Southern African Linguistics and Applied Language Studies

ISSN: 1607-3614 (Print) 1727-9461 (Online) Journal homepage: http://www.tandfonline.com/loi/rall20

Grammar rules for the isiZulu complex verb

C. Maria Keet & Langa Khumalo

To cite this article: C. Maria Keet & Langa Khumalo (2017) Grammar rules for the isiZulu complex verb, Southern African Linguistics and Applied Language Studies, 35:2, 183-200

To link to this article: https://doi.org/10.2989/16073614.2017.1358097

Published online: 15 Dec 2017.

Southern African Linguistics and Applied Language Studies 2017, 35(2): 183–200

SOUTHERN AFRICAN LINGUISTICS
AND APPLIED LANGUAGE STUDIES
ISSN 1607-3614 EISSN 1727-9461
http://dx.doi.org/10.2989/16073614.2017.1358097

Grammar rules for the isiZulu complex verb

C. Maria Keet[1*] and Langa Khumalo[2]
[1]Department of Computer Science, University of Cape Town, Cape Town, South Africa
[2]Linguistics Programme, School of Arts, University of KwaZulu-Natal, Durban, South Africa
**Corresponding author, email: mkeet@cs.uct.ac.za.*

Abstract: The isiZulu verb is known for its morphological complexity, which is a subject of on-going linguistics research, as well as for prospects of computational use, such as controlled natural language interfaces, machine translation and spellcheckers. To this end, we seek to answer the question as to what the precise grammar rules for the isiZulu complex verb are (and, by extension, the Bantu verb morphology). To this end, we iteratively specify the grammar as a Context-Free Grammar, and evaluate it computationally. The grammar presented in this paper covers the subject and object concords, negation, present tense, aspect, mood, and the causative, applicative, stative, and the reciprocal verbal extensions, politeness, the wh-question modifiers, and aspect doubling, ensuring their correct order as they appear in verbs. The grammar conforms to specification.

Introduction

While South Africa recognises eleven official languages, significant investments into computational resources have gone mainly to English and Afrikaans. IsiZulu, which is the most widely spoken language in South Africa, still remains under-resourced. In this article we focus on the development of perfect grammar rules for the isiZulu verb and by extension Bantu verb morphology. The morphology of the verb is widely regarded as the most interesting theoretically. The next two sections provide a brief discussion on this grammatical category whose complexity presents challenges to the computation and generation of grammar rules. Traditional accounts on isiZulu grammar are based on dated sources (Doke 1927; 1935) and limited accounts on Wikipedia and there is no comprehensive synchronic grammar of isiZulu yet. A small Definite Clause Grammar and POS tagger for isiZulu has been proposed and is available online (Spiegler et al. 2010). It covers only a fraction of the complexities of the isiZulu verb, for example, it addresses only one verbal extension at a time to the exclusion of the causative, applicative, and the reciprocal extensions. Other formal approaches to isiZulu morphology focus predominately on nouns rather than verbs (Pretorius and Bosch 2009; 2012) or describe only a few sample regular expressions that cover a very small fraction of the verb (Bosch and Eiselen 2005).

We present a morphological analysis of the isiZulu verbal extension and rules for that. This is done in order to create a spell checking and part-of-speech tagging of the verb in isiZulu. We explore the means to automate the checking of the complex verb morphology. We ultimately address the following question: *What are the precise grammar rules for the isiZulu verb (and, by extension, the Bantu verb morphology)*? We thus formalise the grammar for the isiZulu verb as a Context-Free Grammar. This grammar is subsequently represented computationally so as to test its correctness with respect to specification, using a set of words and generating their derivations in the JFlap tool. The grammar covers not only the usual subject and object concords, but also negation, present tense, aspect, mood, and the verbal extensions such as the causative, applicative, stative and the reciprocal, politeness, the wh-questions modifiers, and aspect doubling.

The paper is structured as follows. The next section gives a synchronic outline of the isiZulu verb morphology and highlights comparative salient features that are characteristic of Bantu languages. It is followed by a section that discusses related works. The main contribution − the formalised account of the isiZulu verb for present tense − is presented in the section after that and evaluated in the penultimate section. We conclude in the final section.

Basics of the isiZulu verb

IsiZulu is a Bantu language that belongs to the Nguni[1] group of languages. It has close affinity to other Nguni language varieties. Bantu languages have a characteristically agglutinating morphology, which makes their structure rich and complex. The agglutinating typology is not unique to Bantu languages; other agglutinating languages with extremely complex morphology include Turkish, Hungarian and Finnish (Durrant 2013). In characterising the complexity of the verbal constructions in Bantu languages, Wald (1987: 291) states that the morphology of the verb shows '[...] the fullest extent of the agglutinative nature of the Bantu language family'. Such complex morphology presents a lot of challenges in attempts to develop computational technologies in isiZulu.

The isiZulu verbal morphology typically comprises a verb root (VR) to which extensions such as the causative, applicative, reciprocal, passive, etc. are suffixed and to which morphemes that encode negation (NEG), subject concord (SC) and object concord (OC) that cross-reference noun phrases (NPs), tense/aspect, modality, etc. are prefixed.

At the core of the verbal structure is a root morpheme, which is called the verb root (VR). The VR forms the nucleus of the verbal morphology. This core element supports a number of affixes. Each affix type occupies a specific position in the verbal morphology. The verb is characteristically terminated with a final vowel (FV) and this final vowel of the verb may encode mood, tense, polarity and potential modality. Figure 1 illustrates the complex verb in Bantu.

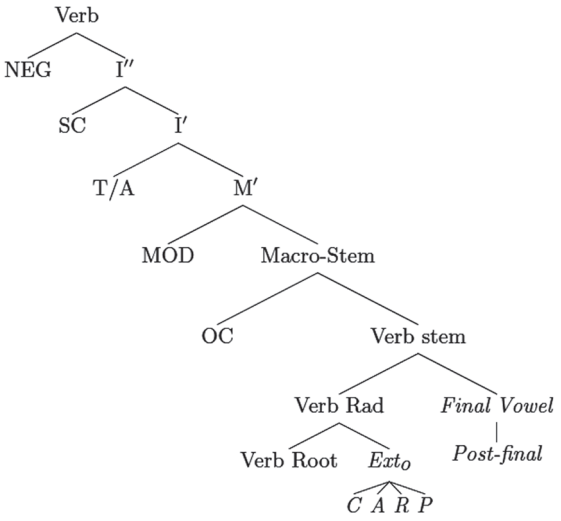

Figure 1: The structure of a complex verb in Bantu, where the elements not in italics font are considered to be the canonical verb structure. NEG: negative; SC: subject concord; T/A: tense/aspect; MOD: mood; OC: object concord; Verb Rad: verb radical; C: causative; A: applicative; R: reciprocal; P: passive

Table 1: Bantu verb slot template, adapted from Meeussen (1967) (Source: Khumalo 2007: 79) .

Slot	Pre-initial	Initial	Post-initial	Pre-radical	Radical	Pre-final	Final	Post-final
Function	TAM, NEG, clause type	SC	TAM, NEG, SC	OC	VR	TAM, valence change (CARP)	FV	Participant, NEG, clause type
Example	a	ngi	za,nga	ba	khal	is (el; an; w)	a	(ni ~ nini)[1]

~: also realised as

[1] The plural suffixes denote both general plurality and honorific plurality.

Table 2: Verbal extensions in Proto Bantu, Kiswahili, isiZulu and isiNdebele (adapted from Schadeberg 2003: 72).

Derivational Extension	Proto Bantu	KiSwahili	IsiZulu	IsiNdebele
Causative	*-i-/-ici	-ish-, -esh-	-is-	-is-
Applicative/dative	*-il-	-i-, -e-	-el-	-el-
Reciprocal/associative	*-an-	-an-	-an-	-an-
Passive	*-ʊ-/-ibʊ	-w-	-iw-, -w-	-iw-, -w-
Stative/neutro-passive/ positional	*-am-	-ik-, -ek-	-ek-	-ek-, -akal-
Reversive/separative	*-ʊl-, ʊk-	-u-, -o-	-ul-	-ul-,-ulul-,-uk-
Neuter	*-ik-			-akal-
Extensive	*-al-			-isis-
Repetitive	*-ag- ~ -ang-			
Impositive	*-ik-			
Tentive/contative	*-at-			

*: morphemes belong to an ancestor language or proto form, ʊ: precise phonetic form of the vowel.

Khumalo (2007: 79) proposes a verb slot system for the complex verbal form[2] in isiNdebele, which is applicable to isiZulu, as included in Table 1. The prefixed morphemes differ from suffixed extensions in both form and function. Formally the suffixes have a -VC- structure, as opposed to the regular CV syllable structure. Functionally, the verbal extensions affect the argument structure (Mchombo 2007: 203). Example (V1) shows the morphological organisation of the verb in isiZulu.

(V1) *Aba-shana ba-ya-zi-theng-is-el-an-a izimpahla*
2.Children 2.SC-Pres-8OC-buy$_{VR}$ -CAUS-APPL-REC-FV 8.clothes
'The children are selling the clothes to each other'

The VR *-theng-* 'buy' supports the extensions *-is-* for the causative, *-el-* for the applicative, *-an-* for the reciprocal, and the prefix clitics *ba-* for the 'subject concord', - *ya-* for the 'tense', and *-zi-* for the 'object concord'.

The verb extensions interact in complex ways with the valency of the base verb. The extensions for several languages are listed in Table 2. Semantically (with the exception of the passive extension), they alter the number of participants expressed by the verb. Grammatically, they alter the number of arguments present expressed by an NP or a pronominal element.

Concordial agreement

The term agreement in Bantu is often used alongside the term concord and the terms are sometimes used interchangeably. Agreement occurs when grammatical information appears on a verb that typically is not the source of that information. This is done through a series of agreement markers called concords that are affixed to the verb. The noun or pronoun is said to govern the agreement of all words associated with it in a syntactical relationship (Zawawi 1979: 8). Agreement is thus a cross-referencing device for subjects and objects. Table 3 shows the conjugation of the verb in isiZulu for all the noun classes and persons. As shown in the table, the verb not only takes the subject and object concords, but also the negative subject concord.

The verbal structure consists entirely of bound morphemes. These are the VR and a number of affixes such as the subject concord (SC), the object concord (OC), Tense Aspect Mood (TAM), and various other derivations (CARP), typically terminated with a final vowel (FV). The example below is a Chishona verb *ndichaenda* 'I will go':

(V2) *ndi - cha - end - a* Chishona: *ndichaenda*
1.SC - 1.TM - Root - FV
'I' 'will' go 'I will go'

Table 3: Basic verb conjugation.

Conjugation for noun classes				Conjugation for persons			
NC	SC	NEG SC	OC	Pers. Pron	SC	NEG SC	OC
1	u	aka	m	I	ngi	angi	ngi
2	ba	aba	ba	you (sg.)	u	awu	ku
1a	u	aka	m	he/she	u	awu	m
2a	ba	aba	ba	we	si	asi	si
3a	u	aka	wu	you (pl.)	ni	ani	ni
2a	ba	aba	ba	they	ba	aba	ba
3	u	awu	wu				
4	i	ayi	yi				
5	li	ali	li				
6	a	awa	wa				
7	si	asi	si				
8	zi	azi	zi				
9a	i	ayi	yi				
6	a	awa	wa				
9	i	ayi	yi				
10	zi	azi	zi				
11	lu	alu	lu				
10	zi	azi	zi				
14	bu	abu	bu				
15	ku	aku	ku				

NC: noun class; SC: subject concord; OC: object concord; NEG SC: negative subject concord.

More on clitics of the verb

In isiZulu verb morphology the term clitic typically refers to elements that attach to the verb stem. The verb stem is made up of the verb root (VR) and the final vowel (FV). Clitics can be pre-or post-verbal elements that attach to the verb stem to form complex verb forms. Clitics are therefore identifiable syntactic elements that cannot stand on their own but must be part of the host word. These identifiable elements coalesce in a morphological process to appear phonologically as part of a derived word. Example 3 demonstrates the process:

(V3)	*u*	-	*ya-*	*ba*	-	*thand* - *a*	*kakhulu*	*Uyabathanda kakhulu.*
	1aSC	- TAM	2aOC	- Root	-	FV		
	'S/he'	(Tense)	'them			'love'	'a lot'	'S/he loves them a lot.'

The prefixes *u-*, *ya-* and *ba-* coalesce with the verb stem *-thanda* to form a derived word *uyabathanda* 's/he loves them'. Clitics are thus identifiable grammatical units without the capacity to stand on their own as phonologically independent elements. Syntactically they are words as units of grammar but phonologically they lack word hood status. It is in this sense that they do not satisfy the basic criteria for being a word in Bantu languages. A word in Bantu languages must have at least two syllables. Looking at the elements in example (V3) *u-* is monosyllabic, *ya-* is monosyllabic, and *ba-* is also a single syllable.

IsiZulu verb extensions are also elements that attach to the verb form after the VR. These verb extensions must be affixed to the VR in order to form a complex verb form. They are in this sense bound elements that lack phonological independence. The coalescing of the VR and the verb extension affects the argument structure of the verb. The isiZulu verb extensions include the causative, applicative, reciprocal, passive, stative, etc. These extensions are attached to the VR and the resultant complex form is terminated by the FV *-a*. When the VR is suffixed with a verb extension it forms, a VS as shown in Figure 1. Example 4 below (from Keet and Khumalo 2017) shows the VR plus the verb extensions.

(V4)	*bon* -	*a*	*bona*	'see'	un-extended verb
	VR -	FV			
	bon - is -	*a*	*bonisa*	'make see'	extended verb
	bon - el -	*a*	*bonela*	'see for'	extended verb
	bon - an -	*a*	*bonana*	'see each other'	extended verb

The introduction of the verb extension to form a complex verb has the effect of introducing an expressible NP within a sentence. However, this does not mean that it is in itself an independent element. Just like prefixes, verb extensions cannot stand as phonological words on their own. It is important to note that clitics can be attached to an extended verb. This is conditional in that the clitic must however come after the FV. It would seem that the VR and the verb extension sequence in a complex verb is not breakable. However and more crucially, while the VS is the domain of a number of linguistic processes, its influence is not extended to the suffixed clitics. It is thus assumed that the VS has lexical integrity. This makes the VS an important subdomain in the morphological structure of the verb, and is thus the domain of lexical processes in Bantu (Mchombo 2004).

Aspects of isiZulu tense

Bantu languages typically consist of rich tense and aspect systems, characterised by various temporal distinctions (Lindfors 2003). The complexity of grammaticalised tense and aspect in isiZulu is exemplified by its five tenses. The tenses include the remote past, recent past, present, immediate future and remote future tense. The three aspectual forms are the simple, progressive, and exclusive aspect.

IsiZulu makes productive use of its grammatical aspect system. The progressive aspect in isiZulu is denoted by the affix *-sa-*. Whilst conveying an ongoing action/state/event, the morpheme also carries an inherent adverbial meaning of 'still' as shown in example (V5).

(V5)	*Ngi* -	*sa* -	*fund* - *a*	*isiZulu*
	1SG -	PROG -	VR - FV	isiZulu
	'Even now I am still studying Zulu'			

There is no direct adverb (lexical item) for the English word 'still' in isiZulu. Instead it is expressed using the adverb *namanje*, which directly translated means 'even now', as shown in (V6). The rich expression of temporal events and situations in isiZulu is further highlighted in example (V6):

(V6)	*Na -manje*	*ngi* -	*sa* -	*fund* -	*a*	*isiZulu*
	Cl-ADV	1SG -	PROG -	VR-	FV	isiZulu
	'I am still studying isiZulu'					

Similarly, the exclusive morpheme *se-* expresses an inherent adverbial aspect, meaning 'now'. This morpheme may be used with the adverb *manje* 'now', thereby expressing double aspect comprising of the grammatical aspect (*se-*, now) + grammatical aspect (*manje*, 'now'). This phenomenon has been referred to as aspect doubling, and is illustrated example (V7).

(V7)	*Se* -	*ngi* -	*ya* -	*fund* - *a*	*manje*	
	EXCL-	1SG -	CONT -	VR -	FV	ADV
	'Now, I am now studying'					

The productive nature of exclusive and progressive aspect morphemes in isiZulu has not received considerable attention. The exclusive morpheme *se-* may be used with adverbial structures conveying similar meanings in isiZulu, while this is proscribed in the English language.

This section has thus shown the complexity of the morphology of the verb. It has shown that not only does the isiZulu verb get inflected before the VR but also after the VR through a whole gamut of clitics that have an effect on the construction of a whole sentence. This is not unique to isiZulu but is characteristic of other Bantu languages, such as Chishona. It is thus this

complexity of verbal morphology that presents challenges in the development of computational technologies in isiZulu.

Related work on isiZulu verbs

The verb in Bantu has received considerable attention (cf. Hyman 1991; Mabugu 2001; Mchombo 2004, etc.). This is because it is arguably the most interesting grammatical category in linguistic theory. Many accounts in Bantu have sought to explicate the many salient morphosyntactic properties of the verb using different generative theoretical approaches. Buell (2005) is the most recent comprehensive study of the isiZulu verb. Buell discusses the isiZulu verb using a restrictive theory of syntax, which is premised on the assumption that there is a close relation between the morphology and the syntax. His account covers an array of inflectional elements such as mood, sub-mood, and polarity, subject and object agreement. Buell (2005) also briefly makes reference to the verbal suffixes such as the applicative. In a study such as his, it is impossible to be exhaustive. In this study, however, we cover the causative, applicative, reciprocal and the passive. Earlier studies on the Zulu verb are (Beuchat 1966), whose study focuses on the verb and its conjugation of various subject concords and their allomorphs, tense and mood conjugational morphemes. Beuchat (1966) does not make reference to derivational extensions.

As we seek to have a precise, formal, representation of the isiZulu verb, we also consider computational processing of the verbs, for they require a structured representation to work computationally. Regarding controlled natural languages and natural language generation, there are only two recent papers (Keet and Khumalo 2014a; 2014b), which cover verbs only to the extent of noun class-appropriate singular present tense when verbalising simple existential quantification involving object properties. Some literature on computational linguistics for isiZulu exists that is relevant to some extent, being morphological analysers. Among these works, the Ukwabelana corpus and related materials (Spiegler et al. 2010) is the most comprehensive and is the only one with online source material. Besides the corpus and limited semi-automated POS tagging, Spiegler et al. developed a basic Definite Clause Grammar (DCG),[3] of which a relevant section is shown in Figure 2. The first to note is that while it has each of the 'CARP' (`xc` etc.; bottom part of Figure 2), it has only ever one of them. This constitutes a subset of the possibilities, as multiple ones can be appended and they appear in a certain order. Also, the passive (`xp` in the CFG in Figure 2) causes changes in the concords in the verb, yet this is not

```
v --> neg, spfn, asp, opf, vr1, vsf_neg.
v --> neg, spfn, asp, vr1, vsf_neg.
...
v --> spfi, asp, opf, vr1, vsf.
v --> spfi, asp, vr1, vsf.
...
v --> spfp, vr1, vsf.
...
v --> spfs, opf, vr1, vs.
...
vr1 --> vr, xa.
vr1 --> vr, xc.
vr1 --> vr, xn.
vr1 --> vr, xp.
vr1 --> vr, xr.
vr1 --> vr.
```

Figure 2: Selection of DCG statements from the online supplementary material to Spiegler et al (2010); "..." means line(s) omitted here

catered for, nor are the politeness prefixes (*aw*-, a.o.) and tenses other than present tense, nor imperative. That is, it covers a subset. That said, it is already useful and at least it can be extended, unlike related works such as (Bosch and Eiselen 2005; Pretorius and Bosch 2009; 2012).

Bosch and Eiselen (2005) report on a basic spelling checker that is based on a set of regular expressions. They illustrate 4 examples that show a few permutations for a verb, e.g.,

```
/^(ba)(ya)?(ngi)?(.+)(el)?(a|c)(ni|phi)?$/
```

which is a subset of the conjugation (*ba*- for 3rd person plural) and CARP (-*el*-) and no details of its implementation is provided (Bosch and Eiselen (2005). A related work on morphological rules focuses on nouns (Pretorius and Bosch 2009).

The bootstrapping approach presented in Pretorius and Bosch (2012) considers the copulative (and a few other word categories) but not verbs in general. Assuming that the lexc and xfst rules as described in Pretorius and Bosch (2003) do exist, then its coverage of verb features is incomplete, notably missing mood and aspect, applicative, reciprocal, stative, politeness, and wh-ending. While their approach of figuring out which CARP extensions are permitted with a verb root is interesting (relying on the noun forms), it results in rules that are too restrictive: 'by explicitly listing the noun stems of the verb root -fund- no suffixes other than -a, -el-o, -i, -is-an-o, -is-i, -is-o, -is-wa, and -o will occur with -fund-' (emphases omitted) (Pretorius and Bosch 2003), but words such as *awufunde* '[could we/you] please study' and *usafundaphi* 'where are you [still] studying?' are valid verb forms.

Concerning verbs in other Bantu languages, several rules for Setswana (also an official language in South Africa) verbs have been implemented in xfst (Pretorius et al. 2009), but it is not clear how much of the grammar of the verb was covered. Further afield from the languages in South Africa, there are exploratory results for Ekegusii (a Bantu language spoken in Western Kenya) with several regular expressions in xfst zooming in on the difficulties of tone in relation to verbs (Elwell 2005), and there is a systematic account of the Runyakitara (a Bantu language spoken in Uganda) verb implemented in fsm2, including both the grammar and context-dependent rewriting rules that handle morpho-phonological and orthographical issues (Katushemererwe and Hanneforth 2010).

From a scientific methodological viewpoint, there is no clear 'winner' between the data-oriented approach and the knowledge and rules-based approach to obtain the grammar; or the empirical and the rational paradigms. The data-based techniques, notably machine learning (Spiegler et al. 2010; Getao and Miriti 2000), have the hurdle of finding or creating a corpus that is representative enough and at least some rules to process them, whereas the rules-based techniques face the issue of a dearth of up-to-date, structured, grammar books, having to start afresh with formalising the grammar as grammar or regular expressions. Our literature survey indicates the latter approach is used considerably more often for Bantu languages (Pretorius and Bosch 2009; 2003; Elwell 2005; Katushemererwe and Henneforth 2010; Pretorius et al. 2009). However, use/preference does not imply greater effectiveness.

Structured representation of the isiZulu verb

Methodologically, theoretically and technically there are multiple ways of specifying the grammar of a POS category; e.g., using a grammar such as a DCG, regular expressions, or their more abstract representation with an automaton (PDA for a CFG). While for the small subset of prefixes for noun classes and some simple verb forms it certainly is easier to design an NFA, transform it into a DFA and from there into a RE, there are so many options with the verbs that the automaton would become too large and wieldy. Moreover, the cross-dependencies of elements before and after the verb root indicate that a regular expression is not expressive enough and may need a CFG rather than an RG. To create the structured representation of the isiZulu verb that is computationally useful, we build it up stepwise from a linguistic pattern, to some quasi regular expressions that in turn revealed a pattern, and from there to a basic grammar, which in turn was extended with other verb features. For reason of exposing this incremental methodological approach to the design of the grammar, we

report on the component-steps of one cycle, and subsequently only the outcome of the subsequent cycles, which amount to extensions of the grammar obtained in the first round. The additions to the first cycle were − and can be − done in arbitrary order.

First iteration

From the general linguistic structure of the isiZulu verb as depicted in Figure 1, we obtain the full set of 'slots' of the verb's basic components as follows:

R0: [NEG] [SC] [T/A] [MOD] [OC] [VR] [C] [A] [R] [P] [FV]

with [VR] being the verb root at the centre. Each NEG, SC etc. has its own set of characters for each noun class; see Table 3. For the CARP, we have, as a general rule, C = *is*, A = *el*, R = *an*, and for P = *w*, though there is some phonological conditioning for A and P.

First part before the VR

Let us consider first what comes before the verb root (VR), with the subject present and active, and both in the positive (thus FV = *a*) and in the negative (FV = *i*), and assuming there is an object after the verb, so that OC can be omitted (see below for OC inclusion). Then the following patterns are permissible (italicised):

- *[SC]* [VR] [FV=a]
- *[SC] [MOD]* [VR] [FV=a]
- *[SC] [T/A] [MOD]* [VR] [FV=a]
- *[NEG] [SC]* [VR] [FV=i]
- *[NEG] [SC] [MOD]* [VR] [FV=i]
- *[NEG] [SC] [T/A] [MOD]* [VR] [FV=i]

This can be captured by the following two quasi regular expressions (where the NEG, SC, T/A, MOD, and VR variables are to be replaced by the actual strings):

R1: `[SC][T/A]`$^{0..1}$`[MOD]`$^{0..1}$`[VR]a`
R2: `[NEG][SC][T/A]`$^{0..1}$`[MOD]`$^{0..1}$`[VR]i`

Or, if the software to implement it allows for REs+rules, then:

R3: `[NEG]`$^{0..1}$`[SC][T/A]`$^{0..1}$`[MOD]`$^{0..1}$`[VR][FV]`
R4: **if** NEG **then** FV=i, **else** FV=a

The OC is used if there is no explicit object named after the verb. Then we have the following options:

- *[SC] [OC]* [VR] [FV=a]
- *[SC] [MOD] [OC]* [VR] [FV=a]
- *[SC] [T/A] [MOD] [OC]* [VR] [FV=a]
- *[NEG] [SC] [OC]* [VR] [FV=i]
- *[NEG] [SC] [MOD] [OC]* [VR] [FV=i]
- *[NEG] [SC] [T/A] [MOD] [OC]* [VR] [FV=i]

This amounts to the following two rules:

R5: `[SC][T/A]`$^{0..1}$`[MOD]`$^{0..1}$`[OC][VR]a`
R6: `[NEG][SC][T/A]`$^{0..1}$`[MOD]`$^{0..1}$`[OC][VR]i`

While we could combine R1, R2, R5 and R6, it then will have to go through a whole set of permutations to either check correct syntax or generate it. We currently expect it to be quicker to look ahead to the tag of the next phrase to determine whether an OC is needed, and then choose either the rules with OC or without; that is:

R7: **if** next word==Ø **or** next word !=noun **then** use R5 or R6,
else use R1 or R2

where the 'next word'==Ø essentially means that the verb is the last word in the sentence.

Second part after the VR
The extension is added to the verb root (VR), and comes before the FV. We show a section of the rather long list of all options:

- *[some prefix]* [VR] *[C]* [FV]
- *[some prefix]* [VR] *[C] [A]* [FV]
- *[some prefix]* [VR] *[C] [A] [R]* [FV]
- *[some prefix]* [VR] *[C] [A] [P]* [FV]
- *[some prefix]* [VR] *[C] [R]* [FV]
- *[some prefix]* [VR] *[C] [R] [P]* [FV]
- *[some prefix]* [VR] *[C] [P]* [FV]
- *[some prefix]* [VR] *[C] [A] [R] [P]* [FV]
- *[some prefix]* [VR] *[A]* [FV]
- etc.

That is, the CARP stay in that order, but any one or more of them can be used, so the following quasi regular expression can be specified:

R8: *[some prefix]*`[VR][C]`$^{0..1}$`[A]`$^{0..1}$`[R]`$^{0..1}$`[P]`$^{0..1}$`[FV]`

to be implemented by filling in the actual strings in the places of the VR, C, A, R, and P, and the *[some prefix]* following the rules as outlined above.

From quasi RE to grammar
The quasi REs show some repetition, and especially the '*[some prefix]*' makes it look clumsy. It also can be seen that there are four components: what comes before the VR, the VR, what comes after the VR, and the final vowel. This can be addressed more easily and succinctly with a generative grammar. To design that, let us first convert R1, R2, R5 and R6 into grammar notation, using the following abbreviations: **v** = verb (with its adornments), **n** = negation, **s** = subject concord, **t** = tense, **asp** = aspect, **o** = object concord, **m** = mood, **c** = causative, **a** = applicative, **r** = reciprocative, **p** = passive, **vr** = verb root, and where text in true type font are terminals, and spaces in the rules are not spaces in the word, but added for readability:

R1 in CFG notation:
v → **s vr** a | **s m vr** a | **s t m vr** a | **s asp m vr a** a
R2 in CFG notation:
v → **n s vr** i | **n s m vr** i | **n s t m vr** i | **n s asp m vr** i
R5 in CFG notation:
v → **s o vr** a | **s m o vr** a | **s t m o vr** a | **s asp m o vr** a
R6 in CFG notation:
v → **n s o vr** i | **n s m o vr** i | **n s t m o vr** i | **n s asp m o vr** i

This still will result in duplications, for these will have to be reused for CARP. To this end, we create **pre** and its negated variant **npre** and a **post** (that can be empty, ε), that will surround the verb root.

pre → **s** | **s m** | **s t m** | **s asp m** | **s o** | **s m o** | **s t m o** | **s asp m o**
npre → **ns** | **ns m** | **ns t m** | **ns asp m** | **ns o** | **ns m o** | **ns t m o** | **ns asp m o**
post → **c** | **c a** | **c a r** | **c a p** | **c r** | **c r p** | **c p** | **c a r p** | **a** | **a r** | **a r p** | **a p** | **r** | **r p** | **p** | ε

This is then put together with the verb root and final vowel:

v → **pre vr post** `a` | **npre vr post** `i`

Let us now complete the grammar so far with the terminals.

(1) List of subject concords and negative subject concords:

s → `ngi` | `u` | `si` | `ni` | `ba` | `i` | `li` | `a` | `zi` | `lu` | `bu` | `ku` | ε
ns → `angi` | `awu` | `aka` | `ali` | `asi` | `ayi` | `alu` | `abu` | `aku` | `ani` | `aba` | `awa` | `azi` | ε

(2) List of mood:

m → `a` | `e` | `ka` | `ma` | `nga` | ε

(3) List of tense (nothing for the simple present tense):

t → ε

(4) List of aspect (additional rules omitted in this first iteration):

asp → `sa` | `se` | `be` | `ile` | ε

(5) List of object concords:

o → `ngi` | `si` | `ku` | `ni` | `m` | `ba` | `wu` | `yi` | `li` | `wa` | `zi` | `lu` | `bu` | ε

(6) Causative:

c → `is`

(7) Applicative:

a → `el`

(8) Reciprocative:

r → `an`

(9) Passive (with phonological conditioning options):

p → `iw` | `w`

(10) Lexicon of verb root:

vr → `ab` | `...` | `zwib`

This completes the first iteration: the core possibilities for present tense are completed with respect to R0 mentioned at the start of the section.[4] It can be optimised, but this is left for the implementation; here, we aimed to be as explicit as feasible.

Subsequent iterations

The outcome of the first iteration does not fully cover all verb options. Further extensions and refinements can be made, which are introduced now in their final version, being politeness, stative verbs, wh-questions, and aspect doubling.

Politeness—The please and polite permissive questions have their own prefix system and a FV = `e`. This amounts to adding a new rule:

ppre → **pl s**

with the following terminals:

(11) Please prefix, permissive prefix (none), and polite proposal doing something together, indicated with **pl**:

pl → `aw` | `awu` | `mawu` | `ma` | ε

and extending the grammar rule for **v** with the extra option:

v → **pre vr post** `a` | **npre vr post** `i` | **ppre vr** `e`

Stative verbs—The stative refers to the state of being of something; e.g. *vula* (‘open’) with its stative variant *vuleka* ‘be opened’, and *mbula* (‘reveal’) results in *mbuleka* ‘be revealed’. This insertion of the *-ek-* between the VR and the FV is also referred to as the neuter extension. As it is conceptually different from the verb extension (i.e., CARP), we create a separate **st** and update the rule for **v** with it:

v → **pre vr post** `a` | **npre vr post** `i` | **ppre vr** `e` | **vr st** `a`

with the following single terminal:

(12) Stative verb, indicated with **st**:

st → `ek`

Because there is only one terminal for **st**, the ‘**vr st** `a`’-part of **v** can also be written as ‘**vr** `eka`’.

Wh-questions—The optional wh-questions fall in the post-final slot (see Table 1) and are added at the end of the verb, being *-ni* ‘what’/‘who’/‘why’/‘how’, *-nini* ‘when’, and *-phi* ‘where’. We create a separate **wh** variable for them and update the rule for **v** with it:

v → **pre vr post** `a` **wh** | **npre vr post** `i` **wh** | **ppre vr** `e` | **vr st** `a`

with the following terminals for the new variable:

(13) Wh-questions, indicated with **wh**:

wh → `ni` | `nini` | `phi` | ε

Aspect doubling—What is normally referred to as aspect doubling is a construction of aspect with continuous tense, i.e., the 'second aspect' is not an aspect in the strict sense of the meaning of aspect. Decomposed, we have EXCL-SC-CONT-(OC-)VR-(post)-a, where the exclusive can only be *se*- and continuous tense only *-ya-*. Because it is a regular exception, we add another 'or' to **v** rather than complicate **pre**:

v → **pre vr post** a **wh** | **npre vr post** i **wh** | **ppre vr** e | **vr st** a | **excl s cont o vr post** a

with the following terminals for the new variables:

(14) 'Double aspect', indicated with **excl** for exclusive (with **excl** ⊂ **asp**)

excl → se

(15) With **cont** ⊂ **t** and **cont** for continuous tense:

cont → ya

The previous extension implies that **t** (item 3, above) also has to be updated:

t → ya | ε

Finally, there is only one terminal for each, so the '**excl s cont o vr post a**'-part of **v** can also be written as '**se s** ya **o vr post** a'.

Other rules

While the CFG may seem like a relatively free combination of anything, there are several constraints that are not covered by these grammar rules, as they would obfuscate the general patterns, not all of them are linguistically accounted for, and they are easier to implement as separate rules. Notably, there is an interaction between the two sides of the VR, which is ruled by the semantics of the CARP extension. For instance, for a construction to be causative and applicative, there have to be at least two things involved. The first participant is already catered for with the SC, the second is catered for with the OC. Typically, the causative and applicative will have an OC but the reciprocal, passive and the stative would not. Further, for the passive, the object moves to the subject position, and so also with SC and OC.

The following set of rules (in pseudocode-style notation) is a first attempt at specifying them, and more will be added in due course:

(16) only **if p** ∈ **post**, **then**: **if pre then s** → ε, **else ns** ε
(17) **if c** ∈ **post**, **then s**,**o** ∈ **pre** or **npre**
(18) **if a** ∈ **post**, **then s**,**o** ∈ **pre** or **npre**
(19) **if p** ∈ **post**, **then o** ∈ (**pre** or **npre**) and **s** ∉ (**pre** or **npre**)
(20) **if vr** ∈ **Intransitive**, **then r** ∉ post and **o** ∉ (**pre** or **npre**)
(21) **if vr** ∈ **Monosyllabic**, **then post** → ε

The second set of rules has to do with phonological and morphological conditioning, such as:

(22) **if s==**u **then pl=**aw, **else pl=**aw or mawu or ma

which we consider orthogonal to coverage of the different elements of the verb, and is therefore left for further work.

Extensions and other considerations

While the 'other rules' indicate intricate interactions between the various elements of a verb that might be addressed either with extra-CFG rules or a blow-up of CFG rules (by splitting the current ones in various ways) once fully known, one of a different kind is the treatment of the elements themselves. For instance, TAM can be at the start of the verb and at the end, but when at the start or end, only a *subset* of TAM is permissible, i.e., FV $\subset$? $\subset$ TAM, where that subset '?' exists, but is not well-documented yet as to why, what, and how.

The formal approach taken in the previous section lends itself well to a rigorous assessment of measurable distance or difference with verbs in other Bantu languages, as well as bootstrapping a CFG for some of the closely related but even lesser-resourced languages, such as isiNdebele and isiXhosa. That said, we are well aware it will not be the same. Take, for instance, the Chishona − a neighbouring language − example in V8.

(V8) *mukomana a-ri-ku-gur-ir-a-**zve** chisikana*
1.boy 1.SC-T-M-break-APPL-FV-**too** 7.girl
'The boy is breaking (something) for the girl **too'**

The order of the extensions and clitics in the example above is worth noting. The clitic *-zve* comes after the FV *-a*. While this phenomenon is not found in the isiZulu verb complex, it shows that there are unique features of the Bantu verb that further complicate the grammar, which will need to be accounted for.

Finally, incorporation of phonological and morphological conditioning, while being an orthogonal aspect to the structure of the components of the verb and order thereof, may, for practical reasons, have an effect on the rules itself. For instance, possibly splitting the terminals of the **vr** into one set for vowel-commencing roots and one for consonant-commencing roots, and then for ease of processing, some of the terminals of the **pre** and **npre** could be split into two as well.

Evaluation of the grammar

We first illustrate manually the functioning of the CFG with three use cases, and subsequently test it systematically with a computational version of it.

Use cases

Three examples are selected that also give a hint toward the CFG's usability for a range of applications: generation of a word from the grammar (useful for machine translation and controlled natural languages), the checking of a correct word whether it is in the language of the grammar (spellchecking), and one misspelled word that gets rejected (correcting).

Let us step through the grammar in the least amount of steps (least amount of components) to 'generate' a word in the language, where each numbered subscript of the arrow is added for explanatory purpose afterward: **v** $\Rightarrow_1$ **pre vr post** a **wh** $\Rightarrow_2$ **s vr post** a **wh** $\Rightarrow_3$ **ngi vr post** a **wh** $\Rightarrow_4$ ngi vel **post** a **wh** $\Rightarrow_5$ ngi vel a **wh** $\Rightarrow_6$ ngi vel a; 1) substitute **v** for the first option; 2) substitute **pre** for the first option (**s**); 3) substitute **s** with the first terminal (ngi); 4) take one of the **vrs** (vel); 5) process **post**, choosing empty (ε); 5) process **wh**, choosing empty (ε). Thus, the word generated is: ngivela 'I come from'. Stepping through the grammar, using more slots at the end, we can check that niboniselana is in the language: **v** $\Rightarrow$ **pre vr post** a **wh** $\Rightarrow$ ni **vr post** a **wh** $\Rightarrow$ ni bon **post** a **wh** $\Rightarrow$ ni bon **c a r** a **wh** $\Rightarrow$ ni bon is **a r** a **wh** $\Rightarrow$ ni bon **is el r** a **wh** $\Rightarrow$ ni bon is el an a **wh** $\Rightarrow$ ni bon is el an a.

The grammar thus also can be used to recognise misspelled words. For instance, a user types *usafundapi, then it rejects it because of the -pi end: **v** $\Rightarrow$ **pre vr post** a $\Rightarrow$ **s asp vr post** a **wh** $\Rightarrow$ u **asp vr post** a **wh** $\Rightarrow$ u sa **vr post** a **wh** $\Rightarrow$ u sa fund **post** a **wh** $\Rightarrow$ u sa fund a **wh** $\Rightarrow$ ×. The trace/tree cannot be completed because pi $\notin$ **wh** (phi and ni are in **wh**), thus *usafundapi is misspelled with respect to the grammar rules as introduced in the previous sections. Proposing a correction can be done by suggesting to complete usafunda- with any of the **wh** terminals, or,

when using the minimum edit distance as an extra service in the spellchecker, it would suggest `usafundaphi` 'where are you still studying?' and `usafundani` 'what/why are you still studying?' as the two options to choose from to correct the misspelled word.

Computational evaluation of the grammar

There are many tools that are candidates to implement the grammar to the point of testing whether the rules are the right ones; that is, the scope is *validation* ('are we building the right grammar?') and *verification* ('are we building the grammar right?'), not end-user tool building.

Implementation considerations

Most computational linguistics papers for Bantu languages use one of the tools for building a morphological analyser. Xfst and lexc has been used to encode a subset of the rules for verbs in isiZulu, Setswana, and Ekegusii (Pretorius and Bosch 2003, Pretorius et al. 2009, Elwell 2005), whereas Fsm2 could not be found online anymore. However, they are problematic theoretically. Xfst, and similar tools such as SFST[5] and OpenFST,[6] are transducers, and therewith limited to regular grammars (The surface syntax gives the impression of accepting a CFG, but that is syntactic sugar and is transformed behind-the-scenes into a (very) large FSA). While most of the rules in the previous section look indeed regular, for being in a fixed order at least, when $\mathbf{p} \in \mathbf{post}$, then the $\mathbf{o} \in \mathbf{pre}$ takes the position of the $\mathbf{s}$. This already indicates that the grammar for the verb on its own is beyond a regular grammar, hence, beyond a FSM, so a transducer is insufficiently expressive. This is unsurprising, as natural languages tend all to be context-free (Pullum 1982). Another option is to take a programmatic approach. Python programming language is popular, used by (Spiegler et al. 2010), and the NLTK (Bird et al. 2009) has a CFG grammar module. However, the latter requires the word already to be segmented, but this is precisely what needs to happen automatically, and building a regular expression grammar faces the same issue as mentioned above. Spiegler et al.'s DCG for the Ukwabelana tagging is in Prolog, but at this validation and verification stage a full-fledged tool is not needed. Therefore, we used the JFlap tool,[7] which can check string membership and generate words in the language.

Testing in JFlap

Transferring the written grammar into JFlap (v8 beta) ironed out two glitches in variable abbreviations (corrected version is included in the previous section), and some of the variable names are different, because the tool allows only single-character variable names. The JFlap file, conversion annotations, and the screenshots of the outputs are available online at http://www.meteck.org/

Table 4: Strings selected for testing

String	Reason	A/R	Correct
`ngivela`	Simple present tense	A	+
`angiveli`	Simple negation	A	+
`angivela`	**pre** with **s** and **o**	A	+
`asingabaveli`	Testing **npre**	A	+
`niboniselana`	Testing **post**	A	+
`vuleka`	Stative	A	+
`usafundaphi`	Wh-extension	A	+
`sengiyafunda`	Aspect doubling	A	+
`awusidle`	Politeness	A	+
`*ngiveli`	Mixing positive with negative FV	A	−
`*usafundapi`	Typo; wrong wh-extension	R	+
`*nibonelisana`	Wrong CARP order	R	+
`*sangiyafunda`	Wrong aspect in aspect doubling	R	+
`*kabevela`	Wrong order in **pre**, no **sc**	R	+

A/R: accepted/rejected.

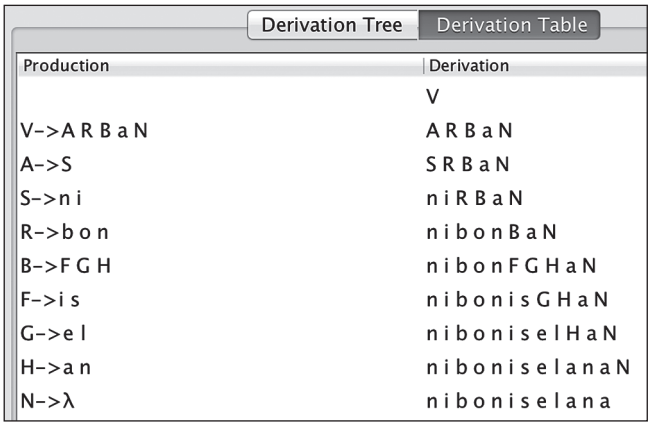

Figure 3: Screenshot of the derivation table as computed by JFlap (brute force parser), with `niboniselana`

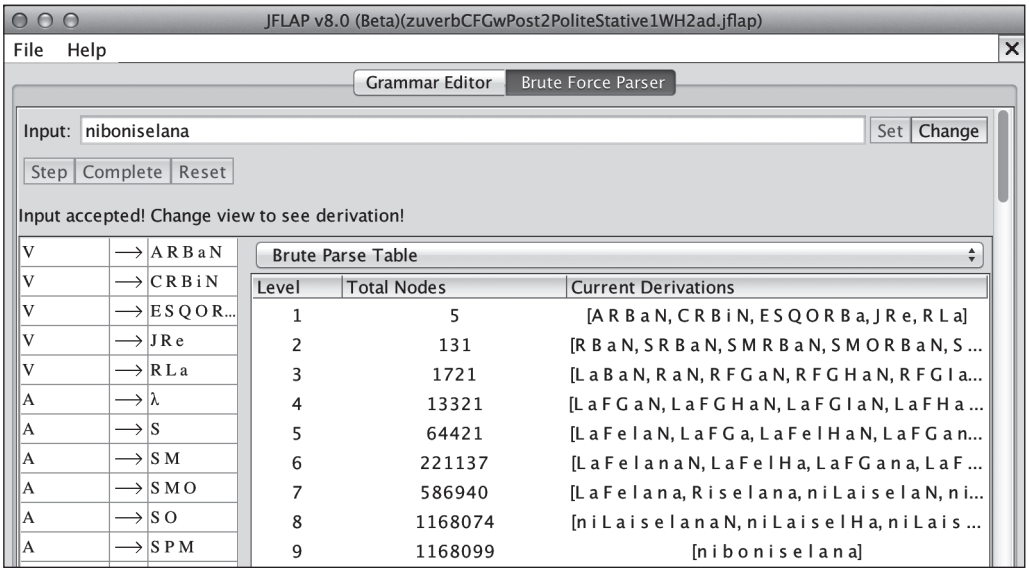

Figure 4: Screenshot of JFlap's output with the brute force parser for the final CFG, on `niboniselana`

files/geni/zuluVerbImpl.zip. We selected a set of verbs that cover the principal permutations of the rules, and some verbs that ought to be rejected, as indicated in Table 4. The strings in the first set were all accepted; a screenshot of the derivation table of `niboniselana` is shown in Figure 3 and the screenshots for the others are in the online material. Thus, the CFG recognises what it should recognise, and hence indicates correctness of the grammar specified. Of the terms one would have liked to have it reject, only `ngiveli` was (incorrectly) accepted, which is due to the εs that are in the grammar because of the absence of the extra rules in the JFlap CFG (recall the section, 'Other rules'), so **npre** is decomposed as **ns o**, with **ns** ⇒ and **o** ⇒ `ngi`. Thus, the grammar accepts/generates more strings than there are in the isiZulu language (i.e., there is some overgeneration). This can be seen also with the 'generate strings' feature in JFlap. Setting the number of strings to

a subset (100) to check, showed that this is due to not only not including the additional constraints but also not catering yet for phonological conditioning; e.g., the aforementioned *ngiveli and *aabaphi. For the latter, a = SC nc:6, a = Mood, ba = VR 'distribute', and phi = wh-question 'where', but it requires a consonant between the two as. Addressing this issue is orthogonal to the grammar, and therefore left for future work.

Further, we checked our grammar against the Zulu finite state morphology demo of the Academy of African Languages and Science.[8] As one may expect, it accepts niboniselana, but it also accepts *nibonelisana, which has the wrong CARP order, and accepts *kabevela (MOD-ASP-VR-FV), which has MOD and ASP in the wrong order (and lacks SC/OC) and is therefore rejected by our CFG. That is, that FSM does not handle any order of components of the verb.

While using a CFG computationally is an error-proof method for 'finding' a derivation, the CFG up to the wh-extension used up 773 240 nodes in the brute force parser to accept niboniselana, due to exploring all potential paths to completion. This was with five verb roots for testing the grammar; adding another five verb roots generated 1 086 109 nodes in order to accept niboniselana. With the aspect doubling extension, this increased further to 1 168 099 nodes; see Figure 4. Computationally, this is clearly not sustainable with a brute force parser, and a practical implementation that uses the CYK algorithm instead will be needed. Nevertheless, the brute-force parser is useful for evaluating the grammar.

Conclusions

We presented the precise specification of the isiZulu verb present tense as a Context-Free Grammar. It covers not only the usual subject and object concords, but also negation, present tense, aspect, mood, and the verbal extensions such as the causative, applicative, stative and the reciprocal, politeness, the wh-questions modifiers, and aspect doubling, all in their correct order as they appear in verbs. In addition to a paper-based specification, it was represented computationally as a CFG in the JFlap tool and tested on correctness of specification, using a set of words and generating their derivations in the JFlap tool. The grammar conforms to specification, though still exhibits some overgeneration. This is due to the absence of additional rules in the implementation and the orthogonal issue of phonological conditioning, which are aspects of future work.

Acknowledgements

This work is based on the research supported in part by the National Research Foundation of South Africa (CMK: Grant Number 93397).

Notes

[1] A term used by Guthrie (1971) to classify isiZulu, isiXhosa, isiNdebele and siSwati in Group S, Zone 40.

[2] The following abbreviations are used: A=aspect; ADV=adverb; APPL=applicative; CONT=continuous tense; EXCL=exclusive aspect; Ext=extension; FV=final vowel; M=mood; NEG=negative tense; OC=object concord; PROG=progressive tense; Rad=radical; SG=singular; SC=subject concord; T=tense; VR=verb root; VS=verb stem

[3] Available from: http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/resources.jsp; last accessed on 19 August 2015.

[4] Except that it does not take into account the swapping with OC and SC in case of P

[5] http://www.cis.uni-muenchen.de/~schmid/tools/SFST/

[6] http://www.openfst.org/twiki/bin/view/FST/WebHome

[7] http://www.cs.duke.edu/csed/jflap/

[8] http://gama.unisa.ac.za/demo/demo/zulmorph; tested with the version online d.d. 17-12-2015.

References

Beuchat PD. 1966. The verb in Zulu. Johannesburg: Witwatersrand University Press (1966), reprinted from *African Studies* 22: 137−169, 1963; 23: 35−49, 67−87, 1964; 25: 61−71, 1966.

Bird S, Klein E, Loper E. 2009. *Natural language processing with Python*. O'Reilly Media Inc.

Bosch SE, Eiselen R. 2005. *The effectiveness of morphological rules for an isiZulu spelling checker. South African Journal of African Languages* 25: 25−36.

Buell LC. 2005. Issues in Zulu verbal morphosyntax. PhD thesis, University of California, Los Angeles.

Doke C. 1927. *Text book of Zulu grammar*. Johannesburg: Witwatersrand University Press.

Doke C. 1935. *Bantu linguistic terminology*. London: Longman, Green and Co.

Durrant P. 2013. Formulaicity in an agglutinating language: the case of Turkish. *Corpus Linguistics and Linguistic Theory* 9: 1−38.

Elwell R. 2005. Finite-state methods for Bantu verb morphology. In: Gaylord N, Hilderbrand S, Lyu H, Palmer A, Ponvert E (eds), *Texas Linguistics Society 10: Computational Linguistics for Less-Studied Languages*. CSLI Publications.

Getao K, Miriti E. 2000. Special Topics in Computing and ICT Research: Computational modelling in Bantu language. Advances in Systems Modelling and ICT Applications. pp. 128–138.

Guthrie M. 1971. Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages. No. v. 1–2, Gregg.

Hyman L. 1991. Conceptual issues in the comparative study of the Bantu verb stem. In: Mufwene SS, Moshi L (eds), *Topics in African linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Co. pp 3−34.

Katushemererwe F, Hanneforth T. 2010. Finite state methods in morphological analysis of Runyakitara verbs. *Nordic Journal of African Studies* 19: 1−22.

Keet CM, Khumalo L. 2014. Basics for a grammar engine to verbalize logical theories in isiZulu. In: Bikakis A, et al. (eds), *Proceedings of the 8th International Web Rule Symposium (RuleML'14). LNCS, vol. 8620, 18–20 August, Prague, Czech Republic*. Springer: Prague. pp 216−225.

Keet CM, Khumalo L. 2014. Toward verbalizing logical theories in isiZulu. In: Davis B, Kuhn T, Kaljurand K (eds), *Proceedings of the 4th Workshop on Controlled Natural Language (CNL'14). LNAI, vol. 8625, 20−22 August, Galway, Ireland*. LNAI: Springer. pp 78{89.

Keet CM, Khumalo, L. 2017. Toward a knowledge-to-text controlled natural language of isiZulu. *Language Resources and Evaluation* 51: 131−157.

Khumalo L. 2007. An analysis of the Ndebele passive construction. PhD thesis, University of Oslo, Norway.

Lindfors AL. 2003. Tense and aspect in Swahili. Technical report, Institutionen for Lingvistik, Uppsalla Universitet (2003), http://www2.lingfil.uu.se/ling/semfiler/Swa_TAM.pdf

Mabugu P. 2001. Polysemy and the applicative verb construction in Chishona. PhD dissertation, University of Edinburgh, Edinburgh.

Mchombo S. 2004. *The syntax of Chichewa*. Cambridge, UK: Cambridge University Press.

Mchombo S. 2007. Argument binding and morphology in Chichewa. In: Hoyt F, Seifert N, Teoderescu A, White J (eds), *Texas Linguistics Society 9: Morphosyntax of Underrepresented Languages*. Texas Linguistics Society. pp 203−221.

Meeussen AE. 1967. Bantu grammatical reconstructions. *Africana Linguistica* 3: 79−121.

Pretorius R, Berg A, Pretorius L, Viljoen B. 2009. Setswana tokenisation and computational verb morphology: Facing the challenge of a disjunctive orthography. In: *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages (AfLaT'09) 31 March, Athens, Greece*. Athens: EACL. pp 66−73.

Pretorius L, Bosch ES. 2003. Finite-state computational morphology: An analyzer prototype for Zulu. *Machine Translation* 18: 195−216.

Pretorius L, Bosch SE. 2009. Finite state morphology of the Nguni language cluster: modelling and implementation issues. In: Yli-Jyra A, Kornai A, Sakarovitch J, Watson B (eds), *Finite-state methods and natural language processing 8th International Workshop (FSMNLP'09). LNCS, vol. 6062*. Springer. pp 123−130.

Pretorius L, Bosch S. 2012. Semi-automated extraction of morphological grammars for Nguni with special reference to Southern Ndebele. In: Workshop on Language Technology for Normalisation of Less-Resourced Languages SALTMIL8/AfLaT2012. pp 73−78.

Pullum GK, Gazdar G. 1982. Natural languages and context-free languages. *Linguistics and Philosophy* 4: 471−504.

Schadeberg TC. 2003. Derivation. In: Nurse D, Philippson G (eds), *The Bantu languages*. Routledge.
Spiegler S, van der Spuy A, Flach PA. 2010. Ukwabelana − an open-source morphological Zulu corpus. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*. Beijing: Association for Computational Linguistics. pp 1020−1028.
Wald B. 1987. Swahili and the Bantu languages. In: Comrie B. (ed.), *The world's major languages*. Oxford: Oxford University Press. pp 991−1014.
Zawawi SM. 1979. *Loan words and their effect on the classification of Swahili nominals*. Leiden: E. J. Brill.