Experimental fast-tracking of morphological analysers for Nguni languages

Sonja Bosch, Laurette Pretorius, Kholisa Podile, Axel Fleisch

University of South Africa
PO Box 392, UNISA 0003, South Africa
boschse@unisa.ac.za, pretol@unisa.ac.za, podilk@unisa.ac.za, axel.fleisch@helsinki.fi

Abstract

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. This paper addresses the experimental fast-tracking of the development of finite-state morphological analysers for Xhosa, Swati and (Southern) Ndebele by using an existing prototype of a morphological analyser for Zulu. The research question is whether fast-tracking is feasible across the language boundaries between these closely related varieties. The objective is a thorough assessment of the recognition rates yielded by the Zulu morphological analyser for the three related languages. The strategy is to use fast-tracking techniques that consist of several cycles 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 tests show that the high degree of shared typological properties and formal similarities among the Nguni varieties warrants a modular fast-tracking approach. Those word forms in these languages that were recognized by the Zulu analyser were mostly adequately interpreted. 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. Introduction

It is well-known that the development of natural language processing (NLP) components is resource-intensive. Rule-based approaches usually require the writing of large numbers of language grammar rules while statistical 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.

In this paper we discuss the experimental fast-tracking of the development of 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. The pertinent research question, therefore, is whether the existing morphological analyser prototype for Zulu (ZulMorph) may be used effectively for fast-tracking the development of morphological analysers for the other three mentioned languages[1]. All four languages are very closely related. Their internal maximum linguistic distance compares, e.g., to that of Spanish and Portuguese. The greatest challenge for the morphological analysis of the Nguni languages lies with the nominal (complex gender system with formally marked noun classes) and the verbal morphology (rich morphology, both inflectional and derivational). Fortunately, despite the complexities of these domains, they are comparable across language boundaries with a high degree of formal similarity. We are therefore 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 (sa) and Oflazer, Nirenburg and McShane (2001). However, the approach described in this paper differs from bootstrapping in the sense that we do not use a Nguni morphological analyser to improve itself, but rather use the Zulu morphological analyser to develop analysers for closely related languages.

The structure of the paper is as follows: Section 2 briefly discusses the general approach with attention to the role of the small parallel development corpus. In section 3 this general approach is unpacked as a sequence of steps in which the baseline analyser, ZulMorph, is applied and then systematically extended

[1] A first attempt at the morphological analysis of Xhosa may be found in (Theron & Cloete, 1997) while Swahili is, as far as we know, the first Bantu language for which a complete morphological analyser has been developed (Hurskainen, 1992).

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 4, 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. In the final section we reflect on the feasibility and suitability of the approach, draw a number of conclusions and map out possible future research and development directions for the morphological analysers under discussion, as well as the fast-tracking of NLP components for these languages in general.

2. Approach

We address the research question by performing experiments of increasing scope in order to assess the feasibility of the approach. The fast-tracking 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.

The suitability of finite-state approaches to computational morphology has been proven 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 modelling the morphotactics is lexc. 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 modelled with the Xerox regular expression language. These regular expressions are then compiled into a finite-state network by means of the xfst tool.

Finally, the two mentioned finite-state networks are combined (composed) together into a single network, a so-called lexical transducer, which constitutes the 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).

So what do we in fact have for ZulMorph, that we can use in the fast-tracking process?

In the first place we have the *morphotactics* component, i.e. an accurate specification of the Zulu word structure. The rich agglutinating morphological structure, which characterises a language such as Zulu, is 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, 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. The morphotactics component therefore 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 we have 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.

Morpho-tactics (lexc)	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 combina-tions and orders of morphemes (e.g. u-ya-ngi-thand-a and not *ya-u-a-thand-ngi)
Morpho-phono-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

[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.

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]
```

3. Procedure

3.1 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 with approximately 200 types for each language, i.e. unique "words" in running text converted into a word list. The analysis of the Zulu 200-type word list was perfected to 100% before the experiment commenced. The success rate of analysis for the other languages is: 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	Rel/adj	Prons	Conj
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 word (26.92 %)
Corpus size: 182 words

Verbs	Nouns	Rel/adj	Prons	Conj
ubatjela warhaba zatjeheja bekabhuda- nga	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 the 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.

3.2 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: 42 words (23.07 %)
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.

3.3 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 2 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)

As can be gleaned from the failures, instances of roots peculiar to a single language or identical to Zulu, simultaneously demonstrate prefixes or suffixes which do not conform to the Zulu equivalent. 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. For instance in the Xhosa *umntu* (a human being) the class prefix *um-* differs from the Zulu *umu-*, that has been modelled 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 modelled 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 modelled in the analyser for polysyllabic noun stems in class 6.

3.4 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,.
The experiment resulted in bringing the three additional languages on a par 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 (lexc). 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 lexc. 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 modelled separately.

3.5 Step 4

Step 4 concerns the adjustment of rules in the rule component of the morphological analyser. The rules are of a dual nature: firstly the rules that model morphophonological alternations, and secondly, auxiliary rules that are introduced for technical reasons. Regarding morphophonological alternations, it was decided to concentrate only on class 9 and 10 rules for this experiment. We describe the relevant Zulu rules and then indicate how the other languages deviate from them. Examples of xfst 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/l/m/n/ny*, e.g.
Zulu: `inkambo i[NPrePre9]n[BPre9]hambo[NStem]`
Xhosa: `ihambo [NPrePre9]hambo[NStem]`

In Swati class 9 aspiration (*h*) remains when *n* is followed by *kh*, *ph*, e.g.
Zulu: `inkulumo [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 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: Cons + a + izin > Cons + ezin
Xhosa and Ndebele Cons + a + iin > Cons + een;
Swati Cons + a + tin > Cons + 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. In the case of auxiliary rules introduced for technical reasons in the rule component, we refer to the following example where the notation %^YY denotes a multi-character symbol (in xfst) introduced in lexc (as ^YY) to mark a particular morpheme yy (say) for use in the rule modelling. These symbols are used in managing alternations and their contexts. Once the symbol has played its discriminatory strategic role, an 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 in figure 1, which is realised as either *zi*, *zim* or *zin*, depending on the context. A detailled explanation of the xfst syntax falls outside the scope of this article (see Beesley and Karttunen, 2003).

```
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;
```

Figure 1: Fragment of xfst scipt for Xhosa class 10 rule

However, briefly, in figures 1, 2 and 3 % 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.
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 %^X is introduced to prevent the rule for Zulu vowel combinations (figure 2) to change *ii* to *i*. The symbol %^X 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;
```

Figure 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;
```

Figure 3: Xhosa and Ndebele rules to preserve *ii* and *ee*

The adjustment of rules in the ZulMorph rule component (xfst) as 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.

4. 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, (sa)). 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:

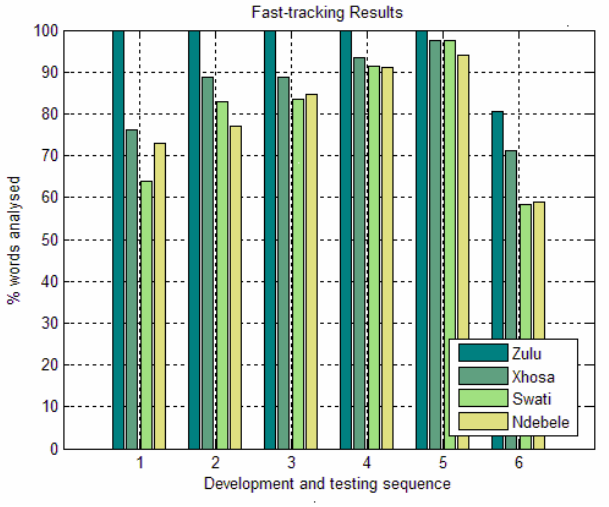
Analysed: 3971 words (58.26 %)
Failed: 2845 words (41.74 %)
Corpus size: 6816 words

- 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 fast-tracking 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).



(1 = step 1; 2 = step 2a; 3 = step 2b; 4 = step 3;
5 = step 4; 6 = preliminary evaluation)

Figure 4: Results of development and testing sequence

5. Conclusion and future work

We return to the research question, namely whether the existing morphological analyser prototype for Zulu may be used effectively for fast-tracking the development of morphological analysers for the other three Nguni languages. Our goal is the development of accurate, usable broad-coverage morphological analysers for the Nguni languages, and therefore the significance of the experiment is:

- There are obvious benefits with regard to development time. Taking into consideration that the development of the Zulu analyser prototype commenced in 2001, whereas experiments with regard to the current development of Xhosa, Swati and Ndebele analyser prototypes took only 3 to 4 months to develop, the procedure may certainly be regarded as "fast-tracking".
- 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 optimising the software process for the further development of these software artefacts. All the phases of the software life cycle, including linguistic design and modelling, implementation, testing, documentation, verification, validation, maintenance and improvement, will benefit. A unified approach is here understood as modular. Besides the benefits in terms of maintenance, this has another great advantage over fully independently developed analysers. Codeswitching and extensive borrowing between each other are common among the Nguni languages. If the respective language-specific tools are compatible, it may at some point allow for an easier integration of these components to build more "permissive" tools.

- 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.
- If this approach proves successful, it can in future also be used for the development of other tools for these languages.

Future work 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. 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, i.e. adding morphological information to lexc, and adapting rules in a systematic manner. Certain areas in the grammar of the individual languages need to be modelled 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 importantly, 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.

6. 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.

7. References

Beesley, Kenneth.R., Karttunen, L. (2003). *Finite state morphology*. Stanford, CA: CSLI Publications.

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

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

Oflazer, K., Nirenburg, S. (sa) [O]. Practical Bootstrapping of Morphological Analyzers. Available: http://citeseer.ist.psu.edu/270795.html Accessed on 30 May 2007.

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

Pretorius, L., Bosch, S.E. (2003). Finite-State Computational Morphology: An Analyzer Prototype for Zulu. *Machine Translation* 18, pp. 195-216.

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

Theron, P., Cloete, I. (1997). Automatic acquisition of two-level morphological rules. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* (Washington, DC, March 31 - April 03, 1997). Applied Natural Language Conferences. Morgan Kaufmann Publishers, San Francisco, CA, pp. 103-110.