# Normalization of Myanmar Grammatical Categories for Part-of-Speech Tagging

Phyu Hninn Myint
University of Computer Studies
Yangon
Myanmar

Tin Myat Htwe
University of Computer Studies
Yangon
Myanmar

Ni Lar Thein
University of Computer Studies
Yangon
Myanmar

## ABSTRACT
In this paper, we analyze the syntactic structure of Myanmar grammatical categories to be able to use in tagging Myanmar text with standard Part-of-Speech (POS) tags. In Myanmar lexicon, all words are annotated with basic tags and these words can be called as stem words or root words. The Myanmar POS tagged corpus creation, which has been proposed in [11], used basic POS tagging for each word. Therefore, all words in this corpus have been tagged with only basic tags as in lexicon. For standard POS tagging, normalization step is needed to form more meaningful words and annotate some words with more appropriate finer POS tags and categories. The finer tags can be called as standard POS tags and these can be used to directly concatenate with English POS tags. These tags are very useful in Myanmar to English Machine Translation System. Hence, the main aim of this study is to develop the customized lexical rules in order to deduce finer or standard POS tag from basic POS tags combinations. By analyzing Myanmar grammatical categories, 27 rules are defined to normalize them. Evaluation has been made on a basic POS tagged corpus which contains 1000 basic POS tagged sentences and it yields full satisfaction for all words in these sentences.

## General Terms
Natural Language Processing, Machine Translation, Syntactic Analysis

## Keywords
Part-of-Speech tagging, Normalization of Grammatical Categories

## 1. INTRODUCTION
Myanmar is a country which is situated in South East Asia and it is a member of ASEAN. In the past, Myanmar was called Burma and its language was called Burmese. Nowadays, in our country, our native language is officially called Myanmar Language.

Myanmar language is the official language in Myanmar. It is a tonal and syllable-based language. Myanmar scripts are adopted from Mon script (one of the Myanmar main national races) that is derived from India Brahmi script. Myanmar language is the native language of the Bamar (main nationality of Myanmar) and related sub-ethnic groups of the Bamar, as well as that of some ethnic minorities in Myanmar like the Mon. It is spoken by 32 million as a first language and as a second language by 10 million, particularly ethnic minorities in Myanmar and those in neighboring countries. Myanmar language is a tonal, pitch-register, and syllable-timed language, largely monosyllabic and analytic language, with a subject–object–verb word order. It is a member of the Tibeto-Burman language family, which is a subfamily of the Sino-Tibetan family of languages. The language uses the Myanmar script, derived from the Old Mon script and ultimately from the Brāhmī script [4].

The basic word order of the Myanmar language is subject-object-verb. Pronouns in Myanmar vary according to the gender and status of the audience. Myanmar is monosyllabic (i.e., every word is a root to which a particle but not another word may be prefixed). Sentence structure determines syntactical relations and verbs are not conjugated. Instead they have particles suffixed to them. For example, the verb "to eat," စား (-sar) is itself unchanged when modified.

We have to analyze the syntactic structure of Myanmar grammatical categories to be able to use in tagging Myanmar text with standard Part-of-Speech tags. In Myanmar lexicon, all words are defined with basic tags and these words can be called as stem words or root words. There are nine Part-of-Speech classes for all Myanmar words since it is described by Myanmar Language Commission [3]. These are Noun, Pronoun, Verb, Adjective, Adverb, Postpositional Marker, Particles and Interjection. In English language, only eighth Part-of-Speech classes are classified. Preposition class in English is mostly the same with Postpositional Marker in Myanmar. The additional class in Myanmar is Particles class. In Myanmar language, the words in Particles class are meaningless words, so one Particles word has no meaning and it cannot stand by itself as a meaningful word. But it can be used as an affix to other Part-of-Speech classes to be formed a meaningful word. Moreover, it is possible to change the Part-of-Speech class of a meaningful word by affixing one or more Particles word with this word. Also, some Part-of-Speech classes can be combined to form another Part-of-Speech class.

Normalizing which is forming finer tag can be done to make the analysis to understand the nature of Myanmar language. Normalization step is needed to form more meaningful words and annotate with more appropriate finer POS tags and categories. In our language, Myanmar, there are many "Particles" in the text. These can be appeared in binding with Noun, Verb, Adjective and Adverb in the text. Moreover, these can convert the type of POS tag, that is, Noun attached with some particles can become Verb or Adjective. Also, Verb or Adjective with some particles can create new POS tag, which is Adjective with superlative or comparative degree. There are the same pattern and particle to transform from one POS tag to another. Therefore, some lexical rules have to be developed to deduce more finer and standard POS tag. The normalization rules and the detail examples are explained in Section IV.

The rest of the paper is organized as follows: Section II introduces a brief of Myanmar language. Section III describes the Part-of-Speech tagging. Section IV discusses our analysis for developing rules. Section V presents performance analysis. Finally, some conclusions on this work are given in section VI.

## 2. MYANMAR LANGUAGE

In Myanmar language, there are 34 basic consonants, 8 basic vowels, 4 medial consonants or dependent consonant signs, dependent various signs, 2 punctuation marks and 10 digits. Vowels can be divided into independent and dependent vowels. Independent vowels can stand alone and dependent vowels are written with a consonant [4].

Basically, there are two types of Myanmar sentences: formal and informal. Formal sentences are used in official letter, newspaper, online news, etc. and can be called as literary (written) form. Informal sentences are used in spoken language and can be defined as colloquial form.

There are nine types of Part-of-Speech in Myanmar language [3].

### 2.1 Adjective

Myanmar language does not have adjectives per sentence. Rather, it has verbs that carry the meaning "to be X", where X is an English adjective. These verbs can modify a noun by means of the grammatical particle တဲ့ (-dae`) in colloquial Myanmar language, or သော (-thaw) in literary Myanmar language. For example, "beautiful person" is ordered: "be beautiful" + adjective particle + "person" (colloquially ချောတဲ့လူ [chaw-dae`-lu], formally ချောသောလူ [chaw-thaw-lu]).

Adjectives may also form a compound with the noun (e.g. လူချော [lu-chaw] "person" + "be beautiful"). Comparatives are usually ordered: X + ထက်ပို [htet-po] + adjective, where X is the object being compared to. Superlatives are indicated with the prefix အ [-a] + adjective + ဆုံး [-sone] [4].

### 2.2 Verb

The roots of Myanmar language verbs are almost always suffixed with at least one particle which conveys such information as tense, intention, politeness, mood, etc. Many of these particles also have formal/literary and colloquial equivalents. In fact, the only time in which no particle is attached to a verb is in imperative commands. However, Myanmar language verbs are not conjugated in the same way as most European languages; the root of the Myanmar verb always remains unchanged and does not have to agree with the subject in person, number or gender.

Verbs are negated by the particle မ [-ma], which is prefixed to the verb [4].

### 2.3 Noun

Nouns in Myanmar language are pluralized by suffixing the particle တွေ [-dway] in colloquial Myanmar language or များ [-myar] in formal Myanmar language. The particle တို့ [-doh], which indicates a group of persons or things, is also suffixed to the modified noun [4].

### 2.4 Particle

The Myanmar language makes prominent usage of particles (called ပစ္စည်း in Myanmar), which are untranslatable words that are suffixed or prefixed to words to indicate level of respect, grammatical tense, or mood. According to the Myanmar-English Dictionary, there are 449 particles in the Myanmar language. For example, ပေး [-pay] is a grammatical particle used to indicate the imperative mood. While လုပ်ပါ ("work" + particle indicating politeness) does not indicate the imperative, လုပ်ပေးပါ ("work" + particle indicating imperative mood + particle indicating politeness) does. Particles may be combined in some cases, especially those modifying verbs.

Some particles modify the word's part of speech. Among the most prominent of these is the particle အ [-a], which is prefixed to verbs and adjectives to form nouns or adverbs. For instance, the word ဝင် means "to enter," but combined with အ, it means "entrance" (အဝင်). Also, the second အ in words can follow the pattern အ + noun/adverb + အ + noun/adverb, like အဆောက်အအုံ, which is formally pronounced [-a-sought-a-ohn] [4].

### 2.5 Pronoun

Subject pronouns begin sentences, though the subject is generally omitted in the imperative forms and in conversation. Grammatically speaking, subject marker particles (က ([-kạ] in colloquial, သည် [-thi] in formal) must be attached to the subject pronoun, although they are also generally omitted in conversation. Object pronouns must have an object marker particle (ကို [-ko] in colloquial, အား [-arr] in formal) attached immediately after the pronoun. Proper nouns are often substituted for pronouns. One's status in relation to the audience determines the pronouns used, with certain pronouns used for different audiences.

The contraction also occurs in some low toned nouns, making them possessive nouns (e.g. အမေ့ or မြန်မာ့, "mother's" and "Myanmar's" respectively) [4].

### 2.6 Reduplication

Reduplication is prevalent in Myanmar and is used to intensify or weaken adjectives' meanings. For example, ချော ([-chaw] "beautiful") is reduplicated, the intensity of the adjective's meaning increases. Many Myanmar words, especially adjectives with two syllables, such as လှပ ([-hl̥ạ-pạ] "beautiful"), when reduplicated (လှပ → လှလှပပ [hla-hla-pa-pa]) become adverbs. This is also true of some Myanmar verbs and nouns (e.g. ခဏ [kha-na] "a moment" → ခဏခဏ [kha-na-kha-na] "frequently"), which become adverbs when reduplicated.

Some nouns are also reduplicated to indicate plurality. For instance, ပြည် ([-pyi] "country"), but when reduplicated to အပြည်ပြည် ([-a-pyi-pyi] "country"), means "many countries," as in အပြည်ပြည်ဆိုင်ရာ ([-a-pyi-pyi-sine-yar] "international"). Another example is အမျိုး[a-myo], which means "a kind," but the reduplicated form အမျိုးမျိုး [a-myo-myo] means "multiple kinds."

A few measure words can also be reduplicated to indicate "one or the other": ယောက် [yout] (measure word for people) → တစ်ယောက်ယောက် [ta-yout-yout] (someone) and ခု [khu] (measure word for things) → တစ်ခုခု [ta-khu-khu] (something)[4].

# 3. PART-OF-SPEECH TAGGING

Part-of-speech (POS) tagging is the act of assigning each word in sentences a tag that describes how that word is used in the sentences. That means POS tagging assigns whether a given word is used as a noun, adjective, verb, etc. One of the most well-known disambiguation problems is POS tagging. A POS tagger attempts to assign the corresponding POS tag to each word in sentences, taking into account the context in which this word appears.

## 3.1 Basic Part-of-Speech Tagging

Because of the data sparseness, we use stem words in the lexicon for POS tagging. According to Myanmar grammar book and dictionary book [3][2], there are nine Part-of-Speech tags in Myanmar language. Basically, we have annotated each word with these nine basic POS tags and we have created a POS tagged corpus in which every word is tagged with appropriate basic POS tags. Moreover, we have developed a lexicon which stores stem words with basic POS tags. This lexicon is used for tagging the input untagged words with all possible tags. As shown in figure 1, each word is attached with its possible basic POS tags. The POS tag has two parts: first is basic POS tag and second is specific category of this POS tag.

```
မိန်းကလေး <girl> # NN.Person
မှာ        <in> # PPM.Place # PPM.Time
```

**Fig 1: Example of some words in the lexicon**

The basic POS tags are depicted in the table, Table 1.

## 3.2 Specific Category Designating

In each basic POS tag, it is possible to have more than one subclass for designating specific category. For instance, "မိန်းကလေး" <girl> is tagged as a noun<NN> in general. Moreover, it is also categorized as a person in specific. Therefore, we have categorized every word with general class and specific class. The general classes are nine basic POS classes and the specific classes are dedicated to describe the detail of each POS class. Table 2 shows the specific categories for basic tags.

**Table 1. Basic Part-of-Speech Tags**

| No. | POS Description | Example |
|---|---|---|
| 1. | Singular Noun | မိန်းကလေး *<girl>* |
| 2. | Singular Pronoun | သူ *<he>* |
| 3. | Adjective | ကောင်း *<good>* |
| 4. | Adverb | အလွန် *<very>* |
| 5. | Verb | ပြေး *<run>* |
| 6. | Conjunction | နှင့် *<and>* |
| 7. | Particles | သော *<-thaw>*၊ အ *<-a>* |
| 8. | Postpositional Marker | သို့ *<to>* |
| 9. | Interjection | ဘုရားရေ *<oh my god>* |

**Table 2. Specific Categories for Basic Tags**

| No. | POS Name | Category Name |
|---|---|---|
| 1. | Noun | • Common<br>• Person<br>• Animals<br>• Time<br>• Body<br>• Building<br>• Objects<br>• Location<br>• Cognition<br>• Attribute |
| 2. | Pronoun | • Person<br>• Distplace<br>• Disttime<br>• Distobj<br>• Question<br>• Reflexive<br>• Possessive |
| 3. | Adjective | • Demonstrative<br>• Distplace<br>• Disttime<br>• Distobj<br>• Quantity<br>• Question |
| 4. | Adverb | • Time<br>• Manner<br>• State<br>• Quantity<br>• Question |
| 5. | Verb | • Common<br>• Compound |
| 6. | Postpositional Marker | • Subject<br>• Object<br>• Leave<br>• Direction<br>• Arrive<br>• Used<br>• Cause<br>• Accept<br>• Place<br>• Time<br>• Agree<br>• Extract<br>• Possessive<br>• Time Start<br>• Time End |
| 7. | Particles | • Type<br>• Common<br>• Number<br>• Support<br>• Interjection<br>• Negative<br>• Quantity<br>• Example |
| 8. | Conjunction | • Sentence<br>• Mean<br>• Chunk |
| 9. | Interjection | • Common |

## 3.3 Standard Part-of-Speech Tagging

The POS tags in the lexicon which tag stem words are denoted as basic POS tags. However, most of the basic level POS tags cannot cope with all words which are formed from the combination of stem words. Since it is possible to do the direct translation to English language or other language if the POS tags of the combined word are known and only basic tags are not enough for all words combination, the development of the standard POS tags are needed. After disambiguation, standard POS tags have to be used to tag to some combinations of stem words if the combination patterns are matched. The standard POS tags are depicted in the following table, Table 3.

**Table 3. Standard Part-of-Speech Tags**

| No. | Description | Example |
|---|---|---|
| 1. | Plural Noun | လူများ *<people>* |
| 2. | Plural Pronoun | သူတို့ *<they>* |
| 3. | Comparative Adjective | ပိုကောင်း*<better>* |
| 4. | Superlative Adjective | အကောင်းဆုံး *<best>* |
| 5. | Comparative Adverb | ပိုမြန် *<quicker>* |
| 6. | Superlative Adverb | အမြန်ဆုံး*<quickest>* |
| 7. | Negative Verb | မလုပ် *<do not>* |
| 8. | Negative Adjective | မကောင်း *<not good>* |
| 9. | Noun (Adjective Convert) | အကောင်း *<goodness>* |
| 10. | Noun (Verb Convert) | ပြုလုပ်ခြင်း *<doing>* |
| 11. | Adverb (Verb Convert) | လျင်လျင်မြန်မြန် *<quickly>* |
| 12. | Adverb (Adjective Convert) | ပျော်ပျော်ပါးပါး *<happily>* |

In addition, it needs to define new categories for standard POS tags. Table 4 describes the additional specific categories for standard tagging.

**Table 4. Specific Categories for Standard Tags**

| No. | POS Name | Category Name |
|---|---|---|
| 1. | Noun | • JJConvert<br>• VBConvert |
| 2. | Adjective | • VBConvert<br>• NNConvert |
| 3. | Adverb | • VBConvert<br>• JJConvert |
| 4. | Negative Verb | • Common |
| 5. | Negative Adjective | • Demonstrative |

# 4. ANALYSIS FOR DEVELOPING RULES

After disambiguation, lexical rules have to be created for finer POS tagging and, using these rules, finer and standard POS tags can be produced for some words. These finer tags are able to be applied in the later steps of NLP applications. It is possible that word with finer tag can be directly translated to other language. We have to analyze "Particles" which are functional words to develop most of the lexical rules.

In Myanmar language, there are many particles which can be called affixes of the word and can cause the changes of sense or type of that word. The prefixes are "မ-"(ma-), "အ-" (a-) and "တ-"(ta-). The prefix "မ-" (ma-) is an immediate constituent of the verb, which is the head of the word construction as in: ma-swa: မ-သွား: 'not go'; ma-kaung: မ-ကောင်း : 'not good'. It changes the positive sense to negative sense of the word. The scope of verbal negation extends to the whole compound of a compound verb, as in ma-tang pra: မ-တင်ပြ : 'not submit'; ma-saung-ywat : မ-ဆောင်ရွက် : 'not carry out'. Another pattern of negation is possible with verb compounds or verb phrases by individualized negation of each portion of the compound, as in: ma-ip ma-ne : မ-အိပ် မ -နေ : 'not sleep at all'; ma-tang ma-kya: မ-တင် မ -ကျ : 'noncommittal'.

The prefix "အ-" (a-) is a type converter which is the head word of the verb or adjective as in: a-lote: အ-လုပ် : 'work or job'; a-hla : အ-လှ : 'beauty'. The prefix "တ-" ( ta-) can also be seen as a type converter, as in ta-lwal ta-chaw: တ-လွဲ တ -ချော် : 'wrongly'.

The postfixes are "-မှု" (-mhu), "-ခြင်း" (-ching), "-ချက်" (-chat), "-ရေး" (-yay), "-နည်း" (-nee), "-စွာ" (-swar), "-သော" (-thaw), "သည့်" (-thi), "-မည့် " (-myi), etc. The postfixes "-မှု" (-mhu), "-ခြင်း" (-ching), "-ချက်" (-chat), "-ရေး" (-yay), "-နည်း" (-nee) change the type of the previous POS tag from verb or adjective or adverb to noun. The words ended with these postfixes are in the noun form. Also, the postfixes "-သော" (-thaw), "သည့်" (-thi), "-မည့် " (-myi) convert to the adjective form from adjective or adverb or verb. The postfixes "-စွာ" (-swar) alters the type of adjective or verb or adverb to form adverb. In noun form, the postfixes "-များ" (-myar), "-တို့" (-doh) change the singular noun to plural noun.

Moreover, in adjective, if JJ tag is lied between two affixes "အ-" (a-) and "ဆုံး-" (-sone), this tag JJ become to JJS (superlative degree), i.e., " အ JJ ဆုံး " is equal to "JJS".

There are many rules up to 27 rules for normalization of our language. Some of the normalization rules are described as follows ::

Rule (1)

- Singular Noun+ ( များ | တို့ | တွေ) = Plural Noun
- Singular Pronoun + ( များ | တို့ | တွေ) = Plural Pronoun

For rule(1), there are some particles "များ" <-myar>, "တွေ"<-doh>, "တို့"<-dway>, which can make some changes on a word tagged as a singular noun and it is possible to create a new word that is a singular noun attached with a particle and it is tagged as plural noun.

Example (1)

သူ <*he*> **Pronoun.Person**
+
တို့ <*-doh*>
**Particle.Number**

→ သူတို့ <*they*> **PluralPronoun.Person**

Example (1) shows an instance for rule (1), that is, "he" (Singular Noun) attached with "တို့" <-doh> becomes "they" (Plural Noun).

For rule (2), there are some particles "သော"<-thaw>, "သည့်"<-thi>, "မည့်"<-myi> which can attached with a verb or an adjective and this combination can become an adjective.

Rule (2)

- ( Verb | Adjective ) + ( သော | သည့် | မည့်) = Adjective
- ( Verb | Adjective ) + Part.Support* + ( သော | သည့် | မည့်) = Adjective

In example (2), when an adjective is attached with "သော"<-thaw>, new word "ကောင်းသော (kg-thaw)"<good> is also an adjective. If a verb is attached with "သော"<-thaw>, it gives a new word that is an adjective but it has new category, VerbConvert. Also, between the verb and "သော"<-thaw>, one or more Part.Support tags can be seen and they can be combined to form a new word.

Example (2)

ကောင်း <*good*>
**Adjective.Demonstrative**
+
သော <*-thaw*>
**Particle.Common**

→ ကောင်းသော <*good*>
**Adjective.Demonstrative**

---

ပေး <*give*>
**Verb.Common**
+
ထား <*-htar*>
**Particle.Support**
+
သော <*-thaw*>
**Particle.Common**

→ ပေးထားသော <*given*>
**Adjective.VerbConvert**

For rule (3), the postfixes Particles "-မှု"<-mhu>, "-ခြင်း"<-ching>, "-ချက်"<-chat>, "-ရေး"<-yay>, "-နည်း"<-nee> can be attached with a verb or adjective and it makes to form a new word tagged with noun.

Rule (3)

- ( Verb | Adjective ) + ( မှု | ခြင်း | ချက် | ရေး | နည်း ) = Noun
- ( Verb | Adjective ) + Part.Support* + ( မှု | ခြင်း | ချက် | ရေး | နည်း ) = Noun

In example (3), when an adjective is attached with "-ခြင်း"<-chin>, new word "ကောင်းခြင်း (kg-chin)" is tagged with noun.

Example (3)

ကောင်း <*good*>
**Adjective.Demonstrative**
+
ခြင်း <*-thaw*>
**Particle.Common**

→ ကောင်းခြင်း <*goodness*>
**Noun.AdjectiveConvert**

---

ပေး <*give*>
**Verb.Common**
+
ထား <*-htar*>
**Particle.Support**
+
ခဲ့ <*-khae*>
**Particle.Support**
+
ခြင်း <*-thaw*>
**Particle.Common**

→ ပေးထားခဲ့ခြင်း <*giving*>
**Noun.VerbConvert**

For rule (4), the prefix "အ-" <a-> can change a verb or an adjective to form a new word tagged with noun.

Rule (4)

- အ + ( Verb | Adjective ) = Noun

Example (4)

အ <*-a*> **Particle.Common**
+
ကောင်း <*good*>
**Adjective.Demonstrative**

→ အကောင်း <*goodness*>
**Noun.AdjConvert**

---

အ <*-a*> **Particle.Common**
+
ရောင်း <*sell*>
**Verb.Common**

→ အရောင်း <*sale*>
**Noun.VerbConvert**

Example (4) shows that when an adjective or a verb is prefixed with "အ-" <a->, a new word can be tagged with noun but it has different categories for adjective or verb such as AdjConvert and VerbConvert.

Rule (5)

- ( Verb | Adjective | Adverb ) + စွာ = Adverb

For rule (5), the postfix "စွာ" <-swar> can change a verb or an adjective or an adverb to form a new word tagged with adverb.

Example (5)

ကောင်း *<good>*
**Adjective.Demonstrative**
**+**
စွာ *<-a>* **Particle.Common**

→ ကောင်းစွာ *<good>* **Adverb.AdjConvert**

လျင်မြန် *<quick>*
**Verb.Common**
**+**
စွာ *<-a>* **Particle.Common**

→ လျင်မြန်စွာ *<quickly>* **Adverb.VerbConvert**

Example (5) shows that when an adjective or a verb is postfixed with "စွာ" <-swar>, a new word can be tagged with adverb tag but it has different categories for adjective or verb such as AdjConvert and VerbConvert.

Rule (6)

- တ + Verb + တ + Verb = Adverb
- အ + Verb + အ + Verb = Adverb
- မ + Verb + မ + Verb = Adverb
- အ + Verb + တ + Verb = Adverb

For rule (6), there are some particles such as "အ-"<a->, "တ-" <ta->, "မ-" <ma->, which can create an adverb when they lie between verbs or adjectives.

In example (6), two "အ-" <a-> are situated before two verbs and " 'အ-' <a-> + verb + 'အ-' <a-> + verb " pattern makes a new word tagged with Adverb.Convert.

Example (6)

မ *<-ma>*
**Particle.Common**
**+**
ပြော *<talk>*
**Verb.Common**
**+**
မ *<-ma>*
**Particle.Common**
**+**
ဆို *<talk>* **Verb.Common**

→ မပြောမဆို *<without talking>* **Adverb.Convert**

တ *<-ta>*
**Particle.Common**
**+**
လွဲ *<miss>* **Verb.Common**
**+**
တ *<-ta>*
**Particle.Common**
**+**
ချော် *<miss>*
**Verb.Common**

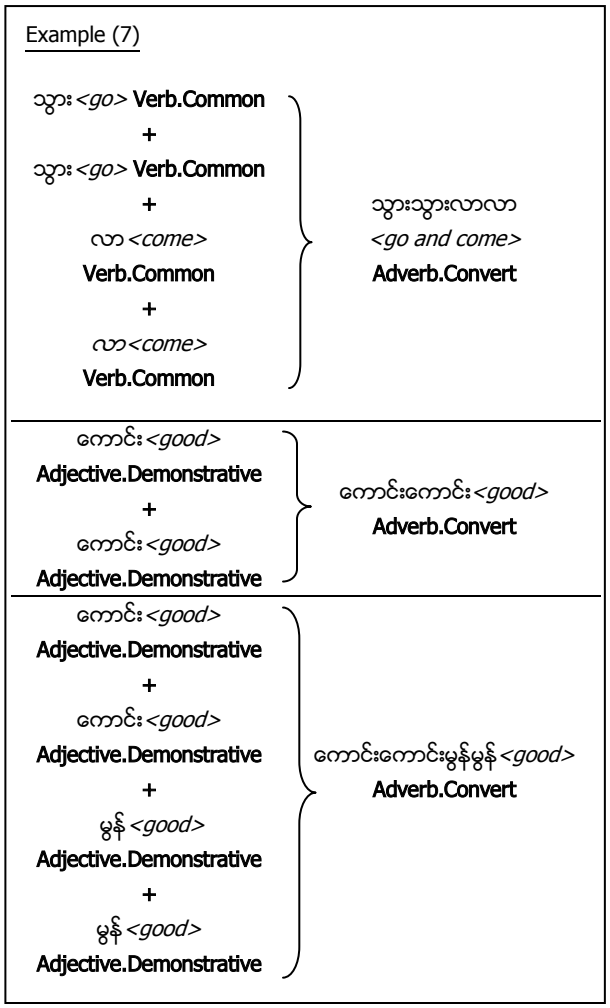
→ တလွဲတချော် *<wrongly>* **Adverb.Convert**

In rule (7), there are some patterns by duplication stem words. When a verb or an adjective can be appeared twice, it can create an adverb. A pair of the same two verbs or adjectives can also produce an adverb.

Rule (7)

- Verb1 + Verb2 = Adverb *<Verb1==Verb2>*
- Verb1 + Verb2 + Verb3 + Verb4 = Adverb
  *<Verb1==Verb2 && Verb3==Verb4 >*
- Adjective1 + Adjective2 = Adverb
  *<Adjective1==Adjective2>*
- Adjective1 + Adjective2 + Adjective3 + Adjective4 = Adverb
  *< Adjective1 == Adjective2 && Adjective3== Adjective4>*

In example (7), "သွား" <go> and "လာ" <come> are verbs but they are duplicated as "သွားသွားလာလာ" <go and come> and it becomes an adverb.

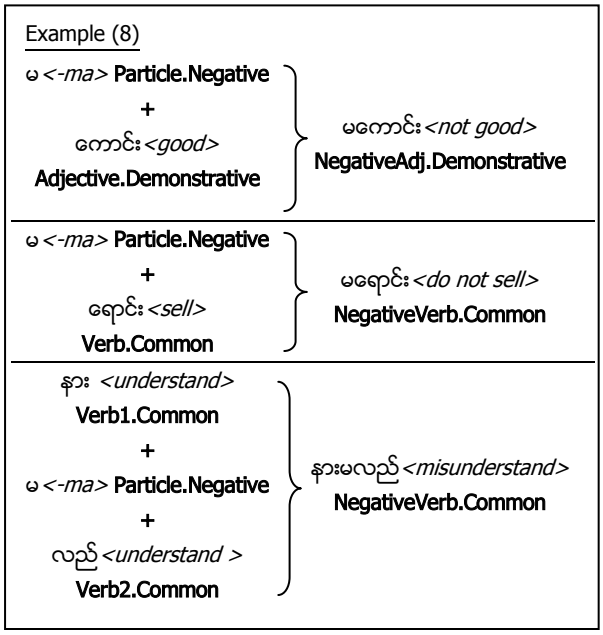



In rule (8), the prefix "မ-"<ma-> can cause a positive verb or adjective to form a negative one. Moreover, some peculiar negative verbs have a strange pattern such as "verb1 + negative particle + verb2 → negative verb". In these verbs, there are two separated words and one negative particle lied in the middle of two words. These two words are tagged as verb1 and verb2 as basic tags.

Rule (8)

- မ + ( Verb | Adjective ) = Negative Verb | Negative Adjective
- Verb1 + မ + Verb2 = Negative Verb

In example (8), "ကောင်း" <good> is a positive adjective and " ရောင်း " <sell> is a positive verb and when "မ" <-ma> is prefixed to them, negative adjective "မကောင်း" <not good> and negative verb "မရောင်း" <not sell> are formed. Moreover, "နားမလည်"<understand> is a positive verb and it has an exceptional case for negation. The negative particle "မ" <-ma> can only be added in the middle of the word to form a negative verb. So such verbs must be separately tagged with verb1 and verb2 as basic tags such as "နား"(Verb1) + "မ"(Negative Particle) + "လည်"(Verb2).




## 5. PERFORMANCE ANALYSIS

In order to measure the performance of the system, we have tested many experiments using our approach on different types of sentences till we get the best accuracy. We can evaluate the result how many wrong words are tagged and how many words can be correctly tagged. Therefore, the performance of our lexical rules is evaluated in terms of the problems that can be encountered in Myanmar sentences because of some peculiar word combination patterns. The sentences that have peculiar patterns are entered into the system and check the accuracy of our rules.

Some errors can occur for some words especially for negative word because unusual pattern of verbal negation is found in such patterns where the second verb of a compound is marked with the negative prefix, as in ne-ma-kaung: နေ-မ-ကောင်း: 'unwell', nar-ma-lal: နား-မ-လည် : 'misunderstand', etc. Although most of the negative verbs can be formed by combining a negative particle "မ-" (ma-) with a verb, some negative verbs can be formed from a combination pattern which has a negative particle "မ-" (ma-) lied between two words of verb. To alleviate these errors, we have to be tagged these words with their basic tags such as verb1 + negative particle + verb2 and then, these three words can be combined together and tagged as a negative verb. If not so, only two words, that is, negative particle + verb, can be combined and tagged as a negative verb.

We have done our testing on the tagged sentences from the corpus which is created and proposed in [11]. All words in the sentences of this corpus are tagged with basic POS tags. We used these sentences as input text for testing the performance of our rules. The corpus has around 1000 Myanmar sentences and its average sentence length is about 10 words.

Our lexical rules can solve all basic POS tags combinations and produce the right standard POS tags for all combinations. After testing with all sentences in the corpus, we have done on testing with the assorted formal sentences from retrieving Myanmar newspapers and online journals and tagging manually with basic POS tags. Most of the words can be determined by rules. However, when we encountered new structure of word combinations, we have added or updated our lexical rules and tested again with these new rules. Therefore, our rules set become larger and it can be able to handle more words combinations. But we need to create additional rules for solving the words combinations in the informal sentences.

## 6. CONCLUSION

This paper introduces our lexical rules for normalizing grammatical categories in Myanmar language. Our lexical rules have to be applied to normalize some words and basic tags in order to produce more accurate and finer tags called standard tags. Therefore, these rules can be used to develop the proposed standard POS tagging. The standard POS tags can be directly concatenated with English POS tags and they are very useful to be used in Myanmar to English Machine Translation System. Total 27 rules have been developed that can cope with all the sentences in the corpus created in [11]. These sentences are well formed and formal sentences. Furthermore, we have to analyze informal structure of the sentence and try to create more rules that can solve all types of word combinations in the informal sentence in the future.

"Myanmar-English Dictionary" [3] and "Myanmar Grammar" [2] books published by Myanmar Language Commission are used as references for POS tagging and analyzing Myanmar words. One of the improvements to be done is adding more lexical rules in order to do more accurate normalization.

For future work, we hope to conduct more experiments to examine how different types of input affect the performance. This approach can be used in a number of NLP applications. In Myanmar to English machine translation system, Grammatical Function Assignment, Word Sense Disambiguation, Translation Model and Reordering systems have to use these standard POS tags for analyzing Myanmar words in order to translate Myanmar text to English text.

## 7. ACKNOWLEDGMENTS

The support of University of Computer Studies, Yangon, Myanmar is greatly acknowledged. The first author would also like to acknowledge the invaluable help and support from her supervisor, co-supervisor, teachers, friends and family.

## 8. REFERENCES

[1] Bradley, D. 2010. The Characteristics of the Burmic family of Tibeto-Burman. The International Symposium on Sino-Tibetan Comparative Studies in the 21st Century. Institute of Linguistics. Academia Sinica. Taipei. Taiwan.

[2] Department of the Myanmar Language Commission. 2005. Myanmar Grammar. Ministry of Education. Myanmar.

[3] Department of the Myanmar Language Commission. 2006. Myanmar-English Dictionary. Ministry of Education. Myanmar.

[4] Grammar. Burmese language. http://en.wikipedia.org/wiki/Burmese_Language

[5] Hopple, P. 1999. Nominalization in Burmese - sentence patterns. The 32nd International Conference on Sino-Tibetan Languages and Linguistics. Urbana-Champaign.

[6] Hopple, P. 2003. The structure of nominalization in burmese. Ph.D Dissertation. University of Texas. Arlington.

[7] Hopple, P. Burmese Particles as Boundary Marking Units of Text. http:/ic.payap.ac.th/graduate/linguistics/papers/Burmese_Particles.pdf?v=1289545363

[8] Judson, A. 1842. Grammatical Notices of the Burmese Language. Maulmain: American Baptist Mission Press.

[9] Ko, Taw Sein. 1924. Elementary handbook of the Burmese language. Rangoon: American Baptist Mission Press.

[10] Latter, T. 1845. A grammar of the language of burmah. Baptist Mission Press.

[11] Myint, P. H. 2010. Assigning automatically Part-of-Speech tags to build tagged corpus for Myanmar language. The Fifth Conference on Parallel Soft Computing. Yangon. Myanmar.

[12] Soe, M. 1999. A grammar of Burmese. Ph.D. dissertation. University of Oregon.

[13] Thurgood, G. 1977. Burmese Historical Morphology. Proceedings of the 3rd Annual Meeting of the Berkeley Linguistics Society.

[14] Wright, E. 1877. Anglo-Burmese Student's Assistant. Tenasserim Press.