Automatic Syllable Segmentation of Myanmar Texts using Finite State Transducer

3/4/2013

ICTer Submission

Tin Htay Hlaing and Yoshiki MIKAMI,

Nagaoka University of Technology, JAPAN.

Automatic Syllable Segmentation of Myanmar Texts using Finite State Transducer

Tin Htay Hlaing
tinhtayhlaing@gmail.com

Yoshiki MIKAMI
mikami@kjs.nagaokaut.ac.jp

Nagaoka University of Technology, JAPAN

***Abstract* — Automatic syllabification lies at the heart of script processing especially for the South East Asian scripts like Myanmar. Myanmar syllabification algorithms implemented so far are either rule-based or dictionary-based approach. This paper proposes a new method for Myanmar syllabification which deploys formal grammar and un-weighted finite state transducers (FST). Our proposed method focuses on orthographic way of syllabification for the input texts encoded in Unicode. We tackle syllabification of Myanmar words with standard syllable structure as well as words with *irregular* structures such as kinzi, consonant stacking which have not been resolved by previous methods. Our FST based syllabifier was tested on 11,732 distinct words contained in Myanmar Orthography Corpus. These words yielded 32,238 syllables and are compared with correctly hand syllabified words. Our FST based syllabification method performs with 99.93% accuracy on Stuttgart FST (SFST) tools.**

***Index Terms*— Automatic Syllabification, Finite State Transducer, Myanmar syllabification, formal description of syllable structure**

I. INTRODUCTION

Syllabication is the task of breaking words into syllables. This is called automatic syllabication when performed using computer algorithms instead of linguists. Knowledge of the syllable boundaries in words is very useful in a number of areas because it is impossible to store syllable information for all words in a language (new vocabulary are continually being added), and thus automatic syllabication algorithms are necessary [3].

Languages differ considerably in the syllable structures that they permit. For most languages, syllabification can be achieved either by writing a set of rules which explain the location of syllable boundaries of words step-by-step or using annotated corpus. Syllabification algorithms have been proposed for different languages by using different approaches. Up to our knowledge, rule-based approaches have been used for Asian scripts, for example, Lao [16] and Sinhala [2]. And also, corpus-based syllabification approach is done for Uyghur [12] and Urdu [1]. Moreover, many language-specific syllabification methods have been modeled by using finite state machines or neural networks and Finite State Transducers for multilingula syllabifiation [9].

These algorithms are mainly used in text-to-speech (TTS) systems in producing natural sounding speech, and in speech recognizers for dealing with out-of-vocabulary words. Also for Myanmar script, syllabification algorithm for Myanmar Text-to-Speech (TTS) system has been developed [20]. However, such phonetic syllabification cannot be used for some major Myanmar language processing tasks such as lexicographic sorting, word breaking, line breaking and spelling checking. In other words, it is necessary to use orthographic syllabification for these tasks.

Although a few researchers have documented attempts at syllabifying Myanmar words orthographically, this is the first known documented method for Myanmar orthographic syllabification using finite state transducers (FST). The objectives of this study are to represent Myanmar syllable structure in formal description (e.g, regular grammar) by taking advantage of its structuredness and unambiguity. Another objective of our study is to achieve correct syllabification without applying step-by-step rules and the need of corpus. In other words, our proposed method is neither heuristic approach nor annotated corpus-based approach.

Thus, in this research, we explain Myanmar syllabification in both orthographic and phonetic views. Un-weighted finite state transducer to divide Myanmar words into syllables is proposed. Syllable struture model is represented in Chomsky`s regular grammar and deploy finite state transducers for automatic syllabification of Myanmar Unicode texts. Our transducer accepts input string (also known as surface string) in Unicode encoding and in generation mode, the transducer produces syllabified texts with boundary marker notation.

The method was tested by using a text corpus containing 11,732 distinct words yielding 32,238 syllables on Stuttgart Finite State Transducer Tool (SFST). And its performance was then meaured in terms of the percentage of correctly syllabified words. Having limited resources for Myanmar language processing, the syllabified results are checked manually. Our result reports 99.93% of accuracy for the test set of 32,238 syllables. The rest of this paper is organized as follows: section 2 introduces syllable segmenation of Myanmar script by highlighting its challenges. We cover related works in section 3 and then discuss overview of Myanmar syllable strucutre in both phonetic and orthographic ways and Unicode cannonical order are in section 4. Section 5 describes our proposed Finte State Transducer approach, and our experiments and results are in section 6.

II. CHALLENGES IN MYANMAR SYLLABIFICATION

Myanmar language, formerly known as Burmese, is an official language of Myanmar and a member of Burmese-Lolo group of the Sino-Tibetan language spoken by about 32 million as a first language and as a second language by 10 million. The Burmese script, attested in stone inscriptions at least as far back as the early twelfth century C.E., is a phonologically based script, adapted from Mon and ultimately based on an Indian (Brahmi) prototype [15].

In Myanmar language, syllable is the smallest linguistic unit and one word consists of one or more syllables. Generally, Myanmar words can be classified into (1) Standard words (i.e, words with standard syllable structure) and (2) irregular words (i.e, words with abbreviated characters or words written in special traditional writing forms which is discussed in detailed in section 4.3). And each word type needs different orthographic ways of syllabifications as follows.

Word Type	Example Word	Meaning	Syllabified Output
Standard Word	သမီး	Daughter	သ # မီး
Irregular Word	တက္ကသိုလ်	University	တက်# က# သိုလ်

In this example, the word “Daughter, သမီး” has two syllables, သ and မီး where the former syllable has only one sub-syllabic element, consonant သ but the latter has three sub-syllabic elements of consonant မ, vowel ◌ီ and diacritic ◌း. For instance, syllables in alphabetic script are composed of only two alphabets, consonants and vowels whereas in Myanmar script, each syllable is composed of at most five sub-syllabic groups and each group has its own members. We will address about this in later section.

For the word “University, တက္ကသိုလ်” has three syllables and it is written in one of the special traditional writing formats known as consonant stacking. It is syllabified by extra insertion of Myanmar Sign ASAT, U+103A between two stacked consonants to get the syllable boundary. For instance, the first character တ is combined with the upper character from the stack and form one syllable and the lower character in the stack itself becomes a syllable. Thus such kind of language-specific features make Myanmar syllable segmentation task complicated. Myanmar language has five different forms of *irregular* words which are explained in section 4.3.

Besides, many Pali words and English loan words that refer to people, places, abbreviation of foreign words, currency units etc., can be found in Myanmar texts and we need to tackle segmentation of such words.

In handling such *irregular* words under-resourced language, Myanmar, rule-based approach and corpus-based approach have shown some failures. Nevertheless, in our FST-based approach, both standard and *irregular* words are correctly syllabified.

III. RELATED WORK

Scripts can be classified into five categories namely: alphabet, abjad, abugida, syllabary, and logosyllabary. There is no need to explain the first category, alphabet. The second and third categories do not sound familiar to us, but abjad corresponds to Semitic scripts, such as Arabic script and Hebraic script, while abugida corresponds to Ethiopian script and Indian-based script. The third category, abugida, “consists of consonant letters with specific vowels attached to them, and expresses other vowels (or a syllable with vowels) by modifying consonant letters in a consistent manner.

The fourth category, syllabary, can be expressed as “letters with no graphical relationship or rules between letters with similar sounds.” A typical example is Kana in Japanese. The fifth category is logosyllabary. Of course, a typical example is Chinese characters. Chinese characters are so classified because they are both logography and syllabary at the same time [22].

syllable segmentation of most Alphabetic and Arabic scripts is done phonetically, i.e, the input surface string is first converted into phonetic symbols and then syllabify by using suitable approach such as rule-based or statistical approach. For the Indian-based script Abugida, syllabification can be done either phonetically or orthographically.

There are many approaches tackling the syllable segmentation task. Generally, these can be divided into two broad categories namely rule-based and data-driven approaches. The rule-based method effectively embodies some theoretical position regarding the syllable, whereas the data-driven paradigm infers “new” syllabifications from examples assumed to be correctly-syllabified already. However, it is difficult to determine correct syllabification in all cases and so to establish the quality of the “gold-standard” corpus used either to quantitatively evaluated the output of an algorithm or as the example-set on which data-driven methods crucially depend [11].

Besides these two categories, statistical methods and finite state methods are also applied for automatic syllabification and we will introduce previous approaches briefly.

In syllabification of written Uyghur [12], a rule-based approach that uses the Principle of Maximum Onset is applied. Experiment on a random sample shows that the syllabification algorithm achieves 98.7 percent word accuracy on word tokens, 99.2 percent on word types, and 99.1 percent syllable accuracy. In [17], the authors described rule based syllabification algorithm for Sinhala after analyzing the syllable structure and linguistic rules for syllabification of Sinhala words. Rule-based syllabification algorithm for Malay is proposed based on Maximum Onset principle for text to speech system in [21].

In [11], one rule-based approach, and three data-driven approaches are evaluated (A Look-up Procedure, an Exemplar-based generalization technique and the syllabification by Analogy (SbA)). The results on the three databases show consistent and robust patterns: the data-driven techniques outperform the rule-based system in word and juncture accuracies by a very significant margin and best results are obtained with SbA.

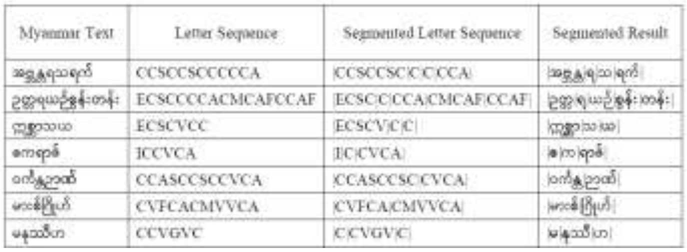
In [9], a weighted finite-state-based approach to syllabification is presented. Their language-independent method builds an automaton for each of onsets, nuclei, and codas, by counting occurrences in training data. These automatons are then composed into a transducer accepting sequences of one or more syllables. They do not report quantitative results for their method. Syllabification of Middle Dutch texts is done by the method which combines a rule-based finite-state component and data-driven error-correction rules. The authors adapt an existing method for hyphenating (Modern) Dutch words by modifying the definition of nucleus and onset, and by adding a number of rules for dealing with spelling variation [4].

Regarding Myanmar syllabification, two of the above mentioned approaches have already been done. In the corpus-based longest matching approach [6], the authors collected 4,550 syllables from different resources. The input texts are syllabified by using longest matching algorithm over their syllable list. They observed that only 0.04% of the actual syllables were not detected and described their failures as three facts:

- differing combinations of writing sequences
- loan words borrowed from foreign languages
- rarely used syllables not listed in their syllable list

Rule-based Myanmar syllable segmentation is done by [23] in which input text strings are converted into equivalent sequence of category form (e.g. CMCACV for the word "Myanmar") and compares the converted character sequence with the syllable rule table to determine syllable boundaries. The authors tested 32,238 syllables in the Myanmar Orthography [14] and the experimental results show an accuracy rate of 99.96% for segmentation. However, their approach cannot solve for the segmentation of *irregular* words with traditional writing forms namely kinzi, consonant stacking, great SA and English loan words with *irregular* forms as shown in table 1. However, in our approach, these kinds of failures in [6] and [23] are addressed the syllabification of *irregular* words and managed correctly.

Table 1. Syllable Segmentation Examples and Results[1]



Myanmar Text	Letter Sequence	Segmented Letter Sequence	Segmented Result
အစ္ဆန္တရသရက်	CCSCCSCCCCCA	|CCSCCSC|C|C|CCA|	|အစ္ဆန္တ|ရ|သ|ရက်|
ဥဒ္ဒရယဉ်စွန်းတန်း	ECSCCCCACMCAFCCAF	|ECSC|C|CCA|CMCAF|CCAF|	|ဥဒ္ဒ|ရ|ယဉ်|စွန်း|တန်း|
ဣန္ဒာသယ	ECSCVCC	|ECSCV|C|C|	|ဣန္ဒာ|သ|ယ|
ဧကရာဇ်	ICCVCA	|I|C|CVCA|	|ဧ|က|ရာဇ်|
ဝင်္ကန္တဒါတ်	CCASCCSCCVCA	|CCASCCSC|CVCA|	|ဝင်္ကန္တ|ဒါတ်|
[illegible]	CVFCACMVVCA	|CVFCA|CMVVCA|	[illegible]
မနုဿီဟ	CCVGVC	|C|CVGV|C|	|မ|နုဿီ|ဟ|

[1] This table is directly taken from the original manuscript of [23].

In [13] Unicode Technical Note, diacritic storage order of Myanmar characters in Unicode (which we refer as sub-syllabic components) is explained in detail and it is highlighted that diacritic storage order does not define a phonetic syllable. Further, automated syllable breaking approach for Myanmar script is mentioned. It is said that the syllable break may occur before any character cluster so long as the kinzi, asat and stacked slots remain empty in the cluster following the possible break point. It is mentioned that their algorithm does not require dictionary but still needs more refinement, for example, sequence of digits should be kept together and visible virama needs more complex analysis. The author also stated that the result of syllable breaking can be applied for line breaking and the same syllable breaking rules can be applied for lexicographic sorting.

Finite state methods have been applied in syllabification of alphabetic scripts but it has not yet been done in Myanmar script. Therefore, in our study, we describe syllable structure model in regular grammar and write regular expressions which are used as input to Finite State Transducer. In our experiment, we use 32,238 syllables covering all possible syllable structure in standard words and *irregular* words from Myanmar Orthography published by Myanmar Language Authority [14] and achieved the accuracy of 99.93%. We implemented the syllabification system by using the programming language for finite state transducer [18].

IV. MYANMAR SYLLABLE STRUCTURE

According to the Unicode standard version 6.2, Myanmar characters range from U+1000 to U+109F. Basically, Myanmar script, formerly known as Burmese, has 33 consonants, 8 vowels (free standing and attached), 2 tone marks, 11 medial consonants, a vowel killer or ASAT, 10 digits, 5 abbreviated syllables and 2 punctuation marks. Among them, consonants, medial consonants and attached vowels, a vowel killer or ASAT and tone marks can be combined in different ways to form a syllable and thus we refer these groups as sub-syllabic elements. Others, free standing vowels, digits and abbreviated syllables can be syllables by themselves.

Further, Myanmar syllable structure can be represented in two ways namely phonetic and orthographic representations which are briefly explained in following section.

4.1 Myanmar Syllable Structure in Phonetic Representation

Myanmar script is known as diacritically modified consonant syllabic scripts or alphasyllabaries as it is derived from Brahmi and it inherited systemic features of Brahmi. The Brahmi system based on the unit of the graphic "syllable" or aksara, which by definition always ends with a vowel (type V, CV, CVV, etc). Further, the basic consonantal character without any diacritic modification is understood to automatically denote the consonant with the inherent vowel [15].

A Myanmar sentence is composed of words, and words written in Myanmar script are made of a series of distinct single syllables. In phonological representation, each syllable is made up of two parts:

(a) A consonant or sometimes two consonant together and
(b) A vowel or a vowel and a final consonant together.

The first part of the syllable is known as *head* and the second part the *rhyme*. The rhyme may also contain tone.

As an example, here is the same principle applied to some English words:

(a) head [consonants or two consonants]	+	(b) rhyme [vowel or vowel and consonant]	=	Syllable
T	+	EE	=	TEE
T	+	ICK	=	TICK
TR	+	ICK	=	TRICK
TR	+	EE	=	TREE

In Burmese script the head of a syllable may be either

(1) an "initial consonant"; for example, the consonants
written: ပ လ န သ
pronounced: p- l- n- thor
or
(2) an initial consonant combined with a second consonant, referred to below as a "medial consonant"; e.g.
written: ပြ လျ နှ သွ
pronounced: py- ly- hn- thw-

There are only four medial consonants in Burmese script.

The rhyme of a syllable may be written with either

(1) an attached vowel symbol; e.g.
written: ပီ လူ နာ သို
pronounced: pi lu na tho
or
(2) a consonant marked as a final consonant by carrying the "killer" symbol ် ; e.g.
written: ပန် လန် နတ် သက်
pronounced: pan lan naq theq
or
(3) a combination of an attached vowel symbol and a final consonant;e.g.
written: ပုန် လိန် နောင် သိုက်
pronounced: poun lein naun thaiq

In addition, tones are part of the rhyme and are mostly represented by the two tone marks ့ and း; e.g.
written: ပုန့် လိန်း နား သို့
pronounced: pou´n leı´n na´ tho´
Other ways of representing tone are used for certain rhymes.

It is also good to note that vowels in the Myanmar script are always attached to their heads (consonants) they are not letters in their own right, like the *a*, *e*, *i*, *o*, *u* of the roman alphabet, and they are not normally written independently. However, a Myanmar letter A "အ" (U+1021) is used to write syllables that have no initial consonant, such as
"i" written အိ, "an" written အန်, "oun" written အုန်

The Myanmar letter A "အ" occupies the position of the initial consonant in the written syllable, but is read aloud as "no initial consonant" [8].

From the view of phonology, Myanmar syllable has the phonemic shape of

C(M)V(CK)D | V(CK)D

where C refers an initial consonant, M for medial consonant, V for attached vowels, CK stands for a final consonant (a combination of consonant C and vowel killer K), and D for the tones respectively. The symbol ( ) means optional.

Therefore, minimal syllable in phonetic view is CVD or VD.

4.2 Myanmar Syllable Structure in Orthographic Representation

As described in previous section, Myanmar script is derived from Brahmi script of ancient India and there are other Indian-based scripts such as Sinhala, Bengali, Dzongka, Thai, Khmer and so on. Myanmar script is based on "a-vowel accompanying consonant syllabics", i.e, this syllabary consists of consonant letters accompanied by an inherent vowel. Since, in this syllabary, a consonant associated with its inherent vowel can indicate a standalone syllable, it would be appropriate to call it a consonant syllable, but we simply call it a consonant letter for simplicity [15]. Thus, only a consonant can be syllable breaking point in orthographic syllabification which means minimal syllable in orthographic view is C which stands for consonant.

Basically, a Myanmar syllable can be described as

S = I | N | P | X where
S = Syllable
| = logical operator OR
I = Free-standing vowel syllables
N = digits
P = Abbreviated syllables
X = a syllable formed by the combination of up to 5 sub-syllabic groups

An additional complexity to Myanmar syllable structure is that there are syllables (X) containing up to 5 sub-syllabic groups namely consonant (C), medial consonant (M), dependent vowel (V), *Asat* or a vowel killer (K) and tones (D) and these groups can appear in a syllable as one of the following combinations.

Table 2. Possible Combinations within a Syllable (X)

Consonant only	Consonant followed by Vowel	Consonant followed by Consonant	Consonant followed by Medial
C	CV	CCK	CM
	CVCK	CCKD	CMV
	CVD		CMVD
	CVCKD		CMVCK
			CMVCKD
			CMCK
			CMCKD

These combinations can be described as a regular expression as

X = C M? V? (C K)? D?

where the symbol " ? " stands for 0 or 1 occurrence of the character. In this expression, the combination of consonant and *Asat* or vowel killer (CK) is called final consonant and the syllables end with this combination are known as closed syllables.

Further, as with the multiple-component vowels, the user reads the entire syllable as an entity. In Myanmar script, 3 vowels and 7 medials which are formed by the combination of other vowels or medials defined in the Unicode character chart. And, if we use this characteristics of vowels and medials in writing regular expression, the expression for syllable structure becomes like this X= C M* V* (CK)? D? where the notation ? means 0 or 1 occurrence and * means 0 or more occurrence[19].

4.3 Myanmar Words in Irregular Forms

There are also traditional as well as particular writing forms for some Myanmar words which are commonly used in the literature. The Burmese script is adapted from Mon, and ultimately based on Indian (Brahmi) prototype. Burmese scribes have the convention of preserving the original spelling of (mostly) Indian loan words and also follow the Indian practice of stacking geminate and homorganic consonants [15]. Therefore, we generally named the group of words with the above mentioned particular forms and English loan words as *Irregular* throughout the paper for readablilty and simplicity.

The details of each irregular form and the correct syllabification for these forms are discussed as follows.

1) **Kinzi.** As mentioned in the previous section, final consonant (CK) can follow the main consonant and the resulting syllable is called closed syllable (CCK). In a few words (many of them loanwords), the combination of consonant Nga "င" and vowel killer "◌်" is " င်" which is not written on the line as the usual way, but is placed above the first consonant of the next syllable. The name of the reduced form is called kinzi and meaning "forehead rider" for obvious reasons [7]. Some examples are shown in the table below.

Table 3. Syllabification of Kinzi

Words with reduced form of "င် "	Correct orthographic syllabification	Meaning
အင်္ဂလန်	အင်#ဂ#လန်	England
သင်္ဘော	သင်#ဘော	Ship
အင်္ဂါနေ့	အင် #ဂါ # နေ့	Tuesday

(Note: The symbol # is used to show the syllable boundary throughout this paper)

From the phonetic view, sometimes, the kinzi is pronounced with high tone, for example, the word သင်္ဘော (ship) is pronounced as သင်း#ဘော by adding diacritic mark "◌း" but in sorting order, it is sorted as without diacritic mark. The spelling of the words gives no indication that they are pronounced with a high tone on the kinzi.

2) **Consonant Stacking.** There are some final consonants which are not written in the reduced form like kinzi. With all final consonants other than kinzi (င်), instead of the final consonant being forced up and over the next consonant, it is the next consonant that is forced down and under the final consonant. So, if we have a pair of syllables like သန်and တာ and they appear in a word that requires them to be compressed, then force the main consonant of second syllable တ under the န်, and write as သန္တာ. The same consonants can be stacked and it is known as **consonant repetition**.

3) **Loan Words.** Loan words mean the words which adopt the English pronunciation directly and sometimes adding Myanmar pronunciation together with English pronunciation. For example, English word "car" is written as "ဘတ်စ်ကား" where the first syllable "ဘတ်စ် " is English pronunciation and the second syllable "ကား" is Myanmar pronunciation.

4) **Great SA.** Myanmar consonant Great SA "ဿ " is commonly used and the words with great SA is syllabified as သ+ ◌်+ သ . For example, the word "problem" in Myanmar language is ပြ ဿ နာ which is syllabified as three syllables ပြ သ် #သ# နာ .

5) **Contraction.** There are usages of double-acting consonants in Myanmar and as the name give which acts both the final consonant of one syllable and the initial consonant of the following syllable. For example, the word ယောက်ျား which means "man" is syllabified as ယောက်#ကျား. There are also some words which can be written in both in standard syllable structure and contracted form, for example, the word "daughter" is written as သမီး and သ္မီး.

4.4 Myanmar Canonical Ordering

It is possible for a Myanmar syllable to have a number of sub-syllabic elements (also known as diacritics in [13]) surrounding a base consonant, independent vowel or digit. Since all these diacritics are not spacing, it is important that there is consistent way of storing strings so that applications can work consistently. Further, our proposed method is for orthographic syllabification and thus canonical order of the script should be taken into account.

Myanmar canonical order in Unicode encoding is described in [13]. There is a general order of: initial consonant, medial consonant, vowel, final and tone. The following is the example of how Myanmar strings are stored according to Unicode canonical order and their syllable boundaries.

Before syllabification :

Input String	Stored order sequence
အိပ်ခန်းတံခါးကို (To the bed room door)	1021 102D 1015 103A 1001 1014 103A 1038 1010 1036 1001 102B 1038 1000 102D 102F

After syllabification :

Output String	Stored order sequence with logical syllable break points
အိပ်#ခန်း#တံ#ခါး# ကို	1021 102D 1015 103A # 1001 1014 103A 1038 # 1010 1036 # 1001 102B 1038 # 1000 102D 102F

V. OUR APPROACH

Automatic syllabification of word is challenging, not least because the syllable is not easy to define precisely. Consequently, no accepted standard algorithm for automatic syllabification exists [10]. However, syllabification can be achieved by writing a declarative grammar of possible locations of syllable boundaries in polysyllabic words [8]. On the other hand, Finite state machines are widely used in the field of natural language processing. Finite state transducers (FSTs) have been used ubiquitously in the domain of phonology as well as in morphology for all sorts of string mapping between string descriptions. Again, Finite state automata and transducers have been used in natural language processing of Asian languages for example, morphological analysis of Urdu [5] and likewise, formal grammar is used to express Sinhala computational grammar [2].

In our approach, first we write syllable structure in regular grammar and then converted it into regular expression as FST program is a set of regular expression. We adopt two-level morphology approach in which the surface string, for example, “boys” is analyzed and produced together with lexical information as “boy<Noun><Plural>”. In our case, our syllabification transducer accepts input string/surface string and outputs the string together with syllable boundary information.

Grammar rules for independent vowels, digits and abbreviated syllable can be described as

S → ဣ |ဤ | ဥ | ဦ |ဧ | ဩ | ဪ
S → ၀ | ၁ | ၂ | ၃ | ၄ | ၅ | ၆ | ၇ | ၈ | ၉
S → ၌|၍ |၎|၏

and the syllable with five sub-syllabic elements can be written as

S→ က X |…..|အ X	# 33 rules for all consonants
X→ ◌ျ A |…… | ◌ြွှ A	# 11 rules for all medials
X→ ◌ာ B |…… |◌ို B	# 12 rules for all vowels
X→ က T |….. |အ T	# 33 rules for ending consonants (consonant + vowel killer)
X→ ε	
A → ◌ာ B	
A→ က T |….. |အ T	# 33 rules for all consonants
A→ε	
B → က T |….. |အ T	# 33 rules for all consonants
B → ◌့| ◌း	# 2 ruels for tones
B→ε	
T→ ◌် D	
D → ◌့| ◌း	# 2 ruels for tones
D→ε	

Thus, we developed the transducer which accepts input Unicode strings and then output the strings with correct syllable boundaries.

Firstly, based on the regular grammar as mentioned above, we write the regular expression to recognize Myanmar syllables and then construct the orthographic automata for a Myanmar syllable A_{mm} as

$$A_{mm} = C\ Opt(M)\ Opt(V)\ Opt(CK)\ Opt(D)\ |\ P\ |\ I\ |\ N$$

where Opt is the just acronym for “optional”.

The above automaton accepts one syllable at a time and it can check the combinations of sub-syllabic elements orthographically.

Then, we construct the syllabification automaton, denoted by A_{syl} , which accepts a sequence of syllables and finds the syllable boundaries correctly. This is achieved by the expression

$$A_{syl} = A_{mm}\ (\ \#\ A_{mm})^*$$

In this expression, syllable structure represented by A_{mm} is followed by zero or more occurrence of the boundary marker (#) and a syllable form A_{mm}. The automaton A_{mm} accepts the sequence of syllables but we need to transform this into a transducer which inserts a boundary marker `#` after each syllable but not after the last syllable. This is simply achieved by computing the identity transducer for A_{mm} and replacing `#` with a mapping `ε : # ` in A_{mm}. Now, the syllabification transducer becomes

$$T_{mmsyl} = Id(A_{mm}\ )\ \ (\varepsilon : \#\ )\ Id(A_{mm})$$

The syllabification Transducer for words with standard syllable structure is shown in figure 1.

For *irregular* words, the stored character sequence is special. It usually uses invisible Myanmar sign *VIRAMA* (U+1039) in the input sequence encoding but it is required to output different character or characters according to the types of *irregular* words in section 4.3. Though *irregular* words are written in tradition forms and complicated, the result of their syllabification turned into standard word syllable structure.

We construct the finite state transducers (FST) for each type of syllabification of *irregular* words respectively and we also show finite state transducer for standard syllable structure in Figure 1.

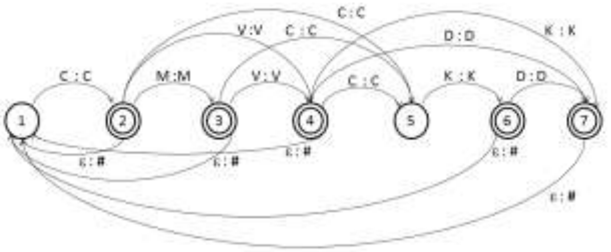


Fig. 1. Finite State Transducer for Myanmar syllabification

VI. EXPERIMENTAL RESULTS AND CONCLUSION

6.1 Experiments and Results

We use Stuttgart Finite State Transducer (SFST) Tools for our syllabification transducers although it is primarily concerned with morphology, for example, SMOR (a large German Morphological Analyzer). The specification of our syllabification transducer is written in SFST-PL, the programming language of the SFST tools which is a set of regular expressions.

As a technical element, it is noticed that the default Myanmar keyboard layout in Ubuntu is in Unicode 4.1 based "Myanmar 1" Font and the code point values of some characters are different from the Unicode 6.1 Myanmar character code table. Thus it is required to customize the keyboard layout file namely /usr/share/m17n/my-kdb.mim so as to get correct syllabification result.

For the test data set in our experiment, we use Myanmar Orthography published by Myanmar Language Commission which is a standardized system to write Myanmar words including rules of spelling [14].

Based on the regular grammar for syllable structure of Myanmar, we can identify correct syllable boundaries in the given texts. We tested all 11,732 distinct words contained in Myanmar Orthography corpus yielding 32,283 syllables covering standard and *irregular* words. Details of the results based on the types of word can be found as follows.

Table 4. Details of Experimental Results

No.	Type of Words	No. of words	Correctly Syllabified Words	% of correct syllabification for each word type
1.	Standard Words	**11,092**	11,092	100%
2.	Irregular Words	**640**		
	2.1 Consonant Stacking	253	245	96.83%
	2.2 Consonant Repetition	266	266	100%
	2.3 Kinzi	71	71	100%
	2.4 Great SA	26	26	100%
	2.5 Contraction	3	3	100%
	2.6 Loan Words	21	21	100%
	TOTAL	**11,732**		

By checking manually, we found that only syllabifcation of 9 stacked words out of 11,732 words are erroneous. Therefore, we received 99.93% of overall accuracy covering both types of word, standard and *irregular* words.

By doing error analysis, the errors are caused by those words in which free standing vowels ဣ (U+1023), consonant stacking and other sub-syllabic elements are mixed and our FST cannot find correct syllable boundaries for these words. The free standing vowels ဣ (U+1023) which is the combination of consonant letter အ (U+1021) and the vowel ိ (U+102D). However, this kind of error can be improved in our transducer and we will solve this issue in our future experiment.

We also analyze the syllabification results of our approach and other existing approaches on *irregular* words.

Table 5. Comparison of Syllabification Results on *Irregular* Words

Irregular Words	Corpus-based Method [6]	Rule-based Method [23]	Finite State Transducer Method
1. Consonant Stacking	NO	NO	YES
2. Consonant Repetition	NO	NO	YES
3. Kinzi	NO	NO	YES
4. Great SA	NO	NO	YES
5. Contraction	NO	NO	YES
6. Loan Words	YES but with some failures	YES for regular words	YES

Further, we summarize and compare the accuracy of the developed methods on Myanmar syllabification as follows.

Table 6. Summary of Myanmar Syllabification Methods

Method	Source Data	Total no. of Syllables	Syllabfication of Standard Words	Syllabification of *Irregular* Words	Accuracy (%)
Corpus-based Longest Matching Method [6]	11 Short Novels	70,384	YES	Pleases refer to Table 5.	99.96%
Rule-based Method [23]	Myanmar Orthography	32,238	YES	Please refer to Table 5.	99.96%
Finite State Transducer Method	Myanmar Orthography	32,238	YES	YES	99.93%

According to the above table, our approach promises the accuracy with 99.93% covering both standard and *irregular* words.

6.2 Conclusion

Syllabification is an important component of many speech and language processing systems, and this FST-based approach is expected to be a significant contribution to the field, and especially to researchers working on various aspects of Myanmar language and other Asian scripts.

For automatic syllabification of alphabetic languages, spelling does not have well-defined structure. Thus, the input texts are transliterated into phonetic symbols and syllabification is done on these transliterated texts, not on the input texts directly. Moreover, it is necessary to apply additional information using dictionary or annotated corpus and even FSA-based approach could be applied in statistical way.

Myanmar script is Indic script in origin like Lao and Thai. Myanmar syllable structure is well-defined and unambiguous. And thus, it could be represented in finite state model. In general, finite state language processing becomes popular because of their simple, elegant and efficient computational power. From computational point of view, finite state based methods achieve a good performance and they are suitable for application more interested in speed and memory footprint. Many works have been done for Myanmar syllabification but FSA approach has not yet been applied to Myanmar.

This paper proved that FSA approach gives significant solution for syllabification of Myanmar language as we achieve correct syllabification without applying step-by-step rules and the need of corpus. In other words, our proposed method is neither heuristic approach nor annotated corpus-based approach. Further, it could handle both regular and *irregular* syllable structures of Myanmar with acceptable performance. And we hope that it could be applicable to automatic syllabification of other syllabic writing systems.

In our further study, we will test our syllable segmentation FST on real online texts to evaluate the accuracy of the proposed approach and the evaluation process will be automated as a future improvement.

References

1. Bilal Arram. *Analysis of Urdu syllabification using Maximum Onset Principle*. http://crulp.org/Publication/Crulp_report/CR02_18E.pdf
2. Chamila, L., Randil, P., Dulip, L. Herath and Ruvan W. (2012): A Computational Grammar of Sinhala. *In the proceeding of 13th International Conference on Computational Linguistic and Intelligent Text Processing (CICLING)*, Mumbai, India.
3. Connie R. Adsett. (2008) : Automatic Syllabification in European Languages: A Comparison of Data-driven Methods, Master Thesis Dissertation, Dalhousie University, Halifax, Nova Scotia.
4. Gosse Bouma, Ben Hermans. *Syllabification of Middle Dutch.* http://alfclul.clul.ul.pt/crpc/acrh2/ACRH-2_papers/Bouma-Hermans.pdf
5. Hassain, S. (2004): Finite State Morphological Analyzer for Urdu. Master Thesis, National University of Computer and Applied Science, Lahore, Pakistan.
6. Hla Hla Htay, Murphy Kavi Narayana. (2008). Myanmar Word Segmentation using Syllable level Longest Matching. *In Proceedings of the 6 th Workshop on Asian Language Resources*, January 11-12, Hyderabad, India.
7. John Okell. (1994): *Burmese, An Introduction to the Script*. Northern Illinois University Press.
8. John Okell. (2002): *Burmese By Ear*, Audio-Forum, Sussex Publications Limited, Microworld House, 4 Foscote Mews, London W9 2HH. Downloaded at http://www.soas.ac.uk/bbe/ (2013, February 23)
9. Kiraz, G.A., M¨obius, B. (1998): Multilingual syllabification using weighted finite-state transducers. *Proceedings of the Third International Workshop on Speech Synthesis*. Jenolan Caves, Australia, pp. 71–76.
10. Le Hong P., Nguyen Thi M.H., Azim R., Ho Tuong V. (2008): A Hybrid approach to Word Segmentation of Vietnamese Texts. *In Proceeding of the 2nd International conference on Language, Automata Theory and Application, LATA 2008*, 240-249.
11. Marchand, Y., Connie A., Damper R. (2007) : Evaluation of automatic syllabication algorithms for English. *In Proceedings of the 6th International Speech Communication Association (ISCA) Workshop on Speech Synthesis*.
12. Maimaitimin Saimaiti , Zhiwei Feng, *A Syllabification Algorithm and SyllableStatistics of Written Uyghur*, *ucrel.lancs.ac.uk/publications/CL2007/paper/153_Paper.pdf*
13. Martin Hosken. (2012): Representing Myanmar in Unicode, Details and Example. Available at http://www.unicode.org/notes/tn11/ (2013, February 23)
14. Myanmar Language Commission. (2006).: *Myanmar Orthography*. Third Edition, University Press, Yangon, Myanmar.
15. Peter T. Daniels, William B. (1996): *The World Writing Systems*, Oxford University Press.
16. Phonpasit and et.al. *Syllabification of Lao Script for Line Breaking.* http://www.panl10n.net/english/outputs/Working%20Papers/Laos/Microsoft%20Word%20-%206_E_N_296.pdf
17. Ruvan W., Asanka W., and Kumudu G. (2005): A Rule Based Syllabification Algorithm for Sinhala, *Proceedings of 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, p. 438-449, Jeju Island, Korea.
18. Schmid, H. (2005): A programming language for finite-state transducers. In Yli-Jyrä, A., Karttunen, L., and Karhumäki, J., editors, Finite-State Methods and Natural Language Processing FSMNLP 2005.
19. Tin Htay Hlaing. (2012): Manually Constructed Context-Free Grammar for Myanmar Syllable Structure. *In the proceeding of the European Chapter of the Association of the Computational Linguistics(EACL), Student Research Workshop*.
20. Win, Kyawt Yin. (2011): Myanmar Text-To-Speech System with Rule-based Tone Analysis. PhD Dissertation, University of RyuKyus, Okinawa, JAPAN.
21. Y.A. El-Imam and Z.M. Don. (2000) *Text-to-Speech conversion of Standard Malay*.International Journal of Speech Technology 3, Kluwer Academic Publishers, pp. 129-146.
22. Yoshiki Mikami.: World of Scripts in Asia available at http://gii2.nagaokaut.ac.jp/ws/indic.html (2013, January 23)
23. Zin Maung Maung, Mikami Yoshiki. (2008): Rule-based Syllable Segmentation of Myanmar Texts. *In Proceedings of the 6 th Workshop on Asian Language Resources*, January 11-12, Hyderabad, India.