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.

Acknowledgements

This report is a result of joint project at National Museum of Ethnology. The member are Y.Ishii, Y.Kaga, 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 Chitthisatcha, 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 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 contextual 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.

Examples shown below mean that there exist several different segmentation. Fig.1 Examples of different segmentations

Segmentation affects to the meaning of sentence and retrieval efficiency.

Examples shown below mean that there exist several different segmentation. Segmentation affects to the meaning of sentence and retrieval efficiency.

Examples shown below mean that there exist several different segmentation.

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 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 casset 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

ORIGINAL TEXT

SEGMENTATION

INPUTTING

CORRECTION

THAI EDITOR

STTISTICAL

DATA

KWIC MAKING

SORTING

PRINTING

Laser beam printer

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.


Example text

VOL. 4 PAGE=250

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 supposes 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 subcommands
*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 position are displayed
*END

editing session is completed

HELLO!! HOW ARE YOU?

ENTER THE VOLUME NUMBER=5

page,200

lget,10

rep

↑ ร ป BY ↑ ณ

 WHICH?=33

see

lget,15

lins

↑ ก BEFORE ↑ ว

see

page,199

lget,12

lpart,100

lade

see

lsee

show,13

*skip*

show,14

*skip*

lget,13

lline,12

12.5↑ พื้น/วaghลรู/พิษารน/wเป

show,13

พื้น/วaghลรู/พิษารน/wเป

show,13

lins

↑ ก BEFORE ↑ ก

see

page,203

lget,13

show

show

lget,15

lpart,100

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.

Sorting

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

Computer algorithm
1) Every occurrence of pre-positioned vowel (ิ ุ อ ิ) 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.

Table 1 Code table of Thai letter

| Thai | ASCII | code |
|------|-------|------|
| ฃ ะ | ถ ะ | ถ ะ |
| ร ะ | ร ะ | ร ะ |
| ผ ะ | ผ ะ | ผ ะ |
| ฝ ะ | ฝ ะ | ฝ ะ |
| ด ะ | ด ะ | ด ะ |
| ต ะ | ต ะ | ต ะ |
| ท ะ | ท ะ | ท ะ |
| น ะ | น ะ | น ะ |
| ไ ะ | ไ ะ | ไ ะ |
| อ ะ | อ ะ | อ ะ |
| ย ะ | ย ะ | ย ะ |
| ร ะ | ร ะ | ร ะ |
| ช ะ | ช ะ | ช ะ |
| ซ ะ | ซ ะ | ซ ะ |
| ฌ ะ | ฌ ะ | ฌ ะ |
| ญ ะ | ญ ะ | ญ ะ |
| ป ะ | ป ะ | ป ะ |
| บ ะ | บ ะ | บ ะ |
| ผ ะ | ผ ะ | ผ ะ |
| ฝ ะ | ฝ ะ | ฝ ะ |
| ง ะ | ง ะ | ง ะ |
| จ ะ | จ ะ | จ ะ |
| ฉ ะ | ฉ ะ | ฉ ะ |
| ช ะ | ช ะ | ช ะ |
| ซ ะ | ซ ะ | ซ ะ |
| ฌ ะ | ฌ ะ | ฌ ะ |
| ญ ะ | ญ ะ | ญ ะ |
| ป ะ | ป ะ | ป ะ |
| บ ะ | บ ะ | บ ะ |
| ผ ะ | ผ ะ | ผ ะ |
| ฝ ะ | ฝ ะ | ฝ ะ |
| ง ะ | ง ะ | ง ะ |
| จ ะ | จ ะ | จ ะ |
| ฉ ะ | ฉ ะ | ฉ ะ |
| ช ะ | ช ะ | ช ะ |
| ซ ะ | ซ ะ | ซ ะ |
| ฌ ะ | ฌ ะ | ฌ ะ |
| ญ ะ | ญ ะ | ญ ะ |
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.

Table 2 Occurrence frequency of single letter

| Letter | Frequency |
|--------|-----------|
| 92754  | 27053     | 14502 | 4233 | 369 |
| 70137  | 22768     | 13916 | 3169 | 365 |
| 62913  | 21840     | 11407 | 3006 | 276 |
| 55392  | 20112     | 10844 | 2930 | 150 |
| 41798  | 19316     | 10739 | 2652 | 125 |
| 41407  | 18866     | 10018 | 1900 | 104 |
| 39532  | 18116     | 8275  | 1708 | 63  |
| 38624  | 18049     | 8069  | 1493 | 52  |
| 37497  | 17658     | 7989  | 1455 | 28  |
| 37310  | 17549     | 7810  | 1186 | 18  |
| 37819  | 17421     | 6303  | 865  | 17  |
| 33185  | 16903     | 5661  | 774  | 7   |
| 29376  | 15405     | 4485  | 698  | 6   |
| 28653  | 15403     | 4241  | 691  | 5   |
| 27657  |           |       |      |     |
Table 3  Occurrence frequency of compound vowel
- : consonant position

|     | 1-ך | 1-ן | 1-ך | 1-ך | 1-ך |
|-----|-----|-----|-----|-----|-----|
| 1-ך | 9947| 2672| 545 | 53  | 0   |
| 1-ן | 9268| 2622| 412 | 49  | 0   |
| 1-ך | 5020| 2134| 406 | 22  | 0   |
| 1-ך | 3617| 1885| 339 | 8   | 0   |
| 1-ך | 3434| 1067| 235 | 6   | 0   |
| 1-ך | 3228| 1056| 107 | 2   | 0   |
| 1-ך | 3085| 955 | 90  | 0   | 0   |

Table 4  Occurrence frequency of connected letters
/: segmentation symbol, 造血 means 造血,  SP : space

|     | 33472| 8248 | 6034 | 4488 | 3599 |
|-----|-----|-----|-----|-----|-----|
| 1-ך | 28370| 8173 | 5943 | 4482 | 3480 |
| 1-ך | 24672| 8025 | 5930 | 4428 | 3479 |
| 1-ך | 21972| 7856 | 5928 | 4427 | 3470 |
| 1-ך | 18666| 7854 | 5798 | 4325 | 3447 |
| 1-ך | 14322| 7670 | 5606 | 4286 | 3442 |
| 1-ך | 13239| 7584 | 5370 | 4162 | 3430 |
| 1-ך | 11695| 7498 | 5310 | 4118 | 3400 |
| 1-ך | 11636| 7363 | 5176 | 4094 | 3373 |
| 1-ך | 11508| 7330 | 5086 | 4066 | 3368 |
| 1-ך | 10842| 7110 | 5016 | 4034 | 3358 |
| 1-ך | 10511| 7021 | 4997 | 3998 | 3339 |
| 1-ך | 9924 | 6829 | 4973 | 3966 | 3339 |
| 1-ך | 9692 | 6756 | 4918 | 3958 | 3337 |
| 1-ך | 9305 | 6676 | 4868 | 3925 | 3249 |
| 1-ך | 9088 | 6484 | 4732 | 3908 | 3204 |
| 1-ך | 8836 | 6398 | 4705 | 3890 | 3204 |
| 1-ך | 8827 | 6395 | 4688 | 3873 | 3204 |
| 1-ך | 8752 | 6299 | 4601 | 3663 | 3160 |
| 1-ך | 8494 | 6285 | 4503 | 3651 | 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 zones. 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 200,000 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, Kazuhiro & Sugita, Shigebaru 1979 "Printing of Thai letters by laser beam printer", the 20th annual meeting of information processing society of Japan

4) Sugita, Shigebaru 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"