ISO/IEC JTC1/SC2/WG2 N4xxx
UTC L2/15-257
2015-11-02

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

Doc Type: **Working Group Document**
Title: **Proposal to Disunify Khamti Letters from Myanmar (Revision 3)**
Source: **Martin Hosken**
Status: **Individual contribution**
Action: **For consideration by UTC and WG2**
Date: **2015-11-02**

Executive Summary This proposal is to disunify the Khamti and Aiton and Phake style Myanmar consonants (those with the dots) into their own letters as part of the Myanmar script block, and thus create a new Myanmar Extended block to hold them in. The effect is to add 16 characters. In addition the representative glyphs for Khamti specific letters in the Myanmar Extended block are adjusted to have the Khamti style.

Introduction 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
ဎ	ဎ

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 style[1] 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.

An alternative analysis of the Myanmar script would consider the dots to be productive. In this analysis, the dotted base characters have their own codepoints and are considered separate characters. In addition the characters that were given representative glyphs without dots, in order for them to fit with the encoding model, would regain their dots. Khamti Shan also has tone marks which are solid dots. This might be considered for disunification. But there are also styles of Burmese where the tone mark dots are filled in. If the tone marks were disunified, it could introduce confusion as to which tone mark a Burmese user should use, especially with the tendency of people to follow form over function when encoding. For this reason, there is no intention to disunify the tone marks. In addition, using the 'wrong' style of tone marks in a plain text context, is considered acceptable.

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'

In summary, the proposal is to promote dots on base characters from being modelled as being stylisting to being productive.

Rationale It can be argued that such a disunification is a form of glyph encoding given that it is visually motivated. But it is a question of which model to use, and in choosing which model, it is worth considering the wider problem. Technology has impact on language and particular on writing. For example, the Thai alphabet has two unused characters that were once used, but were consciously abandoned because there was insufficient room on the typewriter keyboard for them. A number of language groups in Southeast Asia consciously decided to use a Latin based script in order to mitigate technological problems with non-Roman scripts. In the case of the Khamti and Aiton and Phake, they are seeing that technological limitations will have a significant impact on how their script will change in the future, and they are concerned about it. On the one hand, Unicode has made it possible for them to work with their writing on computer and they are very grateful. But they also see that as less and less information is printed and therefore has styling as a strong consideration, along with the effort to get that right, the technology for informal textual interaction is not capable of supporting their cultural forms. This proposal, therefore, arises out of a desire to address those limitations.

In addition to the user community wanting to work with their language, these users engage in active code switching between languages when communicating with each other and those outside their local language communities. Thus a user may well type a text to a friend in Khamti and then immediately type another text to a different friend (or the same one) in Burmese. They may even mix languages in the same text. It is this variety of language switching that causes the problems and calls for an encoding model that facilitates the distinction.

Unification The problem, therefore, becomes one of appropriate rendering of what looks like plain text in a multi-style context. There are different mechanisms that, in theory could support the appropriate rendering (this does not mean perfect fidelity) of such communications.

The keyboard could interact with the application to indicate the language of the text, and the application could send out of band information regarding the language of the text requiring these pieces to be in place:

- Mechanism for browsers and other applications to query the keyboard as to the intended language the user is typing.
- Mechanism for web applications to access said keyboard information.
- A way to encode character styling in plain text.

This would require the whole communications chain to facilitate this out of band information: transmission, storage, etc. This is in the hands of the service providers. It would involve a change to Facebook, Twitter, SMS, chat, etc. and would require a huge level of industry commitment to see happen. This is highly doubtful given the low value of the business case in monetary terms. In addition there are various difficulties with this approach:

- People use the same keyboard to type different languages, e.g. English and Finnish or German, especially when typing language fragments.
- Do we really want to tag every string for language, with its inherent language management issues, just to support some minority language text styling?

Language marking would have to be stored within the text to keep the text plain (are the Plane 14 characters about to have their day?) and applications could then style appropriately. But the language markup characters were abandoned for a reason, and it is unlikely that those reasons have gone away. An alternative approach of using automatic language identification processes does not

help here since users will want the change in styling to be evident as they type the text and not just after it has been processed by some central system.

If plain text is to stay plain, then either text must be marked as to its styling (if the difference is considered a style difference) or that style must be inherent in the characters used.

Disunification carries potential problems. But it is expected that these problems would not arise in this particular case

There is no reason or expectation that users of the other languages (Burmese, etc.) would use or even want to use the disunified characters. Their form is so obviously inappropriate for these other languages, that people would not consider using them. Equally, there is no ambiguity on the part of the Khamti, etc. users, in that they always use this form and so would always use the disunified characters.

The characters that do not disunify but have their representative glyphs changed are only used in languages which use the dotted form. So changing them to have their correct default improves their definition.

An alternative to full disunification is to half disunify using Variation Selectors, which provides fallback rendering and would provide a much faster route to implementation. It may be that using Variation Selectors would allow for a less obvious disunification reflecting the etymology of the characters and would provide a gentler way of addressing the 'flood of need' if no other solution can be found.

The impact on existing Myanmar text of this disunification is minimal. There is, as yet, little text in these languages, using Unicode. There is sufficient interest in using Unicode, this situation will soon change. Existing text, using the Burmese codepoints, will continue to render as before using language specific styling. The only situation is where old text needs to interact with new text in terms of searching. Such long term texts will need to be transcoded with simple search and replace. The greater concern is how long the transition will take before implementation can effectively occur.

Proposal 1 One proposal is to solve the problem through disunification. The disunification creates 16 new characters thus requiring a new block allocation of 1 column. Since this is not a new script and the existing script is all in the BMP, it is proposed that the new columns come from the BMP. One option for finding such space is to place the block immediately preceding the existing Myanmar block, either as a new block or as a change to the existing block structure. While it is odd starting a major block not on a XX00 boundary, the BMP is rather full. This proposal will be written in terms of such a new block, but the block can and probably will be moved.

The new characters inherit their properties from the existing characters and will have compatibility decompositions to those characters.

```
0FF0;MYANMAR LETTER KHAMTI KA;Lo;0;L;<compat> 1000;;;;N;;;;;
0FF1;MYANMAR LETTER KHAMTI KHA;Lo;0;L;<compat> 1002;;;;N;;;;;
0FF2;MYANMAR LETTER KHAMTI NGA;Lo;0;L;<compat> 1004;;;;N;;;;;
0FF3;MYANMAR LETTER KHAMTI TA;L0;0;L;<compat> 1010;;;;N;;;;;
0FF4;MYANMAR LETTER AITON THA;L0,0,L;<compat> 1011;;;;N;;;;;
0FF5;MYANMAR LETTER KHAMTI PA;Lo;0;L;<compat> 1015;;;;N;;;;;
0FF6;MYANMAR LETTER KHAMTI MA;Lo;0;L;<compat> 1019;;;;N;;;;;
0FF7;MYANMAR LETTER KHAMTI YA;Lo;0;L;<compat> 101A;;;;N;;;;;
0FF8;MYANMAR LETTER KHAMTI LA;Lo;0;L;<compat> 101C;;;;N;;;;;
0FF9;MYANMAR LETTER AITON WA;Lo;0;L;<compat> 101D;;;;N;;;;;
0FFA;MYANMAR LETTER KHAMTI A;Lo;0;L;<compat> 1022;;;;N;;;;;
0FFB;MYANMAR VOWEL SIGN AITON E;Mc;0;L;<compat> 1031;;;;N;;;;;
0FFC;MYANMAR LETTER KHAMTI SHAN KA;Lo;0;L;<compat> 1075;;;;N;;;;;
0FFD;MYANMAR LETTER KHAMTI SHAN CA;Lo;0;L;<compat> 1078;;;;N;;;;;
```

```
0FFE;MYANMAR LETTER AITON SHAN NYA;Lo;0;L;<compat> 107A;;;;N;;;;;
0FFF;MYANMAR LETTER KHAMTI SHAN THA;Lo;0;L;<compat> 1080;;;;N;;;;;
```

There are 4 characters in the list which are not used by Khamti, where Khamti uses the undotted Burmese form, while Aiton and Phake use dots in their characters. The intent is that the Khamti would use only the codes required for Khamti, while the Aiton and Phake would use all of them. Also the Khamti style their characters slightly differently from Aiton and Phake. There is no intent to provide a plain text contrast between these languages.

The E Vowel sign is proposed for disunification on the grounds of keeping an appropriate style with the consonants.

In addition to the disunified characters, we propose updates to the representative glyphs of characters only used in Khamti style. The characters involved are: AA61, AA62, AA63, AA64, AA65, AA66, AA6B, AA6C, AA6F. The proposed revised glyphs are shown in black on the chart, with the existing glyphs shown in grey. While this does not necessarily cover all the styling changes for Aiton and Phake, showing the Khamti style indicates which glyphs would be changed for a typical font. The Aiton and Phake variants of Khamti, can be implemented without having to follow the representative glyphs slavishly. The communities are happy with this arrangement.

If this proposal is accepted, it is requested that the proposal be fast tracked due to the block change. The characters interact with others in other Myanmar blocks and therefore it is important that the block be allocated and implementations be updated to support the block as part of the Myanmar script otherwise text runs will be broken within a text.

Proposal 2 An alternative proposal is 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:

Replacement Glyph	Character Sequence	Alternate Glyph	Description
ꩠ	1000 FE00	ꩠ	myanmar letter khamti ka
ꩡ	1002 FE00	ꩡ	myanmar letter khamti kha
ꩢ	1004 FE00	ꩢ	myanmar letter khamti nga
ꩣ	1010 FE00	ꩣ	myanmar letter khamti ta
ꩤ	1011 FE00	ꩤ	myanmar letter aiton tha
ꩥ	1015 FE00	ꩥ	myanmar letter khamti pa
ꩦ	1019 FE00	ꩦ	myanmar letter khamti ma
ꩧ	101A FE00	ꩧ	myanmar letter khamti ya
ꩨ	101C FE00	ꩨ	myanmar letter khamti la
ꩩ	101D FE00	ꩩ	myanmar letter aiton wa
ꩪ	1022 FE00	ꩪ	myanmar letter khamti a
ꩫ◌	1031 FE00	ꩫ◌	myanmar vowel sign aiton e
ꩬ	1075 FE00	ꩬ	myanmar letter khamti shan ka
ꩭ	1078 FE00	ꩭ	myanmar letter khamti shan ca
ꩮ	107A FE00	ꩮ	myanmar letter aiton shan nya

Replacement Glyph	Character Sequence	Alternate Glyph	Description
ခ	1080 FE00	ခ	myanmar letter khamti shan tha
ဂ	AA60 FE00	ဂ	myanmar letter khamti ga dotted
ဈ	AA61 FE00	ဈ	myanmar letter khamti ca dotted
ဨ	AA62 FE00	ဨ	myanmar letter khamti cha dotted
ဈ	AA63 FE00	ဈ	myanmar letter khamti ja dotted
ဝ	AA64 FE00	ဝ	myanmar letter khamti jha dotted
ဉ	AA65 FE00	ဉ	myanmar letter khamti nya dotted
ဏ	AA66 FE00	ဏ	myanmar letter khamti tta dotted
ဂ	AA6B FE00	ဂ	myanmar letter khamti na dotted
ဟ	AA6C FE00	ဟ	myanmar letter khamti sa dotted
ဠ	AA6F FE00	ဠ	myanmar letter khamti fa dotted

The difficulty with this proposal is that for text in Khamti, a high proportion of letters would have a variation selector following them. No text processes would be any harder in such a situation, given there would have to be collation keys for each of the above variation selector sequences. Searching would have a consistent encoding in either case. The advantage is that fonts not designed for the sequences will continue to render something legible.

Keyboarding Given that there are too many characters to type both Burmese and Khamti to fit on a single keyboard, language based keyboard switching is almost inevitable. And even if a single keyboard were usable, there is a visual distinction that is clear and unambiguous.

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.

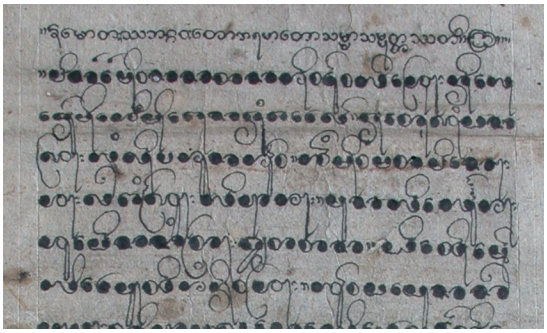

Illustration 1: Old style Phake with an initial line in Burmese.

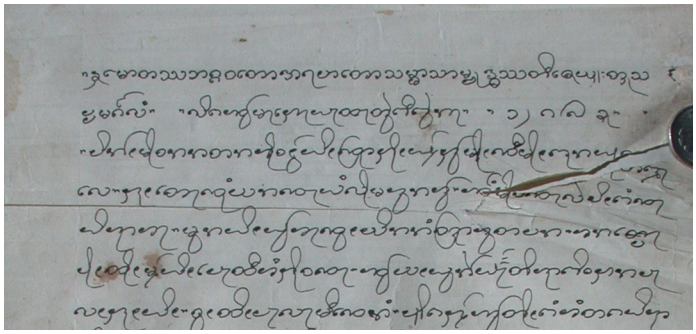

Illustration 2: Modern handwritten Phake, with Burmese heading

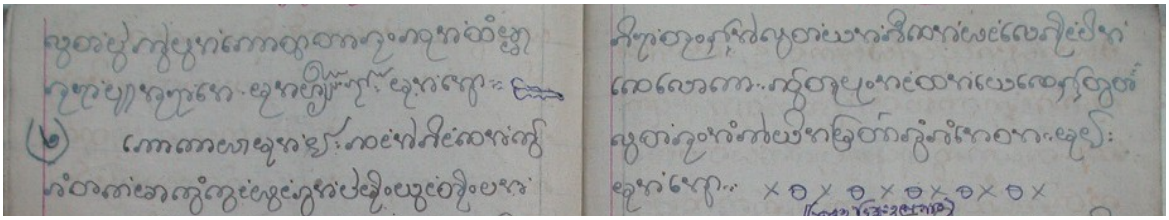

Illustration 3: Handwritten Khamti Shan

The following two illustrations show a plain text environment (Facebook) with a transcription of a text.

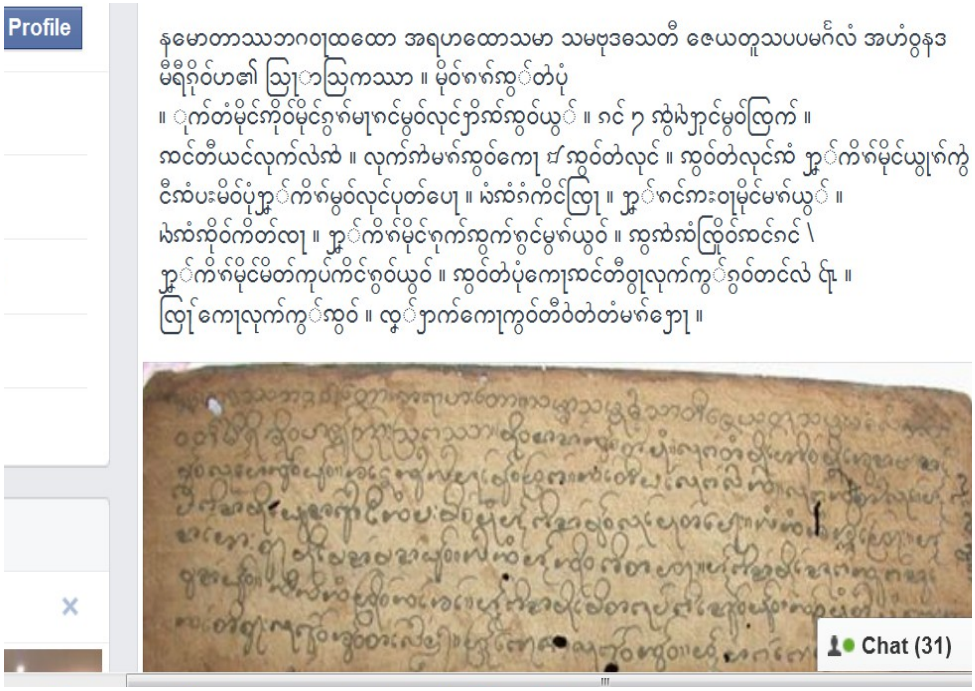

Illustration 4: Transcription using Burmese rendering and Padauk font. Notice the first two lines are Burmese text

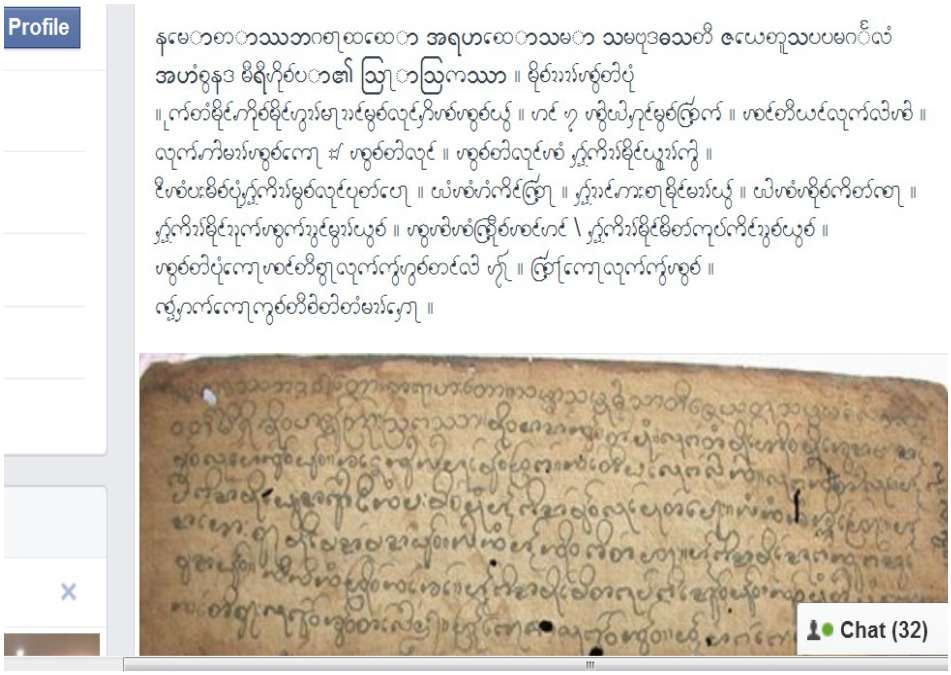

Illustration 5: Transcription using Aiton rendering and Aiton Unicode font. Notice the first two lines are Burmese text, even to font linking.

	0FF
0	ꩰ
1	ꩱ
2	ꩲ
3	ꩳ
4	ꩴ
5	ꩵ
6	ꩶ
7	꩷
8	꩸
9	꩹
A	ꩺ
B	ꩻ◌
C	ꩼ
D	ꩽ
E	ꩾ
F	ꩿ

	AA6	AA7
0	ꩠ	ꩰ
1	ꩡ	ꩱ
2	ꩢ	ꩲ
3	ꩣ	ꩳ
4	ꩤ	ꩴ
5	ꩥ	ꩵ
6	ꩦ	ꩶ
7	ꩧ	꩷
8	ꩨ	꩸
9	ꩩ	꩹
A	ꩪ	ꩺ
B	ꩫ	◌ꩻ
C	ꩬ	◌ꩼ
D	ꩭ	◌ꩽ
E	ꩮ	ꩾ
F	ꩯ	ꩿ

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[2]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** ***Khamti Variants***
2. Requester's name: *Martin Hosken*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *30/Apr/2014*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *X*
   (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters):
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *X*
      Name of the existing block: *Myanmar*
2. Number of characters in proposal: *16/0*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary X | B.1-Specialized (small collection) | B.2-Specialized (large collection)
   C-Major extinct | D-Attested extinct | E-Minor extinct
   F-Archaic Hieroglyphic or Ideographic | G-Obscure or questionable usage symbols
4. Is a repertoire including character names provided? *yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *yes*
   b. Are the character shapes attached in a legible form suitable for review? *yes*
5. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *SIL*
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *no*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *no*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *no*

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

2 Form number: N3102-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *no*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *yes*
   If YES, with whom? *Stephen Morey, Khamti community*
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *no*
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *common*
   Reference:
5. Are the proposed characters in current use by the user community? yes
   If YES, where? Reference:
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *yes*
   If YES, is a rationale provided? *yes*
   If YES, reference: *addition to existing BMP*
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *no*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *yes*
   If YES, is a rationale for its inclusion provided? *yes*
   If YES, reference: *This document*
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *no*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? *yes*
    If YES, is a rationale for its inclusion provided? *yes*
    If YES, reference: *this document*
11. Does the proposal include use of combining characters and/or use of composite sequences? *no*
    If YES, is a rationale for such use provided?
    If YES, reference: *no*
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *no*
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *no*
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility character(s)? *no*
    If YES, is the equivalent corresponding unified ideographic character(s) identified?
    If YES, reference: