

# Response to Eric Muller's comments on draft Sinhala standard

Gihan Dias, ICT Agency of Sri Lanka

2004-06-14

L2/04-248

## 1. Names of characters

N1. page 5, in the table of consonants, you use the names “ndja” and “nda” where TUS 4.0 gives the annotations “nyja” and “nda” in the code charts. The first is repeated in the first note, so at least that one is probably not a typo. Could you confirm that those are the names you prefer? I am confident the UTC would have no problem updating the annotations for the next version.

The romanised representations of Sinhala letters need a bit more work. This item was not discussed at the last Sinhala WG meeting, but deferred to the next meeting.

## 2. Spacing vowel signs

Please see separate document.

## 3. Joiners to control shaping

I thank Eric, Paul, and others who contributed to the discussion on this issue. Let me respond to the points individually.

### a) conjunct letters:

In Sinhala, showing an explicit al-lakuna (i.e., virama) is by far the more common representation when a pure (i.e., unvowelled) consonant is followed by another consonant. However, conjunct forms may also be used for some combinations. As conjunction is the uncommon case, we decided to encode it with a zwj.

### b) rephaya:

Usage of the rephaya too is less common, and we use a zwj for this case.

### c) yansaya and rakaransaya:

These constructs *are* common. The rationale for our decision is as follows:

“A frequent criticism of Unicode is the lack of codes for the *yansaya* and *rakaransaya*. The Unicode documents fail to mention these two symbols. However, we realised that this omission was deliberate, as neither of these symbols are Sinhala letters. Rather, they are abbreviations for the letters  $\text{ය}$  and  $\text{ර}$  respectively, when they follow a pure consonant. E.g.,  $\text{සකය}$  is the conventional way of writing  $\text{සකය}$  and  $\text{මිතු}$  represents the sequence  $\text{මිතර}$ .

As such words are generally spelled using the *yansaya* or *rakaransaya*, it was initially proposed to represent a  $\text{ය}$  or a  $\text{ර}$  following a pure consonant using the relevant symbol. However, some words, such as  $\text{මලේරාජ}$  and  $\text{පස්සාල}$ , do not use the constructs. We were thus faced with two alternatives:

- encode the common case without any special codes; e.g.,  $\text{ක} + \text{ර} = \text{කර}$  and use the code *zero-width non-joiner* (zwnj) to indicate when the construct should *not* be formed; e.g.,  $\text{ක} + \text{ර} + \text{zwnj} = \text{කර}$
- use the code *zero-width joiner* (zwj) to indicate when the *yansaya* or *rakaransaya* should be formed; e.g.,  $\text{ක} + \text{ර} + \text{zwj} = \text{කර}$ ,  $\text{ක} + \text{ර} = \text{කර}$ .

The first alternative yields a shorter code sequence for the more common case, and also follows the Unicode convention that the common case is encoded without special codes. However, the committee selected the second alternative for two reasons.

1. Keying in the sequence  $\text{ᩃ} + \text{᩵} + \text{ᩁ}$  would otherwise have automatically produced a  $\text{ᩃ᩵ᩁ}$ , even if not desired by the user. As the recommended keyboard has specific keys for the  $\text{ᩃ}$  and  $\text{᩵}$ , users would use these keys to generate the *yansaya* and *rakaransaya*, and the above key sequence for producing  $\text{ᩃᩁ}$ .

2. Using the *zwj* to produce the *yansaya* and *rakaransaya*, which are forms of conjunct letters, allows us to use the same representation for all conjunct letters.”

#### d) touching letters

After much discussion, we decided that these are distinct from conjunct letters, and should be encoded differently. My suggestion (open to your recommendations) is:

- normal case (explicit virama)  $C + \text{al-lakuna} + C$
- conjunct letters (half-form combined with a full-form)  $C + \text{al-lakuna} + \text{zwj} + C$
- “touching” letters (two full forms touching)  $C + \text{al-lakuna} + \text{zwnj} + C$

#### 4. Typos

T1. page 4, 1st paragraph, 3rd line: change “syllabury” to “syllabary”. Note that Unicode actually characterizes Sinhala as an abugida rather than a syllabary; see TUS 4.0, section 6.1.

corrected

T2. page 6, note 1, 1st line: change “al-lakana” to “al-lakuna”.

corrected

T3. page 6, note 3, 3rd line: change “symbolises <glyph for U+0DBB> when by a consonant,” to “symbolises <glyph for U+0DBB> when preceded by a consonant,”

modified

T4, page 6, note 3, the glyph for rayanna + al-lakuna has in incorrect placement of the al-lakuna. The same occurs p 15, section 5.7, 1st paragraph.

The glyph looks OK to me. The al-lakuna on the rayanna (and a couple of other letters) is shifted to the left.

T5. page 8, table 3, row 6, column 2: the first glyph is for U+0DAD, it should be the glyph for U+0DD9 instead.

corrected

T6. page 9, 3rd paragraph, 2nd/3rd line: after “and follow a consonant” add “in memory”, to make sure that we are speaking about the coded characters rather than their visual rendering.

modified

T7. page 10, representative glyph for U+0DDA: the al-lakuna should be on the right of the dotted circle.

Due to an error in the font, this glyph was encoded as a sequence of codes, which are probably being rendered inaccurately on your system.

T8. page 11, bottom of the table: change “Concluded” to “Continued”

This is an SLSI idiosyncrasy.

T9. page 14, note 2 (top of the page), second line, list of vowel sign characters: add “+” between them to clearly show the characters; end of fourth line, the last glyph is that of U+0DDC, change to the glyph for U+0DDD.

done

T10. page 15, 1st paragraph of section 5.7, 1st line: change “The repaya <rakaransaya glyph> represents the letter” to “The repaya <repaya glyph> represents the letter”.

corrected

T11. page 15, section 5.8, end of 4th line: the al-lakuna is incorrectly positioned.

This is an error in the font.

T9. page 15, 3rd paragraph of section 5.7, 1st line: change “in words wich as” to “in words such as”.

corrected

## 5. Some questions

Q1. Throughout the document, you speak of consonant clusters as if the only possibility is two consonants (for example in section 5.8: “conjunct letters are represented by the sequence Cons + al-lakuna + zwnj + Cons. The second consonant may optionally be followed by a vowel sign”). Is it just because the vast majority of clusters have only two consonants, but all the situations you describe are really meant to also apply to clusters that have more than two consonants?

The only case where more than two consonants are conjuncted is when a conjunct letter is followed by a yansaya or rakaransaya as specified at the end of Sec. 5.8.


Q2. You mention specifically the conjuncts of the form rayanna/yayanna/yayanna,...

I quote from a paper we wrote:

“The sequence ඊය presents an interesting case. It may be written as ඊය, යී or යී [6]; e.g., ආඊය, ආයී or ආයී. The second form is represented by ට + <sup>ෆ</sup> + zwj + ය. The sequence for the third form was not obvious, but was finally defined as ට + <sup>ෆ</sup> + zwj + ය + <sup>ෆ</sup> + zwj + ය.”

Q3. In the description of touching letters in L2/04-231, you seems to consider a conjunct alparapraana dayanna/mahaapraana dayanna (U+0DAF, U+0DB0) is essentially equivalent to sanyaka ddayanna (U+0DAC).

No. These two glyphs are similar, but lexically distinct.

alparapraana dayanna + mahaapraana dayanna =  sanyaka ddayanna = 

[note that the 1<sup>st</sup> glyph above is wrongly coded in the font used. Please consider only the shape.]

Similarly with dantaja sayanna/dantaja sayanna (U+0DC3, U+0DC3) and muurdhaja sayanna. How strong is that equivalence?

This is not so. If it appears that way, it's an artifact in the document I sent.