සිංහල (Sinhala) Orthography: Ola Leaf to the Computer

හර්ෂුල ජයසූරිය (Harshula Jayasuriya)
<harshula at gmail dot com>

Language Technology Research Laboratory
University of Colombo School of Computing

Content

- Three Orthographies
- Encoding
- Input
- Output

Three Orthographies

- Modern – separate letters
  - No strict ligation of consonant clusters. e.g. ද්ධ, ද්ව
- Conjuncts – combined letters
  - Ligation of consonant clusters. e.g. ක්ෂ, න්ද, ක්‍ර, ක්‍ය, කී
  - Some conjuncts still used in modern writing.
- Pali - touching letters
  - Ligation of consonant clusters. e.g. ඤඤ

Encoding

- Phonetic
  - Similar to Indic
  - But need the same phonemes to be represented in 3 different orthographies
- No implicit conjunct ligatures formed for consonant clusters
  - Conjunct ligatures for consonant clusters must be formed explicitly with the use of Zero Width Joiner (ZWJ)

Encoding (2)

- Note:
  - cons = consonant + inherent vowel
  - al = al-lakuna = remove inherent vowel
- Modern – separate letters
  - cons + al + cons
- Conjuncts – combined letters
  - cons + al + ZWJ + cons
- Pali – touching letters
  - cons + ZWJ + al + cons

Input

- Input Method Technologies
- Wijesekera - Layout from the Sinhala typewriter
- Transliteration
- Phonetic

Input Method Technologies

- XKB (X Windows)
- XIM (X Windows)
- GTK/QT IM
- SCIM/m17n

Wijesekera

- Need surrounding text support or buffering
  - Syllable segmentation rules to detect the start of a syllable, required for keeping the buffer small
    - All independent vowels (U+0d85 - U+0d96)
    - Kombuva (U+0dd9) - except if preceded by a kombuva.
    - All consonants (U+0d9a - U+0dc6) - except if preceded by kombuva or kombuva deka (U+0ddb)
    - Kunddaliya (U+0df4)
    - All non-Sinhala characters/codepoints - except ZWJ (U+200D)

Wijesekera (2)

- Normalisation: composing and decomposing codepoints
- Reordering

Transliteration

- Need surrounding text support or buffering
- Normalisation: composing and decomposing codepoints

Phonetic

- Generally one-to-one mapping
- 3 cases of one-to-many mapping

Output

- Renderer
  - Needs to be aware of ZWJ and pass it to the GSUB stage
  - Ensure pre-base dependent vowels precede consonant clusters
- Fonts
  - repaya + consonant + dependent vowel -> repaya + doubled-touching-consonant + dependent vowel. e.g. සිසි