Syllable Level Segmentation between Myanmar Text and Phonetics Transcriptions

Kyaw Kyaw Maung

Abstract**—**To translate between Myanmar text and phonetics transcriptions, the system needs a segmentation system to break the sequence of text to tokens. Each token represents consonants, consonant clusters and vowels symbols defined in Myanmar language and phonetics transcriptions. Phonetics transcriptions for Myanmar language represented the pronunciations of Myanmar language. In a speech recognition system or text to speech synthesis systems, the phonetics transcriptions can be performed as an intermediate level between acoustics models and language models for translation. Both of phonetics transcriptions and Myanmar language have no space to identify the boundary of syllables. According to Myanmar language syllable structure, the Myanmar words are formed by combining syllables. And syllables are constructed by combining consonants/consonant clusters plus vowels or only the vowels. These system proposed the segmentation between both Myanmar language and phonetics transcriptions for Myanmar language.

Keywords**—**Consonants, Consonant Clusters, Myanmar Text, Phonetics Transcriptions, Syllables, Segmentations, Vowels.

I. Introduction

THERE are about 135 national races in Myanmar. Some of the ethnic groups have only the spoken language. But they do not have their own written language. Some of the researcher try to develop the written language that is based on their spoken language. Most of the pronunciation of the ethnic group's spoken language can be written by using our Myanmar's written language. In Myanmar language's writing system, there are many complex rules in writing. By combining some consonants and vowels, some of the pronunciation may be changed according to the combination of it. Some combination is complex and some of the combination cannot explained why the pronunciation has changed to other pronunciations. To develop a new written language system for some ethnic group, the researcher must convert their pronunciation into a sequence of phonetics transcriptions. Then the phonetics transcriptions are converted into Myanmar writing system. The main motivation of the proposed system is intended to help to solve these problems. In the proposed system used Unicode fonts for both Myanmar text and phonetics transcriptions. Myanmar3 Unicode fonts and Charis SIL fonts are used for Myanmar writing text and phonetics transcriptions.

For hard to detect the word boundary in both languages, the proposed system used to break the syllables according their Unicode encoding. These paper contributes the segmentation sections between Myanmar text and phonetic transcriptions for Myanmar language. The first part of the paper described about the Myanmar written language's consonants, vowels, consonant clusters and their combination rules. The second part described the segmentation in Myanmar text and how to match with the corresponding phonetics transcriptions..

II. Myanmar Language And Its Nature

Myanmar language, also known as Burmese, is an first and an official language of Myanmar. All of researchers who have studied the origin and development of the Burmese script accept that its source was the Brahmi script which flourished in India from 500 B.C to over 300 A.D. [1], [2]. Myanmar writing system constructed from consonants, consonant combination symbols called consonant clusters, vowels symbols related to relevant consonants and diacritic marks indicating tone level. Burmese is a tonal language and it using the Burmese script. Burmese characters are rounded in shape and the scripts is written from left to right. There is no space between each words. The phrases are separated by using space but it has not exact rules to use it [7], [8].

A. Myanmar Consonants

There are thirty-three consonants in Myanmar language. The another an additional consonant symbol called MYANMAR LETTER NNYA (U+101A) is comes from Pali language and it is widely used in Myanmar language. Another consonant is MYANMAR LETTER GREAT SA (U+103F). It is not commonly used in Myanmar language. It is comes from Pali language. These consonants are encoded in Unicode from (U+1000 to U+1021). The latest consonants called MYANMAR LETTER A (U+1021) has a specific nature among the consonants. It can be served as consonant as well as vowels. Some technician assumes that it is a vowel. But in writing system of Myanmar language, it is performed as a consonant and it can also performed like a vowel [1], [2], [9]. To form a word, Myanmar consonant must combined with Myanmar vowels. Myanmar consonant combined with another dependent consonant signs, it become combination of consonants called consonant clusters [3].

From the phonetics points of view, there are only 22 symbols to represent the Myanmar consonants. Some Myanmar consonants has the same speech sounds. The same

Kyaw Kyaw Maung is with the University of Computer Studies, Mandalay, Union of Myanmar (e-mail: kyawkyawmaung.ucsm@gmail.com).

http://dx.doi.org/10.15242/IIE.E0215040

pronunciation consonants can be described as follow:

1) MYANMAR LETTER GHA (U+1003) and MYANMAR LETTER GA (U+1002).

2) MYANMAR LETTER CHA (U+1006) and MYANMAR LETTER CA (U+1005).

3) MYANMAR LETTER JHA (U+1008) and MYANMAR LETTER JA (U+1007),

4) MYANMAR LETTER TTA (U+100B) and MYANMAR LETTER TA (U+1010),

5) MYANMAR LETTER TTA (U+1011) and MYANMAR LETTER TTHA (U+100C),

6) MYANMAR LETTER DA (U+1012) and MYANMAR LETTER DHA (U+1013), MYANMAR LETTER DDA (U+100D) and MYANMAR LETTER DDHA (U+100E),

7) MYANMAR LETTER BA (U+1017) and MYANMAR LETTER BHA (U+1018),

8) MYANMAR LETTER NYA (U+1009) and MYANMAR LETTER NNYA (U+100A).

The above consonants has the same representation symbols in phonetics transcriptions. MYANMAR LETTER GREAT SA (U+103F) is very rare to used. It comes from Pali. These consonants and its corresponded phonetics symbols are show in TABLE I.

TABLE I
MYANMAR CONSONANTS PHONETICS TABLE

No	Phonetics Symbols	Phonetics Unicode Points	Myanmar Consonants	Myanmar Unicode Points
1	k	006B	□	1000
2	k^h	006B+02B0	□	1001
3	ɡ	0261	□	1002
4	ɡ	0261	□	1003
5	ŋ	014B	□	1004
6	s	0073	□	1005
7	s	0073	□	1006
8	z	007A	□	1007
9	z	007A	□	1008
10	ɲ	0272	□	1009
11	ɲ	0272	□	100A
12	t	0074	□	100B
13	t^h	0074+02B0	□	100C
14	d	0064	□	100D
15	d	0064	□	100E
16	n	006E	□	100F
17	t	0074	□	1010
18	t^h	0074+02B0	□	1011
19	d	0064	□	1012
20	d	0064	□	1013
21	n	006E	□	1014
22	p	0070	□	1015
23	p^h	0070+02B0	□	1016
24	b	0062	□	1017
25	b	0062	□	1018
26	m	006D	□	1019
27	j	006A	□	101A
28	ɹ	0279	□	101B
29	l	006C	□	101C
30	w	0077	□	101D
31	θ	03B8	□	101E
32	ð	00F0	□	103F
33	h	0068	□	101F
34	l	006C	□	1020

B. Myanmar Dependent Vowels Signs

Dependent vowels signs are encoded in (U+102B to U+1031). They are dependent vowels called because they must combined with another consonants, or combination of consonants or another dependent vowels sings or independent vowel signs and various signs group. Various signs is also important in Myanmar language. They can combined with consonants and vowels. They show the tone level of the pronunciation. Because Myanmar language is a tonal language, the meaning may be change according to their tone level. Various signs are defined in Unicode from (U+1036 to U+103A). Among them, MYANMAR SIGN VIRAMA (U+1039) has specific in nature. When the font render, it cannot be seen. It can combined two consonants as a top-down positions. Its combination is very complex and there are many rules to used it. Another group of Myanmar sings has defined in Unicode (U+104C, U+104D, U+104E and U+104F). They can used independently without combination with another consonants and vowels except U+104E. These signs are one kind of independent word[1], [2], [9], [6].

C. Myanmar Independent Vowel Signs

In Unicode encoding, the Independent vowels are encoded from (U+1022 to U+102A). The pronunciation of these vowels are same with other some dependent vowels. They are called as independent vowels because they can read or write alone. It means it can stand without combination with another consonants or it can combined with another vowels or diacritic marks that indicating tone level. So, they can called independent vowels[1], [2], [9].

III. CLASSIFICATION OF MYANMAR VOWELS

Basic Dependent Vowels Signs has been presented in above section. There are 12 basic vowels defined in Myanmar language according to the Myanmar Primer. According to their tone, these vowels can expand to use in Myanmar language, there are more than 50 vowel sounds are exist in Myanmar language. Myanmar vowels sounds can be categorized into three groups. They are nasalized vowels, non-nasalized vowels and glottal stop vowels. They are 21 nasalized vowels, 22 non-nasalized vowels and 8 glottal stop vowels. These vowels are essential when construct a syllable. [3], [4].

TABLE II shows Myanmar Non-nasalized Vowels.

TABLE II
MYANMAR NON-NASALIZED VOWELS

No	Phonetics Symbols	Phonetics Unicode Points	Non-nasalized vowels	Myanmar Unicode Points
1	aˉ	0061+02C9	□□	1021+102C
2	aˆ	0061+02C6	□□□	1021+102C+1038
3	aʹ	0061+02B9	□□□	1021+102C+1037
4	ə	0259	□	1021
5	iˉ	0069+02C9	□□	1021+102E
6	iˆ	0069+02C6	□□□	1021+102E+1038
7	iʹ	0069+02B9	□□	1021+102D
8	uˉ	0075+02C9	□□	1021+1030
9	uˆ	0075+02C6	□□□	1021+1030+1038
10	uʹ	0075+02B9	□□	1021+102F
11	eˉ	0065+02C9	□□	1021+1031
12	eˆ	0065+02C6	□□□	1021+1031
13	eʹ	0065+02B9	□□□	1021+1031+1037
14	ɛˉ	025B+02C9	□□□	1021+101A+103A
15	ɛˆ	025B+02C6	□□	1021+1032
16	ɛʹ	025B+02B9	□□□	1021+1032+1037
17	ɔˉ	0254+02C9	□□□□	1021+1031+102C+103A
18	ɔˆ	0254+02C6	□□□	1021+1031+102C
19	ɔʹ	0254+02B9	□□□□	1021+1031+102C+1037
20	oˉ	0065+02C9	□□□	1021+1031
21	oˆ	0065+02C6	□□□□	1021+1031+1038
22	oʹ	0065+02B9	□□□□	1021+1031+1037

TABLE III shows Myanmar nasalized vowels. There are 7 nasalized vowels for higher tone, 7 nasalized vowels for lower tone and 7 nasalized are mid-tone vowels.

TABLE III
MYANMAR NASALIZED VOWELS

No	Phonetics Symbols	Phonetics Unicode Points	Nasalized Vowels	Myanmar Unicode Points
1	ãˉ	0061+0303+02C9	□□□	1021+1014+103A
2	ãˆ	0061+0303+02C6	□□□□	1021+1014+103A+1038
3	ãʹ	0061+0303+02B9	□□□□	1021+1014+103A+1037
4	ĩˉ	026A+0303+02C9	□□□	1021+1004+103A
5	ĩˆ	026A+0303+02C6	□□□□	1021+1004+103A+1038
6	ĩʹ	026A+0303+02B9	□□□□	1021+1004+103A+1037
7	eĩˉ	0065+026A+0303+02C9	□□□□	1021+102D+1014+103A
8	eĩˆ	0065+026A+0303+02C6	□□□□□	1021+102D+1014+103A+1038
9	eĩʹ	0065+026A+0303+02B9	□□□□□	1021+102D+1014+103A+1037
10	oʊ̃ˉ	006F+028A+0303+02C9	□□□□	1021+102F+1014+103A
11	oʊ̃ˆ	006F+028A+0303+02C6	□□□□□	1021+102F+1014+103A+1038
12	oʊ̃ʹ	006F+028A+0303+02B9	□□□□□	1021+102F+1014+103A+1037
13	aĩˉ	0061+026A+0303+02C9	□□□□□	1021+102D+102F+1004+103A
14	aĩˆ	0061+026A+0303+02C6	□□□□□□	1021+102D+102F+1004+103A+1038
15	aĩʹ	0061+026A+0303+02B9	□□□□□□	1021+102D+102F+1004+103A+1037
16	aʊ̃ˉ	0061+028A+0303+02C9	□□□□□	1021+1031+102C+1004+103A
17	aʊ̃ˆ	0061+028A+0303+02C6	□□□□□□	1021+1031+102C+1004+103A+1038
18	aʊ̃ʹ	0061+028A+0303+02B9	□□□□□□	1021+1031+102C+1004+103A+1037
19	ʊ̃ˉ	028A+0303+02C9	□□□□	1021+103D+1014+103A
20	ʊ̃ˆ	028A+0303+02C6	□□□□□	1021+103D+1014+103A+1038
21	ʊ̃ʹ	028A+0303+02B9	□□□□□	1021+103D+1014+103A+1037

TABLE IV shows the glottal stop vowels. There are only 8 types of glottal stop vowels.

TABLE IV
GLOTTAL STOP VOWELS

No	Phonetics Symbols	Phonetics Unicode Points	Glottal Stop Vowels	Myanmar Unicode Points
1	aʔ / ʌʔ	0061+0294 1D27+0294	□□□	1021+1010+103A
2	ɪʔ	026A+0294	□□□	1021+1005+103A
3	eɪʔ	0065+026A+0294	□□□□	1021+102D+1010+103A
4	ɛʔ	025B+0294	□□□	1021+1000+103A
5	oʊʔ	006F+028A+0294	□□□□	1021+102F+1010+103A
6	aɪʔ	0061+026A+0294	□□□□□	1021+102D+102F+1000+103A
7	aʊʔ	0061+028A+0294	□□□□□	1021+1031+102C+1000+103A
8	ʊʔ	028A+0294	□□□□	1021+103D+1010+103A
9	aʔ / ʌʔ	0061+0294 1D27+0294	□□□	1021+1010+103A

IV. Segmentation Method For Proposed System

Unicode encoding formats for Myanmar language is based on the Phonetic Order of Myanmar language. Unicode standard defined the logical order of Myanmar letters. The font render system then display the correct forms of Myanmar letters. In general, the Unicode sorted order in the following order, (1) consonants (U+1000 to U+1021 and U+103F) (2) consonant clusters (U+103B to U+103E) (3) vowels (U+102B to U+1031) (4) killer symbols of vowels (U+1039 and U+103A), (5) the tone level signs (U+1036 to U+103A).[9] A syllable may be formed (1) only a vowel, or (2) the combination of a vowel and consonants, or (3) combination of vowels and consonant clusters. Vowels may be nasalized vowels or nasalized vowels or non-nasalized vowels. [10] In TABLES I, II, III and IV shows the number of Unicode points takes for the consonants, consonants clusters and vowels.

Segmentation procedure for Myanmar text is the following:

1) For segment for Myanmar text, input is a sequence of strings in Unicode Myanmar text.
2) The second step is to construct the Unicode-Phonetic character tables in five groups. The first group is for a string units that takes five Unicode character group. The second group is for a group that takes four Unicode characters. The third group is for three Unicode characters. The fourth group is for two Unicode characters and finally is for the one Unicode character table. This step is defining for the input Myanmar text tokens.
3) Take the five Unicode characters from input sequence of strings in order from left to right and match with the group table that hold the five Unicode character table constructed in step two. If not found, reduced to the input characters into four Unicode characters from left to right to match the four character Unicode table. If not found reduced to three input characters and then match with three character Unicode table. If not found and then match with two character Unicode table. If not found in two, it reduced to one character and then match with one Unicode character table. If not found in one Unicode character table, it may be white space or that character that not defined in Myanmar language and ignore it.
4) If found in one type of group five or four or three or two or one Unicode character tables, the input characters can be identify of its types such as consonants or consonant clusters or nasalized vowels or non-nasalized vowels or glottal stop vowels or independent vowels.
5) repeat to step 1 until reach end of the input Myanmar strings.

After finished segmentation steps, it can be combined as consonant plus vowels, or consonant clusters plus vowels or independent vowels or vowels can formed as a syllables.

V.Conclusion

For the phonetics transcriptions to Myanmar text is the same procedure with the Myanmar text segmentation. But the maximum length of Unicode character is four and the minimum length is one. It can be seen in the above TABLES I, II, III and IV. From Phonetics to Myanmar text, it may need to the language model to construct a word level and it is a future work for the proposed system.

References

[1] Myanmar Language Commission, "Myanmar-English Dictionary",, 11th ed., University Press, 2011.

[2] Myanmar Language Commission, "Myanmar Dictionary", 2nd ed., University Press 2008.

[3] D. T. Tun, "Acoustic Phonetics and Phonology of the Myanmar Language". Win Yadanar Press, 1st ed. 2007.

[4] D. T. Tun, "The subtleties of the Myanmar Language, Grammar, Segments and Prosody in the sound system of the language and spellings" Win Yadanar Press, 1st ed. 2012.

[5] M. Sun. "Myanmar Unicode", YoYar Press, 1st ed. January, 2003.

[6] U.N. Maung, "Myanmar Thinbongyi (Primer) Basic", Yang Aung Press, 1st ed, December, 2013.

[7] Z.M.Maung and Y.Makami,"A rule-based syllable segmentation of Myanmar Text", Proceeding of the IJCNLP-08 workshop of NLP for Less Privileged Language , Pg. 51-58, Hyderabad, India, January, 2008.

[8] The Unicode Consortium, 2012, The Unicode Standard 6.2.

[9] Myanmar Language Commission, "Myanmar Grammar", 1st ed., University Press, June, 2005.

http://dx.doi.org/10.15242/IIE.E0215040