Developing a Sandhi Lexicon (SandhiLex) for Sinhala: Understanding and Formalizing Morphophonology of Sinhala Language

Chamila Liyanage and **Randil Pushpananda**

Language Technology Research Laboratory,
University of Colombo School of Computing,
Sri Lanka
{cml,rpn}@ucsc.cmb.ac.lk

Abstract

Sandhi, a grammatical feature in Sinhala inherited from Old Indo-Aryan, has been discussed in all Sinhala grammar books, beginning with *Sidat Saṅgarāva*, reportedly the first Sinhala grammar book. This paper presents a study of Sandhi in the Sinhala language and introduces a novel classification based on linguistic analysis. The study identifies three primary lexical units involved in sandhi formation and six lexical entries related to the Sandhi process. Based on this analysis, morphophonological variations in Sinhala are classified into four categories: Lexicalized Sandhi, Derivational Sandhi, Etymological Sandhi, and Affixational Sandhi. Accordingly, a Sandhi Lexicon (SandhiLex) for the Sinhala language was compiled using a semi-automatic method. The SandhiLex includes approximately 4,500 Sandhi lexemes for the Lexicalized Sandhi dataset and over 300k lexical units for the Affixational Sandhi dataset, contributing significantly to advancing research in Natural Language Processing for the Sinhala language.

1 Introduction

Sandhi refers to the process of phonological changes that occur at word boundaries. This particularly refers to the morphophonological changes occur at the point of joining two words or characters (Devadath et al., 2014). Sandhi, as a morphophonological phenomenon, is challenging in word boundary detection, leading to difficulties in many NLP tasks such as tokenization, morphological analysis, parts-of-speech tagging, and machine translation.

Sinhala, as an Indo-Aryan language, exhibits the morphophonological feature called Sandhi, making it particularly challenging for NLP tasks. Therefore, a treatment is required to address the recognition of word boundaries. Further, as this grammatical feature has been derived from Old Indo-Aryan phonology (Jain and Cardona, 2007), Sandhi has evolved into a complex grammatical phenomenon, with both historical and contemporary forms occurring in the language. Accordingly, a study of Sandhi in Sinhala is beneficial for understanding the language's phonological structure, linguistic evolution, and interaction between historical and contemporary forms. From a linguistic resource compilation perspective, De Silva (2019) notes that Sinhala is a low-resource language, requiring more language resources for many NLP tasks. However, no reported work has been carried out to develop a language resource for Sandhi in Sinhala language. Hence, this paper reports a study of Sandhi in Sinhala language and the process of developing a Sandhi lexicon for Sinhala.

Text processing tasks in agglutinative languages are not trivial for several reasons, one of which is the concatenation of multiple lexical entries into a single word. For instance, in the following example, වම් *vam* (left) and අත *ata* (hand) are two distinct words. They can be concatenated into a single word, a Sandhi: වමත *vamata* (lefthand), with only minor morphophonological changes.

e.g. වම් *vam* (left) + අත *ata* (hand)
වමත *vamata* (lefthand)

The challenge with Sandhi as a natural language phenomenon lies in the difficulty of recognizing word boundaries. For instance, පරිගණක *parigaṇaka* (computing), අධ්‍යයන *adhyayana* (studies) and ආයතනය *āyatanaya* (institute) are three distinct words in the Sinhala language, each corresponding to different lexical meanings. Figure 1 shows how these three Sinhala words can be arranged in four different structures while maintaining the same meaning.

As exemplified by the four lexical combinations in Figure 1, the same lexical entries can be presented in multiple ways, making it challenging to identify word boundaries and leading to several difficulties in language processing tasks. Accurately

i. පරිගණක අධ්‍යයන ආයතනය
parigaṇaka adhyayana āyatanaya

ii. පරිගණක අධ්‍යයනායතනය
parigaṇaka adhyayanāyatanaya

iii. පරිගණකාධ්‍යයන ආයතනය
parigaṇakādhyayana āyatanaya

iv. පරිගණකාධ්‍යයනායතනය
parigaṇakādhyayanāyatanaya

Figure 1: Four sequences using three lexical units to indicate the meaning 'Institute of Computer Studies'

identifying individual words within concatenated forms can enhance the effectiveness of tasks such as information retrieval, syntactic or grammatical parsing, machine translation, sentiment analysis, and linguistic annotation. Additionally, this complexity poses challenges for language learning and teaching.

Recognition of Sandhi formation can be analyzed through two primary methods: rule-based methods and machine learning methods. Despite the challenges in finding resource persons with relevant linguistic expertise, Priyanga et al. (2017) has attempted to develop a rule-based model of a Sinhala word joiner. However, the actual requirement lies in the opposite direction: recognizing word boundaries to segment Sandhi words. Although the machine learning approach would presumably be more accurate, no research has reportedly been conducted in this direction due to the lack of available datasets. Therefore, Priyanga et al. (2017) primarily focuses on implementing Sandhi rules found in the *Sidat Saṅgarāva*, a 13th-century text, without exploring modern linguistic methods that could be more beneficial. Consequently, the present research was conducted to understand Sandhi in Sinhala language and develop a Sandhi lexicon for the particular language.

2 Sandhi in Sinhala Language

Sinhala, an Indo-Aryan language, is one of the two official languages of Sri Lanka, spoken by the majority of the population, with about 20 million speakers worldwide. Sinhala has been in contact with Tamil, which belongs to the Dravidian language family, for a long time within the country. Due to colonization, Sinhala has also been influenced by Portuguese, Dutch, and English languages.

Jain and Cardona (2007) notes that Sandhi is a feature of Old Indo-Aryan (OIA) phonology. As Sinhala is an Indo-Aryan language, a sub-branch of the Indo-European language family, grammatical features of OIA have been inherited by the language. Thus, Sandhi is one of the major grammatical features discussed in every grammar book since the *Sidat Saṅgarāva*, that became a reference for all subsequent grammar books, such as Gunasekara (1891); Gunawardhana (1924); and Thilakasiri (1997).

Given the complexity of Sandhi as a grammatical phenomenon in Sinhala, it has not only been discussed as a topic in traditional grammar books but has also been the subject of separate works. Several books have been written on Sandhi, including Coperahewa (2014), a compilation of a dictionary of Sandhi words in Sinhala; Ekanayake (2016), an analysis of the Sandhi phenomenon in Sinhala, particularly with reference to Old Sinhala; and Disanayaka (1997), an analysis based on (a kind of) structural linguistics.

2.1 Classification of Sandhi

In the literature, Sandhi in the Sinhala language has been classified based on three criteria: i. morphophonological process, ii. occupying lexical units and iii. diglossic variants.

2.1.1 Morphophonological Process

The *Sidat Saṅgarāva* has classified Sandhi into nine categories based on morphophonological functions. Although the term Sandhi is now commonly used in English, it was referred to as 'Permutation' (*Pt*) in De Alwis (1852), an English translation of the *Sidat Saṅgarāva*. The 9 classes are mentioned below.

i. *Pt* by the elision of the first vowel

ii. *Pt* by the elision of the second vowel

iii. *Pt* of vowels

iv. *Pt* by substitution of vowels

v. *Pt* by substitution of consonants

vi. *Pt* by reduplication of first letter

vii. *Pt* by elision

viii. *Pt* by substitution

	Sandhi	Segmented
i.	අත්‍යන්ත *atyanta* (Absolute)	අති + අන්ත *ati* + *anta*
ii.	අභ්‍යන්තර *abhyantara* (internal)	අභි + අන්තර *abhi* + *antara*
iii.	නිරාහාර *nirāhāra* (Starving)	නිර් + ආහාර *nir* + *āhāra*
iv.	නුදුටු *nuduṭu* (unseen)	නො + දුටු *no* + *duṭu*
v.	මිනිසෙක් *minisek* (a man)	මිනිස් + එක් *minis* + *ek*
vi.	පොතේ *potē* (in the book)	පොත + ඒ *pota* + *ē*
vii.	පොතෙන් *poten* (from the book)	පොත + එන් *pota* + *en*

Table 1: Examples of lexical units for internal Sandhi

ix. *Pt* by reduplication of letters

In Gunawardhana (1924), the author analyzes the classification presented in *Sidat Saṅgarāva*. Considering the nuances of phonological processing, he offers his own analysis of Sandhi classes, expanding the nine categories found in *Sidat Saṅgarāva* to a total of fifteen classes.

As Sandhi is a common grammatical phenomenon in Indo-Aryan languages, Allen (1972) has classified Sandhi in Sanskrit into five distinct classes: i. Vowel + Vowel, ii. Vowel + Consonant, iii. Consonant + Vowel, iv. Consonant + Consonant, and v. Terminal Sandhi. For Sinhala Meegaskumbura (2020) identifies only (first) four classes, omitting the fifth class, Terminal Sandhi.

2.1.2 Occupying Lexical Units

Gunawardhana (1924) and subsequently Kumaranathunga (1937) have classified Sandhi into two categories based on the occurrence of lexical units in the Sandhi process: (i) Internal Sandhi and (ii) External Sandhi.

(i) **Internal Sandhi**

Internal Sandhi refers to morphophonemic changes that occur within a stem or when a stem is joined with an inflectional affix (Gunawardhana, 1924; Kumaranathunga, 1937). These changes can involve all types of affixes, including suffixes and prefixes (with the note that Sinhala does not use infixes). This method of Sandhi formation leads to a large set of new lexical entries in the language. While suffixes typically lead to inflections, prefixes often result in derivations, which are generally included as separate lemmas in dictionaries as depicted in Table 1.

	Sandhi	Segmented
i.	අංගෝපාංග *aṁgōpāṁga* (components)	අංග + උපාංග *aṁga* + *upāṁga* (element) + (accessories)
ii.	උත්තමායුෂ *uttamāyuṣa* (highest age)	උත්තම + ආයුෂ *uttama* + *āyuṣa* (highest) + (age)
iii.	කලායතනය *kalāyatanaya* (art institute)	කලා + ආයතනය *kalā* + *āyatanaya* (art) + (institute)
iv.	නීත්‍යනුකූල *nītyanukūla* (legal)	නීති + අනුකූල *nīti* + *anukūla* (law) + (compliant)
v.	ලේඛනාගාර *lēkhanāgāra* (archives)	ලේඛන + ආගාර *lēkhana* + *āgāra* (records) + (house)

Table 2: Examples of lexical units for external Sandhi

(ii) **External Sandhi**

External Sandhi occurs between either two stems or two words (Gunawardhana, 1924; Kumaranathunga, 1937). For instance, all the lexical entries in Table 2 are distinct words. Significantly, both the Sandhi words and their segmented components appear as separate lemmas in Sinhala dictionaries.

2.1.3 Diglossic Variants

Sandhi, as a natural language phenomenon, can occur in both spoken and written aspects of a language. In the spoken aspect of the Sinhala language, පොත් ටික *pot ṭika* (the small set of books) becomes පොට්ටික *poṭṭika*, and බත් ඵුට්ටක් *bat cuṭṭak* (a small amount of rice) becomes බච්චුට්ටක් *baccuṭṭak*, indicating morphophonological changes at the point where two morphemes join. However, Sandhi in spoken language is not of much concern, since lexical entries with these particular morphophonological changes, such as පොට්ටික *poṭṭika* or බච්චුට්ටක් *baccuṭṭak*, do not typically occur in written form. Consequently, they do not appear in text corpora and do not pose

significant challenges in Sinhala language computing tasks.

2.2 New Classification of Sandhi

As discussed in Section 1 Sandhi refers to a morphophonological process that occurs in several instances. There are three primary lexical units involved in the formation of Sandhi in Sinhala: noun lemmas, prefixes, and suffixes. However, there are six lexical entries involved in the Sandhi process in Sinhala, as illustrated below.

i. Lemma [L]
Noun lemmas are the most frequently used lexical units in the formation of Sandhi words.
e.g. වම *vama* (left), දකුණ *dakuna* (right), අත *ata* (hand)

ii. Prefix I [P1]
In the formation of Sandhi in Sinhala, prefixes can be classified into two categories, with the first category containing prefixes that generate new lemmas.
e.g. නිර් *nir*, සත් *sat*, අති *ati*

iii. Prefix II [P2]
The second category of prefixes includes those that do not generate new lemmas in the formation of Sandhi words.
e.g. නො *no*

iv. Suffixes [S]
In the agglutinative process, adding suffixes to a particular word may cause morphophonemic changes. Thus, suffixes can be recognized as one of the lexical units involved in Sinhala Sandhi formation.
e.g. ඉන් *in*, එන් *en*, එහි *ehi*

v. Unchanged Lemma [UL]
After the concatenation of lexical entries, some Sandhi words remain lemma unchanged. In other words, these Sandhi words do not appear as lemmas in dictionaries.
e.g. ඔවුනොවුන් *ovunovun* (each other), වමත *vamata* (left hand)

vi. New Lemma [NL]
After the concatenation process, certain Sandhi words acquire new meanings and appear as new lemmas in dictionaries.
e.g. අභ්‍යන්තර *abhyantara* (internal), කලායතනය *kalāyatanaya* (art institute)

Sandhi words in Sinhala are formed by combining two or more lexical units from the first four of the six lexical types mentioned above. Analysis reveals five possible types of concatenation using these categories.

- $\mathrm{L} + \mathrm{L} = \mathrm{UL}$
- $\mathrm{L} + \mathrm{L} = \mathrm{NL}$
- $\mathrm{P1} + \mathrm{L} = \mathrm{NL}$
- $\mathrm{P2} + \mathrm{L} = \mathrm{UL}$
- $\mathrm{L} + \mathrm{S} = \mathrm{UL}$

Accordingly, Sandhi can be identified as a morphophonological process that occurs in several instances. Based on these occurrences, we classify Sinhala Sandhi words into four classes:

i. Lexicalized Sandhi ($\mathrm{L}+\mathrm{L} = \mathrm{UL}$)

ii. Derivational Sandhi ($\mathrm{P1}+\mathrm{L} = \mathrm{NL}$)

iii. Etymological Sandhi ($\mathrm{L}+\mathrm{L} = \mathrm{NL}$)

iv. Affixational Sandhi ($\mathrm{P2}+\mathrm{L} = \mathrm{UL} \mid \mathrm{L}+\mathrm{S} = \mathrm{UL}$)

These four distinct categories are discussed below.

2.2.1 Category 1: Lexicalized Sandhi

The most challenging aspect of the Sandhi phenomenon is when two distinct words concatenate to create a new form in which the word boundary cannot be easily identified. For instance, දකුණු *dakuṇu* (right) and අත *ata* (hand) are two distinct words that can be concatenated to form දකුණත *dakuṇata*, a Sandhi word where the boundary between the original words is not clear. Accordingly, in this category, we treat Sandhi forms that are created from two distinct words but maintain their original meaning, where both the separate forms and the concatenated form convey the same meaning. Therefore, they should not appear in dictionaries as distinct entries for the same meaning.

2.2.2 Category 2: Derivational Sandhi

Some of the Sandhi words appear as lemmas in dictionaries, having taken on a referential meaning in their concatenated form, although the Sandhi phenomenon occurs as a result of a morphophonological process. For instance, all five lexical entries in Table 2 are included in this category, where they

are formed as a result of concatenation but have derived new lexical forms with distinct meanings. In each of these five examples, the two forms used to concatenate have distinct meanings and have derived into different forms. For instance, ලේඛන *lēkhana* (writings) and ආගාර *āgāra* (house) are two words with distinct meanings that can be concatenated to form ලේඛනාගාර *lēkhanāgāra* (archives), a new word with distinct meaning, which is thus included in dictionaries.

Further, there is another set of forms that can be included in this category, consisting of cases where one part does not occur as a distinct word in the language. For instance, in the concatenated form සමුපකාර *samupakāra* (co-operative), සං *saṁ* is not a separate word but a prefix, while උපකාර *upakāra* (help) occurs as a distinct word. The morphophonological process has applied as a result of derivation, and thus such words can also be included in this category.

2.2.3 Category 3: Etymological Sandhi

The Sandhi phenomenon can also occur in the etymology of words and in the derivation of two particular morphemes into one lexical form. For example, ප්‍රත්‍යුත්තර *pratyuttara* (Response) is a Sandhi word with ප්‍රති *prati* + උත්තර *uttara* (Answer) as two separate morphemes. Its corresponding Sinhala derived form පිළිතුරු *piḷituru* (Answer) is also split into two morphemes as පිළි *piḷi* + උතුරු *uturu*; however, the latter morpheme උතුරු *uturu* cannot be found in the language with that particular meaning. Thus, the word පිළිතුරු is split only for etymological reasoning.

Furthermore, the word කම්මල *kammala* (smithy) is considered a Sandhi word composed of two distinct words: කම් *kam* (work) and හල *hala* (shop). Although කම්මල *kammala* is derived from these two particular forms, the original lexical meanings of the two forms have disappeared, resulting in a different meaning. Thus, the Sandhi phenomenon occurs here as a result of etymological reasoning. Therefore, such words are treated under the third category.

2.2.4 Category 4: Affixational Sandhi

Internal Sandhi forms discussed in Section 2.1.2 are treated into this category, including Sandhi phenomena that occur between a lexeme and either a prefix or suffix. For instance, the lexical entries in Table 1 are examples for affixational Sandhi. Since the lexical entries in this category involve one word combined with prefix or suffix, they do not present challenges in word boundary detection and are therefore not explored in depth in this work.

	Sandhi	Segmented
i.	අන්‍යෝන්‍යාධාර *anyōnyādhāra* (mutual aid)	අන්‍යෝන්‍ය + ආධාර *anyōnya* + *ādhāra* (mutual) + (aid)
ii.	ඔවුනොවුන් *ovunovun* (each other)	ඔවුන් + ඔවුන් *ovun* + *ovun* (they) + (they)
iii.	එකිනෙක *ekineka* (one by one)	එකින් + එක *ekin* + *eka* (from one) + (one)
iv.	කුටෝපක්‍රම *kūṭōpakrama* (tricks)	කුට + උපක්‍රම *kūṭa* + *upakrama* (crafty) + (plan)
v.	පුණ්‍යෝත්සව *puṇyōtsava* (meritorious ceremony)	පුණ්‍ය + උත්සව *puṇya* + *utsava* (merit) + (ceremony)
vi.	නමැති *namæti* (named)	නම් + ඇති *nam* + *æti* (name) + (having)
vii.	නැණැස *næṇæsa* (wisdom Eye)	නැණ + ඇස *næṇa* + *æsa* (wisdom) + (eye)

Table 3: A sample set of lexemes occur in SandhiLex

3 SandhiLex Compilation

As per the study conducted on the Sinhala Sandhi system, the compilation of SandhiLex, the Sandhi lexicon for Sinhala, was conducted in several steps using both manual and semi-automatic methods. The approach used to develop the Sandhi lexicon was as follows:

i. Collecting Sandhi lexemes from Sinhala grammar books.

ii. Collecting Sandhi lexemes from Sinhala dictionaries.

iii. Extracting Sandhi lexemes from distinct word lists.

iv. Extracting sandhi lexemes for less frequent phonemic combinations

v. Preparing Affixational Sandhi dataset

Accordingly, several types of Sandhi forms were not included in the lexicon for three reasons, such as: (i) etymological Sandhi, (ii) derivational Sandhi, and (iii) those forms appear in the spoken aspect of the language, as discussed in section 2.1.3. A sample set of Sandhi words included in SandhiLex is illustrated in Table 3.

3.1 Sandhi Lexeme

As mentioned in section 2, Sinhala, as an agglutinative language, allows one form to be inflected for many unique lexical elements. Since the lexicon becomes complex when compiled with inflected forms, the core dataset of lexical items of Sandhi (which does not include inflectional Sandhi forms) was denoted only with stem-like lexical units. These units can be considered the most common forms in the compilation of the respective lexical items. Accordingly, in this initiative, Sandhi lexemes (SiLx) refer to those specific lexical elements with no inflections.

3.2 Collecting SiLx from Sinhala grammar books

One of the easier ways of collecting Sandhi words is by reviewing the literature and manually collecting the specific lexical entries, since it is more accurate method of collecting Sandhi lexemes. Further, Sandhi, as a common topic, is addressed in nearly all traditional and contemporary Sinhala grammar books. However, since these resources are only available in print, the data must be collected manually. Thus, as the first step of the initiative, we collected Sinhala Sandhi words manually from Sinhala grammar books. Among the books utilized for collecting manually the sandhi lexemes included Derivative Grammar Books: Pannasara Thero (2004); Gunawardhana (1924), traditional grammar books: Perera (1985); Thilakasiri (1997); Sumanasara (2007); Non-Traditional Prescriptive Grammar Books: Kumaranathunga (1937); De Seram and Gunawardhana (1971); Sampath (2013); and Disanayaka (1997).

3.3 Collecting SiLx from dictionaries and glossaries

Coperahewa (2014) is a dictionary compiled of Sinhala Sandhi words. This dictionary consists of around 1,600 entries, which include all types of Sandhi words, including affixational Sandhi, etymological Sandhi, and derivational Sandhi. As in traditional grammar books, the list of Sandhi words in Coperahewa (2014) includes lexical entries that are not used in contemporary Sinhala language. Furthermore, Sinhala language dictionaries such as Wijethunga (2005), Soratha Thero (1952), and Soratha Thero (1956) were also referred, and Sandhi lexemes were manually collected from these.

3.4 Extracting SiLx from a text corpus

LTRL-UCSC (2007) is a Sinhala text corpus which includes modern Sinhala novels, short stories, and critiques written by renowned Sinhala authors. It also contains news articles collected from mainstream Sinhala newspapers published between 2004 and 2010. This corpus represents contemporary Sinhala language usage across various contexts and genres, making it a balanced text corpus suitable for NLP research and development for the language.

In this initiative, we use the distinct word list from LTRL-UCSC (2007) since it includes the most frequent words in the language. Although manually collecting the particular lexical entries would be more accurate, it is a tedious task due to several reasons. Firstly, it is time-consuming, and secondly, it requires substantial human resources and a high level of linguistic and grammatical knowledge of the language. Therefore, we need efficient methods for extracting lexical entries from relevant resources. Accordingly, a list of Sandhi words was extracted and cleaned through several steps:

i. Utilizing the list of distinct words from LTRL-UCSC (2007) and filtering the words beginning with vowels.

ii. Extracting words for certain character clusters as illustrated in Table 4.

iii. Removing irrelevant words.

This method proved to be more effective.

3.5 Extracting sandhi lexemes for less frequent phonemic combinations

In the process of compiling the lexicon, this step was employed to count the phonemic combinations for which morphophonemic changes were applied. For this task, the entire dataset (only category 1) was transliterated using the ISO 15919 standard for Sinhala. This was done to simplify the

character clusters	Occurrences in the corpus	Remains in the SandhiLex
ාර්ථ *ārtha*	1009	214
ංක *ṁka*	2323	304
ක්ෂ *kṣa*	4317	396
ත්‍ය *tya*	1986	388
ාචාර *ācāra*	649	142
ාලෝක *ālōka*	125	48
ාකාර *ākāra*	1138	102
ාංග *āṁga*	557	110
පදේශ *padēśa*	165	44
න්තර *ntara*	762	136

Table 4: A sample of character clusters extracted from the distinct word list

process of counting the phonemic combinations. After reiterating the process, the phonemic combination frequencies of the final version are presented in Table 5.

As per the statistics given in Table 5, the most frequent phonemic combinations in the list are a a and a ā, which reported frequency counts of 1555 and 1276 respectively. However, none of the other combinations reach a count of 1,000 occurrences. Furthermore, out of 144 phonemic combinations, 68 of them do not appear in the list, whereas another 37 reported fewer than 5 occurrences in the list.

4 Affixational Sandhi dataset

The SiLx entries treated under category 4, which was discussed in Section 2.2.4, are included in the affixational Sandhi dataset. This dataset was compiled using LTRL-UCSC (2007) and LTRL-UCSC (2008) developed by the Language Technology Research Laboratory of the University of Colombo School of Computing, Sri Lanka. Since the data consisted of affixes along with lexemes, the number of data samples is much larger compared to the main set of data, which includes the first three categories. For instance, the dataset consists of 73,620, 18,434, 16,985, 7,520, 2,569, and 2,561 lexical entries for the suffixes උත් *ut*, ඉන් *in*, එන් *en*, එහි *ehi*, එකු *eku*, and එක් *ek* respectively.

5 Conclusion

Sandhi, as a morphophonological process, has been a topic in all grammar books. Considering the inadequacy of studies in traditional grammar books, this paper reports a new classification of Sandhi in Sinhala by classifying them according to their morphophonological processes and occurrences in the language. Accordingly, Sinhala Sandhi has been classified into four categories: Lexicalized Sandhi, Derivational Sandhi, Etymological Sandhi, and Affixational Sandhi. Based on the study, a Sandhi Lexicon (SandhiLex) for the Sinhala language was compiled, comprising around 4,500 Sandhi lexemes for Lexicalized Sandhi data and more than 300k lexical units of affixational Sandhi dataset which will contribute to the advancement of research in NLP for the Sinhala language.

Phonemic combinations	Frequency Count
a a	1555
a ā	1276
a u	327
a i	172
ā a	164
ā ā	122
i a	94
u a	45
i i	41
i ā	35
i u	28
ā u	23
u u	23
ā i	11
a ī	10

Table 5: Phonemic combination frequencies in the SandhiLex

6 Limitations

Sandhi is one of the main grammatical phenomena in the Sinhala language, the morphophonemic nuances can be studied further. However, this research focused specifically on understanding Sandhi phenomena in Sinhala, recognizing its significance as a grammatical feature that affects many NLP applications. Thus, one objective of the paper was to report the process of developing a Sandhi lexicon for Sinhala. As Sandhi has been classified into several categories, the initiative was to collect Sandhi words particularly for the most significant category of Sandhi words. Further, the study was limited to analyzing Sandhi in the Sinhala language. The study can be further advanced

by analyzing the Sandhi categories in other Indo-Aryan languages as well. Additionally, the nuances of morphophonological features can be explored in greater depth in future research.

Acknowledgements

This research was financially supported by the University of Colombo School of Computing through the Research Allocation for Research and Development. The authors gratefully acknowledge Mr. Vincent Halahakone for his assistance in data collection for the Sandhi dataset and for proofreading the paper. We also thank Prof. W.M. Wijeratne for reviewing the paper and providing insightful feedback. Special thanks go to Prof. Sandagomi Coparahewa and Ms. Chanika Dayarathna for their support in finding several books required for the research. Finally, we thank all the members of the Language Technology Research Laboratory of the University of Colombo School of Computing for their various contributions in making this work a success.

References

W Sidney Allen. 1972. *Sandhi: the theoretical, phonetic, and historical bases of word-junction in Sanskrit*. Mouton, The Hague, Paris.

Sandagomi Coperahewa. 2014. *Dictionary of Sinhala Sandhi*. S. Godage Brothers, Colombo 10, Sri Lanka.

James De Alwis. 1852. *The Sidath Sangarawa, a grammar of the Singhalese language, translated into english, with introduction, notes, and appendices, by James de Alwis*. Skeen.

E. De Seram and H.D.J. Gunawardhana. 1971. *vyākaraṇaya vimarśanaya*.

Nisansa De Silva. 2019. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

VV Devadath, Litton J Kurisinkel, Dipti Misra Sharma, and Vasudeva Varma. 2014. A sandhi splitter for malayalam. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 156–161.

J.B. Disanayaka. 1997. *samakālīna sinhala lēkhana vyākaraṇaya - sandhi vigrahaya*. S. Godage Brothers, Colombo 10, Sri Lanka.

Punchibanda Ekanayake. 2016. *sandhi vimarśana*. Samayawardhana Bookshop (Pvt) Ltd., Colombo 10, Sri Lanka.

Abraham Mendis Gunasekara. 1891. *A comprehensive grammar of the Sinhalese language: adapted for the use of English readers and prescribed for the Civil Service examinations*. GJA Skeen.

W.F. Gunawardhana. 1924. *siddhānta parīkṣaṇaya*. Associated Newspapers of Ceylon Limited - ANCL, D.R. Wijewardena Mawatha, Colombo-10, Sri Lnka.

Danesh Jain and George Cardona. 2007. *The Indo-Aryan Languages*. Routledge.

Munidasa Kumaranathunga. 1937. *vyākaraṇa vivaraṇaya*.

LTRL-UCSC. 2007. Language resources of ltrl-ucsc: Ucsc 10m word sinhala text corpus.

LTRL-UCSC. 2008. Language resources of ltrl-ucsc: Ucsc 700k word morphological lexicon for sinhala.

P.B. Meegaskumbura. 2020. *sandhi parisara hā sandhi-vidhi*. In *Lekhanawali*, pages 61–69. Vidarshana Publishers (Pvt) Ltd., Colombo, Sri Lanka.

Okkampitiye Pannasara Thero. 2004. *sidatsaṅgarā vimasuma*. Okkampitiye Pannasara Thero.

Theodore G. Perera. 1985. *Siṃhala bhāṣāva*. M.D. Gunasena Co. (Pvt.) Ltd. Olcott Mawatha, Colombo 11, Sri Lanka.

Rajith Priyanga, Surangika Ranatunga, and Gihan Dias. 2017. Sinhala word joiner. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 220–226.

He.Wa. Bihesh Indika Sampath. 2013. *viyaraṇa vivaraṇa - prathama bhāgaya*. S. Godage Brothers, Colombo 10, Sri Lanka.

Weliwitiye Soratha Thero. 1952. *śrī sumaṁgala śabdakōṣaya : prathama bhāgaya*.

Weliwitiye Soratha Thero. 1956. *śrī sumaṁgala śabdakōṣaya : dvitīya bhāgaya*.

Thimbiriwewa Sumanasara. 2007. *siṁhala bhāṣāvē vyākaraṇaya*. Wijesooriya Grantha Kendraya, Maradana Road, Punchi Borella, Sri Lanka.

Siri Thilakasiri. 1997. *siṁhala viyaraṇa vidi*. Rathna Book Publishers (Pvt) Ltd, Maradana Road, Colombo 10, Sri Lanka.

Harishchandra Wijethunga. 2005. *mahā siṁhala śabdakōṣaya*. M.D. Gunasena Co. (Pvt.) Ltd. Olcott Mawatha, Colombo 11, Sri Lanka.