L2/04-110

ISO/IEC JTC 1/SC2/WG2 N2717

ISO/IEC JTC 1/SC2 Nmmmm

Title: Ordering rules for Khmer, Thai, and Lao; CTT suggestion

Source: Kent Karlsson
Date: 2004-03-09
Status: Expert contribution
Document Type: Working group document, regarding 14651 amd 2
Action: For consideration by the UTC and JTC 1/SC 2

1 Introduction

The text of current annex C.2 should be changed as detailed in a companion paper ("Ordering rules for Khmer, Thai, and Lao; suggestions for annex C.2 changes and a new annex B.5"). Here is a summary:

Consonants:

ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท
ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ฤ ล ฦ ว ศ ษ ส ห ฬ อ ฮ

Dependent vowels (though ◌ฺ is a dependent consonant, and ◌ฺ is a halant):

◌ำ -ะ ◌ั -ๅ -า ◌ิ◌ำ ◌ิ ◌ี ◌ึ ◌ื ◌ุ ◌ู เ- แ- โ- ใ- ไ- ◌ฺ

(where - is a consonant or consonant pair)

ๅ should be ordered as a compatibility variant of า. And ◌ำ should be ordered as า followed by ◌ิ (which is the logical sequence, despite the Unicode compatibility decomposition). Note that ◌ฺ is a virama/halant.

Tone marks, diacritics (◌ฺ◌ is a halant, ◌์ is a cancellation mark, ◌็ is a dependent vowel):

◌ฺ ◌์ ◌็ ◌่ ◌้ ◌๊ ◌๋ (where - is a consonant or vowel)

Punctuation:

ฯ ๆ ๏ ๚ ๛

While annex C.2 is concentrated on Thai, much of the same applies also to Lao and Khmer. The suggestion below includes what is deemed suitable as changes for the CTT. To get even better ordering for Thai, Lao and Khmer, a tailoring must be applied. It should be the same tailoring for all Thai, Lao and Khmer locales (or similar) and cover all of the three scripts.

2 Suggestion for the CTT for the Khmer, Thai, and Lao scripts

Note that the data for Lao has not been confirmed. However, corrections can be done in a tailoring.

For brevity, the declaration and weighting of the collation symbols are omitted, but the weighting given here (as if weighting was done directly on the here unweighted symbols) is that implied by the proper weighting of the collation symbols.

```
%%% Khmer/Thai/Lao punctuation:

<U17D3> IGNORE;IGNORE;IGNORE;<U17D3> %  , KHMER SIGN BATHAMASAT            (deprecated; lunar...)
<U17B5> IGNORE;IGNORE;IGNORE;<U17B5> % (glyphless) KHMER VOWEL INHERENT AA (deprecated)
<U17B4> IGNORE;IGNORE;IGNORE;<U17B4> % (glyphless) KHMER VOWEL INHERENT AQ (deprecated)

<U17DA> IGNORE;IGNORE;IGNORE;<U17DA> %   , KHMER SIGN KOOMUUT              end mark
<U0E5B> IGNORE;IGNORE;IGNORE;<U0E5B> % THAI CHARACTER KHOMUT         Po    end mark

<U17D4> IGNORE;IGNORE;IGNORE;<U17D4> %  , KHMER SIGN KHAN                 danda
<U0E2F> IGNORE;IGNORE;IGNORE;<U0E2F> % THAI CHARACTER PAIYANNOI      Lo    danda (not a letter!)
<U0EAF> IGNORE;IGNORE;IGNORE;<U0EAF> % LAO ELLIPSIS                  Lo    danda (not a letter!)
<U17D5> IGNORE;IGNORE;IGNORE;<U17D5> %  , KHMER SIGN BARIYOOSAN           double danda
<U0E5A> IGNORE;IGNORE;IGNORE;<U0E5A> % THAI CHARACTER ANGKHANKHU     Po    double danda

<U17D6> IGNORE;IGNORE;IGNORE;<U17D6> %  , KHMER SIGN CAMNUC PII KUUH       colon
<U17D9> IGNORE;IGNORE;IGNORE;<U17D9> %  , KHMER SIGN PHNAEK MUAN           bullet
<U0E4F> IGNORE;IGNORE;IGNORE;<U0E4F> % THAI CHARACTER FONGMAN        Po    bullet


%%% Repeat marks; not modifier letters, more like apostrophe and hyphen:
<U17D7> IGNORE;IGNORE;IGNORE;<U17D7> %  , KHMER SIGN LEK TOO % repetition sign
<U0E46> IGNORE;IGNORE;IGNORE;<U0E46> % THAI CHARACTER MAIYAMOK       Lm    repetition
<U0EC6> IGNORE;IGNORE;IGNORE;<U0EC6> % LAO KO LA                     Lm    repetition

<U17DC> IGNORE;IGNORE;IGNORE;<U17DC> %  , KHMER SIGN AVAKRAHASANYA
       % rare, shows a deleted Sanskrit vowel, like an apostrophe



%%% Khmer/Thai/Lao tone marks and other diacritics.
%%% Placed after a consonant or a dependent vowel.

<U17DD> IGNORE;<D17DD>;<MIN>;<U17DD> %    KHMER SIGN ATTHACAN
<U17D1> IGNORE;<D17D1>;<MIN>;<U17D1> %  , KHMER SIGN VIRIAM
<U0E4E> IGNORE;<D0E4E>;<MIN>;<U0E4E> % THAI CHARACTER YAMAKKAN       virama, but ordered thus

<U17CD> IGNORE;<D17CD>;<MIN>;<U17CD> %  , KHMER SIGN TOANDAKHIAT     cancellation mark
<U0E4C> IGNORE;<D0E4C>;<MIN>;<U0E4C> % THAI CHARACTER THANTHAKHAT    cancellation mark
<U0ECC> IGNORE;<D0ECC>;<MIN>;<U0ECC> % LAO CANCELLATION MARK         cancellation mark
```

```
<U17CF> IGNORE;<D17CF>;<MIN>;<U17CF> %  , KHMER SIGN AHSDA % sign used for single-consonant words
<U0E47> IGNORE;<D0E47>;<MIN>;<U0E47> % THAI CHARACTER MAITAIKHU, vowel sign, but ordered thus

<U17C9> IGNORE;<D17C9>;<MIN>;<U17C9> %  , KHMER SIGN MUUSIKATOAN
<U17CA> IGNORE;<D17CA>;<MIN>;<U17CA> %  , KHMER SIGN TRIISAP

<U17D0> IGNORE;<D17D0>;<MIN>;<U17D0> %  , KHMER SIGN SAMYOK SANNYA
       % used to indicate shortened inherent vowel
<U17C8> IGNORE;<D17C8>;<MIN>;<U17C8> % , KHMER SIGN YUUKALEAPINTU
       % makes the inherent vowel short and with an abrupt glottal stop


<U17CB> IGNORE;<D17CB>;<MIN>;<U17CB> %  , KHMER SIGN BANTOC % shortens preceding dependent vowel
<U17CE> IGNORE;<D17CE>;<MIN>;<U17CE> %  , KHMER SIGN KAKABAT % sign used with some exclamations

<U0E48> IGNORE;<D0E48>;<MIN>;<U0E48> % THAI CHARACTER MAI EK          Mn    above
<U0E49> IGNORE;<D0E49>;<MIN>;<U0E49> % THAI CHARACTER MAI THO         Mn    above
<U0E4A> IGNORE;<D0E4A>;<MIN>;<U0E4A> % THAI CHARACTER MAI TRI         Mn    above
<U0E4B> IGNORE;<D0E4B>;<MIN>;<U0E4B> % THAI CHARACTER MAI CHATTAWA    Mn    above

<U0EC8> IGNORE;<D0EC8>;<MIN>;<U0EC8> % LAO TONE MAI EK                Mn    above
<U0EC9> IGNORE;<D0EC9>;<MIN>;<U0EC9> % LAO TONE MAI THO               Mn    above
<U0ECA> IGNORE;<D0ECA>;<MIN>;<U0ECA> % LAO TONE MAI TI                Mn    above
<U0ECB> IGNORE;<D0ECB>;<MIN>;<U0ECB> % LAO TONE MAI CATAWA            Mn    above



%%% Thai consonants:

<U0E01>..<U0E2E> <S0E01>..<S0E2E>;<BASE>;<MIN>;<U0E01>..<U0E2E>
               % THAI CHARACTER KO KAI..THAI CHARACTER HO NOKHUK


%%% Lao consonants:
%%% (The code point range contains some code points which are not assigned to any character.)
%%% TUS 4.0: “Lao contains fewer letters than Thai because by 1960 it was simplified to be
%%% fairly phonemic, while Thai maintains many etymological spellings that are homonyms.”
%%% This quote suggests that there are many missing historic Lao letterforms in Unicode/10646.

<U0E81>..<U0EAE> <S0E81>..<S0EAE>;<BASE>;<MIN>;<U0E81>..<U0EAE>
               % LAO LETTER KO..LAO LETTER HO TAM
<U0EDC> "<S0EAB><S0E99>";"<BASE><BASE>";"<COMPAT><COMPAT>";<U0EDC> % LAO HO NO
<U0EDD> "<S0EAB><S0EA1>";"<BASE><BASE>";"<COMPAT><COMPAT>";<U0EDD> % LAO HO MO



% Khmer consonants:
<U1780>..<U1794> <S1780>..<S1794>;<BASE>;<MIN>;<U1780>..<U1794>
       % KHMER LETTER KA..KHMER LETTER BA
<U1795>..<U179A> <S1795>..<S179A>;<BASE>;<MIN>;<U1795>..<U179A>
       % KHMER LETTER PHA..KHMER LETTER RO
<U17CC> <S179A>;"<BASE><VRNT1>";"<MIN><MIN>";<U17CC> %  , KHMER SIGN ROBAT (combining)
       % robat is a syllable initial r in Indic loan words; written as a diacritic
<U17AB> <S17AB>;<BASE>;<MIN>;<U17AB> %  , KHMER INDEPENDENT VOWEL RY
       % glyph based on glyph for 1794
<U17AC> <S17AC>;<BASE>;<MIN>;<U17AC> %  , KHMER INDEPENDENT VOWEL RYY
       % glyph based on glyph for 1794
<U179B> <S179B>;<BASE>;<MIN>;<U179B> %  , KHMER LETTER LO
<U17D8> <S179B>;<BASE>;<COMPAT>;<U17D8> %   , KHMER SIGN BEYYAL (use is discouraged)
<U17AD> <S17AD>;<BASE>;<MIN>;<U17AD> %  , KHMER INDEPENDENT VOWEL LY
       % glyphs based on glyph for 1796
<U17AE> <S17AE>;<BASE>;<MIN>;<U17AE> %  , KHMER INDEPENDENT VOWEL LYY
       % glyphs based on glyph for 1796
<U179C>..<U17A2> <S179C>..<S17A2>;<BASE>;<MIN>;<U179C>..<U17A2>
       %  , KHMER LETTER VO.. , KHMER LETTER QA (glottal stop)


% Khmer independent vowels.
```

```
% They are collated as variants of the glottal stop + dependent vowel combination.

<U17A3> <S17A2>;<BASE>;<COMPAT>;<U17A3>
       %  , KHMER INDEPENDENT VOWEL QAQ (use is discouraged)
<U17A4> "<S17A2><S17B6>";"<BASE><BASE>";"<COMPAT><COMPAT>";<U17A4>
       %  , KHMER INDEPENDENT VOWEL QAA (use is discouraged)

<U17A5> "<S17A2><S17B7>";"<BASE><VRNT1><BASE>";"<MIN><MIN><MIN>";<U17A5>
       %  , KHMER INDEPENDENT VOWEL QI
<U17A6> "<S17A2><S17B8>";"<BASE><VRNT1><BASE>";"<MIN><MIN><MIN>";<U17A6>
       %  , KHMER INDEPENDENT VOWEL QII
<U17A7> "<S17A2><S17BB>";"<BASE><VRNT1><BASE>";"<MIN><MIN><MIN>";<U17A7>
       %  , KHMER INDEPENDENT VOWEL QU
<U17A8> "<S17A2><S17BB><S1780>";"<BASE><VRNT2><BASE><BASE>";"<MIN><MIN><MIN><MIN>"; <U17A8>
       %  , KHMER INDEPENDENT VOWEL QUK (should that be "<S17A2><S17BB><S17D2><S1780>" instead?)
<U17A9> "<S17A2><S17BC>";"<BASE><VRNT1><BASE>";"<MIN><MIN><MIN>";<U17A9>
       %  , KHMER INDEPENDENT VOWEL QUU
<U17AA> "<S17A2><S17BC>";"<BASE><VRNT2><BASE>";"<MIN><MIN><MIN>";<U17AA>
       %  , KHMER INDEPENDENT VOWEL QUUV (??)
<U17AF> "<S17A2><S17C2>";"<BASE><VRNT1><BASE>";"<MIN><MIN><MIN>";<U17AF>
       %  , KHMER INDEPENDENT VOWEL QE
<U17B0> "<S17A2><S17C3>";"<BASE><VRNT1><BASE>";"<MIN><MIN><MIN>";<U17B0>
       %  , KHMER INDEPENDENT VOWEL QAI
<U17B1> "<S17A2><S17C4>";"<BASE><VRNT1><BASE>";"<MIN><MIN><MIN>";<U17B1>
       %  , KHMER INDEPENDENT VOWEL QOO TYPE ONE
<U17B2> "<S17A2><S17C4>";"<BASE><VRNT2><BASE>";"<MIN><MIN><MIN>";<U17B2>
       %  , KHMER INDEPENDENT VOWEL QOO TYPE TWO
<U17B3> "<S17A2><S17C5>";"<BASE><VRNT1><BASE>";"<MIN><MIN><MIN>";<U17B3>
       %  , KHMER INDEPENDENT VOWEL QAU


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%% After all scripts, among the Hangul vowels and other Indic vowels %%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Dependent vowels (in collation order *after* all scripts):

%%% Thai final nasal. Logical placement in orthographic syllable is: last.

<U0E4D> <S0E4D>;<BASE>;<MIN>;<U0E4D> % THAI CHARACTER NIKHAHIT, Mn, above, * final nasal

%%% Thai (dependent) vowels (note that SARA AM is ordered as <SARA AA, NIKHAHIT>):

<U0E30> <S0E30>;<BASE>;<MIN>;<U0E30> % THAI CHARACTER SARA A
<U0E31> <S0E31>;<BASE>;<MIN>;<U0E31> % THAI CHARACTER MAI HAN-AKAT
<U0E32> <S0E32>;<BASE>;<MIN>;<U0E32> % THAI CHARACTER SARA AA
<U0E45> <S0E32>;<BASE>;<COMPAT>;<U0E45> % THAI CHARACTER LAKKHANGYAO   long stem AA vowel
<U0E33> "<S0E32><S0E4D>";"<BASE><BASE>";"<COMPAT><COMPAT>";<U0E33>
              % THAI CHARACTER SARA AM (compatibility decomposed as: <NIKHAHIT, SARA AA>)
<U0E34> <S0E34>;<BASE>;<MIN>;<U0E34> % THAI CHARACTER SARA I
<U0E35> <S0E35>;<BASE>;<MIN>;<U0E35> % THAI CHARACTER SARA II
<U0E36> <S0E36>;<BASE>;<MIN>;<U0E36> % THAI CHARACTER SARA UE
<U0E37> <S0E37>;<BASE>;<MIN>;<U0E37> % THAI CHARACTER SARA UEE
<U0E38> <S0E38>;<BASE>;<MIN>;<U0E38> % THAI CHARACTER SARA U
<U0E39> <S0E39>;<BASE>;<MIN>;<U0E39> % THAI CHARACTER SARA UU

<U0E40> <S0E40>;<BASE>;<MIN>;<U0E40> % THAI CHARACTER SARA E
<U0E41> "<S0E40><S0E40>";"<BASE><BASE>";"<COMPAT><COMPAT>";<U0E41> % THAI CHARACTER SARA AE
<U0E42> <S0E42>;<BASE>;<MIN>;<U0E42> % THAI CHARACTER SARA O
<U0E43> <S0E43>;<BASE>;<MIN>;<U0E43> % THAI CHARACTER SARA AI MAIMUAN
<U0E44> <S0E44>;<BASE>;<MIN>;<U0E44> % THAI CHARACTER SARA AI MAIMALAI


%%% Lao (dependent) vowels:

<U0EB0> <S0EB0>;<BASE>;<MIN>;<U0EB0> % LAO VOWEL SIGN A
<U0EB1> <S0EB1>;<BASE>;<MIN>;<U0EB1> % LAO VOWEL SIGN MAI KAN
<U0EBB> <S0EBB>;<BASE>;<MIN>;<U0EBB> % LAO VOWEL SIGN MAI KON
<U0EB2> <S0EB2>;<BASE>;<MIN>;<U0EB2> % LAO VOWEL SIGN AA
<U0EB3> "<S0EB><S0ECD>";"<BASE><BASE>";"<COMPAT><COMPAT>";<U0EB3>
              % LAO VOWEL SIGN AM (reordered in the tailoring)
```

```
<U0EB4> <S0EB4>;<BASE>;<MIN>;<U0EB4> % LAO VOWEL SIGN I
<U0EB5> <S0EB5>;<BASE>;<MIN>;<U0EB5> % LAO VOWEL SIGN II
<U0EB6> <S0EB6>;<BASE>;<MIN>;<U0EB6> % LAO VOWEL SIGN Y
<U0EB7> <S0EB7>;<BASE>;<MIN>;<U0EB7> % LAO VOWEL SIGN YY
<U0EB8> <S0EB8>;<BASE>;<MIN>;<U0EB8> % LAO VOWEL SIGN U
<U0EB9> <S0EB9>;<BASE>;<MIN>;<U0EB9> % LAO VOWEL SIGN UU

<U0EC0> <S0EC0>;<BASE>;<MIN>;<U0EC0> % LAO VOWEL SIGN E
<U0EC1> "<S0EC0><S0EC0>";"<BASE><BASE>";"<COMPAT><COMPAT>";<U0EC1> % LAO VOWEL SIGN EI
<U0EC2> <S0EC2>;<BASE>;<MIN>;<U0EC2> % LAO VOWEL SIGN O
<U0EC3> <S0EC3>;<BASE>;<MIN>;<U0EC3> % LAO VOWEL SIGN AY
<U0EC4> <S0EC4>;<BASE>;<MIN>;<U0EC4> % LAO VOWEL SIGN AI

<U0ECD> <S0ECD>;<BASE>;<MIN>;<U0ECD> % LAO NIGGAHITA, Mn, above, * final nasal



%%% Khmer dependent vowels:
<U17B6>..<U17C5> <S17B6>..<S17C5>;<BASE>;<MIN>;<U17B6>..<U17C5>
       % KHMER VOWEL SIGN AA..KHMER VOWEL SIGN AU

%%% Khmer pseudo-vowels:
<U17C6> <S17C6>;<BASE>;<MIN>;<U17C6> %  , KHMER SIGN NIKAHIT
<U17C7> <S17C7>;<BASE>;<MIN>;<U17C7> %  , KHMER SIGN REAHMUK



%% Viramas/halants, and subjoined consonants:

% The Khmer COENG, the consonant gluer (order among other viramas):
<U17D2> <S17D2>;<BASE>;<MIN>;<U17D2> % KHMER SIGN COENG (combining; glyphless; virama).


<U0E3A> <S0E3A>;<BASE>;<MIN>;<U0E3A> % THAI CHARACTER PHINTHU, Mn, below, virama/halant (newer)


<U0EBC> <S0EBC>;<BASE>;<MIN>;<U0EBC> % LAO SEMIVOWEL SIGN LO
<U0EBD> <S0EBD>;<BASE>;<MIN>;<U0EBD> % LAO SEMIVOWEL SIGN NYO
%%% TUS 4.0: “Myanmar and Khmer include a full set of subscript consonant forms used
%%% for conjuncts. Thai no longer uses any of these forms; Lao has just the two.”
%%% This quote suggests that there are many missing historic Thai and Lao letterforms
%%% in Unicode/10646 (or that PHINTHU should form conjuncts for historic Thai).
```

Acknowledgements

Thanks to Maurice Bauhahn for explaining the principles of Khmer collation. Thanks to Theppitak Karoonboonyanan for helping me with the Thai rules and Anousak Souphavanh for helping me with the Lao rules. Any errors or shortcomings here are of course mine (especially since I’ve done a number of interpretations and changes).