Recent Developments in Local Language Computing in Sri Lanka

Gihan Dias
University of Moratuwa
gihan@uom.lk

Aruni Goonetilleke
ICT Agency of Sri Lanka
aruni@icta.lk

ABSTRACT

Computing in Sinhala and Tamil has evolved from a vision to a reality in the past twenty years.

We describe the work done to make this happen at the ICT Agency and other organisations, starting from the initial work in the 1980s, but concentrating on the work done in the past five years.

The work comprises the encoding of Sinhala and Tamil, keyboards, fonts, collation, terminology and tools.

Hardware, operating system and applications support, as well as awareness and dissemination activities are also described.

1. INTRODUCTION

We are currently at the beginning of a new era of computing in Sri Lanka - one where our people do not have to use a foreign language to benefit from information technology. Computers, phones and the Internet now work in our own languages, and very soon we may forget the era where they were only in English.

Popular computing platforms such as MS-Windows and Linux, as well as some phones, now support our national languages - Sinhala and Tamil. We can create and exchange electronic documents, spreadsheets and databases in our languages. We now have operating systems and applications with a localised interface, enabling people without English knowledge to use computers easily. The world-wide web and other information sources are steadily gaining more and more local-language content.

This journey has been a long one, starting in the early 1980's, and is not yet complete.

This paper briefly looks at the history of local-language computing in Sri Lanka, and developments in the last five years - in Tamil and Sinhala - in more detail. We focus on infra-structural developments, and do not consider localised applications and content, which will be covered in another paper.

2. HISTORY

The need for computers to support Sinhala and Tamil was identified early, but the available hardware did not support these scripts. Two pioneering efforts by DMS and Metropolitan to introduce Sinhala computing in the 1980s did not gain widespread use. The main problem was that bit-mapped displays and printers were not in widespread use at that time, and firmware-based fonts needed to be installed on output devices.

Thereafter, Wijeya Graphics produced a Sinhala font for the Macintosh, which was widely used in publishing. The University of Colombo developed a Sinhala screen output for television displays that was used to provide election result displays in the three languages Sinhala, Tamil and English.

2.1 The Sinhala Alphabet and Alphabetical Order

The requirement for a standard code to represent Sinhala letters in computing was identified in the mid-eighties and the Computer and Information Technology Council (CINTEC) together with the Natural Resources, Energy and Science Authority (NARESA) formed the Committee on Adaptation of National Languages in IT (CANLIT), which agreed on a unique Sinhala alphabet and alphabetical order [1]. No immediate action was taken on Tamil, due to the work being undertaken in India. A Sinhala encoding was defined as Sri Lanka Standard 1134 [2] in 1996. This encoding followed the basic principles of the Indian ISCII encoding, modified to accommodate the special features of the Sinhala script.

2.2 Unicode Compatible Sinhala Code

Unicode is a standard encoding to represent all the world's scripts, including Sinhala and Tamil [3]. In 1997, Sri Lanka submitted a proposal for the Sinhala character code to Unicode, which competed with proposals from UK, Ireland and the USA. The Sri Lankan proposal was accepted by Unicode with slight modifications [4].

However, Unicode Sinhala was very slow in its adoption, and there were no implementations for over five years. Some of the perceived shortcomings of the Sinhala encoding, described in [5] were:

- lack of encodings for bandi akuru such as ඹ්,
- lack of encodings for the yansaya (්ය), rakaransaya and rephaya,
- lack of guidance on the use of multiple vowel modifiers and
- lack of guidance on the encoding of non-standard letters, such as ඛ, ඍ and ඓ.

The encodings for the above were documented as a revision to SLS1134 [6]. This standard provides for the complete and accurate encoding of all contemporary and classical Sinhala texts.

Details of the Unicode encoding of Sinhala can be found in [7]. More information on this and the other initiatives presented in Section 2 are in [5].

2.3 Sinhala Computer Keyboard

A number of different computer keyboard layouts, including "phonetic" layouts based on the English keyboard, were popular in Sri Lanka [8]. However, the government approved Wijesekera keyboard is used by many Sinhala typists. Therefore, it was decided to base the Standard Sinhala computer keyboard on the Wijesekera layout, with a few modifications to take advantage of computer processing [6].

2.4 Fonts

Unlike legacy fonts, in which a series of glyphs (pictures of letters or modifiers) are displayed left-to right across a line, Unicode Sinhala has a complex relationship between the codes stored in the computer and the shapes displayed on the screen or printer.

For example, although the kombuwa (ෙ) is displayed to the left of a consonant, in Unicode it is stored *after* the consonant. The display driver and font are responsible for handling this correctly. Conjunct letters such as ඤ are represented by a sequences of codes which are displayed by a single glyph [5].

Only recent technologies such as OpenType [9] can correctly handle such operations, and they still need operating system support. The ICTA Local Language Working Group liaised with companies such as Microsoft to ensure that their fonts and operating systems correctly supported Sinhala and Tamil. While Tamil was supported in Mirosoft Windows XP, Sinhala was natively supported only in Windows Vista.

3. TAMIL INITIATIVES

Initially, local language efforts in Sri Lanka were not directed towards Tamil, as it was expected that this work will be carried out in India, and Tamil Nadu in particular. However, it was observed that progress in India was slow, and that Indian national-level initiatives and Tamil Nadu initiatives diverged.

Also we realised that the use of Tamil in Sri Lanka often diverged considerably from that in Tamil Nadu, and that independent standards and initiatives are needed in this country.

3.1 Tamil Keyboard

Tamil uses a number of keyboards such as the typewriter-based Renganathan, Romainsed and the Indian Govt. approved Inscript layouts [8]. However, a group of experts proposed an optimised Tamil keyboard (Tamil99 Keyboard) at the TamilNet conference in 1999 [10], which was approved by the Govt. of Tamil Nadu in Government Order 17 of 1999.

The ICTA Local Language working group, after studying the available options, recommended in 2004 that the Tamil99 keyboard be adopted for use in Sri Lanka. This was accepted by the ICTA and the Ministry of Education, and Tamil99 keyboards were produced and distributed to government offices and schools.

However, there was resistance to this keyboard, especially by professional typists, and the adoption was low. Therefore, the ICTA decided to adopt a generally acceptable keyboard layout.

A consultative process was followed in agreeing on the standard keyboard layout. ICTA set up a team comprising Tamil linguists and those proficient in the Tamil language to work on the keyboard layout.

The team first held a workshop in October 2006 for stakeholders and the consensus was to use the Renganathan keyboard layout, which is based on the Tamil typewriter and extensively used on computers. However, about 10 variations of the

Renganathan layout are in use. After meetings with other key stakeholders and further analysis, a layout based on Renganathan with some modifications was defined. Key sequences are defined on the "type as you write" method. All Tamil letters and symbols may be typed using this layout.

The proposed a layout was presented to and accepted by a wider audience in January 2007.

This layout was consequently included in the Sri Lanka Standard SLS 1326:2008 – Tamil Character Code for Information Interchange [11].

3.2 Sri Lanka Standard Tamil Character Code

Although the Tamil encoding defined by Unicode [3] is acceptable, a need was felt for an official standard to which implementations may adhere. Therefore, the Tamil sub-committee of the LLWG embarked on a programme to define the Tamil encoding, which was subsequently standardised as SLS 1326:2008 [11].

The encoding of Tamil is similar to other Indic languages, comprising consonants (both Tamil letters and Grantha letters used in Tamil), vowels and modifiers.

The āytam (ஃ) occurs only after a short initial letter and before a hard consonant and never at the beginning or end of words. However it is in contemporary use to represent 'f' and 'ph' sounds in Tamil, e.g. ஃபக்ஸ் (fax). Here the letter ப (PA) is modified as "FA" when the āytam appears front of the "PA".

The Grantha conjunct syllable ஸ்ரீ (shri) and the Grantha letter க்ஷ (ksha) are commonly used in Tamil, but are not assigned Unicode code points. The character sequences used to produce these, as well as all other Tamil letters and syllables are defined in the standard.

SLS1326 also defines Tamil numerals, Tamil symbols and the Tamil OOM sign.

3.3 Tamil Collation

Collation is the order in which a list of words or phrases (e.g. names) are to be sorted. A number of different collation sequences have been used by different authors. The ICTA appointed Mr. G. Balachandran to study these sequences, and to recommend a standard for use in Sri Lanka.

The recommendation was to use the following collation order for Tamil: the vowels first, then the Tamil consonants, followed by the Grantha consonants, and then the ஃப (fa) sequence. The symbols and Tamil numerals will come last in the collation order. This recommendation was accepted by the LLWG and the SLSI, and published as the Sri Lanka Standard Tamil Collation Sequence, SLS1326:2008 Part 1 [12].

4. SINHALA INITIATIVES

Although the basic technologies for Sinhala were already available by 2004, we introduce the additional work on Sinhala during the past 5 years.

4.1 Sinhala Collation Sequence

Sinhala collation is based on the order of Indic letters derived from Sanskrit, but has evolved its own conventions over the years. The dictionaries and other reference works which have been published since the 19th Century agree on the basic Sinhala collation sequence, but disagree in details.

The Local Language Working Group studied the issue of Sinhala collation and requested the University of Colombo School of Computing (UCSC) to study the issue of Sinhala collation and recommend a suitable collation algorithm.

The LLWG studied the report submitted by the UCSC and concluded that although classically, the letter ඳ is a conjunct formed by the letters ඬ and ද it is currently used as an independent letter and is never decomposed.

Therefore, the working group recommended two collation sequences for Sinhala.

The *dictionary collation sequence* - which treats ඳ as the conjunct ඬ+ද - should be used in compiling dictionaries and other scholarly works.

The *simple collation sequence* - which treats ඳ as a single letter - should be used in data processing and for lists of names. This collation would not confuse a naive user who is not aware of the subtleties of the language.

The two collations will produce different results only between words with the letters ඟ or ඹ and the letter ඳ in a given position.

This recommendation was accepted by the SLSI in the Sri Lanka Standard Sinhala Collation Sequence, SLS1134:2007 Part 1 [13].

4.2 Fonts

One issue with the use of Unicode Sinhala is the lack of a variety of fonts. Initially, only the Microsoft Iskoola Pota font and the Malithi Web font from Fontmaster (distributed by ICTA) were available.

A Linux font by the Lanka Linux User Group, Sarasavi from UCSC and several proprietary fonts were also developed, but users still complained of a lack of fonts.

ICTA therefore decided to promote the development of Sinhala fonts based on standard Unicode. One problem is that most font designers are not conversant font rules. Therefore, it was decided to develop a set of font rules for Sinhala and make them available for font designers for creating new Sinhala fonts.

To meet the above objectives ICTA has developed an SLS 1134 Level 2 Sinhala font, භාෂිත, for general applications such as documents, books, etc.

An SLS 1134 Level 3-compliant font, which supports touching letters used in Pali and Sanskrit texts and for historical documents, has also been produced. This is the first Sinhala font which can support this style of writing.

As a large body of Sinhala text has been produced using various 7-bit and 8-bit fonts, converters from such fonts to Unicode are required. The UCSC as well as the University of Moratuwa have developed a number of font converters to convert such legacy Sinhala text to Unicode.

4.3 Sinhala and Tamil Computers

Our vision is that a person should be able to walk into the local computer shop and buy a computer which will start up in Sinhala or Tamil.

This has become a reality in the Linux world, with several Sinhala Linux distributions being available.

Microsoft Corp., as part of its Language Interface Pack (LIP) programme, released a Sinhala version of windows Vista as well as MS-Office in July 2009. The development was done by UCSC and Science Land under the guidance of ICTA.

ICTA also collaborated with Intel to promote the installation of the Sinhala and Tamil keyboard and display drivers (developed by Microimage in conjunction with Microsoft) on new PCs.

With these programmes, it is now possible for a non-English-speaking person to effectively use a computer.

4.4 Sinhala Terminology

A recurring question in developing local language applications and content is: "What is the Sinhala word for ...?". A number of glossaries have been compiled by the Ministry of Education, Official Languages Commission, CINTEC, and other bodies over the years. The Dept. of Official Languages, with the assistance of ICTA, has now provided these glossaries on-line [14].

However, in many cases, the words in the glossaries are not suitable for use in computer applications [15].

The University of Moratuwa initiated the creation of a Sinhala terminology for computer applications. Anther list was developed by the Lanka Localization group. The ICTA, in partnership with a several other organisations, compiled a terminology which is available on its website [16].

4.5 Tools and Utilities

A number of other tools and utilities have also been developed. The Language Technology Research Lab (LTRL) of the University of Colombo School of Computing (UCSC) has built a 10 million word corpus of Sinhala, a Sinhala lexicon, Optical Character Recognition (OCR), and Text to Speech software [17].

Methods of transliteration among Sinhala, Tamil and English have been developed by the University of Moratuwa, UCSC and Science Land.

The popular four color logo for the buttons “Get Sinhala” and “Get Tamil” with the four colours depicting the colors of the Sri Lankan flag were designed by Mr. Winnie Hettigoda at the UCSC.

5. CONCLUSION

As a result of over 20 years of work by many people and institutions, we now have extensive use of Sinhala on the web, as well as in documents and applications.

All the basic technologies needed to use not only computers but information systems in Sinhala and Tamil are now in place.

More work needs to be done to add more features, extend local language support to other platforms, and increase awareness and usage.

Some work has already been done on localised applications and content. However, we consider that these areas will require our attention in the coming years.

6. ACKNOWLEDGMENTS

We have reached this stage due to hard and on-going work by many people and institutions. In addition to the ICTA, much of the work has been undertaken by the University of Colombo School of Computing (UCSC) and the University of Moratuwa.

We also acknowledge the exceptional efforts by Mr. Harsha Wijayawardhana and Mr. G. Balachandran, as well as all the members of the LLWG.

REFERENCES

1) 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*, CINTEC, March 1990.

2) Sri Lanka Standards Institute, *Sri Lanka Standard SLS 1134: 1996 – Sinhala Character Code for Information Interchange*, SLSI, 1996.

3) The Unicode Consortium. *The Unicode Standard.* available at: http://www.unicode.org/standard/standard.html

4) The Unicode Consortium. *The Unicode Standard, Version 3.0*. Addison-Wesley, Reading, MA, 2000. ISBN 0-201-61633-5.

5) Gihan Dias and Aruni Goonetilleke, "Development of Standards for Sinhala Computing", in *Proc. 1st Regional Conference on ICT and E-Paradigms*, Colombo, June 2004. Available at: http://www.siyabas.lk/docs/sinhala%20standards.pdf

6) Sri Lanka Standards Institute, *Sri Lanka Standard SLS 1134: 2004 – Sinhala Character Code for Information Interchange (2nd Revision)*, SLSI, 2004.

7) Gihan V. Dias, Representation of Sinhala in Unicode, ICTA, October 2004. Available at: http://www.siyabas.lk/docs/Representation%20of%20Sinhala%20in%20Unicode.pdf

8) Gihan V. Dias and G. Balachandran, "Keyboards for Indic Languages", in *Proc. 12th Internationalisation and Localisation Conference*, Dublin, Ireland, September 2007. Available at: http://www.gnanam.info/tamil/essay/Keyboards_for_Indic_Languages.pdf

9) Microsoft Corp., OpenType specification. Available at: http://www.microsoft.com/typography/otspec/

10) TamilNet99, Keyboard Standards, 1999. Available at: http://www.tamilvu.org/Tamilnet99/keystand.htm

11) Sri Lanka Standards Institute, *Sri Lanka Standard SLS 1326: 2008 – Tamil Character Code for Information Interchange*, SLSI, 2008.

12) Sri Lanka Standards Institute, *Sri Lanka Standard SLS 1326: Part 1: 2008 - Tamil Character Code for Information Interchange Part 1 – Collation Sequence*, SLSI, 2008.

13) Sri Lanka Standards Institute, *Sri Lanka Standard SLS 1134: Part 1: 2007 - Sinhala Character Code for Information Interchange Part 1 – Collation Sequence*, SLSI, 2007.

14) Dept. of Official Languages, Online Glossary System. Available at: http://languagesdept.gov.lk/glossary/

15) Chamara Disanayake and Gihan Dias, "Sinhala Terms for Computer Applications" in *Proc. 24th National Information Technology Conference,* 2005. Available at: http://www.mrt.ac.lk/sinhala/sinhalawordlist/word_lists/CSSLpaper.pdf

16) ICT Agency of Sri Lanka, *Sinhala IT Glossary.* Available at: http://www.siyabas.lk/docs/Glossary-v-1.2.pdf

17) University of Colombo School of Computing, *PAN Localization Project - Phase I.* Available at: http://www.ucsc.cmb.ac.lk/ltrl/?page=panl10n_p1