On the development of a tagset for Northern Sotho with special reference to the issue of standardisation

E. Taljard, G. Faaß, U. Heid & D.J. Prinsloo
Department of African Languages
University of Pretoria
PRETORIA
E-mail: elsabe.taljard@up.ac.za
gertrud.faasz@up.ac.za
heid@ims.uni-stuttgart.de
danie.prinsloo@up.ac.za

Abstract

On the development of a tagset for Northern Sotho with special reference to the issue of standardisation

Working with corpora in the South African Bantu languages has up till now been limited to the utilisation of raw corpora. Such corpora, however, have limited functionality. Thus the next logical step in any NLP application is the development of software for automatic tagging of electronic texts. The development of a tagset is one of the first steps in corpus annotation. The authors of this article argue that the design of a tagset cannot be isolated from the purpose of the tagset, or from the place of the tagset and its design within the bigger picture of the architecture of corpus annotation. Usage-related aspects therefore feature prominently in the design of the tagset for Northern Sotho. It is explained why this proposed tagset is biased towards human readability, rather than machine readability; this choice of a stochastic tagger is motivated, and the relationship between tokenising, tagging, morphological analysis and parsing is discussed. In order to account at least to some extent for the morphological complexity of Northern Sotho at the tagging level, a multilevel annotation is opted for: the first level comprising obligatory information and the second optional and recommended information. Finally, aspects of standardisation are considered against the background of reuse, of sharing of resources, and of possible adaptation for use by other disjunc-

tively written South African Bantu languages. It is not the aim of this article to evaluate the results of any tagging procedure using the proposed tagset. It only describes the design and motivates the choices made with regard to the tagset design. However, an evaluation is in process and results will be published in the near future (cf. Faaß et al., s.a.).

Opsomming

Die ontwikkeling van 'n stel annoteringsmerkers vir Noord-Sotho, met spesiale verwysing na standaardiseringsaangeleenthede

Tot dusver was die gebruik van korpora in die Suid-Afrikaanse Bantoetale beperk tot die ontginning van rou korpora. Die gebruiksmoontlikhede van hierdie tipe korpora is egter beperk. Die volgende logiese stap in enige toepassing van natuurliketaalprosessering is dus die ontwikkeling van sagteware vir outomatiese teksannotering. Die ontwikkeling van 'n stel annoteringsmerkers is een van die eerste stappe in korpusannotering. Die outeurs van hierdie artikel meen dat die ontwerp van 'n annoteringstel direk verband hou met die doel van so 'n stel, en die posisie daarvan binne die groter raamwerk van die argitektuur van korpusannotasie. Gebruiksaspekte staan daarom sentraal in die ontwerp van 'n annoteringstel vir Noord-Sotho. Daar word verduidelik waarom hierdie stel eerder vir menslike leesbaarheid as vir masjienleesbaarheid voorsiening maak; die keuse van 'n stokastiese annoteerder word gemotiveer, en die verhouding tussen tokenisering, annotasie, en morfologiese en sintaktiese analise word bespreek. Ten einde op annoteringsvlak gedeeltelik voorsiening te maak vir die morfologiese kompleksiteit van Noord-Sotho, is 'n veelvlakkige annotasie verkies waar die eerste annotasievlak verpligte inligting bevat, en die tweede vlak opsionele en aanbevole inligting. Ten slotte word aspekte rondom standaardisering beskou teen die agtergrond van herbruikbaarheid, die deel van hulpbronne en moontlike aanpassing vir gebruik deur ander disjunktief-geskrewe Suid-Afrikaanse Bantoetale. Dit is nie die doel van hierdie artikel om enige annoteringsproses waarin hierdie stel annoteringsmerkers gebruik word, te evalueer nie. Dit beskryf slegs die ontwerp en motiveer die keuses wat tydens die ontwerp van die annoteringsmerkstel gemaak is. 'n Evalueringsproses word tans onderneem en die resultate sal in Faaß et al., (s.a.) gepubliseer word.

1. Introduction

1.1 Context and objectives

South Africa is a relative newcomer to the field of HLT, therefore the pool of expertise and skills in this regard is still rather small. Hence it is of the utmost importance that existing expertise is utilised in the most effective manner. This is particularly important with regard to the development of software tools to be used in NLP applications; it is similarly true of African language resources. Such resources and the pertaining tools are costly to produce and thus reusability should be one of the main concerns in their development.

Within the broader South African perspective, the use of electronic corpora is no longer a novelty. The compilation of electronic corpora for the eleven official languages was started during the late nineties at the University of Pretoria, the initial aim being to utilise these corpora mainly for lexicographic purposes (cf. De Schryver & Prinsloo (2000) in this regard). The University of Pretoria Sepedi Corpus (PSC) being one of these organic corpora, currently stands at about 6,2 million words. Since their initial conception, these corpora have been used in many different applications, e.g. the compilation of wordlists used for the building of spellcheckers, linguistic and terminological research, and translation studies. Raw corpora, however, have limited application possibilities. Therefore the next logical step would be the annotation of these corpora in order to increase their (re)usability and multi-functionality (cf. Snyman *et al*., 2007).

1.2 On standardisation

With regard to corpus annotation, Leech (1997:5) points out that POS-tagging a corpus adds value to the original corpus since the resulting tagged corpus is a reusable resource that can be handed on to other users. Reusability in turn requires some kind of standardisation in order to enable researchers to exchange data and resources, such as annotated corpora. When an annotated corpus is reused for a purpose different from the original one, having a standard way of annotating the data (content-wise) and of representing the annotated text (formally) would contribute greatly to minimising the need for manual adaptation by the new user. This would prevent what Leech and Wilson (1999:55) call a “free-for-all” or “re-invention of the wheel” every time a new project is started. Secondly, standardisation should be aimed at integrating annotation with NLP components such as grammars and lexicons, and with tasks such as parsing, i.e. with a view to later use of the corpus in the NLP

pipeline. In the third instance, standardisation of annotation practices will facilitate cross-language usability – an aspect that is of particular importance in the multilingual South African setup. In a situation where NLP development needs to be done for eleven languages, it is only reasonable to expect that tools developed for one language should be usable for other languages as well, especially in the case of the nine South African Bantu languages, since all of these languages are genetically related and thus share common linguistic features. However, sharing of electronic resources and procedures for software development is only feasible if at least some measure of standardisation is adhered to.

Despite the strong case in favour of standardisation, Leech and Wilson (1999:57) also voice some caution in this regard. They warn that rigid adherence to the principles of standardisation may impose a "straitjacket on scientific and intellectual endeavour". Rather than imposing inflexible standardisation principles, Leech and Wilson (1999:57, 58) therefore advocate an approach in which provisional guidelines are set up, with the expectation that a standard will naturally evolve. They propose the offering of a "default specification which can be adopted where there are no overriding reasons for departing from it". The linguistic reality often makes standardisation an ideal worth striving for, but one difficult to attain. With particular reference to the South African situation, this has already been pointed out by Taljard and Bosch (2006), who indicated that different approaches to word class tagging are needed for Zulu and Northern Sotho. This is mainly necessitated by the difference in writing systems utilised by these two languages: Zulu is a conjunctively written language, whereas Northern Sotho makes use of a disjunctive writing system. Even so, although sequencing of procedures for POS-tagging might differ, tools and the procedures themselves might very well be interchangeable. The degree of interchangeability would naturally be highest between languages which are closely related genetically, and which share the same writing system. This implies that a tagset functioning on the morpheme level for Northern Sotho, should with minor adaptations also be useful for languages such as Tswana and Southern Sotho, and possibly with a larger degree of adaptation to Venda and Tsonga, which are also written disjunctively. Furthermore, provided that conjunctively written languages such as Zulu, Southern Ndebele, et cetera are pre-processed using a morphological analyser, at least the principles underlying the tagset will also be reusable for these languages.

We furthermore believe that annotation practices, which include the development of a tagset, cannot be divorced from the purpose of the tagset, and the place of the tagset and its design within the bigger picture of the architecture of corpus annotation and whatever computational treatment of corpora that may follow. In this article, we recapitulate a few principles and choices for work towards standard tagsets, as well as for tagset design, and we explain the choices we made in our particular context.

2. Design of a tagset for Northern Sotho

In this section, we concentrate on the general aspects of tagset design, of which most are language-independent, and try to establish to what extent these principles can be applied to Northern Sotho. To our knowledge, only one tagset has been described for this language, viz. by De Schryver and De Pauw (2007). This tagset shows a number of similarities with ours; however, it contains no information on noun class numbers – neither for any of the nominal categories (nouns, pronouns, adjectives, etc.), nor for any of the sets of agreement morphemes. In this particular instance we have opted for a higher degree of linguistic granularity, specifically with possible further applications of the tagset, e.g. grammar development in mind.

Tagsets for related languages, e.g. Setswana by Van Rooy and Pretorius (2003) in which the EAGLES standard has also been used as a basis, have indeed been described, however, with a different outcome (cf. paragraph 4). Allwood *et al*. (2003) also describe a tagset for Xhosa. However, a full description of the tagset is lacking, and, as far as we understand, it does not conform to the EAGLES standard. It would therefore be inappropriate to compare the approach followed by the said authors to ours.

As indicated above, we regard the inclusion of noun class information as an essential component of our tagset. As a result, the number of possible labels for each class-dependent element is multiplied by 13,[1] i.e. the number of noun classes defined in our tagset for Northern Sotho. However, elements belonging to the

1 Northern Sotho makes use of eighteen noun classes. However, in our tagset all locative classes are contained in the “LOC” class. We describe thirteen classes: 1 to 10, 14, 15, and LOC. Additionally, agreement information concerning persons is described with the PERS class. However, this category does only contain possible pronominals like pronouns and concords.

different noun classes do not have the same frequency of occurrence. They even differ with regard to distribution patterns, therefore a statistical tagger would need a fair amount of training data. Correct identification of especially those labels that have a low frequency of occurrence therefore necessitates the introduction of a rule-based component (cf. Faaß *et al*., s.a.).

2.1 Principles of tagset design

With regard to general guidelines for the design of a tagset, Leech and Wilson (1999:59) refer to the set of recommendations formulated by EAGLES, the former Expert Advisory Group on Linguistic Engineering Standards (http://www.ilc.cnr.it/EAGLES96/browse.html). They indicate that the choice of the device which is to visually encode any given linguistic phenomenon is an arbitrary one, but propose that the following criteria should be adhered to: *non-ambiguity*, *compactness*, *readability* and *processability*. They do make provision for the possibility that the priority assigned to any of these criteria might differ from one project to the next.

It is by now generally accepted that tagset design is a "trade-off between what is linguistically most desirable and computationally feasible" (Leech, 1997:25). For Northern Sotho, the aspect of linguistic correctness/desirability is in itself problematic, since linguists do not agree on the number of word categories to be distinguished for Northern Sotho, neither are they in agreement on the contents of the different word categories. To cite but one example: Van Wyk (Kosch 1993:61) classifies the possessive concord of Northern Sotho as a particle, thus assigning it the status of a linguistic word, whereas Poulos and Louwrens (1994:96) regard these forms as concords, i.e. bound morphemes. As a result, some of the linguistic distinctions contained in the tagset may seem rather arbitrary. Furthermore, a tagset compiled on sound linguistic principles may not always meet the criterion of, for example processability. Therefore high linguistic granularity needs to be balanced with ease of computational processing: it may turn out that a specific tag which embodies some grammatical distinction cannot be assigned automatically with any degree of accuracy. If priority were given in such a situation to processability, the better option would be to sacrifice some linguistic granularity in favour of ease of automatic processing. To illustrate: a number of verbal moods are distinguished in the grammars of the South African Bantu languages. For Northern Sotho a total of eight different moods are traditionally distinguished, the mood being morphologically encoded in the verb. However,

morphological marking of any particular mood is not always realised in the verb, and in many cases the surface realisation of verbs belonging to different moods is (coincidentally) identical. The verb *a ngwala* "while (s)he writes/then (s)he wrote" can for example be either in the situative (participial) or consecutive mood – one possible way to determine which of the possibilities is the correct one, is to make a semantic analysis of the discourse context in which the verb appears. Another is to take the context into account by searching for the conjunction *ge* "when" in the left cotext, as this conjunction may introduce the situative. However, the better option would probably be to disregard the modal distinction in the design of tags for verbs, since it cannot be automatically assigned with high precision, at least not on the tagging level.

Generally, there are two choices with regard to annotation of linguistic versus orthographic units. The first choice would be to use a morphological analyser as a preprocessing step to part of speech tagging, which would eventually result in linguistically accurate word labels. For a conjunctively written language such as Zulu, this is a mandatory step. Concerning the disjunctively written languages such as Northern Sotho, processing can start with an annotation of orthographic units, i.e. tokens, followed by a morphosyntactic analysis. (For a more detailed discussion, cf. Taljard & Bosch, 2006.)

From a computational perspective, the latter order seems to be less expensive, taking into account that, for example, a verb like *ke a mo rata* "I like him/her" would be analysed by the morphological analyser as one linguistic word, i.e. a verb, although it contains a functional syntactic unit, the objectival pronoun *mo*, anaphorically representing the object of the verb. A syntactic analyser, i.e. a parser would have to "extract" this objectival pronoun at a later stage in order to correctly analyse the verb and its object as two equal parts of one verbal phrase. An inclusion of elements to form a linguistically correct word, followed by an extraction of one of these elements in order to form a linguistically correct phrase, is computationally expensive. On the other hand, a lexicon-based annotation of all elements, irrespective of whether they are bound or free morphemes, enables us to, at a later stage, implement an effective morpho-syntactic analysis in one step.

We are conscious of the problem of the merged object concords and other fused forms (cf. 4.1). These will, however, not be handled during the part of speech tagging stage, but rather during the process of morpho-syntactical analysis.

The discussion that follows will focus on a number of usage-related aspects, specifically on the purpose for which the tagset is designed; on the place of tagging in the computational treatment of corpora; on ways to define tags from a linguistic point of view, as well as on formal aspects of tagsets and on ways to proceed in the creation of standardised tagging resources.

2.2 Usage-related aspects

Two issues are relevant in this regard: the first pertaining to ease of human processing (i.e. readability) vs. ease of machine processability – the second referring to the proposed use of the tagged material.

Most POS-tagging is performed with both machine use and human use in mind. This explains why Leech and Wilson (1999) require tagsets to be both human-readable (i.e. somehow mnemonic) and computer-processable (i.e. unambiguous). If linguists are to read tagged data, some resemblance of tag names with names of types of linguistic phenomena is an advantage. A tagset like the EAGLES intermediate tagset (cf. Leech & Wilson (1999) and http://www.ilc.cnr.it/EAGLES96/annotate/node9.html), however, is only constructed for machine use. Its tags are numeric, composed of numbers (0 and 1) indicating the presence or the absence of certain features. As tagset mapping is nowadays performed equally easily by a comparison of feature structures as by comparing numeric vectors, there does not seem to be an immediate need for an approach in line with the intermediate tagset of EAGLES: machine processability is also ensured with non-numeric tagsets. The criterion of readability therefore enjoys priority for this particular project. Compare the following sample of the tags devised to represent some of the major categories that are distinguished:

N02	Noun, followed by a numeral indicating class number
ADV	Adverb
CS05	Subject concord, followed by a numeral indicating class number
CO05	Object concord, followed by a numeral indicating class number

As indicated above, the design of a tagset is directly linked to its proposed use, therefore any discussion about human readability vs. processability of tagsets and tagged data should take the proposed usage of the tagset and of the tagged material into account, since

this may pose constraints on the shape of the tagset. If a tagged text is for example directly fed into a syntactic parser, compatibility between the tagset and the classification of single word forms in the parser's lexicon is needed. Similarly, if the tagged material is to be used in (interactive) corpus query, the possibility to use underspecified querying is a valuable attribute, e.g. searching for “C.*05” in material tagged with the attached list of tags provides any kinds of concord of class 05, i.e. subject as well as object (and possibly other) concords. This kind of underspecification is in turn supported by a logical tagset. See the discussion on underspecification in the section on linguistic and formal aspects below.

Tagging technology also plays an important role for tagset design: many statistical taggers work best with a tagset of a given size, mainly because they need training material for disambiguation purposes, and the size of the training texts needed minimally to allow the tool to learn probabilities, increases with the size of the tagset. This is caused by the fact that the number of tags is directly related to the degree of ambiguity of the elements of the lexicon. To illustrate: without the addition of the noun class number to the label *SC*, the subject concord *a* would simply be labeled *CS*. Adding the noun class number implies that it has to be ambiguously labeled *CS01:CS06*. However, occurrences of *a* as the subject concord of class 6 are far less frequent than occurrences as the subject concord of class 1. Additionally, the cotext of class 1 and class 6 may be often similar, hence a statistical tagger, as it works frequency- and cotext-based prefers labeling most occurrences of *a* with class 1. To avoid this situation, either training data containing a high number of class 6 occurrences of *a* must be added, or a rule-based component must be introduced as an intermediate tagging step to cover all clear-cut cases of *a* belonging to class 6. In other words, the bigger the tagset, the higher the number of labels that can be assigned to one token, and the bigger the size of the training data needs to be.

The proposed Northern Sotho tagset is also to be used for stochastic tagging, making use of Schmid’s TreeTagger (Schmid, 1994). The choice of Schmid’s TreeTagger was motivated by the fact that for a tagset of ca. 60-100 tags, it only needs 30 000 to 40 000 words as a training corpus, as opposed to most other stochastic taggers such as TNT (Brants, 2000) or MBT (Daelemans *et al*., 2003) which need a significantly higher number of words as training corpora. Unlike these taggers, the TreeTagger also makes use of an external lexicon informing the tool about all possible annotations of tokens

that do occur and tokens that might not occur in training data. Therefore, preprocessing steps (cf. 2.3) can include a guessing system to identify the correct label(s) for unknown words of a new text and addition of those entries to the lexicon, gaining a tagging recall of 100%. Choosing this particular tagger has, however, direct implications for the design of the tagset, one of them being a restriction on the size of the tagset. Prinsloo and Heid (2006) indicate that the current tagset for Northern Sotho which has 141 tags, represents the upper limit of possibilities of the TreeTagger. Proliferation of tags therefore needs to be avoided, since a tagset with a large number of tags may lead to data sparseness, i.e. a situation where a few tags occur very rarely in texts. This may affect the possibility for the tool to learn discriminative contexts of these rare tags. It thus affects the performance of statistical tagging tools and may also have a negative impact on the complexity of pattern-based parsing rules to be defined at a later stage of data processing, when a tool for syntactic analysis is designed. On the other hand, the morphological richness of Northern Sotho favours a rather large tagset, since "there is a general tendency for tagsets to increase in size proportionate to the richness of a language's inflectional morphology" (Leech, 1997:29). At least some measure of its morphological complexity should therefore be accounted for and reflected in the tagset.

Therefore, there is a need for accommodating the morphological complexity of Northern Sotho on the one hand, and the restriction on the size of the tagset dictated by the statistical tagger on the other hand. This need can be met by using multilevel annotation. We opted for a two-level system that consists of a variety of attributes for each orthographic token in a text. On the first level there are 141 different tags that can be assigned. These are used by the TreeTagger. The second level distinctions lead to 262 possible annotations, which account for at least part of the morphological complexity. Since most of these annotations are used for closed class items completely listed in the tagger lexicon, this information can be added in a later processing step by lexicon-based tagging. Consider as an example particles in this regard for which only one tag, PART is defined on the first level. This is the only tag that will be assigned by the TreeTagger. The appropriate feature, e.g. *que* (question particle), *agen* (agentive) or *con* (connective), will then be selected by means of lexicon-based tagging.

2.3 Tagset design within the context of corpus annotation

Two issues are relevant when discussing the place of tagging in the larger process of corpus linguistic treatment of any language: one is the relationship between tokenising, tagging, morphological analysis and parsing, and the other one concerns the tagging technology used. Both have major implications for tagset design.

An approach that has become almost classical in the computational treatment of European languages involves a pipeline-based sequential processing of texts by means of tokenising, tagging, morphological analysis and syntactic parsing. Our intention in some of the work presented here is to investigate to which extent such processing can be utilised for corpus annotation in Northern Sotho. This approach assumes the first elements of the process, i.e. tokenising and tagging to be "local" in scope, i.e. mainly covering word forms, whereas morphological analysis and especially syntactic parsing would be dependent on the clausal context. For tagset design, this implies that the designer should first analyse the situation with respect to the distribution of tasks between the different components of an analysis chain, before deciding on the kinds of distinctions to be made in a tagset. Some such distinctions may be irrelevant at the tagging stage, others may be redundant at earlier or later processing steps. Knowledge for introducing certain distinctions may lack at the time of tagging.

In order to position the design of our tagset within the broader context of corpus annotation, the project team opted for a bootstrapping approach known to be a successful strategy when beginning with few resources. Here, every processing step leads to an intermediate result which can already be utilised for research purposes. A brief overview of the annotation procedure is necessary. Compare Figure 1, taken from Prinsloo and Heid (2006) in this regard.

Figure 1: Architecture for the creation of training material and for the tagging of new texts

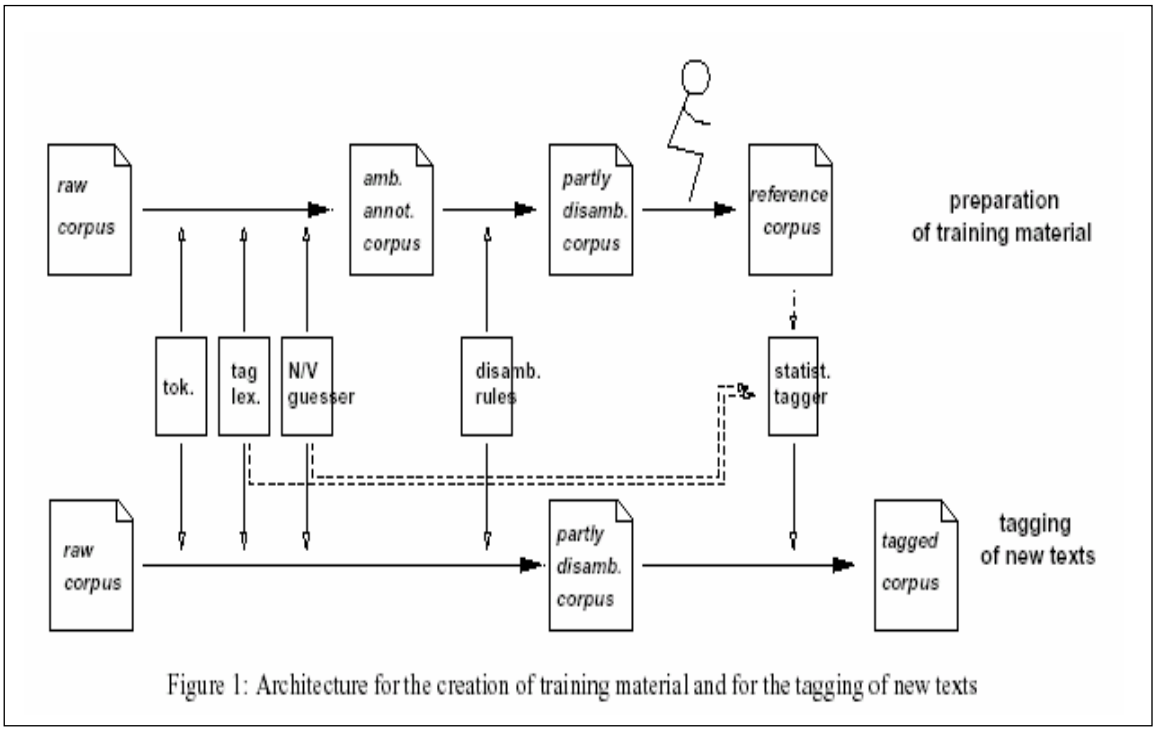


Figure 1: Architecture for the creation of training material and for the tagging of new texts

The starting point of the annotation process is a raw corpus consisting of about 40 000 tokens that is tokenised on a word form level. Thus, for the purposes of tokenisation, we interpret Northern Sotho's disjunctive writing system in a mechanical, non-morphological way. Tokenisation is followed by lexicon-based pretagging, using a tagger lexicon that currently contains about 7 000 known items and their annotations. This system's lexicon consists of a manually tagged inventory of all closed class items (concords, pronouns, conjunctions, etc.), a list of approximately 3 700 top-frequency verb stems extracted from the 6,2 million *PSC*, a manually tagged list of the 1 000 most frequent word forms from the same corpus, and a name lexicon, currently containing 335 personal and place names. Lexicon-based pretagging results in a partially and ambiguously tagged corpus. Items left untagged are assumed to be nouns or verbs, for these are open class items and are therefore not fully covered by the tagger lexicon. A specially designed noun and verb guessing tool is then used to guess the category of these remaining untagged tokens. Guessing relies mainly on singular: plural prefix matching, identification of nominal derivations and consideration of the syntactic environment for nouns, and longest string matching of suffixes for verbs. (See Prinsloo *et al*. (s.a.) and Heid *et al*. (s.a.) for a detailed discussion.) The pretagging and guessing procedures offer a list of possible annotations to be reviewed by the

user, and correct guesses are added to the tagger lexicon. Since ambiguous items are tagged as such (e.g. *ke* will be tagged as a subject concord of a person, a particle, a verb (auxiliary) and a copulative; *CSPERS:PART:V:VCOP*), the result of the pretagging and noun/verb guessing procedures is an ambiguously tagged corpus. Context dependent disambiguation rules may then be used for further automatic disambiguation (cf. Faaß, s.a.). The TreeTagger will then disambiguate any new ambiguously tagged corpora.

This bootstrapping strategy stands in contrast to other tagging strategies, e.g. the Brill-tagger, which firstly assigns the most frequent label, and in a later processing step, may modify this decision by applying linguistically motivated rules. However, already after our computationally "cheap" annotation of all possible labels, we have a fully and correctly, but ambiguously annotated corpus as an intermediate representation. Again, after utilising the statistical tagger (eventually enhanced by a rule-based component), we have a disambiguated corpus, at least on the first level of annotation. The lexicon-based enhancement of these annotations, which adds the second level, then leads to a fully annotated corpus. All intermediate results are usable for research.

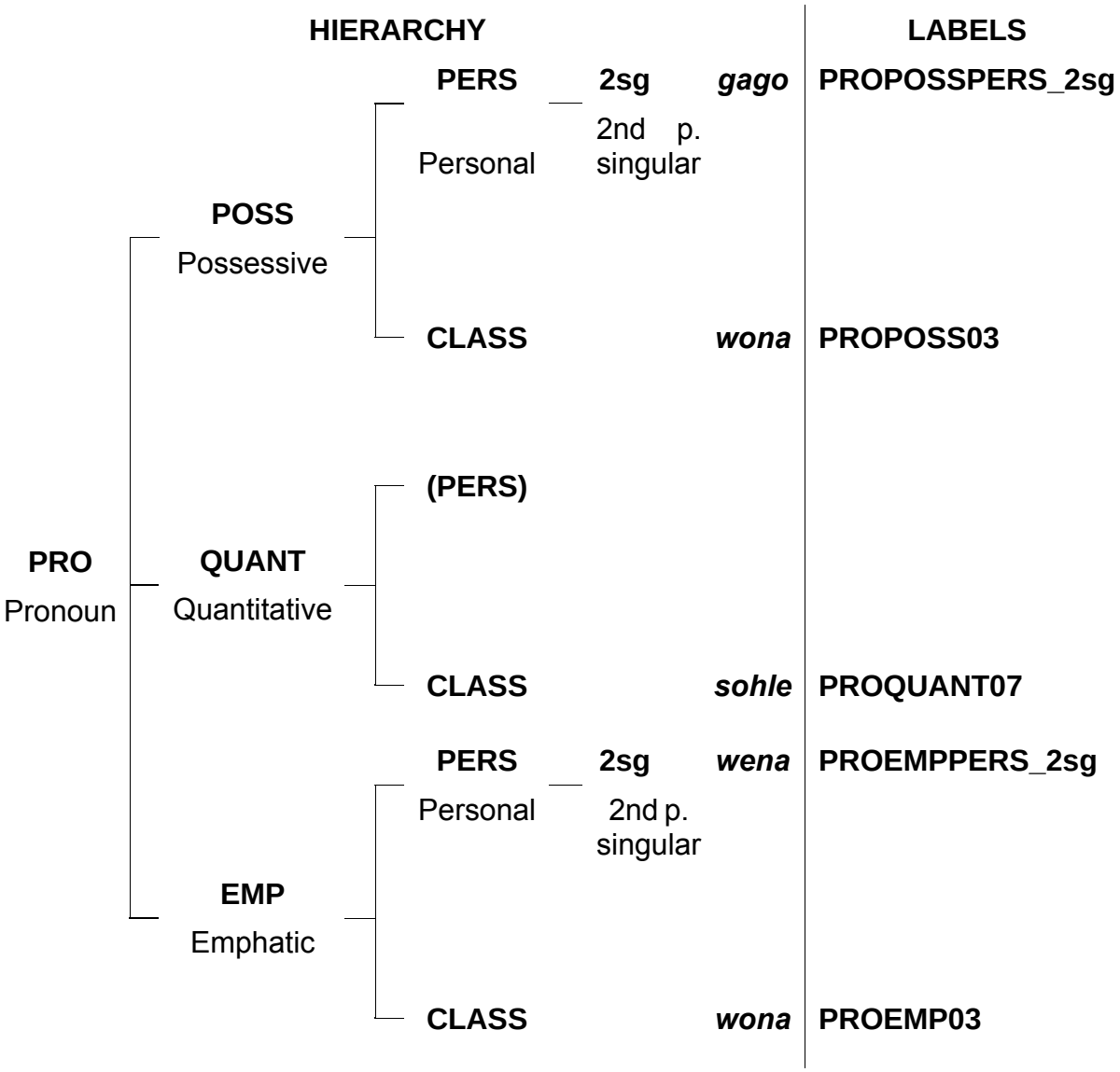
2.4 Linguistic and formal aspects of the design of a Northern Sotho tagset

Tagging can be seen as a task of classifying linguistic objects according to a set of linguistic criteria. For *POS*-tagging, such criteria are mostly (but not exclusively) related to morphosyntax:

- The most frequently used distinctive criteria are morphological criteria. Singular:plural pairings of noun classes is one such a distinctive feature, as indicated by the class prefixes, e.g. ***mo****humagadi* (*lady*) (sing.) (N01) vs. ***ba****humagadi* (*ladies*) (pl.) (*N02*).
- Lexical criteria are often introduced into tagsets next to morphological ones. In Northern Sotho, a distinction is made between emphatic and possessive pronouns (*PROEMP* vs. *PROPOSS*), despite the fact that these are morphologically similar.
- Distributional criteria: for certain forms, appearance in a given context is indicative of one reading, appearance in another context is indicative of another reading. The tagset therefore lists all possible annotations for each item, the lexicon-based tagging procedure hence results in an ambiguously tagged corpus.

With regard to the formal structuring, the Northern Sotho tagset is a logical tagset. This implies that the relations between the word categories can be represented as a hierarchical tree. Such an arrangement therefore reflects the relations between word categories. The attributes of a word category are inherited from one level of the hierarchy to the next. Compare Figure 2 in this regard:

Figure 2: Hierarchical representation of the word category *PRONOUN*



The word class of pronouns (*PRO*) is subdivided into possessive, quantitative and emphatic pronouns (*PROPOSS*, *PROQUANT* and *PROEMP* respectively). The system foresees a further subdivision according to the dimensions of noun classes (e.g. *PROPOSS01*, *PROPOSS14*, *PROPOSSLOC*, et cetera) and, for the emphatic and possessive pronouns, first or second person; the person values are only specified at the second level of the tagset, whereas the presence of the person dimension is expressed in the tag names of the first level, e.g. *PROPOSSPERS*. As there is no dedicated form to express a quantitative pronoun of the 1st or 2nd person, no corresponding tag is necessary, and the most specific tag for quantitative pronouns therefore is *PROQUANT*.

As already indicated, such a logical tagset supports underspecified queries, which is a valuable instrument in corpus processing. It is therefore useful to make sure that the tag naming convention mirrors the logical structure of the classification covered by the tagset. Furthermore, it is important to make sure that the hierarchy of the chosen names of the categories is logically structured. Naming all pronouns *PRO* with a further specification *POSS* for possessive, *QUANT* for quantitative and *EMP* for emphatic pronoun, followed by the class number, allows for an underspecified query in the annotated corpus, such as “search for PRO[POSS|QUANT|EMP][0-9]+” or just “PRO.+” to find all occurrences of pronouns (cf. Prinsloo & Heid, 2006).

It has already been pointed out by Taljard and Bosch (2005) that POS-tagging in Northern Sotho results in hybrid annotations, in which both morphemes and lexemes are tagged, without distinguishing between them. Due to the disjunctive method of writing utilised in Northern Sotho, there is not necessarily a one-to-one correlation between the word as orthographically distinct unit and the word as morphosyntactic unit. In many instances bound morphemes, e.g. verbal prefixes appear as orthographically distinct units. The disjunctive method of writing does offer a wealth of morphological information and it was therefore decided to do morphological tagging parallel to word class tagging. This is also reflected in the tagset – some tags refer to purely morphological features, whereas others are more syntactically oriented. Tags that refer to morphological features, are typically the ones used to annotate verbal prefixes, which are written disjunctively from the verb stem, e.g. subject concords (*CS*) and object concords (*CO*), as well as the present tense morpheme *a* (*MORPH_pres*). On the other hand, the tags *PROEMP* (emphatic pronoun) and *POSSPRO*

(possessive pronoun) reflect a distinction that is purely on the syntactic level. The advantage of this approach is that it can with very little adaptation also be applied to the other disjunctively written languages, i.e. Tswana and Southern Sotho, especially since these languages are so closely related to Northern Sotho. Tsonga and Venda, which also follow a disjunctive writing could also benefit from this approach.

3. Current version of the *POS* tagset for Northern Sotho

Hereafter, we discuss the current (September 2007) form and structure of our tagset for Northern Sotho. A listing of the complete tagset can be found in the appendix.

The tagset is mainly based on the lexical and morphological criteria defined by Lombard (1985) and Louwrens (1991). As described above, the logical structure of the tagset is divided into two layers of linguistic description (annotation levels):

- The first annotation level includes all mandatory, or, according to EAGLES, obligatory information, namely up to three elements: an element hinting at the word class, a second one specifying functional or syntactic properties, and a third one giving morphological specifics, cf. e.g. *PRO(noun)EMP(hatic)PERS(on)*.
- The second level of annotation includes recommended and optional information. This level is in most cases used for a detailed description of closed class items described in the tagger lexicon. Compare the following excerpt:

Figure 3: Annotation levels

Description	**Tag 1st level (mandatory information)**	**Tag 2nd level (optional/ recommended information)**
Pronouns:		
emphatic personal	PROEMPPERS	1sg, 2sg, 1pl, 2pl
Verbals:	V	--, aux
Morphemes:		
deficient	MORPH	def

At the topmost level, our tagset for Northern Sotho distinguishes nine different classes, e.g. concords, pronouns, nouns, adjectives, verbals, morpheme particles, question words, and others.

In addition to functional and lexico-semantic subtypes of these classes (cf. Appendix), the following five types of morphological and lexical distinctions are made; we decided to encode numbers 1 and 5 at the first level of the tagset (i.e. in the tag names used by the statistical tagger) and numbers 2-4 as second level features:

- **The class membership feature:**

The classes 01-15 and the locative classes 16-18 (*LOC*) are all assigned at the first level of annotation, except the so-called copulative subject concords, which actially function as full copulative and verbs are tagged as such, the question words (*QUE*), see below.

- **The personal attribute feature:**

The first level feature (*PERS*) is described in a finer granularity on the second level: the specifications *first person singular and plural (1sg and 1pl)* and *second person singular and plural (2sg and 2pl)* are added on this level.

- **The feature set of morphemes:**

The possible values for morphemes (*MORPH*) are described at the second level of annotation: question (*que*), deficient (*def*), negation (*neg*), potential (*pot*), future (*fut*), present (*pres*), progressive (*prog*), infinitive prefix (*cp15*). Compare in this regard *tlo* (*shall, will*), annotated as *MORPH_fut*, where the underscore separates level 1 and 2 annotations.

- **The feature set of particles:**

The set of possible values for particles (*PART*) includes agentive (*agen*), connective (*con*), copulative (*cop*), hortative (*hort*), instrumental (*ins*), locative (*loc*), question (*que*) and temporal (*temp*), cf. *ka* (*with*) *PART_ins*.

- **Others:**

Most of the other annotations are fully described on the first level of annotation, with the exception of question words (*QUE*), and *VCOP*, where it is possible to indicate whether a copulative is negated (*VCOP_neg*), or if it is a copulative subject concord of a certain class, compare for example *o* ((*it*) *is*), which is annotated as *VCOP_N03*. Some adverbs (*ADV*) have a clear locative character

and are therefore annotated *ADV_loc*, e.g. *pele* (*front*, *ahead*, *initial*), just to mention some of the possible translations of this word; *pele* is tagged *NLOC* in the few cases where it is used as a noun. Considering question words, three categories are distinguished. The first category is nominal question words for which a noun class is assigned on the second level, thus *eng* (*what*) is annotated as *QUE_N9*. For the second category, as it is not nominal but class dependent, a class number is assigned, e.g. *bofe* (*which*) *QUE_14*, *sefe* (*which*) *QUE_07*, et cetera. For all other question words, no annotation is assigned on the second level, therefore *goreng* (*why*) is tagged as *QUE_nil*.

Word categories that use the class membership feature are nouns (*N*), adjectives (*ADJ*) and pronouns (*PRO*), which include emphatic (*PROEMP*), possessive (*PROPOSS*) or quantitative (*PROQUANT*) pronouns. The subject and object concords (*CS* and *CO*) and the demonstrative concords (*CDEM*) also belong to this group. The subject concord also carries the features indefinite (*INDEF*) or neutral (*NEUT*) in addition to the class membership feature. Concerning concords, one might argue that they are morphemes; however, note that all non-nominal elements carrying class information, i.e. concords, can acquire a (pro-)nominal status, hence class information should already be included on the first level of annotation to ensure an equal status to other nominals. Secondly, the demonstrative concord is often described as a demonstrative pronoun. However, as this element fulfils a purely concordial function in relative and adjectival constructions, we opted for categorising it as a concord.

The personal attribute feature *PERS* is assigned only to concords and emphatic and possessive pronouns, as only these can refer to the 1st or 2nd person.

Nominal derivations that describe a locative, an augmentative or a diminutive can be described on the second level of annotation as well: for *bahumagadi* (*ladies*) *N02_aug* is a possible annotation.

4. More issues in the application of the tagset

The nouns of Northern Sotho are categorised into several groups. The first group belongs to the noun classes 1 to 10 and 14. It should be noted, that “nouns” of class 15 consist at least of two tokens, usually a verb form which is preceded by the class prefix of class 15, i.e. *go*. The parts of this construction are annotated on a morpheme level, cf. *go*$_{\text{MORPH\_cp15/to}}$ *goroga*$_{\text{V/arrive}}$ *ga*$_{\text{CPOSS15\_nil/of}}$ *bona*$_{\text{PROPOSS02\_nil/they}}$ (*their arrival*). Following this approach, cases like *go*$_{\text{MORPH\_cp15/to}}$

go$_{\text{COPERS\_2sg/you}}$ *hloya*$_{\text{V/hate}}$ (*to hate you*) do not constitute any problem to the annotator. Enhancing the annotation with information stating that the construction belongs to class 15 would actually mean doing the morphological analysis at the tagging stage. This task will hence be done by a morphosyntactic analyser, which will detect the many possible ways to construct an infinitive, e.g. the case of the negated infinitive *go*$_{\text{MORPH\_cp15/to}}$ *se*$_{\text{MORPH\_neg}}$ *dirwa*$_{\text{V/done}}$ (*not to be done*) or of the copulative construction *go*$_{\text{MORPH\_cp15/to}}$ *ba*$_{\text{VCOP\_nil/be}}$ *mmago*$_{\text{N01\_nil/mother of}}$ *setšhaba*$_{\text{N07\_nil/nation}}$ (*to be mother to (the) nation*). Other nouns are annotated on the first level of annotation according to their class membership, e.g. *monna* (*man*) *N01*, *mahlatse* (*luck*), *N06*, and *toropo* (*town*), *N09*. The second level, as described above, is used only for additional information on derivational forms, i.e. *toropong* (*at/in/to/from town*) being the locative derivate of *toropo* (*town*), is annotated as *N09_loc*, while a noun like *pukwana* (*small book, booklet*) is annotated as *N09_dim*, to also provide for the diminutive suffix-*ana*.

Because of the richness of the Northern Sotho verbal morphology, one could mirror this richness in the tagset by introducing a large number of tags, as was done in the EAGLES-based tagset by Van Rooy and Pretorius (2003). However, as described in Prinsloo and Heid (2006), we automatically annotate verb forms simply as *V* on the first level of annotation. On the second level auxiliary verbs are being annotated *V_aux* at the current stage of implementation.

4.1 Multi-unit tokens

A multi-unit token refers to cases where one single orthographic token would receive more than one tag. According to Cloeren (1999:44) these are typically word forms which result from fusing two tokens, but with each of the original ones preserving its grammatical properties. This usually happens when one of the tokens is shortened and as a result, becomes phonologically dependent on the preceding word. It is a generally recognised orthographical rule that phonologically shortened forms are written conjunctively to the form they are dependent on, resulting in a single fused token. Compare the following examples of such enclitic elements:

1. *kang* < *ka* PART_ins (*with*) + *eng* QUE_N09 (*what*)

2. *bonang* < *bona* V (*see*) + *eng* QUE_N09 (*what*)

Three possible solutions present themselves in this regard. The first would be to run the tokens with their tags together, without intervening spaces, cf:

3. *ka*_PART_ins.*ng*_QUE_N09

4. *bona*_V_nil.*ng*_QUE_N09

The first obvious problem with such an approach is the creation of what Leech (1997:22) calls "phantom words", i.e. words that do not exist. In the above examples *ng* is an example of such a phantom word. Cloeren (1999:44) furthermore points out that keeping tokens together might impede further processing.

A second option would be the use of portmanteau tags, i.e. keeping the fused tokens together, followed by the relevant tags, which could be separated by means of some symbol or punctuation mark, e.g.:

5. *kang*_PART_ins.QUE_N09

6. *bonang*_V _nil.QUE_N09

A portmanteau tag, whether the individual tags are separated by some symbolic means or not, is actually a new tag. As this tag would only be used for the fused forms, it would likely lead to data sparseness problems in the training of the statistical tagger. Taking into account that the first level tagset for Northern Sotho is already size-wise at its maximum, this does not seem to be a satisfactory solution to the problem.

A last option that could be considered would be to separate these fused forms during lexicon-based pretagging, using a special lexicon as a stoplist. This would indeed work well for fused forms that are unambiguous, such as *kang*, *lang (< la + eng)*, *keng* (< *ke* + *eng*), et cetera. As a starting point, the tokenised text (column 1; Figure 4) is scanned for merged forms listed in the stoplist. All lines containing these forms are duplicated and labelled according to the number of underlying components. As a next step, the tool inserts an additional column (column 2), where the underlying forms are listed, and then tagged (column 3).

Figure 4: Tagging of multi-unit tokens

Text tokenised on word level	**Unmerging**	**Lexicon-based pretagging**	
		1st level annotation	2nd level annotation
tlo	tlo	MORPH	fut
sepela	sepela	V	nil
kang#1_2	**ka**	**PART**	**ins**
kang#2_2	**eng**	**QUE**	**N09**

It needs to be taken into account though, that coalescence which results in these fused forms is a productive phonological process, and that any verb can theoretically speaking for example coalesce with a following question word *eng*, resulting in a verbal form ending in *-ang.* A stoplist compiled for these merged forms would therefore never be complete. Furthermore, it cannot be assumed that all verbs ending in *-ang* constitute fused forms. The form *bolelang* is for example three ways ambiguous. Apart from constituting the fused form of *bolela* (*speak, say*), followed by the question word *eng* (*what*), it could also be a verb stem followed by the suffix *-ng,* which functions as a plural marker in the imperative mood, or it could be a relative verb, marked by the suffix *-ng,* a variant of the relative marker *-go*. In cases where the suffix *-ng* is indeed a verbal one, no unmerging is necessary, since the verbal status of *bolelang* is not influenced by either of the two suffixes. Therefore, a lexicon-based unmerging would be problematic at this stage of processing. This problem can only be solved at the parsing stage, where the cotext can be utilised for disambiguation purposes. To illustrate: if the form *bolelang* is not preceded by a subject concord, the parser would recognise this form as being a verb to which a plural marker has been suffixed. For the moment, there is no tag defined for such ambiguous forms, therefore they are annotated *V* only.

4.2 Yet undefined forms

It needs to be taken into account that Northern Sotho has not yet been fully standardised. As a result, linguists continuously come across linguistic phenomena that have not yet been described and/or classified. Although correspondence between the grammatical category to which an item belongs and the tag used for annotation purposes is not a prerequisite in a tagset that aims to be readable by humans, this convention is usually followed. Therefore,

in order to select a tag, it is preferable that the designers of the tagset have some inkling of the grammatical category to which a particular token belongs. An example of a so far non-existent definition is the particle *ga* which appears together with the instrumental particle *ka* (*with*) in the combination *ka ga* (*about*). As yet, there is no annotation defined for *ga* in this case.

5. Conclusions and future work

In this article we present a proposal for a part of speech tagset for Northern Sotho, aimed at human readability (e.g. for interactive corpus query) and at the use with a statistical POS-tagger. We consider aspects of standardisation, against the background of reuse (e.g. tagset vs. grammar), of sharing of resources (reinterpretation at another place), and of adaptation (using Northern Sotho as a model for other disjunctively written South African Bantu languages).

In the design of our tagset, we opted for a logical tagset, i.e. one which has a certain number of distinctions at the top level and which includes finer grained distinctions further down the hierarchy. To keep the overall number of tags (i.e. of criteria which must be distinguished by the automatic tagger) to a manageable size, we opted for the introduction of a second layer of annotation, which contains lexically defined refinements of the more general categories, and which can be projected from the lexicon.

We describe the main distinctions underlying the tagset, its formal properties, its use and application. Current work includes quantitative assessment of the tagging performance first tests lead to only 92-93% correctness; improvements are to follow; progress towards an electronic grammar of Northern Sotho, and the application of both in lexicography. Assigning the suggested part-of-speech tags to a given Northern Sotho corpus is the logical first step when developing a grammar of this language. Such a grammar will describe linguistic hypotheses about Northern Sotho and can form the core of a (morphosyntactic) model of the language itself. We are aware of the fact that when assigning rules to combinations of tokens it is fairly possible that the tagset itself will have to be adapted to any future use.

List of 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.

BRANTS, T. 2000. TnT – a statistical part-of-speech tagger. (*In* Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP), Seattle, WA, p. 224-231.)

CLOEREN, J. 1999. Tagsets. (*In* Van Halteren, H., *ed*. Syntactic word-class tagging. Dordrecht: Kluwer. p. 37-54.)

DAELEMANS, W., ZAVREL, J., VAN DEN BOSCH, A. & VAN DER SLOOT, K. 2003. MBT: memory based tagger, version 2.0, Reference Guide. Tilburg: ILK Research Group. (ILK Research Group Technical Report Series 03-13.)

DE SCHRYVER, G-M. & DE PAUW, G. 2007. Dictionary Writing System (DWS) + Corpus Query Package (CQP): the case of *TshwaneLex*. *Lexikos,* 17:226-246. (AFRILEX-reeks/series 17.)

DE SCHRYVER, G-M. & PRINSLOO, D.J. 2000. The compilation of electronic corpora, with special reference to the African languages. *Southern African linguistics and applied language studies*, 18(1-4):89-106.

FAAß, G., HEID, U., TALJARD, E. & PRINSLOO, D.J. s.a. Creating word class tagged corpora for Northern Sotho – report on statistical tagging. (Not yet published.)

HEID, U., PRINSLOO, D.J., FAAß, G. & TALJARD, E. s.a. The compilation of a noun guesser for part of speech tagging in Northern Sotho.

KOSCH, I. 1993. A historical perspective on Northern Sotho linguistics. Pretoria: Via Afrika. (Via Afrika Monograph Series, 5.)

LEECH, G. 1997. Introducing corpus annotation. (*In* Garside, R., Leech, G. & McEnery, A., *eds*. Corpus annotation: linguistic information from computer text corpora. London: Longman. p. 1-18.)

LEECH, G. & WILSON, A. 1999. Standards for tagsets. (*In* Van Halteren, H., *ed*. Syntactic word-class tagging. Dordrecht: Kluwer. p. 55-80.)

LOMBARD, D.P. 1985. Introduction to the grammar of Northern Sotho. Pretoria: Van Schaik.

LOUWRENS, L.J. 1991. Aspects of Northern Sotho grammar. Pretoria: Via Afrika.

POULOS, G. & LOUWRENS, L.J. 1994. A linguistic analysis of Northern Sotho. Pretoria: Via Afrika.

PRINSLOO, D.J., FAAß, G., TALJARD, E. & HEID, U. s.a. The compilation of a verb guesser for part of speech tagging in Northern Sotho. (Not yet published.)

PRINSLOO, D.J. & HEID, U. 2006. Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping. Conference for Lesser Used Languages and Computer Linguistics, EURAC Research, European Academy, Bolzano, Italy, 27th October-28th October 2005.

SCHMID, H. 1994. Probabilistic part-of-speech tagging using decision trees. Proceedings of the International Conference on New Methods in Language Processing Manchester, UK, p. 44-49.

SNYMAN, G., EHLERS, L. & NAUDÉ, J.A. 2007. Development of the EtsaTrans translation system prototype and its integration into the Parnassus meeting administration system. *Southern African linguistics and applied language studies*, 25(2):225-238.

TALJARD, E. & BOSCH, S.E. 2006. A comparison of approaches towards word class tagging: disjunctively vs. conjunctively written Bantu languages. *Nordic journal of African studies*, 15(4):428-442.

VAN ROOY, B. & PRETORIUS, R. 2003. A word-class tagset for Setswana. *Southern African linguistics and applied language studies*, 21(4):203-222.

Key concepts:

NLP application
part-of-speech tagging (POS-tagging)
tagset: Northern Sotho

Kernbegrippe:

annoteringstel: Noord-Sotho
toepassing vir natuurliketaalprosessering
woordklasannotering

Appendix

The tagset of Northern Sotho 1 / 3

Description	Tag 1st level	Tag 2nd level
Concords:		
subject class 1-10,14, 15	CS01 – CS10, CS14, CS15	--
personal subject	CSPERS	1sg, 2sg, 1pl, 2pl
locative subject	CSLOC	--
indefinite subject	CSINDEF	--
neutral subject	CSNEUT	--
object class 1-10, 14, 15	CO01 – CO10, CO14, CO15	--
personal object	COPERS	2sg, 1pl, 2pl
locative object	COLOC	--
possessive class 1-10, 14, 15	CPOSS01 – 10, CPOSS14, CPOSS15	--
possessive locative	CPOSSLOC	--
demonstrative class 1-10, 14	CDEM01 – CDEM10, CDEM14	--
demonstrative locative	CDEMLOC	--
demonstrative copulative	CDEMCOP	01-10, 14, loc
Pronouns:		
emphatic class 1-10, 14	PROEMP01-10, 14, 15	--, loc
emphatic personal	PROEMPPERS	1sg, 2sg, 1pl, 2pl
emphatic locative	PROEMPLOC	--
possessive class 1-10, 14	PROPOSS01-10, 14, 15	--
possessive personal	PROPOSSPERS	1sg, 2sg, 1pl, 2pl

possessive locative	PROPOSSLOC	--
quantitative class 1-10, 14	PROQUANT01-10, 14, 15	--
quantitative locative	PROQUANTLOC	--
Nouns:		
class 1-10, 14, 15	N01-N10, N14	--, dim, aug, loc
locative	NLOC	--, dim
names of persons singular	N01a	--, name
names of persons plural/respect form	N02b	--, name
names of places	NPP	loc
Adjectives:		
class 1-10, 14, 15	ADJ01-10, ADJ14, ADJ15	--, dim
locative	ADJLOC	--
Verbals:		
verb stem	V	--
copulatives	VCOP	--, neg, N01-10, N14,
Morphemes:		
deficient	MORPH	def
negation	MORPH	neg
potential	MORPH	pot
future	MORPH	fut
present	MORPH	pres
progressive	MORPH	prog
class 15 marker	MORPH	cp15
Particles:		
agentive	PART	agen

connective	PART	con
copulative	PART	cop
hortative	PART	hort
instrumental	PART	ins
locative	PART	loc
question	PART	que
temporal	PART	temp
Question words:		
nominal	QUE	N01-N10, N14
others	QUE	--, 01-10, 14, 15, loc
Others:		
abbreviation	ABBR	--
adverb	ADV	--, loc
conjunction	CONJ	--
enumerative	ENUM	--
numeral	NUM	--
ordinal	ORD	--
ideophone	IDEO	--
interjection	INT	--, neg
punctuation	$., $”, $-	--