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 anthlopology,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 printted 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 recentry,at least in Japan,most cases of computer processing of natural language are distored to indo-europian 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 transfered 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 contexial data in case of OCR.

Segmentation

There is no segmentation problems in case of indo-europian 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

오 늘 밤 나 무 사 온 다

오 늘 밤 나 무 사 온 다
오 늘 밤 나 무 사 온 다
오 늘 밤 나 무 사 온 다
오 늘 밤 나 무 사 온 다
오 늘 밤 나 무 사 온 다
오 늘 밤 나 무 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.

Inputting

Terminal

We use a popular intelligent graphic terminal TEKTRONIX 4051 which has usual alphabet keyboard. We sticked Thai letter labels on the side of each key as if it looks like Thai typewriter. A code table of Thai letters and coresponding 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 expediencies are adopted.

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

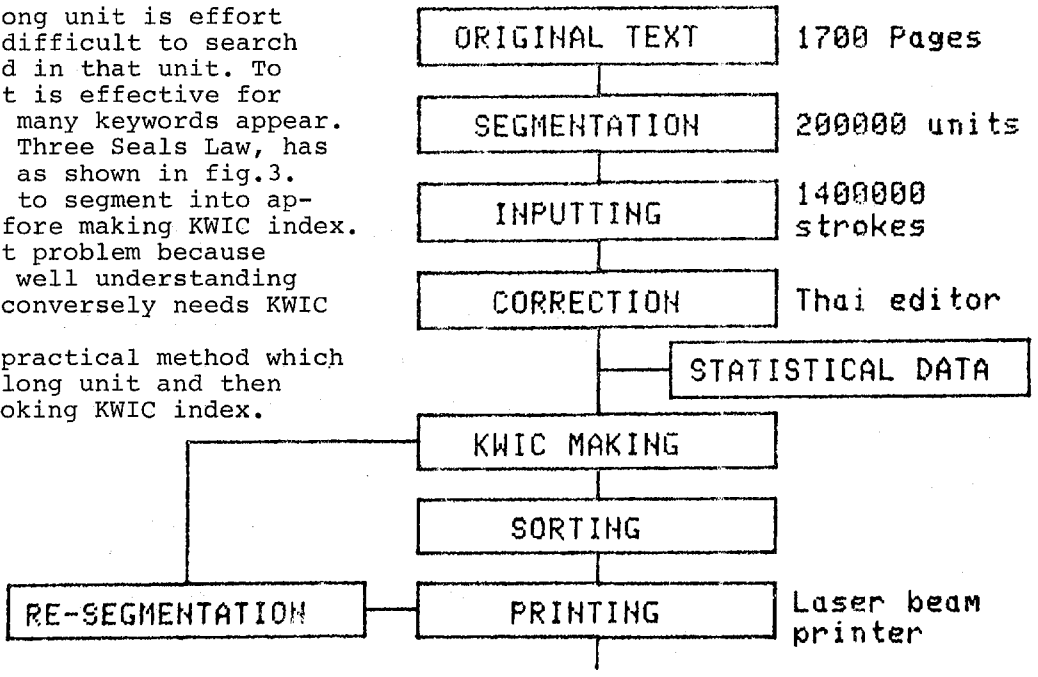


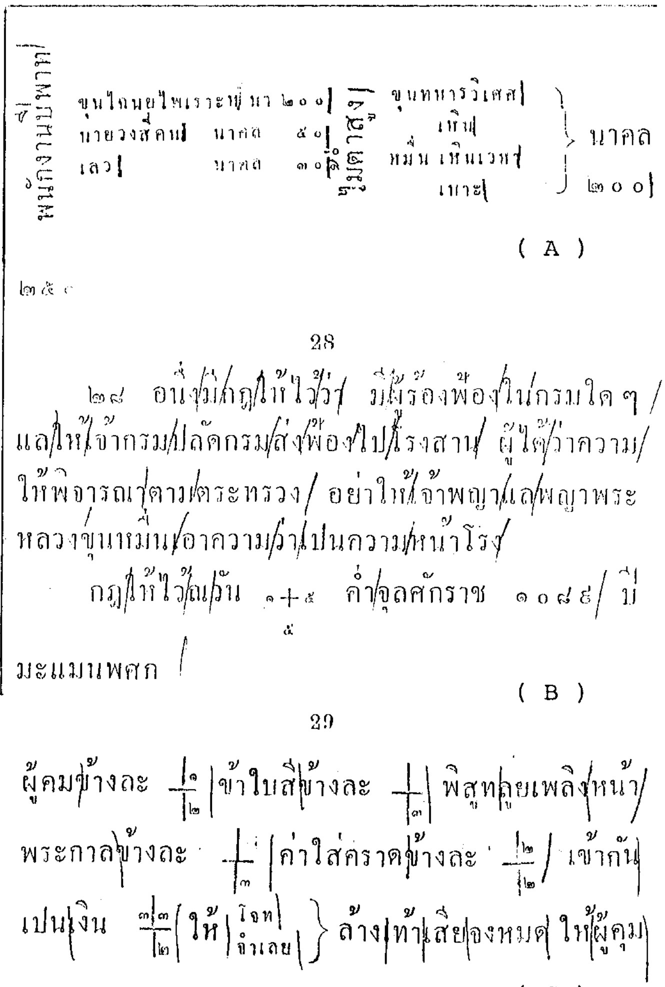
Fig.2 Flow diagram of KWIC making

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.

รู่ → ร ู ่
นี้ → น ี ้
เล้าเบี้ย → เ ล ้ า เ บ ี ้ ย



(A)

(B)

(C)

Example text

```
8  ↑๓พนักงานปี่พาท๓↑
9  ขุนไถนยไพเราะห์/ นา ๒๐๐/
10 นายวงสี่คน/ นาคล ๔๐/
11 เลว/ นาคล ๓๐/
12 ↑๓ไม้ต่ำสูง๓↑
13 (ขุนทหารวิเศศ/,
14  เทิน/,
15 หมื่น เทินเวหา/,
16  เหาะ/) นาคล,๒๐๐/
```
(a)

```
VOL.=4   PAGE=250
1  ๒๘
2  ๒๘ อนึ่ง/มี/กฎ/ให้ไว้/ว่า/ มี/ผู้ร้องฟ้อง/ใน/กรมใด ๆ/
3  แล/ให้/เจ้ากรม/ปลัดกรม/ส่ง/ฟ้อง/ไป/โรงสาน/ ผู้ได้/ว่าความ/
4  ให้พิจารณา/ตาม/ตระหรวง/ อย่าให้/เจ้าพญา/แล/พญาพระ
5  หลวง/ขุนหมื่น/เอาความ/ว่า/เปนความ/หน้าโรง/
6  กฎ/ให้ไว้/ณ/วันอาทิตย์ เดือน ๔ แรม ๔ ค่ำ/จุลศักราช ๑๐๘๙/
```
(b)

```
P=109(CONT)
 ง/หน้า/
16 พระกาล/ข้างละ ๓ สลึง/ ค่าใส่คราด/ข้างละ ๒ บาท ๒ สลึง/ เข้า
 ัน/
17 เปน/เงิน/ ๓ ตำลึง ๓ บาท ๒ สลึง/ ให้/(โจท/,จำเลย/) ล้าง/เห
 า/เสีย/จงหมด/ ให้/ผู้คุม/
```
(c)

Fig. 3 Examples of text and inputted form

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 suporse 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
```

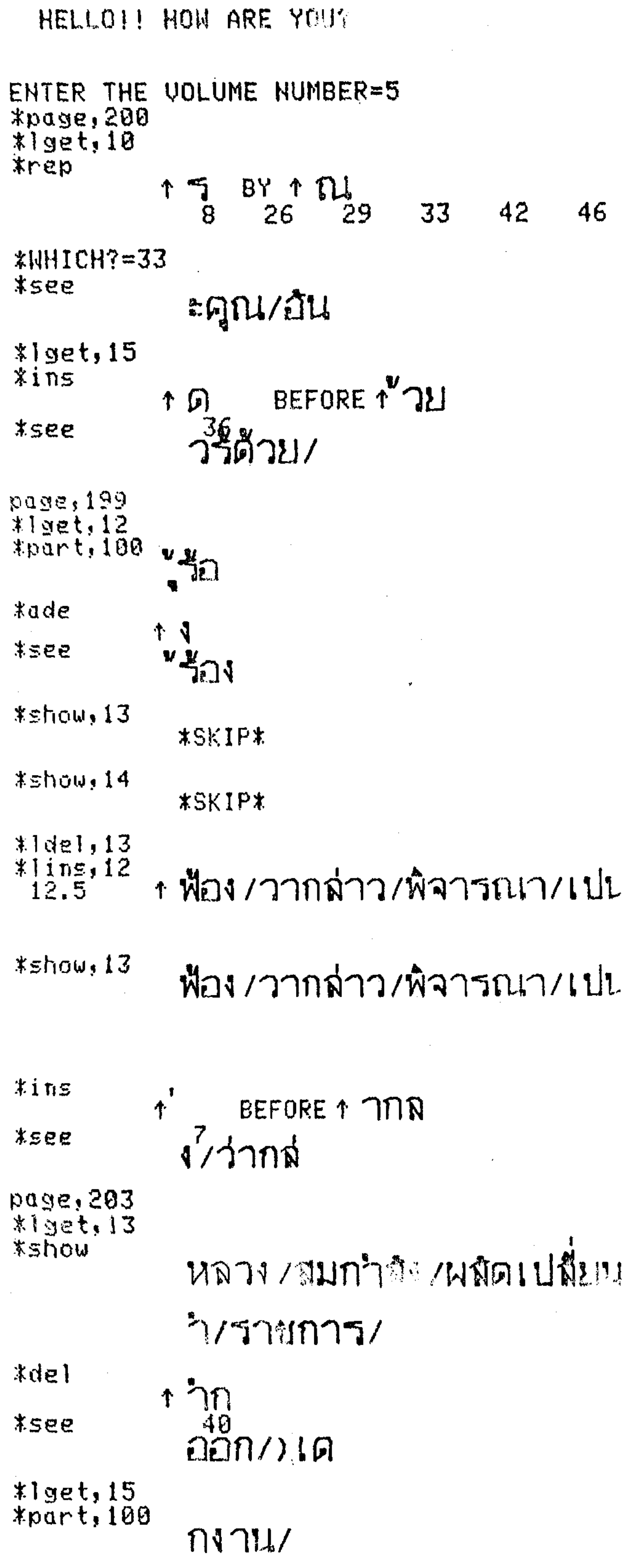


Fig. 4 Examples of editing

KWIC making

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

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

Sorting

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

Computer algorithm

1) Every occurrence of pre-positioned vowel (เ แ โ ใ ไ) is moved to a position immediately following consonant it preceeds.
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	กะลโ0001'	08567146000103
งั่ง	งัง 0002'	15571500020300

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

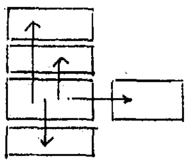
Table 1 Code table of Thai letter

Thai	ASCII	code	Thai	ASCII	code	Thai	ASCII	code	Thai	ASCII	code	Thai	ASCII	code
ๆ	q	01	ฌ	G	21	ภ	4	41	ํ	[	61	[illegible]	)	81
็	H	02	ญ	P	22	ม	,	42	ำ	e	62	[illegible]	}	82
่	j	03	ฎ	E	23	ย	p	43	ิ	b	63	[illegible]	*	83
้	h	04	ฏ	D	24	ร	i	44	ี	u	64	[illegible]	=	84
๊	U	05	ฐ	`	25	ฤ	A	45	ึ	7	65			85
๋	J	06	ฑ	R	26	ล	1	46	ื	n	66			86
์	N	07	ฒ	<	27	ฦ	?	47	ุ	6	67			87
ก	d	08	ณ	I	28	ว	;	48	ู	&	68			88
[illegible]	!	09	ด	f	29	ศ	L	49	เ	g	69		.	89
ข	:	10	ต	9	30	ษ	K	50	แ	c	70	SP		90
ฃ	B	11	ถ	5	31	ส	l	51	โ	F	71			91
ค	8	12	ท	m	32	ห	s	52	ใ	.	72	–	2	92
ฅ	Y	13	ธ	T	33	ฬ	>	53	ไ	w	73	(	Z	93
ฆ	S	14	น	o	34	อ	v	54	ฯ	0	74	)	X	94
ง	`	15	บ	@	35	ฮ	V	55	ฺ	Q	75	,	M	95
จ	0	16	ป	x	36	ะ	t	56	[illegible]	"	76	*	↑	96
ฉ	C	17	ผ	z	37	ั	y	57	[illegible]	#	77	/	3	97
ช	–	18	ฝ	/	38	[illegible]	´	58	[illegible]	$	78			98
ซ	W	19	พ	r	39	า	k	59	[illegible]	%	79			99
[illegible]	+	20	ฟ	a	40	ๅ	{	60	[illegible]	<	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 culculation 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)X100
Q2=(S/M)X100

mean value of Q1=19%
" Q2=23%

Fig. 5

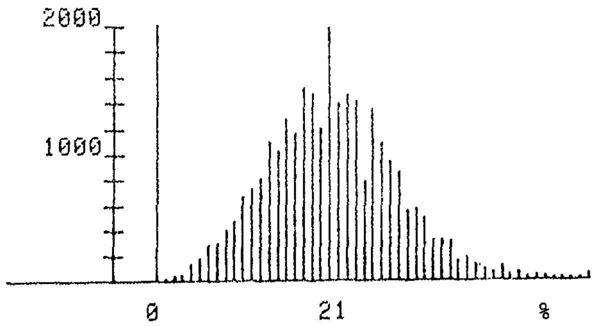


Table 2 Occurrence frequency of single letter

า	92754	ด	27053	ุ	14502	ศ	4233	ฯ	369
น	70137	ท	22768	ค	13916	ธ	3169	ฐ	365
้	62913	ี	21840	ช	11407	ู	3006	[illegible]	276
ร	55392	ั	20112	[illegible]	10844	ภ	2938	ฏ	150
่	41798	ส	19316	ื	10739	ษ	2652	ฬ	125
ก	41407	พ	18866	ข	10018	ฎ	1900	ฅ	104
ง	39532	ะ	18116	ำ	8279	์	1708	ฒ	63
ม	38624	ไ	18049	ผ	8069	ฟ	1493	[illegible]	52
ิ	37497	ต	17658	ถ	7989	ฝ	1455	ฌ	28
อ	37310	แ	17549	็	7810	ๆ	1186	ฤ	18
เ	37019	จ	17421	ใ	6303	ฉ	865	+	17
ห	33185	บ	16903	[illegible]	5661	ฑ	774	๋	7
ว	29376	ป	15405	ณ	4485	[illegible]	698	[illegible]	6
ย	28653	โ	15403	ญ	4241	[illegible]	691	ฮ	5
ล	27657								

Table 3 Occurrence frequency of compound vowel

- : consonant position

เ-า	9947	-ัย	2672	เ-ือ	545	เ-ีย	55	เ-ือะ	0
-ัย	9268	-าว	2622	เ-ียว	412	ำย	49	เ-ีว	0
เ-ือ	5020	แ-ว	2134	็ว	406	เ-ือย	22	เ-ียะ	0
เ-ีย	3617	เ-ะ	1885	เ-ว	339	ฦๅ	8	แ-ะ	0
-ือย	3434	เ-ย	1067	เ-าะ	235	-ือ	6	โ-ะ	0
-ือ	3228	ไ-ย	1056	ฤๅ	107	เ-ะ	2	-ัวะ	0
-ัว	3085	ใ-ย	955	แ-ะ	90	เ-ือะ	0		

Table 4 Occurrence frequency of connected letters

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

น/	33472	แล	8248	นน	6034	ตร	4488	อย	3599
/ SP	28370	ล/	8173	าม	5943	/ค	4482	ย์	3480
า/	24672	/ท	8025	/พ	5930	ในง	4428	ม่	3479
ง /	21972	องค์	7856	นา	5928	ได	4427	มน	3470
/เ	18666	พร	7854	กร	5798	ลา	4325	ป/	3447
ย/	14322	วา	7670	/ต	5606	ม์	4286	ไป	3442
ด/	13239	/น	7584	อน	5370	ท/	4162	ปร	3430
ห์	11695	/พ	7498	/ท	5310	อ/	4118	/ด	3400
ให	11636	/ใ	7363	/อ	5176	/ข	4094	น์	3373
ระ	11508	าร	7330	/จ	5086	น'	4066	ลว	3368
ม/	10842	ว/	7110	หม	5016	ถ์	4034	ขา	3350
ก/	10511	น ็	7021	จา	4997	ว'	3998	ม์	3339
/แ	9924	/ล	6829	กา	4973	ก'	3966	แก	3339
/ใ	9692	านน	6756	ปน	4918	/,	3950	SPถ	3337
รา	9305	ด์	6676	ถา	4860	าช	3925	ป/	3249
าย	9088	มา	6484	เป	4732	SPแ	3908	/ป	3204
นห์	8836	อา	6398	ทา	4705	าก	3810	ล'	3179
/ม	8827	ฟ์	6395	เอ	4611	ด์	3774	หา	3173
/ก	8752	ฟ	6299	ร ็	4601	าง	3663	SPข	3160
ร/	8494	หน	6285	ท'	4508	ะ /	3651	าม	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 anual 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/ระยะบ้านทางไกลกัน/อยู่/เปน /ทางชั่วเที่ยง /ก็ดี/วันหนึ่ง/ก็ดี/จึ่ง/จะ/ถึง/บ้านค่างหน้า/นั้น/แล้ว

๒/๐๕๒/๐๖ กความ/ว่า/แก่/พญาณ/ให้/ว่าตาม /รู้ตาม/เหน /ก็ดี/ว่า/แต่/ตามจริง/ก็ดี/ ท่านว่า/จะ/เอา/เปน/เจ

๒/๐๕๒/๐๗ า/พญาณ/กล่าว/คำ/แก่/ผู้อ้าง/ก็ดี / ผู้คู่ความ /ก็ดี/ว่า/มี/ผู้รู้เหน/ผุ้อ้าง/กล่าว/ข้อ/เนื้อความ/ชี้เจ

๒/๑๓๔/๑๓ ร้อมกัน/จะ/ว่าความ/ต่อกัน/โจท /ก็ดี/จำเลย /ก็ดี/ว่า/หา/สมุด/ดินสอ/มิได้/ใช้/ ท่านว่า/ให้เอา/

๒/๑๐๓/๑๕ ๖ มาตราหนึ่ง/ วิวาท/กัน/ใน/สถาน /แห่ง/มิด /ก็ดี/วิวาท/กัน/ใน/ทาง/สามแพร่ง/ก็ดี/ หา/สักขิพ

๕/๓๓๘/๐๕ ง/ด้วย/วาจา/ก็ดี/พะยัก/เอา/ด้วย /กาย/หน้าตา /ก็ดี/ศิลมุสา/ขาด/ ถ้า/แต่/คิด/ในใจ/มิได้/ขวลขวา

๕/๓๓๑/๑๘ าง/ตาย/ โดย/อันน้อย/แต่/เรือด/แล /เสน/ตาย /ก็ดี/ศิล/นั้น/ขาด/ แล/ศิลปานาติปาตา/จะ/ขาด/ต่

๓/๒๗๒/๐๕ กกล่าว/ให้เปน/คำนับ/ แล/เอา/ของ /ไป/ขาย /ก็ดี/ส่ง/ไป/ไกล/ก็ดี/ อยู่มา/เจ้าของ/มา/ภบเหน/จึ่

๑/๑๗๕/๐๖ ขบกัน/ให้ตาย/ก็ดี/เปน/กระสือ/กิน /กัน/ให้ตาย /ก็ดี/สรรพ/กระทำ/ให้ตาย/ก็ดี/ หา/กัน/ว่า/จับ/หอ

๓/๑๐๕/๐๓ /ไถ/ลง/ก็ดี/ ไถ/เหนือ/ตรุย/ก็ดี /เหนือ/เชิงชัง /ก็ดี/สวนแท้/แพ้จริง/ใช้/ ให้ไหม ๓ รอย/ๆ ละ ๒

volume page line

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