

ISO/IEC JTC 1/SC 2/WG 2 Universal Multiple-Octet Coded Character Set (UCS)
--

DOC TYPE: National Body Contribution  
TITLE: Cambodian official objection to the existing Khmer block in UCS  
SOURCE: Committee for Standardization of Khmer Characters in Computers\*  
PROJECT: JTC 1.02.18 – ISO/IEC 10646  
STATUS: For discussion at the 41<sup>st</sup> WG2 Meeting in Singapore  
ACTION ID: ACT  
DISTRIBUTION: ISO/IEC JTC 1/SC 2/WG 2  
MEDIUM: Electronic  
NO. OF PAGES: 4

\* Standard Organization accredited by Industrial Standards Bureau of Cambodia (ISC)

## **Cambodian official objection to the existing Khmer block in UCS**

We are pleased to have this opportunity to present the official views of Cambodia to the relevant Working Group and Sub-Committee of JTC1.

The Committee for Standardization of Khmer Characters in Computers seeks a rescission of the Khmer Code Table as published in ISO/IEC 10646-1 2<sup>nd</sup> edition, 2000, and its complete replacement by the character set being prepared as a Cambodian national standard.

We base our request on the following grounds:

- 1) no appropriate official Cambodian representative participated in any of the discussions leading to the adoption of the current code table by ISO/IEC, and this code table has never been officially endorsed within Cambodia;
- 2) the present code table contains major deficiencies as outlined in Appendix, of which the most significant is the decision not to allocate individual code points for the subscript consonants, but instead to follow the “virama model”, presenting severe inconsistency in the light of Khmer orthography and causing unnecessary inefficiency for Khmer character processing, which would not be faced by the replacement table we will submit later. We are also against the further attempt to force the virama model proposed in N2359.

We realize that this is an unusual request in the light of the stated position of ISO/IEC 10646-1 that “the names and allocation of the characters ... will remain unchanged” (p.9) . However, it is our contention that as due process was not followed in the adoption of the Khmer code table it is entirely within the norms of international standard-setting for the issue to be revisited now that a request has been formally submitted by the appropriate national body.

Furthermore, we understand that a complete replacement of a published code table within ISO/IEC 10646-1 is not without precedent.

We look forward to the opportunity for the Cambodian delegation to discuss this issue at the WG2 and SC2 meetings in Singapore, 15-19 October 2001.

Dr Pan Sorasak,

Under Secretary of State of the Council of Ministers, Royal Government of Cambodia  
Deputy Chairman of the Committee for Standardization of Khmer Characters in Computers  
Phnom Penh, 8 October 2001

## Appendix

### Comments on the contents of the existing Khmer character code table in ISO/IEC 10646-1:2000 and the Unicode Standard 3.1

- 1) The independent vowels (SRAK PENH TUA) 𑄛 (17A3) and 𑄛𑄌 (17A4) are included in the character table, but such characters do not actually exist in Khmer script. According to the Unicode Standard, they are used for transliteration of Pali/Sanskrit words. However, it is not an enough reason to include them, because they can be represented by the consonant 𑄛 (17A2), and by the consonant 𑄛 (17A2) + the vowel 𑄌 (17B6) respectively, if necessary.
- 2) 𑄛𑄌 (17A8) is included in the character table, but the Chuon Nath's dictionary ("Dictionnaire Cambodgien", 5<sup>e</sup> édition, Institut Bouddhique, 1967-1968) specifically says that it is a ligature of 𑄛 (17A7) + 𑄌 (1780).
- 3) The independent vowel 𑄌 (17B2) is included in the character table, but it is a variant of 𑄌 (17B1).
- 4) Two inherent vowels (17B4) and (17B5) are included in the character table. In fact, such characters have never been used in Khmer and do not actually exist.
- 5) The dependent vowels (SRAK NISSAD) 𑄛̣ and 𑄛̣̣ are regarded not as single vowel signs but as combinations of NIKAHIT 𑄛̣ (17C6) and a vowel sign in ISO/IEC 10646-1 and the Unicode Standard. This is against the stance of the Chuon Nath's dictionary.
- 6) Subscript consonants (COENG PYUNHCANA) are not assigned independent code points, but are instead represented by a control character 𑄛̣ (17D2) plus the corresponding consonant from the character code table, based on the Indic (ISCII) "VIRAMA MODEL" (see "Khmer and Burmese Ad-Hoc Meeting Report", ISO/IEC JTC1/SC2/WG2 N1729, 1998-03-18).

Behind this determination, there seems to be the idea that a subscript consonant is just a different glyph of its corresponding consonant. However there is more than that between them.

First, a consonant can constitute an independent syllable by itself, but a subscript

consonant cannot. In other words, if the former is a character in a narrow sense, the latter is a diacritic. Their relation is similar to that of an independent vowel (SRAK PENH TUA) and a dependent vowel (SRAK NISSAI) of the same pronunciation. As long as each of these vowels has its own code point, each of the two types of consonants should have an independent code point.

Second, it cannot be determined automatically in Khmer script whether a character code value for a consonant should be presented in a normal form or in a subscript form. This is not the case with Arabic script, where the presentation form of a character is automatically determined by its position and situation in a word. As long as the two have to be distinguished at the character code level, they are different characters. Then different characters should have different code points according to one of the principles of ISO/IEC 10646-1 and the Unicode Standard.

- 7) The BATHAMASAT (17D3), (in Khmer: PATHAMASAT) is presumably included to represent the first August of leap year in lunar calendar, but we cannot find any code point assigned for the second August (TUTEYASAT).
- 8) An independent code point is assigned to ័័័័ (17D8), an abbreviation for the Khmer word meaning “et cetera”. Like “etc.” expressed by “e+t+c+.” in English, it can be written as a combination of “័+័+័+័” (17D4+179B+17D4), so there is no need for a special code point to represent it. Furthermore, there are other ways of abbreviating this word, and it would be inconsistent to include only one of them in the character code table.
- 9) The same character or the same combination of characters often has more than one presentation form in Khmer (for example: ឌ and ឌ̣, ឍ and ឍ̣, ណ and ណ̣, ត and ត̣, etc.). However no consistent way to deal with them can be found in the existing table.
- 10) As a consequence of some of the above decisions, the normal sequence of characters as used in the Chuon Nath’s dictionary has been violated, and this in turn presents unnecessary difficulties for sorting algorithms.