Burmese Phrase Segmentation

May Thu Win maythuwin85@gmail.com
Moet Moet Win moetmoetwin.ucsy@gmail.com
Moh Moh Than mohmohthanster@gmail.com
Research Programmer, Myanmar Natural Language Processing Lab

Dr.Myint Myint Than Myanmar Computer Federation
Dr.Khin Aye Member of Myanmar Language Commission

Abstract

Phrase segmentation is the process of determination of phrase boundaries in a piece of text. When it comes to machine translation, phrase segmentation should be computerized. This is the first attempt at automatic phrase segmentation in Burmese (Myanmar). This paper aims to express how to segment phrases in a Burmese sentence and how to formulate rules. The system has been tested by developing a phrase segmentation system using CRF++.

1 Introduction

Burmese Language is the national and official language of Myanmar, and is a member of the Tibeto-Burman language family, which is a subfamily of the Sino-Tibetan family of languages. Its written form uses a script that consists of circular and semi-circular letters, adapted from the Mon script, which in turn was developed from a southern Indian script in the 8th century.

Burmese language users normally use space as they see fit, some write with no space at all. There is no fixed rule for phrase segmentation. In this paper, we propose phrase segmentation rules, in linguistics point of view, which will help Natural Language Processing tasks such as Machine Translation, Text Summarization, Text Categorization, Information Extraction and Information Retrieval and so on.

2 Nature of Burmese Language

There are two types of language style - one is literary or written style used in formal, literary works, official publications, radio broadcasts and formal speeches and the other is colloquial or spoken style used in daily communication, both conversation and writing, in literary works, radio and TV broadcasts, weekly and monthly magazines. Literary Burmese is not so much different from colloquial Burmese. Grammar pattern is the same in both, and so is the essential vocabulary. Some particles are used unchanged in both but a few others are found in one style only. Regional variation is seen in both styles.

2.1 Sentence Construction

One morpheme or a combination of two or more morphemes will give rise to one word; combination of two or more words becomes a phrase; combination of two or more phrases will be one sentence. The following figure shows the hierarchical structure of sentence construction.

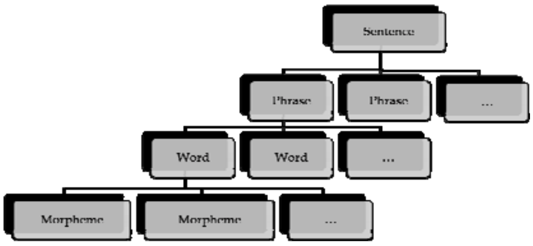


Figure 1: Hierarchical structure of sentence construction

Sentence	သူက ပန်းလေးကို နမ်းတယ် ။						
Phrase	သူက		ပန်းလေးကို			နမ်းတယ်	
Word	သူ	က	ပန်းလေး		ကို	နမ်း	တယ်
Morpheme	သူ	က	ပန်း	လေး	ကို	နမ်း	တယ်

Table 1: Sentence construction of a Burmese sentence

In this table, သူက ပန်းလေးကို နမ်းတယ် means "she kisses the little flower". And သူ is she, က is subject marker, ပန်း is the flower, လေး is little, ကို is object marker, နမ်း is kisses and တယ် is verb marker.

Morpheme: Morpheme is the smallest syntactic unit that has semantic meaning. The sentence shown in Table.1 has seven morphemes.

Word: The word is the basic item in a sentence.

It consists of one or more morphemes that are linked by close juncture. A word consisting of two or more stems joined together is a compound word. Words that carry meaning are lexical words and words that only show grammatical relation are grammatical words. The sentence shown in Table .1 has six words.

Phrase: Two or more words come together to form a phrase. A phrase is a syntactic unit that forms part of a complete sentence and has its particular place in that sentence. Phrase boundary is characterized by what is called a phrase marker or a pause in speech – open juncture. The phrase marker can be omitted, in which case we say there is a zero marker. Markers show different functions, like subject, verb, object, complement, qualifier, time and place adverb, etc. The sentence shown in Table.1 has three phrases.

Sentence: Finally, we want to say something about the sentence. A sentence is organized with one or more phrases in Subject Object [Complement] Verb or Object Subject Verb order. It is a sequence of phrases capable of standing alone to make an assertion, a question, or a command.

2.2 Syntax

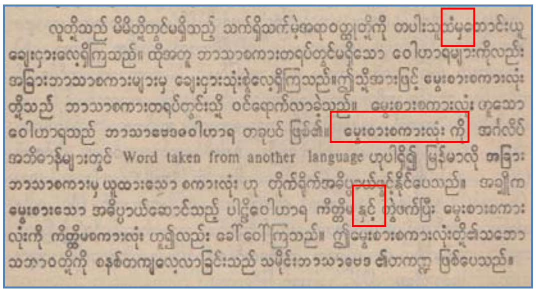
Syntax is the study of the rules and principles found in the construction of sentences in Burmese language. A Burmese sentence is composed of NP+...+NP+VP (where, NP = noun phrase and VP = verb phrase). Noun phrases and verb phrases are marked off by markers but some can be omitted.

3 Parts of Speech

Myanmar Language Commission opines that Burmese has nouns, pronouns, adjectives, verbs, adverbs, postpositions, particles, conjunctions and interjections. In fact, the four really important parts are Nouns, Verbs, Qualifiers or Modifiers and Particles. Pronouns are just nouns. Qualifiers are the equivalents of adjectives and adverbs that are obtained by subordinated use of nouns and verbs. Postpositions and affixes can be considered as markers or particles. Interjections do not count in the parts of speech in Burmese.

4 Phrase Segmentation by Writer's Whim

In Figure.2, sentences are broken into phrases with space in a random way. Phrase segmentation is employed at the writer's whim; it is not guided by rules.



လူတို့သည် မိမိတို့တွင်မရှိသည့် သက်ရှိသက်မဲ့အရာဝတ္ထုတို့ကို တပါးသူထံမှတောင်းယူ
ချေးငှားလေ့ရှိကြသည်။ ထိုအတူ ဘာသာစကားတရပ်တွင်မရှိသော ဝေါဟာရများကိုလည်း
အခြားဘာသာစကားများမှ ချေးငှားသုံးစွဲလေ့ရှိကြသည်။ ဤသို့အားဖြင့် မွေးစားစကားလုံး
တို့သည် ဘာသာစကားတရပ်တွင်းသို့ ဝင်ရောက်လာခဲ့သည်။ မွေးစားစကားလုံး ဟူသော
ဝေါဟာရသည် ဘာသာဗေဒဝေါဟာရ တခုပင် ဖြစ်၏။ မွေးစားစကားလုံး ကို အင်္ဂလိပ်
အဘိဓာန်များတွင် Word taken from another language ဟုပါရှိ၍ မြန်မာလို အခြား
ဘာသာစကားမှ ယူထားသော စကားလုံး ဟု တိုက်ရိုက်အဓိပ္ပာယ်ဖွင့်နိုင်ပေသည်။ အချို့က
မွေးစားသော အဓိပ္ပာယ်ဆောင်သည့် ပါဠိဝေါဟာရ ကိတ္တိမ နှင့် တွဲဖက်ပြီး မွေးစားစကား
လုံးကို ကိတ္တိမစကားလုံး ဟူ၍လည်း ခေါ်ဝေါ်ကြသည်။ ဤမွေးစားစကားလုံးတို့၏သဘော
သဘာဝတို့ကို စနစ်တကျလေ့လာခြင်းသည် သမိုင်းဘာသာဗေဒ ၏ကဏ္ဍ ဖြစ်ပေသည်။

Figure 2: Phrase segmentation by writer's whim

Segmentation may be suggested by pause, or length variety, or clustering of words to bring about meaning. Segmentation in a casual careless way will not be of any help. This paper tries to point out the places where we break sentences with consistency. The boxes shown in the figure are places to break. We will explain how and why we should break at these places in section 6.

5 Particles

We do not normally use lexical words (nouns, verbs and qualifiers) by themselves and they have to be joined by grammatical words (particles or markers) to form a phrase in both literary and colloquial styles. There are three types of particles - formatives, markers and phrase particles.

5.1 Formatives

Formation derives a new word by attaching particles to root morphemes or stems. It may also change the grammatical class of a word by adding affix (prefix or suffix). Adding "စရာ" to the verb "စား eat" gives rise to "စားစရာ food", a noun, and there are many derivational morphemes that change verbs to nouns, verbs to adverbs, and so on.

Reduplication is a word formation process in which some part of a base (a morpheme or two) is repeated, and is found in a wide range of Burmese words. Example; လိုက်လှဲ warm → လိုက်လိုက်လှဲလှဲ warmly.

It can be obviously seen that formation is a way of word structure. It can be useful sometimes in phrase segmentation, as it can easily be marked off as an independent phrase.

5.2 Markers

A marker or a particle is a grammatical word that indicates the grammatical function of the marked word, phrase, or sentence.

When we break up a sentence we first break it into noun phrases and verb phrases. Verb phrases must be followed by verb markers (sentence-ending or subordinating markers). Noun phrase will be followed by various noun markers, also called postpositions, denoting its syntactic role in the sentence. If we want to show a noun is the subject, a marker that indicates the subject function will be strung with this noun. If we want to indicate a noun to be the object, a marker that indicates the object function will be strung. The distinctive feature of markers is that they show the role of a phrase in the sentence.

noun phrase	noun phrase	noun phrase	verb phrase
ကျောင်းသားများသည်	ပုဂံသို့	လေ့လာရေးခရီး	သွားသည်
The students	to Bagan	an excursion	make

"The students make an excursion to Bagan."

Table 2: A Burmese sentence with markers

In this Table.2, we find three markers,
- သည် marking the subject of the sentence
- သို့ marking the place of destination in the sentence
- သည် marking the verb tense of the sentence

Sometimes, we can construct a noun or verb phrase without adding any visible markers to them. In this case, we say we are using zero markers, symbolized by ø after the noun.
- လေ့လာရေးခရီး suffixing ø marker

We, therefore, use markers as pointers to individual phrases in phrase segmentation of Burmese texts.

5.3 Phrase Particles

Phrase particles are suffixes that can be attached to a phrase in a sentence without having any effect on its role in the sentence. They serve only to add emphasis to particular phrases or to the whole sentence or to indicate the relation of one sentence to another. They are attached to nouns or verbs or even to phrases that already contain markers. Phrase particles are of two types: sentence medial and sentence final.

Example: မင်းကတော့ လာမှာပေါ့နော်။
You will come, right?

In this example; မင်း is you (subject), က - subject marker, တော့ - "as for" (sentence medial phrase particle), လာ - come (verb), မှာ - future (verb marker), ပေါ့ - "of course" (sentence final phrase particle) and နော် - right? (sentence final phrase particle).

6 Markers

Some suffixes mark the role of a phrase in the sentence. Suffixes that perform this function are called "markers". So, markers can be seen as phrase boundaries. Markers can be split into two groups: (1) noun markers and (2) verb markers.

6.1 Noun Markers

Markers that are attached to nouns are called "noun markers". A noun marker shows its function as subject, object or complement, instrument, accompaniment, destination, departure, and others in the sentence. We can sometimes construct a phrase without noun markers. Such a phrase is said to be fixed with zero markers symbolized by ø. Its meaning is the same as that of a phrase with markers. A phrase will be segmented where we consider there is a zero marker in the sentence.

6.1.1 Subject Markers

Marker that marks a noun as a subject of sentence can be defined as subject marker.

Literary Style	Colloquial Style	Translation
သည်၊ က၊ မှာ၊ø	က၊ ဟာ၊ ø	no English equivalent

Table 3: Subject markers and their meaning

Example: (with subject marker)
| ကျွန်တော်က | သေချာမှ လုပ်တယ်။
| I | like to be sure before I act.
Example: (with zero markers)
| ကျွန်တော် ø | သေချာမှ လုပ်တယ်။
| I | like to be sure before I act.

6.1.2 Object Markers

Markers that specify the object of the sentence can be defined as object markers.

Literary Style	Colloquial Style	Translation
ကို၊ အား၊ ø	ကို၊ ø	no English equivalent

Table 4: Subject markers and their meaning

Example: သူ |မိန်းကလေးတစ်ဦးကို | ချစ်ဖူးသည်။
He had once fallen in love with | a girl |.

6.1.3 Place Markers

Markers that specify the place and directions can be defined as place markers.

Place Markers	Literary Style	Colloquial Style	Translation
Location	၌၊မှာ၊တွင်၊ ဝယ်၊က	မှာ၊ က	at, on, in
Departure	မှ၊က	က	from
Destination	သို့၊ကို၊ဆီ၊ ø...	ကို၊ဆီ၊ဆီကို ၊ ø	to
Continuation of place	တိုင်တိုင်၊ အထိ ၊ ø ၊ ...	ထိ၊ အထိ၊ ø	until, till

Table 5: Place markers and their meaning

Example: (Departure)
|နေပြည်တော်မှ| ထွက်လာသည်။
I left | from NayPyiDaw | .

6.1.4 Time Marker

Markers that specify the time can be defined as time markers.

Time Markers	Literary Style	Colloquial Style	Translation
Time	မှာ၊တွင်၊ ø..	မှာ၊ က၊ ø	at, on, in
Continuation of time	တိုင်တိုင်၊ အထိ၊ ø၊...	ထိ၊အထိ၊ ထက်တိုင်၊ ø ...	up to, till

Table 6: Time markers and their meaning

Example: (Continuation of time)
| ယခုထက်တိုင် | လွမ်းနေဆဲပါ ခိုင်။
I miss you |up to the present, | Khaing.

6.1.5 Instrumentality Markers

Markers that specify how the action takes place or indicate the manner, or the background condition of the action can be defined as instrumentality markers.

Literary Style	Colloquial Style	Translation
ဖြင့်၊နှင့်	နဲ့	by, with

Table 7: Instrumentality markers and their meaning

Example:| ကားဖြင့် | သွားသည်။
They went | by bus | .

6.1.6 Cause Markers

Markers that specify the reason or cause can be defined as cause markers.

Literary Style	Colloquial Style	Translation
ကြောင့်၊ သဖြင့်၊ ...	ကြောင့်၊နဲ့	because, because of

Table 8: Cause markers and their meaning

Example:| ဝမ်းရောဂါကြောင့်| သေသည်။
He died | of cholera | .

6.1.7 Possessive Markers

Markers that show a possessive phrase or a modifier phrase can be called possessive markers.

Literary Style	Colloquial Style	Translation
၏၊ ရဲ့၊ ့ (tone mark)	ရဲ့၊ ့ (tone mark)	's

Table 9: Possessive markers and their meaning

Example:| မေမေရဲ့ | ကျေးဇူးကို အောက်မေ့ပါ သည်။
I remember | mother's | kindness.

6.1.8 Accordance Markers

Markers that specify an action or event occurs in accordance with something can be defined accordance markers.

Literary Style	Colloquial Style	Translation
အလိုက်၊အရ၊ အလျောက်၊ ...	အရ၊အတိုင်း၊ အညီ၊ ...	as, according to

Table 10: Accordance markers and their meaning

Example:| ရေစီးအလိုက် |သွားခြင်းကို ရေစုန်ဟု ခေါ်သည်။
Going | according to the current | is called "downstream".

6.1.9 Accompaniment [coordinate] Markers

Markers that denote accompaniment and two or more items being together with two or more items are accompaniment markers.

Literary Style	Colloquial	Translation

	Style	
နှင့်၊နှင့်အတူ၊ နှင့်အညီ၊ ...	နဲ့၊ နဲ့အတူ၊ ရော...ရော၊ ...	and, with

Table 11: Coordinate markers and their meaning

Example:| မိဘ**နှင့်အတူ** | နေသည်။
She lives together | with her parents |.

6.1.10 Choice Markers

Markers that specify numbers [of persons or things] to make a choice from can be defined as choice markers.

Literary Style	Colloquial Style	Translation
တွင်၊အနက်၊အထဲမှ၊...	တွင်၊အနက်၊မှ၊ ထဲမှ၊ ...	between, among

Table 12: Choice markers and their meaning

Example: | အဖွဲ့ဝင်များ**ထဲမှ** | တစ်ယောက်ကို ခေါင်းဆောင်အဖြစ် ရွေးချယ်သည်။
One person | from among the members | is chosen as leader of the group.

6.1.11 Purpose Markers

Markers that specify the purpose and are used to denote for, for the sake of, can be defined as purpose markers.

Literary Style	Colloquial Style	Translation
အလို့ငှာ၊ဖို့၊အတွက်၊ ...	ဖို့၊အတွက်၊ရန်၊ ...	to, for

Table 13: Purpose markers and their meaning

Example: ကျောင်းသားများသည် | ဗဟုသုတ**အလို့ငှာ** | လေ့လာရေးခရီးထွက်ကြသည်။
The students set out on a study tour | to gain experience |.

6.1.12 Demonstratives and interrogatives

Demonstratives and interrogatives may be used in subordination to other nouns, as သည်အိမ်, ဟိုအိမ်, ဘယ်အိမ် (this, that, which house). They serve as adjectives followed by nouns. And they can also be used as independent nouns that can take noun markers as ဘာကို, ဘယ်မှာ (what, where). They can be segmented as noun phrases.
Example:| ဘာ | လုပ်ပေးရမလဲ။
| What | can I do for you?

6.2 Verb Markers

Markers that are attached to the verbs are called "verb markers".

6.2.1 Subordinating Markers

In simple sentences, they are generally at the end of the sentence and can be seen as independent markers. We have no need to consider how to break the sentence into phrases with these markers because their position plainly shows it. But in complex sentences, they are in the middle of the sentence and are known as dependent or subordinating markers. Subordinating markers need to be considered before breaking a sentence into phrases. We can break a set of verb and verb markers attached to it as a verb phrase. Some of subordinating markers are လျှင် (if), မ---လျှင် (unless), ကတည်းက (since), သောကြောင့် (because), သောအခါ (when) and so on.

6.2.2 Adjectival Markers

Adjectives are formed by attaching adjectival markers to verbs and they can be segmented as noun modifier phrases.

Literary Style	Colloquial Style	Translation
သော၊သည့်၊မည့်	တဲ့၊မဲ့	no English equivalent

Table 14: Adjectival markers and their meaning

Example: သူ | ပြော**သည့်** | စကားကို ကျွန်မ နားမလည်ချေ။
I didn't understand the words | he spoke |.

6.2.3 Adverbial Marker

Adverbs are formed by adding adverbial marker "စွာ -ly " to verbs and they can be segmented as verb modifier phrases. Adverbs can also be obtained by derivation (prefix and suffix) and reduplication of verbs.
Example: (adverbial marker)
| ငြိမ်သက်စွာ | နားထောင်နေကြသည်။
Listen | quietly | .
Example: (reduplication)
| ငြိမ်ငြိမ်သက်သက် | နားထောင်နေကြသည်။

Listen | quietly | .

7 Other Techniques

We can break the sentences into phrases with noun and verb markers. Moreover, we can also segment the following conditions as phrases.

7.1 Complement

A word or a group of word that serve as the subject/object complement can be considered a phrase with zero ø in Burmese.
Example: ဦးညိုမြက | သတင်းစာဆရာ ø | ဖြစ်တယ်။
U Nyo Mya is | a journalist | .

7.2 Time Phrase

A word or a group of words that show the time can be defined as a time phrase and can be segmented as a phrase (e.g., မကြာမီ soon).

7.3 Sentence Connector

Grammatical words that are used for linking two or more sentences are called sentence connectors. They are generally placed at the beginning of the second sentence. Some are သို့သော် (but), ဒါကြောင့် (therefore), သို့ရာတွင် (however), ထို့အပြင် (moreover) and so on. We regard them as sentence connectors and break them.

7.4 Interjections

A lexical word or phrase used to express an isolated emotion is called an interjection, for example; အလို (Alas!), အမလေး (Oh God) and so on. They are typically placed at the beginning of a sentence. Interjections may be word level or phrase level or sentence level. Whatever level it is, they can be considered a phrase and can be so segmented.

8 Methodology

CRF++ tool is a simple, customizable, and open source implementation of Conditional Random Fields (CRFs) for segmenting/labeling sequential data. CRF++ will be applied to a variety of NLP tasks.

In our system, we have two phases. The first one is encoding phase and the second one is decoding phase. In encoding phase, at first, we collect and normalize raw text from online and offline journals, newspapers and e-books. When we have sufficient corpus, as preprocessing task, we manually break un-segmented text by the rules mentioned above. Next, we train these sentences decoding with CRF++ tool to get Burmese phrase model. According to our Burmese language nature, we employ unigram template features of CRF implementations.

In decoding phase, un-segmented Burmese sentences are inputted to the system and then automatically encoded with Burmese phrase. As a result, we can achieve Burmese sentences that have been segmented into phrases.

9 Experimental Result

Maximum correctness of phrase segmentation performs when the test and training data come from the same category of corpus. The probability of correctness may be worse if we trained on the data from one category and tested on the data from the other one. Here we tested phrase segmentation of various types of corpus with 5000 and 50000 phrase-model of general corpus respectively.

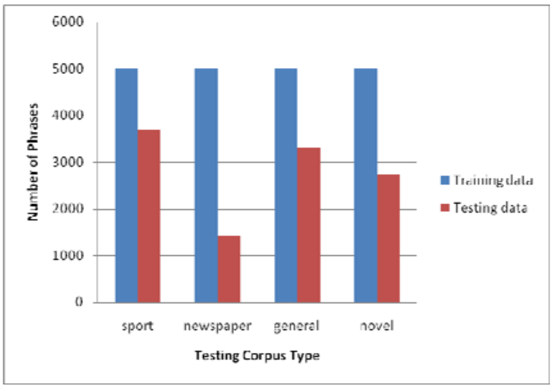


Figure 3: Result of Phrase Segmentation with 5000 phrase-model using CRF++ toolkit

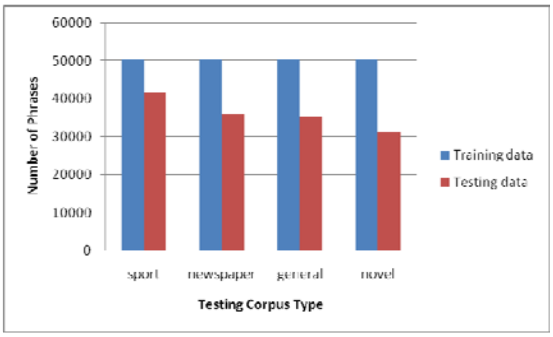


Figure 4: Result of Phrase Segmentation with 50000 phrase-model using CRF++ toolkit

It can be seen that the more sufficient training data, the more efficiency we get. Average scores of phrase segmentation are above 70% according to the F-Measure. The corresponding scores are:

Corpus Type	Score

sport	83%
newspaper	72%
general	70%
novel	62%

Table 15: Various corpus types and their scores

10 Known Issues

Although scores are highly efficient, we face some difficulties that we cannot solve. For example, we can manually segment a sentence into phrases with zero markers such that complement, time and adverbs formed by derivation as a phrase whether it has been attached with markers or not. But in our system, it is difficult to achieve best results because of zero markers. We need more and more training data to cover these zero marker phrases. Boundaries of these phrases may be various. So, we can only get about 50% accuracy for these types of phrases.

Another problem is homonyms. For example: ကို 'Ko' is object marker but it may also be the title of a name like ကိုချမ်းမြေ့ 'Ko Chan Myae'. As a title of a name, we do not need to segment ကိုချမ်းမြေ့ 'Ko Chan Myae'. But CRF++ tool will segment this phrase as ကို 'Ko' and ချမ်းမြေ့ 'Chan Myae' depending on the probability of training data.

11 Conclusion

In this study, we have developed an automatic phrase segmentation system for Burmese language. The segmentation of sentences into phrases is an important task in NLP. So, we have described how we can segment sentences into phrases with noun markers, verb markers, zero markers and other techniques in this paper. We hope this work will help accelerate NLP processing of Burmese language such as Machine Translation, Text summarization, Text Categorization, Information Extraction and Information Retrieval and so on.

12 Further Extension

As we mentioned in section 2.1, the combination of two or more words becomes a phrase. It is easier to segment words after decomposing the phrases of sentence. The result of phrase segmentation will help the word segmentation. Moreover, we can build Burmese parser based on phrase segmentation.

Acknowledgment

The authors are happy to acknowledge their debt and offer grateful thanks Mr. Tha Noe, linguist and Mr. Ngwe Tun, CEO, Solveware Solution.

The authors sincerely want to express acknowledgement to their colleagues of Myanmar Natural Language Processing Team and technicians who helped them and guided on this paper.

References

J.A. Stewart. 1955. *Manual of Colloquial Burmese*. London.

J. Lafferty, A. McCallum, and F. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, In Proc. of ICML, pp.282-289. USA.

John Okell and Anna Allott. 2001. *Burmese/Myanmar Dictionary of Grammatical Form*. Curzon Press,UK.

John Okell.1969. *A Reference Grammar of Colloquial Burmese. London*. Oxford University Press.

Kathleen Forbes.1969.*The Parts of Speech In Burmese and The Burmese Qualifiers*. JBRS, LIII, ii. Arts & Sciences University, Mandalay.

Pe Maung Tin, U. 1954. *Some Features of the Burmese Language*. Myanmar Book Centre & Book Promotion & Service Ltd. Bangkok, Thailand.

Willian Cornyn. 1944. *Outline of Burmese Grammar* , Language Dissertation No.38, Supplement to Language, volume 20, No,4.

http://crfpp.sourceforge.net

ဦးခင်အေး. *မြန်မာသဒ္ဒါနှင့် ဝါ စင်္ဂရှစ်ပါးပြဿနာ*, အတွဲ (၁၃), အပိုင်း (၅), တက္ကသိုလ်ပညာပဒေသာ စာစောင်.

ဦးဖေမောင်တင်. ၁၉၆၅ *အလယ်တန်းမြန်မာသဒ္ဒါ* စာပေဗိမာန်.

မြန်မာသဒ္ဒါ. ၂၀၀၅.မြန်မာစာအဖွဲ့ဦးစီးဌာန.