Modifying NOVA-annotated Myanmar Data to Universal Part-of-Speech Tagset

Sann Su Su Yee[†‡], Chenchen Ding[†], Khin Mar Soe[‡]
Masao Utiyama[†], Eiichiro Sumita[†]
[†]*Advanced Translation Technology Laboratory, ASTREC, NICT, Kyoto, Japan*
[‡]*Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar*
[‡]*{sannsusuyee, khinmarsoe}@ucsy.edu.mm,*
[†]*{chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp*

Abstract

We present our work for morphological annotation on Myanmar part of the Asian Language Treebank project. Former NOVA annotation is a feasible and flexible annotation system for joint tokenization and POS tagging for under-resource languages to bridge Universal part-of-speech tagsets. And this work relate to transform of NOVA to Universal part-of-speech tagset which aims to develop cross-language consistent annotation for Myanmar, furthermore, syntax annotation to construct Treebank. This annotation is more detailed than previous one that adpositions, conjunctions, and particles are further specified.

1. Introduction

Basic linguistically annotated corpora of natural language data are required by state-of-the-art Natural Language Processing (NLP) techniques. In order to develop Asian languages in NLP techniques, Asian Language Treebank (ALT) [1] project [1] provides word segmentation, part-of-speech (POS) tagging and syntactic annotation for eight official languages used in ASEAN, including Indonesian, Khmer, Malay, Myanmar, Vietnamese, Filipino, Laotian, and Thai.

In the ALT, there were 20000-sentences, which were randomly selected news and 1888 articles from English Wikinews (Wikinews, 2014), has been created for Myanmar corpus that was manually translated from the English corpus by using the ALT web-based software tool that created user-friendly working environment for users to perform word segmentation, word alignment, POS tagging and tree building on the target side. Myanmar ALT with basic POS tags was published on 2017 and will be available in website[1]. POS annotation, which is the base of further syntactic annotation, is an important part of the treebank construction. The annotation is a significant challenge for low-resource Asian languages as Myanmar.

The Myanmar language, also known as Burmese, is overwhelmingly monosyllabic where one or more monosyllabic morphemes can be combined with different levels of strength. Words combined with multiple morphemes include a stem with zero or more affixes to form a meaningful unit. Affixes, which are mainly suffixes play a prominent role in the grammar of Myanmar, because they carry almost all the grammatical information contained in a sentence [2].

In this paper, we describe the procedure that we employed in mapping the original NOVA annotation to the Universal POS [3] annotation which aims to develop cross-language annotation for Myanmar. Table 1 gives an example of modifying a NOVA-annotated Myanmar sentence in ALT Treebank into the Universal POS. We separated two steps for modifying process, 1) trivial mapping, and 2) non-trivial classification where the annotation is updated into more specific sub-categories. Some problematic cases on using the tags of adposition, conjunction, and particles are also examined carefully.

Table 1. The meaning of the sentence is "The third match will be played at Birmingham, on March." The original NOVA annotation and Universal POS tags after modification are listed.

Myanmar	NOVA POS	Universal POS
တတိယ	a	adj
ပွဲ	n	noun
ကို့	o-	adp
မတ်	n	noun[noun
လ	n	noun]noun
တွင်	o-	adp
ဘာမင်ဂန်	n	noun
မှာ	o-	adp
ကစား	v[v	verb[verb
လိမ့်မယ်	o-]v	part]verb
။	.	punct

[1] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/

2. Universal POS Tags

To develop cross-linguistically consistent Treebank annotation with other languages, we require datasets that have been annotation in multiple languages with such a consistent schema. A Universal POS tagsets [3] is suitable for that and it has been used in the induction of syntactic structure of 25 different tagsets of 22 different languages. These tagsets based-on the original Google Universal POS 12 tags: noun (noun), adj (adjectives), verb (verb), adv (adverbs), pron (pronouns), det (determiners and articles), adp (prepositions and postpositions), num (numeral), conj (conjunctions), prt (particles), '.'(punctuation marks) and x (other categories). In the extended version, UD v2, it only used 17 tagsets in the Universal Dependencies[2] (UD) project [4]. When we updating from NOVA to Universal POS on Myanmar ALT data, we used the original 12 tags with two creak-tone generative case-marker "noun-adp" and "pron-adp".

The designation of the former NOVA tags is simplified and modularized version of Universal POS tagsets [5] and it can easily exchangeable to other language-specific tagsets. In the annotation, it provided four basic tags: noun "n", verb "v", adjective "a", and "o" which tag for non-n, non-v and non-a tags for other modifications or complements, to represent fundamental word classes, with three auxiliary tags: "1", "." and "+" to represent numbers, punctuations marks, and tokens with weak syntactic roles and after that modified four tags by attaching "-" minus sign to represent functional word classes: "n-", "a-", "o-", and "/o-" which are aim to get detail information on confusing cases, as in [6] [7]. A list of the Universal POS with NOVA is given in Table 2.

Table 2. Universal POS categories with NOVA

NOVA Tag	Universal POS Tag	Universal POS Description
n	noun	noun
n-	pron	pronoun
a	adj	adjective
a-	det	determiner
v	verb	verb
o	adv	adverb
	adp	adposition
o-	conj	conjunction
	part	particle
/o- n/o-	noun-adp	creak tone genetive
/o- n-/o-	pron-adp	case-marker
1	num	numeral
.	punct	punctuation
+	x	other

3. Trivial Mapping

In this paper, the tags are attached to correspondent tokens by an underline "_" and used the brackets "[ ]" for various linguistic phenomena in all of the example such as compound words, agglutinative suffixes, etc.

$$m[m_1\ m_2\ \ldots\ m_n]m \qquad (1)$$

where, m is an integrated tag and m_n is the tags for each morpheme token. "$m[m_1$" and "$m_n]m$" are single tag for first and last morpheme of the word.

When we mapping from NOVA [4] tagset to Universal POS tags, the following tags have not any significantly changes but we reveal with some example to convey all of the Universal POS tagsets in this paper.

3.1. NOUN "noun"

"noun" tag is applied for all the nominal tokens, including common nouns, proper nouns, and collective nouns. *ရွှေတိဂုံ_noun* (Shwedagon) *ဘုရား_noun* (pagoda), where the first token is proper noun and the last one is a common noun. In Myanmar language, there are four types of noun constructions: original noun, qualitative, verb modification and combination noun. In our annotation, we didn't separate the original noun based on the Myanmar grammar if it was done, words are not meaningful, for example *စပါး _noun* (puddy).

Qualitative noun and verb modification, those are modified from adjective and verb by adding prefix and suffix. "အ" is a very common prefix but neglect segmentation to avoid over-segmenting, as in [4]. The noun transform particles: မှု, ခြင်း, ရာ, ဖွယ် are suffixes to change the word classed in derivation. These derivational affixes are mainly used to form nouns:

အမြင်_noun (sight)

ကောင်း_noun[adj မှု_part]noun (good deed)

3.2. PRONOUN "pron"

The "pron" tag includes personal pronouns, interrogative indefinites, called wh-words, and demonstrative pronouns. ကျွန်တော် ("I" for male), ကျွန်မ ("I' for female), ကျွန်ုပ် ("I" for not only male but also female), သူ ("you" for male), သူမ ("you" for female) are personal pronouns and tagged as (pron).

[2] http://universaldependencies.org/

And then, some interrogative-indefinites are tagged as pronouns: ဘယ် (where), ဘာ (what), မည်သူ (who).

မင်း *ဘာ_pron* ကို ကြိုက် သလဲ ။

(what do you like ?)

But sometimes the annotation lead to confuse with adjective: "မင်း *ဘာ_adj* မုန့်*_noun* ကို ကြိုက် သလဲ ။" (What snack do you like ?). In this example "ဘာ" is adjective which modify the word "မုန့်".

Demonstrative pronouns, can be used in place of noun, that are identical, but demonstrative adjective qualify nouns, whereas demonstrative pronouns stand alone. ဤ, သည်, ထို and ၎င်း are common demonstrative pronouns. For example,

သည်_pron မှာ ထိုင် ပါ ။ (Sit this place)

In the above example, "သည်" is demonstrative pronoun that referred the current place but if we translate "သည် နေရာ"(here) instead of "သည်" , in this case, that "*သည်*" is demonstrative adjective modify the word "*နေရာ_noun*".

3.3. ADJECTIVE "adj"

The "adj" tag is annotated for adjective tokens modifying or describing a noun. Qualitative adjective, numeral adjective and interrogative adjective, those are types of adjective construction in Myanmar language. Morphologically, the stem of "verb" and "adj" are identical but "adj" always used to modify nominal tokens by adding သော, သည့်, မည့်, တဲ့, those are transform particles of suffixed to become adjective, and describe the quality of the noun are called qualitative adjective. For example,

ချမ်းသာ_adj[verb သော_part]adj သူ_noun

(rich man)

Numeral adjective expresses the ordinal number and amount of noun by affixing particles to modify noun and then by affixing it to noun. Some of the numerical classifier are: ကောင်, ခု, ခွန်း, စောင်, ထည်, ပင် and ယောက်. The former annotation of NOVA annotated the counters by "n" tag as *a[1 n]a* which means a number and a noun-derived classifier combine together as an adjective to modify a noun, to annotate analytically and naturally [5]. But in this annotation, we annotated by "part" that is generally according to the instruction of the Myanmar grammar [8] and used several common dictionaries [9] [10].

ငါး_noun *တစ်_adj[num ကောင်_part]adj* (a fish)

Interrogative adjectives are words like interrogative pronouns, but they can't stand on their own that means, they serve to modify another term, specifically a noun.

ဤ_adj အဖွဲ့_noun (this group)

3.4. DETERMINER "det"

It is modify nominal token that comes before a noun, it's also demonstrative adjectives to modify a noun, which specific person, place, or thing is mentioned. Common determiners are "ဤ", "သည်", "ထို" and "ယင်း".

သည်_det ကြော်ငြာ_noun[noun ချက်_part]

(this advertisement)

3.5. VERB "verb"

The "verb" tag is annotated in all the verbal tokens, such as dynamic verbs, static verbs and copula. Myanmar is a subject-object-verb (SOV) language where the root of the verb always come at the end of the sentence-final. Generally, verb root composed of a main verb with other verbal morphemes. For example,

လေ့လာ _verb[verb နေ _part သည်_part]verb

(am studying)

ရောက်_verb[verb လာ_part ခဲ့_part သည်_part]verb

(arrived)

In the above examples, the verbal morpheme "သည်" which indicate the verb ending a sentence, is post positional marker in Myanmar grammar but we annotated that token by "part" to form a consistency scheme with other languages.

3.6. ADVERB "adv":

It includes the word end with "စွာ (-ly)", also known as adverbial affix, as well as degree words like "အလွန် (very)", reduplication word "မြန်မြန် (quick)" and negative markers like "မ (not)".

မြောက်မြား_adj[adj စွာ_part သော_part]adj

(many)

In this case, the word should be adverb if we look only the token "စွာ" but there is another adjective transform particle "သော" includes in the word and it becomes adjective.

3.7. Auxiliary Tags

The other three tags “num”, “punct” and “x” are used trivially to represent number, punctuations marks, and word that cannot be annotated part of speech category. Common Myanmar punctuation marks annotated with “punct” tag is “၊” and “။”. “num” tag annotated not only Arabic numbers but also Myanmar common number and number scripts in this annotation. For example,

တစ်_num	(one)	*၁_num*	(1)
တစ်ဆယ်_num	(ten)	*၁၀_num*	(10)
တစ်ရာ_num	(hundred)	*၁၀၀_num*	(100)
တစ်ထောင်_num	(thousand)	*၁၀၀၀_num*	(1000)

4. Non-trivial Modification

Most of the Southeast Asian languages are highly analytic in morphology and overwhelmingly use free morphemes. Myanmar is one of them but, syntactically, it is a strong head-final language. In NOVA annotation, all functional morphemes were annotated “o-” and this section will be discussed with non-trivial classification where the annotation is updated into more specific sub-categories from “o-” to adposition “adp”, particle “part”, and conjunction “conj”, those are easily confused to annotate the token and they are depend on the context of the sentence.

4.1. ADPOSITION “adp”

This tag should be used for word suffixed of noun and pronoun to designate the subject or object.

သူ_pron[pron တို့_part]pron *သည်_adp* (they)

မြို့_noun *၌_adp* (in city)

In this example, the token “သည်” indicate the nominative case and “၌” expresses the location of condition. The most common “adp” in our corpus are listed in Table 3.

Table 3. Most common “adp” token. We selected over 10k frequency count.

Adp	Frequency Count
ကို	18397
၏	12498
က	12374
သည်	12191
တွင်	10371

4.1.1. Creak Tone Genitive Case-marker

In Myanmar language, high creaky tone frequently signals genitive case, which is otherwise realized as a discrete morpheme “ရဲ့” pronounced with high creaky tone. For example, “*ကျွန်တော့်_pron-adp အိမ်_noun*” (my house), in this phrase “ကျွန်တော့်” is already showing possession and inclusion of “ရဲ့” is redundant. “*အများပြည်သူ့_noun-adp အသံ_noun*” (public’s voice), in this example, noun token with a contracted genitive case-marker.

4.2. CONJUNCTION “conj”

This tag connects either other words or sentence or meaning reciprocally. “နှင့်” (with/and) and “သို့မဟုတ်” (either or/else), where they are connected words and followed nouns and pronouns such as things, objects and people. Specific example is follow:

လူ_noun[noun မိုက်_adj]noun *နှင့်_conj* ငွေ_noun

(A fool and his money)

“၍” (and/because), “လျှင်” (if), “သောကြောင့်” (because/as a result) and “သော်လည်း” (although) are conjunctions token that connect sentences.

Table 4. List of the most common “conj”

Adp	Frequency Count
နှင့်	2489
ရန်	1706
၍	1022
သော်လည်း	933
သို့မဟုတ်	703

4.3. PARTICLE “part”

The “part” tag includes a number of mostly monosyllabic words and that may be confused with “adp” and “conj”. Some of them are: “ကို”, “နှင့်” and “ရန်”. The numerical classifier are “part” which are used when counting or measuring nouns.

For example,

“*လေး_adj[num ယောက်_part]adj*” (four people)

where, “ယောက်” is numeral classifier for counting person.

"များ" and "တို့" suffixes are also particles and used to express as the plural noun. We used brackets for a nominal constituent with a plural suffix [5]. For specific example are as follow:

ကလေး_noun[noun များ_part]noun (children)

သူ_pron[pron တို့_part]pron (they)

Table 5. Common "part" tag in ALT corpus

part	Frequency Count
ခဲ့	22890
သည်	21716
များ	18574
သော	13236
မှ	11311

4.4. Problematic Cases

In this section, we discusses some problematic cases on using the tags of adposition, particle, and conjunction are also examined carefully. Those are easily confused and gives guidelines on how to decide such case. It is also shown the common usages of words with examples.

4.4.1. "ကတည်းက"

Both adpositions and conjunctions occur with this word "ကတည်းက" and are often difficult to distinguish from one another. In our annotation, we annotated "conj" if that word is suffixed of the verb and connective with other sentence, and also, when that word is suffixed of nouns to indicate past occurrence ("since" and "from the time of" are equivalent in English preposition), annotated by "adp". Specific examples are as follow:

သဘောပေါက်_verb ပြီး_part *ကတည်းက_conj*

(realize)

where, "ပြီး" that token indicate the completion of an act and "ကတည်းက" comes after verb, just like "conj".

These are some of the usage of "ကတည်းက _adp":

၂၀၁၈ ခုနှစ် *ကတည်းက_adp* , since 2018

စောစော *ကတည်းက_adp* , since early

မနေ့ *ကတည်းက_adp* , since yesterday

4.4.2. "ကို"

Sometimes, it is unclear where "ကို" is "adp" or "part". In general, it should be tagged as an "adp" if it's indicate objective case.

For example,

သူ့ *ကို_adp* သွား ပြော (Go and tell him)

သူ က သိပ် *ကို_adp* တော် တာ ပဲ (He's incredibly skilled)

In this example, the first suffix "ကို" indicate in front of that is the object. Another one in second example setence that used to emphasize words meaning "very" and similar.

4.4.3. "နှင့်"

While the word has been found in "adp", "conj" and "part", it is often very difficult to annotate. If the word suffixed of a noun to indicate the instrumental case (equivalent "with" that is adverb in English usage), that is annotated by "adp".

For example,

ကတ်ကြေး_noun *နှင့်_adp* အပိုင်း လေး တွေ ဖြတ် ပါ

(please cut this small piece with scissors)

where, "နှင့်" is connective affix of noun.

It is also suffixed to a verb and collocating with prohibitory, "မ" being prefixed to the verb to convey and imperative sense as:

မ_adv သွား_verb *နှင့်_part* (not go)

Be aware that in case, "မ (not)" is the negative particle in Myanmar grammar, but it is annotated by "adv" in this annotation scheme for the cross-language consistent.

Prior-past, or prior-future, according to the connection as annotated as "part". For example,

သွား_verb *နှင့်_part* ပြီ_part (he had gone)

စား_verb *နှင့်_part* ပါ_part (go ahead first and eat)

4.4.4. "မှာ"

That word, though usually equivalent to English locative prepositions such as 'at', 'on', 'in', 'under', 'by', etc., is suffixed to a verb to denote genitive case.

သီရိလင်္ကာ_noun *မှာ _adp* ရေတွက်_verb *မှာ_part*

(being counted in Sri Lanka)

where, the first token "မှာ" indicate the locative and the last one is affixed of verb.

4.4.5. "သို့"

In this case, "သို့" is an objective affix, denoting the object towards which or unto which motion is directed.

For example,

ကျောင်း_noun သို့_adp သွားတော့ (go to the school!)

But in colloquial style, it is used by "ကို" instead of "သို့" or sometimes it may be dismissed, as above. The other cases may seem to require the aid of various locative prepositions in English to indicate direction, as:

မြောက် မှ တောင်_noun သို့_adp စီးဆင်း သည်
(flows from North to South)

အလို_noun သို့_adp လိုက်သည်
(to follow (his) will, or according to (his) will)

အိမ်_noun သို့_adp ရောက်သည်
(to arrive at home)

"သို့ (such)", it also follows pronominal adjectives, that are determiner in this guideline, by suffixing with the usual connectives: ဤသို့, သည်သို့ (this) and ထိုသို့, ယင်းသို့ (that).

4.4.6. "သည်"

That word can be annotated "adp" for word indicating nominative case and "part" for word indicating the verb ending a sentence.
For example,

သူ_pron သည်_adp (he)

ဖြစ်_verb[verb သည်_part]verb (to-be)

In the above example, the first token "သည်" shows that the pronoun is subject and the second one is sentence marker.

Be aware that, "သည်" of the end of a sentence is post positional marker in Myanmar grammar but this guideline annotated that case to "part" and other sentence marker: "၏", "ပြီ", "မည်", "တယ်", "စို့", "လိမ့်မည်" are also annotated by "part".

5. Statistical Analysis

The ability of former NOVA annotation system is a feasible and flexible annotation system, and it is also designated as an interface annotation from raw data to a relatively informative level, and it is easy to switch other system just, like Universal POS annotation system, for under-resource languages. Most of the basic and modified nova tags are covered by the identical tag in Universal POS tags except in "adp", "conj", and "part". And we have already explained the modifying process by two parts: trivial mapping and non-trivial mapping in section 3 and 4. This section discuss with statistical analysis between two annotation systems.

Table 6 shows the overall percentage of the non-trivial mapping between NOVA and Universal POS annotation. Most of the percentage in our data are "part" because it is not only serve as boundary units identifying the terminus of conceptual units within sentences but also include as morpheme tokens in the compose of words.

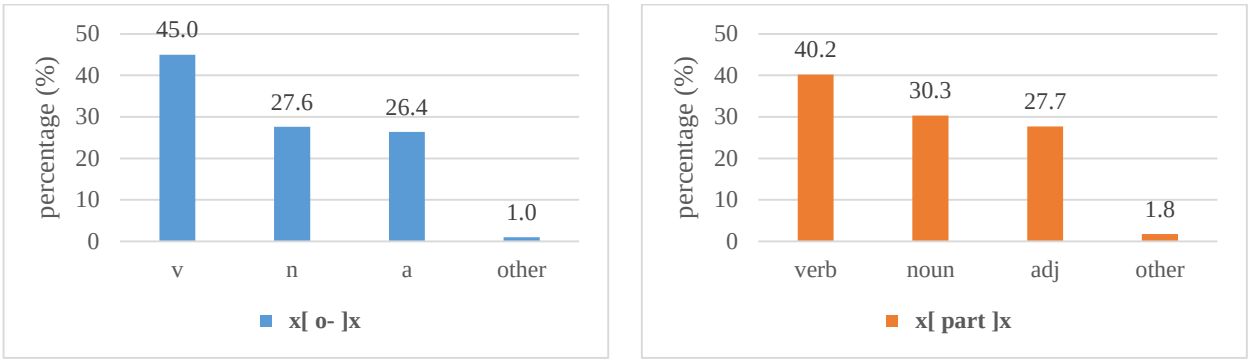
Table 6. Count and percentage of non-trivial mapping NOVA "o-" to Universal POS

Universal POS		
tag	count	%
part	195,839	64.95
adp	93,715	31.09
conj	11,927	3.96
Total	301,481	100.00

Generally, compounds in Myanmar have various patterns, in there, nominal and verbal morphemes can be combined freely. There are four main classifier morphemes: noun classifier, numeral classifier, verbal classifier and genitive classifier, among them, most of the morpheme are particles which has been used the brackets to emphasize the integration of a nominal and verbal constituent in our annotation.

Figure 1 presented for the involvement percentage of the particle in compounds, which express by using brackets "x[" and "]x" as root tag. Where, we listed the statistic data for NOVA "o-" and Universal POS "part" be contained in the root tag of "adj", "noun", "verb" and other tags in Figure 1 (a) and (b), respectively. More than one time of particle may be comprised in one root tag. In the figure, although the percentage of the "adj" and "noun" are slightly increased, the integration of particle morphemes in "verb" decrease after modifying annotation.

Table 7 indicates each tag's relative frequency in the full annotated data of Universal POS compare with former NOVA tags, and our illustration statistic is listed over 1.0% frequency. The percentage between "noun" has significant change because we annotated some word by using bracket. For example, ဇန်နဝါရီ လ (January), where the last token means (month) in English, annotation in NOVA is "n" for both of them but in this modification, we transformed such kind of case to "noun[noun noun]noun" because the second nominal morpheme refer to first token is month. The most common supportive nominal morpheme in our data are ဇနေ့ (day), မြို့ (city), ပြည်နယ် (state), နိုင်ငံ (country) and လူမျိုး (nation).



(a) **(b)**

Figure 1. Illustration of "o-" and "part" involvement percentage in compound words, where "x[" and "]x" represent root tag, and more than one time of "o-" and "part" may be involved in there[3].

Table 7. Comparison of each tag's relative frequency over 1.0% between NOVA and Universal POS

NOVA tag	%	Universal POS tag	%
n	30.9	noun	24.4
o-	27.1	adp	18.5
.	6.4	part	7.2
v	4.2	punct	9.7
n[v o-]n	3.1	verb	3.4
n[n o-]n	2.9	noun[verb part]noun	3.0
a[1 n]a	2.3	noun[noun part]noun	2.9
v[v o- o-]v	2.3	conj	2.5
v[v o-]v	1.8	verb[verb part part]verb	2.5
o	1.8	adj[num part]adj	2.1
n-	1.7	noun[noun noun]noun	2.0
1	1.6	verb[verb part]verb	1.9
a[v o-]a	1.5	adv	1.8
		pron	1.7
		adj[verb part]adj	1.5
		adj	1.2
		num	1.0
Total	87.6		87.3

The frequency of bigrams between 1st word tags: "adj", "adv", "det", "noun", "num", "pron" and "verb", and 2nd word tags: non-trivial mapping tags in an ALT corpus are presented in Table 8. It shown the affiliation of seven tags with "adp", "conj", and "part" in sentences. Among them, 82.9% of "adp" followed on after "noun", that is reasonable, because adpositions is a suffix of nominal morpheme to designate the subject or object. Continuously, 61.9% of "conj" and 64.5% of "part" comes after "verb". In ALT data, almost two parts of three are long sentences and "conj" connected two sentences after the first sentence of "verb".

Table 8. Distribution of bigrams (percentage - %)

1st word tag \ 2nd word tag	adp	conj	part
adj	0.7	1.3	2.5
adv	0.9	1.6	0.5
det	0.9	0.1	0
noun	**82.9**	32.3	19
num	1.8	1.6	9
pron	11.7	0.7	4.3
verb	0.7	**61.9**	**64.5**

6. Conclusion

This morphological annotation paper is intended for modifying NOVA-annotated Myanmar data to Universal POS tagset which aims to develop cross-language consistent annotation for Myanmar with other languages. And we used the original categories 12 tags of Google Universal POS. This annotation is more detail than the former annotation NOVA and we also illustrated some problematic case with example and expressed how we decided about that. And also, we presented statistical analysis between two annotation system. Our annotation is the base for further syntax annotation which is important for Asian Language Treebank (ALT) and it conveys morpheme annotation for Myanmar Language.

Acknowledgements

We offer grateful thanks to the anonymous reviewers for their valuable comments and writers who published the "Myanmar Grammar" and "Myanmar-English Dictionary".

[3] For example: ကလေး_n[n များ_o-]n in NOVA and ကစား_verb[verb နေ_part ကြ_part သည်_part]verb in Universal POS.
(children) (are playing)

References

[1] H. Riza, M. Purwoadi, T. Uliniansyah, Aw Ai T., S. M. Aljunied, L. Chi Mai, V. T. Thang, N. P. Thai, R. Sun, Vichet C., Khin M. S, K. Thandar N., M. Utiyama, C. Ding. "Introduction of the Asian Lanugage Treebank". 2016 O-COCOSDA, 26-28 October, Bali, Indonesia.

[2] John Okell and Anna Allott, "Burmese/Myanmar Dictionary of Grammatical Forms", 2009.

[3] S. Petrov, D. Das, R. McDonald. "A Universal Part-of-Speech Tagset".

[4] M. de Marneffe, T. Dozat, N. Silveria, K. Haverinen, F. Ginter, J. Nivre, C. D. Manning. "Universal dependencies: a cross-linguistic typology". In proceedings of LREC. 2014.

[5] C. Ding, M. Utiyama, E. Sumita. "NOVA: A Feasible and Flexible Annotation System for Joint Tokenization and Part-of-Speech Tagging", ACM Trans. Asian Low-Resour. Lang. Inf. Process, Vol. 18, No.2, Article 17, December 2018.

[6] C. Ding, Hnin .T.Z.A, M. Utiyama, Win P.P, E. Sumita. "Tokenization and Part-of-Speech Annotation Guidelines for Myanmar (Burmese)", Version 0.1, November 2016.
http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/Myanmar-annotation-guideline.pdf

[7] C. Ding, Hnin .T.Z.A, M. Utiyama, Win P.P, E. Sumita. "Supplementary Instructions for Tokenization and Part-of-Speech Annotation Guidelines for Myanmar (Burmese)". Version 0.2, March 2017.
http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/Myanmar-annotation-guideline-supplemantary.pdf

[8] "Myanmar Grammar". Department of the Myanmar Language Commission, Ministry of Education, The Republic of the Union of Myanmar, Edition: 2011 and 2016.

[9] "Myanmar-English Dictionary". Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar. Edition: 2014.