Nordic Journal of African Studies 17(2): 66–88 (2008)

Experimental Bootstrapping of Morphological Analysers for Nguni Languages

SONJA BOSCH
University of South Africa
LAURETTE PRETORIUS
University of South Africa and Meraka Institute, CSIR
AXEL FLEISCH
University of South Africa and University of Helsinki

ABSTRACT

This paper addresses the experimental bootstrapping of the development of broad-coverage finite-state morphological analysers for Xhosa, Swati and (Southern) Ndebele by using an existing prototype of a morphological analyser for Zulu. These languages are both morphologically complex and resource-scarce. The research question is whether bootstrapping is feasible across the language boundaries between these closely related varieties. The objective is an assessment of the recognition rates yielded by the Zulu morphological analyser for the three related languages. The strategy is to use bootstrapping techniques that consist of the following steps: applying the analyser to corpus data from all languages, identifying (types of) failures, and implementing the respective changes in the analyser. The results show that the high degree of shared typological properties and formal similarities among the Nguni varieties warrants a modular bootstrapping approach. Word forms in these languages that were recognized by the Zulu analyser were mostly adequately analysed. Therefore, the focus lies on providing the necessary adaptations based on an analysis of the failure output for each language. As a result, the development of analysers for Xhosa, Swati and Ndebele is considerably faster than the creation of the Zulu prototype. The paper concludes with comments on the feasibility of the experiment, and the results of the evaluation.[1]

Keywords: *Nguni languages, broad-coverage finite-state morphological analysis, agglutinating morphological structure, resource-scarce languages.*

1. INTRODUCTION

The development of natural language processing (NLP) components is resource-intensive and therefore justifies exploring ways of reducing development time and effort when building NLP components. Rule-based approaches usually require the writing of large numbers of language grammar rules while statistical

---

[1] Published with the permission of ELRA. An earlier version of this paper was published within the proceedings of the LREC 2008 Conference.

and machine-learning methods are based on large amounts of annotated data, or even larger amounts of raw data, the acquisition of which is a non-trivial task. It is therefore justified to explore any and all possible ways of reducing development time and effort when building NLP components. This is arguably more relevant in the case of resource-scarce languages. The languages discussed in this paper are not only resource-scarce in terms of funding resources and human resources, but most important, also in terms of language resources, such as exhaustive linguistic description, corpora and machine-readable lexicons, language processing tools and applications. In short, data sparseness and language coverage pose definite challenges in and for the technological development of these languages.

In this paper we discuss the experimental bootstrapping of the development of broad-coverage finite-state (i.e. rule-based) morphological analysers for a group of languages belonging to the South-eastern Bantu zone, namely the Nguni languages Xhosa, Swati and (Southern) Ndebele by using an existing prototype of a morphological analyser for yet another Nguni language, Zulu. These languages are characterized, among others, by their rich morphological structure. The pertinent research question, therefore, is whether the existing morphological analyser prototype for Zulu (ZulMorph) may be used effectively for bootstrapping the development of morphological analysers for the other three mentioned languages. All four languages are very closely related. Their internal maximum linguistic distance compares, for instance to that of Galician and Portuguese, or that of language varieties belonging to the Low Franconian and Low German clusters.

We are confident that the linguistic relatedness of the Nguni languages may be systematically exploited, and the expectation is that useful results and benefits will be forthcoming, in particular, that the development time of morphological analysers for Xhosa, Swati and Ndebele may be significantly reduced without compromising their accuracy.

It is worth mentioning that an approach often used in this context is bootstrapping, an iterative technique of using a tool, in this case an NLP tool, to enhance itself. The bootstrapping of morphological analysers, specifically, is addressed by Oflazer and Nirenburg (1999) and Oflazer et al. (2001). In our context bootstrapping should be understood as follows: The existing Zulu analyser is viewed as the starting point for the bootstrapping process. It represents a first rudimentary morphological analyser for the other closely related languages. By means of systematic and stepwise enhancement improved analysers for Xhosa, Swati and Ndebele are developed.

The structure of the paper is as follows: In section II, the morphological complexity of the Nguni languages is briefly discussed. Section III is devoted to a short discussion of the various approaches in the computational morphological analysis of the Bantu languages. The purpose of this section is to place in context the supervised approach that is used in this paper. Section IV introduces the general approach with attention to the role of the small parallel development corpus. In section V this general approach is unpacked as a sequence of steps in

which the baseline analyser, ZulMorph, is applied and then systematically extended to include the morphology of the other languages. The extensions concern the word roots lexicon, followed by the grammatical morpheme lexicons and finally by the appropriate morphophonological rules. The discussion of each extension includes statistics about the failures and the successful analyses obtained, as well as an interpretation of these first results. Having constructed prototype analysers for Xhosa, Swati and Ndebele, the question as to their accuracy and validity then arises. Section V, therefore, focuses on the application of the four analysers to larger parallel test corpora. As before, statistics about the failures and the successful analyses are given and the results discussed. Future research and development directions for the morphological analysers under discussion are mapped out. In the final section we reflect on the feasibility and suitability of the approach and draw a number of conclusions concerning the bootstrapping of NLP components for these languages in general.

2. Approaches to Computational Morphology of the Bantu Languages

Computational morphological analysis can be divided broadly into two main themes: rule-based approaches and approaches that are based on some form of machine learning. These approaches may be viewed as complementary and are often successfully combined to form (real) hybrid systems in which the advantages of both approaches are exploited.

The question now is: What is the status quo for the Bantu languages? In terms of rule-based systems Swahili is, as far as we know, the first Bantu language for which a *complete* morphological analyser has been developed (Hurskainen 1992). The results of the Swahili analyser have since been used in various machine learning initiatives (de Pauw and Wagacha 2007; de Pauw and de Schryver 2008). Since the finite-state Zulu morphological analyser ZulMorph, as discussed in section IV, is rule-based it 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. It is true that its development may be considered to have been somewhat time-consuming (approximately 3000 highly skilled person hours), but the benefits of accurate modeling and implementation will be forthcoming in all future work based on ZulMorph. These benefits are already seen in the bootstrapping process reported on in this paper where analysers of acceptable quality have been developed for Xhosa, Swati and Ndebele within a total of approximately 300 hours – on average 100 hours per language. We envisage using the (linguistically sound and accurate) results obtained by means of these analysers as training data in a variety of machine learning approaches across all constructions in the near future. The expectation

is that a hybrid system would result that would be usable in a wide range of NLP applications.

Machine learning techniques have been used successfully in building finite-state morphological analysers. The induction of a stem lexicon to be used in a finite-state morphological analyser for Portuguese is discussed in (de Lima 1998) while the automatic learning of formal rewrite or replace rules for morphographemic changes is the focus of Oflazer and Nirenburg (1999).

The machine learning of morphological analysis for the Bantu languages has, on the other hand, been focused on specific morphological constructions, mainly to build small proof-of-concept solutions. The novelty of these preliminary results seems to lie in applying well established machine learning techniques to the exotic, morphologically rich Bantu languages. As expected these approaches face the limitations imposed by the scarcity of language data in the form of either annotated or unannotated corpora, depending on whether the approach is supervised or unsupervised. The constructions that are selected to be learnt are often not particularly complex, for example the noun morphology. Verb morphology on the other hand, is significantly more complex and has not been addressed in any significant way. De Pauw and Wagacha (2007) report on "a proof-of-the-principle experiment" in which a list of noun prefixes for the Gĩkũyũ language was extracted while Shalonova (pers. comm.) focuses on the use of word-pairs to learn noun suffix morphology for Zulu. Indeed, these proof-of-concept approaches have yet to culminate in usable broad-coverage systems.

3. MORPHOLOGICAL COMPLEXITY

The greatest challenge for the morphological analysis of the Nguni languages lies with the nominal and the verbal morphology. The linguistic description of these areas is fairly comprehensive for Zulu (Doke 1973, Taljaard and Bosch 1988, Poulus and Msimang 1998), but rather scarce for the other languages (cf. Pahl 1978, du Plessis 1983 for Xhosa; Taljaard et al. 1991 for Swati; Jiyane 1994 for Ndebele). Note that no grammar book exists for the latter. Also in terms of formal analysis, most progress has been made with regard to Zulu (van der Spuy 2001, 2006, Buell 2005, Zeller 2005).

Nguni has a complex gender system which is formally marked by noun class prefixes. Lexical nominal roots are associated with one of these classes on the basis of semantic properties (although the system is now fully lexicalized) and carry the respective prefix. Typically, these prefixes consist of a consonant-vowel-sequence. With the exception of Swati, these can be preceded by a so-called augment, a preceding copy vowel that fulfils different grammatical functions, e.g. definiteness and specificity, and is subject to morphophonological processes such as vowel deletion and coalescence of average complexity. Compare the following Zulu example for illustration: *u-mu-ntu* (cl. 1) 'person',

a-ba-ntu (cl. 2) 'people' (cf. the cognates in Xhosa *umntu*, *abantu*; Swati *muntfu*, *bantfu*; Ndebele *umuntu*, *abantu*).

Noun class prefixes are operative to a certain degree. Grammatical number (sg./pl.) and notions such as count versus mass nouns are expressed by means of noun classes, and some noun classes, usually as disjunct morphemes in combination with specific endings, serve to derive new word forms. These word forms often entail a category-shift from verb to noun. The productivity of these mechanisms is, however, limited.

The morphological make-up of the verb phrase is considerably more complex. A number of slots, both preceding and following the verb root may contain numerous morphemes serving derivative functions, inflection for tense-aspect, marking of nominal arguments. Examples are cross-reference of the subject and direct object by means of class- (or person-/number-) specific object markers, locative clitics, morphemes distinguishing verb forms in clause-final and non-final position (conjoint/disjoint distinction relating to information structure of the clause), negation/polarity, etc. The degree of morphophonological complexity is, generally-speaking, average. Many of the phonological adjustments at morpheme boundaries are predictable and rule-based. There are exceptions such as the palatalization of certain consonant-vowel combinations. What is problematic about these is the fact that they represent instances of "morphologically-conditioned phonology". For instance, only the diminutive ending *-ana* triggers the palatalization of preceding alveolar consonants (in addition to preceding labials which undergo this change also in combination with certain other endings): *ikhanda* 'head' > diminutive *ikhanjana*, although no plain phonological rule as such would disallow the unpalatalized sequence **ikhandana*.

In summary, the rich agglutinating morphological structure, which characterizes the Nguni languages, is based on two principles, namely the nominal classification system, and the concordial agreement system. According to the nominal classification system, nouns are categorized by prefixal morphemes, which for analysis purposes have been put into classes and given numbers. These noun class prefixes bring about concordial agreement that links the noun to other words in the sentence such as verbs, adjectives, pronouns and so forth.

Fortunately, despite the complexities of these domains, they are comparable across language boundaries with a high degree of formal similarity.

4. COMPUTATIONAL APPROACH

The suitability of finite-state approaches to computational morphology has been shown convincingly (Koskenniemi 1997, Karttunen 2001, Beesley and Karttunen 2003) and has resulted in numerous software toolkits and development environments for this purpose. For the work reported on in this

paper the state-of-the-art Xerox finite-state toolkit (Beesley and Karttunen 2003) is used.

The Xerox software tool for modeling 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 words generated by this network are morphotactically well-formed, but still rather abstract lexical or morphophonemic words.

The morphophonological (phonological and orthographical) alternations are modeled 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 morphological analyser. It is note-worthy that these finite-state networks (transducers) are bi-directional devices, which facilitate morphological analysis in the one direction and morphological generation in the other. 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 and Bosch 2003, Bosch and Pretorius 2006).

What does ZulMorph offer in terms of the bootstrapping process? In the first place it provides the *morphotactics* component, i.e. the accurate specification of the Zulu word structure. The morphotactics component includes 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 Zulu. Word roots include nouns (15 800), verbs (7 600), relatives (408), adjectives (48), ideophones (2 735), conjunctions (176)[2]. Secondly it also offers the *morphophonological* (phonological and orthographical) *alternations* component, i.e. the changes (orthographic/spelling) that take place between the lexical and surface words when morphemes are combined to form new words/word forms, are described. This is summarized in table 1.

[2] Note the small number of adjectives. This is a common feature in Bantu languages. The Nguni languages in particular, have innovated a specific word class, relatives, which makes up for the functional deficiency caused by the lack of adjectives. Relatives are morphologically moderately complex (e.g. they agree in noun class with the head noun they modify). Ideophones are also a common feature in Bantu languages while being virtually inexistent in European languages. They are numerous, but usually morphologically simple. Therefore, they need to be included in the lexicon but are not that relevant for morphological analysis.

Morphotactics (morphological lexicon)	Affixes for all parts-of-speech (e.g. subject & object concords [=inflectional morphology serving to cross-reference nominal arguments on the verb], noun class prefixes, verb extensions etc.)	Word roots (e.g. nouns, verbs, relatives, adjectives, ideophones, conjunctions)	Rules for legal combinations and orders of morphemes (e.g. u-ya-ngi-thand-a and not *ya-u-a-thand-ngi)
Morphophono-logical 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. *ZulMorph Components.*

Some examples of the output of ZulMorph are:

```
ungesabi  u[SC1]nga[NegPre]esab[VRoot]i[VerbTermNeg]

emlonyeni e[LocPre]u[NPrePre3]mu[BPre3]lomo[NStem] ini[LocSuf]
```

5. BOOTSTRAPPING PROCEDURE

This section addresses the research question by performing experiments of increasing scope in order to assess the feasibility of the approach. The bootstrapping is done in various stages for the three additional Nguni languages, viz. Xhosa, Swati and Ndebele, in parallel. Although the process relies on a high level of automation, human intervention i.e. elicitation of linguistic information from humans, is essential in order to maintain linguistic accuracy.

It should be noted that the bootstrapping process scales well. Steps 1 and 2 depend on increasingly larger corpora that are used in the bootstrapping. These steps scale up in the following sense: Although the roots and stems in any language constitute a so-called open class and are subject to continuous maintenance and growth, the number of roots and stems at any given point remains finite. It is expected that the convergence of the bootstrapping process will depend on whether and how many new roots and stems will continue to be discovered in each successive iteration. Semi-automated approaches to mining larger corpora may include the application of a so-called guesser variant (Beesley and Karttunen 2003) of the morphological analyser under development. The necessity of scaling up will terminate with convergence, i.e. if and when the difference between two successive iterations becomes negligible in terms of new information added. A detailed investigation of this aspect forms part of future work. Steps 3 and 4 concern the so-called closed classes and the morphophonological alternation rules. Although instances of the closed classes may be discovered from corpora, these steps are not conceptually reliant on a (large) corpus and should therefore scale well.

5.1 STEP 1

The process used in the experiment starts by applying ZulMorph to a small manageable parallel corpus of Zulu, Xhosa, Swati and Ndebele running text respectively, converted into a wordlist of approximately 200 types (i.e. unique tokens) for each language. The analysis of the Zulu 200-type word list was perfected to 100% before the experiment commenced. The success rates of analysis for the other languages were: Xhosa 76.29%, Swati 64% and Ndebele 73.08%.

The types of failures encountered for the different languages were as follows:

- Xhosa Statistics:
  Analysed: 148 words (76.29 %)
  Failed: 46 words (23.71 %)
  Corpus size: 194 words

Verbs	Nouns	Relatives / adjectives	Pronouns	Conjunctions
andisayi *ndiyathanda* *zagxotha* *ukulumka* *ukutya*	*umntu* *iindlebe* *neendlebe* *impungutye* *umqhagi*	*omde* *ezininzi* *zikufutshane* *ezitsolo*	*ngaloo*	

Table 2. *Examples of failures in Xhosa (step 1).*

- Swati Statistics:
  Analysed: 128 words (64.00 %)
  Failed: 72 words (36.00 %)
  Corpus size: 200 words

Verbs	Nouns	Rel/adj	Prons	Conj
tabaleka *utawubona* *achachatela* *ngiyamati*	*liphupho* *tinsuku* *umuntfu* *lechudze* *netinja*	*umudze* *amnandzi*	*tonkhe* *lonkhe*	*futsi* *kodvwa*

Table 3. *Examples of failures in Swati (step 1).*

- Ndebele Statistics:
  Analysed: 133 words (73.08 %)
  Failed: 49 words (26.92 %)
  Corpus size: 182 words

Verbs	Nouns	Rel/adj	Prons	Conj
ubatjela *warhaba* *zatjeheja* *bekabhudanga*	*amezwi* *nabentwana* *iinkukhu* *ipungutjha* *umsilaso* *neenkukhu*	*amanengi* *ezijamileko*	*loke* *soke*	*nangabe* *khuyini* *ukobana*

Table 4. *Examples of failures in Ndebele (step 1).*

On the one hand, we have forms of verb roots such as (Xh) *-ty-* (eat) and *-gxoth-* (defeat); (Sw) *-chachathel-* (shiver) and *-at-* (know); (Nd) *-tjel-* (tell) and *-rhab-* (hurry), as well as forms of noun stems such as (Xh) *-mpungutye* (jackal) and *-qhagi* (rooster); (Sw) *-chudze* (rooster) and *-ntfu* (person); (Nd) *-pungutjha* (jackal) and *-kukhu* (fowl) which do not feature in the Zulu lexicon. Although the subject concords, verb terminatives and class prefixes concur with those in ZulMorph, these words fail to be analysed because of the missing roots/stems. The same applies to relative stems such as (Xh) *-ninzi* (many), (Sw) *-mnandzi* (pleasant) and the (Nd) adjective stem *-nengi* (many).

On the other hand, we find roots/stems that are identical to their Zulu counterparts, but whenever the prefixes or suffixes differ from the Zulu word structure as specified in the morphotactics component of the analyser, analysis is not possible yet. Examples are (Xh) *ndiyathanda* (I like), *umntu* (human being); (Sw) *tabaleka* (they ran away), *tinsuku* (days); and (Nd) *amezwi* (words) and *nabentwana* (with the children).

Based on the results of this experiment the process is continued by adding linguistic information.

5.2 STEP 2A

In step 2a, the word root lexicon of ZulMorph was enhanced firstly by the addition of an extensive Xhosa lexicon extracted from a prototype paper dictionary that includes noun stems (5 600), verb roots (6 066), relatives (26), adjectives (17), ideophones (30), conjunctions (28); secondly by applying regular Swati sound changes to the Zulu lexicon (i.e. noun stems, verb roots, relative stems and adjective stems). Such sound changes are shown in table 5.

do > dvo, du > dvu, dw > dvw
da > dza, de > dze, di > dzi
to > tfo, tu > tfu, tw > tfw
tho > tfo, thu > tfu, thw > tfw
ta > tsa, te > tse, ti > tsi
tha > tsa, the > tse, thi > tsi
za > ta, ze > te, zi > ti
tsh > tj

Table 5. *Regular sound changes between Zulu and Swati.*

Since no lexicon is available for Ndebele, the identification of Ndebele roots/stems still needs to rely on Zulu, Xhosa and Swati. Following the process described above, the results obtained were:

- Xhosa Statistics:
  Analysed: 172 words (88.66 %)
  Failed: 22 words (11.34 %)
  Corpus size: 194 words

Verbs	Nouns	Rel/adj	Prons	Conj
andisayi *ndiyathanda*	*umntu* *iindlebe* *neendlebe*	*omde* *zikufutshane* *ezitsolo*	*ngaloo*	

Table 6. *Examples of failures in Xhosa (step 2a).*

- Swati Statistics:
  Analysed: 166 words (83.00 %)
  Failed: 34 words (17.00 %)
  Corpus size: 200 words

Verbs	Nouns	Rel/adj	Prons	Conj
tabaleka *utawubona*	*liphupho* *tinsuku* *lichudze* *netinja*	*umudze*	*tonkhe* *lonkhe*	*futsi* *kodvwa*

Table 7. *Examples of failures in Swati (step 2a).*

- Ndebele Statistics:
  Analysed: 136 words (76.92 %)
  Failed: 46 words (23.08 %)
  Corpus size: 182 words

Verbs	Nouns	Rel/adj	Prons	Conj
warhaba *zatjheja* *bekabhudanga*	*amezwi* *iinkukhu* *ipungutjha* *umsilaso* *neenkukhu*	*amanengi* *ezijamileko*	*loke* *soke*	*nangabe* *khuyini* *ukobana*

Table 8. *Examples of failures in Ndebele (step 2a).*

From the statistics it becomes clear that Xhosa, Swati as well as Ndebele have an increased rate of analysis in this step. It is not surprising that with the addition of the extensive Xhosa lexicon and the regular sound changes towards the Swati lexicon, the success rate has increased dramatically, by approximately 12% and 19% respectively to reach 88.66% and 83%. As expected, the success rate of the Ndebele analysis has only increased marginally (by 1.65%). The

marginal increase can be ascribed to a verb root such as *–tjela* (tell) that Ndebele shares with Swati.

In order to maintain language specific information all Xhosa, Swati and Ndebele roots and stems are indexed by means of `[Xh]`, `[Sw]` and `[Nd]`respectively in the ZulMorph word root lexicon. These indices facilitate language specific analyses.

5.3 STEP 2B

For step 2b, all word root/stem lexicons were used as for step 2a, but were all expanded to include the missing roots for the 200 word corpus, i.e. verb roots, noun stems, relative stems, adjective stems etc. The reasoning behind this step was that once all roots had been included, a clearer picture would emerge concerning the other two aspects of the morphotactics component namely the prefixes and suffixes, as well as the valid combinations and orders of morphemes.

In line with expectations, there was no change in the Xhosa results since the root/stem lexicons had already been included in step 2a and no new roots had been identified. However, a significant increase in the success of analyses was recorded for Ndebele (8.8%).

- Xhosa Statistics:
  Analysed: 172 words (88.66 %)
  Failed: 22 words (11.34 %)
  Corpus size: 194 words

Verbs	Nouns	Rel/adj	Prons	Conj
andisayi *ndiyathanda*	*umntu* *iindlebe*	*zikufutshane*	*ngaloo*	

Table 9. *Examples of failures in Xhosa (step 2b).*

- Swati Statistics:
  Analysed: 167 words (83.50 %)
  Failed: 33 words (16.50 %)
  Corpus size: 200 words

Verbs	Nouns	Rel/adj	Prons	Conj
tabaleka *utawubona*	*liphupho* *tinsuku* *lichudze*		*tonkhe*	*futsi* *kodvwa*

Table 10. *Examples of failures in Swati (step 2b).*

- Ndebele Statistics:
  Analysed: 154 words (84.62 %)
  Failed: 28 words (15.38 %)
  Corpus size: 182 words

Verbs	Nouns	Rel/adj	Prons	Conj
bekabhudanga	*Amezwi* *iinkukhu* *ipungutjha* *umsilaso* *neenkukhu*	*ezijamileko*	*loke* *soke*	*nangabe* *khuyini* *ukobana*

Table 11. *Examples of failures in Ndebele (step 2b).*

It should be remembered that at this stage of the experiment, prefix and suffix morpheme structures still depend on the Zulu version of the analyser. Therefore, no matter whether the root is peculiar to a specific language or identical to the Zulu root, the analysis fails if the prefixes or suffixes do not conform to the Zulu equivalent. For instance, in the following examples the roots are identical to their Zulu counterparts. However, in the Xhosa *umntu* (a human being) the class prefix *um-* differs from the Zulu *umu-*, that has been modeled in the analyser for monosyllabic noun stems in class 1; in the Swati *liphupho* (a dream) the class prefix *li-* differs from the Zulu *i-* as has been modeled in the analyser for polysyllabic noun stems in class 5; and in the Ndebele *amezwi* (words) the class prefix *ame-* differs from the Zulu *ama-*, as has been modeled in the analyser for polysyllabic noun stems in class 6.

5.4 STEP 3

Step 3 consisted of adding to the morphological analyser "closed" class information (morphotactics) for Xhosa, Swati and Ndebele, such as noun prefixes, subject concords, object concords, relative concords, absolute, quantitative and demonstrative pronouns, demonstrative copula, conjunctives, ideophones, adjective stems and concords. As in the case of roots and stems all non-Zulu language specific morphemes are indexed by means of either `[Xh]`, `[Sw]` or `[Nd]`.

The experiment resulted in bringing the three additional languages to just over 90% success in each case.

- Xhosa Statistics:
  Analysed: 181 words (93.30 %)
  Failed: 13 words (6.70 %)
  Corpus size: 194 words

Verbs	Nouns	Rel/Adj	Prons	Conj
	umntu *iindlebe* *neendlebe*	*zikufutshane*		

Table 12. *Examples of failures in Xhosa (step 3).*

- Swati Statistics:
  Analysed: 183 words (91.50 %)
  Failed: 17 words (8.50 %)
  Corpus size: 200 words

Verbs	Nouns	Rel/Adj	Prons	Conj
batawubona	*liphupho* *netinja* *tinsuku*			

Table 13. *Examples of failures in Swati (step 3).*

- Ndebele Statistics:
  Analysed: 166 words (91.21 %)
  Failed: 16 words (8.79 %)
  Corpus size: 182 words

Verbs	Nouns	Rel/adj	Prons	Conj
bekabhudanga	*amezwi* *iinkukhu* *umsilaso* *neenkukhu*	*ezijamileko*		

Table 14. *Examples of failures in Ndebele (step 3).*

It is significant that the failures clearly indicated the need for attention to rules that determine the form of morphemes, more specifically class prefixes, as addressed in step 4.

In addition, the failures reveal language specific morphological differences that need to be modelled separately in the morphotactics component (morphological lexicon). For instance, in the case of Swati *batawuthula* (they will be quiet), the future tense construction *–tawu-*, and in the case of Ndebele *bekabhudanga* (he had been dreaming), the continuous tense prefix construction (*be-ka-*) need to be included in the morphological lexicon. Two other examples in Ndebele are *ezijamileko* (that are sharp) and *umsilaso* (his tail). In the first example the suffix *-ko* (relative) differs from the Zulu *-yo*; while *-so* in the second example indicates a possessive construction that differs considerably from the other Nguni languages with regard to morpheme order, and therefore needs to be modeled separately. This will form part of future work.

5.5 STEP 4

Step 4 concerns the extension and refinement of the morphophonological alternation rules. For the purposes of this experiment, the focus is on rules that apply to classes 9 and 10. After describing these rules for Zulu, we indicate how the comparable rules for the other languages deviate from those for Zulu. Examples of such morphonological alternation rules, as well as analyses are given.

The class 9/10 (Singular/Plural, *in/izin*) rules for Zulu are given in the form **Preprefix + Basic prefix + Noun stem > Surface noun.**

- Zulu Rule 1:
  i + n/zin + dlozi > indlozi/izindlozi

  i + n/zin + ja > inja/izinja

- Zulu Rule 2:

The *n* of the basic prefix changes to *m* before labial sounds *b*, *p*, *f*, *v*:

i + m/zim + philo > impilo/izimpilo

i + m/zim + fundo > imfundo/izimfundo

i + m/zim + vula > imvula/izimvula

i + m/zim + bhuzi > imbuzi/izimbuzi

- Zulu Rule 3:

Aspiration (*h*) is removed when n is followed by *kh*, *ph*, *th*, *bh*:

i + m/zim + bhuzi > imbuzi/izimbuzi

i + n/zin + tho > into/izinto

i + m/zim + philo > impilo/izimpilo

i + n/zin + khonzo > inkhonzo/izinkonzo

In Xhosa class 9 only *i* is used before stems beginning with *h*, *i*, *m*, *n*, *ny*, e.g.:

Zulu: `inkambo i[NPrePre9]n[BPre9]hambo[NStem]`

Xhosa: `ihambo i[NPrePre9]hambo[NStem]`

In Swati class 9 aspiration (*h*) remains when n is followed by *kh*, *ph*, e.g.

Zulu: `inkulumo i[NPrePre9]n[BPre9]khulumo[NStem]`

Swati: `inkhulumo i[NPrePre9]n[BPre9]khulumo[NStem]`

In Ndebele class 9 only *i* is used before stems beginning with *p*, *k*, *hl*, *h*, *f*, *s*, *tj* (if the root consists of more than one syllable), e.g.

Zulu: `inkuku i[NPrePre9]n[BPre9]khuku[NStem]`

Ndebele: `ikukhu i[NPrePre9]kukhu[NStem]`

In Xhosa and Ndebele the (surface) class 10 class prefix *iin* before polysyllabic stems needs to be made provision for, since in the case of Zulu, *izin* occurs before monosyllabic as well as polysyllabic stems.

The following rule, related to class 10, was implemented as well, viz.

Zulu: Consonant + *a* + *izin* > Consonant + *ezin*
Xhosa and Ndebele Consonant + *a* + *iin* > Consonant + *een*;
Swati Consonant + *a* + *tin* > Consonant + *etin*, e.g.

Zulu: `nezinja` `na[AdvPre]i[NPrePre10]zin[BPre10]ja[NStem]`

Xhosa: `neenkuku` `na[AdvPre]i[NPrePre10]zin[BPre10]kuku[NStem]`

Ndebele: `neendlebe` `na[AdvPre]i[NPrePre10]zin[BPre10]dlebe[NStem]`

Swati: `netinja` `na[AdvPre]tin[BPre10]ja[NStem]`

The xfst implementation of the above rules is outlined by means of examples. Morphophonological rules are often combined with so-called auxiliary rules, which are introduced to ensure the correct firing and sequencing of the morphophonological rules. We illustrate this by means of the following: The notation `%^YY` denotes a multi-character symbol in xfst, introduced in the morphological lexicon as `^YY` to mark a particular morpheme yy (for example) for use in the rule modeling. These symbols are used in managing alternations and their contexts. Once the symbol has played its discriminatory strategic role, another auxiliary rule is used to eventually remove the symbol or replace it with a string in the surface language.

A particular example is `%^ZINXh` in the xfst fragment for the Xhosa class 10 rule in figure 1, which is realized as either *zi*, *zim* or *zin*, depending on the context. While a detailed explanation of the xfst syntax falls outside the scope of this article (see Beesley and Karttunen 2003), we explain the notation used in figures 1, 2 and 3. The symbol `%` is used to literalize `^`, `||` indicates context in the xfst replace rules, `.o.` is rule composition, `|` is the choice operator, and `+` and `*` are the Kleene plus and star operators. Specific multi-character symbols

used, are %^ZINXh, %^XX, %^BR and %^ER. The general form of the replacement rule, used in the examples, is

```
A -> B || L _ R;
```

which means that every string of language A is replaced by all strings of language B if and only if it occurs between a string of language L to the left and a string of language R to the right.

```
define Syllable [Cons+ Vowel Cons* | Vowel Cons*];
...
define ruleizin1Xh %^ZINXh -> %^XX %^ZINXh || _ [%^BR Syllable
Syllable %^ER | %^BR Syllable %^ER [Vowel | Syllable] | %^BR
Syllable Syllable];
define ruleizin2Xh %^ZINXh ->  z i || _ %^BR [h | l | m | n |
n y];
define ruleizin3Xh %^ZINXh ->  z i m || _ %^BR [p | b | f |
v];
define ruleizin4Xh %^ZINXh -> z i n;
define ruleizinXh ruleizin1Xh .o. ruleizin2Xh .o. ruleizin3Xh
.o. ruleizin4Xh;
```

Fig. 1. *Fragment of xfst scipt for Xhosa class 10 rule.*

Since the vowel combination *ii* does not occur in Zulu, special care should also been taken to preserve the vowel combination *ii* in Xhosa and Ndebele. In the implementation of the Xhosa class 10 rule in figure 1 an auxiliary symbol %^XX is introduced to prevent the rule for Zulu vowel combinations (figure 2) to change *ii* to *i*. The symbol %^XX is eventually (after the Zulu rule in figure 2 was allowed to fire) removed by auxiliary rules. This highlights another important issue namely the order in which rules are allowed to fire. For example, one of the last Zulu rules to fire is the rule that takes care of vowel combinations, as shown in figure 2.

```
define VowelCombs1 a a -> a , a e -> e ,
                   a i -> e , a o -> o ,
                   a u -> o , e a -> e ,
                   e i -> e , e u -> e ,
                   i i -> i , u a -> a ,
                   u o -> o , u u -> u;
```

Fig. 2. *Zulu rule for vowel combinations.*

The Xhosa and Ndebele rules in figure 3 are only allowed to fire after the rule in figure 2 in order to preserve *ii*.

```
define VowelCombs1XhNd %^XX z -> %^XX || [%^IXh | %^INd] _ i;

define VowelCombs2XhNd a [%^IXh | %^INd] %^XX i -> e %^XX e || Cons _;

define VowelCombsXhNd VowelCombs1XhNd .o. VowelCombs2XhNd;

define ruleXX %^XX -> [. 0 .];

define ruleIXhNd [%^IXh | %^INd] -> i;
```

Fig. 3. *Xhosa and Ndebele rules to preserve ii and ee.*

The extension and refinement of rules in the ZulMorph rule component, implemented in xfst and described in step 4, results in:

- Xhosa Statistics:
  Analysed: 189 words (97.42 %)
  Failed: 5 words (2.58 %)
  Corpus size: 194 words

Verbs	Nouns	Rel/Adj	Prons	Conj
		zikufutshane		

Table 15. *Examples of failures in Xhosa (step 4).*

- Swati Statistics:
  Analysed: 195 words (97.50 %)
  Failed: 5 words (2.50 %)
  Corpus size: 200 words

Verbs	Nouns	Rel/Adj	Prons	Conj
utawubona	*liphupho*			

Table 16. *Examples of failures in Swati (step 4).*

- Ndebele Statistics:
  Analysed: 171 words (93.96 %)
  Failed: 11 words (6.04 %)
  Corpus size: 182 words

Verbs	Nouns	Rel/Adj	Prons	Conj
bekabhudanga	*amezwi umsilaso*	*ezijamileko*		

Table 17. *Examples of failures in Ndebele (step 4).*

A slight but steady increase in the success rates for all three languages is evident.

6. PRELIMINARY EVALUATION

The preliminary evaluation is based on the use of parallel test corpora of approximately 7000 types each for the four languages taken from a domain different to the development corpus (The Constitution, s.a.). The results obtained are as follows:

- Zulu Statistics:
  Analysed: 5653 words (80.68 %)
  Failed: 1354 words (19.32 %)
  Corpus size: 7007 words

- Xhosa Statistics:
  Analysed: 5250 words (71.10 %)
  Failed: 2134 words (28.90 %)
  Corpus size: 7384 words

- Swati Statistics:
  Analysed: 3971 words (58.26 %)
  Failed: 2845 words (41.74 %)
  Corpus size: 6816 words

- Ndebele Statistics:
  Analysed: 3994 words (58.96 %)
  Failed: 2780 words (41.04 %)
  Corpus size: 6774 words

In comparison to the results of the development corpus, the success rates for the four languages in the test corpora decreased between 20% and 40%. This can be ascribed among others to "new" roots including newly coined terms and loan words, which are not yet included in the lexicon. Examples in the case of Zulu are *-bhajethi* (budget), *-komidi* (committee), etc. An orthographic discrepancy also contributes to failures in the Swati corpus in the sense that certain demonstrative pronouns in Swati are written conjunctively with the noun, as opposed to the disjunctive orthographic treatment in the case of Zulu. For instance in Swati *lelilungelo* (this right) occurs as *leli lungelo* (this right) in Zulu.

A summary of the improvement of the morphological analysers across the three additional Nguni languages in the bootstrapping process as described so far, is illustrated in figure 4. The preliminary evaluation based on larger parallel test corpora is indicated in the last column (6).

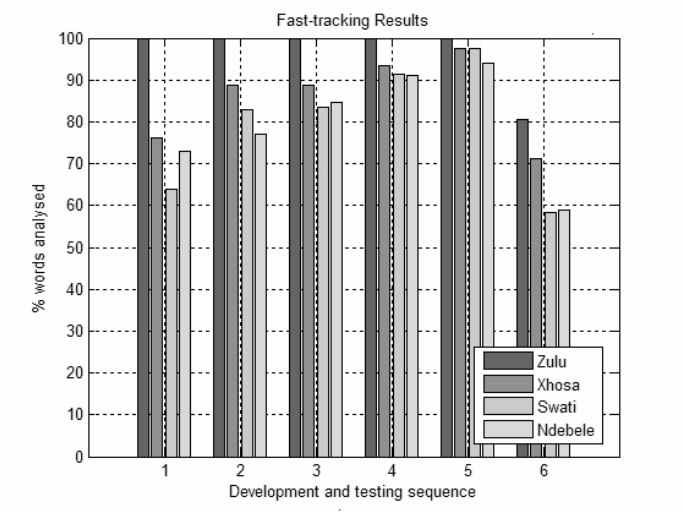

Fig. 4. *Results of development and testing sequence.*
(1=step 1; 2=step 2a; 3=step 2b; 4=step 3; 5=step 4; 6=preliminary evaluation)

Future work regarding the morphological analysers entails systematically scaling up and refining all aspects addressed in the experiment, both with respect to similarities and differences between the various languages. Step 2a clearly shows a marked improvement in the Xhosa morphological analyser after addition of an extensive lexicon. The aim is to follow the same approach for Swati and Ndebele, namely refining the improvised Swati lexicon and adding a Ndebele lexicon. The combined lexicon increases recognition considerably. Items in this comprehensive lexicon are indexed for the specific languages so that accuracy is maintained. This is important in order to ensure that the capacity of the lexicon to analyse language-specific forms is not lost.

In step 2b, where missing roots/stems were added, there was a good improvement for Ndebele, which proves the importance of the lexicon. As in the case of steps 3 and 4, we intend to follow the same procedure as with the Zulu 200 type corpus, that is adding morphological information to the morphological lexicon, and adapting rules in a systematic manner.

Certain areas in the grammar of the individual languages need to be modeled independently and then built into the analyser as an additional component, such as the formation of copulatives. Ndebele copula constructions for instance, differ substantially from the mechanism applicable in Zulu (and the other Nguni languages).

Additions and corrections are then fed back into the analyser on an iterative basis. Once the rate of recognition and accuracy has reached 100% for the various 200 type corpora, the test corpus will be gradually increased to cover more so-called "new" constructions. Even more important, language-specific requirements will be identified by going through the inventory of recognition failures of step 4. The promising results obtained therefore suggest the extension of the approach to larger corpora, which will also stimulate the development of

basic language resources in the form of word root lists, machine-readable lexicons and language corpora for these languages.

7. CONCLUSION

In conclusion, we return to the research question, namely whether the existing morphological analyser prototype for Zulu may be used effectively for bootstrapping the development of accurate, usable broad-coverage morphological analysers for the other three Nguni languages. On the basis of the experiment reported on in this article we conclude the following:

- There are obvious benefits with regard to development time. Taking into consideration that the development of the Zulu analyser prototype required approximately 3000 highly skilled person hours, whereas experiments with regard to the current development of Xhosa, Swati and Ndebele analyser prototypes took only about 300 hours in total to develop.
- Preliminary results are promising as has been illustrated. A systematic assessment and validation of the analyses and also of the linguistic accuracy and coverage of the various analysers are in progress.
- The unified approach to the development of these four morphological analysers has significant advantages in terms of optimizing the software process for the further development of these software artefacts. All the phases of the software life cycle, including linguistic design and modeling, implementation, testing, documentation, verification, validation, maintenance and improvement, will benefit. Besides the benefits in terms of maintenance, this has another advantage over fully independently developed analysers. Code-switching and extensive borrowing between languages are common phenomena in the Nguni languages. The compatibility that results from this unified approach may allow easier integration of these components.
- In order to retain the benefits of the unified development approach in maintaining the analysers, we envisage the design of an automated procedure for extracting a language specific morphological analyser on demand if and when required for a specific application.
- By exploiting correspondences and linguistic relatedness, more effort may be spent on those aspects in which the languages differ, ensuring end products of superior quality, both linguistically and computationally.
- Since this approach proved to be successful, it can in future also be used for the development of other tools for these resource-scarce, data sparse languages. Moreover, the approach followed in this paper in principle applies to any group of languages that exhibits systematic similarities for example, the Sotho group of Bantu languages consisting of Northern and Southern Sotho and Tswana. It provides a structured way of exploiting all the available linguistic knowledge before exploring the additional benefits that machine learning approaches offer.

ACKNOWLEDGMENT

The authors would like to acknowledge the assistance rendered by the following persons with regard to the preparation of the test corpora and grammaticality judgements: Buti Skhosana and Sponono Mahlangu (Ndebele), Kholisa Podile (Xhosa) and Basie Taljaard (Swati); as well as the constructive feedback of anonymous reviewers.

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 L. Karttunen 2003.
Finite state morphology. Stanford, CA: CSLI Publications.

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

Buell, L. C. 2005.
Issues in Zulu Verbal Morphosyntax. Ph.D. dissertation, University of California, Los Angeles. Available: http://www.fizzylogic.com/users/bulbul/school/buell-dissertation-single-space.pdf.

de Lima, E. F. 1998.
Induction of a stem lexicon for two-level morphological analysis. In: D. M. W. Powers (ed.), *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning*, ACL, pp. 267–268.

de Pauw, G. and G-M. De Schryver 2008.
Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. **Lexicos** 18: 303–318.

de Pauw, G., and P. W. Wagacha 2007.
Bootstrapping morphological analysis of Gĩkũyũ using unsupervised maximum entropy learning. In: *Proceedings of the eighth INTERSPEECH conference*, Antwerp, Belgium, 2007.

Doke, C. M. 1973.
Textbook of Zulu grammar. Cape Town: Maskew Miller Longman.

du Plessis, J. A. 1983.
Isixhosa. Goodwood: Oudiovista.

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.

Jiyane, D. M. 1994.
Aspects of isiNdebele grammar. Unpublished M.A. thesis, University of Pretoria.

Karttunen, L. 2001.
Applications of finite-state transducers in Natural Language Processing. In: S. Yu and A. Paun (eds.), *Implementation and application of automata*, pp. 34–46. Lecture Notes in Computer Science, vol. 2088. Heidelberg: Springer.

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

Oflazer, K. and S. Nirenburg 1999.
Practical Bootstrapping of Morphological Analyzers. In: *Proceedings of the Workshop on Computational Natural Language Learning at EACC '99*, Bergen, Norway, 1999.

Oflazer, K., S. Nirenburg, and M. McShane 2001.
Bootstrapping morphological analyzers by combining human elicitation and machine learning. **Computational Linguistics** 27(1): 59–85.

Pahl, H. W. 1978.
Isixhosa. King Williams Town: Thandapers.

Poulos, G. and C. T. Msimang 1998.
A linguistic analysis of Zulu. Cape Town: Via Afrika.

Pretorius, L. and S. E. Bosch 2003.
Finite-state computational morphology: an analyzer prototype for Zulu. **Machine Translation** 18: 195–216.

Shalonova, K. Personal communication, 2008.

Taljaard, P. C. and S. E. Bosch 1988.
Handbook of IsiZulu. Pretoria: Van Schaik.

Taljaard, P. C., J. N. Khumalo, and S. E. Bosch 1991.
Handbook of SiSwati. Pretoria: Van Schaik.

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

van der Spuy, A. 2001.
Grammatical structure and Zulu morphology'. Ph.D. dissertation, University of the Witwatersrand, Johannesburg.

van der Spuy, A. 2006.
Wordhood in Zulu. **South African Linguistics and Applied Language Studies** 24(3): 311–329.

Zeller, J. 2005.
Universal principles and parametric variation: remarks on formal linguistics and the grammar of Zulu. **Ingede Journal of African Scholarship** 1(3). Available: http://linguistics.ukzn.ac.za/Uploads/137d3f2e-0edc-463f-95af-68b9e21267ed/IngedeContribution.pdf.

About the authors: *Sonja E. Bosch* is professor in the Department of African Languages at the University of South Africa (UNISA). Her main field of interest is Nguni natural language processing, with specialization in morphological analysis.

Laurette Pretorius is professor in the School of Computing, University of South Africa, and also works with the Knowledge Systems Group, Meraka Institute, CSIR, Pretoria, South Africa, Her research interests include the natural language processing of various South African languages.

Axel Fleisch is professor at the Institute for Asian and African Studies, University of Helsinki, Finland. His research background is in descriptive linguistics and the typology of Bantu languages. His approach relies on cognitive theories, and he has started to work on Nguni, mainly Ndebele, while gathering data for a comparative semantic study on these languages