Abstracts of the Symposium on Localised Systems and Applications – 2009

September 2, 2009

Centre of Excellence on Localised Applications

University of Moratuwa

Sri Lanka

Abstracts of the Symposium on
Localised Systems and Applications – 2009

Organised by :

Centre of Excellence on Localised Applications,
University of Moratuwa
Sri Lanka
www.lakapps.lk

Strategic partner :

Information and Communication Technology Agency of Sri Lanka,
160/24, Kirimandala Mawatha,
Colombo 5,
Sri Lanka
www.icta.lk

The Inaugural
Symposium on Localised Systems and Applications

Sri Lanka has made impressive progress in developing localised software, systems and applications. Today, our citizens can use computer operating systems, application software, keyboards and mobile phones in their mother tongues of Sinhala and Tamil.

We can be proud that this has been achieved in Sri Lanka, by practitioners in the emerging field of local language computing. However, a forum where they can meet and share their experiences has been a long-felt need.

A discussion on the local-language applications was organised by the ICTA on 14th November 2008 at the BMICH. It provided a valuable opportunity for the people in this area to meet each other and find out what others were doing, as well as publicise their work.

Following up on this meeting, the Centre of Excellence on Localised Applications at the University of Moratuwa decided to organise a regular event to provide a forum for those working in localisation and related fields. The inaugural Symposium on Localised Systems and Applications is being held at the University of Moratuwa on 2nd September 2009.

We are very impressed with the response to our call for presentations. We have a full programme, with presentations by government, university, commercial and non-profit organisations as well as individuals.

It shows that many people in this country are developing Sinhala and Tamil technologies and software, and that our people now have an unprecedented opportunity to use IT in their own language.

Some of the areas being presented at this symposium are:

- Standards
- Linguistic Resources
- Sinhala and Tamil Terminology
- Localised Operating Systems
- Localised Applications
- Local Content
- Documentation and Training
- e-Learning in local languages

We are extremely grateful to all the staff at the University of Moratuwa and LK Domain Registry who worked to make this symposium a success.

Organising Committee

Prof. Gihan Dias
Prof. Asoka Karunananda
Dr. Sanath Jayasena
Dr. Chulantha Kulasekera
Mr. Chamara Disanayake
Mr. K Sarveswaran
Mr. Rohana Dassanayake
Mr. Nilhan Niles
Mr. Dihan Morawaka
Dr. Ruwan Gamage

Centre of Excellence on Localised Applications

(LAKapps-Centre)

University of Moratuwa

The University of Moratuwa has been a pioneer in the field of local-language technologies, localisation, and localised systems and applications development.

In recent years the university has successfully undertaken a number of projects related to local-language technology, training, dissemination, and applications development. Several undergraduate and postgraduate projects have been carried out in these areas, which have been published both locally and internationally. UoM staff members have also been advising the government in these areas.

Accordingly the University of Moratuwa decided set up a centre of excellence to utilize the staff of the university to work as a team to provide expert knowledge and to carry out research work related to localised applications, thereby contributing to the national efforts in this respect.

The objectives of this Centre are as follows:

- to develop, retain and disseminate knowledge and expertise in the field of Localised Applications and its related subject areas including but not limited to; localisation policy and standards development, localisation technology, systems and software globalisation, Internationalisation and localisation, local content development, localisation management and training.
- to formulate and submit proposals to solicit funding for research and study in any one or more of the above areas and to utilise such funding.
- to organize, coordinate and conduct research, training and consulting activities that could include staff across different departments of study.
- to liaise with other relevant institutions, both locally and internationally, in pursuit of common objectives.
- to acquire, retain and make available the knowledge on localised applications and related subjects through studies, publications, seminars and other such means.
- to advocate and make inputs into localisation standards, policies and regulations as may be required by the relevant authorities from time to time.
- to procure goods and services necessary for the pursuit of the above objectives.
- to conduct research at undergraduate, postgraduate and post-doctoral level in relevant areas in collaboration with the relevant departments of the University.

Table of abstracts

Meaning: A Linguistic Analysis

J.B.Disanayaka

jbdisanayaka@yahoo.com

Translation, machine or otherwise, involves the transfer of meaning from one language (source language) to another (target language).

Meaning of a linguistic expression exists on a number of levels, which include the following:

1. **Grammatical Meaning**

   Analysis in terms of grammatical categories such as:

   a. word class (parts of speech, noun, verb, adjective etc)

   b. animate vs. inanimate; masculine vs. feminine; singular vs. plural; definite vs. indefinite; past vs.present etc.

2. **Phrasal Meaning**

   Analysis in terms of the grammatical function in phrases: As head and modifier etc.

3. **Contextual Meaning**

   Analysis in terms of verbal contexts: what occurs before, after or both. Formation of compounds etc.

4. **Idiomatic Meaning**

   Meaning beyond individual words

5. **Restricted Meaning**

   Analysis of meaning in terms of 'semantic fields' or 'domains' such as scientific writing, folklore, literature etc

6. **Proverbial Meaning**

   Meaning of proverbs

The LAKapps Project: Localisation, User Guides and Training

Sanath Jayasena
University of Moratuwa
sanath@uom.lk

Dihan Morawaka
University of Sri J' pura and LKiNG
dihan@sjp.ac.lk

The LAKapps (Local Language Internet) Project was a pilot project to demonstrate the use of Sinhala and Tamil Internet Applications. It was carried out July 2007 – December 2008 in seven locations, across Sri Lanka, namely:

- Nenasala at Sevanagala (Moneragala District)
- Embilipitiya Central College (Ratnapura District)
- Nenasala at Kothmale Community Radio (Nuwaraeliya District)
- Computer Resource Centre (CRC) at Bandaragama Maha Vidyalaya (Kalutara District)
- Kottawa Dharmapala Maha Vidyalaya (Colombo District)
- St. Sylvester's College, Kandy (Kandy District)
- Ramanathan Hindu Ladies College, Colombo (Colombo District)

In addition to the installation of the Sinhala and/or Tamil language pack and keyboard, the following localised software were installed at each of the above sites.

- Sinhala Firefox web browser
- Tamil Firefox web browser
- Sinhala Thunderbird e-mail client
- Tamil Thunderbird e-mail client
- Sinhala Joomla content manager (not localised by us)
- Tamil Joomla content manager
- Sinhala Moodle Learning Management System (not localised by us)
- Tamil Moodle Learning Management System

Before the installation, localized software were improved from the status they were in. In addition, a number of other software such as a webmail client and a Sinhala/Tamil typing tutor were also produced. A noteworthy component of the project was the preparation of user guides and training material in both Sinhala and Tamil for all the software above. Further, we also created multimedia learning material for OpenOffice Writer (word processor) and Calc (spreadsheet).

Two rounds of training - on local language computing and on localised applications - were conducted at each of the sites. We were able to train nearly 150 adults and 300 students in the effective usage of IT applications in their own language. The co-coordinators at each site were given special training to support the end-users of the localized applications and to give us feedback to improve software continually. We achieved successful completion of the objectives ahead of the deadline.

This project was lead by the University of Moratuwa in partnership with the Arthur C. Clarke Institute, LK Domain Registry and the Lanka Internet Networking Group, and administered by Uni-Consultancy Services. It was partly funded by the ICTA under the e-Society Development Initiative, Partnership Assistance Programme (PAP).

Tamil Encoding, Keyboard Layout and Collation Standard for ICT Sri Lanka

G Balachandran

ICT Tamil Consultant – ICTA, Sri Lanka

balag.lk@gmail.com

One of the need of localisation is developing standards for operating environments. When the standardisation process is through, it will be easy for developers and vendors to build applications and products for the target environment.

The SLS 1326 : 2008 standard for Tamil ICT, was approved by the Sectoral Committee on Information Technology and was authorised for adoption and publication as a Sri Lanka Standard by the Council of the Sri Lanka Standards Institution. This standard defines mainly three(3) standards for Tamil ICT; viz character encoding, keyboard layout and collation sequence.

Character encoding defines codes for the vowels, consonants, āytam, vowel modifiers, numerals, and symbols in the Tamil language. Some formations of the language are not represented by individual codes, but are generally constructed as a sequence of one or more consonants followed by a vowel modifier which forms a syllable.

Standard also provides a **keyboard layout**, which in turn is based on the "Rengenathan" typewriter keyboard. Key sequences are defined on the principle "type as you write". Each symbol is typed in the order it is written in, which may be different from the encoding sequence or the display order.

In **collation sequence** standard, an effort has been made to preserve the alphabetical order of the Tamil language to a great extent. Tamil collation has similarities with other Indic languages which follow the Sanskrit collation order. From the 16th century to the early 20th century most Tamil dictionaries followed the Sanskrit collation sequence which includes the Grantha letters in between the Tamil letters. However most dictionaries and scholars follow the unique collation sequence which collates the Grantha letters after all Tamil letters and which has been mostly accepted as the de facto standard among Tamil communities. Moreover this standard provides the place for 'fa' letter in Tami collation sequence first time.

Tamil Localisation Process

K.Sarveswaran

Department of CSE, University of Moratuwa

iamsarves@gmail.com

Localisation has become an active area in the field of computing and many organisations and individuals are localising software into their preferred languages. Different people use different localisation processes to do localisation. We cannot see much difference in the existing localisation processes. There can exist more than one localisations for a language. This difference may occur if localisers follow different glossaries and style guides. Many software are localised to Tamil language. The aim of this abstract is to describe a Tamil localisation process that is followed in Sri Lanka.

The Tamil localisation works are being done in India and Sri Lanka mainly. However because the glossary and style guide that are being used in two of these countries are different, normally Sri Lankan Tamil localisation efforts are carried out under the ta-LK locale. Some organisations follow ISO 639-1and ISO 3166 standard to name the locale. This also leads us to use ta-LK. Furthermore when we say locale that also includes some local properties, for example Mozilla Firefox provides the site searches based on a country's demand. These are the added reason that we had to go with ta-LK locale.

Extracting the source files, translation, testing and packaging are the primary steps that are there in a typical software localisaton process. In each of these steps various tools and technologies are been used to ease the task. The studies have shown that human translation takes lot of time and may lead for lot of mistakes. Also the Machine translation is not always correct. Therefore the intermediate technology called Translation Memory(TM) is used in this translation phase. According to this TM, first a partial translation will be done automatically. Then a translator will do the rest. Using TM the existing translation knowledge can be reused. There are different tools used for TM.

In Sri Lanka also, a localisation process is being followed in Tamil localisation. Part of this process is done through a software called Pootle. The detailed steps are :

- Identify the Software that needs to be localised and do feasibility study. Contact the respective organisations and inform them if it is planned to work on that localisaton
- Get the language source files and identify the appropriate tools to do the translation
- Assign tasks to translators. Initially they use TM to translate the language strings. Then they will attend to the un-translated words. POEdit is used as TM to carry out the initial translation.
- Review the translated strings, mainly for spelling mistakes and policy mismatch.
- Package the translated files, compile the localised version and get the localised version of the software reviewed by people who are really going to use that software
- Submit the localised version to the respective organisation and make the localised version available to the public
- Prepare the required supporting materials like user manual or short guides
- Maintain the language packs and update them with the new versions and new feedbacks
- Spread local applications to the end users

Compared to other localisation processes, this process has certain unique features. According to this process the software are localised and taken to the targeted users. Using this localisation process Mozilla Firefox, Mozilla Thunderbird, Moodle, Joomla!, SquirrelMail, Horde, GeoGebra are being locaised successfully.

Demonstration of Tamil GNU/Linux Live-CD and Dialog about Tamil on GNU/Linux

M.Mauran

Noolaham Foundation

mmauran@gmail.com

Tamil on GNU/Linux and its brief history

As I referred the sources and records I found on Net, Tamil GNU/Linux localization efforts started in 1999. Since its included in KDE 2.0 (released in 2000), Tamil is the first among Indian languages to have GNU/Linux Interface. This made GNU/Linux become first in including Tamil interface among the OS ecosystem. TSCII was adopted as primary encoding system because Unicode was not fully standardized and many controversies and discussions about Unicode standards went on those days. UTF-8 support with backward compatibility became primary Tamil encoding system of GNU/Linux later. Availability of source code and localization packages helped Tamil community to contribute and produce Tamil support for GNU/Linux time to time. Mandrake Linux 9.0 is the very first commercially supported GNU/Linux Distro which included Tamil interface and support by default. Mandrake Linux 9.0 was released on 2002. Today almost all famous GNU/Linux dstros are having Tamil packages and support.

Adoption of Tamil and GNU/Linux in Sri Lanka

Adoption of GNU/Linux in Srilanka is limited among few Computer enthusiastics and Computer technicians. This explains how Tamil GNU/Linux adoption in Sri Lanka is. Commonly, products of foreign commercial private companies are being taught in school curriculum and private computer education institutions. This makes properietory software products become familiar and famous among average computer users in SriLanka. It is an irony that, the commercial products are being taught by government schools for children when we have great FOSS alternatives to teach and use.

Needs of GNU/Linux-based i18n and l10n efforts

The need to put our community labour regards FOSS, especially GNU/Linux is significant. This is very much fit to localization efforts & Language-based software development efforts as well. Using our limited human & capital resource in a useful way and investing them in long lasting sustainable processes is, only possible in this way. This will make sure our efforts could be used by people as their Own. Also this will give the chance for the community itself to benefit from the community's volunteer efforts forever.

Live-CD to demonstrate Tamil on GNU/Linux

Since Lacking of awareness and introduction of GNU/Linux among our society, We need an Out-Of-The-Box, Ready-To-Use product to introduce Tamil GNU/Linux.

A LiveCD with almost all Tamil facilities installed could fulfill this need. Now users can see, feel and use Tamil on GNU/Linux by just boot the CD in their computer.

Experiences and notes on Tamil GNU/Linux creation

- Tamil GNU/Linux Distro is based on Ubuntu GNU/linux. This choice made for the sake of wide range of support base, availability of binary packages etc.
- Major challenge I faced is managing the CD image size. I gave top preference to off-line softwares to use in day to day PC tasks and gave up the space allocation for online service clients such as VOIP Clients, RSS aggregators etc which can be installed via internet with just two or three clicks.
- Non-properietory fonts, All available Tamil Input methods, Localization packages, Unicode renderers, Modifications and settings for Tamil use are included in this LiveCD.
- Tamil is the default interface language for this Distro. This decision compromises the friendlinesses (for the first timers) and pluralism for the sake of it's primary intention which is demonstrating Tamil on GNU/Linux and saving CD space.
- A Collection of bookmarks of useful Tamil resources on Internet is included in it's default Web Browser.

Creating games with a local flavour: From mobile to console

W.G.T.Fernando

GTS, Sri Lanka

gihan.fernando@gtslk.com

Video games started way back in 1971 when Atari released 'Computer Space'. Pong and other games soon followed. Initial games were mostly concentrated for coin operated machines before computers were affordable. Now you get games in almost all interactive electronic devices. You get games preinstalled on your pc and also your mobile. A number of consoles (Xbox360, PlayStation) also exist which have dedicated processing resources to render high end graphics of current games.

So where does Sri Lanka stand in the gaming industry? More importantly, do we see much games with a local flavour? Which platform is best to start out in the gaming industry?

Sri Lanka has a population of 20 million, and of that, close to 9 million have mobile phones. With prices coming down each day, most are able to afford phone models with extensive features such as camera and Java enabled capabilities. Hence the reason to create a game on a mobile so that many people can easily get access to.

With the launch of the first mobile game on the auspicious day of 8th August 2008 (8-8-8), a gaming interest spurred in the country. Never before has anyone played a Sinhala game on a mobile until that fortunate day.

The game was such a hit that award winning actor/director Jackson Anthony himself requested a game for his film 'Aba'. Then the 2nd version of Colombo Ride followed with a 2 player game.

What can we learn from these games and where can we go? Sri Lanka has a great history and one which the gaming world could easily adopt into games. It is our responsibility to use new applications such as games so that our history is not forgotten. Xbox360, PlayStation, Wii are gaming consoles that are widely used in the world. Microsoft and Nintendo have made it possible for even

the smallest company to create games for their consoles. As a result, I am happy to say that in the very near future, you will hear of games based on Saradiel and King Dutugemunu coming to the consoles and PC platform.

A lot of request has been made by the game fans asking how we came up with the mobile game and whether they too can create a similar local game. On the 19th September 2009, any Sri Lankan would easily be able to create their own mobile game as we will be launching 'Colomba Wate: The book' which will explain stage by stage on how we created the game. The complete source of the game will also be released with the book. It's a small step that we can make so that Sri Lanka can make a giant leap in the gaming industry in the near future.

Localizing the Thunderbird e-Mail Client

Rohana Dasanayaka
Arthur C Clarke Institute
rohana@accmt.ac.lk

The highly used language in computing is English and that is also the same for the software systems used in e-mail communication. The language is a barrier for widening the use of email among majority of the people in Sri Lanka. As a result of using ICT applications based on English, the most of the people are unable to obtain full benefit of the technology. So, having a very high literacy rate in the country, the availability of a local-language software will be lead to wide usage of email communication as a means of communication.

The Mozilla Thunderbird is an email client application developed by Mozilla Foundation. This is a Free and Open-Source Software which is designed with internationalization features which can be used for localization purposes. It is found that this email client has been localized to many other languages. Most of the localization tools which are freely available are developed to localize Open-Source applications.

In the process of localizing this email application, the menus, dialog boxes, error messages, and tool tips were also localized. The most suitable words in Sinhala were used to replace the English terms.

Using localized Mozilla Thunderbird, people can send and receive emails with a system which completely works in Sinhala.

English – Sinhala Glossary Project of the Official Languages Department (OLD)

Senarath Gunasena
Commissioner
Official Languages

Anura Lokugamage
Deputy Commissioner –Finance
Official Languages

The English Sinhala Glossary Project of the OLD was started in 2005 in terms of a decision by the Cabinet. This project envisages to revise the existing glossaries (last revised in 1973 by the OLD) and to add new subjects for the areas that were emerged as independent disciplines during the past few decades. The work process of this project was to retain the existing definitions as much as possible and to add new definitions where revisions are needed and to find definitions for the new terms added to the collection. The total number of subject areas to be covered under the project is 35.

At present 5 glossaries have been printed under this project and are available for sale. Their particulars and prices can be obtained by visiting the OLD's website www.languagesdept.gov.lk.

Since the printed versions of glossaries possess several limitations of their own, the OLD decided to computerize them with the view of adding new features such as sorting for different combinations and audio-visual facets wherever necessary, expanding the accessibility by releasing them to a website and making available a versatile consolidated CD version for sale.

With the assistance of the ICTA, the OLD has now released 20 glossaries to its website (vide above) on a testing basis. These glossaries can only be read as separate books at present, but they will be developed in accordance with the parameters given above in the time to come. Also it is expected to establish a common database to cater to all three languages of the country, i.e. Sinhala, Tamil and English as regards the definition of terms.

An XML Interface to the Madura English-Sinhala Dictionary

Madura Kulathunga

info@maduraonline.com

මධුර ඉංග්‍රීසි-සිංහල ශබ්දකෝෂයේ XML දත්ත ලබා දීම

මධුර කුලතුංග

මධුර ඉංග්‍රීසි-සිංහල ශබ්දකෝෂයේ තොරතුරු දේශීය නිර්මාණකරුවන්ට ඔවුන්ගේ වෙබ් අඩවි සහ මෘදුකාංග නිර්මාණ සඳහා පහසුවෙන් භාවිතා කිරීමට හැකිවන අයුරින් XML දත්ත වශයෙන් ලබා දීම.

ඒ සඳහා යෝජිත XML දත්ත ආකෘතිය හඳුන්වා දීම සහ එම යෝජිත ආකෘතිය දේශීය නිර්මාණ කරුවන්ගේ නිර්මාණ සඳහා භාවිත කිරීමට උචිත වන අයුරින් සිදුකල යුතු වෙනස් කම් ඇතොත් ඒවා හඳුනා ගැනීම.

Madura Online වෙබ් අඩවිය වෙත ලබා දිය යුතු Parameters සහ ඒවා යොමු කල යුතු වෙබ් ලිපිනය පිළිබඳව පැහැදිලි කිරීමක් සිදුකිරීම.

එම XML දත්ත AJAX තාක්ෂණය භාවිතා කරමින් තම වෙබ් අඩවි තුලට ලබා ගන්නා ආකාරය පිළිබඳව සරළ හැඳින්වීමක් ලබා දීම.

Interoperable Transliteration Software

Rohan Manamudali
ScienceLand
armanamudali@gmail.com

Transliteration is the representation of a word or phrase in the closest corresponding letters or characters of a different alphabet or language so that the pronunciation is as close as possible to the original word or phrase.

This "INTEROPERABLE TRANSLITERATION SOFTWARE" transliterates local names, addresses and places from Sinhala to Tamil; Sinhala to English; Tamil to Sinhala; Tamil to English; and English to Sinhala; and English to Tamil.

In our country, especially, the Government organizations are expected to hold lists of local names, addresses and places, in databases in Sinhala, Tamil and English. When names are input in one language it is necessary to give such names in another alphabet. It is of national importance, that when the source language is one of the three, the transliteration into the other two, is accurate, as far as possible.

Transliterating local names, addresses and places has very unique differences than transliteration based text input methods in which the users type text of a language using characters of another language.

Mapping from one system of writing to another should be done very carefully. It is very essential to take into consideration the sociological issues prevalent in the country, and the consequences which may arise due to the incorrect spelling of a name. It is also very important to take into consideration the rules for handling Sinhala and Tamil names and the usage patterns.

The issue of the spelling-variants used in local names, and how they may be transliterated to/from the three languages (Sinhala/Tamil/English) should be very carefully taken into consideration. As stated above unacceptable spelling may have serious sociological consequences.

This transliteration software has been created by considering the above matters as one of ICTA e-services projects and it is given as a web service in order to incorporate it in to any other software which requires transliteration facility (e.g. e-National Identity card project) . In addition to that, a public interface is available for quick access, for citizens to ascertain the way in which their names are given.

A comprehensive set of rules have been defined to convert names of persons, places, addresses, in Sri Lanka. An "exceptions word" dictionary has been created to be used in cases where the logic does nor apply.

Software has a self learning facility to build its own "exceptions word" dictionary. If a name is entered for the first time and if it is an exception to the set of rules defined, the system have the ability to include it in the database with an audit-trail being generated to be confirmed by a supervisor.

This system suggests the most likely variant based on previous selections. In other words, in cases of multiple solutions for a given name as list of solutions is provided ranked according to the usage.

Localization Of Ms Windows Vista & Ms Office

Sampath Godamunne
ScienceLand

Localization can be viewed as a process of targeting and adapting particular software and its components as well as the documentation to a specific locale that is, a countries language, standards, culture and even legal requirements. Software localization implies more than just the mere translation of the product's user interface. It includes translations and other conversion, and takes into account local practices and culture, providing a product which is comfortable to use in the target country.

This localization project was a joint effort of Science Land and University of Colombo School of Computing as the phase 2 of an ICTA project for which ICTA has signed a contract with Microsoft on the localization of Windows Vista and Office, termed Language Interface Pack (LIP) project.

Phase 1 of this effort comprised the creation of a language style-guide that provides guidelines on the tone of localization, a Policheck which provides a list of terms that should be avoided, such as profanity or politically incorrect terms, and a glossary of about 2,000 terms. This was previously done by a group of scholars of this field in Sri Lanka and been handed over to Microsoft in 2007.

In this project, Phase 2 consist of translating from English to Sinhala, approximately one million words, phrases in MS Windows Vista (software words, user assistance files etc..) and in MS Office (software words, user assistance files, user assistance tutorials etc..)

A well selected team of language & IT experts were involved in this effort. Technical professionals (IT), Translators and Consultants (Language and IT) were among them. The group of technical professionals consists of experts in Microsoft Windows Vista and MS Office. This group was responsible for providing the proper meaning

context of the technical words/phrases used in MS Windows Vista and MS Office to the translators and training the translation team on tools and processors. The translators were responsible for translating words and phrases provided to the local (Sinhala) language according to the Style Guide, Policheck and Glossary. Required guidance and the consultancy were continually provided by the language experts and IT experts. These experts were responsible for the ultimate outcome of the project. In addition to that, a moderator was appointed for linguistic review.

The software has been made truly accessible to anyone, anywhere as it can be downloaded free of charge from www.microsoft.com/srilanka.

This language interface pack will help most of the rural area people to familiarize with computer technology. Those who are not comfortable with English but familiar with Sinhala can get the benefit out of this product. Those who already adapted to English interface may have hard time adapting to this Sinhala interface.

Internationalized Domain Names in Sri Lanka

Chamara Disanayake
LK Domain Registry
chamara@nic.lk

In the early stages of Internet it published information in English; could be accessed by locating with Latin characters and supported only for English data input and output. Tradition of considering English as the language of the Internet started to changed when the most popular Internet application such as web and e-mail enabled the support for non Latin characters. With this support, the number of web sites published on Internet as well as the number of users who used Internet showed a rapid growth. People were able to access thousands of web sites published information in many languages other than English. Even though Internet supported for other languages, till 2003 the addressing and naming systems of the Internet was limited for Latin characters. In 2003, ICANN introduced a mechanism for supporting non English Internet naming in networking applications identified as Internationalizing Domain Names in Applications (IDNA). As the domain name systems (DNS) were designed to support restricted ASCII character set, an encoding mechanism were designed to encode/decode non ASCII characters to DNS supported ASCII and vice-versa. This encoding syntax is defined as **Punycode**.

With the motivated environment for IT localization, in Sri Lanka the relevant authorities have identified the requirement and importance of enabling IDN in Sri Lanka. ICTA formed a task force for IDN including expertise in the field. As the initial step the task force proposed two IDN ccTLDs in Sinhala and Tamil as .ලංකා and .இலங்கை. Until the ICANN finalized the IDN ccTLD for each country and implement the IDN support in root servers, the IDN task force has continued IDN implementation within the country. As the ccTLD domain registry for Sri Lanka, the LK Domain Registry played a key role in IDN testing and implementations. It used its resources and expertise to enabled IDN in their systems and implemented the

primary DNS zones. Initially the IDN systems were tested with two ISPs. With the successful testing, the IDN task force requested other ISPs to enable IDN support in their systems. Currently more than 5 ISPs who posses over 60% of the Internet service market share have enabled their customers to use Sinhala & Tamil IDNA.

After enabling IDN in ISPs networks there are few issues to be addressed when using IDN in general networking applications. Currently in some web browsing applications having Unicode – Punycode conversion problems when a Sinhala word having 'Rakaransaya' or 'Yansaya' within it. This will be handled in their future versions. In defining URLs in Sinhala or Tamil, the length of a label would be a limitation as it has a predefined 64 byte length. Even though this length can not be altered in DNS Systems, Sinhal/Tamil users will adopt it as the smaller URLs are much easier and meaningful in use. As the concept of defining URLs in Sinahal/Tamil is new to the domain registries and registrants, proper guidance process should be implemented to support uniform systematic IDN hierarchy. Especially when choosing a domain name in Sinahala/Tamil, the registrant has to decide a meaningful name which gives some understanding about the registrant organization/personal to the domain users where as in English the acronyms are used in general.

Common repository for localized technical terms

Wasantha Deshapriya

Director, Re-engineering Government Programme, ICTA

wasantha@icta.lk

Introduction

Over the last 10 years, there has been many developments in localization of computing. Institutionalization of Sinhala and Tamil Unicode as localized computing standards is the main achievement of all which paved way to openning many paths for localizing ICT environment. www.fonts.lk served as the initial center for providing the information related to Sinhala localization. Once the siyabas.lk, emathumozhihal.lk and locallanguages.lk started assisting and disseminating SU and TU technologies, SU and TU won a huge popularity. Since the above technologies could be used for all sorts of sw development initiative, localization of ICT products was begun.

For example OSs including Microsoft Vista and Linux flavours, Productivity suits such as MS Office 2007 has been partially localized. Also IE 8 and Mozilla Firefox has been fully localized. Moodle which is an open collaborative tool and Joomla which is Open CMT have also been localized.

Problem

The technologists who are involved in localization usually depend on Sinhala language skills of themselves and their who are assisting them. Although www.siyabas.lk had a glossary it became a rarely used source as it is not updated and it is not an collaborative tool. Due to this reason each party has selected and selecting the technical terms in an ad-hoc manner for localizing ICT products. This has lead to a situation where many different term for the same English term being used making the ICT users confused than ever.

Solution

ICTA realized that due to the absence of a reliable, central repository of localized terms the ocaliztion community is using various and varied terms for localization of ICT products. Therefore ICTA is currently working to develop a collaborative tool to capture and display the relevant localized terms for original ICT terms.

Experience in English to Sinhala Machine Translations

Asoka S Karunananda[1], Budditha Hettige

[1]Faculty of Information Technology

University of Moratuwa

asokakaru@yahoo.com

Since the classical era of Artificial Intelligence (AI), the filed of Natural Language Processing (NLP), especially, Machine Translation has been a research challenge. Perhaps, this is one of the least achieved areas of AI to date. According to literature, even major machine translation systems such as EDR by Japanese have still not met the expectations fully. Among others, many European and Asian countries have been working on projects that translate from English to respective mother tongues. Due to very high level of diversifications of constructs of different natural languages, it is rather impractical to develop machine translation systems by adapting an existing translation system. This paper reports on our experience/issues in English to Sinhala machine translation project conducted at the Faculty of Information Technology, University of Moratuwa.

The project on English to Sinhala machine translation was launched in 2006, through an MPhil research project. The results of the project include a comprehensive parser for Sinhala language, Sinhala morphological analyzer, translation engine, transliteration system and several prototype dictionaries. This translation system has been incrementally tested and used for several applications. Some of the major research challenges identified by the projects can be stated as follows. At the outset, we could not largely depend on freely available English parsers due to several reasons. Firstly, such parsers offer only the basic grammar, and full parsers are expensive to buy. Secondly, due to the differences in the construct of English and Sinala languages, parsers could not be linked up directly. For instance, Sinhal language deals with just three tenses, while English has several versions of three major tenses. Furthermore, prepositions is a major language construct for English, while Sinhala has no such concept, yet handles the need for prepositions in a

different manner. Developing a parser for Sinhala language has been the major task of the project. The Sinhala parser has been powered by the concept of *varanagema* in Sinhala language. This approach has drastically reduced the need for storing a large number of word forms in the Sinhala dictionary and also handled the role of prepositions and the concept of persons in English language. Lack of comprehensive dictionaries has also been a major barrier for the translation system. Therefore, we have to develop several prototypes dictionaries for the testing of the machine translation system. In fact, we have developed a special transliteration module for handling the out-of-vocabulary issue and translation of proper nouns. Semantic handling is yet another research challenge in the project. We have introduced the concept of intermediate editing to acquire user's feedback as a means of handling those issues. At present, we are also working on handling of English sentences that are created through complex grammatical structures.

Building Linguistic Resources for Sinhala and Tamil

Dulip Herath
University of Colombo School of Computing
dlh@ucsc.cmb.ac.lk

Language resources are extremely important for computer-based language processing as it requires machine readable dictionaries (lexica) and corpora. Usually, these resources are manually annotated according to predefined linguistic formalisms for example, dictionaries with part of speech tags and phonetic transcriptions using IPA, part of speech tagged corpora, and parallel corpora with sentences aligned. These resources are not only important for language processing but also for linguistic research and language studies for example, corpus linguistics, computer-aided language teaching, and computational linguistics.

Language Technology Research Laboratory (LTRL) of the University of Colombo School of Computing has developed various types of language resources for Sinhala and Tamil. Among them, 10 million word Sinhala text corpus and 30,000 word Sinhala lexicon with English and Tamil translation equivalents, Part of Speech Tagged Newspaper corpus, English-Sinhala Parallel Corpus with Part of Speech tags, Corpus-based Sinhala lexicon are more prominent. It is currently building a Sinhala WordNet application based on English WordNet synsets and most frequently occurring words of the Sinhala corpus, morphological analyzer/generator, pronunciation dictionary and a Sinhala speech corpus for telephone query domain. LTRL has also developed part of speech tagsets and transliteration schemes for different annotation and implementation purposes.

LTRL is also focusing on the issues related to disseminating these resources to the researchers and the general public under a proper licencing scheme that protects the intellectual property rights of the contributors.

Building and extending TTS for Sinhala

Viraj Welgama
University of Colombo School of Computing
www@ucsc.cmb.ac.lk

Text-to-Speech (TTS) system for Sinhala is a very useful application especially for the people who are visually handicapped. This tool was developed using the Festival framework. Construction of a diphone database and implementation of the natural language processing modules were designed after doing a comprehensive study of Sinhala phonology. The TTS system accepts direct Sinhala Unicode text input and then it is converted to Festival's context sensitive rule format by rewriting letter-to-sound rules. Sinhala syllabification algorithm was implemented as required for letter-to-sound rules. A Modified Rhyme Test (MRT) was conducted to evaluate the intelligibility of the synthesized speech and yielded a score of 71.5% for the TTS system. Future work will mainly focus on improving the naturalness of the synthesizer and expanding the pronunciation lexicon. Work is in progress to improve the prosody modules.

Sinhala Language based projects developed by the DSCS

R. Bandara
ravimal97@gmail.com
G. S. Makalanda
gnaka@sjp.ac.lk
P. Indranama
pravin@indranama.com
T. G. I. Fernando
gishantha@lankamail.com
B. Hettige
budditha@yahoo.com

Department of Statistics and Computer Science,
Faculty of Applied Science, University of Sri Jayewardenepura

Sinhala language is one of the constitutionally recognized official languages of Sri Lanka, along with Tamil. Development of the local language systems (Sinhala & Tamil) is gain challenges for many researchers. At present variety of Sinhala language systems have been developed including Sinhala Unicode System, number of Local language dictionaries (Madhura, EnSiTip). English to Sinhala Machine Translation System (BEES), Chatbot etc. However, according to the web-based survey we have identified that, usage of the local language technologies are not comparative. This is because many computer users do not have idea about new local language technologies and many of these Sinhala language users can't use the Sinhala Wejesekara keyboard. According to over survey, these are some critical problems in the localization. This paper presents new development of the English/Sinhala popup dictionary (Vidudaya dictionary) and Sinhala typing tutor (Siyapatha) that are developed by the Department of Statistics and Computer Science, for the fiftieth anniversary of the University of Sri Jayewardenepura.

The Vidudaya dictionary is a software, that can be used to find a suitable Sinhala words for the given an English word and vice versa. It has two operating modes namely keyboard enable mode and mouse events enable mode. In the keyboard enable mode, user can select a word and need to press some given key combination to find the suitable Sinhala words. The mouse events mode needs only the select a suitable word by using the mouse. When the user selected an English word or a Sinhala word in anywhere in a file (E.g., Word

document, PDF document, any web page, etc.), Vidudaya popup dictionary shows the related words for the selected English or Sinhala word. This is the most usable feature of the Vidudaya dictionary. The system has been developed by using a Microsoft Visual Basic and dictionary has been implemented using a Microsoft Access database.

The Siyapatha is a Sinhala typing tutor that can be used to improve user's typing performance of the Sinhala Wijesekara keyboard. The system structure is based on the typing master, which is most popular English typing tutor in the world. The system provides four lesions; with several activities. Using this software the user can improve the typing skills to a reasonable level.

The Vidudaya dictionary and Siyapatha Sinhala typing tutor are freeware and anybody can download from departmental website.

The Making of a Web Site of Sinhala Song Lyrics

Dr. S.C. Jayewardena
University of Peradeniya
sakzje@gmail.com

The standards needed for creating, storing and sharing information in local languages have come a long way. This is largely the case for the Sinhala language, too. Today we can, in fact, use most of the popular productivity software like word processors, spreadsheets, etc. with Sinhala language. E-mail can be read and written using Sinhala. The ability to create local language content for the web and store information in databases immesely add to the usability of computers with local languages. All this can be done using the standards. This is important if we are to make certain that information is accessible to every one with ease and to ensure compatibility with future developments. However, it appears that the use of the standards is not as widespread as one would like to see. Therefore, any remaining minor issues with the standards must be resolved and the adoption of the standards and their use to create useful content need to be encouraged.

To this end, more concrete examples of local language use that go beyond the simple 'proof of concept' must be provided. The www.LankaSongBook.com web site actually provides a real life example of the utilization of Sinhala language standards on the web. The web site, although it has a '.com' top level domain, is a non-commercial web site consistsing of a large collection of lyrics of popular Sinhala songs. It uses the typical, and powerful, combination of a web front end and a database back end to present itself to the user. The Sinhala version of the site has all information presented using the Sinhala language. All the vital information is stored in the database using the utf-8 (Unicode) character set and Sinhala collation. The lyrics themselves are available as Unicode text. Using the web front end, users can perform searches using the local language. Furthermore, the GUI tools developed for the maintenance of the database also use utf-8 characters.

From the web site user's perspective, the site's strength lies in its amount of useful content and the ease of access to that content. It was necessary to have an easy-to-use input method to get the large volume of content. As a developer only familiar with the English standard keyboad, the author used a particular input scheme called 'sumihiri' for this task. This scheme was developed as far back as 1995 in connection with a project that used TeX/LaTeX for typesetting Sinhala documents. A module that is added to the SCIM input scheme (on a Linux system) allows fast and efficient direct typing of Sinhala Unicode using 'sumihiri'. The site also has a link for users who are interested in getting a taste of using Sinhala on the computer for reading and writing.

Local Language Content Preservation through Digitization and Archiving

Shaseevan Ganeshanathan
Executive Director, Noolaham Foundation
shaseevang@gmail.com

"The ready availability of relevant local language content is critical for the development of productive capacity in new media… Without locally relevant content in local languages, immediate uses of ICTs for day-to-day activities may not be apparent."

- Danny Butt and Madanmohan Rao

(PAN Localisation - Authors' Working Group on Content)

Different categories of tamil speaking communities exist in Sri Lanka. Jaffna tamils, Batticaloa tamils, Upcountry tamils and Muslims in the country speak their own mother tongue. These communities are considered as minorities in Sri Lanka, but they are dispersed all over the world in different countries and states. They have the need to preserve their language content through various means. In order to conduct research in ancient Sri Lankan local language content and to make researchers involve in the welfare of their communities, local content has to be digitized and archived. These dispersed communities can then be brought together by making this research and development online. "Project Noolaham (PN)" is a non-profitable, collective and voluntary endeavor aimed at preserving Sri Lankan tamil related publications and research papers in digital format in order to make them available to and for the benefit of all those who are able to access the internet. PN receives contribution in many ways from many people and organizations, like dedicated volunteers and the Noolaham Foundation who are the main contributors. Their direct participation in digitization, their monetary donations, wiki contribution and coordination of projects are the most obvious ones.

PN chooses open and free softwares whenever possible. Joomla and Mediawiki are the content management systems used so far and both

were non-proprietary ones. These Open Source systems enable users to contribute easily to the development of the site by registering. The system stores tamil books, magazines, newspapers, pamphlets and research papers in digitized format. They can be searched according to the author, year of publication, publishers and subjects which make research very easy for the researchers and people who are interested in it. In future PN looks forward to implement web 2.0 concepts in to the system where researchers/users can come and discuss issues about the content through forums and provide their comments. New efforts on communicating this to schools, universities, libraries and private institutions are being taken. We are also looking forward to introduce international standards like the "Dewy Decimal Classification" to MediaWiki to improve the reliability and efficiency of the system at an international level.

e Fusions Contribution in the Sinhala Technology in Sri Lanka

Thusitha Rangunuge, Suranga Sampath Jayakodi
e Fusion
thusitha@info.lk

e Fusion was one of the Pioneers in font development & Sinhala technology in Sri Lanka. The first Sinhala True T font was developed in 1992 by Niranjan Meegammana & Mr. Micheal Grober. From there on, e Fusion started to contribute to the local native language technology with developing its first Phonetic font "Kandy" in 1995 & developed its first Sinhala Word Application with a Rich Text Editor in 1996.

In 1997 Lankadeepa started using e Fusion's fonts for the first online Sri Lankan Newspaper. With the success of the Sinhala Word Application, it was upgraded with new features & the Tamil language Editor "TamilWord98" was developed & was continuously upgraded & finally "SinhalaWord 2000" with new features & more fonts with Tamil language support was introduced in 1999.

One of the more famous local Phonetic fonts "Kaputa.com" font was developed in 2000 & most of the Sri Lankan Online News papers like Divaina, Lakbima & Ravaya was using it in their online Newspapers.

After the development of "Kaputa2004", e Fusion developed its first Unicode Font "Kaputa Unicode". e Fusion also played its part on popularizing the usage of Sinhala Unicode & contributed to the development of the Sinhala Unicode Technology by creating Discussion groups like "Sinhala Unicode Group"

From then on e Fusion has developed several True Multilingual Unicode supported dynamic websites & www.gov.lk is also one of them.

Farmer education in Sinhala through Online Distance Learning

Vidanapathirana N.P[*]., Hirimburegama K.

[*]Universitry of Colombo Institute of Agro-technology and Rural Sciences

**nisansalavp@yahoo.com*

Sri Lanka is an agricultural country with huge potential to develop agriculture sector to new era. Farmer plays a major role of this event. But Sri Lankan farmers do not get maximum profit from agriculture products due to lack of updated knowledge on new techniques and lack of information to go for a better production. Introduction of a new method for farmer education fills up the knowledge gap between modern world and farming community. This will help to transfer knowledge, information and technology to the farmers. The Institute for Agro-technology and Rural Sciences University of Colombo has a mandate to "providing, promoting and developing higher education in the discipline of Agro-Technology and Rural Sciences among persons presently engaged in Agriculture and Agro-Technology". It will cater to the changing individual and social needs by taking education through a blended learning mode to the doorsteps of the farming community who never dreamt of entering the portals of higher education in their local language. However they have aspiration and potential to pursue higher education but could not utilize the opportunities for personal, family or economic reasons. With the flexibility in the entry requirements, choice of the language of learning and choice of courses, the Institute demonstrates that it is possible to impart quality higher education using modern information communication technologies (ICT) to learners of the farming community of poor rural sectors to pursue their education at a pace and place convenient, breaking language barrier to them and realize their academic, professional and social aspirations. There are lots of courses in on –line internet based but in English language. But the online Diploma Programme in Agro – Technology is delivered in Sinhala medium.

The institute follows a well organized plan of on-line course development put in place by Distance Education Modernization Project (DEMP) to ensure the quality and the effectiveness of learning material in Sinhala language encouraging self-directed learning. Moodle Learning Management System (LMS) is a state - of - the art software which is designed to enable the student to navigate easily and it can be designed and moderated in local languages also. In our case it is designed in Sinhala medium. Learning out comes; introduction, course navigation path ways and video audio lessons are developed in Sinhala language in well organized manner for all lessons of the course. The students are supported by dedicated on-line tutors and mentors using discussion forums. Tutors and learners are discussing through online forums in the LMS in Sinhala language. At every phase of the on-line education process student receives on-going support. Students of the course (Farmers) also post their problems and comments relating to the course and also regarding their cultivations to the forums and discuss with other colleagues in Sinhala medium. Therefore, online education is turned for a new path of giving knowledge not only for the people who can use international language but also for the people who used localized language. Online Agro-technology Diploma of the Institute for Agro-technology and Rural Sciences University of Colombo is the best evidence for the above statement.

A Digital Educational Bridge to connect 1000 Rural Schools in Sri Lanka

Sameera Wijeratne
NenasaTV
Sameera.Wijerathna@dialog.lk

In an effort to bridge the rural urban digital divide in the education system, Dialog Telekom has gifted the Ministry of Education and Sri Lanka's student population a Digital Bridge – a satellite based television channel designed for distance learning. Following the successful pilot in year 2006/2007, the main broadcast medium was altered from broadband technology to Satellite Digital Video Broadcast (DVB) to reach the most remote schools whilst maintaining cost efficiencies. As a holistic educational programme the Digital Bridge project has weighed all three aspects of connectivity, access and content to make it more effective. 'Nenasa' will connect 1,000 rural schools in Sri Lanka to high-quality rendition of the national curriculum, developed by the National Institute of Education (NIE) over a digital satellite television broadcast medium.

'Nenasa' is a joint effort of Ministry of Education (MoE) National Institute of Education (NIE) and Dialog Telekom. The project has been funded by Dialog Telekom as a part of its Corporate Responsibility programme under the theme of Information and Communication Technology for Development (ICT4D).

The NIE develops the contents for 'Nenasa', which in turn telecasts by Dialog to 1,000 identified schools connected to the network. 'Nenasa' will be dedicated towards broadcasting educational content and cater to the Ordinary Level (O/L) and Advanced Level (A/L) syllabus in Sinhala and Tamil. In addition, teacher-training and skills development would also be included in the programme line-up. These lessons will be supplemented by a Learning Management System (LMS) through which the interactive element of 'Nenasa' is facilitated. The LMS enables children who have access to the internet to download educational material on Nenasa from the

internet via www.nenasa.lk. The Learning Management system was developed by the Dialog – University of Moratuwa (UoM) Mobile Communications Research Lab. As part of this programme, Dialog Telekom will also donate television reception equipment and audio visual devices to the rural schools identified by the Ministry of Education for connection to the Nenasa network.

H.E. the President together with Minister of Education, Susil Premjayantha, launched the NenasaTV from a classroom in a Ranjan Wijerathne Maha Vidyalaya in Pelwatta Moneragala on 20th July 2009 opening a new vista in access to education.