IP Feedback on Devanagari LGR Draft

Date: 2017-09-15

# Overview

This document contains some Integration Panel (IP) comments on the LGR-Proposal\_Devanagari\_20170908.docx document for Devanagari and the associated XML file. The comments are keyed by section number and place on page.

As this constitutes an intermediate level draft for the Devanagari LGR the focus by the IP has been on giving quick feedback, instead of the full scrutiny that would be given to an LGR about to be finalized for submission. In particular, the actual members of the repertoire will warrant a deeper look (at this point, the IP is missing some information to come to a full conclusion).

As for the XML file and the rules: a fuller review of these would require access to a corpus of Devanagari words or labels (covering most or all of the supported languages) for some automated tests. The IP comment reflects what can be done without such data.

# Comments on main document

§3.2: (p4, top)

As per the requirement of the LGR procedure, languages belonging to the EGIDS scale 1 to 4 have been considered.

The use of EGIDS scale is not mandated by the procedure. Code points required for languages that rank 5 at the scale have been considered by some GPs and, given the appropriate overall circumstances may be permissible in the root. (And in Boro, there is the corresponding exception for this LGR).

Perhaps instead: “In developing this LGR, all known languages with a level between 1 and 4 on the EGIDS scale have been considered. (See [MSR-2] for a discussion of the EGIDS scale).”

§3.3.1 (p. 5, line 2nd from last)

... is not empirical.

Misuse of "empirical" to mean its opposite. Would be better

either "emerges empirically from actual usage."

or "is not required by rule, but is rather a constraint that has emerged in practice."

§3.3.3 (p. 6, middle)

their corresponding Matras.

No correponding matras (which are essentially combining marks) are given in the illustration of four characters which follows. They should be inserted, perhaps as an extension of table 5, to preserve parallelism of table 5 immediately above.

§4.1.2.4 (p. 10, 2nd-to-last paragraph)

the Letter principle

This appeal to the Letter principle in order to exclude "Rare and Obsolete Characters" is not really in agreement with that principle as stated in the Procedure:

Only Assigned Code Points normally used to write words should be permitted. Assigned Code Points normally used for both words and other purposes should not be permitted.

The letter principle intends to restrict the repertoire to the space needed for mnemonics, without relying on the Unicode letter property, which is drawn both too narrowly and too widely. The principle does not state that all letters must be permitted.

These characters are, after all, "normally used to write words". In fact, the relevant rule is at B.3.4.1 ("Defining the Repertoire" on p. 29):

 The panel must exclude any code points used only for archaic or historical purposes

Therefore, the action is in accord with B.3.4.1 and also consistent with the Conservatism principle (which supports limiting the repertoire to those code points that are definitely needed).

§5.1 Code Point Repertoire (pp. 12-16)

There are a number of code points present in MSR2 which have been tacitly omitted from the repertoire proposed here.

In particular 0904, 090E, 0912, 0946, 094A all involve glyphs including the "short e" matra (which is seen naked in 0946). Others tacitly omitted are 090C, 0929, 0934.

Although omission is the default under the Conservatism principle, it would be useful if the GP indicated the motivation for all these omissions.

On p. 16, the mention of "DEVANAGARI LETTER RRA" might be glossed as "(also known eyelash RA)" or Reph.

Note: Unicode uses “eyelash RA” whereas the LGR here uses “Reph” instead of “RA”. There’s nothing wrong with that choice and the GP is consistent. For the non-native and non-specialist reviewer it would be nice if the LGR could note and explain that despite the variance in usage, the concepts are the same.

In the table: references to [Inscript]. By providing a single reference, it is not possible to understand from the table in section 5 which code points are required for which languages. Also, it would be useful if, in addition to the normative reference, the GP could provide alternative informative references, perhaps to openly available sources (e.g. omniglot etc, that would help the casual reviewer of the LGR). These might be usefully organized by language. (See Arabic LGR for an example).

§5.2 (p. 17)

"These rules need slight changes for different languages"

This formulation might invite a misunderstanding of the status of characters and rules proposed for the root zone. In this zone, characters and rules which have been admitted on the strength of their use in any language will be acceptable. It does not really matter, therefore that U+093C NUKTA is applicable for Hindi but not Marathi, or Eyelash RA in Marathi, Konkani, Nepali but not Hindi. If the character or sequence is admissible at all, it is admissible in root Devanagari labels without distinction.

Perhaps: “Starting from these Hindi Rules, several slight changes are needed to accommodate the additional languages to be supported in this LGR:”

§5.3.4: the term "decomposition" is fraught with possible misinterpretation given that it also occurs in the context of canonical normalization.

§6.2 about labels ending in Halant. See item ⑫ below.

§10 Bibliography

The 31 items listed here are not actually cited as references in the document. They may be useful to flesh out the bald inclusion/exclusion distinctions of INSCRIPT, but no such links or citations are actually stated.

While a bibliography is undoubtedly helpful to anyone wishing to further educate themselves on the background of the script and its usage, it does not, as such, do anything in documenting the basis for specific decisions made by the GP. In keeping with precedent from other LGRs, the IP expects that LGR documents identify specifically which (principal) languages require a given code point (with a reference identifying a source that provides that specific information, e.g. a listing of that language’s list of code points). A citation of a document that collectively lists many languages is not sufficient, the specific section or table should be cited in case the document covers multiple languages.

Online references are acceptable as are references that are perhaps slightly less authoritative than national standard (but more widely available). The authoritative references can always be given in addition; another option is to provide an appendix with excerpts from documents that cannot, as a whole, be shared freely. (See Lao/Khmer/Thai LGRs for examples of this).

Comments on the XML file

① The XML supplied is not valid: Tags cannot be assigned to sequences. (Lines 110, 113, 116, 119)
(This error should have been caught by any tool that validates against the schema from RFC 7940)

This has been "fixed" by removing the tags in question - it is not clear whether that affects the way the authors were expecting the WLE rules to function. (See attached XML with \_IP in filename)

② The XML supplied is not valid: "<look-behind>" must be first child element of enclosing <rule>" (line 178)
(This error should have been caught by any tool that validates against the schema from RFC 7940)

This has been "fixed" by moving the <class by-ref="nukta"> inside the <look-behind> element -- it is not clear whether that reflects the intent of the LGR authors. (See attached XML with \_IP in filename) — but see also item ④ which replaces this construct entirely.

③ The XML supplied is not valid: "Rule name contains an invalid character (space)" (Lines 37-38, 41-50, 53, 110, 113, 200)
(This error should have been caught by any tool that validates against the schema from RFC 7940)

This has been "fixed" by replacing the space by hyphen -- it appears that doing so would reflect the intent of the LGR authors as all instances of the rule name suffer from the same defect. (See attached XML with \_IP)

④XML: rule "follows-only-C-or-CN". As specified, this rule violates the restriction in RFC 7940 on use of the <anchor /> element:

 "A "when" rule (or context rule) is a named rule that contains any
 combination of "look-behind", "anchor", and "look-ahead" elements, in

 that order. Each of these elements occurs at most once, except if

 nested inside a "choice" element in such a way that in matching each

 alternative at most one occurrence of each is encountered.

 Otherwise, the result is undefined."

In the current LGR, there is a nested rule that contains another <anchor /> element, thus violating the restriction that only one occurrence of an <anchor /> element is encountered.

The nested anonymous rule can be restated as:

    <rule comment="nukta-preceded-by-consonant">
      <class by-ref="consonant" />
      <class by-ref="nukta" />
    </rule>

and will then be matched by any combination of CN. This modification has been carried out. (See attached XML file with \_IP).

The rule now matches the description in section 7 - note that a single context rule matches two numbered rules in the main document, because the contexts for H and M are identical.

⑤ XML: an action is defined as triggered by a context rule. (A context rule is a rule containing an <anchor>). A context rule MUST only be used with "when" or "not-when" attributes. (An action is not necessary for the evaluation of "not-when" or "when" conditions).

(This error is usually not caught by simple schema validation, but tools aware of RFC 7940 should be able to check for additional restrictions like these).

This has been "fixed" by deleting the superfluous action -- it appears that doing so would reflect the intent of the LGR authors. (See attached XML with \_IP).

⑥ XML: nuktated sequences should be sorted in code point order (Lines 109ff)

(The specification of the sequences reciprocal definitions of the variants is otherwise fully in line with how a symmetric LGR is specified - see RFC 8228).

⑦ XML: comments for nuktated vowels could be improved (Lines:
The comments should probably read like "nuktated vowel sing o" etc. as they
a) apply to the full sequence (and therefore should describe the sequence)
b) are not character names (and therefore should be in lower case)

(see suggested change in attached XML with \_IP in filename).

Note that XML-style comments (<!-- -->) like the one for the nuktated sequences in this LGR are discouraged: they are not preserved by some of the tools. If the information is important, it should be moved to "comment" attributes or to the <description> element. (In this case, a short remark about the nuktated sequences would be appropriate in the <h2>Repertoire</h2> and <h2>Variants</h2> section of the <description> element.)

(this is left to the GP, no suggestions in the XML).

⑧ XML <description> element is incomplete:

a) the <description> is missing an <h2>References</h2> section with the usual table of references cited by the description, in particular a (placeholder) reference to [Proposal]. (See any LGR proposals accepted for LGR-2 for examples)

b) Also missing a statement in the description that the XML follows RFC 7940 (and reference to [RFC7940]).

c) As there are in-script variants in the proposal (the nuktated sequences), the <h2>Variants</h2> section in the description should summarize these.

d) Likewise the <h2>Character classes</h2> and <h2>Whole Label Evaluation (WLE) rules</h2> sections should contain summary information on the classes and rules in the LGR (see Khmer, Thai, Lao for examples of how to do this).

⑨ XML: It would help if the comments on <rule> elements would number the rules using the same numbers 1-7 as found in section 7 of the LGR document.

It is OK for a single <rule> element to match more than one numbered rule in the main document, in that case, the comment should mention both numbers, e.g. 2 and 3 for the case of "follows-C-or-CN".

Note: One rule has an programming language escape sequence instead of a character code or numeric character reference: " \u253F" .

(The character reference has been changed to a code point to match the other rules; see attached XML file with \_IP).

⑩ XML: Named classes defined but not used:
    visarga
    anusvara
    chandrabindu

These class definitions could be removed.

⑪ The WLE rule 1 (follows-only-specific-C-or-V-or-M) could *optionally* be reformulated:

 <class name="consonants-before-nukta" comment="consonants that may be nuktated" >

       0915-0917 091C 0921-0922 092B
   </class>
   <class name="vowels-and-matras-before-nukta" comment="vowels and matras that may be
 nuktated in the Santhali language" >
       0906-0913 093E 0948
   </class>

    <rule name="follows-only-specific-C-or-V-or-M" comment="Rule for Nukta which restricts
 its preceding characters">
      <look-behind>
        <choice>
            <class by-ref="consonants-before-nukta" />
            <class by-ref="vowels-and-matras-before-nukta" />
        </choice>
      </look-behind>
      <anchor />
    </rule>

Alternatively, instead of defining a listed class, a tag could be added that indicates a letter is "nuktatable", the rule then changes to:

   <class name="nuktatable" from-tag="nuktatable" />

   <rule name="follows-only-specific-C-or-V-or-M" comment="Rule for Nukta which restricts its preceding characters">
      <look-behind>
      <class by-ref="nuktatable" />
      </look-behind>
      <anchor />
    </rule>

Note that the "tag" attribute on a <char> element may have multiple values, therefore the following is a legal example:

<char cp="0916" tag="consonant nuktatable" ref="0 101" comment="DEVANAGARI LETTER KHA" />

(How to state WLE Rule 1 in the XML format is a matter of taste - the existing specification is certainly valid and unobjectionable. The Lao/Khmer/Thai GPs have mostly decided to use this alternate style using multiple tags, but there is no hard reason why the NeoBrahmi GP would need to make a change in its approach; we are simply pointing out the availability of an alternative).

⑫ Document section 6.2/XML: the proposed rule for Halant ending is not reflected in the XML.

In theory, it is possible to create variants between Halant and the "null" element at the end of a label, and the RFC7940 format supports that. However, it should be investigated whether a more conservative approach could be used, such as simply disallowing the final halant. It would certainly be simpler and if the majority of users (as claimed in the document) do not expect a trailing halant, then it would seem that this feature cannot be one that is so strongly required that it must be provided at the cost of a complex workaround.

(Note that creating a variant mapping to the "null" element is at this point without precedent in the root zone LGR. It is not even known whether the tools would support this case correctly. It is also in some ways a slightly surprising feature of RFC 7940. Therefore, the GP is strongly encouraged to review whether this feature is really required — and the IP at this point withholds a final decision as to the circumstances under which this feature might be deemed acceptable for the root.)

## Comments on Repertoire and Rules

The IP has investigated the XML file for the LGR using several corpora of words in several languages using the Devanagari script. The aim of this investigation is to flag code points that may be in the LGR but do not show up in the samples or vice versa, so that the GP may be able to confirm that these are classified correctly.

From a Nepali corpus, there were half a dozen instances of 0912 Devanagari Vowel Sign Short O. For example,

शऒ्ढया‎ (0936 0912 094D 0922 092F 093E)

शऒ्ढयाका‎ (0936 0912 094D 0922 092F 093E 0915 093E)

शऒ्ढयाको‎ (0936 0912 094D 0922 092F 093E 0915 094B)

शऒ्ढयालगायत‎ (0936 0912 094D 0922 092F 093E 0932 0917 093E 092F 0924)

‎शैऒ्ढयाको‎ (0936 0948 0912 094D 0922 092F 093E 0915 094B)

It would be nice to confirm whether these represent typos in the data sample or a genuine omission in the LGR. (The text of the LGR does not address the reason for excluding 0912).

There were 192 instances of V following H and ~1300 examples of the some other character than C or CN preceding matras or halant. This may be due to a lack of quality in the corpus as it contained many “labels” that were displayed with visible “dotted circles” indicating that the layout engine encountered a combining mark in an unexpected context.

It is probably not useful to try to form any conclusions based on these particular invalid labels, other than that the LGR appears successful in identifying them. However, one of them appears to be the name “Afghanistan” अफगानिस्तान which has a 093E in an invalid context. It would be great if the GP could confirm that this is not an indication of a systematic issue with Nepali.

From a smaller set of Hindi data, only two invalid contexts were found:

 ‎सू़जन‎ (0938 0942 **093C** 091C 0928)

 ‎अरब़‎ (0905 0930 092C **093C**)

Both cases involve a 093C Nukta being applied to an invalid context. One involves VOWEL SIGN UU (0942) the other involves a LETTER BA (092C). It would be good if the GP could confirm that these are indeed typos.

In terms of code point coverage, the Nepali sample covered 61 of 65 code points, the Hindi sample somewhat fewer.

The IP will share additional data generated by this evaluation via ICANN staff.