UCSC Technical Report 03/01

University of Colombo School of Computing

An Introduction to UNICODE for Sinhala Characters

Samaranayake, V. K., Nandasara, S. T., Dissanayake, J. B.*, Weerasinghe, A.R.,
Wijayawardhana, H.

University of Colombo School of Computing
** Sinhala Department, University of Colombo*

Abstract

This paper introduces the background, steps taken and eventual adoption of a Standard Code for the Sinhala Character set and the UNICODE/ISO10646 standard for Sinhala together with clarifications on some of the technical and linguistic issues involved in using the code for implementation.

1. Background

With the introduction of microcomputers in the early eighties, Sri Lanka too embarked on the use of computers with local language input and output. The University of Colombo developed a Sinhala screen output for television displays and went on to provide election result displays in the three languages Sinhala, Tamil and English within a few years. However, the requirement for a standard code was identified and steps were taken by the Computer and Information Technology Council of Sri Lanka (CINTEC) to establish a committee for the use of Sinhala & Tamil in Computer Technology in 1985, soon after its inception. This committee quite correctly took steps to meet the immediate need to agree on an acceptable Sinhala alphabet and an alphabetical order. Thus this committee joined with a committee appointed by the Natural Resources, Energy and Science Authority of Sri Lanka (NARESA) to form the Committee on Adaptation of National Languages in IT (CANLIT), which agreed on a unique Sinhala alphabet and alphabetical order. As for Tamil, no immediate action was taken due to the work being undertaken in India. CANLIT consisted of experts in the Sinhala language as well as IT.

It is of historic importance that a major set back for the development of Sinhala language computing was averted when an injunction on the development of Sinhala word processors taken by one developer against another based on a disputable patent was settled out of court after years of litigation.

2. The Sinhala Alphabet and Alphabetical Order

CANLIT arrived at defining the Sinhala alphabet as having 16 vowels, 2 semi consonants and 41 consonants as shown in the CINTEC publication of 1990 [2]. 13 consonant modifiers were also identified. A new character to denote "fa" (ෆ) was introduced. CANLIT also agreed on the alphabetical order as given in [2] with a slight modification as referred to in section 9 below.

It should be noted that this exercise took a representative group of language and technology experts several months to arrive at a consensus solution.

3. The Standard Sinhala Character Set

In developing the Sinhala Character set for use in IT, the work already done in Thailand for the Thai language, which is somewhat similar to Sinhala, was studied with Dr Thaweesak Koanantakool of Tammasat University, Bangkok. At this stage the aim was to develop a 7-bit code to fill the positions A0 to FF in the single byte ASCII code table (ISO 646). Work towards this was reported in [1,2] and the draft standard code was approved by the Council of CINTEC on the advice of its Working Committee for Recommending Standards for the use of Sinhala and Tamil Script in Computer Technology [2].

4. The Sinhala Standard Code for Information Interchange SLASCII

The standard as approved above (SLASCII) differs in many aspects with the Unicode for Sinhala approved later in 1998 and all such cases are discussed later on in this paper.

At this stage, it is important to indicate the development of the appropriate keyboard layout where again CINTEC took the initiative. Having agreed that a large number of Sinhala typists were using the government approved Wijesekera Keyboard, CINTEC first developed and obtained government approval for the "Extended Wijesekera Keyboard for Electronic Typewriters", the intention being the introduction of Daisywheel and Golf-ball electronic typewriters then used as an interface for microcomputer output. The draft included the new character ෆ (fa) and 3 other additional key positions as explained in [1]. As indicated later on, this layout has once again been modified for use of the 101 Key Standard English Keyboard [2].

This code table and keyboard layout were used in Wadan Tharuwa – one of the earliest commercial Sinhala word processors released in Sri Lanka and later on in Sarasavi the trilingual application package developed by the University of Colombo.

5. What is UNICODE

Text information represented in computers have traditionally been using the American Standard Code for Information Interchange (ASCII) since that standard was made for the English alphabet. This 7-bit code was able to represent 128 characters and sufficed for the purpose it was designed for. The later 8-bit extension allowed an extended ASCII representation of 256 characters, which allowed certain other mainly Roman characters to be included in the code.

As other, especially non-Latin characters were needed to be represented in the computer, there was a need for a standardization effort, so as to avoid multiple characters using the same code. Many such languages however were already supported through proprietary character encodings in application software, most notably in text processing applications. This was normally done by preserving the common codes ASCII had with the given language (e.g. digits and punctuation marks) and 'overwriting' the code points assigned to other Latin characters with the given language's 'fonts'. This meant however, that any such character could be encoded in different ways in different software, and thus could not be exchanged among applications or users.

The UNICODE standard is an attempt to get out of the chaos thus caused, and assigns a unique number (code point) for every character of every conceivable language independent of the application and the computer platform on which such textual data is to be stored and used (see Annex A for definition of terms). UNICODE is based on the ISO/IEC 10646 standard adopted by the International

Standards Organisation. The newest release of the UNICODE standard is version 3.0 and can be obtained from www.unicode.org.

Owing to the large amount of data already stored in ASCII, the first code pages of the UNICODE encoding, are equivalent to their ASCII counterparts, except that the first (empty) byte is padded at the beginning to form a 16-bit code. Thus for example, while 'A' in ASCII has the Hex code 41, it has the 16-bit UNICODE code of 0041 (Hex) represented in UNICODE as 'U+0041'.

Since UNICODE provides a unique number for each character in general, not all characters relevant to any language may be found in its own 'code page'. For instance, the digits 0 through 9 are common to many languages, but are assigned only ONCE in the first code page. Similarly, certain punctuation marks also occupy a common location in UNICODE even though they may be relevant to many languages.

Owing to its 16-bit encoding, UNICODE is theoretically able to support over 65,000 unique character code points. In fact, since this may be not enough at some point, there is UTF-16 extension mechanism in UNICODE that will allow almost 1 million character code points to be assigned for future expansion. Part of this space is also reserved as 'private' in order to allow hardware and software developers to assign codes temporarily for various purposes.

In addition to this 16-bit encoding, UNICODE also provides an 8-bit transformation into UTF-8. This results in a variable length byte encoding that is able to still uniquely represent every known UNICODE character represented so far. Apart from making the characters in the ASCII code correspond exactly to the original ASCII, it also allows UNICODE characters to be used with existing legacy software. Unicode is the official way to implement the ISO/IEC 10646 standard.

While UNICODE specifies a unique code point (number) for each character of any language, it does NOT specify the actual shape of the character that is thus represented. While for demonstration purposes, a representative glyph image is usually shown in the code, what it really represents is its abstract form using a unique upper case name such as "LATIN CHARACTER CAPITAL A" or "SINHALA LETTER AYANNA".

UNICODE provides for both 'precomposed characters' AND 'composite character sequences' for representing characters. Precomposed characters are those taking a single character position, while composite character sequences are where a base character code may be followed by codes for one or more 'non-spacing marks', which 'modify' the character glyph without taking 'additional character space'. The 'SINHALA SIGN AL-LAKUNA is an example of a non-spacing mark in the Sinhala code page.

The URL http://www.unicode.org/standard/where/ indicates how one could find a specific character in the code chart. Characters in Unicode are grouped into blocks. For example Sinhala is in the code page 0D80 to 0DFF [4].

The Charts do not specify the exact shape. They only provide a representative shape for identification. Characters may also take on different shapes in different contexts. Furthermore, the character you are looking for may be represented as a sequence of code points.

6. The Proposals for ISO/UNICODE 10646

The existence of a draft code for Sinhala proposed to the ISO/Unicode 10646 Working Group, by researchers based in Europe was first brought to our notice in the late eighties when an IBM delegation visited the Institute of Computer Technology (ICT) of the University of Colombo and showed the draft code table. This represented a distorted Sinhala character set with several glaring errors and omissions. For example some of the major shortcomings were :

(i) Inclusion of a set of symbols to represent numerals 0-9 based on an obscure document.

(ii) The shift of the vowels (ඏ) and (ඐ) from its natural location between (ඎ) and (එ) to the end of the character set in order to be consistent with Indic languages. (It should be noted that Sinhala is not a subset or equivalent set to any of the Indic languages).

(iii) Non inclusion of some of the important characters such as ඳ

Immediate steps were taken to request ISO directly and through the Sri Lanka Standards Institute (SLSI) to suspend approval of the draft until representations were made by CINTEC and SLSI. The work in Sri Lanka regarding the standard code was thereafter speeded up.

It is interesting to note that there were no Sinhalese or even Sri Lankans in the Working Committee as then constituted by ISO/Unicode.

The seven bit Draft Standard SLASCII was submitted to the working group (WG) and comments on this draft were then received from the members of the WG. Members of the CINTEC committee were included in the ISO/Unicode WG and much correspondence followed. Meanwhile a Unicode based Sinhala Standard was formulated by the CINTEC and thereafter by a SLSI Committee. Public comments were also obtained, as is the case with any Sri Lanka Standard. Finally the Sri Lankan Standard SLS 1134:1996 was approved and published in 1996 [3].

In 1997, CINTEC with the assistance of NARESA sent two of the authors to the 1997 working group meeting held in Crete, Greece where the draft Sinhala Code was discussed intensively. Our two delegates argued for the draft submitted by Sri Lanka opposing several competing proposals from UK, Ireland and the USA. With the support of the majority of delegates to the WG, the Sri Lankan draft was finally agreed on with slight modifications. This was ratified at the 1998 meeting of the WG held at Seattle, USA and the Sinhala Code Chart was included in Unicode Version 3.0 [4]. The SLSI 1134 was also accordingly modified.

7. UNICODE Code Page for Sinhala

The UNICODE chart table as appearing in UNICODE version 3.0 is reproduced below, indicating code positions, abstract character names and explanatory notes. This can be downloaded from the UNICODE Consortium website at the URL: http://www.unicode.org/charts/PDF/U0D80.pdf

Sinhala

Range: 0D80–0DFF

This file contains an excerpt from the character code tables and list of character names for *The Unicode Standard, Version 3.0.*

Disclaimer

The shapes of the reference glyphs used in these code charts are not prescriptive. Considerable variation is to be expected in actual fonts.

For a complete understanding of the use of the characters contained in this excerpt file, please consult the appropriate sections of The Unicode Standard, Version 3.0 (ISBN 0–201–61633–5), as well as the Unicode Technical Reports and the Unicode Character Database, which are available online.

See ftp://ftp.unicode.org/Public/UNIDATA/UnicodeCharacterDatabase.html and http://www.unicode.org/unicode/reports

A thorough understanding of the information contained in these additional sources is required for a successful implementation.

Fonts

The fonts used in these charts were provided to the Unicode Consortium by a number of different font designers

See http://www.unicode.org/unicode/uni2book/u2fonts.html for a list.

Terms of Use

These charts are provided as a convenient online reference to the character contents of the Unicode Standard, Version 3.0. Proper Unicode support requires considerably more than just providing glyphs for characters, and requires consulting the Unicode Standard and the Unicode Technical Reports.

You may freely use these code charts for personal or internal business uses only. You may not incorporate them into any product or publication, or otherwise distribute them without express written permission from the Unicode Consortium.

The information in this file may be updated from time to time. The Unicode Consortium is not liable for errors or omissions in this excerpt file or the standard itself. Information on characters added to the Unicode Standard since the publication of version 3.0 as well as on characters currently being considered for addition to the Unicode Standard can be found on the Unicode website.

See http://www.unicode.org/pending/pending.html and http://www.unicode.org/unicode/alloc/Pipeline.html.

0D80 Sinhala 0DFF

	0D8	0D9	0DA	0DB	0DC	0DD	0DE	0DF
0		ඐ 0D90	ච 0DA0	ධ 0DB0	ව 0DC0	◌ැ 0DD0		
1		එ 0D91	ඡ 0DA1	න 0DB1	ශ 0DC1	◌ෑ 0DD1		
2	◌ං 0D82	ඒ 0D92	ජ 0DA2		ෂ 0DC2	◌ි 0DD2		◌ෲ 0DF2
3	◌ඃ 0D83	ඓ 0D93	ඣ 0DA3	ඳ 0DB3	ස 0DC3	◌ී 0DD3		◌ෳ 0DF3
4		ඔ 0D94	ඤ 0DA4	ප 0DB4	හ 0DC4	◌ු 0DD4		෴ 0DF4
5	අ 0D85	ඕ 0D95	ඥ 0DA5	ඵ 0DB5	ළ 0DC5			
6	ආ 0D86	ඖ 0D96	ඦ 0DA6	බ 0DB6	ෆ 0DC6	◌ූ 0DD6		
7	ඇ 0D87		ට 0DA7	භ 0DB7				
8	ඈ 0D88		ඨ 0DA8	ම 0DB8		◌ෘ 0DD8		
9	ඉ 0D89		ඩ 0DA9	ඹ 0DB9		ෙ◌ 0DD9		
A	ඊ 0D8A	ක 0D9A	ඪ 0DAA	ය 0DBA	◌් 0DCA	ෙ◌් 0DDA		
B	උ 0D8B	ඛ 0D9B	ණ 0DAB	ර 0DBB		ෛ◌ 0DDB		
C	ඌ 0D8C	ග 0D9C	ඬ 0DAC			ෙ◌ා 0DDC		
D	ඍ 0D8D	ඝ 0D9D	ත 0DAD	ල 0DBD		ෙ◌ා් 0DDD		
E	ඎ 0D8E	ඞ 0D9E	ථ 0DAE			ෙ◌ෟ 0DDE		
F	ඏ 0D8F	ඟ 0D9F	ද 0DAF		◌ා 0DCF	◌ෟ 0DDF		

Various signs

0D82 ◌ං SINHALA SIGN ANUSVARAYA
= anusvara

0D83 ◌ඃ SINHALA SIGN VISARGAYA
= visarga

Independent vowels

0D85 අ SINHALA LETTER AYANNA
= sinhala letter a

0D86 ආ SINHALA LETTER AAYANNA
= sinhala letter aa

0D87 ඇ SINHALA LETTER AEYANNA
= sinhala letter ae

0D88 ඈ SINHALA LETTER AEEYANNA
= sinhala letter aae

0D89 ඉ SINHALA LETTER IYANNA
= sinhala letter i

0D8A ඊ SINHALA LETTER IIYANNA
= sinhala letter ii

0D8B උ SINHALA LETTER UYANNA
= sinhala letter u

0D8C ඌ SINHALA LETTER UUYANNA
= sinhala letter uu

0D8D ඍ SINHALA LETTER IRUYANNA
= sinhala letter vocalic r

0D8E ඎ SINHALA LETTER IRUUYANNA
= sinhala letter vocalic rr

0D8F ඏ SINHALA LETTER ILUYANNA
= sinhala letter vocalic l

0D90 ඐ SINHALA LETTER ILUUYANNA
= sinhala letter vocalic ll

0D91 එ SINHALA LETTER EYANNA
= sinhala letter e

0D92 ඒ SINHALA LETTER EEYANNA
= sinhala letter ee

0D93 ඓ SINHALA LETTER AIYANNA
= sinhala letter ai

0D94 ඔ SINHALA LETTER OYANNA
= sinhala letter o

0D95 ඕ SINHALA LETTER OOYANNA
= sinhala letter oo

0D96 ඖ SINHALA LETTER AUYANNA
= sinhala letter au

Consonants

0D9A ක SINHALA LETTER ALPAPRAANA KAYANNA
= sinhala letter ka

0D9B ඛ SINHALA LETTER MAHAAPRAANA KAYANNA
= sinhala letter kha

0D9C ග SINHALA LETTER ALPAPRAANA GAYANNA
= sinhala letter ga

0D9D ඝ SINHALA LETTER MAHAAPRAANA GAYANNA
= sinhala letter gha

0D9E ඞ SINHALA LETTER KANTAJA NAASIKYAYA
= sinhala letter nga

0D9F ඟ SINHALA LETTER SANYAKA GAYANNA
= sinhala letter nnga

0DA0 ච SINHALA LETTER ALPAPRAANA CAYANNA
= sinhala letter ca

0DA1 ඡ SINHALA LETTER MAHAAPRAANA CAYANNA
= sinhala letter cha

0DA2 ජ SINHALA LETTER ALPAPRAANA JAYANNA
= sinhala letter ja

0DA3 ඣ SINHALA LETTER MAHAAPRAANA JAYANNA
= sinhala letter jha

0DA4 ඤ SINHALA LETTER TAALUJA NAASIKYAYA
= sinhala letter nya

0DA5 ඥ SINHALA LETTER TAALUJA SANYOOGA NAAKSIKYAYA
= sinhala letter jnya

0DA6 ඦ SINHALA LETTER SANYAKA JAYANNA
= sinhala letter nyja

0DA7 ට SINHALA LETTER ALPAPRAANA TTAYANNA
= sinhala letter tta

0DA8 ඨ SINHALA LETTER MAHAAPRAANA TTAYANNA
= sinhala letter ttha

0DA9 ඩ SINHALA LETTER ALPAPRAANA DDAYANNA
= sinhala letter dda

0DAA ඪ SINHALA LETTER MAHAAPRAANA DDAYANNA
= sinhala letter ddha

0DAB ණ SINHALA LETTER MUURDHAJA NAYANNA
= sinhala letter nna

0DAC ඬ SINHALA LETTER SANYAKA DDAYANNA
= sinhala letter nndda

0DAD ත SINHALA LETTER ALPAPRAANA TAYANNA
= sinhala letter ta

0DAE ථ SINHALA LETTER MAHAAPRAANA TAYANNA
= sinhala letter tha

0DAF ද SINHALA LETTER ALPAPRAANA DAYANNA
= sinhala letter da

0DB0 ධ SINHALA LETTER MAHAAPRAANA DAYANNA
= sinhala letter dha

0DB1 න SINHALA LETTER DANTAJA NAYANNA
= sinhala letter na

0DB2 ▨ <reserved>

0DB3 ඳ SINHALA LETTER SANYAKA DAYANNA
= sinhala letter nda

0DB4 ප SINHALA LETTER ALPAPRAANA PAYANNA
= sinhala letter pa

0DB5 ඵ SINHALA LETTER MAHAAPRAANA PAYANNA
= sinhala letter pha

0DB6 බ SINHALA LETTER ALPAPRAANA BAYANNA
= sinhala letter ba

0DB7 භ SINHALA LETTER MAHAAPRAANA BAYANNA
= sinhala letter bha

0DB8 ම SINHALA LETTER MAYANNA
= sinhala letter ma

0DB9 ඹ SINHALA LETTER AMBA BAYANNA
= sinhala letter mba

0DBA ය SINHALA LETTER YAYANNA
= sinhala letter ya

0DBB ර SINHALA LETTER RAYANNA
= sinhala letter ra

0DBC ▨ <reserved>

0DBD ල SINHALA LETTER DANTAJA LAYANNA
= sinhala letter la
• dental

0DBE ▨ <reserved>

0DBF ▨ <reserved>

0DC0 ව SINHALA LETTER VAYANNA
= sinhala letter va

0DC1 ශ SINHALA LETTER TAALUJA SAYANNA
= sinhala letter sha

0DC2 ෂ SINHALA LETTER MUURDHAJA SAYANNA
= sinhala letter ssa
• retroflex

0DC3 ස SINHALA LETTER DANTAJA SAYANNA
= sinhala letter sa
• dental

0DC4 හ SINHALA LETTER HAYANNA
= sinhala letter ha

0DC5 ළ SINHALA LETTER MUURDHAJA LAYANNA
= sinhala letter lla
• retroflex

0DC6 ෆ SINHALA LETTER FAYANNA
= sinhala letter fa

Sign

0DCA ◌් SINHALA SIGN AL-LAKUNA
= virama

Dependent vowel signs

0DCF ◌ා SINHALA VOWEL SIGN AELA-PILLA
= sinhala vowel sign aa

0DD0 ◌ැ SINHALA VOWEL SIGN KETTI AEDA-PILLA
= sinhala vowel sign ae

0DD1 ◌ෑ SINHALA VOWEL SIGN DIGA AEDA-PILLA
= sinhala vowel sign aae

0DD2 ◌ි SINHALA VOWEL SIGN KETTI IS-PILLA
= sinhala vowel sign i

0DD3 ◌ී SINHALA VOWEL SIGN DIGA IS-PILLA
= sinhala vowel sign ii

0DD4 ◌ු SINHALA VOWEL SIGN KETTI PAA-PILLA
= sinhala vowel sign u

0DD5 ▨ <reserved>

0DD6 ◌ූ SINHALA VOWEL SIGN DIGA PAA-PILLA
= sinhala vowel sign uu

0DD7 ▨ <reserved>

0DD8 ◌ෘ SINHALA VOWEL SIGN GAETTA-PILLA
= sinhala vowel sign vocalic r

0DD9 ◌ෙ SINHALA VOWEL SIGN KOMBUVA
= sinhala vowel sign e

0DDA ◌ේ SINHALA VOWEL SIGN DIGA KOMBUVA
= sinhala vowel sign ee
≡ 0DD9 ◌ෙ 0DCA ◌්

0DDB ◌ෛ SINHALA VOWEL SIGN KOMBU DEKA
= sinhala vowel sign ai

0DDC ◌ො SINHALA VOWEL SIGN KOMBUVA HAA AELA-PILLA
= sinhala vowel sign o
≡ 0DD9 ◌ෙ 0DCF ◌ා

0DDD ◌ෝ SINHALA VOWEL SIGN KOMBUVA HAA DIGA AELA-PILLA
= sinhala vowel sign oo
≡ 0DDC ◌ො 0DCA ◌්

0DDE ◌ෞ SINHALA VOWEL SIGN KOMBUVA HAA GAYANUKITTA
= sinhala vowel sign au
≡ 0DD9 ◌ෙ 0DDF ◌ෟ

0DDF ◌ෟ SINHALA VOWEL SIGN GAYANUKITTA
= sinhala vowel sign vocalic l

Additional dependent vowel signs

0DF2 ◌ෲ SINHALA VOWEL SIGN DIGA GAETTA-PILLA
= sinhala vowel sign vocalic rr

0DF3 ◌ෳ SINHALA VOWEL SIGN DIGA GAYANUKITTA
= sinhala vowel sign vocalic ll

8. Special Features and Justification

In the Unicode chart for Sinhala, care has been taken to arrange the vowels, consonants and consonant modifiers in such a way as to facilitate automatic code based sorting as far as possible. For example, කෙ is coded 0D9A+0DD9 while කො is coded 0D9A+0DDC so that කෙ is followed by කො and not for instance by ගෙ (which has the code 0D9C+0DD9).

The code positions 0D97-0D99 are left blank for new vowels that may be introduced in the future if necessary. The code positions 0DB2, 0DBC, 0DBE and 0DBF have been reserved to facilitate transliteration between Sinhala and Tamil due to the one to many correspondence between Sinhala and Tamil. Codes 0DC7-0DC9 are reserved for additional consonants that may be introduced. In addition, codes 0DD5 and 0DD7 are for accommodating alternate forms of 0DD4 and 0DD6 respectively. Code positions between 0DE0 and 0DE1 are available for future expansion. Use of all these blank positions requires acceptance by ISO/Unicode. All other unused code positions are reserved by Unicode.

In writing Sinhala, the vowel sign "AL-LAKUNA" at 0DCA is used in two forms '්' and '්' . For example ක (ka) with the "AL-LAKUNA" will be shown as ක් while ම (ma) with the "AL-LAKUNA" will be shown as ම්. However, in Unicode there is a unique code for "AL-LAKUNA" at 0DCA and according to the combined consonant, the glyph will differ. The same applies to the vowel signs "KETTI PAA-PILLA" 0DD4 and "DIGA PAA-PILLA" 0DD6 represented respectively by 'ු' or 'ු' and 'ූ' or 'ූ' according to the combined consonant.

There are a few more instances in the usage of Sinhala where the standard practice of writing has been followed. They are:

0dbb (ර) + 0dd0 (ැ) gives රැ and not රැ as expected
0dbb (ර) + 0dd1 (ෑ) gives රෑ and not රෑ as expected
0dbb (ර) + 0dd4 (ු) gives රු
0dbb (ර) + 0dd6 (ූ) gives රූ
0dc5 (ළ) + 0dd4 (ු) gives ළු
0dc5 (ළ) + 0dd6 (ූ) gives ළූ

These are conventions adopted and the corresponding glyphs should represent these.

Another issue is the non inclusion of "YANSAYA" (්‍ය) and "RAKARANSHAYA" (්‍ර) in the code chart and how they are represented. These two characters are considered consonant modifiers different from others, as they do not represent vowels but a combination of vowel '්' and consonant 'ය' and vowel '්' and consonant 'ර' respectively. Although the keyboard will have these for ease of use, the respective glyphs will have to be constructed for the codes as shown in the example below:

ක +' + ZWJ + ය = 0D9A+0DCA+ZWJ+0DBA to give the glyph ක්‍ය
ත +' + ZWJ + ර = 0DAD+0DCA+ZWJ+0DBB to give the glyph ත්‍ර

The "REPAYA" ( ) while no longer used in standard Sinhala (and thus not included in the chart table, can be coded as the sequence "ර +' + ZWJ" when required.

The code for ZWJ is not in the Sinhala code chart but elsewhere in Unicode. For example, 0DBB + ZWJ + 0DCA + 0DB8 would give the composite character, ර්‍ම.

Annex B details the main types of Sinhala UNICODE character sequences that are used to represent composite Sinhala characters.

9. Aspects outside the UNICODE standard

The following issues are strictly outside the scope of the UNICODE standard. They however affect the implementation and use of Sinhala UNICODE. They are needed to provide the main interfaces needed by humans to interact with the computer. Fundamentally they involve the development of specialised software called Device Drivers traditionally supplied by font developers and vendors.

(a) Input

The keyboard driver, the main form of inputting text to a computer, is responsible for providing a valid translation between the keys represented on a keyboard and the internal UNICODE representation. This is not as straight forward to achieve for Sinhala as for Latin-based languages, where the correspondence between keys on the keyboard and the code point to be generated relates almost one-to-one. For example, when the keys marked 'A' and 'B' are pressed in succession, the code points 0041 and 0042 are generated. However, in Sinhala, when the keys 'ෙ' and 'ක' are pressed what should be generated is the opposite ordering u0D9A followed by 0DD9 (This is even more clear when the keys 'ෙ', 'ක' and 'ා' are pressed in succession. The code generated should be 0D9A followed by u0DDC.

All this however does not directly affect the keyboard layout itself. The keyboard layout does not need to change to accommodate Sinhala UNICODE. The two common kinds of Sinhala keyboard in use today, the Wijesekera (or Extended Wijesekara) keyboard and the more variable Phonetic keyboard, will continue to be the main method used to input Sinhala characters. It is however possible to enhance these in the light of the provisions of UNICODE as well as the enhanced keyboard support provided for non-Lain character input by the ALT-GR (LEFT-ALT) key on modern keyboards.

Other forms of input such as Optical Character Recognition (OCR) and Speech Recognition will also need to consider the Sinhala UNICODE representation in order to provide valid storage and exchange of Sinhala text.

(b) Rendering

The two primary methods of rendering text stored in a computer for human consumption are displaying it on the screen and printing it on a printer.

The display driver is responsible for translating the internally represented (UNICODE) code into recognizable Sinhala character glyphs on the screen. In languages such as Sinhala this is not as straight forward as in Latin-based languages where the correspondence between codes and glyphs is nearly one-to-one. For Sinhala for instance, the appearance of a code 0DDD immediately followed by 0D9A needs to cause the display driver to display on the screen the Sinhala composite ‘කෝ’.

A similar role is played by the printer driver, which converts valid Sinhala UNICODE character sequences to a form suitable for printing.

10. Further work in progress

With the availability of Unicode for Sinhala the scope for application development in Sinhala has increased. However, the definition of the shaping engine support in operating systems need to be standardized urgently in order that further problems do not occur in the implementation of Unicode complaint fonts for universal use.

In addition areas such as translation, speech and text recognition among others need to be encouraged in order to catch up on lost time, and some of these are being addressed at the research level at the UCSC.

References

1. V.K. Samaranayake, J.B. Dissanayake and S.T. Nandasara – A standard Code for Sinhala Characters – Paper presented at the 9th Annual Sessions of the Computer Society of Sri Lanka. (1989).
2. S.T. Nandasara, J.B. Dissanayake, V.K. Samaranayake, E.K. Seneviratne and T. Koannantakool – 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. (March 1990)
3. Sri Lanka Standard SLS 1134:1996 – Sinhala Character Code for Information Interchange (SLSI publication 1996)
4. The Unicode Standard 3.0 (www.unicode.org) (Addison-Wesley Pub Co.), ISBN 02001616335.
5. S.T. Nandasara, K.Y. Leong, V.K. Samaranayake and T.W. Tan – Trilingual Sinhala Tamil English National Web Site of Sri Lanka, INET97, (http://www.isoc.org/inet97/proceedings/EI/E1_3.HTM) (1997)
6. S.T. Nandasara and V.K. Samaranayake – A Standard Code for Information Interchange in Sinhalese : Proceedings of the International Conference on the Standardization of Asian Languages, CICC, Tokyo, Japan (1994).
7. S.T. Nandasara and V.K. Samaranayake – Current developments of Sinhala/Tamil/English Trilingual Processing in Sri Lanka. Paper presented at the Second International Symposium on Standardization of Multilingual Information Technology. November 1997, Tokyo, Japan.

Annex A: Definition of Terms Relating to Character Representation

Character: An abstract representation of a letter of a given written language
e.g. Latin-letter-uppercase-A (not 'A' itself)
Sinhala-letter-ayanna (not 'a')

Code Point: A number assigned to represent an abstract character in a computer
e.g. Decimal 65 (Hex 41) represents Latin-letter-uppercase-A in the ASCII coding scheme
Hexadecimal 0D85 represents Sinhala-letter-ayanna in the UNICODE coding scheme

Code: A set of code points defining a coding scheme
e.g. EBCDIC, ASCII, UNICODE

Glyph: The graphical shape of an abstract character
e.g. Latin-letter-uppercase-A has the glyph 'A'
Sinhala-letter-ayanna has the glyph 'අ'

Font: A set of graphical shapes (typeface) representing a particular style of print
e.g. Latin-letter-uppercase-A in Times font takes the shape 'A'; in Courier font takes the shape `'A'` and in Comic Sana font takes the shape 'A'.
Soon Sinhala will have equivalent font styles…

Composites: A base character followed by a sequence of modifiers belonging together as a composite character in a given language
e.g. In Sinhala 'ෙ' + 'ක' + 'ා' + '්' → කෝ

Conjuncts: Conjunct consonants (bendi akuru) are composites formed by joining together two independent characters (e.g. ක්‍ෂ)

Ligatures: These are combined character glyphs that cannot be separated (e.g. කු). These are provided as unique characters with their own code point in UNICODE

Some Implications:

1. All UNICODE Sinhala fonts will be *real* fonts – *not encodings*, as they now are. So, even though their styles may be different, all characters (e.g. Sinhala-letter-ayanna) will have the same encoding in all fonts.
2. The sequence in which we **input** Sinhala characters through the keyboard may be actually *different* from the way it is represented (**stored**) inside the computer. e.g. though we store 'ක' as 'ක'; we store 'ෙ' + 'ක' + 'ා' as 'ක' + 'ෙ ා'. The way we type will be more or less the same as we've been used to (e.g. using Wijesekera keyboard and the left-to-right, bottom-to-top convention).

Annex B: Representing Sinhala Characters in UNICODE

Sinhala characters including composites can be represented using 1, 2 or 3, 16-bit codes in UNICODE.

1 Code letters: අ එ ඉ ක ඹ ඞ ඌ ශ ෆ ඬ ඎ
2 Code composites: කා කං කී කු කෘ කෟ කෙ කෲ කෛ කො කෞ
3 Code composites: කාං කීං කුං කෲං කෑං

The following are three types of special composite characters that are represented internally by employing the special zero-width joiner (ZWJ) character.

Character ‘modifiers’:
ත්‍ය → ‘ත’ + ‘ ් ’ + ZWJ + ‘ය’
ක්‍ර → ‘ක’ + ‘ ් ’ + ZWJ + ‘ර’ (so for instance ක්‍රීං will consist of 6 codes)

Conjunct consonants (bendi akuru):
→ ‘න’ + ‘ ් ’ + ZWJ + ‘ද’
ක්‍ෂ → ‘ක’ + ‘ ් ’ + ZWJ + ‘ෂ’ (so for instance ක්‍ෂො will consist of 5 codes)

Non-typical ‘modifiers’:
ර්‍ග → ‘ර’ + ‘ ් ’ + ZWJ + ‘ග’
ද්‍ධ → ‘ද’ + ‘ ් ’ + ZWJ + ‘ධ’ (so for instance ද්‍ධං will consist of 5 codes)

Note:

(a) The above does not represent the keyboard sequence needed to input these sequences (as mentioned before, this will be according to the usual conventions and keyboards).
(b) The character-shapes ඬ, ඳ, and ඟ are ligatures that have their own code points in UNICODE and are not considered conjunct consonants (bendi akuru).