Ukwabelana - An open-source morphological Zulu corpus

Sebastian Spiegler
Intelligent Systems Group
University of Bristol
spiegler@cs.bris.ac.uk

Andrew van der Spuy
Linguistics Department
University of the Witwatersrand
andrew.vanderspuy@wits.ac.za

Peter A. Flach
Intelligent Systems Group
University of Bristol
peter.flach@bris.ac.uk

Abstract

Zulu is an indigenous language of South Africa, and one of the eleven official languages of that country. It is spoken by about 11 million speakers. Although it is similar in size to some Western languages, e.g. Swedish, it is considerably under-resourced. This paper presents a new open-source morphological corpus for Zulu named *Ukwabelana corpus*. We describe the agglutinating morphology of Zulu with its multiple prefixation and suffixation, and also introduce our labeling scheme. Further, the annotation process is described and all single resources are explained. These comprise a list of 10,000 labeled and 100,000 unlabeled word types, 3,000 part-of-speech (POS) tagged and 30,000 raw sentences as well as a morphological Zulu grammar, and a parsing algorithm which hypothesizes possible word roots and enumerates parses that conform to the Zulu grammar. We also provide a POS tagger which assigns the grammatical category to a morphologically analyzed word type. As it is hoped that the corpus and all resources will be of benefit to any person doing research on Zulu or on computer-aided analysis of languages, they will be made available in the public domain from http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/Resources/.

1 Introduction

Zulu (also known as isiZulu) is a Bantu language of South Africa, classified as S.30 in Guthrie's classification scheme (Guthrie, 1971). Since 1994, it has been recognized as one of the eleven official languages of South Africa. It has a written history of about 150 years: the first grammar was published by Grout (1859), and the first dictionary by Colenso (1905). There are about 11 million mother-tongue speakers, who constitute approximately 23% of South Africa's population, making Zulu the country's largest language.

Zulu is highly mutually intelligible with the Xhosa, Swati and Southern Ndebele languages, and with Ndebele of Zimbabwe (Lanham, 1960), to the extent that all of these can be considered dialects or varieties of a single language, Nguni. Despite its size, Zulu is considerably *under-resourced*, compared to Western languages with similar numbers of speakers, e.g. Swedish. There are only about four regular publications in Zulu, there are few published books, and the language is not used as a medium of instruction.

This of course is partly due to the short timespan of its written history, but the main reason, of course, is the apartheid history of South Africa: for most of the twentieth century resources were allocated to Afrikaans and English, the two former official languages, and relatively few resources to the indigenous Bantu languages. Since 1994, Zulu has had a much larger presence in the media, with several television programs being broadcast in Zulu every day. Yet much needs to be done in order to improve the resources available to Zulu speakers and students of Zulu.

The aim of the project reported in this paper was to establish a Zulu corpus, named the *Ukwabelana corpus*[1], consisting of morphologically labeled words (that is, word types) and part-of-speech (POS) tagged sentences. Along with the labeled corpus, unlabeled words and sentences, a morphological grammar, a semi-automatic mor-

[1] *Ukwabelana* means 'to share' in Zulu where the 'k' is pronounced voiced like a [g].

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1020–1028,
Beijing, August 2010

phological analyzer and a *POS tagger* for morphologically analyzed words will be provided.

The sources used for the corpus were limited to fictional works and the Zulu Bible. This means that there is not a wide variety of registers, and perhaps even of vocabulary items. This defect will have to be corrected in future work.

The *Ukwabelana corpus* can be used to develop and train automatic morphological analyzers, which in turn tag a large corpus of written Zulu, similar to the Brown corpus or the British National Corpus. Moreover, the list of POS tagged sentences is an essential step towards building an automatic syntactic tagger, which still does not exist for Zulu, and a tagged corpus of Zulu. Such a corpus would be beneficial to language researchers as it provides them with examples of actual usage, as opposed to elicited or invented examples, which may be artificial or unlikely to occur in real discourse. This would greatly improve the quality of Zulu dictionaries and grammars, most of which rely heavily on the work of Doke (1927) and Doke, Malcom and Sikakana (1958), with little in the way of innovation. Morphological tagging is also useful for practical computational applications like predictive text, spell-checking, grammar checking and machine translation; in the case of Zulu, where a large percentage of grammatical information is conveyed by prefixes and suffixes rather than by separate words, it is essential. For example, in English, the negative is expressed by means of a separate word *'not'*, but in Zulu the negative is constructed using a prefix-and-suffix combination on the verb, and this combination differs according to the mood of the verb (indicative, participial or subjunctive). The practical computational applications mentioned could have a very great impact on the use of Zulu as a written language, as spell-checking and grammar checking would benefit proofreaders, editors and writers. Machine translation could aid in increasing the number of texts available in Zulu, thus making it more of a literary language, and allowing it to become established as a language of education. The use of Zulu in public life could also increase. Currently, the tendency is to use English, as this is the language that reaches the widest audience. If high-quality automatic translation becomes available, this would no longer be necessary. As it is hoped that the *Ukwabelana corpus* will be of benefit to any person doing research on Zulu or on computer-aided analysis of languages, it will be made available as the first morphologically analysed corpus of Zulu in the public domain.

2 Related work

In this section, we will give an overview of linguistic research on Nguni languages, following the discussions in van der Spuy (2001), and thereafter a summary of computational approaches to the analysis of Zulu.

2.1 Linguistic research on Nguni languages

The five Nguni languages Zulu, Xhosa, South African Ndebele, Swati, and Zimbabwean Ndebele are highly mutually intelligible, and for this reason, works on any of the other Nguni languages are directly relevant to an analysis of Zulu.

There have been numerous studies of Nguni grammar, especially its morphology; in fact, the Nguni languages probably rival Swahili and Chewa for the title of most-studied Bantu language. The generative approach to morphological description (as developed by Aronoff (1976), Selkirk (1982), Lieber (1980), Lieber (1992)) has had very little influence on most of the work that has been done on Nguni morphology.

Usually, the descriptions have been atheoretical or structuralist. Doke's paradigmatic description of the morphology (Doke, 1927; Doke, 1935) has remained the basis for linguistic work in the Southern Bantu languages. Doke (1935) criticized previous writers on Bantu grammars for basing their classification, treatment and terminology on their own mother tongue or Latin. His intention was to create a grammatical structure for Bantu which did not conform to European or classical standards. Nevertheless, Doke himself could not shake off the European mindset: he treated the languages as if they had inflectional paradigms, with characteristics like subjunctive or indicative belonging to the whole word, rather than to identifiable affixes; in fact, he claimed (1950) that Bantu languages are "inflectional with [just] a tendency to agglutination", and assumed that the morphol-

ogy was linear not hierarchical. Most subsequent linguistic studies and reference grammars of the Southern Bantu languages have been directed at refining or redefining Doke's categories from a paradigmatic perspective.

Important Nguni examples are Van Eeden (1956), Van Wyk (1958), Beuchat (1966), Wilkes (1971), Nkabinde (1975), Cope (1984), Davey (1984), Louw (1984), Ziervogel et al. (1985), Gauton (1990), Gauton (1994), Khumalo (1992), Poulos and Msimang (1998), Posthumus (1987), Posthumus (1988), Posthumus (1988) and Posthumus (2000). Among the very few generative morphological descriptions of Nguni are Lanham (1971), Mbadi (1988) and Du Plessis (1993). Lanham (1971) gives a transformational analysis of Zulu adjectival and relative forms. This analysis can be viewed as diachronic rather than synchronic. Mbadi (1988) applies Lieber (1980) and Selkirk's percolation theory (Selkirk, 1982) to a few Xhosa morphological forms. Du Plessis (1993) gives a hierarchical description of the morphology of the verb, but he assumes that derivation is syntactical rather than lexical.

In short, there has been no thorough-going generative analysis of the morphology which has treated the Nguni languages as agglutinative rather than inflectional.

2.2 Computational approaches to analyzing Zulu

In the last decade, various computational approaches for Zulu have been reported. Based on the *Xerox finite-state toolbox* by Beesley and Karttunen (2003), Pretorius and Bosch (2003) developed a prototype of a computational morphological analyzer for Zulu. Using a semi-automated process, a morphological lexicon and a rule-base were built incrementally. Later work (Pretorius and Bosch, 2007) dealt with overgeneration of the Zulu finite-state tool concerning locative formation from nouns and verbal extensions to verb roots. Pretorius and Bosch (2009) also used cross-linguistic similarities and dissimilarities of Zulu to bootstrap a morphological analyser for Xhosa. Joubert et al. (2004) followed a bootstrapping approach to morphological analysis. A simple framework uses morpheme lists, morphophonological and morphosyntactic rules which are learnt by consulting an *oracle*, in their case a linguistic expert who corrects analyses. The framework then revises its grammar so that the updated morpheme lists and rules do not contradict previously found analyses. Botha and Barnard (2005) compared two approaches for gathering Zulu text corpora from the World Wide Web. They drew the conclusion that using commercial search engines for finding Zulu websites outperforms web-crawlers even with a carefully selected starting point. They saw the reason for that in the fact that most documents on the internet are in one of the world's dominant languages. Bosch and Eiselen (2005) presented a spell checker for Zulu based on morphological analysis and regular expressions. It was shown that after a certain threshold for the lexicon size performance could only be improved by incrementally extending morphological rules. Experiments were performed for basic and complex Zulu verbs and nouns, and large numbers of words still were not recognized. Spiegler et al. (2008) performed experiments where they tested four machine learning algorithms for morphological analysis with different degrees of supervision. An unsupervised algorithm analyzed a raw word list, two semi-supervised algorithms were provided with word stems and subsequently segmented prefix and suffix sequences, and the supervised algorithm used a language model of analysed words which was applied to new words. They experimentally showed that there is a certain trade-off between the usage of labeled data and performance. They also reckoned that computational analysis improves if words of different grammatical categories are analysed separately since there exist homographic morphemes across different word categories.

3 Zulu morphology

Zulu is an agglutinative language, with a complex morphology. It presents an especial problem for computational analysis, because words usually incorporate both prefixes and suffixes, and there can be several of each. This makes it hard to identify the root by mechanical means, as the root could be the first, second, third, or even a later morpheme in a word. The complexities involved are

exacerbated by the fact that a considerable number of affixes, especially prefixes, have allomorphic forms. This is largely brought about by the fact that Zulu has a prohibition against sequences of vowels, so that a prefix whose canonical form is *nga-* will have an allomorph *ng-* before roots that begin with vowels. Given a sequence *nga-*, then, it is possible that it constitutes an entire morpheme, or the beginning of a morpheme like the verb root *ngabaz-* 'to be uncertain', or a morpheme *ng-* followed by a vowel-commencing root like *and-* 'to increase'. Furthermore, many morphemes are homographs, so that the prefix *nga-* could represent either the potential mood morpheme or a form of the negative that occurs in subordinate clauses; and the sequence *ng-* could be the allomorph of either of these, or of a number of homographic morphemes *ngi-*, which represent the first person singular in various moods. Besides these phonologically conditioned allomorphs, there are also morphologically conditioned ones, for example the locative prefix *e-* has an allomorph *o-* that occurs in certain morphological circumstances. Certain morpheme sequences also exhibit syncretism, so that while most nouns take a sequence of prefixes known as the initial vowel and the noun prefix, as in *i-mi-zi* 'villages', nouns of certain classes, like class 5, syncretise these two prefixes, as in *i-gama* 'name', where the prefix *i-* represents both the initial vowel and the noun prefix.

Like all other Bantu languages, Zulu divides its nouns into a number of classes. The class is often identifiable from the noun prefix that is attached to the noun, and it governs the *agreement* of all words that modify the noun, as well as of predicates of which the noun is a subject. Object agreement may also be marked on the predicate. Two examples of this agreement are given below.

Example 1.

Leso si-tshudeni e-si-hle e-ngi-si-fundis-ile si-phas-e kahle.
that student who-AGR-good who-I-him-teach-PAST AGR-pass-PAST well.
'That good student whom I taught passed well.'

Example 2.

Lowo m-fundi o-mu-hle e-ngi-m-fundis-ile u-phas-e kahle.
that learner who-AGR-good who-I-him-teach-PAST AGR-pass-PAST well.
'That good learner whom I taught passed well.'

The differences in agreement morphology in the two sentences is brought about because the nouns *sitshudeni* and *mfundi* belong to different classes. Canonici (1996) argues that a noun should be assigned to a class by virtue of the agreement that it takes. In terms of this criterion, there are twelve noun classes in Zulu. These classes are numbered 1–7, 9, 10, 11, 14, 15. The numbering system was devised by Meinhof (1906), and reflects the historical affinities between Zulu and other Bantu languages: Zulu lacks classes 8, 12 and 13, which are found in other Bantu languages. In the labels used on the database, morphemes that command or show agreement have been labeled as *<xn>*, where *x* is a letter or sequence of letters, and *n* is a number: thus the morpheme *m-* in *mfundi* is labeled *<n1>*, as it marks the noun as belonging to noun class *1*. The morpheme *si-* in *engisifundisile* is marked *<o7>*, as it shows object agreement with a noun of class *7*.

Zulu *predicatives* may be either verbal or nonverbal – the latter are referred to in the literature as copulatives. Copulatives usually consist of a predicative prefix and a base, which may be a noun, an adjective, or a prepositional, locative or adverbial form. There may also be various tense, aspect and polarity markers. They translate the English verb 'be', plus its complement – Zulu has no direct equivalent of 'be'; the verb *-ba*, which has the closest meaning, is probably better translated as 'become'. Examples of copulative forms are *ubenguthisha* 'he was a teacher', *zimandla* 'they are strong', *basekhaya* 'they are at home'. Predicatives may occur in a variety of moods, tenses, aspects and polarities; these are usually distinguished by the affixes attached to the base form. Thus in *engasesendlini* '(s)he no longer being in the house', the initial prefix *e-* indicates third person singular, class 1, participial mood; the prefix *nga-* denotes negative; the first prefix *se-* denotes continuative aspect; the second prefix *se-* is the locative prefix; *n-* shows that the noun belongs to class 9; *dl-* is the noun root meaning 'house', an allomorph of the canonical form *-dlu*; and *-ini* is the locative suffix. Thus in typical agglutinative manner, each affix contributes a distinctive part of

the meaning of the word as a whole. This characteristic of the language was exploited in the labeling system used for the morphological corpus: labels were designed so as to indicate the grammatical function of the morpheme. A person searching for past tense negative verbs, for example, could simply search for the combination of *<past >*, *<neg>* and *<vr>*. A complete list of morphemes, allomorphs and their labels is provided along with the corpus and other resources.

According to the Dokean grammatical tradition (Doke, 1927), Zulu has a large number of parts of speech. This is because what would be separate words in other languages are often prefixes in Zulu, and also because various subtypes of determiner are given individual names. The parts of speech recognised in the corpus are: noun, verb, adjective, pronoun, adverb, conjunction, prepositional, possessive, locative, demonstrative, presentative, quantitative, copulative and relative.

Adjective includes the traditional Dokean adjective (a closed class of roots which take noun prefixes as their agreement prefixes) and the predicative form of the Dokean relative, which is seen as an open class of adjectives (cf. van der Spuy (2006)). *Pronouns* are the personal pronouns, which may also (sometimes in allomorphic form) be used as agreement morphemes in quantifiers. Adverbs may be forms derived from adjectives by prefixing *ka-* to the root, or morphologically unanalysable forms like *phansi* 'in front, forward'. Ideophones have been included as adverbs. *Prepositionals* are words that incorporate the Dokean "adverbials" *na-* 'with', *nga-* 'by means of', *njenga-* 'like', *kuna-* 'more than', etc., which are better analysed as prepositions. The presentative is Doke's "locative demonstrative copulative" - the briefer name was suggested by van der Spuy (2001). *Copulatives* are all Doke's copulatives, excluding the adjectives mentioned above. *Relatives* are all predicative forms incorporating a relative prefix.

4 The labeling scheme

The labeling scheme has been based on the idea that each morpheme in a word should be labeled, even when words belong to a very restricted class. For example, the demonstratives could have been labeled as composite forms, but instead it is assumed that demonstratives contain between one and three morphemes, e.g. *le<d>si<d7>ya<po3>* 'a demonstrative of the third position referring to class 7' - i.e.. 'that one yonder, class 7'. It should be possible from this detailed labeling to build up an amalgam of the morphological structure of the word. The labels have been chosen to be both as brief as possible and as transparent as possible, though transparency was often sacrificed for brevity. Thus indicative subject prefixes are labeled *<i1-15>*, relative prefixes are labeled *<r>*, and noun prefixes are labeled *<n1-15>*; but negative subject prefixes are labeled *<g1-15>* and possessive agreement prefixes are labeled *<z1-15>*. Sometimes a single label was used for several different forms, when these are orthographically distinct, so for example *<asp>* (aspect) is used as a label for the following, among others: the continuative prefix *sa-* and its allomorph *se-*, the exclusive prefix *se-*, and the potential prefix *nga-* and its allomorph *ng-*. A person searching for forms containing the potential aspect would have to search for *'nga<asp> + ng<asp>'*. However, there should be no ambiguity, as the orthographic form would eliminate this. The detailed description of the scheme is provided by Spiegler et al. (2010).

5 Annotation process

The goal of this project was to build a reasonably sized corpus of morphologically annotated words of high quality which could be later used for developing and training automatic morphological analyzers. For this reason, we had gathered a list of the commonest Zulu word types, defined a partial grammar and parsed Zulu words with a logic algorithm which proposes possible parses based on the partial grammar. Compared to a completely manual approach, this framework provided possible annotations to choose from or the option to type in an annotation if none of the suggestions was the correct one. This semi-automatic process speeded up the labeling by an estimated factor of 3-4, compared to a purely manual approach. In Figure 1 we illustrate the annotation process and in the following subsections each step is detailed.

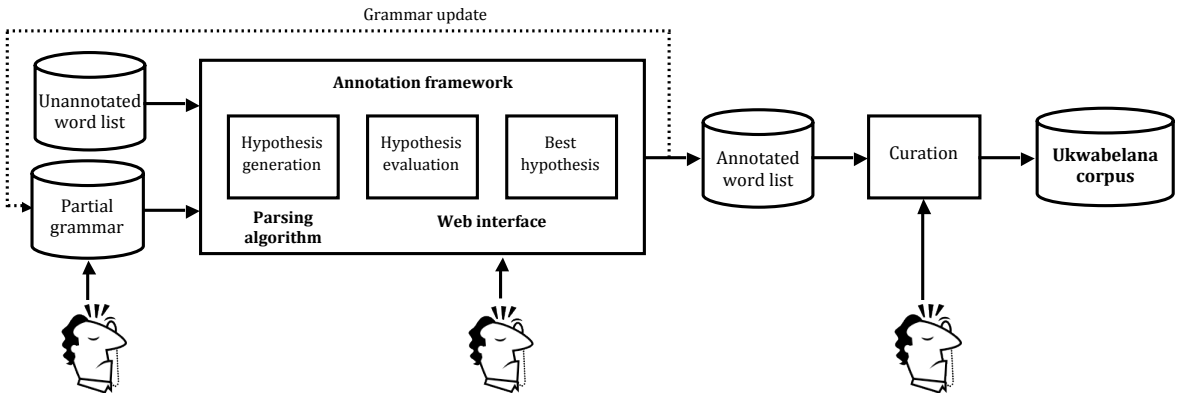

Figure 1: Process view of the annotation.

5.1 Unannotated word list

A list of unannotated Zulu words has been compiled from fictional works and the Zulu Bible. The original list comprises around 100,000 of the commonest Zulu word types. No information, morphological or syntactic, was given along with the words. We selected an initial subset of 10,000 words although our long-term goal is the complete analysis of the entire word list.

5.2 Partial grammar

Our choice for representing the morphological Zulu grammar was the formalism of *Definite Clause Grammars (DCGs)* used in the logic programming language *Prolog*. Although we defined our grammar as a simple context-free grammar, DCGs can also express context-sensitive grammars by associating variables as arguments to non-terminal symbols (Gazdar and Mellish, 1989). When defining our morphological grammar, we assumed that a linguistic expert could enumerate all or at least the most important morphological rules and morphemes of 'closed' morpheme categories, e.g. prefixes and suffixes of nouns and verbs. Morphemes of 'open' categories like noun and verb roots, however, would need to be hypothesized during the semi-automatic analysis and confirmed by the linguistic expert. Our final grammar comprised around 240 morphological rules and almost 300 entries in the morpheme dictionary. Since we did not only want to recognize admissible Zulu words but also obtain their morphological structure, we needed to extend our DCG by adding parse construction arguments as shown in the example below.

Example 3.

```
w((X)) --> n(X).
n((X,Y,Z)) --> iv(X),n2(Y),nr(Z).
iv(iv(a)) --> [a].
n2(n2(ba))--> [ba].
```

A possible parse for the word `abantu` 'people' could be `iv(a),n2(ba),*nr(ntu)` where '`*`' marks the hypothesized noun root.

With our partial grammar we could not directly use the inbuilt Prolog parser since we had to account for missing dictionary entries: Zulu verb and noun roots. We therefore implemented an algorithm which would generate hypotheses for possible parses according to our grammar. The algorithm will be described in the next subsection.

5.3 Hypothesis generation

For the hypothesis generation we reverted to logic programming and *abductive reasoning*. Abduction is a method of reasoning which is used with incomplete information. It generates possible hypotheses (parses) for an observation (word) and a given theory (grammar). Depending on the implementation, abduction finds the best hypothesis by evaluating all possible explanations. Our abductive algorithm is an extension of the meta-interpreter designed by Flach (1994) which only enumerates possible parses based on the grammar. A linguistic expert would then choose the best hypothesis. The algorithm invokes rules *top-down* starting with the most general until it reaches the last level of syntactic variables. These variables

are then matched against their dictionary entries from the left to the right of the word. A possible parse is found if either all syntactic variables can be matched to existing dictionary entries or if an unmatched variable is listed as *abducible*. Abducibles are predefined non-terminal symbols whose dictionary entry can be hypothesized. In our case, abducibles were noun and verb roots.

5.4 Evaluation and best hypothesis

Our annotation framework only enumerated allowable parses for a given word, therefore a linguistic expert needed to evaluate hypotheses. We provided a *web-interface* to the annotation framework, so that multiple users could participate in the annotation process. They would choose either a single or multiple correct parses. If none of the hypotheses were correct, the user would provide the correct analysis. Although our grammar was incomplete it still generated a substantial number of hypotheses per word. These were in no particular order and a result of the inherent ambiguity of Zulu morphology. We therefore experimented with different ways of improving the presentation of parses. The most promising approach was structural sorting. Parses were alphabetically re-ordered according to their morphemes and labels such that similar results were presented next to each other.

5.5 Grammar update

The grammar was defined in an iterative process and extended if the linguistic expert found morphemes of closed categories which had not been listed yet or certain patterns of incomplete or incorrect parses caused by either missing or inaccurate rules. The updated rules and dictionary were considered for newly parsed words.

5.6 Annotated word list and curation process

Although there had been great effort in improving the hypothesis generation of the parsing algorithm, a reasonable number of morphological analyses still had to be provided manually. During the curation process, we therefore had to deal with removing typos and standardizing morpheme labels provided by different experts. In order to guarantee a high quality of the morphological corpus, we also inspected single labels and analyses for their correctness. This was done by examining frequencies of labels and label combinations assuming that infrequent labels and combinations were likely to be incorrect and needed to be manually examined again. The finally curated corpus has an estimated error of 0.4 ± 0.5 incorrect single labels and 2.8 ± 2.1 incorrect complete analyses per 100 parses. Along with each word's analysis we wanted to provide part-of-speech (POS) tags. This was done by using a set of rules which determine the POS tag based on the morphological structure. We developed a prototype of a POS tagger which would assign the part-of-speech to a given morphological analysis based on a set of 34 rules. A summary of morphological analyses and words is given in Table 1. The rules are provided in Spiegler et al. (2010).

Category	**# Analyses**	**# Word types**
Verb	6965	4825
Noun	1437	1420
Relative	1042	988
Prepositional	969	951
Possessive	711	647
Copulative	558	545
Locative	380	379
Adverb	156	155
Modal	113	113
Demonstrative	63	61
Pronoun	38	31
Interjection	24	24
Presentative	15	15
Adjective	14	14
Conjunction	3	3
Total #	12488	10171

Table 1: Categories of labeled words.

5.7 POS tagging of sentences

In addition to the list of morphologically labeled words, we assigned parts-of-speech to a subset of 30,000 Zulu sentences. This task is straightforward if each word of a sentence only belongs to a single grammatical category. This was the case for 2595 sentences. For 431 sentences, however, we needed to disambiguate POS tags. We achieved this by analysing the left and right context of a word form and selecting the most probable part-of-speech from a given list of possible tags.

The overall error is estimated at 3.1 ± 0.3 incorrect POS tags per 100 words for the 3,000 sen-

Dataset	# Sentences	# Word tokens	#Word types	# Words per sentence	Word length
Raw	29,424	288,106	87,154	9.79±6.74	7.49±2.91
Tagged	3,026	21,416	7,858	7.08±3.75	6.81±2.68

Table 2: Statistics of raw and POS-tagged sentences.

tences we tagged. The summary statistics for raw and tagged sentences are shown in Table 2.

6 The Ukwabelana corpus - a resource description

The *Ukwabelana corpus* is three-fold:

1. It contains 10,000 morphologically labeled words and 3,000 POS-tagged sentences.
2. The corpus also comprises around 100,000 common Zulu word types and 30,000 Zulu sentences compiled from fictional works and the Zulu Bible, from which the labeled words and sentences have been sampled.
3. Furthermore, all software and additional data used during the annotation process is provided: the partial grammar in DCG format, the abductive algorithm for parsing with incomplete information and a prototype for a POS tagger which assigns word categories to morphologically analyzed words.

We are making these resources publicly available from `http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/Resources/` so that they will be of benefit to any person doing research on Zulu or on computer-aided analysis of languages.

7 Conclusions and future work

In this paper, we have given an overview of the morphology of the language Zulu, which is spoken by 23% and understood by more than half of the South African population. As an indigenous language with a written history of 150 years which was only recognised as an official languages in 1994, it is considerably under-resourced. We have spent considerable effort to compile the first open-source corpus of labeled and unlabeled words as well as POS-tagged and untagged sentences to promote research on this Bantu language. We have described the annotation process and the tools for compiling this corpus. We see this work as a first step in an ongoing effort to ultimately label the entire word and sentence corpus.

Our future work includes further automation of the annotation process by extending the described abductive algorithm with a more sophisticated hypothesis evaluation and by combining syntactical and morphological information during the decision process. Our research interest also lies in the field of automatic grammar induction which will help to refine our partial grammar. Another aspect is interactive labeling where a linguistic expert directs the search of an online parsing algorithm by providing additional information. Apart from the benefits to language researchers, we foresee an application of the corpus by machine learners which can develop and train their algorithms for morphological analysis.

Acknowledgements

We would like to thank Etienne Barnard and the Human Language Technologies Research Group from the Meraka Institute for their support during this project. Furthermore, we want to acknowledge Johannes Magwaza, Bruno Golénia, Ksenia Shalonova and Roger Tucker. The research work was sponsored by EPSRC grant EP/E010857/1 *Learning the morphology of complex synthetic languages* and a grant from the NRF (S. Africa).

References

Aronoff. 1976. *Word Formation in Generative Grammar*. The MIT Press.

Beesley and Karttunen. 2003. *Finite State Morphology*. University of Chicago Press.

Beuchat. 1966. The Verb in Zulu. *African Studies*, 22:137–169.

Bosch and Eiselen. 2005. The Effectiveness of Morphological Rules for an isiZulu Spelling Checker. *S. African Journal of African Lang.*, 25:25–36.

Botha and Barnard. 2005. Two Approaches to Gathering Text Corpora from the World Wide Web. *16th Ann. Symp. of the Pattern Recog. Ass. of S. Africa*.

Canonici. 1996. *Zulu Grammatical Structure.* Zulu Lang. and Literature, University of Natal, Durban.

Colenso. 1905. *Zulu-English Dictionary*. Natal, Vause, Slatter & Co.

Cope. 1984. An Outline of Zulu Grammars. *African Studies*, 43(2):83–102.

Davey. 1984. Adjectives and Relatives in Zulu. *S. African Journal of African Lang.*, 4:125–138.

Doke. 1927. *Text Book of Zulu Grammar*. Witwatersrand University Press.

Doke. 1935. *Bantu Linguistic Terminology*. Longman, Green and Co, London.

Doke. 1954. *Handbook of African Lang.*, chapter The S.ern Bantu Lang. Oxford University Press.

Doke, Malcom and Sikakana. 1958. *Zulu-English vocabulary*. Witwatersrand Uni. Press.

Du Plessis. 1993. *Linguistica: Festschrift EB van Wyk*, chapter Inflection in Syntax, pp. 61–66. Van Schaik, Pretoria.

Flach. 1994. *Simply Logical*. John Wiley.

Gauton. 1990. Adjektiewe en Relatiewe in Zulu. Master's thesis, University of Pretoria.

Gauton. 1994. Towards the Recognition of a Word Class 'adjective' for Zulu. *S. African Journal of African Lang.*, 14:62–71.

Gazdar and Mellish. 1989. *Natural Language Processing in Prolog*. Addison-Wesley.

Grout. 1859. *The Isizulu: A Grammar Of The Zulu Lang.* Kessinger Publishing.

Guthrie. 1971. *Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Lang.* Farnborough, Gregg International Publishers.

Joubert, Zimu, Davel, and Barnard. 2004. A Framework for Bootstrapping Morphological Decomposition. Tech. report, CSIR/University of Pretoria, S. Africa.

Khumalo. 1992. *African Linguistic Contributions*, chapter The morphology of the direct relative in Zulu. Via Afrika.

Lanham. 1960. *The Comparative Phonology of Nguni*. Ph.D. thesis, Witwatersrand Uni., Jo'burg, S. Africa.

Lanham. 1971. The Noun as Deep-Structure Source for Nguni Adjectives and Relatives. *African Studies*, 30:294–311.

Lieber. 1980. *On the Organization of the Lexicon*. Ph.D. thesis, Massachusetts Institute of Technology.

Lieber. 1992. *Deconstructing Morphology*. The University of Chicago Press.

Louw. 1984. Word Categories in Southern Bantu. *African Studies*, 43(2):231–239.

Mbadi. 1988. *Anthology of Articles on African Linguistics and Literature*, chapter The Percolation Theory in Xhosa Morphology. Lexicon, Jo'burg.

Meinhof. 1906. *Grundzüge einer Vergleichenden Grammatik der Bantusprachen*. Reimer, Berlin.

Nkabinde. 1975. *A Revision of the Word Categories in Zulu*. Ph.D. thesis, University of S. Africa.

Posthumus. 1987. Relevancy and Applicability of Terminology Concerning the Essential Verb Categories in African Lang. *Logos*, 7:185–212.

Posthumus. 1988. Identifying Copulatives in Zulu and S.ern Sotho. *S. African Journal of African Lang.*, 8:61–64.

Posthumus. 2000. The So-Called Adjective in Zulu. *S. African Journal of African Lang.*, 20:148–158.

Poulos and Msimang. 1998. *A Linguistic Analysis of Zulu*. Via Afrika.

Pretorius and Bosch. 2003. Finite-State Computational Morphology: An Analyzer Prototype For Zulu. *Machine Translation*, 18:195–216.

Pretorius and Bosch. 2007. Containing Overgeneration in Zulu Computational Morphology. *Proceedings of 3rd Lang. and Technology Conference*, pp. 54 – 58, Poznan.

Pretorius and Bosch. 2009. Exploiting Cross-Linguistic Similarities in Zulu and Xhosa Computational Morphology. *Workshop on Lang. Technologies for African Lang. (AfLaT)*, pp. 96–103.

Selkirk. 1982. *The Syntax of Words*. MIT Press.

Spiegler, Golenia, Shalonova, Flach, and Tucker. 2008. Learning the Morphology of Zulu with Different Degrees of Supervision. *IEEE Workshop on Spoken Lang. Tech.*

Spiegler, van der Spuy, Flach. 2010. Additional material for the Ukwabelana Zulu corpus. Tech. report, University of Bristol, U.K.

van der Spuy. 2001. *Grammatical Structure and Zulu Morphology*. Ph.D. thesis, University of the Witwatersrand, Jo'burg, S. Africa.

van der Spuy. 2006. Wordhood in Zulu. *S.ern African Linguistics and Applied Lang. Studies*, 24(3):311–329.

Van Eeden. 1956. *Zoeloe-Grammatika*. Pro Ecclesia, Stellenbosch.

Van Wyk. 1958. *Woordverdeling in Noord-Sotho en Zulu: 'n bydrae tot die vraagstuk van word-identifikasie in die Bantoetale*. Ph.D. thesis, University of Pretoria.

Wilkes. 1971. *Agtervoegsels van die werkwoord in Zulu*. Ph.D. thesis, Rand Afrikaans University.

Ziervogel, Louw, and Taljaard. 1985. *A Handbook of the Zulu Lang.* Van Schaik, Pretoria.