Tamil at Unicode – the encoding challenge

ILUGC, 9th December 2006

Ramanraj K <ramanraj.k@gmail.com>

What is Encoding?

To change information
into a form that can be
processed by a computer.

OALD

Representing data with b00ls

- ON-OFF states represent boolean values
  - OFF represents 0 or FALSE
  - ON represents 1 or TRUE
- Each 0 or 1 is a “binary digit” or “bit” of information
- A BYTE (BinarY TablE) is a contiguous sequence of a fixed number of bits which has come to mean 8 bits “octet” capable of holding 256 values from 00000000 to 11111111
- ASCII – American Standard Code for Information Interchange character encoding based on the English Alphabet is the widely used standard
- The 95 printable ASCII characters are:
  !"#$%&'()*+,-./0123456789:;<=>? @ABCDEFGHIJKLMNOPQRSTUVWXYZ [\]^_`abcdefghijklmnopqrstuvwxyz{|}~
- Source code by programmers is converted to machine code which computers understand natively

ASCII Chart

ASCII Code	**Most Significant Bits [MSB]**							
LSB	000	001	010	011	100	101	110	111
0000	NUL,^@	DLE,^P	spc	0	@	P	`	p
0001	SOH,^A	DC1,^Q	!	1	A	Q	a	q
0010	STX,^B	DC2,^R	“	2	B	R	b	r
0011	ETX,^C	DC3,^S	#	3	C	S	c	s
0100	EOT,^D	DC4,^T	$	4	D	T	d	t
0101	ENQ,^E	NAK,^U	%	5	E	U	e	u
0110	ACK,^F	SYN,^V	&	6	F	V	f	v
0111	BEL,^G	ETB,^W	'	7	G	W	g	w
1000	BS,^H	CAN,^X	(	8	H	X	h	x
1001	HT,^I	EM,^Y	)	9	I	Y	i	y
1010	LF,^J	SUB,^Z	*	:	J	Z	j	z
1011	VT,^K	ESC,^[	+	;	K	[	k	{
1100	FF,^L	FS,^\	,	<	L	\	l	|
1101	CR,^M	GS,^[	–	=	M	]	m	}
1110	SO,^N	RS,^^	.	>	N	^	n	~
1111	SI,^O	US,^_	/	?	O	–	o	DEL

Unicode – the emerging encoding standard

Unicode is an industry standard designed to allow text and symbols from all of the writing systems of the world to be consistently represented and manipulated by computers. [Wikipedia]

- Unicode consists of:
  - Character repertoire
  - Encoding methodology
  - Set of standard character encodings
  - Sets of code charts for visual reference,
  - Enumeration of character properties
  - Rules for normalization, decomposition, collation and rendering.
- Unicode features:
  - Range 0000 to 10FFFF
  - 65535 chars in the Basic Multilingual Plane 0 range from U+0000 to U+FFFF
  - Supplementary Planes 1 to 16 for code points U+10000 to U+10FFFF
  - Script based encoding
  - Each character is assigned a unique code point that is an integer value between 0 and 1114112

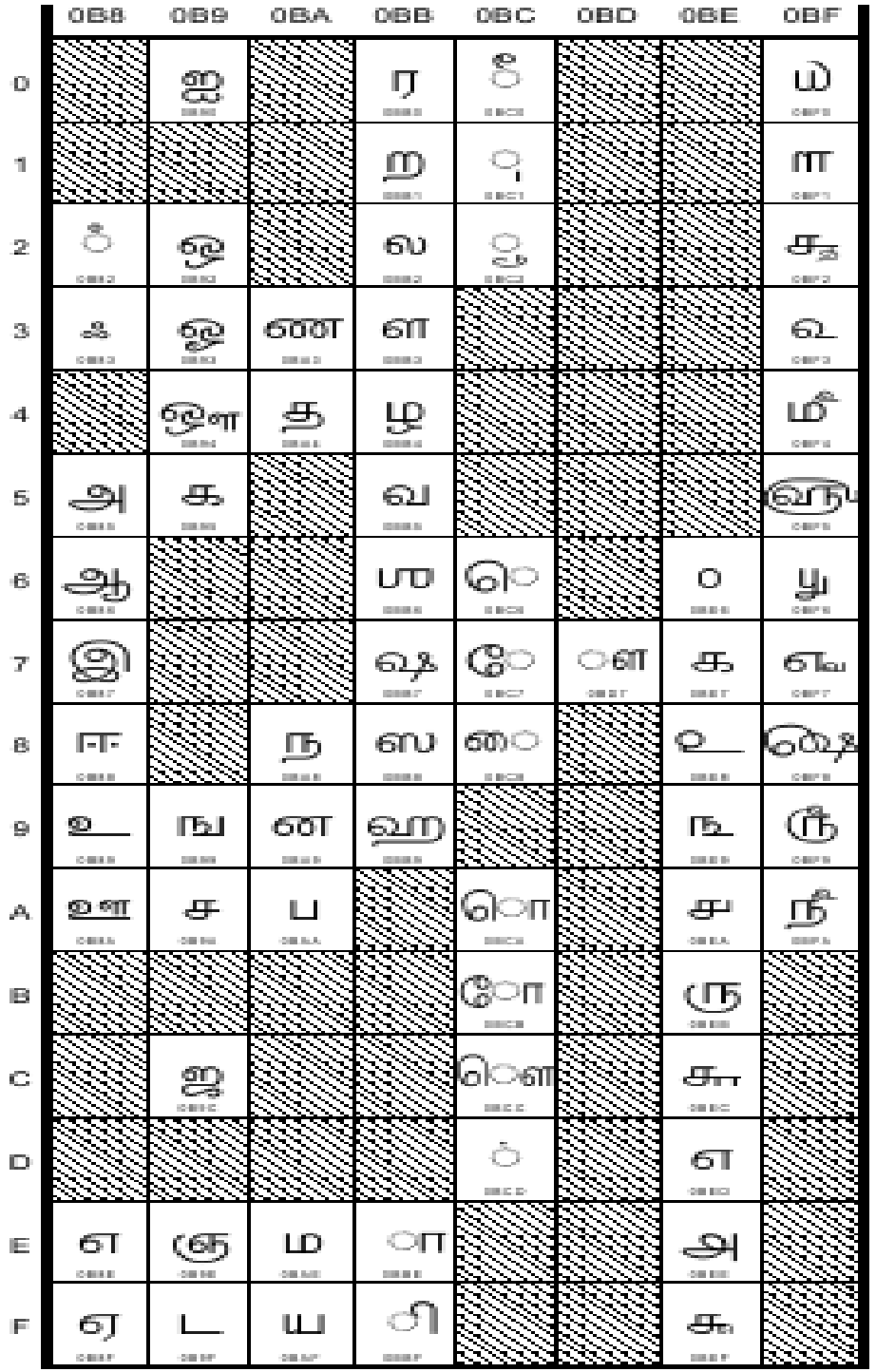
Unicode: Supported Indic Scripts

- # Property: Block
  - 0000..007F; Basic Latin
  - 0900..097F; Devanagari
  - 0980..09FF; Bengali
  - 0A00..0A7F; Gurmukhi
  - 0A80..0AFF; Gujarati
  - 0B00..0B7F; Oriya
  - 0B80..0BFF; Tamil
  - 0C00..0C7F; Telugu
  - 0C80..0CFF; Kannada
  - 0D00..0D7F; Malayalam

All languages in the VIII Schedule of the Constitution of India are supported, except Manipuri (item 9)

தமிழ் யுனிக்கோட்

- B80-BFF Range
  - (128 :: 2944-3071)
- Utilisation:
  - மொத்தம் 71
    - உயிர் 12
    - மெய் 18
    - ஆய்தம் 1
    - இதர குறிகள் 40
  - காலி இடம் 57



Unicode Collation Algorithm :: DUCET

UCA Tamil Collation Chart

source:

- http://developer.mimer.com/charts/UCA_tamil.htm
- http://www.unicode.org/Public/UCA/4.1.0/allkeys.txt

Issues with Unicode Tamil Range B80-BFF

- Sort order not given correctly in DUCET
- Expensive encoding for Tamil chars
  - Three to twelve bytes to store chars
  - Databases could take three times more space
  - Heavy burden on a light language
- Missing chars
  - No code points for உயிர் மெய்
  - Fraction and measurement symbols missing
- Rigid Scheme
  - Brahmi -> Vatteluthu -> ??? Evolution ???

தமிழ் எழுத்துக்கள் – TUNE

(Font : TAUN_Elango_Barathi)

UNICODE_NEW

	E20	E21	E22	E23	E24	E25	E26	E27	E28	E29	E2A	E2B	E2C	E2D	E2E	E2F	E30	E31	E32	E33	E34	E35	E36	E37	E38	E39	E3A	E3B	E3C	E3D	E3E	E3F
0	Null	க்	ங்	ச்	ஞ்	ட்	ண்	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்	ன்	ஜ்	ஶ்	ஷ்	ஸ்	ஹ்	க்ஷ்					ணு	உத	
1	அ	க	ங	ச	ஞ	ட	ண	த	ந	ப	ம	ய	ர	ல	வ	ழ	ள	ற	ன	ஜ	ஶ	ஷ	ஸ	ஹ	க்ஷ					று	சு	
2	ஆ	கா	ஙா	சா	ஞா	டா	ணா	தா	நா	பா	மா	யா	ரா	லா	வா	ழா	ளா	றா	னா	ஜா	ஶா	ஷா	ஸா	ஹா	க்ஷா					ணூ	பீ	
3	இ	கி	ஙி	சி	ஞி	டி	ணி	தி	நி	பி	மி	யி	ரி	லி	வி	ழி	ளி	றி	னி	ஜி	ஶி	ஷி	ஸி	ஹி	க்ஷி					உ	சூ	
4	ஈ	கீ	ஙீ	சீ	ஞீ	டீ	ணீ	தீ	நீ	பீ	மீ	யீ	ரீ	லீ	வீ	ழீ	ளீ	றீ	னீ	ஜீ	ஶீ	ஷீ	ஸீ	ஹீ	க்ஷீ					ணை	சு	
5	உ	கு	ங	சு	ஞ	டு	ணு	து	நு	பு	மு	யு	ரு	லு	வு	ழு	ளு	று	னு	ஜு	ஶு	ஷு	ஸு	ஹு	க்ஷு					லை	ப	
6	ஊ	கூ	ஙூ	சூ	ஞூ	டூ	ணூ	தூ	நூ	பூ	மூ	யூ	ரூ	லூ	வூ	ழூ	ளூ	றூ	னூ	ஜூ	ஶூ	ஷூ	ஸூ	ஹூ	க்ஷூ					ளை	பூ	
7	எ	கெ	ஙெ	செ	ஞெ	டெ	ணெ	தெ	நெ	பெ	மெ	யெ	ரெ	லெ	வெ	ழெ	ளெ	றெ	னெ	ஜெ	ஶெ	ஷெ	ஸெ	ஹெ	க்ஷெ					னை	நி	
8	ஏ	கே	ஙே	சே	ஞே	டே	ணே	தே	நே	பே	மே	யே	ரே	லே	வே	ழே	ளே	றே	னே	ஜே	ஶே	ஷே	ஸே	ஹே	க்ஷே					○	நூ	
9	ஐ	கை	ஙை	சை	ஞை	டை	ணை	தை	நை	பை	மை	யை	ரை	லை	வை	ழை	ளை	றை	னை	ஜை	ஶை	ஷை	ஸை	ஹை	க்ஷை					●	வு	
A	ஒ	கொ	ஙொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ	ளொ	றொ	னொ	ஜொ	ஶொ	ஷொ	ஸொ	ஹொ	க்ஷொ					★	ரி	
B	ஓ	கோ	ஙோ	சோ	ஞோ	டோ	ணோ	தோ	நோ	போ	மோ	யோ	ரோ	லோ	வோ	ழோ	ளோ	றோ	னோ	ஜோ	ஶோ	ஷோ	ஸோ	ஹோ	க்ஷோ					ஃ	உ	
C	ஔ	கௌ	ஙௌ	சௌ	ஞௌ	டௌ	ணௌ	தௌ	நௌ	பௌ	மௌ	யௌ	ரௌ	லௌ	வௌ	ழௌ	ளௌ	றௌ	னௌ	ஜௌ	ஶௌ	ஷௌ	ஸௌ	ஹௌ	க்ஷௌ					ஔ	சூ	
D	ஃ																								ஶ்ரீ					று		
E																														வ		
F																														?		

Note : E39, E3A, E3B & E3C are reserved for TAMIL accent, Punctuation marks & any feature characters.
E3D to E3F are reserved for Tamil Symbols

TUNE

- The GOOD
  - Code points for almost all characters
  - Symbols for Fractions included
  - Table in natural sort order (almost)
- The BAD
  - Encoding in Unicode Private Use Area Range E000–F8FF
    - Reason enough to reject TUNE
  - DUCET Sort order ???
  - Archaic orthographical Tamil chars missing
  - Measurement symbols missing

Unicode Private Use Area
Range: E000–F8FF

- The Private Use Area does not contain any character assignments, consequently no character code charts or namelists are provided for this area.
- Q. What about using private-use characters for encoding Tamil clusters?
  - A: This is a fine solution for internal processing, if an alternate representation is useful for the particular process. For example, a text-to-speech program might use a private-use encoding for English, whereby letters were separated according to pronunciations -- the 'o' in 'love', 'rove', and 'move' all getting different private-use characters.
  - However, such implementations have limited usage. Private-use characters may overlap between different implementations, so general purpose programs cannot assume any particular interpretation of such characters. In general interchange, such as in search engines, private-use characters are typically treated as unknown characters or ignored. As a result private-use characters are inappropriate for open interchange.
- Source: http://www.unicode.org Tamil FAQ and PUA pages

Fixing issues with UNICODE Tamil Range 0B80 - 0BFF

- Give correct collation order of Tamil Chars
  - http://www.unicode.org/Public/UCA/4.1.0/allkeys.txt
  - patch at http://freeshell.in/~ramanraj/allkeys_patch.txt
- Include ALL missing characters
  - Tamil is the oldest language in active use
  - Send chars after exhaustive orthographical research
  - Seek a suitable range for Tamil in the Unicode standard
- Reject TUNE
  - Refer to Unicode Tamil FAQ
  - Send representations to Gov, ELCOT & TVU

Unicode Planes

* Plane 0 (0000–FFFF): Basic Multilingual Plane (BMP)
* Plane 1 (10000–1FFFF): Supplementary Multilingual Plane (SMP)
* Plane 2 (20000–2FFFF): Supplementary Ideographic Plane (SIP)
* Planes 3 to 13 (30000–DFFFF) are unassigned
* Plane 14 (E0000–EFFFF): Supplementary Special-purpose Plane (SSP)
* Plane 15 (F0000–FFFFF) reserved for the Private Use Area (PUA)
* Plane 16 (100000–10FFFF), reserved for the Private Use Area (PUA)

Notes:

- 40k Chinese ideographs in SIP (Plane 2)
- Chinese language uses a logographic script — that is, a script where one or two "characters" corresponds roughly to one "word" or meaning
- GB18030 is the Chinese National Standard for information interchange, the Chinese equivalent of UTF-8 and as of August 1, 2006, support for this character set is officially mandatory for all software products sold in the PRC.
- Indic scripts and root words may seek Plane 3 and enforce it.

The Encoding Challenge

- ¼ = கால் 
- ½ = அரை 
- 1/8 = அரைக்கால்
- ¾ = முக்கால் 
- 1/16 = வீசம்
- 1/80 = காணி
- 1/320 = முந்திரி
- 1/20 = மா
- 1/5 = ?
  - நாலுமா
- 3/20 = ?
  - மும்மா
- 1/10 = ?
  - இருமா
- 1/40 = ?
  - அரைமா
- 1/640 = கீழ்ழரை
- 1/5120 = கீழ்வீசம்
- 1/2560 = கீழ்ழரைக்கால்
- 1/102400 = கீழ்முந்திரி
- 1/1075200 = இம்மி
- 8 = எட்டு = byte = ?
- 16 = ?
- 256, 512, 1024 .... ??

How to encode all this information ?

யுனிக்கோட் எண்வகை

நன்னூல் நூற்பா 193 : எட்டு என்பதற்கு சிறப்பு விதி

"எட்டன் உடம்புணவ் வாகும் என்ப"

விளக்கம்: இறுதி உயிர்மெய் கெட்டு நின்ற எட்டு என்னும் எண்ணினது டகர மெய் நாற்கணமும் வரின் ணகர மெய்யாகத் திரியும்

எட்டு + ஆயிரம் = எண்ணாயிரம்

எட்டு + வகை = எண்வகை

எட்டு + நாள் = எண்ணாள்

எட்டு + கலம் = எண்கலம்

UTF-8 = யுனிக்கோட் எண்வகை மாற்று

UTF-16 = யுனிக்கோட் ஈரெண்வகை மாற்று

UTF-32 = யுனிக்கோட் நாலெண்வகை மாற்று

UTF-8 யுனிக்கோட் எண்வகை மாற்று

Unicode Transformation Format :: UTF-8 :: Code range

hexadecimal	**UTF-8**
000000 – 00007F 128 codes	0zzzzzzz
000080 – 0007FF 1920 codes	110yyyyy 10zzzzzz
000800 – 00FFFF 63488 codes	1110xxxx 10yyyyyy 10zzzzzz
010000 – 10FFFF 1048576 codes	11110www 10xxxxxx 10yyyyyy 10zzzzzz

- Notes:
  - The first 128 characters (Basic Latin) need only one byte for encoding.
  - The next 1920 characters need two bytes to encode.
  - This includes Latin alphabet characters with diacritics, Greek, Cyrillic, Coptic, Armenian, Hebrew, and Arabic characters.
  - The rest of the BMP characters use three bytes, and additional characters are encoded in four bytes. [Wikipedia on UTF-8]

Encoding Challenge ?FUCCT?

- Free Universal Character and Content Table
- FUCCT {Natural Encoding :-} object includes
  - Letters
  - Root words
  - Content
  - Pronunciation
  - Musical scale
  - Mathematics
  - Geometry
  - Relationship with others
  - Scheme with unlimited space and scope
- The Challenge is Open

References

- http://www.unicode.org/reports/tr10/
- http://unicode.org/Public/cldr/1.4.0/
- http://unicode.org/faq/tamil.html
- http://www.unicode.org/reports/tr6/
- http://people.w3.org/rishida/scripts/pickers/tamil/
- http://unifont.org/fontguide/
- http://www.tunerfc.tn.nic.in/
- http://developer.mimer.com/charts/tamil.htm
- நன்னூல் எழுத்ததிகாரம், நன்னூல் சொல்லதிகாரம்
- தொல்காப்பியம்
- கொடுந்தமிழ் - Joseph Beschi