

Proposal for a Tamil Script Root Zone Label Generation Rule-Set (LGR)

LGR Version: 3.0

Date: 24TH MAY 2018

Document version: 2.0

Authors: Neo-Brahmi Generation Panel [NBGP]

1 General Information/ Overview/ Abstract

This document lays down the Label Generation Rule Set for the Tamil script. The three main components of the Tamil Script LGR, Code point repertoire, Variants, and Whole Label Evaluation Rules have been described in detail here. These components have been incorporated in a machine-readable format in the accompanying XML file named "Proposed-LGR-Tamil-20180524.xml".

In addition, a document named "Tamil_Test_Labels_20180524.txt" has been provided. It provides a list of valid and invalid labels as per the Whole Label Evaluation laid down in Section 7 of this document. In addition, set of labels which can produce variants as laid down in Section 6 of this document. The labels have been tagged as valid and invalid under the specific rules¹.

2 Script for which the LGR is proposed

ISO 15924 Code: Tamil

ISO 15924 Key N°: 346

ISO 15924 English Name: Tamil

Latin transliteration of native script name: **tamil**

Native name of the script: தமிழ்

Maximal Starting Repertoire [MSR] version: 2

¹ The categorization of invalid labels under specific rules is given as per the general understanding of the LGR Tool by the NBGP. During testing with LGR tool, whether a particular label gets flagged under the same rule or the different one is totally dependent on the internal implementation of the LGR Tool. In case of discrepancy among the same, the fact that it is an invalid label should only be considered.

3 Background on Script and Principal Languages Using It

Tamil is one of the oldest Dravidian languages which has a continuous history since the age of tholkəppiyəm. The earliest known inscriptions in Tamil date back to 2,200 BC. Tamil literature emerged in around 300 BC, and the language used from then until the 700 AD is known as Old Tamil. From 700-1600 AD the language is known as Middle Tamil, and since 1600 the language has been known as Modern Tamil. Tamil is mainly spoken in the southern part of India, known as Tamilnadu. It is also spoken in Pondyicherry, Andaman & Nicobar island and other states of India. It is one the official languages of Sri Lanka and Singapore. Tamil speaking community is found in countries such as Malaysia, Mauritius, South Africa, UK, Canada, the USA, France and Réunion.

3.1 The Evolution of the Script

Tamil was originally written with a version of the Brahmi script known as Tamil Brahmi, and from 3rd century to 10th century AD this script had become more rounded and developed into the *vaṭṭeluttu* [1004] script. Over time the script has changed somewhat, and it was simplified in the 19th and 20th centuries. The image below shows how *vaṭṭeluttu* got transformed as Tamil letters²

² <https://ta.wikipedia.org/s/jt1>

வட்டெழுத்தாகவும் தமிழ் எழுத்தாகவும் மாற்றம் பெற்றதை விளக்கும் படம்

வட்டெழுத்தாக வளர்ந்த விதம்						தமிழ் எழுத்து	தமிழாக வளர்ந்த விதம்					
கீ.ம	கீ.ம	கீ.ம	கீ.ம	கீ.ம	கீ.ம	தமிழ் எழுத்து	கீ.ம	கீ.ம	கீ.ம	கீ.ம	கீ.ம	கீ.ம
17-தா.6	15-தா.6	14-தா.6	13-தா.6	9-தா.6	7-தா.6	3-தா.6	3-தா.6	7-தா.6	9-தா.6	11-தா.6	16-தா.6	20-தா.6
+	+	+	+	+	+	+	+	+	+	+	+	+
உ	உ	உ	உ	உ	உ	உ	உ	உ	உ	உ	உ	உ
ச	ச	ச	ச	ச	ச	ச	ச	ச	ச	ச	ச	ச
ஓ	ஓ	ஓ	ஓ	ஓ	ஓ	ஓ	ஓ	ஓ	ஓ	ஓ	ஓ	ஓ
ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ
஋	஋	஋	஋	஋	஋	஋	஋	஋	஋	஋	஋	஋
஌	஌	஌	஌	஌	஌	஌	஌	஌	஌	஌	஌	஌
஍	஍	஍	஍	஍	஍	஍	஍	஍	஍	஍	஍	஍
எ	எ	எ	எ	எ	எ	எ	எ	எ	எ	எ	எ	எ
ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ
உ	உ	உ	உ	உ	உ	உ	உ	உ	உ	உ	உ	உ
ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ	ஊ
஋	஋	஋	஋	஋	஋	஋	஋	஋	஋	஋	஋	஋
஌	஌	஌	஌	஌	஌	஌	஌	஌	஌	஌	஌	஌
஍	஍	஍	஍	஍	஍	஍	஍	஍	஍	஍	஍	஍
எ	எ	எ	எ	எ	எ	எ	எ	எ	எ	எ	எ	எ
ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ	ஏ
உ	உ	உ	உ	உ	உ	உ	உ	உ	உ	உ	உ	உ

Figure 1: vaṭṭeluttu to Tamil letters transformation

The central column of the above image indicates (oldest) Tamil Brahmi characters, diverging to vaṭṭeluttu towards left, and to Tamil towards the right. Tamil is also written with a version of the Arabic script known as [Arwi](#) by Tamil-speaking Muslims.

3.2 Languages considered

The Tamil script is mainly used to write the Tamil Language. However there are some tribal languages such as Badaga, Iruka, Kurumba Betta, Kurumba Kannada, Paniya, and Saurashtra which also use the Tamil Script; but since the EGIDS [EGIDS] value of those languages is above four they have not been considered in the present analysis.

EGIDS Scale 1	EGIDS Scale 2	EGIDS Scale 3	EGIDS Scale 4
Tamil (Sri Lanka)	Tamil (India)		Tamil (Singapore, Malaysia)

Table 1: Languages considered under Devanagari LGR

3.3 The structure of written Tamil

The Tamil script is an alphasyllabary and the heart of the writing system is the akshar. It is this unit, which is instinctively recognized by users of the script. To understand the notion of akshar, a brief overview of the writing system is provided in this Section and the akshar itself will be treated in depth in Section 5.4.

The writing system of Tamil could be summed up as composed of the following:

3.3.1 The Consonants

As per traditional grammar classification, Tamil consonants have been categorized in 3 groups according to their phonetic properties (especially in terms of place and manner of articulation with voiced and voiceless nature). They are Stops (vəllməm), Medial (idəyməm) and Nasal (mellməm). It should also be noted that as per Tamil traditional grammar, "Tamil Consonant" is ideally a combination of consonants (as defined in Unicode) + Virama combination. e.g க் (*TAMIL LETTER KA + TAMIL SIGN VIRAMA*) is actually a consonant in Tamil grammar. On the other hand, what Unicode designates as consonant is termed as Vowel-Consonant in Tamil Traditional grammar. However, for the sake of uniformity across all the LGRs under NBGP the Unicode naming convention has been followed.

The Unicode Consonant set of Tamil comprises the following characters:

STOP	க TAMIL LETTER KA (U+0B95)	ச TAMIL LETTER CA (U+0B9A)	ட TAMIL LETTER TTA (U+0B9F)	த TAMIL LETTER TA (U+0BA4)	ப TAMIL LETTER PA (U+0BAA)	ற TAMIL LETTER RRA (U+0BB1)
MEDIAL	ங் TAMIL LETTER NGA (U+0B99)	ஞ் TAMIL LETTER NYA (U+0B9E)	ண் TAMIL LETTER NNA (U+0BA3)	ந் TAMIL LETTER NA (U+0BA8)	ம் TAMIL LETTER MA (U+0BAE)	ன் TAMIL LETTER NNA (U+0BA9)
NASAL	ய் TAMIL LETTER YA (U+0BAF)	ர் TAMIL LETTER RA (U+0BB0)	ல் TAMIL LETTER LA (U+0BB2)	வ் TAMIL LETTER VA (U+0BB5)	ழ் TAMIL LETTER LLA (U+0BB4)	ள் TAMIL LETTER LLA (U+0BB3)

GRANTH A	ஸ TAMIL LETTER SA (U+0BB8)	ஷ TAMIL LETTER SSA (U+0BB7)	ஐ TAMIL LETTER JA (U+0B9C)	ஹ TAMIL LETTER HA (U+0BB9)	ஸ TAMIL LETTER SHA (U+0BB6)	
-------------	--	---	-------------------------------------	-------------------------------------	---	--

Table 2: Group classification of consonants

The IPA of Tamil Consonants as follows:

	Bilabial	Lab-Dental	Dental	Alv	Post-Alv	Retroflex	Palatal	Velar	Uvu	Glottal
Plosive	p (ப) b (ப)		த (த) த் (த)			ட (ட) ட (ட)		க (க) க (க)		
Nasal	m (ம)		ந (ந)	ன (ன)		ண (ண)	ள (ள)	ங (ங) (ங)		
Tap/Flap				ர (ர)						
Trill				ர (ர)						
Fricative				ச (ச)						ஃ (ஃ)
Approx		வ (வ)				ழ (ழ)	ய (ய)			
Lat Approx				ல (ல)		ள (ள)				
Affricate							ஞ (ஞ) ஞ (ஞ)			

Table 3: IPA classification of Tamil consonants

3.3.2 Virama3/Pulli

All consonants contain an implicit vowel (a) within them. A special sign is needed to denote that this implicit vowel is stripped off. This is known as the virama "◌்" (U+0BCD). The virama thus joins two adjacent consonants. In Tamil, unlike other scripts under Neo-Brahmi GP, there are only three instances where this results in the formation of conjunct. Example 1 shows the same and Example 2 shows the non-formation of conjunct.

Example 1

க+ஷ TAMIL LETTER KA TAMIL SIGN VIRAMA+ ஃஷ

³ Unicode (cf. Unicode 3.0 and above) prefers the term Virama. In this report both the terms have been used to denote the character that suppresses the inherent vowel.

TAMIL LETTER SSA

ஸ்+ரீ	TAMIL LETTER SA TAMIL SIGN VIRAMA+	ஸ்ரீ
	TAMIL LETTER RA TAMIL VOWEL SIGN II	
ஸ்+ரீ	TAMIL LETTER SHA TAMIL SIGN VIRAMA+	ஸ்ரீ
	TAMIL LETTER RA TAMIL VOWEL SIGN II	

Example 2

க்+க	TAMIL LETTER KA TAMIL SIGN	க்க
	VIRAMA+ TAMIL LETTER KA	

3.3.3 Vowels

Separate symbols exist for all Vowels, which are pronounced independently either at the beginning or after a vowel sound. To indicate a Vowel sound other than the implicit one, a Vowel sign (Matra) is attached to the consonant. Since the consonant has a built in 'a', there are equivalent Matras for all vowels excepting the அ (VOWEL LETTER A).

The correlation is shown:

Vowel	Corresponding vowel sign (Matra)
அ U+0B85	
ஆ U+0B86	ா U+0BBE
இ U+0B87	ி U+0BBF
ஈ U+0B88	ீ U+0BC0
உ U+0B89	ு U+0BC1

ஊ U+0B8A	ஃ U+0BC2
எ U+0B8E	ெ U+0BC6
ஏ U+0B8F	ே U+0BC7
ஐ U+0B90	ை U+0BC8
ஓ U+0B92	ொ U+0BCA
ஔ U+0B93	ோ U+0BCB
ஔள U+0B94	ௌ U+0BCC

Table 4: Vowels with corresponding Matras

3.3.4 Visarga / Aytham (ஃ - U+ 0B83)

The Visarga is also used in Tamil and represents a sound very close to /k/. அஃறிணை /akṛiṇai/ Non-human (U+0B85 U+0B83 U+0BB1 U+0BBF U+0BA3 U+0BC8).

A Visarga must always be followed by a stop consonants and few from grantha consonants list.

4 Overall Development Process and Methodology

Under the Neo-Brahmi Generation Panel, there are many different scripts belonging to separate Unicode blocks. Each of these scripts will be assigned a separate LGR; however, the 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 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 or archival purposes only will not be considered for inclusion in the code point repertoire.

4.1.1.2 *Unambiguous use:*

Every character proposed should have an unambiguous understanding among the linguistic community about its usage in the language. However MSR has already restricted these characters.

4.1.2 Exclusion principles:

The main exclusion principle is that of External Limits on Scope. These comprise of protocols or standards which are pre-requisites 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, up the ladder in the protocol hierarchies, the canvas of available characters for selection as a part of the Root Zone code point repertoire is already constrained by various protocol layers beneath it. 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 Unicode consortium.

ii. IDNA Protocol:

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

Example: Tamil Number Ten "ௐ" (U+0BF0) is not allowed to be a part of domain name.

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.

Examples: TAMIL OM "ஐ" (U+0BD0) and TAMIL SIGN ANUSVARA (U+0B82), even if allowed by IDNA protocol, are not permitted in the Root Zone Repertoire as per the [MSR].

To sum up, the restrictions start off with admitting only such characters as are part of the code-block of the given script/language. This is further narrowed down by the IDNA 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 Punctuation Marks:

The TLDs being identifiers, punctuation markers present in Brahmi based languages such as Danda "।" (U+0964) and double Danda "॥" (U+0965) will not be included.

4.1.2.3 No Symbols and Abbreviations:

Abbreviations, weights and measures and other such characters like Tamil Debit Sign "ஹ" (U+0BF6) etc. will not be included.

4.1.2.4 No Rare and Obsolete Characters:

AU LENGTH MARK "◌ᳵ" (U+0BD7) is a character in Tamil which has been added to the Unicode and is very rarely used in Modern Tamil. As it is very rarely used by the language community the same character will not be included in the proposed repertoire. This is in compliance with the Conservatism principle as laid down in the Root Zone LGR procedure.

4.1.2.5 No Stress Markers of Classical Sanskrit and Vedic:

Stress markers for classical Sanskrit e.g. DEVANAGARI STRESS SIGN UDATTA "◌̄" (U+0951) and DEVANAGARI STRESS SIGN ANUDATTA "◌̎" (U+0952) will not be included. Since Tamil has no stress, there are no such cases found in Tamil. This is also in compliance with the Letter principle as laid down in the Root Zone LGR procedure.

5 Repertoire

Section 5.1 shows the section of the [MSR] applicable to the Tamil script on which the Tamil code-point repertoire is based.

Section 5.2 details the code-point repertoire that the Neo-Brahmi Generation Panel [NBGP] proposes to be included in the Tamil LGR.

5.1 Tamil section of Maximal Starting Repertoire [MSR] Version 2

	0B8	0B9	0BA	0BB	0BC	0BD	0BE	0BF
0		ஐ		ர	ீ	ஓ		ய
1				ற	ி			ள
2	ஃ	ஔ		ல	ஔ			஑
3	ஐ	ஓ	ண	ள				உ
4		ஔ	த	ழ				ம்
5	அ	க		வ				ஶ
6	ஆ			ஸ	ெ		ஃ	யு
7	ஐ			ஷ	ே	ௌ	க	ௌ
8	ஈ		ந	ஸ	ை		உ	ஶ
9	உ	ங	ன	ஹ			ந	ஶ
A	ஊ	ச	ப		ொ		ச	நீ
B					ோ		ஶ	
C		ஐ			ௌ		ஶ	
D					ஃ		எ	
E	எ	ஶ	ம	ா			அ	
F	ஏ	ஶ	ய	ி			ஶ	

Color convention⁴:

All characters that are included in the [MSR]

- Yellow background

PVALID in IDNA2008 but excluded from the

[MSR] - Pinkish background

Not PVALID in IDNA2008, or are ineligible for the root zone (digits, hyphen) - White background

Figure 2: Tamil Code Page from [MSR]

5.2 Code Point Repertoire:

For each of the code points, language references have been given in the last column titled "Reference". The examples chosen for referencing, together cover all the code-points required for Tamil Language that NBGP has considered as given in 3.2.

Sr. No.	Unicode Code Point	Glyph	Character Name	Unicode General Category (gc)	Indic syllabic category	Example language(s) using the code-point (Not exhaustive list)	Language with lowest EGIDS scale using the code point	Reference
1	0B83	ஃ	TAMIL SIGN VISARGA	Lo	Visarga	Tamil	Tamil	[1003]
2	0B85	அ	TAMIL LETTER A	Lo	Vowel	Tamil	Tamil	[1001]
3	0B86	ஆ	TAMIL LETTER AA	Lo	Vowel	Tamil	Tamil	[1001]
4	0B87	இ	TAMIL LETTER I	Lo	Vowel	Tamil	Tamil	[1001]
5	0B88	ஈ	TAMIL LETTER II	Lo	Vowel	Tamil	Tamil	[1001]
6	0B89	உ	TAMIL LETTER U	Lo	Vowel	Tamil	Tamil	[1001]
7	0B8A	ஊ	TAMIL LETTER UU	Lo	Vowel	Tamil	Tamil	[1001]
8	0B8E	எ	TAMIL LETTER E	Lo	Vowel	Tamil	Tamil	[1001]
9	0B8F	ஏ	TAMIL LETTER EE	Lo	Vowel	Tamil	Tamil	[1001]
10	0B90	ஐ	TAMIL LETTER AI	Lo	Vowel	Tamil	Tamil	[1001]

11	OB92	ஓ	TAMIL LETTER O	Lo	Vowel	Tamil	Tamil	[1001]
12	OB93	ஓ	TAMIL LETTER OO	Lo	Vowel	Tamil	Tamil	[1001]
13	OB94	ஓள	TAMIL LETTER AU	Lo	Vowel	Tamil	Tamil	[1001]
14	OB95	க	TAMIL LETTER KA	Lo	Consonant	Tamil	Tamil	[1002]
15	OB99	ங	TAMIL LETTER NGA	Lo	Consonant	Tamil	Tamil	[1002]
16	OB9A	ச	TAMIL LETTER CA	Lo	Consonant	Tamil	Tamil	[1002]
17	OB9C	ஐ	TAMIL LETTER JA	Lo	Consonant	Tamil	Tamil	[1002]
18	OB9E	ஞ	TAMIL LETTER NYA	Lo	Consonant	Tamil	Tamil	[1002]
19	OB9F	ட	TAMIL LETTER TTA	Lo	Consonant	Tamil	Tamil	[1002]
20	OBA3	ண	TAMIL LETTER NNA	Lo	Consonant	Tamil	Tamil	[1002]
21	OBA4	த	TAMIL LETTER TA	Lo	Consonant	Tamil	Tamil	[1002]
22	OBA8	ந	TAMIL LETTER NA	Lo	Consonant	Tamil	Tamil	[1002]
23	OBA9	ன	TAMIL LETTER NNNA	Lo	Consonant	Tamil	Tamil	[1002]

24	OBAA	ப	TAMIL LETTER PA	Lo	Consonant	Tamil	Tamil	[1002]
25	OBAB	ம	TAMIL LETTER MA	Lo	Consonant	Tamil	Tamil	[1002]
26	OBAB	ய	TAMIL LETTER YA	Lo	Consonant	Tamil	Tamil	[1002]
27	OBBO	ர	TAMIL LETTER RA	Lo	Consonant	Tamil	Tamil	[1002]
28	OBBI	ற	TAMIL LETTER RRA	Lo	Consonant	Tamil	Tamil	[1002]
29	OBBI	ல	TAMIL LETTER LA	Lo	Consonant	Tamil	Tamil	[1002]
30	OBBI	ள	TAMIL LETTER LLA	Lo	Consonant	Tamil	Tamil	[1002]
31	OBBI	ழ	TAMIL LETTER LLLA	Lo	Consonant	Tamil	Tamil	[1002]
32	OBBI	வ	TAMIL LETTER VA	Lo	Consonant	Tamil	Tamil	[1002]
33	OBBI	ஸ	TAMIL LETTER SHA	Lo	Consonant	Tamil	Tamil	[1002]
34	OBBI	ஷ	TAMIL LETTER SSA	Lo	Consonant	Tamil	Tamil	[1002]
35	OBBI	ச	TAMIL LETTER SA	Lo	Consonant	Tamil	Tamil	[1002]
36	OBBI	ஹ	TAMIL LETTER HA	Lo	Consonant	Tamil	Tamil	[1002]

37	OBBE	ஶ	TAMIL VOWEL SIGN AA	Mc	Matra	Tamil	Tamil	[1002]
38	OBBF	ஶி	TAMIL VOWEL SIGN I	Mc	Matra	Tamil	Tamil	[1002]
39	OBC0	ஶீ	TAMIL VOWEL SIGN II	Mn	Matra	Tamil	Tamil	[1002]
40	OBC1	ஶு	TAMIL VOWEL SIGN U	Mc	Matra	Tamil	Tamil	[1002]
41	OBC2	ஶூ	TAMIL VOWEL SIGN UU	Mc	Matra	Tamil	Tamil	[1002]
42	OBC6	ஶெ	TAMIL VOWEL SIGN E	Mc	Matra	Tamil	Tamil	[1002]
43	OBC7	ஶே	TAMIL VOWEL SIGN EE	Mc	Matra	Tamil	Tamil	[1002]
44	OBC8	ஶை	TAMIL VOWEL SIGN AI	Mc	Matra	Tamil	Tamil	[1002]
45	OBCA	ஶொ	TAMIL VOWEL SIGN O	Mc	Matra	Tamil	Tamil	[1002]
46	OBCB	ஶோ	TAMIL VOWEL SIGN OO	Mc	Matra	Tamil	Tamil	[1002]
47	OBCC	ஶொள	TAMIL VOWEL SIGN AU	Mc	Matra	Tamil	Tamil	[1002]

48	OBCD	◌̣	TAMIL SIGN VIRAMA	Mn	Matra	Tamil	Tamil	[1002]
----	------	----	-------------------------	----	-------	-------	-------	--------

Table 5: Code point repertoire

5.3 Code points not included:

The following code points have not been included in the repertoire.

Sr. No.	Unicode Code Point	Glyph	Character Name	Reason for exclusion
1.	U+0BD7	◌̣	TAMIL AU LENGTH MARK	Not in modern usage. Excluded as per conservatism principle.

Table 6: Code points not included

5.4 Structural Formation of Tamil:

All the languages written in any Brahmi-derived scripts follow a particular way of formation of their words, known as "akshar". In the next section, there are detailed akshar formation rules applicable to the representation of the "Tamil" language when written in the Tamil script.

5.5 Akshar formation rules for Tamil:

This section details the Akshar formation rules as applicable to Tamil. The first section lists the categories of the characters in the form of variables. In the rules, instead of their full descriptive names, abbreviated variable names are used. The second section lists four operators along with their functions which are assumed while specifying the rules. The final two sections describe the formation of the two major categories of Akshar; the first of which begins with the vowels and the second one with the consonants.

5.5.1 Variables involved

- Dash → Hyphen -
- Digit → Indo-Arabid digits [0-9]
- C → Consonant
- M → Matra
- V → Vowel

X → Visarga / Aytham
H → Virama / Pulli

5.5.2 Operators used:

Symbol	Function
	Alternative
[]	Optional
*	Variable Repetition
()	Sequence Group

Table 7: Symbol functions

In what follows, the Vowel Sequence and the Consonant Sequence pertinent to Tamil, when used to write Tamil, are given.

5.5.3 Vowel Sequence

A vowel sequence begins with a vowel. It may be optionally followed by a Visarga (X). The number of X which can follow a V in Tamil are restricted to one.

The vowel sequence in Tamil is therefore V [X]

Examples:

Sequence Description	Sequence	Example	Constituting characters
Vowel	V	அ /a/ U+0B85	
Vowel + Visarga	V[X]	அஃ /ak/ U+0B85 U+0B83	அஃ U+0B85 U+0B83

Table 8

5.5.4 Consonant Sequence

A consonant sequence begins with a consonant. It may be optionally followed by a Matra (M), Visarga (X) or a Virama/Pulli (H). The number of instances of these characters occurring after a consonant is restricted to one. There is a possibility of further extension of the Consonant sequence after the M and H. Each of these has been discussed in the following sections:

1. A single consonant (C)

Examples:

Sequence Description	Sequence	Example	Constituting characters
Consonant	C	க /ka/ U+0B95	<single character>

Table 9

2. A consonant optionally followed by dependent vowel sign/Matra [M], Visarga [X] or Virama/Pulli [H]

C [M| H|X]

Examples:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Matra	C[M]	கி /ki/	க ி 0B95 0BBF
Consonant + Virama/Pulli	C[H]	க் /k/ (Pure Consonant)	க ு U+0B95 U+0BCD
Consonant + Visarga	C[X]	கஃ /kk̄/	க ஃ U+0B95 U+0B83

Table 10

2. A. A CM sequence can be optionally followed by X

(CM)[X]

Example:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Matra + Visarga	CM[X]	முஃ /muk̄/	ம ஁ ஃ U+0BAE U+0BC1 U+0B83

Table 11

3. A sequence of consonants (up to 3) joined by Virama/Pulli *2(CH)C

Example:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Virama/Pulli + Consonant + Virama/Pulli +	CHCHC	ழ்த்த / litta /	ழ ு த ு த U+0BB4 U+0BCD U+0BA4 U+0BCD

Consonant			U+0BA4
-----------	--	--	--------

Table 12

Subsets:

3. A. The combination may be followed by M or X

Example:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Virama/Pulli + Consonant + Matra	CHC[M]	க்சு /kku/	க ு க ு U+0B95 U+0BCD U+0B95 U+0BC1
Consonant + Virama/Pulli + Consonant + Visarga	CHC[X]	க்சு /kkak/	க ு க ு U+0B95 U+0BCD U+0B95 U+0B83

Table 13

3. B. *3(CH)CM may be followed by an X

Example:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Virama/Pulli + Consonant + Matra + Visarga	CHCM[X]	க்சு /kkīḥ/	க ு க ு U+0BAE U+0BCD U+0BAE U+0BC1 U+0B83

Table 14

These are the basic akshar rules on which the overall Tamil LGR is based. There are some additional finer aspects to these rules as one takes into account the digits, punctuations and special standalone characters like Avagraha. Those aspects are not discussed here as the [MSR] on which the LGRs are supposed to be based, excludes those characters.

6 Variants

There are some characters/character sequences in Tamil that can be created by using the characters permitted as per the [MSR] and that look exactly alike. The NBGP categorizes these confusingly similar variants in three groups

- ◆ Group 1: Confusing due to being exact homoglyphs
- ◆ Group 2: Confusing due to partial similarity
- ◆ Group 3: Confusing due to similar appearance but actually not valid as per akshar formation rules

6.1 Group 1: Confusing due to being exact homoglyphs

Cases which belong to Group 1 are proposed to be considered as variants. There are three such cases.

6.1.1 TAMIL LETTER AU with TAMIL LETTER O followed by TAMIL LETTER LLA:

This variant pair involves the pure vowel TAMIL LETTER AU (ஓள U+0B94) which looks exactly similar to the vowel + Consonant TAMIL LETTER O + TAMIL LETTER LLA (ஓள U+0B92 U+0BB3) combination. These two cases can cause confusion even to a careful observer and hence are being proposed as variants.

Variant 1	Variant 2
ஓள U+0B94	ஓ ள U+0B92 U+0BB3

Table 15: Proposed Variants - Set 1

6.1.2 TAMIL VOWEL SIGN AU with TAMIL VOWEL SIGN E followed by TAMIL LETTER LLA:

This variant pair involves the split matra TAMIL VOWEL SIGN AU (ொௌ U+0BCC) having left and right side catenators which sit on the preceding consonant. It looks exactly alike to a combination of matra TAMIL VOWEL SIGN E (ொ U + 0BC6) followed by consonant TAMIL LETTER LLA (ள U+0BB3).

Variant 1	Variant 2
ொௌ U+0BCC	ொ ள U+0BC6 U+0BB3

Table 16: Proposed Variants - Set 2

However, it must be noted that the above variant pair needs a preceding consonant to make it a valid akshar formation.

6.1.3 Alternate representation for Shri

This variant pair involves forming “Shri” ligature by inputting two different consonants. Prior to Unicode 4.1, the best mapping to represent the ligature Shri was to the sequence ஶ்ரீ + றீ TAMIL LETTER SA TAMIL SIGN VIRAMA + TAMIL LETTER RA TAMIL VOWEL SIGN II <U+0BB8, U+0BCD+ U+0BB0, U+0BC0>. Unicode 4.1 in 2005 added the character U+0BB6 TAMIL LETTER SHA and as a consequence, the best mapping became TAMIL LETTER SHA TAMIL SIGN VIRAMA +TAMIL LETTER RA TAMIL VOWEL SIGN II <U+0BB6, U+0BCD, U+0BB0, U+0BC0>.

Due to slow updates to implementations, both representations are widespread in existing text. Therefore, in the present situation, Unicode recommends treating both representations as equivalent sequences. All the Tamil fonts which support both the combinations, represent both the sequences in exactly similar form (glyph).

Thus these representations should be treated as allocatable variants of each other as they don’t cause any semantic change of the labels and also the display of the labels would remain the same in both cases, they are being proposed as variants allocatable. Following is the brief description of these variants followed by variants in Table 15 , Table 16 and Table 17.

Variant 1	Variant 2
ஸ ு ர ு = ஸ்ரீ U+0BB6, U+0BCD, U+0BB0, U+0BC0	ஸ ு ர ு = ஸ்ரீ U+0BB8, U+0BCD, U+0BB0, U+0BC0

Table 17: Proposed Variants - Set 3

6.2 Group 2: Confusing due to partial similarity

This happens with the partial similarity of the characters appearance of TAMIL LETTER JA “ஜ” (U+0B9C) with TAMIL LETTER AI “ஐ” (U+0B9C). However, no cases belonging to Group 2 are proposed, as there is another panel (String similarity assessment panel) entrusted to deal with such cases.

Confusable 1	Confusable 2
ஜ (U+0B9C)	ஐ U+0B9C

Table 18: Not Proposed as Variants - Set 1

6.3 Group 3: Confusing due to similar looking but actually not valid as per akshar formation rules.

This happens with wrong formation of consonant followed by two continuous matras. The TAMIL VOWEL SIGN O “ஊ” (U+ 0BCA) looks exactly same as TAMIL VOWEL SIGN E “ஊ” (U+0BC6) followed by TAMIL VOWEL SIGN AA “ஊ” (U+0BBE). However as the formation is not valid as per akshar formation rules, this case is not proposed as variant.

Confusable 1	Confusable 2
ஊ (U+0BCA)	ஊஊ (U+0BC6) (U+0BBE).

Table 19: Not Proposed as Variants - Set 2

6.4 Cross script variants:

A cross-script variant, also sometimes referred to as "Whole Label confusable", is the variant case where one label in one script can be composed in such a way that it can resemble an entire label in a different script. Tamil script has a set of possible cross-script variants only with the Malayalam script. Table 21: Proposed Cross-script variants

lists the variants that are proposed as cross-script variants between Tamil and Malayalam. It is to be noted that none of the combinations listed in Table 21: Proposed Cross-script variants

are termed to be equivalents of each other semantically or otherwise. They are only grouped based on possible visual confusability. Here are some of examples of variant labels.

Tamil label	Malayalam label
<p>வமி U+0BB5 U+0BAE U+0BBF</p>	<p>ഖമി U+0D16 U+0D25 U+0D3F</p>
<p>ஜெமி U+0B9C U+0BC6 U+0BAE U+0BBF</p>	<p>ജെമി U+0D1C U+0D46 U+0D25 U+0D3F</p>

Table 20: Cross-script Variant examples

A label can be considered to have a cross-script variant label only if "all" the constituent characters/aksharas have an equivalent confusable in the other script. If there is even one single character/akshara which does not have an equivalent visual confusable in another script, it essentially provides a visual distinction and hence a non-confusable string.

The following table gives the set of proposed cross-script variants between Tamil and Malayalam.

Tamil	Malayalam
<p>ஜ U+0B9C</p>	<p>ജ U+0D1C</p>
<p>வ U+0BB5</p>	<p>ഖ U+0D16</p>

𑌞 U+0BAE	𑌞 U+0D25
𑌟 U+0BBF	𑌟 0D3F
𑌠 U+0BC6	𑌠 0D46
𑌡 U+0BC7	𑌡 0D47

Table 21: Proposed Cross-script variants

In addition to the above cases, Tamil and Malayalam scripts have a possible set of code points which look similar but not similar enough to be recommended as cross-script variants. They are listed in Table 21: Tamil Variants based on pure visual similarity, in Appendix A.

6.5 Variant Disposition:

6.5.1 Blocked variant

Variants mentioned in Table 15 and Table 16 are cases of homoglyphs and hence it is proposed that these be "blocked" variants.

There is no preference among these variants. Whichever label containing either of these variants is chosen earlier, the other one equivalent variant label should be "blocked".

6.5.2 Allocatable variants

The variant "Shri" described in section 6.1.3 is a case of variant where exactly same visual form is rendered with two distinct sequences. Also, in the minds of the user, regardless of which sequence they choose to input, both are intended to be the same akshar i.e. "Shri". Hence, it is imperative that both the sequences be treated as same in terms of Variant analysis and any label formed with either form should be made available to the same entity. This variant pair is thus being proposed as an "Allocatable" variant.

7 Whole Label Evaluation Rules (WLE)

This section provides the WLEs that are required by all the languages mentioned in section 3.2 when written in Devanagari Script. The rules have been drafted in such a way that they can be easily translated into the LGR specification.

Below are the symbols used in the WLE rules, for each of the "Indic Syllabic Category" as mentioned in the Table 5: Code point repertoire.

C	→	Consonant
M	→	Matra
V	→	Vowel
X	→	Visarga / Aytham
H	→	Virama / Pulli

Below are the specific WLE rules:

1. H: must be preceded by C
2. M: must be preceded by C
3. X: cannot be preceded by X
4. X: must only be succeeded by any of the following consonants,
 - **க** (TAMIL LETTER KA, U+0B95)
 - **ச** (TAMIL LETTER CA, U+0B9A)
 - **ஜ** (TAMIL LETTER JA, U+0B9C) (Used by Arwi community)
 - **ட** (TAMIL LETTER TTA, U+0B9F)
 - **த** (TAMIL LETTER TA, U+0BA4)
 - **ப** (TAMIL LETTER PA, U+0BAA)
 - **ற** (TAMIL LETTER RRA, U+0BB1)
 - **ஸ** (TAMIL LETTER SA, U+0BB8) (Used by Tamil community in Sri Lanka)

8 Contributors

NBGP Co-chairs: Dr. Uday Narayan Singh, Mr. Mahesh D Kulkarni and Dr. Ajay Data

Following is the full list of NBGP members with their Language expertise.

Name	Language Expertise
Udaya Narayana Singh	Bengali, Maithili, Hindi, English
Ajay Data	Hindi
Mahesh D. Kulkarni	Marathi, Hindi
Anupam Agrawal	Hindi, Bengali
Akshat S. Joshi	Hindi, Marathi
Abhijit Dutta	Bengali, Hindi
Neha Gupta	Hindi
Nishit Jain	Hindi
Prabhakar Pandey	Hindi
Raiomond Doctor	English, Hindi, Marathi, Gujarati
N. DeivaSundaram	Tamil
Shantaram S. Warde Walawalikar	Konkani
Bal Krishna Bal	Nepali
Ganesh Murmu	Santali
Balaram Prasain	Nepali
Rajib Chakraborty	Bangla (Bengali)
Gurpreet Singh Lehal	Panjabi
Saroja Bhate	Sanskrit
Shambhu Kumar Singh	Maithili
SwarnaPrabha Chainary	Bodo
Ghanashyam Nepal	Nepali

Kalyan Vasudeo Kale	Marathi
Shashi Pathania	Dogri
Santhosh Thottingal	Malayalam, Sourashtra, Tamil
Uma Maheshwar G	Telugu
Girish Chandra Mishra	Odia
K. C. Tikayat ray	Odia
Debajit Sharma	Assamese
Basanta Kumar Panda	Odia
Arvind Bhandari	Gujarati
Harish Chowdhary	Hindi
Chitrita Chatterjee	Multiple languages represented by members of IAMAI
U.B. Pavanaja	Kannada
Hempal Shrestha	Nepali, Newari
Suraj Adhikari	Nepali
Gangadhar Panday	Telugu
Vinay Murarka	Hindi
Mukesh Saini	Hindi
Jay Paudyal	Hindi
Pawan Chitrakar	Nepali
Nirajan Parajuli	Nepali
Uttam Shrestha Rana	Nepali
Dev Dass Manandhar	Nepali, Newari
Bhim Dhoj Shrestha	Nepali, Newari
Rajiv Kumar	Hindi
Shubham Saran	Hindi

Anivar A. Aravind	Malayalam
Shanmugam R	Tamil
Prasad PK	Malayalam
Cinnathambi Shanmugaraja	Tamil
Sarweshwaran	Tamil

In addition, following members externally gave inputs to NBGP for the respective languages/scripts.

Name	Language/Script Expertise
Ajit Kumar	Awadhi, Braj Language
Basil Baa	Sadri Language
Basil Kiro	Kharia Language
Biswa Limbu	Limbu Language
Devendra Kumar Devesh	Bhojpuri Language
Dinbandhu Mahto	Panchpargania Language
Dr. Birendra Kumar Soy	Mundari Language
Dr. Dinesh Kumar Shrivastav	Magahi Language
Dr. Harvinder Kaur	Gurmukhi Script
Dr. Laxmi Prasad Khatiwada	Nepali Language
Jagannath Singh	Panchpargania Language
Narendra Kumar Negi	Kinnauri Language
Prateek Harshwal	Wagdi and Dhundhari Language

Rayem Olem Dungdung	Sadri Language
Tej Man Angdembe	Limbu Language

Full Updated list of NBGP members is available at :

<https://community.icann.org/display/croscomlgrprocedure/Neo-Brahmi+GP>

9 References

[MSR] Integration Panel, "Maximal Starting Repertoire — MSR-2 Overview and Rationale",
14 April 2015 <https://www.icann.org/en/system/files/files/msr-2-overview-14apr15-en.pdf>

[EGIDS] Expanded Graded Intergenerational Disruption Scale,
<https://www.ethnologue.com/about/language-status> (Accessed on 13th Nov. 2017)

[NBGP] Neo-Brahmi Generation Panel

[gTLD] generic Top Level Domain

[1001] Omniglot, "Tamil", <https://www.omniglot.com/writing/Tamil.htm> (Accessed on 21th Nov. 2017)

[1002] Unicode 10.0.0, "South and Central Asia-I, Page 488-493 (R5 and R5a) ",
<http://www.unicode.org/versions/Unicode10.0.0/ch12.pdf> (Accessed on 21th Nov. 2017)

[1003] "Tamil Paper Website ", <http://www.tamilpaper.net/?p=7931> (Accessed on 27th Nov. 2017)

[1004] Tamil <https://ta.wikipedia.org/s/jt1> (Accessed on 28th Nov. 2018)

10 Books, articles and webographies consulted

Following is a thematically sorted set of documents, books, articles and webographies consulted in the drafting of this report

1. Karunakaran K [2000], Simplified grammar of Tamil. Suvitha Publishers.
2. Kothandaraman Pon [1997], A Grammar of contemporary Literary Tamil. International Institute of Tamil Studies.
3. Kothandaraman Pon [2001], Tamil studies. Ambuli publications
4. Meenakshi Sundaranar Te.Po [1965], A History of Tamil Literature. Annamalai University
5. Vaiyapuripillai [1988], Vaiyapuripillai's History of Tamil language and literature. New Century Book House
6. Varadharajan Mu. [1988], History of Tamil Literature. Sahitya Akademi.
7. Tamil Script Evolution <http://www.virtualvinodh.com/wp/tamil-script-evolution/> (Accessed on 28th Nov. 2018)

11 Appendix A: Cross-script Confusables

As discussed earlier, Tamil script has a major set of possible cross-script confusables with the Malayalam script. The Table 20 lists them.

It is to be noted that none of the combinations listed in Table 20 are termed to be equivalents of each other semantically or otherwise. They are only grouped based on possible visual confusability.

At first they may not look exactly the same, however, in the given context e.g. in browser bar as a part of a domain name, or as a single word where there is no surrounding text from the same script for distinguishing, they can create visual confusion.

A label can be considered to have a cross-script confusables label only if "all" the constituent characters/aksharas have an equivalent confusable in the other script. If there is even one single character/akshara which does not have an equivalent visual confusable in other script, it essentially provides a visually distinguishability and hence a non-confusable string.

Tamil	Malayalam
ஸ U+0BB8	സ U+0D38
ய U+0BAF	ഡ U+0D27
க U+0B95	ക U+0D15

Table 22: Tamil Variants based on pure visual similarity

12 Appendix B: A NOTE ON ZERO WIDTH NON-JOINER

This note is pertinent to the use of Zero Width Non Joiner (ZWNJ) as used in Tamil. ZWJ (U+0200D) and ZWNJ (U+0200C) are code points that have been provided by the Unicode standard to instruct the rendering of a string where the script has the option between joining and non-joining characters. Without the use of these control codes, the string may be rendered in an alternate form from what is intended.

In the case of Tamil, ZWJ does not play an important role insofar as functionality is concerned. But ZWNJ plays a role in the following combinations for example : ஸ்ரீ/srī/(U+092C U+0922 U+093C) ,க்ஷ /kshə/(U+0B95 U+0BCD U+0BB7 U+0BAF) The word " அக்ஷய்" /əkshəy/(U+0B85 U+0B95 U+0BCD U+200C U+0BB7 U+0BAF U+0BCD) can be written with the Unicode values .

U+0B85 U+0B95 U+0BCD **U+200C** U+0BB7 U+0BAF U+0BCD (அக்ஷய் with ZWNJ)

as well as

U+0B85 U+0B95 U+0BCD U+0BB7 U+0BAF U+0BCD (அகஷ்ய without ZWNJ)

Insofar as Tamil is concerned ZWNJ is used to render alternate rendering of ligatures. The use of ZWNJ in Tamil is restricted to representing a dead consonant within a string. Thus to show the combination of க்+ஷ /k+shə/ (U+0B95 U+0BCD U+0BB7) as a single word and retain the shape of the consonant followed by the halant; ZWNJ is used. This practice is followed to represent Sanskrit loan words or proper names demanding a “dead” consonant. As ZWNJ is not part of the MSR, representing the above words in the specific forms would not be possible.