L2/07-124

JTC1/SC2/WG2 N*

Date: 2007-04-22

A response to N3240 (dated 2007-04-11), the National Proposal (China) entitled "Proposal on Adding 3 Tibetan Characters and a symbol for ISO/IEC 10646 in BMP"

Submitted by: Robert R. Chilton (individual expert contributor)
Technical Director, Asian Classics Input Project
Email: acip@well.com

==========

The three new characters proposed in N3240 are:

1. "(TIBETAN ABOVE TRANSFORMED LETTER RAGU) to 0F6B"
2. "(TIBETAN ABOVE TRANSFORMED LETTER LA) to 0F6C"
3. "(TIBETAN ABOVE LETTER SA) to 0F6D"

These correspond to the three "head" letters/glyphs, of RA, LA, and SA, defined by Tibetan orthography and grammar. I will refer to these three as RA-head, LA-head, and SA-head.

The proposal N3240 asserts three reasons for why these three glyph-characters should be added to the BMP. In what follows I examine these three reasons and explain why the stated reasons are **not** sufficient to justify adding the proposed new characters.

It should be noted that international experts on Tibetan-language computing, including members of the Chinese national delegation, already considered whether or not to encode these three "head" letter glyphs as separate characters when the Tibetan script was added to the ISO/IEC 10646 BMP. The consensus of these experts was to adopt an encoding model for Tibetan script that represents the elements of Tibetan script according to their relative spatial position rather than their lexical meaning.[1] It was therefore decided at that time (1995) that no specifically defined lexical "head" characters are needed.

The above-referenced proposal, N3240, calls into question the earlier decision reached by international experts regarding the fact that there is no need for encoding specific Tibetan characters representing RA-head, LA-head and SA-head.

Here I will explain each of the reasons given in the N3240 proposal as to why these three characters are needed, together with analysis showing why these reasons are neither wholly valid nor sufficient to justify adding these proposed Tibetan characters.

---

[1] In forming syllable-like units, Tibetan letters can be combined in both left-to-right sequence and top-to-bottom sequence. For this reason, the BMP for Tibetan includes two sets of the consonant characters: in nominal position the letter follows to the right of the preceding character(s); in subjoined position the letter combines with and appears below the preceding character(s).

N3240 Proposal Reason #1. ***It is suggested that the proposed new characters are needed for demonstrating (i.e., in Tibetan-language textbooks) the orthographic structure of the script.***[2] Specifically, in teaching students how to write the script, there is a need to present, in isolation, the abbreviated "head" forms of the letters RA, LA and SA. Typically (i.e., in the "u-chen" script), the RA-head has a different shape from a nominal letter RA whereas the LA-head and SA-head are of the same shape as the corresponding nominal letters LA and SA but smaller in size.

While it might be tempting to define separate characters for this purpose, this is not strictly necessary. Similar needs for showing contextual presentation forms of various characters are encountered in other combining scripts such as Arabic. In such cases, the characters Zero-Width Joiner (U+200D) and Zero-Width Non-Joiner (U+200C) are used in combination with the existing characters of the script to produce contextual variant forms as needed.

Therefore, what is needed here are not three new Tibetan characters but rather a consensus on how to implement display of these needed "head" letter glyphs in their isolated presentation forms. This would be a simple matter of defining how Tibetan fonts should display the isolated presentation forms via a sequence of the desired letter, whether TIBETAN LETTER RA [U+0F62], TIBETAN LETTER LA [U+0F63], or TIBETAN LETTER SA [U+0F66], in combination with ZERO-WIDTH JOINER [U+200D].

Alternatively, these three "head" letter glyphs could be newly encoded explicitly and solely for such isolated presentation purposes. In this case, they should be clearly defined as standalone symbols and not as alphabetic letters. That is, they should be described as individual, spacing, non-combining symbols. This would prevent any misunderstanding or misuse that would wrongly treat these presentation-form glyphs (symbols) as alphabetic (letter) characters. Likewise, the names should clearly show that these characters are intended as standalone symbolic glyphs, i.e., as:

TIBETAN SYMBOL RA MGO
TIBETAN SYMBOL LA MGO
TIBETAN SYMBOL SA MGO

N3240 Proposal Reason #2. ***It is suggested that because the three "head" letters occur with high frequency in Tibetan, separate head-letter characters are desirable to support data entry such as keyboarding.***[3] The apparent rationale is that Tibetan keyboarding would be improved if there are separate dedicated keys for these head-letter characters.

---

[2] This is the presumed intent of passages in the N3240 proposal that mention "description of glyph" and "describe the characteristics of Modern Tibetan character".

[3] This is the presumed intent of passages in the N3240 proposal that mention "application software development" and "these 3 characters are high frequency character in the Modern Tibetan and usually used as a combining element to form Tibetan composite character".

It is easy to show that even if such a keyboard behavior is desired, there is no need to encode specific head-letter characters in the BMP in order to achieve it. Various systems for keyboard entry of Tibetan are already in use and deemed acceptable; and virtually any desired keyboard layout can be implemented without any need for defining new Tibetan characters.

Moreover, there are at least two strong arguments **against** this particular rationale:

Firstly, if the proposed new characters are encoded as combining characters, they would need to be typed ***after*** the "root" letter that they surmount or modify. This goes against the traditional manner in which words are spelled verbally and also against the order in which the letters are written out by hand. Traditionally the "head" letters are spelled and written ***before*** the "root" letters that they surmount or modify.

Secondly, from the perspective of data processing and interchange in digital environments, adding these proposed characters as Tibetan letters will create many unnecessary problems and complications. Current fonts and rendering systems for Tibetan follow the specification for the Tibetan script that has been in place since 1996. These fonts and rendering systems are now in widespread use in the Microsoft Windows, Macintosh OS X, and Linux operating environments. They operate on the assumption that vertical stacks of Tibetan letters will be encoded as a sequence of a single nominal letter followed by one or more combining subjoined letters. Adding the three proposed characters will force significant changes, and add increased complexity, for all Tibetan fonts and rendering systems since they would have to additionally recognize alternate (and lexically equivalent) sequences for Tibetan letter-stacks which could be encoded as a nominal letter, followed by zero or one "head" combining letter, followed by one or more combining subjoined letters.

Even more disconcerting, adding these proposed new characters will require additional steps of normalization between Tibetan data encoded using the current methodology (character sequences ordered according to relative spatial position) and Tibetan data encoded using the proposed new methodology (character sequences ordered according to both lexical and spatial parameters). Such normalization will require the re-ordering of characters in the data sequence. The current methodology encodes the lexical "head" letter in sequence prior to the "root" letter that it modifies; the "root" letter in turn is encoded as a combining character in subjoined position. In contrast, the proposed new methodology would encode the "root" letter first—in nominal position—followed by the lexical "head" letter which attaches as a combining character.

N3240 Proposal Reason #3. ***It is suggested that use of these proposed new characters will help in sorting (collation) of Tibetan data.***[4] Sorting Tibetan-script data in culturally expected order is not simple. Fortunately, the problem of machine collation of Tibetan data has already

[4] This is the clear intent of passages in the N3240 proposal mentioning "solve the problems with Tibetan sorting based on traditional Tibetan order" and "sort order of Tibetan character base on the root letter, so the rightly find out root letter of every Tibetan glyph, then Tibetan character has sorted is possible based on BMP".

been solved (more than once).[5] LDML collation data for sorting Tibetan is attached as Appendix A.

This supposed rationale has little if any merit. Beyond the fact that the sorting "problem" for Tibetan-character data has already been solved, adding the proposed new characters would do little to simplify the "problem". The general difficulty in sorting Tibetan is due to the fact that the "root" letter, which is given primary weight in determining the relative sort order of Tibetan syllables, can be preceded by one or two pre-radical letters.

Listed below are the 11 combinations of letter(s) that can occur ahead of the "root" letter:

ག <TIBETAN LETTER GA [U+0F42]>
ད <TIBETAN LETTER DA [U+0F51]>
བ <TIBETAN LETTER BA [U+0F56]>
མ <TIBETAN LETTER MA [U+0F58]>
འ <TIBETAN LETTER -A [U+0F60]>
ར <TIBETAN LETTER RA [U+0F62]>
ལ <TIBETAN LETTER LA [U+0F63]>
ས <TIBETAN LETTER SA [U+0F66]>
བར <TIBETAN LETTER BA [U+0F56], TIBETAN LETTER RA [U+0F62]>
བལ <TIBETAN LETTER BA [U+0F56], TIBETAN LETTER LA [U+0F63]>
བས <TIBETAN LETTER BA [U+0F56], TIBETAN LETTER SA [U+0F66]>

Although one might argue that adding the three proposed new characters would simplify sorting for three of the cases listed above, it would do nothing to address the other eight cases. Further, some of the most difficult syllables to parse with regard to determining the "root" letter do not involve any "head" letters. Here are two examples:

བགས་ <BA[U+0F56], GA[U+0F42], SA U+0F66>
མངས་ <MA[U+0F58], NGA[U+0F44], SA[U+0F66]>

The ambiguity lies in the fact that, by the rules of Tibetan grammar, either the first or the second letter could reasonably serve as the "root" letter in these syllables.

---

[5] For example, culturally expected sort order for Tibetan-character data has been implemented within Microsoft Windows Vista, Linux, and Mimer SQL. (LDML data for Tibetan collation is attached as Appendix A.) The schema described for sorting Tibetan-character data is universal and general in the sense that it works equally well whether the Tibetan-script data contains modern Tibetan, classical Tibetan, Dzongkha (Bhutanese), Ladakhi, Balti (Yige), foreign words transliterated into Tibetan script, or any combination of these. For an overview, see my presentation to the Tibetan Information Technology Panel at the 2003 meeting of the *International Association for Tibetan Studies* (IATS) entitled "Sorting Unicode Tibetan using a Multi-Weight Collation Algorithm" <http://www.columbia.edu/~ph2046/iats/it/Chilton_slides.pdf>.

Any solution to the "problem" of sorting Tibetan that proposes the encoding of new characters as a means of explicitly identifying the "root" letter would need to define at least five additional new characters, for the five prefix letters, in addition to the three head letters proposed in N3240. One can imagine other proposals, such as defining separate lexical characters for various of the letters in the Tibetan alphabet—depending on whether they represent a "prefix", "head", "root", "subjoined" or "suffix" letter—that might be set forth. However, such "solutions" are neither necessary nor desirable.

Finally, adding the three new characters proposed in N3240—or any other characters similarly defined according to lexical category—would actually hinder sorting to the extent that existing solutions for sorting Tibetan would need to be modified in order to account for the new and alternative way(s) to represent Tibetan syllables that would be introduced with such additions to the character repertoire.

Summary. The reasons given in N3240 do not support the contention that the current character repertoire for Tibetan is lacking and that three new "head" characters are required. However, the N3240 proposal reminds us that there is occasional need for displaying contextual presentation forms, in isolation, of the Tibetan letters RA, LA and SA. It remains to be decided whether this requirement can best be accomplished through (1) combining, in sequence, the nominal character for the letter together with Zero-Width Joiner, or else by way of (2) encoding three new **symbol** characters which would serve solely and explicitly for the purpose of displaying these contextual glyph-forms.

Appendix A
LDML collation for sorting Tibetan-character data

```
<collation>
<rules>
<reset>&#x0F40;</reset>
    <s>&#x0F88;&#x0F90;</s>
    <p>&#x0F51;&#x0F40;</p>
    <p>&#x0F56;&#x0F40;</p>
    <p>&#x0F62;&#x0F90;</p>
    <p>&#x0F63;&#x0F90;</p>
    <p>&#x0F66;&#x0F90;</p>
    <p>&#x0F56;&#x0F62;&#x0F90;</p>
    <p>&#x0F56;&#x0F66;&#x0F90;</p>
<reset>&#x0F41;</reset>
    <s>&#x0F88;&#x0F91;</s>
    <p>&#x0F58;&#x0F41;</p>
    <p>&#x0F60;&#x0F41;</p>
<reset>&#x0F42;</reset>
    <p>&#x0F51;&#x0F42;&#x0F42;</p>
    <p>&#x0F51;&#x0F42;&#x0F44;</p>
    <p>&#x0F51;&#x0F42;&#x0F51;</p>
    <p>&#x0F51;&#x0F42;&#x0F56;</p>
    <p>&#x0F51;&#x0F42;&#x0F5D;</p>
    <p>&#x0F51;&#x0F42;&#x0F60;</p>
    <p>&#x0F51;&#x0F42;&#x0F62;</p>
    <p>&#x0F51;&#x0F42;&#x0F63;</p>
    <p>&#x0F51;&#x0F42;&#x0F66;</p>
    <p>&#x0F51;&#x0F42;&#x0F74;</p>
    <p>&#x0F51;&#x0F42;&#x0F7A;</p>
    <p>&#x0F51;&#x0F42;&#x0F7C;</p>
    <p>&#x0F51;&#x0F42;&#x0FB1;</p>
    <p>&#x0F51;&#x0F42;&#x0FB2;</p>
    <p>&#x0F56;&#x0F42;&#x0F42;</p>
    <p>&#x0F56;&#x0F42;&#x0F51;</p>
    <p>&#x0F56;&#x0F42;&#x0F58;</p>
    <t>&#x0F56;&#x0F42;&#x0F7E;</t>
    <p>&#x0F56;&#x0F42;&#x0F5D;</p>
    <p>&#x0F56;&#x0F42;&#x0F60;</p>
    <p>&#x0F56;&#x0F42;&#x0F62;</p>
    <p>&#x0F56;&#x0F42;&#x0F7A;</p>
    <p>&#x0F56;&#x0F42;&#x0F7C;</p>
    <p>&#x0F56;&#x0F42;&#x0FB1;</p>
    <p>&#x0F56;&#x0F42;&#x0FB2;</p>
    <p>&#x0F56;&#x0F42;&#x0FB3;</p>
```

མགར
མགལ
མགུ
མགེ
མགོ
མྱ
མགྲ
འགག
འགང
འགད
འགན
འགབ
འགམ
<t>འགཾ</t>
འགའ
འགར
འགལ
འགས
འགི
འགུ
འགེ
འགོ
འགྱ
འགྲ
རྒ
ལྒ
སྒ
བརྒ
བསྒ
<reset>ང</reset>
དངག
དངང
དངན
དངར
དངུ
དངོ
མངག
མངན
མངའ
མངར
མངལ
མངོ
རྔ
ལྔ
སྔ
བརྔ

བསྔ
<reset>ཅ</reset>
གཅ
བཅ
ལྕ
བལྕ
<reset>ཆ</reset>
མཆ
འཆ
<reset>ཇ</reset>
མཇ
འཇ
རྗ
ལྗ
བརྗ
<reset>ཉ</reset>
<s>ྋྙ</s>
གཉ
མཉ
རྙ
སྙ
བརྙ
བསྙ
<reset>ཏ</reset>
<t>ཊ</t>
གཏ
བཏ
རྟ
ལྟ
སྟ
བརྟ
བལྟ
བསྟ
<reset>ཐ</reset>
<t>ཋ</t>
མཐ
འཐ
<reset>ད</reset>
<t>ཌ</t>
གདག
གདང
གདན
གདབ
གདམ
<t>གདཾ</t>
གདའ

གདར
གདལ
གདས
གདི
གདུ
གདེ
གདོ
བདག
བདམ
<t>བདཾ</t>
བདའ
བདར
བདལ
བདས
བདུ
བདེ
བདོ
མདག
མདང
མདན
མདའ
མདར
མདུ
མདེ
མདོ
འདག
འདང
འདད
འདན
འདབ
འདམ
<t>འདཾ</t>
འདཝ
འདའ
འདར
འདལ
འདས
འདི
འདུ
འདེ
འདོ
འདྲ
རྡ
ལྡ
སྡ
བརྡ

བལྡ
བསྡ
<reset>ན</reset>
<t>ཎ</t>
གནག
གནང
གནད
གནན
གནམ
<t>གནཾ</t>
གནཝ
གནའ
གནས
གནུ
གནོ
མནག
མནང
མནན
མནབ
མནམ
<t>མནཾ</t>
མནའ
མནར
མནལ
མནུ
མནེ
མནོ
རྣ
སྣ
བརྣ
བསྣ
<reset>པ</reset>
<s>ྉྤ</s>
དཔག
དཔང
དཔད
དཔའ
དཔར
དཔལ
དཔས
དཔུ
དཔེ
དཔོག
དཔོང
དཔོད
དཔོན

དཔོར
དཔྱ
དཔྲ
ལྤ
སྤ
<reset>ཕ</reset>
<s>ྉྥ</s>
འཕ
<reset>བ</reset>
དབག
དབང
དབད
དབན
དབབ
དབའ
དབར
དབལ
དབས
དབུ
དབེ
དབོ
དབྱ
དབྲ
འབག
འབང
འབད
འབན
འབབ
འབམ
<t>འབཾ</t>
འབའ
འབར
འབལ
འབི
འབུ
འབེ
འབོ
འབྱ
འབྲ
རྦ
ལྦ
སྦ
<reset>མ</reset>
<t>ཾ</t>
<t>ྂ</t>
<t>ྃ</t>

དམག
དམང
དམན
དམཝ
དམའ
དམར
དམས
དམི
དམུ
དམེ
དམོད
དམྱ
རྨ
སྨ
<reset>ཙ</reset>
གཙ
བཙ
རྩ
སྩ
བརྩ
བསྩ
<reset>ཚ</reset>
མཚ
འཚ
<reset>ཛ</reset>
མཛ
འཛ
རྫ
བརྫ
<reset>ཞ</reset>
གཞ
བཞ
<reset>ཟ</reset>
གཟ
བཟ
<reset>ཡ</reset>
གཡ
<reset>ར</reset>
<t>ཪ</t>
བརླ
<reset>ཤ</reset>
<t>ཥ</t>
གཤ
བཤ
<reset>ས</reset>
གསག

གསང
གསད
གསན
གསབ
གསའ
གསར
གསལ
གསས
གསི
གསུ
གསེ
གསོ
བསག
བསང
བསད
བསབ
བསམ
<t>བསཾ</t>
བསར
བསལ
བསི
བསུ
བསེ
བསོ
བསྭ
བསྲ
བསླ
<reset>ཧ</reset>
ལྷ
<reset>ི</reset>
<t>ྀ</t>
<reset>ེ</reset>
<t>ཻ</t>
<reset>ོ</reset>
<t>ཽ</t>
ྐ
ྑ
ྒ
ྔ
ྕ
ྖ
ྗ
ྙ
ྟ
<t>ྚ</t>
ྠ

<t>ྛ</t>
ྡ
<t>ྜ</t>
ྣ
<t>ྞ</t>
ྤ
ྥ
ྦ
ྨ
ྩ
ྪ
ྫ
ྭ
<t>ྺ</t>
ྮ
ྯ
ྰ
ྱ
<t>ྻ</t>
ྲ
<t>ྼ</t>
ླ
ྴ
<t>ྵ</t>
ྶ
ྷ
ྸ
<reset>༹</reset>
<s>྄</s>
<s>ཱ</s>
<s>༹</s>
<s>ཿ</s>
<s>྅</s>
<s>ྈ</s>
<s>ྉ</s>
<s>ྊ</s>
<s>ྋ</s>
<reset>ྲཱྀ</reset>
ཷ
<reset>ླཱྀ</reset>
ཹ
<reset>དགགས</reset>
<t>དགཊ</t>
<t>དགཌ</t>
<reset>བགགས</reset>
<t>བགཊ</t>

<t>བགཌ</t>
<reset>འགགས</reset>
<t>འགཊ</t>
<t>འགཌ</t>
<reset>དངགས</reset>
<t>དངཊ</t>
<t>དངཌ</t>
<reset>མངགས</reset>
<t>མངཊ</t>
<t>མངཌ</t>
<reset>གདགས</reset>
<t>གདཊ</t>
<t>གདཌ</t>
<reset>བདགས</reset>
<t>བདཊ</t>
<t>བདཌ</t>
<reset>མདགས</reset>
<t>མདཊ</t>
<t>མདཌ</t>
<reset>འདགས</reset>
<t>འདཊ</t>
<t>འདཌ</t>
<reset>གནགས</reset>
<t>གནཊ</t>
<t>གནཌ</t>
<reset>མནགས</reset>
<t>མནཊ</t>
<t>མནཌ</t>
<reset>དཔགས</reset>
<t>དཔཊ</t>
<t>དཔཌ</t>
<reset>དབགས</reset>
<t>དབཊ</t>
<t>དབཌ</t>
<reset>འབགས</reset>
<t>འབཊ</t>
<t>འབཌ</t>
<reset>དམགས</reset>
<t>དམཊ</t>
<t>དམཌ</t>
<reset>གསགས</reset>
<t>གསཊ</t>
<t>གསཌ</t>
<reset>བསགས</reset>
<t>བསཊ</t>
<t>བསཌ</t>

<reset>྿</reset>
ༀ
<reset>ໆ</reset>
་
<t>༌</t>
<s>།</s>
<s>༎</s>
<s>༏</s>
<s>༐</s>
<s>༑</s>
<s>༔</s>
</rules>
</collation>

==========