Finite state tokenisation of an orthographical disjunctive agglutinative language: The verbal segment of Northern Sotho

Winston N Anderson

Bytes Technology Group & School of Computing, University of South Africa

Bytes Office Park, 241 3rd Road, Halfway Gardens, Midrand

E-mail: winston.anderson@btgroup.co.za

and

Petronella M Kotzé

Department of African Languages, University of South Africa

P O Box 392, Pretoria, 0003

E-mail: kotzepm@unisa.ac.za

Finite state tokenisation of an orthographical disjunctive agglutinative language: The verbal segment of Northern Sotho

Tokenisation is an important first pre-processing step required to adequately test finite-state morphological analysers. In agglutinative languages each morpheme is concatinatively added on to form a complete morphological structure. Disjunctive agglutinative languages like Northern Sotho write these morphemes, for certain morphological categories only, as separate words separated by spaces or line breaks. These breaks are, by their nature, different from breaks that separate ``words'' that are written conjunctively. A tokeniser is required to isolate categories, like a verb, from raw text before they can be correctly morphologically analysed. The authors have successfully produced a finite state tokeniser for Northern Sotho, where verb segments are written disjunctively but nominal segments conjunctively. The authors show that since reduplication in the Northern Sotho language does not affect the pre-processing tokeniser, the disjunctive standard verbal segment as a construct in Northern Sotho is deterministic, finite-state and a regular Type 0 language in the Chomsky hierarchy and that the copulative verbal segment, due to its semi-disjunctivism, is ambiguously non-deterministic.

1. INTRODUCTION

The research for this paper is part of a project funded by the National Research Foundation in South Africa (The development of a computational morphological analyser for Northern Sotho). This project is part of the focus area Information and Communication Technology and the Information Society in South Africa, which recognises the central importance of research in human language technologies (HLT). Using Xerox finite-state lexical transducer software, a number of Northern Sotho morphological generation/analysis projects have been undertaken in the last two years. A pre-requisite to adequate testing and pre-processing has been the design of a tokeniser for Northern Sotho.

2. LEXICAL UNIT IN THE AFRICAN LANGUAGES OF SOUTHERN AFRICA

Tokenisation is a fundamental task of almost all HLT systems. Tokenisation of the Bantu languages presents a particular problem in that its history is based on different orthographical decisions made by linguists from different backgrounds in the last two centuries. Louwrens (1991) describes two methods of word division which emerged during the early stages of the writing history of the South African Bantu languages[i], namely the disjunctive method according to which relatively simple linguistic units are written separately from each other, e.g. the verb *ke a leboga* 'thank you' (Northern Sotho), and the conjunctive method according to which simple units are joined together to form words, e.g. *ngiyabonga* 'thank you' in Zulu. Nowadays the disjunctive method of word division is used for the Sotho languages (Northern Sotho, Southern Sotho and Tswana) as well as for Venda and Tsonga, while the conjunctive method is used for the Nguni languages (Zulu, Xhosa, Ndebele and Swati). Louwrens explains that the reasoning behind using either the one or the other method of word division is a practical one since it mainly concerns the fundamental differences between the phonological systems of the Sotho and Nguni language groups. He states that

> "Phonological processes such as vowel elision, vowel coalescence and consonantalisation which are very much less productive in the Sotho languages than is the case in the Nguni languages, render the disjunctive method of

> word division a highly impractical proposition in Nguni... In the Sotho languages, on the other hand, disjunctivism presents very few problems, since most formatives in these languages constitute syllables and can therefore easily be written disjunctively." (Louwrens, 1991:2)

Louwrens (1991:2) also points out that

> "A further reason why the conjunctive method of writing was not as acceptable to the Sotho languages as the disjunctive one, is because of their lack of semi-vowels between syllables which consist of a vowel only."

Dixon and Aikhenvald (2002) state that:

> "There is no inherent grammatical difference between these languages; it is just that different writing conventions are followed … This may have been influenced by the fact that some of the prefixes are bound pronouns and case-type markers, corresponding to free pronouns and prepositions in languages such as English and Dutch (the languages of the Europeans who helped devise these writing systems), which are there written as separate words."

In a comprehensive study of Northern Sotho grammatical descriptions ranging from the 1800's to the early 1990's, Kosch (1991) discusses all the stages of linguistic descriptions of this Bantu language. Apart from the predicative word category, she mentions the following word categories of Northern Sotho: Nouns, Pronouns, Qualificatives, Adverbs, Interrogatives, Ideophones, Interjections, and Conjunctions. The non-predicative word categories do not pose as many difficulties in the area of word identification and will therefore not be dealt with here. For the purposes of this paper, the focus will be on the verbal segment of Northern Sotho.

Poulos and Louwrens (1994:115) state that a Northern Sotho verb consists of a number of morphemes – elements that make up a word and represent the constituent parts of a word – which are put together. They say that these morphemes may be

> " a subject concord which refers to the subject of the verb; a tense marker or formative which expresses a particular tense; an object concord which refers to some or other object; a verb root which expresses the basic meaning of the action or state; and a vowel ending which comes at the end and which sometimes gives us an indication of the tense of the verb."

They also mention that not all of these morphemes are obligatory in the verb since, for instance, not all verbs include a subject or object concord morpheme. The only obligatory part of the verb is the root which represents the core of the word.

3. TOKENISATION FOR MORPHOLOGICAL ANALYSIS

The authors started work on a morphological analyser for Northern Sotho in 2003 based on the finite-state techniques and software described in Beesley and Kartunnen (2003). The morphological analysis and generation of the concatinative aspects of all forms of the verb (including reduplication) was completed in 2003. In 2004 the non-concatenative aspects (e.g. the past tense) of the verb were completed which involve more complex morpho-phonological changes (Kotzé (nd)). The year 2004 also marked the completion of analysis of the deverbative noun (Kotzé 2005 and Kotzé 2005(a)). In 2005, the other complex morpho-phonological changes around verbal extension suffixes were completed (Kotzé 2005) , as were all the rest of the parts of speech excluding the noun.

The grammars do not always cover the morphological rules and the rules required to complete tokenisation adequately, as detailed in some of the references mentioned in the last paragraph. Only actual testing highlights these inadequacies. Significant testing was done on the authors' corpora to correctly document the morphological rules.

Prinsloo and De Schryver (2002) have previously made similar findings regarding Northern Sotho. Detailed studies on much larger corpora than the ones we are currently using indicate that real life examples of conjunctivism are very important. They particularly highlight how many of the "created" examples of tokenisation do not conform to real world corpora, but state that the argument for aspects of conjunctivism are still sound. Based on examples from grammar texts such as:

(1) *gaaaapee* (<u>ga_a_a_apee</u> (*mae*)) (She doesn't boil them (the eggs))
(2) *oaoômiša* (*o_a_o_ômiša* (*morôgô*)) She causes it (the morôgô) to become dry

Prinsloo and De Schryver (2002) argue:

> "Linguistic words such as *gaaaapee* and *oaoomiša* are of course typical example creations by grammarians who base their arguments on introspection. Although neither *ga a a apee* or *o a o omiša* occur in the 5.8-million-word Pretoria Sepedi (*Northern Sotho*) Corpus, no one will dispute the sound arguments quoted above."

The arguments they refer to are arguments favouring disjunctive writing for Northern Sotho. Of course, as explained by us, this does present morphological analysis issues if conjunctive tokenisation is not completed as a pre-processing phase.

Once extensive description of the morphological rules of the language had been done and testing had commenced in

2004, 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. This was evident particularly by our experiences with real corpora of text that we were using for test purposes. The linguistic justification for this problem has been described above. It is particularly difficult to unambiguously analyse any Northern Sotho text morphologically without first completing multi-word tokenisation as discussed in the computational linguistic literature. Tokenisation techniques described by Schiller (1996), Kartunnen, Chanod, Grefenstette and Schiller (1997) and Beesley and Kartunnen (2003) in particular were examined for the implementation of multi word tokenisation.

Schiller (1996) explains how tokenisation is the process of dividing input characters into *tokens*. A tokenising transducer matches input text with the lower side of a transducer (the universal language) and outputs text corresponding to the upper side (tokens in a specifically defined language dependant format). The tokeniser deploys the directed replace operator and utilises the longest match, left to right replace operator described in Kartunnen (1996). The longest match operator in our case, ensures that the longest match for the full verbal segment is tokenised, rather than the individual morphemes.

Once this process is complete the morphological analyser can further analyse each morpheme into its correct category and full analysis of the verb stem, noun and any other part of speech can be completed by the morphological analyser.

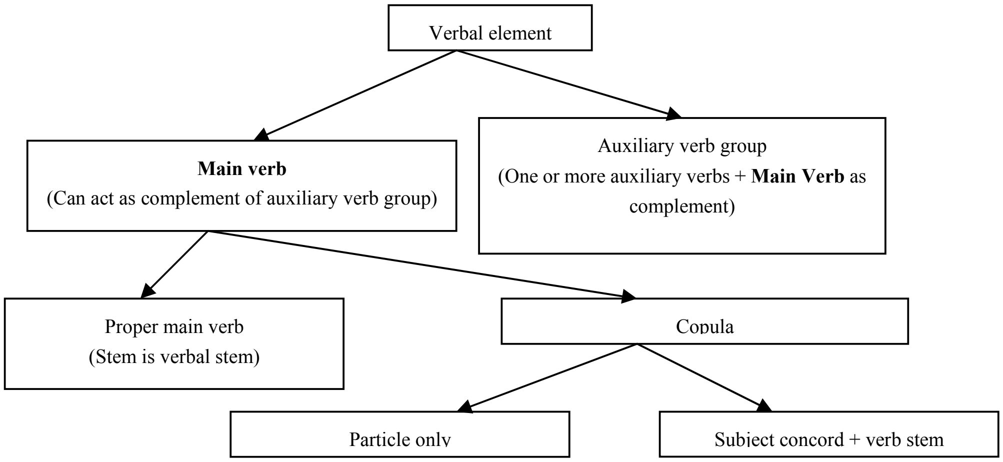
4. THE NORTHERN SOTHO TOKENISER

In Northern Sotho, a tokeniser is required to isolate categories, like a verb, from raw text before they can be correctly morphologically analysed.

The reason for this is to prevent over-analysis and unnecessary morphological ambiguity. For example, a morpheme that is not first tokenised could ambiguously be analysed as an object concord, a subject concord, a hortative prefix or a number of other morphemes. The alliterative nature of the agglutinative Bantu languages evidence repetition of similar sounding morphemes for agreement, but each morpheme has a morphologically different function. Without tokenisation the orthographic word could be analysed as a variety of different possible morphemes. With tokenisation, this ambiguity is removed as the position of the morpheme in the token allows for more accurate analysis of the morpheme. Furthermore, most linguists do not believe these morphemes should be written as separate ‘‘words’’, and hence use devices like underscores, hyphens etc. in the standard grammars to indicate the inherent conjunctivity of these morphemes.

Lombard et al (1988) discuss Van Wyk’s (1958) word tests of isolatability and mobility in order to determine the inherent stability of words. Lombard et al (1988:12) use hyphens between parts of words “in order to bridge the difference between orthographic and linguistic (autonomous) words…” So, for instance, is *ba-a-bereka* ‘they work’ inherently stable as no autonomous word can be used somewhere within this word (the parts of the word are immobile), and furthermore can neither *ba-* nor *-a-*, nor *-nyak-* nor *-a* be used alone in a sentence (the parts are not isolatable). The different parts consisting of a subject concord of the second person plural *ba-* , the imperfect tense marker *-a-,* the verb root *-berek-* and the ending *-a* all form part of one linguistic word.

When one wants to include all variables of the Northern Sotho word category “verb” or what we have termed verbal segment (to include copulative predicates), it has to be taken into account that these verbs can take many different shapes. The Northern Sotho predicate can comprise either a main verb, which in turn can be either a proper main verb or a copula, or an auxiliary word group which has a main verb as a complement. Schematically it can be illustrated as follows:


Verbal element
Main verb
(Can act as complement of auxiliary verb group)
Auxiliary verb group
(One or more auxiliary verbs + Main Verb as complement)
Proper main verb
(Stem is verbal stem)
Copula
Particle only
Subject concord + verb stem

If one takes only one of the four types of copulatives, namely the identifying copulative, the variables are as follows in the positive:

Identifying copulative: (invariable) (positive)
ke Copulative Prefix
e ba Subject Concord + Verb Stem
e ka ba Subject Concord + Potential Marker + Verb Stem
e sa ba Subject Concord + Persistive Marker + Verb Stem
e sa le Subject Concord + Persistive Marker + Verb Stem

In the case of auxiliary verb groups, the linguistic word can comprise up to seven different parts as in the following example:

le ka no se kgone go direla 'you cannot just do for*'*

Subject Concord + Potential Marker + Aspect Prefix + Negative Marker + Auxiliary Verb Stem + Subject Concord + Verb Stem

Hurskainen, Louwrens and Poulos (2005) discuss two approaches to tokenisation and morphological analysis of disjunctive verbal segments. They use Kwanyama as a tokenizing test language and then describe a subsequent morphological parser developed for Northern Sotho verb tests. They similarly highlight the same problems we have experienced with tokenization. Our approach was that our verb morphological analyzer for Northern Sotho designed in 2003 was inadequate for larger scale testing so we saw the need to create a similar but separate finite state tokenizing transducer for testing purposes. i.e. our finite state tokenising transducer is a separate finite-state machine from our finite-state morphological analyser. The Northern Sotho tokenizing transducer was then embarked on in 2004, as more texts and corpora were gathered, and built to cater for the full verbal element including copulative forms as well as other multi word tokens.

5. TOKENISER ANALYSIS

Tokenisation rules were determined by examining the standard Northern Sotho grammars, particularly Poulos and Louwrens (1994).

Tokenisation tests were run against the Northern Sotho Bible (*Bibele: Taba ye botse* 2002), poetry works, literature works, legal and other documents to test that the full verbal element is correctly isolated and tokenised.

In a text such as the Northern Sotho Bible which consists of over 700 000 tokens, almost 12 000 of these are multi word verbal element tokens that are longer than 3 words and take a verb stem as base. Of these multi-word tokens there are 11 that are 7 or more disjunctive words long, close to 60 that are 6 disjunctive words longs, hundreds that are 5 and 4 disjunctive words long, and thousands of 3 disjunctive multi word verbal segments, 3 being the mathematical mode of disjunctive verbal segments with verb as stem.

Tokenisation was fully implemented using the Xerox finite-state tools, and the compilation of a tokeniser to tokenise all words and, particularly for multi-word tokens, all verbal elements (including copulative forms and those forms containing verb stems as main complement) takes 22 minutes to compile on a 2G RAM Intel Pentium IV machine running Fedora Core 3 Linux.

Illustrative longer multi-word token examples (from the Northern Sotho Bible only) follow to demonstrate the tokeniser results. The underlined portion is the finite-state transducer tokenised verbal element.

Consider a complex verbal element that has a verbal stem isolated from the Bible[ii] (Romans 9:29):

(1) "Ge Modimo Ramaatlaohle *a ka be a se a re šadišetša* ditlogolo …
(Unless the Lord Almighty *had left us* descendants …)

a ka be a se a re šadišetša

Subject Concord + Third Person + Singular + Class1 (*a*) + Potential Marker (*ka*) + Auxiliary Verb (*be*) + Subject Concord + Third Person + Singular + Class1 (*a*) + Copulative Verb Stem (*se*) + Subject Concord + Third Person + Singular + Class1 (*a*) + Object Concord + First Person + Plural (*re*) + Verb Root (*šala* - remain) + Verbal Causative Extension (*iš*) + Verbal Applied Extension (*êl*) + Verb Suffix (*a*)

This is one verb meaning "he had caused it to leave with".

Note that the tokeniser analysis is unambiguous, but there could be ambiguity in the morphological analysis (e.g. the Subject Concord *a* could be analysed as Class1 or Class 1A).

The example shows 7 prefix morphemes that are not isolatable words followed by a verb stem. The verb stem itself consists of a verb root and extension suffixes. The above example is regarded as a single verb conjunctively written, traditionally equivalently conjunctively written in one word in other African languages, e.g. Zulu. There is only one possible tokenisation of this construct and it therefore has an unambiguous tokenisation.

Consider a complex verbal element tokenised that has a copulative base (i.e. a predicate that does not contain a proper main verb stem but another word category as its base) (1 John 2:19):

(2) … fela gabotse e be e se ba rena, gobane ge *e ka be e be e le ba rena* ba ka be ba sa dutše ba na le rena
(… for if they would have belonged to us, they *would have remained* with us …)

e ka be e be e le

Invariable Copulative Prefix (*e*) + Potential Marker (*ka*) + Auxiliary Verb (*be*) + Invariable Copulative Prefix (*e*) + Auxiliary Verb (*be*) + Invariable Copulative Prefix (*e*) + Copulative Verb Stem (*le*)

Note that in this example of a copulative, the base is not a verb but a pronoun. In this case there are either two words (Copula and Copulative Base) or one word (full verbal element). Traditionally this is regarded as two words by linguists, since they are isolatable, but they could be tokenised as two or one token for more rigorous morphological analysis. For this reason, the copulative is ambiguous in its tokenisation.

There is only one tokenisation that is the longest match for all these “words”. The verbal element consisting of a verb stem is therefore fully unambiguous and deterministic. It can be implemented by a deterministic tokeniser that is an unambiguous transducer built using the Xerox finite-state tools.

Reduplication is regarded as not being implemented in finite state. In Northern Sotho, the reduplication does not ever occur across space/tab/newline boundaries. Reduplication only occurs within the verb stem itself by reduplication of the verb root or portions of the verb root, and does not affect the tokenisation process. Since the surrounding morphemes are unaffected, it is demonstrated that tokenisation of the verb is a fully concatinative finite state process, and can be implemented in finite state tools to produce a finite state transducer.

Thus Northern Sotho tokenisation is a problem of a Type 0 language in the Chomsky hierarchy (a regular language), for tokenisation, but is context-sensitive for full morphological analysis.

The multi-word copulative verbal segment, due to its semi-conjunctive, semi-disjunctive nature, is a non-deterministic ambiguous tokenisation problem, as illustrated above. The nominal segment, since it is fully conjunctive, is a single word token and is not examined here due to trivial tokenisation.

There are other elements of Northern Sotho (pronominal such as the example *ba rena* above) that are also multi-word tokens but the tokenisation solution of these is also a relatively trivial problem to solve, as the number of separately written disjunctive morphemes are typically only two or three.

An area not yet covered by our tokeniser includes what are termed “deficient verbs” in Northern Sotho (Ziervogel & Mokgokong 1985). Deficient verbs have a semantic function similar to adverbs in languages such as English, but behave morphologically like auxiliary verbs and fit the regular expressions (in terms of tokenisation) of auxiliary verbs. Adverbs in Northern Sotho have a different morphology (for example, an adverb typically occurs after the verb, the deficient verbs occur before the verb in exactly the same place and with similar morphotactics as the auxiliary verbs). Since the grammars do not cover this adequately as highlighted by actual corpora texts, further linguistic research first has to be completed before application to the tokeniser and morphological analyser..

6. ACKNOWLEDGEMENTS

The authors wish to express their sincere gratitude to the Bible Society of South Africa for granting them permission to use an electronic version of the Northern Sotho translation of the Good News Bible for testing purposes.

7. BIBLIOGRAPHICAL REFERENCES

Beesley, K.R. (2004). Tokenizing Transducers. Xerox Research Centre. Europe. Course notes presented in Pretoria September 2004.

Beesley, K.R. & Karttunen, L. (2003). Finite State Morphology. Series: CSLI Studies in Computational Linguistics. CSLI Publications, Stanford

Bibele: Taba Ye Botse (2002). Cape Town: Bible Society of South Africa.

Dixon, R.M.W. and Aikhenvald, A.Y. 2002. Word: A cross-linguistic typology. Cambridge University Press, Cambridge.

Doke, C.M. (1929). The problem of word-division in Bantu, with special reference to the languages of Mashonaland. Department of Native Development, Southern Rhodesia.

Esterhuyse, C.J. (1974). Die ontwikkeling van die Noord-Sothoskryftaal. Unpublished MA-dissertation. Pretoria: University of Pretoria.

Guthrie, M. (1948). Bantu Word Division: a new study of an old problem. International African Institute. Memorandum 22. London: Oxford University Press.

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

Kartunnen, L. (1996). Directed replacement. proceedings of the ACL-96. Santa Cruz, California.

Kosch, I.M. (1991). A survey of Northern Sotho grammatical descriptions since 1876. Unpublished DLitt-thesis. Pretoria: University of South Africa.

Kotzé, A.E. (2005). Towards a morphological analyser for past tense forms in Northern Sotho: Verb stems with final *m* and *n*. Southern African Linguistics and Applied Language Studies 23(3).

Kotzé, A.E. (nd). Making sense of irregular realisations of the past tense suffix in Northern Sotho: An investigation focused on the phonology-morphology interface. To be submitted for publication in Southern African Linguistics and Applied Language Studies.

Kotzé, P.M. (2005a) Towards a finite-state network for Northern Sotho deverbative nouns: The morphotactic rules. Southern African Linguistics and Applied

Language Studies 23(3).

Kotzé, P.M. (2005b) A finite-state transducer for Northern Sotho deverbative nouns: The morphophonemic rules. Southern African Linguistics and Applied Language Studies 23(4).

Kotzé, P.M. & Anderson, W.N. (2005). A computational morphological analyser for Northern Sotho deverbative nouns: Applying Xerox finite-state software to traditional grammar. South African Journal of African Languages 25(1).

Lombard, D.P., Van Wyk, E.B. & Mokgokong, P.C. (1988) Introduction to the grammar of Northern Sotho. Pretoria: J L van Schaik.

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

Poulos, G & Louwrens, L.J. (1994). A linguistic analysis of Northern Sotho. Pretoria: Via Afrika Limited.

Prinsloo, D. &.De Schryver, G. (2002). Towards an 11 x 11 array for the degree of conjunctivism / disjunctivism of the South African languages. Nordic Journal of African Studies 11(2): 249-265 (2002).

Schiller, A. (1996). Multilingual Finite-State Noun Phase Extraction. ECAI-96 Workshop on Extended Finite State Models of Language. Budapest.

Van Wyk, E.B. (1958). Woordverdeling in Noord-Sotho en Zoeloe: 'n Bydrae tot die vraagstuk van woord-identifikasie in die Bantoetale. DLitt thesis, University of Pretoria, Pretoria.

Ziervogel, D & Mokgokong, P.C. (1985). Comprehensive Northern Sotho dictionary. 2nd corrected edition. Pretoria: J.L. van Schaik.

[i] Refer to Guthrie (1948) and Doke (1929) for earlier descriptions of Bantu word division.

[ii] The Bible is a particularly good text to use for testing as it is one of the oldest written Northern Sotho texts, is electronically accessible, and is known to be without any grammatical and spelling errors after years of revision. Note that other mother-tongue written edited literature texts and other translated texts were also used as tests, but the Bible yielded good illustrative examples.