Rule-Based Pāḷi Romanization System for Myanmar Language

Lei Lei Win

[*] Faculty of Computer Science
[**] University of Computer Studies, Meiktila

DOI: 10.29322/IJSRP.8.9.2018.p8158
http://dx.doi.org/10.29322/IJSRP.8.9.2018.p8158

Abstract- Typically, Myanmar is the most religious Buddhist country with regard to the percentage of the population living as monks and the amount of money spent on religion. Pāḷi is a language that has been widely used in the Buddhist scriptures. Generally, the Pāḷi word are expressed with stacked consonant so that there may be some difficulties to pronounce the Myanmar Pāḷi word. Therefore, this paper presents the text to speech system for Myanmar Pāḷi word by using Romanization rules. Firstly, the input words or sentences in Pāḷi are accepted. Then, these words are checked as it is Pāḷi words or not by using rules for Pāḷi. After that, these Pāḷi words are converted into their corresponding roman symbols by using Romanization rules for Pāḷi. Finally, the system generates the speech of Pāḷi words. The aim of this paper is to help the Buddhists and the new Buddhist students who are unfamiliar with some of the Pāḷi words offer used in the study of Buddhism. According the experimental result, the system achieved the acceptable level of accuracy.

Index Terms- Pāḷi language, Text-to-speech, Romanization rules

I. Introduction

Text to speech system is the conversion of the input text into their corresponding output speech. Myanmar language is the official language and it is tonal, pitch-register and syllable time language. Moreover, the nature of Myanmar language is monosyllabic and analytic language. The sentence order is “subject-object-verb”. It also known as Burmese and it is a member of the Lolo-Burmese grouping of the Sino-Tibetan language family. The language uses a Brahmic script called the Burmese script. The Burmese alphabet consists of 33 letters and 12 vowels, and is written from left to right. It requires no spaces between words, although modern writing usually contains spaces after each clause to enhance readability. Besides the Burmese language, the Burmese alphabet is also used for the liturgical languages of Pāḷi and Sanskrit. Pāḷi is the language used to preserve the Buddhist canon of the Theravada (ထေရဝါဒ) Buddhist tradition (is a branch of Buddhism that uses the teaching of the Pāḷi Canon, a collection of the oldest recorded Buddhist texts, as its doctrinal core), which is regarded as the oldest complete collection of Buddhist texts surviving in an Indian language. Pāḷi is closely related to Sanskrit, but its grammar and structure are simpler. Traditional Theravadins regard Pāḷi as the language spoken by the Buddha himself, but in the opinion of leading linguistic scholars, Pāḷi was probably a synthetic language created from several vernaculars to make the Buddhist texts comprehensible to Buddhist monks living in different parts of northern India. As Theravada Buddhism spread to other parts of southern Asia, the use of Pāḷi as the language of the texts spread along with it, and thus Pāḷi became a sacred language in Sri Lanka, Myanmar, Thailand, Laos, Cambodia and Vietnam. Pāḷi has been used almost exclusively for Buddhist teachings, although many religious and literary works related to Buddhism were written in Pāḷi at a time. So Pāḷi is a spoken language, written in the script of the land where it is used: for example, in Myanmar, it is written in Myanmar script. Strictly speaking Myanmar and Indic scripts are abugidas (alpha-syllabaries) and not alphabets [1].

Myanmar language has been greatly influenced by the Pāḷi language due to the widespread practice of Buddhism and the study of Buddhist scriptures in Myanmar. [2] expressed that the essentially Indian genius, the psychological subtleties, and high thoughts of Buddhism have forced the Burmese language to grow, deepen and expand continually. As a consequence of Pāḷi influence on Myanmar language, usages of Pāḷi and Pāḷi derived words are wide and frequent in Myanmar text. Some Pāḷi words were directly incorporated into Myanmar language. Later era, the re-searchers focus on speech processing. Therefore, this paper presents text to speech system for Myanmar Pāḷi words by using Romanization rules to help the people who are interested in Pāḷi that is writing in roman script. It can also help the students and teachers of oriental studies.

[3] proposed the system to use for geographical names in the democratic people's republic of Korea. Its contains the Romanization of Hangeul. The way to write Thai language using roman alphabets are proposed by [4]. It could be performed on the basi of orthographic form (transliteration) or pronumciation (transcription) or both. Romanization system for Japanese Kana is proposed by [5]. It has been in use by the U.S. Board on Geographic names and the U.K. [6] is the system of the Romanization of Shan. It has been developed for use in Romanizing names written in the Shan script.

The rest of paper is organized as follow: in Section II, Pali language is explained. The nature of Pali word in Myanmar language is presented in Section III. The architecture of the proposed system is explained in Section IV with the step by step explanation. The paper is concluded in Section V.

II. PĀḶI LANGUAGE

Language is the speech, spoken by the people for communication, composed of letters (akkharā) or alphabet. Pāḷi is the language in which is composed the Tipiṭaka. The word Pāḷi is used in the sense of "Text", sacred Text and the same thing for the etymology of Pāḷi is the Holy Text, the Scriptures or the canon. Pāḷi language is a branch of Indo-European fami-y and a sister language of Sanskrit. Pāli was first committed to writing in Srīlaṅkā in the 1st century AD for the Buddhist Canon. It is the spoken language. It has no own script but only sound. So, Pāḷi is transliterated into various local scripts. Pāḷi is an inflectional language (declension, conjugation, assimilation). Pāḷi had contributed mostly to the growth of Myanmar as a national language. Brahmanism, Hinduism, Buddhism, Sanskrit and Pāli have been well introduced to Myanmar from the beginning of our history of 4th century AD. The Myanmar invented Myanmar Script using Pallava Script. Moreover, phonetics and ideas were taken from Pāḷi language and literature. As Theravada Buddhism flourished in Myanmar, Pāḷi became the medium of writing [7].

III. PĀḶI WORD NATURE IN MYANMAR LANGUAGE

Every language has its own alphabet which contains letters of that language. They are called akkharā, lipi, script or writing. The Akkahara (alphabet) means one which is eternam or imperishable, however, they are pronounced or used. Therefore, the original meaning of alphabet is “sound”. As they describe the quality of sound, they are also called syllable (vanna). Ka, kha, ga, gha, na, etc. are characters of alpha-bet, these characters are called script (lipi). Generally, Pāḷi is the name of a language and it has no own script. When its alphabets are written, various scripts are used: in Myanmar the Myanmar, in India Devanāgarī, in Śrilaṅkā the Sinha-lese, in Thailand the Siamese, etc. The speech spoken in a language can be written with symbol that are so called alphabet or script. Languages such as English, German used Roman script in writing. Because of Roman origin, it is called Roman script. Romanization is the representation of a language written in a non-Roman script using the Roman alphabet. In the Myanmar writing system, as example, က်ခ်ဂ်ယ်င် are Myanmar script and the corresponding Ro-man scrips are k, kh, g, gh, ṅ. The Pāḷi alphabet contains 8 vowels, 33 consonants, and 1 nasal sound. All vowels can produce their sounds by themselves. Consonants cannot produce their sounds by themselves. So they are called mutes which produce their respective sounds only in combination with the vowels. Pāḷi language contains 41 letters. The Pāḷi words are pronounced with defined Ṭhān [8]. Ṭhān means the organ of articulation. There are six organ of articulation in Pāḷi. They are –

- Kaṇṭha- Gutturals throat
- Tālu- Palatals, hard plate
- Muddha- Cerebrals, soft plate
- Danta- Dentals, teeth
- Oṭṭha- Labials, lips
- Nāsā- Nasals, nose

The 33 consonant can be classified into six group according to the above six Ṭhān. Therefore, the Pāḷi alphabet for both Myanmar script and Roman script classified with Ṭhān are described in Table I and II. Moreover, it can also be classified with Vagga and Avagga sound as shown in Table III and IV.

TABLE I. Vowel in Myanmar script

Vowels							
အ	အာ	ကြ	ဤ	ဥ	ဦ	ဧ	ြ

TABLE II. Consonant in Myanmar script

Consonants				
က်	ခ်	ဂ်	ဃ်	င်
စ်	ဆ်	ဇ်	စျ်	ည်
ဋ်	ဌ်	ဍ်	ဎ	ဏ်
တ်	ထ်	ဒ်	ဓ်	န်
ပ်	ဖ်	ဗ်	ဘ်	မ်
ယ်	ရ်	လ်	ဝ်	သ်
	ဟ်	ဠ်	အံ()	

TABLE III. Vowel in Roamn script

Ṭhān	Short (Rassa Sara)	Long (Dhīgha Sara)
Gutturals	a	ā
Palatals	i	ī
Labials	u	ū
Gutturals+ Palatals	e	
Gutturals+ Labials		o

TABLE IV. Vowel in Roman script

Ṭhān	Vagga (Classified)					Avagga (Unclassified)
	Unaspi-rate	Aspirate	Unaspi-rate	Aspirate	Nasals	h
	First	Second	Third	Fourth	Fifth	
Kaṇṭha	k	kh	g	gh	ṅ	y
Tālu	c	ch	j	jh	ñ	r, ḷ
Muddha	ṭ	ṭh	ḍ	ḍh	ṇ	l, s
Danta	t	th	d	dh	n	v
Oṭṭha	p	ph	b	bh	m	y
Nāsā					ṃ	

IV. PROPOSED SYSTEM ARCHITECTURE

There are four main modules in the proposed Pāḷi Romanization system: (i) text normalization, (ii) check the input text is Pāḷi word or not, (iii) Romanized the Pāḷi word and finally (iv) generate the speech of the Romanized Pāḷi words. Firstly, the input Myanmar Pāḷi words are preprocessed as syllable segmentation for the next processing. Then, the normalized words are checked which are Pāḷi words or not. If it is Pāḷi words, these are transformed into roman script by using proposed Romanization rules. Finally, the system generated the speech output of the Myanmar Pāḷi words. The system architecture of the proposed system is shown in Fig 1.

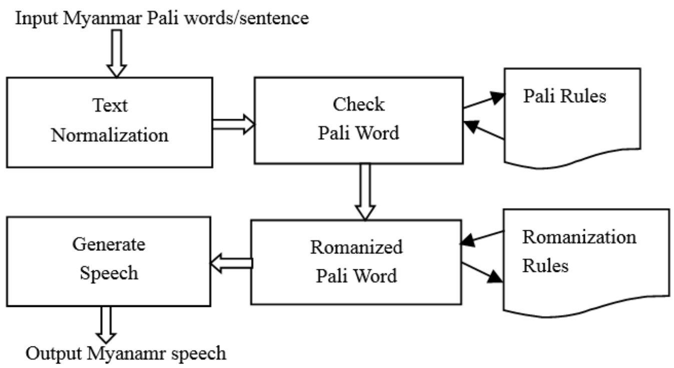


Fig.1. Block diagram of rule based Romanization system

A. Text Normalization

Myanmar language writhing system goes from left to right and there is no space between syllable, word or phrase. Syllable is the combination of consonant phoneme and vowel phoneme [9]. It is the basic unit and syllabication is one of the most essential steps in Myanmar natural language processing. In our system, the input Pāḷi sentences are segmented by rules proposed by [10]. There is no problem for the simple consonant because the Pāḷi words can also be written with the same script for the regular Myanmar texts. Otherwise, the Myanmar Pāḷi words nature has stacked consonants or subscripted consonant form. It means two consonant letters are stacked together and the two consonants are loaded together and the second consonant is subscripted below the first consonant, then deleting the inherent vowel sound of the first one such as ဗုဒ္ဓ (Buddha). In this word, the consonant "ဒ" and "ဓ" is stacked together. According to the Myanmar Unicode format, conjunction of two consonants is specified by the insertion of a virama ("္" - U+1039) between them like (ဒ + ္ + ဓ = ဒ္ဓ). This virama cannot be seen and it can occur before the second consonant to be display in a smaller form under the first one. Firstly, this kind of stacked consonants have to be normalized to become the normal form by using the ALGORITHM I shown in the following Fig 2.

ALGORITHM II : Stacked Consonant Normalization

Begin
Accept the sequence of Myanmar Pāḷi words
If found (consonant + ("္" –shit+f) + consonant) then
Delete "္"
End if
End

Fig.2. The stack consonants normalization algorithm

B. Check Pāḷi Words

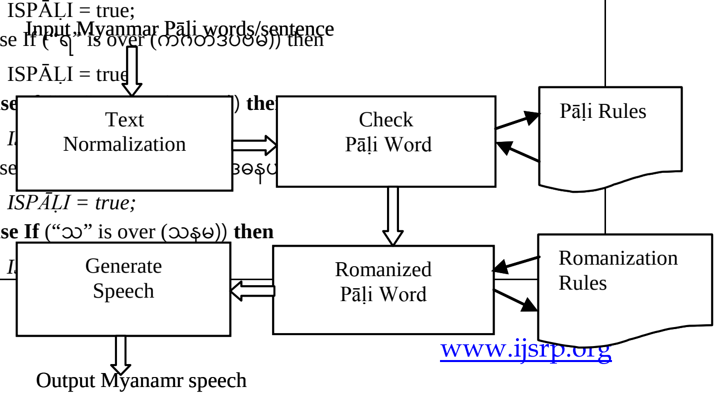
Generally, when someone study the Pāḷi language, they have to be aware that some excluded words in Pāḷi. In general, the syllable with some vowel combination and medials combination cannot allow in Myanmar Pāḷi writing system. For instance - the exclude symbols combined with sample consonant "က" as "ကေး, ကေ့, ကဲ, ကင်, ကက်, ကော်, ကို, ကျွ, ကွ,ကြ, ကျွ, ကျ,ကြွ, ကွ,ကြ ကွှ". Such combinations are excluded words in Myanmar Pāḷi.

According to the Table IV, the 33 consonants are grouped according to Ṭhān. These groups are called vagga. Generally, the consonants that have different Ṭhān cannot be grouped so they are called Avagga. There are 25 Vagga which are also called first, second, third, fourth and fifth group accorded to the vertical value in Table 4. In this case, the first, third and fifth group consonants can stack with each other such as "က္က", "ခ္ခ" and "မ္မ". The last group cannot have this kind of chance. However, the second group can be stacked below the first one. Likewise, the third group can stack over the forth group. Besides, the fifth group can stack over the first four consonants in each group.

In the case of Avagga, except the consonant "ရ", the others can stack each other and the limited and corresponding Avagga can also stack over the Avagga. For the consonant "ယ", other 18 consonants can stack over it and other 9 consonants can do over the "ဝ". "က" and "ပ" can only stack over the consonant "လ". Besides, "န" and "မ" can do over "သ". The last consonant of the Vagga "ဉ, ဏ, န, မ, ယ, ဝ, ဠ" can also stack over the "ဟ". Except the above condition, the other stacked condition cannot allow in Myanmar Pāḷi word. As conclusion, the algorithm for checking for the input word is Pāḷi word or not is shown in the ALGORITHM II.

Algorithm II : CHECK PĀḶI WORD

Input: Myanmar Syllable S = {S1, S2, S3, ….}
{Output: Condition it is Pāḷi word or not, ISPĀḶI = {true, false}
Begin
Procedure check_Pāḷi (MS)
1stGroup = "က,စ,ဋ,တ,ပ";
2ndGroup = "ခ,ဆ,ဌ,ထ,ဖ";
3rdGroup = "ဂ,ဇ,ဍ,ဒ,ဗ";
4thGroup = "ဃ,ဈ,ဎ,ဓ,ဘ";
5thGroup = "ဏ,န,မ";
AwatGroup = "ယ,လ,ဝ,သ,ဟ,ဠ,အ";
excludeWord= "း, ့, ဲ, ် ,ော်, ို, ု,ျ,ွ,ြ,ှ"
If (1stGroup is over 1stGroup || 3rdGroup is over 3rdGroup || 5thGroup is over 5thGroup) **then**
ISPĀḶI = true;
Else if (1stGroup is over 2ndGroup || 3rdGroup is over 4thGroup) **then**
ISPĀḶI = true;
Else If ("ယ" is over (ယကခဂဃစဍဏတဒဓနပဗဘမလသဟ)) then
ISPĀḶI = true;
Else If ("ရ" is over (ကဂတဒပဗမ)) then
ISPĀḶI = true
Else …) **the**…
I…
Else …
ISPĀḶI = true;
Else If ("သ" is over (သ္နမ)) **then**
I…



```
22.  Else If ("ဟ" is over (ဥဏနမယဝဠ)) then
23.      ISPĀḶI = true;
24.  Else If (current syllable does not contain (excludeWord)) then
25.      ISPĀḶI=true;
26.  Else
27.      ISPĀḶI=false;
28.  End if
29.  End procedure
End
```

Fig.3. The algorithm for checking Myanmar Pāḷi words

C. *Romanized Pāḷi Words*

In linguistics, Romanizaion is the conversion of writing from a different writing system to the Roman (Latin) script, or a system for doing so. Methods of Romanization include transliteration, for representing written text, and transcription, for representing the spoken word, and combinations of both. Transcription methods can be subdivided into phonemic transcription, which records the phonemes or units of semantic meaning in speech, and stricter phonetic transcription, which records speech sounds with precision [11].

Myanmar language is one of few alphabets capable of transcribing Pāḷi text with 100% orthographic fidelity. However, because Pāḷi is no longer a spoken, but a written language, the standard pronunciation of Pāḷi text occurs in agreement with the phonetic values and inherent rules of the corresponding alphabets used. Accordingly, like other language such as Thai, Sinhala, Lao and Khmer, speakers, the Myanmar have a very distinct accent when using Pāḷi words.

After checking the input word is a Pāḷi or non-Pāḷi word, if the input word is Pāḷi, the system converted into their corresponding Romanized symbol. Therefore, we created the Romanized table as shown in Table V. Then, the input Pāḷi words are transformed into Roman script by using the following 10 Romanization rules with example words.

TABLE V. Romanization table

Consonant	Unicode	Roman_Symbol
က	U+1000	k
ခ	U+1001	kh
ဂ	U+1002	g
…		
ဥ	U+102f	u
ဦ	U+1030	ū
ဧ	U+1031	e

Rule 1. Combination of Consonant with Vowels

k+ a= ka (က), k+ i= ki (ကိ), k+ u= ku (ကု), k+ e= ke (ကေ)

k+ ā= kā (ကာ), k+ ī= kī (ကီ), k+ ū= kū (ကူ), k+ o= ko (ကော)

Rule 2. Combination of the vowels with the niggahita

a+ ṃ= aṃ (အံ), i+ ṃ= iṃ (အိံ), u+ ṃ= uṃ (အုံ)

ā+ ṃ= aṃ (အံ), ī+ ṃ= iṃ (အိံ), ū+ ṃ= uṃ (အုံ)

Rule 3. Combination of the consonants, vowels with the niggahita

k+ a+ ṃ= kaṃ (ကံ), k+ i+ ṃ= kiṃ (ကိံ), k+ u+ ṃ= kuṃ (ကုံ)

k+ ā+ ṃ= kaṃ (ကံ), k+ ī+ ṃ= kiṃ (ကိံ), k+ ū+ ṃ= kuṃ (ကုံ)

Rule 4. First Group alphabet+ First Group alphabet

k+ k= kk sakka (သက္က)

c+ c= cc sacca (သစ္စ)

ṭ+ ṭ= ṭṭ vaṭṭa (ဝဋ္ဋ)

t+ t= tt satta (သတ္တ)

p+ p= pp sappa (သပ္ပ)

Rule 5. First Group alphabet+ Second Group alphabet

k+ kh= kkh yakkha (ယက္ခ)

c+ ch= cch accha (အစ္ဆ)

t+ th= tth sattha (သတ္ထ)

p+ ph= pph puppha (ပုပ္ဖ)

Rule 6. Third group alphabet + Third group alphabet

g+ g= gg agga (အဂ္ဂ)

j+ j= jj ajja (အဇ္ဇ)

d+ d= dd bhadda (ဘဒ္ဒ)

b+ b= bb sabba (သဗ္ဗ)

Rule 7. Third group alphabet+ Fourth group alphabet

g+ gh= ggh byaggha (ဗျဂ္ဃ)

j+ jh= jjh majjha (မဇ္ဈ)

d+ dh= ddh buddha (ဗုဒ္ဓ)

b+ bh= bbh labbhati (လဗ္ဘတိ)

8. Fifth group alphabet+ Consonant of same Ṭhān

ṅ+ g= ṅg maṅgala (မင်္ဂလ)

ñ+ c= ñc pañca (ပဉ္စ)

ṇ+ ḍ= ṇḍ kaṇḍa (ကဏ္ဍ)

n+ t= nt ananta (အနန္တ)

m+ bh= mbh sambhūla (သမ္ဘူလ)

9. Combination of Avagga letters

y+ y= yy seyya (သယ္ယ)

y+ v= yv	yvāhaṃ (ယွာဟံ)
y+ h= yh	paggayha (ပဂ္ဂယှ)
v+ h= vh	avhayati (အဝှါ ယတိ)
l+ l= ll	vallari (ဝလ္လရိ)
l+ y= ly	kalyā (ကလျာ)
s+ y= sy	nisya (နိသျ)
s+ v= sv	svāhaṃ (သွာဟံ)

10. Combination of the Vagga with Avagga

k+ y= ky	sakyamuni (သကျမုနိ)
k+ r= kr	cakra (စက္ကြ)
d+ v= dv	dvāra (ဒွါရ)
n+ h= nh	nhāna (ဏှာန)

The example Romanized Pāḷi sentence is shown in Table by using the mentioned Romanization rules.

TABLE VI. Sample Romanization result

Input Pāḷi Sentence	နမော တဿ ဘဂဝတော အရဟတော သမ္မာ သမ္ဗုဒ္ဓဿ
Romanized Sentecne	Namo tassa bhagavato arahato saṃmā ssaṃbuddhassa

D. Generated Pāḷi Speech

In the speech generation step, the converted roman scripts are transformed again into speech output. MaryTTS speech engine is used in this potion. MaryTTS is a multilingual Text-to-Speech Synthesis platform written in Java and it is an open-source platform. It was originally developed as a collaborative project of DFKI's Language Technology Lab and the Institute of Phonetics at Saarland University. Now, Multimodal Speech Processing Group in the Cluster of Excellence MMCI and DFK maintain the MaryTTS. As of version 5.2, MaryTTS supports German, British and American English, French, Italian, Luxembourgish, Russian, Swedish, Telugu, and Turkish; more languages are in preparation.

MaryTTS comes with toolkits for quickly adding support for new languages and for building unit selection and HMM-based synthesis voices [12].

V. Experimental Result

Generally, the performance of Myanmar Pāḷi word Romanization can be calculated in different ways. In this system, the goodness of transformation is measured by four types of outcomes: (1) Correct Transformed (CT): A Pāḷi word was converted correctly and is detected to be correct; (2) Correct Rejection (CR): A Pāḷi word was transformed incorrectly and is detected to be incorrect; (3) False Transformed (FA): A Pāḷi word was converted incorrectly and is detected to be correct; (4) False Rejection (FR): A Pāḷi word was transformed correctly and is detected to be incorrect. For this four measure, 500 Pāḷi sentences are tested. The experimental results are shown in Fig 4. Typically, the performance of an error detection algorithm can be calculated in different ways. One way is to measure the scoring accuracy (SA), which is calculated by formula shown below:

$$SA = ((CT+CR) / (CT + CR + FT + FR)) * 100;$$

The ratio of CTs and CRs can be calculated by the classification algorithm: precision, recall, and F-measure metrics. These measurements are as follows:

$$\text{Precision of CT} = (CT / (CT + FA)) * 100;$$

$$\text{Precision of CR} = (CR / (CR + FR)) * 100;$$

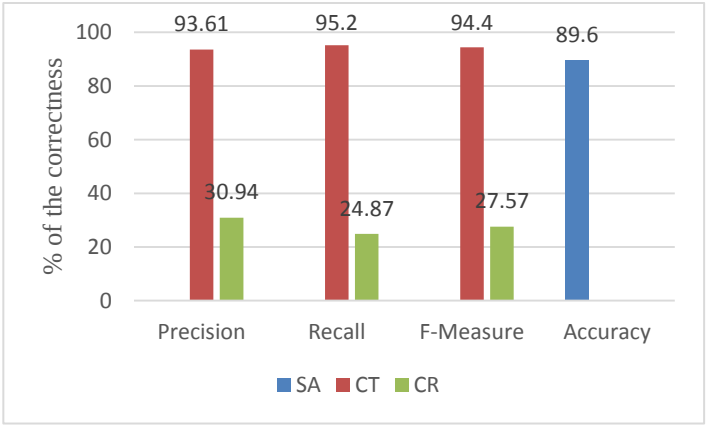
$$\text{Recall of CT} = (CT / (CT + FR)) * 100;$$

$$\text{Recall of CR} = (CR / (CR + FA)) * 100;$$

$$\text{F-measure} = \frac{2*(Precision*Recall)}{(Precision+Recall}$$

VI. CONCLUSION

This paper presented the rule based Myanmar Pāḷi word Romanization system. Therefore, the ten Romanization rules are discussed. Before the Pāḷi words are Romanized, the input words are checked that they are Pāḷi word or not. Consequently, the Myanmar Pāḷi word checking algorithm is presented. This paper is mainly focus on Myanmar Pāḷi word checking and Romanized these Pāḷi words so that the MaryTTS engine is used for speech generation. According to the experimental result for Romanization, the system achieved the overall accuracy is 89.6. For some words, such as "အောင်မင်္ဂလာ". In this word, although "မင်္ဂလာ" is Pāḷi word, the syllalbe "အောင်" is not Pāḷi word. In this case, the system may wrongly check as the Pāḷi words. In this case, the accuracy may be decreased. Nowadays, in Myanmar, the researchers focus on the speech processing in Myanmar natural language processing. In the future, the high quality speech output will be generated by using other speech synthesis methods like concatenation speech synthesis approach by recording own voice for Pāḷi words.



http://dx.doi.org/10.29322/IJSRP.8.9.2018.p8158

Fig.4. The experimental result of the Myanmar Pāḷi word Romanization

REFERENCES

[1] C. Duroiselle, “Practical Grammar of Pali Language”, Third Edition 1997

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

[3] Ryang, Sonia, ed. Koreans in Japan: Critical voices from the margin. Routledge, 2013.

[4] Aroonmanakun, Wirote. "A unified model of Thai romanization and word segmentation." In Proceedings of The 18th Pacific Asia Conference on Language, Information and Computation, pp. 205-214. 2004.

[5] Kudo, Yoko. "Modified Hepburn Romanization System in Japanese Language Cataloging: Where to Look, What to Follow." Cataloging & Classification Quarterly 49, no. 2 (2011): 97-120.

[6] “Romanization System for Shan”, BGN/PCGN 2011 System.

[7] “Fundamental of the Pāḷi Language”, University of Mandalay, Department of Oriental Studies.

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

[9] C.S. Hlaing and A. Thida. "Phoneme based Myanmar text to speech system." International Journal of Advanced Computer Research 8, no. 34 (2018): 47-58.

[10] Maung, Zin Maung, and Yoshiki Mikami. "A rule-based syllable segmentation of Myanmar text." In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages. 2008

[11] https://en.wikipedia.org/wiki/Romanization

[12] http://mary.dfki.de/index.html