L2/04-192

FEEDBACK ON PR-30: Encoding of Bangla Khanda Ta in Unicode

The crux of the matter is that ta-overt hasanta and Khanda ta “need to be distinguished in encoded representation” [PR-30, p.2, bullet 3]. Neither model A nor model B fully satisfies this crucial requirement, and hence, both are inadequate. This is why proposal L2/04-060 was withdrawn. In either model some instances of Khanda-ta would be encoded as <ta, virama>, making it impossible to search exclusively for Khanda-ta in any text. Assume that the following Bangla words occur in some text:

Word	*Encoding in models A and B*
তৎপর	<ta, ta, virama, pa, ra>
তত্র	<ta, ta, virama, ra>
পত্‌পত্‌	<pa, ta, virama, zwnj, pa, ta, virama, zwnj>

How would one search exclusively for Khanda ta in the text? Searching for <ta, virama> would pull out all three words from the text, and searching for <ta, virama, zwj> or <ta, virama, zwj, zwnj> would miss many Khanda ta’s (such as the one in তৎপর). This would be totally unacceptable to any end-user.

Model C (the model that is currently intended by the specification in section 9.2 of Unicode 4.0 for representing Khanda ta) is not susceptible to this problem. The problem arises only if we misconstrue section 9.2 by assuming that it endorses model A. However, model C, does have some of the other disadvantages listed on p.8 of PR-30 which make it unacceptable.

This leaves us with model D which requires Khanda ta to be encoded as a distinct character with a code point of its own. Let us examine the pros and cons of adopting model D.

PROS:

- Khanda ta is a distinct grapheme of the Bangla-Asamiya script as indicated by the fact that it is in contrastive distribution with other graphemes in the script, resulting in minimal pairs such as the following:

মত	/mOt/	“opinion”
মৎ	/mOt/	“by me”
*মত্‌প্রণীত	/mOtpronito/	(* indicates unacceptability/ungrammaticality)
মৎপ্রণীত	/mOtpronito/	“composed by me”

  Model D conforms to the standard convention of encoding each grapheme as a distinct abstract character. This convention has been followed without exception in encoding all Indian scripts.

- The aforementioned search-related problem disappears.

- No “overriding” ZWJ/ZWNJ needed, and therefore any problem associated with their use does not arise.

- Khanda ta treated on par with anusvar and visarga with which it forms a natural class: all three represent dead consonants, none is able to bear a matra or other modifier and none can conjoin with a following consonant.

CONS:

None, as far as I can see, but let’s consider those that have been listed in PR-30:

- “It abandons the current specification in Unicode 4.0 in relation to Khanda ta.” [PR-30, p.9, bullet 3].

  This is hardly a matter of any concern since the current specification, as construed in PR-30, will have to be revised anyway in view of the search-related problem raised in the first paragraph of this document. Furthermore, even the most advanced implementations based upon the “current specification” are broken. For example, this is how Microsoft’s Uniscribe renders a character sequence containing Khanda ta irrespective of whether it is encoded as <ta, virama> or <ta, virama, zwj>:

  উৎেকাচ

  The correct rendering is:

  উৎকোচ

  So, no acceptable implementation will need to be revised or abandoned as a result of adopting model D.

- “It introduces a new character to support a distinction that can be represented using existing mechanisms (albeit with some cost); this may lead to expectations that user communities can request characters for reasons related to cultural perceptions of a script rather than any technical requirements.” [PR-30, bullet 4]

  Though the objection is carefully worded to appeal to the UTC, it is in fact without any substance. Khanda ta should be encoded as a distinct character not because it is culturally perceived to be so but because it is in fact a distinct grapheme of the Bangla-Asamiya script, and because doing so would be the least expensive solution to the problems we are confronted with. The arguments presented here are technical and scientific. They have nothing to do with cultural perceptions (though the latter can NEVER be completely ignored in issues related to language and script).
  The claim that Khanda ta can be represented using existing mechanisms has hopefully been shown to be untenable by now.

- “It results in Khanda ta being represented as an entirely different spelling from other forms of ta. This breaks the phonological and graphological connection between various forms of ta, with some negative impact for searching and for presentation of text in work involving historical texts; e.g. switching presentation between Khanda ta and other presentations cannot be done by formatting but requires conversion of the data itself. (From a cultural perspective, this also breaks the historical connection between Khanda ta and other forms of ta.)” [PR-30, bullet 5]

Khanda ta should be represented with a spelling entirely distinct from the other forms of ta because it is known to be a distinct grapheme, and NOT a distributional variant of these other forms of ta. Phonological and historical considerations are of minimal relevance to encoding. This is why the Bangla anusvar is represented with a spelling distinct from the other form of the velar nasal <nga> and encoded with a distinct code point. This is why the three distinct sa’s in Bangla orthography are encoded as distinct abstract characters in Unicode despite the fact that there is only one sibilant phoneme underlying the three. This is why the two i’s in Bangla, <i> and <ii> are assigned distinct code points even though there is in fact only one phonological /i/ in the language. The same observation holds for the two u’s, <u> and <uu>.

The primary concern of encoding is certainly not to make explicit the phonological and historical connections between the phonemes and graphemes of a language/script. These concerns can be addressed AFTER the primary concerns of round-trip conversion and end-user acceptability have been satisfactorily resolved.

Gautam Sengupta
gsgju@yahoo.com