DATE: 2002-5-18

DOC TYPE: Expert contribution
TITLE: Comments on Requests to Deprecate Khmer Unicode U+17A3, U+17A4, U+17A8, U+17B4, U+17B5, U+17D3, U+17D8
SOURCE: Maurice Bauhahn
PROJECT: Khmer Unicode
STATUS: Comment on proposal
ACTION ID:
DUE DATE:
DISTRIBUTION: Worldwide
MEDIUM: PDF file
NO. OF PAGES: 5

A. Administrative	
1. Title	Comment on Proposals to Deprecate Khmer Unicode Characters
2. Requester's name	Maurice Bauhahn
3. Requester type	Expert
4. Submission date	18 May 2002
5. Requester's reference	ISO/IETC JTC 1/SC 2/WG2
6a. Completion	
6b. More information to be provided?	Only as required.

SUMMARY: Unicode/ISO have been asked to deprecate several characters that should not in fact be deprecated (U+17A4, U+17D8, U+17B4, U+17B5, U+17A). But do deprecate U+17A3, U+17D3.

The Khmer script grew from Indic roots, developed over time, retained complexity, is susceptible to multiple interpretations, and is used for a variety of languages. Hence an official group of nine expert Khmer linguists brought together by the Cambodian Government deliberated over a period of days to come to studied conclusions on how to handle Unicode encoding. They issued an official government four page report dated 14 August 1996. These linguists understood the needs of dictionaries, multiple languages using the same script, and transliteration. Proposals to deprecate several characters in Khmer Unicode would take away much of the multilanguage versatility which was deliberately incorporated into the standard in the first place.

INHERENT VOWELS (U+17B4 KHMER VOWEL INHERENT AQ; U+17B5 KHMER VOWEL INHERENT AA) {circled entities show absence of vowel}

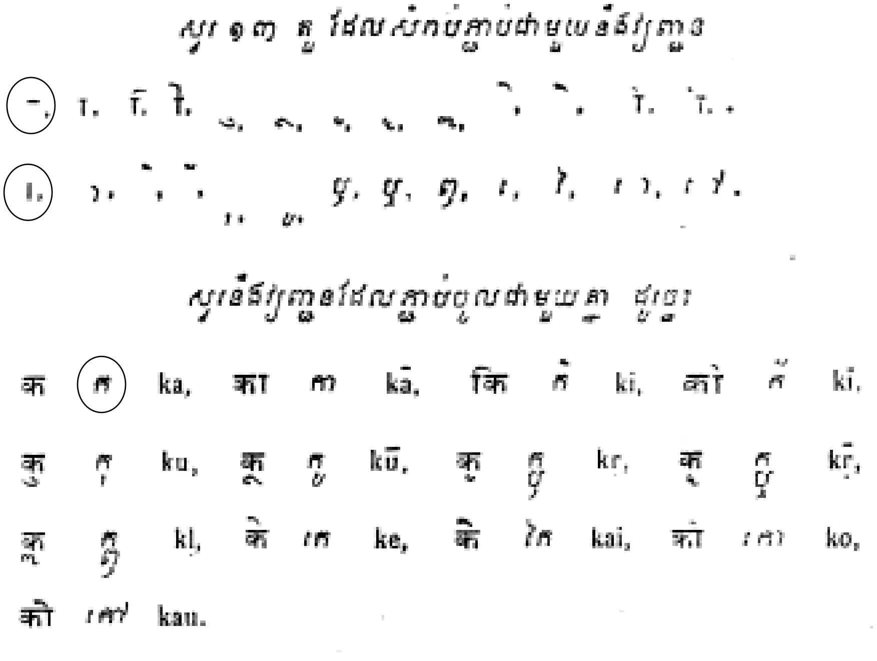
Note that in the Chuon Nath dictionary as shown below[1] there is explicitly displayed an unwritten/inherent vowel as one of the legitimate members of dependent vowels:

យកតួ អ ជាប្រធានដើមហើយបំបែកផ្សំជាមួយនឹងស្រៈនិស្ស័យទាំង ២១ តួជា

អ អា អិ អី អឹ អឺ អុ អូ អួ អើ អឿ អៀ អេ អែ អៃ អោ អៅ

អុំ អំ អាំ អះ

Similarly in a Sanskrit grammar as illustrated in the scan below[2] an inherent vowel is circled one in Sanskrit script and then twice in Khmer script where placeholders merely indicate its otherwise unwritten presence. Circled objects are various placeholders...illustrating unwrittenness of vowel:



ស្រៈ ១៣ តួ ដែលសំរាប់ផ្សំជាមួយនឹងព្យញ្ជនៈ

ស្រៈនិងព្យញ្ជនៈដែលផ្សំចូលជាមួយគ្នា ដូចជា

क ក ka, का កា kā, कि កិ ki, की កី kī,

कु កុ ku, कू កូ kū, कृ ក្ឫ kṛ, कॄ ក្ឬ kṝ,

कॢ ក្ឭ kḷ, के កេ ke, कै កៃ kai, को កោ ko,

कौ កៅ kau.

One could rightly argue that an inherent vowel is merely that...inherent in the consonant with which it is associated. Indic scripts have complex rules whereby each consonant is typically associated with a vowel sound and can lose those inherent vowels by a system of subscripting or virama or other signs. In Khmer the verbal COENG serves the function of killing the preceding inherent vowel sound just as in other Indic scripts an explicit or inherent virama facilitates that.

[1] Dictionnaire Cambodgien. Tome I. K.M. Cinquieme Edition. Phnom Penh: Éditions de l'Institut Bouddhique, 1967. p. 15

[2] Préah Sakyavong H. Tath, p. 3.

However, due to the cross-fertilisation of Indic loan words with Khmer words two contrasting streams of inherent vowels came into vogue. With the Indic loan words inherent vowels are: (1) Short and (2) Attach to final consonants in words. In Khmer language (and some tribal languages using the Khmer script) the inherent vowel is (1) Long and (2) Not pronounced in final consonants. This poses two problems:

1. Khmer dictionary compilers need to distinguish the pronunciations of these contrasting systems. Chuon Nath does this by arbitrarily inserting a KHMER SIGN YUUKALEAPINTU U+17C8 for the exceptional Indic inherent vowel. Obviously, this is ambiguous as that character is also used in the spelling of words. Note on page 1672 of the Chuon Nath dictionary[1] that the inherent vowel appears to affect sort order for the word series: អប   អប់   អប   (អៈប៉ៈ)   Note that the final word sorts separate from the apparently identically spelled first word...but the pronunciation given in parentheses shows that it elaborates two short inherent vowels (the earlier two words have a long and semi-long vowel respectively of typical pronunciation so do not have an explicit pronunciation guide).

2. There is increasing use of transliteration when the same language needs to be displayed in alternative scripts. For consistency of data store this transliteration may well be intelligent font based. No transliteration scheme for Khmer would omit explicit representation of the inherent vowel sounds. Hence one transliteration book[2] represents the Khmer long inherent vowels as â or ô (depending on the register series of the consonant associated with it). Another transliteration book[3] represents the Khmer long inherent vowels as ’a’. However a vowel sound is not dependably associated with every consonant. In some cases there are rules: (1) A consonant followed by a subscript (COENG) loses its inherent vowel, (2) A final consonant usually loses its inherent vowel, (3) A native inherent vowel is long. However, for Indic loan words the second and third rules usually do not apply: There is an inherent vowel sound on the final consonant and that Indic vowel is short.

In addition, it was anticipated that the encoding would double for phonetic transcription purposes of text to speech or speech to text (Khmer vowels play an extraordinarily important role in distinguishing words [whereas final consonants are often indistinguishable] ...hence reportedly Khmer has more vowels than any other language). So, there are specialised uses for these inherent vowel codes. The use of these codes is already narrowed and limited[4] in Unicode 3.0 to the small audience that needs them. There is no need to deprecate them. On the other hand, their excessive

[1] Dictionnaire Cambodgien. Tome II. K.M. Cinquieme Edition. Phnom Penh: Éditions de l’Institut Bouddhique, 1967. p. 1672. On page 1583 there is a further distinguishing of words having differing short versus long inherent vowels.

[2] U.S. Board on Geographic Names. Romanization Systems and Roman-Script Spelling Conventions. Defence Mapping Agency, 1994. p. 51.

[3] Randall K. Barry, Compiler. ALA-LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts. Washington: Library of Congress, 1997.

[4] “These are for phonetic transcription to distinguish Indic language inherent vowels from Khmer inherent vowels.” Khmer Range: 1780-17FF. The Unicode Consortium. The Unicode Standard Version 3.0. Reading, Massachusetts: Addison-Wesley, 2000. p. 472-475.

use should be discouraged as they would present problems if used inconsistently in cases of binary word comparisons.

INDIC INDEPENDENT VOWELS (U+17A3 KHMER INDEPENDENT VOWEL QAQ and U+17A4 KHMER INDEPENDENT VOWEL QAA)

Some have felt that insufficient attention was paid to Cambodians in their official capacities. The presence of these two characters is a direct result of a written request by an official government committee of the leading Khmer linguists in Cambodia. They distinguished U+17A3 (independent vowel; sorted first in Indic languages) and U+17A2 (consonant; sorted last in Khmer) on the basis of sort order and part of speech. Hence in a Khmer script Pali dictionary the ឣ section runs from page one to page forty-three. The separately headed ឤ section runs from page forty-three to page fifty-six. Other independent vowels follow before proceeding to the consonant ក. My preference (a) after hearing many discussions on the Unicode mailing list that sorting is a language-dependency and not a script dependency and (b) actually seeing the Sanskrit and Pali originals, is to lend less weight to the sorting differentiation for U+17A3 and allow its deprecation. U+17A4, however, is a unique construction not recognised in the Khmer language and should be retained with the limitation stated in Unicode v. 3.0: “used only for Pali/Sanskrit transliteration”. One scenario in which it would be adventageous to keep it as a unified character and not transliterated out into the Khmer language as two characters would be if it should ever occur as a subscript. All Khmer independent vowels can be transliterated into roughly equivalent consonant/vowel combinations, but for spelling purposes it is not valid to do so. A similar situation holds for U+17A4.

KHMER SIGN BEYYAL (U+17D8)

This punctuation sign is a direct result of a written request by an official government committee of the leading Khmer linguists in Cambodia. The committee noted its various forms are functionally equivalent. It is not a direct analogue to *et cetera* used in Western scripts because first of all it is called by two different names. Secondly it adopts many glyph forms (which would make it very difficult to search for if it were encoded in glyph parts). Thirdly deprecating it in favour of constituent parts it would confuse searches for sentences and word counts with its unusual mix of glyphs. Fourth, and most authoritatively the Chuon Nath Khmer Dictionary[1] states that *beyyal* is the name of the punctuation (*vannayut*) not that it is an abbreviation as such): *បេយ្យាលៈ (ប៉ៃយ៉ាល់ៈ) ន. (បា.) ឈ្មោះវណ្ណយុត្តមួយយ៉ាង សម្រាប់ប្រើបំប្រួញសេចក្ដីវែងឲ្យខ្លីឬច្រើនឲ្យតិច, ប្រើតែអក្សរ១តួខាងដើមថា ។ បេ ។ (ប៉េ) ក៏បាន, ប្រើតែអក្សរខាងចុងថា ។ល ។ ក៏មាន ។* Furthermore, the ALA-LC Romanization tables also classify it as a sign[2].

[1] Dictionnaire Cambodgien. Tome I. K.M. Cinquieme Edition. Phnom Penh: Éditions de l’Institut Bouddhique, 1967. p. 657.

OBSOLETE INDEPENDENT VOWEL (U+17A8 KHMER INDEPENDENT VOWEL QUK)

There was some request in the past to also deprecate this character. However a transliteration document[1] states “The independent character ឨ is romanized either ŏ or ŭ. Consult a reference source in case of uncertainty.” The presence of this character is also a direct result of a written request by an official government committee of the leading Khmer linguists in Cambodia.

KHMER SIGN BATHAMASAT (U+17D3)

This entry in Khmer Unicode was a mistake partly due to problems of communicating with one of the members of the Khmer linguists committee who proposed it after the committee’s report. It should be deprecated. The date formats of which BATHAMASAT is a part are only rarely used in modern times. The elaboration of these and their meanings by the Cambodian delegation is greatly appreciated. However, an even greater mistake would be to separately encode all the lunar dates, for they are definately ligatures of KHAN with Khmer numbers. It would be the functional equivalent of adding further vulgar fractions to Unicode. Since (a) their use is rare, (b) the character COENG U+17D2 literally means ‘foot’, (c) neither KHAN nor Khmer numbers are used in words associated with the use of COENG, I would suggest that there already exists in Khmer Unicode a means to represent these (number COENG KHAN or KHAN COENG number). Intelligent font designers would have to create some ligature glyphs to represent the series 0 to 15 (០__១៥) and KHAN (។) for numerator and denominator position. Admittedly I have seen Khmer numbers used as subscripts, but this was in a tagging context where the ‘ruby’ standard would be more appropriate.

2. Randall K. Barry, Compiler. ALA-LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts. Washington: Library of Congress, 1997.
1. U.S. Board on Geographic Names. Romanization Systems and Roman-Script Spelling Conventions. Defence Mapping Agency, 1994. p. 54.