Rendering of Unicode Sinhala Characters

Harsha Wijayawardhana
University of Colombo School of Computing (UCSC)

Introduction

The Sinhala language is a member of the Indic family of languages. Indic writing has been considered as Orthographic, which simply means it is a mixture of Phonemic and Syllabic forms. Sinhala like the other Indic languages has syllable units and each syllable unit forms an individual visual unit or a glyph. For some words, glyphs are reconstructed in toto which may not have a resemblance at all to the original syllable units. When writing Sinhala, two glyphs can tie up to form conjunct consonants (bendi akuru); this also adds to the complexity of the language. In this backdrop, it is imperative to consider how these individual syllable units are stored and retrieved in Unicode. When retrieving, the viewer should be able to obtain what was written, as it was keyed into the Computer. Sorting of same words, which are written in different styles (Conjunct consonants as opposed to ordinary writing) should be done according to the sorting rules of the Sinhala language. In other words, words which are written in different styles should occur in the same place in the sorting order (c.f. "Some Salient Features of the Sinhala Alphabet" by Prof. J.B. Dissanyake).

ZWJ and ZWNJ (Zero Width Joiner and Zero Width Non Joiner)

It has been suggested by the Unicode consortium that ZWJ and ZWNJ should be introduced in Orthographic languages like Sinhala to achieve the following:

1. ZWJ joins two or more consonants to form a single unit (conjunct consonants).
2. ZWJ can also alter shape of preceding consonants (cursiveness of the consonant).
3. ZWNJ can be used to disjoin a single ligature into two or more units.

(Reference: http://www.unicode.org/standard/versions/Unicode3.0.1.html)

The storage of ZWJ and ZWNJ with letters (Unicode) will not alter the sorting order according to the collation algorithm of Unicode Technical Standards (UNICODE TECHNICAL STANDARD # 10: http://www.unicode.org/reports/tr10/tr10-8.html), though ZWJ and ZWNJ can be utilized to determine modifiers in Sinhala language such as Yansaya, Rakaransaya and to form conjunct consonants.

YANSAYA, RAKARANSYA and REPAYA

Yansaya and Rakaransaya are consonant modifiers (please refer to the section 2.3 of "Salient features of Sinhala Alphabet, Prof J.B. Dissanayake). Yansaya can be stored as

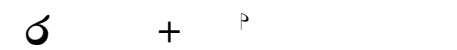
ප + ්‍ය

ZWJ +U+0DCA+U+0DBA:Yansaya

Rakaransaya can be stored as the following:

ZWJ+U+0DCA+U+0DBB:Rakaransaya

If one follows the above, Repaya is stored as



U+0DBB+U+0DCA+ZWJ

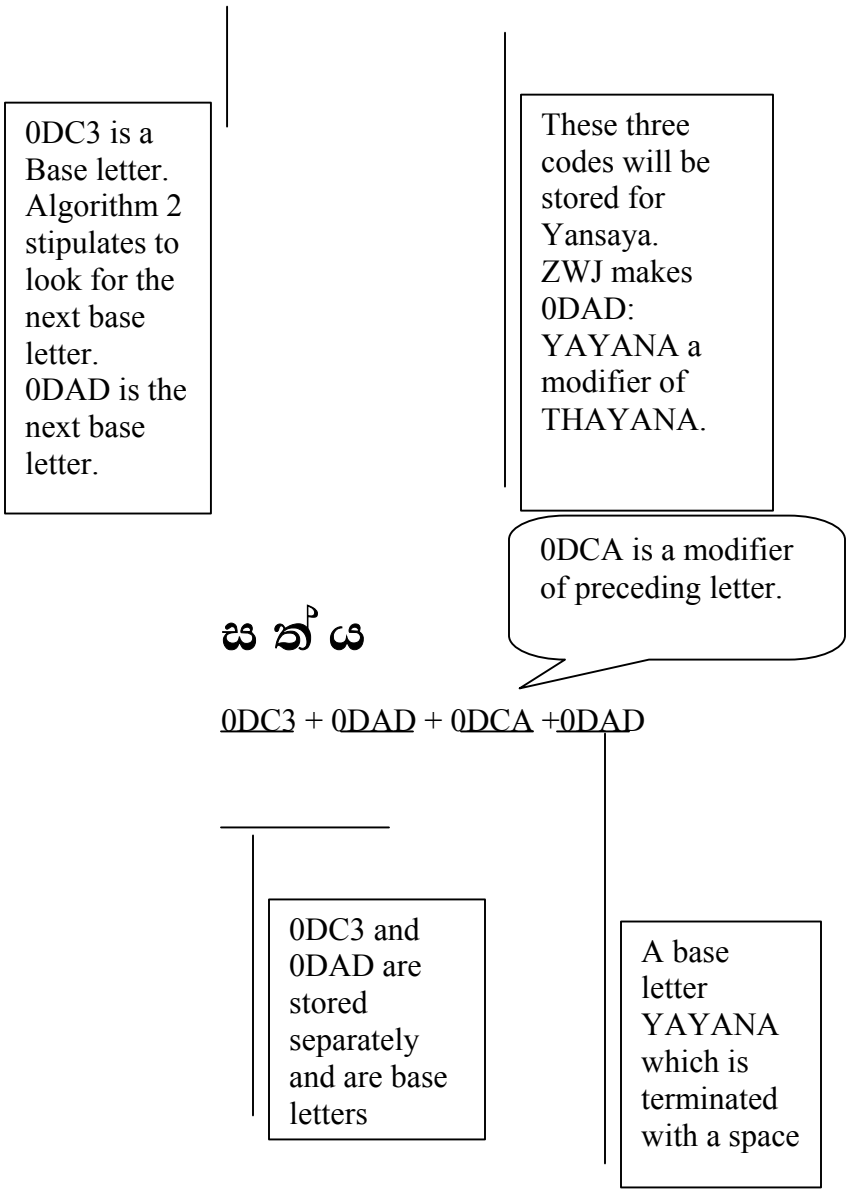
Possible rendering (displaying) algorithm:

1. A base letter is looked for from left to right when displaying is carried out. This letter can be a consonant or a vowel.
2. A base letter is followed by one or more modifiers terminating with a base letter or a valid delimiter character (e.g. space, tab, new line, punctuation etc.).
3. In the event of ZWJ preceding a base letter, that letter is considered a modifier of the previous base letter, which occurs to its left. Examples are YANSAYA and RAKARANSAYA.

NB: In an apparent exception to step 3 of the above algorithm, is the fact that in the Repaya and ‘sangnakaya’ and conjunct consonants. The ZWJ seems to separate a preceding modifier from its base character This however is owing to the fact that the resulting character glyph is visually more similar to the trailing base character. In fact, the behaviour of such phenomena with respect to sort order for instance seems to indicate that the real base character even in these two cases is the preceding base.

To display the above, the following code should be stored:

0DC3 + 0DAD +ZWJ+0DAD+0DCA



In the second example 0DCA hal lakuna follows thayana. Since the Zero width joiner is not here, this will be displayed without Yansaya.

The following can be written in two different ways and the both are shown below with codes:

බුද්ධ

0DB6+0DD4+0DAF+0DCA+0DB0

බුඬ

0DB6+0DD4+0DAF+0DCA+ZWJ+0DB0 (Please refer to the Sinhala Code Page)