# Dictionary Based Machine Translation System for Pali to Sinhala

**R. M. M. Shalini[1], B. Hettige[2]**
[1]Faculty of Humanities and Social Sciences, University of Sri Jayewardenepura,
Nugegoda, Sri Lanka
[2]Department of Computer Science, Faculty of Computing,
General Sir John Kotelawala Defence University, Rathmalana, Sri Lanka
rmmshalini@gmail.com[1], budditha@kdu.ac.lk[2]

## *Abstract*

*Machine translation systems are language translation tools that are also capable to be used as language learning tools. This paper presents a machine translation system that has been developed as a language learning tool for the Pali to Sinhala. This Pali to Sinhala translation tool can translate simple Pali sentences into Sinhala through the dictionary-based approach. The translation system comprises of three modules, namely the Pali morphological analyzer, the dictionary-based translator, and the Sinhala morphological generator. The Pali morphological analyzer uses an affix-spiriting approach to identify a relevant Pali root word for the existing Pali word. The Pali to Sinhala translator identifies an available Sinhala-based word for the existing Pali-based word with the support of the Pali-Sinhala dictionary. The Sinhala morphological generator generates appropriate Sinhala words by using the Sinhala root word and the relevant morphological information of the Pali word. This translator uses word-level translation and gives attention only to source and target morphology. The Pali Sinhala translator has been successfully tested for simple Pali sentences as a language learning tool for the grade 6-9 Dhamma schools.*

***Key Words:*** *Language Learning Tool, Machine Translation System, Morphological Analyzer, Translator, Morphological Generator.*

## 1. Introduction

Language Translation is a communication of the meaning of a source-language text by means of an equivalent target-language text through the human concern [1]. This language translation process can be automated through the computers and still is a challenging research task of the Natural Language Processing. Automated Language translation process is normally called as machine translation, and that can be done through the machine translation systems [2]. There are thousands of machine translation systems are available with different translation approaches including google translator [3], Anusaaraka [4], Systran [5] etc. Theses translation systems gives different performance based on the language pair and the approach used to translate. These Machine Translation approaches are normally categorized with considering the level of attention made on morphology, syntax, and semantics of the source and target [6]. According to the translation pyramid, the direct translation is the lowest level machine translation approach. It is not much level of considering syntax and semantics on both source and target languages. Therefore, direct translation approach is more suitable for language pairs with a closed relationship in between syntax semantics and morphology. Figure 1 shows the translation pyramid of the machine translation.

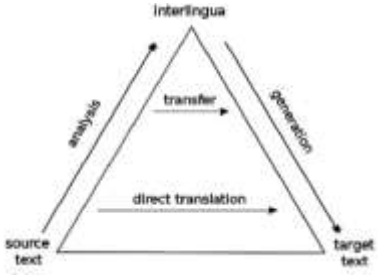


Figure 1: Translation Pyramid

For instance, Sinhala and Pali are the two related languages and has closed relation on syntax and morphology.

Pali is one of the Indo-Aryan languages that has some syntactical, semantical, and morphological relations with Sinhala. Further, Pali is widely studied by the Sri Lankan dhamma school students; thus, it is the language of Buddhism. With this viewpoint, a machine translation system has been developed as a language learning tool for the Pali to Sinhala. This Pali to Sinhala translation tool can translate simple Pali sentences into Sinhala through the dictionary-based approach. This tool would be

beneficial for the Buddhist priests in Pirivenas, Students and who wish to learn Pali language.

The rest of the paper is organized as follows. Section 2 reports brief introduction on Machine Translation systems, morphological analysis and morphological generation with some note on existing systems. Then section 3 reports computational grammar on Pali and Sinhala language, which is required to develop dictionary, based machine translation system. The section 4 presents design and implementation of the Pali-Sinhala machine translation system and section 5 demonstrate how systems work on Simple Pali sentence. Finally, section 6 present conclusions with a note on further works.

## 2. Related Work

This section briefly describes related works for the machine translation, including Morphological analysis and generation.

### *A. Machine Translation*

The Machine Translation is a sub-area of the Natural language processing which is identified during early days of Artificial Intelligence (AI) [2]. The process of the Machine Translation can be described simply as decoding the meaning of the source text and re-encoding this meaning in the target language. However, due to the complexity of the natural languages, development of machine translation system has become a research challenge.

Further, according to the translation methods, Machine translation systems can be broadly categorized into two groups, namely, the direct translation system and the indirect translation system. The direct translation system translates source language to target language by using word-to-word or phrase-to-phrase mapping. Indirect translation systems use an Interlingua or transfer approach. In addition to the above, in general, machine translation approaches, machine translation systems can be classified into seven categories, namely Human-assisted, Rule-based, Statistical, Example-based, Knowledge-based, Hybrid and Agent-based [6].

Dictionary-based translation is taken as a basic approach to machine translation. This type of machine translation applications is easy to implement and that attempt quite competent tools for retrieving translations for unknown search keys. Commonly, dictionary-based translation is basically a translation with a help of a bilingual dictionary. In addition to the bilingual dictionary most of the systems consists of Morphological analyzer and a generator for the source and target languages. For instance, Sindhu and others have developed a dictionary-based machine translation from Kannada to Telugu [7]. The translation system comprises with five components namely Morph analyzer, dictionary, transliteration, transfer grammar and the morph generator.

Consider the Pali language only very limited research has conducted to translate Pali text into other languages. Phoson and others have developed a prototype Rule-based Machine Translation system for Pali to Thai [8]. This system capable to analysis Pali language structure of the input sentence and re-generate Thai language structure. The next section briefly describes Morphological analysis and generation for Machine translation.

### *B. Morphological Analysis*

Morphological analysis is one of the key components in the rule-based machine translation systems. That helps to identify the based word and the grammar for the existing source language words. In addition to the above morphological analyzer is must essential when language has a strong morphology. In the view of this, the morphological analyzer is one of the essential tools for the Asian language processing systems. Thus, some Asian countries including India, Japan, and Thailand have also developed morphological analyzers for their language processing [9]. For Instance, Anusaaraka has been designed to translate among major Indian languages and its morphological analysis is based on the word paradigms [4]. A Sinhala morphological analyzer has been developed for purpose of machine translation [10]. The system uses affix stripping approach to analyze the Sinhala words [11].

### *C. Morphological Generation*

Morphological generation is the backward process of the Morphological analysis. A morphological generator is also an essential tool used in Natural Language Processing (NLP) applications for the generation of final fully inflected words from the root and a set of morphological properties. Note that, highly inflected languages like the Dravidian languages of Tamil, Malayalam, Telugu, and Sinhala requires a morphological generator for produce grammatically correct results. Further,

morphological generator plays an indispensable role in target language sentence generation in Machine Translation (MT) systems. It is also used in Information Retrieval (IR) systems in the front-end for query expansion [12]. Further, the aim of morphological generation is to produce the inflected form of a word according to the features and values in the Feature Structure. It is also necessary to reuse the linguistic resources created for analysis purpose. From a practical point of view, morphological generation is the inverse process of analysis, namely the process of converting the internal representation of a word to its surface form. The same rule definitions can be used to generate the desired word form as used for analysis. Affix stripping approach is an approach to morphological analysis that checks the root form from the available word by adding or removing letters. This type of approach can be easily applying for grammatically rich languages like Sinhala and Pali. With the above view, a research has been conducted to develop Pali to Sinhala machine translation system through the direct transfer approach.

## 3. Computational Grammar for Pali and Sinhala

This section briefly introduces Pali and Sinhala Morphology, which are used to analyses and generate the morphology of both languages.

### *A. The Pali language*

Pali is an Indo- European language derived from Sanskrit that for centuries has been the canonical and liturgical language of Theravada Buddhism. Pali continues to be used throughout Southeast Asia as a religious language - in prayer and ritual, as well as in the study scripture and the composition of commentaries and other religious documents. Even into the 20th Century. Some Theravada elders continued to be able to converse in the ancient tongue. The term ‘Pali' originally had the meaning of ‘canonical'; over the years of use preserving the Pali Canon [13].

The Pali alphabet consists of 41 letters, eight vowels and thirty- three consonants [14]. Long and short vowels are only contrastive syllables: in closed syllables, all vowels are always short. Short and long ඉ and ඔ are in complementary distribution: the short variants occur only in closed syllables; the long variants occur only in open syllables. Short and long ඉ and ඔ are therefore not distinct phonemes

Pali is a highly inflected language, in which almost every word contains, besides the root conveying the basic meaning, one or more affixes (usually suffixes) which modify in some way. Nouns are inflected for gender, number, and case; verbal inflections convey information about a person, number, tense, and mood.

Pali nouns inflect for three grammatical genders (masculine, feminine and neuter) and two numbers (Singular and plural). The nouns also, in principle, display nine cases.

However, in many instances, two or more of these cases are identical in form; this is especially true of the genitive and dative cases.

Table 1: inflection for the Pali Nouns

| | පුරුෂ ලිංග (බුද්ධ ශබ්දය) | |
|---|---|---|
| | ඒක වචන | බහු වචන |
| ප්‍රථමා | බුදෙධා | බුද්ධා |
| දුතියා | බුද්ධං | බුදේධ |
| තතියා | බුද්ධෙන | බුද්ධේහි |
| චතුර්ථී | බුද්ධාය | බුද්ධානං |
| පඤ්චමී | බුද්ධා | බුද්ධේහි |
| ෂෂ්ඨී | බුද්ධස්ස | බුද්ධානං |
| සප්තමී | බුදෙධ | බුදේධසු |

| | ස්ත්‍රී ලිංග (කඤ්ඤා ශබ්දය) | |
|---|---|---|
| | ඒක වචන | බහු වචන |
| ප්‍රථමා | කඤ්ඤා | කඤ්ඤා |
| දුතියා | කඤ්ඤං | කඤ්ඤා |
| තතියා | කඤ්ඤාය | කඤ්ඤාහි |
| චතුර්ථී | කඤ්ඤාය | කඤ්ඤානං |
| පඤ්චමී | කඤ්ඤාය | කඤ්ඤාහි |
| ෂෂ්ඨී | කඤ්ඤාය | කඤ්ඤානං |
| සප්තමී | කඤ්ඤායං | කඤ්ඤාසු |

| | නපුංසක ලිංග (ඵලං ශබ්දය) | |
|---|---|---|
| | ඒක වචන | බහු වචන |
| ප්‍රථමා | ඵලං | ඵලානි |
| දුතියා | ඵලං | ඵලං |
| තතියා | ඵලේන | ඵලේහි |
| චතුර්ථී | ඵලාය | ඵලානං |
| පඤ්චමී | ඵලේන | ඵලේහි |

There are Three Tenses, two voices, two numbers, and three Persons in the conjugation of Pali verbs.

The table 2 shows some inflection form for the Pali word.

Table 2: Pali verb inflections

| Person | Number | Present tense | Past tense | Future tense |
|---|---|---|---|---|
| First | Singular | ගච්ඡාමි | ගවජි | ගච්ජිස්සාමි |
| First | Plural | ගච්ඡාම | ගවජිංසු | ගච්ජිස්සාම |
| Second | Singular | ගච්ඡසි | ගවෙජා | ගච්ජිස්සසි |
| Second | Plural | ගච්ඡථ | ගච්ජිත්ථ | ගච්ජිස්සථ |
| Third | Singular | ගච්ඡති | ගවජිං | ගච්ජිස්සති |
| Third | Plural | ගච්ඡන්ති | ගවජිංම්හ | ගච්ජිස්සන්ති |

Further, the Sinhala language is closely related to the Sinhala language. The next section briefly reports some required computational grammar for Sinhala.

### *B. The Sinhala language*

Sinhala is the main language of Sri Lanka. Approximately 21 million speak Sinhala as their mother tongue and many others in the country such as Tamils and Muslims speak it as a second language. There are also considerable numbers of Sinhala speakers in Singapore, Thailand, Canada and the United Arab Emirates. The Sinhala language belongs to the Indo-Aryan subdivision of the Indo-European language family. It was developed with the help of Pali and Sanskrit that can be mentioned as sacred languages of the Sri Lankan Buddhists. Because of the European colonization, Portuguese, Dutch, and English have also influenced the Sinhala language.

The Sinhala language has its own alphabet and writing system, which was developed for far too long. Sinhala is the only language in the world that uses the Sinhalese alphabet. There are 61 characters in Sinhala alphabet including 18 vowels, 41 consonants, and 2 semi-consonants.

There is a rich morphology system in the Sinhala language. These systems were inherited from Brahmi script to the Sinhala language [16]. Both nouns and verbs are generated into many forms with morphology in Sinhala language usage. Sinhala verb is divided into general classes, namely, a transitive verb (sakarmaka) and intransitive (Akarmaka). Further, these two verb categories are inflected for voice (karaka), mood (vidi), tense (kala), number (wachana) and person (purusha). Voice can be either active or passive. There are four types of moods, namely, indicative, optative, imperative and conditional. The Sinhala language has only three tenses [16]. They are Past tense, Present tense, and future tense. Main verb (Akkyathaya) participate three types of inflections namely person, number, and sex. Table 3 shows inflection forms of a verb in the active voice and the passive voice.

Table 3: inflection forms of a verb in the active voice and the passive voice

| **Person** | **Number** | **Present** | **Past** | **Future** |
|---|---|---|---|---|
| First | Singular | බලමි | බැලීමි | බලන්නෙමි |
| First | Plural | බලමු | බැලීමු | බලන්නෙමු |
| Second | Singular | බලහි | බැලීහි | බලන්නෙහි |
| Second | Plural | බලහු | බැලීහු | බලන්නෙහු |
| Third | Singular | බලයි | බැලී | බලන්නේය |
| Third | Plural | බලති | බැලූ | බලන්නෝය |

Inflection form of the Sinhala verbs (active)

| Person | Number | Present | Past | Future |
|---|---|---|---|---|
| First | Singular | බැලෙමි | බැලිණිමි | බැලෙන්නෙමි |
| First | Plural | බැලෙමු | බැලිණිමු | බැලෙන්නෙමු |
| Second | Singular | බැලෙහි | බැලිණිහි | බැලෙන්නෙහි |
| Second | Plural | බැලෙහු | බැලිණිහු | බැලෙන්නෙහු |
| Third | Singular | බැලෙයි | බැලිණි | බැලෙන්නේය |
| Third | Plural | බැලෙති | බැලුණු | බැලෙන්නෝය |

Inflection form of the Sinhala verbs (passive)

## 4. Design and Implementation

This section briefly describes design and implementation of the Pali to Sinhala translator. Figure 2 shows the design of the Pali Sinhala translator. The translator consists of three modules namely Pali Morphological Analyzer, Pali to Sinhala translator and Sinhala Morphological generator. The task of each module is reported in the below.

### *A. Pali Morphological Generator*

Pali Morphological generator is the key module in the translator that reads Pali word as an input and provide root word and the grammatical information for each word. This Pali Morphological generator uses affix spiriting approach to handle the Pali morphology. Under this research, we have identified irregular and regular forms of the Pali noun and words. Irregular words are put into Pali dictionary and root word indentation table has been developed for the regular words. In the Pali language normally noun participate gender-based conjugation. In addition to that, conjugation can be divided into the last sound of the noun call "Anthaya" The following Table shows the last sound ‘ඇ' for the Pali word "බුද්ධ"

Table 4: Apexes table for the Singular plural noun with last sound ‘අ.

| Case | Example | Add | Remove |
|---|---|---|---|
| පඨමා | බුද්ධො | ධො | ධ |
| දුතයා | බුද්ධං | ං | |
| තතයා | බුද්ධෙන | ධෙන | ධ |
| චතුත්ථ | බුද්ධාය | ාය | |
| කරණ | බුද්ධෙන | ධෙන | ධ |
| පඤ්චම | බුද්ධා | ා | |
| ඡට්ඨ | බුද්ධස්ස | ස්ස | |
| සත්තම | බුද්ධෙ | ධෙ | ධ |
| ආලපන | බුද්ධ | | |

With the same method, we have identified affix spiriting table for the Pali Noun and Verbs. By using these grammatical rules, the root word of the Pali word has been identified. This analysis has been done with the support of the Pali lexical database that consists of regular nouns, regular verbs, irregular nouns, verbs and other words.

### *B. Pali to Sinhala Translator*

The Pali to Sinhala Translator uses Pali-Sinhala dictionary and identify suitable based word for the existing Pali word. This translator does not consider word level ambiguity and semantics issues. Word is not available in the dictionary translator provides an error message and recorded it as an out-of-vocabulary word. The bilingual dictionary consists of Pali words and related Sinhala words with their grammatical category.

### *C. Sinhala Morphological Generator*

The Sinhala Morphological generator uses to generate appropriate Sinhala word for the given Sinhala based word. This Morphological generator uses existing affix spiriting rules that are developed under the BEES project [5] [13]. In here, we use Sinhala Pali sentence, which is available on the Dhamma schoolbooks, therefore, limited numbers of rules are used to generate Sinhala words.

After morphological generating, each word is shown as a translated output of the system. Note that word order of the Pali and Sinhala are closely related to each other. Thus, syntax level generation has not been applied in the current system.

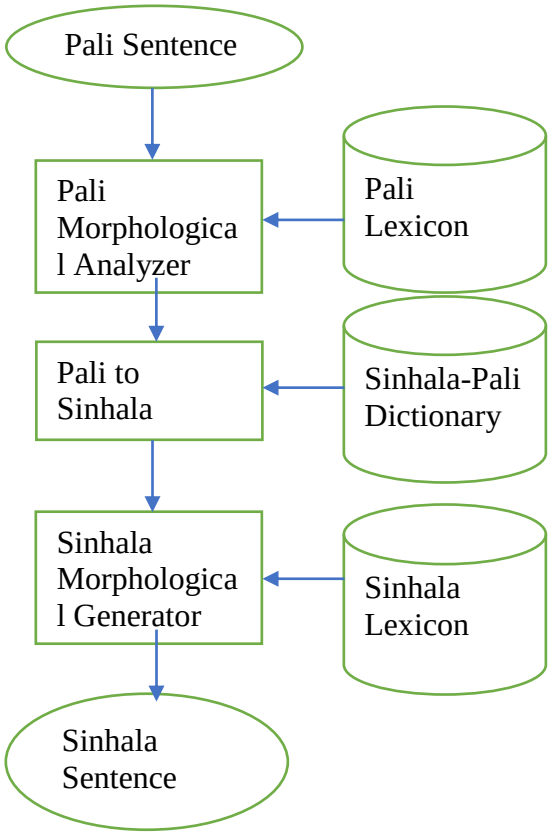


Figure 2: Top-level Design of the Pali to Sinhala Translator


Figure 3: Graphical user interface of the Pali to Sinhala Translator

## 5. How System works

This section reports how system translates a given Pali sentence into Sinhala. For Instance, System reads "අහං ගාමං ගච්ඡාමි" (Ahng gamang gachchami) as the input sentence. Then Pali

Morphological analyzer reads words "අහං", "ගාමං" and "ගච්ඡාමි" as the input and identifies based word and the morphology of each word. The table xx shows the result of the Pali morphological analysis. Then Pali-Sinhala bilingual dictionary used to identify each based word for the input base word. Because of the dictionary, based translation system reserved the මම ගම යනවා as the corresponding Sinhala based words. Then Sinhala Morphological generators read these Sinhala words and grammatical information identified by the Pali Morphological generator, generates appropriate Sinhala words for the existing Sinhala based word. After Morphological generation, the translated sentence is shown as "මම ගමට ගමි". Note that, Pali and Sinhala languages are related to each other and syntax of both languages are almost same. Therefore, syntax level attention for translation is not required.

## 5. Conclusion and Further works

This paper presented design and implementation of the Pali Sinhala translator that was translated Pali sentence into Sinhala with the dictionary-based approach. The system also comprises of Pali Morphological analyzer and Sinhala Morphological generator to analyze and generate Pali and Sinhala words to provide an accurate translation. The Pali Morphological generator has been developed to analyze Pali words. The Pali Morphological analyzer uses affix spiriting rules that are used for word declension. Sinhala Morphological generator also uses Sinhala Morphological rules that are available on the BEES project. This translator uses word-level translation and gives attention only to source and target morphology. The Pali Sinhala translator has been successfully tested for simple Pali sentences as a language learning tool for the grade 6-9 Dhamma schools.

## References


[1] "Translation," *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Translation [Accessed: 18-Oct-2017].

[2] J.Hutchins, "Machine Translation: past, present, future" , New York: Halsted Press, 1986.

[3] "Google Translate." [Online]. Available: https://translate.google.com/. [Accessed: 24-Oct-2017].

[4] A. Bharati, V. Chaitanya, A. P. Kulkarni, and R. Sangal, "Anusaaraka: Machine Translation in Stages," ArXiv Prepr. Cs0306130, 2003

[5] Systran: Past and Present, [Online]. Available: https://lilab.unibas.ch/staff/tenhacken/Applied-CL/3_Systran/3_Systran.html#history, [Accessed: 18-Oct-2017].

[6] B. Hettige and A. S. Karunananda, "Existing Systems and Approaches for Machine Translation: A Review," in *Sri Lanka Association for Artificial Intelligence (SLAAI)*, Moratuwa, 2011.

[7] D. V. Sindhu and B. M. Sagar, "Dictionary Based Machine Translation from Kannada to Telugu," IOP Conf. Ser. Mater. Sci. Eng., vol. 225, p. 012182, Aug. 2017.

[8] N Phonson, a Rule-Based Machine Translation system from Pali to Thai, MSc, Thesis, URL: https://www.ict.mahidol.ac.th /research/thesis/files/2001_25THE.pdf

[9] A. Hatem, N. Omar, and K. Shaker, "Morphological analysis for rule based machine translation," in 2011 International Conference on Semantic Technology and Information Retrieval (STAIR), 2011, pp. 260–263.

[10] B. Hettige and A. S. Karunananda, "A Morphological Analyzer to Enable English to Sinhala Machine Translation," in *International Conference on Information and Automation, 2006. ICIA 2006*, 2006, pp. 21–26.

[11] B. Hettige, "A Computational grammar of Sinhala for English-Sinhala machine translation," M.Phil Thesis, University of Moratuwa,Sri Lanka, Moratuwa, 2011.

[12] A. Pirkola, T. Hedlund, H. Keskustalo, and K. Järvelin, "Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings," Inf. Retr., vol. 4, no. 3–4, pp. 209–230, Sep. 2001.

[13] "The Pali Language and Literature." [Online]. Available: http://www.palitext.com/subpages/lan_lite.htm. [Accessed: 18-Oct-2017].

[14] P. Buddadhatha, *Pali Bshawa Tharanaya*. Rathna Book Prakashakayo, 2015.

[15] J. B. Disanayake, "BasakaMahima 6: Prakurthi", Colombo 10, Sri Lanka: S. Godage and Brothers, 2000.

[16] A. M. Gunasekera"A Comprehensive Grammar of the Sinhalese Language", New Delhi, India : AES Reprint, 1986.

[17] B. Hettige and A. S. Karunananda, "Computational model of grammar for English to Sinhala Machine Translation," in 2011 International Conference on Advances in ICT for Emerging Regions (ICTer), 2011, pp. 26–31.