Dissociating Myanmar Medials
A Proposal to Encode Separate Myanmar Medials

Martin Hosken,
SIL International and Payap University

Introduction

It seems that whenever the Unicode encoding of the Myanmar script is discussed, the issue of whether the medials should have their own codes comes up. This proposal examines the issue and also looks at how a dissociation may be achieved.

Reasons for Dissociating:

L2/04-273 argues that medials are different from syllable chained letters and need different handling.

L2/05-178 recommends dissociating medials to enable simpler representation of other Myanmar based writing systems.

While there have been a number of requests to the UTC to dissociate the medials, all such requests have been rejected on the principle that what currently exists, works. There may well be a recognition that if the Myanmar block were being encoded from scratch today, the medials would be separated. But given the very strong requirement for stability placed upon Unicode, there needs to be sufficient need to justify the upheaval that a deprecation process would require.

Advantages of Dissociating

There are a number of advantages to dissociating the Myanmar medials, we examine some here.

Other Languages

It is very difficult to resolve script model issues when two models both adequately cover the data within the scope of a single language. But when other languages use the same script in slightly differing ways, new issues and requirements arise. This is where models based on encoding language rather than script can face difficulties. And so it is with the Myanmar encoding. Sgaw Karen is a Myanmar based script with some extra characters and also a different association between one particular medial and its base consonant. Mon has other medials than those listed here which need their own analysis:

Medial	*Myanmar/Unicode base*	*Sgaw Karen base*	*Mon base*
◌ျ	**အ (U+101A)**	ဝ (U+101C)	အ (U+101A)
◌ွ		**အ (U+101A)**	ဝ (U+101C)

Medial	*Myanmar/Unicode base*	*Sgaw Karen base*	*Mon base*
◌ျ	ရ (U+101B)	ရ (U+101B)	ရ (U+101B)
◌ြ	ဝ (U+101D)	ဝ (U+101D)	ဝ (U+101D)
◌ၠ	ဟ (U+101F)	ဟ (U+101F)	ဟ (U+101F)

Notice how the shape used for medial -ya in Myanmar is used for medial -la in Sgaw Karen. And that medial -ya is represented by a completely different shape in Sgaw Karen. Having the wrong medial be displayed (not just a variant) breaks the legibility requirement of plain text. Something has to give. Further the shape that is used for -ya in Sgaw Karen is used for -la in Mon.

A possible solution is a complex set of variant selectors, but this would add considerably to an already complex encoding block.

Sorting

The Myanmar language has a particularly complex sorting process. The basic categories for sorting are: [initial consonant], [medials[1]], [final], [vowel], [tone]. Whereas the stored order of a syllable is [initial consonant], [medials], [vowel], [final], [tone].

There are two approaches to sorting Myanmar data: The first is to pre-process the string to bring the relevant categories into the right order and to add information regarding which category a sequence of codes are in. The second is to create large sets of collating elements corresponding to: [initial consonant], [medials], [rhyme (vowel + final)] and [tone]. Currently, due to ambiguity between whether the sequence C `U+1039` C constitutes a final consonant followed by an initial or an initial followed by a medial, it is necessary to merge the tables for initial consonant and medials. For example is `U+1000 MYANMAR LETTER KA U+1039 MYANMAR SIGN VIRAMA U+101D MYANMAR LETTER WA` a final followed by an initial consonant or an initial consonant followed by a medial?

While this is an implementation issue, the current encoding of Myanmar requires either that sorting engines support relatively large merged collating sets or the ability to backtrack while converting codes to collation elements. Dissociating the medials would not fix the rhyme issue, but would alleviate the need for backtracking.

Kinzi

The current encoding of Kinzi is particularly problematic with regard to medials. The only reason that `U+1004 MYANMAR LETTER NGA U+1039 MYANMAR SIGN VIRAMA U+101B MYANMAR LETTER RA` is rendered as ◌ instead of ◌် is that the rendering process has specific knowledge of what is a medial and what is not. In UTN#11 page 5, we see that this issue is resolved by marking the medial using an initial `U+200D ZERO WIDTH JOINER`. Thus the encoding of a medial becomes context dependent. And searching for a medial becomes a question of a regular expression (`(U+1004 U+200D|[^U+1004]) U+1039 U+101A`, say) rather than a simple string search. Dissociating the medials would make the need for `U+200D ZERO WIDTH JOINER` redundant in this context, and simplify their identification.

[1] Since more than one medial may occur.

Word	Current encoding	Proposed encoding
	U+1021 **U+1004 U+1039** U+101D U+1031 U+1037	U+1021 **U+1004 U+1039** U+101D U+1031 U+1037
	U+1021 U+1004 **U+200D U+1039 U+101D** U+1031 U+1037	U+1021 U+1004 **U+103C** U+1031 U+1037
	U+1021 **U+1004 U+1039 U+200C** U+101D U+1031 U+1037	U+1021 **U+1004 U+1039 U+200C** U+101D U+1031 U+1037

Other Advantages

Some other advantages of note are:

- Dissociated medials better reflect the understanding of the script by native speakers of Myanmar. Medials are seen as being separate letters, not as forms of their base character.
- Implementation of basic rendering is greatly simplified if backward compatibility is not required.

Disadvantages of Dissociating

The disadvantages of dissociating are not so much introduced problems as the lack of anticipated advantages. The process of deprecation means that code sequences currently used for medials cannot be reappropriated for other uses. For example, the sequence U+1039 U+101A will not be usable for stacked ya (as opposed to medial -ya).

Dissociation Process

The basic dissociation process is to encode four new letters corresponding to the four medial letters and then to formally deprecate the existing model of virama plus base character for these four medials. The new characters will have no decomposition to include the previous consonants on which they were based.

Since the combining orders for Myanmar diacritics do not enable the normalization algorithm to enforce the listed combining order in the rubric for the Myanmar script in the Unicode Standard 4.0 table 10-3, it is best to give these new diacritics a combining order of 0, rather than risk making matters worse and having the normalization algorithm re-order characters in a string into the wrong order.

The following characters are proposed:

```
103A;MYANMAR SEMIVOWEL SIGN YA;Mn;0;L;;;;N;;;;;
103B;MYANMAR SEMIVOWEL SIGN RA;Mn;0;L;;;;N;;;;;
103C;MYANMAR SEMIVOWEL SIGN WA;Mn;0;L;;;;N;;;;;
103D;MYANMAR SEMIVOWEL SIGN HA;Mn;0;L;;;;N;;;;;
```

Bibliography

Hosken, Martin "A Sgaw Karen Unicode Proposal" (Unicode Document L2/05-178)

Hosken, Martin and Maung TunTunLwin "Representing Myanmar in Unicode" (Unicode Technical Note #11)

Myanmar Unicode and NLP Reaserch Center "Proposal of 4 Myanmar Semivowels" (Unicode Document L2/04-273)

O'Kell, John — *Burmese (Myanmar): An Introduction to the Script* (Northern Illinois University, 1994)

The Unicode Consortium — *The Unicode Standard 4.0* (Addison-Wesley, 2003)