Response of IP to Devanagari LGR of 29th Apr. 2018

DATE: 2018-05-16

# Overview

This document provides IP response to the Devanagari proposal dated 2018-04-29.

## General comments

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| **Previous responses:** | The IP has reviewed ....2018-04-29.DOCX, .... XML, the test label files and the responses by the GP to ....DOCx which were updated in response to the previous feedback by the IP.  The document entitled [Response of NBGP to IP Response: Devanagari LGR Proposal on 2018-03-29] appears to the IP to close all previous issues, hence will not be further referenced. | IP thanks the GP for their ready responses.  We note the progress made by resolving all the issues the IP had raised. |
| **Current situation** | Now that IP has seen the complete set of other LGR proposals for Indic scripts, there are a few additional issues that IP thinks should be addressed, in addition to some further editorial suggestions. | Details as below. |
| **Basis for labels in phrases as a well as words of ordinary language** | A common issue in labels is the lack of interw0rd punctuation or space. IP notes that the GP addressed the issue of an ungrammatical H | V juxtaposition as result of a label representing a multi-word expression written without the benefit of spaces, punctuation or joiners. IP would like the GP to state explicitly whether this is the only such case of concern for the LGR, or, if not, discuss the other cases in a similar manner.  This issue is perhaps exacerbated by the fact that the test files / corpora that the IP has been using in reviewing LGRs generally contains single or compound words, but not multi-word expressions written without the expected inter-word separators. | It would be useful if the script proposals addressed this point directly, presumably in § 4.1.2  (We note that the case of Halant in final position before a vowel is already addressed – this feedback aims to ensure that other, analogous cases are covered as well). |
| **Policy on Joiner characters** | 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. | Ideally, all NeoB LGRs should have a paragraph in that chapter about (a) whether the script commonly uses ZWNJ / ZWJ (b) what the effects are (i.e. what they control) (c) whether their use is 'optional' or stylistic, or not (i.e. semantic) (d) mentions that ZWJ / ZWNJ are out of scope for the Root Zone  When we run tests on the corpora, we will see many failing labels in some scripts that are due to U+200C ZWNJ.  (We note that there are some mentions of ZWj/XWNJ being excluded in sections 4 and 7, but we think they don’t adequately address the issue.) |
| **§ 5. Repertoire** | The IP has been discussing the question of whether the 13 allowable NUKTA sequences should be enumerated, and NUKTA itself removed.  While this would match the way similar diacritics are treated in other scripts, there is still the need to create a class N (for Nukta) for use with the rules. While classes can contain code points that are only elements of sequences, not having the NUKTA listed would create an apparent discrepancy. | No change is recommended in terms of repertoire, but see WLE rule change. |
| **Code chart** | MSR-3 changed conventions in coloring, when updating to MSR-3, note correct conventions. | Suggested: update to MSR-3 |
| **Table** | In the table, the column giving the Unicode general category is somewhat uninformative compared to the Indic Syllable categories. Suggest to drop this column. | Suggested to drop one table column. |
| **§ 6. Variants** | NUKTA isn't accounted for in the cross-script variants, specifically re 0917 and 0A17.  0A1C is also fit for a nukta in Gurmukhi, but corresponds to a geminate 0924 in Devanagari, which cannot be followed by a Devanagari nukta, so can be ignored.  It appears that for cross-script variants for Bengali, the case of VISARGA may need to be investigated. It appears that this mark can be applied following the proposed cross-script variant in each script. | There is a problematic absence only in the situation where 0917 or 0A17 are followed by the respective nukta symbol, viz Devanagari 093C and Gurmukhi 0A3C. This is best addressed by making the NUKTAs variants of each other. |
|  | More generally : | In defining cross-script variants, the IP would like to call the GP's attention to the effect of NUKTA, VISARGA and similar marks of identical appearance across scripts. Normally, cross script variants are not needed for diacritical marks because they cannot stand alone, so they cannot, by themselves, create variant labels. However, the situation changes if the respective base characters are variants. Because of the variant relation for 0917/0A17 any sequences of these followed by a NUKTA should also be considered variants (0917 093C / 0A17 0A3C). The remedy would be to make the NUKTA a cross-script variant for Devanagari/Gurmukhi -- there is no need to restrict this relationship to specific sequences because the context rule and variant relation for the consonants must also agree to form a variant label, therefore implicitly providing that restriction. For Devanagari/Bengali it appears that a variant relation for VISARGA may need to be considered for similar reasons.  Please verify that there are no other such cases (and alert the other script working groups to look out for this issue). |
| **WLE (and corresponding changes in the XML**) | After reviewing Devanagari in the context of the other Indic Script LGR proposals the IP suggests to use uppercase abbreviated class names e.g. "N" instead of "Nukta".  The Devanagari LGR is currently the only one that uses full names, which makes for a bit of an inconsistency (even though the names do not matter for how the rules work). | Use abbreviations in class and rule names. |
|  | XML: Comments on <class> element: please use this comment field to describe the content of each *named* class. | (For suggested comments see accompanying XML) |
|  | Nukta rule: Devanagari LGR is the only one to simply list all code points preceding Nukta inside the rule. For consistency, the IP suggests creating three tags: C1, V1 and M1 that can be applied to the special consonant, vowel and matra code points that are to be used with Nukta. | Also add three new named classes such as:  <class name="C1" from-tag="C1" comment="Consonants that may precede Nukta" />  The Nukta rule then changes to something like "follows-either-V1-C1-or-M1" (see similar rule in Tamil, except that there the full sets V, C and M are used). The IP feels that the slight gain in consistency and documentation is worth the effort. |
| **Nepali** | A reference for Nepali is missing from Table 5 and the XML.  Konkani is another language listed in the table in Section 5 but has no reference. | Please add (best to not renumber the existing reference.) |
| **References** | [MSR] | this citation should be updated to MSR-3 (no other change or review by GP is required as MSR-3 is identical to MSR-2 for Devanagari). |
| **Appendix** |  | "visually distinguishability" --> "visual distinction” |
| **Detailed editing** | Copy of .DOCX is attached With “Track Changes” suggestions | Please note the comments in the document – some of the suggested edits affect passages where the IP suggest a bit of a rewrite for better understanding. Please review to make sure the changed text is agreeable to the GP. |
|  |  |  |

## XML points

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **Comment** |
| Repertoire in <description> | The repertoire count is given as 92, but the IP’s tool reports a count of 96. The discrepancy appears to be due to additional sequences for variants, giving a total of 13 sequences. | Fix 92 🡪 96 (see accompanying XML) |
| **“These characters”** | When reviewing the Devanagari proposal in context, the IP noted that two of the comments in the XML use "These characters" while in the same situation all other Indic LGRs, as well as other comments in the Devanagari rules now explicitly indicate the categories affected (either by long name or abbreviation). | Please replace "these characters" in the comment text by a description of the affected code point categories. This will make these comments more consistent. |
| **Standardizing names of variables in Classes, Rules, Actions, BNF for WLE etc.** | Each of the NB script LGRs has a set of Variables (usually initial capital letters) identifying the major classes of phonetic entity involved in the rules for the script.  There is similarity, but no consistency, among the Variable-names used. | IP recommends that the following words, as they appear in **rule and class names** etc., be replaced by the following capital letters in the Devanagari LGR XML: consonant-halant-matra-nukta-vowel 🡪 C-H-M-N-V |
|  |  | For all NeoBGP LGRs: we note that most of the LGRs are consistent in the use of these abbreviations (the same letters stand for the same categories). However, there are some isolated exceptions. Encourage NeoB GP to review these and recommend an alignment if warranted. |
| **Furhther Editorial issues** | a) Languages covered should be listed in comments. See suggested edits in accompanying XML file  b) for each character class a comment should indicate the contents of that class  c) Santhali spelling change to Santali to match feedback on document  c) detail edits in the description to align with feedback on other NeoB proposals, see accompanying file.  d) several typos corrected | Please review and use the provided file as basis for making further changes. |
| **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

The supplied test files evaluate as expected. Full analysis of Cross-script variants and associated test file has not yet been done, but cross-script labels report as “invalid” because out of repertoire code points which is as expected.