Automatic Diacritic Restoration for Northern Sotho

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

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

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

Abstract

Diacritic markers are usually not inserted in text for convenience as users type text. However, text without diacritic markers could affect the quality of its analysis as it may affect how it is pronounced and its meaning among others. The number of diacritics and the impact of not inserting them vary from language to language. The processes of restoring diacritics in the text can be looked at as language-dependent and language-independent and also as word-based or syllable based. Northern Sotho language uses two diacritic markers to indicate pronunciation and also distinguish between homographs in some cases. Very little research has been done on diacritics restoration in the Northern Sotho language. In this paper, we show that morphological word transformations are consistent in how they insert or do not insert diacritics in derived words. We focus on the caron diacritic marker.An input word is reduced to its root form by a morphological analyzer. The accented form of the root word is retrieved from the diacritic dictionary. This word, together with morphological rules is used to determine the diacritics of the input word. The implemented tool gave a recall performance of 86% on test data. Most errors were due to failures in the morphological analysis of the input word.

Keywords: Diacritic Restoration, Northern Sotho.

1. INTRODUCTION

Many languages use diacritics in their orthography. Diacritics are used to indicate how words or syllables are to be pronounced. The number of diacritics varies from language to language. In many languages diacritics are not used in everyday writing as they are inconvenient to insert as one is typing. Lack of diacritics in the text could result in difficulties in reading or understanding of the text (Shaikh, 2017; Ezeani, 2019). The severity of the problem varies from language to language and word to word. Lack of diacritics in the text may negatively affect the complex analysis of text for applications such as machine translation, sentimental analysis, and speech synthesis as diacritics provide extra information that reduces ambiguity in performing these various tasks (Ungurean, 2008; Stankevicius, 2022). Languages in which restoration of diacritics is successfully performed have seen improvement in some high-level applications (Shaik, 2017) (Mihalcea, 2002).

Diacritic restoration is the process of inserting diacritics in the text that did not have them. There are different approaches used in diacritic restoration. Some are language-independent while some are language-dependent. Diacritic restoration could be looked at as a disambiguation problem in some cases. In this case, the tool determines which word was intended (Asahiah, 2018). In some cases, diacritic restoration can be viewed as a spell-checker tool. Insertion of accents in a word, indicates how the word should have been spelled in the first place.

Diacritics restoration techniques could be used to provide such applications (spell checker, disambiguation) with more information or high-quality data to improve their performance.

Like many tasks in natural language processing (NLP), diacritic restoration techniques or tools are faced with the problem of ambiguity. It is not always obvious where or how diacritics should be inserted. Commondiacritic restoration general approaches of statistical, rule-based, and hybrid approaches are often applied in diacritic restoration tools at a word or character level (Asahiah, 2018; Ezeani, 2019). Some methods are pure statistical methods that do not rely on any language knowledge and therefore are language-independent (Asahiah, 2018). In general, proposed diacritic restoration tools in literature achieve high-performance results with recall rates in the upper ninety percent.

In this study, we propose a rule-based diacritic restoration tool for Northern Sotho. Northern Sotho is one of the 11 official languages in South Africa. It is related to Setswana and Sesotho which form part of the Sotho group of languages. For the Northern Sotho language, there has been little done to automatically restore or insert diacritics in existing text or as text is typed. In (Pauw,2007) a machine learning approach was used to restore diacritics for several resource-scarce African languages including Northern Sotho also referred to as Sesotho Sa Leboa. The character-based approach achieved a 70% accuracy rate. In (Malema,2018), a rule-based diacritic restoration tool was developed for Setswana. The tool is based on a dictionary of root words and morphological rules. The approach gave a 77% recall rate with failures coming from the morphological analyzer. Northern Sotho is very close to Setswana in words and word morphology. However, the two languages differ in pronunciation of words mostly due to the use of the caron marker with *s* (š). In Setswana, only a few words or syllables require the use of š. Motivated by the frequent use of diacritics to determine pronunciation, especially the caron marker, more work needs to be done on diacritic restoration for Northern Sotho. In this paper, we describe and evaluate a word-based diacritics restoration tool designed for the Northern Sotho language. Our premise is that since Northern Sotho is closer to Setswana in word morphology its morphology is also regular and consistent as that of Setswana such that a rule-based determination of diacritics is successful. The proposed approach uses a dictionary of root words, a morphological analyzer, and word morphological rules to determine the diacritics of a given word as was done in (Malema, 2018) for Setswana. Furthermore, Northern Sotho is a low-resourced language with limited resources to use in statistical approaches. The proposed rule-based approach provides a baseline study for future studies in diacritic restoration for Northern Sotho.

2. NORTHERN SOTHO DIACRITICS

In Northern Sotho, diacritics play an important role in determining the meaning and pronunciation of words. Northern Sotho applies the circumflex (^) marker on vowels *o* and *e* and the caron marker on *s* to have š. The circumflex (^) on vowels e and o is used to show how the word should be pronounced. The caron marker is applied on letter s to also change the pronunciation of the syllable. Northern Sotho texts like in Setswana usually do not include the circumflex marker. However, the caron marker is usually included in Northern Sotho text. Probably it is because text without the caron marker makes the text very ambiguous to read. We have not come across any studies on the frequency and impact of accent markers in Northern Sotho. However, as shown by some examples in this paper, accent markers may distinguish words in some cases by specifying how they should be pronounced. Without a marker, a reader may choose the unintended word while reading leading to ambiguity, wrong semantics, or rereading which takes more time. This is illustrated by examples of Northern Sotho homographs below. Accent markers are used to distinguish the words.

tshela/tšhêla (jump over/pour)
noka/nôka (river/waist)
tsêbê/tsebê(ear/know)
letše/lêtše(call,slept)
pêo/pêô(seed/the act of installing or putting something)

In some cases, accents are inserted to correct the spelling of a word such that it is pronounced appropriately. In Northern Sotho, this is particularly the case with the use of š in some syllables. Some words in their root form use accents and should be written that way otherwise no such words exist. Although accents are not commonly inserted in text, most writers and documents at least insert the s accent. Writing a word/syllable without š may confuse the reader as to whether the intended syllables was the one without the accent or not.Some of the syllables that are spelled and pronounced differently in Northern Sotho as compared to Setswana are as follows:

sa>>ša
sa>>ši
so >>šo
tsa>>tša
tse>>tše
tsha>>tšha
tshi>>tšhi

There are some word transformations that result in syllables that require the use of š. Just like in Setswana we have noted that Northern Sotho morphological transformations and the use of š are consistent or regular across the different word transformations that may occur. We outline the different transformations below that require the use of accents.

Causative Transformations

Such transformations cause something to happen or help something to perform an action. The suffix of a given verb is transformed to result in the causative form of the word. Transformations which depend on the suffix of the transformed word are shown below.

-a >> -iša

In general, a verb is converted to its causative form by replacing the suffix *-a* with *-iša.* Examples are.

loka>>lokiša (good/make good)
thuba>>thubiša (break/cause to break)

-la >>tša / la >>diša

In some verbs which end with *-la* the suffix is changed to *-tša*and in some verbs to *-diša.* Examples are:
lala>>latša (sleep/cause to sleep)
laola>>laodiša (control/cause to control)
matlafala>>matlafatša (be strong/make strong)

ga >>ša / ga >>giša

In some verbs the suffix -ga is converted to *-ša* and in some to *-giša*. Examples are:
goga>>gogiša (pull/help to pull)
tsoga>>tsoša (wake up/cause wake up)

-nya/-na>>ntšha

Verb suffixes -na and -nya are changed to -nt*šha*
Examples are:
akanya>>akantšha (think/make to think)
bona >>bontšha (see/ show)

The general observation is that all transformations to causative that introduce a syllable that has s in it, the *s* must be accented (š).

Applicative Transformations
This form generally indicates something is done on behalf of something. It also transforms a given verb depending on its suffix as shown below.

-nya>>nyetša
Examples are:
baakanya>>baakanyetša (fix/fix for)
akanya>>akanyetša (think/think for)

-sa>>setša /ša>>šetša
Examples are:
hlaloša>>hlalošetša (explain/explain for)

-tšha>> -tšhetša
Examples are:
ntšha>>ntšhetša (remove/remove for)

-tšwa>>tšwetša
Examples are:
tlhatšwa>>tlhatšwetša (clean/clean for)

-tsa>>letša.
Examples are:
-bitsa>>biletša (call/call for)
-botsa>>boletša (ask/ask for)

Perfect Tense
The transformation of a verb to its perfect form depends on its suffix. Below are some of the transformations according to different suffixes.

sa>>sitše / ša>>šitše
Examples are:
hlaloša>>hlalošitšê (explain/explained)
tliša>>tlišitšê (bring/brought)

tsa/ tša>>ditše
Examples are:
nosetša>>noseditše(water/watered)
bitša>>biditše (call/called)
fetša>>feditše(finish/finished)

-tšwa>> -tswitše
Examples are:
hlatšwa>>hlatšwitše (clean/cleaned)

-na/-nya>>ntše
Examples are:

nna>>ntše (sit/seated)
senya>>sentše(damage/damaged)
baakanya>>baakantše(fix/fixed)
hlakanya>>hlakantše(mix/mixed)

-la >> -etše
Examples are:
lala>>letše(sleep/slept)
sepela>>sepetše(go/went)

Neuter-Passive
Neuter-Passive implies the task is doable or easy to carry out. It is formed by changing the verb suffix -a to *-ega* or *-ešega*. Examples are:

sega>>segega(cut/easier to cut)
boloka>>bolokešega(save/savable)

Verb to Noun Transformations
In Northern Sotho, verbs can be transformed into nouns by changing the suffix *-a* to *-o*, and adding prefixes and suffixes. For example:

oketša>>mooketšo(increase/ the increase)
thuša>>thušo(help/the help)
šoma>>bašomi (work/ workers)

These transformations also are regular and consistent. That is, they follow the same patterns for the prefixes and suffixes added to the word. This is also true for other noun transformations such as singular to plural.

Locative Nouns
Nouns could be used to indicate location by adding the suffix *-ng*. In cases where the suffix of the noun is *-a* the *-a* is changed to *-eng*.

Examples are:
noka>>nokeng (river/at the river)

The purpose of the examples above is to demonstrate that word morphology especially verb morphology which is more productive than noun morphology in Northern Sotho follows a regular pattern. The examples given are not exhaustive. A comprehensive introduction to Northern Sotho word morphology is found in (Lombard,1985;Poulos,1994). These transformations are generalized and used to restore diacritics of a given word as explained in the next section.

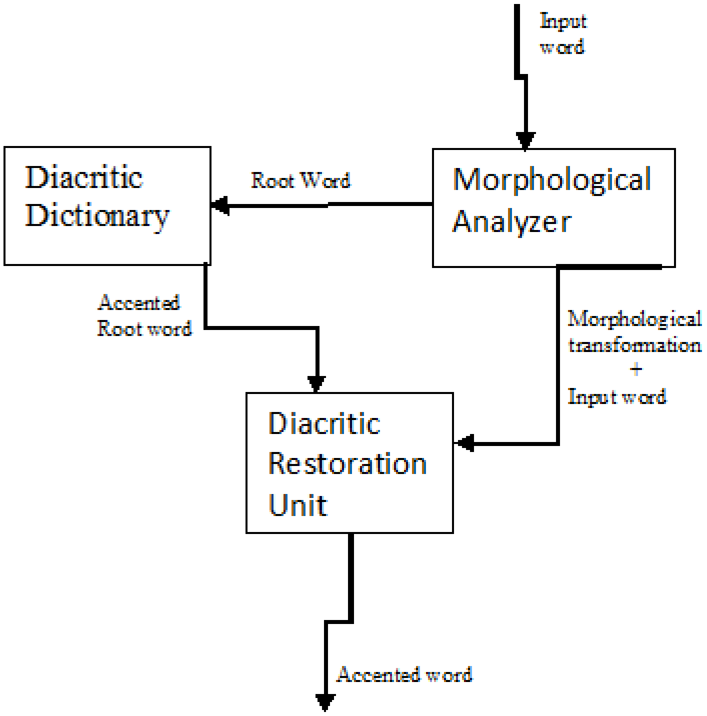

FIGURE 1: Block diagram of diacritic restoration process.

3. AUTOMATIC RESTORATION OF DIACRITICS

We have implemented a tool to automatically restore š accents in Northern Sotho text as illustrated by the block diagram in Figure 1. Input text is tokenized into words which are then fed into a morphological analyzer. The morphological analyzer reduces the input word to its root form. The root word is fed into the dictionary to retrieve the diacritic form of the root word. As shown in Figure 1 the morphological analyzer also feeds the input word and morphological transformations to the Diacritic Restoration Unit which combines these inputs with the diacritic form of the root word from the diacritic dictionary to determine where diacritics should be inserted in the input word. The morphological analyzer determines the transformations made in the derivation of the input word from the root word. The diacritic restoration unit has rules which follow morphological transformations shown in the previous section. For example, given *lokisa,* which is from *loka* through a *causative* transformation, the rule is that the suffix *-isa* should be written as *-iša*. The rules are as per observations or conclusions made on word morphological transformations. If there is no rule for a transformation in the Restoration Unit, then the word is left as it is. That is, no insertion of diacritics is required.

4. PERFORMANCE

The proposed Northern Sotho diacritic restoration tool was implemented in Python. A rule-based morphological analyzer developed in (Malema,2016) was modified to break down Northern Sotho words to their root form. A dictionary of Northern Sotho root words with diacritics was developed using (Louwrens,1994;Jehovah's Witness,2022;Schryver,2022). For root words that are not in the diacritic dictionary, the tool assumes that they are written as they appear. That is, without any diacritics. This significantly reduces the size of the dictionary. All morphological transformations requiring the insertion of diacritics were written as rules in the Diacritic Restoration Unit.

The tool was given a test file with 180 words of different forms and transformations some requiring insertion of the caron marker on s (š) and some not. The tool correctly restored accents in 154 words and incorrectly inserted accents in 13 words resulting in a recall rate of 86% and a precision rate of 92%. Both false positives and false negatives equally affect the quality of the data as they distort how the affected words are to be pronounced and possibly their meaning. The results show that Northern Sotho verb morphology is regular with regard to diacritics. The rule-based approach is predictable and gives better results compared to the machine learning approach in (Pauw,2007). The tool fails to restore targeted accents due to a number of reasons as explained below.

Failure of the Morphological Analyzer
The tool relies on the morphological analyzer to give the correct transformation and the root word. If the analyzer fails, then it affects the quality of the restoration process. It was also noted that there are some words that have unusual transformations which the morphological analyzer has not implemented.

Missing rules
Like other rule-based approaches, the tool fails when there is a rule missing.

Disambiguation of Homographs
It has to be noted that the proposed tool processes a word without context and therefore does not disambiguate homographs. In this paper, the tool gave all possible homographs it could find and accented them accordingly. However, in some cases, the tool did not find all homographs.

5. CONCLUSIONS

This paper presents a rule-based diacritic restoration approach for Northern Sotho. It was demonstrated that Northern word morphology is regular and therefore knowing the transformation between words and the root word one could easily determine where diacritics should be inserted. The approach gives a high performance from the text given. With this performance, the approach could be used in the pre-processing phase of applications such as word processors, editors, text-to-speech, and machine translation among others. However, it heavily relies on the accuracy of the morphological analyzer. The analyzer affects the accuracy of the tool and also the speed of the tool. For such a tool to be used as users are typing it needs to be fast. Further studies are needed to improve the different parts of the approach such as the morphological analyzer, word transformation rules, disambiguation rules and diacritic dictionary.

6. REFERENCES

Asahiah F. O., Odejobi O. A., Adagunodo E.R., (2018), A survey of Approaches to Diacritic Restoration, Natural Language Engineering 1 (1): 1 – 23, 2018.

Ezeani I. M., (2019), Corpus-Based Approaches to Igbo Diacritic Restoration. PhD Thesis, The University of Sheffield, 2019.

Jehovah's Witness (2022), jw.org/nso.

Lombard D.P,(1985). Introduction to the Grammar of Northern Sotho, J.L Schaik, 1985.

Louwrens L. J (1994), Dictionary of Northern Sotho Grammatical Terms, Via Afrika, 1994.

Malema G. and MotlhankaM,Okgetheng B., Motlogelwa N.P, Rammidi G., (2018), Rule Based Setswana Diacritic Restoration, The 20th International Conference on Linguistics and Languages, Nov 15-16, 2018, pp. 1059 – 1062.

MihalceaR.,(2002), Diacritics Restoration: Learning from Letters Versus Learning from Words, Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), 2002, pp. 339 – 348.

Poulos G,(1994). A linguistic Analysis of Northern Sotho, Via Afrika, 1994.

SchryverG. and Mogodi M., (2009). Oxford bilingual school dictionary: Northern Sotho and English.

Shaikh H, Mahar J.A, Mahar M. H, (2017), Instant Diacritics Restoration System for Sindhi Accent Prediction using N-Gram and Memory-Based Learning, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 8 No.4, 2017.

Stankevicius L, LukoševiciusM., DzikienJ.K.,Briediene M.,Krilaviˇcius T. (2022), Correcting Diacritics and Typos with a ByT5 Transformer Model, Applied Sciences, 2022,12,2636.

Ungurean C., Urileanu D., Popescu V.,Nfegrescu C.,DervisA.,(2008). Automatic Diacritic RestorationFor A TTS-Based E-Mail Reader Application, U.P.B. Sci. Bull., Series C, Vol. 70, Iss. 4, 2008.