Spellcheckers for the South African languages, Part 1: The status quo and options for improvement

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

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

May 2003

In this article an annotated diachronic overview is presented of the field of spelling and grammar checkers with specific reference to the underlying computational techniques. Where appropriate, the various methods are illustrated with data drawn from the official South African languages. The performance of the current South African spellcheckers is subsequently studied, which leads to the conclusion that improvements are needed for especially the Nguni group. Various potential future options to that intent are then looked into. Most illustrations and calculations are carried out on an authentic set of parallel texts entitled "What is the African National Congress?" (ANC, [sa]).

Spellcheckers for the South African languages: Genesis and beyond

Spellcheckers for the South African languages were first developed by D.J. Prinsloo at the end of the 1990s, and proofing tools containing components for isiXhosa, isiZulu, Sesotho sa Leboa and Setswana were made available in *WordPerfect 9*, within the *WordPerfect Office 2000* suite. Basically, each of those spellcheckers consisted of around thirty thousand top-frequency orthographic words only (Prinsloo and De Schryver, 2001:129). In 2003, corpus-based spellcheckers, commissioned by the *Department of Arts and Culture* (DAC), were released for all official South African languages – except for (South African) English (Prinsloo and De Schryver, 2003a). In this second attempt the wordlists are typically several hundreds of thousands of words long, wordlists that can simply be loaded as 'custom dictionaries' into commercial word processors such as *Microsoft Word*.

In Prinsloo and De Schryver (2003b) it is shown that wordlists consisting of a few hundred thousand valid orthographic words successfully push the recall of correctly written text – the so-called *lexical recall* – up to 99% for the disjunctively written African languages (Sesotho sa Leboa, Sesotho, Setswana, Xitsonga and Tshivenda),[1] as well as for Afrikaans. Non-words are thus rather easily picked up for these languages, with only minor confusion as a result of valid words also being flagged as misspellings. When this same approach is followed for the conjunctively written African languages (i.e. the Nguni group: isiXhosa, isiZulu, isiNdebele and siSwati), however, even lexica containing up to half a million orthographic words do not result in lexical recall values higher than 90%.

In the present article, the first in a series of two, the scene is set for the introduction of the utilisation of so-called 'clusters of circumfixes' (the topic of Part 2, cf. Prinsloo and De Schryver, 2004), which are aimed at improving the lexical recall of especially the Nguni group of spellcheckers. In order to fully appreciate the need for the utilisation of such clusters of circumfixes, the present article (Part 1) consists of three main sections laying the groundwork. Firstly, a diachronic overview of the various algorithms and techniques that

* Author to whom correspondence should be addressed.

have been suggested worldwide as candidates for the successful detection and correction of erroneous words in running text is presented. In the second section current spellchecker recall and precision values are calculated and cross-compared for Sesotho sa Leboa, isiZulu and Afrikaans. A parallel text entitled "What is the African National Congress?" (ANC, [sa]) is used for this purpose. In the third and final section several proposals are made to improve the current wordlist-only status quo.

Spellcheckers: A theoretical conspectus

State-of-the-art surveys of spellchecking methods can be found in Kukich (1992), Vosse (1994) and chapters five and six of Jurafsky and Martin's textbook (2000: 141-234). To this day, the work of Kukich remains the definitive reference work, with even Jurafsky and Martin – although writing a decade later – heavily drawing on her work for their spellchecking sections. The discussion that follows here will also follow Kukich. She begins by pointing out that three types of distinctions must be made: (i) error *detection* versus error *correction*; (ii) *interactive* spelling checkers versus *automatic* correction; and (iii) attention to *isolated* words versus linguistic or textual *context*. These distinctions result in the fact that research in this field "has focused on three progressively more difficult problems: (1) nonword error detection; (2) isolated-word error correction; and (3) context-dependent word correction" (Kukich, 1992:377). Each of these problem areas will now be reviewed and exemplified with South African language data where appropriate.

Non-word error detection (early 1970s – early 1980s)

A non-word can be defined as a continuous string of letters and/or numbers that does not appear in a given lexicon or that is not a valid orthographic word form. From the early 1970s to the early 1980s, research focused on the detection of such non-words, and the two main techniques from that period, n-*gram analysis* and *dictionary lookup*, are still in use today (either as a basis, in combination, or together with more complex approaches). Current spellchecking efforts for Afrikaans, for instance, include at least one research team that departs from a combination of *n*-gram analysis and dictionary lookup, to which rule-based morphological analysis modules are added (Van Huyssteen and Van Zaanen, 2003).

'*N*-gram analysis' will be looked into first. Letter *n*-grams are sequences of *n* letters, with *n* usually 1, 2 or 3, respectively called letter unigrams, bigrams and trigrams. (Likewise, word *n*-grams are sequences of *n* words. These will be discussed further below.) In general, a method based on (letter) *n*-grams will determine whether or not each *n*-gram of an input string is attested in the language. To do so, comparisons are made with a precompiled table of *n*-gram statistics, where those statistics can be binary values or frequency counts. Such tables may take the position, such as beginning, end, etc., of the *n*-gram within a word into account (positional *n*-gram arrays), or they may be general (non-positional *n*-gram arrays). As an illustration of a simple *n*-gram array, a section of a *non-positional binary bigram* array for Sesotho sa Leboa is shown in Table 1.

Table 1: Non-positional binary bigram array for Sesotho sa Leboa

	a	**b**	**c**	**d**	**e**	**f**	**...**
a	1	1	0	1	1	1	
b	1	0	0	0	1	0	
c	0	0	0	0	0	0	
d	1	0	0	0	1	0	
e	1	1	0	1	1	1	
f	1	0	0	0	1	0	
...							

A letter sequence **bb** in a Sesotho sa Leboa text, for instance, will immediately be *detected* as an error, as the two-dimensional bigram array shows a binary value '0' for this combination. Conversely, letter sequences such as **ba** or **be** are perfectly acceptable. An analogous array for isiZulu is shown in Table 2.

Table 2: Non-positional binary bigram array for isiZulu

	a	**b**	**c**	**d**	**e**	**f**	**...**
a	1	1	1	1	1	1	
b	1	0	0	0	1	0	
c	1	0	0	0	1	0	
d	1	0	0	0	1	0	
e	1	1	1	1	1	1	
f	1	0	0	0	1	0	
...							

Binary *n*-gram analysis techniques have been used with success in optical character recognition (OCR) devices, yet for spellchecking applications *probabilities* instead of binary values should at least be used. Such probabilities are derived from frequency counts in large electronic corpora (of at least a million words), and are more fine-grained. Tables 3 and 4 show a section of the occurrence probability of non-positional bigrams per 1,000 running words for Sesotho sa Leboa and isiZulu.

Table 3: Non-positional frequency bigram array for Sesotho sa Leboa (per 1,000 running words in a 5.8-million-word Sesotho sa Leboa corpus)

	a	**b**	**c**	**d**	**e**	**f**	**...**
a	2.91	16.41	0	8.93	6.45	3.64	
b	74.90	0	0	0	25.66	0	
c	0	0	0	0	0	0	
d	1.13	0	0	0	0.14	0	
e	3.52	14.71	0	8.92	4.91	4.20	
f	8.60	0	0	0	12.45	0	
...							

Table 4: Non-positional frequency bigram array for isiZulu (per 1,000 running words in a 5.0-million-word isiZulu corpus)

	a	**b**	**c**	**d**	**e**	**f**	**...**
a	0.18	75.26	3.84	11.42	0.18	8.94	
b	102.05	0	0	0	51.25	0	
c	8.48	0	0	0	5.10	0	
d	23.47	0	0	0	17.65	0	
e	0.52	18.53	1.70	8.56	0.08	3.20	
f	10.70	0	0	0	1.92	0	
...							

Upon comparing Table 3 with Table 1, or Table 4 with Table 2, it is clear that one can achieve more with frequency counts than with binary values. For example, although the letter sequences **ba** and **da** are both 'valid' in Sesotho sa Leboa (cf. Table 1), **ba** is many times more frequent than **da** (cf. Table 3). Actually, the latter combination only occurs in loanwords, foreign place names, and the like, such as in **da**enamo 'dynamo' or **Kanada** 'Canada' (cf. e.g. Departmental Northern Sotho Language Board, 1988[4]:161 and 114 respectively), **Mojuda** 'Jew', **radara** 'radar', etc. From the non-positional frequency bigram array for isiZulu one can for example deduce that the frequency of the combination **da** (23.47 times per 1,000 running words) is roughly twice that of **ad** (11.42 times per 1,000 running words). IsiZulu words containing **aa** (such as in **aaa!** or **aaah!**, exclamations of pain) or **ee** (such as in **yeheeeeni**, used to mimic the wailing sound the person in question is making) are exceptional, which explains the low values for these sequences in Table 4. If such juxtapositions of vowels are encountered while spellchecking isiZulu text, the likelihood that one is dealing with a misspelling is (extremely) high indeed.

Higher-order *n*-gram tables are obviously even richer in information, and will for example show that,

for Sesotho sa Leboa, the trigram **bae** is relatively infrequent compared to the trigram **abe** (0.27 vs. 3.34 times per 1,000 running words), or that, for isiZulu, the trigram **ceb** is three times more frequent than the trigram **bec** (0.48 vs. 0.16 times per 1,000 running words). For Afrikaans, Van Huyssteen and Van Zaanen report that there are 129,374 valid positional quadrigrams, which they store according to the beginning, middle or end of words (2003: 191). Generally, thresholds are set for the various *n*-grams, and when the probability goes below that threshold, the word being spellchecked is flagged as a potential error (Hirst and St.-Onge, 1998:320).

The second non-word error detection technique is known as 'dictionary lookup'. Taken at face value this method is straightforward, as an input string is simply compared to items in a spellchecker lexicon, and if this string is absent from that lexicon, it is considered to be a misspelling. For many languages the situation is obviously a little more complex as, even today with soaring storing capabilities and processor speeds, it is still not – and will probably never be – feasible to include *all potential* orthographic word forms of a particular language (i.e. all inflections, all derivations, all compounds, and in turn all inflected and derived compounds, etc.). It is furthermore true that, as spellchecker lexica grow larger, the response time goes down. Research has therefore also focussed on reducing dictionary search time, and resulted in now-common techniques such as: (i) hash tables, (ii) tries (this term is derived from 'information retrieval'), (iii) frequency-ordered binary search trees, (iv) finite-state automata,[2] (v) dictionary-partitioning techniques, and (vi) morphological-processing routines such as for example stemming (also know as affix-stripping).

It is well known that listing *all* orthographic forms is for instance untenable for Finnish, where word roots may easily have thousands of inflectional forms each. This very fact led Finnish researchers to design the first finite-state morphological analysers/generators (Koskenniemi, 1983; Karttunen, 1983). Pure dictionary lookup techniques are thus not feasible for any morphologically complex language, especially if the orthography is characterised by a high degree of conjunctivism. One of the extreme examples is Turkish, which is estimated to have a few hundred forms per noun root, a few million forms per verb root, and *200 billion* orthographic words in all (Hankamer, 1989).

Isolated-word error correction (1960s – present)

It is one thing to detect a non-word or misspelling but another thing entirely to suggest an alternative or even to correct the error automatically. Most present-day end users will expect the software to be able to correct typos automatically and will want to be provided with a series of plausible alternatives for an error; so reviewing the common techniques in this regard is required.

Research into isolated-word error correction already started in the 1960s (cf. e.g. Blair, 1960) and continues up to this day. Often, the design of a particular correction technique is based on a study of spelling error *patterns*. In this respect distinctions are generally made between typographic, cognitive and phonetic errors. The latter are obviously much less of a problem for the African languages (when compared to, say, English or French), since there is always a close and systematic correspondence between the orthography and the pronunciation of African languages. Summarising data for English one may conclude that:

> (1) most errors (i.e., roughly 80%) tend to be single instances of insertions, deletions, substitutions, or transpositions; (2) as a corollary, most errors tend to be within one letter in length of the intended word; and (3) few misspellings occur in the first letter of a word. These findings have been exploited by implementing fast algorithms for correcting single-error misspellings and/or by partitioning dictionaries according to first letter and/or word length to reduce search time. (Kukich, 1992:392)

Preliminary studies confirm this finding for the African languages at large. Kukich finds it convenient to group the approaches for isolated-word error correction into six main classes, with each entailing "three subproblems: (1) detection of an error; (2) generation of candidate corrections; and (3) ranking of candidate corrections" (1992: 392). Each of these six classes will now be described concisely, yet it should be remembered that today's trend is towards hybrid approaches, combining several of the mentioned classes.

Firstly, in *minimum edit distance* techniques a 'minimum edit distance' (i.e. the smallest possible number of insertions, deletions and substitutions) is calculated to transform an erroneous word into an item from the spellchecker lexicon (Levenshtein, 1965; Wagner and Fisher, 1974). This is an application of the dynamic programming technique (Jurafsky and Martin, 2000:153-156). For example, a spellchecker that is run on the isiZulu version of the document "What is the African National Congress?", will flag the incorrectly

spelled word *ayisishayagalolunye and suggest **ayisishiyagalolunye** 'nine'. Just one single *substitution* (**a** → **i**) is required to turn a non-word into a valid word, making this suggestion highly probable indeed.

Secondly, in *similarity key* techniques, similarly spelled strings are 'mapped' onto identical or similar 'keys'. Already around the 1920s Odell and Russell (1918-1922) created *Soundex* on this principle, for use in English *phonetic spelling correction*. In their system the key consists of the first letter of the input, followed by three digits (where zeros are eliminated, consecutive duplicate digits collapsed, and, if necessary, digits beyond the third one dropped or trailing zeros added) according to:

A, E, I, O, U, H, W, Y → 0
B, F, P, V → 1
C, G, J, K, Q, S, X, Z → 2
D, T → 3
L → 4
M, N → 5
R → 6

Assuming that Sesotho sa Leboa would follow the sound pattern of English, and thus that the above system would have been designed for it, one could for example have: **motho** 'person, human being' → M0300 → M300 and also *moto → M030 → M300; or **bohlokwa** 'scarcity, preciousness' → B0040200 → B420 and also *bohlokua → B0040200 → B420. The idea is thus that the key of a misspelled string would be identical or similar to the key of the correct item. This Soundex system is only used as an illustration here, as it is far from perfect; yet it exemplifies the notion of a similarity key technique in a simple way. For more information, see Jurafsky and Martin (2000: 89) and Parsons (2003).

Thirdly, in *rule-based* techniques knowledge about spelling error patterns is used to write rules with which misspellings can be rewritten as correct items. Spelling error patterns are, for instance, currently being collected for Sesotho sa Leboa (cf. below).

Fourthly, in n-*gram-based* techniques letter *n*-grams are combined in several ways. For example, above it was indicated that the Sesotho sa Leboa trigram **bae** is relatively infrequent compared to the trigram **abe**. A misspelling such as *thabae will be flagged by a spellchecker, and the highest-ranking suggestion is **thaabe** 'hiccup'. In this case the error would have been the result of a single *transposition* (**a** ↔ **b**). (The second-highest suggestion is **thaba** 'mountain; be happy', in which case the error would have been the result of a single *insertion* (+ **e**).[3])

Fifthly, there are two types of *probabilistic* techniques. *Transition* or *Markov* probabilities deal with the likelihood that a certain letter (sequence) will be followed by a given letter. *Confusion* or *error* probabilities deal with the likelihood that a certain letter replaces another given letter. All these techniques use *n*-grams. Candidates for replacing the misspelled word are generally ranked by means of a Bayesian (or noisy channel) algorithm, as this is usually the easiest one to compute (Kernighan, Church and Gale, 1990; Jurafsky and Martin, 2000:149-153). Note that the noisy channel approach can also be used to model context-dependent errors (cf. below).

Lastly, *neural nets* have the potential (if problems such as linear separability and training dependency can be overcome) to be trained on real user-errors, and thus to 'adapt' to the specific problems of a user community (cf. e.g. Vijaykumar, 1992).

Context-dependent word correction (early 1980s – present)

During the last two decades increasing efforts have been devoted to the correction of a third type of errors, namely real-words that are wrong only when the context (and by extension the grammar) is taken into account. A mistake is thus made whereby a valid word substitutes an intended word. Studying data for English, Kukich concludes that "real-word errors account for anywhere from 25% to well over 50% of all textual errors depending on the application" (1992: 429), and thus that for 'good' spellcheckers, "the greatest gain is to be found in the detection and correction [...] of real-word errors" (1992: 413). Again, preliminary data for the African languages (cf. below) confirm this breakdown. These high percentages thus suggest that context-dependent word correction will also be vital for the creation of state-of-the-art African-language spelling (and grammar) checkers.

Before reviewing the main techniques of context-dependent word correction, it is appropriate to look into a few typical real-word errors. This is done in the list below; in each case examples are drawn from Sesotho sa Leboa:

- **Typographic errors**, for short 'typos' (the result of a motor coordination slip): reka 'buy' → reta 'praise'; monna 'man' → mona 'envy'; bana 'children' → banna 'men'
- **Cognitive and phonetic lapses** (the result of a misconception or a lack of knowledge): boa 'come back' → bua 'skin, operate'
- **Syntactic or grammatical mistakes**, such as:
  - wrong inflected forms: reka 'buy' → reke '(must) buy'
  - wrong function words: yo 'this (with class 1 nouns)' → wo 'this (with class 3 nouns)'; gagwe 'his/hers' → gago 'yours'
- **Semantic anomalies**: gotša mollo 'make a fire' → *lotšha mollo '*greet + fire'; boa gosasa 'come back tomorrow' → *goa gosasa '*shout + morning'
- **Insertions or deletions** of whole words: *Motho wa batho, ba lekile go mo thuša motho fela ba paletšwe ruri. '*The poor guy, they tried to help him the man but they failed dismally.'
- **Improper spacing**, such as:
  - splits: Tšea seripa se. 'Take this slice.' → *Tšea se ripa se. '*Take cut it this.'; A laetše gore ba mmolaye. 'He gave the command that they should kill her.' → *A laetše go re ba mmolaye. '*He gave the command to say they should kill her.'
  - run-ons: Ba tlo se bala. 'They will read it.' → *Ba tlo sebala. '*They will small white spot.'
  - splits / run-ons: Nka se lebale. 'I shall not forget.' ↔ Nka se le bale. 'I shall not read it.'
- **Orthographic intrusion errors** (i.e. substitutions with nearby graphemes): Batho ba palelwa ke dilo tše dintši gobane ba dumela go palelwa. 'People are conquered by many things because they agree to be conquered.' → *Batho ba palelwa ke dilo tše dintši gobane ba dumela go palela. (palelwa 'be conquered' → palela 'conquer'); E tlo mo lekana gabotse. 'It will fit her very well.' → *E tlo mo gakana gabotse. (lekana 'fit' → gakana 'mislead')

Without *contextual* information, such real-word errors cannot be detected nor corrected. There are three main domains of techniques that attempt to handle context-sensitive spelling errors. These are natural language processing, statistical language modelling and neural net modelling techniques.

In *natural language processing* (NLP) techniques, five increasingly more difficult levels of processing constraints can be identified, viz. lexical, syntactic, semantic, discourse and pragmatic. Most NLP systems are parser-driven; hence they try to chart *syntactic* structure. In *acceptance*-based approaches errors are ignored as long as 'some' interpretation can be given; in *relaxation*-based approaches the opposite is true, thus no errors can be ignored; and in *expectation*-based approaches a list is built of words the parser expects to see in the next position.

Kukich defines *statistical language modelling* (SLM) techniques as being "essentially tables of conditional probability estimates for some or all words in a language that specify a word's likelihood to occur within the context of other words" (1992: 423). She gives the following examples:

- **word trigram SLM**: the probability distribution for the next word is conditioned by the two previous words;
- **part-of-speech (POS) bigram SLM**: the probability distribution for the POS of the next word is conditioned by the POS of the previous word;
- **collocation SLM**: the probability distributions for certain words to occur within one another's vicinity.

For example, a well-known application of a POS bigram SLM is the *Constituent Likelihood Automatic Word-tagging System* (CLAWS, [sa]), a part-of-speech tagger for English text with which the 100-million-word *British National Corpus* (BNC, 2002) was POS-tagged. Jurafsky and Martin indicate that there are many other SLM approaches to context-dependent word correction – including Bayesian classifiers, decision lists, transformation-based learning, latent semantic analysis and Winnow – yet point out that all "these algorithms are very similar in many ways; they are all based on features like word and part-of-speech *N*-grams" (2000: 222). As a 'word *n*-gram model' uses the previous *n-1* words to predict the next one, a word trigram SLM is also known as a second-order Markov model since it looks two words into the past. What all this means in

simple terms is that spellcheckers employing SLM techniques will generally flag low word and/or POS probability combinations as potential errors, and suggest more likely replacements (Hirst and St.-Onge, 1998:321).

Lastly, it can be pointed out that *neural net modelling* (NNM) techniques are still in an experimental phase.

While a strict division between natural language processing and statistical language modelling methods might still have been appropriate a decade ago, Jurafsky and Martin rightly point out that "[b]y the last five years of the millennium it was clear that the field was vastly changing [as] probabilistic and data-driven models had become quite standard throughout natural language processing" (2000: 14). Today, the fields have indeed come together as is evident from for example cases where semantic information is used to enrich *n*-grams or from the fact that a number of augmentations to *n*-grams are based on discourse knowledge (Jurafsky and Martin, 2000:231).

Spellcheckers for the South African languages: Current recall and precision values

Summarising the theoretical conspectus in a nutshell, one can say that during the past few decades most spelling-error detection algorithms have been based on *dictionaries*, while most spelling-error correction algorithms relied on *dynamic programming* (such as the 'minimum edit distance' technique). In the nineties, *probabilistic* algorithms have been employed in addition for correcting spelling errors. Turning to South Africa: As excellent language-independent software for suggesting alternatives to non-words is already linked to current word processor spellcheckers, the focus in South Africa so far has been on non-word error detection, and not on the correction of those errors. In the present section the performance of the current wordlist-based spellcheckers will be briefly evaluated.

In the series of tests that are to follow three languages will be used, viz. Sesotho sa Leboa, isiZulu and Afrikaans. IsiZulu represents the Nguni group, and Sesotho sa Leboa the disjunctively written South African languages. In order to be able to somehow cross-compare the results directly, different language versions of the same source text, namely a randomly chosen document, available over the Internet and entitled "What is the African National Congress?", will be spellchecked. The spellcheck-methodology is comparable to the one expounded in Van der Veken and De Schryver (2003) and Prinsloo and De Schryver (2003b), and basically entails that top-frequency types are extracted from corpora and then used as spellchecker lexica. For the tests below, spellcheckers with an increasing number of attested orthographic words will be applied one after the other: first containing only the top 10,000 orthographic words of the language, then the top 30,000, next the top 50,000, etc. Lexical recall values will be calculated in each case.

The data for Sesotho sa Leboa are shown in Addendum A. From these data it can be seen that this language version of the ANC text is 1,382 words long, and that 4 non-words were uncovered. These 4 errors are flagged by the various spellcheckers, so the *error recall* is 100% in each case.[4] When the spellchecker is loaded with just the top 10,000 words, 64 types are falsely flagged as misspellings. These types have been marked with the following underscores: 'token', 'token', 'token' and 'token'. When the spellchecker is loaded with the top 30,000 words, the number of falsely flagged types is cut in half, to 32. The words marked as 'token' are now recognised. With the top 50,000 words in the spellchecker lexicon, the words marked as 'token' are now also recognised, leaving 25 types unrecognised. Finally, with the top 100,000 words, only the 16 types marked as 'token' remain falsely flagged. Both lexical recall and precision values may be calculated from the user's point of view (p.v.). This means that the number of flagged types is compared to the number of running words (i.e. the tokens) in the document (cf. Van der Veken and De Schryver, 2003:38, and Prinsloo and De Schryver, 2003b:318). Given that all 4 non-words in the text *are* flagged by the spellchecker, and given that 16 valid types remain flagged even with the largest spellchecker in this test, the maximum *precision* from the user's point of view is only 4 out of 20 (= 16 + 4), or thus 20.00%. The *lexical recall* from the user's point of view increases from 95.36% with just 10,000 words in the spellchecker lexicon, up to 98.84% with 100,000 words. For the time being such recall and precision values for Sesotho sa Leboa, values which are also found for a variety of text types and for the other disjunctively written African languages, are satisfactory. It should nonetheless be clear that better results could be obtained with larger wordlists – which simply implies the construction of larger corpora for these languages.

Recall and precision values with wordlist-based spellcheckers for the Nguni languages are less satisfactory. This is immediately clear from the data for isiZulu presented in Addendum B. The total number of orthographic words for the isiZulu version is 1,008 – thus, as a result of the conjunctive orthography, lower

than for Sesotho sa Leboa. The text contains 3 non-words, and all are flagged by the spellcheckers. The *non-word error recall* is thus 100%. One could naively assume that all the words in an official document of this nature would be both *correctly spelled* and *correctly used*. The fact that only 3 words were misspelled is satisfying. However, a careful reading also reveals that 3 correctly spelled words are used incorrectly. Recall that Kukich suggested that 25% to well over 50% of all textual errors are real-word errors (1992: 429). With 3 real-word errors out of 6 textual errors, or thus 50%, this proportion is confirmed here. The *overall error recall* is thus only 50.00% (3 out of 6 errors). Detecting and correcting context-dependent errors is of course not possible with the modules in the current wordlist-only spellcheckers. Whereas 10,000 spellchecker items push the *lexical recall* of a Sesotho sa Leboa spellchecker to over 95%, this same number of items in an isiZulu spellchecker is 30% less effective. Even 100,000 items only push the lexical recall to around 80%, 200,000 items to around 85%, and as many as 600,000 items still only up to around 90%. Consequently, *precision* values are also extremely low. Here, with 3 non-words correctly flagged and 101 types falsely flagged in the best case, the precision is still only 2.88%. As comparable values are found for tests with other texts and for the other Nguni languages, there is a great need indeed for ways to substantially increase the lexical recall in spellcheckers for these languages. Larger wordlists, and thus much larger corpora, could be a solution, although this option does not seem to be the ideal one.

In order to complete the South African picture, an analogous set of tests was also performed on the Afrikaans version of the ANC text. The data can be found in Addendum C. The Afrikaans text has 1,287 tokens, of which 2 non-words. Both these errors are again flagged by the spellcheckers, so the *error recall* is 100%. *Lexical recall* values are comparable to the Sesotho sa Leboa data, achieving a recall of 98.37% with 100,000 items. Doubling the number of items pushes this recall to 99.07%. The maximum *precision* from the user's point of view, with the 2 non-words flagged and 12 types falsely flagged, is 14.29%. Including more orthographic words in an Afrikaans spellchecker lexicon could further improve the performance.

Spellcheckers for the South African languages: Forthcoming developments

The tests described in the previous section summarise the current status quo. In the present section various forthcoming developments are briefly looked into, from trivial to more advanced developments.

Automatic error correction

In the discussion so far, only *existing* texts were spellchecked, but spellcheckers are of course first and foremost employed *while typing* text. In this context Kukich's distinction between *interactive* spelling checkers versus *automatic* correction (cf. the theoretical conspectus above) is highly relevant. Indeed, a first addition to the current wordlist-based status quo will be to add a feature that can *automatically* detect and correct frequent typographical errors users are known to make. Basically this implies that a corpus of errors is studied from which the frequent typos are extracted. These typos are then loaded, paired with their correct spelling, into the *autocorrect* feature of a spellchecker. For Sesotho sa Leboa, for example, one of the ways that is currently being used for the collection of authentic misspellings is through a study of the log files of an online Sesotho sa Leboa dictionary (De Schryver and Joffe, 2003). It is a trivial matter to extract and to calculate the occurrence frequencies of all the errors typed in by the users of that dictionary. Compare in this regard Van Huyssteen's *Speltoetserkompetisie* for Afrikaans (2002).

Language identification

A second development consists of a module to automatically detect the language when an existing document is opened or when text is being entered. Many documents in South Africa are of a multilingual nature, and it would indeed be ideal if a spellchecker could recognise where, say, sections in Tshivenda begin and end, or where those in isiXhosa and Setswana do. A system "based on calculating and comparing profiles of N-gram frequencies" (Cavnar and Trenkle, 1994:161) has already been developed for the African languages by Jacky Maniacky and is available online (Maniacky, 2003). At the time of writing, this program – which is appropriately called *Umqageli*, isiZulu for 'diviner' – already gives excellent results for Sesotho sa Leboa,

Sesotho, Setswana, Tshivenda, isiXhosa and siSwati. For isiZulu and isiNdebele this software hesitates between various Nguni languages, while Xitsonga is not yet supported. Dozens of other African languages are also correctly recognised by Umqageli.

Capitalisation

A third development concerns capitalisation. This is best illustrated with the isiZulu version of the ANC text. Recall that, with the largest spellchecker engaged, 101 correct isiZulu types remain flagged. Of these 101, 22 types are however recognised by the spellchecker if they are presented either in lower- or uppercase but not in mixed case, such as **Ekomidi** or **EKOMIDI** 'committee' but not **eKomidi**. In this specific text there is a general tendency to *over-capitalisation* of words. Over-capitalisation here ranges from preference for writing, in mid-sentence, the phrase **abantu bayobusa** 'the people shall govern' as **Abantu Bayobusa**, to instances where capitalisation is performed in an unsystematic way, such as **eMhlanganweni KaWonke-wonke wonyaka** followed by "(Annual General Meeting)". At least **wonyaka** should then also have been capitalised to be on a par with its equivalent 'Annual'. False flags that are the result of an imprecise treatment of capitalisation are not uncommon in spellcheckers, and have for instance been reported by Paggio (2000: 260). Apart from the fact that it is doubtful whether spelling rules were applied correctly in the above isiZulu examples, a filter accepting any mixed lower- and uppercase could be added to the spellchecker. End users could then choose whether or not to engage this filter. With such a filter, only 79 types would remain wrongly flagged, with the lexical recall going up to 92.14% (and the precision to 3.66%).

N-grams

From the above presentation of the status quo, it is clear that although pure dictionary lookup techniques are 'valuable' for the construction of spellcheckers for the Nguni languages, they are not sufficient. Alternative or additional solutions have to be found, and for instance include the utilisation of *n*-grams to recognise more valid words (even though this approach also implies that non-words may then also happen to be accepted as correct). In this regard, two sets of tests were undertaken for isiZulu, and will be briefly explained here. In the first test various valid *n*-grams were first generated from a sample of randomly selected recent corpus material. Table 5 illustrates how the words were broken down into *n*-grams, in this case quinquegrams.

Table 5: Generating valid *n*-grams (here quinquegrams) from valid words in isiZulu

Word	Quinquegram Beginning	Quinquegram Middle # 1	Quinquegram Middle # 2	(Quinquegram Middle # …)	Quinquegram End
kakhulu	kakhu	akhul			khulu
sengathi	senga	engat	ngath		gathi
ekhaya	ekhay				khaya
njengoba	njeng	jengo	engob		ngoba

The number of positional occurrences of the various *n*-grams were then counted. Tables 6 to 8, for example, show the resulting top 10 lists for the trigrams, quadrigrams and quinquegrams respectively, with an indication of the occurrences (for the sample corpus material) broken down into frequencies at the beginning, in the middle or at the end of a word.

As the next step, random paragraphs were taken from the isiZulu *Bona* magazine of April 2003, and spellchecked with the commercial WordPerfect 9 spellchecker. Of the 1,038 words in the Bona test paragraphs, the spellchecker only recognised 607 words, or thus only 58%. The 431 words not recognised are shown in Addendum D. This is a clear example of the fact that this wordlist-based spellchecker's recall needs to be improved. The question was thus whether or not the utilisation of the generated sets of valid *n*-grams, of which Tables 6 to 8 are extracts, could contribute to improved recall. In principle this was done by breaking up the 431 words into their respective tri-, quadri- or quinquegrams, as indicated in Table 9 for a number of these words in respect of quinquegrams.

Table 6: Top 10 list of trigrams in isiZulu

Trigram	Beginning	Middle	End	Total
nga	697	1526	538	2761
ngi	1070	877	101	2047
aba	533	836	313	1682
ang	304	1309	1	1614
ela	42	214	1302	1557
ele	13	452	1029	1494
nge	489	740	130	1358
ama	352	751	248	1350
ath	48	1087	2	1137
ezi	373	661	95	1128

Table 7: Top 10 list of quadrigrams in isiZulu

Quadrigram	Beginning	Middle	End	Total
anga	112	359	301	771
ngen	108	372	0	480
inga	37	356	51	444
ezin	130	310	0	440
khul	5	384	0	389
bona	2	122	252	375
unga	112	179	80	371
atha	10	198	152	360
ngok	288	70	0	358
aban	97	257	0	354

Table 8: Top 10 list of quinquegrams in isiZulu

Quinquegram	Beginning	Middle	End	Total
khulu	4	197	101	302
ngoku	211	48	1	259
ngiya	164	28	3	194
kwenz	18	172	0	190
njeng	168	21	0	189
thath	2	187	0	189
thand	2	184	0	186
langa	3	102	77	181
abang	50	120	0	170
ngama	99	65	1	164

Table 9: Generating and validating *n*-grams (here quinquegrams) from/for unrecognised words in isiZulu

Word	Quinquegrams
esasisizwa	esasi (12), sasis (1), asisi (1), sisiz (6), isizw (3), sizwa (37)
ezakhaya	ezahk (3), zakha (1), akhay (10), khaya (31)
ezikhombise	ezikh (14), zikho (16), ikhom (30), khomb (76), hombi (52), ombis (52), mbise (9)
ezingomeni	ezing (41), zingo (10), ingom (3), ngome (2), gomen (3), omeni (6)
ikhishiwe	ikhis (1), khish (17), hishi (3), ishiw (3), shiwe (12)

Each value between brackets in Table 9 is an indication of how often the specific quinquegram occurred in the original isiZulu corpus sample (cf. Table 8). In this test this value merely had to be at least 1. This means that each and every quinquegram in each of the words had to have an occurrence of at least 1 in the original list; if not, the whole word would (again) be flagged as incorrect. The outcome of this test was that as many as 99

words were now also flagged as correct. This is equivalent to an improvement of 10% in lexical recall. Compare Addendum E in this regard. This test thus indicates that a spellchecker for a Nguni language will likely improve by complementing the dictionary lookup module with an *n*-gram module.

In the previous test, the lexical recall of correctly spelled isiZulu was checked. It is of course also important to check the value of an *n*-gram speller to detect typical typing errors, or thus to study the error recall. In this regard, ten typical African-language errors were introduced (as indicated below) in the following passage:

wayeneminyaka	**wayenemminyaka**	(typing a letter twice)
ngesikhathi	**ngeesikhathi**	..
odokotela	**okodotela**	(switching two syllables around)
ukuthi	**ukukuthi**	(typing a syllable twice)
unomdlavuza	**unomdlavuvuza**	..
eminyakeni	**eminyakini**	(typing a vowel incorrectly under the influence of the following vowel)
emithathu	**emithuthu**	..
uvamile	**vamile**	(omitting the first letter)
beminyaka	**beminyak**	(omitting the last letter)
uyahlasela	**uyahasela**	(omitting a letter)

uQetello Zeka ongummeli wasecape Town **wayenemminyaka** engu-28 **ngeesikhathi okodotela** bethola **ukukuthi unomdlavuvuza** a wamabele **eminyakini emithuthu** eyedlule - lokho kwamenza wazibona etana nenhlanzi eshelwe amanzi. umdlavuza **vamile** ukuhlasela abesimame **beminyak** eyevile kwengu50 kodwa futhi nakwabancane **uyahasela**.

With $n = 5$, all words with the exception of ***ukukuthi** were detected as errors; with n = 4 ***eminyakini**, ***emithuthu** and ***uyahasela**, in addition to ***ukukuthi**, were not detected as errors; and with n = 3 one more word was not detected. This test again indicates that the current wordlist-only spellcheckers will indeed likely gain from the addition of *n*-gram modules, and thus that the creation of *n*-gram modules for application on the African languages and/or the adaptation of internationally available programs will be a worthwhile venture.

Finite-state morphology

Despite the fact that wordlists cum *n*-gram spellers seem to have a lot of potential, algorithms for detecting and correcting spelling errors are increasingly being worked out within the *finite-state* paradigm of natural language processing. Indeed, the need to include *linguistically motivated* morphological representations in spellcheckers – and thus to be able to decompose orthographic words morphologically – has become especially apparent for morphologically complex languages with a high degree of conjunctivism. The Nguni group belongs to those languages. It is now a particularly exciting time to write morphological analysers/generators for orthographic words in natural languages, since, at the time of writing, the Xerox finite-state tools are being released (Beesley and Karttunen, 2003). Many regard these tools as the most sophisticated available (cf. e.g. Daciuk, 2003),[5] and they are likely to enable the so-called 'lesser-studied languages' to leap into the new millennium. As far as linking Xerox finite-state technology to a word-processor spellchecker is concerned, Ken Beesley nonetheless points out that: "Finite-state techniques have been used, commercially, to implement both spelling checkers and spelling correction, but I don't think that we've actually published the techniques" (*personal communication*, 8 March, 2002). Compare, however, with Solak and Oflazer (1993), Oflazer and Güzey (1994), Oflazer (1996), and Kaplan and Newman (1997).

Even so, some African-language research teams have been at the forefront of applying finite-state methods, and while no spellchecker-specific publications have appeared so far, articles describing general finite-state morphological endeavours for the African languages have. The earliest effort comes from Finland, where Arvi Hurskainen already started work in 1987 on a two-level formalism for Kiswahili – thus fairly soon after Koskenniemi (1983) had described the first finite-state morphological rules for Finnish. Results were published by Hurskainen (1992; 1995; 1996; 1999) for Kiswahili, and Hurskainen and Halme (2001) for Kwanyama. More recently, other African-language finite-state teams have been formed in Norway/Sweden/Zimbabwe for ChiShona and SiNdebele, and in South Africa for isiZulu, Sesotho sa Leboa,

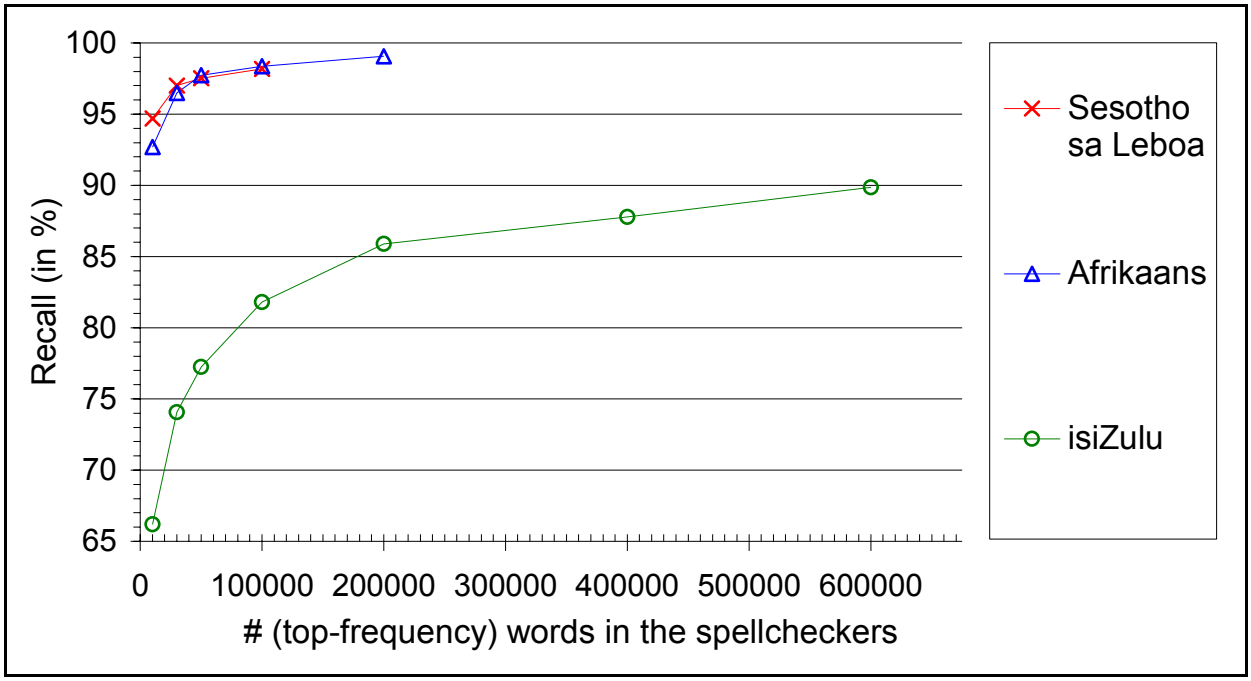
isiXhosa, etc. The efforts of the former team are also based on the somewhat dated two-level model, while all South African projects employ the Xerox finite-state programming languages *xfst* and *lexc*. Several members of all these teams gathered at the *7th International Conference of the African Association for Lexicography* in July 2002, for a special session devoted to finite-state morphological analysers/generators (Bosch and Pretorius, 2002; De Schryver, 2002; Hurskainen, 2002; Maphosa, 2002; Ridings and Mavhu, 2002). In addition, the isiZulu team has been particularly active and reported various results (Pretorius and Bosch, 2001 [2002a]; 2002b; 2003; Bosch, Pretorius and Van Huyssteen, 2003). To this date, the only commercially available African-language spellchecker that resulted from this finite-state work, however, is Hurskainen's *Orthografix 2 for Swahili* (Lingsoft, 1999). With this spellchecker, some 45,000 word roots cum morphological components, allow for the recognition of tens of millions of orthographic words.

The ultimate goal of the South African teams is to design fully functional finite-state networks for both morphological analysis and generation, or thus to design 'finite-state transducers', that could also be put to good use in spellcheckers for the South African languages. While awaiting the completion of these projects, however, the need was felt to come up with an approach to morphological decomposition that (i) could be developed without delay, and (ii) could also be designed with basic software. This approach, which makes use of clusters of circumfixes, deserves a discussion of its own, however, and will be the topic of Part 2 (cf. Prinsloo and De Schryver, 2004).

Conclusion

In a comprehensive theoretical conspectus at the start of this article the three major spelling and grammar checking research fields, viz. (i) non-word error detection, (ii) isolated-word error correction, and (iii) context-dependent word correction, were reviewed. Past and current South African endeavours have mainly concentrated on the first field. Recall and precision values for current (wordlist-based) South African spellcheckers were then calculated on a set of parallel texts, namely for different language versions of the document "What is the African National Congress?" The main results of these tests have been summarised in Figure 1.

Figure 1: Current lexical recall values for three spellcheckers for the South African languages



It was indicated (and this may also be deduced from Figure 1) that especially the Nguni languages, which are both morphologically complex and written conjunctively, would benefit from improved spellcheckers. Several proposals were then made to improve the current wordlist-only status quo. These proposals include automatic

error correction, automatic language identification, mixed capitalisation filters, the use of *n*-gram modules, and the development of (finite-state) morphological analysis and generation components. Given that the development of fully-functioning finite-state transducers for the South African languages will not be achieved anytime soon, the development of an approach revolving around so-called 'clusters of circumfixes' was announced.

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 Note that 'tries' (ii) can be seen as a type of 'finite-state automaton' (iv).
3 Observe that, in this case, *n*-grams guide the edit distance technique.
4 Of course, the overall error recall is not 100% for each randomly chosen text.
5 Other tools include those developed by Jan Daciuk (http://juggernaut.eti.pg.gda.pl/~jandac/), as well as AT&T's *FSM Library* (http://www.research.att.com/sw/tools/fsm/) or Canoo's *WMTrans* (http://www.canoo.com/wmtrans/home/index.html). (All last accessed: 26 April 2003.)

References

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

Beesley, K.R. & Karttunen, L. 2003. *Finite State Morphology.* Stanford: CSLI Publications.

Blair, C.R. 1960. A program for correcting spelling errors. *Information and Control* 3:60–67.

BNC. 2002. *British National Corpus*. http://www.natcorp.ox.ac.uk/ (Last accessed: 26 April 2003).

Bona. Bona magazine in isiZulu, April 2003.

Bosch, S.E. & Pretorius, L. 2002. Using finite-state computational morphology to enhance a machine-readable lexicon, in *AFRILEX 2002, Culture and Dictionaries, Programme & Abstracts*, edited by G.-M. de Schryver. Pretoria: $(SF)^2$ Press:20–22.

Bosch, S.E., Pretorius, L. & Van Huyssteen, L. 2003. Computational morphological analysis as an aid for term extraction, in *TAMA 2003 South Africa: CONFERENCE PROCEEDINGS*, edited by G.-M. de Schryver. Pretoria: $(SF)^2$ Press:65–71.

Cavnar, W.B. & Trenkle, J.M. 1994. N-gram-based text categorization, in *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas:161–175.

CLAWS. [sa]. *CLAWS part-of-speech tagger for English*. http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/ (Last accessed: 26 April 2003).

Daciuk, J. 2003. Home page. http://juggernaut.eti.pg.gda.pl/~jandac/ (Last accessed: 26 April 2003).

Departmental Northern Sotho Language Board. 1988[4]. *Northern Sotho Terminology and Orthography No. 4 / Noord-Sotho terminologie en spelreëls No.4 / Sesotho sa Leboa mareo le mongwalo No.4*. Pretoria: Government Printer.

De Schryver, G.-M. 2002. First steps in the finite-state morphological analysis of Northern Sotho, in *AFRILEX 2002, Culture and Dictionaries, Programme & Abstracts*, edited by G.-M. de Schryver. Pretoria: $(SF)^2$ Press:22–23.

De Schryver, G.-M. & Joffe, D. 2003. *Online Sesotho sa Leboa (Northern Sotho) – English Dictionary*. http://africanlanguages.com/sdp/ (Last accessed: 26 April 2003).

Hankamer, J. 1989. Morphological parsing and the lexicon, in *Lexical Representation and Process*, edited by W.D. Marslen-Wilson. Cambridge: The MIT Press:392–408.

Hirst, G. & St.-Onge, D. 1998. Lexical chains as representations of context for the detection and correction of malapropisms, in *WordNet: An Electronic Lexical Database*, edited by C. Fellbaum. Cambridge: The MIT Press:305–332.

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

Hurskainen, A. 1995. Information retrieval and two-directional word formation. *Nordic Journal of African*

Studies 4(2):81–92.

Hurskainen, A. 1996. Disambiguation of morphological analysis in Bantu languages, in *COLING-96: Proceedings of the Sixteenth International Conference on Computational Linguistics*. Copenhagen: 568–573.

Hurskainen, A. 1999. SALAMA: Swahili language manager. *Nordic Journal of African Studies* 8(2):139–157.

Hurskainen, A. 2002. New advances in corpus-based lexicography, in *AFRILEX 2002, Culture and Dictionaries, Programme & Abstracts*, edited by G.-M. de Schryver. Pretoria: $(SF)^2$ Press:18–19.

Hurskainen, A. & Halme, R. 2001. Mapping between disjoining and conjoining writing systems in Bantu languages: Implementation on Kwanyama. *Nordic Journal of African Studies* 10(3):399–414.

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.

Kaplan, R.M. & Newman, P.S. 1997. Lexical resource reconciliation in the Xerox linguistic environment, in *Proceedings of the ACL Workshop on Computational Environments for Grammar Development and Linguistic Engineering*. Madrid:54–61.

Karttunen, L. 1983. KIMMO: A general morphological processor. *Texas Linguistics Forum* 22:165–186.

Kernighan, M.D., Church, K.W. & Gale, W.A. 1990. A spelling correction program based on a noisy channel model, in *COLING-90: Proceedings of the Thirteenth International Conference on Computational Linguistics*. Helsinki:205–210.

Koskenniemi, K. 1983. *Two-level Morphology: A general computational model for word-form recognition and production* (PhD thesis). Helsinki: Department of General Linguistics (Publication No. 11), University of Helsinki.

Kukich, K. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys* 24(4):377–439.

Levenshtein, V.I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR* 163(4):845–848. [Translation from Russian into English (1966), in *Soviet Physics Doklady* 10(8):707–710.]

Lingsoft. 1999. *Orthografix 2 for Swahili.* http://www.lingsoft.fi/news/1999/o2-swahili.html (Last accessed: 26 April 2003).

Maniacky, J. 2003. *Umqageli (Automatic identification of Bantu languages)*. http://www.bantu-languages.com/en/tools/identification.php (Last accessed: 26 April 2003).

Maphosa, M. 2002. Word division and orthography as some of the factors posing challenges in the development of the Ndebele grammatical parser, in *AFRILEX 2002, Culture and Dictionaries, Programme & Abstracts*, edited by G.-M. de Schryver. Pretoria: $(SF)^2$ Press:23–24.

Odell, M.K. & Russell, R.C. 1918-1922. U.S. patent numbers 1,261,167 (1918) and 1,435,663 (1922). Washington: U.S. Patent Office.

Oflazer, K. 1996. Error-tolerant finite state recognition with applications to morphological analysis and spelling correction. *Computational Linguistics* 22(1):73–89.

Oflazer, K. & Güzey, C. 1994. Spelling correction in agglutinative languages, in *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing*. Stuttgart:194–195.

Paggio, P. 2000. Spelling and grammar correction for Danish in SCARRIE, in *6th Applied Natural Language Processing Conference Proceedings*. Seattle:255–261.

Parsons, R. 2003. *Soundex – the True Story*. http://west-penwith.org.uk/misc/soundex.htm (Last accessed: 26 April 2003).

Pretorius, L. & Bosch, S.E. 2001. Finite-state computational morphology – Treatment of the Zulu noun, in *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*, edited by K. Renaud, P. Kotzé & A. Barnard. Pretoria: Unisa Press:45–53.

Pretorius, L. & Bosch, S.E. 2002a. Finite-state computational morphology – Treatment of the Zulu noun. *South African Computer Journal* 28:30–38. [Also published as Pretorius & Bosch, 2001.]

Pretorius, L. & Bosch, S.E. 2002b. Regular expressions: enabling the development of computational aids for Zulu natural language processing, in *Proceedings of the 2002 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology* (ACM International Conference Proceeding Series). South African Institute for Computer Scientists and Information Technologists:254.

Pretorius, L. & Bosch, S.E. 2003. Enabling computer interaction in the indigenous languages of South Africa: the central role of computational morphology. *ACM interactions* 10(2) (Special Issue: HCI in the developing world):56–63.

Prinsloo, D.J. & De Schryver, G.-M. 2001. Corpus applications for the African languages, with special reference to research, teaching, learning and software. *Southern African Linguistics and Applied Language Studies* 19(1-2):111–131.

Prinsloo, D.J. & De Schryver, G.-M. 2003a. Towards second-generation spellcheckers for the South African languages, in *TAMA 2003 South Africa: CONFERENCE PROCEEDINGS*, edited by G.-M. de Schryver. Pretoria: $(SF)^2$ Press:135–141.

Prinsloo, D.J. & De Schryver, G.-M. 2003b. Non-word error detection in current South African spellcheckers. *Southern African Linguistics and Applied Language Studies* 21(4) (Special issue on 'Human Language Technology in South Africa: Resources and Applications'):307–326.

Prinsloo, D.J. & De Schryver, G.-M. 2004. Spellcheckers for the South African languages, Part 2: The utilisation of clusters of circumfixes. *South African Journal of African Languages* 24(1):83–94.

Ridings, D. & Mavhu, W. 2002. Problems and challenges encountered when developing a morphological parser for the Shona language, in *AFRILEX 2002, Culture and Dictionaries, Programme & Abstracts*, edited by G.-M. de Schryver. Pretoria: $(SF)^2$ Press:24–26.

Solak, A. & Oflazer, K. 1993. Design and implementation of a spelling checker for Turkish. *Literary and Linguistic Computing* 8(3):113–130.

Van der Veken, A. & De Schryver, G.-M. 2003. Les langues africaines sur la Toile: Étude des cas haoussa, somali, lingala et isixhosa. *Cahiers du Rifal* 23 (Thème: Le traitement informatique des langues africaines):33–45.

Van Huyssteen, G.B. 2002. *Speltoetserkompetisie*. http://www.puk.ac.za/fakulteite/lettere/fokusarea04/taaltegnologie/speltoetser/ (Last accessed: 26 April 2003).

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.

Vijaykumar, R. 1992. *Word recognition, spell checking and correction using ART-based neural networks* (M.Tech. dissertation). Bombay: Indian Institute of Technology.

Vosse, T.G. 1994. *The Word Connection: Grammar-based Spelling Error Correction in Dutch* (PhD thesis). Leiden: Rijksuniversiteit Leiden.

Wagner, R.A. & Fisher, M.J. 1974. The string-to-string correction problem. *Journal of the Association for Computing Machinery* 21(1):168–173.

Addendum A: "What is the African National Congress?" in Sesotho sa Leboa (ANC, [sa])

1,382 words in all
1,378 valid words
4 non-words (*kgathelelo → **kgatelelo**; *yha → **ya**; *boikarabela → **boikarabelo**; *šomiši → **šomiše**)
Maximum precision from the user's point of view (p.v.): 4 errors flagged & 16 types falsely flagged
⇒ 20.00%

# words in spellchecker	Underscore coding for false flaggings (valid words which are not recognised):	# types falsely flagged	Lexical recall from user's p.v.
10,000	token token token token	64	95.36 %
30,000	token token token	32	97.68 %
50,000	token token	25	98.19 %
100,000	token	16	98.84 %
100,000 + affix patterns	*token [cf. Prinsloo and De Schryver, 2004]*	*8*	*99.42 %*

NA AFRICAN NATIONAL CONGRESS KE ENG?

- ANC ke seboka sa setšhaba sa go lwela tokologo. E hlomilwe ka 1912 gore e kopanye batho ba Afrika le go eta pele ntwa ya go lwela phethogo ye e bonagalago ya sepolotiki, ya batho le ya ikonomi.
- Mo mengwageng ye masome-senyane ye e fetilego ANC e bile ketapele kgahlanong le ntwa ya semorafe le kgathelelo, gomme e rulaganya masole a kgahlanong le tše ka moka, e kalatša mafase a ka ntle kgahlanong le aparteiti le go ye lwa ka marumo
- ANC e ile ya tloga e fihlela phenyo ye kgolo go dikgetho tša 1994, fao e ilego ya fiwa borongwa bjo bo tiilego bja go rerišana ka Molaotheo wo moswa wa demokrase, Afrika Borwa. Molaotheo wo moswa wo o amogetšwego ka 1996
- Ka ngwaga wa 1999, ANC e ile ya kgethwa gape go mmušo wa setšhaba le wa diprofensi ka palo yha godimo ya bavouti
- Dipholisi tša ANC di dirwa ke maloko, gomme boetapele bo boikarabela go maloka
- Boleloko bja ANC bo buletšwe batho ka moka ba Afrika Borwa, ba mengwaga ya go feta ye 18, go sa *kgathalelwe* morafe le mmala, gomme batho bao ba swanetše go amogela ditheo, dipholisi le diprograma tša ANC.

NA MAIKALELO LE MAIKEMIŠETŠO A ANC KE AFE?

- Maikemišetšo a magolo a ANC ke go bopa setšhaba se se kopanego, se se se nago kgethologanyo ya bong le ya morafe, e bile e le sa demokrase.
- Se se ra gore go swanetše go ba le tokologo ya batho ba Afrika gagolo ba baso go tšwa go kgatelelo ya sepolotiki le ya ikonomi, gora go kaonafatša bophelo bja batho ba Afrika Borwa, gagolo bao ba tlogago ba dila.
- Go katanela go fihlela tše, go bitšwa Phetošo ya Sedemokrati ya Setšhaba

NA PHOLISI YA ANC E LATELA ENG?

Tokomane ya Freedom Charter yeo e ilego ya amogelwa ke Kopano *kgothe-kgothe* ya Batho ya 1955, ke yona tokomane ya Motheo ya ANC.
Freedom Charter e bolela gore:

- Batho ba tla buša
- Dihlopa ka moka tša merafe di tla ba le ditokelo tše di lekanago
- Batho ba tla abelanwa go mahumo a naga
- Lefase le tla abelana magareng ga bao ba le šomelago
- Batho ka moka ba tla lekana pele ga molao
- Batho ka moka ba tla ipshina ka ditokelo tše di swanago
- Go tla ba le mešomo le tšhireletšo
- Mejako ya thuto le setšo e tla bulelwa bohle
- Go tla ba le dintlo, tšhireletšo le boiketlo
- Go tla ba le khutšo le setswalle

Ka ngwaga wa 1994 ANC e ile ya amogela Lenaneo la Kago leswa le Tlhabollo (RDP), bjalo ka motheo wo o tlago *tlhahla* ANC go diphethogo tša Afrika Borwa. Mananeo a bohlokwa a RDP ke:

- go fihlela dinyakwa tša motheo tša batho
- tlhabollo ya bokgoni bja batho
- kago ya ikonomi
- go tliša demokrase go pušo le naga.

NA MASWAO A ANC A BOLELA ENG?

- Folaga ya ANC e dirilwe ka methala yeo e lekanago ya mmala wo moso, wo motala-morogo le wa gauta. Mmala wo moso o emela batho ba Afrika Borwa bao ba bego ba lwanela tokologo mo mengwageng ye mentšhi ye e fetilego. Mmala wo motala o emetše mabu ao a re fago bophelo gomme re bile re tšeetšwe wona ke mebušo ya bokoloniale. Gauta e emela diminerale le mahumo a mangwe a naga, ao e lego a batho ka moka, gomme wona a ile a šomišetšwa fela ke sehlophana se se nnyane.
- Maswao a na le lerumo le kotse tšeo di emelago dintwa tša go lwantšha bokoloniale, ntwa ya marumo ya *lephaga* la sešole la ANC, Umkhonto we Sizwe, le ntwa ye e tšwelago pele ya ANC ya go lwantšha kgatelelo ka sehlopa se se nnyane sa morafe. Leotwana le tlogela go Kopano *Kgothe-kgothe* ya Batho, yeo e amogetšego Freedom Charter, gomme lona le emela kopano ya ditšhaba ka moka tša Afrika Borwa go ntwa ya go katanela tokologo. Gape le bontšha setšo sa ANC sa go se kgethologanye go ya ka mmala. Seatla se se sweregо lerumo se emetše maatla a batho bao ba kopanego go ntwa ya go katanela tokologo le tekatekano.
- Mmolelwana wa ANC wa Amandla ngawethu goba Matla ke a rona o ra gore 'maatla a swanetše go fiwa batho', gomme se se bontšha maikemišetšo a magolo a Freedom Charter a gore batho ba swanetše go buša. Ke *tsebagatšo* ya boithapo bja ANC bja go aga le go tiiša demokrase le go tšea karolo ga batho go ntwa ya go katanela kaonafatšo ya maphelo a bona.

NA MAIKARABELO A LELOKO LA ANC KE AFE?

Ditlwaedi le ditheo tša leloko la ANC di akaretša:

- boikokobetšo le go ineela ka botlalo go lesolo la go katanela setšhaba sa demokrase seo se se nago kgethollo ya mmala le ya bong
- go ela hloko dinyakwa le dikgahlego tša batho, tšeo di lego go tokomane ya Batho Pele,
- boithapo bja go phethagatša dipholisi tša seboka le diphetho tša sona.

Maloko a ANC a letelwa go:

- go ba maloko a porantšhe ya ANC, go lefela tšhelete ya boleloko le go thuša go aga ANC,
- go tšea karolo go ditherišano tša go dira pholisi le diprograma tša ANC gammogo le phethagatšo ya tšona,
- go amogela le go emela diphetho tša dibopego tša mokgatlo,
- go aga kopano ya ANC le seboka sa demokrase le go lwantšha bomenemene, go šoma go ya setswalle le go hlola dihlophana-hlophana
- go lwantšha semorafe, kgethollo ya bong, go se kgotlelelane ga sedumedi le ga sepolotiki goba kgethollo efe goba efe,
- go dula o na le tsebo ya ditiragalo tša sepolotiki, go kaonafatša tsebo ya bona bjale ka karolo ya go ithuta ga go ya go ile,
- go dula o le kgauswi le batho le go tšea karolo go merero ya motse,
- go itshwara ka tsela ya maleba tšatši ka tšatši le go se šomiši maemo ka mabaka a go ikhola.

NA DIBOPEGO TŠA ANC KE DIFE?

- Porantšhe ke uniti ya mathomo ya ANC, fao maloko a tšeago karolo go ditiro tša ANC le ditherišano tša sepolotiki. Porantšhe ke ketapele ya motse, gomme e emela dikgahlego, dinyakwa le le go šomela tlhabollo ya motse. Porantšhe ye nngwe le ye nngwe e kgetha Komiti-Phethiši ya Porantšhe go Kopano Kgothe-Kgothe ya Ngwaga ka Ngwaga.
- Komiti Phethiši ya *Rejine* e kgethwa ka morago ga mengwaga e mebedi go Kopano ya *Rejine*, gomme yona e kgethwa ke dikemedi tša diporantšhe tše di bopago *rejine*.
- Komiti Phethiši ya Profensi e kgethwa ka morago ga mengwaga e meraro ke maloko a diporantšhe ka profensing.

- Komiti Phethiši ya Setšhaba ke sebopego sa ka godimo sa ANC gomme yona e sepediša merero ya mokgatlo ge go se no swarwa dikopano. E kgethwa ka morago ga mengwaga ya ye mehlano go Kopano ya Setšhaba. Dikemedi tša diporantšhe di bopa dipersente tša go ka ba 90 tša dikemedi tše di voutago go Kopano ya Bosetšhaba. Khansele ya Setšhaba ya Bohle e swarwa ge go se no swarwa Kopano ya Setšhaba ya go lekola diprograma tša mokgatlo.
- *Lligi* ya Basadi ya ANC e šoma ka go ikemela ka gare ga ANC. Maikemišetšo a yona ke go tšwetša pele le go šireletša ditokelo tša basadi le go netefatša gore basadi ba raloka karolo ya bona ka botlalo ka gare ga mokgatlo. Boleloko bja *Lligi* bo buletšwe basadi ka moka bao ba lego maloko a ANC.
- *Lligi* ya Baswa ya ANC le yona e šoma bjalo ka sebopego se se ikemetšego, gomme maikemišetšo a yona ke go etela baswa pele le go šomana le mathata a a ba lebanego. *Lligi* gape e netefatša gore baswa ba ba le seabe go ANC. Boleloko bja *Lligi* ya baswa bo buletšwe batho ba magareng ga mengwaga ye 14 le 35.

NA SEBOKA SA MAKALA A MARARO KE ENG?
ANC e šomišana le South African Communist Party (SACP) le Congress of South African Trade Unions (COSATU). Lekala le lengwe le lengwe le ikemetše, gomme le na le molaotheo wa lona le diprograma tša lona. Seboka se hlomilwe godimo ga boithapo go Diphethogo tša Setšhaba tša Demokrase, le hlokego ya go kopanya bontšhi bja batho ba Afrika Borwa go ya ka boithapo bjo.

NA ANC O KA E FIHLELA KAE?
Makala a ANC a hwetšwa go ditoropo ka moka, ditoropo tše nnyane le mebotwaneng ya Afrika Borwa. ANC e na le ofisi ye kgolo, diofisi tše senyane tša diprofensi le diofisi tše mmalwa tša *direjine*.

Addendum B: "What is the African National Congress?" in isiZulu (ANC, [sa])

1,008 words in all
1,005 valid words
3 non-words (*ayisishayagalolunye → **ayisishiyagalolunye**; *abampomfu → **abampofu**; *labho → **lapho**)
Maximum precision from the user's point of view (p.v.): 3 errors flagged & 101 types falsely flagged
⇒ 2.88%
Note that there are also 3 context-dependent errors (**igqugquzela → egqugquzela**; **kumkhankaso → umkhankaso**; **ngaphandle → angaphandle**)

# words in spellchecker	Underscore coding for false flaggings (valid words which are not recognised):	# types falsely flagged	Lexical recall from user's p.v.
10,000	token token token token token token token	339	66.27 %
30,000	token token token token token token	260	74.13 %
50,000	token token token token token	228	77.31 %
100,000	token token token token	182	81.89 %
200,000	token token token	141	85.97 %
400,000	token token	122	87.86 %
600,000	token	101	89.95 %
600,000 + affix patterns	*token* *[cf. Prinsloo and De Schryver, 2004]*	*32*	*96.82 %*

IYINI I-AFRICAN NATIONAL CONGRESS (UKHONGOLOSE, I-ANC)
- UKhongolose (u-ANC) uyinhlangano noma umbuthano olwela amalungelo aso sonke isizwe. Inhlangano kaKhongolose (u-ANC) yasungulwa ngonyaka ka 1912 ukuze ihlanganise onke ama-Afrika kanye nokuba ngumholi kumzabalazo wezinguquko ezijulile kwezombusazwe, izinguquko ezimpilweni zabantu kanye nakwezomnotho.
- Iminyaka engamashumi ayisishayagalolunye (90) uKhongolose ehola umzabalazo wokulwa nobandlululo ngokwebala, kanye nokulwa nengcindezelo, futhi igqugquzela abantu ngobuningi ukungenela kumkhankaso wokulwa nobandlululo, kanye nokugqugquzela amazwe ngaphandle ukusekela umzabalazo kanye nokuqala umzabalazo wezikhali wokulwa nombuso wobandlululo.

- I-ANC *yanqoba ngokuyingqayizivele* okhethweni lwenqubo yentando yeningi lango 1994, lapho yathola khona igunya elinqala *nelinzulu* lokungenela izingxoxo ngoMthethosisekelo omusha wentando yeningi eNingizimu Afrika. Umthethosisekelo omusha wamukelwa ngonyaka ka 1996.
- I-ANC yabuye yakhethwa futhi okhethweni lukazwelonke kanye *nolwezifundazwe lwangonyaka* ka 1999, lapho eyathola khona igunya lokubusa lezinga eliphezulu.
- Imigomo ye-ANC ibekwa ngamalunga kanye nabaholi, abachaza inqubo yomsebenzi kumalunga.
- Ubulunga be-ANC buvuleleke kuzo zonke izakhamizi zaseNingizimu Afrika ezineminyaka engu 18 nengaphezulu, izakhamizi zazo zonke izinhlanga, imibala, kanye nezinkolo inqobo nje uma zamukela izimiso (principles), imigomo kanye nezinhlelo zika-ANC.

NGABE YINI IZINHLOSO NEZINJONGO ZE-ANC (ZIKAKHONGOLOSE)?

- Injongo ebalulekile kaKhongolose ukwakha isizwe esibumbene *nesihlangene*, esinobunye, esingenalo ubandlululo ngokwebala, nangobulili nesiqhuba ngentando yeningi.
- Lokhu kusho inkululeko yama-Afrika kungcindezelo *yezombusazwe* kanye nakwezomnotho. Kanti futhi lokhu kusho ukuthuthukisa izinga lempilo yabo bonke abantu baseNingizimu Afrika, ikakhulukazi abantu abampomfu.
- Umzabalazo wokufezekisa lenjongo ubizwa ngokuthi yi-National Democratic Revolution.

YINI *OKUWUMKHOMBANDLELA* WEMIGOMO YE-ANC?

Umqulu wamalungelo enkululeko (i-Freedom Charter), eyamukelwa kwiNgqungquthela yabantu ngonyaka ka 1955, kuseyiwona Mkhombandlela kaKhongolose (ye-ANC).
Umqulu wamalungelo enkululeko (i-Freedom Charter) uthi:

- Abantu bayobusa
- Zonke izinhlanga zabantu ziyoba namalungelo alinganayo
- Abantu *bayokwabelana* ngomnotho wezwe
- Abantu abasebenza ngomhlaba *bayokwabelana* ngawo
- Bonke abantu bayolingana phambi komthetho
- Bonke abantu bayoba namalungelo obuntu
- Kuyoba nemisebenzi kanye nokuvikeleka
- Iminyango yemfundo kanye namasiko iyovuleleka kubo bonke abantu
- Kuyoba nezindlu, ukuvikeleka kanye nokunethezeka
- Kuyoba noxolo nobungani

Ngonyaka ka 1994, i-ANC yamukela *nokongamela* Uhlelo lweMvuselelo kanye neNtuthuko i-Reconstruction and Development Programme (RDP), njengesisekelo semigomo ye-ANC, eyiyona *nkombandlela* ekuguquleni iNingizimu Afrika. Izinhlelo ezibalulekile ze-RDP yilezi:

- ukufezekisa izidingo ezibalulekile zabantu
- ukuthuthukisa amakhono abantu
- ukwakha umnotho wezwe
- ukuqikelela ukuqhutshwa kwezinto kumbuso nakusizwe sonkana ngentando yeningi.

NGABE IZIMPAWU ZE-ANC ZIMELENI?

- Ifulegi le-ANC linemibala elinganayo eqondile emnyama, eluhlaza *okusatshani*, kanye nesagolide. Uphawu olumnyama lumele abantu baseNingizimu Afrika, *esekuyizizukulwana* belwela inkululeko. Umbala oluhlaza umele umhlaba, okuyiwona *obuwondla* abantu iminyaka eminingi labho abasuswa khona ngohulumeni *bobukoloni* kanye nobandlululo. Kanti igolide limele umcebo ombiwa phansi kanye nomnotho wemvelo waseNingizimu Afrika, umnotho ongowabantu bonke, kodwa usetshenziswa idlanzana labamhlophe ukuzizuzela bona bodwa.
- I-logo inomkhonto kanye nehawu, okumele izimpi *zokulwela* izwe nokulwa nenqubo yobukoloni kudala eyaphuca abantu imihlaba yabo, kanye nomzabalazo wezikhali we-ANC, uMkhonto weSizwe kanye nomzabalazo wabantu wesikhathi eside belwa nobandlululo ngokwebala kanye nengcindezelo. Lomzabalazo sekuyisikhathi udonsa, ukusukela eNgqungqutheleni Yabantu (Peoples Congress), eyamukela Umqulu wamalungelo (i-Freedom Charter), kanti futhi *owawuluphawu* lokuhlangana kwemiphakathi eyahlukene yabantu baseNingizimu Afrika ilwela into eyodwa, okuyinkululeko.

Kuwuphawu losiko lwe-ANC lobumbano lwezinhlanga ezehlukene, eNingizimu Afrika. Inqindi ebambe umkhonto, imele amandla abantu *ababumbene* emzabalazweni wokuthuthukisa izimpilo zabo.

- Isiqubulo se-ANC esithi, Amandla ngawethu noma Matla ke ya rona, sisho ukuthi amandla awabe kubantu, kanti lokhu kususelwa kusidingo esinqala se-Freedom Charter esithi, Abantu Bayobusa. Yisitatimende se-ANC sokuzimisela ukwakha kanye nokuzinzisa inqubo yentando yeningi, kanye nokubamba kwabantu iqhaza emzabalazweni wokuthuthukisa izimpilo zabo.

NGABE YINI IMISEBENZI YELUNGA LE-ANC?

Izinto nezimiso ezibalulekile zelunga le-ANC zibandakanya okulandelayo:

- Ukuzithoba kanye nokuzinikela ngokuphelele emzabalazweni *wokulwela* izwe elingenabandlululo ngokwebala, ubulili kanye *neliqhuba* ngentando yeningi,
- Ukuzinikela ekusebenzeleni izifiso kanye nezidingo zabantu, izinto ezitholakala kumgomo wokubeka abantu phambili we-Batho Pele (Abantu Kuqala),
- Ukuzimisela ukuqhuba izinto ngokulandela imigomo yenhlangano kanye nezinqumo ezithathwa yiningi ngobunye.

Amalunga ka-ANC alindeleke ukuthi:

- Abe ngamalunga amagatsha e-ANC, akhokhele imali yobulunga kanye nokusiza ekwakheni i-ANC,
- Ukubamba iqhaza elibonakalayo ezingxoxweni, ekubunjweni kwemigomo kanye nasekusetshenzisweni kwemigomo kanye nezinhlelo ze-ANC,
- Ukwamukela kanye nokuvikela izinqumo zezakhiwo ezifanele zenhlangano,
- Ukwakha ubunye nobumbano kanye nenqubo yokulandela intando yeningi enhlanganweni, ukulwa nenkohlakalo kanye nokuvuna izihlobo, kanye nokulwa *nokubangwa* koqhekeko olubangwa izigejana enhlanganweni,
- Ukulwa nenqubo yobandlululo ngebala, ubandlululo ngokobuhlanga, ubandlululo ngokobulili, inkolo kanye nokungabekezelelani kwezombusazwe, kanye nokulwa nanoma iyiphi inhlobo yobandlululo noqhekeko,
- Ahlale enolwazi ngezombusazwe kanye nezinye izinto *eziqhubekayo*, kanye nokuthuthukisa amakhono abo njengengxenye yokuqhubekela phambili nokufunda kuyo yonke impilo yomuntu,
- Ukuhlala bexhumene nabantu kanye nokudlala indima ebonakalayo ezindabeni zemiphakathi, kanye
- Nokuziphatha ngendlela eyisibonelo esihle nsuku zonke, kanye nokungasebenzisi izikhundla ukuzizuzela ngokwabo, ukuzicebisa kanye *nokuzibonelela* ziqu zabo.

NGABE YINI IZAKHIWO ZE-ANC?

- Igatsha yisona sakhiwo esiyisisekelo nempande ye-ANC, lapho amalunga okumele abambe khona iqhaza emisebenzini ye-ANC kanye nezingxoxo ngezombusazwe. Igatsha yilona eliyiso nezindlebe zomphakathi, elimele izidingo zomphakathi, kanye nokuba yiwona mnyombo maqondana nezidingo nezifiso zomphakathi, kanye nokuqqugquzela imiphakathi ukuze isebenzele intuthuko ezindaweni. Igatsha negatsha likhetha ikomiti eliphezulu legatsha (Branch Executive Committe) eMhlanganweni *KaWonke*-wonke wonyaka (Annual General Meeting).
- Ikomidi eliphezulu lesiFunda (i-Regional Executive Committee-REC) likhethwa kwiNgqungquthela yesiFunda njalo eminyakeni emibili, ngabameli bamagatsha kusiFunda.
- Ikomiti eliphezulu *kusiFundazwe* (Provincial Executive Committee- PEC), likhethwa njalo eNgqungquthelenі yesiFundazwe eminyakeni emithathu, likhethwa ngabameli bamagatsha esiFundazweni.
- Ikomidi eliphezulu likaZwelonke (National Executive Committee-NEC) yisona sakhiwo esiphezulu esithatha izinqumo ku-ANC phakathi kweziNgqungquthela kanti futhi linomsebenzi wokuhola inhlangano. *Lelikomidi* le-NEC, likhethwa njalo eminyakeni emihlanu eNgqungquthelenі kaZwelonke. I-NEC ikhetha iKomidi elisebenzayo likaZwelonke, i-National Working Committee (NWC) phakathi kwamalunga eKomidi likaZwelonke, ukubhekana nemisebenzi yenhlangano yansuku zonke.
- INgqungquthela kaZwelonke (National Conference), ebanjwa njalo eminyakeni emihlanu, yisona sakhiwo esiphezulu esithatha izinqumo ku-ANC. Abameli bamagatsha bangamaphesenti angu 90 ezithunywa ezivotayo kwiNgqungquthela kaZwelonke. Umkhandlu kaZwelonke Kawonke-wonke (National General Council-NGC), ibanjwa njalo phakathi kweziNgqungquthela zikaZwelonke ukubuyekeza izinhlelo zenhlangano.

- I-ANC Women's League, isebenza njengenhlangano ezimele kunhlangano yonkana ye-ANC. Injongo yayo enkulu wukuqhubela phambili nokulwela amalungelo abesimame kuzo zonke izinhlobo zengcindezelo kanye nokuqinisekisa ukuthi abesimame badlala indima ebalulekile enhlanganweni. Ubulunga be-Women's League buvuleleke kubo bonke abesimame abangamalunga e-ANC.
- I-ANC Youth League, nayo isebenza njengenhlangano ezimele, enenhloso yokuhlanganisa kanye nokuhola abasha ekubhekaneni nezinkinga zabo, kanye nokuqinisekisa ukuthi intsha idlala indima egcwele *nejulile* kumisebenzi ye-ANC. Ubulunga be-Youth League, buvuleleke kubo bonke abantu abaneminyaka ephakathi kuka 14 no 35.

NGABE BUYINI UBUDLELWANO *OBUNXA*-NTATHU (TRIPARTATE ALLIANCE)?
I-ANC inombimbi noma umfelandawonye nenhlangano yamakhomanisi, i-South African Communist Party (SACP), kanye noKhongolose wabasebenzi i-Congress of South African Trade Unions (COSATU). Ingxenye nengxenye yombimbi iyinhlangano ezimele, *enomthethosisekelo* wayo, amalunga ayo kanye nezinhlelo zayo. Umbimbi *lwakhelwe* phansi kwezinjongo ezifanayo, zeNational Democratic Revolution, kanye nokuhlanganisa imikhakha ehlukene yabantu baseNingizimu Afrika ngaphansi kwalezinjongo.

NGABE ITHOLAKALA KUPHI I-ANC
Amagatsha e-ANC atholakala *kumadolobhakazi* amakhulu, emadolobheni, kanye nasezindaweni zasemakhaya eNingizimu Afrika. I-ANC inekomkhulu likazwelonke, amahhovisi *eziFundazwe* ayisishiyagalolunye kanye namahhovisi eziFunda.

Addendum C: "What is the African National Congress?" in Afrikaans (ANC, [sa])

1,287 words in all
1,285 valid words
2 non-words (*Konferense → **Konferensie**; *Afika → **Afrika**)
Maximum precision from the user's point of view (p.v.): 2 errors flagged & 12 types falsely flagged
⇒ 14.29%
Note that there is also 1 context-dependent error (**sale → sal**)

# words in spellchecker	Underscore coding for false flaggings (valid words which are not recognised):	# types falsely flagged	Lexical recall from user's p.v.
10,000	token token token token token	94	92.68 %
30,000	token token token token	45	96.50 %
50,000	token token token	29	97.74 %
100,000	token token	21	98.37 %
200,000	token	12	99.07 %
200,000 + affix patterns	*(all valid words are recognised) [cf. Prinsloo and De Schryver, 2004]*	*0*	*100.00 %*

WAT IS DIE AFRICAN NATIONAL CONGRESS?
- Die ANC is 'n nasionale bevrydingsbeweging. Dit is in 1912 gestig om Afrikane te verenig en die stryd om fundamentele politieke, sosiale en ekonomiese verandering te lei.
- Vir nege dekades het die ANC die stryd teen rassisme en verdrukking gelei, deur middel van georganiseerde massa-teenstand, die mobilisering van die internasionale gemeenskap en die opneming van die gewapende stryd teen apartheid.
- Die ANC het in die 1994-verkiesing 'n groot demokratiese deurbraak gemaak, toe dit 'n ferm mandaat gekry het om 'n nuwe demokratiese Grondwet vir Suid-Afrika te beding. Die nuwe Grondwet is in 1996 aanvaar.
- Die ANC is in 1999 weer tot die nasionale en provinsiale regering verkies met 'n vergrote mandaat.
- Die beleid van die ANC word bepaal deur die lede, en die leierskap doen rekenskap aan die lede.
- Lidmaatskap van die ANC is oop vir alle Suid-Afrikaners bo die ouderdom van 18 jaar, ongeag ras, kleur of geloof, wat die beginsels, beleid en programme van die ANC onderskryf.

WAT IS DIE OOGMERKE EN DOELSTELLINGS VAN DIE ANC?

- Die ANC se hoofdoelstelling is om 'n verenigde, nie-rassige, nie-seksistiese en demokratiese samelewing te skep.
- Dit behels die bevryding van Afrikane in die besonder en swartmense in die algemeen van politieke en ekonomiese onderdrukking. Dit behels die opheffing van die lewensgehalte van alle Suid-Afrikaners, veral die armes.
- Die stryd om hierdie oogmerk te bereik, staan bekend as die Nasionale Demokratiese Revolusie.

WAARDEUR WORD ANC-BELEID GELEI?

Die Vryheidsmanifes, wat in 1955 deur die Congress of the People aanvaar is, bly steeds die basiese beleidsdokument van die ANC.
Die Vryheidsmanifes verklaar dat:

- Die mense sal regeer
- Alle nasionale groepe gelyke regte sale hê
- Die mense sal deel in die land se rykdom
- Die grond gedeel sal word tussen die mense wat dit bewerk
- Almal sal gelyk voor die reg wees
- Almal sal gelyke menseregte geniet
- Daar werk en sekuriteit sal wees
- Die deure van onderwys en kultuur sal oop wees
- Daar huise, sekuriteit en gerief sal wees
- Daar vrede en vriendskap sal heers

In 1994 het die ANC die Heropbou - en Ontwikkelingsprogram (HOP) aanvaar as die basiese beleidsraamwerk wat die ANC sou lei in die transformasie van Suid-Afrika. Die sleutelprogramme van die HOP is:

- om in basiese behoeftes te voorsien
- om ons menslike hulpbronne te ontwikkel
- om die ekonomie op te bou
- om die staat en samelewing te demokratiseer

WAT BETEKEN DIE SIMBOLE VAN DIE ANC?

- Die ANC-vlag bestaan uit gelyke horisontale bane van swart, groen en goud. Die swart simboliseer die mense van Suid-Afrika wat geslagte lank geveg het vir vryheid. Die groen simboliseer die grond, wat ons mense eeue lank onderhou het en waarvan hulle verwyder is deur koloniale en apartheidsregerings. Die goud verteenwoordig die minerale en ander natuurlike rykdom van Suid-Afrika, wat aan al die mense behoort, maar gebruik is om net 'n klein rasse-minderheid te bevoordeel.
- Die logo bevat 'n spies en 'n skild wat die vroeë versetoorloë teen koloniale oorheersing, die gewapende stryd van die ANC se voormalige gewapende vleuel, Umkhonto we Sizwe, en die ANC se voortdurende stryd teen rasse-voorreg en verdrukking verteenwoordig. Die wiel dateer terug tot die veldtog vir die Congress of the People, wat die Vryheidsmanifes aanvaar het, en herdenk die gesamentlike vryheidstryd van al Suid-Afrika se gemeenskappe. Dit is 'n simbool van die sterk nie-rassige tradisies van die ANC. Die vuis wat die spies vashou, verteenwoordig die krag van 'n volk wat verenig is in die stryd om vryheid en gelykheid.
- Die ANC se strydkreet, "Amandla ngawethu" of "Matla ke arona" beteken "mag aan die mense", wat die sentrale eis van die ANC weerspieël dat die mense moet regeer. Dit is 'n stelling van die ANC se verbintenis om populêre demokrasie op te bou en te versterk, en die aktiewe betrokkenheid van die mense in die stryd om hul lewens te verbeter.

WAT IS DIE VERANTWOORDELIKHEDE VAN 'N ANC-LID?

Die waardes en beginsels van 'n ANC-lid sluit die volgende in:

- nederigheid en 'n selflose verbintenis tot die stryd om 'n nie-rassige, nie-seksistiese en demokratiese samelewing;
- besorgdheid oor die wil en belange van die mense, wat vasgevang word in die beginsels van Batho Pele - mense eerste;
- 'n verbintenis om die beginsels van die beweging en die besluite van die kollektief te implementeer.

Daar word van ANC-lede verwag om:

- aan 'n ANC-tak te behoort, ledegeld te betaal en te help om die ANC uit te bou;
- aktief deel te neem aan die bespreking, formulering en implementering van ANC-beleid en -programme;
- die besluite van die betrokke strukture van die beweging te aanvaar en te verdedig;
- die eenheid van die ANC en die demokratiese beweging te versterk en om korrupsie, nepotisme en faksievorming te beveg;
- te stry teen rassisme, stam-chauvinisme, seksisme, godsdienstige en politieke onverdraagsaamheid of enige vorm van diskriminasie;
- deurlopend op hoogte te bly van politieke en ander ontwikkelings, en hul eie vermoëns op te bou as deel van 'n proses van lewenslange onderrig;
- in voeling te bly met die mense en 'n aktiewe rol te speel in gemeenskapsake;
- op 'n voorbeeldige wyse op te tree in die alledaagse lewe, en nie verantwoordelike posisies te misbruik vir selfverryking of persoonlike voordeel nie.

WAT IS DIE STRUKTURE VAN DIE ANC?

- Die tak is die basiese eenheid van die ANC, waar lede deelneem aan ANC-aktiwiteite en politieke besprekings. Die tak staan aan die voorpunt van die gemeenskap, dien hul belange, druk hul strewes uit en mobiliseer hulle om saam te werk vir plaaslike ontwikkeling. Elke tak verkies 'n Tak- Uitvoerende Komitee by 'n Algemene Jaarvergadering.
- Die Streek- Uitvoerende Komitee (SUK) word elke tweede jaar deur die verteenwoordigers van die takke in die streek verkies op 'n Streekskonferensie.
- Die Provinsiale Uitvoerende Komitee (PUK) word elke drie jaar deur die verteenwoordigers van die takke in die provinsie verkies op 'n Provinsiale Konferensie.
- Die Nasionale Uitvoerende Komitee (NUK) is die hoogste orgaan van die ANC tussen konferensies, en is belas daarmee om die organisasie te lei. Dit word elke vyf jaar op die Nasionale Konferensie verkies. Die NUK verkies 'n Nasionale Werkskomitee (NWK) uit hul geledere om die werk van die organisasie op 'n dag-tot-dag basis te koördineer.
- Die Nasionale Konferensie, wat elke vyf jaar plaasvind, is die hoogste besluitnemingsliggaam van die ANC. Takverteenwoordigers maak ten minste 90 persent van stemgeregtigde afgevaardigdes op die Nasionale Konferense uit. 'n Nasionale Algemene Raad (NAG) word tussen Nasionale Konferensies byeengeroep om die programme van die beweging te evalueer.
- Die ANC Vroueliga tree as 'n outonome liggaam binne die algehele struktuur van die ANC op. Die oogmerk van die Vroueliga is om vroueregte te beskerm en te versterk teen alle vorms van verdrukking, en om toe te sien dat vroue 'n volle rol in die lewe van die organisasie vervul. Die Vroueliga is oop vir alle vroue wat ANC-lede is.
- Die ANC Jeugliga tree ook as outonome liggaam op, met die doel om jongmense te verenig en te lei ten einde probleme wat die jeug in die gesig staar, aan te spreek, en om toe te sien dat die jeug 'n volle en ryke bydrae tot die werk van die ANC lewer. Lidmaatskap van die Jeugliga is oop vir almal tussen die ouderdomme van 14 en 35 jaar.

WAT IS DIE DRIELEDIGE ALLIANSIE?

Die ANC is in 'n alliansie met die Suid-Afrikaanse Kommunisteparty (SAKP) en die Congress of South African Trade Unions (Cosatu). Elke alliansie-vennoot is 'n onafhanklike organisasie met sy eie grondwet, lidmaatskap en programme. Die Alliansie is gebou op 'n gesamentlike verbintenis tot die oogmerke van die Nasionale Demokratiese Revolusie, en die noodsaaklikheid om die grootste moontlike deursnit van Suid-Afrikaners te verenig rondom hierdie oogmerke.

WAAR OM DIE ANC TE VIND

ANC-takke is in elke stad, dorp en stat in Suid-Afika te vinde. Die ANC het 'n nasionale hoofkantoor, nege provinsiale kantore en verskeie streekskantore.

Addendum D: IsiZulu words not recognised by the WordPerfect 9 spellchecker

000, abango, abangayenza-, abangamawele, imakethe, 000, 000, eseNingizimu, 000, eyikhombisa, ezakhaya, abakuxoxayo, ikhishiwe, ezibakhathazayo, 000, ezibaphethe, esezingoDJ, ezikhombise, ethengwe, esteji, e-stage, i-mask, ..., i-mask, esingangonyaka, , esingakaze, abanjengoMirriam, i-mask, abangu-15, ezitudyo, iCD, iCD, iCD, iCD, i-American, 1999, i-, i-, House, HIV/AIDS, GORDON, 200, 1, Franklin, ezishubile, idonsa, 11, ihambisana, ihlanganise, eziphusile, ezinkingeni, iHouse, ijazz, ezingu-9, ezingomeni, abahlangabezane, I-CD, ezidumile, 60, bawuthanda, akuhambanga, amakhophi, amakhophi, bazizwele, ama-drums, Alexandra, albhamu, bebafaka, bazibiza, bekhulumabakhulume, bangamawele, ekwi-mask, bangakwenza, bamxoshe, eLimpopo, eliselokishini, esasifuna, albhamu, angakusho, besingalindele, beRevolution, bematasa, be-kwaito, bukhoma, bukhoma, Caiphus, bazoyamukela, i-mask, emabhange, aneminyaka, Cleo, e-albhamu, e-Alexandra, Eastern, ebakhona, Bekungelula, Cape, enokuncintisana, bamkhumule, engaqhamuka, engidla, abayingxenye, engiwenzayo, engize-nzayobagcine, engu-21, eneminyakazo, abayamukele, abehlukene, abaphezulu, ephathini, eSA, esabenza, Besingazi, esafake, afinyelela, engu-25, Afro-Dance, Emakhonsathini, emakhonsathini, babeshoda, babekungabaza, azibandakanye, ayifakayo, ayenenqwaba, abazowuthokozela, axoxisane, esasisizwa, afinyelela, asethengwe, ase-Afrika, angu-100, adume, adubule, abehlukene, emizamweni, lase-, i-mask, ngoDJ, ngohlobo, uGeorge, le-, le-, uGeorge, ngokulingisa, uGeorge, ngokunganqeni, ngokusinika, uGeorge, uGeorge, uGeorge, ngokuythaka, ngomculi, lawoma-, uJoseph, linabantu, likaMzekezeke, likahulumeni, lifune, ngo-, lezimanga, ngoDJ, Letta, u-Alice, uJoseph, lesisihlabani, lesisibalo, Leope, uhlela, U-George, uGeorge, ngo-4, kwakheka, nguS'bu, nguS'bu, sizophendula, ningaceli, sizethemba, siyihlanganise, uDJ, kwalezizinhlelo, kwelishumi, kwaito, siyanda, njengokufaka, kuZola, njengomkhulumeli, no-Amu, no-Aretha, kwalezizinhlelo, kwe-show, zoba, Thabo, kwiYFM, kwiYFM, kwiThe, kwi-, Ngo-October, nguRudeboy, kwe-show, kwehliswe, kweRevolution, sokucula, sokubamba, nguDOUG, slye, nguMzekezeke, ucwaningo, Tabane, yakhishwa, wonyaka, wuhlelo, nezinhlelo, nezicukuthwana, wukwenza, nesitayela, yayingakawulungeli, nese-Afrobeat, weYFM, ne-mask, ne-mask, yalandelwa, negciwane, yama-disc, yase-Alberton, , yageqa, ngayo, ngempelasonto, ngembotshana, ngelikaMzekezeke, ngelikaMzekezeke, wagqoka, walolohlobo, wase-Afrika, WHITFIELD, waseNigeria, wezihlabani, ngayo, ngayo, wayebafundisa, wehlukile, we-kwaito, we-kwaito, nasezinhlelweni, ngegalelo, zingathathi, zeYFM, Mdlongwa, Mbulu, Mbeki, Makeba, zikamabonakude, yaselokishini, lukaPhat, Mnisi, ziwazise, lowaphesheya, loMzekezeke, ziyamhlonipha, lokufeza, ziyobaqopha, Kuti, lukaRudeboy, yokufuna, naseSwazini, nasemabhilidini, yemenenja, yezingoma, yi-Afro-Dance, yiRevolution, namakhonsathi, ze-kwaito, naloMzekezeke, Zazifundisa, zalelizwe, nabalandeli, zawujwayezwa, Mothiba, Mothiba, Mothiba, zokuqopha, nalwazi, izingxenye, izithombenguPETER, izithatha, UMzekezeke, umnandi, ngimatasa-, ngingasithola, izifihle, ukuzihambela, uMzekezeke, ukuwuthanda, izingqinamba, Ukuvumile, Ukusunguleka, izimpilo, ukuqhakambisa, ongusomlomo, ukuziphilela, uMzekezeke, noBheki, ongumkhulumeli, onguDJ, ngesitsotsi, kaMongameli, Kamongameli, ngezinkinga, uMzekezeke, ngezinto, izoxoxisana, uMzekezeke, UMzekezeke, uMzekezeke, uMzekezeke, Journey, Joe, iyaluma, ngezinto, ingcwenga, uSbu, uSbu, ngeRevolution, uRucula, inguMzekezeke, ngesidla, izikhulumela, ingeyaseTembisa, U-Joseph, ineminyaka, ungumculi, ngesitsotsi, uMzekezeke, ngesitsotsi, imishanguzo, Wa, u-Oscar, ngiyikhulume, ngisithole, ngixoxisana, ngixoxisana, isitsotsi, ngiyabuya, isithengise, ukukhiqiza, uSbu, ngiyalithanda, i-overall, isifundazwe, isenesasasa, I-Revolution, i-rap, nglye, Ngiyikhumula, oMzekezeke, ukudlalelwa, sase-Afrika, selizakhele, Saxhumana, sasinephupho, Sasesiside, noMafikizolo, noMandoza, kulesikhathi, nomhobholo, sengiyimina, kunenikuzwa, kunendoda, noMzekezeke, nongumlandeli, nophromotha, kuluhlelo, onganqeni, nomaskanda, S'guqa, I-mask, noJoseph, noJoseph, noJoseph, nokukhomba, silula, sikutshela, seMalombo, Sihlose, Semenya, sezingeni, sesiwahambele, nokuyizinto, nokuzizwa, nokuzothi, sePhasika-, kulemboni, sikulemboni, owasekhaya, ozokhuluma, okuxakayo, okuyigama, okuyindawo, Okwaziwayo, okwehlukile, Kulomculo, owavumbulula, koMapaputsi, omgumzali, overall, overall, kaPhillip, Oscar, ophromotha, ongihlanganisa, ozifihle, oDJ, noSelimathunzi, no-Unathi, kulaphoke, sabambela, Rona, kuFela, noYizo, oGeorge, pop, odumile, Paul, kubaculi, ku-100, ku-, ku-, Paul, noDJ, 'qhwasha'

Addendum E: IsiZulu words flagged as valid by a simulation of an isiZulu 'quinquegram speller'

abahlangabezane, 18, 16, 39, 141, 102, 41, 21, 17, 6, 5, 3
abangamawele, 50, 51, 20, 65, 2, 2, 2, 4
abango, 50, 16
abaphezulu, 18, 10, 17, 12, 13, 22
abehlukene, 1, 3, 22, 12, 12, 11
adubule, 1, 3, 1
afinyelela, 1, 12, 12, 13, 14, 86
angakusho, 11, 19, 8, 19, 28
ase-Afrika, 3, 1, 1, 3, 3, 4
asethengwe, 2, 6, 7, 43, 2, 4
azibandakanye, 1, 6, 5, 7, 5, 7, 3, 16, 12
bamkhumule, 8, 12, 6, 9, 27, 11
bamxoshe, 1, 4, 7, 2
bangakwenza, 32, 39, 26, 18, 51, 172, 77
bangamawele, 32, 20, 65, 2, 2, 2, 4
bazibiza, 6, 1, 3, 14
bazizwele, 1, 8, 8, 7, 2
besingalindele, 4, 3, 23, 25, 28, 7, 26, 33, 57, 58
eliselokishini, 7, 7, 3, 14, 1, 4, 4, 2, 8, 21
emabhange, 2, 12, 3, 4, 6
eneminyakazo, 2, 13, 32, 38, 37, 19, 12, 2
engidla, 2, 5, 4
engiwenzayo, 2, 2, 2, 5, 23, 20, 20
ephathini, 3, 70, 17, 34, 61
esabenza, 4, 2, 5, 54
esasifuna, 12, 2, 3, 21, 23
esasisizwa, 12, 1, 1, 6, 3, 37
esingakaze, 27, 23, 47, 59, 6, 11
ethengwe, 3, 43, 2, 4
eyikhombisa, 1, 12, 30, 76, 52, 52, 51
ezakhaya, 3, 1, 10, 31
ezikhombise, 14, 16, 30, 76, 52, 52, 9
ezingomeni, 41, 10, 3, 2, 3, 6
idonsa, 1, 9
ihambisana, 3, 40, 19, 12, 13, 30
ihlanganise, 3, 141, 102, 98, 33, 30, 10
ikhishiwe, 1, 17, 3, 3, 12
ineminyaka, 1, 13, 32, 38, 37, 51
isithengise, 27, 30, 19, 43, 11, 24, 17
izikhulumela, 17, 28, 99, 197, 151, 19, 8, 60
izimpilo, 16, 4, 2, 17
izingxenye, 22, 2, 1, 5, 6, 10
izithatha, 23, 19, 65, 187, 102
kulaphoke, 4, 5, 3, 1, 1
kulesikhathi, 8, 6, 1, 47, 39, 39, 78, 44
kunendoda, 4, 2, 1, 29, 28
kwehliswe, 10, 4, 12, 1, 3
kwelishumi, 4, 9, 4, 3, 23, 57
lesisibalo, 21, 27, 3, 1, 3, 2
lezimanga, 4, 7, 11, 24, 28
lifune, 1, 4
likahulumeni, 1, 1, 1, 1, 151, 19, 11, 21
nabalandeli, 5, 13, 6, 27, 84, 89, 5

naseSwazini, 3, 3, 3, 4, 4, 8, 18
nasezinhlelweni, 17, 23, 20, 21, 61, 2, 2, 12, 6, 23, 84
nezinhlelo, 79, 21, 61, 2, 2, 7
ngayo, 3, 11
ngempelasonto, 10, 8, 22, 5, 1, 1, 1, 6, 12
ngesidla, 74, 3, 1, 1
ngezinkinga, 102, 69, 54, 3, 5, 9, 10
ngezinto, 102, 69, 36, 19
ngingasithola, 93, 94, 26, 5, 1, 10, 9, 57, 65
ngiyabuya, 164, 12, 1, 1, 17
ngiyalithanda, 164, 6, 4, 1, 10, 8, 57, 184, 116
ngiyikhulume, 42, 11, 5, 16, 99, 197, 151, 16
ngokulingisa, 211, 22, 11, 18, 11, 1, 9, 20
ngokusinika, 211, 23, 11, 17, 1, 8, 26
ngomculi, 3, 1, 9, 2
njengokufaka, 168, 69, 38, 48, 15, 7, 5, 3
njengomkhulumeli, 168, 69, 21, 2, 7, 14, 25, 197, 151, 19, 8, 3
noBheki, 3, 1, 1
nokukhomba, 14, 52, 46, 12, 76, 18
nokuyizinto, 6, 7, 8, 7, 26, 5, 19
nokuzizwa, 5, 9, 1, 1, 14
odumile, 1, 1, 26
okuyigama, 13, 2, 2, 1, 7
okuyindawo, 13, 10, 5, 2, 14, 33
okwehlukile, 1, 16, 9, 22, 1, 8, 23
owasekhaya, 2, 10, 28, 20, 18, 31
sase-Afrika, 3, 14, 1, 1, 3, 3, 4
selizakhele, 2, 1, 1, 1, 7, 35, 11
sengiyimina, 107, 20, 22, 1, 4, 9, 7
sezingeni, 11, 45, 29, 57, 39
sikutshela, 1, 4, 20, 22, 148, 81
sizophendula, 2, 5, 2, 1, 108, 111, 80, 45
sokubamba, 5, 7, 6, 7, 39
uhlela, 3, 21
ukudlalelwa, 6, 7, 8, 3, 63, 13, 35
ukuziphilela, 41, 4, 4, 8, 34, 8, 2, 3
wagqoka, 1, 3, 3
wase-Afrika, 4, 14, 1, 1, 3, 3, 4
wehlukile, 6, 22, 1, 8, 23
wonyaka, 1, 5, 51
yakhishwa, 4, 5, 17, 14, 11
yalandelwa, 1, 27, 84, 89, 5, 3
yezingoma, 25, 45, 10, 3, 15
yokufuna, 1, 12, 29, 6
zingathathi, 22, 16, 74, 12, 45, 187, 44
zokuqopha, 3, 5, 3, 3, 3