

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

LGR Version: 3.0

Date: 1 October 2018

*Document version:*1.5

Authors: Sinhala Generation Panel

1. General Information/ Overview/ Abstract

This document lays down the Label Generation Ruleset for Sinhala script. Three main components of the Sinhala Script LGR, i.e. Code Point Repertoire, Variant Code Points and Whole Label Evaluation Rules, have been described in detail here following the historical background of the Script in Section 3.

All these components have been incorporated in a machine-readable format in the accompanying XML file named "Proposal-LGR-Sinh-20181001.xml".

In addition, a document named "Sinhala-Test-Labels-20181001.txt" has been provided, containing a list of labels covering the repertoire and which can produce variants as laid down in Section 6 of this document and it also provides valid and invalid labels as per the Whole Label Evaluation Rules laid down in Section 7.

2. Script for which the LGR is Proposed

ISO 15924 Code: Sinh

ISO 15924 Key N°: 348

ISO 15924 English Name: Sinhala

Latin transliteration of native script name: Simhala

Native name of the script: සිංහල

Maximal Starting Repertoire (MSR) version: 3 [MSR]

3. Background on Script and Principal Languages Using It

The Sinhala language belongs to the Indo-European language family with its roots deeply associated with Indo-Aryan sub-family to which the languages such as Persian and Hindi belong. Although it is not very clear whether people in Sri Lanka spoke a dialect of Prakrit at the time of arrival of Buddhism in the island, there is enough evidence that Sinhala evolved from mixing of Sanskrit, Magadhi (the language which was spoken in Magadha Province of India where Lord Buddha was born) and local language which was spoken by people of Sri Lanka prior to the arrival of Vijaya, the founder of the Sinhala Kingdom. It is also surmised that Sinhala had evolved from an ancient variant of Apabhraṃśa (middle Indic) which is known as 'Elu'. Historically Elu was preceded by Hela or Pali Sihala.

Sinhala, though it has close relationships with Indo Aryan languages which are spoken primarily in northern, north-eastern and central India, was very much influenced by Tamil which belongs to the Dravidian family of languages. Though Sinhala is related closely to Indic languages, it also has its own unique characteristics: Sinhala uses symbols for two vowels which are not found in any other Indic languages in India: 'æ' (අඞ) and 'æ:' (අඞඞ).

3.1. The Evolution of the Script

The Sinhala script evolved from the Southern Brahmi script from which almost all the Southern Indic Scripts, such as Telugu and Oriya, had evolved. Later Sinhala was influenced by Pallava Grantha writing of Southern India. Since 1250 AD, the Sinhala script has remained the same with few changes. Although some scholars are of the view that the Brahmi Script arrived with Buddhism, *Mahavansa* (Great Chronicle) speaks of written language even right after the arrival of *Vijaya*. Archeologists have found pottery fragments in Anuradhapura, Sri Lanka, with older Brahmi script inscriptions, which have been carbon dated to 5th century BC. The earliest artifacts with Brahmi script found in India have been dated to 6th Century BC in Tamil Nadu though most of the early Brahmi writing found in India has been attributed to emperor Ashoka in the 3rd century BC.

Sinhala letters are round-shaped and are written from left to right and they form the most circular-shaped script found among the Indic scripts. The evolution of the script to the present shapes may have taken place due to writing on Ola leaves. Unlike chiseling on a rock, writing on palm leaves has to be more round-shaped to avoid the stylus ripping the Palm leaf while writing on it. When drawing vertical or horizontal straight lines on Ola leaf, the leaves would have been ripped and this also may have influenced Sinhala not to have a period or full stop. Instead a stylistic stop which was known as 'Kundaliya' is used. Period and commas were later introduced into the Sinhala script after the introduction of paper due to the influence of Western languages.

The following Figure 1 shows the evolution of the Sinhala Script over the years in different major periods.¹

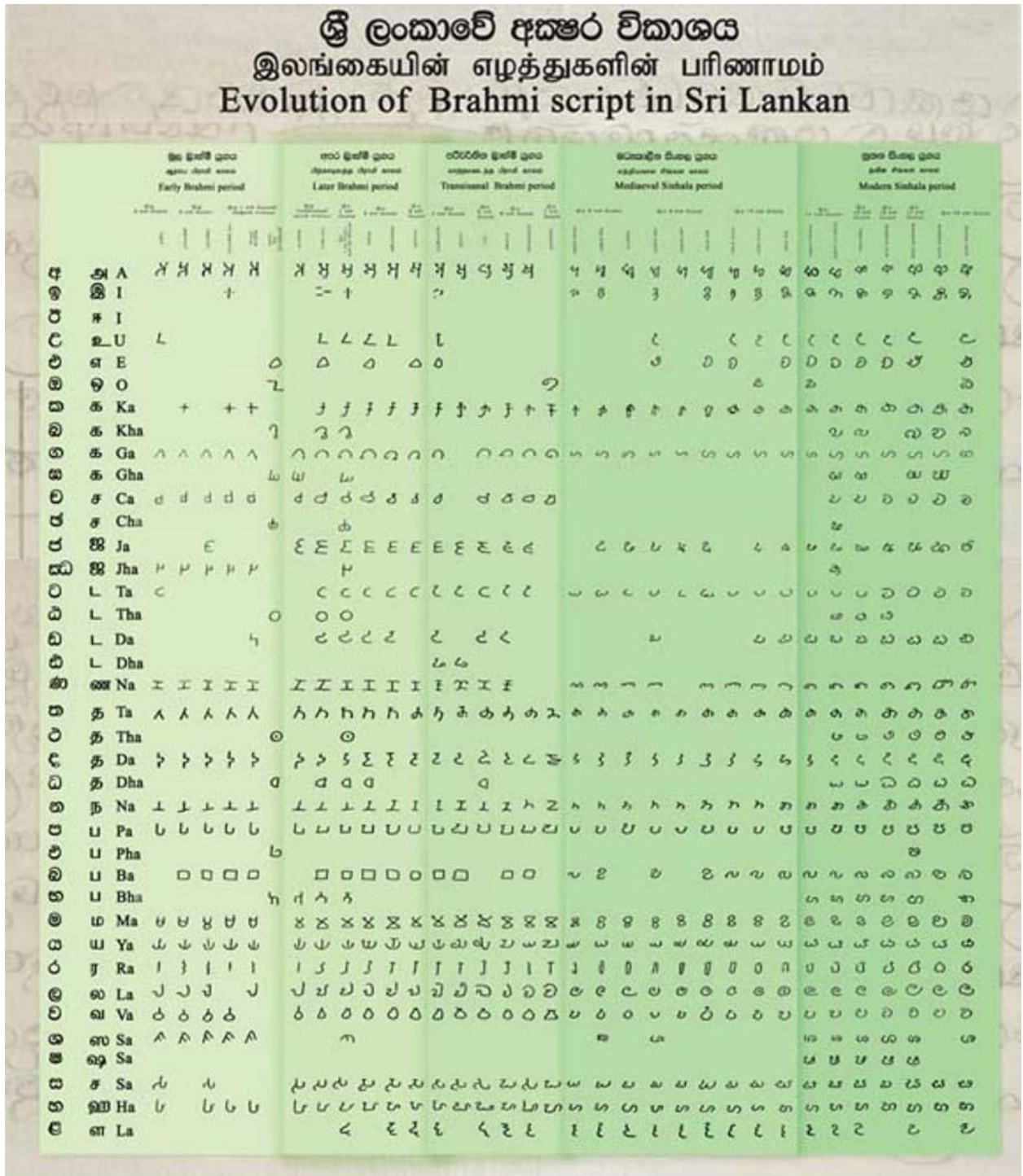


Figure 1: Evolution of Sinhala Script

¹ Source: <http://www.archaeology.gov.lk/web/images/stories/gallery/alphabet/Alphabet.jpg>

3.2. Languages Considered

The Sinhala script is used to write the Sinhala (sin) language, which is one of the official languages of Sri Lanka. In addition, it is used to write Pali (pli) and Sanskrit (san) languages in Sri Lanka. The Sinhala script is used on the Island of Sri Lanka (predominantly in the south) and Sinhala Diaspora in Middle East (Saudi Arabia, Kuwait, Qatar, and UAE), Britain, USA, Australia and Canada. The scripts covered by the Neo-Brahmi GP are related to the Sinhala script. Based on an initial analysis, the Sinhala GP has found script similarity with Malayalam, Kannada and Telugu scripts. In addition, Myanmar script is also related. The Sinhala GP has investigated cross-script variants with these scripts.

3.3. The Structure of Written Sinhala

As most Brahmi-derived scripts, Sinhala is an alpha-syllabary writing system and written from left to right. All the categories of Consonants, Vowels, Sannjakas, Matras, Halant, Anusvara and Visarga are discussed below.

3.3.1. The Consonants

There are 40 consonants in the Sinhala alphabet and 38 of them are selected for inclusion. Its consonants imply an inherent vowel a (අ) when they are used without dependent vowels. Absence of the inherent vowel is marked by adding *halkirima* or halanta (remover of the inherent vowel) to the consonant; thus ක [ka] becomes ක් [k], and ව [va] becomes ව් [v] with *halkirima*.

In addition, conjunct characters and touching letters are features of Sinhala text, but do not require representation in the root-zone for labels. There are conjunct characters used for writing consonant clusters. Though these characters do not have separate code points, ඥ (jna) the symbol is considered as representing ජ්+ඤ (j+na), identical to the consonant in contemporary Sinhala ඥ which has a code point U+0DA5. Other conjunct characters include ක්ඝ (kSa), ක්ව (kva), ක්ඛ (nda), ක්ඪ (ndha), ක්ඹ (ntha), ක්ඪ (ttha) etc. the few conjunct consonants that are not used in contemporary writing include ධ (ddha), ධ (dva), ධ (TTha) and ධ (njca). Moreover, there are touching letters used in old Sinhala writing but not in contemporary writing. However, touching letters are frequently used to write in Pali. These touching combinations are formed by deleting white space between two characters, e.g.: ක්ක (kka), ක්ඛ (kkha), ග්ග (gga), ච්ච (ccha), ජ්ජ (jja), ජ්ඪ (jjha), ධ්ධ (TTha), ජ්ඵ (ppha), ම්ම (mma), etc.

When modifiers are added to any of the above categories, including: (i) individual consonants, (ii) conjunct consonants, or (iii) touching consonants, they will be formed as follows: if ක්

(modifier for vowel ෙ) is added to ක (k), කෞ (kSa) or ක්කි (kkha) they become කෝ, කෞ or ක්කෝ respectively.

Special symbols ඌර (rakaranshaya) for ර (ra) and ඌය (yanshaya) for ය (ya) are used in Sinhala writing when they occur after a consonant (from which the inherent vowel has been removed). For instance, ක්‍රරම (krama) and වාක්‍යය (vakya) are not accepted forms in Sinhala and they are written as ක්‍රම and වාක්‍යයි. Further, ෙ (rephaya) is used to denote ර (r, i.e. ර without its inherent vowel) before a consonant and added on top of the consonants with an inherent vowel: තරක can be written as තර්කී, and both these forms are accepted. However, after ර (r) not yanshaya but ය (ya) is used; so කාර්ය or කාර්යී are not accepted but කාර්යී or කාර්යී are.

1.1 සිංහල වර්ණ මාලාව

ඈ	ඈ	ඈ	ඈ	ඉ	ඊ
උ	උ	උ	උ	ඌ	ඌ
ඍ	ඍ	ඍ	ඎ	ඏ	ඏ
(ඈ)ං	(ඈ)ඃ				
ක	ක	ක	ක	ක	ක
ඛ	ඛ	ඛ	ඛ	ඛ	ඛ
ඃ	ඃ	ඃ	ඃ	ඃ	ඃ
ආ	ආ	ආ	ආ	ආ	ආ
ඈ	ඈ	ඈ	ඈ	ඈ	ඈ
ඉ	ඉ	ඉ	ඉ	ඉ	ඉ
ඊ	ඊ	ඊ	ඊ	ඊ	ඊ
උ	උ	උ	උ	උ	උ
ඌ	ඌ	ඌ	ඌ	ඌ	ඌ

Figure 2: Sinhala Alphabet from *Sinhala Lekhana Rithiya* by NIE² Sri Lanka

² National Institute of Education - Sri Lanka.

			Labial	Dental	Alveolar	Retroflex	Palatal	Velar	Glottal
Stops	-Voice	-Asp	ප/p/	ත/t/		ට/t/		ක/k/	
		+Asp ³	ඵ/p ^h /	ඵ/t ^h /		ඨ/t ^h /		ඛ/k ^h /	
	+Voice	-Asp	බ/b/	ද/d/		ඞ/d/		ග/g/	
		+Asp	ඞ/b ^h /	ධ/d ^h /		ඬ/d ^h /		ඝ/g ^h /	
Affricates	-Voice	-Asp					ච/c/		
		+Asp					ඡ/c ^h /		
	+Voice	-Asp					ජ/j/		
		+Asp					ඣ/j ^h /		
Pre-nasalized voiced stops			ඹ / ^m b/	ඳ / ⁿ d/		ඞ / ^ɳ d/	ඡ / ^ɲ j/	ඟ / ^ŋ g/	
Nasals			ම /m/		න, ඞ/n/		ඤ/ɲ/	ඞ/ŋ/	
Trill						ර/r/			
Lateral					ල, ඳ/l/				
Spirants			ෆ/f/	ස/s/			ශ, ෂ/ʃ/		හ/h/
Semivowels			ව/v/				ය/y/		

Table 1: Sinhala Consonant Classification with Pronunciation

3.3.2. The Vowels

Independent vowels are used at the beginning of a word and dependent vowels are used after consonants. There are separate symbols for dependent vowel forms of all the vowels except the inherent vowel අ in Sinhala. Some characters not used in contemporary writing have not been selected for inclusion. The correlation of the independent and dependent vowels is listed in Table 2.

Independent Vowels			Matra (Dependent Vowels)	
අ	/a/	0D85		
ආ	/a:/	0D86	ආ	0DCF
ඇ	/æ/	0D87	ඈ	0DD0
ඈ	/æ:/	0D88	ඉ	0DD1
ඉ	/i/	0D89	ඊ	0DD2
ඊ	/i:/	0D8A	උ	0DD3
උ	/u/	0D8B	ඌ	0DD4
ඌ	/u:/	0D8C	ඍ	0DD6
ඍ	/ri/	0D8D	ඎ	0DD8

³Aspirated letters are only pronounced in particular use of the language. Ex: in dhamma chanting by the Buddhist monks and some announcers of radio or TV channels.

ඞaa ⁴ /ri:/ 0D8E	ාaa 0DF2
ඞ /ilu/ 0D8F	ාඞ 0DDF
ඞඞ /ilu:/ 0D90	ාඞ 0DF3
ඞ /e/ 0D91	ෙ 0DD9
ඞ /e:/ 0D92	ෙඞ 0DDA
ඞඞ /ai/ 0D93	ෙෙ 0DDB
ඞ /o/ 0D94	ො 0DDC
ඞ /o:/ 0D95	ොඞ 0DDD
ඞඞ /au/ 0D96	ොඞ 0DDE

Table 2: Vowels with Corresponding Matras

3.3.3. Halanta: The Inherent Vowel Remover

Halanta (ඞ 0DCA), which is also called *halkirima* or *hallakuna*, is used to remove the inherent vowel of the consonants in Sinhala. This is thus used to join consonants and form conjunct characters.

Ex: ඞ (U+0DAD) + ඞ (U+0DCA) = ඞ (U+0DAD\U+0DCA)
 ඞ (U+0DC0) + ඞ (U+0DCA) = ඞ (U+0DC0\U+0DCA)
 ඞ (U+0D9A) + ඞ (U+0DCA) + ZWJ (200D) + ඞ (U+0DC2) =
 ඞ (U+0D9A\U+0DCA\U+200D\U+0DC2)

3.3.4. The Anusvara (ං)

The *anusvara* (U+0D82), pronounced /ŋ/, represents all the nasals. It can be preceded by any sign except *halanta* (U+0DCA).

Ex: ඞ (U+0D85) + ඞ (U+0D82) = ඞං (U+0D85\U+0D82)
 ඞ (U+0DB4) + ඞ (U+0DD2) + ඞ (U+0D82) = ඞං (U+0DB4\U+0DD2\U+0D82)

3.3.5. The Visarga (ඞ)

The *visarga* (U+0D83) is a rarely used sign and pronounced as /h/. Most of the Sinhala words with *visarga* are borrowings from Sanskrit.

Ex: ඞඞඞඞඞ /antahpurə/

⁴ Code points 0D8E, 0D8F and 0D90 are not selected: see Section 5.4.

3.3.6. Sannjakas

As given in Table 1 there are five separate letters for pre-nasalized voiced stops called sannjakas in Sinhala. From among these, ජ is not frequently used. One constraint for Sannjakas is that they cannot be followed by *halanta*.

4. Overall Development Process and Methodology

The Sinhala LGR proposal has been developed by the Sinhala Generation Panel (GP) following the principles given in the LGR Procedure. The GP was formed from expert members from multiple backgrounds, with expertise in Sinhala linguistics, Sinhala language processing, Sinhala standardization, (IDN) ccTLD operations and policy development. Many of the members have been active in Sinhala standardization and participated in Sinhala Unicode standardization. The GP was coordinated and supported by Theekshana (which is a not-for-profit company managed by staff of UCSC) and University and Colombo School of Computing (UCSC). The group was organized by the co-chairs, and started its work after a face-to-face training conducted by ICANN in December 2017. Following the training, the GP members met face to face at UCSC regularly to discuss the repertoire, variant code point and whole label evaluation rules.

During the training, the Sinhala GP also met with the Neo-Brahmi GP to discuss cross-script variants with the scripts covered by Neo-Brahmi GP, and to coordinate whole label evaluation rules.

Based on these discussions, the Sinhala GP has finalized its proposal for the Root Zone LGR.

5. Repertoire

Sinhala code point repertoire is discussed in this section.

5.1. Sinhala Section of Maximal Starting Repertoire (MSR)

The Sinhala Unicode chart is given below, highlighting the characters included and excluded in the Sinhala script by the [MSR].

	0D8	0D9	0DA	0DB	0DC	0DD	0DE	0DF
0		ඵ ො 0D90	ඵ ො 0DA0	ඵ ො 0DB0	ඵ ො 0DC0	ඵ ො 0DD0		
1		ඵ ො 0D91	ඵ ො 0DA1	ඵ ො 0DB1	ඵ ො 0DC1	ඵ ො 0DD1		
2	ඵ ො 0D82	ඵ ො 0D92	ඵ ො 0DA2		ඵ ො 0DC2	ඵ ො 0DD2		ඵ ො 0DF2
3	ඵ ො 0D83	ඵ ො 0D93	ඵ ො 0DA3	ඵ ො 0DB3	ඵ ො 0DC3	ඵ ො 0DD3		ඵ ො 0DF3
4		ඵ ො 0D94	ඵ ො 0DA4	ඵ ො 0DB4	ඵ ො 0DC4	ඵ ො 0DD4		ඵ ො 0DF4
5	ඵ ො 0D85	ඵ ො 0D95	ඵ ො 0DA5	ඵ ො 0DB5	ඵ ො 0DC5			
6	ඵ ො 0D86	ඵ ො 0D96	ඵ ො 0DA6	ඵ ො 0DB6	ඵ ො 0DC6	ඵ ො 0DD6	ඵ ො 0DE6	
7	ඵ ො 0D87		ඵ ො 0DA7	ඵ ො 0DB7			ඵ ො 0DE7	
8	ඵ ො 0D88		ඵ ො 0DA8	ඵ ො 0DB8		ඵ ො 0DD8	ඵ ො 0DE8	
9	ඵ ො 0D89		ඵ ො 0DA9	ඵ ො 0DB9		ඵ ො 0DD9	ඵ ො 0DE9	
A	ඵ ො 0D8A	ඵ ො 0D9A	ඵ ො 0DA A	ඵ ො 0DB A	ඵ ො 0DC A	ඵ ො 0DD A	ඵ ො 0DE A	
B	ඵ ො 0D8B	ඵ ො 0D9B	ඵ ො 0DA B	ඵ ො 0DB B		ඵ ො 0DD B	ඵ ො 0DE B	
C	ඵ ො 0D8C	ඵ ො 0D9C	ඵ ො 0DA C			ඵ ො 0DD C	ඵ ො 0DE C	
D	ඵ ො 0D8D	ඵ ො 0D9D	ඵ ො 0DA D	ඵ ො 0DB D		ඵ ො 0DD D	ඵ ො 0DE D	
E	ඵ ො 0D8E	ඵ ො 0D9E	ඵ ො 0DA E			ඵ ො 0DD E	ඵ ො 0DE E	
F	ඵ ො 0D8F	ඵ ො 0D9F	ඵ ො 0DA F		ඵ ො 0DC F	ඵ ො 0DD F	ඵ ො 0DE F	

Color convention:

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**

Figure 3: MSR3 for Sinhala Script

5.2. Code Point Repertoire

This section provides the code point repertoire that Sinhala GP proposes to be included in the Sinhala LGR for use with the Sinhala language, based on the references listed in Section 9, e.g. [102] and [201].

#	Unicode Code Point	Glyph	Character Name	Category
1	0D82	◌◌	SINHALA SIGN ANUSVARAYA	Anusvara
2	0D83	◌◌◌	SINHALA SIGN VISARGAYA	Visarga
3	0D85	අ	SINHALA LETTER AYANNA	Vowel
4	0D86	ආ	SINHALA LETTER AAYANNA	Vowel
5	0D87	ඇ	SINHALA LETTER AEYANNA	Vowel
6	0D88	ඈ	SINHALA LETTER AEEYANNA	Vowel
7	0D89	ඉ	SINHALA LETTER IYANNA	Vowel
8	0D8A	ඊ	SINHALA LETTER IYANNA	Vowel
9	0D8B	උ	SINHALA LETTER UYANNA	Vowel
10	0D8C	ඌ	SINHALA LETTER UUYANNA	Vowel
11	0D8D	ඍ	SINHALA LETTER IRUYANNA	Vowel
12	0D91	එ	SINHALA LETTER EYANNA	Vowel
13	0D92	ඒ	SINHALA LETTER EEYANNA	Vowel
14	0D93	ඓ	SINHALA LETTER AIYANNA	Vowel
15	0D94	ඔ	SINHALA LETTER OYANNA	Vowel
16	0D95	ඕ	SINHALA LETTER OYANNA	Vowel
17	0D96	ඖ	SINHALA LETTER AUYANNA	Vowel
18	0D9A	ක	SINHALA LETTER ALPAPRAANA KAYANNA	Consonant
19	0D9B	ඛ	SINHALA LETTER MAHAAPRAANA KAYANNA	Consonant
20	0D9C	ග	SINHALA LETTER ALPAPRAANA GAYANNA	Consonant
21	0D9D	ඝ	SINHALA LETTER MAHAAPRAANA GAYANNA	Consonant
22	0D9F	ඞ	SINHALA LETTER SANYAKA GAYANNA	Sannjaka
23	0DA0	ච	SINHALA LETTER ALPAPRAANA CAYANNA	Consonant
24	0DA1	ඡ	SINHALA LETTER MAHAAPRAANA CAYANNA	Consonant

25	0DA2	ජ	SINHALA LETTER ALPAPRAANA JAYANNA	Consonant
26	0DA3	ඤ	SINHALA LETTER MAHAAPRAANA JAYANNA	Consonant
27	0DA4	ඤ	SINHALA LETTER TAALUJA NAASIKYAYA	Consonant
28	0DA5	ඤ	SINHALA LETTER TAALUJA SANYOOGA NAAKSIKYAYA	Consonant
29	0DA7	ට	SINHALA LETTER ALPAPRAANA TTAYANNA	Consonant
30	0DA8	ට	SINHALA LETTER MAHAAPRAANA TTAYANNA	Consonant
31	0DA9	ඨ	SINHALA LETTER ALPAPRAANA DDAYANNA	Consonant
32	0DAA	ඨ	SINHALA LETTER MAHAAPRAANA DDAYANNA	Consonant
33	0DAB	ඩ	SINHALA LETTER MUURDHAJA NAYANNA	Consonant
34	0DAC	ඨ	SINHALA LETTER SANYAKA DDAYANNA	Sannjaka
35	0DAD	ඨ	SINHALA LETTER ALPAPRAANA TAYANNA	Consonant
36	0DAE	ඨ	SINHALA LETTER MAHAAPRAANA TAYANNA	Consonant
37	0DAF	ඨ	SINHALA LETTER ALPAPRAANA DAYANNA	Consonant
38	0DB0	ඨ	SINHALA LETTER MAHAAPRAANA DAYANNA	Consonant
39	0DB1	ඨ	SINHALA LETTER DANTAJA NAYANNA	Consonant
40	0DB3	ඨ	SINHALA LETTER SANYAKA DAYANNA	Sannjaka
41	0DB4	ඨ	SINHALA LETTER ALPAPRAANA PAYANNA	Consonant
42	0DB5	ඨ	SINHALA LETTER MAHAAPRAANA PAYANNA	Consonant
43	0DB6	ඨ	SINHALA LETTER ALPAPRAANA BAYANNA	Consonant
44	0DB7	ඨ	SINHALA LETTER MAHAAPRAANA BAYANNA	Consonant
45	0DB8	ඨ	SINHALA LETTER MAYANNA	Consonant
46	0DB9	ඨ	SINHALA LETTER AMBA BAYANNA	Sannjaka
47	0DBA	ඨ	SINHALA LETTER YAYANNA	Consonant
48	0DBB	ඨ	SINHALA LETTER RAYANNA	Consonant

49	0DBD	ඳ	SINHALA LETTER DANTAJA LAYANNA	Consonant
50	0DC0	ච	SINHALA LETTER VAYANNA	Consonant
51	0DC1	ඟ	SINHALA LETTER TAALUJA SAYANNA	Consonant
52	0DC2	ඡ	SINHALA LETTER MUURDHHAJA SAYANNA	Consonant
53	0DC3	ඪ	SINHALA LETTER DANTAJA SAYANNA	Consonant
54	0DC4	ඞ	SINHALA LETTER HAYANNA	Consonant
55	0DC5	ඳ	SINHALA LETTER MUURDHHAJA LAYANNA	Consonant
56	0DC6	ඟ	SINHALA LETTER FAYANNA	Consonant
57	0DCA	ඳ̣	SINHALA SIGN AL-LAKUNA	Halant
58	0DCF	ඳ̣	SINHALA VOWEL SIGN AELA-PILLA	Matra
59	0DD0	ඳ̣	SINHALA VOWEL SIGN KETTI AEDA-PILLA	Matra
60	0DD1	ඳ̣	SINHALA VOWEL SIGN DIGA AEDA-PILLA	Matra
61	0DD2	ඳ̣	SINHALA VOWEL SIGN KETTI IS-PILLA	Matra
62	0DD3	ඳ̣	SINHALA VOWEL SIGN DIGA IS-PILLA	Matra
63	0DD4	ඳ̣	SINHALA VOWEL SIGN KETTI PAA-PILLA	Matra
64	0DD6	ඳ̣	SINHALA VOWEL SIGN DIGA PAA-PILLA	Matra
65	0DD8	ඳ̣	SINHALA VOWEL SIGN GAETTA-PILLA	Matra
66	0DD9	ඳ̣	SINHALA VOWEL SIGN KOMBUVA	Matra
67	0DDA	ඳ̣	SINHALA VOWEL SIGN DIGA KOMBUVA	Matra
68	0DDB	ඳ̣	SINHALA VOWEL SIGN KOMBU DEKA	Matra
69	0DDC	ඳ̣	SINHALA VOWEL SIGN KOMBUVA HAA AELA-PILLA	Matra
70	0DDD	ඳ̣	SINHALA VOWEL SIGN KOMBUVA HAA DIGA AELA-PILLA	Matra
71	0DDE	ඳ̣	SINHALA VOWEL SIGN KOMBUVA HAA GAYANUKITTA	Matra
72	0DF2	ඳ̣	SINHALA VOWEL SIGN DIGA GAETTA-PILLA	Matra

Table 3: Code Point Repertoire

5.3. Code point sequences

The following sequences are defined for the purposes of variants rules in Sections 6 below.

#	Unicode Code Point	Glyph	Character Name
1	U+0DC3 U+0DD8	සා	SINHALA LETTER DANTAJA SAYANNA + SINHALA VOWEL SIGN GAETTA-PILLA
2	U+0DB5 U+0DD9	ඵ	SINHALA LETTER MAHAAPRAANA PAYANNA + SINHALA VOWEL SIGN KOMBUVA
3	U+0DB5 U+0DCA	ඵ	SINHALA LETTER MAHAAPRAANA PAYANNA + SINHALA SIGN AL-LAKUNA
4	U+0DB9 U+0DCA	ඵ	SINHALA LETTER AMBA BAYANNA + SINHALA SIGN AL-LAKUNA

Table 3a: Code Point Sequences

5.4. Code point not included

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

#	Unicode Code Point	Glyph	Character Name	Reason for exclusion
1	0D8E	සා	SINHALA LETTER IRUUYANNA	Usage unknown
2	0D8F	ඵ	SINHALA LETTER ILUYANNA	Usage unknown
3	0D90	ඵ	SINHALA LETTER ILUUYANNA	Usage unknown
4	0D9E	ඵ	SINHALA LETTER KANTAJA	Not in modern usage
5	0DA6	ඵ	SINHALA LETTER SANYAKA	Only used in the word ‘ඉඡු’ (this word is used to call dogs)
6	0DDF	ඵ	SINHALA VOWEL SIGN GAYANUKITTA	Usage unknown
7	0DF3	ඵ	SINHALA VOWEL SIGN DIGA GAYANUKITTA	Usage unknown

Table 4: Code Points Not Included

5.5. Structural Formation of Sinhala

As written in most Brāhmi-derived scripts, Sinhala follows a particular way of formation of its words, known as "akshar". In Sinhala they are called "akshara".

ZWJ is specifically used for rendering of Rakar (Halanta+Ra), Yansa (Halanta+Ya) and Reph forms in Sinhala as well as conjuncts as in most of Brahmi derived scripts. (Please refer to Page 5.) One of the most important deficiencies of not being able to have Top Level Domain with Rakar form is that one cannot have “ඡ්‍රී” (Shri) in a top level domain name, which is an important and

hallowed sound in Sinhala. In order to write the name of the country, Sri Lanka in Sinhala, ශ්‍රී is used.

5.6. Akshar Formation Rules for Sinhala

This section details the Akshar formation rules as applicable to Sinhala. First the categories of characters are given in the form of variables. Then use of two major categories, vowels and consonants, for Akshar formation is discussed.

5.6.1. Variables involved

C	→	Consonant
V	→	Vowel
M	→	Matras / Vowel Signs
B	→	Anusvara (Bindu)
X	→	Visarga
H	→	Halanta / Virama
J	→	Sannjakas

5.6.2. Operators Used

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

Table 5: Operators Used for Rules

5.6.3. The Vowel Sequence

A vowel sequence begins with a vowel in Sinhala. It may optionally be followed by an Anusvara (B), or a Visarga (X).

Sequence Description	Sequence	Example	Example Decomposition
----------------------	----------	---------	-----------------------

Vowel	V	අ /a/ U+0D85	
Vowel + Anusvara	V[B]	අං /aŋ/ U+0D85\U+0D82	අං U+0D85\U+0D82
Vowel + Visarga	V[X]	අඃ /ah/ U+0D85\U+0D83	අඃ U+0D85\U+0D83

Table 6: Structure of Vowel Sequences

5.6.4. Consonant Sequence

A consonant sequence begins with a consonant. It may optionally be followed by a Matra (M), Anusvara (D), Visarga (X) or a Halanta (H). Examples are given in the Table 7.

Sequence Description	Sequence	Example	Example Decomposition
Consonant	C	ක /ka/ U+0D9A	
Consonant + Matra	C[M]	කො /ko/ U+0D9A U+0DDC	කො U+0D9A U+0DDC
Consonant + Halanta	C[H]	ක් /k/ U+0D9A U+0DCA	ක් U+0D9A U+0DCA
Consonant + Anusvara	C[B]	කං /kaŋ/ U+0D9A U+0D82	කං U+0D9A U+0D82
Consonant + Visarga	C[X]	කඃ /kah/ U+0D9A U+0D83	කඃ U+0D9A U+0D83
Consonant + Matra + Anusvara	C[MB]	කෝං /ko:ŋ/ U+0D9A U+0DDD U+0D82	කෝං U+0D9A U+0DDD U+0D82
Consonant + Matra + Visarga	C[MX]	කීඃ /kih/ U+0D9A U+0DD2U+0D83	කීඃ U+0D9A U+0DD2U+0D83

Table 7: Structure of Consonant Sequences

5.6.5. Sannjaka Sequence

A Sannjaka sequence begins with a Sannjaka. It may optionally be followed by a Matra (M) or an Anusvara (D). Though Visarga is not followed by Sannjakas in Sinhala writing, there are few words (Ex: දාදා /iⁿdah/) in colloquial Sinhala with this formation. Examples of Sannjaka sequences are given in the Table 8.

Sequence Description	Sequence	Example	Example Decomposition
Consonant	J	ද / ⁿ da/ U+0DB3	
Consonant + Matra	J[M]	ද් / ⁿ di/ U+0DB3 U+0DD2	ද් U+0DB3 U+0DD2
Consonant + Anusvara	J[B]	ද් / ⁿ dan/ U+0DB3 U+0D82	ද් U+0DB3 U+0D82

Table 7: Structure on Sannjaka Sequences

6. Variants

This section discusses the variants for Sinhala script.

6.1. In-Script Variants

Having considered similar shapes and characters which could be used interchangeably, Sinhala GP decided the following are in-script variant code points:

- ස (U+0DC3) and ස (U+0D9D)
- ඛ (U+0DB6) and ඛ (U+0D9B)
- ඞ (U+0DC4) and ඞ (U+0DB7)
- ච (U+0DA0) and ච (U+0DC0)
- ඡ (U+0D94) and ඡ (U+0DB9)
- ජ (U+0D91) and ජ (U+0DB5)
- සා (U+0D8D) and සා (U+0DC3 U+0DD8)
- ච් (U+0D93) and ච් (U+0DB5 U+0DD9)
- ඡ් (U+0D92) and ඡ් (U+0DB5 U+0DCA)
- ඞ් (U+0D95) and ඞ් (U+0DB9 U+0DCA)

6.2. Cross-Script Variants





The Sinhala GP considered a range of South Indian and Southeast Asian scripts. Considerations and work by the New-Brahmi GP were used as a base for analysis of scripts covered by the Neo-Brahmi GP. Apart from the code page charts from the Unicode Standard, the Sinhala GP used a set of common default fonts in operating systems for cross-script variant analysis and concluded the following cases.

Though there are visually similar cases, as most of these are only for combining marks, except for Malayalam, similar labels with Telugu, Kannada, Devanagari and Gujarati cannot be formed. So Sinhala GP does not propose cross-script variants for these scripts.







6.3. Cross-Script Confusables

Sinhala panel has found that the following code points are visually similar, and they are listed as confusable code points, but not as cross-script variants.

6.3.1. U+0D82 (SINHALA SIGN ANUSVARAYA, ඌ)



Sinhala	Telugu	Kannada	Malayalam
 (U+0D82)	 (U+0C02)	 (U+0C82)	 (U+0D02)

6.3.2. U+0D83 (SINHALA SIGN VISARGAYA, ඌഃ)

Sinhala	Devanagari	Gujarati	Telugu	Kannada	Malayalam
 (U+0D83)	 (U+0903)	 (U+0A83)	 (U+0C03)	 (U+0C83)	 (U+0D03)

6.3.3. Sinhala and Malayalam

Additional Sinhala and Malayalam confusable code points are defined as follows, in addition to those in the tables above.

Sinhala	Malayalam
 (U+0D9C)	 (U+0D17)

ඔ (U+0DC1)	ඔ (U+0D36)
ඔ (U+0DCF)	ඔ (U+0D3E)

6.3.4. Sinhala and Myanmar

Sinhala has the following confusable code points with Myanmar script.

Sinhala	Myanmar
ඔ (U+0D9C)	၀ (U+1010)
ඔ (U+0DC1)	၀ (U+107B)

7. Whole Label Evaluation (WLE) Rules

This section provides the WLE rules that are required by all the languages mentioned in section 3.2 when written in Sinhala 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 Table 3: Code Point Repertoire.

C	→	Consonant
V	→	Vowel
M	→	Matras / Vowel Signs
B	→	Anusvara (Bindu)
X	→	Visarga
H	→	Halanta / Virama
J	→	Sannjakas

Below are the specific WLE rules:

1. H: must be preceded by C
Ex: CH - ක්, ච
2. M: must be preceded by C or J
Ex: CM - මකා
JM - ජි
3. X: must be preceded by either V, C, or M
Ex: VX - ආ:
CX - අන්ත:පුර
MX - බුද්ධි:
4. B: must be preceded by either V, C, J or M
Ex: VB - ආං
CB - කං
JB - කඳං
MB - ජං

8. Contributors

8.1. Sinhala GP

Dr. Ruvan Weerasinghe

Mr. Harsha Wijayawardhana

Mr. Chamila Liyanage

Mr. Pathum Egodawatha

Mr. Viraj Welgama

Ms. Aruni Goonathilake

Mr. Chamara Dissanayake

Ms. Sagarika Wickramasekara

Prof. J.B.Dissanayake

Mr. Champika Wijayathunga

Mr. Rajeewa Abeygoonaratne

Rev. Mettavihari

Ms. Nimasha Dilshani

8.2. Non-members

Prof. Rohini Paranavithana

Mr. Narada Karunatilaka

Mr. Namal Udalamatta

9. Materials and References

The following is a list of books, journals and webographies referred to while drafting this document.

[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>

9.1. Books and Journals

[101] Daniels, Peter T., and William Bright, eds. 1996. *The World's Writing Systems*. New York: Oxford University Press. ISBN 0-19-507993-0

[102] Disanayaka, JB. 2006. *Sinhala Akshara Vicharaya (Sinhala Graphology)*, Sumitha Publishers, Kalubovila. ISBN: 955-1146-44-1

[103] Fernando, PEE. 2008. *Origin and Development of Sinhalese Script*. Ministry of Education, Battaramulla: Sri Lanka National Book Development Council. ISBN: 978-955-602-044-1

[104] Gair, JW., and Karunatilaka, WS. *Literary Sinhala inflected forms: A Synopsis with a Translation Guide to Sinhala script*. Cornell University, New York

[105] *Sinhala lekhana rithiya*. 1989. National Institute of Education – Sri Lanka. ISBN 955-597-059-9

9.2. Webography

[201] Omniglot: The on-line encyclopedia of writing system and Languages, "Sinhala" <https://www.omniglot.com/writing/sinhala.htm> [Accessed on 02/09/2018]

[202] Wikipedia, “Sinhala (Unicode Block) [https://en.wikipedia.org/wiki/Sinhala_\(Unicode_block\)](https://en.wikipedia.org/wiki/Sinhala_(Unicode_block))
[Accessed on 02/09/2018]

[203] Wikipedia, “Sinhala Alphabet”, https://en.wikipedia.org/wiki/Sinhalese_alphabet
[Accessed on 02/09/2018]

[204] Wikipedia, “Sinhalese language”, https://en.wikipedia.org/wiki/Sinhalese_language
(Accessed on 02/09/2018)

[205] Department of Archaeology, www.archaeology.gov.lk [Accessed on 02/09/2018]

[206] Wikipedia, “Sinhala numerals”, https://en.wikipedia.org/wiki/Sinhala_numerals
[Accessed on 02/09/2018]