NBGP Response to IP Comment:   
Initial Draft Gujarati LGR Proposal

DATE: 2018-02-22

# Overview

This is the response of NBGP to the IP Comments received on 6th Feb. 2018 on Gujarati LGR proposal contained in the following files

* LGR-Proposal\_Gujarati\_20180115.docx
* Proposed-LGR-Gujarati-20180107.xml
* test-labels-gujarati.txt

In addition to this document, a separate changes tracked version of the Gujarati LGR was attached alongwith comments and edits suggested from the IP. Most of those comments have been incorporated in the current version of the Gujarati LGR.

# General comments

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Issue** | **IP Recommendation** | **NBGP Response** |
| **a)** | Discussion of individual characters should include the Unicode names and where practical the Unicode code point value.  As well as making look-up easier, and precise reference clearer, this would also give informal guidance on the phonetic attributes of the characters to non-native readers. | Please apply where appropriate | Carried out in most of the cases. However this will further be improved in the next version to have uniform way of representation. |
| **b)** | In general, it would be helpful to outsiders if the example words in Gujarati were also spelt out in Latin script. This is a general requirement on examples written in non-Latin scripts, where practicable. | Please apply where appropriate | Has been done wherever applicable. |
| **c)** | Since it has been adopted by Unicode, the term VIRAMA, (rather than "Halant", "vowel-killer" etc.), should be used consistently to identify any characters within an abugida script which suppress the implicit vowel after consonants that appear without any vowel affix.  HOWEVER, the IP had already accepted the use of “halant” in the context of the Devanagari proposal. There are two ways forward: the NeoBGP may choose to “standardize” on “Halant” or it may adopt the Unicode name “Virama” (and gloss it with the native name in each script).   In either case, the IP expects to see only a single, common term in use across the NeoGP’s scripts. | Please apply where appropriate | This has been clarified in footnote on Page 14.  Whether the community calls it Halant or Virama is dependent on it's shape in the particular script.  The word Halant is actually made up of Hal+Ant where Hal means Plough and Ant meant end. So it describes the shape of character like a plough. However in Tamil as an example languages it does not look like a plough so it is called as Virama which describes its functional aspect (Viram i.e. Pause)  So, giving precedence to the way community calls it, it has been ensured to have uniform way of referring to it. within an LGR document. It may not be wise to have one single term to be used across the board in all the LGR. Request IP to consider it for now.  We will however try to have a broad consensus on the same within the community and update in the next version. |

# Test labels

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Issue** | **IP Recommendation** | **NBGP Response** |
| Test Labels File | The GP supplied test labels in the requested plain text format and the IP performed verification of label status.  The number and variety of “invalid” labels is totally insufficient for meaningful testing. | The IP has verified that all valid labels are accepted by the LGR and all invalid labels are rejected.  The IP appreciates getting the test files in plain text format, which simplifies automated test protocols. | NBGP requests IP to provide some more time to provide exhaustive file with valid and invalid words. Once the Descriptive document is tentatively finalized, it will be provided. |
|  |  |  |  |
| Annotation file | The annotation file appears to contain the same data as the test file. It is not in a format that the IP can use with existing tools | In the future, the format as found in the Test-Labels file is preferred. |  |
|  |  |  |  |
| Test Coverage | The IP collected data on test coverage.  The test files do not cover all elements used in the rules and are far from covering all possible combinations of character classes.  The file covers 4 out of 5 characters classes, 6 out of 7 tags for code points. Only about 50-60% of each context rules were matched (valid labels) or failed to match (invalid labels).  The test files did not contain an example of a nukta.  It is not necessary or desirable for the test files to cover all combinations of code points, but it is generally good to have at least one instance of each code point in some test label. | A file for verifying variant resolution should be added.  The fact that the test file did not flag the incorrect rule for Nukta (below) is a clear indicator of how insufficient it is. |  |
| Labels with Nukta | The IP created a revised LGR for testing that correctly implements the intent of the Nukta rule (correct code points for special consonants)  With this revised LGR, the number of invalid labels in a 100,000 corpus fell from about 40 to about 10.  Note that this means that the incidence of Nukta in a corpus can be estimated at 0.03% of distinct labels. However, that while this data is not conclusive for the expected incidence among TLD labels it is also not very high.  Because the treatment of Nukta is the same as in other NeoB scripts, the total complexity added to the Root Zone LGR by adding a rule that is so rarely used is less than it would be if this was the only script. | Please provide test cases that test the corrected rule for Nukta (both pass and fail). |  |
|  |  |  |  |
| Language coverage | If possible, the valid label file should provide examples of labels in all the languages considered (if that number is small, otherwise a cross section of languages with different requirements).  By doing this, the test label file serves a secondary purpose of documenting to the reviewers that the LGR indeed covers the languages in question. | This does not affect the use of the label files for regression testing, but it would be useful in demonstrating that the LGR caters to typical akshara patterns in ALL of these languages. |  |

# Comments on main document (.docx)

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Issue** | **IP Recommendation** | **NBGP Response** |
| Overall | Missing Page Numbers, wrong script in Header | Please add pagination so it is easier to review the document, and fix the header. | Added. |
|  |  |  |  |
| §1 | Any filenames cited must match the file names/dates of accompanying files.  This section should list all files that are part of the proposal. (The IP understands that in this instance not all files received originated with the same people, but as the proposals get more refined, this information should be in sync). | Please provide in future versions | Will be ensured. |
|  |  |  |  |
| §3 | The term “varga” is proving difficult to non-native, non-linguist, technical reviewers. A formal definition may help. | Provide formal definition for certain terms, e.g. “varga” | It has been elaborated. |
|  |  |  |  |
| §5.1 | The table was a bit wide for the page, we suggest some different font sizes | Please review |  |
|  |  |  |  |
| §5.5.4.3 | The rather complex 'rules' expressed in 5.5.4.3 (limit of 4 consonants) and then followed in 5.5.4.3 A and 5.5.4.3.B don't are not fully enforced in the WLE rules in Section 7 (or at least only partially in rules 3 and 4).  It would make for a rather complex WLE, so it may not be desired. However, in that case the text in section 5.5.4.3 should say something about not fully implementing these well-formed rules, at least as far as the maximal count of consonants is concerned. | Note the difference between the full linguistic Akshara formation rules and the practical subset of the restrictions enforced in the WLE rules in LGR | Difference is noted. |
|  |  |  |  |
| §7.1 Nukta | The list of consonants to be used with Nukta is incorrect. The code points are those for a different script.  This list could be given a symbol as well. | Please fix | Fixed. |
|  |  |  |  |
| §7 | The WLE rules do not implement any restrictions on the number of consonants in a conjunct. That is probably for the best, but this and any other deviations from the structure rules in Section 5 need to be discussed here.  The last rule, 7.6,  V: Can NOT be preceded by H  does not follow from any of the discussion in Section 5. An explanation for this rule is needed. | Please revise  Please add explanation in Section 7 or section 5. | Revised. Relevant footnote added.  Added. |
|  |  |  |  |
| §9 and §10 References and bibliographies | Several references (e.g. [NBGP] do not provide any URLs or other means of locating information)  Item 18 should be transliterated for non-native readers. (and or the title translated)  The “webography” references onlyprovide a URL, without a title, nor do they indicate the date accessed.  Many references appear not to have been cited in the document, at least searching for [2] (etc.) did not locate any instances.  Finally, it is confusing to have all three sets of references (§9, §10.1 and §10.2) numbered the same way. At least the references that are actually cited, e.g. [NGBP] should **not** have numbers. | Please fix and complete | URL to the [NBGP] can be provided once ICANN confirms permanent link to the NBGP Webpage. For now, link of the current homepage of NBGP has been provided.  Item 18 updated.  Webography references updated.  References which have not been cited have not been given numbers and numbering has been fixed now. |
|  |  |  |  |
| **Appendix A In-script variants** | Some discussion should be added to Appendix A, summarizing the conclusion arrived in Section 6. | Please add. | Added. |
|  |  |  |  |
| **Other Editorial issues** | |  |  |
|  | A marked-up version of the proposal file is attached to this response. It contains a large number of individual editorial suggestions. | Please review and apply the editorial suggestions provided. | Done. |
| **Common Text Across NeoB LGR proposals** | Some editorial comments apply to text that is a common “boilerplate” across all NeoBrahmi proposals. It would be good if the same issues could be fixed in parallel in all of them, even if the IP notices them only during the review of one of the LGRs | Fix boilerplate across all NeoB LGR documents. | It has been fixed for GJ document. |

# Comments on LGR Specification (XML)

## <meta> Element in XML

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Issue** | **IP Recommendation** | |
| Meta data | No issues with the meta data.  For <description> see separate section. | |  |

## <description> Element in XML:

|  |  |  |
| --- | --- | --- |
|  | |  |
| Repertoire | The description states: "NBGP, in Section 5 "Repertoire" proposes 65 unique code-points to be made part of the Gujarati LGR [Proposal]."  This wording is weird, because "Section 5" is not a unique location, but specific to the proposal. Also, we don't otherwise use language like "The GP proposes..." in these summaries, because then they have to be rewritten for the actual LGR.  "According to Section 5, "Repertoire" in [Proposal] the Gujarati LGR contains 65 unique code points." | Change wording |
|  |  |  |
| Character Classes | The <description> element in the LGR should be reviewed with an eye towards streamlining the text slightly. (It is a bit wordy but not as much as the Gurmukhi proposal)  The priority should be on information that is needed by an implementer to understand the general purpose and function of the various elements in the LGR (from repertoire to rules), but it is appropriate to refer to the proposal document for any deep background.  An additional class might be added for the specific consonants use with Nukta. See elsewhere in this document. | In particular in the section on “character classes” the information on pronunciation is most likely “too much” for the XML file – however, the summary of which code point contexts a class is used with would seem to be helpful.  Please consider some redrafting. |
|  |  |  |
| Whole Label Rules | In the definition of symbols, the file uses “<br\>” which is invalid HTML syntax, the correct syntax would have been <br />  The rule for Nukta uses the wrong set of code points. As a result, Nukta would not have been allowed in any label. This error would have been avoided, had the list of code point been implemented via tagging members of the repertoire, as opposed to an explicit list of code points in the rules section. | Please fix the syntax  Please ensure that all three places (Section 7, <description> and the “comment” attributes on the<rule> element use the same numbering.  For the next draft, please implement all sets by tagging code points in the repertoire instead of listing explicit code values in the rules. |
|  |  |  |
| References Section | The “References” section in the <description> defines some references that are not used elsewhere in the LGR: [NBGP] or [gTLD]  A reference to MSR-1 will not be needed  [Proposal] place holder is missing | Remove any unused references  remove entry for MSR-1 and add a URL to MSR-2  Add a placeholder for proposal |

## <references> Element

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Recommendation** |
| Unused references | The following references are unused:  **(none)** |  |
|  |  |  |
| References without Title | (none) |  |
|  |  |  |

## <data> element

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | |
| **Item** | **Issue** | | **IP Recommendation** |
| Tags | See below | | Please update the tag values as suggested below |
|  |  | |  |
| Comments | (no issues: the comments match the Unicode names identically) | |  |
|  |  | |  |
| Variants | (none) | |  |
|  |  | |  |

## <rules element> in XML:

|  |  |  |
| --- | --- | --- |
| **Item** | **Issue** | **IP Recommendation** |
| Classes | Define a tag-based class for the “special” consonants that can be followed by a Nukta. | Define named, tag-based class, for example named “CN” |
|  |  |  |
| Factoring WLE rules: Nukta | Rule WLE1 enforces the restriction that nukta may only follow specific consonants. Therefore, in all other rules, the construct ([:**consonant**:]|([:**consonant**:][:**nukta**:]))  can be replaced by  ([:**consonant**:]|[:**nukta**:])  This reduces the depth of the left context | Please update the rule (or the one rule, after conflating the Visarga and Matra rule). |
|  |  |  |
| Incomplete Rule | The rule for Nukta is not yet fully specified. It currently contains an empty context, matching ANY label.  In regular expression notation the context should be [:CN:] where “CN” is a class specified by a CN tag added to any of the specific consonants (see below) | Please fix |
|  |  |  |
| Specific characters | The rule for Nukta contains a list of specific characters. It might be useful to consistently across the NeoBGP LGRs touse tags like C. in the list of repertoire and then define a class for these special consonants based on this tag. (This follows the convention established for the SEA scripts in LGR-2).  Please note that each “char” may have multiple tags; they are space separated in the value of the “tag” attribute, e.g. tag=”c1 consonant”.  If multiple tag values are used per “tag” attribute, make sure that they are ordered alphabetically. | Tag code points belonging to lists of “special” consonants, vowels etc. with additional tags and then create named classes and use them in the rules. |
|  |  |  |
| Rule naming | Rule names include the “target” of the rule. This leads to duplicate rules (see below). Change “Visarga-follows-either-V-C-N-or-M” to ”Follows-only-V-C-N-or-M”. | Please fix. |
|  |  |  |
| Duplicate Rules | The following rules are duplicates: **[Anusvara-follows-either-V-C-N-or-M, Visarga-follows-either-V-C-N-or-M], [Matra-follows-only-C-or-CN, Halant-follows-only-C-or-CN]**. | Update the names of the rules to eliminate the “target” and fold identical rules.  It is OK to keep these rules separate in the DOCX file, and give them separate numbers. Simply reference **both** numbers in the comment for the combined rule. |