Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation

Doc Type: **Working Group Document**
Title: **Response to L2/20-018 W. Cham response**
Source: **Martin Hosken**
Status: **Individual contribution**
Action: **For consideration by UTC and ISO**
Date: **2020-01-20**

Introduction. This document is a response to L2/20-018 which in its turn is a response to L2/19-217r3.

L2/20-018 makes the claim that only 10% of the Western Cham community is part of the Imam San community, which is accurate, but then only 10% of the Western Cham use the Kakkhak script. The rest use Jawi (Arabic script). Therefore the script as presented by the Imam San community is the script used by nearly all Kakkhak script users.

With regard to whether characters should be added to the standard, there is a relatively low bar set in whether there is a user community that would use this character. The difficulty in this case is that the Imam San community considers the analysis in L2/20-018 faulty and so the addition of the proposed characters as redundant at this time. There is interest from the Imam San community in engaging with the Cham Language Advisory Committee to see what might be done to address real issues in the script and help it to move forward. On that basis, it is the author's recommendation that the extra characters proposed by the CLA Committee not be added at this time and for the wider community to have its discussions and then to bring a unified proposal for these characters. By leaving the gaps in the chart as they are, the way is left open for the potential additions to be added easily.

In discussing the various proposed additions, only those examples that are not simply publications from the CLA Committee will be discussed, since the other examples beg the question.

Final G (U+1E241) There is no need to introduce another distinction given that 'coin' may be encoded ꩲꩲ U+1E206 (ka) U+1E240 (final k) and 'tie' as ꩲ U+1E206 (ka) U+1E237 (tkaj ka).

Final B (U+1E23C) There is no ambiguity to resolve since 'smoke' is spelled ꨠꨳ U+1E227 (sa) U+1E22F (e) U+1E231 (au) U+1E247 (final p) and 'sound' is spelled ꨳꩲ U+1E227 (sa) U+1E247 (final p). 'smoke' is never spelled the same way as 'sound'.

Figure 3 shows ꨁ U+1E21D (ba) U+1E237 (tkaj ka) as presented in L2/19-217r3 as figure 15. The presence of a tkaj ka does not require the utterance of a vowel. It is an acceptable way of marking a final.

Final M (U+1E24C) The Cham people never refer to themselves or the region or ever use the long vowel for ca:m, including people from the Kampong province. The long vowel is used in Khmer to refer to the Cham, but that is a Khmer script issue, not a Cham one. 'meet' is spelled ꨠꨕ U+1E213 (ta) U+1E22F (e) U+1E231 (au) U+1E24C (final m) while 'to transplant' is spelled ꨕꩺ U+1E213 (ta) U+1E232 (u) U+1E24C (final m).

Figure 4 shows ꨋ U+1E21F (ma) U+1E237 (tkaj ka) as presented in L2/19-217r3 as figure 16. Again the presence of tkaj ka does not require its utterance and is an appropriate final marker.

Final NG (U+1E243) The question here is whether we encode two characters or one. The core question is whether a single codepoint can be used because what is visually indistinct in one variety can be

algorithmically distinguished such that a single codepoint can be used for the two forms in another variety. The question of contrast when there are two characters is irrelevant. The question is what happens when there is only one and the shapes are different.

The linguistic rule is that the oe/ng character acts as the vowel oe when followed by another final and as the final ng when not.

Having a rule that allows the disambiguation of a character does not of itself require that there be only one code. The usability issue is that for the variety that has no visual distinction between the characters, that there be only one key on the keyboard. The question therefore becomes one of whether the disambiguation occurs in the font for those varieties with contrastive shape or in the keyboard for those varieties with no contrast. It is probable that sorting calls for different codes, but using different codes with a dominant variety not distinguishing them introduces bad confusability. The current proposal comes down on the side of a single code for this to encourage the creation of keyboards which need relatively little complexity otherwise.

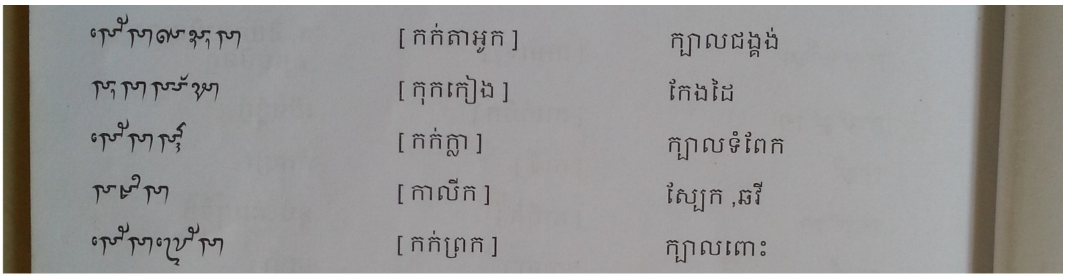
Modifed Nasal Consonants No justification is given for the inclusion of the extra characters beyond "we want them". Hence their removal from the current proposal. The Imam San community uses U+1E20A (ngue) U+1E235 (sign la) U+1E206 (ka) instead of nuge + takay klak, which has never been used by them.

Figure 15 is U+1E210 (nhue) U+1E235 (sign la) U+1E229 (aa). Here the U+1E235 (sign la) is used here to mark the vowel as /a/ as opposed to the default /aʔ/ and U+1E229 (aa) to give it length.

The Kakkhak script does support ɲ, via U+1E210 (nhue). The desire for U+1E211 comes from parity with Eastern Cham. The problem with the proposed character is that it is confusible with as in figure 16 which has which is very close visually to . The Imam San community are not against adding a character, per se, but do not consider the proposed shape appropriate.

Vowel Sign u Lack of a separate u vowel for transcribing Eastern Cham contrasts is certainly a lack in the Western Cham orthography. There is an openness to perhaps adding a character for this, but there would need to be agreement over what it should look like.

Collation Different collations are available via collation tailoring. The collation used in the 2011 dictionary referenced is effectively random. No standard order was used. For example, you can see the problems here:



[កក់តាអូក]	ក្បាលជង្គង់
[កុកកៀង]	កែងដៃ
[កក់ក្លា]	ក្បាលទំពែក
[កាលឹក]	ស្បែក ,ឆវី
[កក់ព្រក]	ក្បាលពោះ

where the first, third and fifth entries all have the same initial syllable but the interleaving words have a very different first syllable.

On this basis, there is agreement to go with the Eastern Cham based ordering as proposed.