Southeast Asian Studies, Vol. 25, No. 2, September 1987

Input/Output Methods for Thai

——Development of a Database and a Computer Concordance for the Three Seals Law of Thailand——

Mamoru SHIBAYAMA*

Abstract

An intelligent Thai computer terminal and a Thai text editor with the function of automatic and consecutive conversion from Roman spelling to Thai letters have been developed which are operable on a micro computer. These employ the Transliteration Method (TM) or the Simplified Transliteration Method (STM), which are based on a newly devised transliteration table from Roman spelling. We are now developing a database and a computer concordance of the Three Seals Law (*Kotmai Tra Sam Duang*), and making a machine-readable Thai dictionary using this terminal and editor.

The Transliteration and Simplified Transliteration Methods were both estimated to require a greater number of key strokes in the making of the machine-readable Thai dictionary than the method used for the ordinary IBM electronic Thai typewriter, here called the Direct Mapping Method (DMM). However, an evaluation of learning effects from the number of key strokes and the measurement of learning curves in the input of the Thai dictionary indicated that although the Transliteration Method required a 42.9% greater number of key strokes than the Direct Mapping Method, a 9.8% higher input rate in terms of characters per minute.

For the output of Thai letters, the design and implementation of a printing system for a Japanese laser beam printer run from a main-frame computer and a CRT display for a micro computer are described.

I Introduction

In the Center for Southeast Asian Studies, Kyoto University, we are now developing a database for the Three Seals Law (*Kotmai Tra Sam Duang*, compiled in 1805, about 1700 pages, about 32,500 lines in the five-volume version published in 1962) on the main-frame computer in the Data Processing Center, Kyoto University with a view to making a computer concordance, which will provide an important resource for studying the history, law, sociology, and linguistics of Thailand [Ishii 1969]. For processing this Thai text, the input/output methods for Thai text and Thai letters are of major importance and present sophisticated problems.

Two input methods are considered. One is the Direct Mapping Method (DMM) by which each Thai letter corresponds uniquely to one of the keys on the keyboard [Sugita 1980]. This scheme requires an intelligent terminal to identify which Thai letter is input; in other words, the ordinary teletype terminal cannot be used for the input/output

* 柴山 守, The Center for Southeast Asian Studies, Kyoto University

of Thai letters because it is impossible to display/print Thai letters on the terminal. The other method is a transliteration approach using a table by which Thai letters are generated from Roman letters according to their pronunciation, like the Roman-Kanji conversion in Japanese, and is a method, here called Hartmann's Transliteration Method (HTM), proposed by J. F. Hartmann and G. M. Henry [Hartmann and Henry 1983]. This approach requires a greater number of key strokes for input than the Direct Mapping Method, but it is readily operable by non-native speakers of Thai. Also, it does not require an intelligent terminal if the transliteration approach is used for displaying Thai letters and if the main-frame computer performs the function of transliteration.

In 1984, we implemented a Thai text editor in which we adopted the DMM as one of the input methods [Shibayama *et al.* 1984]. In 1985, we modified the table proposed by J. F. Hartmann *et al.* to produce a new transliteration table, here called the Transliteration Method (TM), which was incorporated into the text editor [Shibayama *et al.* 1985]. We have also developed a method Roman-Thai conversion called the Simplified Transliteration method (STM) [Shibayama and Hoshino 1987].

For output, we have developed a display system of Thai letters on the micro computer and a printing system of Thai letters on a laser beam printer run from the main-frame computer at the Data Processing Center, Kyoto University.

This paper describes the characteristics of the DMM and the TM implemented on the Thai text editor, compares the DMM, TM, and HTM by estimating the number of key strokes required to input text of the Three Seals Law, and shows the results of typing speed and learning curves measured for the work of inputting the main entries of a Thai-Thai dictionary by the DMM and TM methods. The basic idea of the STM, which should be more readily operable for non-native speakers of Thai, is also presented.

For output, the characteristics and structure of display/printing controls in the system are described.

Lastly, an appendix presents an outline for developing a database and a computer concordance of the Three Seals Law and making a machine-readable Thai dictionary.

II Characteristics of Thai

The Thai writing system differs from that of western languages in several points:

(a) Thai letters are phonetic. Thai has 5 tones. Each Thai syllable is composed either of *consonant*+*vowel* or *consonant*+*vowel*+*consonant*.

(b) The consonants, vowels, and tonal marks must be positioned appropriately.

(c) Words in a sentence are not separated from each other.

(d) Vowels may be placed before, after, above, or below a consonant.

(e) Punctuation is scarcely used.

Fig. 1 shows an example of Thai script printed by our Thai printing system using the laser beam printer. Given the charac-

530807 ทรงพระราชดำริ/เหนว่า/ บท/พระอายการ/ซึ่ง/ให้เอาโทษ/
530808 แก่/ผู้ทำทอง(พราง/.เคลือบ/.อาบ/) แล/ผู้รู้เหน/เปนใจ/ช่วย/ขาย/ช่วย/
530809 จำนำ/นั้น/ ก็/ควรด้วย/โทษ/อยู่แล้ว/ แต่ทะว่า/ให้เรียก
530810 เอา/เงินทุน/ให้แก่/ผู้นั้น/ ครั้น/จะ/ยืน/คงไว้/ตาม/บท/พระ
530811 อายการ/ ฝ่ายผู้มีทรัพย/ซึ่ง/รับ(ซื้อ/.ขาย/.จำนำ/) นั้น/ หา/เขด
530812 หลาบ/ไม่/ ด้วย/ใจโลภ/เจตนา/มิได้/พินิจ/พิจารณา/ก็/จะ/
530813 (ซื้อหา/.รับจำนำ/)ไว้/อีก/ ถึง/เปน/ทอง(เคลือบ/.อาบ/) ก็/ไว้ใจ/ว่า/จะ/

Fig. 1 Example of Thai Letters: A Part of Text of the Three Seals Law

teristics of Thai noted above and shown in Fig. 1, several points must be considered in processing Thai text using the computer:

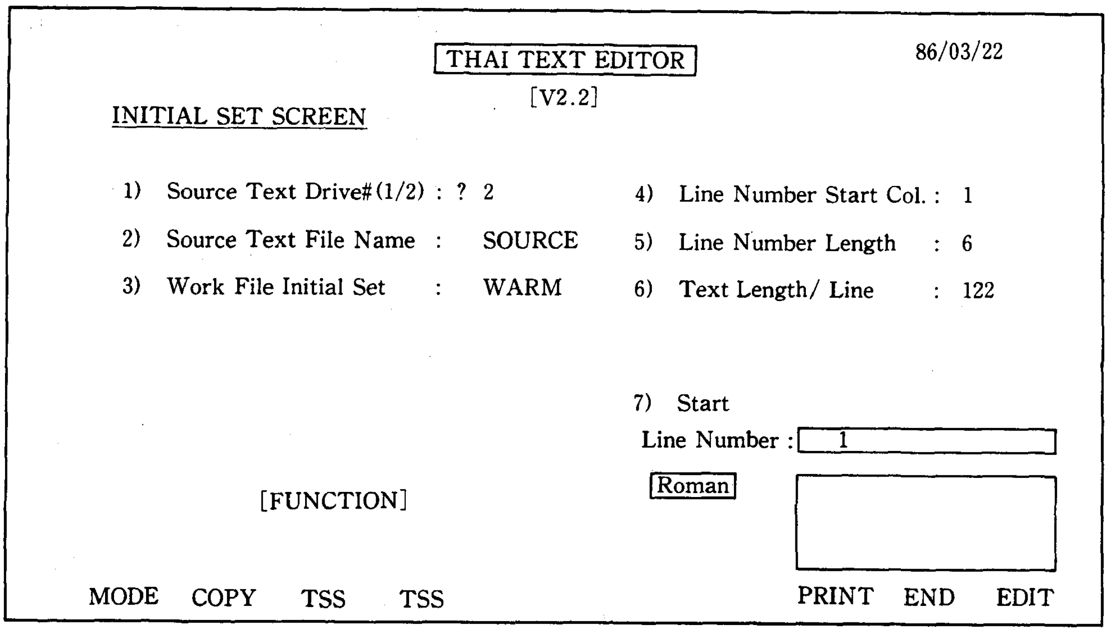
(a) How to input Thai letters.
(b) How to control the display and printing positions for each consonant, vowel, and tonal mark.
(c) How to divide a text into sentences and a sentence into words, namely, segmentation.

Several studies into the complex question of input/output of Thai have been made. For input, typical methods include the keyboard of the ordinary IBM electronic Thai typewriter and the keyboard connected to the computer used in Thailand, which employs the DMM. For inputting Thai text of the Three Seals Law, Sugita adopted the DMM in using a graphic terminal connected to an IBM host computer [Sugita 1980]. Hartmann and Henry, who have been using the computer to study Thai language in the field of library information, have proposed the transliteration approach mentioned in Section 1 [Hartmann and Henry 1983].

For output the typing head of the IBM electronic Thai typewriter and the use of ROM (Read Only Memory) for display and printing Thai letters on the computer are generally used. Sakamoto has developed integrated computer programs by which several Southeast Asian and African can be printed out with a laser beam printer [Sakamoto 1979]. Also, a printing program which can output Thai letters has been developed in the Center for Information Processing of Tsukuba University.

III Thai Text Editor

Fig. 2 shows file control and editing screens of the Thai text editor implemented on a micro computer. This editor employs the functions shown in Table 1. The column "Screen" in Table 1 shows whether the screen is in (a) the file control or (b) the editing mode in Fig. 2. The screen in Fig. 2(b) is in the TM mode described in section IV.*2* and IV.*4.2*, and corresponds to a record, namely, a line in the text which has a maximum of 160 characters in Thai, and which is divided into 4 lines in order to display Thai characters. Below each display line is the area of movement of cursor. The cursor can be moved to any directions



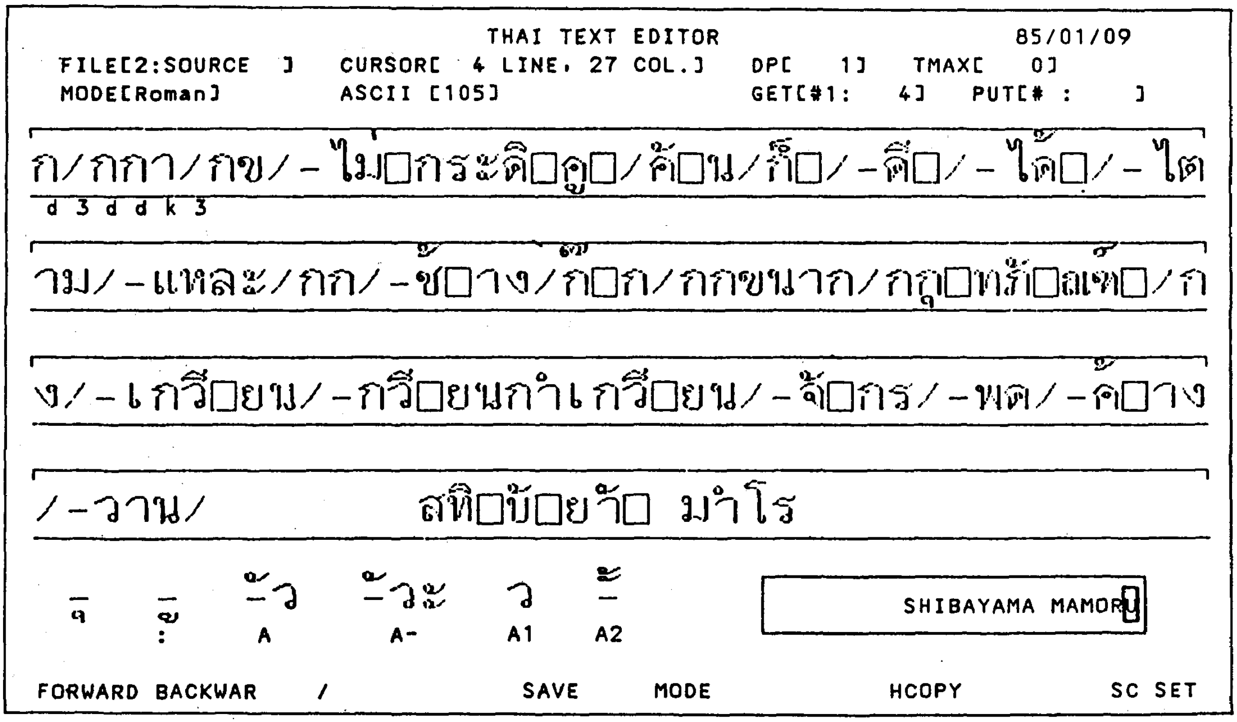
```
                         THAI TEXT EDITOR                 86/03/22
                              [V2.2]
  INITIAL SET SCREEN

  1) Source Text Drive#(1/2) : ? 2        4) Line Number Start Col. : 1
  2) Source Text File Name   :  SOURCE    5) Line Number Length     : 6
  3) Work File Initial Set   :  WARM      6) Text Length/ Line      : 122

                                          7) Start
                                           Line Number : 1
                                          Roman
                    [FUNCTION]

 MODE   COPY   TSS   TSS                         PRINT   END   EDIT
```

(a) File Control Screen



```
                        THAI TEXT EDITOR                         85/01/09
 FILE[2:SOURCE  ]   CURSOR[ 4 LINE, 27 COL.]   DP[   1]   TMAX[   0]
 MODE[Roman]        ASCII [105]                GET[#1:   4]   PUT[# :    ]

ก/กกา/กข/-ไม□กระดิ□ดู□/ค้□น/ก็□/-ดี□/-ได□/-ได
 d 3 d d k 3

าม/-แหละ/กก/-ช□าง/ก□ก/กกขนาก/กฤ□ทรั□ลเพ□/ก

ง/-เกวี□ยน/-กวี□ยนกำเกวี□ยน/-จั้□กร/-พด/-ค□าง

/-วาน/            สทิ□บั□ยา□ มำโร

  ฤ    ฦ    ั-ว    ั-วะ    ว    ะ                SHIBAYAMA MAMORU
       :     A      A-     A1   A2

FORWARD BACKWAR   /          SAVE     MODE          HCOPY         SC SET
```

(b) Editing Screen

Fig. 2 Screens of the Thai Text Editor

using the arrow keys, and Thai characters are displayed using graphic instructions on the micro computer.

We found that a character pattern composed of 16(W)*32(H) mesh stored in RAM (Random Access Memory) can be displayed without appreciable delay by using the *PUT@* statement in the graphic instructions. We synthesized a new pattern on the GVRAM (Graphic Video RAM) by using the *OR* operation for all bits of the patterns shown in Fig. 3. This figure also shows that patterns other than tones should be shifted by 5 dots downward in the *Y* direction from the standard position in order to save the main memory area for storing all patterns.

Table 1 Functions of the Thai Text Editor

Screen	Key	Indication	Function
(a) (b)	f. 1 f. 6	MODE MODE	Input Method is Specified for Thai
(a)	f. 2	COPY	Back-up of Current File is Executed
(a)	f. 3, f. 4	TSS	TSS Emulator is Invoked
(a)	f. 8	PRINT	File Printing
(a)	f. 9	END	Quit the Editor
(b)	f. 1	FORWARD	Move to Next Record
(b)	f. 2	BACKWARD	Move to Previous Record
(b)	f. 3	"/"	"/" is Inserted
(b)	f. 5	SAVE	Editing File is Saved
(b)	f. 8	HCOPY	A Record is Printed
(b)	f. 10	SC SET	Screen (a) is Invoked
(a)	f. 10	EDIT	Editing is Restarted

IV Input Methods

1. *Direct Mapping Method (DMM)*

In the Thai text editor implemented on the micro computer, we adopted the keyboard assignment, here called the Direct Mapping Method (DMM), as shown in Fig. 4. Since this keyboard assignment is almost equivalent to that of the IBM electronic Thai typewriter, and the editor employs the dead key control, whereby a character pattern overlaps the preceding patterns without the carriage moving, so that the consonants, vowels, and tones can be displayed in their appropriate positions on the CRT, the editor can be used like the IBM electronic Thai typewriter.

2. *Transliteration Method (TM)*

The input method of Thai letters by means of the Roman letters representing the Thai pronunciation, namely, the Transliteration Method (TM), has the benefit of improving the operability of typing for non-native speakers of Thai and for people

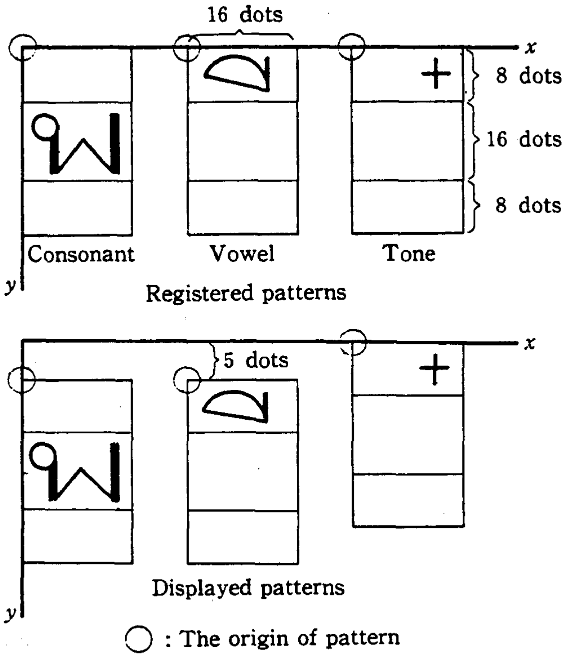
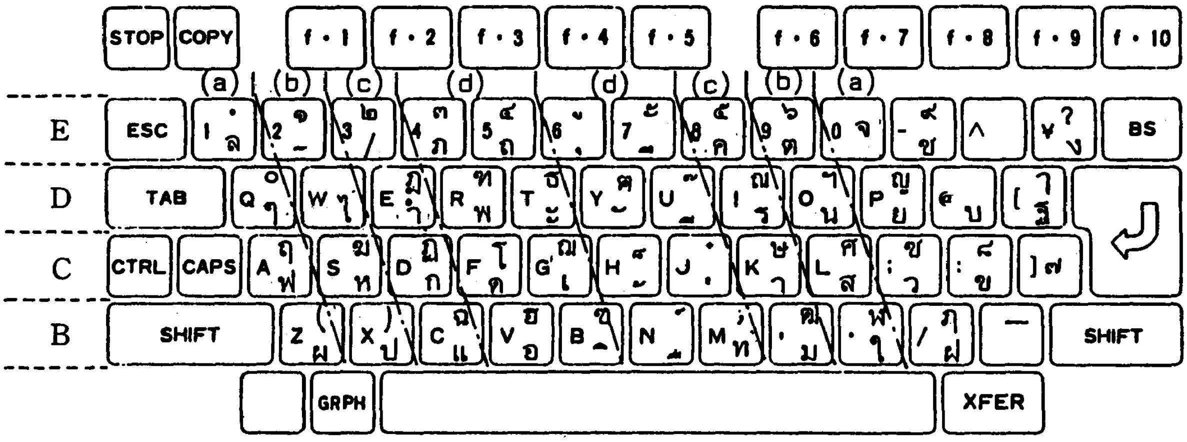


Fig. 3 The Display on the CRT



(a): Little finger　(b): Third finger
(c): Middle finger　(d): Forefinger

Fig. 4 Keyboard Assignment of DMM

accustomed to the normal keyboard assignment of Roman letters. At the same time, the transliteration table should be simple for typists and must be designed to decrease the number of key strokes and the sphere of movement of the fingers. To this end, we have proposed the revised transliteration table from Roman to Thai letters shown in Table 2 and implemented a function capable of automatic and consecutive conversion according to this table.

The characteristics of the transliteration table are as follows:

(a) The Roman spellings of the Thai consonants and vowels are classified into 21 groups according to their pronunciations, each group comprising character strings headed by same Roman character. The numerals, tones, special symbols, and control codes are classified into 28 groups. A group number, *GN*, is assigned to each of these 49 groups, and each character string in a group is discriminated by a local classification number, *LCN*.

(b) The transliteration approach by Hartmann and Henry distinguishes different Thai letters with the same pronunciation by use of apostrophes, for example, *TH, TH', TH'', TH''', TH'''',* and *TH'''''*.

In our system, the distinction is represented by adding a number to the Roman spelling. The Thai letters are arranged, moreover, in order of decreasing frequency of occurrence in the text of the Three Seals Law. In this way, the number of key strokes required by the operator is decreased. For example, *TH, TH1, TH2, TH3, TH4,* and *TH5* are used for ท, ถ, ธ, ฐ, ฑ, and ฒ, instead of *TH, TH', TH'', TH''', TH'''',* and *TH'''''*. An advantage of this scheme is that the ordinary teletype terminal can be used for input/output of Thai letters if the function of interconversion of Roman and Thai letters is implemented on the host computer.

Table 2 Transliteration Table

(a) Consonants

GN	LCN 1	2	3	4	5	6	7	8
1	K	KH	KH1	KH2	KH3	KH4		
	ก	ค	ข	ฆ	ฃ	ฅ		
2	C	CH	CH1	CH2				
	จ	ช	ฌ	ฉ				
3	D	D1						
	ด	ฎ						
4	T	TH	TH1	TH2	TH3	TH4	TH5	T1
	ต	ท	ถ	ธ	ฐ	ฑ	ฒ	ฏ
5	N	N1	NG					
	น	ณ	ง					
6	P	PH	PH1	PH2				
	ป	พ	ผ	ภ				
7	F	F1						
	ฝ	ฟ						
8	L	L1	L2	LEU				
	ล	ฬ	ฦ *	ฦๅ*				

GN : Group Number
LCN : Local Classification Number

GN	LCN 1	2	3	4
9	R	R1	REU	
	ร	ฤ *	ฤๅ*	
10	Y	Y1		
	ย	ญ		
11	S	S1	S2	S3
	ส	ษ	ศ	ซ
12	H	H1		
	ห	ฮ		
13	B			
	บ			
14	M			
	ม			
15	W			
	ว			
16	?			
	อ			

*: Vowel

(b) Vowels

GN	LCN 1	2	3	4	5	6	7	8	9	10	11
17	A	A-	A:	AI	AI1	AE	AE-	AE:	AM	AW	A.
	-ั-	-ะ	-า	ไ-	ใ-	แ-็	แ-ะ	แ-	-ำ	เ-า	-
18	I	I:	IA	IA-							
	-ิ	-ี	เ-ีย	เ-ียะ							
19	U	U:	UA	UA-	UA1						
	-ุ	-ู	-ัว	-ัวะ	-ว-						
20	E	E-	E:	EU	EU:	EU1:	EUA	EUA-			
	เ-็	เ-ะ	เ-	-ึ	-ื	-ือ	เ-ือ	เ-ือะ			
21	O	O:	OU	OU-	OU:	OE	OE:	OE-	O.		
	โ-ะ	โ-	-็อ	เ-าะ	-อ	เ-ิ	เ-อ	เ-อะ	-		

(c) Special Symbols and Control Codes

LCN \ GN	22	23	24	25	26	27	28	29	30	31
1	0	1	2	3	4	5	6	7	8	9
	๐	๑	๒	๓	๔	๕	๖	๗	๘	๙
LCN \ GN	32	33	34	35	36	37	38			
1	'	<	>	+	Q	Z	V			
	-่	-้	-๊	-๋	ฯ	ๆ	-์			
LCN \ GN	39	40	41	42	43	44	45	46	47	48
1	/	*	(	)	,	.	–	:	sp	@
	/	*	(	)	,	.	-	:	sp	@

GN:49, LCN:1 Carriage Return

3. *Comparison of DMM, TM, and HTM*

While the TM requires at most 49 keys to be used on the keyboard, the DMM, in which each Thai letter corresponds uniquely to one key, requires the use of 92 keys. Thus the TM allows the number of keys to be reduced by 46.7%. However, the number of key strokes required to input all Thai letters by the TM is 176, which is 91.3% higher than the number required by the DMM.

Compared with the DMM, the number of key strokes required to input a text with the same frequency of occurrence of Thai letters as the text of the Three Seals Law, it was estimated that the HTM, which is the transliteration method proposed by J. F. Hartmann and G. M. Henry, requires 32.0% more strokes and the TM 21.9% more strokes [Shibayama and Hoshino 1986a].

4. *Measurement and Evaluation of Learning Effect*

Input of the main entries of the Thai-Thai dictionary published by the Thai Royal Institute [Photchana nukrom Thai 1982], a total of 31,202 words, was completed in about 6 weeks by 3 persons (about 9 man-weeks). The frequency of occurrence of each Thai letter in the dictionary, the learning effect measured for the elapsed time of the input work, and its evaluation are as follows.

4.1 *Frequency of Occurrence of Thai Letters*

The main entries in the dictionary contained a total of 217,926 letters, which included all 72 character patterns. The percentages of consonants, vowels, tones, and others were 63.5%, 29.9%, 5.4%, and 1.2% respectively.

Table 3 shows the frequency of occurrence of Thai letters in the main entries of the dictionary. For inputting this text, the ratio of the number of key strokes, *T.D.*, required by the TM and DMM can be represented as follows:

$$T.D.=\frac{\sum f_i r_i}{\sum f_i}$$

where f_i is the frequency of occurrence of the i-th Thai letter indicated in the NO. column in Table 3, r_i is the number of characters in its Roman spelling, and the suffix i ranges from 1 to 70. The $\sum f_i r_i$ represents the total number of key strokes for the text. It was found that the number of key strokes required by the TM was 42.9% higher than by the DMM.

4.2 *Environment of Measurement*

The model of behavior in the input work by the typist in the making of the database for the Thai dictionary is shown in Fig. 5. We have measured the learning effect for two persons in the actual input work by the DMM and TM in conjunction with the text editor implemented by both methods. On the editing screen for this input work, of which an example is shown in Fig. 2(b), the slash (/) indicates the division between the words, and the hyphen (-) means that the previous character string with no hyphen is duplicated in this position.

The Thai character string in the second row from the bottom in Fig. 2(b) is a prompt for the next input in Roman spelling,

Table 3 Frequency of Occurrence of Thai Letters in the Thai Dictionary

(a) Frequency of Occurrence of Consonants

NO.	Letter	Freq.	NO.	Letter	Freq.	NO.	Letter	Freq.	NO.	Letter	Freq.
1	ก	11,015	12	ฌ	11	23	ท	3,956	34	ย	7,024
2	ข	2,477	13	ญ	821	24	ธ	1,263	35	ร	13,058
3	ฃ	1	14	ฎ	125	25	น	11,350	36	ล	6,389
4	ค	3,437	15	ฏ	232	26	บ	4,164	37	ว	6,158
5	ฅ	2	16	ฐ	311	27	ป	3,898	38	ศ	1,358
6	ฆ	190	17	ฑ	243	28	ผ	900	39	ษ	918
7	ง	7,487	18	ฒ	63	29	ฝ	278	40	ส	5,427
8	จ	2,841	19	ณ	1,295	30	พ	3,455	41	ห	4,748
9	ฉ	544	20	ด	4,841	31	ฟ	463	42	ฬ	99
10	ช	2,519	21	ต	5,640	32	ภ	1,168	43	อ	8,599
11	ซ	627	22	ถ	993	33	ม	7,946	44	ฮ	157

(b) Frequency of Occurrence of Vowels and Tones

NO.	Letter	Freq.	NO.	Letter	Freq.	NO.	Letter	Freq.
45	ะ	5,315	54	เ	7,854	63	-้	6,211
46	-ั	7,893	55	แ	2,415	64	-่	5,308
47	า	14,874	56	โ	2,118	65	-๋	205
48	-ิ	6,596	57	ฤ	301	66	-๊	221
49	-ี	4,421	58	ฦ	4	67	-์	2,302
50	-ึ	627	59	ำ	1,625	68	ฯ	292
51	-ื	1,798	60	ใ	652	69	ๆ	2
52	-ุ	4,092	61	ไ	1,495	70	ๅ	20
53	-ู	1,988	62	-็	806			

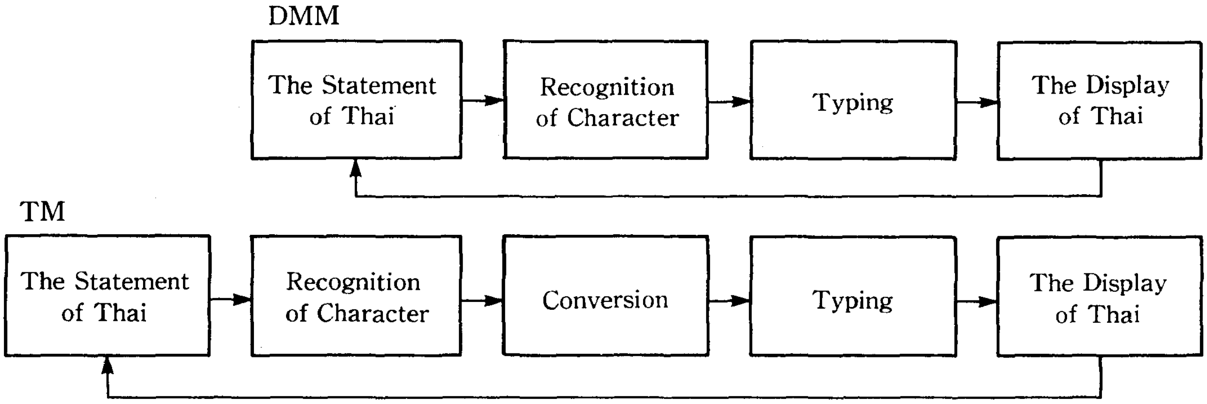


Fig. 5 Model of Behavior for Typing of Thai

Table 4 Enumeration of Main Entries in the Dictionary

Input Method	Operator (A)		Operator (B)	
	Number of Words	Number of Char.	Number of Words	Number of Char.
DMM	12,455	78,340	4,478	29,181
TM	8,490	54,527	4,661	29,200
Total	20,945	132,867	9,139	58,381

displaying the group of Thai letters above and their corresponding Roman spellings below. Fig. 2(b) shows the prompt when "U" was typed as the next input. The Thai character string in the center of the third row from the bottom represents the result of transliteration of the input of the Roman spelling inside the box in the second row from the bottom.

Table 4 shows the amount of text input by two operators, the total number of words and characters input being 30,084, and 191,248 respectively. These two operators had no prior knowledge of Thai letters or Thai language, but were able to input about 200 letters per minute of Roman script.

4.3 Measurement and Its Evaluation

It was assumed that the operators used their fingers and hands in accordance with the assignment shown in Fig. 4. The frequency of the use of fingers, hands, and each row of the keyboard by the typist is illustrated in Fig. 6. It is noticeable that the little fingers, which are considered least effective, are used frequently in both methods, and that the right hand works more than left hand by 24.8% and 17.0% respectively in the DMM and the TM. Of the rows of the keyboard, the home row *C* is used most frequently, which is considered to be effective, and the average

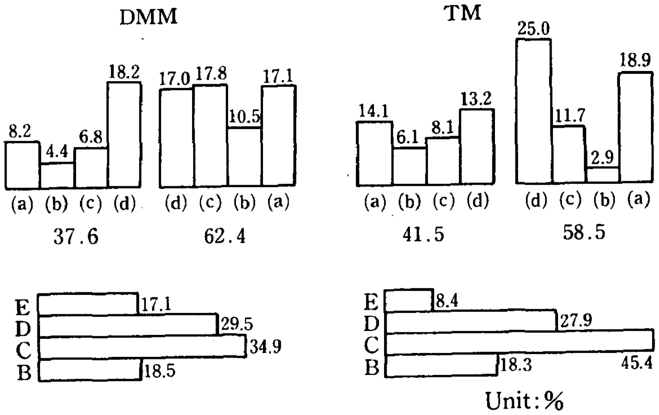


Fig. 6 Utilization of the Fingers, Hands, and Rows of Keyboard

distance of movement of fingers is follows:

$$d_{DMM}=0.185*1+0.349*0+0.295*1+0.171*2$$
$$=0.882$$
$$d_{TM}\ =0.63$$

where d_{DMM} and d_{TM} are the average distance of movement in the DMM and the TM estimated from the utilization in Fig. 6. In comparison, the figures for the standard English keyboard d_E for Roman script input, RICOH d_2 (two-strokes method) for Japanese input, and JIS d_j (JIS keyboard) for Japanese input are 0.66, 0.60, and 0.91 respectively.

Fig. 7 shows the result of the measurement of typing speed by both methods and the learning curves for operator (A). The *X* axis in Fig. 7 indicates the elapsed time in the actual input work, and the *Y* axis indicates the number of characters input per minute. To represent the learning curves, the speed of typing *S(t)* is fitted to a function of the elapsed time t as follows:

$$S(t)=M(1-e^{-Gt})$$

where *M* is the superior limit of the typing speed, and *G* is the coefficient of training efficiency. Fitting of the learning curves to the measured values by use of this relation gives $M=37.25$, $G=0.0527$ for the DMM, and $M=40.9$, $G=0.0797$ for the TM. Despite the time required to consult the transliteration table in the TM, and the 42.9% greater number of key strokes than the DMM, the typing speed is 9.8% higher by the TM than the DMM. Consequently, we found that the TM is more readily operable by non-native speakers of Thai accustomed to inputting Roman script. It is also expected that the typing speed would increase if the elapsed time could be extended.

5. *Simplified Transliteration Method (STM)*

As shown in section IV.*2* and from the experiment just described, to input any Thai letter by the TM, the typist has to memorize or consult the transliteration table to identify the Roman spelling, namely, the *LCN* in Table 2. It is difficult, however, to memorize the transliteration table in a short time, especially for non-native speakers of Thai, and the need to consult it reduces the speed of typing.

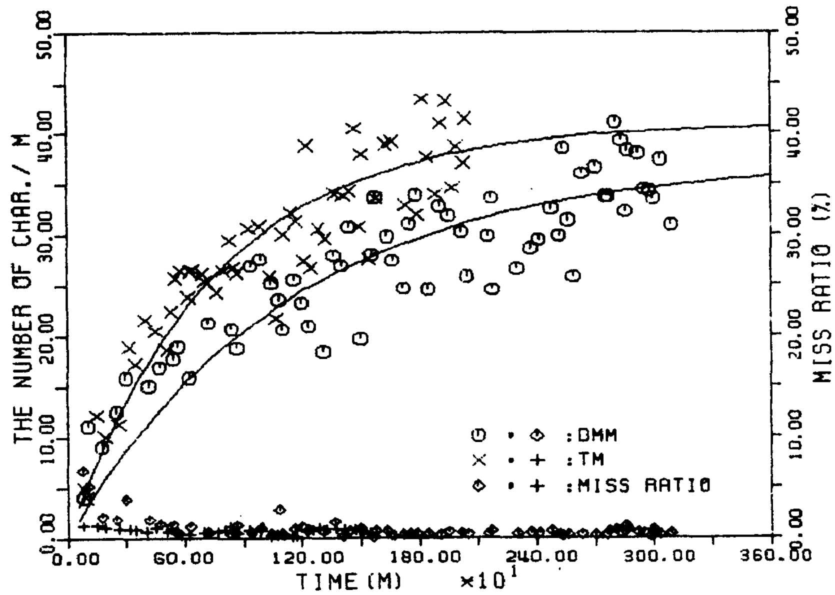


Fig. 7 Typing Speed and the Learning Curve for Operator (A)

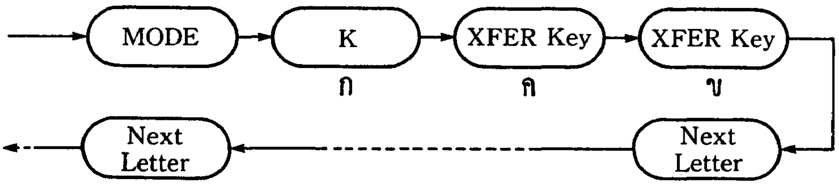


Fig. 8 Simplified Transliteration Method

To eliminate the overhead time for memorizing and consulting the transliteration table in the TM, we have devised a simplified transliteration table composed of only *GN*'s groups, without the distinction of *LCN*, namely, the Simplified Transliteration Method (STM, See Fig. 8). For example, the Roman spelling "K" corresponds to "ก", "ค", "ข", "ฆ", "ฃ", and "ฅ". After pressing "K", the typist then selects the appropriate Thai letter from the group by pressing the "XFER" key (See Fig. 4) on the keyboard, which causes the Thai letters to appear one by one cyclically, and by pressing any key except the "XFER" key for the next input when the appropriate Thai letter appears.

To estimate the number of key strokes of the STM using the frequency of occurrence of Thai letters as shown in Table 3, the ratio of the number of key strokes, *S.D.*, required by the STM and the DMM can be represented as follows:

$$S.D. = \frac{\sum f_i n_j}{\sum f_i}$$

where f_i is the frequency of occurrence of the i-th Thai letter indicated in the NO. column in Table 3 and the suffix i ranges from 1 to 70. The n_j is the number of key strokes for extracting the i-th Thai letter indicated in Table 3, namely, the value of *LCN*, in the j-th group belonging its i-th Thai letter and the suffix j corresponds to the value of *GN* shown in Table 2. For example, the number of key strokes required for "ค" is 2 ("K" and "XFER" keys) which corresponds the value of *LCN* in $GN=1$. And $\sum f_i n_j$ represents the total number of key strokes for the text.

It is estimated that the number of key strokes required by the STM for inputting a text with same frequency of occurrence of Thai letters as the main entries of the dictionary is 61.6% (43.0% for the consonants and 93.9% for the vowels and tonal marks) higher than by the DMM. Compared with the TM, the STM requires 18.7% more strokes. However, the STM is more readily operable by non-native speakers of Thai than the TM, and the number of key strokes can be reduced if the sequence of appearance of Thai letters, especially the vowels, in a group is changed according to the text, like the learning function for the Roman-Kanji conversion in Japanese. This scheme can also be implemented on an intelligent terminal capable of displaying Thai letters, such as a micro computer.

V Thai Printing

1. *Characteristics of Thai Letters*

Thai letters have the following characteristics:

(a) They differ from each other in size as well as shape, for example, า, ญ, ำ, ฤ, and โ.

(b) Several letters are overlapped in printing, for example, ป้ is composed of ป and ้.

(c) Several letters are located above and between adjacent letters, like อัก.

(d) The printing positiors of the same letter are sometimes different, like ขั้ and ข้.

Sakamoto has proposed that such problems can be solved for the printing of almost all Asian and African letters by dividing letter patterns into sub patterns, comprising the *basic character* with special information on the *basic line* and the *added character* with information on its *basic point* [Sakamoto 1979]. We have adopted this idea in the design of a Thai printing system and developed more controllable programs that allow any line spacing with an integral number of dots by adding functions for overlap control of the character patterns, line position control for each line, and vertical position control of a character depending on the position of the previous character. The schemes of these controls are as follows:

(1) Discrimination of *Basic Character* and *Added Character*

Thai letters in combinations of *consonant* +*vowel* or *consonant*+*vowel*+*consonant* are composed of 6 regions centered on the first consonant of a word, as shown in Fig. 9. The characters located at (1), (2), and (3) in Fig. 9 are categorized as *basic characters*, and those at (4), (5), and (6) as *added characters*. It is assumed that the sequence of appearance of characters must be *basic character* before *added character*.

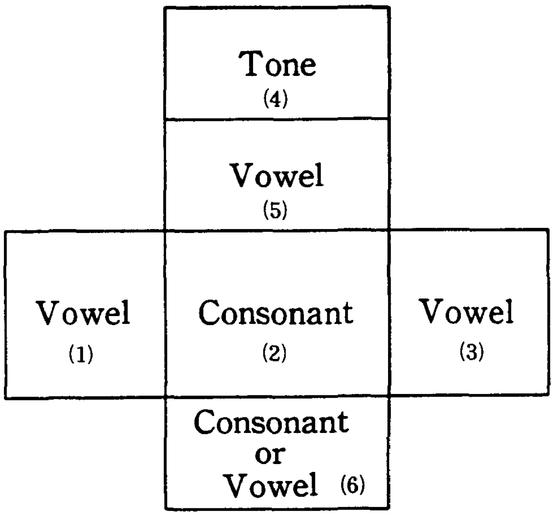


Fig. 9 Positions of Consonants, Vowels, and Tones

(2) Control of *Basic Character*

Basic characters have the attribute *basic line*, which shows the width of the character, and which determines the horizontal printing position of the letter relative to the preceding letter. By employing this scheme, the width of letters can be controlled closely, and printing with proportional spacing is possible.

/ข้างตะเภา/

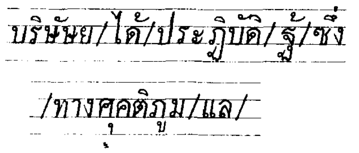
(3) Control of *Added Character*

An *added character* has a *basic point* rather than a *basic line*, which together with information on its *setting position* serves to locate the character relative to the *basic line* of the preceding *basic character*. The printing position for a Thai letter with an *added character* is thus determined by the attribute *setting position*. This control is the same as the dead key control of a typewriter.

/ขาพับ/　พั

(4) *Parental* and *Child Patterns* and the Line Position Control

Satisfactory printing quality of consonants can generally be achieved if a character is represented by 40*40 dots. However, consonants like ฎ and ป require 80*40 dots. We have therefore split the string of Thai letters into three levels, and divided the consonants and vowels represented by 80*40 dots into two patterns, here called *parental* and *child patterns*. The printing position for each pattern in a Thai letter is decided by the attribute *line position*, which shows the level of the *parental* and *child patterns*.



(5) Overlap Control

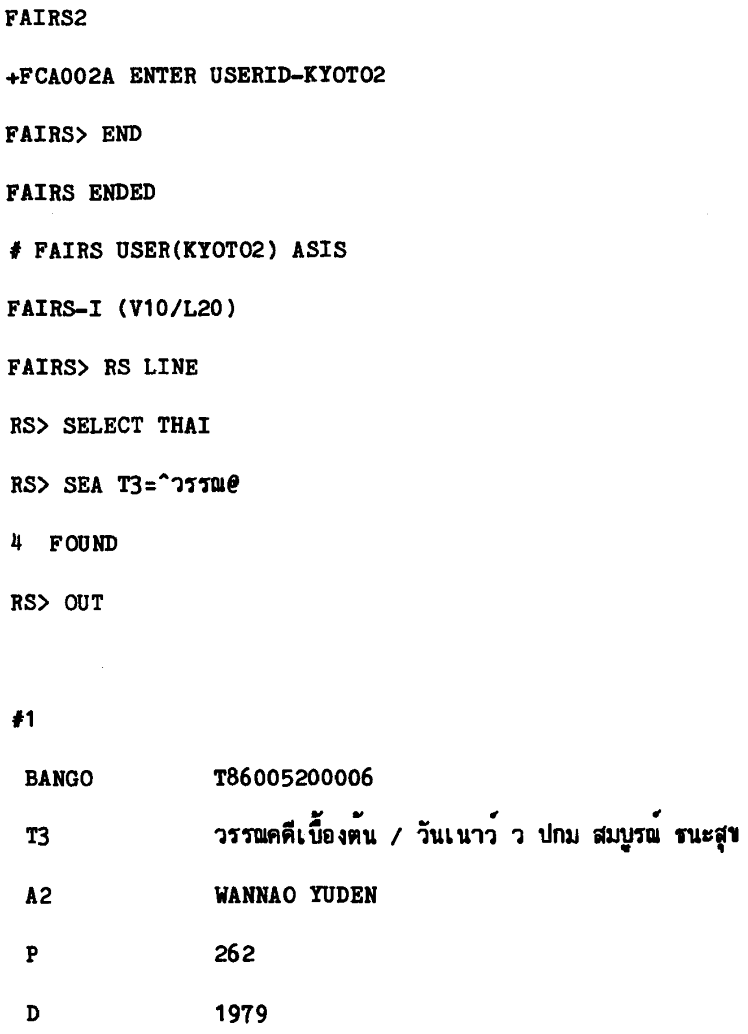
Every character pattern is synthesized by an *OR* operation for all dots. This scheme is necessary for such characters as ปี, ฝี, and ขี้.

(6) Vertical Control of *Added Characters*

To improve printing quality, four tonal marks in the vertical position need to be repositioned if the previous letter has a vowel like ิ, ี, or ื. In this case, the *setting position* of tonal mark is shifted vertically upward by an appropriate number of dots, and the tonal mark is printed overlapping the preceding vowel.

2. *Example*

Fig. 10 shows an example of the output for retrieving a Thai bibliography on the intelligent terminal connected to the host computer. *FAIRS 2* and *FAIRS* in Figure are commands for invoking the information retrieval system on the host computer. We have developed a program such that the printing program runs mainly by operating the dead key and vertical position controls for each character, as



```
FAIRS2

+FCA002A ENTER USERID-KYOTO2

FAIRS> END

FAIRS ENDED

# FAIRS USER(KYOTO2) ASIS

FAIRS-I (V10/L20)

FAIRS> RS LINE

RS> SELECT THAI

RS> SEA T3=^วรรณ@

4  FOUND

RS> OUT


#1

 BANGO          T86005200006

 T3             วรรณคดีเบื้องต้น / วันเนาว์ ว ปกม สมบูรณ์ ธนะสุข

 A2             WANNAO YUDEN

 P              262

 D              1979
```

Fig. 10 Example of Retrieval of Thai Bibliography Using the Intelligent Terminal

described previously.

VI Conclusion

The Thai text editor and the intelligent Thai terminal designed have the functions for inputting the Thai text by the Direct Mapping, Transliteration, and Simplified Transliteration Methods. Using these, we are now developing a database and a computer concordance of the Three Seals Law.

The structure and characteristics of the input methods have been compared by measuring the speed of typing and learning curves in the actual input work for making a machine-readable Thai dictionary. The Transliteration Method has the advantage of requiring fewer keys on the keyboard than the Direct Mapping Method, and if transliteration from Roman spelling to Thai letters is implemented on a host computer connected to the terminal, an ordinary teletype terminal can be used to input Thai letters. Although the number of key strokes required for text input will normally be higher by the Transliteration Method, this method was found to be more readily operable by those accustomed to inputting Roman spelling.

This scheme is applicable to design of terminals and editors for other Southeast Asian languages, like Laotian and Burmese.

We have also developed output methods for Thai letters of high quality, and have described the structure and characteristics of methods for controlling the printing and display positions of Thai characters on a laser beam printer and a micro computer.

We are now working to develop a database and a computer concordance of the Three Seals Law at the Center for Southeast Asian Studies and the Data Processing Center, Kyoto University using this editor and terminal.

Acknowledgment

Thanks are due to Prof. Yoneo Ishii of the Center for Southeast Asian Studies and Prof. Satoshi Hoshino of the Data Processing Center, Kyoto University who led us to this study, and to Dr. Shigeharu Sugita of the National Museum of Ethnology, who made many valuable comments and provided the original text of the Three Seals Law in machine readable form. Thanks are also due to Prof. Yasuyuki Sakamoto of Tokyo Foreign Languages University, Prof. Tetsuya Katsumura of the Research Institute for Humanistic Studies, Kyoto University, Mr. Yukio Hayashi of Ryukoku University, Mrs. Aroonrut Wichienkeeo of Chiang Mai Educational University, Dr. Supamard Panichsakpatana of Kasetsart University, Dr. Jitti Pinthong of Chiang Mai University, and Dr. Sukanya Nitungkorn of Thammasat University, who made many valuable comments on the transliteration table in the editor and terminal design.

This study has carried out as part of a project to develop database for the Three Seals Law of Thailand in the Data Processing Center, Kyoto University, and has been supported by grants for scientific research from the Japanese Ministry of Education.

Appendix

——*Outline of Development of a Database and a Computer Concordance for the Three Seals Law of Thailand*——

The process for implementing on-line information retrieval and a computer concordance of the Three Seals Law is shown in Fig. A–1. The process in Fig. A–1 advances into two flows: on the left, a database of Thai dictionary on the computer has been made in order to verify all the words in the text of the Three Seals Law. This can be used for the studies of natural language processing, in other words, morpheme analysis, syntactic analysis, and semantic analysis as basic research into natural

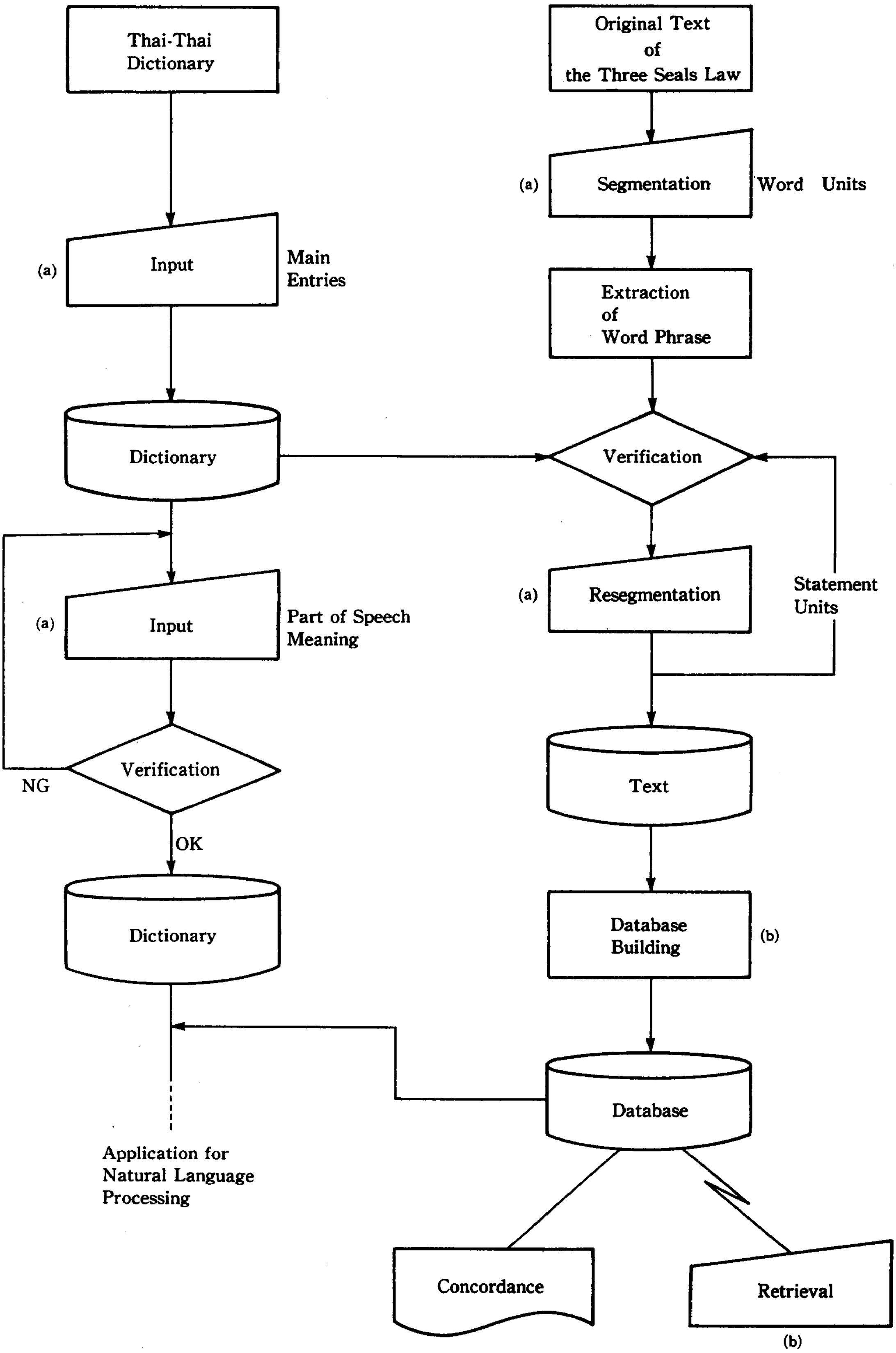


Fig. A-1 Outline for Developing a Database and a Computer Concordance of the Three Seals Law language understanding, for a question-answering system, and for machine translation involving Thai.

The right-hand flow shows the process of building a database from the original text of the Three Seals Law, which had made such segmentation that a statement is divided into the words, provided by the National Museum of Ethnology into a database on the host computer capable of retrieving it through the on-line terminal, and it also shows that a computer concordance is constructed simultaneously.

In making the database of the Thai dictionary, the main entries (about 32,000 words) of the dictionary published by the Thai Royal Institute [Photchana nukrom Thai 1982] were input using the Thai text editor. To complete the machine-readable dictionary, the main entries need to be supplemented with details of part of speech, meanings, and other rules necessary for the machine processing of Thai. Finally, when the database of the Thai dictionary is complete, it can be used for syntactic analysis of Thai statements.

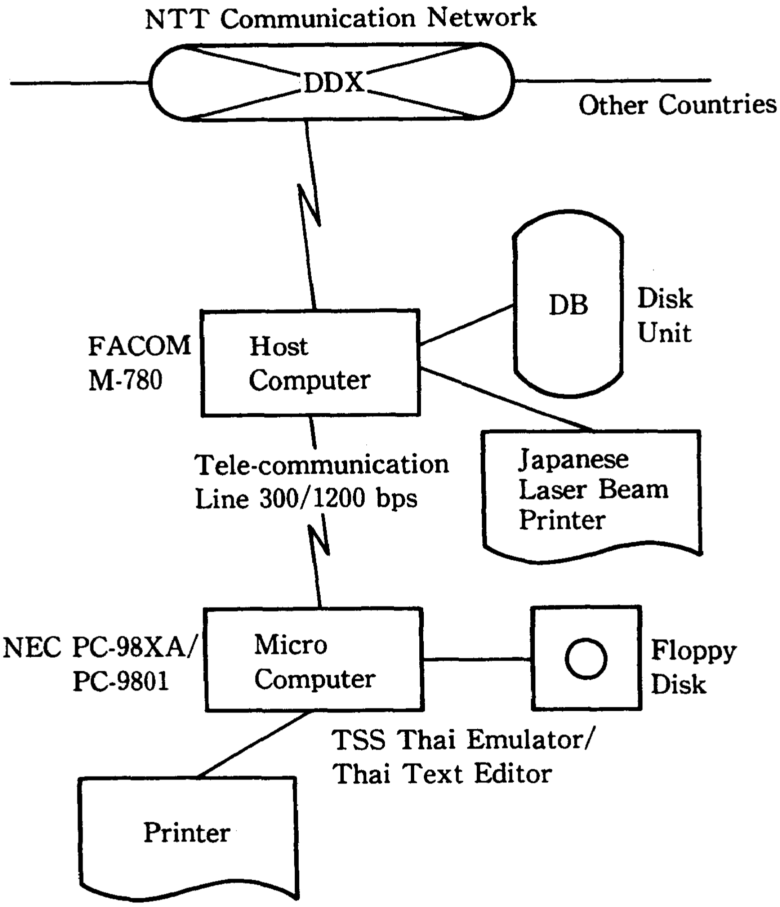

NTT Communication Network
DDX
Other Countries
DB
Disk Unit
FACOM M-780
Host Computer
Tele-communication Line 300/1200 bps
Japanese Laser Beam Printer
NEC PC-98XA/ PC-9801
Micro Computer
Floppy Disk
TSS Thai Emulator/ Thai Text Editor
Printer

Fig. A-2 System Configuration for Retrieving the Database of the Three Seals Law

In the building of the database and the computer concordance of the Three Seals Law, resegmentation has first to be accomplished, which includes the reading of text to confirm the existing segmentation and the input work using the Thai text editor. Then each different word is extracted from the text file and the verified against the dictionary on the host computer. If any mistakes are found in either the dictionary or the text, feedback into the appropriate positions must be attempted. After the corrective work, the text is segmented into the sentences to make the database retrieval more efficient.

By building the database using the information retrieval system, *IRS*, as an application software on the computer, on-line information retrieval of the Three Seals Law is possible, and computer concordance also can be accomplished by using a function in the *IRS*.

As shown in Fig. A-1, (a) shows that the Thai text editor is used for editing the text of the Three Seals Law and the dictionary in the off-line status. The edited files are transferred into the host computer using the file transfer function on the micro computer. Also, (b) shows that the ordinary terminal and the intelligent terminal connected to the host computer are used. The overall configuration of the system for building and retrieving the database is shown in Fig. A-2. The database of the Three Seals Law and the dictionary of Thai are stored on the disk unit and managed by the *IRS* in the host computer.

For building or retrieval, the database is accessed by invoking the *IRS* from the terminal, and the results can be output to a Japanese laser beam printer or to the terminal.

The terminal shown in Fig. A-2 is made up 2 components: a micro computer and a TSS[1]

[1] TSS: Time-sharing system, namely, the mode of communication with the host computer on an interactive basis.

terminal emulator, which is operated on the micro computer, and is connected with the host computer through a telecommunication line with the speed of 300/1200 bps.[2] The micro computer can, of course, be used for editing the Thai text in the off-line status.

References

Hartmann, J. F.; and Henry, G. M. 1983a. Thai Script Computer-converted from a Precise, Pronounceable Transliteration for Bibliographic Management. *Bulletin of Committee on Research Materials on Southeast Asia (CORMOSEA)* 11(2).

————. 1983b. The Processing of Thai Language Text Using a Personal Computer. *Bulletin of Committee on Research Materials on Southeast Asia (CORMOSEA)* 11(2).

Ishii, Y. 1969. Sanin Hoten ni Tsuite [Introductory Remarks on the Law of Three Seals]. *Tonan Ajia Kenkyu* [Southeast Asian Studies] 6(4): 155–178.

Murayama, N. 1982. 2 Sutorooku-ho [2 Strokes Method]. *Jyoho Syori* [Information Processing] 23(6).

Photchana nukrom Thai. 1982 (Thai 2525). *Chabap Ratchabandit-sathan.* Krungthep: Samnakphim Aksonchaoenthat.

Sakamoto, Y. 1979. Ajia Afurica Gengo no Konpyuuta Syori [Computer Processing for Asian and African Languages]. *Jyoho Kanri* [Information Management] 22(7).

Shibayama, M.; Sugita, S.; and Ishii, Y. 1984. Pasokon ni yoru Taigo Tekisuto no Syori [Processing of Thai Text Using a Personal Computer]. *Dai 28 Kai Jyoho Syori Gakkai Zenkoku Taikai Ronbun-syuu* [Proceedings on Japan Information Processing Society 28th National Conference].

————. 1985. Romaji Hyoki ni yoru Taimoji no Nyuryoku Hoshiki [Input Methods for Thai Using Roman Spelling]. *Dai 30 kai Jyoho Syori Gakkai Zenkoku Taikai Ronbun-syuu* [Proceedings of JIPS30].

Shibayama, M.; and Hoshino, S. 1986a. Implementation of an Intelligent Thai Computer Terminal. *Journal of Information Processing* 8(4): 300–306.

————. 1986b. Taiji Jisyo no Nyuryoku-ho to Nyuryoku-tokusei [Methods and Characteristics of Thai Dictionary Inputs]. *Dai 33 kai Jyoho Syori Gakkai Zenkoku Taikai Ronbun-syuu* [Proceedings of JIPS33].

————. 1987. A comparative Study of the Characteristics of Input Methods for Thai. In *Proceedings of the Regional Symposium on Computer Science and Its Application.* Thailand: NRCT. pp. 19·1–19·18.

Sugita, S. 1980. Text processing of Thai language: The Three Seals Law. In *Kagaku Kenkyuu-hi Shiken Kenkyuu Seika Hokoku-syo: Jinbun Kagaku Kenkyuu Shien no tameno Konpyuuta Apurikeesyon no Kaihatsu* [Development of Computer Application for Assisting the Studies of Cultural Science], edited by Tadao Umesao, pp. 122–129. National Museum of Ethnology.

[2] bps: Bits per second, namely, the unit of communication rate.