NOVA: A Feasible and Flexible Annotation System for Joint Tokenization and Part-of-Speech Tagging

CHENCHEN DING, MASAO UTIYAMA, and EIICHIRO SUMITA,

National Institute of Information and Communications Technology, Japan

A feasible and flexible annotation system is designed for joint tokenization and part-of-speech (POS) tagging to annotate those languages without natural definitions of *words*. This design was motivated by the fact that *word separators* are not used in many highly analytic East and Southeast Asian languages. Although several of the languages are well-studied, e.g., Chinese and Japanese, many are understudied with low resources, e.g., Burmese (Myanmar) and Khmer. In the first part of the article, the proposed annotation system, named *nova*, is introduced. *nova* contains only four basic tags (*n*, *v*, *a*, and *o*); these tags can be further modified and combined to adapt complex linguistic phenomena in tokenization and POS tagging. In the second part of the article, the feasibility and flexibility of *nova* is illustrated from the annotation practice on Burmese and Khmer. The relation between *nova* and two universal POS tagsets is discussed in the final part of the article.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Phonology/morphology*;

Additional Key Words and Phrases: Annotation, tokenization, part-of-speech tagging, Asian languages

ACM Reference format:

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-Resour. Lang. Inf. Process.* 18, 2, Article 17 (December 2018), 18 pages.

<https://doi.org/10.1145/3276773>

1 INTRODUCTION

Tokenization and part-of-speech (POS) tagging are two of the most basic tasks applied to raw textual data in sophisticated natural language processing (NLP) applications. For most Indo-European languages, the tokenization process is relatively trivial because their writing systems already use spaces to separate *words*, which can be used directly as *tokens*.¹ Thus, the tokenization process for these languages mainly handles punctuation marks, compounds, abbreviations, and capitalizations [11, 14]. For these languages, the POS tagging process is usually regarded as a separate task from tokenization. However, in East and Southeast Asia, most languages do not always use

¹In this article, we use “word” for conventionally segmented units and “token” for artificially segmented ones. The term of “morpheme” used in the article stands for its original linguistic meaning, i.e., smallest meaningful unit of a language. Hence, a morpheme is always a unit no larger than a word or a token.

Authors’ addresses: C. Ding, M. Utiyama, and E. Sumita, 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.

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.

2018 Copyright is held by the owner/author(s). Publication rights licensed to ACM.

ACM 2375-4699/2018/12-ART17 \$15.00

<https://doi.org/10.1145/3276773>

explicit separators to separate words, as the definition of words is not a natural concept. Typical examples are Chinese, Japanese, Burmese (Myanmar), Khmer, Laotian, and Thai. Therefore, tokenization and POS tagging of these languages are inseparable because the definition and principle of one task unavoidably affect the other. Taking well-studied Chinese and Japanese as examples, the detailed segmentation guidelines of the *Penn Chinese Treebank* is unavoidably based on its POS system [15–17]; there is a two-volume, nearly 400-page manual [7] based on morphological analysis to cover the two tasks for *Balanced Corpus of Contemporary Written Japanese*.

A universal POS tagset for cross-lingual processes has been initially proposed by researchers from *Google* [12]. Their inductive methodology is based on observations of 25 different tagsets of 22 different languages. The original POS tags of the studied languages are deterministically mapped to one certain *Google* universal (G-UNI) POS tag. Finally, the G-UNI tagset contains 12 different coarse tags. An extended version of the G-UNI tagset used in the *Universal Dependencies* (UNI-DEP) project² contains 17 tags [6], where specific sub-classes are further identified from the original 12 tags. As mentioned, tokenization and POS tagging cannot be separated when processing languages without a natural definition of words; a problem of the G-UNI tagset is that it completely covers the details in tokenization. The problem is not serious in the work of the G-UNI tagset, because most of the languages it studied are Indo-European, with several well-studied East Asian languages, i.e., Chinese, Japanese, and Korean, for which tokenization standards and corresponding POS tag systems have already been practically and solidly established. However, the G-UNI tagset is not flexible to be applied to many understudied Southeast Asian languages of our interest, e.g., Burmese and Khmer. Because tokenization standards for practical NLP have not been studied and established on these languages, directly applying a tagset like G-UNI to them will cause confusion and inconsistencies among native speakers. Considering the case of above-mentioned Chinese and Japanese, this is equivalent to omitting details that should be clarified by manuals with hundreds of pages. We thus intend to build a feasible and flexible annotation system to adapt these understudied languages for joint tokenization and POS tagging.

In this article, we propose an annotation system called *nova* for joint tokenization and POS tagging. *nova* uses a compact basic tagset of only four tags: *n*, *v*, *a*, and *o*, roughly representing four basic word classes: *noun*, *verb*, *adjective*, and *adverb*, respectively. The tags can be further modified and combined by extra marks and brackets to form complex tags and tag sequences to represent more informative annotations. Consequently, *nova* can use its combinability to handle ambiguities and difficulties in tokenization and POS tagging. Furthermore, the exchangeability between *nova* and the G-UNI tagset is retained to a large extent. Thus, *nova* is designed to accommodate for those understudied languages without a natural definition of words; and, it is nevertheless universally applicable and can provide a fundamental and efficient annotation for further sophisticated annotation using G-UNI, UNI-DEP, or more language-specific tagsets.

The remainder of the article mainly contains three parts. Descriptions of *nova* are provided in Section 2. In Section 3, we provide linguistic discussions to illustrate the feasibility and flexibility of the *nova* system through our practice of annotating two typical Southeast Asian languages: Burmese and Khmer. The relation between *nova* and the G-UNI tagset is discussed in Section 4. Conclusions are provided in Section 5.

2 NOVA ANNOTATION

2.1 Motivation

In contrast to most of the Indo-European languages, word separators are seldom (or inconsistently) used in different writing systems in East and Southeast Asia. Specifically, word separators are

²<http://universaldependencies.org/>.

completely absent in the Chinese-character based writing systems used by Chinese and Japanese. Many Southeast Asian languages using Brahmi scripts, which are abugidas, only apply spaces inconsistently for easier reading when necessary, e.g., spaces are used to separate long phrases in Burmese, and are more seldom used in Khmer to separate long clauses. Korean and Vietnamese are two exceptional languages in East and Southeast Asia, which regularly use spaces in their writing systems. In Korean, a main reason is the abandoning of Chinese characters in the orthography, where spaces are required for easier reading. In Vietnamese, the spaces are used to separate phonetic syllables rather than morphologically meaningful units. Consequently, Korean and Vietnamese still need a non-trivial tokenization process compared with Indo-European languages.

An underlying reason for the phenomenon can be attributed to the types of languages. Most Indo-European languages are fusional languages with inflections represented by *bound morphemes*, which indicate the boundary of meaningful units. Hence, the concept of words is relatively natural in these languages, and the role of words within a sentence is relatively independent and easy to figure out. In contrast, many Asian languages are highly analytic, i.e., Chinese, Vietnamese, Burmese, Khmer, Laotian, and Thai. They overwhelmingly use *free morphemes*. Consequently, the boundary of meaningful units is non-intuitive in these languages and the role of these free morphemes is much more affected by their contexts. Though bound morphemes in some agglutinative Asian languages, i.e., Japanese and Korean, are clear, the nature of agglutinateness causes long sequences of bound morphemes, on which the segmentation principles turn not clear.³ Furthermore, the POS system is not a less confusing and totally detachable issue regarding the concept of words.⁴ Although they vary by language, *noun* and *verb* are considered as two ultimate and universal classes,⁵ and *adjectives* may be taken as a third class [2].

Based on the above-mentioned linguistic grounds, we consider a feasible and flexible annotation system for NLP practice that should include the following three features.

- (1) a compact tagset for cross-lingual word classes, e.g., nouns and verbs
- (2) extensibility to language-specific word classes, e.g., pronouns, articles, and the like
- (3) flexibility to the concept of words, i.e., can tolerate and represent the ambiguities

Therefore, nova is motivated by these three features and related descriptions are provided, respectively, in the following three sections.

2.2 Basic Tags

We design four basic tags to represent fundamental word classes with three auxiliary tags to represent numbers (1), punctuation marks (.), and tokens with weak syntactic roles (+), e.g., interjections and symbols. Specific descriptions of the basic and auxiliary tags are listed in Table 1.

As nouns and verbs are two indisputably cross-lingual word classes, we use *n* and *v* tags to represent them. To facilitate exchangeability with the G-UNI tagset, we also use a tag to represent adjectives because it is a common (though not universal) word class. The *o* tag is defined in a negative way, which can be interpreted as a non-*n*, non-*v*, and non-*a* tag for all the modification or complement tokens.

In order to provide an intuitive illustration, we show a virtual example that assumes that English is a language that never uses spaces in its writing system. Then, for a piece of text such as:

oncewedreamtthatwewerestrangers.wewakeuptofindthatweweredeartoeachother.

³Classic discussions can be referenced in 11.6 of Bloomfield [1].

⁴A joint discussion is provided in Chapter 16 “Form-Classes and Lexicon” of Bloomfield [1].

⁵As discussed in 12.13 of Bloomfield [1].

Table 1. Basic and Auxiliary Tags in nova

Tag	Description
n	general <u>n</u> ouns, can be subjects or objects of tokens tagged by v
v	general <u>v</u> erbs, can take tokens tagged by n as arguments
a	general <u>a</u> djectives, can directly describe or modify tokens tagged by n
o	other modifications or complements, for which n, v, and a tags are not applicable
1	general numbers
.	general punctuation marks
+	a catch-all category, for tokens with weak syntactic roles

we may obtain the following annotated results using the basic tags.⁶

once_o we_n dream_v t_o? that_o we_n were_v stranger_n s_a? ._. we_n wake_v up_o? to_o
find_v that_o we_n were_v dear_a to_o? each_a other_n ._.

However, there may be difficulties in annotating, which are marked by question marks in the example, assuming the annotators had only a limited dictionary containing base forms and simple idioms. It may be difficult to determine whether some very small parts (i.e., *t* and *s*) or an idiom (i.e., *wake up*) should be segmented, and whether the two *to*-s should be tagged identically (the roles seem different). Because the expressiveness of the basic tags is limited, we design modification and combination schemes in nova to make annotations more flexible and expressive.

2.3 Modification

The basic tags can be modified by attaching a minus sign “-” to represent functional word classes that play a similar role to the corresponding original tags. As functional words vary between languages, we prefer to annotate them in a derived manner rather than design independent tags. For example, *pronouns*, common in Indo-European languages, can be represented by n-, considering they can take the place of nouns. In addition, *prepositions*, which appear in many subject-verb-object languages, can be represented by v- considering that, similar to verbs (but without inflections), they take nouns as arguments. For the same reason, a *determiner* or an *article* of a noun can be annotated as a-, and a *particle* of a verb can be tagged as o-. Thus, with the modifications, the previous example can be further annotated as follows.

once_o we_n- dream_v t_o-? that_o- we_n- were_v stranger_n s_a-? ._. we_n- wake_v
up_o-? to_o- find_v that_o- we_n- were_v dear_a to_v- each_a- other_n- ._.

The differences between pronouns and nouns, or between prepositions and verbs, are clear in English. However, the differences between functional and content words (morphemes) are not so obvious in many highly analytic languages. We believe the modification scheme is effective because language-specific word classes are connected and represented by those basic cross-lingual word classes in a flexible manner, which can be utilized (or omitted) depending on the requirements.

In practice, a modified basic tag (especially v-) can be represented by the o tag for simplicity; in cases, we prefer to use the extra modification to address the concept of *functional*, which can provide sub-categories for functional tokens such as v-, a-, and o-.⁷ An important reason is the

⁶In the presentation of this article, tags are attached to tokens by an underline.

⁷Thus, a bare o tag can be used for *adverbs*, from the description.

consideration of the exchangeability for G-UNI tagset, which contains relatively detailed tags for functional words, which will be discussed in Section 4.

2.4 Combination

We further design a mechanism that allows the tags described in Sections 2.2 and 2.3 to be combined to form more complex and expressive tags and tag sequences, in order to provide a flexible adaption of tokenization ambiguities. The combination is annotated in two ways.

- (1) a pair of brackets “[” and “]” to show one or multiple tags “*working (together) as*”
- (2) a slash “/” to show *non-detachable components* within one token

When aiming to annotate a language lacking a clear definition of words, tokenization can initially be conducted under simple and clear principles largely based on the sense of native speakers; subsequently, the brackets can be used to handle diverse and complicated exceptions. The philosophy here is the same as that of the modification: the ambiguities in tokenization are mainly afforded by the brackets. If the tokenization cannot be conducted because of special usages, the slash can be used for more detailed annotation. With the combination scheme, we can produce a more informative annotation of the example, as follows.

once_o we_n- dream_v[v t_o-]v that_o- we_n- were_v/o- stranger_n[n s_a-]n ._. we_n- wake_v[v up_o-]v to_o- find_v that_o- we_n- were_v/o- dear_a to_v- each_a- other_n- ._.

It can be seen that the suffixes (i.e., *t* and *s*), irregular or merged forms (i.e., *were*), and idioms (i.e., *wake up*) are covered with no difficulty.

In the example and the following examples illustrated in this article, the working-together-as tag is added after the right bracket (]) and before the left bracket ([) in pairs. The annotation is thus similar to the BIE notation used in many chunking or named entity recognition tasks in NLP, where shallow, two-layered information is restored. The example is, after all, a virtual and carefully annotated one, which may not always be required and achieved in practice. On the other hand, if we always pursue a scrupulous annotation, nested brackets may appear even in simple expressions, such as *go_v[v[v ing_o-]v on_o-]v*, which reserves too deep and detailed information regarding the common sense of POS tagging. The example here is only used to demonstrate nova’s flexibility and precise annotation ability. A conservative and temperate application of the brackets should always be considered in practice.

More properly than the English example illustrated here, the slash is reserved for phenomena of contractions, where an unbreakable token actually cannot be described by a single tag. A typical example is the contraction between prepositions and articles, which is common in many Indo-European languages. So, a tag as *v-/a-* can be conveniently used for *am*, *im*, *vom*, *zur*, and *zum* in German, or *au*, *aux*, *du*, and *des* in French. This mechanism can also be applied for annotating informal colloquial expressions, such as *gonna_v/o-* (*going to*), *wanna_v/o-* (*want to*), *wanna_v/a-* (*want a*), *kinda_n/o-* (*kind of*), *coulda_v/v* (*could have*), and so forth.

3 NOVA IN PRACTICE

3.1 Two Typologically Typical Languages

We are undertaking the *Asian Language Treebank* (ALT) project since 2015 [13], including tokenization, POS tagging, phrase structure tree building, and token alignment with English translations over several low-resource Southeast Asian languages. The first two involved languages are Burmese and Khmer. Both of them are highly analytic in morphology, which is a common feature in Southeast Asian languages. Syntactically, Burmese is a strong head-final language, which is a

relatively rare type in Southeast Asia, where most languages have a pure head-initial structure; examples include Khmer, Laotian, Thai, and Vietnamese. Therefore, we consider that Burmese and Khmer, two typologically typical languages that can represent Southeast Asian languages to a considerable extent. We believe that the practical application of nova on these two languages is also applicable to Laotian, Thai, and Vietnamese, as they share certain similar linguistic features.

For the grammar outlines of the two languages, well-edited Burmese grammar in English is provided by Okell [8, 9], and Okell and Allott [10]; an English introduction to Khmer grammar is provided by Ehrman et al. [3]. Traditional analyses distinguish *verbs* and *nouns* for Burmese and Khmer, but not always for *adjectives*, which are sometimes classified as (stative) verbs. *Adverbs* are also marginal, which may be considered as a special usage of nouns or verbs. All functional morphemes are generally classified as *particles* (or *relators*), which have complicated sub-categories such as nominal, verbal, phrasal, and sentential. Because of the highly analytic features, many particles are actually grammaticalized nouns and verbs, which make the boundary between functional and content morphemes not so clear in some cases. Most morphemes in the two languages can be combined freely to form *stable expressions*. Different combination patterns have different levels of strength, and special linguistic phenomena may break in such combination patterns. As a result, the sense of tokenization (and the POS tag of some tokens) is not consistent among native speakers. We show examples of simple Burmese and Khmer sentences in Figures 1 and 2 for intuitive illustration.

As shown in the examples, a morpheme-level analysis is too detailed to represent the sense of native speakers. Thus, *constituents* (or *phrases*, *slots* in the literature) are more commonly used in practice for syntactic and semantic analyses of these languages. An important reason is that constituents are easy to recognize within a sentence, and complicated compositional and grammaticalization phenomena within constituents do not necessarily need to be figured out so clearly. On the other hand, for NLP, the constituent units are too large to be taken as tokens and should be segmented further. However, the concept of words, which is a unit between the two clear extremes of morpheme and constituent, is not always consistent among native speakers. Therefore, we consider intuitive and operational principles of words (i.e., artificial tokens for NLP practice) to be the most important issue in corpus construction for these Southeast Asian languages.

Detailed guidelines on tokenization and POS tagging of Burmese and Khmer using nova have been released with annotated data.⁸ In Section 3.2, we first introduce typical examples to show analytic features of the two languages. In Section 3.3, descriptions of using nova to annotate the two languages are presented, which is a summary and explanation of the released guidelines. Statistics and examples from released annotated data are provided in Section 3.4 to show how nova accommodates linguistic phenomena in the two languages.

3.2 Typical Analytic Features

Head-Initial Nominal Structure in Burmese

As mentioned, Burmese has a general head-final order, in which attributes are mostly placed before the nouns they modify. However, a small set of simple adjectival morphemes can be directly added after nouns. These adjectival morphemes are overwhelmingly single-syllabled and frequently used. Native speakers may not achieve an identical sense of this special structure because only a portion of frequently used expressions have become stable. Figure 3 shows unstable and stable examples. The tokenization and POS tagging around this phenomenon are inconsistent that stable expressions may not be segmented, and a frequently used post-positioned adjectival

⁸ Accessible at the homepage of ALT project: <http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/>. The releasing is under a CC BY-NC-SA 4.0 license.

Fig. 1. A Burmese sentence that means “on the morning of the next day the present writer received two letters” (p. 281 in Okell and Allott [10]). The segmented units are morphemes, attached with English glosses. The gray blocks stand for the conventional positions of using spaces in writing. The three underlined and other single morphemes are recognized as “words” by native speakers.

Fig. 2. A Khmer sentence that means “Osaka is the second largest city in Japan.” The morphemes are spaced and attached by English glosses. Spaces are used less frequently in Khmer’s writing system, so the sentence actually does not have any space when written. The gray block-inserted positions may be considered as boundary of “phrases” and the underlined groups and other single morphemes are recognized as “words.” Notice that the first morpheme in the sentence can be both a single word or a component in a combined expression.

Fig. 3. A typical post-positional, adjectival morpheme in Burmese, with a meaning of *great* or *large*. This morpheme can be used as an ordinary adjectival stem but can also be added after nouns freely. The left example of “large sun-glasses” is an unstable expression and the right one of “adult” is a stable expression. For consistency, we prefer to harmoniously annotate both to two tokens with the help of brackets.

Fig. 4. A Khmer expression consisting of a nominal and a verbal morpheme, which can be explained either endocentrically as *waterfall* (i.e., *water which falls*) or exocentrically as *water falls*. However, the expression is still tended to be considered as two tokens by native speakers, even when it means *waterfall*.

morpheme may be considered as a particle when segmented. In our solution, we prefer to segment out these postpositional morphemes and to tag the sequences as $n[n \ a]n$, which clearly and naturally illustrates the components as well as the combination. The annotation is consistent and provides sufficient information on this structure.

Endo- and Exo-Centric Nominal-Verbal Structure in Khmer

Khmer has a strict subject-verb-object and head-attribute word order, but lacks inflections of verbs. An important result of this feature is that the relation of a bi-gram consisting of a noun and a verb can not be determined using only the POS tag; i.e., the relation can be either exocentric (subject-predicate) or endocentric (head-attribute), as illustrated in Figure 4. The ambiguity rarely occurs with native speakers because of contexts and common sense knowledge, but an overall verb tag is not sufficient for tagging because very important structure information is lost. However, using brackets under the framework of nova can provide clear and unified tagging. For the example in Figure 4, we annotate it as $n[n \ v]n$ if it means *waterfall*, and as $n \ v$ if it means *water falls*.

Complex Verbal Constituents

In both Burmese and Khmer, a strong tendency is to use simple nominal and verbal morphemes to form complex verbal expressions. The meaning of morphemes in such an expression is retained completely in some cases, but the meaning of morphemes playing an auxiliary role may turn vague and nearly functional. However, it is difficult to draw a clear line for different levels of grammaticalization. Examples of Burmese and Khmer are shown in Figures 5 and 6, respectively. According to our principles, we use nova to annotate the two expressions in Figure 5 as $v[n \ v]v$ and

Fig. 5. A Burmese verbal expression consisting of a nominal and a verbal morpheme. The left portion shows the original positive form and the right shows the negative form. Although it is a stable combination, the negation particle is still inserted in the two morphemes.

Fig. 6. A Khmer verbal expression consisting of two verbal morphemes. The left portion shows the positive form and the right shows the negative form. As in Figure 5, a negation particle is inserted. The second verb serves as a complement to show the result of the preceding verb.

$v[n \ o \ v]v$, respectively; and the two expressions in Figure 6 as $v[v \ v]v$ and $v[n \ o \ v]v$, respectively. An advantage of this method is that each token within the verbal expression has identical tags regardless of the extra morphemes that are inserted, and the integration is addressed by the brackets at the same time.

Other Issues

Several other linguistic phenomena should also be discussed. One important issue involves counter words, i.e., classifiers in counting nouns, which is a common word class in Asian languages but rare in Indo-European languages, and is not supported by the G-UNI tagset. Most classifiers are actually nouns, which causes confusion when deciding whether to use noun or prt (particle) tags under G-UNI. Moreover, classifiers may be regarded as adjectives in some analyses because of modifying nouns. This causes the use of the adj (adjective) tag in some cases. Under nova, however, tags such as `a[1 n]a` can express *a number and a noun-derived classifier working together as an adjective to modify a noun*. Hence, the classifiers are annotated analytically and naturally.

Another issue is *reduplication*, which is common in Asian languages and plays various grammatical roles. The reduplication can be used for forming adverbs in Burmese and for emphasis and pluralization in Khmer. To the best of our knowledge, none of the available POS systems widely used in NLP has a scheme for annotating this phenomenon, and currently, we do not provide such a scheme in nova. Because this phenomenon is related to detailed and complicated morphological analysis, we plan to investigate further in practice to decide whether (and how) to handle it specifically.

3.3 Annotation Guidelines

Tokenization

The Burmese and Khmer texts are segmented into morphemes in principle. As a practical instruction, the texts should be split into meaningful tokens as small as possible. Stable multi-morpheme expressions may remain unsegmented, which is decided by native-speaker annotators' common sense and entries listed in dictionaries.⁹ As the two languages are highly analytic, most single morphemes can be treated as words directly. For complicated expressions applying more than one morpheme for an entire concept, the brackets in nova are applied to address the integration. The complicated cases mainly include various compounds, and, especially in Burmese, suffixes.

Usage of Basic and Auxiliary Tags

The usage of basic and auxiliary nova tags is generally according to the instructions in Section 2.2 and Table 1. The application of *n* tag is intuitive; it can be applied for all the nominal tokens, including common nouns, proper nouns, and various pronouns. The three auxiliary tags are applied trivially to represent numbers, punctuations marks, and tokens with weak syntactic roles. However, the usage of *v*, *a*, and *o* tags are more affected by language-specific features.

Because the verbal and adjectival morphemes in Burmese share an identical set of functional suffixes in morphology, the adjective can be analyzed as a stative verb and thus is not a necessary word class. Strictly, the *v* tag can be applied on all the verbal and adjectival morphemes, while the use of *v* and *a* tags in practice is dependent on the annotator's common sense of how "dynamic" the morpheme is. For the determiners, the *a* tag is applied as they are always used to modify nominal tokens. The *o* tag is applied for all the functional tokens, including conjunction, particles, various suffixes, and a few prefixes. Generally, adverbs in Burmese can be classified as nominal tokens, and in practice, adverbs can be annotated by *o* or *n*, based on the annotator's common sense.

In Khmer, the difficulty is mainly around the usage of *v* and *o* tags. Due to the lack of inflection and relatively loose syntactic structures, verbs and prepositions have no obvious distinction in sentences when taking nominal constituents as arguments. Although the decision is largely dependent on the native speakers common sense, the *v* tag is more preferable than *o* in annotation, as a token should be tagged as *v* unless it can never be used as an independent verbal morpheme.

Combination

As the combination of simple nominal and verbal morphemes is free in Burmese and especially in Khmer, there are various patterns of compounds. According to the motivation in the design of nova, one compound will be segmented and wrapped by brackets, if the meaning of each component within the compound is relatively retained and related to the meaning of the entire compound. The typical linguistic examples have been introduced in Section 3.2, and real examples in annotating ALT data will be provided in Section 3.4.

An important issue for Burmese is the annotation of affixes, which afford most syntax and derivation information. In general, nominal suffixes, which mainly serve as case-markers, are simply segmented without using the brackets. The exception is on several plural markers, for which brackets are used to show the integration of a nominal constituent. Verbal suffixes, which add various tense and aspect information, are wrapped with brackets to emphasize the integration of a verbal constituent; suffixes for larger syntactic constituents as phrasal and sentential ones, are treated as completely independent tokens, where no brackets are applied. Besides the affixes to add syntactic information, there are also affixes to change the word classes in derivation. These

⁹Several common dictionaries, including those we used, have been digitalized at <http://sealang.net/>.

derivational affixes are mainly used for nominalization. Brackets annotation is generally applied on these affixes to wrap a derived nominal constituent.

The brackets may wrap more than two tokens if multiple components are used. When the composition of a constituent becomes complex, the brackets may cover a relatively long range. The usage of brackets is restricted to be shallow, and crossed or nested brackets are avoided.

Modification

The above-mentioned guidelines have covered an initial annotation from scratch. As to the motivation of the design of nova, the modification scheme is an optional refinement of the basic tags, to address the functionality of a part of tokens. In practice, the process is creating lists of functional tokens, conducting modification to the modified tags, and processing specific linguistic phenomena.

Generally, the grammatical forms in Burmese are mainly afforded by a small set of original functional morphemes along with a large set of grammaticalized suffixes. Although the boundary of grammaticalization is not so clear, it can be sorted out to a certain extent. We conducted the modification process for Burmese mainly with reference to Okell and Allott [10]. Specifically, the *n-* tags are used for general pronouns, including personal, demonstrative, interrogative, and numeral ones, the *a-* tag for general determiners, most of which derived from *n-* tagged tokens as a direct modifier for *n* or *n-* tagged tokens, and the *o-* tag for general particles, case-markers, and conjunctions. Combined tokens of *n/o-* and *n-/o-* are used for nouns and pronouns with a contracted genitive case-marker that cannot be tokenized. The *v-* tag is not used in order to avoid confusion. Grammaticalized tokens are consistently annotated as *o-* because the grammaticalization of morphemes is contextually dependent, which requires sophisticated analysis.

Khmer is more analytic than Burmese, with more flexible phenomena of grammaticalization. In contrast to the case of Burmese, where the *v-* is reserved, the *v-* tag is used frequently in the modification process on Khmer data. This is because most functional morphemes in Khmer can be ultimately analyzed as grammaticalized verbal morphemes. The *v-* tag is thus preferred over the *o-* tag to indicate the origin of the corresponding morphemes.

3.4 Statistics and Examples on Annotated Data

We further present statistics on the released Burmese and Khmer ALT data annotated by nova.¹⁰ Two ways of statistics are conducted. One is over *short* units (tags and tokens) and another is on *long* units. The statistics on short ones ignore the brackets, and those on long ones take the bracketed sequences of short units as single units.

Figure 7 shows Zipfian features on annotated data. Although Zipf's law is not a “*deep law in natural language*” [5], the overall Zipfian distribution on tokens suggests there is no obvious singularity in the tokenization. It should be noticed that a series of *long-tag* on the two languages, say, combination patterns facilitated by nova, also shows a Zipfian-like tendency, though the log-log plots are steeper than those on tokens. Compared with the non-Zipfian series of *short-tag*, it suggests that the bracketing annotation covers plenty of phenomena in a heavy-tailed distribution, rather than a simple classification of tokens.¹¹

Detailed distributions of tags and patterns are listed in Tables 2, 3, and 4. In Table 2 of short tags, it is obvious that *n*, *v*, and *o* tags nearly cover 90% of the tokens, which is in accordance with the

¹⁰The latest version of annotated data when the article is prepared: Burmese: <http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/my-nova-170405.zip> Khmer: <http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/km-nova-180803.zip>.

¹¹The modified short- and long-tags show similar tendency as those of the basic tags shown in Figure 7.

Fig. 7. Rank-frequency plots based on annotated Burmese (left) and Khmer (right) ALT data.

Table 2. Count and Percentage of nova Tags on Short Tokens from Annotated Burmese and Khmer ALT Data

Burmese				Khmer			
Tag	Count	%	Note	Tag	Count	%	Note
o	306,464	41.6	95.9% of o is o-	n	296,447	41.5	5.7% of n is n-
n	216,190	29.3	8.1% of n is n-	v	215,945	30.2	34.8% of v is v-
v	120,255	16.3		o	108,544	15.2	12.6% of o is o-
.	56,881	7.7		a	37,563	5.3	22.0% of a is a-
1	22,507	3.1		.	37,384	5.2	
a	14,839	2.0	39.8% of a is a-	1	18,234	2.6	
+	1	0.0		+	10	0.0	
Total	737,137	100.0		Total	714,127	100.0	

5.0% of n- tags are further combined with /o- in Burmese data.

Table 3. Count and Percentage of nova Tags on Long Tokens from Annotated Burmese and Khmer ALT Data

Burmese			Khmer		
Tag	Count	%	Tag	Count	%
n	176,695	32.0	v	138,595	26.1
o	155,612	28.1	n	110,713	20.9
.	56,794	10.3	o	79,035	14.9
v	22,733	4.1	n[n n]n	38,739	7.3
n[n o]n	21,029	3.8	.	35,356	6.7
n[v o]n	17,043	3.1	a	22,663	4.3
a[1 n]a	12,733	2.3	n[n v]n	11,100	2.1
v[v o o]v	12,610	2.3	v[v v]v	10,947	2.1
a	11,809	2.1	n[n n n]n	8,014	1.5
v[v o]v	9,900	1.8	1	5,748	1.1
1	9,355	1.7	n[n v v]n	5,658	1.1
a[v o]a	8,440	1.5	o[o o]o	5,524	1.0
Total	514,753	93.1	Total	472,092	89.0

Only tags and patterns with a frequency larger than 1.0% are listed.

Table 4. Portions of Modified nova Tags on Long Tokens from Annotated Burmese and Khmer ALT Data

Burmese			
Basic Tag	Modified Tag	Modified %	Note
n	n-	6.0	
o	o-	93.5	
v	—	—	not used
n[n o]n	n[n o-]n	77.4	o- mostly for plural suffixes
	n-[n- o-]n-	22.2	
n[v o]n	n[v o-]n	99.8	o- mostly for derivational suffixes
a[1 n]a	—	—	exclusively for number-counter combination
v[v o o]v	v[v o- o-]v	99.9	o- mostly for various verbal suffixes
a	a-	50.0	
v[v o]v	v[v o-]v	99.6	o- mostly for various verbal suffixes
a[v o]a	a[v o-]a	99.8	o- mostly for attribute suffixes

Khmer			
Basic Tag	Modified Tag	Modified %	Note
v	v-	53.3	
n	n-	11.6	
o	o-	14.5	
n[n n]n	n-[n n-]n-	6.2	93.8% of n[n n]n is not modified
a	a-	31.8	
n[n v]n	—	—	none of n[n v]n is modified
v[v v]v	—	—	none of v[v v]v is modified
n[n n n]n	—	—	99.8% of n[n n n]n is not modified
n[n v v]n	—	—	99.1% of n[n v v]n is not modified
o[o o]o	o-[o- o-]o-	6.4	88.1% of o[o o]o is not modified
	o-[o- o-]o-	5.2	

Only modified tags and patterns with a frequency larger than 1.0% are listed.

linguistic motivations to provide basic word classes by nova. Most of the o tags in the Burmese data are actually o-, which are those functional tokens affording syntactic information; however, a considerable portion of functional tokens are tagged as v- in the Khmer data, where more than one third of v tags are further modified. In Table 3, the distribution of frequent basic long tag patterns are presented. An obvious phenomenon in Burmese is that a large portion of o tags are bracketed, which are actually o- for various suffixes from Table 4. The a tag is not used frequently in Burmese, but it more appears in a[1 n]a and also in a[v o]a. In the case of Khmer, there are more patterns directly formed by basic n and v tags. This suggests that Khmer is more analytic than Burmese, where compounding is more used than derivation in word formation. Notice the endocentric nominal-verbal structures (n[n v]n and n[n v v]n) in Khmer ALT data also have a certain portion, where the brackets provide an exact annotation.

As shown in Table 3, the 12 most frequent basic tags and combination patterns cover around 90% of the long tokens in both languages, among which six patterns in each are annotated by brackets. We further show the five most frequent examples of these bracketed patterns. The total of 30 examples in Burmese and Khmer are listed in Figures 8 and 9, respectively. In Figure 8 of Burmese, the examples of 1- are the pattern of n[n o]n, where most are nominal tokens attached with a plural suffix. Notice that the plural form of pronouns is also analytic in Burmese, so the n-[n- o-]n-

	Annotated Burmese	English Gloss of Components	English Gloss
1-1	သူ_n-[n- တို့_ဝ-]n-	he/she + -s (a group of)	they
1-2	ကျွန်တော်_n-[n- တို့_ဝ-]n-	I (male) + -s (a group of)	we
1-3	ကျွန်ုပ်_n-[n- တို့_ဝ-]n-	I + -s (a group of)	we
1-4	လူ_n[n- များ_ဝ-]n	man + -s (to be many)	people
1-5	အရာရှိ_n[n များ_ဝ-]n	official + -s (to be many)	officials
2-1	ဆို_n[v တာ_ဝ-]n	to say + (verb-to-noun suffix)	saying
2-2	တိုက်ခိုက်_n[v မှု_ဝ-]n	to attack + -ing (a deed of)	attacking
2-3	လုံခြုံ_n[v ရေး_ဝ-]n	to be secure + -ing (an affair of)	security
2-4	ကျန်းမာ_n[v ရေး_ဝ-]n	to be healthy + -ing (an affair of)	health
2-5	ဆုံးဖြတ်_n[v ချက်_ဝ-]n	to decide + (verb-to-noun suffix)	decision
3-1	တစ်_a[1 ခု_n]a	one + unit	one (anything)
3-2	တစ်_a[1 ယောက်_n]a	one + (counter for person)	one person
3-3	တစ်_a[1 ဦး_n]a	one + (counter for person)	one person
3-4	နှစ်_a[1 ခု_n]a	two + unit	two (anything)
3-5	တစ်_a[1 ရပ်_n]a	one + (counter for abstrac items)	one (idea, subject, etc.)
4-1	ပြော_v[v ခဲ့_ဝ- သည်_ဝ-]v	to talk + (definitiveness) + (non-future)	has/have taked (finite)
4-2	ဖြစ်_v[v ခဲ့_ဝ- သည်_ဝ-]v	to become + (definitiveness) + (non-future)	has/have become (finite)
4-3	ဖော်ပြ_v[v ခဲ့_ဝ- သည်_ဝ-]v	to show + (definitiveness) + (non-future)	has/have shown (finite)
4-4	ရှိ_v[v ခဲ့_ဝ- သည်_ဝ-]v	to be + (definitiveness) + (non-future)	there has/have been
4-5	ဖြစ်_v[v ပါ_ဝ- တယ်_ဝ-]v	to become + (politeness) + (non-future)	has/have become (finite)
5-1	ဖြစ်_v[v သည်_ဝ-]v	to become + (non-future)	become/becomes (finite)
5-2	ရှိ_v[v သည်_ဝ-]v	to be + (non-future)	there is/are
5-3	ပြော_v[v သည်_ဝ-]v	to talk + (non-future)	take/talks (finite)
5-4	ဖြစ်_v[v တယ်_ဝ-]v	to become + (non-future)	become/becomes (finite)
5-5	ဖြစ်_v[v ခဲ့_ဝ-]v	to become + (definitiveness)	to have become
6-1	ဖြစ်_a[v သော_ဝ-]a	to become + -ing (adjectival)	becoming
6-2	ရှိ_a[v သော_ဝ-]a	to be + -ing (adjectival)	being
6-3	ဆိုင်_a[v သော_ဝ-]a	to meet + -ing (adjectival)	concerning
6-4	ဖြစ်_a[v သည့်_ဝ-]a	to become + -ing (non-future adjectival)	becoming
6-5	ပါဝင်_a[v သော_ဝ-]a	to include + -ing (adjectival)	including

Fig. 8. Examples of frequent bracketed patterns on Burmese ALT data.

	Annotated Khmer	English Gloss of Components	English Gloss
1-1	ពួក_n-[n គេ_n-]n-	group + one	they
1-2	ក្រុម_n[n ហ៊ុន_n]n	group + capital	company
1-3	ប្រជា_n[n ជន_n]n	people + people	people/population
1-4	រយៈ_n[n ពេល_n]n	part + time/moment	duration
1-5	កម្ម_n[n វិធី_n]n	action + method	program
2-1	ការ_n[n ប្រកួត_v]n	act + to compete	competition
2-2	យន្ត_n[n ហោះ_v]n	device + to fly	airplane
2-3	ទី_n[n តាំង_v]n	place + to establish	installation/headquarter
2-4	ការ_n[n ផ្ទុះ_v]n	act + to explode	explosion
2-5	គ្រាប់_n[n បែក_v]n	seed + to break	bomb
3-1	ចាប់_v[v ផ្ដើម_v]v	to catch + to start/begin	to start/begin
3-2	ចូល_v[v រួម_v]v	to enter + to join/union	to participate
3-3	រៀប_v[v ចំ_v]v	to prepare + to coincide with	to organize
3-4	ក្លាយ_v[v ជា_v]v	to become + to be as	to become
3-5	ចោទ_v[v ប្រកាន់_v]v	to question + to claim	to accuse
4-1	នាយក_n[n រដ្ឋ_n មន្ត្រី_n]n	chief + state/nation + official	prime minister
4-2	ចក្រ_n[n ភព_n អង់គ្លេស_n]n	circle + world + English	United Kingdom
4-3	សហ_n[n រដ្ឋ_n អាមេរិក_n]n	unity + state/nation + America	United States
4-4	ភាគ_n[n ខាង_n ជើង_n]n	part + side + north	North (Korea)
4-5	ភាគ_n[n ខាង_n ត្បូង_n]n	part + side + south	South (Korea)
5-1	ការ_n[n វាយ_v ប្រហារ_v]n	act + to beat + to kill	attack
5-2	ការ_n[n ស៊ើប_v អង្កេត_v]n	act + to inquire + to inquire	investigation
5-3	ការ_n[n ផ្លាស់_v ប្តូរ_v]n	act + to replace + to exchange	exchange
5-4	ការ_n[n ចោទ_v ប្រកាន់_v]n	act + to question + to claim	accusation
5-5	មេ_n[n ដឹក_v នាំ_v]n	chief + to lead + to lead	leader
6-1	បន្ទាប់_០[០ ពី_០]០	to be next + from/about	after/next
6-2	ជាមួយ_០[០ នឹង_០]០	along with + by means of	along with
6-3	ផង_០[០ ដែរ_០]០	also/too + also/too	also/too
6-4	ក៏_០[០ ដោយ_០]០	also + in accordance with	whether or not
6-5	ប្រសិន_០-[០- បើ_០-]០-	if + if	if

Fig. 9. Examples of frequent bracketed patterns on Khmer ALT data.

pattern is common. The pattern of $n[v \ o]n$ of example 2- is verb-derived nominal expressions, where various nominalization suffixes are addressed. Examples of 3- are the number-counter patterns, and the following examples are all about verbal expressions, where suffixes are tokenized and annotated by brackets. In Figure 9 of Khmer, more compounding phenomena can be observed. In the example of 1-1, the plural marker of a pronoun is even a general nominal token with the meaning of “group.” In the examples of 2-1, 2-4, and from 5-1 to 5-4, the first component is nearly a grammaticalized morpheme for nominalization; however, the original meaning of “act” is more or less retained. There are less *o* tags used in the Khmer data than in the Burmese data, as the original meaning of grammaticalized nominal and verbal tokens is addressed in annotation, and thus, *n* and *v* tags are preferred. The pattern of $o[o \ o]o$ in examples of 6- show that many functional expressions in Khmer are still analytic. Because native speakers are still aware of the compositional substructure within these expressions, brackets are applied conveniently for an analytic annotation.

4 NOVA, G-UNI, AND UNI-DEP

4.1 Descriptive Ability of Tagset

Generally, the basic and modified nova tags can be regarded as a simplified and modularized version of G-UNI, and the exchangeability between the two systems is reserved to a large extent. Table 5 shows a naïve mapping from G-UNI (and from the extended UNI-DEP) tagset to nova’s basic and modified tags without considering language-specific details. In the two middle rows, the mapping is almost one-to-one between G-UNI and modified nova, except that the *conj* (conjunction) and *prt* (particle) in the G-UNI tagset are both mapped to nova’s *o-*. This is reasonable because the conjunction is actually not such a universal class of words and the particle is somewhat of a catch-all class for functional words that are difficult to classify. Thus, they are covered by an identical tag in nova, which can be interpreted as a general functional modifier. In the upper row, UNI-DEP provides more detailed sub-classes in tagging. Specifically, the *conj* tag is substituted by two tags of *cconj* and *sconj* to distinguish *coordinating* and *subordinating* conjunctions, respectively. Proper nouns are separated from general nouns by a *propn* tag. Other introduced tags are *aux* (auxiliary verb), *sym* (symbol), and *intj* (interjection). Basic nova tags in the lower row in Table 5 provide a basic and coarse annotation, where the information to address functional words is lost. However, this kind of information can be partly compensated by the lexicalized frequency in real data. The comparison shows that the descriptive ability of nova is nearly equal to that of G-UNI, with the flexibility on switching between a generalized version and a modified version.

A well-balanced point is important in designing a universal POS tagging system, where the tradeoff between generality and specificity should be measured. A detailed tagset, although providing informative annotation, may turn too specific regarding one or a group of particular languages, and thus, not consistently applicable on languages with different features. The tradeoff between generality and specificity motivates the overall designing of nova, although certain information may be lost in specific cases compared with G-UNI and UNI-DEP. Reviewing our original motivation, the position of nova can be located at facilitating an interface for annotation from raw data to a relatively informative level, from scratch to a basically linguistically aware level. Hence, it can be considered nova provides an efficient and (nearly) seamless basis for G-UNI or UNI-DEP tags, and for further development on sophisticated language-specific annotation. The G-UNI tagset bridges established POS tag systems on different languages in a top-down manner,¹² while nova provides a fundamental ground for those unstudied languages in a bottom-up manner.

¹²The UNI-DEP system is embedded in the *Universal Dependency* project to bridge syntactic structures of different languages, which is oriented to a complex and difficult task at a further higher level.

Table 5. Naïve Mapping between G-UNI/UNI-DEP and nova Tagsets, Organized in a Fine-to-Coarse Way

UNI-DEP	ADJ	ADP	ADV	CCONJ SCONJ	DET	NOUN PROPN	NUM	PRON	PART	VERB AUX	X INTJ	PUNCT SYM
G-UNI	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X	.
modified nova	a	v-	o	o-	a-	n	1	n-	o-	v	+	.
basic nova	a	v	o	o	a	n	1	n	o	v	+	.

4.2 Flexibility in Tokenization

As mentioned, word boundary is not so intuitive for many highly analytic languages. Besides definitions and explanations offered by linguistic theories, a more practical way in NLP engineering is to create large annotated corpora, and to take those corpora themselves as a “data-driven standard *de facto*” for the concept of words. An example is the tokenization for Chinese, where a practical solution has been solidly established by (1) segmentation guidelines, (2) lexicon, and (3) segmented corpus [4]. Here, (1) is a linguistically driven component and (2) is a dictionary-based component, both of which stay in the scope of manual rules. Segmented corpus (3) is a real data-driven component to overcome the limitation of (1) and (2). More essentially, the concept of words in terms of NLP is more of an artificially created unit based on practice, rather than a pure linguistic or conventional consensus. Consequently, the solid basis for tokenization must be established from adequate NLP practices.

For low-resource languages, nova can accelerate the process. As ambiguities and inconsistency are unavoidable in manually annotated data, especially in those data preliminarily annotated from scratch, an important issue is, thus, to keep those problematic cases controllable in regard to practical tasks and further standardization. The combination mechanism in nova thus integrates the information of tokenization and POS tagging in a unified representation, i.e., the information on tokenization is enriched by POS tags, and the POS tagging can be launched and carried on with tokenization simultaneously. Compared with the three components in the case of Chinese tokenization, nova actually provides a way to realize (3), i.e., annotated corpus, not only tokenized, but further POS tagged, when there is not a well-prepared (1), i.e., detailed guidelines. nova bridges the two components to mutually improve each other in the process of corpus construction.

Facilitated by nova, we consider an annotation process of joint tokenization, and POS tagging from scratch can be decomposed into the following coarse-to-fine stages:

- (i) to realize initial POS tagging under unstable tokenization by combined basic nova tags
- (ii) to realize more informative functional POS tagging by modified nova tags
- (iii) to realize stable tokenization and POS tagging by converting nova to G-UNI and UNI-DEP
- (iv) to realize language-specific annotation

Generally, the original motivation of G-UNI is located at (iv) → (iii), i.e., inter-lingual process for different languages with available annotated data, while the motivation of nova is in a reversed direction starting from (i). It can be considered that nova is placed at a “pre-G-UNI” stage, from realized annotated data up to the level of G-UNI in a more efficient way.

5 CONCLUSION AND FUTURE WORK

In this article, we proposed nova, a feasible and flexible annotation system for joint tokenization and POS tagging. nova provides an interface of fundamental and efficient annotation from scratch for low-resource languages. More informative annotation can be established gradually on the basic annotation provided by nova.

To illustrate the practical usage of nova, descriptions on two typical Southeast Asian languages, i.e., Burmese and Khmer, are provided, including guidelines, statistics, and examples of the annotation. We have released the guidelines and annotated data of Burmese and Khmer by nova. Broadly, we plan to apply nova in joint tokenization and POS tagging on other low-resource languages, which share similar linguistic features with the two languages discussed in this article.

ACKNOWLEDGMENTS

We thank many native-speaker annotators in the construction of annotated Burmese and Khmer corpora. We thank Dr. Ye Kyaw Thu for his discussions on Burmese. We thank Mr. Hour Kaing and Mr. Vichet Chea for their discussions on Khmer. We thank Prof. Win Pa Pa and Ms. Hnin Thuzar Kyaw for their help in checking the Burmese data. We thank Mr. Kamthong Ley and Mr. Vanna Chuon for their help in processing the Khmer data. We thank Dr. Atsushi Fujita for his comments on the draft of the article. We thank anonymous reviewers for their insightful comments.

REFERENCES

- [1] L. Bloomfield and C. F. Hockett. 1984. *Language*. <https://books.google.co.jp/books?id=87BCDVsmFE4C>.
- [2] Noam Chomsky. 1970. Remarks on nominalization. *Readings in English Transformational Grammar*, Roderick A. Jacobs and Peter S. Rosenbaum (Eds.). Ginn, 184–221.
- [3] Madeline Elizabeth Ehrman, Kem Sos, and Lim Hak Kheang. 1974. *Contemporary Cambodian—Grammatical Sketch*. Foreign Service Institute, Department of State. Retrieved from <https://www.livelingua.com/fsi/Fsi-Contemporary-Cambodian-GrammaticalSketch.pdf>.
- [4] Chang-Ning Huang and Hai Zhao. 2007. Chinese word segmentation: A decade review. *Journal of Chinese Information Processing* 21, 3 (2007), 8–19. (in Chinese).
- [5] Wentian Li. 1992. Random texts exhibit Zipf’s-law-like word frequency distribution. *IEEE Transactions on Information Theory* 38, 6 (1992), 1842–1845.
- [6] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proc. of LREC*. 1659–1666.
- [7] Hideki Ogura, Hanae Koiso, Yumi Fujiike, Sayaka Miyauchi, Hikari Konishi, and Yutaka Hara. 2011. Retrieved from http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-01.pdf and http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf (in Japanese).
- [8] John Okell. 2010. *Burmese—An Introduction to the Spoken Language, Book 1*. Northern Illinois University Press.
- [9] John Okell. 2010. *Burmese—An Introduction to the Spoken Language, Book 2*. Northern Illinois University Press.
- [10] John Okell and Anna Allott. 2001. *Burmese/Myanmar Dictionary of Grammatical Forms*. Routledge.
- [11] Thomas Pachunke, Oliver Mertineit, Klaus Wothke, and Rudolf Schmidt. 1992. Broad coverage automatic morphological segmentation of German words. In *Proc. of COLING*. 1218–1222.
- [12] Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proc. of LREC*. 2089–2096.
- [13] 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 *Proc. of O-COCOSDA*. 1–6.
- [14] Jonathan J. Webster and Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In *Proc. of COLING*. 1106–1110.
- [15] Fei Xia. 2000. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). Retrieved from <http://www.cis.upenn.edu/~chinese/posguide.3rd.ch.pdf>.
- [16] Fei Xia. 2000. The segmentation guidelines for the Penn Chinese Treebank (3.0). Retrieved from <http://www.cis.upenn.edu/~chinese/segguide.3rd.ch.pdf>.
- [17] Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitchell P. Marcus. 2000. Developing guidelines and ensuring consistency for Chinese text annotation. In *Proc. of LREC*.

Received May 2016; revised August 2018; accepted September 2018