EXPLORING THE TREATMENT OF SELECTED TYPOLOGICAL CHARACTERISTICS OF TSWANA IN LFG

Ansu Berg
North-West University

Rigardt Pretorius
North-West University

Laurette Pretorius
University of South Africa

Proceedings of the LFG12 Conference

Miriam Butt and Tracy Holloway King
(Editors)

2012

CSLI Publications

http://csli-publications.stanford.edu

Abstract

Tswana is a Bantu language in the south eastern zone of Bantu languages and one of the eleven official languages of South Africa. From a linguistic as well as a computational point of view the language is a lesser-studied and under-resourced language. Recently a project was undertaken to describe the syntactic structure of Tswana in the LFG formalism and to develop parser support for this using the XLE parser toolkit. In this paper a short overview of selected distinctive typological characteristics of Tswana, including the word order, agglutinative character, disjunctive orthography and agreement, is presented. The aim is to demonstrate how selected instances of agreement in Tswana can be modeled in LFG and XLE.

1 Introduction

Tswana is a Bantu language in the south eastern zone (zone S in Doke's classification) of Bantu languages and is one of the three languages in the Sotho language group (Cole, 1959; Guthrie, 1971). The other two Sotho languages are Northern Sotho (Sepedi) and Southern Sotho. Tswana is predominantly spoken in South Africa and Botswana. It is one of the eleven official languages of South Africa with approximately 3,272,720 (8.2% of the population) first language speakers (Statistics South Africa Census 2001, 2004: 9). In Botswana, Tswana is the only national language and approximately 1,070,000 (79,06% of the population) in Botswana speak Tswana as first language (Central Statistics Office, 2009:14).

Existing grammatical descriptions of Tswana, mainly focussing on the morphology, are to a large extent based on the structural functional approach (Cole, 1955; Krüger, 2006). Recently a description of the syntactic structure of Tswana in the constraint-based LFG formalism (Dalrymple, 2001) was commenced.

In recent years the following core technologies in Natural Language Processing (NLP) were developed for Tswana:

- a proposed word-class tagset (Van Rooy and Pretorius, 2003)
- a lemmatiser (Brits *et al.*, 2005; Brits, 2006)
- a morphological analyser and a tokeniser (Pretorius *et al.*, 2009)

This work is part of a bigger project to develop a syntactic parser for Tswana, which will in due course form part of the NLP pipeline for Tswana. An LFG description of the syntactic structures of Tswana is undertaken, which will serve as basis for the parser development, using the XLE parser toolkit (Crouch *et al*., 2011).

Agreement is a typical typological characteristic of Tswana. The aim of this paper is to describe the employment of selected instances of subject-verb and noun phrase internal agreement in LFG and XLE. For this purpose an overview of the agglutinative character, word order, and disjunctive orthography of Tswana is provided.

2 Typological features of Tswana

2.1 Agglutinative language

Tswana is an agglutinative language. It is characterised by a complex morphology where affixation is prominent. Affixes modify or extend the meaning of words (Krüger, 2006:40-41).

The noun class prefixes provide essential information regarding class and number features of nouns. Noun suffixes extend the meaning of nouns and provide information regarding certain characteristics (Krüger, 2006: 73-96). In example (1) the meaning of the noun **setlhare** is extended by adding the diminutive and locative suffixes:

(1) **(mo) setlharenyaneng** '(here) in the little tree'
setlharenyaneng
se- + -tlhare + -ana + -ing
NPre7+tree+Dim+Loc

The verbal prefixes and suffixes provide essential information regarding type, tense, aspect and mood (Krüger, 2006: 198-243). The verb in example (2) comprises a verbal root **-thus-** to which prefixes (subject and object agreement morphemes) and suffixes (causative, perfect, verbal ending) are added:

(2) **ba re thusitse (go fetsa tiro ya rona)**
'they helped us (to finish our work)'
ba re thusitse
ba + re + thus- + -is- + -il- + -e
AgrSubj-Cl2+AgrObj-p1-Pl+help+Caus+Perf+VerbEnd

2.2 Word classes

Tswana words are divided into word classes on the basis of similarities between certain words. Words in Tswana have autonomous word status. This has been proven through the application of one or more of the word tests, namely isolatability, separability, transposability and replaceability (Van Wyk, 1967:230-261; Krüger, 2006:12-16; Louwrens and Poulos, 2006:392). The major word categories in Tswana are nouns, verbs, pronouns, particles, adverbs, idiophones and interjections. Within these categories nouns and verbs are open word classes on the basis of their productive morphology. Pronouns, particles, adverbs, interjections and

idiophones are closed word classes because they can be exhaustively listed and are morphologically unproductive.

Sub-categories of the word classes are also distinguished on the grounds of similarities between words within a specific word category. The sub-categories of the nouns include amongst others basic nouns, adjectival nouns, deverbative nouns and locative nouns. An adjective is a special kind of noun because it is characterised by a class prefix which changes in accordance with the class of the qualified noun (Lombard *et al*., 1993:57). Proper verbs, auxiliary verbs and copulative verbs are sub-categories of verbs (Krüger, 2006:24). Absolute pronouns, demonstrative pronouns, quantitative pronouns and possessive pronouns are sub-categories of pronouns (Krüger, 2006:24). The associative particle, instrumental particle, locative particle, possessive particles, qualificative particles as well as the conjunctions are sub-categories of the word class particle (Krüger, 2006:25).

Basic Tswana nouns consist morphologically of a class prefix and a root. Twenty noun classes are distinguished in Tswana (Krüger, 2006:57-70). In classes 1 to 14 the even classes contain singular nouns and the odd classes contain plural nouns. Classes 15 to 20 do not refer to singular or plural (Krüger, 2006:57-124).

Example (3) shows a singular and plural noun with the same root. The singular noun class prefix ***mo-*** indicates a class 1 noun, while the plural class prefix ***ba-*** indicates a noun in class 2:

(3)	**monna** 'man'	>	**banna** 'men'
	mo- + *-nna*		***ba-*** + *-nna*
	NPre1+man		**NPre2**+men

The agreement system in Tswana is based on the noun class prefixes (Louwrens, 1994:9-10). Words in a syntactic relation to a specific noun exhibit formal similarities with the class prefix of that noun. For example, the class prefix of the subject noun determines the form of the subject agreement morpheme on the verb:

(4) **Banna ba bua Setswana.** 'The men speak Tswana.'

banna	**ba** bua	Setswana
ba- +*-nna*	***ba*** + *bu-*+ *-a*	*se-* + *-tswana*
NPre2+men	**AgrSubj-Cl2**+speak+VerbEnd	NPre7+tswana

Nouns are also used in relation to other words which modify them. The noun class of the head noun determines the agreement affixes on modifiers within the NP as illustrated in example (5):

(5) **banna ba** ‘these men’ (‘men these’)

banna	**ba**
ba- *+ -nna*	***ba***
NPre2+men	**DemPro-Cl2**

Words in a syntactic relation to nouns therefore present formal similarities to that noun (Louwrens, 1994:9-10). Various morphemes and words are derived from the noun class prefixes. This applies to, for example, the forming of subject agreement morphemes, object agreement morphemes, pronouns such as absolute pronouns, possessive pronouns, demonstrative pronouns and particles such as possessive particles and qualificative particles, etc. In example (6) agreement regarding noun class 8 is illustrated:

(6) An example of agreement in noun class 8:

• ***noun class prefix***

di: ditlhako ‘shoes’

ditlhako
di- *+ -tlhako*
NPre8+shoes

• ***subject agreement morpheme***

di: Ditlhako di latlhegile. ‘The shoes got lost.’

ditlhako	di latlhegile
di- + -tlhako	***di*** *+ latlheg- + -il- + -e*
NPre8+shoes	**AgrSubj-Cl8**+lose+Perf+VerbEnd

• ***consecutive subject agreement morpheme***

tsa: Ditlhako tsa latlhega. ‘The shoes then got lost.’

ditlhako	tsa latlhega
di- + -tlhako	***tsa*** *+ latlheg- + -a*
NPre8+shoes	**AgrSubjCons-Cl8**+lose+VerbEnd

• ***object agreement morpheme***

di: Basadi ba a di reka. ‘The women buy it.’

basadi	ba a di reka
ba- + -sadi	*ba + a +* ***di*** *+ rek- + -a*
NPre2+women	AgrSubj-Cl2+AspPr+**AgrObj-Cl8**+buy+VerbEnd

• ***absolute pronoun***

tsone: Ditlhako tsone re di rekile. ‘As for the shoes, we bought them.’

ditlhako	tsone	re di rekile
di- + -tlhako	***tsone***	*re + di + rek- + -il- + -e*
NPre8+shoes	**AbsPro-Cl8**	AgrSubj-p1-Pl+AgrObj-Cl8+buy +Perf+VerbEnd

• ***demonstrative pronoun (distance 1)***

tse:	ditlhako tse ‘these shoes’	
	ditlhako	tse
	di- + -tlhako	***tse***
	NPre8+shoes	**DemPro-Cl8-d1**

• ***demonstrative pronoun (distance 2)***

tseo:	ditlhako tseo ‘those shoes’	
	ditlhako	tseo
	di- + -tlhako	***tseo***
	NPre8+shoes	**DemPro-Cl8-d2**

• ***demonstrative pronoun (distance 3)***

tsele:	ditlhako tsele ‘those shoes “over there” ’	
	ditlhako	tsele
	di- + -tlhako	***tsele***
	NPre8+shoes	**DemPro-Cl8-d3**

• ***possessive particle***

tsa:	ditlhako tsa basadi ‘women’s shoes’		
	ditlhako	tsa	basadi
	di- + -tlhako	***tsa***	*ba- + -sadi*
	NPre8+shoes	**PosPart-Cl8**	NPre2+women

• ***possessive pronoun***

tsone:	mebala ya tsone ‘their colours’ (‘colours of them’)		
	mebala	ya	tsone
	me- + -bala	*ya*	***tsone***
	NPre4+colours	PosPart-Cl4	**PosPro-Cl8**

• ***qualificative particle***

tse:	ditlhako tse dintsi ‘many shoes’ (‘shoes that are many’)		
	ditlhako	tse	dintsi
	di- + -tlhako	***tse***	*di- + -ntsi*
	NPre8+shoes	**QualPart-Cl8**	NPre8+many

2.3 Disjunctive orthography

A disjunctive orthography is used for Tswana verbs (Kosch, 1993:43). The prefixes are usually written disjunctively but the suffixes are written conjunctively. The consequence of this writing style is that Tswana words cannot be tokenised only on white space (Pretorius *et al.*, 2009). The correct identification of Tswana word boundaries is essential in the identification of the constituents of Tswana sentences. This is illustrated and explained in examples (7) and (8):

(7) **o a e reka** 'she buys it'
o + *a* + *e* + *rek-* + *-a*
AgrSubj-Cl1+Asp+AgrObj-Cl9+buy+VerbEnd

The verb **o a e reka** in example (7) represents one token and this token consists of four orthographic items.

(8) **Mosadi o badile dibuka.** 'The woman read the book.'

mosadi	o badile	dibuka
mo- + *-sadi*	*o* + *bal-* + *-il-* + *-e*	*di-* + *-buka*
NPre1+woman	AgrSubj-Cl1+read+Perf+VerbEnd	NPre10+book

The Tswana sentence in example (8) consists of three tokens, namely *mosadi /o badile / dibuka*. While the verb **o badile** consists of two orthographic items it represents only one token.

In Tswana a linguistic verb can be a sequence of orthographic items that together function as members of that verb. These orthographic items are also referred to as orthographic words (Louwrens and Poulos, 2006:393).

2.4 Word order

2.4.1 Word order in a simple Tswana sentence

The basic word order in simple Tswana sentences is SVO where the subject precedes a verb and an object appears post verbally (Krüger, 2006:11-12). This word order is illustrated in the following example:

(9) **Banna ba bua Setswana.** 'The men speak Tswana.'

Subject	Verb	Object
banna	ba bua	Setswana
ba- + *-nna*	*ba* + *bu-* + *-a*	*se-* + *-tswana*
NPre2+men	AgrSubj-Cl2+speak+VerbEnd	NPre7+tswana

The phrase structure rule indicating the word order followed in a simple Tswana sentence is as follows:

S →	NP	VP
	(↑SUBJ)=↓	↑=↓

Tswana phrases are head initial. The Tswana verb phrase (VP) can be made up of a verb and a noun phrase (NP) and the following phrase structure rule is then followed (Department of African Languages and Literature, 2000:10):

VP →	V	NP
	↑=↓	(↑OBJ)=↓

One or more obliques or adjuncts referring to place, time, manner, etc. can be incorporated in a simple Tswana sentence. An SVOX word order is then

followed, where 'X' represents the obliques and adjuncts (Creissels, 2000:250-252).

2.4.2 Word order in Tswana noun phrases

In a Tswana NP the head appears in initial position and it is followed by a variable number of modifiers (determiners) (Creissels, 2000:232). Several nominal words or word phrases can act as the head of the NP. If the head is a noun, the following phrase structure rule is used (King and Dalrymple, 2004:71):

NP →	N	Det
	↑=↓	↑=↓

Within the NP, all of the modifiers follow the head noun and modify the head regarding some quality or characteristic (Krüger, 2006:301). Examples of noun modifiers are:

- **a pronoun (personal, absolute, demonstrative, quantitative)**

(10) **banna ba** 'these men'

banna	ba
ba- + -nna	*ba*
NPre2+men	DemPro-Cl2-d1

- ***a possessive phrase (consists of a possessive particle and a complement such as a noun)***

(11) **sekolo sa basimane** 'the boys' school'

sekolo	sa	basimane
se- + -kolo	*sa*	*ba- + -simane*
NPre7+school	PosPart-Cl7	NPre2+boys

- **a qualificative phrase (consists of qualificative particle and a complement such as an adjective)**

(12) **banna ba bagolo** 'big men'

banna	ba	bagolo
ba- + -nna	*ba*	*ba- + -golo*
NPre2+men	QualPart-Cl2	NPre2+big

2.5 Agreement

Agreement in Tswana is observed in the relationship between verbs and nouns and all instances where nouns occur in relationship with other words that modify (qualify) them (Watters, 2000:202).

2.5.1 Subject-verb agreement in a Tswana sentence

Subject verb agreement is established through the subject agreement morpheme on the verb. Subject agreement morphemes agreeing with nouns indicate class. The class feature implies number (singular or plural). Subject

agreement morphemes agreeing with personal pronouns indicate person and number (Krüger, 2006: 171-175).

The following sentence (example (13)) has a noun as subject. The CLASS feature is included in both the subject (noun) and the verb. The subject agreement morpheme agrees with the class of the noun and therefore obtains a valid f-string:

(13) **Basadi ba reka dijo.** 'The women buy food.'

basadi	**ba** reka	dijo
ba- **+** *-sadi*	***ba*** *+ rek- + -a*	*di- + -jo*
NPre2+women	**AgrSubj-Cl2**+buy+VerbEnd	NPre8+food

basadi: (↑PRED) = 'BASADI'
(↑CLASS) = 2
ba reka: (↑PRED) = 'BA REKA <SUBJ OBJ>'
(↑SUBJ CLASS) = 2
dijo: (↑PRED) = 'DIJO'
(↑CLASS) = 8

2.5.2 NP-internal agreement in Tswana

Modifiers (determiners) in a NP agree with the head noun (Louwrens, 1994:52). Apart from locative noun phrases where the demonstrative pronoun precedes the noun, all other nominal modifiers are post modifiers (Louwrens, 1994:10). They are the different pronouns, the possessive phrase, the adjectival phrase and the verbal relative phrase.

The CLASS feature is used to validate NP internal agreement at the level of functional structure. If the specification from the head noun can unify with the specification of the determiner then a valid f-structure is obtained.

Examples (14), (15) and (16) show NP internal agreement:

- ***Noun (Head) + demonstrative pronoun***

In a NP in which the head noun is modified by a demonstrative pronoun, both the noun and demonstrative pronoun must show the same class information:

(14) **sekolo seo** 'that school' ('school that')

sekolo	seo
se- *+ -kolo*	***seo***
NPre7+school	DemPro-Cl7

sekolo: (↑PRED) = 'SEKOLO'
(↑CLASS) = 7
seo: (↑PRED) = 'SEO'
(↑CLASS) = 7

- ***Noun (Head) + possessive phrase***

Tswana only has a long possessive construction (Krüger, 2006: 139-145). The possessive particle agrees with the head noun (example (15)). Should the CLASS specification from the head unify with that of the possessive particle then a valid f-structure is obtained.

(15) **sekolo sa basimane** 'the boys' school' ('school of boys')

sekolo	sa	basimane
se- + *-kolo*	*sa*	*ba-* + *-simane*
NPre7+school	**PosPart-Cl7**	NPre2+boys

sekolo: (↑PRED) = 'SEKOLO'
(↑CLASS) = 7
sa: (↑PRED) = 'SA'
(↑CLASS) = 7
bone: (↑PRED) = 'BASIMANE'
(↑CLASS) = 2

- ***Noun (Head) + adjectival phrase***

NP's in which an adjective modifies a head noun in Tswana differ in internal structure from the equivalent NP's in English (example (16)) (Krüger, 2006:150). In these phrases both the qualificative particle and the adjective agree with the head noun. The head noun, qualificative particle and the adjective must show the same class information.

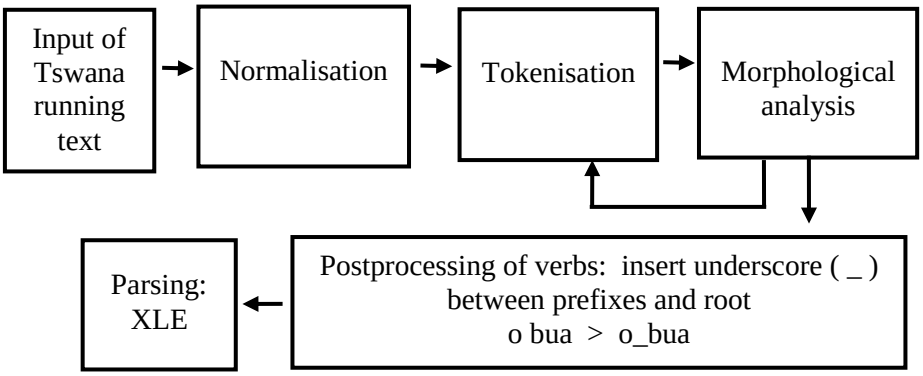
(16) **banna ba bagolo** 'big men' ('men which are big')

banna	ba	bagolo
ba- + *-nna*	***ba***	***ba-*** + *-golo*
NPre2+men	QualPart-Cl2	NPre2+big

banna: (↑PRED) = 'BANNA'
(↑CLASS) = 2
ba: (↑PRED) = 'BA'
(↑CLASS) = 2
bagolo: (↑PRED) = 'BAGOLO'
(↑CLASS) = 2

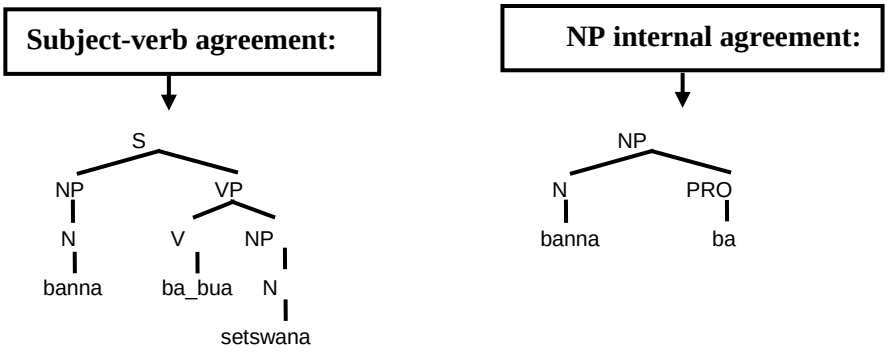
3 The treatment of Tswana agreement in XLE

The Tswana NLP pipeline can be presented schematically as follows:



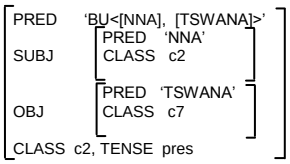
(Crouch *et al*., 2011; Pretorius *et al*., 2009)

The following section shows the LFG description of examples (9) and (10), and their modelling in XLE. Agreement is validated by the CLASS specification of the subject and the verb (example 9) and that of the head and modifier in (example 10):



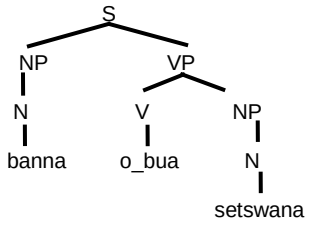
1 valid f-structure

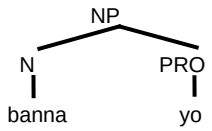
'banna ba_bua Setswana"



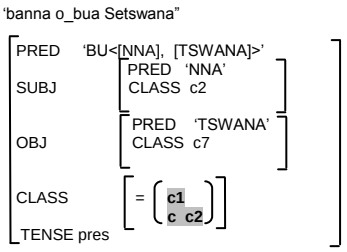
1 valid f-structure

'banna ba"

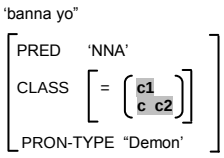




0 valid f-structure (inconsistent)



0 valid f-structure (inconsistent)



4 Conclusion

Tswana has distinctive typological characteristics regarding agglutinative character, word classes, disjunctive orthography, word order and agreement. Agreement is based on the noun classes and personal pronouns. Only a limited number of instances of agreement are presented in this paper. The agreement features of LFG are found to be appropriate for the modelling of some instances of subject-verb and NP-internal agreement in Tswana. Their implementation in XLE could also be done satisfactorily.

As a next step a description of the structure of the simple Tswana sentence will be attempted. All instances of subject-verb and NP-internal agreement will then be described in LFG and XLE.

References

Brits, Karien. 2006. Outomatiese Setswana Lemma-identifisering. Potchefstroom: NWU. (Dissertation – MA). 205p.

Brits, Karien, Pretorius, Rigardt and Van Huyssteen, Gerhard. 2005. Automatic Lemmatization in Setswana: Towards a Prototype. *South African Journal of African Languages*, 25(11):37-47.

Central Statistics Office. 2009. Botswana Demographic Survey 2006. Gaborone: Government Printer. 144p.

Cole, Desmond. 1955. An Introduction to Tswana Grammar. Cape Town: Longman. 473p.

Cole, Desmond. 1959. Doke's classification of Bantu Languages. *African Studies*, 18(4):197-213.

Creissels, Dennis. 2000. Typology. (*In* Heine, Bernd and Nurse, Derek, *eds*. African Languages. An Introduction. Cambridge: Cambridge University Press. p.231-258.)

Crouch, Dick, Dalrymple, Mary, Kaplan, Ron, King, Tracy Holloway, Maxwell, John and Newman, Paula. XLE Documentation. California: PARC. http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html.

Dalrymple, Mary. 2001. Lexical Functional Grammar. Syntax and Semantics Series, Volume 34. New York: Academic Press. 473p.

Department of African Languages and Literature (University of Botswana). 2000. The structure of Setswana sentences. An introduction. Gaborone: Lightbooks. 73p.

Guthrie, Malcolm. 1971. Comparative Bantu: An introduction to the Comparative Linguistics and Prehistory of the Bantu languages. Volume 2. Farnborough: Greg International Publishers. 121p.

King, Tracy Holloway and Dalrymple, Mary. 2004. Determiner agreement and noun conjunction. *Journal of Linguistics*, 40:69-104.

Kosch, Inge. 1993. A historical perspective on Northern Sotho linguistics. Pretoria: UP. (Thesis – PhD). 216p.

Krüger, Caspar. 2006. Introduction to the Morphology of Tswana. München: Lincom. 314p.

Lombard, Daan, Van Wyk, Egidius and Mokgokong, Pothinus. 1993. Introduction to the grammar of Northern Sotho. Pretoria: J.L. van Schaik. 203p.

Louwrens. Louis. 1994. Dictionary of Northern Sotho Grammatical Terms. Pretoria: Via Afrika. 257p.

Louwrens, Louis and Poulos, George. 2006. The status of the word in selected conventional writing systems - the case of disjunctive writing. *Southern African Linguistics and Applied Language Studies*, 24(3):389-401.

Pretorius, Rigardt, Berg, Ansu, Pretorius, Laurette and Viljoen, Biffie. 2009. Setswana Tokenisation and Computational Verb Morphology: Facing the Challenge of a Disjunctive Orthography. (*In* AfLaT2009. Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages. Hosted by the European Association for Computer Linguistics. Athens, Greece. p.66-73.)

Statistics South Africa. 2004. Census 2001: Primary tables South Africa. http://www.statssa.gov.za/census01.html.

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

Van Wyk, Egidius. 1967. The word classes of Northern Sotho. *Lingua*, 17: 230-261.

Watters, John. 2000. Syntax. (*In* Heine, Bernd and Nurse, Derek, *eds*. African Languages: An introduction. Cambridge: Cambridge University Press. p.194-230.)

Appendix: Tags in the text

Tag	**Meaning**
AbsPro	Absolute pronoun
AgrObj	Object agreement morpheme
AgrSubj	Subject agreement morpheme
AgrSubjCons	Consecutive subject agreement morpheme
AspPr	Aspectual prefix
Caus	Causative suffix
Cl	Noun class
DemPro	Demonstrative pronoun
Dim	Diminutive suffix
d1	Distance 1
d2	Distance 2
d3	Distance 3
Loc	Locative suffix
NPre	Noun prefix
Perf	Perfect
Pl	Plural
PosPart	Possessive particle
QualPart	Qualificative particle
VerbEnd	Verbal ending