Response of IP to Public Comment on NBGP proposals

DATE: 2018-11-06

# Documents reviewed

The responses which have entered into this consideration:

|  |  |
| --- | --- |
| Devanagari  | Liang Hai, Sanjeev Goyal, Vivekananda Pani (also prior comment by IP) |
| Gujarati  | Liang Hai  |
| Gurmukhi  | Liang Hai |
| Kannada, Telugu | Liang Hai ((also prior comment by IP) |
| Tamil  | Liang Hai  |
| Malayalam  | Thin Zar Phyo, Cibu |
| Oriya  | Liang Hai, Subhashish Panigrahi, O Foundation, Santosh Mohanty, Sahoo Anindita |

These IP considerations are not intended to replace the review of public comments by the NBGP. In fact the IP would like to encourage members of NBGP to read closely these various public comments (accessible at <https://icann.org/idn>), and implement changes which are deemed appropriate.

# Devanagari

Liang’s discussion of section 3 is largely a critique of generalizations about the role and relation of certain characters within Devanagari as a whole, and of the transcriptions used in describing them; not very important for the business end, as long as code points that are distinct are listed as distinct in the repertoire – e.g. anusvara and bindu are distinct, even if confused in much informal use of Devanagari.

But Liang is understandably confused about the status of ZWJ and ZWNJ, given § 3.3.8’s rather vague and indeterminate description of their status within the LGR. § 3.3.8 should state clearly, at least in summary at end, that both characters are excluded from the Root Zone. (The IP maintains that describing how these characters are used is useful, even if they are not allowed in the Root Zone – with the information users and reviewers will better be able to gauge the effect of the omission)

Furthermore, he is right to require consistency on the exclusion (of purely Sanskrit, vocalic characters , and as matras as well as vowels) in § 4.1.2.4. This is correct in the repertoire, allowing

|  |  |  |  |
| --- | --- | --- | --- |
| 090B | ऋ | DEVANAGARI LETTER VOCALIC R | Vowel |
| 0943 | ृ | DEVANAGARI VOWEL SIGN VOCALIC R | Matra |

but not long vocalic R, or long or short vocalic L. But it is not the same (and so perhaps confusing) in the charts of MSR in §5.1. (There is some failure in the labelling of charts too: e.g., presumably the charts of MSR in §5.1 should given as Table 5.)

Liang notes that the list of Devanagari cross-script homoglyphs listed for Bengali is far shorter than that for Gurmukhi, calling for additional recognition of candrabindu, nukta, vowel signs aa, ii and uu, virama ... and certain consonant letters. IP has also noted the particular absence of visarga. Evidently, the inclusion of most of these (and perhaps all of them, if no “certain consonant letters” can be identified) depend on the presence of just one letter homoglyph (Deva 092E म = Beng 09AE ম). Otherwise (since all the other candidates are not letters) no homoglyph labels can be constructed. But in principle, it would be possible to accept all that Liang is suggesting.

Sanjeev Goyal draws attention to the almost imperceptible presence of nukta in two contrasting labels from the .Bharat registry: the Deva equivalents of nyūj.bhārat and nyūz.bhārat . These are distinct, but perhaps on the border of visibility. As it stands, nuktas have not been involved in in-script variants (except in the marginal cases where Nukta+ Vowel = Vowel): it would require a major change in the variant rules to accept pairs *with and without nukta as virtual homoglyphs*.

Vivek Pani’s comments are thoughtful, and probably represent the view of the “literate layman” in coming to the regulation of the root-zone for Deva. It is noticeable how concerned he is about nukta and halant – possibly emphasizing how hard these are to distinguish from their (respective) absence. He also is relaxed about the absence of ZWJ, but concerned at the absence of ZWNJ. This cannot change anything, but is worth notice. It may be beneficial for NeoB GP to describe the issue, so that anyone using their document as source can know how to prioritize these two in other zones, e.g. when laying down conditions for Second-Level domains. (Likewise, the Cyrilliv GP has listed confusables “for information”, going beyond what might be justified as homoglyphs, hence proposed variants.)

While there is little that requires change in his arguments, his points are all worth pondering, and items 7 and 8 do require specific corrections in the LGR text.

# Gujarati

Liang’s points all seem reasonable, although only the comment at §5.2 will affect the repertoire. (We might look to NBGP to note and explain why Gujarati differs from Devanagari in its choice of these vocal R and L sounds, and in consistency between the vowel letters and the vowel sounds?)

# Gurmukhi

As ever, all points raised by Liang Hai should be accommodated as far as they seem reasonable and/or helpful to NBGP. In particular, requests for clarification probably need active attention from NBGP

§3.3 “Has the well-kown post-base form of YA already fallen out of use in common text? Probably should mention this.” This is a mystery to this IP reader. But we assume that the writers of the Gurmukhi LGR assumed this was not relevant to modern Gurmukhi. If the reference is to YAKASH U+0A75, it is explicitly excluded at §4.1.4. This looks like there’s a need for positive clarification of the decision taken by the GP.

§3.3.2. (2nd para.) is a strange summary of the Gurmukhi case: “Both Hindi and Punjabi-Gurmukhi orthographies allow implicit dead consonants. It’s just that Punjabi-Gurmukhi allows more.” But Liang is presumably referring to the fact that, at the end of Hindi and Punjabi words, consonants automatically lose their final implicit vowels in pronunciation (while in Punjabi this can also happen at syllable boundaries within words too).

But as he goes on to say, “this level of spelling and reading rules are not really relevant” here, where we are concerned only with the incidence of halant/virama CPs and not with what the effects are on pronunciation. It would only be relevant if users would ever ignore the presence of a virama (if seen as redundant) and therefore equate two labels. (After all, all the apparent combinations with final HA, whether as part of aspirated consonants (GHA, JHA, DDHA, DHA, BHA), or in pairing with NA, MA, RA, LA, RRA) actually give altered tone, rather than a phonetic [h]. But this is not mentioned, nor need it be.)

§ 3.3.4 “Suprasegmental” is a bit strange as a term (≠ linguistic use, to describe phonetic features such as tone, nasality which cover whole syllables or words), but can stand, if confusingly, especially because the three signs Addak, Tippi and Bindi are placed above their affected segment.

# Kannada and Telugu

in answer, IP has raised the issue of the unresolved difference between Kannada-Telugu on the one hand, and Sinhala on the other over, over cross-script variants. IP pointed out that this should be directly discussed between the parties, and if a decision was made to accept cross-script variants as suggested by Kannada-Telugu, more specific evidence or reasoning must be presented to justify it.

## Kannada

The corrections at 2 and 3.4.7.4 are well-taken, it seems. Substitute Kannaḍa for Kannada in 2, and “Halant before a vowel...” in 3.4.7.4 (not “after a vowel...”).

General transcription of the language name can remain as “Kannada” throughout the document.

## Telugu

In section 5, various queries are made about inclusion/exclusion of CP from repertoire.

Where characters are missing from MSR, GP does not need to provide further evidence for exclusion – though if this was a mistake, it is the critic’s duty to provide explicit evidence why. There are still changes of repertoire as Unicode develops, e.g. in § 5.3, while OC58 and 0C59 are in pink in the MSR, 0C5A TELUGU LETTER RRRA is not even mentioned in the MSR, since it was only added in Unicode 10. Still, exclusion from the repertoire is equally clear in all three cases.

In section 6, Liang is correct to point out that “blocking” of a variant should not be confused with exclusion of a CP from the repertoire (and the GP should take note of that).

It is prima facie puzzling that U+0C44 TELUGU VOWEL SIGN VOCALIC RR as been included while U+0C60 TELUGU VOWEL LETTER VOCALIC RR has been excluded. Does this reflect the claim that no words start with an RR ? A correction, or explicit justification should be given.

## General

From the reference[[1]](#footnote-1) given under §6.1, it emerges that the actual practice in rendering short o and long o in Kannada and Telugu is much more complex and less predictable than the repertoire for these two CPs makes clear. But Liang’s point to NBGP remains obscure, unless it is to remind the NBGP that they need to indicate this difference between IDN coding and practical spelling/rendering encountered more generally in the languages, i.e. in the LGR for each of Kannada and Telugu. (It is completely absent from the Kannada case. Telugu pp. 16-17 of §6.1 points out implicitly (but not explicitly) that the variety of rendering/CP realization is excluded by involving a matra + matra sequence. But there might be more to say, if the glyphs on Table 9a were more fully glossed (with phonetic names or symbols, as well as CP numbers).

Otherwise, all points raised by Liang Hai should be accommodated as far as they seem reasonable and/or helpful to NBGP. In particular, requests for clarification probably need active attention from NBGP.

Note, however, that it is no part of NBGP’s duty to justify (or reason against) the policy of not admitting ZWJ and ZWNJ within the top-level zone. As it stands, applicants for labels will not have the option of including either zero-width character in their applied-for labels. Such advice might, however, be useful as an input to policy on other zones, so IP welcomes its inclusion.

# Tamil

In general and and as usual, most of the points made by Liang Hai are worthy of consideration. Here IP merely comments on ones that may need extra clarification.

§3.3.1’s point is perhaps less important than others, since the LGR was alluding to the unique Tamil way of referring to consonants which have a following characteristic vowel (evidently a common kind of Neo-Brahmi symbol) as “vowel consonants”, reserving the use of “consonants” to such symbols where their vowels have been “killed” by a PULLI sign. The difference between grapheme and phoneme is something else.

§5.2 (+ §3.2 and §7) This point is about which languages have been used as evidence for characters in use within the Tamil script (although it does not mention the words “languages” or “dialects”). In practice, only Tamil proper has been accepted as providing evidence, regardless of what other languages or dialects actually use it. This judgement is made in §3.2, but only by implication, not as a direct statement. The judgement is again apparently left open at §7. Still, the LGR does seem to be consistent on this.

The GP would expect this decision to be made explicit and for it to be clearly justfied based on the particular facts surrounding the use of Tamil (presence or absence of other significant languages using the script/presence or lack of differences in the use of the Tamil script).

§5.2.1, Table 6a : This item is absent from the LGR .docx, but present in the PDF.

https://www.icann.org/en/system/files/files/proposal-tamil-lgr-25sep18-en.pdf

§11 Appendix A, Table 2. Since the items in this table are are all somewhat similar but not enough to be “homoglyphs”, and so do not qualify as cross-script variables, there is not much point in adding more examples to this table. The NeoB GP might want to add a statement that cases that were “obviously distinguishable” even if somewhat confusable are not listed. As it stands, the section clearly explains that the listing covers cases that saw some discussion.

The GP is free to list additional confusables here, but this is not required.

# Malayalam

There is a comment from Thin Zar Phyo, which comments from the perspective of Myanmar (Misleadingly, it has been submitted by Pitinan with heading Comments-kannada-oriya-telugu-08aug18.)

Thin Zar Phyo points out that U+1002 and U+101D (Myanmar GA and WA) have been accepted in Malayalam LGR as cross-script variants of U+0D31 and U+0D20 (Malayalam RRA and TTHA).

Thin Zar Phyo then points out that U+1000 and U+101A (Myanmar KA and YA) are also confusable with U+0D28 and U+0D27 (Malayalam NA and DHA), but are not proposed (apparently) as cross-script variants. This is strange, since from the point of view of glyphs, they are just as similar pairwise (Myanmar to Malayalam) as the preceding set (and indeed, there are now enough homoglyphs between the scripts to propose realistic confusable labels.

Evidently more thought is needed on the possible cross-script homoglyphs between Myanmar and Malayalam.

(As long as the additional Myanmar cross-script variants do not impose in-script variants in Malayalam, it would be possible to process the latter, but with the proviso that the remaining variants would be imposed as soon as Myanmar gets integrated – this is all according to how the integration process is defined).

The argument does not seem so strong for the last pair to which Thin Zar Phyo draws attention, namely the Myanmar sequence U+1000 and U+102C (KA and VOWEL SIGN A) and the Malayalam sequence U+0D28 and U+0D4D and U+0D28 (NA and SIGN VIRAMA and NA), since these are seldom (if ever?) identical in their renderings.

(This is something that can perhaps be reviewed in the context of reviewing the Myanmar LGR. Note also, that some Myanmar code points are, as of yet, not in the MSR. This could be a reason for discrepancies, if any of these code points were involved.)

Cibu points out:

§6.1 \*In-script variants\* is proposing to disallow <chillu-n, virama, rra>. However, as per Unicode (Standard Version 11.0.0 §12.9 page 506 table 12-38) <chillu-n, virama, rra> is the prescribed sequence for the form {chillu-n base, rra below-base}. Because of this conflict with Unicode, the sequence <chillu-n, virama, rra> should not be disallowed.

The point of the LGR’s in-script variant (equating the three sequences <chillu-NA; NA and VIRAMA; chillu-NA and VIRAMA>, all before RRA) is to pre-empt the possibility that unpredictable renderings for these sequences might set up false oppositions, and so create a security issue. But the security issue could still arise, even if Unicode 11 explicitly mentions one of the sequences and its desired rendering in glyphs.

When it comes to prescriptions of preferred sequences, the IP generally prefers to reference the latest version of the Unicode Standard, even though the LGR it still technically based in Unicode 6.3.0. The reason is, that any changes made since then, unless controversial, are likely to remain on the books going forward. There’s no benefit in an LGR creating an exception.

In this particular issue, it would behove the NeoB GP to review the issue and to review the prescription made in Unicode 11.0.0. If the outcome of that review is that the local experts feel that Unicode is incorrect, then a proposal for modification should be sent to the Unicode Consortium. Otherwise, the LGR should be modified to match the Unicode Standard, or, alternatively, the GP could document why disallowing that sequence is correct in the context of top level domains.

**Issue noted by IP after the close of comments:**

In the XML the rule named “follows-0D33-immediately” should be renamed for clarity to “precedes-0D33-immediately” and the comment changed to “Section 7, WLE 7: 0D33 cannot immediately precede 0D33”.

This is because the rule is evaluated for each 0D33 looking ahead. This would follow more closely the naming conventions for look-ahead rules of this type.

# Oriya

This set of critiques is particularly difficult to reconcile, since they seem to reflect deep dissatisfaction with the sources used for section 3 by NBGP, and hence (though here the arguments are weaker) that the repertoire should change in some details.

As usual, all points raised by Liang Hai should be accommodated as far as they seem reasonable and/or helpful to NBGP. (Two other critics – O Foundation and Subhashish Panigrahi – claim to agree with him anyway.) In particular, requests for clarification probably need active attention from NBGP.

One specific concern is how to position Oriya historically between its similarities with Devanagari and Bengali. Since this issue is likely to generate more heat than light, it may be better to eliminate all (evaluative) statements on this, although the tree diagram in Figure 1 may be useful still.

Another concern would be to have a phonetic (i.e. essentially Latin) transliteration of all example phrases.

Speaking of §6.1, Liang Hai has opened a previously secret issue on two homophones for [ya/yya] and three for [ba/va/wa]. If there is consensus that spelling is fluid as among these two sets of options, there will be a case for some type of variant (cf Ethiopic). If however these just happen to be homophones, but most literate Oriya readers/writers are clear in any case which option is correct, the status quo may be retained. (Note that Liang appears to have a maximalist bias in reflecting “ambiguities” as variants that may be inappropriate. The NBGP need to take a line on all this, one buttressed by argument.)

Santosh Kumar Mohanty seems concerned about §3.4 (some aspect of the repertoire?) and §3.8 (nukta), but unfortunately he does not even hint what problems he sees. He can only be ignored, therefore, unless the NBGP can intuit what is concerning him.

The O Foundation are largely concerned about the true history of the Oriya script, details which would naturally be accommodated in §3. However, this may have implications for membership of the repertoire – although the general tenor of this critque is rather purist, and may try to exclude modern innovations (if such they be) on sentimental or historical grounds. Since the existing section 3 has largely been compacted from parts of Wikipedia, there is a case for extending the close of public comment, and letting the O Foundation offer a new version of section 3, stating explicitly what implication (if any) this has for the repertoire (and even, just possibly, WLE and variants).

There is evidently concern about the incidence of nukta (especially strong in the critique of Subhashish Panigrahi, although his remarks are particularly difficult to distil into practical proposals). The IP is concerned that the features of the script not be restricted in a way that is biased against minor languages.

Anindita Sahoo takes issue with a passage in §4.1.2.1 ii (IDNA Protocol), where the nuktated versions of eight named consonants are termed “allophones” of the same, rather than as separate phones. The use of the term “allophones” is indeed unfortunate and mistaken, but the sequences of code point + nukta (U+0B3C) are indeed composed written entities. (The explanation is a little clearer at §3.8.) No action needs taking, except perhaps to replace the term “allophone” with “approximate fricative”.

1. http://www.unicode.org/L2/L2014/14005-telugu-kannada-vs-o-oo--UTN.pdf [↑](#footnote-ref-1)