**Title:** **Proposal not to encode 4 minority Thai letters for Patani Malay**
**Author:** **Martin Hosken**
**Action:** **For consideration by UTC**
**Date:** **2010-11-01**

**Proposal**: This document proposes not to add 4 new characters that are currently covered by U+02BC MODIFIER LETTER APOSTROPHE, U+02D7 MODIFIER LETTER MINUS, U+0303 COMBINING TILDE, U+0331 COMBINING MACRON.

**Introduction**: The Patani Malay orthography has just been officially recognised by the Thai government and they desire to create implementations for it. An overview description of that orthography is given.

Consonants:

| ก | กฺ | ค | คฺ | ง | ง̲ | จ | ช | ซ | ซฺ | ญ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0E01 | 0E01 0E3A | 0E04 | 0E04 0E3A | 0E07 | 0E07 0331 | 0E08 | 0E0A | 0E0B | 0E0B 0E3A | 0E0D |

| ญ̲ | ด | ต | ท | น | น̲ | บ | ป | พ | ฟ | ม | ม̲ | ย |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0E0D 0331 | 0E14 | 0E15 | 0E17 | 0E18 | 0E19 0331 | 0E1A | 0E1B | 0E1E | 0E1F | 0E21 | 0E21 0331 | 0E22 |

| ยฺ | ร | รฺ | ล | ว | อ | ฮ | ' | - |
|---|---|---|---|---|---|---|---|---|
| 0E22 0E3A | 0E23 | 0E23 0E3A | 0E25 | 0E27 | 0E2D | 0E2E | 02BC | 02D7 |

The initial consonant ' U+02BC MODIFIER LETTER APOSTROPHE marks glottalisation while the consonant - U+02D7 MODIFIER LETTER MINUS SIGN occurs between two vowels to indicate elision. Notice the glyph change with U+0E0D U+0331 compared to the single U+0E0D character. This follows the glyph shape changing convention in Thai when U+0E0D is followed by a lower diacritic.

Vowels:

| ◌ะ | ◌ั | ◌า | ◌ิ | ◌ี | ◌ือ | ◌ุ | ◌ู | เ◌ะ | เ◌็ | เ◌ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0E30 | 0E31 | 0E32 | 0E34 | 0E35 | 0E37 0E30 | 0E38 | 0E39 | 0E40 0E30 | 0E40 0E47 | 0E40 |

| แ◌ะ | แ◌็ | แ◌ | เ◌าะ | ◌็อ | ◌อ | โ◌ะ | โ◌ |
|---|---|---|---|---|---|---|---|
| 0E41 0E30 | 0E41 0E47 | 0E41 | 0E40 0E32 0E30 | 0E47 0E2D | 0E2D | 0E45 0E30 | 0E45 |

| ◌̃ะ | ◌̃า | ◌ุ̃ | เ◌̃าะ | แ◌̃ะ | แ◌̃ | ◌̃อ |
|---|---|---|---|---|---|---|
| 0303 0E30 | 0303 0E32 | 0E38 0303 | 0E40 0303 0E32 0E30 | 0E41 0303 0E30 | 0E41 0303 | 0303 0E2D |

This completes the list of valid character sequences used in Patani Malay. There are no tone marks, punctuation and numbers follow Thai.

**Rationale**: We discuss each of the non proposed characters in turn giving the various encoding options. Undefined behaviour is discussed later in this proposal.

| | |
|---|---|
| ◌̲ | There are a number of options open for encoding the underlined letters. One option is to encode each character separately. The problem here is that for every new orthography that gets created where this approach is used, new characters will need to be encoded. There seems little to be gained by encoding individual letters which are, in effect, presentation forms of an existing sequence. Using a sequence also fits with the use of U+0E3A THAI CHARACTER PHINTHU for generating new consonants, found elsewhere in the orthography. Encoding a new character for this character seems redundant, although there is one minor issue with regard to normalisation which will put a lower vowel (U+0E38 THAI CHARACTER SARA U or U+0E39 THAI CHARACTER SARA UU) before the U+0331 COMBINING MACRON BELOW even though the typing order would have the lower vowel following.<br>This diacritic is only known to occur with the characters listed here. Any other sequences are undefined. |
| ◌̃ | The nasalisation mark does not occur in conjunction with any other upper diacritics and so such sequences are undefined. Encoding options are U+0303 COMBINING TILDE or a script specific character yet to be encoded. Normalisation will place U+0303 following any lower vowels. |
| ' | The preglottalisation mark follows other such marks found in Latin. U+02BC MODIFIER LETTER APOSTROPHE is proposed here as the encoding. It has a script property of common which implies it is appropriate for use in other scripts than Latin. |
| - | This character is used to distinguish elided two vowel sequences from sequences of two separate syllables or words. U+02D7 MODIFIER LETTER MINUS is proposed as the encoding. Since this would mean that such a character is used within a word and not breaking it, then it is recommended that the general category for this character be changed from Sk to Lm. In addition this character is used as such in a number of other orthographies both in Thai and Latin script. |

**Undefined Behaviour**: A number of sequences have been described as being undefined. This means that there is no other language or orthography that gives any indication as to appropriate behaviour. Therefore an implementation is free to interpret the sequence as it will either to mark it as erroneous in some way (for example inserting a dotted circle) or to render the erroneous sequence allowing the user to see it naturally with the reader seeing the error, like any other spelling error.

**Technical Issues**: From an encoding perspective, it is no problem to use characters from different blocks, especially if those blocks are indicated as being for that purpose. But implementations have to deal with various complex details. For example, if a run segmentation process does not recognise that Thai may be used with characters from some other blocks, it may cause segmentation problems, in turn leading to display problems. A common problem is the use of U+02BC at the start of a run will indicate the run as being defaulting to Latin, even though following letters indicate it as being Thai. Likewise if a final letter is U+0303 or U+0331 then this may cause a run break as it is being typed if there is no following Thai letter.

**Preferred Conclusion**: While it is possible to encode any new characters for the encoding of a new writing system, the Patani Malay would prefer not to go through a full proposal process unless there are technical reasons that will cause problems to them in the future which would best be resolved by encoding new characters now.