ISO/IEC JTC1/SC2/WG2 N3384
L2/08-0xx
28-Jan-2008

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation internationale de normalisation
Международная организация по стандартизации

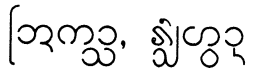
Doc Type: **Working Group Document**
Title: **Tai Tham Subjoined Variants**
Source: **Martin Hosken**
Action: **For consideration by JTC1/SC2/WG2**
References: **N3379, N3207R, SC2/N3982**
Date: **2008-Jan-28**

Introduction: Following the ad hoc meeting held in ChiangMai 21-22 Jan 2008 and reported in N3379, some questions arose regarding subjoined forms and variants. This document will discuss subjoined forms and their variants in Tai Tham. It will also propose the addition of two extra characters for addition to FPDAM5 (SC2/N3982) as modified by N3379 to address modern use of subjoined forms in Tai Tham. All character codes listed are in terms of FPDAM5 as modified in N3379. A complete chart may be found in N3379.

There are a number of subjoined characters in Tai Tham that take two contrastive forms. Examples thare are already encoded are TAI THAM CONSONANT SIGN MEDIAL LA, TAI THAM CONSONANT SIGN HIGH RATHA OR LOW PA, TAI THAM CONSONANT SIGN MA and even TAI THAM CONSONANT SIGN MEDIAL RA. Current analysis that presents these characters as existing due to a single set of spelling rules while sufficient for their encoding, is somewhat limited. This is higlighted when other subjoined variants are claimed. It is true that the medial consonants are primarily used as such, and as a result, their names are adequate for the task. But, for example, the sequence U+1A60 TAI THAM SIGN SAKOT, U+1A43 TAI THAM LETTER LA may also occur in medial position. The difference is in shape.

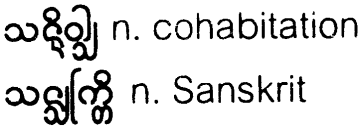
As soon as contrast based on shape is exposed the question needs to be raised as to whether these are merely glyph variants of each other whether across or within the same font. It is not hard to show that the currently encoded characters need to be separately encoded because they occur in contrast in the same textual environment (same document with same font) across different words and may also be shown to represent a consistent spelling. As extra subjoined forms are claimed, the same kinds of tests must be done. This is harder to achieve because such subjoined forms occur much more rarely and the variants even more rarely and then to find documents with them contrasting becomes nearly impossible. Two though, are more common and contrasts may be found in modern documents.

Subjoined Sa: Subjoined High Sa (U+1A60 TAI THAM SIGN SAKOT, U+1A48 TAI THAM LETTER HIGH SA) takes two forms, both of which may be used in the same text.



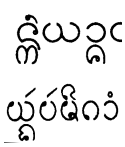
This sample taken from A Khün Reader by Anatole Peltier shows the two forms of subjoined sa. The first, at the end of the first word is the short form and the second at the start of the second word is the tall form.

The following samples are taken from a Khuen/English dictionary published this year.



On the left we see the short form and the right examples show the long form. In discussion with users of the script, the differences in character shape for these examples were expressed in terms of spelling rather than glyph (or style) ariation. This is in distinct contrast to the glyph variation surrounding subjoined nga

(U+1A26 TAI THAM LETTER NGA) where the following two examples are simply glyph variation.



Notice the two forms of subjoined nga. The first at the end of the top line and the second at the start of the second line. There is a contrast within the same word in the same book between the second form and a different scan with the first form:

The conclusion is that there are two independent spelling variants for subjoined sa. Of the two, the tall form is more prevalent and follows a pattern for other subjoined variants that are being claimed, that the tall form should be the default and be used for the subjoined character sequence (following U+1A60 TAI THAM SIGN SAKOT).

Subjoined Ba: Subjoined ba (U+1A37 TAI THAM LETTER BA) is shown in N3207R as taking the form of a small ba below the baseline. But the more common case is that it takes the form listed for a subjoined pha (U+1A38 TAI THAM LETTER HIGH PHA) which also takes the same form. The subjoined pha only occurs rarely, in words borrowed from Thai. The form listed for subjoined ba is used only occasionally in one or two words, but it is clearly a spelling difference. Thus we can compare:

 and 

The nature of the spelling rules surrounding the two forms of subjoined ba are clear even across language boundaries. Again we can conclude that there are two independent spelling variants for subjoined ba. Of the two the second form is by far the most prevalent and should be the one represented by U+1A60 TAI THAM SIGN SAKOT U+1A37 TAI THAM LETTER BA.

Encoding: Having established the separate variants, how do we encode them? The two alternatives are separate codes and variant selectors. On one level the variant selector approach is preferable in that it keeps the strong idea that the variants are in fact variants of the same subjoined characters. But acknowledging the extra complexity of using variant selectors, it is possible to encode these as simply extra characters in the block following the lead of the characters listed earlier. It would be wise though to give all of these characters compatibility decompositions into the sequences they are variants of.

The principle followed in the other characters of this form that have been encoded (apart from those called MEDIAL) is that the default character is subjoined in the normal way following U+1A60 TAI THAM SIGN SAKOT and that the irregularity is given its own code.

Recommendation: Add the following two characters:

- U+1A5D TAI THAM CONSONANT SIGN BA with representative glyph of a small ba (U+1A37 TAI THAM LETTER BA) occuring below a dotted circle
- U+1A5E TAI THAM CONSONANT SIGN SA with representative glyph of a small sa (U+1A48 TAI THAM LETTER HIGH SA) occurring below a dotted circle.

Issue: How many more of these are going to have to be encoded? On investigation of this phenomenon, claims were made for another 6 variants of subjoined characters. In searching for evidence to support such claims it is clear that any that do really need to be encoded are going to be from older texts and while the variants may gain in popularity, there is no need to rush to support them until clear evidence is found of their need. Even if and when they are attested it will need to be made clear that these characters are only used when contrastive spelling is required. This is in contrast to the characters proposed here which will take standard form regardless of inherent style.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** *Tai Tham Subjoined Variants*
2. Requester's name: *Martin Hosken*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date:
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *X*
   (or) More information will be provided later:

B. Technical General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters):
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *X*
      Name of the existing block: *Tai Tham*
2. Number of characters in proposal: *2*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary X | B.1-Specialized (small collection) | B.2-Specialized (large collection)
   C-Major extinct | D-Attested extinct | E-Minor extinct
   F-Archaic Hieroglyphic or Ideographic | G-Obscure or questionable usage symbols
4. Is a repertoire including character names provided? *yes*
   a. If YES, are the names in accordance with the character naming guidelines in Annex L of P&P document? *yes*
   b. Are the character shapes attached in a legible form suitable for review? *no*
5. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *Michael Everson*
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *no*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *yes*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *no*

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *no*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *yes*
   If YES, with whom? *Chiang Mai University, Khuen user community*
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *no*

1 Form number: N3102-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03)

Reference:

4. The context of use for the proposed characters (type of use; common or rare) *common*

Reference:

5. Are the proposed characters in current use by the user community? yes

If YES, where? Reference:

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *yes*

If YES, is a rationale provided? *yes*

If YES, reference: *addition to existing BMP block*

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *yes*

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *no*

If YES, is a rationale for its inclusion provided?

If YES, reference:

9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *no*

If YES, is a rationale for its inclusion provided?

If YES, reference:

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? *yes*

If YES, is a rationale for its inclusion provided? *yes*

If YES, reference: *this document*

11. Does the proposal include use of combining characters and/or use of composite sequences? *yes*

If YES, is a rationale for such use provided? *yes*

If YES, reference: *N3207R*

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *no*

If YES, reference:

12. Does the proposal contain characters with any special properties such as control function or similar semantics? *no*

If YES, describe in detail (include attachment if necessary)

13. Does the proposal contain any Ideographic compatibility character(s)? *no*

If YES, is the equivalent corresponding unified ideographic character(s) identified?

If YES, reference: