Survey of Language Computing in Asia 2005

Sarmad Hussain
Nadir Durrani
Sana Gul

Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences

www.nu.edu.pk

www.idrc.ca

Published by

Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences
Lahore, Pakistan

Copyrights © International Development Research Center, Canada

Printed by Walayatsons, Pakistan

ISBN: 969-8961-00-3

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Ottawa, Canada, administered through the Centre for Research in Urdu Language Processing (CRULP), National University of Computer and Emerging Sciences (NUCES), Pakistan.

Thai

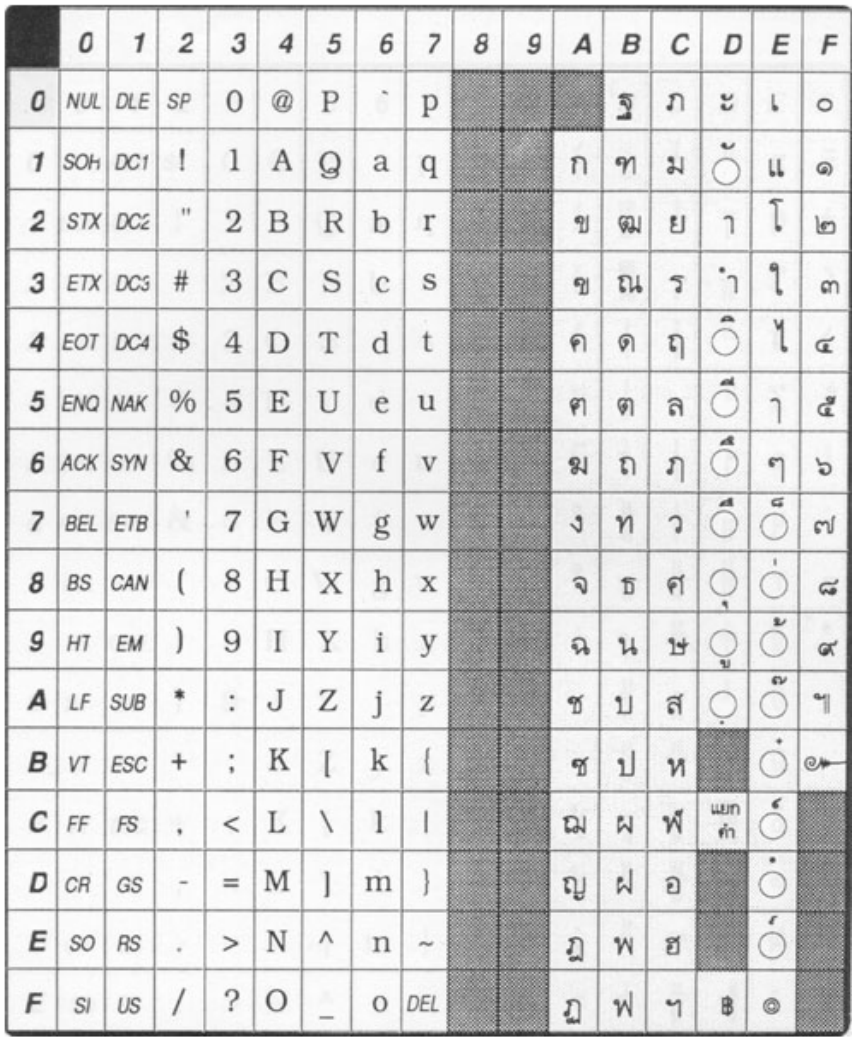
Thai belongs to the Tai language family of the Kadai or Kam-Tai family, latter arguably regarded, along with Austronesian, as a branch of Austro-Tai [1]. About 20 million people speak Thai in Thailand, where it also the official language of the country. Figure 1 shows the language tree for Thai [1].

Figure 1: Language Family Tree for Thai

Thai is written using the Thai script, which was derived from Brahmi scripts around 12th century AD. Thai script is known to have been influenced by Khmer script [2].

Character Set and Encoding

Thailand Industrial Standards Institute (TISI) develops standards in response to the government policy [3]. TISI has approved many standards for local language computing, e.g. for characters set, keyboard, encoding, etc. Before Unicode more than a score different Thai code pages, defined by local vendors, were in use. This led to the lack of interoperability between applications. To counter the problem TISI introduced a national encoding standard TIS 620-2529/1986, later upgraded in TIS 620-2533/1990 [4, 5]. TIS 620-2533/1990 was used as a basis for IBM code page 874 (cp-874), Microsoft code page 874 (windows-874) and Apple Thai (MacThai) [4]. Eventually, Unicode character set was introduced, which is now widely used. The encoding for Thai ranges from 0E00-0E7F in Unicode. Figure 2 shows Thai code chart for TIS 620-2533 [6]. A comparison of TIS 620 and Unicode is presented in [7]. In 1999, the international standard ISO/IEC 8859-11 Latin/Thai characters was also reactivated but not accepted [8]. Also see [8] for a comprehensive coverage of historical development of standards and [14] for a complete list of standards. Microsoft has been using another encoding standard for Thai [31].



	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		◌๋	เ◌า
C	FF	FS	,	<	L	\	l	|		ฌ	ผ	ฬ	เ◌ ◌่	◌์	
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	๏	๚	

Figure 2: Thai Character Code Chart TIS 620-2533 [7]

Fonts and Rendering

Microsoft Platform

Microsoft provides rendering support for Thai, ships Thai fonts with its Thai version of Windows and Office [11], and also provides guidelines for Thai font development [12]. Figure 3 shows rendering results of some Thai fonts [9]. Many more Thai fonts are available through other organizations (e.g. [10]).

สวัสดี	Angsana
ส วั ส ดี	Courier Mono Thai
ส วั ส ดี	Courier Proportional Thai

Figure 3: Thai Fonts on Microsoft Platform [9]

Linux Platform

Thai rendering is also supported on Linux and open source platforms, e.g. in Pango, the rendering engine of GNOME, as shown in Figure 4.

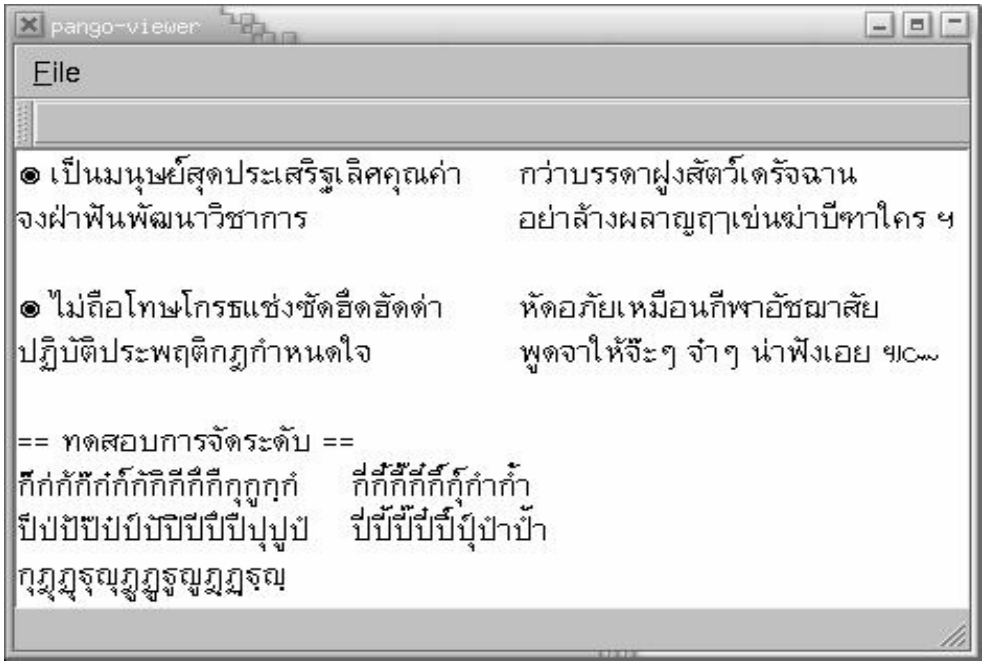


Figure 4: Thai Font Rendering on Pango

The following Thai fonts are supported on Linux. The encoding standard that has been used for the specific fonts is also given along with each font specification.

TIS-620 BDF fonts
Manop
Phaisarn
Yenbut
NECTEC
Type1 fonts
DearBook
Omega /NECTEC
ISO10646 BDF fonts
XFree86
TrueType fonts
Omega/NECTEC

Figure 5: Thai Fonts and Encodings

Keyboard

TIS 820-2531 was the initial national keyboard standard [13], later modified to TIS 820-2538. Both these layouts are based on Ketmanee layout [8].

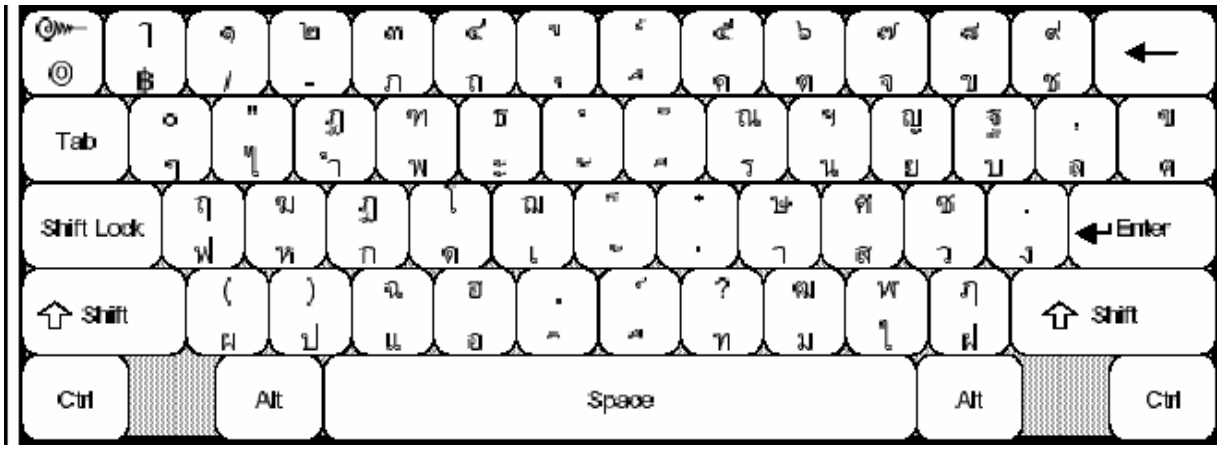


Figure 6: Standard Keyboard Layout TIS 820-2538 [8]

Thai language requires a complex input method, to use the keyboard and give adequate output. Though there is no official standard, there is now a single ad hoc solution supported by most vendors, e.g. Microsoft and Thai Language Environment (TLE) on Solaris. Details of this proposed standard, WTT 2.0, are given in [8].

Microsoft Platform

Microsoft Windows provides built-in support for Thai keyboard. Two different Thai keyboard layouts, Thai Ketmanee and Thai Pattachote, are available on Windows XP, shown in Figure 7.

(a)

(b)

(c)

(d)

Figure 7: Microsoft On-Screen Keyboard Layout for (a) Ketmanee (Normal Version), (b) Ketmanee (Shift Version), (c) Pattachote (Normal Version), and (d) Pattachote (Shift Version)

Linux Platform

Thai keyboard is available in Linux Red Hat version 9 and many other Thai Linux distributions. Ketmanee keymap has been supported as the standard keyboard layout in Thai Linux distributions.

Collation

Thai collation standard is defined, and is based on Thai Royal Institute Dictionary 2525 B.E. edition, the official Thai dictionary. Thai encoding standards TIS 620 is based on this dictionary.

Thai collation rules based on this collation order are also defined [8, 15, 16]. Thai collation is available on Linux platform, and also works on Microsoft platform.

Locale

Work started in 1990's for definition of Thai locale, which is now available and supported in Microsoft and Linux platforms [see 17]. Thai locale (th_TH) is completely supported in the IBM ICU locale data repository [18]. Both the Gregorian calendar and the Thai official Buddhist calendar are supported within the locale definition. In addition, the date format (long, medium, short), number formats, currency symbol, and weight and measures specific to the Thai conventions have been defined in the IBM ICU locale [18]. Thai locale is also available within CLDR 1.3. Also see [28, 29, 30] for further details.

Microsoft Platform

Microsoft provides support for Thai locale in Windows XP editions. If the system locale is switched to Thai, changes in date, time, and currency symbol of all application of Microsoft are displayed. This is shown in Figure 8.

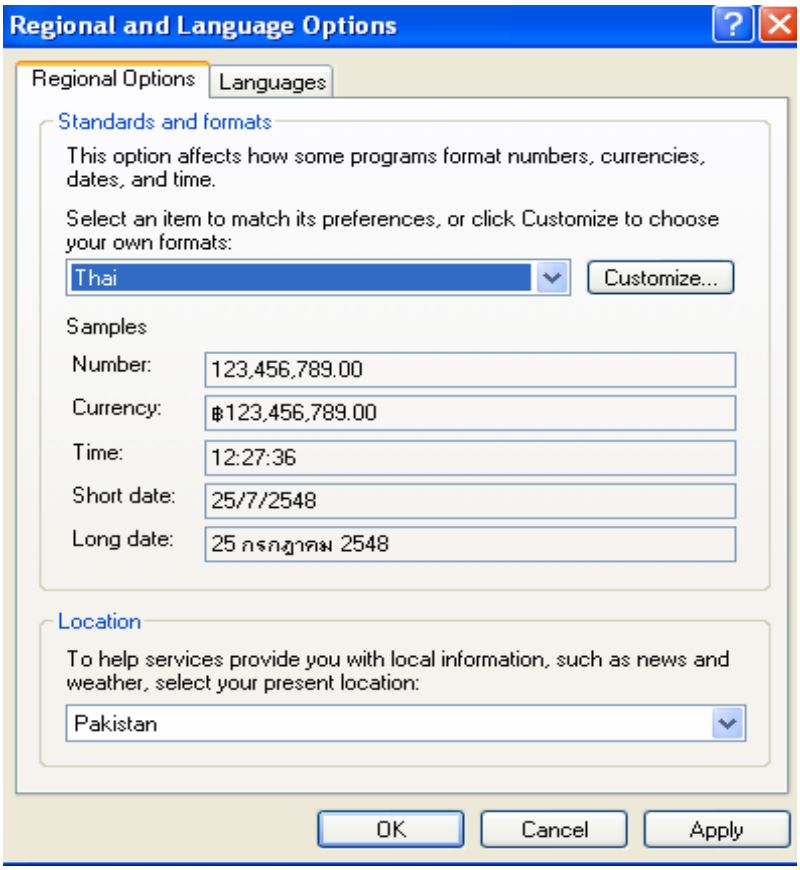


Figure 8: Thai Locale Settings on Microsoft Windows

Linux Platform

Thai locale is available on multiple platforms within Linux, including definition in Open Office and GNOME platforms [4, 19, 20].

Interface Terminology Translation

Microsoft Platform

Localized versions of Microsoft Windows XP and Microsoft Office are available [11]. Proofing tools for the MUI pack of Office 2003 are also available to facilitate advanced desktop processing in Microsoft platform.

Linux Platform

Multiple Linux distributions are available. These are Kaiwal Linux, Linux School Internet Server (Linux SIS) with Thai Language Extension (Linux-TLE), and Burapha Linux [4]. In addition, the latest version of GNOME 2.12 and Firefox have been released, and work is also in progress on KDE [20]. Thai version of Open Office is also available [19]. Figure 9 below shows Thai Linux developed by the Thai Linux Working Group (TLWG).

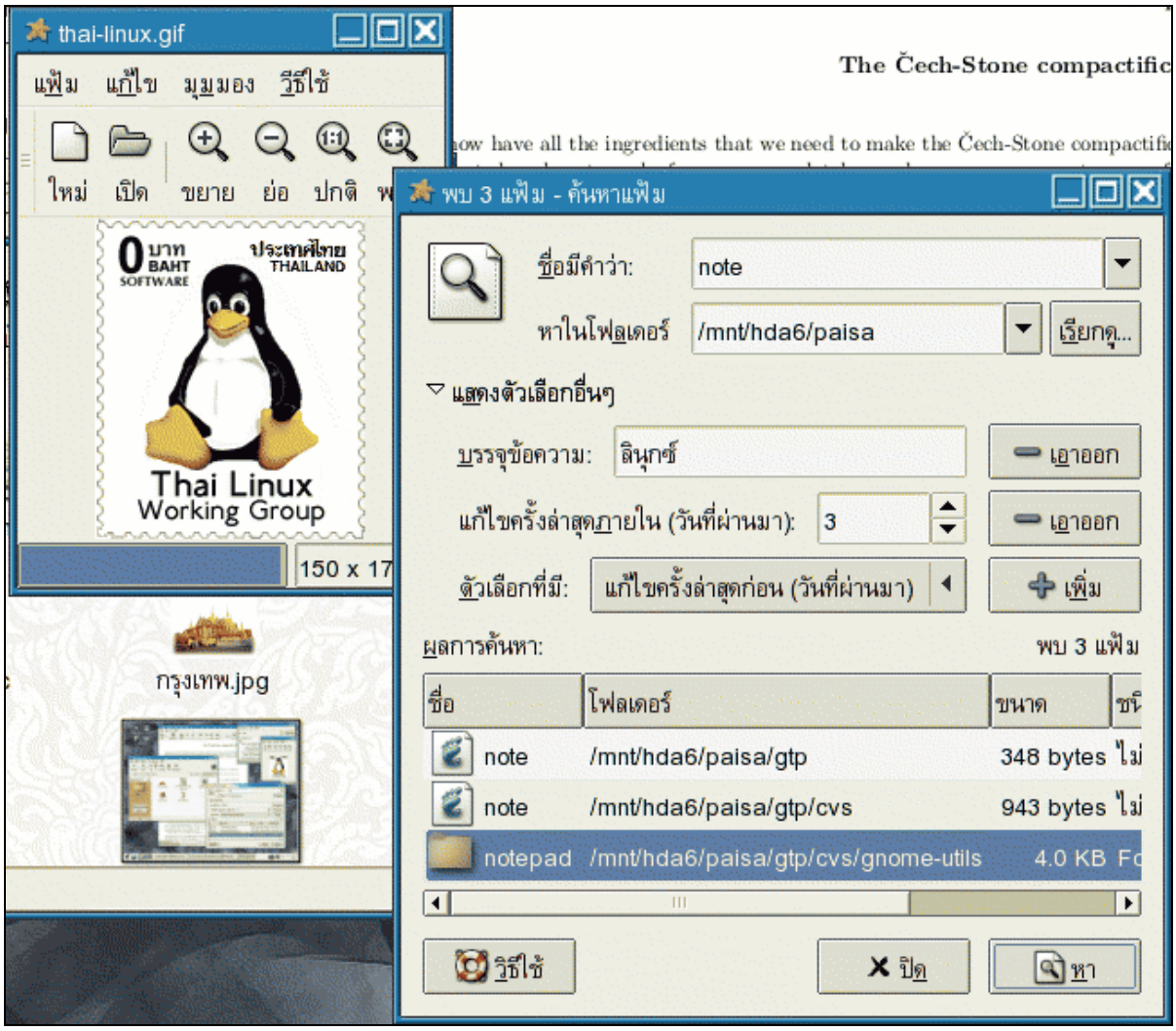


Figure 9: Thai Linux Distribution Developed by Thai Linux Working Group (TLWG) [20]

Status of Advanced Applications

Thailand has significantly progressed in Thai language processing, with R&D at NECTEC, the statutory government organization under Ministry of Science and Technology. The Center has established R&D cell to develop the fundamental resources such as sorting, line breaking and word breaking, lexicons and font development to support use of Thai language in software and operating systems.

NECTEC has developed the following projects (also see [23]):

- Khian Thai and Khat Thai, two Thai word processors
- Thai Sorting Program
- Thai word segmentation and line breaking
- Orchid, Thai text corpus [21] and POS annotated corpus
- Thai part-of-speech tagging program based on Orchid
- Lexitron, [22], an online English-Thai-English dictionary which is based on the Royal Institute Electronic dictionary also developed by NECTEC.
- Initial work on thesaurus and grammar checker for Thai has also been done
- ARN Thai, a Thai optical character recognition application [24]
- Parsit, an English-to-Thai machine translation system [25]
- Vaja, a Thai text to speech system [26]
- Sansarn, a Thai-English search engine [27]

Significant more work continues on further maturing these applications at NECTEC, universities and organizations across Thailand.

References

[1] http://www.ethnologue.com/show_language.asp?code=tha
[2] http://www.omniglot.com/writing/thai.htm
[3] http://www.tisi.go.th/eng/tisi.html
[4] Karoonboonyanan, T. and Koanantakool, T. "Standardization Activities and Open Source Movements in Thailand." http://www.nectec.or.th/it-standards/mlit99/mlit99-country.html
[5] http://www.nectec.or.th/it-standards/std620/std620.html
[6] http://www.nectec.or.th/it-standards/mlit97/country/gii2.htm
[7] Koanantakool, T., Tanprasert, C. and Viravan, C. "Country Report – Thailand." In the Proceedings of International Symposium on Standardization of Multilingual Information Technology, Singapore. http://mozart.inet.co.th/cyberclub/trin/thairef/tis620-iso10646.html, 1997
[8] Karoonboonyanan, T. "Standardization and Implementations of Thai Language." http://www.nectec.or.th/it-standards/thaistd.pdf.
[9] http://www.into-asia.com/thai_language/thaifont/?PHPSESSID=a9df4eab71d3e863f06aa8dc17f89641
[10] http://www.travelphrases.info/gallery/fonts_Thai.html
[11] Windows XP Thai LIP, http://www.microsoft.com/downloads/details.aspx?displaylang=th&FamilyID=0db2e8f9-79c4-4625-a07a-0cc1b341be7c
[12] http://www.microsoft.com/typography/otfntdev/thaiot/default.htm
[13] http://www.nectec.or.th/users/htk/it-standard/TISKB551.gif
[14] http://www.nectec.or.th/it-standards/
[15] http://linux.thai.net/~thep/
[16] http://www.open-std.org/jtc1/sc22/wg20/docs/n668.pdf
[17] http://linux.thai.net/~thep/th-locale/
[18] http://www-306.ibm.com/software/globalization/topics/thai/locale.jsp
[19] Open TLE, http://www.opentle.org/
[20] http://linux.thai.net/

[21] Orchid Thai Corpus, http://www.links.nectec.or.th/orchid/
[22] Lexitron Thai Dictionary, http://lexitron.nectec.or.th/index.php
[23] http://www.links.nectec.or.th/web_service.php
[24] ARN Thai OCR System, http://arnthai.nectec.or.th/arnthai.htm
[25] Parsit, http://suparsit.com/index1.php
[26] Vaja, Thai Text-to-Speech, http://vaja.nectec.or.th/onlinedemo/
[27] Sansarn, Thai Search Engine , http://sansarn.com/
[28] http://software.thai.net/locale/locale/14651/n537e.pdf
[29] ISO/IEC 14652 Cultural Convention Specification,
http://software.thai.net/locale/locale/14652/14652fcd.doc
[30] ISO/IEC 15435 Internationalization APIs, http://software.thai.net/locale/locale/15435/n536.pdf
[31] http://www.microsoft.com/globaldev/reference/wincp.mspx