Setswana Tokenisation and Computational Verb Morphology: Facing the Challenge of a Disjunctive Orthography

Rigardt Pretorius
School of Languages
North-West University
Potchefstroom, South Africa
Rigardt.Pretorius@nwu.ac.za

Ansu Berg
School of Languages
North-West University
Potchefstroom, South Africa
Ansu.Berg@nwu.ac.za

Laurette Pretorius
School of Computing
University of South Africa
and
Meraka Institute, CSIR
Pretoria, South Africa
pretol@unisa.ac.za

Biffie Viljoen
School of Computing
University of South Africa
Pretoria, South Africa
viljoe@unisa.ac.za

Abstract

Setswana, a Bantu language in the Sotho group, is one of the eleven official languages of South Africa. The language is characterised by a disjunctive orthography, mainly affecting the important word category of verbs. In particular, verbal prefixal morphemes are usually written disjunctively, while suffixal morphemes follow a conjunctive writing style. Therefore, Setswana tokenisation cannot be based solely on whitespace, as is the case in many alphabetic, segmented languages, including the conjunctively written Nguni group of South African Bantu languages. This paper shows how a combination of two tokeniser transducers and a finite-state (rule-based) morphological analyser may be combined to effectively solve the Setswana tokenisation problem. The approach has the important advantage of bringing the processing of Setswana beyond the morphological analysis level in line with what is appropriate for the Nguni languages. This means that the challenge of the disjunctive orthography is met at the tokenisation/morphological analysis level and does not in principle propagate to subsequent levels of analysis such as POS tagging and shallow parsing, etc. Indeed, the approach ensures that an aspect such as orthography does not obfuscate sound linguistics and, ultimately, proper semantic analysis, which remains the ultimate aim of linguistic analysis and therefore also computational linguistic analysis.

1 Introduction

Words, syntactic groups, clauses, sentences, paragraphs, etc. usually form the basis of the analysis and processing of natural language text. However, texts in electronic form are just sequences of characters, including letters of the alphabet, numbers, punctuation, special symbols, whitespace, etc. The identification of word and sentence boundaries is therefore essential for any further processing of an electronic text. *Tokenisation* or word segmentation may be defined as the process of breaking up the sequence of characters in a text at the word boundaries (see, for example, Palmer, 2000). Tokenisation may therefore be regarded as a core technology in natural language processing.

Since disjunctive orthography is our focus, we distinguish between an orthographic word, that is a unit of text bounded by whitespace, but not containing whitespace, and a linguistic word, that is a sequence of orthographic words that together functions as a member of a word category such as, for example, nouns, pronouns, verbs and adverbs (Kosch, 2006). Therefore, tokenisation may also be described as the process of identifying linguistic words, henceforth referred to as tokens.

While the Bantu languages are all agglutinative and exhibit significant inherent structural similarity, they differ substantially in terms of their orthography. The reasons for this difference are both historical and phonological. A detailed

Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009, pages 66–73,
Athens, Greece, 31 March 2009.

discussion of this aspect falls outside the scope of this article, but the interested reader is referred to Cole (1955), Van Wyk (1958 & 1967) and Krüger (2006).

Setswana, Northern Sotho and Southern Sotho form the Sotho group belonging to the South-Eastern zone of Bantu languages. These languages are characterised by a disjunctive (also referred to as semi-conjunctive) orthography, affecting mainly the word category of verbs (Krüger, 2006:12-28). In particular, verbal prefixal morphemes are usually written disjunctively, while suffixal morphemes follow a conjunctive writing style. For this reason Setswana tokenisation cannot be based solely on whitespace, as is the case in many alphabetic, segmented languages, including the conjunctively written Nguni group of South African Bantu languages, which includes Zulu, Xhosa, Swati and Ndebele.

The following research question arises: Can the development and application of a precise tokeniser and morphological analyser for Setswana resolve the issue of disjunctive orthography? If so, subsequent levels of processing could exploit the inherent structural similarities between the Bantu languages (Dixon and Aikhenvald, 2002:8) and allow a uniform approach.

The structure of the paper is as follows: The introduction states and contextualises the research question. The following section discusses tokenisation in the context of the South African Bantu languages. Since the morphological structure of the Setswana verb is central to the tokenisation problem, the next section comprises a brief exposition thereof. The paper then proceeds to discuss the finite-state computational approach that is followed. This entails the combination of two tokeniser transducers and a finite-state (rule-based) morphological analyser. The penultimate section concerns a discussion of the computational results and insights gained. Possibilities for future work conclude the paper.

2 Tokenisation

Tokenisation for alphabetic, segmented languages such as English is considered a relatively simple process where linguistic words are usually delimited by whitespace and punctuation. This task is effectively handled by means of regular expression scripts. Mikeev (2003) however warns that "errors made at such an early stage are very likely to induce more errors at later stages of text processing and are therefore very dangerous." The importance of accurate tokenisation is also emphasised by Forst and Kaplan (2006). While Setswana is also an alphabetic segmented language, its disjunctive orthography causes token internal whitespace in a number of constructions of which the verb is the most important and widely occurring. Since the standard tokenisation issues of languages such as English have been extensively discussed (Farghaly, 2003; Mikeev, 2003; Palmer, 2000), our focus is on the challenge of Setswana verb tokenisation specifically. We illustrate this by means of two examples:

Example 1: In the English sentence "I shall buy meat" the four tokens (separated by "/") are I / shall / buy / meat. However, in the Setswana sentence *Ke tla reka nama* (I shall buy meat) the two tokens are *Ke tla reka* / *nama*.

Example 2: Improper tokenisation may distort corpus linguistic conclusions and statistics. In a study on corpus design for Setswana lexicography Otlogetswe (2007) claims that *a* is the most frequent "word" in his 1.3 million "words" Setswana corpus (Otlogetswe, 2007:125). In reality, the orthographic word *a* in Setswana could be any of several linguistic words or morphemes. Compare the following:

***A**/ o itse/ rre/ yo/?* (Do you know this gentleman?) Interrogative particle;

*Re bone/ makau/ **a**/ maabane/.* (We saw these young men yesterday.) Demonstrative pronoun;

*Metsi/ **a**/ bollo/.* (The water is hot.) Descriptive copulative;

*Madi/ **a**/ rona/ a/ mo/ bankeng/.* (Our money (the money of us) is in the bank.) Possessive particle and descriptive copulative;

*Mosadi/ **a** ba bitsa/.* (The woman (then) called them.) Subject agreement morpheme;

*Dintswa/ ga di **a** re bona/.* (The dogs did not see us.) Negative morpheme, which is concomitant with the negative morpheme *ga* when the negative of the perfect is indicated, thus an example of a separated dependency.

In the six occurrences of *a* above only four represent orthographic words that should form part of a word frequency count for *a*.

The above examples emphasise the importance of correct tokenisation of corpora, particularly in the light of the increased exploitation of electronic corpora for linguistic and lexicographic research. In particular, the correct tokenisation of verbs in disjunctively written languages is crucial for all reliable and accurate corpus-based research. Hurskeinen et al. (2005:450) confirm this by stating that "a care-

fully designed tokeniser is a prerequisite for identifying verb structure in text".

3 Morphological Structure of the Verb in Setswana

A complete exposition of Setswana verb morphology falls outside the scope of this article (see Krüger, 2006). Main aspects of interest are briefly introduced and illustrated by means of examples.

The most basic form of the verb in Setswana consists of an infinitive prefix + a root + a verb-final suffix, for example, *go bona* (to see) consists of the infinitive prefix *go*, the root *bon-* and the verb-final suffix *-a*.

While verbs in Setswana may also include various other prefixes and suffixes, the root always forms the lexical core of a word. Krüger (2006:36) describes the root as "a lexical morpheme [that] can be defined as that part of a word which does not include a grammatical morpheme; cannot occur independently as in the case with words; constitutes the lexical meaning of a word and belongs quantitatively to an open class".

3.1 Prefixes of the Setswana verb

The verbal root can be preceded by several prefixes (cf. Krüger (2006:171-183):

Subject agreement morphemes: The subject agreement morphemes, written disjunctively, include non-consecutive subject agreement morphemes and consecutive subject agreement morphemes. This is the only modal distinction that influences the form of the subject morpheme. The same subject agreement morpheme therefore has a consecutive as well as a non-consecutive form. For example, the non-consecutive subject agreement morpheme for class 5 is *le* as in *lekau* ***le*** *a tshega* (the young man is laughing), while the consecutive subject agreement morpheme for class 5 is *la* as in *lekau* ***la*** *tshega* (the young man **then** laughed).

Object agreement morphemes: The object agreement morpheme is written disjunctively in most instances, for example *ba* ***di*** *bona* (they see it).

The reflexive morpheme: The reflexive morpheme *i-* (-self) is always written conjunctively to the root, for example *o* ***i****pona* (he sees himself).

The aspectual morphemes: The aspectual morphemes are written disjunctively and include the present tense morpheme *a*, the progressive morpheme *sa* (still) and the potential morpheme *ka* (can). Examples are *o* ***a*** *araba* (he answers), *ba* ***sa*** *ithuta* (they are still learning) and *ba* ***ka*** *ithuta* (they can learn).

The temporal morpheme: The temporal morpheme *tla* (indicating the future tense) is written disjunctively, for example *ba* ***tla*** *ithuta* (they shall learn).

The negative morphemes *ga, sa* and *se*: The negative morphemes *ga*, *sa* and *se* are written disjunctively. Examples are ***ga*** *ba ithute* (they do not learn), *re* ***sa*** *mo thuse* (we do not help him), *o* ***se*** *mo rome* (do not send him).

3.2 Suffixes of the Setswana verb

Various morphemes may be suffixed to the verbal root and follow the conjunctive writing style:

Verb-final morphemes: Verbal-final suffixes *a*, *e*, the relative *-ng* and the imperative *–ng,* for example, *ga ba ithut****e*** (they are not learning).

The causative suffix *-is-*: Example, *o rek****is****a* (he sells (he causes to buy)).

The applicative suffix *-el-*: Example, *o bal****el****a* (she reads for).

The reciprocal suffix *-an-*: Example, *re a thus****an****a* (we help each other).

The perfect suffix *-il-*: Example, *ba utlw****il****e* (they heard).

The passive suffix *-w-*: Example, *o rom****i****wa* (he is sent).

3.3 Auxiliary verbs and copulatives

Krüger (2006:273) states that "Syntactically an auxiliary verb is a verb which must be followed by a complementary predicate, which can be a verb or verbal group or a copulative group or an auxiliary verbal group, because it cannot function in isolation". Consider the following example of the auxiliary verb **tlhola**: *re tlhola/ re ba thusa/* (we always help them). For a more detailed discussion of auxiliary verbs in Setswana refer to Pretorius (1997).

Copulatives function as introductory members to non-verbal complements. The morphological forms of copula are determined by the copulative relation and the type of modal category in which they occur. These factors give rise to a large variety of morphological forms (Krüger, 2006: 275-281).

3.4 Formation of verbs

The formation of Setswana verbs is governed by a set of linguistic rules according to which the various prefixes and suffixes may be sequenced

to form valid verb forms (so-called morhotactics) and by a set of morphophonological alternation rules that model the sound changes that occur at morpheme boundaries. These formation rules constitute a model of Setswana morphology that forms the basis of the finite-state morphological analyser, discussed in subsequent sections.

This model, supported by a complete set of known, attested Setswana roots, may be used to recognise valid words, including verbs. It will not recognise either incorrectly formed or partial strings as words. The significance of this for tokenisation specifically is that, in principle, the model and therefore also the morphological analyser based on it can and should recognise only (valid) tokens.

Morphotactics: While verbs may be analysed linearly or hierarchically, our computational analysis follows the former approach, for example:

ba a kwala (they write)
`Verb(INDmode),(PREStense,Pos):AgrSubj-Cl2+AspPre + [kwal]+Term`
o tla reka (he will buy)
`Verb(INDmode),(FUTtense,Pos):AgrSubj-Cl1+TmpPre+[rek]+Term`
ke dirile (I have worked)
`Verb(INDmode),(PERFtense,Pos):AgrSubj-1P-Sg+[dir]+Perf+Term`

The above analyses indicate the part-of-speech (verb), the mode (indicative) and the tense (present, future or perfect), followed by a ':' and then the morphological analyses. The tags are chosen to be self-explanatory and the verb root appears in square brackets. For example the first analysis is *ba*: subject agreement class 2; *a*: aspectual prefix; *kwal*: verb root; *a*: verb terminative (verb-final suffix). The notation used in the presentation of the morphological analyses is user-defined.

In linear analyses the prefixes and suffixes have a specific sequencing with regard to the verbal root. We illustrate this by means of a number of examples. A detailed exposition of the rules governing the order and valid combinations of the various prefixes and suffixes may be found in Krüger (2006).

Object agreement morphemes and the reflexive morpheme always appear directly in front of the verbal root, for example *le a **di** reka* (he buys it). No other prefix can be placed between the object agreement morpheme and the verbal root or between the reflexive morpheme and the verbal root.

The position of the negative morpheme *ga* is always directly in front of the subject agreement morpheme, for example, ***ga ke** di bône*. (I do not see it/them).

The negative morpheme *sa* follows the subject agreement morpheme, for example, *(fa) **le sa** dire* ((while) he is not working).

The negative morpheme *se* also follows the subject agreement morpheme, for example, *(gore) **re se** di je* ((so that) we do not eat it). However, if the verb is in the imperative mood the negative morpheme *se* is used before the verbal root, for example, ***Se** kwale!* (Do not write!).

The aspectual morphemes always follow the subject agreement morpheme, for example, ***ba sa** dira* (they are still working).

The temporal morpheme also follows the subject agreement morpheme, for example, ***ba tla** dira* (they shall work).

Due to the agglutinating nature of the language and the presence of long distance dependencies, the combinatorial complexity of possible morpheme combinations makes the identification of the underlying verb rather difficult. Examples of rules that assist in limiting possible combinations are as follows:

The object agreement morpheme is a prefix that can be used simultaneously with the other prefixes in the verb, for example, *ba a **di** bona* (they see it/them).

The aspectual morphemes and the temporal morpheme cannot be used simultaneously, for example, *le ka ithuta* (he can learn) and l*e tla ithuta* (he will learn).

Since (combinations of) suffixes are written conjunctively, they do not add to the complexity of the disjunctive writing style prevalent in verb tokenisation.

Morphophonological alternation rules: Sound changes can occur when morphemes are affixed to the verbal root.

The prefixes: The object agreement morpheme of the first person singular *ni/n* in combination with the root causes a sound change and this combination is written conjunctively, for example *ba ni-bon-a* > *ba mpona* (they see me). In some instances the object agreement morpheme of the third person singular and class 1 causes sound changes when used with verbal roots beginning with *b-*. They are then written conjunctively, for example, *ba mo-bon-a* > *ba mmona* (they see him).

When the subject agreement morpheme *ke* (the first person singular) and the progressive morpheme *ka* are used in the same verb, the sound change *ke ka* > *nka* appears, for example, *ke ka opela* > *nka opela* (I can sing).

The suffixes: Sound changes also occur under certain circumstances, but do not affect the conjunctive writing style.

Summarising, the processing of electronic Setswana text requires precise tokenisation; the disjunctive writing style followed for verb constructions renders tokenisation on whitespace inappropriate; morphological structure is crucial in identifying valid verbs in text; due to the regularity of word formation, linguistic rules (morphotactics and morphophonological alternation rules) suggest a rule-based model of Setswana morphology that may form the basis of a tokeniser transducer, and together with an extensive word root lexicon, also the basis for a rule-based morphological analyser. Since the Bantu languages exhibit similar linguistic structure, differences in orthography should be addressed at tokenisation / morphological analysis level so that subsequent levels of computational (syntactic and semantic) analysis may benefit optimally from prevalent structural similarities.

4 Facing the Computational Challenge

Apart from tokenisation, computational morphological analysis is regarded as central to the processing of the (agglutinating) South African Bantu languages (Bosch & Pretorius, 2002, Pretorius & Bosch, 2003). Moreover, standards and standardisation are pertinent to the development of appropriate software tools and language resources (Van Rooy & Pretorius, 2003), particularly for languages that are similar in structure. While such standardisation is an ideal worth striving for, it remains difficult to attain. Indeed, the non-standard writing styles pose a definite challenge.

4.1 Other approaches to Bantu tokenisation

Taljard and Bosch (2005) advocate an approach to word class identification that makes no mention of tokenisation as a central issue in the processing of Northern Sotho and Zulu text. For Northern Sotho they propose a hybrid system (consisting of a tagger, a morphological analyser and a grammar) "containing information on both morphological and syntactic aspects, although biased towards morphology. This approach is dictated at least in part, by the disjunctive method of writing." In contrast, Hurskainen et al. (2005) in their work on the computational description of verbs of Kwanjama and Northern Sotho, concludes that "a carefully designed tokeniser is a prerequisite for identifying verb structures in text". Anderson and Kotzé (2006) concur that in their development of a Northern Sotho morphological analyser "it became obvious that tokenisation was a problem that needed to be overcome for the Northern Sotho language as distinct from the ongoing morphological and morpho-phonological analysis".

4.2 Our approach

Our underlying assumption is that the Bantu languages are structurally very closely related. Our contention is that precise tokenisation will result in comparable morphological analyses, and that the similarities and structural agreement between Setswana and languages such as Zulu will prevail at subsequent levels of syntactic analysis, which could and should then also be computationally exploited.

Our approach is based on the novel combination of two tokeniser transducers and a morphological analyser for Setswana.

4.3 Morphological analyser

The finite-state morphological analyser prototype for Setswana, developed with the Xerox finite state toolkit (Beesley and Karttunen, 2003), implements Setswana morpheme sequencing (morphotactics) by means of a lexc script containing cascades of so-called `lexicons`, each of which represents a specific type of prefix, suffix or root. Sound changes at morpheme boundaries (morphophonological alternation rules) are implemented by means of xfst regular expressions. These lexc and xfst scripts are then compiled and subsequently composed into a single finite state transducer, constituting the morphological analyser (Pretorius et al., 2005 and 2008). While the implementation of the morphotactics and alternation rules is, in principle, complete, the word root lexicons still need to be extended to include all known and valid Setswana roots. The verb morphology is based on the assumption that valid verb structures are disjunctively written. For example, the verb token *re tla dula* (we will sit/stay) is analysed as follows:

```
Verb(INDmode),(FUTtense,Pos): AgrSubj-
1p-Pl+TmpPre+[dul]+Term
```

or

```
Verb(PARmode),(FUTtense,Pos): AgrSubj-
1p-Pl+TmpPre+[dul]+Term
```

Both modes, indicative and participial, constitute valid analyses. The occurrence of multiple valid morphological analyses is typical and would require (context dependent) disambiguation at subsequent levels of processing.

4.4 Tokeniser

Since the focus is on verb constructions, the Setswana tokeniser prototype makes provision for punctuation and alphabetic text, but not yet for the usual non-alphabetic tokens such as dates, numbers, hyphenation, abbreviations, etc. A grammar for linguistically valid verb constructions is implemented with xfst regular expressions. By way of illustration we show a fragment thereof, where SP represents a single blank character, WS is general whitespace and SYMBOL is punctuation. In the fragment of xfst below '...' indicates that other options have been removed for conciseness and is not strict xfst syntax :

```
define WORD [Char]+[SP | SYMBOL];
define WORDwithVERBEnding [Char]+[a | e
| n g] [SP | SYMBOL];
echo >>> define object concords
define OBJ [g o | r e | l o | l e | m o
| b a | o | e | a | s e | d i | b o]
WS+;
echo >>> define subject concords
define SUBJ [k e | o | r e | l o | l e |
a | b a | e | s e | d i | b o | g o]
WS+;
echo >>> define verb prefixes
echo >>> define indicative mode
define INDPREF [(g a WS+) SUBJ ([a | s a
] WS+) ([a | k a | s a] WS+) (t l a WS+)
(OBJ)];
define VPREF [...| INDPREF | ...];
echo >>> define verb groups
define VGROUP [VPREF WORDwithVERBEnd-
ing];
echo >>> define tokens
define Token [VGROUP | WORD | ...];
```

Finally, whitespace is normalised to a single blank character and the right-arrow, right-to-left, longest match rule for verb tokens is built on the template

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

where A, B, L and R are regular expressions denoting languages, and L and R are optional (Beesley and Karttunen, 2003:174).

We note that (i) it may happen that a longest match does not constitute a valid verb construct; (ii) the right-to-left strategy is appropriate since the verb root and suffixes are written conjunctively and therefore should not be individually identified at the tokenisation stage while disjunctively written prefixes need to be recognised.

The two aspects that need further clarification are (i) How do we determine whether a morpheme sequence is valid? (ii) How do we recognise disjunctively written prefixes? Both these questions are discussed in the subsequent section.

4.5 Methodology

Our methodology is based on a combination of a comprehensive and reliable morphological analyser for Setswana catering for disjunctively written verb constructions (see section 5.3), a verb tokeniser transducer (see section 5.4) and a tokeniser transducer that tokenises on whitespace. The process is illustrated in Figure 1. Central to our approach is the assumption that only analysed tokens are valid tokens and strings that could not be analysed are not valid linguistic words.

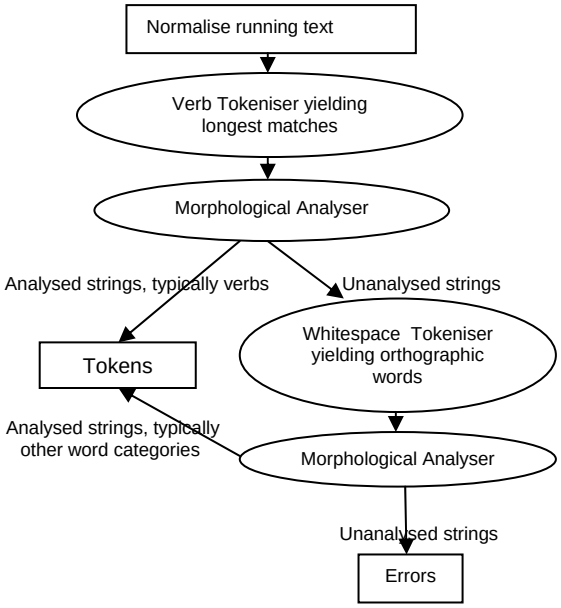

Figure 1: Tokenisation procedure

Tokenisation procedure:

Step 1: Normalise test data (running text) by removing capitalisation and punctuation;

Step 2: Tokenise on longest match right-to-left;

Step 3: Perform a morphological analysis of the "tokens" from step 2;

Step 4: Separate the tokens that were successfully analysed in step 3 from those that could not be analysed;

Step 5: Tokenise all unanalysed "tokens" from step 4 on whitespace;

[Example: unanalysed *wa me* becomes *wa* and *me*.]

Step 6: Perform a morphological analysis of the "tokens" in step 5;

Step 7: Again, as in step 4, separate the analysed and unanalysed strings resulting from step 6;

Step 8: Combine all the valid tokens from steps 4 and 7.

This procedure yields the tokens obtained by computational means. Errors are typically strings that could not be analysed by the morphological analyser and should be rare. These strings should be subjected to human elicitation. Finally a comparison of the correspondences and differences

between the hand-tokenised tokens (hand-tokens) and the tokens obtained by computational means (auto-tokens) is necessary in order to assess the reliability of the described tokenisation approach.

The test data: Since the purpose was to establish the validity of the tokenisation approach, we made use of a short Setswana text of 547 orthographic words, containing a variety of verb constructions (see Table 1). The text was tokenised by hand and checked by a linguist in order to provide a means to measure the success of the tokenisation approach. Furthermore, the text was normalised not to contain capitalisation and punctuation. All word roots occurring in the text were added to the root lexicon of the morphological analyser to ensure that limitations in the analyser would not influence the tokenisation experiment.

Examples of output of step 2:
ke tla nna
o tla go kopa
le ditsebe

Examples of output of step 3:
Based on the morphological analysis, the first two of the above longest matches are tokens and the third is not. The relevant analyses are:
ke tla nna
`Verb(INDmode), (FUTtense,Pos): AgrSubj-1p-Sg+TmpPre+[nn]+Term`
o tla go kopa
`Verb(INDmode), (FUTtense,Pos): AgrSubj-Cl1+TmpPre+AgrObj-2p-Sg+[kop]+Term`

Examples of output of step 5:
le, ditsebe

Examples of output of step 6:
le
`CopVerb(Descr), (INDmode), (FUT-tense,Neg): AgrSubj-Cl5`
ditsebe
`NPre10+[tsebe]`

5 Results and Discussion

The results of the tokenisation procedure applied to the test data, is summarised in Tables 1 and 2.

Token length (in orthographic words)	Test data	Correctly tokenised
2	84	68
3	25	25
4	2	2

Table 1. Verb constructions

Table 1 shows that 111 of the 409 tokens in the test data consist of more than one orthographic word (i.e. verb constructions) of which 95 are correctly tokenised. Moreover, it suggests that the tokenisation improves with the length of the tokens.

	Tokens	Types
Hand-tokens, *H*	409	208
Auto-tokens, *A*	412	202
$H \cap A$	383 (93.6%)	193 (92.8%)
$A \setminus H$	29	9
$H \setminus A$	26	15
Precision, P	0.93	0.96
Recall, R	0.94	0.93
F-score, 2PR/(P+R)	0.93	0.94

Table 2. Tokenisation results

The F-score of 0.93 in Table 2 may be considered a promising result, given that it was obtained on the most challenging aspect of Setswana tokenisation. The approach scales well and may form the basis for a full scale, broad coverage tokeniser for Setswana. A limiting factor is the as yet incomplete root lexicon of the morphological analyser. However, this may be addressed by making use of a guesser variant of the morphological analyser that contains consonant/vowel patterns for phonologically possible roots to cater for absent roots.

It should be noted that the procedure presented in this paper yields correctly tokenised and morphologically analysed linguistic words, ready for subsequent levels of parsing.

We identify two issues that warrant future investigation:

- Longest matches that allow morphological analysis, but do not constitute tokens. Examples are *ba ba neng*, *e e siameng* and *o o fetileng*. In these instances the tokeniser did not recognise the qualificative particle. The tokenisation should have been *ba/ ba neng*, *e/ e siameng* and *o/ o fetileng*.

- Longest matches that do not allow morphological analysis and are directly split up into single orthographic words instead of allowing verb constructions of intermediate length. An example is *e le monna*, which was finally tokenised as *e/ le/ monna* instead of *e le/ monna*.

Finally, perfect tokenisation is context sensitive. The string *ke tsala* should have been tokenised as *ke/ tsala* (noun), and not as the verb construction *ke tsala*. In another context it can however be a verb with *tsal-* as the verb root.

In conclusion, we have successfully demonstrated that the novel combination of a precise tokeniser and morphological analyser for

Setswana could indeed form the basis for resolving the issue of disjunctive orthography.

6 Future work

- The extension of the morphological analyser to include complete coverage of the so-called closed word categories, as well as comprehensive noun and verb root lexicons;
- The refinement of the verb tokeniser to cater for a more extensive grammar of Setswana verb constructions and more sophisticated ways of reducing the length of invalid longest right-to-left matches;
- The application of the procedure to large text corpora.

Acknowledgements

This material is based upon work supported by the South African National Research Foundation under grant number 2053403. Any opinion, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Research Foundation.

References

Anderson, W.N. and Kotzé, P.M. Finite state tokenisation of an orthographical disjunctive agglutinative language: The verbal segment of Northern Sotho. In *Proceedings of the 5th International Conference on Language Resources and Evalution*, Genoa, Italy, May 22-28, 2006.

Bosch, S.E. and Pretorius, L. 2002. The significance of computational morphology for Zulu lexicography. *South African Journal of African Languages*, 22(1):11-20.

Cole, D.T. 1955. *An Introduction to Tswana Grammar*. Longman, Cape Town, South Africa.

Dixon, R.M.W. and Aikhenvald, A.Y. 2002. *Word: A Cross-linguistic Typology*. Cambridge University Press, Cambridge, UK.

Forst, M. and Kaplan, R.M. 2006. The importance of precise tokenization for deep grammars. In *Proceedings of the 5th International Conference on Language Resources and Evalution*, Genoa, Italy, May 22-28, 2006.

Hurskeinen, A., Louwrens, L. and Poulos, G. 2005 Computational description of verbs in disjoining writing systems. *Nordic Journal of African Studies*, 14(4): 438-451.

Kosch, I.M. 2006. *Topics in Morphology in the African Language Context*. Unisa Press, Pretoria, South Africa.

Krüger, C.J.H. 2006. *Introduction to the Morphology of Setswana*. Lincom Europe, München, Germany.

Megerdoomian, K. 2003. Text mining, corpus building and testing. *In Handbook for Language Engineers*, Farghaly, A. (Ed.). CSLI Publications, California, USA.

Mikheev, A. 2003. Text segmentation. In *The Oxford Handbook of Computational Linguistics*, Mitkov, R. (Ed.) Oxford University Press, Oxford, UK.

Otlogetswe, T.J. 2007. *Corpus Design for Setswana Lexicography*. PhD thesis. University of Pretoria, Pretoria, South Africa.

Palmer, D.D. 2000. Tokenisation and sentence segmentation. *In Handbook of natural Language Processing*, Dale, R., Moisl, H. And Somers, H. (Eds.). Marcel Dekker, Inc., New York, USA.

Pretorius, R.S. 1997. *Auxiliary Verbs as a Subcategory of the Verb in Tswana*. PhD thesis. PU for CHE, Potchefstroom, South Africa.

Pretorius, L and Bosch, S.E. 2003. Computational aids for Zulu natural language processing. *South African Linguistics and Applied Language Studies*, 21(4):267-281.

Pretorius, R., Viljoen, B. and Pretorius, L. 2005. A finite-state morphological analysis of Setswana nouns. *South African Journal of African Languages*, 25(1):48-58.

Pretorius, L., Viljoen, B., Pretorius, R. and Berg, A. 2008. Towards a computational morphological analysis of Setswana compounds. *Literator*, 29(1):1-20.

Schiller, A. 1996. Multilingual finite-state noun-phrase extraction. In *Proceedings of the ECAI 96 Workshop on Extended Finite State Models of Lang*uage, Kornai, A. (Ed.).

Taljard, E. 2006 Corpus based linguistic investigation for the South African Bantu languages: a Northern Sotho case study. *South African journal of African languages*, 26(4):165-183.

Taljard, E. and Bosch, S.E. 2006. A Comparison of Approaches towards Word Class Tagging: Disjunctively versus Conjunctively Written Bantu Languages. *Nordic Journal of African Studies*, 15(4): 428-442.

Van Wyk, E.B. 1958. *Woordverdeling in Noord-Sotho en Zoeloe. 'n Bydrae tot die Vraagstuk van Woordidentifikasie in die Bantoetale.* University of Pretoria, Pretoria, South Africa.

Van Wyk, E.B. 1967. The word classes of Northern Sotho. *Lingua*, 17(2):230-261.