Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation internationale de normalisation
Международная организация по стандартизации

L2/02-131

Doc Type:	**Working Group Document**
Title:	**On the suitability of the COENG encoding model for Khmer**
Source:	**Michael Everson**
Status:	**For discussion**
Action ID:	**ACT**
Date:	**2002-03-31**
Distribution: WG2 and UTC	

It is my belief that the COENG encoding model links the Khmer script to its sister scripts, which use the VIRAMA encoding model, and reflects, therefore, a long and venerable history of writing.

In the middle of the third century BCE – some 2250 years ago – the Mauryan emperor Aśoka is thought to have directed the creation of the Brahmi script. Brahmi was used in all parts of India to write Prakrit and Sanskrit (except for the north-western regions where Kharoshthi was used to write Prakrit and Gandhari, and occasionally Sanskrit). Brahmi shows all the features found in the scripts descended from it:

1 There are a number of unique initial vowels.
2 Each consonant has an inherent vowel.
3 Vowel signs are added to change the inherent vowel.
4 Graphic syllables consisting of a cluster of two or more consonants require that the consonants be joined together in a conjunct character cluster to indicate the cancellation of the inherent vowel in the preceding consonant.
5 The relationship of the initial consonant and the other consonants in the conjunct is *in general* shown by reducing and subscripting the subsequent consonants.

These *structural* features are maintained in almost all of the Brahmi-derived, or Brahmic scripts – even where the pronunciation of the basic characters has changed greatly, as it has, for example, in Myanmar and Khmer.

NOTE: The glyph relationship described in (5) above is not always the case in scripts descended from Brahmi – in Devanagari, sometimes the first consonant in a conjunct is halved. The syllable which had an *original* pronunciation *kta* may have different outer form in different Brahmic scripts, but the structure is the same: there is a KA, with a vowel killed, joined somehow with a TA. In some scripts, the KA continues to look like a KA and the TA is greatly altered. In others, the reverse is true. In still others neither are greatly changed in form. But *kta* is still *kta* – a KA, with a vowel killed, joined somehow with a TA.

Brahmi	✛ ka	+	⋏ ta	=	⟊ kta
Devanagari	क ka	+	त ta	=	क्त kta
Kannada	ಕ ka	+	ತ ta	=	ಕ್ತ kta
Myanmar	က ka	+	တ ta	=	က္တ kta
Khmer	ក ka	+	ត ta	=	ក្ត kta

2002-02-28

The glyphs used in Brahmi regional varieties diversified, and two major divisions had emerged by the first and second centuries CE: Northern and Southern styles, which subsequently developed into Western and Eastern styles in the north, and into Deccan and Peninsular styles in the south. By about the sixth century CE, a southern Indian script of the Pallava dynasty had spread to Southeast Asia, where it too formed regional varieties in Fu-nan (southern Vietnam and Cambodia), Champa (central and southern Vietnam), in Cambodia, in the Mon area of Thailand, in Sunda (western Java), in Central and East Java, in East Kalimantan (Borneo), in Sumatra, and in the Malay peninsula. A script called by de Casperis "Cambodian Nagari" had developed unique features by about 900 CE.
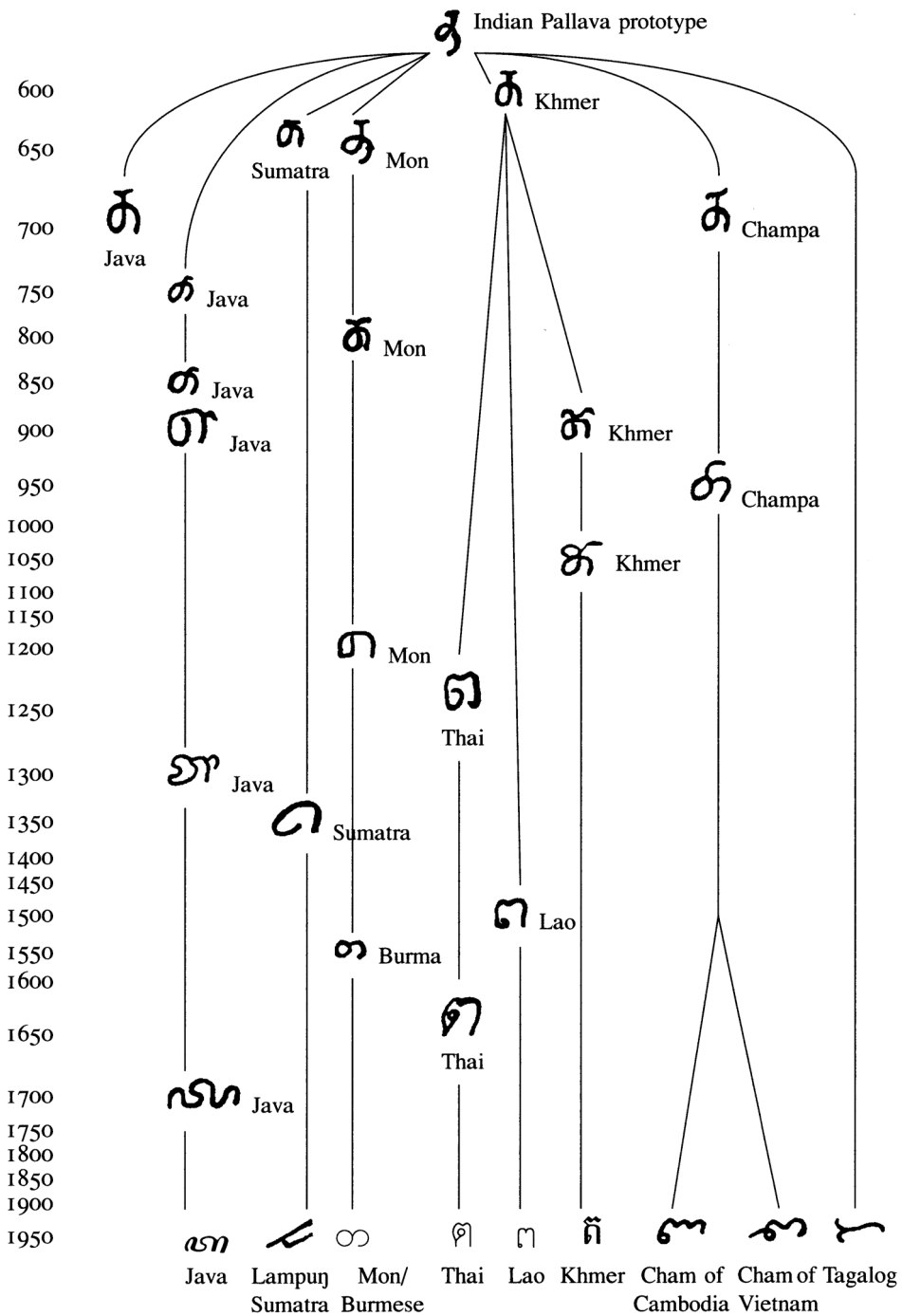

FIGURE 55. Family tree of the akshara *ta* in Southeast Asian scripts. Pallava prototype from Filliozat 1953: 697; other forms from Damais 1955, fig. 15.

Figure from Daniels & Bright 1990 showing Pallava TA and its descendants.

Why have some of our Cambodian colleagues said that they believe the Brahmic encoding model to be inappropriate for their script? It is, in my view, because the *reading rules* of Sanskrit and Pali don't apply in the same way as the reading rules of Khmer, and this has, historically, affected how the script is *taught* in Cambodian schools. The Pallava scripts came as vehicles of Sanskrit and Pali, languages which have very different phonetic structure from medieval Khmer or medieval Myanmar. Pali words changed their pronunciation, quite naturally, in Cambodia and Myanmar, just as they did in other parts of East and Southeast Asia. Further changes in the Khmer language made the reading rules of the script even further removed from the pronunciation of the original letters. The Sanskrit word *dharma* ['dʰʌrmʌ] (Devanagari धर्म, Khmer ធម៌) is now pronounced [toə] in Khmer. (Phonetic change occurred in Prakrit too, where the pronunciation became ['dʰʌm:ʌ], which the Buddhists came to respell as *dhamma* (Devanagari धम्म); in Myanmar this is also written ဓမ္မ *dhamma* and pronounced ['dʌma.].) Still, the structure of the word and its letters is the same: DHA + RA + killed A + MA (Prakrit DHA + MA + killed A + MA). This structure has been encoded with a character VIRAMA, which, in Unicode, marks the killing of the vowel and also indicates to the font that some kind of special glyph arrangement applies to the two consonants on either side of the VIRAMA.

It does have to be said that Khmer has a descendent of a visible vowel-killing sign, seen in the Devanagari VIRAMA ्, and in the Myanmar killer ်: this is the Khmer WIRIAM ៌ – although WIRIAM has taken on a very different function in Khmer which cannot be equated with the conjunct-forming underlying COENG. Nevertheless, this conceptual COENG has a place in Khmer encoding, in preserving the *structural* relationship of encoded words – particularly in Sanskrit and Pali and in loanwords in Khmer from those languages. It should also be said that a visible VIRAMA does not occur in the earliest Brahmi texts – making the COENG model even more Brahmic than the VIRAMA model! And modern Hindi has also departed in its *reading rules* from the pure Brahmic model: often, as in Khmer, a killed vowel is not represented by either a conjunct consonant, consonant with subscript, or by a visible VIRAMA. Examples of this are Hindi गनेश *Ganeś* (compare Sanskrit *Ganeśa*) and देवनागरी *devnāgrī* (compare Sanskrit *devanāgarī*). In this encoding model, the VIRAMA and COENG's primary function is to cause one of a pair of consonants flanking it to change its shape – in *both* Devanagari and Khmer.

It has been argued that Cambodians perceive KA and COENG KA as different. Well, they *are* different. ្ម COENG MA has a different glyph and may be, today, pronounced differently from ម MA in at least some contexts. But Khmer linguists and monks, pronouncing Sanskrit as it was pronounced historically – ['dʰʌrmʌ] rather than [toə] – must use different *reading rules,* just as different reading rules are used by linguists in pronouncing Middle English *make* ['makə] and Modern English *make* [meik], or Ancient Greek βῆτα ['beta] and Modern Greek βῆτα ['vita]. I am told that when a Cambodian is spelling a word, ក្ក *kka* is read out as "KA COENG KA". The Brahmic model analyses this as [KA][COENG][KA] instead of [KA][COENG KA] as in the subscript. *Either* model *could* have been used, in principle. It has even been argued recently on the Unicode list that Devanagari क्क *kka* could also be reanalysed as [KA VIRAMA][KA] and क्क *kka* as either [KA][VIRAMA][KA] or [KA][VIRAMA KA]. Conceptually, either a VIRAMA/COENG model or a subscript model could have been applied to all Brahmic scripts, and it's not necessarily obvious what is best depending on how one *feels* about conjunct formation. Here, the structural model KA + killed A + KA in the encoding is expressed with the encoded VIRAMA/COENG characters, with the glyph implementation left to the font.

> NOTE: The problem of representing early pre-Khmer and early Khmer texts is related to the encoding model. If the subscript model were chosen for modern Khmer, we would have to deal with deciding when exactly "modern" Khmer began, and then we would have to encode a separate script, or several separate scripts, for that which preceded.

Mr Svay Leng has said that Cambodian programmers can work with the current COENG model if they have to. He has presented arguments about they *perceive* the script's structure, but he has *not* shown that the Brahmic analysis is unsuited to Khmer. Indeed, as shown below, the Brahmic structure is obvious when one compares a single word written in several scripts.

Here I have written the word "Sanskrit" in five Brahmic scripts. What we see is a set of letter shapes, which we know to be historically related, written in a linear fashion which is surprisingly similar given the changes in the basic shapes themselves. The words on the right are all in the Sanskrit language, *saṃskṛta*, and are all spelled the same with the phonetic Brahmic model: the Brahmi, Devanagari, Kannada, and Myanmar examples are written SA + ANUSVARA + SA + VIRAMA + KA + VOWEL SIGN VOCALIC R + TA. The Khmer example is written SA + NIKAHIT + SA + COENG + KA + COENG + INDEPENDENT VOWEL RY + TA. (Khmer no longer has a special sign for the vowel sign VOCALIC R and uses the independent vowel RY in a reduced subjoined form.) The words on the left are slightly different from one another, because an orthographic change has occurred in both Myanmar and Khmer. The Myanmar is written *saṅsakaruit*, SA + NGA + VIRAMA + SA + KA + RA + VOWEL SIGN U + VOWEL SIGN I + TA + VIRAMA + ZWNJ and is pronounced [θinθʌkʌraiʔ]. The Khmer is written *saṃskrita*, SA + NIKAHIT + COENG + SA + KA + RO + VOWEL SIGN Y + TA + COENG + ZWNJ and is pronounced [səŋskrit]. The Brahmi, Devanagari, and Kannada examples are fictitious, since those scripts did not substitute RA + vowel for VOCALIC R; the example is written *saṃskrīta*, SA + ANUSVARA + SA + VIRAMA + KA + VIRAMA + RA + VOWEL SIGN II + TA; if the Myanmar were written as these last examples are, it would be သံြွ္စ္ဝိတ – note how similar this is, structurally and glyphically, to the Khmer សំ ស្ក្រ៊ិ ត! The Brahmic model holds completely and consistently for everything represented here despite superficial orthographic differences.

A note on the history of the COENG *encoding*
An official committee of Cambodian linguists assembled by the Cambodian government took part in laying the foundations of the encoding of Khmer in Unicode. This committee consisted of the members of the National Working Group on language as well as Khmer researchers at the Research Institute, colleges, and universities. These eminent Khmer scholars, Professor Thoang Thel, Dr Long Seam, His Excellency Chhorn Eam, Mr Pit Chamnan, Mr Can Mono, Mr So Muy Kieng, Mr Ly Sovy, Dr Neou Sun, and Mr Ok Cuan, met together to discuss and exchange ideas together five times at the Secretariat of the National Higher Education Task Force, room 8C, on 21 March, 28 March, 24 May, 31 May, and 18 July in 1996. The merits of the COENG model and the subscript model were discussed by this committee. The linguists understood the structural nature of the COENG model and the advantages it would have for multiscript publication of Buddhist texts, and the usefulness of such a model for comparative linguistics.

Structurally, Khmer can easily be considered to be a Brahmic script despite the way that Cambodians *pronounce* the written forms of their language. The experts consulted in 1996 recognized that a Brahmic model for encoding a Brahmic script was appropriate for Khmer, and was based in a structural system more than two thousand years old, first introduced into Cambodia some 1400 years ago, where it developed, as noted above, into unique Cambodian glyphs 1100 years ago. Although the COENG model was discussed with the government committee, it was not specifically referenced in their final report. What is clear, however, is that they raised no objections to it. It should also be noted that when UNESCO invited Khmer government representatives to discuss Unicode for Khmer around 1993 with Peter Lofting, they were encouraged to favour the COENG model.

It has been claimed that the COENG model results in larger text and file sizes, and that this is an unacceptable penalty, particularly in a developing country which may not be able to afford immediate infrastructure upgrades to the latest equipment.

It is certainly true that the COENG model requires storage of two characters for each subscript, rather than the single character which would be required if each subscript consonant (or independent vowel) were separately encoded. However, the difference in size of text and files is not as significant as it might appear to be. First of all, the overall expansion that results for basic Khmer text is not a doubling in size, but more like 12%, based on the frequency of subscripts in text. Second, in nearly all modern applications, the raw size of text data is now swamped by the size of markup text, formatting data, graphics, and the like. For example, on many web pages, the body text itself is a very small percentage of the overall data. Third, the cost of hardware for storage, the cost of bandwidth for transmission, and similar data size bottlenecks continues to drop. In 1991 I bought an external 200 MB hard drive for $1000 ($5 per MB). In 1997 I bought an external 4 GB (4,000 MB) hard drive for $400 (10¢ per MB). An external 30 GB (30,000 MB) hard drive costs $220 today (less than a penny ($.0073) per MB), a megabyte of disk space now costing 685 times less than it did in since 1991. Disk space prices continue to fall worldwide. Fourth, the need for upgrading systems to the level of operating system and application support that can handle the complex rendering of Khmer will be driven by other considerations, rather than the minor differences in encoding efficiency between a COENG or an explicit subscript encoding, in any case – the ability to connect to the Internet and make use of current-level browsers and other state-of-the-art software such as databases is far more important in forcing such infrastructure upgrade decisions.

Rendering for Khmer must be font-based no matter what, if the positioning of vowel signs and marks is to be handled correctly. Rendering with the COENG model is no worse than rendering for other Brahmic scripts, or indeed for Arabic, which is also quite complex. Exactly the same kinds of rendering rules apply to all Brahmic scripts. The glyphs are different, but the same structural relationship between strings of characters and glyph cells applies. And it would be no different if we had encoded Khmer with a subscript model.

Khmer has been encoded with a Brahmic model in the International Standard, with participation of a team of experts, a number of whom were Cambodian linguists. A number of companies have used this standard encoding as the basis of software development. Deprecating the current model in favour of another will destabilize all support for Khmer, and would open the door to argument from people who have not understood the Brahmic encoding model and who want it changed for *all* of the Brahmi-encoded scripts in the standard. If we were to allow this to happen, the whole value of the UCS as a *universal* encoding standard and interchange platform would be compromised. Software development for *every* Brahmic script would be jeopardized.

It has been argued that the COENG model facilitates programmers worldwide in creating solutions for Khmer. This is true, and it works just as well for Cambodians as it does for everyone else. If Cambodian

programmers and font developers are conversant in the COENG model for their own script, they will find it far easier to port their own software to other Brahmic scripts. An economic advantage for *everyone.*

The question of ROBAT

Assuming that the COENG model will be retained for Khmer, one outstanding problem is what to do about ROBAT. This character is in origin a special form of RA, which, because of the frequency of pre-consonantal -R in Sanskrit, came to have a symbol of its own as a way of facilitating manuscript hand-writing. Comparing the word *dharma* shown above in Devanagari (धर्म) and Khmer (ធម៌), it is easy to see that the same mark is used for the same function; indeed the name *repha* is related to the name *robat.* In both the COENG and VIRAMA models, this *should* be encoded as DHA + RA + VIRAMA + MA (Khmer THO + RO + COENG + MO). With the current encoding, however, the difficulty is that the Khmer word ធម៌ could also be written THO + MO + ROBAT. Because the COENG model is based on the Brahmic phonetic model, however, the situation arises that we may have two ways of representing the same word ធម៌. This question should concern us. For example, if ROBAT is not deprecated, then the logical consequence would be that THO + RO + COENG + MO would yield the form ធ្រម, which, as far as I know, is not permitted. It would also introduce a difficulty in transliteration of Sanskrit text from other Brahmic scripts into Khmer. In my opinion, the explicitly-encoded combining character ROBAT was an error based on an inadequate analysis of Khmer, and it should be deprecated in favour of a consistent application of the COENG model. The COENG model implies that where two consonants flank the COENG character, one of them changes its shape to form a conjunct; in the case of RO, it takes one of two special shapes, ○ or ○, depending on whether it precedes or follows the COENG. This is logical, and it is consistent with the historical development of Brahmic RA ligatures in many scripts as well as in Khmer. Deprecating the ROBAT and keeping to the COENG model here will also greatly facilitate ordering. Khmer orders the ROBAT form of RO phonetically – in accordance with the Brahmic model. In the Chuon Nath dictionary, ទុក្កត *duggata* precedes ទុរា *durā* precedes ទុគ៌ត *durgata*, and if this is encoded <u>TO-U-KO</u>-COENG-KO-TA > <u>TO-U-RO</u>-AA > <u>TO-U-RO</u>-COENG-KO-TA, it is simpler for the sorting algorithm than if it were <u>TO-U-KO</u>-COENG-KO-TA > <u>TO-U-RO</u>-AA > <u>TO-U-KO</u>-ROBAT-TA. Thus the ordering practice of the Chuon Nath dictionary is in complete accord on this matter with the traditional Brahmic ordering practice, as can be seen by comparing the Khmer samples with the Devanagari samples below.

458	duka	ទុក	दुक
461	duggata	ទុក្កត	दुग्गत
462	duṅa	ទុង	दुङ
462	dutiya	ទុតិយ	दुतिय
463	duna	ទុន	दुन
463	dubūla	ទុពូល	दुबूल
465	du*r*abala	ទុរពល	दुरबल
465	du*r*abhiksa	ទុរភិក្ស	दुरभिक्स
465	du*r*ā	ទុរា	दुरा
465	du*r*āgaman	ទុរាគមន៌	दुरागमन्
465	du*r*ācāra	ទុរាចារ	दुराचार
465	du*r*ena	ទុរេន	दुरेन
465	du*r*gaka	ទុគ៌ក	दुर्गक
465	du*r*gama	ទុគ៌ម	दुर्गम
465	du*r*jana	ទុជ៌ន	दुर्जन
465	du*r*yasa	ទុយ៌ស	दुर्यस
466	dula	ទុល	दुल

It seems to me that the explicit ROBAT should be deprecated in order to ensure a consistent application of the COENG model as a Brahmic model.