

A toy LFG grammar describing some of the morpho-syntactics of Northern Sotho verbs

Gertrud Faab

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12
D 70174 Stuttgart

Spring Pargram Meeting
25th March 2010

Faah (IMS)

A toy grammar

Northern Sotho 1 / 32

Project introduction

- PhD Thesis:
A morphosyntactic description of Northern Sotho as a basis for an automated translation from Northern Sotho into English
- Supervisors:
Main: Prof. Danie J Prinsloo, University of Pretoria, South Africa
Co: PD Dr. phil. habil. Ulrich Heid, apl. Prof., Universität Stuttgart, Germany
- Dates/Status:
 - Sep 2006 – Sep 2009
 - Submitted Jan 2010 for the degree of Ph.D. in African Languages
 - University of Pretoria, South Africa



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA



Faah (IMS)

A toy grammar

Northern Sotho 3 / 32

Project introduction

Contents of the thesis

- Begins with prescriptive studies, summarises and moves the perspective to computational processing
- Describes a substantial grammar fragment
 - Language units are identified (tokens and words) and sorted into POS categories
 - Formal relationships between these units are defined in the framework of generative grammar
- Attempts to find generalisations on the contextual distributions of the language units
- Designs a toy grammar containing the constellations of the indicative mood
- Adds basic definitions for a proposed machine translation into English (a first attempt for a contrastive description)

Faah (IMS)

A toy grammar

Northern Sotho 5 / 32

This Talk

- Project introduction
- Language introduction
- Some word classes (parts of speech)
- The verbal phrase: basic layout
- A toy LFG grammar of Northern Sotho
- I got some questions
- Discussion?

Faah (IMS)

A toy grammar

Northern Sotho 2 / 32

Project introduction

View on previous works - resources

- Available so far
 - Prescriptive comprehensive study books (last publication: 1994)
 - Some noisy corpora: ca. 5 million orthographic tokens (not available for the public, but accessible on request)
 - Descriptive articles on singled out phenomena
 - * Tokenization (orthographic versus linguistic tokens)
 - * Morphological issues (mainly about verbs and nouns)
 - * Tagging (so far, two approaches):
One tagger is available on line (www.aflat.org)
- Not available:
 - Clean corpora of considerable size (current gold standard: 45,000 tokens)
 - Morphological analyser: finite state machinery (however described by UNISA, dept. of African Languages)
 - Any comprehensive grammar from a computational perspective

Faah (IMS)

A toy grammar

Northern Sotho 4 / 32

Language Introduction

The "Bantu" Languages (Niger Congo)

- Central, East and Southern Africa (a couple of thousand languages)
- Several writing systems (conjunctive versus disjunctive)
- Rather rigid word order: morphological marking mostly absent
- Noun class system (see below)

Faah (IMS)

A toy grammar

Northern Sotho 6 / 32

Language Introduction

The Sotho Group: spoken in South Africa and its neighbouring countries
 source: David Joffe's African Languages page: <http://afzicoolanguages.com/>

- SeSotho (Southern Sotho):
 - Official language of South Africa: ca. 3.5 million speakers
 - Official language of Lesotho: ca. 2.1 million speakers
- SeTswana (Tswana, Western Sotho):
 - Official language of South Africa, ca. 3.6 million speakers
 - Official language of Botswana, ca. 1.1 million speakers
- SeSotho sa Leboa (Northern Sotho, #Sepedi):
 - Official language of South Africa (one of eleven): ca. 4.2 million speakers
 - A standardised written form of about 30 dialects

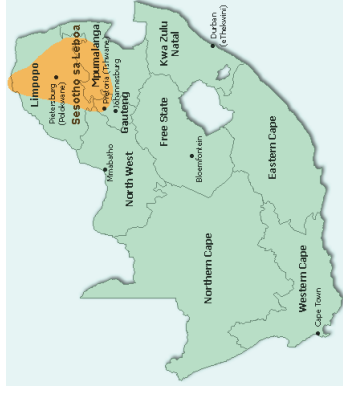
Faaf (IMS)

A toy grammar

Northern Sotho 7 / 32

Language Introduction

Northern Sotho: Geographical distribution
 source: David Joffe's African Languages page: <http://afzicoolanguages.com/>



Faaf (IMS)

A toy grammar

Northern Sotho 9 / 32

Language Introduction

Nguni Languages
 source: Tajiri and Bosch: A Comparison of Approaches to Word Class Tagging: Disjunctively vs. Conjunctively Written Bantu Languages (2006), N.J.A.S. [Tajiri and Bosch(2006)]
 Nguni

- Similar grammars and lexicons
- However: Conjunctive writing system
- Example: Comparison of a verb of Northern Sotho and the same verb of IsiZulu

Northern Sotho	ba	a	mo	thuša
IsiZulu	they	pres	him/her	help
	ba	-ya-	m-	-siza
				<i>bayamsiza</i>

- Similarity on the level of morpheme
- Finite state processing of IsiZulu: UNISA, Profs. Bosch/Pretorius

Faaf (IMS)

A toy grammar

Northern Sotho 11 / 32

Language Introduction

The Sotho Group: similarities

- Similar grammars and lexicons
 Northern Sotho: *lapile* = [be] tired (rarely: [be] hungry)
 Southern Sotho: *lapile* = [be] hungry (not: [be] tired)
- Closed class items very similar
- Disjunctive writing system
 → Transfer is possible

Faaf (IMS)

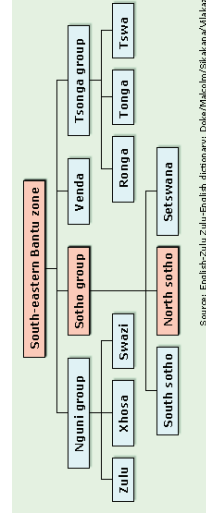
A toy grammar

Northern Sotho 8 / 32

Language Introduction

Sotho vs. Nguni vs. others

source: David Joffe's African Languages page: <http://afzicoolanguages.com/>



Faaf (IMS)

A toy grammar

Northern Sotho 10 / 32

Some word classes (parts of speech)

Bantu Languages in general make use of a noun class system

Properties of nominal items

- No gender (num can be retrieved in most cases, see next slide)
- Noun classes
 - Up to 24 noun classes ([Mutaka(2000), p. 151])
 - Northern Sotho makes use of 1 – 10, 14, (15), 16 – 18, nasal- and *ga*-locative classes
 - Class 15 = infinitive (verbal properties - separate issue)
 - Classes 16 – 18, nasal- an *ga*- are all locative classes (we summarise them as "loc")
- Lack of determiners (NP → N)

Faaf (IMS)

A toy grammar

Northern Sotho 12 / 32

Some word classes (parts of speech)

Noun classes - examples

class	noun	translation
1	<i>Modimo</i>	God (common noun?)
1	<i>modimo</i>	ghost, spirit of a deceased
2	<i>badimo</i>	ghosts/spirits of the deceased
3	<i>modimo</i>	evil spirit
4	<i>medimo</i>	evil spirits
5	<i>ledimo</i>	thunderstorm
6	<i>madimo</i>	thunderstorms
7	<i>sedimo</i>	sacrifice
8	<i>didimo</i>	sacrifices
14	<i>bodimo</i>	cannibalism
LOC	<i>godimo</i>	high above, in the air

[Heid et al.(2009)Heid, D.J., Faaß, and Tajjard] describe the design of a noun guesser for POS-tagging (accuracy ca. 92%)

Faaß (IMS)

A toy grammar

Northern Sotho 13 / 32

Some word classes (parts of speech)

Concordial items

Table: Some subject concords

set	PERS.1sg	PERS.2sg	class					
1	<i>ke</i>	<i>o</i>	1	2	3	4	5	6
2	<i>ke</i>	<i>o</i>	..	<i>o</i>	<i>ba</i>	<i>o</i>	<i>e</i>	<i>le</i>
3	<i>ka</i>	<i>wa</i>	..	<i>a</i>	<i>ba</i>	<i>o</i>	<i>e</i>	<i>le</i>
			..	<i>a</i>	<i>ba</i>	<i>wa</i>	<i>ya</i>	<i>la</i>

Different moods make use of different sets of subject concords.

Most concordial elements are highly ambiguous (Faaß et al.(2009)Faaß, Heid, Tajjard, and Prinsloo)

!a a mo thušá

Faaß (IMS)

A toy grammar

Northern Sotho 15 / 32

Some word classes (parts of speech)

object concords

- Object concords are the only “pronouns” in the traditional sense of the word: replacing an omitted (or topicalised) noun/NP
- Double transitive verbs: Only one of the objects may be pronominalised (usually the indirect one)
- May fuse with the verb:
 - *bona* [to] see
 - + *mo* him/her
 - *mpona* see him/her
- Always precede the verb stem directly

Faaß (IMS)

A toy grammar

Northern Sotho 17 / 32

Some word classes (parts of speech)

Pronominal and Concordial items

‘... strictly speaking, any linguistic element which agrees with a noun can acquire a pronominal function when that noun is deleted [...] these words do not stand in place of the deleted noun.’ (Louwrens(1991), p. 154)

→ Null-Subject language

Table: Emphatic pronoun: example of use

<i>dimpšáN10</i>	<i>tšémp-3rd-cl0</i>
these dogs	
<i>tšémp-3rd-cl0</i>	<i>dimpšáN10</i>
these (specific) dogs (not the others)	
<i>tšémp-3rd-cl0</i>	
these ones	

Faaß (IMS)

A toy grammar

Northern Sotho 14 / 32

Some word classes (parts of speech)

Tense morphemes

Table: Tense/Aspect morphemes

Morpheme	Indication	Comments
<i>tlo/ tla</i>	future	<i>tlo</i> and <i>tla</i> appear interchangeably
<i>a</i>	pres.	only appears in the indicative mood (if verb ends in verb stem)
<i>a</i>	past	only appears in the negated perfect indicative mood
<i>ka</i>	potential	used in the sense of “may possibly”

pres.*ba a mo thušá* They help them
future *ba tšémp-3rd-cl0* They will/shall help them
potential *ba !a mo thušá* They might help them

Faaß (IMS)

A toy grammar

Northern Sotho 16 / 32

Some word classes (parts of speech)

verbs

ba a mo thušá → They help him/her
ba thušá monna → They help (a/the) man

Table: Short excerpt Northern Sotho verb stems (cf. e.g. [Prinsloo et al.(2008)Prinsloo, Faaß, Tajjard, and Heid])

base	[to]	rule	modification	result
<i>sepela</i>	walk	-a	imp.	-a- <i>ang</i> <i>sepela!</i>
<i>le bu!a</i>	open it	-a	imp. (oc)	-e- <i>eng</i> <i>le bu!e!</i>
<i>bona</i>	see		pass.	-W- <i>bonwa</i>
<i>ngwala</i>	write	-a	*	<i>ngwadišišwego</i>
<i>re</i>	say	-a	past	-!- <i>itše</i>

* = verb + causative (!š) + perfect (le) + passive (w) + relative marker (go) ** = who are registered/emailed

Faaß (IMS)

A toy grammar

Northern Sotho 18 / 32

Some word classes (parts of speech)

Negation morpheme clusters

Table: Negated indicative mood: constellations

pres.	<ga>	2CS _{catag}	verb+object(s)
past 1	<ga se>	3CS _{catag}	verb+object(s)
past 2	<ga se>	2CS _{catag}	verb+object(s)
past 3	<ga>	3CS _{catag}	verb+object(s)
past 4	<ga>	1CS _{catag}	verb+object(s)
fut	2CS _{catag}	<ka se>	verb+object(s)

The verbal phrase: basic layout

The Verbal Phrase: VBP+VIE

Idea: Split the linguistic verb into two parts:

- The verb stem and its arguments (constellations are dependent on the verbal semantics) ;
- All other morphemes/concords (constellations define the mood)

Definitions:

- Slot system:
 - Verb and objects = Verbal Basic Phrase = slot "zero"
 - (VP in the case of the positive imperative mood)
 - Add two preceding "slots":
 - slot "zero-1" optional, contains one aspect/tense morpheme
 - slot "zero-2" optional, contains negation morpheme(s) and/or subject concord
- Slots "zero-1" and "zero-2" :
The **Verbal Inflectional Element** defines the mood of the VP

except for positive imperative

A toy grammar of Northern Sotho

VP-rules

S → { NP: (↑ SUBJ) = I;
VP
| VP: (↑ SUBJ PRED) = 'null_pro'
(↑ SUBJ PRON TYPE) = null
(↑ TNS-ASP MOOD) = c indicative
...
| VP: (↑ SUBJ PRED) = 'null_pro'
(↑ SUBJ PRON TYPE) = null
(↑ SUBJ PERS) = 2
(↑ TNS-ASP MOOD) = imperative
};
VP → VPind
VPind → VIE VBP.

"cases where subj is deleted"

"imperative"

The verbal phrase: basic layout

Basic Verbal Phrase VBP

- General rules for morpho-syntactics of the Northern Sotho verb
 - Objects always follow the verb stem
 - One of them may be pronominalised;
 - the resp. object concord precedes the verb stem
 - All other parts are optional, and they appear in front of verb and object(s)
 - Negation and subject concord appear as a group
 - Tense/aspect morphemes appear as a group

A toy grammar of Northern Sotho

Parametrisation of V-Stem ending and VIE

- Parametrisation of V-Stem ending "vend":
mainly caters for allomorphy / irregular forms
 - re vend = "a" (to say)
 - bontshitše vend = "ile" (showed)
 - ditše vend = "a" (to sit, dula (go to sit))
 - May support identification of mood/tense
- Parametrisation of all elements of the VIE "MOOD", "clause type"
 - Subject concord (noun class for subject-verb agreement) "class"
 - Tense / Aspect morphemes (tense, modal information) "TENSE"
 - positive/negated: "po(arity)"

A toy grammar of Northern Sotho







VIE-rules (excerpt)

VIE → {"General Verbal Inflectional Elements : VIE"
{ICS : (↑ TNS-ASP FORM) = short; 'short present tense form"
e : (↑ TNS-ASP TENSE) = pres
(↑ VEND) = c a
ICS
MORPH: (↑ TNS-ASP FORM) = long 'long form"
(↑ TNS-ASP TENSE) = pres;
e : (↑ VEND) = c a
...
ICS
e : (↑ VEND) = c ile
...
ICS
MORPH : (↑ TNS-ASP TENSE) = c fut
(↑ TNS-ASP POL) = pos
...
}.

Done...

Thank you!

References

-  G. Faßl, U. Heid, E. Tajjard, and D.J. Prinsloo. **Part-of-Speech tagging in Northern Sotho: disambiguating polysemous function words.** In *Proceedings of the EACL2009 Workshop on Language Technologies for African Languages – ALaT 2009*, pages 38 – 45. The 12th Conference of the European Chapter of the Association for Computational Linguistics, 30th March to 3rd April 2009.
-  U. Heid, Prinsloo D.J., G. Faßl, and E. Tajjard. **Designing a noun guesser for part of speech tagging in Northern Sotho.** *South African Journal of African Languages (SAJAL)*, 29(1):1 – 19, 2009.
-  L.J. Louwrens. **Aspects of the Northern Sotho Grammar.** Via Afrika, Pretoria, South Africa, 1991.
-  N.N. Mutaka. **An Introduction to African Linguistics.** LINCOM Handbooks in Linguistics 16. LINCOM EUROPA, München, 2000.
-  D.J. Prinsloo, G. Faßl, E. Tajjard, and U. Heid. **Designing a word guesser for part of speech tagging in Northern Sotho.** *South African Linguistics and Applied Language Studies (SALALS)*, 26(2):185 – 196, 2008.
-  E. Tajjard and S.E. Bosch. **A Comparison of Approaches to Word Class Tagging: Distinctively Versus Conjectively Written Bantu Languages.** *Nordic Journal of African Studies*, 15(4):428 – 442, 2006.