A Rule-based Syllable Segmentation of Myanmar Text

Zin Maung Maung
Management Information Systems
Engineering Department
Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka, Japan
s065400@ics.nagaokaut.ac.jp

Yoshiki Mikami
Management Information Systems
Engineering Department
Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka, Japan
mikami@kjs.nagaokaut.ac.jp

Abstract

Myanmar script uses no space between words and syllable segmentation represents a significant process in many NLP tasks such as word segmentation, sorting, line breaking and so on. In this study, a rule-based approach of syllable segmentation algorithm for Myanmar text is proposed. Segmentation rules were created based on the syllable structure of Myanmar script and a syllable segmentation algorithm was designed based on the created rules. A segmentation program was developed to evaluate the algorithm. A training corpus containing 32,283 Myanmar syllables was tested in the program and the experimental results show an accuracy rate of 99.96% for segmentation.

1 Introduction

Myanmar language, also known as Burmese, is the official language of the Union of Myanmar. It is spoken by 32 million as a first language, and as a second language by ethnic minorities in Myanmar (Ethnologue, 2005). Burmese is a member of the Tibeto-Burman languages, which is a subfamily of the Sino-Tibetan family of languages. Burmese is a tonal and analytic language using the Burmese script. This is a phonologically based script, adapted from Mon, and ultimately based on an Indian (Brahmi) prototype (Daniels and Bright, 1996). Burmese characters are rounded in shape and the script is written from left to right. No space is used between words but spaces are usually used to separate phrases.

The Myanmar language still remains as one of the less privileged Asian languages in cyberspace. Many people have put considerable effort into the computerization of the Myanmar script. However, Myanmar still lacks support on computers and not many NLP tools and applications are available for this language. A standard encoding is needed for the language processing of Myanmar script; however, there is not yet any official national standard encoding for Myanmar script.

This study focuses on the syllable segmentation of Myanmar text based on the UTN11-2[1] encoding model for Myanmar script. Myanmar script has been granted space in Unicode (U+1000-U+109F) since version 3.0. In Unicode version 4.0, the Unicode consortium defined standards for encoding Myanmar script and canonical order. The current version of Unicode is 5.0. However, there are only a few Unicode-compliant Myanmar fonts that fully follow the Unicode encoding standard. Local font developers and implementers have produced fonts that follow only part of the Unicode standards and many of these partially-compliant fonts are widely used in cyberspace. In 2006, Myanmar proposed additional characters[2] to be added to the Unicode version 5.0. The proposed characters for the Burmese script are as follows:

- 102B MYANMAR VOWEL SIGN TALL AA
- 1039 MYANMAR SIGN VIRAMA [Glyph change and note change]

[1] Unicode Technical Note 11-2, Martin Hosken & Maung Tuntun Lwin, Representing Myanmar in Unicode: Details and Examples, http://www.unicode.org/notes/tn11/

[2] Proposal to Encode Seven Additional Myanmar Characters in the UCS, Myanmar Computer Federation, Myanmar Language Commission

Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 51–58,
Hyderabad, India, January 2008.

- 103A MYANMAR SIGN ASAT
- 103B MYANMAR CONSONANT SIGN MEDIAL YA
- 103C MYANMAR CONSONANT SIGN MEDIAL RA
- 103D MYANMAR CONSONANT SIGN MEDIAL WA
- 103E MYANMAR CONSONANT SIGN MEDIAL HA
- 103F MYANMAR LETTER GREAT SA
- 104E MYANMAR SYMBOL AFOREMENTIONED [Glyph change]

The Unicode technical committee has accepted these proposed characters for inclusion in future versions of the Unicode standard.[3] If the proposal is adopted, this will become the standard encoding for Myanmar script. Therefore, this paper employs the proposed encoding model for the syllable segmentation of Myanmar text.

2 Related Work

The lack of official standard encoding hinders localization of Myanmar language and no previous work on the syllable segmentation of Myanmar script was found. Although character codes for Myanmar languages have been allocated in UCS/Unicode (U+1000–U+109F), lack of implementation makes them unavailable to local end users (Ko Ko and Mikami, 2005). We can learn, however, from related works done for other languages which have similarities to Myanmar. Many attempts have been made in Thai language processing for syllable and word segmentation. Poowarawan (1986) proposed a dictionary-based approach to Thai syllable separation. Thai syllable segmentation was considered as the first step towards word segmentation and many of word segmentation ambiguities were resolved at the level of syllable segmentation (Aroonmanakun, 2002). Thai syllable segmentation can be viewed as the problem of inserting spaces between pairs of characters in the text and the character-level ambiguity of word segmentation can be reduced by extracting syllables whose structures are more well-defined (Sornil and Chaiwanarom, 2004). Most approaches to Thai word segmentation use a dictionary as their basis. However, the segmentation accuracy depends on the quality of the dictionary used for analysis and unknown words can reduce the performance. Theeramunkong and Usanavasin (2001) proposed a non dictionary-based approach to Thai word segmentation. A method based on decision tree models was proposed and their approach claimed to outperform some well-known dictionary-dependent techniques of word segmentation such as the maximum and the longest matching methods.

3 Myanmar Alphabets

In order to clarify the syllable structure, characters of the Myanmar script are classified into twelve categories. Each category is given a name and the glyphs and Unicode code points of characters belonging to each category are shown in Table 1.
The Myanmar script consists of a total of 75 characters. There are 34 consonant letters in Consonants group, four medials in the Medials group and eight vowels in the Dependent Vowels group. Myanmar Sign Virama is used for stacking consonant letters and it does not have a glyph, while Myanmar Sign Asat is used in devowelising process (e.g. ဆင်). There are three dependent various signs in Group F. The Group I consists of three independent vowels (ဣ, ဧ, ဪ) and three independent various signs (၌, ၍, ၏). The characters in Group I can act as stand-alone syllables. Group E consists of four independent vowels (ဤ, ဥ, ဦ, ဩ) and Myanmar Symbol Aforementioned (၎). Each of the independent vowels in group E has its own syllable but they can also combine with other signs to form a syllable (e.g. ဥက္ကာ). Myanmar Symbol Aforementioned in Group E can never stand alone and it is always written as ၎င်း as a short form of လည်းကောင်း. Myanmar Letter Great Sa is always preceded by a consonant and is never written alone (e.g. မနုဿ). There are ten Myanmar digits in the Digits group. The group P consists of two Myanmar punctuation marks. Myanmar script uses white space between phrases, which is taken into account in this study. Non-Myanmar characters are not included in this study.

[3] http://www.unicode.org/alloc/Pipeline.html

Category Name	Name	Glyph	Unicode Code Point
C	Consonants	ကခဂဃငစဆဇဈဉညဋဌဍဎဏတ ထဒဓနပဖဗဘမယရလဝသဟဠအ	U+1000…U+1021
M	Medials	◌ျ ◌ြ ◌ွ ◌ှ	U+103B…U+103E
V	Dependent Vowel Signs	◌ါ ◌ာ ◌ိ ◌ီ ◌ု ◌ူ ◌ေ ◌ဲ	U+102B…U+1032
S	Myanmar Sign Virama	◌္	U+1039
A	Myanmar Sign Asat	◌်	U+103A
F	Dependent Various Signs	◌ံ ◌့ ◌း	U+1036…U+1038

I	Independent Vowels, Independent Various Signs	ဤ ဧ ဪ
		၌ ၍ ၏
			U+1024; U+1027
			U+102A; U+104C;
			U+104D; U+104F;

E	Independent Vowels, Myanmar Symbol Aforementioned	ဣ ဥ ဦ ဩ
		၎
			U+1023; U+1025;
			U+1026; U+1029;
			U+104E;

G	Myanmar Letter Great Sa	ဿ	U+103F
D	Myanmar Digits	၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉	U+1040…U+1049
P	Punctuation Marks	၊ ။	U+104A…U+104B
W	White space		U+0020

Table 1. Classification of Myanmar Script

4 Syllable Structure

A Myanmar syllable consists of one initial consonant, zero or more medials, zero or more vowels and optional dependent various signs. Independent vowels, independent various signs and digits can act as stand-alone syllables. According to the Unicode standard, vowels are stored after the consonant. Therefore, Myanmar vowel sign E (U+1031) is stored after the consonant although it is placed before the consonant in rendering (e.g. နေ). Medials may appear at most three times in a syllable (e.g. မြွှာ). Vowels may appear twice in a syllable (e.g. စော). In a syllable, a second consonant may come together with an Asat for devowelising (e.g. ခင်). Each of the independent vowels in group E has its own syllable but they can also combine with other signs (consonants, dependent vowels, dependent various signs) to form a syllable (e.g. ဣန္ဒြေ, ဥက္ကာ, ဦး, ဩာင်း). The syllable structure of Myanmar script can be written in BNF (Backus-Naur Form) as follows:

$$\text{Syllable} ::= C\{M\}\{V\}\{F\} \mid C\{M\}V^{+}A \mid$$
$$C\{M\}\{V\}CA[F] \mid E[CA][F] \mid I \mid D$$

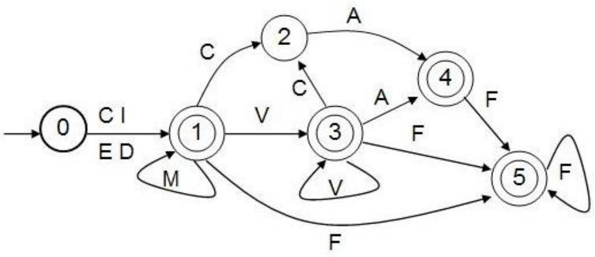


Figure 1. FSA for Syllable Structure

A finite state machine or finite state automaton (FSA) can be employed to demonstrate the syllable structure of Myanmar script. A finite state machine is a model of behavior composed of a finite number of states, transitions between those states, and actions. The starting state is shown by a bold circle and double circles indicate final or accepting states. The above figure shows a finite state automaton that can realize a Myanmar syllable. Examples of Myanmar syllables and their equivalent Unicode code points are shown in Table 2.

Syllable	Example	Unicode Point
C	က	U+1000
CF	ကံ	U+1000 U+1036
CCA	ကင်	U+1000 U+1004 U+103A
CCAF	ကင်း	U+1000 U+1004 U+103A U+1038
CV	ကာ	U+1000 U+102C
CVF	ကား	U+1000 U+102C U+1038
CVVA	ကော်	U+1000 U+1031 U+102C U+103A
CVVCA	ကောင်	U+1000 U+1031 U+102C U+1004 U+103A
CVVCAF	ကောင်း	U+1000 U+1031 U+102C U+1004 U+103A U+1038
CM	ကျ	U+1000 U+103B
CMF	ကျံ	U+1000 U+103B U+1036
CMCA	ကျင်	U+1000 U+103B U+1004 103A
CMCAF	ကျင်း	U+1000 U+103B U+1004 103A U+1038
CMV	ကျာ	U+1000 U+103B U+102C
CMVF	ကျား	U+1000 U+103B U+102C U+1038
CMVVA	ကျော်	U+1000 U+103B U+1031 U+102C U+103A
CMVVCA	ကြောင်	U+1000 U+103C U+1031 U+102C U+1004 U+103A
CMVVCAF	ကျောင်း	U+1000 U+103B U+1031 U+102C U+1004 U+103A U+1038
I	ဪ	U+102A
E	ဣ	U+1023

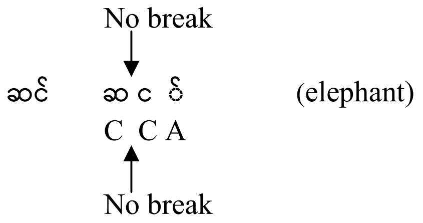
Table 2. Syllable Structure with Examples

5 Syllable Segmentation Rules

Typically, a syllable boundary can be determined by comparing pairs of characters to find whether a break is possible or not between them. However, in some cases it is not sufficient to determine a syllable boundary by just comparing two characters. The following sections explain these cases and give examples.

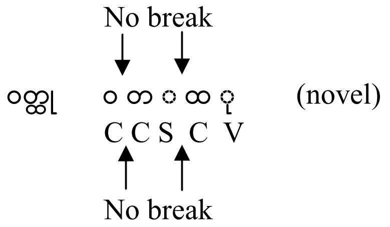
5.1 Devowelising

In one syllable, a consonant may appear twice but the second consonant is used for the devowelising process in conjunction with an Asat (U+103A MYANMAR SIGN ASAT). Therefore the character after the second consonant should be further checked for an Asat. If the character after the second consonant is an Asat, there should be no syllable break before the second consonant.



5.2 Syllable Chaining

Subjoined characters are shown by using an invisible Virama sign (U+1039 MYANMAR SIGN VIRAMA) to indicate that the following character is subjoined and should take a subjoined form. In this case, if the character after the second consonant is an invisible Virama sign, there should be no syllable break before the second and third consonant. Although there are two syllables in a subjoined form, it is not possible to separate them in written form and they are therefore treated as one syllable.



5.3 Kinzi

Kinzi is a special form of devowelised Nga (U+1004 MYANMAR LETTER NGA) with the following letter underneath, i.e., subjoined. In this case, if the character after the second consonant is an Asat and the next character after Asat is an invisible Virama sign (U+1039 MYANMAR SIGN VIRAMA) then there should be no syllable break before the second and third consonant. Kinzi also consists of two syllables but it is treated as one syllable in written form.

No break
↓ ↓
မင်္ဂလာ မ င ◌် ◌္ ဂ - လ ◌ာ (blessing)
C C A S C - C V
↑ ↑
No break

5.4 Loan Words

Usage of loan words can be found in Myanmar text. Although loan words do not follow the Myanmar syllable structure, their usage is common and the segmentation rules for these words are considered in this study.

No break
↓
မားစ်ဂြိုဟ် မ ◌ာ ◌း စ ◌် - ဂ ◌ြ ◌ိ ◌ု ဟ ◌် (Mars)
C V F C A - C M V V C A
↑
No break

5.5 Great Sa

There should be no syllable break before great Sa (U+103F MYANMAR LETTER GREAT SA) as great Sa acts like a stacked သ္သ and devowelises the preceding consonant.

No break
↓
မနုဿ မ - န ◌ု ဿ (human)
C - C V G
↑
No break

5.6 Contractions

There are usages of double-acting consonants in Myanmar text. The double-acting consonant acts as both the final consonant of one syllable and the initial consonant of the following syllable. There are two syllables in a contracted form but they cannot be segmented in written form and there should be no syllable break between them.

No break
↓
ယောက်ျား ယ ေ◌ ◌ာ က ◌် ◌ျ ◌ာ ◌း (man)
C V V C A M V F
↑
No break

6 Implementation

Syllable segmentation rules are presented in the form of letter sequence tables (Tables 4-6). The tables were created by comparing each pair of character categories. However, it is not sufficient to determine all syllable breaks by comparing only two characters. In some cases, a maximum of four consecutive characters need to be considered to determine a possible syllable boundary. Two additional letter sequence tables were created for this purpose (Tables 5 and 6).

Table 4 defines the break status for each pair of two consecutive characters. Table 5 and 6 define the break status for each pair of three and four consecutive characters, respectively. The symbol U in the Table 4 and 5 stands for undefined cases. Cases undefined in Table 4 are defined in the Table 5, and those undefined in Table 5 are then defined in Table 6.

The syllable segmentation program obtains the break status for each pair of characters by comparing the input character sequence with the letter sequence tables. The syllable break status and definitions are shown in Table 3. The break status -1 indicates a breach of canonical spelling order and a question mark is appended after the ambiguous character pair. The status 0 means there should be no syllable break after the first character. For break cases, a syllable breaking symbol (i.e. B in the flowchart) is inserted at each syllable boundary of the input string. The syllable segmentation process is shown in the flowchart in Figure 2.

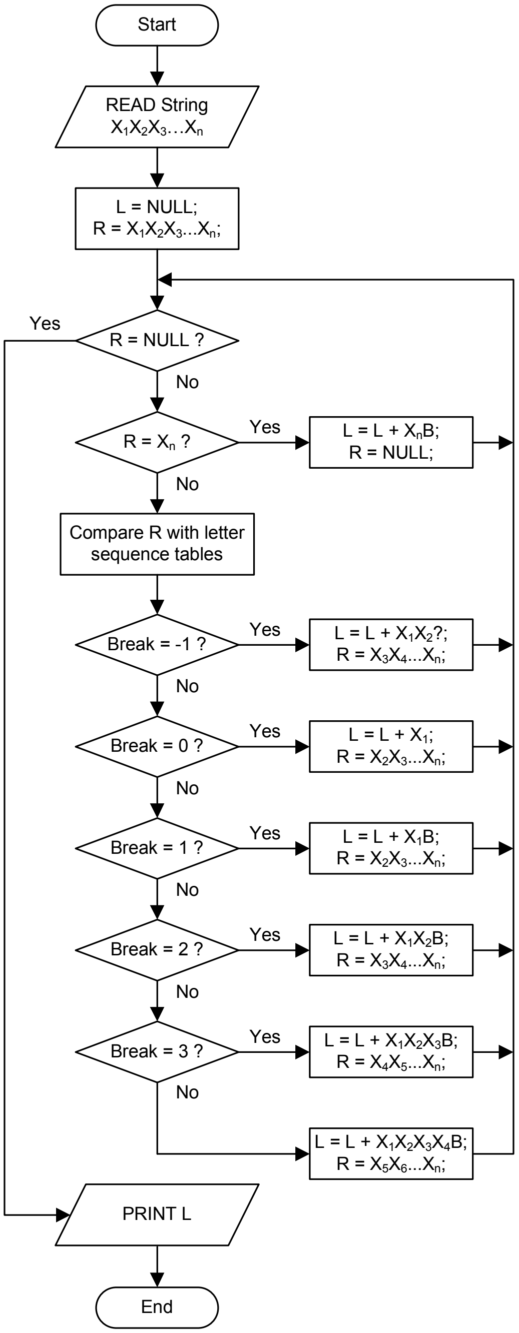


Figure 2. Syllable Segmentation Flowchart

Break Status	Definition
-1	Illegal spelling order
0	No break after 1st character
1	Break after 1st character
2	Break after 2nd character
3	Break after 3rd character
4	Break after 4th character

Table 3. Syllable Break Status and Definition

7 Method and Results

A syllable segmentation program was developed to evaluate the algorithm and segmentation rules. The program accepts the Myanmar text string and shows the output string in a segmented form. The program converts the input text string into equivalent sequence of category form (e.g. CMCACV for မြန်မာ) and compares the converted character sequence with the letter sequence tables to determine syllable boundaries. A syllable segmented Myanmar text string is shown as the output of the program. The symbol "|" is used to represent the syllable breaking point. In order to evaluate the accuracy of the algorithm, a training corpus was developed by extracting 11,732 headwords from Myanmar Orthography (Myanmar Language Commission, 2003). The corpus contains a total of 32,238 Myanmar syllables. These syllables were tested in the program and the segmented results were manually checked. The results showed 12 errors of incorrectly segmented syllables, thus achieving accuracy of 99.96% for segmentation. The few errors occur with the Myanmar Letter Great Sa 'ဿ' and the Independent Vowel 'ဥ'. The errors can be fixed by updating the segmentation rules of these two characters in letter sequence tables. Some examples of input text strings and their segmented results are shown in Table 7.

8 Conclusion

Syllables are building blocks of words and syllable segmentation is essential for the language processing of Myanmar script. In this study, a rule-based approach of syllable segmentation algorithm for Myanmar script is presented. The segmentation rules were created based on the characteristics of Myanmar syllable structure. A segmentation program was developed to evaluate the algorithm. A test corpus containing 32,238 Myanmar syllables was tested in the program and 99.96% accuracy was achieved. From this study, we can conclude that syllable segmentation of Myanmar text can be implemented by a rule-based approach. While characters of non-Myanmar script are not considered in this study, the segmentation rules can be further extended to cover these characters. A complete syllable segmentation algorithm for Myanmar script can be further implemented by applying this algorithm.

	2nd Character												
1st Character		A	C	D	E	F	G	I	M	P	S	V	W
	A	-1	U	1	1	0	-1	1	0	1	0	0	1
	C	0	U	1	1	0	0	1	0	1	0	0	1
	D	-1	1	0	1	-1	-1	1	-1	1	-1	-1	1
	E	-1	U	1	1	2	0	1	-1	1	-1	0	1
	F	-1	U	1	1	2	-1	1	-1	1	-1	-1	1
	G	-1	1	1	1	0	-1	1	-1	1	-1	0	1
	I	-1	1	1	1	-1	-1	1	-1	1	-1	-1	1
	M	2	U	1	1	0	0	1	0	1	-1	0	1
	P	-1	1	1	1	-1	-1	1	-1	1	-1	-1	1
	S	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	V	2	U	1	1	0	0	1	-1	1	-1	0	1
	W	-1	1	1	1	-1	-1	1	-1	1	-1	-1	0

Table 4. Letter Sequence Table 1

	3rd Character												
First 2 Characters		A	C	D	E	F	G	I	M	P	S	V	W
	AC	3	1	1	1	1	1	1	U	1	1	1	1
	CC	0	1	1	1	1	1	1	1	1	0	1	1
	EC	0	1	1	1	1	1	1	1	1	0	1	1
	FC	3	1	1	1	1	1	1	U	1	1	1	1
	MC	0	1	1	1	1	1	1	1	1	0	1	1
	VC	0	1	1	1	1	1	1	U	1	0	1	1

Table 5. Letter Sequence Table 2

	4th Character												
First 3 Characters		A	C	D	E	F	G	I	M	P	S	V	W
	ACM	4	1	1	1	1	1	1	1	1	1	1	1
	FCM	4	1	1	1	1	1	1	1	1	1	1	1
	VCM	4	1	1	1	1	1	1	1	1	1	1	1

Table 6. Letter Sequence Table 3

Myanmar Text	Letter Sequence	Segmented Letter Sequence	Segmented Result
အဗ္ဘန္တရသရက်	CCSCCSCCCCCA	|CCSCCSC|C|C|CCA|	|အဗ္ဘန္တ|ရ|သ|ရက်|
ဥတ္တရယဉ်စွန်းတန်း	ECSCCCCACMCAFCCAF	|ECSC|C|CCA|CMCAF|CCAF|	|ဥတ္တ|ရ|ယဉ်|စွန်း|တန်း|
ဣဋ္ဌာသယ	ECSCVCC	|ECSCV|C|C|	|ဣဋ္ဌာ|သ|ယ|
ဧကရာဇ်	ICCVCA	|I|C|CVCA|	|ဧ|က|ရာဇ်|
ဝင်္ကန္တဉာဏ်	CCASCCSCCVCA	|CCASCCSC|CVCA|	|ဝင်္ကန္တ|ဉာဏ်|
မားစ်ဂြိုဟ်	CVFCACMVVCA	|CVFCA|CMVVCA|	|မားစ်|ဂြိုဟ်|
မနုဿီဟ	CCVGVC	|C|CVGV|C|	|မ|နုဿီ|ဟ|
တာဝတိံသာ	CVCCVFCV	|CV|C|CVF|CV|	|တာ|ဝ|တိံ|သာ|
ကျွန်ုပ်၏ကား	CMMCAVCAICAF	|CMMCAVCA|I|CAF|	|ကျွန်ုပ်|၏|ကား|
ကက်ရှ်မီးယား	CCACMACVFCVF	|CCACMA|CVF|CVF|	|ကက်ရှ်|မီး|ယား|
လ္ဘက်ရည်ဆိုင်	CSCCACCACVVCA	|CSCCA|CCA|CVVCA|	|လ္ဘက်|ရည်|ဆိုင်|

Table 7. Syllable Segmentation Examples and Results

Acknowledgement

The study was made possible by the sponsorship of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). The authors wish to express special thanks to Myanmar Unicode and NLP Research Center and its members for their help during this research.

References

Ethnologue. 2005. *Languages of the World, Fifteenth edition.* Online version: http://www.ethnologue.com/, Edited by Raymond G. Gordon, Jr. Dallas, Tex.: SIL International.

Martin Hosken and Maung Tuntunlwin. 2007. *Representing Myanmar in Unicode: Details and Examples.* http://www.unicode.org/notes/tn11/

Myanmar Computer Federation, Myanmar Language Commission. 2006. *Proposal to Encode Seven Additional Myanmar Characters in the UCS.* http://www.myanmarnlp.net.mm/doc/updateOnDec/20060228_ProposaltoEncodeSevenAdditionalMyanmarCharsinUCS.pdf

Myanmar Language Commission. 2003. *Myanmar Orthography, 2nd Edition.* University Press, Yangon, Myanmar.

Ohm Sornil and Paweena Chaiwanarom. 2004. *Combining Prediction by Partial Matching and Logistic Regression for Thai Word Segmentation.* Proceedings of the 20th International Conference on Computational Linguistics.

Peter T. Daniels and William Bright. 1996. *The World's Writing Systems.* Oxford University Press.

Thanaruk Theeramunkong and Sasiporn Usanavasin. 2001. *Non-Dictionary-Based Thai Word Segmentation Using Decision Trees.* Proceedings of the First International Conference on Human Language Technology Research.

The Unicode Consortium. 2003. *The Unicode Standard Version 4.0.* Addison-Wesley.

The Unicode Consortium. 2006. *The Unicode Standard Version 5.0.* Addison-Wesley.

Wirote Aroonmanakun. 2002. *Collocation and Thai Word Segmentation.* Proceedings of SNLP-Oriental COCOSDA.

Wunna Ko Ko and Yoshiki Mikami. 2005. *Languages of Myanmar in Cyberspace.* Nagaoka University of Technology, Bulletin on Language Science and Humanity, Vol. 19.pp.249-264.

Yuen Poowarawan. 1986. *Dictionary-based Thai Syllable Separation.* Proceedings of the Ninth Electronics Engineering Conference.