An Algorithm to Use Wijesekara Keyboard for Sinhala Unicode Input

Anuradha Ratnaweera
Lanka Linux User Group[*] *and Virtusa Corporation*[†]

October 12, 2003

This is *not* original work. Similar concepts are found in abundance in many input modules. This text is an attempt to formalize the algorithm we will implement[1] for Sinhala keyboard input. The terminology and definitions are also specific to this text.

1 Glyph, Groups and Characters

Glyphs are visual representation of characters. They can be interpreted as visual units of text.

We define three types of glyphs: *simple glyphs*, *compound glyphs* and *modifiers*. A simple glyph or a modifier corresponds exactly to one keystroke, but only the former can be stand-alone. A compound glyph is a combination of one simple glyph and one or more modifiers. Examples for simple glyphs are අ, ක and ◌. ඒ and කු are compound glyphs.

ආ and කෝ are *not* compound glyphs, but *groups* or glyphs. The former is made up of two simple glyphs අ and ා, whereas the latter is made up of two simple glyphs ක and ෝ, and compound glyph ෝ.

[*] `www.lug.lk` and `www.linux.lk`

[†] `www.virtusa.com`

[1] Hopefully, this will be converted to past tence soon ;-)

Glyph group	Character group
ඇ	U+0D86
ඇෘ	U+0D87
ඇෲ	U+0D88
ඇ	U+0D85

Table 1: Equivalence maps for ඇ

A *visual text* is a sequence of glyph groups. The internal representation of a visual text is by means of a sequence of *characters*. One or more sequencial characters is called a *character group*.

Here are the grammer definitions for compound glyph, glyph and latter group. Simple glyphs and modifiers are basic (terminal) constructs.

compound_glyph :
  simple_glyph *modifier*
  compound_glyph *modifier*

glyph :
  simple_glyph
  compound_glyph

glyph_group :
  glyph
  glyph_group *glyph*

2 Equivalence Maps

Glyph groups are mapped to character groups. Maps are stored in such a way that all the maps related to a given glyph or a given character can be retrived quickly.

Let's start with a simple example. Table 1 lists the equivalence maps for the glyph ඇ. The most important point to notice is that the glyph group with a single ඇ is listed last.

Glyph group	Character group
ඇ	U+0D86
කෝ	U+0D9A U+ODDD
ඛෝ	U+0D9B U+0DDD
ගෝ	U+0D9C U+ODDD
...	...
කො	U+0D9C U+ODDD
ඛො	U+0D9C U+0DDD
ගො	U+0D9C U+ODDD
...	...
කා	U+0D9A U+ODCF
ඛා	U+0D9B U+0DCF
ගා	U+0D9C U+ODCF
...	...
ඉ	U+0D85

Table 2: Equivalence maps for ඉ

Table 2 lists the equivalence maps for the glyph ඉ. Again, notice that the longer groups are liste earlier than others. Also, one map, the first one, is common to both ඇ and ඉ. The list is ratehr long, but later, we will simply this list by introducing *character classes* and *glyph classes*.

The ordering of the set of equivalence maps for a given glyph is known as *partial ordering* [1]. If a glyph group is a superset of another glyph[2], the former should come before the latter. Howerver, independent glyph groups can appear in any order.

Table 3 lists the equvalence maps for the glyph ක .

3 Simple Example

Before we continue furthur, let's look at an example. Table 4 lists the status of the characters and the glyphs after each key press.

[2]This definition is vague and technically incorrect. But the intended meaning should be clear.

Glyph group	Character group
කෝ	U+0D9A U+ODDD
කො	U+0D9C U+ODDC
කෙ	U+0D9A U+ODDB
කෞ	U+0D9A U+ODDE
කේ	U+0D9A U+ODDA
කෙ	U+0D9A U+ODD9
කෲ	U+0D9A U+ODF2
කෘ	U+0D9A U+ODD9
කූ	U+0D9A U+0DD6
කු	U+0D9A U+0DD4
කී	U+0D9A U+0DD3
කි	U+0D9A U+0DD2
කෑ	U+0D9A U+0DD1
කැ	U+0D9A U+0DD0
කා	U+0D9A U+ODCF
ක	U+0D9A

Table 3: Equivalence maps for ක

Step	Glyphs	Characters
1	ක	U+0D9A
2	කා	U+0D9A U0DCF
3	කා	U+0D9A U0DCF
4	කෑ	U+0D9A U0D86

Table 4: Simple example

- Step 1: User presses ක . The only valid map is ක.
- Step 2: User presses ා. Looking up through the list of maps for ා gives කා as the *first* match.
- Step 3: User presses left arrow key. The cursor moves between ක and ා. FIXME 1: needs to add the cursor to the table. FIXME 2: This is not how Pango behaves.
- Step 4: User presses ෑ . A lookup for the relevent glyph brings up ෑ as the first match. However, this also result in another lookup for the glyph ක , because ා is now taken away from it, which brings the last one, ක itself.

4 Classes

5 Cursor Positioning

References

[1] Donald Knuth, The Art of Computer Programming, Volume 3 (Sorting and Searching)