Proc. of The Third Intl. Conf. on Advances in Computing, Control and Networking - ACCN 2015.

ISBN: 978-1-63248-082-8 doi: 10.15224/ 978-1-63248-082-8-08

A Rule-Based Setswana Verb Lemmatizer

G. A Malema, N.P Motlogelwa, M. Lefoane

***Abstract*— Lemmatization is a pre-processing stage in several natural language processing applications such as data retrieval. There are a few attempts on Setswana word lemmatization. Developed Setswana lemmatizers do not show in details where lemmatization fails to work well leading to reduced performance. This paper presents a detailed rule-based Setswana verb lemmatizer. Challenges in verb lemmatization are pointed out by word category. The overall results show that rule based Setswana verb lemmatization gives a good performance of 87%. However, reflexive verbs have a significant large percentage of exceptions.**

***Keywords*— Setswana, Verb lemmatization,rule-based lemmatization.**

I. Introduction

Setswana is an official and main language spoken in Botswana. It is also spoken in neighbouring countries such as South Africa and Zimbabwe. Compared to other languages not much has been developed in terms of language analytical tools. The field of natural language processing covers a wide range of topics to develop tools in speech synthesis, tokenizers, parsers, dictionaries and lemmatizers. Some of these tools are pre-processing of the first phases of a larger system.

This paper investigates a rule-based Setswana verb lemmatizer. Lemmatization is a process of finding a root-word of a given word. It has been shown to improve other systems such as information retrieval [1]. It is also used to study morphology of a language. There are different approaches to lemmatization, the most prominent being statistical and rule-based approaches. Statistical lemmatizers apply statistical techniques on a sample/training data to derive lemmatization rules. 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 verb morphology has been looked at in a number of works in the literature including [2][3] We therefore use the established rules or patterns to implement the proposed lemmatizer.

The first Setswana lemmatizer in [4] had a low performance of 62%. An improved rule-based lemmatizer has been developed in [5] with a 94% performance. However, there is little information on which categories of words were used, the main challenges encountered and how they were overcome if any. Also only a random sample of 500 words was used. This paper presents a rule-based Setswana verb lemmatizer. In this paper only basic verbs are considered, with over 6000 input words. Section II presents Setswana verb morphology by word category. In Section III we present the architecture and implementation of the lemmatizer, the results obtained by implementing the morphological rules in Section IV. Section V is conclusion of the paper and suggestions for further work.

G.A Malema, N.P Motlogelwa, M. Lefoane
Department of Computer Science, University of Botswana
Botswana

II. Setswana Morphology

A. *Setswana language*

Setswana is a scarce resourced language. Setswana is considered to have seven major classes of speech [2]. These are pronouns, adverbs, interjections, adjectives, idiophones, particles, verbs and nouns. The first 5 classes are considered closed and are not affected by prefix and suffix transformations. They could therefore be processed through a lookup table. The last two classes, verbs and nouns are affected by morphological transformations through the application of prefixes and suffixes. Verbs and nouns in Setswana are mostly created by affixing prefixes and suffixes to a root word. The affixes change or extend the meaning of the word [2]. In this paper we use Setswana verb morphological rules to implement verb lemmatization. The subsections below look at various categories of verbs and their transformations.

B. *Setswana Verb Morphology*

In Setswana verbs prefixes and suffixes provide essential information regarding type, tense and mood [2]. For example the verb *bua* (*talk/speak*) could be changed in meaning by using different suffixes as below:

bua (speak)

buisa (speak to)

buisiwa (spoken to)

buile (spoken)

buisana (talk to each other)

Below we look at the application of prefixes and suffixes in different verb forms or categories. Although the application of prefixes and suffixes is regular for the most part there are cases where they do not give a valid word or do not work at all. We investigate for each category to what extend the general morphological rules work, exceptions to the rules, how significant in number are the exceptions and how they could

Proc. of The Third Intl. Conf. on Advances in Computing, Control and Networking - ACCN 2015.

ISBN: 978-1-63248-082-8 doi: 10.15224/ 978-1-63248-082-8-08

be handled. In this study we cover infinite tense, perfect tense, plural and mood verbs.

C. *Infinite Tense*

The Infinitive Setswana verbs end with an *–a* and have a combination of distinctive suffixes and prefixes. Prefixes and suffixes are used to indicate passive, causative and other forms. Below we look at how prefixes and suffixes are used to form various verb forms.

The passive (Tirwa in Setswana): Suffix –w-

Passive verbs indicate that some action is performed on you or on your behalf. Passive verbs are created by replacing the suffix *–a* in the root verb with the suffix *-iwa*. For example:

supa(point) → *supiwa(pointed to)*

loga(weave) → *logiwa (being weaved)*

bopa(mould) → *bopiwa(being moulded)*

The reverse transformation therefore will remove *–iw-* to get the base form of the word. There are several suffixes that are used to show passivity and are as outlined in Table 1.

TABLE I. SETSWANA PASSIVITY SUFFIXES

Infinite suffix	Passive suffix	examples
-ba	-biwa/-jwa	roba → rojwa
-fa	-fiwa/-swa	bofa → bofiwa
-ga	-giwa/-gwa	loga → logiwa
-pa	-piwa/tswa	kopa → kopiwa
-ma	-miwa/-ngwa	loma → longwa
-na	-niwa/-nwa	nna → nnwa
-nya	-nyiwa/-nywa	anya → anywa
-tsa	-diwa/-tswa	botsa → botswa
-tlha	-tlhiwa/-tlhwa	latlha → latlhiwa
-tla	-tliwa/-tlwa	batla → batliwa
-ta	-tiwa/-twa	feta → fetwa
-sa	-siwa/-swa	lesa → leswa
-wa	-wiwa	latswa → latswiwa
-a	-wa	lora → lorwa

Generally the *–a* is replaced by *–wa* or *-iwa* in the passive form. The given suffixes in Table 1 indicate passivity for the most part. However, there are some verbs that have the passivity suffix but are not passive verbs. Examples are *ungwa, wa, swa, nwa, lwa*. We treat these words as exceptions and put them in a lookup table. Out of the total number of causative verbs considered only 5% were exceptions.

Causative (Tirisa) suffix –is-

Causative verbs suggest that one is been caused or helped to perform some action. Causative verbs are created by attaching the suffix *–is-* to the verb. For example:

Supa(point) → *supisa (help/cause to point)*

loga(weave) → *logisa(help/cause to weave)*

bopa(mould) → *bopisa(help/cause to mould)*

The reverse transformation removes *–is-* to get the base form of the word. However, there are exceptions, which use the *–is-* suffix but do not mean causativity. Examples of such words are *tataisa, gaisa (out perform), laisa(load),fisa(burn).* Such words are considered as exceptions in the implementation. Out of the total number of causative verbs considered only 2% were exceptions. There are also causative words that end with suffix *–tsha*. For example, the causative form of the word *bona (see)* is *bontsha*, not *bonisa*. Such unique transformations are implemented in the transformation rules.

Intensity (Tirisisa): suffix –isis-

Intensity verbs suggest the action is performed with some intensity. It could be thought of as a double application of the causative suffix to a root verb. They are created by affixing the suffix *–isis-* to a verb. For example:

loma(bite) → *lomisisa(bite harder)*

batla(search) → *batlisisa(search harder)*

We did not come across exceptions of the intensity form. Therefore all verbs with the *–isis-* suffix and lemmatized by removing the suffix. All the intensity verbs considered result in the correct root word. There are a few cases where the causative form could be applied three times to a root word. We only considered two applications of the suffix.

The applicative (Tiredi in Setswana): suffix *–el-*

Applicative verbs suggest the work is done for someone. The applicative verbs are created by attaching the suffix *–el-* to a verb. Examples include:

Proc. of The Third Intl. Conf. on Advances in Computing, Control and Networking - ACCN 2015.

ISBN: 978-1-63248-082-8 doi: 10.15224/ 978-1-63248-082-8-08

supa(point) → *supela(point for or witness for in some context)*

loga(weave) → *logela(weave for)*

bopa(mould) → *bopela(mould for)*

The reverse transformation removes *–el-*. Exceptions include *bela, sela, tlhatlhela.* It was found out that there are few exceptions in this category. Out of all the verbs considered about 3% were exceptions.

Reciprocal (Tirana): suffix –an-

The verbs suggest that objects are doing the action on one another. Reciprocal verbs are created using the *–an-* suffix. Examples are:

supa(point) → *supana(point to each other)*

loga(weave) → *logana(weave one another)*

bopa(mould) → *bopana(mould one another)*

dumela(agree) → *dumelana(in agreement)*

Exceptions include *pana, gana, fapaana, rulagana.* Out of all the verbs considered only 2% were exceptions.

The Neuter-Passive (Tiregi): indicated by suffixes –eg-, -al-, -agal-, -eseg-.

The verbs suggest that the work or task is doable. For example:

apaya(cook) → *apeega(cookable)*

loma(bite) → *lomega(biteable)*

supa(point) → *supega(pointable or shown)*

sega(cut) → *segega(easy to cut)*

There are also exceptions. Some verbs have these suffixes on their root form. Examples are

sega, bega, anega, pega. Out of all the verbs considered only 3% were exceptions.

Reversal (Tirolola): suffix –olol-

Reversal verbs suggests the action is being undone or repeated. Examples are

bofa(tie) → *bofolola(untie)*

kopa(copy) → *kopolola(repeat copying)*

We have found out that most verbs in Setswana in this category have the reversal form but do not indicate reversal. Out of all the verbs considered only 20% required lemmatization transformation. The rest were in their basic form. We therefore, have assumed that verbs with the *–olol-* suffix will not need any transformation to root verb. Verbs that need the transformation are put in a lookup table. Examples of verbs that have the reversal suffix but are in their basic form include *tlhabolola, simolola, phamola*.

Iterative (Tiraka is Setswana): suffix –ka-

Iterative verbs end with *–ka-* suffix. Iterative verbs indicate that an action was or is being repeated. Examples are:

roba → *robakaka*

thuba → *thubakaka*

raga → *ragakaka*

We found a very few verbs of these form. We found 100 verbs from our sources. However, it seems most verbs could be affixed with the *–ka* suffixes although it seems there are not commonly used. Lemmatization removes the *–ka* suffix to form the root verb.

Reflexive (Itira): prefix i-

Reflexive verbs start with prefix *i-*. In some cases there is some transformation at the beginning of the simple verb. There are different transformations when a verb is converted to reflexive depending on the starting alphabet of the verb. Table 2 below summarises some of the transformations.

Verbs starting with other alphabets and combinations just introduce the reflexive prefix *i-*. For example,

tlola → *itlola*

kala → *ikala*

Verbs starting with vowels introduce *–k-* as shown in the table above. The reverse transformation (lemmatization) therefore removes *–ik* to get the base form of the verb. However, verbs starting with *k-* just insert the reflexive prefix *i-* without any further transformation. For example:

Proc. of The Third Intl. Conf. on Advances in Computing, Control and Networking - ACCN 2015.

ISBN: 978-1-63248-082-8 doi: 10.15224/ 978-1-63248-082-8-08

TABLE II. SETSWANA REFLEXIVE PREFIXES

Staring with	Reflexive prefix	Examples
-a	-ika	akela → ikakela
-e	-ike	emela → ikemela
-i	-iki	itsa → ikitsa
-o	-iko	oba → ikoba
-u	-iku	utswa → ikutswa
-w	-ikw	wela → ikwela
-g	-ikg	goga → ikgoga
-b	-ip	bona → ipona
-l	-it	loma → itoma
-r	-ith	ruta → ithuta
-s	-itsh	sotla → itshotla
-d	-it	dumela → itumela
-h	-iph, -ikh	hula → iphula, huma → ikhumisa
-f	-ph	fenya → iphenya

katela → ikatela

ketola → iketola

kiba → ikiba

kokometsa → ikokometsa.

kuka → ikuka

kwala → ikwala

How do we then differentiate verbs starting with *k*- in the base form and those that start with a vowel during lemmatization? How many of such words are there? From our investigations there are fewer verbs starting with *k*- than with vowels. Since we have not found any pattern to distinguish the two, we have put verbs starting with *k*- as exceptions in a lookup table.

The same challenge is faced with verbs starting with *l*- and t-, *g*- and *kg*- and so on as in the table above. As with verbs starting with vowels, we could not find a pattern to distinguish the two transformations. We therefore, have put the smaller groups between the two in an exception table for each transformation.

Not all verbs starting with *h*- change *h*- to *ph*- as in verbs starting with *f*-. Verbs starting with *h*- are transformed to *ph*- if there is an alternative verb starting with *f*-. Some Setswana verbs could be written starting with *h*- or *f*- but meaning the same thing. For example:

hisa and fisa (iphisa for both of them)

hata and fata (iphata for both of them)

Verbs that start with *h*- and do not have an equivalent verb starting with *f*-, transform to *kh*- when the prefix *i*- is affixed to the verb. For example,

huma(become rich) → ikhumisa(make yourself rich)

hema → ikhemisa.

The reverse transformation removes *ikh*- and replaces it with *h*-. However, there are verbs starting with *kh*- that affix the reflexive prefix *i*- without any further transformation. For example:

khurumela → ikhurumela

khontsha → ikhontsha

In this paper, we transformed verbs with –*ph*- to those starting with *f*- even if there is an equivalent of *h*-. There are few exceptions that start with *h*- and have equivalent words starting with *g*-. For example: *hakgamatsa* and *gakgamatsa*. The transformation is handled under words starting with *g*-. Words starting with *kh*- are transformed to *h*-.

Object markers

The first-person and third-person object markers are sometimes affixed to the verb. The first-object marker *n*- results in transformations similar to those of the reflexive prefix *i*-. However, the prefix *n*- becomes *m*- for verbs starting with *b*-, *p*-, *ph*-, and *f*-. Examples are:

Alola(chase) → nkalola(chase me)

bona(see) → mpona(see me)

fisa(burn) → mphisa(burn me)

phamola → mphamola

The third-person object marker *mo*- is affixed to verbs starting with *b*-. It is contracted to *m*- and the *b*- becomes –*m*. Examples are *betsa → mmetsa*, *bega → mmega*. Out of the verbs considered only 5% exceptions were found.

Proc. of The Third Intl. Conf. on Advances in Computing, Control and Networking - ACCN 2015.

ISBN: 978-1-63248-082-8 doi: 10.15224/ 978-1-63248-082-8-08

Plural Verbs

Setswana verbs could also indicate plural which is achieved by adding a suffix *–ng* to the verb. For example

rapela(pray) → *rapelang*

nthebola → *nthebolang*

aga(build) → *agang*

ipopa → *ipopeng* (note: *a* changes to *e* with reflexive verbs)

The root form is found by removing the suffix *–ng*. The plural suffix *–ng* is removed first.

D. *Perfect Verbs*

Perfect verbs behave in a similar way as in infinitive tense. Most perfect tense verbs use the suffix *–ile*. Table 3 below shows suffix conversions from infinitive to perfect form. Lemmatization removes the perfect suffixes and replaces them with infinite suffixes.

TABLE III. SETSWANA PERFECT SUFFIXES

perfect	infinite	examples
-etswe,-otswe,-utswe	*-lwa*	*epetswe → epelwa, butswe → bulwa*
-ditse, -tswitse	*-tsa*	*onaditse → onatsa*
-eile	*-aa*	*reile → raa*
-ntse	*-nya*	*omantse → omanya*
-tshitswe	*-tshwa*	*bontshwa → bontshitswe*
-sitswe	*-siwa/swa*	*gasitswe → gasiwa/gaswa*
-tshitse	*-tsha*	*bontsha → bontshitse*
-ntswe	*-nngwa*	*omantswe → omanngwa*
-sitse	*-sa*	*gasitse → gasa*
-dile, -tse	*-la*	*adile → ala, setse → sala*
-lwe	*-wa*	
-ditswe,-tswitswe,-tsitswe	*-tswa*	*onaditswe → onatswa atswitswe → atswa phatsitswe → phatswa*
-ile	*-a*	*agile → aga*
-nne	*-na*	*ganne → gana*
-nwe	*-nwa*	*bonwe → bonwa*
-ne	*-na*	*Bopagana → popagane*
-ame	*-ama*	*Palama → palame*
-ere	*-ara*	*Apere,sikere, hulere,itshopere*

Reflexive transformations are as in infinite forms. For example:

supa → itshupa → itshupile.

loga → itoga → itogile.

bopa → ipopa → ipopile.

The perfect tense rules as stated in Table 3 above work for most of the verbs. There were few exceptions. Out of the verbs considered only 1% were exceptions.

Moods

Setswana like other languages has ways to indicate aspect and mood [2]. Setswana verbs generally indicate mood by changing the basic verb ending *–a* to *–e*. Examples:

palama → palame

loga → loge

soba → sobe

In this case the lemmatization process replaces the *–e* with *–a*. However, there are few instances where it could indicate past tense. Examples are *nole*, *same*. There seems to be no transformation in this category except of *–a* to *–e*. This category was therefore not implemented.

E. *Implementation of Setswana Lemmatization rules*

The lemmatization program comprises of a set of transformation rules as described in the above sub Section. However, in lemmatization the process is reversed, given a word find the root word by reversing the transformation done on the root word. As shown above the transformation rules mostly work and therefore applied to an unknown word should produce a good form of its lemma.

Setswana language is not standardised and therefore there are different ways of writing the same word. In most cases there are two ways of writing the same word. We considered both ways in the implementation of the proposed lemmatizer. Examples of such cases are described below.

Proc. of The Third Intl. Conf. on Advances in Computing, Control and Networking - ACCN 2015.

ISBN: 978-1-63248-082-8 doi: 10.15224/ 978-1-63248-082-8-08

In Setswana, some words could be spelled differently with an introduction of -l-. For example: *ta and tla; latha and latlha; thapa and tlhapa;thaga and tlhaga*.

As shown in the passive table, two suffixes could be used for a particular transformation. Examples are: *gamiwa* and *gangwa* mean the same thing.

There are cases where some words could be transformed into more than one verb. For example:

ntshwaa → *swaa* | *tshwaa*

ntika → *tika* | *dika*

In these cases we performed one transformation. We did not cater for alternative transformations.

Homographs are words that have identical pronunciations but different meanings. In Setswana there are many homographs but usually they belong in different word classes. Therefore any word encountered in this case was assumed to be a verb. Example: *nama* (verb, stretch legs out) and *nama* (noun, meat)

equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

III. Architecture of the Lemmatizer

The proposed lemmatizer is implemented in Python. The morphological rules descrbed above are implemented as patterns. Exceptions are put in arrays. Each verb category is represented by a class or submodule. Prefixes are removed following a sequential order proposed in [6][7]. However, the sequencing is slightly modified with the introduction of reversal and iterative verb forms. Prefix –ng is removed first. The-ng prefix is used for both plural and past continuous tense. For example: *tlhapileng, batlang*. It is then followed by passive (Tirwa) –iw-, the perfective (Paka Pheti) –il-, the reciprocal (Tirana) –an-, the applicative (Tiredi) –el-, the neuter-passive (Tiregi) –eg, the causative (Tirisa) –is- , and finally the reversal (Tirolola) –olol-. The major reason in our opinion for the sequence is that in some cases you can apply one transformation over another but not verse versa. For example, you can apply passive transformation on a causative verb but not verse versa. For example, *dirisa* → *dirisiwa*. This example shows that we should consider removing passive before causative prefixes in Setswana verbs. However, there are many examples that show that some forms could be interchanged. For example we could apply the applicative + reciprocal suffixes on the verb *utlwa* (*hear, taste, understand*) to get *utlwelana* (*taste for one another*). We could apply the two transformations in reverse to get *utlwanela* (*friendship with one another*). We therefore run our sequence of transformations more than once to catch all the transformations to the root word.

you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. *Implementation*

An iterative Setswana rule based lemmatizer was implemented using python. The implementation is based on the famous pipes and filter architecture [8] as shown in figure 1 below. Advantages of this architecture include flexibility and robustness. Word category transformations are independent and could be rearranged. Nine Setswana verb categories were considered and for each verb category a filter was implemented. They transform a word from one verb category to another or from a verb category to the root word. Because of the former, filters were chained as shown in figure 1.

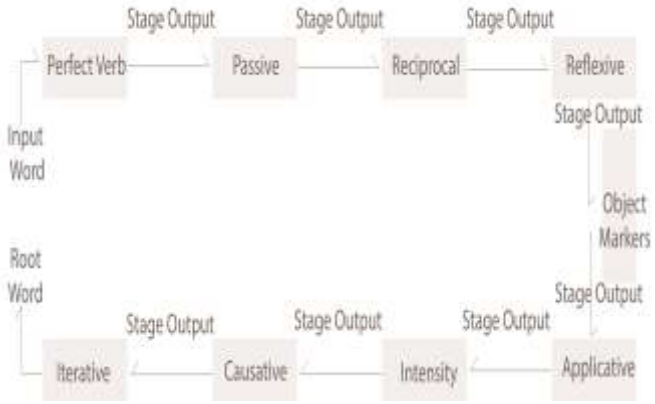

Figure 1. Sequencing of word transformations by word category

Our preliminary analysis of the transformations based on the structure of Setswana suggested the transformation sequence adopted. However, opportunity for optimising the transformation sequence exists. A finer granularity of the verb categories can lead to more focused highly accurate

Proc. of The Third Intl. Conf. on Advances in Computing, Control and Networking - ACCN 2015.

ISBN: 978-1-63248-082-8 doi: 10.15224/ 978-1-63248-082-8-08

transformations. A sequencing of these finer transformations will provide a highly optimised accurate lemmatiser.

IV. Performance of Proposed Lemmatizer

We present our results focusing on: accuracy statistics per verb category, observed problems, and transformation sequence deficiencies.

The evaluation of the lemmatizer was based on how well the lemmatizer maps derived forms of a verb to its basic form, mapped verbs are genuine linguistic variants and does not lemmatize basic word forms.

A corpus of 6000 verbs was created mainly from [9][10] comprising of verbs from all categories considered above. With use of the lookup tables the overall performance is improved to 87%. Incorrect words did not follow the rules and were not in exceptions.

Although the verbs found in [9] and [10] may not be exhaustive they represent a fairly large proportion of Setswana verbs. Many verbs follow the same rules; we considered as many verbs as we could in an effort to find out if we could catch obscure exceptions to the rules.

Accuracy statistics

The graph below in figure 2 shows the accuracy level for each morphological category. It shows that, in overall, the iterative rule based Setswana lemmatiser archives accuracy levels of over 80%. It further shows that the best morphological category with over 90% accuracy is tirisa (Causative) and the worst with over 65% accuracy is itira (Reflexive). The low level of accuracy is due to a number of factors.

1). Some words in lookup table, but others that could be derived from those words not in lookup table.

2). Some words have similar structure as a certain morphological category, but being root words.

3). Difficult to place all deserving words in lookup tables

We identified the primary causes to be the transformation sequence and the fact that some words that are not causative have a similar structure as causative. We are currently analysis the different sub-morphological categories to better understand internal morphological category dynamics. This will provide opportunity for further optimisations.

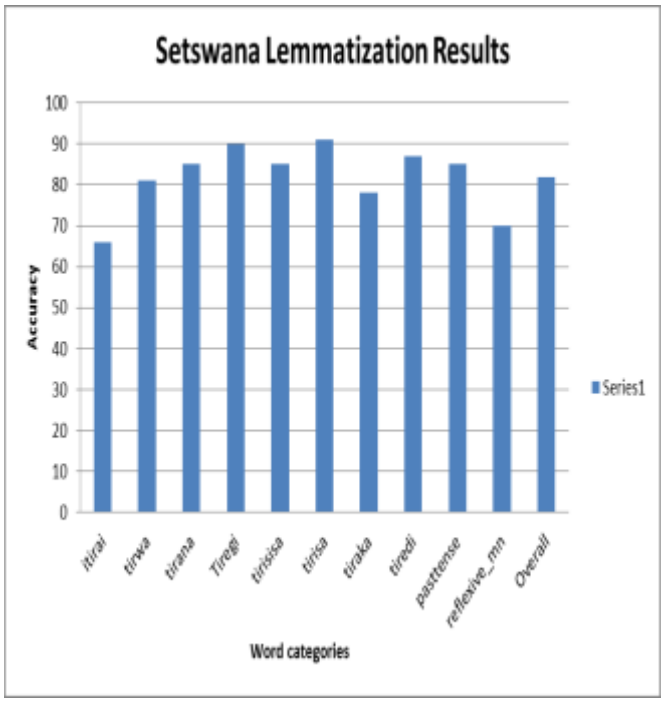

Figure 2. Lemmatization results by word category

Problematic verb categories

The following cases were identified as problem cases that need further analysis and implementation.

- The first is tirolola. Many words have structural similarity to tirolola but do not belong to that morphological category. Many words have structural similarity to tirolola, but are actually root words, so tirolola will transform them to non-existing words.
- Most of the word patterns in the object markers word category can be transformed to two different forms. Further research need to be done to see how two words with similar pattern can be distinguished. For example, mpota => pota and mpona => bona. This was observed in a number of words in this category and we couldn't find an appropriate mechanism to handle such.
- The reflexive morphological category exhibits similar problems to object markers. They are two different transformations that could be applied to words with similar patterns. For example, Words beginning with iko- can either be transformed by deleting the ik- or deleting the i- alone. Example, ikoketsa => oketsa whereas ikopa => kopa. Using the word pattern alone, it is not clear which transformation to apply. Other patterns exist which show similar problems.
- For passive morphological group, same problems as for reflexive word categories are observed. As an example words ending with –diwa have two different valid transformations. First –diwa can be replaced by la as in: badiwa => bala. In other words, -diwa is

Proc. of The Third Intl. Conf. on Advances in Computing, Control and Networking - ACCN 2015.

ISBN: 978-1-63248-082-8 doi: 10.15224/ 978-1-63248-082-8-08

replaced by –tsa as in: kgothadiwa => kgothatsa. Many other word patterns exist in this category that has the same challenge.

Transformation sequence deficiencies

The current transformation sequence was derived based on the structure of Setswana words. The basic idea was to sequence them such that enabling transformations are performed first. The following is an extract from our output: robalanwa => tirwa => robalana => tirana => robala => **robala . It shows the input word (robalanwa) and the sequence of transformations it goes through until it reaches the root word. For an insight into some of our transformation sequence, see figure 2 below. The input word is in the morphological category of tirwa (passive). The first transformation produces a word in the morphological category of Tirana (reciprocal) and the final transformation produces a root word. The number of transformations varies between and within morphological categories. We conclude that this variable is dynamic and solely depended on the word being processed. We foresee opportunities for improvement based on transformation order and granularity of transformations.

- ribegile => **pasttense** => ribega => **tirege** => riba => **riba =>
- humanegile => **pasttense** => humanega => **tirege** => humana => **tirana** => huma => **huma =>
- iphaphile => **pasttense** => iphapha => **reflexive_i** => fapha => **fapha =>

v. Conclusions

A rule-based lemmatizer for Setswana verbs has been developed. The verbs' morphology is fairly regular for most categories resulting in a high lemmatization rate of 87%. Challenges in rule-based lemmatization include words with similar suffixes and prefixes, different spellings, homographs and a large number of exceptions in some categories. Only infinite and perfect forms were considered. Further investigation of exceptions is needed and implementation of other forms.

Acknowledgment

We acknowledgment and thank Prof. T. Otlogetswe, Department of English, University of Botswana for helping us with root words and basic Setswana morphology principles.

References

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

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

[3] I Mogapi, K, "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] Jeanette H. Brits "Outomatiese Setswana Lemma-identifisering: Automatic Setswana Lemmatization", Master's Thesis. North West University, Potchefstroom, South Africa.2006.

[6] Kruger, Capser, "Introduction to the morphology of Tswana", Munchean, Lincon, pp314, 2006.

[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] Van Vliet, H, "Software Engineering: Principles and Practice", Second Edition, Wiley, 1999.

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

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