Spellcheckers for the South African languages, Part 2: The utilisation of clusters of circumfixes

DJ Prinsloo*
Department of African Languages, University of Pretoria, Pretoria 0002, South Africa
E-mail: danie.prinsloo@up.ac.za

Gilles-Maurice de Schryver
Department of African Languages and Cultures, Ghent University, Rozier 44, B-9000 Ghent, Belgium
Department of African Languages, University of Pretoria, Pretoria 0002, South Africa
E-mail: gillesmaurice.deschryver@UGent.be

May 2003

The aim of this article is to introduce the utilisation of 'clusters of circumfixes' as a new strategy to increase the lexical recall in spellcheckers for especially African languages. This strategy is particularly useful in the compilation of spellcheckers for conjunctively written languages such as those in the Nguni group, where even extensive lexica render lexical recall values of only 90%. It is not disputed that a full morphological decomposition, for instance by means of finite-state analysis, is a good strategy for obtaining improved lexical recall, especially for the Nguni group, and that it should be pursued as an excellent solution in the long run. The utilisation of clusters of circumfixes, however, offers a less comprehensive alternative or in-between strategy to increase the lexical recall. These clusters can furthermore be compiled in a short space of time and with standard word processing and database software.

The need to improve the lexical recall of spellcheckers for the Nguni languages

Although all official African languages of South Africa are so-called *agglutinating* languages,[1] in which morphemes are juxtaposed to form linguistic words, this agglutination is not always visible on the orthographic level. As a result of distinct phonological processes, formatives are written together in the Nguni languages (i.e. isiXhosa, isiZulu, isiNdebele and siSwati), yet mostly separately in the disjunctively written languages (i.e. Sesotho sa Leboa, Sesotho, Setswana, Xitsonga and Tshivenda). A succinct overview of the underlying practical motivations for the development of these two traditions can be found in Louwrens (1991: 1-12). From a spellchecker angle, the Nguni languages present a considerable challenge. Indeed, in De Schryver and Prinsloo (2004), henceforth referred to as Part 1, it is for instance shown that an isiZulu spellchecker that merely consists of the most frequent 600,000 orthographic words drawn from a corpus, barely results in lexical recall values attaining 90%. What one actually needs to do in order to be able to validate more correctly spelled words is therefore to *parse* the Nguni words. 'Parsing' is the second of only eight key concepts in Jurafsky and Martin's 1,000-page textbook: "Parsing means taking an input and producing some sort of structure for it" (2000: 57). Here, the obtained 'structure' may of course be seen as a *morphological decomposition*.

In other words, given the morphological and morphophonological features of the Nguni group and the way these features are represented in a conjunctive orthography, the ideal would be to design spellchecker modules that could do a full morphological parsing of the input texts. Although good progress has been made for some African languages in this regard (cf. Part 1), developing the necessary tools for *full* morphological parsing proves to be a complicated and time consuming process. Since the urge is to build better spellcheckers

[*] Author to whom correspondence should be addressed.

right away, the process can hardly be delayed, and any suggested alternative solution should be workable at this point in time, that is, should result in products for immediate use. This article therefore proposes such an alternative solution, whereby the lexical recall can be substantially improved. This solution, which can be characterised as a *partial and statistically motivated morphological decomposition*, can be employed until such time as full morphological analysers/generators, finite-state or otherwise, will have been developed. The approach is *partial* because there is no insistence on decomposing down to each formative morpheme, as one rather works with 'clusters of circumfixes';[2] it is *statistically motivated* because the very choice of these affix chunks is heavily based on frequency data. This methodology, which could also be termed 'frequency-driven clumped morphotactics', is currently being developed on a large scale for all the Nguni languages. Also note that the suggested approach can be fine-grained over time, meaning that clusters of circumfixes can be used as an alternative point of departure for morphological analysis when full morphological parsing for spellchecking purposes is attempted at a later stage.

The main aim of this article is thus to introduce a new strategy to improve the lexical recall of spellcheckers for especially the Nguni languages. One language from this group, namely isiZulu, will be used for the various illustrations throughout. For the case study presented in Part 1, where the document entitled "What is the African National Congress?" (ANC, [sa]) was spellchecked with a spellchecker lexicon comprising the top 600,000 isiZulu orthographic words, it was shown that 79 correctly spelled types remained unrecognised, even after the addition of a filter to take care of mixed capitalisation. With the addition of this filter, the lexical recall for the ANC text went up to around 92%. One of the practical tasks in the present article, Part 2, will then also be to check if more words can indeed be recognised in the ANC text with the suggested approach, so that the lexical recall may improve further. For the discussion, attention will be devoted to nominal and verbal forms, and to a lesser extent also to compounds.

The complexity of the Nguni languages

Noun and verb roots that occur with long strings of affixes (i.e. prefixes, suffixes and circumfixes) are by far the largest category of words responsible for most of the non-recognised but correctly spelled words in texts for isiXhosa, isiZulu, isiNdebele and siSwati. The main challenge is thus to find a way to strengthen spellcheckers for these languages so that a greater percentage of especially nouns and verbs in which affixes are stacked will be recognised. The best practical way in seeking a solution to this challenge seems to be to deal with (non-compounded) single verbs and nouns first and to analyse their *affix patterns*.

The traditional approach in grammatical analyses of the Nguni languages, in the broader sense of the word, is to meticulously formulate sets of rules in order to try to cater for every single and all of the possible combinations and permutations of affixes and to describe processes such as sound changes, assimilation, deletion, elision, etc. The rules governing affixation in the Nguni languages are rather complicated and many combinations and permutations are likely to occur when combining, for example, *all* subject concords with *all* object concords with *all* possible verbal extensions. According to Van Wyk more than 4,000 prefix combinations could be found with a single verb root in isiZulu:

> Any verb root can be combined with any subject marker, any modal or aspectual morpheme plus a compatible final vowel, and any appropriate negative morpheme[.] If it is a transitive root, it can moreover be combined with any object marker (or the reflexive morpheme). The number of combinations possible for a suitable transitive verb stem is, therefore, 18 x 19 x 6 x 2. (Van Wyk, 1995:87)

Note that this calculation does not include the numerous verbal extensions and combinations of extensions with which these roots cum verbal prefixes could occur. As an illustration, consider Tables 1 and 2, which respectively list a random selection of affixes co-occurring with the verb stem **-khuluma** 'speak, talk' and the noun **abantu** 'people' in a 5.0-million-word isiZulu corpus, together with their frequencies in that corpus.

Any attempt to comprehensively describe *all* possible combinations of affixes necessarily leads to multi-tiered rule systems that may for instance be approached from a finite-state angle. As a less complex alternative, one could consider studying the bolded sections such as those seen in Tables 1 and 2, and see if using the observed 'patterns' could not lead to simpler solutions to parsing, which could be implemented in a relatively short space of time and which would not require training in and access to sophisticated computer programs. This attempt will now be pursued.

Table 1: Extract for **-khuluma** ‘speak, talk’ in a 5.0-million-word isiZulu corpus

Word	Freq.	Word	Freq.	Word	Freq.
ukukhulum**a**	1,082	**be**khulum**a**	270	**kwa**khulum**a**	138
esekhulum**ile**	24	**noku**khulum**a**	159	**be**khulum**ela**	39
ekukhulum**eni**	48	**zoku**khulum**a**	212	**abawa**khulum**ayo**	6
njengokukhulum**a**	21	**aka**khulum**anga**	11	**waye**khulum**a**	100
ukukhulum**isa**	12	**a**khulum**a**	207	**esa**khulum**a**	93
engasakhulum**anga**	9	**uya**khulum**a**	184	**o**khulum**ayo**	171

Table 2: Extract for **abantu** ‘people’ in a 5.0-million-word isiZulu corpus

Word	Freq.	Word	Freq.	Word	Freq.
ngabant**u**	950	**kano**bant**u**	62	**yilaba**bant**u**	18
kwabant**u**	691	**banga**bant**u**	59	**nalaba**bant**u**	16
okungabant**wanyana**	3	**naku**bant**u**	52	**kwakunge**bant**u**	12
lababant**u**	511	**ko**bant**u**	49	**babenga**bant**u**	11
singabant**ukazana**	3	**okwa**bant**u**	49	**singa**bant**u**	49

An example-driven needs analysis

Consider, as a point of departure for the discussion, the following randomly chosen examples of orthographic words that were not recognised by a 600,000-words strong wordlist-only isiZulu spellchecker (cf. Part 1, Addendum B): **ukuzizuzela** (< *uku-zi-zuz-el-a*) ‘to gain for oneself’, **nokungasebenzisi** (< *na-(u)ku-nga-sebenz-is-i*) ‘and by not utilising’, and **wokufezekisa** (< *wa-(u)ku-fez-ek-is-a*) ‘of making something possible’. Numerous examples of the ‘pattern’ *ukuzi- + verbal root + -ela* exist in the corpus, of which the isiZulu words printed in Roman in Column 1 of Table 3 are but a few examples.

Table 3: Comparing *patterns* and *paradigms* for a selection of isiZulu words

uku+zi+<V_root>+el+a	na+(u)ku+nga+<V_root>+is+i	wa+(u)ku+<V_root>+ek+is+a
ukuzi**zuz**ela		
	nokunga**sebenz**isi	
		woku**fez**ekisa
ukuzibamb**ela**		
	nokungagcul**isi**	
		wokufan**ekisa**
ukuzifik**ela**		
	nokungahlek**isi**	
		wokuy**ekisa**
ukuzibon**ela**		
	nokungalobol**isi**	
ukuzihlal**ela**		
ukuzibhal**ela**		
…		
	…	
		…

For the other two patterns shown in Table 3, Row 1, Columns 2 and 3, only the listed examples printed in Roman could be found in the corpus. Thus Columns 1-3 show three *incomplete paradigms* with missing derivational forms represented by *every empty space* in the table. If these forms are missing in a certain corpus, so will they be in spellchecker lexica culled from that corpus. Given a situation as illustrated in Table 3, the following options are available to the compiler for improving the spellchecker. The compiler could:

(i) add more words manually to the spellchecker lexicon; and/or

(ii) increase the corpus on the assumption that there will always be a direct positive relation 'bigger corpus → larger list of types that can be used as wordlist for the spellchecker lexicon → improved lexical recall'; and/or

(iii) attempt to apply a set of rules consisting of affix patterns on noun and verb roots (and stems) in the lexicon to complete the nominal and verbal paradigms.

It is true that the utilisation of options (i) and (ii) will gradually fill some or even many of the gaps but a much quicker and decisive process will be to *generate* these missing forms.

For spellchecking purposes option (iii) could either mean to physically generate the missing forms in Table 3 and to add those forms to the spellchecker lexicon, or else to add their patterns (Row 1, Columns 1-3) as rule components, together with lexica of nominal and verbal roots/stems to which these patterns apply, so that if the user types in such forms, they will be validated as acceptable by the software.

For the first three examples in Table 3 it means that the spellchecker should not only be able to recognise **ukuzizuzela**, **nokungasebenzisi** and **wokufezekisa**, but should also be able to recognise other verbs with the same affix patterns. For the paradigm represented by **ukuzizuzela**, with pattern *uku+zi+<V_root>+el+a*, the <V_root> = **-zuz-** should at that point be replaced by other verbal roots.

One thus sees that a substantial pattern simplification and reduction is achieved when the affix patterns are considered as large affix clusters, here *ukuzi* and *ela* respectively prefixed and suffixed to the root. A further simplification and reduction is achieved when circumfix clusters are considered, namely *ukuzi<V_root>ela* where <V_root> is a placeholder for insertion of a verbal root. The same could be done for **nokungasebenzisi**, thus a circumfix cluster *nokunga<V_root>isi*, and *woku<V_root>ekisa* for **wokufezekisa**. If these three clusters of circumfixes, *ukuzi<V_root>ela*, *nokunga<V_root>isi* and *woku<V_root>ekisa* are used as *generators* on for instance the verbal roots **-zuz-**, **-sebenz-** and **-fez-** the result is the full 3 x 3 paradigm shown in Table 4.

Table 4: Illustration of a 3 x 3 generation matrix (circumfix: prefix(es)<V_root>suffix(es))

	ukuzi+<V_root>+ela	nokunga+<V_root>+isi	woku+<V_root>+ekisa
<zuz>	**ukuzi+<zuz>+ela**	nokunga+<zuz>+isi	woku+<zuz>+ekisa
<sebenz>	ukuzi+<sebenz>+ela	**nokunga+<sebenz>+isi**	woku+<sebenz>+ekisa
<fez>	ukuzi+<fez>+ela	nokunga+<fez>+isi	**woku+<fez>+ekisa**

Departing from the three words in bold, this 3 x 3 matrix generated the six non-boldface words, none of which appears in the corpus, and all of which could be added to the spellchecker lexicon. Performing the same procedure for the remainder of Table 3 will render a complete 10 x 3 matrix that will add 20 new words to the spellchecker lexicon, even though those 20 new forms did not occur in the corpus. Hundreds of thousands of new words, which are very likely to be correct/possible forms as a result of the fact that such large clusters of affixes are employed, may thus be generated by an *m x n* matrix. The larger *m* and *n*, the more words are added.

This approach can thus more formally be described as the utilisation of clusters of circumfixes as fixed form *generators* in the process of generating especially nominal and verbal derivational forms for the purpose of increased lexical recall. This definition implies that clusters of circumfixes will in most cases consist of more than one morpheme for which synchronic combination of morphemes and possible resulting sound changes have been completed, hence these clusters may be viewed as artificially fossilised units often containing both strings of prefixes and suffixes. Likewise, the nominal and verbal forms that form the *objects* of the generation process do not necessarily conform to what is grammatically regarded as a nominal or verbal root/stem. The sole aim is to design the *generators* (clusters of circumfixes) and the *objects* on which the generators operate (the nominal/verbal forms) in such a way that nominal and verbal *derivational forms* likely to exist in the language are generated. Formulated differently, the approach entails a process whereby generators operate on objects to render nominal and verbal derivational forms that could either be physically or potentially generated orthographic words. For the actual generation, no sophisticated computer program is required – it can even be done by the mail merge function in *Microsoft Word*.

The compiler should of course at all times closely monitor over-generation, meaning that the number of erroneous and unnatural words constructed in this way should be kept to a minimum. Viewed from the angle of the user of a spellchecker, it should be determined to what extent *error recall* will be influenced negatively, through the 'acceptance' of unnatural words and even errors by the spellchecker. Note, however, that most if not all morphological generation systems are burdened with over-generation. Indeed, comparable generation or acceptance strategies as the one illustrated in Table 4 are commonly used in spellcheckers, but the efforts to combat or filter out unnatural and incorrect forms vary from rather bad to good. Especially for semi-agglutinative languages with productive compound formation such as for instance Afrikaans, Dutch or German, the problem may be rather severe. The spellchecker for Dutch built into Microsoft Word 2000, for instance, does not flag any error in the following phrase:

> *begeleidingdoor jullie is vanbelang voor mijnkwalitatieve luidsprekeryoghurtijsjes*
> 'guidanceby you is ofimportance for myqualitative smallloudspeakeryoghurticecreams'

The three run-ons, *begeleidingdoor, *vanbelang and *mijnkwalitatieve have not been flagged as errors, presumably because the words **door**, **van** and **mijn** are listed in a sub-lexicon of productive affixes. The non-word *luidsprekeryoghurtijsjes illustrates the fact that any two (compound) nouns may be glued together in this spellchecker of Dutch, with nominal inflection still applicable to the result.

The quest for enhanced precision will always be a factor. On the one hand a compiler of a spellchecker should check if, on the whole, *generated* orthographic words are correct, natural isiZulu words. Conversely, one could argue that if users of a spellchecker compose what are basically generated forms, the likelihood that these forms also exist is rather high, and so the spellchecker would best not flag and thus *accept* them.

There is nevertheless a way to try and stem the encroachment of over-generation. Indeed, the *frequency* of occurrence of clusters of circumfixes and the *frequency* of nouns and verbs could also be studied. It is reasonable to expect that a sub-lexicon consisting of frequently used nouns and verbs subjected to a set of highly frequent clusters of circumfixes will generate a high percentage of correct, natural words and thus improved recall and precision. The opposite might also be true, namely that a set of infrequent clusters of circumfixes computed over infrequently used nouns and verbs would render a larger percentage of erroneous or unnatural words and thus lower recall and precision values. Support for this hypothesis can be found in Van Huyssteen and Van Zaanen's observation, for Afrikaans, that "one could prevent over-stemming mistakes by limiting the stemming algorithm to only the most frequent affixes" (2003: 194). (Observe that stemming, whereby one cuts off affixes prior to dictionary lookup, can be seen as the inverse of the generation of words by means of cluster circumfixation.)

A top-down frequency-level approach to cluster circumfixation

The suggested approach to cluster circumfixation as a generation strategy is a *top-down approach* in terms of frequency of both clusters of circumfixes and nominal or verbal roots/stems as the objects of the generation.

The clusters/generators as well as the nominal and verbal objects for the generation process were extracted from a 5.0-million-word isiZulu corpus. Firstly, noun and verb roots/stems were selected from a frequency list of types generated from the corpus. Clusters of circumfixes co-occurring with these nominal and verbal roots/stems were then isolated. This, for example, means that clusters of circumfixes that generally occur with all or most of these selected nouns/verbs are regarded as high frequency generators, and noun/verb roots/stems occurring with a high frequency in the corpus are regarded as high frequency objects for generation of nominal or verbal forms. Likewise, clusters of circumfixes that only occur with a limited number of nouns/verbs in the corpus are regarded as low frequency generators and noun/verb roots/stems occurring with a low frequency in the corpus are regarded as low frequency objects for generation of nominal and verbal forms.

A good strategy for any compiler of a spellchecker seems to be to test the increase in lexical recall obtained by using top frequency generators and top frequency objects first, and to then extend this to lower frequency generators and objects until the desired recall percentage has been reached.

As a case study, all the orthographic forms of the top 20 verbs were extracted, from which the most frequent *verbal clusters of circumfixes* were deduced. Compare, for example, the most frequent patterns for two frequently used isiZulu verbs in Tables 5 and 6.

Table 5: Top frequent orthographic forms for the verb stem **-sebenza** 'work'

Word	Freq.	Word	Freq.	Word	Freq.
umsebenzi	5,053	ukusebenzisa	237	abasebenza	128
emsebenzini	1,375	izisebenzi	223	komsebenzi	127
imisebenzi	931	sebenzisa	206	osebenza	120
ukusebenza	513	ngokusebenzisa	197	wasebenza	117
lomsebenzi	420	ngumsebenzi	169	asebenze	116
msebenzi	392	usebenzise	166	ngemisebenzi	111
nomsebenzi	293	nemisebenzi	165	somsebenzi	111
esebenza	274	abasebenzi	150	ngisebenza	110
usebenza	266	emisebenzini	148	nokusebenza	100
ngomsebenzi	238	usebenzisa	140		

Table 6: Top frequent orthographic forms for the verb stem **-khuluma** 'speak, talk'

Word	Freq.	Word	Freq.	Word	Freq.
ukukhuluma	1,082	bakhuluma	288	nokukhuluma	159
ukhuluma	1,081	ikhuluma	275	ngikhulume	146
ekhuluma	954	bekhuluma	270	esekhuluma	140
wakhuluma	730	ukhulume	252	kwakhuluma	138
ngikhuluma	622	zokukhuluma	212	sikhulume	111
khuluma	555	okhuluma	209	yakhuluma	109
akhulume	472	akhuluma	207	wayekhuluma	100
kukhuluma	337	uyakhuluma	184		
sikhuluma	317	okhulumayo	171		

From the selection of 20 frequently used verbs, the 48 most common verbal clusters of circumfixes were isolated, as listed in Table 7.

Table 7: Top frequency verbal generators (derived from 20 frequently used verbs)

Cluster of circumfixes	Freq.	Cluster of circumfixes	Freq.	Cluster of circumfixes	Freq.
u<>a	19	i<>a	17	esi<>a	16
nga<>a	19	wawu<>a	17	ba<>e	16
ku<>a	18	ngi<>e	17	uzo<>a	16
koku<>a	18	ya<>a	17	soku<>a	16
ba<>a	18	o<>a	17	wayese<>a	16
la<>a	18	noku<>a	17	waye<>a	16
uku<>a	18	ngoku<>a	17	ngizo<>a	16
si<>a	18	unga<>i	17	ni<>e	16
uya<>a	18	iya<>a	16	sezi<>a	16
wa<>a	18	baya<>a	16	uyo<>a	16
ngi<>a	18	babe<>a	16	zi<>a	16
ngiya<>a	18	li<>a	16	o<>ayo	16
a<>e	17	azo<>a	16	sesi<>a	16
be<>a	17	esa<>a	16	u<>e	16
ese<>a	17	loku<>a	16	use<>a	16
e<>a	17	e<>e	16	sa<>a	16

Consider now the application of the 48 verbal generators from Table 7 on five randomly selected verbs, that is the generation objects, rendering the generation result shown in Table 8.

Table 8: Generated verbal derivational forms for **-chaz-** ‘explain’, **-phel-** ‘live’, **-limal-** ‘injure’, **-thath-** ‘take’ and **-duduz-** ‘comfort’

uchaza	uphela	ulimala	uthatha	ududuza
ngachaza	ngaphela	ngalimala	ngathatha	ngaduduza
kuchaza	kuphela	kulimala	kuthatha	kududuza
kokuchaza	kokuphela	kokulimala	kokuthatha	kokududuza
bachaza	baphela	balimala	bathatha	baduduza
lachaza	laphela	lalimala	lathatha	laduduza
ukuchaza	ukuphela	ukulimala	ukuthatha	ukududuza
sichaza	siphela	silimala	sithatha	siduduza
uyachaza	uyaphela	uyalimala	uyathatha	uyaduduza
wachaza	waphela	walimala	wathatha	waduduza
ngichaza	ngiphela	ngilimala	ngithatha	ngiduduza
ngiyachaza	ngiyaphela	ngiyalimala	ngiyathatha	ngiyaduduza
achaze	aphele	alimale	athathe	aduduze
bechaza	bephela	belimala	bethatha	beduduza
esechaza	esephela	eselimala	esethatha	eseduduza
echaza	ephela	elimala	ethatha	eduduza
ichaza	iphela	ilimala	ithatha	iduduza
wawuchaza	wawuphela	wawulimala	wawuthatha	wawududuza
ngichaze	ngiphele	ngilimale	ngithathe	ngiduduze
yachaza	yaphela	yalimala	yathatha	yaduduza
ochaza	ophela	olimala	othatha	oduduza
nokuchaza	nokuphela	nokulimala	nokuthatha	nokududuza
ngokuchaza	ngokuphela	ngokulimala	ngokuthatha	ngokududuza
ungachazi	ungapheli	ungalimali	ungathathi	ungaduduzi
iyachaza	iyaphela	iyalimala	iyathatha	iyaduduza
bayachaza	bayaphela	bayalimala	bayathatha	bayaduduza
babechaza	babephela	babelimala	babethatha	babeduduza
lichaza	liphela	lilimala	lithatha	liduduza
azochaza	azophela	azolimala	azothatha	azoduduza
esachaza	esaphela	esalimala	esathatha	esaduduza
lokuchaza	lokuphela	lokulimala	lokuthatha	lokududuza
echaze	ephele	elimale	ethathe	eduduze
esichaza	esiphela	esilimala	esithatha	esiduduza
bachaze	baphele	balimale	bathathe	baduduze
uzochaza	uzophela	uzolimala	uzothatha	uzoduduza
sokuchaza	sokuphela	sokulimala	sokuthatha	sokududuza
wayesechaza	wayesephela	wayeselimala	wayesethatha	wayeseduduza
wayechaza	wayephela	wayelimala	wayethatha	wayeduduza
ngizochaza	ngizophela	ngizolimala	ngizothatha	ngizoduduza
nichaze	niphele	nilimale	nithathe	niduduze
sezichaza	seziphela	sezilimala	sezithatha	seziduduza
uyochaza	uyophela	uyolimala	uyothatha	uyoduduza
zichaza	ziphela	zilimala	zithatha	ziduduza
ochazayo	ophelayo	olimalayo	othathayo	oduduzayo
sesichaza	sesiphela	sesilimala	sesithatha	sesiduduza
uchaze	uphele	ulimale	uthathe	ududuze
usechaza	usephela	uselimala	usethatha	useduduza
sachaza	saphela	salimala	sathatha	saduduza

In order to determine whether improvement in lexical recall has been obtained by this limited experiment, the generated words in Table 8 were tested against the *WordPerfect 9* isiZulu spellchecker, the only commercially available spellchecker for isiZulu (cf. Part 1). Table 9 gives an extract from the list of generated words that are not in the WordPerfect spellchecker.

Table 9: A selection of generated verbal forms not in the WordPerfect 9 spellchecker

ichaza	ngokuphela	siphela	uyathatha	azothatha
sichaza	zichaza	ngichaze	lokuthatha	yalimala
sithatha	ulimale	uzochaza	bachaze	elimala
uphela	wayethatha	uyababa	iyababa	sachaza
walimala	othatha	uyachaza	ochaza	uyophela
lithatha	nokuchaza	uzolimala	lokuchaza	babethatha
lathatha	ngokuchaza	esichaza	ngiphele	kokuchaza
lichaza	usethatha	ethathe	lokuphela	bephela
ephela	alimale	esechaza	ngiphela	

A similar process as the one illustrated for verbs above was undertaken for nouns. In the case of nouns, all the orthographic forms of the top 80 nouns were extracted, from which the most frequent *nominal clusters of circumfixes* were then deduced. Consider the following examples of typical clusters of circumfixes occurring with the nouns **isikhathi** 'day' and **umuntu** 'human being', in Tables 10 and 11 respectively.

Table 10: Top frequent orthographic forms for the noun **isikhathi** 'day'

Word	Freq.	Word	Freq.	Word	Freq.
isikhathi	4,728	lesisikhathi	215	ngalesikhathi	51
ngesikhathi	1,693	yisikhathi	188	lesikhathi	44
sikhathi	1,155	ngalesisikhathi	185	kuyisikhathi	32
kwesikhathi	582	lesosikhathi	169	sikhathisimbe	30
ngalesosikhathi	274	ngasikhathi	95	kulesisikhathi	26
esikhathini	257	sekuyisikhathi	91	kusenesikhathi	18
nesikhathi	222	nangesikhathi	56		

Table 11: Top frequent orthographic forms for the noun **umuntu** 'human being'

Word	Freq.	Word	Freq.	Word	Freq.
umuntu	13,418	ungumuntu	386	namuntu	138
muntu	3,086	komuntu	333	engumuntu	137
lomuntu	1,610	okomuntu	292	ngingumuntu	132
ngumuntu	848	ngomuntu	238	umuntukaziwa	108
nomuntu	613	somuntu	224	zomuntu	103
kumuntu	570	womuntu	176	wayengumuntu	102
yomuntu	530	mntanomuntu	152		
njengomuntu	468	lowomuntu	140		

The nominal generators themselves can be designed in several ways, so as to arrive at for example generators specific to each noun class, or generators that reflect grammatical or semantic subdivisions within each noun class. For this experiment nouns were divided into the 20 types shown in Table 12, taking the pre-prefix and the final vowel of the noun into account.

Table 12: Noun types used for nominal generation

	a<>	**i<>**	**o<>**	**u<>**
<>a	a<N_root>a	i<N_root>a	o<N_root>a	u<N_root>a
<>e	a<N_root>e	i<N_root>e	o<N_root>e	u<N_root>e
<>i	a<N_root>i	i<N_root>i	o<N_root>i	u<N_root>i
<>o	a<N_root>o	i<N_root>o	o<N_root>o	u<N_root>o
<>u	a<N_root>u	i<N_root>u	o<N_root>u	u<N_root>u

From Table 10 and other frequent nouns of the type *i<N_root>i* the top frequency generators seen in Table 13 were isolated (which is thus one of the twenty tables for nouns).

Table 13: Top frequency generators for the noun type *i<N_root>i*

kuyi<>i	okuyi<>i	we<>i	kwa<>i	okwe<>i
nase<>ini	kwakuyi<>i	bene<>i	nayi<>i	sine<>i
e<>i	iyi<>i	abe<>i	se<>i	kwe<>i
le<>i	na<>i	kune<>i	kule<>i	lwe<>i
yile<>i	ye<>i	yi<>i	ne<>i	ngale<>i
ze<>i	be<>i	nge<>i	njenge<>i	
e<>ini	ngokwe<>i	u<>i	e<>	

If the generators from Table 13 are applied to nouns of the type *i<N_root>i*, being the obligatory generation objects, then orthographic forms such as those displayed in Table 14 are obtained.

Table 14: Generated nominal derivational forms for **isibindi** 'liver; courage', **isibili** 'second time', **isithunzi** 'shadow', **isigijimi** 'runner' and **isikhali** 'weapon'

kuyisibindi	kuyisibili	kuyisithunzi	kuyisigijimi	kuyisikhali
nasesibindini	nasesibilini	nasesithunzini	nasesigijimini	nasesikhalini
esibindi	esibili	esithunzi	esigijimi	esikhali
lesibindi	lesibili	lesithunzi	lesigijimi	lesikhali
yilesibindi	yilesibili	yilesithunzi	yilesigijimi	yilesikhali
zesibindi	zesibili	zesithunzi	zesigijimi	zesikhali
esibindini	esibilini	esithunzini	esigijimini	esikhalini
okuyisibindi	okuyisibili	okuyisithunzi	okuyisigijimi	okuyisikhali
kwakuyisibindi	kwakuyisibili	kwakuyisithunzi	kwakuyisigijimi	kwakuyisikhali
iyisibindi	iyisibili	iyisithunzi	iyisigijimi	iyisikhali
nasibindi	nasibili	nasithunzi	nasigijimi	nasikhali
yesibindi	yesibili	yesithunzi	yesigijimi	yesikhali
besibindi	besibili	besithunzi	besigijimi	besikhali
ngokwesibindi	ngokwesibili	ngokwesithunzi	ngokwesigijimi	ngokwesikhali
wesibindi	wesibili	wesithunzi	wesigijimi	wesikhali
benesibindi	benesibili	benesithunzi	benesigijimi	benesikhali
abesibindi	abesibili	abesithunzi	abesigijimi	abesikhali
kunesibindi	kunesibili	kunesithunzi	kunesigijimi	kunesikhali
yisibindi	yisibili	yisithunzi	yisigijimi	yisikhali
ngesibindi	ngesibili	ngesithunzi	ngesigijimi	ngesikhali
usibindi	usibili	usithunzi	usigijimi	usikhali
kwasibindi	kwasibili	kwasithunzi	kwasigijimi	kwasikhali
nayisibindi	nayisibili	nayisithunzi	nayisigijimi	nayisikhali
sesibindi	sesibili	sesithunzi	sesigijimi	sesikhali
kulesibindi	kulesibili	kulesithunzi	kulesigijimi	kulesikhali
nesibindi	nesibili	nesithunzi	nesigijimi	nesikhali
njengesibindi	njengesibili	njengesithunzi	njengesigijimi	njengesikhali
esibind	esibil	esithunz	esigijim	esikhal
okwesibindi	okwesibili	okwesithunzi	okwesigijimi	okwesikhali
sinesibindi	sinesibili	sinesithunzi	sinesigijimi	sinesikhali
kwesibindi	kwesibili	kwesithunzi	kwesigijimi	kwesikhali
lwesibindi	lwesibili	lwesithunzi	lwesigijimi	lwesikhali
ngalesibindi	ngalesibili	ngalesithunzi	ngalesigijimi	ngalesikhali

As in the case of the verbal forms, quite a number of nominal forms were generated that are not in the WordPerfect lexicon (cf. Table 15), thus suggesting a potential improvement in lexical recall if these forms are added to that spellchecker lexicon.

Table 15: A selection of generated nominal forms not in the WordPerfect 9 spellchecker

zesibili
sithunzi
ngesikhali
sikhali
yesibindi
nasibindi
yisikhali
iwesingisi
besibili
zesibindi
nayisithunzi
sibili
yesikhali
kunesithunzi
nasithunzi
lesibindi
sinesibindi
kwakuyisibindi
wesibindi
njengesikhali
lesithunzi

In Tables 9 and 15 it was shown how, in principle, generated words can extend a spellchecker lexicon, words which may then in turn contribute to an improved lexical recall. It is however important to establish to what extent the suggested generation strategy could increase the lexical recall in actual cases where correctly typed isiZulu words were flagged as incorrect by a wordlist-only spellchecker. A second important aspect that needs to be studied is the effect of frequency in respect of the generators and objects of the generators, hence to see if the flagged words will be recognised as correct if only high frequency generators are used on high frequency objects or whether most will only be recognised once the generation process has been extended to lower frequency ranks of generators and/or objects. Formulated differently, one needs to determine whether or not frequent generators computed over frequent nominal and verbal objects are sufficient to increase the recall on real texts, and to what extent lower ranking generators operating on lower frequency nominal and verbal objects need to be employed to achieve a substantial improvement in recall.

This process is briefly illustrated in Table 16 for a random selection of some of the 79 nouns and verbs from the ANC text that had not been recognised by the wordlist-only spellchecker.

Table 16: Success and failure to recognise extra words by means of clusters of circumfixes

Word in the ANC text	Cluster of circumfixes		Current item in the spellchecker lexicon				Success
	Pattern	Freq.	Item	Freq.	POS	Equivalent	
bayolingana	bayo<>a	medium	lingana	14	v.	be equal	Y
bobukoloni	bo<>i	low	ubukoloni	—	n.	colony	—
esiyisisekelo	esiyi<>o	low	isisekelo	59	n.	foundation	Y
kusidingo	ku<>o	high	isidingo	169	n.	necessity	Y
kusiFunda	ku<>a	high	isifunda	56	n.	district	Y
kwalezinjongo	kwale<>o	medium	izinjongo	72	n.	aims	Y
lwangonyaka	lwango<>a	—	unyaka	418	n.	year	—
njengesisekelo	njenge<>o	high	isisekelo	59	n.	foundation	Y
nokuhola	noku<>a	high	hola	11	v.	draw along	Y
uyinhlangano	uyi<>o	medium	inhlangano	163	n.	meeting	Y

The overall improvement in lexical recall turns out to be substantial, as only 32 types remain unrecognised (shown in italics in Part 1, Addendum B, i.e. as '*token*'). This thus means that with the utilisation of clusters of circumfixes as an approach to generate plausible orthographic words, in conjunction with a filter that takes care of mixed capitalisation, the *lexical recall increases with nearly 7%*, from around 90% to nearly 97%.[3] Clearly, the actual implementation of this alternative solution should be seriously considered for the Nguni languages.

Observe that one could furthermore also design such a strategy for the disjunctively written African languages, as well as for Afrikaans. If this is done and applied to the respective ANC texts, then one notices that the lexical recall value rises from 98.84% to 99.42% for the Sesotho sa Leboa version of the ANC text (cf. the words in italics in Part 1, Addendum A, i.e. '*token*'), and from 99.07% to 100.00% for the Afrikaans version of the ANC text (with, however, considerable over-generation in this case).[4]

Compounding

A final observation relates to the issue of compounding. For Afrikaans, for example, the compiler has to cater for connectors such as the **s** in **navorsingsuitsette** 'research outputs' and **ambassadeursvrou** 'ambassador's wife' but has to filter (remove or prevent) in one way or another, such as by implementing error lists, the generation or acceptance of this connector in words such as *seunsskool 'school for boys' or *meisiesskool 'school for girls'.

On this level the situation for the African languages, in casu for isiZulu, is apparently much more problematic. The components that make up a compound, mostly nouns and verbs, often include their own sets of complicated affixes (as discussed in the previous paragraphs) and not merely a single connector as in the Afrikaans examples given above. Furthermore, in these cases, even formulated in an over-simplified way rather than in a strict morphological analysis notation, the situation is much more complex. In isiZulu it is not just a matter of merely conjoining a verb V with its prefixes PV and suffixes SV (thus PV-V-SV), with a noun N with its own prefixes PN and suffixes SN (thus PN-N-SN), in a simplistic consecutive linear order as PV-V-SV-PN-N-SN. Most often the entire compound is nominalised, thus resulting in the structure PN-V-N-SN, or even more complicated as PN-PV-V-SV-PN-N-SN. Consider in this regard the analysis of **okuwumkhombandlela** 'something which shows the way' in Figure 1, one of the 32 types still unrecognised in the ANC text, as PV(V)-PN-(V)-(PN)-N in terms of this notation.

Figure 1: Analysis of the isiZulu compound **okuwumkhombandlela**

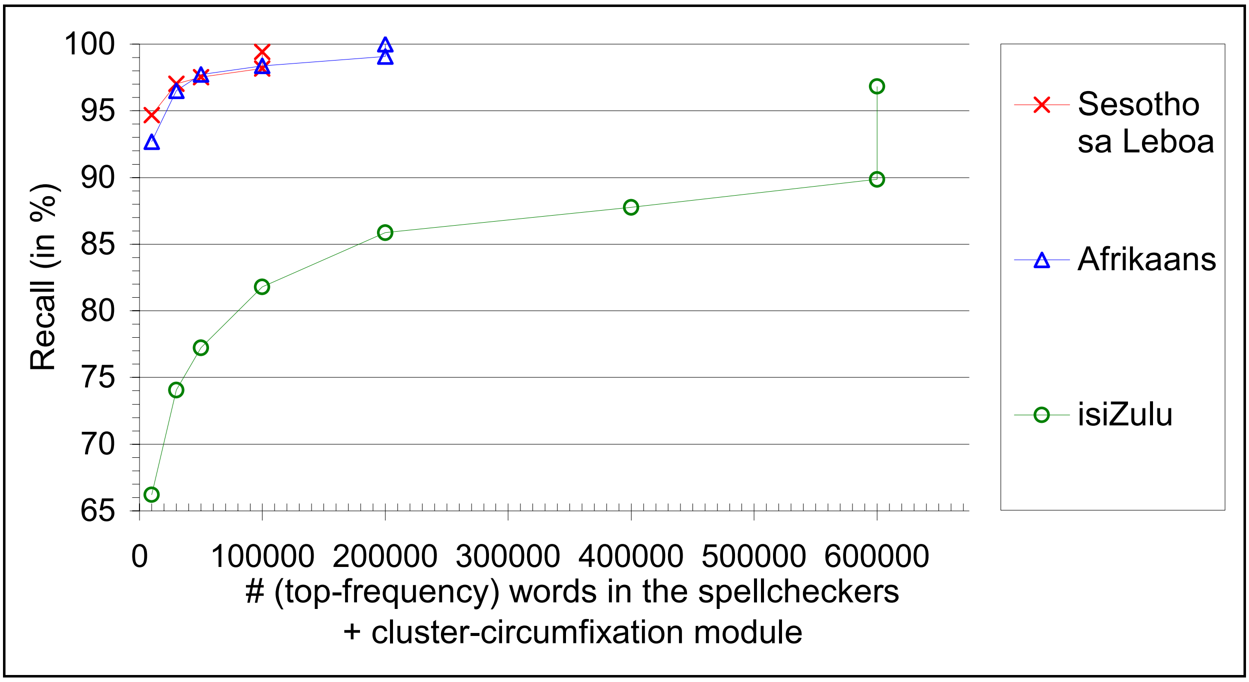
Component 1				
oku-	+	w(u)-	→	**okuw(u)-**
PV		(V)		PV(V)
		(copulative[5])		
Component 2				
-khomba	→	**-umkhomba**		
V		PN-(V)		
		nominalised		
Component 3				
indlela	→	**-ndlela**		
PN-N		(PN)-N		
		minus pre-prefix		

Apart from other difficulties, special attention should be given to sound changes, assimilation, deletion, elision, etc. when affixes are conjoined with nouns or verbs, as in the case of the final vowel of Component 1 versus the initial vowel of Component 2. From this it is clear that an exact morphological, say, finite-state type of description of **okuwumkhombandlela** will be *very* complex, many times more complex than the mentioned Dutch non-word compound *luidsprekeryoghurtijsjes. Therefore, as in the case of single verbs and nouns above, the utilisation of *fewer* rules based on *larger* components should be considered as a possible solution to the quest for improved recall and precision in the case of compounds. This remains to be implemented.

Conclusion

Especially the Nguni languages, which are both morphologically complex and conjunctively written, would benefit from spellcheckers that include morphological analysis and/or generation modules. Given that the development of fully functional transducers for the South African languages is a complicated and time-consuming process, an alternative strategy, namely a partial and statistically motivated morphological decomposition by means of clusters of circumfixes, was introduced. It was indicated that with this approach many more words can either be physically generated and added to a spellchecker lexicon, or that the rules (i.e. the patterns + root lexica on which the patterns apply) could be incorporated in the spellchecker software. The effects of the utilisation of clusters of circumfixes on spellchecking real texts were studied and it was found that a substantial increase in lexical recall could be achieved. The results are summarised in Figure 2. The graph for isiZulu, which represents the Nguni group, clearly indicates that developing the proposal to utilise clusters of circumfixes for spellchecking purposes will indeed be a worthwhile venture.

Figure 2: Increase in lexical recall values for three spellcheckers for the South African languages, as a result of the addition of a basic cluster-circumfixation module (as well as a mixed capitalisation filter for isiZulu)



Notes

1. Since this article is being submitted for publication in South Africa, necessary sensitivity with regard to the term 'Bantu' languages is exercised in the authors' choice rather to use the term African languages. Keep in mind, however, that the latter includes more than just the 'Bantu Language Family'.
2. Circumfixation entails a process of simultaneous affixation of prefix(es) and suffix(es) to a root/stem. In this article the use of the term 'circumfix' is extended to include prefixes in combination with vowel endings of *underived* nouns.
3. The precision also improves, from 2.88% to 8.57%.
4. The precision for Sesotho sa Leboa improves from 20.00% to 33.33%, and for Afrikaans from 14.29% to 100%.
5. The **-w-** is generally regarded as the 'copulative prefix' (variant of the more common **-ng-**) while **u-** is regarded as the pre-prefix of the noun.

References

ANC. [sa]. *What is the African National Congress?* http://www.anc.org.za/about/anc.html (Last accessed: 24 August 2002).

De Schryver, G.-M. & Prinsloo, D.J. 2004. Spellcheckers for the South African languages, Part 1: The status quo and options for improvement. *South African Journal of African Languages* 24(1):57–82.

Jurafsky, D.S. & Martin, J.H. 2000. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River: Prentice-Hall.

Louwrens, L.J. 1991. *Aspects of Northern Sotho Grammar*. Pretoria: Via Afrika Limited.

Van Huyssteen, G.B. & Van Zaanen, M.M. 2003. A spellchecker for Afrikaans, based on morphological analysis, in *TAMA 2003 South Africa: CONFERENCE PROCEEDINGS*, edited by G.-M. de Schryver. Pretoria: (SF)^2 Press:189–194.

Van Wyk, E.B. 1995. Linguistic assumptions and lexicographical traditions in the African languages. *Lexikos* 5 (AFRILEX-reeks/series 5B: 1995):82–96.