Supplementary Instructions for Tokenization and Part-of-Speech Annotation Guidelines for Khmer (Cambodian) (Version 0.2, August 2018)

Chenchen Ding¹, Hour Kaing¹, Kamthong Ley¹,², Masao Utiyama¹, Eiichiro Sumita¹
¹Advanced Translation Technology Laboratory, ASTREC, NICT, Japan
²National 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

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>n-</td>
<td>general pronouns, including personal and demonstrative</td>
</tr>
<tr>
<td>v-</td>
<td>grammaticalized auxiliary verbs modifying other content verbs</td>
</tr>
<tr>
<td>a-</td>
<td>general functional noun-modifiers, including possessive and demonstrative</td>
</tr>
<tr>
<td>o-</td>
<td>general particles, e.g., negators, conjunctions, mood particles, etc.</td>
</tr>
</tbody>
</table>

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_n- _v[ v _n]v _n[n _n]n _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”, “គា ់”, “តើង”, and “តេ”.

- Example 2
  Annotated Khmer: តេះ_n- _v v ចុំណែក_ n ឯង_ a - ។.
  English gloss: this to-be part you.
  English translation: This is your part.
  note: “តេះ” is a demonstrative pronoun.
  Common demonstrative pronouns are “តេះ” and “មី”.

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- _v[ v_ n]v ភាសា_ n[ n ខែ_ n]n ។.
  English gloss: I to-get to-learn language Khmer.
  English translation: I learnt the Khmer language.
  note: “បា” is grammaticalized as a past tense marker.
  Common pre-positioned grammaticalized verbs are: “បាន”, “តើ”, and “បាន”.
- **Example 4**

  **Annotated Khmer:** ងនត្វីឱ្យខ្ញុំពិបាក។
  
  **English gloss:** she to-do to-let I to-suffer heart.
  
  **English translation:** She makes me unhappy.
  
  **note:** “ត្វី” is grammaticalized as a causative marker.

  Common post-positioned grammaticalized verbs are: “តៅ”, “ថាស”, “រង”, “ពេញ”, and “ធ្វើ”,

- **Example 5**

  **Annotated Khmer:** ការ្កាយផ្តល់ជុំួខ្ញុំ។
  
  **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 “ជុំួ” is in between them.

### 2.3. Usage of “a-” Tag

- **Example 6**

  **Annotated Khmer:** ផ្ទេះខ្ញុំសង់តលើវាលណរស។
  
  **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 “ផ្ទេះ”.

- **Example 7**

  **Annotated Khmer:** ប្រការស្តីៗៗម៉ូន្តីស្តីវាសាម្របាទរាយ។
  
  **English gloss:** mission this to-have importance very.
  
  **English translation:** This mission is very important.
  
  **note:** “ម៉ូន្តី” is a demonstrative adjective modifying “ប្រការស្តីៗៗ”.
2.4. Usage of “o-” Tag

- Example 8
  Annotated Khmer: តររេះ _o- ខ្ញុំ _n- មិ _o- ដឹង _v តេ _o-
  English gloss: because I not to-go n/a
  English translation: because I don’t know
  note: “តររេះ” is a subordinating conjunction.
  “មិ” is a particle for negation.
  “តេ” is a final particle for negated sentences.

- Example 9
  Annotated Khmer: តើ _o- អ្នក _n- សុខ _v [a បាំ_v] v តេ _o- _?.
  English gloss: n/a you happy to-be-happy n/a ?
  English translation: How are you?
  note: “តើ” is an interrogative particle.
  “តេ” is a final particle for interrogative sentences.

3. Modification for Multi-Token Compounds

<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>ស្វោ_n- [n- ខ្ញុំ_n-]n- she I → I (female)</td>
<td>gender</td>
</tr>
<tr>
<td>ស្វោ_n- [n- បាំ_o]n- I yes → I (male)</td>
<td></td>
</tr>
<tr>
<td>បូប_n-[n ក្រក_n-]n- oneself he → I</td>
<td>emphasis</td>
</tr>
<tr>
<td>បូប_n-[n បាំ_n]n- body he → he</td>
<td></td>
</tr>
<tr>
<td>អ្នក_n- [n- បាំ_n]n- you female → you (female)</td>
<td></td>
</tr>
<tr>
<td>អ្នក_n- [n- បាំ_n]n- you female → you (female)</td>
<td>politeness</td>
</tr>
<tr>
<td>អ្នក_n- [n- បាំ_n]n- you male → you (male)</td>
<td></td>
</tr>
<tr>
<td>ក្រក_n- [n ខ្ញុំ_n-]n- group I → we (exclusive)</td>
<td></td>
</tr>
<tr>
<td>ក្រក_n- [n បាំ_n]n- group we → we (inclusive)</td>
<td>plurality</td>
</tr>
<tr>
<td>អ្នក_n- [o បាំ_n]n- all this → these</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Examples of two-token compound auxiliary verbs

<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>រគា_v- {v- ខ្លួ_o}v-</td>
<td>to-be-enough only</td>
</tr>
</tbody>
</table>

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

<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>របសិ_o- {o- តបើ_o}o-</td>
<td>if if</td>
</tr>
</tbody>
</table>

4. Confusion Cases

- Example 10
  
  **Annotated Khmer:** តៅ_n- ិយា_v តាម_v- តា_n- ។.
  
  **English gloss:** you say follow I.
  
  **English translation:** You repeat after me.
  
  **note:** “តៅ” (grandson) and “តា” (grandpa) are nouns. Here they are used as pronouns for “you” and “me”.

- Example 11
  
  **Annotated Khmer:** តោក_n [n តដិី_n ោ ់ដឺសុ_n]n
  
  **English gloss:** Mr. David Andersen
  
  **English translation:** Mr. David Andersen
  
  **note:** “តោក” is a pronoun for “you”. Here it is used as a title.
Example 12

Annotated Khmer: ផ្ទេះ _n បស់ _o ខ្ញុំ _n-

English gloss: house of my

English translation: my house

note: “ខ្ញុំ” is considered as a pronoun.

“បស់” is a possessive marker.