Tokenization and Part-of-Speech Annotation Guidelines for Khmer (Cambodian)

(Version 0.2, December 2016)

Chenchen Ding[1], Hour Kaing[1, 2], Masao Utiyama[1], Vichet Chea[2], Eiichiro Sumita[1]

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

[2]National Institute of Posts, Telecommunications and ICT, Cambodia

chenchen.ding@nict.go.jp

1. Introduction

The Khmer (Cambodian) language is the official language of the Kingdom of Cambodia. The language is a member of the Austroasiatic language family. Morphologically, Khmer is an analytic, isolating language. Morphemes can be combined freely with no changes, and particles and auxiliary words are used to indicate grammatical relationships. Syntactically, Khmer is typically head-initial where modifiers follow the word they modify. General word order is subject–verb–object and topic-comment structure is common in Khmer.

This manual provides detailed guidelines for the surface annotation of the Khmer 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 annotating single tokens and Section 5 describes the annotation for compounds. Section 6 provides descriptions on confusing 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, mainly for compounds in the case of Khmer.

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

No word-separators are used in Khmer texts to show the word boundary. This section describes the principles used to segment Khmer texts into tokens.

Tokens are classified into two types, (1) word and (2) compound in annotation. A word is a token which cannot be further segmented without losing the meaning of its own. As Khmer is highly analytic, the words are equal to morphemes in most cases. A compound is then a unit with integrated meaning composed by two or more words. In tokenization, words, and components with in compounds are separated by spaces, i.e., Khmer texts are generally tokenized into morpheme-level. In practice, the Chuon Nath Khmer dictionary is referred to in tokenizing compounds. For example, “តេជគុណ”, which means “Your Excellency”, is considered as a compound and segmented into “តេជ” and “គុណ”, which means “power” and “kindness”, respectively, because both “តេជ”and “គុណ” are meaningful entries listed in Chuon Nath Khmer Dictionary. A further tokenization example is illustrated in the following Example 1, where “សាលារៀន” is a compound composed by two words “សាលា” and “រៀន”.

- Example 1

Tokenized Khmer:	នេះ	ជា	សាលា	រៀន	ខ្ញុំ	។
English gloss:	this	to-be	school	to-study	my	.
English translation:	This is my school.					

The compounds are segmented into tokens and will be finally annotated by the brackets in NOVA. Generally, the annotation with brackets is in a form of “$\mathtt{x[x_1\ x_2}$ … $\mathtt{x_n]x}$”, where $\mathtt{x_k}$ are the tags for each component morpheme and $\mathtt{x}$ is the tag for the integrated expression. Here the “$\mathtt{x[x_1}$” and “$\mathtt{x_n]x}$” 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. The details on identification and annotation of compounds will be introduced in Section 5.

4. Part-of-Speech Annotation for Single Tokens

4.1. Usage of "n" Tag

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 are illustrated in the following. For all the examples in this manual, the tags are attached to correspondent tokens by an underline ("_").

- Example 2

Annotated Khmer:	កម្ពុជា_n	សម្បូរ_v	ប្រាសាទ_n	ណាស់_o	។_.
English gloss:	Cambodia	to-have-plenty-of	temple	very	.
English translation:	Cambodia has many temples.				
note:	"កម្ពុជា" is a proper noun; "ប្រាសាទ" is a common noun.				

- Example 3

Annotated Khmer:	តើ	ឯង	ចូល	ចិត្ត	មុខ	វិជ្ជា	អ្វី	?
	_o	_n	_v[v	_n]v	_n[n	_n]n	_a	_.
English gloss:	n/a	you	to-enter	heart	head	subject	what	?
English translation:	What subject do you like?							
note:	"ឯង" is a personal pronoun. Common personal pronouns are: "ខ្ញុំ", "ឯង", "វា", "គាត់", and "គេ".							

- Example 4

Annotated Khmer:	នេះ_n	ជា_v	ចំណែក_n	ឯង_a	។_.
English gloss:	this	to-be	part	your	.
English translation:	This is yours.				
note:	"នេះ" is a demonstrative pronoun. Common demonstrative pronouns are: នេះ", "នោះ", "ទាំងនេះ", and "ទាំងនោះ".				

- Example 5

Annotated Khmer:	តើ	អ្វី	មាន	តំលៃ	ថ្លៃ	ជាង	គេ	?
	_o	_n	_v	_n	_a	_o	_n	_.
English gloss:	n/a	what	to-have	cost	dear	more-than	other	?

English translation: What is the most expensive one?

note: "អ្វី" is an interrogative pronoun. Common interrogative pronouns are: "អ្នកណា", "នរណា", "អ្វី", "ណា", and "ណាខ្លះ".

- Example 6

Annotated Khmer:	កល្យាណ	មិត្ត	ដែល	ខ្ញុំ	ស្រលាញ់	ជាង	គេ
	_n[n	_n]n	_n	_n	_v	_o	_n
English gloss:	virtuous	friend	who	I	to-like	more-than	other

English translation: intimate friend who I like the most

note: "ដែល" is a relative pronoun standing for the antecedent noun.

4.2. Usage of "v" Tag

The "v" tag is applied for all the verbal tokens, including general verbs and copula. Specific examples are illustrated in the following.

- Example 7

Annotated Khmer:	ខ្ញុំ_n	រៀន_v	ភាសា_n[n	ខ្មែរ_n]n	។_.
English gloss:	I	to-learn	language	Khmer	.

English translation: I learn the Khmer language.

note: "រៀន" is a transitive verb.

- Example 8

Annotated Khmer:	កូន_n	ដេក_v	ក្នុង_o	អង្រឹង_n	។_.
English gloss:	child	to-sleep	in	hammock	.

English translation: The child sleeps in a hammock.

note: "ដេក" is an intransitive verb.

- Example 9

Annotated Khmer:	ខ្ញុំ_n	ជា_v	ប៉ូលីស_n	។_.
English gloss:	I	to-be	police	.
English translation:	I am a police.			
note:	“ជា” is a copula. Another common copula is “គឺ”.			

4.3. Usage of “a” Tag

The “a” tag is applied for adjective tokens modifying or describing a noun. Specific examples are illustrated in the following.

- Example 10

Annotated Khmer:	កង់ _n	តូច _a	ជិះ _v	មិន _o	បាន _v	លឿន _o	ទេ _o	។ _.
English gloss:	bicycle	small	to-ride	not	can	fast	n/a	.
English translation:	Riding a small bicycle cannot be fast.							
note:	“តូច” is a common adjective.							

- Example 11

Annotated Khmer:	ផ្ទះ _n	ខ្ញុំ _a	សង់ _v	នៅ _o[o	កណ្ដាល _o]o	វាល _n[n	ស្រែ _n]n	។ _.
English gloss:	house	my	to-build	in	middle	field	paddy	.
English translation:	My house is built in the middle of the rice field.							
note:	“គាត់” is a possessive adjective. Most personal pronouns can be used as possessive adjectives directly.							

- Example 12

Annotated Khmer:	សាលា _n[n	រៀន _v]n	នេះ _a	ជា _v	សាលា _n[n	រៀន _v]n	ខ្ញុំ _a	។ _.
English gloss:	school	to-study	this	to-be	school	to-study	my	.
English translation:	This school is my school.							
note:	“នេះ” is a demonstrative adjective. Most demonstrative pronoun can be used as demonstrative adjectives directly.							

As Khmer has head-initial structure, the adjective usually comes after the nouns they modify, as illustrated by the examples. However, adjectives borrowed from Sanskrit or Pali may precede the nouns they modify to form head-final structures. Typical examples are សាធារណ_n[a រដ្ឋ_n]n (republic), which is composed by សាធារណ_a (public) and រដ្ឋ_n (state), and បុព្វ_n[a បុរស_n]n (ancestor), which is composed by បុព្វ_a (ancient) and បុរស_n (person).

4.4. Usage of "o" Tag

The "o" tag is applied for all the functional tokens, including adverbs, auxiliary verbs, prepositions, conjunctions, and various particles. Generally, the "o" tag can be applied for any ambiguous tokens with a certain syntactic role (or the "+" tag will be applied). Specific examples are illustrated in the following.

- Example 13

Annotated Khmer:	ឯង _n	កុំ _o	បើក _v	ម៉ូតូ _n	លឿន _o	ពេក _o	។ _.
English gloss:	you	not	to-ride	bike	fast	too	.
English translation:	Do not ride a bike too fast.						
note:	"កុំ" is a particle for negation. "លឿន" is a common adverb to show the manner of an activity. "ពេក" is an adverb for degree.						

- Example 14

Annotated Khmer:	វា_n	នៅ_v	ឆ្ងាយ_o	។_.
English gloss:	it	to-stay	far	.
English translation:	It is far from here.			
note:	"ឆ្ងាយ" is an adverb to show place.			

- Example 15

Annotated Khmer:	ជួនជា _o	ខ្ញុំ _n	ដើរ _v	កាត់ _v	ទី _n	នោះ _a	។ _.
English gloss:	just	I	to-walk	to-pass-through	place	that	.
English translation:	I just passed there.						
note:	"ជួនជា" is an adverb to show time.						

- Example 16

Annotated Khmer:	តើ_o	ឯង_n	មក_v	អង្កាល_o	?_.
English gloss:	n/a	you	to-come	when	?

English translation: When will you arrive?

note: “អង្កាល” is an interrogative adverb.
“តើ” is a particle for interrogative sentences.
Other common interrogative adverbs are
“ដូចម្ដេច” (how), and “ប៉ុន្មាន” (how much).

- Example 17

Annotated Khmer:	ខ្ញុំ _n	បាន _o	អាន _v	សៀវភៅ _n	នេះ _a	ចប់ _v	ហើយ _o	។ _.
English gloss:	I	already	to-read	book	this	to-finish	n/a	.

English translation: I have finished the book.

note: “បាន” is an aspect marker, considered as auxiliary verb.
“ហើយ” is a final particle indicating completed action.

- Example 18

Annotated Khmer:	ខ្ញុំ _n	ធ្លាប់ _o	និយាយ _v	រឿង _n	នេះ _a	ដែរ _o	។ _.
English gloss:	I	to-be-used-to	to-tell	story	this	too	.

English translation: I am used to tell this story too.

note: “ធ្លាប់” is an auxiliary verb. “ដែរ” is a final particle.

- Example 19

Annotated Khmer:	គាត់_n	អាច_o	ច្រៀង_v[v	បាន_v]v	ពីរោះ_o	។_.
English gloss:	she/he	can	to-sing	to-get	sweetly	.

English translation: She (He) can sing sweetly.

note: “អាច” is a modal verb. “ពីរោះ” is an adverb.

- Example 20

Annotated Khmer:	យើង _n	ធ្វើ _v	ដំណើរ _n	តាម _o	យន្ត _n[n	ហោះ _n]n	។ _.
English gloss:	we	to-do	travel	by	machine	to-fly	.

English translation: We travel by plan.

note: “តាម” is a preposition.

- Example 21

Annotated Khmer:	គាត់_n	សរសេរ_v	ស្អាត_o	ហើយ_o	ត្រឹមត្រូវ_o	។_.
English gloss:	she/he	to-write	neatly	and	correctly	.

English translation: She (He) writes neatly and correctly.

note: “ហើយ” is a coordinating conjunction.
“ស្អាត” and “ត្រឹមត្រូវ” are common adverbs.
Another common coordinating conjunctions is “ឬ” (or).

- Example 22

Annotated Khmer:	ខ្ញុំ _n	មិន _o	ទៅ _v	ធ្វើ _v[v	ការ _n]v	ព្រោះ _o	ខ្ញុំ _n	ឈឺ _v	។ _.
English gloss:	I	not	to-go	to-do	work	because	I	sick	.

English translation: I do not go to work because I am sick.

note: “មិន” is a particle for negation.
“ព្រោះ” is a subordinating conjunction.

- Example 23

Annotated Khmer:	ខ្ញុំ_n	មិន_o	ដឹង_v	ទេ_o	។_.
English gloss:	I	not	to-know	n/a	.

English translation: I do not know.

note: “មិន” is a particle for negation.
“ទេ” is a final particle for negated sentences.
Other common particles for negation are “គ្មាន” and “កុំ”

4.5. Usage of Auxiliary Tags

The three auxiliary tags “1”, “.”, and “+” are used trivially to represent numbers, punctuations marks, and tokens with weak syntactic roles, e.g., interjections.

Specifically, numbers represented in Arabic numbers, Khmer numbers, and in lexical forms are all annotated by the “1” tag, e.g., 123_1, ប្រាំ_1, and ១២៣៩_1. Common Khmer punctuation marks annotated with the “.” tag are “។”, “៕”, “៛”, “៖”, and “។ល។”. Common Khmer particles annotated with the “+” tag are “អាដេញ”, “អា”, “ណ៉ែ”, “អើ”, “ណេះ”, “មើលី៍”, “បាទ”, and “ចា៎”, to expressing hesitation or confirmation in conversation.

5. Part-of-Speech Annotation for Multi-Token Compounds

5.1. Identification

As Khmer has a head-initial structure and adjectives are usually placed after the noun they modify, the integration of nominal compounds composed of multiple tokens can be identified by the place of the adjective placed. For example, four morphemes can be identified with the expression of “ឡានដឹកទំនិញស”, which are “ឡាន” (vehicle), “ដឹក” (to-carry), “ទំនិញ” (goods), and “ស” (white). If a further adjectival morpheme like “តូច” (small) is used to modify the expression, the most natural place of insertion is between “ឡានដឹកទំនិញ” and “ស”, from which we can identify the “ឡានដឹកទំនិញ” (truck) is an integrated compound. As a further example, “ដីស” composed of two morphemes “ដី” (soil) and “ស” (white), while the “តូច” (small) cannot be placed between them but be attached to the whole expression for modification. So “ដីស” should be considered as an integrated compound. It should be notice that the compound identification cannot be decided only by specific morphemes from the two examples, that the relation and integrity among components should be considered.

The combination between verbal morphemes are looser than that of nominal ones. We can use the same insertion approach to judge whether a sequence of verbal morphemes is compound or not. For example, the expression of “ទៅលេង” contains two morphemes “ទៅ” (to-go) and “លេង” (to-visit), and a further object “ផ្ទះ” (home) can be inserted between them. So, “ទៅលេង” will not be treated as a compound. Another examples is “ព្យាយាមការពារ”, where two morphemes “ព្យាយាម” (to-try) and “ការពារ” (to-protect) can be identified. When the expression is negated, the negation particle “មិន” (not) is placed between them. So, this expression is still not a compound.

5.2. Common Patten

For the annotation of the compound, each token within the compound is annotated separately according to its own tag, while the first and last token are wrapped by a pair of brackets as mentioned in Section 3. Generally, there is no restriction on the number of tokens in a compound, while in Khmer a compound composed of two or three tokens (morphemes) are common. Examples of common compound patterns are listed in the following Tables 2 and 3.

Table 2. Examples of two-token compound

Annotated Khmer		English gloss		
1st-token	**2nd-token**	**1st-token**	**2nd-token**	**compound**
ចំណេះ_n[n	វិជ្ជា_n]n	knowledge	knowledge	→ knowledge
មុខ_n[n	មាត់_n]n	face	mouth	→ appearance
ចំនួន_n[n	លេខ_n]n	number	number	→ number
ដៃ_n[n	ជើង_n]n	hand	leg	→ stooge
រទេះ_n[n	ភ្លើង_n]n	cart	fire	→ train
ឆ្នេរ_n[n	សមុទ្រ_n]n	beach	sea	→ beach
ស្រោម_n[n	ដៃ_n]n	cover	hand	→ gloves
ជាតិ_n[n	ដែក_n]n	substance	iron	→ iron
គិត_v[v	គូរ_v]v	to-think	to-draw	→ to-think-about
ស៊ី_v[v	ផឹក_v]v	to-eat	to-drink	→ to-eat-and-drink
បែង_v[v	ចែក_v]v	to-divide	to-divide	→ to-divide
ខ្ពង់_a[a	ខ្ពស់_a]a	high	high	→ high
ថោក_a[a	ទាប_a]a	cheap	low	→ cheap
ស្ងៀម_a[a	ស្ងាត់_a]a	silent	silent	→ silent
ងាយ_a[a	ស្រួល_a]a	easy	easy	→ easy
ហែល_v[v	ទឹក_n]v	to-swim	water	→ to-swim
លោត_v[v	ទឹក_n]v	to-jump	water	→ to-dive
វង្វេង_v[v	ផ្លូវ_n]v	to-be-lost	way	→ to-lose-the-way
វិល_v[v	មុខ_n]v	to-spin	face	→ to-feel-dizzy
ទាល់_v[v	គំនិត_n]v	to-exhaust	idea	→ to-be-stuck
មាស_n[n	ឆ្អិន_a]n	gold	pure	→ pure-gold
ដី_n[n	ជាំ_a]n	soil	wet	→ wet-ground
រាង_n[n	ជ្រលម_a]n	shape	tapered	→ cone-shape
មនុស្ស_n[n	ចាស់_a]n	person	old	→ adult
ទឹក_n[n	ខ្មៅ_a]n	water	black	→ ink
ដី_n[n	ស_a]n	soil	white	→ chalk

Annotated Khmer		English gloss			
1st-token	2nd-token	1st-token	2nd-token		compound
រដូវ_n[n	ក្ដៅ_a]n	season	hot	→	summer
អាវ_n[n	ធំ_a]n	shirt	big	→	coat
ត្រី_n[n	អាំង_v]n	fish	broil	→	broiled-fish
សាលា_n[n	រៀន_v]n	school	study	→	school
យន្ត_n[n	ហោះ_v]n	machine	fly	→	plane
បន្ទប់_n[n	ដេក_v]n	room	sleep	→	bed-room

Table 3. Examples of three-token compound

Annotated Khmer			English gloss				
1st-tok.	2nd-tok.	3rd-tok.	1st-tok.	2nd-tok.	3rd-tok.		compound
ផ្កាយ_n[n	ដុះ_v	កន្ទុយ_n]n	star	to-grow	tail	→	comet
សេះ_n[n	ពាក់_v	បង្កង់_n]n	horse	to-wear	ring	→	zebra
បន្ទប់_n[n	ទទួល_v	ភ្ញៀវ_n]n	room	to-greet	guest	→	living-room

5.3. Affixation

Many very frequent components in compounds may be considered as affixes. Most of them are borrowed from Sanskrit or Pali, and a part of prefixes are originally Khmer nouns. These affixes are segmented as token and annotated by the brackets as compounds. Typical examples are shown in Tables 4 and 5.

Table 4. Examples of two-token compound, where the first token is a prefix

Annotated Khmer		English gloss			
1st-token	2nd-token	1st-token	2nd-token		compound
អ្នក_n[n	ជំនួញ_n]n	person (-er/-or)	trade	→	businessman
អ្នក_n[n	សៀមរាប_n]n		Siem Reap	→	Siem-Reap-people
អ្នក_n[n	ក្រ_a]n		poor	→	the-poor
ការ_n[n	ពិត_a]n	work/act	real	→	reality
ការ_n[n	ផ្ទះ_n]n		house	→	housework
សេចក្ដី_n[n	ក្លាហាន_a]n	case/state	brave	→	courage
សេចក្ដី_n[n	សង្ឃឹម_v]n		to-hope	→	hope
អនុ_n[a	មន្ត្រី_n]n	vice-	minister	→	junior-minister
អនុ_n[a	ប្រធាន_n]n		president	→	vice-president

Annotated Khmer		English gloss		
1st-token	2nd-token	1st-token	2nd-token	compound
កិច្ច_n[n	សន្យា_v]n	act	to-agree	→ agreement
កិច្ច_n[n	ប្រជុំ_v]n		to-meet	→ meeting
ឯក_a[a	រាជ_n]a	single/alone	reign	→ independent
ឯក_a[a	ទិស_n]a		direction	→ one-way
អគ្គ_n[a	នាយក_n]n	first/best	director	→ general-director
អគ្គ_n[a	វាចា_n]n		word	→ sophisticated-words

Table 5. Examples of two-token compound, where the second token is a suffix

Annotated Khmer		English gloss		
1st-token	2nd-token	1st-token	2nd-token	compound
សង្គម_n[n	និយម_n]n	society	principle	→ socialism
ប្រាកដ_n[a	និយម_n]n	real	(-ism)	→ realism
ទាស_n[n	ភាព_n]n	slave	state/condition	→ slavery
សេរី_n[n	ភាព_n]n	free-person		→ liberty
ប្រវត្តិ_n[n	វិទូ_n]n	history	scholar/expert	→ historian
ទស្សន_n[n	វិទូ_n]n	concept		→ philosopher
ភូមិ_n[n	សាស្ត្រ_n]n	earth	science/knowledge	→ geography
អក្សរ_n[n	សាស្ត្រ_n]n	letter		→ literature
រដ្ឋ_n[n	សភា_n]n	nation	assembly	→ parliament
ព្រឹទ្ធ_n[n	សភា_n]n	senior		→ senate
វិទ្យា_n[n	ស្ថាន_n]n	knowledge	marketplace	→ institute
ឱសថ_n[n	ស្ថាន_n]n	medicine		→ pharmacy

6. Confusion Cases

6.1. Functional Multi-Token Expression

In most cases, the brackets are applied for compounds with substantial meanings, most of which are nominal expressions. The bracket can also be used for functional expressions, as long as the meaning of components within them can be identified clearly, as the tokenization principle is to segment out small tokens as possible. Specific cases include some interrogative or indefinite pronouns, adverbs, and complex prepositions and conjunctions. Examples are listed as follows.

- Example 24

Annotated Khmer:	តើ	អ្នក	ណា	ទើប	នឹង	មក	?
	_o	_n[n	_n]n	_o[o	_o]o	_v	_.
English gloss:	n/a	person	which	then	with	to-come	?

English translation: Who has just come?

note: “អ្នក_n[n ណា_n]n” means “who”, an interrogative pronoun.
“ទើប_n[n នឹង_n]n means “just”, an adverb.

- Example 25

Annotated Khmer:	តើ	អ្នក	រាល់	គ្នា	ចង់	បាន	អ្វី	?
	_o	_n[n	_o	_n]n	_o	_v	_n	_.
English gloss:	n/a	person	every	n/a	to-want	to-get	what	?

English translation: What does everyone what?

note: “អ្នក_n[n រាល់_o គ្នា_n]n” means “everyone”, an indefinite pronoun, where “គ្នា” is a counter of person.

- Example 26

Annotated Khmer:	ក្រោយ	មក	ពិធី	ជប់លៀង	ក៏	បញ្ចប់	។
	_o[o	_o]o	_n[n	_v]n	_o	_v	_.
English gloss:	after	toward	ceremony	then	them	to-end	.

English translation: Afterward, the party ended.

note: “ក្រោយ_o[o មក_o]o” means “afterward”, an adverb.

- Example 27

Annotated Khmer:	មនុស្ស	ខ្លះ	នៅ	ខាង	ក្នុង	បន្ទប់	។
	_n	_a	_o[o	_o	_o]o	_n	_.
English gloss:	people	some	in	side	in	room	.

English translation: There are some people in the room.

note: “នៅ_o[o ខាង_o ក្នុង_o]o” is a complex preposition, used to show the place.

- Example 28

Annotated Khmer:	ប្រសិន _o[o	បើ _o]o	អ្នក _n	នៅ _o[o	តែ _o]o	មិន _o	យល់ _v
English gloss:	if	if	you	in	just	not	to-understand

English translation: If you still not understand…

note: "ប្រសិន_o[o បើ_o]o" means "if", a conjunction.
"នៅ_o[o តែ_o]o" means "still", an adverb.

6.2. Annotation Around Number and Reduplication

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 are listed as follows.

- Example 29

Annotated Khmer:	ខ្ញុំ _n	បាន _o	ទិញ _v	សៀវភៅ _n	ដប់ _a[1	ក្បាល _n]a	។ _.
English gloss:	I	have	to-buy	book	ten	unit	.

English translation: I have bought ten books.

note: "ក្បាល" is a counter for counting books.

- Example 30

Annotated Khmer:	ខ្ញុំ _n	រៀន _v	ដល់ _o	មេរៀន _n	ទី _a[n	បួន _1]a	។ _.
English gloss:	I	to-study	up-to	lesson	place	four	.

English translation: I study the lesson fourth.

note: "ទី" originally means "place", used to form an ordinal number.

- Example 31

Annotated Khmer:	ដាណេ _n	រៀន _v	ថ្នាក់ _n	ទី _a[n	ប្រាំ _1	បី _1]a	។ _.
English gloss:	Danei	to-study	class	place	five	three	.
English translation:	Danei is an eighth-grade student.						
note:	As Khmer uses a quinary numeral system, digits are segmented.						

The phenomenon of reduplication is common in Khmer, which can be applied on nouns, adjectives, and adverbs. The reduplication is annotated by a special “ៗ” mark, which is annotated by “.” and wrapped with the preceding main token by brackets.

- Example 32

Annotated Khmer:	ក្មេង _n[n	ៗ _.]n	កំពុង _o	លេង _v	បិទ _n[v	ពួន _v]n	។ _.
English gloss:	child	-s	-ing	to-play	to-stick	to-hide	.
English translation:	children are playing hide-and-seek.						
note:	“ៗ” is used to address plural here.						

6.3. Context Depended Tagging

The annotation around functionalized or grammaticalized tokens should depend on the contexts, i.e., the specific role they played. A typical case is nominal pronouns playing attributive rather than a substitutive role, where “a” should be used instead of “n”. Specific examples are listed as follows.

- Example 33

Annotated Khmer: នេះ_n ជា_v សៀវភៅ_n ។_.

English gloss: this is book .

English translation: This is a book.

note: “នេះ” is used as a nominal token.

- Example 34

Annotated Khmer: ខ្ញុំ_n ស្រឡាញ់_v សៀវភៅ_n នេះ_a ។_.

English gloss: I to-like book this .

English translation: I like this book.

note: “នេះ” is used as an adjectival token.

- Example 35

Annotated Khmer: មួយ_n[1 ណា_n]n ជា_v របស់_n ឯង_a ?_.

English gloss: one which is thing you ?

English translation: Which is yours?

note: “មួយ_n[1 ណា_n]n” is a complex interrogative pronoun.
“ឯង” is a pronoun used as adjective for possession.

- Example 36

Annotated Khmer: តើ ម៉ូតូ មួយ ណា ជា របស់ ឯង ?
_o _n _a[1 _n]a _v _n _a _.

English gloss: n/a bike one which is thing you ?

English translation: Which bike is yours?

note: “មួយ_a[1 ណា_n]a” here is used as an adjective token.

Another phenomenon is the relatively free changing among nominal, adjectival, and verbal tokens. Specific examples are listed as follows. The tagging must conducted with a consideration of the phrase or sentence structure and the specific role the token plays.

- Example 37

Annotated Khmer:	ផ្លែ_n[n	ស្វាយ_n]n	នេះ_a	ទុំ_v	ហើយ_o	។_.
English gloss:	fruit	mango	this	to-ripen	already	.

English translation: Mangos ripen.

note: “ផ្លែ” is a noun with a meaning of “fruit”.
“ទុំ” is a verb with a meaning of “to ripen”.

- Example 38

Annotated Khmer:	ស្វាយ_n	ផ្លែ_v	ច្រើន_o	ណាស់_o	។_.
English gloss:	mango	to-bear-fruit	much	very	.

English translation: Mango trees bear a lot of fruits.

note: “ផ្លែ” is used as a verbal token here to address “bear fruit”.

- Example 39

Annotated Khmer:	ស្រូវ_n	ទុំ_a	ត្រូវ_o	បាន_o	ដក_v	។_.
English gloss:	rice	ripe	n/a	already	to-pull	.

English translation: Ripe rice has been pulled.

note: “ទុំ” is used as an adjectival token here to address “ripe”.