The Lemmatisation of Nouns in Tshivenda Dictionaries

Mashudu Nthambeleni (mashudu.nthambeleni@univen.ac.za) and Nelson Mbulaheni Musehane (musehanen@univen.ac.za)
M.E.R. Mathutha Centre for African Languages, Arts and Culture, School of Human and Social Sciences, University of Venda, Thohoyandou, Limpopo, South Africa

Abstract: Tshivenda is one of the official languages of South Africa, mostly spoken in Limpopo Province at Vhembe District. From 1994, when the Republic of South Africa became a democratic government, speakers of the Tshivenda language spread to all nine provinces of South Africa. Tshivenda has different word categories, like nouns, verbs, adjectives and many more. When words are looked up in dictionaries, one should know what types of categories are given, e.g. their spelling, pronunciation and meaning. If this is not clearly represented one would not be able to use such a dictionary. This paper seeks to investigate how nouns are lemmatised in Tshivenda dictionaries. Attention is given to word and stem lemmatisation. It also looks at the lemmatisation of singular and plural nouns, and the lemmatisation of deverbal and diminutive nouns. This will be accomplished by analysing published Tshivenda dictionaries in circulation.

Keywords: LEMMATISATION, NOUNS, TSHIVENDA, DEVERBATIVE NOUNS, DIMINUTIVE NOUNS, SINGULAR, PLURAL

1. Introduction

If dictionaries are not compiled according to the required acceptable standard,
users of such dictionaries would find it difficult to use them. In other words, such dictionaries will not be user-friendly. There should be proper planning to determine what information about spelling, pronunciation as well as meaning should be included. This process is termed lemmatisation. Hartmann and James (1998: 83) define lemmatisation as "the reduction of a paradigm of variant word forms to a canonical form; e.g. the inflected forms of English verbs to the infinitive." They add that lemmatisation is the process by which the compilers of dictionaries establish the canonical forms of headwords either by removing or retaining inflections.

On the other hand, Plisson et al. (2005: 369) define lemmatisation as "the process of finding the normalized form of words" as their arrangement into alphabetical order whereas Sinclair (1991: 173) defines lemmatisation as "the process of gathering word-forms and arranging them into lemma or lemmata".

From these definitions, it can be deduced that lemmatisation is the selection of words or data to be included in a dictionary. According to Gove (1961: Preface 4a), the lemmatisation of words is determined by the degree to which they are most likely to be looked for.

A noun in Tshivenda is dzina "name". According to the Macmillan English Dictionary for Advanced Learners (2002: 967) a noun is a word or group of words used for referring to a person, thing, place or quantity whereas the Chambers-Macmillan South African Student’s Dictionary (1996: 639) defines a noun as a word used to refer to a person, thing or quality. By that is meant that it can refer to anything that can occupy a space and have weight as well as some abstract qualities.

Several wordlists and dictionaries have been published in Tshivenda, e.g. those by Marole (1955), Marole and De Gama (1958), Joubert and Rapea (1972), Wentzel and Muloiwa (1982), Hartshorne (1984), Van Warmelo (1989) and that by the Tshivenda National Lexicography Unit, Tshivenda/English Thalusamaipfi/Dictionary (henceforth abbreviated as TETD) (2006). This article will analyse the lemmatisation of nouns, particularly in Wentzel and Muloiwa (1982), Van Warmelo (1989) and TETD (2006).

When the current dictionaries of South African languages, including Tshivenda, are investigated, it is found that nouns are lemmatised principally in two ways:

(a) using the whole word (noun), and
(b) using the noun stems.

2. Lemmatisation of nouns

Lemmatising the whole word in languages such as Tshivenda, Sesotho sa Leboa and Xitsonga has become tradition. Lemmatising according to the noun stem is found in the Nguni languages, where it originated, although it also occurs in Sesotho sa Leboa.
2.1 Lemmatisation of the whole word

From the following examples, it can be seen that words such as *musadzi* in Tshivenda, *mosadi* in Sesotho sa Leboa and *wansati* in Xitsonga have been lemmatised as whole words.


(b) Sesotho sa Leboa: the word *mosadi* 'woman' in Kriel and Van Wyk (1983: 172)

(c) Xitsonga: the word *wansati* 'woman' in Cuenod (1967: 224)

Every language has its own way of writing words. In other words, Tshivenda, Sesotho sa Leboa and Xitsonga have a disjunctive way of writing, as shown in the following sentences, e.g.

Tshivenda: *Musadzi u amba na n'wana* 'The woman talks to the child'

Sesotho sa Leboa: *Mosadi o bolela le ngwana* 'The woman talks to the child'

Xitsonga: *Wansati u vulavula na n'wana* 'The woman talks to the child'

In the Nguni languages, a conjunctive way of writing is traditional, as shown in the following sentences, e.g.

IsiZulu: *Umfazi ukhuluma nomntwana* 'The woman talks to the child'

The preceding examples indicate that in Tshivenda, Sesotho sa Leboa and Xitsonga, there are five words in a sentence, including morphemes, whereas in the Nguni languages there are only three words. This is seen in Van Warmelo (1989: 235 and 289), where the verb stem *amba* 'talks' and nouns *musadzi* 'woman' and *n'wana* 'child' are lemmatised as individual words. This is also found in Wentzel and Muloika (1982: 7, 43, 58). In the preceding isiZulu example, however, the sentence is formed by only three words. All these words have been lemmatised as stems, as can be seen in Doke et al. (1971: 201 and 609) where the nouns *umfazi* and *nomntwana* are lemmatised as *-fazi* and *-ntwana* respectively.

The conjunctive way of writing results in lemmatisation using stems in the Nguni languages, whereas the disjunctive way of writing brings about lemmatisation of nouns using the whole word. The difference between the two ways of lemmatisation lies in the use of a hyphen before the stem, while the word can stand on its own.

Although a dictionary is not a grammar book, it includes the morphology of words because the dictionary cannot be used without some knowledge of their morphology. This is why users should be aware that at some point they would have to be able to find the stems of verbs, nouns and adjectives of the
specific language, so that they can utilise the dictionary. It is also because compilers of dictionaries use morphology when writing their dictionaries, which indicates the correct way of constructing words. This method provides the meaning of such words.

When the three dictionaries by Wentzel and Muloiwa (1982), Van Warmelo (1989), and TETD (2006) were examined, it was found that they have lemmatised the nouns using the complete word and not just the stems.

2.2 Lemmatisation of singular and plural nouns

Nouns can be lemmatised in singular or plural form. In Tshivenda, there are different class noun prefixes, which indicate the singular or plural. Such prefixes are affixed to the noun stem, as shown in the following examples:

(a) mushumi ‘worker’ < mu- + -shumi in Van Warmelo (1989: 237),

(b) musidzana ‘girl’ < mu- + -sidzana in Van Warmelo (1989: 237),

(c) vhashumi ‘workers’ < vha- + -shumi in Van Warmelo (1989),

(d) mushumi ‘worker’ < mu + -shumi in TETD (2006: 49),

(e) vhashumi ‘workers’ < vha- + -shumi in TETD (2006: 83),

(f) musidzana ‘girl’ < vha- + -sidzana in TETD (2006: 49),

(g) vhasidzana ‘girls’ < vha- + -sidzana in TETD (2006: 83),

(h) mushumi ‘worker’ < mu- + -shumi in Wentzel and Muloiwa (1982: 43), and


When the preceding examples are considered, it can be observed that Van Warmelo (1989) and Wentzel and Muloiwa (1982) have lemmatised the noun mushumi ‘worker’ and musidzana ‘girl’ in the singular only; whereas TETD (2006) has lemmatised both musidzana and mushumi in the singular and vhasidzana and vhashumi in the plural form. For learners who are not mother-tongue speakers of Tshivenda it will be problematic when they come across the nouns vhashumi ‘workers’ and vhasidzana ‘girls’ but cannot find them in the dictionaries.

TETD tried to solve this problem by lemmatising nouns in the singular and plural forms, as shown above. However, they did not lemmatise other Tshivenda nouns in their plural forms, e.g. mulilo ‘fire’, is lemmatised in TETD (2006: 48). However, its plural, mulilo ‘fires’, is not. Also, mulambo ‘river’, is lemmatised in TETD (2006: 47). However, its plural milambo ‘rivers’, is not.

The problem here is that there is no consistency in the procedure of the compilers of these dictionaries, because some words have been lemmatised in the singular and plural, and others in the singular only.
In Tshivenda, there are 21 noun class prefixes in the singular and plural forms. Among these prefixes, some are infinite and some locative class prefixes. The noun class prefixes are as follows:

<table>
<thead>
<tr>
<th>Class</th>
<th>Singular</th>
<th>Class</th>
<th>Singular</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>mu-</td>
<td>Class 2</td>
<td>vha-</td>
</tr>
<tr>
<td>Class 3</td>
<td>mu-</td>
<td>Class 4</td>
<td>mi-</td>
</tr>
<tr>
<td>Class 5</td>
<td>ǐ-</td>
<td>Class 6</td>
<td>ma-</td>
</tr>
<tr>
<td>Class 7</td>
<td>tshi-</td>
<td>Class 8</td>
<td>zwi-</td>
</tr>
<tr>
<td>Class 9</td>
<td>n-</td>
<td>Class 10</td>
<td>dzi-</td>
</tr>
<tr>
<td>Class 11</td>
<td>u-</td>
<td>Classes 10/14</td>
<td>dzi/-vhu-</td>
</tr>
<tr>
<td>Class 14</td>
<td>vhu-</td>
<td>Class 6</td>
<td>ma-</td>
</tr>
<tr>
<td>Class 15</td>
<td>u</td>
<td>Infinite noun class</td>
<td></td>
</tr>
<tr>
<td>Classes 16/17/18</td>
<td>fha/-ku/-mu-</td>
<td>Locative noun classes</td>
<td></td>
</tr>
<tr>
<td>Class 20</td>
<td>ku-</td>
<td>Class 8</td>
<td>zwi-</td>
</tr>
<tr>
<td>Class 21</td>
<td>ąda-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A noun in Tshivenda is formed by two morphemes; namely the noun class prefix plus the noun stem. The singular or plural prefix indicates the singular or plural form. In addition, every noun in Tshivenda is categorized under these classes, e.g.,

- **Singular** > **Plural**
  - musadzi ‘woman’ > vhasadzi ‘women’
  - muthu ‘person’ > vhathu ‘persons’
  - mulilo ‘fire’ > mililo ‘fires’

The examples above show that the noun musadzi falls under the prefix mu- in the singular and under the prefix vha- in the plural vhasadzi. The noun mulilo falls under the singular class prefix mu-, whereas it comes under the plural prefix mi- in mililo. Among these nouns there are instances where the prefix is not visible in the singular form whereas in the plural it is, e.g.,

<table>
<thead>
<tr>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>mme ‘mother’ class 1a</td>
<td>vhomme ‘mothers’ class 2b</td>
</tr>
<tr>
<td>khotsi ‘father’ class 1a</td>
<td>vhokhotsi ‘fathers’ class 2b</td>
</tr>
</tbody>
</table>

In the preceding examples, it may appear that there is no class prefix attached to the nouns. What is seen here are the stems mme and khotsi. To these stems, which are recognized nouns, the plural prefix vho- can be affixed to form the plural nouns vhomme and vhokhotsi. This prefix vho- indicates that the noun to which it is affixed is in the plural form. These words should be interpreted as two different words. Lemmatising nouns using a singular or a plural prefix started some time ago; for example, Van Warmelo (1989) and Wentzel and Muloiwa (1982). In Wentzel and Muloiwa (1982: 43) lemmatisation has been
done as follows:

\textit{musadzi (vha-)}

This shows that the noun \textit{musadzi} is in the singular form, and its plural form is indicated as \textit{vha-} in brackets. Van Warmelo (1989: 235) has not indicated the plural of the noun \textit{musadzi}. He has given it as follows in the singular form only:

\textit{musadzi 1 (Class 1)}

This merely shows that the noun \textit{musadzi} falls under the first class prefix of nouns, and is in the singular form. Not all nouns in the plural form are lemmatised in this way; others are represented as follows:

\textit{Mid\textlast{-\textacuten}ali} is the plural form of \textit{mu\textlast{-\textacuten}ali}

\textit{mililo} is the plural form of \textit{mulilo}

These nouns appear to have been lemmatised in Van Warmelo (1989: 196 and 197) in alphabetical order. However, \textit{mu\textlast{-\textacuten}ali} in the singular and \textit{mi\textlast{-\textacuten}ali} in the plural form have been omitted in TETD (2006). The difference between \textit{mulilo} and \textit{mililo} is brought about by the class prefix \textit{mu-} of the singular, meaning that the word refers to a single entity, while the plural class prefix \textit{mi-} indicates that the word refers to more than one thing. This way of lemmatisation on the basis of the singular and plural forms has advantages and disadvantages.

### 2.2.1 Advantages of the lemmatisation of singular and plural nouns

It is user-friendly when all nouns will be lemmatised in the singular and the plural. This has been emphasised by Prinsloo and De Schryver (1999: 267) as follows:

The major advantage of this lemmatization procedure is user-friendliness rendering a practical theory.

With this way of lemmatisation, the dictionary is easy to use, because no intensive knowledge of the language is expected from users. What is required is just knowledge of the alphabet of Tshivenda. When we look at the Tshivenda noun \textit{mushumi} 'worker/labourer', it is lemmatised in the singular by Van Warmelo (1989: 237). He has also lemmatised the plural form of \textit{mushumi} 'worker/labourer', which is \textit{vhashumi} 'workers/labourers'. This is a user-friendly format even for a person who is not Muvenda.

Wentzel and Muloiwa (1982: 43) also lemmatised the noun \textit{mushonga} 'medicine', and the plural for \textit{mushonga} 'medicine' is shown by the prefix \textit{mi-}. By doing this, he is indicating that when one wants to get the plural form of \textit{mushonga} 'medicine' one should just add the prefix \textit{mi-} to the noun stem -\textit{shonga} to make it \textit{mishonga} 'medicines'. This is a user-friendly approach.
2.2.2 Disadvantages of the lemmatisation of singular and plural nouns

There are also problems associated with the lemmatisation of both singular and plural nouns. Firstly, because plurality appears in all Tshivenda nouns, compilers of dictionaries will have to pose the question: Which nouns should be selected for lemmatisation? As a result compilers often fail to lemmatise important, frequently occurring nouns, thus rendering the dictionary ineffective to its users.

The following nouns have been lemmatised in the singular form in Van Warmelo (1989: 237 and 235):

- *musidzana* 'girl'
- *musadzi* 'woman'

However, the following plurals of these nouns have not been lemmatised:

- *vhasidzana* 'girls'
- *vhasadzi* 'women'

The dictionary seems to have been written only for users who are conversant in Tshivenda; that is users who would know that the plural of the noun *musidzana* is *vhasidzana*. By lemmatising the nouns in the singular only, the plurals of important nouns in Tshivenda are left unlemmatised.

In Wentzel and Muloiwa (1982: 42 and 44) the following nouns have been lemmatised in both the singular and plural, and the plural is indicated by *vha*- as shown below:

- *munna (vha-)*
- *musidzana (vha-)*

This means that the noun *musidzana* is the singular form, and to form the plural the prefix *vha-* must be added. This way of lemmatisation is very effective for those who know Tshivenda. However, users who do not know Tshivenda cannot consult the dictionary to check plurality. When they come across *vhasidzana*, they would not know that it is the plural of *musidzana*. When they come across the plural *vhasidzana*, instead of the singular *musidzana*, the users would be unaware that it is the plural of *musidzana*. It therefore renders the dictionary ineffective to users who are not conversant in Tshivenda. In this way, dictionaries omit important words in the plural.

2.3 Lemmatisation of deverbative nouns

Poulos (1990: 14) defines *deverbative nouns* as follows:

A number of nouns in some of the classes are derived from other parts of speech, for example, from verbal radicals (or verb roots).
In Tshivenda, the derivation of nouns from verb stems or adjective stems occurs by affixing singular and plural prefixes to these stems and suffixes or endings -i, -e, -o, -u.

### Nouns derived from verbs

<table>
<thead>
<tr>
<th>Nouns</th>
<th>derived from verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ndima</td>
<td>-lima 'to plough'</td>
</tr>
<tr>
<td>murengi</td>
<td>-renga 'to purchase'</td>
</tr>
<tr>
<td>mulimi</td>
<td>-lima 'to plough'</td>
</tr>
<tr>
<td>muumbi</td>
<td>-imba 'to sing'</td>
</tr>
<tr>
<td>tsibiiki</td>
<td>-bika 'to cook'</td>
</tr>
<tr>
<td>vhuhulu</td>
<td>-hula 'to be big'</td>
</tr>
<tr>
<td>vhusekene</td>
<td>-sekena 'to be thin'</td>
</tr>
<tr>
<td>vhulapfu</td>
<td>-lapfa 'to be long'</td>
</tr>
</tbody>
</table>

(Nthambeleni 2008: 49)

Other deverbatives formed from verbs with suffixes or endings added are the following:

- murungo 'seam'
- muhuno 'government'

Because they are problematic, the preceding nouns should have been lemmatised in the dictionary. Compilers should include deverbatives like these in their dictionaries.

In African languages, nominal prefixes are therefore important to such an extent that prefixes of class nouns and suffixes or endings such as -a, -e, -i, -o, and -u should be considered when deverbatives are discussed.

Nouns formed in this manner are not lemmatised in great numbers in dictionaries, as is the case in Tshivenda dictionaries. This results in ineffective dictionaries being produced, which disadvantages dictionary users. Lemmatisation of these words in the dictionary depends on the way speakers use them. When they are used in language, compilers of dictionaries should select them for lemmatisation.

Focusing again on the above-mentioned items, compilers of Tshivenda dictionaries have fulfilled an important function in lemmatising deverbatives. However, it is problematic when some deverbative nouns have not been lemmatised as in Van Warmelo (1989: 52). Here the verb stem -fha ‘to build’ has been given a deverbative noun by affixing the class prefix mu- to a verb which is not given. This shortcoming causes problems for non-native users of the dictionary. Such users will find the noun mufhati 'builder' but not know its meaning. This is also the case with the noun murengi 'buyer'. Although they are used frequently in Tshivenda, these nouns have not been lemmatised.

### 2.4 Lemmatisation of diminutive nouns

Poulos (1990: 81) defines diminutives as follows:
Those expressions in a language that basically express the idea of ‘smallness’, ‘shortness’ and where appropriate the ‘young of’ some or other noun.

As users would look up diminutives in the dictionary, it is important to lemmatise them. In Tshivenda, diminutives can be formed by affixing the different class prefixes and suffixes to the nouns, such as the suffixes -ana, -nyana; and class prefixes 7 tshi-, 11 lu-, 20 ku- in the singular and class 8 zwi- and 14 vhu- in the plural, as in the following examples (not from the dictionaries mentioned above).

Class 7 tshi- + -budzi > tshibudzi (small fat goat)
Class 20 ku- + -budzi > kubudzi (small young goat)

Each of these nouns tshibudzi and kubudzi has a different meaning. They are therefore lemmatised in dictionaries.

When the suffix -ana is added to nouns, it forms diminutive nouns

mbudzi + -ana > mbudzana

The suffix -nyana also forms diminutives and, when added to nouns, it results in the following:

(a) When it is added to nouns it may refer to a liquid which is little in quantity.
   (i) madì + -nyana > maðinyana (little water)
   (ii) mafhi + -nyana > mafhinyana (little milk)
   (iii) malofha + -nyana > malofhinyana (little blood)

(b) When it is used with a class prefix and a verb stem, it may indicate that there is not much of something.
   (i) mu + -shum + -o + -nyana > mushumonyana (little work)

(c) When it is added to nouns, it may indicate smallness in size.
   (i) musidzana + -nyana > musidzanyana (small girl)

The noun prefix plus suffix function together when added to the noun stem to form a diminutive noun, as in the following examples:

  tshi- + -budz- + -ana > tshibudzana (small goat)
  zwi- + -budz- + -ana > zwibudzana (small goats)
  vhu- + -budz- + -ana > vhubudzana (thin goats)

From the examples shown above, it becomes evident that, just like other nouns, diminutives are important in a language. In addition, in a language like Tshivenda, they are used frequently.
When the different dictionaries in Tshivenda are examined, one will find that they have lemmatised diminutive nouns. The noun mbudzi (goat) has been lemmatised, whereas its diminutive form tshibudzana (small goat) has not been listed in Van Warmelo (1989), although it is used frequently. Omitting it from the dictionaries is problematic for the users who do not know Tshivenda because they would not be able to find the word tshibudzana and its meaning. When nouns are looked up in Tshivenda dictionaries such as Wentzel and Muloiwa (1982), Van Warmelo (1989) and TETD (2006), it becomes clear that several diminutive nouns have not been lemmatised. Compilers should ensure that all such nouns are lemmatised. Only when this is done, will good dictionaries be compiled that will satisfy users' needs.

3. Conclusion

When the current Tshivenda dictionaries are examined, it becomes clear that all nouns should be lemmatised. The main aim of lemmatisation of nouns is to enable compilers of dictionaries to produce useful dictionaries, and not repeat the same mistakes made in previous dictionaries. For the production of useful dictionaries that will benefit users, nouns should be lemmatised in a well-considered manner. This will also support De Schryver and Prinsloo’s (2000) idea of ‘Simultaneous Feedback’, since a dictionary should be written in such a way that users can find information easily. It has been established that Tshivenda nouns are lemmatised in both the singular and plural forms; as diminutives and deverbatives. Including nouns in singular and plural form, as well as diminutive and deverbative nouns, should be formulated as a lexicographic principle. Failure to meet this requirement would render the dictionary not user-friendly at all.

References

Mashudu Nthambeleni and Nelson Mbuheleni Musehane


