Response of IP to Malayalam Draft LGR of 11th May 2018

DATE: 2018-05-24

# Overview

This document provides IP response to the Malayalam LGR proposal dated 2018-05-11.

# General Comments

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| Current situation | This marks the first response by the IP to a draft LGR for the Malayalam script. We understand that this is an initial draft that may therefore take significant edits and additional rounds of feedback before completion. |  |
| Font Size | A general problem with the Proposal as submitted is the size of font used for Malayalam. It makes the Malayalam examples hard to read for those unfamiliar with the script. Although it is formally 10-point (comparable with the English script used) in fact it seems less than half the height as rendered. (And of course, Malayalam characters are typically much more intricate than Latin characters.) Also, wherever possible, code points should be provided as well. | Please fix. |
| § 3.1 | Please change the reference into a standard numbered reference with full entry in the Reference section | Please fix. |
| § 3.1 Joiner | Zero-width joiner and Zero-width non-joiner are not part of the MSR, because the Procedure rules out any CONTEXTJ code points.  However, these code points are commonly used in Indic scripts to control the formation of conjuncts. That means that not incorporating them into an LGR prevents certain terms from being displayed and/or represented correctly.  We do have in all NeoB LGRs a part (§ 3) that purports to give the background information for the script; these chapters go into great details on things like the early history, but are curiously silent on the joiner characters.  We understand that there is a placeholder for this information. | Please complete the planned text. |
| §4 Languages considered | 1. If Malayalam is the only relevant language, then this can be stated in a single sentence (w/o EGIDS and using table). Otherwise, please complete the information. 2. It might be useful to have a subsection on “The Structure of Written Malayalam,” containing tables of the different classes of character (vowels and characters) plus anything distinctive about ways they are structured or presented. This would be particularly helpful since it appears that there has been some change in the system of characters in practical use, so the GP should show which system it supports. (As it is, we only get two schematic paragraphs on "Orthography reform" in sub-section 3.1, without any specifics.) 3. The  CHILLU series could usefully be introduced here, and shown in interaction with VIRAMA and possibly also ZWJ/ZWNJ (these have been promised already); it would also be useful to see the extent of conjunct consonants in use, and how the GP's proposal represents them. Also how does "CHILLU" relate to "CHILLAKSHARAM" and "SAMVRUTHOKARAM"? 4. Also, we need explanation of what consonant diacritics are (as per sub-section 5.8) and what samvruthokarams (as per sub-section 5.9) 5. The content of the tables in subsections 5.3 Vowels thru 5.9 Samvruthokarams could be incorporated here to separate description of the writing system from description of the repertoire of the LGR | Please consider these suggestions and make appropriate edits. |
| § 5, Table of repetoire | 1. The introductory text appears to be the wrong one. Expected: a short introduction introducing the selection of code points and exclusions documented in Section 5. 2. Unicode general category column: the data in this column are not very interesting in the context of Malayalam, especially as more relevant data (Syllabic category) is provided. Suggest removing the Colum to make more space in the table. 3. References for code points: the reference given is the MSR – what was except is an independent reference documenting code points in modern use. 4. It would be useful to include a table of the current MSR representation of Malayalam block, before the major tabulated repertoire. | Please edit |
| § 5.2 | 1. Give explanation of why VOCALIC L, VOCALIC RR and LETTER NNNA are excluded. | Please provide explanation |
| § 5.6 Chillu letters | 1. No code points are given 2. Chillus have a pre-Unicode 5.0 encoding using sequences using Joiners. This older encoding is still prevalent in data, such as corpora and may even be in current use. The basic facts of this change of encoding should be mentioned.   C:\Users\asmusf\AppData\Local\Temp\okeeilcanoeinale.png   1. It is understood that the sequences are ruled out because the root does not allow joiners, so there is no issue with duplicate encoding (this could be mentioned). 2. However, because the atomic encoding for chillus is apparently not universally used, there should be a prominent note that the RZ-LGR only allows the atomic encoding. 3. A citation of and reference to the relevant chapter (12) in the Unicode standard. 4. IP notes that many references to Chillu-related issues exist in the references section, and they were very helpful in understanding the issue. However, none of them were cited from Section 5.6. | please provide code points and extend description as suggested |
| §6.1 | Give motivation for the variant sets (here repeated in the way they will appear in the LGR HTML document).   1. **Variant Set 10 — 3 Members**  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Source** | **Glyph** | **Target** | **Glyph** |  | **Type** | | [0D28 0D4D 0D31](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Malayalam-Script-2018-05-20-en.html#0D28_0D4D_0D31) | ന്റ | [0D7B 0D31](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Malayalam-Script-2018-05-20-en.html#0D7B_0D31) | ൻറ | ↔ | blocked | | [0D28 0D4D 0D31](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Malayalam-Script-2018-05-20-en.html#0D28_0D4D_0D31) | ന്റ | [0D7B 0D4D 0D31](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Malayalam-Script-2018-05-20-en.html#0D7B_0D4D_0D31) | ൻ്റ | ↔ | blocked | | [0D7B 0D31](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Malayalam-Script-2018-05-20-en.html#0D7B_0D31) | ൻറ | [0D7B 0D4D 0D31](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Malayalam-Script-2018-05-20-en.html#0D7B_0D4D_0D31) | ൻ്റ | ↔ | blocked |   The IP notes that for many systems the sequence 0D7B + Halant does not render properly (shown with a dotted circle). Would it not be better to disallow?  Here’s the rendering on the latest Safari, which differs from that used in some versions of MS Word or other browsers (also note the dotted circles):    Note that LGR-2 disallows some grammatical sequences in Thai, because of lack of universal support in rendering. So there is a precedent.    Also, the various displayed shapes are not altogether that close in appearance, so the IP would normally question the choice of making these variants.   1. **Variant Set 11 — 2 Members**  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Source** | **Glyph** | **Target** | **Glyph** |  | **Type** | | [0D33 0D33](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Malayalam-Script-2018-05-20-en.html#0D33_0D33) | ളള | [0D33 0D4D 0D33](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Malayalam-Script-2018-05-20-en.html#0D33_0D4D_0D33) | ള്ള | ↔ | blocked |   While the IP agrees that the two “Glyphs” while not 100% identical are close enough to warrant variant treatment, there is the issue that the pattern “CC is variant of CHC” creates more complexity than might otherwise appear:  A sequence CCC can have as variants CHCC CHCHC and CCHC. This sort of combinatorial explosion can perhaps be tolerated for “blocked” variants, but only if there is no alternative. The IP would like to understand (from language to be added to the proposal) why oD33 0D4D 0D33 needs to be permitted. (This may be obvious to a native reader, but needs to be made explicit). |  |
| §6.2 | The document states that variants are defined for code points that “look somewhat similar ” to characters in other scripts.  This is too low a bar. The actual requirement for defining variants is for code points that users may substitute for one another as equivalent. Where the equivalence is based on appearance, this means identical, or very nearly identical appearance, such that, for example, a variant label entirely constructed from the Tamil cr0ss-script variants would be taken by the user without question as a Malayalam label.  The issue is with the description more than the selection of code points. In all cases listed, it would probably be fair to say that the code points either look identical or very near identical. | Please fix opening paragraph of Section 6.2 to better reflect the actual criteria for selecting these variants. |
| References | 1. None of the references provided are actually cited from anywhere in the document. 2. Some reference descriptions appear split across two reference items (for example 102 and 103) 3. For websites, please indicate the name of the Website (e.g. Omniglot) as well as page title in addition to the URL. As appropriate, prove an “accessed on” date. If a document has an identified author, publishing date, or serial number, please list these, even if that information is somehow also encoded in the URL. 4. Reference [131] is not a static link, but a search expression. Please replace with a proper reference and URL 5. the document <https://www.unicode.org/L2/L2005/05213-samvruktokaram.pdf> seems to address Samvruthokaram and Chandrakkala which otherwise appear somewhat mysterious to the non-native reader. Perhaps useful to add as reference? | Please make edits as suggested |
| §10.1 | Please note that the Unicode Standard explicitly suggests that the sequence equivalents for 0D4A, 0D4B and 0D57 should not be used. This fact should be documented by reference, even though it doesn’t change the result, which is that the sequences are excluded by akshar formation rules. | Please review the suggested edit and adjust as appropriate |
| §10.3.3 | This section is misnumbered as 6.2.3. | please fix |

# Whole Label Evaluation Rules

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| §7 | 1. There is the feeling that the abbreviation CH can be misread for C+H. Perhaps use a different letter?   (This would apply to the XML, including names of rules and classes).   1. At this point the IP has no other issues with the content or formulation of the WLE and context rules. (See report of test results below). | Please consider replacing the abbreviation. |

# XML file specific

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| XML: <description> | 1. There is no explanatory section on "Character Classes" between sections on "Variants" and "Background of Malayalam Script". [A placeholder has been added, which the GP needs to replace with its own text]. 2. Minor typos, Explicit place holder for script LGR proposal | Replace placeholders and review suggested edits in accompanying XML. |
| XML: Comments on <class> element: | There are no comments on the class elements. | Please use this comment field to describe the contents of each *named* class. |
| **Detailed editing** | Copy of XML included with suggested changes. | Please compare to the version submitted for feedback and note suggested changes, review and use as basis for further edits. |

# Test files

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| Test file | 1. the file submitted was more in the nature of a corpus file, useful for investigating the sufficiency of the WLE rules. 2. the file contained the pre-Unicode 5.1 Chillu sequences, not the atomic chillus 3. what the IP is looking for in a **test** file is    1. modest length < 1K entries    2. good coverage of all code point classes    3. good coverage of all code point contexts    4. invalid and valid labels all one label per line    5. a file organized in sections (with full line section headers using # at the start of the line)    6. each section to give one type if invalid label    7. a separate file listing cross-script variant labels | Please provide test file according to specification. |

## Corpus file analysis results

The IP evaluated both the corpus file delivered “Malayalam-test-labels.txt” as well as a separate corpus.

Neither used the atomic chillu representation, so the IP converted the files and normalized them to NFC.

**Corpus U**: 9861 valid Labels (single words)

This corpus had no invalid labels other than out-of-repertoire due to use of 04DC in corpus.

The following code points were not found: {0D0B 0D1D 0D57 0D7F}

The only contexts recorded from this corpus were of the form

[:Consonant:] ⚓=[:Halant:]

[:Consonant:] ⚓=[:Matra:]

[:Consonant:] ⚓=[:Anusvara:]

[:Consonant:] ⚓=[:Chillu:]

[:Matra:] ⚓=[:Anusvara:]

[:Matra:] ⚓=[:Visarga:]

[:Matra:] ⚓=[:Chillu:]

[:Vowel:] ⚓=[:Anusvara:]

[:Vowel:] ⚓=[:Chillu:]

**Corpus M**: this corpus has considerably more labels, did not use 0D4C, and contains “contaminants”, such as ASCII digits and punctuation. Nevertheless the number of invalid labels is rather low, indicating that the LGR rules do not appear to be overly restrictive.

Total Labels processed: 59803 of which

valid labels: 55044

invalid labels: 1512

Number of invalid labels by reason:

1 instances of invalid context (follows-B-X-or-H)

10 instances of invalid context (follows-only-C-V-or-M)

1454 instances of not in repertoire

39 instances of invalid context (follows-only-C-or-specific-CH-or-M)

skipped labels: 3247 of which

duplicate labels: 3157

broken labels: 39

contain join controls: 51

start w/ wrong script: 0

Test Label Coverage:

Repertoire (code points): 69 of 70. {0D02-0D03 0D05-0D0B 0D0E-0D10 0D12-0D28 0D2A-0D39 0D3E-0D43 0D46-0D48 0D4A-0D4B 0D4D 0D57 0D7A-...}

Repertoire not covered: 1 of 70. {0D7F}

Out of Repertoire: 18. [{0020 0022 0029 002B 002D-002F 0031-0037 003A 003C 003E-003F}]

Contexts encountered:

Contexts matching rule "follows-only-C-V-or-M":

[:Consonant:] ⚓=[:Anusvara:]

[:Consonant:] ⚓=[:Visarga:]

[:Matra:] ⚓=[:Anusvara:]

[:Matra:] ⚓=[:Visarga:]

[:Vowel:] ⚓=[:Anusvara:]

Contexts matching rule "follows-only-C":

[:Consonant:] ⚓=[:Matra:]

Contexts matching rule "follows-only-C-or-specific-CH-or-M":

[:Consonant:] ⚓=[:Halant:]

[:Matra:] ⚓=[:Halant:]

Contexts matching rule "follows-B-X-or-H":

[:Anusvara:] ⚓=[:Chillu:]

Contexts not matching rule "follows-B-X-or-H":

[:Chillu:] ⚓=[:Chillu:]

[:Consonant:] ⚓=[:Chillu:]

[:Matra:] ⚓=[:Chillu:]

[:Vowel:] ⚓=[:Chillu:]

Contexts not matching rule "follows-only-C-or-specific-CH-or-M":

[:Chillu:] ⚓=[:Halant:]

[:Matra:] ⚓=[:Halant:]

Contexts not matching rule "follows-only-C-V-or-M":

[:Chillu:] ⚓=[:Anusvara:]

[:Halant:] ⚓=[:Anusvara:]