1 Thai Language

Thai, spoken by almost 100 percent of the population, is the official national language of Thailand. It is a tonal, uninflected, and predominantly monosyllabic language. Most polysyllabic words in the vocabulary have been borrowed, mainly from Khmer, Pali or Sanskrit. [1]

The ancient Thai language itself belongs to the Tai language family, which includes language spoken in Assam, northern Myanmar, Thailand, Laos, northern Vietnam and the Chinese provinces of Yunnan, Guizhou and Guangxi. [2]

The oldest Thai script found in the history dated over 700 years ago, in Sukhothai age. Over years, Thai script has been gradually changed. The contemporary Thai alphabet comprises 44 consonants, 32 vowels, 4 tone marks, 2 diacritics, and 10 decimal digits. These cover only the minimal set of symbols to write Thai words. There are additional 3 symbols particularly used for writing Pali and Sanskrit words, and 6 typographical symbols.

Consonants

Thai consonants include 34 of 35 Sanskrit equivalences, plus 10 invented ones. Thai consonants are classified into three classes—namely, high, middle, and low consonants—which can affect the syllable tone when functioning as initial sound.

Table 1 is the consonant chart laid after the grammar of Sanskrit. The first 5 rows are classified according to their articulators. The last row is the remainder of the classification. The parenthesized consonants are invented by Thai. Within the first 5 rows, according to Thai grammar, the first column contains middle consonants, the second contains high consonants, the next two columns contain paired low consonants (low consonant with the same sound as the high consonant in the same row), and the last column contains the single low consonants, whose sounds are nasal.

<table>
<thead>
<tr>
<th>middle</th>
<th>high</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>paired</td>
<td>single</td>
</tr>
<tr>
<td>ก [k]</td>
<td>ข [kh]</td>
<td>ฃ [kh]</td>
</tr>
<tr>
<td>ก [k]</td>
<td>ฅ [kh]</td>
<td>ฆ [kh]</td>
</tr>
<tr>
<td>ฉ [tɕ]</td>
<td>ฎ [tɕh]</td>
<td>ฏ [tɕh]</td>
</tr>
<tr>
<td>ฉ [tɕ]</td>
<td>ฏ [tɕh]</td>
<td>ฐ [tɕh]</td>
</tr>
<tr>
<td>ฑ [d]</td>
<td>ฒ [d]</td>
<td>ฒ [d]</td>
</tr>
<tr>
<td>ฑ [d]</td>
<td>ฒ [d]</td>
<td>ฒ [d]</td>
</tr>
<tr>
<td>ป [b]</td>
<td>ป [p]</td>
<td>ป [p]</td>
</tr>
</tbody>
</table>

Table 1—Thai consonant chart inherited from Sanskrit
Vowels

There are 32 vowel forms in Thai, which could be written by combinations of 18 vowel symbols and 3 consonants. The 32 vowel forms are:

```
[ɨː] [ɔː] [i] [ɨː] [uː] [ɯː] [ʊ] [uː] [eː] [e] [ɛː] [o] [ɔː]
[ɨː] [ɔː] [i] [ɨː] [uː] [ɯː] [ʊ] [uː] [eː] [e] [ɛː] [o] [ɔː]
```

The 18 vowel symbols are:

```
[ɨː] [ɔː] [i] [ɨː] [uː] [ɯː] [ʊ] [uː] [eː] [e] [ɛː] [o] [ɔː]
```

And the 3 consonants used to compose vowels are ːə ːп and ː.

Tones

Thai is a tonal language. There are 5 tones in Thai syllables, namely Saman (middle), Ek (low), Tho (falling), Tri (high) and Chattawa (rising).

There are 4 tone marks in Thai script, namely Mai Ek (⁀), Mai Tho (è), Mai Tri (⁁) and Mai Chattawa (⁂). Despite their names, the tone marks do not directly imply the actual tone of the syllables in which they appear. The initial consonant and the syllable ending sound are also factors which determine the actual tone.

Diacritics

There are 2 diacritics for writing Thai words: Mai Taikhu (⁀) or sound shortener, and Thanthakhat (⁁) or sound killer; and additional 3 marks for Pali/Sanskrit words: Nikhahit (⁃), Pinthu (⁄), and Yamakkhan (⁅).

Digits

Thai has ten decimal digits (0-9): ๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙.

Typographical Symbols

Paiyan Noi (¶) or omission sign, and Mai Yamok (§) or repetition sign, are contemporary punctuation marks generally used in documents. Thai Baht currency sign is written as ฿.

In addition, there are three ancient typographical symbols used in some documents: Fongman (⊕) or sentence/stanza/paragraph beginner, Angkhandeaw (¶) or sentence/stanza terminator (normally written using the same glyph as Paiyan Noi), Angkhankhu (¶) or chapter/episode terminator, and Khomut (⊕) or story terminator. [24]
2 Character Sets

The only standard for Thai character set for information interchange is **TIS 620-2533 (1990)**, defined by the Thai Industrial Standard Institute (TISI), Ministry of Industry. It defines two eight-bit character code sets by extending ISO 646-1983 and IBM GX20-1850-4 (EBCDIC), respectively. TIS 620-2533 (1990) is indeed a minor correction of the previous one, **TIS 620-2529 (1986)**. The character set table remains the same. Only some detailed descriptions are added for compliance with international standards.

TIS 620 extends the 7-bit character set of ISO 646 by using 8-bit code table and defining 87 Thai characters in the GR area, namely A1-DA and DF-FB, as shown in Table 2. And it has also been submitted to the European Computing Manufacturers’ Association (ECMA) for registering in the ISO 2375 repertoire. ECMA registered and assigned it ISO-IR-166. Therefore, it can be used with ISO/IEC 2022, and an example of such implementation is GNU Emacs.

Around 1996, a TISI working group had drafted the Latin/Thai part (part 11) of ISO/IEC 8859 based on TIS 620, but, due to the prohibition of combining characters in the sense of ISO/IEC 2022, it did not pass, and the work was suspended.

In Thailand, most Thai documents are encoded using the extension of ISO 646 in TIS 620. (The use of the EBCDIC extension part is as rare as the use of EBCDIC itself.) In the internet, there has been a confusion of MIME character sets for Thai e-mails and web pages among the ad hoc solutions, such as using “iso-8859-1”, “x-user-defined” and “windows-874”, until, in 1998, Trin Tantsetthi [21] requested for “tis-620” character set registration with the Internet Assigned Number Authority (IANA), according to RFC 2278, for information interchange on the internet. IANA registered it in the same year. Now the MIME character set is under usage campaign. Mozilla project [30] has already supported it.

Thai part in ISO/IEC 10646-1 and Unicode 2.1 (U+0E00 – U+0E5F) is the code-by-code mapping with the characters A0-FF of TIS 620.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>ก</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ข</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ค</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ด</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ต</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ง</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>จ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>ฉ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>ช</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>ซ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>ฌ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>ฉ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>ญ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>ฎ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>ฏ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

■ = same as ISO 646
shaded = unspecified

Table 2—TIS 620-2533 character set (ISO 646 extension)
3 Input Method

Thai software industry had suffered a diversity of Thai language implementations until a group of professionals had joined together to form the Thai API Consortium (TAPIC), headed by Thaweesak Koanantakool and sponsored by National Electronics and Computer Technology Center (NECTEC), for drafting a proposal for Thai API standard called WTT 2.0 [14] in 1991. (The name WTT comes from its Thai abbreviation “ททท”, or “Wor Tor Tor”, of which the full name is “วิถีทั่วไป”, meaning “runs everywhere”.) The proposal comprises 3 parts, defining the general facilities, Thai input/output methods, and printer identification number, respectively. The proposal was submitted to the TISI/TC536/SC2 for adopting as a standard in 1991.

WTT 2.0 was a consequence of WTT 1.0 system developed by Thaweesak Koanantakool and his team at Thammasat University. Although not proclaimed as a standard, due to the wide cooperation among industrial companies, such as Digital, Sun, Microsoft and IBM, WTT 2.0 has been adopted in many systems, such as Digital UNIX, Thai Language Environment (TLE) for Solaris [16], Microsoft Windows 95/98 and Windows NT [17]. This makes WTT 2.0 become de facto.

According to WTT 2.0, Thai words are input and stored letter-by-letter from left to right. For the cells which contain symbols above or below baseline, the base character will be stored first, followed by the lower or upper vowel, and then the tone mark or diacritic at the top-most level. For example, the phrase ‘ภูทั้งป่า’ will be input and stored as ‘ภู-ทั้ง-ป่า’. A WTT 2.0 compliant input method can provide a sequence checking mechanism to ensure the validity of the input sequence in one of the three levels of strictness: level 0: pass through, level 1: basic check, level 2: strict check.

WTT 2.0 has also specified the cursor movements and editing behavior of Thai text editors and word processors. Cursor must be moved from cells to cells, that is, all characters in other levels than the base line must be skipped. Text deletion using the “Delete” key must also remove all characters in the current cell, including the above, below and top symbols. Meanwhile, character-by-character, right-to-left, removal is still possible by using the “Backspace” key, where the order of removal is considered by the order they are stored.

Thai computer keyboard layout is specified by TIS 820-2538 (1995), as shown in Figure 1. TIS 820-2538 is a modification of the previous one, TIS 820-2531 (1988). TIS 820 are modified from Ketmanee, a famous layout for typewriters. There was, however, another keyboard layout resulted from a character frequency distribution research, named Pattachote. But it is not as familiar for typists as Ketmanee.

![Figure 1—Layout of Thai Computer Keyboards (TIS 820-2538)](image-url)
4 Output Method

Thai is written in 4 levels, as shown in Figure 2.

![Thai 4-level writing system](image)

Figure 2—Thai 4-level writing system

The multiple levels of Thai texts are rendered by means of combining characters. WTT 2.0 classifies printable characters in TIS 620 as forward characters and dead characters. Forward characters are those characters written on the base line and occupy horizontal space, while dead characters are those written below or above base line and whose widths are zero.

Thai quality display and printing has been enabled by the font technologies. Among these, TrueType is the most widely used. New codes for presentation purpose have been added. For example, since some Thai consonants with long tails occupy a little space in the upper level, new codes for upper vowels and tone marks shifted left have been added. Another need is the adjustment of tone mark position when upper vowel is absent. Thus, new codes for lower tone marks are also added. (See Figure 3 for other adjustments.)

![Vowels and tone marks adjustment](image)

(a) Vowels and tone marks adjustment

![Base removal when combined with below vowel](image)

(b) Base removal when combined with below vowel

![Lowered below vowel](image)

(c) Lowered below vowel

Figure 3—Added glyphs for quality text representations

However, there are two major different code tables for Thai TrueType fonts. One is defined for Mac OS Thai, another is for Microsoft Windows. The one for Mac OS is based on MacThai character set [19], while that for Microsoft is an extension of code page 874 [18]. This makes the two kinds of TrueType fonts unable to be used together.

5 Lexicographical Ordering

The widely accepted standard for Thai lexicographical ordering is defined in the Royal Institute Dictionary 2525 B.E. Edition [23], the official standard Thai dictionary. And TIS 620 code table has been carefully specified according to this principle. The ordering principle is:

1. Words are ordered alphabetically, not by sounds. Consonants are ordered according to traditional order. TIS 620 has defined codes for 44 consonants and 2 vowels ɿ, ɿ according to this order in the range A1-CE.
2. Vowels are also ordered by written forms, not by sounds. TIS 620 has defined codes for vowels according to this order in the range D0-D9 and E0-E4.

3. Consonants always precede vowels. Word comparison is processed from left to right by considering initial consonants before vowels in the same syllable.

4. Tones and diacritics are normally ignored, unless all other parts are equal, in which case the order is as traditional order. TIS 620 defines order for these marks in the range E7-EB.

The principle is sufficient for use in the dictionary. And TIS 620 code specification almost enables C standard library function `strcmp()` to work with Thai strings, with two exceptions [26]:

1. **Leading vowels** (E0-E4 in TIS 620), which are written before the consonants, must be considered after the consonants. Therefore a reordering is needed before the actual comparison. This is also described as rearrangement in the draft of Unicode Technical Report #10 (Unicode Collation Algorithm) [25].

2. **Tone marks and diacritics** (E7-EB in TIS 620) must be ignored in the first pass, and be considered at later pass if the first pass yields equality.

Several Thai developers have designed sorting algorithms after this consideration, some of which have been summarized in [26]. However, the Royal Institute principle does not cover all cases in TIS 620 table. Therefore, a group of developers [27] have worked out the non-covered part and have proposed a generic string collation principle for Thai, based on ISO/IEC 14651 sorting model. Table 3 summarizes the proposed principle.

Table 3—Thai lexicographical ordering proposed by T. Karoonboonyanan et. al. [27]

<table>
<thead>
<tr>
<th>Level</th>
<th>Order</th>
<th>TIS 620</th>
<th>ISO/IEC 10646</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>digits (language-insensitive), English alphabets (case-insensitive), Thai consonants, Nikkhahit ’, Thai vowels</td>
<td>30-39</td>
<td>U0030-U0039</td>
<td>• U0030=;U0039, ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F0-F9</td>
<td>U0E50-U0E59</td>
<td>U0039=;U0039</td>
</tr>
<tr>
<td></td>
<td></td>
<td>41-5A</td>
<td>U0041-U005A</td>
<td>U0041=;U0051, ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>61-7A</td>
<td>U0061-U007A</td>
<td>U005A=;U007A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1-CE</td>
<td>U0E01-U0E0E</td>
<td>U0E32=;U0E45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ED</td>
<td>U0E4D</td>
<td>rearrange</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D0-D9</td>
<td>U0E30-U0E39</td>
<td>U0E40–U0E44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E0-E5</td>
<td>U0E40-U0E45</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Latin-script property, Thai-script property, Yamakkan ’, Pinthu ,, Thanthakhat ’, Mai Taikhu ‘, Thai tone marks</td>
<td>30-39</td>
<td>U0030-U0039</td>
<td>• U0030&lt;;U0039, ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F0-F9</td>
<td>U0E50-U0E59</td>
<td>U0039&lt;;U0039</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EE</td>
<td>U0E4E</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DA</td>
<td>U0E3A</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EC</td>
<td>U0E4C</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E7</td>
<td>U0E47</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E8-EB</td>
<td>U0E33-;U0E45</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>English-lower-case property, English-upper-case property, Thai-extra property</td>
<td>61-7A</td>
<td>U0061-U007A</td>
<td>• U0041&lt;;U0061, ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>41-5A</td>
<td>U0041-U005A</td>
<td>U005A=;U007A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E5</td>
<td>U0E45</td>
<td>• U0039&lt;;U0039</td>
</tr>
<tr>
<td>4</td>
<td>space, non-breaking space, low line _, hyphen -, comma ,, semicolon ;, colon :</td>
<td>20</td>
<td>U0020</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>A0</td>
<td>U00A0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5F</td>
<td>U005F</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2D</td>
<td>U002D</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2C</td>
<td>U002C</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3B</td>
<td>U003B</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3A</td>
<td>U003A</td>
<td></td>
</tr>
</tbody>
</table>
Here are the explanations of some cases [27] [24] :

**Paiyan Noi (U0E2F) and Mai Y amok (U0E46)**

Paiyan Noi (ʼ) is an abbreviation sign, representing omitted text. Its function is like ellipsis, although the more correspondent Thai punctuation mark for ellipsis is Paiyan Yai (ʼʼʼʼ). Therefore, it makes sense to order it after ellipsis category in the Common Template, after Devanagari abbreviation sign <U0970>.

Mai Y amok (ʻ) is a duplication sign, representing a repeat of the preceding word. Therefore, it deserves to be sorted in the duplication marks category in the Common Template.

**Nikhahit (U0E4D)**

According to traditional Thai principle, Nikhahit (ʼ) is a name of a vowel symbol. It, with Pin I (ʼ) compose Sara Ue (ʼʼʼʼ), with Lakkhang (ʼ) compose Sara Am (ʼʼʼʼ).

However, in Thai document, such function is ineffective, because TIS 620 has defined codes for such combined vowels, and they were input with a single keystroke. Instead, Nikhahit is normally used as Pali-Sanskrit-inherited consonant. For example, following Pali, पालि is pronounced [pʰāḷhī]; and following Sanskrit, फ़्ति is pronounced [tʃh umnuː]. Therefore,
Nikhahit arguably deserves to be classified as a consonant, ordered after the last Thai consonant (وء).

Yamakkan (U0E4E), Pinthu (U0E3A), Thanthakhat (U0E4C), Mai Taikhu (U0E47) and Tone marks (U0E48-U0E4B)

Yamakkan (²) is an ancient punctuation mark used to mark cluster, such as ทรายธน [thraθtn]. Pinthu (: ) has two functions, one is the same as Yamakkan, such as a Sanskrit word ทษธน [thṣṭhn], the other is to mark final consonant in Pali writing system, such as ท ช [抔สะ].

Thanthakhat (ٱ) is used for killing letter’s sounds, such as ทษธน [thṣṭnh], in which the consonant ษ is killed (not pronounced). สกหลน [āk], in which the three consonants งน are killed, นทมี [nāmī], in which the two letters นม are killed.

The three marks can be considered as having less effect in changing words than tone marks. For example, the meaning of ทษธน and ทษธน are not different, while ทษ and ทษ are. Therefore, they deserved to be ordered before tone marks.

Pinthu may come between Yamakkan and Thanthakhat, because another function of Thanthakhat is to mark final consonant in ancient writing of Pali words, which is in common with Pinthu.

In summary, the order should be Yamakkan, Pinthu, Thanthakhat, Mai Taikhu, Mai Ek, Mai Tho, Mai Tri, and Mai Chattawa.

Lakkhang Yao (U0E45) and Sara Aa (U0E32)

Lakkhang Yao (¹) is always written after Ru (:not pronounced), and Lu (اء), to produce new symbols Rue (ت) and Lue (ل). Traditional Thai principle treats each of these two new symbols as an atomic entity. Therefore, in general, Lakkhang Yao is never written after any letters other than Ru (ر) and Lu (اء). Meanwhile, the two symbols are never followed by a vowel. Hence, the occurrence of Lakkhang Yao (¹) and Sara Aa (اء — a vowel) are mutually exclusive, and assigning order between them seems not necessary. However, due to their similar shapes, they are occasionally confused (in a reasonable way). So, they should be treated as weakly identical, that is, their weights should be different at the last level of comparison.

Fongman (U0E4F), Angkhankhu (U0E5A), and Khomut (U0E5B)

As described earlier, Fongman (¶) is used in ancient books as paragraph, sentence, or poem stanza begin marker. Angkhandeaw (¶) is used to end a sentence or a stanza. Angkhankhu (¶) is used to end a chapter or episode. Khomut (¶) ends a story.

These marks can be classified as typographic symbols in the Common Template, where Fongman functions like a bullet. Unfortunately, there is no specific code for Angkhandeaw ( ¶) and Paiyan Noi is always used instead. Therefore, we should ignore the existence of Angkhandeaw and the similarity between Paiyan Noi (¶) and Angkhankhu (¶) when considering their orders.
6 Word Boundary

There is no word delimiter in Thai writing system. Therefore, a computer program that processes Thai documents needs to be able to tokenize Thai strings into words, using certain kinds of linguistic knowledge.

There have been several approaches to this famous problem of Thai language. The first approaches were rule-based. This needed human knowledge to find out rules. But there were too much exceptions and an accurate rule set seems impossible. Dictionary-based approaches were thus proposed. Several Thai word processors nowadays are using these approaches and users are happy at some degree. However, it is still too coarse for more fine-grained applications like machine translation. Therefore, modern research are now focusing on statistical and machine learning approaches.

However, it still lacks standard API for Thai word tokenization, although there was awareness of requirement among software developers to mediate the diversity of the existing word break API’s, so that the word boundary finding task of an internationalized software can be split and left to the pluggable libraries.

Existing word break APIs include:

1. Line wrapping within hot zone

   Example
   
   ```
   int LineWrap(const char* text, const int charWidths[], int rightMargin);
   ```
   returns last word position that is not greater than the rightMargin for the given text string, of which the width of each character is given in charWidths[] correspondingly.

2. Word delimiters insertion

   Example
   
   ```
   int InsertWBR(const char* text, char* output, const char* delimiter);
   ```
   reads text, copies each word into output buffer, putting delimiter at the end of each word, and returns the number of words found.

3. Filling array of pointers to word break positions

   Example
   
   ```
   int FindWBR(const char* text, int* output);
   ```
   reads text, finds the word boundaries within it, fills the word break positions in output buffer, and returns the number of words found.

4. Finite state machine [28]

   Example
   
   ```
   int Reset();
   int AddChar(char c);
   ```
   Reset() initializes the internal abstract state machine. AddChar(c) feeds c to the internal machine, returns word length if reaching a word break position, or zero if the word is not complete yet.

5. Word break iterator [29]

   Example
   
   ```
   class ThBreakIterator {
   ```
7 Minority Scripts in Thailand

There is no thorough research about contemporary scripts uses in Thailand. It is clear that almost 100 percent of Thai population are using Thai script, as they are taught in public schools and it is the only official national language.

However, in the four Muslim provinces of southern Thailand, Pattani dialect of Jawi with script is being used in everyday life by the local people. And the script is different from Malay Jawi script used in Malaysia.

In the northern and north-eastern regions, there are two ancient scripts used in old writing materials like stones and palm leaves. One is Fagkham script, the old Tham script closely blended with Sukhothai script. The other is Muang script, which looks closer to Tham script used in Laos [31] [32]. Nowadays, Muang script is found being used for religious study materials.

However, the need of the computerization of the scripts, including what other scripts being used, need further research and survey.

References


Sun Microsystems, Inc. TLE 2.2 Release Overview. 1994.


Peter Edberg, Julio Gonzalez, and John Jenkins. Map (external version) from Mac OS Thai character set to Unicode 2.0. ftp://ftp.unicode.org/Public/MAPPINGS/VENDORS/APPLE/THAI.TXT.


Prasert Na Nakhon. Lai Sue Thai. 700 Years of Lai Sue Thai. Dhurakit Bhundit. 1983. (in Thai)