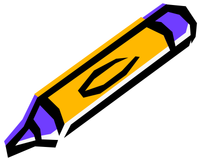
What is UNICODE?

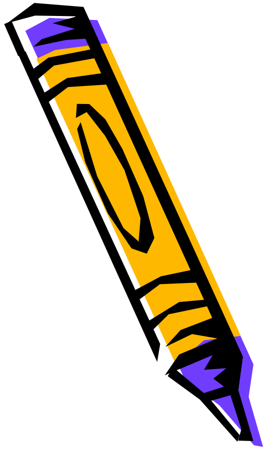
Ruvan Weerasinghe

University of Colombo School of Computing

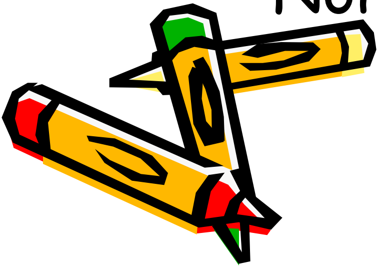


ICTA Local Language Working Group Workshop – March 2004

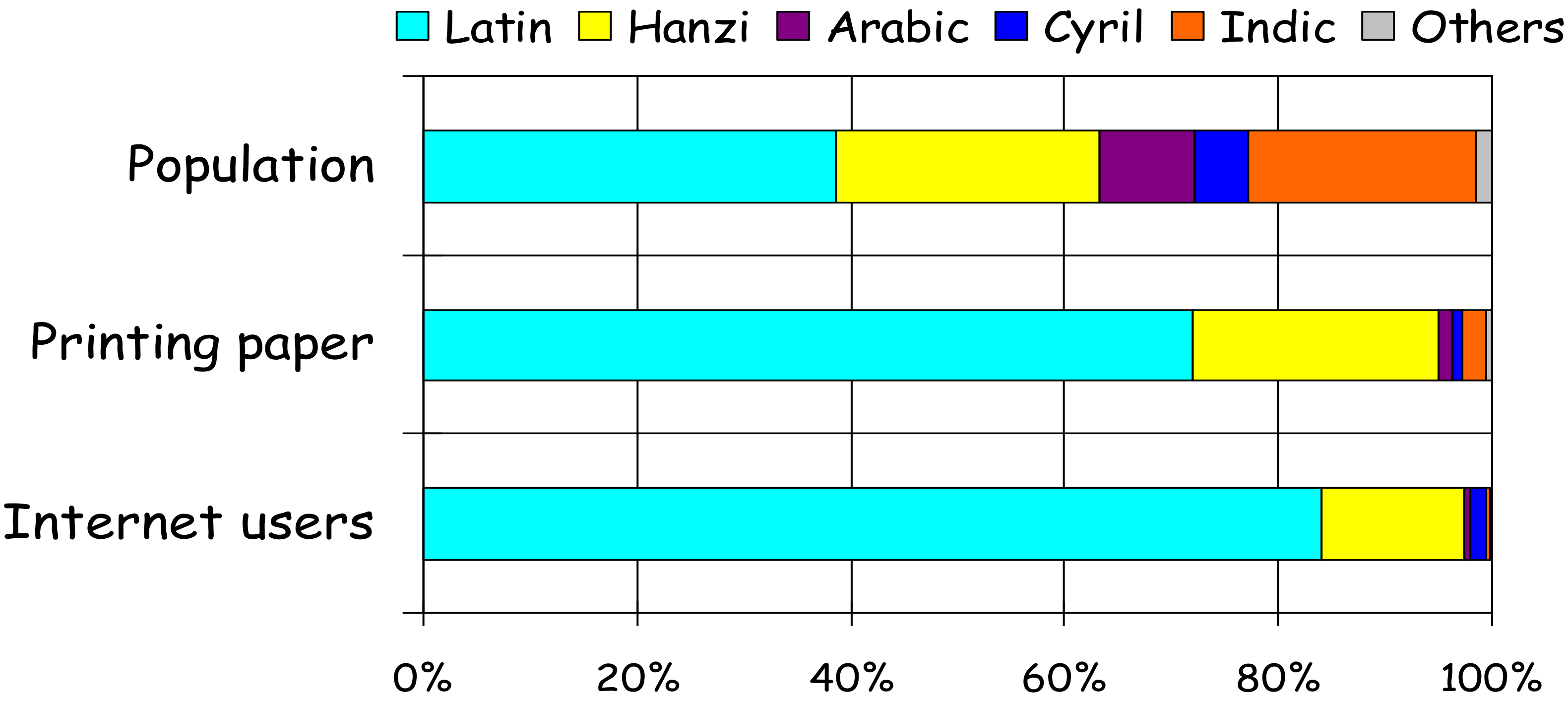
Why UNICODE?



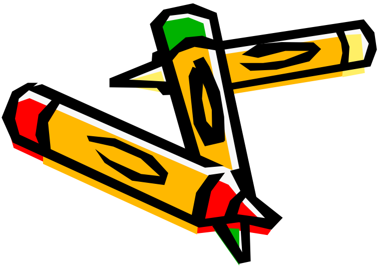
- Because we need Internationalization!
  - Western scripts are dominating the e-world
- Because we need Multilingualization!!
  - Not to be able to have Sinhala only
  - But to have Sinhala with Tamil, English etc.
- Because it is the best supported standard
  - UNICODE supports all the above needs
  - It is the single widely supported framework for Non-Latin support (e.g. Java, ORACLE, XML…)



Global Digital Divide 1999
- by script grouping -



Source: ITU, UNESCO

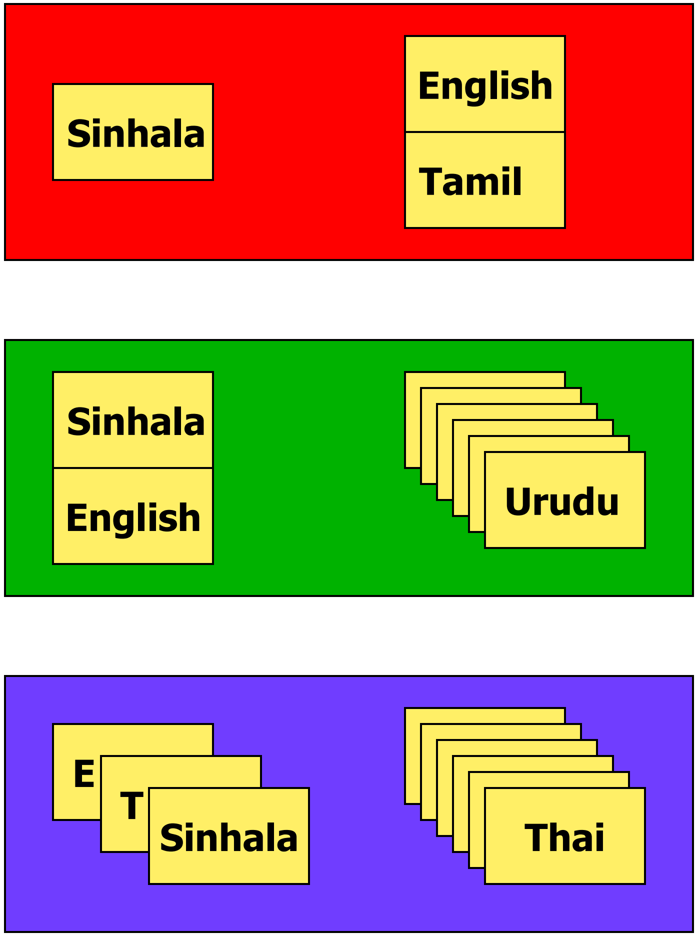


L10N, I18N and M17N

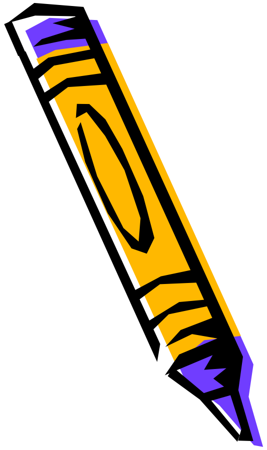
Localization

Internationalization

Multilingualization



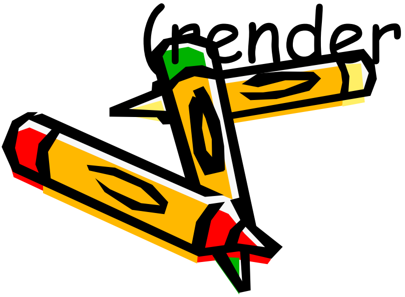
What UNICODE is NOT



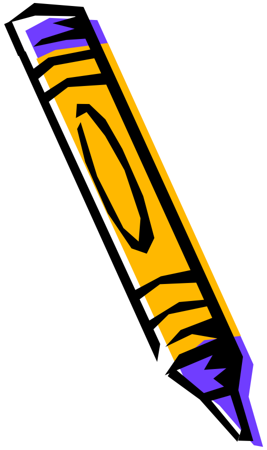
- It is NOT another *font.*
- It is NOT another *keyboard*

……

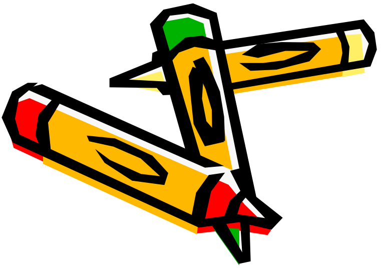
- It only defines a unique *internal* representation of characters
  - e.g. LATIN-CHARACTER-UPPERCASE-A (is at u+0041), SINHALA-LETTER-AYANNA (is at u+0D85)
- It makes no assumptions about how characters are input nor how characters will be displayed (rendered on screen or printer)



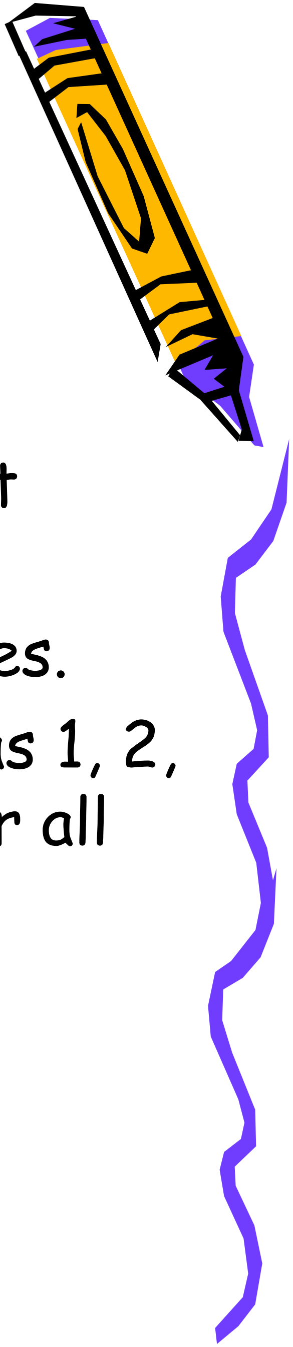
So what IS UNICODE?



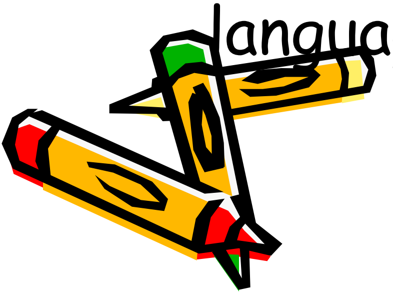
- The 3 aspects of Language Support:
  - Input method
  - Storage (Representation) scheme
  - Output (Rendering) format
- UNICODE primarily concerns the *internal representation* mechanism:
  - Unique codes for the essential characters.
  - Composite characters stored as base character followed by modifier(s)



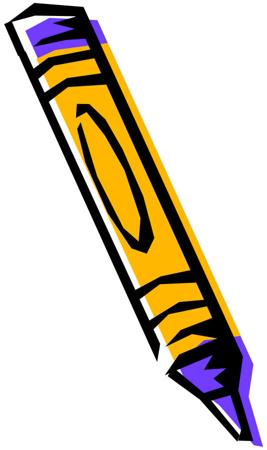
So what IS UNICODE?



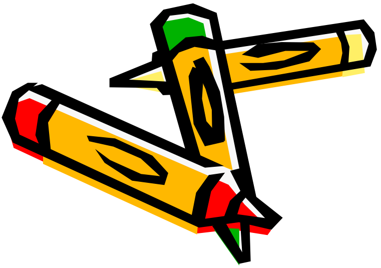
- A 'UNIversal CODE' designed to:
  - Provide adequate space for each (even the most complex) language.
  - Avoid the use of special character/control codes.
  - No duplicate characters (e.g. characters such as 1, 2, 3, >, ? + etc. are in ONE single common place for all languages)
  - Supports multiple languages simultaneously.
  - Implementations do not force users to load all languages of the world!



So what IS UNICODE?



- Main features:
  - Is based on a 20-bit pattern (can represent > 1m 'code points'): 8 bits could store 128/256
  - Provides 8-bit, 16-bit and 32-bit representations for backward compatibility
  - UTF-8 equivalent to ASCII
  - Most Common form is UTF-16



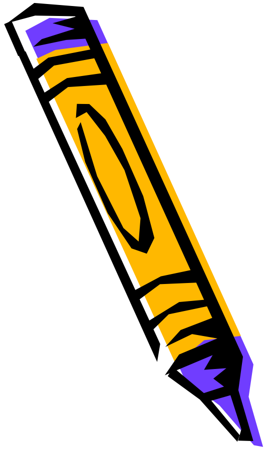
So what IS UNICODE?

- Main features (contd.):
  - It allows pre-composed and composite characters
  - It uses single and multi-word codes
  - It always stores the 'base' first, followed by 'modifier(s)'

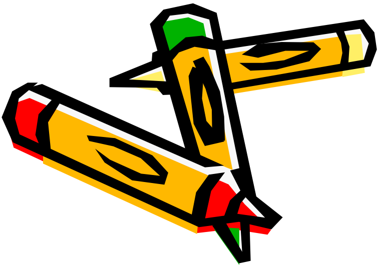
So what IS UNICODE?

- References:
  - See Sinhala code page for UNICODE
    - http://www.unicode.org/charts/PDF/U0D80.pdf
  - And Chapter 9 (South Asian Scripts) of the UNICODE standard
    - http://www.unicode.org/versions/Unicode4.0.0/ch09.pdf
  - Important FAQ on Indic scripts (ongoing issues)
    - http://www.unicode.org/faq/indic.html

How does it all happen?



- Enabling environment
  - Open Type Table (Adobe)
    - An extension of TTF
    - Allows rules in addition to glyphs
  - Rendering/shaping engine to interpret rules
    - Uniscribe in Microsoft OS's
    - Pango, ICU etc. in Linux



How does it all happen?

- Completely transparent to the User:
  - Still types kombuva (ෙ◌), kayanna (ක) and aela-pilla (◌ා) to get කො
  - But can be assured that it will remain කො in any other system and be stored as කො
- The need for a standard input scheme
  - Not crucial for UNICODE
  - But it is important for training, government
  - Wijesekera standard keyboard based (but can use also romanized, phonetic keyboards,…)

So what now?

- Converters for converting legacy (proprietary) encoded texts to UNICODE
- Menus and icons (UI) and help files in Sinhala, Tamil.
- Think about dictionaries and spell checking
- Work on grammar and translation
- What about TTS, OCR and ASR