

タイ語古典テキストの自動分かち書きについて

柴山 守  
(大阪国際大学 経営情報学部)

星野 聡  
(京都大学大型計算機センター)

タイ語古典テキストの語単位への自動分かち書きについて述べる。タイ語の音節構造や正書法の分析から、タイ語音節形成規則を定義し、音節の区切りまでバックトラックを起こす音節最長一致法を提案、実験・評価し、98.0% (文単位) の結果を得た。また、音節形成規則に基づいて、有限オートマトンを用いたタイ語音節認識機構モデルを構築し、実現した。タイ語「三印法典」(The Three Seals Law、約1,700頁、約2万文、1805年編纂)テキストに対して、辞書を使用しない単音節単位への自動分割の実験を行った。失敗例の解析からの発見的的手法による改良型モデルでは、同一テキストの自動分割において93.9%の分かち書き率を示した。

Automatic Segmentation of Thai Ancient Text

Mamoru SHIBAYAMA

Satoshi HOSHINO

Faculty of Management  
and Information,  
Osaka International  
University

Data Processing Center,  
Kyoto University

3-50-1 Sugi, Hirakata,  
Osaka 573-01, Japan

Yoshida Honmachi, Sakyo,  
Kyoto 606, Japan

Thai automatic segmentation which separates a sentence into words for an ancient text is discussed. Based on the syllable formation rules derived from linguistic analysis of the syllabic structure and the orthography of Thai, a Syllable Longest Match (SLM) method with back-tracking function has been proposed. The ratio of correct segmentation using this method was a 98.0% in terms of sentence. Furthermore, a Thai syllable recognizer which segments a sentence into monosyllables without reference to the dictionary also has been proposed and implemented. A revised recognizer based on the heuristic approach derived from the analysis of unsuccessful cases gives a ratio of correct segmentation of 93.9% for the Three Seals Law.

## 1. はじめに

自然言語処理における形態素解析やデータベース構築、計算機索引作成の際には文や語の適切な単位への分かち書きが必要となる。この分かち書き（セグメンテーション）は、手作業に依存すると莫大な時間や費用を必要とし、結果が統一性に欠けるなどの問題を生ずる。計算機の利用によって短時間に正確に分かち書きできればその効果は極めて大きい。また、分かち書きは、対象とする言語が日本語のような膠着的・未分割的特徴をもつ場合特に重要であり、従来から辞書を使用しての最長一致法が用いられてきた。[1]

タイ語は、語あるいは文間には空白が入れられず連続して表記され、日本語と同様非分割的である。また、句読点はほとんど使用されず、正書法では子音と母音が逆の順序に書かれたり、字母の大きさや形は互いにまちまちであるなど、計算機処理において前述の諸問題に直面する。[2]

本稿では、このタイ語の特徴について分析を行い、最初に音節形成規則を導出、定義する。そして、タイ語「三印法典」(The Three Seals Law、KTSD: Kotmai Tra Sam Duang、約1,700頁、約2万文、1805年編纂)[3] テキストについて、従来から用いられてきた最長一致法を用いて分かち書きの実験を行った。この結果について分析を行い、音節形成規則に基づいて文字素単位にバックトラックを起し、音節単位に分かち書きする音節最長一致法(SLM: Syllable Longest-Match method)を提案し、その実験結果について示す。

また、分かち書きの処理過程において辞書の利用は一般的であるが、辞書のデータ構造やアクセス時間の短縮は解析効率を向上させる上で重要である。解析に辞書が不必要ならば解析時間の短縮に大きく貢献する。この実現のために、単音節単位への分かち書きを行うタイ語の音節認識機構モデルを構築し、タイ語音節認識機構を実現した。このモデルを用いた分かち書き実験について報告する。また、結果の解析によって得られたヒューリスティックな規則に基づいて音節認識機構を改良、その実験結果について報告する。

## 2. 音節形成規則

タイ語音節を形成する特徴は、つぎの通りである。(1)タイ語は、表音文字で、“子音+母音”、または“子音+母音+子音”に5つの声調が加わる。(2)タイ文字は、子音44字、母音32字で表されるが、母音は複数の字母で表されることがある。タイ語の字母について図1に示す。(3)子音字母をC、母音字母をVで表すと、音節は[C+V]または[C+V+C]で形成される。これを文単位で音節の列として見てみると

- ① [C+V] + [C+V] + . . .
- ② [C+V] + [C+V+C] . . .
- ③ [C+V+C] + [C+V] . . .
- ④ [C+V+C] + [C+V+C] . . .

の4種類になる。(4)つぎに、図1中でc33とv9、c43とv12、c44とv15は同一の字母である。

(5)また、母音32字について、音節構造と正書法による字母の出現順序をみると必ずしも一致しない。第1字目が子音字母で始まる(一致する)場合の母音字母の出現順序を図2(a)に、母音字母で始まる(一致しない)場合を図2(b)に示す。図中のハイフン‘-’は、その位置に子音が挿入されることを示している。また、本稿では、各々の字母を音節の音素に対応して、文字素と呼ぶことがある。

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
ก	ข	ช	ฌ	ฎ	ฏ	ด	ต	ถ	ศ
C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
ค	ฆ	ฉ	ซ	ญ	ฎ	ฏ	ด	ต	ถ
C21	C22	C23	C24	C25	C26	C27	C28	C29	C30
ท	ธ	น	บ	ป	ผ	ฝ	ภ	ผ	ล
C31	C32	C33	C34	C35	C36	C37	C38	C39	C40
ฬ	ร	ย	ว	ศ	ษ	ส	ห	ฬ	อ
C41	C42	C43	C44						
ุ	ู	ัว	อ						

(a) 子音

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
null	◌	◌	◌	◌	◌	◌	◌	◌	'
V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
◌	◌	◌	◌	◌	◌	◌	◌	◌	◌

(b) 母音

t1	t2	t3	t4
,	◌	◌	+

(c) 声調記号

S1	S2	S3	S4
◌	◌	◌	sp

(d) 特殊記号

n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
◌	◌	◌	◌	◌	◌	◌	◌	◌	◌

(e) 数字

図1 タイ語字母表

ケース	1st	2nd	3rd	4th	5th
1	-	v2	-		
2	-	v2	v12		
3	-	v2	v12	v6	
4	-	v3			
5	-	v5			
6	-	v6			
7	-	v7			
8	-	v8			
9	-	v10			
10	-	v11			
11	-	v12			
12	-	v13			
13	-	v14			
14	-	v14	v15		
15	-	v4	-		
16	-	v4	v15		
17	-	v15			

(a) 音素・文字素順が一致する場合

1	v16	-			
2	v17	-			
3	v18	-			
4	v18	-	v6		
5	v19	-			
6	v19	-	v4		
7	v19	-	v6		
8	v20	-			
9	v20	-	v3		
10	v20	-	v3	v6	
11	v20	-	v4		
12	v20	-	v6		
13	v20	-	v7		
14	v20	-	v8	v9	
15	v20	-	v8	v9	v6
16	v20	-	v14	v15	
17	v20	-	v14	v15	v6
18	v20	-	v15		
19	v20	-	v15	v6	

(b) 音素・文字素順が一致しない場合

図2 母音字母の出現順序

以上のような子音、母音の特徴から以下に示す音声形成規則が導かれる。

【規則1】音節を形成する音素順序から見て、一般に母音字母の直前で語境界を形成しない。すなわち、[C+V]または[C+V+C]のシーケンスにおいて、子音Cと母音Vの間で分かち書きされない。

ここで、子音Cに続いて出現する母音Vは、その母音が書かれる位置によって4つのグループに分けられる。それらは、①母音字母Vが子音Cの直前に書かれる場合、②子音Cの直後に書かれる場合、③子音Cの上部に書かれる場合、④子音Cの下部に書かれる場合である。

【規則2】母音字母v16, v17, v18, v19, v20は音節の文字素列の中で先頭の文字素となる。例えば、 $\text{p}^{\text{h}}\text{ai}$  (pai:行く)はv16, c24である。

【規則3】規則2を除いて、他の3グループの母音字母は子音の直後に出現する。例えば、 $\text{ca}$  (未来形を示す)は、c7, v6である。

【規則4】同一字母により子音、母音が表される場合、その区別は文字素が出現する位置によって決定される。例えば、 $\text{waa}$  (接尾辞)の先頭文字は、子音c43であり、母音ではない。続く母音はv3である。

【規則5】声調記号の直前で語境界を形成しない。

【規則6】もし接続する2文字が、次に示す文字列のいずれかであれば、2重子音であり一つの子音とみなす。2重子音は、c1c32, c1c30, c3c32, c2c32, c3c30, c2c30, c13c32, c24c32, c24c30, c25c32, c26c30, c25c30, c3c43, c1c43の文字列である。

【規則7】接続する2文字素が二重子音でなければ、音節[C+V+C]形において母音が省略されている1音節を示す。例えば、 $\text{khon}$  (人)は、c2, c21である。

また、タイ語正書法においては、次のような特殊記号が使用される。

- (1) 記号 $\text{h}$ は、長い名前やタイトルを省略する時に使用される。
- (2) 記号 $\text{h}$ は、直前に出現した語あるいは語の連なりを繰り返す場合に使用される。
- (3) 記号 $\text{h}$ は、直前の文字素が発声されないことを示す。

【規則8】特殊記号 $\text{h}$  (s3)の直前で分かち書きされない。

以上に示した、8つの音節形成規則に沿って以下に示す音節最長一致法や音節認識機構のモデル化に適用し、実験を進める。

### 3. 音節形成規則による最長一致法

#### 3.1 入力テキストと実験方法

日本語の形態素解析では従来から右向き最長一致法が用いられてきた。これをタイ語テキストの分かち書きに使用する。実験に用いたタイ語テキストはタイ語「三印法典」(KTSD: Kotmai Tra Sam Duang, 約1,700頁、約34,000行、1805年編纂)で約450年間の法律・布告を含む古典テキスト[2]である。また、この実験で参照する辞書は、すでに法典の専門家によって分かち書き・構成されたKTSDに基づいて作成した見出し語のみから構成される。これらの概要を表1に示す。

実験方法は、2通りである。

- (1) 音節形成規則を使わないで最長一致法でおこなう。
- (2) 上の実験結果を分析、つぎに音節形成規則に基づいたバックトラック機能を組み込んだ最長

表1 入力テキストの特徴  
タイ語「三印法典」

ステートメント数	20,631	文
平均語数/文	11.8	語
文字数/語	4.7	字
辞書エントリー数	20,475	語

表2 最長一致法による実験結果

辞書参照回数	243,918	回
分割失敗文数	998	文
分かち書き率	95.27	%

一致法について実験する。

なお、辞書は主記憶上に置かれ、データ構造は子音44字に対応するインデックス部と当該子音で始まる語が格納されるデータ域で構成される。

### 3.2 最長一致法

音節形成規則を使わない最長一致法による分かち書きの実験結果を表2に示す。

表2の実験結果が示すように、辞書を用いた右向き最長一致法において95.27%（文単位）の結果を得た。その結果は、日本語文でほぼ80%である[1]ことから比較すると約15%高い値である。

この実験における不成功の例を図3に示す。不成功例の特徴は、主に2つに分けられる。

(1) 語順が[C+V]+[C+V]または[C+V]+[C+V+C]に対して、[C+V+C]で分かち書きされた(図3(a)など)。(2) 辞書エントリーが複数音節または複合語から成る場合、複数音節で分かち書きされた(図3(h),(g)など)。

(a) / v19 c30 / c1 c32 v6 ... / v19 c30 c1 / c32 v6 ...	(b) / v19 c30 / c1 v7 c41 c32 ... / v19 v32 c1 / v7 c41 c32
(c) / v20 c35 v8 v9 / c11 t2 v12 c1 v2 c21 / / v20 c35 v8 v9 c11 / t2	(d) / c43 t1 v3 / c41 v3 c1 / / c43 t1 v3 c41 v3 / c1
(e) / v16 c41 t1 / c35 c41 v2 c1 / / v16 c41 t1 c35 c41 / v2	(f) / c14 v10 c1 / c14 v3 t1 / / c14 v10 c1 c14 v3 / t1
(h) / c21 v15 c1 / c32 v2 t2 v12 c14 v13 c41 / / c21 v15 c1 c32 v2 t2 v12 / c14 v13 c41	(g) / c15 t2 v3 / c39 v3 c1 v2 c21 / / c15 t2 v3 c39 v3 c1 / v2

図3 最長一致法における分かち書き失敗例 [上段:正しい分かち書き、下段:失敗例]

つぎに分かち書きの失敗例について、失敗例におけるずれの文字数をd(例えば、図3の(a),(b),(c)のように1文字の違い(ずれ)をd=1)として、その集計結果を表3に示す。

### 3.3 音節最長一致法

前節の最長一致法による分かち書きの実験結果、図3及び表3から、音節形成規則を適用して分かち書きを行うことにより、分かち書き率が向上することが判る。そこで、セグメンテーション過程において、解析不能になった時点で音節形成規則を適用してバックトラックを起こす音節最長一致(SLM: Syllable Longest-Match)法を提案する。この音節形成規則の規則1、3、5、8を適用したSLM法を用いた結果、98.0%の分かち書き率を得た。そこで、この方法は、タイ

語文の分かち書きに有効であることが確認された。

表3 失敗例の集計結果

	d=1	d=2	d=3	d=4
(A) 子音(Consonant)	227	42	29	22
(B) 母音(Vowel)	451	11	9	13
(C) 声調、その他	135	31	0	2
(A)+(B)+(C) 合計	813	84	29	37
(B)+(C)	586	42	9	15

#### 4. タイ語音節認識機構(Thai Syllable Recognizer)

図2の母音字母の出現順序及び音節形成規則に基づいて、タイ語テキストを自動・連続的に音節単位にセグメンテーションを行うタイ語音節認識機構のモデルを作成した。このモデルではセグメント時に一切の辞書を参照することなく分かち書きできる。モデルには有限オートマトンを用い、5つのサブモデルから構成される。(1) [C

+C] (2重子音、または母音省略)モデル、(2) [C+V]または[C+V+C]モデル、(3)変形[C+V]モデル[母音 v16,v17]、(4)変形[C+V]モデル[母音 v18]、(5)変形[C+V]モデル[母音v19 ,v20]である。概略モデルをを図5に示す。

図5において、記号CC, DC, C, Vは各々母音省略型語、2重子音、子音、母音を表し、また、円内の数字は状態を、2重円は終了状態を示す。

各サブモデルの代表的な例と図5に示したモデルについて述べる。

サブモデル(1)では、2重子音または母音省略語を扱う。พม (phom: 僕)は、母音省略型で規則7を適用し、図5では、状態0→1で遷移する。พธ (phra: 接頭語)は、先頭2文字が2重子音 c25c32でサブモデル(1)によって一つの子音とみなし、続いて'ธ'は母音v6で[C+V]型のためサブモデル(2)を適用して、図5では状態0→(DC)→2→3の状態遷移となる。また、サブモデル(2)では、พท (ma:k :たいへん)のような[C+V+C]型を扱い、状態0→(C)→2→3→6の状態遷移をたどる。サブモデル(3)は母音v16,v17に、サブモデル(4)は母音 v18、サブモデル(5)は母音 v19,v20で始まる音節、すなわち音素・文字素順が一致していないモデルである。ไป (pai :行く)、ใจ (cai : 心)は各々v16c24、v17c7でサブモデル(3)となり、図5では状態0→4→5となる。ใจ (do:i : ~で)はv18,c11,c33の[C+V+C]型であり、図5では状態0→4→5→6の状態遷移となる。ใจ (pen : ~である)はv20,c24,c21の[C+V+C]型であり、図5では(4)と同じ状態0→4→5→6の状態遷移となる。

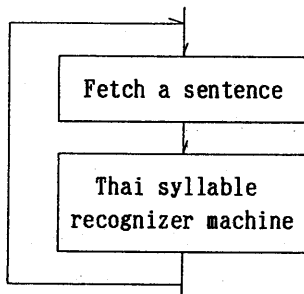


図4 タイ語音節認識機構の構造

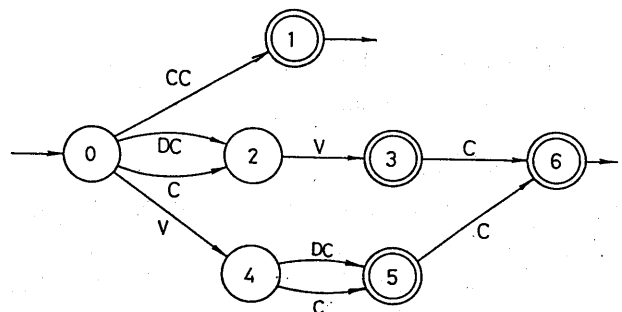


図5 音節認識機構モデル

#### 4. 1 音節認識機構モデルによる実験

実験には最長一致法と同じタイ語「三印法典」を使用する。原始文（専門家が文に分割している）、入力文と分かち書きされた文は次のようになる。

原始文： / v19 c30 / v16 c42 t2 c1 c32 v6 c14 v10 t1 c42 /  
 入力文： v19 c30 v16 c42 t2 c1 c32 v6 c14 v10 t1 c42  
 分かち書き文： / v19 c30 / v16 c42 t2 / c1 c32 v6 / c14 v10 t1 c42 /  
 実験結果を表4に示す。

表4 音節認識機構による実験結果

入力文	20, 631文
入力語数	252, 619語
分割されなかった文	10, 406文
生成語数	417, 179語
辞書に未登録	178, 915語 (42.9%)
分割されなかった語	59, 571語
分かち書き率(文単位)	49.6%

また、セグメントされた結果である単音節が辞書に存在するか否かを確認するために、タイタイ辞書の見出し語31、202語から作成した計算機可読形辞書を参照する。

この実験結果から判るように、最長一致法と比較して悪い結果である。この主な理由は、(1)語形が[C+V+C]であるにも関わらず、[C+V]で分かち書きされた。例えば、図2(a)中のケース9、17におけるc2 v10 c22、c44 v15 c1などである。(2)音節形成規則に存在しないような、紛らわしいケースである。例えば、c44 c33 v11 t1のような語である。

#### 4. 2 ヒューリスティクスによる改良型音節認識機構

分かち書きの失敗例では、[C+V+C]形においてタイ語の文法規則から規則1-8を導いたにも関わらず、上記の(1)のように一致しない、また、(2)のように規則から外れた形の語である。これらの結果から、次のような規則が発見される。

- (1) 母音v3、v5、v6が出現したとき、これらの母音は音節の最右端に於けるデリミタを表す。
- (2) 母音v7 v8 v10 v11 v13 v15の母音が出現したとき、この母音に引き続いて子音が付加されることがある。このヒューリスティクスを導入した[C+V]または[C+V+C]の改良型モデルを提案する。

この改良型モデルによる実験を行った。この実験結果を表5に示す。

表5 改良型モデルによる実験結果

音節単位に	
分割されなかった文	1, 269文
生成語数	401, 577語
辞書に登録されていない語	
タイタイ辞書	98, 021語(24.4%)
KTSD辞書	112, 095語(27.9%)
分かち書き率	93.9%

この実験結果から、音節形成規則のみのモデルより、44.3%高い、93.9%の分かち書き率を示した。

分かち書きされなかった1, 269文の内訳を図6に示す。

なお、結果の音節を確認するために2種類の辞書を用いた。

ケース	不成功例	文数	ケース	不成功例	文数
(1)	/v20 c1/ c30 t1 v3/ /v20 c1 c30 t1/ v3 ...	362	(2)	/c7 c23 c32 v2 c1 c36/ c44 v2 c21/ /c7 c23 c32 v2 c1/ c36 v15/ v2 ...	343
(3)	/c41 v10 c13 c32 c44 v8/ /c41 v10 c13 c32 v15/ v8 ...	37	(4)	/c35 c43 v2 c11 v8/ /c35 v12/ v2 ...	278
(5)	/c32 v2 c25 c33/ c1 c43 v3/ /c32 v2 c25/ c33 c1/ v43 ...	25	(6)	... c16 c32 c32 c42/ c44 v2 c21/ ...	74
(7)	/c41 v10 v6/ .....	61	(8)	/c7 v8/ c33 v2 c23/ /c7 v8 v9/ v2 ...	39
(9)	/c16 c32 c17 c44 v7 c37 c43 ... .....	4	(10)	/c1 c32 c42 c3 c43 v3/ .....	46

上段：正しい分かち書き、下段：不成功例

1,269

図6 改良型モデルによる分かち書き不成功例

改良型モデルの実験結果から、不成功の1,269文についてセグメンテーションが正しく実行されなかったケースは、概ねつぎのような種類に分けられる。

- (1) いくつかの文字素が、母音、子音の両字母として用いられる為、正しい認識がおこなわれな  
ない。例えば、子音字母c33, c43, c44は、各々v9, v12, v15に等しい。図6のケース(4)の場合である。
- (2) 語の音節構造が[C+V]+[C+V]であるが、文字素構造が[V+C]+[C+V]で  
あり、正しく認識されない。例えば、図6のケース(1)で連続する子音が2重子音となる。
- (3) 語が固有名詞であり、音節が[C+V]または[C+V+C]形をもたない場合、認識され  
ない。例えば、図6のケース(6)のような場合である。

## 5. おわりに

タイ語「三印法典」テキストの20,631文を入力文として、未分割的性格を持つタイ語の自動分かち書きの実験を行った。第1には、日本語形態素解析に由来から用いられてきた最長一致法であり、提案した音節最長一致法によって98.0%の高い分かち書き率を得た。この結果は、他のタイ語テキストの分かち書きにも適用でき、有効であることを示した。第2に、タイ語音節認識機構のモデル化とその実現についてである。形態素解析において辞書の探索回数を減らし、解析効率を上げることは一つの重要な課題であるが、本機構では辞書を全く必要としない特徴を備えつつ、93.9%(文単位)の分かち書き率を得た。分かち書きの結果は語単位ではなく、音節単位であることから語単位への合成には辞書が必要であり、今後の研究課題でもある。なお、タイ語「三印法典」は、すでに京都大学大型計算機センターの共用データベースKTSDとして公開している。

### 【参考文献】

- (1) 田中穂積：自然言語解析の基礎、産業図書、pp.133-139、1989
- (2) 柴山 守：東南アジア学術支援多言語テキスト処理システムの研究、科研報告書、1988
- (3) 石井米雄：三印法典について、東南アジア研究、Vol.6、No.4、pp.155-178、1969
- (4) 柴山 守：データベースKTSDについて、京大大型センター広報、Vol.23、No.2、1990