L2/21-054

Interaction of Vedic svara markers with post-base spacing marks

Shriramana Sharma, samjnaa-at-gmail-dot-com, India
2021-Mar-13

దేవేభ్యఁ: అస్మభ్యఁం

There is a problem with the current encoding of Vedic text involving a combination of svara markers and post-base spacing marks. The spacing marks are mainly visarga characters and anusvara characters of Bengali and South Indian scripts.

The native user expectation is that these sequences should be encoded as LETTER_SYLLABLE + SVARA_MARKERS* + SPACING_MARK where LETTER_SYLLABLE is an independent vowel or [CONSONANT + VIRAMA]* + CONSONANT. The main point here is that the svara markers come between the syllable and the spacing mark.

However, the above is what happens with existing text shaping engines when one inputs such sequences: they are marked as illegal sequences by the insertion of dotted circle.

Please refer to the discussion at https://github.com/harfbuzz/harfbuzz/issues/2017. It says that the expected sequence is to put the svara markers *after* the spacing mark but the font and/or shaping engine should place it before the same.

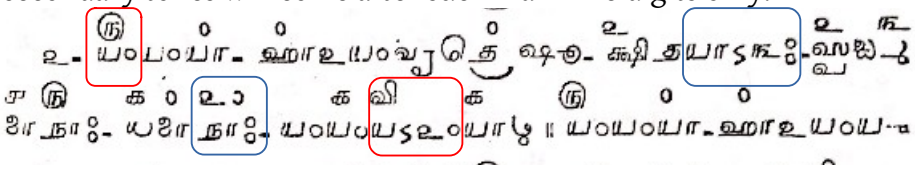
The same discussion shows that this expectation/behaviour began with old versions of Microsoft text shaping engines which possibly did the early work on supporting Indic. Apparently other engines feel obligated to mimic the same behaviour whether it is meaningful or not. They thus ask that it be resolve at the Unicode standard level before they are asked to support the sequence expected by native users.

Now the reasons we native users expect the svara markers to be input before the spacing mark are as follows:

1) First of all in an LTR script, the visual scanning of any writing identifies the above/below-base marks first and spacing post-base marks only after that. The same bug report above notes that in all other scripts above/below-base marks are expected to be input before post-base marks. This is obviously for this reason. Indic is the only aberration and there is no justification for the same.
2) It is self-evidently meaningless to insist on a text to be input in one order and then have the font or shaping engine to reorder them before displaying when there is no point in having them input in the wrong order in the first place.
3) In the case of reordrant marks such as Vowel Sign E in East and South Indic scripts, the reason for inputting them as CONSONANT + VOWEL_SIGN in contradiction to visual order VOWEL_SIGN + CONSONANT is that linguistically the vowel sound they represent comes after the consonant sound. In the present case, in fact the tone that is indicated by the svara marker applies to the vowel preceding it, and the spacing mark visarga or anusvara denotes a separate[1] sound after the vowel. Thus logically also the svara marker should precede the spacing mark.

1 I am aware that there are some descriptions of the anusvara as denoting nasality in Unicode related publications. Presumably this nasality is construed as belonging to the previous vowel. However, native linguistics recognizes the anusvara as an independent nasal sound similar to “m” and in many cases the written anusvara is used as a shortcut for representing the homorganic nasal sound of the following consonant. At any rate, the anusvara sign does not denote the nasality of the previous vowel. That is the role of the candrabindu.

4) Sama Vedic texts denote primary tones of vowels by above-base digits and secondary tones by digits placed on the main line after the vowel. Any visarga or anusvara after such secondary tones will come after such main line digits only:



Two visarga examples are marked in blue and two anusvara examples in red. In each pair, one shows a usage with secondary tone indicated by a main line digit and the other is a usage without secondary tone. It is illogical to declare that in the case where there is no secondary tone, the encoding should be LETTER_SYLLABLE + VISARGA/ANUSVARA + SVARA_MARKER and when there is a secondary tone, the visarga/anusvara should be at the end of the sequence.

5) There are rules described in Sama Vedic ancillary texts indicating how syllables with given primary tones will be converted to syllables with a sequence of primary and secondary tones in singing. Any Unicode-based implementation of such rules would need to unnecessarily special-case contexts where visarga/anusvara occur.

6) Likewise, implementation of sandhi rules is only straightforward when the anusvara/visarga are placed at the end. For instance, a visarga at the end of a word when followed by TA is converted to SA + VIRAMA, and an anusvara in the same context is converted to NA + VIRAMA.

देवेभ्यः + त्वा > देवेभ्यस्त्वा । तुभ्यं + ताः > तुभ्यन्ताः ।

A simple rule without special casing will do this sandhi both in the presence and absence of svara markers only if the visarga/anusvara are placed at the end.

7) Note that while most svara markers are above/below-base, there are two 1CE1 ◌᳁ and 1CF7 ◌᳷ which are post-base. If these need to be used with the spacing visarga/anusvara, it is all the more illogical to require them to be input *after* the visarga/anusvara but then display them *before*. Note that if these two characters were to be used in combination with any other above/below-base svara markers, they would be expected to follow such markers in input. Why should the visarga/anusvara be any different?

A concern was raised in the above bug report that declaring a new sequence will invalidate existing text corpuses. However, the problems above do need to be addressed and I do not see another way to do it. Further, there are not too many such corpuses and doing a regex search replace would be straightforward.

In fact overall Unicode-based Vedic support is still in its infancy with only the presence of requisite characters in the standard. Very few fonts actually provide glyphs for these characters. Smart font tables to properly support their use in the myriad combinations do not exist. Finally the support of text shaping engines for these combinations is rudimentary at best. Hence we still are at a point to make good design decisions and get them implemented.

Hence I request that the Unicode standard should recommend the usage of svara markers before any post-base visarga and anusvara. Note that for consistency, this should also apply to scripts like Devanagari where the anusvara is non-spacing. Smart font tables should take care of proper positioning of the various non-spacing marks in combination including any anusvara.

-o-