Tamil in Unicode

Helmut Steeb, dated: 2005-01-14.

Inhaltsverzeichnis

Introduction

As described in the Unicode Design Priniciples in [U3] page 13ff., Unicode defines „Characters, Not Glyphs“, the number of glyphs needed to display a given script may be significantly larger than the number of characters encoded the basic units of that script. The process of mapping from characters in the memory representation to glyphs involves complex rules. The memory representation contains text in a logical order, roughly corresponding to the order in which text is typed in via the keyboard. Characters such as the short i in Devanagari are displayed before the characters that they logically follow in the memory representation.

Therefore the conversion of a „logical“ in memory representation of a Unicode encoded text requires information from fonts and from software library algorithms. This information may be missing in „older“ fonts and software libraries.

Tamil in Unicode

[U3] chapter 9.1 „Devanagari“ (p. 212ff.) describes the principles for rendering a Unicode encoded text. The introductions to the other Indic scripts are based on the Devanagari section and just highlight any differences from Devanagari. Some of the Indic scripts have single dependent vowels that are rendered as two or more glyph components. Those glyph components may surround a consonant letter both to the left and right.

[U3] ch. 9.6 „Tamil“ (p. 228ff.): in the process of rendering a Unicode encoded Tamil text, first vowel reordering and splitting applies, then ligatures are formed. These are depicted in ch. 9.6.

[U3] ch. 14 „Code Charts“ (p. 415ff.) contains the table of nominal forms of the Tamil Unicode characters, which is also available in the web [T]. A font for rendering Tamil must contain many additional glyphs.

[T2U] 4.2 Note b mentions that the shaping of the ligatures is handled in the font through substitution tables. This means that a font may be inappropriate for rendering Tamil even if it contains enough ofthe glyphs.

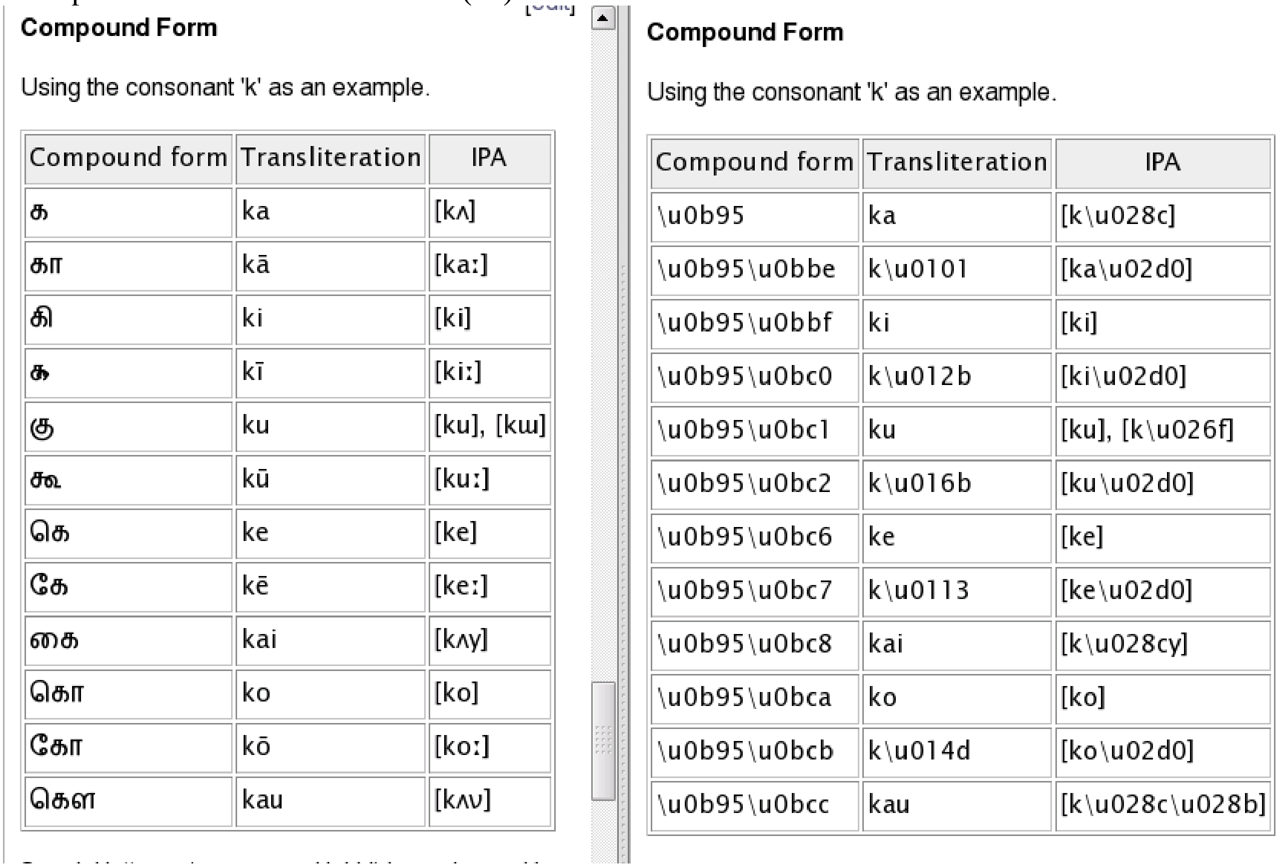
Investigation using a Unicode encoded web page

Test files

#	***File***	***Description***
1	wikipedia_Tamil_script.part.htm	Extract from [W]
2	wikipedia_Tamil_script.part.ucn.htm	created by SuSE Linux 10.0 command line: recode utf8..java <wikipedia_Tamil_script.part.htm > wikipedia_Tamil_script.part.ucn.htm
3	wikipedia-ucn-2column.gif	Screen shot for #1 and #2
4	wikipedia_Tamil_script.part.fonts.htm	Based on #1, duplicated columns
5	wikipedia_Tamil_script.part.fonts.gif	Screen shot for #4

Base functionality

The following screen shot (#3) displays to the left the text from [W] (shortened in #1, displayed using Konqueror under SuSE Linux 10.0), to the left for comparison the same text with the Unicode code points of the Tamil characters (#2).



Compound Form

Using the consonant 'k' as an example.

Compound form	Transliteration	IPA
க	ka	[kʌ]
கா	kā	[kaː]
கி	ki	[ki]
கீ	kī	[kiː]
கு	ku	[ku], [kɯ]
கூ	kū	[kuː]
கெ	ke	[ke]
கே	kē	[keː]
கை	kai	[kʌy]
கொ	ko	[ko]
கோ	kō	[koː]
கௌ	kau	[kʌʋ]

Compound Form

Using the consonant 'k' as an example.

Compound form	Transliteration	IPA
\u0b95	ka	[k\u028c]
\u0b95\u0bbe	k\u0101	[ka\u02d0]
\u0b95\u0bbf	ki	[ki]
\u0b95\u0bc0	k\u012b	[ki\u02d0]
\u0b95\u0bc1	ku	[ku], [k\u026f]
\u0b95\u0bc2	k\u016b	[ku\u02d0]
\u0b95\u0bc6	ke	[ke]
\u0b95\u0bc7	k\u0113	[ke\u02d0]
\u0b95\u0bc8	kai	[k\u028cy]
\u0b95\u0bca	ko	[ko]
\u0b95\u0bcb	k\u014d	[ko\u02d0]
\u0b95\u0bcc	kau	[k\u028c\u028b]

This example shows that

- some sequences of Unicode characters are rendered by a single glyph, some others by multiple glyphs;
- \u0bbe (TAMIL VOWEL SIGN AA) is rendered correctly;
- some vowels form a ligature: \u0bbf (TAMIL VOWEL SIGN I), \u0bc0 (TAMIL VOWEL SIGN II), \u0bc1 (TAMIL VOWEL SIGN U), \u0bc2 (TAMIL VOWEL SIGN UU);
- some vowels that follow a consonant in memory are rendered to the left of the consonant: \u0bc6 (TAMIL VOWEL SIGN E), \u0bc7 (TAMIL VOWEL SIGN EE), \u0bc8 (TAMIL VOWEL SIGN AI);
- some vowels that follow a consonant in memory are rendered both left and right of the consonant: \u0bca (TAMIL VOWEL SIGN O), \u0bcb (TAMIL VOWEL SIGN OO), \u0bcc (TAMIL VOWEL SIGN AU).

Conditions for correct rendering

Whether Tamil Unicode encoded characters are rendered correctly depends at least on the operating system and the fonts involved. The following example screen shot (=#5) shows file #4 in OpenOffice.org 1.1.3 with different fonts (captured with OpenOffice.org 1.1.3 under SuSE Linux 10.0, File|Open, then assign different fonts to the columns):

Thorndale AMT	**Latha**	**Arial**	***GIST-TMOTAbhirami***	**Transliteration**
க	க	க	க	ka
கா	கா	கா	கா	kā
கி	கி	கி	கி	ki
கீ	கீ	கீ	கீ	kī
கு	கு	கு	கு	ku
கூ	கூ	கூ	கூ	kū
ெ◌க	கெ	ெ◌க	கெ	ke
ே◌க	கே	ே◌க	கே	kē
ை◌க	கை	ை◌க	கை	kai
ெ◌ாகொ	கொ	ெ◌ாகொ	கொ	ko
ே◌ாகோ	கோ	ே◌ாகோ	கோ	kō
ெ◌ளகௌ	கௌ	ெ◌ளகௌ	கௌ	kau

This illustrates that vowels are rendered correctly with fonts Latha and GIST-TMOTAbhirami [ILDC], but not with fonts „Thorndale AMT“ and „Arial“. With the latter, e.g.

- for /ke/, the vowel ெ◌that follows the consonant க in memory is actually rendered to the left of the consonant angezeigt, but with an additional place-holder symbol for the consonant;
- for /ko/, both the left glyph and the right glyph of the two-part vowel are first rendered with the place-holder symbol, then again surrounding the consonant.

Additional Notes

Unicode fonts for Tamil

Using the font „TMOTABBI_Ship.ttf“ from [ILDC], the line from ta.openoffice.org

```
OpenOffice.org - \u0b87\u0ba4\u0bc1 \u0b92\u0bb0\u0bc1
\u0bb5\u0bbf\u0b9f\u0bc1\u0ba4\u0bb2\u0bc8
\u0bae\u0ba9\u0baa\u0bcd\u0baa\u0bbe\u0b99\u0bcd\u0b95\u0bcb\u0b9f\u0bc1
\u0b89\u0bb0\u0bc1\u0bb5\u0bbe\u0ba9
```

is rendered as:

OpenOffice.org - இது ஒரு விடுதலை மனப்பாங்கோடு உருவான

Latha

Using font Latha from [L], the line from ta.openoffice.org is rendered as:

OpenOffice.org – இது ஒரு விடுதலை மனப்பாங்கோடு உருவான

Rendering problems in Windows

See e.g. [TN]:

> Windows XP comes with a Unicode Tamil Font (Latha) and you should **not** need to download/install a unicode font.

1. You need to have Unicode Tamil fonts installed on your computer and the Operating System capable of rendering Tamil Scripts. Windows XP comes with a Unicode Tamil Font (Latha) and you should **not** need to download/install a unicode font.

2. In the Control Panel, in Regional/Languages Options you will need to ensure that Indic/Asian Language option is checked.

3. Use a browser that is capable of handling UTF-8 based pages (Netscape 6, Internet Explorer 5) with the Unicode Tamil font

...

Some users of Windows 98/NT/ME/2000 Windows NT may find that even with a Tamil Unicode font installed and the browser correctly set to View > Encoding > Unicode, some of the Tamil letters are incorrectly displayed.

...

This may be due to the earlier version of usp10.dll used in these operating systems. To correct this, it may be necessary to update the usp10.dll file in Windows.

Text input

See [IN]:

Ekalappai 1.0 - a Bundled Software for Tamil Entry in PC

Ekalappai 1.0 is one of the new pieces of software that has revolutionized Tamil computing this year. Mugunthraj and his Erumbugal friends introduced it on January 16, 2002.

The default keyboard layout is anjal layout. Most of us are familiar with the anjal keyboard layout. It is phonetic. Anyone can learn to input in Tamil using anjal layout in 5 minutes. I understand that you can install Tamilnet99, Mylai, and Tamil typewriter layouts as well with this program.

Tamil bible in Unicode

[TB] offers a website that allows query (into an HTML page) of single bible chapters in „encoding“ either „ASCII“ or „Unicode“.

[BIB] offers online bibles in many languages. For Tamil:

> These Bibles are free, either to read online or for download:

- The Tamil Bible (http://www.wbtc.com/articles/downloads/default.htm) from the World Bible Translation Center.
- Tamil Bible (http://tamilbible.net/index.html) in audio and text versions
- Tamil Bible and search engine (http://members.tripod.com/~ksiva/Bible/)
- Tamilchristian.com (http://www.tamilchristian.com/) offers gospel resources, including an online Bible.

References

[ILDC] http://www.ildc.in/GIST/htm/otfonts.htm, e.g. Fonts.tar.gz, font „TMOTABBI_Ship.ttf“ (=„GIST-TMOTAbhirami“)

[L] http://www.tamilnation.org/fonts/latha.ttf

[BIB] http://www.ethnicharvest.org/bibles/

[IN] www.infitt.org/paper9.html

[T] http://www.unicode.org/charts/PDF/U0B80.pdf

[T2U] http://www.unicode.org/notes/tn15/Tscii2Unicode.pdf

[TB] http://www.tamil-bible.com/tabletype.php?Type=Eng

[TN] http://www.tamilnation.org/digital/Tamil%20Fonts%20&%20Software.htm

[U3] The Unicode Standard Version 3.0, The Unicode Consortium, Addison-Wesley 2000.

[W] http://en.wikipedia.org/wiki/Tamil_script