

A Framework for Identifying Pali Words in Myanmar Text

Zin Maung Maung

University of Computer Studies, Mandalay, Myanmar

zmmaung@gmail.com

Abstract

Myanmar language has long been influenced by the usage of Pali words since ancient times. This study proposes a framework for identifying Pali words in Myanmar text to analyze the influence of Pali words on Myanmar language. The framework applies a combination of rule-based and dictionary look-up approach to detect Pali words written in Myanmar text. The system proposes a Pali word inventory process for collecting Pali words from real-world Myanmar online articles. After detecting the Pali words, Part-of-Speeches (POS) and corresponding meanings can be provided with the help of a Pali-Myanmar dictionary. By being able to detect Pali words in Myanmar text, the findings of this study are expected to benefit statistical analysis of Pali words influencing on Myanmar language and this, in turn, will improve further NLP tasks of Myanmar language such as spelling checking, machine translation etc.

Keywords: Myanmar Pali Words Identification, Syllable Segmentation, Syllable Merging, Chained Syllables, Normalization

1. Introduction

Myanmar language, also known as Burmese, is the official language of the Union of Myanmar. Myanmar is a member of Tibeto-Burman languages, which is a sub-family of the Sino-Tibetan family of languages. Myanmar is a

tonal and analytic language. It is spoken by 32 million people as their first language and also spoken as a second language by ethnic minorities living in Myanmar [3]. Myanmar language is written using the Myanmar or Burmese script. This is a phonologically based script, adapted from Mon, and ultimately based on an Indian Brahmi prototype [6]. Myanmar characters are rounded in appearance and the script is written from left to right. No space is used between words but spaces are usually used to separate phrases.

Pali is a Middle Indo-Aryan language (or Prakrit) of the Indian subcontinent. It is best known as the language of many of the earliest extant Buddhist scriptures, as collected in the Pali Canon or Tipitaka, and as the liturgical language of Theravada Buddhism [1]. It is spoken in Theravada nations of Southeast Asia to chant in a ritual context. Pali language has great influence on Myanmar language due to the widespread teaching and practicing of Buddhism and Buddhist literature in Myanmar. Many words from the Buddhist Canon have been directly or indirectly used in Myanmar literature since ancient times. Some of them have already become native Myanmar words and are widely used in Myanmar text such as မေတ္တာ၊ သုခ။ There can also be seen Myanmar words which were originally derived from Pali language such as တက္ကသိုလ်၊ မာန်။ In Myanmar, Pali language is written by using the Myanmar script. Table 1 describes examples of Pali words frequently used in Myanmar language.

Table 1. Examples of Pali words

Pali Words	Meaning in English
လောက	world
ဓမ္မ	doctrine
ပညာ	wisdom
ဥတ္တရ	northern
ရသာ	taste

2. Related Works

Since Myanmar text are written without using spaces between words, word segmentation of Myanmar text becomes a challenging task. Myanmar words are formed by combining one or more syllables and as a consequence syllable segmentation plays as the basic and essential role in natural language processing of Myanmar text. Syllable segmentation or syllabification can be defined as the process of identifying syllable boundaries in a piece of text.

Zin Maung Maung and Yoshiki Mikami (2008) developed a rule-based approach of syllable segmentation algorithm for Myanmar text [10]. In their study, syllable segmentation rules were created based on the characteristics and syllable structure of Myanmar script and syllable segmentation is carried out by comparing the input text string with the defined syllable segmentation rules. The input text string is first converted into equivalent category code string and then matched against the pre-defined segmentation rules to identify syllable boundaries. The algorithm was tested on a corpus containing 32,283 Myanmar syllables and an accuracy of 99.96% was achieved.

Yuzana and Khin Mar Lar Tun (2009) proposed a syllabification algorithm for Myanmar Pali characters by pattern generating method [9]. The proposed method attempts to achieve Pali word syllabification by generating user defined patterns for syllables.

3. Nature of Myanmar Pali Words

In Myanmar language, Pali words are written using the Myanmar or Burmese script. This section explains the nature of Pali words written in Myanmar text and discusses the need to normalize some Myanmar Pali words for language processing purposes.

3.1. Types of Pali Words

Myanmar Pali words can be generally classified into two groups. The first group contains Pali words written in conjunct consonants or subscripted consonants form. In this form, two consonant letters are stacked together and the second consonant is subscripted below the first consonant. There are at least two syllables involved in this form of Pali words and hence they can be called chained syllables. In Unicode encoding of Myanmar script, conjunction of consonant letters is indicated by the insertion of a virama (U+1039 MYANMAR SIGN VIRAMA) between them. It causes the second consonant to be displayed in a smaller form below the first; the virama is not visibly rendered. Examples of this group of Pali words are shown in Table 2.

Table 2. Pali words with conjunct consonants

Pali Word	Meaning
စက္ခူ	eye
ဒက္ခိဏ	south
မိတ္တ	friend
ရုက္ခ	tree
သိပ္ပ	arts and science

The second group consists of Pali words that are written without using conjunct consonants. These Pali words are similar to Myanmar words in writing. They usually contain at least one syllable and follow the structure of Myanmar syllables. They cannot be differentiated from

Myanmar words, by means of writing, without having knowledge of the Pali language. Some Pali words have long been used in Myanmar language and literature so that they become native Myanmar words in usage. Examples of the second group of Pali words are shown in Table 3.

Table 3. Pali words without stacked consonants

Pali Word	Meaning
ပဌမ	first
မဟေသီ	queen
ရာဇ	king
သုရိယ	sun
အဓိပတိ	chief

3.2. Normalization of Chained Syllables

The Pali word identification framework of this study is developed based on the Unicode encoding of Myanmar script [4]. In Unicode encoding, subjoined characters are shown by using an invisible Virama sign (U+1039 MYANMAR SIGN VIRAMA) to indicate that the following character is subjoined and should take a subjoined form.

In Myanmar text, many Pali words can be found in conjunct consonants forms as explained in previous section. This form contains more than one syllable and these syllables will be referred to as chained syllables in this study. In chained syllables writings, the upper consonant acts as the devoweliser for the preceding consonant and the lower consonant acts as the base consonant of the second syllable. In some cases, native Myanmar words can also be found with conjunct consonants forms. They are not Pali words however they were written by simulating the Pali style of conjunct consonants (e.g., လွတ်၊ လိမ္မော်). Table 4 describes examples of Pali words that contain chained syllables.

In some cases, the syllable chained Pali words need to be normalized for language processing operations. Normalization here refers to separating the stacked consonants of Pali words into separate syllables. Myanmar word

sorting requires syllable level analysis during the word ordering process [7] and Myanmar Text-to-Speech processing (TTS) requires producing sound for each syllable unit [2]. In such situations, stacked consonants present a barrier for syllable boundary detection and they need to be normalized to get separate syllables. Normalizing the chained syllabled would simplify such NLP tasks requiring syllable-level analysis. Myanmar Pali words are also very important for text analysis and text input method interface of Myanmar language because such words use various consonant clusters [8].

Normalization of conjunct consonants also plays an important role in analysis of Pali words in Myanmar text. In Myanmar text, some of the Pali words are read differently from their written form. For example, the Pali word ဝါကျ is read as ဝတ်ကျ. Here, the consonant က acts not only as the devoweliser (together with an Asat) for the first syllable ဝတ် but also as the base consonant for the second syllable ကျ. There are specific rules on how to write and read Pali words in Myanmar grammar [5]. By combining the Pali word reading rules into the normalizing process, the system can show simplified and correct pronunciation for identified Pali words from the given input text.

Table 4. Examples of chained syllables and their normalized forms

Chained Syllable	Normalized Form	Syllable Segmented Form
ခတ္တိယ	ခတ်တိယ	ခတ် တိ ယ
စာရိတ္တ	စာရိတ်တ	စာ ရိတ် တ
တက္ကသိုလ်	တက်ကသိုလ်	တက် က သိုလ်
ဝတ္ထု	ဝတ်ထု	ဝတ် ထု
သမ္မတ	သမ်မတ	သမ် မ တ

4. Proposed Framework

The framework contains two parts: (1) Pali word inventory process and (2) Pali word identification process. The first section reviews the rule-based syllable segmentation algorithm

which is essential for both Pali word inventory and identification processes. The second section explains the Pali word inventory process for improving the system dictionary and the last section explains the proposed framework for identifying Pali words in Myanmar text.

4.1. Syllable Segmentation Process

Syllables are building blocks of words and syllable segmentation plays an essential role in many NLP application of Myanmar text, including line breaking, word segmentation and sorting. In Myanmar script, each consonant carries an inherent ‘a’ vowel hence forming a syllable. A syllable can be formed by combining the base consonant with optional medials, dependent vowels and dependent various signs. There are also independent vowels that act as stand-alone syllables without combining with consonants. The proposed framework applies rule-based syllable segmentation algorithm developed by Zin Maung Maung and Yoshiki Mikami [10] and this section will briefly review the aforementioned algorithm.

မြန်မာစကား = မ+ြ+န+ိ+မ+ာ+စ+က+ာ+း

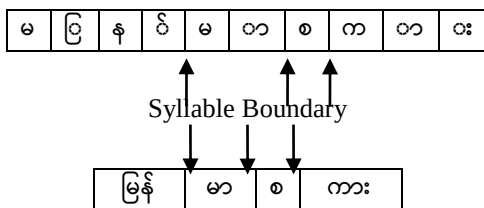


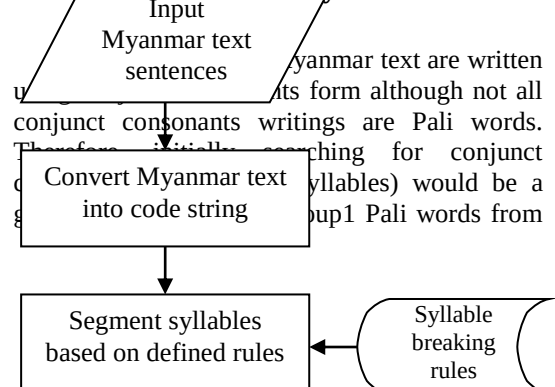
Figure 1. Example of syllable segmentation

Figure 2. Syllable segmentation process

First of all, the input Myanmar text string is converted into equivalent category code string according to the pre-defined category codes for Myanmar script. Syllable breaking rules were defined based on the Myanmar syllable structure and the converted code string is matched against the breaking rules to identify the syllable boundary for each character pair. Then, syllable boundary signs are inserted in the code sting according to the breaking rules. Finally, the syllable segmented code string is converted into equivalent Myanmar text string and displayed as the system output.

By doing syllable segmentation as a pre-processing step, the system can efficiently use syllable-level longest matching method when comparing the input text with dictionary contents, effectively eliminating the character-level ambiguity problems.

4.2. Syllable Segmentation Process



the input text. This will reduce the processing time of scanning every character in the input text to match with stored Pali words in the dictionary.

First of all, Myanmar articles from different sources such as blogs, online magazines and news sites will be collected and converted into Unicode encoded text. The collected articles will be syllable segmented by the algorithm developed in [10]. From the syllable segmented text, the system will scan chained syllables to get potential Pali words. At this stage, syllable merging needs to be carried out since some Pali words contain additional syllables before or after the chained syllables. The candidate Pali words are then compared with 3,559 Pali words initially stored in the system dictionary whether they are Pali words or not.

The rule-based system will detect every word containing chained syllables and this include (1) Pali and Pali derived words that contain chained syllables and (2) non-Pali words written in chained syllables form. By manually checking the identified potential Pali words, Pali words can be differentiated from non-Pali words that contain chained syllables. From this step, newly found Pali words can be added to the Pali dictionary and Myanmar or other words that contain chained syllables can be added to the non-Pali words list. The list is useful for automatically differentiating potential Pali words from Myanmar and other words. The manual checking process may be eliminated if the system can collect a significant amount of non-Pali words written in chained syllables form.

The rule-based collection of Pali words from online Myanmar text will improve the system dictionary and the dictionary contents will gradually grow as the system is trained with more and more Myanmar text. The proposed Pali word inventory process is shown in Figure 3.

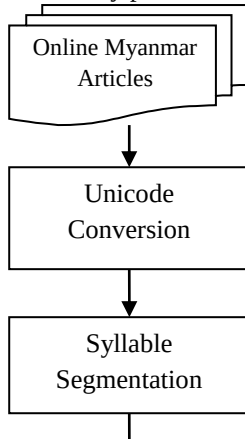


Figure 3. Pali word inventory process

4.3. Pali Word Identification Process

Myanmar text includes a mixture of native Myanmar words together with foreign loan words, digits, punctuation marks and white spaces. The input of the system is paragraph-level of Myanmar text and the purpose of the system is to identify Pali words from a mixture of these different kinds of characters from the input text. The flowchart for Pali word identification process is shown in Figure 4.

First, the input text will be syllable segmented according to the algorithm developed in [10]. The segmented syllables are then matched against the Pali words stored in the system dictionary. The dictionary contains pre-stored Pali and Pali derived words together with corresponding part of speeches and meanings. The syllable level longest matching method is used to identify Pali words from the input text.

The syllable level longest matching method is a variation of the maximum matching or greedy algorithm using syllable units rather than characters. The longest matching method starts at the beginning of the sentence by reading each syllable from the input text and comparing with the stored Pali words until a match with the longest word is found. If a longest match is found, the algorithm marks a boundary at the end of the longest word and starts reading from the next syllable. If no match is found, it just puts a word boundary after the read syllable and scans the next syllable. It repeats the process until the end of sentence is reached.

During the identification process, newly found Pali words can be added to the dictionary with the help of user verification. After identifying the Pali words, corresponding part of speeches and their meanings can be extracted from the Pali dictionary. Among the identified Pali words, the ones with conjunct consonants will be normalized by applying the normalization rules. Finally, the system will display a summary of Pali words identified in the input text to users, together with their corresponding part of speeches, meanings and normalized forms. The following example shows step by step processing of the proposed system and the expected system output is shown in Table 5.

Step 1: Input Text

သိဒ္ဓတ္ထမင်းသားသည် မဇ္ဈိမဒေသတွင် ဗုဒ္ဓအဖြစ်
ပေါ်ထွန်းခဲ့သည်။

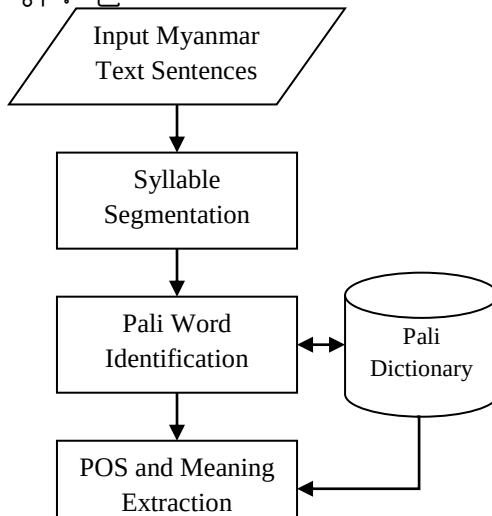


Figure 4. Pali word identification process

Step 2: Syllable Segmentation

|သိဒ္ဓတ္ထ|မင်း|သား|သည်| |မဇ္ဈိ|မ|ဒေ|သ|တွင်| |ဗုဒ္ဓ|အ|ဖြစ်|
|ပေါ်|ထွန်း|ခဲ့|သည်|။|
|CVCSCCSC|CCAF|CVF|CCA|W|CCSCV|C|CV|C|C
MCA|W|CVCSC|C|CMCA|W|CVVA|CMCAF|CVF|C
CA|P|

Step 3: Pali Words Identification

|သိဒ္ဓတ္ထ|
|မဇ္ဈိ|မ|
|ဗုဒ္ဓ|

Where,

C = Consonant, M = Medial, V = Dependent Vowel Sign, F = Dependent Various Sign, A = Myanmar Sign Asat, W = White Space, P = Punctuation Mark, S = Myanmar Sign Virama

5. Conclusion

This paper proposes a framework for identifying Pali words in Myanmar text. This research attempts to find out the influence of Pali words on Myanmar literature by searching the frequently used Pali words from real-world Myanmar text. The framework employs a combination of rule-based and dictionary look-up approach to identify Pali words. The system

dictionary initially contains 3,559 Pali and Pali derived words and the rule-based method is applied to improve the system dictionary by collecting Pali words from real-world online articles. The identification process applies syllable-level longest matching method to detect Pali words from the input text. The accuracy of the proposed system depends highly on the completeness of the system dictionary. The

quality of the dictionary will improve as the system is trained with more and more Myanmar text. This research is in progress and hopes to find out the usage and frequency distribution of Pali and Pali derived words used in Myanmar language after doing experiments. The findings of this research will greatly benefit further NLP steps of Myanmar language such as sorting, spelling checking, machine translation etc.

Table 5. Expected system output

Pali Word	Coded Form	Syllable Count	Normalized Form	Part-of-Speech	Meaning
သိဒ္ဓတ္ထ	CVCSCCSC	3	သိဒ်ဇတ်ထ	Noun	name of the Buddha
မဇ္ဈိမ	CCSCV C	3	မဇ်ဇိမ	Adjective	middle
ဗုဒ္ဓ	CVCSC	2	ဗုဒ်ဇ	Noun	the enlightened one

References

- [1] <http://en.wikipedia.org/wiki/Pali.htm>
- [2] Ko Ko Oo and Nilar Thein, “Myanmar Language Text-to-Speech (TTS) Synthesis”, Proceedings of the 5th International Conference on Computer Applications (ICCA2007), 2007, Yangon, Myanmar.
- [3] Lewis, M. Paul (ed.) 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex: SIL International. Online version: <http://www.ethnologue.com/>
- [4] Martin Hosken and Maung Tuntun Lwin, “Representing Myanmar in Unicode: Details and Examples”, Unicode Technical Note 11-2, <http://www.unicode.org/notes/tn11/>
- [5] Myanmar Language Commission, “Myanmar Grammar, Thirty Years Special Edition”, University Press, 2005, Yangon, Myanmar.
- [6] Peter T. Daniels and William Bright, 1996, “The World’s Writing Systems”, Oxford University Press.
- [7] Tin Htay Hlaing and Yoshiki Mikami, “Collation Weight Design for Myanmar Unicode Text”, Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.
- [8] Ye Kyaw Thu and Yoshiyori Urano, “Grapheme Cluster Segmentation Tool for Myanmar Text Based on Positional Prediction Text Input Concept, Proceedings of the 9th International Conference on Computer Applications (ICCA2011), 2011, Yangon, Myanmar.
- [9] Yuzana and Khin Mar Lar Tun, “Myanmar-Pali Text Syllabification by Utilizing Pattern Generation Tactic”, Proceedings of the 7th International Conference on Computer Applications (ICCA2009), 2009, Yangon, Myanmar.
- [10] Zin Maung Maung and Yoshiki Mikami, “A Rule-based Syllable Segmentation of Myanmar Text”, Proceedings of the IJCNLP-2008 Workshop on NLP for Less Privileged Languages, pages 51-58, Hyderabad, India, January 2008.