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

DATE: 2018-05-24

# Overview

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

# General Comments

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| Status | The May 11 version of the Bengali LGR proposal represents an update to the proposal reviewed earlier by the IP (See Response on-Bengali-20180302-v3-IP.docx) dated April 12, 2018. |  |
| Other communi-cation | After we started looking at the 2018-05-11 version of the proposal, we received additional communication that indicated that there were a few known issues with the proposal.   * Variants are not fully updated to consensus * 0994 consonant -> vowel * 098C and 09D7 should be excluded * the VHCM sequences effectively create new vowels and rule “S” is an attempt to enforce that constraint via context rules | We have taken note of this, and in our response, we treat these items as if the information to the right represents the actual intent of the GP. |
| 3.3.6 Nukta | Section 3.3.6 "Nukta" is very confusing as it repeats the description of Nukta from the Devanagari proposal verbatim with the Bengali-relevant issues casually inserted. This needs to be cleaned up. | Remove text not specific to Bengali. |
| 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 there is a brief mention in section 7.1.1, but it is not solely focused on joiner characters. | Please add. |
| 5.1 | The Unicode General Category is rather uninteresting and can be removed from the table. | Remove one column |
| 098C and 09D7 | We understand from separate communication that 098C and 09D7 should no longer be members of the repertoire and the category of 0994 should be Vowel, not Consonant.  We do understand that 098C is a VOCALIC L , used only for (very few) Sanskrit words , so naturally a candidate for omission.  However we have a question regarding the supposed decision to exclude 09D7:  IP notes that 09D7 is an AU LENGTH MARK, and according to Unicode 10.0 p. 470 it would be required to write Kok-Borok, a language which according to the Bengali LGR Proposal p. 9 is of EGIDS 4, and so would be expected to be supported in the LGR. In the Proposal document on p. 35, 09D7 is also said to be needed for Bengali, Manipuri and Assamese. So why is it being dropped?  Also, any code point dropped should be listed in the section on excluded code points together with some sort of explanation why it is inappropriate for Root Zone IDNs. | Please make sure the next draft correctly reflects the intent of the GP |
| Sequences | Section 5.1 mentions sequences but they are not found in the current draft of the XML file. | See separate discussion below. |
| Khanda Ta | There is a very specific rule defined for the Khanda Ta in Section 7 and in the XML file; the earlier discussion in the document appears not to match it; it is difficult to see that the XML implements what the document states as being required.  Sections 5.4.5 and 5.4.6 state simply  **5.4.5.** A single ‘Khanda’-Ta (Z)  …  **5.4.6.** A Khanda Ta can be preceded by a consonant and Halant (also known as Virama)  The latter is then (apparently) paraphrased by a BNF formula  [CH]Z  and both are followed by equations which are presented as “Examples”.  The former just means a Z cannot follow a Z and the latter is even less restrictive than rule P (C2H combination – see below in the XML section under <rules>).  (Appendix item 2 resolves the difference between [CH]Z and [C2H]Z, but the other details cannot be immediately understood.  Perhaps the real rule here is:  7: must be preceded by some code point that is not Z or H, or by P.  In a pseudo code XML formulation this could be rendered as  follows(complement (union(Z, H)), or P)  which may or may not be easier to follow that than  follows(V, C, N, M, D, B, X, or P)  Best might be to keep the rule as is, but just add a footnote that relates the formulation in section 5 to the implementation in section 7 – with pointer to the appendix). | Please resolve the discrepancy between the very general restriction in section 5.4.5 and the actual very detailed rule 7 in 7.1 which is also expressed in a manner not easily related back to the way it’s stated in section 5.45 (made more complicated by the fact that the appendix modifies CHZ). |
| Writing Bengali Syllable tta | Still under the heading of Khanda Ta, it is possible that a variant might need to be added. On p. 473 of Unicode Standard 10.0 , there are said to be three ways of writing tta (though perhaps the second of these is excluded from the root zone, since it involves ZWNJ). Are these in free variation, i.e. a given word can be spelt either way? Or do different words choose one spelling or the other? If the former, 09A4 + 09CD + 09A4 should be made a variant of 09CE + 09A4. (They are homophones, not homoglyphs.) | Please review and resolve |
| Multi-word or phrase-based labels | A common issue in labels is the lack of interw0rd punctuation or space. Apart from a brief mention in Section 7.1.1, issues related to placing words or phrases together without any separation appear not to be addressed. Because multi-word domain names are common for other scripts, they can be expected for the Bengali script. Therefore, a more complete discussion of the issue in the context of describing the script behavior in Section 5 would be advisable.   This issue is 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 explicitly addressed any potential issues caused by concatenating words directly. |
| References | 1. Please change the few bare URLs in the text to actual numbered references 2. 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. 3. [108] this seems to be incomplete – same issue in XML | Please make edits as suggested |
| § 5, Table of repetoire | 1. 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. | Please edit |
| § 4.2.1 | ABNF formalism is described, but not used in Section 4. Perhaps indicate where in the document this is used. |  |
| Ya-phala, Ra-phala an Ref | The spelllings here are non-standard : the Unicode Standard (p. 473) uses ya-phalaa, Repha and Ra-phalaa. |  |

# Ya-phalaa and Ra-phala and Repha sequences

These sequences are special in that they allow some code points to appear in contexts not otherwise allowed by the rules. The IP proposes to the GP to merge the description of these from section 7 with the description in section 5, making new sub-sections “3.3.8 Ya-phala Sequences” and “3.3.9 Ra-phala and Ref Sequences”. (See editorial suggestions on pp 46-47 of the marked up document)

This would be the best place to describe how these sequences work in the context of the script.

As we will see, the ya-phala or VHCM sequences should be listed in the repertoire. They are currently listed in a separate table; that table can be merged with the main table or kept separate, as long as it is clear that these sequences are members of the repertoire. Also: Add the sequences to the XML file.

Some of the context rules that will become unnecessary.

The following two rules can be deleted from the XML. V1 H C1 M1 will be allowed because all valid combinations will be listed as a sequence in the repertoire (and context rules are not evaluated inside a sequence).

|  |  |  |
| --- | --- | --- |
| precedes-V1-and-follows-C1-and-M1 | ([:[**V1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_V1):])←⚓→([:[**C1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_C1):][:[**M1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_M1):]) | Section 7, Special case of Halant from WLE 7: H must precede V1 and follow C1 and M1 |
| precedes-only-C-or-N-or-precedes-specific-V-and-follows-C1-and-M1 | (:[follows-only-C-or-N](file:///C:\\src\\idntables\\test\\LGR-3\\staging4-NeoB\\Element\\lgr-3-Bengali-Script-2018-05-15-en.html" \l "rule_follows-only-C-or-N" \o "Rule: follows-only-C-or-N):)|(:[precedes-V1-and-follows-C1-and-M1](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#rule_precedes-V1-and-follows-C1-and-M1):) | Section 7, From WLE 2 and 7, H: must be preceded by C or N, or preceded by V1 and followed by C1 and M1 |

For the context required for the Khanda Ta, we note that “S” ends in M1 (which is a subset of “M”). Therefore, term (:S:) can be removed without any change (and rule “S” can be deleted entirely).

Rule “P” ensures that a H cannot precede a Z unless it is following a C2. This rule can be retained, and term (:P:) can likewise be retained.

|  |  |  |
| --- | --- | --- |
| S | [:[**V1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_V1):][:[**H**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_H):][:[**C1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_C1):][:[**M1**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_M1):] | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P |
| P | [:[**C2**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_C2):][:[**H**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_H):] | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P |
| follows-only-V-C-N-M-D-B-X-S-or-P | ([:[**V**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_V):]|[:[**C**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_C):]|[:[**N**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_N):]|[:[**M**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_M):]|[:[**D**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_D):]|[:[**B**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_B):]|[:[**X**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_X):]|(:[S](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#rule_S):)|(:[P](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#rule_P):))←⚓ | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P |

The ya-phala sequences begin with a “V1”, therefore they need a context rule that is appropriate for their first code point, which is the same rule as for V (as “not-when” rule):

|  |  |  |
| --- | --- | --- |
| preceded-by-H | ([:[**H**](file:///C:\src\idntables\test\LGR-3\staging4-NeoB\Element\lgr-3-Bengali-Script-2018-05-15-en.html#class_H):])←⚓ | Section 7, WLE 8: V: cannot be preceded by H |

M normally does not have a right-hand-side context, so the ya-phala sequences, which end in “M1” also need no right-hand-side context.

The Ra-phala is a C H C2 sequence and thus does not require special treatment in the WLE rules. The Ref is a C2 H C sequence, which also needs no special treatment, other than allowing a “Z” (Khanda Ta) in place of the trailing “C”.

Rule P for the Ra-Halant can be kept in Section 7 and the XML, but all the description text should move to section 3 (as suggested above).

In rule “7” the term “S” can be deleted as can the <rule by-ref=”S” /> in the XML. Also, note to delete the “S” in the names of any rules in the XML. The context rule for 09CD needs to be changed to “follows-C-or-N” and these two rules “precedes-V1-and-follows-C1-and-M1” and “precedes-only-C-or-N-or-precedes-specific-V-and-follows-C1-and-M1” deleted.

Summary:

1. Create two subsections in section 3 and collect all description of the linguistic function of the sequences there.
2. Keep table of ya-phala sequences and mark them as members of the repertoire
3. Add them to the XML with context rule same as for V
4. Delete rule “S” from section 7 and XML; delete all references to “S” from rules and names.
5. Keep rule “P”.
6. Delete two rules in XML related to halant (now taken care of by sequence) and change context rule for 09CD to “follows-only-C-or-N”.

# Variants and WLE Rules

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| Variants | The listed cross-script variants are in disagreement with the corresponding variants from the Deva and Guru proposals  (We understand this may be a “known” issue; we nevertheless list it as a feedback item here, so it doesn’t get lost) | Please update the variants either here or in other proposals until they match. |
| §6 | The IP would like to make sure that the GP makes a careful distinction between the concepts of “confusable” or “similar” on the one hand, and “homograph/homoglyph” and “(near) identical” on the other hand.  Only the latter can be fully substituted for each other and are therefore eligible for variant status.  In this context, the IP likes to reserve the term “variant” for those cases where an actual, formal variant relation has been proposed. Anything else is either a “confusable” or possibly a “candidate variant” (if under consideration for variant status). | Please restate to better clarify these distinctions |
| Section 6 | The discussion in the first part of Section 6 presents “interesting” and “noteworthy” examples, without coming to a firm conclusion on whether these are or are not to be treated as variants.  For the purported dual encoding it would be of interest to know whether under the WLE rules both are possible, or whether one of them is effectively excluded.  For example, the following restriction documented in Chapter 12 of Unicode 10.0 [100], is covered by the rule preventing a Matra from following a V. | Please complete the discussion in section 6 and provide a conclusion on whether in-script variants can exist (after WLE rules have been applied). |
| Missing Variant | **Visarga:** the Visarga would seem to make a candidate for variant between Bengali and Devanagari (it's not used in Gurmukhi). In both, it is relatively unconstrained and therefore could appear in whole-script variant labels made up from other variant code points. | Please investigate the Visarga |
| Variant needed for Bengali/ Assamese RA? | It IP is wondering whether there may be a case for a blocked variant relation between Bengali and Assamese RA, at least when occurring in Repha or Ra-phalaa.  The GP writes: “The point is in both the cases the slot for *ra* could be Bengali *ra* র (U+09B0) or the Assamese *ra* ৰ (U+09F0), followed/ preceded by the common halant (U+09CD), whereas the shapes of *repha* and *ra-phala* in both the cases remain the same. | Please review and document both review and decision in the next draft |
| Section 7 | ” (Please add A Ya phala and e Ya phala cases here)” This appears to be an editing instruction that can be removed after re-organizing the sequences (see separate section above) | please remove |
| Section 7 | Please consider discussing the linguistic details behind rules S and P in Section 5 so that only a brief reference is needed in Section 7. | Please reorganize the description to collect the full description of the script behavior in Section 5. |
| Multi-word labels | IP notes that for some scripts the NeoB 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 these are the only such case of concern for the NeoB LGRs, or, if not, discuss the other cases (including other types of juxtapositions) in a similar manner. | Please state |
| Appendix | This section details 7 restrictions which apply to “Bengali“ (apparently, the script when used to write Bangla, but not necessarily for other languages). As such, they should not figure in the Root Zone, at least not without linguistic evidence of their widespread validity.  Item 5 appears to require use of Avagraha 09BD. But this is excluded from MSR. Presumably, therefore the “permissible combinations” are not permissible in the root zone. (It is fine to document such features of the script, as long as the relation to the Root Zone restriction and/or Root Zone LGR are made clear.)  The status of items 6 and 7 is unclear. The constraints mentioned do not appear in the LGR; are they simply for information or are they supposed to be added somehow to the proposal (.docx, and.xml)? | Please clarify status |

# XML file

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| **<description>** | 1. Because of some sensitivities about the naming of the script, an explanatory sentence is suggested mentioning Assamese in particular as one of the languages covered by this LGR. 2. The section on consonants can be streamlined by removing the details about phonetic classification. Such classification, while traditional, does not affect the way consonants are used in IDNs or in this LGR. | Please note suggested edits. |
| **<references>** | reference 124 is missing; online references are bare URLs, need Website name, Page title, and, if identified authors/editors, plus “accessed on” date. Reference 108 appears incomplete | Please complete references |
| **09D7** | The accompanying XML file has 09D7 removed based on the separate communication received by the IP – however, in light of the comment above, it may be necessary to restore it. | Please review and fix |
| **Comments on <char>** | Generally, the IP like to see the language information in the comments. That information cannot be looked up elsewhere in creating the HTML files, unlike character names. The latter can be omitted.( If its desirable to have both, put character name first, then use “:” as separator) | Please provide the language information from the table in Section 5.1 in the comments for each code point. |
| **<rules>** | **WLE rules:** In the following set of rules defined to constrain the Khanda Ta (Z), the rule "S" is ineffective, because any [:M:] already matches the left-hand context for Z whether or not is also satisfies rule "S".   |  |  |  | | --- | --- | --- | | S | [:**V1**:][:**H**:][:**C1**:][:**M1**:] | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P | | P | [:**C2**:][:**H**:] | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P | | follows-only-V-C-N-M-D-B-X-S-or-P | ([:**V**:]|[:**C**:]|[:**N**:]|[:**M**:]| [:**D**:]|[:**B**:]|[:**X**:]|(:S:)|(:P:))←⚓ | Section 7, WLE 7: Z: must be preceded by V, C, N, M, D, B, X, S or P |   These rules are affected by the recommendation to list ya-phala sequences explicitly as members of the repertoire. (See discussion above) |  |
| **Unused classes** | Several named classes are defined but not used. These should be removed from the XML.  Corresponding Tag values as well as listing in the description section can and should be retained. | Please remove unused rules. |
| **rule for H** | The rule for H is defined in Section 7 as, “H: must be preceded by C or N, or preceded by V1 and followed by C1 and M1”, which is implemented by this rule in the XML:  precedes-only-C-or-N-or-precedes-specific-V-and-follows-C1-and-M1.  As discussed above, this can be simplified to just using  [follows-only-C-or-N](file:///C:\Business\Projects\Dropbox\IP\LGR-3\LGR-3-staging4-NeoB\NeoBTestRelease-2018-05-13\lgr-3-bengali-script-2018-05-13-en.html#rule_follows-only-C-or-N)  plus enumerated sequences. (see above) | Please change rule for H in the XML |
| **Sequences** | The attached XML shows how sequences could be implemented and how the rules would have to be adjusted. | Please review. |
| **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. |
| 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. |

# Test Files

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Comment** |
| Issues | * The test label file is really a corpus - about 40K entries. That makes it too large to use for regression testing. * 4100 entries (or 10%) were **not normalized to NFC**.  That invalidates them for testing without adding information, because in a real-life scenario applicants would submit the same words after normalization. (IP ran them though a small script that normalizes and lowercases corpus entries for the results listed below) * The status was annotated on a line-by line style, which is a format our comparison scripts are not set up for. IP sorted the file by result, to move all "invalid" labels into a single section and all valid ones in to another. * There is no indication why a label should be invalid | The IP would like to request a smaller file (a few hundred entries) that gives good coverage of all tags, classes and contexts, but is small enough to be used |
| Coverage | Even though it is more a corpus than a test file the file does not cover the entire repertoire; While it covers all tags and classes, it does not cover all the context rules (see “findings”) |  |

Findings:

After processing the submitted file to normalize it to NFC and applying the originally submitted XML  
the following results were obtained (the do not change substantially after rewriting the XML to the form that is attached with this feedback).

* All labels that were supposed to be valid, report as valid.
* Found the following breakdown of labels:

Total Labels processed: 40807 of which  
 valid labels:   38632  
 invalid labels: 1465 of which

1286 instances of not in repertoire  
 {0027 002D-002F 09E6-09EA 09EE 2019 201D}

and these instances of invalid context

* 1 preceded-by-H
* 143 follows-only-C-or-N    
  or .3% which is a bit on the high side
* 5 precedes-only-C-or-N-or-precedes-specific-V-and-follows-C1-and-M1

(after the edits to the XML, these 148 cases all identify as follows-only-C-or-N)

* 30 follows-only-V-C-N-or-M

   NOTE: three context rules were not encountered in invalid labels

* follows-only-V-C-N-M-or-D
* follows-only-specific-Cs
* follows-only-V-C-N-M-D-B-X-S-or-P

        710 skipped labels, of which

            duplicate labels:      425  
            contain join controls: 285

Repertoire (code points):  59 of  64. {0981-0983 0985-098B 098F-0990 0993-09A8 09AA-09B0 09B2   
 09B6-09B9 09BC 09BE-09C3 09C7-09C8 09CB-...}  
Repertoire not covered:     5 of  64. {098C 09C4 09D7 09F0-09F1}