L2/19-345R2

Alternative encodings for Malayalam “nta”

മലയാളത്തിന്റെ “ന്റ”–യുടെ വിവിധ എൻകോഡിങ്ങുകൾ

To: Unicode Technical Committee
From: 梁海 Liang Hai <lianghai@gmail.com>
Date: 16 January 2020

1 Proposal

Add the following new sub-subsection right before the sub-subsection “Legacy Chillu Sequences” in the *Core Specification* (page 512 as of 12.0):

> ***Legacy Representations of Conjunct*** **/ṉṯa/.** Prior to Unicode 5.1 when <0D7B *chillu-n,* 0D4D *virama,* 0D31 *rra*> became the recommendation for the conjunct ന്റ /ṉṯa/, 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 *Core Specification* may also, at its discretion, further clarify that the two legacy representations are special cases and they do not suggest any productive rule in the encoding model of Malayalam.

2 Document history

Major changes since L2/19-345 (6 October 2019), the initial version of this document:

- Updated the proposed text in the section 1 for the *Core Specification*, taking into consideration the comments from both the discussion at UTC #161 and L2/19-348 (*Response to L2/19-345:* Alternative encodings for Malayalam “nta”, Cibu C Johny, 6 October 2019). Now the proposed text addresses both legacy representations and how exactly they should be treated as equivalences of the recommended one.
- Editorially improved the format of Table 1, *Encodings supported by platforms and fonts,* for better readability.

3 Background

There are a pair of related written forms that often cause confusion and difficulty, and the stacked form ൻ്റ is known as “nta”:

ൻറ   ൻ്റ

side-by-side vs. *stacked*

Graphically speaking, the side-by-side form ൻറ is ordinary, with two aksharas, a base ൻ *chillu n* [n] (typically encoded as U+0D7B ൻ MALAYALAM LETTER CHILLU N) and a base റ *rra* [r*a*, t*a*] (U+0D31 റ MALAYALAM LETTER RRA; italic [*a*] is inherent vowel). The stacked form ൻ്റ is graphically a single akshara, with a bottom-side sign of റ *rra* (post-base <0D4D ് VIRAMA, 0D31 റ RRA>) stacked under the base ൻ *chillu n*, then as a whole it should be encoded as <0D7B ൻ CHILLU N, 0D4D ് VIRAMA, 0D31 റ RRA> (*the graphic encoding*).

As Malayalam [r] has a plosive variant [t] that can surface when geminated or preceded by its homorganic stop [n], and graphic stacking emphasizes this alternation, the stacked form ൻ്റ explicitly represents [nt*a*] (and [nd*a*], if Dravidian free voicing is taken into consideration). The side-by-side form ൻറ is however ambiguous, representing either [nt*a*] or a literal [nr*a*].

3.1 A chillu-less analysis

Chillus are typically only written on their own as a standalone akshara, and can be alternatively understood as a preceding akshara’s right-side sign (comparable with ◌ം *anusvara* and ◌ഃ *visarga*).

Therefore, instead of being considered to be a graphic, productive composition between ൻ *chillu n* and റ *rra*, this unusual stacked form ൻ്റ [nt*a*] tends to be analyzed as a phonetic, irregular conjunct form between ന് *n* (ന *na* [n̪*a*, n*a*] with inherent vowel suppressed by ◌് *virama*) and റ *rra*, parallel to other conjuncts (see also the section 3.2, *Observations*, on page 8, L2/07-057) such as:

- ങ്ക *ng.ka* [ŋk*a*] = ങ് suppressed *nga* + ക *ka*
- ഞ്ച *ny.ca* [ɲt͡ʃ*a*] = ഞ് suppressed *nya* + ച *ca*
- ണ്ട *nn.tta* [ɳʈ*a*] = ണ് suppressed *nna* + ട *tta*
- ന്ത *n.ta* [n̪t̪*a*] = ന് suppressed *na* + ത *ta*
- മ്പ *m.pa* [mp*a*] = മ് suppressed *ma* + പ *pa*

Then the ന് *n* + റ *rra* conjunct would be systematically encoded as <0D28 ന NA, 0D4D ് VIRAMA, 0D31 റ RRA> (*the phonetic encoding*).

3.2 *Current encoding prescription*

As per the *Core Specification* 12.0 (paragraphs between Table 12-39 and Table 12-40, page 511), the encoding of ൻ്റ is graphic:

> <0D7B ൻ CHILLU N, 0D4D ് VIRAMA, 0D31 റ RRA>

However, the exact specification text talks about rendering, thus does not explicitly preclude alternative representations:

> *… The sequence <0D7B, 0D31> is rendered as* ൻറ*, regardless of the reading of that text. The sequence <0D7B, 0D4D, 0D31> is rendered as* ൻ്റ*. …*

Also, note that in addition to the now preferred atomic encoding U+0D7B ൻ MALAYALAM LETTER CHILLU N for ൻ *chillu n*, there is also a legacy, sequential encoding <0D28 ന NA, 0D4D ് VIRAMA, 200D ZWJ> (see the section "Legacy Chillu Sequences", page 512).

4 Early considerations and decision-making

It was part of the rationale for atomic chillu characters, that the stacked form ൻ്റ would need to be differentiated from the side-by-side form ൻറ at encoding level with a graphic analysis (an unusual sequence <*letter*, 0D4D ് VIRAMA, 200D ZWJ, 0D4D ് VIRAMA, *letter*> would be thus involved if atomic *chillu n* would not be available; see the section 7.16 on page 3–4, L2/06-207):

- Graphic encoding: <0D7B ൻ CHILLU N, 0D4D ് VIRAMA, 0D31 റ RRA>

The graphic encoding proposal received strong pushback from native-user experts, and many of them preferred a phonetic encoding, because of the phonetic analog of other conjuncts (see the section 3.1, *A chillu-less analysis*):

- Phonetic encoding: <0D28 ന NA, 0D4D ് VIRAMA, 0D31 റ RRA>

However their counterarguments were rather weak. Many failed to understand Unicode's fundamental graphic analysis, and kept arguing that it is wrong to append a virama (inherent vowel suppressor) to a chillu (pure consonant, naturally without an inherent vowel) because of some secondary analyses, such as (point 12, L2/08-038):

> *… Chillu's never form conjucts. All proposals for such definitions are linguistically incorrect (function of virama is to create vowel-less and you can't use it with a chillu because these are already vowel-less forms of the underlying consonants) …*

Even Cibu C. Johny at some point analyzed (the section "The need for correction", L2/07-393) in the same way:

> *… in the Indic model, Virama acts as the vowel remover for a consonant with default vowel /a/. The Chillus does not have an inherent vowel. So <chillu, virama> sequence could be violating the Indic model. …*

4.1 The hasty decision

In the midst of discussing various confusing topics including atomic chillu encoding, IDN (internationalized domain name) spoofing, ZWNJ/ZWJ restriction, multi-base implied akshara with left-side vowel sign (e.g., ൻറെ), and dot repha, the encoding issue of the stacked form ന്റ did not actually receive enough attention and clarification.

Eventually the consensus 113-C20 stood, and the graphic encoding became part of the *Core Specification* in Unicode 5.1.0 (April 4, 2008) under the section "Malayalam Chillu Characters".

4.2 Implementational difficulties

Several years later, the document L2/13-036 (Roozbeh Pournader and Cibu C. Johny) pointed out the problem that, by standardizing a seemingly helpful new encoding to replace an existing but unideal solution, "… software implementations are required to support both encodings of Malayalam chillus for eternity …". This is also relevant to the encoding issue of the stacked form ന്റ, as the phonetic encoding had already been working before the graphic analysis and encoding got standardized.

Furthermore, as the most influential platform, Windows never adapted its Malayalam OTL (OpenType Layout) shaper to allow the graphic encoding in an Indic cluster. This failure has greatly contributed to the graphic encoding's unpopularity.

5 Real-world encodings

The following five strings (including two control groups intended for different written forms) have been tested with major platforms and influential fonts:

- *Graphic* for ന്റ (current prescription):
  <0D7B ൻ CHILLU N, 0D4D ് VIRAMA, 0D31 റ RRA>
- *Phonetic* for ന്റ (chillu-less decomposition):
  <0D28 ന NA, 0D4D ് VIRAMA, 0D31 റ RRA>
- *Windows* for ന്റ (using legacy encoding for ൻ *chillu n*; requiring an additional U+200C ZERO WIDTH NON-JOINER after ZWJ for side-by-side form ൻറ; the seemingly alternative *Control 2* does not lead to the same rendering):
  <0D28 ന NA, 0D4D ് VIRAMA, 200D ZWJ, 0D31 റ RRA>
- *Control 1* for നറ:
  <0D28 ന NA, 0D31 റ RRA>
- *Control 2* for ൻറ (see also bullet for the *Windows* encoding):
  <0D7B ൻ CHILLU N, 0D31 റ RRA>

The test results are shown in the table below, with the influential fonts highlighted in yellow. The two control groups are omitted in the table as they did not exhibit unusual behavior in the test. In particular, the *Control 2* encoding for ൻറ does not have a ന്റ rendering with Nirmala UI or Kartika on Windows.

Table 1. Encodings supported by platforms and fonts

			Alternative encodings		
Platform		*Font*	*Graphic* ന്റ	*Phonetic* ന്റ	*Windows* ന്റ
Windows/DirectWrite, OTL (OpenType Layout)		**Nirmala UI**	*supported by font but not platform*		·
		Kartika			·
		any OTL font on this platform	*invalid cluster*	*okay*	*okay*
Android/HarfBuzz, OTL		**Noto Sans Malayalam**	·	·	·
		any OTL font on this platform	*okay*	*okay*	*okay*
iOS, macOS, … / Core Text	**AAT**	**Malayalam Sangam MN**		·	
		any AAT font on this platform	*okay*	*okay*	*okay*
	OTL	*any OTL font on this platform*	*okay*	*okay*	*okay*
***Other platforms*, OTL**		**Lohit Malayalam**		·	
		SMC fonts: **Meera, …**		·	

AAT is Apple Advanced Typography, which, unlike OTL, does not rely on shaper's script-specific knowledge. *SMC* is Swathanthra Malayalam Computing / സ്വതന്ത്ര മലയാളം കമ്പ്യൂട്ടിങ്ങ് (https://smc.org.in).

6 ICANN RZ-LGR situation

In ICANN's now published Root Zone Label Generation Rules (RZ-LGR) Version 3 for Malayalam (see "RZ-LGR-3-Element-LGR-MalayalamScript" on the page), there is a conflict involving the stacked form ന്റ:

- The original Malayalam RZ-LGR proposal suggests the phonetic encoding (<0D28 ന NA, 0D4D ് VIRAMA, 0D31 റ RRA>) should be used for the stacked form ന്റ and disallows the graphic encoding (<0D7B ൻ CHILLU N, 0D4D ് VIRAMA, 0D31 റ RRA>).
- However the eventually published Malayalam RZ-LGR normative XML specification accidentally allows both the phonetic and graphic encodings without variant control between the two (in the more readable HTML version, see rule "follows-C-or-0D41-or-0D7B" in the section 4.2, *Whole label evaluation and context rules*, and "Variant Set 8" in the section 3, *Variant Sets*).

ICANN is still in the process of investigating this issue.

7 Acknowledgements

Cibu C Johny and Santhosh Thottingal / സന്തോഷ് തോട്ടിങ്ങൽ kindly reviewed this document's a couple of revisions. Santhosh also translated the title into Malayalam.

The Malayalam font is Manjari / മഞ്ജരി (version 1.710) from SMC.

* EOF *