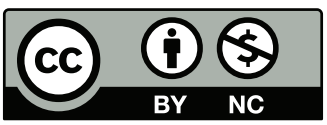
Computational morphology systems for Zulu – a comparison

Sonja Bosch
Department of African Languages
University of South Africa (UNISA)
PO Box 392
0003 UNISA, South Africa
boschse@unisa.ac.za

Abstract

The morphological analysis of Bantu languages, particularly for those with a conjunctive orthography such as Zulu, is crucial not only for the purposes of accurate corpus searches for Bantu linguists, but also as a basic enabling application that facilitates the development of more advanced tools and practical language processing applications, such as tokenising, disambiguation, part-of-speech tagging, parsing and machine translation. In this article, a comparison is made between four freely available computational morphology systems for Zulu, namely *isiZulu.net*, a Zulu–English online dictionary that also offers morphological analysis; *ZulMorph*, a finite-state morphological analyser for Zulu, currently available as a finite-state morphology demo; an open source morphological decomposer (available as modules and data) listed as the *NCHLT (National Centre for HLT) IsiZulu Morphological Decomposer*; and *CHIPMUN*K, a morphological segmenter and stemmer that contains components for modelling Zulu morphotactics. Criteria that are considered for the purposes of this comparison are, among others, accessibility and lookup capacity, embedded lexicons, degree of granularity of morphological analysis or decomposition, and also the documentation of tagsets used for purposes of analysis. Furthermore, the results of an evaluation based on recall and precision are presented. Against this background, this first comparison of four available Zulu computational morphology systems will be presented, based on output examples of a broad range of word categories with varying morphological complexity extracted by means of random sampling from the freely available Leipzig Wortschatz Collection corpus.

Keywords: computational morphology systems, morphological analyser, morphological decomposer, segmentation, Zulu morphology



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Disciplinary field of study: Computational Morphology

About the author

Sonja Bosch is a professor in the Department of African Languages at the University of South Africa (UNISA). Her main research interests are the linguistic aspects of natural language processing of Nguni languages, in particular morphological analysis, as well as the development of electronic lexical resources such as Wordnets.

1 Introduction[1]

Morphological analysis is the process of analysing words into their (potential) constituent morphemes, the smallest linguistic meaning-bearing units. Hammarström and Borin (2011, 310) point out that

> In language technology applications, a morphological component forms a bridge between texts and structured information about the vocabulary of a language. Some kind of morphological analysis and/or generation thus forms a basic component in many natural language processing applications.

Advances in research and in the production of sophisticated higher-level applications for human–computer communication largely depend on automated morphological analysis. Morphological analysis is therefore generally regarded as a basic enabling application that facilitates the development of more advanced tools and practical language processing applications, such as tokenising, part-of-speech tagging, syntactic parsing, and machine translation.

Morphological analysis is particularly relevant to languages belonging to the Bantu language family, with their unique morphological structure, which is heavily based on a nominal class system. Noun prefixes classify nouns into a number of classes which in turn connect a noun with other words in the sentence, such as the verb, adjective, pronoun, and so forth, by means of agreement morphemes, as shown in Table 1.

Table 1: Exemplification of grammatical agreement in Zulu

aba-ntu	***aba-ningi***	***ba-ya-ku[2]-theng-a***	***uku-dla***
2.person	$ADJP_2$-many	SP_2-PRES-OP_{15}-buy-FV	15.food
people	who are many	they buy it	food
‘many people buy food’			

The Bantu languages are essentially agglutinating in nature, since prefixes and suffixes are used extensively in word formation. Furthermore, some of these languages, including Zulu, have a conjunctive writing system “whereby meaningful units (morphemes) of the same linguistic word are joined in practical orthography” (Kosch 2006, 3). Hence a single word in Zulu often represents a complete sentence, as illustrated in Table 2.

[1] The author acknowledges the University of South Africa for granting her leave for research purposes. Dr Gertrud Faaβ’s valuable advice regarding the evaluation metrics of performance and the critical reading of a first draft of the article; Ms Marissa Griesel’s assistance with the installation of the CHIPMUNK segmenter, and three anonymous reviewers’ constructive inputs are thankfully acknowledged.

[2] Prefixes and pre-FV suffixes are traditionally presented with hyphens on both sides in Bantu linguistics, although they are not actually “infixes”.

Table 2: Analysis of *basazomphekela* 'they will still cook for him/her'

morpheme	***ba-***	***-sa-***	***-zo-***	***-m-***	***-phek-***	***-el-***	***-a***
category	SP_2-	-PROG-	-FUT-	$-OP_1$-	-COOK-	-APPLSUF -	-FV
English translation	they	still	will	him/her	cook	for	
	'they will still cook for him/her'						

In order to be of practical use, morphological analysis needs to be automated. The automatic analysis of word forms in a language such as Zulu is crucial for various reasons. For example, Cosijn et al. (2002) report inadequate results in experiments on cross-lingual information retrieval from Zulu to English owing to a lack of electronic resources and tools for morphological analysis. Prinsloo and de Schryver (2003, 323) state that the development of good quality spellchecking for a conjunctively written language such as Zulu requires automated morphological analyses. Subsequently, Bosch and Eiselen (2005) have proven that morphological analysis can contribute to the improved quality of a Zulu spellchecker. Pala et al. (2008) also emphasise that the development of superior mono- and multilingual lexical resources such as Wordnets and other lexical databases is not possible without morphological analysers.

Likewise, morphological analysis is crucial for the purposes of accurate Bantu language corpus searches, as pointed out by Hurskainen (1997, 631–633). Practical examples involve, for instance, the challenge of identifying monosyllabic verbs, surrounded by inflection and even derivation. In Zulu the retrieval of monophonemic verb roots[3] (for example, *-y-* 'go', *-f-* 'die', *-ph-* 'give', *-dl-* 'eat', *-mb-* 'dig', *-z-* 'come', *-w-* 'fall', *-th-* 'say', *-sh-* 'say', and *-lw-* 'fight') in a text corpus for the purposes of lemmatisation, frequency analysis, concordances, dictionary-lookup or language learning, becomes nearly impossible and very time consuming without the assistance of automated analysis.

Against this background of the fundamental need for automated morphological analysis to facilitate the development of more advanced tools and practical language processing applications for the agglutinating Bantu languages, a variety of computational morphology systems have been developed for Zulu. Some use rule-based approaches that usually require the writing of significant numbers of language-specific grammar rules, while others use statistical and machine-learning methods that depend on large amounts of annotated data, or even larger amounts of raw data.

The focus in this article is on a comparison of four freely available morphological analysers or decomposers for Zulu, namely *isiZulu.net*, a Zulu–English online dictionary that also offers morphological analysis; *ZulMorph,*[4] a finite-state morphological analyser for Zulu currently available as a finite-state morphology demo; an open source morphological decomposer catalogued as the *NCHLT (National Centre for HLT) IsiZulu Morphological Decomposer*; and *CHIPMUNK*, a morphological segmentation system that contains components for modelling Zulu morphotactics, among other languages.

Although further computational morphology systems for Zulu have been reported on, they have not been made available for use. One such analyser is described by Spiegler et al. (2008), who indicate that the *Ukwabelana* project "intends to deliver a morphological analyzer for a Zulu TTS system". Two years later Spiegler et al. (2010, 1024) reported as follows:

[3] See Poulos and Msimang (1998, 171) for a description of the phonological make-up of verb roots.
[4] The author of this article is a member of the development team of ZulMorph.

> 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.

The morphological analyser as such was not made available, only an annotated corpus containing 10,000 morphologically labelled words, 3,000 part-of-speech tagged sentences and a prototype of a part-of-speech tagger that assigns word categories to morphologically analysed words (see https://www.aflat.org/zulutag). It is relevant to note that the *Ukwabelana*[5] data was used as training data for the morphological segmentation system CHIPMUNK (Cotterell et al. 2015) which will be discussed in the next section.

A distinction needs to be made between the terms "morphological analyser/analysis" and "morphological decomposer/decomposition". Whereas morphological decomposition or segmentation entails the splitting of words into constituent morphemes, morphological analysis goes one step further by assigning labels ("tags") to the individual morphemes based on their grammatical function (Eiselen and Puttkammer 2014). In the description of the various tools it will be made clear whether the output is a mere decomposition of words into their individual morphemes, or whether the morphemes are also tagged according to their grammatical function.

In the next section an overview will be given of the four freely available computational morphology systems for Zulu, after which, in Section 3, these morphology systems will be explored critically according to qualitative rather than quantitative factors. An evaluation of the four morphological systems will be presented in Section 4, followed by a conclusion and suggestions for future directions.

2 Overview of four freely available computational morphology systems for Zulu

2.1 isiZulu.net

The tool isiZulu.net[6] functions primarily as a Zulu–English online dictionary but also offers automatic morphological analysis. According to the developer, the aim of isiZulu.net is to provide "a reasonably useful, modern Zulu–English online dictionary that anyone can contribute to" (isiZulu.net 2020). Lookups of a maximum of 40 characters (including white spaces) are bidirectional, that is, Zulu–English as well as English–Zulu. Furthermore, this online dictionary offers the conjugation and phonetic spelling of Zulu words, translation of simple Zulu and English phrases and spell checking in the sense that spelling suggestions are offered if a lookup is not successful. There is a section on grammar (a so-called *Zulu grammar cheat sheet*) and pronunciation basics, while a forum is available for suggesting new words and reporting incorrect entries. Little is known about the development of the online dictionary, except that "more than half of the code deals with regular expression based machine translation (aka computational linguistics, to use a buzzword), while a small part also does rule-based translations…" (isiZulu.net 2020).

[5] https://www.aflat.org/ukwabelana
[6] https://isizulu.net/

2.2 ZulMorph

ZulMorph[7] is currently available as a finite-state morphology demonstration system (Pretorius and Bosch 2018). Although the system would process large masses of text without difficulties, it only allows 99 characters (with spaces) to be analysed at once. All input is processed as lowercase text. Alongside the demo system, a brief introduction to Zulu morphology and an exposition of the ZulMorph tagset are also presented.

ZulMorph is reported on in detail in several publications, including Pretorius and Bosch (2010) and Bosch and Pretorius (2011). The morphological analyser addresses the two primary challenges of morphology, namely morphotactics (rules for ordering morphemes) and morphophonological alternation rules (orthographical rules and sound changes). The Zulu morphology is described with regular expressions that are compiled into a finite-state transducer, which constitutes the morphological analyser.

The finite-state computational morphological analyser for Zulu, ZulMorph, was initially developed using the Xerox finite-state toolkit (Beesley and Karttunen 2003). Successful compilation has also been done with *Foma* (Hulden 2009), a Xerox-compatible multipurpose finite-state compiler for constructing finite-state automata and transducers. *Foma* is free software licensed under the GNU general public licence.

The core components of ZulMorph (Bosch 2012, 130) are organised into two parts, namely the morphotactics component, which includes word roots, pronouns, demonstrative copulatives, and affixes for all parts-of-speech, as well as rules for legal combinations and orders of morphemes; and the morphophonological alternations component, which covers the rules that determine the form of each morpheme.

2.3 NCHLT IsiZulu Morphological Decomposer

The NCHLT IsiZulu Morphological Decomposer[8] (one of a series available for South African languages) is an open source morphological decomposer. According to Eiselen and Puttkammer (2014), the decomposers for conjunctively written languages such as Zulu are rule-based implementations, based on morphological analysis work previously done by Bosch et al. (2006). The fundamental methodology is the identification of all affixes recursively until no further affixes can be found. The residual elements are subsequently validated against a list of roots and stems. Analysis is deemed to be successful when a valid combination of affixes occurs with a valid stem or root. The set of affixes includes, among other things, relatives, negatives, verbal extensions, class concords, locatives, and various derivational affixes.

2.4 CHIPMUNK

CHIPMUNK[9] is an open source morphological segmentation system that explicitly models morphotactics, according to Cotterell et al. (2015, 164). It contains components for the modelling of Zulu morphotactics. The output of CHIPMUNK results in both segmentation and stemming (by means of a root and stem detector). The two processes are conducted in series, with

[7] https://portal.sadilar.org/FiniteState
[8] https://repo.sadilar.org/handle/20.500.12185/318
[9] http://cistern.cis.lmu.de/chipmunk/

each using the output of the previous one. CHIPMUNK is the only one of the four morphology systems that is deterministic in the sense that it always opts for only one analysis, as will be discussed further in Section 3.3.3.

3 Comparison

A first comparison of the four available Zulu computational morphology systems described in Section 2 will be presented based on qualitative rather than quantitative factors, such as output examples of a broad range of parts of speech with varying morphological complexity. Owing to restrictions on the number of input words in two of the systems, as well as the predominantly manual comparative analysis, it is not possible at this stage to compare a larger number of word forms than 60 tokens selected by means of random sampling.

The criteria that are considered for this comparison are as follows:

- accessibility and lookup capacity
- embedded lexicon
- output
  - documentation of tagsets
  - degree of granularity of morphological analysis
  - ambiguity
  - glosses/translations
  - non-standardised spelling/orthography
- improvement possibilities

3.1 Accessibility and lookup capacity

Two of the computational morphology systems, isiZulu.net and ZulMorph, are directly accessible online; isiZulu.net is also available as a mobile version. However, isiZulu.net only allows an input of 40 characters (with spaces included) at a time, while ZulMorph allows a short sentence (a maximum of 99 characters with spaces).[10] The NCHLT IsiZulu Morphological Decomposer and CHIPMUNK need to be installed locally, and, although this is not mentioned explicitly in the documentation, are then able to decompose a list of an unlimited number of words (offline). Previous test runs over wordlists of more than 500 words were successfully completed, but system memory limitations may become a consideration for wordlists of over 10,000 words. CHIPMUNK and isiZulu.net accept the input of capital letters; however, isiZulu.net gives an instruction "Please unlock your Caps Lock key" when capital letters are not valid for a specific word. ZulMorph changes capital letters to lowercase letters in the output except in the case of recognised proper nouns, whereas the NCHLT IsiZulu Morphological Decomposer requires the input to be in lowercase letters.

[10] Although ZulMorph is not available for offline use, the developers process lists of words on request (see Faaβ and Bosch 2019).

3.2 Embedded lexicon

The components and size of the embedded lexicon of ZulMorph are enumerated in Bosch and Pretorius (2011, 146), while the embedded lexicon of the NCHLT IsiZulu Morphological Decomposer forms part of the supporting (background) data modules contained in the executable file. These modules can be edited and expanded at the user's discretion. Table 3 compares the components and size of the embedded lexicons of these two systems.

Table 3: Embedded lexicons of ZulMorph and the NCHLT IsiZulu Morphological Decomposer

	ZulMorph	**NCHLT IsiZulu Morphological Decomposer**
Nouns/ noun stems	15,825 noun stems (including class information)	25,631 noun stems (sg. and pl., no class information)
Verb roots	7,597	5,839
Relative stems	408	371
Adjective stems	48	47
Conjunctions	176	56
Ideophones	2,583	0

The size or coverage of the isiZulu.net lexicon is not known; it is presumably a "living" lexicon due to a forum allowing users to contribute entries.

CHIPMUNK reports as follows on the datasets based on the *Ukwebelana* resource, which is a freely available morphologically annotated Zulu corpus (Spiegler et al. 2010):

- test data includes 11,271 words (10,634 stems, 11,392 roots), and
- training data consists of 1,211 words (1,129 stems, 1,200 roots).

3.3 Output

In this section the output of the four morphology systems is considered, with a focus on the documentation of tagsets, degree of granularity of the analyses, ambiguity of analyses, availability of English glosses or translations, and the treatment of words that do not conform to standardised or correct spelling and orthography.

3.3.1 Documentation of tagsets

In the case of isiZulu.net, an alphabetic list of 57 abbreviations is clarified under "Usage"[11], as illustrated in Figure 1.

part.	**pr.c.**
participial mood	pronomial concord
perf.	**pron.**
perfect (= recent past)	pronoun
pl.	**prp.**
plural	preposition
poss.	**QC, q.c.**
possessive	quantitative concord

Figure 1: Excerpt from the isiZulu.net tagset

On the demo webpage of ZulMorph[12], a detailed tagset with a description, an example, and an analysis of the example word is presented. The tagset is divided into two sections, namely those tags dependent on class, person, and/or number (23 tags), and those tags independent of class, person, and/or number (47 tags), as illustrated in Figure 2.

Tag	Description	Example	Analysis
Tags dependent on class, person and/or number			
2pp	second person plural	*ningashada*	ni[SC][2pp]nga[Pot]shad[VRoot]a[VT]
AC	Adjective concord	*obuningi*	obu[AC][14]ningi[AdjStem]
AdjPre	Adjective prefix	*sincane*	si[AdjPre][7]ncane[AdjStem]
BPre	Basic prefix	*abantu*	a[NPrePre][2]ba[BPre][2]ntu[NStem][1-2]
Dem	Demonstrative pronoun	*lena*	le[Dem][4][Pos1]

Tag	Description	Example	Analysis
Tags NOT dependent on class, person and/or number			
ProgPre	Progressive prefix	lisalandela	li[SC][5]sa[ProgPre]landel[VRoot]a[VT]
PronSuf	Pronoun suffix	sona	so[PronStem][7]na[PronSuf]
ProperName	ProperName	ujabulani	u[NPrePre][1a]Jabulani[NStem][1a-2a]
QuantStem	Quantitative stem	zonke	zo[QC][10]nke[QuantStem]

Figure 2: Excerpt from the ZulMorph tagset documentation

[11] https://isizulu.net/usage/
[12] https://portal.sadilar.org/FiniteState/demo/zulmorph/doc.html#tagset

According to Pretorius and Bosch (2003, 204), "tags were devised that consist of intuitive mnemonic character strings that abbreviate the features they are associated with".

Since the NCHLT IsiZulu Morphological Decomposer is not an analyser but a decomposer, there are no tags involved; tokens are merely split into their constituent morphemes and the boundaries between affixes, roots, or stems are marked by means of hyphens (Eiselen and Puttkammer 2014, 3701).

In the README file accompanying CHIPMUNK, the tagsets are described on various levels of granularity, as shown in Table 4. The basic level results in segmentation of the word with the tag SEGMENT, that is, Level 0, doing decomposition only. On Level 1, the identification of each segment as either PREFIX, ROOT, or SUFFIX is presented. On Level 2, prefixes and suffixes are labelled as inflectional INFL or derivational DERIV. Cotterell et al. (2015, 165–166) describe additional tagsets on Levels 3 and 4. On Level 3, labels categorising components as being VERBAL, NOMINAL, or ADJECTIVAL are added. Lastly, on level 4, inflectional features of a suffix, for example, CASE or NUMBER, may be added. However, these last two levels of tagging were not implemented for the Zulu language.

Table 4: Overview of the CHIPMUNK tagset

Level	Tag	Tag	Tag
4	PREFIX:INFL:NOUN PREFIX:DERIV:VERB	ROOT:NOUN ROOT:VERB ROOT:ADJ	SUFFIX:INFL:NOUN: PLURAL SUFFIX:INFL:NOUN: SINGULAR SUFFIX:DERIV:NOUN
3	PREFIX:INFL PREFIX:DERIV	ROOT:NOUN ROOT:VERB ROOT:ADJ	SUFFIX:INFL:NOUN SUFFIX:DERIV:NOUN
2	PREFIX:INFL PREFIX:DERIV	ROOT	SUFFIX:INFL SUFFIX:DERIV
1	PREFIX	ROOT	SUFFIX
0	SEGMENT		

In summary, the tagsets documented for isiZulu.net, as well as for ZulMorph, although fine-grained with detailed analysis, can be described as flat tagsets consisting of relatively large lists of independent categories describing tag sequences, rather than a hierarchical structure. The CHIPMUNK tagset, on the other hand, leans more towards a hierarchical tagset with a smaller number of categories structured relative to one another, rather than a large number of independent categories.

3.3.2 Degree of granularity

In order to compare the degree of granularity of analysis or decomposition, two examples of

relatively complex words, representing a variety of morphemes, were selected randomly. By means of the words *kwakungumuntu* 'it was a person', with a noun stem as lexical core, and *okhulumela* 'who speaks for; they/it will speak for', with a verb root as lexical core, the varying results with respect to the granularity of analysis and tagging practices as produced by the four morphology systems are explained. Table 5 represents the granularity of analysis or decomposition of *kwakungumuntu* 'it was a person'. Please see Appendix 1 for the full analyses including tags or decompositions.

Table 5: Granularity of analysis or decomposition of kwakungumuntu 'it was a person'

	Copulative construction	**Copulative noun prefix**	**Noun**
ZulMorph	kwa[SCPT][15] be[AuxVStem] ku[SC][15]	ngu[CopPre]	u[NPrePre][1] mu[BPre][1] ntu[NStem][1-2]
isiZulu.net	SC: kwaku- (cl. 15, cl. 17)	ID: ngu-	umuntu/abantu n. 1/2 (-ntu)
NCHLT	kwaku	ng	umuntu
CHIPMUNK	kw:INFL|PREFIX| SEGMENT a:INFL|PREFIX| SEGMENT ku:DERI|PREFIX| SEGMENT	ng:DERI|SUFFIX| SEGMENT	u:INFL|PREFIX| SEGMENT mu:DERI|SUFFIX| SEGMENT ntu:ROOT|SEGMENT

The ZulMorph analysis of the copulative construction preceding the noun *umuntu* 'person/human' is extremely fine-grained in the sense that the prefix *kwaku-* is analysed as two subject concords (class 15), the second of which is the concord of an underlying auxiliary verb stem *-be* 'was/became', which is omitted in the surface structure in certain instances (see Taljaard and Bosch 1993, 149). This compound predicate is followed by the copulative noun prefix *ngu-,* involving vowel elision. The nominal part is broken down into the noun pre-prefix *u-,* followed by the class 1 basic noun prefix *-mu-*, as described in grammatical works such as Doke (1973) and Poulos and Msimang (1998).

The isiZulu.net analysis is less fine-grained, since the remote past continuous form *kwaku-* is analysed as single subject concord (in either class 15 or 17). The identifying prefix is indicated as *ngu-,* and the full singular form of the class prefix *umu-* is provided, as well as class information (class 1/2). Vowel elision is taken into account: *ngu-* + *umu-* results in *ngumu-*, as in the analysis of ZulMorph.

The NCHLT IsiZulu Morphological Decomposer presents *kwaku-* as the first decomposed part (in line with the analysis of isiZulu.net), while the copulative prefix is decomposed as *-ng-*, thus avoiding morphophonological issues. The noun *umuntu* 'person/human' is not decomposed at all, although the identification of the class prefix *umu-* and the noun stem *-ntu* would have been expected, particularly since Eiselen and Puttkammer (2014, 3702) state that

The basic approach to decomposition is to identify all affixes recursively until no additional affixes can be found. The remaining constituent is then verified against a lexicon of roots and stems, and only in those instances where a valid combination of affixes along with a valid stem or root is found, will the decomposition be successful. The set of affixes consists of various grammatical classes, including relatives, negatives, verbal extension, concord classes, locatives and various derivational affixes.

The output of CHIPMUNK results firstly in segmentation of the word, and then identification of each segment as either a prefix, a root, or a suffix. Prefixes and suffixes are then labelled as either derivational or inflectional. The root of the noun *umuntu* 'person/human' is correctly recognised as *-ntu*, and the overall segmentation of the compositional part as *u-mu-ntu* is accurate and finely grained.

A summary of the granularity of analysis or decomposition of *okhulumela* 'who speaks for; they/it will speak for' is presented in Table 6. Full analyses, including tags or decompositions, are presented in Appendix 1.

Table 6: Granularity of analysis or decomposition of okhulumela ('that/which/who speaks for; they/it will speak for')

	Verb prefixes	**Verb stem**
ZulMorph	o[RC][1] o[RC][2ps] o[RC][3]	khulum[VRoot] el[ApplExt] a[VT]
	a[SC][6] yo[Fut]	
isiZulu.net	RC: o- (cl. 1, cl. 1a, cl. 3)	khulumela (v/t.)
	SC: a- (cl. 6) zo (future tense)	
NCHLT	o	khulum-el-a
CHIPMUNK	o:DERI|SUFFIX|SEGMENT	khulum:ROOT|SEGMENT el:DERI|SUFFIX|SEGMENT a:DERI|SUFFIX|SEGMENT

ZulMorph analyses the verb *okhulumela* 'who speaks for; they/it will speak for' down to the smallest morpheme. Three class variations of the subject concord (class 1, 3, and 2ps), and the option of a class 6 subject concord assimilated with a phonologically conditioned future tense morpheme (see Doke 1973, 175), are analysed. This is followed by the correctly identified verb root *-khulum-* 'speak'. ZulMorph also analyses the so-called verbal extensions, in this example the valency increasing applied extension *-el-*, which is suffixed to the verb root *-khulum-* 'speak', thereby modifying the basic meaning of the verb root.

The lookup of *okhulumela* 'who speaks for; they/it will speak for' in isiZulu.net results in detailed information on potential decompositions, which include prefixes with class information where applicable, such as subject concord (SC), relative concord (RC), and future tense prefix.

In addition to the identification of the extended verb stem *-khulumela* 'talk for', transitivity is also indicated (v/t). So-called verbal extensions that modify the basic meaning of the verb root, such as the applied extension *-el-* in this example, are not analysed separately. In other words, the analysis does not go further than the extended stem (see Kosch 2006, 10).

In the NCHLT IsiZulu Morphological Decomposer, the verb *okhulumela* 'who speaks for; they/it will speak for' is decomposed into its relevant prefixes and suffixes, as well as the verb root *-khulum-* 'speak', the applied extension *-el-*, and the verb final vowel *-a.* The boundaries are marked by means of hyphens. As can be expected from a morphological decomposer, no tags for the various morphemes are provided.

The output of CHIPMUNK results firstly in the segmentation of the word, and then the identification of each segment as either a prefix, a root, or a suffix. Prefixes and suffixes are then labelled as either derivational or inflectional. The root of the verb *okhulumela* 'who speaks for; they/it will speak for' is correctly identified as *-khulum-* 'speak', and the segmentation is also correct, although the tags are rather odd – the first segment of the word being tagged as a suffix. The segmentation of this verb demonstrates the same granularity as that of ZulMorph.

The lookup of just two examples thus gives a good overall impression of the degree of granularity of analysis by the four morphology systems.

3.3.3 Ambiguity

It is well known that complex words are often ambiguous. Concerning ambiguity in the morphological analysis of Swahili, Hurskainen (1996, 573) observes

> The morphological analysis of Swahili tends to produce a comparatively large number of ambiguous readings. The noun class structure coupled with class agreement marking in dependent constituents, contributes significantly to ambiguity. The phenomenon is particularly evident in verb structures, where different sets of noun class markers add to the ambiguity of the same verb-form.

The same holds true for a language such as Zulu. There is no real solution to ambiguity at the morphological level, and this can only be resolved through semantic context-based disambiguation at a later stage of processing. The extent of ambiguity manifests itself more in the two morphological analysers for obvious reasons, in particular with regard to class information. The fact that several noun classes have identical prefixes leads to one of the main problems of disambiguation. It was reported in Faaß and Bosch (2019, 226) that a word such as *abazi*, with several meanings, resulted in 105 different analyses in ZulMorph. These range from the noun *abazi* 'connoisseurs; ones who know' and *abazi* 'of (the) connoisseurs', to the verb *abazi* in various tenses 'he/she/it/they know(s) them; they/it knew them', in the affirmative as well as the negative ('they do not know'), as a hortative construction ('let them know'), a relative construction ('that/which/who know them'), and with the subject concord in various classes (class 1, 1a, 6 etc.) and moods (indicative and subjunctive). This diversity of analyses is also evident in the output of *abazi* in isiZulu.net, where translations are provided as well. The lists of analyses for *abazi* produced by ZulMorph and isiZulu.net are too extensive to be repeated here, but can be checked online.

The decomposition of *abazi* by the NCHLT IsiZulu Morphological Decomposer results in just three options, one of which is no decomposition at all, while the other two options avoid

morphophonological issues, thus not providing the full morphemes of, for instance, the object concord, as revealed in Figure 3.

abazi	a-b-azi
abazi	ab-azi
abazi	abazi

Figure 3: NCHLT IsiZulu Morphological Decomposer decomposition of abazi

As mentioned earlier, CHIPMUNK is the only one of the examined tools that is deterministic; that is, it always opts for only one segmentation possibility, as Figure 4 shows.

a:DERI|SUFFIX|SEGMENT ba:INFL|PREFIX|SEGMENT zi:ROOT|SEGMENT

Figure 4: CHIPMUNK segmentation of abazi

Overall, morphological analysis or decomposition promises to be a valuable source of information for the identification of contextually valid options as part of the development of a tool for (semi-) automated disambiguation in future work. Considering that CHIPMUNK and the NCHLT tool are both tools that can be integrated into a natural language processing (NLP) chain, their output might be somewhat problematic: the repetition of the full form in the NCHLT tool is not useful, as such an output indicates that there might be no segmentation necessary. Looking at the CHIPMUNK output, it always determines one analysis as the correct one, even if there are several possible analyses for which disambiguation in context by a tagger would be possible. Even though ZulMorph delivers a high rate of morphological ambiguity, it would still be considered the most practical option if made available for local processing. Since isiZulu.net was developed primarily as an online dictionary, it is not necessarily expected that every word in question should be able to be analysed. Instead, the expectation is that users' needs for analyses, usually going hand in hand with the frequency of occurrences in corpora, as shown by de Schryver et al. (2019) in the case of a Swahili–English online dictionary, are satisfied.

3.3.4 English glosses/translations

As a bilingual dictionary, isiZulu.net is the only one of the four morphology systems that provides English translations. Although this system is described as not being a comprehensive text translator, it attempts to translate single words, compound expressions, and simple phrases. In the case of nouns, a translation of the singular form of the noun is also given.

3.3.5 Non-standardised orthography

A general source of inconsistency found in Zulu corpora is the use of outdated or non-standardised orthography. A typical example is the demonstrative pronoun, which, when occurring

before the noun, was written conjunctively in older texts (e.g. *lelikhaya* 'this home/homestead'), but was later officially changed to being written disjunctively (i.e. *leli khaya* 'this home/homestead'), according to IsiZulu terminology and orthography No. 4 (Department of Education and Training 1993, xii).

isiZulu.net is the only one of the four systems that provides spelling suggestions, that is, a default orthography which entails showing entries similar to the input word. In the case of the example of *leli khaya*, the correct orthography is suggested, as shown in Figure 5.

leli **dem. pron.** [ˈle:li]	
cl. 5	this; this one

-khaya **n. stem**	
n. 5/6	ikhaya
n. 3/-	umkhaya

Compound Expressions (verbatim)
leli khaya
this home
this household

Figure 5: Screenshot of spelling suggestion in isiZulu.net after entering *lelikhaya*

Neither ZulMorph nor the NCHLT IsiZulu Morphological Decomposer is capable of analysing the outdated orthographical version of *lelikhaya* 'this home/homestead'. Although CHIPMUNK produces a segmentation, the demonstrative pronoun is incorrectly segmented.

3.4 Improvement possibilities

This section investigates the flexibility of the four computational morphology systems with regard to their output.

isiZulu.net and ZulMorph are packaged in so-called black boxes, that is, the inputs and outputs of the system are known, but the internal workings are not known to the user. Therefore, users are not able to manage any improvements themselves; instead they are offered the opportunity to provide feedback, such as suggesting new words, reporting wrong entries, and so forth. Such suggestions can then be considered for inclusion by the developers.

In the NCHLT IsiZulu Morphological Decomposer, on the other hand, all the resources are accessible as open-source modules and data which can be improved and extended by researchers and developers as they see fit (Eiselen and Puttkammer 2014, 3702). In other words, the supporting modules distributed as an executable file can be edited and expanded at the user's discretion. As an experiment, a new verb root *-bhorek-* 'get bored' was added to the data module *Verb*, and a new class 1a/2a noun stem *-solwazi* 'professor' was added to the data module *Noun*. Prior to these new additions, the output was incorrectly decomposed or not decomposed at all, as indicated by the starred forms under Output 1 in Table 7.

The addition of *bhorek-* 'get bored' resulted in the correct decomposition of *ubhoreka* 'he/she/it gets bored' and *ukubhoreka* 'to get bored', particularly with regard to the verb root. The third example, *bayabhoreka* 'they get bored', was still not decomposed at all. Similarly, notwithstanding the addition of the noun stem *-solwazi* 'professor' with its class 1a/2a information, *usolwazi* was still decomposed incorrectly or not decomposed at all. The only correct decomposition was that of *njengosolwazi* 'like the/a professor'. It would therefore appear that the addition of new verb roots or nouns to the data modules does not contribute significantly to improved output, as illustrated under Output 2 in Table 7.

Table 7: Examples of the output of newly added items to NCHLT IsiZulu Morphological Decomposer data modules

Input	**Output 1**	**Output 2**
ubhoreka 'he/she/it gets bored'	**u-b-horeka*	*u-bhorek-a*
ukubhoreka 'to get bored'	**u-ku-bhoreka*	*u-ku-bhorek-a* *uku-bhorek-a*
bayabhoreka 'they get bored'	**bayabhoreka*	**bayabhoreka*
usolwazi 'professor'	**u-sol-w-azi*	**u-sol-w-azi* **usolwazi*
njengosolwazi 'like the/a professor'	**njengosolwazi*	*njenga-u-solwazi*

(*incorrect decompositions)

Although the datasets in CHIPMUNK are made up of different (human readable) text files which contain test data and training data, it is not made clear how the morphological system can be improved by adding information to these datasets. For a heuristic tool like this, an external user has to become a developer him-/herself in order to enhance results. Cotterell et al. (2015, 172) do state, however, that "a primary goal of future work will be to use CHIPMUNK to attempt to induce higher-quality morphological processing systems".

4 Evaluation of the computational morphology systems

The four computational morphology systems are evaluated according to the metrics of performance termed *recall* and *precision* in this section. Data for the evaluation was extracted from the Wortschatz Universität Leipzig (2020) collection that contains approximately three million tokens with marked sentence boundaries. In total, 149,196 sentences (2,337,566 tokens) were selected for local processing, after deleting noise. Subsequently, twenty types of tokens were selected in each of the following sectors of the dataset, ensuring good coverage in terms of frequent word forms, but also of other randomly selected word forms:

a) most frequent,
b) median (random selection of tokens with occurrences between 1 and 9, but excluding any that might already occur in a)), and
c) random sampling of hapaxes (tokens with only one occurrence).

The tokens selected[13] had to consist of at least two letters, hyphens were allowed, and all words were changed to lower-case letters, as shown in Appendix 2.

For NLP tools, as explained by Faaβ (2011) in her evaluation of SMOR,[14] a morphological analyser for German, recall is calculated from the number of "negative" analyses or decompositions of word forms (not recognised by the tool) and from the number of all analyses or decompositions of word forms included in the test data. The percentage calculated indicates the success rate. Usually, a distinction is made between "true negatives" and "false negatives". Analyses or decompositions of word forms that are not found, and are not expected to be found, for example misspelled words, are considered to be "true negatives". An example is the misspelled word *ababetshona* (see Appendix 3, example 1). On the other hand, "false negatives" refer to analyses or decompositions of word forms that the tools have not recognised, although they are correctly written, for example *ngenxa* 'because of' or 'with a portion/share' (see Appendix 3, example 2). Precision is calculated at the analysis level by Faaß (2011); that is, each analysis for each word form is included in the calculation of the value. The two precision categories are "true positives", which include found and correct analyses or decompositions, and "false positives", which entail found but incorrect analyses, as exemplified in the word *eyayikhombisa* 'that/which/who showed them/him/her/it' (see Appendix 3, example 3).

Accuracy is calculated as follows:

(true positives + true negatives) /
(true positives + false positives + true negatives + false negatives)

The overall results of the evaluation of the four computational morphology systems, based on the output of a broad range of word categories with varying morphological complexity, are summarised in Table 8.

Table 8: Overall result of the evaluation of the four computational morphology systems

OVERALL RESULT	**Number of correct analyses: 245** **Number of possible analyses: 332**							
	found	correct	wrong	true pos	false pos	true neg	false neg	**accuracy**
ZulMorph	283	224	59	226	59	9	25	**73.0%**
isiZulu.net	173	145	28	134	23	8	99	**54.4%**
NCHLT	76	41	33	39	30	5	79	**28.8%**
CHIPMUNK	60	26	31	26	30	4	76	**22.1%**

[13] The random selection of tokens for evaluation is intended to prevent any bias towards any one of the four morphology systems.
[14] https://www.cis.uni-muenchen.de/~schmid/tools/SMOR/

In order to calculate accuracy, the number of technically possible and the number of correct analyses must be known. However, experience shows that humans are usually not able to predict the expected analyses by such tools, as real-world knowledge often hinders us from imagining analyses that are technically possible and might even be correct (but would never occur in reality). To solve this problem, the tool delivering the highest number of analyses is taken as a default for calculating recall; these analyses were, however, manually validated.

The evaluation of the computational morphology systems was carried out semi-automatically, owing to the diversity of output formats of the four systems, as exemplified in Appendix 1. Each analysis or decomposition was carefully inspected and the results documented against the relevant system in a spreadsheet, an excerpt of which is included in Appendix 3. Although space restrictions do not allow a full discussion of all practical matters that emerged during the evaluation, a few noteworthy issues are discussed below.

A distinction had to be made between the results of the two systems producing analyses, that is, isiZulu.net and ZulMorph, and those of the other two systems, producing decompositions and segmentations respectively, that is the NCHLT isiZulu Morphological Decomposer and CHIPMUNK. In the output of CHIPMUNK, only the decomposed segments of Level 1 are taken into consideration for the evaluation. Looking, for instance, at the example *okhulumela* 'that/which/who speaks for' or 'they/it will speak for' (see Appendix 3, example 4), the accuracy of all systems is 100%, although the two analysers have four true positives, and the two decomposers have only one true positive. The reason for this is that although the decomposition displayed in the analysers is identical to that of the decomposers, the granularity of the tags in the analysers adds additional possibilities such as class information.

Although closed classes are usually successfully analysed, it becomes apparent in the evaluation that in the case of some pronouns, the output delivered does not cover all possible analyses. Demonstrative pronouns such as *lokhu* 'this, this one, these' and *lapho* 'there' also function as conjunctions, namely *lokhu* 'since, as, inasmuch as' and *lapho* 'when'. Therefore, analysers are expected to include both analyses with their relevant tags, whereas decomposers are merely evaluated on the word form. This becomes apparent in Appendix 3, example 5.

Absolute pronouns often have more than one analysis owing to the agglutinating nature of the language; for example, *khona* 'it' also functions as a conjunction ('so that; in order that'), as an adverb of place, and even as a noun stem *-khona* 'corner', according to isiZulu.net. The evaluation demonstrated in Appendix 3, example 5, shows that the 100% accuracy of isiZulu.net can, in fact, be ascribed to its additional recognition of the noun *ikhona* 'corner', which is missing in the ZulMorph analyses, as well as in the decompositions of the NCHLT IsiZulu Morphological Decomposer and CHIPMUNK.

The degree of granularity of analyses plays a significant role in the evaluation. For instance, the noun *inkosi* 'chief, king' has an underlying stem *-khosi*, which has undergone a process of deaspiration (see Poulos and Msimang 1998, 526), as may be construed from the plural form *amakhosi* 'chiefs, kings'. The ZulMorph and isiZulu.net analyses recognise the basic form of this noun stem, while the NCHLT IsiZulu Morphological Decomposer and CHIPMUNK do not. The result is reflected in the accuracy score, as shown in Appendix 3, example 7.

With regard to lookup, it is noteworthy that all four systems change capital letters to lowercase, except in the case of well-known proper names such as *uJehova* (Jehovah), where the capital letter of the stem is retained in the output of isiZulu.net and ZulMorph.

Table 9 shows a summary of the features of the four computational morphology systems evaluated, after a comparative test of the systems.

Table 9: Summary of the features of the four evaluated computational morphology systems

Criteria	isiZulu.net	ZulMorph	NCHLT	CHIPMUNK
Accessibility	++	++	+	+
Lookup capacity	+	+	++	++
Embedded lexicon	+	++	++	++
Documentation of tagsets	+	+	-	+
Degree of granularity	++	+++	+	+
Ambiguity	++	+++	+	-
Glosses/translations	+	-	-	-
Non-standard spelling	+	-	+	+
Improvement possibilities	-	-	+	-

5 Conclusion and future directions

In this article, a first comparison between four freely available computational morphology systems for Zulu has been presented. Criteria used for the comparison are, among others, accessibility and lookup capacity, embedded lexicons, degree of granularity of morphological analysis or decomposition, and the documentation of tagsets used for analysis. For the purposes of evaluation, a dataset of 60 tokens was extracted from a freely available corpus by means of random sampling. The dataset includes tokens from diverse word categories with varying morphological complexity. By evaluating the computational morphology systems in terms of recall and precision, an accuracy score for each system could be determined, along with specific strengths and shortcomings of the various systems.

The general finding is that the two most accurate and finely grained computational morphological analysers, namely isiZulu.net and ZulMorph, are easily accessible online, and have detailed tagsets and documentation readily available. However, the main drawback in both cases is the limited lookup possibilities, thereby making it very challenging or even impossible to process large sets of data or to integrate them into an NLP tool chain. The extremely fine-grained output of morphological analysis, as well as the format of the output of ZulMorph, gives this system an advantage over isiZulu.net, but leads to a much higher rate of morphological ambiguities which might be too finely grained for the purposes of certain kinds of further processing, where such details are not relevant. ZulMorph could, however, still provide all necessary information for higher-level processing tasks such as part-of-speech tagging, syntactic parsing, and so forth, if the current demo version were to be extended to a full version and made available for download, that is, for local use and integration into tool chains. In general, offline tools are preferable since they can be utilised by or combined with other NLP tools. isiZulu.net remains an extremely useful and accurate online dictionary and morphological analyser not only for language learning purposes, but also for translation, spellchecking, and grammar.

The output of the NCHLT IsiZulu Morphological Decomposer, as well as that of CHIPMUNK, falls under the category of decomposition or unlabelled morphological segmentation. Both systems have the advantage of sizeable lookup capacity. Although the developers of the NCHLT IsiZulu Morphological Decomposer claim that the system could contribute to the development of processing tasks such as named entity recognition systems, chunkers, parsers, and many more (Eiselen and Puttkammer 2014), inconsistent granularity and flawed decom-

position, as revealed in the accuracy scores, are a critical drawback of this system. The main shortcoming of the CHIPMUNK segmentations is their inflexibility in recognising ambiguities. This becomes clear in the results of the evaluation, where CHIPMUNK produces only one fixed set of segments per word.

All four tools could be functional in different use cases. While isiZulu.net assists Zulu learners to better understand the formation and inflection of Zulu words, ZulMorph provides more finely grained, rich morphological information for linguists and – if made available for local use – could prove valuable when integrated into an NLP tool chain that goes towards part-of-speech tagging and even parsing. With regard to possible other NLP use cases, the NCHLT IsiZulu Morphological Decomposer tool and CHIPMUNK could be employed for the development of stemmers in Information Retrieval or similar tasks.

Finally, a future research goal is the development of a gold standard for Zulu morphological analyses. Such a gold standard can be utilised for further in-depth evaluation and comparison of computational morphology systems. Given a gold standard of significant size, training of new heuristic or neural tools would be possible as well. The ambiguous output of a computational morphology system will naturally remain, but is usually dealt with by a part-of-speech tagger that considers each word in context and assigns it a single reading.

References

Beesley, Kenneth R. and Lauri Karttunen. 2003. *Finite state morphology*. Stanford: CSLI Publications.

Bosch, Sonja E. 2012. "Bootstrapping the Development of Morphological Analysers for 'Dispersed' Nguni Languages: A Linguistic Investigation." *Language Science and Language Technology in Africa: A Festschrift for JC Roux*, edited by H. Steve Ndinga-Koumba-Binza and Sonja Bosch, 123–144. Stellenbosch: SUN Press.

Bosch, Sonja E. and Roald Eiselen. 2005. "The effectiveness of morphological rules for an isiZulu spelling checker." *South African Journal of African Languages* 25, no. 2: 25–36.

Bosch, Sonja, Jackie Jones, Laurette Pretorius and Winston Anderson. 2006. "Computational Morphological Analysers and Machine-Readable Lexicons for South African Bantu Languages." *Localisation Focus: The International Journal of Localisation* 6, no. 1: 22–28.

Bosch, Sonja E. and Laurette Pretorius. 2011. "Towards Zulu Corpus Clean-up, Lexicon Development and Corpus Annotation by Means of Computational Morphological Analysis." *South African Journal of African Languages* 31, no. 1: 138–158.

Cosijn, Erica, Ari Pirkola, Theo Bothma and Kalervo Järvelin. 2002. "Information access in indigenous languages: A Zulu case study." *South African Journal of Libraries and Information Science* 68, no. 2: 103–114.

Cotterell, Ryan, Thomas Müller, Alexander Fraser and Hinrich Schütze. 2015. "Labeled Morphological Segmentation with Semi-Markov Models." *Proceedings of the 19th Conference on Computational Language Learning,* 164–174. Beijing, China.

de Schryver, Gilles-Maurice, Sascha Wolfer and Robert Lew. 2019. The relationship between dictionary look-up frequency and corpus frequency revisited: A log-file analysis of a decade of user interaction with a Swahili-English dictionary. *GEMA Online® Journal of Language Studies* 19, no. 4: 1–27.

Department of Education and Training. 1993. *IsiZulu terminology and orthography No. 4*. Pretoria: The Government Printer.

Doke, Clement. M. 1973. *Textbook of Zulu Grammar.* Cape Town: Maskew Miller Longman.

Eiselen, Roald, and Martin Puttkammer. 2014. "Developing Text Resources for Ten South African Languages." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 3698–3703. Reykjavik, Iceland.

Faaß, Gertrud. 2011. "Stuttgarter Morphologisches Analysewerkzeug (SMOR): Dokumentation." https://www.cis.uni-muenchen.de/~schmid/tools/SMOR/dspin/ (accessed 9 April 2020).

Faaß, Gertrud and Sonja Bosch, 2019. "Towards a gold standard corpus for detecting valencies of Zulu verbs". In *Proceedings of 15th Conference on Natural Language Processing (KONVENS 2019)*, 223–228. Erlangen, Germany.

Hammarström, Harald and Lars Borin. 2011. "Unsupervised Learning of Morphology." *Computational Linguistics* 37, no. 2: 309–350.

Hulden, Mans. 2009. "Foma: A Finite-state Compiler and Library." In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 29–32.

Hurskainen, Arvi. 1996. "Disambiguation of morphological analysis in Bantu languages." *COLING '96 Proceedings of the 16th conference on Computational linguistics, Volume 1*, 568–573.

Hurskainen, Arvi. 1997. "A language sensitive approach to information management and retrieval: the case of Swahili." In *African Linguistics at the Crossroads: Papers from Kwaluseni*, edited by Robert K. Herbert, 629–642. Cologne: Rüdiger Köppe.

isiZulu.net: Bilingual Zulu–English dictionary. 2020. https://isizulu.net/ (accessed 9 April 2020).

Kosch, Inge M. 2006. *Topics in Morphology in the African Language Context*. Pretoria: Unisa Press.

Pala, Karel, Sonja Bosch and Christiane Fellbaum. 2008. "Building resources for African languages." *Proceedings of the 6th International Conference on Language Resources and Evaluation* (Workshop on Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages), Marrakech, Marocco, 13–17.

Poulos, George and Christian T. Msimang, 1998. *A Linguistic Analysis of Zulu.* Pretoria: Via Afrika.

Pretorius, Laurette and Sonja E. Bosch. 2003. "Finite-State Computational Morphology: An Analyzer Prototype for Zulu." *Machine Translation* 18: 195–216.

Pretorius, Laurette and Sonja E. Bosch. 2010. "Finite State Morphology of the Nguni Language Cluster: Modelling and Implementation Issues." *Lecture Notes in Computer Science* Volume 6062/2010, 123–130. Berlin, Heidelberg: Springer.

Pretorius, Laurette and Sonja E. Bosch. 2018. "ZulMorph: Finite State Morphological Analyser for Zulu (Version 20190103) [Software]." Web demo at https://portal.sadilar.org/FiniteState/demo/zulmorph/ (accessed 6 April 2020).

Prinsloo, Daniel J. and Gilles-Maurice de Schryver. 2003. "Non-word error detection in current South African spellcheckers." *Southern African Linguistics and Applied Language Studies* 21, no. 4: 307–325.

Spiegler, Sebastian, Bruno Golenia, Ksenia Shalonova, Peter Flach and Roger Tucker. 2008. "Learning the morphology of Zulu with different degrees of supervision." *IEEE Spoken Language Technology Workshop*, 9–12. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4777827&casa_token=aTKT4qc12McAAAAA:CRZszQnqhzriyfXO0DaCxgHyEYHv2M-EK0Z8--K4fN6f2PGYa9Vfvl58qguq1nAndOnRrJgD&tag=1 (accessed 6 April 2020).

Spiegler, Sebastian, Andrew van der Spuy and Peter Flach. 2010. "Ukwabelana: An Open-source Morphological Zulu Corpus." In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, 1020–1028. https://www.aclweb.org/anthology/C10-1115.pdf (accessed 9 April 2020).

Taljaard, P. C. and S. E. Bosch. 1993. *Handbook of isiZulu*. Pretoria: J.L. van Schaik.

Wortschatz Universität Leipzig. 2020. https://corpora.uni-leipzig.de/de?corpusId=zul_mixed_2016 (accessed 26 October 2019).

APPENDIX 1: EXAMPLES OF OUTPUT

1. Example of the output of isiZulu.net

Potential Decompositions (verbatim)

umuntu/abantu n. 1/2 (-ntu)

person; human; human being

kwakungumuntu ← *kwaku + ngu + umuntu*
Noun with identifying prefix (Remote Past Continuous) [←umuntu (n.)]
SC: kwaku- (cl. 15, cl. 17)
ID: ngu-

it was (a/the) person

Potential Decompositions (verbatim)

okhulumela [okʰuluˈmɛːla] ← *o + khulumela*
Present Tense [←khulumela (v/t.)]
RC: o- (cl. 1, cl. 1a, cl. 3)

that/which/who speaks for

okhulumela [okʰuluˈmɛːla] ← *a + zo + khulumela*
Future tense [←khulumela (v/t.)]
SC: a- (cl. 6)

they will speak for

it will speak for

2. Example of the output of ZulMorph

kwakungumuntu

- kwa[SCPT][15]be[AuxVStem]ku[SC][15]ngu[CopPre]u[NPrePre][1]mu[BPre][1]ntu[NStem][1-2]

okhulumela

- o[RC][1]khulum[VRoot]el[ApplExt]a[VT]
- o[RC][2ps]khulum[VRoot]el[ApplExt]a[VT]
- o[RC][3]khulum[VRoot]el[ApplExt]a[VT]
- a[SC][6]yo[Fut]khulum[VRoot]el[ApplExt]a[VT

3. Example of the output of the NCHLT IsiZulu Morphological Decomposer

kwakungumuntu kwaku-ng-umuntu okhulumela o-khulum-el-a

4. Example of the output of CHIPMUNK

Original	Segmenter	Root	Stem
kwakungumuntu	kw:INFL|PREFIX|SEGMENT a:INFL|PREFIX|SEGMENT ku:DERI|PREFIX|SEGMENT ng:DERI|SUFFIX|SEGMENT u:INFL|PREFIX|SEGMENT mu:DERI|SUFFIX|SEGMENT ntu:ROOT|SEGMENT	ntu	kungumuntu
okhulumela	o:DERI|SUFFIX|SEGMENT khulum:ROOT|SEGMENT el:DERI|SUFFIX|SEGMENT a:DERI|SUFFIX|SEGMENT	khulum	okhulumela

APPENDIX 2: TOKENS IN EVALUATION DATASET

MOST FREQUENT	MEDIAN	RANDOM
ukuthi	phakathi	ukuyiphuthuma
uma	wakhe	*kananasi
kodwa	*ku	oluzitholayo
abantu	bonke	nebuenos
futhi	ujehova	ethani
kusho	ngenxa	yikhondomu
ukuba	inkosi	ayedinga
kanye	ngokuthi	*nomabokela
noma	zonke	eningakwenzanga
uthe	abanye	*ulidzwinyu
khona	okhulumela	eyayikhombisa
ngoba	mina	*yasepulaneng
lokhu	lokho	*kwagold
kakhulu	yini	kwabazoveza
lapho	konke	ucassius
ngesikhathi	phansi	eyokubopha
wathi	isikhathi	ozibonileyo
lo	*ka	ngabafundisayo
uthi	lakhe	bengayidlanga
njengoba	labo	*ababetshona

*invalid Zulu word forms

APPENDIX 3: EXCERPT OF EVALUATION

EXAMPLE 1

ababetshona	no. of correct analyses: 0			no. of possible analyses: 1				
	found	correct	wrong	true pos	false pos	true neg	false neg	accuracy
ZulMorph	0	0	0	0	0	1	0	100.0%
isiZulu.net	0	0	0	0	0	1	0	100.0%
NCHLT	1	0	1	0	1	1	0	50.0%
Chipmunk	1	0	1	0	1	1	0	50.0%

EXAMPLE 2

ngenxa	no. of correct analyses: 2			no. of possible analyses: 4				
	found	correct	wrong	true pos	false pos	true neg	false neg	accuracy
ZulMorph	4	2	2	4	2	0	1	57.1%
isiZulu.net	1	1	0	1	0	0	2	33.3%
NCHLT	1	1	0	1	0	0	1	50.0%
Chipmunk	1	0	1	0	1	0	2	0.0%

EXAMPLE 3

eyayikhombisa	no. of correct analyses: 4			no. of possible analyses: 8				
	found	correct	wrong	true pos	false pos	true neg	false neg	accuracy
ZulMorph	8	4	4	4	4	0	0	50.0%
isiZulu.net	4	4	0	4	0	0	0	100.0%
NCHLT	1	1	0	1	0	0	2	33.3%
Chipmunk	1	0	1	0	1	0	2	0.0%

EXAMPLE 4

okhulumela	no. of correct analyses: 4			no. of possible analyses: 4				
	found	correct	wrong	true pos	false pos	true neg	false neg	accuracy
ZulMorph	4	4	0	4	0	0	0	100.0%
isiZulu.net	4	4	0	4	0	0	0	100.0%
NCHLT	1	1	0	1	0	0	0	100.0%
Chipmunk	1	1	0	1	0	0	0	100.0%

EXAMPLE 5

lo	no. of correct analyses: 3			no. of possible analyses: 3				
	found	correct	wrong	true pos	false pos	true neg	false neg	accuracy
ZulMorph	3	3	0	3	0	0	1	75.0%
isiZulu.net	3	3	0	3	0	0	1	75.0%
NCHLT	1	1	0	1	0	0	0	100.0%
Chipmunk	1	1	0	1	0	0	0	100.0%

EXAMPLE 6

khona	no. of correct analyses: 5			no. of possible analyses: 5				
	found	correct	wrong	true pos	false pos	true neg	false neg	accuracy
ZulMorph	3	3	0	3	0	0	2	60.0%
isiZulu.net	5	5	0	5	0	0	0	100.0%
NCHLT	2	1	1	1	1	0	0	33.3%
Chipmunk	1	1	0	1	0	0	1	50.0%

EXAMPLE 7

inkosi	no. of correct analyses: 1			no. of possible analyses: 1				
	found	correct	wrong	true pos	false pos	true neg	false neg	accuracy
ZulMorph	1	1	0	1	0	0	0	100.0%
isiZulu.net	1	1	0	1	0	0	0	100.0%
NCHLT	1	0	1	0	1	0	1	0.0%
Chipmunk	1	0	1	0	1	0	1	0.0%

List of tags

ADJP	adjective prefix
APPLSUF	applied suffix
FUT	future
FV	final vowel
OP	object prefix
PRES	present
PROG	progressive
SP	subject prefix
1 ... 15	noun classes 1 to 15