Nordic Journal of African Studies 15(4): 428–442 (2006)

A Comparison of Approaches to Word Class Tagging: Disjunctively vs. Conjunctively Written Bantu Languages*

ELSABÉ TALJARD
University of Pretoria, South-Africa
&
SONJA E. BOSCH
University of South Africa, South-Africa

ABSTRACT

Northern Sotho and Zulu are two South African Bantu languages that make use of different writing systems, viz. a disjunctive and a conjunctive writing system respectively. In this article it is argued that the different orthographic systems obscure the morphological similarities and that these systems impact directly on word class tagging for the two languages. It is illustrated that not only different approaches are needed for word class tagging, but also that the sequencing of tasks is to a large extent determined by the difference in writing systems.

Keywords: word class tagging, conjunctive writing system, disjunctive writing system, natural language processing, Bantu languages

1. INTRODUCTION

The aim of this article is to draw a comparison of approaches towards word class tagging in two orthographically distinct Bantu languages. The disjunctive versus conjunctive writing systems in the South African Bantu languages have direct implications for word class tagging. For purposes of this discussion we selected Northern Sotho, representing the disjunctive writing system, and Zulu as an example of a conjunctively written language. These two languages which belong to the South-Eastern zone of Bantu languages are two of the eleven official languages of South Africa. Northern Sotho is spoken by approximately 4,2 million mother-tongue speakers while Zulu is spoken by approximately 10,6 million mother-tongue speakers. Both these languages belong to a larger grouping of languages, i.e. the Sotho and Nguni language groups respectively. Languages belonging to the same language group are closely related and to a large extent mutually intelligible. Furthermore, since all three languages

* An earlier version of this paper was read at the Conference for Lesser Used Languages and Computer Linguistics, EURAC Research, European Academy. Bolzano, Italy. 27 – 28 October 2005 (cf. Taljard & Bosch, 2006).

belonging to the Sotho group follow the disjunctive method of writing, the methodology utilised for part-of-speech tagging in Northern Sotho would to a large extent be applicable to the other two Sotho languages (Southern Sotho and Tswana) as well. The same holds for Zulu with regard to the other Nguni languages, i.e. Xhosa, Swati, and Ndebele, which are also conjunctively written languages. The South African Bantu languages are not yet fully standardised with regard to orthography, terminology and spelling rules and compared to European languages, these languages cannot boast a wealth of linguistic resources. A limited number of grammar books and dictionaries is available for these languages, while computational resources are even scarcer. In terms of natural language processing, the Bantu languages in general undoubtedly belong to the lesser-studied languages of the world.

In this article a concise overview is firstly given of the relevant Bantu morphology and reference is made to the differing orthographical conventions. In the subsequent section the available linguistic and computational resources for the two languages are compared; thereafter a comparison is drawn between the approaches towards word class tagging for Northern Sotho and Zulu. In conclusion, future work regarding word class tagging for Bantu languages is discussed.

2. BANTU MORPHOLOGY AND ORTHOGRAPHY

According to Poulos and Louwrens (1994: 4), “there are [the] numerous similarities that can be seen in the structure (i.e. morphology) as well as the syntax of words and word categories, in the various languages of this family”. These languages are basically agglutinating in nature since prefixes and suffixes are used extensively in word formation.

The focus in this concise discussion on aspects of Bantu morphology is on the two basic morphological systems, namely the noun class system, and the resulting system of concordial agreement.

2.1 NOUN CLASSES AND CONCORDIAL AGREEMENT SYSTEM

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. In general, noun prefixes indicate number, with the uneven class numbers designating singular and the corresponding even class numbers designating plural. The following are examples of Meinhof's (1932: 48) numbering system of some of the noun class prefixes:

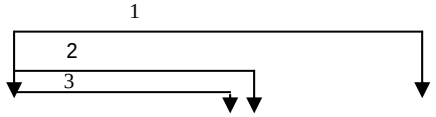
Table 1: Noun class system: illustrative excerpt.

Class #	Northern Sotho		Zulu	
	Prefix	Example	Prefix	Example
1 (sg)	mo-	motho “person”	umu-	umuntu “person”
2 (pl)	ba-	batho “persons”	aba-	abantu “persons”
1a(sg)	Ø-	makgolo “grandmother”	u-	udokotela “doctor”
2b(pl)	bo-	bomakgolo “grandmothers”	o-	odokotela “doctors”
3 (sg)	mo-	mohlare “tree”	umu-	umuthi “tree”
4 (pl)	me-	mehlare “trees”	imi-	imithi “trees”
7 (sg)	se-	setulo “chair”	isi-	isitsha “dish”
8 (pl)	di-	ditulo “chairs”	izi-	izitsha “dishes”
14	bo-	botho “humanity”	ubu-	ubuntu “humanity”

However, the correspondence between singular and plural classes is not perfectly regular, since some nouns in so-called plural classes do not have a singular form; in Zulu, class 11 nouns take their plurals in class 10, while a class such as 14 is not associated with number.

The significance of noun prefixes is not limited to the role they play in indicating the classes to which the different nouns belong. In fact, noun prefixes play a further important role in the morphological structure of the Bantu languages in that they link the noun to other words in the sentence. This linking is manifested by a system of concordial agreement, which is the pivotal constituent of the whole sentence structure, and governs grammatical agreement in verbs, adjectives, possessives, pronouns and so forth. The concordial morphemes are derived from the noun prefixes and usually bear a close resemblance to the noun prefixes, as illustrated by the bold printed morphemes in the following Northern Sotho example:

Figure 1: Concordial agreement – Northern Sotho.

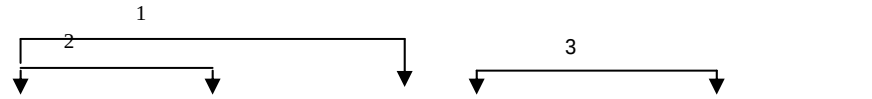


Bašemane	**ba ba**golo	**ba** ka bala	dipuku
Ba-šemane	**ba ba**-golo	**ba**-ka-bal-a	di-puku
Boys	who are big	they may read	books

“(The) big boys may read (the) books.”

In this sentence, three structural relationships can be identified. The class 2 noun *bašemane* "boys" governs the subject concord *ba-* in the verb *ba ka bala* "they may read" (1), as well as the class prefix *ba-* in the adjective *bagolo* "big" (2), and the demonstrative pronoun *ba*, preceding the adjective (3). The corresponding Zulu example would be as follows, where (1) indicates subject-verb agreement and (2) is agreement between the noun and the adjective concord *aba-* in the qualificative *abakhulu*. The class 10 noun *izincwadi* "books" determines concordial agreement of the object concord *-zi-* in the verb (3).

Figure 2: Concordial agreement – Zulu.



Abafana	**aba**khulu	**ba**nga**zi**funda	**izi**ncwadi
Aba-fana	**aba**-khulu	**ba**-nga-**zi**-fund-a	**izi**n-ncwadi
Boys	who are big	they may read them	books

"(The) big boys may read (the) books."

The predominantly agglutinating nature of the Bantu languages is clearly illustrated in the above sentences, each word of which consists of more than one morpheme. This complex morphological structure will be discussed very briefly by referring to two of the most complex word types, namely nouns and verbs.

2.2 MORPHOLOGY OF NOUNS

Nouns as well as verbs in the Bantu languages are constructed by means of the two generally recognized types of morphemes namely roots and affixes, the latter subdivided into prefixes and suffixes. The majority of roots are bound morphemes since they do not constitute words by themselves, but require one or more affixes to complete the word. The root is generally regarded to be "the core element of a word, the part which carries the basic meaning of a word." (Poulos & Msimang, 1996: 170). For instance, in the Northern Sotho example *dipuku* "books", the root that conveys the semantic significance of the word is *–puku* "book", the morpheme *di-* being the class prefix of class 10. In the Zulu word *izincwadi*, the prefixes are *i-* and *-zin-*, with *-ncwadi* carrying the basic meaning "book". By adding the suffixes *-ng* (Northern Sotho) and *-ini* (Zulu), and the prefix *e-* (in the case of Zulu) to the noun, a locative meaning is imparted:

Northern Sotho:	*dipukung*	di-puku-ng	"in the books"
Zulu:	*ezincwadini*	e-(i)-zin-ncwadi-ini	"in the books"

2.3 VERBAL MORPHOLOGY

In the case of the verb, the core element which expresses the basic meaning of the word is the verb root. The essential morphemes of a Bantu verb are a subject concord (except in the imperative and infinitive), a verb root and an inflectional ending. Over and above the subject concord (s.c.), the form of which is determined by the class of the subject noun, a number of other morphemes may be prefixed to a verb root. These include morphemes such as object concords (o.c.), potential and progressive morphemes as well as negative morphemes. Compare the following example in this regard:

Table 2: Verbal morphology - Northern Sotho & Zulu.

N.S	ba ka di bala	ba	ka	di	bal-	-a
Z	bangazifunda	ba-	-nga-	-zi-	-fund-	-a
	"they can read them"	s.c. cl 2	potential morpheme	o.c. cl 10	verb root	inflectional ending

It should be noted that whereas object concords also show concordial agreement with the class of the object noun, all other verbal affixes are class independent. Furthermore, verbal affixes have a fixed order in the construction of verb forms, with the object concord prefixed directly to the verb root.

Derivational suffixes may be inserted between the verb root and the inflectional ending. In the following examples the causative suffix *-iš- / -is-* has been suffixed to the verb root. It will furthermore be noted that the inflectional ending has changed to the negative *-e/-i* in accordance with the negative prefix *ga-/a*, e.g.

Table 3: Verbal derivation by means of suffixes.

N.S	ga ba rekiše	ga	ba	rek-	**-iš-**	-e
Z	abathengisi	a-	-ba-	-theng-	**-is-**	-i
	"they do not sell"	negative morpheme	s.c. cl 2	verb root	suffix	inflectional ending

2.4 CONJUNCTIVE VERSUS DISJUNCTIVE WRITING SYSTEMS

Following this explanation of the morphological structure of the Bantu languages, a few observations will be made regarding the different writing systems which are followed in the Bantu languages, with specific reference to Northern Sotho and Zulu. These different writing systems impact directly on

POS-tagging, as will be explained below (see also Hurskainen et al., 2005: 438). The following example illustrates the difference in writing systems:

Table 4: Conjunctivism vs disjunctivism.

	Orthographical representation	Morphological analysis				
N.S Z	ke a ba rata ngiyabathanda "I like them"	ke ngi- s.c. 1p.sg	a -ya- PRES	ba -ba- o.c. cl 2	rat- -thand- verb root	-a -a inflectional ending

The English translation "I like them" consists of three orthographic words, each of which is also a linguistic word, belonging to a different word category. In the case of the Zulu sentence, where the conjunctive system of writing is adhered to, we observe one orthographic word that corresponds to one linguistic word. This word is classified by Zulu linguists as a verb. The orthographic word *ngiyabathanda* is therefore also a linguistic word, belonging to a particular word category. This correspondence between orthographic and linguistic words is a characteristic feature of Zulu, which distinguishes it from Northern Sotho. In the disjunctively written Northern Sotho sentence, four orthographic words constitute one linguistic word that is again classified as a verb. In other words, in the latter case, four orthographic elements making up one word category are written as separate orthographic entities.

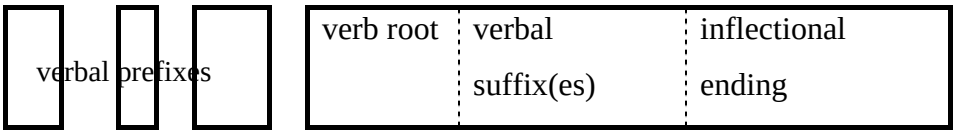
The reason for the utilization of different writing systems is based partly on historical and partly on phonological considerations. When Northern Sotho and Zulu were first put to writing, mainly by missionaries in the second half of the nineteenth century, they intuitively opted for disjunctivism when writing Northern Sotho and conjunctivism when writing Zulu. Thus an orthographic tradition was initiated that prevails even today. Although based on intuition, the decision to adopt either a conjunctive or a disjunctive writing system was probably guided by an underlying realisation that the phonological systems of the two languages necessitated different orthographical systems. As Wilkes (1985: 149) points out, the presence of phonological processes such as vowel elision, vowel coalescence and consonantalization in Zulu makes a disjunctive writing system highly impractical: the disjunctive representation of the sentence *Wayesezofika ekhaya* "He would have arrived at home" as *W a ye s' e zo fika ekhaya* is almost impossible to read and / or to pronounce. In Northern Sotho, these phonological processes are much less prevalent, and furthermore, most morphemes in this language are syllabic and therefore pose no problems for disjunctive writing.

However, what needs to be pointed out at this stage is that there is indeed some overlap with regard to the orthographical systems used by the two

languages and that Northern Sotho and Zulu should rather be viewed as occupying different positions on a continuum ranging from complete conjunctivism to complete disjunctivism. The diagrams below illustrate the degree of overlap between the writing systems of the two languages. (Dashed lines indicate morphological units, solid lines indicate orthographical units.) It can be observed that the disjunctive writing convention in Northern Sotho is mainly applicable to prefixes preceding the class prefix and prefixes preceding the verb root.

Figure 3: Overlap between conjunctivism and disjunctivism.

Northern Sotho verb structure:



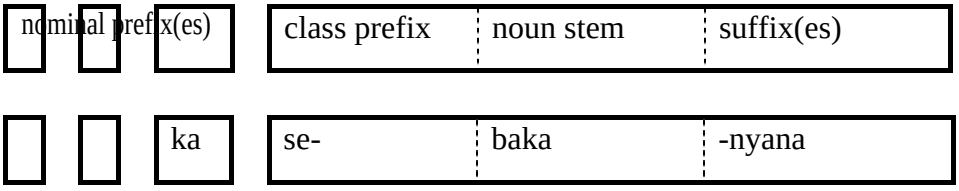
"They sell them"

Zulu verb structure:

verbal prefix(es)	verb root	verbal suffix(es)	inflectional ending
Ba-ya-zi-	-theng-	-is	-a

"They sell them"

Northern Sotho nominal structure:



"In a while"

Zulu nominal structure:

nominal prefix(es)	class prefix	noun stem	suffix(es)
nga-	isi-	khathi	-ana

"In a while"

At this stage it is important to note that the different writing systems utilised by the two languages actually obscure the underlying morphological similarities. These disjunctive versus conjunctive writing systems in the Bantu languages have direct implications for word class tagging, as will be demonstrated later in this article. In the next section the available computational resources for the two languages are compared.

3. COMPUTATIONAL LINGUISTIC RESOURCES

Existing linguistic and computational resources should be exploited as far as possible in order to facilitate the task of word class tagging. Both languages have unannotated electronic corpora at their disposal – approximately 6.5 million tokens for Northern Sotho, and 5.2 million tokens for Zulu. These corpora were compiled in the Department of African Languages at the University of Pretoria and consist of a mixed genre of texts including samples of most of the different literary genres, newspaper reports, academic texts, as well as internet material. Since most of the texts incorporated in the corpora were not available electronically, OCR scanning was done, followed by manual cleaning of scanned material.

The corpora have so far been utilised among others for the generation of frequency lists, which are of specific importance for the development and testing of word class tagging, especially in disjunctively written languages. In Northern Sotho, for instance, the top 10 000 types by frequency in the corpus represent approximately 90% of the tokens, whereas in Zulu the top 10 000 types represent only 62% of the tokens. This observation is directly related to the conjunctive vs disjunctive writing systems. Since frequency counts in an unannotated corpus are based on orthographical units, a large orthographic chunk such as *ngiyabathanda* found in Zulu would have a much lower frequency rate than the corresponding units *ke*, *a*, *ba* and *rata* in Northern Sotho. This implies that the correct tagging of the top 10 000 tokens in Northern Sotho, be it manual, automatic or a combination, results in a 90% correctly tagged corpus. The low relation between types vs tokens in Zulu, however, results in a much smaller percentage, that is, only 62% of the corpus being tagged. It furthermore impacts directly on the methodology used for word class tagging in the two languages: the low type/token relationship in Zulu necessitates the use of an additional tool, such as a morphological analyser prototype as described in Pretorius & Bosch (2003), to achieve a higher percentage in the automatic tagging of the Zulu corpus. Compare the following examples which have been analysed by the above mentioned analyser:

amanzi ”water/that are wet“
a[NPrePre6]ma[BPre6]nzi[NStem]
a[RelConc6]manzi[RelStem]

yimithi "they are trees"
yi[CopPre]i[NPrePre4]mi[BPre4]thi[NStem]

ngomsebenzi "with work"
nga[AdvForm]u[NPrePre3]mu[BPre3]sebenzi[NStem]

bangibona "they see me"
ba[SC2]ngi[OC1ps]bon[VRoot]a[VerbTerm]

abathunjwa "(they) who are taken captive/they are not taken captive"
aba[RelConc2]thumb[VRoot]w[PassExt]a[VerbTerm4]
a[NegPre]ba[SC2]thumb[VRoot]w[PassExt]a[VerbTerm4]

Examples with more than one analysis exhibit morphological ambiguity which in most cases, can only be resolved by contextual information. Nevertheless, a morphologically analysed corpus provides useful clues for determining word class tags, since the output of the morphological analysis is a rich source of significant information that facilitates the identification of word classes. For example, the above morphologically analysed words lead to the following information regarding further processing on word class level:

Table 5: Zulu morphological analysis and word classes.

Output of morpho-logical analysis	Word class	Examples
[NPrePre] and/or [BPre] + [NStem] + …	NOUN	amanzi
[CopPre] + [NStem] + …	COPULATIVE	yimithi
[SC] + [VRoot] + … OR [NegPre] + [SC] + [VRoot] + …	VERB	bangibona abathunjwa
[RelConc] + …	QUALIFICATIVE	abathunjwa; amanzi
[AdvForm] + …	ADVERB	ngomsebenzi

Concerning the tags used in the above morphological analysis, it should be noted that "tags were devised that consist of intuitive mnemonic character strings that abbreviate the features they are associated with." (Pretorius & Bosch, 2003: 208).

The word class tagset for Zulu is based on the classification by Poulos and Msimang (1996: 26). More will be said about this tagset further on in the discussion. The features and tags concerned are as follows:

Table 6: Zulu tags - illustrative excerpt.

Tag	Feature
[AdvForm]	Adverbial formative
[BPre6]	Basic prefix class 6
[CopPre]	Copulative prefix
[NegPre]	Negative prefix
[NPrePre6]	Noun preprefix class 6
[NStem]	Noun stem
[OC1ps]	Object concord 1st pers singular
[PassExt]	Passive extension
[RelStem]	Relative stem
[SC2]	Subject concord class 2
[VRoot]	Verb root
[VerbTerm]	Verb terminative

In this article it is argued that the difference in writing systems dictates the need for different architectures, specifically for a different sequencing of tasks for POS-tagging in Northern Sotho and Zulu. Approaches followed to implement word class taggers for Northern Sotho and Zulu will be presented in the following section.

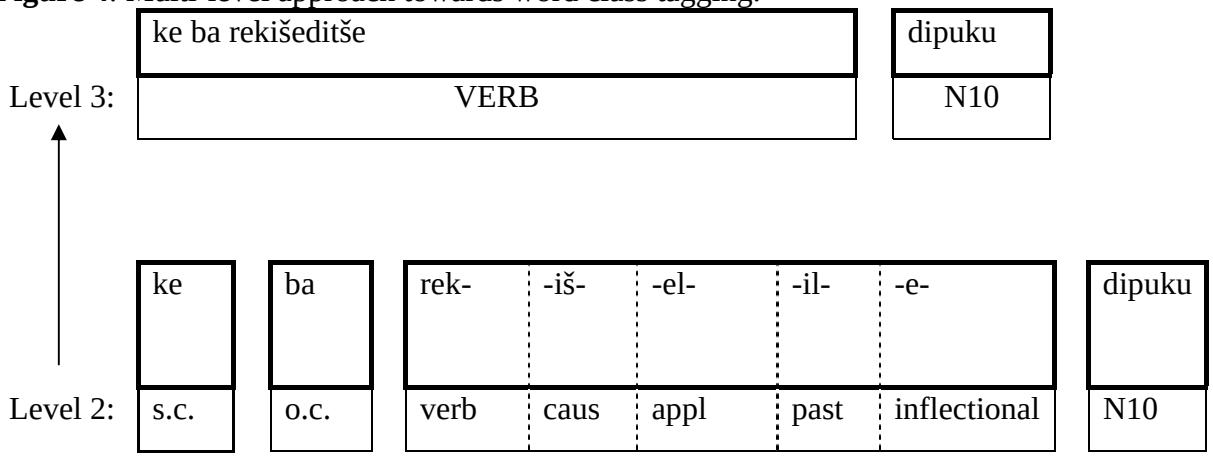
4. Comparison of Approaches Towards Word Class Tagging for Northern Sotho and Zulu

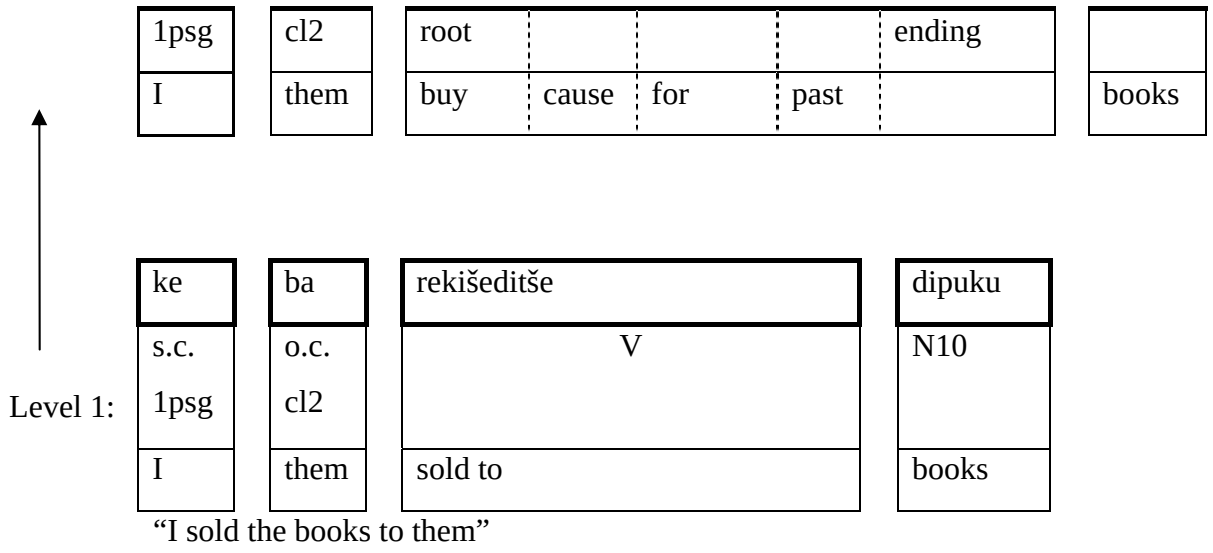
With regard to Northern Sotho, the term POS-tagging is used in a slightly wider sense, following Voutilainen (Mitkov, 2003: 220) who states that POS-taggers usually produce more information than simply parts of speech. He indicates that the term "POS-tagger" is often regarded as being synonymous with "morphological tagger", "word class tagger" or even "lexical tagger". POS-tagging for Northern Sotho results in a hybrid system, containing information on both morphological and syntactic aspects, although biased towards morphology.

This approach is dictated at least in part, by the disjunctive method of writing, in which bound morphemes such as for example verbal prefixes show up as orthographically distinct units. As a result, in Northern Sotho, orthographic words do not always correspond to linguistic words, which traditionally constitute word classes or parts of speech. Rather than to see this as a disadvantage, it was decided to make use of the morphological information already implicit in the orthography, thus doing morphological tagging in parallel to a more syntactically oriented word class tagging. It is therefore not necessary to develop a tool for the separation of morphemes, since this is largely catered for by the disjunctive orthography of Northern Sotho. As a result, all verbal prefixes can for example be tagged by making use of standard tagging technology, even though they are actually bound morphemes belonging to a complex verb form. A further motivation for the tagging of these bound morphemes, is the fact that they are grammatical words or function words, belonging to closed classes, which normally make up a large percentage of any Northern Sotho corpus. Tagging of these forms would therefore result in a large proportion of the corpus being tagged, alhough many forms will be ambiguously tagged. The decision to annotate all orthographically distinct surface forms, regardless of whether these are free or bound morphemes, resulted in a tagset which is rather larger than normal – even though only 9 word classes are traditionally distinguished for Northern Sotho, the proposed tagset contains 141 tags. This number is due to the distinction of class-based subtypes for some of the tags: the category PROEMP (emphatic pronoun) for example, has 17 subtypes in order to account for the pronouns of the first and second person, as well as those of the different noun classes. (For a full discussion of the tagset design, see Prinsloo and Heid, 2005)

However, the existence of complex morphological units whose parts are not realized as surface forms necessitates a multi-level annotation. A separate tool such as a morphological analyser would be needed for the analysis of inter alia verbal derivations of Northern Sotho. Typical examples that would need to be analysed by such a tool would be verbal suffixes. Such a multi-level approach could be represented as follows:

Figure 4: Multi-level approach towards word class tagging.





"I sold the books to them"

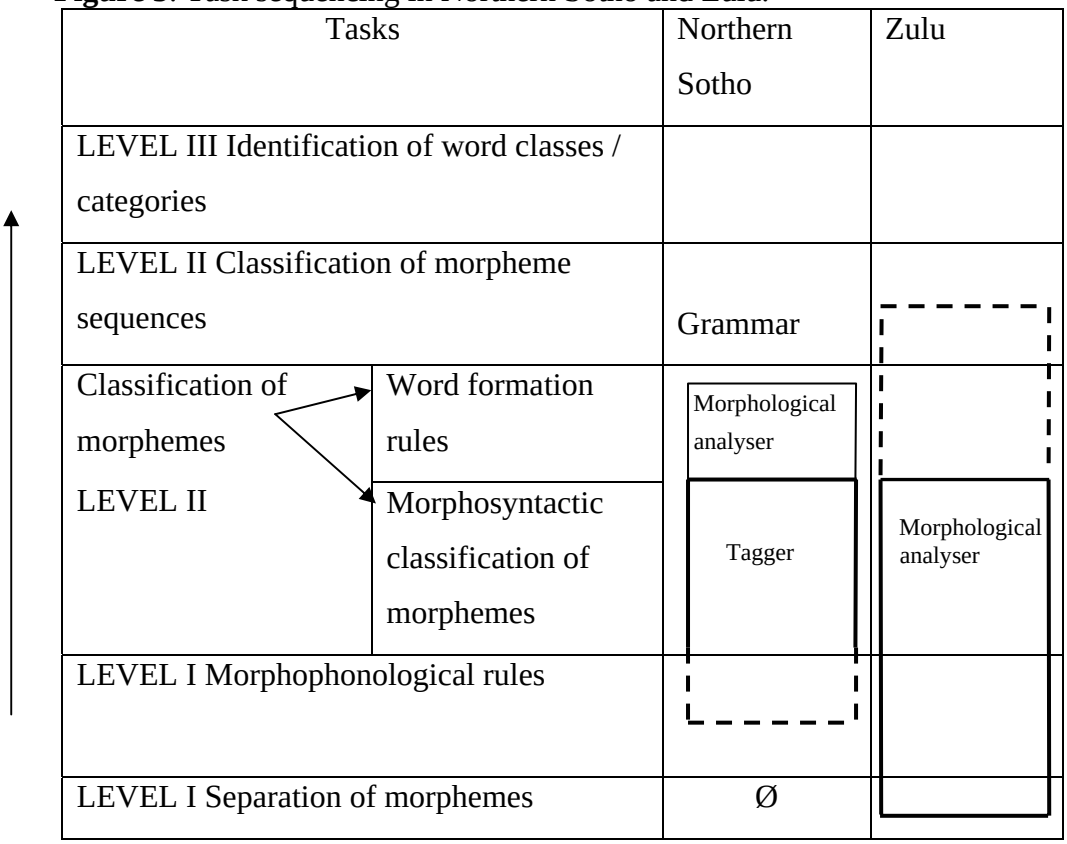
It should be noted that there are cases where the object concord appears within the verbal structure, notably the object concord of the first person singular. This particular object concord distinguishes itself from other object concords in that it is phonologically and orthographically fused to the verbal root. All other object concords are written separately from the verbal root and are thus easily identifiable, except for the object concord of class 1 before verb stems commencing with *b-*, e.g. *mo + bona > mmona* "see him/her". A procedure similar to the one illustrated above would be needed for these cases.

In the case of Zulu, morphological aspects need not be included in the word class tagging since these are already accounted for in the morphological analysis. This difference in approach to the tagsets can be mainly ascribed to the different writing systems. The word class tagset for Zulu used for purposes of illustration above, is based on the classification by Poulos and Msimang (1996: 26), according to which "words which have similar functions, structures and meanings (or significances) would tend to be classified together as members of the same word category…". The tagset comprises the following: Noun; Pronoun; Demonstrative; Qualificative; Verb; Copulative; Adverb; Ideophone; Interjection; Conjunction; Interrrogative. It is well-known that the degree of granularity of a tagset should be appropriate to the purposes of the tagged corpus (Allwood et al., 2003: 230).

The following diagram is a summary of the distinct approaches towards word class tagging as exemplified in the two Bantu languages, Northern Sotho and Zulu. The tasks that need to be performed are similar, but the approaches and sequencing of tasks differ significantly. It is noticeable that in Northern Sotho no dedicated tool is needed for the separation of morphemes, since this is already implicit in the disjunctive writing system. The tagger caters to a certain extent for morphophonological rules, but is especially significant for the second level where morphosyntactic classification of morphemes takes place. Analysis of word formation rules would only need to be done on level II, for which a morphological analyser is needed.

In the case of Zulu, the morphological analyser plays a significant role on levels I and II where constituent roots and affixes are separated and identified by means of the modelling of two general linguistic components. The morphotactics component contains the word formation rules, which determine the construction of words from the inventory of morphemes (roots and affixes). This component includes the classification of morpheme sequences. The morphophonological alternations component describes the morphophonological changes between lexical and surface levels (cf. Pretorius & Bosch, 2003: 273–274). Finally, Northern Sotho and Zulu are on a par on level III, where the identification of word classes, associated with the assigning of tags, takes place.

Figure 5: Task sequencing in Northern Sotho and Zulu.



5. Conclusion and Future Work

In this article a comparison of approaches towards word class tagging in two orthographically distinct Bantu languages, namely Northern Sotho and Zulu, was drawn. The disjunctive versus conjunctive writing systems in these two South African Bantu languages have direct implications for word class tagging. Northern Sotho on the one hand resorts to a hybrid system, which contains information on both morphological and syntactic aspects, although biased towards morphology. In the case of Zulu on the other hand, morphological aspects need not be included in the word class tagging since these are already accounted for in the morphological analysis. Word class tags for Zulu are

associated with syntactic information. The work described in this article is of crucial importance for preprocessing purposes – not only for automatic word class taggers of Northern Sotho and Zulu, but also for the other languages belonging to the Sotho and Nguni language groups.

Regarding future work, two significant issues have been identified. Firstly, cases of ambiguous annotation require the application of disambiguation rules based mainly on surrounding contexts. A typical example of ambiguity is that of class membership, due to the agreement system prevalent in these languages. For instance, in Northern Sotho as well as Zulu, the class prefix of class 1 nouns is morphologically similar to that of class 3 nouns, i.e. *mo-* (N.S) and *umu-* (Z). This similarity makes it impossible to correctly assign class membership of words such as adjectives, which are in concordial agreement with nouns, without taking the context into account. Secondly, the standardisation of tagsets for use in automatic word class taggers of the Bantu languages needs serious attention. A word class tagset based on standards proposed by the Expert Advisory Group on Language Engineering Standards (EAGLES), was recently proposed for Tswana, a Bantu language belonging to the Sotho language group, by Van Rooy and Pretorius (2003). Similarly, Allwood et al. (2003) propose a tagset to be used on a corpus of spoken Xhosa, a member of the Nguni language group. In order to ensure standardisation and therefore achieve reuseability of linguistic resources such as word class tagsets, this initial research on the standardisation of tagsets needs to be extended to all the Bantu languages.

ACKNOWLEDGEMENTS

We would like to thank Dr. Ulrich Heid (IMS, University of Stuttgart) for unselfishly sharing his knowledge and expertise with us. His comments on an earlier version of this article added immeasurable value to our effort.

REFERENCES

Allwood, J., Grönqvist, L. & 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.

Hurskainen, A., Louwrens, L.J. & Poulos, G. 2005.
Computational Description of Verbs in Disjoining Writing Systems. **Nordic Journal of African Studies** 14(4): 438–451.

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

Poulos, G. & Louwrens, L.J. 1994.
A Linguistic Analysis of Northern Sotho. Pretoria: Via Afrika Limited.

Poulos, G. & Msimang, T. 1996.
A Linguistic Analysis of Zulu. Pretoria: Via Afrika Limited.

Pretorius, L. & Bosch, S.E. 2003.
Computational aids for Zulu natural language processing. **Southern African Linguistics and Applied Language Studies** 21(4): 267–282.

Prinsloo, D.J. & Heid, U. 2006.
Creating Word Class Tagged Corpora for Northern Sotho by Linguistically Informed Bootstrapping. In: I. Ties (ed.), **Proceedings of the Lesser Used Languages and Computer Linguistics Conference (LULCL), Bolzano, 27– 28 October 2005**, pp. 97–115. EURAC Research.

Taljard, E. & Bosch, S.E. 2006.
A Comparison of Approaches to Word Class Tagging: Disjunctively Versus Conjunctively Written Bantu Languages. In: I. Ties (ed.), **Proceedings of the Lesser Used Languages and Computer Linguistics Conference (LULCL), Bolzano, 27– 28 October 2005**, pp. 117–131. EURAC Research.

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

Voutilainen, A. 2003.
Part-of-speech tagging. In: R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, pp. 219–232. Oxford: Oxford University Press.

Wilkes, A. 1985.
Words and word division: a study of some orthographical problems in the writing systems of the Nguni and Sotho languages. **South African Journal of African Languages** 5(4): 148–153.

About the authors: *Elsabé Taljard* is an 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.

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