A Burmese (Myanmar) Treebank: Guideline and Analysis

CHENCHEN DING, ASTREC, National Institute of Information and Communications Technology, Japan
SANN SU SU YEE, WIN PA PA, and KHIN MAR SOE, University of Computer Studies,
Yangon, Myanmar
MASAO UTIYAMA and EIICHIRO SUMITA, ASTREC, National Institute of Information and
Communications Technology, Japan

A 20,000-sentence Burmese (Myanmar) treebank on news articles has been released under a CC BY-NC-SA license. Complete phrase structure annotation was developed for each sentence from the morphologically annotated data prepared in previous work of Ding et al. [1]. As the final result of the Burmese component in the *Asian Language Treebank Project*, this is the first large-scale, open-access treebank for the Burmese language. The annotation details and features of this treebank are presented.

CCS Concepts: • **Computing methodologies** → **Language resources**;

Additional Key Words and Phrases: Burmese (Myanmar), treebank, phrase structure

ACM Reference format:

Chenchen Ding, Sann Su Su Yee, Win Pa Pa, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2020. A Burmese (Myanmar) Treebank: Guideline and Analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 19, 3, Article 40 (January 2020), 13 pages.
<https://doi.org/10.1145/3373268>

1 INTRODUCTION

The Burmese (Myanmar) language, although referred to as a low-resource language with limited studies in the natural language processing (NLP) research field, has developed gradually in recent years. In a series of previous works [1, 2, 4], language processing on the Burmese morphological level has been established systematically. A large-scale open-access evaluation campaign for Burmese–English machine translation (MT) was consequently facilitated [7].¹

Based on the previous morphological-level studies, this study explores a deeper syntactic level. We completed a 20,000-sentence Burmese treebank for news articles and released the data under

¹For the latest WAT 2019 Myanmar–English parallel data, see <http://lotus.kuee.kyoto-u.ac.jp/WAT/my-en-data/>.

This work was performed when Sann Su Su Yee was an intern student at ASTREC, National Institute of Information and Communications Technology, Japan, from April 2018 to April 2019.

Authors' addresses: C. Ding, M. Utiyama, and E. Sumita, ASTREC, National Institute of Information and Communications Technology, 3-5, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan; emails: {chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp; Sann Su Su Yee, Win Pa Pa, and Khin Mar Soe, University of Computer Studies, Yangon, No.(4) Main Road, Shwe Pyi Thar Township, Yangon, Myanmar; emails: {sannsusuyee, winpapa, khinmarsoe}@ucsy.edu.mm.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2375-4699/2020/01-ART40 \$15.00

<https://doi.org/10.1145/3373268>

a CC BY-NC-SA license [3],² as the Burmese component of the *Asian Language Treebank (ALT) Project* [11]. This is the first large-scale, open-access treebank for the Burmese language. This article provides the details for this treebank. Section 2 provides a general overview on treebank development and related work on Burmese language processing. Section 3 provides an overview and the annotation guidelines for the treebank. Section 4 provides statistics for the treebank to illustrate its numerical features. Our conclusion and future work are provided in Section 5.

2 RELATED WORK

2.1 Treebank

A treebank is generally considered to be a text corpus with syntactic annotation. A well-known treebank project is the *Penn Treebank* for English [6]. Large-scale treebanks have also been developed successfully for various Asian languages, e.g., the *Penn Chinese Treebank* [15] and the Japanese *Kyoto University Text Corpus* [5]. Different treebanks may apply different syntactic formulations, depending on the features of the language and the purpose of the corpus. Generally, constituency and dependency are two mainstream grammars for syntactic tree annotation. In constituency grammar, sentences are organized in a hierarchical structure of phrases, where linguistic categories for phrases are assigned. The aforementioned *Penn treebanks* for English and Chinese apply constituency grammar. However, the dependency grammar directly organizes sentences on a word level by categorizing the *dependency relationship* of word pairs. Thus, dependency grammar is mainly used for languages with a relatively free word order, where the phrase structure is not rigid. Japanese is a typical language that applies dependency grammar in treebank construction. Furthermore, dependency grammar is applied in the *Universal dependencies*³ cross-lingual annotation framework.

2.2 Burmese Processing

To the best of our knowledge, real data-based, supervised studies at Burmese syntactic level have not been performed in NLP. Under limited resources, some typical early work [13, 14] applied dictionary and manually edited context-free grammar (CRF) rules, but no further substantial progress has been made beyond these rule-based approaches. This work contributes significantly and substantially to the literature by providing a 20,000-sentence Burmese treebank, which can directly facilitate practical Burmese syntactic analysis and downstream NLP tasks that take advantage of syntax.

This article is the final step of a series of our work in developing Burmese linguistic data. The early development of the Burmese ALT treebank was reported previously Ding et al. [1]. As a brief retrospective summary, the development follows:

- Data preparation and preliminary annotation from 2015.
- Systematic morphological annotation from 2016 to 2017.
- Systematic syntactic annotation from 2017 to 2019.

Considering studies of Burmese NLP, our early investigation into Burmese tokenization [4] revealed problems and ambiguities in Burmese morphological data. Therefore, we developed an annotation system called *nova* [2] for consistent morphological annotation and systematically applied it to Burmese [1]. Briefly, tags of *n*, *v*, *a* were used for *nominal*, *verbal*, and *adjectival* tokens. An overall *o* tag was used for all other content tokens, such as *adverbial* ones. The corresponding

²Also accessible at <http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/my-alt-190530.zip>.

³<https://universaldependencies.org/>.

n- and a- tags were used for *pronominal* and *determinative* tokens, respectively, and o- for general functional affixes and particles. A pair of brackets [and] were used to cover tokenization ambiguities on compounds and derivations. More details were discussed previously by Ding et al. [1, 2] and in the released guidelines.⁴

3 ANNOTATION

3.1 Overview

Constituency grammar is applied in the Burmese ALT treebank. One realistic reason is that the standalone tokens used for dependency grammar annotation cannot be identified trivially, because of the implicit word boundary in Burmese orthography. However, by organizing Burmese sentences in nested phrases, a relatively consistent annotation in practice is more feasible.⁵

The morphologically annotated ALT data prepared in Ding et al. [1]⁶ were modified in the following ways during the syntactic annotation process in this work.

- The Burmese data were originally translated from English articles in *Wikinews*. Tens of sentences were lost in the morphological data because of errors in the translation process. These lost sentences were retranslated and added to the treebank. The total number of sentences increased from 19,965 to 20,106. A complete sentence-aligned parallel English–Burmese treebank is available.⁷ In addition to the newly added sentences, hundreds of Burmese sentences were also retranslated or partially modified.
- The POS tag has been systematically transferred from the nova system to the universal POS (Uni-POS) tagset [10].⁸ The tokenization and POS-tagging were further refined along with the syntactic tree annotation. The detailed guidelines and analysis of the POS tag transfer were discussed by Sann Su Su Yee et al. [12].

A subset of the Uni-POS tagset was adapted as the syntactic category for phrase structures. The lower- and uppercases were used to distinguish whether a tag was morphological or syntactic. This manner naturally integrated the two-layer morphological annotation into the syntactic tree, where the leaf nodes of lexical tokens will always be annotated by lowercase tags. Part of the internal nodes taking leaf nodes as children will be annotated by lowercase tags to tolerate the ambiguities at the morphological level. All other internal nodes used uppercase tags as syntactic annotations. Letting leaf nodes of lexical tokens in a syntactic tree at the height of 0, all lowercase nodes are restricted at heights of 1 and 2, and the uppercase nodes are always with a height of at least 2.

Guidelines for the syntactic annotation will be given in the next subsection. Figure 1 illustrates an example of an annotated syntactic tree based on the modification of the previous morphological data. The sentence and the annotation by nova below it are identical to that presented in Figure 3 of Ding et al. [2]. The transferred Uni-POS and syntactic tree are shown above the Burmese sentence. Considering the tokenization, the original token **12** was further segmented into a numerical stem and a suffix for counting instead of an overall cardinal attribute. The brackets for tokens **15** **16** **17** in the original morphological data were merged into the tree structure using a lowercase tag. More of these morphological parent nodes were addressed, although not annotated in the original

⁴<http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/Myanmar-annotation-guideline.pdf>.

⁵<http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/Myanmar-annotation-guideline-supplementary.pdf>.

⁶A preliminary version with rough phrase structure annotation can be accessed at <http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/Myanmar-ALT-20170110.zip>.

⁷<http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/my-nova-170405.zip>.

⁸The ALT English treebank: <http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/English-ALT-20170107.zip>.

⁹The original 12-tag version proposed by Google.

Fig. 1. Identical sentence in Figure 3 of Ding et al. [2]. The original English sentence is "He will officially become the Philippines's fifteenth president on June 30." Above the sentence is the Uni-POS tagging and the syntactic structure prepared by this study.

Table 1. Use of Morphological Tags (Uni-POS Tagset)

Tag	Description
adj	general adjectives, trivially from a in nova
adp	mainly for various case markers, nontrivially modified from o- in nova
adv	general adverb, trivially from o in nova
conj	general conjunctions, nontrivially modified from o- in nova
det	general determiners, trivially from a- in nova
noun	general nouns, trivially from n in nova
noun-adp	nouns with a contracted genitive case marker, trivially from n/o- in nova
num	general numbers, trivially from 1 in nova
part	various affixes (mainly suffixes), nontrivially modified from o- in nova
pron	general pronouns, trivially from n- in nova
pron-adp	pronouns with a contracted genitive case marker, trivially from n-/o- in nova
punct	general punctuation marks, trivially from . in nova
verb	general verbs, trivially from v in nova
x	a catch-all category, trivially from + in nova

Table 2. Use of Syntactic Tags

Tag	Description
ADJ	general adjectival constituents, distinguished from VERB by annotator's commonsense
ADV	general adverbial constituents, distinguished from NOUN by annotator's commonsense
NOUN	general nominal constituents
ROOT	exclusively for the root node of sentences
VERB	general verbal constituents, also for clauses rooted by verbal constituents

data, i.e., parent nodes for **3 4, 5 6 7, 9 10**, and for the segmented token **12**. We do not draw a strict line to distinguish between the morphological and syntactic nodes, while generally the “long tokens” in the original morphological data were reserved by morphological nodes. As the two kinds of nodes were only distinguished by the capitalization of tags, these differences can be easily utilized or omitted in specific downstream NLP tasks. The philosophy of syntactic annotation is consistent with that for morphological annotation: A flexible margin, i.e., a portion of nodes at height of 2, is introduced to tolerate ambiguities.

3.2 Guideline

3.2.1 Tagset. Tables 1 and 2 list the descriptions of the use of tags. Table 1 basically follows the design of the Uni-POS for morphological annotation. Generally, the o- tag in the nova system, which is an overall tag for all functional tokens, is further classified into three tags: adp (adposition), conj (conjunction), and part (particle) in the Uni-POS tagset. The part tag is used by default, unless there are clear reasons to use adp for case markers or conj for conjunctive particles. As for the adp tag, it also appears in hyphenated tags of noun-adp and pron-adp, which are used to adapt nouns with a contracted genitive case marker where the tokenization cannot be done. As a result, the part tag covers all verbal suffixes, a few nominal plural suffixes, and functional affixes that are difficult to classify. Many ambiguous cases arise in practical annotation, which were discussed by Sann Su Su Yee et al. [12]. Here we present an example of the tagging of functional tokens in Figure 2.

Fig. 2. Example for functional adp, conj, and part tags. The original English text was “... live with dementia and/or Alzheimer’s.” The **2** is a nominalization suffix, where the part tag is used. Note that the **1** is tagged as verb instead of adj and the parent node of **1 2** is a morphological noun node. As for **3** and **6**, the token is an instrumental/comitative case marker, which can be translated into English as “with” or “by”; however, it is more freely used as a coordinate marker corresponding to the English “and.” In this example, annotators consider **6** to be more like a case marker, but **3** is more like a conjunction (especially followed by **4**). Note that although **3** and **4** are tagged as conj, annotators still prefer to use them as suffixes and wrap **1 2 3 4** as a suffixed noun phrase. The annotation here is quite neutral in a flattened manner.

In an extreme analysis, Burmese sentences can be organized in nested noun and verb phrases.⁹ The syntactic tags NOUN and VERB in Table 2 are recommended to be used for consistent annotation. The ADJ and ADV tags are also used based on native speakers’ commonsense, which is according to the principle in the morphological annotation (i.e., the use of a and o tags in the preliminary annotation). However, ADJ and ADV are rarely used by annotators in practice, because the adj- and adv-tagged internal morphological nodes cover most cases and these nodes are directly merged into the upper NOUN and VERB nodes. An example is the tree-let of **12 13** in Figure 1. No tag for clauses was used in annotation. For a complete clause rooted in a verb phrase, the clause will be annotated consistently by VERB. An extra ROOT tag was used exclusively for the root node of the syntactic tree for each sentence.

3.2.2 Tree Structure. The syntactic structures are mainly organized into binary trees, with the following exceptional cases: (1) basic phrases, (2) coordinate structures, and (3) punctuation marks. In the example of Figure 1, all internal syntactic nodes of NOUN and VERB are binary branched, which is the basic principle of the tree annotation. The exceptional case (1) is mainly for the internal morphological nodes, such as the tree-lets of **5 6 7** and **15 16 17** in Figure 1. As for case (2), conjunctive tokens may be treated as suffixes in local and short phrases, as shown by the tree-let of **1 2 3 4** in Figure 2. In this situation, case (2) is a variety of case (1). More typical instances are coordinate structures of long clauses, where a flat multibranch tree structure will be applied.

⁹In the appendix of *Outline Grammatical Description* [8]: “The typical Burmese sentence can be analysed into one or more ‘noun phrases’ followed by a ‘verb phrase’ ... For brevity, we use the term ‘noun phrase’ as a cover term to include constituents of sentences that may be translated as adverbs or certain other types of expression.”

The case (3) is trivial; i.e., we generally do not use separate internal nodes for punctuation marks. As the **18** of Burmese period in Figure 1, punctuation marks are covered directly by higher nodes of sentences or clauses. The cases (2) and (3) usually appear simultaneously.

3.2.3 Constituent Identification. Burmese is a typical head-final language where most grammatical functions are afforded by suffixes. Constituents are built bottom-up from the basic morphological nodes. Generally, various functional suffixes are gradually attached to their preceding constituents to form larger constituents. The constituent tags are based on the tags of morphological stem; therefore, the NOUN and VERB syntactic tags are mostly used as stated in Section 3.2.1.¹⁰ Generally, the morphological nodes of adj, adp, det, noun, noun-adp, num, pron, and pron-adp will be covered by higher NOUN nodes, while nodes of adv and verb are covered by higher VERB nodes. Nodes of conj and part can be covered either by NOUN or VERB nodes.

On higher nodes indirectly connected with the local morphological nodes, VERB nodes govern the syntactic structure. Typically, a series of NOUN nodes as arguments of a following VERB node will be wrapped by a higher VERB nodes, and so forth the sentence is gradually organized. Some import issues related to VERB nodes are as follows.

- The identification of VERB nodes is primarily triggered by suffixes at morphological level, and secondarily by contexts at syntactic and semantic level.¹¹
- Normalization for verbal phrases can appear on syntactic level as well as on morphological level. On morphological level, the normalization has been covered by the noun nodes. On syntactic level, a VERB node covering a clause will be alternated into a NOUN node, if the root verbal phrase with the clause is marked by normalization suffixes.¹²

3.2.4 Special Issues. In the annotation, the adj/ADJ and adv/ADV tags are generally applied marginally. However, they are designated for some important uses. Specifically, (1) the postpositional attributes to nominal tokens will be annotated by adj, and (2) the negative particle *ma* in Burmese is annotated as adv.

The use of adj in case (1) has been mentioned previously [1, 2]; it mostly covers number-counter phrases and postpositional simple adjectival morphemes, both of which are special morphosyntactic structures in the general head-final principles of Burmese. Examples of this issue are shown in Figure 3. The negative particle *ma* may be referred to as a particle or a prefix under different analyses. It is consistently segmented and annotated as adv rather than part because of its relatively freer usage than common functional suffixes.¹³

4 STATISTICS AND ANALYSIS

Table 3 lists the distribution of lowercase morphological tags. The column of **height = 1** shows the distribution in POS-tagging. The total number is slightly different from the number of “short

¹⁰In the syntactic analysis of English, e.g., a preposition with a noun may form a prepositional phrase. From this perspective, the Burmese syntactic trees will be full of ADP phrases, which is quite counterintuitive for native-speaker annotators.

¹¹Generally, the morphological and syntactic features are adequate to reveal the verbalness in Burmese. As an extreme case, the imperative in Burmese is formed by a bare verb stem, where there is no suffix at morphological level and no argument at syntactic level; however, the verbalness can still be identified clearly by semantics.

¹²Note that the normalized verbal phrase itself still deserves a VERB (or verb) node within the clause, the normalization is reflected on the higher node covering the entire clause.

¹³Beside the prepositional *ma*, Burmese has another important prefix *a*. *a/ma* is an alternation in some structures to form positive/negative pairs, e.g. the superlative structure of “*a/ma* - adj - *hsoun*” for “*most/least* adj”. In such case of alternation, *a* is tokenized, because *ma* is tokenized all the time; part rather than adv will be used for *a* because of its “placeholder” role. In structures where *a/ma* do not form an alternation, *a* is not tokenized. An example is the passive structure of “*a* - verb - *khan*” for “*to be verb-ed*”. The negation form is not “*ma* - verb - *khan*”, but “*a* - verb - *ma. khan*”.

Fig. 3. The specific use of the `adj` tag. The original English for the left example is “a try.” Here **3** **4** comprise a postpositional number-counter phrase for the normalized verb **1**. The original English for the right example is “central clinical hospital.” Here **8** is a postpositional adjectival morpheme, but it is ambiguous whether **8** is modifying only **7** or the entire phrase of **5** **6** **7**; therefore, a multibranch flat structure was applied.

Table 3. Distribution of Morphological Tags

	height = 1		height = 2			
	frequency		frequency		branch	
					μ	(σ)
adj	10,132	(1.4%)	31,778	(21.3%)	2.4	(0.7)
adp	93,683	(12.7%)	772	(0.5%)	2.0	(0.0)
adv	13,183	(1.8%)	1,493	(1.0%)	2.2	(0.4)
conj	14,873	(2.0%)	603	(0.4%)	3.1	(0.4)
det	5,339	(0.7%)	0	(0.0%)	—	(—)
noun	189,275	(25.7%)	69,996	(47.0%)	2.3	(0.6)
noun-adp	628	(0.1%)	0	(0.0%)	—	(—)
num	22,862	(3.1%)	5	(0.0%)	3.0	(0.0)
part	199,694	(27.1%)	761	(0.5%)	2.9	(0.2)
pron	14,130	(1.9%)	4,783	(3.2%)	2.0	(0.1)
pron-adp	939	(0.1%)	0	(0.0%)	—	(—)
punct	56,101	(7.6%)	0	(0.0%)	—	(—)
verb	114,851	(15.6%)	38,851	(26.1%)	3.0	(0.9)
x	3	(0.0%)	0	(0.0%)	—	(—)
Total	735,693	(100.0%)	149,042	(100.0%)		

token” in Table 2 of Ding et al. [1] because of the modification of Burmese texts. The most used POS-tags are as follows: `part` (a little more than one quarter), `noun` (around one-quarter; if `pron` added, slightly more than `part`), `verb` (around one-sixth), and `adp` (around one-eighth). The four tags have covered around 80% of the tokens in the corpus. Considering the left 20% tokens, around one-third are punctuation marks (`punct`). The column of **height = 2** for internal lowercase tags is the margin between morphological and syntactical annotation. Generally, the `noun` and `verb` tags are mostly used to cover various expression of compounds and derivations. Note that the `adj`

Table 4. Derivation of Internal Morphological Tags

left side	right side			
	type	singleton	perplexity	top-min (3, perplexity)
adj ->	167	61	9	-> num part 34.4%
				-> verb part 24.1%
				-> verb part part 16.7%
adp ->	4	0	3	-> noun adp 45.9%
				-> adp part 41.1%
				-> adp adp 12.8%
adv ->	24	9	4	-> verb part 54.7%
				-> adj part 14.3%
				-> verb conj 13.2%
conj ->	23	9	4	-> adv verb conj 67.7%
				-> adv verb part 11.1%
				-> adv verb part conj 6.1%
noun ->	327	143	11	-> verb part 23.2%
				-> noun part 21.6%
				-> noun noun 19.8%
num ->	1	0	1	-> num punct num 100.0%
part ->	3	0	2	-> part verb part 58.0%
				-> part verb conj 35.5%
pron ->	5	1	1	-> pron part 99.3%
				-> verb part part 37.5%
verb ->	191	71	8	-> verb part 27.2%
				-> verb part part part 11.8%

tag is more often used when height is 2 rather than 1, because it is used more often in suffixed attributive expressions than in tagging single tokens. The branch of these internal morphological nodes at the height of 2, i.e., number of child nodes at the height of 1, is also listed in terms of average (μ) and standard deviation (σ) to show the distributions of the ranges they cover.

Table 4 shows more statistics and examples of the derivation rules for internal morphological nodes. Specifically, for each parent node, which are on the left side of the derivation rule, we list the number of total types on the right side, with the top-3 frequently observed examples, unless there are few varieties. The number of singletons and perplexity of the right side are also presented to depict the distribution. For the most used noun, verb, and adj tags, their right-hand distribution is quite long-tailed. For the use of adj tag, the most frequent pattern “-> num part” is the number-counter expression shown in Figure 3 (i.e., the left example) and the remainder are mostly suffixed verbal expressions. For the noun tag, the part in the pattern of “-> verb part” is generally for various nominalization suffixes, while that in the pattern of “-> noun part” is generally for plural suffixes. The verb tag is mainly used to wrap verbal stems and various suffixes.

Tables 3 and 4 present the annotation characteristics of Burmese morphology. Tables 5 and 6 further depict the syntactic organization in the annotation. In addition to the height and the branch, the depth for the distance between a node and the ROOT node, and the range of words a node covers¹⁴ are also shown. The NOUN and VERB tags are overwhelmingly applied as described in the

¹⁴Note that the range will be identical to the branch at height of 2 in Table 3, because these internal morphological nodes have no nested structure in annotation.

Table 5. Distribution of Syntactic Tags

	frequency		depth		height		branch		range	
			μ	(σ)	μ	(σ)	μ	(σ)	μ	(σ)
ADJ	6,983	(1.6%)	5.0	(2.0)	3.2	(0.8)	2.1	(0.2)	4.5	(2.2)
ADV	1,408	(0.3%)	4.6	(1.8)	2.7	(0.8)	2.1	(0.3)	3.2	(1.7)
NOUN	285,200	(65.4%)	4.2	(1.9)	3.6	(1.4)	2.1	(0.3)	5.2	(4.0)
ROOT	20,106	(4.6%)	0.0	(0.0)	8.7	(1.9)	3.5	(0.7)	36.6	(15.3)
VERB	122,399	(28.1%)	2.9	(1.7)	5.3	(2.3)	2.1	(0.3)	12.4	(10.9)
Total	436,096	(100.0%)								

Table 6. Derivation of Syntactic Tags

left side	right side			
	type	singleton	perplexity	top-3
ADJ ->	216	87	21	-> noun adj 27.9%
				-> adj adj 10.9%
				-> adv adj 10.5%
ADV ->	112	59	18	-> noun adv 31.6%
				-> adv adp 14.1%
				-> adv adv 8.3%
NOUN ->	707	272	20	-> NOUN NOUN 19.0%
				-> NOUN adp 15.2%
				-> noun adp 12.2%
ROOT ->	173	96	6	-> NOUN VERB punct 37.3%
				-> VERB VERB punct 19.6%
				-> NOUN punct VERB punct 17.7%
VERB ->	402	169	13	-> NOUN VERB 35.4%
				-> NOUN verb 11.9%
				-> VERB VERB 9.2%

guidelines described in Section 3.2. The average height of the ROOT node is 8.7, which is the average height of the syntactic trees in the whole treebank. The average range of ROOT is the average sentence length. Generally, the NOUN nodes are located lower than VERB nodes, with larger depths, smaller heights, and shorter ranges. The branch of all node types, except ROOT, are around 2, which is according to the basic binary annotation. Punctuation marks are usually covered directly by the ROOT node, which makes the average branch up to 3.5.

Table 6 shows more detailed information on the syntactic derivation rules. ROOT obviously always derives one or more punct tokens, which leads to a larger branching. For the right side of NOUN and VERB nodes, the NOUN can be observed to generate lowercase morphological nodes more often than VERB. As the height of morphological nodes will not be higher than 2, NOUN nodes are thus lower on average than VERB nodes in syntactic trees. Note that the marginally used ADJ and ADV tags mostly generate two morphological nodes; therefore, it is natural that the average heights of the two types of nodes are around 3. These two marginal tags also work as a further “cushion” over those morphological nodes at a height of 2; therefore, the structures of higher nodes are more purely organized by NOUN and VERB tags.

Table 7. Derivation Patterns

left side	right side	
	derivation pattern	coverage
Noun ->	(Det)*(Noun Pron Verb Adj Num)+(Part)*(Adp)*(Conj)*	93.3%
Verb ->	(Noun)*(Adv)*(Verb Adj)+(Part)*(Conj)*	91.4%
Adj ->	(Adv)*(Noun Verb Adj Num)+(Part)*(Conj)*	94.6%
Root ->	((Conj Punct)*(Noun Verb Adj)+(Conj Punct)*)+	99.7%

Table 8. F-score on Constituent Identification on Development (dev.) and Test Sets

	overall (neutralized)	span ≤ 20 (neutralized)	20 < span ≤ 40	40 < span
dev.	92.6 $\pm$ 0.1 (93.2 $\pm$ 0.1)	92.9 $\pm$ 0.1 (93.6 $\pm$ 0.0)	80.5 $\pm$ 0.8	90.0 $\pm$ 0.9
test	90.3 $\pm$ 0.1 (91.2 $\pm$ 0.1)	90.7 $\pm$ 0.1 (91.6 $\pm$ 0.1)	77.6 $\pm$ 0.8	87.8 $\pm$ 0.9

The morphological and syntactic tags are not distinguished in **neutralized**. As morphological tags are only applied for local nodes, the neutralized results do not affect constituents with large spans.

Common derivation patterns are shown in Table 7, where the uppercase syntactic and lowercase morphological tags are not distinguished, but are denoted in a neutralized form by capitalizing the initial letter. The right side of the derivation was organized in a form of regular expression based on a commonsense understanding of Burmese linguistic features.¹⁵ The right-side patterns have covered more than 90% derivation cases for each left side, while all patterns have covered 91.5% of the derivation instances in the treebank.¹⁶ Table 7 shows that the derivation in the treebank was basically performed according to linguistic intuitions.¹⁷

In terms of NLP application, the Burmese ALT treebank facilitates practical automatic Burmese parsing. Here we report benchmark results of parsing experiments. The entire Burmese ALT treebank is split into training/development/test sets in the same manner of Ding et al. [1]. The BerkeleyParser 1.7 [9] modified by NICT¹⁸ was used as an off-the-shelf tool. The binarization in model training was set to LEFT and all the other options were by default settings. The grammar model was trained on training and development data, and the parsing performance was investigated on development and test sets, i.e., on closed and open data, respectively. Eight models were trained separately with different random seeds. The average and standard deviation of the F-score on constituent identification are listed in Table 8. It can be observed that both the performance on the development data and on the test data are stable. The overall F-score on the test set is over 90%, and the difference between the closed/open data is around 2%. Generally, short and long constituents can be identified more accurately than those in-between ones. This is in accordance of human intuition. Comparing the numerical results of experiments on English Penn treebank in Petrov et al. [9], we consider the quality of Burmese ALT treebank is adequate for NLP practical application.

¹⁵ ()+ for obligatory components; ()* for optional components; and | for multiple options.

¹⁶ For Noun, the right-side patterns cover 331,398 derivation instances, $93.3\% = \frac{331,398}{69,996+285,200}$, where the denominator is the sum of the frequency of noun at a height of 2 in Table 4, and the frequency of NOUN in Table 6. Similarly, $91.4\% = \frac{147,434}{38,851+122,399}$ for Verb, $94.6\% = \frac{36,656}{31,778+6,983}$ for Adj, and $99.7\% = \frac{20,045}{0+20,106}$ for Root, and the overall $91.5\% = \frac{331,398+147,434+36,656+20,045}{149,042+436,096}$.

¹⁷ Punct of punctuation marks may also appear at the right side of Noun, Verb, and Adj. For brevity, Punct-contained derivations are only counted for Root.

¹⁸ <http://www2.nict.go.jp/astrec-att/member/mutiyama/software.html>.

5 CONCLUSION AND FUTURE WORK

In this article, we introduced the newly released 20,000-sentence Burmese ALT treebank. The annotation guidelines and statistics were presented to show the treebank's characteristics.

This work on the Burmese language is instructive for data development and NLP research on other low-source languages. Our work on developing the Khmer language is undergoing in addition to two directions for future work on Burmese (1) to edit large-scale dictionaries for proper nouns and terminologies and (2) to develop large-scale clean monolingual data. Combined with the treebank, it (1) can contribute better analyses of Burmese textual data in practice and (2) can contribute directly to MT for Burmese in the near future.

ACKNOWLEDGMENTS

We thank all of the translators and annotators who contributed to the construction of the Burmese ALT treebank in recent years. We thank Dr. Zaw Myint, Director General of the *Department of Myanmar Nationalities' Languages, Ministry of Education, Myanmar*, and Dr. Mie Mie Thet Thwin, rector of the *University of Computer Studies, Yangon*, for their help in verifying the annotation. We thank anonymous reviewers for their insightful comments.

This work was conducted under the program “*Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology*” of the *Ministry of Internal Affairs and Communications, Japan*.

REFERENCES

- [1] Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2019. Towards Burmese (Myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. *ACM Trans. Asian Low-Resource Lang. Inf. Process.* 19, 1 (2019), 5.
- [2] Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2018. NOVA: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. *ACM Trans. Asian Low-Resource Lang. Inf. Process.* 18, 2 (2018), 17.
- [3] Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2019. *Burmese (Myanmar) Treebank of Asian Language Treebank Project*. Retrieved from DOI: <https://doi.org/10.5281/zenodo.3463010>
- [4] Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, and Eiichiro Sumita. 2016. Word segmentation for Burmese (Myanmar). *ACM Trans. Asian Low-Resource Lang. Inf. Process.* 15, 4 (2016), 22.
- [5] Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the Annual Language Resources and Evaluation Conference (LREC'02)*. 2008–2013.
- [6] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Comput. Linguist.* 19, 2 (1993), 313–330.
- [7] Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. Overview of the 5th workshop on Asian translation. In *Proceedings of the 5th Workshop on Asian Translation (WAT'18)*. 1–41.
- [8] John Okell and Anna Allott. 2001. *Burmese/Myanmar Dictionary of Grammatical Forms*. Routledge.
- [9] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL'06)*. 433–440.
- [10] Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Annual Language Resources and Evaluation Conference (LREC'12)*. 2089–2096.
- [11] Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian language treebank. In *Proceedings of the Oriental International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques Conference (O-COCOSDA'16)*. 1–6.
- [12] Sann Su Su Yee, Chenchen Ding, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2019. Modifying NOVA-annotated Myanmar data to universal part-of-speech tagset. In *Proceedings of the International Conference on Computing Advancements (ICCA'19)*. 230–237.
- [13] Soe Lai Phye and Aye Thida. 2013. Unknown word detection via syntax analyze. *IAES Int. J. Artif. Intell.* 2, 3 (2013), 107–116.

- [14] Win Win Thant, Tin Myat Htwe, and Ni Lar Thein. 2012. Parsing of Myanmar sentences with function tagging. *Int. J. Nat. Lang. Comput.* 1, 1 (2012), 9–27.
- [15] Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Nat. Lang. Eng.* 11, 2 (2005), 207–238.

Received June 2019; revised October 2019; accepted November 2019