Tokenization and Part-of-Speech Annotation Guidelines for Myanmar (Burmese)

(Version 0.2, November 2016)

Chenchen Ding[1], Hnin Thu Zar Aye[1, 2], Masao Utiyama[1], Win Pa Pa[2], Eiichiro Sumita[1]

[1]Advanced Translation Technology Laboratory, ASTREC, NICT, Japan

[2]University of Computer Studies, Yangon, Myanmar

chenchen.ding@nict.go.jp

1. Introduction

The Myanmar (Burmese) language is the official language of the Republic of the Union of Myanmar. The language is a member of the Lolo-Burmese grouping of the Sino-Tibetan language family. Morphologically, Myanmar is highly analytic with no inflection of morphemes. Morphemes can be combined freely with no changes. Syntactically, Myanmar is typically head-final where the functional dependent morphemes succeeding content independent morphemes and the verb constituent working as the root of a sentence always comes at the end of a sentence.

This manual provides detailed guidelines for the surface annotation of the Myanmar texts in Asian language treebank (ALT). The tokenization and part-of-speech (POS) annotation is included in this manual and handled uniformly under an annotation system called NOVA. The manual is organized as follows. Section 2 is the introduction of the NOVA system used for annotation. Section 3 describes the principles of tokenization. Section 4 describes the details in POS annotation, specifically, annotation for simple words, for affixes, and for compounds are described respectively. Section 5 provides descriptions on specific and difficult cases in annotation.

2. NOVA Annotation

NOVA provides four basic tags: "n", "v", "a", and "o" to represent fundamental word classes, with further three auxiliary tags to represent numbers, punctuations marks, and tokens with weak syntactic roles. Specific descriptions of the basic and auxiliary tags are listed in Table 1. Besides the simple tags, a pair of brackets "[" and "]" are further applied to show multiple tags "working (together) as". The brackets are used widely in the annotation to represent various linguistic phenomena, e.g., compound, agglutinative suffix, etc.

Table 1. Basic and auxiliary tags in NOVA

tag	description
n	general nouns, can be subjects or objects of tokens tagged by **v**
v	general verbs, can take tokens tagged by **n** as arguments
a	general adjectives, can directly describe or modify tokens tagged by **n**
o	other modifications or complements for tokens or larger syntactic parts
1	general numbers
.	general punctuation marks
+	a catch-all category, for tokens with weak syntactic roles

3. Tokenization

The Myanmar texts are split into morphemes in principle, i.e., the text are split into meaningful tokens as small as possible. Stable multi-morpheme expressions may be remained unsegmented, which is decided by native-speaker annotators' common-sense and entries listed in dictionaries. As Myanmar is highly analytic, most single morphemes can be treated as words directly. For complicated expressions applying more than one morpheme for an entire concept, the brackets are applied to address the integration. The complicated cases mainly include various compounds and agglutinative suffixes added to stem morphemes, which will be described in the following sections together with POS annotation in details. Generally, the annotation with brackets is in a form of "$x[x_1\ x_2\ \ldots\ x_n]x$", where x_k are the tags for each component morpheme and x is the tag for the integrated expression. Here the "$x[x_1$" and "$x_n]x$" are single tags for the initial and final morphemes in the expression. The usage of brackets are restricted to be shallow, that crossed or nested brackets are avoided.

4. Part-of-Speech Annotation

4.1. Word

This sub-section provides descriptions and examples for the usage of basic and auxiliary tags. The usage of brackets for adapting complicated cases are provided in the following sub-sections. For all the examples illustrated from now on, the tags are attached to correspondent tokens by an underline ("_").

The “n” tag is applied for all the nominal tokens, including common nouns and proper nouns. Various pronouns are also taken as nominal tokens and annotated by “n”. Specific examples of the “n” tag are listed in Table 2.

Table 2. Examples of tokens annotated with the “n” tag.

annotated	Myanmar	English gloss	note
ရလဒ်_n		result	common noun
အလူမီနီယမ်_n		aluminum	common noun
ကိုရီးယား_n		Korean	proper noun (country)
ဖီဖာ_n		FIFA	proper noun (organization)
သူ_n		he	pronoun
ကျွန်တော်_n		I	pronoun

The “v” tag is applied for all the verbal tokens, including common (dynamic) verbs, stative verbs, and copula. The “a” tag is applied for adjective tokens modifying or describing a noun. Because the verbal and adjectival stems in Myanmar share an identical set of functional suffixes in morphology, the adjective can be analyzed as stative verbs and thus is not a necessary word class in Myanmar. Strictly, the “v” tag can be applied on all the verbal and adjectival stems, while the use of “v” and “a” in practice is dependent on the sense of how “dynamic” the verbal stem is. For the determiners, the “a” tag is applied as they are always used to modify nominal tokens. Specific examples of the “v” and “a” tags are listed in Table 3.

Table 3. Examples of tokens annotated with the “v” and “a” tags.

annotated	Myanmar	English gloss	note
သွား_v		to-go	common verb stem
စား_v		to-eat	common verb stem

annotated Myanmar	English gloss	note
ဖြစ်_v	to-be	copula
ကောင်း_a	to-be-good	stative verb (adjective) stem
လှ_a	to-be-beautiful	stative verb (adjective) stem
ဤ_a	this	determiner (demonstrative)
မည်သည်_a	which	determiner (interrogative)

The "o" tag is applied for all the functional tokens, including conjunction, particles, various post-positional nominal, verbal, phrasal, and sentential functional suffixes, and a few prefixes. The adverb in Myanmar can be analyzed as nominal tokens and thus is not a necessary word class as adjectives. In practice, adverbs can be annotated by "o" or "n", based on the annotator's sense. Generally, the "o" tag can be applied for any ambiguous tokens with a certain syntactic role (or the "+" tag will be applied). Specific examples of the "o" tag are listed in Table 4.

Table 4. Examples of tokens annotated with the "o" tag.

annotated Myanmar	English gloss	note
သိပ်_o	very	common adverb
နှင့်_o	and	conjunction
သော်လည်း_o	however	conjunction
ကို_o	n/a	accusative case-marker (nominal)
သော_o	n/a	attributive marker (verbal)
လည်း_o	also	particle (phrasal)

annotated Myanmar	English gloss	note
၏_○	n/a	affirmative marker (sentential)

The three auxiliary tags "1", "．", and "+" are used trivially to represent numbers, punctuations marks, and tokens with weak syntactic roles, e.g., interjections. Specific examples of the three tags are listed in Table 5.

Table 5. Examples of tokens annotated with "1", "．", and "+" tags.

annotated Myanmar	English gloss	note
တစ်_1	one	common number
၃၀_1	30	Myanmar number scripts
။_．	.	Myanmar period mark
ဝိုး_+	wow	interjection for amazing

4.2. Agglutinative Affix

According to the previous descriptions, Myanmar functional suffixes (and less used prefixes) are segmented as tokens and annotated by the "○" tag. The bracket annotation is further used for those suffixes having a relatively strong cohesion with the preceding stem.

In general, nominal suffixes, which mainly serve as case-markers, are simply segmented without using the brackets. The only exception is several plural markers, for which brackets are used to show the integration with the stem; verbal suffixes, which add various tense and aspect information to stems, are wrapped with bracket to emphases the integration of a verbal constituent; suffixes for larger syntactic constituent as phrasal and sentential ones, are treated as completely independent tokens, where no brackets are applied. Specific examples on the usage of brackets around agglutinative affixes are listed in Table 6.

Table 6. Examples on the usage of brackets around agglutinative affixes.

annotated Myanmar	English gloss / English trans.	note
သူမ_n သည်_o	she n/a / “she”	“သည်” is a nominative case-marker
သူမ_n ကို_o	she to / “to her”	“ကို” is a dative/accusative case-marker
သူမ_n ၏_o	she of / “of her”	“၏” is a genitive case-marker
ကလေး_n[n များ_o]n	child -s / “children”	“များ” is a plural marker, brackets used
သွား_v[v သည်_o]v	to-go n/a / “go (finite)”	“သည်” is a suffix for sentence ending
သွား_v[v ကြ_o သည်_o]v	to-go n/a / “go (finite)”	“ကြ” is a suffix for plural subject
မ_v[o သွား_v ဘူး_o]v	not to-go not / “don’t go (finite)”	“မ…ဘူး” is a circumfix for negation

As demonstrated, the brackets may wrap more than two tokens if multiple affixes are used. When the composition of a verbal constituent become complex, the brackets may cover a relative long range. Table 7 illustrates an example, with one token per row.

Table 7. An example of a complex verbal expression, with a meaning of “were killed”.

annotated Myanmar	English gloss	note
သတ်_v[v	to-kill	
ဖြတ်_v	to-cut	a compound verb, mentioned in Sec. 4.4
ခံ_v	to-receive	a common verb, used to form passive voice
ခဲ့_o	were (approx.)	a suffix to emphasize past happening
ကြ_o		a suffix for plural subject

annotated Myanmar	English gloss	note
ရ_ဝ	have (approx.)	a suffix to emphasize occurrence
သည်_ဝ]v	n/a	a suffix for sentence ending

4.3. Derivational Affix

Besides the agglutinative affixes (most suffixes) to add further syntactic information (e.g., case, number, tense, aspect, voice, mood and so on) to stems from specific word class, there are also affixes to change the word classes in derivation. These derivational affixes are mainly used to form nouns, i.e., used for nominalization. Brackets annotation is generally applied on these affixes to wrap a derived nominal expression. Specific examples on the usage of brackets around agglutinative affixes are listed in Table 8.

Table 8. Examples on the usage of brackets around derivational affixes.

annotated Myanmar	English gloss / English trans.	note
ကြိုးစား_n[v မှု_ဝ]n	to-try n/a / “attempt”	“မှု” is a suffix to change verbs to nouns
သင်ကြား_n[v ခြင်း _ဝ]n	to-teach n/a / “teaching”	“ခြင်း” is a suffix to change verbs to nouns
အားနည်း_n[a ချက်_ဝ]n	weak n/a / “weak point”	“ချက်” is a suffix to change adjective to nouns

There is an important and frequent prefix “အ” to form nouns in Myanmar. This prefix is generally kept unsegmented, which will be described in detail in Sec. 5.1.

4.4. Compound

As the combination of simple nominal and verbal morphemes is relatively free in Myanmar, there are various patterns of compounds. Generally, one compound will be segmented and brackets are applied to wrap, if the meaning of each component within the compound is clear retained and related to the meaning of the entire compound. Specific examples of nominal, verbal, and adjectival compounds are listed in Table 9.

Table 9. Examples of nominal, verbal, and adjectival compounds.

annotated Myanmar	English gloss	English translation
ရေ_n[n ပုလင်း_n]n	water bottle	“water bottle”
သောက်_n[v ရေ_n]n	to-drink water	“drinking-water”
စား_n[v သောက်_v ဆိုင်_n]n	to-eat to-drink shop	“restaurant”
လက်_n[n ပတ်_v နာရီ_n]n	hand to-wrap clock	“wrist watch”
စိုက်_v[v ပျိုး_v]v	to-plant to-sew	“to cultivate”
အကြံ_v[n ပေး_v]v	advice to-give	“to suggest”
လေး_a[a နက်_a]a	to-be-heavy to-be-deep	“to be profound”
ရုပ်_a[n ဖြောင့်_a]a	appearance to-be-straight	“to be handsome”

As mentioned in the section of tokenization, the nested brackets are avoided in the annotation. If the compound is further modified by affixes, only the most outside brackets are reserved, as illustrated by the example in Table 7.

5. Specific Examples

5.1. Annotation Around Prefix “အ”

“အ” is a very common prefix used in Myanmar to address the “nominality” of an expression. The prefix can be used freely for verbal or adjectival stem to transform them into nouns, and it also appears at the head of many basic nouns. In practice, the “အ” prefix is

kept together with the following stems to avoid over-segmenting. One reason is the segmentation of "အ" may generate too small tokens even for basic nouns. Another reason is the meaning of the left component in some nouns has become vague and is not directly related to the entire meaning. Specific examples of nominal tokens with the "အ" prefix are listed in Table 10.

Table 10. Examples of nominal tokens with the "အ" prefix

annotated Myanmar	English gloss	note
အစား_n	eating	"စား" is a verb, meaning "to eat"
အလှ_n	beauty	"လှ" is an adjective, meaning "beautiful"
အချိန်_n	time	the meaning of "ချိန်" is vaguely related to "time"
အလိပ်_n	roll	the meaning of "လိပ်" is vaguely related to "roll"
အရောင်_n	color	"ရောင်" means "color" by itself

The "အ" prefix is also applied in specific syntactic structures, such as passive voice for verbs and superlative form for adjectives. It is only segmented in the superlative form for adjectives, where it is actually used in a circumfix form as "အ…ဆုံး". Examples is illustrated in Table 11.

Table 11. Example of the "အ…ဆုံး" in superlative adjective.

annotated Myanmar	English gloss	note
အလှ_n	beauty	derived from adjective "လှ" ("beautiful")
အမြင့်_n	altitude	derived from adjective "မြင့်" ("high")

annotated Myanmar	English gloss	note
အ_a[o လှ_a ဆုံး_o]a	n/a beautiful -est	“most beautiful”, from “လှ”, not “အလှ”
အ_a[o မြင့်_a ဆုံး_o]a	n/a high -est	“highest”, from “မြင့်”, not “အမြင့်”

A further related issue is around compounds with “အ”-leading nouns, where the prefix may be dropped in combination. These compounds with an incomplete “အ”-leading nouns are thus kept together in practice, to avoid over-segmenting. Specific examples of such nouns are listed in Table 12.

Table 12. Examples of compounds with “အ”-leading nouns

annotated Myanmar	English gloss	note
ရပ်နားချိန်_n	break-time	from “ရပ်နား” (“to-stop”) and “အချိန်” (“time”)
စက္ကူလိပ်_n	paper-roll	from “စက္ကူ” (“paper”) and “အလိပ်” (“roll”)
လိမ္မော်ရောင်_n	orange-color	from “လိမ္မော်” (“orange”) and “အရောင်” (“color”)

5.2. “n” or “a” and “v” or “o”

The annotation around functionalized or grammaticalized tokens should depend on the contexts, i.e., the specific role they played. Typical cases are nominal pronouns used as determiners, where “a” should be used instead of “n”, and verbs used as auxiliary roles, where “o” should be used instead of “v”. Specific examples are listed in Table 13.

Table 13. Examples of context-dependent annotation.

annotated Myanmar	English gloss / English trans.	note
ထို_n ကဲ့သို့_o	that as / “as that”	“ထို” is followed by a suffix, so “n” used

annotated Myanmar	English gloss / English trans.	note
ထို_a အရာ_n	that thing / “that thing”	“ထို” is followed by a noun, so “a” used
သွား_v[v ခဲ့_o သည်_o]v	go already n/a / “went”	“သွား” is the verbal stem, so “v[v” used
လုပ်_v[v သွား_o မည်_o]v	do will n/a / “will do”	“သွား” is modifying “လုပ်”, so “o” used

5.3. Brackets for Post-Positional Adjectives

Although Myanmar is a typical head-final language, many simple adjectival stems can be directly added after the nominal expressions they modify to form a relatively “loose nominal compound”. Brackets are used for this phenomenon, just as used for a common compound. Specific examples are listed in Table 14.

Table 14. Examples for brackets used for post-positional adjectives.

annotated Myanmar	English gloss	English translation
တော_n[n ရိုင်း_a]n	forest wild	“wild forest”
ရေ_n[n အေး_a]n	water cool	“cool water”

5.4. Brackets for Counters

The numbers are generally segmented to separate token. For the number-counter constituents for counting nouns, brackets are used. The number-counter constituents are treated as adjectival expressions because of modifying nouns, and the counters are generally annotated by “n”, as most of them are derived from grammaticalized nouns. Specific examples on annotating numbers and counters are listed in Table 15.

Table 15. Examples on annotating numbers and counters

annotated Myanmar	English gloss	English translation
လူ_n ငါး_a[1 ယောက်_n]a	person five n/a	“five people”

annotated Myanmar	English gloss	English translation
စာအုပ်_n တစ်_a[1 အုပ်_n]a	book one n/a	“one book”