L2/06-163

Non-Shan Issues Relating to N3080: Preliminary proposal for encoding Karen, Shan, and Kayah Characters

Submission by: Dr Richard Wordingham (richard.wordingham@ntlworld.com)
Date: 7 May 2006

I would like the UTC to consider the following five issues relating to characters for Karen and Kayah when discussing agenda item C.13.6: Preliminary proposal for encoding Karen, Shan, and Kayah characters.

Contents

The first three issues relate to the disunification of characters. The second two issues relate to the ability to produce a thorough update of the Myanmar section of the Unicode standard, though investigation of Issue 4 may reveal problems.

1. Doubtful Distinctions

The distinction between MYANMAR LETTER NNA and the proposed MYANMAR LETTER EASTERN PWO KAREN NNA seems to depend on the 'plain-text monofont requirement', that a single non-language sensitive font be able to render languages encoded in the Myanmar font as each language's speakers expect. If this requirement is not valid, then the case for making them separate characters is also invalid.

2. Myanmar or Common Tonemarks?

*U+106A MYANMAR SIGN WESTERN PWO KAREN TONE-2
*U+106B MYANMAR SIGN WESTERN PWO KAREN TONE-3
*U+106C MYANMAR SIGN WESTERN PWO KAREN TONE-4

Are not these glyph variants of standard tone marks U+02E8 or U+02E9, U+02E5 or U+02E6, and U+02EA respectively? Note that the latter marks are 'common', not peculiar to any script.

One difference is that these marks are proposed to be of class Mc with combining class 0, whereas the standard tonemarks are of class Sk, also with combining class 0. The only argument I see that the tone marks should be class Mc is that U+1038 MYANMAR SIGN VISARGA is also Mc, and Figure 8 lists it as a Western Pwo Karen tone mark.

Another possibly relevant feature is that MYANMAR SIGN WESTERN PWO KAREN TONE-1, TONE-2 and TONE-3 also occur with a dot below (see Figure 8). The

encoding for this dot should be documented. The dot does not seem to be different to U+1037 MYANMAR SIGN DOT BELOW.

3. MYANMAR VOWEL SIGN GEBA KAREN I

There are two challenges to its appropriateness of this as a new character. For typographical reasons, I prefer the analysis as double anusvara to the interpretation as a combining diaeresis.

3.1 Combining Diaeresis?

The S'gaw Karen use of anusvara described in http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3083.pdf rather suggests a Karen unification of U+102D and U+0307 COMBINING DOT ABOVE, beating the Turks to their (non-Unicode) decomposition of U+0069 LATIN SMALL LETTER I. This in turn rather suggests that this vowel sign is none other than

0308;COMBINING DIAERESIS;Mn;230;NSM;;;;;N;NON-SPACING DIAERESIS;Dialytika;;;

abstracted from U+00EF LATIN SMALL LETTER I WITH DIAERESIS.

As U+0308 does not belong to any specific script, we cannot appeal to the principle of script separation.

3.2 Double Anusvara?

Perhaps it is double anusvara, i.e. its properties should be

1098;MYANMAR VOWEL SIGN GEBA KAYAH TENSE I;Mn;0;NSM;102D 102D;;;;N;;;;;

On the face of it, this violates the stability pact. I wonder, though - is a new character allowed to have as an expansion what was previously an ill-formed sequence?

4. Extra Uses of MYANMAR SIGN DOT BELOW

It seems to be assumed that the third vowel in Figure 7 should be encoded as U+1037 MYANMAR SIGN DOT BELOW. What are the implications on text processing of using it as a vowel symbol, as Martin Hosken says is to be done in S'gaw Karen (http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3083.pdf)?

It might also be used as the dot below some of the Western Pwo Karen tonemarks – see Figure 8.

5. Co-occurring Vowel Signs

Proposed MYANMAR VOWEL SIGN KAYAH OE and MYANMAR VOWEL SIGN KAYAH EE co-occur with others (e.g. Myanmar vowel signs U and UU in Figure 11), and the order in which they should occur needs to be defined. Transposing them results in a canonically inequivalent sequence.