Software Tools for Morphological Tagging of Zulu Corpora and Lexicon Development

Sonja E Bosch[1] and Laurette Pretorius[2]
[1]Department of African Languages and [2]School of Computing
PO Box 392, UNISA, 0003 Pretoria, South Africa
boschse@unisa.ac.za and pretol@unisa.ac.za

Abstract

The aim of this paper is to discuss aspects of an on-going project on the development of grammatical and lexical resources for Zulu with sufficient coverage for unrestricted text. We explain how the basic software tools of computational morphology are used in linguistic processing, more specifically for automatic word form recognition and morphological tagging of the growing stock of electronic text corpora of a Bantu language such as Zulu. It is also shown how a machine-readable lexicon is in turn enhanced with the information acquired and extracted by means of such corpus analysis.

Introduction

The central role that electronic text corpora play in natural language processing is well known. At present such Zulu corpora[1] are predominantly raw, that is unannotated, resulting in obvious limitations when information needs to be extracted. This is particularly the case in a morphologically complex language such as Zulu, which is predominantly agglutinating in nature. Our efforts are therefore focused on methods and tools for automated morphological tagging and annotation of text corpora.

An essential tool for our specific natural language processing application is a morphological analyser, which maps written orthographical Zulu words to possible analyses that separate and identify constituent roots and affixes, while assigning suitable tags to these constituents (see also Pretorius & Bosch, 2003). The morphological analyser contains

a) a finite-state lexical transducer, representing all the morphological information about the language being analysed,

b) an embedded representation of Zulu word roots/stems, and

c) a guessing module for analysing words of which the roots/stems are not represented in the transducer.

Furthermore, a machine-readable Zulu lexicon is needed as a basic resource. In order to ensure portability and re-usability of this resource, it is developed as an XML document. XML is the de facto standard for machine-readable text documents and makes such a lexicon suitable to function as a crucial, integral language resource for a wide range of applications. A present limitation in the natural language processing of Zulu is the fact that a machine-readable Zulu lexicon is not readily available in any form, although we have access to a monolingual word list. The only explicit linguistic information included in the word list is noun class information.

Building A Morphological Analyser

The approach followed in the on-going development and implementation of a finite-state morphological analyser prototype for Zulu is based on the Xerox finite-state tools (Beesley & Karttunen, 2003). In order to automate the computational analysis of word forms, the modelling of two general linguistic components is required. The morphotactics component contains the word formation rules, which determine the construction of words/word forms from the inventory of morphemes (that is roots and affixes). The morphophonological alternations component describes the morphophonological changes between lexical and surface levels.

The modelling of the two linguistic components requires specialised tools for the automatic analysis of word forms, as well as for most other corpus-based analyses. Compared to a language such as English for instance, where the variation of word forms is relatively limited, the situation in a language such as Zulu is quite different. Zulu, being mainly an agglutinating language, entails extensive use of prefixes as well as suffixes in the formation of words.

In Zulu the root is the constant core element from which words or word forms are constructed while the rest is inflection and derivation. Therefore, morphological analysis is essential for any kind of information retrieval from text corpora. Zulu follows the convention of a conjunctive writing system, that is a system in which the morphemes constituting a word appear as a single token. Each linguistic word consists of a number of bound parts or morphemes that can never occur independently as separate words. The two types of morphemes that are generally recognised, are roots and affixes. Roots always form the lexical core of a word, while affixes usually add a grammatical meaning or function to the word. It should be noted that for purposes of convenience as discussed in Poulos & Msimang (1998:170), the uninflected forms of nouns or noun-based words are usually referred to as noun stems.

The complex nature of the monosyllabic noun stem *-phaphu*, meaning 'lung', is exemplified by the stem appearing in various forms within nouns, for instance:

i(li)phaphu 'lung'
amaphaphu 'lungs'
emaphashini 'in the lungs'
amaphashana 'small lungs'.

In the word *emaphashini* for instance, it is not at all obvious that the stem is identical to that of all the other words above, namely *–phaphu,* and that its morphological analysis is as follows:

e- locative prefix
-ma- basic prefix class 6

[1] An example is the Pretoria isiZulu Corpus (PZC), a general language corpus based mainly on written texts but also including internet texts and consisting of 5,783,634 tokens; 674,380 types.

-*phaphu* noun stem 'lung'
-*ini* locative suffix 'in the lungs'.
As already mentioned, the automation of morphological analysis requires the modelling of two general linguistic components, viz. morphotactics and morphophonological alternations.

The Morphotactics

The morphotactics component contains the word-formation rules, which determine the construction of words/word forms from the inventory of morphemes. This inventory of morphemes consists of a) word roots/stems that form an 'open' morpheme class in the sense that new roots/stems can be added continuously, and b) affixes, the 'closed' morpheme classes which model the fixed morphological structure of words. Morphemes that constitute words cannot combine at random, but are confined to certain combinations and sequences. A morphological analyser therefore needs to know the valid combinations of morphemes of the language concerned.
For the implementation of the morphotactics the Xerox tools provide a declarative programming language, namely **lexc** (for **Lex**icon **C**ompiler) for specifying the required natural language lexicon and the morphotactic structure of the words in the language. The **lexc** script, consisting of a cascade of so-called lexicons (morpheme continuation classes that represent the valid morpheme combinations of Zulu), is compiled into a finite-state network.
In the Zulu word *emaphashini* 'in the lungs' for instance, the morphemes *e-*, *-ma-*, *-phaphu-* and *-ini* form the intermediate morphophonemic or lexical string. At the lexical level, morphemes as well as morphological feature tags relating to the word are defined, that is e[LocPre]ma[BPre]phaphu[NRoot]ini[Loc]. In this example, the noun stem is preceded by a locative prefix and a basic prefix of the noun class 6, and is followed by the locative suffix.

The Morphophonological Alternations

The morphophonological alternations component describes the morphophonological changes between lexical and surface levels in order to render the correct surface form of the word. A morphological analyser needs to recognise the correct form of each morpheme since one and the same morpheme may be realised in different ways depending on the environment in which it occurs.
With the suffixation of the locative morpheme *-ini*, various phonological changes may occur, depending on the nature of the final syllable of the noun stem. It is such phonological changes that need to be accounted for by way of alternation rules in the automatic morphological analysis of Zulu. In the word *emaphashini* the locative suffix *-ini* triggers palatalisation, because the final labial consonant *-ph-* of the noun stem is followed by a back labial vowel [u]. In the process *-ph-* changes to a palatal *-sh-* (cf. Poulos & Msimang, 1998:532).
The morphophonological alternations are modelled by means of the Xerox regular expression language **xfst**. The **xfst** script is then compiled into a finite-state network by the **xfst** compiler.
Finally the two networks, representing the morphotactics and the morphophonological alternation rules, are composed into a single network, which constitutes the morphological analyser.

The Current Prototype

The current status of the morphological analyser prototype for Zulu, which includes a representative sample of all word categories, covers most of the morphotactics and morphophonological alternations required for the automated analysis/generation of nouns of all 15 noun classes, the positive and negative forms of verbs in the present, perfect, past and future tenses, absolute and quantitative pronouns, four positions of the demonstrative and copulative demonstrative, underived adverbs, positive and negative forms of relatives in similar tenses as the verb, possessives, conjunctions and ideophones. Word categories that still need to be supplemented are the verb, with compound tenses of the verb, the adverb with derived adverbial constructions, and the adjective in the various tenses.
The ultimate aim is to model the morphological structure of Zulu in such a way that all valid words in the language are included and correctly analysed, but that all character strings that do not represent words in the real language are excluded.

The Guesser Variant Of The Analyser

In order to systematically update and extend the root/stem list of the morphological analyser, reflecting the dynamic nature of the language as well as the well-known principle that lexicon development is an on-going and repeated activity, the Xerox finite-state tools include a useful feature, namely the option of building a so-called guesser.
The guesser is a variant of the morphological analyser and is designed to identify all possible analyses of a word based not on the roots/stems embedded in the lexical transducer, but on all phonologically possible word roots/stems in Zulu. The roots/stems in the analysis returned by the guesser are clearly marked as guesses, and are scrutinised by a lexicographer, before leading to new entries in our evolving machine readable lexicon. The guesser is therefore a powerful tool to drive lexicography (cf. Beesley, 2003:25).

Developing A Machine-Readable Lexicon

By definition the analyser can only recognise and analyse words of which the roots/stems have been explicitly included in its embedded lexicon. Ideally, a comprehensive machine-readable Zulu lexicon in the form of an XML document should be available as a basic resource from which word roots/stems may be obtained. This would involve a process of so-called 'down-translation' from the XML format into a format suitable for inclusion in the morphological analyser (Beesley, 2003).
Since no such machine-readable lexicon for Zulu was available at the start of the project, an electronic monolingual word list was used to build a rudimentary lexicon as an XML document of which each entry contained a lemma from the list. The word list was based on a Zulu paper dictionary (Doke & Vilakazi, 1964) of which the last revised edition dates back to the1950s, and contains a total of over 28 000 lemmas of which about 13 000 are noun stems.

Since the available word list represents the Zulu vocabulary as it existed 50 years ago, it is understandable that the coverage of word roots/stems from a recent corpus of running text may be unsatisfactory, due to the dynamic nature of language. Provision therefore needs to be made for the constant inclusion of new roots/stems, be they newly coined or as yet unlisted foreign roots/stems.

At this point it is worth emphasising that the word list served as a central resource in building first versions of both the analyser and the rudimentary XML lexicon. To enhance, refine and update them requires current and contemporary language resources in the form of text corpora.

The Refinement Cycle

We follow a process by means of which an up-to-date machine-readable lexicon and a computational morphological analyser for Zulu may be maintained and enriched by using the growing body of electronically available text corpora. Automating this process to a large extent, allows the Zulu lexicographer to focus on those linguistic issues that require human examination and judgement. The tools and proposed methods may also be employed in evaluating the lexicon and the analyser on a regular basis in order to ensure their integrity, and thereby create computational tools that can be used in various other areas of Zulu natural language processing.

The availability of a morphological analyser prototype and an XML lexicon prototype can serve a dual purpose. Firstly, the morphological analyser may be systematically applied to new text corpora in order to enhance such corpora with morphological tagging where possible. Secondly, words in the corpora that are not successfully analysed may be considered as candidates for new Zulu words thus providing information regarding the growth and evolution of the language. This may then lead to new entries in the XML lexicon.

Once the text corpora have been explored for new Zulu words, the guesser variant of the analyser may be applied to identify these as yet unknown words and suggest possible associated stems.

In short, value is added to a corpus in the form of morphological tagging of known word roots/stems and value is added to the analyser and XML lexicon in the form of the addition of new word roots/stems. As a final step in the refinement cycle all the valid words in the relevant corpus may now be morphologically tagged, adding value to the raw corpus by enabling further processing such as part-of-speech (POS) tagging and disambiguation.

Exploring A Sample Corpus

The enrichment of text corpora and the refinement cycle of the existing lexicon are demonstrated by means of a raw sample corpus[2] consisting of 4 700 tokens. The sample corpus was selected according to its relevance to language technology applications in the South African context, often involving Text-to-Speech and Speech-to-Speech systems. This is a domain specific corpus relating to primary healthcare, with an emphasis on cholera, tuberculosis and HIV/AIDS. The corpus, based on medical pamphlets in Zulu, does not include highly specialised terms of the domain, but rather words that laypersons need to read and understand as part of their everyday lives. The focus here is on the occurrence of nouns and noun-based words in the corpus.

As explained in previous sections, the morphological analyser is now applied to the sample corpus resulting in the tagging of 360 known roots/stems occurring in nouns or noun-based words.

For the morphological analyser to be able to identify roots/stems, they need to be represented in the analyser. Examples are *negciwane* 'with a germ', *ngumhlengikazi* 'it is a nurse' and *kunezimpawu* 'it has symptoms', and they are tagged as follows:

```
na[AdvForm]i[NPrePre5]li[BPre5]
gciwane[NStem],
ngu[CopPre]u[NPrePre1]mu[BPre1]
hlengi[NStem]kazi[AugSuf]
```

and

```
ku[SC15]na[AdvForm]i[NPrePre10]
zin[BPre10]phawu[NStem]
```

Words constructed from roots/stems that are not represented, will not be analysed. Referring to the sample corpus, should the stem of the noun *isifo* 'disease', for instance, not be represented in the morphological analyser, it would not be possible to tag this stem in any of the following different word forms occurring in the corpus, i.e.

isi-fo
yisi-fo
zesi-fo
ngesi-fo
nesi-fo
kwesi-fo
abanesi-fo
ezi-fweni
izi-fo

Therefore, as a next step the guesser is applied to all the words in the sample corpus that were not successfully analysed and tagged. This resulted in the extraction of 78 new noun roots/stems, which were either newly coined Zulu words or foreign words without standardised spelling.

The output of the guesser when applied, for example, to the words *ilitha* 'litre' and *yingculazi* 'it is AIDS' is as follows:

```
i[NPrePre5]li[BPre5]litha[NStem-Guess]
i[NPrePre9a]litha[NStem-Guess]
```

and

```
yi[CopPre]i[NPrePre9]n[BPre9]
gculazi[NStem-Guess]
yi[CopPre]i[NPrePre5]li[BPre5]
ngculazi[NStem-Guess]
yi[CopPre]i[NPrePre9]n[BPre9]
ngculazi[NStem-Guess]
yi[CopPre]i[NPrePre9a]
ngculazi[NStem-Guess]
```

The expert human lexicographer scrutinises this output and builds lexicon lemmas to be included in the XML machine-readable lexicon. *–litha* is a foreign stem and has been tagged appropriately as a guessed noun stem in both instances. In order to determine the correct noun class (5 or 9a) the lexicographer needs to examine the surrounding context. *Ilitha* 'litre' is qualified by *yamanzi* 'of water'

[2] This corpus was compiled for research on Zulu as a technical language (van Huyssteen, 2003).

where the possessive concord *ya-* clearly indicates class 9a concordial agreement. The correct guess is identified as i[NPrePre9a]litha[NStem-Guess].
The newly coined noun stem *-ngulazi* is tagged as the appropriate guess in the second example above, which leaves the lexicographer to determine the correct noun class of the noun stem, that is 5, 9 or 9a. The possibility of class 9a is already excluded by the fact that this stem is not of foreign origin. Concordial agreement with *ingulazi* in the corpus clearly points to class 9 as the correct class, and therefore the correct guess is: yi[CopPre]i[NPrePre9]n[BPre9]ngculazi[NStem-Guess].
Approximately 21% of the nouns and noun-based words in the sample corpus contain new roots/stems, which are not listed in the electronically available word list of Zulu. This discrepancy means that 21% of the important key words of the domain-specific corpus are not included in the lexicon, which is a clear indication of the need to supplement the lexicon with these specific lemmas.
The next step in the refinement cycle is completed by adding the new lemmas to the XML lexicon, down-translating the updated lexicon to **lexc** format and rebuilding the analyser and guesser.
Finally, the refinement cycle is completed by applying the analyser to the sample corpus. The result is an enriched morphologically tagged sample corpus.

Conclusion

In this paper a semi-automated process for developing, maintaining and enriching a representative XML lexicon is described. The building of a computational morphological analyser/generator for Zulu is described, together with its application to a domain specific sample corpus. In view of the growing body of electronically available language corpora, this process enables the enrichment of raw text corpora with morphological tags, necessary for further natural language processing such as part-of-speech (POS) tagging and disambiguation.
This methodology and the insight gained in the Zulu project are already being applied to the development of automated morphological analysers and XML machine-readable lexicons for other Bantu languages such as Xhosa and Northern Sotho.

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 author and do not necessarily reflect the views of the National Research Foundation.

References

Beesley, K.R. (2003). Finite-State Morphological Analysis and Generation for Aymara. In Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, 10th Conference of the European Chapter of the Association for Computational Linguistics, April 13-14 2003, Budapest, Hungary. ACL http://www.acl-web.org. (pp. 19-26).

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

Doke, C.M. & Vilakazi, B. (1964). Zulu-English Dictionary. Johannesburg: Witwatersrand University Press.

Pretorius, L. & Bosch, S.E. (2003). Computational aids for Zulu natural language processing. Southern African Linguistics and Applied Language Studies, 21(4), 267--282.

Van Eynde, F. & Gibbon, D. (2000). Lexicon development for speech and language processing. Dordrecht: Kluwer Academic Publishers.

Van Huyssteen, L. (2003). A practical approach to the standardisation and elaboration of Zulu as a technical language. Unpublished dissertation. Pretoria: University of South Africa.