L2/20-294

Clarification Request on Malayalam /n̲t̲a/ Conjunct Specification Proposed in L2/19-345r2

Author: Cibu C Johny (cibucj@gmail.com)
Date: 2020 December 25
Action: For consideration by UTC, Script Ad Hoc, and Editorial Committees

Please see the following core specification text proposed in Alternative encodings for Malayalam “nta” (L2/19-345r2):

> ***Legacy Representations of Conjunct /n̲t̲a/***. Prior to Unicode 5.1 when <0D7B *chillu-n*, 0D4D *virama*, 0D31 *rra*> became the recommendation for the conjunct ന്റ /n̲t̲a/, two other representations <0D28 *na*, 0D4D *virama*, 0D31 *rra*> and <0D28 *na*, 0D4D *virama*, 200D ZWJ, 0D31 *rra*> were already in use. Due to slow updates to implementations, all three representations are widespread. It is recommended that implementations be prepared to treat <*na*, *virama*, *rra*> as an equivalent sequence of the recommended representation.
>
> The other legacy representation <*na*, *virama*, ZWJ, *rra*> conflicts with the legacy representation of <0D7B *chillu-n*, *rra*> (see “Legacy Chillu Sequences” later in this section), which represent the side-by-side form ൻറ. Therefore, implementations should treat <*na*, *virama*, ZWJ, *rra*> as a representation of ന്റ only when they know this sequence is not used to represent ൻറ.

The phrase ‘equivalent sequence of the recommended representation’ can be problematic as it does not further clarify whether newly generated text can use <*na, virama, rra*> to represent ന്റ /n̲t̲a/ or not. Without that, this phrase could be interpreted as both the sequences are practically equivalent in all respects including in newly generated text. That is an undesirable outcome as it is essentially recommending double encoding for /n̲t̲a/ conjunct. To prevent such interpretations, the Unicode core specification should make an additional effort to clarify that any newly generated text should be in the recommended representation. This has already been done for Bengali KHANDA TA and arguably for Malayalam *chillus*. As an example, please see the KHANDA TA documentation in Bengali Chapter 12.2:

> U+09CE BENGALI LETTER KHANDA TA should instead be used explicitly in newly generated text, but users are cautioned that instances of the older representation may exist.

Unlike Bengali KHANDA TA and Malayalam *chillus,* this potential double encoding outcome is not due to any specification change, but for legacy reasons only; so, this double encoding situation is avoidable. Therefore, I propose to clarify the above proposal text in the following way:

> ***Legacy Representations of Conjunct /n̪t̪a/***. Prior to Unicode 5.1 when <0D7B *chillu-n*, 0D4D *virama*, 0D31 *rra*> became the recommendation for the conjunct ന്റ /n̪t̪a/, two other representations <0D28 *na*, 0D4D *virama*, 0D31 *rra*> and <0D28 *na*, 0D4D *virama*, 200D ZWJ, 0D31 *rra*> were already in use. Due to slow updates to implementations, all three representations are widespread. It is recommended that implementations be prepared to treat <*na*, *virama*, *rra*> in existing text as an equivalent sequence of the recommended representation <0D7B *chillu-n*, 0D4D *virama*, 0D31 *rra*> which should instead be used explicitly in newly generated text.
>
> The other legacy representation <*na*, *virama*, ZWJ, *rra*> conflicts with the legacy representation of <0D7B *chillu-n*, *rra*> (see "Legacy Chillu Sequences" later in this section), which represent the side-by-side form ൻഠ. Therefore, implementations should treat <*na*, *virama*, ZWJ, *rra*> as a representation of ന്റ only when they know this sequence is not used to represent ൻഠ.

✺ END ✺