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

DOC TYPE:	National Body Contribution
TITLE:	On the unsuitability of the “COENG encoding model” for Khmer (Response to WG2 Document N2412, 2002-03-31)
SOURCE:	Cambodian Committee for Standardization of Khmer Characters in Computers*
STATUS:	For discussion at the 42nd WG2 Meeting in Dublin
DISTRIBUTION:	ISO/IEC JTC 1/SC 2/ WG 2 and UTC
MEDIUM:	Electronic
NO. OF PAGES:	6

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

On the unsuitability of the "COENG encoding model" for Khmer
(Response to WG2 Document N2412, 2002-03-31)

We welcome Mr. Michael Everson's recent submission (ISO/IEC JTC1/SC2/WG2 N2412) on the suitability of the COENG encoding model for Khmer, though we cannot agree with him on the main points. We would also appreciate it if he could bring counterarguments, if any, to the remaining points we raised before in our documents (ISO/IEC JTC1/SC2/WG2 N2380R and N2406).

First of all, we have to reconfirm a basic point. The model he calls "COENG encoding model" had been called "virama model" until recently. The critical decision to adopt the existing model in 1998 was made principally on the reasoning that "(t)he main benefit of the virama model was ease of implementation as it is a well-known model (ISO/IEC JTC1/SC2/WG2 N1729)".
We have previously shown that there is no "virama" sign as a general "killer" in Khmer script, unlike, for example, in Devanagari script. So the proponents of the current model had to invent a fictional character as just a control code, which led to a different model from the virama model. The fact that they had to change the name of the model when applying it to Khmer supports our position that it does not correspond to the Khmer reality. Moreover, the "ease of implementation" of the existing model is even denied by implementers themselves, nullifying the reasoning of N1729. For both rendering and sorting, the explicitly encoded subscript model is better than the existing model. In sum, the existing model was decided based on critical misunderstandings.

Now we wish to turn to refuting the new points raised in N2412.

On the adequateness of the existing model

Mr. Everson quoted a figure from "Daniels & Bright 1990" to show that Khmer script came from Indian Pallava prototype, a descendent of Brahmi. We have found the figure 55 rather on p.448 of Peter T. Daniels and William Bright, eds., *The World's Writing Systems*, Oxford University Press, 1996. Anyway we have never argued against the point that Brahmi script is an ancestor of Khmer script. We can, however, refer to significant differences, too, with regard to each of the five points of similarity advanced to justify utilizing the same model.

1. While Khmer does indeed have independent vowel characters, their use is limited. Khmer script has another way to represent the same initial vowel sound as them by using a consonant character QA for a glottal stop and a dependent vowel sign.
   The independent vowel characters are used almost only for old words. Some of them can be written using QA + dependent vowel sign, too.

   ឱ្យ (give) can be written as អោយ.

   The other words, especially new words, are written using QA + dependent vowel sign only.

អី (what, something) cannot be written as ឦ.

អូតែ់ល (hotel) cannot be written as ឩតែ់ល.

The existence of the consonant character QA is one proof of the unique development of Khmer script.

2. While in Khmer each consonant does have an inherent vowel, the Khmer system introduces a new feature with categorizing the consonant characters into two series, and varying the inherent vowel sound for a consonant character depending on which series it belongs to. There are many pairs of characters whose consonant sounds are the same but whose inherent vowel sounds are different.

3. While vowel signs are added to change the inherent vowel sound, because of the unique system of Khmer script mentioned above, the sound of the same vowel sign changes according to the series of the consonant character it is attached to.

4. & 5. They are important points.

Another figure in p.380 of the 1996 Daniels & Bright book referred to by Mr. Everson shows that Brahmi script diverged into northern scripts and southern scripts before the third century. Pallava is among the southern ones, while Devanagari belongs to the northern group.

The northern scripts generally constitute a ligature-like conjunct to represent a consonant cluster, where the original entities cannot be seen separately. There may be multiple representation forms for a single conjunct. These scripts have utilized a “killer” sign (virama) to suppress the preceding inherent vowel sound. Historically its use was limited to denote the absence of the inherent vowel sound of a final consonant of a syllable, but in the modern age it is also used to suppress the inherent vowel of the first consonant(s) in a consonant cluster in order to simplify complex conjuncts.

It is not always the case with the southern scripts. For them, complex conjuncts to represent consonant clusters are rather exceptional. Tamil script has a real general killer sign (pulli), which makes most of such conjuncts unnecessary. Telugu developed another way. It developed consonant signs independent from consonant characters, and put them to the first consonant character to denote consonant clusters. Such differences between northern and southern scripts can be easily seen in the examples of *kta*, as Mr. Everson showed in p.1 of N2412.

Khmer script came from the southern line, but has had its own history of development for more than 1400 years. It developed another complete system of consonant signs that are positioned below a consonant character. Because of this vertical positioning, a consonant sign is called *coeng* (leg, foot). Please note that *coeng* means a consonant sign as a whole, not a “killer” sign like virama. A consonant character and a (subscript) consonant sign are completely independent

entities. In most cases you can combine them as you like without changing their shapes. Complex conjuncts to represent consonant clusters are not necessary at all. This system also widened the use of the consonant signs. Sometimes they are used to denote a final consonant sound in a syllable, as follows:

ន្ម = នាម (name) ≠ ន្ម

ទ្ងំ = ទាំង (both) ≠ ទ្ងាំ

The existing “COENG encoding model” is based on a fictional general “killer” sign arbitrarily named COENG. This model was invented on the ground that a Khmer subscript consonant sign can be interpreted as a combination of COENG and a consonant character because a subscript consonant kills the preceding inherent vowels like virama. This reasoning, however, is not adequate for Khmer script. Please see the examples above. In these cases, the consonant sign DOES NOT KILL any preceding inherent vowel sound. Subscript consonant signs in Khmer have more roles than was expected by those who invented “COENG encoding model”.
We can refer to another example. Not only a consonant character but also an independent vowel character can have a consonant sign below it. There is no change of the initial vowel sound in the following cases.

ឲ្យ (give) ឱ្ន (exclamation of solemn affirmation)

These features show the uniqueness of Khmer compared with Indic scripts, especially Devanagari.

The logic of the virama model is artificial. As Mr. Everson himself admits, there is no virama in the original Brahmi script itself, which means it is not a common or natural feature of those scripts derived from Brahmi. It is just one possible way to deal with complex conjuncts for consonant clusters efficiently by a system of ligature control paying attention to the phonetic function of the virama to kill the preceding inherent vowels. Thus Mr. Everson’s assertion that all the scripts rooted in Brahmi should use the existing model is groundless.

It is clear that such logic is not adequate for Khmer. As shown above, Khmer script has its own unique structure. The existence of subscript consonant signs independent from consonant characters is the core of the structure. Consequently, the explicitly encoded subscript model is far better than the existing model, not only for storing data but also for sorting, searching and rendering precisely because it fits the structure of the script itself.

On the process

As for the lack of due process that is necessary in making international standards, we

wrote basic important facts in ISO/IEC JTC1/SC2/WG2 N2406, so we will not repeat them here, and will limit ourselves to saying that we stand by our position that an irregular and unacceptable process was followed, without proper consultation with the designated national body.
The tentative results of the five meetings Mr. Everson mentioned were summarized in a private report of National Higher Education Task Force dated on August 14, 1996, addressed to Mr. Maurice Bauhahn. Although it is true that eminent linguists gathered, they did not decide any official or final stance of Cambodia. The report itself says it is not sufficient. This task force was not given a mandate to make an official decision on this issue. It had nothing to do with the national standards body of Cambodia that had already been registered with ISO in 1995.

Nevertheless, it is still useful to confirm here that the report clearly listed subscript consonant signs independently from consonant characters among the necessary characters that should be encoded. While non-Cambodians might have suggested to them to accept "virama model" they evidently refused to do so. Mr. Everson's assertion that they were not explicitly against "virama model" is not supported by the facts shown in the report.

We would like to add that some of the scholars mentioned by Mr. Everson are clearly supporting the current Cambodian stance.

On ROBAT

In modern Khmer script, ROBAT has lost its original meaning as a variant of RO to represent a consonant cluster beginning with RO. In some old loan words from Sanskrit/Pali, it is pronounced according to its original rule i.e. just before the base character it is attached above. In the other such old loan words, however, it is not pronounced at all. It is kept just for information of the original spelling.
It is sure that we can see words containing ROBAT even now, it is not a rule for Khmer script itself to spell a consonant cluster beginning with RO by ROBAT. The rule is to spell a consonant character RO and a subscript consonant sign of another consonant below it. More than a hundred examples can be found in the Cambodian standard Chuon Nath's dictionary. Some words are written in both ways.

អារ្យ = អាយ៌ (civilized), រាជរ្ឫី (king hermit), etc.

Thus Mr. Everson's proposal to deprecate ROBAT based on the premise that the consonant character RO cannot have a subscript consonant sign is not acceptable.

On other points

Mr. Everson is trying to play down some of the strong points of the explicitly encoded subscript model we are proposing, but he cannot deny them. That is enough for us.

The ultimate reasons for not adopting our model seem to be procedural ones. We also have much to say about procedures, as we wrote in N2406.

Mr. Everson asserts that UCS as a universal encoding standard and interchange platform would be compromised if our requests are accepted. We do not think so. “Universal” does not mean “all the same”. It should mean “everyone can enjoy it”. For that purpose, the credibility of Unicode for everyone should be important. Please note that we are making our proposal to make UCS/Unicode better, not to put it down.

We would like to add another point finally. Even if an encoding is not a good one, there is no problem once it was approved by the concerned parties. This Khmer case is an irregular one where such due process was not followed. Cambodia has never approved the existing standard. It was never informed or consulted. We believe this is a special case. So nobody needs to worry about possible changes in the standard of other scripts that were established with the approval of concerned parties.