Sinhala Computing in Early Stage – Sri Lanka Experience

S. T. Nandasara[1] **and Yoshiki Mikami**[2]

[1] University of Colombo School of Computing, Sri Lanka, nandasara@yahoo.com
[2] Nagaoka University of Technology, Japan, mikami@kjs.nagaokaut.ac.jp

Abstract: Sinhala writing system used in Sri Lanka is a syllabic writing system deriving from *Brahmi* and it consists of vowels, consonants, diacritical marks, and special symbols. Several of these are combined to form complex ligatures. Total number of different glyphs is almost close to 2300. Thus, all computer equipment for 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. In this paper, we discuss how Sinhala computing technology has evolved in early personal computers with limited capabilities and resources.

Keywords: Complex script computing, Sinhala language, Standarisation

1 Introduction

Sri Lanka has a population of 20 million of whom the majorities are Sinhalese (74%). Other ethnic groups are made up of Sri Lankan Tamils and Indian Tamils (18%), Moors (7%), Malays and Burghers. There are three living languages in Sri Lanka. They are Sinhala, Tamil, and English, used for general, everyday communication: both interpersonal and mass communication. Written documents, on paper or other materials, appear in one, two, or all of these languages.

This article focuses on key issues and the structure concerning Sinhala writing at the character level. Then we will discuss some of the major issues involved in design of Sinhala computing interface for early character base machines using 8-bit standard for Sinhala scripts.

2 Sinhala Scripts Structure and Major Issues

Sinhala script is used for writing the Sinhala language in Sri Lanka is said to be derivatives from the ancient scripts *Brahmi*, known to have existed since third to second century B.C.E. Literary Sinhala obtained its standard in the 14th century A.D., and this standard is respected by the whole speech community of Sri Lanka. Full Sinhala script includes the symbols necessary for writing loan words from Sanskrit and Pali, notably the aspirated consonants.

Formal description of the Sinhala character set can be defined as follows:
Semi-consonants = {ං, ඃ}
Vowels = {අ, ආ, ඇ, ඈ, ඉ, ඊ, උ, ඌ, ඍ, ඎ, ඏ, ඐ, එ, ඒ, ඓ, ඔ, ඕ, ඖ}
Consonants = {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}
Vowel signs = {ා, ැ, ෑ, ෘ, ෙ, ේ, ෛ, ෟ, ෲ, ෳ, ◌a, ◌aa, ◌ෙ, ◌ේ, ◌ෛ, ◌ො, ◌ෝ, ◌ෞ}
Non-vocalic strokes = {◌්, ◌්‍ය}

2.1 Major Issues in Writing Systems

We must consider the following points for use of the Sinhala writing systems. Firstly, every vowel except the first one has a corresponding vowel sign that can be attached to consonants to make composite characters. Secondly, when vowels appear at the beginning of a word, vowels are written as independent letters. Thirdly, there are two commonly used diacritical marks: '*anuswar'* and '*visarga'*, like most of the Indic languages. Fourthly, unlike in English, vowel signs are attached to the right, left, above or below to its fix position or variable position. When we attach some modifiers, it changes the original shapes of the consonants. Appearances of modifiers are also differed according to the consonants. Next, there are two special symbols (non-vocalic strokes) corresponding to the sound of 'r' and 'y' called *rakaransaya* and *yansaya*. Lastly, when Sanskrit and Pali words are adopted into Sinhala, they are transcribed in the compound manner in which they are 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.

2.2 Complex Ligature and Character Positioning

In Sinhala language, combinations of consonants, vowel signs, and diacritical marks are constructed in a different way according to the shape of the Sinhala

letter. Some would create a rather uneven, irregular, and illogical outer appearance.

Every combination is constructed in the way according to the shape of the Sinhala letter. Forty-one (41) consonants and sixteen (16) vowel signs combined to form glyphs. Thereafter, each united glyphs can further combined with 2 special symbols, rakaransaya and yansaya and then even further it can be combined with 2 diacritical marks (semi-consonants) and after all it will produce more than 2300 "usable" combinations used for Sinhala writing. For example consonant ka (ක) with vowel signs and non-vocalic strokes will produce following combinations;

ක, ක්, ක, කා, කැ, කෑ, කි, කී, කු, කූ, කෘ, කෲ, කෙ, කේ, කෛ, කො, කෝ, කෞ, ක්‍ර, ක්‍රා, ක්‍රැ, ක්‍රෑ, ක්‍රි, ක්‍රී, ක්‍රෙ, ක්‍රේ, ක්‍රෛ, ක්‍රො, ක්‍රෝ, ක්‍රෞ, කඃ, කඃා, කඃැ, කඃෑ, කඃි, කඃී, කඃු, කඃූ, කෙඃ, කෙඃ්, කෛඃ, කෙඃා, කෙඃෝ & කෙඃෞ.

Consequently, Sinhala characters can be divided in to three main groups. (1) Those having a normal 'x' height, (2) Those which have an ascender, similar to 'l' and (3) Those which have a descender, similar to 'g'. However, this positioning is more complicated when single or multiple vowel signs are attached to the same character (see Figure 1).

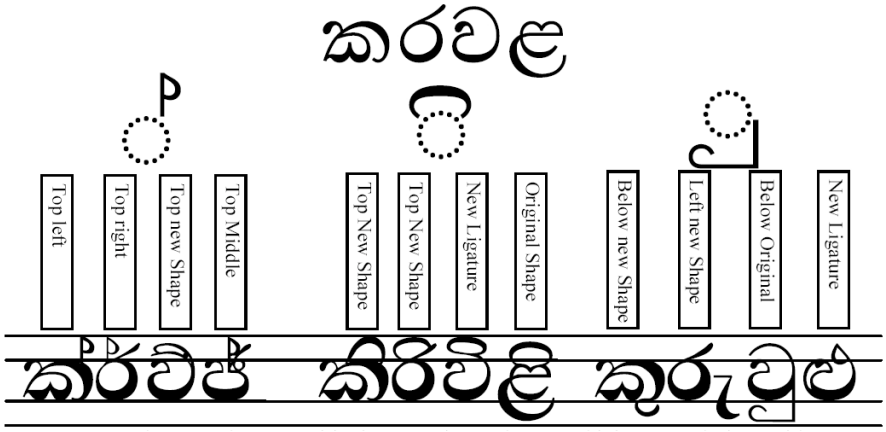

Figure 1 Shapes and New Positioning are given when combining Vowel Signs with Consonants.

3 Major Steps in Sinhala Text Processing

With the introduction of BBC microcomputers to the University of Colombo in 1982, I myself developed a set of Sinhala Bitmap fonts for computers. Using this Sinhala font set, daily TV programme schedule was transmitted for public by the National TV Station and it was the first attempt to use computers with local languages. Introduction of IBM PCs for data processing, need of developing

proper application was the major challenge to language like Sinhala where existing technologies were far behind to handle such complex scripts. The very first Sinhala word processor developed by one Chinese company in 1984 was not successful in Sri Lanka. Thereafter, there was another word processor developed by GIST in India. This was also not successful. Some local computer venders were interested in developing software for IBM compatible personal computer end up with a patent disputes over the software developed by one company against other company.

					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	ජ	ප		්	
0	0	0	1	1	1			!	1	A	Q	a	q			අ	ඣ	ඵ		ං	
0	0	1	0	2	2			"	2	B	R	b	r			ඉ	ඤ	බ		ඃ	
0	0	1	1	3	3			#	3	C	S	c	s			ඊ	ඥ	භ			
0	1	0	0	4	4			$	4	D	T	d	t			උ	ට	ම		ා	්
0	1	0	1	5	5			%	5	E	U	e	u			ඍ	ඨ	ඹ		ැ	ු
0	1	1	0	6	6			&	6	F	V	f	v			එ	ඩ	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	}			ඟ	ධ	හ		ෟ	
1	1	1	0	E	14			.	>	N	^	n	~			ච	න	ළ		ෲ	
1	1	1	1	F	15			/	?	O	_	o				ඡ	ඳ	ෆ		ු	

Figure 2 First ever encoding for Sinhala Character Set submitted for the public comment, 1990.

Since mid 1980s, a number of steps were taken by the government to formulate Sinhala language related discrepancies, such as different alphabetical orders used by different dictionaries. Due to the importance of information interchange among computers in national language and the requirement for a standard code was identified by the Information and Technology Council of Sri

Lanka (CINTEC) in 1985. One of the committee's initial endeavours was to establish a standard code for information interchange in Sinhala.

Because of the collaborative work with the Thammasat University, Thailand, 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 [1] to the public for comments and observations in March 1990.

After receiving the public comments and recommendations, the first ever standard (Figure 2) was approved by the Council of CINTEC and the Sri Lanka Standard Institute on the advice of its Working Committee for Recommending Standards for the use of Sinhala Script in Computer Technology [2][3][4][5].

3.1 Standard Keyboard for Sinhala

At this stage, it is important 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 *Wijesekera*[1] Sinhala Typewriter Keyboard, CINTEC first developed and obtained government approval for the "Extended *Wijesekera* Keyboard for Electronic Typewriters" (*see* Figure 3), the intention being the introduction of electronic typewriters then used as an interface for microcomputer input.

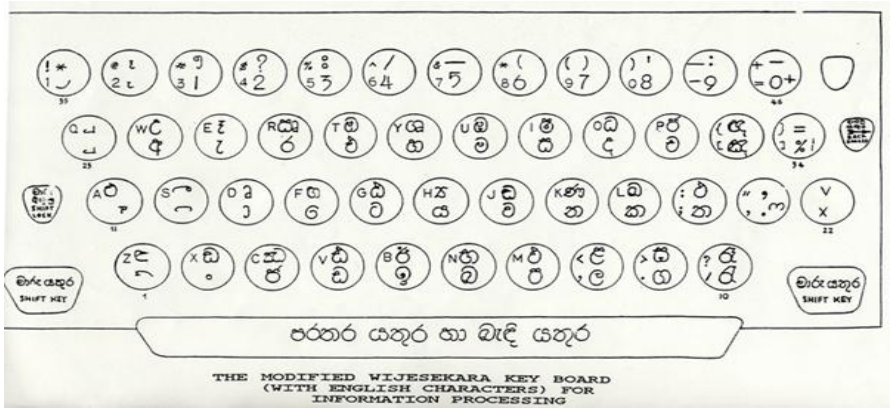

Figure 3 Extended *Wijesekara* Keyboard for Electronic typewriters used interface for microcomputer input (1989).

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

3.2 Input Method for Sinhala Character Set

Sri Lankan software industry had suffered with legal dispute on patent issues of software implementation as mentioned earlier; few individuals had started their own Sinhala word processors. In the meantime, the Institute of Computer Technology (ICT) of the University of Colombo, initiated collaborative work with Thammasat University to incorporate Sinhala capabilities for personal computer. The SLASCII standard was used to create the first ever Sinhala/English bilingual character based API called SBIOS (Sinhala BIOS) and then Sinhala keyboard layout was used with Sinhala word processor WT Ver. 1.0 (well known as "*Wadan Tharuwa*")[2] developed in early 1990s for IBM-PC computers (See Figure 4). According to *Wadan Tharuwa*, Sinhala words are input and stored letter-by-letter from left to right. This is a three-layer system for the cells that contain symbols in base level, above or below levels. The base character will be stored first, followed by the upper and then lower if the case arises. System will alarm for illegal input key sequence such as there cannot be any diacritic at the lower level after the upper level diacritic is placed.

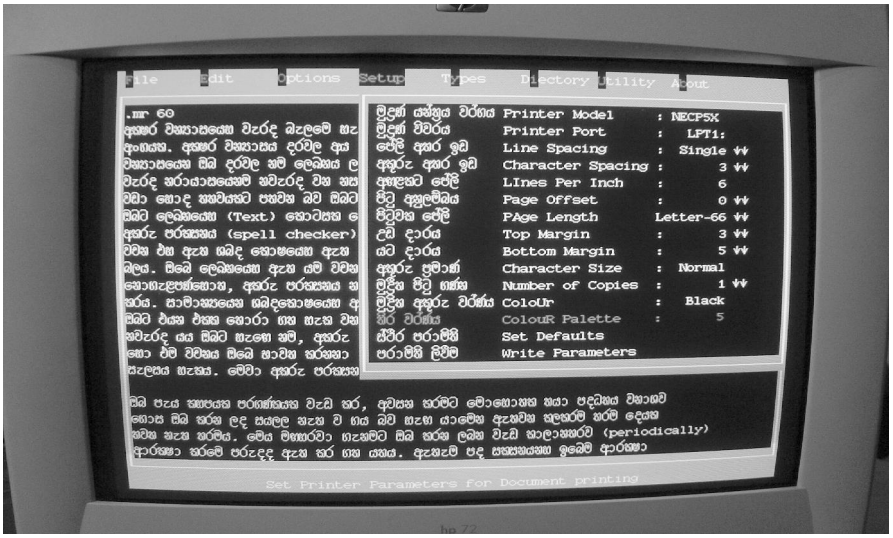

Figure 4 Sample VGA screen shot of Bi-lingual Sinhala/English Word Processor "*Wadan Tharuwa*" in early 1989. (Photo: author's collection)

[2] *Wadan Tharuwa* is a one of the earliest bi-lingual and menu-driven commercial word processor 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 the author, S T Nandasara.

SBIOS API input method provided a sequence checking mechanism to ensure the validity of the input sequence in one of the three levels of strictness; base level: pass through, lower level: basic check, and upper level: strict check.

This mechanism of sequence checking is provided for three reasons.

1. To maintain logical sorting order of the alphabet,
2. To maintain the visual correctness of the character display,
3. To maintain the correctness of the use of diacritical marks.

SBIOS has also specified the cursor movements and editing behaviour for Sinhala WT. We must move the cursor from cells to cells. However, we must skip all characters in other levels than the base level. Text deletion using the "Delete" key must also remove all characters in the current cell, including the above and below levels. Meanwhile, 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 they are stored. At a later stage, extra capabilities added to maintain a diacritical marks, mathematical and phonetic symbols for DOS operating system [6][7]. Language was selected by toggling the Shift-Ctrl key combination whenever is required.

3.3 From Bitmap to Open Font

During mid 1990s introduction of Desk Top Publishing (DTP) with PC, there was a demand for quality printing in desktop computers. The first attempt to introduce Sinhala Desktop publishing for IBM PC was available with Xerox Ventura®. One of the early outline fonts for Roman scripts was available with Xerox Ventura® for WYSIWYG (What You See Is What You Get) DTP in 1994. However, there was no whatsoever technical support given by Xerox Ventura® how non-Roman fonts to be installed with this DTP package. Thanks to the reverse engineering efforts and tag concept was used to format text and paragraphs within the package, *Athwela* was developed in 1994 to support tri-lingual (Sinhala, Tamil & English) DTP with Xerox ventura® (see Figure 5).

This move in character rendering technology with Bit Map Font technology for laser printer opened the way to the next stage of text processing, and it coincides with the emergence of new design of Sinhala character code.

Apple Macintosh® came with their early version of word processors with Sinhala language support with laser printer technology.

3.4 Current Development Platform Status

The extraction and inclusion of Sinhala code page in UCS/Unicode has made possible to connect Sinhala community to global cyberspace. However, in order to be really connected, we should do localization on proprietary or open platforms

accordingly. Currently Microsoft Windows platform is widely used in Sri Lanka. Microsoft does not provide proper Unicode support input method for Sinhala. It is planning to be released with its next version of the operating system. However, Sinhala language kit (beta version) released by Microsoft in early 2005 for Windows XP/SP2 can be used with the third party keyboard input methods.

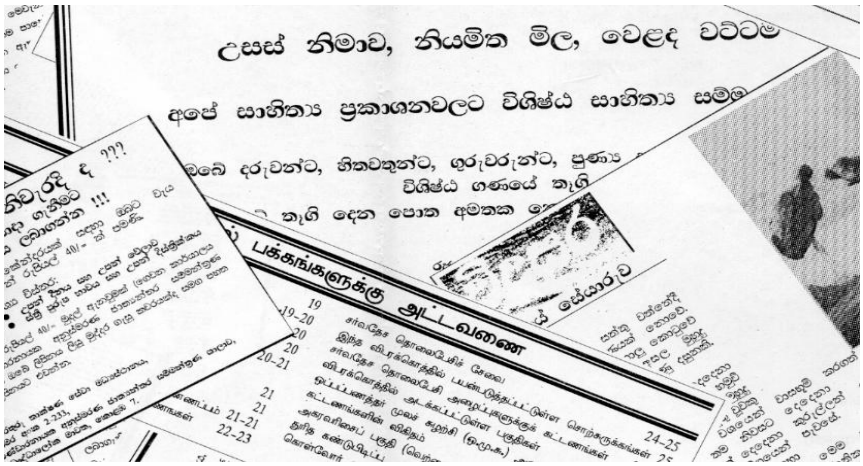

Figure 5 Sample tri-lingual laser documents from *Athwela* DTP Package

4. Conclusion

Sri Lanka has always been independent in her island history. Thus, the Sinhala language remains unique and cannot be automatically handled by technologies originated from the west. However, writing systems for various southeast and south Asian languages, such as Tamil, Thai, Khmer, and Myanmar have a lot of commonalities, and these language communities, could share a lot of common challenges to more ahead in their own text processing.

References

1. 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.
2. Working Paper, 1985. *Order of Alphabet and System of Transliteration,* CANLIT & NARESA.
3. SLS 1134:1996. *Sri Lanka Standard SLS 1134:1996-Sinhala Character Code for Information Interchange,* SLSI publication.
4. Nandasara, S. T., and Samaranayake, V. K., 1991. *A Standard Code for Information Interchange in Sinhalese,* ISO-IEC JTC1/SCL/WG2 N673, October.

5. SLS 1134:2004. *Sri Lanka Standard SLS 1134:2004-Sinhala Character Code for Information Interchange,* SLSI publication.
6. Nandasara, S.T., Sri Lanka Experience of Development of Tamil Input/Output/Display Methods, TAMILNET'97 – International Symposium, Singapore, May, 1997
7. Nandasara, S. T., Samaranayake, V. K., *Current Development of Sinhala / Tamil / English Trilingual Processing in Sri Lanka*, MLIT-2, November 7-8, Tokyo, Japan, pp. 181-192, 1997.