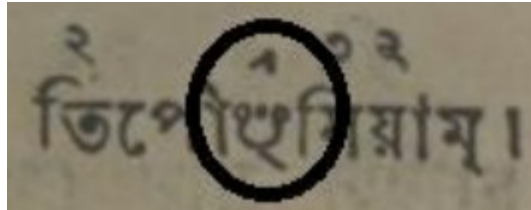


Encoding model issues with the Vedic gomukha characters

Shriramana Sharma, samjnaa-at-gmail-dot-com, India

2017-Aug-06

L2/17-098 p 12 shows a certain written form in the Bengali script:



... and analyses this as a sequence of 1CEA ঞ + 1CED ঞ্. The review committee report L2/17-255 p 8 questions this analysis and asks “what about the dot inside”.

In fact the passage is that of Kauthuma Sāma Gāna 551:

ā haryatāya dhṛṣṇave dhanuṣ ṭanvanti paumsnyam

... with the written form ঞ্ representing the anusvara indicated in bold in the last word. Already in L2/15-161R p 5 I have attested a Vedic anusvara character for Bengali for use in similar Kauthuma Sama Veda contexts:

বিশ্ববেদম্, হব্যবাহমমর্ত্যম্ ।

I had proposed that it be encoded as a script-specific character in comparison to A8F3 ঞ for Devanagari. Effectively, the shape ঞ্ is just a variant style of writing the same anusvara.

One notes that this exact same written form ঞ্ was attested in the GoI's Vedic doc L2/08-043 p 36 for the Yajur Veda and a variant with the candrabindu (moon-dot) instead of just bindu (dot) above was also attested for the same in the Vedic proposal L2/07-343 p 27:



Now the Vedic proposal had proposed that the gomukha characters should be encoded as generic characters in the Vedic Extensions block. On p 9 it clearly stated that these are to be used “with a bindu or candrabindu added on top”. However it did not specify which block these “on top” characters should be used from. Further, no mention of the tiryak (lit. “horizontal”) mark ँ was made though the attestation above clearly shows one. It was the GoI doc which proposed the tiryak mark separately based on the Yajur Veda attestation of ञ्. Later the revised Vedic proposal L2/08-273R3 included the tiryak mark from the GoI doc.

Neither the original nor the revised Vedic proposals seem to have tried to explicitly state an encoding model for these characters. However, the Vedic Extensions section of TUS 10.0 on p 504 of the PDF states:

The gomukha characters from U+1CE9..U+1CEC may be combined with U+0902 DEVANAGARI SIGN ANUSVARA or U+0901 DEVANAGARI SIGN CANDRABINDU. ... U+1CED VEDIC SIGN TIRYAK is the only combining character in this set of nasalization marks. While it appears similar to the U+094D DEVANAGARI SIGN VIRAMA, it is used to render glyph variants of nasal marks that occur in manuscripts and printed texts.

The part about the gomukha characters being combined with 0901 and 0902 is problematic. Since the gomukha-s are also seen in Bengali as demonstrated now, it was right to encode them in Vedic Extensions and not as Devanagari-specific characters. And using the tiryak from the generic block is also fine for Bengali. But what to do about the combining marks for Bengali?

Consider the options:

1. Use the Bengali anusvara and candrabindu-s. Not possible because the Bengali anusvara ং is not shaped as a dot above which is the desired shape.
2. Re-use Devanagari anusvara/candrabindu even amidst Bengali text. Quite unclean! Besides, Bengali has its own candrabindu at 0981 meaning that only the Devanagari anusvara should be re-used adding to the confusion.
3. Encode a separate dot above for Bengali. Possible, but a script-specific character for a generic shape which is not otherwise attested in the script doesn't seem appropriate.
4. Use 0307 DOT ABOVE and 0310 CANDRABINDU from the Combining Diacritics block. Possible, but it is not practice. Lookalikes between international diacritics and Vedic (dot below etc) have always been encoded separately.
5. Encode script=inherited dot above and candrabindu characters newly in the Vedic Extensions block. Possible, but will destabilize the existing recommendation for Devanagari (unavoidable?) and add two more lookalike characters to the standard.
6. A final option is to just encode a script=inherited dot-above character in Vedic Extensions because the candrabindu is uniform through all Indic scripts and only the anusvara shape differs in different scripts. The dot above will also complement the dots below 1CDD~1CDF.

The UTC should decide what is to be done in this situation.