Alternative encodings for Malayalam “nta”

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 Chillum Sequences” in the Core Specification (page 512 as of 12.0):

Legacy Representations of Conjunct /nta/. Prior to Unicode 5.1 when <OD7B chillu-n, 0D4D virama, 0D31 rra> became the recommendation for the conjunct ൙ /nta/, two other representations <OD28 na, 0D4D virama, 0D31 rra> and <OD28 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 <OD7B chillu-n, rra> (see “Legacy Chillum Sequences” later in this section), which represent the side-by-side form ൙. Therefore, implementations should treat <na, virama, ZWJ, rra> as a representation of ൙ only when they know this sequence is not used to represent ൙.

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 ൻി is known as “nta”:

\[
\begin{array}{c}
\text{side-by-side} \\
\text{vs.} \\
\text{stacked}
\end{array}
\]

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

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 ൻ anusvaram and ൻ visarga).

Therefore, instead of being considered to be a graphic, productive composition between ൻ chilu n and ൻ rra, this unusual stacked form ൻ [nta] tends to be analyzed as a phonetic, irregular conjunct form between ൨ n (usahaan, na) with inherent vowel suppressed by ൨ ‘virama) and ൻ rra, parallel to other conjuncts (see also the section 3.2, Observations, on page 8, L2/07-057) such as:

- ൨ ng.ka [ŋka] = ൨‘ suppressed nga + ൨ ka
- ൨ ny.ca [ɲa] = ൨‘ suppressed nya + ൨ ca
- ൨ nn.tta [ṇta] = ൨‘ suppressed nna + ൨ tta
- ൨ n.ta [ṇta] = ൨‘ suppressed na + ൨ ta
- ൨ m.pa [mpa] = ൨‘ suppressed ma + ൨ pa

Then the ൨ n + ൻ rra conjunct would be systematically encoded as <0D28 ൨ NA, 0D4D‘VIRAMA, 0D31 ൻ 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 ೀ is graphic:

\[<0D7B Glyph CHILLU N, 0D4D^`VIRAMA, 0D31 o rra>\]

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

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

Also, note that in addition to the now preferred atomic encoding U+0D7B ೀ MALAYALAM LETTER CHILLU N for ೀ chillu n, there is also a legacy, sequential encoding <0D28 ೀ 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 ೀ would need to be differentiated from the side-by-side form ೀ 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 ೀ CHILLU N, 0D4D^`VIRAMA, 0D31 o 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 ೀ NA, 0D4D^`VIRAMA, 0D31 o 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 (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., എൻറ), and dot repha, the encoding issue of the stacked form ങ്ങ്ങ്ങ 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 ങ്ങ്ങ, 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 ങ്ങ (current prescription):**
  `<\0D79 ങ CHILLU N, \0D4D`virama, \0D31 ങ RRA>

- **Phonetic for ങ (chillu-less decomposition):**
  `<\0D2B ങ NA, \0D4D`virama, \0D31 ങ RRA>

- **Windows for ങ (using legacy encoding for ങ chillu n; requiring an additional U+200C ZERO WIDTH NON-JOINER after ZWJ for side-by-side form ങ; the seemingly alternative Control 2 does not lead to the same rendering):**
  `<\0D2B ങ NA, \0D4D`virama, 200D ZWJ, \0D31 ങ RRA>

- **Control 1 for ങ:**
  `<\0D2B ങ NA, \0D31 ങ RRA>

- **Control 2 for ങ (see also bullet for the Windows encoding):**
  `<\0D79 ങ CHILLU N, \0D31 ങ 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 ങ does not have a ങ rendering with Nirmala UI or Kartika on Windows.
Table 1. Encodings supported by platforms and fonts

<table>
<thead>
<tr>
<th>Platform</th>
<th>Font</th>
<th>Graphic ciągę</th>
<th>Phonetic ciągę</th>
<th>Windows ciągę</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows/DirectWrite, OTL (OpenType Layout)</td>
<td>Nirmala UI</td>
<td>supported by font but not platform</td>
<td>·</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kartika</td>
<td>·</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>any OTL font on this platform</td>
<td>invalid cluster</td>
<td>okay</td>
<td>okay</td>
</tr>
<tr>
<td>Android/HarfBuzz, OTL</td>
<td>Noto Sans Malayalam</td>
<td>·</td>
<td>·</td>
<td>·</td>
</tr>
<tr>
<td></td>
<td>any OTL font on this platform</td>
<td>okay</td>
<td>okay</td>
<td>okay</td>
</tr>
<tr>
<td>iOS, macOS, ... / Core Text</td>
<td>Malayalam Sangam MN</td>
<td>·</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>any AAT font on this platform</td>
<td>okay</td>
<td>okay</td>
<td>okay</td>
</tr>
<tr>
<td>Other platforms, OTL</td>
<td>Lohit Malayalam</td>
<td>·</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SMC fonts: Meera, ...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AAT is Apple Advanced Typography, which, unlike OTL, does not rely on shaper’s script-specific knowledge. SMC is Swathanthra Malayalam Computing / സ്വാതന്ത്ര്യ മലയാളം കംപ്യുട്ടേഴ്സ് (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 닙:

- The original Malayalam RZ-LGR proposal suggests the phonetic encoding (<0D28 0D NA, 0D4D“VIRAMA, 0D31 0 RRA>) should be used for the stacked form 닙 and disallows the graphic encoding (<0D7B 0D CHILLU N, 0D4D“VIRAMA, 0D31 0 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 kindly reviewed this document’s a couple of revisions. Santhosh also translated the title into Malayalam.

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

* EOF *

* EOF *