierarchy of agreement*.

Once the ontology has been fully developed, a glossary-like definition of every category will be provided in the documentation. It will then be possible to extend these definitions to other Bantu languages, because provision will be made for language specific elements not yet known.

3.4 DATABASE LAYOUT

A typical approach for an implementation of an ontology is using a hierarchical structure, usually making use of the programming language XML. However, relations between linguistic items are not always hierarchical, therefore XML implementations nowadays reflect multiple relations between items by using several stand-off annotation files for each of the original items. Spohr (2012: 28) rightfully argues for such an implementation which makes use of a typed formalism for linguistic ontologies for lexicographic purposes. We are designing our ontology in a way implementable with XML; however, for ease of our current first implementation, we decided to begin with a rather simple relational DB implemented with MySQL.

In a relational database, several tables (= **relations**) store the data sets (the linguistic items, identified by a unique id, the *primary key*). A table consists of columns (= **attributes)** and lines (= **tuples**) containing the values for these attributes. Each tuple usually has a unique identity (id) assigned automatically to which other tuples can be linked.

The core of such a database is the relational scheme describing the types of attributes that may appear in each of the tables and the possible values that may appear in the tuples. In our case, Figures 2 to 5 above represent the basic tables of the relations. Figure 6 is interlinked with all the tables containing agreement items. These relations are shown in Figure 7 below.

3.5 METHODOLOGY

There have been some attempts to fully specify the (closed) morphological units of Northern Sotho (e.g. Faaß, 2010) and Zulu (e.g. Spiegler et al., 2010a/b). Our ontology summarizes the morphemes which are identical for both languages, but simultaneously also makes provision for language specific additions. We make use of the afore-mentioned descriptions and other linguistic documentation (e.g. Lombard, 1985) for Northern Sotho and (e.g. Poulos and Msimang, 1998) for Zulu as a basis for our inventory.

While collecting the data, we made use of excel tables that represent the ontology by way of columns as illustrated in Table 6. The column ‘Description’ is used by the linguist to document rules for the planned morphological analyzer/generator.

Level 0,1,2,3	***4***	***5***	***NSO-item***	***ZUL-item***	***Description***
class independent morpheme, bound, verbal, derivation	extensions (Ext)	applied (_appl)	el	el	
class independent morpheme, bound, verbal, derivation	extensions (Ext)	applied (_appl)	letš		verb stems ending in -tš(a)

Table 6. *An excerpt of one of the filled Excel tables.*

The database tables reflect the hierarchy described above with one exception: there is one table describing all of the morpheme forms, assigning a primary key (which is its unique id) to each of them. All tables containing descriptions as shown in Figures 2 to 5 above only refer to these ids in order to assign a specific morpheme to a specific category. In addition to the agreement table, the languages table containing unique ids for each of the described languages, is also being assigned to the items described in the ontology tables, see Figure 7 below.

In Figure 7, arrows between the tables describe their relations. For example, between the table ‘morphemes’ and the tables containing class dependent items, the relation is n:m, meaning that several of the primary keys of the morpheme table may be referred to from several primary keys of the tables containing class dependent items. On the other hand, only one language and only one agreement feature may be assigned to each of those items, though several of those items may be referred to the same language and to the same agreement feature (n:1).

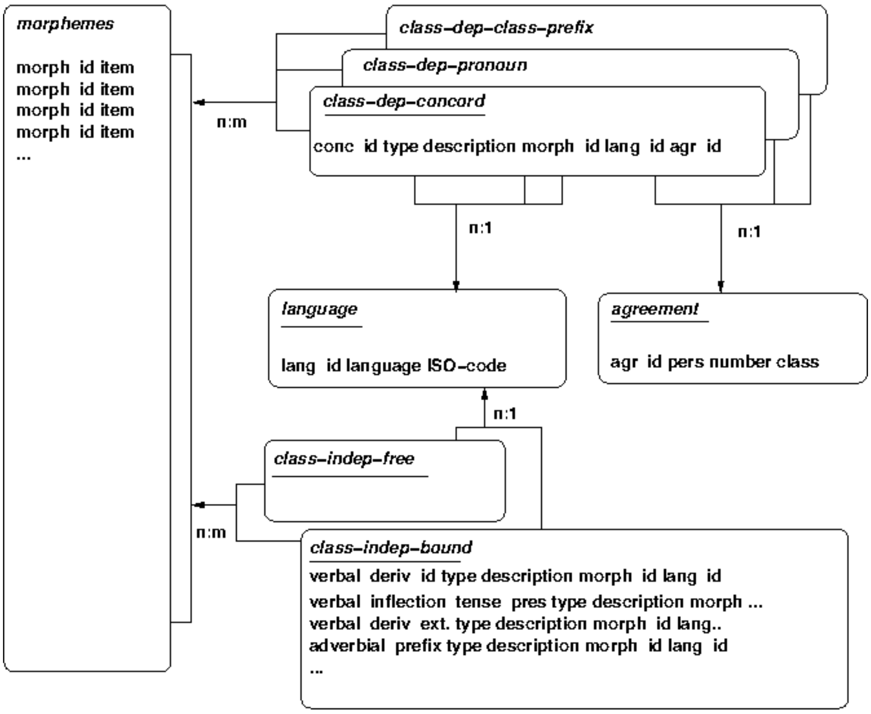
morphemes
morph id item
morph id item
morph id item
morph id item
...
class-dep-class-prefix
class-dep-pronoun
class-dep-concord
conc id type description morph id lang id agr id
n:m
n:1
n:1
language
lang id language ISO-code
agreement
agr id pers number class
n:1
class-indep-free
n:m
class-indep-bound
verbal deriv id type description morph id lang id
verbal inflection tense pres type description morph ...
verbal deriv ext. type description morph id lang..
adverbial prefix type description morph id lang id
...

Figure 7. *The basic design of the SQL-database.*

Note that Figure 7 only shows a simplified structure, as we are still in the process of designing the tables in detail. We will report in a follow-up article on its details when the database has been fully implemented.

4. CONCLUSION AND FUTURE WORK

In this article, a first attempt is made at documenting and encoding morphemic units of two South African Bantu languages. It is shown that it is necessary to describe units on the sub-word level when aiming at tagging each of the morphemes properly. These languages require a description of very finely grained categories, e.g. bound or free morphemes, class dependent or class independent, etc. We illustrate that it is feasible to store and to represent such morphemic units of two Bantu languages jointly in a single ontology. Our ontology is designed in such a way that it makes provision for morphemes common to all Bantu languages, as well as for language specific ones.

Our current implementation is on-going, since we are still in the process of compiling the inventory and fully specifying each category. Thus future work includes the full specification as a first step. This will be presented for discussion to experts of other Nguni and Sotho languages. While under development, interested researchers may get access to the inventory for the purpose of correcting and filling it with the necessary data. When the major parts are agreed upon, we will further specify and encode the ontology which will then be made freely available and extendable to other, closely related Bantu languages via a web interface. Further project steps will entail the development of morphological analyzers/generators that will make use of the database to analyze/generate surface words.

Acknowledgements

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/).

References

Allwood, J., Grönqvist, L. and Hendrikse, A.P. 2003.
Developing a tagset and tagger for the African languages of South Africa with special reference to Xhosa. **Southern African Linguistics and Applied Language Studies** 21(4): 223–237.

Anderson, W.N. & Kotzé, A.E. 2008.
Verbal extension sequencing: an examination from a computational perspective. **Literator** 29(1): 43–64.

Bosch, S., Pretorius, L. and Jones, J. 2007.
Towards Machine-Readable lexicons for South African Bantu languages. **Nordic Journal of African Studies** 16(2): 131–145.

De Schryver, G.M. and Gauton, R. 2002.
The Zulu locative prefix ku- revisited: a corpus-based approach. **Southern African Linguistics and Applied Language Studies** 20: 201–220. ISSN 1607–3614.

EAGLES 1996.
Recommendations for the Morphosyntactic Annotation of Corpora. http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html (Accessed: May 31, 2013).

Faaß, G. 2010.
A morphosyntactic description of Northern Sotho as a basis for an automated translation from Northern Sotho into English. PhD thesis, University of Pretoria, Pretoria, South Africa.

Faaß, G., Heid, U., Taljard, E. and Prinsloo, D. 2009.
Part-of-Speech tagging in Northern Sotho: disambiguating polysemous function words. In: *Proceedings of the EACL2009 Workshop on Language Technologies for African Languages – AfLaT 2009, The 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 38–45.

Fellbaum, C. (ed). 1998.
WordNet. An Electronic Lexical Database. Cambridge, MA: MIT Press.

Heid, U. 2000.
Morphologie und Lexikon. Chapter 17. In: Görz, G., Rollinger, C.R., and Schneeberger, J. (ed.), *Handbuch der künstlichen Intelligenz,* pp. 665–709. München/Wien: Oldenbourg Verlag.

Hendrikse, R. and & Mfusi, M. 2008.
A morphosyntactic tag set for Southern Bantu within a Construction Grammar Approach. **Language Matters** 39(2): 181–203

Kahrel P., Barnett, R. and Leech, G. 1997.
Corpus Annotation. Linguistic Information from Computer Text Corpora. In: *Towards cross-linguistic standards or guidelines for the annotation of corpora*, pp. 231–242. Longman, London/New York.

Khoury, R., Karray, F. and Kamel, M. 2008.
Keyword Extraction Rules Base 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. Via Afrika, Pretoria.

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

Le Roux J., Moropa, K., Bosch, S. and Fellbaum, C. 2007.
Introducing the African Languages Wordnet. In: Tan'acs A,Csendes D, Vincze V, Fellbaum C, Vossen P (eds): *Proceedings of the Fourth Global WordNet Conference*, pp. 269–280. University of Szeged, Dept. of Linguistics, Szeged, Hungary. ISBN 978-963-482-854-9.

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

Louwrens, L.J. 1994.
Dictionary of Northern Sotho linguistic terms. Pretoria, South Africa: Via Afrika.

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

Pretorius, L. and Bosch, S. 2003.
Finite-State Computational Morphology: An Analyzer Prototype for Zulu. **Machine Translation** 18: 195–216.

Schmid, H. and Laws, F. 2008.
Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In: *Proceedings of the 22nd International Conference on Computational Linguistics* (*COLING),* pp. 777–784. Manchester, UK.

Spiegler, S., van der Spuy, A. and Flach, P. 2010a.
Ukwabelana - an open-source morphological Zulu corpus. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pp. 1020–1028. Beijing.

2010b *Additional material for the Ukwabelana Zulu Corpus*. http://www.cs.bris.ac.uk/Publications/Papers/2001225.pdf (Accessed: May 31, 2013).

Spohr, D. 2012.
Towards a Multifunctional Lexical Resource. Berlin/Boston: DeGruyter.

Taljard, E. and Bosch, S. 2006.
A Comparison of Approaches to Word Class Tagging: Distinctively Versus Conjunctively Written Bantu Languages. **Nordic Journal of African Studies** 15(4): 428–442.

Taljard E., Faaß G., Heid, U. and Prinsloo, D. 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.

Van Wyk, E.B. 1961.
Die woordklasse van Noord-Sotho. *Feesbundel vir Prof. Dr. Jan Antonie Engelbrecht*. Johannesburg: Afrikaanse Pers Beperk.

Van Rooy, B. and Pretorius, R. 2003.
A word-class tagset for Setswana. **Southern African Linguistics and Applied Language Studies** 21 (4): 203–222.

About the authors: *Gertrud Faaß* is a Research Fellow in the Department of African Languages University of South Africa (UNISA). She is working towards electronic representation of Northern Sotho linguistic units, textual data and metadata. *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. *Elsabé Taljard* is associate 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.

APPENDIX 1: A COMBINED TAGSET

Description	***NSO***		***ZUL***	
	tag 1st level	***tag 2nd level***	***tag 1st level***	***tag 2nd level***
concords				
subject (all classes/persons)	CS	neut, indef, *agr*	CS	*agr*
object (all classes/persons)	CO	*agr*	CO	*agr*
copulative (all classes/persons)	CCOP	*agr*	-	-
demonstrative in qualificative constructions (all classes/persons)	CQUAL	*agr*	-	-
adjective	-	-	AD	*agr*
relative	-	-	REL	*agr*
enumerative (all classes)	ENUM	*agr*	ENUM	*agr*
possessive (all classes)	CPOSS	*agr*	CPOSS	*agr*
pronouns				
emphatic (all classes/persons)	PROEMP	*agr*	PROEMP	*agr*
quantitative (all classes/persons)	PROQUANT	*agr*	PROQUANT[4]	*agr*
possessive (all classes/persons)	PROPOSS	*agr*	PROPOSS	*agr*
demonstrative (all classes/persons)	PRODEM	*agr*	PRODEM	*agr*
interrogative (all classes/persons)	PROQUE	*agr*	-	-
demonstrative copulative	PRODEMCOP	*agr*	PRODEMCOP	*agr*
noun stems				
common noun	N	*SC*	N	*SC*
proper noun	N	*SP*	N	*SP*
class prefixes				
class prefix	CP	*agr*	CP	*agr*
nominal prefixes				
agentive	NPREF	*agent*	NPREF	*agent*
connective	NPREF	*conn*	NPREF	*conn*
copulative	NPREF	*cop*	NPREF	*cop*
instrumental	NPREF	*instr*	NPREF	*instr*
locative	NPREF	*loc*	NPREF	*loc*
temporal	NPREF	*temp*	-	-
nominal suffixes				
locative	NSUFF	*loc*	NSUFF	*loc*

[4] A distinction is made in Zulu between inclusive and exclusive quantitative pronouns.

augmentative	NSUFF	*aug*	NSUFF	*aug*
diminutive	NSUFF	*dim*	NSUFF	*dim*
verbal root morphemes				
main verb	VR	-	VR	-
copulative	VCOP	-	VCOP	-
auxiliary	VAUX	-	VAUX	-
verbal inflectional morphemes				
present tense	TENSE	*pres*	TENSE	*pres*
past tense	TENSE	*past*	TENSE	*past*
future tense	TENSE	*fut*	TENSE	*fut*
potential	ASPECT	*pot*	ASPECT	*pot*
progressive	ASPECT	*prog*	ASPECT	*prog*
verbal suffix ending	VSUFF	*Vend*	VSUFF	*Vend*
verbal suffix relative	VSUFF	*RelSuff*	VSUFF	*RelSuff*
verbal suffix plural	VSUFF	*PlSuff*	VSUFF	*PlSuff*
verbal negation	NEG	-	NEG	*pref/suff*
verbal affixes				
applied	VEXT	*appl*	VEXT	*appl*
intensive	VEXT	*intens*	VEXT	*intens*
reciprocal	VEXT	*recip*	VEXT	*recip*
causative	VEXT	*caus*	VEXT	*caus*
reflexive	VEXT	*refl*	VEXT	*refl*
passive	VEXT	*pass*	VEXT	*pass*
neuter	VEXT	*neut*	VEXT	*neut*
adverbial morphemes				
prefix	ADVPREF	-	ADVPREF	-
adjectival stems				
adjectival	ADJ	-	ADJ	-
enumerative	ENUM	-	ENUM	-
relative	-	-	REL	-
free morphemes				
particles	PART	-	-	-
adverb	ADV	-	ADV	-
conjunction	CONJ	-	CONJ	-
interjection	INT	-	INT	-
ideophone	IDEO	-	IDEO	-
question word	QUE	-	QUE	-