

Alternative encodings for Malayalam "nta"

മലയാളത്തിന്റെ “ന്റ”-യുടെ വിവിധ എൻകോഡിങ്ങുകൾ

To: Unicode Technical Committee
From: 梁海 Liang Hai <lianghai@gmail.com>
Date: 6 October 2019

1 Proposal

Append the following new paragraph to end of the sub-subsection “Special Cases Involving *rra*” in the *Core Specification* (page 511 as of 12.0):

Prior to Unicode 5.1 when <0D7B, 0D4D, 0D31> became the recommended representation, two other representations <0D28, 0D4D, 0D31> and <0D28, 0D4D, 200D, 0D31> already existed for the stacked ഞ. Due to slow updates to implementations, all three representations are widespread in existing text. Therefore, it is recommended that implementations be prepared to treat the two legacy representations as equivalent sequences of the Unicode 5.1 recommended representation.

The *Core Specification* may also, at its discretion, further clarify that the two additional representations are special cases and they do not suggest any productive rule in the encoding model of Malayalam.

2 Background

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

ന്റ ഞ
side-by-side vs. *stacked*

Graphically speaking, the side-by-side form ന്റ is ordinary, with two aksharas, a base ഞ *chillu n* [n] (typically encoded as U+0D7B ഞ MALAYALAM LETTER CHILLU N) and a base റ *rra* [ra, ta] (U+0D31 റ MALAYALAM LETTER RRA; italic [a] is inherent vowel). The stacked form ഞ is graphically a single akshara, with a bottom-side sign of റ *rra* (post-base <0D4D റ VIRAMA, 0D31 റ RRA>) stacked under the base ഞ *chillu n*, then as a whole it should be encoded as <0D7B ഞ CHILLU N, 0D4D റ 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 ഞ explicitly represents [nta] (and [nda], if Dravidian free voicing is taken into consideration). The side-by-side form ന്റ is however ambiguous, representing either [nta] or a literal [nra].

2.1 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 ള *anusvara* and ഴ *visarga*).

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

- റ്റ *ng.ka* [ṅka] = ി suppressed *nga* + റ *ka*
- റ്റ *ny.ca* [ntʃa] = ി suppressed *nya* + റ *ca*
- റ്റ *nn.tta* [ntta] = ി suppressed *nnā* + റ *tta*
- റ്റ *n.ta* [nta] = ി suppressed *na* + റ *ta*
- റ്റ *m.pa* [mpa] = ി suppressed *ma* + റ *pa*

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

2.2 Current encoding prescription

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

<0D7B റ 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 റ, regardless of the reading of that text.
The sequence <0D7B, 0D4D, 0D31> is rendered as റ്റ. ...

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

3 Early considerations and decision-making

It was part of the rationale for atomic chillu characters, that the stacked form റ്റ would need to be differentiated from the side-by-side form റ 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 section 7.16 on page 3-4, [L2/06-207](#)):

- Graphic encoding: <0D7B റ 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 section 2.1, *A chillu-less analysis*):

- Phonetic encoding: <0D28 ഐ NA, 0D4D ഏ VIRAMA, 0D31 ഐ 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 conjuncts. 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 (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. ...

3.1 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., ഐ), and dot repha, the encoding issue of the stacked form ഐ 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 [section “Malayalam Chillu Characters”](#).

3.2 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 ഐ, 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.

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

The control groups are omitted in the table as they did not exhibit unusual behavior in the test. Especially, the *Control 2* encoding for റ്ററ does not have a റ്റ rendering with Nirmala UI or Kartika on Windows.

Table 1. Encodings supported by platforms and fonts

Platform	Font	Alternative encodings		
		Graphic റ്റ	Phonetic റ്റ	Windows റ്റ
Windows/DirectWrite, OTL (OpenType Layout)	Nirmala UI	supported by font but not platform		.
	Kartika			.
	any font	invalid cluster	okay	okay
Android/HarfBuzz, OTL	Noto Sans Malayalam	.	.	.
	any font	okay	okay	okay
iOS, macOS, ... / Core Text	AAT Malayalam Sangam MN			.
	OTL any font	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>).

5 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 റ്റ:

- The original Malayalam RZ-LGR [proposal](#) suggests the phonetic encoding (<OD28 ഩ NA, OD4D റ് VIRAMA, OD31 റ RRA>) should be used for the stacked form റ് and disallows the graphic encoding (<OD7B റ് CHILLU N, OD4D റ് VIRAMA, OD31 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-OD41-or-OD7B” in section 4.2, *Whole label evaluation and context rules*, and “Variant Set 8” in section 3, *Variant Sets*).

ICANN is still in the process of investigating this issue.

6 Acknowledgements

Cibu C. Johny / കിബു സി. ജോണി 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 *