# Computers and the Thai Language

**Hugh Thaweesak Koanantakool**
*National Science and Technology Development Agency*

**Theppitak Karoonboonyanan**
*Thai Linux Working Group*

**Chai Wutiwiwatchai**
*National Electronics and Computer Technology Center*

This article explains the history of Thai language development for computers, examining such factors as the language, script, and writing system, among others. The article also analyzes characteristics of Thai characters and I/O methods, and addresses key issues involved in Thai text processing. Finally, the article reports on language processing research and provides detailed information on Thai language resources.

Thai is the official language of Thailand. The Thai script system has been used for Thai, Pali, and Sanskrit languages in Buddhist texts all over the country. Standard Thai is used in all schools in Thailand, and most dialects of Thai use the same script.

Thai is the language of 65 million people, and has a number of regional dialects, such as Northeastern Thai (or Isan; 15 million people), Northern Thai (or Kam Meuang or Lanna; 6 million people), Southern Thai (5 million people), Khorat Thai (400,000 people), and many more variations (http://en.wikipedia.org/wiki/Thai_language). Thai language is considered a member of the family of *Tai* languages, the language used in many parts of the Indochina subregion including India, southern China, northern Myanmar, Laos, Thai, Cambodia, and North Vietnam.

The Thai script of today has a history going back about 700 years, with gradual changes in the script's shape and writing system evolving over the years. The script was originally derived from the Khmer script in the sixth century. It is generally thought that the Khmer script developed from the Pallava script of India.[1]

Thai is written left-to-right, without spaces between words. Each character has only one form, that is, no notion of uppercase and lowercase characters. Some vowels are written before and after the main consonant. Certain vowels, all tone marks, and diacritics are written above and below the main character.

Pronunciation of Thai words does not change with their usage, as each word has a fixed tone. Changing the tone of a syllable may lead to a totally different meaning. Thai verbs do not change their forms as with tense, gender, and singular or plural form, as is the case in European languages. Instead, there are other additional words to help with the meaning for tense, gender, and singular or plural. Basic Thai words are typically monosyllabic. Contemporary Thai makes extensive use of adapted Pali, Sanskrit, English, and Chinese words embedded in day-to-day vocabulary. Some words have been in use long enough that people have forgotten that they originated from other languages. At present, most new words are created from English.

A Thai word is typically formed by the combination of one or more consonants, one vowel, one tone mark, and one or more final consonants to make one syllable. Certain words may be polysyllabic and therefore they may consist of many characters in combination. Because of a limited number of characters (see Figure 1), the writing system can be implemented on typewriters by mapping each symbol directly to each key. A typing sequence is exactly the same as a writing sequence.

Published by the IEEE Computer Society 1058-6180/09/$25.00 © 2009 IEEE

| | | |
|---|---|---|
| Consonants | 44 | ก ข ฃ ค ฅ ฆ ง<br>จ ฉ ช ซ ฌ ญ<br>ฎ ฏ ฐ ฑ ฒ ณ<br>ด ต ถ ท ธ น<br>บ ป ผ ฝ พ ฟ ภ ม<br>ย ร ล ว ศ ษ ส<br>ห ฬ อ ฮ |
| Vowels | 18 | ะ ◌ั า ◌ำ ◌ิ ◌ี ◌ึ ◌ื ◌ุ ◌ู<br>เ แ โ ใ ไ ๅ ฤ ฦ |
| Tone marks | 4 | ◌่ ◌้ ◌๊ ◌๋ |
| Diacritics | 5 | ◌ฺ ◌็ ◌์ ◌ํ ◌๎ |
| Numerals | 10 | ๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙ |
| Other symbols | 6 | ฯ ฿ ๆ ๏ ๚ ๛ |
| Total | 87 | |

**Figure 1. Thai characters.**

The computerization of Thai script for the Thai language preserves the input sequence, and assigns one character in storage to correspond to one input keystroke. This concept is simple to understand, and all language processing algorithms are designed to suit this arrangement, despite the fact that some Thai vowels consist of up to three characters (from the vowel group) in that word.

Early IT standards in Thailand defined the code points for computer handling, keyboard layout, and I/O method. Subsequent research projects created many useful functions such as word break, text-to-speech, optical character recognition (OCR), voice recognition, machine translation, and search algorithms, which we describe elsewhere in this article.

## Thai language on typewriters and computers

Printing of the Thai language was first accomplished at Serampore in 1819: according to professor Michael Winship,[2] a *Catechism of Religion* had been prepared by the American Baptist missionary Ann Hasseltine Judson for distribution to a small community of Thai prisoners of war in Burma. By 1836, Dan Beach Bradley, a missionary physician, started printing in Thailand using a donated press by the American Board of Commissioners for Foreign Mission. At the time, publications were intended primarily for spreading Christianity into Thailand.[3]

Thai-language typewriter localization was first shown to the Thai public in 1891, when Edwin McFarland, in cooperation with Smith Premier, a US typewriter company, introduced the first Thai language typewriter. In 1898, son George McFarland opened a typewriter shop in Bangkok. Those first typewriters had a moving carriage without any shift mechanism.[4] Today's typewriter keyboard has five rows, with standard shift keys.

The use of computers with the Thai language started in the late 1960s when IBM introduced card-punch machines and line printers with Thai character capability. In doing so, the 8-bit EBCDIC code assignments were made for Thai characters.[5] With the printer's mechanical limitations, for each line of Thai text an equivalent of four-pass printing was required to complete the multilevel nature of the Thai writing system. In the 1970s, Univac, Control Data, and Wang introduced their computers using the Thai language. In 1979, some companies offered CRT terminals, with a 12-lines-per-screen capability, which could display Thai characters. Interactive computing came to Thailand in 1983 when a Thai inventor modified the Hercules Graphic Card (HGC) circuitry and made it capable of displaying 25 lines per screen in text mode. The full display capability of the Thai language has been adopted widely since then.

The computer industry in Thailand in the 1980s was very much vertically integrated: that is, hardware manufacturers always supported their version of localized operating systems and applications. Every company used different Thai character codes as a result of competition and corporate strategy to retain their customers. Application developers, however, suffered from the codes' incompatibility and from differences in the system interfaces. By 1984, there were at least 20 different character coding conventions used in Thailand.[6] This situation prompted the Thai Industrial Standard Institute (TISI) to establish a standards committee to draft the national standard code for computers. The work, completed in 1986, is known as TIS 620-2529—The Standard for Thai Character Codes for Computers B.E. [Buddhist Era] 2529.[7] Basically, two designs of the code points were adopted in the standard: the IBM-EBCDIC and the extended ASCII (8-bit plane). In 1990, the TIS 620 standard was enhanced with a clearer explanation but without any change to the code points, and this standard became TIS 620-2533. The technical committees of TISI subsequently issued a number of IT standards for Thailand (see Table 1).

**Table 1. Thai industrial standards on information technology. The four digits at the end of the standard number (e.g., 2529) is the year of issue, according to Buddhist Era.**

| Standard number | Standard title |
|---|---|
| TIS 620-2529 (1986) | Standard for Thai Character Codes for Computers |
| TIS 820-2531 (1988) | Layout of Thai Character Keys on Computer Keyboards |
| TIS 620-2533 (1990) | Standard for Thai Character Codes for Computers (Enhanced) |
| TIS 988-2533 (1990) | Recommendation for Thai Combined Character Codes and Symbols for Line Graphics for Dot-Matrix Printers |
| TIS 1074-2535 (1992) | Standard for 6-Bit Teletype Codes |
| TIS 1075-2535 (1992) | Standard for Conversion Between Computer Codes and 6-Bit Teletype Codes |
| TIS 1099-2535 (1992) | Standard for Province Identification Codes for Data Interchange |
| TIS 1111-2535 (1992) | Standard for Representation of Dates and Times |
| TIS 820-2538 (1995) | Layout of Thai Character Keys on Computer Keyboards (Enhanced) |
| TIS 1566-2541 (1988) | Thai Input/Output Methods for Computers |

Source: http://www.nectec.or.th/it-standards/.

It took only one year after the standard's announcement for every computer company to be fully compliant with the TIS 620 standard code. Everyone's data then became interchangeable.

In 1990, the Thai API Consortium, a group of Thailand software developers led by Thaweesak Koanantakool, drafted a common specification for computer handling of the Thai I/O method. The work was funded by the National Electronics and Computer Technology Center (Nectec) and was published in 1991[8] as the WTT 2.0 specification.[9] The specification was widely used by industry, including IBM, Digital Equipment, Hewlett-Packard, and, later, Microsoft. WTT 2.0 was proposed to TISI as a draft national standard and, seven years later, became TIS 1566-2541 (1998). At the same time, many more interesting, natural language processing (NLP) projects with useful results and solutions found their way into industrial use. In the past two decades, Thailand has achieved several milestones for advanced computer processing with the Thai language; we report on most of them in this article.

## Thai alphabets

Encoding of Thai characters into an octet has been made simple, since we need only 87 characters to represent the language. In the code space between 0×80 and 0×FF, TIS 620 occupies only the code space between 0×A1 and 0×FB. The ASCII-compatible version of TIS 620 was interoperable with any ASCII-based computer with true 8-bit storage per character. The actual placement on the code table was placed in the ''most acceptable'' collating sequence. The TIS 620 sequence was subsequently registered with the Universal Character Set defined by the international standard ISO/IEC 10646, *Universal Multiple-Octet Coded Character Set,* in the region U+0E00…U+0E7F, and also as the ISO/IEC 8859-11 standard for 8-bit (single byte) coded graphic character sets, Latin/Thai alphabet. Figure 2 shows the code assignment for Thai characters and their names in TIS 620 and Unicode.

### *Consonants*

Classifications of the Thai consonants can be made as plosives (stops), non-plosives, sibilants, and voiced ''h.'' In the plosive table in Figure 3, each column is also grouped by voiced/unvoiced properties. In the last group, อ is classified as a zero consonant, and it can be used to write with vowels, which cannot be stand-alone. Figure 3 shows the consonants, classifications, and the associated International Phonetic Alphabet (created by the International Phonetic Association, IPA) and the International Alphabet of Sanskrit Transliteration (IAST) representations.

### *Vowels*

The 18 symbols for vowels, together with three consonants—ย, ว, and อ—are used in combination to create 32 vowels for Thai.

The 18 symbols are listed in Figure 4, with the associated Unicode values.

These 18 vowel symbols and 3 consonants (21 in total) combine to form 32 vowels, as Figure 5 shows.

| TIS | 0xAx | 0xBx | 0xCx | 0xDx | 0xEx | 0xFx | | |
|---|---|---|---|---|---|---|---|---|
| Unicode | U+0E0x | U+0E1x | U+0E2x | U+0E3x | U+0E4x | U+0E5x | U+0E6x | U+0E7x |
| 0 | | ฐ<br>tho than | ภ<br>pho samphao | ะ<br>sara a | เ<br>sara e | ๐<br>thai zero | | |
| 1 | ก<br>ko kai | ฑ<br>tho nangmontho | ม<br>mo ma | ั<br>mai han-akat | แ<br>sara ae | ๑<br>thai one | | |
| 2 | ข<br>kho khai | ฒ<br>tho phuthao | ย<br>yo yak | า<br>sara aa | โ<br>sara o | ๒<br>thai two | | |
| 3 | ฃ<br>kho khuat | ณ<br>no nen | ร<br>ro rua | ำ<br>sara am | ใ<br>sara ai maimuan | ๓<br>thai three | | |
| 4 | ค<br>kho khwai | ด<br>do dek | ฤ<br>ru | ิ<br>sara i | ไ<br>sara ai maimalai | ๔<br>thai four | | |
| 5 | ฅ<br>kho khon | ต<br>to tao | ล<br>lo ling | ี<br>sara ii | ๅ<br>lakkhangyao | ๕<br>thai five | | |
| 6 | ฆ<br>kho rakhang | ถ<br>tho thung | ฦ<br>lu | ึ<br>sara ue | ๆ<br>maiyamok | ๖<br>thai six | | |
| 7 | ง<br>ngo ngu | ท<br>tho thahan | ว<br>wo waen | ื<br>sara uee | ็<br>maitaikhu | ๗<br>thai seven | | |
| 8 | จ<br>cho chan | ธ<br>tho thong | ศ<br>so sala | ุ<br>sara u | ่<br>mai ek | ๘<br>thai eight | | |
| 9 | ฉ<br>cho ching | น<br>no nu | ษ<br>so rusi | ู<br>sara uu | ้<br>mai tho | ๙<br>thai nine | | |
| A | ช<br>cho chang | บ<br>bo baimai | ส<br>so sua | ฺ<br>pinthu | ๊<br>mai tri | ๚<br>angkhankhu | | |
| B | ซ<br>so so | ป<br>po pla | ห<br>ho hip | | ๋<br>mai chattawa | ๛<br>khomut | | |
| C | ฌ<br>cho choe | ผ<br>pho phung | ฬ<br>lo chula | | ์<br>thanthakhat | | | |
| D | ญ<br>yo ying | ฝ<br>fo fa | อ<br>o ang | | ํ<br>nikhahit | | | |
| E | ฎ<br>do chada | พ<br>pho phan | ฮ<br>ho nokhuk | | ๎<br>yamakkan | | | |
| F | ฏ<br>to patak | ฟ<br>fo fan | ฯ<br>paiyannoi | ฿<br>baht | ๏<br>fongman | | | |

**Figure 2. Standard codes for Thai characters—TIS 620 and Unicode, with their names as defined by the WTT 2.0 specification.**

## Thai input method and keyboards

Thai is written with the combining marks stacked above or below the base consonant, like diacritics in European languages. However, although the concepts are quite similar, the implementations are significantly different.

### *Thai input method requirements*

First, there are too many possible combinations of base consonants and combining marks in Thai to be enumerated like Latin accents in the ISO/IEC 8859 series standard for 8-bit character encoding. Therefore, base characters and combining characters are encoded separately, rather than precombined.

Second, Thai combining marks are classified into upper or lower vowels, tone marks, and other diacritics. Moreover, a Thai base consonant can be combined with up to two combining marks, that is, zero or one upper or lower vowel and zero or one tone mark or

| Plosive class | Unvoiced | | voiced | | |
|---|---|---|---|---|---|
| | unaspirated | aspirated | unaspirated | aspirated | nasal |
| velar | ก /kà/ [ka] | ข /kʰà/ [kha] | ค /kʰá/ [ga] | ฆ /kʰá/ [gha] | ง /ŋá/ [ṅa] |
| palatal | จ /tɕà/ [ca] | ฉ /tɕʰà/ [cha] | ช /tɕʰá/ [ja] | ฌ /tɕʰá/ jha] | ญ /já/ [ña] |
| retroflex | (ฎ /dà/ [da])<br>ฏ /tà/ [ṭa] | ฐ /tʰà/ [ṭha] | ฑ /tʰá/ [ḍa] | ฒ /tʰá/ ḍha] | ณ /ná/ [ṇa] |
| dental | (ด /dà/ [da])<br>ต /tà/ [ta] | ถ /tʰà/ [tha] | ท /tʰá/ [da] | ธ /tʰá/ [dha] | น /ná/ [na] |
| labial | (บ /bà/ [ba])<br>ป /pà/ [pa] | ผ /pʰà/ pha]<br>(ฝ /fà/ [fa]) | พ /pʰá/ [ba]<br>(ฟ /fá/ [fa]) | ภ /pʰá/ [bha] | ม /má/ [ma] |
| *tone* | *middle* | *high* | *low* | *low* | *low* |

| | palatal | retroflex | dental | labial |
|---|---|---|---|---|
| non-plosive<br>(semi-vowel) | ย /já/ [ya] | ร /rá/ [ra] | ล /lá/ [la]<br>(ฬ /lá/ [ḷa]) | ว /wá/ [wá] |
| sibilants | ศ /sà/ [śa] | ษ /sà/ [ṣa] | ส /sà/ [sa] | |

| | | |
|---|---|---|
| Voiced "h" | ห /hà/ [ha] | ฮ /há/ [ha] |
| zero consonant (never a vowel) | | อ /ʔ/ |

**Figure 3. Classification of Thai consonants. Each cell consists of the character, International Phonetic Alphabet, and the International Alphabet of Sanskrit Transliteration (IAST) representations in square brackets.**

| Vowel | Name | Type | Position relative to the main consonant |
|---|---|---|---|
| ะ | sara a | FV1 | follow |
| ั | mai han-akat | AV2 | above |
| า | sara aa | FV1 | follow |
| ำ | sara am | FV1 | follow |
| ิ | sara i | AV1 | above |
| ี | sara ii | AV3 | above |
| ึ | sara ue | AV2 | above |
| ื | sara uee | AV3 | above |
| ุ | sara ue | BV1 | below |
| ู | sara uu | BV2 | below |
| เ | sara e | LV | leading vowel |
| แ | sara ae | LV | leading vowel |
| โ | sara o | LV | leading vowel |
| ใ | sara ai maimuan | LV | leading vowel |
| ไ | sara ai maimalai | LV | leading vowel |
| ฤ | ru | FV3 | follow |
| ฦ | lu | FV3 | follow |
| ๅ | lakkhangyao | FV2 | follow |

**Figure 4. Eighteen vowel symbols in Thai.**

diacritic. The upper/lower vowel, if present, is always attached to the consonant before the tone/diacritic. These conditions must be governed by some rules to limit the number and order of the combining characters to be put after the base consonant.

As a third difference, Thai users are familiar with typing the combining characters *after* the base consonant, in contrast to many European input methods in which users type the diacritic *before* the base letter to compose an accent. The typing sequence and the internal storage of a Thai string, therefore, are totally consistent. Moreover, a typing sequence is almost exactly the same as a handwriting sequence, with the exception of only one character, *sara am*, a composite vowel that consists of two symbols sharing one code point.

|  | Front |  | Back |  |  |  |
|---|---|---|---|---|---|---|
|  | Unrounded |  | Unrounded |  | Rounded |  |
|  | Short | Long | Short | Long | Short | Long |
| Close | -ิ /i/ | -ี /iː/ | -ึ /ɯ/ | -ื /ɯː/ | -ุ /u/ | -ู /uː/ |
| Close-mid | เ-ะ /e/ | เ- /eː/ | เ-อะ /ɤ/ | เ-อ /ɤː/ | โ-ะ /o/ | โ- /oː/ |
| Open-mid | แ-ะ /ɛ/ | แ- /ɛː/ | – | – | เ-าะ /ɔ/ | -อ /ɔː/ |
| Open | – | – | -ะ, -ั /a/ | -า /aː/ | – | – |

(a) 18 Monophthongs

| Thai | -ัวะ | -ัว | เ-ียะ | เ-ีย | เ-ือะ | เ-ือ | ใ- | ไ- | เ-า |
|---|---|---|---|---|---|---|---|---|---|
| IPA | /ua/ | /uːa/ | /ia/ | /iːa/ | /ɯa/ | /ɯːa/ | /aj/ | /aj/ | /aw/ |

(b) 9 Diphthongs

| Thai | -ำ | ฤ | ฤๅ | ฦ | ฦๅ |
|---|---|---|---|---|---|
| IPA | /am/ | /rɯ/, /ri/, /rɤː/ | /rɯː/ | /lɯ/ | /lɯː/ |

(c) 5 Semi-vowels

**Figure 5. Thirty-two vowels in Thai: 18 monophthongs, 9 diphthongs, and 5 semi-vowels.**

### *Typewriter keyboards*

The keyboard of the first Thai typewriter made by Smith Premier in 1891 consisted of 12 keys in seven rows (Figure 6a shows the layout). When the typewriter industry started using the shift key, the Thai layout was also adapted on the typical QWERTY arrangement. The characters that normally appear above and below the base characters are assigned to individual keys, and they appear above or below the previous position without moving the carriage. This concept is called *dead keys*.

The two main differences between Thai and English typewriters are that there are two different characters on the same key, and that there are eight dead keys. All the dead keys are grouped in the center of the Ketmanee keyboard layout (see Figure 6b), the classic layout named after its inventor.

In 1966, Sarit Pattajoti,[11] an engineer at the Royal Irrigation Department, redesigned the layout to improve the mechanical coordination of human fingers. Statistically, the new layout would eliminate the imbalance of the Ketmanee layout (30% distributed to the left hand and 70% to the right hand), and distribute more of the workload to the index and middle fingers. The Pattajoti layout (see Figure 6c) became the official standard. Manufacturers were encouraged to make them in quantity for government offices. However, this ''official'' standard was abandoned in 1971 because its marginal speed improvement was not significant enough for people to migrate from the de facto Ketmanee.

In 1988, TISI issued the standard layout for the computer keyboard, TIS 820-2531 (1988) (see Figure 6d). The TISI technical committee adopted the Ketmanee layout with only one change. In 1993, the standard was enhanced to utilize the additional three keys available on the computer keyboards. Figure 6e shows the present standard layout as defined by TIS 820-2536 (1993). Four rarely used characters, however, are defined in the TIS 620 standard that are not on the keyboard; they would be entered on the specific program that uses them, typically through a GUI or, in the pre-Microsoft Windows era, via direct hexadecimal code in DOS.

*WTT 2.0 canonical order*

The Thai I/O method specifications described in WTT 2.0 are based on the TIS 620 standard code.[8] The same concept was later extended to the Thai code page in Unicode. TIS 620 defines a code point for each of the Thai symbols, irrespective of their positions when rendered. In the same manner as Unicode *normalization,* the WTT 2.0 specification defines the *canonical order* of Thai character strings. It differs, however, in the implementation details and in terms of syntactic strictness. WTT 2.0 compliance requires that certain input-sequence rules must be met, and most (if not all) input syntactic errors are eliminated at the time of data entry (see Figure 7).

First, WTT 2.0 requires strings to be stored and transmitted in canonical order, not to be normalized at processing time. Therefore, in WTT 2.0, canonical and noncanonical strings mean different things and can yield different results in rendering, sorting, searching, and so on. This helps simplify the string handling routines to some degree, leaving the task of ensuring the order at a single point, the input method.

Second, WTT 2.0 concerns syntactic correctness of the input string. It inhibits multiple tone marks to combine in one cell, while Unicode does not care. Moreover, WTT 2.0 defines three levels of syntactic strictness of input method. Level 0 (Passthrough) does not filter at all. Level 1 (BasicCheck) just ensures that the input sequence complies with canonical order and can be displayed gracefully in the output method. Level 2 (Strict) is more picky in filtering out most problematic sequences. Therefore, the WTT 2.0 input conditioning helps make the storage order of Thai text cleaner than would the raw Unicode concept. The WTT 2.0 input method produces strings of a proper subset of the Unicode specification, and thus it does not negatively affect Unicode-compliant programs. WTT 2.0 eliminates problematic strings that might otherwise fail string matching or that might confuse the output method.

*Technical characteristics of Thai input method*

Technical requirements of the Thai TIS 1566-2541 (WTT 2.0) input method differ from input methods of other languages. The input method

- does not need pre-edit-and-commit stages (as used in some CJK [Chinese, Japanese, and Korean] input methods),

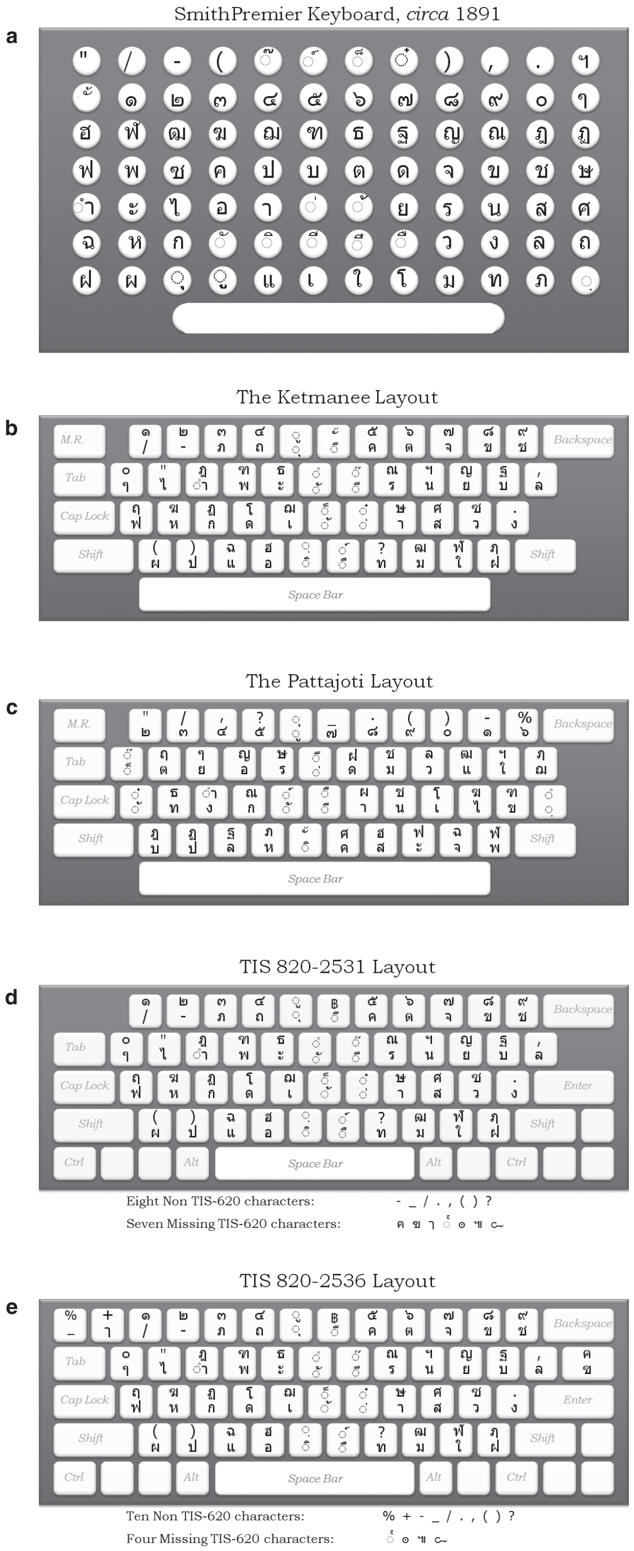

SmithPremier Keyboard, *circa* 1891
a
The Ketmanee Layout
b
M.R.
Backspace
Tab
Cap Lock
Shift
Space Bar
The Pattajoti Layout
c
TIS 820-2531 Layout
d
Enter
Ctrl
Alt
Eight Non TIS-620 characters: - _ / . , ( ) ?
Seven Missing TIS-620 characters: ฅ ฃ ๅ ๎ ๏ ๚ ๛
TIS 820-2536 Layout
e
Ten Non TIS-620 characters: % + - _ / . , ( ) ?
Four Missing TIS-620 characters: ๎ ๏ ๚ ๛

**Figure 6. Evolution of Thai typewriter keyboard layout: from top, (a) Smith Premier layout, ca. 1890, (b) Ketmanee layout (no date), and computer keyboard layouts: (c) Pattajoti layout (1966), (d) TIS 820-2531 (1988), and (e) TIS 820-2536 (1993). (Source: Koanantakool[10])**

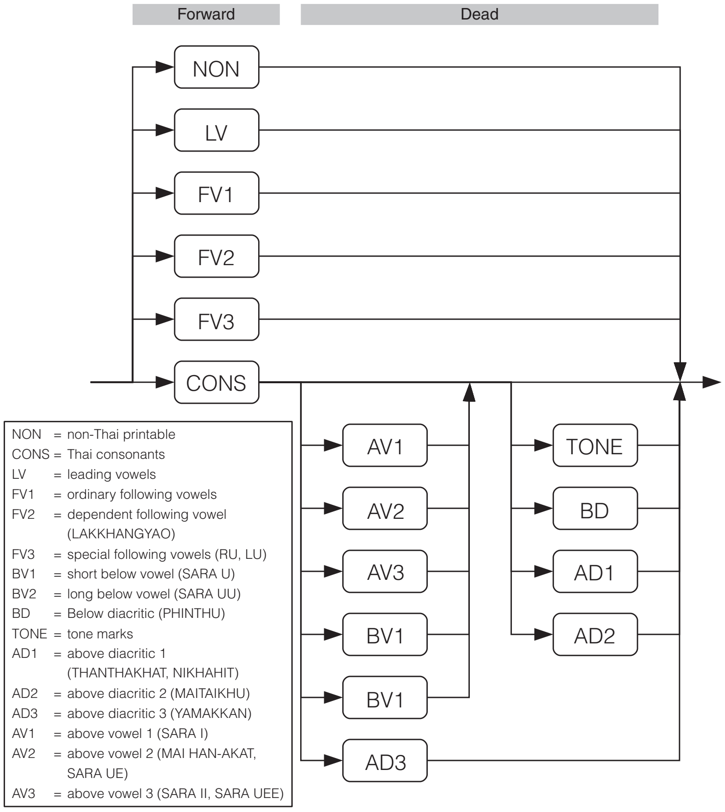

Forward
Dead
NON
LV
FV1
FV2
FV3
CONS
AV1
AV2
AV3
BV1
BV1
AD3
TONE
BD
AD1
AD2
NON = non-Thai printable
CONS = Thai consonants
LV = leading vowels
FV1 = ordinary following vowels
FV2 = dependent following vowel (LAKKHANGYAO)
FV3 = special following vowels (RU, LU)
BV1 = short below vowel (SARA U)
BV2 = long below vowel (SARA UU)
BD = Below diacritic (PHINTHU)
TONE = tone marks
AD1 = above diacritic 1 (THANTHAKHAT, NIKHAHIT)
AD2 = above diacritic 2 (MAITAIKHU)
AD3 = above diacritic 3 (YAMAKKAN)
AV1 = above vowel 1 (SARA I)
AV2 = above vowel 2 (MAI HAN-AKAT, SARA UE)
AV3 = above vowel 3 (SARA II, SARA UEE)

**Figure 7. WTT 2.0 Level 1 I/O state machine.**

- does not use the composing method (as used in Europe), and
- is not a straight key-to-character mapping (as used in ordinary English input method).

Rather, the input method requires the following:

- one-to-one key-to-character mapping (irrespective of the width and position of the character),
- ability to retrieve context character from application input buffer (to verify validity), and
- (optional) write access to application input buffer.

The key-to-character mapping is straightforward. Thailand has an industrial standard for a keyboard map, totally compatible with the traditional (i.e., mechanical) typewriter.

The ability to retrieve context characters from the application input buffer is intended to validate the key events, according to the state machine. Note that other solutions are not adequate, such as these:

- using pre-edit string and committing validated characters as a chunk,
- using composing method to commit strings cell-by-cell, and
- remembering previously typed characters in the input context memory.

These solutions can serve only for the inputting of new text but cannot handle the editing of existing text, especially when the first key to be input is a combining character.

The last requirement of write access to the application input buffer is for input sequence correction enhancement.

*Implementations*

Known implementations of WTT-based Thai input methods include the following:

- Thai Language Environment for Solaris
- Microsoft DOS 6.0 and Windows (all versions)
- X Window: XIM (embedded in Xlib)
- GTK+:
  —Thai-Lao IM module (stock GTK+ IM module)
  —gtk-im-libthai (third party, based on libthai)
- SCIM:
  —scim-thai (third party, based on libthai)
  —scim-m17n (third party, by ETL, Japan)

*Making TIS 620 a part of international standards*

The early 1990s witnessed different opinions concerning the encoding of the Thai block in the making of draft Basic Multilingual Plane (BMP), the code space between 0×0000 and 0×FFFF in Unicode. Trin Tantsetthi, a Thai computer architecture and software expert, developed proposals to relevant standards bodies that were responsible for the internationalization of Thai character codes.[12] In order to have a Thai language character set for the ISO 8859 series, the TIS 620 standard was registered under ISO 2375 by the ECMA (European Computer Manufacturers Association) as ISO-IR-166. The process encountered many obstacles. According to Tantsetthi, the first proposal was to eliminate nonspacing characters, which were nonoptional atomic parts of the script, by precomposing characters into a rectangular bounding box of glyphs. This proposal was later dropped because there were an insufficient number of code spaces to accommodate all precomposed glyphs. Once the industry had learned about this limitation, the long-awaited ISO/IEC 8859-11 Latin/Thai standard was approved. TIS 620 was formally registered by the Internet Assigned Numbers

Authority (IANA) as a legitimate encoding for the Internet in 1998.[12]

Thailand also encountered another problem in defining Thai in the ISO 10646 standard, as some nonnative linguists had strongly proposed the use of *phonetic order* for Thai, following Hindi-based scripts. Although this scheme allows words to be normalized into a "proper form," making word parsing a bit easier, the proposal could not address the following problems:

- it could not handle language exceptions vigorously;
- compounded vowels were not addressed, neither in code-point assignments nor in input method; and
- no backward compatibility was provided.

We opted for the traditional input method (i.e., visual, instead of phonetic, order) because, since Thai localization had begun in 1968, by 1990 the Thai computer industry had equipped itself with libraries and tools to use visual order without a problem. That's basically how the Thai block in BMP/ Unicode has developed. In 1998, IANA (Internet Assigned Numbers Authority) adopted TIS 620 as a legitimate character set on the Internet. The visual order was formally announced as a national standard in TIS 1566-2541 (1998).

### Computer display, printing, and Thai word processors

Like most languages, modern Thai text display on computers has evolved from traditional printing technologies. Mechanical typewriter design has allocated four vertical zones for placing Thai characters: one for base characters, one for combining characters below the baseline, and the other two for combining characters above the base character. Characters are strictly classified on the basis of their assigned positions. This has propagated to a text-mode computer display grid. The WTT 2.0 specification was designed to accommodate exactly this.

Thai typography has a more-refined scheme than that of typewriters. The topmost combining characters can be shifted down slightly when the combining character below them is missing to eliminate the empty gap. In addition, the combining characters can be slightly biased to the left or right of the base characters, which have long ascenders (stems) to deliver aesthetic print results.

The bias technique was necessary for hot-metal typesetting because all the combined symbols had to be molded individually as one piece, and there were about 28 combinations to be implemented. Another set of 28 types precombined with ascenders was also required. In phototypesetting or computer typesetting, these precombination techniques were no longer a concern, but the same concept is useful to improve the speed of dot matrix printing. To print a line of Thai text on an impact printer normally requires four passes. By precombining the top characters (two levels) before printing, the speed improves by 25% because only three passes are required, rather than the four that mechanical typesetting required—one for each of the four vertical zones. To make all dot matrix printers interoperable, TISI issued TIS 988-2533, the standard code for combined characters in dot matrix printers, in 1990.

Thai WTT-2.0–based rendering engines for text-mode display usually handle Thai text in two steps. The text string is first clustered into cells, which are then shaped by proper placements of the combining characters. Printing with an impact printer, on the other hand, requires a conversion of one line of text into three or four strings whose lengths are equal to the line width, with each string corresponding to each pass of printing. The fastest dot matrix printer manages to print a Thai line with only two passes; typical printing requires three passes.

A number of DOS-based Thai-language I/O subsystems were developed at Kasetsart University, Chulalongkorn University, and Thammasat University. Yuen Poovorawan and his team at Kasetsart University created the first Thai word processor called Thai Easy Writer. He also developed the Thai Kernel System as a resident program for DOS to handle the Thai language.[13] The most popular DOS-based Thai word processor was released in 1989. According to professor Wanchai Rivepiboon of Chulalongkorn, CU Writer version 1.1 was first released to the public in April 1989. It was a tremendous success because of the demands for quality printing of Thai documents. The Faculty of Engineering of Chulalongkorn University continued to support and improve the program for about five years until it was superseded in 1994 by commercial word processing software running on the Windows platform.[14]

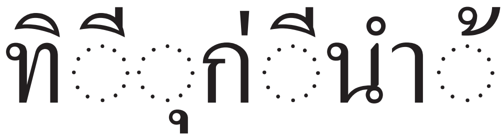


**Figure 8. WTT 2.0 clustering. The dotted circles indicate a rejected sequence.**

*Cell clustering*

Cell clustering shares the same state table with the input method. Composable sequences are tokenized into cells. Rejected sequences are separated, with an optional dotted circle as the base character that can be easily caught by human eyesight, as Figure 8 illustrates.

*Shaping*

For typewriter-style displays, the tokenized cells are rendered in a perfect rectangular bounding box, and all combining marks are placed at fixed vertical positions. But for modern desktop publishing and the Thai language, beginning in 1989, combining marks have been typographically adjusted for the following issues:

- The topmost combining mark without a mark below it is shifted down to eliminate the gap between it and the base character.
- Combining marks that are combined with the base character with long ascenders (stems) are biased slightly to the left to avoid overlapping.
- Consonants with removable descenders (*yo ying* and *tho than*) have the descender removed when combined with a lower vowel or diacritic.

To accommodate these issues, typesetters have used Private Use Area glyphs. PUA glyphs are still widely used in most fonts today. OpenType may eventually make their use obsolete with Thai, although not in the near term.

*PUA glyphs solutions*

Operating system vendors assigned separate, predefined PUA code points to shifted glyphs. Font creators prepared the required glyphs, and the rendering engine knew how to use them. Unfortunately, Microsoft and Apple defined their own PUA code points differently, which meant fonts could not be used across Windows and Mac platforms.

To make matters worse, Adobe did not support any PUA glyphs for Thai. This situation caused developers to create some common hacks with automatic filters to encode text with PUA glyphs so that such applications would render properly. This situation created another mess: when users of such filters transmitted messages in PUA, they were unreadable by anyone using another platform.

Thai users have long suffered from this lack of interoperability. The availability of PUA glyphs, however, is important enough in Thai typography to live on.

Pango, a free software rendering engine, supports both sets of PUA glyphs—Windows and Apple—by detecting them before using. Open standard and open source software has significantly helped the local Thai industry to find useful typographical solutions.

*OpenType solutions*

With OpenType technology, all shaping details can be moved into fonts. All that the rendering engines need do is call the appropriate features for the language.

Glyph processing features in OpenType fonts are represented in two forms: glyph substitution (GSUB) and glyph positioning (GPOS). GSUB contains substitution rules, while GPOS describes anchor points for attaching combining marks to base glyph or combining marks to other combining marks.

According to Microsoft's specification for Thai OpenType fonts,[15] several features are required, as explained next.

*Character Composition/Decomposition (''ccmp'').* This must contain GSUB rules to

1. select the lower variation for topmost marks in the absence of an upper vowel (see Figure 9, left),
2. remove the descender for the consonants *yo ying* and *tho than* when combined with below-base characters (see Figure 9, middle), and
3. decompose the composite vowel *sara am* into the *nikhahit* sign and *sara aa* vowel, plus rearrange them with the combined tone mark if one is present (see Figure 9, right).

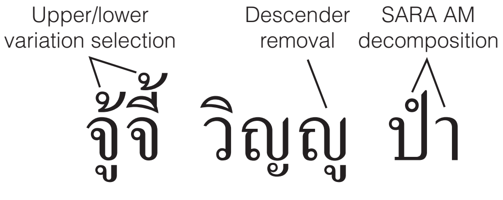


**Figure 9. Glyph substitutions for shaping.**

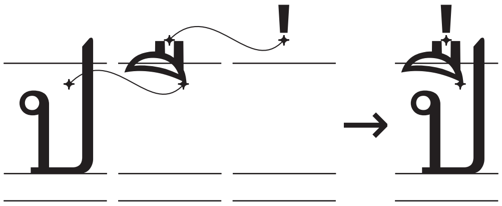


Figure 10. Mark positioning with GPOS.

*Mark to Base Positioning (''mark'').* This is composed of GPOS base anchors in the base characters and mark anchors in the combining marks. This feature places marks above or below the base characters (see Figure 10).

*Mark to Mark Positioning (''mkmk'').* This is composed of GPOS base anchors in the base marks and mark anchors in the combining marks. This feature places topmost marks on upper vowels.

## Thai cultural conventions

The most basic kind of cultural convention is the Posix locale for the standard C library. According to Posix, a number of C functions are defined to be locale-dependent, such as date and time format, string collation, and so on. Many other programming languages also rely on these functions.

The Posix locale specification has been further extended in ISO/IEC TR 14652, *Specification Method for Cultural Conventions,* by adding more categories, which have now been supported by more-modern software.

Theppitak Karoonboonyanan[16] has gathered information on Thai cultural conventions and documented his Thai locale creation for the GNU C library. Most Posix categories were defined and later extended when ISO/IEC TR 14652 was supported.

## Handling of Thai words

According to the Royal Institute of Thailand's *Royal Institute Dictionary* (http://en.wikipedia.org/wiki/The_Royal_Institute_of_Thailand#The_Royal_Institute_Dictionary), Thai strings can be collated by comparison from left to right, with two exceptions:

- Leading vowels are less significant than the initial consonant of the syllable and must be compared after the consonant.
- Tone marks, *thanthakhat* and *maitaikhu,* are ignored unless all other parts are equal.

No syllable structure or word boundary analysis is required. Thai words are sorted alphabetically, not phonetically. Figure 11 shows an example of a sorted list of Thai words.

กก
กรรม
กรรม์
-กระแย่ง
กราบ
กะเกณฑ์
กัก
ก้าว
กำ
กิน
กี่
กึ๋น
กุม
กูด
เก้ง
เกล้า
เกลียว
เกศ
เก้า
เกาทัณฑ์
เกาะ
เกี่ยว
เกี๊ยะ
เกือก
แกง
แกะ
โกน
โกร๋น
ใกล้
ไก่
ไกล
ขั้น
ขนาบ
ข้าง
ข้าง ๆ
ข้างขึ้น
ข้างควาย
ข้าง ๆ คู ๆ
ข้างแรม
ข้างออก
เขน
เข็น
เข่น
แข่ง
แข้ง
แข้งขวา
แข็งขัน
แข่งขัน
แขน
ครรภ-
ครรภ์
จุมพล
จุํพล
ชาย
เฒ่า
เณร
ตลาด
ทูลเกล้า
ทูลเกล้าฯ
ทูลเกล้าทูลกระหม่อม
บุญหลง
บุญ-หลง
ปา
ป่า
ป้า
ป๊า
ป๋า
ปาน
พญา
ฯพณฯ
พณิชย์
รอง
ฤทธิ์
ฤษี
ฤๅษี
ลิง
ฦๅชา
วก
ศาลา
หริภุญชัย
หฤทัย
หลง
แหง่
แห่ง
แหน
แหนม
แหนหวง
แหบ
อาน
ฮา

Figure 11. Example of a sorted list of Thai words.

### String collation solutions

The first known solution for Thai string collation was proposed in 1969 by D. Londe and Udom Warotamasikkhadit,[17] and later referenced by Vichit Lorchirachoonkul[18] in 1979. Londe and Warotamasikkhadit proposed an algorithm for converting Thai strings into left-to-right comparable form.

In 1992, Samphan Khamthaidee (a pen name used by Samphan Raruenrom)[19] published an article in a computer hobbyist journal describing an algorithm for comparing two Thai strings on the fly. The algorithm was well crafted so that the strings are scanned from left to right in a single pass, and the difference is detected as early as possible. Karoonboonyanan[20] later extended Khamthaidee's algorithm to cover punctuation marks by generalizing it to accommodate extra levels of character weights in 1997. Subsequently, details about the order of Thai punctuation marks have been summarized in the literature.[21,22]

### International standards

String collation is based primarily on the Posix standard library functions `strcoll()` and `strxfrm()`. Both are based on the `LC_COLLATE` locale setting. The Posix locale was later extended as ISO/IEC TR 14652, *Specification Method for Cultural Conventions*. For `LC_COLLATE` in particular, a Common Template Table for collating strings encoded in ISO/IEC

10646, *Universal Multiple-Octet Coded Character Set,* in general has been defined as ISO/IEC 14651, *International String Ordering and Comparison.* Every locale can customize the table to fit local needs. Unicode also defines UTS #10 (the Unicode Collation Algorithm) in parallel with the ISO/IEC 14651 standard.

Both standards already supported multiple-level character weights in the first place, which was important for handling Thai tone marks and diacritics. So, all the specification needed at that time was the proper weights definition. The reordering of the leading vowels, however, had been controversial for a while, as we will explain.

Defending the non-Indic style of Thai Unicode encoding had given the Unicode committee sufficient information to address Thai in UTS#10 since the beginning. The reordering was defined as a separate preprocessing stage. It took time, however, to convince the ISO/IEC 14651 committee to include this in the Common Template Table (CTT) as a set of predefined contraction forms as there was no preprocessing concept in it. Rather, the reordering requirement was first described in Annex C in DIS 14651:2000[23] without realization in the CTT, before the actual amendment was implemented in 2003.[24]

### Leading-vowel rearrangement issue

One common misinterpretation of the leading-vowels rearrangement by nonnative linguists, which has been frequently encountered at conferences and meetings, was that Thai string collation required linguistic analysis, as a consequence of a Thai *visual encoding* scheme, as opposed to a *phonetic encoding* scheme of other Indic scripts. This was simply wrong.

A major ambiguity problem would result if one tried to do linguistic analysis. For example, the Thai string เพลา can mean two different words. One is a two-syllable word *phe-la* (meaning "time"), and the other is a one-syllable word *phlao* (meaning "axel" or "abate"). In trying to preprocess the string by rearranging the leading vowel เ with the syllable initial sound, one will find two possible rearrangements: พเลา for the former case, and พลเา for the latter. This makes it impossible to determine a single weight for the string. This is another reason why Thailand completely rejected phonetic encoding in its standards.

| Sample text | Romanization | IPA | SAMPA |
|---|---|---|---|
| สวัสดี | sawatdi | sa wat diː | sa_2 wat_2 di:_1 |
| ขอบคุณ | khopkhun | kʰɔːp kʰun | khOp_2 khun_1 |
| ลาก่อน | lakon | laː kɔːn | la:_1 kOn_2 |

**Figure 12. Examples of Thai sound representations.**

Thai string collation is in fact as simple as its encoding scheme. The leading vowel only needs to be swapped with its immediate succeeding character—nothing more than that.

### Word and syllable boundary

Typical Thai writing has no space between words and syllables. We use space only to separate sentences. Lacking a word or syllable boundary marker has been a major problem for early Thai language processing applications, which has been researched since the 1980s. Early works focused on syllable segmentation due to its potential for being modeled in rule-based fashion.[25,26] Word segmentation, on the other hand, requires a dictionary. Longest or maximal matching incorporated with a well-designed dictionary was the first practical approach.[27] Over time, researchers have proposed advanced algorithms. A major contribution includes the first open-source software for word segmentation called *S*mart *W*ord *A*nalyzer for *TH*ai (SWATH), which NECTEC developed.[28] It is based on a statistical part-of-speech *n*-gram. In 2002, Wirote Aroonmanakun constructed a word segmentation engine using two-pass processing, syllable segmentation, and syllable merging based on collocations.[29] These fundamental tools have been applied in various language processing tasks such as line breaking and word boundary detection in word processing and publishing software.

### Soundex

*Soundex* is a phonetic algorithm for indexing words by sound. It has been widely used in Internet search engines for flexible searching by the same or a similar sound without regard to the written form. An original algorithm, proposed by Vichit Lorchirachoonkul,[30] was based on a set of rules used to segment Thai character strings into syllables, simplify the syllables, and encode them in a five-character code. Another general procedure is to find pronunciations of a given keyword and match them to indexed words in the search database. Standard sound representations are defined by the International Phonetic Association (IPA) and the Speech Assessment Methods Phonetic Alphabet (SAMPA).[31,32] The Thai Royal Institute has defined a standard sound-based romanization of Thai words. Figure 12 demonstrates these standard sound representations.

Automatic letter-to-sound conversion is an important tool that activates the automatic soundex search. One of the first complete tools for letter-to-sound conversion was based on a probabilistic, generalized left-to-right parser trained by a pronunciation dictionary.[33]

## Advanced human–computer I/O

Several forms of human–computer I/O—OCR, handwriting recognition, text-to-speech and speech synthesis, and automatic speech recognition—present different challenges with respect to the Thai language. This section reviews major contributions to these advanced technologies.

### *OCR and handwriting recognition*

Thai OCR is not trivial, as a bounding box may contain up to three symbols stacking together, and the written (or printed) sentences do not have spaces within them. In addition, modern Thai texts do mix with English words. The challenges to OCR developers have been tremendous.

The history of Thai OCR spans well over 20 years. Pipat Hiranvanichakorn et al. and Chom Kimpan et al. were two of the original groups of researchers in this area.[34,35] Their subsequent works primarily concerned recognition of individual (isolated) printed characters, that is, without connections among the characters. A number of commercial products have been launched since the 1990s—for example, *ArnThai* developed by NECTEC was one of the first complete OCR software applications that helped users convert scanned pictures of documents to editable documents, with over 90% conversion accuracy under restricted scanning conditions.[36]

The technology trend has been tailored to the use of machine learning such as an artificial neural network. Although OCR technology has reached a commercial level, its mass utilization has not yet been achieved. Problems include connected characters, which often appear in dirty documents; poorly scanned documents; or documents with rarely used fonts that cause recognition errors. Research on Thai OCR is ongoing: for example, an introduction of postprocessing to recover recognition errors.

Thai handwriting recognition was explored about a decade after OCR. Academic research started also on isolated-character recognition.[37] Current applicable systems still mostly focus on isolated characters with a neural network as a basic classifier, like those of OCR. Some systems have been adopted commercially in mobile devices.[38] On this simple but accurate task, characters are grouped into Thai alphabets, English alphabets, and digits. Modification on character shapes is applied to enlarge the difference of some characters or to simplify the writing.

### *Text-to-speech and speech synthesis*

Research on Thai speech synthesis, chiefly for academic purposes, began in the 1980s. Early work was reported by Sudaporn Luksaneeyanawin.[39] Text-to-speech synthesis (TTS) has resulted in several practical products produced around 1990—for example, CUTalk by Luksaneeyanawin and Vaja by NECTEC.[40] Major efforts required for Thai TTS are *text processing,* where a given text is segmented, normalized, and converted to sound representatives; *prosody prediction* including estimating unit durations, filled pauses, and fundamental frequencies; and *synthesis*, including the speech database and synthesizing algorithms.

The technology for speech synthesis has been incorporated by adopting available TTS engines in applications such as the IBM homepage reader,[41] screen reader, and telephone voice response services. Limitations on speech quality and intelligibility inhibit widespread use. The desire to provide TTS applications on portable devices is increasing, and therefore researchers will have to optimize the size of the speech database and text processing dictionary. Thai TTS has been used widely, however, for persons with visual disabilities in Thailand.[42]

### *Automatic speech recognition*

Development of Thai automatic speech recognition (ASR) began with isolated word recognition (in which speakers utter only a short command, not a sentence) in the mid-1990s, primarily at Chulalongkorn University.[43] A few years ago, the research expanded to applications of continuous speech in limited tasks: for example, email access and telebanking. Several attempts were made on large vocabulary continuous speech recognition (LVCSR) for Thai; Table 2 summarizes the results. The lack of a very large speech database has slowed the advance of Thai speech recognition; however, NECTEC is one of the organizations that has continuously developed Thai speech recognition resources available for research, such as the LOTUS corpus.[44]

Table 2. State-of-the-art Thai large vocabulary continuous speech recognition (LVCSR) performance. The organization that provided the research results is listed after each "Task."

| Task | Vocabulary size (words) | Perplexity | Word error rate (%) |
|---|---|---|---|
| Newspaper reading (Carnegie Mellon University)[45] | ~7,400 | 140 | 14.0 |
| Newspaper reading (Tokyo Institute of Technology)[46] | 5,000 | 140 | 11.6 |

## Search and Semantic Web

Increasingly, as more companies and organizations in Thailand, both public and private, start adopting digital solutions in place of their original paper-based workflow, the need for search technology becomes inevitable. Most of the available software tools for developing information retrieval (IR) systems have not been specifically designed to work with Thai. Consequently, there are two possible approaches for indexing Thai texts, including *suffix array* and *inverted file index*. The former approach treats text as a string of characters and records character positions; the latter approach adopts the conventional word-based paradigm applied for Latin-based languages. Word segmentation is therefore necessary in the latter approach. The stemming process is, however, inapplicable since Thai words are considered noninflectional.

An early publication describing full-text search for Thai was published in 1999 by Pradit Mitrapiyanuruk et al.[47] In 2000, Surapan Meknavin, a co-author of that publication, focused his research on founding an Internet service business called SiamGuru (http://www.siamguru.com), which could be recognized as the first Thai-specific search engine. Today, the research trend in Thai-language search engines is moving toward the natural-language and semantic searches for implementing question-answering (QA) systems.

The Semantic Web research initiatives in Thailand can be classified into four major categories:

- information extraction (IE) and ontology learning,
- metadata/ontology standards and tools,
- knowledge representation (KR) and inferences, and
- applications and services.

A wide spectrum of techniques has been researched and developed since the early 2000s. For example, a Thai translation of the Dublin Core metadata element sets[48] was completed by Science and Technology Knowledge Services (STKS), which also promotes its uses in Thailand, and the XML Declarative Description (XDD)[49] language was proposed by the Asian Institute of Technology to extend the capabilities of XML and RDF to support rule-based reasoning.

## Language resources and industrial standard lists

One of the most important language resources is a dictionary. *So Sethaputra* is considered the first Thai-English dictionary made into a commercial electronic dictionary

Table 3. Available Thai dictionaries.

| Dictionary | Type | Size (words) | Availability | Source |
|---|---|---|---|---|
| *Royal Institute Dictionary* | Monolingual with pronunciation | 33,582 | Web application | http://rirs3.royin.go.th/dictionary.asp |
| NECTEC *LEXiTRON* | Bilingual (En-Th) with pronunciation | 53,000 English,<br>35,000 Thai | Publicly available | http://lexitron.nectec.or.th/ |
| *KDictThai* | Bilingual (En-Th) | 37,018 | Publicly available | http://kdictthai.sourceforge.net/ |
| *Saikam Dictionary* | Bilingual (Jp-Th) | 133,524 | Web application | http://saikam.nii.ac.jp/ |
| *Longdo Dictionary* | Multilingual (En-Th, Th-En, De-Th, Th-De, Jp-Th Th-Jp, Fr-Th, Th-Fr) | >600,000 | Web, gadget, widget | http://dict.longdo.com/ |

**Table 4. Thai text and speech resources.**

| Corpus (Organization) | Type | Purpose |
|---|---|---|
| Orchid (NECTEC)[51]<br>http://www.hlt.nectec.or.th/orchid/ | Text | Annotated 568,316 words of Thai junior encyclopedias and NECTEC technical papers |
| LOTUS (NECTEC)[44] | Speech | Well-designed speech utterances for 5,000-word dictation systems |

product. *LEXiTRON* is another electronic dictionary with a long history.[50] The first version of *LEXiTRON* was launched by NECTEC in both standalone and Web-based systems in 1995. This first version contained approximately 13,000 Thai words and 9,000 English words. Through user collaboration, the number of entries has been more than doubled in the current version, *LEXiTRON 2.2*. Table 3 lists other available dictionaries. Another language resource necessary for language processing is text and speech corpora. Orchid and LOTUS, shown in Table 4, are considered two of the first official language corpora available publicly for Thai language research.

### Concluding remarks

The history of the Thai language on computers dates back to about 1965, but local research involvement began around 1975, when low-cost microcomputers became available. Since the first standard Thai code for computers was developed in 1986, the computer processing speed has improved from 8-bit, 8-MHz to 32-bit, 3.6-GHz (1,800 times), and the RAM size for an entry-level computer has increased from 256 Kbytes to 1 Gbyte (3,900 times). NLP in real time is becoming a reality. Research opportunities are open to any students and researchers who can afford these inexpensive computers. Thailand's I18N (internationalization) and L10N (localization) processes owe much to open standards and many open source projects. It is anticipated that openness in the international IT community will enable the Thai language to be more usable on any computers in the future.

With more IT users, improvement in human-computer interaction for the Thai language will become more desirable. Because computers are also becoming indispensable for the elderly, persons with disabilities, and those who are illiterate, many new developments will be possible. Speech I/O, handwriting input, sign languages, mobile-phone text entry, a voice-command system, speech-to-speech translation, voice search, and so on are the most likely efforts to be developed and commercialized.

## Acknowledgments

The authors would like to thank Trin Tantsetthi and Yuen Poovorawan for their personal communications, and to Choochart Haruechaiyasak and Marut Buranarach from NECTEC for their initial survey on search engines and Semantic Web technology.

## References and Notes

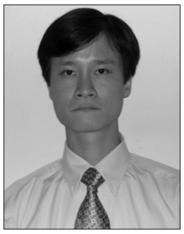
1. P. Soonthornpoct, *From Freedom to Hell: A History of Foreign Interventions in Cambodian Politics and Wars,* Vantage Press, 2006.
2. M. Winship, ''The Printing Press as an Agent of Change? Early missionary printing in Thailand''; http://www.historycooperative.org/journals/cp/vol-08/no-02/tales/.
3. NECTEC National Font Committee, *The Time Line of Thai Printing,* Thai National Font Book (in Thai), National Electronics and Computer Technology Center (NECTEC), Bangkok, 2001.
4. Antique Phonograph and Gramophone Thai Society, ''The Origins of Thai Typewriter''; http://www.talkingmachine.org/smithpremierorigin.html.
5. K. Hensch, ''IBM History of Far Eastern Languages in Computing, Part 1: Requirements and Initial Phonetic Product Solutions in the 1960s,'' *IEEE Annals of the History of Computing,* vol. 27, no. 1, 2005, pp. 17-26.
6. T. Koanantakool, ''Compilation of all Thai Character Codes for Computers,'' *Microcomputer Magazine,* vol. 1, no. 9, Aug. 1984.
7. National Electronics and Computer Technology Center, *Thai Information Technology Standards;* http://www.nectec.or.th/it-standards/.
8. T. Koanantakool and the Thai API Consortium, *Computer Gub Pasa Thai* [Computers and the Thai Language], NECTEC, Oct. 1991 (in Thai).
9. The WTT specification was published in the book *Computers and the Thai Language* in 1991. WTT, a project code name, is a Thai acronym of ''Wing Took Thee,'' meaning ''I run everywhere.''
10. T. Koanantakool, ''The Keyboard Layouts and Input Method of the Thai Language,'' *Proc. Symp. Natural Language Processing,* Chulalongkorn Univ., 1993.
11. S. Pattajoti, *The Evolution of the Typewriter,* National Research Council, The Office of the Prime Minister, 4 Nov. 1966 (in Thai).

12. T. Tantsetthi, ''Thai Software Alert #1: Legitimate Thai Charset on the Internet''; http://software.thai.net/alerts/tis-620/index.html.
13. P. Suriyawong, Y. Poovarawan, and C. Wongchaisuwat, ''A Development of Thai Kernel System,'' *Proc. Regional Workshop on Computer Processing of Asian Languages* (CPAL), 26-28 Sept. 1989.
14. T. Koanantakool, ''A Brief History of ICT in Thailand''; http://www.bangkokpost.com/20th_database/07Feb2007_data00.php.
15. Microsoft, ''Thai OpenType Specification: Creating and Supporting OpenType fonts for the Thai Script''; http://www.microsoft.com/typography/otfntdev/thaiot/default.htm.
16. T. Karoonboonyanan, ''Thai Locale''; http://linux.thai.net/~thep/th-locale/.
17. D. Londe and U. Warotamasikkhadit, ''Computerized Alphabetization of Thai, tech. memo TM-BA-1000/000/01, System Development Corp., 1969.
18. V. Lorchirachoonkul, *Thai Sorting Algorithm, Thai Algorithm: Design, Analysis and Implementation,* tech. report, NIDA and Electrical Generation Authority of Thailand, Bangkok, 1979.
19. S. Khamthaidee, ''Thai Style Algorithm: Sorting Thai,'' *Computer Rev.,* vol. 91, Mar. 1992.
20. T. Karoonboonyanan, ''Thai Sorting Algorithms''; http://linux.thai.net/~thep/tsort.html.
21. T. Karoonboonyanan, S. Raruenrom, and P. Boonma, ''Thai-English Bilingual Sorting''; http://linux.thai.net/~thep/blsort_utf.html.
22. T. Karoonboonyanan, S. Raruenrom, and P. Boonma, ''Sorting Thai Words with Punctuation Marks,'' *Nectec J.,* vol. 14, Jan.–Feb. 1997, Nectec, 1997.
23. A. LaBonté, ed., *ISO/IEC DIS 14651: International String Ordering and Comparison—Method for Comparing Character Strings and Description of the Common Template Tailorable Ordering,* ISO/IEC JTC1/SC22/WG20 N731, 2000.
24. K. Karlsson, *Ordering Rules for Thai and Lao,* ISO/IEC JTC1/SC22/WG20 N1077, 2003.
25. Y. Thairatananond, ''Towards the Design of a Thai Text Syllable Analyzer,'' master's thesis, Asian Inst. of Technology, Pathumthani, 1981.
26. S. Charnyapornpong, ''A Thai Syllable Separation Algorithm,'' master's thesis, Asian Inst. of Technology, Pathumthani, 1983.
27. R. Varakulsiripunth et al., ''Word Segmentation in Thai Sentence by Longest Word Mapping,'' *Papers on Natural Language Processing: Multilingual Machine Translation and Related Topics (1987–1994),* Nectec, 1989.
28. S. Meknavin, P. Charoenpornsawat, and B. Kijsirikul, ''Feature-Based Thai Word Segmentation,'' *Proc. Natural Language Processing Pacific Rim Symp.* (NLPRS 97), 1997, pp. 41-48.
29. W. Aroonmanakun, ''Collocation and Thai Word Segmentation,'' *Proc. Joint Int'l Conf. Symp. Natural Language Processing* (SNLP) *and Oriental Chapter of the Int'l Committee for the Coordination and Standardization of Speech Databases and Assessment Techniques* (O-COCOSDA), 2002, pp. 68-75.
30. V. Lorchirachoonkul, ''A Thai Soundex System,'' *Information Processing and Management,* vol. 18, no. 5, 1982, pp. 243-255.
31. K. Tingsabadh and A.S. Abramson, *Illustrations of the IPA: Thai, Handbook of the International Phonetic Association,* Cambridge Univ. Press, 1999.
32. J.C. Wells, ''SAMPA for Thai''; http://www.phon.ucl.ac.uk/home/sampa/thai.htm.
33. P. Tarsaku, V. Sornlertlamvanich, and R. Thongprasirt, ''Thai Grapheme-to-Phoneme Using Probabilistic GLR Parser,'' *Proc. European Conf. Speech Communication and Technology* (EuroSpeech), ISCA, 2001, pp. 1057-1060.
34. P. Hiranvanichakorn, T. Agui, and M. Nkajima, ''A Recognition of Thai Characters,'' *Trans. IECE Japan,* vol. E 54, no. 12, 1982.
35. C. Kimpan et al., ''Thai Character Pattern Recognition Preliminary,'' *Proc. Conf. L-05,* College of Science and Technology, Nihon Univ., 1980.
36. Nectec, ''Open TLE''; http://www.opentle.org/.
37. V. Tanrudee, ''Thai Character Handwriting Recognition System,'' master's thesis, Chulalongkorn Univ., 1989.
38. G Softbiz Co., Ltd., ''Thai-G''; http://www.thai-g.com/.
39. S. Luksaneeyanawin, ''A Thai Text-to-Speech System,'' *Proc. Regional Workshop Computer Processing of Asian Languages* (CPAL), 1989, pp. 305-315.
40. P. Mittrapiyanurak et al., ''Issues in Thai Text-to-Speech Synthesis: The Nectec Approach,'' *Proc. Nectec Ann. Conf.,* Nectec, 2000, pp. 483-495.
41. IBM, ''IBM Launches IBM Home Page Reader V3.02 Thai, A Thai-speaking Web Browser—The 12th Language of the World,'' press release, 12 Feb. 2003.
42. Thailand Association of the Blind, ''Software for the Blind''; http://www.tab.or.th/.
43. T. Pathumthan, ''Thai Speech Recognition Using Syllable Units,'' master's thesis, Chulalongkorn Univ., 1987 (in Thai).
44. S. Kasuriya et al., ''Thai Speech Corpus for Speech Recognition,'' *Proc. Int'l Conf. Int'l Committee for the Coordination and Standardization of Speech Databases and Assessments* (O-COCOSDA), 2003, pp. 105-111.
45. S. Suebvisai et al., ''Thai Automatic Speech Recognition,'' *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP), IEEE Press, 2005, pp. 857-860.
46. I. Thienlikit, C. Wutiwiwatchai, and S. Furui, ''Language Model Construction for Thai LVCSR,'' *Reports of the Meeting of the Acoustic Soc. of Japan,* ASJ, 2004, pp. 131-132.

47. P. Mitrapiyanuruk et al., ''A Development of Full-Text Search Engine for Large Scale Thai Text Database,'' *Proc. Nat'l Science and Technology Development Agency* (NSTDA), annual meeting, 1999, pp. 246-257 (in Thai).
48. Science and Technology Knowledge Service (STKS), ''Dublin Core Metadata Element Set (Thai translation)''; http://dublin.stks.or.th/.
49. V. Wuwongse et al., ''XML Declarative Description: A Language for the Semantic Web,'' *IEEE Intelligent Systems,* vol. 16, no. 3, 2001, pp. 54-65.
50. NECTEC, ''LEXiTRON Online Dictionary''; http://lexitron.nectec.or.th/.
51. V. Sornlertlamvanich, T. Charoenporn, and H. Isahara, *ORCHID: Thai part-of-speech tagged corpus,* tech. report TR-NECTEC-1997-001, ISBN 9747576-98-8, NECTEC, 1997.

**Hugh Thaweesak Koanantakool** is a vice president of the National Science and Technology Development Agency. He received a PhD in digital communications from the Imperial College of Science and Technology, London University. His major works are in IT standards of Thailand, the establishment of the Internet in Thailand, SchoolNet Thailand, and e-Government. Koanantakool served on Thailand's National IT Committee, the parliamentary commissions on the Electronic Transactions Act of 2001 and the Computer-related Offences Act in 2007. Contact him at htk@nectec.or.th.

**Chai Wutiwiwatchai** received a PhD from Tokyo Institute of Technology in 2004. Currently he is the director of the Human Language Technology Laboratory in NECTEC, Thailand. His research interests include speech processing, natural language processing, and human–machine interaction. Contact him at chai.wutiwiwatchai@nectec.or.th.



**Theppitak Karoonboonyanan** received a B.Eng. (honors) in computer engineering from Chulalongkorn University in 1993. His interests are software engineering, and free and open source software philosophy. He has contributed to the GNOME, Debian, and other projects for Thai supporting infrastructure, and has maintained upstream projects for Thai resources. Contact him at thep@linux.thai.net.

**For further information on this or any other computing topic, please visit our Digital Library at http://computer.org/csdl.**

Security Code Review • The Road to Automatic Security
IEEE
SECURITY&PRIVACY
Building Confidence in a Networked World
WEB SECURITY
computer society
IEEE