# Khmer Sorting Questions

(1) It appears to me that Khmer sorts/orders by syllables in the following decreasing priority:

- Independent vowel or base consonant (which sort either as (a) decomposed consonant 179A ( រ ), 179B ( ល ) , or 17A2 ( អ ) plus a vowel or (b) [as suggested by Robert Headley] as separate entity after 17A2 ( អ ) in the same order in which they appear in the Khmer Unicode encoding: 17A5 ( ឥ ), 17A6 ( ឦ ), 17A7 ( ឧ ), 17AA ( ឪ ), 17AB ( ឫ ), 17AC ( ឬ ), 17AD ( ឭ ), 17AE ( ឮ ), 17AF ( ឯ ), 17B0 ( ឰ ), 17B1 ( ឱ ). There are three independent vowels not accounted for in this list which presumeably slot in at the same locations as the Unicode encoding ( ឲឩឳ ). Presumeably the two other Indic independent vowels ( ឣឤ ) are merged within 17A2 ( អ ) .
- First Subscript consonant (or independent vowel, composed of 17D2 + a base character)
- Second subscript consonant (or independent vowel...not two subscript independent consonants in a row, composed of 17D2 + a base character)
- Vowel entered with one keystroke (it may be composed of two glyphs [like 17C4 ( ◌ោ ) which has a glyph before and a glyph after] or it may an inherent [invisible] (no entry [which is presumed to be 17B5], 17B4, 17B5) entered by zero or one keystroke)
- First sign (if this is a 17C6 or 17C7 it will metamorph for sorting purposes into a vowel, transforming the preceding vowel [if one] into another vowel or replacing it if the vowel was originally of the inherent variety)
- Second sign
- Third Sign (this occurs only in the case the first sign is a 17C6 or 17C7, so for all practical purposes there are only two signs involved in ordering)

(2) In the linguists committee consultation [http://www.bauhahnm.clara.net/Page1.gif Page2.gift Page3.gif Page4.gif] Unicode characters 17C6 and 17C7 were defined as signs so as to avoid ambiguity when keyboarding those characters in the many other vowel-like combinations in which they occur. That was a wise decision (avoiding the addition of at least 14 hard to remember points in typing the keyboard in the future). Chhuan Nath gives those two characters names as signs [p. 532 and p. 1111]...even though he then has them function as vowels. In the Chhuan Nath dictionary there are four vowels which sort in the order 17BB+17C6 ( ◌ុំ ), 17C6 ( ◌ំ ), 17B6+17C6 ( ◌ាំ ), 17C7 ( ◌ះ ) at the END of the dependent Khmer vowels elaborated in Unicode(priority 4)...and these are commonly recited by schoolchildren in a listing of vowels (although all the other vowels created by conjunction with 17C7 ( ះ ) are not so recited). If we simply plugged the Unicode characters 17C6 ( ◌ំ ) and 17C7 ( ◌ះ ) into the signs slot (priority 5) as they are named, they would all sort near the BEGINNING of other vowels (17C6 [ ◌ំ ] before 17B6 [ ◌ា ], etc). We could maintain the present Chhuon Nath ordering by an algorithm which would deterministically merge and transform those signs to function as vowels, giving them increased ordering priority not only in those four cases...but also in the cases of the other vowels created by the combination of 'simple' dependent vowels and those signs ( ◌ាះ ◌ិះ ◌ីះ ◌ឹះ ◌ឺះ ◌ុះ ◌ូះ ◌ើះ ◌ៀះ ◌េះ ◌ែះ [◌េ៉ះ] ◌ោះ ). With the probable addition of [◌េ៉ះ], the following combinations

do not appear to exist in Khmer: ( ◌ួះ ◌ឿះ ◌ោះ ◌ុំះ ◌ាំះ ).

(3) Why are short inherent and inherent vowel shortened  as a result of BANTOC (17CB ◌់) or YUUKALEAPINTU (17C8 ◌ៈ ) not sorted in a similar way? See arabic numbered pages 119, 148, 1590, 1583

(4) Sanskrit/Pali characters have been encoded as 17A3 ( ឣ ) and 17A4 ( ឤ ). See the scans of equivalences in KhmerSanskrit.pdf (1 MB) or the smaller KhmerSanskrit1.jpg KhmerSanskrit2.jpg. A standard ordering routine may also (separately?) have to sort Sanskrit/Pali in Khmer characters. Is the order of those languages in Khmer script different than the ordering of Khmer script by itself? There was an Achar at the Royal Palace (who was also on our linguists committee) who was compiling such a dictionary. What order is it in?

(5) Does 17A1 ( ឡ ) NEVER appear in subscript form in Khmer/Pali/Sanskrit?

Apparently 17A1 only rarely has a subscript (there are several occurances with subscript 17A0 ( ឡ្ហ )).

(6) Do independent vowels 17A5-17B3 never have subscripts attached to them in Khmer/Sanskrit/Pali (with the exceptions of 17B1+17D2+1799  [ ឱ្យ ] or 17B2+17D2+1799 [ ឲ្យ ] in Khmer)? What are the exceptions?

(7) Are 179D-179F (ឝឞស) in the right order for sorting (this does seem to be the order they should be encoded at according to KhmerSankrit1.jpg/KhmerSanskrit1.jpg)? This may be an academic question neither of the Indic transliteration characters here are even represented as sortable entries in Chhuan Nath. On the other hand they might figure significantly in a Sanskrit ordered list! Or should they be merged at the primary level (such as many idependent vowels are merged with 17A2) and distinguished in another pass?

(8) Should numbers or any other signs sort before/after base independent vowels/base consonants in indexes? This is often the case in technical indexes at the back of English books.

(9) In Chhuan Nath's dictionary the doubling sign ( ៗ ) and the long hyphen seem to infer a following consonant so sort after identical (p.1083, p.1728) and sometimes not so identical (p.1590) words. Some entries also have question marks or exclamation points (p. 1886-1887), should they be treated specially?

(10) In my medical dictionary I gave an even weaker weight to signs in the sort than above (shifting the sign to sort in the next syllable). It appears that some signs act that way (17C9 [] and 17CA []), but others (such as BANTOC) appear much stronger (p.148). In newer technology it is possible to give weaker weights in a second pass.We need some guidance on how signs affect sorting of syllables and words.

(11) I have been asked to indicate how a computer algorithm could distinguish the beginning or end of a Khmer syllable. I have given these rules: (a) It appears to be an independent vowel not preceeded by 17D2 (COENG) is a syllable to itself (with the only exceptions of 17B1+17D2+1799 or 17B2+17D2+1799 as indicated in (6) above), (b) Any consonant not preceeded by 17D2 starts a syllable (hence if another syllable starts, the previous syllable is closed), (c) Any character other than 1780-17D2 before a consonant or independent vowel marks the beginning of a syllable if it is followed by a consonant or independent vowel,  (d) Any character other than 1780-17D2 marks the end of a syllable if preceded by any character between 1780-17D2, (e) the first Khmer consonant or independent vowel in a document indicates the beginning of a syllable, and (f) the last Khmer character between 1780-17D1 in a document

marks the end of a syllable. Rules similar to (b), (c), (d), and (e) mark word beginnings and endings. Normally words should be divided by ZERO WIDTH SPACE \u2008 and phrases should be separated by SPACE \U0020. I tried to introduce into Unicode a conditional space to be used to facilitate justification where a true phrase break was not warranted...but the suggestion was ignored after I failed to produce adequate documentation;-) Myanmar would also like such a feature for use with their script. How should hyphenated words be treated in sorting? How should hyphenated words be treated in terms of end of word? Khmer does not appear to use hyphens at the ends of lines even if words are broken, right?

(12) It seems to me that Chhuan Nath's dictionary sorts the following two inconsistent with the principle of decomposition: 17AA and 17A9 ( ឪឩ ). It seems to me that the latter should sort before the former, since the former has an additional implied 179C ( វ although that is not strong enough to decompose the 179C ( វ ) into a subscript vowel position). Similarly the obsolete 17A8 ( ឨ ) has an implied consonant 1780 ( ក p. 1852 and p. 1877).

(13) It appears that Sanskrit has three forms of 'L', whereas Khmer has only two ( លហ ). Is there some convention whereby the third is represented?

(14) Sanskrit and inherent/A vowels.
Sanskrit has:
[1] a written independent short A vowel (to match its
[2] implicit short A vowel) and a
[3] long AA independent vowel (to match its
[4] explicit AA dependent vowel). But it does not seem to have the equivalent of a
[5] long inherent A vowel written explicitly as an independentvowel (neither does it have a
[6] long inherent A vowel such as Khmer has).

Hence the equivalences:

| Sanskrit | Khmer Unicode | Description of character |
|---|---|---|
| [1] | 17A3 ( ឣ ) | written independent short A vowel |
| [2] | 17B4 ( ឴ ) | implicit/unwritten/inherent short A vowel |
| [3] | 17A4 ( ឤ ) | long AA independent vowel |
| [4] | 17B6 ( ◌ា ) | explicit long AA dependent vowel |
| [5] (no) | (no) | independent long A equivalent to inherent long A vowel |
| [6] (no) | 17B5 ( ឵) | inherent long A vowel |

Please feel free to challenge and question me on these various points. A thorough and open discussion is much more likely to yield a valid result than a declaration by any one person.

Maurice Bauhahn, 2 Meadow Way; Dorney Reach; MAIDENHEAD SL6 0DS;
22U.K. Tel: +44(0)1932 626068; Email: bauhahnm@clara.net
Khmer Sorting Questions version 0.4 beta 3 February 2001