Originally published in: De Pauw, Guy/de Schryver, Gilles-Maurice/Levin, Lori (Eds.): Proceedings of the First Workshop on Language Technologies for African Languages. – Athens, Greece: Association for Computational Linguistics, 2009. Pp. 38-45.

Part-of-Speech tagging of Northern Sotho: Disambiguating polysemous function words

Gertrud Faaß[†‡] **Ulrich Heid**[†] **Elsabé Taljard**[‡] **Danie Prinsloo**[‡]
[†] Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Germany
faaszgd@ims.uni-stuttgart.de
heid@ims.uni-stuttgart.de

[‡] University of Pretoria
South Africa
elsabe.taljard@up.ac.za
danie.prinsloo@up.ac.za

Abstract

A major obstacle to part-of-speech (=POS) tagging of Northern Sotho (Bantu, S 32) are ambiguous function words. Many are highly polysemous and very frequent in texts, and their local context is not always distinctive.
With certain taggers, this issue leads to comparatively poor results (between 88 and 92 % accuracy), especially when sizeable tagsets (over 100 tags) are used.
We use the RF-tagger (Schmid and Laws, 2008), which is particularly designed for the annotation of fine-grained tagsets (e.g. including agreement information), and we restructure the 141 tags of the tagset proposed by Taljard et al. (2008) in a way to fit the RF tagger. This leads to over 94 % accuracy. Error analysis in addition shows which types of phenomena cause trouble in the POS-tagging of Northern Sotho.

1 Introduction

In this paper, we discuss issues of the part-of-speech (POS) tagging of Northern Sotho, one of the eleven official languages of South Africa, spoken in the North-east of the country. Northern Sotho is a Bantu language belonging to the Sotho family (Guthrie, 1967: S32). It is written disjunctively (contrary to e.g. Zulu), i.e. certain morphemes appear as character strings separated by blank spaces. It makes use of 18 noun classes (1, 1a, 2, 2b, 3 to 10, 14, 15, and the locative classes 16, 17, 18, *N-*, which may be summarized as LOC for their identical syntactic features). A concordial system helps to verify agreement or resolve ambiguities.

We address questions of the ambiguity of function words, in the framework of an attempt to use "standard" European-style statistical POS taggers on Northern Sotho texts.

In the remainder of this section, we briefly discuss our objectives (section 1.1) and situate our work within the state of the art (section 1.2). Section 2 deals with the main issues at stake, the handling of unknown open class words, and the polysemy of Northern Sotho function words. In section 3, we discuss our methodology, summarizing the tagset and the tagging technique used, and reporting results from other taggers. Section 4 is devoted to details of our own results, the effects of the size of training material (4.2), the effects of polysemy and reading frequency (4.3), and it includes a discussion of proposals for quality improvement (Spoustová et al., 2007). We conclude in section 5.

1.1 Objectives

The long term perspective of our work is to support information extraction, lexicography, as well as grammar development of Northern Sotho with POS-tagged and possibly parsed corpus data. We currently use the 5.5 million word *University of Pretoria Sepedi Corpus* (PSC, cf. de Schryver and Prinsloo (2000)), as well as a 45,000 words training corpus. We aim at high accuracy in the POS-tagging, and at minimizing the amount of unknown word forms in arbitrary unseen corpora, by using guessers for the open word classes.

1.2 Recent work

A few publications, so far, deal with POS-tagging of Northern Sotho; most prominently, de Schryver and de Pauw (2007) have presented the MaxTag method, a tagger based on Maximum Entropy

Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009, pages 38–45,
Athens, Greece, 31 March 2009. ©2009 Association for Computational Linguistics

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-118813

Learning (Berger et al., 1996) as implemented in the machine learning package Maxent (Le, 2004). When trained on manually annotated text, it extracts features such as the first and last letter of each word, or the first two and last two letters or the first three and last three letters of each word; it takes the word and the tag preceding and following the item to be tagged, etc., to decide about word/tag probabilities. De Schryver and de Pauw report an accuracy of 93.5 % on unseen data, using a small training corpus of only ca. 10,000 word forms.

Other work is only partly engaged in POS-tagging, e.g. Kotzé's (2008) finite state analysis of the verb complex of Northern Sotho. This study does not cover all parts of speech and can thus not be directly compared with our work. Taljard et al. (2008) and Van Rooy and Pretorius (2003) present tagsets for Northern Sotho and the closely related language Setswana, but they focus on the definition of the tagsets without discussing their automatic application in detail. In (Prinsloo and Heid, 2005), POS-tagging is mentioned as a step in a corpus processing pipeline for Northern Sotho, but no experimental results are reported.

2 Challenges in tagging Northern Sotho

POS-tagging of Northern Sotho and of any disjunctively written Bantu language has to deal especially with two major issues which are consequences of their morphology and their syntax. One is the presence, in any unseen text, of a number of lexical items which are not covered by the lexicon of the tagger ("unknown words"), and the other is an extraordinarily high number of ambiguous function words.

2.1 Unknown words

In Northern Sotho, nouns, verbs and adverbs are open class items; all other categories are closed word classes: their items can be listed. The open classes are characterized in particular by a rich morphology: nouns can form derivations to express diminutives and augmentatives, as well as locative forms, to name just a few. Adding the suffix *-ng* to *toropo* 'town', for example, forms *toropong*, 'in/at/to town'. For verbs, tense, voice, mood and many other dimensions, as well as nominalization, lead to an even larger number of derived items. Prinsloo (1994) distinguishes 18 clusters of verbal suffixes which give rise to over 260 individual derivation types per verb. Only a few of these derivations are highly frequent in corpus text; however, due to productivity, a large number of verbal derivation types can potentially appear in any unseen text.

For tagging, noun and verb derivations show up as unknown items, and an attempt to cover them within a large lexicon will partly fail due to productivity and recursive applicability of certain affixes. The impact of the unknown material on tagging quality is evident: de Schryver and de Pauw (2007) report 95.1 % accuracy on known items, but only 78.9 % on unknowns; this leads to a total accuracy of 93.5 % on their test corpus. We have carried out experiments with a version of the memory-based tagger, MBT (Daelemans et al., 2007), which arrives at 90.67 % for the known items of our own test corpus (see section 3.2), as opposed to only 59.68 % for unknowns.

To counterbalance the effect of unknown items, we use rule-based and partly heuristic guessers for noun and verb forms (cf. Prinsloo et al. (2008) and (Heid et al., 2008)) and add their results to the tagger lexicon before applying the statistical tagger: the possible annotations for all words contained in the text are thus part of the knowledge available to the tagger.

Adverbs are also an open word class in Northern Sotho; so far, we have no tools for identifying them. In high quality tagging, the suggestions of our guessers are examined manually, before they are added to the tagger lexicon.

2.2 Polysemous function words and ambiguity

Function words of Northern Sotho are highly ambiguous, and because of the disjunctive writing system of the language, a number of bound morphemes are written separately from other words.

A single form can have several functions. For example, the token *-a-* is nine-ways ambiguous: it can be a subject concord of noun class 1 or 6, an object concord of class 6, a possessive concord of class 6, a demonstrative of class 6, a hortative or a question particle or a verbal morpheme indicating present tense or past tense (Appendix A illustrates the ambiguity of *-a-* with example sentences). Furthermore, the most polysemous function words are also the most frequent word types in corpora. The highly ambiguous item *go*[1] alone ac-

[1] 11 different functions of *go* may be distinguished: object

counts for over 5 % of all occurrences in our training corpus, where 88 types of function words, with an average frequency of well over 200, make up for about 40 % of all occurrences.

The different readings of the function words are not evenly distributed: some are highly frequent, others are rare. Furthermore, many ambiguous function words appear in the context of other function words; thus the local context does not necessarily disambiguate individual function words. This issue is particularly significant with ambiguities between concords which can have the same function (e.g. object) in different noun classes. As mentioned, *-a-* can be a subject concord of either noun class 1 or 6: though there are some clearcut cases, like the appearance of a noun of class 6 (indicating class 6), or an auxiliary or the conjunction *ge* in the left context (both rather indicating class 1) there still remain a number of occurrences of *-a-* in the corpora only where a broader context, sometimes even information from preceding sentences, may help to disambiguate this item.

Consequently, comparing tagging performance across different tagsets does not give very clear results: if a tagset, like the one used by de Schryver and de Pauw (2007), does not distinguish noun classes, obviously a large number of difficult disambiguation cases does not appear at all (their tagset distinguishes, for example, subject and object concord, but gives no information on noun class numbers). For the lexicographic application we are interested in, and more generally as a preparatory step to chunking or parsing of Northern Sotho texts, an annotation providing information on noun classes is however highly desirable.

3 Methodology

3.1 Tagset

There are several proposals for tagsets to be used for Northern Sotho and related languages. Van Rooy and Pretorius (2003) propose a detailed tagset for Setswana, which is fully in line with the guidelines stated by the EAGLES project, cf. Leech and Wilson (1999). This tagset encodes a considerable number of semantic distinctions in its nominal and verbal tags. In Kotzé's work on the Northern Sotho verb complex, (Kotzé, 2008), a number of POS tags are utilized to distinguish the elements of the verb, however, due to Kotzé's objectives, her classification does not cover other items. De Schryver and de Pauw (2007) use a tagset of only 56 different tags, whereas the proposal by Van Rooy and Pretorius leads to over 100 tags. Finally, Taljard et al. (2008) propose a rather detailed tagset: contrary to the other authors mentioned, they do encode noun classes in all relevant tags, which leads to a total of 141 tags. Furthermore, they encode a number of additional morphosyntactic distinctions on a second level of their tagset, which leads to a total of 262 different classifications of Northern Sotho morphemes.

Our current tagset is inspired by Taljard et al. (2008). However, we disregard some of their second level information for the moment (which in many cases encodes lexical properties of the items, e.g. the subdivision of particles: hortative, question, instrumental, locative, connective, etc.). We use the RF-tagger (Schmid and Laws, 2008) (cf. section 3.3), which is geared towards the annotation of structured tagsets, by separating information which partitions the inventory of forms (e.g. broad word classes) from feature-like information possibly shared by several classes, such as the Sotho noun classes. With this method, we are able to account for Taljard et al.'s (2008) 141 tags by means of only 25 toplevel tags, plus a number of feature-like labels of lower levels. We summarize the properties of the tagsets considered in table 1.

3.2 Training corpus

Our training corpus consists of ca. 45.000 manually annotated word forms, from two text types. Over 20.000 word forms come from a novel of the South African author Oliver K. Matsepe (Matsepe, 1974); over 10.000 forms come from a Ph.D. dissertation by Raphehli M. Thobakgale (Thobakgale, 2005), and another 10.000 from a second Ph.D. dissertation, by Ramalau R. Maila (Maila, 2006). Obviously, this is not a balanced corpus; it was indeed chosen because of its easy accessibility. We use this corpus to train our taggers and to test them; in a ten-fold cross validation, we split the text into ten slices of roughly equal size, train on 9 of them and test on the tenth. In this article, we give figures for the median of these results.

concord of class 15, object concord of the locative classes, object concord of the 2nd person singular, subject concord of class 15, indefinite subject concord, subject concord of the locative classes, class prefix of class 15, locative particle, copulative indicating either an indefinite subject, or a subject of class 15 or a locative subject.

Authors	No. of tags	± noun class	tool?
(van Rooy and Pretorius, 2003)	106	- noun class	no
(De Schryver and De Pauw, 2007)	56	- noun class	yes
(Kotzé, 2008)	partial	N.R.	yes
(Taljard et al., 2008)	141/262	+ noun class	no
This paper	25/141	+ noun class	yes

Table 1: Tagsets for N. Sotho: authors, # of tags, consideration of the noun class system, use in tools

3.3 Tagging techniques: the RF-tagger

We opt for the RF-tagger (Schmid and Laws, 2008), because it is a Hidden-Markov-Model (HMM) tagger which was developed especially for POS tagsets with a large number of (fine-grained) tags. Tests with our training corpus have shown that this tagger outperforms the Tree-tagger ((Schmid, 1994) and (Schmid, 1995)), as shown in figure 1. An additional external lexicon may serve as input, too. The development of the RF-tagger was based on the widely shared opinion that for languages like German or Czech, agreement information (e.g. case, number or person) should preferably appear as part of all appropriate part of speech tags. However, as tagsets increase immensely in size when such information is part of the tags, the data are decomposed, i.e. split into several levels of processing. The probability of each level is then calculated separately (the joint probability of all levels is afterwards calculated as their product). With such methodology, a tag of the German determiner *das* may contain five levels of information, e.g. ART.Def.Nom.Sg.Neut to define a definite, nominative singular, neutral determiner (article) that appears in the nominative case.

This approach makes sense for the Bantu-languages as well, since information on noun class numbers should be part of the noun tags, too, as in Taljard et al's (2008) tagset. A noun here is not only tagged "N", but `Nclass`, e.g. *mohumagadi* '(married) woman' as N01. All concords, pronouns or other types that concordially agree with nouns are also labelled with a class number, e.g. *o*, the subject concord of class 1, is labelled CS01. This approach makes sense, especially in the view of chunking/parsing and reference resolution, because any of those elements can acquire a pronominal function when the noun that they refer to is deleted (Louwrens, 1991).

To utilize the RF-tagger, we split all tags containing noun class numbers into several levels (e.g. the tag N01 becomes N.01). Emphatic and possessive pronouns are represented on three levels (e.g. PROPOSSPERS becomes PRO.POSS.PERS)[2].

4 Results

In a preliminary experiment, we compared several taggers[3] on our manually annotated data. Apart from the RF-tagger (Schmid and Laws, 2008), we also used the Tree-Tagger (Schmid, 1994), the TnT tagger (Brants, 2000) and MBT (Daelemans et al., 2007).

4.1 Comparing taggers

The results give a relatively homogenous picture, with the RF-tagger achieving a median of 94.16 % when utilising a lexicon containing several thousand nouns and verbs. It leads to 91 % accuracy without this lexicon (to simulate similar conditions as for TnT or MBT where no external lexicon was offered). TnT achieves 91.01 %, and MBT 87.68 %. Data from the Tree-Tagger were not comparable for they had been obtained at an earlier stage using the lexicon (92.46 %).

4.2 Effects of the size of the training corpus on the tagging results

All probabilistic taggers are in need of training data the size of which depends on the size of the tagset and on the frequencies of occurrence of each context. De Schryver and de Pauw (2007) demonstrated that when utilizing a tagset that contains only about a third of the tags (56) contained in Taljard et al.'s (2008) tagset (141), their MaxTag POS-tagger reaches a 93.5 % accuracy with a training corpus of only about 10,000 tokens.

Figure 1 depicts the effects of the size of the training corpus on the accuracy figures of the Tree-tagger and the RF-tagger. Tests with training corpora of the sizes 15,000, 30,000 and 45,000 tokens

[2]Tests have shown that the quantitative pronouns should be treated separately, their tags are thus only split into two levels.

[3]Check the References section for the availability of these taggers.

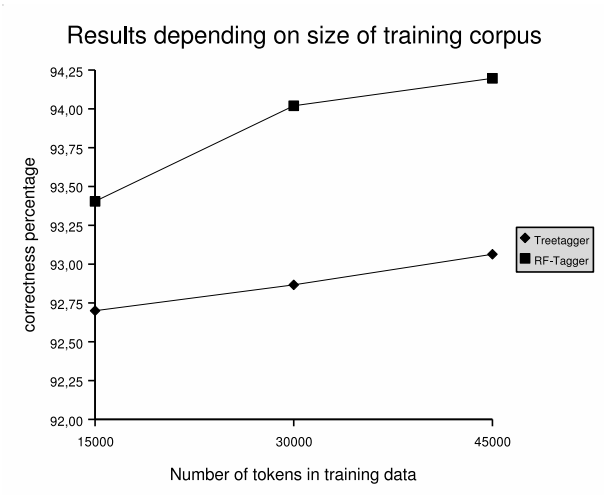

Figure 1: Effects of the size of the training corpus on tagger results.

showed that the results might not improve much if more data is added. The RF-tagger already reaches more than 94 % correctness when utilizing the current 45,000 token training corpus.

4.3 Effects of the highly polysemous function words of Northern Sotho

The less frequently a token-label pair appears in the corpus, the lower is its probability (leading to the sparse data problem, when probability guesses become unreliable because of low numbers of occurrences). This issue poses a problem for Northern Sotho function words: if they occur very frequently with a certain label, the chances of them being detected with another label are fairly low. This effect is demonstrated in table 2, which describes the detection of the parts of speech of the highly ambiguous function word -*a*-. The word -*a*- as PART(icle) occurs only 45 times while -*a*- as CS.01 occurs 1,182 times. More than 50 % of the particle occurrences (23) are wrongly labelled CS.01 by the tagger. In table 2, we list the correct tags of all occurrences of -*a*-, as well as the assigned tags to each of them by our tagger. Each block of table 2 is ordered by decreasing numbers of occurrence of each tag in the output of the RF-tagger. For easier reference, the correct tags assigned by the RF-tagger are printed in bold. Table 2 also clearly shows the effect of ambiguous local context on the tagging result: the accuracy of the CS.06-annotation (subject concord of class 6) is considerably lower than that of the more frequent CS.01 (96.45 % vs. 63.08 %), and CS.01 is the most frequent error in the CS.06 assignment process.

***a* as**	**freq**	**RF-tagger**	**sums**	%
CS.01	1182	**CS.01**	1140	96.4
		CS.06	19	1.6
		MORPH	14	1.2
		CDEM.06	3	0.3
		PART	2	0.2
		CPOSS.06	2	0.2
		CO.06	2	0.2
CS.06	176	**CS.06**	111	63.1
		CS.01	43	24.4
		CPOSS.06	10	5.7
		CDEM.06	5	2.8
		MORPH	3	1.7
		PART	3	1.7
		CO.06	1	0.6
CO.06	18	MORPH	7	38.9
		CS.01	6	33.3
		CO.06	3	16.7
		CS.06	2	11.11
PART	45	CS.01	23	51.1
		MORPH	11	24.4
		CDEM.06	5	11.1
		PART	5	11.1
		CPOSS.06	1	2.2
CDEM.06	97	**CDEM.06**	89	91.8
		CPOSS.06	4	4.1
		CS.06	2	2.1
		CS.01	1	1.0
		PART	1	1.0
CPOSS.06	209	**CPOSS.06**	186	89.0
		CDEM.06	12	5.7
		CS.06	6	2.9
		PART	4	1.9
		CS.01	1	0.5
MORPH	89	**MORPH**	44	49.4
		CO.06	26	29.2
		CS.01	15	16.9
		CPOSS.06	4	4.5
sums	1816		1816	

Table 2: RF-tagger results for -***a***-

4.4 Suggestions for increasing correctness percentages

Spoustová et al. (2007) describe a significant increase of accuracy in statistical tagging when utilizing rule-based macros as a preprocessing, for Czech. We have contemplated, in an earlier stage of our work (Prinsloo and Heid, 2005) to adopt a

similar strategy, i.e. to design rule-based macros for the (partial) disambiguation of high-frequency function words. However, the fact that the local context of many function words is similar (i.e. the ambiguity of this local context (see above)), is a major obstacle to a disambiguation of single function words by means of rules. Rules would interact in many ways, be dependent on the application order, or disambiguate only partially (i.e. leave several tag options). An alternative would be to design rules for the disambiguation of word or morpheme sequences. This would however amount to partial parsing. The status of such rules within a tagging architecture would then be unclear.

4.5 Effects of tagset size and structure

While a preprocessing with rule-based disambiguation does not seem to be promising, there are other methods of improving accuracy, such as, e.g., the adaptation of the tagset. Obviously, types appearing in different contexts should have different labels. For example, in the tagset of Taljard et al. (2008), auxiliary verbs are a sub-class of verbs (V_aux). In typical Northern Sotho contexts, however, auxiliaries are surrounded by subject concords, while verbs are only preceded by them. When 'promoting' the auxiliaries to the first level by labelling them VAUX, the RF-tagger result increases by 0.13 % to 94.16 % accuracy. We still see room for further improvement here. For example, *ga* as PART (either locative particle PART_loc or hortative particle PART_hort) is identified correctly in only 29.2 % of all cases at the moment. The hortative particle usually appears at the beginning of a verbal segment, while the locative in most cases follows the segment. Results may increase to an even higher accuracy when 'promoting' these second level annotations, hort(ative) and loc(ative) to the first annotation level.

5 Conclusions and future work

This article gives an overview of work on POS-tagging for Northern Sotho. Depending on the place of tagging in an overall NLP chain for this language, different choices with respect to the tagset and to the tagging technology may prove adequate.

In our work, which is part of a detailed linguistic annotation of Northern Sotho corpora for linguistic exploration with a view to lexicons and grammars, it is vital to provide a solid basis for chunking and/or parsing, by including information on noun class numbers in the annotation. We found that the RF-tagger (Schmid and Laws, 2008) performs well on this task, partly because it allows us to structure the tagset into layers, and to deal with noun class information in the same way as with agreement features for European languages. We reach over 94 % correctness, which indicates that at least a first attempt at covering the PSC corpus may now be in order.

Our error analysis, however, also highlights a few more general aspects of the POS annotation of Northern Sotho and related languages: obviously, frequent items and items in distinctive local contexts are tagged quite well. When noun class information is part of the distinctions underlying the tagset, function words usable for more than one noun class tend however, to appear in non-distinctive local contexts and thus to lead to a considerable error rate. Furthermore, we found a few cases of uses of, e.g., subject concords that are anaphoric, with antecedents far away and thus not accessible to tagging procedures based on the local context. These facts raise the question whether, to achieve the highest quality of lexical classification of the words and morphemes of a text, chunking/parsing might be required altogether, rather than tagging.

Our experiments also showed that several parameters are involved in fine-tuning a Sotho tagger. The size and structure of the tagset is one such a prominent parameter. Tendencies towards simpler and smaller tagsets obviously conflict with the needs of advanced processing of the texts and of linguistically demanding applications. It seems that tagset design and tool development go hand in hand.

We intend to apply the current version of the RF-tagger to the PSC corpus and to evaluate the results carefully. We expect a substantial gain from the use of the guessers for nouns and verbs, cf. (Prinsloo et. al, 2008) and (Heid et al., 2008). Detailed error analysis should allow us to also design specific rules to correct the output of the tagger. Instead of preprocessing (as proposed by Spoustová et al. (2007)), a partial postprocessing may contribute to further improving the overall quality. Rules would then probably have to be applied to particular sequences of words and/or morphemes which cause difficulties in the statistical process.

References

Adam L. Berger, Stephen Della Peitra, and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics* 22(1): pp. 39 – 71.

Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA*.

Walter Daelemans, Jakub Zavrel, Antal van den Bosch. 2007. *MBT: Memory-Base Tagger, version 3.1*. Reference Guide. ILK Technical Report Series 07–08 [online]. Available : `http://ilk.uvt.nl/mbt` (10th Jan, 2009).

Gilles-Maurice de Schryver and Guy de Pauw. 2007. Dictionary Writing System (DWS) + Corpus Query Package (CQP): The Case of TshwaneLex. *Lexikos 17 AFRILEX-reeks/series* 17:2007: pp. 226 – 246. [Online tagger:] `http://aflat.org/?q=node/177`. (10th Feb, 2009)

Gilles-Maurice de Schryver and Daan 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): pp. 89 – 106.

Malcolm Guthrie. 1971. *Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages, vol 2*, Farnborough: Gregg International.

Ulrich Heid, Daan J. Prinsloo, Gertrud Faaß, and Elsabé Taljard. 2008 *Designing a noun guesser for part of speech tagging in Northern Sotho* (33 pp). ms: University of Pretoria.

Petronella M. Kotzé. 2008. Northern Sotho grammatical descriptions: the design of a tokeniser for the verbal segment. *Southern African Linguistics and Applied Language Studies* 26(2): pp. 197 – 208.

Zhang Le. 2004. *Maximum Entropy Modeling Toolkit for Python and C++* (Technical Report) [online]. Available: `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html` (10th Jan, 2009).

Geoffrey Leech, Andrew Wilson. 1999. Standards for Tagsets. in van Halteren (Ed.) *Syntactic world-class tagging*: pp. 55 – 80 Dordrecht/Boston/London: Kluwer Academic Publishers.

Louis J. Louwrens. 1991. *Aspects of the Northern Sotho Grammar* p. 154. Pretoria:Via Afrika.

Ramalau A. Maila. 2006. *Kgolo ya tiragatso ya Sepedi.* [="Development of the Sepedi Drama"]. Doctoral thesis. University of Pretoria, South Africa.

Oliver K. Matsepe. 1974. Tša Ka Mafuri. [="From the homestead"]. Pretoria: Van Schaik.

Daan J. Prinsloo. 1994. Lemmatization of verbs in Northern Sotho. *SA Journal of African Languages* 14(2): pp. 93 – 102.

Daan J. Prinsloo and Ulrich Heid. 2005. Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping. in: Isabella Ties (Ed.): *LULCL, Lesser used languages and computational linguistics*, 27/28-10-2005, Bozen/Bolzano, (Bozen: Eurac) 2006: pp. 97 – 113.

Daan J. Prinsloo, Gertrud Faaß, Elsabé Taljard, and Ulrich Heid. 2008. Designing a verb guesser for part of speech tagging in Northern Sotho. *Southern African Linguistics and Applied Languages Studies (SALALS) 26(2)*.

Helmut Schmid and Florian Laws. 2008. *Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging* [online]. COLING 2008. Manchester, Great Britain. Available: `http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/` (10th Jan, 2009).

Helmut Schmid. September 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*[online]. Available: `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/` (10th Jan, 2009).

Helmut Schmid. March 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*.

Drahomíra "johanka" Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. Jun 29, 2007. The best of two worlds: Cooperation of Statistical and Rule-based Taggers for Czech. *Balto-Slavonic Natural Language Processing*: pp. 67 – 74 [online]. Available: `http://langtech.jrc.it/BSNLP2007/m/BSNLP-2007-proceedings.pdf` (10th Jan, 2009).

Elsabé Taljard, Gertrud Faaß, Ulrich Heid and Daan J. Prinsloo. 2008. On the development of a tagset for Northern Sotho with special reference to standardisation. *Literator 29(1) 2008*. Potchefstroom, South Africa.

Raphehli M. Thobakgale. *Khuetšo ya OK Matsepe go bangwadi ba Sepedi* [="Influence of OK Matsepe on the writers of Sepedi"]. Doctoral thesis. University of Pretoria, South Africa.

Bertus van Rooy and Rigardt Pretorius. 2003. A word-class tagset for Setswana. *Southern African Linguistics and Applied Language Studies* 21(4): pp. 203 – 222.

Appendix A. The polysemy of *-a-*

	Description	Example
1	Subject concord of	*ge monna* ***a*** *fihla* conjunctive + noun cl. 1 + **subject concord cl. 1** + verb stem if/when + man + **subj-cl1** + arrive "when the man arrives"
2	Subject concord of nominal cl. 6	*masogana* ***a*** *thuša basadi* noun cl. 6 + **subject concord cl. 6** + verb stem + noun cl.2 young men + **subj-cl6** + help women "the young men help the women"
3	Possessive concord of nominal cl. 6	*maoto* ***a*** *gagwe* noun cl. 6 + **possessive concord cl. 6** + possessive pronoun cl. 1 feet + **of** + his "his feet"
4	Present tense morpheme	*morutiši o* ***a*** *bitša* noun cl. 1 + subject concord cl.1 + **present tense marker** + verb stem teacher + subj-cl1 + **pres** + call "the teacher is calling"
5	Past tense morpheme	*morutiši ga o* ***a*** *bitša masogana* noun cl. 1 + negation morpheme + subject concord cl.1 + **past tense marker** + verb stem + noun cl. 6 teacher + neg + subj-cl1 + **past** + call + young men "the teacher did not call the young men"
6	Demonstrative concord of nominal cl. 6	*ba nyaka masogana* ***a*** subject concord cl. 2 + verb stem + noun cl. 6 + **demonstrative concord** they + look for + young men + **these** "they are looking for these young men"
7	Hortative particle	***a*** *ba tsene* **hortative particle** + subject concord cl. 2 + verb stem let + subj-cl2 + come in "let them come in"
8	Interrogative particle	***a*** *o tseba Sepedi* **interrogative particle** + subject concord 2nd pers sg. + verb stem + noun cl. 7 **ques** + subj-2nd-pers-sg + know + Sepedi "do you know Sepedi"
9	Object concord of	*moruti o* ***a*** *biditše* noun cl. 1 + subject concord cl. 1 + **object concord cl. 6** + verb stem teacher + subj-cl1 + **obj-cl6** + called "the teacher called them"