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

Summary

- An existing Zulu morphological analyser prototype (ZulMorph) serves as basis for the bootstrapping of 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.

Background on ZulMorph

Xerox finite-state toolkit:
lexc for modelling the morphotactics;
xfst (regular expression language) for modelling morphophonological alternations
Word roots include 15 800 nouns, 7 600 verbs, 408 relatives, 47 adjectives, 2 735 ideophones, 176 conjunctions.

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

Morphotactics

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.

We focus on Xhosa affixes that differ from their Zulu counterparts. A few examples are given in Table 2.

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 2ps, Class 1 & 3 Class 4 & 9	*ngo-* *wo-* *yo-*	*ndo-* *uyo-* *iyo-*

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

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.

Morphophonological alternations

Differences in morphophonological alternations between Zulu and Xhosa are exemplified in Table 3.

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

Table 3. Examples of variations in Zulu and Xhosa morphophonology

The word root lexicons

Zulu lexicon:

- Based on an extensive word list dating back to the mid 1950s, 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).

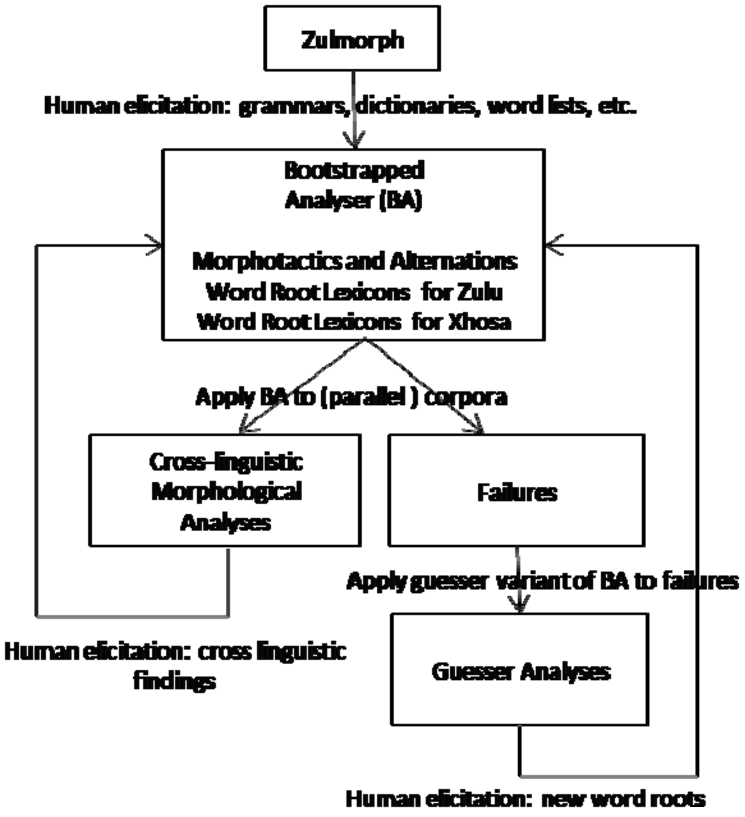
Xhosa lexicon:

- Noun stems with class information (4 959) and verb roots (5 984) 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.

A computational approach to cross-linguistic similarity

- Application of the bootstrapped morphological analyser to parallel test corpora
- Guesser variant of the morphological analyser that uses typical word root patterns for identifying potential new word roots

Bootstrapping procedure


Zulmorph
Human elicitation: grammars, dictionaries, word lists, etc.
Bootstrapped Analyser (BA)
Morphotactics and Alternations
Word Root Lexicons for Zulu
Word Root Lexicons for Xhosa
Apply BA to (parallel) corpora
Cross-linguistic Morphological Analyses
Failures
Apply guesser variant of BA to failures
Human elicitation: cross linguistic findings
Guesser Analyses
Human elicitation: new word roots

The language resources chosen to illustrate this point are parallel corpora in the form of the South African Constitution.

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

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]
```

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]

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

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]
```

Conclusion and Future Work

- 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.
- 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.
- 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.