Developing a tag set and tagger for the African languages of South Africa with special reference to Xhosa

Jens Allwood
Department of Linguistics
University of Gothenburg
Renströmparken
S-312 98 Gothenburg
Sweden
Visiting Scholar
Department of Linguistics
Unisa
jens@ling.gu.se

Leif Grönqvist
Department of Linguistics
University of Gothenburg
Renströmparken
S-312 98 Gothenburg
Sweden
Visiting Scholar
Department of Linguistics
Unisa
leifg@ling.gu.se

A P Hendrikse (Contact)
Department of Linguistics
University of South Africa
P O Box 392
Pretoria 0003
hendrap@unisa.ac.za

There are currently two distinct but not necessarily mutually exclusive approaches to the retrieval of information from linguistic corpora. 'Corpus-driven' approaches rely solely on the corpus itself to yield significant patterns. With the exception of orthographic spacing, no additional annotations to a 'raw' corpus are used to guide searches and the retrieval of information from the corpus. Typically, key word in context (KWIC) analyses are applied to relevant concordance lines to extract statistically significant lexical and grammatical patterns. In 'corpus-based' approaches, on the other hand, information is retrieved from an enriched corpus on the basis of annotations in the form of linguistic tags and annotations. That is, the annotations are used to direct the searches to specific grammatical and lexical phenomena in a corpus.

In this article, we propose a corpus-based approach and a tag set to be used on a corpus of spoken language for the African languages of South Africa. A number of problematic linguistic phenomena such as fixed expressions, agglutination, morphemic merging and spoken language phenomena such as interrupted words, etc, often have some effect on tagging principles. These problematic phenomena are discussed and illustrated. The development of the tag set is based on the morphosyntactic properties of Xhosa for reasons that are outlined in the article.

Manual tagging of a large corpus would be quite a daunting and time-consuming task, not to mention the potential for various kinds of errors. This problem is solved in a two-step process. Firstly, a computer-based drag-and-drop tagger was developed to facilitate the manual tagging of a so-called training corpus. This training corpus then forms the input to the development of an automatic tagger. The principles and procedures for the development of an automatic tagger for African languages are also discussed.

Introduction

The purpose of this article is threefold. Firstly, we would like to make the tag set that we have developed for the tagging of the text samples in the spoken language corpora publicly available for comments and utilisation by other colleagues also involved in the tagging of African language corpora. A sample of the tags representative of the range of morphosyntactic categories captured in the tags appears in an appendix to this article. Secondly, our article aims at drawing attention to the range of theoretical and practical issues and problematic phenomena that must be accounted for in the design of a tag set for the transcription of spoken language corpora of South African African languages. Finally, manual tagging in any corpus is a laborious, time-consuming and not very cost-effective task which is further complicated by the potential for typing errors and tag assignment errors. The development of automated taggers is the obvious solution to these problems and there are several different approaches to the development of such taggers. Our article therefore also gives an outline of the particular approach that will be subsumed in the development of a fully computational tagging tool. Before we deal with these issues in more detail a few general comments about the annotation of corpora would be in order.

Annotating a corpus – why?

Generally speaking, two approaches to the retrieval of information from a corpus are distinguished: the corpus-driven approach and the corpus-based approach (cf. Leech 1991, Hunston & Francis 2000, Tognini-Bonelli 1996). Typically, the partly corpus-driven approach uses word frequencies and the KWIC (Key Word In Context) analysis method to retrieve information about the associational patterns, distribution and significances of key words from concordance lines excerpted from a corpus. This approach is ideally suited for the corpus study of lexical patterns by means of computer concordance programs such as WordSmith Tools. But the corpus-driven approach is equally well-suited to grammatical pattern analysis. The major advocate of the corpus-driven approach to the study of grammatical patterns, Sinclair (1987, 1991), eschews any form of linguistic categorial annotation as an aid to the retrieval of grammatical patterns from corpora. He and his co-workers (cf. Sinclair & Renouf 1991) simply take the word-form defined by orthographic space as the key to colligational (the term used for patterned associations between grammatical units) searches. Leech (1991: 14) gives a very apt characterisation of the corpus-driven approach when he says that "[a]t one end of the scale, the

computer program (e.g. a concordance program) is used simply as a tool for sorting and counting data, while all the serious data analysis is performed by the human investigator".

The corpus-based approach, on the other hand, works from the assumption that "to make more linguistically interesting use of a corpus, it is necessary to analyse the corpus itself, and therefore to develop tools for linguistic analysis" (Leech 1991: 12). Annotated corpora serve as the platform for searches and the retrieval of various kinds of information, obviously depending on the level of information encoded in the annotation scheme (grammatical, conceptual-functional, discourse features, etc). The annotation of a corpus therefore is a way of adding value to the raw corpus. The decision on what kind of information should be encoded in the annotation therefore pivots on what kind of information one wants to retrieve from the corpus (cf. Leech 1999).

What to encode in annotation schemes?

Broadly speaking, many different kinds of information and significances can be captured in annotations schemes (cf. Van Halteren 1999). In our multilingual corpus project, we are primarily interested in enriching the corpus with morphological and syntactic information. There is a fairly widely used standard for the annotations of corpora, known as eXtensible Markup Language (XML). In the annotation of our corpora we use a format which conforms to XML.

This article concerns the morphosyntactic tagging of the transcriptions in our spoken language corpora. We have reported on the more general annotations used in the transcriptions of the corpora in another article in this volume (cf. Allwood & Hendrikse this volume). Two of these more general annotations warrant some comment here since they have a direct bearing on the morphosyntactic tagging of the transcriptions in the corpora. Firstly, spoken language in the multilingual context of South Africa (irrespective of the language in question) typically involves code switching (to varying degrees) and loan words. Since it would complicate matters unnecessarily to develop a tag set that accounts for the language used in code switching as well, we account for code switching instead by means of an annotation whereby the instance of code switching is enclosed in angle brackets and commented on in a comment line in the transcription. The same annotation procedure also applies to loan words that have not been (partially or fully) indigenised. Indigenised words present no problem; such words are simply treated as standard items of the language in question and tagged accordingly. For example,

loans such as *itafile* (Afrikaans: *tafel*) or *esitishini* (English: *station* 'at the station') are tagged as Xhosa expressions as illustrated below:

i«p9»tafile«n»

e«locgen»si«n7»tish«n»ini«locsuf»«adv»

Unconventional loans (i.e. those instances of mixed language use that are arbitrary and idiosyncratic) create rather peculiar tagging problems especially in cases where the vernacular morphology is integrated with an adapted form of a loan. Consider the example below where the vernacular morphology is tagged in terms of standard morphosyntactic categories, but the foreign material is left unanalysed and tagged as residual material.

Ndi«indIs»*ya*«ipresd»*claimish*«res»*a*«basicv»«v» (English: *claim* 'I am claiming')

The second phenomenon typical of spoken language that has a bearing on tagging is the contraction of expressions in spoken language. Once again, we use an annotational convention, namely curly brackets to indicate the corresponding written language elements which could then be tagged in the conventional way.

m«n1»*hlob{o}*«n»*{w}am*«poss1» (*mhlobam* 'my friend')

Some contractions are so regular that they have become fully lexicalised. Although it is still possible to analyse the morphological elements we treat such contractions both in the transcription and the tagging as unanalysable units. That is we do not recover the contracted elements with the curly bracket convention and thus do not tag these elements.

«n9»*kwedini*«n» (nkwenkwedini 'hey, youngster')

As we noted earlier, it is possible to tag a corpus for various linguistic levels. In this first spoken language corpus project on indigenous languages we have decided to follow the widely practised norm for initial tagging (cf. Cloeren 1999; Leech 1991, 1999 and Voutilainen 1999), namely morphosyntactic tagging. Cloeren (1999: 38), for example, notes:

> "At present, most corpora are raw or have been annotated morphosyntactically only. A reason for the focus on morphosyntax is its relative feasibility when compared to the added value for lexicography and grammar development."

This choice for the initial tagging of the transcriptions in our corpora does not preclude further tagging of the samples with tags for other levels and domains of research needs. For instance, the retrieval of information on semantic subclasses of nouns and verbs or on syntactic

constituents and hierarchies or on functional categories would require more advanced tagging of the corpora.

The design of a tag set – general considerations

Leech (1991: 24-25) highlights the following general, but very important, points regarding the development of tags which we quote here at length because of their significance.

> "[b] The devisers of such schemes of analysis generally seek to incorporate 'consensually approved' features such as (in the simplest case) traditional parts of speech. But ultimately, there is no such thing as a consensus analysis: all schemes are likely to be biased in some way or another – however minor – towards a particular theoretical or descriptive position.
>
> [c] At the same time, there is much to be said for a harmonization of different annotation schemes. As things are, tagging schemes and parsing schemes have arisen piecemeal, and if any standardization has taken place it has been no more than the *de facto* standardization accorded to a widely used scheme ... . It is widely felt that standardization of annotation schemes – in spite of its attraction in the abstract – is too high a goal to aim at; instead, our goal should be of annotation 'harmonization' – using commonly agreed labels where possible, and providing readily available information on the mappings, or partial mappings, between one scheme and another. Such a goal should be easier to attain in a flexible annotation system allowing for both hierarchies of annotation levels and degrees of delicacy in the specification of categories. (Spoken corpora may need special tags for speech-specific items.)."

As far as we know no tagging scheme or tag set exists for the South African African languages. One of the purposes of this article then is to document the morphosyntactic tag set that we have developed for these languages thereby making it available to the wider community of scholars serving as an input to the ultimate goals of consensus, harmonisation and standardisation suggested by Leech as in the quotation above.

In 1993 an initiative towards the setting up of standardisation guidelines for the development of cross-linguistic tag sets known as the EAGLES (Expert Advisory Groups on Language Engineering Standards) project was taken in Europe. Principles such as the following form the basis for the EAGLES guidelines towards standardisation (cf. Leech & Wilson 1999):

- Re-usability

A tag set as a general research resource should be usable by different end-users with varied research goals other than the researchers who developed the tag set. This requires that the tag set should be theory-neutral (as far as possible) so as to avoid the constraints imposed by a specific theoretical framework.

- Interchangeability and compatibility

  Labels in tag set should be sufficiently explicit and comprehensive with regard to the range of morphosyntactic categories that they encode so that mapping between different tag sets should be possible. This principle allows some freedom for the development of different tag sets by different researchers as long as there is sufficient compatibility between the categories and the granularity of scope and depth in the different tag sets so as to allow mappings across tag sets. This would also ensure compatibility between the tag sets for different languages in multilingual corpora.

Endeavours towards setting standards for the development of tag sets are not always favourably received by the scientific community as Leech & Wilson (1999: 57) note:

> "The disadvantages of standardization lie in its imposition of a 'straight-jacket' on scientific and intellectual endeavour. In linguistics, where the immense diversity of human languages continues to challenge and baffle the research community, any attempt to regiment the use of terms, categories or theories is likely to be anathemized."

Guidelines rather than standards seem to be the way whereby this kind of resistance towards standardisation could be overcome. Leech & Wilson (1999: 58ff) give an outline of the guidelines that went into the EAGLES proposals for standardisation. For lack of space we will not consider these guidelines in detail here. Suffice it to say that the development of tag sets for African languages urgently needs something similar to the EAGLES project. We would hope that the tag set proposed in this article and the considerations that guided its development could initiate the establishment of a standardisation project for Southern African African languages along the lines of the EAGLES project. Admittedly, this suggestion may seem like a call for the reinvention of the wheel. In the spirit of the principle of re-usability shouldn't the EAGLES guidelines, standardisation principles and suggested tag set for English not rather be adopted (with appropriate adaptations where relevant) for the African languages of South Africa? There are several reasons why the latter approach would not be feasible for the development of a tag set for the African languages of South Africa. For the purpose of this article a consideration of one of these reasons should be sufficient. The EAGLES project addresses languages that could be typologised as isolating languages. In contrast, the African languages of South Africa are typically agglutinating languages. Needless to say, no language is exclusively isolating or

exclusively agglutinative. Thus the languages which the EAGLES project accounts for, though overwhelmingly isolating, show varying degrees of agglutinative features as well. Similarly, the African languages of South Africa are essentially agglutinative, but also exhibit some isolating features. For example, most of the syntactic and functional significances that would be expressed in isolated words in, say, English, would be expressed in a single agglutinated morphological complex in Xhosa.

English: *He is not doing (this) for my benefit*
Xhosa: *Akandilungiseleli*

At a macro part-of-speech level one could tag the Xhosa expression as a verb, but this would be meaningless as this would simply leave most of the significant information unaccounted for. The Xhosa expression is at the syntactic level of analysis a complete sentence, a predicate phrase and a word, i.e. a verb in the sense that orthographic space would separate it from other words.

The EAGLES guidelines are, however, not without merit or usefulness even in a spoken language corpus project on African languages. In our discussion of the phenomena that we tried to account for in our tag set the EAGLES guidelines have therefore been taken into consideration. Before we come to this discussion a few comments on the principles that should guide the design of the actual coding of tags are in order.

Tag representation in text samples

According to Cloeren (1999: 49) there are several formats for the representation of tags "ranging from fully written out names through mnemonic letter-digit sequences to completely numerical labels". Each format has its own advantages and disadvantages regarding readability, transparency, cross-mapping between tag sets and so on. We have settled for a semi-mnemonic format in our tag set. That is, the tags encode, more or less, directly an abbreviation of the relevant (or at least part of) traditional categorial and morphological significances associated with the various morphosyntactic units. But the tag set is only semi-mnemonic in order to keep each tag unique and distinct from all the others. For example, the subject concords in Xhosa are typical portmanteau morphemes with a whole range of significances, such as mood, tense, number/noun class, polarity accumulated in them. To make a subject concord tag fully mnemonic of all these significances would entail a rather unwieldy tag. We tried to represent at least the most prominent significances which encompass also the other significances in the tags. Consider the tag for the indicative, present, positive, class 1 subject concord below:

u«ind1»……..

Since syntactic significances such as past and negative are generally considered the marked ones, their non-representation in the tag would imply the significances present and positive. Furthermore, since only the subject concords involve predicate significances such as mood or tense, the mnemonic symbol "ind" for the indicative mood is sufficient to suggest that this is a subject concord, rather than an object concord. Therefore these significance need not be represented in the tag, which would have been resulted in a tag such as the following if they were to be represented.

u«indpresposssc1»……..

The ordering of significances expressed in the symbols in a tag

As we have noted above inflectional (as well as some derivational) morphemes are very often portmanteau morphemes, i.e. various significances are cumulated in a single morpheme. In fact, this also applies to stem morphemes or, for that matter, word categories if one wants to distinguish sub-classes within the main lexical categories. For instance, one may want to sub-classify nouns into proper nouns, count nouns, mass nouns etc. and encode these sub-classifications in the make-up of the relevant tags. But this would entail some form of hierarchy in the structure of tags. The way to deal with this problem in isolating languages is the ordering of the constituent elements of a tag. For example *nprop*, *nmass*, *ncount* represent hierarchical relations between the category noun and its subclasses by ordering the categorial symbol before the sub-class symbol.

The tags in our tag set were structured on a different basis than the hierarchical relations between the encoded significances. We have, more or less, tried to capture the traditional terminology for the agglutinative morphosyntactic categories used in current grammatical descriptions of the African languages of South Africa (cf. Doke & Mofokeng 1974; du Plessis 1978). For example, subcategories of nouns such as proper names, mass nouns and count nouns are overtly represented in the noun class prefix system, but not necessarily in these terms. Proper names in Xhosa occur in noun class 1a, *uThemba* (Themba), but also kinship terms *umalume* (uncle) and loanwords *uloliwe* (train). Mass nouns occur in noun class 6, *amanzi* (water), but so do other subcategories of nouns, *amadoda* (men). The capturing of intra-categorial hierarchies typical of a language such as English tags would therefore not make much sense in the tags for nouns in an African language.

Needless to say, there are phenomena and linguistic categories for which there are no appropriate linguistic labels or no label at all. Furthermore, there are linguistic phenomena that occur in spoken language that have not been given any linguistic status in traditional terminology. We therefore had to create some forms of linguistic terminology in order to account for these phenomena in our tag set. For example, the significances of the morphemes *kwa-* and *na-* in the expression *kwanotata* (*kwa-no-tata* 'and also my father') have been noted in descriptive grammars, but without any appropriate grammatical term(s) for the morphemes. In this case we proposed the term **inclusive associative** with the tag «inclass». Similarly, the constructional pronoun, say, *ngokwethu* ('ourselves') involving the instrumental morpheme *nga-* ('by'), the relative formative *o-* ('which'), the possessive concord of noun class 15 *kwa-* ('of') and an appropriate possessive pronoun does not seem to have received a term in descriptive grammars of the African languages of South Africa. We proposed the term **reflexive pronoun** with the tag «reflproIp».

Spoken language typically involve two types of discourse management expressions, namely own communication management expressions such as hesitation markers *e.. e..* or change-of-mind expressions such as the Xhosa expression *mandithe ke* ('let me say', 'I mean') and feedback expressions. These expressions have not been properly described in the descriptive grammars of the African languages of South Africa partly because they are considered to be instances of disfluencies and therefore not grammatical expressions and partly because they do not belong to written language. Although various subcategories of these expressions could be distinguished, we use only two tags for these expressions due to the lack of knowledge about the range and subcategories of own communication management and feed back expressions. The two tags are «ocm» and «feedb». Hopefully, the searches of corpora for expressions tagged in this way will lead to an analysis and classification of these expressions as a result of which more refined tags could be developed.

The positioning of tags in a transcribed text

Once again, there are several possibilities (cf. Cloeren 1999). Interlinear tagging, whereby the tags appear on a line immediately below the relevant text, seems to be fairly common practice. The problem here is the alignment of the tags with the relevant elements in the text – a problem that is further exacerbated in agglutinating languages with excessive accumulation of significances in a single morpheme.

We have opted in our tagging practice for an approach in which each morpheme is immediately followed by its tag. Consider, for example, the tagging in the Xhosa expression *usesikolweni* ('he is at school')

u«ind1»*se*«locgen»*si*«n7»*kolw*«n»*eni*«locsuf»

The tag for the word class where appropriate appears at the end of the series of tagged morphemes in the orthographic word. Thus, the final tag in a tagged string always indicates the word category of the whole string. In some instances the final tag may thus follow another tag rather than a morpheme. Consider, for example, the tagging of the locative adverbial expression *emlanjeni* ('in the river') derived from the noun base *umlambo* ('river') where the final tag «adv» indicates the word category of the whole expression.

e«locgen»*m*«n3»*lanj*«n»*eni*«locsuf»«adv»

Special problems to be accounted for in a tag set for African languages

There is a whole range of phenomena that, one way or another, need to be addressed in the design of a tag set for the indigenous languages of South Africa. It is therefore perhaps best to deal with each one of these phenomena separately.

The language bias of the tag set

Our spoken language corpus project deals with multilingual corpora potentially involving 9 languages. Ideally, one should perhaps develop a tag set for each of the languages and then harmonise these tag sets into a generalised tag set for all these languages. This option was not feasible for the development of a tag set on account of the time-frame and budget constraints. Furthermore, the pervasive morphosyntactic similarities between the African languages of South Africa make it possible to develop a reasonably representative tag set on the basis of one language depending on the choice of the language.

A language from the Nguni branch of Southern Bantu seems to be a justifiable choice. The Nguni languages, particularly Zulu and Xhosa, seem to have had the advantage of the longest tradition of descriptive studies and their morphosyntax has therefore been fairly comprehensively described. But there are more compelling reasons for choosing a Nguni language as the basis for the development of a tagset. The Nguni languages, in contrast with the languages from other branches (e.g. Southern Sotho, Northern Sotho and Venda) retained the preprefix in the noun class prefix:

Zulu: *umuntu* (‘person’)

Southern Sotho: *motho* (‘person’)

The occurrence of the preprefix in nouns has a profound effect on the morphophonological shape of certain morphemes because of the preferred CV syllable structure. A whole range of phonological processes such as vowel elision, coalescence and the insertion of a semivowel or a consonant has rather significant ramifications for tagging. If, say, Southern Sotho was used as the basis for the development of a tag set this problem would not have been encountered and therefore not accounted for in the tagging guidelines. We return to this specific issue with illustrations further on.

The Nguni languages also show wider ranges of morphosyntactic distinctions than, for instance, the Sotho languages. In locative expressions of a certain type, the Sotho languages use a suffix only while the Nguni languages use both a prefix and a suffix:

Southern Sotho: *se*«n7»*fate*«n»*ng*«locsuf»«adv» (‘in the tree’)

Xhosa: *e*«locgen»*m*«n3»*th*«n»*ini*«locsuf»«adv» (‘in the tree’)

Finally, the conjunctive orthographic tradition followed in the Nguni languages as opposed to the disjunctive tradition followed in the other languages seems to be a closer approximation in the orthographic representation of the nature of agglutinating languages. In agglutinating languages bound morphemes are normally adjoined in the orthography. In the Sotho languages this orthographic principle is violated in the case of certain prefixes, though not in the case of suffixes. In a sense, the disjunctive orthography used for prefixes is simply a practical orthographic means used to prevent the violation of the CV syllable structure. Unfortunately, this entails that orthographic space is used between bound morphemes. Orthographic space is, however, very often used in computer-based tools to retrieve information about words which would normally be separated by means of such spaces. In order to prevent the incorrect application of such tools in the case of a disjunctively written language one would have to get rid of the spaces between bound morphemes first. Thus the Southern Sotho expression *o a tsamaya* where the morphemes *o* and *a* are bound morphemes that should be have been adjoined to *tsamaya,* which in this case is not a free morpheme either, will have to be orthographically amended to *oatsamaya* in order to enable the correct application of certain computer-based text tools. Even in an isolating language such as Afrikaans unacceptable vowel clusters resulting from the adjoining of bound morphemes which may cause misreadings of the adjacent vowels are not resolved by means of a disjunctive writing convention, but rather by means of a diaeresis.

Afrikaans: *geëer* (‘honoured’)

Among the Nguni languages, Xhosa seems to have retained finer granular distinctions in its morphological paradigms than the other Nguni languages. For example, in the copulative construction Zulu uses mainly two generalised copulative morphemes, namely *ng-* for noun classes with preprefix *u-*, *a-*, or o- and *y-* for noun classes with preprefix *i-*, or alternatively, simply a low tone on the preprefix of a noun irrespective of the noun class. In Xhosa, however, class distinctions are largely maintained in the copulative morphemes.

Zulu: *ngumuntu* or *umuntu* (‘it is a person’)
ngamadoda or *amadoda* (‘they are men’)
ngobaba or *obaba* (‘it is father and company’)
yinja or *inja* (‘it is a dog’)
yisitsha or *isitsha* (‘it is a dish’)
yizinja or *izinja* (‘they are dogs’)

Xhosa: *ngumntu* (‘it is a person’)
sisitya (‘it is a dish’)
zizitya (‘they are dishes’)
zizinja (‘they are dogs’)

Apart from these rather insignificant differences between Xhosa and Zulu, these two languages are so similar in their morphosyntax that it would not have made much of a difference to have used either one as the basis for the development of the tag set. Nevertheless, we chose Xhosa as basis for the development of the prototype tag set. Needless to say, the tag set should be able to account for the morphosyntax of the 9 African languages in the spoken language corpus, with, hopefully, minor adjustments according to the idiosyncrasies of individual languages. We expect that the need for such adaptations and modifications will become clear when we begin to do the tagging of text samples from all the languages in the multilingual corpora on the basis of the prototype.

The morphosyntactic granularity of the tagset and of tagging

Depending on the purpose of the tagging of a corpus, tag sets differ with regard to the morphosyntac scope and depth, also known as the granularity, encoded in the tags (cf. Cloeren 1999; Leech & Wilson 1999; Voutilainen 1999). Some tagsets will only encode the fairly generally accepted parts-of-speech such as noun, verb, adjective, etc. Others may encode morphosyntactic distinctions at the syntagmatic level, i.e. a tag for each syntagmatic slot in the

linearly ordered sequence of morphemes in a morphological complex bounded by space, for example preprefix-(gender/number)prefix-stem-(diminutive)suffix for the Xhosa morphological complex, *a-ba-ntw-ana* ('children') with tags such as *a*«prepref»*ba*«pref»*ntw*«nstem»*ana* «dim».

The granularity of our tag set goes a bit further than the two levels described above, namely it also distinguishes tags at the paradigmatic level. That is, the tags encode all the varieties within a paradigmatic slot. For the sake of clarity let us take a closer look at how the different levels mentioned here differ with respect to granularity and why we regard the degree of granularity of our tag set appropriate for our purposes.

Level 1: Parts of speech tags.

At this level the major syntactic categories of a language would be encoded in the tag set. There will therefore be tags for word categories such as Noun, Verb, Adjective, Adverb, Pronoun and Auxiliary. Subcategories within each major category will not be represented in this tag set. For example, in the category Pronoun different types of pronouns, such as Absolute Pronoun, Possessive Pronoun and Demonstrative Pronoun will not be distinguished. This tag set will therefore be very general and probably suitable for most languages, but not very rich in morphosyntactic information. Obviously, the tag set could be enriched with various levels of subcategories, but the more detailed the differentiations, the less language-inclusive the tag set will become.

The Xhosa example above would in this approach simply be tagged as *abantwana*«N»

Level 2: Syntagmatic morphological categories

In an agglutinating language the sequential slots within a morphosyntactic unit would be encoded in the tag set. At the most general degree of granularity in this case prefixal, stem and suffixal morphemes may be distinguished. Once again, this would account for most languages, i.e. agglutinating as well as isolating types but not yielding particularly interesting morphosyntactic information. One could increase the degree of granularity by tagging each one of the syntagmatic slots in a morphologically complex sequence. Thus the Xhosa example above could either be tagged as *a*«pref»*ba*«pref»*ntw*«nstem»*ana* «suf» or as *a*«prepref»*ba*«pref»*ntw*«nstem»*ana* «dim».

Level 3: Paradigmatic distinctions

Within specific syntagmatic morphological slots one may find internal paradigmatic distinctions. A very obvious example is the various noun class prefixes of, say, Xhosa. While on Level 2 above, only the morphological category, prefix, is represented in the tag set, the various distinct morphological instantiations of this category are also

distinguished in the tag set of this level. Thus, instead of the tag «pref» we now find a distinct tag for each one of the noun class prefixes listed below:

um- «n1»
aba- «n2»
um- «n3»
imi- «n4»
ili- «n5»
ama- «n6»
isi- «n7»
izi- «n8»
in- «n9»
izin- «10»
ulu- «11»
ubu- «14»
uku- «15»

Unfortunately, this led to a rather ungainly tag set as is to be expected given the range of noun class distinctions and concomitant concords typical of the African language of South Africa. It would be only natural to question this degree of granularity in a tag set. There is some justification for this level of granularity in our proposed tag set. All languages, particularly the spoken language forms of languages, are constantly changing for language-internal reasons (e.g. language acquisition) but also because of external influences (such as cross-linguistic influences). Some spoken language forms of Xhosa such as verbal negative forms used with certain adjective stems have not been recorded in grammars because they are considered non-standard forms of Xhosa. Thus, the forms *akadanga* ('he/she is not tall') and *akabanga* ('he/she is not pretty') rather than the written forms *akamde* and *akambi* occur freely in spoken language.

Similarly, there are indications of the reduction of noun class distinction in the African languages of South Africa, particularly in spoken language. Unless the tag set distinguishes the various morphological forms within a particular morphological category, facts such as those mentioned above will remain unrecorded and unacknowledged. If the spoken corpora that we are developing are to serve a meaningful role in language development and other applications, we do indeed need this level of morphosyntactic detail captured in the tag set.

What has been said above of the granularity of morphosyntactic differentiation in different tag sets applies equally well to the actual tagging of text samples. For example, to what extent should one assign separate tags to the constituent elements of a compound name such *South Africa*? Should one tag *South* as adjective and *Africa* as proper noun or should one simply treat the compound as a unit and tag it as proper noun? In the African languages of South Africa this problem is even more complex. Very often words can function as several parts of speech. For example, the exclusive quantifier *kodwa* ('only') in Xhosa functions also as the conjunctive 'but'. Or the infinitive form *ukuba* ('to be'), which, incidentally, is also a class 15 noun, functions as the conditional conjunctive 'if' and the complementizer 'that'. To what degree should these various morphosyntactic relations between these functions be represented in the tagging of a corpus? Once again the answer lies in the purpose and research aims envisaged for the tagged corpus. We have settled for tagging according to the relevant function of the expression in question in a particular context, ignoring all the other possible taggable information. Thus, we will simply tag *ukuba* ('that'/'if') as a conjunctive or as an infinitive depending on the context, without regard for the fact that the conjunctive form may have originated from the infinitive form which may have originated from a class 15 noun. In the latter case the conjunctive use of *ukuba* will have to be tagged as:

[[*u*«p15»*ku*«n15»]«inf»*ba*«v»]«conj»

This degree of granularity lies outside the scope of our project, but may very well be useful in a corpus-based study of grammaticalisation processes.

Unfortunately, this procedural simplification cannot be applied consistently. Important inflectional and derivational information would be lost with such coarse granular tagging in instances that Cloeren (1999: 44) calls multi-unit tokens and multi-token units. For instance, the Xhosa locative expression, *esikolweni* ('at the school') is an example of a multi-unit token. It could be simply tagged as an adverb, but only at the expense of significant morphosyntactic information, namely the derivational processes whereby the locative prefix *e-* and the locative suffix *–ini* are affixed to the noun of class 7, *isikolo* (school). This kind of information, we believe, should be explicated in the tagging.

On the other hand, multi-token units are somewhat anomalous in the sense that tokens of word categories are supposed to be individuated by means of orthographic space. There are several types of such multi-token expressions which we cannot discuss in detail here for lack of space. According to the tagging procedures followed in our corpus project we distinguish between two broad categories of multi-token expressions: fixed phrases and multi-token units that acquired

unitary categorial status. Fixed phrases are a reflection of what Sinclair (1987, 1991) ascribes to the idiom principle underlying language use. According to the idiom principle, a large amount of language use is simply a matter of managing non-compositional and thus memorised phrases, the best example of which is idioms, hence the name of the principle, but also phrases such as the English expressions *of course, as a matter of fact, you know* etc. Moon (1998) gives an excellent detailed survey of the range and degree of non-compositionality of fixed expressions in English. Ideally, fixed phrases should be tagged as if they were morphosyntactic simplexes. Clearly, this would presuppose extensive prior knowledge of such expressions or at least the possibility of lexical look-up facilities for such expressions and such facilities are unavailable for most languages and even more so for the African languages of South Africa. It is therefore currently not possible to appropriately tag such multi-token units in our transcription samples.

The other type of multi-token units, that is, phrasal expressions which have acquired categorial status, are tagged accordingly in our corpora. For this purpose we have introduced a specific convention whereby the various tokens of such units are linked by a subscripted line. For example, the Xhosa expression *kutheni na ukuze* (lit. 'it happened what that') is functioning as a non-compositional unit meaning 'why'. This expression will be tagged as *kutheni_na_ukuze*«q» where the subscripted lines indicate the unitary status of the linked tokens and the tag «q» that the unit is a question.

Morpheme merging and other syllable structure maintenance processes

Violations of the canonical CV syllable structure in the African languages of South Africa are resolved in various phonological ways, one of which is the coalescence of the vowels of adjoining morphemes. For example, the final vowel of the associative morpheme *na-* merges with the preprefix *i-* of the noun *inja ('dog')* in the expression *nenja* ('with the dog'). Tagging the underlying morphemes would be unnatural and impractical. The tagging of text samples in our corpora therefore assigns as closely as possible tags to the surface forms according to the canonical syllable structure. Thus, *ne* (CV) will be regarded as a variant of *na-* rather than *-e-* being considered a variant of the preprefix *-i-*. The tagged version of *nenja* will therefore be *ne*«ass»*n*«n9»*ja*«n»«n».

The same principle applies in the tagging of consonant insertions to keep the vowels of adjoining morphemes apart. That is, for tagging purposes the inserted consonant is considered part of the morpheme where the consonant in question helps to maintain the preferred syllable structure. Thus, in a locative predicative expression such as *basemthini* ('they are in the tree')

the consonant *s* keeps the vowel *a* of the concord *ba-* apart from the vowel *e* of the locative prefix *e-*. Following the principle discussed above, *-se-* will be considered a variant of the locative prefix *e-* and will be tagged accordingly, that is as *se*«locgen» which is the same tag used for the locative prefix *e-*.

The structure and organisation of the tag set

As we have said earlier, the fine granularity of the distinctions made in the tag set demands numerous tags. Needless to say, the number of tags is not going to be easily managed by human annotators of the corpora. One way of overcoming this problem is to develop a computerised tagger. We will return to this possibility and what it entails further on. Be that as it may, manual tagging is still required at various stages (tagging a training corpus and tag editing and corrections) even in an automated tagging procedure. We have tried to make the tag set more accessible by organising the tags according to the question of whether they are concordial or non-concordial. Furthermore, the tags are classified according to the major word categories with which the morphemes are associated, rather than the functional or derivational effect of the relevant morphemes. For example, although locative affixes turn nominals into locative adverbs, the tags for these affixes are grouped with the category noun, rather than with adverbs. Similarly, copulative prefixes should, strictly speaking, be classified with the verbal category (i.e. copulative verbs). However, since these prefixes typically occur with nominal lexical bases they have been classified accordingly. These unconventional groupings are simply a measure of making the tag set more accessible for manual tagging.

The development of an automatic tagger

As we have mentioned before, it would be virtually impossible to tag the corpora for the various languages manually. In this section we give an overview of the approach that will be followed in the development of an automatic tagger that will be based on the tag set presented in this article.

All automatic taggers have one thing in common, namely they all rely on linguistic regularities. One possible way of identifying such morphosyntactic regularities is to tag a substantial part of the corpus by hand. This tagged sub-corpus is called the training corpus. The development of an automatic tagger proceeds in two phases, both of which are performed automatically:

1. **The training phase**: A statistical procedure identifies enough useful regularities in the training corpus and store them in an effectively searchable structure.
2. **The tagging phase**: Based on the results from the training phase and the tag set, the prototype tagger places a tag in each tag position, in a way that should result in the highest possible tagging accuracy.

Part of the tagging phase involves the testing of the accuracy of the tagger on a test corpus which is different from the training corpus. As part of the testing of the tagger, it performs a tokenisation procedure on the test corpus. Retraining of the tagger is done until all the tokens in the test corpus are properly identified.

What kind of tagger should we use?

The various approaches that may be subsumed for the development of a tagger can be divided into two main categories:

- **Stochastic tagging**: Based on estimated probabilities extracted from the training corpus, the probabilities for different tag sequences in a sentence from the test corpus may be calculated and the sequence with the highest posterior probability (depending on the training corpus and the test corpus), is selected by the tagger.
- **Rule based tagging**: During the tagging phase, rules for correcting incorrect tags are extracted from the training corpus. These rules are then used to correct a random (or guessed) tag sequence for the test corpus.

There are also ways of training a stochastic tagger without the need for training data, i.e. the Baum-Welch algorithm, but this is known to give unacceptable success rates (cf. Merialdo 1994).

Stochastic tagging

A stochastic tagger (Nivre & Grönqvist 2001) uses a so-called Hidden Markov Model (HMM) as a way to implement Shannon's noisy channel modelling of tagging. A tagged sequence of a linguistic expression serves as the input and the actual word sequence of the expression constitutes the noisy output of noisy channel model. Actually, this rather unintuitive way to model tagging is very useful and thanks to many other applications, there is an effective algorithm, called the Viterbi Algorithm whereby the most probable input sequence (the tags) is calculated on the basis of the output sequence (the words). Unfortunately, the tokenisation and the tag set designed for the African languages of South Africa do not fit very well into this

framework. The HMM consists of a number of states corresponding to tag-values (or possible pairs of tag-values), and we need probabilities for jumping between states (transition probabilities) and for generating (emission probabilities), i.e. an output symbol (a word or a part of a word). One condition for our model to be meaningful is that all tokens in the token sequences are on the same level. This is not the case in the corpus of South African languages. Here instead the following conditions hold:

- One token may be a part of a word or a complete word
- Parts of a word may be tagged, as well as the complete word
- Some tag values may be attached to a phrase (a sequence of words)

This problem could be solved by using an extended HMM containing different kinds of states, or some kind of multi-layer HMM which could account for the different levels of tokens. However, since an HMM tagger is not very flexible as regards different kinds of morphosyntacic regularities in a language we have decided to consider other alternative approaches.

Rule based tagging

A rule based tagger (Brill 1992) will be capable of overcoming the problems of tokens on different levels noted above. The rule formats must simply be design in such a way that they will account for the structures in the corpus. Since we aim an automatic training of the tagger we have to make a choice between two basic approaches to the development of rule based tagging:

- Constraint based tagging (CG = Constraint based Grammar)
- Transformation based tagging (TBL = Transformation Based Learning)

According to the literature on this issue the best CG taggers (Karlsson et al 1995) use a large set (say, a 1000) of hand-written rules, which is too complex for effective machine learning. Experiments have shown that many of these rules may be learned automatically from a tagged corpus but the results are not as good as for the hand-written rules. TBL-rules, on the other hand, are difficult to write but the approach is very well suited to a machine-learning-task. Given that we aim at the automatic training of a tagger, it would seem that a transformation based tagging approach would suit our purposes best. Let us therefore take a closer look at TBL.

Transformation based tagging (TBL tagging)

The basic idea with TBL tagging may be described as follows:

1. For each token in the test corpus, insert the most probable tag without taking any contextual information into consideration.
2. Go through the list of rules and update the tags where the rules are applicable.
3. If any updates were performed in step 2, go back and run it again.

The training is a bit more complicated:

1. Write a set of rule templates of the type: "If the word at position *i-1* has the tag t_1 and the word at position *i+1* has the tag t_2, then change the tag for the word at position *i* to t_3."
2. Among the set of all possible rules generated by the set of rule templates, find the one with the highest score. The score is calculated as function of the number of updates and the number of correct updates, when the rule is run on the training corpus.
3. Go back to step 2 until a lowest score threshold is passed.
4. Keep the ordered set of rules

It is important to note that the rules might update tags to incorrect values but these will in many cases be corrected by later rules.

An example of TBL tagging

This example shows a simple part-of-speech tagger. Let us assume that we have trained a simple tagger with the tagset: noun, verb, pronoun, article, infinitive marker and preposition. Now when feeding this tagger with the small test corpus: "you have to book a chair on deck", step 1 in the tagging procedure will be to assign the most common tag to each word:

you	have	to	book	a	chair	on	deck
pronoun	verb	infinitive marker	noun	article	noun	preposition	noun

The initial tagging error for the word *book* will be corrected by a rule and hopefully no errors are introduced by the rules.

Designing a TBL tagger for Xhosa morphosyntax

In order to design a TNL tagger we have to:

- continue with the manual tagging of text samples to get a big enough training corpus. A part of the tagged corpus should also be left for evaluation later on;
- design the rule templates, which have to match the structure of the tokenized corpus;

- train and test a tagger, and try to find needs for more complicated rules.

The two last steps will be repeated several times until the results are good enough and the rules are simple enough not to make the training process too slow. The rule templates will then probably look much like Brill's original rules (Brill 1992) with extensions for the tags on different levels.

Conclusion

In this paper we have presented and discussed how to best develop a morphosyntactic tagset for Xhosa and potentially for all the nine indigenous languages of South Africa. We have suggested that, at present, this might best be done through a corpus-based rather than a corpus-driven approach.

We have presented a suggestion for such a tagset which will now be tested and tried for a spoken language corpus of Xhosa. We have also discussed different options in developing a computer supported tagger and found that probably transformation based learning will be the best approach.