Celebrating PAN L10N Impact Sri Lankan Component

Language Technology Research Lab
University of Colombo School of Computing
Sri Lanka

Overview

- Introduction
  - Sri Lanka, Sinhala Language, UCSC, LTRL
- PAN L10N Project
- Derivatives
- Trainings
- Impact
  - Human Resources
  - LTRL
- Lessons Learned
- Publications

Introduction

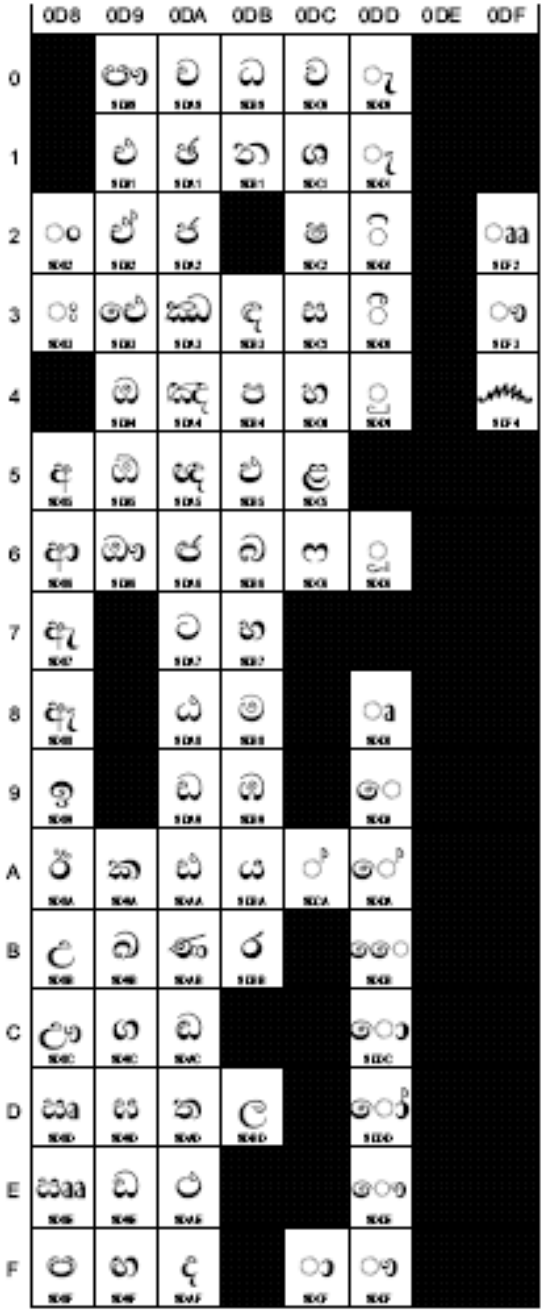
- Sri Lanka
  - An island in the Indian ocean
  - Area - 65,610 km2 (122nd)
  - Population - Above 20M (56th)
  - Two official languages
    - Sinhala – 74%
    - Tamil – 18%

Introduction

- Sinhala Language
  - Mother tongue of the Sinhalese
  - 15.6 million speakers
  - has its own writing system
    - 18 Vowels
    - 2 Semi Vowels
    - 40 Consonants
  - belongs to the Indo-Aryan branch of the Indo-European languages
    - Morphologically rich (Nouns-110, Verbs-282)
    - Use vowel modifiers

Introduction

Sinhala



Introduction

- University of Colombo School of Computing (UCSC)
  - Established in 2002 by merging the Institute of Computer Technology and The Department of Computer Science of the University of Colombo
  - First centre of higher learning of computing in Sri Lanka
  - Involved in many research areas
    - Natural & Local Language Processing, Human Computer Interfaces, Image Processing and Vision, Cryptographic Systems, Multi Media and Virtual Reality, Intelligent Agent Systems, Pattern Recognition, Distributed Systems, Information Retrieval and Data Mining, Process Broker Modeling Systems, Web Based Business Services, Multi Media Database Systems, e-Learning, Strategic Planning & Management of IT, IT Policy and Multi Database Systems.

Introduction

- Language Technology Research Laboratory (LTRL)
  - Was established in 2004
    - Under the PAN L10N project funded by IDRC, Canada
  - Aim is to address the growing need of local language computing in Sri Lanka by doing Localization and Language Processing research and development

PAN L10N Projects

- PAN L10N phase 01
  - March 2004 – March 2007
- PAN L10N phase 02
  - April 2007 – March 2010

PAN L10N Projects

- PAN L10N phase 01
  - March 2004 – March 2007
- PAN L10N phase 02
  - April 2007 – March 2010

Phase 01 (March 2004 - March 2007)

- Linguistic Resources
  - 10M Words Contemporary Sinhala Corpus
    - Considered text published after 1948
    - 3 major categories
      - Newspaper (33.88%)
      - Technical Writing (43.03%)
      - Creative writing (23.11%)
    - ~440,000 distinct words
    - Derivatives
      - Most Frequent words
      - Sinhala Character Distribution
      - List of Sinhala Stop words
      - List of Sinhala Nouns and Verbs
      - List of Proper Names

Phase 01 (March 2004 - March 2007)

- Linguistic Resources
  - Tri-Lingual (Sinhala-Tamil-English) Dictionary
    - Based on a Sinhala dictionary complied by the National Institute of Education
    - Added more words from a Tri-lingual dictionary compiled by the department of official languages, Sri Lanka
    - Contains more that 25,000 words
    - Included grammatical details

Phase 01 (March 2004 - March 2007)

- Linguistic Resources
  - UCSC Sinhala POS TAG set
    - Referring other Indic language Tagsets
    - 22 unique tags
    - Provides the necessary and sufficient information to a morphological parser that can generate enriched detailed tags for any given word
  - 500k Sinhala Tagged Corpus
    - Text Selected from 10M UCSC Sinhala Corpus
    - Manually Tagged by Language Experts

Phase 01 (March 2004 - March 2007)

- Language Tools
  - Sinhala Text to Speech System
    - Diphone concatenation method
    - Used the Festival framework
    - Compiled the Sinhala TTS as MSAPI compatible
    - Most innovative Product 2008 – Infotel
  - Sinhala Screen Reader
    - Embedded Sinhala TTS with Thunder
    - Widely use by the blind community

Phase 01 (March 2004 - March 2007)

- Language Tools
  - Optical Character Recognizer for Sinhala
    - Font dependent
    - High accuracy with cleaned images
  - Encoding Conversation Utilities
    - Converts more than 20 most common proprietary encodings to Sinhala Unicode standard
    - Online and offline versions
    - Widely using by the community

PAN L10N Projects

- PAN L10N phase 01
  - March 2004 – March 2007
- PAN L10N phase 02
  - April 2007 – March 2010

PAN L10N Projects

- PAN L10N phase 01
  - March 2004 – March 2007

- PAN L10N phase 02
  - April 2007 – March 2010

Phase 02 (April 2007 - March 2010)

- Linguistic Resources
  - 100k English-Sinhala Parallel Corpus
    - English Text selected from PENN Corpus
    - All other partner countries of PAN L10N project translated the same English text to their languages
    - Rich resource for various kinds of inter-language processing activates
  - 1000 words Sinhala Wordnet
    - Based on a PWN
    - Aimed to form a valuable linguistic resource for various NLP tasks
    - Started with most frequent Sinhala words extracted from 10M UCSC Sinhala corpus
    - Contributing to the AWN

Phase 02 (April 2007 - March 2010)

- Language Resources
  - Localized URL
    - Defined gTLDs and ccTLDs for Sinhala
    - Approved by the local authority (ICTA)
  - Teaching materials for local languages teaching and learning
    - Developed teaching materials to teach Sinhala
    - Contents developed for situation based learning
    - Focused on Adult learners
    - Implemented using the Language Teaching Framework

Phase 02 (April 2007 - March 2010)

- Language Tools
  - Language Teaching Framework
    - Independent from language
    - All the data are stored in XML files
    - Developed contents to teach Tamil in Sinhala and Sinhala in English
    - Special Merit - e-Swabhimani 2010
      - In the category of e-Learning an Education
  - Translation Memory for Machine Assisted Translations
    - Based on Omega T
    - Used for traslating teaching materials

Phase 02 (April 2007 - March 2010)

- Services
  - Training on Sinhala web content development
    - Developed training materials
    - Conducted trainings
      - Universities
      - Professional institutes
      - Government offices
    - Helped to increase the local language content in the web

PAN L10N Project - Derivatives

- Sinhala Lexicon
  - Words extracted from 10M UCSC Sinhala Corpus
  - Manually categorized in to main POS categories
  - Introduced a novel classification to each category which essential for computational linguistics
  - Used standard policies for word separation and spellings
  - Coverage over 95% for refine text
- Sinhala Spell Checker
  - Rule based
  - Used statistics of UCSC 10M Sinhala corpus

PAN L10N Project - Derivatives

- Subasa.lk
  - For Sinhala language computing services
    - Spell checker
    - Keyboards
      - Phonetic, Online keyboard
    - Encoding converters
    - Ingiya – A Si-En pop-up dictionary
    - Training videos
    - Related links
  - Winner - e-Swabhimani 2010
    - One of the country's best e-content applications

PAN L10N Project - Derivatives

- Compiled the glossary for IT
- Localized windows Vista, Windows 7, Office 2007 & Office 2010
- Spell Checker application for MS Word
- Defining Sinhala Collation sequence
- Defining Sinhala Locale for CLDR
- Defining official country names and country codes in Sinhala for CLDR

PAN L10N Project - Trainings

- Unicode awareness programs
  - Government offices
    - Election Commissioner office, Ministry of Samurdi, Ministry of National languages, National House Development Authority, Land Ministry
  - Universities
    - Sabaragamuwa University of Sri Lanka (academic and non-academic staff, students in many faculties), University of Peradeniya (academic and non-academic staff), University of Colombo (students in the Arts faculty)
  - Professional Institutes
    - Sri Lanka Collage of Journalism
  - Organizations
    - Kotapola Multi Purpose Co-operative Society

PAN L10N Project - Trainings

- Web Content Development trainings
  - University students and staff
  - Students at other Professional institutes

- On-call help for local language related issues
  - Government offices, web developers, journalists, organizations, individuals, research students

PAN L10N Project - Impact

- Rapid increase of using computers in local languages
  - Services at Subasa.lk
- Rapid increase of local language contents in the web
  - Sinhala Wikipedia articles
    - 5 articles (2004) to 6,300 articles (2011)
  - Sinhala Blogs
  - Sinhala newspapers
  - All government websites
  - Sinhala Glossaries & Dictionaries
- Accessibility of the blind community

PAN L10N Project - Impact

- Human Resources
  - Produced computational linguists who have many experiences & knowledge in both Computer Science and Linguistics
    - 3 computer science graduates followed masters in linguistics
    - 3 linguistics graduates have been trained many years in computer labs
    - 3 computer science graduates have gained an intensive training on computational linguistics at Lahore in 2006
  - 5 Mphils in the field of NLP
  - 6 candidates are reading for PhDs
  - Undergraduate programs in NLP & Software L10N

PAN L10N Project - Impact

- Language Lab @ UCSC
  - LIP Project (Localizing Windows Vista & 7 and Office 2007 & 2010)
  - Microsoft Sinhala Speller Project
  - Speech Recognition with OnMobile Ltd., India
  - Teaching Materials Translation
  - Mini projects with Local Entities
    - Defining Sinhala Collation order
    - Standardizing country names and country codes in Sinhala
    - Defining Sinhala terminology glossary for Microsoft

PAN L10N – Lesson Learned

- Technology needs to be user-centric
  - Holistic planning: not just how to develop (and hope for downloads) → how to deploy widely
  - Need to be ready to learn: from teaching UNICODE, keyboards and fonts → training on blogging & wiki
  - Being prepared to take a back seat: from organized training → becoming a 'background service' in bigger movements (Sinhala UNICODE and Bloggers groups)
  - Importance of help-desk access: from implicit → explicit service?

(Successful in Sri Lankan context)

PAN L10N – Lesson Learned

- Importance of linking with strategic 'boundary' partners
  - Top-down: ICT Agency, Standards Institution, Ministries of Language, Education
  - Bottom-up:
    - Vendors (the 'frontline') – limited success in SL
    - Community based groups – LUGs, NGOs
  - New lesson: different messages for different 'boundary' partners

(Relatively successful in Sri Lanka)

PAN L10N – Lesson Learned

- Need to broad-base dissemination and adoption
  - Starting young: school curricula (NIE)
  - Being pervasive: incorporating into all IT Awareness, Driving License and university IT training
  - Initiating new discipline (Comp. Linguistics): through courses in university CS programmes, guest lectures and student projects in Linguistics programmes

(Limited success in SL, but gaining ground)

PAN L10N – Lesson Learned

- Targetting technology as low in the stack as possible
  - From Hacks → Applications → Platforms → Hardware
  - Not standalone LL apps, but libraries and 'platforms' (current porting to iBus for Linux input; Android input methods instead of separate apps)
  - Web as platform (OS independent): 'any' browser input method, spell checker etc (Javascript works offliine!), web services
  - Target out-of-the-box 'experience'

(Some success in SL, room to go further)

PAN L10N Project - Publications

1. Jeevanthi Uthpala Liyanapathirana and Ruvan Weerasinghe (2011). *English to Sinhala Machine Translation: Towards Better information access for Sri Lankans*. Conference on Human Language Technology for Development 2011. Alexandria, Egypt, 3-5 May 2011.
2. Thilini Nadungodage and Ruvan Weerasinghe (2011). *Continuous Sinhala Speech Recognizer*. Conference on Human Language Technology for Development 2011. Alexandria, Egypt, 3-5 May 2011.
3. Randil Pushpananda, Chamila Liyanage, Namal Udalamatta and Ruvan Weerasinghe (2011). *Content independent open-source language teaching framework*. Conference on Human Language Technology for Development 2011. Alexandria, Egypt, 3-5 May 2011.
4. Viraj Welgama, Dulip Lakmal Herath, Chamila Liyanage, Namal Udalamatta, Ruvan Weerasinghe and Tissa Jayawardana (2011). *Towards a Sinhala Wordnet*. Conference on Human Language Technology for Development 2011. Alexandria, Egypt, 3-5 May 2011.
5. Tissa Jayawardena, Chamila Liyanage, Namal Udalamatta, Randil Pushpananda and Ruwan Weerasinghe (2011). *A Study on the Course of Study of Sinhala*. 4th Research Conference, Royal Asiatic Society, Sri Lanka. 26-26 March 2011.
6. Ruvan Weerasinghe, Tissa Jayawardhane, Vincent Halahakoon, Dulip Herath, Chamila Liyanage, Viraj Welgama and Namal Udalamatta (2010). *A Semantic Study of Sinhala Words*. Annual Research Symposium, Faculty of Graduate Studies, University of Kelaniya, Sri Lanka. 30th November 2010.

PAN L10N Project – Publications

7. Asanka Wasala, Ruvan Weerasinghe, Randil Pushpananda, Chamila Liyanage and Eranga Jayalatharachchi (2010). *A Data-Driven Approach to Checking and Correcting Spelling Errors in Sinhala*. The International Journal on Advances in ICT for Emerging Regions 2010.
8. Asanka Wasala, Ruvan Weerasinghe, Randil Pushpananda, Chamila Liyanage and Eranga Jayalatharachchi (2009). *An Open-Source Data Driven Spell Checker for Sinhala*. e-Asia 2009. Colombo, Sri Lanka, 2-4 December 2009.
9. Weerasinghe, A. R., Liyanapathirana, J. U., Asanka Wasala, Dulip Herath, Viraj Welgama (2009). *OpenTM: A Translation Memory System for Complex Script Languages*. International Conference on Machine Translation Twenty-Five Years On, Bedfordshire, UK, 21-22 November 2009.
10. Ruvan Weerasinghe, Dulip Herath and Viraj Welgama (2009). *A Corpus-based Sinhala Lexicon*. In Proceedings of the 07th Workshop on Asian Language Resources, Singapore, 6-7 August 2009.
11. Asanka Wasala, Ruvan Weerasinghe (2008). *EnSiTip: A Tool to Unlock the English Web*. 11th International Conference on Humans and Computers, Nagaoka University of Technology, Nagaoka, Japan, 20-23 November 2008.
12. Harsha Wijayawardana, Asanka Wasala, Ruvan Weerasinghe and Chamila Liyanage (2008). *Implementation of Internet Domain Names in Sinhala*. International Symposium on Country Domain Governance. Nagaoka, Japan, 20-23 November 2008.

(*http://www.ucsc.cmb.ac.lk/ltrl/?page=publications(=en&style=default*)

PAN L10N Project - Publications

13. Silva, A. M. and Weerasinghe, A. R. (2008). *Example Based Machine Translation for English-Sinhala Translations*. In Proceedings of the 09th International IT Conference (IITC 2008), Colombo, Sri Lanka, 27-28 October 2008.
14. Ruvan Weerasinghe, Asanka Wasala, Dulip Herath and Viraj Welgama (2008). *NLP Applications of Sinhala: TTS & OCR*. Proceedings of the 03rd International Joint Conference on Natural Language Processing (IJCNLP 2008), Hyderabad, India, 7-12 January 2008.
15. Ruvan Weerasinghe, Asanka Wasala, Viraj Welgama and Kumudu Gamage (2007). *Festival-si: A Sinhala Text to Speech System*. Proceedings of Text, Speech and Dialogue, 10th International Conference, TSD 2007, Pilsen, Czech Republic, 3-7 September 2007. (2007) 472-479.

    (http://www.ucsc.cmb.ac.lk/ltrl/?page=publications(=en&style=default)
16. Ruvan Weerasinghe, Asanka Wasala, Samantha Mathara Arachchi (2007). *Facilitating Information Accessibility for the Print Disabled*. Diriya 2007 - a conference on "Mainstreaming Disability into Development". Colombo, Sri Lanka.

    (http://www.ucsc.cmb.ac.lk/ltrl/?page=publications(=en&style=default)
17. Kandeepan S. and Weerasinghe A.R (2007). *Text Based Tamil SMS for Mobile Devices*. Proceedings of the e-Asia Conference 2007, Kuala Lumpur, Malaysia.
18. Ruvan Weerasinghe, Dulip Herath, Nishantha Medagoda (2006). *A KNN based Algorithm for Printed Sinhala Character Recognition*. 08th International Information Technology Conference. Colombo, Sri Lanka, 12-13 October 2006.

PAN L10N Project - Publications

19. Asanka Wasala, Ruvan Weerasinghe, Kumudu Gamage (2006). *Sinhala Grapheme to Phoneme Conversion and Rules for Schwa Epenthesis*. COLING/ACL 2006 Main Conference Poster Sessions. Sydney, Australia (2006) 890-897
(*http://www.ucsc.cmb.ac.lk/ltrl/?page=publications(=en&style=default*)
20. Ruvan Weerasinghe, Dulip Herath, Kumudu Gamage (2006). *The Sinhala Collation Sequence and its Representation in UNICODE*. Localisation Focus - The International Journal for Localisation. March 2006. 13-19
(*http://www.ucsc.cmb.ac.lk/ltrl/?page=publications(=en&style=default*)
21. Ruvan Weerasinghe, Asanka Wasala, Kumudu Gamage (2006). *A Rule Based Syllabification Algorithm for Sinhala*. 02nd International Joint Conference on Natural Language Processing (IJCNLP-05). Jeju Island, Korea, 11-13 October 2005. 438-449
(*http://www.ucsc.cmb.ac.lk/ltrl/?page=publications(=en&style=default*)
22. Herath, D.L. and Weerasinghe, A.R. (2004). *A Stochastic Part-of-Speech Tagger for Sinhala*. Proceedings of the 06th International Information Technology Conference (IITC'04), Colombo, Sri Lanka, 27-28 December 2004.