ISO/IEC JTC1/SC2/WG2 N1578
Date: 1997-04-06

Title: Cham encoding discussion
Source: Michael Everson, Everson Gunn Teoranta (IE)
Status: Expert contribution
Action: For consideration by WG2

Cham has some features unique to Brahmic scripts, particularly as regards the vowel signs and initial vowels. Following are the initial vowels as they appear in the dictionary *Từ Điển Chăm-Việt*.

ꨀ a, ꨀꨩ ā, ꨀꨪ i, ꨀꨭ u, ꨀꨮ ơ, ꨯꨀ o, ꨰꨀ ai, ꨯꨀꨱ au, ꨀꨲ ư, ꨀꨳ ya, ꨀꨳꨩ yā, ꨀꨳꨮ yơ, ꨯꨀꨳꨩ ye, ꨯꨀꨳꨱ yau, ꨀꨶ wa, ꨀꨶꨪ wi, ꨀꨶꨮ wơ; ꨁ i, ꨁꨩ ī; ꨂ u; ꨃ e/ē; ꨄ ai, ꨰꨄ ai, ꨯꨄꨩ e; ꨅ o, ꨅꨩ ō.

Firstly it is interesting that ꨀꨪ and ꨁ both mean *i*, that ꨯꨀ and ꨅ both mean *o*, and that ꨀꨭ and ꨂ both mean *u*. In this case it is certain that the independent vowel must be coded independently -- the words are sorted separately in the dictionary, and the independent sign has a unique form. A number of the vowels given as independent vowels in the French source that Hugh Ross provided gave forms which do *not* occur initially in the dictionary (such as ꨂꨩ *ū*). The problem which this causes is in terms of inputting. It will be much simpler to encode only the vowel signs which do not have unique independent forms (that is, those that are not attested as ꨀ plus a vowel sign), as this may well be the native perception (which would not be the case could *i*, *o* and *u* only be coded in way). Independent vowels can occur word-internally, as in ꨚꨀꨩꨘ *paan* and ꨚꨃꨤ *pael*. It is possible that there is a glottal onset *pa'an* and *pa'el* in these words. In that case ꨀ is really a consonant, which makes the treatment different than in other Brahmic scripts. The Cham dictionary gives a phonetic inventory in which does include a glottal stop (p. XXXIII).

The vowels ꨀꨳ *ya*, ꨀꨳꨩ *yā*, ꨀꨳꨮ *yơ*, ꨯꨀꨳꨩ *ye*, ꨯꨀꨳꨱ *yau*, ꨀꨶ *wa*, ꨀꨶꨪ *wi*, ꨀꨶꨮ *wơ* are not really vowels; they have a conjoined *y* and *w*, just as other consonants do: ꨀꨳꨩ *yā*, ꨝꨳꨩ *byā*, ꨀꨶ *wa*, ꨝꨶ *bwa*. Again, ꨀ is functioning as a consonant. Note the near-minimal pairs ꨓꨀꨳꨀꨱ *tayak* (*ta'yak*?) and ꨓꨢꨨ *tayah*.

Both ꨣꨮꨨ *rơh* and ꨤꨮꨨ *lơh* can be coded by sequences, and I could not find any evidence of Sanskrit loanwords containing *ṛ*, *ḷ*, *ṝ*, or *ḹ* in the dictionary. No vowel signs for these are provided even in the French source; I do not believe that these should be coded either as independent vowels or as vowel signs.

A number of consonants have special final forms which should not be coded separately. Three finals, however, operate in the same way as in other Brahmic scripts and accordingly are given in positions 1901, 1902, and 1903 here.

I can find no evidence in the Cham dictionary for the NASAL RHYMER character (*takai dak*) of N1126. Combinations of the vowel signs with the three characters at 1901-1903 should be realized as combinations. Note the near-homoglyphs ꨆꨭꩃ *kung* and ꨆꨮꨭ *kơu* (coded 1915 + 1941 + 1901 and 1915 + 1970 + 1941 respectively).

Virama or halant is used to create conjuncts, which occur internally and finally: ꨕꩀꨦꨗꨭꩀ *dakssanưk* (1926 + 1915 + 194D + 1936 + 1928 + 1915 + 194D); ꨆꨭꩍꨵꨯꨱꩌ *kuhlaum* (1919 + 1941 + 1939 + 194D + 1932 + 194C + 1902) where 1932 ꨤ *la* has its conjoined form "ꨵ" and not its final form since it is *preceded* by virama. Note that rendering can be complex in such instances: ꨆꨴꨰꩀ *kraik* is 1915 + 194D + 1930 + 1948 + 1915 + 194D, but only the second ꨆ has its final form ꩀ.

Attached is a revised code table and names list for Cham. In the code table some of the consonant conjunct glyphs appear in the grey boxes; these can be ignored but their presence provides information as to how Cham could be encoded in an 8-bit system with the positions used for font representation.