Proceedings of the Second Workshop on African Language Technology
- AfLaT 2010 -

The verbal phrase of Northern Sotho: A morpho-syntactic perspective

Gertrud Faaß

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
Azenbergstraße 12, 70174 Stuttgart, Germany
gertrud.faasz@ims.uni-stuttgart.de

Abstract

So far, comprehensive grammar descriptions of Northern Sotho have only been available in the form of prescriptive books aiming at teaching the language. This paper describes parts of the first morpho-syntactic description of Northern Sotho from a computational perspective (Faaß, 2010a). Such a description is necessary for implementing rule based, operational grammars. It is also essential for the annotation of training data to be utilised by statistical parsers. The work that we partially present here may hence provide a resource for computational processing of the language in order to proceed with producing linguistic representations beyond tagging, may it be chunking or parsing. The paper begins with describing significant Northern Sotho verbal morpho-syntactics (section 2). It is shown that the topology of the verb can be depicted as a slot system which may form the basis for computational processing (section 3). Note that the implementation of the described rules (section 4) and also coverage tests are ongoing processes upon that we will report in more detail at a later stage.

1. Language introduction and overview

Northern Sotho is one of the eleven national languages of South Africa and one of the four Sotho languages of the South-eastern Language Zone group (S.30 in the classification of Guthrie (1971)): Northern and Southern Sotho, Tswana ("Western Sotho") and Lozi (Silozi, Rozi). The Sotho languages are written disjunctively (Van Wyk, 1958), i.e. a number of (mostly inflectional) affixes (henceforth "morphemes") are written separately instead of being merged with the stems that they belong to. Especially Northern Sotho verbs differ significantly from that of other languages, not only in respect to their moods, but also in the many different morpheme constellations that they may appear in.

Concerning its computational processing, Northern Sotho language resources and tools have been the subject of a few publications during recent years of which, for space reasons, we can only list a few: a methodology for tagging corpora of Northern Sotho has been designed by Prinsloo and Heid (2005), a number of finite-state tokenizers have been reported upon, concerning its verbs e.g. by Anderson and Kotzé (2006), a first tagset and a tagging experiment have been described by De Schryver and De Pauw (2007). Taljard et al. (2008) defined a more finely granulated tagset of which we partially make use in this paper. Lastly, Faaß et al. (2009) reported on results of tagging Northern Sotho with an approach of disambiguation of polysemous items.[1] This paper aims at designing a first morpho-syntactic framework for a number of Northern Sotho verbal constellations.[2] The paper does not claim to present a complete description, it is however shown here that a significant fragment of the verbal grammar can be depicted with a basic distinction made between elements syntactically subcategorised by the verb stem on the basis of its lexical semantic properties on the one hand, and its inflectional elements on the other.

The rules defined here are already in the process of being implemented in a rule-based parser (cf. (Faaß, 2010b) for a first report), however, they may also be used for other purposes, e.g. to produce training data for a statistical parser. Coverage tests are currently in process.

2. Northern Sotho verbs: An Introduction

Northern Sotho verbs appear in a variety of moods; for space reasons, we however focus on the independent predicative moods in this paper: indicative, situative and relative (for an overview of the Northern Sotho moods, see e.g. (Lombard, 1985, p. 144)). These moods all appear in the three tenses future, present and past and all may be negated. To mark tense or polarity, each of the moods makes use of specific morphemes to appear in front of the verb stem. Note that except for the missing tense marking, the dependent grammatical moods (consecutive, habitual and subjunctive) basically make use of the same morphemes as described here. The verb stem itself may show specific affixes fused to it, e.g. indicating some of the past constellations (allomorphs of the past tense morpheme *-il-*), and it appears with certain endings, inter alia *-a* or *-e*, that each are predefined by the constellation they must occur in. The ending *-a*, for example, usually appears in the positive constellations, while some negative ones require *-e* to appear. Other verbal affixes indicate certain semantic changes (e.g. passive: *-w-*, applicative voice: *-el-*, or reciprocal: *-an-*), of which most change its argument structure, however, these are not relevant for mood detection and thus not furtherly described here.

The indicative mood differs from the situative mood insofar as it forms a matrix sentence, cf. (1-a), while the situative

[1] For an overview of NLP resources of the African languages of South Africa, cf. (Roux and Bosch, 2006).

[2] This paper focuses on main verbs, it is developed on the grounds of (Faaß, 2010a).

mood usually does not, cf. (1-b). The relative mood depicts relative clauses, e.g. as in (1-c).

(1) a. *monna* *o* *reka dipuku*
man$_{N01}$ subj-3rd-01 buy books$_{N10}$
'(a) man buys books'

b. *ge* *monna* *a* *reka dipuku*
when man$_{N01}$ subj-3rd-01 buy books$_{N10}$
'when (a) man buys books'

c. *monna* *yo* *a*
man$_{N01}$ dem-3rd-01 subj-3rd-01
rekago *dipuku*
buy-rel books$_{N10}$
(a) man who buys books'

2.1. Elements of the Northern Sotho verb

This section is focusing on the following contents of the Northern Sotho verbs: The subject concord (relevant for subject-verb agreement, the verb stem and its objects, some negation clusters, and tense markers. These present the elements present in all grammatical moods of Northern Sotho.

2.1.1. Subject-verb agreement

Northern Sotho predicative verbs[3] all have to agree with their subject (in terms of either noun class[4] or person/number). A class specific subject concord marks this (usually anaphoric) relationship to the referent; if the relationship cannot be established, the neutral concord *e* is used.

In (1-a), $o_{\text{subj-3rd-01}}$ appears, a class 1 subject concord which is to be used for the positive indicative. In (1-a) (indicative) and (1-b) (situative), the verb stem (V) and the subject concord (CS_{categ}[5]) together form the linguistic verb. In the relative mood, cf. (1-c), a demonstrative concord $\text{CDEM}_{\text{categ}}$ is added. This concord – from a morpho-syntactic perspective – is not part of the verb itself, but precedes it (forming a CP containing a VP[6]). It is thus not described furtherly in this article. The ending *-go* internally marks the relative mood (see also paragraph 2.1.4.).

Like other Bantu languages, Northern Sotho is a null-subject language: in the case that the subject noun is omitted (e.g. because it is known in the discourse), the subject concord will acquire its grammatical function, cf. (2). For indefinite cases, similar to the English expletive 'it', the indefinite subject concord *go* appears with the grammatical function of a subject.

(2) *o* *reka dipuku*
subj-3rd-01 buy books$_{N10}$
'(s)he buys books'

[3]Non-predicative verbs of Northern Sotho are infinitives and imperatives, cf. e.g. Lombard's modal system (Lombard, 1985, p. 144 table 7.5.4).

[4]The noun classes of Northern Sotho are described in detail by e.g. (Lombard, 1985, p. 29 et seq.) and (Faaß, 2010a).

[5]$_{\text{categ}}$ stands for classes as defined by (Taljard et al., 2008): 01 - 10, 14, 15, LOC and the persons PERS_1sg to PERS_3pl respectively.

[6]The demonstrative concord also heads other constellations, like, e.g. adjectival phrases.

Each mood makes use of a specific type of subject concord. It is thus mandatory to define several sets of these so that they can be depicted as morpho-syntactic rules. Poulos and Louwrens (1994, p. 168 et seq.) define two sets, ignoring the fact that *o* and *a* (both described as being in set 1) occur in specific moods and are therefore not interchangable. We therefore opt for the definition of three sets, hence extending the labels, where set 1 contains *o* for class 1, and set 2 containing *a* instead (the subject concords of all other classes remain identical), and set 3 containing what is described as the "consecutive" (e.g. by Lombard (1985, p. 152 et seq.)). This set of subject concords also appears in non-consecutive moods, e.g. in one of the negated past tense forms of the relative mood, therefore we decide for a more neutral naming of the category, "3", instead.

Out three sets of subject concords are labeled accordingly as 1CS⟨categ⟩, 2CS⟨categ⟩ and 3CS⟨categ⟩; cf. examples in (3-a), where we repeat (1-a), (3-b), where we repeat (1-c), and (3-c).

(3) a. *monna* $o_{\text{1CS01(set1)}}$ *reka dipuku*
man$_{N01}$ subj-3rd-01 buy books$_{N10}$
'(a) man buys books'

b. *monna* *yo* $a_{\text{2CS01(set2)}}$
man$_{N01}$ dem-3rd-01 subj-3rd-01
rekago *dipuku*
buy-rel books$_{N10}$
'(a) man who buys books'

c. *lesogana* le_{CDEM05} $le_{\text{1CS05(set1)}}$
young man dem-3rd-05 subj-3rd-05
sego $la_{\text{3CS05(set3)}}$ *go toropong*
neg subj-3rd-05 go town-loc
(a) young man who did not go to town.

2.1.2. The verb stem and its arguments

The verb stem *reka* '[to] buy' in the examples above is transitive and followed by its object, *dipuku* 'books', hence one could indeed assume that Northern Sotho is an SVO language. However, this order of functional elements is only valid for overt objects and independent of the moods the verb appears in: in the case of an object being omitted, like e.g. in a discourse where it is already known, a pronominal object concord (CO_{categ}) has to appear in front of the verb stem, as shown in (4). Note that in such a case (i.e. where a positive indicative present tense ends in the verb stem), a tense marker is to be inserted after the subject concord (cf. paragraph 2.1.4.) .

Concerning double transitives, both objects may not be simultaneously substituted by means of a pronominal object concord. The Northern Sotho verb, unlike, e.g. Setswana verbs, may contain only one object concord (usually, the indirect object is pronominalized), an example of a pronominalisation process is shown in (5).

(4) *monna* *o* *a* *di* *reka*
man$_{N01}$ subj-3rd-01 pres obj-3rd-10 buy
'(a) man buys them'

(5) a. *ke* *diretše* *monna*
subj-1st-sg make-appl-perf man$_{N01}$
kofi
coffee$_{N09}$
'I made coffee for (the) man'

b. *ke* *diretše* *monna*
subj-1st-sg make-appl-perf man$_{N01}$

yona
emp-3rd-09
'I made it for the man'

c. *ke* *mo*
subj-1st-sg obj-3rd-01
diretše *yona*
make-appl-perf emp-3rd-09
'I made it for him/her'

The subject concord preceding this constellation can be set separately from the VBP, not only because it is an inflectional element that is relevant for the subject-verb agreement, but also because it can appear with any other verb stem+object(s) constellation. We therefore define an optional "Verbal Inflectional Element" (henceforth VIE) containing it. Note that the positive imperative VP of Northern Sotho contains a VBP only.

Though Northern Sotho is classified as a disjunctively written language, there are some object concords that are merged to the verb stem, like *N–*, referring to the first person singular. The verb stem $bona_{\text{V\_tr}}$ '[to] see', merges with this object concord to *mpona* '[to] see me', as in (6). From a syntactic perspective, such verbs have their argument structure saturated, therefore we use respective labels when tagging such tokens: "V_sat-tr" to indicate a saturated transitive verb. Half saturated double transitive verbs are labelled "V_hsat-dtr", cf. (7) containing the double transitive verb stem *fa* '[to] give', again merged with the object concord of the first person singular, *N–*, forming *mpha* 'give me'. Such specific annotations for fused forms, however, may become obsolete in the future, as a morphological analyser can split them, annotate them with their respective part of speech and report the two elements separately to the parser. Such an analyser already exists, cf. e.g. Anderson and Kotzé (2006) and some of its implemented routines could thus be utilised as a pre-processing element to parsing.

(6) *monna* *o* *a* ***mpona***$_{\text{V\_sat-tr}}$
man_{N01} subj-3rd-01 pres obj-1st-sg-see
'(a) man sees me'

(7) *monna* *o* ***mpha***$_{\text{V\_hsat-dtr}}$
man_{N01} subj-3rd-01 obj-1st-sg-give
dipuku
$\text{books}_{\text{N10}}$
'(a) man gives me books.'

2.1.3. Negation

The kinds of negation affixes (or clusters thereof) that are to appear in a verbal constellation of Northern Sotho vary significantly in the different moods (and tenses). Two examples are shown here: a negated indicative (8) and a negated situative in (9).

(8) *monna* ***ga*** ***a*** *reke dipuku*
man_{N01} neg subj-3rd-01 buy $\text{books}_{\text{N10}}$
'(a) man does not buy books.'

(9) *ge* *monna* ***a*** ***sa*** *reke dipuku*
when man_{N01} subj-3rd-01 neg buy $\text{books}_{\text{N10}}$
'when (a) man does not buy books'

(8) and (9) demonstrate that the order in which subject concord and negation morpheme(s) appear depends on the mood of the verb. Some negated forms, e.g. one of the negated forms of the relative, as shown in (3-c) above, even contain two subject concords surrounding the negation morpheme. Thus we need to define one slot containing both. Secondly, these examples show that the verb stem in a negated phrase often ends in *–e* instead of *–a*. The verbal ending in general, however, may itself be seen as underspecified information, because it does not determine a specific mood or tense, but a set of them. Together with certain contents of the VIE, it may be distinctive when identifying a certain mood/tense. We therefore add an attribute to the VBP: "Vend" describing the verbal ending that the verb stem has to appear with (cf. paragraph 2.). This attribute is to be stored with the correct value for each verb stem in the lexicon of the parser. Any morpho-syntactic rule defined thus can utilise the attribute "Vend" and assign a default value to it (cf. table 4 demonstrating the morpho-syntactic rules describing the indicative VP).

Such a method makes also sense in the case of irregular verbs: the so-called "defective" verb stem *re* '[to] say', for example, behaves like any regular verbs ending in *–a*. Following this method, we can define *re* as ending in *–a*.

2.1.4. Tense marking

The present tense morpheme *a* appearing in (4) and (6) above, is explicitly defined for the positive indicative verbal phrases that end in the verb stem (the so-called "long" form of the verb, cf. e.g. (Poulos and Louwrens, 1994, p. 208 et seq.)). All other present tense constellations do not contain a specific tense marker. The past tense is often marked with an allomorph of the verbal ending *-il–e*, which may appear inter alia as *-etš–e, -itš–e, -tš–e, -š–e* etc. Again, we can make use of the attribute "Vend" in describing all past tense forms as ending in *-ile*.

Consider examples (10) and (11), where (11) contains the subject concord of class 2 (usually referring to humans in the plural) and the past tense form of *sepela* 'walk', *sepetše*.

(10) *monna* *o* *rek**ile*** *dipuku*
man_{N01} subj-3rd-01 buy-past $\text{books}_{\text{N10}}$
'(a) man bought books'

(11) *ba* *sepe**tše***.
subj-3rd-02 walk-past
'they walked'

For other past tense forms, like the negated past relative shown in example (3-c), however, the constellation of subject concords and negation morpheme in the VIE is specific enough to mark the past tense. Here, the verb stem appears without any specific tense infix, and with the verbal ending *a*.

There are basically two interchangable future tense affixes found in Northern Sotho texts: *tlo* and *tla*, the relative mood of the future tense makes additional use of *tlogo* and *tlago* which appear whenever the verb stem does not have the relative suffix *-go* added. Any positive predicative mood can use one of these affixes which appear after the subject concord (or the cluster containing negation and subject concord) and before the object concord, as e.g. in (12).

(12) *monna* *o* ***tlo*** *di*
man_{N01} subj-3rd-01 fut obj-3rd-10
reka
$\text{books}_{\text{N10}}$
'(a) man will buy books.'

The slot system					
VIE		**VBP**			
zero-2	**zero-1**	**slot zero verb stem and its object(s)**			
pos-n to pos-3 subject conc. and/or negation	**pos-2** tense marker	**pos-1** object concord	**pos-0**$_{\text{Vend}}$ **verb stem**	**pos+1** object 1	**pos+2** object 2

Table 1: A schematic representation of the slot system

3. A slot system to describe the Northern Sotho verbal constellations

Table 1 demonstrates our knowledge of the Northern Sotho topology of the verb in a slot system: the "core" of the Northern Sotho verb contains the verb stem and its object(s), we call this part of the verb "Verbal Basic Phrase" (VBP). It is split into four positions: pos-1 can contain a separate object concord (or is empty), pos-0 contains the verb stem (possibly merged with an object concord). The optional pos+1 and pos+2 finally may contain overt objects.

To cater for the present and future tense affixes which always occur between the subject concord/negation cluster and VBP, the slot VIE contains two positions which we call zero-1 and zero-2 (while the four positions of the VBP slot are summarised as zero-0), cf. table 1. Slot zero-2 can contain several morphemes, zero-1 has only one position defined.

4. Computational processing

As stated above, inflectional (tense/negation) and subject-verb-agreement information are provided by subject concords and morphemes that precede the verb stem with the exception of the past tense and passive affixes which are merged to the verb stem, changing its ending. Such information can be described as attribute/value pairs in the lexicon. Concerning the examples stated above, the lexicon entries will appear as shown in table 2, where the parameter "Vend" appears in column 3. Such handling makes sense moreover in the case of irregular verbs: *re* '[to] say', for example, behaves like any other regular verb ending in *–a*. Any morpho-syntactic rule defined for a parser can utilise this parameter. Table 3 shows a simplified Lexical Functional Grammar (LFG) lexicon, containing some sample entries of morphemes and nouns. Their parts of speech and noun class are also encoded as attribute/value pairs. Additionally, number and person attribute/value pairs are listed for sake of information.

A typical morpho-syntactic rule describing the indicative mood fills the slot system as shown in table 4. The parentheses in *(–w)* indicate optionality, i.e. all verbs contained in these constellations may appear in their passive form.

In the framework of LFG, we can define e.g. a VIE-rule for the positive indicative (the "short" and the "long"' forms) as follows (the following shows a simplified version of the so far implemented operational grammar (Faaß, 2010b) for the sake of demonstration):

```
VP   →   VIE VBP.
         "General Verbal Inflectional Elements : VIE"
VIE  →   "short present tense form:"
         "subject concord of the first set"
         {   1CS : (↑ TNS-ASP FORM) = short;
         "no specific tense element"
         e : (↑ TNS-ASP TENSE) = pres
         "constraint: verbal ending in lexicon to be"
         "defined as a"
         (↑ VEND) =c a

         "long present tense form:"
         "subject concord of the first set"
         |   1CS
         "present tense morpheme indicates the long form"
         MORPH: (↑ TNS-ASP FORM) = long
         (↑ TNS-ASP TENSE) = pres;
         "constraint: verbal ending in lexicon must"
         "be defined as a"
         e : (↑VEND) =c a
         ...
         }
```

A fragment of these morpho-syntactic rules has been processed in the framework of LFG[7], and implemented in the Xerox Linguistic Environment provided by the Xerox Palo Alto Research Centre (PARC, `http://www2.parc.com/isl/groups/nltt/xle`) in the framework of a research license. It is planned to make this grammar available as a free web service.

5. Conclusions and future work

Concerning computational processing of Northern Sotho text, so far only linguistic "essentials" like tagsets (De Schryver and De Pauw, 2007; Taljard et al., 2008) and tagging procedures have been developed. Furthermore, finite-state machinery approaches that describe Northern Sotho's linguistic words have been delineated. As the next logical step in producing high level computational linguistic representations, morpho-syntactic descriptions of part-of-speech constellations forming verbal phrases are provided by this paper.

[7] See e.g. `http://www-lfg.stanford.edu/lfg`.

verb stem	label (transitivity)	verbal ending	comments
reka	V_tr	–a	'[to] buy'
reke	V_tr	–e	'[to] buy'
rekile	V_tr	–ile	'bought'
rekago	V_tr	–a-rel	'who buy(s)'
bona	V_tr	–a	'[to] see'
mpona	V_sat-tr	–a	'[to] see me', Obj=1st-sg
mphe	V_hsat-dtr	–a	'[to] give me', Obj=1st-sg
sepetše	V_itr	–il–e	'walked'
longwa	V_itr	-w–a	'[to] bite' passive form
ya	V_itr	-a	'walk"
ile	V_itr	–il–e	'walked'
re	V_tr	–a	'[to] say'

Table 2: Lexicon entries of verb stems

monna	n(oun)	class = 1, pers = 3, num = sg
lesogana	n(oun)	class = 5, pers = 3, num = sg
a	2CS	class = 1
	3CS	class = 1
	...	
MORPH	(↑ TNS-ASP TENSE) = pres, (↑ TNS-ASP FORM) = long	
le	1CS	class = 5
	CDEM	class = 5
	...	
tlo	MORPH	(↑ TNS-ASP TENSE) = fut.

Table 3: Sample lexicon entries of other parts of speech

$^{\text{INDPRES}}$VP

	VIE		VBP	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
pres.pos.long	$1CS_{categ}$	MORPH_pres	VBP	*(–w)–a*
pres.pos.short	$1CS_{categ}$		VBP	*(–w)–a*
pres.neg.	ga_{MORPH_neg} $2CS_{categ}$		VBP	*(–w) –e*
perf.pos.	$1CS_{categ}$		VBP	*–il(–w)–e*
perf.neg. 1	ga_{MORPH_neg} se_{MORPH_neg} $3CS_{categ}$		VBP	*(–w) –a*
perf.neg. 2	ga_{MORPH_neg} se_{MORPH_neg} $2CS_{categ}$		VBP	*(–w)–e*
perf.neg. 3	ga_{MORPH_neg} $3CS_{categ}$		VBP	*(–w)–a*
perf.neg. 4	ga_{MORPH_neg} $1CS_{categ}$ $–a_{MORPH\_past}$		VBP	*(–w) –a*
fut.pos	$1CS_{categ}$	*tlo/tla* $_{MORPH\_fut}$	VBP	*(–w)–a*
fut.neg	$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}		VBP	*(–w) –e*

Table 4: A summary of the independent indicative forms

Here, the verb's topology, i.e. the rather fixed order of the tokens that form the Northern Sotho verb allows for the definition of a slot system, fulfilling the following conditions:

- The central slot contains the verb and its object(s) or object concord (and second object in the case of a double transitive verb). Its contents solely depend on the lexical semantics of the verb stem in question. Standing alone, it describes an imperative VP.
- In the case of a negated imperative or a predicative VP[8], one to two preceding slots are to be defined of which
  - the first slot contains a subject concord of a predefined set and/or a specific negation morpheme (cluster)
  - the second slot is optional, however, if it occurs, it contains at maximum one element indicating tense.

Information on the verbal ending is often crucial when identifying certain moods, tenses or negated forms. There-

[8]We exclude the infinitive here, though it as well can be described in the slot system. The infinitive will be described in a separate publication, cf. (Faaß and Prinsloo, forthcoming)

fore, the lexical attribute "Vend" is introduced, which may be used as constraints by morpho-syntactic rules. It caters however not only for "regular" forms (like, e.g. "Vend" = *-a* for the verb rek*a*), but also for irregular forms ("Vend" = *-a* for the verb stem r*e*) and for allomorphs of the past tense morpheme *-il-*, e.g. *-etš-)*, as in *sepetše*.

Different moods make use of different sets of subject concords. So far, only two such sets have been described (e.g. by Poulos and Louwrens (1994)). The first of these sets contains two different subject concords for noun class 1, *o* and *a*. In order to be able to identify the precise mood/tense/polarity of a verb, we define two sets (1CS and 2CS) instead. The third set, so far beeing called the "consecutive" set, is renamed to the more neutral name "3CS" as it also occurs in moods other than the consecutive. Morpho-syntactic rules can now make use of these more precisely defined sets of parts of speech.

The design of a slot system for verbal phrases described in this paper is part of a project developing an operational LFG grammar of Northern Sotho (Faaß, 2010a) in the framework of XLE. Current work entails the extension of the grammar, and an implementation of a Northern Sotho noun phrase chunker making use of the CASS parser (Abney, 1996) and doing coverage tests. For utilising the descriptions for statistical parsing, morpho-syntactic analyses of these may be taken as a start point for the development of training data.

6. References

Abney, S. (1996). Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4), pp. 337–344.

Anderson, W. & Kotzé, P.M. (2006). Finite state tokenisation of an orthographical disjunctive agglutinative language: The verbal segment of Northern Sotho. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC-2006*. Genova, Italia: ELRA. [CD-ROM].

De Schryver, G-M. & De Pauw, G. (2007). Dictionary writing systems (DWS) + corpus query package (CQP): The case of Tshwanelex. *Lexikos*, 17, pp. 226 – 246.

Faaß, G. & Prinsloo, D.J. (forthcoming). A morpho-syntactic view on the infinitive of Northern Sotho.

Faaß, G., Heid, U., Taljard, E. & Prinsloo, D.J. (2009). Part-of-Speech tagging in Northern Sotho: disambiguating polysemous function words. In *Proceedings of the EACL2009 Workshop on Language Technologies for African Languages–AfLaT 2009*. Athens, Greece: Association for Computational Linguistics, pp. 38–45.

Faaß, G. (2010a). *A morphosyntactic description of Northern Sotho as a basis for an automated translation from Northern Sotho into English*. Ph.D. thesis, University of Pretoria, Pretoria, South Africa. (submitted).

Faaß, G. (2010b). A toy LFG grammar describing some of the morpho-syntactics of Northern Sotho verbs. Presented at Spring Pargram Meeting, University of Constance. [Online]. Available: http://www.ims.uni-stuttgart.de/~faaszgd/spring-pargram-2010-NSotho-handout.pdf (accessed March 2010).

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

Lombard, D.P. (1985). *Introduction to the Grammar of Northern Sotho*. Pretoria, South Africa: J.L. van Schaik.

Poulos, G. & Louwrens, L.J. (1994). *A Linguistic Analysis of Northern Sotho*. Pretoria, South Africa: Via Afrika.

Prinsloo, D.J. & Heid, U. (2005). Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping. In I. Ties (Ed.), *LULCL Lesser used languages and computational linguistics*. Bozen, Italy: Eurac, pp. 97–113.

Roux, J. & Bosch, S.E. (2006). Language resources and tools in Southern Africa. In *Proceedings of the Workshop on Networking the Development of Language Resources for African Languages. 5th International Conference on Language Resources and Evaluation*. Genova, Italy: ELRA, pp. 11–15.

Taljard, E., Faaß, G., Heid, U. & Prinsloo, D.J. (2008). On the development of a tagset for Northern Sotho with special reference to the issue of standardization. *Literator – special edition on Human Language Technologies*, 29(1), pp. 111–137.

Van Wyk, E.B. (1958). *Woordverdeling in Noord-Sotho en Zulu (Word division in Northern Sotho and Zulu)*. Ph.D. thesis, DLitt thesis. University of Pretoria, Pretoria, South Africa.