

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

Doc Type: Working Group Document
Title: Proposal to Create Variation Sequences for Khamti Characters
Source: Martin Hosken
Status: Individual contribution
Action: For consideration by UTC and WG2
Date: 2015-11-03

Executive Summary This proposal is to encode variation sequences for the Khamti and Aiton and Phake style Myanmar consonants (those with the dots).

Introduction For more details on the discussion that led to this document, see L2/15-257.

In the encoding of Khamti, Aiton and Phake, the decision was made to unify the dotted characters with their undotted forms. The differences were considered stylistic:

Burmese Style	Khamti Style
၀	၀

Most Khamti, Aiton and Phake users living in Burma are also fluent in Burmese and Shan, and use those languages, as well as their own language, on a computer. In a plain text context (such as is most commonly used, including Facebook, SMS, email) where these languages are being used, the Burmese style of characters gets used exclusively. This is because it makes even less sense to view Burmese using Khamti style¹ characters than to view Khamti using the dotless Burmese style. This has the effect of users rarely seeing their language written in an appropriate style. There is no option to select an appropriate font since the same codepoints are being used for Burmese as for Khamti and so, in a plain text context, there is no way to see the two styles. The communities, therefore, have a real concern that a significant aspect of their cultural heritage, tied up in their script, will be lost. They are therefore requesting that the characters that have a Khamti style be disunified from their Burmese equivalents, so that in multi-lingual plain text, the contrast may be conserved.

Variation Sequences: Rather than a full disunification, an alternative approach is to use variation sequences to provide a unique encoding for each of the Khamti style characters. Variation sequences are designed precisely for this kind of problem. In all respects the characters are the same as the Burmese form characters, except for their visual form. A variation selector is stored after the character to indicate to rendering processes that the alternate shape should be used.

Khamti Shan also has tone marks which are solid dots. This might be considered for variation. But there are also styles of Burmese where the tone mark dots are filled in. The contrast between these two forms is not considered important to the communities, so if the wrong style is used in a context, this is not considered ideal, but neither would it cause a problem. The default style of whether tone marks are filled or not is often language motivated, but can take either style somewhat arbitrarily. Clearly in Khamti they should always be filled, but if they are not, the community does not consider this as the same level of problem as the dotted forms of the base characters.

¹ For brevity we use the term 'Khamti style' to cover Khamti, Aiton and Phake styles. There is some difference, but in general they are the same. Likewise 'Burmese' for 'Burmese and Shan'

Rationale Why should this community have particular support for its character styling where many other communities are limited to using different fonts or language marking? The user community for which this is a problem is one that uses both Burmese and Khamti intermixed. This is in contrast to other styling situations where users are not expecting to use both styles intermixed in a plain text situation. In addition, while the Khamti community is very grateful to Unicode for giving them the ability to type in their language. But they are deeply concerned that in plain text situations, their particular language styling will be lost, along with it an important part of their culture. With less and less text being printed and more and more text being used in plain text contexts, they want to be able to see the dots on their characters. But if all characters are dotted, then Burmese will look ridiculous with dotted form characters. Likewise Khamti text looks anaemic without them.

Proposal: to use Variation Selector characters. U+FE00 VARIATION SELECTOR-1 follows the compatibility decomposition character listed in the database entries above to result in the corresponding glyphs:

Original Glyph	Character Sequence	Sequence Glyph
က	1000 FE00	က
ဂ	1002 FE00	ဂ
င	1004 FE00	င
စ	1010 FE00	စ
ဆ	1011 FE00	ဆ
ဗ	1015 FE00	ဗ
မ	1019 FE00	မ
ယ	101A FE00	ယ
ရ	101C FE00	ရ
ဝ	101D FE00	ဝ
က	1022 FE00	က
ဇ	1031 FE00	ဇ
၈	1075 FE00	၈
၉	1078 FE00	၉
၀	107A FE00	၀
၁	1080 FE00	၁
၂	AA60 FE00	၂
၃	AA61 FE00	၃
၄	AA62 FE00	၄
၅	AA63 FE00	၅
၆	AA64 FE00	၆
၇	AA65 FE00	၇

Original Glyph	Character Sequence	Sequence Glyph
၀	AA66 FE00	၀
၁	AA6B FE00	၁
၂	AA6C FE00	၂
၃	AA6F FE00	၃
၄	AA7A FE00	၄

This proposal will result in a growth of storage needs for Khamti text, for which SMS is probably the only situation where this is a problem. There is some ambiguity between text stored without variation selectors but is marked as being Khamti for rendering purposes and text that is stored using variation selectors. But good search should be able to ignore variation selectors.

There are different dotted forms for Aiton, Phake and Khamti, but only one dotted form needs to be encoded since users are happy to not have an encoded contrast between these languages, but just between the dotted and undotted forms. Notice that the dotted forms for AA7A only has dots in Aiton.

The following lines need to be added to the Unicode Database file StandardizedVariants.txt

```
# Myanmar
1000 FE00; dotted form;      # MYANMAR LETTER KA
1002 FE00; dotted form;      # MYANMAR LETTER KHA
1004 FE00; dotted form;      # MYANMAR LETTER NGA
1010 FE00; dotted form;      # MYANMAR LETTER TA
1011 FE00; dotted form;      # MYANMAR LETTER THA
1015 FE00; dotted form;      # MYANMAR LETTER PA
1019 FE00; dotted form;      # MYANMAR LETTER MA
101A FE00; dotted form;      # MYANMAR LETTER YA
101C FE00; dotted form;      # MYANMAR LETTER LA
101D FE00; dotted form;      # MYANMAR LETTER WA
1022 FE00; dotted form;      # MYANMAR LETTER SHAN A
1031 FE00; dotted form;      # MYANMAR LETTER SIGN E
1075 FE00; dotted form;      # MYANMAR LETTER SHAN KA
1078 FE00; dotted form;      # MYANMAR LETTER SHAN CA
107A FE00; dotted form;      # MYANMAR LETTER SHAN NYA
1080 FE00; dotted form;      # MYANMAR LETTER SHAN THA
AA60 FE00; dotted form;      # MYANMAR LETTER KHAMTI GA
AA61 FE00; dotted form;      # MYANMAR LETTER KHAMTI CA
AA62 FE00; dotted form;      # MYANMAR LETTER KHAMTI CHA
AA63 FE00; dotted form;      # MYANMAR LETTER KHAMTI JA
AA64 FE00; dotted form;      # MYANMAR LETTER KHAMTI JHA
AA65 FE00; dotted form;      # MYANMAR LETTER KHAMTI NYA
AA66 FE00; dotted form;      # MYANMAR LETTER KHAMTI TTA
AA6B FE00; dotted form;      # MYANMAR LETTER KHAMTI NA
AA6C FE00; dotted form;      # MYANMAR LETTER KHAMTI SA
AA6F FE00; dotted form;      # MYANMAR LETTER KHAMTI FA
AA7A FE00; dotted form;      # MYANMAR LETTER AITON RA
```

Acknowledgements Thanks go to Payap University Linguistics Institute, Chiang Mai, Thailand, under whose auspices this work is done.

Samples These samples courtesy of Stephen Morey.

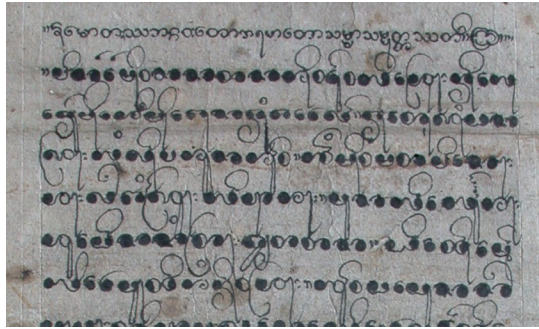


Illustration 1: Old style Phake with an initial line in Burmese.

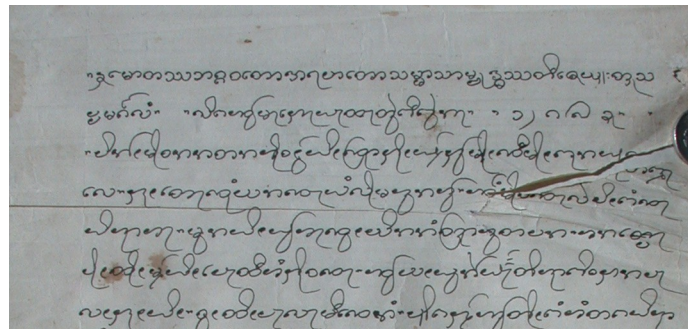


Illustration 2: Modern handwritten Phake, with Burmese heading

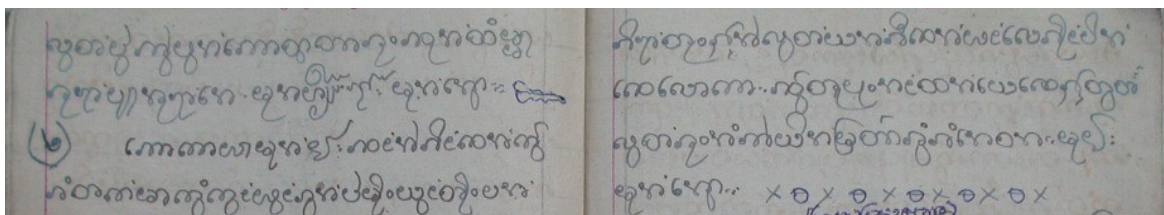



Illustration 3: Handwritten Khamti Shan

[illegible]

Profile

နမောတဿဘဂဝါဒဏေဝံ အရဟတဿသမ္မာ သမုဗ္ဗဓေဝံ ဇေယျသုပပဓဂီဝံ
 အဟိဂ္ဂနဒ မိရိဟိတဏါ ဩဝါဠကဿာ။ မိဉ်ဘာ်ဟွံတိပုံ
 ။ ဣကံမိုင်ကိုင်မိုင်ဟွံဘာ်မိုင်သုဉ်ဟွံဟွံယွံ ။ ဟင် ၇ ဟွံမိဂုဉ်မိုင်ဣကံ ။ ဟင်တင်ယင်သုဉ်သိမိမိ ။
 သုဉ်ကံမိဉ်ဟွံတင်ကော # ဟွံတင်သုဉ် ။ ဟွံတင်သုဉ်ဟွံ ဟွံကိဉ်မိုင်ယွံကွဲ ။
 ငါမံပမိဉ်ပုံဟွံကိဉ်မိုင်သုဉ်သုဉ်ဟေ ။ ယံမံကံကိုင်ဣကံ ။ ဟွံဘာ်ကောဣမိုင်ယွံ ။ ယံမံမိုင်ကံကိုင်ဟေ ။
 ဟွံကိဉ်မိုင်ဟွံကံကိုင်ဟွံဘာ်မိုင်ယွံ ။ ဟွံမံမိဉ်ဣကံမိုင်ဟင်ဟင် ။ ဟွံကိဉ်မိုင်မိဉ်ကံကိုင်ဟွံဘာ်မိုင်ယွံ ။
 ဟွံတင်သုဉ်ကောဟင်တိဣသုဉ်ကံကိုင်ဟွံဘာ်မိုင် ဟွံ ။ ဣကံကောသုဉ်ကံကိုင်ဟွံဘာ်မိုင် ။
 ဣကံကောကွံတင်မိဉ်တင်မိုင်ဟေ ။

 Chat (32)

5