

# Encoding Myanmar Language To Unicode

Author/s: Zaw Win Aung

Myanmar Burmese Language Support@sourceforge.net

VERSION: 1.0.1

Date: 19 MAY 2003

Prerequisite: This document assume that you have already studied the design document by Maw, A.(2001).

## Introduction

This document was started after we faced difficulties implementing Myanmar language engine using the encoding and implementation method by Dr.Aung Maw; Maw, A. (2001). We tried to incorporate his implementation to the Pango language processing engine (www.pango.org) on Linux, but, the processing complexities became paramount and we set off to create an easier method.

## Complexities in implementing Dr.Aung Maw's implementation

### Complexities of Virama

In Myanmar language, cases of (consonant + consonant + Virama + consonant) e.g. အတိတ်; can sometimes be written as Conjunct where Virama ( <sup>၎</sup> ) will disappear and the second consonant will go under the first consonant meaning that the Virama is implicit. The final implicit Virama form will look like this အတ္တ.

According to Maw, A. (2001) implementation, Virama is explicitly added and the engine has to sort out the rendering by suppressing the Virama and getting the smaller character of the second consonant and putting it under the first one (Note: Normal character can not go under because it is big, thus, font designer has to design a smaller version of normal character for the going-under case). To cater for cases where Virama has to be explicitly rendered, Dr.Aung Maw used Zero Width Non-Joiner (ZWNJ) U+200C, thus, Virama + ZWNJ means that Virama has to be rendered by the engine. See Maw, A. (2001) slide number 8.

Myanmar "medial consonant" YA' PIN (၂), YA' YIT (၆), WA' ZWE (၀) and HA' HTOO (၂) are derived from Pali. YA' PIN is the representation of Pali form that consonant YA (ယ) going under consonant KA (က) which will look like this (ယံ). YA' YIT is representation of Pali form that consonant RA (ရ) going under consonant KA (က) which will look like this (ရံ). WA' ZWE is the representation of Pali form that consonant WA (ဝ) going under KA (က) which will look like this (ဝံ). HA' HTOO is the representation of Pali form that consonant HA (ဟ) going under MA (မ) which will look like this (ဟံ).

Dr.Aung Maw incorporated that fact directly into the implementation and the result is that YA PIN is encoded as KA + Virama + YA and the engine has to replaced it with YA' PIN . And the same goes for YA' YIT, WA' ZWE and HA' HTOO. See Maw, A. (2001) slide number 9 and 10.

**Our implementation based on Dr.Aung Maw's implementation**

We separated the Virama from mingling with Conjunct as explicit Virama is used way more times than implicit Virama; Conjunct. You will find one or two or none conjunct in a page. It is rare in modern Myanmar as it is based on Pali. However, explicit Virama always make atleast 30% of characters in a single page.

Conjunct (Implicit Virama)

To encode the implicit Virama, we added 19 consonants which normally go under the previous consonant in the case of implicit Virama. Other consonants do not go under. These 19 consonants and their corresponding UNICODE are stated below.

Character	Pronunciation	UNICODE
ယံ	KA	U+105A
ရံ	KHA	U+105B
ဝံ	GA	U+105C

𑄣	GHA	U+105D
𑄤	CA/SA	U+105E
𑄥	CHA/SAA	U+105F
𑄦	JA/ZA	U+1060
𑄧	DDHA	U+1061
𑄨	TA	U+1062
𑄩	THA/HTA	U+1063
𑄪	DA	U+1064
𑄫	DHA	U+1065
𑄬	NA	U+1066
𑄭	PA	U+1067
𑄮	PHA	U+1068
𑄯	BA	U+1069
𑄰	BHA	U+106A
𑄱	MA	U+106B
𑄲	LA	U+106C

*Justification to our implementation:* Myanmar language is in writing script since 12<sup>th</sup> century; Ager, S. (2002); and it is in rock stable state. We could not see any further changes to it. In our UNICODE character set range, there are a lot of free character codes which are sitting idle. I think that adding another 19 characters is not disadvantage. It is an advantage because processing complexity will reduce. Even English character set use different code point for capital and small letter.

#### Dependent vowels

Before we start talking about dependent vowels, we have to talk about "medial consonants". In Dr.Aung Maw's implementation, "medial consonants"; YA' PIN, YA' YIT, WA' ZWE and HA' HTOO; are encoded to their Pali term using Virama. We, however, break out from this form and add separate code point to represent the four characters. Below is the table of these characters and their corresponding UNICODE.

Character	Pronunciation	UNICODE
↓	YA' PIN	U+103A
┌	YA' YIT	U+103B
o	WA' ZWE	U+103C
└	HA' HTOO	U+103D

These four "medial consonant" surrounds visually to the consonant and can not stand themselves without a consonant. As this characteristic is exactly the same as dependent vowels, we put them into the dependent vowels group which has 11-characters. Now, we have 15 dependent vowels (11 dependent vowels + 4 "medial consonants"). Based on this, we have 2 dependent vowels which go to left of the consonant, 3 to the right, 5 each to the top and bottom.

To be able to call them easier we will name them as below

- Dependent vowels which go to left of consonant - Leftie
- Dependent vowels which go to right of consonant - Rightie
- Dependent vowels which go to top of consonant - Topie
- Dependent vowels which go to bottom of consonant - Undie

Below is the table of 15 dependent vowels and their corresponding UNICODE.

Character	Pronunciation	UNICODE
<b>Leftie</b>		
⊖	THA WE HTOO	U+1031
┌	YA' YIT	U+103B
<b>Rightie</b>		
⋮	WIT SA POT/VISARGA	U+1038
↓	YA' PIN	U+103A
↻	YEE CHAR	U+102C
<b>Topie</b>		
⌢	AH THA' /VIRAMA	U+1039
o	LONE GEE TIN	U+102D
⋅	THE' THE' TIN/ANUSVARA	U+1036
⋒	NOT PYIT	U+1032
e	LONE GEE TIN SAN KAT	U+102E
<b>Undie</b>		
o	WA' ZWE	U+103C

ၵ	TA CHON NG' IN	U+102F
ၶ	HA' HTOO	U+103D
ၷ	HN'A CHON NG' IN	U+1030
ၸ	AUT KA MYINT	U+1037

We came up with a way to encode these 15 dependent vowels in a 16-bit UNICODE character. Myanmar UNICODE range starts from U+1000 and ends at U+109F. In binary it starts from 0001000000000000 and ends at 0001000010011111. That means that Myanmar UNICODE will never have bit-15 '1' in binary if we take from the left. And, we can encode remaining 15 bits; from bit-14 to bit-0; to represent the 15 dependent vowels, thus, all the dependent vowels surrounding a consonant can be encoded in just a character. Now, all we need to test is the 15th bit of the character. We will call this character DEpendent Vowel Encoding Character (DEVEC). The table below shows the Leftie, Rightie, Topie and Undie bit range.

Bit Range	
14 - 13	2 Lefties
12 - 10	3 Righties
9 - 5	5 Topies
4 - 0	5 Undies

The table below shows the bit position of the dependent vowels in a DEVEC and their corresponding binary.

Dependent Vowel	Bit position	Binary
ၵ	14	1100 0000 0000 0000
ၶ	13	1010 0000 0000 0000
ၷ	12	1001 0000 0000 0000
ၸ	11	1000 1000 0000 0000
ၹ	10	1000 0100 0000 0000
ၺ	9	1000 0010 0000 0000
ၻ	8	1000 0001 0000 0000
ၼ	7	1000 0000 1000 0000
ၽ	6	1000 0000 0100 0000
ၾ	5	1000 0000 0010 0000
ၿ	4	1000 0000 0001 0000
ႀ	3	1000 0000 0000 1000
ႁ	2	1000 0000 0000 0100

<u>၀</u>	1	1000 0000 0000 0010
<u>၀</u>	0	1000 0000 0000 0001

By analysing the above table, it is obvious that we can differentiate between normal Myanmar character and DEVEC by testing the presence of 15-th bit. For individual dependent vowel, all we need to do is test the presence of the respectable bit in the DEVEC.

## **REFERENCES**

Maw, A. (2001). Encoding of Myanmar Character Set and Implementation, Myanmar Standardization Committee

Ager, S. (2002). Burmese Script, Omniglot a guide to writing systems, Retrieved May 27, 2003, from <http://www.omniglot.com/writing/burmese.htm>

UNICODE Chart (2003). Myanmar UNICODE chart, UNICODE Consortium, Retrieved May 19, 2003, from <http://www.unicode.org/charts/PDF/U1000.pdf>

Whistler, K., Davis, M. (2003). Character Encoding Model, UNICODE Consortium, Retrieved May 19, 2003, from <http://www.unicode.org/reports/tr17/>