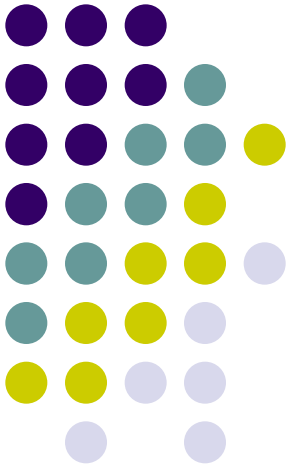
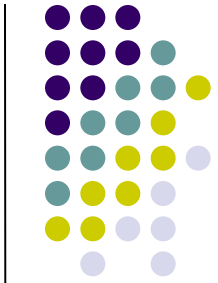
Development of Standards for Sinhala Computing

Gihan Dias
Aruni Goonetilleke
ICT Agency

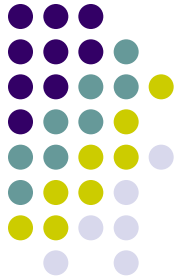


Introduction



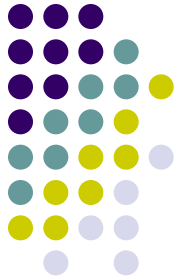
- Most computer use in Sri Lanka is in English
- Most people in Sri Lanka prefer to use Sinhala or Tamil
- For people to benefit from the IT revolution, they should be able to use computers in their own language
  - not only computers but mobile phones, games, and other electronic devices

Multilingual computing



- Computers originated in English speaking countries
- Initially, they all ran in English
- Countries such as Japan, Thailand introduced local character sets
  - were standardised and became common
- Did not happen in Sri Lanka
  - why?

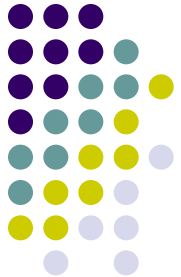
What is required for local language support?



- Character Encoding
  - how letters and words are encoded in a system
  - required for document portability
  - not limited to a specific font
  - not limited to a specific appliaction (e.g. MS Word)
- Fonts
  - how text is represented on a screen or printer
  - many ways of writing a letter

    e.g.: g 𝔤 𝕘 g $\boldsymbol{g}$ $\mathbf{g}$
  - scope for artistic expression

Local Language Support



- Text input
  - from keyboard, pen, voice recognition, etc.
  - keyboard layout
  - key sequences
- Application support
  - each application must have local language menus, error messages, help screens, etc.
- Utilities
  - spelling checkers

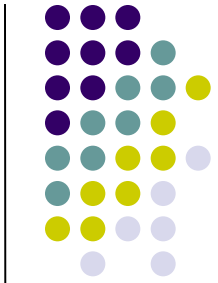
Current Sinhala Technology

- Fonts
- Packages
  - Bundle a font with applications (e.g., a word processor)

Features of Current Systems:

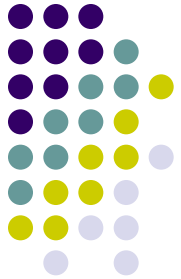
- 8-bit Character Set
- Character codes based on keyboard layout
  - e.g. ක may be represented by the same code as k

Issues with existing packages



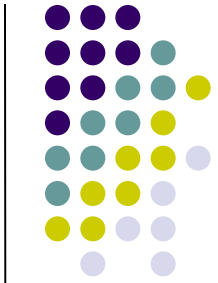
- Lack of needed letters
- Problems with *pili*
- *Collation and Searching*
- *Lack of a Standard – causes problems with*
  - *Transferrring Documents*
  - *e-mail*
  - *Web*
  - *Databases*

Review of Sandardisation Work



- CANLIT – defined Sinhala alphabet
- SLASCII – Character Encoding – 1996
  - SLS1134
- Introduction of Sinhala into Unicode – 1997
- Revision of SLS1134 to conform with Unicode – 2001
- CINTEC Sinhala Fonts committee – 2003
  - To understand why Standard Sinhala was not being used, and take remedial measures

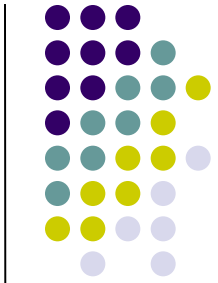
Unicode



- Standard method to represent many languages
- Supported by most modern computer systems
- Tamil has been included for some time

- International standard for language representation

Why no Unicode adoption?

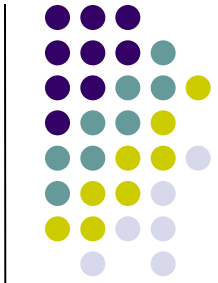


- Although Sinhala has been included in Unicode, adoption has been slow

Why?

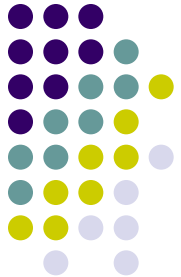
- A lack of awareness of Unicode
- Incompatibility with legacy systems
- Unicode's complexity
- Lack of support in MS-Windows (till now)

Work done by CINTEC, SLSI and ICTA



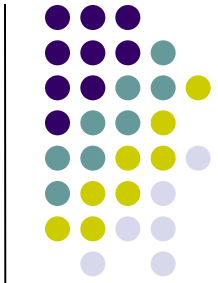
- Encoding Standards
- Fonts
- Keyboard
- Dissemination
  - Sinhala Language pack for Windows XP
  - Joint effort of ICTA, Microsoft, ANCL, Microimage and others

Encoding



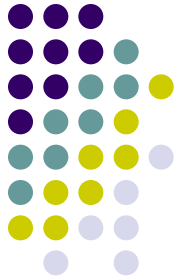
- Must be able to represent all contemporary and classical Sinhala text
- Should facilitate collation and searching
- Should be efficient

Encoding methods



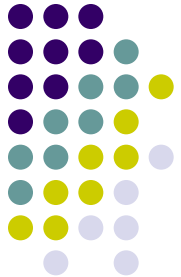
- One code per symbol
  - e.g. ක = 100, ා = 150, ෙ = 152
- One code per letter
  - කෝ = 54
- Unicode uses one code for each consonant and another for each vowel modifier
  - e.g., කෙ is represented by two codes, ක + ෙ

Shortcomings of Unicode



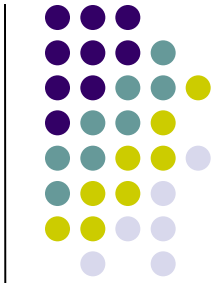
- Lack of encodings for *bandi akuru* such as ක්ෂ,
- Lack of encodings for the *yansaya*, *rakaransaya* and *rephaya*,
- Lack of guidance on the use of multiple vowel modifiers and
- Lack of guidance on the encoding of non-standard letters, such as කූ, රෑ and ළු.

Conjunct Letters (බැඳී අකුරු)



- Shorthand for writing a pure consonant followed by another letter
  - e.g., න්ද = ඳ
- use the Unicode zero width joiner (ZWJ) to indicate a conjunct
  - e.g., න + ් + ද = න්ද
  - න + ් + zwj + ද = ඳ

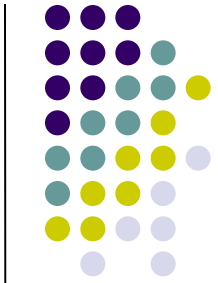
Yansaya and Rakaransaya



- These Symbols are not included in Unicode

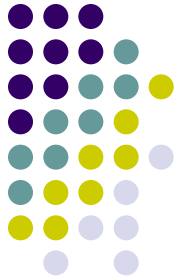
- Why?
  - They are not Sinhala letters
  - are a shorthand for a ය or a ර following a pure consonant
- How are they represented?
  - e.g., ක + ් + zwj + ර = ක්‍ර

Non-Standard Letters



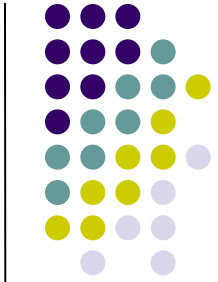
- Pa-pili
  - Use the same code for all pa-pili e.g. කූ
- රෑ
  - Represent as ර + ූ
  - ර + ෑ = ඍ
- ළු
  - Represent as ළ + ු.

Fonts



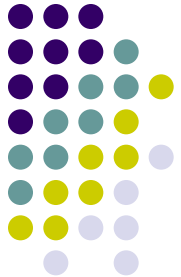
- Older fonts did not support complex scripts
- Newer font technologies such as OpenType contain rules specifying what glyphs to display for character sequences.
- Worked with Font Developers to introduce them to the new technologies

Operating System Support



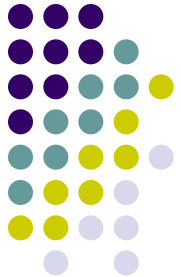
- Windows 95, 98 etc. do not support Unicode
- Windows 2000 supports Unicode but not *complex scripts*
- Windows XP supports Tamil, etc. but not Sinhala
- Microsoft has introduced support for Sinhala, but not yet officially
- Linux uses Pango for complex Scripts
  - Version of Pango with Sinhala support is available

Types of Sinhala keyboards



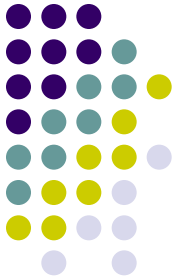
- Wijesekera keyboard
  - used with both typewriters and computers
- “Phonetic” keyboards
  - key assignment is based on the English key layout.
- Transliteration schemes
  - text is typed as a sequence of English letters.
- Consonant-vowel sequence keyboards
  - consonant typed first, followed by vowel modifier

Standard Keyboard



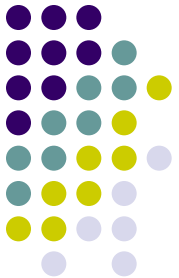
- Need of a Standard keyboard
  - Manufacturers can produce pre-printed keyboards
  - Students can learn typing
  - Users can move from one computer to another
- Wijesekera keyboard may not be optimum
  - but no other Standard was available
- Decided to use Wijesekera keyboard with some modifications

Keyboard Design Principles



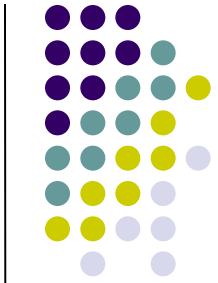
- Common letters in same places as typewriter keyboard.
- Number keys same as in US-ASCII keyboard.
- Only one form of the *al-lakuna* and each *pilla*
- No “half letters” on the keyboard
  - *Bandi akuru* constructed by pressing a *join* key between the two consonants.
- Use same sequence in typing as in writing;
  - e.g., ෙ + ක + ා + ් = කෝ; ක + ්‍ර + ි = ක්‍රි.

Other Keyboards



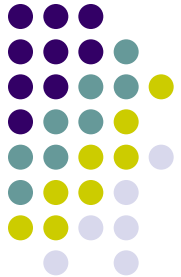
- Standardise a “phonetic” input method
- Design an “optimised” Sinhala keyboard

Conclusion



- Our Objective:
  - using computers, and other devices such as mobile phones, in Sinhala should be as convenient and obvious as in English
- Our work in standardisation of encoding and keyboards is only a first step
- Next step is to disseminate use of Sinhala throughout the country
- A parallel effort in Tamil

Acknowlegements



members of

- the CINTEC Sinhala Fonts Committee
- the ICTA Language Requirements Working Group
- the SLSI Sinhala Working Group
- Font Developers, Lake House, Microsoft
- all others who provided assistance