Standardization and Implementations of Thai Language

Theppitak Karoonboonyanan

National Electronics and Computer Technology Center, THAILAND.

Overview

- ◎ Thai Language
- ◎ Thai Character Set
- ◎ WTT 2.0
- ◎ Input Method
- ◎ Output Method
- ◎ Lexicographical Ordering
- ◎ Word Boundary
- ◎ Minority’s Scripts

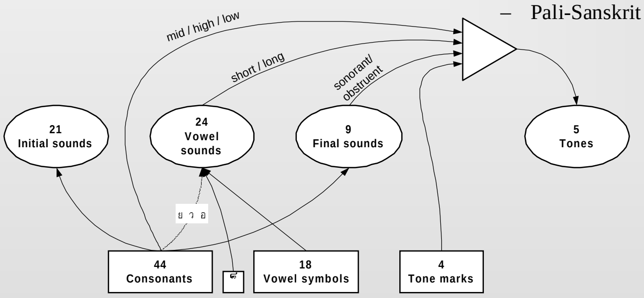
Thai Language (1)

◎ Syllable Components

– Initial Sound	[21]
– Vowel	[24]
– Final Sound	[9]
– Tone	[5]

◎ Writing System Components

– Consonants	[44]
– Vowel Symbols	[18]
– Tone Marks	[4]
– Sound Marks	[2]
– Pali-Sanskrit Marks	[3]



Thai Language (2)

44 Consonants → 21 Initial sounds + 9 Final sounds

	middle	high	low			
			paired		single	
/-k/	ก [k]	ข (ฃ) [kh]	ค (ฅ) [kh]	ฆ [kh]	ง [ŋ]	/-ŋ/
	จ [tɕ]	ฉ [tɕh]	ช (ซ) [tɕh] [s]	ฌ [tɕh]	ญ [j]	
/-t/	(ฎ) ฏ [d] [t]	ฐ [th]	ฑ [th]	ฒ [th]	ณ [n]	/-n/
	(ด) ต [d] [t]	ถ [th]	ท [th]	ธ [th]	น [n]	
/-p/	(บ) ป [b] [p]	ผ (ฝ) [ph] [f]	พ (ฟ) [ph] [f]	ภ [ph]	ม [m]	/-m/

ย	ร	ล	ว	ศ	ษ	ส	ห	ฬ	(อ)	(ฮ)
[j]	[r]	[l]	[w]	[s]	[s]	[s]	[h]	[l]	[ʔ]	[h]
/-j/	/-n/	/-n/	/-w/	/-t/	/-t/	/-t/		/-n/		

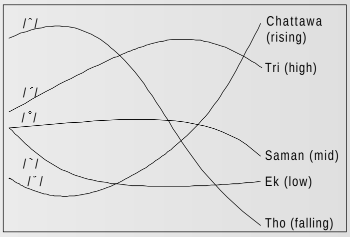
18 Vowel Symbols

ะ ั า ำ ิ ี ึ ื ุ ู เ แ โ ใ ไ ฤ ฦ ๅ

4 Tone Marks

่ ้ ๊ ๋

5 Tones



2 Diacritics + 3 Pali-Sanskrit Diacritics

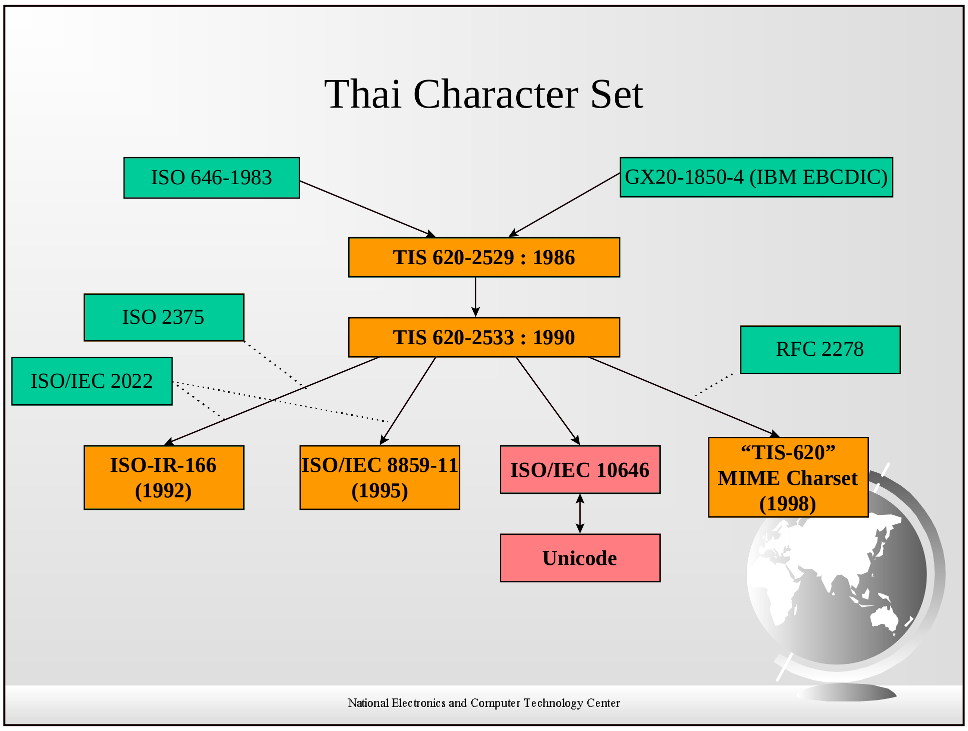
์ ็ ฺ ๎ ํ

10 Digits

๐๑๒๓๔๕๖๗๘๙

6 Typographical Symbols

ฯ ๆ ฿ ๏ ๚ ๛



Thai Character Set

ISO 646-1983

GX20-1850-4 (IBM EBCDIC)

TIS 620-2529 : 1986

ISO 2375

TIS 620-2533 : 1990

RFC 2278

ISO/IEC 2022

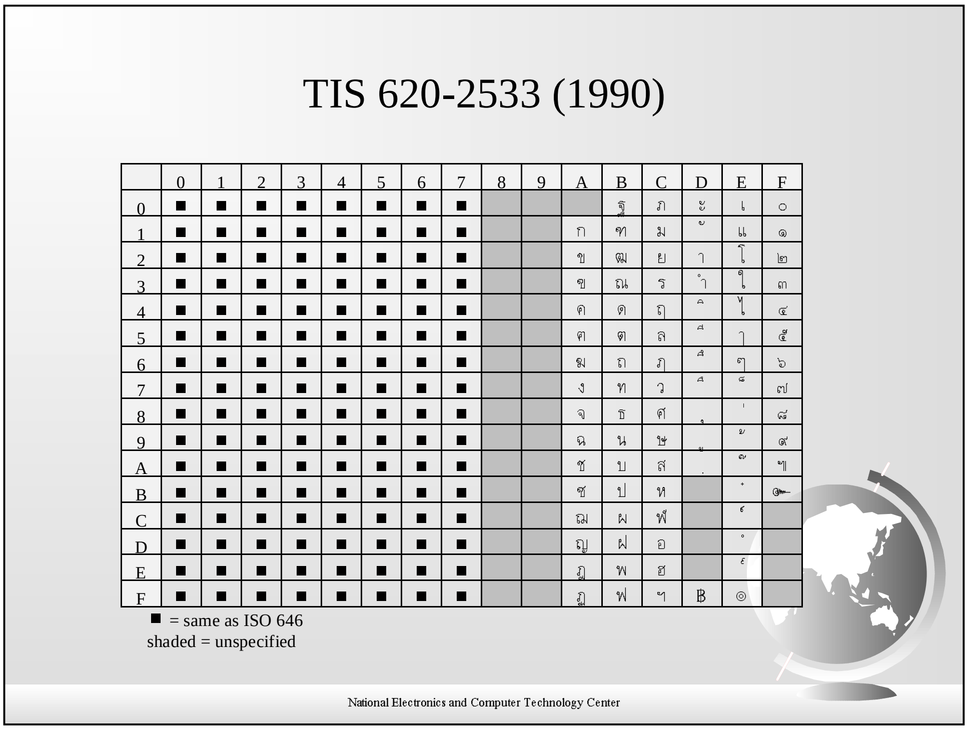
ISO-IR-166 (1992)

ISO/IEC 8859-11 (1995)

ISO/IEC 10646

"TIS-620" MIME Charset (1998)

Unicode



TIS 620-2533 (1990)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	■	■	■	■	■	■	■	■	▨	▨	▨	ฐ	ภ	ะ	เ	๐
1	■	■	■	■	■	■	■	■	▨	▨	ก	ฑ	ม	ั	แ	๑
2	■	■	■	■	■	■	■	■	▨	▨	ข	ฒ	ย	า	โ	๒
3	■	■	■	■	■	■	■	■	▨	▨	ฃ	ณ	ร	ำ	ใ	๓
4	■	■	■	■	■	■	■	■	▨	▨	ค	ด	ฤ	ิ	ไ	๔
5	■	■	■	■	■	■	■	■	▨	▨	ฅ	ต	ล	ี	ๅ	๕
6	■	■	■	■	■	■	■	■	▨	▨	ฆ	ถ	ฦ	ึ	ๆ	๖
7	■	■	■	■	■	■	■	■	▨	▨	ง	ท	ว	ื	็	๗
8	■	■	■	■	■	■	■	■	▨	▨	จ	ธ	ศ	ุ	่	๘
9	■	■	■	■	■	■	■	■	▨	▨	ฉ	น	ษ	ู	้	๙
A	■	■	■	■	■	■	■	■	▨	▨	ช	บ	ส	ฺ	๊	๚
B	■	■	■	■	■	■	■	■	▨	▨	ซ	ป	ห	▨	๋	๛
C	■	■	■	■	■	■	■	■	▨	▨	ฌ	ผ	ฬ	▨	์	▨
D	■	■	■	■	■	■	■	■	▨	▨	ญ	ฝ	อ	▨	ํ	▨
E	■	■	■	■	■	■	■	■	▨	▨	ฎ	พ	ฮ	▨	๎	▨
F	■	■	■	■	■	■	■	■	▨	▨	ฏ	ฟ	ฯ	฿	๏	▨

■ = same as ISO 646
shaded = unspecified

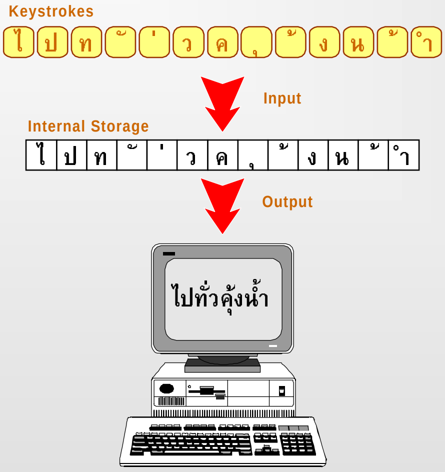
WTT 2.0 (1)

- About **WTT 2.0** (1991)
  - A Thai standard API project
  - WTT = Wor Tor Tor = วทท = วิ่งทุกที่ = Runs Everywhere
  - WTT 1.0 + Thai API Consortium (TAPIC)
    - **Sponsor :** NECTEC
    - **Head :** Dr. Thaweesak Koanantakool
    - **Members :** Digital, OCT & Datamat (Sun), Microwiz (Microsoft), IBM, etc.

WTT 2.0 (2)

- **Status :**
  - Submitted to TISI for adopting as a standard in 1991.
  - Although not endorsed yet, WTT 2.0 has been widely implemented by the alliance companies, such as Digital UNIX and TLE for Solaris.
  - Therefore, *de facto.*
- **WTT 2.0** Contents
  - 3 Specification Drafts
    - General Programming Facilities (char type, char name, etc.)
    - *Thai Input/Output Method*
    - Printer ID

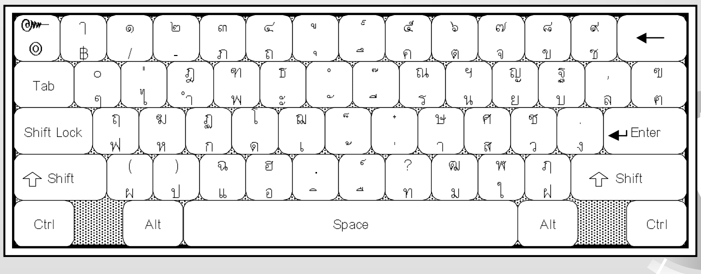
Thai Input Method (1)

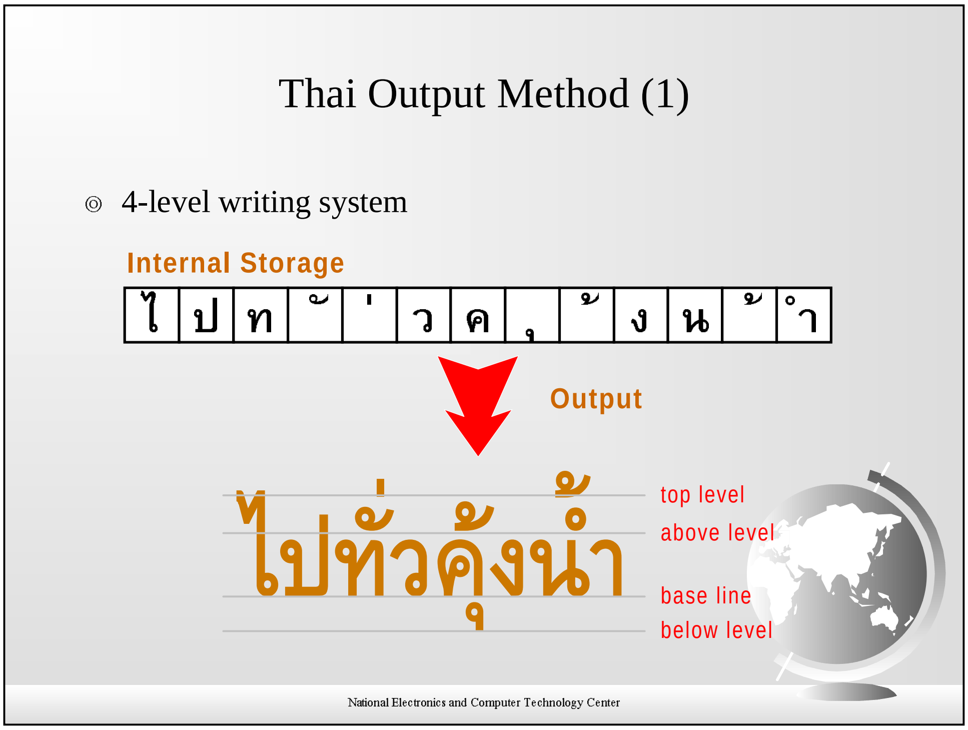


- Left to right
- One keystroke per one character
- Input sequence per cell (column):
  - Base level
  - Above/Below level
  - Top level
- 3 levels of sequence check : pass through, basic, strict

Thai Input Method (2)

- 2 kinds of keyboard layouts
  - Ketmanee (traditional typewriter layout)
  - Pattachote (from character frequency distribution research)
- **TIS 820-2538 (1995)**, modified from **TIS 820-2531** (1988), which is based on Ketmanee





Thai Output Method (1)

◎ 4-level writing system

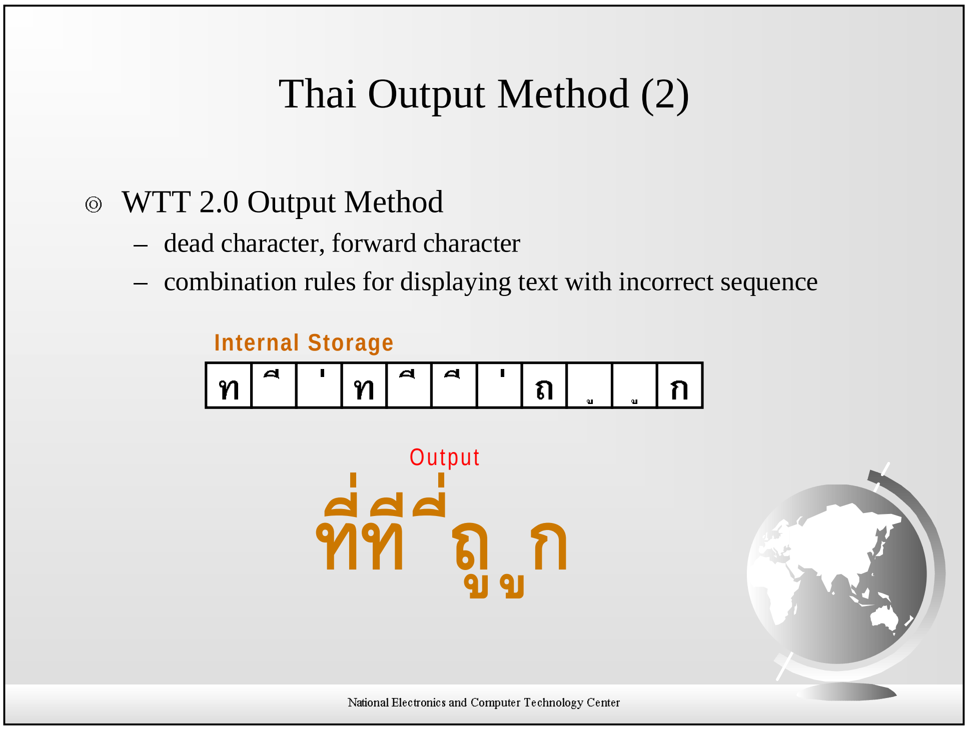
Internal Storage

ไ	ป	ท	ั	่	ว	ค	ุ	้	ง	น	้	ำ

Output

ไปทั่วคุ้งน้ำ

top level
above level
base line
below level



Thai Output Method (2)

◎ WTT 2.0 Output Method
- – dead character, forward character
- – combination rules for displaying text with incorrect sequence

Internal Storage

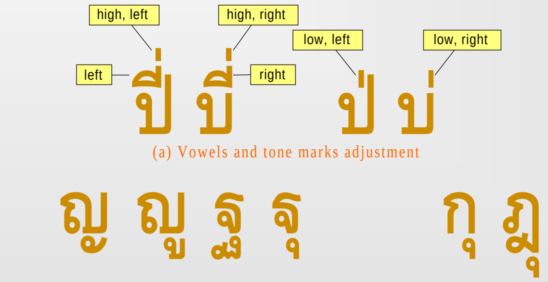
ท	ี	่	ท	ี	ี	่	ถ	ู	ู	ก

Output

ที่ทีี่ถููก

Thai Output Method (3)

- Glyph adjustments for quality publishing



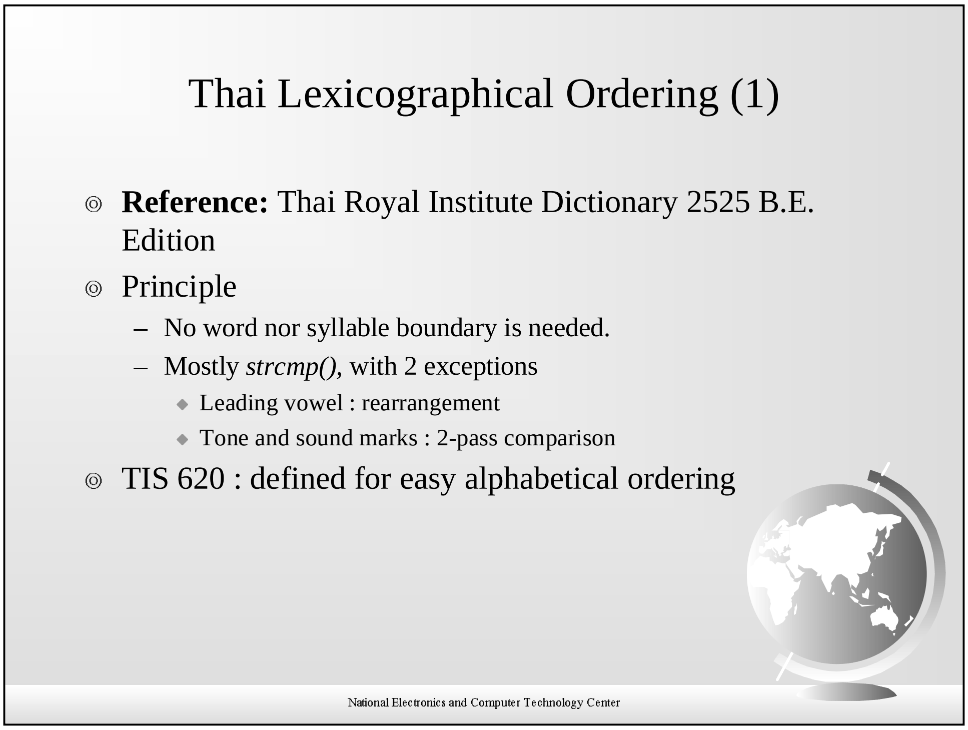
(a) Vowels and tone marks adjustment

(b) Base removal when combined with below vowel

(c) Lowered below vowel

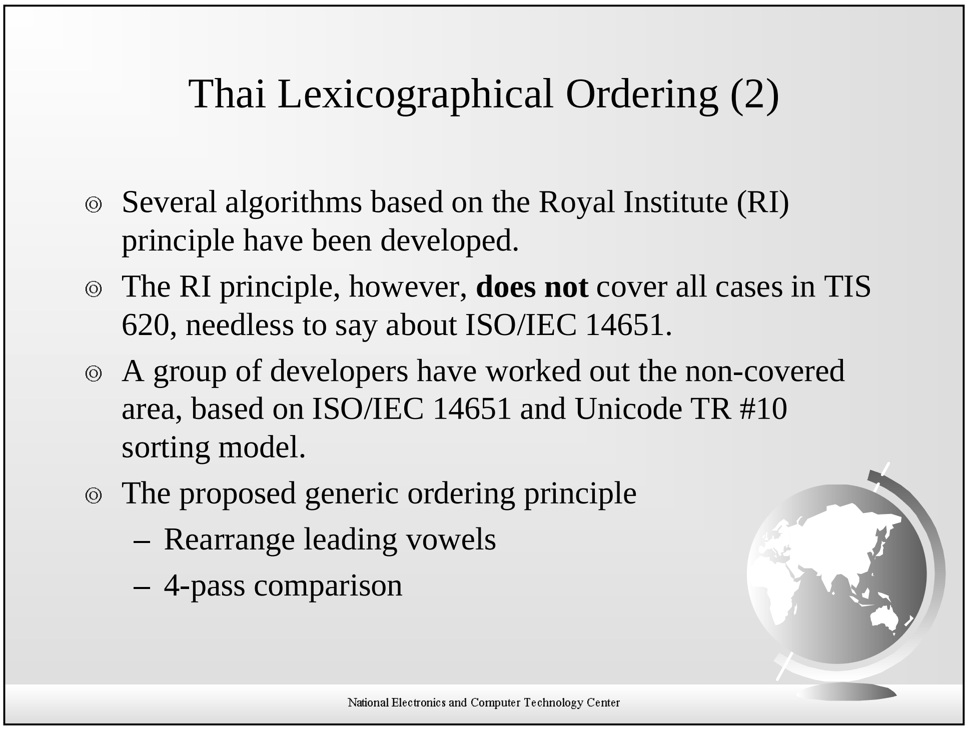
Thai Output Method (4)

- Glyph adjustments in TrueType fonts
  - Mac OS Thai : based on MacThai character set
  - Microsoft Windows : extends Codepage 874
- Problem : Incompatibility between the two kinds of fonts



Thai Lexicographical Ordering (1)

- **Reference:** Thai Royal Institute Dictionary 2525 B.E. Edition
- Principle
  - No word nor syllable boundary is needed.
  - Mostly *strcmp()*, with 2 exceptions
    - Leading vowel : rearrangement
    - Tone and sound marks : 2-pass comparison
- TIS 620 : defined for easy alphabetical ordering



Thai Lexicographical Ordering (2)

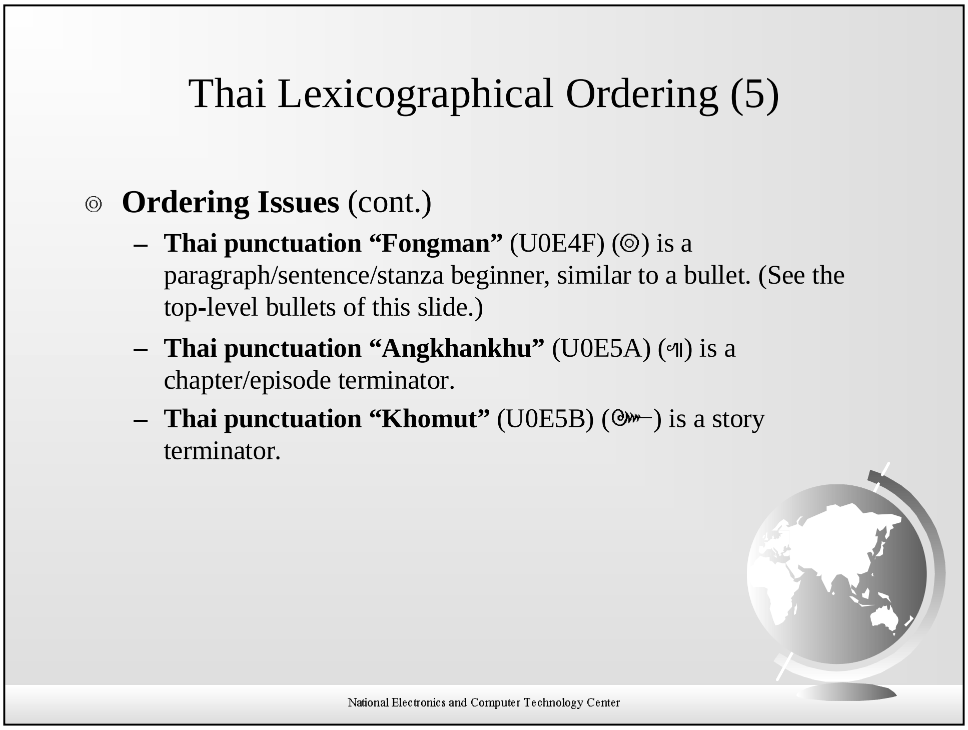
- Several algorithms based on the Royal Institute (RI) principle have been developed.
- The RI principle, however, **does not** cover all cases in TIS 620, needless to say about ISO/IEC 14651.
- A group of developers have worked out the non-covered area, based on ISO/IEC 14651 and Unicode TR #10 sorting model.
- The proposed generic ordering principle
  - Rearrange leading vowels
  - 4-pass comparison

Thai Lexicographical Ordering (3)

- **Ordering Issues**
  - **Digits** Corresponding digits in different languages are treated equal in level 1, discriminated in level 2. [~14651]
  - **Latin Alphabets** are case-insensitive in level 1, discriminated in level 3. [~14651]
  - **Thai character “Nikhahit”** (U0E4D) comes after the last consonant (U0E2E) and before the fist vowel (U0E30) in level 1.
  - **Thai leading vowels** (U0E40-U0E44) is rearranged (swapped with the next character) in level 1.
  - **Thai vowel “Sara Aa”** (U0E32) (า) is treated equal to **Thai character “Lakkang Yao”** (U0E45) (ๅ) in level 1, discriminated in level 3.

Thai Lexicographical Ordering (4)

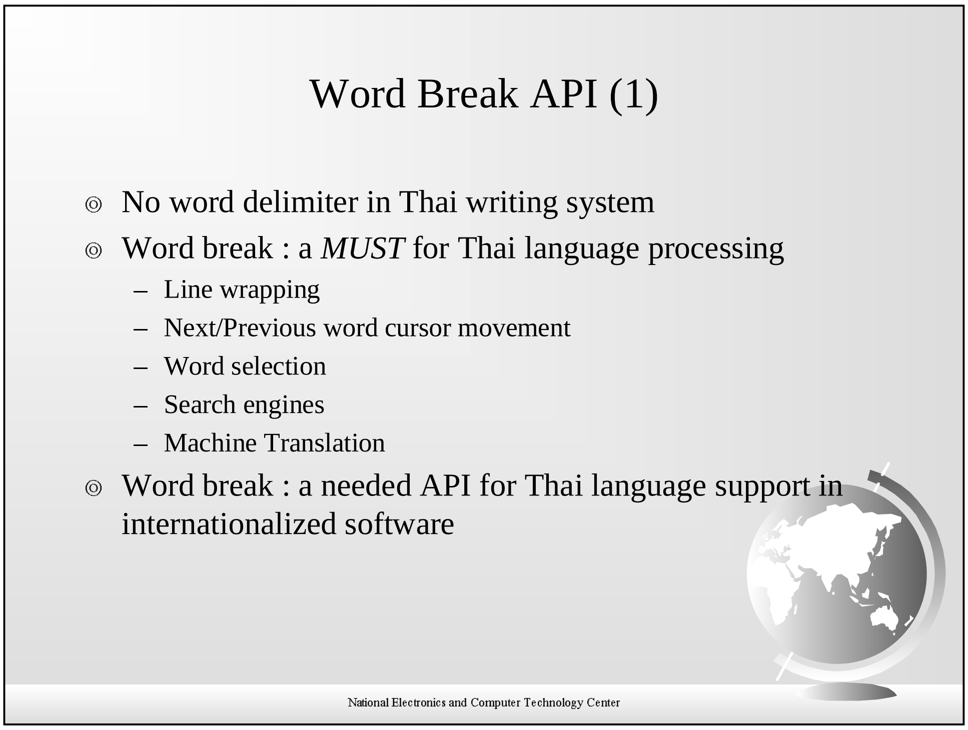
- **Ordering Issues** (cont.)
  - **Thai diacritics and tone marks** are sorted in this order in level 2 :
    - **Yamakkan** (U0E4E), **Pintu** (U0E3A), **Thanthakhat** (U0E4C), **Mai Taikhu** (U0E47), **Mai Ek** (U0E48), **Mai Tho** (U0E49), **Mai Tri** (U0E4A), **Mai Chattawa** (U0E4B)
  - **Thai punctuation “Paiyan Noi”** (U0E2F) (ฯ) is an abbreviation sign, representing omitted parts of a word.
  - **Thai punctuation “Mai Yamok”** (U0E46) (ๆ) is a word/phrase repettition sign.



Thai Lexicographical Ordering (5)

๏ **Ordering Issues** (cont.)

- **Thai punctuation “Fongman”** (U0E4F) (๏) is a paragraph/sentence/stanza beginner, similar to a bullet. (See the top-level bullets of this slide.)
- **Thai punctuation “Angkhankhu”** (U0E5A) (๚) is a chapter/episode terminator.
- **Thai punctuation “Khomut”** (U0E5B) (๛) is a story terminator.



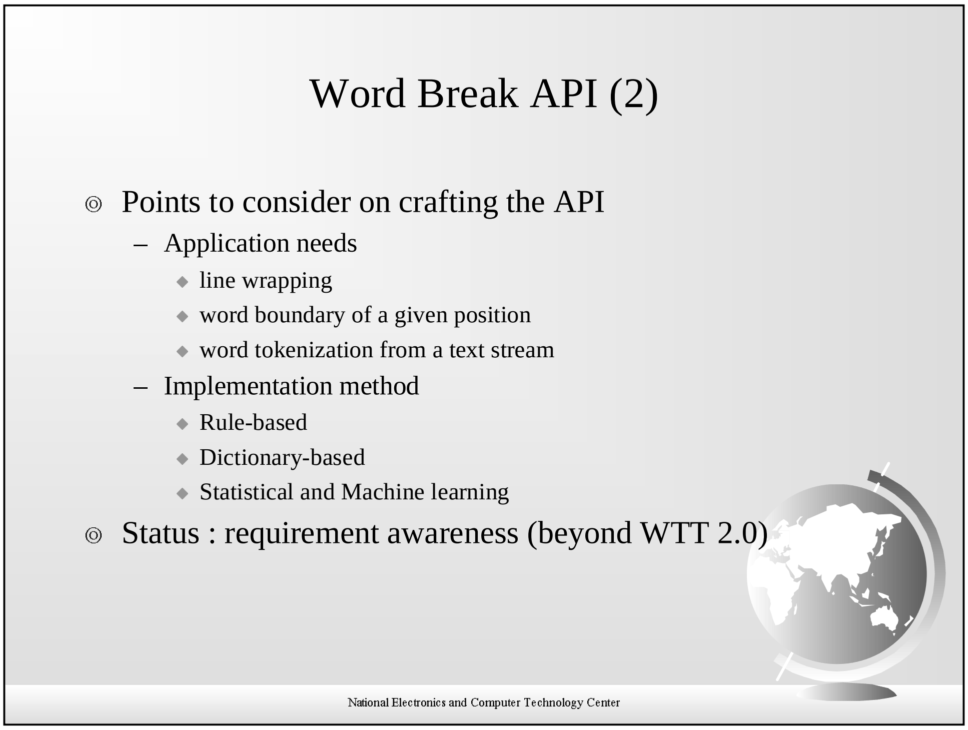
Word Break API (1)

๏ No word delimiter in Thai writing system

๏ Word break : a *MUST* for Thai language processing

- Line wrapping
- Next/Previous word cursor movement
- Word selection
- Search engines
- Machine Translation

๏ Word break : a needed API for Thai language support in internationalized software



Word Break API (2)

- Points to consider on crafting the API
  - Application needs
    - line wrapping
    - word boundary of a given position
    - word tokenization from a text stream
  - Implementation method
    - Rule-based
    - Dictionary-based
    - Statistical and Machine learning
- Status : requirement awareness (beyond WTT 2.0)

Minority’s Scripts

- No thourough research on contemporary script uses in Thailand.
- Known script being used
  - Jawi (Pattani dialect)
    - **Region:** the 4 Muslim provinces of southern Thailand
    - **Usage:** in everyday life
    - **Characteristics:** close to but different from Malay Jawi
  - Muang Script
    - **Region:** northern region of Thailand
    - **Usage:** in religious books
    - **Characteristics:** close to Tham script
- Further research is needed. ๚ะ๛