Nordic Journal of African Studies 26(4): 403–430 (2017)

Setswana Syllable Structure and Distribution

Thapelo J. OTLOGETSWE
University of Botswana, Botswana

ABSTRACT

This paper investigates the frequency and distribution of Setswana syllables by analysing the frequency and distribution of (orthographic representations of) Setswana syllables in a wordlist of about 59,000 words extracted from a corpus of over 13.6 million words. The study uses Wordsmith tools to calculate syllables found at word initial, word medial and word final positions of Setswana words. Each orthographic consonant is then paired to each of the five Setswana orthographic vowels to generate a list of potential Setswana syllables. Each potential syllable is then tested for frequency and distribution in the extracted wordlist. Non-occurring syllables are discarded, leaving an inventory of all Setswana syllables that make up Setswana words. The final part investigates the distribution of the allophones [l] and [d], as well as the phoneme /g/ and confirms, with some modification, previous analyses.

***Keywords**: Setswana, syllable frequency, syllable distribution, phoneme, orthography.*

1. INTRODUCTION

Not much advanced analysis of Setswana syllable exists. Only basic analysis of the Setswana syllable structure has been undertaken (DALL, 1999; Cole, 1955:52). Much of the Setswana syllable research, however, is based on a theoretical assumption of what constitutes a Setswana syllable and largely ignores the large amount of borrowed terms which raise new challenges to long-held views. The paper studies a 13.6 million Setswana corpus of orthographic words using Oxford Wordsmith Tools to determine what constitutes a syllable in Setswana and how such syllables are distributed at either word initial, medial or final position. While syllabicity is usually perceived as a characteristic of speech, this study uses written text mainly because it is cheap to compile and easy to compute. The syllable position measured count is therefore of the Setswana orthographic word.

The study uses quantitative and computational approaches to study language data to reveal patterns in language (Kessler & Treiman, 1997; Munthuli et al., 2015). The study measures syllable frequency and distribution and contributes towards a better understanding of Setswana phonotactics and syllable preference in Setswana word formation. Measuring syllable frequencies in a language is critical for many areas of linguistics, phonetics and speech technology.

In this study orthographic data is investigated since it lends itself to statistical analysis at a large scale. It assumes that all Setswana consonants can pair with any Setswana vowel to form a syllable until corpus data proves otherwise. For instance the consonantal phonemes /tlh/ and /ph/ can be paired to each of the Setswana vowels in the following manner:

tl^h →tl^h a, tl^he, tl^hi, tl^ho, tl^hu.
p^h →p^ha, p^he, p^hi, p^ho, p^hu

A list of all potential Setswana syllables is generated then we test the occurrence of each of the potential Setswana syllables against the corpus-generated wordlist. The study uses Chargrams software tool of Oxford Wordsmith Tools (version 7) (Scott, 2017) to run the list of potential Setswana syllables against a frequency list.

2. SETSWANA SYLLABLE STRUCTURE

The structure of Setswana syllable is fairly straightforward. It can take three forms:

1. It can be formed of a consonantal onset that occupies syllable initial position before a vowel nucleus. This is also known as a CV pattern. Orthographically this single phonological consonantal position may be composed of a single orthographical character consonant, digraph, trigraph or quadgraph as in Table 1. It is essential to emphasise that all these complex orthographic consonants constitute single phonological consonants.

Table 1. *Setswana syllables*.

Type	**Orthographic form**	**phonetic transcription**
Single consonant onset	ba, ga, na, sa, le, fa	ba, ga, na, sa, le, fa
Digraph onset	tsa, tla, kga, pha, nga, nya	tsa, tɬa, q^ha, p^ha, ŋa, ɲa
Trigraph onset	tlha, tshe, tšhu, kgwa	tl^ha, ts^he, $ʧ^h$u, q^{hw}a
Quadgraph onset	tlhwa, tshwe, tšhwu	tl^{hw}a, ts^{hw}e, $ʧ^{hw}$u

2. A syllable can also be formed of a single vowel such as /a/ in a.ba (‘give away’) and /o/ in pe.o (‘seed’).
3. A Setswana syllable may also be formed of any of the four syllabic consonants when they occur as the first character of a homogeneous digraph (a digraph comprising two instances of the same character) m, n, l, r such as in mme (‘but’), nna (‘me’), sello (‘cry’), rre (‘father’) and the syllabic /ŋ/ when it occurs before other consonants or at word final position as in mang (‘who’), nngate (‘throw at me’) and nkatle (‘kiss me’). It should be noted that the orthographic form <ng> never occurs in both word-initial and word-

medial positions as a distinct syllable, though it does occur in such positions as a syllable onset in words such as ngaka /ŋàkà/ (‘doctor’) and ngongorega /ŋɔ́ŋɔ́réχà/ “(‘complain’). The orthographic <ng> only occurs at the end of words as a distinct syllable. However, the syllabic nasal /ŋ/ does occur in word initial and word medial positions in Setswana for example in words such as nkuku /ŋ̀kùkú/ (‘grandmother’), bankane /bàŋ̀kànì/ (‘age mates’). This observation is significant since the syllabic consonant /ŋ/ in Setswana is represented orthographically by <n> and <ng>. <n> is pronounced /ŋ/ when it comes before the voiceless velar plosive /k/ as in nkatle /ŋ̀kàtɬɛ́/ (‘kiss me’).

2.1 CONSONANT INVENTORY

To better understand the Setswana syllable structure and distribution the inventory of Setswana vowels and consonants is discussed briefly since it has a bearing on what a syllable is. Setswana has 28 phonemic consonants (Batibo, 2013 and DALL, 1999:12) as shown on Table 2.

Table 2. *Setswana phonetic symbols*.

		Labial	Alveolar		Post alveolar	Palatal	Velar	Uvular	Glottal
			Central	Lateral					
NASAL		/m/	/n/			/ɲ/	/ŋ/		
PLOSIVE	Unaspirated	/p/ /b/	/t/ /d/				/k/		
	Aspirated	/p^h/	/t^h/				/k^h/	/q^h/	
AFFRICATE	Unaspirated		/ts/	/tɬ/	/ʧ/ /ʤ/				
	Aspirated		/ts^h/	/$tɬ^h$/	/$ʧ^h$/				
FRICATIVE		/f/	/s/		/ʃ/			/χ/	/h/
Trill			/r/						
Approximant		/w/		/l/		/j/			

Four of these (p^h, t^h, k^h and q^h) are aspirated voiceless plosives, three are non-aspirated voiceless plosives (p, t and k), one voiced plosive /b/, three aspirated voiceless affricates ($ʧ^h$, $tɬ^h$, ts^h), three non- aspirated voiceless affricates (ts, tɬ, and ʧ), one voiced affricate (ʤ), five voiceless fricatives (f, s, ʃ, χ, and h), four nasals (m, n, ɲ, and ŋ), four sonorants (r, l, j and w).

The voiced plosive [d] is usually considered an allophone of /l/ where [d] is usually followed by close vowels such as /u/ and /i/ as in *dika* (‘surround’) and *huduga* (‘relocate’) while /l/ is followed by open vowels such as /a/ and /o/ as in *lala* (‘spend the night’) and *lole* (‘fought’). This paper revisits this matter of allophony later and tests it against corpus evidence.

The Setswana consonantal phonemes are represented orthographically with their equivalent forms as demonstrated on Table 3.

Table 3. *Phonetic consonants and their orthographic forms*.

Phonetic	Orthography
/m/	m
/p/	p
/pʰ/	ph
/b/	b
/f/	f
w/	w
/n/	n
/t/	t
/d/	d
/tʰ/	th
/ts/	ts
/tsʰ/	tsh
/s/	s
/r/	r
/tɬ/	tl
/tɬʰ/	tlh
/l/	l
/ʧ/	tš
/ʤ/	j
/ʧʰ/	tšh
/ʃ/	š
/ɲ/	ny
/j/	y
/ŋ/	ng
/k/	k
/kʰ/	kh
/qʰ/	kg
/χ/	g
/h/	h
Total	**29**

This orthographic representation is important to this study since it can be counted computationally to measure the frequency and distribution of Setswana consonantal phonemes amongst Setswana words.

2.2 CONSONANT + /W/ CLUSTER

The list of consonants as outlined above does not account for all Setswana consonants. All the twenty orthographic Consonant + /w/ (Cw) clusters which form digraphs, trigraphs and quadgraphs are not accounted for by the consonantal inventory. They are not captured in the Setswana phoneme inventory, where they are seen as combinations of phonemes. We list the orthographic form with their phonemic forms in Table 4.

Table 4. *Cw clusters*.

Orthography	Phoneme
gw	χ^w
jw	$ʤ^w$
kgw	q^w
khw	k^{hw}
kw	k^w
lw	l^w
ngw	$ŋ^w$
nw	n^w
nyw	$ɲ^w$
rw	r^w
sw	s^w
šw	$ʃ^w$
thw	t^{hw}
tlhw	tl^{hw}
tlw	$tɬ^w$
tshw	ts^{hw}
tšhw	$ʧ^{hw}$
tsw	ts^w
tšw	$ʧ^w$
tw	t^w

The question that arises is whether Cw clusters should be considered as single labialized consonantal phonemes or whether each should be considered as a sequence of consonantal phonemes since C and /w/ exist separately as consonantal phonemes in Setswana.

Following Gouskova et al. (2011), in this study we adopt an approach that Cw clusters form distinct phonemes and are recognised as single consonants.

With Setswana borrowing words from other languages, the assumptions regarding which sounds form part of the inventory of the Setswana phonology must be constantly under review as the language grows. This is so that linguistic analysis reflects language phonotactics without bias. While in general there is no [g] sound in Setswana, it is important to note that because of borrowing this sound is now pronounced in certain loan words such as *mmengu* (/m̀míŋχu/) ('mango'), *legwinya* (/lìgwìŋá/) ('fat cake'), ***ge**iti* (/geiti/) ('gate') and *digumbagumba* (/digumbagumba/) ('loud music stereo'). The <g> in these words is not pronounced as [χ] but as [g]. These terms are also borrowings and largely colloquial and informal.

Additionally, while /z/ is not a recognised Setswana phoneme, a number of words with [z] have been borrowed from other African languages such as Zulu. Amongst these are words such as *zama* ('try'), *nzamela* /ǹzàmɛ́là/ ('try for me; get me airtime'), *bazelwana* /bàzɛ̀lʷánà/ ('Christian believers') and *banuza* /bànúzà/ ('girls'). While these words are informal and colloquial, they are Setswana words which need to be accounted for within Setswana phonology.

2.2.1 Setswana Vowels

Setswana has seven vowels: /i, e, ɛ, o, ɔ, u, a/ (DALL, 1999:17) (See Chebanne et al. 1997 for a different view). The seven vowels may be represented in the following manner in a table:

Table 5. *Setswana vowels*.

Height	**Localisation**	
	Front	Back
Close	i	u
Half-close	e	o
Half-Open	ɛ	ɔ
Open	a	

Orthographically Setswana doesn't mark a distinction between /ɛ/ and /e/. They are both represented by <e>. Additionally Setswana orthography doesn't mark a distinction between /o/ and /ɔ/. They are both orthographically represented by <o>. Since Setswana doesn't mark the (ɛ, e) and (o, ɔ) distinction orthographically, this study uses only the five orthographic vowels <a, e, i, o, u> since they lend themselves to easy computation. This is a necessary limitation of the study, resulting from the use of a written corpus.

This paper will therefore attempt to answer the following set of questions:

1. What are the most/least frequent syllables in Setswana?
2. What are the most/least frequent syllables in orthographic word initial position?
3. What are the most/least frequent syllables in word medial position?
4. What are the most/least frequent syllables in word final position?
5. Which "potential" Setswana syllables are not found in Setswana words?
6. Which syllables though present in Setswana are not found in word initial position?
7. Which syllables though present in Setswana are not found in word medial position?
8. Which syllables though present in Setswana are not found in word final position?

3. STUDY LIMITATIONS

The current study has a number of limitations. It makes phonological claims based on a largely written corpus. It attempts to establish certain parallels between the orthographic word and Setswana phonology. This correlation is not always accurate. However, a large orthographic corpus's advantage is that it lends itself to computational testing which has not been attempted before. The results are therefore illuminating though they cannot be said to be definitive of Setswana phonology.

The study also does not consider the sonorous liquid /l/ which is common in certain Setswana dialects such as the Sekgatla (Kgafela) dialect in words such as *mollo* ('fire'), *sello* ('cry') and *kgollo* ('redemption'). It only analyses orthographic words and Setswana orthography doesn't allow /ll/ consonantal cluster. We therefore only analyse words such as *molelo* ('fire'), *selelo* ('cry') and *kgololo* ('redemption'). This study also doesn't differentiate between /e/ and /ɛ/ and /o/ and /ɔ/ since orthographically they are rendered as *e* and *o* respectively. Because of such limitations, although this study advances scholarship in better understanding Setswana syllabicity, it is not exhaustive.

4. METHODOLOGY

This study uses a Setswana corpus comprising 13,672,159 Setswana tokens. It includes Setswana text from various domains from both Botswana and South Africa. The corpus is discussed in detail in Otlogetswe (2011).

The corpus comprises both the written and spoken elements of Setswana text. Table 6 gives the different components of the whole corpus on the basis of tokens, types, type token ratio (TTR) and standardised type token ratio (STTR). The TTR is calculated by dividing types by tokens and multiplying by 100. By

types we refer to the *different types* of orthographic words that occur in a document while by tokens we refer to the count of every word regardless of its repetition. Thus if the word *gore* occurs in a document 75 times, it is said to constitute a single type but 75 tokens. The ratio for STTR is calculated at every specified number of tokens and an average of the different ratios computed. STTR is computed every *n* words as Wordlist goes through each text file. For instance *n* may be 1,000. In other words the ratio is calculated for the first 1,000 running tokens, and then calculated afresh for the next 1,000, and so on to the end of the text or corpus. A running average is computed, which means that an average type/token ratio based on consecutive 1,000-word chunks of text would be calculated. Texts with less than 1,000 words get a standardized type/token ratio of 0. STTR measures are attractive since they can compare type/token ratios across texts of differing lengths since what they do is segment a corpus into comparable chunks and calculate the type/token ratio for each.

Table 6. *The corpus written and spoken components*.

Text type	Tokens	Types	TTR	STTR
Written language	12,831,759	358,182	2.90	33.63
Spoken language	840,400	38,118	4.54	32.94

Table 6 reveals the corpus components divisions with the bulk of the corpus being material from the written language. While there are huge numerical differences between spoken and written language, both in terms of tokens and types, the differences on the basis of STTR between the two are minor (33.63 for the written language and 32.94 for spoken language). Ninety four percent of the corpus is written language material while six percent is transcribed language. This means that the corpus is skewed towards written language.

4.1 THE WRITTEN LANGUAGE COMPONENTS

The written component of the corpus occupies the largest part of the corpus at 94%. It comprises about 12,831,795 tokens, 358,182 types, with a STTR of 33.63. The scope of Setswana texts is limited. Most Setswana texts are published for the school curriculum. The majority of them are therefore grammar books and literature material (novels, plays and poetry books) for Setswana classes at both primary and secondary school levels. The texts are limited to materials for language and literature classes. Other subjects like Mathematics, Science, Agriculture and Art are taught in English, and therefore use texts written in English. Material in such subjects could therefore not be included in the corpus. Hardly ever do people read leisure texts in Setswana, partly because they are rare and partly because there is no literacy culture in the Setswana language, beyond secondary school education. School and public

libraries and bookshops have small numbers of Setswana books. There are neither bestsellers lists nor literary prizes which could be inspected for potential text inclusion. Most novels, plays and poetry added to the corpus had either been in the curriculum or were currently used in schools. All the texts were published after 1980. This date was not an intentional cut-off date, texts in Setswana published before 1980 are hard to find. The general rarity of texts, and their small size (in terms of number of words), necessitated the inclusion of whole texts in the corpus.

The corpus includes texts from two newspapers: *Mokgosi* and *Naledi*. *Naledi* is an insert in the largest private daily, *Mmegi*, while *Mokgosi* was the only weekly newspaper that wrote exclusively in Setswana. The *Mokgosi* newspaper closed down in 2005. The Setswana corpus also contains miscellaneous texts including student essays and letters from junior secondary schools, and the complete text of the national vision. There is also religious text (Christian, Bahai, Islamic texts). There are also political texts, Science texts, Business texts (e.g. from Botswana Meat Commission and Botswana Telecommunication Authority). Magazines in Setswana are hard to find. However, the *Kutlwano* magazine, which is predominantly written in English, has stories in Setswana which we were able to include in the corpus.

Table 7 reveals that Prose has the largest number of tokens and types, followed by Newspaper text. Science text has the smallest number of tokens. Although Science text has the smallest number of tokens, it is Politics that has the smallest vocabulary with the lowest number of types. The Prose section of the corpus has the largest number of tokens partly because of the large number of published Setswana novels included in this section. However Prose does not only include novels. Included in this section are folklores/folktales, collections of short stories, children's literature, cultural texts such as anthology of proverbs, sayings and riddles. Included also are cultural books about chieftaincy and the Setswana culture and language in general. The texts also comprise some online documents such as student tests and some online Setswana newsletters.

Table 7. *Overall statistics of the written subcorpus*.

Text Types	Tokens	Types	TTR	STTR
Prose Text	4,772,704	289,270	6.00	38.55
Newspaper Text	2,870,300	74,497	2.60	27.20
Religious Text	735,061	30,539	4.15	34.87
Chat-site Text	712,445	37,403	5.26	44.89
Miscellaneous Text	616,181	49,725	8.07	34.30
Poetry Text	530,261	47,235	8.91	43.43
Grammar books	504,559	35,386	7.01	37.05
Politics Text	262,652	10,782	4.11	30.23
Science Text	154,398	10,878	6.87	33.30

4.2 THE SPOKEN LANGUAGE COMPONENTS

The overall spoken component of the Setswana corpus has 840,400 tokens and 38,118 types. It has a type/token ratio of 4.54 and the STTR of 32.94. It represents the various spoken language collected from different sources such as radio call-in programs, parliamentary Hansards, sport commentary and unscripted dialogues. Table 8 shows the breakdown parts of the spoken component.

Table 8. *Spoken components statistics*.

Text-type	Tokens	Types	TTR	STTR	%
Hansards	616,695	33,581	5.45	35.51	73
Call-in	72,634	4,264	5.87	27.05	8.63
Interview	42,882	3,795	8.85	26.66	5
Sport	26,618	2,162	8.12	30.12	3.16
Open program	25,194	3,968	15.75	35.16	3
Education	23,545	1,329	5.64	25.20	2.80
Religious	17,736	2,210	12.46	29.14	2.11
Court	12,216	1,829	14.97	34.59	1.45
Dialogues	4,207	599	14.24	25.07	0.50

All spoken language is spontaneous speech and not scripted. As with any sampling, some compromise had to be achieved between what was "theoretically desirable and what was feasible" (Burnard, 1995: 21) for the corpus compilation. Indeed our approach to the compilation of the spoken text is characterised accurately by Atkins et al.

> The difficulty and high cost of recording and transcribing natural speech events lead the corpus linguist to adopt a more open strategy in collecting spoken language (Atkins et al., 1992: 3).

The corpus contains recordings of sermons, family dialogues, funeral services, classroom interactions, radio and television debates, court transcriptions and other spoken text, recorded using micro-cassette tape recorders. Conversations, speeches and other dialogues were recorded as unobtrusively as possible ensuring that the material gathered was as natural and as spontaneous as possible. For instance in classroom recordings, teachers were trained on how to record themselves and were given tape recorders to take to class. The researcher avoided going into a classroom to record a teacher since it was felt that this could create tension and make the teacher feel under observation which could lead them to modify their speech. In other cases a different approach was used. For instance in funeral recordings, permission was sought from the family in advance and different speakers in the service/ceremony.

5. TEXT PROCESSING

A frequency list was generated from the corpus with the most frequent words at the top and the least frequent at the bottom. The wordlist was then cleaned to eliminate misspelt words, English words, abbreviations, acronyms, place names, personal names and other textual material that was not needed. Personal and place names were eliminated because many of them are complete sentences which may distort the position of certain syllables at word initial, medial or final position. For instance: *Gaborone* is really three words in standard Setswana (disjunctive) spelling: *Ga bo rone* ('it doesn't ill-fit') and the personal name *Kemmone* is a sentence *Ke mmone* ('I have seen him/her'). This syllable position-count is therefore based on a disjunctive orthography and not a conjunctive one.

The principal aim of the study is to test "possible syllables" against a Setswana wordlist to determine the number of syllables that are used in the formation of Setswana words. Additionally, the aim of the study is to determine which syllables occur in the different positions of word initial, medial and word final positions. To generate the Setswana possible syllables every Setswana consonant is paired with each Setswana vowel and then tested against a wordlist to determine if it is found in word initial, word medial or word final orthographic positions. This study assumes that all Setswana consonants can pair with any Setswana vowel to form a syllable until corpus data proves otherwise. The limitation of this study must be emphasised here that since the orthography doesn't make a distinction between (ɛ, e) and (o, ɔ) this study will count orthographic [e] and [o] as if they represented a single sound.

The following are all the potential 253 Setswana syllables. These were generated by pairing each of the 48 consonants to each of the 5 Setswana vowels (49 x 5 = 245), three syllabic consonants *m*, *n*, *ng* which occur without being attached to a vowel were added and finally the five Setswana vowels were added (a, e, i, o, and u).

Table 9. *No of syllables that start with certain consonants*.

Articulation manner	**Consonantal onsets**	**Orthography**	**Freq.**	**%**
Voiced Plosive	ba, be, bi, bo, bu, da, de, di, do, du	ba, be, bi, bo, bu, da, de, di, do, du	10	4%
Aspirated Voiceless Plosive	p^ha, p^he, p^hi, p^ho, p^hu, t^ha, t^he, t^hi, t^ho, t^hu, k^ha, k^he, k^hi, k^ho, k^hu, $k^{hw}a$, $k^{hw}e$, $k^{hw}i$, $k^{hw}o$, $k^{hw}u$, $t^{hw}a$, $t^{hw}e$, $t^{hw}i$, $t^{hw}o$, $t^{hw}u$, q^ha, q^he, q^hi, q^ho, q^hu, $q^{hw}a$, $q^{hw}e$, $q^{hw}i$, $q^{hw}o$, $q^{hw}u$.	pha, phe, phi, pho, phu, tha, the, thi, tho, thu, kha, khe, khi, kho, khu, khwa, khwe, khwi, khwo, khwu, thwa, thwe, thwi, thwo, thwu, kga, kge, kgi, kgo, kgu, kgwa, kgwe, kgwi, kgwo, kgwu.	35	13.8%
Non-Aspirated	pa, pe, pi, po, pu, ta, te,	pa, pe, pi, po, pu, ta, te,	25	9.9%

Voiceless Plosives	ti, to, tu, t^{w}a, t^{w}e, t^{w}i, t^{w}o, t^{w}u, ka, ke, ki, ko, ku, k^{w}a, k^{w}e, k^{w}i, k^{w}o, k^{w}u.	ti, to, tu, twa, twe, twi, two, twu, ka, ke, ki, ko, ku, kwa, kwe, kwi, kwo, kwu.		
Aspirated Voiceless Affricates	ts^{h}a, ts^{h}i, ts^{h}e, ts^{h}o, ts^{h}u, tɬ^{h}a, tɬ^{h}e, tɬ^{h}i, tɬ^{h}o, tɬ^{h}u, ts^{hw}a, ts^{hw}e, ts^{hw}i, ts^{hw}o, ts^{hw}u, ʧ^{h}a, ʧ^{h}e, ʧ^{h}i, ʧ^{h}o, ʧ^{h}u, ʧ^{hw}a, ʧ^{hw}e, ʧ^{hw}i, ʧ^{hw}o, ʧ^{hw}u, tɬ^{hw}a, tɬ^{hw}e, tɬ^{hw}i, tɬ^{hw}o, tɬ^{hw}u.	tsha, tshi, tshe, tsho, tshu, tlha, tlhe, tlhi, tlho, tlhu, tshwa, tshwe, tshwi, tshwo, tshwu, tša, tše, tši, tšo, tšu, tšhwa, tšhwe, tšhwi, tšhwo, tšhwu, tlhwa, tlhwe, tlhwi, tlhwo, tlhwu.	30	11.9%
Non-Aspirated Voiceless Affricates	tsa, tse, tsi, tso, tsu, ʧa, ʧe, ʧi, ʧo, ʧu, tɬa, tɬe, tɬi, tɬo, tɬu, tɬ^{w}a, tɬ^{w}e, tɬ^{w}i, tɬ^{w}o, tɬ^{w}u, ts^{w}a, ts^{w}e, tswi, ts^{w}o, ts^{w}u, ʧ^{w}a, ʧ^{w}e, ʧ^{w}i, ʧ^{w}o, ʧ^{w}u.	tsa, tse, tsi, tso, tsu, tša, tše, tši, tšo, tšu, tla, tle, tli, tlo, tlu, tlwa, tlwe, tlwi, tlwo, tlwu, tswa, tswe, tswi, tswo, tswu, tšwa, tšwe, tšwi, tšwo, tšwu.	30	11.9%
Fricatives	fa, fe, fi, fo, fu, χa, χe, χi, χo, χu, χ^{w}a, χ^{w}e, χ^{w}i, χ^{w}o, χ^{w}u, ha, he, hi, ho, hu, sa, se, si, so, su, s^{w}a, s^{w}e, s^{w}i, s^{w}o, s^{w}u, ʃa, ʃe, ʃi, ʃo, ʃu, ʃ^{w}a, ʃ^{w}e, ʃ^{w}i, ʃ^{w}o, ʃ^{w}u.	fa, fe, fi, fo, fu, ga, ge, gi, go, gu, gwa, gwe, gwi, gwo, gwu, ha, he, hi, ho, hu, sa, se, si, so, su, swa, swe, swi, swo, swu, ša, še, ši, šo, šu, šwa, šwe, šwi, šwo, šwu.	40	15.8%
Voiced Affricate	ʤa, ʤe, ʤi, ʤo, ʤu, ʤ^{w}a, ʤ^{w}e, ʤ^{w}i, ʤ^{w}o, ʤ^{w}u.	ja, je, ji, jo, ju, jwa, jwe, jwi, jwo, jwu.	10	4%
Nasals	ma, me, mi, mo, mu, na, ne, ni, no, nu, ŋa, ŋe, ŋi, ŋo, ŋu, ŋ^{w}a, ŋ^{w}e, ŋ^{w}i, ŋ^{w}o, ŋ^{w}u, n^{w}a, n^{w}e, n^{w}i, n^{w}o, n^{w}u, ɲa, ɲe, ɲi, ɲo, ɲu, ɲ^{w}a, ɲ^{w}e, ɲ^{w}i, ɲ^{w}o, ɲ^{w}u, m, n, ŋ.	ma, me, mi, mo, mu, na, ne, ni, no, nu, nga, nge, ngi, ngo, ngu, ngwa, ngwe, ngwi, ngwo, ngwu, nwa, nwe, nwi, nwo, nwu, nya, nye, nyi, nyo, nyu, nywa, nywe, nywi, nywo, nywu, m, n, ŋ.	38	15%
Sonorants	ra, re, ri, ro, ru, r^{w}a, r^{w}e, r^{w}i, r^{w}o, r^{w}u, la, le, li, lo, lu, l^{w}a, l^{w}e, l^{w}i, l^{w}o, l^{w}u, wa, we, wi, wo, wu, ya, ye, yi, yo, yu.	ra, re, ri, ro, ru, rwa, rwe, rwi, rwo, rwu, la, le, li, lo, lu, lwa, lwe, lwi, lwo, lwu, wa, we, wi, wo, wu, ya, ye, yi, yo, yu.	30	11.9%
Vowels	a, e, i, o, u.	a, e, i, o, u.	5	2%
Total			**253**	**100%**

The study then uses Chargrams software tool of Oxford Wordsmith Tools version 7 (Scott, 2017) to analyse the list of potential Setswana syllables against a frequency list generated from a corpus. For the experiment the study uses 59,782 tokens/words. It checks and calculates the occurrence of each of the syllable on the word list extracted.

6. RESULTS

Below we discuss the different results from the study. First, the most frequent syllables and their frequencies in Setswana are presented. They reveal the kind of syllables that Setswana prefers in its word formation processes. Table 10 presents the most frequent 20 Setswana syllables. The most frequent 100 Setswana syllables are listed on Appendix 1.

Table 10. *Twenty most frequent Setswana syllables*.

Syllable	Freq.				
le	12536	mo	5925	ka	4498
ng	10974	i	5868	tse	4393
di	9471	n	5792	ba	4182
la	7978	se	5767	ne	4055
ma	7084	bo	5531	a	3572
lo	6343	na	5129	ko	3268
		ga	4637	me	3220

The results show that *le, ng, di, la, ma, lo, mo, i, n, se, bo, na, ga, ka, tse, ba, ne, a, ko,* and *me* are the most commonly used Setswana syllables across the various three positions of word initial, medial and word final. Most of these are digraphs while only three /*i, n* and *a*/ are single phoneme syllables. /*ng*/ is always syllabic in word final position and occurs as an onset of a syllable in word initial and word medial positions.

The study also calculated the most frequent syllables at the different word positions of word initial, word medial and word final. The results are presented on Tables 11, 12 and 13.

Table 11. *Twenty most frequent initial syllables*.

Syllable	Frequency				
di	4211	le	2668	kga	890
bo	3899	ba	2214	kgo	838
ma	3666	n	1968	ga	829
i	3581	m	1459	ko	822
mo	3288	a	1222	ka	783
se	3067	tlha	1041	ra	761
		me	928	tlho	729

Table 12. *Twenty most frequent medial syllables.*

Syllable	Frequency				
le	6843	ma	2885	a	1994
di	4116	ga	2604	me	1983
lo	3876	e	2437	ra	1907
n	3793	mo	2204	te	1743
la	3494	i	2077	ne	1692
ka	3000	se	2061	si	1645
		ko	2033	tse	1565

Table 13. *Twenty most frequent syllables in word final position.*

Syllable	Frequency				
ng	10848	lo	1779	lwa	907
la	4083	tsa	1316	ka	715
na	3475	ga	1204	go	692
le	3025	di	1144	lwe	666
tse	2572	wa	1086	se	639
ne	2229	tswe	1014	ra	547
		sa	978	tso	541

6.1 POTENTIAL UNATTESTED SYLLABLES IN THE CORPUS

Since this study started with potential syllables by considering a group of Setswana consonants and pairing them with Setswana vowels to create Setswana potential syllables. This study found out that there are certain consonant and vowel combinations which were not found in the entire word list studied. The thirty two (32) syllables below were not found as part of any Setswana word.

1. gwi
2. gwo
3. gwu
4. jwi
5. kgu
6. kgwo
7. kwo
8. lwo
9. lwu
10. nwo
11. nyu
12. rwi
13. rwo
14. šu
15. šwi
16. swo
17. thwi
18. tlhu
19. tlhwo
20. tlwo
21. tlwu
22. tše
23. tšhwi
24. tshwo
25. tši
26. tšo
27. tšu
28. tswo
29. tswu
30. two
31. wo
32. wu

Of the 32 syllables, 25 are labialised consonants, while 5 are followed by the close back vowel /u/. The absence of these syllables in Setswana words means that Setswana words are formed from an inventory of 221 syllables. However, with Setswana coming into contact with various languages and borrowing from such languages, it is conceivable that a syllable inventory from which Setswana words are formed will continue to increase. It also appears that any vowel produced with lip rounding such as /u/, /o/ and /ɔ/ is not preceded by the semi-vowel /w/ which is itself produced with strong lip-rounding.

6.2 SYLLABLES THAT NEVER OCCUR IN INITIAL, MEDIAL AND FINAL WORD POSITIONS

The next set of tests was to determine which syllables were never found in word initial position, word medial and word final positions. This was to determine if Setswana has a preference for syllable occurrence in certain positions or whether all Setswana syllables could occupy any position. Twenty four syllables were never found in word initial orthographic position though they were found in either mid or word final positions of Setswana orthographic words. These are:

1. tɬʰi (tlhi)
2. ɲi (nyi)
3. li
4. lu
5. tɬʷi (tlwi)
6. ʤi (ji)
7. ŋe (nge)
8. de
9. qʰi (kgi)
10. ʧʷa (tšwa)
11. ŋe (ngi)
12. ɲʷe (nywe)
13. tɬu (tlu)
14. wi
15. ʤu (ju)
16. ŋʷi (ngwi)
17. yi
18. kʰʷi (khwi)
19. ʧʷe (tšwe)
20. lʷi (lwi)
21. nʷi (nwi)
22. tʷi (twi)
23. ɲʷa (nywa)
24. ɲʷi (nywi)

Certain Setswana syllables were never found in word medial position of Setswana words. There are eight of such syllables:

1. kgwi
2. šwe
3. tšhwe
4. šwa
5. khwe
6. wi
7. yi
8. khwi

This study also found that eighteen syllables were never found in word final position.

1. m
2. n
3. tlwi
4. kgi
5. da
6. ngi
7. šo
8. khwa
9. tshwi
10. tšho
11. ju
12. yu
13. gu
14. nywa
15. nywi
16. kgwi
17. tšhwe
18. khwe

These results demonstrate that although Setswana forms its words from an inventory of 221 syllables, there are restrictions on where some syllables occur in Setswana words. Some never occur in word initial, word medial or word final position.

The following are syllables which are very rare in Setswana. By rare it is meant those syllables which occurred less than ten times in the corpus.

Table 14. *Rare Setswana syllables*.

Syllable	freq.	lu	6	ngu	3
khi	9	nywe	6	ngwi	3
nu	9	thwe	6	nwi	3
tsu	9	tšhi	6	šwe	3
tšwa	9	kha	5	tlu	3
wi	9	swi	5	tšhu	3
tshu	8	tlhwe	5	tšwe	3
tshwa	8	yi	5	do	2
hi	7	khe	3	gu	2
phu	7	khwi	3	kho	2
ši	7	lwi	3	twi	2
tša	7	nge	3	de	1

This study also computed the frequencies of consonants that form a Setswana onset. The results are presented on Table 15.

Table 15. *No of syllables that start with certain consonants*.

Articulation manner	Consonantal onsets	Freq.	%
Voiced Plosive	b, d	23,717	11.8%
Aspirated Voiceless Plosive	p^h, t^h, k^h, k^{hw}, t^{hw}, q^h, q^{hw}	12,300	6.1%
Non-Aspirated Voiceless Plosives	p, t, t^w, k, k^w	31,255	15.6%
Aspirated Voiceless Affricates	ts^h, $tɬ^h$, ts^{hw}, $ʧ^h$, $ʧ^{hw}$	8,867	4.4%
Non-Aspirated Voiceless Affricates	ts, ʧ, tɬ	13,042	6.5%
Fricatives	f, χ, $χ^w$, s, s^w, ʃ, $ʃ^w$	15,555	7.8%
Voiced Affricate	ʤ, $ʤ^w$	964	0.5%
Nasals	m, n, ɲ, n^w, ŋ, $ŋ^w$, $ɲ^w$	53,109	26.5%
Sonorants	r, r^w, l, l^w, w, y	41,487	20.7%
Total		**200,296**	**100%**

By far, most Setswana syllable onsets are plosives. Setswana plosives account for 33.5% of the syllable onsets while nasals account for 26.5% of the Setswana syllables onsets. This means that 60% of Setswana onsets are plosives and nasals. 20.7 of the syllables onsets are sonorants while 11.4% of the syllables onsets are affricates. Fricatives are rare as syllable onsets. They account for only 7.8% of all the syllable onsets. The evidence is therefore that there is a clear preference for plosives and nasals in the syllable onset position.

6.3 SYLLABLE NUCLEUS

Since the Setswana syllable can take any one of the following structures: CV, V or syllabic-consonant, the study then analysed the CV kind of syllables to determine how distributed their vowel endings are. The analysis calculated the number of syllables that ended with each one of the vowels. The results are

presented on Table 16. Our findings are that most syllables end with /a/ while fewer end with /u/. This result is consistent with the results of Otlogetswe (2016) which demonstrated that /a/ is the most frequent vowel in Setswana while /u/ is the least frequent. The total on Table 16 refers to the addition of all syllables that end with a certain vowel. For instance Table 16 demonstrates that there are 50 syllables that end with /a/ and that such syllables have a frequency of 72,118.

Table 16. *Syllables that end with a certain vowel*.

Vowel ending	Total	Frequency
a	50	72,118
e	49	57,483
o	28	43,250
i	41	28,714
u	24	8,666
Total	**192**	**210,231**

6.4 AN ANALYSIS OF SETSWANA SYLLABIC CONSONANTS

Syllabicity is not only a vowel quality in Setswana. Setswana has five syllabic consonants which are /m, n, r, ŋ and l/. As stated previously, this study doesn't consider the syllabic /l/ since it is never found in formal written text and is more pronounced in Sekgatla dialect speech. Only four syllabic consonants are studied. These are *m, n, r* and *ng*. The statistics for the consonants are presented on Table 17.

Table 17. *Frequency results of syllabic consonants*.

Syllable	Initial	Medial	Final	TOTAL
ng	577	560	10848	11,985
n	1897	3050	0	4947
m	1398	724	0	2122
r	128	98	0	226
Total	**3423**	**3872**	**10848**	**19,280**

The most frequent syllabic consonant is /ŋ/ which predominantly occurs at the end of verbs and nouns (to form adverbials). While /ŋ/ is the most frequent syllabic consonant, it occurs mostly at word final position where the other syllabic consonants never occur. /n/ is the most frequent of all syllabic consonants in word initial and medial positions.

6.5 TESTING LINGUISTIC CLAIMS

With Setswana borrowing words from other languages, the assumptions regarding which sounds form part of the inventory of the Setswana phonology

must be constantly under review as the language grows. This is so that linguistic analysis reflects language phonotactics without bias. The study of syllables allows us to test certain linguistic claims using fairly large text. In this section two claims are tested. The first one is that the lateral phoneme /l/ is realised as [d] whenever followed by the high tense vowels /i/ and /u/. In the present orthography, /l/ is spelt as *d* before these vowels (DALL, 1999:14). This claim is repeated elsewhere,

> In Setswana, when the phoneme /l/ is followed by the vowels /i/ and /u/, it becomes [d]… In this case, we can state that the sounds [l] and [d] belong to the same class in Setswana. While the sound [l] appears before the vowels [e] [ʊ], [ɛ], [ɔ] and [a], the sound [d] occurs before the vowels [i] and [u] (DALL, 1999:3).

The frequency analysis of Setswana supports this claim since the syllable *la* occurs 28,121 times, *le* occurs 45,123 and *lo* 23,180. However there is evidence of Setswana syllables where the [l] precedes [i] and [u] as well as where [d] precedes [*e*, *o* or *a*] as shown in the data that follows. The list of exceptions is composed of borrowings from English, Afrikaans or other African languages such as Kalanga.

Words where [d] precedes [o]

1. *lephondo* (large hair puff hair style; from Xhosa ‘iphondo’)
2. *domi* (a nickname for the Botswana Democratic Party; from Afrikaans ‘domkrag’ meaning ‘a jack’)
3. *madongwana* (nickname for members of the Botswana Democratic Party, from Afrikaans ‘domkrag’ meaning ‘a jack’.
4. *ledombi* (from English ‘dumbling’)
5. *khondomo* (from English ‘condom’)

Words where [d] precedes [a]

1. *khondae* (from English ‘conductor’)
2. *didatse* (from English ‘darts’)
3. *tandabala* (old age pension money; from Kalanga ‘tandabala’)

Words where [d] precedes [e]

1. *poresidente* (from English ‘president’)
2. *modemone* (from English ‘demons’)
3. *demokerasi* (from English ‘democracy’)

Words where [l] precedes [i]

1. *lithara* (from English ‘litre’)
2. *Bafilipi* (from English ‘Philippians’)
3. *Bafilisitia* (from English ‘Philistines’)
4. *baselini* (from English ‘vaseline’)

Words where [l] precedes [u]

1. *Lutere* (from English 'Lutheran')
2. *bolumara* (from English 'bloomers')
3. *baluni* (from English 'baloon')
4. *folutu* (from English 'flute')
5. *dihaleluja* (from English 'hallelujah')
6. *dithulusu* (from English 'tools')
7. *diulu* (from English 'wool')
8. *malutu* (yellow sorghum powder from English 'multi-(vitamin)')
9. *saluni* (from English 'salon')
10. *saluti* (from English 'salute')

The second claim that this paper tests is that "there is no [g] sound in Setswana, hence the grapheme <g> is used to represent /χ/" (DALL, 1999:15).

While in general <g> in Setswana is pronounced as /χ/, because of borrowing, <g> is now pronounced in a limited number of loanwords, as /g/ in words such as *mmengu* /m̀míŋ̀χu/ 'mango', *legwinya* /lìgwìŋá/ 'fat cake', *geiti* /geiti/ 'gate' and *digumbagumba* /digumbagumba/ 'loud music stereo'. The <g> in these words is not pronounced as /χ/ but as /g/. These borrowings are also colloquial and informal and therefore rarely occur in written text.

A similar occurrence has been observed in relation to /z/ which is not a recognised Setswana phoneme. However, a number of words with /z/ have been borrowed from other African languages such as Zulu and now appear on the frequency list of the corpus. Amongst these are *zama* (try), *nzamela* /ǹzàmɛ́là/ 'try for me; get me airtime', *bazelwana* /bàzɛ̀lʷánà/ 'Christian believers' and *banuza* /bànúzà/ 'girls'. While these words are informal and colloquial, they are Setswana words which need to be accounted for within Setswana phonology.

7. FINDINGS AND CONCLUSIONS

The aim of this article was to provide a frequency analysis of different aspects of Setswana syllable inventory using a corpus of orthographic Setswana words. Although there are potential limitations in terms of the imperfect correspondence between orthographic and phonological representations, the methodology was adopted because it makes it possible to study large number of words (over 13 million) computationally. It is hoped that this study, regardless of its limitations, sheds some light on the syllable distribution of Setswana. The study revealed that the most frequent syllables in Setswana are: *le, ng, di, la, ma, lo, mo, i, n, se, bo, na, ga, ka, tse, ba, ne, a, ko,* and *me.* It also investigated the frequency of Setswana syllable onsets and established that the most frequent word initial syllables are: *di, bo, ma, i, mo, se, le, ba, n, m, a, tlha, me, kga, kgo, ga, ko, ka, ra,* and *tlho*. The most frequent word medial syllables were established to be: *le, di, lo, n, la, ka, ma, ga, e, mo, i, se, ko, a, me, ra, te, ne, si,*

and *tse*. It has also been established that the most frequent word final syllables are: *ng, la, na, le, tse, ne, lo, tsa, ga, di, wa, tswe, sa, lwa, ka, go, lwe, se, ra,* and *tso*. This study developed a list of what it called *potential syllables* based on the combinations of all Setswana's consonantal phonemes paired with the five Setswana orthographic vowels to form 253 potential syllables. The study found that the *potential* syllables which are not part of Setswana syllable inventory are: *gwi, gwo, gwu, jwi, kgu, kgwo, kwo, lwo, lwu, now, nyu, rwi, rwo, šu, šwi, swo, thwi, tlhu, tlhwo, tlwo, tlwu, tše, tšhwi, tshwo, tši, tšo, tšu, tswo, tswu, two, wo* and *wu*. This leaves Setswana with an inventory of 221 attested syllables in the corpus under study.

The study also tested for the occurrence of all Setswana syllables in word initial, medial and final. It established that syllables that don't occur in word-initial position are: *tlhi, nyi, tlwi, ji, nge, de, kgi, tšwa, ngi, nywe* and *tlu*. Syllables that don't occur in word-medial position are: *kgwi, šwe, tšhwe, šwa, khwe, wi, yi,* and *khwi* while syllables that don't occur in word-final position are: *m, tlwi, kgi, da, ngi, šo, khwa* and *tshwi*. The inspection of the data also made it possible to test certain linguistic claims (such as the /l/ and /d/ allophony). It was also determined that most Setswana syllables end in /a/ while fewer end in /u/. /ng/ was found to be the most common syllabic consonant while /r/ was the least common. The importance of this study lies in identifying an inventory of Setswana syllables and how they are distributed in the various Setswana words. The findings of this study will therefore prove to be important to other related studies in psycholinguistics and computational linguistics. The paper contributes to a better understanding of Setswana phonemes and syllable inventory.

REFERENCES

Batibo, H. M. 2013.
The evolution and adaptation of Swahili and Tswana syllable structures. In: K. Legère (ed.), *Bantu Languages and Linguistics: Papers in Memory of Dr Rugatiri D. K. Mekacha,* **Bayreuth African Studies** 91: 13–34. Bayreuth: Bayreuth University Press.

Cole, D. 1955.
An Introduction to Tswana Grammar. Longman. Johannesburg.

DALL (Department of African Languages and Literature) University of Botswana. 1999.
The Sound System of Setswana. Lightbooks. Gaborone.

Gouskova, M., Zsiga, E., Tlale-Boyer, O. 2011.
Grounded constraints and the consonants of Setswana. **Lingua** 121: 2120–2150.

Kessler, B. and Treiman R. 1997.
Syllable structure and the distribution of phonemes in English Syllables. **Journal of memory and language** 37: 295–311.

Munthuli, C. Tantibundhit, C. Onsuwan, K. Kosawat, and C. Wutiwiwatchai. 2015.
Frequency of Occurrence of Phonemes and Syllables in Thai: Analysis of Spoken and Written Corpora. In: *Proceedings of the 18th International Congress of Phonetic Sciences,* The Scottish Consortium for ICPhS 2015 (Ed.), Glasgow: The University of Glasgow. Paper number 1041.1–9 retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS1013.pdf

Otlogetswe T. J. 2016.
The design of Setswana Scrabble. **South African Journal of African Languages** 36(2): 153–161.

2011 *Text variability measures in corpus design for Setswana lexicography*. Cambridge: Cambridge Scholars Publishing.

Scott, M. 2017.
Oxford Wordsmith Tools. Stroud: Lexical Analysis Software.

About the author: *Thapelo J. Otlogetswe* is Associate Professor of linguistics and lexicography at the University of Botswana. He is a recipient of The Presidential Order of Honour and the foremost expert on corpus linguistics and lexicography of the Setswana language. His research is in lexical computing and corpus lexicography, in particular, that of Setswana. His research includes work on Setswana names, rhyming patterns in Setswana, genre and text type analysis. He has compiled a number of dictionaries including "Tlhalosi ya Medi ya Setswana", "English-Setswana Dictionary", "Oxford English Setswana Setswana English School Dictionary" and 'Poeletso-medumo ya Setswana: a Setswana Rhyming dictionary'. He is a member of the African Academy of Languages, African Association of Lexicography and sits on the editorial boards of Lexikos and Marang Journals.

APPENDIX 1: THE MOST REQUENT 100 SETSWANA SYLLABLES

Syllable	**Total**
le	12536
ng	10974
di	8330
la	7978
ma	7084
lo	6343
mo	5925
i	5868
n	5792
se	5767
bo	5531
na	5129
ga	4637
ka	4498
tse	4393
ba	4182
ne	4055
a	3572
ko	3268
me	3220
ra	3215
e	3047
go	2886
tsa	2704
m	2657
tlha	2608
te	2545
si	2440
sa	2356
pa	2330
be	2164
ke	2162
nya	2066
ta	2002
pe	1995
o	1972
po	1972
kga	1947
tlho	1933
re	1896
kgo	1808
ge	1769
ro	1757
tswe	1746
ti	1645
tla	1589
wa	1555
lwa	1546
fa	1530
to	1514
no	1446
so	1202
ri	1177
ki	1161
tso	1156
tshe	1128
tha	1121
gi	1079
nye	1072
pha	1013
fe	990
tlo	970
pi	962
tsi	959
lwe	951
bi	894
du	876
mi	858
thu	856
ru	845
tle	837
u	826
the	754
tsho	734
tswa	729
kwa	715
tho	707
tshwa	686
ku	676
fo	654
we	643
tsha	614
tu	592
tshi	574
ngwa	552
phe	549
su	547
bu	537
khu	534
pho	492
tlhe	492
pu	491
fi	470
ja	447
hu	424
nga	415
thi	405
phu	401
ni	343
swa	341
phi	322
gwa	315
rwa	315
ya	311
nyo	304
kge	296
fu	291
ngwe	262
gwe	244
ha	238
tlwa	238
ngo	226
kgwa	224
kwe	221
tli	219
mu	217
tshu	216
je	205
twa	198
he	184
tšhwa	179
tsu	177
kgwe	168
kha	168
tshwe	167
swe	165
nwa	158
hi	156
tlhi	136
jwa	129
nyi	118
tša	118
kho	117
še	117
li	116
nwe	116
ša	110
ho	99
khi	99
thwa	92
jo	89
rwe	84
tlwe	83
ye	80
tswi	75
twe	71
tlhwa	62
yo	57
jwe	54
tše	51
khe	48
lu	48
ši	48
tlwi	46
do	39
tši	33
ji	32
nge	31
kwi	30
tšho	30
tša	27
nu	26
de	24
ngu	23
da	21
kgi	21
šo	19
thwe	19
ngi	18
tšwa	18
yu	18
khwa	17
šwa	16
tšhu	16
tshwi	16
swi	14
kgwi	13
nywe	13
tlhwe	13
tlu	12
ju	9
šwe	9
wi	9
gu	8
ngwi	7
tšhwe	5
yi	5

khwi	4
tšwe	4
khwe	3
lwi	3
nwi	3
nywa	3
nywi	3
twi	3
gwi	0
gwo	0
gwu	0

jwi	0
kgu	0
kgwo	0
kwo	0
lwo	0
lwu	0
nwo	0
nyu	0
rwi	0
rwo	0
šu	0

šwi	0
swo	0
thwi	0
tlhu	0
tlhwo	0
tlwo	0
tlwu	0
tše	0
tšhwi	0
tshwo	0
tši	0

tšo	0
tšu	0
tswo	0
tswu	0
two	0
wo	0
wu	0
TOTAL	229619

APPENDIX 2: FREQUENCY OF SYLLABLES THAT OCCUR IN WORD INITIAL POSITION

Syllable	**Initial**
di	4211
bo	3899
ma	3666
i	3581
mo	3288
se	3067
le	2668
ba	2214
n	1968
m	1459
a	1222
tlha	1041
me	928
kga	890
kgo	838
ga	829
ko	822
ka	783
ra	761
tlho	729
go	704
lo	688
be	444
pha	444
pa	429
o	425
thu	414
po	413
te	403
la	401
re	397
tha	395
ke	352
ro	345
tshe	340
e	338
so	326
si	322
ta	312
fe	307
ti	306
su	297
na	289
tla	282
tlo	276
to	271
tsho	266
ru	260
tse	256
fa	248
tshwa	246
pe	242
fo	238
phu	235
bu	233
khu	233
ku	230
u	225
the	225
hu	214
tsa	210
du	203
phe	202
sa	194
tu	191
pho	188
tho	185
no	177
tsi	175
tso	174
nga	169
tshi	164
bi	163
tsha	162
ne	134
fu	133
kwa	131
fi	130
thi	129
pi	126
kge	121
ki	115
pu	115
nye	114
ngo	104
phi	101
tswe	98
nya	94
tshu	91
kgwa	90
tswa	87
kgwe	83
kha	80
swa	79
ja	78
tsu	77
ngwa	76
ri	74
je	74
tšhwa	74
ha	71
he	67
tshwe	67
kho	63
hi	56
tlwa	51
tle	45
swe	43
ša	43
lwa	40
we	39
kwe	38
ho	38
gwe	37
rwa	36
tli	36
khi	36
ge	35
še	35
jo	35
gwa	33
tša	27
thwa	26
nyo	25
mi	24
jwa	22
rwe	22
nwe	21
tšho	21
wa	20
tlhwa	19
tswi	18
mu	17
khe	17
ng	15
ngwe	15
tlhe	13
yo	13
jwe	13
kgwi	13
tšhe	12
ni	11
lwe	10
do	10
ngu	10
yu	10
tši	9
twa	8
ši	7
gi	6
nwa	6
tlwe	6
kwi	6
šo	6
šwe	6
ya	5
twe	5
tša	5
thwe	5
tšhu	5
tshwi	5
tšhwe	5
khwa	4
swi	4
tlhwe	4
ye	3
nu	3
da	3
šwa	3
gu	3
khwe	3
tlhi	0
nyi	0
li	0
lu	0
tlwi	0
ji	0
nge	0
de	0
kgi	0
ngi	0
tšwa	0
nywe	0

tlu	0
ju	0
wi	0
ngwi	0
yi	0
khwi	0
tšwe	0
lwi	0
nwi	0
nywa	0
nywi	0
twi	0

gwi	0
gwo	0
gwu	0
jwi	0
kgu	0
kgwo	0
kwo	0
lwo	0
lwu	0
nwo	0
nyu	0
rwi	0

rwo	0
šu	0
šwi	0
swo	0
thwi	0
tlhu	0
tlhwo	0
tlwo	0
tlwu	0
tše	0
tšhwi	0
tshwo	0

tši	0
tšo	0
tšu	0
tswo	0
tswu	0
two	0
wo	0
wu	0
TOTAL	58793

APPENDIX 3: FREQUENCY OF SYLLABLES THAT OCCUR IN WORD MEDIAL POSITION

Syllable	**Medial**
le	6843
di	4116
lo	3876
n	3793
la	3494
ka	3000
ma	2885
ga	2604
e	2437
mo	2204
i	2077
se	2061
ko	2033
a	1994
me	1983
ra	1907
te	1743
ne	1692
si	1645
tse	1565
ba	1554
ke	1518
go	1490
pa	1477
pe	1475
be	1454
nya	1452
bo	1448
ge	1402
na	1365
po	1335
tlha	1334
o	1222
ta	1211
m	1198
re	1197
sa	1184
tsa	1178
fa	1175
tlho	1163
ro	1158
ti	1092
kga	1005
tla	980
to	968
kgo	943
gi	908
ri	850
ki	802
no	789
mi	732
nye	699
tle	660
pi	659
tshe	651
bi	639
tswe	634
fe	610
tha	602
lwa	599
tlo	574
du	533
ru	522
pha	498
so	472
kwa	454
the	450
wa	449
tso	441
u	437
tshwa	432
thu	426
tsi	404
tho	403
ku	390
tlhe	386
tswa	372
tshi	356
tu	349
tsho	347
fo	346
pu	341
ja	326
phe	313
tsha	310
ngwa	292
khu	281
pho	279
lwe	275
fi	267
bu	243
nga	223
thi	219
su	217
rwa	206
hu	198
phi	198
gwa	191
mu	181
tli	165
ni	165
kge	160
phu	159
ha	151
ya	149
tlwa	140
swa	135
fu	134
we	127
kgwa	119
tshu	117
twa	117
ng	111
ngo	110
tlhi	110
nyo	108
he	105
je	101
li	99
hi	93
tsu	91
kha	83
swe	82
gwe	80
tshwe	79
kwe	79
ngwe	79
tša	77
kgwe	70
še	65
nwa	65
jwa	64
khi	54
kho	52
tlwe	52
tšhwa	49
ša	48
ho	48
thwa	46
tswi	46
tlwi	46
nwe	42
lu	42
ši	34
jo	32
ye	32
rwe	29
khe	28
nge	28
yo	27
nyi	27
do	26
twe	26
de	21
kgi	21
tšhi	18
da	18
ji	18
ngi	18
tlhwa	17
tša	15
tše	13
šo	13
khwa	13
nu	13
kwi	12
tshwi	11
ngu	10
tšho	9
tšwa	9
tlu	9
ju	9
jwe	8
yu	8
thwe	8
tšhu	8
nywe	7
swi	5
gu	5
tlhwe	4
ngwi	4
nywa	3
nywi	3

kgwi	0
šwe	0
tšhwe	0
šwa	0
khwe	0
wi	0
yi	0
khwi	0
tšwe	0
lwi	0
nwi	0
twi	0

gwi	0
gwo	0
gwu	0
jwi	0
kgu	0
kgwo	0
kwo	0
lwo	0
lwu	0
nwo	0
nyu	0
rwi	0

rwo	0
šu	0
šwi	0
swo	0
thwi	0
tlhu	0
tlhwo	0
tlwo	0
tlwu	0
tše	0
tšhwi	0
tshwo	0

tši	0
tšo	0
tšu	0
tswo	0
tswu	0
two	0
wo	0
wu	0
TOTAL	115098

APPENDIX 4: FREQUENCY OF SYLLABLES THAT OCCUR IN WORD FINAL POSITION

Syllable	**Final**
ng	10848
la	4083
na	3475
le	3025
tse	2572
ne	2229
lo	1779
tsa	1316
ga	1204
wa	1086
tswe	1014
sa	978
lwa	907
ka	715
go	692
lwe	666
se	639
ra	547
tso	541
ma	533
nya	520
no	480
ta	479
we	477
si	473
mo	433
pa	424
ba	414
ko	413
so	404
te	399
tsi	380
ge	367
a	356
tla	327
o	325
me	309
re	302
ke	292
pe	278
to	275
e	272
tswa	270
be	266
nye	259
ro	254
ri	253
ti	247
ki	244
tlha	233
po	224
i	210
bo	184
ngwa	184
pi	177
nyo	171
ngwe	168
ni	167
gi	165
u	164
ya	157
tsha	142
du	140
tshe	137
tle	132
kwa	130
swa	127
gwe	127
tha	124
tsho	121
tlo	120
tho	119
fa	107
kwe	104
mi	102
tlhe	93
bi	92
gwa	91
nyi	91
nwa	87
the	79
fe	73
fi	73
rwa	73
twa	73
pha	71
fo	70
ru	63
bu	61
thi	57
ku	56
tšhwa	56
tshi	54
nwe	53
kga	52
tu	52
tlwa	47
ye	45
ja	43
jwa	43
tlho	41
swe	40
twe	40
pu	35
phe	34
su	33
rwe	33
jwe	33
n	31
je	30
kgo	27
tlhi	26
tlhwa	26
tšhe	26
pho	25
tlwe	25
fu	24
nga	23
phi	23
jo	22
tshwe	21
khu	20
thwa	20
mu	19
ša	19
tli	18
li	17
še	17
yo	17
thu	16
ha	16
kge	15
kgwa	15
kgwe	15
tša	14
ji	14
ho	13
šwa	13
hu	12
ngo	12
he	12
kwi	12
tswi	11
nu	10
tsu	9
khi	9
tšwa	9
wi	9
tshwa	8
tshu	8
phu	7
hi	7
ši	7
tša	7
lu	6
tši	6
thwe	6
nywe	6
kha	5
swi	5
tlhwe	5
yi	5
khwi	4
tšwe	4
di	3
khe	3
nge	3
do	3
de	3
ngu	3
tlu	3
tšhu	3
ngwi	3
šwe	3
lwi	3
nwi	3
twi	3
kho	2
m	0
tlwi	0
kgi	0
da	0
ngi	0

šo	0
khwa	0
tshwi	0
tšho	0
ju	0
yu	0
gu	0
nywa	0
nywi	0
kgwi	0
tšhwe	0

khwe	0
gwi	0
gwo	0
gwu	0
jwi	0
kgu	0
kgwo	0
kwo	0
lwo	0
lwu	0
nwo	0

nyu	0
rwi	0
rwo	0
šu	0
šwi	0
swo	0
thwi	0
tlhu	0
tlhwo	0
tlwo	0
tlwu	0

tše	0
tšhwi	0
tshwo	0
tši	0
tšo	0
tšu	0
tswo	0
tswu	0
two	0
wo	0
wu	0