Invariants between Indic positional and syllabic categories and general category

Norbert Lindenberg, 2023-06-25

Proposal

This document proposes to document several invariants between the Unicode properties Indic syllabic category (InSC), Indic positional category (InPC), and general category (gc), to correct a few property value assignments that currently violate these invariants, update a comment on two affected characters in IndicPositionalCategory.txt, and to add informative notes to two affected characters.

The following invariants should be documented in the introduction in IndicPositionalCategory.txt:

- The name of the InPC value contains "Left" or "Right" → gc is Mc or Lo.
- gc is Mc → the name of the InPC value contains "Left" or "Right".
- The name of the InPC value does not contain "Left" or "Right" → gc is Mn or Lo.
- gc is Mn → the name of the InPC value does not contain "Left" or "Right".
- gc is Lo and the name of the InPC value does not contain "Left" or "Right" → InSC is Consonant_Preceding_Repha or Consonant_Prefixed.

The following property value assignments need to be corrected:

- For U+1171E ◌ AHOM CONSONANT SIGN MEDIAL RA, change gc from Mn to Mc, and the Bidi_Class value from NSM to L to match the new gc value.
- For U+0D41 ◌ MALAYALAM VOWEL SIGN U and U+0D42 ◌ MALAYALAM VOWEL SIGN UU, change InPC from Right to Bottom.

The following comment in IndicPositionalCategory.txt should be updated:

- The note on U+0D41 ◌ MALAYALAM VOWEL SIGN U and U+0D42 ◌ MALAYALAM VOWEL SIGN UU in Notes section 1 should be moved to section 3 (the one covering "Bottom"), and changed to "U+0D41 MALAYALAM VOWEL SIGN U and U+0D42 MALAYALAM VOWEL SIGN UU are spacing marks to the right of the base in reformed Malayalam orthography.".

The following informative notes should be added in NamesList.txt:

- For U+0D41 ◌ MALAYALAM VOWEL SIGN U and U+0D42 ◌ MALAYALAM VOWEL SIGN UU, add informative notes "The representative glyph is for reformed orthography; the Unicode properties for traditional orthography".

Invariants

In general, the intent of the Indic positional category substrings "Left" or "Right" is to convey that the representative glyph requires space to the left or the right of the base character, which implies general category Mc. The absence of these substrings then conveys that the representative glyph requires no space to the left or the right of the base character, which implies general category Mn.

This simple equivalence is complicated somewhat by characters with Indic syllabic categories Consonant_Preceding_Repha or Consonant_Prefixed. These characters behave like combining marks, but are encoded before their base characters and therefore can't have general category Mn or Mc. They are given general category Lo instead. Characters with general category Lo are normally spacing, but when they have Indic syllabic categories Consonant_Preceding_Repha or Consonant_Prefixed, they can be nonspacing.

Character properties, comments, and informative notes

For U+1171E ◌ AHOM CONSONANT SIGN MEDIAL RA, it seems pretty clear that gc=Mn was a mistake. The representative glyph wraps around the base's top, left, and bottom, and this requires extra space on the left. Early Ahom proposals had gc=Mn; the one that was accepted, L2/12-309R, changed to gc=Mc but still had Bidi_Class=NSM, resulting in an invalid combination, and apparently that conflict was resolved the wrong way.

U+0D41 ◌ MALAYALAM VOWEL SIGN U and U+0D42 ◌ MALAYALAM VOWEL SIGN UU were affected by the reform of Malayalam orthography in the 1970es. As table 12-33 in section 12.9 of the Unicode Standard, version 15.0, shows, they're spacing marks in the reformed orthography of Malayalam, but form ligatures with consonants in the older orthography. The current representative glyphs are those of the reformed orthography, spacing marks to the right of the base, and this led to InPC=Right. In Unicode 1.0, the representative glyphs were only the circular components, sitting to the lower right of the base, and this may have influenced the choice of gc=Mn. Because the character has been encoded for such a long time, the Script Ad Hoc recommended not to change its general category, and instead to align the Indic positional category with the general category.

The proposed comment change reflects this change in InPC, and the proposed informative notes serve to explain why the data for these characters doesn't match the representative glyphs.