# Exploiting Cross-linguistic Similarities in Zulu and Xhosa Computational Morphology

**Laurette Pretorius**
School of Computing
University of South Africa &
Meraka Institute, CSIR
Pretoria, South Africa
pretol@unisa.ac.za

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

## Abstract

This paper investigates the possibilities that cross-linguistic similarities and dissimilarities between related languages offer in terms of bootstrapping a morphological analyser. In this case an existing Zulu morphological analyser prototype (ZulMorph) serves as basis for a Xhosa analyser. The investigation is structured around the morphotactics and the morphophonological alternations of the languages involved. Special attention is given to the so-called "open" class, which represents the word root lexicons for specifically nouns and verbs. The acquisition and coverage of these lexicons prove to be crucial for the success of the analysers under development. The bootstrapped morphological analyser is applied to parallel test corpora and the results are discussed. A variety of cross-linguistic effects is illustrated with examples from the corpora. It is found that bootstrapping morphological analysers for languages that exhibit significant structural and lexical similarities may be fruitfully exploited for developing analysers for lesser-resourced languages.

## 1 Introduction

Zulu and Xhosa belong to the Nguni languages, a group of languages from the South-eastern Bantu zone and, as two of the eleven official languages of South Africa, are spoken by approximately 9 and 8 million mother-tongue speakers, respectively. In terms of natural language processing, particularly computational morphology, the Bantu languages including Zulu and Xhosa certainly belong to the lesser-studied languages of the world.

One of the few Bantu languages for which computational morphological analysers have been fully developed so far is Swahili (Hurskainen, 1992; De Pauw and De Schryver, 2008). A computational morphological analyser prototype for Zulu (ZulMorph) is in an advanced stage of development, the results of which have already been used in other applications. Preliminary experiments and results towards obtaining morphological analysers for Xhosa, Swati and Ndebele by bootstrapping ZulMorph were particularly encouraging (Bosch et al., 2008). This bootstrapping process may be briefly summarised as a sequence of steps in which the baseline analyser, ZulMorph, is applied to the new language (in this case Xhosa) and then systematically extended to include the morphology of the other language. The extensions concern the word root lexicon, followed by the grammatical morpheme lexicons and finally by the appropriate morphophonological rules. The guiding principle in this process is as follows: Use the Zulu morphological structure wherever applicable and only extend the analyser to accommodate differences between the source language (Zulu) and the target language (in this case Xhosa). So far the question as to whether the bootstrapped analyser, extended to include Xhosa morphology, could also improve the coverage of the Zulu analyser was not specifically addressed in Bosch et al. (2008).

Cross-linguistic similarity and its exploitation is a rather wide concept. In its broadest sense it aims at investigating and developing resources and technologies that can be compared and linked, used and analysed with common approaches, and that contain linguistic information for the same or comparable phenomena. In this paper the focus is on the morphological similarities and dissimilarities between Zulu and Xhosa and how these cross-linguistic similarities and dissimilarities inform the bootstrapping of a morphological analyser for Zulu and Xhosa. In particular, issues such as open versus closed classes, and language specific morphotactics and alternation rules are discussed. Special attention

is given to the word root lexicons. In addition, the procedure for bootstrapping is broadened to include a guesser variant of the morphological analyser.

The structure of the paper is as follows: Section 2 gives a general overview of the morphological structure of the languages concerned. The modelling and implementation approach is also discussed. This is followed in sections 3 and 4 by a systematic exposition of the cross-linguistic dissimilarities pertaining to morphotactics and morphophonological alternations. Section 5 focuses on the so-called "open" class, which represents the word root lexicons for specifically nouns and verbs. The acquisition and coverage of these lexicons prove to be crucial for the success of the analysers under development. Section 6 addresses the use of the guesser variant of the morphological analyser as well as the application of the bootstrapped morphological analyser to parallel test corpora. A variety of cross-linguistic effects is illustrated with examples from the corpora. This provides novel insights into the investigation and exploitation of cross-linguistic similarities and their significance for bootstrapping purposes. Section 7 concerns future work and a conclusion.

## 2 General overview

### 2.1 Morphological structure

Bantu languages are characterised by a rich agglutinating morphological structure, based on two principles, namely the nominal classification system, and the concordial agreement system. According to the nominal classification system, nouns are categorised by prefixal morphemes. These noun prefixes have, for ease of analysis, been assigned numbers by scholars who have worked within the field of Bantu linguistics, In Zulu a noun such as *umuntu* 'person' for instance, consists of a noun prefix *umu-* followed by the noun stem *-ntu* and is classified as a class 1 noun, while the noun *isitha* 'rival' consists of a noun prefix *isi-* and the noun stem *-tha* and is classified as a class 7 noun. Noun prefixes generally indicate number, with the uneven class numbers designating singular and the corresponding even class numbers designating plural. The plural forms of the above examples would therefore respectively be the class 2 noun *abantu* 'persons' and the class 8 noun *izitha* 'rivals'. We follow Meinhof's (1932:48) numbering system which distinguishes between 23 noun prefixes altogether in the various Bantu languages.

The concordial agreement system is significant in the Bantu languages because it forms the backbone of the whole sentence structure. Concordial agreement is brought about by the various noun classes in the sense that their prefixes link the noun to other words in the sentence. This linking is manifested by a concordial morpheme that is derived from the noun prefix, and usually bears a close resemblance to the noun prefix, as illustrated in the following example:

***Izi**tsha le**zi** **ezi**ne zephukile*
'These four plates are broken'

This concordial agreement system governs grammatical correlation in verbs, adjectives, possessives, pronouns, and so forth. Bantu languages are predominantly agglutinating and polymorphematic in nature, with affixes attached to the root or core of the word.

The morphological make-up of the verb is considerably more complex than that of the noun. A number of slots, both preceding and following the verb root may contain numerous morphemes with functions such as derivations, inflection for tense-aspect and marking of nominal arguments. Examples are cross-reference of the subject and object by means of class- (or person-/number-)specific object markers, locative affixes, morphemes distinguishing verb forms in clause-final and non-final position, negation etc.

Despite the complexities of these domains, they are comparable across language boundaries, specifically Nguni language boundaries, with a degree of formal similarity that lends itself to exploitation for bootstrapping purposes.

### 2.2 Modelling and Implementation

In the modelling and implementation of the morphological structure a finite-state approach is followed. The suitability of finite-state approaches to computational morphology is well known and has resulted in numerous software toolkits and development environments for this purpose (cf. Koskenniemi, 1997 and Karttunen, 2001). Yli-Jyrä (2005) discusses the importance of a finite-state morphology toolkit for lesser-studies languages. He maintains that "[a]lthough some lexicons and morphological grammars can be learned automatically from texts ... fully automatic or unsupervised methods are not sufficient. This is due to two reasons. First, the amount of freely available corpora is limited for many of the less studied languages. Second, many of the less studied languages have rich morphologies that are difficult to learn accurately with unsupervised methods".

The Xerox finite-state tools (Beesley and Karttunen, 2003) as one of the preferred toolkits for modelling and implementing natural language morphology, is used in this work.

The morphological challenges in computational morphological analysis comprise the modelling of two general linguistic components, namely morphotactics (word formation rules) as well as morphophonological alternations.

Ideally, the morphotactics component should include all and only word roots 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 the language concerned. Moreover, the morphophonological alternations rules should constitute all known sound changes that occur at morpheme boundaries. The combination of these two components constitutes an accurate model of the morphology of the language(s) under consideration.

The Xerox lexicon compiler, **lexc**, is well-suited to capturing the morphotactics of Zulu. A **lexc** script, consisting of cascades of so-called continuation classes (of morpheme lexicons) representing the (concatenative) morpheme sequencing, is compiled into a finite-state network. The Xerox regular expression language, **xfst**, provides an extended regular expression calculus with sophisticated Replace Rules for describing the morphophonological alternations rules of Zulu. The **xfst** script is also compiled into a finite-state network. These networks are finally combined by means of the operation of composition into a so-called Lexical Transducer that constitutes the morphological analyser and contains all the morphological information of Zulu, including derivation, inflection, alternation and compounding. Pretorius and Bosch (2002) address the suitability of this approach to Zulu morphology and illustrate it by means of examples of **lexc** and **xfst** scripts for modelling the Zulu noun.

A detailed exposition of the design and implementation of ZulMorph may be found in Pretorius and Bosch (2003). In addition to considering both the accurate modelling of the morphotactics and the morphophonological alternation rules, they also address implementation and other issues that need to be resolved in order to produce a useful software artefact for automated morphological analysis. Issues of implementation include a justification for the finite-state approach followed, designing for accuracy and correctness and decisions regarding the analyser's interface with its environment and its usage.

Particular attention is paid to the handling of exceptions; the modelling of separated dependencies by means of so-called flag-diacritics; the specification of lexical forms (analyses) in terms of morphological granularity and feature information; the choice of an associated and appropriate morphological tag set and also the positioning of these tags in relation to the morphemes they are associated with in the morphological analyses (lexical forms) that are rendered.

The components of ZulMorph, including its scope in terms of word categories and their morphological structure, are summarised in Table 1 while its lexical coverage as reflected by the number of different noun stems, verb roots etc. is discussed in section 5.

The bootstrapping of ZulMorph to provide for Xhosa as well requires a careful investigation of the cross-linguistic similarities and dissimilarities and how they are best modelled and implemented. This aspect will be discussed in more detail in the following section.

| | | | |
|---|---|---|---|
| Morphotactics (**lexc**) | Affixes for all parts-of-speech (e.g. subject & object concords, noun class prefixes, verb extensions etc.) | Word roots (e.g. nouns, verbs, relatives, ideophones) | Rules for legal combinations and orders of morphemes (e.g. *u-ya-ngi-thand-a* and not **ya-u-a-thand-ngi*) |
| Morpho-phonological alternations (**xfst**) | Rules that determine the form of each morpheme (e.g. *ku-lob-w-a* > *ku-lotsh-w-a*, *u-mu-lomo* > *u-m-lomo*) | | |

Table 1: Zulu Morphological Analyser Components

## 3 Morphotactics

In word formation we distinguish between so-called closed and open classes. The open class accepts the addition of new items by means of processes such as borrowing, coining, compounding and derivation. In the context of this paper, the open class represents word roots including verb roots and noun stems. The closed class represents affixes that model the fixed morphological structure of words, as well as items such as conjunctions, pronouns etc. Typically no new items can be added to the closed class (Fromkin et al., 2003:74).

Since our point of departure is ZulMorph, we focus on Xhosa affixes that differ from their Zulu

counterparts. A few examples are given in Table 2.

Certain areas in the Xhosa grammar need to be modelled independently and then built into the

| Morpheme | Zulu | Xhosa |
|---|---|---|
| **Noun Class Prefixes** | | |
| Class 1 and 3 *um(u)-* | full form *umu-* with monosyllabic noun stems, shortened form with polysyllabic noun stems: *umu-ntu*, *um-fana* | um- with all noun stems: *um-ntu, um-fana* |
| Class 2a | *o-: o-baba* | *oo-: oo-bawo* |
| Class 9 | *in-* with all noun stems: *in-nyama* | *i-* with noun stems beginning with *h*, *i*, *m*, *n*, *ny*: *i-hambo* |
| Class 10 | *izin-* with monosyllabic and polysyllabic stems. *izin-ja; izin-dlebe* | *iin-* with polysyllabic stems: *iin-dlebe* |
| **Contracted subject concords (future tense).** Examples: | | |
| 1ps<br>2ps, Class 1 & 3<br>Class 4 & 9 | *ngo-*<br>*wo-*<br>*yo-* | *ndo-*<br>*uyo-*<br>*iyo-* |
| **Object concords** | | |
| 1ps | *ngi-* | *ndi-* |
| **Absolute pronouns** | | |
| 1ps<br>Class 15 | *mina*<br>*khona* | *mna*<br>*kona* |
| **Demonstrative Pronouns:** Three positional types of the demonstrative pronouns are listed separately for each language. Examples: | | |
| Class 1<br>Class 5 | Pos. 1 *lo*; Pos. 2 *lowo*; Pos. 3 *lowaya*<br>Pos. 1 *leli*; Pos. 2 *lelo*; Pos. 3 *leliya* | Pos. 1 *lo*; Pos. 2 *lowo/loo*; Pos. 3 *lowa*<br>Pos. 1 *eli*; Pos. 2 *elo*; Pos. 3 *eliya* |
| **Adjective basic prefixes** | | |
| 1ps<br>2ps<br>Class1& 3<br>Class 8 | *ngim(u-)*<br>*umu-*<br>*mu-*<br>*zin-* | *nim-*<br>*um-*<br>*m-*<br>*zi-* |
| **Locative demonstrative copulatives :** Three positional types of the so-called locative demonstrative copulatives differ considerably for Zulu and Xhosa and are therefore listed separately for each language. Examples: | | |
| Class 1<br>Class 5 | Pos. 1 *nangu*; Pos. 2 *nango*; Pos. 3 *nanguya*<br>Pos. 1 *nanti*; Pos. 2 *nanto*; Pos. 3 *nantiya* | Pos. 1 *nanku*; Pos. 2 *nanko*; Pos. 3 *nankuya*<br>Pos. 1 *nali*; Pos. 2 *nalo*; Pos. 3 *naliya* |
| **Copulatives :** Formation of copulatives derived from Xhosa nouns differs considerably from Zulu. This construction is class dependent in Xhosa and is modelled differently to its Zulu counterpart. Examples: | | |
| | *yi-* combines with noun prefixes i-:<br>*yi-indoda > yindoda*<br>*ngu-* combines with noun prefixes *u-, o-, a*:<br>*ngu-umuntu > ngumuntu*<br>*ngu-obaba > ngobaba*<br>*ngu-amakati > ngamakati*<br>*wu* combines with noun prefixes *u-, o-:*<br>*wu-muntu > wumuntu,*<br>*wu-obaba > wobaba* | *ngu-* combines with classes 1, 1a, 2, 2a, 3 & 6, e.g. *ngu-umntu > ngumntu*<br>*yi-* combines with classes 4 *imi-* and 9 *in-*, e.g. *yi-imithi > yimithi*<br>*li-* combines with class 5 *i(li)-:*<br>*li-ihashe > lihashe*<br>*si-* combines with class 7 *isi-:*<br>*si-isitya > sisitya*<br>etc. |

Table 2. Examples of variations in Zulu and Xhosa 'closed' morpheme information

analyser, for instance the formation of the so-called temporal form that does not occur in Zulu. The temporal form is an indication of when an action takes place or when a process is carried out, and has a present or past tense form (Louw, et al., 1984:163). The simple form consists of a subject concord plus *-a-* followed by the verb stem in the infinitive, the preprefix of which has been elided, for example *si-a-uku-buya > sa-kubuya* 'when we return'. In terms of the word formation rules this means that an additional Xhosa specific morpheme lexicon (continuation class) needs to be included. To facilitate accurate modelling appropriate constraints also need to be formulated.

The bootstrapping process is iterative and new information regarding dissimilar morphological constructions is incorporated systematically in the morphotactics component. Similarly, rules are adapted in a systematic manner. The process

also inherently relies on similarities between the languages, and therefore the challenge is to model the dissimilarities accurately. The carefully conceptualised and appropriately structured (**lexc**) continuation classes embodying the Zulu morphotactics provide a suitable framework for including all the closed class dissimilarities discussed above.

## 4 Morphophonological alternations

Differences in morphophonological alternations between Zulu and Xhosa are exemplified in Table 3. Some occur in noun class prefixes of class 10 and associated constructions, such as prefixing of adverbial morphemes (*na-*, *nga-*, etc.). Others are found in instances of palatalisation, "a sound change whereby a bilabial sound in passive formation, locativisation and diminutive formation is replaced by a palatal sound" (Poulos and Msimang, 1998:531).

| Zulu | Xhosa |
|---|---|
| Class 10 class prefix *izin-* occurs before monosyllabic as well as polysyllabic stems, e.g. *izinja,* ***izindlebe*** <br> Adverb prefix *na + i > ne*, e.g. ***nezindlebe*** (*na-izin-ndlebe*) | Class 10 class prefix *izin-* changes to *iin-* before polysyllabic stems, e.g. *izinja, iindlebe* <br> Adverb prefix *na + ii > nee*; e.g. ***neendlebe*** (*na-iin-ndlebe*) |
| Palatalisation with passive, diminutive & locative formation: <br> ***b > tsh*** <br> *-hlab-w-a > -hlatsh-w-a, intaba-ana > intatsh-ana,* <br> *indaba > entdatsheni* <br> ***ph > sh*** <br> *-boph-w-a > -bosh-w-a, iphaphu-ana > iphash-ana* <br> *iphaphu > ephasheni* | Palatalisation with passive, diminutive & locative formation: <br> ***b > ty*** <br> *-hlab-w-a > -hlaty-w-a, intaba-ana > intaty-a na* <br> *ihlobo > ehlotyeni* <br> ***ph > tsh*** <br> *–boph-w-a > -botsh-w-a, iphaphu-ana > iphatsh-ana,* <br> *usapho > elusatsheni* |

Table 3. Examples of variations in Zulu and Xhosa morphophonology

As before, the Zulu alternations are assumed to apply to Xhosa unless otherwise modelled. Regarding language-specific alternations special care is taken to ensure that the rules fire only in the desired contexts and order. For example, Xhosa-specific sound changes should not fire between Zulu-specific morphemes, and vice versa. This applies, for instance, to the vowel combination *ii*, which does not occur in Zulu. While the general rule *ii > i* holds for Zulu, the vowel combination *ii* needs to be preserved in Xhosa.

## 5 The word root lexicons

Compiling sufficiently extensive and complete word root lexicons (i.e. populating the "open" word classes) is a major challenge, particularly for lesser-resourced languages (Yli-Jyrä, 2005:2). A pragmatic approach of harvesting roots from all readily available sources is followed. The Zulu lexicon is based on an extensive word list dating back to the mid 1950s (cf. Doke and Vilakazi, 1964), but significant improvements and additions are regularly made. At present the Zulu word roots include noun stems with class information (15 759), verb roots (7 567), relative stems (406), adjective stems (48), ideophones (1 360), conjunctions (176). Noun stems with class information (4 959) and verb roots (5 984) for the Xhosa lexicon were extracted from various recent prototype paper dictionaries whereas relative stems (27), adjective stems (17), ideophones (30) and conjunctions (28) were only included as representative samples at this stage.

The most obvious difference between the two word root lexicons is the sparse coverage of nouns for Xhosa. A typical shortcoming in the current Xhosa lexicon is limited class information for noun stems.

Observations are firstly occurrences of shared noun stems (mainly loan words) but different class information, typically class 5/6 for Zulu versus class 9/10 for Xhosa, for example

'box' *-bhokisi* (Xhosa 9/10; Zulu 5/6)
'duster' *-dasta* (Xhosa 9/10; Zulu 5/6)
'pinafore' *-fasikoti* (Xhosa 9/10; Zulu 5/6).

It should be noted that although a Xhosa noun stem may be identical to its Zulu counterpart, analysis is not possible if the class prefix differs from the specified Zulu class prefix + noun stem combination in the morphotactics component of the analyser.

A second observation is identical noun stems with correct class information, valid for both languages, but so far only appearing in the Xhosa lexicon, for example

'number' *-namba* (Xhosa and Zulu 9/10)
'dice' *-dayisi* (Xhosa and Zulu 5/6).

This phenomenon occurs mainly with borrowed nouns that are more prevalent in the Xhosa lexicon than in the more outdated Zulu lexicon.

A closer look at the contents of the lexicons reveals that the two languages have the following in common: 1027 noun stems with corresponding class information, 1722 verb roots, 20 relative stems, 11 adjective stems, 10 ideophones and 9 conjunctions.

## 6 A computational approach to cross-linguistic similarity

This section discusses the extension of the bootstrapping procedure of the morphological analyser to include the use of the guesser variant of the morphological analyser. In addition the application of the bootstrapped morphological analyser to parallel test corpora is addressed. A variety of cross-linguistic effects is illustrated with examples from the corpora.

Even in languages where extensive word root lexicons are available, new word roots may occur from time to time. The Xerox toolkit makes provision for a **guesser variant** of the morphological analyser that uses typical word root patterns for identifying potential new word roots (Beesley and Karttunen, 2003:444). By exploiting the morphotactics and morphophonological alternations of the analyser prototype, the guesser is able to analyse morphologically valid words of which the roots match the specified pattern. Therefore, in cases where both the Zulu and Xhosa word root lexicons do not contain a root, the guesser may facilitate the bootstrapping process.

The extended **bootstrapping procedure** is schematically represented in Figure 1.

Since the available Zulu word list represents a rather outdated vocabulary, it is to be expected that the coverage of word roots/stems from a recent corpus of running Zulu text may be unsatisfactory, due to the dynamic nature of language. For example the word list contains no entry of the loan word *utoliki* 'interpreter' since 'interpreter' is rendered only as *i(li)humusha* 'translator', the traditional term derived from the verb stem *–humusha* 'to translate, interpret'. Provision therefore needs to be made for the constant inclusion of new roots/stems, be they newly coined, compounds or as yet unlisted foreign roots/stems.

Updating and refining the lexicon requires the availability of current and contemporary lan

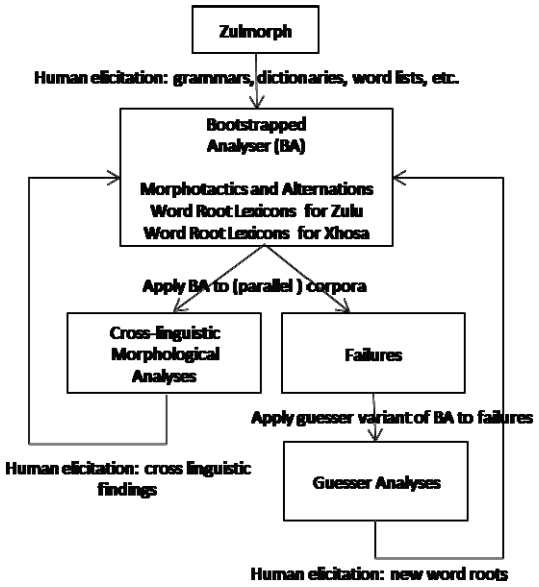


Figure 1. Bootstrapping procedure

guage resources in the form of **text corpora** as well as human intervention in the form of expert lexicographers or linguists to determine the eligibility of such words.

The language resources chosen to illustrate this point are parallel corpora in the form of the South African Constitution (The Constitution, (sa). The reason for the choice of these corpora is that they are easily accessible on-line, and it is assumed that the nature of the contents ensures accurate translations.

The **results** of the application of the bootstrapped morphological analyser to this corpus are as follows:

**Zulu Statistics**

Corpus size: 7057 types
Analysed: 5748 types (81.45 %)
Failures: 1309 types (18.55%)
Failures analysed by guesser: 1239 types
Failures not analysed by guesser: 70 types

**Xhosa Statistics**

Corpus size: 7423 types.
Analysed: 5380 types (72.48 %)
Failures: 2043 types (27.52%)
Failures analysed by guesser: 1772 types
Failures not analysed by guesser: 271 types

The output of the combined morphological analyser enables a detailed investigation into cross-linguistic features pertaining to the morphology of Zulu and Xhosa. The outcome of this investigation is illustrated by means of typical examples from the corpora. This provides novel insights into the investigation and exploitation of

cross-linguistic similarities and their significance for bootstrapping purposes, as shown in Figure 1.

Notational conventions include `[Xh]` for Xhosa specific morphemes, numbers indicate noun class information, e.g. `[NPrePre9]` tags the noun preprefix of a class 9 noun while `[RelConc8]` tags the relative concord of a class 8 noun.

**Examples from the Zulu corpus:**

The analysis of the Zulu word *ifomu* 'form' uses the Xhosa noun stem *–fomu* (9/10) in the Xhosa lexicon in the absence of the Zulu stem:

```
ifomu  i[NPrePre9]fomu[Xh][NStem]
```

The analysis of the Zulu word *ukutolikwa* 'to interpret' uses the Xhosa verb root *-tolik-* in the Xhosa lexicon:

```
ukutolikwa
u[NPrePre15]ku[BPre15]
tolik[Xh][VRoot]w[PassExt]a[VerbTerm]
```

**Examples from the Xhosa corpus:**

The analysis of the Xhosa words *bephondo* 'of the province' and *esikhundleni* 'in the office' use the Zulu noun stems *–phondo* (5/6) and *-khundleni* (7/8) respectively in the Zulu lexicon:

```
bephondo
ba[PossConc14]i[NPrePre5]li[BPre5]
phondo[NStem]

bephondo
ba[PossConc2]i[NPrePre5]li[BPre5]
phondo[NStem]

esikhundleni
e[LocPre]i[NPrePre7]si[BPre7]
khundla[NStem]ini[LocSuf]
```

The analysis of the Xhosa words *ekukhethweni* 'in the election' and *esihlonyelweyo* 'amended' use the Zulu verb roots *–kheth-* and *–hlom-* respectively in the Xhosa lexicon:

```
ekukhethweni
e[LocPre]u[NPrePre15]ku[BPre15]
kheth[VRoot]w[PassExt]a[VerbTerm]
ini[LocSuf]

esihlonyelweyo
esi[RelConc7]hlom[VRoot]el[ApplExt]
w[PassExt]e[VerbTermPerf]yo[RelSuf]
```

Ideophones used from the Zulu lexicon are:

```
ga[Ideoph]    qho[Ideoph]
sa[Ideoph]    tu[Ideoph]
ya[Ideoph]
```

Relative stems used from the Zulu lexicon are:

```
mandla[RelStem]
mdaka[RelStem]
njalo[RelStem]
mcimbi[RelStem]
```

Conjunctions used from the Zulu lexicon are:

```
futhi[Conj]
ukuthi[Conj]
```

**Examples of the guesser output from the Zulu corpus:**

The compound noun *-shayamthetho* (7/8) 'legislature' is not listed in the Zulu lexicon, but was guessed correctly:

```
isishayamthetho
i[NPrePre7]si[BPre7]
shayamthetho-Guess[NStem]
```

The following are two examples of borrowed nouns (*amabhajethi* 'budgets' and *amakhemikali* 'chemicals') not in the Zulu lexicon, but guessed correctly:

```
amabhajethi
a[NPrePre6]ma[BPre6]
bhajethi-Guess[NStem]

amakhemikali
a[NPrePre6]ma[BPre6]
khemikali-Guess[NStem]
```

The borrowed verb root *-rejest-* 'register' is not listed in the Zulu lexicon, but was guessed correctly:

```
ezirejestiwe
ezi[RelConc8]rejest-Guess[VRoot]
iw[PassExt]e[VerbTermPerf]

ezi[RelConc10]rejest-Guess[VRoot]
iw[PassExt]e[VerbTermPerf]
```

The relatively small number of failures that are not analysed by the guesser and for which no guessed verb roots or noun stems are offered, simply do not match the word root patterns as specified for Zulu and Xhosa in the analyser prototype, namely

```
[C (C (C)) V]+ C (C (C))
```

for verb roots and

```
[C (C (C)) V]+ C (C (C)) V
```

for noun stems. The majority of such failures is caused by spelling errors and foreign words in the test corpus.

## 7 Conclusion and Future Work

In this paper we focused on two aspects of cross-linguistic similarity between Zulu and Xhosa, namely the morphological structure (morphotactics and alternation rules) and the word root lexicons.

Regarding the morphological structure only differences between Zulu and Xhosa were added.

Therefore, Zulu informed Xhosa in the sense that the systematically developed grammar for ZulMorph was directly available for the Xhosa analyser development, which significantly reduced the development time for the Xhosa prototype compared to that for ZulMorph.

Special attention was also given to the so-called "open" class, which represents the word root lexicons for specifically nouns and verbs. The acquisition and coverage of these lexicons proved to be crucial for the success of the analysers under development. Since we were fortunate in having access to word root lexicons for both Zulu and Xhosa we included what was available in such a way that word roots could be shared between the languages. Here, although to a lesser extent, Xhosa also informed Zulu by providing a current (more up to date) Xhosa lexicon. In addition, the guesser variant was employed in identifying possible new roots in the test corpora, both for Zulu and for Xhosa.

In general it is concluded that bootstrapping morphological analysers for languages that exhibit significant structural and lexical similarities may be fruitfully exploited for developing analysers for lesser-resourced languages.

Future work includes the application of the approach followed in this work to the other Nguni languages, namely Swati and Ndebele (Southern and Zimbabwe); the application to larger corpora, and the subsequent construction of stand-alone versions. Finally, the combined analyser could also be used for (corpus-based) quantitative studies in cross-linguistic similarity.

## Acknowledgements

This material is based upon work supported by the South African 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.

## References

Beesley, K.R. and Karttunen, L. 2003. *Finite State Morphology*. CSLI Publications, Stanford, CA.

Bosch, S., Pretorius, L., Podile, K. and Fleisch, A. 2008. Experimental fast-tracking of morphological analysers for Nguni languages. *Proceedings of the 6th International Conference on Language Resources and Evaluation,* Marrakech, Morocco. ISBN 2-9517408-4-0.

De Pauw, G. and de Schryver, G-M. 2008. Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. *Lexikos 18* (AFRILEX-reeks/series 18: 2008): 303–318.

Doke, C.M. and Vilakazi, B.W. 1964. *Zulu–English Dictionary*. Witwatersrand University Press, Johannesburg.

Fromkin, V., Rodman, R. and Hyams, N. 2007. *An Introduction to Language*. Thomson Heinle, Massachusetts, USA.

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.

Koskenniemi, K. 1997. Representations and finite-state components in natural language, in *Finite-State Language Processing* E. Roche and Y. Schabes (eds.), pp. 99–116. MIT Press, Boston.

Karttunen, L. 2001. Applications of finite-state trransducers in natural language processing, in *Implementation and application of automata*, S. Yu and A. Paun (eds.). Lecture Notes in Computer Science, 2088:34-46. Springer, Heidelberg.

Louw, J.A., Finlayson, R. and Satyo, S.C. 1984. *Xhosa Guide 3 for XHA100-F*. University of South Africa, Pretoria.

Meinhof, C. 1932. *Introduction to the phonology of the Bantu languages*. Dietrich Reimer/Ernst Vohsen, Berlin.

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

Pretorius, L. and Bosch, S.E. 2002. Finite state computational morphology: Treatment of the Zulu noun. *South African Computer Journal*, 28:30-38.

Pretorius, L. and Bosch, S.E. 2003. Finite state computational morphology: An analyzer prototype for Zulu. *Machine Translation – Special issue on finite-state language resources and language processing*, 18:195-216.

The Constitution. (sa). [O]. Available: http://www.concourt.gov.za/site/theconstitution/thetext.htm. Accessed on 31 January 2008.

Yli-Jyrä, A. 2005. Toward a widely usable finite-state morphology workbench for less studied languages — Part I: Desiderata. *Nordic Journal of African Studies*, 14(4): 479 – 491.