Identification of Adopted Pali Words in Myanmar Text

Zin Maung Maung

University of Computer Studies
Mandalay, Myanmar

Abstract

Myanmar language has been significantly influenced by Pali language due to the practice of Buddhism and study of Buddhist literature in Myanmar. As a result, Pali words have been widely adopted and used in Myanmar language. This study presents an algorithm for identifying Myanmar-adopted Pali words in Myanmar text. The system employs a combination of rule-based syllable segmentation and a dictionary-based longest matching method. A program was developed and trained on a corpus containing 8,895 sentences. It recognized 579 unique Pali words. The accuracy of the system was tested on a different corpus containing 3,641 sentences and the system correctly identified 279 unique Pali words, achieving a Precision of 97.59%, Recall of 99.04% and F-measure of 98.31%. Usages of Pali words are inevitable in Myanmar text and the results of this study will improve many NLP tasks of Myanmar language such as spelling checking, text categorization and text-to-speech synthesis etc.

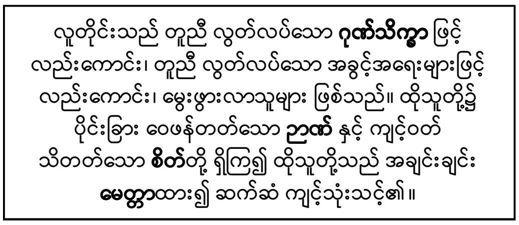
Keywords: *Myanmar Pali Words, Pali Words Identification, Syllable Segmentation, Longest Matching, Conjunct Consonants.*

1. Introduction

The Myanmar language, also known as Burmese, is the official language of the Republic of the Union of Myanmar. It is spoken by the majority of people, including ethnic groups, living in Myanmar. Myanmar is a member of Tibeto-Burman language family, which is a subfamily of the Sino-Tibetan family of languages. Myanmar is a tonal and analytic language and it is written using the Myanmar or Burmese script. Burmese is a phonologically based script, adapted from Mon, and ultimately based on an Indian Brahmi prototype [8]. The earliest known inscriptions in the Burmese script date from the 11th century. It is also used to write Pali, the sacred language of Theravada Buddhism, and other several ethnic minority languages of Myanmar with the addition of specialized characters and diacritics for each language. Burmese 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. There is no specific rule on usage of spaces in Burmese script.

Pali is a Middle Indo-Aryan language (of Prakrit group) 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. It is mostly spoken in Theravada nations of Southeast Asia and frequently chanted in a ritual context. Various scripts, including Sinhalese, Khmer, Lao, Devanagari, Asokan Brahmi and Roman, have been used to write the Pali language in different nations. In Myanmar, Burmese script has been used to write the Pali language.

Myanmar language has been greatly influenced by the Pali language due to the widespread practice of Buddhism and the study of Buddhist scriptures in Myanmar. M. H. Bode (1909) stated in his book "The Pali Literature of Burma" that the essentially Indian genius, the psychological subtleties, and high thoughts of Buddhism have forced the Burmese language to grow, deepen and expand continually. He mentioned several facts about the influence of Pali on Burmese language that 1) Burmese was raised to the level of a literary language (in or about the fourteenth century) by the addition of a great body of Indian words necessary to express ideas beyond the scope of that picturesque vernacular 2) Burmese, being an agglutinative language, lacks the force, terseness, and delicacy that Pali owes to its nominal and verbal inflections and its power of forming elaborate compounds and 3) thus before the translating period, authors of Burmese race had studied Pali and learned to use it: ever since the twelfth century it has been a tradition of Burmese scholars to produce literary work in Pali [5]. As a consequence of Pali influence on Myanmar language, usages of Pali and Pali-derived words are wide and frequent in Myanmar text. Some Pali words were directly incorporated into Myanmar language (e.g. မေတ္တာ၊ သုခ) and some of them are derivatives of the original Pali words (e.g. တက္ကသိုလ်၊ မာန်). As an example of Pali words usages in Myanmar text, an excerpt of the UDHR (Universal Declaration of Human Rights) written in Myanmar language, is shown in Figure 1. Myanmar-adopted Pali words are shown in bold and their meanings in English and Pali origins are shown in Table 1.



လူတိုင်းသည် တူညီ လွတ်လပ်သော **ဂုဏ်သိက္ခာ** ဖြင့်
လည်းကောင်း၊ တူညီ လွတ်လပ်သော အခွင့်အရေးများဖြင့်
လည်းကောင်း၊ မွေးဖွားလာသူများ ဖြစ်သည်။ ထိုသူတို့၌
ပိုင်းခြား ဝေဖန်တတ်သော **ဉာဏ်** နှင့် ကျင့်ဝတ်
သိတတ်သော **စိတ်**တို့ ရှိကြ၍ ထိုသူတို့သည် အချင်းချင်း
မေတ္တာထား၍ ဆက်ဆံ ကျင့်သုံးသင့်၏။

Fig. 1 Pali words in UDHR

Table 1: Pali words in UDHR

Pali Word	Pali root-word	Meaning
ဂုဏ်	ဂုဏ (guṇa)	virtue
သိက္ခာ	သိက္ခာ (sikkhā)	discipline
ဉာဏ်	ဉာဏ (ñāṇa)	knowledge
စိတ်	စိတ္တ (citta)	mind
မေတ္တာ	မေတ္တာ (metta)	loving kindness

2. Related Works

Pali language had a significant influence on Myanmar language since ancient times. Buddhist monks and scholars studied the Pali language mainly to gain access to the Buddhist Canon and many religious works were written using the Pali language. As a result, a large number of Pali words had been adopted and incorporated into Myanmar language. Today, usages of Pali-derived words are frequent and unavoidable in Myanmar language. Myanmar is still a less computerized language compared to others and the availability of tools and resources essential for NLP operations such as spelling checkers, comprehensive machine readable dictionaries, standardized balanced corpora etc., are still inadequate. This motivates a dedicated study of Myanmar-adopted Pali words by identifying them from real-world Myanmar text and producing electronically available language resources of them to be used in NLP tasks.

No previous work was found on the study of Myanmar-adopted Pali words in a real-world context and this research is the first attempt to measure the influence of Pali language on Myanmar language (by computational means). Burmese script, similar in nature to some south-east Asia scripts of Brahmi descendants, uses no space between words and this creates challenges for identifying word boundaries in Myanmar text. Identifying adopted Pali words in Myanmar text also requires word boundary detection and previous works relating to the word segmentation of Myanmar and Thai script are reviewed in the following sections.

2.1 Related Works on Myanmar Language

Syllabification of Pali words in Myanmar text was first presented by Yuzana and Khin Mar Lar Tun (2009) [14]. A rule-based pattern generation method for syllabification is presented in their research. The algorithm was tested on normal and Pali syllables. Unique code values were first defined for each of the consonants, medials, vowels, kinzi and final characters of the input text and the algorithm generates corresponding code patterns for input text by matching the input character with the defined coded patterns. The proposed algorithm can accept word-level Myanmar text and generates only corresponding patterns or codes of the input words.

Myanmar doesn't use inter-word spacing and word tokenization or segmentation plays a vital role in most NLP applications. Syllables are building blocks of words and knowing syllable boundaries is very helpful for identifying word boundaries. Therefore, syllable segmentation step is a foundation step for many NLP operations, including word segmentation, sorting and text-to-speech synthesis.

Zin Maung Maung and Yoshiki Mikami (2008) presented a rule-based approach of syllable segmentation algorithm for Myanmar text [15]. 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 each character pair of the input text string with the pre-defined syllable segmentation rules. The algorithm was tested on a corpus containing 32,283 Myanmar syllables and an accuracy of 99.96% was achieved.

Syllable segmentation was usually carried out as a basic step in previous approaches of Myanmar word segmentation. Hla Hla Htay and K. N. Murthy (2008) presented a Myanmar word segmentation using syllable-level longest matching approach [2]. Syllable segmentation in their research was performed by matching the input text with pre-stored Myanmar syllables list collected from various sources. Word segmentation was carried out by performing syllable-level longest matching method utilizing segmented syllables. The longest matching method was used together with a list of Myanmar words collected from various sources. Their research reported 98.95% F-measure for word segmentation. The encoding of Myanmar text used in their research was in WinInnwa font, which is a glyph-substituted Latin font used to write Myanmar text. An encoding conversion may be necessary to perform the proposed algorithm on Unicode encoded Myanmar text.

Another work focusing on Myanmar word segmentation based on the Unicode standard encoding was reported by Htun Thura Thet, et al. (2008) [3]. The research conducts word segmentation using a two-phase method: syllable segmentation and syllable merging. The first step performs syllable segmentation using a rule-based heuristic approach. A dictionary-based statistical syllable merging approach is carried out for word segmentation. Pre-defined syllable segmentation rules were initially applied to the Myanmar text and the dictionary-based syllable merging method, together with the collocation strengths of a sentence or phrase, was used for word segmentation. In their research, F-measure of 98.99% was reported for word segmentation. Word segmentation errors were mainly caused by the occurrences of unknown words in Myanmar text, including Myanmar-adopted Pali words.

2.2 Related Works on Thai Language

Many attempts have been made in Thai language processing efforts for syllable and word segmentation. A dictionary-based approach to Thai syllable separation was first proposed by Poowarawan (1986) [13]. In Thai Language, syllable segmentation was considered as the first step towards word segmentation and research results showed that many of word segmentation ambiguities were resolved at the level of syllable segmentation (Aroonmanakun, 2002) [12]. Thai syllable segmentation can be viewed as the problem of inserting spaces between pairs of characters in the text and Sornil and Chaiwanarom (2004) reported that the character-level ambiguities of word segmentation can be reduced by extracting syllables whose structures are more well-defined [7]. Most approaches of Thai word segmentation use a dictionary as their basis. However, in such approaches, the segmentation accuracy depends on the quality of the dictionary used for analysis and unknown words in the text can affect the performance. Theeramunkong and Usanavasin (2001) proposed a non-dictionary-based approach to Thai word segmentation by using a method based on decision tree models [9]. Their approach claimed to outperform some well-known dictionary-dependent techniques of word segmentation such as the maximum and the longest matching methods.

3. Nature of Myanmar Pali Words

In Myanmar language, Pali words are written using the same script for writing Myanmar language. Myanmar Pali words can be generally classified into two groups from an encoding perspective. The definition of Myanmar Pali words in this research refers to Myanmar-adopted Pali words i.e., Pali loanwords used in Myanmar language. The following sections explain the characteristics of Myanmar Pali words found in Myanmar language.

3.1 Conjunct Consonants

A conjunct consonant contains two consonants letters coming together form what is called a conjunct or double consonant. For instance, in vassa, kattha and pandapeti, the ss, tth nd are conjunct consonants [1]. In Myanmar script, conjunct consonants are shown in a subscripted consonant form.

ဓမ္မ (dhamma)
↑ ↑
Conjunct Consonants

3.2 Pali Words with Conjunct Consonants

The first group of Pali words contains Pali words written in conjunct consonants or subscripted consonant form. In this form, two consonant letters are stacked together and the second consonant is subscripted below the first consonant, killing the inherent vowel sound of the first consonant. There are at least two syllables joined together in this form of Pali words and hence they can be called chained syllables. In Unicode encoding for Myanmar script, conjunction of two 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 [6], [10]. Encoding of a Myanmar Pali word written in conjunct consonants form is shown below.

သမ္မတ 101E 1019 1039 1019 1010 president
↑
Virama Sign

Table 2: Pali words with conjunct consonants

Pali Word	Meaning
စက္ခု	eye
ဒက္ခိဏာ	south
မိတ္တ	friend
ရုက္ခ	tree
သိပ္ပ	arts and science

3.3 Pali Words without Conjunct Consonants

The second group consists of Pali words that are written without using a conjunct consonants mechanism. These Pali words are similar in encoding to normal Myanmar words. The Pali words in this group contain one or more syllables and they follow the Myanmar syllable structure. Thus, these Pali words similar to Myanmar words in syntax and cannot be distinguished from Myanmar ones without knowledge of the Pali language. Some Pali words had so long been incorporated into Myanmar language that they have almost become native Myanmar words in usage. Encoding of a Myanmar Pali word without a subscripted consonant form is shown below.

မဟာ (mahā) 1019 101F 102C great

Table 3: Pali words without conjunct consonants

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

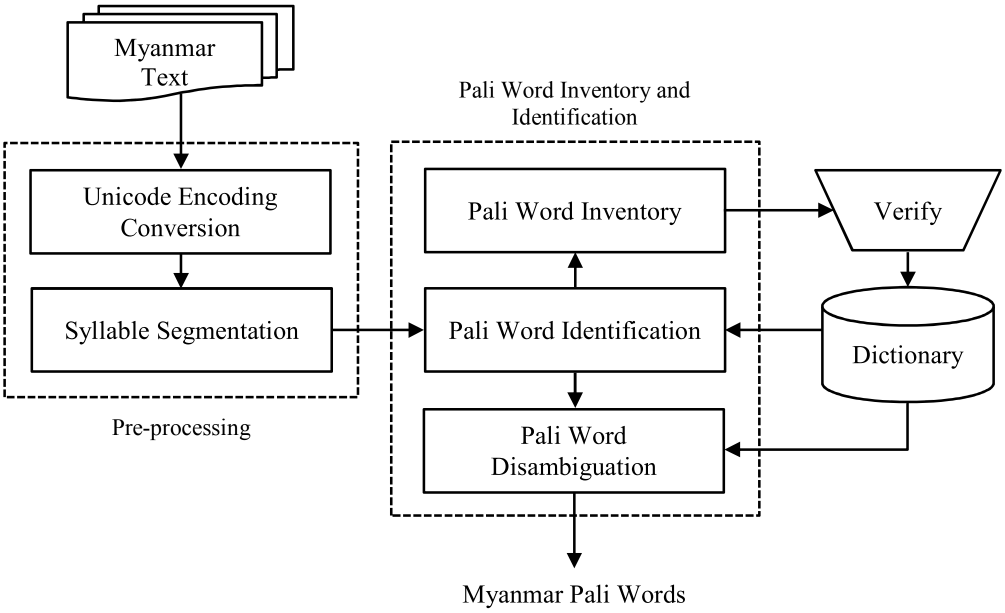


Fig. 2 System Architecture for Pali Word Identification

4. System Architecture

This section presents the overview design of the Pali word identification system. There are two major components in the proposed system. The first one is the pre-processing step. The pre-processing step is composed of 1) encoding conversion process and 2) syllable segmentation process. The second step contains three major components of the system: 1) Pali word identification process 2) Pali word inventory process and 3) Pali word disambiguation process. Details of each process are explained in the following sections.

4.1 Encoding Conversion

Most of the Myanmar articles available on the Internet are written in "Zawgyi-one" font, which is a Pseudo-Unicode font partially following the Unicode encoding standard for Myanmar script. The identification algorithms of this research is designed based on the Unicode encoding standard for Myanmar script [6] and hence, most of the Myanmar texts collected over the Internet need to be converted to a fully Unicode-compliant font for language processing purpose. In this research, a fully Unicode-compliant Myanmar font, known as Myanmar3, developed by Myanmar Unicode and NLP Research Centre has been used.

4.2 Syllable Segmentation

Since Myanmar script does not use space or other delimiters to separate words, word boundary detection presents a challenge in many NLP tasks. Syllables are building blocks of words and knowing syllable boundaries is very helpful in determining word boundaries of Myanmar text.

This research applies the left to right syllable-level longest matching approach when searching the input text against the Pali words stored in the dictionary. Myanmar words are composed of one or more syllables and performing the longest matching method on syllables not only eliminates the character-level ambiguities but also speeds up the matching of words in a dictionary-based look-up approach.

A Myanmar syllable is composed of one initial consonant together with optional medials, vowels and dependent various signs (e.g., က, ကာ, ကား). There are also independent vowels which can stand as a single syllable without joining with other characters (e.g., ဦ, ဤ, ဪ).

Syllable segmentation is carried out as a pre-processing step in this research. Input Myanmar texts are syllable segmented by using a syllable segmentation program developed by Z. M. Maung and Y. Mikami (2008). The program accepts paragraph-level Myanmar text as input and generates syllable segmented Myanmar text as output. The input text is first converted to equivalent category code string and the code string is syllable segmented according to the pre-defined syllable breaking rules. The syllable breaking rules give a break status, whether to mark a syllable boundary or not, for each pair of character codes. Then, the segmented code string is converted back into Myanmar text string and the syllable segmented Myanmar text is shown as output of the system. The flowchart for the syllable segmentation process is shown in Figure 3.

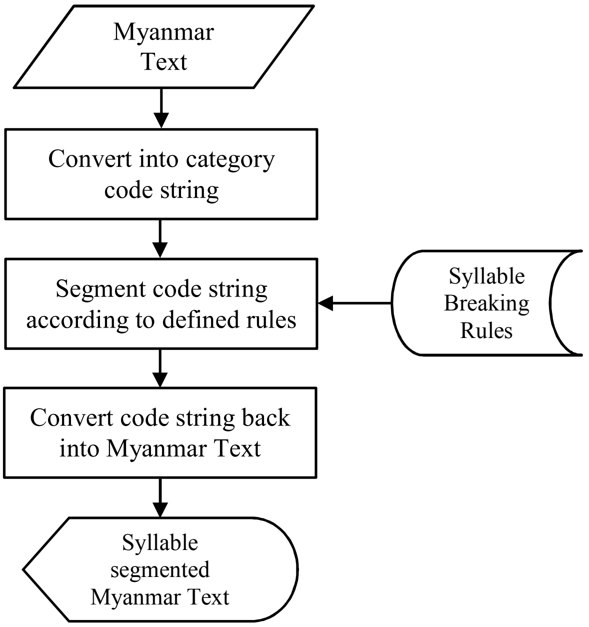

Myanmar Text
Convert into category code string
Segment code string according to defined rules
Syllable Breaking Rules
Convert code string back into Myanmar Text
Syllable segmented Myanmar Text

Fig. 3 Syllable Segmentation Process

4.3 Pali Word Identification

Pali words in Myanmar text are identified by using a dictionary-based syllable-level longest matching method. The longest matching method scans the input text by sequentially reading each syllable from the input text and matching the read syllable against stored Pali words in the dictionary. If a match is found, the method reads next syllable from the input text, compares it to the stored word list until a longest matched word is found and marks it as a Pali word. If there is no match, the method just skips the last read syllable, starts reading the next syllable and begins the longest matching process again until it reaches the end of the text. The longest matching method is used together with a dictionary containing 3477 Myanmar-adopted Pali words.

4.4 Pali Word Inventory

The Pali word inventory process is used to collect newly found Pali words from the input text. The Pali dictionary initially contains 3477 Pali words collected from the "Dictionary of Pali-derivatives" compiled by U Htun Myint (1986) [11]. However, this is not enough and usages of Myanmar Pali words, not in the dictionary, were detected during experiments. The Pali word inventory process is used to collect such newly found Pali words to the dictionary. The process has the ability to recognize Pali words with conjunct consonants since conjunct consonants are encoded using a special Unicode character "U+1039" (Invisible Virama Sign). The inventory process detects them by applying a simple rule to check the invisible virama sign in the input text and performing a forward and backward matching on the input syllables with pre-stored Pali words. If there is no match in the Pali dictionary, then the inventory process shows the text phrase containing the potential Pali word for user verification.

There are usages of Myanmar words written in conjunct consonants form (e.g., လိမ္မော်၊ စက္ကူ) and hence it is necessary to distinguish them from the Pali conjunct consonants. During initial trainings of the system, manual verification step is added to solve this problem. After the verification, only valid Pali words were added to the Pali dictionary and Myanmar words with conjunct consonants were saved in a separate wordlist. During experiments, 215 Myanmar words in conjunct consonants forms were collected and saved to the Myanmar wordlist. The wordlist is useful for distinguishing Pali conjunct consonants from Myanmar ones and the manual verification step could be eliminated depending on the comprehensiveness of the Myanmar wordlist collected during the training and test of the system.

4.5 Pali Word Disambiguation

During experiments, some Myanmar words were incorrectly identified as Pali words. Majority of them are mono-syllabic Pali words that have a different meaning in Myanmar language (See Table 4). The accuracy of the identification system slightly dropped due to the incorrect identification of such ambiguous Pali words. The Pali word disambiguation process was designed to solve such problems. Since most of the ambiguous Pali words are mono-syllabic, the mono-syllabic Pali words are first filtered from the identified result. Then, these words are matched with a Myanmar wordlist before identifying them as Pali words. The Myanmar wordlist contains compound words that include ambiguous mono-syllabic Pali words. A Myanmar word is formed by combining one or more syllables and a mono-syllabic Pali word can become a subset syllable of a compound Myanmar word. Therefore, ambiguous Pali words can be eliminated by matching them with Myanmar words. If a mono-syllabic Pali word is found in a compound Myanmar word, then this word is eliminated from the list of identified Pali words. The syllable-level forward and backward matching method is used to match ambiguous Pali words with the Myanmar wordlist.

Table 4: Examples of ambiguous Pali words

Ambiguous Pali Word (Pali Origin)	Pali Meaning	Myanmar Meaning	Myanmar Compound Word
နည်း(နယ)	way	little	အ+နည်း+ငယ်
ဝါ (ဝဿ)	rain	yellow	အ+ဝါ +ရောင်
စိတ်(စိတ္တ)	mind	piece	အ+စိတ်+အ+ပိုင်း

5. Experiments and Results

5.1 Corpus Creation

The main purpose of this research is to identify adopted Pali words from the real-world Myanmar text. Therefore, various Myanmar articles were collected from sources such as online magazines, blogs and social sites. The articles contain different categories of text such as news, sports, editorials, novels, religious, health and interviews etc. Two different corpora were created for training and evaluation purposes. The training corpus contains 163 Myanmar articles containing 8895 sentences and the test corpus contains 100 Myanmar articles containing 3641 sentences. Most of the articles collected over the Internet were written in a non-standard encoded font and they were converted to the Unicode standard encoded font for the research purpose. The collected Myanmar articles were syllable segmented according to the rule-based method mentioned in previous sections [15]. The syllable segmented corpora were then given as input to the program for identifying Myanmar Pali words. Details of the corpora used in this research are shown in Table 5.

Table 5: Details of training and test corpus

Corpus Type	Articles	Sentences	Syllables	Unique/Total Pali words identified
Training	163	8895	431439	579/12230
Test	100	3641	140390	279/4017

5.2 Evaluation

The program was first trained on a corpus containing 163 Myanmar articles collected from the Internet. The corpus contains 8895 Myanmar sentences, containing a total of 431439 Myanmar syllables. The program recognized 579 unique Pali words, from a total of 12230 Pali words in the corpus. It is found that some Pali words are used very frequently in Myanmar language. The frequency distributions of the most widely used Pali words in the corpora are shown in Table 7. During the training, Pali words that were not already in the dictionary were found and 215 newly found Pali words were added to the Pali dictionary by using the Pali word identification process.

Testing of the program was carried out on a different corpus containing 100 articles collected over the Internet. The test corpus contains 3641 sentences with a total of 140390 Myanmar syllables. The initial test achieved a Precision of 90.45%. During experiments, the program wrongly identified some Pali words that have the same spelling but a different meaning in Myanmar language. The precision of the program dropped slightly due to such ambiguous cases and the program was modified by adding the Pali word disambiguation process to the system. The disambiguation process utilized a Myanmar wordlist to distinguish ambiguous Pali words from the valid Pali words. The wordlist was developed by collecting Myanmar words containing Pali syllables during the training and test phases of the system. The program was tested again after adding the disambiguation process, together with a Myanmar wordlist collected during experiments. The precision of the system improved from 90.45% to 97.59% after doing disambiguation on the identified Pali words. The Pali word inventory process of the system collected 215 new Pali words during experiments and they were added to the Pali dictionary. It is found that the recall of the system increased from 97.42% to 99.04% after updating the dictionary with newly found Pali words. The details of the system performance on different test settings and their results are shown in Table 6.

$$Precision = \frac{no.of\ correctly\ identified\ Pali\ words}{no.of\ identified\ Pali\ words}$$

$$Recall = \frac{no.of\ correctly\ identified\ Pali\ words}{no.of\ valid\ Pali\ words}$$

$$Fmeasure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Table 6: Identification results on test corpus

Disambiguated?	Dictionary Updated?	Precision	Recall	F-measure
No	No	90.45%	97.42%	93.81%
Yes	No	97.59%	97.42%	97.50%
No	Yes	90.45%	99.04%	94.55%
Yes	Yes	97.59%	99.04%	98.31%

6. Error Analysis

The errors of Pali words identification system were mainly caused by 1) missing Pali words that were not in the dictionary 2) ambiguous Pali words that have the syntax but carry different meanings for compound Myanmar words.

6.1 Missing Pali Words in the Dictionary

The system employs a dictionary initially containing 3477 Myanmar Pali words. However, this is not enough for detecting Pali words from the real-world Myanmar text. Missing Pali words include 1) proper nouns including names of monks, people and places etc. and 2) occurrences of pure Pali words in religious texts). In this research, the Pali word inventory process was used to identify conjunct consonant Pali words, as an effective way to detect missing Pali words from the input text, and add them to the dictionary.

6.2 Ambiguous Pali Words

Some Pali words have the same syntax but different meanings in Myanmar and Pali language. Most of the ambiguous words are mono-syllabic Pali words. In Myanmar language, Myanmar words are formed by joining one or more syllables. During experiments, the system wrongly identified some Myanmar syllables as Pali words. In such cases, mono-syllabic Pali words were actually part of a compound Myanmar word, carrying a totally different meaning in Myanmar language. Examples of mono-syllabic ambiguous Pali words can be seen in Table 4. Ambiguous Pali words cannot be identified by referencing their syntax alone. They have the same syntax but different semantic representations when combined with other syllables. In this research, the disambiguation process for mono-syllabic Pali words was designed and implemented, with the use of a special Myanmar wordlist. There were few cases of bi-syllabic ambiguous Pali words. However, occurrences of such cases are few and negligible.

7. Conclusion

Pali language has a great influence on Myanmar language and usages of Myanmar-adopted Pali words are inevitable in Myanmar text. In this research, Myanmar-adopted Pali words from the real-world Myanmar text were identified by using a dictionary-based syllable-level longest matching approach. The algorithm was trained on a corpus containing 8895 sentences and newly found Pali words were added to the Pali dictionary by using an inventory process. During experiments, cases of ambiguous Pali words having a different meaning in Myanmar language were found and this problem was solved by performing the disambiguation on the identified Pali words using a special wordlist of Myanmar words. The accuracy of the system was tested on a different corpus containing 3641 sentences and the program achieved a precision of 97.59%, recall 99.04% and F-measure of 98.31%. Among the identified Pali words, the top most frequently used Myanmar Pali words are reported in this paper, together with their word-lengths in terms of syllables. In this research, the percentage of Pali words taking up the overall Myanmar text was calculated. According to the experiment results, it can be concluded that Pali text represent 5-8% of the Myanmar text (counting was done in terms of syllables and whitespaces and non-Myanmar characters were removed from the input text). The percentage of Pali words taking up in Myanmar text may vary depending on the type or genre of the article used and it is found that the Pali text percentage increased to approximately 20% of the total text in some religious articles. Pali words are unavoidable in Myanmar text and identifying the Myanmar-adopted Pali words from the real-world Myanmar text and developing electronically available language resources of them will improve many NLP tasks of Myanmar language such as spelling checking, text categorization, text-to-speech synthesis and machine translation etc.

Appendix

The frequency distributions of the most widely used Pali words identified in this research are shown in Table 7. The Pali words were identified from a corpus developed by combining the training and test corpus of this research. Currently, there is no standardized corpus for Myanmar language and the availability of a balanced, standard corpus representing the Myanmar language will make the findings of this research more comprehensive.

Table 7: Top ten Myanmar Pali words identified in corpus

No.	Pali Word	Syllable Count	Frequency
1	ဥပဒေ	3	987
2	ကိုယ်	1	911
3	သမ္မတ	3	752
4	ပညာ	2	555
5	စိတ်	1	553
6	ကိစ္စ	2	503
7	သဘော	2	414
8	ဆရာ	2	351
9	ဌာန	2	332
10	နည်း	1	331

Table 8: Percentage of Pali words in Myanmar text

Corpus Type	Unique/Total Pali words identified	% of Pali (Syllables)	% of Pali (Characters)
Training	579/12230	7.75%	5.71%
Test	279/4017	6.09%	4.26%

Table 9: Top ten Myanmar Pali words by syllable-lengths

#	1-gram	2-gram	3-gram	4-gram	5-gram
1	ကိုယ်	ပညာ	ဥပဒေ	ပဋိပက္ခ	ကုလသမဂ္ဂ
2	စိတ်	ကိစ္စ	သမ္မတ	ပုဂ္ဂလိက	သတိပဋ္ဌာန
3	နည်း	သဘော	ဥက္ကဋ္ဌ	ဗုဒ္ဓဘာသာ	ပုဂ္ဂလဓိဋ္ဌာန်
4	ဗိုလ်	ဆရာ	တက္ကသိုလ်	ဗဟုသုတ	သမဝါ ယမ
5	ခေတ်	ဌာန	အဓိက	သယံဇာတ	အရဟတ္တဖိုလ်
6	ဓာတ်	ဒေသ	ပထမ	ဝိပဿနာ	ဝိသမလောဘ
7	ပုဒ်	ဘဝ	ပြဿနာ	သုတေသန	သေနင်္ဗျူဟာ
8	စက်	ဆန္ဒ	ဒုတိယ	အနုပညာ	ဟီရိဩတ္တပ္ပ
9	လာဘ်	အာဏာ	သဘာဝ	အနေကဇာ	ဓမ္မာစရိယ
10	ဉာဏ်	ဘာသာ	အဓိပ္ပါယ်	မဟာဗျူဟာ	သဗ္ဗညုတဉာဏ်

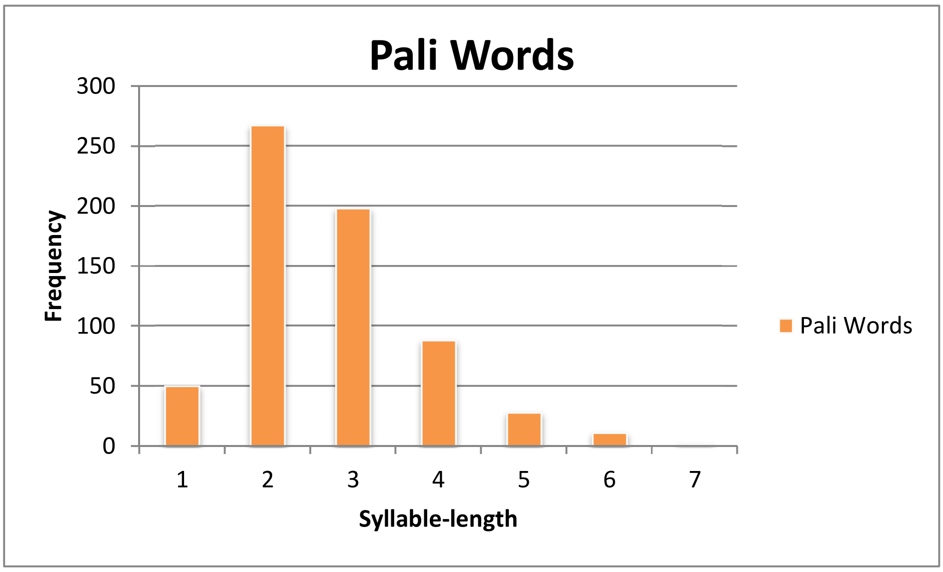


Fig. 4 Frequency Distribution of Myanmar Pali Words by Syllable-lengths

References

[1] C. Duroiselle, A Practical Pali Grammar of the Pali Language, Third Edition, 1997, originally printed at the British Burma Press, 1921

[2] H. H. Htay and K. N. Murthy, "Myanmar Word Segmentation using Syllable level Longest Matching", Proceedings of the IJCNLP-2008 Workshop on Asian Language Resources, Jan 11-12, 2008, Hyderabad, India.

[3] H. T. Thet, et al., "Word Segmentation for the Myanmar Language", Journal of Information Science, October 2008 volume 34, Issue 5, pp. 688-704.

[5] M. H. Bode, The Pali Literature of Burma, The Royal Asiatic Society of Great Britain and Ireland, first published 1909, reprinted 1966.

[6] M. Hosken, "Representing Myanmar in Unicode: Details and Examples Version 3", Unicode Technical Note #11, http://www.unicode.org/notes/tn11/tn11-3.html

[7] O. Sornil and P. Chaiwanarom, "Combining Prediction by Partial Matching and Logistic Regression for Thai Word Segmentation", Proceedings of the 20th International Conference on Computational Linguistics, 2004.

[8] Peter T. Daniels and William Bright, 1996, “The World’s Writing Systems”, Oxford University Press.

[9] T. Theeramunkong and S. Usanavasin, "Non-Dictionary-Based Thai Word Segmentation Using Decision Trees", Proceedings of the First International Conference on Human Language Technology Research, 2001.

[10] The Unicode Consortium, The Unicode Standard, Version 6.2.0, (Mountain View, CA: The Unicode Consortium, 2012. ISBN978-1-936213-07-8), http://www.unicode.org/versions/Unicode6.2.0/

[11] U. H. Myint, Dictionary of Pali-derived Words, First Edition, Universities' Press, Yangon, Myanmar, 1986.

[12] W. Aroonmanakun, "Collocation and Thai Word Segmentation", Proceedings of Joint International Conference of SNLP-Oriental, COCOSDA-2002.

[13] Y. Poowarawan, "Dictionary-based Thai Syllable Separation", Proceedings of the Ninth Electronics Engineering Conference, 1986.

[14] Yuzana and K. M. L. Tun, “Myanmar-Pali Text Syllabification by Utilizing Pattern Generation Tactic”, Proceedings of the 7th International Conference on Computer Applications (ICCA2009), 2009, Yangon, Myanmar.

[15] Z. M. Maung and Y. 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.

First Author Zin Maung Maung received bachelor and master degrees of Computer Science from the University of Computer Studies, Mandalay, Myanmar. He has also obtained a master of engineering degree, M.E., from the Nagaoka University of Technology, Japan, specializing in Management and Information Systems Engineering. Currently, he is pursuing a Ph.D. course at the University of Computer Studies, Mandalay, Myanmar and his current research interests includes Natural Language Processing and Web Mining, especially in the areas of localization and information retrieval techniques applicable on Myanmar language.