Nagaoka University of Technology

Graduate School of Engineering

Development and Standardization of Sinhala Script Code for Digital Inclusion of Native Computer Users

(母語によるコンピュータ利用者のすそ野拡大を目指したシンハラ文字コード標準の開発)

This dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Engineering

SHAKRANGE TURRANCE NANDASARA

September 2019

This page intentionally left blank

Short Summary of the Dissertation

The international scope of computing, information interchange, and electronic publishing created the need for a standardized character-encoding scheme. This research dissertation is dedicated to studying 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 their scripts. In this dissertation, the author reports the outcome of the extensive studies on significant issues involved in machine representation of textual information in graphical and phonetic approach for Sinhala scripts. A comprehensive evaluation of the possible character representations is also presented in this work. Further, the design philosophy of Sri Lanka Sinhala Standard Code for Information Interchange (SLASCII) based on ISO 646 considered along with the phonetic model of UCS/Unicode and ISO/IEC 10646 philosophies are discussed. Sinhala UCS/Unicode code standard has been published in its version 3.0, and Sinhala enabled operating systems, and Sinhala application compatibilities are becoming available since then. Character codes require fonts that provide visual images—glyphs—corresponding to the codes in both 8 bits and UCS/Unicode, and it should appear on the screen or paper by the language of Sinhala, Pali and Sanskrit using this script with an acceptable and comprehensive manner. This dissertation concludes by identifying critical issues concerning standardization at the character level. Lastly, the context of SLASCII and SLS 1134:1996 and philosophy behind the design of ISO/IEC 10646 proposal is also discussed.

This page intentionally left blank

Acknowledgment

I would like to make it an opportunity to express my deepest gratitude to the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan for partly sponsoring this research work. Further, I would like to extend my humble gratitude and thank my supervisor Prof. Takashi Yukawa, Nagaoka University of Technology, for his invaluable advice, guidance, the enthusiasm of this thesis. It is also incumbent on me to express my utmost gratitude to Prof. Yoshiki Mikami, Nagaoka University of Technology, my long-lasting collaborator and have supervised and mentored this doctoral study to the end, for his advice, fruitful discussions, and continuous encouragement during the preparation of this thesis. I take great pleasure in acknowledging Dr. AIC. Mohideen, School of Engineering, University of Aberdeen, UK, who kindly read through this thesis and provided valuable suggestions and corrections. Finally, a word of thanks goes to the technical staff and secretaries of the Nagaoka University of Technology, who assisted me during this work at some point.

This page intentionally left blank

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification to this, or any other university, other institutes of learning, or industrial organization.

Shakrange Turrance Nandasara
September 2019

This page intentionally left blank

List of Abbreviations

ASCII	– American Standards Code for Informmation Interchange
ANSI	– American National Standards Institute
BMP	– Basic Multilingual Plane
CANLIT	– The Committee on Adaptation of National Languages in Information Technology
CGA/EGA	– Colour Graphics Adapter/Enhanced Graphics Adapter
CINTEC	– Computer and Information Technology Council of Sri Lanka
DOS	– Disk Operating System
DTP	– Desktop Publishing
FpDAM	– Final Preliminary Draft Amendment
GTS	– Government Technical School
GIST	– Graphics and Intelligence based Script Technology
IBM	– International Business Machine
ICT	– Institute of Computer Technolgy
IEC	– International Electrotechnical Commission
ISO	– International Organization for Standardization
LRTL	– Language Research Technology Laboratory
NARESA	– National Resources, Energy, and Science Authority of Sri Lanka
OCLC	– Online Computer Library Centre
pDAM	– Preliminary Draft Amendment
SBIOS	– Sinhala Basic Input/Output System
SLASCII	– Sri Lanka Sinhala Standard Code for Information Interchange
SLS	– Sri Lanka Standards
SLSI	– Sri Lanka Standard Institute
TSR	– Terminate and Stay Residence
UCD	– Unicode Character Database
UCS	– Universal Coded Character Set
UTF	– Unicode Transformation Format
VGA	– Video Graphics Array
ZWJ	– Zero Width Joiner
ZWNJ	– Zero Width Non-Joiner

This page intentionally left blank

Abstract of the Dissertation

The Sinhala writing system, which is used in Sri Lanka, is a syllabic writing system derived from Brahmi. Sinhala is consist of vowels, consonants, diacritical marks and special symbols constructs. Several of these constructs are combined to form complex ligatures. In the Sinhala language, the total number of different glyphs is almost close to 2300. Thus, all computer equipment that supports the Sinhala language needs to support a greater degree of complexity in memory management, output, display and printing with near minimal changes to the keyboard or the input systems.

The International scope of computing, information interchange, and electronic publishing have created a need for a standardized character-encoding scheme. In support of the standardization momentum, in order to enable Sinhala language and their scripts, 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, requires an extensive study. This research presents an extensive study on the significant 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 which also is based on the typewriter metaphor, considered along with the phonetic model of the SLS 1134:1996 following the guidelines of UCS/Unicode and ISO/IEC 10646 philosophies are also presented.

One of the outputs of this research, the Unicode based Sinhala standard was formulated through a comprehensive and theoretical manner, the same was submitted to the Sri Lanka Standard Institute and forwarded to the international committees for further investigate. Having had lengthy discussions with international language experts, and also considering a public comment received from the local experts, Sinhala UCS/Unicode code standard was published in its version 3.0, and Sinhala enabled operating systems, and Sinhala application compatibilities are becoming available since then. Character codes require fonts that provide visual images—glyphs—corresponding to the codes in both 8 bits and UCS/Unicode, and it should appear on the screen or paper by the language of Sinhala, Pali and Sanskrit using this script with an acceptable and in a comprehensive manner. This research concludes by identifying some of the critical issues concerning the standardization at the character level and therefore, designed the standard guideline principles for standardization. Further, the context of SLASCII and SLS 1134:1996 and philosophy behind the design of ISO/IEC 10646 proposals are also studied.

This research, identifies (1) the rich diversity of the languages in the Asian region, (2) discusses the historical background of the Sinhala writing system, (3) studies the Sinhala scripts' characteristics and complexities, (4) illustrate how Sinhala computing technology has evolved over the last quarter-century and analyzed (5) design of standards for Sinhala scripts for deferent technological stages from single-byte to multi-bytes environment. The design of Sinhala character code standards marks major steps as a cornerstone of whole architecture for text processing in Sri Lanka. Additionally, this study shows how small communities of non-Roman script users can connect to ("Digital Inclusion") the Romanised system dominated cyberspace over the past 30 years.

Contents

Figures

This page intentionally left blank

Tables

This page intentionally left blank

Chapter 1: Digital Inclusion, Benefits and Disadvantages in the Asia Region

The chapter introduces the Digital Inclusion in general and moves onto briefing on Inclusion by Language in the world: Talks about the origin and differences in languages and the challenges in digitizing them. Provides an introduction to Asian Languages and shows how they differ from western languages, dedicates a whole chapter for Sinhala to compare mainly the differences and similarities of Sinhala with other languages, and the benefits of digitizing Sinhala.

1.1 Introduction

Multilingualism is becoming a vital social concept amongst many advanced communities around the globe. One of the significant attributes of multilingualism on the global scale is the ubiquitous e-content and the knowledge explosion. Over the past three to four decades, working groups from the west to east, north to south, south-east and Pacific region have been exploring the possibilities of joint international agenda in the field of multilingual and multicultural society with sufficient emphasis on interoperability issues, global resource discovery, metadata issues, multilingual information access, and multilingual multimedia databases. Out of these areas, relating to its rich diversity of languages and commonality issues, special consideration to the standards, interoperability and multilingual resources creation, is given in the Asian region.

1.2 Language Diversity

In the Asian continent, several language families exist. Among them exist the Austroasiatic, Austronesian, Dravidian, Indo-Iranian, Mongolian, Semitic, Sino-Tibetan, Thai-Kadai, Turkic, and Tungus. However, some of the language families not firmly established; hence, they could be regrouped into the broader languages or could divide into smaller groups. As an example, Altacic family can be a family for languages such as Turkic, Mongolian, and Tungus. The Indo-Iranian language family can divide into the sub-family such as Indo-Aryan, Iranian, and Kafiri. In addition to the above language families, there are also some isolated languages around the Asian continent, e.g. Korean, Japanese, Ainu, and Burushaski. Some European languages – English, Russian, French, and Portuguese – are also used in the region as official languages, and from the mixture of an indigenous language and an introduced language, pidgins or creoles have emerged. Among those language families, Sino-Tibetan has the most significant number of speakers, estimated at 1.3 billion in China alone. Next family, Indo-Iranian, with at least 700 million speakers in India, and more than 200 million people in Pakistan, Bangladesh, Iran and other South and Middle East Asian countries. Malay in the Austronesian language family has around 250 million speakers in Indonesia, Malaysia, Brunei, Singapore, the southern Philippines, and Thailand. Tamil, a Dravidian family, has speakers in India. Semitic includes a language of many speakers, that is, Arabic, the number of which is estimated to be about 200 million. Other language families have a relatively small number of speakers. Among the isolated languages, the Japanese have the most significant number of speakers with about 128 million and Korean follows with about 75 million. When describing the Asian languages, one cannot avoid mentioning the diversity of scripts they use. Contrasted with Western Europe, diversity is outstanding. In Southeast and South Asian countries, most of the scripts derived from the Brāhmī, and in the East and Near East Asian countries, Hanzi script and some other indigenous scripts used. Latin, Arabic and Cyrillic script also used with some additional letters and diacritical marks.

1.3 Language Diversity in the Internet

In profiling characteristics of Internet users versus world population of that language in 2019, the available data is that the number of English-speaking users recorded at 1,105 million (25.2%), followed by Chinese-speakers, at 863 million (19.3%) and then drops to 344 million (7.9%) for Spanish users—in a total user base of 4.3 billion—see Figure 1. (World Stats, 2019; W3Techs, 2018).

Further, according to the Internet World Stats-2019, out of the estimated 97 million individuals in the world that speak German, 95.1% are internet users, out of the estimated 126 million individuals speak Japanese, 93.5% are internet users, out of the estimated 143 million persons speak Russian, 76.1% are internet users and out of the estimated 1,485 million persons speak English, 74.5% are internet users. The content available on the internet to these users, English leads at 54%, with an immediate plunge to Russian at 6%, German at 5.9%, Spanish at 4.9%, French at 4% and Chinese at 1.7% which is much less compared to the 3% in 2015. Similarly, the English content is at 2% less compared to the 2015 statistics (World Stats, 2015).

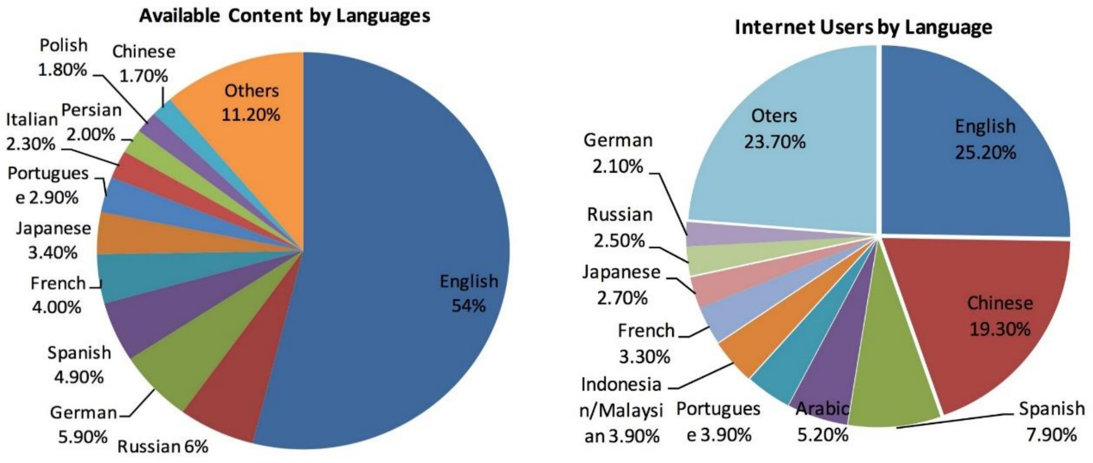


Figure** 1*: ***Internet content available by languages and Internet users by languages

The statistics on the growth rates of the use of languages also provide more insights into Internet users. The growth rate in the number of English-speaking users has continued steadily at a rate of about 685.7% from 2000 to 2019, and other global languages are significantly overshadowing it. Arabic grew by 8,917.3%, and Russian and Chinese grew by 3,434.0% and 2,572.3%, respectively, with other languages showing considerable growth in the same period—for example, Portuguese at 2,164.8%, and Indonesian/Malaysian at 2,861.4% and Chinese at 2,572.3% (World Stats, 2019; W3Techs, 2018). This trend mirrors the composition of the automatic language translation domain during the same period.

1.4 Asian Writing Systems and Script Diversity

In the Asian region, it displays a wide variety of genetically and typologically diverse languages. Some language families have spread through different parts of Asia. According to Ethnologue (Ethnologue, 2005), the number of languages used worldwide runs more close to seven thousand (6912). Out of this more than 3070 languages (44.35%) are spoken in Asia & Pacific countries (*see* Table 1). Even in a single country, the Papua New Guinea (PNG), only using 0.30% of the land from the earth, is said to have about 820 indigenous languages (11.86%). Second to the PNG, in the South East Asian region, 742 languages are used in Indonesia, while they mostly use Roman script to write their language. In South Asia, more than six per cent (6.18%) of the world languages exits in India alone, and where many scripts are used to write, but mainly Devanagari. The Devanagari script used along with other scripts or exclusively used to write several languages including Hindi, Sanskrit, Marathi, and Nepali. (*see* Table 2)

Table 1: Top 10 countries with the richest language diversity by continents

Country	Number of Languages & Total Land Area (in Thousand sq km.)	Percentage of languages & (Land Area)
PNG	820 (452,860)	11.86% (0.30%)
Indonesia	742 (1,826,440)	10.73% (1.23%)
India	427 (2,973,190)	6.18% (2.00%)
Australia	275 (7,617,930)	3.98% (5.11%)
China	241 (9,326,410)	3.49% (6.26%)
Philippines	180 (298,170)	2.60% (0.20%)
Malaysia	147 (328,550)	2.13% (0.22%)
Russia	129 (16,995,800)	1.87% (11.41%)
Nepal	125 (136,800)	1.81% (0.09%)
Myanmar	109 (657,740)	1.58% (0.44%)

In addition to its rich language diversity, Asia is especially rich in scripts and writing style. Three types of scripts originated in three separate geographic areas. Brāhmī script originated in

South Asia in and around the Indian subcontinent and the continental part of Southeast Asia; Chinese ideographs originated in East Asia, and the Aramaic script originated in Southwest Asia and spread to West and Central Asia.

Table 2: Languages using Devanagari script

Language	**User Population**
Hindi	More than 480 million speakers
Marathi & Nepali	More than 10 million speakers
Awadhi, Bhojpuri, Braj-Dhasha, Chahattsigarhi, Konkani, Kachchhi, Marwani, Maithali & Magahi	More than 1 million speakers
Bagheli, Bathi, Bateri, Bhatneri, Bhili, Bihari, Garhwali, Gondi, Harauti, Ho, Jaipuri, Khadiya, Khorhti, Kanauji, Kankan, Kului, Kumaoni, Kurku, Kurukh, Kurmali, Limbu, Mundari, Nagpuri, Newari, Palpa, Panchpargania, Santali & Sherpa	Less than 1 million speakers
Sanskrit	Scholars’ language

Language writing systems played significantly different cultural roles in the Asian region than in many other parts of the ancient world. In general scripts and writing systems in traditional South Asian culture never achieved the status and influence that in attained in many others such as those of the ancient South East, the Islamic world, or China. Many languages of contemporary South Asia belong to the Indo-Aryan (or Indic) branch of the Indo-Iranian family. The earliest – definitely be datable written documents in South Asia are the inscription of the Emperor Aśoka (mid-3rd century B.C.). The history of South Asian scripts consists mainly of the development and regional diversification of Brāhmī script (Falk, 1993), which became the ancestor of many of scripts of South, South-east, and Inner Asia. Eventually, Brāhmī script developed into distinct scripts, often associated with particular languages of the Indo-Aryan, Dravidian, and other families, however, in systemic terms, the Indic scripts typically share the same basic principles of the *akṣara* system, i.e. a modified consonantal syllabary representing most vowels by diacritic signs attached to the consonants. In recent centuries, under Islamic and European influence, writing style and use of Roman script have been introduced.

John Clews says, South Asian and Southeast Asian scripts also represent vowel sounds by the vowel-signs left, right, above or below the consonants, but in this case, their use is mandatory (John Clews, 1997). They derive from the Brāhmī script used on the Indian sub-continent many centuries ago and written from left-to-right. Unlike European scripts, several of these scripts also combine letters as ligatures or conjunct consonants.

Ultimately from the same Brāhmī script, particularly true for Sinhala, few other languages in the region also add a considerable number of additional consonants and vowels to the repertoire. By comparison, some of these scripts lack the independent vowels that all other South Asian scripts have. Indeed, Tamil has far fewer vowels and consonants than most other scripts used in the Indian sub-continent. The later history of South Asian scripts consists mainly of the Brāhmī script, which was the ancestor of literally dozens of Asian scripts. Early regional varieties of Brāhmī eventually developed into distinct scripts often associated with particular languages of the Indo-Aryan, Dravidian, and other families. However, in precise terms, these scripts typically share the same basic principles of the script system—that is, “diacritical modified consonant syllabic scripts.”

Brāhmī, as developed in India and as exported to other parts of Asia in the first millennium C.E., is the ultimate source not only of all of the indigenous scripts of South Asia (i.e. Telugu, Malayalam, Sinhala) but also of the major Southeast Asian scripts (Burmese, Thai, Lao, Khmer, etc.), or Tibetan.

The spread of Brāhmī script into Southeast Asia generally established by the earliest known examples of some brief inscriptions on precious objects discovered in southern Vietnam. However, no concrete evidence that the objects inscribed in Southeast Asia since the area was an important trading center during the period from the second to the fifth century A.D. Note that the types of script that have spread in Southeast Asia all seem to have originated from the Pallava and Grantha scripts of the south of India. These include those of the Burmese, the Lao, the Khmer, the Thai; the obsolescent national scripts of the Mon, the Javanese, the Sundanese, the Balinese, and the Cham, among others, in addition to tribal scripts in Sumatra and the Philippines. (Daniels et al.1996)

The Brāhmī script is written from left to right (though several specimens are running from right to left have been found). By about third century A.D. several distinct regional sub-varieties had arisen, and these continued to differentiate until, by around 1000 A.D, the situation approximated the modern picture in which the Brāhmī-derived scripts have developed to the point that they are in effect independent scripts whose common ancestry may not be apparent to the casual observer. Figure 2, below show how 3rd century B.C. Brāhmī script evolved to stylish modern shapes of many languages used in the contemporary writing style mainly in the Indian region.

On the other hand, one can see these complex script languages are belonging to several language families on the Asian continent: Austroasiatic, Austronesian, Dravidian, Indo-Iranian, Mongolian, Semitic, Sino-Tibetan, Thai-Kadai, Turkic, and Tungus. Some of these language families are not firmly established and could be regrouped into larger language groups or divided into smaller sub-groups. For example, the Turkic, Mongolian and Tungus language families regrouped into a larger language family. Altaic and the Indo-Iranian language family divided into Indo-Aryan, Iranian, and Kafiri. There are some isolated languages around the Asian continent, e.g., Korean, Japanese, Ainu and Burushaski. Some European languages – English, Russian, French, and Portuguese – are also used in the region as official languages, and from the mixture of an indigenous language and an introduced language, pidgins or creoles have emerged.

In recent centuries, under Islamic influence, the Arabic script has become the written tools for some South Asian languages (i.e., Urdu, Sindi, Kashmiri, and Divehi) (Nandasaa, Mikami, 2009). This situation resulted in part from the exploration and widening economic interest of many European – Portugues, Dutch, English, French, Spaniard, Russians, and others. As a result, the roman script introduced for Sout Asian countries such as Vietnam, Malaysia, and the Philippines, and Indonesian language scripts such as the Javanese, the Sundanese and the Balinese. Table 3 shows the selected world languages used to write their language name using language scripts belong to their language and speaking population.

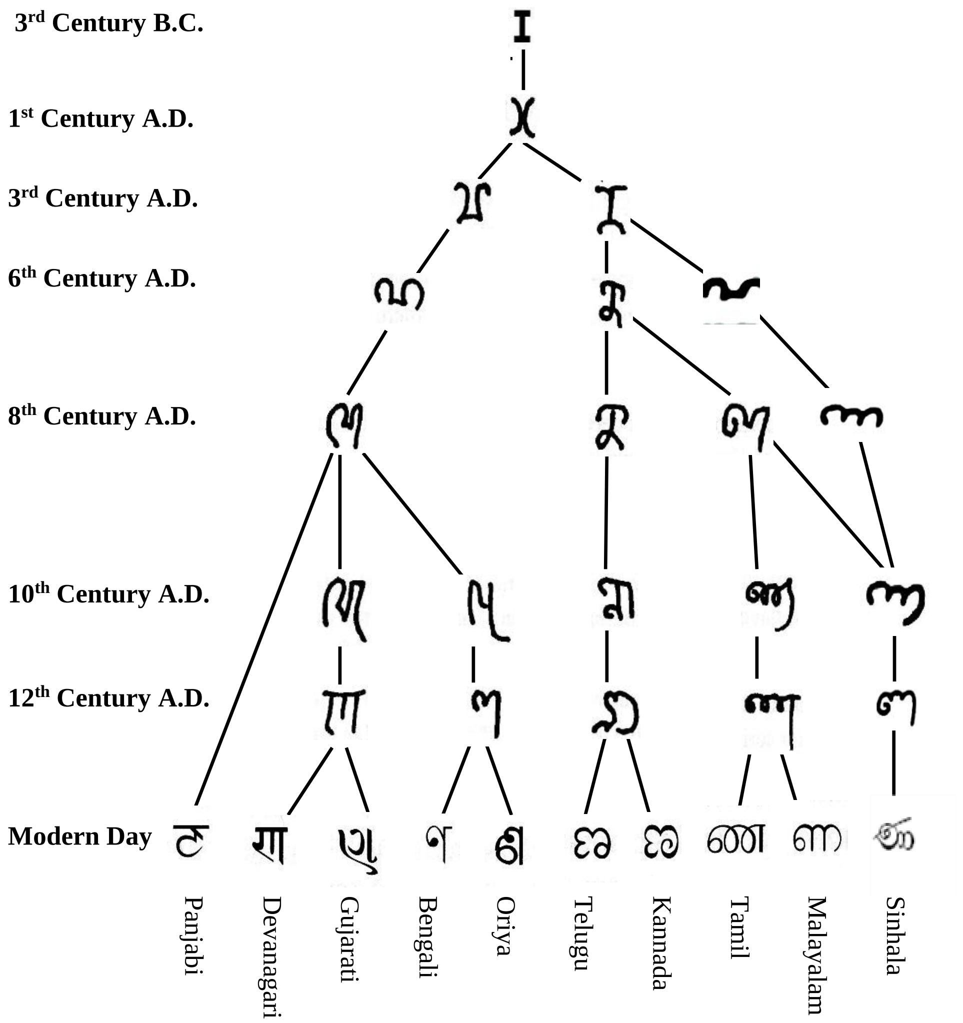


Figure 2: Evolution of Asian language scripts

Table 3: Selected world languages and language name written in their script[1]

Language Name	Speaking Population	Script	Language Name	Speaking Population	Script
Chinese	885,000,000	普通話	Nepali	16,200,000	नेपाली
English	322,000,000	English	Filipino	14,850,000	Tagalog
Arabic	280,000,000	لعربية	Assamese	14,604,000	অসমীয়া
Bengali	196,000,000	বাংলা	Azerbaijani	13,869,000	Азәрбајҹан дили
Hindi	182,000,000	हिन्दी	Sinhala	13,218,000	සිංහල
Portuguese	182,000,000	português	Zhuang	10,000,000	Saw cuengh
Indonesian	140,000,000	Indonesea	Pashto	9,585,000	پښتو
Japanese	125,000,000	日本語	Kazakh	8,000,000	Қазақ / قازاق
Hankuko	75,000,000	한국어	Uyghur	7,464,000	ئۇيغۇر
Telugu	73,000,000	తెలుగు	Khmer	7,063,200	ភាសា◌ែខ្មរ
Vietnamese	66,897,000	Tiếng Việt	Dari	7,000,000	دَرِي
Marathi	64,783,000	मराठी	Tatar	7,000,000	تاتارچا
Tamil	62,000,000	தமிழ்	Turkmen	5,397,500	түркменче
Turkish	59,000,000	Türkçe	Kashmiri	4,381,000	काऽशुर
Urdu	54,000,000	اردو	Lao	4,000,000	ພາສາລາວ
Gujarati	44,000,000	ગુજરાતી	Balinese	3,800,000	Bahasa Bali
Malayalam	34,014,000	മലയാളം	Kyrgyz	2,631,420	Кыргыз
Kannada	33,663,000	ಕನ್ನಡ	Fijian	650,000	vaka-Viti
Punjabi	25,700,000	ਪੰਜਾਬੀ	Dhivehi	280000	ދިވެހި
Thai	21,000,000	ภาษาไทย	Sanskrit	194,433	संस्कृतम्
Sindhi	19,675,000	سنڌي	Tahitian	150,000	Te Reo Tahiti
Uzbek	18,386,000	Ўзбек	Maori	70,000	Te Reo Māori
Bahasa Melayu	17,600,000	Bahasa melayu	Hawaiian	8,000	Ōlelo Hawai'i

Present records show the existence of more than 700 spoken languages in a region that now consists of India, Pakistan, Nepal, Bangladesh, Sri Lanka, and Afghanistan. According to the Summer Institute of Languages, India has 454, Nepal 122, Pakistan 77, Afghanistan 41, and Bangladesh 44 actively spoken languages (Lewis, Simons, & Fennig, 2013). The Greenberg diversity index for India is 0.916 (ranking it fourteenth in the world), while for Pakistan it is 0.775 (ranking it at number 40).

[1] Data source: Ethnologue: Languages of the World, 15th ed. (2005) .

1.5 Lack of Standard and Legacy Encoding

The most common character sets, ASCII is known as American Standard Code for Information Interchange, contains 96 character codes for the primary Latin alphanumeric characters. While excellent for English, ASCII does not support the accented letters used in Western European languages, nor does it define combined conjuncts in non-Latin scripts. These character codes conflict between 8 bit coded characters sets used in Asian languages. The same character code is being mapping to different characters, and the same character mapped to different character codes.

Until recently, many languages not based on the Latin alphabet have represented on computers with 8-bit extended ASCII encoding, including the ISO/IEC 8859 character set (SAORA's Report, 2005). These languages; many of them are in the South & South-East Asian region, and Sri Lanka used the 8-bit extended ASCII encoding method in cyberspace.

Legacy encoding is a common problem faced by any nation today. So far language recourses have been collected based on the needs on individual nations, or more relevant to Roman scripts, but in the real world, there are many cases in which it is necessary to process multiple language recourses in a single environment. For example, the cyberspace needs to use multiple language scripts in foreign language dictionaries and textbooks, and now in the 21st century in the age of the Internet, people will come into contact with foreign language data written in the character set of another language. However, lack of standards, legacy encoding schemes and other script related issues such as rendering in various languages standstill.

1.6 Interoperability Issues

1.6.1 Encoding Standards

To resolve interoperability problem that is spread all over the world languages' scripts, the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), and the Unicode Consortium made efforts to develop a single comprehensive universal character set, and they published the first version of "The Universal Multiple Octet Coded Character Set (UCS)" in 1991. Although the ISO/IEC standard (ISO/IEC 10646) and the Consortium standard (Unicode) remain separate, their character repertoires are effectively identical. However, computer systems are now able to create, store, display, process, and transmit textual data independent of hardware platforms, software applications, or languages, but remains an unresolved problem of specific rendering issues, identification of languages, phonetic ordering and sorting. However, the progress of the encoding standards are much appreciated and the current version 12.1 of the Unicode Standard, developed by ISO/IEC and the Unicode Consortium, assigns a unique identifier to each of 1,114,112 possible characters.

Unicode has defined a small number of character encoding formats to make it easier for software applications to use or store Unicode character codes. The most commonly used forms (UTF -

Unicode Transformation Format) are UTF-8 and UTF-16. UTF-8 (Rob Pike & Ken Thompson) transforms every Unicode character into a sequence of one to six 8-bit code values. ASCII characters remain encoded with single byte code values. All other characters encoded with two or more bytes. Because UTF-8 expresses code values in 1-byte, pre-Unicode software applications that expect single-byte ASCII code values will continue to work with UTF-8 data.

Moreover, if the text contains mostly ASCII characters, using UTF-8 encoding saves computer storage space, because of characters stored with 8-bit rather than 16- or 32-bit code values. UTF-16, on the other hand, uses 16-bit (or two-byte) code values for every Unicode character. As a result, all 65,536 characters from the Basic Multilingual Plane (BMP) can represent the same value found in the Unicode code charts.

1.6.2 Language Names and Codes

The number of languages in the cyberspace grows considerable manner, and essential accessibility demand ever increases for the taxonomy of human languages to adequately support language data, corpus, corpora, and databases. In order for an application to get the right content and present it in the right format, language code and naming tag should provide comprehensive coverage, and some standardized meaning. ISO standards on language names and codes (ISO 639-2, 1998) providing two or three letter codes (“Codes for the representation of names of languages" in ISO 639, ISO/DIS 639-1 and ISO 639-2)[2], seems inadequate to meet the application requirements being levied by users in a growing number of domains. Taking into account the fact that language pages are available or written in cyberspace for an increasing number of minority languages, the limited availability of codes, 139 codes of ISO 639:1988, or 405 codes of ISO 639-2:1998, or 437 in ANSI/NISO Z39.53-2001 (ANSI/NISO, 2001). Even the 457 in MARC: “The Network Development and MARC Standards Office” (US Library of Congress, 2003) are too few to support metadata description. Linguists applying language codes at a low level within natural language texts may discover that the ISO codes do not sufficiently distinguish regional, social, or dialectical variation. As many of the information providers or software developers using two-letter language codes find it somewhat obsolete and ISO is promoting three-letter language code standards. Although users in general, may find that the two-letter code identifiers are still heavy, nevertheless further classifications for 7,000+/- languages/dialects are needed for future purposes.

1.6.3 Script Names and Codes

Language identification is one of the core areas of language resource corpus building. The majority of characters encoded in the Unicode Standard (Unicode)[3] are elements of collections called scripts. The script is a collection of symbols used to represent textual information in one or more writing systems.

[2] ISO 639/Joint Advisory Committee (ISO 639/JAC), https://www.loc.gov/standards/iso639-2/iso639jac.html
[3] http://www.unicode.org/reports/tr24/

ISO has appointed the Unicode Consortium as the Registration Authority for International Standard Codes for the representation of names of scripts (ISO TC46/WG3 and M. Everson, 2003). The first version of ISO 15924 four-letter script codes and the corresponding numerical value was published on 2004-02-04 but not yet used in markup languages.

The Unicode Character Database (UCD) provides a mapping from Unicode characters to script name values (UTR #24)[4]. The mapping from character to script is useful for a variety of tasks that need to analyze a piece of text and determine what parts of it are in which script.

1.7 Digital Divide

The "digital divide" is the gap in technology usage and access. The digital divide has been investigated by scholars (Hoffman et al., 1998) and policymakers (NTIA, 1999) mainly as an economy-specific issue that permeates the population across all demographic profiles, such as income, gender, age, education, race, and region, but not specific to the languages of different communities (Nandasara and Yoshiki Mikami 2015). The lack of native language-driven Information and Communication Technology is a major conducive factor in the digital divide.

1.8 Conclusion

In the Asian region, due to its rich diversity of languages and commonality issues with special consideration to standards, interoperability and coordination issues are still addressable by any nation in this region. So far, Asian language recourses collected by only individual nations; however, they are away from the visible spectrum of the real users in the real world. There are many cases in which they need to use multiple language scripts in the 21st century. Building up of networks between other organizations exists in the region and worldwide that are involving in such activities would benefit by all parties involved in this timely development efforts.

Providing UTF-8 based encoding schemes, standard evaluation criteria and the generalized benchmark tools will build up equality among languages and their people. Broader sharing of technical achievements and know-how in this area will bring to minimize the digital language divide in the region and address the United Nations Educational, Scientific and Cultural Organization (UNESCO) recommendation concerning the “Promotion and Use of Multilingualism and Universal Access to Cyberspace” (UNESCO, 2003).

Author wises to concluding the first chapter of this thesis in general presented as an introductory background to the historical development of languages in the Asian region and further discusses the ideas relating to the rich diversity of languages and commonality issues with special consideration to the standards, interpretability and multilingual recourses creation in the Asian region. Further, a broad introduction to the origin and differences in languages and the challenges in digitizing them

[4] http://www.unicode.org/reports/tr24/tr24-5.html

presented in this chapter will examine. A comprehensive introduction to Asian languages and how they differ from the western languages are also presented, eventually dedicating a whole chapter for the Sinhala language, to compare mainly the differences and similarities of Sinhala with other languages, and to argue for the benefits of digitizing Sinhala (Nandasara et al., 2008).

Chapter two of this thesis focuses on the results of the analysis of the presence of the Asian language in web pages (Nandasara et al., 2008). The results of the literature survey review the script, with particular consideration to the encoding standards, interpretability and multilingual recourses creation in the Asian region. This chapter also contributes to the needs of Sinhala computing for education. Computer technology boots up during the late 1980s as an introduction of microcomputers to the universities and schools, and there was tremendous interest in learning computer technology through the native language (Nandasara, 2012). The chapter discusses how computer education generally developed through the country reach the goal in four decades. The case described in this chapter reveals that the digital language divide exists at a severe level in the region, in particular, small communities of non-Roman script users; such as Sinhala speaking community.

Chapter three mainly concentrates on providing an overview of the 2500 years of the historical development of modern Sinhala script, its mechanization and adaptation for different technological generations, such as writing and printing (Nandasara, Mikami, 2015). The chapter also discovers how western communities were inventing mechanisms on Sinhala type fonts for manual composing, printing and mechanical and electronic typewriting in their printing industries.

Chapter four describes the Sinhala writing systems, their structure, features, characteristics, complexities, explore the rendering issues, and consideration for the use of existing typewrite as a computer keyboard, and 8-bit standards (similar to ASCII called SLASCII) of the language scripts in respect to the early attempts of developing Sinhala interfaces for 8-bit computers (Nandasara, 2009).

The last chapter focused on the eight-year-long exercise of implementing the SLS 1134:1996 initiative and achieving the UCS/Unicode implementation for the Sinhala scripts. The latter part of the chapter comprises the cultural, sociological and technological background of the entire process related to USC/Unicode implementation. Rendering issues related to the modern writing style of Sinhala together with problem-solving methodologies relating to Sanskrit and Pali writing styles and how UCS/Unicode help to solve the problem of the Digital divide in the society (Nandasara, 2009; Nandasra, Mikami, 2015).

This page intentionally left blank

Chapter 2: Script and Encoding Issues of Asian Language Text on the Web and Needs of Sinhala Computing in the Country

Information and communication have now become such an integral part of societies that its access is considered a fundamental human right. Although the availability of computers is improving in Asia, very few people are Internet users. The major limiting factor is the ability of most people to speak and write in English or in any other major languages that dominate the Internet. On the Internet, there is another divide that separates the world's population into almost equal halves, those who understand English and those who do not. Therefore, any serious discussion of universal access to the Internet must confront the language divide in cyberspace, to ascertain how it adversely affects a large segment of the population who do not use English. Only one in ten people in the world can read English. Most of the websites in the cyberspace are written in English. Another problem is illiteracy; most of the people in their country cannot read so no matter what language the website is using.

2.1 Introduction

According to the survey data published in 2008 (Nandasara et al., 2008), based on the number of web pages per 1000 population, as this is the reflection of the degree of presence of a country on the web, shows that Israel is the highest (3757 pages per 1000 population) in the rank, and Singapore (1040) and Cyprus (693) follow, respectively. The survey also shows that Kazakhstan and Azerbaijan respectively have the highest web page size per 1000 population among Central Asian countries, whereas, Cambodia, Afghanistan, Pakistan, India, Syria, Yemen, Bangladesh, and the last, Myanmar, have the least number of pages presence on the web (between 5 (4.54%) to 0 (0.35%) pages per 1000 population). It is worth noting that Myanmar, the neighboring country to Thailand, has the least (0.35%) among all the Asian countries.

The Online Computer Library Center's (OCLC) Web Characterisations Project (O'Neill, Lavoie and Bennett, 2003) covers a large number of European languages. Most of these surveys have evolved along with multilingual search engines like Yahoo, Google and Alltheweb. The language-specific search capability of the search engines has provided a means of surveying for researchers. Although these surveys have given us reasonably good pictures about European language presence on the Web, far less attention has been paid to Asian languages, among them "less computerized languages" in particular. With the exceptions of Chinese, Japanese, Korean, Thai, Malay, Turkish, Arabic, and Hebrew, nothing known about the extent of the presence of Asian languages on the Web. The UNESCO report, presented to the Tunis phase of the World Summit on the Information

Society, “Measuring Language Diversity on the Internet” (Paolillo, Pimienta, Prado, 2005; Annual Project Report-2005), shares precisely the same concerns.

As far as Asia is concerned, most of the countries are far behind in developing their language written websites on the Internet. Most of the websites geared toward English. There are a few million people out there who have ideas and cannot express them on through the Internet because any other cultures will not read it. Sri Lanka, though started developing multilingual websites (Nandasara, Samaranayake, 1997; Nandasara et al., 1997; Nandasara, 1997) facilitating trilingual capabilities (say Sinhala, Tamil & English) within the National website (www.lk), inaugurated on 08th July, 1996 was fully supported by the government and thereafter taken many attempt to overcome the problems such as common understanding of language character set, their alphabetical order, and standards for the scripts. Sinhala scripts represent vowel sounds by adding vowel-signs before, above, below or after the consonants. They derive from the Brāhmī script used in the Indian sub-continent many centuries ago and written from left-to-right. These entire alphabets follow a very logical phonetic order. Thus, all computer equipment for the Sinhala language needs to provide for this degree of complexity in both display and printing but without adding any extra complexity to the keyboard or other input systems (Nandasara S. T. and Yoshiki Mikami, 2008).

The Sinhala writing system used in Sri Lanka is a syllabic writing system deriving from Brāhmī and consists of vowels, consonants and semi-consonants. Several of these are combined to form complex ligatures. The total number of different glyphs is almost close to 2300. Someone can argue that it every possible letters, composites, conjuncts and ligature are to be represented will exceed 4000. In recent years, some newspaper and printing organizations and individuals have developed such fonts, as it allowed them to have full control over the appearance of each letter. Although these fonts are large, consequently, all computer equipment for the Sinhala language needs to provide for this degree of complexity in both display and printing but without adding any extra complexity to the keyboard or the input systems.

Asia is concerned, once again, most of the countries are far behind in developing content in their native languages. Millions of people out there who have ideas are not able to express them digitally “Digital-Contents” due to the lack of native language support while many are not able to read digital content due to the same reason. In Sri Lanka; though the development of multilingual computing facilities (for example Sinhala & Tamil) traced back to as early as the mid-80s, a survey shows (Mikami et al., 2005) that the web has only 0.02% of the Sinhala content pages for 1000 Sinhala speaking inhabitants.

2.2 Sinhala Language as the Official Language and Free Education

The Sinhala language or, (to use the native term now frequently used among linguists and grammarians), *Hela* (Fernando, 1949) a unique script which is used only in Sri Lanka. This language entered a new phase in its history in 1956 when Sinhala recognized as the official national language and medium of education of the Island spoken by about sixteen million people. The new status accorded to Sinhala brought about changes not only its communication and sociological role but also its linguistic structure.

In 1805, after the British invasion, the education system in Sri Lanka was developed based on the British System. Sri Lanka received independence from the British on the 4th of February 1948. Hon. C. W. W. Kannangara, the then Minister of Education in the State Council (1931-1947) introduced several far-reaching policies in education, a few years before receiving independence from the British. They included primary education in the mother tongue, free education for primary, secondary and higher education, the establishment of many quality schools in all regions of the country and the provision of a free mid-day meal in schools. The striding measures earned him the name “The Father of Free Education” and have contributed immensely towards opening up primary, secondary and higher education to the masses as opposed to the elite that benefited until then.

In Sri Lanka, though the government plays a primary role in education, it does not claim a monopoly over it. There are many Buddhist *pansala* and *pirivena*, Muslim schools, and Christian schools in the country. Roman Catholic Churches alone operates several hundred schools enrolling over 80,000 children. The language of teaching in Buddist *pansala* and *pirivena* is mainly Sinhala and for writing Sanskrit and Pali also used Sinhala script.

The Government Technical School (GTS), also established in 1893 in Colombo and provided technical education in Civil, Electrical and Mechanical Engineering fields.

The university system in Sri Lanka operates within the framework laid down in the Universities Act 16 of 1978 and reintroduced the system of separate universities functioning under the direction of the University Grants Commission (UGC). UGC has governed 15 public universities, seven postgraduate institutes, and nine other higher education institutions by 2010. Besides, there are two universities for Buddhist studies established under the Ministry of Higher Education. Most of the Universities in Sri Lanka teach non-technical courses in Sinhal and Tamil Medium.

2.3 “Computer Revolution” in Sri Lanka

The teaching of computer programming and the use of computer applications for research and teaching at the University of Colombo commenced in 1967 the first time in Sri Lanka. Computing facilities provided by the State Engineering Corporation in the ICL 1901 mini computer and it was given free of charge. A few years later in 1971, the Department of Census and Statistics allowed the University of Colombo free computer time on their IBM 360/25. The fact that these installations were close to the University of Colombo and the interest of both organizations in statistical and scientific applications helped the university researchers to make excellent use of this invaluable gesture.

2.4 Computer Programming Course Units for Non-Science Students

Under the Higher Education Reforms that took place in 1972 (Jayarthne Report), the Department of Mathematics and the Statistical Unit of the University of Colombo made a remarkable attempt to initiate new course units in Mathematics, Statistics and FORTRAN Programming to the Statistical Services job stream for Faculty of Art degree students following the newly introduced Special Degree in Development Studies. Thirty students were selected from 210 reading for the above degree and allowed to follow a special degree stream, which was specially designed and managed by the Department of Mathematics of the University of Colombo. The first time of the history of university education, all and above, the computer programming in FORTRAN course conducted in Sinhala medium.

2.5 Computers go Public

One of the main demarcation points of the history of Sinhala computing in Sri Lanka was the introduction of computers for assisting the Commissioner of Elections to process the results of the National Presidential Election in November 1982. In late 1981, purchase a few BBC microcomputers, due to be released to the market in 1982. When they did arrive, in October 1982, this remarkable microcomputer was an immediate success, and the Computer Centre received much publicity among the public by their computer display which enabled the telecast of the Presidential Elections of 1982 over National Television (*Rupavahini*). The process of release of results of every national election has continued without a single break after that, with technological improvement and use of local languages, Sinhala and Tamil, at every stage.

With high publicity received by the University of Colombo due to the release of the Computer-Assisted Presidential Election results, the Computer Centre was able to inaugurate a Certificate Course in Computer Applications in Sinhala medium for the public, to be held during weekends, not interfering with the undergraduate courses. These courses primarily meant for the employed to gain knowledge of computer applications rather than to those wanting to learn computer programming for employment.

Much software developed in terms of national interest includes the graphical display for the media such as the TV Programme Parade, the Cricket Scoreboard, South Asian Federation (SAF) Games Display and National Quiz programs organized, managed and conducted by the University of Colombo in early 1980s. All these programs use Sinhala and Tamil as a language for display.

The computing staff, mainly the author, at Colombo University, has been engaged in the dissemination of knowledge to the nation’s citizens through public and private media channels for more than three decades in the Sinhala medium. These television and radio programs received the highest ratings and were the most popular programs all the time. These programs, namely, the Computing for Schools (5 episodes, 1995), Day-to-day use of computers, the program directed by Daya Liyanage from MTV TV channel (5 episodes, 1996), ‘අන්තර්ජාලය ඔබේ නිවසට’ (Internet to your Home) – weekly one hour LIVE program for IT related technical discussion on Sri Lanka Rupavahini Corporation (SLRC) (National Television), Sundays 7.00pm – 8.00pm (- more than 150 episodes, 1997-2001), ‘Internet and You’ a weekly one hour LIVE program for IT and internet related technical discussion on Sri Lanka Broadcasting Corporation (SLBC) Wednesdays 9.00am to 10.00am (- more than 120 Programs, 1997 – 2001), ‘e@ශනිදා’ (e-Sunday), weekly programme on wendsday evening for children, nation-wide IT quiz competition (IT-Quiz 2000) program telecasted on National Television, weekly one hour LIVE program for IT related technical discussion on SLRC Sundays 7.00am - 8.00am, (5 Programs, 2003), ‘FORUM for BIT’ – a weekly 30 minutes educational TV program for Bachelor Degree in Information Technology (BIT) telecasted on SLRC and TVLanka, (132 Programs from Oct. 2003 to Sep. 2005), *Siyabasa Alankaraya* (සියබස් අලංකාරය; Beauty of the Mother Tongue) weekly television program on Independent Television Network (ITN) (22 Episodes, 2005), ‘පරිගණක පරාදීසය’ (Computer Paradise), weekly 30 minutes children program on SLRC (Thursday 5.00pm - 5.30pm, 2008), ‘e-මිතුරෝ’; e-friends, weekly 30 minutes program on SLRC (Thursday from 5.00pm to 5.30pm, 2009), ‘ආසයි ලේසියි IT’, weekly 30 minutes program in Information and Communication Technology for students in schools and universities, and for the general public on SLRC, Saturdays from 6.00pm to 6.30pm, (102 episodes, from May 8th, 2010 and Aplral 2012), ‘ජාතික පාසල’ (National School), 24 programs for ‘Information and Communication Technology’ for A/L syllabus were recorded and telecased from 1st October 2011 to 31st of July, 2012 on Rupavahihi and many other programs on National and other Television channels in IT related discussions.

At present, in Sri Lanka, which has a population of 22.5 million people, the majority are Sinhalese (74.9%). Other ethnic groups include Sri Lankan Tamil and Indian Tamil (15.4%), Sri Lankan Moors (9.2%), Malays, and Burghers. (World Factbook, 2018)

There are also four classical languages used mainly for religious purposes. Sri Lanka has all the major religions of the world: Buddhism, Hinduism, Islam and Christianity. Each of these religions has a classical language used in the conduct of religious transactions, including traditional rites and rituals.

1. *Pali*[5] (Florian, 1989) is also a member of the Indo-Aryan family closely related to Sanskrit, and its origins go back to the ancient Indian language Magadhi used in the state of Magadha where the Gautama Buddha spent a more significant part of his life. Most of the

[5] Pali, the language of the Buddhist canonical writings, is the oldest literary Prakrit. It remains in liturgical use in Sri Lanka, Thailand, Myanmar, Lao and Cambodia (*Florian Coulmas, 1989: p190*).

Buddhist scriptures such as the Tripitaka (The Three Baskets) written in Pali. Pali has no script of its own: in each of the countries where it used as a religious language, written in the local scripts. These were committed to writing for the first time in Sri Lanka in the 1st century A.D.

2. *Sanskrit*[6] (Florian, 1989) is the classical language of Hinduism. Some of the ancient Buddhist Sanskrit texts used to write in Sinhala.
3. Classical *Arabic* used to recite the Holy Quran by Sri Lankan Muslim communities in the mosque, although their daily communication both at home and mosques are in Tamil and Sinhala languages.
4. *Latin* is the classical language of the Catholic Church. It is still used in the chanting of certain hymns although there is a trend to replace it with the national languages.

There are, also, a few languages and creoles spoken by small, almost microscopic, communities. These include the following.

5. *Malay* used by a small community of people of Malay origin. They migrated to the Island during the eighteenth and nineteenth centuries from Java when the western coastal areas of Sri Lanka. It may be considered a dialect of Malaysian.
6. The creole of the *Veddas*, known as the aboriginals of the Island, is used by them in addition to Sinhala. The creole of the *Veddas* based on the grammar of Sinhala and a vocabulary derived from the original language of the *Veddas*.
7. The creole of the '*Gadi'* community of the island was also used in addition to Sinhala. This community now assimilated into the Sinhala ethnic group. Their creole also has a Sinhala grammatical basis.
8. The speech of the '*Ahikuntika'*, the so-called gypsies of the Island, is a dialect that has links with India, particularly the language of the gypsies who live in Andhra Pradesh.
9. Sri Lanka was also the home of a *Portuguese* creole that was in use in the Eastern province by a small community of descendants of Portuguese origin. The Island was under the Portuguese rule for a century and a half beginning from the dawn of the sixteenth century.

Figure 3 shows the areas of language speakers spread around the island, and obviously, the majority of Sinhalese are living in more than the southern part of the island and Tamils are living in north and central part of the island. Other minority groups Vaddas, Portuguese, Maly, Sinhala Tamil, are scattered around the island.

[6] Sanskrit (from *sąm-skŗta* 'elaborated') is that phase of the literary language of ancient India witch is described in the grammar of Panini (*Florian Coulmas, 1989: p186*).

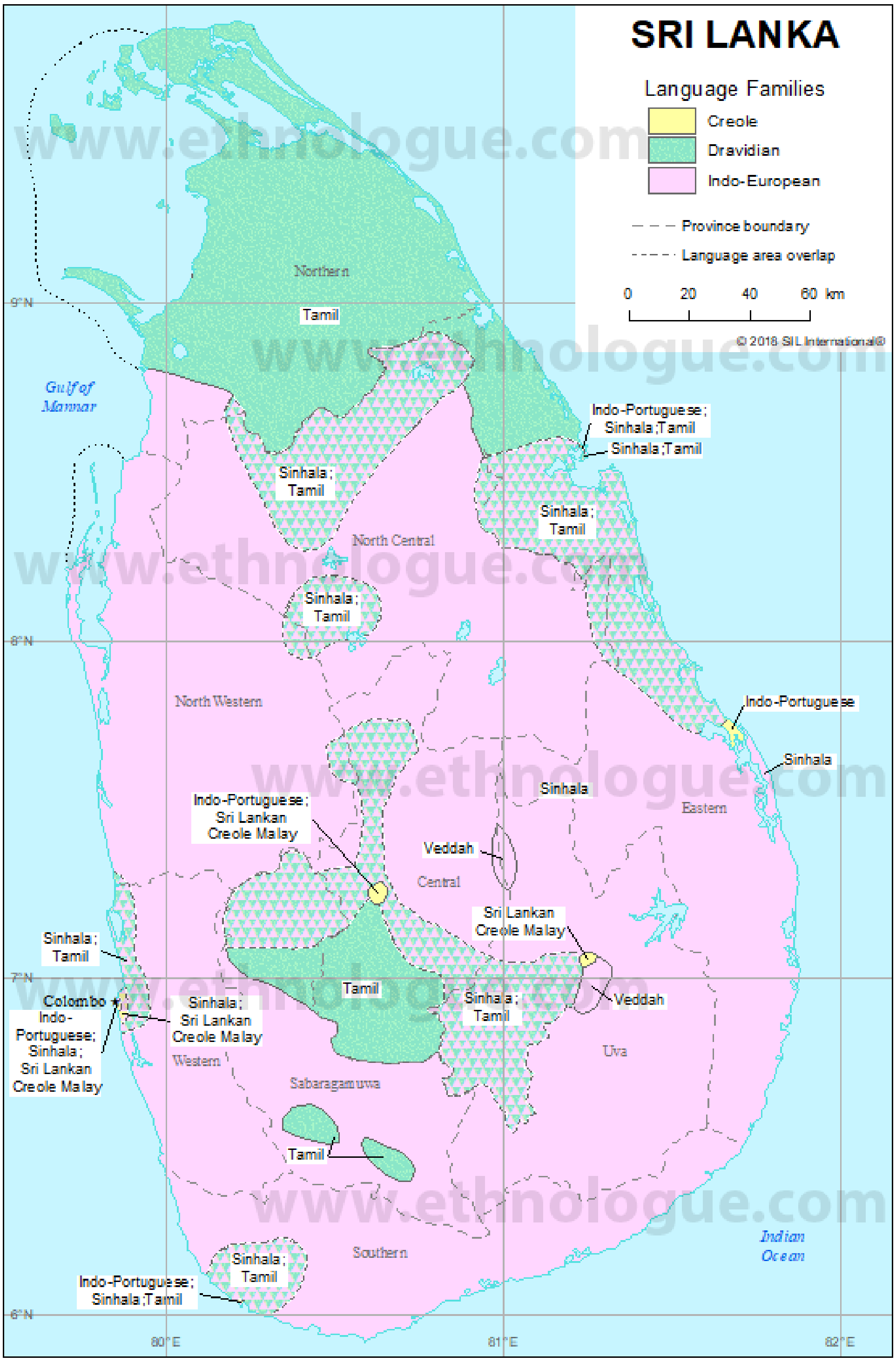


Figure 3: A Map of Sri Lanka showing areas by use of languages.

2.6 Conclusion

The next chapter, the chapter three discusses (1) historical background of the Sinhala writing system; (2) essential features of Sinhala language and chapter four illustrates; how Sinhala text processing technology has evolved over the last quarter-century. The design of character code standards marks significant steps as the cornerstone of whole architecture for the Sinhala text processing. A case described in the next chapter three and four discusses how isolated island communities of non-Roman script users could connect itself to the Romanised system dominated cyberspace.

Chapter 3: Sinhala Language and Challenges - Evolutions of Sinhala Script and Related Problems

This chapter focuses on providing an overview of the historical development of modern Sinhala scripts, their mechanization and adaptation for different technological generations (such as rock inscriptions, an era of "ola" manuscript writing and printing). Preceded by a brief historical background on Sinhala script, this chapter further elaborates on the structure, features, characteristics and complexities of the Sinhala scripts. An overview of early attempts of mechanization of Sinhala scripts—covering aspects of manual composing, printing, and mechanical and electronic typewriting—is also presented. Followed by language features, syllable structure, this chapter also explores the significant issues of Sinhala writing system. Some of the complex ligature and their shapes examined since the aforesaid is a significant challenge faced by the Sinhala language interface developers.

3.1 Historical Background

The recorded history of Sri Lanka traced back to the pre-Christian era. The early history of Sri Lanka is contemporaneous with that of the early North Indian Buddhist civilizations. In the 3rd century B.C., during the reign of Emperor Asoka in India, Buddhism was formally introduced to Sri Lanka by venerable *Arahat Mahinda*, on the conclusion of the Third Council held at *Pataliputra*, under the Chairmanship of *Moggaliputta Tissa Thera*. It is the accepted view that the Venerable *Arahat Mahinda* preached the *Dhamma* to the people of Sri Lanka in their tongue. Supposing however that the noble *Arahat Mahinda Thera* was not already conversant with the *Hela*[7] language (*dipa-bhasa*) he would have acquired a sound knowledge of the language to suit the demands of his mission.

The early stages of linguistic history mainly marked by either cave or rock inscriptions[8] found in almost all parts of the Island. Usually, these cave inscriptions found below the drip-ledge where the script protected from water. In some cases, the writing continues as one line for about forty to fifty feet and written from right to left. Besides the available inscriptions, literary works begin to appear during the medieval period. There were two distinct genera of literature produced at the time. The first was the exegetical and philosophical treatises, and then the other was the creative Sinhala literary works. The efforts of some of these highly talented scholars have referred to both in the historical and epigraphical records of the Island.

[7] Hela Language, Elu, Hela, Sihala and Sinhala mean the same language that has gone through various stages of evolution as shown in numerous evidences from inscriptions and early Sinhala works.

[8] The University of Cambridge, England has 274 volumes of 'Epigraphica Zeylanica' with over 3000 inscriptions from Sri Lanka (that is more inscriptions than the whole of mainland China has), including one dating back to 6th century B.C. Over 2000 of these have been deciphered, indicating the consistent development of the Sinhalese language.

The period commencing from the 8th century A.D. and ending in the middle of the 13th century A.D. is considered as the old Sinhalese age. During this age, many of the features of the modern Sinhala language sprang up and evolved satisfactorily. However, during this period, the only existing works are *Dhampiya Atuva getapadaya* (10th century A.D.), *Amavatura* (12th century A.D.), *Sadharmaratnavaliya* and *Pujavaliya* (13th century A.D.). In addition to these, there are in the present works numerous quotations supposed to be from lost books that furnish exciting insights into the literary activity of the ancient period. From the 13th century A.D. onwards, the production of literature became more prolific. Besides, there are the *Sigiri* Graffiti scribed on the gallery wall of the famous rock by that name. These depict an exceptional picture of not only the literary sensitivity and the aesthetic value of those ancient people, but also their educational and social background, particularly of those people who visited *Sigiriya*, during 8th to 10th centuries A.D.

During this period, the rock inscriptions, which had previously been brief, began to grow in size and number. The language specialists believe that there were significant changes in the spoken language during the medieval period more rapidly than written language. (Jayathilaka, 1937).

The birth of the style of written language, known as mixed Sinhalese language today, has its origin in the 12th century. The predominance of Sanskrit found in the style of the Sinhala language used in both inscriptions and literature belonging to this period (Kulasuriya, 1962).

The geographical situation of Sri Lanka has made it a haven for various sea-farming nations. This association was also responsible for the Sinhala language acquiring the features of a mixed dialect. The rich Sri Lankan literature both in Sinhala and Pali attracted many scholars to the Island from time immemorial. Among these were a large number of South Indian Dravidian scholars of no mean repute. Their contribution, particularly to Sri Lanka Pali literature, is significant. Sri Lanka, some experts claim as having one of the world's oldest continuous written records of history called *Mahavamsa* (more than 2500 years) and numerous such references to the literary activities of ancient Sri Lanka in the historical and literary record left behind.

Sinhala is a uniquely spoken, and written language in Sri Lanka and a unique script used to write the Sinhala language. Sinhala is said to have derivatives from the ancient scripts of Brāhmī, known to have existed since the third to the second century B.C. Subsequently, alphabet and writing systems have changed considerably with notable influence by the *Kadamba* and *Pallava Grantha* script of south India (Florian 1989; Florian 1996; Fernando 1949). The full Sinhala script includes the symbols necessary for writing *loaned* words-that is, words that originated from the Sanskrit and the Pali, notably the aspirated consonants.

According to Jayasiri Lankage, Sinhala scripts used in ancient inscriptions and literature have categorized into five periods between 3rd to 10th centuries A.D. (Jayasiri Lankage 1996). This categorization is very much aligned with our study too. The evolution of the Sinhala script for the standardization study was based on mainly the usage of the scripts, and shapes and patterns of scripts in various historical inscriptions and literature. According to the evolution of script by their shapes and the usage, the periods can categorize into seven as follows: (*see* Figure 10 for details)

1. Ancient Brahmi Scripts from the 3rd to 1st century B.C.
2. The first period of simplified Brahmi Scripts from 1st to 3rd century A.D.
3. The second period of simplified Brahmi Scripts from 4th to 5th century A.D.
4. Pallava period from 6th to 7th century A.D.
5. Mediaeval Sinhala Scripts from the 8th A.D. to 10th A.D.
6. Ola Leaf and other literature found in 11th to 17th century A.D.
7. Printed Sinhala Script from 18th century A.D. to Today

3.2 Evolution of Sinhala Scripts

The oldest writing of Sinhala can trace back to about 3rd century B.C. These are inscriptions[9] mainly marked in either cave or rocks are found in almost all parts of the Island. Usually, these cave inscriptions found below the drip-ledge where the script has protected from water. In some cases, the writing continues as one line for about forty to fifty feet from left to right, and in some cases, inscriptions found written from right to left.

3.2.1 Golden Era of Cave and Rock Inscriptions

The Sinhala writing system derived from the ancient North Indian scripts, Brāhmī, where forty symbols occurred in the edicts of *Asoka*. The study shows (see Figure 10) that thirty-eight appeared in Sri Lankan inscriptions from the 3rd century B.C. to 3rd century A.D. Earliest inscriptions were geometrical in shape (Figure 4 and Figure 5) and resembled the Roman and Greek scripts. However, as the time passed the geometric straight-line scripts gradually became rounded at the edges by 1st century A.D.; as Figure 6 shows and subsequently alphabet and writing patterns have changed almost circular form by 15th century A.D. with notable influenced by the *Kadamba* and *Pallava Grantha* writing system of South India (Florian, 1996).

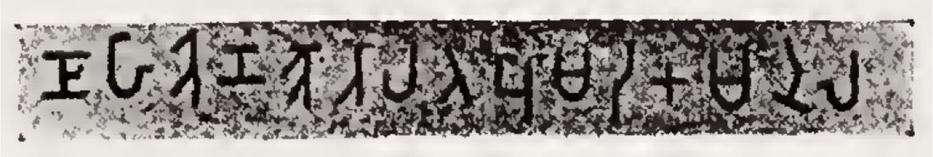


Figure 4: An Ancient Cave Inscription at Ambalakanda. (One of the earliest to be found in the Island, Date before fifth century A.D.)

[9] The University of Cambridge, England has 274 volumes of 'Epigraphica Zeylanica' with over 3000 inscriptions from Sri Lanka (that is more inscriptions than the whole of mainland China has), including one dating back to 6th century B.C. Over 2000 of these have been deciphered, indicating the consistent development of the Sinhalese language.

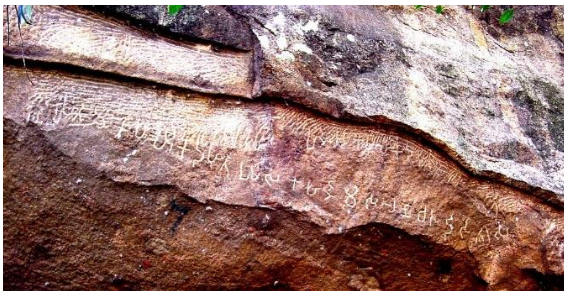


Figure 5: Vessagiri Cave Inscription in 2nd century B.C.

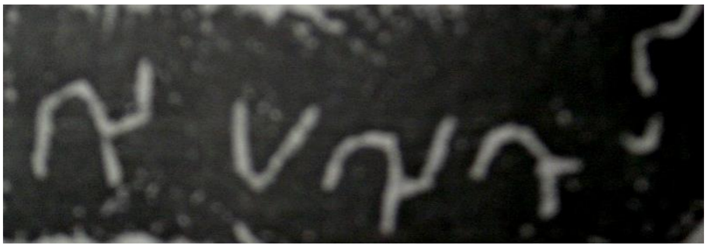


Figure 6: Cave Inscription 1st century A.D.

3.2.2 Golden Era of Ola Manuscripts

National Museum in Sri Lanka has a collection of about 3600 ola leaf manuscripts. Among them, about half of these are on Buddhism. The oldest palm leaf manuscripts in existence are *Dhampiyā Aṭuvā Geṭapadaya* (10th century A.D.), *Amāvatura* (12th century A.D.), *Chūla Vagga* (12th century A.D.), *Saddharmaratnāvaliya* and *Pujāvaliya* (13th century A.D.). From the 13th century A.D. onwards, the production of literature becomes more prolific.

Furthermore, during the 8th to10th centuries, there are *Sigiri* Graffiti scribed on the gallery wall of the famous rock palace and kingdom by that name. These show an excellent picture of not only the literary sensitivity and the aesthetic values of those people, but also their educational and social background particularly people who visited many parts of the Island.

By about the 17th century A.D. a vibrant book industry was in operation. Books were written on varied subjects such as History (*Mahāvamsa*), Buddhism (*Chūla Vagga*), Grammar (*Sidhaṭ Sangarā*), Poetry (*Jātaka Poṭa*), Art, Medicine, Astrology and Rituals. The efforts of some of these highly talented scholars have referred to both in the historical and epigraphical records of the Island.

Figure 7 & 8 show some manuscript pages from the early 18th century to the late 19th century. The oldest of this (Figure 7) has more rectangular letters than Figure 8, and each letter is separated. Figures 8 and 9 are from the 19th century and observed that these characters are written with more confidence and flourishers.

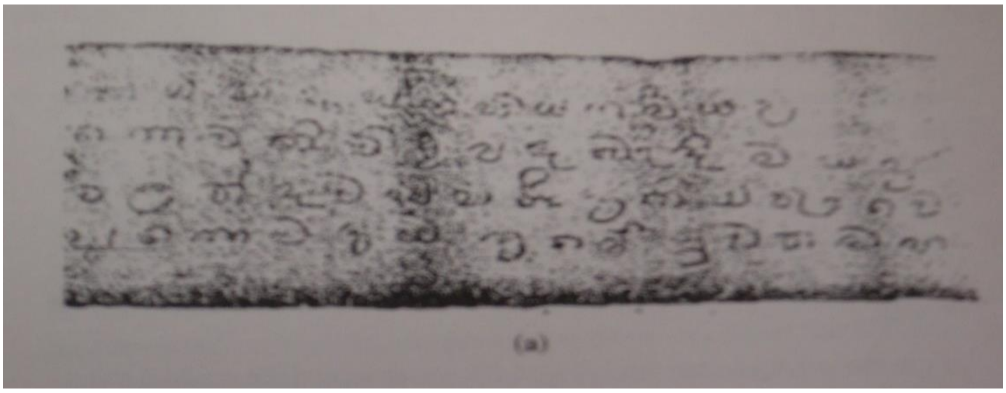


Figure 7: Early 18th Century Ola Manuscript

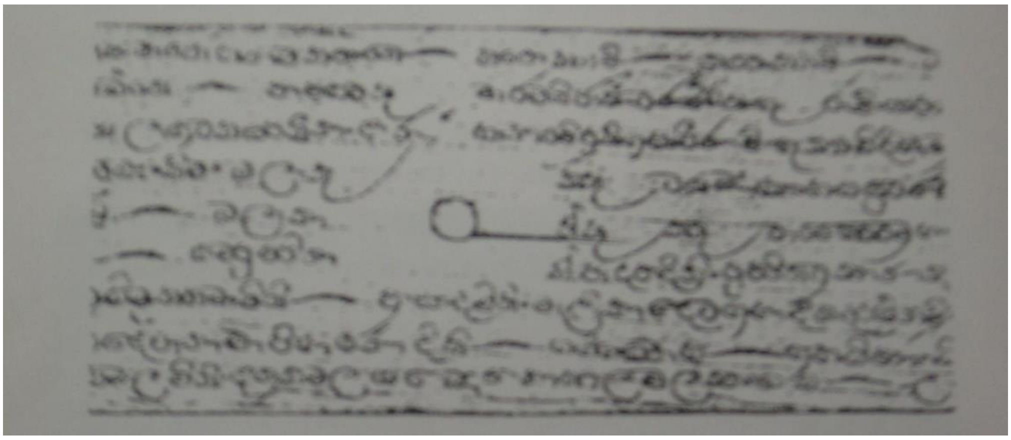


Figure 8: 19th Century Ola Manuscript

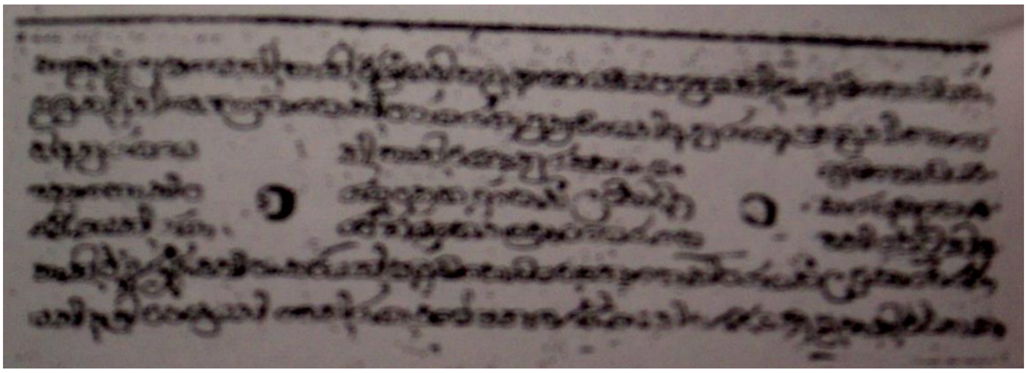


Figure 9: Late 19th Century highly literary Jātaka Pota of 550 stories of Buddha's Prior Incarnations written in Sinhala on Ola Manuscript

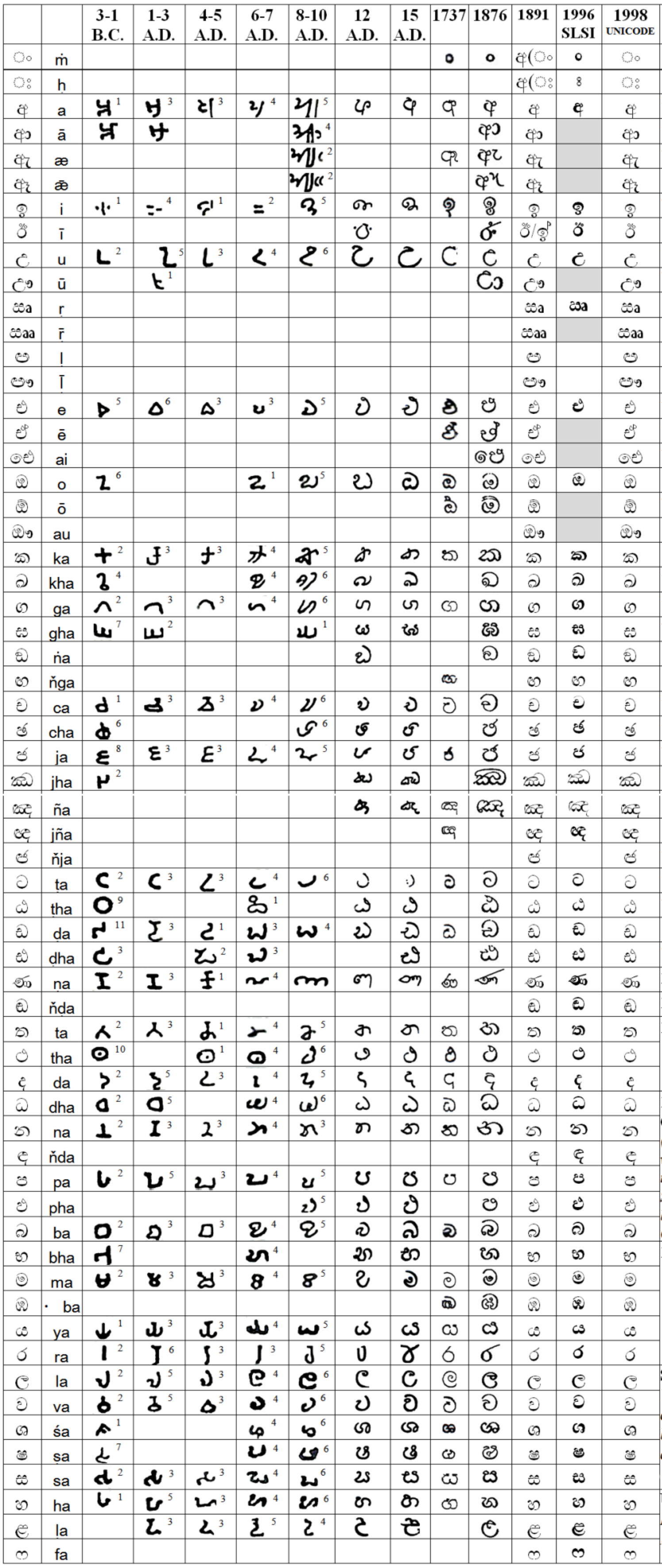
Rounded shapes of the Sinhala characters evolved mainly due to the use of *ola leaf* (from *Palmyra* tree) from the very early time. According to the *Alutveva* Pillar Inscription (*900 A.D.*), *Palmyra* was extensively cultivated during the early days and referred to four kinds of land in the country, and one of the lands reserve for cultivation with *Palmyra*. Due to the use of ola leaf with sharper steel stylus point, the original *Brahmi* script of the horizontal and vertical straight line was not suitable for scribing on palm leaves as the leaf gets torn (Jayasili Lankage, 1996). As a result of the *Brahmi* script gradually took the present rounded shape to form the modern Sinhala alphabet. John Davy (John Devy 1990) says "*the Sinhalese write very neatly and expeditiously, with the sharp-pointed style*". Davy, further says, "*Their (Sinhala) books are pretty numerous, and though* *much more expensive than printed works are very much cheaper than manuscripts were in Europe before invented in printing. The insect of ink curves into a shape that is almost sickle, spoon, eyelid....*". Michael Ondaatje[10], in his book, "*Running in the Family"* (Michael 1984), he said, "*I still believe the most beautiful alphabet was created by the Sinhalese. The insect of ink curves into a shape that is almost sickle, spoon, eyelid....*". The length of this stylus varies from ten inches to twenty inches. A stylus has six distinct parts, and the names suggest their general shape. Different kinds of metal used for making styluses, for example, gold, silver, copper and bronze, but the sharper writing point always made of steel. Figure 10 shows the evolution of Sinhala script from 3rd century B.C. to date; more importantly, it examines the evolution that happened until Germany introduced the Sinhala printing in 1876.

In Sri Lanka, it is a traditional practice to enshrine books and other precious material in *Dagobas*. In *Haguranketa Vihara,* there have listed some books awaiting enshrinement; says D.M. de Z. Wickramasinghe, Assistant Librarian, Colombo Museum Library, in the letter to the Librarian, F.H.M. Corber, reported in 1889 (Piyadasa 1985). They are:

1. Five Prakarana books of the *Vinaya Pitaka* written on silver plates
2. Seven *Prakaranas of Abhidhamma* on silver plates
3. *Digha Nikaya* of *Sutra Pitaka* on silver plates
4. *Majjhima*, *Samyukta and Anguttara Nikayas* on Ola
5. *Satipatthana Pratimoksha* and other *books* on 37 plates of gold
6. *Jataka Atuwa* on 900 copper plates.

This act compares favorably with the ages-old tradition that the books written at *Aluvihara* in Matale were copied on plates of gold and enshrined under a rock at *Alu Lena*. *Mahavamsa* is the only text written in Pali language by *Buddhagosha,* monks after the 5th century A.D, which carried in written form as excerpts of some of the early Sinhala text. However, on the other hand, inscriptions of ancient Sri Lanka dating from the 3rd century B.C. and right down to the 12th century B.C. series were found on rocks, pillars and sometimes on slabs. These inscriptions demonstrated that the use of the language dealt with the people, Kings' orders and their determination of the taxes (*Bojakapati*) and donations made to the Buddhist *Viharas* (temples) giving a glimpse to the day to day activities of the people during this era. The University of Cambridge, England has 274 volumes of 'Epigraphica Zeylanica' with over 3000 inscriptions from Sri Lanka (that is more inscriptions than the whole of mainland China has), including one dating back to 6th century B.C., indicating the consistent development of the Sinhalese language. However, over 2000 of these have disappeared.

[10] Michael Ondaatje is a novelist and poet who was born in Sri Lanka and now lives in Toronto, Canada. He is the author of The English Patient (for which he was awarded the Booker Prize)..


3-1 B.C.
1-3 A.D.
4-5 A.D.
6-7 A.D.
8-10 A.D.
12 A.D.
15 A.D.
1737
1876
1891
1996 SLSI
1998 UNICODE
ṁ
h
a
ā
æ
ǣ
i
ī
u
ū
r
ṝ
l
ḹ
e
ē
ai
o
ō
au
ka
kha
ga
gha
ṅa
ňga
ca
cha
ja
jha
ña
jña
ňja
ta
tha
da
dha
na
ňda
ta
tha
da
dha
na
ňda
pa
pha
ba
bha
ma
ba
ya
ra
la
va
śa
sa
sa
ha
la
fa

Notes

3-1 B.C. (Inscriptions)

[1] Periyankulama (207-197 B.C.)
[2] Mihintale (207-197 B.C.)
[3] Situlpawwa (161-137 B.C.)
[4] Korawakgaka (77-63 B.C.)
[5] Ritigala Weweltanne (22-7 B.C.)
[6] Yatahalena Vihara (22-7 B.C.)
[7] Gallena Vihara (22-7 B.C.)
[8] Nuwaragala (22-7 B.C.)
[9] Ritigala Andiayakanna (22-7 B.C.)
[10] Boowattegala (22-7 B.C.)
[11] Rajagala (44-22 B.C)

1-3 A.D. (Inscriptions)

[1] Anuradhapura (1-7 A.D)
[2] Situlpawwa (1-7 A.D.)
[3] Maharatmale (7-18 A.D.)
[4] Wallipuram (67-111 A.D.)
[5] Viharagala (60-67 A.D.)
[6] Pahala Kainattama (60-67 A.D.)

4-5 A.D. (Inscriptions)

[1] Tonilaga (301-328 A.D.)
[2] Ruwanweliseya (337-365 A.D.)
[3] Tissamaharama (406-428 A.D.)
[4] Anuradhapura (437-452 A.D.)

6-7 A.D. (Inscriptions & Pillars)

[1] Kandanadu (517-518 A.D.)
[2] Dhakshinathupa (639-650 A.D.)
[3] Baron Paviliyan (639-650 A.D.)
[4] Kuchchaweli (639-650 A.D.)
[5] Murutawa (639-650 A.D.)

8-10 A.D. (Inscriptions & Pillars)

[1] Thiriyaya (733-771 A.D.)
[2] Viyaulpotha (853-887 A.D.)
[3] Dorabewila (915-923 A.D.)
[4] Baddulla (946-954 A.D.)
[5] Polonnaruwa (982-1029 A.D.)
[6] Indikatuseya (982-1029 A.D.)

12 A.D. (Inscriptions & Pillars)

15 A.D. (Inscriptions & Pillars)

1737 (Printed Characters)

Sinhala in First Printed Book (1737) (see Figures 5 for sample page)

1876 *"ALFABETE DES GESAMMTEN ERDKREISES" ("Alphabet of all race of the world") which publishes the letter printing type of the world (K.K. HOF- UND STAATSDRUCKEREI IN WIEN, 1876) by The Royal Print Shop in 1876, Vienna, Germany (29 sheets).*

1891

Alphabet listed in "A Comprehensive Grammar of the Sinhala Language" by A. M. Gunasekara. (1891)

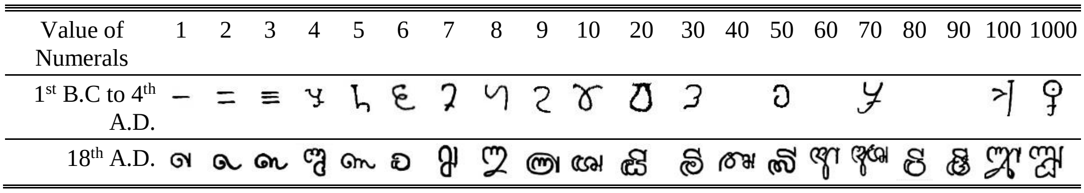
SLSI Character Set *'Sarasavi'* Font – 10 points (1996), *Characters in Gray cells were not included due to they can be produced by combining with consonant modifiers*

UNICODE Character Set *'Iskoola Pota'* Font – 10 points (1998)

***Figure 10**: Illustration of the Evolution of Sinhala Script*

3.2.3 Use of Numerals in Early Days

Numerals played an important role in the development of the writing system in Sri Lanka from the early days. For instance, in Sinhala books, manuscript book number, number of pages of a book, number of sections, number of verses, number of words, and, even number of letters in a book were marked on them. For this purpose, the historians used the characters of the alphabets, normally starting first with consonant letter ක (ka). Figure 11 shows the numerals used from 1st to 18th century A.D. Old Sinhala civilization before the 1st century B.C. initiates Brāhmī letters as numerals. By 1st century B.C. Sri Lankan invented their symbols as numerals (Paranavitana 1970), which were important for the development of irrigation, particularly mathematics for their engineering works and also for management and administration. The value of the numerals varies according to its place. The first book printed in Sri Lanka by Dutch indicated the year of the book printed as 1737 in Arabic numerals and 𑇡𑇴𑇧𑇳𑇬𑇧 as in Sinhala numerals (see Figure 11). The sequence of 1737 is interpreted as the result of adding 1 (𑇡) times 1000 (𑇴) to 7 (𑇧) times 100 (𑇳) to 30 (𑇬) and 7 (𑇧).



Value of Numerals	1	2	3	4	5	6	7	8	9	10	20	30	40	50	60	70	80	90	100	1000
1st B.C to 4th A.D.	𑁒	𑁓	𑁔	𑁕	𑁖	𑁗	𑁘	𑁙	𑁚	𑁛	𑁜	𑁝		𑁟		𑁡			𑁤	𑁥
18th A.D.	𑇡	𑇢	𑇣	𑇤	𑇥	𑇦	𑇧	𑇨	𑇩	𑇪	𑇫	𑇬	𑇭	𑇮	𑇯	𑇰	𑇱	𑇲	𑇳	𑇴

Figure 11: The Numerals used in Sinhala writing system

3.2.4 Technologies other than Language

Sri Lankan technology level before the 16th century A.D. was nothing comparable to most societies of the time. It is believed that in ancient time’s immense knowledge of trigonometry, some practical geometry and astronomy were well known to Sri Lanka.

Sri Lanka was excellent in irrigation technology including many trans-basin diversions, multi-purpose irrigation, drainage, flood control and conservation. The vast human-made ancient dams and lakes situated in the North Central Province since over 2,000 years are fully interconnected by a vast network of canals. One of these canals (*Yōda æla*; Giant Canal) has the surprisingly small gradient of 6 inches per mile for 10 miles and still in operation. This network of canals and lakes provides water for the irrigation of the paddy fields, which provide rice. Some writers reported that Sri Lankan 12th century A.D. ancient irrigation technology was unique and such technology could not seen in the rest of the world till 17th century A.D.

The massive civil engineering monuments, such as the *Jethwana Dagoba*, standing to this day as the tallest brick structure humankind has ever made are comparable with the highest pyramids of Egypt.

The world's first museum and library were built in Sri Lanka 2200 years ago. The museum housed the parts of the ship that brought the *Bodhi* sapling to Sri Lanka from India in 3rd century B.C. At one time, the Sinhalese ships were the biggest at Shanghai harbor (Chinese records), and history records a time when the representative of the Sinhalese sat on the right-hand side seat of Claudius Caesar.

In the health sector, many hospitals built by many Kings (*Parakramabahu* and *Dutugemunu* are appreciable) and therefore, most writers reported that Sri Lanka was very advanced concerning contemporary indigenous medicine technology.

3.2.5 Era of Type Printing

There was remarkable progress in documentation system after Portuguese who arrived in Sri Lanka in 1505 A.D. First European power seized control of the maritime districts; Portuguese was setting up of the administration in Sri Lanka with the proper documentation. The Portuguese carefully compiled the lists of villages so that the task of collecting taxes would be made easier. These lists (*tombōs*[11]) prepared by the Portuguese during their time was an essential contribution to the writing system of the country. Meanwhile, the Roman Catholic missionaries were interested in learning the Sinhala language to help them in their missionary activities in the country. As a result, the Colombo-born Jesuit *Reverent Emanuel* de Costa has written Sinhala Grammar book in Latin and the *Jesuit Pierre Berguin* who wrote one in Portuguese language and used to write religious books in Sinhala for them to use in their Church for religious education.

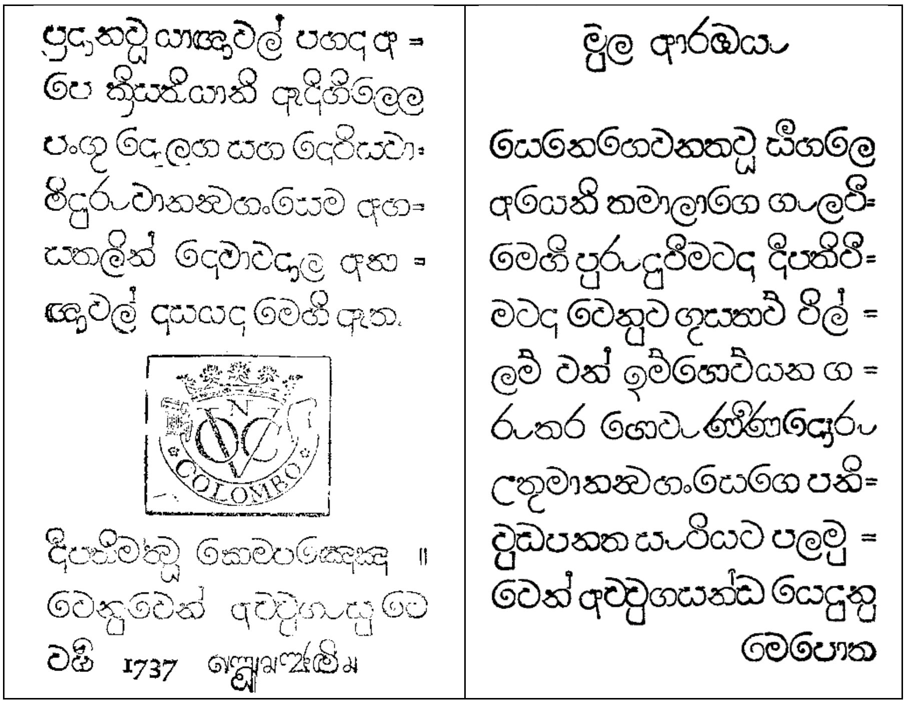
However, Portuguese rule came into the end in 1658, and next European power, Dutch became the masters of the Maritime Provinces (1656-1796) and maintained those records (tombōs), and also, they made a more significant contribution of charting the area on maps. Dutch started schools for Europeans and also for local people. In these schools, the language of the medium was Sinhala and Tamil, that is their mother tongue. Seminaries were established in *Jaffna* (City of the northern parts of Sri Lanka) in the 1760s for higher education, where languages were the subjects among the other, and hence, these educational activities demanded books in their languages. In the view of the fact that the production and issuing of handwritten ola leaf books for schools were not sufficient and as a result the circulation was poor. The lack of books and other documents necessary for education led to the necessity of printing press established in Sri Lanka.

As a result, the establishment of a printing press in Sri Lanka was seriously taken up first by Governor *Jacob Christian Pielat* in 1734. In his records, he said "*It is a fact that there are no better means of conviction than the learning and reading of God's Holy Word in one's language. I have greatly interested myself in this matter, and intend to bring it before the Honorable the Indian Government*" (Piyadasa, T. G., 1985). As a result, the Council of India approved the establishment

[11] The *Thombo* (from the Portuguese word '*tombo*', a register) contained names, detailed descriptions of the location and extent of each village as well as of the agricultural produce, including timber and fruit trees, fount there.

of first Sinhala printing office in Galle, Sri Lanka, during the period of Governor *Van Imhoff* (1736-1740) and in his memoir, he says: “The printing press established during my administration will be a useful instrument.”

In 1737, the first book published by *Gabriel Schade* was the Sinhala prayer book (41 pages), Figure 12 shows the sample two pages of the first of any size ever printed in Sinhala. *Gabriel Schade* became the first printer in Sri Lanka too. It was hand composed using founders types, cast abroad and brought over to Sri Lanka. These fonts were designed by foreigners, who lacked a proper understanding of the subtle variations in some characters. Most prominent similarities seen between (ට ව), (න ත), (භ හ), (ඇ ද) and (පී ජී). The Second printing press was established by *Van Imhoff* to print Sinhala and Tamil Christian literature. Printing of law and order notifications for public awareness was the other type of essential documents printed at that time (see Figure 13). Therefore, the starting of the printing trade in Sri Lanka by the Dutch was an important step taken towards the literature produced in the country.



පුදානවූ යාඥාවල් පශද ඇ =
පෙ ක්‍රිස්තියානි ඇදිහිල්ලෙ
පංගු දෙකලහ සහ දෙවිඝවා:
මිදුරුවානකවහංසෙම ඇගා=
සකලින් දෙමවදුල ඇසා =
ඥාවල් දසයද මෙහි ඇත.

දීපසිම්තු කොම්පඤ්ඤ ॥
මෙනුමෙන් අච්චුගැසු මෙ
වහි 1737

මූල ආරම්භය

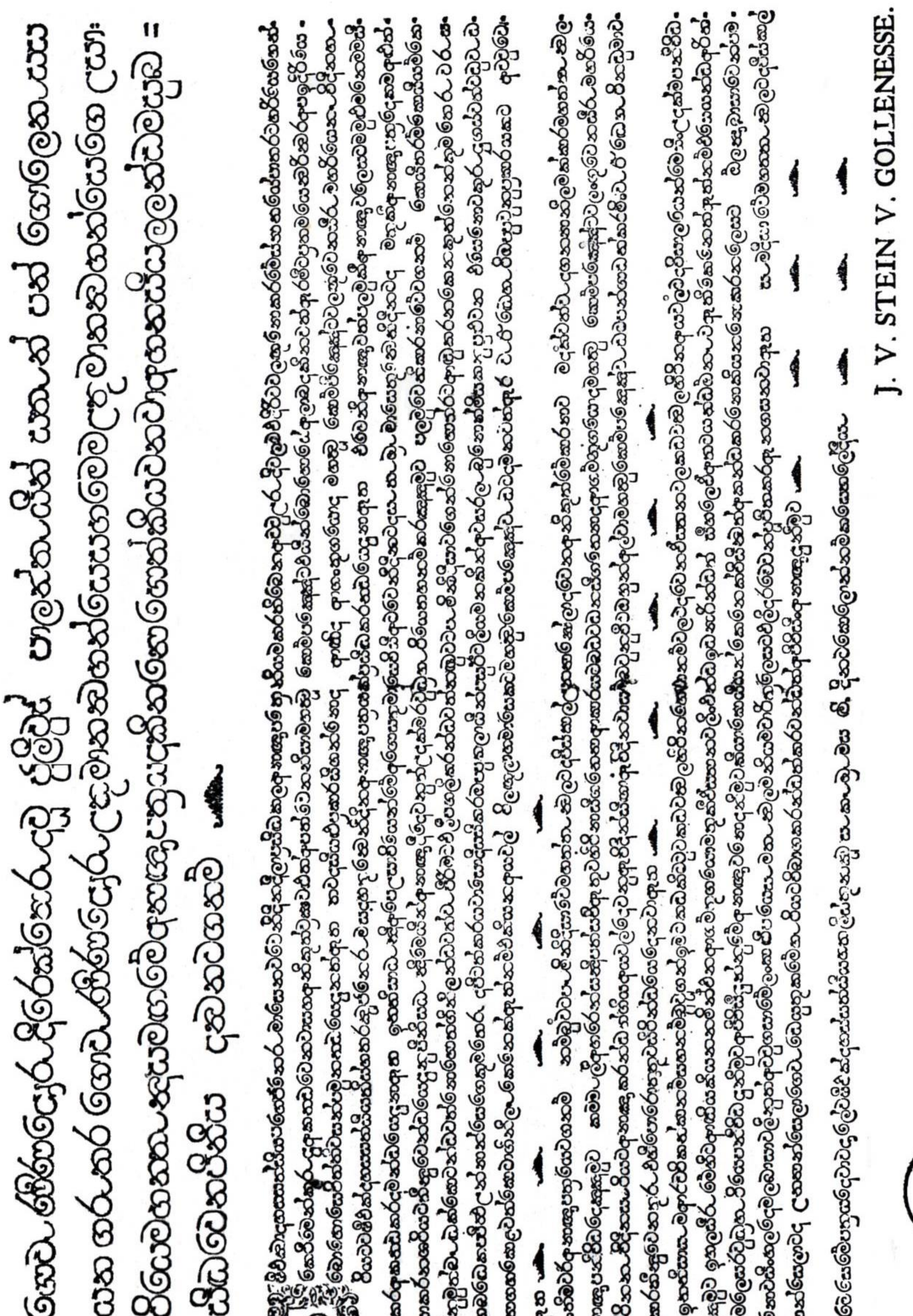
සොනෙගොවනතවූ සීහලෙ
අයෙති තමාලාගෙ භාලුවී=
මෙහි පුරුදුවීමටද දීපතිවී=
මටද වෙනුව ගැසනට වීල් =
ලම් එන් ඉම්හොවයන ග =
රුතර හොවු ඛ්‍යීඛෙද්‍යරු
උතුමානකවහංසෙගෙ පතී=
වූඩපනත සාරියට පලමු =
වෙත් අච්චුගැසන්ඩ සොදුනු
මෙපොත

Figure 12: Two pages from the first book printed in Sri Lanka (1737)

In 1876, the comprehensive samples register which is entitled “*ALFABETE DES GESAMMTEN ERDKREISES”* (“Alphabet of all race of the world”) which publishes the letter printing type of the world (Hof K. K., 1876) by The Royal Print Shop in 1876, Vienna, Germany (29 sheets). The printing type of 76 languages recorded in this sample register, among those that exists Sinhala and recorded as CINGALESISCH, which shows the European interest in Sinhala scripts. It is important to note that; even at that time the font width was a very significant factor for printing, and therefore they have paid due attention for the width of each font (Figure 14). Breakdown of the printing type, which published in this document, found as 13 vowels, 35 consonants, one semi consonant, 13 consonants modifiers, and 41 ligatures. The five vowels (ඍ, ඎ, ඏ, ඐ, ඖ), five consonants (ඟ, ඡ, ඞ, ඥ, ඦ), two medial signs (ෳ, ◌ෟ) and one diacritical mark (◌ඃ) were not found in this list though they are essential for the Sinhala printing.

In the meantime in Bengali, when the Serampore Missionaries[12] founded by William Carey on January, 10th 1800 (Daniel and Hedlund, 1993), He realized the complexity of socio-cultural problems due to its multilingual characters. William Carey and its Serampore Missionaries engaged, from the beginning of the 19th century, advocating the development and improvements of all the famous Indian languages including Sinhalese. As a result, the Sinhala printed letter shape and quality improved by its circularity and individuality. Figure 15 shows the early version of specimens of the versions of the sacred scriptures, of some of the Asian languages, including Sinhala.

[12] Serampore was then a Danish colony and it was a small town known also as Fredricksnagore after the name of King Fredrick V of Denmark.



ගෞව. ශ්‍රීෆයෝරු දිරෙක්ටෙරුදුවූ ජූලිවුස් වාලෙන්තැයින් ස්තැන් පන් ගොලෙනැස්ස
යන ගරුතර ගෞව. ශ්‍රීෆයෝරු උදුමානනවගන්සෙයගමෙමලදුමානනවගන්සෙගෙ උයාඃ
පියෙමගනන. නුදසමගමේඅනඥාපත්‍රයකිනගෙනගෙත්කියවනවාඅගනසියල්ලන්ඩමදුබ =
සීඩවෙනපිනිස අනනටගනම්

[illegible]

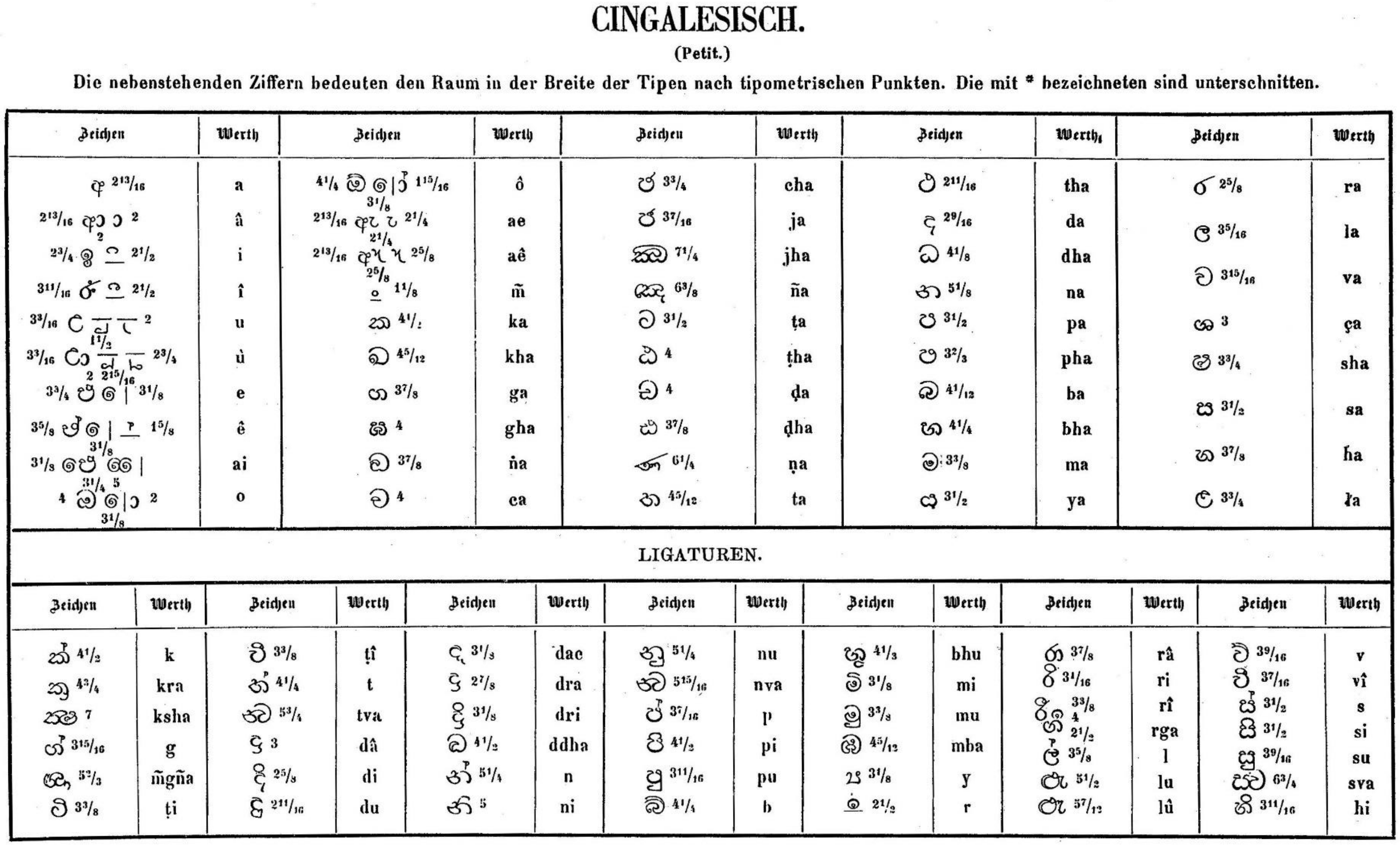
J. V. STEIN V. GOLLENESSE.

LS

Ter Ordonnantie van Welgemelte zyn Edele,
en den Raad.

J. v. D. SLEYDEN,
Secret.

Figure 13: The Gazette Notification issued on September 30th, 1743 (Original image was scaled down to 48%)



CINGALESISCH.

(Petit.)

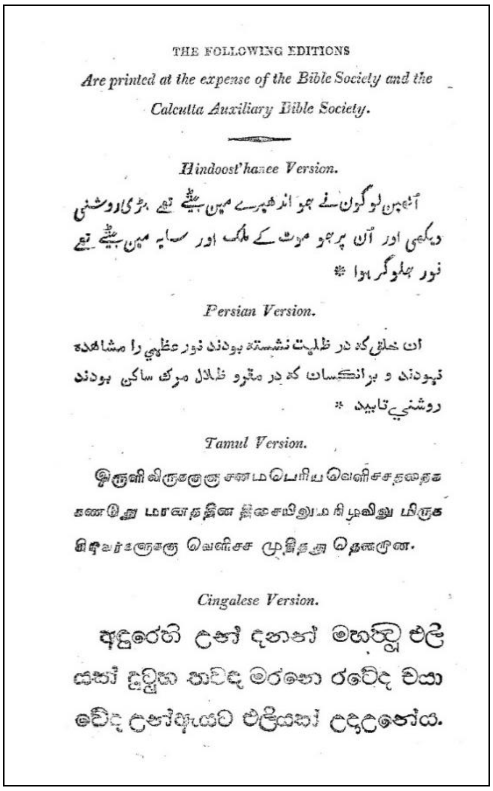
Die nebenstehenden Ziffern bedeuten den Raum in der Breite der Tipen nach tipometrischen Punkten. Die mit * bezeichneten sind unterschnitten.

Zeichen	Werth	Zeichen	Werth	Zeichen	Werth	Zeichen	Werth,	Zeichen	Werth
අ $2\frac{13}{16}$	a	$4\frac{1}{4}$ ඕ ෙ|ෝ $1\frac{15}{16}$ $3\frac{1}{8}$	ô	ඡ $3\frac{3}{4}$	cha	ථ $2\frac{11}{16}$	tha	ර $2\frac{5}{8}$	ra
$2\frac{13}{16}$ ආ ා 2 2	â	$2\frac{13}{16}$ ඇ ැ $2\frac{1}{4}$ $2\frac{1}{4}$	ae	ජ $3\frac{7}{16}$	ja	ද $2\frac{9}{16}$	da	ල $3\frac{5}{16}$	la
$2\frac{3}{4}$ ඉ ි $2\frac{1}{2}$	i	$2\frac{13}{16}$ ඈ ෑ $2\frac{5}{8}$ $2\frac{5}{8}$	aê	ඣ $7\frac{1}{4}$	jha	ධ $4\frac{1}{8}$	dha	ව $3\frac{15}{16}$	va
$3\frac{11}{16}$ ඊ ී $2\frac{1}{2}$	î	ං $1\frac{1}{8}$	m̃	ඤ $6\frac{3}{8}$	ña	න $5\frac{1}{8}$	na		
$3\frac{3}{16}$ උ ු ු 2 $1\frac{1}{2}$	u	ක $4\frac{1}{2}$	ka	ට $3\frac{1}{2}$	ṭa	ප $3\frac{1}{2}$	pa	ශ 3	ça
$3\frac{3}{16}$ ඌ ූ ූ $2\frac{3}{4}$ 2 $2\frac{15}{16}$	ù	ඛ $4\frac{5}{12}$	kha	ඨ 4	ṭha	ඵ $3\frac{2}{3}$	pha	ෂ $3\frac{3}{4}$	sha
$3\frac{3}{4}$ එ ෙ | $3\frac{1}{8}$	e	ග $3\frac{7}{8}$	ga	ඩ 4	ḍa	බ $4\frac{1}{12}$	ba	ස $3\frac{1}{2}$	sa
$3\frac{5}{8}$ ඒ ෙ | ් $1\frac{5}{8}$ $3\frac{1}{8}$	ê	ඝ 4	gha	ඪ $3\frac{7}{8}$	ḍha	භ $4\frac{1}{4}$	bha		
$3\frac{1}{8}$ ඓ ෛ | $3\frac{1}{4}$ 5	ai	ඞ $3\frac{7}{8}$	ṅa	ණ $6\frac{1}{4}$	ṇa	ම $3\frac{3}{8}$	ma	හ $3\frac{7}{8}$	ha
4 ඔ ෙ|ා 2 $3\frac{1}{8}$	o	ච 4	ca	ත $4\frac{5}{12}$	ta	ය $3\frac{1}{2}$	ya	ළ $3\frac{3}{4}$	ḷa

LIGATUREN.

Zeichen	Werth	Zeichen	Werth	Zeichen	Werth	Zeichen	Werth	Zeichen	Werth	Zeichen	Werth	Zeichen	Werth
ක් $4\frac{1}{2}$	k	ටී $3\frac{3}{8}$	ṭî	දැ $3\frac{1}{3}$	dac	නු $5\frac{1}{4}$	nu	භූ $4\frac{1}{3}$	bhu	රා $3\frac{7}{8}$	râ	ව් $3\frac{9}{16}$	v
ක්‍ර $4\frac{3}{4}$	kra	ත් $4\frac{1}{4}$	t	ද්‍ර $2\frac{7}{8}$	dra	න්ව $5\frac{15}{16}$	nva	මි $3\frac{1}{8}$	mi	රි $3\frac{1}{16}$	ri	වී $3\frac{7}{16}$	vî
ක්‍ෂ 7	ksha	ත්ව $5\frac{3}{4}$	tva	දෘ $3\frac{1}{8}$	dri	ප් $3\frac{7}{16}$	p	මු $3\frac{3}{8}$	mu	රී $3\frac{3}{8}$ 4	rî	ස් $3\frac{1}{2}$	s
ග් $3\frac{15}{16}$	g	දා 3	dâ	ද්ධ $4\frac{1}{2}$	ddha	පි $4\frac{1}{2}$	pi	ඹ $4\frac{5}{12}$	mba	ර්ග $2\frac{1}{2}$	rga	සි $3\frac{1}{2}$	si
ඥ $5\frac{2}{3}$	m̃gña	දි $2\frac{5}{8}$	di	න් $5\frac{1}{4}$	n	පු $3\frac{11}{16}$	pu	්‍ය $3\frac{1}{8}$	y	ල් $3\frac{5}{8}$	l	සු $3\frac{9}{16}$	su
ටි $3\frac{3}{8}$	ṭi	දු $2\frac{11}{16}$	du	නි 5	ni	බ් $4\frac{1}{4}$	b	ර්‍ $2\frac{1}{2}$	r	ලු $5\frac{1}{2}$	lu	ස්ව $6\frac{3}{4}$	sva
										ලූ $5\frac{7}{12}$	lû	හි $3\frac{11}{16}$	hi

Figure 14: Sinhala printing types from the ALFABETE DES GESAMMTEN ERDKREISES AUS DER k.k. HOF- UND STAATS DRUCKEREI INWIEN, 2nd edn (Vienna, 1876) (The accompanying figures signify the space in width in points for each symbol. Original image was scaled down to 75%)



THE FOLLOWING EDITIONS

Are printed at the expense of the Bible Society and the Calcutta Auxiliary Bible Society.

Hindoost'hanee Version.

اُنہین لوگون نے جو اندھیرے مین بیٹھے تھے بڑی روشنی
دیکھی اور اُن پر جو موت کے ملک اور سایہ مین بیٹھے تھے
نور جلوہ گر ہوا *

Persian Version.

ان خلق كه در ظلمت نشسته بودند نور عظيمي را مشاهده
نمودند و برانكسان كه در مقرو ظلال مرك ساكن بودند
روشني تابيد *

Tamul Version.

இருளி லிருக்கிற சனம பெரிய வெளிச்சத்தைக
கண்டது மரணத்தின் திசையிலுமந் நிழலிலு மிருக
கிறவர்களுக்கு வெளிச்ச முதித்தது தென்றுண.

Cingalese Version.

අඳුරෙහි උන් දනන් මහත්වූ එලි
යක් දුටුහ තවද මරණෙ රටේද ඡායා
වේද උන්නෑයට එලියක් උදාවෝය.

Figure 15: Specimens of translations of the version of the Sacred Scriptures CINGALESE (Sinhala) and TAMUL (Tamil) printed by Serampore Missionaries in 1813.

3.2.6 Printing Establishments in Sri Lanka

Until the time of the ***Colebrooke Commission*** (1930), there were few printing establishments. These included the Government Printing Establishment and a few other presses owned by English Missionary Societies. The first Sinhala newspaper registered under the ordinance was "*Lakmini Pahana*". This newspaper commenced publication on 17th September 1862. *Gunatilake Athapattu Mudiyar* of *Galle, Pandit. L.W. Batuwatudawe, Koggala Pandit, Ven. Walane Siddhartha* was instrumental in giving birth to this paper. The First editor was *Koggala Panditatilake* followed by *Pandit Batuwatudawe* from 1st July 1866. *Matara Dharmaratne* succeeded him on 28th July 1883 up to the year 1924. *Munidasa Kumaratunga* became its editor on 20th June 1934, and as a result, this newspaper gained popularity. "*Lakrivikirana*” was a Sinhala Buddhist paper, which fought for the rights of the Buddhists. In the year 1891 *Lakrivikirana* became a daily newspaper. The first daily Sinhala newspaper was "*Dinapatha Pravurthi*".

The productions of wooden types used Tar tree wooden (see Figure 16) used for printing of posters for headlines in a couple of tabloid news.

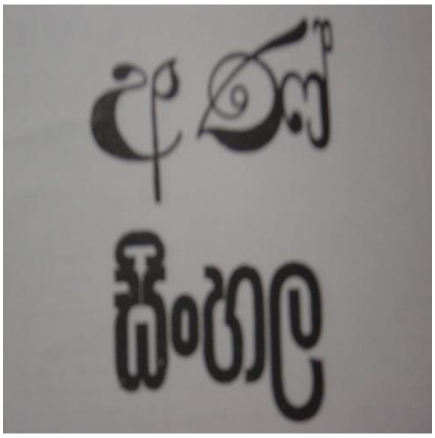


Figure 16: Wooden Types used to print headlines

3.2.7 Sinhala Monotype Type Faces

The Sinhala Printing was well established in Sri Lanka by the mid 19th century. A Monotype Sinhalese font Series No. 557 (*Sinhalese*), No. 657 (*Sinhalese Bold*), No. 698 (*Sinhalese Italic*) and No. 699 (*Sinhalese Bold Italic*) consisted of 302 characters and 26 punctuation marks and numbers in each font types (Monotype, 1959). Figure 17 gives the complete set of Monotype Sinhala character Series No. 557, which was in use since 1960, in Monotype machines. The first machine[13] of this type was established in Sri Lanka in 1904 at the Sri Lanka Government Printing Department, and such machines were used until the year 2005 (see Figure 18 & 19).

[13] This machine composes type by casting new, single types in correct order, ready for printing and it perpetuating the concept of a separate keyboard and caster interfaced by a 31-channel punched paper tape. The keyboard consisted of a two-alphabet layout augmented by four shift keys. In the caster, the matrix-case contained 324 characters arranged in 18x18 rows. Spaces between words are varied by the system (in steps of 0.0005 inch!) to exactly justify each line. The system was devised by *Tolbert Lanston* and others in the USA with British cooperation about 1890

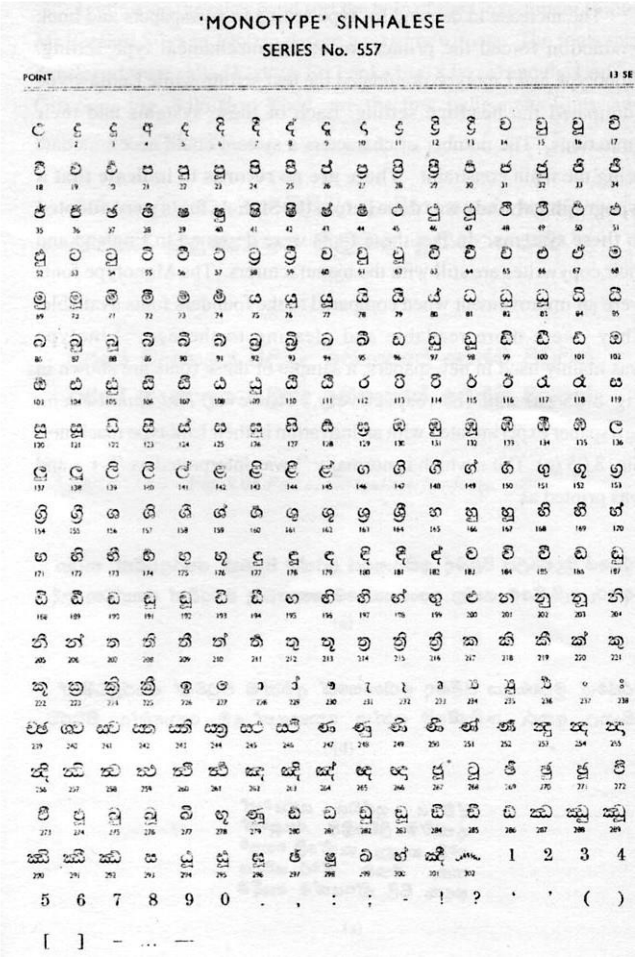


Figure 17: Complete Sinhala Monotype Font used in Monotype Composition Caster

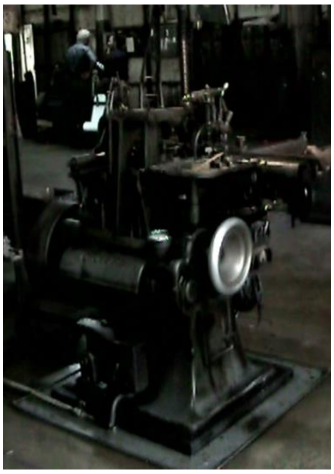


Figure 18: Monotype Composition Caster (Courtesy of Sri Lanka Government Printing Department)

The increase in demand for typesetting in newspapers and book production forced printing to utilize mechanical typesetting. Monotype and Linotype dominated the text setting, while the Ludlow type dominated the headline set.

Ludlow, on the other hand, sorts the help of the Government Printer Mr Bernard Silva in 1960 to design a set of new fonts. The fonts thus developed were called Ludlow Sri Lanka. They are the first indigenous fonts and became the trendsetter.

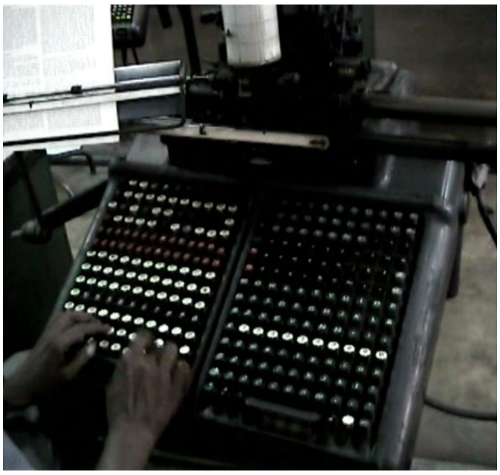


Figure 19: Monotype Composition Keyboard (Courtesy of Sri Lanka Government Printing Department)

Literary Sinhala achieved its standard in the 14th century A.D., and this standard is respected and observed throughout Sri Lanka. Figures 20 provide some examples of Sinhala, along with their meaning in English. As Figure 20 (bottom), shows Sinhala differs not only in its forms and structure, but also in its use and function—that is, in Sanskrit, Pali, classical writing, general writing, and in regional and contemporary speaking.

ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රීය සමාජවාදී ජනරජය
Democratic Socialist Republic of Sri Lanka

ශ්‍රී ජයවර්ධන පුර කෝට්ටේ
Sri Jayawardhana Pura Kotte

ශක්‍ර දේවේන්ද්‍ර යන් විසින් කුස රාජෝත්තමයානන් වහන්සේට
දුන් මාණික්‍යය තෙම අට කොනකින් වක් වූයේය.
The gem that was given to King Kusa by Sakra, the Load of gods, was curved in eight places.

Figure 20: (Top) Official name of the government of Sri Lanka; (Middle) capital city of Sri Lanka; (Bottom) a classical Sinhala sentence showing the mixed alphabet.

3.3 Languages Used in Present Society

The three living languages in Sri Lanka are Sinhala, Tamil and English. They are used for general, everyday communication: both interpersonal and mass communication. Of them, Sinhala and Tamil are considered 'national languages' while English is considered as a 'link language' to link the major ethnic groups of the Island. Thus written documents, on paper or other materials, appear in one, two or all of these languages.

Sinhala is a language of the Indo-Aryan sub-family of the larger Indo-European family. Tamil is a language of the Dravidian family used in south India. In Sri Lanka, the Tamil language used by Sri Lankan Tamils and Indian Tamils. The Indian Tamils who arrived in Sri Lanka and they work in the tea plantations, and the Muslims who use Arabic for religious activities also speak in Tamil and Sinhala.

There are also four classical languages, as mentioned earlier used mainly for religious purposes. In terms of religion, Sri Lanka has all the major religions of the world: Buddhism, Hinduism, Islam and Christianity. Each of these religions has a classical language used in the conduct of religious transactions, including traditional rites and rituals (see Figure 3).

3.4 Sinhala Language

The Sinhala writing system is a syllabary derived from the ancient North Indian scripts, Brāhmī, started to appear in inscriptions in the 3rd centuries B.C. as indicated before, and subsequently alphabet and writing patterns have changed considerably with notable influenced by the *Kabamba* and *Pallawa Grantha* script of South India respectively (*Florian Coulmas, 1996*). However, the modern script used in writing Sinhala is unique to this language. The Sinhala language has distinct and unique features. Sinhala has 61 scripts, the most extensive vowel set in the world. The Sinhala character set is consist of vowels, consonants and semi-consonants. Sinhala is a syllabic alphabet in which all consonants have an inherent vowel /a/. This alphabet differs from all other Indo-Aryan languages in that it contains distinctive sounds that are unique to it since 8th century A.D. Literary Sinhala obtained its standard in the 14th century A.D., and this standard respected by the whole Sinhala speaking community of Sri Lanka. Sinhala alphabet has a pair of unique letters (ඇ *æ* and ඈ *ǣ*) to represent two vowel sounds (*see* Figure 10, *Viyaulpotha Pillar Inscription*). The presence of a set of five nasal sounds known as "half nasal" or "prenasalized stops" in modern writing is unique to the Sinhala language too (ඟ *ňga*, ඦ *ňja*, ඬ *ňḍa*, ඳ *ňda* and ඹ *mba*). Diacritics can appear before, below, above or after the consonant they belong to, are used to change the inherent vowel. When diacritics appear the beginning of a syllable, vowels written as independent letters.

When certain consonants occur together, i.e. නන්ද, special conjunct symbols are used which combine the essential parts of each letter like නන්‍ද. However, this is not the general practice of the modern writing system. A subset of letters was used to write classical Sinhala, and some additional letters added to the Sinhala alphabet, they know as '*miśra hōdiya'* (මිශ්‍ර හෝඩිය; mix alphabet) is used to write Sanskrit and Pali loan words. To end with, De Silva (Silva, 1982) describes the language, "The belief that Literary Sinhalese is superior, more beautiful, more logical and more correct prevails at every level of the society".

The Sinhala language may manifest itself in two media; sound medium in spoken Sinhala and graphical medium in written Sinhala. Spoken Sinhala uses thirty-nine sound (39), of which twelve (12) of them are vowels, twenty-two (22) of them are consonants (nasals), and five (05) of them are half-nasals (*see* Figure 21). Out of 5 half-nasals, one ňja (ඦ) half-nasals not used in spoken Sinhala. When carefully examined the usage of Sinhala language from 3rd century B.C. to 12th century A.D. (*see* Figure 10), there have computerization only minimum 26 or maximum 39 scripts for writing too, which of the nine are aspirations used when writing partly due to the influence from Brāhmī scripts and Pali. It also notices that only five (05) vowel sound (*a, i, u, e, o*) used for writing too. Many of the aspirations used in 3rd to 1st century B.C., which browed from Brāhmī script have ignored after 1st century A.D. shows that the

Sinhala language does not use them. During the 1st to 5th century A.D. only four (04) aspiration letters were pound in *Situlpawwa*, *Anuradhapura* and *Viharagala* inscriptions to write *Dham-ma* (ධම්ම; religious merit), *ve-dha* (වෙඪ; interests), *va-tha* (වථ; land), and *san-gha* (සංඝ; Monk) are more common in Pali writing. Uses of half-nasals were not exits till 18th century A.D. and are used only in modern Sinhala.

18 Vowels in the Sinhala Language	
අ ආ ඇ ඈ ඉ ඊ උ ඌ සa සaa ඏ ඐ එ ඒ ඓ ඔ ඕ ඖ	
a ā æ ǣ i ī u ū ṛ ṝ ḷ ḹ e ē ai o ō au	
Used in Sinhala writing	12 {අ ආ ඇ ඈ ඉ ඊ උ ඌ එ ඒ ඔ ඕ }
Used in Sanskrit writing	06 {සa සaa ඏ ඐ ඓ ඖ}
41 Consonants in the Sinhala Language	
ක ඛ ග ඝ ඞ ඟ ච ඡ ජ ඣ ඤ ඥ ඦ ට ඨ ඩ ඪ ණ ඬ ත ථ ද ධ න ඳ ප ඵ බ භ ම ඹ ය ර ල ව ශ ෂ ස හ ළ ෆ	
ka kha ga gha ṅa ňga ca cha ja jha ña jña ňja ṭa ṭha ḍa ḍha ṇa ňḍa ta tha da dha na ňda pa pha ba bha ma mba ya ra la va śa ṣa sa ha ḷa fa	
Used in Sinhala writing (Spoken Sinhala [nasals])	22 {ක ග ච ජ ට ඩ ණ ත ද න ප බ ම ය ර ල ව ශ ස හ ළ ෆ}
Used in Sinhala writing (Spoken Sinhala [Half-nasals])	05 {ඟ ඦ ඬ ඳ ඹ}
Used to write Sanskrit and Pali ([Aspirations])	10 {ඛ ඝ ඡ ඣ ඨ ඪ ථ ධ ඵ භ}
Used to write Sanskrit, Pali and Sinhala	04 {ඞ ඤ ඥ ෂ}
02 Diacritical Marks	
◌ං ◌ඃ	
ṅ ḥ	
16 Vocalic and Combined Vocalic Strokes to represent Vowels (Vowels ඏ and ඐ do not use in modern Sinhala)	
◌් ◌ා ◌ැ ◌ෑ ◌ි ◌ී ◌ු ◌ූ ◌a ◌aa ◌ෙ ◌ේ ◌ෛ ◌ො ◌ෝ ◌ෞ	
ā æ ǣ i ī u ū ṛ ṝ e ē ai o ō au	
03 Medial Signs	
ෟ ◌ෳ ◌	

Figure 21: Sinhala alphabet and their usage

Even during the 6th to 10th century A.D., the Sinhala language well illustrated in the rock fortress of *Sigiriya*, in the *Sigiri* Graffiti (*Sigiri Gee)*, people used only spoken Sinhala to express and write their insatiable taste for beauty, came from all four corners of Island. They saw "*Sigiri Apsara"*, and their ecstasy poured out in spontaneous "*Prathibha*", people's expressions of an enthusiastic intelligent imagination. These they scribed on the gallery wall as a memento of their visit to Sigiriya. These compositions are simple in style, harmonious in rhythm, proportioned and finished according to rules of poesy. In one Graffiti, the writer says, "*The ladies who wear golden chains on their breasts beckon to me. Now I have seen these resplendent ladies. Heaven has lost its appeal for me*" *(S. Paranawithana)*. These graffiti belong to the form of Sinhala writing (*see* Figure 22). These appear by the time Pali and Sanskrit had become almost the language of some degree of the educated class. *Dr Gunadasa Amarasekera* in his *Sinhala Kavya Sampradaya* reminds us that the language of Sinhala poetry is still Sinhala in the sense that the Sinhala poets in general use a language that does not have Sanskrit words or Pali way of writing (*see* Figure 23). The detailed analysis of occurrences of the characters with 70129 words list compiled by the UCSC/LRTL[14] and results listed according to the categories in Figure 21. The results are indicating that the use of consonants to write Sanskrit words in modern Sinhala is decreasing, and that is less than 2% of the total usage (*see* Table 13). The divergence has been continuing even today and therefore can agree with what *Florian Coulmas* says, (1996), "As a result of the continuing divergence of the spoken varieties form the highly *Sanskritized* writing variety, modern Sinhala exits in a situation of Diglossia[15]." (Florian Coulmas, 1996)

ස්වශ්ති! බුද්ගමු වෙහෙර නියමජෙත් න සගස්මි ලෙඑමි
බලන්නට ගහනුන් බලුමො සිහිගිරි
අවුත් තුට් සිත් කල මරෙනන් දක නො මුස්නෙයෙ මෙය

Sva-śti!   Bud-ga-mu ve-he-ra ni-yam-jet na- sa-gas-mi le-e-mi
Ba-lan-na-ṭa ga-ha-nun ba-lu-mo si-hi-gi-ri
A-*vuj* tu-ṭ si-t ka-la ma-ra-nen da-ka no mus-ne-ye ma-ye

Hail! I am Na Sagasmi, the chief administrator (ni-yam-jet) of
Budgamu vehera is writing (le-e-mi)
We looked at Sihigiri in order to look at women
Having come here, having seen them, death does not worry me

Figure 22: Sigiri Graffiti

මෙරට හැටි - මුනිදාස කුමාරතුංග

පැර   කුම්   දුටු ගැමුණු නම කියමින්   නිත   ර
පැර   දුම්   විදිති ගැහැනුන් අතිනුදු   නිත   ර
සැර   සුම්   කළ මෙකල වැව් වනමින්   පත   ර
ගැර   හුම්   ලබති සිද බඩගින්නේ   නිත   ර

Merata Heti - Munidasa Kumaratunga

මගේ රට - පී. බී. අල්විස් පෙරේරා

රුහුණේ කුඹුරු සරු සාරයි එක   හී   යයි
මින්නේරියේ ජනපදයයි වෙල්   යා   යයි
රජරට පලාතේ පමණක් වැව්   සී   යයි
මේවා බලන්නට මෙන් ඉර හඳ   පි   යයි

My country (1947) by P.B. Alwis Perera

Figure 23: Sample of Sinhala poets

3.4.1 Major Features of the Writing System

The following features are notable in the use of Sinhala writing systems.

[14] Language Research Technology Laboratory (LRTL) of the UCSC create a Sinhala Corpus Beta Version and it consist of 70142 distinct Sinhala worlds.

[15] A language situation characterized by a conspicuous divergence between a literary or 'high' variety and the colloquial or 'low' variety. In many cases such divergence is a consequence of the attempt to conserve a writing variety in a from recognized as classic or associated with religious significance (*Florian Coulmas, 1996*)

Every vowel except the first one has a corresponding vowel modifier symbol, which can be attached to consonants to make composite characters.

When vowels appear at the beginning of a word, vowels written as independent letters, on the other hand, when a pure consonant combined with the first vowel (අ), it leads to the common consonant form, i.e. the consonant sign has an inherent vowel /*a*/ associated with it.

Figures 24, 25 and Figure 26 illustrate how some consonants and vowel signs are combined to form syllable blocks (glyphs). According to the shape of the Sinhala letter, some glyphs constructed differently. Some of the constructed glyphs would create a somewhat uneven, irregular and illogical outer appearance.

Figure 24 depicts some of the consonants combined with vowel signs. It should be noted that the visual appearances of the final form may be wholly different from their constituents. Appearances of modifiers also vary according to the consonants.

Consonants & Vowel Signs				Conjuncts Forms	
ර	◌්			=	ර්
ක	◌්			=	ක්
ජ	◌්			=	ජ්
ච	◌්			=	ච්
ර	◌ැ			=	රැ
ර	◌ෑ			=	රෑ
ක	◌ැ			=	කැ
ත	◌ු			=	තු
ර	◌ු			=	රු
ර	◌ූ			=	රූ
ළ	◌ු			=	ළු
ළ	◌ූ			=	ළූ
ක	◌ු			=	කු
ක	◌ූ			=	කූ
ෙ◌	ච	◌්		=	චේ
ෙ◌	ර	◌්		=	රේ
ෙ◌	ක	◌ා	◌්	=	කෝ

Figure 24: Illustration of some of the conjuncts in Sinhala with their vowel signs attached to it. Conjuncts forms vary depending on the consonants and vowel signs.

There are two commonly used diacritical marks: '*anuswara*' and '*visarga*', like most of the Indic languages. They may appear only followed by a vowel or a consonant with an implicit or explicit vowel.

Also, three special symbols are corresponding to the letter of ◌්‍ර (*rak*), ◌්‍ය (*yan*) and ර්‍◌ (*rep*) called *rakaransaya, yansaya,* and *repaya,* respectively. The symbols *repaya* and *rakaransaya* represent the sound *ra* (ර), and *yansaya* represent *ya* (ය) respectively, following a pure consonant. Some examples are listed in Figure 25. They may appear alone with consonants, or consonants together with a vowel sign.

Semi-consonant	Consonants & Vowel Sings							Conjuncts Forms
rakaransaya	ක	◌්	ර				=	ක්‍ර
rakaransaya	ක	◌්	ර	◌ී			=	ක්‍රී
rakaransaya	ක	◌්	ර	◌ී	◌ං		=	ක්‍රීං
rakaransaya	ක	◌්	ෙ◌	ර	◌ා	◌්	=	ක්‍රෝ
rakaransaya	ද	◌්	ෙ◌	ර	◌ා	◌්	=	ද්‍රෝ
yansaya	ක	◌්	ය	◌ු			=	ක්‍යු
yansaya	ක	◌්	ෙ◌	ය	◌ා	◌්	=	ක්‍යෝ
repaya	ධ	ර	◌්	ම			=	ධර්‍ම

Figure 25: Example of a few Sinhala characters with semi-consonant 'rakaransaya', 'yansaya' and 'repaya' attached to it.

When Sanskrit and Pali words adopted into Sinhala, they transcribed in the compound manner in which they have written in Sanskrit and Pali. This composition is effected by the union of one or more consonants, or their parts or symbols, with vowels-consonants or its parts or symbols, and vice versa (see Figure 26).

Keystroke Sequence										**A modern way of writing**	**Pali and Sanskrit way of writing**
ක	◌්	ෙ◌	ෂ	◌්	න	◌්	ද	◌්	ර	ක්ෂේන්ද්ර	ක්‍ෂේන්‍ද්‍ර
බ	◌ු	ද	◌්	ධ						බුද්ධ	බුද්‍ධ
ඉ	න	◌්	ද	◌්	ර	◌ි	ය	න	◌්	ඉන්ද්රියන්	ඉන්‍ද්‍රියන්

Figure 26: Illustration of some of the Sinhala consonants and vowel sign combinations that form different glyphs in Sinhala, Pali & Sanskrit respectively

Having looked at the basic structure and significant features of Sinhala scripts, in the following section, the author examines the complexity involved in the machine representation.

In Pali writing, it has a set of conjunct-consonants amount to 68 (see Figure 27).

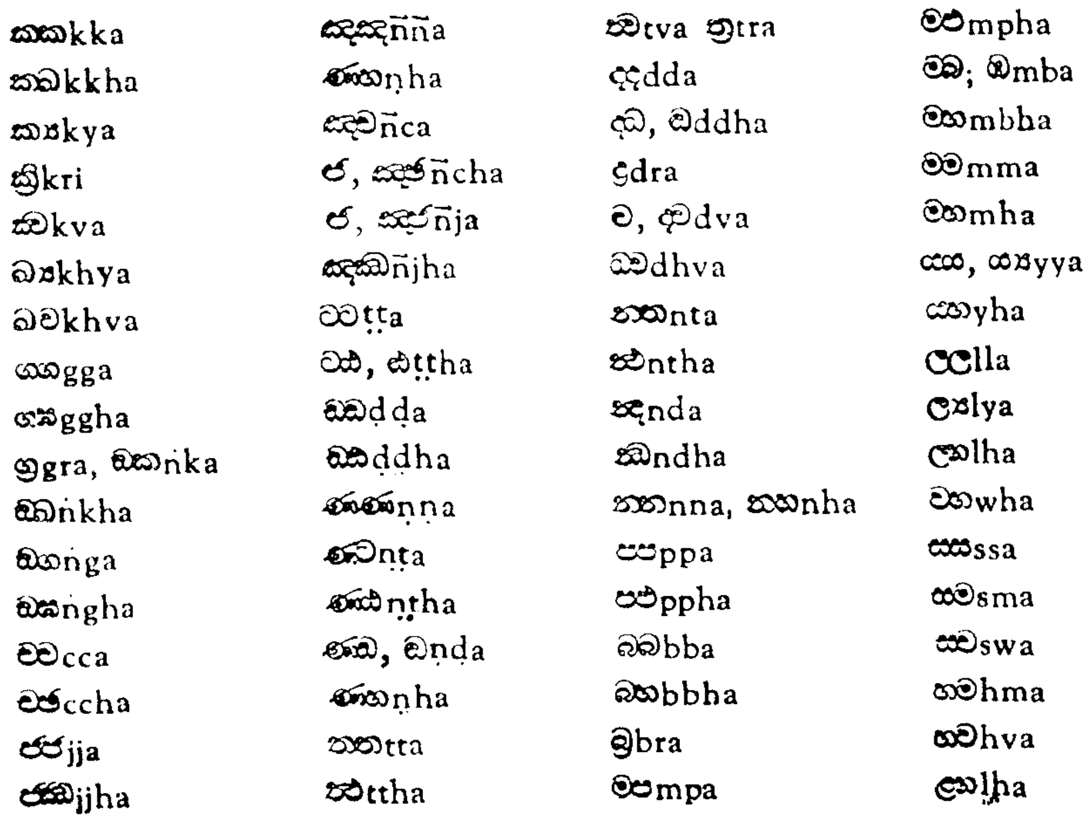
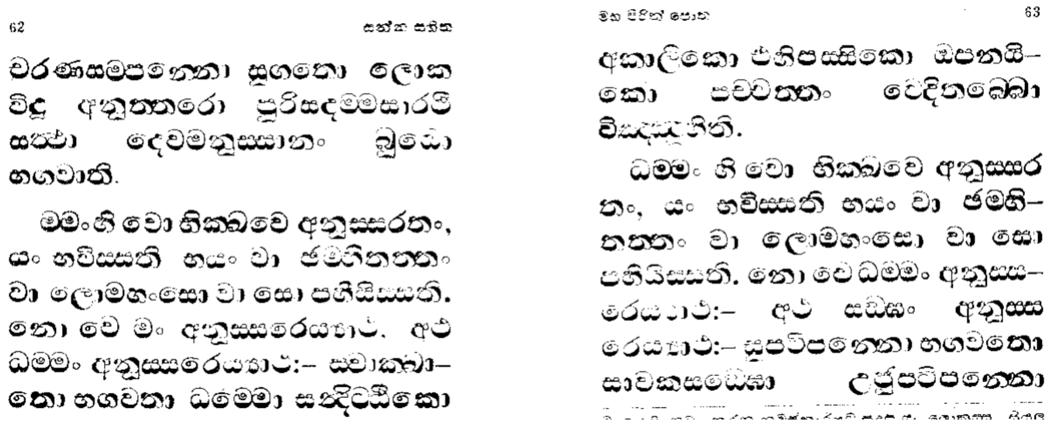

kka
kkha
kya
kri
kva
khya
khva
gga
ggha
gra, ṅka
ṅkha
ṅga
ṅgha
cca
ccha
jja
jjha
ññа
ṇha
ñca
ñcha
ñja
ñjha
ṭṭa
ṭṭha
ḍḍa
ḍḍha
ṇṇa
ṇṭa
ṇṭha
ṇḍa
ṇha
tta
ttha
tva tra
dda
ddha
dra
dva
dhva
nta
ntha
nda
ndha
nna, nha
ppa
ppha
bba
bbha
bra
mpa
mpha
mba
mbha
mma
mha
yya
yha
lla
lya
lha
wha
ssa
sma
swa
hma
hva
ḷha

Figure 27: Complete set of Pali conjunct-consonants used in the Sinhala language[16]

When Pali words adopted into Sinhala, they transcribed in the compound manner in which they have written in Pali. These conjunct letters are written in Pali language convention, a pure consonant (letter with Halant) followed by another consonant can be represented by writing the consonants touching each other. If the preceding pure consonant has half letter form following consonant can be combined and formed conjunct letters. The special symbols *rakaransaya*, *yansaya* and *repaya* can be written using the conjunct or non-conjunct letters (see Figure 28).



62 සත්ත සුත්ත

චරණසම්පන්නො සුගතො ලොක
විදූ අනුත්තරො පුරිසදම්මසාරථි
සත්ථා දෙවමනුස්සානං බුද්ධො
භගවාති.

මමං හි වො භික්ඛවෙ අනුස්සරතං,
යං භවිස්සති භයං වා ඡම්භිතත්තං
වා ලොමහංසො වා සො පහීයිස්සති.
නො චෙ මං අනුස්සරෙය්‍යාථ, අථ
ධම්මං අනුස්සරෙය්‍යාථ:– ස්වාක්ඛා–
තො භගවතා ධම්මො සන්දිට්ඨිකො

මහ පිරිත් පොත 63

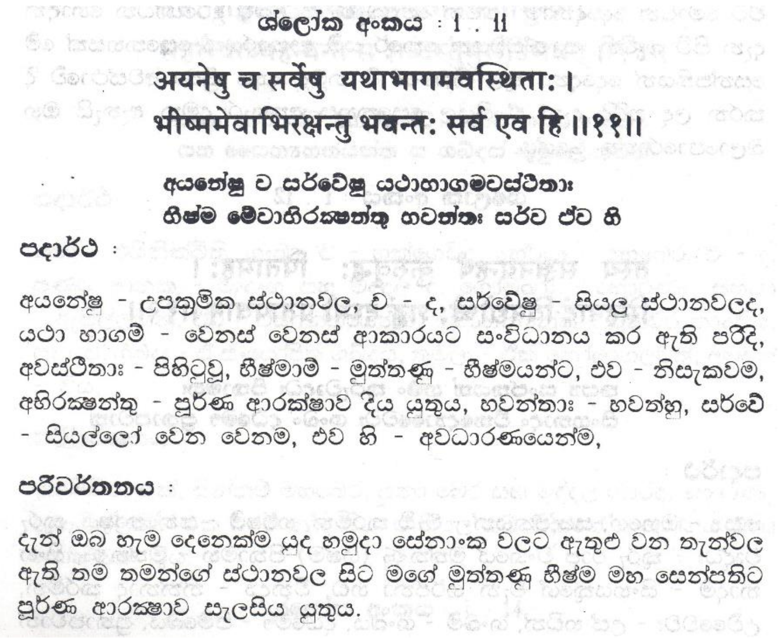
අකාලිකො ඒහිපස්සිකො ඕපනයි–
කො පච්චත්තං වෙදිතබ්බො
විඤ්ඤූහීති.

ධම්මං හි වො භික්ඛවෙ අනුස්සර
තං, යං භවිස්සති භයං වා ඡම්භි–
තත්තං වා ලොමහංසො වා සො
පහීයිස්සති. නො චෙ ධම්මං අනුස්ස–
රෙය්‍යාථ:– අථ සඞ්ඝං අනුස්ස
රෙය්‍යාථ:– සුපටිපන්නො භගවතො
සාවකසඞ්ඝො උජුපටිපන්නො

Figure 28: Sample specimen of Pali Sutra

[16] A language situation characterised by a conspicuous divergence between a literary or ‘high’ variety and the colloquial or ‘low’ variety. In many cases such divergence is a consequence of the attempt to conserve a writing variety in a from recognized as classic or associated with religious significance (*Florian Coulmas, 1996*)

As shown in Figure 29, when letters written in Sanskrit, pure consonants can represent a conjunct form or non-conjunct form.



ශ්ලෝක අංකය : 1 . 11

अयनेषु च सर्वेषु यथाभागमवस्थिताः ।
भीष्ममेवाभिरक्षन्तु भवन्तः सर्व एव हि ॥११॥

අයනේෂු ච සර්වේෂු යථාභාගමවස්ථිතාඃ
භීෂ්ම මේවාභිරක්ෂන්තු භවන්තඃ සර්ව ඒව හි

පදාර්ථ :

අයනේෂු - උපක්‍රමික ස්ථානවල, ච - ද, සර්වේෂු - සියලු ස්ථානවලද, යථා භාගම් - වෙනස් වෙනස් ආකාරයට සංවිධානය කර ඇති පරිදි, අවස්ථිතාඃ - පිහිටුවූ, භීෂ්මාම් - මුත්තණු - භීෂ්මයන්ට, එව - නිසැකවම, අභිරක්ෂන්තු - පූර්ණ ආරක්ෂාව දිය යුතුය, භවන්තාඃ - භවත්හු, සර්වේ - සියල්ලෝ වෙන වෙනම, එව හි - අවධාරණයෙන්ම,

පරිවර්තනය :

දැන් ඔබ හැම දෙනෙක්ම යුද හමුදා සේනාංක වලට ඇතුළු වන තැන්වල ඇති තම තමන්ගේ ස්ථානවල සිට මගේ මුත්තණු භීෂ්ම මහ සෙන්පතිට පූර්ණ ආරක්ෂාව සැලසිය යුතුය.

Figure 29: Section of a page from the Bhagavatgeetava written in Devanagari and Sanskrit (with Sinhala letters) and translation in Sinhala

3.5 Sinhala Scripts Structure and Major Issues

3.5.1 Introduction

Sinhala script which is used for writing in Sinhala language in Sri Lanka is said to have derived from the ancient scripts of Brāhmī, known to have existed since the third to the second century B.C. Subsequently, alphabet and writing system have changed considerably with notable influenced by the *Kadamba* and *Pallawa Grantha* script of South India (Florian Coulmas, 1989). Full Sinhala script includes the symbols necessary for writing loan words from Sanskrit and Pali, mainly the aspirated consonants.

There are two alphabets in the current Sinhala writing system. It contains an "alphabet within an alphabet" —namely, the *Eḷu*[17] alphabet and the "Mixed" alphabet (see Figure 31). The *Eḷu* alphabet, a subset of full language as described in the classical grammar *Sidatsaňgarā* (1300 A.D.) is made up of letters used in writing pure Sinhala words and is still widely regarded as authoritative. It contains 33 letters, of which 12 are vowels and 21 consonants. The "mixed" alphabet comprises letters of the *Eḷu* alphabet and the Sanskrit alphabet. It contains 61 letters, of which 18 are vowels, 41 consonants and two diacritical marks. Writing Sinhala maintains ten aspiration letters (10) and four (04) more consonants in order to fulfill the usage of loan words from Sanskrit and Pali. The mixed alphabet used in writing *Eḷu*, Pali, Sanskrit and foreign words naturalized in the language. Table 4 is the Sinhala consonants syllabics laid in order of organs and place of articulation.

[17] The term *Eḷu* is given to the pure dialect of Sinhala unmixed with foreign words, and *Siṅhala t*o the mixed dialect, though in point of signification the two terms have not the least difference. *Sihaḷa* in Pali, *Siṅhala* in Sanskrit and *heḷa in Eḷu.*

The Sinhala language is a syllabic alphabet in which all consonants have an inherent vowel /a/. This alphabet differs from all other Indo-Aryan languages, and it contains distinctive sounds that are unique to itself since the 8th century A.D.

The presence of a set of five nasal sounds known as "half nasal" or "prenasalized stops" in Sinhala writing is unique too (ඟ *nnga*, ඦ *nja*, ඬ *nndda*, ඳ *nda* and ඹ *mba*).

Table 4: Sinhala consonant syllabics in order of organs and place of articulation

	Gutturals		Palatals		Cerebrals		Dentals		Labials	
Voiceless non-Aspirate	ක	*ka*	ච	*ca*	ට	*tta*	ත	*ta*	ප	*pa*
Voiceless Aspirates	ඛ	*kha*	ඡ	*cha*	ඨ	*ttha*	ථ	*tha*	ඵ	*pha*
Voiced non-Aspirates	ග	*ga*	ජ	*ja*	ඩ	*dda*	ද	*da*	බ	*ba*
Voiced Aspirates	ඝ	*gha*	ඣ	*jha*	ඪ	*ddha*	ධ	*dha*	භ	*bha*
Pure-Nasals	ඞ	*nga*	ඤ, ඥ	*nya,* *nyja*	ණ	*nna*	න	*na*	ම	*ma*
Half-Nasals	ඟ	*nnga*	ඦ	*nja*	ඬ	*nndda*	ඳ	*nda*	ඹ	*mba*
Semi-Vowels			ය	*ya*			ව	*va*		
Trills					ර	*ra*				
Spirants			ශ	*sha*	ෂ	*ssa*	ස	*sa*	හ, ෆ	*ha,* *fa*
Laterals					ළ	*lla*	ල	*la*		

Modern Sinhala writing system has the alphabet of sixty-one (61) characters (see Figure 30). Out of 61 characters, eighteen of them are vowels (18), forty-one of them are consonants (41), and two of them are diacritical marks (02). Two vowels, *ḷ* (ඏ) and *ḹ* (ඐ) not used in present Sinhala writing. In addition to the 61 characters, the Sinhala language has 17 vowel sings and a *virama*. Besides, the Sinhala language uses a unique punctuation mark and 03 medial sings in the language. The pair of unique vowel letters (ඇ *æ* and ඈ *ǣ*) to represent two vowel sounds, and these two appears since 8th century A.D. *(Viyaulpota* Pillar Inscription, see Figure 10). These two vowel letters not seen in any other languages in the world (Disanayaka, 1991).

Vowel signs (known as vowel modifier) are unique, and they used in conjunction with consonants. The vowel signs also used in writing some vowels (e.g. ආ ඕ ඖ). The vowel signs of the Sinhala occur in two different forms; i.e. vowel signs and medial signs, as shown in Figure 30. Like other Indic languages, these vowel signs positioned on any of the four sides of the base consonant. These vowel signs classified as follows.

<Vowel Signs>::= <Left Vowel Signs><RightVowel Signs><Upper Vowel Signs><Lower Vowel Signs>

<Left Vowel Signs>:= ෙ◌ ෛ◌

<Right Vowel Signs>::= ◌ා ◌ැ ◌ෑ ◌ෘ ◌ෲ ◌ෟ ◌ෳ (and Medial Sign ◌්‍ය - *yansaya*)

<Upper Vowel Signs>::= ◌් ◌ි ◌ී (and Medial Sign ◌ - *repaya*)

<Lower Vowel Signs>::=◌ු ◌ූ (and Medial Sign ◌්‍ර - *rakaransaya*)

The vowel, their associated names, and the vowels represented by them gave in Table 5 below.

Table 5: Vowel Signs, Name of the Vowl Signs and Vowel Representations

Serial Number	Vowel Signs	Name of the Vowel Signs	Vowel Representation
1	◌්	Sinhala al-lakuna[1]	
2	◌ා	Sinhala vowel sign aela-pilla	ආ
3	◌ැ	Sinhala vowel sign ketti aeda-pilla	ඇ
4	◌ෑ	Sinhala vowel sign diga aeda-pilla	ඈ
5	◌ි	Sinhala vowel sign ketti is-pilla	ඉ
6	◌ී	Sinhala vowel sign diga is-pilla	ඊ
7	◌ු	Sinhala vowel sign ketti pa-pilla	උ
8	◌ූ	Sinhala vowel sign diga pa-pilla	ඌ
9	◌ෘ	Sinhala vowel sign ketti gaetta-pilla	ඍ
10	◌ෲ	Sinhala vowel sign diga gaetta-pilla	ඎ
11	ෙ◌	Sinhala vowel sign kommbuva	එ
12	ෙ◌්	Sinhala vowel sign diga kommbuva	ඒ
13	ෛ◌	Sinhala vowel sign kommbo deka	ඓ
14	ෙ◌ා	Sinhala vowel sign kommbuva haa aela-pilla	ඔ
15	ෙ◌ා්	Sinhala vowel sign kommbuva haa diga aela-pilla	ඕ
16	◌ෟ	Sinhala vowel sign ketti gayanukitta	ඖ
17	◌ෳ	Sinhala vowel sign diga gayanukitta	ඐ
	Medial Signs		
18	◌්‍ය	Sinhala yansaya	ක්‍ය
19	◌	Sinhala repaya	ර්‍ක
20	◌්‍ර	Sinhala rakaransaya	ක්‍ර

[1] The al-launa removes the implicit vowel associated with a consonant, forming a pure consonant.

Independent Vowel Syllable (V)	18	අ (a) ආ (aa) ඇ (ae) ඈ (aae) ඉ (i) ඊ (ii) උ (u) ඌ (uu) ඍ (r) ඎ (rr) ඏ (l) ඐ (ll) එ (e) ඒ (ee) ඓ (ai) ඔ (o) ඕ (oo) ඖ (au)
Inherent Vowel Sign (Ispilla) (X)	1	◌් (virama)
Vowel Signs ($\underline{V}$)	17	◌ා (aa) ◌ැ (ae) ◌ෑ (aae) ◌ි (i) ◌ී (ii) ◌ු (u) ◌ූ (uu) ◌ෘ (r) ◌ෲ (rr) ◌ෟ (au) ◌ෳ (ll) ◌ෙ (e) ◌ේ (ee) ◌ෛ (ai)◌ො (o) ◌ෝ (oo) ◌ෞ (au)
Consonants (C)	41	ක (ka) ඛ (kha) ග (ga)ඝ (gha) ඞ (nga) ඟ (nnga) ච (ca) ඡ (cha) ජ (ja) ඣ (jha) ඤ (nya) ඥ (jnya) ඦ (nyja) ට (tta) ඨ (ttha)ඩ (dda)ඪ (ddha) ණ (nna) ඬ (nndda) ත (ta) ථ (tha) ද (da) ධ (dha) න (na) ඳ (nda) ප (pa) ඵ (pha)බ (ba) භ (bha) ම (ma) ඹ (mba) ය (ya) ර (ra) ල (la) ව (va) ශ (sha)ෂ (ssa) ස (sa) හ (ha) ළ (lla) ෆ (fa)
Diacritical Marks (D)	2	◌ං (anus) ◌ඃ (visar) (anuswara) (visarga)
Punctuation Mark (P)	1	෴)kundaliya(
Medial Signs ($\underline{C}_{ryr}$)	3	◌්‍ර (rak) ◌්‍ය (yan) ◌ (rep) (rakaransaya) (yansaya) (repaya)

Figure 30: Sinhala character set (Three Medial Signs not included in the UCS/Unicode character set since they are not part of the alphabet.)

As illustrated in Figure 31, Sinhala differ not only in their forms and structure, but also in their uses and functions, i.e., in Sanskrit, Pali, classical writing, general writing, and in regional and contemporary speaking. Some writers who are committed to preserving the purity of the classical idioms, they used different styles, different spellings and the original rules of word-formation like *vid-yā-la-ya-ya* (විද්‍යාලයය; College), however, modernists written with only single '*ya'*.

English Meaning	Old Sinhala (*Eḷu* Alphabet)	Modern Sinhala	Classical Sinhala (Mixed Alphabet)
Household	ගහපති *(ga-ha-pa-ti)*	ගෙයපති *(ge-ya-pa-ti)*	ගෘහපති *(gr-ha-pa-ti)*
Crypt	ලෙණ *(le-na)*	ලෙණ *(le-na)*	ලයන *(la-ya-na)*
Road	මග *(ma-ga)*	මග *(ma-ga)*	මාග *(mār-ga)*
Reverent	තෙර *(te-ra)*	තෙර *(te-ra)*	ස්ථවීර *(stha-vī-ra)*
Teacher	අවරිය *(ava-ri-ya)*	ගුරු *(gu-ru)*	ආචාර්ය *(ā-chār-ya)*
Wife	බරීය *(ba-ri-ya)*	බිරිඳ *(bi-ri-ňda)*	භාර්යා *(bhār-ya-yā)*
Brigadier	සෙනපිති *(se-na-pi-ti)*	සේනාපති *(sē-nā-pa-ti)*	සෙනාපති *(se-nā-pa-ti)*
First	පරමුක *(pa-ra-mu-ka)*	පළමු *(pa-la-mu)*	ප්‍රමුඛ *(pra-mu-kha)*
Minister	අමති *(a-ma-ti)*	ඇමති *(æ-ma-ti)*	අමාත්‍ය *(a-māt-ya)*
Buddhist Monk	සගෙය *(sa-ga-ye)*	සග *(sa-ga)*	සංඝාය *(sań-gha-yā)*

Figure 31: Use of Sinhala in old, modern and classical society

Vowel signs or medial signs can appear before, below, above or after the consonant they belong to, and are used to change the inherent vowel (see Figure 24, 25, 26). When vocalic strokes appear at the beginning of a word, vowels written as independent letters. When the vocalic or non-vocalic strokes appear on above or below of the characters, in most cases the original shape of the letter will be changed. When certain consonants, i.e. නන්ද, occur together, special conjunct symbols are used to combine the essential parts of each letter (i.e. නන්‍ද).

3.5.2 Basic Order of the Character Set

3.5.2.1 Consonants Order

Consonants (C) are the sounds that are produced by the organs of speech, with obstruction in the mouth, of the stream of air that flows from the lungs (see Table 4). There are as such 41 different sounds, which produced 41 consonants in Sinhala. There is syllabic consonant in which all consonants have an inherent vowel /a/. Sinhala consonant set has a unique set of five nasal sounds known as “half nasal” or “prenasalized stops” in modern writing (ඟ *nnga,* ඦ *jnya*, ඬ *nndda*, ඳ *nda,* and ඹ *mba*) and these five consonants have no equivalent in any Indic languages. The phonetic architecture of the Sinhala shown in Table 4. This phonetic architecture is similar to the other languages in the region, such as Devanagari, Tamil and Myanmar.

In Sinhala, two types of consonants occur, consonants and semi-consonants. Consonants do not carry any vowel sound and have two kinds of occurrences, one with its inherent vowel /a/ (i.e. ‘ක’) called open consonants and the other is formed, in general, by adding a stroke /◌්/ to the open consonant letter (i.e. ‘ක්’) is called close consonant or pure-consonant. An open consonant letter is one, which represents a consonants sound that occurs before a vowel sound, i.e. 'ක', in an open syllable. A closed consonant letter, on the other hand, is one which represents a consonant's sound that occurs at the end of a syllable or before another consonant's sound, (i.e. 'කා'), in a closed syllable.

Though the Sinhala language having the complex phonetic structure, the alphabetical order of the consonants (C1 – C41) are well defined as $C_{1\text{-}41}$ = {ක, ඛ, ග, ඝ, ඞ, ඟ, ච, ඡ, ජ, ඣ, ඤ, ඥ, ඦ, ට, ඨ, ඩ, ඪ, ණ, ඬ, ත, ථ, ද, ධ, න, ඳ, ප, ඵ, බ, භ, ම, ඹ, ය, ර, ල, ව, ශ, ෂ, ස, හ, ළ, ෆ}.

However, the consonants 'ඥ' has two phonetic values: at the beginning of a word it has the same phonetic value as 'ඤ', but the order has been arranged as 'ඤ ≤ ඥ', and elsewhere it has phonetic values of 'ජ්' and 'ඤ'.

3.5.2.2 Vowel Characters and Vowel Signs

Sinhala writing uses 18 vowel syllables and its corresponding 17 vowel strokes or vowel signs, as shown in raw number 3 of Figure 30. In representing an independent vowel syllable, vowel character form used. Vowel strokes form used when it is used to change the inherent vowel sound of consonant syllabics. The vowel syllables can occur only at the beginning of the word. Vowel strokes, on the other hand, are one with representing a vowel sound, which occurs after a consonant's sound in words. Each vowel syllable has its form of graphic representation. This set has a pair of unique vowel syllables ඇ (*æ*) and ඈ (*ǣ*) to represent two sounds, and that has been using since the 7th century. Vowel signs 'ඏ' (*ḷ*) and 'ඐ' (*ḹ*) do not occur in modern usage and are not found in any words in many dictionaries, but allow the vowel sign to appear itself.

For this reason, they also included in UCS/Unicode standards for completeness of the vowel set. The vowel 'ඍaa' (*ṝ*) also does not occur in modern usage but its corresponding vocalic stroke, '◌aa' is used; for example, කර්තaa. The usage of two forms of vowels and vowel strokes are standard features of Indian origin writing systems.

The order of the vowels was arranged by tradition and contemporary style. It can define as $V_{1\text{-}18}$ = {අ, ආ, ඇ, ඈ, ඉ, ඊ, උ, ඌ, ඍa, ඍaa, ඏ, ඐ, එ, ඒ, ඓ, ඔ, ඕ, ඖ}, and when a binary relation ≤ is defined on V as a Sinhala vowel syllable order, such as අ ≤ ආ ≤ ඇ ≤ ඈ ≤ ඉ ≤ ඊ ≤ ..... ≤ ඕ ≤ ඖ.

In the case of vowel strokes (V), they are graphical signs always use in conjunction with a consonant. Vowel signs are composed of one, two, three or four-stroke in writing. The representation of a syllable such as 'කෝ' is a form a consonant 'ක' followed by several vowel signs 'ෙ◌', '◌ා' and '◌්', will be keying in its graphical order from left to right and some other cases from bottom to up too. In UCS/Unicode Standard (Unicode, 1998) a combined form of five vowel signs with a single code position for each has been defined. In such cases, the use of single composite vowel sign code following a consonant is encouraged in SLS 1134:1996 Standard. In Sinhala, those strokes are referred to as '*pili*' whereas in technical literature these strokes are called "consonant modifiers". The consonant modifiers of the Sinhala scripts will modify the inherent behavior of the open-consonant such as ක /ka/. Some of the behaviors of such consonants are listed in raw numbers 2 and 3 in Figure 30, for example. When we consider the vowel stroke set $\underline{V}$ = {◌ා, ◌ැ, ◌ෑ, ◌ි, ◌ී, ◌ු, ◌ූ, ◌ෘ, ◌ෲ, ◌ෟ, ◌ෳ, ෙ◌, ෙ◌්, ෙෙ◌, ෙ◌ා, ෙ◌ා්, ෙ◌ෟ} together with '◌්' (X = {◌්}) in this order is acceptable by Standard Institute (SLS 1134:1996, 1998).

In the case of *Ispilla* in raw number 2 in Figure 30 ('◌්'; is also called *virama*) used to remove the inherent sound /a/ from the consonant and it has no corresponding vowel syllable. Thence, it will be treated as a special vowel stroke within the Sinhala language and defined as X = {◌්}.

3.5.2.3 Semi-consonants

Sinhala alphabet has two semi-consonants, known as diacritical marts and they are used only in conjunction with vowels or consonants. The diacritical marts appear only flowed by a vowel or a consonant with an implicit or explicit vowel. Therefore, they consider separate consonant modifiers in the lexicographical order as D = {Ø, ◌ං, ◌ඃ}. The corresponding phonetic notations of semi-consonants are ◌ං (*anuswara*) and ◌ඃ (*visarga*). These two characters have placed at the beginning of the code set, considering the proper lexicographical order.

The consonant syllable 'ඞ' (*ṅa*) and semi-consonants ◌ං (ṅ) have the same phonetic behavior, but their positioning has no relation in lexicographical order and as such, they behave independently.

Taking into consideration of the $\underline{V}$ (vowel strokes) and D (semi-consonants), the target order of the encoring pattern with the relations of D and $\underline{V}$, therefore, can be described as $(D, \leq_d) \times (\underline{V} \leq_{\underline{v}})$.

3.5.2.4 Non-vowel Strokes

The non-vowel strokes ◌්‍ය (*yansaya*) and ◌්‍ර (*rakaransaya*) represent the letter ය (*ya*) and ර (*ra*) respectively, following a pure consonant. The *yansaya* (◌්‍ය) symbolizes ය when preceded by a consonant, i.e. ක් + ය = ක්‍ය, and *rakaransaya* (◌්‍ර) symbolizes ර when by a consonant, i.e. ක් + ර = ක්‍ර. They may appear alone with consonants, or consonants together with vowel strokes. If so, the vowel applies to the ය (*ya*) or ර (*ra*) and not for the initial consonant. The UCS/Unicode standard does not provide code positions for the ◌්‍ය (*yansaya*), ◌ (*repaya)* and ◌්‍ර (*rakaransaya*), are represented by code sequences with zero-width joiner (ZWJ), which is used to form conjunct letters, for example 'ක + ◌් + ZWJ + ය = ක්‍ය' and 'ක + ◌් + ZWJ + ර = ක්‍ර'. Since this is a set of unique conjunct syllables and the relation ≤ is defined for both on Cyr as a standard ascending order for ≤‿, ($C_y \leq C_r$). Lexicographical order of the syllable 'ක' with '◌්‍ය' and '◌්‍ර' with vowel strokes are given in Figure 25.

The non-vowel stroke Sinhala *repaya* (◌) symbolizes a ර් when by preceding a consonant, e.g. ක + ර් + ම = කර්‍ම. The use of this is not mandatory in Sinhala and will not be considered as a case in lexicographical order since the code will be converted simply to a ර්.

Finally, the formal description of the order of the Sinhala character set can be defined as follows:

D = {ø , ◌ං, ◌ඃ}

V = {අ, ආ, ඇ, ඈ, ඉ, ඊ, උ, ඌ, සa, සaa, ඏ, ඐ, එ, ඒ, ඓ, ඔ, ඕ, ඖ}

C = {ක, ඛ, ග, ඝ, ඞ, ඟ, ච, ඡ, ජ, ඣ, ඤ, ඥ, ඦ, ට, ඨ, ඩ, ඪ, ණ, ඬ, ත, ථ, ද, ධ, න,
ඳ, ප, ඵ, බ, භ, ම, ඹ, ය, ර, ල, ව, ශ, ෂ, ස, හ, ළ, ෆ}

$\underline{V}$ = {◌ා, ◌ැ, ◌ෑ, ◌ි, ◌ී, ◌ු, ◌ූ, ◌a, ◌aa, ◌ෙ, ◌ේ, ◌ෛ, ◌ො, ◌ෝ, ◌ෞ}

Cyrr = {◌්‍ය , ◌ර්‍, ◌්‍ර }

X = {◌්}.

Figure 32 is giving the State Transition Chart for Sinhala Syllable Structure, and Figure 33 lists the Order Check Table to reflect the alphabetical order of the Sinhala character set.

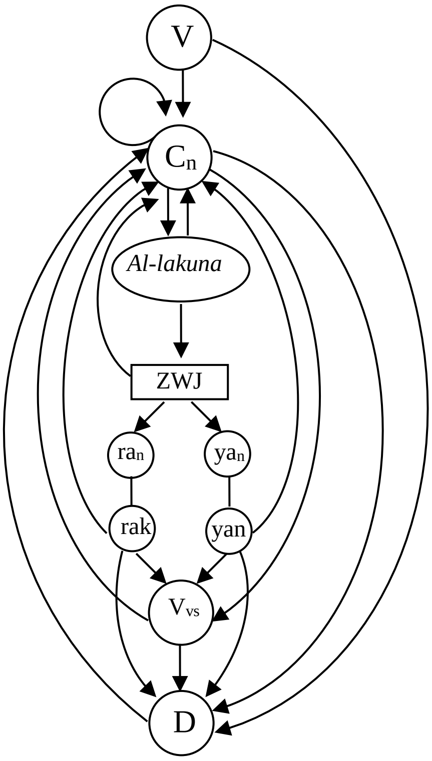


Figure 32: State Transition Chart for Sinhala Syllable Structure

V → C ② V → D ①	D < V
C → C ② C → X ④ C → $\underline{V}$ ③ C → D ①	D < C < $\underline{V}$ < X
X → ZWJ D ① X → C ②	ZWJ D < C
ZWJ→Cry ①	
$\underline{V}$ → C ② D ①	D < C
D → C ②	D < C

Figure 33: Order Check Table

3.6 Syllable Structure

3.6.1 Open Syllable

Disanayaka says in his book (Disanayaka, 1991), syllables in spoken Sinhala are two types; Open Syllables and Closed Syllables. Open Syllables are composed of a vowel (V) or which ends in a vowel (CV). In V structure vowel can occur only at the beginning of a word and they can be short as *i* in *i-ra* (ඉර; sun) or long as *ī* in *ī-ye* (ඊයෙ; yesterday) and single as *a* in *a-da* (අද; today) or clusters (VV) as *iu* in *iu-va* (ඉවුව; cooked).

CV structure occurs initially, medially and finally in words. All consonants occur in such syllables except the half-nasals and nasals (i.e., ago).

3.6.2 Closed Syllable

A syllable ends with a consonant is called Close Syllable. It preceded by a vowel only (VC) or a consonant and a vowel (CVC).

Syllables in VC structure occur only at the beginning of the word. The vowel is usually short and single. Half-nasals (ඟ, ඦ, ඬ, ඳ, ඹ) and '◌ං' (ṅ) are not occurring in such syllables.

In the CVC structure, close syllables occur beginning, middle and the end in words. The middle vowel occurs only in loan words and few proper nouns. All consonants occur as the final consonant, but ‘ඞ’ (*ṅa*) does not occur as the initial consonant.

3.6.3 Syllabic Writing System

The functional unit of these writing systems is the orthographic syllable, consisting of two types; open syllables and closed syllables as indicated above. Figure 34 illustrates some example of open and closed syllable structures and Figure 35 shows Sinhala syllabic writing system in Sinhala. The syllable is built up of alphabetic pieces, the actual letters of the Sinhala script. These pieces consist of four distinct character types: consonants (C), vowels (V), vowel signs ($\underline{V}$) and Diacritical Marks (D) as shown in Figure 30. The sequence of text is stored in visual order of the keying.

V	ඌ	uu	Him
VV	ඉවුවා	i-u-va	Cooked
VC	ඉර	i-ra	Sun
VC	ඊයෙ	ii-ye	Yesterday
VC	අද	a-da	Today
CV	කටකතා	ka-ta-ka-taa	Rumor
CV	පිරිසිදු	pi-ri-si-du	Clean
CV	දහය	da-ha-ya	Ten
CVV	රැවුල	ræ-u-la	Beard

Figure 34: Example of an open and closed syllable structure

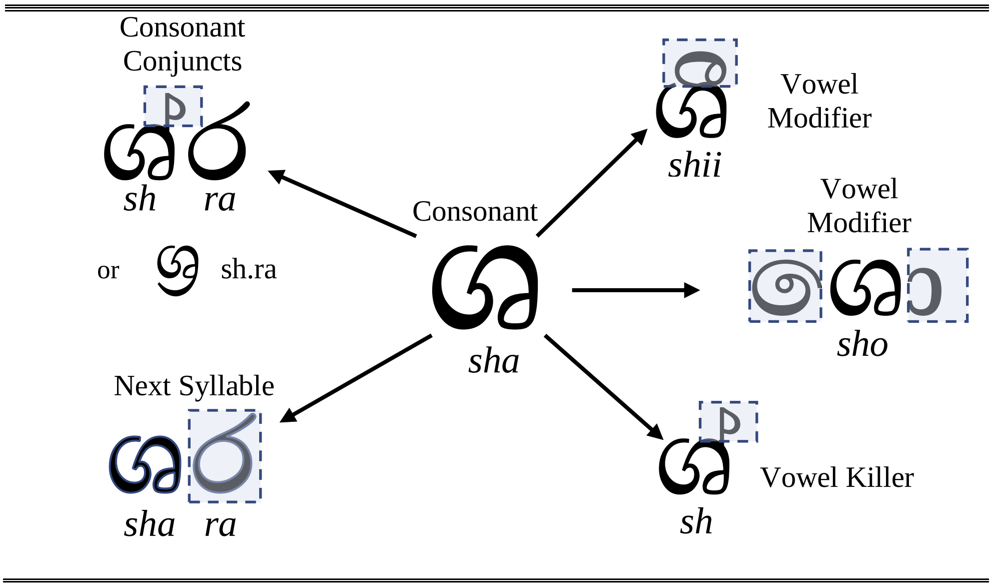


Figure 35: Syllabic writing system in Sinhala

According to the above syllable structure, the Sinhala character set approved for standardization (see Table 6) can be grouped into 06 parts: vowels (V), inherent vowel sign (X), vowel signs ($\underline{V}$), consonants (C), diacritical marks (D) and medial signs (C_{ryr}). A Sinhala word is typically formed by a combination of one or more consonants, initial vowel and one or more consonants, one or more consonants with a combination of vowel signs, or diacritical marks and special symbols to make a syllable.

However, not all productions of the above definition are valid composite characters, though all valid composite characters follow the above definition. Moreover, the total number of valid and usable composite glyphs is roughly 2300. Table 7 shows some of the composite glyphs, which combined with consonant and vowel sings only — however, the combination with diacritical marks and medial signs not formed in this table.

Table 6: Sinhala Character Set (CANLIT & NARESA, 1985) (Dotted circle denotes a consonant)

Vowel (V)	16	අ ආ ඇ ඈ ඉ ඊ උ ඌ සa සaa එ ඒ ඓ ඔ ඕ ඖ
Inherent Vowel Sign (X)	1	◌්
Vowel Signs ($\underline{V}$)	12	◌ා ◌ැ ◌ෑ ◌ි ◌ී ◌ු ◌ූ ◌a ◌aa ෙ◌ ෛ◌ ◌ෟ
Consonants (C)	41	ක ඛ ග ඝ ඞ ඟ ච ඡ ජ ඣ ඤ ඦ ඥ ට ඨ ඩ ඪ ණ ඬ ත ථ ද ධ න ඳ ප ඵ බ භ ම ඹ ය ර ල ව ශ ෂ ස හ ළ ෆ
Diacritical marks (D)	2	◌ං ◌ඃ
Medial Signs ($\underline{C}_{ryr}$)	3	◌්‍ර *rakaransaya* ◌්‍ය *yansaya* ර්‍◌ *repaya*

Table 7: Sinhala vowels, consonants and composite glyphs

	◌ං	◌ඃ	අ	ආ	ඇ	ඈ	ඉ	ඊ	උ	ඌ	සa	සaa	එ	ඒ	ඓ	ඔ	ඕ	ඖ
ක	කං	කඃ	ක	කා	කැ	කෑ	කි	කී	කු	කූ	කa	කaa	කෙ	කේ	කෛ	කො	කෝ	කෞ
ඛ	ඛං	ඛඃ	ඛ	ඛා	ඛැ	ඛෑ	ඛි	ඛී	ඛු	ඛූ	ඛa	ඛaa	ඛෙ	ඛේ	ඛෛ	ඛො	ඛෝ	ඛෞ
ග	ගං	ගඃ	ග	ගා	ගැ	ගෑ	ගි	ගී	ගු	ගූ	ගa	ගaa	ගෙ	ගේ	ගෛ	ගො	ගෝ	ගෞ
ඝ	ඝං	ඝඃ	ඝ	ඝා	ඝැ	ඝෑ	ඝි	ඝී	ඝු	ඝූ	ඝa	ඝaa	ඝෙ	ඝේ	ඝෛ	ඝො	ඝෝ	ඝෞ
ඞ	ඞං	ඞඃ	ඞ	ඞා	ඞැ	ඞෑ	ඞි	ඞී	ඞු	ඞූ	ඞa	ඞaa	ෙ◌	ෙ◌	ෛ◌	ෙ◌	ෙ◌	ෙ◌
ඟ	ඟං	ඟඃ	ඟ	ඟා	ඟැ	ඟෑ	ඟි	ඟී	ඟු	ඟූ	ඟa	ඟaa	ඟෙ	ඟේ	ඟෛ	ඟො	ඟෝ	ඟෞ
ච	චං	චඃ	ච	චා	චැ	චෑ	චි	චී	චු	චූ	චa	චaa	චෙ	චේ	චෛ	චො	චෝ	චෞ
ඡ	ඡං	ඡඃ	ඡ	ඡා	ඡැ	ඡෑ	ඡි	ඡී	ඡු	ඡූ	ඡa	ඡaa	ඡෙ	ඡේ	ඡෛ	ඡො	ඡෝ	ඡෞ
ජ	ජං	ජඃ	ජ	ජා	ජැ	ජෑ	ජි	ජී	ජු	ජූ	ජa	ජaa	ජෙ	ජේ	ජෛ	ජො	ජෝ	ජෞ
ඣ	ඣං	ඣඃ	ඣ	ඣා	ඣැ	ඣෑ	ඣි	ඣී	ඣු	ඣූ	ඣa	ඣaa	ඣෙ	ඣේ	ඣෛ	ඣො	ඣෝ	ඣෞ
ඤ‹	ඤං	ඤඃ	ඤ	ඤා	ඤැ	ඤෑ	ඤි	ඤී	ඤු	ඤූ	ඤa	ඤaa	ඤෙ	ඤේ	ඤෛ	ඤො	ඤෝ	ඤෞ
ඥ	ඥං	ඥඃ	ඥ	ඥා	ඥැ	ඥෑ	ඥි	ඥී	ඥු	ඥූ	ඥa	ඥaa	ඥෙ	ඥේ	ඥෛ	ඥො	ඥෝ	ඥෞ
ඦ	ඦං	ඦඃ	ඦ	ඦා	ඦැ	ඦෑ	ඦි	ඦී	ඦු	ඦූ	ඦa	ඦaa	ඦෙ	ඦේ	ඦෛ	ඦො	ඦෝ	ඦෞ
ට	ටං	ටඃ	ට	ටා	ටැ	ටෑ	ටි	ටී	ටු	ටූ	ටa	ටaa	ටෙ	ටේ	ටෛ	ටො	ටෝ	ටෞ
ඨ	ඨං	ඨඃ	ඨ	ඨා	ඨැ	ඨෑ	ඨි	ඨී	ඨු	ඨූ	ඨa	ඨaa	ඨෙ	ඨේ	ඨෛ	ඨො	ඨෝ	ඨෞ

ඩ	ඩං	ඩඃ	ඩ	ඩා	ඩැ	ඩෑ	ඩි	ඩී	ඩු	ඩූ	ඩa	ඩaa	ඩෙ	ඩේ	ඩෛ	ඩො	ඩෝ	ඩෞ
ඪ	ඪං	ඪඃ	ඪ	ඪා	ඪැ	ඪෑ	ඪි	ඪී	ඪු	ඪූ	ඪa	ඪaa	ඪෙ	ඪේ	ඪෛ	ඪො	ඪෝ	ඪෞ
ණ	ණං	ණඃ	ණ	ණා	ණැ	ණෑ	ණි	ණී	ණු	ණූ	ණa	ණaa	ණෙ	ණේ	ණෛ	ණො	ණෝ	ණෞ
ඬ	ඬං	ඬඃ	ඬ	ඬා	ඬැ	ඬෑ	ඬි	ඬී	ඬු	ඬූ	ඬa	ඬaa	ඬෙ	ඬේ	ඬෛ	ඬො	ඬෝ	ඬෞ
ත	තං	තඃ	ත	තා	තැ	තෑ	ති	තී	තු	තූ	තa	තaa	තෙ	තේ	තෛ	තො	තෝ	තෞ
ථ	ථං	ථඃ	ථ	ථා	ථැ	ථෑ	ථි	ථී	ථු	ථූ	ථa	ථaa	ථෙ	ථේ	ථෛ	ථො	ථෝ	ථෞ
ද	දං	දඃ	ද	දා	දැ	දෑ	දි	දී	දු	දූ	දa	දaa	දෙ	දේ	දෛ	දො	දෝ	දෞ
ධ	ධං	ධඃ	ධ	ධා	ධැ	ධෑ	ධි	ධී	ධු	ධූ	ධa	ධaa	ධෙ	ධේ	ධෛ	ධො	ධෝ	ධෞ
න	නං	නඃ	න	නා	නැ	නෑ	නි	නී	නු	නූ	නa	නaa	නෙ	නේ	නෛ	නො	නෝ	නෞ
ඳ	ඳං	ඳඃ	ඳ	ඳා	ඳැ	ඳෑ	ඳි	ඳී	ඳු	ඳූ	ඳa	ඳaa	ඳෙ	ඳේ	ඳෛ	ඳො	ඳෝ	ඳෞ
ප	පං	පඃ	ප	පා	පැ	පෑ	පි	පී	පු	පූ	පa	පaa	පෙ	පේ	පෛ	පො	පෝ	පෞ
ඵ	ඵං	ඵඃ	ඵ	ඵා	ඵැ	ඵෑ	ඵි	ඵී	ඵු	ඵූ	ඵa	ඵaa	ඵෙ	ඵේ	ඵෛ	ඵො	ඵෝ	ඵෞ
බ	බං	බඃ	බ	බා	බැ	බෑ	බි	බී	බු	බූ	බa	බaa	බෙ	බේ	බෛ	බො	බෝ	බෞ
භ	භං	භඃ	භ	භා	භැ	භෑ	භි	භී	භු	භූ	භa	භaa	භෙ	භේ	භෛ	භො	භෝ	භෞ
ම	මං	මඃ	ම	මා	මැ	මෑ	මි	මී	මු	මූ	මa	මaa	මෙ	මේ	මෛ	මො	මෝ	මෞ
ඹ	ඹං	ඹඃ	ඹ	ඹා	ඹැ	ඹෑ	ඹි	ඹී	ඹු	ඹූ	ඹa	ඹaa	ඹෙ	ඹේ	ඹෛ	ඹො	ඹෝ	ඹෞ
ය	යං	යඃ	ය	යා	යැ	යෑ	යි	යී	යු	යූ	යa	යaa	යෙ	යේ	යෛ	යො	යෝ	යෞ
ර	රං	රඃ	ර	රා	රැ	රෑ	රි	රී	රු	රූ	රa	රaa	රෙ	රේ	රෛ	රො	රෝ	රෞ
ල	ලං	ලඃ	ල	ලා	ලැ	ලෑ	ලි	ලී	ලු	ලූ	ලa	ලaa	ලෙ	ලේ	ලෛ	ලො	ලෝ	ලෞ
ව	වං	වඃ	ව	වා	වැ	වෑ	වි	වී	වු	වූ	වa	වaa	වෙ	වේ	වෛ	වො	වෝ	වෞ
ශ	ශං	ශඃ	ශ	ශා	ශැ	ශෑ	ශි	ශී	ශු	ශූ	ශa	ශaa	ශෙ	ශේ	ශෛ	ශො	ශෝ	ශෞ
ෂ	ෂං	ෂඃ	ෂ	ෂා	ෂැ	ෂෑ	ෂි	ෂී	ෂු	ෂූ	ෂa	ෂaa	ෂෙ	ෂේ	ෂෛ	ෂො	ෂෝ	ෂෞ
ස	සං	සඃ	ස	සා	සැ	සෑ	සි	සී	සු	සූ	සa	සaa	සෙ	සේ	සෛ	සො	සෝ	සෞ
හ	හං	හඃ	හ	හා	හැ	හෑ	හි	හී	හු	හූ	හa	හaa	හෙ	හේ	හෛ	හො	හෝ	හෞ
ළ	ළං	ළඃ	ළ	ළා	ළැ	ළෑ	ළි	ළී	ළු	ළූ	ළa	ළaa	ළෙ	ළේ	ළෛ	ළො	ළෝ	ළෞ
ෆ	ෆං	ෆඃ	ෆ	ෆා	ෆැ	ෆෑ	ෆි	ෆී	ෆු	ෆූ	ෆa	ෆaa	ෆෙ	ෆේ	ෆෛ	ෆො	ෆෝ	ෆෞ

3.7 Major Issues in Writing System

The Sinhala alphabet and its complexities were developing intensively since the 5th century A.D.; the following points considered for the use of the Sinhala writing systems.

1. Every vowel except the first one has a corresponding vowel modifier symbol, which can be attached to consonants to make composite characters.
2. When vowels appear at the beginning of a word, vowels written as independent letters, on the other hand, when a pure consonant combined with the first vowel, it leads to the common consonant form, i.e., the consonant sign has an inherent vowel /*a*/ associated with it.
3. There were two commonly used diacritical marks: '*anuswar*' (◌ං *ṅ*) and '*visarga*' (◌ඃ *ḥ*), like most of the Indic languages. They may appear only followed by a vowel or a consonant with an implicit or explicit vowel. The diacritical marks have placed in order at the beginning of the character set.
4. Vowel signs are attached to the left, right, above or below to its fix position or variable positions, ten vowel signs fixed to the right of the consonants, four fixed above the consonants, five vowel signs are fixed below to the consonants, and one sign fixed to the left of the consonants. Vowel signs have composed a form of one, two, three or four-stroke in writing. The representation of a syllable such as

'කෝ' is a form a consonant 'ක' with several vowel signs 'ෙ'. 'ා' and '්, will be keying in its graphical order (visual order) from left to right. (see Figure 38)

5. In some cases, they attached above or below the consonants. When some modifiers are attached, it changes the original shapes of the consonants. Appearances of modifiers also differed according to the consonants. (see Figure 38)
6. Also, two special symbols are corresponding to the sound of 'r' and 'y' called *rakaransaya* and *yansaya*. The symbols ්‍ය (*yansaya*) (C_y) and ්‍ර (*rakaransaya*) (C_r) represent the letter ය (*ya*) and ර (*ra*) respectively, following a pure consonant. The *yansaya* (්‍ය) symbolizes ය when preceded by a consonant, i.e. ක් + ය = ක්‍ය, and *rakaransaya* (්‍ර) symbolizes ර when by a consonant, i.e. ක් + ර = ක්‍ර. They may appear along with consonants, or consonants together with vowel sign.
7. The non-vowel stroke *repaya* (ර්‍) symbolizes an 'r' (ර්) when by preceding a consonant, e.g. ka (ක) + r (ර්) + ma (ම) = *karma* (කර්ම).
8. When Sanskrit and Pali words adopted into Sinhala, they transcribed in the compound manner in which they have written in Sanskrit and Pali. This composition is effected by the union of one or more consonants, or their parts or symbols, with a vowels-consonant or its parts or symbols, and vice versa (i.e., a word like *Buddha* is being written in බුද්ධ, බුද‍්ධ and බුඪ).
9. In Pali and Sanskrit writing with Sinhala characters, some glyph combined further with any one of the consonants with half-consonants or consonants such as න්ද like න්‍ද or 'න්ද්රෝ' like න්‍ද්‍රෝ or 'ස්ත්රෝත්ර' like **ස්ත්‍රෝත්‍ර** or ප්රත්යක්ෂ like ප්‍රත්‍යක්‍ෂ or පච්චුප්පන්නා like පච්චුප්පන්නා. They are pronounced as *nda*, *ndro*, *strōtra*, *pratyaksa*, and *pachchuppanna* respectively.

Having looked at the basic structure and significant issues of Sinhala scripts, in the following section, the author examines the complexity involved in their machine representation.

3.8 Complex Ligatures

In Sinhala language, combinations of consonants, consonant modifiers and semi-consonants produce different phonetic sounds.

Table 8 illustrates how some consonants and vowel strokes are combined to form syllable blocks (glyphs). Some glyphs constructed differently according to the shape of the Sinhala letter (they marked with †). Some glyphs are un-pronounceable (they marked *). Some would create a somewhat uneven, irregular and illogical outer appearance (they marked with ‡). For example consonant *ka* (ක) with vocalic sing *æ* (ැ) represent syllable *kæ* (කැ) whereas consonant *ra* (ර) with vocalic sing *u* (ු) represent the same *ru* (රු) vocalic sing in appearance like *ru* (රු) which is used in syllable representation with other consonants like *ka* (ක). Similarly, consonant *lla* (ළ) with vocalic sing *u* (◌ු) represent the new shape like *llu* (ළු), further consonant *lla* (ළ) with vocalic sing *uu* (◌ූ) represent the new shape like *lluu* (ළූ).

Table 8: Vowel syllables, corresponding vowel strokes and their usage with selected consonants

Sound		*a*	*ā*	*æ*	*ǣ*	*i*	*ī*	*u*	*ū*	*ṛ*	*ṝ*	*ḷ*	*ḹ*	*e*	*ē*	*ai*	*o*	*ō*	*au*
Vowel Syllable(V)		අ	ආ	ඇ	ඈ	ඉ	ඊ	උ	ඌ	සa	සaa	ඏ	ඐ	එ	ඒ	ඓ	ඔ	ඕ	ඖ
Vowel Signs(V)	◌්		◌ා	◌ැ	◌ෑ	◌ි	◌ී	◌ු	◌ූ	◌a	◌aa			ෙ◌	ෙ◌්	ෛ◌	ෙ◌ා	ෙ◌ා්	ෙ◌ෟ
C_1V	ක්	ක	කා	කැ	කෑ	කි	කී	කු	කූ	කa	කaa			කෙ	කේ	කෛ	කො	කෝ	කෞ
Regular Key in order with Consonant ‘ක’ (C_1)	ක ◌්	ක	ක, ◌ා	ක, ◌ැ	ක, ◌ෑ	ක, ◌ි	ක, ◌ී	ක, ◌ු	ක, ◌ූ	ක, ◌a	ක, ◌aa			ෙ◌, ක	ෙ◌, ක, ◌්	ෙ◌, ෙ◌, ක	ෙ◌, ක, ◌ා	ෙ◌, ක, ◌ා, ◌්	ෙ◌, ක, ◌ෟ
Outer appearances with C_{16},C_{33} & C_{40} with V	ඩ්	ඩ	ඩා	ඩැ	ඩෑ	ඩි†	ඩී†	ඩු†	ඩූ†	ඩa	ඩaa			ඩෙ	ඩේ†	ඩෛ	ඩො	ඩෝ	ඩෞ
	ර්	ර	රා	රැ‡	රෑ‡	රි	රී	රු‡	රූ‡	රa*	රaa*			රෙ	රේ	රෛ	රො	රෝ	රෞ
	ළ්	ළ	ළා	ළැ	ළෑ	ළි	ළී	ළු‡	ළූ‡	ළa*	ළaa*			ළෙ	ළේ	ළෛ	ළො	ළෝ	ළෞ

Every syllabic frame constructed in the way according to the shape of the Sinhala letter. In every frame, 41 consonants (C) and 16 vowel strokes (V) combined to form glyphs. After that, each united glyphs in a frame can further be combined with two non-vocalic strokes (C_{yrr}) and then even further it can be combined with two semi-consonants (D), and after all, it will produce more than 2300 “pronounceable” glyphs used for Sinhala writing. For example consonant *ka* (ක) with vowel strokes and non-vocalic strokes will produce following glyphs; ක, ක්, ක, කා, කැ, කෑ, කි, කී, කු, කූ, කa, කaa, කෙ, කේ, කෛ, කො, කෝ, කෞ, ක්‍ර, ක්, ක්‍ර, ක්‍රා, ක්‍රැ, ක්‍රෑ, ක්‍රි, ක්‍රී, ක්‍රෙ, ක්‍රේ, ක්‍රෛ, ක්‍රො, ක්‍රෝ, ක්‍රෞ, ක්‍ය, ක්‍යා, ක්‍යැ, ක්‍යෑ, ක්‍යි, ක්‍යී, ක්‍යු, ක්‍යූ, ක්‍යෙ, ක්‍යේ, ක්‍යෛ, ක්‍යො, ක්‍යෝ & ක්‍යෞ.

3.9 Character Appearance and Positioning in Reading and Writing

Sinhala character positioning is not laid out in a neat linear line from left to right. Thus Sinhala characters can be divided into three main groups. (1) Those are having a regular ‘x’ height such as ක, ග, ය, ස, න. (2) Those are having an ascender, similar to ‘l’ such as ට, ච, ර, ජ, බ, ම, and (3) Those are having a descender, similar to ‘g’ such as අ, ඳ, ද, ඤ, ඥ, උ, ළ (See Figure 36). The positioning can be even more complicated when multiple vocalic strokes are attached to the same character. For example, a consonant with a vocalic stroke will be positioned in three levels for the word like නේත්‍රාලෝක තෛල as given in Figure 37, and Figure 38 shows new shapes and positions.

Level 2 and 3 are the base level, and vowels, consonants, non-vocalic and some of the vocalic strokes are used in these planes. However, some parts of the vowels or consonants appear in levels 2, 3 and 1 or level 2, 3 and 4. Vocalic strokes may attach not just to the top or bottom of the consonants, but also to the left or right to the consonants.

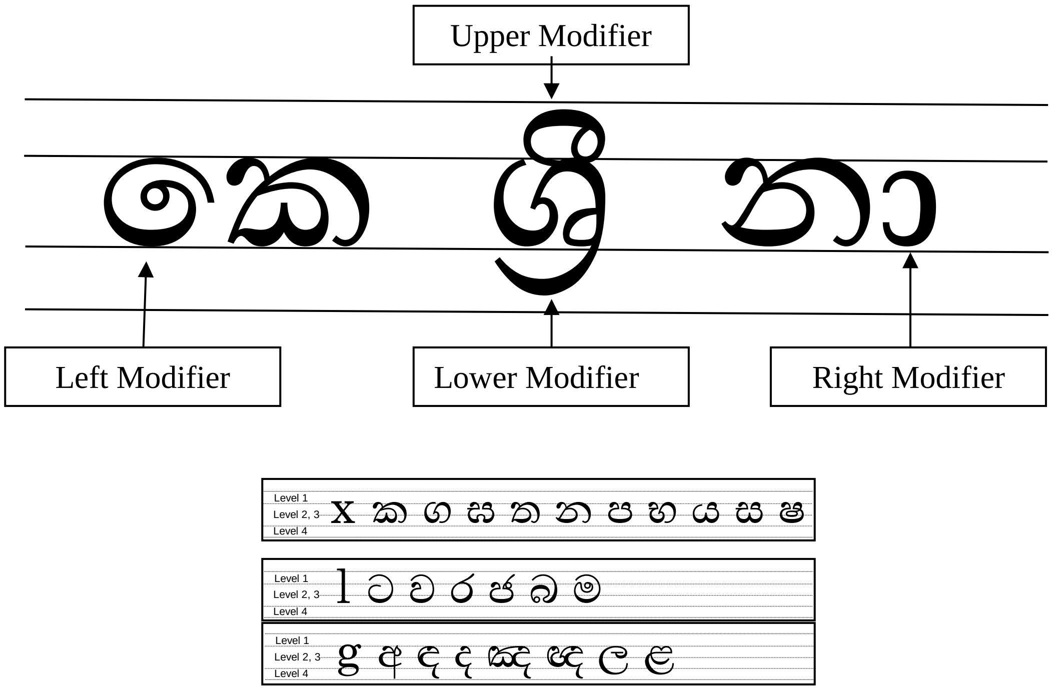


Figure 36: Sinhala Character Positioning

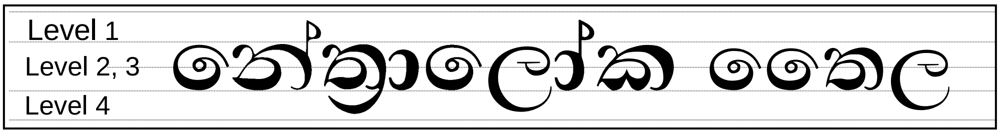


Figure 37: Sinhala Consonant & Vowel Signs Positioning

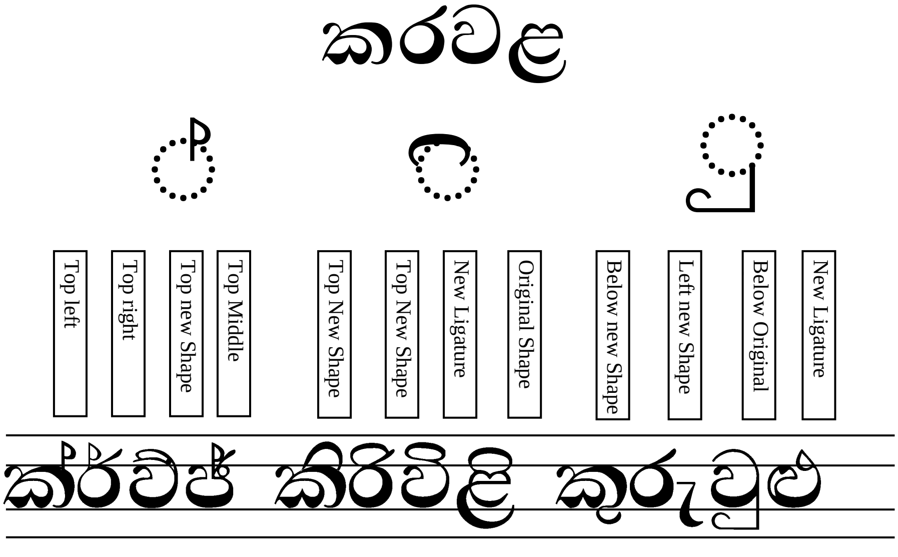


Figure 38: New shapes and new positioning are given when combining vowel signs with consonants

3.10 Characters Shapes and Writing Style

In the Sinhala alphabet, the consonants take different shapes and have to differentiate one consonant from the other by its particular shape. The consonants formed with the combination of several vowel modifiers

When writing a consonant, like ක. ත, and ග curve down at the rear end, and consonants like ට, ඛ and ච curve to the left at the upper end. In letters ඵ, ළු and ථ the ending part bends to the left with a small hook-like thing at the end. The nasalized consonants ඳ and ඹ and the nasal consonant ඤ special kind of a subordinate part joins at the end.

Every consonant classified into three, namely, base consonants, ascending consonants and descending consonants. The writing area can be divided into three phases horizontally. The consonants written in the middle or base phase are base. ග, හ, ස, ප are base consonants. The consonants that spread up from the middle phase to the upper phase are called ascending consonants. They are ට, ච, ඵ, ඩ, ඪ. The consonants that spread down to the lower phase are called descending consonants, like ඇ, උ, ද.

There are two kinds of ascending consonants, namely, consonant without a stroke at the top and the consonants with a stroke at the top. The consonants of the first category are a little shorter than the second category, i.e. ට, ධ, ච, ම්, ඔ, ඕ and ඛ.

The consonant ඣ looks like formed by combining two consonants, one medial letter and an ascending letter. The consonant ළු which represent cerebral ළ and vowel sound උ is also an ascending consonant. The descending letters are written in the middle phase and spreads to the lower phase.

The two consonants such as ඤ and ඥ, like in above, is written with the combination of a base consonant and descending consonant.

The vowel modifiers that are written as a combination with consonants to represent a vowel sound can be divided into three categories. The vowel modifiers are smaller than the letters. The vowel modifiers appeared in base phase can be written in the same height as the base letters දා, දැ, දෑ, දෘ, දෲ, දෞ. The vowel modifiers are written usually after the consonants except for the (ෙ◌) *kombuwa*. It comes before the consonant. The *yansaya* (◌්‍ය) that represents a consonant sound also categorized as a base stroke.

3.11 Conclusion

Two thousand five hundred years of recorded history written in cave inscriptions, ola manuscripts, and silver, copper, and gold plates, shows thirty-eight Brāhmī scripts appeared in these records. This thirty-eight Brāhmī scripts derived from forty symbols occurred in the edicts of Asoka. Earliest writing styles of inscriptions were geometrical in shape and the time passed the geometric straight-line scripts gradually became rounded at the edges and subsequently alphabet and writing patterns have changed almost circular, with notable influenced by the Kadamba and Pallava Grantha writing system of South India. Subsequently, with the influence of western technologies such as font casting, manual composing, typesetting and mechanical and electronic typewriting, development of Sinhala script and its language became a challenge for an adaptation to the roman script-oriented computer technology to handle the characteristics and complexities of the Sinhala scripts. Some of the significant issues in writing system, controversial issues in alphabetical order, complex ligature and their shapes used not only in Sanskrit and Pali, use of Sinhala in old, modern and classical society also examined since the aforesaid is a significant challenge faced by the Sinhala language interface developers and experts involving in standardizing the language character set. The next chapter discusses how computer technologists manage this complexity when introducing Sinhala to early computers with limited recourses.

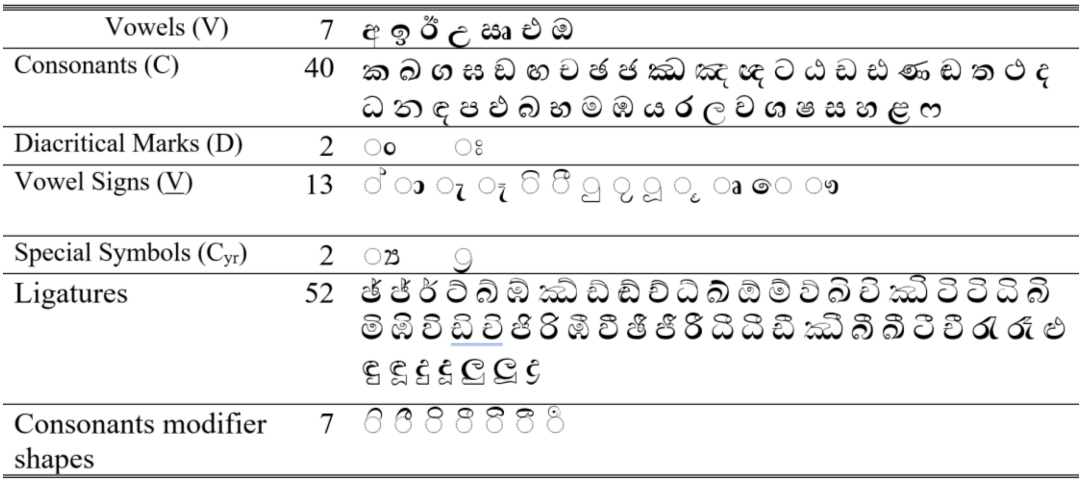
Chapter 4: Digitizing Sinhala - Experience and Benefits From Early Computer Technology

The Sinhala writing system, which is used in Sri Lanka, is a syllabic writing system derived from ancient Brahmi and consists of vowels, consonants, diacritical marks and special symbols. Several of these characters are combined to form complex ligatures. The total number of different glyphs is nearly 2,300. Computer equipment used to represent the Sinhala language needs to facilitate this complexity, in both display and printing, without adding extra complexity to the keyboard or the input systems. In this chapter, the historical background of the Sinhala writing system and Sinhala scripts' characteristics and complexities are discussed; and the evolution of Sinhala computing technology over the last quarter of a century is surveyed. The design of character code standards is a cornerstone of the whole architecture for text processing presented. A case study described in this chapter also shows how small communities of non-Roman script users -interact with the Romanized system dominated cyberspace.

4.1 Milestones in Sinhala Text Processing

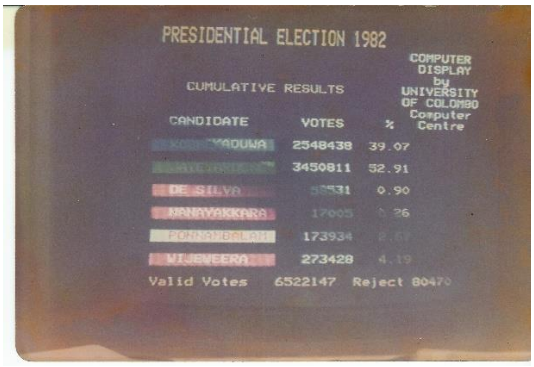
The introduction of BBC (British Broadcasting Corporation) microcomputer systems to the University of Colombo in 1982, kindled the thinking of educational institutes and schools. This microcomputer was notable for its ruggedness, expandability, and the quality of its operating system, which could be programmed using BBC assembly language and BBC Basic interpreter language. The experts of the University of Colombo worked on this small machine and used to disseminate computer technology, especially to schools using mother tongue of Sinhala or Tamil around the country. As a result, the graphics-based set of Sinhala bitmap fonts were developed and used with the machine, and this was the seminal attempt to use computers for the local language. Using Sinhala bitmap font, daily TV program schedule was transmitted for the public by the National TV Station Independent Television Network (ITN) and it was the first attempt in Sri Lanka to use computers with local languages for the public. An early set of bitmap font called "*Sinhala INet"* was to display the TV program schedule during the years 1983 to 1985.

The BBC 6502 microcomputer running Acorn MOS (Arcorn Machine Operating System) with 32 Kbytes of memory machine used for the above activity came with color capabilities. BBC Basic interpreter software to develop Sinhala characters and to link the output with the PAL vision mixer at the TV station; the schedule was broadcast this way for more than three years. In November 1982, the same facility was used to transmit general election results on SLRC (Sri Lanka *Rupavahini* Corporation), which is the leading national television station. This broadcast was the breakthrough that created awareness about computers capable of using languages other than English in Sri Lanka. Figure 39 shows the improved bitmap font set called "*Sinhala INet Font Ver. 0.1*", was created by the author and used to display computer-generated election results transmitted to the public through national television. Sample computer screenshots were shown in Figure 40 below. The graphical model font set has the capability of constructing remaining nine vowels ‘*ā, æ, ǣ, ū, ṝ, ai, ō* and *au’* (ආ ඇ ඈ ඌ ඎ ඒ ඓ ඕ ඖ) using available vowels & vowel signs. The ligature like ‘*k.roo*’ (ක්‍රෝ) can be constructed using constant, vowel signs and non-vocalic strokes and similarly, ligature like ‘*vee*’ (”වේ”) can construct using vowel sings and ligature. The ligatures such as ‘*tii, nii, pi, and pii*’ තී නී පි පී constructed using consonants and consonant modifier shapes. The combined ligatures used in Sanskrit (*n.da* ඬ, *k.sha* ක්‍ෂ *t.va* ත්‍ව *k.va* ක්‍ව) and touching ligature used in pali (*d.va* ද්‍ව *d,dha* ද්‍ධ *b.ba* බ්‍බ) was not implemented and, however, it should note that during this time standards were not defined.



Vowels (V)	7	අ ඉ ඍ උ ඎ එ ඔ
Consonants (C)	40	ක ඛ ග ඝ ඞ ඟ ච ඡ ජ ඣ ඤ ඥ ට ඨ ඩ ඪ ණ ඬ ත ථ ද ධ න ඳ ප ඵ බ භ ම ඹ ය ර ල ව ශ ෂ ස හ ළ ෆ
Diacritical Marks (D)	2	◌ං ◌ඃ
Vowel Signs (V)	13	◌් ◌ා ◌ැ ◌ෑ ◌ි ◌ී ◌ු ◌ූ ◌ු ◌ෘ ◌ෲ ෙ◌ ◌ෟ
Special Symbols (C_{yr})	2	◌ෳ ්‍ර
Ligatures	52	ජි ජී රි ටි බි ඹි ඣි ඬි ඬී චි ධි බී ඹී මි චි බී චි ඣී ටි ටී ධී බී මී ඹී චී ඩි ඩී ජි රී ඹී චී ජී ජී රී ධි ධී ඩී ඣී නි නී ටී චී රු රූ ළු දු දූ ඳු ඳූ ලු ලූ ශු
Consonants modifier shapes	7	◌ි ◌ී ◌ි ◌ී ◌ි ◌ී ◌ි

Figure 39: “Sinhala INet Font Ver. 0.1”[18] used to display computer-generated election results transmitted via national television.

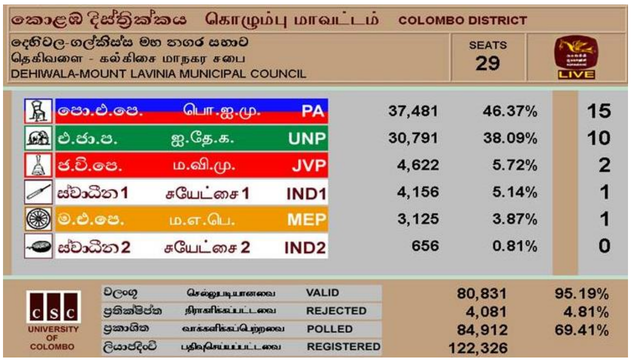


PRESIDENTIAL ELECTION 1982

COMPUTER DISPLAY by UNIVERSITY OF COLOMBO Computer Centre

CUMULATIVE RESULTS

```
CANDIDATE        VOTES      %
[illegible]ADUWA 2548438  39.07
[illegible]      3450811  52.91
DE SILVA           58531   0.90
NANAYAKKARA        17005   0.26
PONNAMBALAM       173934   2.67
WIJEWEERA         273428   4.19
Valid Votes  6522147  Reject 80470
```



කොළඹ දිස්ත්‍රික්කය கொழும்பு மாவட்டம் COLOMBO DISTRICT

දෙහිවල-ගල්කිස්ස මහ නගර සභාව
தெகிவளை - கல்கிசை மாநகர சபை
DEHIWALA-MOUNT LAVINIA MUNICIPAL COUNCIL

SEATS 29

LIVE

පො.එ.පෙ.	பொ.ஐ.மு.	PA	37,481	46.37%	15
එ.ජා.ප.	ஐ.தே.க.	UNP	30,791	38.09%	10
ජ.වි.පෙ.	ம.வி.மு.	JVP	4,622	5.72%	2
ස්වාධීන 1	சுயேட்சை 1	IND1	4,156	5.14%	1
ම.එ.පෙ.	ம.எ.பெ.	MEP	3,125	3.87%	1
ස්වාධීන 2	சுயேட்சை 2	IND2	656	0.81%	0

වලංගු	செல்லுபடியானவை	VALID	80,831	95.19%
ප්‍රතික්ෂිප්ත	நிராகரிக்கப்பட்டவை	REJECTED	4,081	4.81%
ප්‍රකාශිත	வாக்களிக்கப்பெற்றவை	POLLED	84,912	69.41%
ලියාපදිංචි	பதிவுசெய்யப்பட்டவை	REGISTERED	122,326	

CSC UNIVERSITY OF COLOMBO

Figure 40: Screenshots of the broadcasted election result in 1982 and 1992 via national television. “Sinhala INet Font” Set used to display computer-generated election results.

4.2 Early Unsuccessful Attempts

The introduction of IBM PCs for data processing and the need for developing appropriate applications were significant challenges in computer-processing languages like Sinhala. Existing technologies lagged

[18] “Sinhala INet Font” developed by S T Nandasara and later, he developed further this font set and used for his Wordprocessor package – “*Wadan Tharuwa*”.

far behind in proper handling of such complex scripts. The first Sinhala word processor, developed by a Chinese company in 1984, was not successful in Sri Lanka because of the input method's unacceptable behaviour (such as having to use two keystrokes for single characters like *ṇa* (ණ) as shown in the Figure 41, and the 12-line, 40 column character display). After that, another two Sinhala word processor was introduced, developed by the GIST (Graphics and Intelligence-based Script Technology) group in India (Vinod and Jivesh, 2007) and AEM (Applied Electro-Magnetics Pvt Ltd), needed an additional hardware plug-in board to display 80 columns and 24 lines of Sinhala text. Further, these GIST and AEM input methods based on the phonetic keyboard, which was not acceptable to Sri Lanka. Therefore, these efforts were not successful.

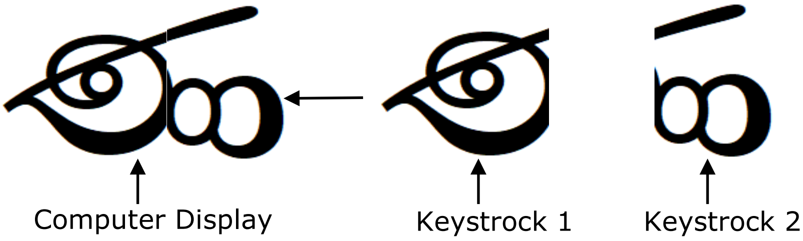


Figure 41: First foreign company from China developed Sinhala Wordprocessor used more one key stocks to display some consonants like nna (ණ) in the computer display.

Some local computer venders were interested in developing software for IBM compatible personal computer end up with a patent dispute over the software developed by one company against other company and it was a remarkable historical and significant setback for the development of Sinhala language computing in Sri Lanka. This averted when an injunction on the development of Sinhala word processors taken by one developer against another based on a disputable patent settled out of court after more than two years of litigation.

4.3 Long Debates on Alphabet and Alphabetical Order

Since the mid-1980s, several steps were taken by the government to formulate Sinhala language-related discrepancies, one of which was the different alphabetical orders used by different dictionaries. As a result, some of the committees involved have been working on the Sinhala-language character set and its alphabetical order. Because of the importance of information interchange among computers in the national language, the Information and Technology Council of Sri Lanka (CINTEC) identified the requirement for a standard code in 1985. Soon after, a working committee was formed on the use of Sinhala and Tamil in computer technology headed by the CINTEC chairman, professor V.K. Samaranayake. One of the committee's initial aims was to establish a standard code for information interchange in Sinhala to avoid the difficulties that would arise if a large number of different coding systems were permitted. The committee shortly agreed on an acceptable Sinhala alphabet, an alphabetical order, and the keyboard layout. In the same year, 1985, the above committee joined with a committee appointed by the Natural Resources, Energy, and Science Authority of Sri Lanka (NARESA) to form yet another committee: CANLIT (the Committee on Adaptation of National Languages in Information

Technology). The initial discussion based on a document prepared by the CANLIT committee's working paper (Working Paper, 1985). After that, work progressed on the alphabetical order and standard codes, the associated problems of character addition, and the essential features and shape of each character. For the Tamil, which is the second National language of Sri Lanka, no action was taken due to the work that has been progressing in South India.

CINTEC/CANLIT arrived at defining the Sinhala alphabet as having 16 vowels, 02 diacritical marks, 41 consonants, 13 consonant modifiers and 02 medial signs, which the paper presented to the 9th Annual Conference of Computer Society of Sri Lanka (CSSL) (Samaranayake, Disanayaka, and Nandasara, 1989). A new character to denote *fa* (ෆ) was introduced formally to the standard as a last character in the consonant set. CINTEC/CANLIT also agreed on the alphabetical order as given in same publication with a slight modification. It should be noted that this exercise had formed a collaborative work atmosphere among language and technology experts. However, they had lengthy discussions for several years to agree at a consensus solution.

4.4 Era of SLASCII – Typewriter Metaphor

During the late-1980s, technology author made several research visits to Thailand to study Thai language development for information processing[19]. The development of the Sinhala character set for use in IT, the similar work already progressing in Thailand for the Thai language, which is the language, derived from Brahmi family, and some similarities noticed during the visits. The aim was to develop, at that stage, an 8-bit code to fill the positions from A0 to FF in the single-byte ISO/IEC 8859 similar code table based on the keyboard's character set.

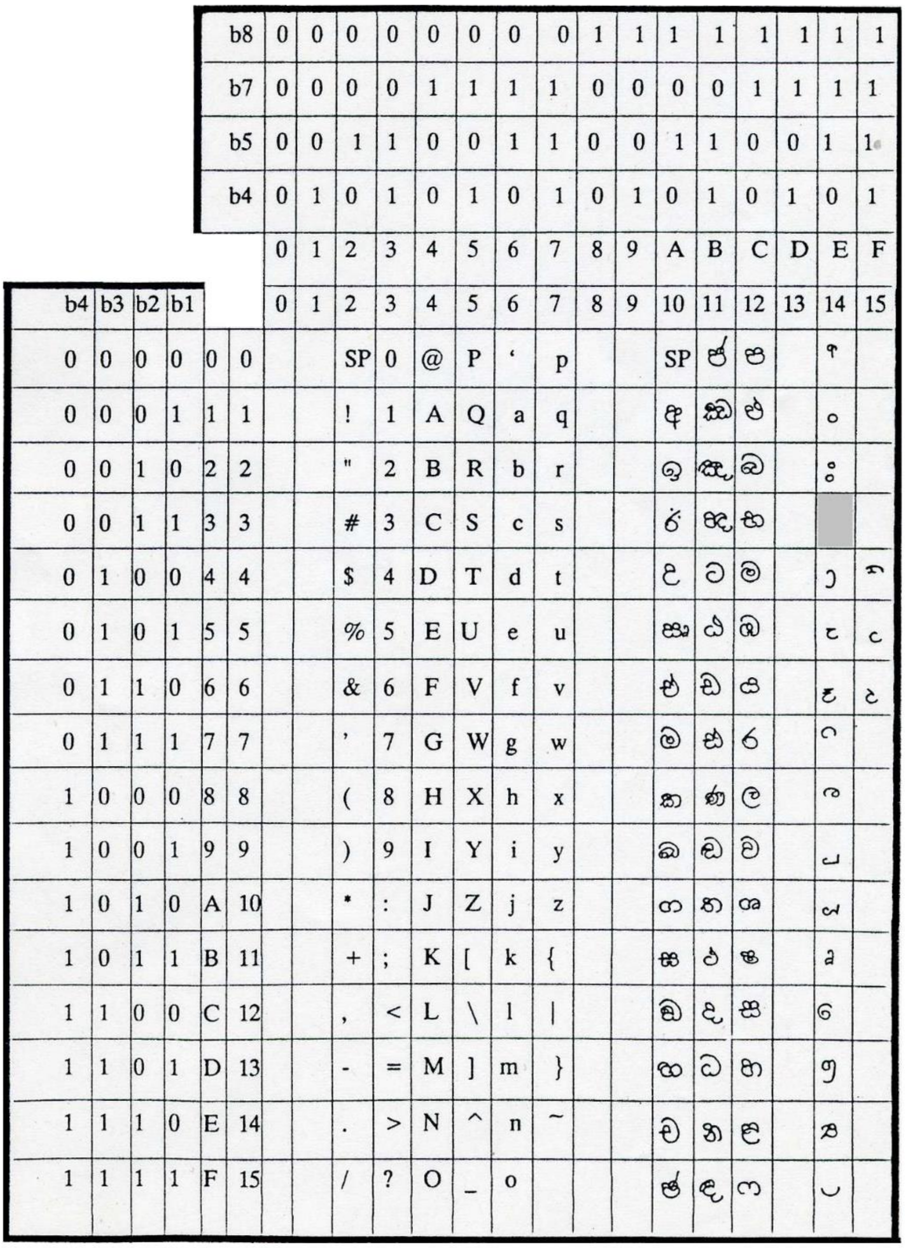
While designing the codes for internal representation, the following points have considered:

- The sorting and collating are one of the most frequently required operations; assignment of code should maintain the collating sequence of the language. Maintaining the collating sequence was not very straightforward 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. Rendering is another critical factor to consider for easiness of sorting. The vowel modifier 'ෙ◌' appearing before the consonant should always take move after the consonant to minimized the hazels of sorting in Sinhala.
- The amount of storage required in the memory and the number of code positions in the code table was another critical consideration. The total number of vowels reduced from 18 to 7, assuming others could be composed using vowel signs. Therefore, vowel signs arranged according to the sorting order.

[19] In this study, S. T. Nandasara closely worked with Dr. Thaweesak Konantakol, Information Processing Institute for Education and Development (IPIED), Thammasat University, Bangkok, Thailand.

- The Sinhala code should base itself on the ASCII standards adopted for the English language, especially in terms of control characters, line-drawing characters, special symbols and escape sequences. This would enable existing software and communication links to be totally compatible.
- As far as possible there should be a direct correspondence to existing Sinhala typewriter keyboard.

As a result of the collaborative work with the Thammasat University and the inputs from the CINTEC Working Committee on the Use of Sinhala and Tamil in Computer Technology, the draft standard was released as a CINTEC publication (Nandasara, et al. 1990) to the general public for comments and observations in March 1990 (see Figure 42).



					b8	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
					b7	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
					b5	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
					b4	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
						0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
b4	b3	b2	b1			0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	0			SP	0	@	P	‘	p			SP	ඣ	ප		්	
0	0	0	1	1	1			!	1	A	Q	a	q			අ	ඤ	ඵ		ං	
0	0	1	0	2	2			"	2	B	R	b	r			ඉ	ඥ	බ		ඃ	
0	0	1	1	3	3			#	3	C	S	c	s			ඍ	ඦ	භ			
0	1	0	0	4	4			$	4	D	T	d	t			උ	ට	ම		ා	ර්‍
0	1	0	1	5	5			%	5	E	U	e	u			ඏ	ඨ	ඹ		ැ	්‍ර
0	1	1	0	6	6			&	6	F	V	f	v			එ	ඩ	ය		ෑ	්‍ය
0	1	1	1	7	7			’	7	G	W	g	w			ඔ	ඪ	ර		ි	
1	0	0	0	8	8			(	8	H	X	h	x			ක	ණ	ල		ී	
1	0	0	1	9	9			)	9	I	Y	i	y			ඛ	ඬ	ව		ු	
1	0	1	0	A	10			•	:	J	Z	j	z			ග	ත	ශ		ූ	
1	0	1	1	B	11			+	;	K	[	k	{			ඝ	ථ	ෂ		ෘ	
1	1	0	0	C	12			,	<	L	\	l	|			ඞ	ද	ස		ෙ	
1	1	0	1	D	13			-	=	M	]	m	}			ච	ධ	හ		ෟ	
1	1	1	0	E	14			.	>	N	^	n	~			ඡ	න	ළ		ෲ	
1	1	1	1	F	15			/	?	O	_	o				ජ	ඳ	ෆ		ු	

Figure 42: First-ever encoding for Sinhala character set submitted for the public comment, 1990 (SLASCII code for Sinhala character set (ISO/IEC 646 extension)

Having several workshops and seminars about the draft, received comments and recommendation from the public comments and recommendations, the CINTEC Council approved the draft standard on the advice of its working committee for recommending standards for the use of Sinhala and Tamil Script in Computer Technology. The 8-bit code table defining 65 Sinhala characters in the A1-A7 (allocated for vowels), A8-CF (allocated for consonants), E0-E2 (allocated for virama and two diacritical marks), E3 kept reserve for used only with sorting algorithms, E4-ED (allocated for consonant modifier) EE-EF (two medial signs) and F4-F6 (allocated for additional consonants modifiers) as shown in Figure 42 along with extended Wijesekara keyboard (*see* Figure 43) was submitted to the Sri Lanka Standards Institute (SLSI) as the Sri Lanka Sinhala Standard Code for Information Interchange (SLASCII) for implementation (Nandasara, Samaranayake and Disanayaka, 1991; Nandasara,1991).

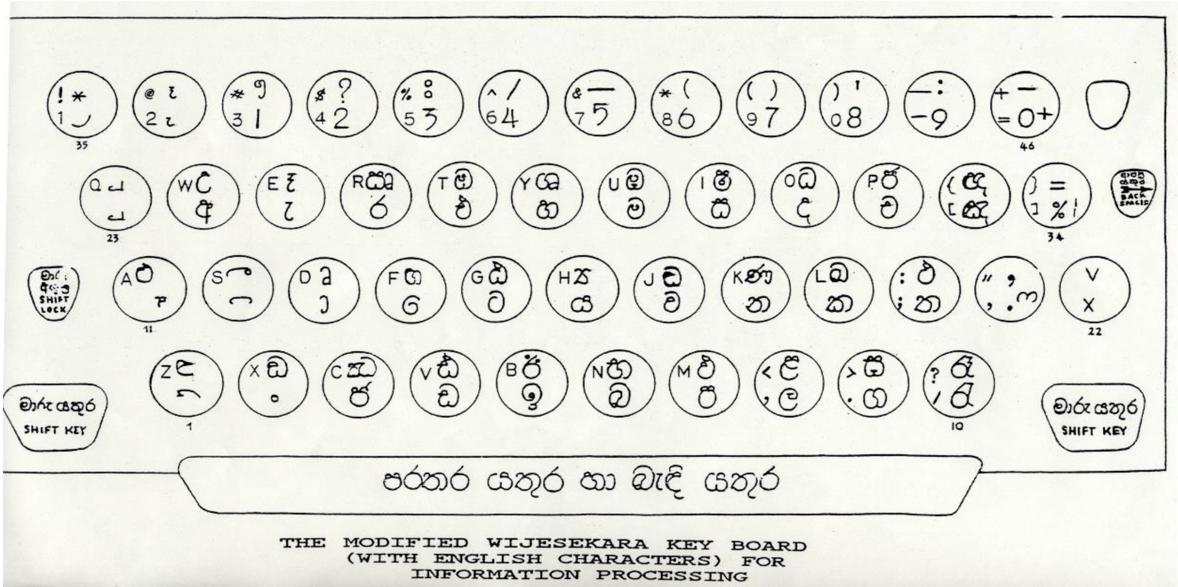


Figure 43: Extended Wijesekara keyboard for electronic typewriters (1989)

4.5 Keyboard for Sinhala

At this stage, it is essential to indicate that for the development of the appropriate electronic keyboard layout where again CINTEC took the initiative. Having agreed that a large number of Sinhala typists were using the government-approved *Wijesekara*[20] Sinhala Typewriter Keyboard (Figure 44), CINTEC first developed and obtained government approval for the “Extended *Wijesekara* Keyboard for Computers” (*see* Figure 45), the intention being the introduction of electronic typewriters then used as an interface for microcomputer output.

[20] *Wijesekara* Typewriter Keyboard was approved by the government of Sri Lanka as a National Sinhala Typewriter in 1968.

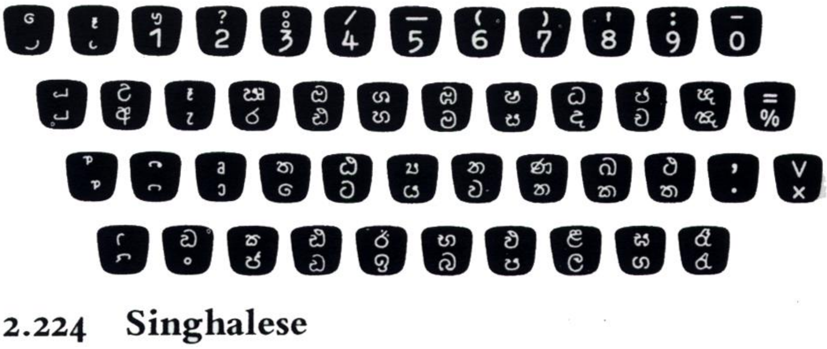


Figure 44: Wijesekara Sinhala typewriter keyboard layout (1964) (Courtesy: Olympia International) [21]

Based upon SLASCII standard and having agreed that a large number of Sinhala typists were using the government-approved *Wijesekara* typewriter keyboard, the draft included the new character *fa* (ෆ) and four other additional key positions (ඟ *ňga*, ඬ *ňḍa,*ඣ *jha* and ඳ *ňda*) by removing three half characters; *half-ka*, *half-ta* and *half-na* (ක්, ත් and න්) and two symbols that were no longer needed. Initially approved this keyboard layout, later on, once again been modified for the use of the 101-key Standard English Keyboard (Samaranayake et al. 1994).

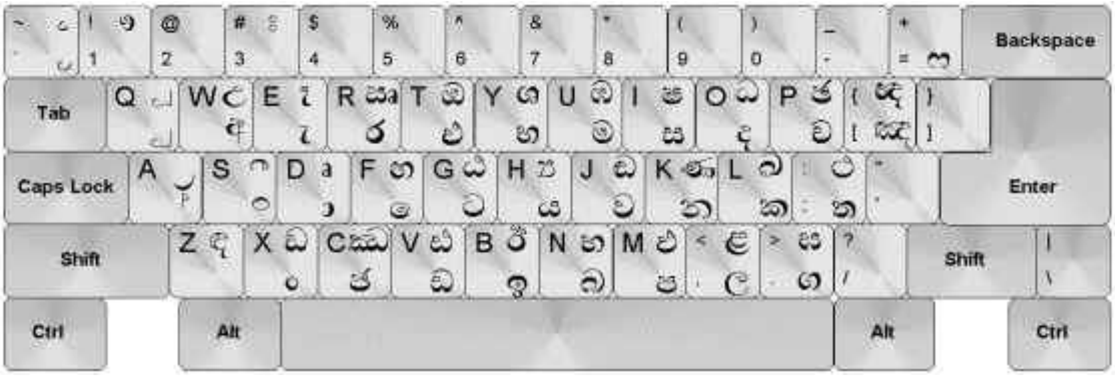


Figure 45: Extended Wijesekara keyboard for computers

4.6 Input Method for Sinhala Character Set

The author attached to the Institute of Computer Technology (ICT) of the University of Colombo, initiated collaborative work with Dr Thaweesak Koanantakool, Thammasat University to incorporate Sinhala capabilities for the personal computer based on the approved SLASCII standard. The research was used to create the first-ever Sinhala/English bilingual character-based API called SBIOS (SBIOS-Sinhala Basic Input/Output System). See Figure 46 for the architecture and the functions of the SBIOS systems. How the font rendering works will be discussed later in this chapter. The approved *wijesekara* keyboard layout was used with Sinhala word processor WT Ver. 1.0 (well known as "*Wadan Tharuwa*")[22]

[21] "*Century of the Typewriters*", 1974. Wilfred A. Beeching, British Typewriter Museum. pp. 66

[22] *Wadan Tharuwa* is a one of the earliest bi-lingual and menu-driven commercial word processors released in Sri Lanka to run on IBM-PC and it was conformed to SLASCII. The name "*Wadan Tharuwa"* meaning "Word Star" was developed by S. T. Nandasara at ICT.

developed by the author in the early 1990s for IBM-PC computers (see Figure 47). According to *Wadan Tharuwa*, running on SBIOS API architecture, Sinhala words are input and stored letter-by-letter from left to right. The architecture of this system used the three-layer system for the cells which contained symbols in base level, above or below levels. The base character was stored first, followed by the lower and then upper if the case arises. The system alarmed for illegal input key sequence such as there could not be any diacritic at the lower level after the upper-level diacritic is placed. For example, "ශ්‍රී නේත්‍රාලෝක තෛලය" (*Srī Neththālōka Thailaya*) will be input and stored as "ශ ු ී නෙ ් ත ු ා ෙ ල ා ් ක තෛලය".

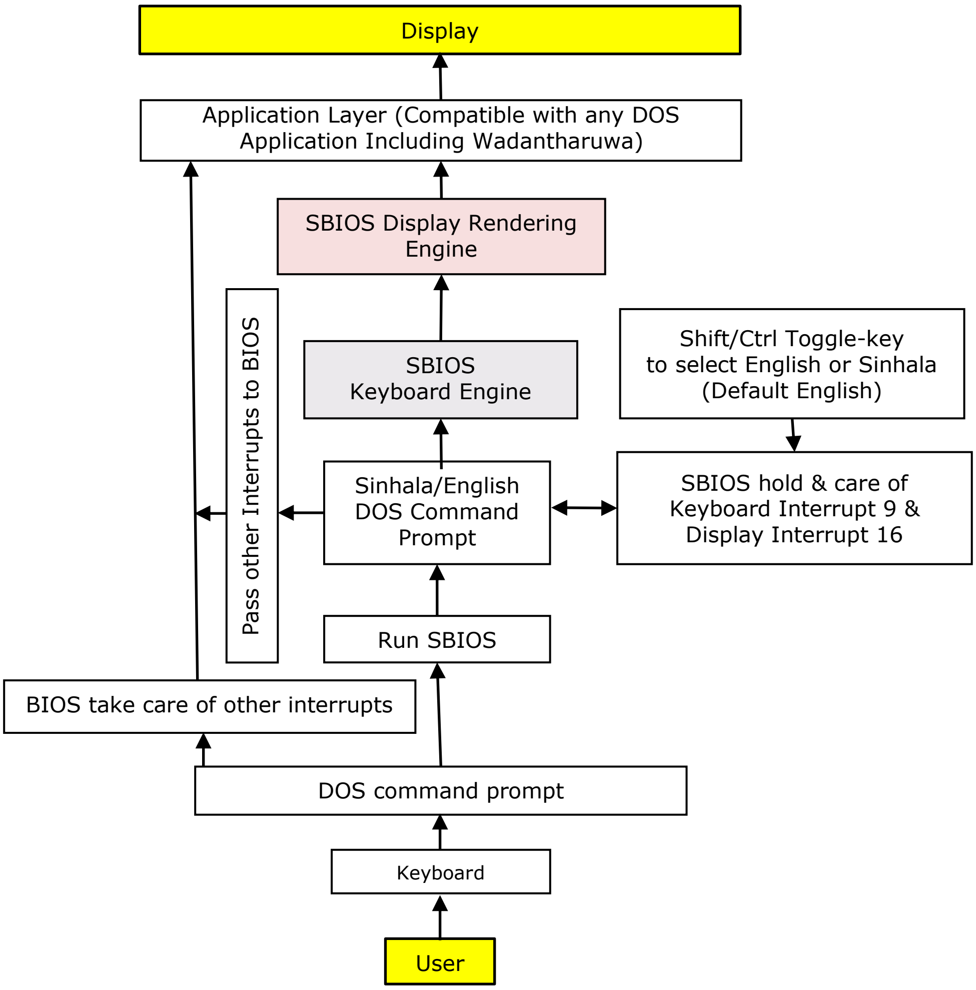

Display
Application Layer (Compatible with any DOS Application Including Wadantharuwa)
SBIOS Display Rendering Engine
Pass other Interrupts to BIOS
SBIOS Keyboard Engine
Shift/Ctrl Toggle-key to select English or Sinhala (Default English)
Sinhala/English DOS Command Prompt
SBIOS hold & care of Keyboard Interrupt 9 & Display Interrupt 16
Run SBIOS
BIOS take care of other interrupts
DOS command prompt
Keyboard
User

Figure 46: System architecture and function of the SBIOS system

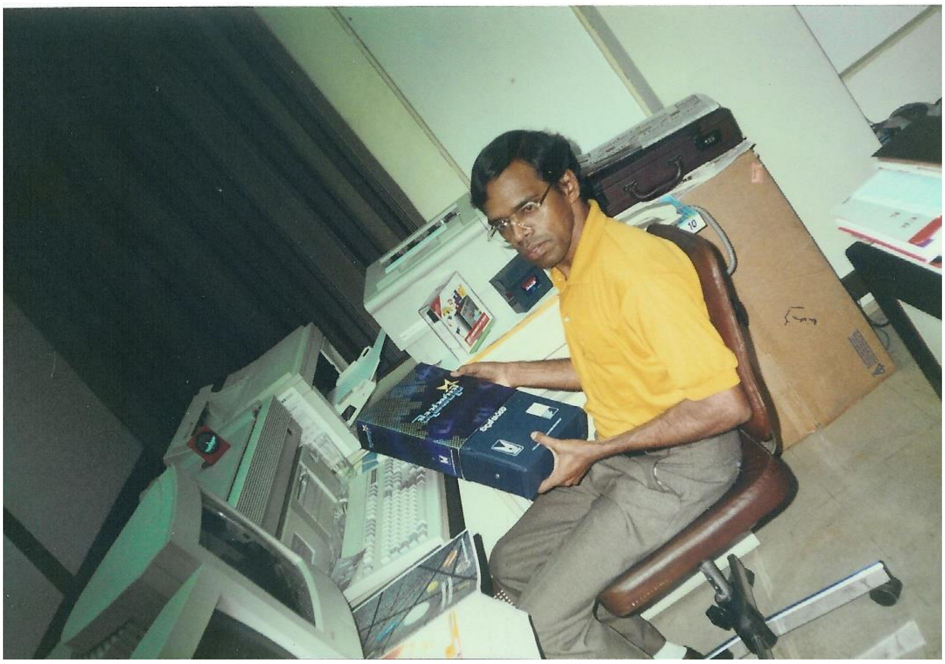
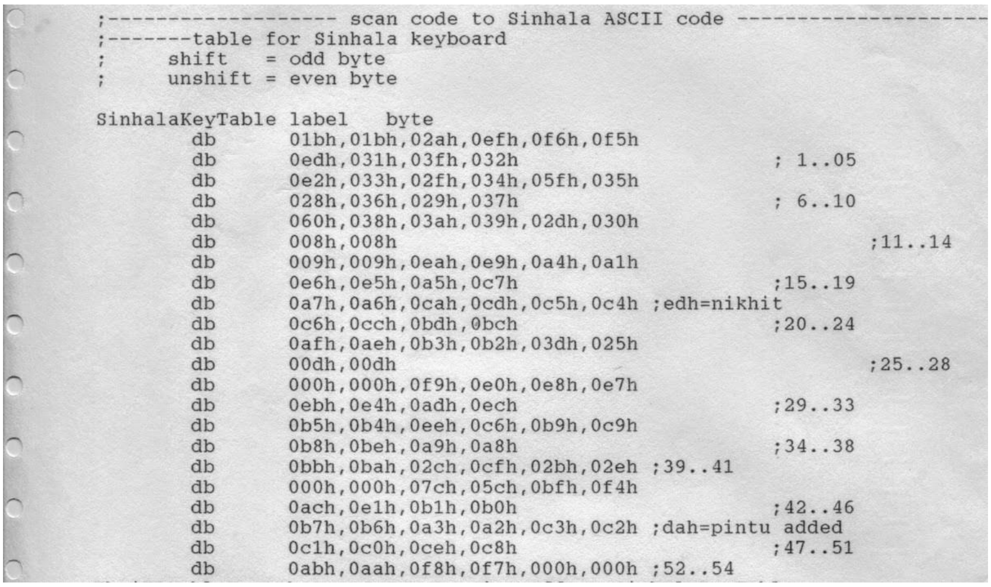


Figure 47: Sinhala Word Processor Wadantharuwa (වදන් තරුව) package with the developer (Author of this dissertation).

Base upon the modification approved by the CINTEC for “Extended Wijesekara Keyboard” for computers, Figure 48 shows the SBIOS Assembly source code which assigned the scancode table data with SLASCII code and registers keyboard controller on the motherboard. The scancode table (or scan code) is the data that most computer keyboards send to a computer to report which keys have pressed. The Assembly routing of SBIOS assigned two codes that perform *shift* and *unshift* functions in general, for example, key numbers 17 (which is the ‘W’ key on “QWERTY” keyboard) assigned two code which equal to Sinhala vowels උ (HxA4) and අ (HxA1) with *shift* and *unshift* key-press.



```
;------------------- scan code to Sinhala ASCII code --------------------
;-------table for Sinhala keyboard
;      shift    = odd byte
;      unshift  = even byte

SinhalaKeyTable label    byte
        db       01bh,01bh,02ah,0efh,0f6h,0f5h
        db       0edh,031h,03fh,032h                   ; 1..05
        db       0e2h,033h,02fh,034h,05fh,035h
        db       028h,036h,029h,037h                   ; 6..10
        db       060h,038h,03ah,039h,02dh,030h
        db       008h,008h                                     ;11..14
        db       009h,009h,0eah,0e9h,0a4h,0a1h
        db       0e6h,0e5h,0a5h,0c7h                   ;15..19
        db       0a7h,0a6h,0cah,0cdh,0c5h,0c4h ;edh=nikhit
        db       0c6h,0cch,0bdh,0bch                   ;20..24
        db       0afh,0aeh,0b3h,0b2h,03dh,025h
        db       00dh,00dh                                     ;25..28
        db       000h,000h,0f9h,0e0h,0e8h,0e7h
        db       0ebh,0e4h,0adh,0ech                   ;29..33
        db       0b5h,0b4h,0eeh,0c6h,0b9h,0c9h
        db       0b8h,0beh,0a9h,0a8h                   ;34..38
        db       0bbh,0bah,02ch,0cfh,02bh,02eh ;39..41
        db       000h,000h,07ch,05ch,0bfh,0f4h
        db       0ach,0e1h,0b1h,0b0h                   ;42..46
        db       0b7h,0b6h,0a3h,0a2h,0c3h,0c2h ;dah=pintu added
        db       0c1h,0c0h,0ceh,0c8h                   ;47..51
        db       0abh,0aah,0f8h,0f7h,000h,000h ;52..54
```

Figure 48: SBIOS API Assembly language routing used to register SLASCII codes to the keyboard controller.

The architecture of the SBIOS API input method provided a *sequence checking mechanism* to ensure the validity of the input sequence. The *sequence checking mechanism* is capable of monitoring of the key-in sequences of three levels; that is of strictness; base level: pass through, lower level: basic check, and upper level: strict check.

This mechanism of sequence checking provided for three reasons.

- The first is to maintain logical sorting order of the alphabet, i.e. ශ ්‍රී is rendered as ශ්‍රී *(Srī)* or ශ්රී *(Srī)* which is the valid order for the sorting as well as reading, but not ශ ී ්‍ර which gives ශීර් *(Sīr)*. This is the wrong interpretation for internal sorting sequence, and as a result, it will place in different positions. If it is the case, the correct input sequence would be as ශ ් ර ී.
- The second is to maintain the visual correctness of the character display, i.e. the diacritical mark ් should be placed in different positions depend on the consonants such as for ක it should be placed in left most like ක්, for ජ it should be placed in the middle like ජ් and for ර it should be placed in the rightmost like ර්. This is handling by SBIOS API without human intervention (see Figure 38).
- The third is to maintain the correctness of the use of vowel signs, i.e. ු, ූ, ු, ූ, ු, ෘ and ෲ used for different consonants (කු බු) for the same purpose. Unlike in traditional typewriter keyboard, Standard Sinhala keyboard not providing two keys for this purpose. API will select the correct diacritical mark from the font table and will be used for the consonants.

The SBIOS has also specified the cursor movements and editing behavior of the Sinhala application running on SBIOS such as word processor (*Wadan Tharuwa*). The cursor must be moved from cell to cell. However, all characters in levels other than the base level must be skipped. The *delete-key* must also remove all characters in the current cell, including levels above and below the cell. Additionally, character-by-character, right-to-left removal is still possible by using the *backspace-key*, where the order of removal is considered by the order in which they are stored.

As mentioned earlier, the SBIOS API featured to handle almost all the ligatures used in Sinhala writing; however, SBIOS does not facilitate to generate a joint ligature, i.e., ක්‍ව, ත්‍ව and න්‍ද, *repaya* (ර්‍) and other conjunct formations. The font rendering engine uses the base, upper and lower level shapes to construct vowels, consonants and complex ligatures. Table 12 illustrates this mechanism of construction of ligatures. As shown in Tables 9, the vowel signs are always reordered in front of the previous consonant cluster in SBIOS API to maintain the internal sorting sequence.

Table 9: Key-in sequence and vowel signs re-ordering process in Sinhala SBIOS API

Key-in Sequence	Memory representation after re-ordering	Display
ෙ ක	ක ෙ	කෙ
ෙ ක ්	ක ෙ ්	කේ
ෙ ක ්‍ර ්	ක ්‍ර ෙ ්	ක්‍රේ
ෙ ක ්‍ර ා ්	ක ්‍ර ෙ ා ්	ක්‍රෝ

4.7 Designing Font Matrix & Development of Font Display Engine

There were four main reasons for the development of the new font-display engine for the Sinhala language.

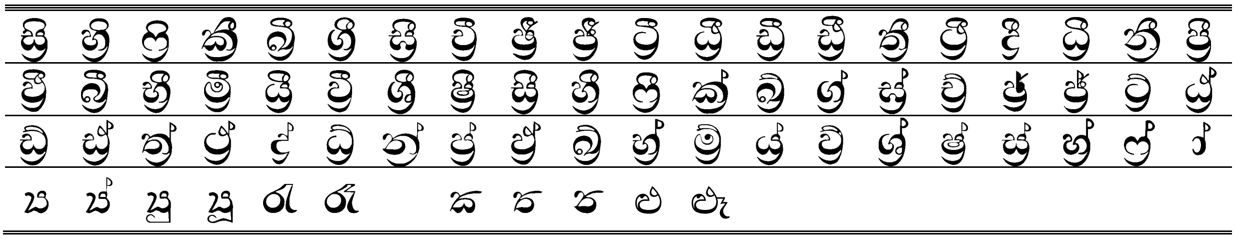
- Owing to the nature of the Sinhala glyphs, vowel signs appeared above and below the consonants and had to be reshaped accordingly.
- The set brought the total number of distinct glyphs to almost 390 (see Table 10), including the combinations with the special symbol *ra* (◌්‍ර).
- Using existing CGA/EGA graphics adapter, this number of glyphs was hard to accommodate in the internal font matrix, which was limited by hardware to a 4-Kbyte memory (256 locations with 16x8 bits per location—the cell size cannot be reprogrammed because of the hardware limitation). Therefore some degree of expansion of resolution was desired[23].

Table 10: Minimum set of glyphs for Sinhala[24]

◌ං	◌ඃ	අ	ආ	ඇ	ඈ	ඉ	ඊ	උ	ඌ	ඍ	ඎ	ඏ	ඐ	එ	ඒ	ඓ	ඔ	ඕ	ඖ
◌්	◌ා	◌ැ	◌ෑ	◌ි	◌ී	◌ු	◌ූ	◌ෘ	ෙ◌	◌ෟ	◌ෲ	◌ෳ	ක	ඛ	ග	ඝ	ඞ	ඟ	ච
ඡ	ජ	ඣ	ඤ	ඥ	ඦ	ට	ඨ	ඩ	ඪ	ණ	ඬ	ත	ථ	ද	ධ	න	ඳ	ප	ඵ
බ	භ	ම	ඹ	ය	ර	ල	ව	ශ	ෂ	ස	හ	ළ	ෆ	ක්	ඛ්	ග්	ඝ්	ඞ්	ච්
ඡ්	ජ්	ඣ්	ඤ්	ඥ්	ට්	ඨ්	ඩ්	ඪ්	ණ්	ඬ්	ත්	ථ්	ද්	ධ්	න්	ඳ්	ප්	ඵ්	බ්
භ්	ම්	ඹ්	ය්	ර්	ල්	ව්	ශ්	ෂ්	ස්	හ්	ළ්	ෆ්	කි	ඛි	ගි	ඝි	ඟි	චි	ඡි
ජි	ඣි	ඤි	ඥි	ටි	ඨි	ඩි	ඪි	ණි	ඬි	ති	ථි	දි	ධි	නි	ඳි	පි	ඵි	බි	භි
මි	ඹි	යි	රි	ලි	වි	ශි	ෂි	සි	හි	ළි	ෆි	කී	ඛී	ගී	ඝී	ඟී	චී	ඡී	ජී
ඣී	ඤී	ඥී	ටී	ඨී	ඩී	ඪී	ණී	ඬී	තී	ථී	දී	ධී	නී	ඳී	පී	ඵී	බී	භී	මී
ඹී	යී	රී	ලී	වී	ශී	ෂී	සී	හී	ළී	ෆී	කු	ඛු	ගු	ඝු	ඟු	චු	ඡු	ජු	ඣු
ඤු	ටු	ඨු	ඩු	ඪු	ණු	ඬු	තු	ථු	දු	ධු	නු	ඳු	පු	ඵු	බු	භු	මු	ඹු	යු
රු	ලු	වු	ශු	ෂු	සු	හු	ළු	ෆු	කූ	ඛූ	ගූ	ඝූ	ඟූ	චූ	ඡූ	ජූ	ඣූ	ඤූ	ටූ
ඨූ	ඩූ	ඪූ	ණූ	ඬූ	තූ	ථූ	දූ	ධූ	නූ	ඳූ	පූ	ඵූ	බූ	භූ	මූ	ඹූ	යූ	රූ	ලූ
වූ	ශූ	ෂූ	සූ	හූ	ළූ	ෆූ	ක්‍ර	ඛ්‍ර	ග්‍ර	ඝ්‍ර	ච්‍ර	ඡ්‍ර	ජ්‍ර	ට්‍ර	ඨ්‍ර	ඩ්‍ර	ඪ්‍ර	ත්‍ර	ථ්‍ර
ද්‍ර	ධ්‍ර	න්‍ර	ප්‍ර	ඵ්‍ර	බ්‍ර	භ්‍ර	ම්‍ර	ය්‍ර	ව්‍ර	ශ්‍ර	ෂ්‍ර	ස්‍ර	හ්‍ර	ෆ්‍ර	ක්‍රි	ඛ්‍රි	ග්‍රි	ඝ්‍රි	ච්‍රි
ඡ්‍රි	ජ්‍රි	ට්‍රි	ඨ්‍රි	ඩ්‍රි	ඪ්‍රි	ත්‍රි	ථ්‍රි	ද්‍රි	ධ්‍රි	න්‍රි	ප්‍රි	ඵ්‍රි	බ්‍රි	භ්‍රි	ම්‍රි	ය්‍රි	ව්‍රි	ශ්‍රි	ෂ්‍රි

[23] This was possible only after Video Graphics Array (VGA) was released to the market by IBM in 1987.

[24] S T Nandasara developed "Sarasavi" font for his Tri-lingual Wordprocessor marketed in early 19s.



Two other noteworthy complications were that, first, using the existing CGA/EGA graphics adapter resolution for a single character (16x8), it was not possible to design the rounded shape complex ligature. Therefore, some degree of expansion of resolution was needed. The expansion of the resolution became possible only after IBM released the Video Graphics Array (VGA) card in 1987. Using VGA graphics adapter, internal font matrix, which was limited by hardware to a 4-Kbyte memory increases to 5Kbyte using software (256 locations with 20x8 bits per location—the cell size reprogrammed because of the VGA character matrix memory was not limited by hardware). Therefore, such a degree of expansion of resolution was a tremendous technological help for developing complex glyphs for the Sinhala language.

Table 11: Vowels & consonants groups according to the basic appearance of the character shapes

Individual Vowels and Consonants Grouped according to their shapes														
1	අ													
2	ඉ													
3	ඊ													
4	උ	ල												
5	එ	ට	ඨ	ඪ	ඵ									
6	ඔ	ඛ	ඩ	ච	ට	ඞ	ඬ	ධ	බ	ම	ඹ	ව		
7	ක	ග	ඝ	ඟ	ත	න	ප	භ	ය	ශ	ෂ	ස	හ	ෆ
8	ඡ	ජ	ඦ											
9	ඤ	ඥ	ද	ඳ										
10	ණ	ර												
11	ඣ													
12	ළ													

Second, to maintain the existing ISO/IEC 646 character-set as it was, in addition to Sinhala characters, the line-drawing characters, mathematical symbols, and most useful European characters were stored in the extended area. The SBIOS display engine overcame these two limitations with the ‘‘modular concept’’ approach, by considering the shapes of Sinhala characters and using the VGA architecture. To illustrate, note that the Sinhala character set can be grouped according to the basic appearance of character shapes, as shown in Table 11. There can see some similarities in the upper and lower parts of the character-set itself (some isolated cases, of course, have to be treated separately, like ඉ, ඨ, ඎ, and ළු).

The three-layer system discussed earlier encouraged developing the font display engine to maintain the proper character display in the IBM-PC environment. Figure 49 shows how the font memory, which has been arranged to accommodate the Sinhala character-shapes for display. Accordingly, font-rendering methodologies were implemented (as Figure 49, and Figure 51, show) for Sinhala/English bilingual-character-based systems (see Figure 52 for the sample *Wadantharuwa* screenshot taken from an IBM PC-compatible computer).

At a later stage, remaining cells in the extended area were used to maintain line drawing characters, diacritical marks and mathematical and phonetic symbols need for the DOS operating system. The methodology and technology used to develop and introduce Sinhala were versatile, it was easy to introduce to any other Indic languages; as an example Tamil was introduced in the same manner to used *Wadantharuwa* in Sri Lanka, where, user selected the language by toggling the Shift-Ctrl key combination when required (Nandasara and Samaranayake, 1994; Nandasara and Samaranayake, 1997; Nandasara, 1997; Nandasara et al., 1997).

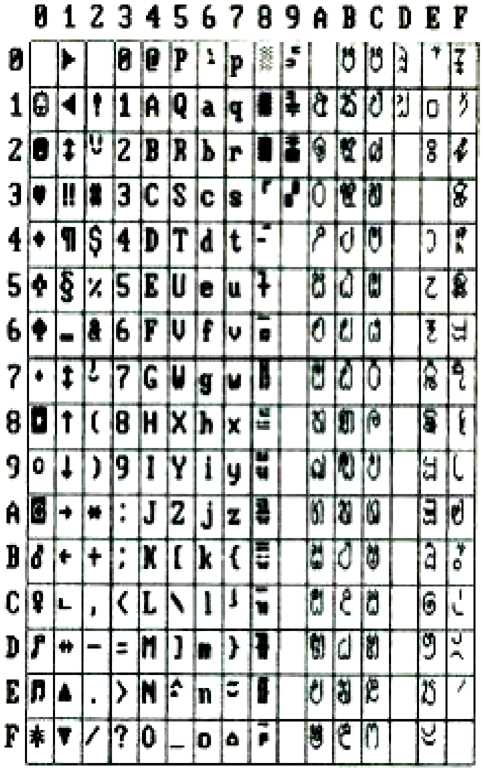

0 1 2 3 4 5 6 7 8 9 A B C D E F
0
1
2
3
4
5
6
7
8
9
A
B
C
D
E
F

Figure 49: Memory dump of the Sinhala text shapes as stored in character matrix memory (Source: Author’s collection, 1988)

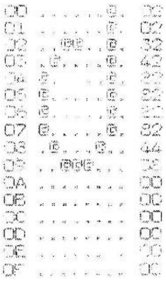
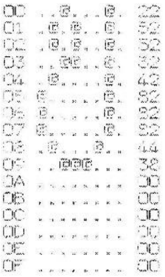
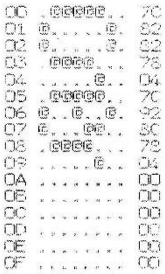
The base part of the character ට (raw 0 - 9) – Base Level	The base part of the character ඩ (raw 0 - 9) – Base Level	Upperparts of the ligatures ඩි & ඩී (raw 0-4 and 5-9)
 00 ......@. 02 01 ......@. 02 02 ..@@..@. 32 03 .@....@. 42 04 @.....@. 82 05 @.....@. 82 06 @.....@. 82 07 @.....@. 82 08 .@...@.. 44 09 ..@@@... 38 0A .... 00 0B .... 00 0C .... 00 0D .... 00 0E 00 0F 00	 00 ..@...@. 22 01 .@.@..@. 52 02 .@.@..@. 52 03 ..@@..@. 32 04 .@....@. 42 05 @.....@. 82 06 @.....@. 82 07 @.....@. 82 08 .@...@.. 44 09 ..@@@... 38 0A .... 00 0B .... 00 0C .... 00 0D .... 00 0E 00 0F 00	 00 .@@@@@.. 7C 01 @.....@. 82 02 @.....@. 82 03 .@@@@... 78 04 .....@.. 04 05 .@@@@@.. 7C 06 @..@..@. 92 07 @...@@.. 8C 08 .@@@@... 78 09 .....@.. 04 0A .... 00 0B .... 00 0C .... 00 0D .... 00 0E 00 0F 00

Figure 50: Enlarged Version of the Sinhala Text Shapes Stored in Character Matrix Memory. (see B4, C9 and E8 in Figure 51)

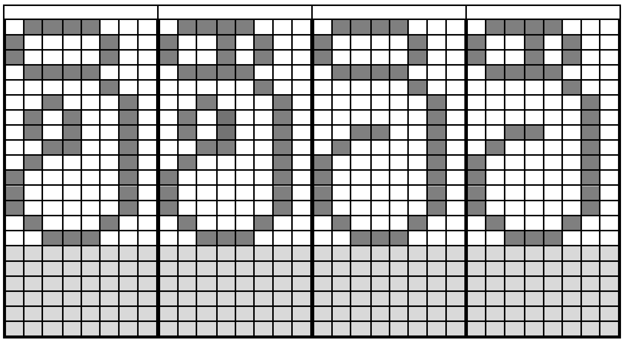


Figure 51: Font rendering engine displaying the Characters ඩි, ඩී, ට and ඩ using the parts of the characters and ligatures illustrated in Figure 48 (8 bytes, 4 cells) on IBM-PC VGA Compatible Computer by Increasing the Resolution of the Character Cell up to 20x8. The Author re-created this. (See Figure 52 for Wadantharuwa Sample Screen)

Table 12: Demonstrate how font rendering engine construct vowels, consonants and ligatures in SBIOS API using shapes stored in character matrix memory

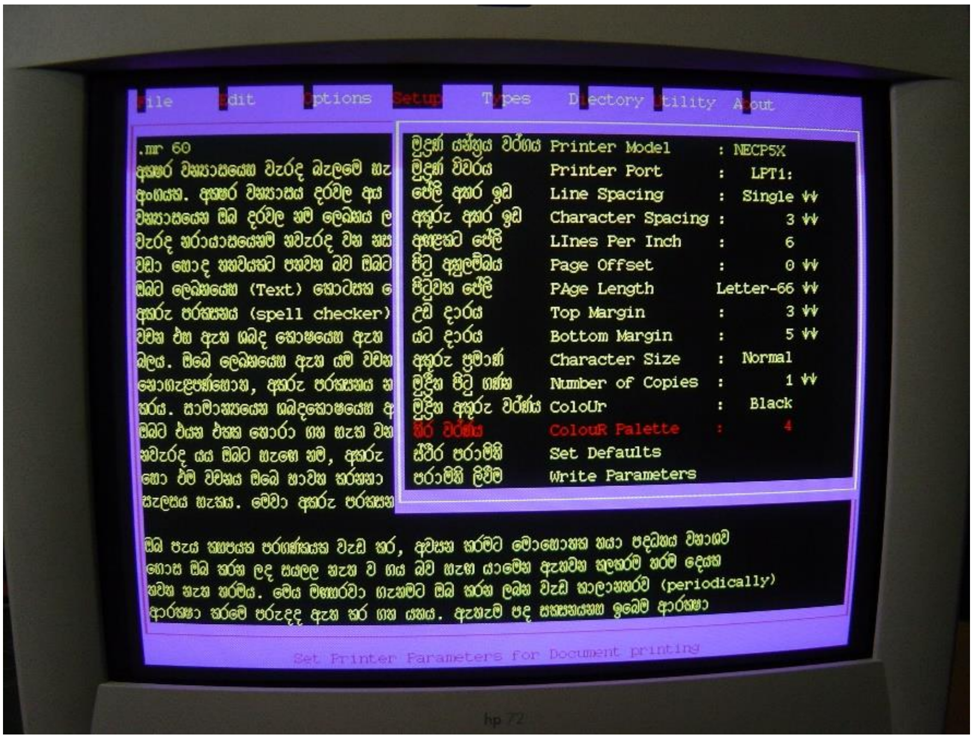
Vowels	
අ = a1$_b$ + fc$_u$	ආ = a1$_b$ + fc$_u$ + e4$_b$
ඇ = a1$_b$ + fc$_u$ + e5$_b$	ඈ = a1$_b$ + fc$_u$ + e6$_b$
ඉ = a2$_b$ + fd$_u$	ඊ = a3$_b$ + f2$_u$
උ = a4$_b$ + ef$_l$	ඌ = a4$_b$ + ef$_l$ + ed$_b$
ඍ = a5$_b$ + eb$_b$	ඎ = a5$_b$ + eb$_b$ + eb$_b$
එ = a6$_b$ + f4$_l$	ඒ = a6$_b$ + f4$_u$
ඓ = ec$_b$ + a6$_b$ +f4$_l$	ඔ = a7$_b$ + fd$_l$
ඕ = a7$_b$ + f7$_u$	ඖ = a7$_b$ + fd$_l$ + ed$_b$
Some Consonants and Simple Ligature	
ක = a8$_b$	ෂ = (b8$_b$ + f1$_l$)
කූ = (b2$_b$ + f7$_l$)	ලූ = (b3$_b$ + f7$_l$)
ඊ = (c7$_b$ + f3$_u$)	ඊ = (c7$_b$ + f3$_l$)
Some Complex Ligatures	

ක්‍රෝ = $ec_b + a8_b + ef_u + e4_b + e0_u$

ක්‍යෝ = $ec_b + a8_b + ee_b + e4_b + e0_u$

බ්‍රෝ = $ec_b + (c2_b + fd_l) + ef_u + e4_b + e0_u$

Character Matrix Memory levels: *b – base level; u – upper level; l – lower level*



.mr 60
අක්ෂර වින්‍යාසයෙහි වැරදි බැලමේ හැ
අංගයක. අක්ෂර වින්‍යාසය දරවල අය
වින්‍යාසයෙන ඔබ දරවල නම ලෙඛනය ල
වැරදි නරායාසයෙනම නවැරදි වන නස
වඩා හොද නනවයකට පනවන බව ඔබට
ඔබට ලෙඛනයෙහ (Text) කොටසක ෙ
අකුරු පරීක්ෂනය (spell checker)
වචන එන ඇත ශබ්ද කොෂයෙහ ඇත
බලය. ඔබෙ ලෙඛනයෙහ ඇත යම වචන
නොගැළපණහොත, අකුරු පරීක්ෂනය න
කරය. සාමාන්‍යයෙන ශබ්දකොෂයෙහ අ
ඔබට එයත එකක තොරා ගත හැක වන
නවැරදි යය ඔබට හැඟෙ නම, අකුරු
නො එම වචනය ඔබෙ භාවත කරනනා
සැලසිය හැකය. මෙවා අකුරු පරීක්ෂන

මුද්‍රණ යන්ත්‍රය වර්ගය	Printer Model	: NECP5X
මුද්‍රණ විවරය	Printer Port	: LPT1:
පේළි අතර ඉඩ	Line Spacing	: Single ⇊
අකුරු අතර ඉඩ	Character Spacing	: 3 ⇊
අඟලකට පේළි	LInes Per Inch	: 6
පිටු අනුලම්බය	Page Offset	: 0 ⇊
පිටුවක පේළි	PAge Length	Letter-66 ⇊
උඩ දාරය	Top Margin	: 3 ⇊
යට දාරය	Bottom Margin	: 5 ⇊
අකුරු ප්‍රමාණ	Character Size	: Normal
මුද්‍රිත පිටු ගණන	Number of Copies	: 1 ⇊
මුද්‍රිත අකුරු වර්ණය	ColoUr	: Black
තීර වර්ණය	ColouR Palette	: 4
ස්ථිර පරාමිති	Set Defaults	
පරාමිති ලිවීම	Write Parameters	

ඔබ පැය ගණනයක පරිගණකයක වැඩ කර, අවසන කරමට මොහොතක තයා පද්ධතය විනාශව
හොස ඔබ කරන ලද සියලල හැන ව ගය බව හැඟ යාමෙන ඇතවන කලකරම කරම දෙයක
තවන හැන කරමය. මෙය මඟහරවා ගැනමට ඔබ කරන ලබන වැඩ කාලානතරව (periodically)
ආරක්ෂා කරමෙ පරුද්ද ඇත කර ගත යතය. ඇතැම පද සකසනයනහ ඉබෙම ආරක්ෂා

Set Printer Parameters for Document printing

Figure 52: Sample screenshot of bi-lingual Sinhala/English Word Processor, "Wadantharuwa"

This phenomenon illustrated in Figure 52 (as a sample case) to demonstrate how it was generated by using their relative parts from each plane as a case for *mbī* (ඹී) and by limiting the use of 18 cells in extended character generator chip and construct 80 glyphs using font rendering technology (Figure 54).

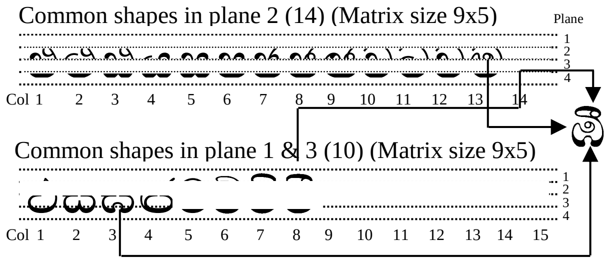

Common shapes in plane 2 (14) (Matrix size 9x5)
Plane
1
2
3
4
Col 1 2 3 4 5 6 7 8 9 10 11 12 13 14
ඹී
Common shapes in plane 1 & 3 (10) (Matrix size 9x5)
1
2
3
4
Col 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Figure 53: How mbī (ඹී) was generated

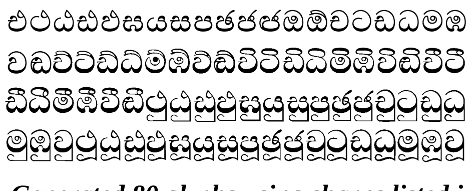


Figure 54: Generated 80 glyphs using shapes listed in Figure 53

4.8 From Bitmap to Open Font

During the mid-1990s and the introduction of desktop publishing (DTP) with PCs, there was a demand for quality printing on desktop computers. Introduction of Sinhala desktop publishing for the IBM PC was made available with Xerox Ventura as a very first attempt. However, Xerox offered no technical support for installing non-roman fonts. Tag concept, which Ventura featured for formatting text and paragraphs and the reverse engineering efforts, help to develop the *Athwela*[25] software in 1994 to support trilingual (Sinhala, Tamil, and English) DTP (see Figure 55 and Figure 56). This move-in character-rendering technology with bitmapped font technology for a laser printer opened the way for the new stage of text processing, and it coincided with the emergence of new designs for the Sinhala character code.

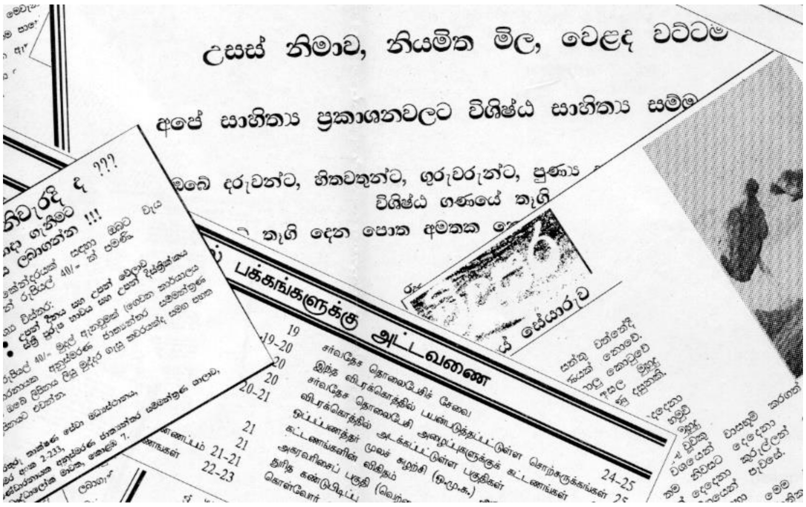


Figure 55: Tri-lingual laser output from Athwela Desktop Publishing package

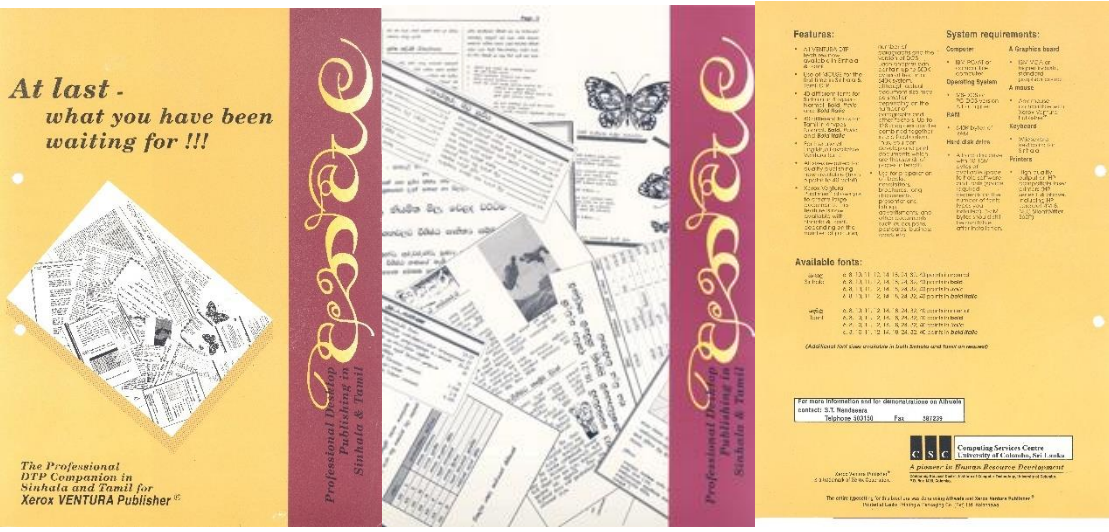
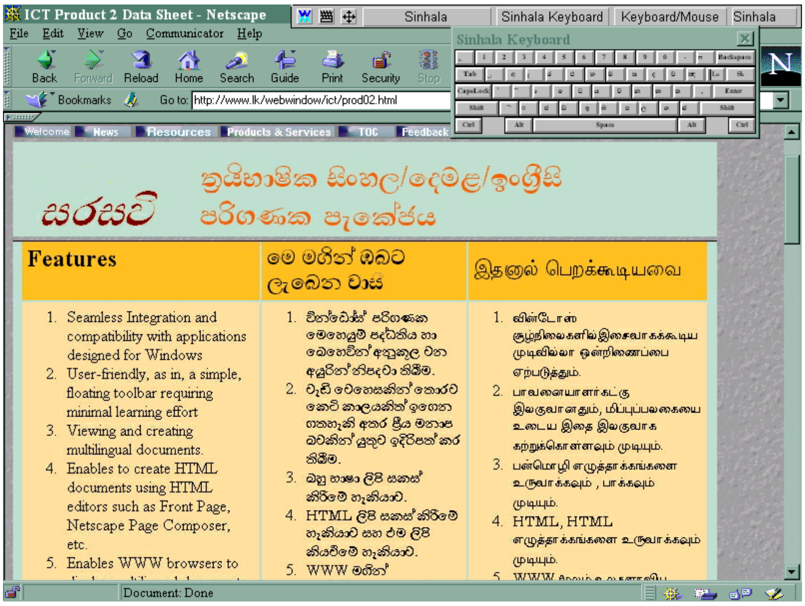


Figure 56: Three sides of 'Athwela' brochure (The samples of tri-lingual laser outputs taken from 'Athwela' were used to prepare this brochure. Image: author's collection).

[25] *Athwela* Desktop publishing software package and "*Sarasavi*" fonts developed by S T Nandasara used to typeset the document with tri-lingual features in and marketed during 1994 and onverd in SriLanka..

4.9 Tri-lingual Solution for World-Wide-Web Running of Personal Computers

The tri-lingual system, with Sinhala, Tamil and English were introduced around mid 1997s (see Figure 57 for tri-lingual website running on Netscape web browser in 1996). More applications with both language interfaces had become available. Users can switch from one language to another very easily by using toggle key in the keyboard.


ICT Product 2 Data Sheet - Netscape
Sinhala
Sinhala Keyboard
Keyboard/Mouse
Sinhala
File Edit View Go Communicator Help
Sinhala Keyboard
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks
Go to: http://www.lk/webwindow/ict/prod02.html
Welcome News Resources Products & Services TOC Feedback
සරසවි
ත්‍රයිභාෂික සිංහල/දෙමළ/ඉංග්‍රීසි
පරිගණක පැකේජය
Document: Done

Features	මෙ මගින් ඔබට ලැබෙන වාසි	இதனால் பெறக்கூடியவை
1. Seamless Integration and compatibility with applications designed for Windows 2. User-friendly, as in, a simple, floating toolbar requiring minimal learning effort 3. Viewing and creating multilingual documents. 4. Enables to create HTML documents using HTML editors such as Front Page, Netscape Page Composer, etc. 5. Enables WWW browsers to	1. වින්ඩෝස් පරිගණක මෙහෙයුම් පද්ධතිය හා බෙහෙවින් අනුකූල වන අයුරින් නිපදවා තිබීම. 2. වැඩි වෙහෙසකින් තොරව කෙටි කාලයකින් ඉගෙන ගතහැකි අතර ප්‍රිය මනාප බටකින් යුතුව ඉදිරිපත් කර තිබීම. 3. බහු භාෂා ලිපි සකස් කිරීමේ හැකියාව. 4. HTML ලිපි සකස් කිරීමේ හැකියාව සහ එම ලිපි කියවීමේ හැකියාව. 5. WWW මගින්	1. விண்டோஸ் சூழ்நிலைகளில் இசைவாக்கக்கூடிய முடிவில்லா ஒன்றிணைப்பை ஏற்படுத்தும். 2. பாவனையாளர்கட்கு இலகுவானதும், மிப்புப்பலகையை உடைய இதை இலகுவாக கற்றுக்கொள்ளவும் முடியும். 3. பன்மொழி எழுத்தாக்கங்களை உருவாக்கவும் , பாக்கவும் முடியும். 4. HTML, HTML எழுத்தாக்கங்களை உருவாக்கவும் முடியும். 5. WWW

Figure 57: Trilingual National web site of Sri Lanka - Screenshots (www.lk)

4.10 Conclusion

Sinhala BIOS was an early and famous terminate and stay residence program (STR) run for MS-DOS which enabled computer users to activate the Sinhala or English multilingual command prompt using a hotkey combination (Ctrl-Shift) on IBM PCs. Any application running on user command prompt can work on the language selection and for example Wadantharuwa (WT), a text-mode and windows interface base allowed to crate wordprocessing document in mixed language scripts. Development based on the IBM PC open architecture allowed further development stage of desktop publishing through Athwela, another simplified DTP application interface for Xerox Ventura create the opportunity to create trilingual document very easily. Time passes through till the year 2000, Sarasavi trilingual 16-bit application based on SLS 1134:1996 standard was running on Windows 95, Windows 98, Windows 2000 and Windows NT, till the year 2001. By 1995, in the world of business computing, what followed was a dark period of standardization around technology as well as the multilingualism. The IBM PC, most famous around the Asian region, may have been a sad starting point for those multilingual applications, but the credibility of IBM dictated it as the "Big Blue" master of computing. In decades, the MS-DOS command line executed, only in 1995 as an introduction of innovative developments, mostly the introduction of Graphical User Interface (GUI), "go-it-alone" computers appeared. Microsoft Corporation changes the policy of development application based on DOS into Windows, a new development era began. As a result os kernel platform and kept the dos running as an emulator. This makes all the DOS application went on sleep forever. The need for multilingual applications for society considered as an essential requirement. Standardization of the universal language character-sets was in demand. The next chapter, chapter five, discusses how small communities of non-Roman script users could connect to the Romanised system dominated cyberspace with next generation of 16-bit encoding based on Sri Lanka proposal SLS 1134:1996, which supports multilingualism.

Chapter 5: From Formation to Publication - Design of 16-bit Standards for Sinhala Script

The fifth chapter is dedicated to the eight-year-long exercise of implementing the ISO/IEC 10646 initiative and achieving the UCS/Unicode implementation for Sinhala scripts. The international scope of computing, information interchange, and electronic publishing created a need for a worldwide character-encoding scheme. UCS/Unicode is a comprehensive standard designed by Unicode consortium and ISO/IEC, developed over the last two and half decades, for standardizing more than 130,000 scripts in the world. Sinhala code standard was published in its version 3.0 in 1998, and operating systems and its application compatibilities are becoming available since then. The design of a standard to provide a portion of the Unicode—Sinhala—is described and comprehensively discussed in this chapter concerning the cultural, sociological and technological background of the entire process related to UCS/Unicode implementation.

5.1 Introduction

The new design of character code is based on a different design concept from the very early "graphical concept" to SLASCII standards, which was based on "typewriter metaphor", then replaced by the one based on the "phonetic model". This process, however, was realized through intensive interaction between international standard developing bodies and Sinhala experts.

According to the R.M.K. Sinha (Sinha, 1992), some of the primary considerations in the design of code for internal representation are one and one code for semantically equivalent characters, uniqueness of cording, uniformity in assigning and usage of control characters. Based on the discussions had in earlier chapters, the guidelines and principles are envisaged in the design of the codes for information interchange for Sinhala scripts.

Further, consideration for the internal and external memory economy, acceptable and easiness of sorting and collating sequence, easiness of rendering, the direct and unambiguous transformation of keyboard symbols to the internal representation, ease of editing, and the ease of composing the complex script on the output devices were the primary areas considered for developing the standards for the Sinhala. The scope of the standardization is not limited to the character code itself, but also the aesthetic appearances of the shapes and size of the scripts.

Standard character sets of Asian languages contained vowels, consonants, diacritical marks, numerals, punctuations, ligatures, and sometimes also contained special symbols. One of such language used in Sri Lanka, the Sinhala has been recognized under UCS/Unicode standard, and the range from U+0D80 to U+0DFF is allocated. This code page comprises codes for the diacritical marks, vowels, vowel signs, consonants, special symbols, punctuation marks, and numerals.

In the modern Information age, the exchange of information can happen if we can communicate effectively in any language and script of the world. This, in turn, demands easy entry of **linguistic information into the computers and** natural way of **communicating with each other** through devices. The first language coding schemes of Sri Lanka (Nandasara, 2009) classified into the following two major categories:

- **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. However, the preparation of multilingual text is 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) (Samaranayake V. K., Disanayaka J. B. and Nandasara S. T., 1989) are the examples for this category. 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/IEC-8859-1similar code table based on the keyboard's character set.
- **Multi-lingual user category:** Here, an attempt is made to exploit standard features of the language/scripts, and the unique needs are dealt with separately. The preparation of the multilingual document, multilingual dictionaries, and transliteration from one language to another are essential aspects for consideration. SLS 1134:1996 phonetic model design for the Sinhala character code (SLS 1134:1996, 1996) replaced the typewriter metaphor concept from the previous SLASCII standard (Nandasara, 1991). It could also be made universally applicable to all languages (UCS/Unicode or International Standards Organization (ISO) 10646 come under the universal class).

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

5.2 Evaluation of Frequency of Occurrence of Diacritical Marks, Vowels, Consonant, and "Medial" Signs

At the time of the development of an 8-bit coded character set for SLASCII, the occurrence of the diacritical marks, vowels, consonants, and "medial" signs considered when storage requirements and the code point limitation as the ASCII pages are concerned. At this point, it is worth to raise the frequency of occurrence of groups of symbols in Sinhala and that data is given in Table 13 below based on the UCS/Unicode compatible word list consists of 70142 Sinhala words extracted from Sinhala corpus beta version developed by LTRL project of the University of Colombo School of Computing (UCSC) in April 2005. Sinhala characters in Figure 30 have divided into four groups, and they consist of two (02) diacritical marks (D_n), eighteen (18) vowels (V_n), forty-one (41) consonants (C_n), and three medial signs

(<u>C</u>$_n$). Vowels (V$_n$) will appear only at the beginning of the words. The vowel signs (<u>V</u>$_n$) 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 13, the occurrences of the vowels V$_{11}$ (ඏ) and V$_{12}$ (ඐ) are zero per cent in modern usage and are not found in any words or any dictionaries, but allow the vowel sign <u>V</u>$_{11}$ (◌ෟ) and <u>V</u>$_{12}$ (◌ෟ) to appear itself. The vowel V$_{10}$ (ඍaa) also does not occur in modern usage, but its corresponding vowel sign <u>V</u>$_{10}$ (◌aa) used. The occurrence of the consonant C$_{12}$ (ඦ) is also zero per cent in modern usage and not found in any dictionary (see Table 13). According to the results in Table 13, the usage of consonants C$_5$ (ඞ) and C$_{10}$ (ඣ) also occurred zero per cent in the present writing system, however, the consonant C$_{10}$ (ඣ) used to write *Pali* and *Sanskrit* in the language. The use of medial sign C$_3$ also not occurred in modern writing; however, it used only in contemporary writing. Medial signs <u>C</u>$_1$ and <u>C</u>$_2$ are essential though they are not a part of the Sinhala alphabet and used to write wherever they are needed.

The other important consideration of the Sinhala language concerned, Sinhala alphabet differed from all other Indo-Aryan languages and contained the distinctive sound that has been unique to itself since the 8th century A.D. 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 (Disanayaka, 1991).

Table 13: **Detailed analysis of occurrences of Sinhala characters in modern writing**

No.	**Characters in Alphabetical Order**				**Vowels Occurrences***					**Total**	**%**
	Diacritical Marks (D)				*Used to write Sinhala Words*						
1	ṁ	D_1	◌ං	2,469	1	a	V_1	අ	4,698		
2	ḥ	D_2	◌ඃ	7	2	ā	V_2	ආ	1,138		
	Vowels				3	æ	V_3	ඇ	1295		
3	a	V_1	අ	4,698	4	ǣ	V_4	ඈ	41		
4	ā	V_2	ආ	1,138	5	i	V_5	ඉ	1,194		
5	æ	V_3	ඇ	1,295	6	ī	V_6	ඊ	92		
6	ǣ	V_4	ඈ	41	7	u	V_7	උ	1,219		
7	i	V_5	ඉ	1,194	8	ū	V_8	ඌ	35		
8	ī	V_6	ඊ	92	9	e	V_{13}	එ	1,029		
9	u	V_7	උ	1,219	10	ē	V_{14}	ඒ	190		
10	ū	V_8	ඌ	35	11	o	V_{16}	ඔ	401		
11	ṛ	V_9	ඍ	19	12	ō	V_{17}	ඕ	146	***11,478***	***3.43***
12	ṝ	V_{10}	ඎ	0	*Used to write Sanskrit Words*						
13	ḷ	V_{11}	ඏ	0	1	ṛ	V_9	ඍ	19		
14	ḹ	V_{12}	ඐ	0	2	ṝ	V_{10}	ඎ	0		
15	e	V_{13}	එ	1,029	3	ḷ	V_{11}	ඏ	0		
16	ē	V_{14}	ඒ	190	4	ḹ	V_{12}	ඐ	0		
17	ai	V_{15}	ඓ	12	5	ai	V_{15}	ඓ	12		
18	o	V_{16}	ඔ	401	6	au	V_{18}	ඖ	11	***42***	***0.01***
19	ō	V_{17}	ඕ	146	**Consonants Occurrences**						
20	au	V_{18}	ඖ	11	*Nasals used to write Sinhala Words*						
	Consonants (C)				1	ka	C_1	ක	27,289		
21	ka	C_1	ක	27,289	2	ga	C_3	ග	10,542		
22	kha	C_2	ඛ	313	3	ca	C_7	ච	1,759		
23	ga	C_3	ග	10,542	4	ja	C_9	ජ	2,621		
24	gha	C_4	ඝ	206	5	ṭa	C_{14}	ට	12,553		
25	ṅa	C_5	ඞ	0	6	ḍa	C_{16}	ඩ	3,673		
26	ňga	C_6	ඟ	662	7	ṇa	C_{18}	ණ	4,630		
27	ca	C_7	ච	1,759	8	ta	C_{20}	ත	19,929		
28	cha	C_8	ඡ	142	9	da	C_{22}	ද	14,898		
29	ja	C_9	ජ	2,621	10	na	C_{24}	න	35,634		
30	jha	C_{10}	ඣ	0	11	pa	C_{26}	ප	14,173		
31	ña	C_{11}	ඤ	67	12	ba	C_{28}	බ	5,432		
32	jña	C_{12}	ඥ	268	13	ma	C_{30}	ම	20,848		
33	ňja	C_{13}	ඦ	0	14	ya	C_{32}	ය	27,943		
34	ṭa	C_{14}	ට	12,553	15	ra	C_{33}	ර	25,087		
35	ṭha	C_{15}	ඨ	175	16	la	C_{34}	ල	16,459		
36	ḍa	C_{16}	ඩ	3,673	17	va	C_{35}	ව	29,230		
37	ḍha	C_{17}	ඪ	23	18	śa	C_{36}	ශ	2,498		
38	ṇa	C_{18}	ණ	4,630	19	sa	C_{38}	ස	17,154		
39	ňḍa	C_{19}	ඬ	171	20	ha	C_{39}	හ	8,883		
40	ta	C_{20}	ත	19,929	21	ḷa	C_{40}	ළ	3,107		
41	tha	C_{21}	ථ	786	22	fa	C_{41}	ෆ	415	***304,757***	***91.02***
42	da	C_{22}	ද	14,898	*Half Nasals used to write Sinhala Words*						
43	dha	C_{23}	ධ	2,231	1	ňga	C_6	ඟ	662		
44	na	C_{24}	න	35,634	2	ňja	C_{13}	ඦ	0		
45	ňda	C_{25}	ඳ	1,237	3	ňḍa	C_{19}	ඬ	171		
46	pa	C_{26}	ප	14,173	4	ňda	C_{25}	ඳ	1,237		
47	pha	C_{27}	ඵ	71	5	mba	C_{31}	ඹ	595	***2,665***	***0.80***
48	ba	C_{28}	බ	5,432	*Aspirations used to write Sanskrit and Pali Words*						
49	bha	C_{29}	භ	1,269	1	kha	C_2	ඛ	313		
50	ma	C_{30}	ම	20,848	2	gha	C_4	ඝ	206		
51	mba	C_{31}	ඹ	595	3	cha	C_8	ඡ	142		
52	ya	C_{32}	ය	27,943	4	jha	C_{10}	ඣ	0		
53	ra	C_{33}	ර	25,087	5	ṭha	C_{15}	ඨ	175		
54	la	C_{34}	ල	16,459	6	ḍha	C_{17}	ඪ	23		
55	va	C_{35}	ව	29,230	7	tha	C_{21}	ථ	786		
56	śa	C_{36}	ශ	2,498	8	dha	C_{23}	ධ	2,231		
57	ṣa	C_{37}	ෂ	2,186	9	pha	C_{27}	ඵ	71		
58	sa	C_{38}	ස	17,154	10	bha	C_{29}	භ	1,269	***5,216***	***1.56***
59	ha	C_{39}	හ	8,883	*Other consonants used to write Sanskrit, Pali & Sinhala*						
60	ḷa	C_{40}	ළ	3,107	1	ṅa	C_5	ඞ	0		
61	fa	C_{42}	ෆ	415	2	ña	C_{11}	ඤ	67		
	Non-Vocalic Strokes				3	jña	C_{12}	ඥ	268		
62	yan	C_{yan}	◌්‍ය	1,660	4	ṣa	C_{32}	ෂ	2,186	***2,521***	***0.75***
63	rep	C_{rep}	ර්‍	0	**Diacritical Marks Occurrences**						
64	rak	C_{rak}	◌්‍ර	3,999	1	ṅ	D_1	◌ං	2,469		
			Total	**334,814**	2	ḥ	D_2	◌ඃ	7	***2,476***	***0.74***
Vowels will appear only at the beginning of the words. Otherwise, diacritics are used to change the inherent vowel. Diacritics are ignored and not counted as vowels in this analysis.					**Non-vocalic Strokes Occurrences**						
					1	yan	C_{yan}	◌්‍ය	1,660		
					2	rep	C_{rep}	ර්‍	0		
					3	rak	C_{rak}	◌්‍ර	3,999	***5,659***	***1.69***
									Total	**334,814**	**100.00**

5.3 Early Contributions and SLS 1134:1996 Proposals

The interested of IBM for Sinhala character set and the existence of a draft code for Sinhala (shown in Figure 58) proposed to the ISO/IEC 10646, Working Group, was first brought to the notice in the late eighties when an IBM delegation visited the Institute of Computer Technology (ICT) of the University of Colombo.

The year Later, IBM again showed the Unicode code table for Sinhala (see Figure 59) that published in the first publication, Unicode 1.0. in 1989 with some modifications to the repertories shown in Figure 60.

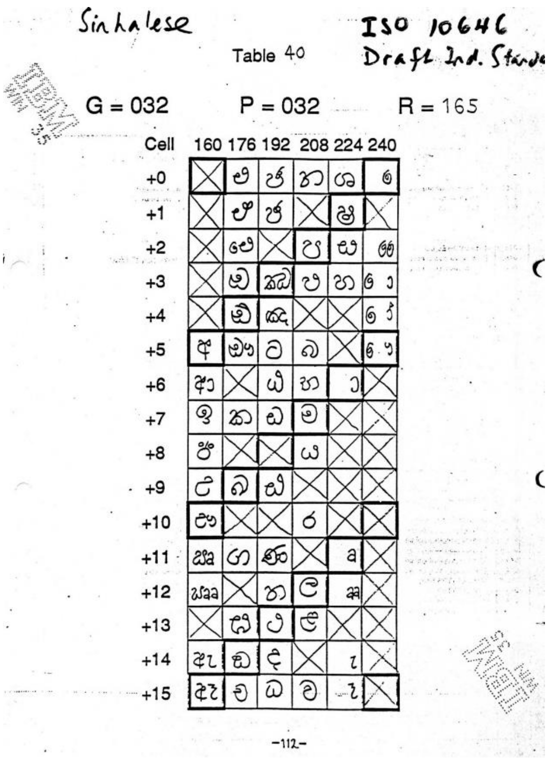

Sinhalese
ISO 10646
Draft Ind. Stand
Table 40
G = 032
P = 032
R = 165
Cell
160 176 192 208 224 240
+0
+1
+2
+3
+4
+5
+6
+7
+8
+9
+10
+11
+12
+13
+14
+15
-112-

Figure 58: Sinhala code page for ISO/IEC 10646 as an expert contribution from IBM (1987)

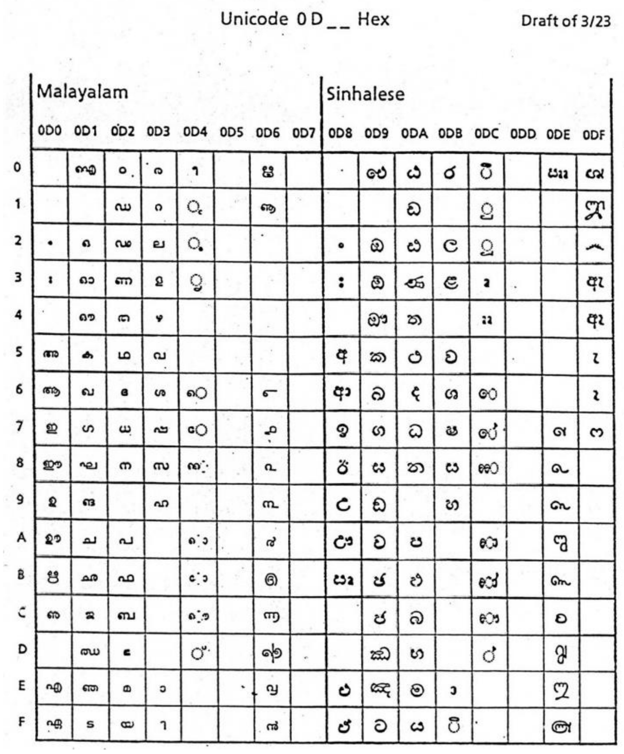
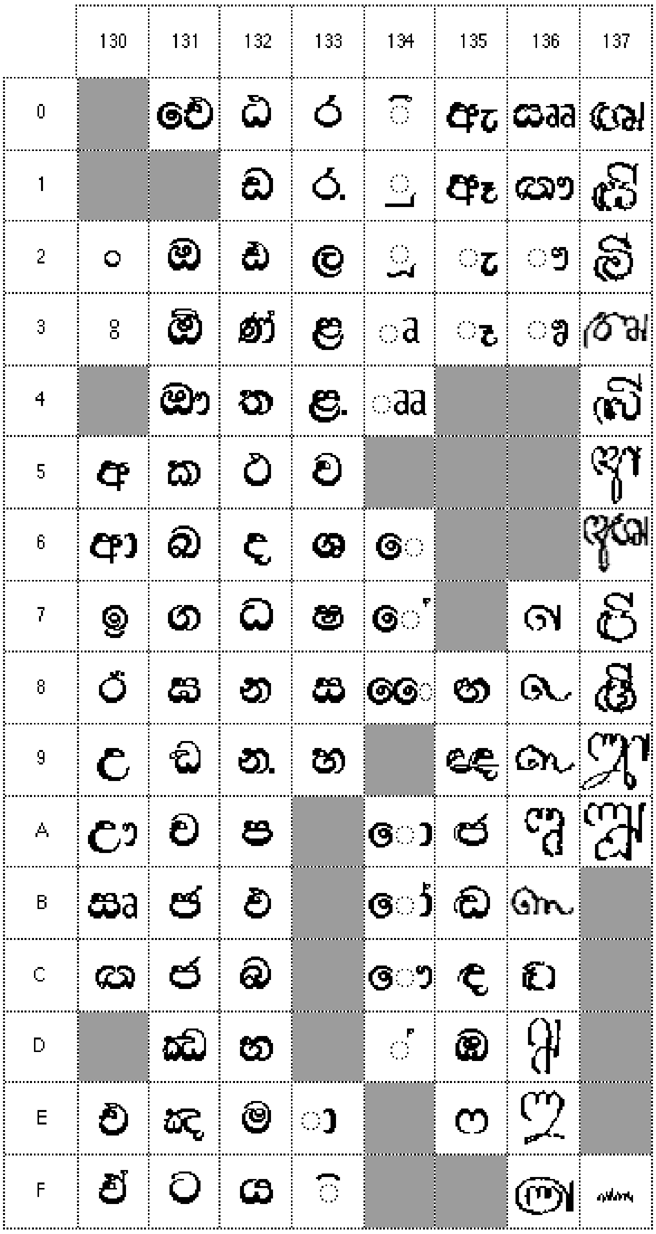

Unicode 0D__ Hex
Draft of 3/23
Malayalam
Sinhalese
0D0 0D1 0D2 0D3 0D4 0D5 0D6 0D7 0D8 0D9 0DA 0DB 0DC 0DD 0DE 0DF
0
1
2
3
4
5
6
7
8
9
A
B
C
D
E
F

Figure 59: Early version of Unicode Sinhala code page

Formulation of the Sinhala Unicode Standard sworn in Figure 59 was based upon the proposal submitted by the early contributor from Ireland (Michael Everson, 1989) and the United Kingdom (Hugh McGregor Ross, 1996). It is interesting to note that neither National bodies nor Sinhala language experts were consulted to prepare this Unicode Sinhala Code Page. Code table shown in Figure 59, represented a distorted Sinhala character set with several glaring errors and omissions. For example, some of the significant shortcomings were the distorted alphabetical order, notable character omission and the inclusion of a non-exiting set of characters and some symbols to represent numerals (Michael Everson, 1999) based on an early grammar book (Gunasekara, 1891). Moreover, it noticed this code page very much biased on *Devanagari* (Indic) repertories.

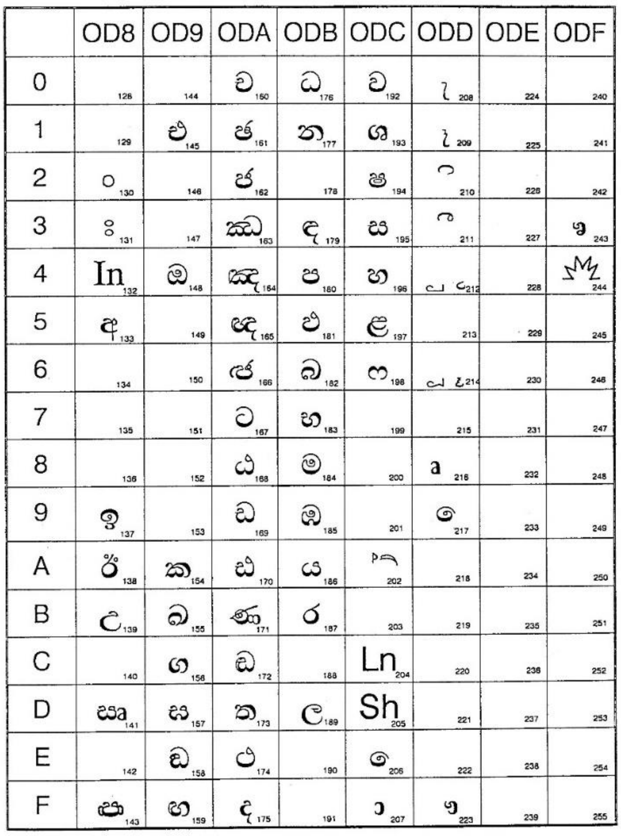
Immediate steps are taken to request subcommittee 2 (SC2) of the joint technical committee 1 (JTC1) of the ISO/IEC and Unicode Technical Committee (UTC), directly through the Sri Lanka Standards Institute (SLSI) to suspend approval of the draft code page until the CINTEC and SLSI made representations. The request was accepted by the ISO/IEC/JTC/SC2, and the UTC and accordingly Sinhala code page had been removed from their next edited version, Unicode 2.0, released on July 1996. Because of the urgent situation, the work regarding the Sinhala standard code process was after that speeded up.



	130	131	132	133	134	135	136	137
0		ඓ	ඨ	ර	ී	ඇ	ඎ	[illegible]
1			ඩ	ර.	ු	ඈ	ඐ	[illegible]
2	ං	ඔ	ඪ	ල	ූ	ැ	ෟ	[illegible]
3	ඃ	ඕ	ණ	ළ	ෘ	ෑ	ෳ	[illegible]
4		ඖ	ත	ළ.	ෲ			[illegible]
5	අ	ක	ථ	ව				[illegible]
6	ආ	ඛ	ද	ශ	ෙ			[illegible]
7	ඉ	ග	ධ	ෂ	ේ		[illegible]	[illegible]
8	ඊ	ඝ	න	ස	ෛ	ඟ	[illegible]	[illegible]
9	උ	ඞ	න.	හ		ඥ	[illegible]	[illegible]
A	ඌ	ච	ප		ො	ඦ	[illegible]	[illegible]
B	ඍ	ඡ	ඵ		ෝ	ඬ	[illegible]	
C	ඏ	ජ	බ		ෞ	ඳ	[illegible]	
D		ඣ	භ		්	ඹ	[illegible]	
E	එ	ඤ	ම	ා		ෆ	[illegible]	
F	ඒ	ට	ය	ි			[illegible]	[illegible]

Figure 60: Proposed Sinhala code page for ISO/IEC 10646 as an expert contribution from Ireland (1989, 1996)

The 8-bit SLASCII standard was submitted to the working group (WG) formed by the Sri Lanka Standard Institute on the recommendation of CNTEC, and comments on the SLASCII standards then received from the members of the WG of the SLSI. At the same time, a SLASCII based Sinhala Standard was formulated by the CINTEC (Figure 42) forwarded to the ISO/IEC/JTC1/SC2 is a draft version for consideration (Samaranayake, Nandasara, 1990) and after that by the SLSI Committee speed up the work for UCS/Unicode standards. As the case with any Sri Lanka Standard, public comments were obtained and amalgamate to the document. It took a few years to conclude the final version of Sinhala Standard for ISO/IEC. The SLS 1134:1996 Sinhala standard had prepared to fall in line with the requirements laid down in ISO/IEC 10646 and maintaining the logical sequence of the alphabet. An effort had also been made to preserve the alphabetical order (not the phonetic sorting order) of the Sinhala character set to a greater extent, which has been marinating for centuries. Finally, the Sri Lanka Standard Sinhala Character Code for Information Interchange (SLS 1134:1996; Samaranayake et al., 1996) approved by the Sectoral Committee on Information Technology[26] and was authorized for the adoption and publication as a Sri Lanka Standard by the Council of the Sri Lanka Standards Institute on 19th September 1996. The proposal then sent to the ISO/IEC/JTC1/SC2 is a Sri Lanka proposal for the UCS/Unicode Standard (Figure 61). Two members of the CINTEC committee were included in the ISO/IEC/JTC1/SC2/WG2, and much correspondence followed after that.



	OD8	OD9	ODA	ODB	ODC	ODD	ODE	ODF
0	128	144	ච 160	ධ 176	ව 192	ැ 208	224	240
1	129	එ 145	ඡ 161	න 177	ශ 193	ෑ 209	225	241
2	ං 130	146	ජ 162	178	ෂ 194	ි 210	226	242
3	ඃ 131	147	ඣ 163	ඳ 179	ස 195	ී 211	227	ෳ 243
4	In 132	ඔ 148	ඤ 164	ප 180	හ 196	ු ු 212	228	෴ 244
5	අ 133	149	ඥ 165	ඵ 181	ළ 197	213	229	245
6	134	150	ඦ 166	බ 182	ෆ 198	ූ ූ 214	230	246
7	135	151	ට 167	භ 183	199	215	231	247
8	136	152	ඨ 168	ම 184	200	ෘ 216	232	248
9	ඉ 137	153	ඩ 169	ඹ 185	201	ෙ 217	233	249
A	ඊ 138	ක 154	ඪ 170	ය 186	් 202	218	234	250
B	උ 139	ඛ 155	ණ 171	ර 187	203	219	235	251
C	140	ග 156	ඬ 172	188	Ln 204	220	236	252
D	ඍ 141	ඝ 157	ත 173	ල 189	Sh 205	221	237	253
E	142	ඞ 158	ථ 174	190	ෙ 206	222	238	254
F	ඏ 143	ඟ 159	ද 175	191	ා 207	ෟ 223	239	255

NOTE : *In - Invisible* *Ln - Link* *Sh - Short*

Figure 61: Early UCS/Unicode proposal for standard code for information interchange in Sinhala (ISO-IEC JTC1/SCL/WG2 N673, October 1996)

[26] The Sectoral Committee on Information Technology was established in 1993 by SLSI and Mr. Rohan Wijeratne is the continuing Chairman for this committee since then.

The SLS 1134:1996 featured to handle ligatures used in Sinhala writing. For this reason, the Link key or link code is used to combine two coded characters to generate a link or joint ligature i.e., ක්‍ව, ත්‍ව and න්‍ද. The short key or shortcode is used to create *repaya* (ර්‍) and other conjunct formations. The invisible key or invisible code is used to delete a particular character or part thereof. The SLS 1134:1996 standard compatible 16-bit application called “*Sarasavi*” developed by the author displayed the mechanism of handling conjunct ligatures using above mentioned these control codes as given in Table 14.

Table 14: Special control codes used with SLS 1134:1996 to display complex ligature used in Sanskrit and Pali

ද	◌් (*al-lakuna*)	ධ	=	ද්ධ	Modern
ද	Link key/Link code	ධ	=	ද්‍ධ	Pali
ක	Link key/Link code	ව	=	ක්‍ව	Sanskrit
ත	Link key/Link code	ව	=	ත්‍ව	Sanskrit
න	Link key/Link code	ද	=	න්‍ද	Sanskrit
ධ	Short key/Short code	ර් ම	=	ධර්ම	Sanskrit
○ ්	Invisible key/ Inv. code		=	්	Modern

This Sinhala standard proposal first discussed at the Singapore ISO/IEC JTC 1/SC 2/WG 2 Meeting #32 on 1997-04-01 (Mike Ksar, Umamaheswaran, 1997) and was recommended to examine in detail with Sri Lankan national body representatives in the upcoming meeting in June 1997 in Greece.

5.4 The UCS/Unicode Standard

Funding source from CINTEC with the assistance of NARESA, two national representatives[27] attended to the ISO/IEC JTC 1/SC 2/WG 2 Meeting #33 held in Crete, Greece on June 1997 (Mike Ksar, Umamaheswaran V.S., 1997), where the Draft Sinhala Code was discussed intensively. National delegates argued for the draft submitted by Sri Lanka opposing several competing proposals from the Unicode Inc. (Andy Daniels, 1992), USA (Lloyd et al. 1992), Ireland (Michael, 1996, Michael Everson, 1997a, Michael Everson, 1997b) and the UK (Hugh McGregor Ross, 1996). The Unicode Inc. (Andy Daniels, 1992) proposal somewhat similar to the Sri Lankan proposal had vowels and consonants, were arranged according to the alphabetical order (Samaranayake and Nandasara. 1990 & SLS 1134:1996, 1996) while other proposals arrangement based on *Devanagari* repertories. After few ad-hoc committee meetings with National delegates and other nominated country delegates concluded to accept (Figure 62) the repertoire, names, and arrangement for Sinhala script (Michael, et al. #1613, 1997), based on Sri Lankan proposal with slight modification with the support of the majority delegates from Canada, Greece, United Kindom, United State of America and Japan (Takayuki, 1998; NAITO, 1998).

In the ad-hoc committee meetings, each proposal examined in terms of the specified repertoire, names, arrangement and encoding principles. The result of this examination and the resolution of differences described in the following.

[27] S. T. Nandasara and J. B. Disanayaka were the two members attended for WG2 #33 from Sri Lanka.

There was no difference between the proposals regarding the primary letters commonly employed in modern Sinhala orthography. The difference arises regarding; (1) ten independent vowel letters and six vowel signs employed in transcribing Sinhala and Pali and in transliterating Sanskrit and Pali text written in other scripts; (2) three extended letters used when writing Tamil language proposed by M. Everson (JTC1/SC2/WG2 N1473); (3) nine digits and eleven number signs employed in earlier (non-modern) Sinhala orthographies proposed by M. Everson (JTC1/SC2/WG2 N1473); (4) three special control characters employed by SLS 1134:1996 for encoding conjuncts and other special orthographic features.

After having lengthy discussions among the members of the committee, these differences resolved as follows; (a) accept ten vowel letters and six vowel signs for Sanskrit and Pali usage; (b) postpone inclusion of three letters for Tamil usage pending further study; (c) postpone inclusion of nine digits and eleven number signs pending further study (d) postpone inclusion of three control characters pending further study of the applicability of existing mechanisms (namely implicit conjuncts, ZWJ, ZWNJ, ZWNBSP).

The differences among the proposals regarding character naming are that one based on native Sinhala character names (as employed by SLS 1134:1996) and the other is based on harmonization with names employed with the Indian scripts. Also, different spelling conventions for native Sinhala names were employed by the different proposals. These differences were resolved by determining to use native Sinhala names with a normalized spelling convention.

The difference among the proposals regarding character arrangement is that one is based on native Sinhal ordering (as employed by SLS 1134:1996) and the other is based on harmonization with the arrangement employed with Indian scripts. These differences were resolved by determining to use native Sinhala ordering in the arrangement found in SLS 114:1996. The use of this arrangement more closely follows actual practice and facilitates better interoperation with the Sri Lanka character set standard.

The encoding of conjuncts, half forms, and other special orthographic features is accommodated in SLS 1134:1996 by use of three special control characters: LINK, SHORT, and INV. However, these features are encoded in ISO/IEC 10646 by means of particular conventions using VIRAMA (*al-lakuna*) to perform implicit conjunct formation, using ZERO-WITH JOINER to effect half consonant formation, using ZERO WIDTH NON-JOINER to effect explicit VIRAMA (*al-lakuna*), and using ZERO WIDTH NO-BREAK SPACE to effect a consonant filter function (i.e. the function served by INV in SLS 1134:1996). Because these functions already are served by these conventions, the three control characters found in SLS 1134:1996 are not thought to be needed. However, some further study is requested by Sri Lanka to determine the adequacy of these conventions. Consequently, the inclusion of the three SLS 1134:1996 controls postponed pending completion of the study.

After considering the above amendment document was forwarded to the WG-2 for adoption and processing at a Preliminary Draft Amendment (pDAM) stage. The pDAM ratified[28] at the WG-2 Meeting #34 on 16th March 1998 held at Redmond, Seattle, USA (Mike Ksar, Umamaheswaran V.S., 1998). The Final Preliminary Draft Amendment (FpDAM) submitted to the meeting #35 held in London, UK (Mike Ksar, Umamaheswaran V.S. 1998) and the final revised Sinhala Code Chart (Revised text of ISO/IEC 10646-1/FPDAM 21, 1998) included in the Unicode Version 3.0 (Nandasara, 2000). The SLS 1134:1996 was also accordingly modified.

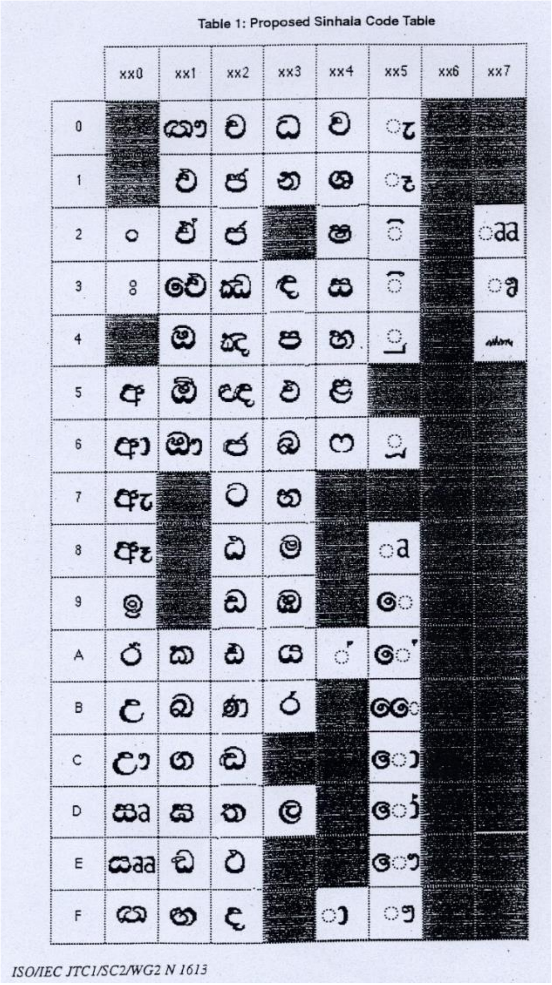


Table 1: Proposed Sinhala Code Table

	xx0	xx1	xx2	xx3	xx4	xx5	xx6	xx7
0		ඐ	ච	ධ	ව	◌ැ		
1		එ	ඡ	න	ශ	◌ෑ		
2	ං	ඒ	ජ		ෂ	◌ි		◌ෲ
3	ඃ	ඓ	ඣ	ඳ	ස	◌ී		◌ෳ
4		ඔ	ඤ	ප	හ	◌ු		෴
5	අ	ඕ	ඥ	ඵ	ළ			
6	ආ	ඖ	ඦ	බ	ෆ	◌ූ		
7	ඇ		ට	භ				
8	ඈ		ඨ	ම		◌ෘ		
9	ඉ		ඩ	ඹ		ෙ◌		
A	ඊ	ක	ඪ	ය	◌්	ෙ◌්		
B	උ	ඛ	ණ	ර		ෛ◌		
C	ඌ	ග	ඬ			ෙ◌ා		
D	ඍ	ඝ	ත	ල		ෙ◌ා්		
E	ඎ	ඞ	ථ			ෙ◌ෟ		
F	ඏ	ඟ	ද		◌ා	◌ෟ		

ISO/IEC JTC1/SC2/WG2 N 1613

Figure 62: The Repertoire and arrangement for Sinhala script based on Sri Lankan proposal with slight modification (ISO/IEC Doc. #1613, 1997)

Consequently, the standard Sinhala keyboard layout, which was first standardized in 1996, was modified to some extent in 2001. In 2004 keyboard layout was further modified (SLS 1134:2004), and now it is used as the national standard keyboard layout for Sinhala.

5.5 Rendering Sinhala Characters in UCS/Unicode

Sinhala characters can combine or change the shape depending on their context, as mentioned earlier, mainly used to write words borrowed from Sanskrit and Pali. These combined shapes can cause the

[28] S. T. Nandasara was attended for WG2 #34 at Seattle, USA.

appearance of Sinhala characters to differ from their nominal glyphs (used in the code charts). Additionally, a few Sinhala characters cause a change in the order of the displayed characters (Figure 63).

ක◌් ◌ෙෂ ◌් න◌් ද ◌් ර	=	ක්ෂේන්ද්ර	-	ක්‍ෂේන්‍ද්‍ර
බ + ◌ු + ද + ◌් + ධ	=	බුද්ධ	බුද්‍ධ	බුද්‍ධ
න + න + ◌් + ද	=	නන්ද	-	නන්‍ද

Figure 63: Illustration of some of the Sinhala consonants and vowel sign combination form different glyphs in Sinhala, Pali & Sanskrit respectively

5.5.1 Sinhala al-lakuna (Virama)

Sinhala employs a sign U+0DCA SINHALA AL-LAKUNA, known in Sanskrit as the *virama* or vowel omission sign. The *al-lakuna* sign nominally serves to cancel (or kill) the inherent vowel /a/ of the consonant to which it is applied, i.e. ka$_n$ + *al-lakuna*$_n$ → ka$_d$. When a consonant has lost its inherent vowel by the application of *al-lakuna*, it is known as a dead consonant, i.e. ka$_d$.

In the USC/Unicode Standard, a dead consonant is defined as a sequence consisting of a consonant letter followed by a *virama* or vowel omission sign.

ka$_n$ + *al-lakuna$_n$* → ka$_d$	ක + ◌් → ක්
ka$_n$ + *al-lakuna*$_n$ + ssa$_n$→ k.ssa$_n$	ක + ◌් + ෂ → ක්‍ෂ

5.5.2 Yansaya, Rakaransaya and Rapay

A frequent criticism of Unicode is the lack of codes for the *yansaya*, *rakaransaya* and *repaya*. The Unicode document considered representing these three special symbols using ZWJ, since these symbols are not Sinhala letters. Instead, they are abbreviations for the letters *ya*$_n$ and *ra*$_n$ when they follow a dead consonant. For example, sa$_n$ + ta$_d$ + ya$_n$ → sa$_n$ + t.ya$_{yan}$ (සත්‍ය) is the traditional way of writing සත්ය, ma$_n$ + i$_{vs}$ + ta$_n$ + ra$_n$ → ma$_n$ + i$_{vs}$ + t.ra$_{rak}$ (මිත්‍ර) represents the sequence මිත්ර, and ka$_n$ + ra$_d$ + ma$_n$ → ka$_n$ + r.ma$_{rep}$ (කර්‍ම) represents the sequence කර්ම.

5.5.3 Consonant Conjuncts

The Sinhala scripts are noted for a large number of consonant conjunct forms that serve as ligatures of two or more adjacent letterforms. These ligatures are standard in Sanskrit and Pali writing which is more common in Sri Lanka. An orthographic consonant cluster is defined as a sequence of characters that represents one or more dead consonants (denoted C$_d$) followed by a regular, live consonant letter (denoted C$_l$). Under pause circumstances, a consonant cluster is depicted with a conjunct glyph if such a glyph is available in the current font. In the absence of a conjunct glyph, the one or more dead consonants that form part of the cluster are depicted using half-form glyphs (denoted by C$_h$). In the absence of half-form glyphs, the dead consonants are depicted using the simple consonant forms combined with visible al-lakuna signs. Under normal circumstances, simple consonant forms will not combine with al-lakuna signs (see Figure 63). Simple glyph form of consonant C appears in the code table is denoted by C$_{n}$. The *rakaransaya* (ra$_{rak}$) and *yansaya* (ya$_{yan}$) are forms of conjunct letters. The ra$_{rak}$ represents a nominal ra (ra$_{n}$), which follows a dead consonant. In turn, it can be followed by a vowel sign. Similarly, the ya$_{yan}$ represent a nominal ya (ya$_{n}$), which follows a dead consonant.

A number of types of conjunct formations appear in Figure 64: (1) a half-forms of na, ta and ka (nah, tah, kah) in its combination with the full forms of dha, tha and ssa (n.dha$_{n}$, t.tha$_{n}$, k.ssa$_{n}$); (2) a touch conjunct d.dha; and (3) a fully ligated conjunct n.da with ra$_{rak}$, in which the components ra$_{l}$ are no longer distinct (n.d.ra$_{rak}$). In example (4) in Figure 64, the dead consonant ya$_{d}$ followed by ya$_{l}$ is formed to ya$_{yan}$ ligature, and in (5) dead consonant ya$_{d}$ converts the following ya$_{l}$ to ya$_{yan}$ and place the non-spacing combining mark ra$_{rep}$.

1	na$_{d}$ + dha$_{l}$ → na$_{h}$ + dha$_{n}$
	න් + ධ → න්‍ධ
	ta$_{d}$ + tha$_{l}$ → ta$_{h}$ + tha$_{n}$
	ත් + ථ → ත්‍ථ
	ka$_{d}$ + ssa$_{l}$ → ka$_{h}$ + ssa$_{n}$
	ක් + ෂ → ක්‍ෂ
2	da$_{d}$ + dha$_{l}$ → da$_{t}$ + dha$_{n}$
	ද් + ධ → ද්‍ධ
3	na$_{d}$ + da$_{d}$ + ra$_{l}$ → na$_{h}$ + da$_{h}$ + ra$_{rak}$
	න් + ද් + ර → න්‍ද්‍ර
4	ya$_{d}$ + ya$_{l}$ → ya$_{n}$ + ya$_{yan}$
	ය් + ය → ය්‍ය
5	ya$_{d}$ + ya$_{l}$ + ra$_{d}$→ ra$_{rep}$ + ya$_{yan}$ + ya$_{n}$
	ය් + ය + ර් → ර්‍ය්‍ය

Figure 64: Consonant conjuncts in Sinhala

5.5.4 Function of al-lakuna in Sinhala Writing

Usually, an *al-lakuna* sign serves to create dead consonants that are, in turn, visibly rendered and did not combine with subsequent consonants to form conjuncts like most of the Indian languages. This behavior usually results in an *al-lakuna* sign is always depicted as appropriate for the consonant and is being shown visually. Occasionally, this default behavior is not desired when a dead consonant should be rendered as conjunct formation, in which case the *la-lakuna* sign is not visibly rendered, especially, when writing Sanskrit and Pali words in Sinhala text. To incorporate this behavior, the UCS/Unicode Standard adopts the convention of placing the character U+200D ZWJ (ZERO WIDTH JOINER) immediately after the encoded dead consonant that is to render as a conjunct form. In this case, the *al-lakuna* signs not being depicted visually.

ka_n + *al-lakuna*$_n$ + ssa_l → ka_d + ssa_n	ක + ් + ෂ → ක්ෂ (1) Preventing the conjunct form in Sinhala
ka_n + *al-lakuna*$_n$ + ZWJ + ssa_l → ka_h + ssa_n	ක + ් + ZWJ + ෂ → ක්‍ෂ (2) Use of ZWJ for Half consonant in Sinhala
da_n + *al-lakuna*$_n$ + ZWJ + dha_l → da_h + dha_n	ද + ් + ZWJ + ධ → ද්‍ධ (3) Use of ZWJ to form a new ligature

In general, as mentioned above, in the Sinhala writing system, *al-lakuna* (or virama) is being used for everyday writing, except for scholarly or Sanskrit writing where it is appropriate. However, in traditional Pali writing, use of al-lakuna is omitted. Inserted either it is being combined and forms a combined conjunct or is being attached to the preceding consonants. Since the Sanskrit is widely being used in Sri Lanka for centuries, use of combined conjunct formation is the general practice for Pali writing too. For example, writing a combined conjunct form of ද්‍ධ with ZWJ rule as used in (3) above can not be used to formulate ද්ධ in (5) below.

5.6 Rules for Rendering Sinhala Characters

Vowels

Each vowel is represented by one character in the range U+0D85 – U+0D96, i.e. අ = U+0D85 or ආ = U+0D86. Note that a vowel such as ආ should not be represented as a character sequence such as U+0D9A U+0DCF.

Consonants

1. When consonant C_n that does not precedes *al-lakuna*$_n$ it is considered to be a live consonant C_l.

ka_n → ka_l	ක → ක

2. When consonant C_n precedes *al-lakuna*$_n$ it is considered to be a dead consonant C_d.

ka_n + *al-lakuna*$_n$ → ka_d	ක + ් → ක්

3. When consonant C_n precedes with *al-lakuna*$_n$ and ZWJ, it is considered to be a half consonant C_h.

ka_n + *al-lakuna*$_n$ + ZWJ → ka_h	ක + ් + ZWJ → ක්‍

***Consonant ra* (ර)**

4. The character 0DBB SINHALA LETTER RAYANNA takes one of the numbers of forms depending on its context in a consonant cluster. By default, the letter is depicted with its simple glyph form (as shown in the code charts). In some content, it is depicted using one of two nonspacing glyph form that combines with a base letterform.

ka$_n$ + ra$_n$ + *al-lakuna$_n$* + ma$_n$ → ka$_n$ + ra$_d$ + ma$_n$ (*karma*) ක + ර + ් + ම → ක + ර් + ම (කර්ම)
ka$_n$ + ra$_n$ + *al-lakuna*$_n$ + ZWJ + ma$_n$ → ka$_n$ + ra$_{rep}$ + ma$_n$ (*karma*) ක + ර + ් + ZWJ + ම → ක + ර්‍ + ම (කර්‍ම)
ka$_n$ + *al-lakuna*$_n$ + ZWJ + ra$_n$ + ma$_n$ → ka$_n$ + ra$_{rak}$ + ma$_n$ ක + ් + ZWJ + ර + ම → ක + ්‍ර + ම (ක්‍රම)

5. If the superscripted mark ra$_{rep}$ is to be applied to a dead consonant and that dead consonant is combined with another consonant to form a conjunct ligature, then the ra$_{rep}$ mark is positioned so that it appears to the conjunct ligature form as a whole.

a$_v$ + ra$_n$ + *al-lakuna*$_n$ + ZWJ + ta$_n$ + *al-lakuna*$_n$ + ZWJ + tha$_n$ → a$_v$ + ta$_h$ + tha$_n$ + ra$_{rep}$ අ + ර + ් + ZWJ + ත + ් + ZWJ + ථ → අ + ත්‍ + ථ + ර්‍ (අර්‍ත්‍ථ)

6. When a dead consonant (other than ra$_d$) C$_d$ precedes the live consonant ra$_l$, then C$_d$ is replaced with its simple form C$_n$, and ra is replaced by the nonspacing mark ra$_{rak}$, which is positioned so that it is applied to C$_n$.

pa$_n$ + *al-lakuna*$_n$ + ZWJ + ra$_l$ → pa$_n$ + ra$_{rak}$ ප + ් + ZWJ + ර → ප + ්‍ර

7. If the subscripted mark ra$_{rak}$ is to be applied to a dead consonant and that dead consonant is combined with another consonant to form a conjunct ligature, then the ra$_{rak}$ mark is positioned to the conjunct ligature as a whole.

ca$_n$ + na$_n$ + *al-lakuna*$_n$ +ZWJ+ da$_n$ + *al-lakuna*$_n$ +ZWJ+ ra$_n$ → ca$_n$ + na$_h$ + da$_h$ + ra$_{rak}$ ච + න + ් + ZWJ + න + ් + ZWJ + ර → ච + න්‍ + ද්‍ + ්‍ර (චන්‍ද්‍ර)

***Consonant ya* (ය)**

8. The character 0DBA SINHALA LETTER YAYANNA takes one of two numbers of forms depending on its context in a consonant cluster. By default, the letter is depicted with its simple glyph form (as shown in the code charts). In some content, if the dead consonant C$_d$ precedes a nominal consonant ya$_n$, then ya$_n$ is replaced by the *yansaya* (ya$_{yan}$).

ka$_n$ + *al-lakuna$_n$* + ya$_n$ → ka$_d$ + ya$_n$	ක + ් + ය → ක් + ය (ක්ය)
ka$_n$ + *al-lakuna*$_n$ + ZWJ + ya$_n$ → ka$_n$ + ya$_{yan}$	ක + ් + ZWJ + ය → ක + ◌්‍ය (ක්‍ය)

9. If the *al-lakuna*$_n$ precedes with ZERO WIDTH JOINER followed by ya$_n$, the special symbol yansaya (ya$_{yan}$), is used instead of ya$_d$.

al-lakuna$_n$ + ZWJ + ya$_n$ → ya$_{yan}$	් + ZWJ + ය → ◌්‍ය

10. If the subscripted mark ra$_{rak}$ is to be applied to a consonant, which is ya$_n$ followed by ya$_{yan}$, then the ra$_{rak}$ mark is positioned so that it applies to the ya$_{yan}$.

ca$_n$ + ra$_n$ + *al-lakuna*$_n$ +ZWJ+ ya$_n$ + *al-lakuna*$_n$+ZWJ+ ya$_n$ → ca$_n$ + ya$_n$ + ya$_{yan}$ + ra$_{rak}$
ච + ර + ◌් + ZWJ + ය + ◌් + ZWJ + ය ⟶ ච + ය් + ◌්‍ය + ර්‍ (චය්‍ය්‍ර)

In normal circumstances, this can be written in two forms:

either ca$_n$ + ra$_n$ + *al-lakuna*$_n$ + ya$_n$ + *al-lakuna*$_n$ + ya$_n$ → ca$_n$ + ra$_d$ + ya$_d$ + ya$_n$
ච + ර + ◌් + ය + ◌් + ය ⟶ ච + ර + ◌් + ය + ◌් + ය (චර්ය්ය)
or in more general writing, ca$_n$ + ra$_n$ + *al-lakuna*$_n$ + ya$_n$ → ca$_n$ + ra$_d$ + ya$_n$
ච + ර + ◌් + ය ⟶ ච + ර + ◌් + ය (චර්ය)

As such words are generally spelled using the *yansaya* or *rakaransaya*; it was proposed to represent a ya$_n$ or ra$_n$ following a dead consonant using the relevant symbol. However, some words, such as මල්රාජ් (name of a person) and පස්යාල (name of the city), do not use the consonant constructs. The alternatives are:

- encode the default case Sinhala does; i.e. ta$_n$ + al-lakuna + ra$_n$ and did not use the code ZWNJ to indicate when the construct should not be formed like most of the Indic does; i.e. ta$_d$ + ra$_n$
- encode the default case like most of the Indic does; i.e. ta$_d$ + ra$_{rak}$ and use the code ZWNJ to indicate when the construct should not be formed; i.e. ta$_d$ + ZWNJ + ra$_n$
- to use the code ZWJ to indicate when the *yansaya* and *rakaransaya* should be formed; i.e. ta$_d$ + ZWJ + ra$_n$

The second alternative yields a shorter code sequence for the more common case like in Indic, and also follows the Unicode convention that the typical case is encoded without unique code. However, there are not many consonant conjuncts in Sinhala like in Devanagari or Malayalam languages, the first and third alternatives have been selected for two reasons.

- Keying the sequence ta$_n$ + *al-lakuka*$_n$ + ra$_n$ would otherwise have automatically produced a t.ra$_{rak}$ (ත්‍ර), even if not desired by the user. As the recommended keyboard has a unique key for the *yansaya* and *rakaransaya* users would use these keys to generate the ◌්‍ය and ◌්‍ර, the key sequence ta$_d$ + ZWJ + ra$_n$ produced ත්‍ර and ta$_d$ + ZWJ + ya$_n$ produced ත්‍ය.
- Using the ZWJ to produce the yansaya and rakaransaya, which are forms of consonant conjuncts, allows the user to use the same representation for all consonant conjuncts as desired.

However, in some cases, the use of deferent conjoint for the same code sequences, i.e. da$_n$ + *al-lakuna*$_n$ + ZWJ + dha$_l$ → da$_t$ + dha$_n$ / da$_h$ + dha$_n$ for ද්‍ධ and ද්ධ should be considered separately as user request ligature form.

Consonants with Vowel Signs

11. A consonant and vowel sign represented by a two code sequence C$_n$ + V$_{vs}$ where C$_n$ represents a

consonant and V$_{vs}$ represent a vowel sign.

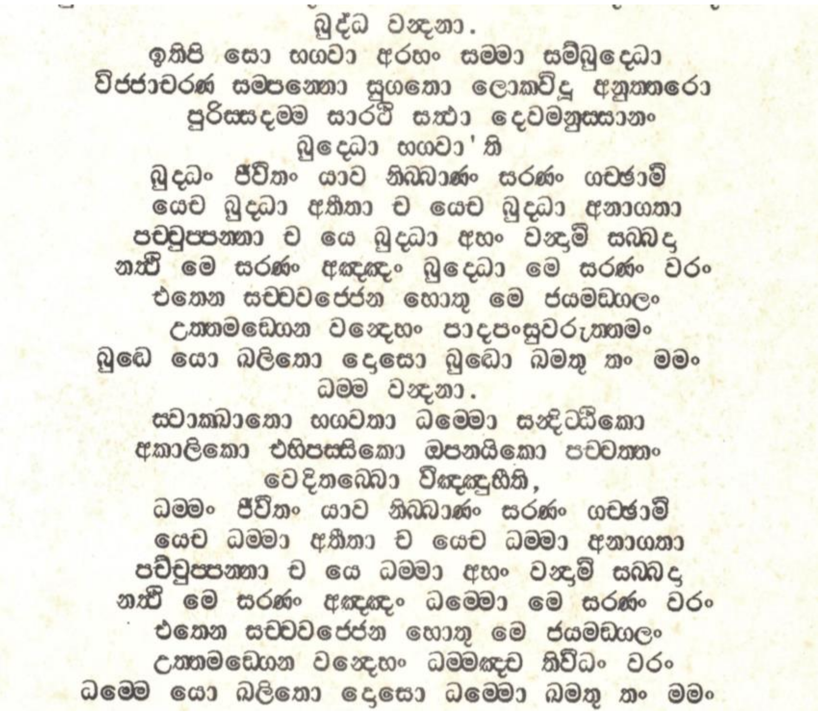
ka$_{n}$ + *al-lakuna*$_{n}$ → ka$_{d}$	ක	+	◌්	→	ක්
ka$_{n}$ + aa$_{vs}$ → kaa$_{n}$	ක	+	◌ා	→	කා
ra$_{n}$ + oo$_{vs}$ → roo$_{n}$	ර	+	◌ෝ	→	රෝ
ra$_{n}$ + uu$_{vs}$ → ruu$_{n}$	ර	+	◌ූ	→	රූ
ka$_{n}$ + *al-lakuna*$_{n}$ + ra$_{n}$ + oo$_{vs}$ → k.roo$_{n}$	ක + ◌් + ර + ◌ෝ			→	ක්‍රෝ

12. *al-lakuna* has two forms, depending on the associated consonant, the same code point represents both.

ka$_{n}$ + *al-lakuna*$_{n}$ → ka$_{d}$	ක	+	◌්	→	ක්
dha$_{n}$ + *al-lakuna*$_{n}$ → dha$_{d}$	ධ	+	◌්	→	ධ්

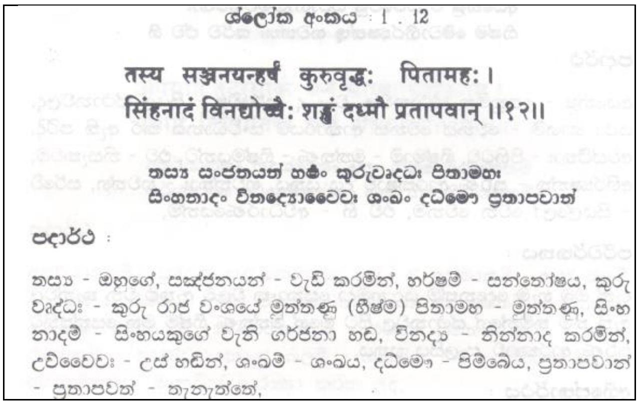
Consonant conjunct in Pali

When Pali words are adopted into Sinhala, they are transcribed in the compound manner in which they are written in Pali. These conjunct letters are written in the Pali language convention, a dead consonant (letter with *al-lakuna*) followed by another consonant can be represented by writing that consonants are touching each other. If the preceding pure consonant has half letter form following consonant can be combined and formed conjunct letters. As a result, the *al-lakuna* sign not being depicted visually. However, in some contemporary Pali writing, use of half consonants or toughing behavior is not the general practice. The special symbols *rakaransaya* and *yansaya* can be written using the conjunct or non-conjunct letters (see Figure 65). In *Pali* writing, it has 68 sets of such conjunct-consonants, including touched consonants (see Figure 27). On the other hand, when letters are written in Sanskrit, dead consonants can be represented as a conjunct form or non-conjunct form, as shown in Figure 66.



බුද්ධ වන්දනා.
ඉතිපි සො භගවා අරහං සම්මා සම්බුද්ධො
විජ්ජාචරණ සම්පන්නො සුගතො ලොකවිදූ අනුත්තරො
පුරිසදම්ම සාරථී සත්ථා දෙවමනුස්සානං
බුද්ධො භගවා'ති
බුද්ධං ජීවිතං යාව නිබ්බාණං සරණං ගච්ඡාමි
යෙච බුද්ධා අතීතා ච යෙච බුද්ධා අනාගතා
පච්චුප්පන්නා ච යෙ බුද්ධා අහං වන්දාමි සබ්බදා
නත්ථි මෙ සරණං අඤ්ඤං බුද්ධො මෙ සරණං වරං
එතෙන සච්චවජ්ජෙන හොතු මෙ ජයමඞ්ගලං
උත්තමඞ්ගෙන වන්දෙහං පාදපංසුවරුත්තමං
බුද්ධෙ යො ඛලිතො දොසො බුද්ධො ඛමතු තං මමං
ධම්ම වන්දනා.
ස්වාක්ඛාතො භගවතා ධම්මො සන්දිට්ඨිකො
අකාලිකො එහිපස්සිකො ඔපනයිකො පච්චත්තං
වෙදිතබ්බො විඤ්ඤූහීති,
ධම්මං ජීවිතං යාව නිබ්බාණං සරණං ගච්ඡාමි
යෙච ධම්මා අතීතා ච යෙච ධම්මා අනාගතා
පච්චුප්පන්නා ච යෙ ධම්මා අහං වන්දාමි සබ්බදා
නත්ථි මෙ සරණං අඤ්ඤං ධම්මො මෙ සරණං වරං
එතෙන සච්චවජ්ජෙන හොතු මෙ ජයමඞ්ගලං
උත්තමඞ්ගෙන වන්දෙහං ධම්මඤ්ච තිවිධං වරං
ධම්මෙ යො ඛලිතො දොසො ධම්මො ඛමතු තං මමං

Figure 65: Sample specimen of Pali text



ශ්ලෝක අංකය : 1 . 12

तस्य सञ्जनयन्हर्षं कुरुवृद्धः पितामहः ।
सिंहनादं विनद्योच्चैः शङ्खं दध्मौ प्रतापवान् ॥१२॥

තස්‍ය සංජනයන් හර්ෂං කුරුවෘද්ධඃ පිතාමහඃ
සිංහනාදං විනද්‍යෝච්චෛඃ ශංඛං දධ්මෞ ප්‍රතාපවාන්

පදාර්ථ :

තස්‍ය - ඔහුගේ, සඤ්ජනයන් - වැඩි කරමින්, හර්ෂම් - සන්තෝෂය, කුරු වෘද්ධඃ - කුරු රාජ වංශයේ මුත්තණු (භීෂ්ම) පිතාමහ - මුත්තණු, සිංහ නාදම් - සිංහයකුගේ වැනි ගර්ජනා හඬ, විනද්‍ය - නින්නාද කරමින්, උච්චෛඃ - උස් හඬින්, ශංඛම් - ශංඛය, දධ්මෞ - පිඹියේය, ප්‍රතාපවාන් - ප්‍රතාපවත් - තැනැත්තේ,

Figure 66: Section of a page from the Bhagavatgeetava written in Devanagari, Sanskrit and Sinhala with translation in Sinhala

13. When conjunct letters are written in Pali language convention, a dead consonant (letter with *al-lakuna*) followed by another consonant can be represented by writing that consonants as touching each other. If the preceding pure consonant has half letter form, the following consonant can be combined and formed as conjunct letters.

sa$_n$ + ZWJ + *al-lakuna*$_n$ + sa$_l$ ⟶ sa$_t$ + sa$_n$
ස් + ස ⟶ ස්‍ස (4) Touching consonant in Pali writing

da$_n$ + ZWJ + *al-lakuna*$_n$ + dha$_l$ ⟶ da$_t$ + dha$_n$
ද් + ධ → ද්‍ධ (5) Touching consonant similar to (4) above

ta$_n$ + *al-lakuna*$_n$ + ZWJ + tha$_l$ ⟶ ta$_h$ + tha$_n$
ත් + ථ → ත්‍ථ (6) Use of half consonant in Pali similar to (2) above

sa$_n$ + *al-lakuna*$_n$ + sa$_l$ ⟶ sa$_n$ + sa$_n$
ස් + ස → ස් + ස (7) Contemporary Pali writing

5.7 Designing Methodology for UCS/Unicode

5.7.1 Designing character codes – Guideline principles

According to the R.M.K. Sinha (Sinha, 1992), some of the primary 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 had in earlier, 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 critical consideration, especially for large information systems. Internal representation needing less space without some overhead of processing time preferred over those who try to reduce processing time at the cost of storage.

5) ***Compatibility* -** The punctuation marks, numerals, and operators & other universal symbols 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) ***Easiness of transliteration* -** Transliteration from Sinhala to Tamil should consider.
8) ***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 consider.
9) ***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.
10) ***Keyboard sequence compatibility* -** As far as possible, there should be the direct and unambiguous transformation of keyboard symbols to the internal representation.

Standard 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 standardized under UCS/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 as discussed earlier, 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 again as follows:

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

<C> := consonants; (41)
<V> := vowels; (18)
<X> := *al-lakuna*/*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 similar graphical shapes for four vowel signs included to the early SLASCII and SLS 1134:1996 and they defined as follows:

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

Table 15: 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	16 bit		8-bit	16 bit		8-bit	16 bit		8-bit	16 bit	
		SLASCII	SLS	UCS	SLASCII	SLS	UCS	SLASCII	SLS	UCS	SLASCII	SLS	UCS
	Hex	8x	0D8x		9x	0D9x		Ax	0DAx		Bx	0DBx	
1	0						ඐ		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	ඍ	ඥ	ඥ	ඨ	ඵ	ඵ
7	6			ආ			ඖ	එ	ඦ	ඦ	ඩ	බ	බ
8	7			ඇ				ඔ	ට	ට	ඪ	භ	භ
9	8			ඈ				ක	ඨ	ඨ	ණ	ම	ම
10	9		ඉ	ඉ				ඛ	ඩ	ඩ	ඬ	ඹ	ඹ
11	A		ඊ	ඊ		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	ඞ	ඬ	ඬ	ද		
14	D		ඍ	ඍ		ඝ	ඝ	ඟ	ත	ත	ධ	ල	ල
15	E			ඎ		ඞ	ඞ	ච	ථ	ථ	න		
16	F		ඏ	ඏ		ඟ	ඟ	ඡ	ද	ද	ඳ		

Table 16: 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 15)

Row No.		8-bit	16 bit		8-bit	16 bit		8-bit	16 bit		8-bit	16 bit	
		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		◌ැ	◌ැ	්			ර්‍		
2	1	ඵ	ශ	ශ		◌ෑ	◌ෑ	ං					
3	2	බ	ෂ	ෂ		◌ි	◌ි	ඃ					◌ෲ
4	3	භ	ස	ස		◌ී	◌ී					◌ෲ	◌ෳ
5	4	ම	හ	හ		◌ු	◌ු	ා				෴	෴
6	5	ඹ	ළ	ළ				ැ					
7	6	ය	ෆ	ෆ		◌ූ	◌ූ	ෑ					
8	7	ර						ි					
9	8	ල				◌ෘ	◌ෘ	ී					
10	9	ව				ෙ◌	ෙ◌	ු			◌ු		
11	A	ශ	◌් ර්‍	◌්			ේ	ූ			◌ූ		
12	B	ෂ					ෛ	ෘ					
13	C	ස	Ln				ො	ෙ					
14	D	හ	Sh				ෝ	ෟ					
15	E	ළ	ෙ◌				ෞ	ෲ					
16	F	ෆ	◌ා	◌ා		◌ෟ	◌ෟ	්‍ය					

Tables 15 and 16 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.

In the next section, we examine and discuss the guideline principle, which mentioned above sections, has led to the formation of standard codes assigned by the SLASCII, SLS 1134:1996 and UCS/Unicode. Table 17 below demonstrates the comparison of the formation of standard codes assigned by the SLASCII, SLS 1134:1996 and UCS/Unicode.

Table 17: Comparison table of the formation of standard codes assigned by the SLASCII, SLS 1134:1996 and UCS/Unicode

Grouping based on UCS/Unicode Code Table		No. of Sinhala Symbols	SLASCII	SLS 1134:1996	UCS/Unicode
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) (◌ෲ and ◌ෳ)	2	0	1	2
Punctuations (P) (෴)		1	0	1	1
Sinhala Numerals (N) included in SLS 1134 : 2011 and published in UCS/Unicode Ver. 7		10	0	0	10
Medial Sign (<u>C</u>)	*rakaransaya* (<u>C₁</u> = X + C₃₃) (◌්‍ර) *yansaya* (<u>C₂</u> = X + C₃₂) (◌්‍ය) *repaya* (<u>C₃</u> = C₃₃ + X) (ර්‍)	3	2 Graphically coded C_1 and C_2	0	0
			0	3	0
	Control Code			In, Ln, Sh	ZWJ ZWNJ
Total code points/Total composed symbols (without Numerals)		83	65/79	68/83	80/80

5.8 Comparison of Guideline Principles in Standardizing of Sinhala Scripts

5.8.1 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 of the extended ASCII area from Hex 80 to 9F 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 seven vowels.
4) Positions A8 through CF are assigned for 40 consonants.
5) Positions E0 is assigned for *virama* (◌්), E1 assigned for *anusvara* (◌ං), and E2 is assigned for *visarga* (◌ඃ) are assigned for diacritical marks. Positions E4 to ED are assigned for ten vowel signs (◌ා ◌ැ ◌ෑ ◌ි ◌ී ◌ු ◌ූ ◌ෘ ෙ◌ ◌ෟ).
6) Two (02) code points EE and EF are used to assigned code points for medial sings $\underline{C}_1$ (◌්‍ය) and $\underline{C}_2$ (◌්‍ර) (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 the 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 + ◌ෟ)
$V_{10} = V_9 + \underline{V}_9$	(ඎ = ඍ + ◌ෘ)
$V_{14} = V_{13} + X$	(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

		Examples
$\underline{V}_{14}$ = $\underline{V}_{13}$+X	(ෙ◌් = ෙ◌+◌්)	කේ = ෙ◌ + ක + ◌්
V_{15}= V_{13}+V_{13}	(ෙෙ◌ = ෙ◌+ෙ◌)	කෛ = ෙ◌ + ෙ◌ + ක
$\underline{V}_{16}$ =$\underline{V}_{13}$+$\underline{V}_{2}$	(ෙ◌ා = ෙ◌+◌ා)	කො = ෙ◌ + ක + ◌ා
$\underline{V}_{17}$=$\underline{V}_{13}$+$\underline{V}_{2}$+X	(ෙ◌ා් = ෙ◌+◌ා+◌්)	කෝ = ෙ◌ + ක + ◌ා + ◌්
$\underline{V}_{18}$=$\underline{V}_{13}$+$\underline{V}_{11}$	(ෙ◌ෟ = ෙ◌+◌ෟ)	කෞ = ෙ◌ + 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} (ඦ) consonants. In a total, 79 symbols were able to use this standard.

NOTES FOR SLS 1134:1996

1) Composite vowel signs 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 assigned for *virama* (◌්), 0D82 (◌ං) and 0D83 (◌ඃ) assigned for diacritical marks. Positions 0DCF to 0DDF assigned for ten 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

5.8.2 Unification - There should be one and only one code for semantically equivalent characters

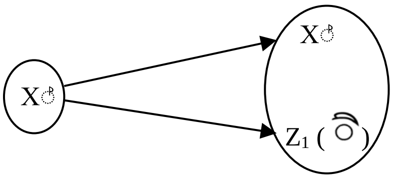
NOTES FOR SLASCII:

C_{11}	C_{13}
nya	*nyja*
ඤ	ඥ

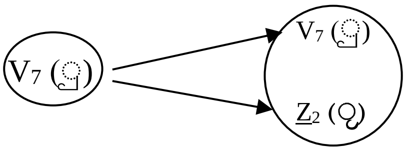
1) The letters C_{11} (ඤ) (*nya*) and C_{13} (ඥ) (*nyja*) are identical in sound only in the initial position of a word, for example, ඤාණ (*nya+aa+nna*) and ඥාන (*nyja+aa+na*). They are not identical in non-initial

positions, where C_{13} (ඥ) behaves like a combination of two consonant sounds, for example, ප්‍රඥා (*p+ra+k+nyja*). Therefore, two codes are assigned for these two consonants.

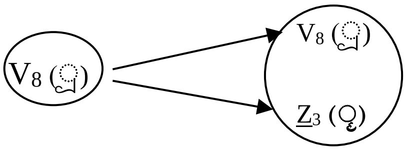
2) The following *virama* sign X (◌්) has assigned two codes, one for X (◌්) and second for Z_1 (◌්), although they are semantically equivalent vowel signs in the writing system.



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.



5.8.3 Uniqueness

No two characters that differ in their meaning be assigned to the same code.

5.8.4 Memory Economy

The required amount of storage is another critical consideration, especially for large information systems. Internal representation needing less space without some overhead of processing time is preferred over those who try to reduce processing time at the cost of storage.

5.8.5 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

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

5.8.6 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* (Code point A0) 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 1134:1996 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 + ්‍ර) (*rakaransaya)*
$C_1\underline{C}_2 = C_1 + \underline{C}_2$	(ක්‍ය = ක + ්‍ය) (*yansaya)*
$\underline{C}_3$ (ර්‍) (*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:

$C_{24}C_{22} = C_{24} + X + Ln + C_{22}$	(න්‍ද=න+්+Ln+ද)
$C_1C_{37} = C_1 + X + Ln + C_{37}$	(ක්‍ෂ=ක+්+Ln+ෂ)
$C_{20}C_{35} = C_{20} + X + Ln + C_{35}$	(ත්‍ව=ත+්+Ln+ව)
$\underline{C}_3 = C_{33} + X + Sh$	(ර්‍ =ර+්+Sh)

In UCS/Unicode

In this version 'Link Key' or 'Link code' (Ln), 'Short Key' or 'Shortcode' (Sh) and 'Invisible Key' or 'Invisible code' (In) are omitted as a function of 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:

$C_{24}C_{22} = C_{24} + X + ZWJ + C_{22}$	(න්‍ද = න + ් + ZWJ + ද)
$C_1C_{37} = C_1 + X + ZWJ + C_{37}$	(ක්‍ෂ = ක + ් + ZWJ + ෂ)
$C_{20}C_{35} = C_{20} + X + ZWJ + C_{35}$	(ත්‍ව = ත + ් + ZWJ + ව)
$S_3 = C_{33} + X + ZWJ$	(ර්‍ = ර + ් + ZWJ)

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

$C_{24}XC_{22} = C_{24} + X + ZWNJ + C_{22}$	(න්‌ද = න + ් + 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 generally happen.

5.8.7 Separate code where necessary - Diacritical marks should be assigned separate codes

SLASCII	SLS 1134:1996	UCS/Unicode
D_1 (◌ං) D_2 (◌ඃ)	D_1 (◌ං) D_2 (◌ඃ)	D_1 (◌ං) D_2 (◌ඃ)
V→*C*→X →*D*→V̲→S	*D*→*V*→*C*→V̲→P	*D*→*V*→*C*→V̲→P
Prevent by entering the wrong order. The sorting needs additional code E3 reserved for use with second '*combwa'* (ෙෙ◌).	Sorting makes it easier.	Sorting makes it easier.

5.8.8 Easiness of transliteration - Transliteration from Sinhala to Tamil should be considered

SLASCII	SLS 1134:1996	UCS/Unicode
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 1134:1996 and UCS/Unicode for Tamil letters ண (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.

5.8.9 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.

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 ($C_1\underline{V}_{16}$ = කො), consonant with more than individual vowel sign as coded in SLASCII and after re-ordering process implies ($C_1V_{13}V_2X$ = කෝ), a conjunct ligature used in UCS/Unicode (C_1X+ZWJ+C_{37} = ක්‍ෂ) or conjunct ligature used in SLS 1134:1996 (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 ($C_1\underline{V}_2D_1$ = 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 usage. It can define as:

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

In the case of vowel signs (V), they are graphical signs 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 (C) order define as:

Sort_key_6: {$\underline{C}_1$ < $\underline{C}_2$ < $\underline{C}_3$}

5.8.10 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 are 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	**Example 1**	**Example** 2
Key-in sequence	ක්‍රෝං = ෙ ක ්‍ර ා ් ං	ක්‍ෂේම = (Not implemented)
After Re-ordering	ක්‍රෝං = ක ් ෙ ්‍ර ා ං	ක්‍ෂේම = (Not implemented)
For SLS 1134:1996		
Key-in sequence	ක්‍රෝං = ෙ ක ්‍ර ා ් ං	ක්‍ෂේම = ෙ ක Ln ෂ ් ම
After Re-ordering	ක්‍රෝං = ක ෙ ්‍ර ා ් ං	ක්‍ෂේම = ක ් Ln ෂ ේ ම
For UCS/Unicode		
Key-in sequence	ක්‍රෝං = ෙ ක ්‍ර ා ් ං	ක්‍ෂේම = ෙ ක ෟ ෂ ් ම
After Re-ordering	ක්‍රෝං = ෙ ක ්‍ර ා ් ං	ක්‍ෂේම = ක ් zwj ෂ ේ ම

NOTES FOR SLASCII

1) Only seven 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} (ඏ) and V_{12} (ඐ) do not occur in modern usage; therefore, these two were not included in the code table.
3) Independent vowel letter V_{10} (ඎaa) also does not occur in modern usage, but its corresponding vowel sign $\underline{V}_{10}$ (◌aa) is used, for example, C_{20} (ත) and it represented by a combination of $\underline{V}_9$ (◌a); for example $C_{20}\underline{V}_9\underline{V}_9$ = තaa.
4) The consonant C_{12} (ඦ) does not occur in modern 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.

NOTES SLS 1134:1996

1) Independent vowel letter V_{11} (ඏ) and V_{12} (ඐ) do not occur in modern 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} (ඏ) and V_{12} (ඐ) included in the code set for completeness of the code.
3) Independent vowel letter V_{10} (ඎaa) also does not occur in modern usage, but its corresponding vowel sign V_{10} (◌aa) are used, for example, කර්තaaහු.

NOTES UCS/Unicode

1) Complete consonant set considered, and every consonant in the alphabet has its codes.
2) Complete vowel set considered and every vowel in the alphabet has its codes.
3) The *virama* X (◌්) has a higher order of the vowel signs.
4) No alternative vowel signs and Before the vowels
5) There are five vowel signs ($\underline{V}_{10}$ to $\underline{V}_{14}$) have glyph pieces that 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.

5.8.11 Keyboard Sequence Compatibility

The most direct and unambiguous transformation of keyboard symbols to the internal representation is integral.

There have been several studies for the 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 designed in such a way that all characters that are to used for interchanging Sinhala to be represented by 46 keys as given in Figure 44. The following principles 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, maybe 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.

5.9 Discussion

Using ZWJ for conjunct ligatures that usually formed is acceptable; such as ක්‍ෂ (k.ssa$_n$) is a special ligature which according to the Sinhala orthography used to write Sanskrit words. Some of these conjunct forms used in writing Sanskrit and Pali words in Sinhala. On the other hand, in other Indic script k.sha fall under language feature which means k.sha ligature is unbreakable. In Sinhala writing, the word like පක්‍ෂ (pa$_n$ + k.ssa$_n$) in Sanskrit writing or පක්ෂ (pa$_n$ + ka$_n$ + ssa$_n$) in Sinhala contemporary writing is acceptable behaviors. Figure 67 shows some examples of Sinhala and Sanskrit syllable cluster behavior in UCS/Unicode. However, according to the Pali writing, conjunct ligatures listed as k,va$_n$, t.tha$_n$, t.va$_{n,}$ n.tha$_n$, n.da$_n$, n.dha$_n$ and sixty-eight touching letters have to be implemented as unbreakable ligatures.

Words	UCS/Unicode Code Sequence	Syllable Clusters
අත්‍ම	0d85 0dad 0dca 200d 0dae	අ ත ◌් ZWJ ම
අර්ත්‍ම	0d85 0dbb 0dca 0dad 0dca 200d 0dae	අ ර ◌් ත ◌් ZWJ ම
අර්‍ත්‍ම	0d85 0dbb 0dca 200d 0dad 0dca 200d 0dae	අ ර ◌් ZWJ ත ◌් ZWJ ම
අත්‍මර්‍	0d85 0dad 0dca 200d 0dae 0dbb 0dca 200d	අ ත ◌් ZWJ ම ර ◌් ZWJ
අත්‍මර්‍ය	0d85 0dad 0dca 200d 0dae 0dbb 0dca 200d 0dba	අ ත ◌් ZWJ ම ර ◌් ZWJ ය
අත්‍මර්‍‌ය	0d85 0dad 0dca 200d 0dae 0dbb 0dca 200d 200c 0dba	අ ත ◌් ZWJ ම ර ◌් ZWJ ZWJN ය
ආර්ය	0d86 0dbb 0dca 0dba	ආ ර ◌් ය
ආර්‍ය	0d86 0dbb 0dca 200d 0dba	ආ ර ◌් ZWJ ය
ආර්‍ය්‍ය	0d86 0dbb 0dca 200d 0dba 0dca 200d 0dba	ආ ර ◌් ZWJ ය ◌් ZWJ ය
වක්‍ර	0dc0 0d9a 0dca 200d 0dbb	ව ක ◌් ZWJ ර
ව්‍යංග	0dc0 0dca 200d 0dba 0d82 0d9c	ව ◌් ZWJ ය ◌ං ග

Figure 67: Some examples of Sinhala and Sanskrit syllable clusters in UCS/Unicode

The convention of writing a pure consonant touching the following letter, instead of using an al-lakuna$_n$, is standard in Pali text written in Sinhala script. This represents by using the sequence 200D (ZWJ) between the two letters after the al-lakuna$_n$ like acceptable conjunct ligature in Sinhala orthography used to write Sanskrit words. This approach sounds logical as this will allow someone to encode both Pali and Sinhala in the same document without having rich-text or markup. However, the independence of language and script needs to be clarified.

However, writing Pali text in Indic languages has not recognized Pali as an independent language. For example, in Malayalam, to form a conjunct consonant for *Buddho* (ബുദ്ധോ) use the default rendering behavior, where ZWJ is not inserted. As shown in Figure 67, *Buddho* (බුද්ධො) written in Pali in Sinhala text to form an independence ligatures, as default behaviors of placing ZWNJ in Sinhala text.

The allocation of the code for the ZWNJ in writing Pali in Sinhala is appreciated and if someone needs to be used conjunct ligature form of බුද්‍ධො (*Buddho*) instead of nominal glyphs form බුද්ධො (*Buddho*), ZWNJ can replace with ZWJ. However, one thing should be noticed here, by doing so now required to encode as extra two byte-character to render a mandatory orthographic construct which could have been easily handled by the rendering engine.

Another problem with this arrangement is that the Indic scripts are not all the same; in fact, there are some very significant differences between scripts. It is intended that similar specifications should be applied as appropriate to Sinhala as well. However, as the Figure 68 shows, combined or touching cases have been handling in the default Indic way in www.tipitaka.org and ZWNJ inserted to indicate when the consonant conjunct should not be formed; i.e. බුද්ධො. In Sinhala script used with Pali writing, there are ambiguities with regards to which consonants or consonants conjuncts are to be displayed (see Figure 69).

Word Buddha	UCS/Unicode code sequence	Language	With ZWNJ
බුද්ධො	0db6 0dd4 0daf 0dca **200c** 0db0 0ddc	Sinhala	බුද්‌ධො
ബുദ്ധോ	0d2c 0d41 0d26 0d4d 0d27 0d4b	Malayalam	ബുദ്‌ധോ
बुद्धो	092c 0941 0926 094d 0927 094b	Devanagari	बुद्‌धो
বুদ্ধো	09ac 09c1 09a6 09cd 09a7 09cb	Bengali	বুদ্‌ধো
ಬುದ್ಧೋ	0cac 0cca 0ca6 0ccd 0ca7 0ccb	Kannada	ಬುದ್‌ಧೋ
బుద్ధో	0c2c 0c41 0c26 0c4d 0c27 0c4b	Telugu	బుద్‌ధో
พุโทฺธ	0e1e 0e38 0e42 0e17 0e3a 0e18	Thai	พุโท‌ฺธ

Figure 68: UCS/Unicode code sequence used for the word Buddha appeared in www.tipitaka.org

Giving explicit control of conjunct formation is sensible, but by doing so we are now required to encode an extra 2-byte-character to render a mandatory orthographic construct that could have been easily handled by the rendering engine. UCS/Unicode standard recommends using ZWJ and ZWNJ characters for the following purposes; “ZWJ and ZWNJ produce a ligature or cursive connection where one would otherwise not occur. These two characters can be used to override the default choice of glyphs.” (Unicode Standard, 1998)

ත්ථ	0dad 0dca 0dae	Display isolated consonants cluster (by default) in Sinhala without using ZWJ or ZWNJ
ද්ධ	0daf 0dca 0db0	
ත්‌ථ	0dad 0dca 200c 0dae	ZWNJ used to display as isolated ligatures in www.tipitaka.org
ද්‌ධ	0daf 0dca 200c 0db0	ZWNJ used to display as isolated ligatures in www.tipitaka.org
ත්‍ථ	0dad 0dca 200d 0dae	ZWJ used to display a half form in a consonant cluster
ද‍්ධ	0daf 200d 0dca 0db0	ZWJ used to display a touching form in Pali

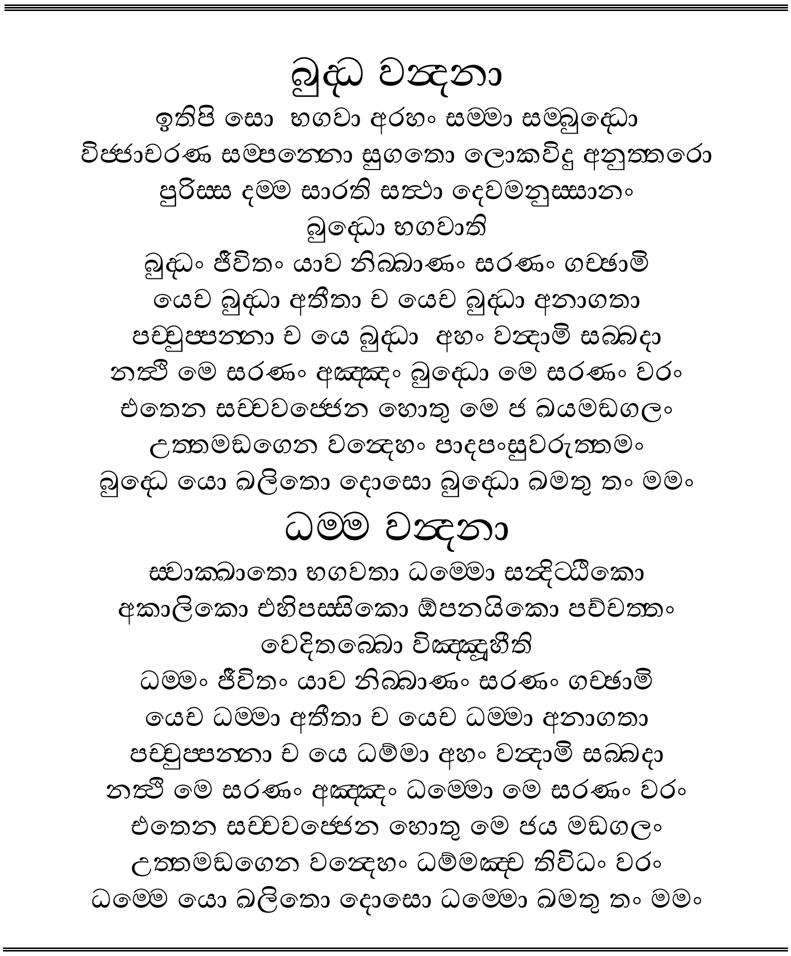
Figure 69: Ambiguous in conjoining sequences

Let’s look at the ZWJ *al-lakuna* character sequence for the formation of touching letters (cursively connected). The need for this code sequence for ‘cursively connected’ letters will arise only if for example n.da conjunct can be rendered as a ligature and also be rendered a cursively joined. But, according to Pali alphabet (Amaramoli, 1959), there are not many cases exist in contemporary writing Pali in Sinhala. The only conjunct ligatures are k,va$_{n}$, n.da$_{n}$, n.dha$_{n}$, n.tha$_{n}$, t.tha$_{n}$ and t.va$_{n}$ (Amaramoli, 1959). In this case, we can use the code sequence al-lakuna + ZWJ for conjunct ligatures. On the other hand, there are many cases are existing for touching letters in Pali as shown in Figure 65. For this case, we may require some way of using ZWJ for the touching letters.

However, this was recognized by Michael Everson (Michael et al., 1997) and according to that one can have both conjunct ligature and in simple ligature form in writing Sinhala and Pali language. The interesting thing to notice here is that even though Michael and the members of the ad-hoc committee have clearly recognized the distinction of script and language. They recommended using a ZWNJ when *Buddho* is written in Sinhala, which implies that they trying to cater to both languages simultaneously.

Nevertheless, the suggested way of encoding would be ideal if Pali is recognized as an independent language which has its orthographic conventions such as the *al-lakuna* sign not being depicted visually.

Alternatively, as Peter Constable (Constable, 2004) suggested as a possible solution like ZWJ and VIRAMA, just as for conjoining conjuncts, rather than some other sequence. For example, touching conjuncts like 'ද්ධ' (d.dha$_n$) can have a sequence such as da$_n$ + ZWJ + *al-lakuna*$_n$ + dha$_n$. Peter Constable's suggestion executed and Figure 70 shows that sample specimen of Pali text in Figure 64 been created using the sequences such as C$_n$ + ZWJ + *al-lakuna*$_n$ + C$_n$.



බුද්ධ වන්දනා
ඉතිපි සො භගවා අරහං සම්මා සම්බුද්ධො
විජ්ජාචරණ සම්පන්නො සුගතො ලොකවිදූ අනුත්තරො
පුරිස්ස දම්ම සාරති සත්ථා දෙවමනුස්සානං
බුද්ධො භගවාති
බුද්ධං ජීවිතං යාව නිබ්බාණං සරණං ගච්ඡාමි
යෙච බුද්ධා අතීතා ච යෙච බුද්ධා අනාගතා
පච්චුප්පන්නා ච යෙ බුද්ධා අහං වන්දාමි සබ්බදා
නත්ථි මෙ සරණං අඤ්ඤං බුද්ධො මෙ සරණං වරං
එතෙන සච්චවජ්ජෙන හොතු මෙ ජ ඛයමඞ්ගලං
උත්තමඞ්ගෙන වන්දෙහං පාදපංසුවරුත්තමං
බුද්ධෙ යො ඛලිතො දොසො බුද්ධො ඛමතු තං මමං
ධම්ම වන්දනා
ස්වාක්ඛාතො භගවතා ධම්මො සන්දිට්ඨිකො
අකාලිකො එහිපස්සිකො ඕපනයිකො පච්චත්තං
වෙදිතබ්බො විඤ්ඤූහීති
ධම්මං ජීවිතං යාව නිබ්බාණං සරණං ගච්ඡාමි
යෙච ධම්මා අතීතා ච යෙච ධම්මා අනාගතා
පච්චුප්පන්නා ච යෙ ධම්මා අහං වන්දාමි සබ්බදා
නත්ථි මෙ සරණං අඤ්ඤං ධම්මො මෙ සරණං වරං
එතෙන සච්චවජ්ජෙන හොතු මෙ ජය මඞ්ගලං
උත්තමඞ්ගෙන වන්දෙහං ධම්මඤ්ච තිවිධං වරං
ධම්මෙ යො ඛලිතො දොසො ධම්මො ඛමතු තං මමං

Figure 70: Sample specimen of Pali text shown in Figure 65 representing with the different conjoining sequences of "Consonant + ZWJ + al-lakuna"

5.10 Conclusion

The Sinhala language is more complicated than some of the other Indic scripts since it is used with Pali and Sanskrit writing. Sri Lanka being an Island, and the 2500 years continues development of the language character set was more independent than any other languages in the Indian sub-continent, However, both conceptually and in processing can be treated as set of elements: consonants, stand-alone vowels, vowel signs, diacritical marks, non-vocalic strokes, ligatures, and its combined forms. In a large set of Sinhala character designs and rendering multiple orthographic environments, many details can be discussed, but in this research, standardization of character set for the different technological stages including addressing the issues in development and implementation has touched several of the more exciting and general issues.

Finally, the best practice for composing digital texts in the Sinhala language is to follow the UCS/Unicode standard and to use a UCS/Unicode font, either of the OpenType or Apple Advanced Technology (AAT) variety. This practice will ensure that documents will be easily accessible for years to come. Because UCS/Unicode is an industry-standard, infinitely expandable, and used in all contemporary operating systems (Windows, OS X, Linux), as well as the World Wide Web, UCS/Unicode is the best character encoding standard. It is here to stay. Font technology, on the other hand, will continue to evolve as new display engines and fonts are continually developed. By composing texts according, the UCS/Unicode standard, texts will be able to easily take advantage of evolutions in font technology that will only improve the display of any documents.

Chapter 6: Conclusion

Asian languages are especially rich in scripts. The five basic scripts are: Ideographic, Brahmi, Latin, Arabic, and Cyrillic. These languages grew up in the region, each separated mainly by mountains, ocean, or deserts. In East Asia, the influence of Chinese ideographic script (hanzi) is remarkable. In South Asia, in and around the Indian subcontinent and in the continental part of Southeast Asia, the scripts originated from Brahmi-scripts. The islands of Southeast Asia and Australasia have mostly adopted Latin scripts (some islands in the region still use Brahmi-originated scripts such as the Balinese script, or *aksara* Bali). In Central Asia, historically, languages were written in the Arabic script under the influence of the Ottoman Empire but later transformed into Cyrillic. Lastly, in the western part of Asia, Arabic script is widely used not only by Arabic speakers but also by non-Arabic speakers.

Sri Lanka is an island country and has always been independent in her 2500 years of recorded history. Thus, the Sinhala language of Sri Lanka remains unique and cannot automatically be handled by technologies developed for the roman script region. However, southeast and south Asian languages, such as Tamil, Thai, Khmer and Myanmar have many standard features in writing systems, and these language communities could share ample of common challenges to more ahead in their text processing.

On the other hand, more than 2500 years of cultural, social, and religious influence originated from the 3rd century B.C.; the original Brahmi script developed into the varieties of complex scripting systems. For example, the island of Sri Lanka's writing style uniquely spoken, and written language of Sinhala due to the country is an Island and scripts were heavily influenced by Buddhism, and in some stages of history, primarily from the 8th century to the 10th century, alphabet and writing systems have changed considerably with influence by the *Kadamba* and *Pallawa Grantha* script of South India. As discussed in early chapters, the Sinhala character set has unique features such as its rounded shape, more vowels than any other languages in the world, out of which two vowels do not exist in any other languages in the world. Further, the unique five consonants known as half-nasals that do not exist in the world languages and, has many complex ligatures in the writing system, and therefore, implementation of writing system for the early foundry font-based industry such as printing and later character-based computing were hard goals to achieve.

Besides, the Sinhala language used in Sri Lanka known as a mixed language is used to write *Pali* and *Sanskrit* words that embedded in the modern Sinhala known as loan words. In the Sinhala writing system, *Pali* and *Sanskrit* could be written without using their consonant conjuncts form or touching form, which is acceptable to some extent. Designing of 8-bit SLASCII and the development of systems and applications based on 8-bit standards, writing styles of *Sanskrit* and *Pali* were ignored due to the limitation of the technology available during that period. However, with 16-bit SLS 1134:1996 and UCS/IEC UNICODE standards, writing styles of *Pali* and *Sanskrit* were considered during the process of implementation and development of standards and applications for the Sinhala language.

In the early seventies century, the introduction of the printing industry to the country by Dutch, and foundry font technology by many European countries, changed the ages-old tradition of books writing style on *ola-leaf* (from *Palmyra* tree) with a sharper steel stylus point. The development of foundry technology with the European interest in Sinhala script was a base of the theoretical and mathematical way of designing the Sinhala script, which shows that even at that time the font width was a very significant factor for printing, and therefore, they have paid due attention for the variable width of each font. Subsequently, from the beginning of the 19th century, advocating the development and improvement of Sinhala printed letter shapes and quality improved by its circularity and individuality. The font technological enhancement was the main technological breakthrough for the design of font for the computers at the later stage.

Since the programmable computer technology introduced in the micro-computing era in the early 1980s and with the introduction of low-cost 8-bit micro-computers, the demand increases from the public for knowledge sharing & enhancement and information gathering through computer technology by their languages, such as Sinhala and Tamil; therefore, there was an urgency for the development of native language computing to the country. Furthermore, generation changes in the computing industry, language-related issues such as standards for the character set, keyboard standards for computers, development of font sets for display terminals and for the printers, development of algorithms to handle keyboard inputs and rendering issues for proper character display were to be addressed independently based on the hardware resources available in every stages of the technology. All these issues and related problems addressed carefully with assistance from the language and linguistic experts, national and international technologists, governing authorities, international organizations and more precisely from the users and public.

In the late-1980s, at the time of design 8-bit standards for the Sinhala character set, the primary consideration was given to the following guidelines. Alphabetical order to be recognized by the national-level authorities. After that, designing the standard codes for the internal representation, sorting and collating sequence was one of the other considerations. The amount of storage requirement in the memory and the number of code positions in the code table was another critical consideration. In addition to the arrangement made according to the alphabetical order for the diacritical marks, vowels and consonants, the order of the vowel sing arrangement should be arranged according to the sorting order, which is essential to maintain the sorting order required for the dictionary users. Rendering is another critical factor considered. For example, for easiness of shorting, which means any vowel modifies appearing before the consonant should always be moved after the consonant to minimize the hazels of sorting in Sinhala. Also, rendering was essential to maintain the proper appearance on the display and the printing. Standard codes should directly correspondence with existing Sinhala keyboard, was the other consideration. The architecture of the Sinhala Basic Input and Output System known as SBIOS API input method provided a sequence checking mechanism to ensure the validity of the input sequence. The sequence checking mechanism is capable of monitoring the keying sequences of three levels; that is of the strictness of the upper level, base level, and lower level. This mechanism ensures the correct appearance of the glyphs including the implementation of consonants conjuncts and joined ligatures used in *Sanskrit* and *Pali* writing systems. Besides, ensure the correctness of sorting order on the Sinhala language. This mechanism of sequence checking applied not only in early 8-bit SLASCII API architecture but also every stage of development in the 16-bit cording standards related font technologies. These guidelines were maintained throughout the process of standardization from 8-bit SLASCII to 16-bit SLS 1134:1996 and finally to the UCS/IEC UNICODE standards.

The Sinhala language is more complicated than some of the other Indic scripts since it is used with Pali and Sanskrit writing. Development of the language character set was more independent than any other language in the Indian sub-continent; however, both conceptually and in processing can be treated as set of elements: consonants, stand-alone vowels, vowels signs, diacritical marks, non-vocalic strokes, ligatures, and its combined forms. In a broad set of Sinhala character designs and rendering multiple orthographic environments, many details discussed, but in this research, standardization of character set for the different technological stages, including addressing the issues in development and implementation has touched.

The identity of the Sinhala language and its recognition must be maintained. This can be done through newspapers, local radio stations, broadcasting stations, or social media. It could also be done on a low-cost technology (UNESCO 2001) and worldwide scale with the help of UNESCO, which is the organization committed to preserve human culture and languages and narrow down the digital language divide. Delivering local Education in the mother tongue is another way of promoting and preserving languages of the digital world. Building multilingual word dictionaries, maintaining everyday social context and standard oral corpus would deem useful amongst the multilingual communities. Thus, the creation of digitized corpora is a fundamental task in the effort to preserve the world’s languages. The corpora can be multimodal, spoken or written, depending on what type of linguistic material and recording equipment is available. This will bring to a discussion of the role of language technology once again. Probably a key factor here is the reuse of technology for similar languages. This will bring the solution as mention in the publication 'E-commerce, and Development Report 2003' by the United Nations Conference on Trade and Development (UNCTAD) provides an insight into the software that developing countries can use for bridging the digital divide. It recommends the use of free or open-source software against the proprietary license-to-use software.

One of the majors concerned by UNESCO and other related organizations is the "digital divide" which is the gap in technology usage and access to the information. The digital divide has been investigated by scholars and policymakers mainly as an economy-specific issue that permeates the population across all demographic profiles, such as income, gender, age, education, race, and region, but not specific to the languages of different communities. The lack of native language-driven ICT is a major conducive factor in the digital divide.

Once again, the Sinhala writing system used in Sri Lanka is a syllabic writing system derived from Brahmi which consists of vowels, consonants, diacritical marks, and special symbols constructs. Several of these constructs are combined to form complex ligatures. Because the Sinhala language evolved independently and uniquely as a result and cannot be automatically handled by technologies originating from the west or for that matter those from the Indian subcontinent, however, the writing system for various Southeast and South Asian languages have many commonalities could share an array of common challenges in moving ahead in their text processing techniques. Having the total number of different glyphs is almost close to 2300 in the Sinhala language, sharing achievements and know-how in this area will help us to minimize the digital language divide in the region.

In the end, the author would like to quote Michael Ondaatje, a renowned novelist "*I still believe the most beautiful alphabet was created by the Sinhalese. The insect of ink curves into a shape that is almost sickle, spoon, eyelid. The letters are washed blunt glass which betray no jaggedness. open-source, but its sharp grid features were not possible in Ceylon. Here the Ola leaves which people wrote on were too brittle. A straight line would cut apart the leaf and so a curling alphabet was derived from its Indian cousin* — Moon coconut. The bones of a lover's spine" Michael Ondaatje (The Booker Prize-winning author of *The English Patient*), "*Running in the Family", (1982: p83-84*).

References

Amaramoli, 1959. Venerable Pandita Veragoda Amaramoli Thera, Vinayapitaka-Parajika Pali, Budda Jayanti Tripitaka Series, Volume 1, The Government of Ceylon, 1959.

ANSI/NISO Z39.53-2001. Revision of ANSI/NISO Z39.53-1994. “Codes for the Representation of Languages for Information Interchange”, A National Information Standards Organization, ISSN: 104105653.

Andy Daniels and Joe Becker, 1992, “Unicode Technical Report #2 for Sinhala, Tibetan and Mongolian”, 03-11-1992.

Annual Project Report-2005, 2006. “Asian Language Resource Network Project”, Nagaoka University of Technology, Japan, March 2006.

CANLIT & NARESA, 1985. Working Paper, Order of Alphabet and System of Transliteration, 1985.

Constable, P., 2004. *Proposal on Clarification and Consolidation of the Function of ZERO WIDTH JOINER in Indic Scripts*, Public Review Issue #37, Microsoft, USA, 2004-06-30.

Daniels P. T. and Bright W., The World’s Writing Systems, Oxford Univ. Press, 1996. pp. 445-484.

Daniel J. T. K. & Hedlund R. E., 1993. Carey’s Obligatio and India’s Renaissance, Council of Serampore College, Seranpore, West Bengali, 1993.

De Silva, M. W. Sugatha pala, 1982. Some Consequences of Diglossia, 1982.

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

Ethnologue: Languages of the World, 15th ed., SIL Int’l, 2005.

Falk, Harry, 1993. Schift im altenIndien, Ein Forschiungsbericht mit Anmerkungen (ScriptOralia 56), Tübinger: Narr.

Fernando, P. E. E., 1949. *Paleographical Development of the Brahmi Script in Ceylon from 3*rd Century B.C. to 7th Century A.D., University of Ceylon Review 7. p282-301.

Florian Coulman, 1996. *The Blackwell Encyclopedia of Writing Systems*, Blackwell Publishers Ltd. P. 469.

Florian Coulmas, 1989. *Sanskrit* (from sạm-skṛta ‘elaborated’) is that phase of the literary language of ancient India which is described in the grammar of Panini, *p186*.

Gunasekara, Abraham Mendis, 1891. *A comprehensive Sinhalese Language,* Asian Educational Services, New Delhi. pp. 3-30.

Hof K. K., 1876. *Alphabet of all races of the world,* Royal Print Shop, Vienna, Germany.

Hoffman, Donna L.; Novak, Thomas P., 1998. Bridging the Digital Divide: The Impact of Race on Computer Access and Internet Use. Educational Resources Information Center, Department of Education, U.S.A.

Hugh McGregor Ross, 1996. *Sinhala proposal,* ISO/IEC JTC1/SC2/WG2 N1376, 1996

ISO 639, 1988, International Organization for Standardization, “Code for the representation of names of languages, 1st edition”, ISO Standard 639.

ISO 639-2:1998, 1998. “Codes for the representation of names of languages -- Part 2: Alpha-3 code - edition 1”, International Organization for Standardization.

ISO TC46/WG3 and M. Everson, Ed., 2003, "ISO 15924:2003 (E/F) - Codes for the representation of names of scripts".

Jayasiri Lankage, 1996, Evolution of Sinhala Scripts, S. Godage & Pvt Ltd, Colombo, SriLanka.

Jayathilaka, D. B., 1937, “*Sinhala Shabdha Koshya”*, (The Sinhalese Dictionary), RASCB, p. ix.

John Clews, 1997. Digital Language Access: Scripts, Transliteration, and Computer Access, SESAME Computer Projects, D-Lib Magazine, ISSN 1082-9873, March 1997.

John Devy, 1990. "An Account of the interior of Ceylon", First published in 1812. New Delhi, AES.

Kulasuriya, A., 1962, "*Sinhala Sahitya"*, (The Sinhalese literature) part I, Saman Printers, Maharagama, pp. 67,68.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (Eds.), 2013. Ethnologue Languages of the World (17th ed.). Dallas, TX SIL International.

Lloyd Anderson, Ken Whistler, Peter Lofting, Rick McGowan (Contributors), 1992. *Draft Proposals Unicode Technical Report #2,* Unicode Inc., USA.

Monotype Corp, 1959. *Book of Information,* London & Salfords.

Michael Everson,1989. Proposal for encoding the Sinhala script in ISO/IEC 10646 (revision 1). http://www.evertype.com/standards/si/si.html

Michael Everson, 1996. *Report of the Sinhala Standards,* ISO/IEC JTC1/SC2/WG2 N1473R.

Michael Everson, Takayuki Sato, Kohji Shibano, Disanayaka J. B., Nandasara S. T., Johan van Wingen & Glenn Adams, 1997. *Minutes of Sinhala Ad-Hoc Committee, Doc. # 1613, ISO Meeting No. 33, Iraklion, Crete, Greece.*

Michael Everson, 1997a. Proposal for encoding the Sinhala script in ISO/IEC 10646 (revision 1), Expert Contribution ISO/IEC JTC1/SC2/WG2 N1473R, Date: 1997-03-01

Michael Everson, 1997b. Mapping of Sinhala between ISO/IEC 10646 and SLS 1134:1996, Expert Contribution, Everson Gunn Teoranta (IE), ISO/IEC JTC1/SC2/WG2 N_______ Date: 1997-06-1.

Michael Everson, 1999. Response to N2084 (Sources of characters), Contributing Editor, Personal contribution, ISO/IEC JTC1/SC2/WG2 N2106R 1999-09-30.

Michael Ondaatje, 1984, "Running in the Family", pp 83-84.

Mikami Y, Zavarsky P., Rozan M. L., Suzuki I., Takahashi M., Maki T., Nizan Ayob I., Boddi P., Santini M., Vigna S., 2005. "The Language Observatory Project (LOP)", www 2005, Proceeding, Chiba, Japan, 2005.

Mike Ksar, Umamaheswaran V.S., 1997. Unconfirmed Minutes of WG 2 Meeting # 32, Singapore; 1997-01-20--24, ISO/IEC JTC 1/SC 2/WG 2 N1503 Date: 1997-04-01.

Mike Ksar, Umamaheswaran V.S., 1997. Unconfirmed Minutes of WG 2 Meeting # 33, Heraklion, Crete, Greece; 1997-06-30/07-04, ISO/IEC JTC 1/SC 2/WG 2 N1603 Date: 1997-10-24.

Mike Ksar, Umamaheswaran V.S., 1998. "Meeting Report", JTC 1.02.18 – ISO/IEC 10646, Unconfirmed Meeting Minutes, ISO/IEC JTC 1/SC 2/WG 2 Meeting # 34, Redmond, WA, USA; 1998-03-16/20, ISO/IEC JTC 1/SC 2 N 3148, 1998-07-22.

Mike Ksar, Umamaheswaran V.S. 1998. Unconfirmed Minutes of WG 2 meeting #35, London, U.K.; 1998-09-21--25, ISO/IEC JTC 1/SC 2/WG 2 N1903, DATE: 1998-12-30.

Naito Eisuke, 1998. *Progress Report of the MLIT Project,* AFSIT-12, Ha Noi, Vietnam, October 1998.

Nandasara, S. T., Disanayaka, J. B., Samaranayake, V. K., Seneviratne, E. K., and Koannantakool, T., 1990. *– Draft* Standard *for the Use of Sinhala in Computer Technology,* approved by the CINTEC on the advice of its working committee for recommending Standards for the Use of Sinhala and Tamil Script in Computer Technology, 1990.

Nandasara S.T., Samaranayake V. K., Disanayaka J. B., 1991. "*A Standard Code for Information Interchange in Sinhala*", Proposal submitted to Sri Lanka Standard Institute (SLSI), 1991.

Nandasara, S. T., 1991. "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, January 1991.

Nandasara, S. T., Samaranayake, V. K., 1994. "A Standard Code for Information Interchange in Sinhalese" Proceeding of the International Conference on the Standardization of Asian Languages, CICC, Tokyo, Japan, 1994.

Nandasara, S. T., 1996. "*Draft Sri Lankan Standard proposal for Sinhala Character Code for Information Interchange*", Proposal submitted to Sri Lanka Standard Institute (SLSI), May 1996.

Nandasara, S. T., Samaranayake, V. K., 1997. *Current Development of Sinhala / Tamil / English Trilingual Processing in Sri Lanka*, Second Internation Symposium on Standardization of Multilingual Information Technology, MLIT-2, November 7-8, Tokyo, Japan, 1997, pp. 181-192.

Nandasara, S.T., 1997. *Sri Lanka Experience of Development of Tamil Input/Output/Display Methods,* TAMILNET'97 – International Symposium, Singapore, May 1997, pp 113-121.

Nandasara, S. T., Leong, K. Y., Samaranayake, V. K., and Tan, T. W., 1997. *Trilingual Sinhala-Tamil-English National Web Site of Sri Lanka*, 1997, INET97, http://www.isoc.org/inet97/proceedings/EI/E1_3.HTM, INET97, Kuala Lumpur, Malaysia, June 1997,

Nandasara, S. T., 2000, (*Contributing author)*, "The Unicode Standard 3.0", Chapter 9, Unicode Consortium, (www.unicode.org), Addison-Wesley Publishing Co., ISBN-13: 9780201616330, ISBN-0-201-61633-5, Copyright © 1999-2000 by Unicode, Inc. pp. 209-254.

Nandasara, S. T., Shigeaki Kodama, Chew Yew Choong, Rizza Caminero, Ahmed Tarcan, Hammam Riza, Robin Lee Nagano and, Yoshiki Mikami, 2008. An Analysis of Asian Language Web Pages, The International Journal on Advances in ICT for Emerging Regions, ISSN 1800-4156, Volume 1, Number 1, 2008 01 (01), pp. 12-23.

Nandasara S. T**.**, Yoshiki Mikami, 2008. "*Sinhala Computing in Early Stage – Sri Lanka Experience*", in IFIP International Federation for Information Processing, History of Computing Education 3, ed. John Impagliazzo, ISBN: 978-0-387-09656-8, (Boston, Springer), 2008, pp. 157-169.

Nandasara S. T., 2009. "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.

Nandasara, S. T. and Yoshiki Mikami, 2009, "*Guest Editors' Introduction: Asian Language Processing: History and Perspectives*," *IEEE Annals of the History of Computing, Special Issue: Asian Language Processing,* ISSN 1058-6180, Volume. 31, Number. 1, January-March, 2009, IEEE Computer Society, pp. 4-7.

Nandasara, S. T**.**, 2012. *"Evolution of Computer Science Education in the purview of Free Education",* A. Tatnall (Ed.): Reflections on the History of Computing-*Preserving Memories and Sharing Stories*, IFIP AICT 387, ISBN: 978-3-642-33898-4, Springer-Verlag GmbH Berlin, 2012, pp. 127–142.

Nandasara, S. T., Yoshiki Mikami, 2015. *Bridging the Digital Divide in Sri Lanka: Some Challenges and Opportunities in using Sinhala in ICT,* The International Journal on Advances in ICT for Emerging Regions, ISSN 1800-4156, Volume 8, Number 1, May 2015. pp. 85-97.

NTIA, 1999. Falling Through the Net: Defining the Digital Divide, National Telecommunications and Information Administration, U.S. Department of Commerce, U.S.A.

O'Neill E.T., Lavoie B.F., Bennett R., April 2003. Trends in the Evolution of the Public Web 1998 - 2002, D-Lib Magazine, Volume 9, April 2003.

Paranavitana S, 1970, *Inscriptions of Ceylon*, Vol. I, 1970.

Paolillo J., Pimienta D., Prado D, et al., 2005. Measuring Linguistic Diversity on the Internet, UNESCO Institute for Statistics, Montreal, Canada.

Piyadasa T. G., 1985, Libraries in Sri Lanka, Their Origin and History from Ancient Times to the Present Time, Sri Satguru Publication, India, pp. 1-18.

Revised text of ISO/IEC 10646-1/FPDAM 21, 1998. Universal Multiple-Octet Coded Character set (UCS) -- Part 1: Architecture and Basic Multilingual Plane -- AMENDMENT 21: Sinhala, ISO/IEC JTC 1/SC 2 N 3190 Date: 1998-10-22.

Rob Pike & Ken Thompson, Bell Labs, rob,ken@plan9.bell−labs.com.

Samaranayake, V. K., Disanayaka, J. B. and Nandasara, S. T., 1989. *A standard Code for Sinhala Characters,* Proceedings, 9th Annual Sessions of the Computer Society of Sri Lanka, Colombo, 1989.

Samaranayake, V. K., Nandasara, S. T., 1990. *A Standard Code for Information Interchange in Sinhalese,* ISO-IEC JTC 1/ SC 2/WG 2 N673, Unicode Inc., USA, October. 1990.

Samaranayake, V. K., et al., 1996. Sri Lanka Standard for the Sinhala Character Code for Information Interchange, Presented by Chairman, Working Group on Sinhala Character Code for Information Interchange of the Sri Lanka Standards Institute, on behalf of the Working Group, International Symposium on Multilingual Information Processing, Institute of Industrial Technology, 25-26 March 1996, Tsukuba, Japan, pp. 42-64.

SLS 1134:1996. Sri Lanka Standard SLS 1134:1996, 1996. Sinhala Character Code for Information Interchange, Document Np. N 1480, 1996.

SAORA's Survey Report, March 2005. Japan.

Sinha R. M. K., 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.

SLS 1134:1996. Sri Lanka Standard SLS 1134:1996-Sinhala Character Code for Information Interchange, SLSI Publication, Sri Lanka Standards Institute, Sri Lanka, 1996.

Takayuki K. Sato, 1998. Status of Cooperative Activities for the Missing Characters, MLIT Secretariat, CICC, Japan, 1998.

UNESCO, 2003. (Adopted by the UNESCO General Conference at its 32nd session) "Promotion and Use of Multilingualism and Universal Access to Cyberspace", 2003.

UNESCO, 2001, "Integrating Modern and Traditional Information and Communication Technologies for Community Development", An International Seminar addressing the digital divide in some of the poorest communities of the developing world, Kothmale, Sri Lanka, January 22 - 27,200l.

UNESCO Publication, 2005. "*Diversity and Endangerment of Languages in Nepal*", United Nations Educational, Scientific and Cultural Organization, Katmandu Office, Nepal, 2005.

Unicode Standard 3.0, The, 1998. *(*www.unicode.org*),* Addison-Wesley Pub Co., ISBN 02001616335, 1998.

US Library of Congress, "A standardized 3-character code to indicate the language in the exchange of information is defined". ISSN: 1041-5653, 2003.

Vinod Kumar Garg and Jivesh Bansal, 2007. Automation of Indian Languages through GIST: A Study of Panjab University Library,5th International CALIBER-2007, Panjab University, Chandigarh, 08-10 February 2007 © INFLIBNET Centre, Ahmedabad, India, 2007, pp. 486-490.

W3Techs, 2018. Internet Users by Languages, [Online] Available at https://w3techs.com/technologies/overview/content_language/all (Accessed: 31st June 2019)

Working Paper, 1985. Order of Alphabet and System of Transliteration, CANLIT & NARESA Working Committee, Sri Lanka, 1985.

World Factbook, 2018. US, Central Intelligence Agency, 2018.

World Stats, Internet, 2019. Internet content available by language, [Online] Available at https://www.internetworldstats.com/stats7.htm (Accessed: 31st June 2019).

World Stats, Internet, 2015. Internet content available by language, 2015.