DATE: 2002-05-18

DOC TYPE: Expert contribution
TITLE: Comments on 02098-KhmerSrakOm_Aam (Proposal to add Om and Am)
SOURCE: Maurice Bauhahn
PROJECT: Khmer Unicode
STATUS: Comment on proposal
ACTION ID:
DUE DATE:
DISTRIBUTION: Worldwide
MEDIUM: PDF file
NO. OF PAGES:

A. Administrative	
1. Title	Comment on Proposal to add Om and Am
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: The proposed KHMER VOWEL SIGN OM and KHMER VOWEL SIGN AM characters are in fact ligatures which reflect only a small part of vowel/sign combinations needed for languages expressed using the Khmer script. They should be rejected.

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. Furthermore, Khmer reportedly has more vowels than any other language. 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 a four page report dated 14 August 1996. These linguists understood the need (and many of the complications) of

expressing Sanskrit, Pali, and minority languages in the Khmer script. The following excerpt[1] shows the vowels they stated should be in Khmer Unicode:

ខ.១-សញ្ញាស្រៈនិស្ស័យមាន ១៦ តួ :

-ា ិ ី ឹ ឺ ុ ូ ួ ើ ឿ ៀ េ ែ ៃ ោ ៅ

The authors of the 02098 document have come to very different conclusions with regard to Khmer script vowels. They did not document their rationale, but evidences they could use include the list of vowels from the authoritative Chuon Nath dictionary as shown below[2] (note the two items proposed are circled):

យកតួ អ ជាប្រធានដើមហើយបំបែកផ្សំជាមួយនឹងស្រៈនំស្ស័យទាំង ២១ តួជា

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

(អុំ)អំ (អាំ)អះ

Furthermore these vowels are spelled with single syllable sounds. In addition, the derived Thai and Laotian scripts have registered such characters in their respective sections of Unicode. The NIKAHIT is not spelled separately in Khmer in these two instances (spelled OM and AM). And to top it all off: alphabetic ordering includes these in the sequence.

In the face of such obvious evidence, what rationale could there have been for a group of the leading Khmer linguists to not include these two items (or any of the four last *vowels* above) among the vowels in their recommendation for Khmer Unicode?

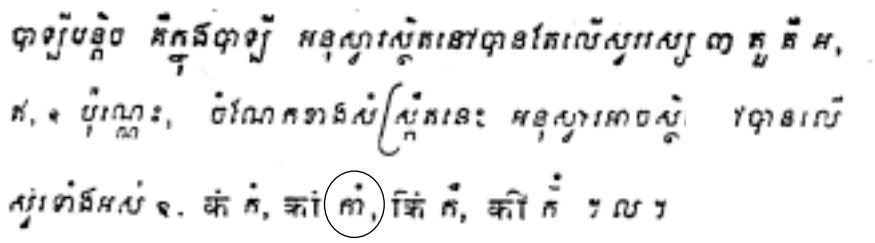
This document seeks to demonstrate that (1) NIKAHIT should be regarded as a distinct character, (2) that the combination of NIKAHIT with vowels would introduce ligatures into Khmer Unicode, (3) adding such ligatures would likely lead to inconsistent data entry/data store, and (4) adding such ligatures would lead to inadequate handling of non-Khmer language texts (especially Sanskrit).

I. NIKAHIT combines with vowels but is not essentially a vowel

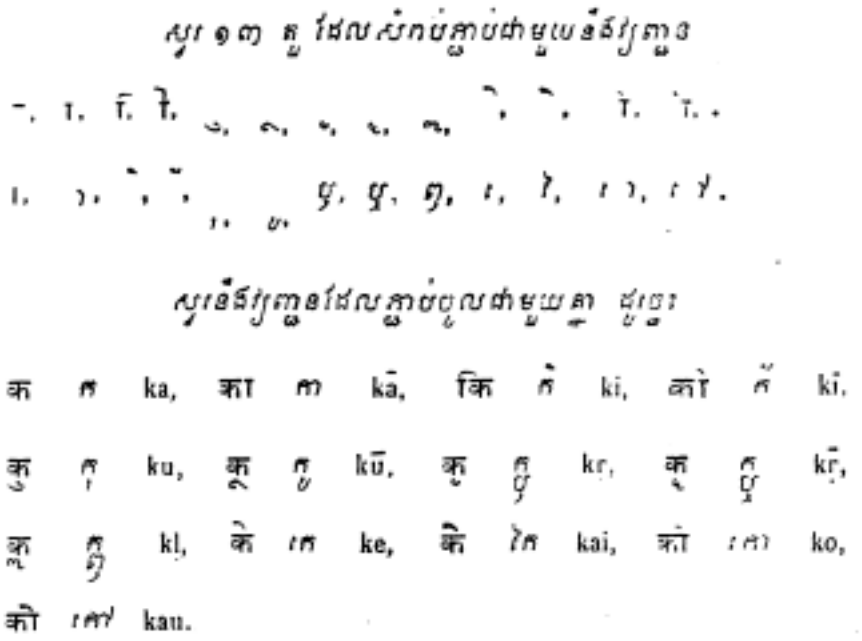
---

1. National Higher Education Task Force. Decisions of the secretariat of the National Higher Education Task Force regarding encoding of the Khmer language into Unicode/computer. p. 2. Available at http://www.bauhahnm.clara.net/Khmer/Page1.JPG Page2.JPG Page3.JPG Page4.JPG and in translation TaskForce.html
2. Dictionnaire Cambodgien. Tome I. K.M. Cinquieme Edition. Phnom Penh: Éditions de l'Institut Boud-dhique, 1967. p. 15

1. The scan below from a Khmer script grammer of Sanskrit[1] states “on the part of this

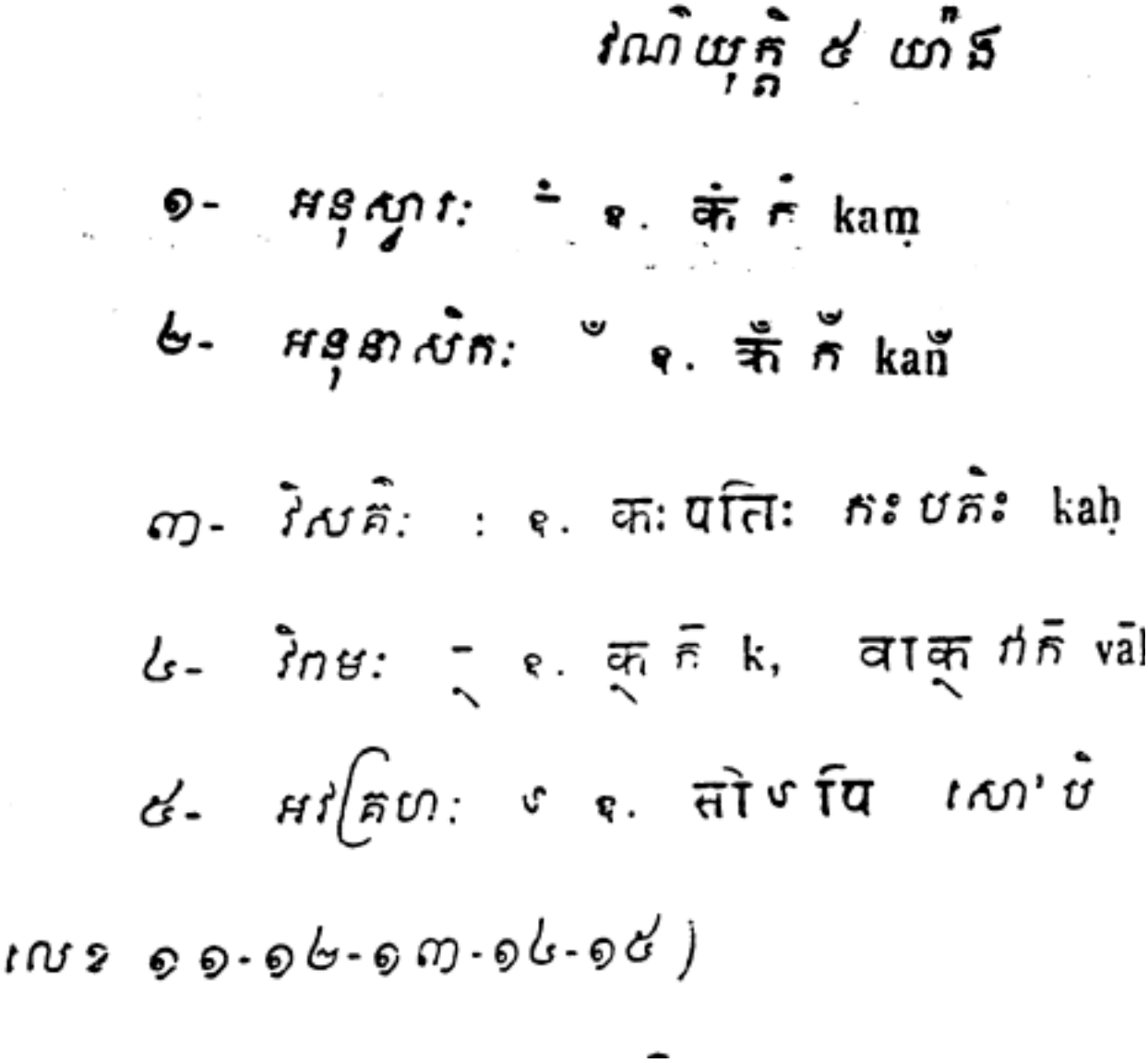


Sanskrit, NIKAHIT may be situated above all the vowels...” The text then illustrates four initial vowels and then the punctuation sign equivalent to ‘et cetera’. Note that the proposed KHMER VOWEL SIGN AM (proposed for 17DF, and circled above) is only one of many combinations possible. Note the scan below[2] of all Sanskrit vowels NIKAHIT may rest above:



1. Préah Sakyavong H. Tath, Grammaire Sanscrite: Tome II. Deuxième Édition. Phnom Penh: Éditions de l’Institut Bouddhique, 1956. p. 19. Please excuse the quality of the scans (they were scanned as TIFF files at 600dpi from a reproduction, imported into FrameMaker Mac, and exported as PDF).
2. Préah Sakyavong H. Tath, p. 3.

2. The same Sanskrit grammar[1] extracted below identifies five signs of which the first is NIKAHIT.



3. This is not confined to Sanskrit, however. The authoritative Chuon Nath dictionary defines NIKAHIT (but does not give such an entry to dependent vowels).

4. The sign NIKAHIT typically has an ‘m’ or ‘ng’ sound. This sound is preserved in both of the proposed items. Although it appears to be lost when followed by KHMER LETTER NGO, that is largely because it is redundant with the sound of the following consonant. Hence NIKAHIT has strong consonant qualities as opposed to vowel ones.

---

1. Préah Sakyavong H. Tath, p. 3-4

5. The Chuon Nath dictionary identifies ANUSVARA (another name for NIKAHIT) as derivative of Sanskirt and Pali in a separate entry[1] (there is no separate entry for dependent vowels).

> អនុស្វារ ឬ អនុស្សារ (អៈនុស-ស្វារ៑ ឬ -សារ៑) ន.
> (សំ. ឬ បា.) និគ្គហិតឬដំលើ (ំ) : *អនុស្វារមានសំឡេង*
> *ជាគរុ* (ប្រើជា-រៈ ក៏បាន) ។
> អនុស្សាវនក ឬ -កៈ (ម.ព. អនុស្សាវនា) ។

In a further entry for NIKAHIT[2], it is noted that the sign is virtually a consonant in Sanskrit.

> និគ្គហិត (និក-គៈហិត) ន. (បា. *និគ្គហិត* ឬ *និគ្គហីត*)
> ឈ្មោះគ្រឿងសម្គាល់មានរូបសណ្ឋានជាសូន្យមូលយ៉ាងនេះ
> “o” រាប់ចូលជាព្យញ្ជនៈដែរ, មានសំឡេងដូចតួ “*ង*”
> ប្រកបក៏មានដូចតួ“*ម*” ប្រកបក៏មាន ។ សំ. និង បា. រាប់
> និគ្គហិតនេះចូលក្នុងពួកព្យញ្ជនៈសេសវគ្គផងដែរ អានថា
> *អ័ង* ឬ *អ័ម* ។ ឯទម្លាប់នៃជនបរទេសខ្លះ គេអានថា *អ្ងឹ* ឬ
> *អ្ងឹម* ឲ្យគ្រាន់តែឮសំឡេងដូចគេគ្រហឹមងំក្នុងច្រមុះប៉ុណ្ណោះ
> ដោយគេយល់ថាអានកុំឲ្យឮច្បាស់ថា *អ័ង* ឬ *អ័ម* ដើម្បីឲ្យ
> និគ្គហិតនេះអាចរលាស់សំឡេងផ្សំនឹងស្រៈ *ឥ, ឧ* ហើយនិង
> ស្រៈឯទៀតខ្លះបានដោយងាយ សម្រាប់ភាសារបស់
> គេ ... ។

6. NIKAHIT is already recognised as a sign (not a vowel) in Khmer Unicode, set apart with the similar characters KHMER SIGN REAHMUK and KHMER SIGN YUUKALEAPINTU.

7. Note that the last four ‘vowels’ in the Chuon Nath dictionary are of a different nature from those that precede: (a) They all involve combinations with signs (two with the unwritten inherent vowel), (b) They follow the last vowel which matches Sanskrit/ Pali vowels, (c) they do not match Sanskrit or Pali vowels, and (d) they all have following consonantal sounds (ng, m, or h).

---

1. Dictionnaire Cambodgien, p. 1658.
2. Dictionnaire Cambodgien, p. 532.

II. The combination of NIKAHIT with vowels creates ligatures.

1. Not only is the NIKAHIT sound preserved in the two items proposed, the vowel sound is also preserved. Hence, two distinct characters retain their identity when used in combination. This is not the case in other Khmer vowels formed without signs that nevertheless share glyphs.

2. The Khmer and Laotian analogues which were brought into Unicode from legacy encodings have presented problems in their [insert thailaoam.tif] decompositions[1].

> The categories discussed above have covered essentially all of the characters with compatibility decompositions defined in TUS 3.1, leaving only a dozen or less. I cannot give an exact count since it is so unclear (at least to me) how the remaining characters are best characterised that another person might have already included them in one of the above sets and perhaps have a different residue. I will just mention two that I have found hard to categorise: U+0E33 THAI CHARACTER SARA AM “ ำ”, and U+0EB3 LAO VOWEL SIGN AM “ ຳ” Because these both decompose into a combining mark followed by a base characters, with the former combining with some preceding characters, I have not know which of the preceding categories, if any, to put them into.

3. By way of analogy there are many other combinations of dependent vowels with the signs KHMER SIGN REAHMUK and KHMER SIGN YUUKALEAPINTU which sort with the same weight as separate vowels in the Chuon Nath dictionary (see the document at http://www.bauhahnm.clara.net/Khmer/KhmerSortingUnicodegamma.pdf) but are not ‘officially’ recognised as vowels.

III. Adding these two items would introduce ambiguity into Khmer Unicode.

1. Were the two proposed items to be accepted, they would introduce alternative ways to enter the same material in addition to that originally intended. This, unfortunately, would lead to identical words entered by alternative means being treated as separate and distinct in binary searches and spell checks. A compatibility algorithm could, of course, resolve such difficulties...but given the speed required of these basic functions over a growing corpus of data, it is unlikely that a speed degrading compatibility algorithm will be introduced in the foreseeable future.

2. Whenever there is ambiguity, data entry would be slowed while the typist seeks to mentally resolve which alternative to choose.

IV. Adding such ligatures would lead to inadequate handling of non-Khmer language texts (especially Sanskrit).

---

1. Peter Constable. “A review of characters with compatibility decompositions”. Implementing Writing Systems: An Introduction. Melinda Lyons and NRSI Team, Editors. Preliminary edition. Dallas: SIL International, 2001. p. 212.

1. One of the rules of Khmer/Sanskrit/Pali is that at most one dependent vowel is allowed in a single consonantal cluster. By defining NIKAHIT as a vowel, we would encounter the awkward situation where combinations of NIKAHIT with ‘other’ vowels (and in the case of Sanskrit that could be any vowel) in a single cluster violates that grammatical rule.

2. In order to facilitate proper intra-cluster ordering of Khmer script characters, fonts and keyboard drivers should enforce a sequence: one consonant or independent vowel, zero or one register shifter, zero one or two COENG + consonant/independent vowel pairs, zero or one dependent vowel, zero or one sign, and zero or one vowel-like sign (this has not been standardised yet...but something like this would be the end result). It would damage categorisation to have NIKAHIT function in two different categories (i.e., dependent vowel and vowel-like sign).

In conclusion, NIKAHIT should remain in Khmer Unicode as a separate sign. Ligatures such as the proposed KHMER VOWEL SIGN OM/KHMER VOWEL SIGN AM should not be added. Nevertheless, the combinations of KHMER SIGN NIKAHIT, KHMER SIGN REAHMUK or KHMER SIGN YUUKALEAPINTU with inherent or dependent vowels may be treated as unique vowels in future Khmer sorting algorithms. To explicitly encode these two ligatures, however, would unnecessarily complicate data entry, code tables, and display algorithms. Encoding the proposed two characters would be inconsistent with the handling of the KHMER SIGN REAHMUK or KHMER SIGN YUUKALEAPINTU which function similarly in combination with dependent vowels.