L2/09-274

Proposals for discussion on Bengali Script in Unicode Meeting dated 10-14 August, 2009.

(A) Most important drawback of current Bengali Unicode is the scheme to represent the 'compound characters' which phonetically represent combination of two, three and four consonants. The present scheme of rendering them are through the 'hasant'(09CD) code point. Thus, 'hasant' serves two purposes: one, to appear as 'hasant' at the end of a word or after a consonant ; two, as a joiner to form 'compound characters'. However, to render 'hasant' after a consonant, a 'Zero Width Non joiner' [ZWNJ] code should accompany the 'hasant' code. Yet there is a problem to generate two types of compounding of R(09B0) and Ya(09AF). To distinguish the two, another code named 'Zero Width joiner' [ZWJ] is employed.

The whole process is cumbersome, unnatural and takes more bits to represent Bengali text. Also, people working on NLP and OCR of Bengali will find it inconvenient to write algorithms with such scheme. Instead we propose the following two modifications, one of which may be accepted by the consortium:

Proposal 1 (Our earlier proposal):

1. The code for 'hasant' (09CD) should be used only to render hasant after a consonant.
2. The 'Zero Width joiner' [ZWJ] code should be used to construct the "compound characters' only (except 'Ya phala' to be rendered as follows.
3. Introduce a code for 'Ya phala' which when used after the code for a basic character will render 'Ya phala' after that character.

This scheme will take care of all the "compound character" formation in a more straightforward and compact way. If you are interested, I can give a more elaborate explanation.

Proposal 2 (Modified proposal):

1. Since a lot of documents have been already prepared using the hasant for conjunct formation, we can continue the practice. The practice of using ZWNJ for explicit rendering of hasant nay also continue.
2. However, as before, we propose to introduce a code point for “Ya phala” (which may create conjunct with consonant without needing ‘hasant’. Then the use of ZWJ becomes redundant. Thus we can make

অ + ্য + া --> অ্যা

ত + ্য + া --> ত্যা

র + ্ + য --> র্য

র + ্য --> র্য্য

দ + ্ + ZWNJ + গ --> দ্‌গ

ন + ্ + ত + ্ + র + ্য --> ন্ত্র্য

(B) We do not understand why there is a code point 09D7 assigned to a shape that is neither a character nor its modifier (eg vowel sign) form. The shape shown in your table is the right part of the vowel sign 'Ou kar'(09CC). But such right part alone has no utility (not even historically). I assume it has been introduced with the influence of Devanagari 'O-kar'(094B) which is visually identical.

(C) Many code points are unnecessarily filled by some old (historical) character or modifier signs which are not used. I discussed with scholars and none of them have found any running text containing 098C (li), 09C4 (Double ri-kar), 09E0 (Double ri), 09E1 (Double li), 09E2 (li-kar), 09E3 (double li-kar) We can place the code points for such shapes in the private use plane of Unicode, which is so empty. I am sure text already coded with them are negligibly small, that may be corrected easily.

(D). In the Unicode 5.0 book we noted a request for two form of Consonant-vowel ligature. We do not think that this is in right spirit of Unicode where characters are to be coded, not glyph shapes.

গ + ু --> গু

গ + ZWJ + ু --> গু

(E) In the Unicode space of Devanagari these is the code point 0950 for ‘Om’ character. This character, though rarely used, is there in Bengali too. We can reserve the code point 09D0 (at the corresponding position of Devanagari) for this character the Bengali script.

(F) A conjunct formed by 0915 followed by 0937 in Devanagari is considered as a basic character in Bengali called “Khinya”. Unlike Devanagari we should treat this as basic character and I propose to reserve the code point 09CA for the same.

(G) A symbol like upper comma has a different role than apostrophie in Bengali, that represents missed character in between. This is called ‘urdha comma’. Till 1970’s it was well used in text. We can introduce a suitable code point for the same.

(H) There is no explicit 'full stop' code point for Bengali. The points reserved for Devanagari 'single' and 'double' full stop (Danda) is to be used for Bengali also. Why separate code for single and double full-stop? There is no harm if single danda code is used twice to depict double stop (danda) and we can save one code point. In fact, the use of double danda is obsolete.

(I) Some discussion about acronym and other signs necessary is needed.

Such issues may be discussed and resolved in the August meeting. I shall be glad to hear form your experts also.

Regards.

Bidyut Baran Chaudhuri
Vice President,
Society for Natural Language Technology Research, kolkata