Response of NBGP to IP feedback of Proposal on Tamil (Proposal of 2018-03-02)

DATE: 2018-04-22

# Overview

This documents covers the response of NBGP on the feedback of IP on Tamil LGR proposal dated 2018-03-02.

**IP Comment:** The draft .docx file constitutes an initial draft, with some sections that are fairly complete, while others are left unfinished. The included XML document allowed review of the specification against an available very large corpus of Tamil labels; that file, however, is unwieldy to use and there was there no purpose-driven file of test labels.

**NBGP Response:** Test labels have now been made available which cover all the valid-invalid sequences of characters.

**IP Comment:** The IP analyzed the BNF rules presented in section 7 and the repertoire date in section 5. The IP also reviewed the text of the proposal document and a marked-up version of that document is attached, as well as a version of the XML file with suggested changes in annotations/descriptions.

Notified changes have been made in section 7, changes have also been made in XML file.

The results of the test label review have been separately communicated to and discussed with the NeoB GP, they are listed here for the record. The IP understands that the GP has not fully finalized the details around some of the proposed context rules.

**NBGP Response:** The Visarga rule was still under consideration at the time of submission of the Proposal. Hence the large number of Visarga cases in the test labels. However, all the other invalid labels were properly invalidated and they had improper character sequences.

Now the context rules have also been fully finalized including those for the Visarga.

# Conclusion

The biggest issue in reviewing the Tamil LGR proposal draft rests in the fact that the Tamil script is related to Malayalam, yet, at this date, there is no corresponding proposal draft for the Malayalam script. Final evaluation of the proposed cross-script variants will have to wait until drafts for all related scripts are in hand.

In the meantime, the IP has identified a number of issues, many of them editorial, that could usefully be addressed by the GP in a revised draft. Such a revised draft should also complete any open items and hopefully finalize the proposed WLE rules. One of the issues concerns the treatment of alternate/nonstandard sequences for the SHRI ligature.

## General comments

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
|  | Final progress on Tamil will depend on getting Malayalam, as there are cross-script variants expected. |  |

## Comments on main document (.docx)

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Issue** | **IP Comment** | **NBGP Response** |
| Editorial and other issues | Proposed emendations attached as a separate file. | LGR\_Proposal\_Tamil\_20180302\_IP | LGR document has been amended accordingly. |
| References | The references section is left incomplete | Please complete | This has now been completed. |
| WLE rules | The GP has indicated that the WLE are not final | Please finalize the analysis | WLE rules have been finalized now. |
| Tamil Ligature Shri | Summarized immediately below. | This has been inserted with tentative wording into the marked .docx file, but any variant definitions added will of course need to be included in the XML. (Including listing any of the affected sequences).  While Unicode recommends the treatment of both sequences as equivalent for general purposes, the GP should evaluate whether making these variants is the most appropriate treatment for the Root Zone; alternatively, one of them might be disallowed instead. | This ligature was under consideration. After due discussions it was decided that they definitely need to be made variants instead of blocking one. Also, they have to be allocatable variants as both mean the same and there is a user community for both.  NBGP is pleased to acknowledge that IP has also suggested the same with justifications. |
|  |  |  |  |

### Tamil Ligature Shri.

Prior to Unicode 4.1, the best mapping to represent the ligature Shri was to the sequence <U+0BB8, U+0BCD, U+0BB0, U+0BC0>. Unicode 4.1 in 2005 added the character U+0BB6 TAMIL LETTER SHA and as a consequence, the best mapping became <U+0BB6, U+0BCD, U+0BB0, U+0BC0>. Due to slow updates to implementations, both representations are widespread in existing text. Therefore, treating both representations as equivalent sequences is recommended.

The two sequences are listed here:

U+0BB6, U+0BCD, U+0BB0, U+0BC0

SHA, VIRAMA, RA, II

ஶ ◌் ர ◌ீ = ஶ்ரீ

U+0BB8, U+0BCD, U+0BB0, U+0BC0

SA, VIRAMA, RA, II

ஸ ◌் ர ◌ீ = ஸ்ரீ

## Comments on LGR specification (XML)

The attached XML file shows an implementation of our suggestions. (None of them are normative).

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Issue** | **IP Comment** | **NBGP Comments** |
| <description>  section on Repertoire | The description states: "NBGP, in Section 5 "Repertoire" proposes 48 unique code-points to be made part of the Tamil LGR [Proposal]."  This wording is weird, because "Section 5" is not a unique location, but specific to the proposal. Also, we don't otherwise use language like "The GP proposes..." in these summaries, because then they have to be rewritten for the actual LGR.  Also, the LGR contains two sequences; this fact needs to be mentioned. | Change wording to  "According to Section 5, "Repertoire" in [Proposal] the Tamil LGR contains 48 unique code points and 2 sequences."  (The number depends on whether any variant sequences are defined for SHRI) | The XML is amended as per the comments. |
| <description> section on Variants | The current text (quoted below) is generic boilerplate that reiterates the principle, but does not provide information on what is actually in the LGR. We expect the <description> to be more of a human readable summary rather than general policy statement.  “…this LGR defines variants which are "confusing due to deviation from normally perceived character formations by larger linguistic community". These cases are not of mere visual similarity. They can cause confusion even to a careful observer and hence being proposed as variants.” | Suggest instead:  “…this LGR defines two sequences as variants to single code points because they look exactly alike and can cause confusion even to a careful observer.”  In the paragraphs following change  “…The Tamil script has a set of possible cross-script variants…”  To: “This LGR defines six cross-script variants…”  See attached XML file. (This text will need to change if SHRI is added as a variant.) | The XML is amended as per the comments and the inclusion of SHRI has also been done. |
| Data section | In the comments for VISARGA and VIRAMA, it would be OK to to give the native in lower case following " = " (space, equals, space) . (See attached XML)  The comment for U+0BBE does not match the Unicode character name. (TAMIL VOWEL SIGN AA)  The two sequences that are defined due to homoglyph variants should be identified as such in a comment. | Add a comment “homoglyph” to the “var” elements. | IP’s recommendation has been followed. |
|  |  |  |  |
| Variants | The text in this section needs some wordsmithing (see attached XML)   The in-script variant mappings should be commented in the data section (see attached XML) | See attached XML file for suggested wording. Please review and adapt as desired. | Reviewed and changed. |
|  |  |  |  |
| Character Classes | The <description> of the character classes provides useful information; there are just a few places where this strays into detail not needed to understand the specification, and which could be deleted to streamline this text.  There are also a few minor wording issues, they match the issues found in the equivalent text in the main LGR document. | Please consider some redrafting.   See attached XML file for suggested wording. Please review and adapt as desired. | Reviewed and changed. |
|  |  |  |  |
| Whole Label Rules | In the definition of symbols, the file uses “<br\>” which is invalid HTML syntax, the correct syntax would have been <br />  (The actual rule names in the <rules> section should be lowercased – except for any variable names like “C”.) | Please fix the syntax (see attached XML)  The rules would be further affected if the GP chose to make changes based on further analysis as indicated in the discussion on test labels. | Reviewed and changed. |
|  |  |  |  |
| References Section | [Proposal] place holder is missing | Add a placeholder for proposal (see attached XML) | Added |

## Comments on Test Labels

We do not yet have a designated focused test-label file, so the report below represents an analysis of a very large (1.1M entries) corpus. Despite its large size, it does not cover three of the code points. We do not have any tests regarding variants. Please provide specific test files (or a single combined file) for valid/invalid and variant labels.

Our detailed observations have been submitted before; they are listed here for reference, together with the response received.

**NBGP Response:** Test labels have been made and shared now.

### Detailed observations

We only have 1.1M entry word list. This is pretty unwieldy to work with, as it takes almost an hour to process. Only .2% of the labels contain one of 106 code points not in the repertoire, and a similar number fails one of the two context rules. The other context rule (Follows-either-V-C-or-M) is failed by 7K labels or about 3.5 times as often. Almost 1% of the putative labels fail the context rules - that is moderately high and would benefit from some research.  
  
Test Coverage is very good. Despite the huge number of putative labels, 3 out of 48 code points were never encountered. This is either because these code points are really rare, or because they are used in some language not covered by the sample. Both would be issues for concern. The letters in question are the three independent vowels

U+0B86     ஆ     TAMIL LETTER AA     
U+0B87     இ     TAMIL LETTER I            
U+0B88     ஈ     TAMIL LETTER II

It is unclear why they should not occur even in a sample of 1.1M words.

The putative labels cover all 5 defined tags and 3 named classes, as well as both context rules (match and failure). As with other test files, there is no label that matches the leading-combining -mark WLE rule.   
  
The Out of Repertoire code points which are covered include some of the code points excluded from the Tamil repertoire. (This is desirable for a test file).

Special test: the following four labels are processed correctly as "valid" by our tool. One of the IP members’ tools apparently does not have the same bug as the ICANN tool. (Note, however, that two labels are identical).

Validation: ‎ஷௌஸௌஹௌ‎ (0BB7 0BCC 0BB8 0BCC 0BB9 0BCC) : valid  
Validation: ‎ழௌளௌறௌ‎ (0BB4 0BCC 0BB3 0BCC 0BB1 0BCC) : valid  
IDNTools: \*\*\* WARNING \*\*\* label ‎ழௌளௌறௌ‎ (0BB4 0BCC 0BB3 0BCC 0BB1 0BCC) is repeated (Skipped): (Line 6)  
Validation: ‎கௌஙௌசௌஞௌடௌணௌ‎ (0B95 0BCC 0B99 0BCC 0B9A 0BCC 0B9E 0BCC 0B9F 0BCC 0BA3 0BCC) : valid

invalid context (Follows-either-V-C-or-M) : (0B83) :   
  
There are about .48% of all labels starting: (0B83 0BAA...  ) , or 4774 out of 1.1M; there are an additional 3-400 labels starting with (08B3) and continuing with a different letter. A few of those combinations have more than a dozen examples using the same second letter.

This points to the need to investigate to what degree leading Visarga should be relaxed.

Visarga follows Virama, about 2100 times (or in .2% of all labels). This seems a high rate, unless this is a common pattern for misspellings.

Altogether, we found:

   7145 instances of invalid context (Follows-either-V-C-or-M)  
   2097 instances of invalid context (Follows-only-C)  
   For comparison:    2056 instances of not in repertoire

(detailed results have been shared separately with the GP)

# Conclusion

The absence of the three vowel signs and the moderately high level of failure for one of the two context rules points to avenues for further investigation / confirmation before we can conclude whether the LGR is a satisfactory match for the corpus.

### Exchange between IP and NBGP on preliminary Tamil label testing

This is based on the Tamil corpus results, sorted by context rule and affected code point.

It is indicated in the LGR document that the GP is investigating whether to relax the rules around leading visarga.  This data could help.

About 70% of all failures for the visarga context rule are due to a word-initial visarga. (5.2K out of 1.1M) - in the majority, the following letter is 0BAA, but one or two other letters have notable frequencies as well.

A further 28%  are due to visarga following a virama (0.2% of all labels, or 2.1k out of 1.1M). This was not anticipated by the GP; if this is a common pattern of misspelling, it should be mentioned explicitly that it is not supported by the LGR. Alternatively, the GP may want to find out, why it occurs so often.

There are also many failures of the (Foll0ws-C) rule, listed in the file.

**From:** Dr. Shanmugam R. [mailto:shanmugamr@cdac.in]   
**Sent:** Tuesday, March 27, 2018 3:15 PM  
**Subject:** [Ext] Re: Tamil corpus results - invalid sorted by cause and code point

 Dear Dr Sarmad,

  Thank you for the interim feedback on Tamil document.

  I have gone through all the labels that failed the tests and all are valid cases which should have failed. The current Tamil LGR recommendations need not be changed for any of those except the case of Visarga which was already noted in the document.

I am giving below my analysis of various cases observed:

All those labels got tagged as invalid because they have the following combinations which are not recommended currently.

I. Visarga Combinations:

As you pointed out many of them are visarga + 0BAA combination which is currently under consideration and the same has been mentioned in the proposal chapter number 7 (WLE). The new thing I found from this sample is there are few other characters combination with visarga. As It is not a general practice in Tamil writing I feel the same should not be considered. This decision may also be taken after the due consultation with experts of Tamil.

II. Non Visarga Combination :

a. Matra + Matra (Violation of Left context rule for Matra)

b. Halant + Matra (Violation of Left context rule for Matra)

c. Vowel  + Matra (Violation of Left context rule for Matra)

d. Halant in the Beginning of the word. (Violation of Left context rule for Halant)

After filtering on the basis of above parameters I could not find any valid labels in the text file.

Thanks and regards, Dr.Shanmugam R.