Regular Paper

Thai Morphological Analyses Based on the Syllable Formation Rules

MAMORU SHIBAYAMA* and SATOSHI HOSHINO**

The Thai syllable formation rules were deduced from an orthographical analysis of Thai. At the morphological level, segmentation was analyzed by the ordinary longest-match method for the input of Thai text (the Law of Three Seals: 20,631 sentences), a revised method of segmentation, called the Syllable Longest-Match method (SLM), which incorporated a mechanism of back-tracking for each phoneme based on the syllable formation rules when the segmentation failed, was then devised to reduce the number of unsuccessful cases. This method indicated that the ratio of segmentation is a 98.0%, which is 2.8% greater than the ordinary method in terms of sentences.

A finite automaton model which employs the automatic segmentation from a sentence into monosyllables without reference to a dictionary, called Thai syllable recognizer, was also proposed. A revised Thai syllable recognizer was also devised, in which knowledge rules based on the heuristics derived from the analysis of unsuccessful cases were adapted the existing syllable formation rules. This gave a ratio of segmentation is 93.9% in terms of sentences for the input of same text.

1. Introduction

Written Thai differs from western languages in several points; (a) Thai letters are phonetic; they consist of 44 consonants and 32 vowels as shown in Fig. 1. (b) Least unit of Thai word is monosyllable accompanying one of 5 tones. (c) The words in a sentence are not separated from each other, here called being unsegmental. (d) Some vowels are placed before the consonant, and the tones must be placed in appropriate positions in the writing system. (e) Punctuation is scarcely used [1, 2].

Mechanical processing of Thai with emphasis on linguistic approaches to natural language processing therefore presents problems relating to successive levels of language processing such as input/output, morphological, lexical, syntactic, and semantic levels [3–7]. At the morphological level, the segmentation, especially, has a very similar property to that of Japanese and Chinese characteristics which is unsegmental [8]. In that case, for making a concordance or KWIC index [9], or building a database [10], the segmentation into the appropriate units as a word, a sentence, or a syllable for the original text may be required. The practical work of segmentation is too time-consuming and excessive, and reduces the accuracy and consistency of the result if the work depends on the effort and handiwork of an expert without the assistance of a computer or if the automatic segmentation using the computer is impossible [11, 12].

Segmentation in the case of an unsegmental language generally uses the longest-match method on the basis of the fact that the longest-match is most simple and popular method [13]. However, few studies on natural language processing of Thai have been carried out.

Thai morphological analyses described in this paper are based on several phonetic rules and definitions derived from linguistic analysis of the phonemes embedded in the syllabic structure and the orthography of Thai, which are here called the syllable formation rules. After introducing those rules first, the ordinary longest-match method is adopted, and this presents the outcome of how well it is segmented for Thai text of the Law of Three Seals (KTSD: Kotmai Tra Sam duang, 20,631 sentences, compiled in 1805) [14, 15]. A revised longest-match method, called the Syllable Longest-Match method (SLM), which incorporated a mechanism of back-tracking for each phoneme based on the syllable formation rules when the segmentation failed, is then revised to reduce the number of unsuccessful cases.

Furthermore, a nondeterministic finite automaton model for recognizing Thai syllables is proposed, using the morphological knowledge which is based on the syllable formation rules of symbols corresponding to the phonemes embedded in a syllable. According to the model, a Thai syllable recognizer also is implemented,

*Faculty of Management of Information Science, Osaka International University.
**Data Processing Center, Kyoto University.

NO.	Letter	Pronun-ciation	NO.	Letter	Pronun-ciation	NO.	Letter	Pronun-ciation	NO.	Letter	Pronun-ciation
1	ก	kɔ:	12	ฉ	chɔ:	23	ท	thɔ:	34	ย	jɔ:
2	ข	khɔ̌:	13	ช	jɔ:	24	ธ	ṭhɔ:	35	ร	rɔ:
3	ฃ	khɔ̌:	14	ฌ	dɔ:	25	น	nɔ:	36	ล	lɔ:
4	ค	khɔ:	15	ฎ	tɔ:	26	บ	bɔ:	37	ว	wɔ:
5	ฅ	khɔ:	16	ฏ	thɔ̌:	27	ป	pɔ:	38	ศ	sɔ̌:
6	ฆ	khɔ:	17	ฐ	thɔ:	28	ผ	phɔ̌:	39	ษ	sɔ̌:
7	ง	ŋɔ:	18	ฑ	thɔ:	29	ฝ	fɔ̌:	40	ส	sɔ̌:
8	จ	cɔ:	19	ฒ	nɔ:	30	พ	phɔ:	41	ห	hɔ̌:
9	ฉ	chɔ̌:	20	ณ	dɔ:	31	ฟ	fɔ:	42	ฬ	lɔ:
10	ช	chɔ:	21	ด	tɔ:	32	ภ	phɔ:	43	อ	ʔɔ:
11	ซ	sɔ:	22	ต	thɔ̌:	33	ม	mɔ:	44	ฮ	hɔ:

(a) Consonant letters

NO.	Letter	Pronunciation	NO.	Letter	Pronunciation
1	-ะ, -ั-	(a?), (a)	17	เ-	(aj)
2	ิ	(i?), (i)	18	เ-า	(aw)
3	ึ	(ɯ?) (ɯ)	19	-า	(a:)
4	ุ	(u?), (u)	20	ี	(i:)
5	เ-ะ, เ็	(e?), (e)	21	ี-, ี-ย	(ɯ:)
6	แ-ะ, แ็-	(ɛ?), (ɛ)	22	ู	(u:)
7	โ-ะ, - -	(o?), (o)	23	เ-	(e:)
8	เ-าะ, ็อ	(ɔ?), (ɔ)	24	แ-	(ɛ:)
9	เ-อะ	(ə?)	25	โ-	(o:)
10	เ็ียะ	(ia?)	26	-อ	(ɔ:)
11	เ็อะ	(ɯa?)	27	เ-อ, เ็	(ə:)
12	็วะ	(ua?)	28	เ็ย	(i:a)
13	ฤ	(rɯ), (ri)	29	เ็อ	(ɯa)
14	ฦ	(lɯ)	30	็ว, -ว-	(ua)
15	็ำ	(am)	31	ฤๅ	(rɯ:)
16	ไ-	(aj)	32	ฦๅ	(lɯ:)

(b) Vowel letters

Fig. 1 Thai letters.

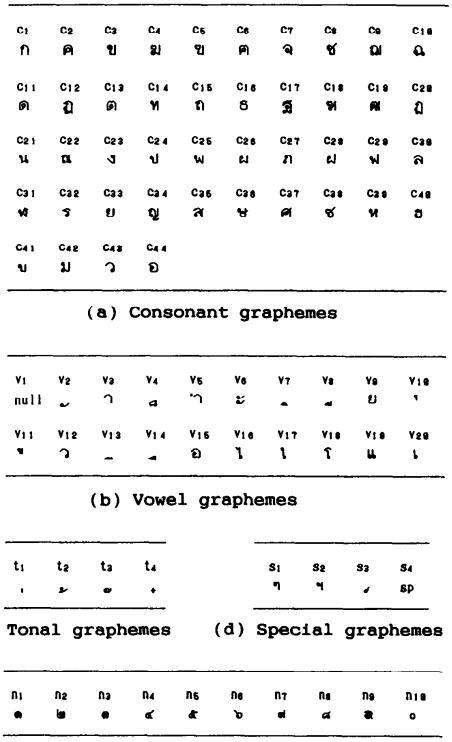

Fig. 2 Thai graphemes.

and can be run by which no dictionary is used for segmenting the sentences. And a result of automatic segmentation obtained from the experiment by the recognizer are presented.

Finally, the consideration as to what is attempted to reduce the error ratio by incorporating the heuristic knowledge into the syllable formation rules is presented, then a revised recognizer is adopted. The detailed analysis and evaluation based on the outcome derived from the behavior of improperly segmented cases by the revised recognizer are described.

2. Syllable Formation Rules

Characteristics of formalizing the syllable, which is defined as the word because Thai language is the phonogram, are listed as follows: (a) Thai words are phonetic; they are basically monosyllables with tones and are composed of "Consonant+Vowel" or "Consonant+Vowel+Consonant". Then monosyllable is defined as the syllable in this paper. (b) Thai word also is formed by the combination of several words, called compound word. (c) Thai letters consist of 44 consonants and 32 vowels. However some vowels are represented with a combination of several symbols based on Fig. 1(b) [16].

The smallest recognizable unit on the Thai syllable recognizer or the least unit of character in machine readable form is defined as a symbol, and also Thai alphabet expression corresponding to a symbol is defined as a grapheme. Therefore, each grapheme derived from the Thai letters in Fig. 1 can be expressed as shown in Fig. 2. Some vowel letters in Fig. 1(b) are composed of several graphemes; for example, No. 9 in Fig. 1(b) is represented with a combination of 3 vowel graphemes, and a consonant grapheme is placed at the "—" position for formalizing the syllable.

From the characteristics of above mentioned (a), we denote the syllable as follows:

$$[C+V] \tag{1}$$

$$[C+V+C] \tag{2}$$

where C and V correspond to a consonant and vowel respectively, "[ ]" means a words unit, and "+" means the concatenating operator among the symbols.

Hence, it is found that the sentence formation, namely, the individual words which are embedded in a sentence, is specialized from the expression (1) and (2) as the following sentence patterns;

$$[C+V]+[C+V]+ \ldots\ldots \quad (3)$$

$$[C+V]+[C+V+C]+ \ldots\ldots \quad (4)$$

$$[C+V+C]+[C+V]+ \ldots\ldots \quad (5)$$

$$[C+V+C]+[C+V+C]+ \ldots\ldots \quad (6)$$

Next, the Thai symbols based on Thai letters are defined.

Thai symbols are classified as follows:

(1) Subset CS: Consonant symbols

$$CS=\{c_1, c_2, c_3, \ldots, c_{44}\} \quad (7)$$

Each consonant grapheme in Fig. 2(a) is discriminated by c_i which is each element in set CS, and i varies from 1 to 44.

(2) Subset VS: Vowel symbols

$$VS=\{v_1, v_2, \ldots, v_{20}\} \quad (8)$$

Each symbol corresponds to each vowel grapheme in Fig. 2(b) respectively.

(3) Subset TS: Tonal symbols

$$TS=\{t_1, t_2, t_3, t_4\} \quad (9)$$

Each symbol corresponds to each tonal grapheme in Fig. 2(c) respectively.

(4) Subset NS: Numeral symbols

$$NS=\{n_1, n_2, n_3, \ldots, n_{10}\} \quad (10)$$

(5) Subset SS: Special Symbols

$$SS=\{s_1, s_2, s_3, s_4\} \quad (11)$$

where such vowel symbols are equivalent to the consonant symbol as follows;

(a) $v_9=c_{33}$

(b) $v_{12}=c_{43}$

(c) $v_{15}=c_{44}$

Consequently, the set of symbol which forms Thai language is defined as;

$$T=\{CS, VS, TS, NS, SS\},$$

namely, T implies Thai alphabet, and all the string formed by arbitrary finite sequence of symbols from T including null string is designated by T^*, and T^* also is finite.

By analyzing the order of appearance of grapheme, in other words, symbols according to their phonemic rules for Thai, the consonant symbols at end of a syllable in the form [C+V+C] are 41. And the vowel symbols which coincide with phonetic rule in sequence as shown in Fig. 3(a) introduced from Fig. 1(b) are defined. Contrary to its sequence, the vowel symbols that the sequence of appearance of grapheme embedded in a syllable differs from its phonetic rule also are defined as shown in Fig. 3(b). In both Fig. 3(a) and (b), a consonant grapheme in Fig. 2(a) must be inserted in "−" position.

Case	1st	2nd	3rd	4th
1	-	v_2	-	
2	-	v_2	v_{12}	
3	-	v_2	v_{12}	v_6
4	-	v_3		
5	-	v_4		
6	-	v_4	v_{15}	
7	-	v_5		
8	-	v_6		
9	-	v_7		
10	-	v_8		
11	-	v_{10}		
12	-	v_{11}		
13	-	v_{12}	-	
14	-	v_{13}		
15	-	v_{14}		
16	-	v_{14}	v_{15}	
17	-	v_{15}		

(a) Vowel symbols coinciding with phonemic rule

Case	1st	2nd	3rd	4th	5th
1	v_{16}	-			
2	v_{17}	-			
3	v_{18}	-			
4	v_{18}	-	v_6		
5	v_{19}	-			
6	v_{19}	-	v_4		
7	v_{19}	-	v_6		
8	v_{20}	-			
9	v_{20}	-	v_3		
10	v_{20}	-	v_3	v_6	
11	v_{20}	-	v_4		
12	v_{20}	-	v_6		
13	v_{20}	-	v_7		
14	v_{20}	-	v_8	v_9	
15	v_{20}	-	v_8	v_9	v_6
16	v_{20}	-	v_{14}	v_{15}	
17	v_{20}	-	v_{14}	v_{15}	v_6
18	v_{20}	-	v_{15}		
19	v_{20}	-	v_{15}	v_6	

(b) Vowel symbols uncoinciding with phonemic rule

Fig. 3 Sequence of appearance of graphemes in the vowels.

Consonant and vowel symbols are represented as follows:

(6) Subset CS_e: Consonant symbols at end of the syllable

$$CS_e=CS-\{c_{39}, c_{40}, c_{44}\} \quad (12)$$

where an arbitrary element in set CS_e is c_e $\{c_e: c_e \in CS_e\}$.

(7) Subset VS_n: Vowel symbols coinciding with phonemic rule

$$VS_n=VS-\{v_1, v_9, v_{16}, v_{17}, v_{18}, v_{19}, v_{20}\} \quad (13)$$

where an arbitrary element in set VS_n is v_n $\{v_n: v_n \in VS_n\}$.

The syllable formation rules based on the Thai grammatical rules are introduced as below:

The expressions (3), (4), (5), and (6) indicated previously deduce to one rule.

[Rule 1] No boundary of word unit is at a point immediately before the vowel symbol. In other words, the segmentation should not be made between symbols of C and V in the [C+V] or [C+V+C] sequence.

Look at the table of vowel letters as shown in Fig. 1(b) again. In the Thai orthography, the sequence of appearance of grapheme for the letter No. 5-11, 17-18,

23–25, and 27–29 in Fig. 1(b) is reversed as shown in Fig. 3(b) contrary to the sequence of phonemes. Those graphemes concern with such vowel letters are 5; เ , ไ , ใ , แ , and โ , then they are denoted by v_{16}, v_{17}, v_{18}, v_{19}, and v_{20} respectively.

This result leads to a following rule.

[Rule 2] Symbol v_{16}, v_{17}, v_{18}, v_{19}, and v_{20} are the first symbol embedded in a word, for example, ไป (pai) is denoted by v_{16} and c_{24}.

From the table of vowel letters, the following graphemes consisting vowel letter are classified into 3 groups. The first group is when the consonant grapheme is followed by a vowel grapheme; which consists of the symbols v_3, v_5, v_6, v_9, v_{12}, and v_{15}. The second is when a vowel grapheme is placed on appropriate position above the consonant; which consists of the symbols v_2, v_4, v_7, v_8, v_{13}, and v_{14}. The third is when a vowel grapheme is placed on appropriate position below the consonant; which consists of symbols v_{10} and v_{11} respectively.

Assume that the consonant symbol is followed by a vowel symbol for all of above groups in the sequence of appearance; for example, ก็ is a string c_1v_4 in the machine readable form, where ก is denoted by c_1.

[Rule 3] No boundary of word unit is at a point immediately before all symbols of V belonging above 3 groups; for example, a word จะ (ca) is denoted by c_7 and v_6.

When either v_9, v_{12}, or v_{15} is not a vowel symbol, an attribute of that consonant or vowel corresponds to a symbol is decided by the position where it is embedded in a word based on [Rule 1], for example, a word อออก (oo: k) is a string $c_{44}v_{15}c_1$, where first อ is denoted by c_{44}.

[Rule 4] Discrimination between a consonant and a vowel is determined based on the position of its symbol embedded in a word; for example, first symbol of วว (wa:) must be recognized as a consonant c_{43} instead of the vowel v_{12}.

If a symbol equals to one of the tonal graphemes, then such a symbol is a following one after a consonant or vowel. Hence, by only one symbol of any tonal grapheme, the word cannot be formed.

[Rule 5] No boundary of word unit is at a point immediately before any tonal symbol.

In the expression (1) or (2), it is known that the consonant of C consist of two letters, which are represented by the two symbols for the following combination: กร.กล.กว.กร.ตร. คล.คร.ปร.ปล.พร.ผล.ผร.พล.บร.คว, and ทร, where that symbol combination is called a double consonant, and the combinations above are denoted by c_1c_{32}, c_1c_{30}, c_3c_{32}, c_2c_{32}, c_3c_{30}, c_2c_{30}, $c_{13}c_{32}$, $c_{24}c_{32}$, $c_{24}c_{30}$, $c_{25}c_{32}$, $c_{26}c_{30}$, $c_{25}c_{30}$, c_3c_{43}, c_2c_{43}, and c_1c_{43} respectively.

Consequently, if a set of the double consonant above is denoted by C_d, then

$$C_d = \{c_1c_{32}, c_1c_{30}, \ldots, c_1c_{43}\} \quad (14)$$

[Rule 6] If two adjacent symbols are a double consonant, those two symbols are recognized and formed as a consonant. Hence, C in expression (1) or (2) is equivalent to two symbols of C; for example, พระ (pra) is denoted by the double consonant $c_{25}c_{32}$ and a vowel v_6.

Consider another adjacent two symbols, here denoted by c_ic_j. If $c_ic_j \notin C_d$, then c_ic_j should be considered as a word in which a vowel is omitted in the expression (2) when the symbols are a recognizable word in the lexical unit.

[Rule 7] If the two adjacent symbols are not a double consonant and form a recognizable word as a morpheme, the two symbols are a word without the vowel symbol in the form [C+V+C]; for example, a word คน (khon) is a string c_2c_{21}, where no vowel symbol is used.

In Thai writing system, other miscellaneous marks including numeral as shown in Fig. 2(d) and (e) are used as follows:

(1) To abbreviate a long name or title, a grapheme ' ๆ ' is used. However, it is ignored in our Thai syllable recognizer.

(2) To indicate that the preceding word or group of words should be repeated, a grapheme ' ๆ ' is used, However, it also may be ignored in the recognizer.

(3) To show one, sometimes two or more silent letters for the consonant, a grapheme ' ์ ' is written in a position above a preceding letter. The grapheme is denoted by a symbol s_3.

(4) Numeral, punctuation, and other graphemes are ignored in the recognizer.

Above explanation (3) leads to next rule.

[Rule 8] No boundary of word unit is in a point immediately before a symbol s_3.

The morphological analyses in the following sections are advanced and evaluated on the basis of the rules defined above.

3. Longest-Match Method based on Syllable Formation Rules

In the first phase of experiment of segmentation for Thai sentences, the right-directed longest-match method has been used. The number of Thai input sentences for an experiment consist of 20,631, which is full-text of KTSD inputted from a copy of Khurusapha [12, 14]. It has already been segmented correctly by the handiwork of an expert, and the delimiter "/" has been inserted between the words. This delimiter is used for determining whether the sentence has properly segmented or not. The characteristics of input text is shown in Table 1.

The dictionary used for the experiment is made up arbitrarily from the previously segmented statements. A data structure on the main memory is shown in Fig. 4. Figure 4 shows that an index of consonants in the dictionary consists of only 44 entries with each pointer pointing to a group of words corresponding to each con-

Table 1 Characteristics of input text of KTSD.

Number of sentences	20,631
Average words per a sentence	11.8
Average symbols per word	4.7
Number of main entries in the dictionary	20,475

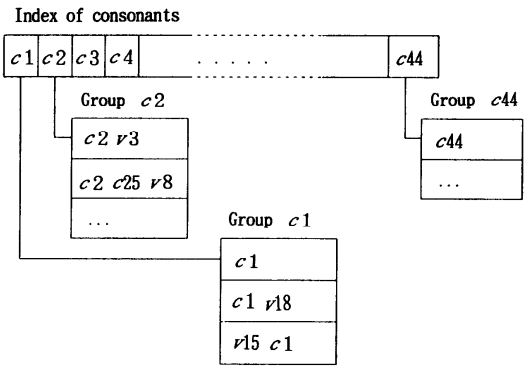

Fig. 4 Data structure of dictionary.

sonant. For the each region in the group, the words which the first consonant is belonging to its group have been stored. The number of regions in each group is variable in size, and the length of each word also is variable. When the word of a reversed type based on Rule 2 is referred to the dictionary, the search must be performed using the first consonant of the second symbol.

Two experiments for segmenting Thai sentences of the KTSD are carried out as follows; (1) Using the ordinary longest-match method with reference to a dictionary first, that is not based on the syllable formation rules defined in Section 2, (2) Using the longest-match method which employs the back-tracking function based on the syllable formation rules with necessarily reference to a dictionary.

3.1 Ordinary Longest-Match Method

From the experiment of segmentation using the longest-match method, the result as shown in Table 2 was derived.

A detailed example of segmentation for unsuccessful cases is shown in Fig. 5. As a result of segmentation, it is found that the ratio of automatic segmentation by using the ordinary right-directed longest-match is 95.2% in terms of sentences, and its ratio is about 15% higher than the result of Japanese sentences by approximately 80% in general [13].

Table 2 Result of segmentation for KTSD.

Input sentences	20,631
Frequency of reference of dictionary	243,918
Not segmented sentences	998
Ratio of segmentation	95.2%

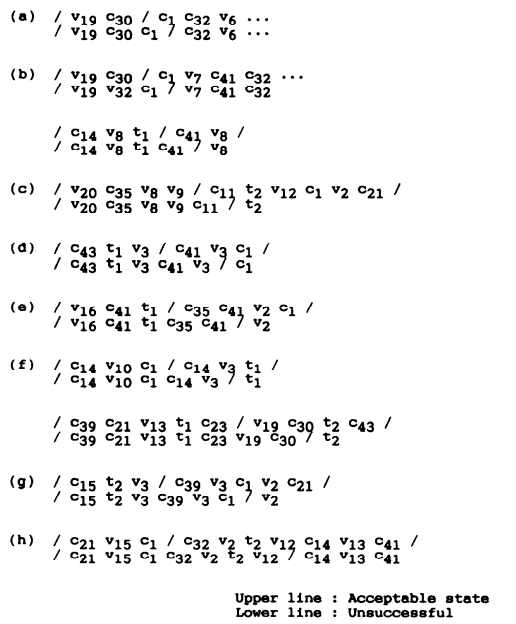

Fig. 5 An example of unsuccessful cases by the ordinary longest-match method.

The characteristics of unsuccessful cases are mainly classified into two categories; one is segmented by the word form [C+V+C] for the sequences [C+V] plus [C+V] or [C+V] plus [C+V+C]; for example, (a), (b), and (c) in Fig. 5. The second is when a syllable completely corresponds to the most left-side syllable embedded in a word in which the word is formed by several syllables or the compound word like (d), (e), (f), (g), and (h) in Fig. 5.

3.2 Syllable Longest-Match Method

Table 3 summarizes the behavior of unsuccessful cases in the experiment using the ordinary longest-match method.

Let $d=m$ when the result is segmented correctly except m symbols; for example, $d=1$ in the case (a), (b), and (c) in Fig. 5.

By analyzing Table 3, it is found that the ratio of correct segmentation can be increased by the method based on the syllable formation rules in Section 2. Consider the longest-match method incorporated with the function of back-tracking when the segmentation cannot be carried out any more. A Syllable Longest-Match method (SLM), which employs the back-tracking function for bringing the analysis back to the preceding symbol based on syllable formation rules, especially, Rule 1, 3, 5, and 8 with respect to the rules of word boundary, has been proposed. this SLM also refers to a dictionary.

It is found that a ratio of segmentation by the Syllable Longest-Match method is 98.0%, which is

Table 3 Summary of unsuccessful cases.
d: Difference in the number of symbols

	d=1	d=2	d=3	d=4
(A) Consonant	227	42	29	22
(B) Vowel	451	11	9	13
(C) Tone or others	135	31	0	2
(A)+(B)+(C) Total	813	84	29	37
(B)+(C)	586	42	9	15

d>4: 35, Total 998

2.8% higher in ratio of the number of sentences than that by the previous method. Therefore, the Syllable Longest-Match method is found to be an effective scheme for Thai word segmentation.

4. Thai Syllable Recognizer Model

To implement a machine for recognizing Thai syllable automatically, a syntax-directed program for Thai is devised, here called Thai syllable recognizer, which employs the function of automatic and consecutive segmentation for Thai sentences based on Thai syllable formation rules as shown in Section 2 and without reference to a dictionary.

In the first place, the two types which forms formulas (1) and (2) are classified into 6 models depending on the rules of appearance of symbols, and each model is built up by the non-deterministic finite automaton.

Each element in set CS of consonant symbols is defined as c_c $\{c_c: c_c \in CS\}$. Also, the double consonant for two adjacent symbols $c_i c_j$ is defined as a set C_d (14). Therefore, if c_i c_j is not an element of set C_d; c_i $c_j \notin C_d$, then c_i and c_j correspond to a vowel abbreviated model according to Rule 7. Because a symbol is fetched by the recognizer with one by one, set C_d must be separated to the subsets composed of each symbol in the set. The following subsets are introduced for two adjacent consonants.

(1) The sets are classified into 3 groups depending on the kind of combination of two symbols as follows:

In the double consonant,

(i) Subsets C_{i1} and C_{j1}: First group
If $C_{i1}=\{c_1, c_2, c_3\}$
then $C_{j1}=\{c_{30}, c_{32}, c_{43}\}$.

(ii) Subsets C_{i2} and C_{j2}: Second group
If $C_{i2}=\{c_{13}, c_{24}, c_{25}\}$,
then $C_{j2}=\{c_{30}, c_{32}\}$.

(iii) Subsets C_{i3} and C_{j3}: Third group
If $C_{i3}=\{c_{26}\}$,
then $C_{j3}=\{c_{30}\}$.

(2) First consonant does not exist in that double consonant.

(i) Subset C_w: $C_w=CC-\{C_{i1} \cup C_{i2} \cup C_{i3}\}$

(3) Second consonant does not exist in the set of second symbols within double consonant.

(i) Subset C_{r1}: $C_{r1}=CS_e-C_{j1}$
(ii) Subset C_{r2}: $C_{r2}=CS_e-C_{j2}$
(iii) Subset C_{r3}: $C_{r3}=CS_e-C_{j3}$

Here, each elements is discriminated as follows;

a_1: $a_1 \in C_{i1}$, b_1: $b_1 \in C_{j1}$, a_2: $a_2 \in C_{i2}$, b_2: $b_2 \in C_{j2}$,

a_3: $a_3 \in C_{i3}$, b_3: $b_3 \in C_{j3}$, and w: $w \in C_w$,

r_1: $r_1 \in C_{r1}$, r_2: $r_2 \in C_{r2}$, r_3: $r_3 \in C_{r3}$ respectively.

Assume that the tonal marks are included to a preceding consonant or vowel, and represented by the denotation of the preceding symbol.

In the following state diagrams for the recognizer model, q_n ($n=0, 1, \ldots, 36$) and q_0 in the circles represent a state and a start state respectively. The accept states also are distinguished by being drawn with a double circle. And, the transition functions are omitted in this paper.

Each model in the recognizer is represented as follows:

[A] Double consonants and abbreviated vowel model: [C+C]

This model provides the [C+C] case either two consonants are a double consonant or the vowel is abbreviated as shown in Fig. 6.

[B] General Model: [C+V+C] or [C+V]

This model provides [C+V] and [C+V+C] cases of which the symbol at the beginning of monosyllable is the consonant as shown in Fig. 7. Where VS_{n2} is defined as follows:

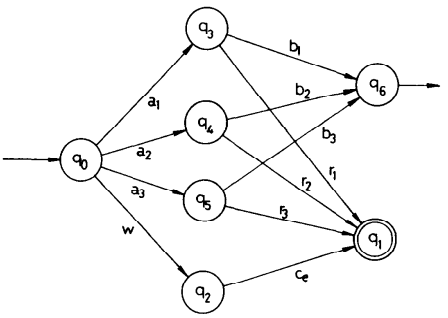

Fig. 6 [A] The double consonant or abbreviated vowel model.

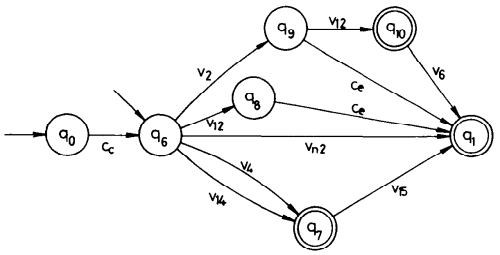

Fig. 7 [B] The general model.

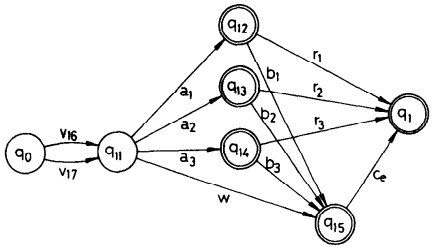

Fig. 8 [C] Deformed model (1).

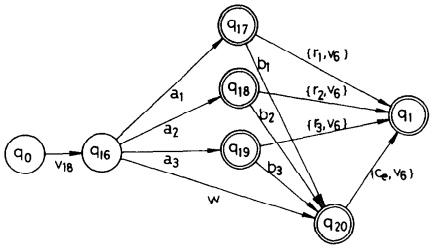

Fig. 9 [D] Deformed model (2).

$$VS_{n2} = VS_n - \{v_2, v_4, v_{12}, v_{14}\}$$

Each element in set VS_{n2} is v_{n2}, $[v_{n2}: v_{n2} \in VS_{n2}]$.

[C] Deformed model (1): [C+V] and [C+V+C]

This model provides the [C+V] and [C+V+C] cases of which the symbol at the beginning of monosyllable is v_{16} or v_{17} as shown in Fig. 8.

[D] Deformed model (2): [C+V] and [C+V+C]

This model provides the [C+V] and [C+V+C] cases of which the symbol at the beginning of monosyllable is v_{18} as shown in Fig. 9.

[E] Deformed model (3): [C+V] and [C+V+C]

This model provides the [C+V] and [C+V+C] cases of which the symbol at the beginning of monosyllable is v_{19} as shown in Fig. 10.

[F] Deformed model (4): [C+V] and [C+V+C]

This model provides the [C+V] and [C+V+C] cases of which the symbol at the beginning of monosyllable is v_{20} as shown in Fig. 11.

5. Automatic Segmentation using Thai Syllable Recognizer

5.1 Implementation

On the basis of the previous models, a syntax-directed recognizer, which enables to automatic and consecutive segmentation for recognizable portion of words in Thai sentence has been designed. A outline of recognizer machine is shown in Fig. 12. The machine is implemented as a program (about 1200 steps) which is coded by the PL/I programming language, and it can

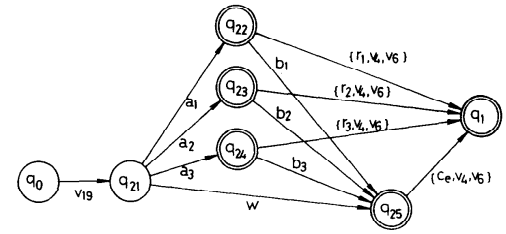

Fig. 10 [E] Deformed model (3).

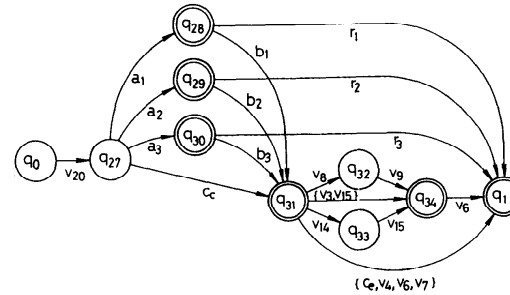

Fig. 11 [F] Deformed model (4).

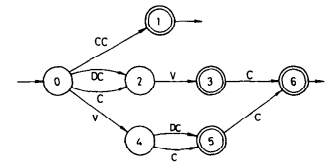

Fig. 12 Model of thai syllable recognizer.

be operated on FACOM M-780 main-frame computer.

The model of the Thai syllable recognizer was introduced in the preceding section, where CC, DC, C, and V imply words with abbreviated vowel, double consonant, general consonant and vowel(s) respectively. Also, each numeral in the circles represents a state, and the accept states are distinguished by being drawn with a double circle in the recognizer.

One of the most important characteristics in this recognizer is that no dictionary is used, whereas the unsegmental characteristics generally need the use of dictionary for segmentation.

5.2 Segmentation

The use of previous recognizer which segments a sentence for the same text, KTSD, in an experiment of the longest-match method has been attempted. To analyze the result of segmentation in detail, especially, about the ratio of segmentation and the sequence of the appearance of symbols in the unsuccessful cases, a Thai-Thai dictionary composed of 31,202 main entries has been used [5, 18].

Segmented words obtained as the result of segmenta-

tion by the recognizer are specialized as monosyllables composed of [C+V] or [C+V+C], whereas the words before segmenting the sentences for KTSD consists of monosyllables or compound words. Therefore, a criterion of a recognizable unit as a syllable is;

Original sentence: /v_{19} c_{30} /v_{16} c_{42} t_2 c_1 c_{32} v_6 c_{14} v_{10} t_1 c_{42}/

Input sentence: v_{19} c_{30} v_{16} c_{42} t_2 c_1 c_{32} v_6 c_{14} v_{10} t_1 v_{42}

Segmented syllable by recognizer: /v_{19} c_{30}/ v_{16} c_{42} t_2/ c_1 c_{32} v_6/ c_{14} v_{10} t_1 c_{42}/

Criteria of recognition are as follows:

(1) Delimiters at position in the original sentence are corresponded to the delimiters of the segmented sentence.

(2) Each symbol embedded in a string surrounded by both delimiters,, "/" and "/", has the sequence [C+V] or [C+V+C] based on the rules in Section 2. If the above criteria are satisfied, a string between both delimiters "/" is to be recognized as a syllable, and this syllable is searched in the main entries of dictionary in order to confirm whether it is registered or not. The result of experiment is shown in Table 4.

From the result, the segmentation ratio in terms of sentences becomes very much lower than by 95.2% of the ordinary longest-match method. Some of the biggest reasons for unsuccessful cases is as follows:

(1) Segmentation by string type [C+V] according to Fig. 3(a) is happened in spite of type [C+V+C]; for example, จาก and คุณ are case 4 and 11 respectively.

(2) Segmentation by string type [C+V] according to Fig. 3(b) happens in spite of the type [C+V+C]; for example, เป็น and เขียว are case 11 and 14 respectively.

5.3 Heuristic Approach

As for the key reasons of improperly segmented words, it is found that the rule for type [C+V+C] is not matched since all of rules are only deduced from Thai grammatical rules and neither the characteristics in the case of above (1), nor the case of (2) above has not been installed into the recognizer.

On the basis of heuristics, the following rules are deduced:

(1) When v_5, or v_6 appeared, those symbols imply a last symbol at end of the monosyllable. Then the following symbol, of course, should be a consonant in the top position in the monosyllable.

(2) Let subset VS_{n3} be Vowel Symbols given by

$$VS_{n3} = VS_n - \{v_2, v_4, v_5, v_6, v_{12}, v_{14}\}$$
$$= \{v_3, v_7, v_8, v_{10}, v_{11}, v_{13}, v_{15}\},$$

instead of the set VS_{n2} in [B] General model.

If input symbol s is equal to V_{n3} $\{V_{n3}: V_{n3} \in VS_{n3}\}$, then, a following consonant C_e in the set CS_e (12) sometimes is added.

(3) After the vowel v_4 and v_{14} in the general model: [C+V] or [C+V+C], a consonant c_e in the set CS_e sometimes is added.

(4) The following symbol after the vowel v_4 in the deformed model (3) sometimes is a consonant c_e in the set CS_e.

(5) In the deformed model (4), a consonant c_e in the set CS_e sometimes is added after the vowels v_4, v_7, v_9, and v_{15}.

Here, the first revised model, [B'] Revised model (1), which incorporates three features of above rules (1), (2), and (3) for the general model as shown in Fig. 13 is proposed. The second revised model, [E'] Revised model (2), which incorporates two features of the above rules (1) and (4) for the deformed model (3), and the third revised model, [F'] Revised model (3), which incorporates two features of the above rules (1) and (5) for the deformed model (4) are proposed as shown in Fig.

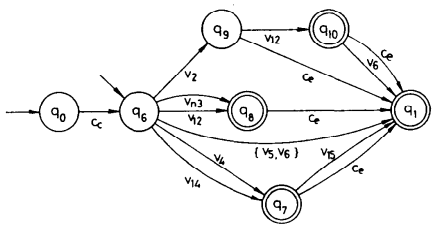

Fig. 13 [B'] Revised model (1) for the general model.

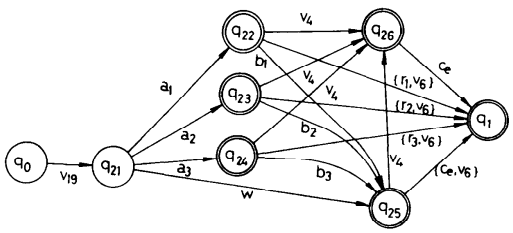

Fig. 14 [E'] Revised model (2) for the deformed model (3).

Table 4 Result of segmentation by recognizer.

Input sentences	20,631	
Number of words	252,619	
Not segmented sentences	10,406	
Number of generated words	417,179	
Not found words	178,915	(42.9%)
Segmentation Ratio (Sentence unit)		49.6%

Table 5 Result of experiment of revised model.

Not segmented sentences	1,269	
Number of generated words	401,577	
Not found words		
Thai-Thai dictionary	98,021	(24.4%)
KTSD dictionary	112,095	(27.9%)
Segmentation Ratio (Sentence unit)		(93.9%)

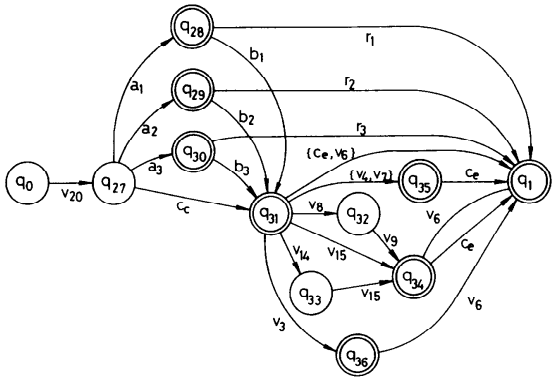

Fig. 15 [F'] Revised model (3) for the deformed model (4).

Class	Example	Frequency
(1)	/v_{20} c_1/ c_{30} t_1 v_{12}/	
	/v_{20} c_1 c_{30} t_1/ v_{12} ...	362
(2)	/c_7 c_{23} c_{32} v_2 c_1 c_{36}/ c_{44} v_2 c_{21}/	
	/c_7 c_{23}/c_{32} v_2 c_1/ c_{36} v_{15}/ v_2 ...	343
(3)	/c_{41} v_{10} c_{13} c_{32} c_{44} v_8/	
	/c_{41} v_{10}/c_{13} c_{32} v_{15}/ v_8 ...	37
(4)	/c_{35} c_{43} v_2 c_{11} v_8/	
	/c_{35} v_{12}/ v_2 ...	278
(5)	/c_{32} v_2 c_{25} c_{33}/ c_1 c_{43} v_3/	
	/c_{32} v_2 c_{25}/ c_{33} c_1/ v_{12} ...	25
(6)	... c_{16} c_{32} c_{32} c_{42}/ c_{44} v_2 c_{21}/	
	.....	74
(7)	/c_{41} v_{10} v_6/	
	.....	61
(8)	/v_{20} c_7 v_8/ c_{33} v_2 c_{23}/	
	/v_{20} c_7 v_8 v_9/ v_2 ...	39
(9)	/c_{16} c_{32} c_{17} c_{44} v_7 c_{37} c_{43} c_{32} ...	
	.....	4
(10)	/c_1 c_{32} c_{42} c_3 c_{43} v_3/	
	.....	46
		1269

Unit of frequency: Sentences
Upper line: Input string
Below line: Recognized pattern

Fig. 16 Enumeration of unsuccessful cases by the revised recognizer.

14 and Fig. 15 respectively.

The result of experiment using the revised model, is shown in Table 5. From the experiment, it is found that the segmentation ratio has attained at most to 49.6% by the recognizer based on the syllable formation rules only. By adapting the knowledge rules based on heuristics derived from the analysis of unsuccessful cases into exiting syllable formation rules, it is found that the ratio of segmentation has improved by 93.9%, and it is 44.3% higher than the previous experiment by recognizer whereas 4.1% lower than the method by the Syllable Longest-Match method.

Figure 16 shows the classification and its enumeration of improperly segmented sentences. The key reasons in the column 'class' in Fig. 16, which are classified into 10 classes, are as follows:

(1) The following two consonants after a vowel of [Rule 2], namely, by forming a type of '[V+C]+[C+V]', was recognized as the double consonant [C+C], whereas its word was formed by a combination of both monosyllable [C+V]s.

(2) The consonant c_{44} was recognized as a vowel v_{15} because the preceding syllable was [C+V+C+C] type, and the following one is [C+V].

(3) The following consonant c_{44} after the double consonant was recognized as a vowel v_{15}, namely, [C+C+V+C].

(4) The second consonant c_{43} embedded in the monosyllable was recognized as the vowel v_{12}.

(5) The second consonant c_{43} in the syllable "Double consonant+Vowel" was recognized as a vowel v_{12} when the preceding syllable was [C+V+C+C].

(6) Successive two consonants of c_{32} were appeared. It should be replaced by a vowel v_2 based on the grammatical rule of Thai in advance.

(7) The following two vowels after a consonant were appeared. They were not in Fig. 2 and syllable formation rules.

(8) The consonant c_{33} was recognized as a vowel v_9 according to $\delta(q_{32}, v_9)=q_{34}$ in Fig. 15, because the preceding symbol was a vowel v_8.

(9) They are the proper noun, which have no [C+V] or [C+V+C] type.

(10) They are the other cases like "Double consonant+Consonant+Double consonant+Vowel".

6. Concluding Remarks

One of Problems in mechanical processing for Thai is the language processing of recognizable portions of words in the morphological level. The characteristic of being unsegmental increases the complexity in the morphological analysis. The segmentation in the unsegmental or agglutinative language is carried out using the grammaticcal structure rules and the algorithms derived from heuristics depend on individual language along with the use of dictionary necessarily.

An experiment of segmentation as the input of 20,631 sentences based on the ordinary longest-match method using the dictionary consisting of 20,475 main entries of KTSD has been performed, then a Syllable Longest-Match method (SLM) based on the analysis derived from the result of preceding experiment has been devised. It is found that the back-tracking for one character is most effective by 98.0% in the ratio of segmentation.

Finite automaton model, called Thai syllable recognizer, for segmenting a sentence into monosyllables without references to the dictionary only has been proposed. By adapting the knowledge rules depending on heuristics derived from the analysis of unsuccessful cases into existing syllable formation rules, it is found that the ratio of segmentation can be obtained by 93.9%, and showed that the adaption of knowledge derived from the heuristics to the language processing depend on the individual language plays major role.

In the segmentation based on the syllable formation rules and its syllabic symbols, it is necessary to improve the efficiency of segmentation by using the dictionary along with the heuristic knowledge eliminating the

monosyllable with no meaning and of restoring word by synthesizing monosyllable automatically. Also, the features of dynamic self-learning should be equipped for the segmentation process to improve the ratio of correct segmentation.

References

1. ALLISON, G. H. Simplified Thai, *Nibondh, Thailand* (1973), 1-76.
2. HAAS, M. R. Thai system of writing, *American Council of Learned Societies, Washington, D.C.* (1956) 1-40.
3. WAROTAMASIKKHABIT, V. Problems in using the Thai alphabet in computing, *Proc. of the 1984 Southeast Asia Regional Computer Conference. SEARCC 84, SEACC* (1984) 18/1-8.
4. SHIBAYAMA, M. and HOSHINO, S. Implementation of an intelligent Thai computer terminal, *J. Inf. Process.*, **8**, 4 (March 1985), 300-306.
5. SHIBAYAMA, M., HOSHINO, S. and ISHII, Y. A Comparative Study of the Characteristics of Input Methods for Thai, *Proc. of the Regional Symposium on Computer Science and its Applications, NRCT-JSPS, Thailand* (Feb. 1987), 19.1-19.18.
6. SHIBAYAMA, M. Thai Syntax Analysis using the Case Frame, *OIU Journal of International Studies* (Oct. 1989), 107-117.
7. VORASUCHA, V. and TANAKA, H. Thai syntax analysis based on GPSG, *Jour. of JSAI*, **3** (1988), 78-85.
8. TANAKA, Y. and KOGA, K. Automatic Segmentation of Hiragana Strings Appearing in the Japanese Sentences (in Japanese), *J. IPS Japan*, **22** (1981), 242-247.
9. SUGITA, S. TEXT PROCESSING OF THAI LANGUAGE ==THE THREE SEALS LAW==, *Proc. of COLING80* (1980).
10. SHIBAYAMA, M. Input/Output Methods for Thai, *Southeast Asian Studies*, **25**, 2 (1987), 279-296.
11. SHIBAYAMA, M. and ISHII, Y. Computerization of the Thammasat Version of the Kotmai Tra Sam Duang, Studies on the Multi-Lingual Text Processing for Assisting Southeast Asian Studies edited by Shibayama, M., The Center for Southeast Asian Studies of Kyoto University, Report by grants for scientific research from Japanese Ministry of Education (Mar. 1988), 46-50.
12. ISHII, Y., SHIBAYAMA, M. and AROONRUT, W. Datchani Kotmai Tra Sam Duang (in Thai, The Computer Concordance to The Law of Three Seals), Amarin Publication, **5**, Thailand, p. 4850 (Aug. 1990).
13. TANAKA, H. Fundamentals of Natural Language Analyses (in Japanese), *Sangyo-Tosyo* (1989), 133-137, 138-139.
14. KHRUSAPHA Kotmai Tra Sam Duang (in Thai), Thammasat Univ., **5**, p. 1775 (1962).
15. ISHII, Y. Introductory Remarks on the Law of Three Seals (in Japanese), *Southeast Asian Studies*, **6**, 4 (1969), 155-178.
16. KAWABE, T. Thai Fundamentals (in Japanese), *Daigaku Syorin* (1980), 5-12.
17. NAGAO, M., TSUJII, J., YAMADA, A. and TATEBE, S. Data Structure of a Large Japanese Dictionary and Morphological Analysis by using It (in Japanese), *Trans. IPS Japan*, **19**, 6 (1978), 514-521.
18. Photchana nukorm Thai, Chabap Ratchabandit sathan, (in Thai), Khrungtheep, Samnakphim Askonchaoenthat (1982) (Thai 2525).

(Received September 24, 1991; revised June 1, 1992)