Public Comment on RZ-LGR 6 Update

SUBMITTER: Asmus Freytag, representing the Integration Panel **June 27, 2025**

Variant mappings between Devanagari Candrabindu or Halant and Bangla Candrabindu or Hasant respectively have been identified as missing during the String Similarity Evaluation project. These mappings should be added to this update of the RZ-LGR.

In contrast, the further claim that a variant mapping between Gurmukhi Nukta and Bangla Nukta is also missing, falls apart for the lack of any actual cross-script variant labels due to the constraints on the use of Nukta. Likewise, no variant labels would be possible with a Gurmukhi Virama.

Summary

The Integration Panel has received feedback from the String Similarity Evaluation project that U+0981 ◌ঁ BENGALI CANDRABINDU and U+0901 ◌ँ DEVANAGARI CANDRABINDU are not defined as variants despite being homoglyphs. Further, the viramas U+094D ◌् DEVANAGARI VIRAMA and U+09CD ◌্ BENGALI VIRAMA also are not defined as variants, even though applying them to the base consonants does not render them more distinct.

Admittedly, there are only a limited number of variants between these scripts, but this looks like an oversight. Our conclusion is that even small sets of variants should be complete and consistent. In terms of the eventual availability of labels, this correction changes little, because any paired labels with these code points would fail String Similarity Evaluation.

Even if string similarity does "catch" such cases at a later stage, that treatment is not consistent with similar cases, and not an automatic exclusion like blocked variants. Therefore, these pairs should be added to the pending update of the RZ-LGR.

Background On Missing RZ-LGR Variants for Candrabindu and Halant/Hasant

09AE + 09A1

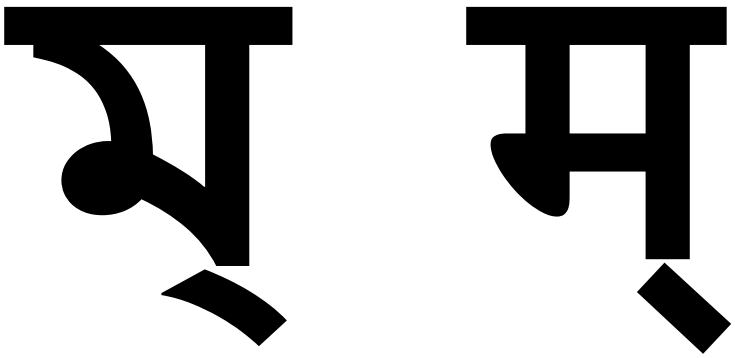
092E+0901

Above is the pair of existing consonants U+09AE and U+092E that are currently defined as variants in the RZ-LGR, each with the CANDRABINDU applied.

- Bengali base + Bengali candrabindu on the left
- Devanagari base + Devanagari candrabindu on the right.

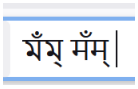
The candrabindus are actually stronger homoglyphs than the consonants.

Here are the same consonants, with a virama (halant) applied to each. Again, the virama is more of a homoglyph than the base consonants (which *are* defined as variants in the RZ-LGR).



09AE + 09CD 092E+094D

Here are some examples of possible combinations of all of these code points as seen in an address bar showing two hypothetical Bengali and Devanagari labels:



Because labels are not restricted to actual words, these cases should be treated like any other, similar cases and a variant definition defined that results in blocked variants.

Nothing in either context or WLE rules as specified in the respective LGRs would prevent labels that contain candrabindus from being variants. There is also an existing variant mapping from 0901 to 0945 0902, but it is a non-issue in this respect.

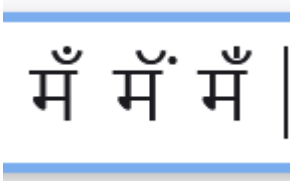
Existing RZ-LGR Variants Involving U+0901

U+0901 ँ DEVANAGARI CANDRABINDU is involved in some other variant relations. Here are some labels that are Devanagari in-script variants:

092E 0901 / 092E 0945 0902 / 092E 0945 093A

मँ / मॅं / मॅ◌ं

In MS Word, the variant using 0945 0902 is shown as unconnected, as is the one with U+093A substituted for U+0902, applying an existing in-script variant mapping. However, if these labels are entered in an address bar in Firefox, the variant with U+09A is shown connected.



Incidentally, on the Mac URL bar, 092E 0901 (left) and 092E 0945 0902 (middle) look identical, whereas 092E 0945 093A (right) doesn't render as combined:

From these examples it can be seen that the existing blocked variants were, in part, defined based on blocking labels that can sometimes render as homographs, even if that isn't consistent.

Nothing here seems to give rise to any issues in terms of these in-script variants conflicting with making a cross-script variant to Bengali. In particular 0902 is the Devanagari ANUSVARA, but that letter is very different in Bengali and therefore U+0902 does not have a cross-script variant.

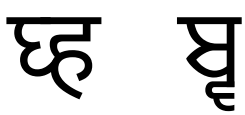
Why no Variant is Missing between Devanagari Halant and Gurmukhi Virama

In reviewing this, it was confirmed that U+094D Devanagari Halant and U+0A4D Gurmukhi Virama do not need a variant mapping. This example shows them applied to a pair of consonants, U+0918 and U+0A2C that are *existing* variants in the RZ-LGR. Adding a virama should not allow a label ending in these consonants to become distinct.

0918 094D 0A2C 0A4D

However, the RZ-LGR has a context rule for Gurmukhi that requires a following consonant from a short list. The example shown here is thus disallowed (as indicated with a star).

There is one possible valid conjunct using code points with cross-script variants and visually these conjuncts look very distinct across scripts:



Why no Variant is Missing for Bengali Nukta

In contrast, a claim that Bengali Nukta is missing RZ-LGR variant assignments with U+0A3C Gurmukhi Nukta falls apart once we look beyond simply the Nukta code points as such. Nukta is visually simply a dot below, and thus the 093C Devanagari Nukta, 09BC Bengali Nukta, 0A3C Gurmukhi Nukta, and finally 0B3C Oriya Nukta would all be homoglyphs of each other.

However, the use of Nukta is typically restricted to just a few consonants (or vowels in the case of Devanagari used for minority languages).

In this case, Bengali Nukta is only allowed with three consonants (the sequences are enumerated in the RZ-LGR): 09A1, 09A2 and 09AF. None of these consonants have any cross-script variants, so it isn't possible to form a variant label in any other script that looks like a Bengali label, if the latter contains a Nukta.

Curiously, the feedback does not make the same claim for U+093C the Devanagari Nukta or U+0B3C the Oriya Nukta. From the Bengali perspective, they also lack credible variant labels, particularly Oriya, which applies the Nukta to only two consonants, neither of which have any variants of their own.

The one exception is the *existing* RZ-LGR variant between U+093C Devanagari Nukta and U+0A3C Gurmukhi Nukta. There are many consonants in these scripts that have mutual cross-script variants and thus it is possible to create cross-script variant labels containing Nuktas (the placement of the dot is font dependent and thus not a consistent distinguishing element).

0918 093C 0A2C 0A3C

Other Examples of Existing Variants Involving Small Combining Marks

For presenting a more complete background, it is useful to consider other small combining marks in NeoBrahmi scripts, such as this case, which relates to an existing RZ-LGR variant definition.

Devanagari-Gurmukhi variant labels that would be blocked variants under existing RZ-LGR variant definitions:

0A08 092A 094D 091F 0940

Adding an Anusvara or Bindi does not make these labels distinct.

0A08 0A02 092A 094D 091F 0940 0902

It is clear that the precedent is to define such marks as variants whenever there's a variant pair of code points or sequences that allows the formation of cross-script variant labels.