

Sinhala, Language Engineering and an Attempt to take Internet to the Masses.

Halahakonege Don Joseph Vincent

Introduction

The oldest written datable Sinhala manuscripts go back to the 3rd century BC written using Brahmi. Since then, a wealth of literature has been documented on various subjects ranging from Ayurveda , indigenous medicine, more than 10,000 Buddhist scripts and a large number of literary works. For all this to be shared and used universally Sinhala needs to be electronically recognized in the sphere of language engineering. On the other hand there is an equally important need for the users of this language to have access to the wealth of knowledge that is universally available in the World Wide Web. This paper discusses two parallel attempts to accomplish this task of crossing the divide.

Brief Note on Sinhala Language

Sinhala language belongs to the Indo-Aryan group of languages just like Bengali, Gujarati, Marathi, Hindi, and Punjabi. As modern European languages trace their ancestry to Greek and Latin, Sinhala and other Indo-Aryan languages trace their origins to Sanskrit. Sri Lanka, some experts claim as having one of the world's oldest continuous written records of history called *Mahavansa* (more than 2500 years).

Sinhala, which developed as an island language, has some unique features, which are not known in any other Indo-Aryan languages. This uniqueness of Sinhala is due to its exposure by way being geographically together to other language families of the region such as Dravidian and Malayo-Polynesian. The Sinhala language came to Sri Lanka with the original migrants from North India who are traditionally considered to be the founders of the Sinhala nation. They spoke Indo-Aryan languages depending on the areas from which they migrated. The early migrants came from Bengal, *Magadha* and *Kalinga*. The languages in all these areas were variants of Indo-Aryan, not too dissimilar to each other, and it is speculated that Sinhala is an amalgam of these languages. The Tamil language, which belongs to the Dravidian group, has influenced the structure and vocabulary of Sinhala to such an extent that some scholars were erroneously led to believe that Sinhala belonged to the Dravidian group of languages.

Sinhala and Corpus building

With the introduction to the Project Enabling Minority Language Engineering (EMILLE) as one of its contributing languages, Sinhala began to become a focus of increasing attention within the circles of computer science experts and linguists of Sri Lanka.

The EMILLE was established to construct a 67 million-word corpus of South Asian languages with the goal of developing written language corpora of at least 9,000,000 words for Bengali, Gujarati, Hindi, Punjabi, Sinhala, Tamil and Urdu. In addition, for those languages with a UK community large enough to sustain spoken corpus collection (Bengali, Gujarati, Hindi, Punjabi and Urdu), the project aimed to produce

spoken corpora of at least 500,000 words per language and 200,000 words of parallel corpus data for each language based on translations from English. (Paul baker et al 2003)

<i>Language</i>	<i>Target word count (millions)</i>	<i>Current word count (millions)</i>
Assamese	2.6	2.6
Bengali	9.0	5.5
Gujarati	10.6	10.6
Hindi	12.0	11.2
Kannada	2.2	2.2
Kashmiri	2.3	2.3
Malayalam	2.3	2.3
Marathi	2.2	2.2
Oriya	2.7	2.7
Punjabi	9.0	4.5
Sinhala	9.0	6.0
Tamil	15.0	13.9
Telegu	4.0	4.0
Urdu	3.0	1.6
Total	85.9	72.1

(From Paul Baker et al 2003)

Sinhala and Coding Attempts

When microcomputers were introduced in Sri Lanka in early 80s there was no proper coding system for Sinhala available but the work on input and output and local languages continued for various purposes such as general election results display on television. (*Samaranayaka V.K, et al, 2003*). The need to establish a standard code for Sinhala was identified and the Computer and Information Technology Council of Sri Lanka (CINTEC) appointed a committee in 1985 for the use of Sinhala and Tamil in communication technology.

Early steps were taken to agree on an acceptable Sinhala alphabet and a sorting order for IT use. This committee worked with close liaison with the Committee on Adaptation of National Languages in IT (CANLIT), which was appointed by the Natural Resources Energy and Science of Sri Lanka. (NARESA).

The next important step was the proposal of a draft code for Sinhala to the ISO/UNICODE 10646 work group by researchers based in Europe in late eighties. Although this proposal contained glaring errors especially in the location of characters apparently caused by the misunderstanding that Sinhala was a subset or equivalent set of other Indic Language, it was a very important step towards Sinhala being accepted in the Unicode and the for opportunities it paved for experts to work on an improved version of the coding system later.

The errors mentioned were immediately put right with the intervention by representations made by the CINTEC and the Sri Lankan Standard Institute (SLSI). During this period there was no Sri Lankan representation at the ISO/UNICODE working group. In 1997 the working group met in Greece, which was duly represented, by Sri Lankan experts and a final agreement was reached with slight modifications to earlier the proposal. Following the work group meeting in Seattle in 1988 the Sinhala code chart was included in the UNICODE version 3.0.

The differences between Sinhala and the other languages in the region is that it has a set of pre nasalized stops and the presence of distinct signs short and long front vowel sounds similar to initial vowel of the English word ‘apple’ as represented in IPA as æ ‘ash’.

As a result, Sinhala does not precisely follow the format assigned to the other Indic scripts such as Devanagari but it does use the general structure. (UNICODE Version 3.0)

This leads us to the technical issues outside the UNICODE standards namely input and rendering. Sinhala is encoded by no less three competing and mutually unintelligible fonts (see McEnery & Gaizauskas in Jayalal 1997, Mohanraj 1998 & Pratheepan 1998). As mentioned by McEnery et al ‘There is a pressing need to develop software to map these many font based representations of the writing systems used by Indic languages to a common standards.’ Mapping these different fonts to UNICODE is being done at Universities at research level.

Meanwhile there is good news that Microsoft is preparing itself to accept Sinhala as one of its languages. According to Microsoft this program called the ‘ Iskoola Pota’ is to be shipped on a future version of Windows or with a service pack that supports Sinhala. It will not be available for older versions of Windows because it relies on a version of *Uniscribe* that has the Sinhala shaping engine. Work is also underway on a commonly accepted keyboard and a set of Unicode compliant set of fonts at the University of Colombo.

Once this work is over it is necessary that we create web pages of local languages in Unicode compliant fonts for universal use. Parallel work is then to be undertaken (some work has already done at research level) on MLT, speech recognition and character recognition relating to Sinhala.

This is the story of how Sinhala is trying to emerge in the NLP panorama and some of the related issues. For every person in Sri Lanka to have the benefit of a computer in trade and in information technology communication it is necessary to have uniformity in the way Sinhala and Tamil characters represented in computers.

Currently most computer operation systems, databases, and applications in Sri Lanka run only in English language, which only 1 million people out of the total population of 19 million people can use. The majority, the 18 million has a need to access this sea of information in their own languages, namely Sinhala and Tamil.

Sri Lanka Kothmale’s Community Radio Internet Project

The Kothmale FM Community Radio Internet Project is an attempt to bridge this gap of extending ‘...the benefits of information and communication technology (ICT) to

some of the remote areas of Sri Lanka through the innovative convergence of two media – the radio and the Internet.’ This radio station was launched in February 1989 by Sri Lanka Broadcasting Corporation. Its objective was to offer the benefits of the Information Communication Technology (ICT) to a population of nearly 60,000 valley dwellers who had to be resettled due to a huge hydrological developmental site - the Mahaweli Irrigation Project.

The Radio Station a 45-minute travel by bus from the nearest town is located at the isolated hilltop in an accessible central region of the island. The place is mountainous and renowned for its beauty, and salubrious air.

The radio station is equipped with high tech gadgetry, a microwave connection, a telephone line, a server, three PCs and a 64 KB dedicated line with Internet access all made possible by a UNESCO grant. Initially transmission there was limited to three hours a day. Expenses of part of this broadcast had to be covered from commercials. In 1996, discussions were held with UNESCO about this novel idea of combining the media facilities of the radio and the Internet.

The Internet project represented firstly, ‘a move to bring computing and Internet to a poor rural area, with the additional aim of establishing outlying Internet access in neighbouring towns.’ It however has another more innovative objective of ‘integrating Internet and radio within community communications.’ (*An Ethnography of Kothmale Community Radio Internet Project* (KCRIP)). The advantage of this facility was the gates it opened to the ‘Information Super Highway’ at a very low connectivity cost. The project grant was an aid by UNESCO of US \$ 30,000 and implemented jointly by Sri Lanka Broadcasting Corporation, the Ministry of Post, Telecommunication and Media, the Sri Lanka Telecom Regulatory Commission and the University of Colombo.

One of the proponents of this idea, MJR David, pointed out its aims as:

- to empower marginalized rural communities to take advantage of new communication technologiesto reduce the rural/urban communication gap.
- to create awareness of new technologies among rural communities.
- to develop a rural community database by interfacing community radio with the internet.
- to provide a test-bed to develop an adaptable model for provision of access to new communication technologies by alternative means. (David 2001)

Programs broadcast are on health, human rights, international news and women and children’s programs. One of the advantages is that the people in the community have access to Internet information in their local language either as answers to the queries on telephone or as printouts of the public domain information from the WEB. Certain government documents such as passport applications are a facility available.

There have been moments of frustration in the growth process emanating from issues institutional, technical and local politics as usual. This situation reached a crisis when the entire Internet connection was disconnected for a year but it is now back on track.

Impact

The project has surely spread the ICT benefits to the remotest areas in the island. It has also paved way for further development of such projects in other parts of the island and the involvement of more and advanced technical and human resources. As reported by a local survey, 90% of its respondents listen to the Internet radio for information. (Notley 2000).

Information gathered at this station is reported to be used for various purposes such as education, business, livelihood and recreation. An average of 20 to 25 people surf the six PCs at the station daily. In addition there are other access points in the nearest towns where daily surfers are calculated be around 10 to 15 a day. The majority – 90% - of the surfers belong to the age groups of 10 to 25 years and 60% of this are between 15 and 20 (Notely 2002).

To end it is best to quote from the ‘Ethnographic Monitoring And Evaluation Of Community Multimedia Centres’ - A Study of Kothmale Community Radio Internet Project, Sri Lanka “ Kotmale project probably with help and support from other organizations and donors – could play a central role in addressing one of the key ICT needs in the area: educating the educators. Training school teachers (from both state and private schools) so that they could make fuller, more confident and more creative use of the ICT facilities that they have, and with more emphasis on developing practical competences and project-oriented approaches, would make a huge difference to ICT use, job and career opportunities and the general level of education of all their pupils.

Extremely useful and effective partnerships could be developed with IT educational establishments at all levels and through such partnerships and the number and quality of locally skilled people this could feed back into the locality through the provision of a range of computer training to suit all levels of need in the community.

References

A Study of Kothmale Community Radio Internet Project, Sri Lanka, Don Slater, LSE Jo Tacchi, QUT Peter Lewis, LSE (Report of a research project funded by Department for International Development, UK (DfID), in collaboration with UNESCO.)

An Introduction to UNICODE for Sinhala characters, (Samaranayaka. V.K, Nandasara, S.T, University of Colombo School of Computing 2003.

Corpus Data for South Asian Language Processing, Paul Baker, Andrew Hardie, Tony McEnery, (Department of Linguistics, Lancaster University, and Sri B.D. Jayaram Central Institute of Indian Languages, Mysore)

Empowerment Case Studies: Sri Lanka’s Kothmale Community Radio Project Internet Project prepared by Prof. Subhash Bhatnagar and Ankita Dewan at the Indian institute of Management (Ahmedabad) and Magüi Moreno Torres and Parameeta Karungp at the Word bank (Washington DC)

Results from a Survey Conducted by Tanya Notely, which are included in her unpublished report on Kothmale (2000). The survey sampled 93 users of the Kothmale Internet facilities over a two-week period.

The Unicode Consortium (2000) The Unicode Standard 3.0, Harlow: Addison Wesley.

Appendix 1

Further Information: References and World Wide Web Resources

Web References

Kothmale Community Radio Website. URL: <www.kothmale.net>.

UNESCO: Communication, Information, Informatics. "Internet Radio in Sri Lanka." URL: <http://www.unesco.org/webworld/highlights/internet_radio_130599.html>.

Jayaweera, Wijayananda. 2001. "Kothmale Community Radio/Internet Project: Expanding the Knowledge Base." The World Bank Group E*Government Website. URL: <http://www1.worldbank.org/publicsector/egov/kothmale_cs.htm>.

The Communication Initiative. 1989. "Kothmale Community Radio, Sri Lanka." URL: <http://www.comminit.com/11-342-case_studies/sld-617.html>.

Hermida, Alfred. "Listening to the Web on the Radio." BBC News. URL: <<http://news.bbc.co.uk/1/hi/sci/tech/1796236.stm>>.

UNESCO. 2001. "Including the Excluded." January. URL: <<http://216.239.33.100/search?q=cache:veubvKLSgHIC:www.kothmale.net/kcrwebsite/english/seminar/kothmale%2520Brochure.pdf+Kothmale+Community+Radio+Challenges&hl=en&ie=UTF-8>>. See also: <<http://www.kothmale.net/kcrwebsite/english/seminar/kothmale%20Brochure.pdf>>.

Stockholm Challenge. "Project Details, Kothmale Internet Community Radio, Sri Lanka." URL: <<http://www.challenge.stockholm.se/projects.asp?ProjectId=2602>>.

Jayaweera, Wijayananda. "Access in Rural Areas, Some Lessons and Issues." URL: <<http://216.239.33.100/search?q=cache:bqkZwMhS4V8C:www.aptsec.org/seminar/meeting-2001/ecs2001/EC-09-UNESCO.doc+Kothmale+Community+Radio+Lessons&hl=en&ie=UTF-8>>. See also: <<http://www.aptsec.org/seminar/meeting-2001/ecs2001/EC-09-UNESCO.doc>>.

Pringle, Ian, and M.J.R. David. 2002. "Rural Community ICT Applications: The Kothmale Model." *The Electronic Journal on Information Systems in Developing Countries* (EJISDC) 8(4): 1–14. URL: <<http://216.239.33.100/search?q=cache:ivL8XjPrDo0C:www.is.cityu.edu.hk/research/ejisd/vol8/v8r4.pdf+Kothmale+Community+Radio+Lessons&hl=en&ie=UTF-8>>. See also: <<http://www.is.cityu.edu.hk/research/ejisd/vol8/v8r4.pdf>>.