Khmer Sorting Analysis

Recent changes are in Red.

Sorting scheme for Khmer

Note that page references in this document are typically to Chhuan Nath's Khmer-Khmer Dictionary, Japanese Reprint Edition with arabic numbers at the bottom of the page.

Priority 1:

(Should Khmer numbers and signs precede the alphabet? Should 17A3/17A4 precede the other letters of the alphabet?)

[1780-1793] The first 20 (of 33) Khmer consonants in the order they are encoded in Unicode: កខគឃងចឆជឈញដឋឌឍណតថទធន

[1794] The next one (of 33) Khmer consonants in the order they are encoded in Unicode: ប It would probably be best to merge this and the next two entries under one heading, words with signs would list immediately after words with identical spelling without said signs. Is that acceptable?

[1794+17C9] A variant of the 21st Khmer consonant with 'p' pronunciation comes next (this is evident when marked as: ប៉ , however, there are hundreds of words whose only distinction from a simple ប is their derivation)

[1794+17CA] A variant of the 21st Khmer consonant comes next (happily this is always marked as: ប៊)

[1795-1799] An additional 5 (of 33) Khmer consonants in the order they are encoded in Unicode: ផពភមយ

[179A] An additional 1 (of 33) Khmer consonants in the order it is encoded in Unicode: រ It would probably be best to merge this and the next two entries under one heading (i.e., including ROBAT and the two independent vowels decomposed into រ and the appropriate dependent vowel). Is that acceptable?

[17CC] The ROBAT sign is (inconsistently in the Chhuan Nath dictionary p. 465, 506, 538, 609, 750-1, 768, 1322, 1339-1340, 1633) treated for ordering purposes as an independent syllable. Should this be entered in phonetic order (as everything else is; I believe that would be appropriate)? What is its writing order when entered by a learned monk? It seems to fill the roll of a superscript consonant and is not written stand-alone. If it is sorted as indicated here and not entered in phonetic order, there will have to be some mechanism to reorder it in the ordering algorithm.

[17AB-17AC] These two independent vowels [ឫឬ] are treated as consonants following 179A as they share a consonantal sound of 'r'

[179B] The next one (of 33) Khmer consonant: ល Should this and the following section be merged with decomposition of the following in 179B plus the appropriate vowel?

[17AD-17AE] These two independent vowels [ឭឮ] are treated as consonants following 179B as they share a consonantal sound of 'l'

[179C] The next one (of 33) Khmer consonant: រ

[17AB-17AC] These two transliteration consonants [ឝឞ] are treated as consonants following 179C. They resemble the following Khmer consonant 179F as they share a sound 's'. (Q: Are these two in the right order for sorting? Should they be integrated within the Khmer 17DC for ordering purposes? None seem to be sorted in the Chhuon Nate dictionary. Could we have examples of the characters they transliterate and the name of the script that character comes from? Have the glyphs and names been switched in Unicode?)

[179F-17A0] The next 2 (of 33) Khmer consonants: សហ

[17A1] The next 1 (of 33) Khmer consonants (this is separated because it is not available in a subscript form): ឡ

[17A2, 17A3-17AA, 17AF-17B3] These characters are merged under one consonant (17A2) by means of decomposition into a glottal stop and a dependent vowel. For there to be a deterministic system this decomposition must be standardised. The resulting system (hopefully) will also sort transliterated Sanskrit/Pali text.

អ	អ	17A2
ឣ	ឣ	17A3 -> 17A2 (?)[1]

[1] There does not appear to be a strong differentiation between short initial inherent vowel words (presumeably 17A3) and long inherent vowel words (presumeably 17A2) in the final section of the Chhuan Nate Khmer dictionary. There is some controversy over the significance of 17A3 and 17A4 in Unicode. The linguist committee in Phnom Penh felt that there needed to be a distinction between the final Khmer consonant 17A2

ឤ	អ+ា	17A4->17A2 + 17B6 (?)
ឥ	អ+ិ	17A5->17A2 + 17B7
ឦ	អ+ី	17A6->17A2 + 17B8
ឧ	អ+ុ	17A7->17A2 + 17BB[2]
ឨ	អ+ុ (+ ក)	17A8->17A2 + 17BB (+ 1780)[3]
ឩ	អ+ូ	17A9->17A2 + 17BC
ឪ	អ+ូ (+ វ)	17AA->17A2 + 17BC (+ 179C)[4]
ឯ	អ+ ែ	17AF->17A2 + 17C2[5]
ឰ	អ+ ៃ	17B0->17A2 + 17C3
ឱ	អ+ ោ	17B1->17A2 + 17C4
ឲ	អ+ ោ	17B2->17A2 + 17C4[6]
ឳ	អ+ ៅ	17B3->17A2 + 17C5

Priority 2: First subscript should include all the characters in Priority 1 with the (possible) exception of a subscript form of ឡ which reportedly does not exist. However for sorting and display purposes it is assumed that any character in the range 1780-17B3 could be a subscript. On the other hand

and the two independent Sanscrit vowels 17A3-17A4. It would be good to clarify this issue if the particular Pali/Sancrit characters these are to represent could be shown.

[2] There are good examples of the equality of 17A2 and the first part of the decomposed independent vowel on pages 1808-1850 (arabic) of the Japanese reprint of Chhuan Nath's dictionary.

[3] The final Khmer consonant sound does not affect the ordering of this extremely rare and obsolete independent vowel. There will be some need of differentiating 17A7 and 17A8, but only at a higher level of sorting. This is referenced at the top of p. 1852 and p. 1877 of Chhuan Nath's dictionary.

[4] The final consonant 179C does not figure in the sorting order, and is presented only for an understanding of the roots of the character. By this analysis there would seem to be an inconsistency on page 1851-1856, particularly with ឪ ... ឪជំ ... ឪឡ ... អូ! ... អ៊ូ ... ឩក If the Chhuan Nath precedent were followed in this case it would seem to contradict the useage of decomposition for the other independent vowels that seem to separate into 17A2 + x.

[5] Note on p. 1860 the independent vowel in Chhuan Nath's dictionary seems to have a secondary priority over the decomposition: ឯ អែ

[6] There are only two words which require the use of this character, the very common ឲ្យ and the very rare .

only a subset of independent vowels are presently known to be subscripts (in addition to the consonant អ): ឫឪឯ (ហ្ឫទ័យ បង្ឪផន ផ្ឯធ)

Priority 3: Theoretically any of the characters under Priority 2 may also sort in the same orders under Priority 3. On the other hand in the Khmer language only about 9 are documented) ្រ៉ ៏ ្រ ្ហ ្រណា ៍ ឫ

Priority 4: Vowel 18 (Unicode: A committee of Khmer linguists voted to move three characters [17C6-17C8] from independent and combining forms of vowel to instead be signs as indicated in the Khmer Unicode section, reducing the number of dependent vowels that would need to be keyboarded.
The vowel/sign combinations which are known to exist using these are as follows:

□឵	17B5	Short inherent p. 1583
□឴	17B4	Long inherent
□ា	17B6	
□ាះ	17B6+17C7	p. 982, 1786, 1793
□ិ	17B7	
□ិះ	17B7+17C7	p. 132, 1237, 1549
□ី	17B8	
□ីះ	17B7+17C7	p. 64, 251
□ឹ	17B9	
□ឹះ	17B9+17C7	p. 760, 743-4, 1239, 1463
□ឺ	17BA	
□ឺះ	17BA+17C7	p. 246, 458, 597, 1887, 1808
□ុ	17BB	
□ុះ	17BB+17C7	p. 224, 542-3, 812, 1451, 1513, 1554
□ូ	17BC	
□ូះ	17BC+17C7	p. 1887
□ួ	17BD	
□ួះ	17BD+17C7	(Invalid? Not in Chhuan Nath dictionary)

□ើ	17BE	
□ើះ	17BE+17C7	p. 743-4, 895, 1878-9
□ឿ	17BF	
□ឿះ	17BF+17C7	(Invalid? Not in Chhuan Nath dictionary)
□ៀ	17C0	
□ៀះ	17C0+17C7	p. 748, 1242
□េ	17C1	
□េះ	17C1+17C7	p. 68, 215, 264, 689, 748 (but p. 1061)
□ែ	17C2	
□ែះ	17C2+17C7	p. 74, 142, 709, 761, 1475
□ៃ	17C3	
□ៃះ	17C3+17C7	(Valid? No example)
□ោ	17C4	
□ោះ	17C4+17C7	p. 76, 134-5, 142, 187
□ៅ	17C5	
□ៅះ	17C5+17C7	(Invalid? Not in Chhuan Nath dictionary)
□ុំ	17BB+17C6	
□ុំះ	17BB+17C6+17C7	(Invalid? Not in Chhuan Nath dictionary)
□ំ	17C6	
□ាំ	17B6+17C6	
□ាំះ	17B6+17C6+17C7	(Invalid? Not in Chhuan Nath dictionary)
□ះ	17C7	

Priority 5: Signs

□៉	17C9	p. 195, 626 (in conjunction with 1794 higher priority?), 1178

□៊	17CA	p. 715 (in conjunction with 1794 higher priority?), 1538-9, 1534-5
□៎	17CE	p. 252, 542-3
!	(exclamation)	p. 1558
□ៈ	17C8	p. 413, 843, 1178, 1492, 1562, 1590, but lower priority to hyphen p. 1392-3!
□់	17CB	p. 119, 133, 148 (higher priority?), 177, 1178, 1544 (?)
-	(hyphen)	p. 1254, but why p. 1538-9
□័	17D0	p. 119, 483, 681, 839, 1254
□៍	17CD	
□៏	17CF	
□៑	17D1	
_	(long hyphen)	p. 504, 1590, 1728, 1392-3
ៗ	17D7	p. 252, 860

Priority 6: Signs as above, relatively rare ណា៎ៈ អ៊ូ៎យ ហ៊ែៈ រ៊ុឹះ

Test collation series

ក	1	\u1780[7]	Single consonant
ក៏	2	\u1780\u17CF	Single consonant and sign
កក	3	\u1780\u1780	Consonant and next base consonant
កក់	4	\u1780\u1780 \u17CB	Consonant and next base consonant and sign

[7] When sorting ignore all spaces inserted into this column; they are purely for presentation/word-wrap purposes.

កករ	5	\u1780\u1780 \u179A	Could also be expressed with inherent vowels encoded \u1780\u17A4 \u1780\u17A4 \u179A (final consonant lacks vowel)
កករ	5	\u1780\u17A4 \u1780\u17A4 \u179A	Identical to previous
កកាត	6	\u1780\u1780 \u17B6\u178F	Vowel on second base resets cycling of third consonant
កកាយ	7	\u1780\u1780 \u17B6\u1799	Third base consonant changes
កកេះ	8	\u1780\u1780 \u17C1\u17C7	Vowel on second base resets cycling, starting with no third base
កកែកករ	9	\u1780\u1780 \u17C2\u1780 \u1780\u179A	ditto (presence of consonant in third base position follows absence of third base consonant)
កកែប	10	\u1780\u1780 \u17C2\u1794	Third base consonant cycle
កកោះ	11	\u1780\u1780 \u17C4\u17C7	Continuing to cycle through vowels on second base consonant
កក្រើក	12	\u1780\u1780 \u17D2\u179A \u17BE\u1780	Start cycling through subscript consonant on second base (reset cycling of vowel on second base)
កក្អាក	13	\u1780\u1780 \u17D2\u17A2 \u17B6\u1780	Continue cycling through subscript consonant on second base (reset cycling of vowel on second base)

ក឵ក្អាក	13	\u1780\u17B5 \u1780\u17D2 \u17A2\u17B6 \u1780	Identical to above (no implicit vowel when there is an explicit dependent vowel)
ខៅតាក	14	\u1781\u17C5 \u178F\u17B6 \u1780	Next consonant; cycling through vowel on first base
ខំ	15	\u1781\u17C6	Cycling through sign turned to vowel on first base
ខាំ	16	\u1781\u17B6 \u17C6	cycling through composed vowel on first base
ខាំង	17	\u1781\u17B6 \u17C6\u1784	Second base
ខះ	18	\u1781\u17C7	Cycling through sign turned to vowel on first base
ឃ្មោះ	19	\u178E\u17D2 \u1798\u17C4 \u17C7	
ឃ្មុំ	20	\u178E\u17D2 \u1798\u17BB \u17C6	Composed vowel starts with subscript part first, then superscript.
ងោង	21	\u1784\u17C4 \u1784	
ងោ៉ង	22	\u1784\u17C4 \u17C9\u1784	Word with sign follows word without sign
ឆា	23	\u1786\u17B6	
ឆា៎	24	\u1786\u17B6 \u17CE	Sign follows vowel in entry order
ឆាៗ	25	\u1786\u17B6 \u17D7	Doubling sign indicates a consonant will follow (but weights as a sign)

For corrections and suggestions please contact:
Maurice Bauhahn, 2 Meadow Way; Dorney Reach; MAIDENHEAD SL6 0DS;
U.K. Tel: +44(0)1628 626068; Email: bauhahnm@clara.net
3 February 2001 version 0.4beta