Setswana Noun Analyzer and Generator

Gabofetswe Malema *malemag@mopipi.ub.bw*
Computer Science
University of Botswana
Gaborone, Botswana

Moffat Motlhanka *mofenyimoffat@gmail.com*
Computer Science
University of Botswana
Gaborone, Botswana

Boago Okgetheng *okgethengb@gmail.com*
Computer Science
University of Botswana
Gaborone, Botswana

Nkwebi Motlogelwa *motlogel@mopipi.ub.bw*
Computer Science
University of Botswana
Gaborone, Botswana

Abstract

Word morphology is a process of analysing word formation. Morphological analysis is one of the pre-processing steps in natural language processing tasks. Few studies have looked at Setswana noun morphology analysis and generation computationally. In this paper we present a rule-based Setswana noun morphological analyzer and generator. The analyser and generator implement morphological rules which are supported by a dictionary of root words with some attributes. Results show that Setswana nouns could mostly be analysed using morphological rules and the rules could also be used to generate other words. Adjectives, pronouns, adverbs and enumeratives are also included. The generator shows that Setswana nouns, adjectives and adverbs are less productive compared to verbs. The analyzer gives a 79% performance rate and the generator 92%. The analyser rules fail when multiple words have the same intermediate word and with homographs. The generator failures are due to over generation and under generation.

Keywords: Setswana, Setswana Noun Morphology, Morphological Analyzer and Generator.

1. INTRODUCTION

Setswana is a Bantu language spoken in Botswana, Namibia, Zimbabwe and South Africa. There have been few attempts in the development of Setswana language processing tools such as spell checkers, grammar checkers and machine translation. This could be attributed to lack of basic language processing tools such as morphological analysers and part of speech taggers. These tools are pre-processing phases of larger and more complex systems such as machine translation information retrieval and extraction and grammar checkers [1]. Therefore, there is need to develop basic Setswana processing tools that are accurate and usable to other systems.
This paper investigates the development of a rule-based Setswana noun morphological analyser and generator. Morphology is the study of word formation in a language. It deals with the generation, reduction and analysis of words in a language. A morphological analyser reduces words to their basic form whereas a generator derives words from a given word using morphological rules of that language. There are different approaches to morphological analysis,

the most prominent been statistical and rule-based approaches. Statistical approaches require test data to learn words formations in a language. They are language independent and less complex compared to rule based approaches. However, statistical approaches rely heavily on available data. In scarcely resources languages such as Setswana, this approach will probably not have good results due to insufficient training data. Rule-based approaches follow morphological language rules. These rules are implemented as a program to transform the words. Unlike statistical algorithms, rule-based algorithms heavily depend on language knowledge. Setswana language morphology has been studied in a number of works including [2] and [3]. We use the established rules or patterns to implement the proposed morphological noun analyser and generator.

A few research works have been done in the development of a Setswana morphological noun analyser and generator. K. Brits et al developed a prototype for automatic lemmatization of Setswana words [4]. The rule based prototype uses finite state automation of rules. The results were good with a performance of 94% for verbs and 93% for nouns. Similar works have been done on Setswana lemmatization in [5][6]. However, we have not seen any developments towards a fully developed and general purpose Setswana morphological analyser and generator.

In this paper a rule-based Setswana noun analyser and generator is presented. We develop a dictionary in which words have attributes which are used to help in generation of new words or reduction of words. We include adjectives, adverbs, pronouns and enumeratives in the study as they are very similar to nouns morphologically. We show why in some cases the rules fail and possible ways of minimizing such errors. Compound words are not covered in this study.

This paper is organized as follows. Section II presents Setswana noun morphology by category. Section III presents the proposed analyzer and generator. Performance results obtained by implementing the morphological rules in Section II are discussed. Section V concludes the paper.

2. SETSWANA WORD MORPHOLOGY

Setswana language is an agglutinative language and Setswana words can be generated from root words by adding appropriate suffixes and prefixes. A word can be used to generate many words using derivational and inflectional morphemes. The affixes change or extend the meaning of the word [2][3][7]. However, nouns do not generate as many words as verbs. In Setswana, nouns prefixes and suffixes provide essential information regarding number, diminutive and location. In this Section we look at Setswana noun, adjective, pronoun and adverb morphology. Nouns can be modified to plural, locative and diminutive forms. In some cases they could also be used to form verbs. Below we look at the different word categories generated from nouns, adjectives and adverbs.

2.1 Locative Nouns: suffix -ng

Nouns plus suffix *–ng* imply a location been talked about. Examples are

ntlo >> ntlong(house/in the house)
lwapa >> lwapeng (home/at home)
sediba >> sedibeng (well/in(at) the well)

For nouns ending with *–a,* the *–a* is changed to *–e* to give *–eng*. This is true also for nouns from verbs. It has to be noted that the suffix *–ng* when used with verbs indicates several subjects are performing an action. Therefore, somehow the analyser has to distinguish between verbs and nouns ending with *–ng.*

2.2 Diminutive Nouns: suffix -nyana

Diminutive nouns refer to smaller/younger objects. Diminutive nouns are formed by attaching suffix *–nyana* to the noun. Examples are

ntlo >> ntlonyana(house/a small house)
setlhare >> setlharenyana (tree/small tree)

However, there are several other suffixes used to form diminutive nouns in Setswana. Below are some of the most common ones.

-tlwana : ntlo >> ntlwana (small house)
-tshana: setlhare >> setlhatshana (small tree)
-na : noka >> nokana (small river)
-ngwana : tshimo >> tshingwana (small farm)
-jwana: kobo >> kojwana (small blanket)

2.3 Plural Nouns

Plural nouns refer to multiples or large amount of an object. In Setswana the prefix used to indicate the plural depends on the noun class the singular nouns belongs to [2]. Plural prefixes include *ma-,di-, me-* and *ba-*. Examples are

sethare >> ditlhare(tree/trees)
lefatshe >> mafatshe(country/countries)

In Setswana too many/much of something could also be shown by repeating the root word twice. For example

setlhare >> ditlharetlhare (lots of trees)
lefatshe >> mafatshefatshe (many many countries)
kgomo >> dikgomokgomo (lots of cattle)

It has to be noted that in some cases the prefixes are attached to the root word and in some cases the prefix replaces the prefix syllable of the singular word.

2.4 Combination of Transformations

The affixes above could be combined in forming a noun. Examples are

selepe >> dilepenyana (plural + diminutive)
selepe >> dilepenyaneng (plural + diminutive + locative)
selepe >> dilepeng (plural + diminutive)

2.5 Nouns to verbs

Some Setswana nouns could be used to form verbs. Examples are

motho >> mothofala(person/personify)

lefifi >> fifala (dark/make dark)

We therefore need a rule that could reverse the formation. That is, given a verb of this form reduce it to a noun. In some cases this could be done by identifying stems such as –fala as in the examples above.

2.6 Noun to Adjectives

Some nouns are transformed to adjectives by adding prefix bo-. Examples are

ngaka >> bongaka (doctor/doctorship)
motho >> botho(person/humanity)

segole >> bogole(disabled person/ disability)

kgosi >> bogosi(chief/chieftainship)

However, there are a few cases where the prefix bo- is used for plural. Examples are

mantshe >> bomantshe (ostrich/ostriches)

mme >> bomme (lady/ladies)

The analyzer therefore has to distinguish between *bo-* prefixes as a plural and as an adjective.

2.7 Nouns from Verbs

Setswana verbs could be used to form nouns and there are many of them. These nouns are the most common and more than root nouns and their derivatives. Below are some ways of forming nouns from verbs.

Case 1: mo/ba + verb (a >> i/o)
These nouns refer to a person who is performing a task just like *teach >> teacher* in English.

Prefix *mo-* is attached to the verb and the -a in the verb changes to *-i.* When *-a* changes to *-o* then it does not refer to a person. Examples are

ruta >> moruti (teach/teacher,preacher)
thusa >> mothusi (help/helper)
goga >> mogogo (pull/the way one pulls or pulled)
roka >> moroko(sew/the way one sews or sewed)

For plural, *ba-* is used. The nouns will be *baruti* and *bathusi* in the examples above.

Case 2: se/di + verb (a >> i/o) and le/ma + verb (a >> i/o)
These nouns refer to an object other than a person that is performing a task just like *point >> pointer* in English. Examples are

ipona >> seiponi (look at oneself/object used to look at oneself like mirror)
fetlha >> lefetlho (stir/stirrer)

Plural of *se-* is *di-* and of *le-* is *ma-*. Although this formation is largely not used to refer to people, there are a few exceptions. For example, *tagwa >> letagwa (drunk/drunkard).*

Case 3: verb (a >> o)
These nouns refer to the act performed by the verb. They are similar to "*a ruling*" from rule and "*a hearing*" from *hear* in English. Examples are

supa >> tshupo (show/the show)
itse >> kitso (know/knowledge)

Plural prefix of these nouns is *di-*.

Case 4: bo + verb (a >> i/o)
These nouns refer to the act performed by the verb. Similar to case 3. Examples are

loa >> boloi (bewitch/witchcraft)
gola >> bogolo (grow/old age)

Case 5: verb >> noun + -ng
These nouns indicate location. The nouns generated in the cases above plus suffix *–ng* result in a locative noun. Examples are

moruti >> moruting (preacher/at the preacher)
potso >> potsong (question/in the question)

Case 6: verb >> noun
Nouns can be formed by attaching prefixes *mo*, *ma, le, se* to a verb. Examples are

tlhatsa >> matlhatsa (preacher/at the preacher)
tagwa(become drunk)>> letagwa(drunkard)

2.8 Adjectives
In Setswana some adjectives could be used as nouns. The function of the word depends on the context. Adjectives have similar transformations as nouns except that some of them do not have the plural form. Most of them start with prefix *bo-*.

Adjectives to verbs
Adjectives could be transformed to nouns by adding the *–fala* suffix. Examples are

bokete >> ketefala(heaviness/become heavy)
bontsi >> ntsifala(plenty, many/increase in size,number)
nnete >> netefala(truth/be truthfull)
bonolo >> nolofala(kindness/become kind)

Adjectives, most of which could be used as nouns behave the same way as nouns when transformed to diminutive and locative forms. Some do not have the plural form but those who have replace the prefix *bo-* with *ma-*. For example, *bosula >> masula*.

Some suffixes (or adjectival stems) are used with noun class prefixes to form adjectives. For example the stem *khutswane (short)* would be

mokhutshwane (where mo noun class prefix)
sekhutshwane (where se is noun class prefix)
lekhuthswane (where le is noun class prefix)

2.9 Adverbs
Adverbs are used to modify verbs. Some adverbs could also be used as adjectives. Adverbs could be modified similarly as nouns. However, they are slight variations. Adverbs do not take plural forms. Instead repetition is used to show extent. Examples are

makuku >> makukukuku (early morning >> very early morning)
thata >> thatathata(much/a lot etc >> very much/a lot more)

For a more complete list of perfect verbs suffixes refer to [2].

2.10 Pronouns
Proper or absolute nouns
Pronouns in Setswana have singular and plural forms. Examples are
ene (him/her) >> bone(them)
nna (me/myself) >> rona (us)

Demonstrative nouns
Demonstrative nouns are used to reference to items in terms of the relative distance from the subject. They also have singular and plural forms. Examples are

yo (this one) >> ba (these ones) – refers to people.

e (this one) >> tse (these ones) – in reference to non-human objects

Prefix *bo-* could be used with nouns to form adverbs. Examples are

bonna(by myself)
borona(by ourselfs)

2.11 Quantitatives

They are used to show the quantity or number of objects. They use the stem –the. Examples are

sotlhe(all of it) >> tshotlhe(all of them)
yotlhe(all of it) >> tsotlhe(all of them)

2.12 Enumeratives

They use enumerative stems such as *–ngwe*, *-pe* and *–fe*. Examples are

ope(no one, where o is the noun class prefix)
sefe(which one, where se is the noun class prefix)

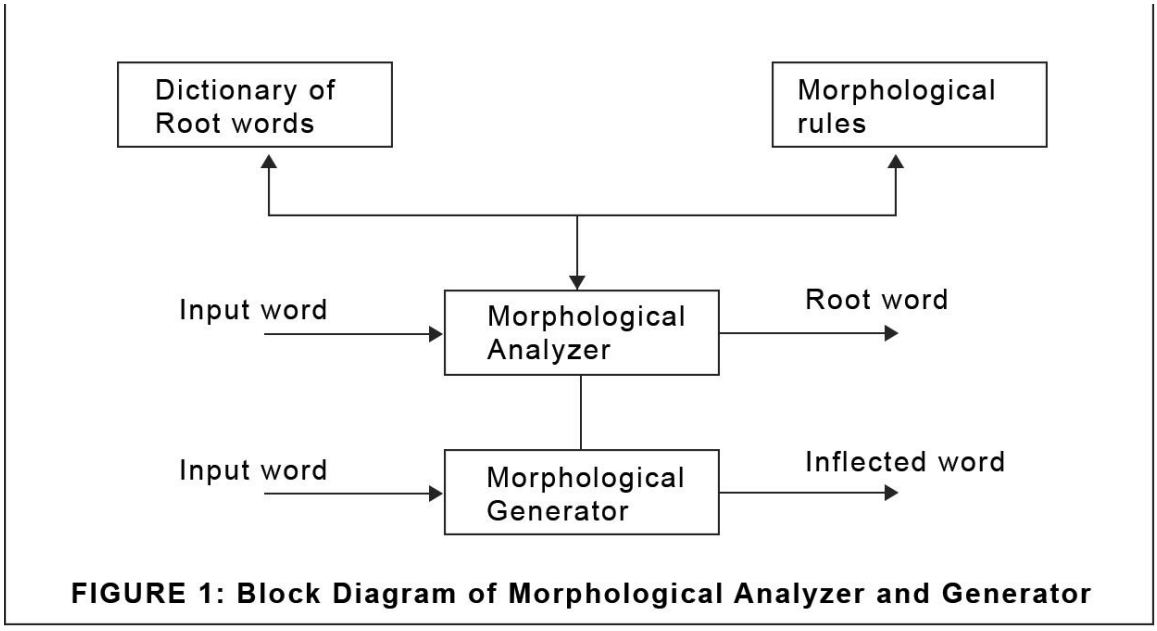

FIGURE 1: Block Diagram of Morphological Analyzer and Generator

FIGURE 1: Block Diagram of Morphological Analyzer and Generator.

3. PROPOSED ANALYZER AND GENERATOR

As shown in Figure 1, the proposed analyzer and generator uses a database of words and transformation rules. The dictionary is a list of Setswana nouns, verbs, adjectives, pronouns, enumeratives and adverbs in their basic form. Since Setswana basic form words are relatively not too many (a few thousands) we have added possible transformations (as attributes) for each word in the dictionary. The attributes help the generator to perform appropriate transformations. In the case of the analyzer, the attributes help the rules in determining whether a particular reduction is appropriate for a given word. For example

selepe(ny,tswana,ng,dir,drp)

means that the noun *selepe (axe)* could take diminutive suffixes *–nyana* and *-tswana*, locative suffix *-ng*, pural suffix *di-* which replaces singular prefix *se-* and plural could be repeated giving *dilepelepe (lots of axes).* Therefore, the generator could generate these words

selepenyana, seletswana, selepeng, dilepe and dilepelepe.

Combinations of affixes are implied. The same approach is done with the dictionary of verbs such that we could generate valid nouns from them.
The analyzer receives a word and transforms it to its basic form if not already in its basic form. It executes the transformations discussed in the above Section in reverse. The input word is recursively transformed and checked in the dictionary until it matches a word in the dictionary or it fails. The analyzer relies on the dictionary to determine if the transformation is successful or not. The attributes help to determine whether such a transformation is valid or not on the root word. It has to be noted that although these attributes help in reducing errors, they do not cover intermediate words. Covering intermediate words would drastically expand the size of the dictionary. This is one of the limitations of this approach.

The proposed Setswana morphological analyzer and generator were implemented in java. The two modules are made up of functions which perform word category transformations. The main module combines several functions to reduce or generate words. In the morphological analyzer affixes are removed in sequential order as proposed in [8]. However, reducing a word to its root form is not as trivial as described above due to the many options in word formation and in some cases a word could be transformed into multiple forms. For example given a word such as *metse*; this word could be reduced in two valid ways. Assuming the prefix *me-* is for plural, then the word is treated as a noun with a singular prefix of *mo-* giving root noun *motse(village)*. The word could also be assumed to be a verb in the mood form. In that case the suffix –e is converted to –a given the verb *metsa(sallow)*.

Basically the morphological analyzer has to try all the morphologically possible combinations and check which ones result in a valid word from the dictionary. The challenge is that some words could be reduced to valid unrelated words (over-generation).

4. EVALUATION AND PERFROMANCE

Analyzer
The evaluation of the proposed analyzer was based on how well it maps derived forms of a word to a valid basic form in the dictionary. The noun dictionary contained 3419 words and the verb dictionary 2767 words. The test data for the analyzer and generator contained 2000 words mainly from [9][10]. The key thing in the test corpus is not only the number of words but also its variety. The corpus included almost all suffixes and prefixes common in Setswana nouns, adjectives and adverbs morphological transformations. Prefixes such as *di-, me-, se-,le-,ma-* result in different transformations when applied to nouns. We also included nouns that are derived from verbs. Such nouns include prefixes such as *ba-, bo-, mo-, ma-, se-* and *le-*. Suffixes covered all word categories; plural, locative and diminutive as mentioned in the sections above.

The proposed morphological analyzer successfully reduced 1478 out of the 2000 words in the corpus giving a performance rate of 79%. The analyzer rules fails mainly due to overlapping word transformations. As pointed out earlier, some words could be reduced morphologically to an unrelated valid word. For nouns derived from verbs, the analyzer depends on the verb analyzer from [11]. The analyzer has challenges in resolving homographs as stated in [11]. For example, given the word *letlalong (on the skin)*, the analyzer will reduce it in to two ways according to morphological rules as shown below.

letlalong >> letlalo (skin)
>> letlalo >> tlala (full)

It is difficult to know the intended root word unless if there is a way to determine that from the context. In this study we have not included disambiguation rules for such cases. In some cases it would be easier to disambiguate words if their diacritics are restored.

Morphological Generator
The generator rules overall work well. Since there are not many transformations for nouns, our dictionary includes the transformations that a noun could take. The analyzer also generates nouns from verbs. However, in this case the generator can derive non-words or words that are valid morphologically but are not used. The generator was given 300 root nouns to form other words from. From the 300 words, 4345 words were generated. A manual inspection of the resulting words resulted with 298 words deemed meaningless and there were 51 words which were found not to be in the list of generated words leading to a performance rate of 92%.

Although the process of generating words in Setswana is relatively simpler than that of the analyzer, there is increased ambiguity in some formations. Some generated words do not make sense. They are morphological correct but are not used or preferred.

For example:

fetlha >> le + fetlha >> lefetlho(stirrer)
>> se + fetlha >> sefetlho
ipona >> le + bona >> leiponi
>> se + bona >> seiponi

In the example above the verb *fetlha(stir)* is combined with *le-* or *se-* to form a noun. In Setswana *lefetlho* is used instead of *sefetlho*. *Se-* and *le-* are used to form nouns from verbs. We could not find any consistency in the use of le- or se-. Therefore our generator produces both forms. There are many such cases with other prefixes such as *bo- , i-, n-* and *m-*. We have reduced such over generations by including transformational attributes in the dictionary. However, as stated in the above section, transformational attributes for intermediate words are not included in the dictionary.

5. CONCLUSIONS

A rule-based morphological analyzer for Setswana nouns has been developed. The nouns' morphology is fairly regular for most categories resulting in a high performance rate of 79%. The loss in performance is mainly due to overlapping intermediate words and homographs. The significance of the errors and error handling methods would heavily depend on the application of the analyzer. We believe the proposed analyzer is adaptable to many applications. We intend to develop disambiguation rules that could help morphological rules resolve conflicts in overlapping words. A generator was also developed which has a high performance rate of 92%. It allows transformation from nouns to other nouns and adjectives to nouns. The generator could be used to form new vocabulary. We plan to do further research in the performance of morphological rules by developing ways to improve performance. These will entail developing ways to handle ambiguity of intermediate words, improving dictionary and develop ways to determine non-words. After that we would like to combine both verb and noun rules and test if there are any conflicting rules that might lead to lower performance compared to when the two modules are separated.

6. REFERENCES

[1] V. Balakrishnan and E. Lloyd-Yemoh "Stemming and Lemmatization: A Comparison of Retrieval Performances", Lecturer Notes on Software Engineering, Vol. 2 No.3, August 2014.

[2] D.T Cole, "An Introduction to Tswana grammar", Longmans and Green, Cape Town.

[3] K. Mogapi, "Thuto Puo ya Setswana", Longman Botswana, 184, ISBN:0582 61903 3.

[4] K. Brits, R. Petorius and G.B van Huyssteen, “Automatic lemmatization in Setswana: towards a prototype”, South African Journal of Languages, 25:1, 27-47, 2013.

[5] J.H. Brits “Outomatiese Setswana Lemma-identifisering: Automatic Setswana Lemmatization”, Master’s Thesis. North West University, Potchefstroom, South Africa 2006.

[6] I. Petrorius and S.E Bosch, “Computational aids for Zulu natural language processing”, Southern African Linguistics and Applied Language Studies 21(4):276-282, 2003.

[7] Anderson Chebanne, “Intersuffixing in Setswana: The case of the perfective –ile, the applicative –ela, and the causative –isa”, Pula: Botswana Journal of African Studies. Vol. 10 No.2 pp. 83 – 94, 1996.

[8] Kruger, Capser, “Introduction to the morphology of Tswana”, Munchean, Lincon, pp314, 2006.

[9] T.J. Otlogetswe,“Poeletso-medumo ya Setswana: The Setswana Rhyming Dictionary”, Centre for Advanced Studies for African Society, 2010 ISBN: 978-1-920287-02-3.

[10] T.J Otlogetswe,“Tlhalosi ya medi ya Setswana”, Medi Publishin, 2012. ISBN: 978-99912-921-3-7.

[11] G. Malema, N. Motlogelwa, B. Okgetheng, O. Mogotlhwane, “Setswana Verb Analyzer and Generator”, International Journal of Computational Linguistics (IJCL), Vol 7, issue 1, 2016.