L2/04-247

Depiction of Combining Marks in Isolation

Gihan Dias, ICT Agency of Sri Lanka
2004-06-14

In the draft Sri Lanka Standard 1134 Rev 2, we have specified code sequences for depicting vowel modifiers in isolation (i.e. without a base letter). Based on the response from Rick McGowan, we decided to change in the base of a sequence from zwnj to <space>. A number of comments were received on this change. Eric Muller has submitted a comprehensive document on this topic entitled "Using SPACE as a base character".

This document contains my comments on the above document, as well as the other responses received.

1. Need for the Depiction of Combining Marks in Isolation

Eric states:

... declare that displaying fragments such as the superscript or subscript form of RA is beyond the realm of plain text, or even styled text for that matter, and is in the realm of graphics.

In Sinhala, The strokes, one or more of which may combine with a consonant to form a letter, are well-defined lexical symbols, which are named, e.g. *ispilla, diga paa-pilla.* It is necessary to depict them in isolation for the following reasons:

1. For pedagogical purposes, in order to refer to the symbol.
2. Because these symbols are representable in isolation using legacy fonts. Not being able to represent them in Unicode would be perceived as a significant drawback of Unicode.

Although I have approached this issue from the perspective of Sinhala, similar considerations will apply for other Indic languages (and others).

Therefore, I recommend that some form of encoding be defined to represent these symbols.

2. Code sequences used to depict such marks

The Unicode documentation gives two alternatives:

Ch2: under ***Spacing Clones of European Diacritical Marks*** – use 0020 space or 00A0 nbsp.

Ch9: under ***Kannada*** – use 200C zwnj

Ch3 (of version 4.0.1): **Definition D14** – specifies the use of a combining character with no base character, but says that this is only when a combining character is at the start of text or follows a control or format character, such as a carriage return, tab, or RIGHT-LEFT MARK.

Current implementations depict the combining mark with a dotted circle, whenever not preceeded by a Sinhala consonant. This convention assists in proofing, as it indicates that the character is probably incorrect. Omitting the dotted circle and depicting the mark by itself, even when preceeded by a non-sinhala-consonant character, may be a solution, but leads to further questions, such as "should vowel re-ordering occur in such a case?"

In L2/04-234, Eric Muller discusses the issues with using space as a base character, and recommends the use of a *placeholder* character. This would be similar to the ISCII INV (invisible) character.

The SLS1134 working group does not have any strong views on the choice of a base character, and considers this a matter in which it will accept the guidance of the Unicode Technical Committee.