Supplementary Instructions for Tokenization and Part-of-Speech Annotation Guidelines for Myanmar (Burmese)

(Version 0.1, March 2017)

Chenchen Ding[1], Hnin Thu Zar Aye[1, 2], Masao Utiyama[1], Win Pa Pa[2], Eiichiro Sumita[1]

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

[2]University 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

tag	Description
n-	general pronouns, including personal, demonstrative, interrogative, and numeral
a-	general determiners, most derived from **n-** as a direct modifier for **n** or **n-**
o-	general particles, case-markers, conjunctions, etc., i.e., functional **o**
/o-	combined with preceding tags for unbreakable contracted particles.

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.

annotated Myanmar	English gloss	Note
သူ_n-	he	personal pronoun
သူ_n-[n- တို့_o-]n-	they	"တို့" is a plural marker
ကလေး_n[n များ_o-]n	children	"များ" is a plural marker
သည်_n- မှာ_o-	at here	"သည်" is a demonstrative pronoun
သည်_a- ခုံ_n	this chair	"သည်" is a demonstrative adjective
သူ့_n-/o-	his	pronoun with a contracted genitive case-marker

3. Specific Examples

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

- Example 1

Annotated Myanmar: သမီး_n[n တို့_o-]n

English gloss: daughter -s

English translation: daughters

note: a nominal constituent with a plural suffix, brackets used

Annotated Myanmar: သမီး_n နှစ်_a[1 ယောက်_n]a တို့_o-

English gloss: daughter two person -s

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: အရာရှိ_n[n များ_o- တို့_o-]n

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: အမေရိကန်_n နှင့်_o- ဥရောပ_n တို့_o-

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: အီတလီ_n နှင့်_o- ဆွစ်ဇာလန်_n နိုင်ငံ_n[n တို့_o-]n

English gloss: Italy and Switzerland country -s

English translation: Italy and Switzerland

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