Response of NBGP to IP feedback on Devanagari received on 29th Mar. 2018

DATE: 2018-04-27

# Overview

This document provides NBGP response and actions taken on the feedback provided by IP on 29th Mar. 2018 to the Devanagari proposal dated 2018-01-30.

## General comments

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Issue** | **IP Comment** | **NBGP Response** |
| Previous responses: | The document entitled [Response of NBGP to IP Response: Devanagari LGR Proposal on 2018-01-02] appears to the IP to close all previous issues, hence will not be further referenced. | IP thanks the GP for their ready responses. | ☺ |
| Spelling | There are two competing spellings of one language name: is it to be "Santali" or "Santhali"?  Different sources disagree on the spelling, however, it is preferable to settle on a single spelling, preferably used in the technical literature or standards. | Standardization is desired.  Suggest “Santali” as this is the spelling in the Unicode CLDR database as well as the ISO 639-2 list. | All changed to Santali. |
| Cross-script variants | There is an interdependence between scripts that share mutual cross-script variants.  Because of the simultaneous presence of in-script variants in some cases, this interdependence is stronger than in the case of the European scripts. | It may be better to hold the integration of mutually related Neo-Brahmi scripts (Deva + Gurmukhi), but also (S. Indian scripts + Sinhala) until there is two-way consensus, on each cross-script variant pair of code points. (It may even be advised to submit any such script pairs to Public Review simultaneously). | It has been done in NBGP F2F meeting in Delhi on 23rd-25th April 2018.  Following actions taken:  1. 0902 and 0A02 have been made cross-script variants in both LGRs.  2. 0A4B has been added as Cross-script variant with 0946 and 0947. |

## Comments on main document (.docx)

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Issue** | **IP Comment** | **NBGP Response** |
| p. 10 | The footnote (no. 7) names explicit source info for the incidence of Nukta in Devanagari-using languages, but does not thereby make it directly accessible (as the Omniglot references are). | It is unfortunate that the citation is so indirect, as it makes could make for problems consulting this reference in the future, if disputes arise, or extensions are sought. But at least the current position, and its motivation, is now explicit.  The IP requests a private extract of the reference with supporting data for its files. | It will require some time to collate such private extracts. NBGP will try to provide the same in the next version.  It should also be noted that not all the cases could be substantiated with private extracts. NBGP still will try to cover maximum instances. |
| p. 32 | This page says that the characters in Appendix A "Variants based on pure visual similarity" are not to be classed as variants. This title is confusing. | IP suggests to identify characters (or sets) as "confusables" or "equivalents"  when their status is being considered, and to list them as "variants" when it is proposed that they be linked by variant rules. (“Candidate variants” may also be appropriate,)  NOTE THAT THIS IS A POINT THAT WILL APPLY TO ALL THE NEO-BRAHMI SCRIPTS. | Changed as required. |
| p. 34, item 1 | This section on Halant refers to " a section of community"  and "this section of community", without identifying it further. It is non-native style to use "community" as a mass-noun in this way, and in any case unclear what community, if any, is being indicated. | Perhaps it would be clearer simply to say "some users" and "these users", where this refers generally and vaguely to users of Devanagari script, whatever language they may be writing. Evidently, it would be more cogent if the GP could characterize these users as a group in some way, but this may not be possible. | Changed as suggested. |
| p. 34, item 2 | "As the final Halant is clearly visible, this is being considered as a variant belonging to Group 1 (i.e. of pure visual confusability)".  This wording presents needless multiple paradox, since these are not to be variants, and furthermore are not visibly confusable at all. If anything, they are cases of pure phonetic confusability, but some visible similarity. | It would be better just to say:  "In these cases, the presence or absence of final Halant is clearly visible, and there is no apparent case to make them variant pairs.  Eventually, in the light of practical experience, future NBGP revision may assess if these cases need to be considered as a variant pairs." | Changed as suggested. |
| pp. 35-37, cross-script homoglyphs with Gurmukhi | This section proposing cross-script homoglyphs with Gurmukhi is quite new as part of a Devanagari proposal.  Its proposal is largely the same as, but sporadically different from, the cross-script homoglyphs proposed in the Gurmukhi  proposal (occurring in §6, Group 2, of the Gurmukhi proposal .docx).  No explanation is given for the discrepancies, or even mention made that the Gurmukhi GP has opened this subject.  There are a number of loose ends (listed in Note 1 below) which suggest that this proposal is not yet in a final agreed state with all those involved (in Devanagari and/or Gurmukhi, let alone others in NBGP). | It is clear that there must be uniformity of opinion between Devanagari and Gurmukhi GPs before the proposal can move forward.  Please review the issues in Note 1 below and give clear picture of the sets of variants, separately addressing the series of in-script sets from that of cross-script sets, and stating explicitly (in the .docx) what its criteria for inclusion and exclusion have been. | It is done in NBGP F2F meeting in Delhi on 23rd-25th April 2018.  Added |
| p. 49 | Appendix A: This table (no. 19) of "Variants based on pure visual similarity" is explicitly not a table of variants (p. 32 above). | This is unnecessarily confusing. Change the titles to reflect GP’s intention.  (Note our earlier comment on when not to call something a“variant”) | Changed as required. |
| p. 50 | Appendix B: This table (no. 20) of "Cross-script Variants" is explicitly not a table of variants, but of pairs of characters, as between Devanagari and other scripts.  They have nothing to do with the immediately comparable Appendix A, which are all Variants within Devanagari (though chosen for confusability). | Their status is not clear, but they are not being proposed as variants, and so should not be called ”Cross-script Variants".  Some instances of the words "variant" may need to be changed there and in section 6 where it points to Appendix B and also the caption for Table 20). | Changed as required. |

## Comments on LGR specification (XML)

**Note:** because some of the issues with the XML file prevented the IP from opening / processing the LGR, the IP has created an XML document that corrects a number of these issues, as noted below. This corrected XML document is attached, and the GP is encouraged to review it, and use it as a base from which to make further edits.

**Note from NBGP:** Changes reviewed. Noted. Response below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Issue** | **IP Comment** | **NBGP Response** |
| Dating error: | Note that the date on the .html and .xml is 2017-11-21, way out of line with the date of the proposal document .docx at 30 January 2018. | Correct the dates | Corrected. |
| Philological error | “words borrowed from Perso-Arabic” | Should be  words borrowed from Perso-Arabic, English and other non-Aryan sources. | Changed. |
| (1), (6) Minor issues | Several sequences were listed twice. There was a minor typo or two in the <description>. | The duplicate listings were merged.  typos fixed | NBGP thanks IP for the modified file. |
| (2) | Several non-reflexive mappings were given "type" of "out-of-repertoire-var " | these have been corrected to "blocked" | Accepted and merged. |
| (3) | Code point U+0902 has both in-script and cross-script variants (to U+093A and U+0A02). | This means that all three code points have to have 2 variant mappings for symmetry and transitivity (and U+0A02 has an extra reflexive mapping). This has been corrected. | This point is being escalated to Gurmukhi LGR and will be subsequently adopted as per agreement with Gurmukhi LGR. For now the mapping change is accepted in Deva LGR. |
| (3a) | one variant mapping was entered incorrectly from code point U+0930 to U+0A15. | This has been corrected to a mapping from U+0935. | Thank you. It was inadvertent. |
| (4) Re automatic generation of .html files from .xml | In order to improve documentation of sequences in the Root Zone LGR, the IP now provides "names" for ad-hoc sequences in the HTML files generated during integration.  These names follow the pattern: NAME + NAME + NAME where NAME is a character name as extracted from the Unicode character database, and spaces surround the "+" signs.  Because "+" is not a valid character in formal Unicode names, there is no confusion as to the status of such ad-hoc names. (We've used "plus" before, but shall deprecate that practice).  (Note: an "ad-hoc" sequence is one that is not defined as a "named sequence" in Unicode, but defined in the LGR).  If the XML contains comments that start with a character name or sequence name or ad-hoc sequence name, the IP automatically strips that part of the comment (plus one more character for a presumed delimiter) and places the remainder of the comment into the "comment" field in the HTML table. | Therefore, any character names given in XML comments should follow the exact Unicode character name (or the convention for sequences described at left) to avoid being listed twice in the HTML.  NOTE THAT THIS IS A POINT THAT WILL APPLY TO ALL THE NEO-BRAHMI SCRIPTS.  IP has corrected a few names to match the convention for ad-hoc sequences.   (If an alias is desired, such as “halant” for VIRAMA, then by convention recognized by our tool, it should be placed after the Unicode name, separated, by “ = “ (SPACE EQUALS SPACE). | Accepted. |
| (5) to implement a convention for commenting variants / reflexive mappings. | For reflexive mappings of type "out-of-repertoire-var" the comment should not be "identity" but "out-of-repertoire". For all cross-script mappings whether to/from a code point in another script should be commented as "cross-script homoglyph" or "cross-script variant" depending on whether the code points are homoglyphs or other variants. For "in-script" mappings, other, relevant comments should be chosen, for example "Latin homoglyph" or "Devanagari variant" to identify these mappings as being within a script. | This commenting convention is especially needed for integration, because for non-reflexive mappings these comments will show in the common (merged) LGR, and all scripts contributing to a cross-script mapping will contribute comments. If the comments match, they can be folded mechanically, otherwise they would have to be manually merged.   IP will have to convert all non-conforming comments anyway. Therefore it is better to align comments in advance. Some corresponding edits have been made in the XML. | Noted. Accepted. |
| (7) Order of elements in XML | In order to be able to compare XML source code, IP needs certain items in a fixed order. GP had re-ordered all the attributes for the "char" elements. | IP has replaced them again in the order that we need them in to be able to do minimal source level diffs.  Please do not change the order of attributes etc. between versions.  NOTE THAT THIS IS A POINT THAT WILL APPLY TO ALL THE NEO-BRAHMI SCRIPTS. | Noted. Accepted. |

## Notes on Proposed Cross-Script Homoglyphs between Devanagari and Gurmukhi.

IP had found a number of inconsistencies between the versions of these proposals received by the IP before Feb 27 2018. Since then, the second version of the Gurmukhi proposal (20180302a), has addressed some of the issues. Nonetheless, the NeoB GP should review both the Devanagari and the Gurmukhi proposal to ensure that the next revision of both proposals will be consistent.In particular, IP notes:

1. There are some instances where one of the scripts maps two distinct code points to a single destination in the other script.

Deva U+0902 🡪 Guru U+0A02; but also Deva U+093A 🡪 Deva +0902; so by transitivity Deva U+093A 🡪 Guru U+0A02

Deva U+0946 and Deva U+0947 🡪Guru U+0A47

1. In the case of Deva U+902 and Deva U+093A, these two are explicitly declared as in-script variants in Table 17, however, in Table 18 an entry for Deva U+093A 🡪 Guru U+0A02 is absent, despite the entry for Deva U+0902🡪 Guru U+0A02.

**NBGP Response: This has been accepted in Devanagari as well as Gurmukhi LGRs. Now 093A is cross-script variant with 0A02. Same is harmonized with Gurmukhi LGR.**

By contrast, In the case of Deva U+0946 and Deva U+0947, these two are explicitly declared as in-script variants in Table 17, motivating the dual mapping.   
  
In this example, two principles collide. One is the desire to map Kashmiri vowel signs to more common ones, to prevent their use in spoofing (in-script variants).

The other is the principle of assigning cross-script variants based on homoglyph (or near homoglyph) relations. The latter would argue for a different mapping, viz.

\* Deva U+0946 🡪 U+0A4B

Following both principles would have introduced an unwelcome in-script variant in Gurmukhi

\*Guru U+0A47 🡨🡪 U+0A4B

While this may be an acceptable reason to avoid mapping 0946 to 0A4B, it does create an inconsistency, and therefore requires an explicit discussion in the LGR with an explanation of the issue and reason for the choice.

**NBGP Response: This has been accepted in Devanagari as well as Gurmukhi LGRs. Now 0A4B is cross-script variant with U+0946 and U+0947. Same is harmonized with Gurmukhi LGR.**

1. In the case of Guru U+0A27 🡪 Deva U+092A and Deva U+092F, these mappings would lead to an in-script variant between Deva U+092A and Deva U+092F. No such in-script variant is defined anywhere in the Devanagari LGR.

This issue must be resolved, whether by updating the Devanagari LGR or the Gurmukhi LGR.

**NBGP Response: It is agreed by both Devanagari and Gurmukhi LGR teams that 092F and 0A27 are not variants of one another. Devanagari LGR mentions it in non-normative section but not as a variant pair. Hence this issue does not arise at this point as per NBGP.**

1. The available version of the Gurmukhi is unfortunately still tentative, including a "temp" table of 4 Dev-Gur code point pairs (again not a disjoint set of pairs, and there is more overlap with code points in the pairs already proposed - U+091F, U+0922) was left "for discussion" by the Gurmukhi proposal (end of §6 ). These are not mentioned at all by the Devanagari proposal.

**NBGP Response: U+091F and U+0922 are part of the Devanagari Cross-script variants. Same is expected to be part of the Gurmukhi LGR.**

1. The IP would like to see included in the LGR proposals some (very brief) summary of the process of cross-script variant resolution among the member of the NeoB GP and the various script panels in particular.

This would help to demonstrate that the resulting sets (after they are fully aligned) do in fact represent the agreement of the communities involved.

**NBGP Response: Added in section 6.5.**

The IP would need to see anboth an updated Devanagari and an updated Gurmukhi proposal to be able to ascertain that the above issues have been mutually resolved.

## Note on Proposed In-Script Homoglyphs in Devanagari.

For Devanagari there are some proposed conjuncts as homoglyphs, set out below.   
(The font used here does not stack the conjunct as tightly as some other fonts. But the IP judges that there is sufficient similarity to justify the proposed variants.)

(Note that variant set numbering is from a trial integration performed by the IP, not from the Devanagari proposal).



In future, IP will need to check whether there are other comparable cases in the other script pairs.

NOTE THAT THIS IS A POINT THAT WILL APPLY TO ALL THE NEO-BRAHMI SCRIPTS.

**NBGP Response: 0A72 - 092A 094D 091F was proposed by Devanagari team. As 0A72 is not part of the repertoire of the Gurmukhi, it has been dropped.**

**0A72 was again added by IP to the XML in this recent review. The same has been removed.**

**Rest of the case are already part of the proposal.**

## Comments on Test Labels

The corpus data available to the IP is split across languages and is therefore a bit unwieldy (and still needs to be fullyadd to test suite). Therefore, no corpus analysis for Devanagari in this round.

There is a Test Labels file which is well categorized by context rules. Our tool matches the suggested failure modes for invalid labels, and returns all intentionally valid labels as such. The test coverage is excellent for such a short file. While only 63 of 84 code points are present, all 8 defined tags and 5 named classes are covered, as are all 4 of the context rules (both match and failure).

IP had been concerned that test labels should be explicitly drawn from the languages besides Hindi, since the proposal mentions at least 10 such languages. Such languages are usefully associated with specific characters in section 5.2 Code Point Repertoire.

IP is pleased to acknowledge that in the latest test file NBGP has given labels in 11 languages

NOTE THAT THIS MULTILINGUAL DIMENSION OF TESTING WILL BE RELEVANT TO ANY OTHER NEO-BRAHMI SCRIPT THAT IS USED WITH MULTIPLE LANGUAGES.

There is no invalid label starting with a combining mark, therefore the WLE rule (leading-combining-mark) is never matched. (Also, most combining marks for Devanagari also have context rules that require them to follow some other code point. Because a leading combining mark for Devanagari fails this context check, the general rule is never invoked).

As the coverage is so good, it's worth looking at the details: checking various combinations of character classes possible under the context rule, compared to the theoretically possible permutations, only about 25% are not covered.

Review of the effects of cross-script variants for Deva/Guru is still outstanding due to the fact that the two proposals were not fully aligned when first reviewed (and the specification of the variants in the Deva XML was broken).

Conclusion: The corpus is not conveniently organized for simple testing of "Devanagari" across all/most languages; individual languages were tested in an earlier round, resulting in high confidence in the appropriateness of the WLE rules. The Test Label file promises to be highly effective for regression testing. (Some testing of the effect of cross-script variants is still needed, but cannot proceed until Deva and Guru proposals have been aligned).