Towards Zulu corpus clean-up, lexicon development and corpus annotation by means of computational morphological analysis

Sonja Bosch

Department of African Languages, University of South Africa
P.O. Box 392, UNISA 0003, Pretoria, South Africa
boschse@unisa.ac.za

Laurette Pretorius

School of Graduate Studies, University of South Africa
P.O. Box 392, UNISA 0003, Pretoria, South Africa
pretol@unisa.ac.za

This article reports on a practical, semi-automated procedure towards creating a clean, morphologically annotated Zulu corpus of tractable size that could eventually serve both as a gold standard for Zulu computational morphology and as basis for further linguistic annotation. A corpus development architecture is proposed which includes the corpus in various stages of development, a pre-processing module, the Zulu morphological analyser and its guesser variant, the machine-readable lexicon that serves as comprehensive lexical database for Zulu, and a human elicitation function for ensuring the integrity of the lexical database. The approach is novel in the sense that an existing rule-based, finite-state Zulu computational morphological analyser is used as a core technology in this procedure to facilitate the complex, agglutinative nature of Zulu morphology. The corpus, at present consisting of the Zulu version of the South African Constitution, will have morphological analysis and tagging as a first level of annotation.

Introduction

Technological advances towards the end of the previous millennium and an increasing interest in the study of language use as opposed to language systems *in abstracto* (see for example, Johansson, 2008) resulted in corpus linguistics and electronic corpora increasing in relevance and importance in language studies and linguistics research. For a language such as English, much has been done in terms of electronic corpus design, development and linguistic annotation, the availability of which has enabled extensive corpus-based linguistics research. For the purposes of this article the term 'corpus-based' is interpreted in the broad sense of McEnery, Xiao and Tono (2006:11), that is, no distinction is made between so-called corpus-based and corpus-driven approaches.

For Zulu and various other South African Bantu languages, corpus building projects have resulted in, among others, general corpora for all eleven official South African languages compiled at the University of Pretoria (Language Corpora compiled at the University of Pretoria, 2008), etc. The authors are, however, not aware of any reports on extensive clean-up or annotation of (a fragment of) the abovementioned corpora or any other Zulu corpus.

This article reports on a practical, semi-automated procedure for creating a clean, morphologically annotated Zulu corpus of tractable size that could eventually serve both as a gold standard for Zulu computational morphology and as basis for further linguistic annotation. The approach is novel in the sense that a Zulu computational morphological analyser is used as a core technology in this procedure due to the complex, agglutinative nature of Zulu morphology. The corpus of tractable size, at present consisting of the Zulu version of the South African Constitution (The Constitution, (sa)), will, after the successful completion of the above-mentioned process, have morphological analysis and tagging as a first level of annotation. This work serves primarily as a proof of concept, but the expectation is also that this procedure will be sufficiently general so as to be applicable to extended Zulu corpora[1].

The proposed corpus development architecture, underlying the above-mentioned procedure, includes the corpus in various stages of development, a pre-processing module, the Zulu morphological analyser and its guesser variant, the machine-readable lexicon that serves as comprehensive lexical database for Zulu, and a human elicitation function for ensuring the integrity of the lexical database. Of specific interest is the role that **ZulMorph**, an existing rule-based, finite-state computational morphological analyser, may play in Zulu corpus annotation and the creation of a gold standard for Zulu computational morphology.

More specifically, we discuss

- the necessary and customary pre-processing, including tokenization of the corpus;
- the application of **ZulMorph** for the purposes of semi-automated corpus clean-up, i.e. the identification of non-words, and the subsequent processing of such non-words;
- the application of the guesser variant of **ZulMorph** for the purposes of enriching the embedded word root/stem lexicons[2] of **ZulMorph** to include all word roots, stems and named entities that occur in the corpus;
- the application of **ZulMorph** for enhancing morphological coverage; and
- a gold standard for Zulu computational morphology and the morphological annotation of the Zulu tokens of the corpus.

The remainder of the paper is organized as follows: The next section introduces the proposed corpus development architecture. This is followed by a section on corpus selection and the necessary pre-processing, which includes corpus normalization, tokenization and appropriate forms of clean-up. In the next section the focus moves to **ZulMorph**, a finite-state morphological analyser prototype for Zulu, followed by a section on its word root/stem guesser variant. The main contribution of the article may be found in the next section. It discusses in some detail results obtained by the systematic application of **ZulMorph** and its guesser variant to the cleaned-up Zulu corpus with the purpose of mining it (in a semi-automated way) for new linguistic and lexical information. Typical kinds of new information are new word roots and stems, new named entities, as well as non-rule based behaviour in Zulu morphology as present in the corpus. Such information is then subjected to human elicitation before inclusion in the embedded word root/stem lexicons of **ZulMorph**. This lexicon development is of core significance for the automated annotation of the corpus and the semi-automated development of the gold standard since morphological analysers such as **ZulMorph** only analyse valid Zulu words, the roots and stems of which are present in its embedded word root/stem lexicon. The penultimate section is devoted to a brief discussion of the final application of the enhanced **ZulMorph** to the clean Zulu corpus to obtain a first version of a gold standard for Zulu morphology and a morphologically annotated Zulu corpus. A sample annotated sentence from the corpus is provided as illustration of the results obtained. Finally we provide some conclusions and ideas for future work.

Corpus development architecture

It is well-known that natural language data or resources in electronic form constitute key components in contemporary language and linguistics research for any language. Two kinds of resource are of specific relevance, viz. electronic corpora (representing language use) and machine-readable lexicons (representing semi-structured lexical information). Moreover, for morphologically complex languages such as Zulu, morphological analysers are enabling technologies in the sense that without such components no serious natural language processing is possible. The corpus development architecture in Figure 1 represents a semi-automated approach by which available rudimentary ‘entry-level’ components may be combined to bootstrap and enhance existing technologies, improve existing lexicons, and clean-up and annotate available corpora. For resource-scarce, lesser-studied languages the development of such approaches is of particular importance.

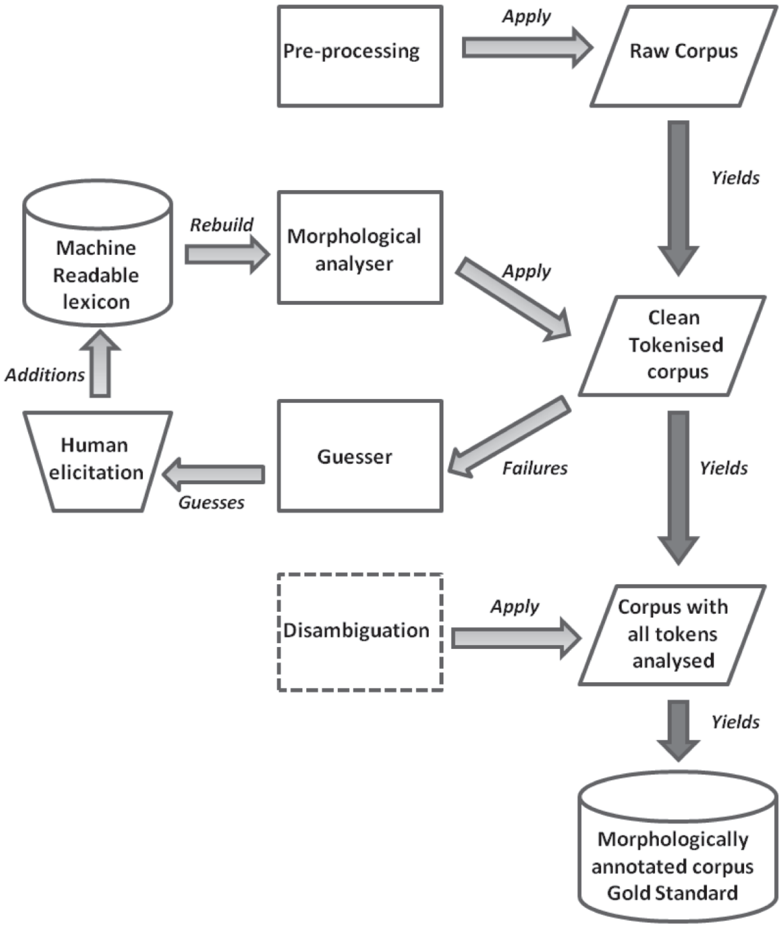
Each component in Figure 1 together with the role that it plays in the proposed corpus development and annotation approach is briefly described here and in more detail in subsequent sections.

Raw or unprocessed corpora in electronic format come in a variety of character and text encodings and typing conventions, which may vary and differ in a multitude of ways, depending on their origin, their age, their original purpose, and so forth. On the other hand morphological analysers and other natural language processing tools and applications usually assume standardized input and standardized output, which is usually decided upon and fixed at design and development time. In the present context *pre-processing* in the form of normalization, tokenization and clean-up is performed in order to prepare a corpus for the successful application of the morphological analyser. In other words, the pre-processing aims at transforming the corpus in accordance with the standard input of the morphological analyser *prior* to the application of the morphological analyser.

A finite-state *morphological analyser* is a (rule-based) model or representation of the word formation processes of a language (i.e. a lexical grammar). In other words, it basically implements the complete and accurate morphotactics and the morphophonological alternation rules of a language by means of finite-state transducers. A broad coverage morphological analyser also contains an extensive word root/stem (in our case we use the terms 'noun stem' and 'verb root') lexicon. Only words, the roots/stems of which are included in this lexicon, can be successfully analysed.

For processing word roots/stems in a corpus that do not yet occur in the embedded lexicon of the morphological analyser, the analyser can be enhanced with a *guesser* variant, which exploits the already existing morphological structure and rules built into the morphological analyser, for 'guessing' new word roots/stems. In the guesser variant of the morphological analyser the embedded word root/stem lexicon is replaced by phonologically possible word roots/stems, written in the form of regular expressions.

Figure 1: The corpus development architecture showing the components and their interrelationships



The *machine-readable lexicon* serves as an all-inclusive, state of the art lexical database and is a core component of any corpus development architecture. Indeed, the intensive use of corpora as a basis for lexicon development (Heid, 2008) has been identified as contemporary best practices in lexicography. For Zulu a machine-readable lexicon, based on a Bantu languages data model (Bosch, Pretorius & Jones, 2007), has been developed to serve as an appropriately structured repository for all the available lexical information, idiosyncratic or otherwise.

Disambiguation is the process by which one analysis among multiple correct analyses is identified as the contextually valid one, which should serve as the morphological annotation of the particular token in the particular position (context) in the corpus. In the present architecture this is (still) a manual process of human elicitation, but the development of a tool for automated disambiguation forms part of future work.

Having introduced the different components in the architecture, it remains to explain how they may be combined to eventually obtain a morphologically annotated version of the corpus under discussion. Once the target corpus has been appropriately standardized in the pre-processing stage, the morphological analyser is applied to all the resulting tokens in the corpus. Tokens that are analysed are considered successes and accepted as valid Zulu words, while unanalysed words represent failures and require further processing. The guesser is then applied to the failures and the resulting output is, in turn, presented to the human lexicographer/linguist in the form of candidate word roots/stems. By means of this (human) linguistic evaluation, word roots/stems are identified for addition to the machine-readable lexicon and subsequently to the embedded word root/stem lexicon of the morphological analyser. Once all the tokens are recognized by the morphological analyser as valid Zulu words the final stage in the architecture in Figure 1 is that of disambiguation by which each token in the corpus is annotated with a single morphological analysis, based on its syntactic context. The process terminates with all the tokens being morphologically analysed and annotated with one correct morphological analysis.

Corpus selection and pre-processing

This work concerns a proof of concept of a general procedure for corpus clean-up and annotation. Therefore it made good sense to select the South African Constitution for the following reasons: the text is of strategic importance, it is expected to employ modern Zulu and possibly be rich in terminology, it is expected to be written according to the official Zulu orthography, and to employ standardized Zulu terminology. Furthermore, it is of tractable size and forms part of a parallel corpus.

On the one hand pre-processing is about choices regarding general standards to enhance portability and re-usability of the chosen corpus and, on the other hand, about actually performing normalization of the text according to such standards, performing segmentation of the text into tokens that are amenable to morphological analysis (in our case specifically) and clean-up. Other levels of segmentation, for example into sentences, may be required for next levels of processing, including disambiguation and syntactic parsing. This falls outside the scope of this article and forms part of future work.

Standardization concerns the awareness and adoption of and adherence to standards for implementation, standards for annotation and standards for metadata. As far as implementation, and more specifically data and rendering format goes, we use XML for text encoding and Unicode for character encoding (Lehmberg & Wörner, 2008:485; Jones, Bosch, Pretorius & Prinsloo, 2005). Linguistic annotation covers any descriptive or analytic notations applied to raw language data. At present the ISO/TC 37/SC4 Linguistic Annotation Framework (LAF) allows for any type of user annotation provided that it is automatically transferable to and from the LAF Dump Format (Lehmberg & Wörner, 2008:491). Since standards for linguistic annotation are still under development, it suffices to note that our representation of Zulu morphological analysis (see subsequent sections) satisfies this LAF requirement. We developed a tag set that consists of intuitive mnemonic tag names and is ideally suited to accurately annotating the morphological structure of Zulu (see Appendix for specific examples). Our basic approach is to remain standard-aware. While metadata (Lehmberg & Wörner, 2008:493) is important in contexts where large numbers of documents are stored, accessed and searched it is not directly relevant in the present context, but will be considered at an appropriate time in the future.

Normalization: Our basic aim is to transform the text into Unicode and to standardize punctuation and other non-alphabetic symbols in the text for the purposes of automated tokenization (Schmid, 2008:532–533). The `iconv` utility was used for codeset conversion (see, for example, The IEEE and The Open Group, 2004). In particular,

```
iconv -t UTF-8 source_text > source_text_utf8
```

converts any given file (e.g. `source_text`) to (`-t`) Unicode UTF-8 and outputs it to another file (e.g. `source_text_utf8`). UTF-8 (8-bit Unicode Transformation Format) has become a dominant character encoding for files, e-mail, web pages, and software that manipulates textual information, largely due to its backward-compatibility with ASCII.

The non-alphabetic symbols that were found in our target text are given in Table 1. The punctuation symbols '?' or '!' did not occur in the text, the circumflex accent occurrences were removed as typing errors, and the hyphen was considered as part of the alphabet and was not treated as punctuation. Inconsistencies in the use of the hyphen and the en dash were manually corrected. Two non-Zulu alphabetic characters, viz. 'ë' and 'š' occurred in the text as part of the national anthem, but were not removed at this stage. The reverse solidus was identified as a typographic error and replaced by the solidus.

Table 1: Non-alphabetic symbols in the Constitution

Character	Unicode (in hexadecimal)	Description	Punctuation
(	0028	LEFT PARENTHESIS	✓
)	0029	RIGHT PARENTHESIS	✓
,	002C	COMMA	✓
-	002D	HYPHEN – MINUS	✗
.	002E	FULL STOP	✓
/	002F	SOLIDUS	✓
:	003A	COLON	✓
;	003B	SEMICOLON	✓
\	005C	REVERSE SOLIDUS	✓
^	005E	CIRCUMFLEX ACCENT	✗
‘	2018	LEFT SINGLE QUOTATION MARK	✓
’	2019	RIGHT SINGLE QUOTATION MARK	✓
“	201C	LEFT DOUBLE QUOTATION MARK	✓
”	201D	RIGHT DOUBLE QUOTATION MARK	✓
–	2013	EN DASH	✓

Tokenization, the segmentation of text into linguistic words, is necessary for the purposes of morphological annotation. For parsing as a next level of processing the text is further segmented into sentences. Sentence boundaries are usually indicated by '.', '?' or '!'. Since our focus is on morphological analysis and morphological annotation, sentence boundaries are mainly interesting in as far as they also indicate token boundaries. Tokens are words, numbers, punctuation marks, parentheses, quotation marks, and similar entities. In alphabetic languages such as English and Zulu words are usually separated by whitespace and optionally preceded and followed by the mentioned entities. Therefore, a simple tokenizer that replaces whitespace with word boundaries and cuts off punctuation marks, parentheses, and quotation marks at both ends of a word is already quite accurate. General tokenization issues that need to be resolved in the context of morphological analysis are the following (Schmid, 2008:529–536):

- Periods: Abbreviations versus sentence-final punctuation (full stops)
- Ordinal numbers
- Multi-word expressions
- Clitics
- Word-internal punctuation
- De-hyphenation
- Missing whitespace
- De-capitalization.

The basic approach was to tokenize on whitespace and punctuation, as alluded to before. As first step, contiguous whitespace (`\s+` in Perl, + denoting the Kleene plus in Perl regular expressions) was replaced by a newline marker (`\n` in Perl) and each `psymbol` in the (UTF-8) source text, shown as punctuation in Table 1, was replaced by `\npsymbol\n`. Next, blank lines were deleted from the resulting text, yielding the list of tokens, one per line.

Regarding other tokenization issues in the bulleted list above, the text of the South African Constitution contains various abbreviations, but they occur without periods (for example, *ANC)*; numbers do occur in the form of page numbers, years and section numbers and are not discussed further as part of the morphological analysis; multi-word expressions were not considered; clitics do not occur except in the national anthem (for example, *Sikelel' iAfrika*); word-internal punctuation does not occur; de-hyphenation does not occur; and missing whitespace is discussed in the subsequent section on clean-up. De-capitalization was done in all cases except word-internal capitals, which denote proper names endowed with morphological structure as is customary in Zulu. The stems of proper names form part of the embedded lexicon of **ZulMorph** and are an important part of the open word classes of **ZulMorph** that require regular updating and maintenance.

A finite-state morphological analyser may be thought of as a rudimentary spelling checker in the following sense: As will be discussed in a subsequent section, such a morphological analyser is a bidirectional finite-state transducer, which analyses surface forms in the one direction and generates surface forms from their given morphological analyses in the other direction. In the analysis direction the transducer acts as a surface form acceptor for correctly spelt words while surface forms/words that are not correctly spelt are not accepted.

Let us consider the following statistics regarding the performance of **ZulMorph** on the (raw, tokenized) corpus under discussion if we ignore all tokens that contain any non-alphabetic symbols (i.e. not amenable to basic spelling checking) and ignore all capitalization (i.e. we do not consider capitalization as a basic spelling issue):

Total number of words:	28245
Analysed (correctly spelt):	23965 (84.85 %)
Not analysed:	4280 (15.15 %)
Total number of word types:	7044
Analysed (correctly spelt):	5877 (83.43 %)
Not analysed:	1167 (16.57 %)

An interpretation of this result is that there are approximately 1200 word types in the corpus that require some form of clean-up in order to transform the corpus into a gold standard for Zulu computational morphological analysis. In terms of spelling checking this result is promising and compares well with a result of Reynaert (2006) who reports that over 21% of the word types in the Reuters Corpus Volume 1 are typographical errors. It also demonstrates the potential of **ZulMorph** as a rudimentary spelling checker.

Clean-up: We consider clean-up as the identification and appropriate processing of non-words, or equivalently, non-attested words in the Zulu language. The obvious question that arises is: What kinds of systematic clean-up are necessary, i.e. what kinds of errors occur, and how can this clean-up be performed? From the point of the application of **ZulMorph** and its guesser variant we distinguish between errors in words, the roots and stems of which are already present in the embedded lexicon of **ZulMorph**, and errors pertaining to the absence of appropriate (i.e. as yet not included) word roots and stems in **ZulMorph**. The first kind of error is discussed under clean-up (below) and the second under lexicon development (see a subsequent section).

It should also be noted that, basically, de-capitalization should not be performed on proper name stems (as was done in the basic spelling checking above), i.e. the corpus should contain only valid proper names with correct capitalization. Moreover, not only misspelt words should be corrected, but orthographic discrepancies (conjunctive instead of disjunctive writing) should be catered for. Table 2 contains typical corrective actions for errors that were identified (semi-automatically) via the human elicitation of the **ZulMorph** failures when applied to the raw, tokenized corpus.

Table 2: Typical corrective actions as obtained from **ZulMorph** failures via human elicitation

Number	Corrective action	Examples from corpus
1	Remove capitalization	Beginning of sentence markers
2	Remove abbreviations	ANC, IFP
3	Remove foreign words	contents, Africa, Afrika, bless
4	Introduce whitespace (to correct orthography, etc.)	See numerous examples below
5	Correct typographical errors	*zezeifundazwe* > *zezifundazwe*
6	Add a word root or stem to the embedded lexicon	*-khomishane* (noun stem) *kambe* (conjunctive)[3] *-vot-* (verb root)

Examples of corrective action 4, taken from the South African Constitution corpus:

- *na-* or *ku-* plus absolute pronouns is written disjunctively from the following quantitative pronoun:

(1) *… othole amavoti amancane* ***kunabobonke*** *…>* ***kunabo bonke***
‘… who receives the lowest number of votes …’
… kubobonke … > … kubo bonke …
‘… to all …’

- Demonstratives

Whereas demonstratives were traditionally written conjunctively with the following nouns, the *IsiZulu Terminology and Orthography No. 4* (1993) prescribes a disjunctive approach:

(2) *kuleyonkantolo* > *kuleyo nkantolo* ‘at that office’
kulezizigaba > *kulezi zigaba* ‘in these sections’
kulowomuntu > *kulowo muntu* ‘to that person’
lelolungu > *lelo lungu* ‘that member’
lesisigaba > *lesi sigaba* ‘this section’
lesisikhathi > *lesi sikhathi* ‘this time’
lezozakhiwo > *lezo zakhiwo* ‘those institutions’

- Copulatives

Copulatives that take on a compound formation consisting of the auxiliary verb stem *-ba* followed by a complement (Poulos & Msimang, 1996:364), are also required to be written disjunctively:

(3) *libekhona* > *libe khona* 'it must be present'
libenomphumela > *libe nomphumela* 'it has effect'

The clean-up process was concluded by performing 803 corrective actions 1-5 on all identified instances in the corpus. The resulting improvement in the performance of **ZulMorph** was as follows:

Total number of types:	6890
Analysed:	5891 (85.50%)
Not analysed:	999 (14.50%)

Corrective action 6 is discussed in the section entitled **Enriching the lexicon of *ZulMorph***.

Zulu computational morphological analysis

Although large coverage morphological analysers exist for various languages of the world, computational morphology to this day remains a challenge for most Bantu languages. The Bantu language that is probably most well-known for its extensive technological development is Swahili (Hurskainen, 1992 & 1997), but computational morphological analysis has also been reported on for a number of other Bantu languages such as Zulu (Pretorius & Bosch, 2003) and three other languages belonging to the Nguni group (Bosch, Pretorius & Fleisch, 2008; Pretorius & Bosch, 2010), as well as for Tswana (Pretorius, Viljoen & Pretorius, 2005), Northern Sotho (Hurskainen, Louwrens & Poulos, 2005) and Kinyarwanda (Muhirwe, 2007).

The finite-state Zulu morphological analyser **ZulMorph** is rule-based and represents an accurate and comprehensive linguistic representation of Zulu morphology, based on a body of available linguistic resources, including grammar texts, paper dictionaries and electronically available corpora.

The suitability of finite-state approaches to computational morphology has been shown convincingly (Koskenniemi, 1997; Karttunen, 2003; Beesley & Karttunen, 2003) and has resulted in numerous software toolkits and development environments for this purpose. For the work reported on in this paper the Xerox finite-state toolkit (Beesley & Karttunen, 2003) is used. A new and promising development is the open source Foma toolkit (Hulden, 2009) that is largely compatible with the Xerox finite-state toolkit. A first experiment confirmed the seamless compilation and execution of our **lexc** and **xfst** code with Foma, which is free software and available under the GNU General Public License.

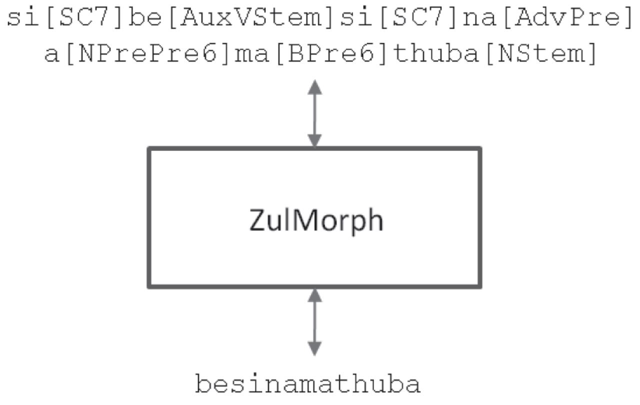
The Xerox software tool for modelling the morphotactics is **lexc** (**lex**icon **c**ompiler). An accurate specification of the Zulu word structure is created as a **lexc** script file and compiled into a so-called finite-state network. The morphotactics component includes all and only word roots/stems in the language, all and only the affixes for all parts of speech (word categories) as well as a complete description of the valid combinations and orders of these morphemes for forming all and only the words of Zulu. The words generated by this network are morphotactically well-formed, but still rather abstract lexical or morphophonemic words.

The morphophonological (phonological and orthographical) alternations, i.e. the changes that take place between the lexical and surface words when certain morphemes are combined, are modelled with the Xerox regular expression language. These regular expressions are then compiled into a finite-state network by means of the **xfst** tool. Finally, the two mentioned finite-state networks are combined (composed) together into a single network, a so-called lexical transducer, which constitutes the **ZulMorph** morphological analyser. It is note-worthy that this finite-state network (transducer) is a bi-directional device, which facilitates morphological analysis in the one direction and morphological generation in the other, as illustrated in Figure 2. It remains a challenge to build such lexical transducers that analyse and generate all and only the words of a given language, in this case Zulu (cf. Pretorius & Bosch, 2003; Bosch & Pretorius, 2006).

We distinguish between closed and open classes of morphemes. The closed classes cover all affixes that model the fixed morphological structure of words, (e.g. subject and object agreement morphemes, class prefixes, verb extensions etc.) as well as items such as pronouns. Typically no new items can be added to the closed class (Fromkin, Rodman & Hyams, 2007:74).

The open classes accept the addition of word roots/stems including verb roots and noun stems. The addition of new items takes place by means of processes such as borrowing, coining, compounding and derivation. Word roots/stems currently include nouns and their class information, verbs, relatives, adjectives, ideophones and conjunctions. The word root/stem list is based on a printed Zulu dictionary (Doke & Vilakazi, 1964) of which the last revised edition dates back to the 1950s, and contains a total of over 27 000 lemmas. We often refer to the word root/stem list as the embedded word root/stem lexicon of the morphological analyser.

Figure 2: Example of the bi-directionality of **ZulMorph** – see Appendix for an explanation of the tags


si[SC7]be[AuxVStem]si[SC7]na[AdvPre]
a[NPrePre6]ma[BPre6]thuba[NStem]
ZulMorph
besinamathuba

The components of **ZulMorph**, including its scope in terms of word categories and their morphological structure, as well as its lexical coverage as reflected by the number of different noun stems, verb roots etc. are summarized in Table 3 below.

Table 3: Zulu morphological analyser components

Morphotactics (**lexc**)	Affixes for all parts of speech (e.g. subject & object agreement morphemes, noun class prefixes, verb extensions etc.); pronouns	Word roots/stems e.g. nouns (15 825), verbs (7597), relatives (408), adjectives (48), ideophones (2583), conjunctions (176)	Rules for legal combinations and orders of morphemes (e.g. *u-ya-ngi-thand-a* and not **ya-u-a-thand-ngi*)
Morphophonological alternations (**xfst**)	Rules that determine the form of each morpheme (e.g. *ku-hamb-w-a* > *ku-hanj-w-a*, *u-mu-lilo* > *u-m-lilo*)		

In general terms the quality of a rule-based finite-state morphological analyser for a particular language is determined by two considerations, namely the care that has been taken in accurately modelling the morphology of the language and the comprehensiveness of the embedded word root/stem lexicon. Conceptually the first aspect is a one-off process, while the second aspect needs constant attention to cater for the dynamic nature of human language. It is in this latter aspect that corpora play an increasingly prominent role.

The *ZulMorph* word root/stem guesser

In this section corrective action pertaining to tokens in the corpus that are valid Zulu words, but are not analysed by **ZulMorph** is addressed. There are two possibilities: either the word root or stem of the unanalysed token is not present in the embedded lexicon of **ZulMorph** (see section entitled **Enriching the lexicon of *ZulMorph***) or the morphological structure of the particular word has not yet been included in **ZulMorph** (see section entitled **Enhancing the morphological coverage of *ZulMorph***).

In the Xerox finite-state toolkit the morphotactics as well as the word root/stem lists are usually described as cascades of lexicons using **lexc** scripts. In principle, the morphological analyser and the guesser share a common **lexc** script in which a placeholder entry in each open word class, viz. `^GUESSNOUNSTEM` and `^GUESSVERBROOT` are inserted. These placeholders are multi-character symbols. In addition to distinguishing between noun stems and verb roots, it is useful and necessary to capture information regarding noun classes as well as the different kinds of verb roots, viz. vowel verb roots, monosyllabic and polysyllabic verb roots.

The following code models phonologically possible noun stems occurring in different noun classes and different kinds of verb roots. Noun stems are guessed together with their noun class information. The placeholders are associated with noun classes and types of verb roots in the same way as attested noun stems and verb roots:

```
LEXICON NounStem
^GUESSNOUNSTEM NClass1-2;
...
^GUESSNOUNSTEM NClass14;
...
ntu   NClass1-2;
ntu   NClass14;
...

LEXICON VerbRoot
^GUESSVERBROOT VVRClass15;
^GUESSVERBROOT VPSClass15;
^GUESSVERBROOT VMSClass15;
...
fund VPSClass15;
```

The **lexc** script is compiled in the usual way into a finite-state network in which the shown placeholders occur. The next step is to define as accurately as possible all phonologically possible stems and roots in Zulu by means of **xfst** regular expressions. These regular expressions are then compiled into finite-state networks and *substituted* for the placeholder symbols in the compiled **lexc** script. In order to distinguish guessed stems and roots from known listed ones the tag “`[Guess]`” is included to appear in the analysis language.

The following regular expressions are used in our experiments for representing the phonologically possible Zulu stems and roots:

```
define VL a|e|o|i|u;
define CL b|c|d|f|g|h|j|k|l|m|n|p|q|r|s|t|v|w|x|y|z;
define CU B|C|D|F|G|H|J|K|L|M|N|P|Q|R|S|T|V|W|X|Y|Z;
define VerbRoot [CL (CL (CL)) VL]+ CL (CL (CL)) “[Guess]”:0;
define NounStem [CL|CU] (CL (CL)) VL [CL (CL (CL)) VL]+ “[Guess]”:0;
```

The '+' denotes the Kleene plus (i.e. one or more occurrences) and '(...)' denotes optionality. Therefore, the simplest verb root allowed by the given regular expression is of the form CVC, the simplest noun stem has the form CVCV and the simplest proper name stem has the form CVCV where the first consonant is in upper case. Clearly, the form CVC excludes vowel verb roots and monosyllabic verb roots – such verb roots will therefore not be guessed.

The basic idea is to look up a given word in the morphological analyser network first. If and only if it is not found, it is passed to the guesser network. The output of the guesser is then subjected to human elicitation after which valid noun stems and/or verb roots are added to the embedded word root/stem lexicon of **ZulMorph**.

Enriching the lexicon of *ZulMorph*

While the lexicon of a language (as mainly reflected by the open classes) is dynamic and may evolve rapidly with time, the morphological structure of a language remains stable although changes in the use of certain morphemes may occur over time. However, when modelling a language, and in particular its morphology, both these processes have to be taken into consideration. The application of a finite-state morphological analyser, as a representation of the attested morphological structure (as, for instance, in grammar textbooks) and the lexicon of a language, to a corpus provides a useful tool for identifying not only newly coined word roots and stems, but also new tendencies in morphological constructions. It should be emphasized that the concept of 'newly coined' presupposes an existing collection of word roots/stems as a basis of comparison – in the case of **ZulMorph** this basis (its embedded lexicon) is the authoritative, but dated paper dictionary by Doke and Vilakazi (1964). In this section the focus is on enriching the embedded lexicon from the corpus under discussion. In the next section the identification of changes in the use of morphological constructions, and specifically the locative prefixes *ku-* and *kwi-* as they occur in the corpus, are addressed.

The application of **ZulMorph** and its guesser variant to the corpus in order to enrich the **ZulMorph** embedded lexicon is illustrated by the examples, together with their guessed (human elicited) analyses, given in Table 4.

Table 4: Examples of guessed analyses

```
ngokuphindaphindwa

nga[AdvPre]u[NPrePre15]ku[BPre15]phindaphind[Guess][VRoot]w[PassExt]
a[VerbTerm]

ikhophi i[NPrePre5]li[BPre5]khophi[Guess][NStem]

eMpumalanga e[LocPre]i[NPrePre9]n[BPre9]Mpumalanga[NStem]

isiHindi i[NPrePre7]si[BPre7]Hindi[Guess][NStem]

isiphathimandla i[NPrePre7]si[BPre7]phathimandla[NStem]

izinkampani i[NPrePre10]zin[BPre10]kampani[Guess][NStem]

lokubhikisha
la[PossConc5]u[NPrePre15]ku[BPre15]bhikish[Guess][VRoot]a[VerbTerm]

noguquko na[AdvPre]u[NPrePre11]lu[BPre11]guquko[Guess][NStem]

ubugqili u[NPrePre14]bu[BPre14]gqili[Guess][NStem]

ubuholi u[NPrePre14]bu[BPre14]holi[Guess][NStem]

umasipala u[NPrePre1a]masipala[Guess][NStem]

umthethosivivinywa u[NPrePre3]mu[BPre3]thethosivivinywa[Guess][NStem]

usheduli u[NPrePre1a]sheduli[Guess][NStem]
```

New stems and roots

Adopted nouns (from English) usually fit in with the linguistic structure of Zulu and appear in the nominal classes that usually accommodate adoptives, namely classes 1a/2a, 5/6 and 9/10 (Poulos & Msimang, 1998:91):

Class 1a/2a
(4) *umasipala* ‘municipality’
usheduli ‘schedule’

Class 5/6
(5) *ama-ambulense* ‘ambulances’
amabhajethi ‘budgets’
amademeshe ‘damages’
amakhasino ‘casinos’
amalevi ‘levies’
amaphesenti ‘percentages’
amasevisi ‘services’
ikhophi ‘copy’

Class 9/10
(6) *izinkampani* ‘companies’

Four orthographic variations of the translation of ‘cabinet’ were found:

(7) *-khabhinethe* ‘cabinet’
-khabhinethi ‘cabinet’
-khabinethe ‘cabinet’
-khabinethi ‘cabinet’

Since none of the above terms for ‘cabinet’ has been standardized, it is also not surprising that a stem such as *-khabhinethi* is used with either a class 5 or class 9 noun prefix, as can be deduced from the following contexts:

(8) *Ikhabhinethi kumele* ***li****sebenze* ...
‘Cabinet must act ...’
IKhabhinethi ***i****noMongameli, njengenhloko yeKhabhinethe…*
‘The Cabinet consists of the President, as head of the Cabinet …’

Names of languages are not all listed in Doke and Vilakazi (1964). In some instances they are listed according to outdated orthography. Names of languages are typically accommodated in class 7 *isi-* (singular only):

(9) *isiHindi* ‘Hindi’
isiNdebele ‘Ndebele’
isiPedi ‘Northern Sotho’
isiTamil ‘Tamil’
isiTswana ‘Tswana’
isiVenda ‘Venda’
isiTsonga ‘Tsonga’
isiXhosa[4] ‘Xhosa’

Place names in which, for the purposes of computational morphological analysis, the place name stem is regarded as the part of the place name that commences with a capital letter (Van Huyssteen & Bosch, 2008:105–112):

(10) -*Mpumalanga* ‘East’
-*Ntshonalanga* ‘West’
-*Mgungundlovu* ‘Pietermaritzburg’
-*Kapa* ‘Cape’
-*Pitoli* ‘Pretoria’

The place name stems above do not appear in Doke and Vilakazi (1964), while the following are not listed as separate entries, but rather as locative forms of nouns:

(11) *i(li)goli* ‘gold’ > loc. *eGoli* ‘Johannesburg’
i(li)theku ‘harbour’ > loc. *eThekwini* ‘Durban’
u(lu)khahlamba ‘broken mountain range’ > *u(lu)Khahlamba* ‘the Drakensberg Mountains’ > loc. *oKhahlamba*

New verb roots that are not listed in Doke and Vilakazi (1964), do occasionally appear in more recent dictionaries such as Mbatha (2006):

(12) -*vot*- ‘vote’ as in
*… kodwa azivunyelwe uku****vot****a*
‘… but may not **vote**’

-*bhikish*- ‘demonstrate’ as in
*Wonke umuntu unelungelo … loku****bhikish****a …*
‘Everyone has the right … to **demonstrate** …’

Reduplicated verb roots that usually express an action that is carried out frequently or repetitively (Poulos & Msimang, 1998:202) and are not listed as main entries in Doke and Vilakazi (1964) are:

(13) *… iqhaza lamaqembu a****hlukahluk****ene …*
‘a **multi**-party system’

The verb root that is reduplicated, is -*ahluk*- ‘differ from one another’.

(14) *… zephulwe kabi kakhulu noma ngoku****phindaphind****wa.*
‘… a serious or **persistent** material breach.’

The verb root that is reduplicated in this case, is -*phind*- ‘repeat’. -*phindaphinda* only appears as sub-entry in Doke and Vilakazi (1964:662). The automatic recognition of reduplicated verb roots forms part of future work.

Compounds are made up on the fly since compounding is a generative, productive process (Bosch & Fellbaum, 2009:39). It is therefore not surprising that numerous new compounds, not listed in Doke and Vilakazi (1964), occur in a corpus such as the South African Constitution:

Class 3/4
(15) *umthethosivivinywa*
‘bill’
umthethosisekelo
‘constitution’
umhlahlandlela
‘guideline’

Class 5/6

(16) *amalungelomvume*
amalungelomvumo
'privileges'

Class 7/8

(17) *isishayamthetho*
'legislature'
isiphathimandla
'authority'
izidingongqangi
'priorities'

Class 9/10

(18) *inqubekelaphambili*[5]
'progress, advancement'
inkulumompikiswano
'debate'

The main patterns of Zulu compounding consist of a leftmost governing head and a rightmost dependent element. Words contributing to a compound may undergo phonological and morphological changes and therefore the process of compounding in Zulu is not merely a concatenation of independent words. Parts of these words 'may be elided, replaced or adapted in some way or another, so that the lexical components do not appear as full words, but as bound stems and roots' (Kosch, 2006:122). Compounding also affects prefixation. A class prefix is added for every compound headed by a verb as in -*shaya* 'hit' + *umthetho* 'law' > *isishayamthetho* 'legislature' to which the class 7 (*isi*-) prefix has been added. The automatic identification of compounds by means of morphological analysis is therefore not a straightforward process, but is indeed on the cards for future work.

Mining of noun stems with new class prefixes

Word class-maintaining derivation of new nouns (Kosch, 2006:120): a number of 'new' nouns with prefixes in class 11/10 and class 14, formed from existing noun stems taking prefixes in other classes in the Zulu lexicon, were identified in the corpus. The following entries of the noun stem -*guquko* are found in Doke and Vilakazi (1964:278):

(19) -*guquko* (*in*-) sg. only 'change in character; change of opinion' *noguquko*
-*guquko* (*um-/imi*-) 'a changing'

In the corpus, a related semantic significance is added by the prefixation of a class 11/10 prefix to the stem -*guquko*:

(20) … *kube no***guquko** *oluzoqinisekisa*
'… there be **reform** that will ensure ...'
… *enganquma ngokusemthethweni noma yiluphi* **uguquko** *lomthethosisekelo;*
'… decide on the constitutionality of any **amendment** to the Constitution;'

… *ne***zinguquko** *ezishiwo* …
'… with **amendments** or **adaptations** referred to …'

Further examples of new nouns prefixing a class 11/10 prefix in the corpus are:

(21) *ucwaningo* 'research'
ushintsho 'amendment'
uthikamezo 'nuisance'

The prefixation of a class 14 prefix to existing noun stems such as *-gqili* and *-holi*, adds a related, though abstract semantic significance as illustrated by the following examples in the corpus:

(22) *ubugqili* 'slavery' derived from
isigqili 'slave' (Doke &Vilakazi, 1964:264)

ubuholi 'leadership' derived from
umholi 'leader' (Doke &Vilakazi, 1964:344)

Further examples of new nouns formed from existing noun stems by prefixing a class 14 prefix, in the corpus are:

(23) *ububhimbi* 'incompetence'
ubundunankulu 'premiership'
ubunhloli 'intelligence'

Mining of non-rule based behaviour

Non-rule based behaviour includes idiosyncrasies where the **rule works differently**, e.g. palatalization across syllable boundaries in the formation of passives as well as instances where there is **no rule**, e.g. irregular locative formation.

Passives causing palatalization across syllable boundaries

According to the palatalization rule caused by the passive extension *-w-,* a sound change applies to a consonant in the final root position. However, in some cases a consonant within a root is affected by a sound change, e.g.

(24) ***Ukusetshenziswa*** *komthetho wamazwe ngamazwe* … (*uku-sebenz-is-w-a*)
'Application of international law …'

in which case ***b*** **>** ***tsh.***

Irregular locative formation

In noun classes 3 to 10 the regular formation of the locative is by prefixing *e-*, followed by a locative suffix *-ini*, e.g. *esikhundleni* (*e-isi-khundla-ini*). In exceptional, morphologically unpredictable cases only the prefix *e-* occurs, e.g.

(25) *ekhaya* (*e-i(li)-khaya*)
'at home'

The machine-readable lexicon under development for Zulu does not at this stage contain exhaustive information on this type of locative derivatives, which is however, in the process of being mined from corpora.

In summary, 117 new noun stems, 66 proper names, 2 true adverbs, and 2 verb stems were added to the **ZulMorph** embedded lexicon, as well as to the machine-readable Zulu lexicon.

Enhancing the morphological coverage of *ZulMorph*

Expansion of the use of morphemes

The occurrence of the locative prefixes *ku-* and *kwi-* in the corpus represents an instance of change in use of morphemes that was largely overlooked in most (traditional) Zulu grammars on which **ZulMorph** is based.

De Schryver and Gauton (2002:218) argue that 'Locativisation by means of the class 17 prefix *ku***-** (and its variant *kwi-*) is an ascendant strategy in Zulu, being increasingly applied in the derivation of locatives from nouns from classes other than 1/2, 1a/2a and [+human] nouns in class 6, i.e. in environments where the use of the locative affixes *e-/o-...-ini* would have been expected'. This finding was confirmed by results obtained when applying **ZulMorph** to our chosen corpus. Table 5 contains some examples in the corpus of the use of locative prefixes *ku-* and *kwi-* instead of the locative affixes *e-/o-...-ini* as expected:

Table 5: Examples from the corpus: locative *kwi-* and *ku-* instead of locative affixes *e-/o-...-ini*

Locative *kwi*-:	**Locative *ku*-:**
kwinkantolo 'in the office'	*kumabhange* 'in banks'
kwiphephandaba 'in the newspaper'	*kumayunivesithi* 'at universities'
kwizinhlangano 'at the meetings'	*kumnyango* 'in the department '
kwizishayamthetho 'in the legislature'	*kusiqephu* 'in section'

This concludes the addition of new word roots or stems to the embedded lexicon (corrective action 6 in Table 2). The following results reflect the status quo of the enhanced prototype of **ZulMorph**:

Total number of types:	6890
Analysed:	6281 (91.16%)
Not analysed:	609 (8.84%)

Attested morphological constructions not yet modelled in *ZulMorph*

The complex nature of the Zulu morphology calls for continuous enhancement of the morphological coverage of **ZulMorph** by the modelling of additional constructions. Such constructions are often identified when morphological analysis still fails after corrective actions 1 to 6, as indicated in Table 2, have been carried out. Table 6 contains examples that involve for instance adverbial prefixes and relative prefixes that combine with a variety of constructions.

Table 6: Examples of morphological constructions not yet modelled in **ZulMorph**

na- (adverbial prefix)	***nga-*** (adverbial prefix)	***njenga-*** (adverbial prefix)	**Relative prefixes**
nabangaphansi	*ngalinye*	*njengelungile*	*angabakhona*
nabezindaba	*ngaluphi*	*njengezijwayelekile*	*elingelona*
nakusishayamthetho	*ngokumaqondana*	*njengobawamukelwe*	*elingemuva*
nakuyilona	*ngokungafanele*	*njengogwetshiwe*	*esingakanani*
nakwisigungu	*ngokuphelele*	*njengokulungile*	*obekungabhekekile*
nangayiphi		*njengokusho*	*okwesihlanu*
nangokobulili			
nangokobuzwe			

The expectation is that the inclusion of these morphological constructions will further improve the success rate of **ZulMorph**.

Towards morphological annotation and a gold standard for Zulu

Once the corpus has been cleaned up, all the roots/stems that occur in the corpus have been added to the root/stem lexicons and the morphological coverage has been maximized, the morphological analyser should yield morphological analyses for all the Zulu tokens in the corpus. Since the finite-state morphological analyser yields all possible analyses, many tokens will have multiple valid morphological analyses, as illustrated in the two analysed words in example (26) from the sentence in Table 7.

(26)

```
uma  uma[Conj]
uma  u[SC1](i)m[VRoot]a[VerbTerm]
uma  u[SC][2ps](i)m[VRoot]a[VerbTerm]
uma  u[SC3](i)m[VRoot]a[VerbTerm]

singaphumeleli si[SC7]nga[NegPre]phum[VRoot]elel[IntensExt]i[VerbTermNeg]
singaphumeleli si[SC][1pp]nga[NegPre]phum[VRoot]elel[IntensExt]i[VerbTermNeg]
```

For the corpus to be amenable to further linguistic analysis, disambiguation based on word context, has to be performed. In the above examples, the context determines that *uma* functions as a conjunction that introduces a clause, and not as a verb with the verb root -*(i)m*-; and the subject of *singaphumeleli* is *isicelo*, therefore the appropriate subject agreement morpheme of the verb in context is that of class 7 `[SC7]` and not of first person plural `[SC][1pp]`.

The (semi-) automation of disambiguation forms part of future work. In Table 7 we show one example of a sentence from the corpus, its English translation and the morphological analyses of the tokens constituting the sentence.

Table 7: Towards a gold standard for Zulu computational morphology: A sentence from the corpus, the tokens and their morphological annotations

Uma isicelo singaphumeleli, iNkantolo yoMthethosisekelo kufanele inqume ukuthi labo abafake isicelo kufanele bakhokhe izindleko ngaphandle uma isicelo besinamathuba anele okuphumelela.	
'If an application is unsuccessful, and did not have a reasonable prospect of success, the Constitutional Court may order the applicants to pay costs.'	
uma	uma[Conj]
isicelo	i[NPrePre7]si[BPre7]celo[NStem]
singaphumeleli	si[SC7]nga[NegPre]phum[VRoot]elel[IntensExt]i[VerbTermNeg]
inkantolo	i[NPrePre9]n[BPre9]kantolo[NStem]
yomthethosisekelo	ya[PossConc9]u[NPrePre3]mu[BPre3]thethosisekelo[NStem]
kufanele	ku[SC15]fan[VRoot]el[ApplExt]e[VerbTermPerf]
inqume	i[SubjSC9]nqum[VRoot]e[VerbTermSubj]
ukuthi	ukuthi[Conj]
labo	labo[Dem2][Pos2]
abafake	aba[RelConc2]fak[VRoot]e[VerbTermPerf]
isicelo	i[NPrePre7]si[BPre7]celo[NStem]
kufanele	ku[SC15]fan[VRoot]el[ApplExt]e[VerbTermPerf]
bakhokhe	ba[SubjSC2]khokh[VRoot]e[VerbTermSubj]
izindleko	i[NPrePre10]zin[BPre10]dleko[NStem]
ngaphandle	nga[AdvPre]phandle[Adv]
uma	uma[Conj]
isicelo	i[NPrePre7]si[BPre7]celo[NStem]
besinamathuba	si[SC7]be[AuxVStem]si[SC7]na[AdvPre]a[NPrePre6]ma[BPre6]thuba[NStem]
anele	a[SubjSC6]anel[VRoot]e[VerbTermSubj]
okuphumelela	a[PossConc6]u[NPrePre15]ku[BPre15]phum[VRoot] elel[IntensExt]a[VerbTerm]

Conclusion and future work

The development of a clean morphologically annotated corpus of tractable size, making use of a Zulu computational morphological analyser, was described in some detail. The semi-automated process of systematically cleaning up a corpus and performing the morphological annotation thereof yields as 'by-products' a refined morphological analyser prototype and an improved and extended machine-readable lexicon. Moreover, human elicitation has been shown to be a key function both in ensuring the quality of the lexical information that is added to the lexicon – the system proposes candidate unknown roots/stems which in turn need to be verified by expert linguists or lexicographers; and in the context-based disambiguation of the often multiple correct analyses of the tokens in the corpus towards a gold standard for Zulu morphological analysis.

Future work includes the following:

- Continuous improvement of the output of the **ZulMorph** prototype, based on the exploitation of real world corpora.
- A refined and further extended XML machine-readable lexicon prototype, enriched and extended by means of inputs from existing electronically available corpora.
- A comprehensive prototype of the Zulu morphological analyser, available on-line for testing and feedback purposes.
- Extension of the corpus annotation to next levels of analysis, including syntactic structure of sentences.
- Application of the approach discussed in this article to develop similar resources (which may serve as annotated parallel corpora, based on translations of the South African Constitution) for the other South African Bantu languages for which morphological analysers similar to **ZulMorph** have also been developed.

Acknowledgements

This material is based upon work supported by the National Research Foundation under grant number 2053403. Any opinion, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Research Foundation.

Notes

1. This is significant in the light of the current establishment of a National Centre for HLT in South Africa that will serve among others as a repository for reusable high-level annotated text data, which includes the development of core technologies such as lemmatizers and morphological analysers, as well as stratified text corpora for all official languages of South Africa (cf. Department of Arts and Culture, 2006; CTexT, 2010).
2. We refer to a 'root/stem lexicon' since the lexicon contains roots such as verb roots e.g. *-hamb-* 'walk', *-bon-* 'see', as well as noun stems consisting of monomorphemic structures (also often called roots), e.g. *-ntu* as in *umuntu* 'person', *-khathi* as in *isikhathi* 'time' and polymorphemic structures such as nouns derived from verbs, e.g. *-cel-o* as in *isicelo* 'request', *-thwal-o* as in *umthwalo* 'load' and *-fund-is-i* as in *umfundisi* 'preacher/teacher'. See also Kosch (2006:6-12) in this regard. In the latter example the noun stem consists of a combination of a root plus its suffixal morphemes.
3. Van Eeden (1956:502) classifies *kambe* 'that is so/true, of course, well' as a conjunctive, but adds that the stem *-mbe* '… may also function as an adverb together with the adverb prefix *ka-*' (op cit. 1956:502).
4. The following entry for the language isiXhosa is found in Doke and Vilakazi (1964:865): '-xhoza (**isixhoza**, 3.2.9.9, sg. only) n. [<i(li)Xhoza.] Xhosa language, mannerisms. *ukukhuluma isixhoza* (to speak Xhosa).'
5. An entry does occur in Doke and Vilakazi (1964:703) but with a hyphen*: inqubekela-phambili*.

Appendix: Key to selected morphological tags

[Adv]	Adverb
[AdvPre]	Adverbial prefix
[ApplExt]	Applied verb extension
[AuxVStem]	Auxiliary verb stem
[BPre3]	Basic prefix, class 3
[Conj]	Conjunction
[Dem2][Pos2]	Demonstrative pronoun, class 2, position 2
[Guess][NStem]	Guessed noun stem
[Guess][VRoot]	Guessed verb root
[IntensExt]	Intensive verb extension
[NegPre]	Negative prefix
[NPrePre3]	Noun preprefix, class 3

[NStem]	Noun stem
[PossConc5]	Possessive morpheme class 5
[RecipExt]	Reciprocal verb extension
[RelConc2]	Relative morpheme class 2
[SC3]	Subject agreement morpheme, class 3
[SC][1pp]	Subject agreement morpheme, first person plural
[SC][2ps]	Subject agreement morpheme, second person singular
[VerbTerm]	Verb terminative
[VerbTermNeg]	Verb terminative negative
[VerbTermPerf]	Verb terminative perfect
[VerbTermSubj]	Verb terminative subjunctive
[VRoot]	Verb root

References

Beesley, K.R. & Karttunen, L. 2003. *Finite state morphology*. Stanford, CA: CSLI Publications.

Bosch, Sonja, & Fellbaum, Christiane. 2009. A comparative view of noun compounds in English and Zulu, in *After half a century of Slavonic natural language processing*, edited by Dana Hlaváčková, Aleš Horák, Klára Osolsobě & Pavel Rychlý. Brno: Masaryk University:35–44.

Bosch, S., Pretorius, L. & Fleisch, A. 2008. Experimental bootstrapping of morphological analysers for Nguni Languages. *Nordic Journal of African Studies* 17(2):66–88.

Bosch, S.E., Pretorius, L. & Jones, J. 2007. Towards machine-readable lexicons for South African Bantu languages. *Nordic Journal of African Studies* 16(2):131–145.

Bosch, S.E., & Pretorius, L. 2006. A finite-state approach to linguistic constraints in Zulu morphological analysis. *Studia Orientalia* 103:205–227.

CTexT. 2010. [O]. Available: http://www.nwu.ac.za/export/sites/default/nwu/p-news/pm_808_a.html Accessed on 2010/06/30.

De Schryver, G-M. & Gauton, R. 2002. The Zulu locative prefix *ku-* revisited: A corpus-based approach. *Southern African Linguistics and Applied Language Studies* 20(4): 201–220.

Department of Arts and Culture. 2006 [O]. Available: http://www.dac.gov.za/chief_directorates/language_services.htm. Accessed on 2010/06/30.

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

Doke, C.M. & Vilakazi, B. 1964. *Zulu-English dictionary*. Johannesburg: Witwatersrand University Press.

Fromkin, V., Rodman, R. & Hyams, N. 2007. *An introduction to language*. Massachusetts: Thomson Heinle.

Heid, U. 2008. Corpus linguistics and lexicography, in *Corpus linguistics: An international handbook, Volume 1*, edited by A. Lüdeling & M. Kytö. Berlin: Walter de Gruyter:131–153.

Hulden, M. 2009. Foma: a Finite-state compiler and library. *Proceedings of the 12th Conference of the European Chapter of the Association of Computational Linguistics*, March 30 – April 3, 2009, Athens, Greece.

Hurskainen, A. 1992. A two-level formalism for the analysis of Bantu morphology: an application to Swahili. *Nordic Journal of African Studies* 1(1):87–122.

Hurskainen, A. 1997. Information management and retrieval in Swahili, in *African linguistics at the crossroads. Papers from Kwaluseni*, edited by R.K. Herbert. Cologne: Rüdiger Köppe Verlag:629-642.

Hurskainen, A., Louwrens, L.J. & Poulos, G. 2005. Computational description of verbs in disjoining writing systems. *Nordic Journal of African Studies* 14(4):438–451.

Johansson, S. 2008. Some aspects of the development of corpus linguistics in the 1970s and 1980s, in *Corpus linguistics: An international handbook, Volume 1*, edited by A. Lüdeling & M. Kytö. Berlin: Walter de Gruyter:15–33.

Jones, J., Bosch, S., Pretorius, L. & Prinsloo, D. 2005. Development of reusable resources for human language technologies (HLT) applications: practice and experience. *South African Journal of African Languages* 25(2):141–159.

Karttunen, L. 2003. Finite-state technology, in *The Oxford handbook of computational linguistics*, edited by R. Mitkov. New York: Oxford University Press Inc.:339–357.

Kosch, I.M. 2006. *Topics in morphology in the African language context*. Pretoria: Unisa Press.

Koskenniemi, K. 1997. Representations and finite-state components in natural language, in *Finite-state language processing*, edited by E. Roche & Y. Schabes. Cambridge, MA & London: The MIT Press:99–116.

Language corpora compiled at the University of Pretoria. 2008. [O]. Available: http://web.up.ac.za/default.asp?ipkCategoryID=1883&subid=1883. Accessed on 2010/06/30.

Lehmberg, T. & Wörner, K. 2008. Annotation standards, in *Corpus linguistics: An international handbook, Volume 1*, edited by A. Lüdeling & M. Kytö. Berlin: Walter de Gruyter:484–500.

Mbatha, M.O. (ed.) 2006. *Isichazamazwi SesiZulu*. Pietermaritzburg: New Dawn Publishers.

McEnery, T., Xiao, R. & Tono, Y. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.

Muhirwe, J. 2007. Computational analysis of Kinyarwanda morphology: The morphological alternations. *International Journal of Computing and ICT Research* 1(1):85– 92. http://www.ijcir.org/volume1-number1/article10.pdf.

Poulos, G. & Msimang, C.T. 1998. *A linguistic analysis of Zulu*. Pretoria: Via Afrika.

Pretorius, L. & Bosch, S.E. 2003. Finite-State Computational Morphology: An analyzer prototype for Zulu. *Machine Translation* 18:195–216.

Pretorius, L. & Bosch, S.E. 2010. Finite state morphology of the Nguni language cluster: Modelling and implementation issues, in *Finite-state methods and natural language processing 8th International Workshop, FSMNLP 2009, Pretoria, South Africa, July 21-24, 2009, Revised Selected Papers*, Lecture Notes in Computer Science Volume 6062/2010, Berlin, Heidelberg: Springer, ISSN 0302-9743 (Print) 1611-3349 (Online):123–130.

Pretorius, R., Viljoen, E. & Pretorius, L. 2005. A finite-state morphological analysis of Tswana nouns. *South African Journal of African Languages* 25(1):48–58.

Reynaert, M. 2006. Corpus-induced corpus clean-up. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, Genoa, Italy, May 22-28, 2010.

Schmid, H. 2008. Tokenizing and part-of-speech tagging, in *Corpus linguistics: An international handbook, Volume 1*, edited by A. Lüdeling & M. Kytö. Berlin: Walter de Gruyter:527–551.

The Constitution. sa. [O]. Available: http://www.constitutionalcourt.org.za/site/theconstitution/zulu.htm. Accessed on 2010/06/30.

The IEEE and The Open Group. 2004. *The Open Group Base Specifications Issue 6*. IEEE Std 1003.1, 2004 Edition. Available: http://www.opengroup.org/onlinepubs/009695399/functions/iconv.html. Accessed on 2010/07/24.

Van Eeden, B.I.C. 1956. *Zoeloe grammatika*. Stellenbosch: Universiteitsuitgewers en Boekhandelaars (Edms.) Beperk.

Van Huyssteen, L. & Bosch, S. 2008. Place names: Challenges for a Zulu computational morphological analyser. *Nomina Africana* 22(1 & 2):105–125.