Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Towards a gold standard corpus for detecting valencies of Zulu verbs

Gertrud Faaß
Hildesheim University
Institute for Information Science and Natural Language Processing
and University of South Africa
Department of African Languages
gertrud.faass@uni-hildesheim.de

Sonja Bosch
University of South Africa
Department of African Languages
boschse@unisa.ac.za

Abstract

We report on a new project building a Natural Language Processing resource for Zulu by making use of resources already available. Combining tagging results with the results of morphological analysis semi-automatically, we expect to reduce the amount of manual work when generating a finely-grained gold standard corpus usable for training a tagger. From the tagged corpus, we plan to extract verb-argument pairs with the aim of compiling a verb valency lexicon for Zulu.

1 Introduction

The observation that all parts of speech in a phrase, clause or sentence interact in some way with each other is one of the most important basics of today's grammars. With regard to Head-Driven Phrase Structure Grammar (HPSG, cf. Pollard and Sag, 1994), Sag et al. (2003:536) state that 'all the parts of a phrase depend directly on its head word'. Looking at the constraint-based Lexical Functional Grammar (LFG, cf. Bresnan, 2001, cited by Butt et al., 1999:43), we note that the 'determination of a verb's subcategorization frame […] constitutes a central part of any grammar development effort'. In accordance with the perspective of (lexical) semantics, a verb that for instance denotes a change of state requires one or more 'participants', the *arguments* of the verb that will represent the actor, and/or the thing or person experiencing the change-of-state described by this verb.

As the type and number of arguments of a verb depend on its use, the availability of large text collections (i.e. corpora) is essential when attempting to generate a lexicon of verb valencies. With regard to an under-resourced language such as Zulu, a relatively large corpus has only been made available recently, hence we can start working towards generating a verb valency lexicon for Zulu, combining known methods and available tools with the aim of a - at least mostly automated – processing chain.

2 Zulu challenges

Zulu is a member of the Bantu language family and is one of the eleven official languages of South Africa. The morphological structure of Zulu is depicted by a nominal classification system according to which nouns have prefixal morphemes (so-called noun prefixes). For ease of reference these noun prefixes have been assigned numbers by scholars working in the field of Bantu linguistics. The various noun prefixes link the noun to other words in the sentence, e.g. verbs, adjectives, possessives, pronouns, and so forth by means of concordial morphemes or concords. Zulu is predominantly agglutinating in nature, with the majority of words consisting of more than one morpheme which is as such, a challenge for NLP processing. Like other languages with a highly informative morphology, Zulu also allows for a relatively free word order (cf. Gowlett, 2003:636).
While examining any Zulu grammar book, readers are often surprised by the hundreds of forms a verb, for instance, can appear in. Ten different tempi can be distinguished. Polarity and modality are encoded in the verb, too. Zulu is not content

with a first, a second and a third person in singular and plural: the third person is split into sixteen *noun classes* (of which two have sub-classes, and about half express the singular while the others stand for plural and abstract forms). In order to express subject-verb congruence a *subject concord* exists for each noun class, as shown in Table 1.

word form	***analysis***	***Translation***
ngihamba	$ngi_{1ps\text{-}sg}$-hamb-a	I walk
uhamba	$u_{2ps\text{-}sg}$-hamb-a	you walk
uhamba	$u_{cl1\text{-}sg}$-hamb-a	he/she/it walks
...		
lihamba	$li_{cl5\text{-}sg}$-hamb-a	he/she/it walks
ahamba	$a_{cl6\text{-}pl}$-hamb-a	they walk
...		

Table 1. *hamba* ("walk"): Partial inflection paradigm of the present tense indicative

Object concords may also appear as part of the orthographic verb and they may either co-occur or substitute an overt object in the sentence. As a demonstration of the latter phenomenon, the orthographic verb *bayakupheka* ("they are cooking it (at the moment)") that actually expresses an entire clause, is explained from a morphological perspective in Table 2.

morph	*ba-*	*-ya-*	*-ku-*	*-phek-*	*-a*
categ.	subj. concord cl. 2	pres. ind. long form	obj. concord cl. 15	verb root	verb ending
Engl.	they	now	it	cook	
transl.	they are cooking it (at the moment).				

Table 2. Analysis of *bayakupheka* ("they are cooking it (at the moment)")

Lastly, one may find suffixes in verbs that modify their valency. These suffixes have meanings similar to prepositions in languages like English or German and, just like these, they require arguments. Adding the applicative suffix *–el* for example to a verb changes its valency: it now needs an additional argument describing a beneficent. This issue is demonstrated in Table 3 for the verb form *ngipheka* ("I cook") becoming *ngiphekela* ("I cook for"). Such a derivation often changes the meaning of the verb as well (cf. Bosch and Pretorius, 2017).

word form	***analysis***	***transl.***
ngipheka	$ngi_{1ps\text{-}sg}$-phek-a	I cook
ngiphekela	$ngi_{1ps\text{-}sg}$-phek-el_{appl}-a	I cook for

Table 3. Application of the applicative suffix *-el*

3 Aims and Resources

Our long-term aim is to compile a corpus-based machine-readable valency lexicon for Zulu verbs which will be freely available for research purposes. By generating this lexicon, we expect to be able to explore the syntax of the Zulu language in use on a bigger scale than previously possible.

However, there is still a long way to go: thus far, Zulu text taggers (Spiegler et al., 2010; Koleva, 2011; De Pauw, 2012; Eiselen and Puttkammer, 2014) are all using a rather coarse tagset not applicable for our purposes. Second, except the tagger by Eiselen and Puttkammer (2014) none of these taggers seems to be available for local use[1].

In summary, the following list shows our primary short term aims:

1. developing a more informative tagset,
2. generating a gold standard corpus fully annotated with the tagset,
3. training and evaluating taggers and tagset, and
4. developing a chunker for extracting verbs and their arguments.

This paper is concerned with the first two steps, and it is describing the corpus and the tagging processes that have been done so far.

In the last decade, a number of NLP resources for Zulu were compiled. Most of them are available at the *South African Centre for Digital Language Resources* (SADiLaR)[2], inter alia a Zulu Tagger (Eiselen and Puttkammer, 2014). This tagger is listed as the NCHLT Tagger.

Another important resource for our project is the 3-million token Zulu corpus compiled in the *Wortschatz* project in Leipzig[3]. This corpus has been extended recently to 15.4 million tokens

[1] De Pauw's (2012) tagger can be applied online via (https://www.aflat.org/zulutag).
[2] https://www.sadilar.org/
[3] Leipzig Corpora Collection (2016)*:* zul_mixed_2016 based on texts of the year 2016. Leipzig Corpora Collection.

which will also be made available for free download. Lastly, we make use of the ZulMorph morphological analyser available as a Finite state morphology demo and reported on in detail in several publications, e.g. Bosch and Pretorius (2011). An attempt was also made to get the other taggers described above (Spiegler et al., 2010; Koleva, 2011; De Pauw, 2012) for local use, although their tagsets are not very useful. However, our requests to the authors of the respective papers were not successful.

4 Application of resources

4.1 Corpus

The currently available Zulu corpus of the Leipzig Wortschatz Collection contains more than 3 million tokens with marked sentence borders. We selected 149,196 sentences (2,337,566 tokens) in total for our local processing after deleting noise. To build our gold standard corpus, we randomly selected 1,500 sentences (about 17k tokens) from this resource.

4.2 Tagset and Tagger

The Zulu tagset used by the NCHLT Tagger (Eiselen and Puttkammer, 2014), includes nouns (N*n*), adjectives (A*n*), and verbs (V) of which the former two have noun classes (*n*) assigned, e.g. N01 or A07.

Because of the agglutinative orthography of the Zulu language, a number of syntactic constructions like copulatives (COP) and possessives (POS*n*) have their own tags assigned. This issue was criticized by e.g. Hendrikse and Mfusi already in 2008, calling for a tagset that marks such constructions as clauses, a suggestion that we will implement.

As to pronouns, there are tags for personal pronouns (PRON*n*), demonstratives (DEM*n*), and quantitatives (QUANT*n*). Unfortunately, the tagger also assigns tags like "PRON", "QUANT" or "REL" ("relative") without naming a noun class and we even find an undescribed tag "P" (we assume that this stands for "any kind of pronoun").

There are tags labelling ideophones (IDEO), though these either function as adverbs or as verbs in a sentence. We also find tags for adverbs (ADV), numeratives (NUM), conjunctions (CONJ), and interjections (INT).

The NCHLT Tagger moreover makes use of the tag "M" for which we do not find any description in the NCHLT Project[4]. The tag labels a variety of items, like (copulative) verbs but also proper nouns and abbreviations.

When developing their tagger, Spiegler et al. (2010) collapsed the two noun classes 8 and 10 into one as their forms are identical. Before tagging, they also deleted punctuation in the text and changed all characters to lower case thus their corpus is not in its original form any longer. Other developers (Koleva, 2011; De Pauw, 2012) based their works on the tagset of Spiegler et al. (2010), but they did not differentiate between noun classes at all (therefore gaining a high precision).

As described above, the NCHLT Tagger, Eiselen and Puttkammer (2014) make use of a tagset that distinguishes noun classes, however all verbs are labelled with the tag "V", which means that a subject-verb congruence cannot be detected.

In a first go, we utilize the NCHLT Tagger as it is the only tagger than can be applied to our corpus (and that seems to be available), however we need to extend the tag "V" with subject and object class information whenever this information is available and we must train a new tagger, as the NCHLT Tagger comes without the possibility to adapt it to other tagsets. We must also be aware of the fact that this tagger has not been evaluated and that it applies tags "P" and "M" which we both define as "miscellaneous" categories not usable for further processing of tagged text.

4.3 ZulMorph

The ZulMorph morphological analyser (Bosch and Pretorius, 2006) is unfortunately not available for offline use, but the developers process lists of words on request. Currently, there are about 36,000 verb roots described (Pretorius and Bosch, 2017) in ZulMorph. When applying it to our 8,625 types (see 5.2), 1,895 types were not analysed**.**

4.4 Combining information provided by the tools

The need for a finer-grained tagset and a procedure allowing us to generate a gold standard leads us to an idea described in Jung's dissertation (ne Eckart, 2018). Jung suggests the combination of information provided by different tools in order to achieve a better result.

Dataset. https://corpora.uni-leipzig.de/de?corpusId=zul_mixed_2016

[4] https://repo.sadilar.org/handle/20.500.12185/351

We hence plan to apply a "voting" procedure: in case the NCHLT Tagger agrees with the ZulMorph analysis for closed class items (like CONJ), we will not check the results again, if the NCHLT Tagger votes for V while ZulMorph offers V and non-V analyses, we will choose the V-analyses. If there are several, a semi-automatic selection of the correct analysis will take place.

5 Intermediate Results

5.1 Tagset

Our preliminary tagset is built on four levels, of which the first two are shown in Table 4. The first level describes the coarse category (to allow for future coarse tagging), the second level describes the part-of-speech in more detail. For verbs, we distinguish regular finite forms and forms with suffixes modifying their valency (NEUT(er), APPL(icative), RECIP(rocal), CAUS(ative) and PASS(ive)).
For nominal items, a third level specifying the noun class will be utilized. For verbs, this third level contains the noun class of the subject noun, the fourth level then describes (if available), the noun class of its object in the cases where an object concord appears. This fourth level is filled with the letters "RF" in case the verb contains a reflexive prefix. The tagset does not distinguish positive from negative polarity for this factor does not change its valency.

1st level	*2nd level*	*Description*
V(erb)	APPL	applicative
	CAUS	causative
	COP	copulative (x is V)
	IMP	imperative
	IDEO	ideophone
	FIN	finite (inflected form)
	NEUT	neuter
	RECIP	reciprocal
	PASS	passive
	RELP	verb containing a relative clause
N(oun)	COPP	nominal copulative clause (N is N)
	INF	infinitive (noun prefix and verbal stem)
	PROP	proper noun
	POSP	noun containing a possessive clause
	REG	regular noun
	RELP	noun containing a relative clause
ADJ(ective)	COPP	adjectival copulative clause (x is A)
	REG	regular
ADV(erb)	IDEO	ideophone
	LOC	locative
	REG	true adverb
P(ronoun)	DEM	demonstrative
	PER	personal
	QUANT	quantitative
CONJ		conjunction
INTJ		interjection
INTR		interrogative
PUNCT		punctuation
CARD		anything containing numbers
FM		foreign language material

Table 4. Preliminary Zulu-tagset

5.2 Tagging

To gain tags from morphological analyses, we first extracted all 8,625 types of our corpus (note that in this number, upper- and lower-case forms were merged) and ran them through the ZulMorph tool. This resulted in 40,458 analyses, as there are types resulting in around 100 analyses in total (e.g. *abazi*, a word with several meanings[5] that resulted in 105 different analyses).

The ZulMorph analyses contain a number of items not relevant for our purposes, we hence simplify those analyses reducing the amount of information provided. To gain a better overview, analyses like (1) of the word *omfundisayo* ("who teaches him/her") are reduced to the relevant information, as in (2).

(1) omfundisayo [RC][1]mu[OC][1]
fund[VRoot]is[CausExt]a[VT]yo[RelSuf][6]

(2) omfundisayo
o[RC][1][OC][1][CausExt][RelSuf]

From there, we can identify the verbal relative phrase, of which the subject is of noun class 1 and the object is of noun class 1 (tag: V-RELP-S01-O01). Note that for the word *omfundisayo*, there are altogether 6 ZulMorph analyses which our

[5] The interested reader is referred to https://isizulu.net/ for the variety of meanings of the word *abazi*

[6] In this ZulMorph analysis, all processing information (unrelated to morphemes) was deleted.

scripts reduce to three: V-RELP-S01-O01 (subject identified as of noun class 1), V-RELP-S02ps-O01 (subject identified as 2nd Person Singular) and V-RELP-S03-O01 (subject identified as of noun class 3). For the word *ungomunye* ("you/he/she are/is the other one"), we find 45 ZulMorph analyses that are collapsed by our tool to 6 possible annotations, and for the above mentioned *abazi*, our implementation reduces the ZulMorph analyses from 105 to 13 possible tags.

However, there is still a need for a human expert to decide upon which of the found analyses is correct in the given context.

The scripts and tools developed so far select most of the analyses generated by ZulMorph fully automatically. There are currently still 3,297 types in our gold standard corpus to be identified. For these, we follow the following actions:

1. Collapsing further ZulMorph analyses to tags that can be annotated automatically;
2. identifying the possible tags of the types for which no ZulMorph analyses are available.

We currently assist the experts working on 2. with an automated detection of possible POS-tags by looking at orthography patterns of types. For example, names usually begin with a lowercase nominal prefix "u" or "i", followed by the name beginning with an uppercase letter (e.g. uSiwela). We can annotate such types as N-PROP-01a automatically in order to avoid the necessity of human intervention.

6 Conclusion and future work

In conclusion, this project on the preparation of a future gold standard corpus for detecting valencies of Zulu verbs is still in its initial stages. So far, we have developed an informative tagset and found a methodology that makes use of available resources like the NCHLT Tagger and ZulMorph to assist us in assigning possible tags for each word of this corpus. This paper serves to describe our path towards achieving our goals and to elicit constructive feedback.

Our next steps entail the selection of the correct POS-Tag as soon as all possible POS-tags have been found for all types occurring in the training corpus, i.e. the future gold standard. In most cases, this selection must be done manually by language experts. As soon as the training corpus is finalized, we will make this corpus and a full description of the preliminary tagset available via the SADiLaR repository.

After completing the gold standard corpus, we will train statistical taggers, evaluate the tagset and find the tagger best suited for the task. With this tagger, we will then annotate a new 20-million token Zulu corpus which will be provided by the University of Leipzig in pursuit of our goal of detecting verbs and their arguments on a bigger scale. We plan to also annotate only the first level of the tags in the course of the validation expecting a higher precision. We will also make the resulting annotated corpus available for other researchers who do not need a finer tagset for their purposes. The next goal is the development of a chunker for the identification of relevant verbal phrases from the corpus. After the phrase annotations have been added to the corpus, we will be able to generate the planned lexicon of verb valencies.

References

Bosch S.E. and Pretorius L. 2006. A Finite-State Approach to Linguistic Constraints in Zulu Morphological Analysis. *Studia Orientalia*, 103, pp 205-227.

Bosch S.E. and Pretorius L. 2011. Towards Zulu corpus clean-up, lexicon development and corpus annotation by means of computational morphological analysis. *South African Journal of African Languages*, 31(1), pp 138-158.

Bosch S.E. and Pretorius L. 2017. A Computational Approach to Zulu Verb Morphology within the Context of Lexical Semantics. *Lexikos* (AFRILEX-reeks/series 27: 2017), pp. 152-182.

Bresnan J. 2001. *Lexical-Functional Syntax.* Blackwell Publishing, Maiden, USA, Oxford, UK, Victoria, Australia.

Butt M., Holloway King T., Niño M.E. and Segond F. 1999. Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Conference on Applied Natural Language Processing,* Washington DC, USA.

De Pauw, G., De Schryver, G-M. and van de Loo, J. 2012. Resource-Light Bantu Part-of-Speech Tagging. *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages* (SALTMIL8/AfLaT2012). Istanbul: European Language Resources Association. pp 55-92.

Eiselen E.R. and Puttkammer M.J. 2014. Developing text resources for ten South African languages. (In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland. pp 3698-3703.

Gowlett, D. 2003. Zone S. In D. Nurse and G. Philippson (eds): *The Bantu languages*, pp. 609-38. Routledge: London & New York.

Hendrikse R. and Mfusi M. 2008. A morphosyntactic tagset for Southern Bantu within a Construction Grammar Approach, *Language Matters* (39:2), pp 181-203.

Jung (ne Eckart), K. 2018. Dissertation: *Task-based parser output combination: workflow and infrastructure*. Universität Stuttgart: Fakultät Informatik. doi:10.18419/opus-9853.

Koleva, M. 2011. M.A. Dissertation: *Towards Adaptation of NLP Tools for Closely-Related Bantu Languages: Building a Part-of-Speech Tagger for Zulu*. Saarbrücken: Universität des Saarlandes.

Pollard C. and Sag I.A. 1994. *Head Driven Phrase Structure Grammar.* Studies in Contemporary Linguistics. The University of Chicago Press, USA.

Pretorius, L. and Bosch, S. E. 2003. Finite-State Computational Morphology: An Analyzer Prototype for Zulu. *Machine Translation,* 18(3), pp. 195-216.

Sag, I.A., Wasow, T. and Bender, E.M. 2003. *Syntactic Theory.* CSLI Lecture Notes Number 152, Stanford, California, USA, 2nd Edition.

Spiegler, S., van der Spuy, A. and Flach, P. A. 2010. Ukwabelana - An open-source morphological Zulu corpus. *Proceedings of the 23rd International Conference on Computational Linguistics* (Coling 2010), 28. Beijing, China. pp. 1020-1028.