Tokenization and Part-of-Speech Annotation
Guidelines for Khmer (Cambodian)
(Version 0.2, December 2016)

Chenchen Ding\textsuperscript{1}, Hour Kaing\textsuperscript{1,2}, Masao Utiyama\textsuperscript{1},
Vichet Chea\textsuperscript{2}, Eiichiro Sumita\textsuperscript{1}
\textsuperscript{1}Advanced Translation Technology Laboratory, ASTREC, NICT, Japan
\textsuperscript{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 “c” 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

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

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 “សាលាតៀន” is a compound composed by two words “សាលា” and “តៀន”.

Example 1

- **Tokenized Khmer:** តនេះជា សាលាតៀន 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 “x [x₁ x₂ ... xₙ] x”, where xₖ are the tags for each component morpheme and x is the tag for the integrated expression. Here the “x [x₁]” and “xₙ] 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: ស្រប់_n មាន់_v ប្រាសាទ_n មាន់_o ។
  English gloss: Cambodia to-have-plenty-of temple very .
  English translation: Cambodia has many temples.
  note: “ស្រប់” is a proper noun; “ប្រាសាទ” is a common noun.

- Example 3
  Annotated Khmer: គ្រូ_o មាន់_v[v_ n]v[n_ n][n_ n]n_a ។
  English gloss: n/a you to-enter heart head subject what ?
  English translation: What subject do you like?
  note: “មាន់” is a personal pronoun. Common personal pronouns are: “ខ”, “ខ្ញុំ”, “វា”, “គាំតិ”, and “ត្រូវ”.

- Example 4
  Annotated Khmer: តនេះ_n ជា_v ប្រាសាទ_n មាន់_a ។
  English gloss: this to-be part your .
  English translation: This is yours.
  note: “តនេះ” is a demonstrative pronoun. Common demonstrative pronouns are: “នេះ”, “នាង”, “អ្វី”, and “នេះខ្ញុំ”.
- Example 5

Annotated Khmer: តេើ អ្វី មាន លលៃ ជាង តគ ?
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”, “ន ណា”, “អ្វី”, “ណា”, and “ណាខៃេះ”.

- Example 6

Annotated Khmer: កសត្តុ ត្រឹម ឈោះ ដែល កន្លែង តែ ប្រសលាញ់ ជាង តគ
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: ខ្ការ ដួល ភាសា កន្លែង តៀន ។.
English gloss: I to-learn language Khmer .
English translation: I learn the Khmer language.

Note: “ដួល” is a transitive verb.

- Example 8

Annotated Khmer: កូន ដឹក កន្លែង នៃីន នៃី ។.
English gloss: child to-sleep in hammock .
English translation: The child sleeps in a hammock.

Note: “ដឹក” is an intransitive verb.
- Example 9

Annotated Khmer: ខ្ញុំ ជា បូលីស។

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 េូច ជិេះ ម្ិន តលឿន។

English gloss: bicycle small to-ride not can fast n/a.

English translation: Riding a small bicycle cannot be fast.

note: “េូច” is a common adjective.

- Example 11

Annotated Khmer: ផ្ទេះ គាេ់ សង់ តៅ កណា ា ល វាល។

English gloss: house my to-build in middle field paddy.

English translation: My house is built in the middle of the rice field.

note: “គាេ់” is a possessive adjective. Most personal pronouns can be used as possessive adjectives directly.

- Example 12

Annotated Khmer: សាលា តៀន ជា សាលា ខ្ញុំ ។

English gloss: school to-study this to-be school to-study my.

English translation: This school is my school.

note: “តៀន” 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 សាធា ណ (public) and ែឋ (state), and បុព្វ (ancient) and បុ ស (person).

4.4. Usage of “ο” Tag

The “ο” tag is applied for all the functional tokens, including adverbs, auxiliary verbs, prepositions, conjunctions, and various particles. Generally, the “ο” 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 កុុំ តបើក ម្ ូេូ តព្ក

  English gloss: you not to-ride bike fast too

  English translation: Do not ride a bike too fast.

  note: “កុុំ” is a particle for negation.

- Example 14

  Annotated Khmer: វា ឆ្ង ា យ

  English gloss: it to-stay far

  English translation: It is far from here.

  note: “ឆ្ង ា យ” is an adverb to show place.

- Example 15

  Annotated Khmer: ជួនជា ខ ្ញុំ តែើ កាេ់ ទី

  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:** តេើ ឯង ម្ក ា ល ្ើ ?

**English gloss:** n/a you to-come when ?

**English translation:** When will you arrive?

**note:** "អ្ង្ក ា ល" is an interrogative adverb.

"តេើ" is a particle for interrogative sentences.

Other common interrogative adverbs are "ែូចតម្តច" (how), and "ប ុ ែ ន" (how much).

- Example 17

**Annotated Khmer:** ខ ្ញុំ ាន អាន តសៀិតៅ តនេះ ចប់ ត ើយ ។

**English gloss:** I already to-read book this to-finish n/a .

**English translation:** I have finished the book.

**note:** “ខ ្ញុំ” is an aspect marker, considered as auxiliary verb.

“ត ើយ” is a final particle indicating completed action.

- Example 18

**Annotated Khmer:** ខ ្ញុំ ធា ៃ ប់ និយាយ ត ឿង តនេះ ណែ ។

**English gloss:** I to-be-used-to to-tell story this too .

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

**note:** “ធា ៃ ប់” is an auxiliary verb. “ណែ” is a final particle.

- Example 19

**Annotated Khmer:** គាេ់ អាច តប្រចៀង ស ទត េះ ។

**English gloss:** she/he can to-sing to-get sweetly .

**English translation:** She (He) can sing sweetly.

**note:** “អាច” is a modal verb. “ទត េះ” is an adverb.
- Example 20

Annotated Khmer: 

English gloss: we to-do travel by machine to-fly.

English translation: We travel by plan.

note: “តាម្” is a preposition.

- Example 21

Annotated Khmer: 

English gloss: she/he to-write neatly and correctly.

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

note: “សា ា េ” and “ប្រេឹម្ប្រេូិ” are common adverbs.

Another common coordinating conjunctions is “ឬ” (or).

- Example 22

Annotated Khmer: 

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: “ប្ររេះ” is a subordinating conjunction.

- Example 23

Annotated Khmer: 

English gloss: I not to-know n/a.

English translation: I do not know.

note: “ប្ររេះ” is a subordinating conjunction.

Other common particles for negation are “គា ែ ន” and “កុុំ”.

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, ៖_1, and ៖_1. Common Khmer punctuation marks annotated with the “.” tag are “。”, “៖”, “ែ”, “ៗ”, and “ៗ”。

Common Khmer particles annotated with the “+” tag are “អាតែញ”, “អា”, “ណណ”, “តអ្ើ”, “តណេះ”, “តម្ើលិ៍”, “ាទ”, 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 “ឡានែឹកទុំនិញស”, which are “ឡាន” (vehicle), “ែឹក” (to-carry), “ទុំនិញ” (goods), and “ស” (white). If a further adjectival morpheme like “េូច” (small) is used to modify the expression, the most natural place of insertion is between “ឡានែឹកទុំនិញ” and “ស”, from which we can identify the “ឡានែឹកទុំនិញស” (truck) is an integrated compound. As a further example, “ែីស” composed of two morphemes “ែី” (soil) and “ស” (white), while the “េូច” (small) cannot be placed between them but be attached to the whole expression for modification. So “ែីស” 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 “តៅតលង” contains two morphemes “តៅ” (to-go) and “តលង” (to-visit), and a further object “ផ្ទេះ” (home) can be inserted between them. So, “តៅតលង” will not be treated as a compound. Another examples is “ព្ាយាម្កា រ”, where two morphemes “ព្ាយាម្កា” (to-try) and “កា រ” (to-protect) can be identified. When the expression is negated, the negation particle “ម្ិន” (not) is placed between them. So, this expression is still not a compound.

5.2. Common Pattern
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>
<thead>
<tr>
<th>1st-token</th>
<th>2nd-token</th>
<th>1st-token</th>
<th>2nd-token</th>
<th>compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>ឆ្នាំង_ _n</td>
<td>ៈ្មន_ _n</td>
<td>knowledge</td>
<td>knowledge</td>
<td>→ knowledge</td>
</tr>
<tr>
<td>ំំំ_ _n</td>
<td>៉ាំ_ _n</td>
<td>face</td>
<td>mouth</td>
<td>→ appearance</td>
</tr>
<tr>
<td>កស្កុត_ _n</td>
<td>ំំំ_ _n</td>
<td>number</td>
<td>number</td>
<td>→ number</td>
</tr>
<tr>
<td>ៃ_ _n</td>
<td>៉ាំ_ _n</td>
<td>hand</td>
<td>leg</td>
<td>→ stooge</td>
</tr>
<tr>
<td>លោះ_ _n</td>
<td>៉ាំ_ _n</td>
<td>cart</td>
<td>fire</td>
<td>→ train</td>
</tr>
<tr>
<td>លោះ_ _n</td>
<td>៉ាំ_ _n</td>
<td>beach</td>
<td>sea</td>
<td>→ beach</td>
</tr>
<tr>
<td>ប៉ាំសន_ _n</td>
<td>៉ាំ_ _n</td>
<td>cover</td>
<td>hand</td>
<td>→ gloves</td>
</tr>
<tr>
<td>ស្ទឹង_ _n</td>
<td>៉ាំ_ _n</td>
<td>substance</td>
<td>iron</td>
<td>→ iron</td>
</tr>
<tr>
<td>េេូ_ _v[ ]v</td>
<td>៍រៀ_ _v[v</td>
<td>to-think</td>
<td>to-draw</td>
<td>→ to-think-about</td>
</tr>
<tr>
<td>េេូ_ _v[ ]v</td>
<td>៍រៀ_ _v[v</td>
<td>to-eat</td>
<td>to-drink</td>
<td>→ to-eat-and-drink</td>
</tr>
<tr>
<td>េេូ_ _v[ ]v</td>
<td>៍រៀ_ _v[v</td>
<td>to-divide</td>
<td>to-divide</td>
<td>→ to-divide</td>
</tr>
<tr>
<td>េេូ_ _a[a</td>
<td>េេូ_ _a[a</td>
<td>high</td>
<td>high</td>
<td>→ high</td>
</tr>
<tr>
<td>េេូ_ _a[a</td>
<td>េេូ_ _a[a</td>
<td>cheap</td>
<td>low</td>
<td>→ cheap</td>
</tr>
<tr>
<td>េេូ_ _a[a</td>
<td>េេូ_ _a[a</td>
<td>silent</td>
<td>silent</td>
<td>→ silent</td>
</tr>
<tr>
<td>េេូ_ _a[a</td>
<td>េេូ_ _a[a</td>
<td>easy</td>
<td>easy</td>
<td>→ easy</td>
</tr>
<tr>
<td>េេូ_ _v[ ]v</td>
<td>៍រៀ_ _n[v</td>
<td>to-swim</td>
<td>water</td>
<td>→ to-swim</td>
</tr>
<tr>
<td>េេូ_ _v[ ]v</td>
<td>៍រៀ_ _n[v</td>
<td>to-jump</td>
<td>water</td>
<td>→ to-dive</td>
</tr>
<tr>
<td>េេូ_ _v[ ]v</td>
<td>៍រៀ_ _n[v</td>
<td>to-be-lost</td>
<td>way</td>
<td>→ to-lose-the-way</td>
</tr>
<tr>
<td>េេូ_ _v[ ]v</td>
<td>៍រៀ_ _n[v</td>
<td>to-spin</td>
<td>face</td>
<td>→ to-feel-dizzy</td>
</tr>
<tr>
<td>េេូ_ _v[ ]v</td>
<td>៍រៀ_ _n[v</td>
<td>to-exhaust</td>
<td>idea</td>
<td>→ to-be-stuck</td>
</tr>
<tr>
<td>េេូ_ _n[n</td>
<td>៍រៀ_ _a[n</td>
<td>gold</td>
<td>pure</td>
<td>→ pure-gold</td>
</tr>
<tr>
<td>េេូ_ _n[n</td>
<td>៍រៀ_ _a[n</td>
<td>soil</td>
<td>wet</td>
<td>→ wet-ground</td>
</tr>
<tr>
<td>េេូ_ _n[n</td>
<td>៍រៀ_ _a[n</td>
<td>shape</td>
<td>tapered</td>
<td>→ cone-shape</td>
</tr>
<tr>
<td>េេូ_ _n[n</td>
<td>៍រៀ_ _a[n</td>
<td>person</td>
<td>old</td>
<td>→ adult</td>
</tr>
<tr>
<td>េេូ_ _n[n</td>
<td>៍រៀ_ _a[n</td>
<td>water</td>
<td>black</td>
<td>→ ink</td>
</tr>
<tr>
<td>េេូ_ _n[n</td>
<td>៍រៀ_ _a[n</td>
<td>soil</td>
<td>white</td>
<td>→ chalk</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>Annotated Khmer</th>
<th>English gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st-token</td>
<td>2nd-token</td>
</tr>
<tr>
<td>កិចច_n[ n សនា_v]n</td>
<td>act to-agree  → agreement</td>
</tr>
<tr>
<td>កិចច_n[ n ប្រប_v]n</td>
<td>to-meet  → meeting</td>
</tr>
<tr>
<td>ប្រប _a[a កាស_n]a</td>
<td>single/alone  → independent</td>
</tr>
<tr>
<td>ប្រប _a[a ប្រប_v]n</td>
<td>direction  → one-way</td>
</tr>
<tr>
<td>ឧក_n[a រស្ត្រ_n]n</td>
<td>first/best  director  → general-director</td>
</tr>
<tr>
<td>ឧក_n[a កាស_n]n</td>
<td>word  → sophisticated-words</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Annotated Khmer</th>
<th>English gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st-token</td>
<td>2nd-token</td>
</tr>
<tr>
<td>សងគម្_n[n និយម្_n]n</td>
<td>society principle  → socialism</td>
</tr>
<tr>
<td>ប្រាកែ_n[a និយម្_n]n</td>
<td>real (-ism)  → realism</td>
</tr>
<tr>
<td>ទស_ n[n ភាព្_n]n</td>
<td>slave state/condition  → slavery</td>
</tr>
<tr>
<td>ទសសន_ n[n ភាព្_n]n</td>
<td>free-person state/condition  → liberty</td>
</tr>
<tr>
<td>ប្រព្ឹទធ_ n[n សភា_n]n</td>
<td>history scholar/expert  → historian</td>
</tr>
<tr>
<td>ប្រព្ឹទធ_ n[n សភा_n]n</td>
<td>concept  → philosopher</td>
</tr>
<tr>
<td>ែឋ_ n[n សភា_n]n</td>
<td>earth science/knowledge  → geography</td>
</tr>
<tr>
<td>អ្កស_ n[n សាន្តសត_n]n</td>
<td>letter  → literature</td>
</tr>
<tr>
<td>ែឋ_ n[n សភា_n]n</td>
<td>nation assembly  → parliament</td>
</tr>
<tr>
<td>កិចច_ n[n សាន្តសត_n]n</td>
<td>senior  → senate</td>
</tr>
<tr>
<td>ឱសល_ n[n សាន្តសត_n]n</td>
<td>knowledge marketplace  → institute</td>
</tr>
<tr>
<td>ឱសល_ n[n សាន្តសត_n]n</td>
<td>medicine marketplace  → pharmacy</td>
</tr>
</tbody>
</table>

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 ត្រូវ នឹង មក ឬ?

English gloss: n/a person which then with to-come ?

English translation: Who has just come?

“ទេើ_n[ n ប្រការ_n]n means “just”, an adverb.

- Example 25

Annotated Khmer: ត្រូវ គារ ញូ ឬ និង រឿង ឬ?

English gloss: n/a person every n/a to-want to-get what ?

English translation: What does everyone what?

note: “ត្រូវ_ n[ n គារ_o ញូ_n]n” means “everyone”, an indefinite pronoun, where “ញូ” is a counter of person.

- Example 26

Annotated Khmer: ឱ្យប្រការ អាច សម្រាប់ ក្នុង កបូជាគៅ ឬ?

English gloss: after toward ceremony then them to-end .

English translation: Afterward, the party ended.


- Example 27

Annotated Khmer: មនុស្ស ស្រុក គណៈ បាន ស្លាប់ ឬ?

English gloss: people some in side in room .

English translation: There are some people in the room.

note: “ស្លាប់_o[ o គណៈ_o បាន_o]o” is a complex preposition, used to show the place.
- Example 28

Annotated Khmer: ប្របសិន បំ ស្រុក ទៅ អ្នក ម្ិន យល់

English gloss: if if you in just not to-understand

English translation: If you still not understand…

note: “ប្របសិន” means “if”, a conjunction.
“តៅ” 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 បាន ទិញ តសៀវកម្ម ែប់ កាល

English gloss: I have to-buy book ten unit.

English translation: I have bought ten books.

note: “កាល” is a counter for counting books.

- Example 30

Annotated Khmer: ខ្ញុំ បាន ទិញ តសៀវកម្ម ទៅ ទី បួន

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.
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: 

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.
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 33

Annotated Khmer: តនេះ ជា តសៀិតៅ។

English gloss: this is book.

English translation: This is a book.

Note: “តនេះ” is used as a nominal token.

Example 34

Annotated Khmer: យុត្តិុត្តិ មើល ដែល តាម កូន នៅ អាយ ។

English gloss: I to-like book this.

English translation: I like this book.

Note: “តនេះ” is used as an adjectival token.

Example 35

Annotated Khmer: ម្ួយណាម្ួយ ម្ួយ ជា បស់ ឯង ។

English gloss: one which is thing you?

English translation: Which is yours?

Note: “ម្ួយណាម្ួយ” is a complex interrogative pronoun. “ឯង” is a pronoun used as adjective for possession.

Example 36

Annotated Khmer: ម្ួយ អាចម្ួយ ម្ួយ ជា បស់ ឯង ។

English gloss: bike one which is thing you?

English translation: Which bike is yours?

Note: “ម្ួយ” here is used as an adjective token.
- Example 37

Annotated Khmer: ណផ្ៃ [n សាស្រ្វ_n] n ្ៃ_a ្ៃ_v ្ៃ_o ្. 

English gloss: fruit mango this to-ripen already .

English translation: Mangos ripen.

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

- Example 38

Annotated Khmer: សា វ យ ណផ្ៃ ្_o ប្រចើន_o ្. 

English gloss: mango to-bear-fruit much very .

English translation: Mango trees bear a lot of fruits.

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

- Example 39

Annotated Khmer: ប្រសូិ ទុុំ ប្រេិ ាន ែក_o ្. 

English gloss: rice ripe n/a already to-pull .

English translation: Ripe rice has been pulled.

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

17