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

DATE: 2018-03-26

# Overview

Although the description in English of the proposal is in some cases hard to follow (some suggestions for improvements are included below) the proposal itself appears to be technically sound.

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 shall be maintained.

# Conclusion

The IP has reviewed the proposal and found several issues related to wording and other editorial matters. Testing against putative labels from a corpus file did not reveal any obvious issues with this proposed set of generation rules for the Kannada script.

## General comments

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Issue** | **IP Comment** | **GP Comment** |
| Previous responses: | Of the IP comments on the Kannada proposal of 2018-01-15, all points appear to have been addressed in the proposal of 2018-03-02, except for the following. | IP thanks the GP for their ready responses to most of the suggestions. |  |
| syllable formation rules | A summary of syllable formation rules (essential a BNF for akshara-structure) was requested (like Appendix B in the Telugu proposal; but only a short section §7 was provided. | The summary of rules is very brief, no answer was given to the question on the status of V (Independent vowel characters), namely if they are not restricted to word-initial position, or position immediately after another V or M. | The rule is provided in section 3.4 |
| §5.1 | an index reference for each code point was requested | A single reference is used for all code points, viz [10] Kannada Madhyama Vyakarana, 2001 |  |
| References | There are a number of requests to clarify and standardize the references.  These should   * be according to the existing scheme whereby each reference is assigned a number (< 100 for Unicode documents, >100 others) * occur in a single list, which should include a place-holder for the Proposal itself. | In fact, there are two separate lists of references (occurring at the end of the description section and at the end of the proposal proper);  no numbering has been assigned (just brief names); and the Proposal placeholder has no explicit name, title or definition.  Please consider using the conventional reference numbering starting at 100 (not 10). This helps in integration.  (see discussion below for XML) | Done |

## Comments on main document (.docx)

A marked-up copy of the proposal is also attached, showing editorial suggestions and other comments from the IP.

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Issue** | **IP Comment** | **GP Comment** |
| Multiple | See Marked-up copy of the proposal  (These range from simple copy edit suggestions to requests to supply missing text). | See marked up copy of the proposal | All incorporated. |

Additional:

The sentence

Most of the public records were written during Mysore kings are in Modi script

should be revised:

Most of the public records written in the period of Mysore kings are in Modi script.

The sentence

This can easily be called the last developed script which taught even now in schools…

should be revised:

This may be called the latest development in the script, and is taught even now in schools…

The sentence

The Konkani language is written in Devanagari, Roman, Malayalam, scripts also.

should be revised:

The Konkani language is also written in the Devanagari, Roman and Malayalam, scripts.

One spelling “Deavanagari” (on p. 6) should be corrected to “Devanagari”.

The first Paragraph in 4.1.2.1 should be re-written:

The code point repertoire for root zone being a very special case, up the ladder in the protocol hierarchies, the canvas of available characters for selection as part of the Root zone code point repertoire is already constrained by the various protocol layers beneath it. Following three main protocols…

might be

The code point repertoire for the Root Zone is a very special case. In fact, the range of available characters for inclusion in part of the Root Zone code point repertoire is already constrained by the various protocol layers on which it rests. The following three main protocols…

## Comments on LGR specification (XML)

The clause:

"which evolved from the 5th-century Kadamba script."

should be deleted - such background belongs in the DOCX file.

The sentence:

"Structure of Kannada is similar to other Indian languages."

should be edited as follows:

"The structure of Kannada is similar to that of other Indian languages; it is particularly closely related to Telugu".

The sentence as follows

"Two categories of consonant characters (vyañjanas) are defined in Kannada: the structured consonants (vargeeya vyanjana) and the unstructured consonants (avargeeya vyanjana). The structured consonants are classified according to where the tongue touches the palate of the mouth and are classified accordingly into five structured groups. The unstructured consonants are consonants that do not fall into any of the five structured groups. "

is in the <description> but it oes into detail not relevant to the design /use/ implementation of the LGR and should be omitted.

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| <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. There is no preference among these variants.</p> |

## Review of Label files

*Corpus*. Various tools concur that in identifying 3690 out of 125K labels as invalid in the corpus file. Almost all of these are due to "out-of-repertoire", but one of the two context rules defined has a moderately high fail rate:

354 instances of invalid context (Follows-only-C)

Test coverage is very good, and includes all 62 code points. Correspondingly all 6 named classes and 6 defined tag values are covered as are all 2 context rules (each has at least one match and one fail). WLE rules: "leading-combining-mark" does not have a match case (i.e. no invalid label starts with one).

Two of the out-of-repertoire code points are from the excluded set of Gurmukhi code points.

Variants. There is no dedicated test file for variants (one should be supplied), but the IP carried out a test on the corpora for Telugu and Kannada to check for the frequency of collisions due to the large number of cross-script variants.

It was found that after “registering” about 45K of Telugu words as labels, only 287 of 120K Kannada words had a conflict (were blocked variants of) some corresponding Telugu entry. This finding is preliminary. If verified, it would indicate that the effective reduction of the combined namespace due to the variants is rather moderate.

Test labels. There also exist a shorter test-labels file with rather less complete test coverage, including not covering all tags and classes. Not attempted to confirm whether this file is a strict subset. It doesn't seem complete enough for regression tests.

*Conclusion*: the invalid context rate is moderate for a not very "clean" sample, however it might be worth verifying that all failures are indeed intentional. Please provide a better file for regression testing (short file with reasonably complete coverage of code points, rules and contexts) and sectioned into valid/invalid or invalid-by-reason areas. Also provide a test file for variants.

Files and versions used:

Input\Proposed-LGR-Knda-20180302.xml;     Labels\test-labels-kannada.txt  
Input\Proposed-LGR-Knda-20180302.xml;     Labels\corpus-labels-kannada.txt