

Analysis of comments for Malayalam LGR Proposal

Revision: June 30, 2019

Neo-Brahmi Generation Panel (NBGP) published the Malayalam script LGR Proposal for the Root Zone for [public comment](#) on 25 September 2018. This document is an additional document of the public comment [report](#), collecting all comments and NBGP analyses as well as the concluded responses. There are 7 (severn) comment analyses as follow:

No.	1	From	Thin Zar Phyo, Myanmar GP Chair																								
Subject	Finalized Information for NBGP LGRs																										
Comment	<p>Dear NBGP members,</p> <p>Myanmar GP would like to congratulate on the complete work of Malayalam LGR proposal.</p> <p>We are currently developing the Myanmar Script LGR proposal. In the Malayalam and Myanmar cross-script variant analysis, Myanmar GP defines the following code points as variants</p> <p>Variant code points:</p> <table border="1"><thead><tr><th>No.</th><th>Glyph</th><th>Code Point</th><th>Myanmar Character Name</th><th>Glyph</th><th>Code Point</th><th>Malayalam Character Name</th></tr></thead><tbody><tr><td>1</td><td>◌</td><td>U+1002</td><td>MYANMAR LETTER GA</td><td>◌</td><td>U+0D31</td><td>MALAYALAM LETTER RRA</td></tr><tr><td>2</td><td>◌</td><td>U+101D</td><td>MYANMAR LETTER WA</td><td>◌</td><td>U+0D20</td><td>MALAYALAM LETTER TTHA</td></tr></tbody></table> <p>We'd like to draw your attention to these sets. They might need to be included in Malayalam variant rules.</p> <p>In addition, Myanmar GP also lists the confusable code points (not variants) as the appendix as follow:</p>						No.	Glyph	Code Point	Myanmar Character Name	Glyph	Code Point	Malayalam Character Name	1	◌	U+1002	MYANMAR LETTER GA	◌	U+0D31	MALAYALAM LETTER RRA	2	◌	U+101D	MYANMAR LETTER WA	◌	U+0D20	MALAYALAM LETTER TTHA
No.	Glyph	Code Point	Myanmar Character Name	Glyph	Code Point	Malayalam Character Name																					
1	◌	U+1002	MYANMAR LETTER GA	◌	U+0D31	MALAYALAM LETTER RRA																					
2	◌	U+101D	MYANMAR LETTER WA	◌	U+0D20	MALAYALAM LETTER TTHA																					

		Confusable code points:				
No.	Glyph	Code Point	Myanmar Character Name	Glyph	Code Point	Malayalam Character Name
1	က	U+1000	MYANMAR LETTER KA	ന	U+0D28	MALAYALAM LETTER NA
2	ഓ	U+101A	MYANMAR LETTER YA	ഘ	U+0D27	MALAYALAM LETTER DHA
3	ꣳ	U+1000, U+102C	Sequence: MYANMAR LETTER KA , MYANMAR VOWEL SIGN AA	ꣳꣳ	U+0D28, U+0D4D, U+0D28	Sequence: MALAYALAM LETTER NA, MALAYALAM SIGN VIRAMA, MALAYALAM LETTER NA
		<p>This is for your information, and if the Appendix of both Malayalam and Myanmar LGR have the same list, it could be useful for the user of these proposals.</p> <p>We'd like to thank you for your good work. And we hope to have further collaboration with the NBGP regarding the LGR proposals.</p>				
NBGP Analysis	Agree.					
NBGP Response	Update the proposal as per the comment.					

No.	2	From	Cibu
Subject	Feedback on §6.1 In-script variants Set #1: sequence <chillu-n, virama, rra>		
Comment	<p>§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.</p>		
NBGP Analysis	Agree.		

NBGP Response	Update the proposal as per the comment.
---------------	---

No.	3	From	Santhosh Thottingal
Subject	Malayalam LGR feedback		
Comment	<p>1. The case of ൠ is similar to ൡ in the document. A font that does not stack the ൠ + ൡ + ൠ can render it in horizontal format. So a word like മീ൬൬ can be spoofed by applying virama to the last two ൠ. It is rare to see a font that does not stack ൠ, but instead of depending on that weak assumption, better add a WLE rule similar to ൡ. (This is also mentioned in Appendix part of the document as community feedback- That feedback was from me.)</p> <p>2. The document should NOT be conflicting with nta as explained in Unicode version 11, Chapter 12. In the table under 6.1, 1a, 1b, 1c - all three should be allowed as variants and should not block any of them. Currently 1b is blocked. 1b is mentioned in Unicode Chapter 12. So it should be allowed. There is a debate going on whether that is a correct sequence or not in Unicode, it may or may not be corrected. As it exists in Chapter 12, it should be treated as variant.</p>		
NBGP Analysis	Agree.		
NBGP Response	Update the proposal as per the comment.		

No.	4	From	Gowtham Raghunathan
Subject	(None)		
Comment	<p>best point to argue for this initiative is...</p> <p>English is a language which has unique letters and unique pronunciation other languages has identical letters and identical pronunciation.</p> <p>it will make an easy way for cyber theft.</p> <p>hope you won't go with this decision falls under serious category</p>		

NBGP Analysis	NBGP acknowledges the comment.
NBGP Response	No action required.

No.	5	From	SELVARAJ R
Subject	(None)		
Comment	<p>മലയാളഭാഷയിലെ അനുസ്വാരം "ം" എന്ന സ്വരം ഇംഗ്ലീഷിലെ "o" സാമ്യമുള്ളത് ഏത് തരത്തിലുള്ള പ്രശ്നങ്ങൾക്കാണ് കാരണമാകുന്നത്? വിശദീകരിക്കാമോ?</p> <p>[Translation by Veena Solomon: Can you explain what are the problems caused by the similarity of Malayalam Anuswara "ം" with English o?]</p>		
NBGP Analysis	The LATIN SMALL LETTER O and LATIN SMALL LETTER S was discussed and concluded that they are out of scope of NBGP.		
NBGP Response	No action required.		

No.	6	From	Ajay . (ajaykerala at hotmail.com)
Subject	very good initiative		
Comment	This is very useful for a lot of people in the state of Kerala.		
NBGP Analysis	NBGP acknowledges the comment.		
NBGP Response	No action required.		

No.	7	From	Liang Hai
-----	---	------	-----------

Subject	A quick review of the Malayalam proposal
Comment	<p>- §3, §3.1–§3.3: Unclear what the point is for having such a lengthy and detailed introduction of the script’s history. Move it to an appendix, or just remove it.</p> <p>- §3.5, “Sanskrit, although it falls under EGIDS 4, is not considered in Malayalam script LGR because Malayalam is rarely used to write Sanskrit.”: The Sanskrit language’s Malayalam writing system should have its own EGIDS rating for such an evaluation.</p> <p>- §3.6, “ICANN’s Maximal Starting Repertoire (MSR) for IDN LGR is based on these exclusion rules for ZWJ and ZWNJ.”: Based on what exactly rules?</p> <p>- §3.6, “But there are no identified cases where a missing ZWNJ forms another valid word with different meaning.”: What’s this discussion of attested “another valid word with different meaning” meant to reflect? A wrong spelling is simply another word, and whether this “another word” means something is a vocabulary problem, which is not really relevant here.</p> <p>- §3.6, “Missing ZWJ means, the word is a different word with different meaning. This is very rare — ...”: This pair is not relevant because the first word uses a ZWJ only because of its chillu, while chillus have atomic encodings.</p> <p>- §3.6, “Missing ZWJ never means a spelling mistake, but just a writing style.”: It’s plausible to try to distinguish “a spelling mistake” and “a writing style” (which can be better put as “a spelling style” though, given what the example implies). However—</p> <p>* The example is not relevant because it uses ZWJ for a chillu.</p> <p>* Basically this whole section of ZWJ and ZWNJ requirement probably needs to be preceded by the section that discusses about some ZWJ-using structures that can also be safely encoded without ZWJ, so this group of ZWJ use cases can be excluded.</p>

	<p>* Also, it's actually unclear why "Missing ZWJ never means a spelling mistake", as ZWJ is specified by the Unicode Standard (see Table 12-36, Use of Joinders in Malayalam, in the Core Specification 11.0) to have the ability of requesting a consonant stack, which is discussed in the first case as a matter of spelling mistakes.</p> <p>* The differentiation between a spelling mistake and a writing/spelling style also largely depends on the exact orthography being followed.</p> <p>* Note the whole situation of when a ZWJ is required and when a ZWNJ is required is highly dependent on fonts. The first case only requires a ZWNJ because Windows's default Malayalam font, Nirmala UI, as an inappropriately produced font, forms a lot of undesired consonant stacks despite being largely a reformed-orthography font. It strikes me as an apparent necessity that, for such an LGR analysis, a survey of commonly used Malayalam fonts should first be carried out.</p>
NBGP Analysis	The NBGP believe the text is reasonable introduction and not too lengthy. Although IDNA2008 allows the use of ZWJ and ZWNJ in domain names, they are not allowed in the root zone labels, due to exclusion from MSR. Also, there was a demonstrative example of the issue in the proposal.
NBGP Response	No action required.
Comment	§3.7, Script and Orthography: Unclear why the consonant letter ᵛᵛ is missing.
NBGP Analysis	Agree.
NBGP Response	Updated the proposal as suggested.
Comment	<p>- §3.7, Anusvaram and Visarga: "... and hence is traditionally treated as a kind of vowel sign.": There's no causality here. Signs like anusvara and visarga are traditionally categorized together with vowel signs (and the category is not necessarily comparable to the modern concept of vowel) because they all are dependent signs that modify a base letter.</p> <p>- §3.7, Chillu letters (Chillaksharam) and Samvruthokarams, "Chillaksharam is an original feature of Malayalam used only with 6</p>

	<p>consonants at present.”: A broader discussion of other rare chillus (in addition to chillu K) should be discussed.</p> <p>- §3.7, Chillu letters (Chillaksharam) and Samvruthokarams, “Any consonant can be followed by consonant ... The chandrakkala alone at the end of a word is treated as Samvruthokaram.”: The paragraph is filled with conflicting statements. Making a clear distinction between the actual written structures and the intended phonetic sequence is important.</p> <p>- §3.7, Chillu letters (Chillaksharam) and Samvruthokarams, “Chandrakkala coming within a word (followed by other character(s) of the word) denotes a conjunct letter formed by the character(s) preceding and following the chandrakkala.”: Unclear if this is talking about written structures (then a visible chandrakkala sign has nothing to do with a written conjunct) or the general conjunct encoding (then the conjunct is not a letter but a sequence of consonant characters and chandrakkalas that can probably be rendered as a visual structure of conjunct).</p> <p>- §3.7, Chillu letters (Chillaksharam) and Samvruthokarams, “Examples of Samvruthokaram:”: The document should use a specific orthography by default and explicitly call out when a non-default orthography is discussed for some reason. Here the examples are in the traditional orthography but all the preceding content in the document is basically in the reformed orthography (eg, the “Vowel diacritics” section), and there’s no any note about this inconsistency.</p> <p>- §3.7, Chillu letters (Chillaksharam) and Samvruthokarams, “For the words that end in chillu, Samvruthokaram is used to make the pronunciation clearer. ...”: Unclear how such a phonetic discussion (as well as the following four cases of “phonological transformations”) is relevant to written structures and encoding. Also unclear why only the orthography that uses an explicit vowel sign u is presented in the examples.</p> <p>- §3.7, A selection of conjunct consonants, Table 5: Adjust column widths to avoid line breaks, which make the NFL row confusing</p>
<p>NBGP Analysis</p>	<p>The text in section 3.7 explains how Samvruthokaram is written in south and north Kerala. It also explains what happens to a chandrakkala in the middle of a word. However for the unclear text the NBGP revise some text for more clarification</p>

<p>NBGP Response</p>	<p>Revise section 3.7 as per some of the comments.</p>
<p>Comment</p>	<p>- §5.3: See the comment below for §6.1.</p> <p>- §6.1, set 1: The analysis is a mess.</p> <p>* Note the case 1a is a non-standard de facto encoding for the written structure <chillu n base, below-base rra sign>. NBGP need to work with the Unicode Consortium and make sure they give consistent recommendations on this problematic issue.</p> <p>* Also, as 1a is rendered as a wrong structure in Windows’ default Malayalam font Nirmala UI, it’s unclear why this encoding is not disallowed because of “rendering problem” (which makes 1b disallowed).</p> <p>* About 1b, note the only working sequence (so the only intended sequence) for Nirmala UI on Windows is <NA, VIRAMA, ZWJ, RRA>, although the standard <CHILLU N, VIRAMA, RRA> is somehow also implemented in the font (therefore can be rendered by it with a shaping engine that supports the sequence, while Windows’ shaping engine does support the sequence).</p> <p>* Then it’s unclear why “it is safe to disallow” <CHILLU N, VIRAMA, RRA> while allowing <NA, VIRAMA, RRA> when both sequences have rendering problems and only the former one is recommended by the Unicode Standard’s Core Specification.</p> <p>* As ordinary fonts shouldn’t render a character sequence intended for <chillu n base, below-base rra sign> as <chillu n base, rra nase>, therefore there isn’t visual confusability despite spelling and phonetic relationship, it’s unclear why this variant is blocked. Are other spelling alternatives to be blocked too?</p> <p>* As ordinary fonts shouldn’t render a character sequence intended for <chillu n base, below-base rra sign> as <chillu n base, rra nase>, therefore there isn’t visual confusability despite spelling and phonetic relationship, it’s unclear why this variant is blocked. Are other spelling alternatives to be blocked too?</p>

NBGP Analysis	Agree.
NBGP Response	The modification has been made to the proposal.
Comment	§6.2.1, Table 10: A bad rendering of the Tamil glyph in the set 6
NBGP Analysis	Agree.
NBGP Response	Updated the proposal as per the comment.
Comment	<p>§7.1.2: Note the document basically suggests such a pattern: `C[M][B X] V[B X] C[U+0D41]H L`</p> <p>* Rule 5 and 6 should be safe but it's really unsettling to restrict something not because of written limitations but phonology and spelling conventions.</p> <p>* Rule 7 doesn't seem to be consistent with the restrictions suggested in §6.1 (which disallows ஐஐ... and allows ஐ). See the comment below for Appendix C.</p> <p>* Note that the Unicode Standard's Core Specification suggests (see Table 12-33, page 504, in the referred Core Spec 10.0) a samvruthokaram not only appears at the end of a word, but can also appear as an independent vowel letter (typically a word-initial structure) or be followed by a anusvaram. The inconsistency between the Core Spec's claim and this document's analysis must be addressed, and the WLE rules might need to be loosened up. Note this is a typical case exhibiting how dangerous it is to set up a restrictive pattern not simply based on written structures but the limited known spelling conventions and phonological theories.</p> <p>- §10, Appendix A, Table A-1: The last column seems to be meant to reflect confusable renderings, then the renderings of sequences and atomic characters can be simply merged if the authors don't have a word processor that allows the sequences to be rendered with correct reordering and without dotted circles.</p> <p>- §10, Appendix A, "Although, Unicode defines this canonical decomposition, the Standard recommends not to use the sequence": The Unicode Standard doesn't recommend "not to use the sequence[s]".</p>

	<p>- §12, Appendix C:</p> <p>* I agree the <u>ഒ</u> vs <u>ഓ</u> pair is indeed worth discussing, since this pair is probably the single most confusable pair in the reformed orthography (while the traditional orthography naturally relies on a greater number of details) because of the structural disadvantage of the letter <u>ഓ</u>, and other comparably confusable pairs (<u>മ</u> vs <u>മ</u>, <u>ത</u> vs <u>ത</u>, <u>ക</u> vs <u>ക</u>, etc) are indeed significantly less confusable.</p> <p>* However, if the NBGP plans to make restrictions for such an issue, a thorough and accurate research must be first finished. I don't think either the NBGP or the IP's current researches and considerations are enough.</p> <p>* From what is presented in the document, it seems both the NBGP and the IP have been analyzing only words but not what combinations can occur when inter-word spaces are removed from a sequence of words. However the latter should be a key topic for the discussion, and apparently it can introduce many more sequences that are previously considered highly limited, eg, a much larger number of <u>ഓ</u>.</p> <p>* Also, it's not appropriate if the authors have been only analyzing the character sequence but not the final glyph sequence (which includes reordered glyphs, such as pre-base vowel signs, which can break an otherwise confusable sequence, eg, <u>ഓ</u> + <u>ഓ</u> → <u>ഓഓ</u>).</p> <p>* "The consonant <u>ഓ</u> (0D33) rarely follows another <u>ഓ</u> in Malayalam, except in the case of some place names.": It's unclear why the NBGP considers attested place names and phrase contractions that contain <u>ഓ</u> can be disallowed. The "Feedback from the community" section makes a pretty clear case to me that the NBGP is again over-restricting a script/language based on limited knowledge and prescriptivist grammar.</p>
NBGP Analysis	The NBGP acknowledges and appreciate the comments. The restriction of <u>ഓ</u> vs <u>ഓ</u> has been revised later on based on by the IP feedback.
NBGP Response	The rules of <u>ഓ</u> vs <u>ഓ</u> has been revised.