ISO/IEC JTC1/SC2/WG2 N3010

L2/05-338

Title	**Comments on N2985 – Balti Tibetan additions**
Source	Andrew C. West
Document Type	Expert Contribution
Date	25th October 2005

In N2985 **Proposal to add four Tibetan characters for Balti to the BMP of the UCS** four additional Tibetan letters are proposed for encoding in order to meet the needs of representing the Balti language in the Tibetan script :

0F6B TIBETAN LETTER KKA
0F6C TIBETAN LETTER KHHA
0F6D TIBETAN LETTER GHHA
0F6E TIBETAN LETTER RRA

Whilst this proposal is welcome, the encoding of two of these four characters is problematic, and should be subject to careful consideration.

The two proposed characters TIBETAN LETTER KHHA ཁྼ and TIBETAN LETTER GHHA ག྽ are composed from the letters KHA [U+0F41] and GA [U+0F42] respectively by the addition of TIBETAN MARK TSA -PHRU [U+0F39]. The justification for encoding these letters as precomposed characters is given as :

> These could in principle be composed with TSA -PHRU, but there is a serious problem with the canonical combining class of TSA -PHRU which makes it a problem to use. TSA -PHRU is added to a base letter when forming contractions and usually indicates TSA, TSHA, DZA, or ZA in the contraction or abbreviation. Here, we are talking about separate (new) letters for transcribing Urdu or Arabic words. Accordingly, the KHHA and GHHA should be considered atomic, as Tibetan TSA, TSHA, and DZA are.

As explained below, I do not believe that these are sufficient reasons to encode these two letters as individual characters.

Use of TSA -PHRU to form Letters

Whilst TSA -PHRU is frequently used to mark contractions and abbreviations, it is also used to form distinct letters in their own right. In particular the sounds [f] and [v], which do not occur in Tibetan, may be represented by the letters PHA [U+0F55] plus TSA -PHRU and BA [U+0F56] plus TSA -PHRU respectively when required in foreign names or loan words :

<0F55, 0F39> = ཕྼ FA

<0F56, 0F39> = བྼ VA

That the TSA -PHRU sign is intended to be used for the composition of letters such as the proposed Balti letters KHHA and GHHA is explicitly stated in the Unicode Standard 4.0 section 9.11 (my highlighting):

The sign U+0F39 TIBETAN MARK TSA -PHRU (tsa-'phru, which is a lenition mark) is the ornamental flaglike mark that is an integral part of the three consonants U+0F59 TIBETAN LETTER TSA, U+0F5A TIBETAN LETTER TSHA, and U+0F5B TIBETAN LETTER DZA. Although those consonants are not decomposable, this mark has been abstracted and may by itself be applied to "pha" and other consonants **to make new letters for use in transliteration and transcription of other languages**. For example, in modern literary Tibetan, it is one of the ways used to transcribe the Chinese "fa" and "va" sounds not represented by the normal Tibetan consonants. Tsa-'phru is also used to represent tsa, tsha, or dza in abbreviations.

Problems with Canonical Combining Class of TSA -PHRU

TIBETAN MARK TSA -PHRU has a combining class of 216, whereas Tibetan vowels have a combining class of 130 or 132. This means that when normalized, a sequence of Tibetan Consonant plus TSA -PHRU plus Vowel becomes reordered to Consonant plus Vowel plus TSA -PHRU. Versions of the Windows Uniscribe rendering engine which support Tibetan (version 1.473.4067.15 used to render Tibetan in this document) only correctly render TSA -PHRU when directly following a consonant, thus <0F41, 0F39, 0F72>  [KHHI]

renders correctly, but the normalized form <0F41, 0F72, 0F39>  [KHHI] does not.

Note that this is only a problem if TSA -PHRU is followed by a vowel, and only if the text has been normalized. Furthermore the inability to correctly render TSA -PHRU when separated from its base consonant by a vowel should be considered a defect in Uniscribe which Microsoft should be encouraged to fix. The **Principles and Procedures** [N2652] clearly state that short-term deficiency of rendering technology is not a sufficient reason to encode precomposed characters.

Furthermore this deficiency applies to all letters that are composed with TSA -PHRU, so if it is felt that KHHA and GHHA need to be encoded as individual characters because of the problem with canonical reordering, then the letters FA and VA (which are far more widely used than the Balti Tibetan letters) would also need to be encoded as individual characters. However, this would be problematic as the new characters would have no canonical equivalence to the existing character sequences that are used to represent these letters in existing Tibetan data.

Multiple Spellings

As it is possible using current technology to correctly render Tibetan KHA plus TSA -PHRU and GA plus TSA -PHRU (as long as the text is not normalized), accepting the letters KHHA and GHHA for encoding will introduce the possibility of multiple spellings, with some people using the new letters and some people using the existing base letters plus TSA -PHRU. This should be avoided if at all possible.
