Response of IP to NBGP Proposal on Telugu (Proposal of 2018-03-14)

DATE: 2018-03-27

# Overview

Although the description in English of the proposal is in some cases hard to follow the proposal itself appears to be technically sound. Some suggestions for improvements are included in the attached edited version of the .docx file.

Given the vast number proposed of cross-script homoglyphs between the Kannada and the Telugu scripts (amounting to just over half the repertoires of either ) , it is reassuring to see that a single list of homoglyphs is proposed in both the proposals. This must be maintained.

# Conclusion

The IP has reviewed the proposal and found several issues related to wording and other editorial matters. These can be discovered in the attached markked-up version of the proposal .docx file, as well as in this Response document.

Testing on corpus date, however, suggests that this may provide a valid basis as a set of generation rules for the Kannada script. The test label file provided is insufficient.

## General comments

|  |  |  |  |
| --- | --- | --- | --- |
| **Item**  | **Issue** | **IP Comment** | **GP comment**  |
| Previous responses:  | Of the IP comments on the Telugu proposal of 2018-01-15, all points appear to have been addressed in the proposal of 2018-03-14. | IP thanks the GP for their ready responses to most of the suggestions. | GP Appreciate IP effort to review the proposal and supporting document. |
|  |  |  |  |

## Comments on main document (.docx)

|  |  |  |  |
| --- | --- | --- | --- |
| **Item**  | **Issue** | **IP Comment** | **GP Comment** |
| Table 7 | The references are numbered 2 and 3. | please consider renumbering these to 101 and 102 and to match the use of these numbers in the XML.Adhering to this convention helps in integration. | Updated |

Some additional suggestions for improvements are included in the attached edited version of the .docx file.

## Comments on XML

|  |  |  |  |
| --- | --- | --- | --- |
| **Item**  | **Issue** | **IP Comment** | **GP Comment** |
| Numbering the references. | In the RZ-LGR we like to reserve [0] - [99] for versions of the Unicode Standard, so that these values have the same meaning across the entire LGR. We will impose that during integration, but it would be nice if the LGR proposal could renumber the script-specific references from [100]... on up. | Follow the RZ-LGR convention, using the same numbering used in the XML and DOCx. | Updated |
| <description> | <p>Variant Disposition: As variants are of confusingly similar, albeit of a peculiar nature, it is proposed that they be considered of "blocking" nature. There is no preference among these variants. Whichever label…blocked .</p>There are two issues. (1) It is unclear what “peculiar nature” implies and their natures has already been described in the paragraph prior.(2) The term used. “blocking” is not defined, so it would be better to use phrasing that allows the term “blocked” to be used.Finally, it might be useful to gloss the meaning of “blocked”.This change was already made in the 2017-12-07 LGR for Devanagari, but not propagated to any LGR that used the Devanagari LGR as template.THIS ISSUE APPLIES ACROSS A NUMBER OF NEO BRAHMI LGRS THAT USE THIS LANGUAGE |  <p>Variant Disposition: All variants are of type &ldquo;blocked&rdquo;, making labels that differ only by these variants mutually exclusive: whichever …. blocked. Ther e is not preference among these variants.</p> | Updated |
| placeholder | <dt>[Proposal]</dt> <dd></dd> | <dt>[Proposal]</dt> <dd>Neo-Brahmi Generation Panel, “Proposal for a Telugu Script Root Zone Label Generation Ruleset (LGR)”, [URL and Date TBD]</dd> | Updated |
| <rules> | The following named classes are defined but not used in this LGR: H, B, X. | Please remove the <class> definitions but do retain the tag values for <char> elements for these. | Removed |

## Comments on Test Labels

**Telugu, revised (2018-03-14)**

The revised proposal changes some rules and provides an additional (annotated) test label file.

The annotation (colon-separated an inline) is not handled well by our tools, so we changed the file into one that is sorted in to an invalid and a valid label section.

The new XML matches the dispositions for all labels.

However, out of 49K labels, only 3 fail due to context rules - this result might mean that either the context rules devised by the GP for Telugu are exceptionally permissive, or the label set is amazingly clean. Looking at the design of the rules, it appears that the reason may be that the rules are very simple and that that matches the way the script is used. Telugu does not appear to have a NUKTA, which greatly simplifies the rules by eliminating many possible permutations. There are only three special code points, of which Anusvara and Visarga at not non-spacing (meaning they occupy a very distinct location in the text and getting them out of order would visually be very noticeable. this makes it believable that out-of-order Anusvara or Visarga are rare to non-existent. Because the rules for Anusvara and Visarga are the same, there are only two context rules.

The annotated test file fails only one of them. It might have been nice to perhaps create some artificial invalid examples, perhaps in the test-labels file. Just to make sure the rule actually fails when it is expected to.

Updated Conclusion: The test coverage for the annotated file and the corpus is very good. If the script is as simple as the rules seem to indicate, then they are probably good as stated. The "test label" file is utterly insufficient (see earlier “Conclusion”).

**GP Comment:** The WLE rules have been revised. Kindly conduct the test again.

**Earlier results: Telugu (2018-02-15)**

There is a corpus file and IP tool  identifies 573 out of 49K labels as invalid (**while the file contains a comment #invalid labels it appears not actually sorted into valid and invalid labels.**) Almost all of the invalid labels are due to "out-of-repertoire" code points. The remainder of invalid labels are distributed across 4 out of 5 failed context rules, of which only the one related to the Visarga (Visarga-follows-only-V-or-M) has notable frequency (.1% or 57 instances).

Test Coverage is very good with all 63 code points covered. However, one context rule is never failed (Anusvara-follows-only-V-or-M) even though it is matched in 52 distinct contexts. [That context rule, as written, allows "Consonant" as well, despite the name, and clearly the majority of consonants do occur before an anusvara in the label]. As with other test files, there is no label that matches the leading-combining -mark WLE rule.

Three of the out-of-repertoire code points are from the excluded set of Telugu code points.

A separate test labels file is totally insufficient for regression testing as it produces only a single invalid label due to ASCII punctuation.
With the one exception noted, our tool **does not match** the putative set of "invalid labels" in that file; perhaps they were selected based on a different sets of context rules?

Conclusion: The test coverage for the corpus is very good, except for one relatively permissive rule (which effectively prohibits only anusvara, halant or visarga in front of an anusvara) which is never failed. The test label file is utterly insufficient. It is expected that a test label file contains at most a few hundred well-chosen labels that test all features WLE rules and repertoire. For scripts defining variants, the test file should provide a test for these. Cross-script variants for script pairs could be handled in a common cross-script test label file.

**An Exploration of Cross-script Collisions between Kannada and Telugu**

IP concatenated the Telugu and Kannada corpora in that order and tried to "register" all of them in sequence to check collisions.

Some statistics

Conflict: Number of conflicts (duplication) = 2,070
Conflict: **Number of conflicts (variants)    = 287**
Conflict: Number of conflicts (this file)   = 2,357
Number of invalid labels by reason:
   1 instances of invalid context (Orya-follows-only-C-or-N-or-start)
   3 instances of invalid context (Telu-Follows-only-C)
   345 instances of invalid context (Knda-Follows-only-C)
   4113 instances of not in repertoire

---End of List--(**174831 IDN Labels Input**,  4481 Invalid Labels)-------------

    Total Labels processed: 174832 of which
         valid labels:   170350
         invalid labels: 4481
         skipped labels: 1 of which
            duplicate labels:      0
            broken labels:         1
            start w/ wrong script: 0

Note the files contained invalid labels which were screened out, and an appallingly large number of exact duplicates that I didn't catch as such before.

Only 287 report as true collisions despite the huge set of cross-script variants.

**GP Comment:** Thank you IP for detailed information. The new test file with almost 200 labels has been provided. It covers valid and invalid case for each rule, and it covers variant labels with Kannada and Sinhala scripts. There is no longer cross-script variant code point with Oriya, and there are three cross-script variant sets with Sinhala in addition to the 34 variant set with Kannada.