Manually Constructed Context-Free Grammar For Myanmar Syllable Structure

Tin Htay Hlaing
Nagaoka University of Technology
Nagaoka, JAPAN
tinhtayhlaing@gmail.com

Abstract

Myanmar language and script are unique and complex. Up to our knowledge, considerable amount of work has not yet been done in describing Myanmar script using formal language theory. This paper presents manually constructed context free grammar (CFG) with "111" productions to describe the Myanmar Syllable Structure. We make our CFG in conformity with the properties of LL(1) grammar so that we can apply conventional parsing technique called predictive top-down parsing to identify Myanmar syllables. We present Myanmar syllable structure according to orthographic rules. We also discuss the preprocessing step called contraction for vowels and consonant conjuncts. We make LL (1) grammar in which "1" does not mean exactly one character of lookahead for parsing because of the above mentioned contracted forms. We use five basic sub syllabic elements to construct CFG and found that all possible syllable combinations in Myanmar Orthography can be parsed correctly using the proposed grammar.

1 Introduction

Formal Language Theory is a common way to represent grammatical structures of natural languages and programming languages. The origin of grammar hierarchy is the pioneering work of Noam Chomsky (Noam Chomsky, 1957). A huge amount of work has been done in Natural Language Processing where Chomsky`s grammar is used to describe the grammatical rules of natural languages. However, formulation rules have not been established for grammar for Myanmar script. The long term goal of this study is to develop automatic syllabification of Myanmar polysyllabic words using regular grammar and/or finite state methods so that syllabified strings can be used for Myanmar sorting.

In this paper, as a preliminary stage, we describe the structure of a Myanmar syllable in context-free grammar and parse the syllables using predictive top-down parsing technique to determine whether a given syllable can be recognized by the proposed grammar or not. Further, the constructed grammar includes linguistic information and follows the traditional writing system of Myanmar script.

2 Myanmar Script

Myanmar is a syllabic script and also one of the languages which have complex orthographic structures. Myanmar words are formed by collection of syllables and each syllable may contain up to seven different sub syllabic elements. Again, each component group has its own members having specific order.

Basically, Myanmar script has 33 consonants, 8 vowels (free standing and attached)[1] , 2 diacritics, 11 medials, a vowel killer or ASAT, 10 digits and 2 punctuation marks.

A Myanmar syllable consists of 7 different components in Backus Normal Form (BNF) is as follows.

S:= C{M}{V}[CK][D] | I[CK] | N

where

S = Syllable

1. C = Consonant
2. M = Medial or Consonant Conjunct or attached consonant

[1] Free standing vowel syllables (eg. ဩ)and attached vowel symbols (eg. ◌ိ)

3. V = Attached Vowel
4. K = Vowel Killer or ASAT
5. D = Diacritic
6. I = Free standing Vowel
7. N = Digit

And the notation [ ] means 0 or 1 occurrence and { } means 0 or more occurrence.
However, in this paper, we ignore digits, free standing vowel and punctuation marks in writing grammar for Myanmar syllable and we focus only on basic and major five sub syllabic groups namely consonants(C), medial(M), attached vowels(V), a vowel killer (K) and diacritics(D). The following subsection will give the details of each sub syllabic group.

2.1 Brief Description of Basic Myanmar Sub Syllabic Elements

Each Myanmar consonant has default vowel sound and itself works as a syllable. The set of consonants in Unicode chart is C={က, ခ, ဂ, ဃ, င, စ, ဆ, ဇ, ဈ, ဉ, ည ,ဋ ,ဌ, ဍ, ဎ, ဏ, တ, ထ ,ဒ ,ဓ ,န ,ပ ,ဖ, ဗ, ဘ ,မ ,ယ ,ရ, လ, ဝ, သ, ဟ, ဠ } having 33 elements. But, the letter အ can act as consonant as well as free standing vowel.

Medials or consonant conjuncts mean the modifiers of the syllables` vowel and they are encoded separately in the Unicode encoding. There are four basic medials in Unicode chart and it is represented as the set M={ျ, ြ, ွ, ှ}.

The set V of Myanmar attached vowel characters in Unicode contains 8 elements { ါ, ာ, ိ, ီ, ု, ူ, ေ, ဲ}. (Peter and William, 1996)

Diacritics alter the vowel sounds of accompanying consonants and they are used to indicate tone level. There are 2 diacritical marks { ့, း } in Myanmar script and the set is represented as D.

The asat, or killer, representing the set K= { ် } is a visibly displayed sign. In some cases it indicates that the inherent vowel sound of a consonant letter is suppressed. In other cases it combines with other characters to form a vowel letter. Regardless of its function, this visible sign is always represented by the character U+103A .[2] [John Okell, 1994]

In Unicode chart, the diacritics group D and the vowel killer or ASAT "K" are included in the group named various signs.

2.2 Preprocessing of Texts - Contraction

In writing formal grammar for a Myanmar syllable, there are some cases where two or more Myanmar characters combine each other and the resulting combined forms are also used in Myanmar traditional writing system though they are not coded directly in the Myanmar Unicode chart. Such combinations of vowel and medials are described in detail below.
Two or more Myanmar attached vowels are combined and formed new three members { ော, ောက်, ိုဲ } in the vowel set.

Glyph	Unicode for Contraction	Description
ေ + ာ	1031+102C	Vowel sign E + AA
ေ + ာ + ်	1031+102C+103A	Vowel sign E +AA+ASAT
ိ + ု	102D + 102F	Vowel sign I + UU

"Table 1. Contractions of vowels"

Similarly, 4 basic Myanmar medials combine each other in some different ways and produce new set of medials { ျွ, ြွ, ျှ, ြှ, ွှ, ျွှ , ြွှ }. [Tin Htay Hlaing and Yoshiki Mikami, 2011]

Glyph	Unicode for Contraction	Description
ျ + ွ	103B + 103D	Consonant Sign Medial YA + WA
ြ + ွ	103C + 103D	Consonant Sign Medial RA + WA
ျ + ှ	103B + 103E	Consonant Sign Medial YA + HA
ြ + ှ	103C + 103E	Consonant Sign Medial RA + HA

[2] http://www.unicode.org/versions/Unicode6.0.0/ch11.pdf

◌ွ + ◌့	103D + 103E	Consonant Sign Medial WA + HA
◌ျ + ◌ွ + ◌့	103B + 103D + 103E	Consonant Sign Medial YA+WA + HA
◌ြ + ◌ွ+ ◌့	103C + 103D + 103E	Consonant Sign Medial YA+WA + HA

"Table 2. Contractions of Medials"

The above mentioned combinations of characters are considered as one vowel or medial in constructing the grammar. The complete sets of elements for vowels and meidals used in writing grammar are depicted in the table below.[3]

Name of Sub Syllabic Component	Elements
Medials or Conjunct Consonants	◌ျ, ◌ြ, ◌ွ, ◌့, ◌ျွ, ◌ြွ, ◌ျ့,◌ြ့, ◌ွ့, ◌ျ ◌ွ ◌့ ,◌ြ ◌ွ ◌့
Attached vowels	◌ာ, ◌ါ, ◌ိ, ◌ီ, ◌ု, ◌ူ, ◌ေ , ◌ဲ, ◌ော, ◌ော်, ◌ံ, ◌ိုး

"Table 3. List of vowels and Medials"

2.3 Combinations of Syllabic Components within a Syllable

As mentioned in the earlier sections, we choose only 5 basic sub syllabic components namely consonants (C), medial (M), attached vowels (V), vowel killer (K) and diacritics (D) to describe Myanmar syllable. As our intended use for syllabification is for sorting, we omit stand-alone vowels and digits in describing Myanmar syllable structure. Further, according to the sorting order of Myanmar Orthography, stand-alone vowels are sorted as the syllable using the above 5 sub syllabic elements having the same pronunciation. For example, stand-alone vowel "ဣ" is sorted as consonant "အ" and attached vowel "◌ိ" combination as "အိ".

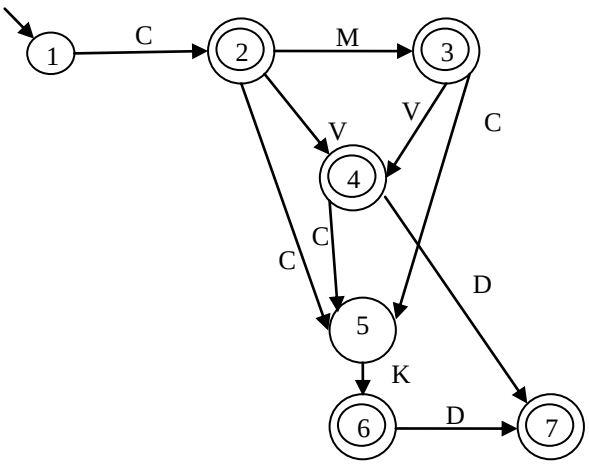
[3] Sorting order of Medials and attached vowels in Myanmar Orthography

In Myanmar language, *a syllable with only one consonant can be taken as one syllable* because Myanmar script is Abugida which means all letters have inherent vowel. And, consonants can be followed by vowels, consonant, vowel killer and medials in different combinations.
One special feature is that if there are two consonants in a given syllable, the second consonant must be followed by vowel killer (K). We found that 1872 combinations of sub-syllabic elements in Myanmar Orthography [Myanmar Language Commission, 2006]. The table below shows top level combinations of these sub-syllabic elements.

Conso-nant only	**Consona-nt followed by Vowel**	**Consona-nt followed by Consona-nt**	**Consonant followed by Medial**
C	CV CVCK CVD CVCKD	CCK CCKD	CM CMV CMVD CMVCK CMVCKD CMCK CMCKD

"Table 4. Possible Combinations within a Syllable"

The combinations among five basic sub syllabic components can also be described using Finite State Automaton. We also find that Myanmar orthographic syllable structure can be described in regular grammar.



"Figure 1. FSA for a Myanmar Syllable"

In the above FSA, an interesting point is that only one consonant can be a syllable because Myanmar consonants have default vowel sounds. That is why, state 2 can be a final state. For instance, a Myanmar Word "မိန်းမ" (means "Woman" in English) has two syllables. In the first syllable "မိန်း", the sub syllabic elements are Consonant(မ) + Vowel(ိ) +Consonant(န)+ Vowel Killer(်)+Diacritics(း). The second syllable has only one consonant "မ".

3 Myanmar Syllable Structure in Context-Free Grammar

3.1 Manually Constructed Context-Free Grammar for Myanmar Syllable Structure

Context free (CF) grammar refers to the grammar rules of languages which are formulated independently of any context. A CF-grammar is defined by:

1. A finite terminal vocabulary V_T.
2. A finite auxiliary vocabulary V_A.
3. An axiom $S \in V_A$.
4. A finite number of context-free rules P of the form $A \rightarrow \phi$ where

$A \in V_A$ and $\phi \in \{V_A \cup V_T\}^*$ (M.Gross and A.Lentin, 1970)

The grammar G to represent all possible structures of a Myanmar syllable can be written as $G= (V_T, V_A, P, S)$ where the elements of P are:

S→ ◌ X

Such production will be expanded for 33 consonants.

X→ ◌ A

Such production will be expanded for 11 medials.

X→ ◌ B

Such production will be expanded for 12 vowels.

X→C ◌ D

X→ε

A→ ◌ B

Such production will be expanded for 12 vowels.

A→ C ◌ D

A→ε

B → C ◌ D

B→ D

B→ε

D→ ◌း # Diacritics

D→ ◌ံ # Diacritics

D→ ε

C→ ◌

Such production will be expanded for 33 consonants.

Total number of productions/rules to recognize Myanmar syllable structure is "111" and we found that the director symbol sets (which is also known as first and follow sets) for same non-terminal symbols with different productions are disjoint.

This is the property of LL(1) grammar which means for each non terminal that appears on the left side of more than one production, the directory symbol sets of all the productions in which it appears on the left side are disjoint. Therefore, our proposed grammar can be said as LL(1) grammar.

The term LL1 is made up as follows. The first L means reading from Left to right, the second L means using Leftmost derivations, and the "1" means with one symbol of lookahead. (Robin Hunter, 1999)

3.2 Parse Table for Myanmar CFG

The following figure is a part of parse table made from the productions of the proposed LL(1) grammar.

	◌	◌	◌	◌	◌း	◌ံ	◌်	$
S	S→ ◌X	S→ ◌ X						
X	X→ C ◌် D	X →C ◌် D	X → ◌ A	X → ◌ B				X →ε
A	A→ C ◌် D	A →C ◌် D		A → ◌ B				A →ε
B	B→ C ◌် D	B→ C ◌် D			B→ D	B→ D		B →ε
D					D → ◌း	D →◌ံ		D →ε
C	C→ ◌	C→ ◌						

"Table 5. Parse Table for Myanmar Syllable"

In the above table, the topmost row represents terminal symbols whereas the leftmost column represents the non terminal symbols. The entries in the table are productions to apply for each pair of non terminal and terminal.

An example of Myanmar syllable having 4 different sub syllabic elements is parsed using proposed grammar and the above parse table. The parsing steps show proper working of the proposed grammar and the detail of parsing a syllable is as follows.

Input Syllable = ကျား =က(C) + ျ(M)+ ာ (V)+း (D)

Parse Stack	Remaining Input	Parser Action
S $	က ျ ာ း $	S→ကX
ကX $	က ျ ာ း $	MATCH
က္က X $	က္က ျ ာ း $	X→ ျA
က္က ျA $	က္က ျ ာ း $	MATCH
က္က ျA $	က္က ျ ာ း $	A→ ာB
က္က ျ ာ B $	က္က ျ ာ း $	MATCH
က္က ျ ာB $	က္က ျ ာ း $	B→D
က္က ျ ာ D $	က္က ျ ာ း $	D→ း
က္က ျ ာ း $	က္က ျ ာ း $	MATCH
က္က ျ ာ း $	း $	SUCCESS

"Table 6. Parsing a Myanmar Syllable using predictive top-down parsing method"

4 Conclusion

This study shows the powerfulness of Chomsky`s context free grammar as it can apply not only to describe the sentence structure but also the syllable structure of an Asian script, Myanmar. Though the number of productions in the proposed grammar for Myanmar syllable is large, the syntactic structure of a Myanmar syllable is correctly recognized and the grammar is not ambiguous.

Further, in parsing Myanmar syllable, it is necessary to do preprocessing called contraction for input sequences of vowels and consonant conjuncts or medials to meet the requirements of traditional writing systems. However, because of these contracted forms, single lookahead symbol in our proposed LL(1) grammar does not refer exactly to one character and it may be a combination of two or more characters in parsing Myanmar syllable.

5 Discussion and Future Work

Myanmar script is syllabic as well as agglutinative script. Every Myanmar word or sentence is composed of series of individual syllables. Thus, it is critical to have efficient way of recognizing syllables in conformity with the rules of Myanmar traditional writing system.

Our intended research is the automatic syllabification of Myanmar polysyllabic words using formal language theory.

One option to do is to modify our current CFG to recognize consecutive syllables as a first step. We found that if the current CFG is changed for sequence of syllables, the grammar can be no longer LL(1). Then, we need to use one of the statistical methods, for example, probabilistic CFG, to choose correct productions or best parse for finding syllable boundaries.

Again, it is necessary to calculate the probability values for each production based on the frequency of occurrence of a syllable in a dictionary we referred or using TreeBank.

We need Myanmar corpus or a tree bank which contains evidence for rule expansions for syllable structure and such a resource does not yet exist for Myanmar. And also, the time and cost for constructing a corpus by ourselves came into consideration.

Another approach is to construct finite state transducer for automatic syllabification of Myanmar words. If we choose this approach, we firstly need to construct regular grammar to recognize Myanmar syllables. We already have Myanmar syllable structure in regular grammar. However, for finite state syllabification using weights, there is a lack of resource for training database.

We still have many language specific issues to be addressed for implementing Myanmar script using CFG or FSA. As a first issue, our current grammar is based on five basic sub-syllabic elements and thus developing the grammar which can handle all seven Myanmar sub syllabic elements will be future study.

Our current grammar is based on the code point values of the input syllables or words. Then, as a second issue, we need to consider about different presentations or code point values of same character. Moreover, we have special writing traditions for some characters, for example, such

as consonant stacking eg. ဗုဒ္ဓ (Buddha), မန္တလေး (Mandalay, second capital of Myanmar), consonant repetition eg. တက္ကသိုလ် (University), kinzi eg. အင်္ဂတေ (Cement), loan words eg. ဘတ်(စ်) (bus). To represent such complex forms in a computer system, we use invisible Virama sign (U+1039). Therefore, it is necessary to construct the productions which have conformity with the stored character code sequence of Myanmar Language.

References

John Okell. " Burmese An Introduction to the Script". Northern Illinois University Press, 1994.

M.Gross, A.Lentin. "Introduction to Formal Grammar". Springer-Verlag, 1970.

Myanmar Language Commission. *Myanmar Orthography*, Third Edition, University Press, Yangon, Myanmar, 2006.

Noam Chomsky. "Syntactic Structures". Mouton De Gruyter, Berlin, 1957.

Peter T. Denials, William Bright. "World`s Writing System". Oxford University Press, 1996.

Robin Hunter. "The Essence of Compilers". Prentice Hall, 1999.

Tin Htay Hlaing, Yoshiki Mikami. " Collation Weight Design for Myanmar Unicode Texts" in Proceedings of Human Language Technology for Development organized by PAN Localization- Asia, AnLoc – Africa, IDRC – Canada. May 2011, Alexandria, EGYPT, Page 1- 6.