

Unicode and its importance to Tamil usage in computers

An article by *Thayumanasamy Somasundaram*

[<http://tamil.somasundaram.us/ta-unicode.html>]

January 14, 2009

What is Unicode?

I will use the definition given by the [Unicode Consortium](#) itself to answer the question:

“Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language.” [From [Unicode Consortium](#)’s webpage]

யூனிக்கோடு என்றால் என்ன?

“யூனிக்கோடு எந்த இயங்குதளம் ஆயினும், எந்த நிரல் ஆயினும், எந்த மொழி ஆயினும் ஒவ்வொரு எழுத்துக்கும் தனித்துவமான எண் ஒன்றை வழங்குகிறது”.
[[யூனிக்கோடு ஒன்றியம் இணையத்திலிருந்து](#)]

Encoding Standards (ASCII)

As we all know computers store, handle, and transmit information using **binary digits** (bits). This idea is famously expressed by the saying that computers store information in “0s” and “1s” or “On” and “Off” or “Up” and “Down” method. So, for example, if you want to type, store, and transmit the English letter “Capital A”, a computer needs to do this in a particular way so that when it is read back, sent off, printed out, by another computer running in a different platform (*think*, PC, Mac, iPhone, web server, color printer, etc), running a different operating system (*think*, Windows, Macintosh, Linux/UNIX, Palm O/S, PostScript, etc), and using a different application (*think*, web browser, word processor, e-mail, pdf-viewer, etc) that system can reproduce exactly the same letter. So computer experts devised a standard that will fulfill the above requirements.

In early 1960s when English language dominated the computing world it was decided that [ASCII](#) (**American Standard Code for Information Interchange**) a 7-bit character encoding system will be used as a standard encoding for storing and exchanging information in and between computers systems (**EBCDIC** an 8-bit character encoding and several other standards existed before and during that time and they are beyond the scope of this article). With 128 possible values ($2^7=128$; the number 2 comes from the binary and the power 7 comes since there 7 positions) 31 control (non-printable) characters, A-Z, a-z, 0-9, and several punctuation marks were assigned from 0 to 127. So the “Capital letter A” was assigned a decimal value of 65 or binary value of 100 0001 or a hexadecimal value of 41 (hexadecimal has values from 0 to 9 and A through F; a total of 16). All

computers then were able to reinterpret this value back to "Capital letter A". Soon it became necessary to add European languages like German, French, and Spanish with their accent marks (Á), umlauts (Ü), graves (Ô), tildes (Ñ), British Pound, Japanese Yen, and Copyright sign. So the computer experts designed an **Extended ASCII** system with its full 8-bit character encoding of a byte ($2^8=256$) thus extending the value to 256 positions (0 to 255; retaining the original 0-127 values). After all, most of the computers at that time read data in bytes which contained 8 bits, why not use the last bit as well!

However, when computers became more common place in 1980s, non-US and non-European countries decided to use the extended set to include the characters of their own languages. This meant that characters assigned for positions 128 to 255 (past the first 0 to 127 values) could indicate or encode different characters depending upon what kind of encoding the input-user used and *may not* be reinterpreted correctly by the output-user, if they *didn't know* the original encoding scheme. To complicate the matter, if the information being exchanged had more than one (non-Latin) language it became almost impossible to input both the languages at the same time.

Tamil Encoding Standards (TSCII)

As the Internet started to become more popular in 1990s several non-Latin based languages wanted to post, spread, and share their information/contents online. But they found it difficult to exchange and display their languages on the web browsers. Soon these pioneering people started to devise their own encoding system assigning certain values to certain letters (glyphs) and including the encoding information in special files (and fonts) with documents or a separate downloads. The early developers/web pioneers for Tamil language included Dr. K. Kalyanasundaram (Kalyan) of Switzerland (www.tamilelibrary.org) and Muthu Nedumaran (Muthu) of Malaysia (www.murasu.com) and their colleagues at the Standard for Tamil Computing ([STC](#)) and [Tamil Net](#).

Dr Kalyan started to develop an 8-bit bilingual character encoding scheme for Tamil in late 1990s and named it **Tamil Standard Code for Information Interchange (TSCII)**. As we discussed before, like other non-Latin language pioneers Kalyan used the positions 128 through 255 to encode or map various Tamil characters like vowels (உயிர்; அ), and consonants (மெய்; ழ), combinations (உயிர்மெய்; ழு), and add-ons like (ி,) (ை), and (ா). The TSCII encoding scheme went through several modifications to accommodate or streamline some features and finally settling with version 1.7 (current) and is shown in **Fig 1**. In 2007 TSCII encoding was registered with [Internet Assigned Names Authority](#). Following [Tamil Net 1999 Conference](#), Tamil Nadu Government came up with two more encoding schemes called TAM (Tamil monolingual; Tamil only glyphs for positions 0-255) and TAB (Tamil bilingual; English for 0-127 and Tamil for 128-255) that were different from TSCII. In addition they gave suggestions for Tamil Type Writer key positions. The TAB encoding is shown in **Fig 2** and the Extended ASCII table is shown in **Fig 3** to compare and contrast with the two encoding schemes discussed above.

A possible 8-bit font encoding scheme for Tamil -v. 1.7
(Tamil standard Code for Information Interchange TSCH)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0 (0-15)																
1 (16-31)																
2 (32-47)			'	#	\$	%	&	'	()	*	+	,	-	.	/
3 (48-63)	0	1	2	3	4	5	6	7	8	9	.	,	<	=	>	?
4 (64-79)	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5 (80-95)	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6 (96-111)	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7 (112-127)	p	q	r	s	t	u	v	w	x	y	z	{		}	~	
8 (128-143)	௦	௧	௨	௩	௪	௫	௬	௭	௮	௹	௰	௱	௲	௳	௴	௵
9 (144-159)	௶	௷	௸	௹	௰	௱	௲	௳	௴	௵	௶	௷	௸	௹	௰	௱
A (160-175)		ஊ	஋	஌	஍	எ	ஏ	உ	ஐ	஑	ஒ	ஓ	ஔ	க	஖	஗
B (176-191)	஘	ங	ஐ	஑	ஒ	ஓ	ஔ	க	஖	஗	஘	ங	ஐ	஑	ஒ	ஓ
C (192-207)	ஔ	க	஖	஗	஘	ங	ஐ	஑	ஒ	ஓ	ஔ	க	஖	஗	஘	ங
D (208-223)	ஐ	஑	ஒ	ஓ	ஔ	க	஖	஗	஘	ங	ஐ	஑	ஒ	ஓ	ஔ	க
E (224-239)	஖	஗	஘	ங	ஐ	஑	ஒ	ஓ	ஔ	க	஖	஗	஘	ங	ஐ	஑
F (240-255)	஑	ஒ	ஓ	ஔ	க	஖	஗	஘	ங	ஐ	஑	ஒ	ஓ	ஔ	க	

Fig 1. TSCII 1.7 from Kalyan and others www.tamil.net/tscii/tscii.html

TABxxx - BILINGUAL CODING SCHEME FOR TAMIL														
	32	48	64	80	96	112	128	144	160	176	192	208	224	240
0		0	@	P	'	p				௦	௧		௨	௩
1	!	1	A	Q	a	q				௪	௫		௬	௭
2	"	2	B	R	b	r				௮	௹		௰	௱
3	#	3	C	S	c	s				௲	௳		௴	௵
4	\$	4	D	T	d	t				௷	௸		௹	௰
5	%	5	E	U	e	u		*	*	௳	௴		௵	௶
6	&	6	F	V	f	v				௷	௸		௹	௰
7	'	7	G	W	g	w				௹	௰		௱	௲
8	(8	H	X	h	x				௳	௴		௵	௶
9)	9	I	Y	i	y				௷	௸		௹	௰
10	*	:	J	Z	j	z				௹	௰		௱	௲
11	+	:	K	[k	(௳	௴		௵	௶
12	,	<	L	\	l					௷	௸		௹	௰
13	-	=	M]	m)				௹	௰		௱	௲
14	.	>	N	^	n	~				௳	௴		௵	௶
15	/	?	O	_	o	DEL				௷	௸		௹	௰

Fig 2. TAB encoding from TN Gov www.tamilvu.org/Tamilnet99/annex4.htm

Friendly	Numeric	Description	Friendly	Numeric	Description	Friendly	Numeric	Description
‘	'	left single quote	R	R	uppercase letter	·	·	middle dot
’	'	right single quote	S	S	uppercase letter	¸	ˆ	cedilla
‚	‚	single low-9 quote	T	T	uppercase letter	¹	ˆ	superscript one
“	“	left double quote	U	U	uppercase letter	º	♂	masculine
”	”	right double quote	V	V	uppercase letter	»	∟	right angle
„	”	double low-9 quote	W	W	uppercase letter	¼	¼	one-fourth
†	†	dagger	X	X	uppercase letter	½	½	one-half
‡	‡	double dagger	Y	Y	uppercase letter	¾	¾	three-fourths
‰	‰	per mill sign	Z	Z	uppercase letter	?	?	inverted
‹	‹	left angle quote	[[left square bracket	À	À	A, grave
›	›	right angle quote	\	\	backslash	&Acute;	Á	A, acute accent
♠	♠	black spade suit]]	right square bracket	Â	Â	A, circumflex
♣	♣	black club suit	^	^	caret	Ã	Ã	A, tilde
♥	♥	black heart suit	_	_	underscore	Ä	Ä	A, umlaut
&diamonds;	♦	black diamond suit	`	˘	grave accent	Å	Å	A, ring
‾	—	overline	a	a	lowercase letter	&AE;	Æ	AE
←	←	leftward arrow	b	b	lowercase letter	Ç	Ç	C, cedilla
↑	↑	upward arrow	c	c	lowercase letter	È	È	E, grave accent
→	→	rightward arrow	d	d	lowercase letter	É	É	E, acute accent
↓	↓	downward arrow	e	e	lowercase letter	Ê	Ê	E, circumflex
™	™	trademark sign	f	f	lowercase letter	Ë	Ë	E, umlaut
			horizontal tab	g	g	lowercase letter	ì	Ì	I, grave accent
	
	line feed	h	h	lowercase letter	í	Í	I, acute accent
	 	space	i	i	lowercase letter	î	Î	I, circumflex
	!	! exclamation mark	j	j	lowercase letter	ï	Ï	I, umlaut
"	“	double quotation	k	k	lowercase letter	Ð	Ð	Eth, Icelandic
	#	# number sign	l	l	lowercase letter	Ñ	Ñ	N, tilde
	$	\$ dollar sign	m	m	lowercase letter	Ò	Ò	O, grave
	%	% percent sign	n	n	lowercase letter	Ó	Ó	O, acute accent
&	&	ampersand	o	o	lowercase letter	Ô	Ô	O, circumflex
	'	' apostrophe	p	p	lowercase letter	Õ	Õ	O, tilde
	((left parenthesis	q	q	lowercase letter	Ö	Ö	O, umlaut
)) right parenthesis	r	r	lowercase letter	×	×	multiplication
	*	* asterisk	s	s	lowercase letter	Ø	Ø	O, slash
	+	+ plus sign	t	t	lowercase letter	Ù	Ù	U, grave accent
	,	, comma	u	u	lowercase letter	Ú	Ú	U, acute accent
	-	- hyphen	v	v	lowercase letter	Û	Û	U, circumflex
	.	. period	w	w	lowercase letter	Ü	Ü	U, umlaut
&frac1;	/	slash	x	x	lowercase letter	Ý	Ý	Y, acute accent
	0	0	y	y	lowercase letter	Þ	Þ	THORN
	1	1	z	z	lowercase letter	ß	ß	sharps
	2	2	{	{	left curly brace	à	à	a, grave accent
	3	3	|		vertical bar	á	á	a, acute accent
	4	4	}	}	right curly brace	â	â	a, circumflex
	5	5	~	~	tilde	ã	ã	a, tilde
	6	6	–	–	en dash	ä	ä	a, umlaut
	7	7	—	—	em dash	å	å	a, ring
	8	8	 		nonbreaking space	æ	æ	ae
	9	9	¡	!	inverted exclamation	ç	ç	c, cedilla
	:	: colon	¢	¢	cent sign	è	è	e, grave accent
	;	; semicolon	£	£	pound sterling	é	é	e, acute accent
<	<	less-than sign	¤t;	¤	general currency sign	ê	ê	e, circumflex
	=	= equals sign	¥	¥	yen sign	ë	ë	e, umlaut
>	>	greater-than sign	¦		broken vertical bar	ì	ì	i, grave accent
	?	? question mark	&brkbar;		broken vertical bar	í	í	i, acute accent
	@	@ at sign	§	§	section sign	î	î	i, circumflex
	A	A uppercase letter	¨	¨	umlaut	ï	ï	i, umlaut
	B	B uppercase letter	¨	¨	umlaut	ð	ð	eth, Icelandic
	C	C uppercase letter	©	©	copyright	ñ	ñ	n, tilde
	D	D uppercase letter	ª	ª	feminine ordinal	ò	ò	o, grave accent
	E	E uppercase letter	«	«	left angle quote	ó	ó	o, acute accent
	F	F uppercase letter	¬	¬	not sign	ô	ô	o, circumflex
	G	G uppercase letter	&soft;	ˆ	soft hyphen	õ	õ	o, tilde
	H	H uppercase letter	®	®	registered trademark	ö	ö	o, umlaut
	I	I uppercase letter	¯	ˉ	macron accent	÷	÷	division sign
	J	J uppercase letter	&shbar;	ˆ	macron accent	ø	ø	o, slash
	K	K uppercase letter	°	°	degree sign	ù	ù	u, grave accent
	L	L uppercase letter	±	±	plus or minus	ú	ú	u, acute accent
	M	M uppercase letter	²	²	superscript two	û	û	u, circumflex
	N	N uppercase letter	³	³	superscript three	ü	ü	u, umlaut
	O	O uppercase letter	´	´	acute accent	ý	ý	y, acute accent
	P	P uppercase letter	µ	µ	micro sign	þ	þ	thorn, Icelandic
	Q	Q uppercase letter	¶	¶	paragraph sign	ÿ	ÿ	y, umlaut

Source: www.LookupTables.com

Fig 3. Extended ASCII encoding www.lookuptables.com

Propriety Encoding/Fonts

The situation became more complicated when leading Tamil magazines and news papers like Ananda Vikatan, Kalki, Kumdum, Dina Mani, Daily Thani, and Dina Malar decided to encode their own proprietary schemes and fonts. So by early 2000s it became difficult to browse, type, exchange e-mails unless everyone had the same encoding/font scheme. Many readers will recall that without the proper fonts installed in their computers reading Tamil magazines and newspaper articles was difficult and at times the computer would display funny characters instead of Tamil alphabets. The user will then be prompted to download the proprietary font and restart the browser before continuing.

As the number of publications, applications like web browsers, e-mail clients, mobile phones increased the number of fonts and schemes mushroomed. The

[Tamil Nadu Government](#) and the [Virtual Tamil University](#) devised their own font and scheme. **[Indian Script Code for Information Interchange \(ISCII\)](#)** is another encoding supported by [Government of India](#). Where as, TSCII encoded the characters in written order, ISCII encoded it in logical order (similar to Unicode). No one wanted to adopt other's scheme, the web pioneers wanted to unite the Tamils and come up with uniform standard. One such solution is Tamil Unicode. **Please note** that there are still lots of active discussions about the advantages and disadvantages of the current Tamil Unicode Standard 5.1.0 and no by no means the situation is finalized. But that is beyond the scope of this article.

Tamil Unicode

The [Unicode Consortium](#) starting with Unicode version 1.0.0 (Oct 1991; 24 scripts) and to the current version 5.1.0 (Apr 2008; 75 scripts) has included [Tamil](#) as one of the scripts it supports (Note that Unicode refers to *script* rather than *language* since scripts like Latin is used by several languages like, English, Italian, etc). The consortium assigned 128 positions for Tamil ([U-0B80 to U-0BFF](#); in hexadecimal) with positions for vowels (ஈ; U-0B88), consonants with inherent vowels (ஐ; U-0BA4), Tamil numerals (௪; U-0BEA), and special Tamil characters (஠; U-0BFA), and ligatures (஡; U-0BFF) and several positions are left open for future inclusions (actually only 72 positions have been used out of the 128)

The Unicode supports several

Tamil Unicode Writers and Editors

I will give some Tamil Unicode Writers and Editors here. There may be more, but I found the following list to be very useful. In some writers you will be asked to type in literal Tamil and press a button to get the words in Tamil language (For example, type *ammaa* or *ammA* to get அம்மா or *en peyar sOmasuwtharam* to get என் பெயர் சோமசுந்தரம் (note some quirks like w for ற், sO for சோ and so for சொ)

- http://ezilnila.com/tane/unicode_Writer.htm (Online)
- <http://www.suratha.com/reader.htm> (Online)
- <http://yesudas.rs.googlepages.com/tamilunicodewriter> (Online)
- http://yesudas.rs.googlepages.com/WOG_UniPad.zip (Download and use it without an Internet connection | Stand-alone product)
- <http://www.higopi.com/ucedit/Tamil.html> (Online)
- <http://software.nhm.in> (Download software and use it offline | Windows XP/Vista)

Unicode Resources

I will list some Unicode general and Tamil specific resources below. By no means is the list exhaustive, but I found the following to be very helpful.

- Unicode Consortium (www.unicode.org)
- Wikipedia Unicode pages (en.wikipedia.org/wiki/Unicode)
- Tamil Unicode Character Charts (<http://unicode.org/charts/PDF/U0B80.pdf>)
- Alan Wood's Unicode Resources (<http://www.alanwood.net/unicode/tamil.html>)
- Acharya- IIT, Madras site (http://acharya.iitm.ac.in/tamil/tamil_unicode.php)
- Tamil Pad Unicode Editor (<http://www.tamilpad.com/>)
- Tamil Script (http://en.wikipedia.org/wiki/Tamil_script)
- Fileformat Info (<http://www.fileformat.info/info/unicode/char/0b85/index.htm>)
- Indic Script at Unicode (<http://unicode.org/faq/indic.html>)
- Unicode Standard for South Asian Scripts ([Ch 9](#))
- Project Madurai ([Tamil Literature Repository](#))
- Kalyan's Tamil Electronic Library (<http://www.tamilelibrary.org/>)
- Unicode Tamil Font Gallery (http://www.wazu.jp/gallery/Fonts_Tamil.html)
- [SALRC at UChicago](#)
- Suratha's Unicode Web Presentation (<http://www.suratha.com/tamilunicode.html>)
- எழில் நிலா ([How TO Unicode](#))
- Gopi's Tools (<http://www.higopi.com/tools/>)
- Penn State's TLT (<http://tlt.its.psu.edu/suggestions/international/bylanguage/tamil.html>)
- ஒருங்குறி (<http://ta.wikipedia.org/wiki/%E0%AE%92%E0%AE%B0%E0%AF%81%E0%AE%99%E0%AF%8D%E0%AE%95%E0%AF%81%E0%AE%B1%E0%AE%BF>)