Survey of Language Computing in Asia
2005

Sarmad Hussain
Nadir Durrani
Sana Gul

Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences

www.nu.edu.pk

www.idrc.ca

Published by

Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences
Lahore, Pakistan

Copyrights © International Development Research Center, Canada

Printed by Walayatsons, Pakistan

ISBN: 969-8961-00-3

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Ottawa, Canada, administered through the Centre for Research in Urdu Language Processing (CRULP), National University of Computer and Emerging Sciences (NUCES), Pakistan.

Sinhala

Sinhala (also called Sinhalese) is an Indo-Aryan language spoken in Sri Lanka by about 13 million people [1]. A considerable number of speakers reside in Singapore, Thailand, Canada and the United Arab Emirates as well [1]. Sinhala is the national language of Sri Lanka. Figure 1 shows the family tree of Sinhala.

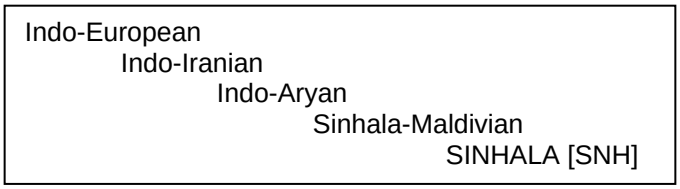

Figure 1: Language Family Tree for Sinhala [1]

Sinhala is written using the Sinhala script. The script is believed to be an off-shoot of Brahmi script [2]. The structure of Sinhala script is similar to that of Sanskrit while it has imported few Tamil characters as well [3].

Character Set and Encoding

SLS1134, developed in 1997 by the Sri Lanka Standards Institute (SLSI), has been the national standard for Sinhala. Unicode code block from 0D80-0DFF is the international standard encoding scheme for Sinhala character set [10], also being adopted at national level but is still not widely used. Prior to the inclusion of Sinhala code page in Unicode in 2003, various character encodings had been developed by local vendors. As reported by University of Colombo School of Computing (UCSC), Sri Lanka, these are still popularly used encoding schemes, and include Thibus, Kaputa, AQNCL, Microimage, Phonetic and SLS1134.

Fonts and Rendering

True Type and Open Type fonts are available for Sinhala. These include free fonts and those developed by various vendors. However, they use different encoding schemes. These include Kaputadotcom, Mi_Dasun Tall (which have Open Type support as well), Dil-Silumina, FM-Basuru, FS-Dilu, Amallee Palin, thibus and Sarasavi. Most of these fonts use phonetic encoding, and Sarasavi Sinhal font uses the national standard SLS1134 encoding. Unicode based fonts are also available [4, 5].

Microsoft Platform

With the inclusion of Sinhala characters in Unicode, Sinhala text can now be rendered on Microsoft platform. Iskola Pota is packaged with the beta version of Sinhala kit for Windows XP which is available for download (through ICT Agency of Sri Lanka) [4, 6]. Outputs of a few fonts for Microsoft platform are given in Figure 2.

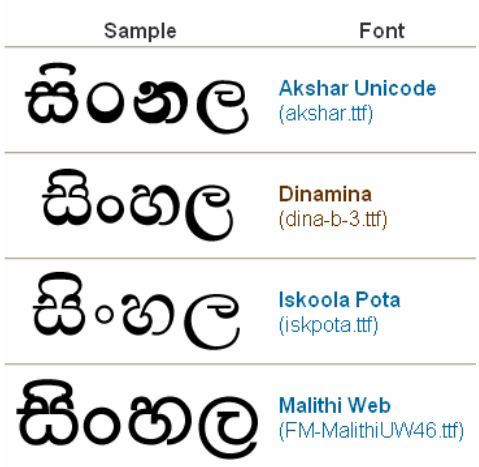

Figure 2: Sinhala Fonts on Microsoft Platform [5]

Linux Platform

Sinhala GNU/Linux 0.1 contains Sinhala language patch for Pango having Sinhala Open Type font support. The figure below shows Sinhala fonts rendering in G-edit and Konquorer.

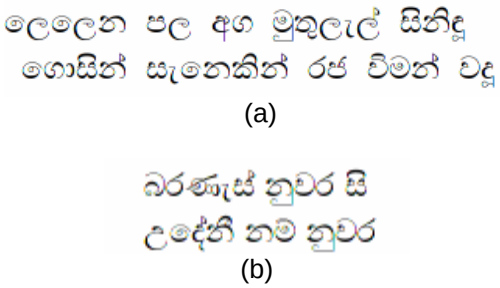

Figure 3: Font Rendering in (a) GNOME and (b) KDE [7]

Keyboard

The standard keyboard layout is based on Wijesekara layout. It was first standardized in 1996 and minor modifications were made in 2001. In 2004 it was further modified and now it is used as the national standard for Sinhala keyboard layout (SLS1134:2004). The following figure shows this keyboard layout (see [9] for a more comprehensive overview of this and other standardization for Sinhala and [11] for the evolution of Sinhala keyboard). Another popular keyboard is based on phonetic mapping of Sinahala characters on QWERTY keyboard of English.

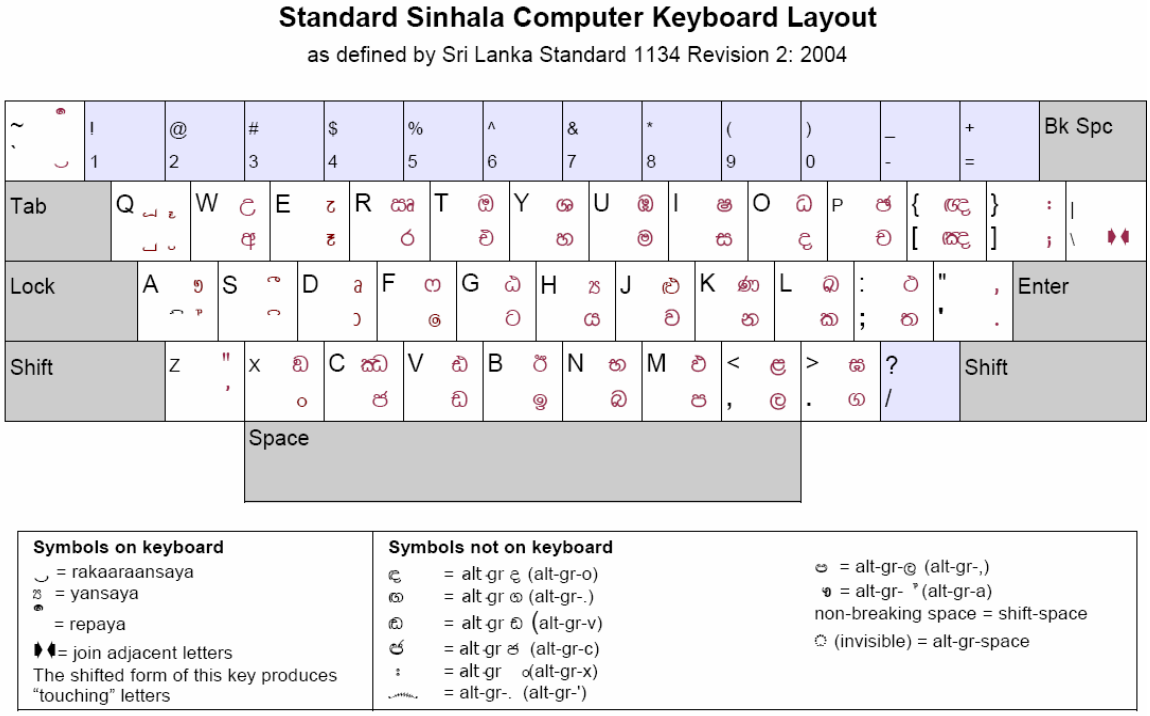

Figure 4: Standard Keyboard Layout [8]

Microsoft Platform

Microsoft does not provide support for Sinhala text input. However, a built-in Sinhala keyboard layout is installed with the beta version of the Sinhala Language Kit [4]. Once the kit is installed, the Sinhala keyboard is available.

Linux Platform

A phonetic-based keyboard layout is shipped with Sinhala Linux distribution.

Collation

To date, no standard collation order exists for Sinhala character set. However, currently a standard is being developed at University of Colombo School of Computing (UCSC) for Information Communication Technologies Agency (ICTA) which will be standardized through SLSI.

As no standard collation exists, there is no collation implemented on Microsoft or Linux platforms.

Locale

Sinhala locale has neither been defined in the IBM ICU nor in the CLDR 1.3 release. Currently Sinhala locale definition for Microsoft is being finalized by ICTA [6]. Locale definition for Sinhala and Sinhala interface terminology translation is also being done in parallel by a Lanka Linux User Group (LkLUG) [7].

Microsoft does not provide a locale for Sinhala. On the Linux platform the Lanka Linux User Group (LkLUG) has included limited support for a Sinhala locale in Glibc in the Sinhala Linux distribution GNU/Linux 0.1 [7].

Interface Terminology Translation

Microsoft Platform

Microsoft developed a Sinhala version [12] at an accelerated pace to help support Tsunami relief in 2005 in collaboration with ICTA, Sri Lanka. A Sinhala LIP for Windows XP and a Sinhala kit for Microsoft Office 2003 is available for download [4]. However, more mature version of interface translation is still under development. ICTA has completed initial phase in translating interface terminology. The LkLUG has done more work on this than is available on Microsoft platform at present. They had a small project to translate the 200 most common UI terms. There is some on-going work on terminology translation (of some 2500 words) at UCSC which will be incorporated in future. This work will also be adopted by ICTA, as reported by University of Colombo School of Computing.

Linux Platform

A complete Sinhala Linux distribution GNU/Linux 0.1 has been developed and released by LkLUG. This distribution includes Sinhala fonts, phonetic keyboard, locale and a language patch for Pango. It is partially localized, as shown in Figure 5.

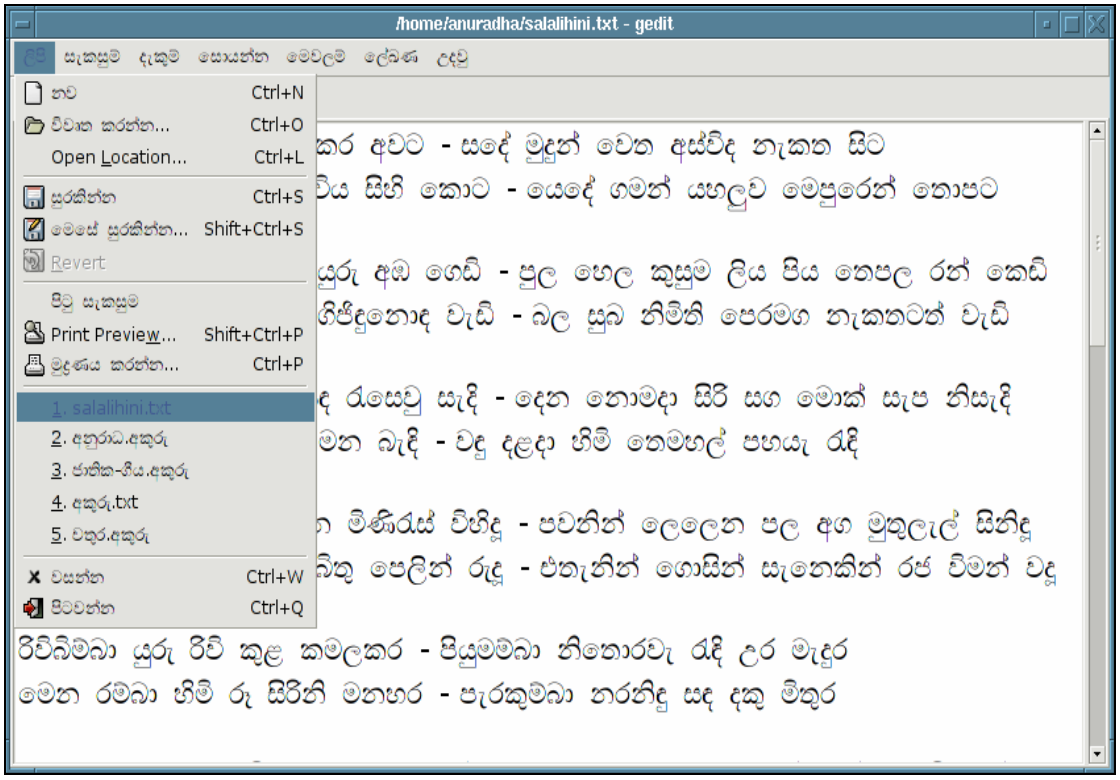

Figure 5: GUI for G-Edit [7]

Translation teams for both KDE and GNOME have been registered on their official websites. Some work on the glossary translation of open source applications is currently underway. Language teams for Sinhala have registered on the official websites of KDE and Mozilla. Glossary translation of GNOME and Open Office has not been initiated yet.

Status of Advanced Applications

As reported by the University of Colombo School Of Computing, work is in progress to develop local language applications in Sinhala. The language processing teams have developed experimental versions of Sinhala Lexicon, corpus and spell checker. Spell checker has also been developed by LkLUG, incorporated in Linux distribution, as shown in Figure 6.

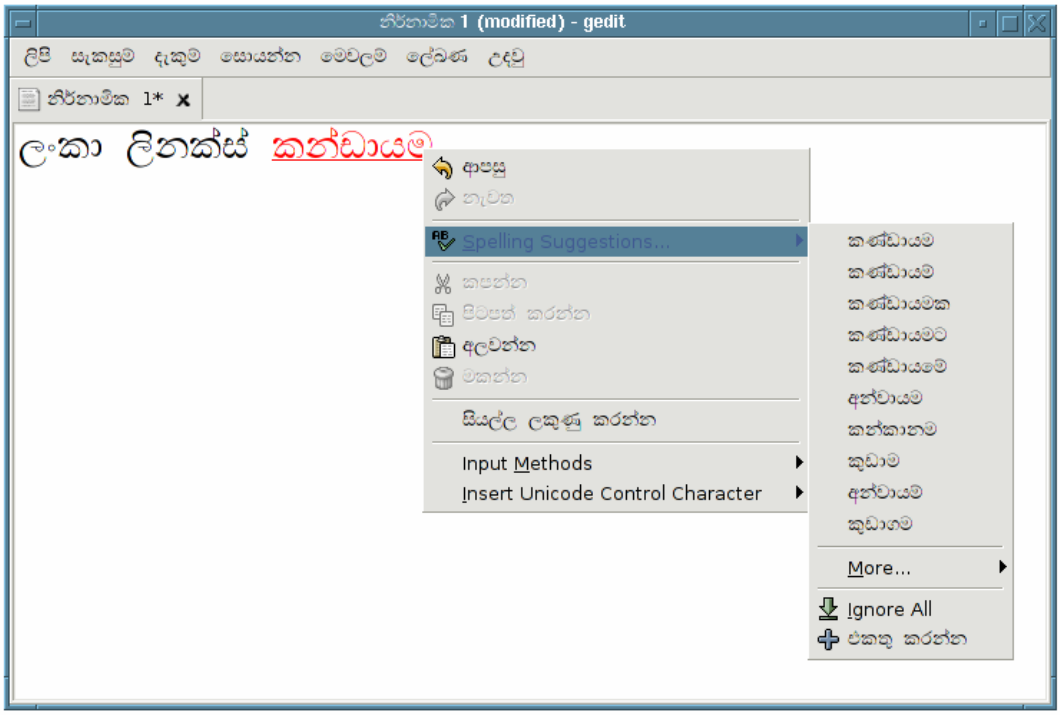

Figure 6: GUI for G-Edit [13]

Through PAN Localization project, UCSC has developed and already released alpha versions of Sinhala OCR and text-to-speech system [14]. The work is underway for maturing these systems.

References

[1] http://www.ethnologue.com/show_language.asp?code=sin
[2] http://www.omniglot.com/writing/sinhala.htm
[3] http://acharya.iitm.ac.in/cgi-bin/script_disp.pl?sinhala
[4] http://www.fonts.lk/index.html
[5] http://salrc.uchicago.edu/resources/fonts/sinhalafonts.html
[6] http://www.icta.lk/
[7] http://sinhala.linux.lk/
[8] http://www.fonts.lk/doc/sin-kbd-layout5.pdf
[9] Dias, G. and Goonitilleke, A., "Development of Standards for Sinhala Computing." 1st Regional Conference on ICT and E-Paradigms, Srilanka. http://www.fonts.lk/doc/sinhala%20standards.pdf. 2004.
[10] Dias, G. "Representation of Sinhala in Unicode." http://www.fonts.lk/doc/Representation%20of%20Sinhala%20in%20Unicode.pdf
[11] http://www.fonts.lk/doc/workshop-sinhalakeyboards-anura.ppt
[12] http://www.dailynews.lk/2005/01/29/bus02.html
[13] http://sinhala.linux.lk/pub/0.2/screenshots/spell.png
[14] http://www.PANL10n.net