ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 788-800 (2020)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Dependency Head Annotation for Myanmar Dependency Treebank

Hnin Thu Zar Aye*, Win Pa Pa

Natural Language Processing Lab, University of Computer Studies, Yangon, Yangon, 11411, Myanmar

ARTICLE INFO

Article history:
Received: 22 September, 2020
Accepted: 17 November, 2020
Online: 24 November, 2020

Keywords:
Dependency head
Universal Dependencies
Treebank
Annotation schemes

ABSTRACT

Complete manual annotation of dependency treebank needs resources like annotators and annotation tools and takes long time and has high possibility of inconsistent annotations for free word order languages such as Myanmar. This paper describes a dependency head annotation scheme with Universal part-of-speech and Universal Dependencies for Myanmar dependency treebank. Currently 22,810 sentences and 680,218 tokens were annotated from three corpora for Myanmar dependency treebank. Some language specific issues are also described with examples. Raw syntactic structures were annotated automatically by UDPipe according to the Universal Dependencies based on Universal-part-of-speech tag scheme. Then unsupervised annotated dependency head structures have been manually updated in post processing. To be reliable and speedy post process with reduced errors for manual updating, selected sentences were added to the training data after being updated. After that the model has been retrained and the remaining sentences were parsed by UDPipe. Post processing was repeated until all sentences were updated. Some specifications of dependency annotation schemes in sentences encountered in post processing are presented with examples. For parsing performance of annotated data, cross validation tests and parsing experiments were performed. Moreover, annotated treebank data have also been evaluated by CoNLL 2017 evaluation script for parsing performance. Results of parsing experiments and evaluation are also reported by unlabeled and labeled attachment scores and demonstrated that the proposed method is a suitable way for building Myanmar dependency trees. Moreover, syntax structures of treebank are also analyzed and syntax information is also presented. This dependency head annotation for dependency treebank is the first work for Myanmar language as far as we know.

1. Introduction

A treebank annotated with syntax or dependency structure is an essential and important resource for natural language processing systems in any language. Treebank annotated dependency structures is called dependency treebank. While phrasal constituents and syntax rules could not provide a direct role in sentences, syntactic dependency information of a sentence can describe directed grammatical relations between words. Moreover, dependency grammar is also able to deal with morphologically rich and relatively free word order languages. Dependency treebank is also critical resource in any language to develop natural language processing applications [1].

Many dependency treebanks have been constructed manually for many languages such as Czech, German, Danish, French, Portuguese, Estonian, Russian, Dutch, Danish, Turkish, Basque, Italian, English [2], Norwegian [3], Finnish [4], Romanian [5], Ancient Greek and Latin [6], Vietnamese [7],

Annotating dependency syntactic information in sentences is still a hard task for Myanmar having free word order nature. Moreover, currently there is still low resource for syntactic information for Myanmar language.

The Myanmar grammar is different from other languages of ASEAN countries such as Thailand, Vietnam, Malaysia and these languages already have treebank resources. Myanmar has been similar structures with the other SOV order languages such as Japanese, Chinese, and Korean and also a head final language. According to these properties, for Myanmar, a dependency-based head finalization has been proposed for statistical machine translation (SMT) in [8]. Although the proposed method was being able to improve a baseline SMT result without requiring parallel

*Corresponding Author: Hnin Thu Zar Aye, hninthuzaraye@ucsy.edu.mm

training information, it depends on dependency parser of source-side language to achieve higher performance than unsupervised baseline result [8]. Having similar syntactic structures of Japanese and Myanmar, a parsing approach has been proposed by applying SMT by using Japanese as pivot in [9]. The proposed method performed well with satisfying results due to the good performance of Japanese parser [9]. Although Myanmar dependency syntax structures were applied competently for SMT [8] and parsing [9] without annotated Myanmar corpus, there was dependence on intermediate language dependency parser like English and Japanese as limitation

Progress of unsupervised dependency parsing researches is increasing with the shared tasks of the tenth conference on computational natural language learning (CoNLL-X) in recent years [10]. A Universal part-of-speech tag (U-POS) set has been proposed as standard for research in unsupervised induction of syntactic structure [11]. Universal Dependencies (UD) is a project developing cross-linguistically consistent treebank annotation for many languages [12]. Currently, there are over 100 treebanks of more than 70 languages available in the UD inventory. Treebanks have derived UD format from existing formats in many languages like Korean [13]. UDPipe has been proposed to easily perform basic natural language processing tasks from tokenization to parsing in CoNLL-U format, the revised version of CoNLL-X format, for treebanks of UD without requiring any other external data [12]. UDPipe has been applied in dependency treebank building [14, 15].

Currently, Asian language treebank (ALT) project has developed a Myanmar syntax treebank with a parallel corpus by annotators using web-based tool [16]. Building treebank needs annotators and applied tools for processes of tree building. Therefore, it becomes hard for low resource language like Myanmar.

Having limitation of related works [8, 9], no resource for dependency parsing and information of Myanmar, improvement of unsupervised parsing researches [10], and simple trainable facility of UDPipe with CoNLL-U format, as our motivation, we annotated a corpus by applying U-POS tags and unsupervised dependency parsing by UDPipe of UD project to get raw syntax information for Myanmar [17]. Then, as future work of unsupervised annotation, manual post processing is carried out on the unsupervised parsed results with reference dependency structures to build dependency treebank in order to apply in parsing Myanmar sentences by deep learning approaches.

This paper presents dependency head annotation to build Myanmar dependency treebank by U-POS tag sets and UD 2.0 guidelines by updating in post processing on unsupervised annotated corpus. Building Myanmar dependency treebank contains two main parts. The first is automatic annotating by unsupervised dependency parsing by UDPipe to get raw universal dependency syntactic information [17]. The second is manual post processing by dependency structures for correct dependency heads. During post processing, UDPipe is applied in a bootstrapping manner to be consistent updating with reduced post processing time..

The organization structure of paper is as follow: Section 2 briefly describes nature of Myanmar language and sentences. Section 3 presents corpus information and overview of annotation scheme. Section 4 describes corpus pre-processing tasks before annotation. Section 5 presents Myanmar language specific tags and their related U-POS tags and mapping scheme between language POS and U-POS tags. Section 6 describes post checking and updating unsupervised dependency heads in sentences with examples. Section 7 reports parsing experiments with post processed data. Section 8 discusses about parsing and evaluation results of treebank. Section 9 concludes all sections and presents future work.

2. Myanmar Language

Myanmar (Burmese) is a member of the Lolo-Burmese grouping of the Sino-Tibetan language belonging to the Southern Burmish branch of the Tibeto-Burman languages. It is an official language in Myanmar. It is also the first language of the Bamar people, the principal ethnic group and related ethnic groups of the country, and a second language of ethnic minorities in Myanmar. Myanmar (Burmese) is a tonal, pitch-register, and syllable-timed language, largely monosyllabic and analytic, with a subject–object–verb word order. It is morphological rich and agglutinative language.

2.1. Nature of Myanmar Sentences

Myanmar sentences can be written with formal or colloquial style. Two types of Myanmar sentence construction are simple and complex sentence types. Simple sentence has only one nominal phrase, action maker or subject, and one verb phrase. Complex sentence has two or more clauses or simple sentences, joined with conjunction, or post positional markers, or particles to modify the followed part which might be phrase or clause of main sentence.

Sentence / (Type)	ကိုယ် တွင် ညို သော်လည်း ၊ မျက်နှာ နှင့် လက် များ မှာ မည်း နေ သည် ။ (Complex sentence)													
Clauses / (Type)	***ကိုယ်***တွင် ***ညို*** သော်လည်း ၊ (Dependent clause)					***မျက်နှာ*** နှင့် ***လက်*** များ မှာ ***မည်း***နေ သည် ။ (Independent root clause)								
Phrases/ (Type)	***ကိုယ်*** တွင် (Noun)		***ညို*** သော်လည်း ၊ **(Adjective)**			***မျက်နှာ*** နှင့် ***လက်*** များ မှာ (Noun)					***မည်း***နေ သည် ။ **(Adjective/Root)**			
Words	***ကိုယ်***	တွင်	***ညို***	သော်လည်း	၊	***မျက်နှာ***	နှင့်	***လက်*** များ		မှာ	***မည်း***	နေ	သည်	။
Morphenes	***ကိုယ်***	တွင်	***ညို***	သော်လည်း	၊	***မျက်နှာ***	နှင့်	***လက်***	များ	မှာ	***မည်း***	နေ	သည်	။
Morpheme translation	body	at	brown	although	,	face	and	leg	-s	-	black	-	-	.
Sentence translation	Although body is brown, face and legs are black .													

Figure 1: Example grammatical hierarchy structure of example sentence

Table 1: Suffixes of example sentence from Figure 1

Types	Nominal marker (post positional marker)	Plural marker (particle)	Verb suffix (particle)	Verb marker (post positional marker)
Morpheme Words	“တွင်”,“များ”, “မှာ”	“များ”	“နေ”	“သည်”

Table 2: Composition of Myanmar Dependency Treebank

Corpus Source	Domain	Sentences	Tokens	Average Length	Remark
myPOS	Wikipedia	11,010	239,534	21.75	manually annotated
Web News	Websites’ News	1,800	66,710	37.06	manually annotated
ALT	Wiki news	10,000	373,974	37.39	Automatic annotated
	Total	22,810	680,218		

Myanmar noun phrases are usually ended with different types of nominal markers, postpositional markers (PPMs), to identify their roles in sentences such as subject, object but most colloquial style noun phrases are written without these markers. Myanmar sentences might have one or more phrases or clauses. Some subordinate clauses are usually used to represent more detail meaning for modified parts and also placed before modified ones. Having these nested parts, defining correct dependency between sub and main parts of the sentence takes more time. Figure 1 illustrates an example sentence structure.

There are four phrases in sentence of Figure 1. In Myanmar language, adjective ended with verb maker suffix is verb phrase. In the example sentence, sentence ended verb phrase is composed of adjective with post positional verb marker Four suffix types of morphemes of the example sentence are described in Table 1. Bold and italic words are main content words of each phrase. Main root phrase of a sentence or clause is final verb phrase of that sentence. Dependent sentence or clause modifies the independent root sentence. Therefore, main root of example sentence in Figure 1 is the last right most verb phrase.

Moreover, most formal sentences are usually ended with verb or verb phrase. However, some sentences may not be ended with verbs because of the Myanmar sentence writing style in which there are presence of hiding verbs hidden for actions of being or having or living or coming actions. Moreover, Myanmar sentences can be written by many forms according to nature of free word order language, and some special cases of sentence construction of Myanmar grammar. Moreover, emphasized noun phrases can be put at the beginning of the sentence according to writer’s idea.

These conditions can also be complicated and time-consuming issues to define correct dependency heads and relations for phrases in sentences. Besides above nested cases, there arises one issue in dependency tree building.

3. Corpus Annotation

This section presents corpus statistic and overall architecture of dependency head annotation.

3.1. Corpus Information

Currently two corpora have been annotated manually by dependency head information and one corpus has been annotated in unsupervised way for Myanmar treebank. The myPOS corpus consists of 11,010 sentences written by formal and colloquial format from the Myanmar Wikipedia including various areas such as economics, history, news, politics and philosophy [18]. The sentence writing style of myPOS corpus is very similar to current mostly used standard formal and colloquial style. Therefore, we selected first this corpus to annotate to get similar syntactic structures of current written styles of Myanmar sentences and annotated them as first standard sentences for other corpora. Web News corpus contains 1,800 sentences written by current modern writing styles in Myanmar news websites. ALT corpus contains 10,000 translated sentences from Wiki news. The statistical information of current Myanmar dependency treebank is presented in Table 2.

3.2. Overview Structure of Annotation

Using difference corpora in treebank, word segmentation and POS tagging style of each corpus is different. Being morphological rich and agglutinative language, most words are segmented to provide morphological level syntax information in this work. Therefore, pre-processing is needed to carry out to be the same word segmentation and POS-tagged scheme among different corpora. U-POS tags, CoNLL-U shared task format, shared Japanese model, and UDPipe of UD project were used to get raw universal dependencies in this work.

The corpora were transformed to CoNLL-U format by adding U-POS tags in pre-processing. After transformation, they were annotated by unsupervised dependency parsing by using shared Japanese model in UD project [17] because of similar conditions in grammar structure of Myanmar and Japanese and dependency structures and UD of Japanese [19,20]. Then unsupervised annotated dependency heads were manually post checked by human annotator to be more correct dependency head nodes. One annotator was done manual post checking on automatic annotated results by learning Myanmar language grammar books and books written by linguistic experts of Myanmar language. Therefore, dependency head annotation of Myanmar dependency treebank has two steps: initial corpus pre-processing and post processing on unsupervised annotated results of pre-processing.

In post processing, to be consistent and fast checking and updating, selected updated data were repeatedly trained with all other not updated data in corpus by UDPipe. After checking and updating manually selected 2,000 sentences, they were added to the training data, Then the training model was retrained to parse the remaining sentences in corpus by the updated model. Parsed

results were updated manually and trained again until all sentences were post processed in corpus. The overall architecture of annotation is illustrated in Figure 2. Detail explanation with examples will be presented in Section 4 and Section 6.

4. Pre-processing

This section presents overview of corpus pre-processing step. Before dependency structures annotation, it is needed to check and update word segmentation and POS tagging of used corpora because some corpora might have different segmentation and POS tagging styles based on their original created purposes. Using unique word segmentation and POS tagging can be easy to transform dependency corpus. It is also better and easier having unique word segmentation and POS tagging than different formats to transform dependency corpus.

To have a consistent Myanmar POS tag scheme, a new general POS tag scheme has been defined. It will be explained in following sub section. Moreover, U-POS tags were also added to the CoNLL-U format in order to get raw universal dependencies syntax structures.

In Myanmar sentences, some words might also have different POS tag forms for the same word. Tagging correctly for each content word in sentences is important for correct dependency head. One example of these issues is presented in Table 3. The quality of word segmentation and POS tagging can mainly impact providing correct dependency information of sentences in corpus. Therefore, rechecking word segmentation and POS tagging of used corpus before annotation is very important.

Then, the correct POS tagged corpus was transformed CoNLL-U format by adding related U-POS tags for most language POS tags by manual python script as UD 2.0 format. In adding U-POS tags by python script, it is needed to check again manually U-POS tags of Myanmar conjunctions because it is also needed to be correct form of two conjunction tags of U-POS as described in Table 5 according to the content words of sentence.

5. Part-of-speech Tagging Scheme

There are ten main POS tags in Myanmar language such as noun, pronoun, adjective, verb, adverb, post-positional marker, particles, conjunctions, interjection and punctuation [21]. For easy mapping to U-POS tags, 16 POS tags have been defined for language specific tags which are presented in Table 4 with examples.

Most tags of these were already defined in ALT [6] except proper noun (PRPN) and text number (TNUM). In the previous work of this, these two tags were defined as noun, “N” [17]. In most Myanmar corpora, proper nouns and text numbers are tagged as noun, but they are already defined as proper noun and number in U-POS tag set. Therefore, these two tags were also added in Myanmar language tag set in order to be fast and easy mapping between Myanmar POS tags and U-POS tags.

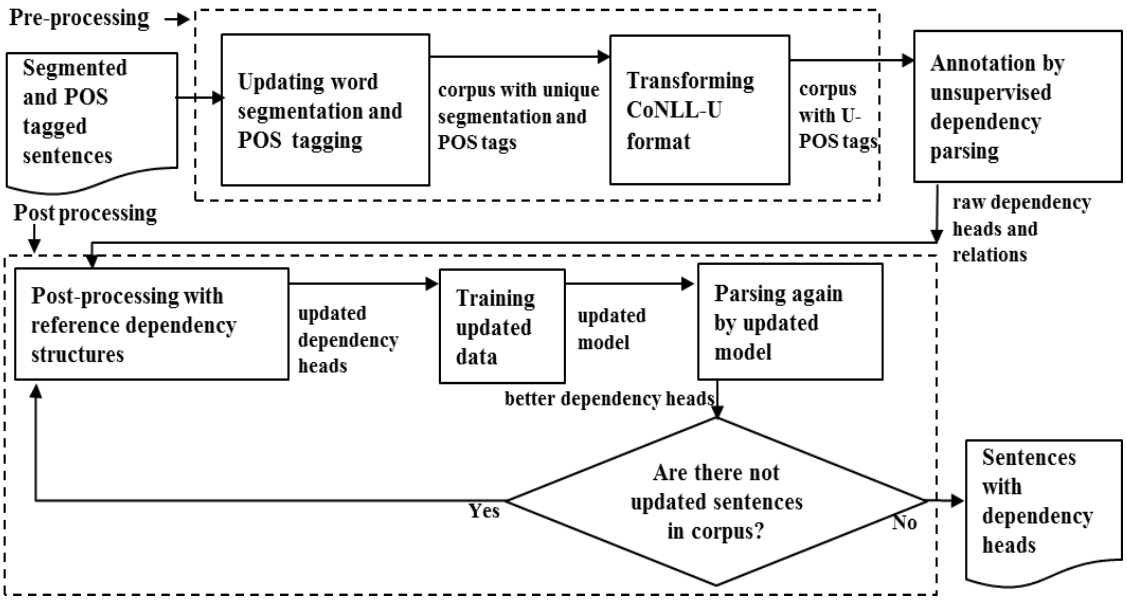

Figure 2: Overview architecture of dependency head annotation

Table 3: Example POS tagging for same word

Words	Correct Tags	Meanings	Examples in phrases
နေ	verb	live	Myanmar : ရန်ကုန် မှာ နေ Words in English: Yangon in live - Translation : Live in Yangon.
နေ	particle	describing continuous action	Myanmar : လေ့လာ နေ သည် Words in English: study - - Translation : is studying

Table 4: Myanmar POS tags for language specific tag set

POS Tag	Description	Example words
N	Noun	သတင်းစာ (Newspaper), ခန်းမ (Hall), ကုမ္ပဏီ (Company)
PRPN	Proper Noun	စက်တင်ဘာ (September), အာရှ (Asia)
NUM	Number	ဝ (0), ၁(1), ၂(2), ၃(3), ၄(4)
TNUM	Number text letter	သုည (zero) , တစ် (one) , နှစ် (two), သုံး (three), လေး(four)
FOR	Foreign word	Carsh, Federal, process, Bit
ABB	Abbreviation	ညွှန်/ချုပ် (Director-General)
PRON	Pronoun	ကျွန်တော်/ ကျွန်မ/ ကျွန်ုပ် (I), ထို (that), မည်သူ (who)
ADJ	Adjective	ကျယ်ဝန်း (wide), ယဉ်ကျေး (polite), မြင့်မား (high), လေးလံ (heavy)
ADV	Adverb	မကြာခဏ (frequently), အလွန် (very), လောလောဆယ် (currently)
V	Verb	စား (eat), သွား (go), ရေးသား (write), ဖြစ် (be), ရှိ (has/have), တည်ရှိ (exist)
CONJ	Conjunction	နှင့် (and), သို့မဟုတ် (or), ၍/သောကြောင့် (because)
PART	Particle	များ/တို့/တွေ (plural marker), ခန့် (about), ခဲ့ (past marker), နိုင်(can), ကြ (plural action marker)
PPM	Post positional marker	သည်/က/မှာ (nominal marker for subject), ၌ (at) , ၏ (of) , ဖြင့် (by)
PUNC	Punctuation	၊ , ။ , “ , (or -LRB-,) or -RRB- , “, ”
SB	Symbol	% , $
INT	Interjection	အိုး (Oh), အမလေး (Oh my god!)

Table 5: Mapping scheme between Universal POS and Myanmar language specific POS tags

U-POS Tag	Description [Examples]	Myanmar Language POS
NOUN	Noun	N
		FOR
PROPN	Proper Noun	PRPN
		ABB
NUM	Number	NUM
		TNUM
PRON	Pronoun	PRON
ADJ	Adjective	ADJ
ADV	Adverb	ADV
VERB	Verb	V
CCONJ	Coordinating Conjunction [နှင့် (and), သို့မဟုတ် (or)]	CONJ
SCONJ	Subordinating Conjunction [၍/ သောကြောင့် (because), နှင့်တစ်ပြိုင်နက် (as soon as), ပါက (if)]	
PART	Particle	PART
ADP	Adposition	PPM
PUNCT	Punctuation	PUNCT
SYM	Symbol	SB
INTJ	Interjection	INT

Table 6: Mapping scheme between Universal POS and Myanmar language specific POS tags

U-POS Tag	Description	myPOS	ALT	Web News	Total
NOUN	Noun	55,590	71,246	18,237	145,073
PRON	Pronoun	2,680	7,864	331	10,875
PROPN	Proper Noun	12,377	23,168	3,375	38,920
NUM	Number	6,261	11,240	2,259	19,760
ADJ	Adjective	7,108	7,776	1,393	16,277
ADV	Adverb	2,868	4,686	852	8,406
VERB	Verb	34,046	62,876	11,057	107,979
CCONJ	Coordinating Conjunction	4,624	6,368	1,584	12,576
SCONJ	Subordinating Conjunction	6,587	7,493	1,452	15,532
PART	Particle	52,413	78,268	13,294	143,975
ADP	Adposition	38,838	64,624	8,674	112,136
PUNCT	Punctuation	15,845	28,255	4,199	48,299
SYM	Symbol	199	116	3	318
INTJ	Interjection	98	0	0	98

All conjunctions in Myanmar sentences were tagged as CONJ which were mapped to one of two conjunctions of U-POS tags: CCONJ and SCONJ, coordinating and subordinating conjunction respectively. Mapping scheme between U-POS tags and Myanmar POS tags can be seen in Table 5. Total frequencies of U-POS tags in each corpus of treebank are listed in Table 6.

6. Post-processing

Annotation method and types of annotation schemes are important in dependency treebank construction. The annotation was based on Universal Dependencies. In the UD annotation scheme, dependency relations are expressed between words and main content words attached leaf words such as function words in sentences [22]. Main parts of syntactic structures are head nodes and relation links called as dependency relation labels between words.

Generally Myanmar nouns, adjectives, and verbs are formally written with suffixes such as post positional markers or particles. Currently, dependency relation labels are automatic unsupervised annotated results without post processing in treebank. Only dependency head information was post processed.

Overview of the linking structures of dependency head node words between dependent words in most occurred phrases in sentences had been presented in our previous work [17]. In that work, only construction of dependency link arc connections between heads and dependents words had been presented. Therefore, the arcs directed to the heads from dependents.

After the previous work, unsupervised dependency annotation results have been post processed to update dependency head links based on the dependency structures described in our previous work. Updating dependency relation labels needs more time because of few annotators and issues of sentence writing styles. Therefore, dependency relations labels are still automatic unsupervised annotated labels after post processing. To describe full referenced dependency information, in this paper, dependency relation labels between head node words and dependent words will be presented with unsupervised annotated labels in sample sentences according to the viewpoint of the language typology.

6.1. Proper Noun and Possessive Phrase

The specific unique names of common nouns are called as proper nouns as in other languages. Myanmar proper nouns are usually found in before or after common nouns [21]. In Figure 3, two proper nouns, following two common nouns can be seen in example sentence. In that sentence, example possessive case can also be seen. Possessive case of a noun or pronoun can be written by a post positional suffix marker, “ ၏”, to show possession of following right noun by left proper noun of it.

6.2. Compound Noun

Myanmar compound nouns might have two or more words and POS tags of these words can be nouns, verbs, adjectives or adverbs [21]. An example compound noun in sentence of Figure 4 contains two consecutive nouns and means “memorial stamp”. In this case, the left noun modifies the right one.

6.3. Numeral Phrase

Most numeral noun phrases can be written by three general formats which are described in Table 7 with their meanings. Counting amount can be written by digit number or text number [21]. Sample numeral phrases in sentence can be seen in Figure 5.

6.4. Adjective Phrase

Adjectives modify nouns and are usually placed before or after noun in Myanmar sentences. Myanmar adjectives can be written as simple or simple adjective with suffix particles or transformed adjective [21]. Suffixes of adjective or example adjective phrases are presented in Table 8. Sample dependency of adjective phrase in sentence can be seen in Figure 6.

6.5. Adverb Phrase and Verb Phrase

Adverbs can be written as simple or transformed adverb with the suffix particle, “ စွာ”, followed by verb, or adjective, or adverb.

Example adverb phrases can be seen in Table 9. Example dependency of adverb phrase in sentence can be seen in Figure 7.

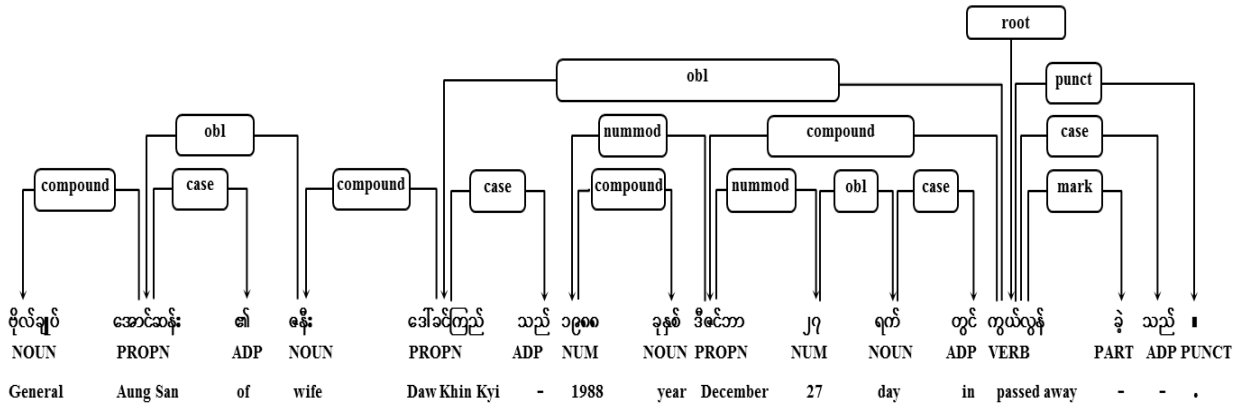

Figure 3: Proper noun and possessive case examples in sentence

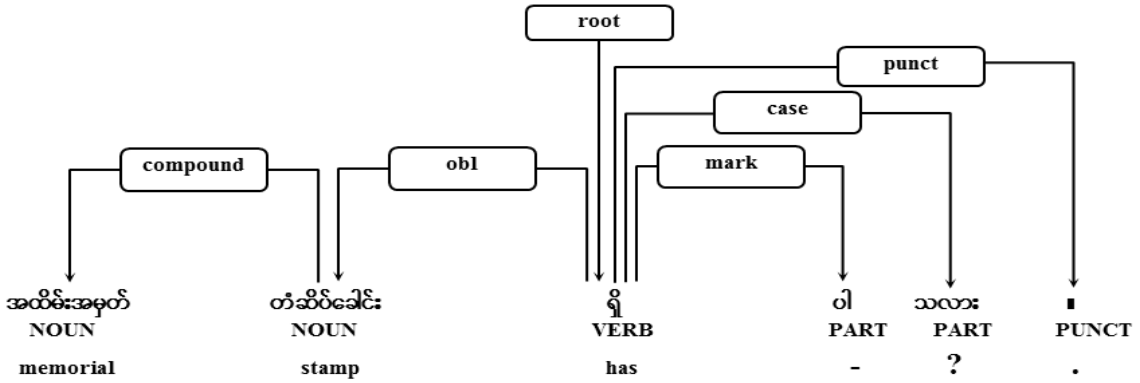

Figure 4: Compound noun example in sentence

Table 7: Numeral phrase formats

Numeral Phrase Forms	Example	Meaning	Remarks
N NUM/TNUM PART/N	မှတ်ချက် ၂/နှစ် ချက် comment 2/two - (Glossary)	2/two comments	
N N (noun affixed particle, "အ", to form common nouns) NUM/TNUM	မှတ်ချက် အချက် နှစ်ဆယ် comment fact twenty (Glossary)	twenty comments	if counted amount is exact numbers of multiple of ten, hundred, thousand, etc.
N N(transformed or common noun) NUM/TNUM PART/N	မှတ်ချက် အချက် နှစ်ဆယ် comment fact twenty two - (Glossary)	twenty two comments	

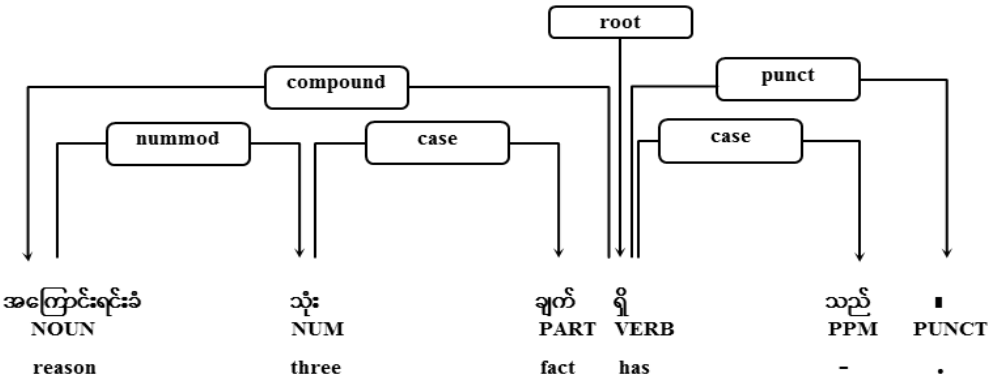

Figure 5: Numeral phrase example in sentence

Table 8: Suffixes of adjective or adjective phrases and examples

Particles used to suffix or transform adjective or adjective phrase		သော, သည့်, မည့်, တဲ့
Example Myanmar adjective	***Translations***	***Remarks***
ကြီး သော အိမ် **big** - house	big house	Simple adjective followed suffix
အိမ် **ကြီး** house **big**	big house	Simple adjective
အစိုးရ ပေး သည့် အိမ် government provide - house	house provided by government	Transformed adjective

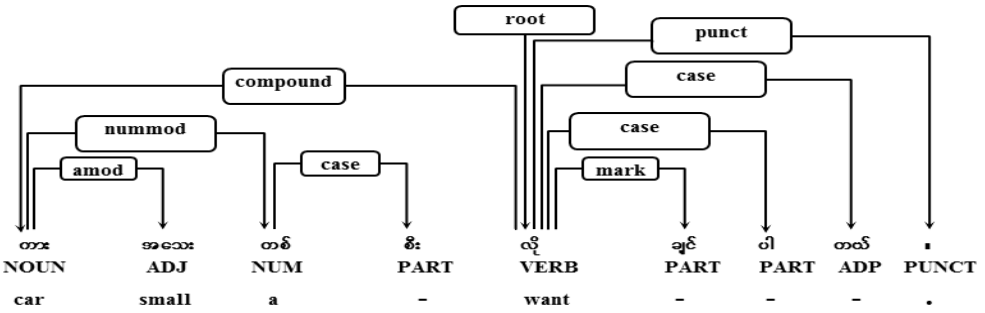

Figure 6: Adjective phrase example in sentence

Table 9: Example of adverb forms

Example Myanmar adverb phrases	**Meanings**	**Remarks**
မြန်မြန်(quickly)	quickly	Simple adverb
မြန်မြန်သွက်သွက် (quickly)	quickly	Simple adverb
လျင်မြန် စွာ quick -	quickly	Transformed adverb with suffix particle

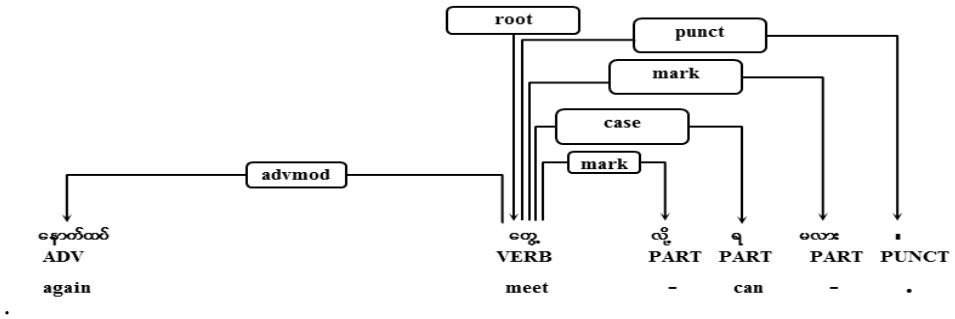

Figure 7: Adverb and verb phrase example in sentence

Table 10: Suffixes of verbs

Suffixes to verb	**Description of usage and example**	**Suffix Type**
သည်, ၏, ပြီ, မည်, မယ်, တယ်	to form verb by showing tense Examples: သွား သည် (go) သွား မယ် (will go)	post positional markers
မှာ, ပါ, စမ်း	to express giving order or answering action Examples: လုပ် ပါ (giving order or requesting “Do” action) မောင်း မှာ (answering or forecasting “drive” action based on the meaning of content words in sentence)	particles

Verb phrases in sentences are usually formed by one of the verb ended post positional markers or particles [21] as presented in Table 10 .

In addition, zero or more particles can be followed by verbs before verb ended markers to express the full state of action. In example sentence of Figure 7, the verb phrase, " တွေ့လို့ရမလား", means "can meet" and "တွေ့" is main verb and the suffix sequence with three particles, "လို့ ရ မလား", means asking politely for the requested action. That main verb phrase is also modified by left adverb phrase, " နောက်ထပ်". As a result, the verb, "တွေ့", is root of that sentence. The whole sentence means "Can meet again?",

6.6. Conjunctions

In Myanmar language, conjunctions are used to combine not only clauses or simple sentences but also related words or phrases. Moreover, simple sentences can be connected by suffixes: post positional markers, " က, မှာ, ကို", or particles "ဟု , သော , သည့် , မည့်, တဲ့" to form noun or adjective clause in sentence. If two or more simple sentences are connected by post positional markers or particles or conjunctions, combined sentence becomes complex sentence and clauses represent the roles as subjects or objects or adverbs. Most clauses ended by conjunctions are usually adverb modifiers in main sentence. Therefore, the role of dependent clauses can be divided into three types: noun clause, adjective clause, and adverb clause based on their roles in sentences. If combined clauses or sentences contains the same subject, subject can be omitted in dependent clause or independent clause or in both. Similarly, object noun can also be omitted in dependent clause or independent clause if it is placed in one [21, 23].

Some conjunctions described in Table 11 are used to connect words, phrases to connect two sentences to give coordinated extra meaning [21, 23].

Table 11: Example conjunctions

Conjunctions	**Usage**
နှင့် , လည်းကောင်း…လည်းကောင်း, ရော…ပါ, ရော… ရော, သို့မဟုတ်, ဖြစ်စေ…ဖြစ်စေ, ဖြစ်ဖြစ်…ဖြစ်ဖြစ်, သော်လည်းကောင်း…သော်လည်းကောင်း, မှတစ်ပါး, ပြီး	to connect words, phrases in sentence for extra meaning
ထို့ပြင်, ထို့အပြင်, ၎င်းပြင်, သည့်ပြင်	to give connection between prior sentence and next by giving coordinated extra meaning.

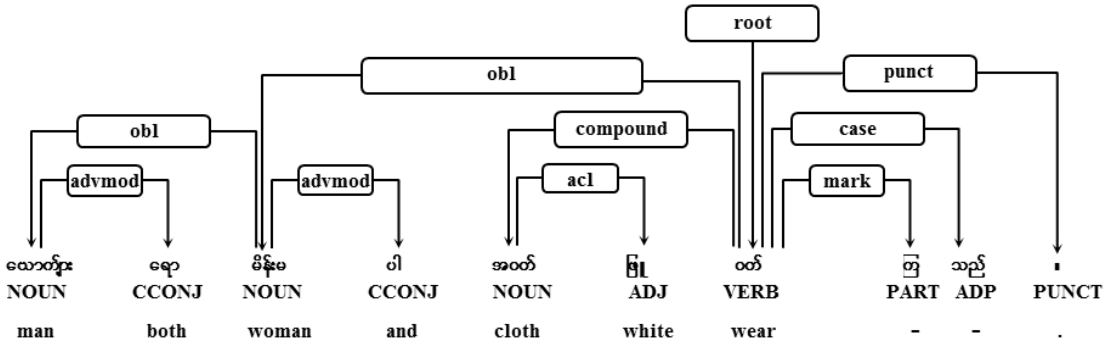

Figure 8: Coordinated phrase example in sentence

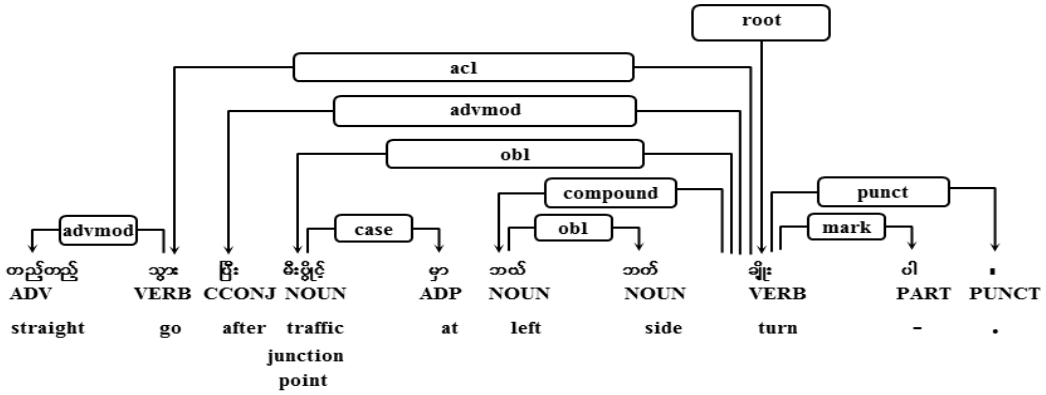

Figure 9: Coordinated clauses example in sentence

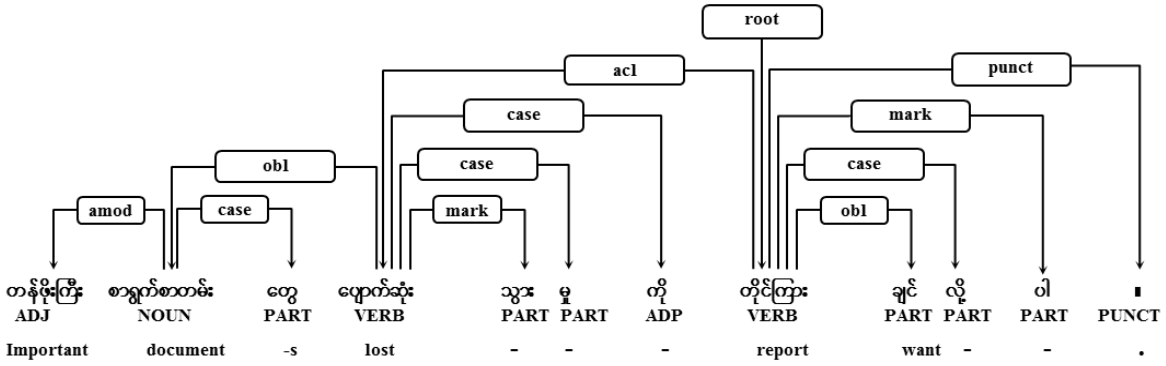

Figure 10: Complex sentence with noun clause

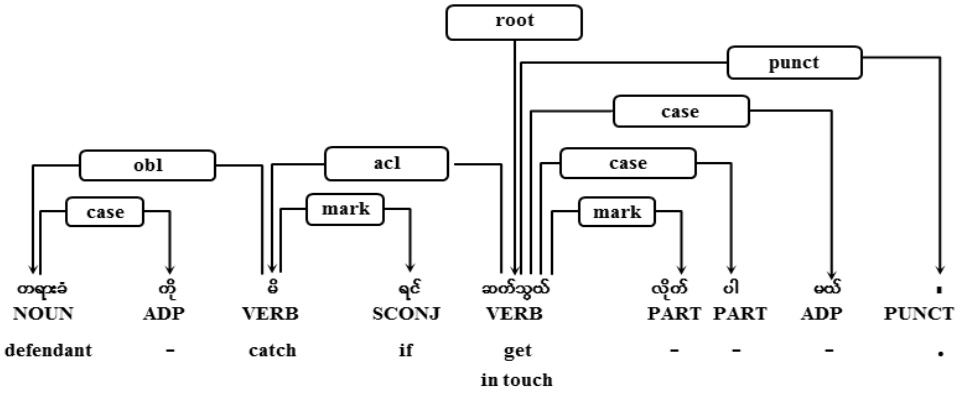

Figure 11: Complex sentence joined with subordinating conjunction

Table 12: Average parsing precision scores of each corpus

Corpus	Sentences		Average Scores	
	Train	Test	UAS	LAS
myPOS	10,010	1,000	89.06	86.67
Web News	1,620	180	87.03	84.83
ALT	9,000	1,000	91.40	90.24

Sample coordinated phrase with conjunctions in sentence can be seen in Figure 8. Moreover, complex sentence combined two simple sentences by coordinating conjunction can also be seen in Figure 9. The left side of conjunction is dependent sentence or clause to give first action and the right side of conjunction is independent clause to give final action as shown in Figure 9. Subjects of clauses are omitted and main sentence means "After go straight, turn left at the traffic junction point.". Therefore, the conjunction is tagged as "CCONJ", coordinating conjunction, for U-POS tag.

Some Myanmar conjunctions such as "လျှင်/ရင်/ပါက, သောကြောင့်, နှင့်တစ်ပြိုင်နက်, သကဲ့သို့, စေရန်, သောအခါ, သော်လည်း,.." , "if, because, as soon as, as/like, in order to, when, although,.." , are frequently used to connect clauses to provide the required meaning for main clauses [21, 23]. Therefore, these types of conjunctions are tagged as "SCONJ", subordinating conjunction, for U-POS tag.

Example complex sentence including noun clause as an object role in sentence can be seen in Figure 10. Main complex sentence means "Lost of important documents is wanted to be reported." The PPM, " ကို", is used to connect the left clause to represent as object noun for the action of root verb, "တိုင်ကြား", of main sentence.

In addition, example complex sentence joined with subordinating conjunction type, " ရင်" (if) , can be seen in Figure 11. Subjects are omitted in both dependent clause and main independent clause. It means that "If defendant is caught, I will get in touch".

7. Parsing Experiment

The main purpose of building treebank is to use in dependency parsing. This section presents parsing experiments executed for performance of treebank. UDPipe is an open-source trainable pipeline processing tool to perform segmentation, POS tagging, lemmatization and dependency parsing without any other external

data for multiple languages. And it is also available under open-source Mozilla Public Licence (MPL) and provides bindings for C++, Python, Perl, Java and C#. UDPipe 1.2 has been used in our parsing experiments.

We divided 90% and 10% of total sentences from each corpus of treebank for training and test data to evaluate each corpus by 10-fold cross validation test. We arranged alternate order of test sentences range and all the rest for train data in each corpus and split them into equally-sized parts for each validation test by python script. We calculated average score of all validation tests to report as total average score of cross validation test for each corpus.

8. Results and Evaluation

In this section, the statistical results of parsing experiments to evaluate parsing performance of treebank and evaluation results will be described. For each corpus performance of treebank, average parsing precision scores are calculated from the total results of cross validation tests in each corpus and they are listed in Table 12. Average precision scores measured by unlabeled attachment score (UAS) and label attachment score (LAS) of each corpus are over 89% and 86%, over 87% and 84 %, and 91% and 90% in myPOS, Web News, and ALT respectively .

Each corpus of treebank was evaluated by the CoNLL 2017 UD parsing evaluation script. The evaluation results will be described in following sub section. In addition, syntax structures of the referenced dependency types were also analyzed by manual python script which counts types of dependency structures being countable types in each sentence. The analyzed results will be described in next sub sections.

8.1. Evaluation

Corpora contained in treebank were evaluated by the CoNLL 2017 UD parsing evaluation script, "conll17_ud_eval.py" [24], to know their parsing accuracy. Gold standard file and system output file are input to the evaluation script to evaluate the data. The parsing model was trained with all current data of treebank to generate system output of each corpus. The current accuracies comparing standard annotated data with system output parsing result of each corpus can be seen in Figure 12.

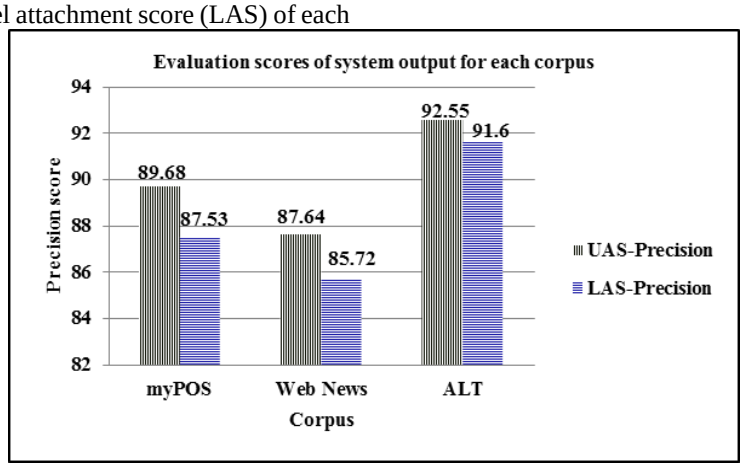

Figure 12: Precision scores of treebank data

Table 13: Information of sentence lengths in treebank

Sentence Type	myPOS	ALT	Web News	Total sentences of each range	Remark
Simple short	2,253	13	74	2,340	<=10 words
Short	6,463	3,858	732	11,053	>=11 and <=30 words
Normal range	1,929	4,398	568	6,895	>=31 and <=50 words
Long	349	1,687	401	2,437	>=51 and <=100 words
Very long	16	44	25	85	>100 words

Table 14: Phrases and clauses structures in treebank

Phrase Types	myPOS	ALT	Web News	Total
Noun	27,083	48,466	6,316	81,865
Proper Noun	10,514	17,750	3,013	31,277
Numeral Noun	5,790	10,852	2,125	18,767
Compound Noun	21,096	25,706	9,537	56,339
Adjective	9,535	11,993	1,581	23,109
Adverb	3,027	4,580	837	8,444
Verb	16,593	19,136	3,584	39,313
Phrases with Conjunctions	2,830	3,534	921	7,285
Clauses	7,027	9,837	1,950	18,814

Precision scores of UAS and LAS in system outputs of myPOS and Web News data are over 89% and 87%, and 87% and 85% respectively in comparing manually updated standard data. System output precision scores of UAS and LAS for ALT data is over 92% and 91% in comparing with automatic unsupervised annotated standard data.

As sentences of ALT corpus are longer than myPOS and Web News corpus, precision scores of unsupervised annotated ALT is more than manual updated standard data of myPOS and Web News corpus. However, system outputs of myPOS and Web News corpus are better and more similar to reference dependency structures than system outputs of ALT in manual random checking on system output trees.

Although currently Myanmar ALT has been developed manually, there is still no dependency resource for Myanmar. Constructing treebank manually is an error-prone and slow process. The high precision scores of cross validation tests and evaluations by post processed model illustrate that the proposed method can produce efficient precision scores for dependency parsing. With consistent and faster annotation, the proposed method will provide fast dependency tree building for Myanmar which has free word order and issues mentioned in Section 2

The length of sentences is also one main part of treebank characteristics for having various types of syntax and syntactic structures. Therefore, the ranges of sentences in treebank were also analyzed by classifying five levels by python script and resulted sentence ranges are listed in Table 13.

8.2. Syntax Analysis

The syntax information is one of the most important characteristics of treebank to know syntax status. It is difficult to count all syntax information of natural language sentences by program script exactly because of the issues of phrases and many sentence construction types discussed in Section 2. However, overview syntax structures of formal sentences from each corpus of treebank could be extracted by python script based on sequence order of words and POS tags information of phrases and clauses written by formal literature written style. The program counts the related phrase and clause types from word and POS tag sequences of sentences annotated as example dependency structure types described in Section 6. However, some informal phrases and clauses not ended with the formal related post positional markers or particles, could not be counted by the program. Overall countable syntax structures are listed in Table 14.

9. Conclusion

This paper has presented annotating dependency heads based on Universal Dependencies framework for Myanmar dependency treebank. Myanmar POS tags and U-POS tags of treebank, issues of tagging, and mapping scheme between two tag sets have also been presented. In addition, dependency head annotation schemes have also been described with sample sentences in line with grammatical point of views. This work is first for Myanmar to the best of our knowledge. Parsing experiments have also been executed for performance of treebank and results have also been presented. To conclude, contribution is first dependency head annotation for building Myanmar dependency treebank and can provide useful information for direct dependency parsing for Myanmar sentences by Myanmar model in future.

As future work, firstly we would intend to add more annotated sentences from same and different domains such as News articles data and Myanmar grammar books data to treebank because correct syntactic forms are able to provide faster annotation and better useful syntax information for Myanmar dependency treebank. The next work is to update unsupervised dependency labels according to the standard of Universal Dependencies based on Myanmar grammatical point of views.

References

[1] N. Green, “Dependency Parsing”, In WDS 2011 Proceedings of Contributed Papers, WDS 2011, 137–142, 2011, doi:10.1007/1-4020-4889-0_3..

[2] T. Kakkonen, “Dependency treebanks: methods, annotation schemes and tools”, In Proceedings of the 15th NODALIDA conference, 94–104, 2005, doi: arXiv:cs/0610124

[3] P. E. Solberg, “Building gold-standard treebanks for Norwegian”, In Proceedings of the 19th Nordic Conference of Computational Linguistics, 2013 (NODALIDA 2013), Linköping University Electronic Press, 459-464, 2013.

[4] K. Haverinen, J. Nyblom, T. Viljanen, V. Laippala, S. Kohonen, A. Missilä, S. Ojala, T. Salakoski, and F. Ginter, “Building the essential resources for Finnish: the Turku Dependency Treebank”, Language Resources & Evaluation, **48**(3), 493-531, 2014, doi: 10.1007/s10579-013-9244-1.

[5] C. MĂRĂNDUC, and C. A. PEREZ “A Romanian Dependency Treebank”, In International Journal of Computational Linguistics and Applications, **6**(2), 83-103,2015.

[6] D. Bamman, and G. Crane, "The Ancient Greek and Latin Dependency Treebanks," In Language Technology for Cultural Heritage 2011, Springer, 79-98, 2011, doi:10.1007/978-3-642-20227-8_5.

[7] K. Nguyen, “BKTreebank: Building a Vietnamese Dependency Treebank”, In Proceedings of the 11th International Conference on Language Resources and Evaluation 2018 (LREC-2018),European Languages Resources Association (ELRA), 2164-2168, 2018.

[8] C. Ding, Y. K. Thu, M. Utiyama, A. M. Finch, and E. Sumita, “Empirical Dependency-Based Head Finalization for Statistical Chinese-, English-, and French-to-Myanmar (Burmese) Machine Translation”, in Proceedings of the 11th International Workshop on Spoken Language Translation, 184-191, 2014.

[9] C. Ding, Y. K. Thu, M. Utiyama, and E. Sumita, “Parsing Myanmar (Burmese) by Using Japanese as a Pivot”, Proceedings of 14th International Conference on Computer Applications, 158-162, 2016

[10] D. Mareˇcek, “Twelve Years of Unsupervised Dependency Parsing”, ITAT 2016 Proceedings, CEUR Workshop Proceedings , **1649**, 56–62, 2016..

[11] S. Petrov, D. Das, and R. McDonald, “A Universal Part-of-Speech Tagset”, In Proceedings of the 8th International Conference on Language Resources and Evaluation, 2089-2096, 2012.

[12] M. Straka, J. Hajic, and J. Strakov ˇ a´, “UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing,” In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 4290–4297, 2016.

[13] J. Chun, N. Han, J. D. Hwang, and J. D. Choi, “Building Universal Dependency Treebanks in Korean”, In Proceedings of the 11th International Conference on Language Resources and Evaluation 2018 (LREC-2018). European Languages Resources Association (ELRA), 2194-2202, 2018.

[14] E. Badmaeva, and F. M. Tyers, “A Dependency Treebank for Buryat”, In 15th Internationtal Workshop on Treebanks and Linguistic Theories (TLT15), 1-12, 2017.

[15] T. Rama, and S. Vajjala, “A Dependency Treebank for Telgu”, In Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), 119–128, 2018.

[16] Y. Kyaw Thu,, W. Pa Pa, M. Utiyama, A. Finch, and E. Sumita, “Introducing the Asian Language Treebank (ALT)”, In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16) 2016 May , 1574-1578, 2016. DOI: 10.1109/ICSDA.2016.7918974

[17] H. T. Z. Aye, W. P. Pa, and Y. K. Thu, “Unsupervised Dependency Corpus Annotation For Myanmar Language”, In Proceedings of the 2018 Oriental COCOSDA-International Conference on Speech Database and Assessments, IEEE, 78-83, 2018, doi:10.1109/ICSDA.2018.8693009

[18] K. W. W. Htike, Y. K. Thu, Z. Zhang, W. P. Pa, Y. Sagisaka, and N. Iwahashi, "Comparison of Six POS Tagging Methods on 10K Sentences Myanmar Language (Burmese) POS Tagged Corpus", In Proceedings of the CICLING 2017, April 2017..

[19] S. Mori, H. Ogura, and T. Sasada, "A Japanese word dependency corpus", Proceedings of the 9th International Conference on Language Resources and Evaluation 2014 May 26, 753–758, 2014.

[20] T. Tanaka, Y. Miyao, M. Asahara, S. Uematsu, H. Kanayama, S. Mori, and Y. Matsumoto, "Universal Dependencies for Japanese", In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16) , 1651-1658, May 2016.

[21] Department of the Myanmar Language Committee, Myanmar Grammar, Ministry of Education, The Republic of the Union of Myanmar, 2013.

[22] P. Osenova, and K. Simov, "Treebanks, Linguistic Theories and Applications Multilingual Treebanks: the Universal Dependencies Case", In 30th European Summer School in Logic, Language and Information, 2018.

[23] U Thaung Lwin, Advanced New Method Myanmar Grammar) Second Edition, New Method Book Store (နည်းသစ်စာအုပ်တိုက်), 2015.

[24] A. Kutuzov, "Dependency Parsing (Project A) ", In INF5830, Fall 2017,University of Oslo, 2017