IP Review of Bengali (Bangla) LGR Draft

DATE: October 24, 2019

VERSION reviewed: 15oct19-version4-4 (xml and supporting document)

# Overview

The proposal appears to be in good shape. However, the IP has not completed its full suite of tests; until then, this conclusion has to remain tentative. While there are quite a few detail comments, none of them affect the core of the proposal; they are largely editorial matters that it should be possible to address quickly. If that is done, and further testing doesn’t identify any hidden technical issues, the LGR proposal could well be ready to meet the projected timeline for LGR-4.

Because of this, the review includes detailed comments appropriate for a late-stage draft.

# Comments

The following summarize various observations made during the review; any suggestions or recommendations derived from these observations are called out explicitly. In no particular order:

(1) In the table of repertoire elements (Table 8), the **sequences are not numbered**. (They appear as unnumbered entries).

**Recommendation:** please number all entries in table 8.
**GP response:** DONE.

(2) The Unicode name for these sequences given in table 8 is "Normalized Form of...<Unicode Name>". That's not something we can do in the **HTML format** for publishing the LGR – in that format the column for Name is programmatically based on Unicode character names. However, for the supporting document things are fine as they are.

**Recommendation:** no change
**GP response:** Noted.

(3) There are two four-member and **three 2-member sequences** that override WLE constraints or extend the repertoire respectively. The former set is documented for that purpose in the proposal, the latter set is **not listed separately** (it consists of the three NFC sequences containing Nukta which are shown in the table of repertoire). While the issue of Nukta and normalization is otherwise adequately addressed in 3.3.6 we do need **some mention of it in the repertoire section** and in the **XML description**.

**Recommendation**: document the fact that these sequences contain a code point that is not listed in isolation. For examples, see U+0931 for Devanagari or U+0E45 for Thai. For comparison, here is the language used in the Thai LGR:

“…The other two sequences were defined to restrict U+0E45 (THAI CHARACTER LAKKHANGYAO) from appearing in any context other than these sequences. Accordingly, while U+0E45 is not listed by itself it brings the total of distinct code points to 69.”

**GP response:** Done. The changes done in WLE and Defining Nukta in ABNF and WLE (as a subset). The description in XML was updated.

 (4) Apparently as a result of simplification of the WLE rules, the four classes C1, V1, M1 and CN in the **XML are unused and should be removed**. The same applies to the corresponding tags. (Unused tags are normally retained only when they describe basic categories of code points, such as "Consonant", but not when they are limited subsets that only are meaningful in conjunction with specific WLE rules.)

**Recommendation:** remove unused classes and tags. Because C1 like C2 is defined in Section 7, perhaps retain C1, but then also mention both of them in the <description> as well as making sure they are commented (see item(16)).
**GP response:** Removed unused classes and tags C1, V1, M1, CN. Using rules to handle CN.
All code points are commented.

(5) The comment for the definition of rule "P" should be changed to **explain what "P" covers**, instead of repeating the comment for the rule it is to be used in. The description in the **XML glosses** P as "Ra-Hasanta", which would be fine. So the comment could be something as simple as: "Ra-Hasanta, defined for use in WLE-7"

**Recommendation:** fix the comment
**GP response:** Fixed as suggested.

(6) The three sequences using Nukta have an annotation in the References and Comment column of the table in the document, e.g. "09DD is the preferred code point, however it is not available for LGR as per the standards **at the moment of this LGR development**". That wording unfortunately implies the counterfactual assumption that the standards' stability polices would allow some future change.

**Recommendation:** use slightly more neutral wording: “09DD is the preferred code point, however it is not available for LGR as per the standards **governing this LGR development**”.
**GP response:** Fixed as suggested, both in document (Table 8) and the XML.

(7) The description in the XML states: "Nukta, Zero Width Non-joiner (ZWNJ), Zero Width Joiner (U+200D), are not included in the repertoire. More details in Section "3.3.6 Nukta" and Section "3.3.8 Zero Width Non-joiner (U+200C) and Zero Width Joiner (U+200D)" of the [Proposal]." This can be seen as misleading as Nukta can occur in sequences. Pointing to the "details" in Section 3.3.6 may be acceptable for the proposal, but it should be possible for any reader to tell directly from the published LGR that labels may contain a particular code point, as is the case for all other such cases.

**Recommendation:** make sure the description identifies the total count of distinct code points that may be found in labels, as well as the count and brief discussion of any sequences and their purpose. See item (3) and other LGRs for ways to handle this.
**GP response:** Split the explanation in XML for Nukta and Joiners into two paragraphs.

(8) There's something weird going on with the **formatting some** of the comments in table 8. In at least some IP members cases we see boxes for all but letters d and r, even though the text is English. Word shows the selected font to be "Helvetica" and it might be that that maps to some strange font on some system; perhaps the format could be changed to use the same font as the table proper.

**Recommendation:** fix the font
**GP response:** (The font used is Cambria all over the TABLE 8).

 (9) The issue about the wording mentioned in (6) also shows up in the **XML comments**.

**Recommendation:** please make a corresponding change  **GP response:** Fixed

(10) The XML comments for the WLE rules refer to "Section 7", which is fine for the draft, and the "Element" LGR, but will not work well, when LGRs are concatenated for the Common LGR as part of integration.

**Recommendation:** No action (IP can address that in the integrated version).
**GP response:** Noted

(11) The variant relation between RA / RA WITH DIAGONAL has no comment. We've been using the convention of writing "<script name> variant" for these cases as that works well in isolation and in the common LGR. So, the comment on the <var> element should be "Bengali variant" using the script identifier, not the common name for the script. (Likewise for the two sequences 09B8 09CD 09A5 and 09B8 09CD 09B9).

**Recommendation:** please fix.
**GP response:** Fixed

(12) The **references** given (in both Table 8 and in the XML) are overly detailed. There is no need to track each letter to half-a-dozen grammar books. A single reference establishing that a code point is required for a given language should be enough. For example, the following references apply to ALL code points: [102], [103], [111], [121] - most of them appear to cover the origin of the writing systems. As such, they should be cited inside the supporting document when the script is discussed, but citations are not needed on every single character in the repertoire. The expectation would be that because there are three languages, that **up to three citations are sufficient for each code point**, one for the use of that code point for a given language.

**Recommendation:** the full set of references can be retained in the document, but they should be cited as part of the discussion of grammatical / linguistic features as appropriate, not for each code point in table 8. For the XML, only the references needed to establish the membership of a code point in some languages alphabet should be present.
**GP response:** Done. (up to three citations per code point added in Table 8 and Table 9), as well as in the XML.

(13) The URLs giving for references 127 and 128 proved unreachable. Ref 124 leads to the page giving the EGIDs level for Assamese (as that is a constant, it doesn't need to be cited for each code point, instead, it needs to be cited where the EGIDS level for the Assamese language is discussed)

**Recommendation:** please fix
**GP response:** Deleted the reference 127 128 as it is no longer referred to. Manipuri and Assamese are covered by other references

(14) The reference for the Ethnologue entry on the Manipuri alphabet is misspelled. Also, because the reference is to only one section, not the whole page it should be edited to read something like this:

[125] Bengali alphabet for Manipuri, found in Ethnologue, Manipuri (Meeteilon/ Meithei), accessed on 20.10.2019 Done

**Recommendation:** please fix
**GP response:** Done.

(15) The description has an unbalanced **HTML** element (the closing </p> is incomplete)

**Recommendation:** please fix
**GP response:** Done.

(16) In the **XML** we normally see comments for each character <class> element, such as comment=” Any Bengali consonant” for class C (look at other Neo-Brahmi LGRs – as published in LGR-3 – for further examples).

**Recommendation:** please add
**GP response:** Comments with the list of language using code points, similar to comments in LGR-3

(17) The sequences that are **variants** have neither comments nor references, they probably should at least get a comment in the **XML** (because there, the context of a separate table is not available). For example: “(added as variant)”

**Recommendation:** please fix
**GP response:** Comments added.

(18) The section on **variants in the XML** description should probably call out (by count, not individually) the four cross-script variants to Gurmukhi and Devanagari (two each) and the two sequences defined for in-script variants as well as the two versions of RA (and then link to the section in the proposal).

**Recommendation:** please add
**GP response:** Added.

(19) In the second-to-last para in p.2 of the DOCX, there is a missing space in the second line, and an extra space in the third, so it should read:

(ii) Then there is Middle Bangla Period - 1200-1800 AD, again divided into three stages: (a) Transitional Middle Bangla (1200-1300 A.D, for which no genuine specimens are found) [147], (b) Early Middle Bangla (1300-1500 A.D), and (c) Late Middle Bangla (1500-1800 A.D).

**Recommendation:** please fix
**GP response:** Done.

(20) In **XML** description “Overview”, after the title of the proposal a citation is missing.

**Recommendation:** please add: [Proposal]
**GP response:** Added.

(21) In **XML** description “Repertoire”, please make the changes suggested in other items above, but also consider replacing the last sentence with: “Each code point is tagged with the script or scripts that the code point is used with, a category value, and one or more references documenting sufficient justification for inclusion in the repertoire, see "References" below. For code points that are part of the repertoire, comments identify the languages using the code point.” to match other, published LGRs.

**Recommendation:** please edit
**GP response:** Done.

(22) In Section 7, S is defined, but it is not referenced in the rules. A statement is missing that **explains** (in Section 7) that **sequences of form S** (that is S1 and S2) are valid, even if not allowed by the other context rules. That needs to be added to allow readers to understand how to apply the rules without a need to read the entire document.

**Recommendation:** please add such language (‘’S’’ defined in WLE Rule no. 9.)
**GP response:** Added in Table 16 that S1 and S2 are valid, even they are not allowed by the other context rules.

(23) The inclusion of “S” in rule 7 as described in the document is redundant, because each “S” ends in an “M” and therefore may already be followed by Z. Suggest to delete “S” from the rule (as was done in the XML), but to add a line after “Example:…” as follows:

 “(Because S ends it M, Z may also follow S)”.

This would better align the XML and document in stating rule 7.

**Recommendation:** please make those changes
**GP response:** Z rules already takes care of such a situation. Deleted ‘’S’’.

(24) In the **XML**, please add at the front of the comment for each of the two Ya-phala sequences: “= ya-phala (S1); ” (or S2), ahead of the language names. (Also mention S1, S2 in the <description> when introducing these (see also item (3)).

**Recommendation:** please add
**GP response:** Done.

(25) There are two definitions of “C1” in Section 7 which are not the same. The second one shows up in Rule 1. Why not reformulate a bit like this:

1. C is a set of C and CN where CN is the set of normalized forms of {ড়, ঢ়, য়}.

This avoids the need for defining another C1 and repurposes the symbol CN in a natural way.

**Recommendation:** please fix
**GP response:** Done.

(26) In Section 7 there are a **few words that are struck out**. Those appear intended as manual indications of prior edits. Please delete. Affected words include “community”, “by”, “some”, “of the”.

**Recommendation:** please fix (and check for possible other cases elsewhere)
**GP response:** Done.

(27) **Section numbering**. The subsections of **section 5** are presented as 1., 2., and so on; it would be less confusing if they were numbered 5.1, 5.2, etc.; likewise for **section 7** (where there’s no 1. or 7.1, only a 2.). Because subsection 5.3.1 is numbered “3.1” it can be confused with actual subsection 3.1 which is inside Section 4. This should be fixed before publication or public comment.

**Recommendation:** please fix Section
**GP response:** Done.

(28) Table 10 “Excluded code points”. This table, to be consistent with all the other LGRs, should be reserved for code points that can ***never*** be found in a label. Therefore, **U+09BC should be presented in its own table**. For example, this is the way a similar situation is handled in the Myanmar LGR:

**Code point not used alone**

… U+1063 is excluded from repertoire since it will never be used alone. It will be used as sequence ST3 (Table 8-A).

|  |  |  |  |
| --- | --- | --- | --- |
| **Unicode Code Point** | **Glyph** | **Character Name** | **Reason for exclusion** |
| U+1063 | ၣ | Myanmar Tone Mark Sgaw Karen Hathi | Never used alone. Only used together with U+103A as …. |

**Recommendation:**
**GP response:** Done.

(29) In the **XML** description, for the WLE rules, please add the rule number for each WLE rule (as was done for the published version of the Devanagari LGR for example). Users report that it makes it easier for them.

**Recommendation:** please add
**GP response:** Done.

(30) **Section 3.2** states: “The number of D, B or X which can follow a V in Bangla may not be restricted to one.” This sentence is followed by a new paragraph (indicating a new though) that discusses the availability of D B or D X, but because of the break, it is not clear whether that statement is in addition to the previous one, or simply a more precise restatement. Note that with the rules as formulated it is *not possible* to have: V D D or V B B or V X X. But it is possible to have V D B or V D X, just as it would be possible to have C D B or C DX. **If it is desirable to allow, V D D or V B B or V X X** or any other combinations for Bangla, the rules would need to be adjusted (or any exceptional sequences would need to be added to the repertoire).

**Recommendation:** please review and change either the explanation, the rules or the defined sequences.

**GP response:** Done. It is not permitted orthographically to allow two chandrabindus, anusvaras or visarga AFTER after a Vowel. Therefore, VDD, VBB and VXX is INVALID. hæncha and hæn examples given and explained to remove the confusion.

(31) The **XML** applies WLE8 to the two sequences S1 and S2. However, the statement of the Rules in Section 7 limits rule 8 to “V”. Please add “S” if WLE8 is desired to apply as well. Alternatively, create rule WLE9: “S cannot be preceded by H.” Please also adjust the comment in the XML to match any changes.

**Recommendation:** please fix (Atiur-Neha)

**GP response:** Added rule 9, XML adjusted accordingly.