ISO/IEC JTC 1/SC 2/WG 2 **N2406**

DATE: 2001-11-04

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

DOC TYPE:	National Body Contribution
TITLE:	Response to WG2 Document N2385, 2001-10-11
SOURCE:	Cambodian Committee for Standardization of Khmer Characters in Computers*
STATUS:	As requested during WG2 meeting in Singapore 2001-10-15/18
DISTRIBUTION:	ISO/IEC JTC 1/SC 2/ WG 2 and UTC (same as N2385)
MEDIUM:	Electronic
NO. OF PAGES:	11

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

Response to WG2 Document N2385, 2001-10-11

This document is being submitted in response to a request from the Convener of WG2 to the Cambodian delegation to document in writing their position as discussed in Ad hoc meetings during the 41st meeting of WG2 held in Singapore, 15-18 September 2001. The Cambodian delegation was pleased to have its first opportunity to participate in these discussions on the Khmer block in UCS, and to pursue the issue further here, providing more detailed written clarification of its stance presented in N2380R, and the objections to N2380 put forward by Maurice Bauhahn and Michael Everson in N2385.

Please note the official Cambodian procedural objection and request for rescission of the previously adopted Khmer code table advanced in N2380R and in the letters of 30 and 31 May 2001 (SC2 N 3571). Our comments here are restricted to responding to points made in N2385.

Background comments
We comment on the long introduction to N2385 only where such history is significant in regard to the scope of the challenge to develop an appropriate encoding system for the Khmer script and related tasks.

Lack of official Cambodian involvement in the standard-setting process
The authors of N2385 state that the existing Khmer block in UCS was finalized, using the virama model, after heated discussions. "Certainly the virama model was one of the issues of controversy. But if there has not been an appropriate official Cambodian representative, it has not been for lack of trying!" – This is a surprising claim, as the history presented in N2385 mentions no communication with any part of the Cambodian government since 1997.

The Industrial Standards Bureau of Cambodia (ISC), the Cambodian national body registered as a subscriber member of ISO since 1995, has never received any communication from JTC1 or ISO regarding the establishment of an international standard for the Cambodian national script.

To finalize a standard for a script without the involvement of the national body of the country in which the script is almost entirely used is completely against the spirit of Article 12.5 of the WTO-TBT Agreement effective since 1995, which determines that WTO members should facilitate the active and representative participation of all the relevant bodies, especially in developing countries, in international standardizing processes. The Genoa Plan of Action endorsed by the latest G8 summit clearly supports this spirit especially in language related standards. We are afraid that the approach followed with regard to Khmer character encoding harms not only the developing country involved, but also the credibility and authority of the standard itself and the international standard organization. Recent letters from the e-ASEAN Task Force to JTC1 etc. clearly support the Cambodian objection made in May 2001.

Several offers to establish contacts with the appropriate official bodies in Cambodia were offered from Cambodia, and an e-mail from Glenn Adams, at that time coordinating Khmer related work for UNICODE encoding, foresaw what would happen as a result of a hasty decision without such involvement:

> -----Original Message-----
> From: Glenn Adams [mailto:glenn@spyglass.com]
> Sent: Wednesday, 13 August 1997 01:25
> To: Lee Collins; Maurice J Bauhahn
> Cc: khmer@unicode.org; averyb@microsoft.com; mleisher@crl.nmsu.edu; Martin Duerst
> Subject: Re: Subscript Consonants in Khmer
>
> [snip]
>
> No matter how nice it would be, not everyone is in a position to adopt new (possibly better) designs. I'm trying to maintain a pragmatic position here which can accommodate a consensus position. I don't want to be in a situation like we were with Korean Hangul and now are in with ISCII-91 incompatibility simply because we rush into a design which appears to meet our current interests and predelictions but fails to obtain a consensus from the primary parties.
>
> Given the absence of any Cambodian authority in these discussions, I do not want to go ahead with the ISCII oriented approach. What would persuade me would be an official or semi-official position paper from Cambodian authorities stating their willingness/desire to go along with one or the other approach to subscript consonants.
>
> In the absence of any paper, we will have no recourse for our decisions if in the future Cambodia decides they don't like the way we designed their script encoding.

It is a pity that similar voices in WG2 were also ignored. The minutes of WG2 London meeting in 1998-09-21--25 (WG2 N1903) include the following:

> B.5 - Cambodia has a large number of subscript forms. The model used is to use the Cambodian Virama to create the subscript form.
> Mr. Takayuki Sato: Cambodian users should be consulted on these to check the model used and the coding used.
> [snip]

> Mr. Takayuki Sato: Based on what I hear from Mr. Maurice Bauhahn, let us go with the new proposal. We need a small note from the governmental organizations that the current proposal is acceptable encoding for them. Why is it so difficult to send some expert in Cambodia a note?
> Mr. Mike Ksar: WG 2 has the expertise - and usually we do not go to the government organization. We would like to add some explanation to the disposition of comments to address the Japanese comments and possibly change their ballot to positive.

Similar concerns can be found in other documents as follows:

-- Israel Votes YES, provided that the relevant(*) national body/bodies also approves this draft - otherwise our vote is ABSTAIN.
(*)relevant national body = the national body for which this character collection describes its national language.
(by the Israeli national body in 1998; SC2 N3160 / WG2 N1835)
-- The vote Approval is on condition of approval from the countries concerned.
(by the Swedish national body in 1998; SC2 N3160 / WG2 N1835)
-- This FDAM is not acceptable to Cambodian natives (by the Japanese national body in 1999; SC2 N3375 / WG2 N2137)
-- AMD-25 is not a proven character set by usage of native user (by a Japanese expert in 2000; WG2 N2164)

The role of the KPP
The description in N2385 of the role of the activities of the Khmer Philology Project is completely misleading. The Royal Academy of Cambodia has organized several conferences with some relevance to this subject, and the Khmer Philology Project was invited to participate in these activities of the Royal Academy of Cambodia. We fail to fully understand the intended insinuation, and we refuse to accept that the Royal Government of Cambodia "has recently been so heavily influenced, in this case, by such short term interests."

Here, we think it necessary to make the role of the Khmer Philology Project clear. Khmer Philology Project coordinated by the Asia Pacific Association of Japan (APA-KPP) is a voluntary, non-profit group Cambodian Japanese, as well as people in Cambodia. It tried to resolve the prevailing situation of improper and incompatible fonts, and developed both a coded character set and an implementation, the Intelligent Khmer Writing System (IKWS), based on it to prove its capacity in relation to input and rendering. Both these phases of their work were carried out in cooperation with Cambodian governmental and academic authorities and other organizations in Cambodia. As IKWS was welcomed by many local people, the KPP planned to distribute IKWS for free, but considering the fact that this would cause yet further incompatibility with a newly defined Khmer block for UCS/Unicode, they stopped and asked the Cambodian government to consider its position. The KPP view was expressed as follows: “Encoding is one thing, implementation is another. We think that our coded character set is better for Cambodian people than the newly defined Khmer block in UCS/Unicode, but if you decide to accept the latter, we can recode our implementation accordingly. If you cannot accept it, however, we will support your efforts to change it. We will follow your decision.” After several months, on 30 May 2001 the Cambodian government lodged an official objection to JTC1 etc. While it is indeed likely that this official Cambodian move might not have occurred without input from APA-KPP (given the fact that the government had received no notification of the JTC1 activities in defining a Khmer code table), the decision was made by the government itself.

Ignorance or overlooking of basic Unicode principles

Unfortunately, baseless assumptions that the work done under the auspices of the Royal Academy and supported by the National Body did not regard the fundamental principles for Unicode encoding are frequently made in N2385.

The authors of N2385 wrote it based on N2380, not on N2380R, and so their comments and criticisms were made without having seen the draft of the coded character set the proposed by the Cambodia National Body. That is the basis of their serious misunderstanding that Cambodia was proposing glyph-based encoding. We hope they now see clearly that the Cambodian proposal is firmly for character-based encoding, so that is not an issue between the two approaches. And, as we showed in N2380R and again show below, we think that, despite its proclaimed intent, the existing standard is actually rather inconsistent in avoiding glyph-based encoding.

Also important in this context is the authors’ own admission that their choice of the virama model “appears to fly in the face” of the standard Chuon Nath’s dictionary. Their assumption that an explicit encoding of subscripts would probably prevent the encoding of dictionaries such as that of Chuon Nath has been proved wrong by our test.

It is mentioned in N2385 that the standard Chuon Nath’s dictionary was previously re-typed (though no information is given in which medium) “partly with the aim of discovering exceptional characters and constructs that might be needed in computer implementations.” The same exercise has also been done again recently in Phnom Penh, using a prototype implementation of the character based model with subscripts, as proposed by the Cambodian National body (Committee for Standardization of Khmer Characters in Computers). This test carried out by Open Forum of Cambodia during 2001, using IKWS, has proved quite capable for keying correctly and easily all of the Chuon Nath headwords.

Responses to the “Responses to the issues raised in the Appendix of N2380” in N2385

(NB. Please note that the following are only comments made in response to N2385. Our original points as made in N2380R should also be referred to in order to see our full position on these issues)

1. Transliteration is intended to represent characters of one script by characters of another script. It is basically against this definition to add new characters for transliteration. Khmer script has no independent vowel characters corresponding to the independent vowel characters QAQ (ឣ) and QAA (ឤ) in Indian scripts that may be used for Pali/Sanskrit texts (eg. Devanagari, Sinhalese etc). This is a feature of Khmer script. Nevertheless, transliteration is quite possible by using the Khmer consonant characters QA (អ) and QA+AA (អ+ា)respectively. As the Indian scripts have no such consonant character, this does not cause any ambiguity. Also, Pali sorting can be done by a collation algorithm with language information, and it is unnecessary to add these characters only for sorting.

2. If we follow the Chuon Nath dictionary, we should not encode QUK (ឨ). When it was used before, it was interchangeable with QU KA (ឧក). The dictionary regards it as a ligature and does not include it among independent vowels. Encoding QUK (ឨ) independently is to resort to glyph-encoding.
   The dictionary surely says that the shapes of QUUV (ឪ) and QAU (ឳ) are based on QU+VO and QOO+VO respectively, but explicitly includes them among independent vowels. They cannot be regarded as ligatures, because they cannot be written as QU VO (ឧវ) or QOO VO (ឱវ) in any case.

3. QOO TYPE TWO (ឲ) has been gradually superseded historically by QOO (ឱ), and can be replaced by QOO (ឱ) without any change in pronunciation or meaning. Accordingly, QOO TYPE TWO (ឲ) should be regarded as a variant of QOO (ឱ) and encoded as QOO+Variation Selector. Encoding QOO TYPE TWO (ឲ) independently is to resort to glyph-encoding.

4. We think the addition of such artificial characters for sound is not needed for text-to-speech applications, because the processing unit for such phonetic software does not match the unit for writing. A sequence of normal written Khmer characters will be converted internally into a sequence of some phonetic signs that suits synthesizing voices anyway.
   As for the “verbal spelling and some dictionary orderings”, we would appreciate any concrete example where those characters are indispensable.

5. It is sure that there are several opinions on this issue among scholars. However, thinking them as single vowels is the stance of Chuon Nath's dictionary as well as of ordinary Cambodian people. It also thinks REAHMUK as a vowel as well as a sign and YUUKALEAPINTU as a sign. The Royal Academy of Cambodia has

declared that they follow the dictionary. And the Committee for Standardization of Khmer Characters in Computers (CSKCC) follows the decision.

The following are the responses to your points.
On your first point, it is also called Srak AM (◌ំ).
On your second point, interpretation based on the stance of Chuon Nath's dictionary mentioned above is quite possible.
On your third point, ambiguity can be avoided by depending only on the stance of Chuon Nath's dictionary.
On your fourth point, it is true that sorting is another problem. However, it is useless to make sorting algorithm unnecessarily complicated.
On your fifth point, the rendering issue is another problem. We have confirmed that it is not difficult even if we base it on OM (◌ុំ), AM (◌ំ), AAM (◌ាំ) as single vowels.

6. As for the virama model, detailed discussion on this is presented below, but it is useful to confirm here that the encoding issue is independent from the key typing issue.

7. Bauhahn and Everson agree on the necessity to represent TUTEYASAT.
A separate issue is whether these characters should be encoded by using combining characters. In the existing Khmer table in UCS, BATHAMASAT is the name of a combining character, which is supposed to be used with KHMER DIGIT EIGHT. In reality, however, BATHAMASAT is the name for a whole one character, not for the upper part of the character. That part itself is not a character in reality, and has no specific name. This means that the existing encoding is glyph-encoding. It is also the case with TUTEYASAT. We think PATHAMASAT (BATHAMASAT) and TUTEYASAT should be encoded as self-contained non-combining characters.

8. What we are saying is that you can produce the necessary glyph as a ligature of KHAN+LO+KHAN (។+ល+។) without defining a special character code value for it. Also, we are saying that other combinations of Khmer characters (ex. hyphen+LO+ hyphen (-ល-), KHAN+Srak E+BA+KHAN (។េប។), hyphen+Srak E+BA+hyphen (-េប-)) are used for similar but not exactly the same purposes, and that they should also be processed as ligatures of the original combinations of Khmer characters, not as variants of a single fictional character. We would appreciate it if those advocating this code point for Khmer could show us an example of a code point of a single coded character for "etc." in Latin script.

9. We still think that the separation of a character and a glyph is not consistent in the existing table based on the arguments above.
The ligatures of KA+VO, RO+VO, etc. can co-occur with non-ligatures in the same text. The difference between them is not of design but of glyph. Then the difference should be represented in the character code level using ZWJ etc.
The new sample glyph for your COENG sign, "plus sign below base character", may be still confusing, because a similar sign is used to denote "insertion point" in Khmer handwriting. However, we understand that any glyph will do because this is an artificial character.

We also reconfirm here that the encoding issue is independent from the key typing issue.

10. We are talking about creating "unnecessary difficulty". Sorting algorithms, if necessary, will become simpler if based on our proposed table.

Responses to the "Specific discussion of explicit subscript vs. virama encoding" in N2385

0. The essential points of our position are the following:
   (a) The virama model is groundless for Khmer script;
   (b) The virama model is inefficient for Khmer because it will increase normal text file size by around 20%. This is a clear demerit for end-users;
   (c) It will impose frequent application of an unnecessary step in the rendering phase (substituting COENG+consonant with a glyph of a subscript consonant). This is a demerit for implementers;
   (d) The rendering rules used for Devanagari cannot be used for Khmer even if its encoding were to be based on the virama model. The merit for implementers is not highly significant;
   (e) The only practical reason for the virama model seems to be to economize on the number of code points, as one of the authors of N2385 clearly said in the WG2 London meeting (WG2 N1903), but this is not a merit for end-users or for implementers.
   (f) In sum, the virama model has nothing that can justify deviation from the natural and efficient approach for Khmer: explicit encoding of subscripts.

We now move on to respond to the points as outlined by Bauhahn and Everson.

1. The authors of N2385 present the difficulty of determining the entire set of possible subscript characters in advance as a reason for adoption of the virama model. However, a character set has to be explicitly determined in a standard. Without it, each implementer might define it as they like. As the UCS is a coded character set, and a subscript consonant is a character (as we argue in N2380R and further below), the Khmer block should clearly define the set of possible subscript characters. If one finds a missing one, one can add it in the future.

   We are afraid that the authors of N2385 have not grasped the significant distinction in Khmer between main characters and subscripts. If LO (ល) and COENG LO (◌្ល) were the same character, as they assert, it would not be necessary to distinguish one from the other in a character code sequence. However this is not the case in reality. You cannot determine whether a character sequence SA LO should be presented as SA LO (សល) or SA COENG LO (ស្ល). This is a very important point, because these two presentation forms represent two different words with two different pronunciations. This means that the difference between them is not that of a glyph. They are two different characters. Therefore COENG LO (◌្ល) should be assigned an independent code point from LO (ល), and does not make sense to make COENG LO (◌្ល) a composed character of COENG+LO (◌្+ល) (according to the virama model). This is not only unnecessary, but also inefficient, because it will increase the string size markedly.

There is no linguistic reason to adopt the virama model for the Khmer script, in contrast to, for example, the Devanagari script. In the Devanagari script, halant (virama in UCS) is a sign to drop an inherent vowel sound of a consonant. There are three ways to represent the same phonetic consonant cluster.

Example: KKA
basic character sequence: KA+VIRAMA+KA
glyph sequence: (based on the explanation in Unicode3.0)
(1) One conjunct consonant of KKA (if available)
(2) Half letter of KA+KA [if (1) is not available, or for KA+VIRAMA+ZWJ+KA]
(3) KA+VIRAMA+KA [if both of (1) and (2) are unavailable, or for KA+VIRAMA+ZWNJ+KA]

Here the virama is an existent sign. Usually it is absorbed in some kind of ligature, but sometimes it is shown as a visible independent sign. It is meaningful to adopt the virama model for Devanagari.

In contrast to this, the Khmer script has no such sign. There is a similar sign called VIRIAM, but its usage is different from the VIRAMA. It is used only to denote the absence of the inherent vowel of a closing consonant in a phonetic (not orthographic) syllable. It cannot be used to denote the absence of the inherent vowel sound of the first consonant in a consonant cluster. Without using the VIRAMA, Khmer uses one or two subscript consonant(s) with a consonant to construct a consonant cluster.

KA VIRIAM KA does not make sense. Cambodians do not identify it with KA COENG KA. COENG KA cannot be thought of as a ligature of VIRIAM KA. That is why the existing table in the Khmer block in UCS had to add a fictional “COENG” sign other than VIRIAM to impose the virama model, and consequently decomposition is required of what actually is a single character into an artificial special character and another character. More strictly speaking, the designers of this table invented a new model different from the original virama model for Devanagari. It is clear that this decision had nothing to do with separation of character and glyph.

The Tibetan script is said to have a north-Indian origin, and, like Khmer, it uses subscript consonants as a way to construct consonant clusters. The editors of the Tibetan block in Unicode refused to adopt the virama model, and instead assigned independent code points for all the subscript consonants. The Khmer script, where the only way to construct a consonant cluster is using subscript consonants, should follow the same encoding model.

2. The Khmer independent vowel letter “I” is called “srak penh tua I” (ឥ), while the Khmer vowel sign “I” is called “srak I” (◌ិ). If the authors of N2385 make a control character according to the nomenclature, why did they not make another control character “penh tua”? We think the nomenclature is not the defining characteristic in the process of identification of separate characters.

3. No need to respond. This is based on the misunderstanding that the Cambodian National Body’s proposal is for glyph-based encoding.

4. No need to respond. The key typing issue is independent from the encoding issue. Proper input methods can provide various key strokes for any necessary character code sequence.

5. Indeed our position is to discard the virama model completely, and to adopt the explicit subscript model. We are not advocating ambiguity.

6. No need to respond. This is based on the misunderstanding that the Cambodian National Body’s proposal is for glyph-based encoding.

7. No need to respond. This is based on the misunderstanding that the Cambodian National Body’s proposal is a glyph-based encoding.

Additional comments

We would appreciate it if the following clarifications could be made by the authors of N2385.

1. Line 21 of page 4: Khmer word ឪ្យ . What does it stand for?
   (We guess that it may be the representation of ឱ្យ).

2. Line 17 of page 5: Khmer word បង្អូផន . What is the meaning of this word?
   (We guess that it may be បងប្អូន : meaning SIBLING)

3. Line 18 of page 5: Khmer word ផ្អធ. What is the meaning of this word?
   (We guess that it may be ផ្អម : meaning SWEET)

4. We would like asking the authors of N2385 to use the author’s name of the Khmer dictionary correctly. The correct name is Chuon Nath, not Chhuan Nath. Please confirm it in the dictionary.

Taking this opportunity, we would like to point out the questionable name of some Khmer characters in UCS, in terms of Khmer nomenclature, as well as Khmer pronunciation. We would appreciate it if consistent rules for character naming can be developed.

1. U+178E (ណ) KHMER LETTER NNO. As the character belongs to the first register, its name should be NA, not NNO.

2. U+179E (ឞ) KHMER LETTER SSO. With the same reason mentioned above, the character name should be SSA, not SSO.

3. U+17C8 (◌ៈ) KHMER SIGN YUUKALEAPINTU. According to Khmer spelling, the character name should be YUKALEAKPINTU, instead of YUUKALEAPINTU.

4. U+17C9 (◌៉) KHMER SIGN MUUSIKATOAN. According to Khmer pronunciation, the character name should be MUUSEKATOAN (commonly spelled MUSEKATOAN).

5. U+17CA (◌៊) KHMER SIGN TRIISAP. According to Khmer pronunciation, the character name should be TREISAP.

6. U+17CB (◌់) KHMER SIGN BANTOC. According to Khmer spelling, and as C is used to stand for CA and CO, the character name should be BANTAK.

7. U+17D3 (◌៓) KHMER SIGN BATHAMASAT. According to Khmer pronunciation, although the initial character is written as the consonant BA (U+1794), the character name should be PATHAMASAT.

8. U+17D8 (។ល។) KHMER SIGN BEYYAL. According to Khmer pronunciation, although the initial character is written as the consonant BA (U+1794), the character name should be PEYYAL. However its common name is LAK.

9. U+17D9 (៙) KHMER SIGN PHNAEK MUAN. According to Khmer pronunciation, the character name should be PHNEK MOAN.

10. U+17DA (៚) KHMER SING KOOMUUT. According to Khmer pronunciation, the character name should be KOOMOOT (commonly spelled KOMOT).