Proposal for a Malayalam Script Root Zone Label Generation Ruleset (LGR)

LGR Version: 3.0
Date: 2018-08-29
Document version: 1.6
Authors: Neo-Brahmi Generation Panel [NBGP]

1. General Information

The purpose of this document is to give an overview of the proposed Malayalam LGR in the XML format and the rationale behind the design decisions taken. It includes a discussion of relevant features of the script, the communities or languages using it, the process and methodology used, the repertoire of code points included, variant code point(s), whole label evaluation rules and information on the contributors. The formal specification of the LGR can be found in the accompanying XML document: proposed-lgr-mlym-20180829.xml. Labels for testing can be found in the accompanying text document: malayalam-test-labels-20180829.txt

2. Script for Which the LGR Is Proposed

ISO 15924 Code: Mlym
ISO 15924 Key N°: 347
ISO 15924 English Name: Malayalam
Latin transliteration of native script name: malayāḷaṁ
Native name of the script: മലയാളം

Maximal Starting Repertoire (MSR) version: MSR-3

3. Background on Script and Principal Languages Using It

Malayalam is a Dravidian language with about 38 million speakers spoken mainly in the south west of India, particularly in Kerala, the Lakshadweep Islands and neighbouring states, and also in Bahrain, Fiji, Israel, Malaysia, Qatar, Singapore, UAE and the UK.

Malayalam was first written with the Vatteluttu alphabet (വട്ടെഴുത്ത് Vaṭṭeḻuttŭ), which means 'round writing' and developed from the Brahmi script. The oldest known written text in Malayalam is known as the Vazhappalli or Vazhappally inscription, is in the Vatteluttu alphabet and dates from about 830 AD.

A version of the Grantha alphabet originally used in the Chola kingdom was brought to the southwest of India in the 8th or 9th century and was adapted to write the Malayalam and Tulu languages. By the early 13th century it is thought that a systemised Malayalam alphabet had

emerged. Some changes were made to the alphabet over the following centuries, and by the middle of the 19th century the Malayalam alphabet had attained its current form.

As a result of the difficulties of printing Malayalam, a simplified or reformed version of the script was introduced during the 1970s and 1980s. The main change involved writing consonants and diacritics separately rather than as complex characters. These changes are not applied consistently so the modern script is often a mixture of traditional and simplified letters.

The script has the following notable features:

- Malayalam script is written left to right in horizontal lines using a syllabic alphabet in which all consonants have an inherent vowel. Diacritics, which can appear above, below, before or after a consonant, are used to change the inherent vowel.
- When they appear at the beginning of a syllable, vowels are written as independent letters.
- Chillaksharam is another feature of Malayalam. A chillu is a pure consonant without the use of a virama, which kills the inherent vowel of a consonant.
- When certain consonants occur together, special conjunct symbols are used which combine the essential parts of each letter.

3.1 The Evolution of Malayalam Script

Malayalam was first written in the Vatteluttu alphabet, an ancient script of Tamil. However, the modern Malayalam script evolved from the Grantha alphabet, which was originally used to write Sanskrit. Both Vatteluttu and Grantha evolved from the Brahmi script, but independently.

3.2 Vatteluttu alphabet

Vatteluttu (Malayalam: വട്ടെഴുത്ത്, Vaṭṭeḻuttŭ, “round writing”) is a script that had evolved from Tamil-Brahmi and was once used extensively in the southern part of present-day Tamil Nadu and in Kerala.

Malayalam was first written in Vatteluttu. The Vazhappally inscription issued by Rajashekhara Varman is the earliest example, dating from about 830 CE. In the Tamil country, the modern Tamil script had supplanted Vatteluttu by the 15th century, but in the Malabar region, Vatteluttu remained in general use up to the 17th century, or the 18th century. A variant form of this script, Kolezhuthu, was used until about the 19th century mainly in the Kochi area and in the Malabar area. Another variant form, Malayanma, was used in the south of Thiruvananthapuram.

3.3 Grantha, Tigalari and Malayalam scripts

According to Arthur Coke Burnell, one form of the Grantha alphabet, originally used in the Chola dynasty, was imported into the southwest coast of India in the 8th or 9th century, which was then modified in course of time in this secluded area, where

communication with the east coast was very limited. It later evolved into Tigalari-Malayalam script was used by the Malayali, Havyaka Brahmins and Tulu Brahmin people, but was originally only applied to write Sanskrit. This script split into two scripts: Tigalari and Malayalam. While Malayalam script was extended and modified to write vernacular Malayalam language, the Tigalari was written for Sanskrit only. In Malabar, this writing system was termed Arya-eluttu (ആര്യ ഏഴുത്ത്, Ārya eḻuttŭ), meaning "Arya writing" (Sanskrit is Indo-Aryan language while Malayalam is a Dravidian language).

Vatteluttu was in general use, but was not suitable for literature where many Sanskrit words were used. Like Tamil-Brahmi, it was originally used to write Tamil, and as such, did not have letters for voiced or aspirated consonants used in Sanskrit but not used in Tamil. For this reason, Vatteluttu and the Grantha alphabet were sometimes mixed, as in the Manipravalam. One of the oldest examples of the Manipravalam literature, Vaishikatantram (വൈശികതന്ത്രം, Vaiśikatantram), dates back to the 12th century, where the earliest form of the Malayalam script was used, which seems to have been systematized to some extent by the first half of the 13th century.

Thunchaththu Ezhuthachan, a poet from around the 17th century, used Arya-eluttu to write his Malayalam poems based on Classical Sanskrit literature. For a few letters missing in Arya-eluttu (ḷa, l̲a, r̲a), he used Vatteluttu. His works became unprecedentedly popular to the point that the Malayali people eventually started to call him the father of the Malayalam language, which also popularized Arya-eluttu as a script to write Malayalam. However, Grantha did not have distinctions between e and ē, and between o and ō, as it was as an alphabet to write a Sanskrit language. The Malayalam script as it is today was modified in the middle of the 19th century when Hermann Gundert invented the new vowel signs to distinguish them.

By the 19th century, old scripts like Kolezhuthu had been supplanted by Arya-eluttu – that is the current Malayalam script. Nowadays, it is widely used in the press of the Malayali population in Kerala.

Malayalam and Tigalari are sister scripts descended from the Grantha alphabet. Both share similar glyphic and orthographic characteristics.

3.4 Orthography reform

In 1971, the Government of Kerala reformed the orthography of Malayalam by a government order to the education department. The objective was to reduce the labour in the process of print and typewriting technology of that time, by reducing the number of glyphs required. In 1967, the government appointed a committee headed by Sooranad Kunjan Pillai, who was the editor of the Malayalam Lexicon project. It reduced number of glyphs required for Malayalam printing from around 1000 to around 250. Above committee's recommendations were further modified by another committee in 1969 [105].

None of the major newspapers implemented it completely. But every newspaper took its own subset from the proposal. The reformed script came into effect on 15 April 1971 (the Kerala New Year), by a government order released on 23 March 1971.

3.5 Languages using the Malayalam script

The script is also used to write several languages such as Paniya, Betta Kurumba, and Ravula (all at EGIDS 5). The Malayalam language itself was historically written in several different scripts.

NBGP considered languages with EGIDS scale 1 to 4 for inclusion. Malayalam is one of the two languages written in Malayalam script (viz Malayalam & Sanskrit) meeting this criterion. Malayalam is placed among the 22 scheduled languages of India. Sanskrit, although falls under EGIDS 4 is not considered in Malayalam script LGR because Malayalam is rarely used to write Sanskrit.

3.6 ZWJ/ZWNJ

Apart from the existing Unicode character codepoints in Malayalam [110], Zero Width Joiner (ZWJ, U+200D) and Zero Width Non-Joiner (ZWNJ, U+200C) are widely used to control how ligatures are formed. Being invisible characters, they are often removed while doing normalization, particularly before doing a string comparison, or collation. ICANN's Maximal Starting Repertoire (MSR) for IDN LGR is based on these exclusion rules for ZWJ and ZWNJ. [101]

Impact of excluding them from domain name system: 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.

Hence it is not possible to register Malayalam domain names with words that contain zwj/zwnj.

There are three cases:

- Missing **ZWNJ** is considered as a spelling mistake. Example: Tamil Nadu (***tamil nadu***) is written as:

  തമിഴ്നാട് [0D24 0D2E 0D3F 0D34 0D4D **200C** 0D28 0D3E 0D1F 0D4D] (correct),
  തമിഴ്നാട് [0D24 0D2E 0D3F 0D34 0D4D 0D28 0D3E 0D1F 0D4D] (incorrect).
  But there are no identified cases where a missing ZWNJ form another valid word with different meaning.
- Missing **ZWJ** means, the word is a different word with different meaning. This is very rare – വൻയവനിക,***avanika*** (meaning: large curtain) വന്യവനിക ***vanyaVanika*** (meaning: wild garden) pair is often cited an example for this. But many people argue this is not a valid case. [102] [103]
- Missing ZWJ never means a spelling mistake, but just a writing style. There are many examples for this. നൻമ - നന്മ (meaning: goodness) is one obvious one.

Historically, ZWJ was used to render chillu in certain fonts but later Unicode included chillu characters as standalone code points and MSR-3 also includes these standalone chillu characters.

Pre-Unicode 5.0, Chillu letters were encoded as a sequence using Joiners. The older encoding is still prevalent in data, such as corpora and may even be in current use.

But this legacy representation of Chillu using Virama and ZWJ is ruled out because the root does not allow joiners, so there is no issue with the duplicate encoding of Chillu. Hence, it is to be noted that although atomic encoding of Chillu letters is not universally used, Root Zone only allows the atomic encoding.

Visual	Legacy Representation (5.0)	Preferred Representation
ൺ	NNA, VIRAMA, ZWJ 0D23, 0D4D, 200D	0D7A MALAYALAM LETTER CHILLU NN
ൻ	NA, VIRAMA, ZWJ 0D28, 0D4D, 200D	0D7B MALAYALAM LETTER CHILLU N
ർ	RA, VIRAMA, ZWJ 0D30, 0D4D, 200D	0D7C MALAYALAM LETTER CHILLU RR
ൽ	LA, VIRAMA, ZWJ 0D32, 0D4D, 200D	0D7D MALAYALAM LETTER CHILLU L
ൾ	LLA, VIRAMA, ZWJ 0D33, 0D4D, 200D	0D7E MALAYALAM LETTER CHILLU LL
ൿ	*undefined*	0D7F MALAYALAM LETTER CHILLU K

Figure 1: Atomic Encoding Malayalam Chillus [107]

ZWNJ, is used to prevent the formation of conjunct ligatures and it is required to avoid spelling mistakes and unnecessary conjuncts. For example, in a 2-word label, the first word ending in virama can form conjunct with the second word starting in a consonant. This causes a spelling mistake.

3.7 The Structure of Malayalam Script

The Malayalam Aksharam or grapheme cluster is based on the Malayalam phonological system, with the following basic phonological template.

Phonology

Vowels: Malayalam has five short and five long vowels. Vowels occur in all positions in a word, except for **o** which is not permitted at the end of it. It also has two diphthongs, **ai**, **au**.

	Front	Central	Back
High	i i:		u u:
Mid	e e:		o o:
Low		a a:	

Figure 2: Malayalam Vowel Phonology [109]

Consonants: Besides a Dravidian consonantal inventory, Malayalam has aspirated stops and supplementary sibilants borrowed from Indo-Aryan. [f] occurs mostly in European borrowings. Voiceless unaspirated stops, nasals and laterals [l], [ɭ] can be germinated. The distinction between single and geminated consonants is phonemic. Only six consonants, [m], [n], [ɳ], [r], [l], and [ɭ], can occur word finally.

		Labial	Dental	Retroflex	Palatal	Velar	Glottal
Stop	*Voiceless*	p pʰ	t tʰ	ʈ ʈʰ		k kʰ	
	Voiced	b bʰ	d dʰ	ɖ ɖʰ		g gʰ	
Affricate	*Voiceless*				ʧ ʧʰ		
	Voiced				ʤ ʤʰ		
Fricative	*Voiceless*	(f)	s	ʂ	ʃ		h
Nasal		m	n	ɳ	ɲ	ŋ	
Liquid			l r	ɭ ɽ			
Glide		w			j		

Figure 3: Malayalam Consonant Phonology [109]

Sandhi: internal and external sandhi are commonplace. They result in vowel and consonant deletion, assimilation of consonants and fusion.

Stress: it falls always on the first syllable of a word

Script and Orthography

Malayalam is written in an abugida script derived ultimately from Brāhmī in which every consonant carries an inherent a. The alphabetic order is based on phonological principles: it begins with the simple vowels and diphthongs followed by 25 stops and nasals arranged in five groups according to their place of articulation. It continues with semivowels (liquids and glides) and fricatives to end in two retroflex liquids which don't exist in Sanskrit and, thus, were not represented in Brāhmī.

Geminated consonants and other consonant clusters are written side by side or one above the other. Below each Malayalam sign appears the standard transliteration in the Latin alphabet, and between square brackets its equivalent in the International Phonetic Alphabet.

The following sections provide details of the Malayalam sounds and how these are written in Malayalam.

അ	ആ	ഇ	ഈ	ഉ	ഊ	ഋ	എ	ഏ	ഐ	ഒ	ഓ	ഔ
a	ā	i	ī	u	ū	ṛ	e	ē	ai	o	ō	au
[a]	[a:]	[i]	[i:]	[u]	[u:]	[rɨ]	[e]	[e:]	[ai]	[o]	[o:]	[au]

ക	ഖ	ഗ	ഘ	ങ	Velar stops and nasal
ka	kha	ga	gha	ṅa	
[ka]	[kʰa]	[ga]	[gʰa]	[ŋa]	
ച	ഛ	ജ	ഝ	ഞ	Palatal affricates and nasal
ca	cha	ja	jha	ña	
[tʃa]	[tʃʰa]	[dʒa]	[dʒʰa]	[ɲa]	
ട	ഠ	ഡ	ഢ	ണ	Retroflex stops and nasal
ṭa	ṭha	ḍa	ḍha	ṇa	
[ʈa]	[ʈʰa]	[ɖa]	[ɖʰa]	[ɳa]	
ത	ഥ	ദ	ധ	ന	Dental stops and nasal
ta	tha	da	dha	na	
[ta]	[tʰa]	[da]	[dʰa]	[na]	
പ	ഫ	ബ	ഭ	മ	Labial stops and nasal
pa	pha	ba	bha	ma	
[pa]	[pʰa/fa]	[ba]	[bʰa]	[ma]	
യ	ര	ല	വ		Semivowels (liquids and glides)
[ja]	[ra]	[la]	[wa]		
ശ	ഷ	സ	ഹ	ഴ	റ Fricatives and retroflex liquids
śa	ṣa	sa	ha	ḷa	ṟa
[ʃa]	[ʂa]	[sa]	[ha]	[ɭa]	[ɽa]

ṛ is a syllabic vowel found only in Sanskrit loanwords.
[f] is found mostly in Urdu and English loanwords and doesn't have a specific sign; it is represented with ph that also serves for [p^h].

Vowels

Vowels are written in this form when they are independently used.

അ U+0D05 A	ആ U+0D06 AA	ഇ U+0D07 I	ഈ U+0D08 II	ഉ U+0D09 U	ഊ U+0D0A UU	ഋ U+0D0B R
ഏ U+0D0E E	ഏ U+0D0F EE	ഐ U+0D10 AI	ഒ U+0D12 O	ഓ U+0D13 OO	ഔ U+0D14 AU	

Table 1: Malayalam Vowels

Vowel diacritics

Vowels can also be written as diacritics referred to as Matras, when these follow consonants. Their forms are given below, illustrated with the letter ക (U+0D15) MALAYALAM LETTER KA.

ക U+0D15 KA	കാ U+0D15 U+0D3E KAA	കി U+0D15 U+0D3F KI	കീ U+0D15 U+0D40 KII	കു U+0D15 U+0D41 KU	കൂ U+0D15 U+0D42 KUU	കൃ U+0D15 U+0D43 KR
കെ U+0D15 U+0D46 KE	കേ U+0D15 U+0D47 KEE	കൈ U+0D15 U+0D48 KAI	കൊ U+0D15 U+0D4A KO	കോ U+0D15 U+0D4B KOO	കൗ U+0D15 U+0D57 KAU	

Table 2: Malayalam Vowel Diacritics

Consonants

Malayalam has the following consonants, generally arranged my manner and place of articulation.

ക U+0D15 KA	ഖ U+0D16 KHA	ഗ U+0D17 GA	ഘ U+0D18 GHA	ങ U+0D19 NGA
ച U+0D1A CA	ഛ U+0D1B CHA	ജ U+0D1C JA	ഝ U+0D1D JHA	ഞ U+0D1E NYA
ട U+0D1F TTA	ഠ U+0D20 TTHA	ഡ U+0D21 DDA	ഢ U+0D22 DDHA	ണ U+0D23 NNA
ത U+0D24 TA	ഥ U+0D25 THA	ദ U+0D26 DA	ധ U+0D27 DHA	ന U+0D28 NA
പ U+0D2A PA	ഫ U+0D2B PHA	ബ U+0D2C BA	ഭ U+0D2D BHA	മ U+0D2E MA
യ U+0D2F YA	ര U+0D30 RA	റ U+0D31 RRA	ല U+0D32 LA	ള U+0D33 LLA

ഴ U+0D34 LLLA	വ U+0D35 VA	ശ U+0D36 SHA	ഷ U+0D37 SSA	സ U+0D38 SA
ഹ U+0D39 HA				

Table 3: Malayalam Consonants

Anusvaram and Visargam

Anusvaram: An anusvaram (അനുസ്വാരം anusvāram), or an anusvara, originally denoted the nasalization where the preceding vowel was changed into a nasalized vowel, and hence is traditionally treated as a kind of vowel sign. In Malayalam, anusvara represented as ◌ം (0D02) however, simply represents a consonant /m/ after a vowel, though this /m/ may be assimilated to another nasal consonant. It is a special consonant letter, different from a "normal" consonant letter, in that it is never followed by an inherent vowel or another vowel. In general, an anusvara at the end of a word in an Indian language is transliterated as ṁ in ISO 15919, but a Malayalam anusvara at the end of a word is transliterated as m without a dot.

Visargam: A visargam (വിസർഗം, visargam), or visarga, represents a consonant /h/ after a vowel, and is transliterated as ḥ. Like the anusvara, it is a special symbol, and is never followed by an inherent vowel or another vowel. In Malayalam, ◌ഃ (0D03) is the visarga symbol.

Chillu letters (Chillaksharam) and Samvruthokarams

In the Indo-European family of languages like Sanskrit, a large number of words end in consonants. But in Dravidian languages like Malayalam majority of words end in vowels. But, the chillaksharams of Malayalam are exceptions to this general feature. Chillaksharams are pure consonants, without any vowel sound. [111]

Chillaksharam is an original feature of Malayalam used only with 6 consonants at present. The consonants are ന (na), ണ (ṇa), ര (ra), ല (la) ള (ḷa) and ക (ka) and their corresponding chillus are ൻ (n̲), ൺ (ṇ), ർ (r), ൽ (l) ൾ (ḷ) and ൿ (k̲) in certain contexts, occur at the end of the word without the implicit vowel.

ൻ U+0D7A NN	ൺ U+0D7B N	ർ U+0D7C RR	ൽ U+0D7D L	ൾ U+0D7E LL	ൿ U+0D7F K

Table 4: Malayalam Chillu letters

Samvruthokaram is a soft ending virama (chandrakkala). Any consonant can be followed by consonant + ◌ു (0D41) + ◌് (0D4D), creating the samvruthokaram form of that consonant. In southern Kerala, the U matra ◌ു (0D41) and chandrakkala (virama) ◌് (0D4D) together form the grapheme for samvruthokaram. However, in northern Kerala, just chandrakkala (visible virama) standing alone is used. The chandrakkala alone at the end of a word is treated as Samvruthokaram.

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.

Examples of Samvruthokaram:

ഏതു് (***ethu*** meaning **which**) , code points - U+0D0F U+0D24 U+0D41 U+0D4D

അതു് (***athu*** meaning **that**) code points - U+0D05 U+0D24 U+0D41 U+0D4D

For the words that end in chillu, Samvruthokaram is used to make the pronunciation clearer. Either samvruthokaram is added directly to the word-ending chillaksharam, or the word-ending chillaksharam is geminated and Samvruthokaram is added to it.

The following are the main phonological transformations of chillaksharam. [113]

1. The word-ending consonant written as chillaksharam, is geminated and a samvrukthokaram is attached:

വിൺ -> വിണ്ടു് (viṇ -> viṇṭu)
മൺ -> മണ്ടു് (maṇ -> maṇṭu)
പൊൻ -> പൊന്നു് (poṉ -> poṉṉu)
പുൽ -> പുല്ലു് (pul -> pullu)

2. To the word-ending consonant written as chillaksharam, a samvrukthokaram is attached:

പാൽ -> പാലു് (pāl -> pālu)
മലർ -> മലരു് (malar -> malaru)
കോൺ -> കോണു് (kōṇ -> kōṇu)
തേൻ -> തേനു് (tēṉ -> tēṉu)

3. The chillaksharam undergoes the same phonological changes (in progressive/ regressive assimilation, gemination, etc) as in the case of other consonants in the context of combination of syllables :

വെൺ + നിലാവു് -> വെണ്ണിലാവു് (veṇ + nilāvu -> veṇṇilāvu)
കൺ + നീർ -> കണ്ണീർ (kaṇ + nīr̲ -> kaṇṇīr̲)
പൊൻ + ഓണം -> പൊന്നോണം (poṉ + ōṇam -> poṉṉōṇam)
വിൺ + തലം -> വിണ്ടലം (viṇ + talam -> viṇṭalam)

4. In sandhi, when a vowel follows a chillaksharam, they join in the same way as when vowels follow other consonants:

അവൻ + ഓടു് -> അവനോടു് (avan + ōṭu -> avanōṭu)
നീർ + ഇൽ -> നീരിൽ (nir̲ + il -> niril)
കവിൾ + ഇൽ -> കവിളിൽ (kaviḷ + il -> kaviḷil)

Even though Samvruthokaram may be seen as derived from the vowels അ (a) or ഉ (u), in fact, it has an independent identity as a vowel. This feature is seen only in Malayalam. [111]

A selection of conjunct consonants

A consonant can be combined with another consonant or conjunct using Virama. Conjuncts with more than four consonants are rare. The conjunct ഗ്ദ്ധ്യ is formed by five consonants.

	kka	ṅka	ṅṅa	cca	ñca	ññа	ṭṭa	ṇṭa	ṇṇa	tta	nta	nna
NLF	ക്ക	ങ്ക	ങ്ങ	ച്ച	ഞ്ച	ഞ്ഞ	ട്ട	ണ്ട	ണ്ണ	ത്ത	ന്ത	ന്ന
LF	ക്ക	ങ്ക	ങ്ങ	ച്ച	ഞ്ച	ഞ്ഞ	ട്ട	ണ്ട	ണ്ണ	ത്ത	ന്ത	ന്ന

Table 5: Malayalam Conjunct Consonants

NLF - Non-ligated form has a visible virama (chandrakkala)
LF- Ligated form in which consonants are conjoined fully or partially (as rendered by fonts)

Conjuncts with diacritics using യ (U+0D2F), ര (U+0D30), ല (U+0D32), വ (U+0D35)

Conjunct consonants formed with യ (0D2F), ര (0D30), ല (0D32) and വ (0D35) are rendered with diacritic marks/signs in the glyph. Examples of these in combination with

ക (0D15) and ഩ (0D2A) are given below. Other consonants can be combined in similar fashion.

Consonant + ്യ	Consonant + ്ര	Consonant + ്ല	Consonant + ്വ
ക്യ (0D15 0D4D 0D2F)	ക്ര (0D15 0D4D 0D30)	ക്ല (0D15 0D4D0D32)	ക്വ (0D15 0D4D 0D35)
ഩ്യ (0D2A 0D4D 0D2F)	ഩ്ര (0D2A 0D4D 0D30)	ഩ്ല (0D2A 0D4D0D32)	ഩ്വ (0D2A 0D4D 0D35)

Table 6: Malayalam Conjuncts with diacritics
using ്യ (U+0D2F), ്ര (U+0D30), ്ല (U+0D32), ്വ (U+0D35)

4. Overall Development Process and Methodology

Neo-Brahmi Generation Panel (NBGP) has been formed by members having experience in linguistics and computational linguistics. Under the Neo-Brahmi Generation Panel, there are nine scripts belonging to separate Unicode blocks. Each of these scripts is assigned a separate LGR; however Neo-Brahmi GP ensures that the fundamental philosophy behind building those LGRs are all in sync with all other Brahmi derived scripts.

4.1 Guiding Principles

The NBGP adopts the following broad principles for the selection of code-points in the code-point repertoire across the board for all the scripts within its ambit.

4.1.1 Inclusion principles:

4.1.1.1 Modern usage:

Every character proposed should be in the everyday usage of a particular linguistic community. Characters which have been encoded in Unicode for transcription purposes only or for archival purposes will not be considered for inclusion in the code-point repertoire.

4.1.1.2 Unambiguous use:

Every character proposed should have unambiguous understanding among the linguistic community about its usage in the language.

4.1.2 Exclusion principles:

The main exclusion principle is that of External Limits on Scope. These comprise protocols or standards which are prerequisites to the Label Generation Rulesets. All

further principles are in fact subsumed under these limitations but have been spelt out separately for the sake of clarity.

4.1.2.1 External Limits on Scope:

The code point repertoire for root zone being a very special case, at the top of the protocol hierarchies, the range of available characters for selection as a part of the Root Zone code point repertoire is already constrained by various protocol layers beneath it. The following three main protocols/standards act as successive filters:

i. The Unicode Chart:

Out of all the characters that are needed by the given script, if the character in question is not encoded in Unicode, it cannot be incorporated in the code point repertoire. Such cases are quite rare, given the elaborate and exhaustive character inclusion efforts made by the Unicode Consortium.

ii. IDNA Protocol:

Unicode being the character encoding standard for providing the maximum possible representation of a given script/language, it has encoded as far as possible all the possible characters needed by the script. However, the Domain name being a specialized case, it is governed by an additional protocol known as IDNA (Internationalized Domain Names in Applications). The IDNA protocol introduces exclusion of some characters out of Unicode repertoire from being part of the domain names.

iii. Maximal Starting Repertoire:

The Root-zone LGR being a repertoire of the characters which are going to be used for creation of the root zone TLDs, which in turn are an even more specialized case of domain names, the ROOT LGR procedure introduces additional exclusions on IDNA allowed set of characters.

Example: MALAYALAM SIGN AVAGRAHA "ഽ " (U+ 0D3D) even if allowed by IDNA protocol, is not permitted in the Root Zone Repertoire as per the [MSR].

To sum up, the restrictions start off by admitting only such characters as are part of the code-block of the given script/language. This is further narrowed down by the IDNA 2008 Protocol and finally an additional filter in the form of Maximal Starting Repertoire restricts the character set associated with the given language even more.

4.1.2.2 No Rare and Obsolete Characters:

There are characters which have been added to Unicode to accommodate rare forms like MALAYALAM LETTER VOCALIC L "ഌ" (U+0D0C), which is an obsolete vowel used to write Sanskrit words and is not considered as part of the modern Malayalam orthography. All such characters will not be included. This is in consonance with the Conservatism principle as laid down in the Root Zone LGR procedure.

5. Repertoire

Based on the LGR Procedure for the Root Zone and the MSR, NBGP conducted the code point analysis of the Malayalam script. The analysis is presented in this section, including the list of code points recommended for inclusion and exclusion from the repertoire.

5.1 Malayalam section of Maximal Starting Repertoire [MSR] Version 3

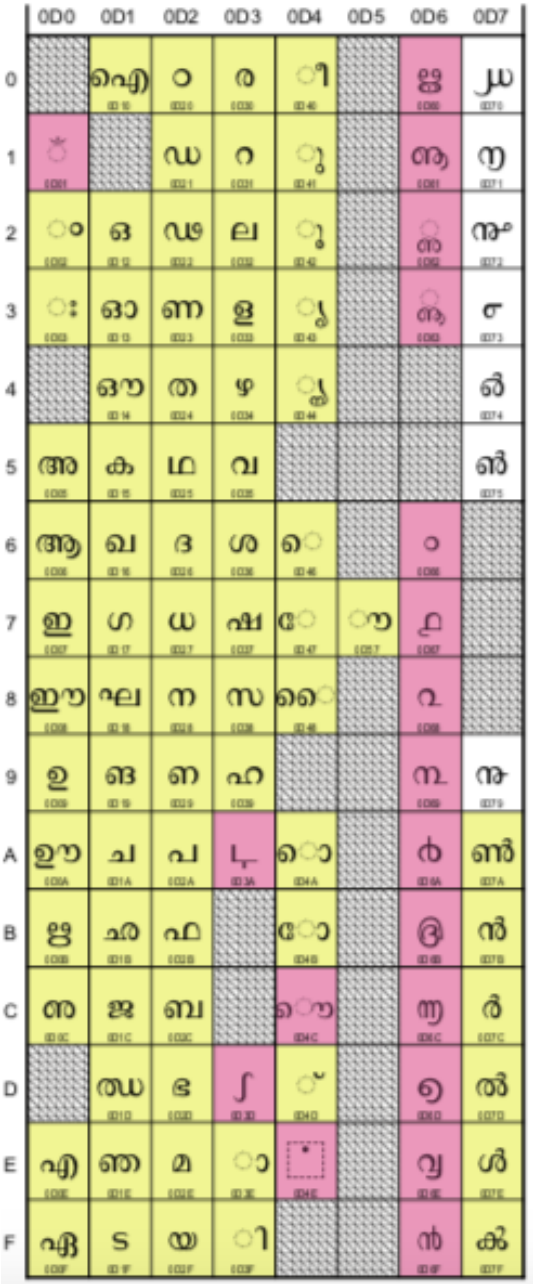


Figure 4: Malayalam Code Page from [MSR]

Color convention[1]:

All characters that are included in the [MSR] - Yellow background

PVALID in IDNA2008 but excluded from the [MSR] - Pinkish background

Not PVALID in IDNA2008 - White background

[1]This document needs to be printed in color for this to be read correctly.

5.2 Unicode Code Points Inclusion

The following code points are included in the repertoire.

Sr. No.	Unicode Code Point	Glyph	Character Name	Indic Syllabic Category	Refs.
1	0D02	◌ം	MALAYALAM SIGN ANUSVARA	Anusvaram	[106]
2	0D03	◌ഃ	MALAYALAM SIGN VISARGA	Visargam	[106]
3	0D05	അ	MALAYALAM LETTER A	Vowel	[106]
4	0D06	ആ	MALAYALAM LETTER AA	Vowel	[106]
5	0D07	ഇ	MALAYALAM LETTER I	Vowel	[106]
6	0D08	ഈ	MALAYALAM LETTER II	Vowel	[106]
7	0D09	ഉ	MALAYALAM LETTER U	Vowel	[106]
8	0D0A	ഊ	MALAYALAM LETTER UU	Vowel	[106]
9	0D0B	ഋ	MALAYALAM LETTER VOCALIC R	Vowel	[106]
10	0D0E	എ	MALAYALAM LETTER E	Vowel	[106]
11	0D0F	ഏ	MALAYALAM LETTER EE	Vowel	[106]
12	0D10	ഐ	MALAYALAM LETTER AI	Vowel	[106]
13	0D12	ഒ	MALAYALAM LETTER O	Vowel	[106]
14	0D13	ഓ	MALAYALAM LETTER OO	Vowel	[106]
15	0D14	ഔ	MALAYALAM LETTER AU	Vowel	[106]
16	0D15	ക	MALAYALAM LETTER KA	Consonant	[106]
17	0D16	ഖ	MALAYALAM LETTER KHA	Consonant	[106]
18	0D17	ഗ	MALAYALAM LETTER GA	Consonant	[106]
19	0D18	ഘ	MALAYALAM LETTER GHA	Consonant	[106]
20	0D19	ങ	MALAYALAM LETTER NGA	Consonant	[106]
21	0D1A	ച	MALAYALAM LETTER CA	Consonant	[106]
22	0D1B	ഛ	MALAYALAM LETTER CHA	Consonant	[106]

23	0D1C	ജ	MALAYALAM LETTER JA	Consonant	[106]
24	0D1D	ഝ	MALAYALAM LETTER JHA	Consonant	[106]
25	0D1E	ഞ	MALAYALAM LETTER NYA	Consonant	[106]
26	0D1F	ട	MALAYALAM LETTER TTA	Consonant	[106]
27	0D20	ഠ	MALAYALAM LETTER TTHA	Consonant	[106]
28	0D21	ഡ	MALAYALAM LETTER DDA	Consonant	[106]
29	0D22	ഢ	MALAYALAM LETTER DDHA	Consonant	[106]
30	0D23	ണ	MALAYALAM LETTER NNA	Consonant	[106]
31	0D24	ത	MALAYALAM LETTER TA	Consonant	[106]
32	0D25	ഥ	MALAYALAM LETTER THA	Consonant	[106]
33	0D26	ദ	MALAYALAM LETTER DA	Consonant	[106]
34	0D27	ധ	MALAYALAM LETTER DHA	Consonant	[106]
35	0D28	ന	MALAYALAM LETTER NA	Consonant	[106]
36	0D2A	പ	MALAYALAM LETTER PA	Consonant	[106]
37	0D2B	ഫ	MALAYALAM LETTER PHA	Consonant	[106]
38	0D2C	ബ	MALAYALAM LETTER BA	Consonant	[106]
39	0D2D	ഭ	MALAYALAM LETTER BHA	Consonant	[106]
40	0D2E	മ	MALAYALAM LETTER MA	Consonant	[106]
41	0D2F	യ	MALAYALAM LETTER YA	Consonant	[106]
42	0D30	ര	MALAYALAM LETTER RA	Consonant	[106]
43	0D31	റ	MALAYALAM LETTER RRA	Consonant	[106]
44	0D32	ല	MALAYALAM LETTER LA	Consonant	[106]
45	0D33	ള	MALAYALAM LETTER LLA	Consonant	[106]
46	0D34	ഴ	MALAYALAM LETTER LLLA	Consonant	[106]
47	0D35	വ	MALAYALAM LETTER VA	Consonant	[106]

48	0D36	ശ	MALAYALAM LETTER SHA	Consonant	[106]
49	0D37	ഷ	MALAYALAM LETTER SSA	Consonant	[106]
50	0D38	സ	MALAYALAM LETTER SA	Consonant	[106]
51	0D39	ഹ	MALAYALAM LETTER HA	Consonant	[106]
52	0D3E	ാ	MALAYALAM VOWEL SIGN AA	Matra	[106]
53	0D3F	ി	MALAYALAM VOWEL SIGN I	Matra	[106]
54	0D40	ീ	MALAYALAM VOWEL SIGN II	Matra	[106]
55	0D41	ു	MALAYALAM VOWEL SIGN U	Matra	[106]
56	0D42	ൂ	MALAYALAM VOWEL SIGN UU	Matra	[106]
57	0D43	ൃ	MALAYALAM VOWEL SIGN VOCALIC R	Matra	[106]
58	0D46	െ	MALAYALAM VOWEL SIGN E	Matra	[106]
59	0D47	േ	MALAYALAM VOWEL SIGN EE	Matra	[106]
60	0D48	ൈ	MALAYALAM VOWEL SIGN AI	Matra	[106]
61	0D4A	ൊ	MALAYALAM VOWEL SIGN O	Matra	[106]
62	0D4B	ോ	MALAYALAM VOWEL SIGN OO	Matra	[106]
63	0D4D	്	MALAYALAM SIGN VIRAMA	Chandrakkala / Virama	[106]
64	0D57	ൗ	MALAYALAM AU LENGTH MARK	Matra	[106]
65	0D7A	ൺ	MALAYALAM LETTER CHILLU NN	Chillu Letters	[106]
66	0D7B	ൻ	MALAYALAM LETTER CHILLU N	Chillu Letters	[106]
67	0D7C	ർ	MALAYALAM LETTER CHILLU RR	Chillu Letters	[106]
68	0D7D	ൽ	MALAYALAM LETTER CHILLU L	Chillu Letters	[106]
69	0D7E	ൾ	MALAYALAM LETTER CHILLU LL	Chillu Letters	[106]
70.	0D7F	ൿ	MALAYALAM LETTER CHILLU K	Chillu Letters	[106]

Table 7: Malayalam Code Point Repertoire

5.3 Code Point Sequence

The following sequences has been defined for the purpose of variant and the WLE rules (see section 6.1).

1.	U+0D28 U+0D4D U+0D31	ന ് റ [ന്റ]	MALAYALAM LETTER NA MALAYALAM SIGN VIRAMA MALAYALAM LETTER RRA
2	U+0D33 U+0D4D U+0D33	ള ് ള [ള്ള]	MALAYALAM LETTER LLA MALAYALAM SIGN VIRAMA MALAYALAM LETTER LLA
3	U+0D7B U+0D31	ൻ റ [ൻറ]	MALAYALAM LETTER CHILLU N MALAYALAM LETTER RRA

Table 7a: Malayalam Code Point Sequences

5.4 Unicode Code Point Exclusion

The following code points are excluded because they are archaic or obsolete in current Malayalam orthography.

Sr. No.	Unicode Code Point	Glyph	Character Name	Indic Syllabic Category	Reason
1.	0D0C	ഌ	MALAYALAM LETTER VOCALIC L	Vowel	ഌ (0D0C) an obsolete vowel used to write Sanskrit words. The letter ഌ is very rare, and are not considered as part of the modern Malayalam orthography.
2.	0D44	◌ൄ	MALAYALAM VOWEL SIGN VOCALIC RR	Matra	◌ൄ (0D44) is the matra sign of obsolete vowel VOCALIC RR ൠ (0D60) which is not among the approved codepoints in MSR-3. It is no longer used in Malayalam orthography.

3.	0D29	൩	MALAYALAM LETTER NNNA	Consonant	൩ (0D29) corresponds to Tamil ṉa ன. Used rarely in scholarly texts to represent the alveolar nasal, as opposed to the dental nasal. [108]. In ordinary texts both are represented by na ന (0D28).

Table 8: Malayalam Excluded Code Point

6. Variants

This section discusses the variant code points found in Malayalam within script and with other related scripts.

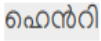
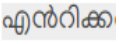
6.1 In-script variants

This section lists sequences that should be considered variants of each other.

Set #		Characters	Code Points	Glyph
1.	a)	ന് + റ	0D28 +0D4D +0D31	ന്റ or ന്റ
	b)	ൻ + ് + റ	0D7B + 0D4D + 0D31	ന്റ
	c)	ൻ + റ	0D7B + 0D31	ൻറ
2.	a)	ള് + ള	0D33 + 0D4D + 0D33	ള്ള
	b)	ള + ള	0D33 + 0D33	ള്ള

Table 9: In-script Variant Analysis

Set 1: These are various ways to write the conjunct "***nta***" in Malayalam. 1 a) Here ***nta*** is encoded as a combination of 0D28 +0D4D +0D31 and it is rendered as ന്റ in most of the Malayalam Unicode fonts and a few of the Microsoft fonts render it as ന്റ.

1 b) is how some Microsoft fonts have encoded ***nta*** 0D7B + 0D4D + 0D31 and it is rendered as in those fonts and as  in other fonts. Because the rendering problem, it is safe to disallow this sequence by WLE. (please see WLE Rule1). Although 1. c) has also been used historically to write ***nta*** and such sequential style of writing is still in use, that combination can also be used to write ***nra*** in words like  (Henry) or  (Enrica). [112] Hence the sequence of 1. c) is allowed. The variants in set 1contains only two sequences: 0D7B + 0D31 and 0D28 +0D4D +0D31. And the disposition is “blocked”.

Set 2: The consonant ള (0D33) rarely follows another ള in Malayalam, except in the case of some place names. The double conjunct of ള (0D33) formed by code points 0D33 + 0D4D + 0D33 is rendered as the glyph ള്ള which looks visually very similar to a ള following another ള. This can result in spoofed labels. For example, in Malayalam we write “***vellam***” as “വെള്ളം” - 0D35 0D46 0D33 0D4D 0D33 0D02 (meaning: water), a spoofed label can write it as “വെളളം” - 0D35 0D46 0D33 0D33 0D02. This should be blocked.

However, this pattern gives rise to some complications because it effectively makes the Halant (0D4D) a variant of a "null position", in this case, whenever it occurs between two instances of 0D33 ള LLA. Variant definitions of that nature can lead to unexpected results because a label: 0D33 **0D4D** 0D33 **0D4D** 0D33 can be analyzed two ways:

{0D33 **0D4D** 0D33} {**0D4D**} {0D33} and {0D33} {**0D4D**} {0D33 **0D4D** 0D33}

NBGP takes into account the data provided by the IP on occurrences of the labels in certain labels where a consonant ള (0D33) follows another ള and found that the percentage is small. However, the community feedback shows an increase in usage due to foreign-language-borrowed words language. The detailed analysis and supporting data can be found in Appendix C.

Therefore, NBGP has decided not to define Set 2 as variants, but to handle this case by using a WLE rule. The rule will not allow a first consonant ള (0D33) in a label to be immediately followed by a second 0D33, but requires an H (0D4D) (or any other eligible code point) to intervene.

A sequence 0D4D 0D33 0D33 has been defined in the repertoire section. Adding such a sequence would have the effect of allowing the case “ള്ള” (0D33 0D4D 0D33 0D33) while continue to disallow 0D33 0D33 everywhere else, including in “ള്ള” (0D33 0D33 0D4D 0D33).

6.2 Cross-Script Variants

The Malayalam characters in tables below are considered variant code points with some characters in Oriya and Tamil as they could be considered visually same for the users. See Appendix A for additional code points for other scripts which are visually similar but not considered as variant code points for the reasons listed.

6.2.1 Cross-script variants for Tamil and Malayalam

Variant Set	Tamil		Malayalam	
	CP	Glyph	CP	Glyph
1.	0B9C	ஜ	0D1C	ജ
2.	0BB5	வ	0D16	വ
3.	0BAE	ம	0D25	ഥ
4.	0BBF	ி	0D3F	ി
5.	0BC6	ெ	0D46	െ
6.	0BC7	ே	0D47	േ

Table 10: Tamil – Malayalam Cross Script Variants

6.2.2 Cross-script variants for Oriya and Malayalam

Case of Malayalam and Odia (Oriya) TTHA Consonant:

This is the case of "Consonant Ttha" which happened to retain the same shape despite being part of different scripts, i.e., Malayalam and Odia. These characters are:

ഠ - MALAYALAM LETTER TTHA (U+0D20)

ଠ - ORIYA LETTER TTHA (U+0B20)

Both characters look exactly alike and belong to a "Consonant" category. As they are consonants, each of them, even in the simplest form i.e. the characters themselves, are valid labels. As per the NBGP cross-script variant inclusion policy (Appendix D), this is a

valid case for inclusion. Also, even if they are single characters, when the same character combines, theoretically they can form an infinite[2] number of cross-script variant labels between the scripts involved. Here are samples of some of those labels:

Malayalam	Oriya
ഠഠഠ U+0D20 U+0D20 U+0D20	ଠଠଠ U+0B20 U+0B20 U+0B20
ഠഠഠഠ U+0D20 U+0D20 U+0D20 U+0D20	ଠଠଠଠ U+0B20 U+0B20 U+0B20 U+0B20
ഠഠഠഠഠ U+0D20 U+0D20 U+0D20 U+0D20 U+0D20	ଠଠଠଠଠ U+0B20 U+0B20 U+0B20 U+0B20 U+0B20

Since, having such labels is a realistic possibility and the corresponding labels look almost exactly alike, NBGP has proposed them (together with similar combining marks) as blocked variants.

Variant Set	Oriya		Malayalam	
	CP	Glyph	CP	Glyph
1.	0B20	ଠ	0D20	ഠ

Table 11: Oriya – Malayalam Cross Script Variants

7. Whole Label Evaluation (WLE) Rules

This section provides the WLE rules that are required by all the languages mentioned in Section 4 when written in Malayalam Script. The rules have been drafted in such a way that they can be easily translated into the LGR specifications.

Below are the symbols used in the WLE rules, for each of the "Indic Syllabic Category" as mentioned in the table provided for code point repertoire in Section 5.

[2]Though theoretically infinite, this number would be limited to the number of such labels whose equivalent punycode string would not exceed 63 characters including the ACE prefix "xn--".

7.1.1 Variables or definitions

V	→	Vowel
M	→	Matra (Vowel Sign)
C	→	Consonant
L	→	Chillu
H	→	Chandrakkala/Halant/Virama (◌് U+0D4D)
B	→	Anusvaram (◌ം U+0D02)
X	→	Visargam (◌ഃ U+0D03)

7.1.2 Rules for Forming Aksharam

Rule 1: H must be preceded by C or the M ◌ൃ (0D41) (Samvruthokaram)

Rule 2: M must be preceded by C

Rule 3: B must be preceded by C, V or M

Rule 4: X must be preceded by C, V or M

Rule 5: L cannot be preceded by B, X or H

Rule 6: Label does not begin with L

Rule 7: The ള (0D33) cannot immediately follow ള (0D33)

8. Contributors

Neo-Brahmi Generation Panel (NBGP)
Veena Solomon (veena.ycet@gmail.com)
Prasad Pattarumadom Kesava Kurup (pkpdelhi@gmail.com)
Santhosh Thottingal (santhosh.thottingal@gmail.com)
Anivar Aravind (anivar@indicproject.org)
Jijo Pappachan (jijospeaks@yahoo.com)

9. References

[MSR] Integration Panel, "Maximal Starting Repertoire — MSR-3 Overview and Rationale", 28 March 2018 https://www.icann.org/en/system/files/files/msr-3-overview- 28mar18-en.pdf

[EGIDS] Expanded Graded Intergenerational Disruption Scale, https://www.ethnologue.com/about/language-status (Accessed on 5th July, 2018)

[101] Unicode® Standard Annex #31 Mark Davis, “Unicode Identifier And Pattern Syntax”: 2.3 Layout and Format Control Characters http://unicode.org/reports/tr31/#Layout_and_Format_Control_Characters (Accessed on 5th July, 2018)

[102] “Report on Malayalam Unicode Issues” (2012) prepared by Santhosh Thottingal (also part of NEGP) and submitted to Unicode via Wikimedia Foundation. It discusses both chillu and nta issues:

http://thottingal.in/documents/ReportonMalayalamUnicodeIssues.pdf (Accessed on 5th July, 2018)
[103] ഓളം Dictionary, https://olam.in/ (Accessed on 5th July, 2018)
[104] Roozbeh Pournader and Cibu Johny, “Old and New Chillus in Malayalam and implications for Sinhala” http://www.unicode.org/L2/L2013/13036-chillus-uptake.pdf (Accessed on 5th July, 2018)
[105] Wikipedia, “Malayalam script” https://en.wikipedia.org/wiki/Malayalam_script (Accessed on 5th July, 2018)
[106] Omniglot, “Malayalam (മലയാളം)” https://www.omniglot.com/writing/malayalam.htm (Accessed on 5th July, 2018)
[107] The Unicode Standard, Version 10.0., Chapter 12 “South and Central Asia I: Official Scripts of India”, https://www.unicode.org/versions/Unicode10.0.0/ch12.pdf#page=65 (Accessed on 5th July, 2018)
[108] Everson, Michael (2007). "Proposal to add two characters for Malayalam to the BMP of the UCS" (PDF). ISO/IEC JTC1/SC2/WG2 N3494. Retrieved 2009-09-09: http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3494.pdf (Accessed on 5th July, 2018)
[109] Alejandro Gutman and Beatriz Avanzati “Malayalam, The Language Gulper” http://www.languagesgulper.com/eng/Malayalam.html (Accessed on 5th July, 2018)
[110] Malayalam Range: 0D00–0D7F, The Unicode Standard, Version 11.0 https://unicode.org/charts/PDF/U0D00.pdf (Accessed on 5th July, 2018)
[111] R. Chitrajakumar , N. Gangadharan Rachana Akshara Vedi“Samvruthokaram and Chandrakkala” https://www.unicode.org/L2/L2005/05213-samvruktokaram.pdf (Accessed on 2nd August, 2018)
[112] Santhosh Thottingal, “ന്റ - ഭാഷ, യുണിക്കോഡ്, ചിത്രീകരണം” https://blog.smc.org.in/nta-rendering-rules/ (Accessed on Aug 2nd, 2018]
[113] R. Chitrajakumar , N. Gangadharan Rachana Akshara Vedi “Chillaksharam of Malayalam Language” https://unicode.org/L2/L2005/05214-chillu.pdf (Accessed on 27th August 2018)

10. Appendix A: Excluded In-Script Variants

As the following formations are not valid as per Aksharam formation rules, these cases are not proposed as variants.

1.	ഌൗ	0D08	ഌൗ
	ഌ + ൗ	0D07 + 0D57	ഌൗ
2.	ൡൗ	0D0A	ൡൗ
	ൡ + ൗ	0D09 + 0D57	ൡൗ
3.	ഔൗ	0D14	ഔൗ
	ഒ + ൗ	0D12 + 0D57	ഒൗ
4.	ഓ	0D13	ഓ
	ഒ + ാ	0D12 + 0D3E	ഒാ
5.	ഐ	0D10	ഐ
	ഏ + ൊ	0D0E + 0D46	ഏൊ

Table A-1: Excluded In-Script Variants Due to Invalid Combination

In Table A-2, Column 1: These vowel signs have glyph pieces which stand on both sides of the consonant; they follow the consonant in logical order, and should be handled as a unit for most processing. Column 2: Although, Unicode defines this canonical decomposition, the Standard recommends not to use the sequence [107], p501. Therefore, it is not advisable to use them in IDN labels; they are blocked here by akshara formation rule.

Code Point 1 + Glyph 1	Code Point 2 + Glyph 2
ൊാ (0D4A)	ൊ (0D46) + ാ (0D3E)
ോാ(0D4B)	ോ (0D47) + ാ (0D3E)
ൗ (0D57)	ൊ (0D46) + ൗ (0D57)

Table A-2: Split Vowel Case

11. Appendix B: Confusable Code Points

The code-points below are visually confusing only in smaller fonts and can be excluded from consideration as variant code points.

Tamil	Malayalam
ஸ (0BB8)	സ (0D38)

Table B-1: Tamil-Malayalam Confusable Code Points

Oriya	Malayalam
ଂ (0B02)	ംം (0D02)
ଃ (0B03)	ംഃ (0D03)

Table B-2: Oriya-Malayalam Confusable Code Points

At Sri Lanka face to face meeting, it was decided to exclude the code points below from variant list as these do not look alike, due to round - square structural differences.

Kannada	Malayalam
ಲ (0CB2)	ഩ (0D32)

Table B-3: Kannada-Malayalam Confusable Code Points

Telugu	Malayalam
ఌ (0C32)	ഩ (0D32)

Table B-4: Telugu-Malayalam Confusable Code Points

Code points in Table B-5, B-6, and B-7 would qualify as cross-script code point variants but there are not enough of them to form a variant labels, therefore these cases can be excluded. (If only combining marks are variants for a given script, no label can be formed without using at least one non-variant code point). In the case of Sinhala, the relevant base character is distinct.

Kannada	Malayalam
ಂ (0C82)	ംം (0D02)
ಃ (0C83)	ംഃ (0D03)

Table B-5: Kannada-Malayalam Too Few Identical Code Points

Telugu	Malayalam
◌ం (0C02)	◌ം (0D02)
◌ః (0C03)	◌ഃ (0D03)

Table B-6: Telugu-Malayalam Too Few Identical Code Points

Sinhala	Malayalam
◌ං (0D82)	◌ം (0D02)
◌ඃ (0D83)	◌ഃ (0D03)

Table B-7: Sinhala-Malayalam Too Few Identical Code Points

NBGP also considers that 0D1F (ട) MALAYALAM LETTER TTA is similar to 0073 (s) LATIN SMALL LETTER S and 0455 (s) CYRILLIC SMALL LETTER DZE. However, Latin script and Cyrillic script are not derived from the Brahmi script. This case is out of scope of NBGP cross script variant analysis.

12. Appendix C: Case of ള (0D33) + ള (0D33)

The consonant ള (0D33) rarely follows another ള in Malayalam, except in the case of some place names. The double conjunct of ള (0D33) formed by code points 0D33 + 0D4D + 0D33 is rendered as the glyph ള്ള which looks visually very similar to a ള following another ള. This can result in spoofed labels. For example, in Malayalam we write "***vellam***" as "വെള്ളം" - 0D35 0D46 0D33 0D4D 0D33 0D02 (meaning: water), a spoofed label can write it as "വെളളം" - 0D35 0D46 0D33 0D33 0D02.

Combination	Code points	Glyph
ള് + ള	0D33 + 0D4D + 0D33	ള്ള
ള + ള	0D33 + 0D33	ളള

Table C-1: Case of ള (0D33) + ള (0D33)

This has been restricted by a WLE rule 7. It allows the combination “ള്ള” (0D33 0D4D 0D33 0D33) which is present in words like “ഉള്ളളവ്” (meaning: inner dimension viz. volume), and blocks the combination “ള്ള” (0D33 0D33 0D4D 0D33) which is rarely found in usage. The existence of “ള്ള” (0D33 0D33) in considerable percentage on the web can be attributed to misspelling due to extreme visual similarity.

==========

Proposed recommendation from the Integration Panel

==========

Proposed recommendation for Malayalam

DATE: 2018-06-12

Overview

The IP recently discovered a technical issue with the proposed variants for Malayalam.

Issue Statement

The Malayalam LGR defines the following variant

0D33 0D33 <-> 0D33 **0D4D** 0D33 (i.e.: ള്ള <--> ള്ള)

This pattern gives rise to some complications because it effectively makes the Halant (0D4D) a variant of a "null position", in this case, whenever it occurs between two instances of 0D33 ള LLA. Variant definitions of that nature can lead to unexpected results because a label 0D33 **0D4D** 0D33 **0D4D** 0D33 can be analyzed two ways:

{0D33 **0D4D** 0D33} {**0D4D**} {0D33} and

{0D33} {**0D4D**} {0D33 **0D4D** 0D33}

As a result of this, variant definitions of this nature, although seemingly well-defined on the ***code point*** level can lead to unexpected variant relations among ***labels.***

Therefore, such kinds of variant sequence definitions cannot be used without some further restriction. Below the IP will suggest two possible approaches and requests that the GP consider them in light of the knowledge of how the script is used.

Background:

Looking at the Malayalam sample file the IP notes:

0D33 0D33 ള്ള exists once (1) in sample of 60K labels

(it's part of the longer pattern: 0D33 **0D4D** 0D33 0D33 or ള്ളള്)

0D33 0D33 0D33 (ള്ളള്) exists (0) times

0D33 **0D4D** 0D33 (ള്ള) exists 523 times, or .9% of the total; of these:

- 1/10 or 52 are followed by an 0D4D (Halant): 0D33 **0D4D** 0D33 **0D4D** (ള്)
- none (0) is of the pattern 0D33 **0D4D** 0D33 **0D4D** 0D33 (or longer)

From this one can conclude:

- ളള is quite frequent and can be spoofed by ളള (which doesn't occur normally or at least not frequently)
- ള്ള also occurs with some frequency and could be spoofed by ള്ള (the latter again not seen in the sample)
- ളളള does occur, if rarely, and can be spoofed by ളളള or ളളള, but not by ള്ളള (where the code points are: 0D33 **0D4D** 0D33 0D33, 0D33 0D33 **0D4D** 0D33 and 0D33 **0D4D** 0D33 **0D4D** 0D33)

Under the definition in the proposed LGR ളളള and ളളള are not actually variant labels of each other, while ള്ളള is a variant of ളളള even though it shouldn't be. (The reason why the last label shouldn't be a variant label is because the second halant would be rendered visibly, making it distinct.)

Longer patterns are either rare or do not occur in standard sample; they seem quite likely to be nonsensical (at least some of them). Therefore, the cases seen so far would appear to be the total set of cases where there is a practical need for some variants or other restriction.

Options

The IP identified two suggested options to resolve the issue.

Option One

Restricting the variant so it cannot occur following an 0D33 ള or Halant.

If the variant can be limited to the beginning of a cluster, that is, a requirement added that it only applies when not following an 0D33 of **0D4D**, then we can take still care of the most frequent and second most frequent case, and these cases produce variant labels that are related in expected ways: longer strings of alternating 0D33 and **0D4D** pose no problems as any alternate grouping of code points into sequences does not lead to any

additional variants. Only the leading {0D33 0D33} or {0D33 **0D4D** 0D33} would cause variants. In particular ൞്൞ (with a visible Halant) would not become a variant of ൞൞, etc. However, cases like ൞൞ / ൞൞ / ൞൞ would still not fully work as intended as the first and second label would not be variants of each other, and only the first would be a variant of the last.

Option Two

Restricting valid labels to exclude ൞൞

Restricting labels from containing two 0D33 ൞ that are not joined by a Halant would robustly prevent any spoofing. However, it would also disallow a small number of potentially meaningful labels. (About 0.0015% of the words in the test file are affected - or 1 in 60K). No variant definition would be needed.

Recommendation

The IP requests the NeoB GP to study these options and to consider them in determining a proposed approach to fixing the issue with the kind of variant mapping mentioned at the head of the document.

We realize that these represent a trade-off. For the Root Zone we feel comfortable that restriction of the allowed labels to avoid some problem cases is definitely appropriate, even if the process contains a String Review phase that would allow the manual weeding out of specific bad cases.

However, we feel that an option that leaves some, if rare, opportunities for spoofing may well be inappropriate for the second and other levels as well: for those levels, human oversight of the process is going to be even less available.

The IP suggests that the GP also weigh the extent to which decisions for the Root Zone affect other zones (by example).

==========

Feedback from community

==========

നീളള്ള മുടി, ***neelalla mudi*** is how people say നീളമുള്ള മുടി, ***neelamulla mudi*** [meaning: long hair, lit. hair with length], locally in Valluvanad area of North Kerala. Similarly,

നല്ല താളള്ള പാട്ട്, ***nalla thaalalla paattu***, is the same as നല്ല താളമുള്ള പാട്ട്, ***nalla thaalamulla paattu*** [meaning: (a) song with good rhythm]

വെളള്ളള്ള കിണറ്, ***vellalla kinaru,*** is വെള്ളമുള്ള കിണറ്, ***vellamulla kinaru*** [meaning: (a) well with water] This label is not blocked because ള്ളള is allowed.

I don't these needs to be considered as the **ള്ളള** part in these labels is a spoken contraction of **ള്ളള,** ulla [meaning: having, with].

In other parts of Kerala, the spoken dialect changes the contraction to "ള്ളള" or ള്ളള which are allowed as per the rule.

Then there are some place names like **മാള്ളള**. On doing a Google search, I got only a single result [google.co.in].

Feedback from the community:

I won't recommend adding such rules based on the existence of current (and popular) vocabulary of 2018. Malayalam has an active practice of borrowing words from other languages than inventing native words (mainly from English nowadays). Because of this anything that is a valid conjunct can come into the language. Here is an example: You may know, I am a typeface designer too. When some of our initial fonts did not have the OpenType rules to handle സ്+ബ , സ്+ബു, it was because nobody could find a word that can have such a combination. Later, around 2010, Facebook became a thing. People

started writing it in Malayalam. Our fonts could not handle the rendering gracefully and then we added the required ligatures and rules and released a new version. While I was working on another typeface, another conjunct ജ്+മ was not supporting thinking, there is no Malayalam word with ജ്മ. But later a friend came and complained he wants to have an error-free rendering for അജ്മീർ.. So that is about the 'reasoning of rare occurrence in Malayalam'. Btw, there are people and places with name മാളള്ള (Malalla) - try a google search. We people from Valluvanad area often has this നല്ല നീളള്ള മുടി, നല്ല താളള്ള പാട്ട് , വെളളള്ള കിണറ്...
A google search for വെളളള്ള shows me that it is a place name in Idukki.

About the visual similarity, again, as a type designer, we consciously make them visually different while designing. ള്+ള -> ള്ള appear very joined with the tails fused together, While ളള appear with enough spacing between the letters and no fusing of tails.

Also, ററ is a similar case where people write two Ra together to get /tta/ , Almost all fonts nowadays stack them if it is for /tta/. But not guaranteed. So similar arguments can be there for that as well.
Misspelling like മീററ്, ലാററെററ് etc comes to my mind.

In all these cases, exclusion rules would be the least preferred choice.

13. Appendix D: NBGP Cross-script Variant Inclusion Policy

If, in any two given scripts, all the potential cross-script variants consist of dependent (e.g. Vowel Signs, Anusvara, Visarga, Chandrabindu etc.) characters **ONLY**, then that entire set can be ignored and no cross-script variants be proposed between those two scripts.

If, in any two given scripts, there is **AT LEAST ONE** non-dependent (e.g. Consonant, Vowel etc.) cross-script variant character/sequence present, all the potential cross-script variants be considered and proposed between the two scripts.

This cross-script analysis has been restricted to the scripts that have descended from the Brahmi as most of them share similar usage patterns. By and large, all of these scripts have a common set of characters that existed in Brahmi script and bear the same identities. However, as the scripts branched out from the Brahmi, depending on various factors, the shapes of the characters changed. This change in the shape was not uniform across all the characters and the scripts. Some characters shapes did change significantly whereas some of them still retained similarity. The cross-script similarity analysis also aims to identify such cases where the same character retained almost the same shape despite being part of the different scripts. These set of characters are variants of each other in true sense than merely of co-incidental visual similarity.

Since, having such labels is a realistic possibility and the corresponding labels look almost exactly alike, NBGP has proposed them as blocked variants.

NBGP acknowledges the concern that this shape is quite generic and may have parallels in other scripts not under its ambit. However, as NBGP does not have any exposure about actual usage of those characters in those particular scripts, NBGP desisted from including them in the analysis. As NBGP has already considered all the related scripts under the cross-script variant analysis, the similarity of the characters belonging to NBGP scripts with other scripts not under the NBGP ambit, may be of a mere co-incidental visual nature.

Additionally, this concern is not limited to these two characters but for all the characters in all the scripts under the scope of the Root LGR procedure. Carrying out this analysis can practically be done only with the Generation Panels that exist while the NBGP is active. This still leaves out those scripts out of the scope which may not have a Generation Panel established yet. Hence, carrying out this exercise in entirety is quite impracticable. This conundrum can be resolved if all the such cases are handled by the "String Similarity Assessment Panel" of ICANN.