Syllable analysis of North Sotho and its effect on computerized hyphenation

C.S. Price and Q.H. Gee

The need to address multiple languages in word processing is becoming increasingly important. Most word processors have some hyphenation facility which will split words occurring at the end of a line into two subwords. However, most word processors' hyphenation facilities cater only for American English, while some cater for British English, and a few for South African English. This means that if one is to process text written in a language other than English, the automatic hyphenation will produce incorrect and maybe even nonsensical subwords. In this case, the hyphenation facility will have to be disabled.

Some work has been performed on the hyphenation of words: Ocker's work is one of the many done on the English language.1 Even though English has many exceptions, an algorithm has been developed to find all possible hyphenation points without much semantic and syntactic knowledge. Mañas2 has formulated an algorithm for the hyphenation (and syllabication) of Spanish words: this algorithm needs only morphological knowledge to determine the hyphenation points. Other work done for languages spoken in South Africa include the hyphenation of Afrikaans by Gee3 and Zulu by Hampton.4

This article presents an algorithm for the hyphenation and syllabication of North Sotho words. It does not require any syntactic or semantic knowledge of the word to be hyphenated. The hyphenation and syllabication of the words are very similar, and their differences are described. The algorithm finds all possible hyphenation points of a word, but does not choose a particular hyphenation point. This will be done by the calling procedure, which is probably part of the word processor package. The results from an implementation of the algorithm are also presented.

Introduction to North Sotho

North Sotho (Sepedi or Pedi) is a language which is spoken predominantly in the northern Transvaal. The work of transcribing the language took place between the mid and late 19th century, with the first reader appearing in 1870, the first excerpts of the New Testament in 1890 and the complete Bible in 1904. The language is written with Roman characters and has undergone three major changes to its orthography (the way the language is written). These three changes are marked by the periods 1870–1951, 1951–1985 and 1985 to the present day, respectively. Most published material in schools and bookshops are typed using the second orthography. This orthography is therefore the one considered here. It is conjectured that the use of other orthographies does not pose problems for the hyphenation and syllabication algorithms presented.

The orthography chosen uses the same characters to represent vowels and consonants as English does. It has an additional consonant, viz. ñ, which is pronounced as the English 'sh', but the diacritic mark makes no difference to the phonetic value of the word when considering it for hyphenation. In literary texts, the two vowels ë and ë can be found in addition, but the circumflex does not change the phonetic value of the word, and the accent can therefore be omitted for our purposes.

Because North Sotho was transcribed recently (in comparison to English), a closer mapping has been achieved between the syllables, sounds and morphological components, and the orthographic characters which represent them. This makes the rules for hyphenation and syllabication much simpler than the rules for English. Rules for the hyphenation of Spanish words are simpler than those of English, but they are more complex than those for North Sotho.

Rules for hyphenation and syllabication

By hyphenation we mean the breaking of written words so that the resulting subwords consist of complete syllables. The syllable refers to acceptable sound sequences of consonants and vowels, together with features such as length and stress, or to single consonants or vowels, which are suitably considered as a group for analysis. For our purposes syllabication means the decomposition of words into these syllables. Thus the difference is that hyphenated subwords always have at least one written vowel per syllable, whereas with syllabication the minimum component can be a single vowel or a consonant cluster.

North Sotho hyphenation and syllabication are identical except where certain consonant combinations occur at the beginning or at the end of a word. In what follows below, 'hyphenation' means both hyphenation and syllabication, unless otherwise stated.

Consonant clusters

A consonant cluster consists of one or more consonants. All consonants forming such a cluster are pronounced as one sound. Hence, even though there are two, three or four consonants consecutively in a word, they can never be separated by hyphens. A table showing some of the most common consonant clusters with examples of words containing these clusters is shown below. Clusters with only one consonant are not shown.

<table>
<thead>
<tr>
<th>Consonant cluster</th>
<th>Example of word</th>
</tr>
</thead>
<tbody>
<tr>
<td>bj</td>
<td>bjalo, bja, bj</td>
</tr>
<tr>
<td>fs</td>
<td>bofsa, kgaušī</td>
</tr>
<tr>
<td>ft</td>
<td>fla</td>
</tr>
<tr>
<td>gw</td>
<td>gagwe</td>
</tr>
<tr>
<td>hl</td>
<td>Moahlodi, kaholo, mehla</td>
</tr>
<tr>
<td>hw</td>
<td>hwetše</td>
</tr>
<tr>
<td>kg</td>
<td>Mokgetsha, kgopo, kgalefo</td>
</tr>
<tr>
<td>kgw</td>
<td>nkxwana, mokgw</td>
</tr>
<tr>
<td>kh</td>
<td>khunollo, khoršag, khutsong</td>
</tr>
<tr>
<td>kr</td>
<td>disakranente</td>
</tr>
<tr>
<td>kw</td>
<td>kwang, kwa, kwena</td>
</tr>
<tr>
<td>ll</td>
<td>khunollo</td>
</tr>
<tr>
<td>lw</td>
<td>ngwadiwlego, bolelwa</td>
</tr>
<tr>
<td>mm</td>
<td>mma, mm</td>
</tr>
<tr>
<td>nd</td>
<td>maikondijo</td>
</tr>
<tr>
<td>ng</td>
<td>kerekeng, melaong, seng</td>
</tr>
<tr>
<td>ngw</td>
<td>ngwadiwlego, ngwana</td>
</tr>
<tr>
<td>nw</td>
<td>nwa</td>
</tr>
<tr>
<td>ny</td>
<td>Bosenyi, tshewneylego, nyatše</td>
</tr>
<tr>
<td>nyw</td>
<td>tsenywango, enywago</td>
</tr>
<tr>
<td>ph</td>
<td>diphetogo, phele, mphetoša, hophelong</td>
</tr>
<tr>
<td>pi</td>
<td>eupilé</td>
</tr>
<tr>
<td>psh</td>
<td>mphafasha, mphafatšwa</td>
</tr>
<tr>
<td>py</td>
<td>Mpye</td>
</tr>
<tr>
<td>rw</td>
<td>rwa, feberawar</td>
</tr>
<tr>
<td>sw</td>
<td>swanetse</td>
</tr>
<tr>
<td>šw</td>
<td>šlišwe</td>
</tr>
<tr>
<td>th</td>
<td>mpetha, motho</td>
</tr>
<tr>
<td>tl</td>
<td>šlišwe, šlogetla, tla, seelegilego</td>
</tr>
<tr>
<td>tlw</td>
<td>mbelitwa</td>
</tr>
<tr>
<td>ts</td>
<td>tsenywango, tsēba</td>
</tr>
<tr>
<td>tsh</td>
<td>boitsholo</td>
</tr>
<tr>
<td>tshw</td>
<td>tshewneylego</td>
</tr>
<tr>
<td>tsw</td>
<td>tsawalela</td>
</tr>
<tr>
<td>tl</td>
<td>nyatše, tla, beetlego, letšatsi</td>
</tr>
<tr>
<td>thh</td>
<td>nthuswe</td>
</tr>
<tr>
<td>tlw</td>
<td>mphafatšwa</td>
</tr>
</tbody>
</table>

The treatment of vowels

The hyphenation and syllabication of North Sotho words revolve mainly around the vowels. When a vowel cluster is followed by a consonant cluster, a hyphen is always inserted between them. However, an exception may occur in words ending in ng, depending on whether hyphenation or syllabication is required. The ng at the end of a word is pronounced as a separate syllable, so if syllabication is required, the above rule does not need to be modified. If hyphenation is required, there should be no hyphenation point between the root of the word and the ng as it is not considered
good practice to allow such a word to be hyphenated before the ng. This is the first differentiation made between hyphenation and syllabication.

Prefixes and suffixes

North Sotho is a language which does not have multiple prefixes and suffixes. It is useful to recognize and remove the suffixes and prefixes (affixes) from the root of the word when hyphenating for two reasons: first, when the affix is attached to the root of the word, the hyphenation rules incorrectly perform the hyphenation. Secondly, there is always a hyphenation point between the affixes and the root of the word, so the affix and the root can be considered to be different entities to be hyphenated.

Prefixes

Most often, prefixes in North Sotho indicate singular or plural words. Other types of prefix exist, which could be classed as part of a consonant cluster, except for the fact that they are pronounced as two syllables. These are

- m followed by p, e.g. m-pha-fa-tsa
- n followed by k, n or t, e.g. n-te-ka

However, it is not considered good practice to separate the m or n from the rest of the word in hyphenating, even though it is syllabically correct. If syllabication is required, a hyphen is inserted after the m or n. The prefixes are shown in the table below. Our notation is that _ represents a blank space, i.e. the start of a word, and that - represents a valid hyphenation point. Upper and lower cases are not significant.

<table>
<thead>
<tr>
<th>Prefixes (Singular/plural prefixes)</th>
<th>Followed by</th>
<th>Hyphenation</th>
<th>Syllabication</th>
</tr>
</thead>
<tbody>
<tr>
<td>_bo</td>
<td>_bo</td>
<td>_bo</td>
<td>_bo</td>
</tr>
<tr>
<td>_le</td>
<td>_le</td>
<td>_le</td>
<td>_le</td>
</tr>
<tr>
<td>_mo</td>
<td>_mo</td>
<td>_mo</td>
<td>_mo</td>
</tr>
<tr>
<td>_se</td>
<td>_se</td>
<td>_se</td>
<td>_se</td>
</tr>
<tr>
<td>_ba</td>
<td>_ba</td>
<td>_ba</td>
<td>_ba</td>
</tr>
<tr>
<td>_di</td>
<td>_di</td>
<td>_di</td>
<td>_di</td>
</tr>
<tr>
<td>_ma</td>
<td>_ma</td>
<td>_ma</td>
<td>_ma</td>
</tr>
<tr>
<td>(Other prefixes)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>_m p</td>
<td>_mp</td>
<td>_m-p</td>
<td>_m-p</td>
</tr>
<tr>
<td>_n k</td>
<td>_nk</td>
<td>_n-k</td>
<td>_n-k</td>
</tr>
<tr>
<td>_n n</td>
<td>_nn</td>
<td>_n-n</td>
<td>_n-n</td>
</tr>
<tr>
<td>_n t</td>
<td>_nt</td>
<td>_n-t</td>
<td>_n-t</td>
</tr>
</tbody>
</table>

Suffixes

There are several suffixes in North Sotho which mostly denote the tense of the verb. An extension to the rule of hyphenating before the suffix -ng is: if two vowels precede the ng, then a hyphen is placed between them. The common suffixes are shown in the table below. Note that _ again represents the blank space following the end of a word. Upper and lower cases are not significant. V denotes any vowel.

<table>
<thead>
<tr>
<th>Suffixes</th>
<th>Hyphenation</th>
<th>Syllabication</th>
</tr>
</thead>
<tbody>
<tr>
<td>ela_</td>
<td><em>-e-la</em></td>
<td><em>-e-la</em></td>
</tr>
<tr>
<td>elsgo_</td>
<td><em>-els-go</em></td>
<td><em>-els-go</em></td>
</tr>
<tr>
<td>elgo_</td>
<td><em>-e-lg</em></td>
<td><em>-e-lg</em></td>
</tr>
<tr>
<td>ilwego_</td>
<td><em>-il-we-go</em></td>
<td><em>-il-we-go</em></td>
</tr>
<tr>
<td>ng_</td>
<td><em>-ng</em></td>
<td><em>-ng</em></td>
</tr>
<tr>
<td>olola_</td>
<td><em>-ol-ola</em></td>
<td><em>-ol-ola</em></td>
</tr>
<tr>
<td>wa_</td>
<td><em>wa</em></td>
<td><em>wa</em></td>
</tr>
<tr>
<td>V-ng_</td>
<td><em>V-ng</em></td>
<td><em>V-ng</em></td>
</tr>
</tbody>
</table>

Double vowels

If there are two consecutive vowels in a word, and if the above rules have not dealt with them, then two further cases must be tested. If these vowel pairs are ai, ee, ei, eu, ii or oo, they should not be separated by a hyphen. If they do not fall under this category, they should be separated. A list of vowel clusters and vowel pairs is shown below.

Non-separable vowel pair Example of word
ai | seeswai, seswaing, letswai
ee | itseela, gotse, meeset
ei | eite, leineng
eu | euple
ii | Titsele
oo | diphoofolo

Separable vowel pair Example of word
ae | tado, Istrael
ao | moloa, moato, laola
au | Dikaguelo
ea | seatleng
eo | yoo, Lenanoe
ia | batiakane
oa | Moahloidi
oi | koloi

Figure 1 summarises these rules. Note that syllabication is an extended form of hyphenation: it differs only in Rules 2a, 7 and 8.

Exceptions

Two exceptions to the above rules have been found so far. They are:

1) lefeela: here the ee should remain unseparated as ela here is not a suffix. lefeela.
2) maatla: here the ma does not denote a plural word; hence the aa should not be separated. maatla.

<table>
<thead>
<tr>
<th>Rule No.</th>
<th>Hyphenation</th>
<th>Syllabication</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CC-CC</td>
<td>CC-CC</td>
</tr>
<tr>
<td>2a</td>
<td>Vng- <em>-Vng</em></td>
<td>Vng- <em>-V-ng</em></td>
</tr>
<tr>
<td>2</td>
<td>VC- _-C</td>
<td>VC- _-C</td>
</tr>
<tr>
<td>3</td>
<td>VS- _-S</td>
<td>VS- _-S</td>
</tr>
<tr>
<td>4</td>
<td>VVng- <em>-V-Vng</em></td>
<td>P(n)-V- P(n)-V</td>
</tr>
<tr>
<td>5</td>
<td>P(n)-C- P(n)-C</td>
<td>mcag - mcag</td>
</tr>
<tr>
<td>6</td>
<td>_-mcag- <em>-mcag</em></td>
<td>nC/vn- <em>-nC/vn</em></td>
</tr>
<tr>
<td>7</td>
<td>V(vp)vp(V)vp _-V(vp)V</td>
<td>V(p)-vp(V)-vp(V)-vp(V)</td>
</tr>
<tr>
<td>8</td>
<td>VV- _-V-V</td>
<td>VV- _-V-V</td>
</tr>
</tbody>
</table>

| Where | | P(n) is a prefix of the singular/plural type
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>C is a consonant cluster beginning with p</td>
</tr>
<tr>
<td>C_vn</td>
<td>C_vn is a consonant cluster beginning with k, n or t</td>
</tr>
<tr>
<td>S</td>
<td>V is one of the vowels in an inseparable vowel pair</td>
</tr>
</tbody>
</table>

Fig. 1. Summary of hyphenation and syllabication rules.
lignonked as wat tans ontsiikel word, kan deur onopgedeelde persone persiel word. Die herstelwerk word dan beperk tot die vervanging van 'n gedrukte stroombaan (koste ongeveer R15) of die vervanging van die meetkabel met die ligtransistors. Die gebruiksuur van die apparaat is fettlik onbeprek en dit behoort eikele jare te hou.

Vervanging van battery. Met die nuwe model kan die battery van die kant af ingedruk word. 'n Blinde persoon kan die battery sonder hulp van 'n siende persoon vervang.

Die personeel van die Technikon Pretoria vervaardig dan die apparaat kosteloos vir die Nasionale Raad vir Blinde. Daar word diesels verwyd dat die Raad vir die komponente sal betaal. Die herstel en instandhouding van onklaar dra so wys ook gratis gedaan. Plaaslike maatskappe kan gerus die projek borg sodat die apparaat kosteloos aan blinde beskikbaar gestel kan word.

Hoe werk die apparaat?
Die uitsig van die ligsenistiewe transistors word eers met behulp van bewerkingsver- sterkers (OP-AMP’s) versterk. Die uitsit van die twee verstekers word deur 'n derde bewerkingsversterker met mekaar vergelyk. Die derde versterker bekrug 'n oscillator waar sleg in werkende tree indien 'n ligsein ontvang word.

Die apparaat is afhanklik van die verskil in lig en nie op die meting van ligsterte nie. Die huidige apparaat kan direk na die son gerig word sonder enige oudiose, maar dit kan enige ligdiode se lig waarneem. Tydens praktiese toets is die diodes se uitstalling vermind onder sonder dat enige probleme ontstaan het.