

Evaluating the Sesotho rule-based syllabification system on Sepedi and Setswana words

Johannes Sibeko

Nelson Mandela University
Linguistics and Applied Linguistics
Port Elizabeth, South Africa
johannes.sibeko@mandela.ac.za

Mmasibidi Setaka

South African Centre
for Digital Language Resources
Potchefstroom, South Africa
mmasibidi.setaka@nwu.ac.za

Abstract

The purpose of this article is to demonstrate that the recently developed automated rule-based syllabification system for Sesotho can be used broadly across the officially recognised South African Sotho-Tswana language group encompassing Sepedi, Sesotho and Setswana. We evaluate the automatic syllabification system on 400 words comprising 100 most frequently used words and 100 least-used words in Sepedi and Setswana as evident in the Autshumato corpus publicly available online. It is found that the Sesotho rule-based syllabification system can be used to correctly identify vowel-only syllables, consonant-vowel syllables and consonant-only syllables in Sepedi and Setswana. Among other findings, it has been demonstrated that words with diacritics as in the case of Sepedi are correctly broken down into syllables. We make two main recommendations. First, the rules for syllabification should be updated so that Sepedi diacritics are accommodated. Second, the syllabification system should be updated so that it reflects the broader Sotho-Tswana language group instead of being limited to Sesotho. Further research is needed to ascertain whether the complex consonant [ɲ] behaves similarly in all three officially recognised Sotho-Tswana languages and evaluate the need for a specific rule for the [ɲ] nasal consonant.

1 Introduction

This article presents an evaluation of the Sesotho rule-based syllabification system for use in Sepedi and Setswana written texts. The Republic of South Africa recognises eleven official languages, namely, Afrikaans, English, isiNdebele, isiXhosa, isiZulu, SiSwati, Sepedi, Sesotho, Setswana, Tshivenda, and Xitsonga (Republic of South Africa, 1996). Of these eleven languages, Afrikaans and English are often identified as Germanic languages (Zulu et al., 2008). Even so, some argue that Afrikaans is an African language as it is spoken in Africa even

though it is a variant of Dutch (Willemse, 2018; Staphorst, 2022). IsiNdebele, isiXhosa, isiZulu and isiSwati are classified under the Nguni language group. Sepedi [also referred to as Northern Sotho (Rakgogo and Zungu, 2021)], Sesotho [also referred to as Southern Sotho (Demuth, 2007)], and Setswana [sometimes referred to as Western Sotho (Mojela, 2016)] are grouped under the Sotho-Tswana language group. These Sotho-Tswana languages have many variations within themselves. Even so, they are mutually intelligible in that the speakers of these languages can understand each other without difficulty (Makalela, 2009). The national language bodies, under the auspices of the Pan South African Language Board (PanSALB) have dictated three different orthographies for these languages. PanSALB develops rules and standards for spelling and orthography for the proper functioning of all official languages¹.

According to the 2011 census², there were at least 4 618 577 Sepedi first language speakers, 3 849 562 Sesotho first language speakers and 4 067 248 Setswana first language speakers in South Africa alone (Lehohla, 2011). Sesotho is also an official language in Lesotho and Zimbabwe. Setswana is also one of the official languages of Botswana. The South African Human Language Technologies surveys indicate that there is a paucity of research in syllabification systems for the indigenous languages of South Africa (Grover et al., 2010, 2011; Barnard et al., 2014; Moors et al., 2018a,b). As far as we are aware, Sesotho is presently the only South African indigenous language to have a publicly accessible rule-based syllabification system. Even so, according to Sibeko and Setaka (2022), there are two systems for syllabification in Sesotho. The machine learning T_EX-based pattern-

¹see mandates for South African language boards accessible at https://static.pmg.org.za/PanSALB_APP_2122_compressed_reviewed_08032021_final.pdf

²South Africa's 2022 census results have not been released

ing system relies on gold-standard corpora with exemplified syllable annotations. The Sesotho T_EX-based syllabification system used a gold standard list of words and their syllables³. Unfortunately, we are not aware of similar lists for Sepedi and Setswana. As such, we cannot evaluate this machine learning system in this article.

This article aims to demonstrate the applicability of the Sesotho rule-based system in identifying syllables in the wider South African Sotho-Tswana language group. We provide a very brief background to orthographies of South African languages and expected syllable types in Sepedi and Setswana in section 2. We then describe our method of data collection in section 3. Finally, we discuss our findings in section 4 and end with a discussion in section 5.

2 Background

2.1 Orthographies

Written languages have systematic rules for spelling words (Matlosa, 2017). Rules for writing in African languages were introduced by religious missionaries. As a result, most orthographies of African languages are modelled after European orthographies (Mahlangu, 2015). Two major writing systems are used in South African languages. First, the Sotho-Tswana language group together with Tshivenda and Xitsonga use disjunctive writing systems, while the Nguni language group composing isiZulu, isiXhosa, isiNdebele, and SiSwati, use a conjunctive writing system (Prinsloo, 2011). The writing systems are illustrated below:

- (a) Bana ba ya matha. - Sesotho
Children they are running. - English
- (b) Abantwana bayagijima. - Isizulu
The+children they+are+running. - English

As exemplified above, languages such as Sotho-Tswana languages with disjunctive writing systems isolate words while languages with conjunctive writing systems [for example, isiZulu] combine some parts of speech such the present tense and the plural subject marker in the word *bayagijima*. Nonetheless, Sotho-Tswana languages are easy to compare as they follow the same phonemic orthography where seven phonemic representations are mutually common [a, e, i, o, u, ε, and ɔ] (Dickens, 1978; Matlosa, 2017).

³The Sesotho syllable information annotated wordlist can be freely accessed at <https://repo.sadilar.org/handle/20.500.12185/556>

Sesotho has two recognised orthographies, namely, (i) the South African Sesotho (SAS) orthography and (ii) the Lesotho Sesotho (LS) orthography. One of the main differences between SAS and LS orthographies is the use of diacritics in LS orthography. For instance, see examples (c) and (d) below:

- (c) Tshepiso. - SAS orthography
Promise. - English
- (d) Tšepiso. - LS orthography
Promise. - English

In the example above, the LS orthography uses the accented š letter while the SAS orthography uses the ‘sh’ digraph. Differences between SAS and LS orthographies are discussed thoroughly in studies such as Demuth (2007) and Matlosa (2017). The rule-based syllabification system evaluated in this article uses the SAS orthography (Sibeko and van Zaanen, 2022a).

According to Suyanto et al. (2021), rule-based systems perform better for low-resourced languages with limited to no gold standard corpora. This was also demonstrated in the case of Sesotho where the rule-based system outperformed the T_EX-based patterning system (Sibeko and van Zaanen, 2022a). Rule-based systems are based on sets of rules carefully designed by language experts. According to Sibeko (2022), when the designed list of rules is properly implemented, syllable boundaries can be identified in any Sesotho word.

The Sesotho rule-based system for syllabification uses rules for vowel-only syllables (v-syllables), consonant-only syllables (c-syllables), and consonant-vowel syllables (cv-syllables). According to Sibeko (2022) and Guma (1982) these are the only syllable types in Sesotho. Sibeko and Van Zaanen’s (2022b) rule-based syllabification system takes one word per line as input and then outputs the syllabified version. The syllable boundaries are indicated by spaces.

The rule-based syllabification system recognises only 26 letters of the alphabet [abcdefghijklmnopqrstuvwxy]. The rules for syllabification identify three main syllable types together with sixteen subtypes. These types, subtypes, and examples are presented in table 1. Interestingly, the Sesotho rule-based syllabification system used in this article demonstrated exceptional accuracy by achieving a rate of 99.6% (Sibeko and van Zaanen, 2022a,b).

The sounds w and y are considered semivowels

Type	Sub-types	Input	Syllabified	English
V	word-initial vowel	<i>oma</i>	<i>o-ma</i>	dry
	consecutive vowels	<i>boena</i>	<i>bo-e-na</i>	brotherhood
	word-final vowel	<i>lemao</i>	<i>le-ma-o</i>	needle
CV	one consonant - one vowel	<i>nama</i>	<i>na-ma</i>	meat
	one consonant - semi-vowel- one vowel	<i>lwetse</i>	<i>lwe-tse</i>	september
	two consonants - one vowel	<i>tlola</i>	<i>tlo-la</i>	skip
	two consonants - semi-vowel- one vowel	<i>shwashwi</i>	<i>shwa-shwi</i>	gossiper
	three consonants - one vowel	<i>tlhapa</i>	<i>tlha-pa</i>	insult
	three consonants - semi-vowel- one vowel	<i>tshweu</i>	<i>tshwe-u</i>	white
C	nasal consonant n, m - non-nasal consonant	<i>ntja</i>	<i>n-tja</i>	dog
	nasal consonant n, m - nasal consonant	<i>mmoho</i>	<i>m-mo-ho</i>	together
	nasal consonant n - complex nasal consonant	<i>nngwe</i>	<i>n-ngwe</i>	one
	complex nasal consonant ŋ - vowel	<i>ngola</i>	<i>ngo-la</i>	write
	complex nasal consonant ŋ- non-nasal consonant	<i>hanghang</i>	<i>ha-ng-ha-ng</i>	immediately
	word-ending complex nasal consonant ŋ	<i>mang</i>	<i>ma-ng</i>	who
	consecutive lateral consonants l	<i>lla</i>	<i>l-la</i>	cry

Table 1: Syllabification rules and examples.

when they occur at the onset of a syllable. However, some studies, such as Nkolola-Wakumelo et al.'s (2012) analysis of Setswana and Sesotho syllables, use the term "glides" instead.

2.2 Sepedi syllables

According to Wilsenach (2019), Sotho-Tswana languages have similar syllable structures. That is, Sepedi words can also be broken down into v-syllables, cv-syllables and c-syllables. The v-syllables are formed using only one vowel either at the beginning, middle or end of a word, or by monosyllabic words formed only of a vowel. The cv-syllable structure can contain between one and four onsets. The four onsets can be composed of three consonants (ccc) as in words like *tlhekišo* [tlheki-šo], and a semi-vowel (w) resulting in the four onsets and a vowel syllable (cccwv) as in words like *tlhwekišo* [tlhweki-šo].

The c-syllables can be formed m, n, l, r, ɲ, and ŋ syllabic consonants (Chokoe, 2020). First, c-syllables are formed when two identical syllabic consonants occur in succession within a single word, for instance in words like *ba-l-li* 'criers', and *wa-r-ra* 'brother' (Chokoe, 2020). Second, when nasal consonants precede any other consonant, for instance in words like *n-tšha* 'draw', and *m-phsa* 'new'. Third, the ŋ c-syllable is formed when the ŋ complex nasal takes the word-final position such as in words like *n-tlo-ng-* (Makaure, 2021; Chokoe, 2020).

2.3 Setswana syllables

There are also three syllable structures in Setswana Otlogetswe (2017). First, Setswana uses the open cv-syllable structure where cv-syllables can contain between one (cv) and four (cccwv) onset consonants (Sebina, 2014). Second, v-syllables can be formed using one vowel either at the word-initial, word-medial or word-final positions such in words like *a-lo-la*, *lo-e-to*, *bo-e-* 'make, trip, return'. Vowel-only monosyllabic words are also used in Setswana, for instance in words like *ao* 'to/of'. Third, like Sepedi, c-syllables can be formed by the m, n, l, r, and ŋ syllabic consonants. The simple syllabic consonants can appear at the word-initial position such as in words like *m-ma* 'mom', and word-medial positions such as in words like *bo-r-re* 'fathers' (Otlogetswe and Ramaeba, 2022). The ŋ nasal consonant can also appear at the word-final position such as in words like *fi-sa-ng-* 'hot'. While other syllabic consonants behave similarly to those in Sesotho, the current Sesotho syllabification system does not account for the representation of the r c-syllable in its rules.

3 Methodology

The South African Centre for Digital Language Resources hosts a publicly available online repository at repo.sadilar.org. For this article, we collected two Autshumato 6 corpora, that is, the Sepedi (McKellar, 2022a) and Setswana (McKel-

word	syllables	word	syllable	word	syllable	word	syllable
go	go	bo	bo	tla	tla	feta	fē ta
ya	ya	mmušo	m mu šo	na	na	barutwana	ba ru twa na
le	le	rena	re na	tše	tše o	mokgwa	mo kgwa
ka	ka	yona	yo na	swanetše	swa ne tše	karolo	ka ro lo
a	a	kudu	ku du	wo	wo	leo	le o
e	e	swanetšego	swa ne tše go	pele	pe le	fela	fe la
ba	ba	godimo	go di mo	bona	bo na	maemo	ma e mo
tša	tša	gagwe	ga gwe	gona	go na	kgopelo	kgo pe lo
o	o	nngwe	n ngwe	gomme	go m me	moo	mo o
di	di	mongwe	mo ngwe	gago	ga go	dingwe	di ngwe
ye	ye	gape	ga pe	be	be	bjo	bjo
se	se	fao	fa o	ao	a o	ngwaga	ngwa ga
ke	ke	ngwala	ngwa la	bjalo	bjalo	ntle	n tle
wa	wa	tshedimošo	tshe di mo šo	batho	ba tho	lebaka	le ba ka
tše	tše	motho	mo tho	dira	di ra	tee	te e
gore	go re	bala	ba la	yo	yo	šomiša	šo mi ša
ga	ga	morago	mo ra go	lego	le go	mešomo	me šo mo
sa	sa	tšwa	tšwa	bao	ba o	nago	na go
la	la	ile	i le	moka	mo ka	latelago	la te la go
ge	ge	mabapi	ma ba pi	seo	se o	maleba	ma le ba
goba	go ba	mošomo	mo šo mo	borwa	bo rwa	tšona	tšo na
re	re	gare	ga re	afrika	a fri ka	lenaneo	le na ne o
mo	mo	naga	na ga	setšhaba	se tšha ba	ditirelo	di ti re lo
bjalo	bjalo	mme	m me	bohlokwa	bo hlo kwa	taolo	ta o lo
yeo	ye o	molao	mo la o	nako	na ko	šoma	šo ma

Table 2: Lists of frequently used words and syllabified counterparts in Sepedi

lar, 2022b) Autshumato monolingual corpora. The Sepedi corpus contained a total of 3 458 067 words while the Setswana corpus contained a total of 5 219 070 words.

We used *bash* to extract four frequency lists. One, a list of one hundred most frequently used words in Sepedi. Two, the hundred most frequently used words in Setswana. Three, the hundred most infrequently used words in Sepedi. Four, the hundred least frequently used words in the Setswana corpus.

We then extracted the syllabification information from all four lists using Sibeko and Van Zaanen’s (2022b) rule-based syllabification system that was also downloaded from SADiLaR’s repository⁴.

4 Results

This section presents the results of the syllabification process. Both the 100 most used words and the 100 least used words from the Autshumato corpora for Sepedi and Setswana are presented. Stop words were not considered for any of the four lists.

4.1 Sepedi

4.1.1 Frequently used words

The hundred most frequently used Sepedi words ranged between 229 028 times [for the word *go*]

⁴see <https://repo.sadilar.org/handle/20.500.12185/556> for the Sesotho syllabification systems

and 3 387 times [for the word *šoma*]. The list of original words and their syllables are presented in table 2. The v-syllable, cv-syllable, and c-syllable types can be observed from the list. Note that we use the dash (-) to indicate syllable boundaries while the syllabification system only uses spaces.

The v-syllables structure was observed for monosyllabic vowel-only words such as a, e, ε, o and o. Furthermore, we observed v-syllables at the word-initial position in words such as *ile* [i-le-] ‘went’, the word-medial position in words such as *taolo* [ta-o-lo-] ‘control’, and the word-final position in words such as *tee* [te-e-] ‘only’. We did not observe any erroneous identification of v-syllables.

At least four cv-syllable types are present on the list. First, the one-consonant-one-vowel structure was observed in words such as *kudu* [ku-du-] ‘a lot’ which was correctly broken into two syllables. Second, the cwv syllable structure was evident in words such as *bohlokwa* [bo-hlo-kwa-] ‘important’ which was broken into three syllables. Third, the ccv structure was evident in words such as *tše* [tše-o-] ‘those’. Fourth, the ccwv structure was evident in words such as *ngwaga* [ngwa-ga-] ‘year’ which was broken into two syllables. Fifth, the cccv structure was observed in words such as *setšhaba* [se-tšha-ba-] ‘nation’ which was broken into three syllables. Unfortunately, there were no instances of cccwv syllable structures on the list.

word	syllables	word	syllable	word	syllables
ac	a c	abakase	a ba ka se	abalanago	a ba la na go
aar	a a r	abakeng	a ba ke ng	abelanang	a be la na ng
acr	a cr	abalobi	a ba lo bi	abelanago	a be la ne go
adi	a di	abapile	a ba pi le	abitafti	a bi ta fi ti
abby	a bby	abelala	a be la la	addictive	a ddi cti ve
abel	a be l	abelano	a be la no	adiolotši	a di o lo tši
abis	a bi s	abeleng	a be le ng	advantage	a dva n ta ge
abiy	a bi y	abetswe	a be tswe	abonagala	a bo na ga la
aesa	a e sa	abganya	a bga nya	aaaahhhhhh	a a a hhhhhh
aces	a ce s	abidjan	a bi dja n	abagantšhe	a ba ga n tšhe
acsa	a csa	abiwego	a bi we go	abaganwego	a ba ga nwe go
acts	a cts	abišana	a bi ša na	ablefatile	a ble fe ti le
adha	a dha	abokato	a bo ka to	aerospeisi	a e ro spe i si
adiš	a di š	aerobic	a e ro bi c	acceptable	a cce pta ble
aakar	a a ka r	adalats	a da la ts	accredited	a ccre di te d
abdel	a bde l	adilego	a di le go	adimišanwa	a di mi ša nwa
abedi	a be di	abattoir	a ba tto i r	abaganywago	a ba ga nywa go
abego	a be go	abdicate	a bdi ca te	abahlankedi	a ba hla n ke di
abeke	a be ke	adminiša	a dmi ni ša	adophilwego	a do pthi lwe go
abjwe	a bjwe	abelanye	a be la nye	aaohegnoboae	a a o he gno bo a e
abona	a bo na	aeration	a e ra ti o n	accessibility	a cce ssi bi li ty
abubi	a bu bi	aeskrimi	a e skri mi	accommodation	a cco m mo da ti o n
abuja	a bu ja	aethiops	a e thi o ps	actinomyces	a cti no myce te s
accom	a cco m	abrahams	a bra ha m s	adumeletšwego	a du me le tšwe go
adira	a di ra	accounts	a cco u n ts	adoption	a do pti o n
adult	a du lt	acidosis	a ci do si s	advocate	a dvo ca te
aeemo	a e e mo	adimišwa	a di mi šwa	abortion	a bo rti o n
abacus	a ba cu s	adimišwe	a di mi šwe	abaganago	a ba ga na go
acacia	a ca ci a	admirale	a dmi ra le	abaganeng	a ba ga ne ng
acdas	a cda sa	aemiše	a e mi še	abagantše	a ba ga n tše
achmat	a chma t	aeneng	a e ne ng	abulela	a bu le la
acquah	a cqu a h				
aakpaorleatsštwikai			a a kpa o rle a tsštwi ka i		
abeahlalošetšamapho			a be a hla lo še tša ma pho		
abonagopotologafaoabego			a bo na go po to lo ga fa o a be go a		
adiraboipiletšobjakagare			a di ra bo i pi le tšo bja ka ga re		
abonakebonabaobafetelwago			a bo na ke bo na ba o ba fe te lwa go		
abelwagokelenaneoedirwemenyetlayagoyagoile			a be lwa go ke le na ne o e di rwe me nye tla ya go ya go i le		

Table 3: Lists of least used words and syllabified counterparts in Sepedi

Furthermore, our list of frequently used words was limited in that it did not reflect all possible consonant-only syllable types. Even so, we were able to investigate the behaviour of the syllabic m and n nasal consonants. For instance, we find words such as *mmušo* [m-mu-šo-] ‘government’ and *nngwe* [n-ngwe-] ‘one’ which were correctly broken into syllables.

4.1.2 Least used words

We also surveyed the hundred least-used words from the Sepedi corpus. Each of the words appeared no more than once in the corpus. The original words and the derived syllables are listed in table 3.

Our Sepedi list of most infrequently used words contained instances of untranslated English words. Some of the English words were left as references for newly coined Sepedi words. We did not clean the list, instead, we fed it into the syllabification

system to see how the system would handle all the different unexpected words.

Fortunately, rule-based syllabification systems are best for unseen words (Adsett et al., 2009). Being able to handle unseen words allows the syllabification system to identify syllable boundaries in unexpected words such as concatenations like *abelwagokelenaneoedirwemenyetlayagoyagoile* and in words from a different language such as the English word ‘Abrahams’ [a-bra-ha-ms].

Three v-syllable structures were observed. First, the word-initial v-syllable structure was observed in words such as *abjwe* [a-bjwe-] ‘shared’ which was broken down into two syllables. Second, the word-final v-syllable structure was observed in the word *aaohegnoboae*⁵ which was broken into eight syllables [a-a-o-he-gno-bo-a-e-]. Finally, the word-

⁵note that this is another instance of a non-Sepedi word. It was used here due to the absence of a proper Sepedi word with the word-final v-syllable

word	syllables	word	syllable	word	syllable	word	syllable
a	a	tswa	tswa	letsatsi	le tsa tsi	re	re
wa	wa	fela	fe la	madi	ma di	rona	ro na
ba	ba	ga	ga	maemo	ma e mo	sa	sa
baagi	ba a gi	gago	ga go	metsi	me tsi	se	se
baitthuti	ba i thu ti	gagwe	ga gwe	mme	m me	sengwe	se ngwe
bana	ba na	gape	ga pe	mmogo	m mo go	seno	se no
batho	ba tho	go	go	mo	mo	teng	te ng
batla	ba tla	godimo	go di mo	mongwe	mo ngwe	thata	tha ta
bile	bi le	gore	go re	morago	mo ra go	thusa	thu sa
bo	bo	haba	ha ba	motho	mo tho	tiro	ti ro
bona	bo na	jaaka	ja a ka	na	na	tla	tla
bone	bo ne	jalo	ja lo	nako	na ko	tlaa	tla a
borwa	bo rwa	jo	jo	nang	na ng	tlase	tla se
ya	ya	jwa	jwa	ne	ne	tsa	tsa
di	di	ka	ka	neng	ne ng	tse	tse
dilo	di lo	karolo	ka ro lo	ngwaga	ngwa ga	tsela	tse la
dingwe	di ngwe	ke	ke	nna	n na	tshedimose tso	tshe di mo se tso
dintlha	di n tlha	kgona	kgo na	nne	n ne	tshwanetse	tshwa ne tse
dira	di ra	kgotsa	kgo tsa	nngwe	n ngwe	tsothle	tso tlhe
dirisa	di ri sa	kwa	kwa	ntlha	n tlha	farologaneng	fa ro lo ga ne ng
ditirelo	di ti re lo	kwala	kwa la	ntse	n tse	aforika	a fo ri ka
ditiro	di ti ro	la	la	o	o	botlhokwa	bo tlho kwa
e	e	latelang	la te la ng	pele	pe le	yo	yo
eno	e no	le	le	puo	pu o	yona	yo na
fa	fa	leng	le ng	puso	pu so	yone	yo ne

Table 4: Lists of frequently used words and syllabified counterparts in Setswana

medial v-syllable structure was observed in words such as *aemiše* [a-e-mi-še-] ‘he stops’ which was broken into four syllables.

Five cv-syllable structures were identified from the word list. First, the one-consonant-one-vowel structure was observed in words such as *abulela* ‘he opened’ which was broken into four syllables [a-bu-le-la-]. Second, the cwv structure was evident in words such as *adimišwe* ‘lend’ which was broken into four syllables [a-di-mi-šwe-]. Third, the ccv structure was observed in words such as *abagantše* ‘divided’ which was broken into five syllables [a-ba-ga-n-tše-]. Fourth, we observed the ccwv structure in words like *adumeletšwego* [a-du-me-le-tšwe-go-] ‘approved’ which was broken into six syllables. Finally, we observed the cccv structure in words such as *abagantšhe* [a-ba-ga-n-tšhe-] ‘separate’ which was broken into five syllables. There were no instances of the ccwv structure in the current word list.

Only two c-syllable structures were observed from the list of one hundred least frequently used words. First, the n syllabic nasal structure was observed in words such as *abagantše* discussed above. Second, the ŋ complex syllabic nasal structure was observed in words such as *abaganeng* [a-ba-ga-ne-ŋ-] where it is in the word-final position.

Although there were no instances of the syllabic nasal m, there is an interesting behaviour of the con-

sonant m in words such as *adminiša* [a-dmi-ni-ša-] ‘administer’ and *admirale* [a-dmi-ra-le-] ‘admiral’ which are respectively broken down into four syllables. This structure of the *dmi* syllable is unexpected in the Sepedi language and it is enabled only by naturalised loaned words. One would expect a vowel between the letters d and m as in *adiminiša* [a-di-mi-ni-ša-] and *adimirale* [a-di-mi-ra-le-].

Nonetheless, Sotho-Tswana languages do not have strict rules for spelling loaned words. As a result, the ‘dmi’ syllable does not break spelling rules as it is a ccv syllable which falls under the cv-syllable structure generally preferred in Bantu languages (Ditsele, 2014). What is important here is that the syllable boundaries are correctly identified.

4.2 Setswana

4.2.1 Frequently used words

The one hundred most frequently used Setswana words ranged between 334 188 times [for the word *go*] and 4 688 times [for the word *yona*].

Vowel-only monosyllabic words such as a, e, and o were frequently used in the corpus. Furthermore, we observed v-syllable structures in word-initial positions in words such as *eno* [e-no-] ‘that one’, in word-final positions such as in words like *puo* [pu-o-] ‘speech’, and in word-medial position in words such as ‘*jaaka*’ [ja-a-ka-] ‘like’. Overall, no

word	syllables	word	syllable	word	syllable
aaa	a a a	acae	a ca e	abiweka	a bi we ka
aabb	a a bb	accelerated	a cce le ra te d	abolition	a bo li ti o n
aaferika	a a fe ri ka	accidental	a cci de n ta l	abolokiwang	a bo lo ki wa ng
aaforika	a a fo ri ka	accom	a cco m	abone	a bo ne
aa Kantse	a a ka n tse	accountancy	a cco u n ta n cy	aboratoring	a bo ra to ri ng
aa karetsang	a a ka re tsa ng	accra	a ccra	abosesebalolang	a bo se se ba lo la ng
aamebitlwa	a a me bi tlwa	accuweather	a ccu we a the r	about	a bo u t
aamogetse	a a mo ge tse	acdas	a cda sa	absorbers	a bso rbe rs
aasa	a a sa	ace	a ce	absorption	a bso rpti o n
aa u	a a u	acesulfame	a ce su lfa me	abueng	a bu e ng
abakhase	a ba kha se	achievable	a chi e va ble	abuiwa	a bu i wa
abalanang	a ba la na ng	acln	a cln	abuja	a bu ja
abapisa	a ba pi sa	actions	a cti o n s	abula	a bu la
abaram	a ba ra m	actives	a cti ve s	abuse	a bu se
abasa	a ba sa	activities	a cti vi ti e s	abutiago	a bu ti a go
abasetsana	a ba se tsa na	actt	a ctt	abutilelapa	a bu ti le la pa
abat	a ba t	actuarial	a ctu a ri a l	acacia	a ca ci a
abbotsford	a bbo tsfo rd	actuary	a ctu a ry	adhanom	a dha no m
abbott	a bbo tt	acumda	a cu m da	adikarabo	a di ka ra bo
abdalla	a bda l la	acwa	a cwa	adikolo	a di ko lo
abdel	a bde l	acwy	a cw y	adileng	a di le ng
abderrahmane	a bde rra hma ne	acyclovir	a cyclo vi r	adimaneng	a di ma ne ng
abeetsweng	a be e tsw e ng	adalats	a da la ts	adimanwe	a di ma nwe
abel	a be l	adama	a da ma	adimelwang	a di me lwa ng
abelanweng	a be la nwe ng	adapotara	a da po ta ra	adimeng	a di me ng
abelweng	a be lwe ng	adb	a db	adimetsweng	a di me tsw e ng
abengditirelo	a be ng di ti re lo	adc	a dc	adiminsanang	a di mi n sa na ng
aberbargoed	a be rba rgo e d	added	a dde d	adimisane	a di mi sa ne
abgn	a bgn	address	a ddre ss	adimisaneng	a di mi sa ne ng
abillweng	a bi l lwe ng	adelaide	a de la i de	adimisanwang	a di mi sa nwa ng
abining	a bi ni ng	adenoviuses	a de no vi u se s	adimiswang	a di mi swa ng
abiotiki	a bi o ti ki	adequate	a de qu a te	adimiwe	a di mi we
abis	a bi s	adha	a dha	adimlweng	a di m lwe ng
adingwe	a di ngwe				

Table 5: List of least used words and syllabified counterparts in Setswana

errors were observed for v-syllable structures.

At least six cv-syllable structures were observed. One, the cv structure was evident in words such as *pele* [pe-le-] ‘following’. Two, the cwv structure was observed in words like *kwala* [kwa-la-] ‘write’. Three, the ccv syllable structure was observed in words such as *tlase* [tla-se-] ‘low’. Four, the ccwv syllable structure was observed in words such as *se ngwe* [se-ngwe-] ‘something’. Five, we observed the cccv syllable structure in words like *botlhokwa* [bo-tlho-kwa-] ‘important’. Six, we observed the cccwv structure in words such as *tshwanetse* [tshwa-ne-tse-] ‘must’.

We also observed three c-syllable types. One, the m syllable was evident in words such as *mmogo* ‘together’ where it appeared at the word-initial position [m-mo-go-]. Two, the n syllable was observed at the word-medial position in words like *dintlha* [di-n-tlha-] ‘details’. Three, the ŋ syllable appeared at the word-final position in words like *neng* [ne-ng-] ‘when’.

4.2.2 Least used words

The rarest words from the Setswana corpus appeared no more than once in the corpus. The original words together with the syllables are presented in table 5. Similar to the Sepedi list, the Setswana list contains some instances of incorrect spelling such as *adiminsanang* [a-di-mi-n-sa-na-ng-]. Even so, the syllabification system was able to insert justifiable syllable boundaries at the expected spaces. For instance, the additional n in *adimi-n-sanang* is followed by a correct syllable boundary. Furthermore, like the Sepedi list, there are numerous instances of non-Setswana words on the list.

Three v-syllable structures were observed. That is, at the word-initial placement in words like *adileng* [a-di-le-ng-] ‘laid out’, the word-medial position in words such as *abiotiki* [a-bi-o-ti-ki-] ‘abiotic’, and the word-final location in the untranslated English acronym for Autism Centers of Excellence, that is *acae* [a-ca-e-].

Five cv-syllable structures were also observed. That is, the cv syllable in words like *adikarabo*

[a-di-ka-ra-bo-] ‘of answers’, the cwv syllable in words like *abelwaneng* [a-be-lwa-ne-ng-] ‘shared’, the ccv syllable in words such as *adimanwe* [a-di-ma-nwe-] ‘borrowed each other’, and the ccwv syllable in words like *adimetsweng* [a-di-me-tsweng-] ‘borrowed for’.

Finally, three c-syllable types were observed. One, the m syllable was observed in words like *adimlweng* [a-di-m-lwe-ng-]. Although the word is incorrectly spelt, the syllable boundaries are in the expected places. Two, the word-medial position l syllable is evident in words such as *abillweng* [a-bi-l-lwe-ng-]. The second lateral in the word is unfortunately a typo. Even so, the syllabification system managed to insert justifiable boundaries following the order of letters in the word. Finally, the ŋ syllable was observed in the word-final position in words such as *adimeng* [a-di-me-ng-] and in the word-medial position in words like *abengditirelo* [a-be-ng-di-ti-re-lo-].

As we expected, the Setswana syllabic ‘r’ is not covered by the Sesotho rules for syllabification as described in Sibeko (2022) and Sibeko and van Zaanen (2022a). Unfortunately, a proper Setswana word containing the ‘r’ c-syllable is not present in both lists of Setswana words. Even so, the word *abderrahmane* [a-bde-rra-hma-ne-] contains consecutive ‘r’ letters. In this occurrence, the expected syllable boundary between the ‘rra’, syllable, i.e. [r-ra-] is missing.

5 Discussion

As stated earlier in this article, the Sotho-Tswana languages are mutually intelligible to a great extent. Even though some vocabulary choices may be ambiguous, the ambiguity does not affect syllable breaks. This article set out to evaluate the Sesotho rule-based syllabification system on both Sepedi and Setswana words. We used the Autshumato machine translation corpora for both Sepedi and Setswana. The texts were translated from English texts as a pivot language. As a result, they contain somewhat similar information.

The v-syllable structures showed consistently correct syllable placement in both Sepedi and Setswana. All v-syllable structures argued by Sibeko (2022) were identified for both Sepedi and Setswana. All word-initial, word-medial, and word-final v-syllable structures were correctly identified. This consistency in the accuracy of the syllable breaks indicates that the current Sesotho syllabifi-

cation system is ideal for identifying v-syllables in both Sepedi and Setswana. Unfortunately, single-letter words cannot be broken down into syllables. Even so, no unexpected outputs were observed for single-letter vowel-only words.

The syllabification system inserted consistently correct syllable breaks in words containing the m and n syllabic consonants on both Sepedi and Setswana texts. Unfortunately, the ŋ could only be identified at the word-end position in Sepedi. As such, we were not able to observe its behaviour when it appears at word-initial and word-medial positions. Even so, the word-medial ŋ syllable was correctly identified in the Setswana list. Furthermore, the Sepedi list did not contain any instances of the l syllable. However, it was observed in the Setswana list. As a result, we can safely assume that the current syllabification system can insert correct syllable boundaries for the l consonant even in Sepedi as the l syllable behaves similarly in all three Sotho-Tswana languages.

The unexpected structure of the *dmi* syllable highlights a need for clear rules governing the behaviour of nasal consonants that follow other consonants. To this point, the rules are only descriptive when the nasal consonant comes before the other consonants. It might be interesting to also investigate this in future studies.

Although the system attempted to identify syllable boundaries in non-Sotho-Tswana words, that is English words, the discord between the rules as implemented in the syllabification system and the structure of the orthography of English words could not be ignored.

All expected syllable boundaries in the correctly spelt words in Setswana were successfully identified by the syllabification system. We however missed an instance of consecutive r syllable in both the Sepedi and the Setswana lists. It would have been interesting to analyse actual Setswana words with such instances. Nonetheless, we noticed the absence of a syllable break between consecutive r letters in the non-Setswana examples. This finding confirms our initial assumption that the current Sesotho syllabification system does not identify the r syllabic consonants.

We also noticed inconsistencies in the spelling of words like *aaforika* and *aaferika* ‘Africa’ in the Setswana list of one hundred rarest used words. Both words were justifiably broken into syllables according to the given incorrect spelling, see table

5. Although this is unimportant in the identification of syllables, it does affect the counts of syllables as it may exaggerate syllable counts and types identified from a text.

The syllabification system's inability to recognise diacritics such as those used in Sepedi proved unproblematic for our selected words. That is, Sepedi words with diacritics were correctly broken into syllables. Even so, we are not aware of all possible placements of letters with accents in the written Sepedi words. As a caution, we recommend that the update to the syllabification system include letters with diacritics.

Overall, we recommend that the Sesotho rule-based syllabification be updated to cover all three standardised Sotho-Tswana languages. We also recommend that diacritics be included and specifically handled in the recommended Sotho-Tswana syllabification system. Equally important, we recommend that updated rules should also cover specific rules for handling the Sepedi and Setswana r syllable.

Limitations

The results of this article are limited by our sampling method which included the use of the hundred most used words and the hundred least used words in each of the languages as evidenced by the Autshumato corpora. Future studies could consider developing gold-standard syllable information annotated corpora for Sepedi and Setswana. The corpora could then be used for evaluating the usability of the T_EX-based Sesotho syllabification system on Sepedi and Setswana texts. In this article, we were limited by the lack of such corpora and were therefore limited only to the evaluation of the rule-based syllabification system. The lists used did not contain correct Sepedi examples of words containing consecutive r consonants. As a result, we are unable to draw concrete conclusions on the rule-based syllabification system's performance on such words.

Ethics Statement

This article utilizes publicly available resources. The authors have taken measures to ensure that the data used is properly cited and attributed to the original sources and that any potential biases or limitations in the data are acknowledged.

References

- Connie R Adsett, Yannick Marchand, Vlado Kes, et al. 2009. Syllabification rules versus data-driven methods in a language with low syllabic complexity: The case of Italian. *Computer Speech & Language*, 23(4):444–463.
- Etienne Barnard, Marelle H Davel, Charl Van Heerden, Febe De Wet, and Jaco Badenhorst. 2014. The NCHLT corpus of the South African languages. In *Proceedings of the 4th International Workshop Spoken Language Technologies for Under-resourced Languages*, pages 194–200.
- Sekgaila Chokoe. 2020. Spell it the way you like: The inconsistencies that prevail in the spelling of Northern Sotho loanwords. *South African Journal of African Languages*, 40(1):130–138.
- Katherine Demuth. 2007. Sesotho speech acquisition. *The international guide to speech acquisition*, pages 526–538.
- Patrick Dickens. 1978. A preliminary report on Kgala-gadi vowels. *African Studies*, 37(1):99–106.
- Thabo Ditsele. 2014. Why not use Sepitori to enrich the vocabularies of Setswana and Sepedi? *Southern African Linguistics and Applied Language Studies*, 32(2):215–228.
- Aditi Sharma Grover, Gerhard B van Huyssteen, and Marthinus W Pretorius. 2010. The South African Human Language Technologies audit. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2847–2850.
- Aditi Sharma Grover, Gerhard Beukes, van Huyssteen, and Marthinus W. Pretorius. 2011. The South African Human Language Technology audit. *Language resources and evaluation*, 45:271–288.
- Samson Mbizo Guma. 1982. *An outline structure of Southern Sotho*, 2nd edition. Shooter and Shuter Publishers, Pietermaritzburg, South Africa.
- Pali Lehohla. 2011. Census in brief. *Statistics South Africa*. government printing works: Pretoria.
- Katjie Sponono Mahlangu. 2015. *The growth and development of isiNdebele orthography and spelling (1921-2010)*. Ph.D. thesis, University of Pretoria.
- Leketi Makalela. 2009. Harmonizing South African Sotho language varieties: Lessons from reading proficiency assessment. *International Multilingual Research Journal*, 3(2):120–133.
- Zvinaiye Patricia Makaure. 2021. *The contribution of phonological processing skills to early literacy development in Northern Sotho-English bilingual children: A longitudinal investigation*. Ph.D. thesis, The University of South Africa, Pretoria.

- Litsépiso Matlosa. 2017. Sesotho orthography called into question: The case of some Sesotho personal names. *Nomina Africana: Journal of African Onomastics*, 31(1):51–58.
- Cindy McKellar. 2022a. *Autshumato Monolingual Sepedi Corpus*. ONLINE. South African Centre for Digital Language Resources. Available at: <https://repo.sadilar.org/handle/20.500.12185/582> Accessed: 28 Jan 2023.
- Cindy McKellar. 2022b. *Autshumato Monolingual Setswana Corpus*. ONLINE. South African Centre for Digital Language Resources. Available at: <https://repo.sadilar.org/handle/20.500.12185/584> Accessed: 28 Jan 2023.
- Victor Mojela. 2016. Etymology & figurative: The role of etymology in the lemmatization of Sotho terminology. In *The 10th International Conference of the Asian Association for Lexicography (AsiaLex2016) 1-3 June 2016 Manila, The Philippines*, page 93.
- Carmen Moors, Illana Wilken, Karen Calteaux, and Tebogo Gumede. 2018a. Human Language Technology audit 2018: Analysing the development trends in resource availability in all South African languages. In *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*, pages 296–304.
- Carmen Moors, Illana Wilken, Tebogo Gumede, and Karen Calteaux. 2018b. [Human Language Technology audit 2017/18](#). Technical report, CSIR Meraka Institute.
- Mildred Nkolola-Wakumelol, Liketso Rantsoz, and Keneilwe Matlhaku. 2012. Syllabification of consonants in Sesotho and Setswana. In H S Nginga-Koumba-Binza and S Bosch, editors, *Language Science and Language technology in Africa: Festschrift for Justus C. Roux*, pages 10–13. Sun Express, Stellenbosch, South Africa.
- Thapelo J Otlogetswwe. 2017. Setswana syllable structure and distribution. *Nordic Journal of African Studies*, 26(4):28–28.
- Thapelo J Otlogetswwe and Goabilwe N Ramaeba. 2022. Nickname creation through shortening Setswana personal names. *South African Journal of African Languages*, 42(2):200–206.
- Danie Prinsloo. 2011. Tribute to Professor Louis Jacobus Louwrens: This issue of the south african journal of african languages is dedicated to Professor Louis Jacobus Louwrens. *South African Journal of African Languages*, 31(1):1–5.
- Tebogo Rakgogo and Evangeline Zungu. 2021. The onomastic possibility of renaming the Sepedi and Sesotho sa Leboa (Northern Sotho) language names to restore peace, dignity and solidarity. *Literator (Potchefstroom. Online)*, 42(1):1–14.
- Republic of South Africa. 1996. *Constitution of the Republic of South Africa*. Department of Justice, Pretoria.
- Boikanyego Sebina. 2014. First language attrition in the native environment. *LANGUAGE*, 6:53–60.
- Johannes Sibeko. 2022. Tshebediso ya melao kabong ya dinoko tsa Sesotho. *Southern African Linguistics and Applied Language Studies*, 40(4):494–506.
- Johannes Sibeko and Mmasibidi Setaka. 2022. An overview of Sesotho blark content. *Journal of the Digital Humanities Association of Southern Africa*, 4(01).
- Johannes Sibeko and Menno van Zaanen. 2022a. Developing a text readability system for Sesotho based on classical readability metrics. In *Responding to Asian diversity*, pages 571–572.
- Johannes Sibeko and Menno van Zaanen. 2022b. Sesotho syllabification systems. *Southern African Centre for Digital Language Resources*. Available at: <https://repo.sadilar.org/handle/20.500.12185/555> [accessed: 3 jan 2023].
- Luan Staphorst. 2022. Ongehoord: Voices unaccented; voices unharmonized. Afrikaans and South Africa’s first peoples in discourses of higher education transformation. *Unpublished MA dissertation, Nelson Mandela University*.
- Suyanto Suyanto, Ade Romadhony, Febryanti Stehvanie, and Rezza Nafi Ismail. 2021. Augmented words to improve a deep learning-based indonesian syllabification. *Heliyon*, 7(10):e08115.
- Hein Willemsse. 2018. The hidden histories of Afrikaans. *Whiteness Afrikaans Afrikaners: Addressing Post-Apartheid Legacies, Privileges and Burdens*, page 115.
- Carien Wilsenach. 2019. Phonological awareness and reading in Northern Sotho—understanding the contribution of phonemes and syllables in grade 3 reading attainment. *South African Journal of Childhood Education*, 9(1):1–10.
- Philimon Nhlanhla Zulu, Gerrit Botha, and Etienne Barnard. 2008. Orthographic measures of language distances between the official South African languages. *Literator: Journal of literary criticism, comparative linguistics and literary studies*, 29(1):185–204.