Development of Thai Encoding and the Implementations

Virach Sornlertlamvanich
Thai Computational Linguistics Laboratory
Communications Research Laboratory, Thailand

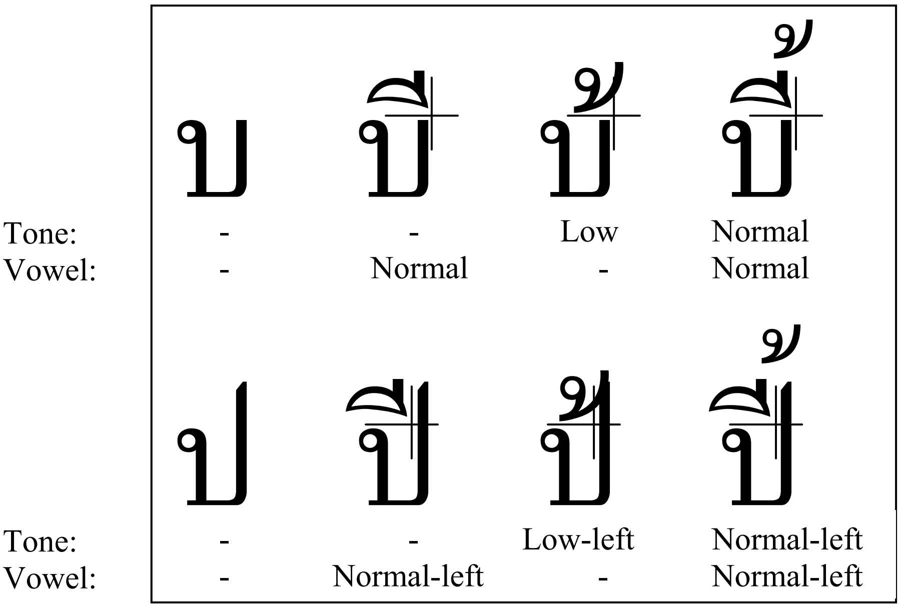
Introduction

The Thai script is classified into the group of Indic and Southeast Asia script by observing its writing system. The base consonant letter is modified by a vowel sign and a tone mark. A vowel sign letter is placed either before (to the left of) or after (to the right of) or around (on both sides of) or above (on top of) or below (under) the base consonant letter. A tone mark is placed on the top of either the base consonant letter or the above vowel letter if exists. The orthography looks complicated and closed to others in the Indic family languages. However, each letter has its own glyph and fixed relative position to the base consonant letter. One letter has one glyph in basis, except for kerning including resizing and space filling to make the combination of the glyphs looks more complex. One letter has one code assigned in basis but however, because of the font, rendering and printing technology, in the beginning of the implementation, multiple glyphs are assigned to different codes. Memory-based code conversion method is introduced to shape the combination of letters. Thai encoding has been done by trial and error in several years and resulted in many versions of character sets. Thai writing system itself has a sophisticated feature that makes the encoding of a character cannot be directly mapped one-to-one to the display character. It needs an additional rendering process to shape the character to realize the high quality of the documentation.

Script, Glyph and Writing System

Thai has its own script as well as most of the languages used in Asia. There are 44 consonant, 21 vowel and 4 tone mark letters in Thai. Each group of the letters has a fixed relative position to the base consonant letter. Thai letter is a consonant possessing an inherent vowel sound. It further features inherent tone. The inherent vowel and tone can be modified by means of vowel signs and tone marks attached to the base consonant letter. Some of the vowel signs and all of the tone marks are rendered in the script as diacritics attached above or below the base consonant.

Glyph is an image used in the visual representation of characters. In the Thai language, one letter has one glyph on basis. But unfortunately, due to the limitation of the font and rendering technology, one letter has more than one glyph assigned. Figure 1 shows 4 different glyphs of an upper tone mark, i.e. normal, low, low-left and normal-left; and 2 different glyphs of an upper vowel sign, i.e. normal and normal-left. Each glyph is prepared and assigned to different code for the sake of glyph selection according to the base consonant in shaping process.



Tone: - - Low Normal
Vowel: - Normal - Normal
Tone: - - Low-left Normal-left
Vowel: - Normal-left - Normal-left

Figure 1 Glyphs of a vowel sign and a tone mark

The consonant letter is placed on the base line. The base consonant letter is modified by a vowel sign and a tone mark. A vowel sign letter is placed either before (to the left of) or after (to the right of) or around (on both sides of) or above (on top of) or below (under) the base consonant letter. A tone mark is placed on the top of either the base consonant letter or the above vowel letter if exists.

Some sounds of vowel are represented by a combination of multiple vowel signs in case that it is a compound vowel. Therefore, a consonant can be modified by more than one vowel sign but not the tone mark. However, the lower vowel sign does not go together with the upper vowel sign when attaching to a consonant. As a result, it is not possible to have a combination of 4-levels letters in any case.

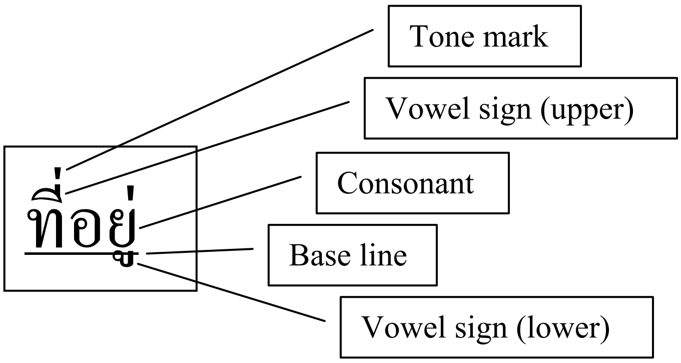


Figure 2 Thai orthography

Most of the languages used in Asia have a particular script which shares some similar features. Considering the writing system, the languages can be classified into 4 major groups. Namely,

1. *Ideographical script*

   The script consists of Chinese ideograms and of characters invented in the areas around China under the influence of Chinese writing. The ideogram represents the idea of a thing rather than the sounds of a word. Each character is enough to represent a meaning of a word, and usually combines with other characters to produce a compound word. It includes Japanese, Khitan, Nuchen, Hsihsia, Korean, Annanese, Lolo, and Moso characters. Characters in this group are generally written from top to bottom on vertical lines shifting from right to left.
2. *Indic and Southeast Asia script*

   The script consists of the Brahmi script that developed in ancient India. It contains many varieties of Indian script as well as Khotanese, Tibetan, Burman, Thai, Shan, Lao, Khmer, Ceylonese, Sumatran, Javanese, Celebes, and Philippine scripts. Scripts of this group are generally written from left to right on a horizontal line.
3. *Islamic script*

   The script consists of the descendants of the Aramaic script that originated in Syria. It contains the Hebrew, South Arabic, Arabic, Avesta, Syriac, Middle Persian, Sogdian, Kok Turki, Uighur, Mongolian, and Manchu scripts. Scripts of this group are mostly written from right to left on a horizontal line.
4. *Roman script with diacritical marks*

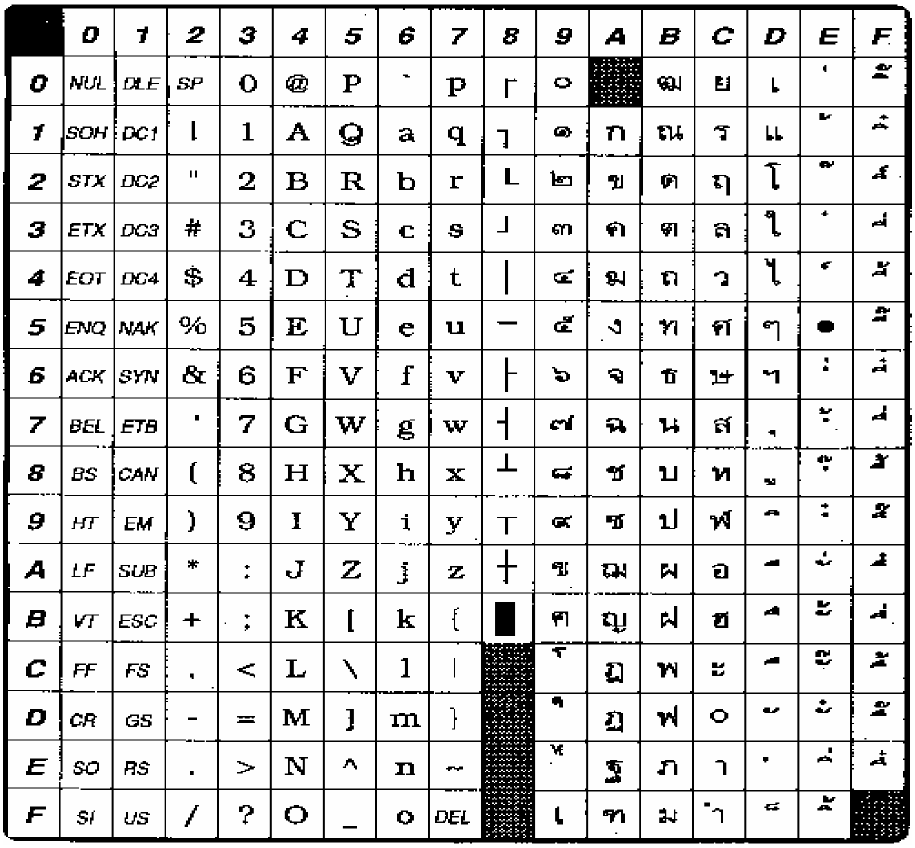
   The languages in this group are modified to use Roman script with additional diacritical marks in case of discriminating the pronunciations or meanings while still keep the original pronunciations and usages. It includes Malay, Indonesian, Philippine, and Vietnamese scripts. Scripts of this group are generally written from left to right on a horizontal line.

Encoding

There are 87 letters in total, including consonants, vowel signs, tone marks, symbols, and Thai digits. Basically, the upper part of the 8-bits ASCII table is large enough to provide a code point to each letter. Due to the past font and rendering technology, extensional code points are assigned to some glyphs to overcome the rendering and printing problems. The special glyphs for the upper tone marks and upper vowel signs are added as discussed in the previous Section.

In this Section, to show the development of the character code table, we make a collection of the code tables that are implemented in several applications. Some are registered as the standards and some are used de facto.

<u>*KU Code*</u>



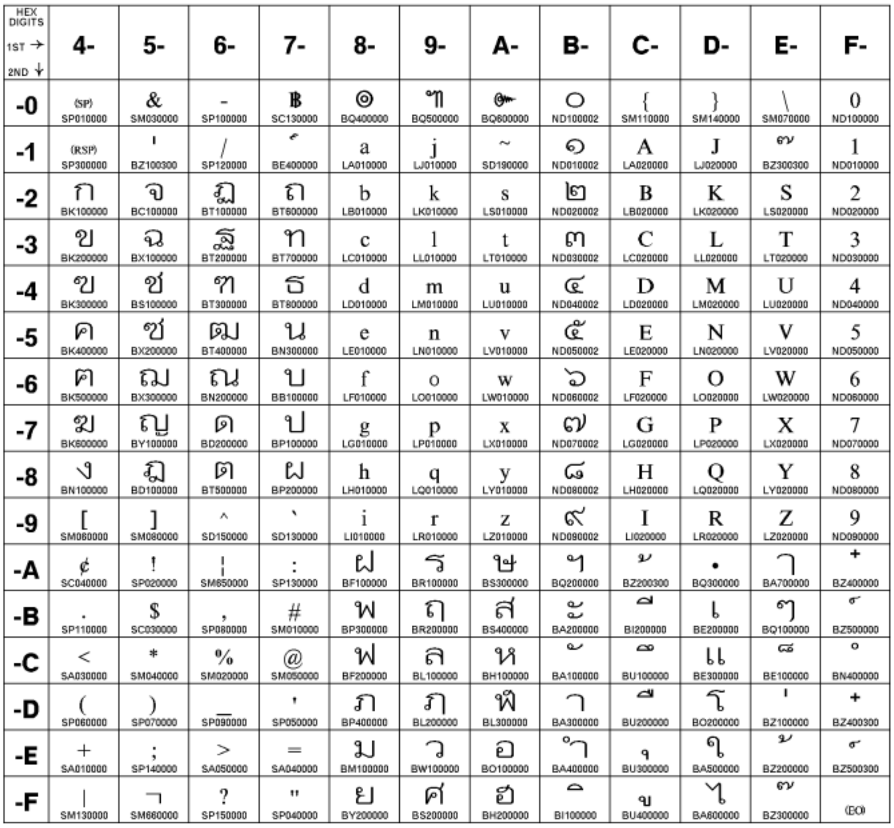
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	DLE	SP	0	@	P	`	p	┌	๐		ฒ	ย	เ	่	[illegible]
1	SOH	DC1	!	1	A	Q	a	q	┐	๑	ก	ณ	ร	แ	้	[illegible]
2	STX	DC2	"	2	B	R	b	r	└	๒	ข	ด	ฤ	โ	๊	[illegible]
3	ETX	DC3	#	3	C	S	c	s	┘	๓	ค	ต	ล	ใ	๋	[illegible]
4	EOT	DC4	$	4	D	T	d	t	│	๔	ฆ	ถ	ว	ไ	์	[illegible]
5	ENQ	NAK	%	5	E	U	e	u	─	๕	ง	ท	ศ	ๆ	●	[illegible]
6	ACK	SYN	&	6	F	V	f	v	├	๖	จ	ธ	ษ	ฯ	[illegible]	[illegible]
7	BEL	ETB	'	7	G	W	g	w	┤	๗	ฉ	น	ส	ฺ	[illegible]	[illegible]
8	BS	CAN	(	8	H	X	h	x	┴	๘	ช	บ	ห	ุ	[illegible]	[illegible]
9	HT	EM	)	9	I	Y	i	y	┬	๙	ซ	ป	ฬ	ั	[illegible]	[illegible]
A	LF	SUB	*	:	J	Z	j	z	┼	ฃ	ฌ	ผ	อ	ิ	[illegible]	[illegible]
B	VT	ESC	+	;	K	[	k	{	█	ฅ	ญ	ฝ	ฮ	ี	[illegible]	[illegible]
C	FF	FS	,	<	L	\	l	|		[illegible]	ฎ	พ	ะ	ึ	[illegible]	[illegible]
D	CR	GS	-	=	M	]	m	}		[illegible]	ฏ	ฟ	○	ื	[illegible]	[illegible]
E	SO	RS	.	>	N	^	n	~		[illegible]	ฐ	ภ	า	ํ	[illegible]	[illegible]
F	SI	US	/	?	O	_	o	DEL		เ	ฑ	ม	ำ	็	[illegible]	

Table 1 KU Code

In the early days, there were many proposals of Thai encoding table. One of the most well-accepted character set was the one that developed at Kasetsart University around the early of 1980s. It was implemented in several programs for Thai text processing. The table had been modified to support dot matrix printer. All possible combinations of the upper vowel sign and tone mark letters are pre-defined to suit the base consonant letters. This character set is scarcely used today.

IBM CP838

IBM Corp had registered its Code Page 00838 (EBCDIC) Thai Extended with the Internet Assigned Numbers Authority, IANA. This character set is also listed as a supported encoding by Sun Microsystems' Java Development Kit version 1.1 (JDK 1.1).



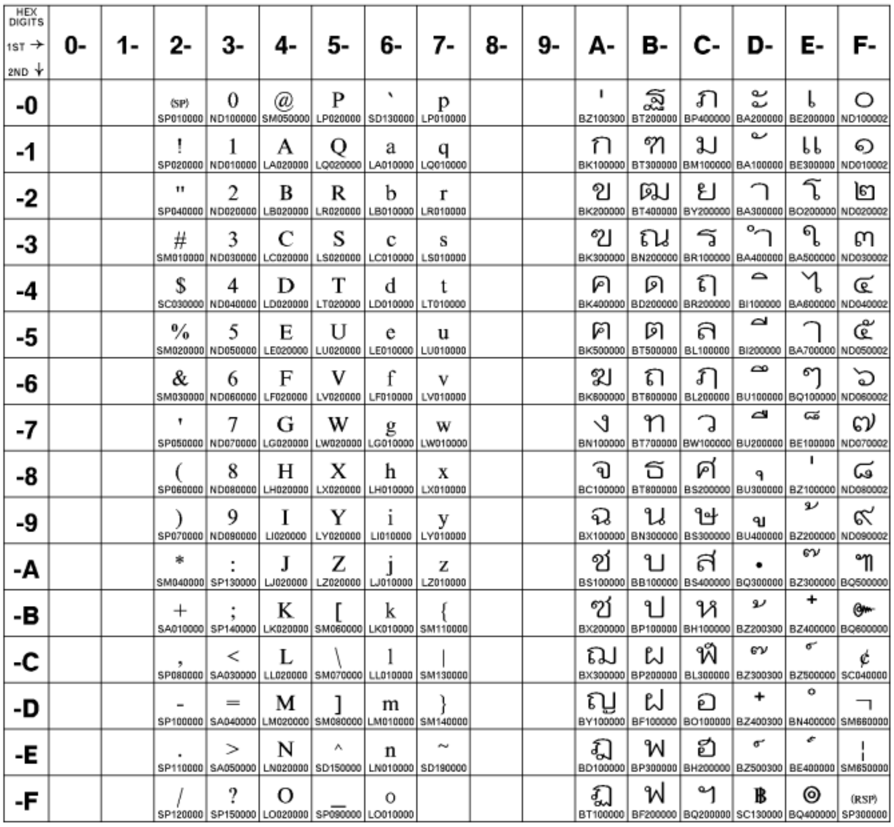
HEX DIGITS 1ST → 2ND ↓	4-	5-	6-	7-	8-	9-	A-	B-	C-	D-	E-	F-
-0	(SP) SP010000	& SM030000	- SP100000	฿ SC130000	๏ BQ400000	๚ BQ500000	๛ BQ600000	๐ ND100002	{ SM110000	} SM140000	\ SM070000	0 ND100000
-1	(RSP) SP300000	่ BZ100300	/ SP120000	๎ BE400000	a LA010000	j LJ010000	~ SD190000	๑ ND010002	A LA020000	J LJ020000	๊ BZ300300	1 ND010000
-2	ก BK100000	จ BC100000	ฏ BT100000	ถ BT600000	b LB010000	k LK010000	s LS010000	๒ ND020002	B LB020000	K LK020000	S LS020000	2 ND020000
-3	ข BK200000	ฉ BX100000	ฐ BT200000	ท BT700000	c LC010000	l LL010000	t LT010000	๓ ND030002	C LC020000	L LL020000	T LT020000	3 ND030000
-4	ฃ BK300000	ช BS100000	ฑ BT300000	ธ BT800000	d LD010000	m LM010000	u LU010000	๔ ND040002	D LD020000	M LM020000	U LU020000	4 ND040000
-5	ค BK400000	ซ BX200000	ฒ BT400000	น BN300000	e LE010000	n LN010000	v LV010000	๕ ND050002	E LE020000	N LN020000	V LV020000	5 ND050000
-6	ฅ BK500000	ฌ BX300000	ณ BN200000	บ BB100000	f LF010000	o LO010000	w LW010000	๖ ND060002	F LF020000	O LO020000	W LW020000	6 ND060000
-7	ฆ BK600000	ญ BY100000	ด BD200000	ป BP100000	g LG010000	p LP010000	x LX010000	๗ ND070002	G LG020000	P LP020000	X LX020000	7 ND070000
-8	ง BN100000	ฎ BD100000	ต BT500000	ผ BP200000	h LH010000	q LQ010000	y LY010000	๘ ND080002	H LH020000	Q LQ020000	Y LY020000	8 ND080000
-9	[SM060000	] SM080000	^ SD150000	` SD130000	i LI010000	r LR010000	z LZ010000	๙ ND090002	I LI020000	R LR020000	Z LZ020000	9 ND090000
-A	¢ SC040000	! SP020000	¦ SM650000	: SP130000	ฝ BF100000	ร BR100000	ษ BS300000	ฯ BQ200000	้ BZ200300	ฺ BQ300000	ๅ BA700000	๋ BZ400000
-B	. SP110000	$ SC030000	, SP080000	# SM010000	พ BP300000	ฤ BR200000	ส BS400000	ะ BA200000	ี BI200000	เ BE200000	ๆ BQ100000	์ BZ500000
-C	< SA030000	* SM040000	% SM020000	@ SM050000	ฟ BF200000	ล BL100000	ห BH100000	ั BA100000	ึ BU100000	แ BE300000	็ BE100000	ํ BN400000
-D	(SP060000	) SP070000	_ SP090000	' SP050000	ภ BP400000	ฦ BL200000	ฬ BL300000	า BA300000	ื BU200000	โ BO200000	่ BZ100000	๋ BZ400300
-E	+ SA010000	; SP140000	> SA050000	= SA040000	ม BM100000	ว BW100000	อ BO100000	ำ BA400000	ุ BU300000	ใ BA500000	้ BZ200000	์ BZ500300
-F	| SM130000	¬ SM660000	? SP150000	" SP040000	ย BY200000	ศ BS200000	ฮ BH200000	ิ BI100000	ู BU400000	ไ BA600000	๊ BZ300000	(EO)

Code Page 00838

Table 2 Code Page 00838 (EBCDIC) Thai Extended

IBM CP874

This character set is listed as a supported encoding by the JDK 1.1. Most of the code points are assigned same as the current standard TIS 620. Some of the unused codes are assigned privately to the tone marks. It is known as the original character table for *Windows-874*.



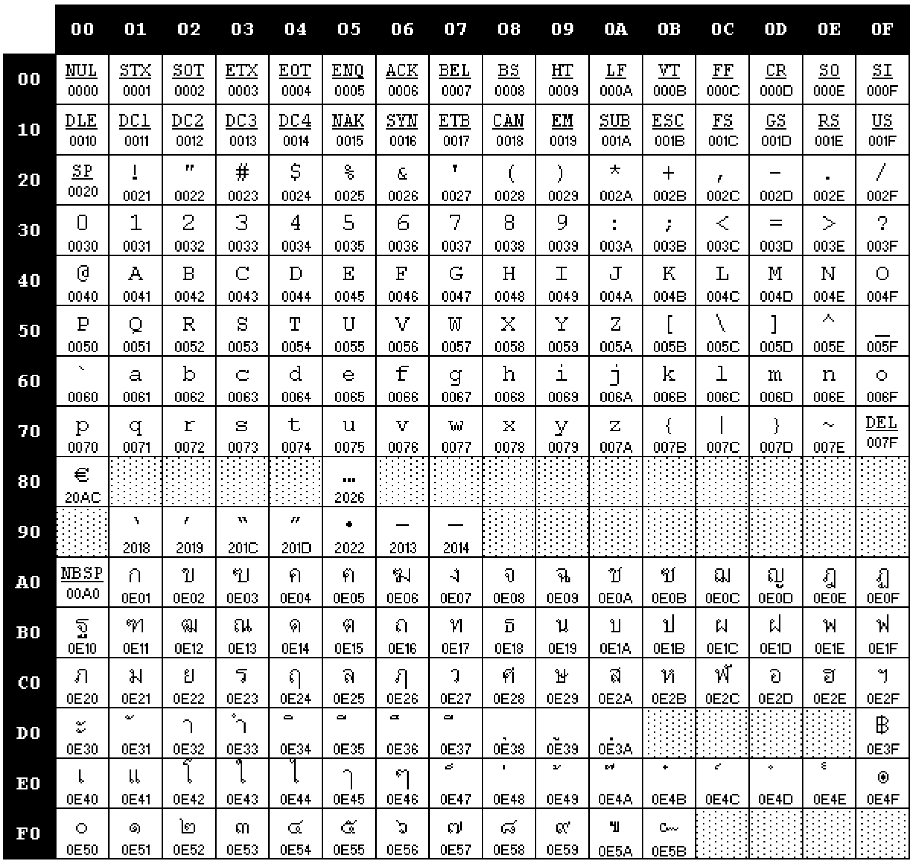
HEX DIGITS 1ST → 2ND ↓	0-	1-	2-	3-	4-	5-	6-	7-	8-	9-	A-	B-	C-	D-	E-	F-
-0			(SP) SP010000	0 ND100000	@ SM050000	P LP020000	` SD130000	p LP010000			่ BZ100300	ฐ BT200000	ภ BP400000	ะ BA200000	เ BE200000	๐ ND100002
-1			! SP020000	1 ND010000	A LA020000	Q LQ020000	a LA010000	q LQ010000			ก BK100000	ฑ BT300000	ม BM100000	ั BA100000	แ BE300000	๑ ND010002
-2			" SP040000	2 ND020000	B LB020000	R LR020000	b LB010000	r LR010000			ข BK200000	ฒ BT400000	ย BY200000	า BA300000	โ BO200000	๒ ND020002
-3			# SM010000	3 ND030000	C LC020000	S LS020000	c LC010000	s LS010000			ฃ BK300000	ณ BN200000	ร BR100000	ำ BA400000	ใ BA500000	๓ ND030002
-4			$ SC030000	4 ND040000	D LD020000	T LT020000	d LD010000	t LT010000			ค BK400000	ด BD200000	ฤ BR200000	ิ BI100000	ไ BA600000	๔ ND040002
-5			% SM020000	5 ND050000	E LE020000	U LU020000	e LE010000	u LU010000			ฅ BK500000	ต BT500000	ล BL100000	ี BI200000	ๅ BA700000	๕ ND050002
-6			& SM030000	6 ND060000	F LF020000	V LV020000	f LF010000	v LV010000			ฆ BK600000	ถ BT600000	ฦ BL200000	ึ BU100000	ๆ BQ100000	๖ ND060002
-7			' SP050000	7 ND070000	G LG020000	W LW020000	g LG010000	w LW010000			ง BN100000	ท BT700000	ว BW100000	ื BU200000	็ BE100000	๗ ND070002
-8			(SP060000	8 ND080000	H LH020000	X LX020000	h LH010000	x LX010000			จ BC100000	ธ BT800000	ศ BS200000	ุ BU300000	่ BZ100000	๘ ND080002
-9			) SP070000	9 ND090000	I LI020000	Y LY020000	i LI010000	y LY010000			ฉ BX100000	น BN300000	ษ BS300000	ู BU400000	้ BZ200000	๙ ND090002
-A			* SM040000	: SP130000	J LJ020000	Z LZ020000	j LJ010000	z LZ010000			ช BS100000	บ BB100000	ส BS400000	• BQ300000	๊ BZ300000	๚ BQ500000
-B			+ SA010000	; SP140000	K LK020000	[SM060000	k LK010000	{ SM110000			ซ BX200000	ป BP100000	ห BH100000	้ BZ200300	๋ BZ400000	๛ BQ600000
-C			, SP080000	< SA030000	L LL020000	\\ SM070000	l LL010000	| SM130000			ฌ BX300000	ผ BP200000	ฬ BL300000	๊ BZ300300	์ BZ500000	¢ SC040000
-D			- SP100000	= SA040000	M LM020000	] SM080000	m LM010000	} SM140000			ญ BY100000	ฝ BF100000	อ BO100000	๋ BZ400300	ํ BN400000	¬ SM660000
-E			. SP110000	> SA050000	N LN020000	^ SD150000	n LN010000	~ SD190000			ฎ BD100000	พ BP300000	ฮ BH200000	์ BZ500300	๎ BE400000	¦ SM650000
-F			/ SP120000	? SP150000	O LO020000	_ SP090000	o LO010000				ฏ BT100000	ฟ BF200000	ฯ BQ200000	฿ SC130000	๏ BQ400000	(RSP) SP300000

Code Page 00874

Table 3 Code Page 00874 (IBM Personal Computer) Thai Extended

Windows-874

This character set is used as a code page for Thai in Microsoft products, i.e. the MS-DOS and the family of the Windows operating system. This character set is listed as a supported encoding by the JDK 1.1. It is widely used as a de facto standard because the name of the code page Windows-874 is recognized in all the Microsoft products which are very popular in recent years. It is completely compatible with TIS 620 but the name is much widely recognized.

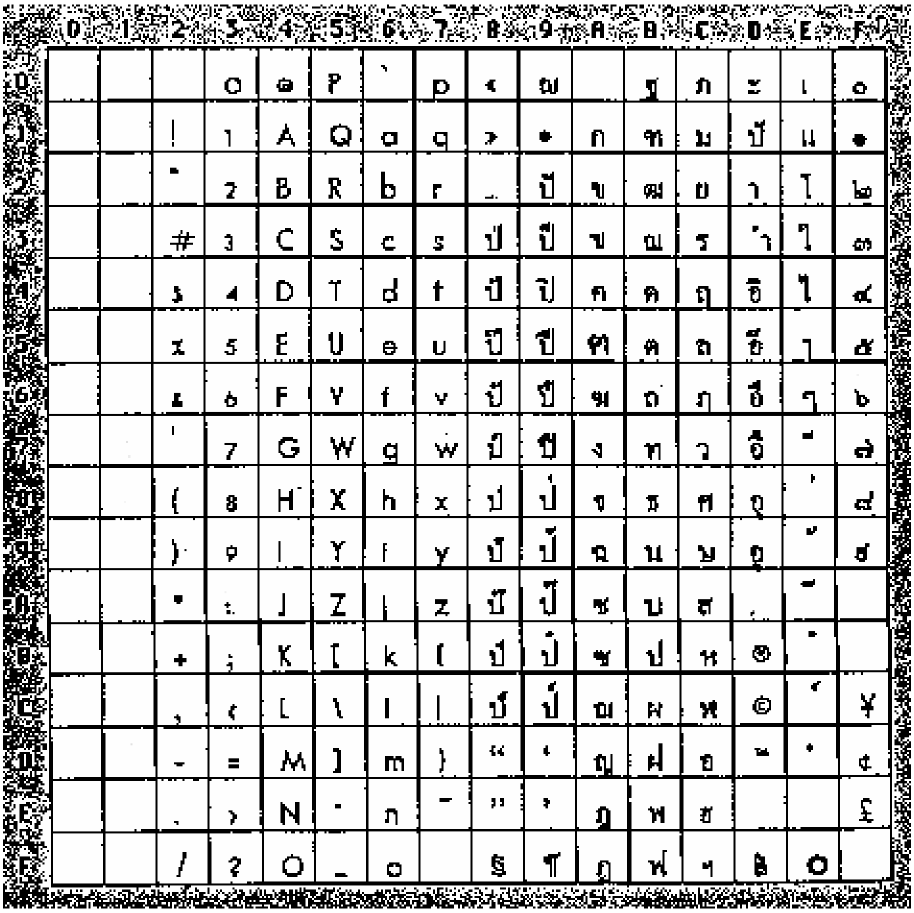


	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL 0000	STX 0001	SOT 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
10	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
20	SP 0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
30	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
40	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
50	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
60	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
70	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	DEL 007F
80	€ 20AC					… 2026										
90		‘ 2018	’ 2019	“ 201C	” 201D	• 2022	– 2013	— 2014								
A0	NBSP 00A0	ก 0E01	ข 0E02	ฃ 0E03	ค 0E04	ฅ 0E05	ฆ 0E06	ง 0E07	จ 0E08	ฉ 0E09	ช 0E0A	ซ 0E0B	ฌ 0E0C	ญ 0E0D	ฎ 0E0E	ฏ 0E0F
B0	ฐ 0E10	ฑ 0E11	ฒ 0E12	ณ 0E13	ด 0E14	ต 0E15	ถ 0E16	ท 0E17	ธ 0E18	น 0E19	บ 0E1A	ป 0E1B	ผ 0E1C	ฝ 0E1D	พ 0E1E	ฟ 0E1F
C0	ภ 0E20	ม 0E21	ย 0E22	ร 0E23	ฤ 0E24	ล 0E25	ฦ 0E26	ว 0E27	ศ 0E28	ษ 0E29	ส 0E2A	ห 0E2B	ฬ 0E2C	อ 0E2D	ฮ 0E2E	ฯ 0E2F
D0	ะ 0E30	ั 0E31	า 0E32	ำ 0E33	ิ 0E34	ี 0E35	ึ 0E36	ื 0E37	ุ 0E38	ู 0E39	ฺ 0E3A					฿ 0E3F
E0	เ 0E40	แ 0E41	โ 0E42	ใ 0E43	ไ 0E44	ๅ 0E45	ๆ 0E46	็ 0E47	่ 0E48	้ 0E49	๊ 0E4A	๋ 0E4B	์ 0E4C	ํ 0E4D	๎ 0E4E	๏ 0E4F
F0	๐ 0E50	๑ 0E51	๒ 0E52	๓ 0E53	๔ 0E54	๕ 0E55	๖ 0E56	๗ 0E57	๘ 0E58	๙ 0E59	๚ 0E5A	๛ 0E5B				

Table 4 Windows-874

MAC Thai

This character set is one of the oldest set defined and used by Apple Computer, Inc. for the Thai implementation under the MacOS operating system. This character set is also a supported encoding in JDK 1.1.

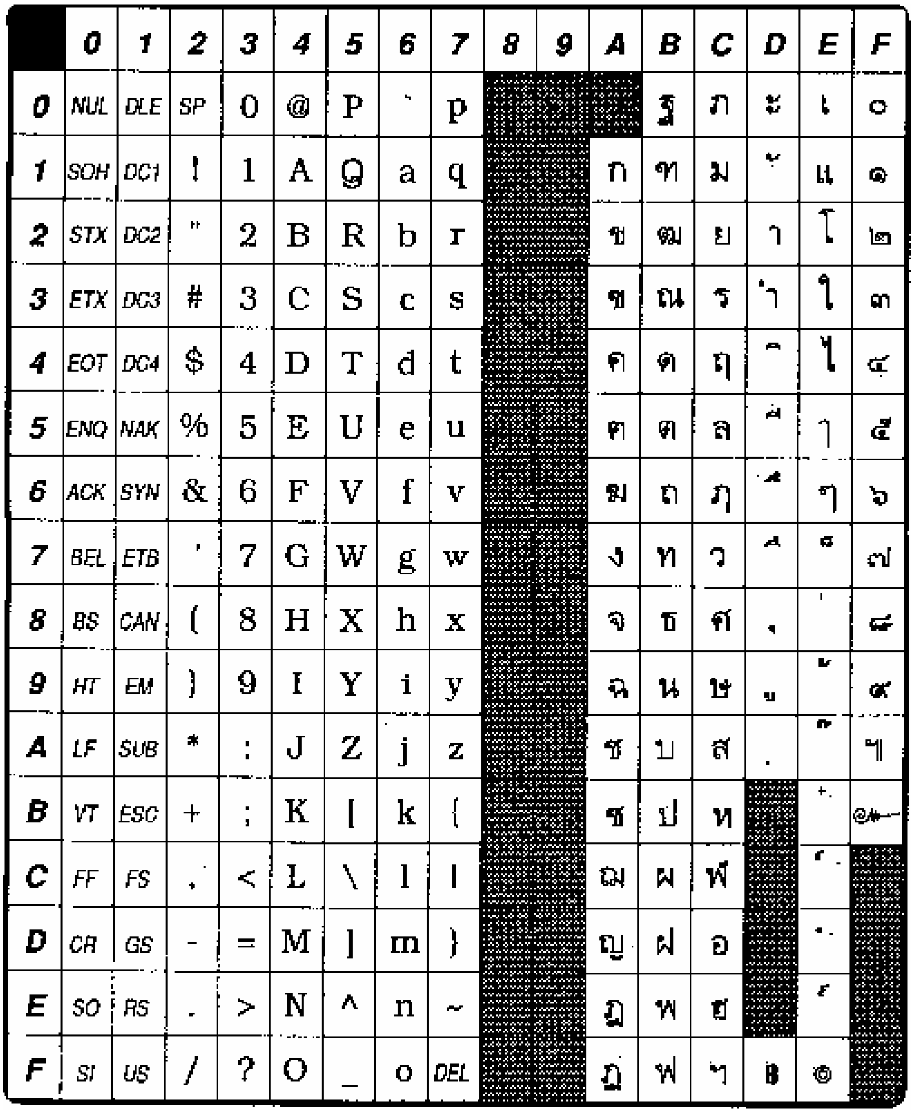


	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0				0	@	P	`	p	«	ญ		ฐ	ภ	ะ	เ	๐
1			!	1	A	Q	a	q	»	•	ก	ฑ	ม	ปั	แ	๑
2			"	2	B	R	b	r	…	ปิ	ข	ฒ	ย	า	โ	๒
3			#	3	C	S	c	s	ป่	ปี	ฃ	ณ	ร	ำ	ใ	๓
4			$	4	D	T	d	t	ป้	ปิ	ค	ด	ฤ	ิ	ไ	๔
5			%	5	E	U	e	u	ป๊	ปี	ฅ	ต	ล	ี	ๅ	๕
6			&	6	F	V	f	v	ป๋	ปึ	ฆ	ถ	ฦ	ึ	ๆ	๖
7			'	7	G	W	g	w	ป์	ปื	ง	ท	ว	ื	็	๗
8			(	8	H	X	h	x	ป่	ป่	จ	ธ	ศ	ุ	่	๘
9			)	9	I	Y	i	y	ป้	ป้	ฉ	น	ษ	ู	้	๙
A			*	:	J	Z	j	z	ป๊	ป๊	ช	บ	ส	ฺ	๊	
B			+	;	K	[	k	{	ป๋	ป๋	ซ	ป	ห	®	๋	
C			,	<	L	\	l	|	ป์	ป์	ฌ	ผ	ฬ	©	์	¥
D			-	=	M	]	m	}	“	‘	ญ	ฝ	อ	™	ํ	¢
E			.	>	N	^	n	~	”	’	ฎ	พ	ฮ			£
F			/	?	O	_	o		§	¶	ฏ	ฟ	ฯ	฿	๏	

Table 5 Mac Thai Code

TIS 620-2533:1990

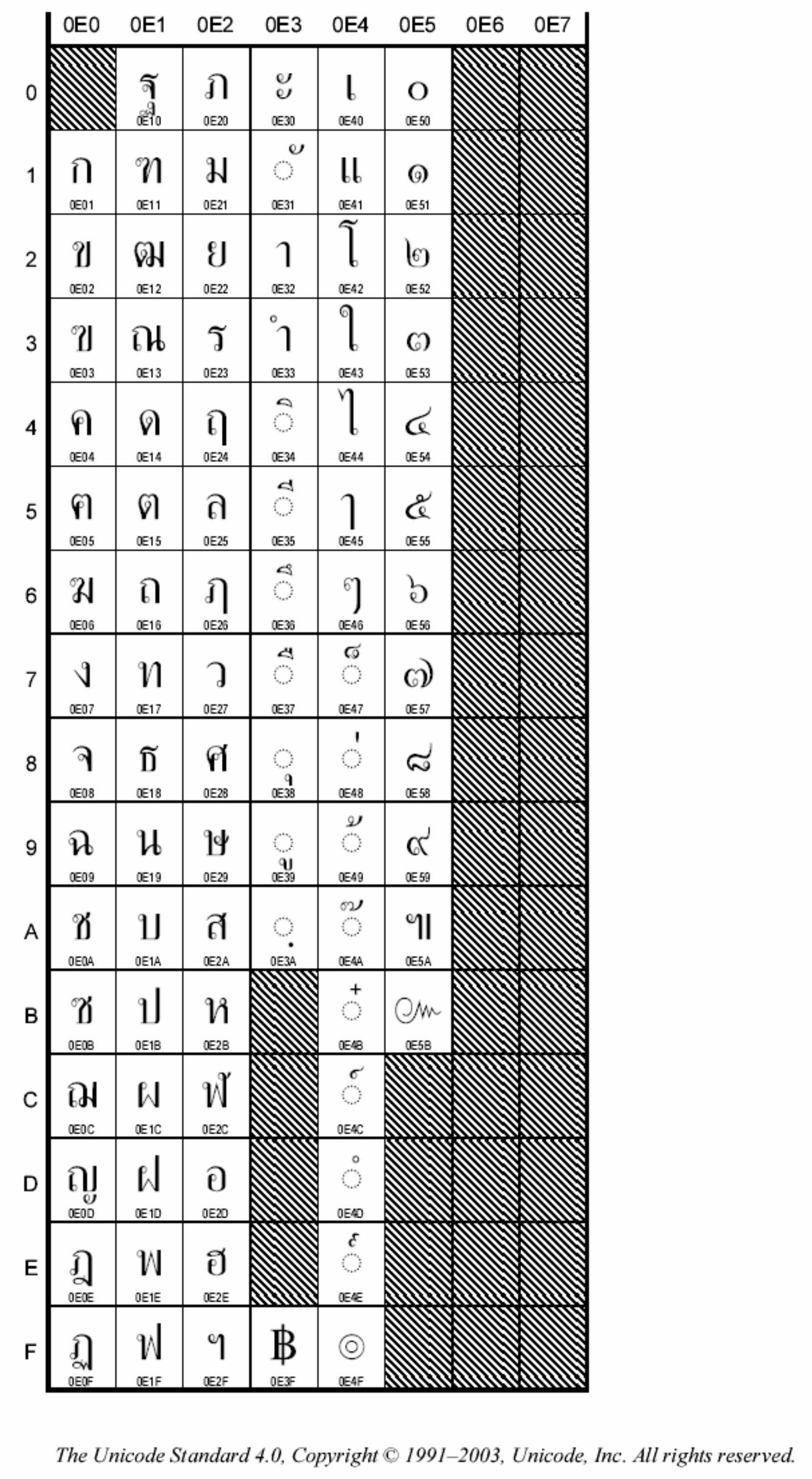
In Thailand, there is the one and only Thai Character Set standard, TIS 620-2533, defined by the Thai Industrial Standards Institute (TISI), Ministry of Industry, Royal Thai Government. It was the work of the 536th Technical Committee, TC536, who was, and is still, in charge of Thai Information Technology Standards. TIS 620-2533 is a revision of the earlier standard TIS 620-2529. Assignments of each code point in these two versions of TIS 620 remains the same.



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	DLE	SP	0	@	P	`	p				ฐ	ภ	ะ	เ	๐
1	SOH	DC1	!	1	A	Q	a	q			ก	ฑ	ม	ั	แ	๑
2	STX	DC2	"	2	B	R	b	r			ข	ฒ	ย	า	โ	๒
3	ETX	DC3	#	3	C	S	c	s			ฃ	ณ	ร	ำ	ใ	๓
4	EOT	DC4	$	4	D	T	d	t			ค	ด	ฤ	ิ	ไ	๔
5	ENQ	NAK	%	5	E	U	e	u			ฅ	ต	ล	ี	ๅ	๕
6	ACK	SYN	&	6	F	V	f	v			ฆ	ถ	ฦ	ึ	ๆ	๖
7	BEL	ETB	'	7	G	W	g	w			ง	ท	ว	ื	็	๗
8	BS	CAN	(	8	H	X	h	x			จ	ธ	ศ	ุ	่	๘
9	HT	EM	)	9	I	Y	i	y			ฉ	น	ษ	ู	้	๙
A	LF	SUB	*	:	J	Z	j	z			ช	บ	ส	ฺ	๊	๚
B	VT	ESC	+	;	K	[	k	{			ซ	ป	ห		๋	๛
C	FF	FS	,	<	L	\	l	|			ฌ	ผ	ฬ		์	
D	CR	GS	-	=	M	]	m	}			ญ	ฝ	อ		ํ	
E	SO	RS	.	>	N	^	n	~			ฎ	พ	ฮ		๎	
F	SI	US	/	?	O	_	o	DEL			ฏ	ฟ	ฯ	฿	๏	

Table 6 TIS 620-2533:1990

<u>*Unicode 4.0*</u>



	0E0	0E1	0E2	0E3	0E4	0E5	0E6	0E7
0	▨	ฐ 0E10	ภ 0E20	ะ 0E30	เ 0E40	๐ 0E50	▨	▨
1	ก 0E01	ฑ 0E11	ม 0E21	◌ั 0E31	แ 0E41	๑ 0E51	▨	▨
2	ข 0E02	ฒ 0E12	ย 0E22	า 0E32	โ 0E42	๒ 0E52	▨	▨
3	ฃ 0E03	ณ 0E13	ร 0E23	◌ำ 0E33	ใ 0E43	๓ 0E53	▨	▨
4	ค 0E04	ด 0E14	ฤ 0E24	◌ิ 0E34	ไ 0E44	๔ 0E54	▨	▨
5	ฅ 0E05	ต 0E15	ล 0E25	◌ี 0E35	ๅ 0E45	๕ 0E55	▨	▨
6	ฆ 0E06	ถ 0E16	ฦ 0E26	◌ึ 0E36	ๆ 0E46	๖ 0E56	▨	▨
7	ง 0E07	ท 0E17	ว 0E27	◌ื 0E37	◌็ 0E47	๗ 0E57	▨	▨
8	จ 0E08	ธ 0E18	ศ 0E28	◌ุ 0E38	◌่ 0E48	๘ 0E58	▨	▨
9	ฉ 0E09	น 0E19	ษ 0E29	◌ู 0E39	◌้ 0E49	๙ 0E59	▨	▨
A	ช 0E0A	บ 0E1A	ส 0E2A	◌ฺ 0E3A	◌๊ 0E4A	๚ 0E5A	▨	▨
B	ซ 0E0B	ป 0E1B	ห 0E2B	▨	◌๋ 0E4B	๛ 0E5B	▨	▨
C	ฌ 0E0C	ผ 0E1C	ฬ 0E2C	▨	◌์ 0E4C	▨	▨	▨
D	ญ 0E0D	ฝ 0E1D	อ 0E2D	▨	◌ํ 0E4D	▨	▨	▨
E	ฎ 0E0E	พ 0E1E	ฮ 0E2E	▨	◌๎ 0E4E	▨	▨	▨
F	ฏ 0E0F	ฟ 0E1F	ฯ 0E2F	฿ 0E3F	๏ 0E4F	▨	▨	▨

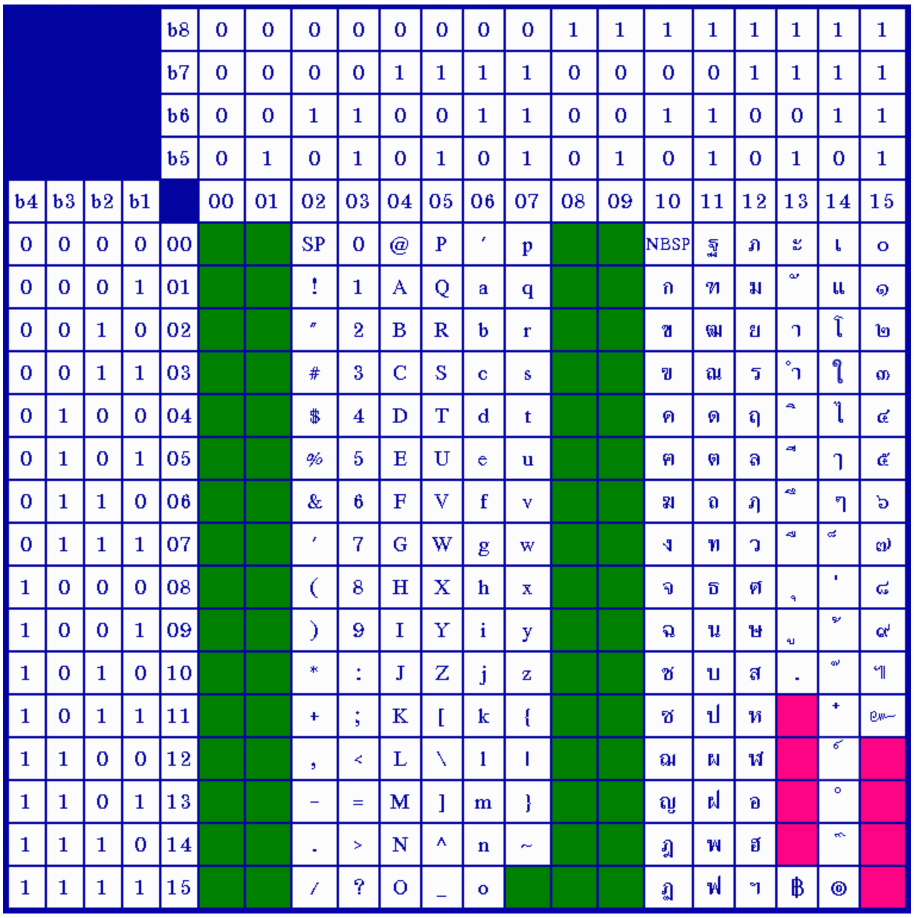
The Unicode Standard 4.0, Copyright © 1991–2003, Unicode, Inc. All rights reserved.

Table 7 Thai Code Page in Unicode 4.0

<u>ISO/IEC 8859-11:2001 – Latin/Thai</u>

ISO/IEC 8859 consists of several parts. Each part specifies a set of up to 191 graphic characters and the coded representation of these characters by means of a single 8-bit bytes. Therefore, a combining character family is not suitable for ISO/IEC 8859. There was a long discussion on whether Thai script has combining characters or not. Some vowel signs and tone marks are somehow attached to the base consonant like a diacritic. In principle each Thai character is a single character. Though some letters are used to modify the base consonant, their shapes do not change and still occupy a single code. Therefore, no combining characters are needed to define in the table.

After a long discussion on the proposal of Thai character set for ISO/IEC 8859, it was finally announced as a standard of ISO/IEC 8859-11 in 2001. It is the same as TIS 620 except for the NBSP (0xA0) which is additionally defined.



				b8	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
				b7	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
				b6	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
				b5	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
b4	b3	b2	b1		00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
0	0	0	0	00	🟩	🟩	SP	0	@	P	`	p	🟩	🟩	NBSP	ฐ	ภ	ะ	เ	๐
0	0	0	1	01	🟩	🟩	!	1	A	Q	a	q	🟩	🟩	ก	ฑ	ม	ั	แ	๑
0	0	1	0	02	🟩	🟩	"	2	B	R	b	r	🟩	🟩	ข	ฒ	ย	า	โ	๒
0	0	1	1	03	🟩	🟩	#	3	C	S	c	s	🟩	🟩	ฃ	ณ	ร	ำ	ใ	๓
0	1	0	0	04	🟩	🟩	$	4	D	T	d	t	🟩	🟩	ค	ด	ฤ	ิ	ไ	๔
0	1	0	1	05	🟩	🟩	%	5	E	U	e	u	🟩	🟩	ฅ	ต	ล	ี	ๅ	๕
0	1	1	0	06	🟩	🟩	&	6	F	V	f	v	🟩	🟩	ฆ	ถ	ฦ	ึ	ๆ	๖
0	1	1	1	07	🟩	🟩	'	7	G	W	g	w	🟩	🟩	ง	ท	ว	ื	็	๗
1	0	0	0	08	🟩	🟩	(	8	H	X	h	x	🟩	🟩	จ	ธ	ศ	ุ	่	๘
1	0	0	1	09	🟩	🟩	)	9	I	Y	i	y	🟩	🟩	ฉ	น	ษ	ู	้	๙
1	0	1	0	10	🟩	🟩	*	:	J	Z	j	z	🟩	🟩	ช	บ	ส	ฺ	๊	๚
1	0	1	1	11	🟩	🟩	+	;	K	[	k	{	🟩	🟩	ซ	ป	ห	🟥	๋	๛
1	1	0	0	12	🟩	🟩	,	<	L	\	l	|	🟩	🟩	ฌ	ผ	ฬ	🟥	์	🟥
1	1	0	1	13	🟩	🟩	-	=	M	]	m	}	🟩	🟩	ญ	ฝ	อ	🟥	ํ	🟥
1	1	1	0	14	🟩	🟩	.	>	N	^	n	~	🟩	🟩	ฎ	พ	ฮ	🟥	๎	🟥
1	1	1	1	15	🟩	🟩	/	?	O	_	o	🟩	🟩	🟩	ฏ	ฟ	ฯ	฿	๏	🟥

Table 8 ISO/IEC 8859-11:2001 – Latin/Thai

Implementation

Encoding is essentially used in emails and web pages. To investigate the utilization of the encoding, we collected the web pages written in Thai as many as possible. Most of them define their content encoding in the meta tag of Charset.

```
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-
8859-1">
<title>Publication</title>
</head>
```

Figure 3 Meta tag in HTML file defining the Charset

From the field of Charset in the meta tag, we accumulate the count and show in Table 9 and 10. Table 9 shows the count of Charset in .th web pages in October 2001 while Table 10 shows the change of the count in July 2003. The counts of the Thai encoding standards (tis-620 and iso-8859-11) significantly increased from 4.89% to 30.36% after the campaign of using the appropriate Thai encoding standards. However, the de facto windows-874 still has the majority (52.06% and 50.55%). It is because of its well acceptance by most of the applications, such as Internet Explorer, MS-Outlook, MS-Word, etc. The encoding iso-8859-1 is wrongly defined in a very ratio because of the using of the modified web page editing tools without the awareness of the encoding.

Charset	Site	%	Charset	Site	%
windows-874	682	52.06	**gb2312**	2	0.15
(blank)	519	39.62	**x-user-defined**	1	0.08
tis-620	61	4.66	**windows874**	1	0.08
iso-8859-1	8	0.61	**Thai(tis-620)**	1	0.08
shift_jis	8	0.61	**thai(Windows)**	1	0.08
window-874	6	0.46	**TIS620**	1	0.08
windows-1252	3	0.23	**tis620)**	1	0.08
utf-8	3	0.23	**window**	1	0.08
euc-kr	3	0.23	**windows-128**	1	0.08
iso-8859-11	3	0.23	**windows-847**	1	0.08
x-sjis	2	0.15	**X-MAC-THAI**	1	0.08
			Total	1310	100

Table 9 Charsets used in Thai web pages (.th), Oct 2001

	.th	.com	.net	.org	Total	%
windows-874	5315	3048	223	39	8625	50.55
Tis-620	2173	2930	40	25	5168	30.29
Iso-8859-1	991	1419	16	38	2464	14.44
Utf-8	51	20	1	3	75	0.44
Iso-8859-11	2	10	0	0	12	0.07
(blank)	496	182	22	20	720	4.22
Total	9028	7609	302	125	17064	100

Table 10 Charsets used in Thai web pages, July 2003

Applications have a strong influence on making use of the standards. To ensure that the documents are compatibly exchanged in the cyber space, we need a standard and the applications that support the encoding standard.

Conclusion

The Thai script is encoded character each. Basically, one character has one code. Due to the feature of the modification of some vowel signs and tone marks to the base consonant letter, the size and the relative position to the base consonant have to adjust. To realize the modification the rendering module needs a shaping process to adjust the size and the relative position according to the context. Since the process is acceptably realized in the memory base manner, there is no need to declare the derived characters or even the composed characters in the standard of the character set table.

Though the standards have been formally registered in the international standard organization (ISO/IEC and Unicode), the dispersal of the de facto standard is conspicuous. We need applications to help supporting the standards and well understanding of the implementers.

References

Akira Nakanishi. 1998. *Writing Systems of the World. –Alphabets, Syllabaries, Pictograms--.* Charles E. Tuttle Company.

Software.thai.net. *An Annotated reference to a Thai Implementations.* http://www.inet.co.th/cyberclub/trin/thairef/#ThaiCharsetStds.

Xencraft. *Character Sets and Code Pages at the Push of a Button.* http://www.i18nguy.com/unicode/codepages.html.

IBM BookManager BookServer Library. *http://publib.boulder.ibm.com/cgi-bin/bookmgr/BOOKS/QB3AQ501/F.0.*

Thaweesak Koanantakool. 1992. *Standard of IT Industry and the Open System.* http://www.nectec.or.th/it-standards/ITStaNew.htm#e9.

The Unicode Consortium. 2003. *The Unicode Standard 4.0.* Addison-Wesley.