IP Feedback on Devanagari LGR Draft

Date: 2017-11-24

Overview

This document contains some Integration Panel (IP) comments on the “NGP LGR for Devanagari ver 2 - 2017.docx” document dated November 17, 2017 and the associated XML file ("Proposed-LGR-Deva-20171117.xml”). The comments are keyed by section number and place on page.

The document reviewed constitutes an advanced level draft for the Devanagari LGR, as evidenced by the attention given by the GP to finalizing the discussion and documentation, as well as the completion of the repertoire based on additional language data. Nevertheless, there are number of items that are still missing for an LGR about to be finalized for submission.

In particular, a section discussing the need for cross-script variants must be added before submission, and any cross-script variants applicable to Devanagari need to be listed as part of the LGR – including in the XML (see discussion below).

Because a full review of the rules and their expression in the XML file would require access to a more complete corpus of Devanagari words or labels (covering most or all of the supported languages) for some automated tests, the IP comments in this round do not reflect any mechanical testing. The IP has been given to understand that test files for valid/invalid labels may be expected.

The IP appreciates that the NeoBrahmi GP was able to accommodate many of its previous comments have been addressed in this iteration of the LGR draft and in a very short amount of time; any comments fully addressed have been removed from this document (and a few new comments added).

In the following, text preceded by ➔ introduces an IP suggestion or comment, whereas regular text usually describes an issue or repeats a previous comment by the IP for discussion.

Comments on main document

Title (p1)

> ➔ in the title of the Devanagari LGR proposal there’s an “[LGR]”, which should be changed to “(LGR)” as it would otherwise be confused with a citation

§3.3.6 (p. 8, l. 2)

Previously, the IP had written:

“Use of nukta in other languages should be described here.”

> ➔ the IP notes that the discussion of Nukta has been revised; generally it appear much improved, however, see additional comments below.

§5.2 (p. 17)

Former §6.2 about labels ending in Halant.

Previously, the IP had written:

“The IP notes that the NeoBrahmi GP has elected to not implement a variant relation between an ending Halant and the “null” element. However, it would be interesting to hear from the NeoBrahmi GP why it is felt that a trailing halant should be supported at all. Would it not be more conservative to restrict this? If there is a good reason to not restrict this instance in spite of possible confusion to the “majority of Devanagari users” that was mentioned in the now deleted section, it would be good if the GP could make that case explicitly, rather than tacitly permitting this case.

Perhaps the previous section 6.2 could be restored (at a different place in the document, as this is no longer a discussion on variants), but changed to an explanation on why permitting trailing Halant is the correct choice for the Devanagari LGR.”

> ➔the new text in Section 6.3 is perhaps acceptable, although it doesn’t answer the question whether it would be a possible alternative to disallow trailing Halant altogether.

> ➔ Spelling of Halant: The IP notes some inconsistency in spelling “Halant”. IP suggests eliminating “halant” (lowercase) and “Halanta”.

§7 Whole Label Evaluation Rules (WLE)

Previously, the IP had written:

“Technically, most of the rules in the Devanagari LGR are context rules, however that is a bit of an implementation detail…

We note in the latest XML file, that there is a new rule called “follows-only-C” (in contrast to “follows-only-C-or-CN”), which is applied to the sequence U+093E U+093C. This new rule is not described in this section.”

> ➔ this rule still present but now unused in favor of “follows-only-C-or-CN”. In the IP-supplied XML it has been removed; if it is required, the GP can restore it and add the necessary source to make sure it is invoked and text to document it.

§7. (pp. 26-7)

Previously, the IP had written:

"It appears that the restrictions on *nukta* in Hindi may not be valid for Devanagari as a whole: e.g., in Konkani, as the Unicode standard 9.0, vol. 2, p. 462, states that *nukta* (U+093C) may be used after U+091A *ca*; this is confirmed by https://www.omniglot.com/writing/konkani.htm, which also adds use of *nukta* after U+091D *jha*). However, other sources (e.g. http://tdil-dc.in/tdildcMain/articles/285368Konkani%20Script%20Grammar.pdf) suggest that *nukta* is not used in Konkani at all."

> ➔ the IP notes the addition of two code points permitted as left-context of nukta, the full set now includes 091A but not 091D. See Rule "1" in section 7 or the HTML/XML; the purported Konkani usage appears to not be supported – which may be fine, but hard to tell based on the evidence presented, which includes only the result, not the data leading up to it. Perhaps section 3.3.6 could point to source material describing the use of Nukta (or absent written sources, could describe how the GP came to identify the needed code points allowed with Nukta).

§3.3.3 (p 9)

> ➔ is there a single or multiple reference documenting the specific set of code points to be used with Nukta? If so, it would be nice to annotate the description.

Comments on the XML file

> ➔ The Integration panel is appending a version of the XML file that carries out all suggested edits mentioned here. The following section is mainly intended to explain the issues found and why the IP believes the suggested edits are a good way to fix them. The IP requests the NeoBrahmi GP to review this section of the feedback document as well as the XML and HTML file supplied before making a decision whether to keep, modify or reject any suggested changes.

① The XML supplied by the GP (Proposed-LGR-Deva-20171117.xml) is valid except for item ②. It appears to match the specification in the document, but a few of the descriptive (human readable) elements and attributes are still deficient and would need to be fixed before this LGR can be finalized (see items ③ and others).

> ➔ there is an unused rule ("follows-C") – this has been removed

> ➔ the IP notes that the <description> is entirely new and substantially improved, see review below.

② The XML as supplied contained citations of undefined references..

➔ this is due to the fact that the renumbering of references was not carried through to the "ref" attributes (a version of the XML that fixes only this issue is appended).

③ XML <description> element :

➔ the IP notes that the <description> has been totally revised, it is far from incomplete, in fact, it may exceed in detail what is required for the XML file. See review below.

④ XML: The <references> element and reference numbering do not follow some of the conventions otherwise applied to the Root Zone LGRs.

➔ the IP notes that the GP started to renumber the references in the <references> element, but did not carry this through to the individual "ref" attributes, introducing a mismatch. The IP supplied XML file, which, in addition to renumbering ref="1 2 3" to ref="101 102 103" etc., contains the following additional modifications that will reduce the need to make changes during integration:

a) gives references to the Unicode version in which each code point was first encoded. These are [0], [8] … [11].

b) adds corresponding source references to Unicode Version 1.1 through Unicode version 6.1 in the format
used in the RZ-LGR.

c) adds boilerplate text in the references section of the description matching the text used in RZ-LGR

➔ The following are optional suggestions to the GP:

d) optional: remove references 106 and 107 from the XML as they are not used in that file
(this has not been carried out; optional for NeoBGP)

e) optional: add references [0], [8], [9], and [11] to the table in section 5 of the main document
(note that the numbering is discontinuous, as there is a fixed relation between each number and
a given Unicode version across the RZ-LGR)

f) if (e) is chosen, add references [0], [8], [9], and [11]to the main document's References section

Using consistent referencing conventions makes it easier to integrate the various LGR.

➔ except as noted, this has been addressed

⑤ XML Repertoire: matra U+0944:

➔ this matra has been removed (the XML and table in Section 5 now appear to match).

<description> Element in XML

➔ The following contains an excerpt of the <description> section with most suggested edits applied in place, followed by an explanation or further discussion of the suggested change introduced with a ➔. All suggested edits (and a few additional edits mostly outside the <description> element) have been carried out in the appended XML file.

Label Generation Rules for Devanagari script

Overview

This file contains Label Generation Rules (LGR) for the Devanagari script as would be appropriate for the Root zone. For more details on this proposal see "Proposal for a Devanagari Script Root Zone Label Generation Rule-Set~~Proposal for Generation Panel for Neo-Brahmi Scripts Label Generation Ruleset for the Root Zone~~ [Proposal]". The format of this file follows [RFC 7940].

➔ wrong proposal document

Repertoire

The NeoBrahmi Generation Panl (NBGP), ~~in Section 5 "Repertoire"~~ proposes 83 unique code-points to be made part of the Devanagari LGR [Proposal] in addition to two sequences i.e. 0931 094D 092F and 0931 094D 0939 which put the character U+0931 (DEVANAGARI LETTER RRA) in its own specific context beyond which it does not stand by itself. Accordingly, while U+0931 (ऱ) is not listed by itself it brings the total of distinct code points to 84.

A number of other sequences have been defined in connection with the definition of variants, bringing the total repertoire entries to 92 (see “Variants” below).

➔ The user will see 92 entries in the table, so this summary should prepare for that; if cross-script variants are found to be necessary, there will be “out-of-repertoire” code points listed as well.

The repertoire includes code points used by languages written in Devanagari that fall within [EGIDS] scale 1 to 4. Boro, Braj, Dhundari, Mundari, Kharia have also been additionally covered. Though listed in EGIDS scale 4, Saraiki is not covered ~~by the NBGP. As per ethnologue, for~~because Devanagari script is "no longer in use" by Saraiki community. ~~Ref: https://www.ethnologue.com/language/skr~~ (For more details, see Section 5 "Repertoire" in [Proposal]).

➔ for these details the reader can refer to section 5

The repertoire is based on [MSR-2], which is a subset of Unicode 6.3 [Unicode 6.3].

~~Each code point has associated Glyph, Character Name, Unicode General Category (gc), Indic Syllabic Category and Reference.~~

➔ this is true for the HTML version of the document, but not for the XML source. Perhaps best to leave this sentence out, for in the HTML version the table comes with a legend generated by the tool that explains all the data.

Variants

According to Section 6 "Variants", in "[Proposal]", this LGR defines variants which are "Confusing due to deviation from normally perceived character formations by larger linguistic community" These cases are not of mere visual similarity as they involve some deviations from the widely accepted norms of Devanagari Akshar formations. These can cause confusion even to a careful observer and are hence being proposed as variants. ~~Following are it's~~They fall into two broad categories:

- Vowel/Vowel sign followed by Nukta
- Unique Vowels and Vowel Signs required for Kashmiri

➔ best perhaps to make this a short list (using <ul> etc. HTML codes)

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.~~ All variants are of type “blocked”, making labels that differ only by these variants mutually exclusive: w~~W~~hichever label containing either of these variants is chosen earlier would be delegated, while the other one ~~equivalent variant~~ label should be blocked.

➔ perhaps a term other than “peculiar” would be preferable and “nature” is used a bit repetitively; see the suggested alternative phrasing.

In addition to these, NBGP plans to do cross-script variant analysis among all the scripts under NBGP ambit. That will be separately released.

➔ any cross script variants affecting Devanagari would need to be added to this proposal before it can go to public comment. This does not preclude the NeoBGP from releasing an overview document that shows all of them together for all scripts in question, but for purposes of Root Zone integration, the IP expects matters to be handled according to "How to specify an out-of-repertoire variant in XML" available online as https://community.icann.org/download/attachments/43989034/Out-of-Repertoire-Variants-2017-09-25.pdf

Character Classes

➔ The IP felt that the GP draft of this section went into unnecessary detail here; the deep background is best kept in the proposal document and simply cited by reference to the appropriate section. However, details on the position of a character in relation to the Akshar are of importance to any user of the XML file and should be summarized (briefly). Suggested edits to that effect have been carried out below.

Devanagari is an alphasyllabary and the heart of the writing system is the *Akshar*. It is this unit, which is instinctively recognized by users of the script. The writing system of Devanagari could be summed up as composed of Consonants, Implicit Vowel Killer: Halant, Vowels,Anusvara, Chandrabindu, Nukta and a Visarga.

Consonants: Devanagari consonants have an implicit schwa /ə/ included in them. To make a full syllable, consonants may be followed by certain code points from one of the other groups (see "WLE rules" below). ~~As per traditional classification they are categorized according to their phonetic properties (especially in terms of place plus manner of articulation). There are 5 Varga groups (classes) and one non-Varga group. Each Varga, which corresponds to Stops, contains five consonants classified as per their properties. The first four consonants are classified on the basis of Voicing and Aspiration and the last is the corresponding nasal. More details in~~See Section "3.3.1 The Consonants" of the [Proposal]

➔ we generally like to encourage GPs to limit the discussion of the background to issues that are directly visible to users of the LGR. The Varga groups in particular are not reflected in any way in the design, however, the fact that consonants can be followed by dependent vowels, halant, nukta etc. is of interest.

Halant: All consonants have an implicit vowel sign (schwa) within them. A special sign is needed to denote that this implicit vowel is stripped off. This is known as the Halant "◌्" (U+094D (◌्)). The Halant thus joins two consonants and creates conjuncts, which can be generally from 2 to 4 consonant combinations. In rare cases, ~~it~~ a conjunct can ~~join~~ contain up to 5 consonants. However, ~~the notion of maximum number of consonants joining to form one akshar is empirical. It is just an observation drawn from the words that have been observed till date. Given the confluence of languages happening in the Internet age, the possibility that one may want a generic Top Level Domain [gTLD] which may have more than the observed maximum cannot be ruled out. Hence, in the LGR work,~~ this LGR will not enforce any limit ~~will not be enforced~~. See section 3.3.2 "The Implicit Vowel Killer: Halant" in [Proposal].

➔ Again, for the purpose here, a bit less detail is probably better, if needed, put a "See section" pointer here, as done for Consonants, etc.

Vowels: There are separate code points for vowels that are pronounced independently at the beginning of a syllable or after a vowel sound. ~~Separate symbols exist for all Vowels, which are pronounced independently either at the beginning or after a vowel sound.~~ To indicate a Vowel sound following a consonant, other than the implicit ~~one~~shwa sound, a Vowel sign (Matra) is attached to the consonant. ~~Since the consonant has a built in schwa, T~~there are equivalent Matras for all vowels excepting the U+0905 (अ). ~~More details in~~See Section "3.3.3 Vowels" of the [Proposal]

➔ suggestions for possible alternate phrasing to make the discussion more compact without losing essentials.

Anusvara : The Anusvara ~~represents a homorganic nasal. It replaces a conjunct group of a Nasal Consonant+Halant+Consonant belonging to that particular varga. Before a non-varga consonant the anusvara represents a nasal sound~~ (showing a nasal at the end of a syllable) follow a vowel, matra, consonant or nukta. ~~More details in~~See Section "3.3.4 The Anusvara" of the [Proposal]

C~~h~~andrabindu : C~~h~~andrabindu denotes nasalization of the preceding vowel. ~~as in आँख /ãkh/ eye (U+0906 U+0901 U+0916 (आँख)).~~ Present-day Hindi users tend to replace the c~~h~~andrabindu by the Anusvara. It can follow a vowel, matra, consonant or nukta ~~More details in~~See Section "3.3.5 Nasalization: C~~h~~andrabindu" of the [Proposal]

➔The document has been change to “candrabindu” from “chandrabindu”; the XML should follow the same convention.

➔Instead of describing in too much detail the function of these two in the writing system, it might be more useful for readers at this level to know where they can occur in the syllable (and therefore in the label).

Nukta : The nukta sign is placed below a certain number of consonants to represent sounds found only in words borrowed from Perso-Arabic. ~~It is pre-dominantly used in this manner in Bodo, Hindi, Kashmiri, Maithili, Santhali and Sindhi.~~ It is also placed under ~~"ड़" (U+0921 (ड़))~~ and ~~"ढ़" (U+0922 (ढ़))~~ to indicate flapped sounds~~. Apart from this,~~ and the Santali language uses Nukta ~~in a unique way~~ adjoined to certain Vowels and Vowel signs. Vowels that are followed by nukta may not be reliably distinct from vowels without nukta by large part of the user community, they should therefore be mutually exclusive in the same position in the label (See Variants, below). ~~More details in~~See Section "3.3.6 Nukta" of the [Proposal]

➔ again, IP suggestion is to focus on the result, that is, now Nukta is treated in this LGR.

➔ Note that U+xxxx in the XML source for <description> will be expanded to to U+xxxx (x) that is, with glyph shown, in the HTML version of the file therefore the usage “x” U+xxxx would lead to the redundant “x” U+xxxx (x).

Visarga ~~and Avagraha~~: The Visarga (U+0903), representing an aspiration at the end of a syllable, is frequently used in Sanskrit. is frequently used in Sanskrit. ~~and represents a sound very close to /h/. दुःख /du:kh/ sorrow, unhappiness (U+0926 U+0941 U+0903 U+0916 (दुःख)). The Avagraha "ऽ" (U+093D (ऽ)) creates an extra stress on the preceding vowel and is used in Sanskrit texts. It is rarely used in other languages using Devanagari. In case of LGR, the Avagraha is not part of the repertoire as it is barred in the Maximal Starting Repertoire. More details in~~See Section "3.3.7 Visarga and Avagraha" of the [Proposal]

➔ some suggested edits to tighten up the text; it is generally not appropriate for this summary description to delve into details of excluded code points, particularly those not even members of the MSR.

Whole Label Evaluaton (WLE) rules

Default Whole Label Evaluation Rules

The LGR includes the set of required default WLE rules and actions applicable to the Root Zone and defined in [MSR-2]. They are marked with ⍟.

Devanagari specific Rules

These rules ~~have been drafted to~~ ensure that the ~~prospective~~ Devanagari label conforms to akshar formation norms ~~as desired in~~for the Devanagari script. In this LGR, t~~T~~hese norms are exclusively presented as context rules.

~~Following~~ The following symbols are used in the comments and names for WLE rules:

- C → Consonant
- M → Matra
- V → Vowel
- B → Anusvara (Bindu)
- D → Chandrabindu
- X → Visarga
- H → Halant / Virama
- N → Nukta
- S → Eyelash Reph (C1HC2) where
    - C1 is 0931 (र - DEVANAGARI LETTER RRA)
    - H is 094D (◌् - DEVANAGARI SIGN VIRAMA)
    - C2 is either - 092F (य - DEVANAGARI LETTER YA) or 0939 (ह - DEVANAGARI LETTER HA)

➔ please consider making the above a list (as shown) using nested <ul> elements in HTML.

The rules are:

- 1. N: must be preceded only by either of specific set of Cs, Vs and Ms
- 2. H: must be preceded by C or CN
- 3. M: must be preceded by C or CN
- 4. X: must be preceded by either of V, C, N or M

- 5. B: must be preceded by either of V, C, N or M
- 6. D: must be preceded by either of V, C, N or M
- 7. V: Can NOT be preceded by H

SeeMore details in Section "7 Whole Label Evaluation Rules (WLE)" of the [Proposal]

Overall Development Process and Methodology

Under the Neo-Brahmi Generation Panel, there are many different scripts belonging to separate Unicode blocks. Each of these scripts will be assigned a separate LGR; however, Neo-Brahmi GP will ensure that the fundamental philosophy behind building those LGRs are all in sync with all other Brahmi derived scripts. This is the Devanagari LGR, which caters to multiple languages written using Devanagari belonging to EGIDS scale 1 to 4.

For additional details and contributors, see Sections 4 and 8 of [Proposal].

➔ there's generally a pointer to the section defining the participants here (see LGR-2).

References

Reference [0] refers to the Unicode Standard version in which corresponding code points were initially encoded. Reference [100] and up correspond to sources given in [ProposalFollowing] for justifying the inclusion of for the corresponding code points. Single code point or ranges may have multiple source reference values.

In addition, the following references are cited in this document:

➔ the above boilerplate is used in all RZ-LGR documents, adding it now, reduces the needs for edits during the integration.

Following references are cited in this document:

[MSR-2]

Integration Panel, "Maximal Starting Repertoire — MSR-2 Overview and Rationale", 14 April 2015
https://www.icann.org/en/system/files/files/msr-2-overview-14apr15-en.pdf

[Proposal]

➔ document name, author,date, url missing for this, GP should at least put something in as a placeholder, so it cannot be forgotten during the publication process:

Neo-Brahmi Generation Panel, "Proposal for a Devanagari Script Root Zone Label Generation Rule-Set (LGR)", [date TBD][link TBD]

[RFC 7940]

Davies, K. and A. Freytag, "Representing Label Generation Rulesets Using XML", RFC 7940, August 2016, http://www.rfc-editor.org/info/rfc7940.

[EGIDS]

Expanded Graded Intergenerational Disruption Scale, https://www.ethnologue.com/about/language-status (Accessed on 13th Nov. 2017)

[Unicode 6.3]

The Unicode Consortium. The Unicode Standard, Version 6.3.0, (Mountain View, CA: The Unicode Consortium, 2013. ISBN 978-1-936213-08-5) http://www.unicode.org/versions/Unicode6.3.0/

For more details for references [100] and up and [0] and up refer to the Table of References below.

➔ another boilerplate sentence added