L2/09-265

Date: August 3, 2009

To: Unicode Technical Committee

From: Peter Edberg on behalf of Apple Inc.

Subject: Grapheme break changes, especially for Thai & Lao

We have had strong feedback from users that the some aspects of the new UAX #29 default grapheme cluster boundary specification introduced in Unicode 5.1 are inappropriate for many or most usages in Thai, such as text editing (cursor movement, character deletion) and default alignment of search strings. This feedback is likely to apply to Lao as well, though we have not had direct feedback concerning Lao. The new specification is the result of consensus 114-C7 from UTC #114 / L2 #211 in February 2008.

This proposal requests either:

- A change to the current default grapheme cluster definition to delete rule GB9b and the Prepend value, and to delete the following characters from Extend: 0E30, 0E32, 0E45, 0EB0, 0EB2. This would treat Thai and Lao spacing vowels (except for Thai SARA AM and Lao AM) as separate units, not included in a cluster with the corresponding consonant.
- Or preferably, definition of a second grapheme cluster type ("spacing units"?) which for Thai and Lao differs from the extended grapheme cluster definition in the manner described above, and which for other scripts and characters may differ in other ways.

Background:

For Unicode 5.1, UAX #29 introduced the notion of extended grapheme cluster, and this is used for the new default grapheme cluster boundary specification. In particular, for Thai and Lao, the relevant changes were as follows.

1. A new grapheme cluster break property value "Prepend" was defined, and rule GB9b was added to prevent breaks after a Prepend character: "Prepend ×". Characters with the Prepend value are exclusively the Thai and Lao spacing vowels that **precede** (in display and memory order) the consonant to which they apply (in Unicode 5.0 these had the default grapheme cluster break value "Other"):

0E40	THAI CHARACTER SARA E
0E41	THAI CHARACTER SARA AE
0E42	THAI CHARACTER SARA O
0E43	THAI CHARACTER SARA AI MAIMUAN
0E44	THAI CHARACTER SARA AI MAIMALAI

0EC0	LAO VOWEL SIGN E
0EC1	LAO VOWEL SIGN EI
0EC2	LAO VOWEL SIGN O
0EC3	LAO VOWEL SIGN AY
0EC4	LAO VOWEL SIGN AI

The feedback we have is that for most or all common editing operations, none of these characters should be included in a grapheme cluster with the associated consonant. That is, effectively, rule GB9b and the Prepend value should both be deleted.

2. The following Thai and Lao characters were changed to have the existing grapheme cluster break property value “Extend”, before which breaks are not allowed; these are the Thai and Lao spacing vowels that **follow** (in display and memory order) the consonant to which they apply (in Unicode 5.0 these had the default grapheme cluster break value “Other”):

0E30	THAI CHARACTER SARA A
0E32	THAI CHARACTER SARA AA
0E33	THAI CHARACTER SARA AM
0E45	THAI CHARACTER LAKKHANGYAO
0EB0	LAO VOWEL SIGN A
0EB2	LAO VOWEL SIGN AA
0EB3	LAO VOWEL SIGN AM

The feedback we have is also that for most or all common editing operations, none of these characters except for THAI CHARACTER SARA AM (and LAO VOWEL SIGN AM) should be included in a grapheme cluster with the associated consonant. The special behavior for Thai SARA AM is because it includes a mark like the nonspacing sign THAI CHARACTER NIKHAHIT (the exception for Lao AM is for a similar reason).

This editing behavior for Thai characters should be used regardless of the user’s current locale. However, there are certain string processing operations for which the current grapheme break behavior may be appropriate in Thai, such as determining safe or appropriate boundaries at which to truncate or concatenate strings and other string processing operations.

The ideal solution is to define multiple types of grapheme clusters (cldrbug #2142 discusses this, see <http://www.unicode.org/cldr/bugs/locale-bugs?findid=2142>). Grapheme cluster types could be defined functionally, based on boundaries appropriate for related sets of tasks; some exemplars of such task sets include:

- backwards deletion
- cursor / arrow-key movement
- truncating strings
- collation

Ideally these functional tasks could be mapped onto cluster types defined solely in terms of character properties (the mapping will be locale-dependent, although the goal would be to create default behavior that would be suitable for most locales), such as:

- code point units
- spacing units, similar to legacy grapheme clusters (this could be defined to obtain the Thai and Lao editing behavior described above)
- extended grapheme clusters

Notes:

- CLDR 1.7 and ICU 4.2 already implement a Thai cluster break tailoring that implements the editing behavior described here, see: <http://www.unicode.org/cldr/bugs/locale-bugs?findid=2161>
- There was also an independently-filed ICU bug to request this: <http://bugs.icu-project.org/trac/ticket/6317>