

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

LGR Version: 3.0

Date: 08 Jan. 2018

Document version: 1.6

Authors: Neo-Brahmi Generation Panel [NBGP]

1 General Information/ Overview/ Abstract

This document lays down the Label Generation Rule Set for Tamil script. Three main components of the Tamil Script LGR i.e. Code point repertoire, Variants and Whole Label Evaluation Rules have been described in detail here. All these components have been incorporated in a machine-readable format in the accompanying XML file named "**Proposed-LGR-Tamil-20180108.xml**". (Not yet completed)

2 Script for which the LGR is proposed

ISO 15924 Code: Taml

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

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 tholkappiyam. 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 has a wide collections of literature which documented the ancient life style of Tamil people through Inscriptions and Manuscripts. Sangam literature of Tamil is well-known for its contents on humanity, braveness, love and other disciplines of Tamil. Tamil is mainly spoken in the southern part of India, known as Tamilnadu. Tamil has also been identified as a classical language by the government of India. It is also spoken in other parts of India such as Pondicherry, Andaman & Nicobar Island and other states of India. It is one the official languages in Sri Lanka, Singapore. The Tamil speaking communities are found in the other countries such as Malaysia, Mauritius, South Africa, UK, Canada, the USA, France and Reunion.

3.1 The Evolution of the Script

Tamil was originally written with a version of the Brahmi script known as Tamil Brahmi, and from 3-rd century to 10-th century AD this script had become more rounded and developed into the *vaṭṭeluttu* script. Over time the script got changed somewhat, and it was simplified in the 19th and 20th centuries through some reformations. The below image shows how *vaṭṭeluttu* got transformed as Tamil letters¹

¹ <https://ta.wikipedia.org/s/jt1>

வட்டெழுத்தாகவும் தமிழ் எழுத்தாகவும் மாற்றம் பெற்றதை விளக்கும் படம்

வட்டெழுத்தாக வளர்ந்த விகிதம்						தமிழ் எழுத்து	தமிழாக வளர்ந்த விகிதம்					
கீ.பி	கீ.பி	கீ.பி	கீ.பி	கீ.பி	கீ.பி	கீ.பி	கீ.பி	கீ.பி	கீ.பி	கீ.பி	கீ.பி	
17.தூ.6	15.தூ.6	14.தூ.6	13.தூ.6	9.தூ.6	7.தூ.6	3.தூ.6	3.தூ.6	7.தூ.6	9.தூ.6	11.தூ.6	16.தூ.6	20.தூ.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
ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ
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
ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ	ஐ

Figure 1: vatteḷuttu to Tamil letters transformation

Tamil is also written with a version of the Arabic script known as [Arwi](#) by Tamil-speaking Muslim community.

3.2 Languages considered

Tamil script is being used to represent only Tamil Language. It falls under [EGIDS] scale 1 in Sri Lanka where it has a national status. It falls under [EGIDS] 2 In Indian state of Tamilnadu and two Union territories of India namely Puducherry and Andaman & Nicobar Islands where it has a provincial status. Tamil is also listed under [EGIDS] 4 with Educational status in Singapore and Malaysia. In addition to this, there are some tribal languages in India such as Badaga, Irula, Kurumba Betta, Kurumba Kannada, Paniya, and Saurashtra which use Tamil script but the [EGIDS] of those languages are above four hence they have not been considered in the current analysis.

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

Table 1: Languages considered under Tamil LGR

3.3 The structure of written Tamil

Tamil 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.

Tamil is somehow different from other scripts which are derived from the same Brahmi scripts in a number of ways. It does not have graphemes for voiced or aspirated stop consonants. Voiced consonants are represented through allophonic variations of the same phoneme and a linguistic context helps to distinguish the variations of the phonemes. Thus the character க் k, for example, represents /k/ but can also be pronounced as /g/ or /x/ based on the contextual rules of Tamil grammar. A separate set of characters appears for these sounds when the Tamil script is used to write Sanskrit or other languages.

Also unlike other Brahmi scripts, the Tamil script rarely uses typographic ligatures to represent conjunct consonants, which are far less frequent in Tamil than in other Indian languages. Conjunct consonants are written by writing the character for the first consonant, adding the pulli to suppress its inherent vowel, and then writing the character for the second consonant. There are a few exceptions, namely க்ஷ kṣa and ஸ்ரீ śrī.

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

3.3.1 The Consonants

As per traditional classification Tamil consonants have been categorized as 3 groups according to their phonetic properties (especially in terms of place and manner of articulation with voiced and voiceless nature). They are Stops (vəllinəm) , Medial (iḍəiyinəm) and Nasal(məllinəm)

The Consonant set of Tamil comprises the following characters:

STOP	க U+0B95	ச U+0B9A	ட U+0B9F	த U+0BA4	ப U+0BAA	ற U+0BB1
MEDIAL	ங U+0B99	ஞ U+0B9E	ண U+0BA3	ந U+0BA8	ம U+0BAE	ள U+0BA9
NASAL	ய U+0BAF	ர U+0BB0	ல U+0BB2	வ U+0BB5	ழ U+0BB4	ள U+0BB3
GRANTHA	ஸ U+0BB8	ஷ U+0BB7	ஐ U+0B9C	ஹ U+0BB9	ஸ 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) <u>b</u> (ப)		t (த) d (த)			ɽ (ட) ɽ (ட)		k (க) g (க)		
Nasal	<u>m</u> (ம)		<u>n</u> (ந)	<u>ɳ</u> (ண)		<u>ɳ</u> (ண)	<u>ɲ</u> (ஞ)	<u>ŋ</u> (ங)		
Tap/Flap				<u>ɾ</u> (ர)						
Trill				<u>r</u> (ற)						
Fricative				<u>s</u> (ச)						<u>h</u> (ஃ)
Approx		<u>ʋ</u> (வ)				<u>ɻ</u> (ழ)	<u>j</u> (ய)			
Lat Approx				<u>l</u> (ல)		<u>ɭ</u> (ள)				
Affricate							<u>tʃ</u> (ச) <u>dʒ</u> (ஜ)			

Table 3: IPA classification of Tamil consonants

3.3.2 The Implicit Vowel Killer: Halant2/Pulli

All consonants have 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 Halant "◌̣" (U+0BCD). The

² 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.

Halant in general joins two consonants and creates conjuncts, However Tamil has only two conjuncts i.e. கூர் (U+0B95 U+0BCD U+0BB7), ஸ்ரீ (U+0BB8 U+0BCD U+0BB0 U+0BC0). In Tamil Halanth/ Pulli majorly is used in stripping the vowel from the consonant.

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.

The correlation is shown as under:

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	ூ 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	ொள 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 /ak̐riṇai/ Non-human (U+0B85 U+0B83 U+0BB1 U+0BBF U+0BA3 U+0BC8).

The condition to use Visarga is it should be always followed by a stop consonant.

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 Neo-Brahmi GP will ensure 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 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. The characters which have been encoded in the 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 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 Acknowledgement to Environmental Limitations. 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 Acknowledgement to Environment Limitations:

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, 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.

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.

Example: TAMIL OM "ॐ" (U+0BD0) 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 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 "५" (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 consonance 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 consonance with the Letter principle as laid down in the Root Zone LGR procedure.

5 Repertoire

Section 5.1 provides 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	ழ			த
3	ஶ	ஷ	ண	ள				உ
4		ஔ	த	ழ				ம்
5	அ	க		வ				ஶ
6	ஆ			ஸ	ஃ		ஃ	பு
7	இ			ஷ	ஃ	ள	க	ஶ
8	ஈ		ந	ஸ	ஃ		உ	ஶ
9	உ	ங	ன	ஶ			ங	ஶ
A	ஊ	ச	ப		ஃ		ச	ஶ
B					ஃ		ஶ	
C		ஐ			ஃ		க	
D					ஃ		எ	
E	எ	ஶ	ம	ஃ			அ	
F	ஏ	ஶ	ய	ி			க	

Color convention³:

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]

3

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

5.2 Code Point Repertoire:

For each of the code points, language references have been given in the last column titled "Reference". For the entire coverage of Tamil code points, references of the same have been given. The examples have been chosen for referencing, they 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	<i>in current and widespread use ? [Yes/No]</i>
1	0B83	ஃ	TAMIL SIGN VISARGA	Lo	Visarga	Tamil	Tamil	[1003]	Yes
2	0B85	அ	TAMIL LETTER A	Lo	Vowel	Tamil	Tamil	[1001]	Yes
3	0B86	ஆ	TAMIL LETTER AA	Lo	Vowel	Tamil	Tamil	[1001]	Yes
4	0B87	இ	TAMIL LETTER I	Lo	Vowel	Tamil	Tamil	[1001]	Yes
5	0B88	ஈ	TAMIL LETTER II	Lo	Vowel	Tamil	Tamil	[1001]	Yes
6	0B89	உ	TAMIL LETTER U	Lo	Vowel	Tamil	Tamil	[1001]	Yes
7	0B8A	ஊ	TAMIL LETTER UU	Lo	Vowel	Tamil	Tamil	[1001]	Yes
8	0B8E	எ	TAMIL LETTER E	Lo	Vowel	Tamil	Tamil	[1001]	Yes

9	0B8F	ஏ	TAMIL LETTER EE	Lo	Vowel	Tamil	Tamil	[1001]	Yes
10	0B90	ஐ	TAMIL LETTER AI	Lo	Vowel	Tamil	Tamil	[1001]	Yes
11	0B92	ஓ	TAMIL LETTER O	Lo	Vowel	Tamil	Tamil	[1001]	Yes
12	0B93	ஔ	TAMIL LETTER OO	Lo	Vowel	Tamil	Tamil	[1001]	Yes
13	0B94	ஔள	TAMIL LETTER AU	Lo	Vowel	Tamil	Tamil	[1001]	Yes
14	0B95	க	TAMIL LETTER KA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
15	0B99	ங	TAMIL LETTER NGA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
16	0B9A	ச	TAMIL LETTER CA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
17	0B9C	ஐ	TAMIL LETTER JA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
18	0B9E	ஞ	TAMIL LETTER NYA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
19	0B9F	ட	TAMIL LETTER TTA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
20	0BA3	ண	TAMIL LETTER NNA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
21	0BA4	த	TAMIL LETTER TA	Lo	Consonant	Tamil	Tamil	[1002]	Yes

22	0BA8	ந	TAMIL LETTER NA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
23	0BA9	ள	TAMIL LETTER NNA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
24	0BAA	ப	TAMIL LETTER PA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
25	0BAE	ம	TAMIL LETTER MA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
26	0BAF	ய	TAMIL LETTER YA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
27	0BB0	ர	TAMIL LETTER RA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
28	0BB1	ற	TAMIL LETTER RRA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
29	0BB2	ல	TAMIL LETTER LA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
30	0BB3	ள	TAMIL LETTER LLA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
31	0BB4	ழ	TAMIL LETTER LLA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
32	0BB5	வ	TAMIL LETTER VA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
33	0BB6	ஷ	TAMIL LETTER SHA	Lo	Consonant	Tamil	Tamil	[1002]	Yes

34	0BB7	ஷ	TAMIL LETTER SSA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
35	0BB8	ஸ	TAMIL LETTER SA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
36	0BB9	ஹ	TAMIL LETTER HA	Lo	Consonant	Tamil	Tamil	[1002]	Yes
37	0BBE	ா	TAMIL VOWEL SIGN AA	Mc	Matra	Tamil	Tamil	[1002]	Yes
38	0BBF	ி	TAMIL VOWEL SIGN I	Mc	Matra	Tamil	Tamil	[1002]	Yes
39	0BC0	ீ	TAMIL VOWEL SIGN II	Mn	Matra	Tamil	Tamil	[1002]	Yes
40	0BC1	ு	TAMIL VOWEL SIGN U	Mc	Matra	Tamil	Tamil	[1002]	Yes
41	0BC2	ூ	TAMIL VOWEL SIGN UU	Mc	Matra	Tamil	Tamil	[1002]	Yes
42	0BC6	ெ	TAMIL VOWEL SIGN E	Mc	Matra	Tamil	Tamil	[1002]	Yes
43	0BC7	ே	TAMIL VOWEL SIGN EE	Mc	Matra	Tamil	Tamil	[1002]	Yes
44	0BC8	ை	TAMIL VOWEL SIGN AI	Mc	Matra	Tamil	Tamil	[1002]	Yes

45	0BCA	ஊ	TAMIL VOWEL SIGN O	Mc	Matra	Tamil	Tamil	[1002]	Yes
46	0BCB	஋	TAMIL VOWEL SIGN OO	Mc	Matra	Tamil	Tamil	[1002]	Yes
47	0BCC	஌	TAMIL VOWEL SIGN AU	Mc	Matra	Tamil	Tamil	[1002]	Yes
48	0BCD	஍	TAMIL SIGN VIRAMA	Mn	Matra	Tamil	Tamil	[1002]	Yes

Table 5: Code point repertoire

5.3 Code points not included:

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

In section 7, the Whole Label Evaluation (WLE) rules are given which covers Tamil Language under the purview of the NBBP for 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 descriptive names, the variable names are used. The second section lists four operators along with their functions which are assumed while specifying the rules. The following

two sections describe the two major categories of the Akshar formations first of which begins with the vowels and the second one with the consonants.

5.5.1 Variables involved

C	→ Consonant
M	→ Matra
V	→ Vowel
X	→ Visarga / Aytham
H	→ Halant / 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 The 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 /aḳ/ U+0B85 U+0B83	அ ஃ 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 Halant (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 Halant [H]

C [M| H|X]

Examples:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Matra	C[M]	கி /ki/	க ி 0B95 0BBF
Consonant + Halant	C[H]	க̣ /k/ (Pure Consonant)	க ̣ U+0B95 U+0BCD
Consonant + Visarga	C[X]	கஃ /kḥ/	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 /muk/	ம ு ஃ U+0BAE U+0BC1 U+0B83

Table 11

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

Example:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Halant + Consonant + Halant + Consonant	CHCHC	ழ்த்த/ <i>lta</i> /	ழ ஂ த ஂ த U+0BB4 U+0BCD U+0BA4 U+0BCD U+0BA4

Table 12

Subsets:

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

Example:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Halant + Consonant + Matra	CHC[M]	க்ரு /kku/	க ஂ க ு U+0B95 U+0BCD U+0B95 U+0BC1
Consonant + Halant + Consonant + Visarga	CHC[X]	க்கஃ /kkak/	க ஂ க ஃ U+0B95 U+0BCD U+0B95 U+0B83

Table 13

3. B. *3(CH)CM may be followed by a B, D or X

Example:

Sequence Description	Sequence	Example	Constituting characters
Consonant + Halant + Consonant + Matra + Visarga	CHCM[X]	ம்முஃ /kkīh/	ம ஂ ம ு ஃ 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 which can be created by using the characters permitted as per the [MSR] and look exactly alike. The NBGP categorizes these confusingly similar variants in three groups

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

6.1 Group 1: Confusing due to exact look

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

First is because of the split matra TAMIL VOWEL SIGN AU (௪ௌ U+0BCC) having left and right side catenators which sit on the preceding consonant. It looks exactly alike to a combination of another matra TAMIL VOWEL SIGN E (௪ U + 0BC6) followed by consonant TAMIL LETTER LLA (ள U+0BB3). The later combination also needs a preceding consonant.

Second one is a pure vowel (௪ௌ U+0B94) which exactly looks similar to a vowel + Consonant (௪ௌ U+0B92 U+0BB3) combination. These can cause confusion even to a careful observer and hence being proposed as variants. Following is the brief description of these variants followed by variants in Table 15 and Table 16.

6.1.1 Pure Vowel and a Vowel followed by consonant Tamil Consonant Lla:

Variant 1	Variant 2
௪ௌ U+0B94	௪ ள U+0B92 U+0BB3

Table 15: Proposed Variants - Set 1

6.1.2 Any Consonant followed by a Split Matra and the same Consonant followed by a Matra and Tamil Consonant Lla

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

Table 16: Proposed Variants - Set 2

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, as advised by ICANN, no cases belonging to Group 2 are proposed, as there is another panel (String similarity assessment panel) entrusted to deal with such cases.

Variant 1	Variant 2
ஐ (U+0B9C)	ஐ U+0B9C

Table 17: 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 “஌” (U+ 0BCA) looks exactly same as TAMIL VOWEL SIGN E “஌” (U+0BC6) followed by TAMIL VOWEL SIGN AA “஌” (U+0BBE). However as the formation is not valid as per akshar formation rules, this case is not proposed as variant.

Variant 1	Variant 2
஌ (U+0BCA)	஌ ஌ (U+0BC6) (U+0BBE).

Table 18: Not Proposed as Variants - Set 2

6.4 Variant Disposition:

As variants mentioned in both (Table 15: Proposed Variants - Set 1

15 and Table 16: Proposed Variants - Set 2

16) categories are of confusingly similar, albeit of a peculiar nature, it is proposed that they be considered of "blocking" nature.

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.

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 Tamil 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	→	Halant / Virama / Pulli

Below are the specific WLE rules:

1. H: must be preceded by C
2. M: must be preceded by C
3. X: must be preceded by either of V, C, or M

Note: As "Fa" character is not present in Tamil, Visarga followed by Pa is used as an alternative of "Fa".

Example: ஃபாஃரிஸ் (*Foreign*)

TAMIL SIGN VISARGA
TAMIL LETTER PA

TAMIL VOWEL SIGN AA
TAMIL LETTER RA
TAMIL VOWEL SIGN I
TAMIL LETTER NNNA
TAMIL SIGN VIRAMA

This is a practice in some Modern Tamil community which has originated from “arwi” influence. Arwi is practiced by Tamil speaking Muslim community. The Tamil NBGP team is still discussing internally on this whether it is a standard practice or not. Once finalized, a clear decision on this will be taken.

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
Dr. Shanmugam R	Tamil
Sinnathambi Shanmugarajah	Tamil
Prasad PK	Malayalam

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)

<https://ta.wikipedia.org/s/jt1>

https://en.wikipedia.org/wiki/Tamil_script

<http://www.virtualvinodh.com/wp/tamil-script-evolution/>

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. Kothandaraman Pon [1997]., A Grammar of contemporary Literary Tamil.
International Institute of Tamil Studies.

(To be completed)

11 Appendix A: Variants based on pure visual similarity







Variant 1	Variant 2
 U+0B94	  U+0B92 U+0BB3
 U+0BCC	  U+0BC6 U+0BB3

Table 19: Tamil Variants based on pure visual similarity

12 Appendix B: Cross-script Variants

Tamil script has a set of possible cross-script variants only with the Malayalam script. The below 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. Here are some of the examples of them.

வமி - வமி

ஜெஸி - ஜெஸி

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 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 other script, it essentially provides a visually distinguishability and hence a non-confusable string.

Tamil	Malayalam
ജ U+0B9C	ജ U+0D1C
ഖ U+0BB5	ഖ U+0D16
ഥ U+0BAE	ഥ U+0D25
സ U+0BB8	സ U+0D38
ി U+0BBF	ി U+0D3F
െ U+0BC6	െ U+0D46
ം U+0BCD	ം U+0D4E
േ U+0BC7	േ U+0D47

Table 20: Cross script variants