Supplementary Instructions for Tokenization and Part-of-Speech Annotation Guidelines for Myanmar (Burmese)
(Version 0.1, March 2017)

Chenchen Ding¹, Hnin Thu Zar Aye¹,², Masao Utiyama¹, Win Pa Pa², Eiichiro Sumita¹
¹Advanced Translation Technology Laboratory, ASTREC, NICT, Japan
²University of Computer Studies, Yangon, Myanmar
chenchen.ding@nict.go.jp

1. Introduction

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

The basic NOVA tags used in preliminarily annotated Myanmar 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. The v- tags is not used, and general particles, case-markers, etc., are all covered by o-. Please refer to the section 5.2 in Tokenization and Part-of-Speech Annotation Guidelines for Myanmar (Burmese). A further “/o-” can be attached to n or n- to form tags as “n/o-” or “n-/o-” to annotate contracted genitive case-marker, which is a creak tone that cannot be detached in the process of tokenization.

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, demonstrative, interrogative, and numeral</td>
</tr>
<tr>
<td>a-</td>
<td>general determiners, most derived from n- as a direct modifier for n or n-</td>
</tr>
<tr>
<td>o-</td>
<td>general particles, case-markers, conjunctions, etc., i.e., functional o</td>
</tr>
<tr>
<td>/o-</td>
<td>combined with preceding tags for unbreakable contracted particles.</td>
</tr>
</tbody>
</table>

2. Modification Examples
The usage of “n-”, “a-”, “o-”, and “/o-” tags are illustrated in Table 2. For all the examples illustrated, the tags are attached to correspondent tokens by an underline (“_”).

Table 2. Examples of tokens annotated with the “n-”, “a-”, “o-”, and “/o-” tags.

<table>
<thead>
<tr>
<th>annotated Myanmar</th>
<th>English gloss</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>ၵိုင်-n-</td>
<td>he</td>
<td>personal pronoun</td>
</tr>
<tr>
<td>ၵိုင်-n-[n- ၏-o-]n-</td>
<td>they</td>
<td>“/&gt;&lt;/o-” is a plural marker</td>
</tr>
<tr>
<td>ၵိုင်-n[ n ၏-o-]n</td>
<td>children</td>
<td>“/&gt;&lt;/o-” is a plural marker</td>
</tr>
<tr>
<td>ၵိုင်-n- ၏-o-</td>
<td>at here</td>
<td>“/&gt;&lt;/o-” is a demonstrative pronoun</td>
</tr>
<tr>
<td>ၵိုင်-a- ၏-n</td>
<td>this chair</td>
<td>“/&gt;&lt;/o-” is a demonstrative adjective</td>
</tr>
<tr>
<td>ၵိုင်-n-/o-</td>
<td>his</td>
<td>pronoun with a contracted genitive case-marker</td>
</tr>
</tbody>
</table>

3. Specific Examples

Specific cases, which may arise confusion in annotation, is listed here.

- Example 1

Annotated Myanmar: ၵိုင်[ n ][ ၏-o- ]n

English gloss: daughter
English translation: daughters
note: a nominal constituent with a plural suffix, brackets used

Annotated Myanmar: ၵိုင် [ n ၏-ا[s [ n ] ] ]

English gloss: daughter two person
English translation: two daughters
note: a numeral constituent inserted into a noun and a plural suffix, no brackets for the whole constituent
- Example 2

**Annotated Myanmar:** အရ ရန်မေးနေ [n မေး- တို့]

**English gloss:** officer -s -s

**English translation:** officers

**note:** a nominal constituent with double plural suffixes, brackets used to cover all plural suffixes

- Example 3

**Annotated Myanmar:** အလမ်ကန်နင့််ရာ ပ [n နင့််- တို့]

**English gloss:** America and Europe -s

**English translation:** American and Europe

**note:** a plural suffix added after a coordinating structure, no brackets used

**Annotated Myanmar:** အတန်နင့််ဆွစ်ဇန်နင်ငန် [n နင့််- ဆွစ်ဇန်]

**English gloss:** Italy and Switzerland

**English translation:** Italy and Switzerland

**note:** a plural suffix not directly modifying a coordinating structure brackets used