From formation to publication – Design of standards for Sinhala script

S T Nandasara[*], Yoshiki Mikami[**]

[*] University of Colombo School of Computing, Colombo, Sri Lanka.
[**] Language Observatory, Nagaoka University of Technology, Niigata, Japan.

DOI: 10.29322/IJSRP.9.07.2019.p9198
http://dx.doi.org/10.29322/IJSRP.9.07.2019.p9198

Abstract-The international scope of computing, information interchange, and electronic publishing created a need for a worldwide character encoding scheme. In this paper, we examine some of the issues involved in the process of standardization of character codes primarily for the 8-bit machine and subsequently for the 16-bit code environments in the context of Sinhala language and scripts. We examined some of the major issues involved in machine representation of textual information in graphical and phonetic approach for Sinhala scripts. A comprehensive evaluation of some of the possible representation is also presented. The design philosophy of Sri Lanka Sinhala Standard Code for Information Interchange (SLASCII) based on ISO 646 has been considered along with the phonetic model of UCS/UNICODE and ISO 10646 philosophies. The character codes require fonts that provide visual images—glyphs—corresponding to the codes in both 8 bits and UCS/UNICODE, it should have appeared on the screen or paper in the language of Sinhala, Pali, and Sanskrit using this script with an acceptable and in a comprehensive manner. This paper discusses some of the design guidelines concerning standardization in detail at the character level. It is examined in the context of SLASCII and SLS 1134:1996 and philosophy behind the design of ISO 10646 proposals.

Index Terms- Sinhala Script, Unicode, Character Encoding,

I. INTRODUCTION

In the modern Information age, the exchange of information can happen only if we are able to communicate effectively in any language and script of the world. This in turn demands easy entry of **linguistic information into the computers**, easy way of **communicating with each other** through devices, easy way of **rendering language information** on different devices to suit to individual aesthetics, easy way of **adapting already existing software** where large investment has been made and also easy way of **adapting one standard to another** through solid theoretical manner. These require standardization at various levels.

The early language coding schemes of Sri Lanka [1] can be classified into the following two major categories:

(i) **Single language user category:** Here, it is considered the efficient way of inputting the text, representing the text internally for processing, and easy way of rendering it on various output devices. But, the preparation of multilingual text is of not being concerned. Other than the control of escape sequence positions in the code table, the rest of the code positions will be used to assign the language scripts in the design. American Standard Code for Information Interchange (ASCII) and some of the national coding schemes such as Sri Lanka Standard Code for Information Interchange (SLASCII) [2] are examples for this class. For example, SLASCII caters the Sinhala scripts during the late-1980s and the aim was to develop an 8-bit code to fill the positions from A0 to FF in the single byte ISO 8859 similar code table based on the keyboard's character set.

(ii) **Multi-lingual user category:** Here an attempt is made to exploit common features of the language/scripts and the special needs are dealt with separately. The preparation of the multilingual document, multilingual dictionaries, and transliteration from one language to another are important aspects for consideration. SLS 1134:1996 phonetic model design for the Sinhala character code [3] has replaced the older typewriter metaphor concept from the previous SLASCII standard [4]. It could also be made universally applicable to all languages (UCS/UNICODE or International Standards Organization (ISO) 10646 come under the universal class).

In the multilingual user category, the 8-bit coding is divided 9into two pages based on the most significant bit (MSB) being 0 or 1. The first page (MSB=0) is left as the ASCII page, and the second page (MSB=1) is used to insert the Sinhala script code. This leaves us with six columns (96 symbols), and it is possible to give an explicit presentation to vowels, in addition to vowel signs and consonants. The first two columns of the second page are reserved for ASCII compatibility as per the recommendations of the ISO/IEC-8859-1.

Rare attempts have been made for standardizing the task of transliteration or translation. However, a good deal of effort has been made in standardizing character codes for different languages. The ISO has come up with a Universal Coded Character Set (UCS) encompassing all living scripts of the globe. On the other hand, UNICODE developed by Unicode Consortium uses a fixed two-byte code to represent all the world's normal text characters for electronic information processing.

II. Background, Objectives and Literature Survey

A. Background

The initiative described in this paper was started in the late 1980s by the University of Colombo and subsequently, standardization activities stared in early 1990s by the Council for Information Technology (CINTEC) and thereafter these activities continued by the Information and Communications Technology Agency (ICTA). Our objective was to ascertain why the usage of Sinhala was so low and to take steps to increase its use. There are a number of criticisms and questions raised by individuals and groups of personals in regards to the development of Sinhala under Unicode environment too.

The Unicode block for Sinhala simply lists the encoding symbols and gives no guidance on implementation. The representational model for encoding Indic scripts in Unicode is described in the Standard [5], with sections describing details for each of the individual scripts. The first section, covering the *Devanagari* script, provides a much greater level of detail than do subsequent sections and is provided as a template on which other scripts are based on. One problem with this arrangement is that the Indic scripts are not all the same; in fact, there are some very significant differences between scripts. Secondly, particular problems resulting from differences are that it is not clear how certain encoding formalisms specified are to be applied in other Indic scripts, and that there are common problems found in other scripts that are not addressed in the section on *Devanagari*. Thirdly, while listing the symbols may be sufficient for encodings in which each symbol is assigned a code, it is not possible to implement Sinhala in Unicode without additional information. Fourthly, the experience gained from 8-bit implementation active from 1989, addressed most of the scripts level problem did not consider at all for the formation of Sinhala in Unicode. As explained by Peter Constable [6], it is not correct to assume that Devanagari is representative of other Indic languages. The encoding of Sinhala in Unicode 12.1 was described just in two pages [5]; a casual reader gets the impression that the implementation of Sinhala Unicode is incomplete or not complicated. Further, it is not guided to sources of further information.

The Working Group formed by CINTEC in early 1990 published the working document and released this document to the public for comments [7] and subsequently modified document submitted to Sri Lankan Standard Institute (SLSI) [8] followed by sending to UNICODE consortium [9] for their consideration. As a result SLSI formed a first working group and developed specifications for encoding all valid constructs in the working document [10], is a revision to the SLS 1134:1996 standards [3], we believe that it is still not possible to represent not only contemporary Sinhala writing, but also classical writing including Pali and Sanskrit text written in Sinhala script. This paper describes the rendering and collating & sorting issues of the language scripts and thereafter, discusses some of the key issues concerning standardization at the character level.

B. Research Objectives

In this paper, after providing a brief background on Sinhala script, we describe the rendering and collating & sorting issues of the language scripts and thereafter, discuss some of the key issues concerning standardization at the character level. In the context of SLASCII design, the philosophy behind the design of ISO/IEC 10646 compatible SLS 1134:1996 is examined and UNICODE proposals are also discussed in a theoretical manner in every stage of standardization.

Some of the major considerations in the design guideline principles of the character codes for internal representation are: storage requirement for font and data, unification and uniqueness of the code, completeness of code table, right collating sequence, ease of editing, direct and unambiguous transformation of keyboard symbols to the internal representation and the ease of composing the script on the output device.

C. Literature Survey

Sinhala script is used for writing the languages of Sinhala, Pali, and Sanskrit in Sri Lanka, which is said to be derivative of the ancient scripts Brahmi. Originally, the Sinhala scripts were a geometrical and straight line in shape, gradually scripts became rounded at the edges, with notable influence by the *Kadamba* and *Pallawa Grantha* script of South India [11, 12]. The oldest palm leaf manuscripts in existence are from the 10th century onward and by about the 17th century, these manuscripts were written on varied subjects. These manuscripts with its thousands of characters and ligatures present a challenge to western type designers unaccustomed to varying shape character sets. For most of the past 500 years, Latin fonts were founded with character sets sufficient to compose text in one or a few European languages. Although some early fonts were designed with many variant characters and ligatures to emulate the rich variety of Asian scribal handwriting, the triumph of printing in Europe was partly due to the efficiency and economy of text composition from small sets of alphabetic characters.

Similarly, in Sri Lanka, during 1501-1658 Portuguese used Sinhala handwritten scripts in their "*tombo"* (list of records of villages to aid with tax collection), included a few of complex ligatures used in traditional "ola" leaf manuscripts. The founders' typefaces, used in the first book of any size ever printed in Sinhala of 1736. In this book, Gabriel Schade uses vowels, consonants, ligatures, touching characters, numerals and punctuation marks (*Kundaliya*) containing some 64 characters and ligatures. Printing of law-and-order notifications for public awareness was the other type of important document printed at that time using the same founders' types. K.K. Hof, Alphabet [13] of all races of the world of his comprehensive samples register, published in 1876, list 102 characters including 41 ligatures. However, many more important ligatures were not found in this 1876 list to reduce the cost of cutting, founding, composing, and distributing type.

The change from manual to mechanical composition technology at the end of the 19th century did not alter the average size of character sets, nor did the changes or improve the quality of the character set till the introduction of Sinhala Monotype composer [14] in the early 1960s. The Monotype matrix case contained 302 characters, 17 punctuation marks, and 10 numbers. The typewriter, a 19th-century invention that became available for Sinhala only in the mid 20th century, likewise offers small character sets. The *Wijesekara* Sinhala typewriter, popular around Sri Lanka, which has to be the government approved

official typewriter introduced in mid 1960s, provide 68 characters and 14 punctuation marks and 10 numbers.

Having looked at the manual and mechanical composition technology used for printing Sinhala scripts form the past to present, in the following section we examine the syllable structure which is being used to write modern Sinhala, Pali and Sanskrit languages, which is the general practice in the country.

III. Sinhala Language and Early Sinhala Standards

A. *Evaluation of frequency of occurrence of diacritical marks, vowels, consonant, and "medial" signs*

At the time of the development of 8-bit coded character set for SLASCII, the occurrence of the diacritical marks, vowels, consonants, and "medial" signs were considered when storage requirements and the code point limitation as the ASCII pages are concerned. The frequency of occurrence of groups of symbols in Sinhala is given in Table 1-a to 1-c below based on the UCS/UNICODE compatible word list consists of 70142 distinct Sinhala words extracted from the UCSC/LTRL, Sinhala corpus beta version, April 2005. Sinhala characters in Table 1-a has been divided into 2 groups and they consist of 02 diacritical marks (D_n), 18 vowels (V_n), and Table 1-b consist of 41 consonants (C_n), and Table 1-c consist of 03 medial signs ($\underline{C}_n$). Vowels (V_n) will appear only at the beginning of the words. The vowel signs ($\underline{V}_n$) are used to change the inherent vowel and therefore, Vowel signs are ignored and not counted as vowels in this analysis.

According to the table 1-a given below, the occurrences of the vowels V_{11} (ඍ) and V_{12} (ඎ) are zero percent in present usage and are not found in any words or any dictionaries, but allow the vowel sign $\underline{V}_{11}$ (ෘ) and $\underline{V}_{12}$ (ෲ) to appear itself. The vowel V_{10} (ඈ) also does not occur in present usage but its corresponding vowel sign $\underline{V}_{10}$ (ාa) is used. The occurrence of the consonant C_{12} (ඏ) is also zero percent in modern usage and not found in any dictionary (see Table 1-b). According to the results in the Table 1-c below, the usage of consonants C_5 (ඩ) and C_{10} (ඹ) also occurred zero percent in the general writing system, however, the consonant C_{10} (ඹ) used to write *Pali* and *Sanskrit* in the language. Use of medial sign C_3 also not occurred in modern writing; however, it was used in contemporary writing. Medial signs $\underline{C}_1$ and $\underline{C}_2$ are essential though they are not a part of the Sinhala alphabet and used to write wherever they are needed.

This alphabet differs from all other Indo-Aryan languages and contained the special sound that has been unique to itself since the 8th century AD. For example, the presence of the set of five nasal sound known as "*half nasal*" C_6, C_{12}, C_{19}, C_{25} and C_{31} in modern writing system they have occurred only 0.80%, however they cannot be omitted except C_{12} (which is not used at all) since these are essential for the writing system [15].

B. *Cording in terms of 8-bit - SLASCII*

According to the editors of an article published in the Communication ACM in 1963, S. Gorn, W. Bemer and J. Green [16], they have stated, while designing the 7-bit codes for internal representation, collating conventions and other considerations and criteria are intended for the interchange of information among information processing and communication systems and associated equipment. Based upon, the following points have been considered when designing the SLASCII:

1) Automatic sorting has necessitated allocating reserved code position in the code table too. The sorting and collating are one of the most frequently required operations; assignment of code should maintain the collating sequence of the language. This is not very straight forward to achieve. Many a time there is a conflict between logical order of the diacritical marks in the character set and its placement at the code table.
2) The amount of storage required and a number of code positions in the code table were another important consideration. A total number of vowels reduced to 7 from 18 assuming the others can be composed using vowel signs. Therefore, vowel signs are also arranged according to the sorting order. Also, other unwanted consonants were removed from the standard too.
3) It should base itself on the ASCII standards adopted for the English language especially in terms of control characters and escape sequences. This enables the existing software and communication links to be totally compatible.
4) As far as possible there should be a direct correspondence to the existing Sinhala typewriter keyboard.

Table 1-a: Detailed analysis of occurrences of diacritical marks and vowels in modern Sinhala writing

Sinhala Vowels and Vowel Signs in Alphabetical Order and Vowel Occurrences							
Diacritical Marks						Occurrences	%
1	*anus*	D_1	◌ං			2,469	0.737
2	*visar*	D_2	◌ඃ			7	0.002
Vowels (V)				Corresponding Vowel Signs ($\underline{V}$)		Vowels Occurrences	%
1	*a*	V_1	අ	$\underline{V}_1$ (X)	්	4,698	1.403
2	*aa*	V_2	ආ	$\underline{V}_2$	◌ා	1,138	0.340
3	*ae*	V_3	ඇ	$\underline{V}_3$	◌ැ	1295	0.387
4	*aae*	V_4	ඈ	$\underline{V}_4$	◌ෑ	41	0.012
5	*i*	V_5	ඉ	$\underline{V}_5$	◌ි	1,194	0.357
6	*ii*	V_6	ඊ	$\underline{V}_6$	◌ී	92	0.027
7	*u*	V_7	උ	$\underline{V}_7$	◌ු	1,219	0.364
8	*uu*	V_8	ඌ	$\underline{V}_8$	◌ූ	35	0.010
9	*r*	V_9	ඍ	$\underline{V}_9$	◌a	19	0.006
10	*rr*	V_{10}	ඎ	$\underline{V}_{10}$	◌aa	0	0.000
11	*l*	V_{11}	ඏ	$\underline{V}_{11}$	◌ෘ	0	0.000
12	*ll*	V_{12}	ඐ	$\underline{V}_{12}$	◌ෲ	0	0.000
13	*e*	V_{13}	එ	$\underline{V}_{13}$	◌ෙ	1,029	0.307

14	*ee*	V_{14}	ේ	$\underline{V}_{14}$	ේ	190	0.057
15	*ai*	V_{15}	ෛ	$\underline{V}_{15}$	ෛ	12	0.004
16	*o*	V_{16}	ඕ	$\underline{V}_{16}$	ෝ	401	0.120
17	*oo*	V_{17}	ඖ	$\underline{V}_{17}$	ෝ	146	0.044
18	*au*	V_{18}	ඖ	$\underline{V}_{18}$	ෞ	11	0.003

Table 1-b: Detailed analysis of occurrences of consonants in modern Sinhala writing

Sinhala Consonants in Alphabetical Order and Occurrences					
Consonants				Occurrences	%
1	*ka*	C_1	ක	27,289	8.150
2	*kha*	C_2	ඛ	313	0.093
3	*ga*	C_3	ග	10,542	3.149
4	*gha*	C_4	ඝ	206	0.062
5	*nga*	C_5	ඞ	0	0.000
6	*nnga*	C_6	ඟ	662	0.198
7	*ca*	C_7	ච	1,759	0.525
8	*cha*	C_8	ඡ	142	0.042
9	*ja*	C_9	ජ	2,621	0.783
10	*jha*	C_{10}	ඣ	0	0.000
11	*nya*	C_{11}	ඤ	67	0.020
12	*jnya*	C_{12}	ඥ	0	0.000
13	*nyja*	C_{13}	ඦ	268	0.080
14	*ṭta*	C_{14}	ට	12,553	3.749
15	*ṭtha*	C_{15}	ඨ	175	0.052
16	*ḍda*	C_{16}	ඩ	3,673	1.097
17	*ḍdha*	C_{17}	ඪ	23	0.007
18	*ṇna*	C_{18}	ණ	4,630	1.383
19	*nndda*	C_{19}	ඬ	171	0.051
20	*ta*	C_{20}	ත	19,929	5.952
21	*tha*	C_{21}	ථ	786	0.235
22	*da*	C_{22}	ද	14,898	4.450
23	*dha*	C_{23}	ධ	2,231	0.666
24	*na*	C_{24}	න	35,634	10.643
25	*nda*	C_{25}	ඳ	1,237	0.369
26	*pa*	C_{26}	ප	14,173	4.233
27	*pha*	C_{27}	ඵ	71	0.021
28	*ba*	C_{28}	බ	5,432	1.622
29	*bha*	C_{29}	භ	1,269	0.379
30	*ma*	C_{30}	ම	20,848	6.227
31	*mba*	C_{31}	ඹ	595	0.178
32	*ya*	C_{32}	ය	27,943	8.346
33	*ra*	C_{33}	ර	25,087	7.493
34	*la*	C_{34}	ල	16,459	4.916
35	*va*	C_{35}	ව	29,230	8.730
36	*sha*	C_{36}	ශ	2,498	0.746
37	*ṣsa*	C_{37}	ෂ	2,186	0.653
38	*sa*	C_{38}	ස	17,154	5.123
39	*ha*	C_{39}	හ	8,883	2.653
40	*ḷa*	C_{40}	ළ	3,107	0.928
41	*fa*	C_{41}	ෆ	415	0.124

Table 1-c: Detailed analysis of occurrences of medial signs in modern Sinhala writing

Sinhala Scripts in Alphabetical Order and Occurrences					
Medial Signs				Occurrences	%
1	*rakaransaya*	$\underline{C}_1$	්‍ර	3,999	1.194
2	*yansaya*	$\underline{C}_2$	්‍ය	1,660	0.496
3	*repaya*	$\underline{C}_3$	ී	0	0.000

C. Phonetic model SLS 1134:1996

The new, phonetic model design for the Sinhala characters code replaced the older typewriter metaphor concept from the previous SLASCII standard and Sinhala has been standardized under Unicode standard and this encoding uses the hexadecimal code in the range U+0Dxx to U+0Dxx. This process was based on the guideline principles for design character codes and going through an intensive interaction between the Sinhala experts in Sri Lanka. Nevertheless, designing of the codes for SLS 1134:1996, in the following points have been considered:

1) The character code set maintains the logical sequence of the alphabet. An effort has been made to preserve the alphabetical order of the Sinhala language to a greater extent.
2) As Sri Lanka has two official languages in the usage, namely Sinhala and Tamil, for the benefit of users who do transliteration from Tamil to Sinhala, some additional codes positions are reserved to accommodate Tamil.
3) In designing, this code set efforts were also made to retain the flexibility of the language to incorporate future development.

IV. Designing Methodology for ISO/IEC 10646

A. Designing character codes – Guideline principles

According to the R.M.K. Sinha [17], some of the major considerations in designing the code for internal representation are one and only one code for semantically equivalent characters, uniqueness of cording, uniformity in assigning, usage of control

characters, etc. and based on the discussions we had in earlier sections in this paper, following guideline principles are envisaged in the designing of the codes for information interchange for Sinhala scripts of Brahmi origin.

1) ***Completeness*** - All characters should be represented in the code table.
2) ***Unification*** - There should be one and only one code for semantically equivalent characters.
3) ***Uniqueness*** - Two characters which differ in their meaning cannot be assigned to the same code.
4) ***Memory economy*** – The required amount of storage is another important consideration, especially for large information systems. Internal representation needing less space without some overhead of processing time is preferred over those which try to reduce processing time at the cost of storage.
5) ***Compatibility*** - The punctuation marks, numerals, and operators & other universal symbols are assigned the same code across the languages. No character other than the control characters is assigned the role of invoking an action or a role.
6) ***Uniformity*** – It should base itself on the present standards adapted for English computers especially in terms of control characters and escape sequences. This will enable the existing English software and communication links to be totally compatible.
7) ***Separate code where necessary*** - Diacritical marks should be assigned separate codes.
8) ***Easiness of transliteration*** - Transliteration from Sinhala to Tamil should be considered.
9) ***Easiness of sorting*** – Sorting and collating is one of the most frequently required operations, assignment of code which should maintain the collating sequence of the language. Therefore, acceptable and easiness of sorting and collating sequence should be considered.
10) ***Easiness of rendering*** – Special control code(s) should be introduced in Orthographic languages like Sinhala to join two or more consonants to form a single unit (conjunct consonants), alter the shape of preceding consonants (cursiveness of the consonant) and disjoin a single ligature into two or more units.
11) ***Keyboard sequence compatibility*** - As far as possible there should be the direct and unambiguous transformation of keyboard symbols to the internal representation.

Typical character sets of Sinhala language contained vowels (V), consonants (C), *virama* or *is-pilla* (X), vowel signs (V), diacritical marks (D), medial signs (C), punctuations (P), numerals (N), and sometimes special symbols. Sinhala has been standardized under Unicode standard and this encoding uses the hexadecimal code in the range U+0D80 to U+0DFF. This code chart comprises codes for the diacritical marks, vowels, vowel signs, consonants, punctuation mark, and numerals. The symbols used in the Sinhala language consist of consonants (C), vowels (V), *virama* or *is-pilla* (X), vowel signs (V), diacritical marks (D), medial signs (C) and punctuation mark (P) representation can be defined as follows:

Sinhala Language: = <C, V, X, V, D, C, P>; where

<C> := consonants; (41)

<V> := vowels; (18)

<X> := *is-pilla*/*virama*; (01)

<V> := vowel signs; (17)

<D> := diacritical Marks; (02)

<C> := medial signs; (03)

<P> := punctuations Mark;(01)

In addition to the above representation, code points for semantically equivalent graphical shapes for four vowel signs were included to the SLASCII and SLS 1134:1996 and they are defined as follows:

<Z> := Alternative Graphical Signs; (04)

The tables 2-a and 2-b below give the summary of the number of code points assigned for symbols of the language Sinhala in terms of the 8-bit SLASCII and 16-bit SLS 1134:1996 and UCS/UNICODE standards to compare the completeness of the implementation.

Table 2-a: Summary of the number of code points assigned for symbols of the language Sinhala in terms of the 8-bit SLASCII and 16-bit SLS 1134:1996 and UCS/UNICODE

Row No.		8-bit	UNICODE		8-bit	UNICODE		8-bit	UNICODE		8-bit	UNICODE	
		SLASCII	SLS	UCS	SLASCII	SLS	UCS	SLASCII	SLS	UCS	SLASCII	SLS	UCS
	Hex	8x	0D8x		9x	0D9x		Ax	0DAx		Bx	0DBx	
1	0						ප෧		ච	ච	ජ	ධ	ධ
2	1					එ	එ	අ	ඡ	ඡ	ඣ	න	න
3	2		◌ං	◌ං			ඒ	ඉ	ජ	ජ	ඤ		
4	3		◌ඃ	◌ඃ			ඓ	ඊ	ඣ	ඣ	ඥ	ඳ	ඳ
5	4		In			ඔ	ඔ	උ	ඤ	ඤ	ට	ප	ප
6	5		අ	අ			ඕ	ඌ	ඥ	ඥ	ඨ	ඵ	ඵ
7	6			ආ			ඖ	එ	ඡ	ඡ	ඩ	බ	බ
8	7			ඇ				ඔ	ට	ට	ඪ	භ	භ
9	8			ඈ				ක	ඨ	ඨ	ණ	ම	ම
10	9		ෙ	ෙ				ඛ	ඩ	ඩ	ඬ	ඹ	ඹ
11	A		ේ	ේ		ක	ක	ග	ඪ	ඪ	ත	ය	ය
12	B		උ	උ		ඛ	ඛ	ඝ	ණ	ණ	ථ	ර	ර
13	C			ඌ		ග	ග	ඩ	ඬ	ඬ	ද		
14	D		ඌa	ඌa		ඝ	ඝ	ඟ	ත	ත	ධ	ල	ල
15	E			ඌa		ඩ	ඩ	ච	ථ	ථ	න		
16	F		ප	ප		ඟ	ඟ	ඡ	ද	ද	ඳ		

Table 2-b: Summary of the number of code points assigned for symbols of the language Sinhala in terms of the 8-bit SLASCII and 16-bit SLS 1134:1996 and UCS/UNICODE (Continued from previous Table 2-a)

Row No.		8-bit	UNICODE		8-bit	UNICODE		8-bit	UNICODE		8-bit	UNICODE	
		SLASCII	SLS	UCS	SLASCII	SLS	UCS	SLASCII	SLS	UCS	SLASCII	SLS	UCS
	Hex	Cx	0DCx		Dx	0DDx		Ex	0DEx		Fx	0DFx	
1	0	ප	ව	ව		◌ෟ	◌ෟ	ᵖ			◌ී		

2	1	ඓ	ශ	ශ		◌ෟ	◌ෲ	ෝ					
3	2	ඔ	ෂ	ෂ		◌ී	◌ී	ඃ					◌ෘ
4	3	ඕ	ස	ස		◌ු	◌ූ					◌ෘ	◌aa
5	4	ඖ	හ	හ		◌ෙ	◌ෘ	ෟ					
6	5	ක	ළ	ළ				෠					
7	6	ඛ	ෆ	ෆ		◌ෛ	◌ෟ	෡					
8	7	ග						෢					
9	8	ඝ				◌a	◌a	෣					
10	9	ඞ				◌ෝ	◌ෝ	෤			◌ෟ		
11	A	ච	◌ෟ◌ෘ	◌ෟ			◌ෟ	෤			◌ෟ		
12	B	ඡ					◌ෝ◌ෝ	a					
13	C	ජ	Ln				◌ෟ◌ෟ	෧					
14	D	ඣ	Sh				◌ෟ◌ෟ	෦					
15	E	ඤ	◌ෝ				◌ෟ◌ෘ	෪					
16	F	ඥ	◌ෟ	◌ෟ		◌ෘ	◌ෘ	෫					

In the next section, we examine and discuss the guideline principle which is mentioned above early sections, has led to the formation of standard codes assigned by the SLASCII, SLS 1134:1996 and UCS/UNICODE. Table 3 below demonstrates the comparison of formation of standard codes assigned by the SLASCII, SLS 1134:1996 and UCS/UNICODE.

Table 3: Comparison table of the formation of standard codes assigned by the SLASCII, SLS 1134:1996 and UCS/UNICODE

Grouping Based on UNICODE Code Table		No. of Sinhala Symbols	SLASCII	SLS	UCS
Diacritical Marks (D)		2	2	2	2
Consonants (C)		41	40	41	41
Vowels (V)	Pre-composed vowels	18	7	8	18
	Combining vowels		9	10	0
Ispilla (*Virama*) (X)		1	1 (E0)	1	1
	Semantically equivalent graphical shapes		1 (F0 is Graphically coded, Z_1)	0	0
Vowel Sign (V)	Coded		10	10	15
	Composite vowel signs		5	5	0
	In addition to the vowel signs, semantically equivalent graphical shapes		2 (F9 and FA are Graphically coded, Z_2 and Z_3)	1 (0DCE) Graphically coded Z_4	0
	Additional Vowel Sign (V) (◌aa and ◌ෘ)	2	0	1	2
Punctuations (P) (෴)		1	0	1	1
Sinhala Numerals (N) included in SLS 1134 : 2011 and published in UNICODE Ver. 7		10	0	0	10
Medial Sign (C)	*rakaransaya* (C_1 = X + C_{33}) (ූ)	3	2 Graphically coded C_1 and C_2	0	0
	yansaya (C_2 = X + C_{32}) (◌ූ)		0	3	0
	repaya (C_3 = C_{33} + X) (◌)				
	Control Code			In, Ln, Sh	ZWJ ZWNJ
Total code points/Total composed symbols (without Numerals)		83	65/79	68/83	80/80

V. Comparison of Guideline Principles in Standardizing of Sinhala Scripts

A. **Completeness** - All characters should be represented in the code table

NOTES FOR SLASCII:

The coded character set of the SLASCII standard has the following structure.

1) Positions 00 to 7F still have the same meaning as the 7bit codes conforming to the ISO 646.
2) The two columns positions of the 2nd page from Hex 80 to 9F are reserved for control characters as per the recommendation of the International Standard Organization. These control characters may be used to cater to the specific needs of the application.
3) Positions from the third column Hex value from A1 through A7 are assigned for 7 vowels.
4) Positions A8 through CF are assigned for 40 consonants.
5) Positions E0 is assigned for *virama* (◌්), E1 is assigned for *anusvara* (◌ං) and E2 is assigned for *visarga* (◌ඃ) are assigned for diacritical marks. Positions E4 to ED are assigned for 10 vowel signs (◌ා ◌ැ ◌ෑ ◌ි ◌ී ◌ු ◌ූ ◌a ◌ෝ ◌ෘ).
6) Two (02) code points EE and EF are used to assigned code points for medial sings C_1 (◌ූ) and C_2 (ූ) (graphically coded).
7) Three (03) code positions F0, F9 and FA [Z_1 (◌ෟ), Z_2 (◌ූ) and Z_3 (◌ෟ◌ෘ)] are semantically equivalent graphical shapes

for E0, E9 and EA [X (්), V_7 (ු) and V_8 (ූ)] vowel signs however, assigned 3 code points in SLASCII standards.

8) Nine (09) combining vowels are created using corresponding vowel and appropriate vowel sings as follows:

$V_2 = V_1 + \underline{V_2}$ (ආ = අ + ා)

$V_3 = V_1 + \underline{V_3}$ (ඇ = අ + ැ)

$V_4 = V_1 + \underline{V_4}$ (ඈ = අ + ෑ)

$V_8 = V_7 + \underline{V_{11}}$ (ඌ = උ + ෟ)

$V_{10} = V_9 + \underline{V_9}$ (සaa = සa + ාa)

$V_{14} = V_{13} + X$ (ඒ = එ + ්)

$V_{15} = \underline{V_{13}} + V_{13}$ (ඓ = ෙ + එ)

$V_{17} = V_{16} + \underline{Z_1}$ (ඕ = ඔ + ෘ)

$V_{18} = V_{16} + \underline{V_{11}}$ (ඖ = ඔ + ෟ)

9) Five (05) composite vowel signs are entered as a sequence of two or more keys in SLASCII

$\underline{V_{14}} = \underline{V_{13}}+X$ (ොේ = ෙ+ේ)

$V_{15}= V_{13}+V_{13}$ (ෞ = ෙ+ෙ)

$\underline{V_{16}} =\underline{V_{13}}+\underline{V_2}$ (ෞ = ෙ+ෟ)

$\underline{V_{17}}=\underline{V_{13}}+\underline{V_2}+X$ (ොෝ = ෙ+ෟ+ේ)

$\underline{V_{18}}=\underline{V_{13}}+\underline{V_{11}}$ (ොෟ = ෙ+ෟ)

Ex. කේ = ෙ + ක + ේ

කෙ = ෙ + ෙ + ක

කෟ = ෙ + ක + ෟ

කෝ = ෙ + ක + ෟ + ේ

කෟ = ෙ + ක + ෟ

This leaves us five columns (65 symbols), and now it is possible to give an explicit representation of vowels, in addition to the vowel modifier symbols, also allowed us to use 40 consonants except for C_{12} (ඥ) consonants. In a total 79 symbols were able to use with this standard.

NOTES FOR SLS 1134:1996

1) Composite vowel signs are entered as a sequence of two or more keys in SLS:1134:1996 similar to SLASCII

$\underline{V_{14}} = \underline{V_{13}}+X$ (ොේ = ෙ+ේ)

$V_{15}= V_{13}+V_{13}$ (ෞ = ෙ+ෙ)

$\underline{V_{16}} =\underline{V_{13}}+\underline{V_2}$ (ෞ = ෙ+ෟ)

$\underline{V_{17}}=\underline{V_{13}}+\underline{V_2}+X$ (ොෝ = ෙ+ෟ+ේ)

$\underline{V_{18}}=\underline{V_{13}}+\underline{V_{11}}$ (ොෟ = ෙ+ෟ)

2) Positions 0DCA is assigned for *virama* (්), 0D82 (ං) and 0D83 (ඃ) are assigned for diacritical marks. Positions 0DCF to 0DDF are assigned for 10 vowel signs (ා ැ ෑ ි ී ු ූ ෘ ෙ ෟ).

3) *0DCE for* $\underline{Z_4}$ (෴)

4) *0D84 for In, 0DCC, and 0DCD for Ln and Sh*

5) *0DF4 for P* (෵)

NOTES FOR UCS/UNICODE

1) Single Coded 17 Vowel Sign for UCS/UNICODE

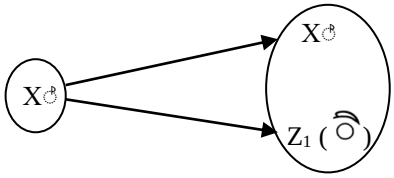
B. **Unification** - There should be one and only one code for semantically equivalent characters

NOTES FOR SLASCII:

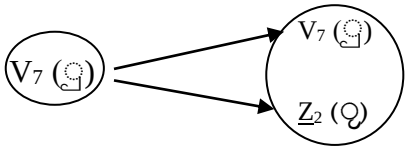
C_{11}	C_{13}
nya	*nyja*
ඤ	ඦ

1) The letters C_{11} (ඤ) (*nya*) and C_{13} (ඦ) (*nyja*) are identical in sound only in the initial position of a word, for example, ඤාණ (*nya+aa+nna*) and ඦාන (*nyja+aa+na*). They are not identical in non-initial positions, where C_{13} (ඦ) behaves like a combination of two consonant sounds, for example, ප්‍රඥා (*p+ra+k+nyja*). Therefore, two codes are assigned for these two consonants.

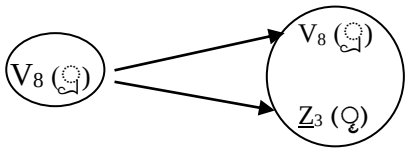
2) The following *virama* sign X (්) has assigned two codes one for X (්) and second for Z_1 (ෘ), although they are semantically equivalent vowel signs in the writing system.



3) The following vowel sign V_7 (ු) has assigned two codes as one for V_7 (ු) and second for $\underline{Z_2}$ (ූ) although they are semantically equivalent vowel signs in the writing system.



4) The following vowel sign V_8 (ූ) has assigned two codes one for V_8 (ූ) and second for $\underline{Z}_3$ (ෟ) although they are semantically equivalent vowel signs in the writing system.



C. **Uniqueness** - No two characters which differ in their meaning be assigned to the same code

D. **Memory economy** – The required amount of storage is another important consideration, especially for large information systems. Internal representation needing less space without some overhead of processing time is preferred over those which try to reduce processing time at the cost of storage.

E. **Compatibility** - The punctuation marks, numerals, and operators & other universal symbols are assigned the same code across the languages. No character other than the control characters is assigned the role of invoking an action or a role

SLASCII	SLS 1134:1996	UCS/UNICPDE
used 8-bit ASCII	Used 16-bit UCS environment	UCS/UNICODE environment

F. **Uniformity** – It should base itself on the present standards adopted for English computers especially in terms of control characters and escape sequences. This will enable the existing English software and communication links to be totally compatible.

SLASCII	SLS 1134:1996	UCS/UNICODE
No additional Sinhala specific control characters. *NBSP* was used for *SP* characters	*In*, *Sh* and *Ln* were introduced	Replaced by ZWJ and ZWNJ inserted of *Ln* and *Sh*. In was missing.

Codes are not provided in the code table in SLS:1134 and UCS/UNICODE for distinct formations in the language for the medial signs $\underline{C}_1$ (*rakaransaya)*, $\underline{C}_2$ (*yansaya)*, $\underline{C}_3$ (*repaya).* However, these shapes could be generated by using relevant combinations given herein.

In SLASCII

$C_1\underline{C}_1 = C_1 + \underline{C}_1$ (ක්‍ර = ක + ෟ) (*rakaransaya)*

$C_1\underline{C}_2 = C_1 + \underline{C}_2$ (කය = ක + ්‍ය) (*yansaya)*

$\underline{C}_3$ (ි) (*repaya)* was not coded in this table.

In SLS 1134:1996

Provision for joint character is made through '*Ln*' (stand for Link) and '*Sh*' (stand for Short) keys; for example:

$C_{24}C_{22} = C_{24} + X + Ln + C_{22}$ (සද=ස+්+Ln+ද)

$C_1C_{37} = C_1 + X + Ln + C_{37}$ (කෂ=ක+්+Ln+ෂ)

$C_{20}C_{35} = C_{20} + X + Ln + C_{35}$ (තව=ත+්+Ln+ව)

$\underline{C}_3 = C_{33} + X + Sh$ (ි=ර+්+Sh)

In UCS/UNICODE:

In this version 'Link Key' and 'Link code' (Ln), 'Short Key' and 'Short code' (Sh) and 'Invisible Key' and 'Invisible code' (In) are omitted as these keys are already included in the international standard, in a different page.

Inserted, the two special characters, U+200C ZERO WIDTH NON-JOINER (ZWNJ for short) and U+200D ZERO WIDTH JOINER (ZWJ for short), can be used as hints of which glyph shape is preferred in a particular situation. ZWNJ prevents the formation of a cursive connection or ligature in situations where one would normally happen, and ZWJ produces a ligature or cursive connection where one would otherwise not occur. These two characters can be used to override the default choice of glyphs. Join characters are made through ZWJ (U+200D) code, for example:

$C_{24}C_{22} = C_{24} + X + ZWJ + C_{22}$

(සද = ස + ් + ZWJ + ද)

$C_1C_{37} = C_1 + X + ZWJ + C_{37}$

(කෂ = ක + ් + ZWJ + ෂ)

$C_{20}C_{35} = C_{20} + X + ZWJ + C_{35}$

(තව = ත + ් + ZWJ + ව)

$S_3 = C_{33} + X + ZWJ$ (ි = ර + ් + ZWJ)

Non Join characters are made through ZWNJ (U+200C) code, for example:

$C_{24}XC_{22} = C_{24} + X + ZWNJ + C_{22}$

(ස්‌ද = ස + ් + ZWNJ +ද)

$C_1XC_{37} = C_1 + X + ZWNJ + C_{37}$

(ක්‌ෂ = ක + ් + ZWNJ +ෂ)

$C_{20}XC_{35} = C_{20} + X + ZWNJ + C_{35}$

(ත්‌ව = ත + ් + ZWNJ +ව)

However, the default case in Sinhala prevents the formation of a cursive connection or ligature in situations where one would normally happen.

G. **Separate code where necessary** - Diacritical marks should be assigned separate codes

SLASCII	SLS:1134	UCS/UNICODE
D_1 (ං) D_2 (ඃ)	D_1 (ං) D_2 (ඃ)	D_1 (ං) D_2 (ඃ)
V→*C*→X →*D*→V→S	*D*→*V*→*C*→V→P	*D*→*V*→*C*→V→P
Prevent by entering the wrong order. Sorting need additional code ordering tools.	Sorting makes easier.	Sorting makes easier.

H. **Easiness of transliteration** - *Transliteration from Sinhala to Tamil should be considered*

SLASCII	SLS:1134	UCS/UNICPDE
Not considered	Considered The 0DB2, 0DBC, 0DBE and 0DBF Code positions are reserved for Tamil characters	Considered The 0DB2, 0DBC, 0DBE and 0DBF Code positions are reserved for Tamil characters

NOTE:

Special codes for Tamil characters could be utilized to transliterate scripts from Tamil to Sinhala. The character positions given in code tables both SLS:1134 and UCS/UNICODE for Tamil letters ன *(nnna)*, ற *(rra)*, ள *(lla)* and ழ (*llla)* could be represented 0DB2, 0DBC, 0DBE and ODBF respectively. However, corresponding Sinhala characters are to be designed and implemented.

I. **Easiness of sorting** – Sorting and collating is one of the most frequently required operations, assignment of code which should maintain the collating sequence of the language. Therefore, acceptable and easiness of sorting and collating sequence should be considered.

Collation is defined as the culturally expected ordering of linguistic characters in a particular language. This culturally expected ordering allows users to define the structure and find data in a way that is consistent for their particular language. This is not very straight forward to achieve. Many a time there is a conflict between logical order of the diacritical marks in the character set and its placement at the code table. Automatic sorting has necessitated allocating reserved code position in the code table too.

Each Sinhala letter is represented by a sequence of symbols in the code tables. A letter may be a vowel (V_{13} = එ), a consonant (C_1 = ක), a consonant followed by a *virama* (C_1X = ක්), consonant with a vowel sign ($C_1\underline{V_2}$ = කා), consonant with a combined vowel sing ($C_1\underline{V_{16}}$ = කෝ), consonant with more than individual vowel sign as coded in SLASCII and after re-ordering process implies ($C_1V_{13}V_2X$ = කෝ්), a conjunct ligature used in UCS/UNICODE ($C_1X+ZWJ+C_{37}$ = ක්ෂ) or conjunct ligature used in SLS 1134 ($C_1X+Ln+C_{37}$ = ක්ෂ) or medial sign used in SLASCII ($C_1\underline{C_1}$ = ක්‍ර) and one of the above followed by a diacritical mark ($C_1\underline{V_2}D_1$ = කාං).

Though the Sinhala language is based on a complex phonetic structure, the alphabetical order of the consonants (C_1 …. C_{41}) are well defined as:

Sort_key_1: {$C_1 < C_2 < C_3$ …… C_{41}}

The orders of the vowels (V) are arranged by tradition and contemporary. It can be defined as:

Sort _key_2: {$V_1 < V_2 < V_3$ …….. V_{18}}

In the case of vowel signs ($\underline{V}$), they are graphical signs which are always used in conjunction with consonants. It can be defined as:

Sort_key_3: {Ø < $\underline{V_1}$ (X) < $\underline{V_2}$ < $\underline{V_3}$ ……. $\underline{V_{18}}$}

In the case of *virama* (X) has no corresponding vowel syllable, but it will be used to remove the inherent sound /a/ from the consonant. Therefore, it will be treated as a special vowel sign within the language and defined as:

Sort_key_4: {X}

The Sinhala alphabet has two diacritical marks (D) and they used only in conjunction with vowels and consonants. They may appear only flowed by a vowel or consonant with an implicit or explicit vowel. Therefore, their lexicographical order defined as:

Sort_key_5: {Ø < D_1 < D_2}

Three special symbols known as Medial Signs ($\underline{C}$) order define as:

Sort_key_6: {$\underline{C_1}$ < $\underline{C_2}$ < $\underline{C_3}$}

J. **Easiness of rendering -**

As shown in the table below, independent vowels combine with consonants in different ways. In single-byte or double bytes Sinhala text, the vowels that append to the left are written first followed by a consonant. The vowels that append to the right, above or below are written after the consonant. However, the logical order in both cases is the same, i.e. the consonant is followed by the vowel. Therefore, the legacy symbol, re-ordering will be required before string comparisons can be performed for sorting.

For example, Key-in sequence and code reordering for C, $\underline{V}$, $\underline{C}$, and D vowel signs combinations for SLASCII, SLS:1134:1996 and UCS:UNICODE as followed:

For SLASCII

Key-in sequence = Example 1: ක්‍රෝං = ෙ ක ් ර ෝ ං

Example 2: ක්ෂ්ම = (Not implemented)

After Re-ordering = Example 1: ක්‍රෝං = ක ් ර ෙ ෝ ං

Example 2: ක්ෂ්ම = (Not implemented)

For SLS 1134:1996

Key-in sequence = Example 1: ක්‍රෝං = ෙ ක ් ර ෝ ං

Example 2: ක්ෂ්ම = ෙ ක Ln ෂ ් ම

After Re-ordering = Example 1: ක්‍රෝං = ක ෙ ර ෝ ් ං

Example 2: ක්ෂ්ම = ක ් Ln ෂ ේ ම

For UCS/UNICODE

Key-in sequence = Example 1: ක්‍රෝං = ෙ ක ් ර ෝ ං

Example 2: ක්ෂ්ම = ෙ ක ්‍ර ෂ ් ම

After Re-ordering = Example 1: ක්‍රෝං = ක ් zwj ර ෝ ං

Example 2: ක්ෂ්ම = ක ් zwj ෂ ේ ම

NOTES:

SLASCII

1) Only 7 vowel letters are represented by single code point (V_1, V_2, V_3, V_4, V_5, V_6, and V_7) and other vowels are represented by the vowel and corresponding vowel signs; V_2 = V_1 + $\underline{V_2}$ (ආ = අ + ා).

2) Independent vowel letters V_{11} (ඵ) and V_{12} (ඵෘ) do not occur in present usage; therefore these two were not included in the code table.

3) Independent vowel letter V_{10} (ඖaa) also does not occur in present usage, but its corresponding vowel sign $\underline{V}_{10}$ (ෞaa) is used, for example, C_{20} (ඨ) and it is represented by a combination of $\underline{V}_9$ (ෟa); for example $C_{20}\underline{V}_9\underline{V}_9$ = ඨaa.
4) The consonant C_{12} (ඬ) does not occur in present usage; therefore this was not included in the code table.
5) Different graphical form $\underline{Z}_2$ (ු) corresponding with vowel sign $\underline{V}_7$ (ු) given separate code.
6) Different graphical form $\underline{Z}_3$ (ූ) corresponding with vowel sign $\underline{V}_8$ (ූ) given separate code.

SLS 1134:1996

1) Independent vowel letter V_{11} (ඔ) and V_{12} (ඖ) do not occur in present usage, but they are included in the code set for completeness of the order of the code table.
2) Corresponding vowel signs $\underline{V}_{11}$ (ෟ) and $\underline{V}_{12}$ (ෟ) for V_{11} (ඔ) and V_{12} (ඖ) are included in the code set for completeness of the code.
3) Independent vowel letter V_{10} (ඖaa) also does not occur in present usage, but its corresponding vowel sign V_{10} (ෞaa) are used, for example, කර්තaaහු.

UCS/UNICODE

4) Complete consonant set considered and every consonant in the alphabet has its own codes.
1) Complete vowel set considered and every vowel in the alphabet has its own codes.
2) The *virama* X (්) has a higher order of the vowel signs.
3) No alternative vowel signs and Before the vowels
4) There are five vowel signs ($\underline{V}_{10}$ to $\underline{V}_{14}$) have glyph pieces which stand on both sides of the consonant; they follow the consonant in a logical order and should be handled as a unit for most processing.

K. **Keyboard sequence compatibility** - Most direct and unambiguous transformation of keyboard symbols to the internal representation is integral.

There have been a number of studies for standardization of Sinhala keyboard for use with electronic Devices. Sri Lanka Standard Institute has published 'the standard keyboard' in 1989. The standard keyboard layout had been designed in such a way that all characters that are to be used for interchanging Sinhala to be represented by 46 keys as given in Figure 1 below. The following principles were considered when designing the standard keyboard for Sinhala. These are as follows: 1. Every key on the keyboard is precious; 2. The keyboard should be easy to remember; 3. Punctuation marks and other editorial characters which are necessary for documenting Sinhala writing, but not given in the keyboard layout, may be to be used using common English plane; 4. Latin signs and symbols used in Sinhala and language should be identified and should have the same keyboard location as in English keyboard; and 5. Each key represents two strokes, one at the normal position and the other at the shift lock position and few characters represent with shift lock with right alt key combination.

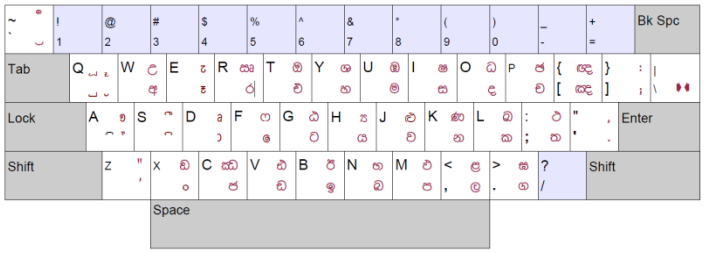


Figure 1. Standard Sinhala Keyboard Layout as defined by the SLS 1134:1996

I. CONCLUSION

We surveyed and compared the designing methodologies, rendering issues, implementations stages of Sinhala scripts and analyzed available OpenType fonts and their rendering schemes for use with modern Sinhala script. The study produced guidelines for designing UCS/IEC code scheme for *Brahmi* based Sinhala languages, which are not documented in international standards. With this design of UCS/IEC code scheme, one of the major opportunities of information interchange is the ability to transliterate from one script to another with little effort, and the other major utility of the unified standard code is to provide the ability to create mixed script documents. Therefore, it is important that there is uniformity in codes for control characters, numerals, and punctuations, mathematical and graphical symbols which are not discussed in this paper. All Sinhala fonts failed to display the variant glyphs with free variant selectors correctly, which were already standardized in Unicode 10646.

ACKNOWLEDGMENT

The study was made possible by the Asian Language Resource Network Project of the Nagaoka University of Technology with financial support of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

REFERENCES

[1] S. T. Nandasara, "From the Past to the Present: Evolution of Computing in the Sinhala Language", IEEE Annals of the History of Computing, IEEE Computer Society, Vol. 31. No. 1, ISSN:1058-61880, 2009. pp. 32-45.
[2] V. K. Samaranayake, J. B. Disanayaka, and S. T. Nandasara, "A Standard Code for Sinhala Characters", presented at the 9th Annual Sessions of the National Computer Conference, Society of Sri Lanka, Colombo, 1989.
[3] SLS 1134:1996, Sri Lanka Standard SLS 1134:1996-Sinhala Character Code for Information Interchange, SLSI Publication, 1996.
[4] Nandasara, S. T., "Proposed Sri Lankan Sinhala Standard Code for Information Interchange (SALASCII)", Approved for the Computer and Information Technology Council of Sri Lanka (CINTEC) and Submitted to the Sri Lanka Standards Institute, 1991
[5] The Unicode Consortium, The Unicode Standard 12.1, Addision-Wesley, 2019.
[6] P. Constable, Proposal on Clarification and Consolidation of the Function of ZERO WIDTH JOINER in Indic Scripts, Public Review Issue #37, Microsoft, USA, 2004-06-30. https://www.unicode.org/review/pr-37.pdf
[7] S. T. Nandasara, V. K. Samaranayake, J. B. Disanayaka, E. K. Seneviratne, T. Koanantakool, "Draft 7 bit Standards Code for the Use of Sinhala in Computer Technology (SLASCII)", 1990.

[8] V.K. Samaranayake and S.T. Nandasara, *A Standard Code for Information Interchange in Sinhalese, Technical Report,* ISO-IEC JTC1/SCL/WG2 N673, Oct. 1990.

[9] Nandasara, S. T., "Proposed Sri Lankan Sinhala Standard Code for Information Interchange (SALASCII)", Approved for the Computer and Information Technology Council of Sri Lanka (CINTEC) and Submitted to the Sri Lanka Standards Institute, 1991.

[10] SLS 1134:2004, Sri Lanka Standard SLS 1134:2004—Sinhala Character Code for Information Interchange, Revision 2, SLSI publication, 2004.

[11] F. Coulman, The Blockwell Encyclopedia of Writing Systems, Blackwell, 1996, p. 469.

[12] P.E.E. Fernando, "Paleographical Development of the Brahmi Script in Ceylon from 3rd Century B.C. to 7th Century A.D.", *University of Ceylon Rev.*, vol. 7, no. 4, 1949, pp.282-301.

[13] K.K. HOF- UND STAATSDRUCKEREI IN WIEN, ("Alphabet of all race of the world"), Royal Print Shop in 1876, Vienna, Germany, 1876.

[14] Monotype, Book of Information. New ed. 1959, London.

[15] J. B. Disanayaka, 1991, "The Structure of Spoken Sinhala Sound and their Patterns", National Institute of Education, Sri Lanka.

[16] S. Gorn, W. Bemer, and J. Green, American Standard for Information Interchange, Communication ACM, Vol. 6. No. 8, 1963.

[17] R. M. K. Sinha, 1992, "Standardizing Linguistic Information – An Overview", Proceeding of the Second Regional Workshop on Computer Processing of Asian Languages, Tata – Mc-Grow Hill, New Delhi, pp. 272-290.

AUTHORS

First Author – S. T. Nandasara, qualifications, University of Colombo School of Computing, Colombo, Sri Lanka, nandasara@yahoo.com.
Second Author – Yoshiki Mikami, Dr., Professor, Nagaoka University of Technology, Nagaoka, Japan, mikami@ksj.nagaokaut.ac.jp.

Correspondence Author – S. T. Nandasara, nandasara@yahoo.com, stn@ucsc.cmb.ac.lk, Telephone Office: +94 11 258 1247, Mobile: +94 77 383 2934.