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

(Version 0.2, August 2018)

Chenchen Ding[1], Hour Kaing[1], Kamthong Ley[1, 2],
Masao Utiyama[1], 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

This is a supplementary document for *Tokenization and Part-of-Speech Annotation Guidelines for Khmer (Cambodian)*. Further modified NOVA tags and instructions on confusing cases are provided.

The basic NOVA tags used in preliminarily annotated Khmer texts can be further modified to add more detailed information. The further introduced tags are listed in Table 1. Generally, a "-" is added to basic NOVA tags to further address the functionality of a token, i.e., to distinguish functional tokens from content tokens.

Table 1. Modified basic tags in NOVA

Tag	Description
`n-`	general pronouns, including personal and demonstrative
`v-`	grammaticalized auxiliary verbs modifying other content verbs
`a-`	general functional noun-modifiers, including possessive and demonstrative
`o-`	general particles, e.g., negators, conjunctions, mood particles, etc.

The usage of "`n-`", "`v-`", "`a-`", and "`o-`" tags are illustrated in following sections. For all the examples illustrated, the tags are attached to corresponding tokens by an underline ("_").

2. Modification for Single Tokens

2.1. Usage of “n-” Tag

- Example 1

Annotated Khmer:	តើ _o-	ឯង _n-	ចូល _v[v	ចិត្ត _n]v	មុខ _n[n	វិជ្ជា _n]n	អ្វី _n	? _.
English gloss:	n/a	you	to-enter	heart	field	subject	what	?

English translation: What subject do you like?

note: “ឯង” is a personal pronoun.

Common personal pronouns are: “ខ្ញុំ”, “ឯង”, “វា”, “គាត់”, “យើង”, and “គេ”.

- Example 2

Annotated Khmer:	នេះ_n-	ជា_v	ចំណែក_n	ឯង_a-	។_.
English gloss:	this	to-be	part	you	.

English translation: This is your part.

note: “នេះ” is a demonstrative pronoun.

Common demonstrative pronouns are “នេះ” and “នោះ”.

2.2. Usage of “v-” Tag

The “v-” tag is used for grammaticalized verbal tokens modifying other content verbal tokens. The grammaticalized verb can precede or follow the content verb it modifies. When the content verb is transitive, the directed object may be inserted in between the two verbs.

- Example 3

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

English translation: I learnt the Khmer language.

note: “បាន” is grammaticalized as a past tense marker.

Common pre-positioned grammaticalized verbs are:

“អោយ”, “គឺ”, and “បន្ត”.

- Example 4

Annotated Khmer:	នាង	ធ្វើ	ឱ្យ	ខ្ញុំ	ពិបាក	ចិត្ត	។
	_n-	_v	_v-	_n-	_v[v	_n]v	_.
English gloss:	she	to-do	to-let	I	to-suffer	heart	.
English translation:	She makes me unhappy.						
note:	“ឱ្យ” is grammaticalized as a causative marker. Common post-positioned grammaticalized verbs are: “ទៅ”, “នៅ”, “អោយ”, “ឡើង”, and “ថា”,						

- Example 5

Annotated Khmer:	គាត់	ផ្តល់	ជំនួយ	ដល់	ខ្ញុំ	។
	_n-	_v	_n	_v-	_n-	_.
English gloss:	he	to-offer	support	to-reach	I	.
English translation:	He offers me the support.					
note:	“ដល់” is a grammaticalized verb modifying “ផ្តល់”. The directed objected “ជំនួយ” is in between them.					

2.3. Usage of “a-” Tag

- Example 6

Annotated Khmer:	ផ្ទះ	ខ្ញុំ	សង់	លើ	វាល	ស្រែ	។
	_n	_a-	_v	_o	_n[n	_n]n	_.
English gloss:	house	my	to-build	over	field	paddy	.
English translation:	My house is built in the rice field.						
note:	“ខ្ញុំ” is a possessive adjective modifying “ផ្ទះ”.						

- Example 7

Annotated Khmer:	បេសកកម្ម	នេះ	មាន	សារៈសំខាន់	ណាស់	។
	_n	_a-	_v	_n	_o	_.
English gloss:	mission	this	to-have	importance	very	.
English translation:	This mission is very important.					
note:	“នេះ” is a demonstrative adjective modifying “បេសកកម្ម”.					

2.4. Usage of “o-” Tag

- Example 8

Annotated Khmer:	ព្រោះ_o-	ខ្ញុំ_n-	មិន_o-	ដឹង_v	ទេ_o-
English gloss:	because	I	not	to-go	n/a

English translation: because I don’t know

note: “ព្រោះ” is a subordinating conjunction.
“មិន” is a particle for negation.
“ទេ” is a final particle for negated sentences.

- Example 9

Annotated Khmer:	តើ_o-	អ្នក_n-	សុខ_v[a	សប្បាយ_v]v	ទេ_o-	?_.
English gloss:	n/a	you	happy	to-be-happy	n/a	?

English translation: How are you?

note: “តើ” is an interrogative particle.
“ទេ” is a final particle for interrogative sentences.

3. Modification for Multi-Token Compounds

Table 2. Examples of two-token compound pronouns

Annotated Khmer		**English gloss**			
1st-token	**2nd-token**	**1st-token**	**2nd-token**	**compound**	**note**
នាង_n-[n-	ខ្ញុំ_n-]n-	she	I	→ I (female)	gender
ខ្ញុំ_n-[n-	បាទ_o]n-	I	yes	→ I (male)	
ខ្លួន_n-[n	គាត់_n-]n-	oneself	he	→ I	emphasis
រូប_n-[n	គាត់_n-]n-	body	he	→ he	
អ្នក_n-[n-	ស្រី_n]n-	you	female	→ you (female)	politeness
លោក_n-[n-	ស្រី_n]n-	you	female	→ you (female)	
លោក_n-[n-	ប្រុស_n]n-	you	male	→ you (male)	
ពួក_n-[n	ខ្ញុំ_n-]n-	group	I	→ we (exclusive)	plurality
ពួក_n-[n	យើង_n-]n-	group	we	→ we (inclusive)	
ទាំង_n-[o	នេះ_n-]n-	all	this	→ these	

Table 3. Examples of two-token compound auxiliary verbs

Annotated Khmer		English gloss		
1st-token	**2nd-token**	**1st-token**	**2nd-token**	**compound**
គ្រាន់_v-[v-	តែ_o]v-	to-be-enough	only	→ to-be-just
ចេះ_v-[v-	តែ_o]v-	to-know	only	→ to-be-always
ចាប់_v-[v-	ផ្ដើម_v-]v-	to-start	to-start	→ to-start-to

Table 4. Examples of two-token compound conjunctions and negation particles

Annotated Khmer		English gloss		
1st-token	**2nd-token**	**1st-token**	**2nd-token**	**compound**
ប្រសិន_o-[o-	បើ_o-]o-	if	if	→ if
មិន_o-[o-	មែន_o]o-	not	really	→ not
ព្រោះ_o-[o-	តែ_o]o-	because	only	→ because

4. Confusion Cases

- Example 10

Annotated Khmer: ចៅ_n- និយាយ_v តាម_v- តា_n- ។_.

English gloss: you say follow I .

English translation: You repeat after me.

note: "ចៅ" (grandson) and "តា" (grandpa) are nouns.
Here they are used as pronouns for "you" and "me".

- Example 11

Annotated Khmer: លោក_n[n ដេវីត_n អាន់ឌឺសុន_n]n

English gloss: Mr. David Andersen

English translation: Mr. David Andersen

note: "លោក" is a pronoun for "you".
Here it is used as a title.

- Example 12

Annotated Khmer:	ផ្ទះ_n	របស់_o	ខ្ញុំ_n-
English gloss:	house	of	my
English translation:	my house		
note:	"ខ្ញុំ" is considered as a pronoun.		
	"របស់" is a possessive marker.		