Title: **Proposal to change grapheme extending properties of various characters**
Author: **Martin Hosken**
Action: **For consideration by UTC**
Date: **2011-04-18**

Proposal: This document proposes:

1. The following characters have their grapheme break properties changed from SpacingMark to Other:
    - U+0E30, U+0E32, U+0E45
    - U+0EB0, U+0EB2
    - U+102B, U+102C, U+1038, U+1062, U+1063, U+1064, U+1067, U+1068, U+1069, U+106A, U+106B, U+106C, U+106D, U+1083, U+1087, U+1088, U+1089, U+108A, U+108B, U+108C, U+108F, U+109A, U+109B, U+109C, U+AA7B
    - U+19B0, U+19B1, U+19B2, U+19B3, U+19B4, U+19B8, U+19B9, U+19BB, U+19BC, U+19BD, U+19BE, U+19BF, U+19C0, U+19C8, U+19C9
    - U+1A61 U+1A63, U+1A64
2. The following characters have their grapheme break properties changed from Prepend to Other:
    - U+0E40, U+0E41, U+0E42, U+0E43, U+0E44
    - U+0EC0, U+0EC1, U+0EC2, U+0EC3, U+0EC4
    - U+AAB5, U+AAB6, U+AAB9, U+AABB, U+AABC

    Notice that this has the same effect as removing rule GB9b from UAX#29, since following this proposal, the set of characters with grapheme break property of Prepend, is empty.

Introduction: This document follows on from L2/10-460 and L2/11-051 after further research.

UAX#29, while not being followed by everyone in the industry, is an important document. Not only is UAX#29 part of the Unicode standard, there are implementations using based upon UAX#29, particularly the ICU library. The way ICU and UAX#29 gets used in applications is in calculating valid cursor positions within a string. If UAX#29 says that a grapheme cluster break cannot occur in a text sequence, then neither can a cursor occur within that sequence. By changing UAX#29, therefore, highly visible editing behaviours can change and this is what has happened for some southeast Asian scripts. Where, before, users could insert a cursor between a prevowel and a following consonant, for example, they cannot now. This highly visible regression (in their eyes) is cause for complaint. It should also be noted that implementations are increasingly relying on UAX#29 in its default interpretation to enforce cursor behaviour in text editing.

Rationale: There are two approaches to grapheme clustering within the scripts of the Indic subcontinent and southeast Asia. The first is that typified within India, that a cluster corresponds to an orthographic syllable. A cluster includes the initial consonant conjunct and the following vowel. The second, which is used in southeast Asia is that wherever you can sensibly, visually, insert a cursor, that corresponds to a cluster break. Thus one may insert a cursor between a pre-vowel and the consonant that is rendered following it, for example. All spacing characters, therefore may take a cursor position before them.

UAX#29 in its original inception, followed this separation of clustering approaches. But recently, for indeterminate reasons, perhaps of consistency, the indic clustering approach has been applied to the scripts of southeast asia and this has a very visible impact on the users of applications applying UAX#29.

The scripts affected are: Thai, Lao, Myanmar, New Tai Lue, Tai Tham and Tai Viet. Batak and Khmer may

both be considered to be in the southeast asian area, but no change was made to the behaviour of these scripts in recent changes to UAX#29 and no definite opinion on behaviour has been received to justify making a change to these scripts, so these two scripts have been excluded from consideration in this proposal.

The proposals at the start of this document specify the changes that give appropriate basic cursor positions for southeast asian scripts. Proposal 1 allows cursor points before spacing characters. Proposal 2 allows cursor points between a prevowel and the following consonant in visually ordered scripts.

Extended Grapheme Clusters: In the case of Myanmar, New Tai Lue and Tai Tham, prevowels are stored following the initial consonant cluster. This situation is covered by the differentiation between extended and non-extended grapheme clusters. In some cases, an implementation will want to allow a cursor point between the consonant cluster and the following prevowel, and in other implementations, it will not. The proposal, on the other hand, explicitly states that cursor positions as specified by the proposal, shall always be allowed whether an implementation is using extended or non-extended grapheme clusters.

Tailoring: One approach that has been suggested for addressing these issues, is to use per language (writing system) tailoring, to achieve the appropriate behaviour that users want. The reason that this is inappropriate is that what would be needed is the same tailoring for all languages using the same script, which is a waste of tailoring. Why not get the script right in the first place and save the tailoring. In addition, far fewer implementations will support per language tailoring (especially if text is not marked with language) than will support default script behaviour. Therefore, the use of per language tailoring to resolve this issue is considered an unacceptable solution.

Legacy: Another approach to solving this problem has been to relegate the required behaviour to a legacy grapheme breaking. But this means that the right behaviour is made non-default and the wrong behaviour is made default. This approach, while allowing the right thing to be done, means that implementors have to know that the default recommendation is wrong and to use the non-default. This is unlikely to happen and so implementations will implement wrong behaviour. This is also an unaccpetable solution.