# Grammar-based tools for the creation of tagging resources for an unresourced language: the case of Northern Sotho

**Ulrich Heid*, Elsabé Taljard†, Danie J. Prinsloo†,**

*Institut für Maschinelle Sprachverarbeitung (IMS)
Universität Stuttgart
Azenbergstr. 12
70174 Stuttgart
Germany
Ulrich.Heid@ims.uni-stuttgart.de

†University of Pretoria - African Languages
0002 Pretoria
South Africa
{danie.prinsloo — elsabe.taljard}@up.ac.za

**Abstract**

We describe an architecture for the parallel construction of a tagger lexicon and an annotated reference corpus for the part-of-speech tagging of Nothern Sotho, a Bantu language of South Africa, for which no tagged resources have been available so far. Our tools make use of grammatical properties (morphological and syntactic) of the language. We use symbolic pretagging, followed by stochastic tagging, an architecture which proves useful not only for the bootstrapping of tagging resources, but also for the tagging of any new text. We discuss the tagset design, the tool architecture and the current state of our ongoing effort.

## 1. Introduction: Situation and objectives

Northern Sotho (alternatively Sepedi) is one of the 11 official languages of South Africa, spoken by around 4.5 million people in the northern and northeastern part of the country. Together with two other languages (Tswana and Southern Sotho), it forms the Sotho language group.

Northern Sotho is one of those languages for which few linguistic resources are available, and the medium term objective of the work which is reported on in this paper is to prepare such resources. All language data available for Northern Sotho are so far are unannotated, as no linguistic tagset was previously available.

A raw corpus of over 6 million words is available for Northern Sotho, called the *University of Pretoria Sepedi Corpus* (PSC), comprising 327 Northern Sotho books and magazines, captured at University of Pretoria by means of optical character recognition (OCR). Subsequent sentence tokenizing and clean-up of OCR errors are still ongoing (for details, see (D.J. Prinsloo, 1991) and (G-M. De Schryver and D.J. Prinsloo, 2000)).

In addition, a frequency-based word form list has been derived from the corpus, comprising ca. 58.000 full forms, as lingware for a spelling checker.

Work described in this paper aims at the creation of different kinds of electronic language resources for Northern Sotho, with a focus on resources for part-of-speech tagging (= pos-tagging). As we opted for symbolic pretagging and subsequent stochastic tagging (for reasons to be discussed in section 3.1. below), we need a tagset, a manually disambiguated pos-tagged corpus of at least 40,000 tokens and a tagger lexicon as a minimal start-up kit. This paper describes how these resources have been created, and which architecture we envisage for the tagging process.

Our objective was to create these resources in parallel, in one go, at least as much as possible. Furthermore, since Northern Sotho, as all Bantu languages, has a highly productive verbal morphology, we assume that the combination of symbolic pretagging and stochastic tagging not only helps to efficiently bootstrap tagging resources, but will also be an advantage in any kind of processing of unknown text.

As we report on ongoing work, no full evaluation of the results is yet available, and the tool suite is not yet completed. We describe the linguistic background and the tagset design in section 2., before discussing our tool design methodology, the intended architecture and examples of our pretagging tools in section 3. Section 4. is devoted to the first results of our work and 5. to a discussion of future work.

## 2. Northern Sotho linguistics and tagset design

Many considerations in the design of a tagset for Northern Sotho are obviously conditioned by specificities of the grammar and lexicon of the language. We discuss aspects of the behaviour of nouns and verbs, as well as some quantitative tendencies, all of which have influenced our design principles. For details about Northern Sotho grammar, see (D.P. Lombard, E.B. Van Wyk, P.C. Mokgokong, 1985), (L.J. Louwrens, 1991) and (G. Poulos and L.J. Louwrens, 1994).

### 2.1. The Northern Sotho noun system

As in all Bantu languages, Northern Sotho nouns are grouped into noun classes. Classes 1 to 10 are paired, such that classes 1, 3, 5, 7 and 9 contain singular forms, their respective plural forms belonging to classes 2, 4, 6, 8 and 10. Bantu classes 11 to 13 are not used for the description

of Northern Sotho, and classes 14 (sortal reading of mass nouns), 15 (nominalized infinitive) and 16 to 18 (locative classes) are comparatively sparsely populated in terms of types in the PSC; examples of the noun classes are given in table 1.

| Class | Prefix | Example | Translation |
|---|---|---|---|
| 1 | mo- | monna | man |
| 2 | ba- | banna | men |
| 1a | ∅ | malome | uncle |
| 2a | bo+ | bomalome | uncles |
| 3 | mo- | monwana | finger |
| 4 | me- | menwana | fingers |
| 5 | le- | lesogana | young man |
| 6 | ma- | masogana | young men |
| 7 | se- | selepe | axe |
| 8 | di- | dilepe | axes |
| 9 | n-/ ∅ | nku | sheep (sg.) |
| 10 | di+ | dinku | sheep (pl.) |
| 11 | | | |
| 12 | | | |
| 13 | | | |
| 14 | bo- | bogobe | porrige |
| 6 | ma- | magobe | different kinds of porrige |
| 15 | go | go bona | to see |
| 16 | fa- | fase | below |
| 17 | go- | godimo | above |
| 18 | mo- | morago | behind |

Table 1: Northern Sotho noun classes and examples

The noun class system is also applied to different types of concords (subject concords, object concords, possessive and demonstrative concords), as well as to emphatic, quantifying and possessive pronouns. As the concords and pronouns play an important role in the constitution of nominal groups and thus in the interpretation of sentences, it makes sense to keep track of their noun class subtypes in pos tagging. In fact, concords and pronouns agree with their respective antecedent nouns, as shown by the sentence in table 2, translated as *this man loves them*[1].

With the exception of the noun class prefix (e.g. *mo-*, in the case of the noun *monna* in table 2), most concords, particles, etc. are written as separate orthographic units (disjunctive writing, contrary to the South African Nguni languages, such as Zulu) and must get pos-tags. For a detailed discussion of the issue of tagging in Sotho vs. Nguni languages see (E. Taljard and S. E. Bosch, 2005). With about 12 noun class values per subtype, this alone leads to over 80 different tags. The concordial phenomena explained above, which are sensitive to noun classes, can be annotated (and used as contexts for data extraction, see below) on this basis.

[1]We use the following abbreviations: Pron(oun), Conc(ord), V(erb); for pronouns and concords, we give subtypes as :dem(onstrative), :sub(ject), :obj(ect); "Cl.1" stands for "noun class 1", "Cl8/10" for "noun class 8 or 10".

| *Monna* | *yo* | *o* | *a* | *di* | *rata* |
|---|---|---|---|---|---|
| Noun | Pron. :dem. | Conc :subj. | Present | Conc :obj. | V :stem |
| Cl.1 | Cl.1 | Cl.1 | tense | Cl.8/10 | |
| man | this | (he) | ( ) | them | love(s) |

Table 2: Northern Sotho sentence illustrating nominal class agreement of the subject noun, pronoun and concord

### 2.2. Northern Sotho verbal morphology

The example in table 2 above, contains a verb which is marked as verb stem. In Northern Sotho, there are numerous verbal derivations (e.g. expressing causativity, perfectivity, passive, reciprocals, etc.), which formally consist of prefix and suffix clusters added to the stem. (D.J. Prinsloo, 1994) distinguishes 18 modules of such verbal derivation clusters, and so far over 260 different individual verbal derivation types have been listed in the comparison of different verb stems.

For example, for the stem *reka* ("buy"), the form *rekana* comprises the stem and a reciprocality marker; it makes up a series with reciprocal and perfect suffixes (*rekane*), reciprocal and passive suffixes (*rekanwa*), reciprocal, perfect and passive suffixes (*rekanwe*).

Not all verbal derivations are highly frequent in corpora (in terms of tokens), but due to productivity, a huge number of verbal derivation types can potentially show up in any text. A lexical listing of these forms is not economic, and only a minimal number of verb forms shows categorial homography with other items (mostly concords and particles). Both factors speak in favour of a rule-based tool to analyze occurrences of the verb derivations.

The morphological richness of verbs can be accounted for in different ways. In the tagset, a number of morphological features (such as active vs. passive, present vs. past, etc.) could be expressed through individual tags. This has been proposed by (B. Van Rooy and R. Pretorius, 2003). With stochastic tagging in mind, we avoid introducing such distinctions in the tagset, to keep the overall number of different tags to be assigned as low as possible.

We consequently only use the tag "V", for verb forms of any kind. We assume that the tagging of Northern Sotho texts will be followed by additional, possibly specialized annotation steps. One of these could be to expand the annotation of word forms by adding morphological information. At this stage, verbal derivations would be treated, and the interpretation of the (mostly unambiguous) affix clusters would lead to the respective morphological data. Verb derivation would then be handled the same way as nominal derivation; nominal diminuitive, augmentative and feminine, as well as locative formation. These nominal derivations are also morphologically constructed by affixation. We do not take these categories into consideration but expect to be able to identify (and annotate) them in a second step.

### 2.3. Ambiguity and frequency in Northern Sotho word forms

In Northern Sotho, only nouns and verbs are open classes with lexical productivity, all other word classes (including adjectives) are regarded as closed classes. Above, we mentioned the productivity of the verbal derivation system of Northern Sotho. It leads to a situation where any text to be automatically analysed may contain a considerable number of word forms not contained even in a well-developed lexicon. As mentioned, this motivates our decision for rule-based pretagging (of verbs and nouns).

As was indicated above, most verb forms are not ambiguous, neither within the verbal paradigms, nor in the sense of categorial homography. There is only a small number of (however very frequent) accidental homographs between verb forms and function words.

With nouns, homography is even less frequent. In a 40,000 words sample from the novel "Tša ka mafuri" (O.K. Matsepe, 1974)), 86% of all noun form occurrences are unambiguous. At type level, this amounts to only a handful of items which have homographs.

The picture is completely different for the closed classes. Around 80% of all occurrences of closed class items are ambiguous. As in most languages, the most frequent words show the highest number of alternatives. In our tagset, the most frequent item (*a,* frequency: 2.304 in 40,000) is ten ways ambiguous. These highly ambiguous items make up for a very large portion of the texts: 88 types of closed class items, with an average frequency of well above 200 in our 40,000 words sample, make up for around 40% of all occurrences.

| Item | Possible tags | Freq. |
|---|---|---|
| a | CDEM6:CO6:CS1:CS6:CPOSS1: CPOSS6:PAHORT:PAQUE:PRES | 2304 |
| go | CO2psg:CO15:COLOC:CP15:CS15: CSLOC:CSindef:PALOC | 2201 |
| ka | CS1psg:PAINS:PATEMP: PALOC:POSSPRO1psg:POT | 1979 |
| le | CDEM5:CO2ppl:CO5: CS2ppl:CS5:PACON:VCOP | 1690 |
| ba | AUX:CDEM2:CO2:CS2: CPOSS2:VCOP | 1509 |

Table 3: Part-of-speech ambiguity of the five topmost words of Northern Sotho by frequency

### 2.4. A tagset for Northern Sotho

The part-of-speech tagset designed for Northern Sotho keeps track of some of the linguistic facts discussed above, in particular the richness of the nominal class system and its impact on nominal categories. Nouns and adjectives, as well as all four subclasses of concords (subject and object concord, posessive and demonstrative concord) and the three subclasses of pronouns (emphatic, possessive and quantifying) all are further subdivided according to noun classes. As mentioned above all verb forms are annotated "V", and only a few copulative verbs ("VCOP") are singled out. As Northern Sotho has a number of morphemes to express temporal (present vs. future), aspectual (progressive) and modal meaning (potential, negation, question), these particle-like words are tagged each with its grammatical meaning. Similarly, we distinguish particles into agentive, connective, copulative, hortative, instrumental, locative and temporal. In addition, adverbs, auxiliaries, ideophones, conjunctions, numerals and enumerative words are distinguished.

In total, the current version of the tagset has 156 different tags (large number due to multiplication of tags for the noun classes, see above). Due to ambiguity at the level of function words, of the 170 tags or tag combinations used in the 40,000 word sample of the PSC, only 66 annotation types are unambiguous. This seems to underline the need for stepwise disambiguation.

The tagset is designed to be a logical tagset, i.e. tags are grouped in a hierarchy, which supports underspecified queries (e.g. search for some concord of class 1:`''C.+01''`, which matches `''CS01''`, `''CO01''`, `''CDEM01''`, `''CPOSS01''`, i.e. subject, object, demonstrative or possessive concords of class 1.

## 3. Towards tools for pos-tagging of Northern Sotho

### 3.1. Options for tagging

To our knowledge, no experiments into pos-tagging of a Sotho language have as yet been undertaken. Thus, no existing technology can be adapted and a new tool setup needs to be designed. We expect however such adaptation to be possible in principle, due to the close relatedness of the Sotho languages.

In the decision for rule-based vs. stochastic tagging, we opted for a mixed solution, comprising partial rule-based pretagging followed by stochastic tagging. An obvious advantage of stochastic taggers is their trainability. Typically the preparation of a disambiguated training corpus requires less effort than the creation of hand-crafted rules for a rule-based tagger. On the other hand, well known problems of stochastic tagging include the threshold in the size of needed trainig data, restrictions on the size of the tagset and problems with large numbers of unknown phenomena which cannot be adequately captured by the typical bigram or trigram approaches to statistical tagging.

We opted for the use of Schmid's TreeTagger (H. Schmid, 1994)), as it provides acceptable results with training corpora of only about 40,000 word forms, whereas most other taggers require a minimum of 100.000 words. The tagset for Northern Sotho, with its 156 tags, is at the upper limit of the possibilities of the TreeTagger. To avoid the unknown words problem, we designed pretagging tools for the open word classes, i.e. nouns and verbs. Non-local effects seem to be relatively rare in Nothern Sotho; on the contrary, most of the concords, pronouns and adjectives that are highly ambiguous are found in the neighbourhood within two to three words from the nouns they agree with.

### 3.2. An architecture for pretagging and main tagging

We decided to decouple the tagging of open class items from that of closed class items, introducing rule-based pretaggers for the open classes. With nouns, this is motivated by the sheer size of the nominal lexicon and by considerations of robustness: any domain of terminology will bring up new nouns, and the possibility to identify nouns (at least in terms of their part of speech) is crucial for the treatment of any new text. For verbs, the motivation is obviously their morphological productivity (cf. above, section 2.2).

Thus we designed and implemented two guessing tools, one to identify verb forms, the other to identify noun candidates. We run both guessers before the actual use of the TreeTagger.

Furthermore, there are numerous sequencing patterns of closed class items which show lexical variation but have a well-defined meaning and thus can be used to (possibly partly) disambiguate the closed class items. We are in the process of carrying out experiments on the usefulness of such disambiguation rules as an additional element of pretagging. They make sense for a (partial) disambiguation of particularly frequent and ambiguous words, especially in patterns that cannot be defined in terms of lexeme sequences (and can thus not easily be learned from a comparatively small training corpus). We will illustrate all three kinds of pretagging devices below, in section 3.3. and 3.4. For other languages, a combination of rule-based pretagging and statistical tagging has also been proposed. Klatt (cf. (S. Klatt, 2005)) shows the improvements achieved by means of grammar-based pretagging on the task of tagging German newspaper text.

The working procedure for the preparation of a tagger lexicon and a disambiguated training corpus for the TreeTagger is semi-automatic. The raw corpus (40,000 words, see above) is sentence-tokenized first. A very first version of a tagger lexicon which comprises about 2,000 closed class items (with all their tag alternatives) and around 5,000 noun and verb forms is then used to ambiguously annotate some of the word forms in the corpus. Thereafter, guessing tools for verb forms and noun forms are applied and their tag hypotheses annotated into the corpus. This leads to an ambiguously annotated corpus. A set of closed class disambiguation rules is then applied. to remove tag alternatives, where possible automatically. In the creation of the training corpus for the TreeTagger, the resulting partly disambiguated corpus is then manually disambiguated.

A very similar architecture can be used in a "regular" tagging situation. Obviously, all results of the verb and noun guessers are manually checked, and the verb and noun forms are included in the tagger lexicon. Thus the size of the lexicon grows continuously. In the "regular" tagging procedure, the trained TreeTagger is used for disambiguation of the partly disambiguated corpus. Figure 1 schematically represents the two strands of the architecture, cf. also (D.J. Prinsloo and U. Heid, 2005).

### 3.3. Verb and noun guessing as pretagging

As mentioned above, most verb and noun forms of Northern Sotho are not ambiguous. This holds especially for verb forms, due to their suffixes. Verb guessing is thus only based on the recognition of one of the ca. 260 different suffix clusters, implemented as pattern matching.

Noun identification is not possible on the basis of affixes alone. Noun suffixes (locatives, augmentatives/feminines; diminuitives) are too rare to lead to acceptable recall. Noun prefixes (class indicators, see table 1) do not only show up in noun forms; the form *letetše* ("waited") is verbal, and *le-* is not, in this form, a prefix of noun class 5, but part of the verb stem.

Thus we opted for the use of combined indicators of noun classes, namely noun prefixes and the adjectives, concords and pronouns showing up in the environment of a word. Furthermore, we use pairings of singular and plural as a third indicator of noun readings. As mentioned above (see section 2.1. and table 2), adjectives, concords and pronouns agree with the nouns they refer to. We exploit this to check every word form in the text which starts with one of the possible noun prefixes (*ba-, bo-, di-, ma-, me-, mo-*, etc., see table 1). For each noun class, we collected lists of the most frequent closed class words which accompany nouns. For classes 5 and 6, examples of such items are listed in table 4, below.

| POS | class 5 | class 6 |
|---|---|---|
| CS | la, le | a |
| CDEM | lekhwi | akhwi |
| CDEM | leno | ano |
| QUANTPRO | lohle | ohle |
| ADJ | leso | maso |
| ADJ | lesese | masese |

Table 4: Closed class words indicative of noun classes 5 and 6

These closed class items are usable as noun indicators, when

(i) one or more of them appear in the neighbourhood of a word form supposedly belonging to one of the noun classes; and

(ii) parallel behaviour is observed for both singular and plural forms.

We thus implemented a tool to guess singular-plural word form pairs for all noun classes which can be called with forms of either form. It applies the knowledge of table 1 and a set of morphophonological rules; subsequently, we test both, the form found in the corpus and the guessed corresponding form, in corpus search patterns on the PSC. These patterns contain the respective closed-class items. Only if both, singular and plural candidates (of classes 1 to 10) fulfill the context tests, we accept (both forms) with their class assignment into the tagger lexicon and pretag the occurrences in the corpus accordingly. Clear cases of non-noun forms are left over for verb guessing, and all other cases are manually decided. To some extent, the context patterns for noun identification obviously also locally disambiguate the closed class items used as indicators: the form *a*, used as a subject concord in the context of a class

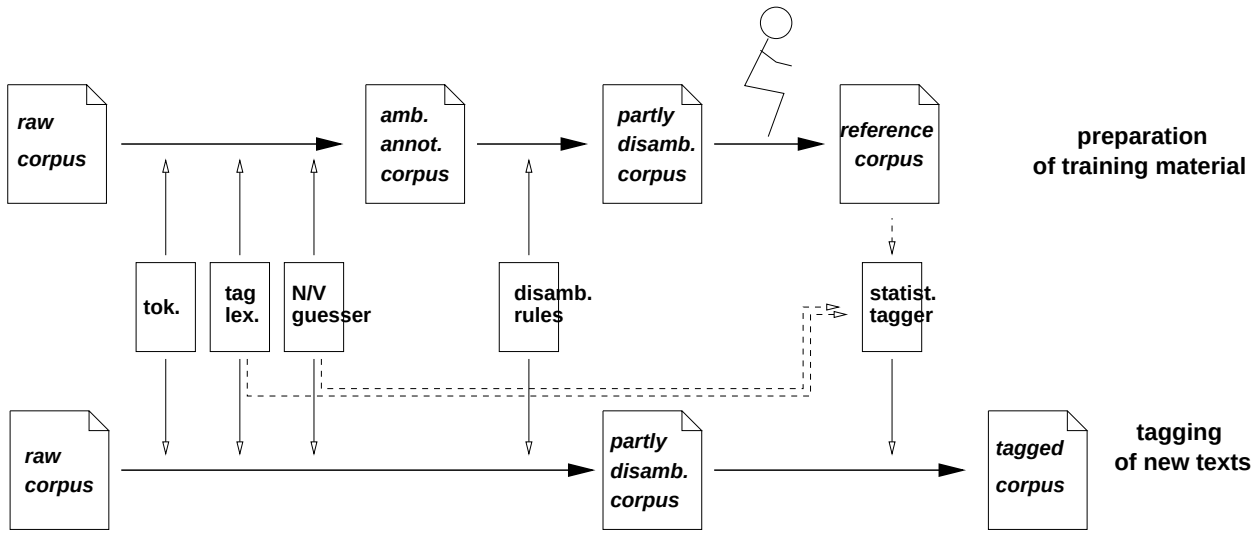


Figure 1: Architecture for the creation of training material and for the tagging of new texts

6 noun, could be among other things, a class 1 subject concord, but only in the context of a class 1 noun. In the context of a class 6 noun, *a* can be any type of class 6 concord; thus the disambiguation is not complete.

### 3.4. Disambiguation of closed class items

To disambiguate closed-class items, at least to some extent, in the pretagging step, we make use of patterns, in terms of word forms, word classes and/or morphological properties. For example, the word *a* is to be tagged as a possessive concord of a class 6 noun, if it appears exactly between two nouns. Similar rules can be constructed for other possessive concords, as the pattern "noun + possessive concord + noun" is a standard pattern to form complex nominals. Similarly, the same word *a* preceding immediately a relative clause verb form (ending in *-go*), cannot be a possessive, but should be a subject or object concord ("`CS01`" or "`CS06`" or "`CO06`" ). The set of rules for the disambiguation of closed class items is currently under development.

## 4. Results and Conclusions

### 4.1. Early results

We have not yet been able to quantitatively evaluate the tools discussed here; however, work on the 40,000 words sample of the PSC used as a training corpus for the TreeTagger suggests that verb and noun guessing lead to both high precision and high recall. This is expectable, as the number of homographs in the nominal and verbal domain is small. Table 5 gives a few examples of noun guessing for classes 7 and 8, showing that the guesser is robust towards nonexistent forms (cf.**dipetše*) as well as towards actual homographs (in table 5: *sekelela* "recommend" vs. *dikelela* ("disappear")), which formally, look like a pair of nouns of classes 7 and 8.

| Cl. 7 cands. | Cl. 8 cands. | N? | Equivalent(s) |
|---|---|---|---|
| selo | dilo | + | thing, things |
| setšhaba | ditšhaba | + | nation, nations |
| sello | dillo | + | (out)cry, outcries |
| sepetše | *dipetše | – | walked |
| sekelela | dikelela | – | recommend, disappear |

Table 5: Sample results of noun guessing for the Northern Sotho classes 7 and 8

### 4.2. Methodological considerations

The creation of tagging resources for Northern Sotho is conceived as a bootstrapping process:

- first, a basic dictionary of closed class items is provided;
- then, morphological knowledge (for verbs) and knowledge about local morphosyntactic patterns (for nouns) is used to guess the word classes of open class items;
- the guessing routines for nouns provide a partial disambiguation of some of the highly ambiguous closed class items; additional specialized patterns serve the same purpose.
- the results of all guessing processes are annotated in the corpus and at the same time integrated in the word form lexicon, thus permitting a parallel resource bootstrapping.

The architecture comprising a rule-based pretagging and a stochastic main-tagging phase can be used both to bootstrap the tagging resources and to constantly enlarge the stock of available resources for the tagging of Northern Sotho.

## 5. Conclusions and future work

We have discussed an approach for the parallel creation of a tagger lexicon and an annotated training text. By exploiting the morphological and syntactic properties of Northern Sotho, we are able to use rule-based pretagging in order to avoid the unknown words problem in the statistical tagging step and to reduce the amount of pos-hypotheses for the highly ambiguous closed class items. The proposed architecture can be used for the creation of training data for the tagger, as well as in any tagging work at a later stage.

It is necessary to quantify the effect of pos-disambiguation with respect to tagging accuracy. To this end, we plan a series of experiments, where we will check tagging quality on the 40,000 words sample of the PSC with and without pretagging. In particular, we intend to systematically check the impact of certain (sets of) rules for high frequency closed class items.

In the medium term, we expect to tag the PSC in slices of ca. 2 million words each, to understand frequency effects involved in lexicon acquisition for Northern Sotho. In a long term perspective, it would be tempting to address the issue ofextending the approach to other Sotho and possibly other disjunctively written Bantu languages.

## 6. Acknowledgements

The work described here was made possible by generous support from the University of Pretoria and from the Universität Stuttgart. In particular, we should like to thank Gertrud Faaß (Stuttgart) for her invaluable help with the implementation of the tools. Without her, this work would not have been possible.

## 7. References

G-M. De Schryver and D.J. Prinsloo. 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.

S. Klatt. 2005. *Textanalyseverfahren für die Korpusannotation und Informationsextraktion*. Shaker, Aachen.

D.P. Lombard, E.B. Van Wyk, P.C. Mokgokong. 1985. *Introduction to the Grammar of Northern Sotho*. J.L. van Schaik , Pretoria

L.J. Louwrens. 1991. *Aspects of Northern Sotho Grammar*. Via Afrika, Pretoria.

O.K. Matsepe. 1974, *Tša ka mafuri*. Van Schaik. Pretoria.

G. Poulos and L.J. Louwrens. 1994. *A Linguistic Analysis of Northern Sotho* Via Afrika. Pretoria.

D.J. Prinsloo. 1991. Towards computer-assisted word frequency studies in Northern Sotho. *SA Journal of African Languages*, 11(2).

D.J. Prinsloo 1994. Lemmatization of verbs in Northern Sotho. *SA Journal of African Languages* , 14(2):93-102.

D.J. Prinsloo and U. Heid. 2005. *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.

H. Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*. 44-49.

E. Taljard and S. E. Bosch. 2005. *A comparison of approaches towards word class tagging: disjunctively vs. conjunctively written Bantu languages*. Conference for Lesser Used Languages and Computer Linguistics, EURAC research, European Academy. Bolzano, Italy. 27th October-28th October 2005.

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