Title: **Proposal to change UAX#29 in response to Thai, Lao, Tai Dam user needs**
Author: **Martin Hosken**
Action: **For consideration by UTC for inclusion in Unicode 6.0**
Date: **2010-09-17**

Proposal: This document proposes the following changes to UAX#29, with immediate effect:

- Remove all characters in the includes list for prepending characters: U+0E40-U+0E44, U+0EC0-U+0EC4, U+AAB5, U+AAB6, U+AAB9, U+AABB, U+AABC.
- Remove listed Thai and Lao characters from Spacing Marks (U+0E30, U+0E32, U+0E33, U+0E45, U+0EB0, U+0EB2, U+0EB3)
- Remove Thai examples from Table 1a, Extended Grapheme Clusters.

Rationale: This proposal undoes changes that were made to UAX#29 with regard to Thai, Lao and Tai Viet for Unicode 5.2 and following in response to requests from the user community.

The change was first introduced with v13 of UAX#29 following a public review (no. 138) where a somewhat unclear flag was raised on this issue http://www.unicode.org/L2/L2009/09123-pubrev.html#pri138. L2/10-281 attempted to rectify the issue but its recommendations actually result in no behavioural change, and so does not fix the problem.

The issue that the change to UAX#29 is causing problems may be seen in this bug report: https://bugzilla.gnome.org/show_bug.cgi?id=576156. As this bug discusses, it is possible for particular implementations to ignore UAX#29 in certain areas, but then what value UAX#29?

The desire for legacy grapheme clustering is not merely a whim of a subset of the user community. In Thailand there is considerable concern over this problem and this arises from a long history of the legacy grapheme cluster model being used, and their having no interest in changing it. WTT2.0, published in 1991 and later ratified as TIS 1566-2541 (1998) is the foundational document for Thai script computing within the country. Part 2 of that standard is concerned with input and output. It talks about different categories of letters:

> **NON** are non-Thai script characters or symbols.
>
> **CONS** are Thai consonants
>
> **LV** are leading vowels (U+0E40-U+0E44)
>
> **FV** are following vowels (U+0E30, U+0E32, U+0E33, U+0E45)
>
> There are also **AV** (above vowel diacritics), **BV** (below vowel diacritics), **AD** (above diacritics including tone marks), **BD** (below diacritics).

It then goes on in section 4 (p68) to describe a cell. A cell consists of a character from classes NON, CONS, LV or FV or a character from class CONS followed by 0-3 dead characters (non-spacing diacritics). In section 8 (p78) on editing text it then states what the effect of arrow keys are. Arrow keys move the cursor forward or back one cell. This means that a cursor may occur between a leading vowel and a following consonant, and also between a consonant and a following vowel.

To the Thai user community, therefore, the change to UAX#29 constitutes an unwelcome regression of behaviour and they strongly request its immediate reversion. They do not see a use case where having leading or following vowels cluster with the consonant would be desirable. Hence the proposed change.

Lao has identical behaviour to Thai in this area. Evidence that they too do not want to change comes from the same bug report and from personal correspondence with Lao users.

Tai Viet is a much smaller and less technically advanced user community. But in correspondence with Jim Brase, the Tai Viet proposal writer, he states that with 35 years experience with the script he has never seen any evidence that they consider a pre or post vowel as being in any way attached to the

consonant. In fact they have resisted any implementations that imply that. For example, an attempt was made to do an implementation that deleted consonant and pre vowel together, but it was resisted by users and the attempt failed. In addition, the Tai Viet encoding is based on the Thai and Lao models for the very issue regarding prevowels.

Recommendation: It is noted that the changes themselves were introduced with no evidence of any contact being made with the user community they affect. It should not be the place of user communities to have to proactively monitor the UTC to ensure that the UTC not introduce seemingly random changes that affect them. Instead, changes to Unicode need to show evidence of interaction with user communities, just as script proposals are so required. This is especially the case where something is being changed rather than added.

Bibliography

ดร. ทวีศักดิ์ กออนันตกูล (Dr. Thaweesak Koanantakool) คอมพิวเตอร์กับภาษาไทย: การพัฒนามาตรฐานเบื้องตัน สำหรับเทคโนโลยีสารสนเทศของไทย (Computers and the Thai Language: Towards developing a standard for Thai Technology) (Thailand 2534) ISBN 974-7570-66-1