TEXT PROCESSING OF THAI LANGUAGE
=THE THREE SEALS LAW=

Shigeharu Sugita

National Museum of Ethnology
Expo Park, Senri, Suita
OSAKA 565, Japan

Abstract

Computer softwares for processing Thai language are developed at National Museum of Ethnology, Osaka, Japan. We use a popular intelligent terminal TEKTRONIX 4051 for inputting and editing, IBM 370 model 138 for KWIC making and sorting, and CANON's laser beam printer for final output.

Using these systems, "Kotmai Tra Sam Duang" (the Three Seals Law) which contains many kind of laws and ordinances proclaimed in Thai between 1350-1805 A.D. is computerized. This text has 1700 pages and about 1400000 letters. KWIC index becomes 200000 lines.

Some statistical data for this text are obtained. They are occurrence frequency data of single letter, group vowel, and letter combination (digram), etc.

Aknowledgements

This report is a result of joint project at National Museum of Ethnology. The member are Y. Ishii, I. Akagi, S. Tanabe, Y. Sakamoto, S. Uemura, A. Ishizawa, M. Sawamura, K. Sasaki, Y. Kurita, and S. Sugita. Their research field are ethnology, linguistics, computer science, and sociology etc.

We thanks Mr. Sophon Chitthasatcha, Miss Sumalee Maungpaisaln and Miss Hiroe Matsumoto for their help in segmentation, inputting and correction.

We also thanks Prof. K. Nakayama and A. Oikawa of Tsukuba University for their support on making Thai letter patterns and output software for laser beam printer.

Introduction

In the field of ethnology or cultural anthropology, ethnographies are very important information sources for comparative study of many different societies. Not only bibliographic data but also contents of text are necessary.

HRAF (Human Relations Area Files), which was developed by Dr. Murdock and now managed by HRAF Inc. at Yale University, is a unique retrieval system.

They use about 800 category codes by which analysts classify the contents of each pages of books.

Though HRAF system is an elaborate work, it is not easy to search necessary data by user terms, that is, natural words. If whole text are fed into computer, it is very easy to retrieve any part of text by the same natural words used in the text.

On-line retrieval system is smart and effective. But sometimes researcher wants printed index like as KWIC which is usable at any time and place. Combining KWIC index and thesaurus dictionary, it gives us a very powerful tools for searching special expression hidden in the text.

Till quite recently, at least in Japan, most cases of computer processing of natural language are distorted to indo-european language or Japanese. In the ethnological studies, we must treat many areas in the world. We need computer softwares which process unfamiliar languages for us, such as Arabic, Korean, Sumerian, Mongolian, Devanagari, Thai, etc.

National Museum of Ethnology at Osaka has introduced several computer systems to encourage humanity study, and now is developing many application softwares which are usable by any researchers who do not know computer programming or how to use computer.

This report describes one of such application softwares which treats Thai letters. The points of our work are as follows;

- 1) A popular computer terminal is used for Thai letter inputting and editing. It is easy to use because dead key operation is not necessary.
- 2) KWIC making and sorting software are implemented using FORTRAN language which can be transferred to any other computer system. The algorithm is not so complex but it was not implemented only because they are not popular language.
- 3) Statistical data of the text are obtained. They are occurrence frequency of single letter, group vowel, and letter combination. These data will help us as a contextual data in case of OCR.

Segmentation

There is no segmentation problems in case of indo-european languages, because they have clear separator for word unit such as space or comma. There are, however, many languages in Asia which have no clear separator. They are Korean (Hangul), Chinese, Japanese, and Thai, etc. Examples shown below mean that there exist several different segmentation. Segmentation affects to the meaning of sentence and retrieval efficiency.

国立民族学博物館

国立 / 民族学博物館

国立 / 民族学 / 博物館

国立 / 民族 / 学 / 博物 / 館

오늘 밤 나무사 온다

오늘	밤	나무	사	온	다
오늘	밤	나무	사	온	다
오늘	밤	나무	사	온	다
오늘	밤	나무	사	온	다
오늘	밤	나무	사	온	다
오늘	밤	나무	사	온	다
오늘	밤	나무	사	온	다
오늘	밤	나무	사	온	다
오늘	밤	나무	사	온	다
오늘	밤	나무	사	온	다

Fig.1 Examples of different segmentations

To cut into long unit is effort saving, but it is difficult to search the string included in that unit. To cut into short unit is effective for searching, but too many keywords appear.

The text, the Three Seals Law, has no word separator, as shown in fig.3. So it is necessary to segment into appropriate units before making KWIC index. But it is difficult problem because segmentation needs well understanding of meaning, which conversely needs KWIC index.

We adopted a practical method which at first cut into long unit and then cut again after looking KWIC index.

Fig.2 Flow diagram of KWIC making

Inputting

Terminal

We use a popular intelligent graphic terminal TEKTRONIX 4051 which has usual alphabet keyboard. We stucked Thai letter labels on the side of each key as if it looks like Thai typewriter. A code table of Thai letters and corresponding english alphabets is shown in Table 1.

The characteristics of this terminal are;

- 1) It generates Thai letter pattern by BASIC program in graphic mode. User can affirm the letter he typed.
- 2) It has local cassett memory, so that user can input and edit data anytime, even when host computer is not working.
- 3) By way of communication line, stored data can be transmitted to host computer for time consuming work.
- 4) It is easy to implement a flexible Thai language editor, which accept alphabet commands and display Thai letters.
- 5) Copy of screen can be taken by the hard copy unit attached to it.

Rules for text inputting

The text has many irregular expressions. So following expedencies are adopted.

- 1) Quotated words or phrases from Pali language are skipped by inserting special symbol to indicate there are skipped words.

- 2) Tables are skipped.
- 3) Special expressions for money, dating, and fractional number are transformed into sentence form.
- 4) Vertical expression shown in Fig.3 are attached special symbol after and before the word.
- 5) Parallel expression in the middle of line, and tree like expression are transformed into linear form from which original form can be reconstructable as much as possible.

Order of input

The order of input of Thai letters to the computer is same with the order in which one would strike keys of typewriter.

ฐ → ร
 ฐ → น
 เล้าเบ๊ย → เล้าเบ๊ย

- 8 พิมพ์งานนี้พอ
- 9 ขุนไฉนโยพระพมา ๒๐๐ /
- 10 นายวงสกล / นาค ๕๐ /
- 11 เลว / นาค ๓๐ /
- 12 ไม่ต่ำสูง ๓๑
- 13 (ขุนทหารวิเศษ /
- 14 เคน /
- 15 หมีน เคนเวหา /
- 16 เหาะ /) นาค ๒๐๐ / (a)

VOL.=4 PAGE=250

- 1 ๒๕
 - 2 ๒๕ อนึ่ง/มิ/กฎ/ให้ไว้/ว่า/ มีผู้ร้องฟ้อง/ใน/กรมใด ๆ /
 - 3 แล/ให้/เจ้ากรม/ปลัดกรม/ส่ง/ฟ้อง/ไป/โรงศาล/ ผู้ได้/ว่าความ/
 - 4 ให้พิจารณา/ตาม/กระทรวง/ อย่าให้/เจ้าพญา/แล/พญาพระ
 - 5 หลวง/ขุนหมื่น/เอาความ/ว่า/เป็นความ/หน้าโรง/
 - 6 กฎ/ให้ไว้/ณ/วันอาทิตย์ เดือน ๕ แรม ๕ ค่ำ/จุลศักราช ๑๐๔๔ / (b)
- F=109 (CONT)
- ง/หน้า/
- 16 พระกาล/ข้างละ ๓ สลึง / ค่าใส่คราด/ข้างละ ๒ บาท ๓ สลึง / เข้า
 - ๑๖ /
 - 17 เป็น/เงิน/ ๓ ตำลึง ๓ บาท ๓ สลึง / ให้/(ใจห/, จาเลข/) ล้าง/เห
 - า/เสีย/จงบมด/ ให้/ผู้คุม/ (c)

Fig. 3 Examples of text and inputted form

ขุนไฉนโยพระพมา ๒๐๐ / ขุนทหารวิเศษ /
 นายวงสกล / นาค ๕๐ / เคน / นาค
 เลว / นาค ๓๐ / หมีน เคนเวหา / ๒๐๐ /
 (A)

๒๕ อนึ่ง/มิ/กฎ/ให้ไว้/ว่า/ มีผู้ร้องฟ้อง/ใน/กรมใด ๆ /
 แล/ให้/เจ้ากรม/ปลัดกรม/ส่ง/ฟ้อง/ไป/โรงศาล/ ผู้ได้/ว่าความ/
 ให้พิจารณา/ตาม/กระทรวง/ อย่าให้/เจ้าพญา/แล/พญาพระ
 หลวง/ขุนหมื่น/เอาความ/ว่า/เป็นความ/หน้าโรง/
 กฎ/ให้ไว้/ณ/วันอาทิตย์ เดือน ๕ แรม ๕ ค่ำ/จุลศักราช ๑๐๔๔ / ๑๖
 มะแมนพสก /
 (B)

๒๙ ผู้คุมข้างละ ๑๒ / ข่าใบสี่ข้างละ ๑๒ / พิศุทล้วยเพลิงหน้า/
 พระกาลข้างละ ๑๒ / ค่าใส่คราดข้างละ ๑๒ / เข้ากัน
 เป็นเงิน ๓๒ (ใจห/ จาเลข) / ล้าง/ทำ/เสีย/จงบมด/ ให้/ผู้คุม
 (C)

Example text

Correction

Thai editor

A line editor for Thai text is implemented on TEKTRONIX 4051 terminal. Commands are english like term and Thai text are displayed by Thai letter.

This editor suppose that there are volume number, page number, and line number.

```

ENTER THE VOLUME NUMBER= v
      ;specify volume number
*PAGE,N      ;specify page number. Until
      ;next page command, this page
      ;is held in memory.
*LADD XX     ;XX is added to this page as
      ;a last line
*LINS,M XX   ;XX is inserted as a new line
      ;after line number m
*LDEL,M      ;line number m is deleted
      ;from this page
*SHOW,M      ;string of line number m is
      ;displayed in Thai letter
*LGET,M      ;line number m is object to
      ;be edited by following sub-
      ;commands
*ADB XX      ;XX is added to the beginning
      ;of line m
*ADE XX      ;XX is added to the end part
      ;of line m
*DEL XX      ;string XX is deleted from
      ;line m. If there are several
      ;XX's in line m, the position
      ;number are displayed. Enter
      ;corresponding number after
      ;prompt "which?"
*INS XX BEFORE YY
      ;string XX is inserted before
      ;string YY. If several YY's
      ;are there, type corresponding
      ;number after prompt "which?"
*REP XX BY YY
      ;string XX is replaced by YY.
      ;If several YY's are there,
      ;type corresponding number
      ;after prompt "which?"
*SEE         ;three letters after and
      ;before changed part are
      ;displayed
*PART,0      ;five letters of beginning
      ;part of line m are displayed
*PART,100    ;last five letters of line m
      ;are displayed
*PART,K      ;five letters from kth posi-
      ;tion are displayed
*END         ;editing session is completed
  
```

HELLO!! HOW ARE YOU?

ENTER THE VOLUME NUMBER=5

```

*page,200
*lget,10
*rep
  
```

↑ ๕ BY ↑ ๖
8 26 29 33 42 46

```

*WHICH?=33
*see
  
```

๕คุณ/อัน

```

*lget,15
*ins
  
```

↑ ๓ BEFORE ↑ ๖
36 ๖
๖ ๖
๖ ๖

```

page,199
*lget,12
*part,100
  
```

```

*ade
  
```

```

*see
  
```

```

*show,13
  
```

SKIP

```

*show,14
  
```

SKIP

```

*ldel,13
*ling,12
12.5
  
```

↑ ฟ้อง/วากล่าว/พิจารณา/แปล

```

*show,13
  
```

ฟ้อง/วากล่าว/พิจารณา/แปล

```

*ins
  
```

```

*see
  
```

↑ ' BEFORE ↑ ๓ ก
๗ /วากล

```

page,203
*lget,13
*show
  
```

หลวง/สมกำลัง/ผลิตเปลี่ยน
๖/ราชการ/

```

*del
  
```

```

*see
  
```

↑ ๓
๔๐
๖๐๐/๖๐๐

```

*lget,15
*part,100
  
```

กงาน/

Fig. 4 Examples of editing

KWIC making

The most obvious complication is the fact that in Thai writing as many as three separate characters can appear at the same horizontal position in four different vertical positions. Therefore number of letters to take as before or after context must be carefully counted.

As a index of every unit, volume number, page number and line number are attached to the left side.

Computer algorithm

- 1) Every occurrence of pre-positioned vowel (๒ ๓ ๔ ๕) is moved to a position immediately following consonant it precedes.
- 2) Diacritic symbols are moved to the end of word with the indication of position counted from the end of word.
- 3) Each letter is replaced by the code given in Table 1.
- 4) Then two words are compared as if they are numerals.

กะโล่ กะล๑๐๐๐1' 08567146000103

งั๋บ บั๋บ ๐๐๐2' 15571500020300

Sorting

Sorting algorithm of Thai words is not so simple as English.

We ignored algorithm 2), because our segmentation units are not necessarily words so that it does not work effectively.

Table 1 Code table of Thai letter

Thai	ASCII	code												
ก	q	01	ข	G	21	ค	4	41	ฅ	£	61	ด	>	81
ข	H	02	ฅ	P	22	ด	,	42	ค	e	62	ด	>	82
ค	j	03	ด	E	23	ด	p	43	ค	b	63	ด	*	83
ด	h	04	ด	D	24	ด	i	44	ค	u	64	ด	=	84
ด	U	05	ด	,	25	ด	A	45	ค	7	65			85
ด	J	06	ด	R	26	ด	1	46	ค	n	66			86
ด	H	07	ด	<	27	ด	?	47	ค	6	67			87
ด	d	08	ด	I	28	ด	;	48	ค	&	68			88
ด	!	09	ด	f	29	ด	L	49	ค	g	69			89
ด	:	10	ด	9	30	ด	K	50	ค	c	70	SP		90
ด	B	11	ด	5	31	ด	l	51	ค	F	71			91
ด	8	12	ด	m	32	ด	s	52	ค	.	72	-	2	92
ด	Y	13	ด	T	33	ด	>	53	ค	w	73	(	2	93
ด	S	14	ด	o	34	ด	v	54	ค	0	74	)	X	94
ด	\	15	ด	@	35	ด	U	55	ค	Q	75	,	M	95
ด	0	16	ด	x	36	ด	t	56	ค	"	76	*	↑	96
ด	C	17	ด	z	37	ด	y	57	ค	#	77	/	3	97
ด	-	18	ด	/	38	ด	<	58	ค	\$	78			98
ด	W	19	ด	r	39	ด	k	59	ค	%	79			99
ด	+	20	ด	a	40	ด	c	60	ค	<	80			

Statistical data

Total number of letters in the machine readable text is 1362602 which include special symbols such as separator, skip symbol, comma, etc. Total line number is 29582. In Table 2 is shown letter occurrence frequency for each letter. Table 3 shows occurrence frequency of compound vowels. Combination frequency of two letters are listed in Table 4. They are taken in order from the highest frequency. The combination is taken as shown below.

Fig.5 show a distribution of the ratio of upper and lower letters to the total number of letters in a line. Average ratio is 19%. A simple calculation give a ratio of 23% which is number of upper and lower letters among the horizontal positions. This means that in a line of Thai letter upper and lower letters is about 23% of normal horizontal positions.

T=total number of letters in one line

S=total number of upper and lower letters in the line

M=T-S=number of horizontal positions in the line

$$Q1=(S/T) \times 100$$

$$Q2=(S/M) \times 100$$

mean value of Q1=19%

" Q2=23%

Fig. 5

Table 2 Occurrence frequency of single letter

ก	92754	ค	27053	น	14502	ด	4233	ร	369
ข	70137	ท	22768	ก	13916	ต	3169	ล	365
ช	62913	ด	21840	ข	11407	น	3006	ว	276
จ	55392	ค	20112	ท	10844	ก	2938	ม	150
ฎ	41798	น	19316	ด	10739	ข	2652	ฟ	125
ก	41407	พ	18866	ข	10018	ด	1900	ม	104
ง	39532	ะ	18116	ร	8279	ร	1708	ม	63
บ	38624	เ	18049	พ	8069	พ	1493	ท	52
ป	37497	ม	17658	ก	7989	พ	1455	ม	28
อ	37310	ล	17549	ด	7810	ก	1186	ก	18
เ	37019	จ	17421	เ	6303	อ	865	+	17
ห	33185	บ	16903	ข	5661	ข	774	ว	7
ว	29376	ป	15405	ด	4485	ด	698	ก	6
ย	28653	เ	15403	ด	4241	ก	691	อ	5
ณ	27657								

Table 3 Occurrence frequency of compound vowel

- : consonant position

เ-า	9947	-า	2672	เ-อ	545	เ-	55	เ-ะ	0
-า	9268	-า	2622	เ-า	412	เ-	49	เ-า	0
เ-อ	5020	เ-า	2134	เ-	406	เ-อ	22	เ-า	0
เ-ย	3617	เ-	1885	เ-า	339	เ-	8	เ-	0
-อ	3434	เ-ย	1067	เ-า	235	-อ	6	เ-	0
-อ	3228	เ-ย	1056	เ-	107	เ-	2	เ-	0
-า	3085	เ-ย	955	เ-	90	เ-	0		

Table 4 Occurrence frequency of connected letters

/ : segmentation symbol, ห means ห , SP : space

น/	33472	น	8248	น	6034	น	4488	น	3599
/ SP	28370	น/	8173	น	5943	/ค	4482	น	3480
า/	24672	/น	8025	/น	5930	น	4428	น	3479
ง /	21972	น	7856	น	5928	น	4427	น	3470
/เ	18666	น	7854	น	5798	น	4325	น/	3447
ย/	14322	น	7670	/ค	5606	น	4286	น	3442
ด/	13239	/น	7584	น	5370	น/	4162	น	3430
น	11695	/น	7498	/น	5310	น/	4118	/ค	3400
น	11636	/เ	7363	/อ	5176	/น	4094	น	3373
น	11508	น	7330	/จ	5086	น	4066	น	3368
น/	10842	น/	7110	น	5016	น	4034	น	3350
ก/	10511	น	7021	น	4997	น	3998	น	3339
/แ	9924	/ส	6829	น	4973	น	3966	น	3339
/เ	9692	น	6756	น	4918	/,	3950	SPน	3337
น	9305	น	6676	น	4860	น	3925	น/	3249
น	9088	น	6484	น	4732	SPแ	3908	/น	3204
น	8836	น	6398	น	4705	น	3810	น	3179
/น	8827	น	6395	น	4611	น	3774	น	3173
/ก	8752	น	6299	น	4601	น	3663	SPน	3160
น/	8494	น	6285	น	4508	น	3651	น	3148

Printing

Laser beam printer

CANON LBP-3500 is a laser beam printer which can print out any kind of figure and characters. In a character mode, character must be defined as a dot matrix of 8X8, 16X16, 24X24, 32X32, etc.

We use 16X16 matrix as a minimum module of Thai letter pattern. Thai characters are classified into fifteen types from the size of dot matrix. The largest pattern has 48X32 matrix which uses 6 modules.

One text line is printed by five horizontal zone. Each zone has 16 dot vertical width. The horizontal width of each letter can be changed character by character. But in a same zone, vertical size can not be changed.

Control of different letter width

The complex part of output program is to control the width so that heading part of KWIC index come in a line vertically.

An example of KWIC index is shown in Fig.6. We have printed about 200000 lines.

Reference

- 1) Ishii, Yoneo
1969 "Introductory remarks on the Law of Three Seals", East Asian Study, Vol.6, No.4, Kyoto University.
- 2) Murdock, George P.
1971 "Outline of cultural materials" Human Relations Area Files, Inc.
- 3) Oikawa, Akifumi & Nakayama, Kazuhiko & Sugita, Shigeharu
1979 "Printing of Thai letters by laser beam printer", the 20th annual meeting of information processing society of Japan
- 4) Sugita, Shigeharu
1979 "Computer use in ethnological studies", Bulletin of the National Museum of Ethnology, Vol.4, No.1
- 5) Udom Warotamasikkhadit & David Londe
"Computerized Alphabetization of Thai"

๔/๓๓๘/๑๓	ถ้า/ระชะบ้านทางไกลกัน/อยู่/เป็น /ทางชั่วเที่ยง	/ก็ดี/วันหนึ่ง/ก็ดี/จึง/จะ/ถึง/บ้านค่างหน้า/นั้น/แล้ว
๒/๐๕๒/๐๖	กความ/ว่า/แก่/พญาน/ให้/ว่าคาม/รู้ตาม/เห	/ก็ดี/ว่า/แต่/ตามจริง/ก็ดี/ ท่านว่า/จะ/เอา/เป็น/เจ
๒/๐๕๒/๐๗	า/พญาน/กล่าว/คำ/แก่/ผู้อ้าง/ก็ดี/ ผู้คู่ความ	/ก็ดี/ว่า/มี/ผู้รู้/เห็น/ผู้อ้าง/กล่าว/ข้อ/เนื้อความ/ขึ้น/เจ
๒/๑๓๔/๑๓	ร่วมกัน/จะ/ว่าความ/ต่อกัน/โจท/ก็ดี/จำเลย	/ก็ดี/ว่า/หา/สมุด/คืน/สอ/มิได้/ใช้/ ท่านว่า/ให้/เอา/
๒/๑๐๓/๑๕	๖ มาตราหนึ่ง/ วิวาท/กัน/ใน/สถาน /แห่ง/มีค	/ก็ดี/วิวาท/กัน/ใน/ทาง/สาม/เพรง/ก็ดี/ หา/สักขีพ
๕/๓๓๘/๐๕	ง/ด้วย/วจา/ก็ดี/พะยัก/เอา/ด้วย/กาย/หน้าคา	/ก็ดี/ศีล/มุสา/ขาด/ ถ้า/แต่/คิด/ใน/ใจ/มิได้/ขาด/ขาด
๕/๓๓๑/๑๘	าง/ตาย/ โดย/อันน้อย/แต่/เรือก/แล/เสน/ตาย	/ก็ดี/ศีล/นั้น/ขาด/ แล/ศีล/ปานาคี/ปาตา/จะ/ขาด/ต่อ
๓/๒๗๒/๐๕	กกล่าว/ให้/เป็น/ค่านับ/ แล/เอา/ของ /ไป/ขาย	/ก็ดี/ส่ง/ไป/ไกล/ก็ดี/ อยู่มา/เจ้า/ของ/มา/ภบ/เห/จี
๑/๑๗๕/๐๖	ขมกัน/ให้/ตาย/ก็ดี/เป็น/กระสือ/กิน /กัน/ให้/ตาย	/ก็ดี/สรรพ/กระทำ/ให้/ตาย/ก็ดี/ หา/กัน/ว่า/จับ/หอ
๓/๑๐๕/๐๓	/ใด/ลง/ก็ดี/ ไถ/เหนือ/ทราย/ก็ดี/เหนือ/เชิง/ข้ง	/ก็ดี/สวน/แท้/แพ้/จริง/ใช้/ ให้/ใหม่ ๓ รอย/ๆ ละ ๒

volume page line

Fig. 6 Example of KWIC index of the Three Seals Law