Final Project Report
Detailed Form

I. Title of Proposed Project:

Open Collaboration for Developing and Using Asian Language Treebank

II. Project Leader:

Full name: Masao Utiyama
Institution: National Institute of Information and Communications Technology, Japan
Address: 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan
Phone: +81-774-98-6343
E-mail: mutiyama@nict.go.jp

III. Project Members:

Name	Position/Degree	Department, Institution, Country	Email Address
Hammam Riza	Deputy Chairman IT,Energy and Material / PhD	BPPT, Indonesia	hammam.riza@bppt.go.id
Aw Ai Ti	Unit Head, Human Language Technologies / Ms.	HLT, I2R, Singapore	aaiti@i2r.a-star.edu.sg
Luong Chi Mai	Assoc. Prof. / PhD	Multimedia Human-Machine Language Technology, IOIT, Vietnam	lcmai@ioit.ac.vn
Sethserey Sam	Vice President of Research / PhD	NIPTICT, Cambodia	sethserey.sam@niptict.edu.kh
Khin Mar Soe	Professor / PhD	NLP Lab,UCSY, Myanmar	khinmarsoe@ucsy.edu.mm

Masao Utiyama	Executive Researcher / PhD	UCRI, NICT, Japan	mutiyama@nict.go.jp
Thepchai Supnithi	Research Team Leader/ Principal Researcher	NECTEC, Thailand	Thepchai.Supnithi@nectec.or.th
Ria A. Sagum	Assoc. Prof. / CCIS Faculty Researcher	PUP, Philippines	rasagum@pup.edu.ph riasagum31@gmail.com

IV. Project Report

i) Introduction

The ASEAN Economic Community (AEC) envisages the following key characteristics [1]:

- a single market and production base,
- a highly competitive economic region,
- a region of equitable economic development,
- a region fully integrated into the global economy.

The population of ACE is over 600 million and they speak many different languages. Consequently, natural language processing (NLP) is necessary to cope with many languages, in order to provide "ICT solutions to the Challenges surrounding Urbanization" and make "Social Renovation in Rural Areas and/or Urban Areas."

NLP is one of the core technologies in ICT. This is because the contents of information are conveyed by natural languages, such as English, Indonesian, Japanese, Khmer, Malay, Myanmar, Vietnamese, and so on. For example, Web search engines, machine translation engines, and input method editors all use NLP to analyze the contents.

The state of the art technologies in NLP are based on treebanks. A treebank is a linguistic knowledge representation of natural language texts. The basic linguistic annotations in treebanks are word segmentation, part-of-speech (POS) tagging, and parsing annotations. Those annotations are used to produce NLP tools, such as word segmenters, POS taggers, and syntax parsers. Almost all NLP researches and tools are based on treebanks in a broad sense [1][2][3][4].

The main problem of the creation of a treebank is that it needs a lot of linguistic knowledge for the language. Each language treebank needs each language expertise. As a result, existing treebanks are limited in their sizes, annotation types and languages. In particular, there have been no publicly available POS-tagged and constituency tree corpora for most of Asian languages before our project. (Though, some corpora are available for some languages).

This background has made us propose this project for developing Asian Language Treebank (ALT). The objective of ALT is developing a parallel treebank for Asian languages. The benefits of ALT to the society is immense. ALT will accelerate research of NLP for Asian languages, such as Indonesian, Vietnamese, Japanese, Khmer, Laos, Malay, Myanmar, Philippine, Thai, and so on. This will result in the better communication in the ASEAN region and the world.

After three years of the ALT project, we have developed treebanks and software, which are available from the project website:

http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html

ii) Project Activities

(1) Development and Implement

One of the characteristics of ALT is it uses the parallel sentences for creating treebanks. This means that each sentence will be translated into several languages and be annotated with several languages. In addition, word alignment links between different language words will also be provided using English as the pivot language.

This characteristics is unique. No other treebanks have this property among Asian languages. It has both scientific and engineering benefits. For the scientific benefits, we can compare various NLP techniques as applied to various languages on the same ground. This will reveal how NLP techniques can be applicable to different languages. For the engineering benefits, we can easily transfer the NLP techniques available in resource rich languages (for example, English and Japanese) into other languages.

The development of ALT for each language has been conducted by each member institute, which is a top-level NLP research institute for that language. They have their own tools for accelerating the development of ALT [1][2][3][4]. Please refer to "Findings and Outcomes" for details. In this section, we describe the common settings for the development of ALT.

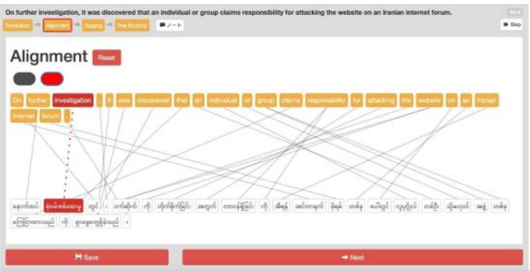
ALT comprises about 20,000 sentences originally sampled from the English Wikinews. English Wikinews was selected as the source texts because its license is Creative Commons Attribution 2.5 License. We can use this corpus freely on the condition that we mention the copyright. 1893 articles were randomly selected from the archive of English Wikinews in 2014. Those 20,000 sentences were translated into Indonesian, Malay, Vietnamese, Khmer, Myanmar, Thai, Filipino and Japanese languages.

Here are sample URLs and sample texts.
Sample URLs:
http://en.wikinews.org/wiki/Steve_Wright,_killer_of_five_women_in_Suffolk,_England,_sentenced_to_life_imprisonment
http://en.wikinews.org/wiki/Wikinews_interviews_U.S._Libertarian_presidential_candidate_James_Burns
http://en.wikinews.org/wiki/Kahne_takes_Checkered_Flag_at_Richmond
http://en.wikinews.org/wiki/Pickens_County,_South_Carolina_sheriff_refuses_to_lower_US_flag_to_honor_Mandela
Sample Texts:
Italy have defeated Portugal 31-5 in Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France.
Andrea Masi opened the scoring in the fourth minute with a try for Italy.
Despite controlling the game for much of the first half, Italy could not score any other tries before the interval but David Bortolussi kicked three penalties to extend their lead.
Portugal never gave up and David Penalva scored a try in the 33rd minute, providing their only points of the match.

NICT built an annotation server for developing ALT using the English texts above. The member institutes used this ALT annotation server or their original tools for building ALT. The annotation server helps all steps of the development; translation, word segmentation, word alignment, POS tagging, and syntax annotation.

The data are represented in an XML format. Thus, the server is flexible enough to accept different POS tags for different languages, for example. It is also possible to mix automatic and manual annotations. That is, at first, automatic analysis is conducted by NLP tools. Then, it can be manually corrected using the server.

This server also supports users, groups and task management and has a multilingual help system. The administrator can assign particular tasks to users considering their linguistic expertise (such as assigning only translation tasks etc.). If the user is assigned all tasks he/she can continuously and efficiently work through all processes for the given sentence. For word alignment, POS tagging and syntax annotation, user input can be performed using only the mouse. Fig. 1 shows the user interface for word alignment annotation between an English sentence and the corresponding translated Myanmar sentence, and Fig. 2 shows the user interface for syntax annotation, or constituency tree building [6].



On further investigation, it was discovered that an individual or group claims responsibility for attacking the website on an Iranian internet forum.

Fig. 1. Word alignment interface

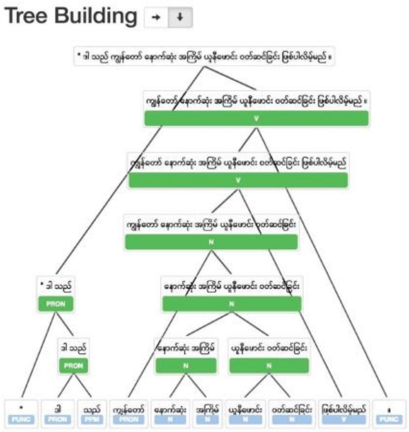


Fig. 2. Tree building interface

(2) Leveraged Resources and Participants

ASEAN IVO is an ideal organization for developing ALT, because it consists of top-level NLP research institutes for Asian languages. Without ASEAN IVO, it would be impossible to corporate and cover main Asian languages for building treebanks.

In this project, BPPT, I2R, IOIT, NIPTICT, UCSY, NECTEC, PUP and NICT have developed ALT for Indonesian, Malay, Vietnamese, Khmer, Myanmar, Thai, Filipino and Japanese languages, respectively. (NICT has also developed English ALT). Those different language treebanks have been built from the translated Wikinews sentences (about 20,000 sentences).

The members of this project are as follows:

- BPPT
  - ➢ Dr. Hammam Riza, Deputy Chairman IT, Energy and Material
  - ➢ Dr. Michael Purwoadi, Director ICT Center, oversee the Intelligent computing and Language Technology activities in BPPT
  - ➢ Gunarso, Leader of Language Technology working group
  - ➢ Dr. Teduh Uliniansyah, Researcher of Language Technology working group
- I2R
  - ➢ Ms Aw Ai Ti, a senior researcher at I^{2}R, who is an expert in NLP and machine translation
  - ➢ Ms Nabilah Binte Md Johan, a research engineer at I^{2}R, who is an expert in English and Malay Linguistics.
- IOIT / VNU UET
  - ➢ Vu Tat Thang, PhD.
  - ➢ Luong Chi Mai, Assoc. Prof., PhD.
  - ➢ Nguyen Phuong Thais, Assoc. Prof, PhD. (VNU University of Engineering and Technology)
- NIPTICT
  - ➢ Dr. Sethserey Sam, Vice President of Research, who is the supervisor of NLP projects
  - ➢ Mr. Ly Rottana, researcher at NIPTICT, who is an expert in NLP and machine translation.
- UCSY
  - ➢ Dr. Khin Mar Soe, a Professor at NLP lab, UCSY, who is currently doing research in NLP and machine translation.
  - ➢ Dr. Khin Thandar Nwet, a researcher at NLP lab, UCSY, who is currently doing research in NLP and machine translation.
- NICT
  - ➢ Dr. Masao Utiyama, an executive researcher at NICT, who is an expert in NLP and machine translation
  - ➢ Dr. Chenchen Ding, a researcher at NICT, who is an expert in NLP and machine translation
- NECTEC
  - ➢ Dr. Thepchai Supnithi, a principal researcher at NECTEC who is the supervisor of Language and Semantic technology team
  - ➢ Dr. Prachya Boonkwan, a researcher at NECTEC, who is currently doing research in NLP and machine translation.
- PUP (all are member of PUP-CCIS faculty, who are currently doing research in NLP and machine translation)
  - ➢ Assoc. Prof. Ria A. Sagum
  - ➢ Assoc. Prof. Michael B. dela Fuente

- Asst. Prof. Carlo G. Inovero
- Ms. Janelle Kyra A. Sagum

(3) Findings and Outcomes

We elaborate the findings and outcomes for each language below from the point of each institute. We also describe how we developed these findings and outcomes here.

- BPPT

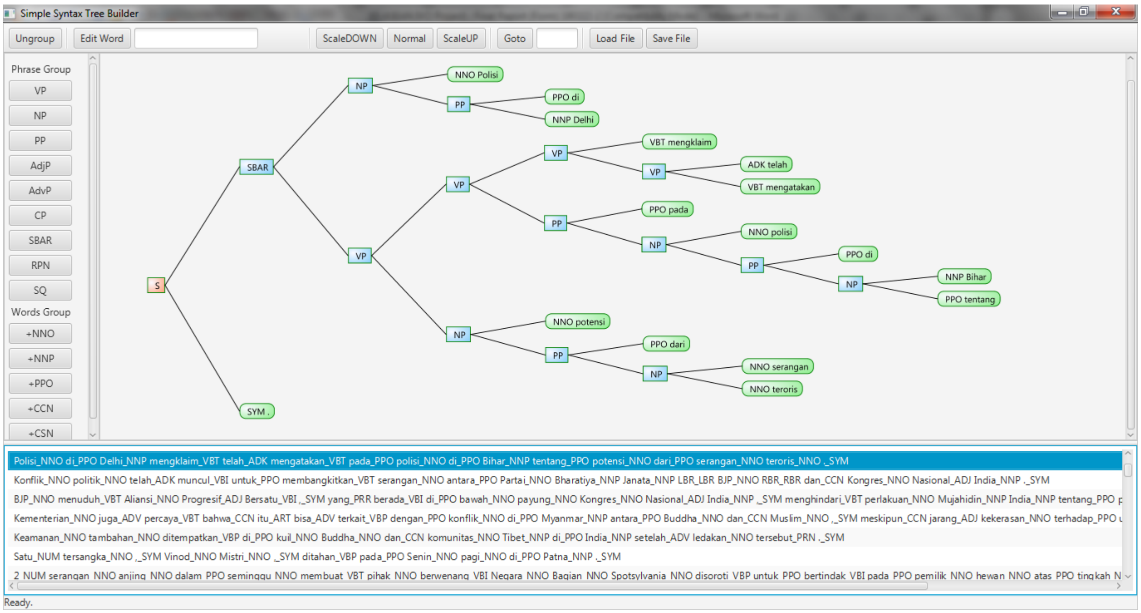
In syntactic analysis for Indonesian, we use a specific POS Tag for Indonesian language compiled by BPPT and INACL (Indonesian Association of Computational Linguistics). The morphological analysis and labeling were carried out using tools developed by BPPT. The Indonesian tree builder tool is called Simple Syntax Tree Builder. The actual works were carried out by INACL members consisting of several researchers from several research institutions and universities in Indonesia.

Word Alignment of English words with Indonesian is done first with a Giza ++ tool, then the results are manually checked.

POS tagging is done with morphological analysis tools. Examples of Indonesian language POS posting outputs are as follows:

```
Di_PPO bulan_NNO Oktober_NNO anggota_NNO dari_PPO Mujahidin_NNP
India_NNP mengklaim_VBT telah_ADK mengunjungi_VBT Ku_PRN il_NNO
Mahabodhi_NNO untuk_PPO merencanakan_VBT serangan_NNO
teroris_NNO ._SYM
```

Below is the snapshot of Simple Syntax Tree Builder.



Polisi_NNO di_PPO Delhi_NNP mengklaim_VBT telah_ADK mengatakan_VBT pada_PPO polisi_NNO di_PPO Bihar_NNP tentang_PPO potensi_NNO dari_PPO serangan_NNO teroris_NNO ._SYM

Konflik_NNO politik_NNO telah_ADK muncul_VBI untuk_PPO membangkitkan_VBT serangan_NNO antara_PPO Partai_NNO Bharatiya_NNP Janata_NNP LBR_LBR BJP_NNO RBR_RBR dan_CCN Kongres_NNO Nasional_ADJ India_NNP ._SYM

BJP_NNO menuduh_VBT Aliansi_NNO Progresif_ADJ Bersatu_VBI ,_SYM yang_PRR berada_VBI di_PPO bawah_NNO payung_NNO Kongres_NNO Nasional_ADJ India_NNP ,_SYM menghindari_VBT perlakuan_NNO Mujahidin_NNP India_NNP tentang_PPO p

Kementerian_NNO juga_ADV percaya_VBT bahwa_CCN itu_ART bisa_ADV terkait_VBP dengan_PPO konflik_NNO di_PPO Myanmar_NNP antara_PPO Buddha_NNO dan_CCN Muslim_NNO ,_SYM meskipun_CCN jarang_ADJ kekerasan_NNO terhadap_PPO u

Keamanan_NNO tambahan_NNO ditempatkan_VBP di_PPO kuil_NNO Buddha_NNO dan_CCN komunitas_NNO Tibet_NNP di_PPO India_NNP setelah_ADV ledakan_NNO tersebut_PRN ._SYM

Satu_NUM tersangka_NNO ,_SYM Vinod_NNO Mistri_NNO ,_SYM ditahan_VBP pada_PPO Senin_NNO pagi_NNO di_PPO Patna_NNP ._SYM

2_NUM serangan_NNO anjing_NNO dalam_PPO seminggu_NNO membuat_VBT pihak_NNO berwenang_VBI Negara_NNO Bagian_NNO Spotsylvania_NNO disoroti_VBP untuk_PPO bertindak_VBI pada_PPO pemilik_NNO hewan_NNO atas_PPO tingkah_N

Example of syntax trees generated from the Symple Syntax Tree Builder are as follows:

(S (SBAR (NP (NNO Polisi)(PP (PPO di)(NNP Delhi)))(VP (VP (VP (VBT mengklaim)(VP (ADK telah)(VBT mengatakan)))(PP (PPO pada)(NP (NNO polisi)(PP (PPO di)(NP (NNP Bihar)(PPO tentang))))))(NP (NNO potensi)(PP (PPO dari)(NP (NNO serangan)(NNO teroris))))))(SYM .))

- I2R

Malay ALT was annotated with word alignment and POS information. Annotation of the dependency relations was partially completed.

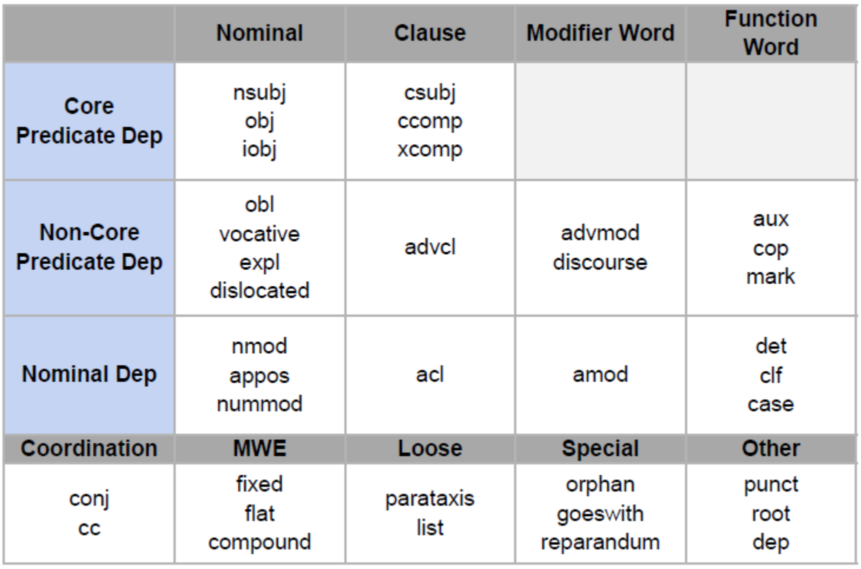
Automatic alignment was performed by GIZA++, assuming aligned-pairs should be bi-directional aligned. To achieve this, we combined ALT 20,106 sentence pairs with I2R 214,463 sentence pairs to get a bigger corpus and hence more accurate alignment estimates. Simple heuristics were then applied to align the left over words. All aligned pairs were manually checked by a Malay linguist.

We proposed 42 Malay POS categories. POS tag for each word was again automatically tagged by I2R POS tagger and manually checked by the linguist. The table below describes the POS category for Malay.

POS Tag	Description
NN	Common nouns
NR	Proper nouns
RPER	Personal pronoun. Some are originally affixed (nya/ku/mu)
RDEM	Demonstrative pronoun
RINT	Interoggative pronoun
RDT	Indefinite and quantifying head words
YANG	All YANG words

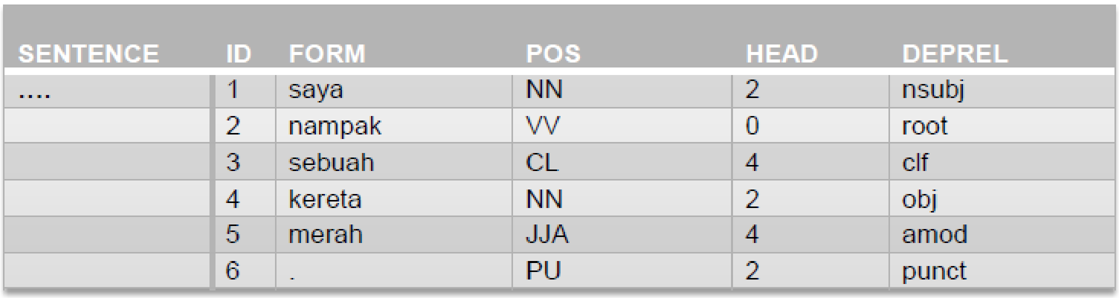
REFX	Reflexive pronoun/noun
RREL	Other relativiser (expected in informal)
DPER	Possesive relation (post-noun). Some are originally affixed (nya/ku/mu)
DDEM	Demonstrative relation
DINT	Interoggative relation (post-noun)
PDT	Quantifying, definite & indefinite determiners/articles (pre-noun)
VV	Main verbs (active)
VVP	Main verbs (passive)
VA	Adjs head of verb-less clause
JJA	Adjs post-modifying noun (same NP, typical ADJ)
JJV	Verbs modifying Nouns, (Yang-less) relative clause like English gerund
JJVP	Modifying verbs, passive
AUX	Auxiliary verbs/modals
CL	Classifiers
ADV	Adverbs (most are ambiguous with adjectives)
AINT	Interoggative adverbs
NEG	Negative words
CD	Cardinal numbers
OD	Ordinal numbers
P	Preposition
CNJ	Conjunctions
IJ	Interjection (some can be ADV, when fits inside clause, eg Arabic words). Includes simleys as well.
PAR	Particle
HYP	Tokenised hyphens joining 2 words/ phrases/numbers (excl. redup and affixes which are not tokenized).
SYM	Symbols (segmented ones only) – see list
PU	Punctuation (All mid-sentence and end-sentence punctuation; incl. non-redup hyphen. Smileys which are tokenized correctly, would get the IJ tag.
PX	Tags for ‘pure’ prefixes which have been split (by space) – Hyphenated affixes are not tokenized separately. These are for unseen ones.
SX	Tags for remaining suffixes.
FW*	Foreign word - in general tag using the regular tags. This is last resort. (need to determine length of sequence)
X	Used when a single word is split up with a space. The first would be assigned the whole word’s tag, while the second word-part would be assigned X
RED2	2nd reduplication element, if separated
PNP	Preposition + N/NP
VPO	Passive verb with object
VO	Active verb with object
JJVO	As VO but modifying

Malay Treebank was manually annotated with the dependency relations category summarized in the following table.



	Nominal	Clause	Modifier Word	Function Word
Core Predicate Dep	nsubj obj iobj	csubj ccomp xcomp		
Non-Core Predicate Dep	obl vocative expl dislocated	advcl	advmod discourse	aux cop mark
Nominal Dep	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	parataxis list	orphan goeswith reparandum	punct root dep

The annotation followed CoNLL format due to its ability to present a flat but meaningful tree. Each sentence has 6 columns described below. Numbers under the HEAD column indicate its dependents. The following shows an annotation for the sentence, "Saya nampak sebuah kereta merah."



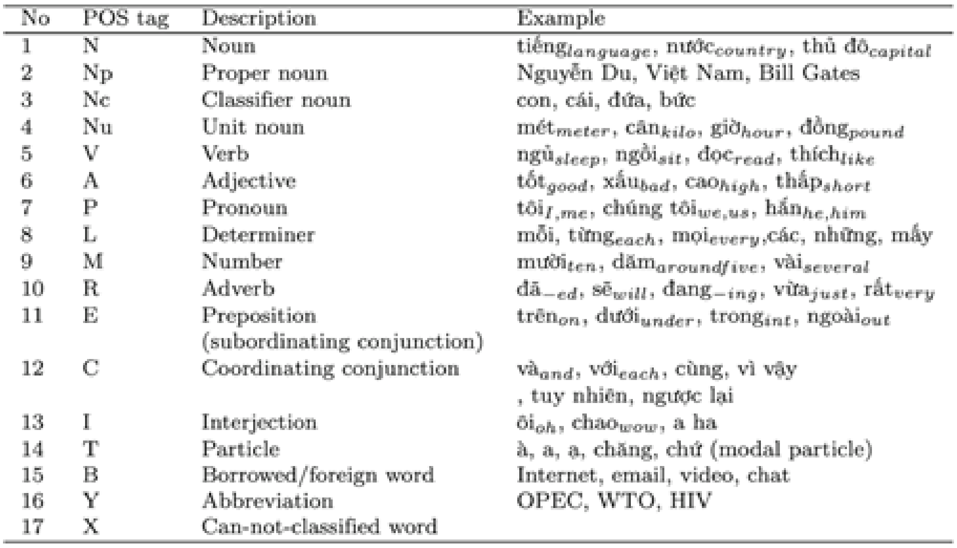
SENTENCE	ID	FORM	POS	HEAD	DEPREL
....	1	saya	NN	2	nsubj
	2	nampak	VV	0	root
	3	sebuah	CL	4	clf
	4	kereta	NN	2	obj
	5	merah	JJA	4	amod
	6	.	PU	2	punct

Where SENTENCE: Sentence number as it appears in data
ID: Word index, integer starting at 1 for each new sentence
FORM: Word form or punctuation symbol
POS: Part-of-speech tag
HEAD: Head of the current word
DEPREL: Universal dependency relation to the HEAD

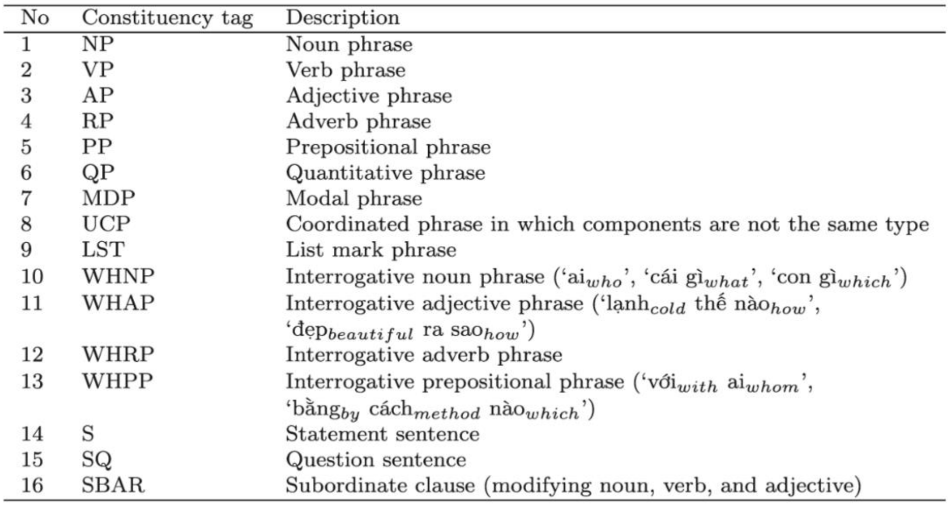
- IOIT / VNU UET
  There are a number of important characteristics of the Vietnamese language that impact greatly on the treebank construction. First, the smallest unit in the formation of Vietnamese words is the syllable. Words can have just one syllable or be a compound of two or more syllables. Like many other Asian languages such as Chinese, Japanese and Thai, there is no word delimiter in Vietnamese. The space is a syllable delimiter but not a word delimiter, so a Vietnamese sentence can often be segmented in many ways.

Second, Vietnamese is an isolating language in which words do not change their forms according to their grammatical function in a sentence. Third, the Vietnamese syntax conforms to the subject-verb-object (SVO) word order.

There are three levels of annotation including word segmentation, part-of-speech tagging, and syntactic analysis. Since Vietnamese has a relatively restrictive word order and often relies on the order of constituents to convey important grammatical information, we chose to use constituency representation of syntactic structures. POS, phrasal, and clausal tag sets are reported in the following figures. In order to deal with ambiguities occurring at various levels of annotation, we systematically applied linguistics analysis tests such as deletion, insertion, substitution, questioning, and transformation. Notions for these techniques were described in the guideline documents with examples, arguments and alternatives. These techniques originated in the literature or were proposed by members of our group.



No	POS tag	Description	Example
1	N	Noun	tiếng$_{language}$, nước$_{country}$, thủ đô$_{capital}$
2	Np	Proper noun	Nguyễn Du, Việt Nam, Bill Gates
3	Nc	Classifier noun	con, cái, đứa, bức
4	Nu	Unit noun	mét$_{meter}$, cân$_{kilo}$, giờ$_{hour}$, đồng$_{pound}$
5	V	Verb	ngủ$_{sleep}$, ngồi$_{sit}$, đọc$_{read}$, thích$_{like}$
6	A	Adjective	tốt$_{good}$, xấu$_{bad}$, cao$_{high}$, thấp$_{short}$
7	P	Pronoun	tôi$_{I,me}$, chúng tôi$_{we,us}$, hắn$_{he,him}$
8	L	Determiner	mỗi, từng$_{each}$, mọi$_{every}$,các, những, mấy
9	M	Number	mười$_{ten}$, dăm$_{aroundfive}$, vài$_{several}$
10	R	Adverb	đã$_{-ed}$, sẽ$_{will}$, đang$_{-ing}$, vừa$_{just}$, rất$_{very}$
11	E	Preposition (subordinating conjunction)	trên$_{on}$, dưới$_{under}$, trong$_{int}$, ngoài$_{out}$
12	C	Coordinating conjunction	và$_{and}$, với$_{each}$, cùng, vì vậy , tuy nhiên, ngược lại
13	I	Interjection	ôi$_{oh}$, chao$_{wow}$, a ha
14	T	Particle	à, a, ạ, chăng, chứ (modal particle)
15	B	Borrowed/foreign word	Internet, email, video, chat
16	Y	Abbreviation	OPEC, WTO, HIV
17	X	Can-not-classified word	

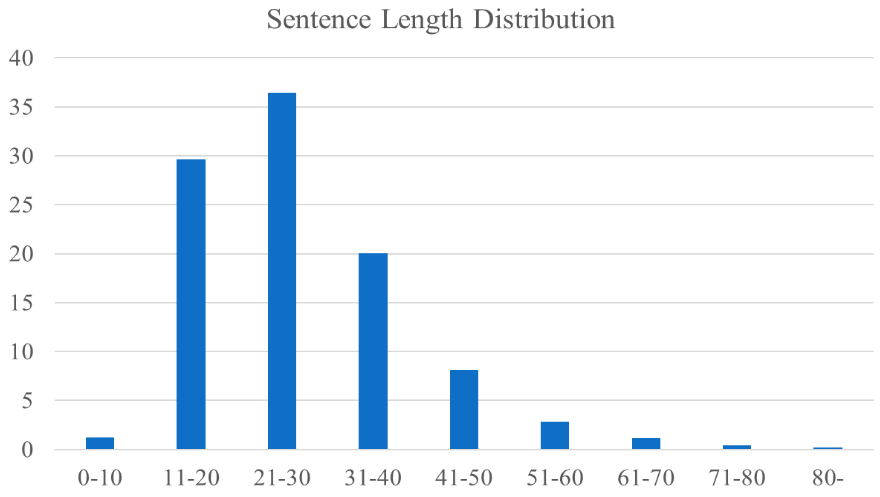


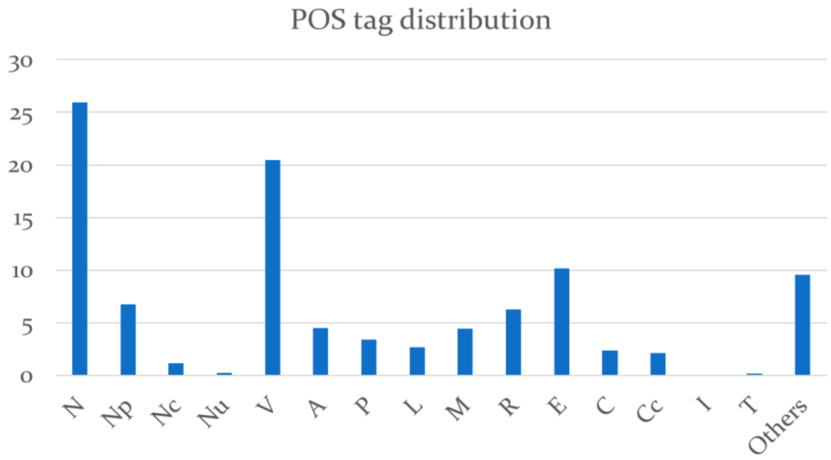
No	Constituency tag	Description
1	NP	Noun phrase
2	VP	Verb phrase
3	AP	Adjective phrase
4	RP	Adverb phrase
5	PP	Prepositional phrase
6	QP	Quantitative phrase
7	MDP	Modal phrase
8	UCP	Coordinated phrase in which components are not the same type
9	LST	List mark phrase
10	WHNP	Interrogative noun phrase (‘ai$_{who}$’, ‘cái gì$_{what}$’, ‘con gì$_{which}$’)
11	WHAP	Interrogative adjective phrase (‘lạnh$_{cold}$ thế nào$_{how}$’, ‘đẹp$_{beautiful}$ ra sao$_{how}$’)
12	WHRP	Interrogative adverb phrase
13	WHPP	Interrogative prepositional phrase (‘với$_{with}$ ai$_{whom}$’, ‘bằng$_{by}$ cách$_{method}$ nào$_{which}$’)
14	S	Statement sentence
15	SQ	Question sentence
16	SBAR	Subordinate clause (modifying noun, verb, and adjective)

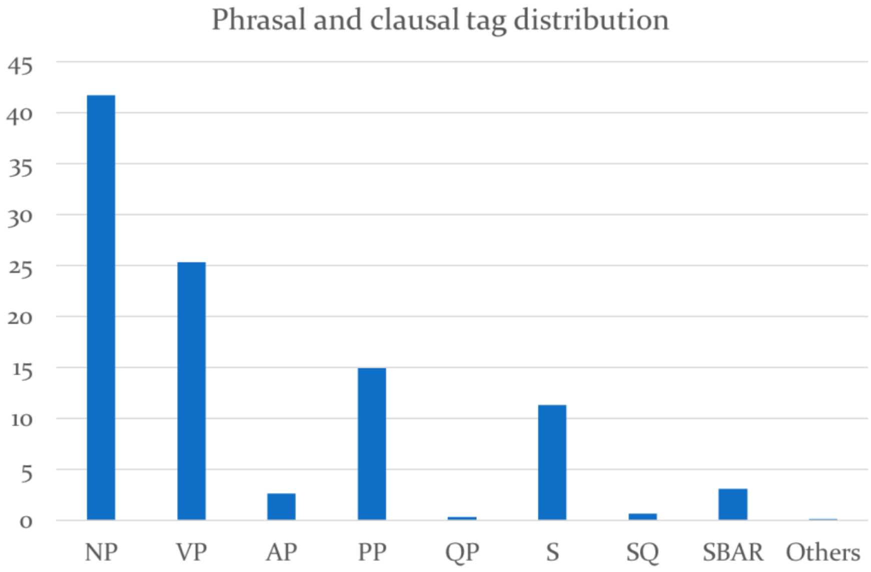
The following table shows the number of sentences at each annotation level.

No	Annotation level	Sentences
	Word segmentation	10,000
	POS tagging	7,000
	Syntax annotation	4,000

A number of data statistics are presented in the following figures.







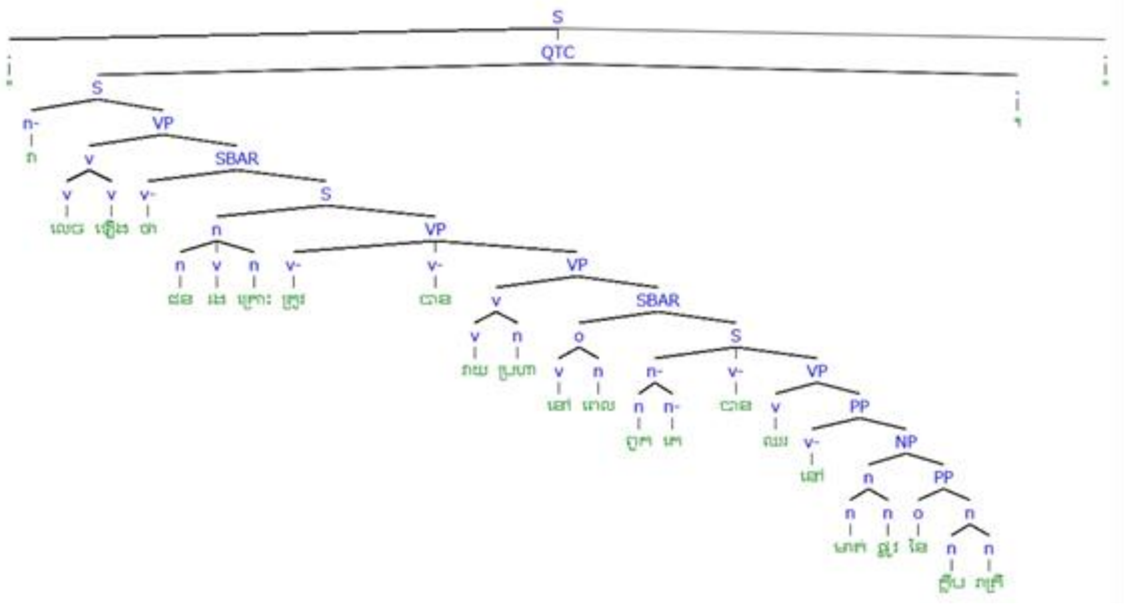
- NIPTICT

Under the cooperation of NIPTICT and NICT, the morphological annotation, including tokenization and POS-tagging, on the entire Khmer ALT data has been done. The annotation applies the nova annotation system. Finally, we have a morphologically annotated Khmer dataset composed of around 20,000 sentences with 700,000 tokens. Both the date and the annotation guidelines have been released on the homepage of the ALT project. The distribution of different POS tags is listed in the following table.

Part-of-Speech	Percentage
Noun	39.1
Verb	19.7
Adverb	13.3
Preposition	10.5

Punctuation mark	5.2
Adjective	4.1
Number	2.6
Pronoun	2.4
Particle	1.9
Determiner	1.2

Based on the morphological annotation, the syntactic annotation on the entire Khmer ALT data has been preliminarily done by mapping the syntactic tree of English sentences onto the corresponding Khmer sentences.
The following example illustrates an annotated Khmer sentence.

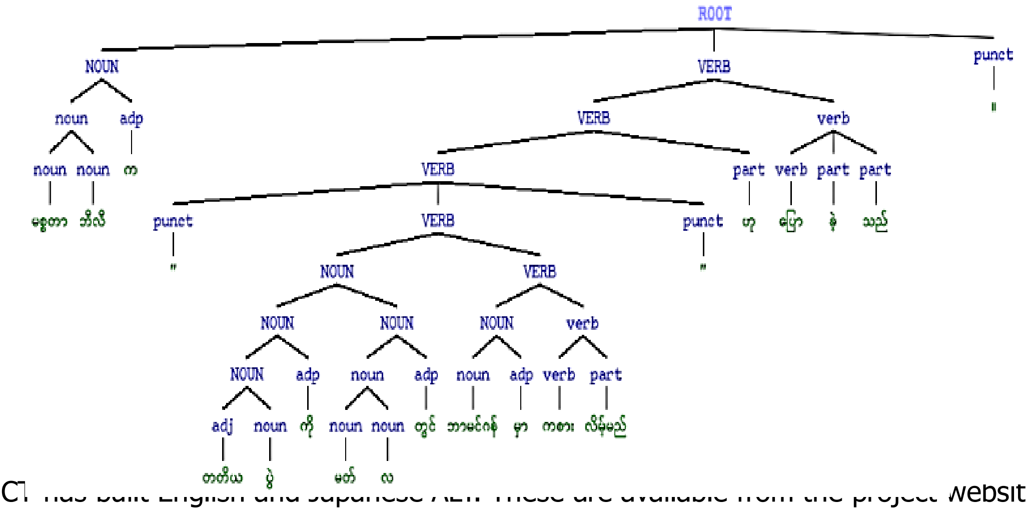


- UCSY

UCSY has built Myanmar and English ALT. These are available from the project website. There are 20K sentences and used for many Myanmar NLP researches such as Word Segmentation, POS Tagging, Name Entity Recognition, Machine Translation, Text Summarization, Spelling Checking and Text Classification.
Myanmar ALT was also recommended by Myanmar Language Commission stated that it is very useful for Myanmar NLP researches.

In construction of the Myanmar ALT, word alignment between Myanmar and English are done manually. Then word segmentation and spell checking are rechecked and corrected. The word alignment and syntactic tree are also refined and cleaned in many ways manually/automatically. Myanmar POS tagging firstly used NOVA tags, but later it was transformed to the universal POS tagging.

The example Myanmar ALT tree for one sentence is as follows:

- 

NICT has built English and Japanese ALT. These are available from the project website. The developed data has already been used for improving machine translation technology.

English ALT:
English texts were first tokenized by using Stanford Tokenizer [7]. Examples are:

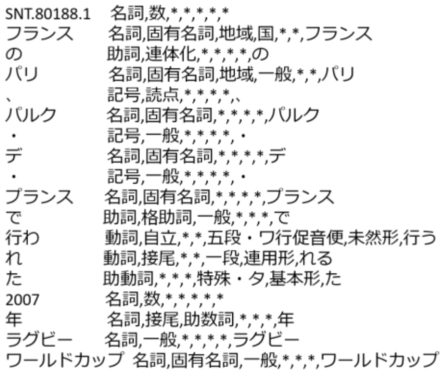
Italy have defeated Portugal 31-5 in Pool C of the 2007 Rugby World Cup at Parc des Princes , Paris , France .
Andrea Masi opened the scoring in the fourth minute with a try for Italy .
Portugal never gave up and David Penalva scored a try in the 33rd minute , providing their only points of the match .

Then, Penn treebank style POS tagging and syntax analysis were conducted. Examples are:

(S (S (BASENP (NNP Italy)) (VP (VBP have) (VP (VP (VP (VBN defeated) (BASENP (NNP Portugal))) (ADVP (RB 31-5))) (PP (IN in) (NP (BASENP (NNP Pool) (NNP C)) (PP (IN of) (NP (BASENP (DT the) (NN 2007) (NNP Rugby) (NNP World) (NNP Cup)) (PP (IN at) (NP (BASENP (NNP Parc) (FW des) (NNP Princes)) (COMMA ,) (BASENP (NNP Paris) (COMMA ,) (NNP France)))))))))) (PERIOD .))

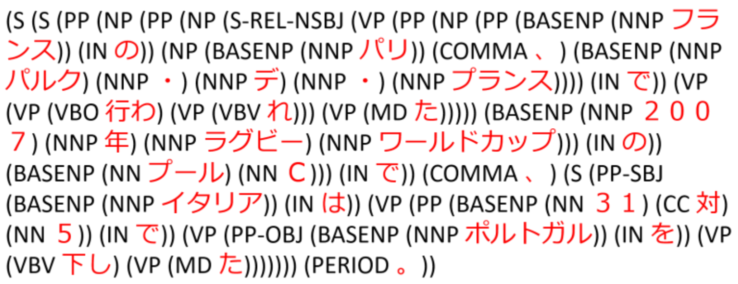
Those were first automatically obtained [8] then corrected manually.

Japanese ALT:
Mecab IPADIC [9] was used for Word segmentation and POS tagging. Examples are:



```
SNT.80188.1   名詞,数,*,*,*,*,*
フランス      名詞,固有名詞,地域,国,*,*,フランス
の            助詞,連体化,*,*,*,*,の
パリ          名詞,固有名詞,地域,一般,*,*,パリ
、            記号,読点,*,*,*,*,、
パルク        名詞,固有名詞,*,*,*,*,パルク
・            記号,一般,*,*,*,*,・
デ            名詞,固有名詞,*,*,*,*,デ
・            記号,一般,*,*,*,*,・
プランス      名詞,固有名詞,*,*,*,*,プランス
で            助詞,格助詞,一般,*,*,*,で
行わ          動詞,自立,*,*,五段・ワ行促音便,未然形,行う
れ            動詞,接尾,*,*,一段,連用形,れる
た            助動詞,*,*,*,特殊・タ,基本形,た
2007          名詞,数,*,*,*,*,*
年            名詞,接尾,助数詞,*,*,*,年
ラグビー      名詞,一般,*,*,*,*,ラグビー
ワールドカップ 名詞,固有名詞,一般,*,*,*,ワールドカップ
```

Word alignment between Japanese and English was done by hand completely, because the structures between Japanese and English are completely different.



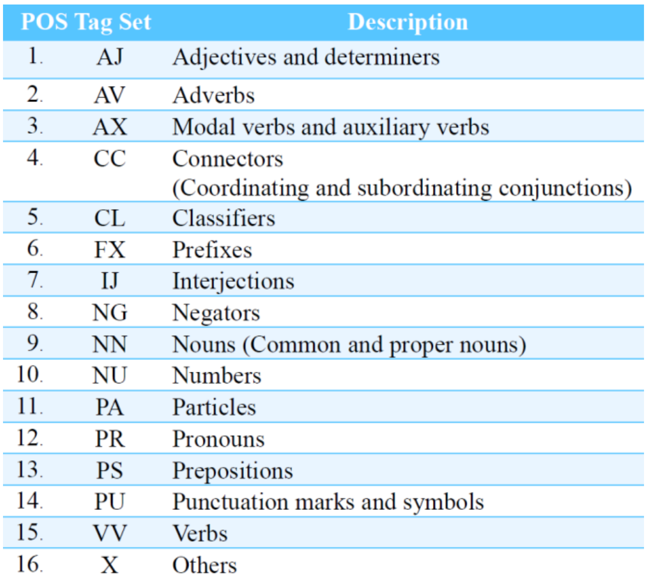
```
(S (S (PP (NP (PP (NP (S-REL-NSBJ (VP (PP (NP (PP (BASENP (NNP フラ
ンス)) (IN の)) (NP (BASENP (NNP パリ)) (COMMA 、 ) (BASENP (NNP
パルク) (NNP ・) (NNP デ) (NNP ・) (NNP プランス)))) (IN で)) (VP
(VP (VBO 行わ) (VP (VBV れ))) (VP (MD た))))) (BASENP (NNP ２００
７) (NNP 年) (NNP ラグビー) (NNP ワールドカップ))) (IN の))
(BASENP (NN プール) (NN Ｃ))) (IN で)) (COMMA 、 ) (S (PP-SBJ
(BASENP (NNP イタリア)) (IN は)) (VP (PP (BASENP (NN ３１) (CC 対)
(NN ５)) (IN で)) (VP (PP-OBJ (BASENP (NNP ポルトガル)) (IN を)) (VP
(VBV 下し) (VP (MD た))))))) (PERIOD 。 ))
```

Penn treebank style syntax analysis was also conducted as English ALT. Examples are:

In syntax analysis for Japanese, we added Japanese specific POS tags. The detailed annotation guidelines for English and Japanese ALTs are available at the project website.

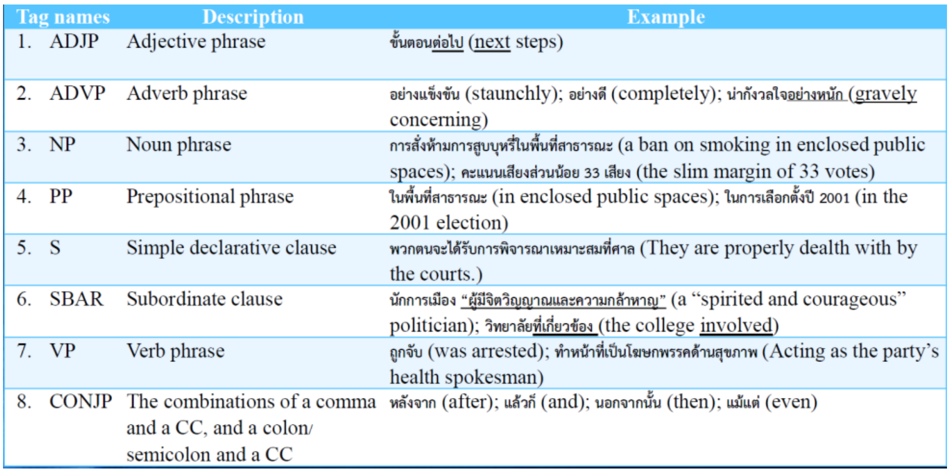
- NECTEC

NECTEC has built Thai ALT. NECTEC segmented Thai sentences into words then tagged POS tags. The list of POS tags is shown below.



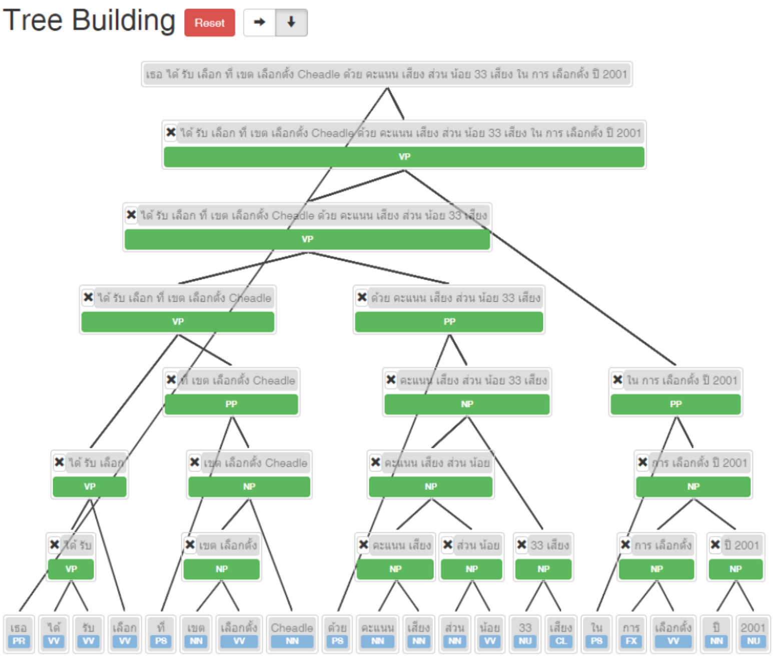
POS Tag Set		Description
1.	AJ	Adjectives and determiners
2.	AV	Adverbs
3.	AX	Modal verbs and auxiliary verbs
4.	CC	Connectors (Coordinating and subordinating conjunctions)
5.	CL	Classifiers
6.	FX	Prefixes
7.	IJ	Interjections
8.	NG	Negators
9.	NN	Nouns (Common and proper nouns)
10.	NU	Numbers
11.	PA	Particles
12.	PR	Pronouns
13.	PS	Prepositions
14.	PU	Punctuation marks and symbols
15.	VV	Verbs
16.	X	Others

Then, syntax analysis was conducted using the syntax tags below.



Tag names		Description	Example
1.	ADJP	Adjective phrase	ขั้นตอนต่อไป (next steps)
2.	ADVP	Adverb phrase	อย่างแข็งขัน (staunchly); อย่างดี (completely); น่ากังวลใจอย่างหนัก (gravely concerning)
3.	NP	Noun phrase	การสั่งห้ามการสูบบุหรี่ในพื้นที่สาธารณะ (a ban on smoking in enclosed public spaces); คะแนนเสียงส่วนน้อย 33 เสียง (the slim margin of 33 votes)
4.	PP	Prepositional phrase	ในพื้นที่สาธารณะ (in enclosed public spaces); ในการเลือกตั้งปี 2001 (in the 2001 election)
5.	S	Simple declarative clause	พวกตนจะได้รับการพิจารณาเหมาะสมที่ศาล (They are properly dealth with by the courts.)
6.	SBAR	Subordinate clause	นักการเมือง "ผู้มีจิตวิญญาณและความกล้าหาญ" (a "spirited and courageous" politician); วิทยาลัยที่เกี่ยวข้อง (the college involved)
7.	VP	Verb phrase	ถูกจับ (was arrested); ทำหน้าที่เป็นโฆษกพรรคด้านสุขภาพ (Acting as the party's health spokesman)
8.	CONJP	The combinations of a comma and a CC, and a colon/ semicolon and a CC	หลังจาก (after); แล้วก็ (and); นอกจากนั้น (then); แม้แต่ (even)

Here is an example of a 20-word sample.


Tree Building
Reset
→
↓
เธอ ได้ รับ เลือก ที่ เขต เลือกตั้ง Cheadle ด้วย คะแนน เสียง ส่วน น้อย 33 เสียง ใน การ เลือกตั้ง ปี 2001
ได้ รับ เลือก ที่ เขต เลือกตั้ง Cheadle ด้วย คะแนน เสียง ส่วน น้อย 33 เสียง ใน การ เลือกตั้ง ปี 2001
VP
ได้ รับ เลือก ที่ เขต เลือกตั้ง Cheadle ด้วย คะแนน เสียง ส่วน น้อย 33 เสียง
VP
ได้ รับ เลือก ที่ เขต เลือกตั้ง Cheadle
VP
ด้วย คะแนน เสียง ส่วน น้อย 33 เสียง
PP
ที่ เขต เลือกตั้ง Cheadle
PP
คะแนน เสียง ส่วน น้อย 33 เสียง
NP
ใน การ เลือกตั้ง ปี 2001
PP
ได้ รับ เลือก
VP
เขต เลือกตั้ง Cheadle
NP
คะแนน เสียง ส่วน น้อย
NP
การ เลือกตั้ง ปี 2001
NP
ได้ รับ
VP
เขต เลือกตั้ง
NP
คะแนน เสียง
NP
ส่วน น้อย
NP
33 เสียง
NP
การ เลือกตั้ง
NP
ปี 2001
NP
เธอ PR
ได้ VV
รับ VV
เลือก VV
ที่ PS
เขต NN
เลือกตั้ง VV
Cheadle NN
ด้วย PS
คะแนน NN
เสียง NN
ส่วน NN
น้อย VV
33 NU
เสียง CL
ใน PS
การ FX
เลือกตั้ง VV
ปี NN
2001 NU

- PUP Filipino-ALT

The Filipino treebank was based from Penn tree bank tagging the group created a tool that will automatically generate a treebank. A Filipino Part of speech tagging (FIL-POST) was created to tag the word's parts of speech. Dependency parsing was implemented n syntax analysis to generate the treebank. Examples are:

Ex.1

DP(D'(D(si)NP(N'(N(andrea))))),'(NP(N'(N(masi)))),'(NP(N'(N(ang)))),'(NP(N'(N(nagsimula)))),'(PP(P'(P(na)))),'(NP(N'(N(makapuntos)))),'(PP(P'(P(sa)))),'(NP(N'(N(italya)))),'(PP(P'(P(sa)))),'(AP(A'(A(ikaapat)))),'(PP(P'(P(na)))),'(NP(N'(N(minuto)DP(D'(D(ng)NP(N'(N(laro)))))))),'(PERIOD(.)))

Ex.2

VP(V'(V(binalaan)DP(D'(D(ng)NP(N'(N(pulisya))))))),'(NP(N'(N(ang)))),'(DP(Q(mga)D'())),'(AP(A'(A(sibilyan)))),'(PP(P'(P(na)))),'(NP(N'(N(ang)))),'(NP(N'(N(suspek)DP(D'(D(ay)NP(N'(N(maaring)))))))),'(NP(N'(N(may)))),'(NP(N'(N(gas)))),'(NP(N'(N(rebolber)))),'(PP(P'(P(at)))),'(NP(N'(N(isang)))),'(NP(N'(N(semi)AP(A'(A(awtomatik)))))),'(PP(P'(P(na)))),'(NP(N'(N(mahabang)))),'(NP(N'(N(baril)))),'(PERIOD(.)))

Ex.3

NP(N'(N(ang)AP(A'(A(pampook))))),'(PP(P'(P(na)))),'(NP(N'(N(pulis)))),'(PP(P'(P(na)))),'(NP(N'(N(tagapagsalita)))),'(PP(P'(P(na)DP(D'(D(si)NP(N'(N(alexei)))))))),'(NP(N'(N(pomorov)DP(D'(D(ay)NP(N'(N(nagsabi)))))))),'(PP(P'(P(sa)))),'(NP(N'(N(afp)DP(D'()))),'(VP(ADV(hindi)V'())),'(NP(N'(N(namin)))),'(NP(N'(N(alam)))),'(C(kung)),'(PP(P'(P(ito)DP(D'(D(ay)NP(N'(N(isang)DP(D'(ADV(hindi)))))))))),'(NP(N'(N(pagkakasundo)))),'(C(dahil)),'(NP(N'(N(wala)))),'(NP(N'(N(kami)))),'(NP(N'(N(mapagtanungan)))),'(DP(Q(lahat)D'(D(ay)NP(AP(A'(A(patay)N'()))))))),'(PP(P'(P(na)))),'(PERIOD(.)))

(4) Broader Impact

As stated in the introduction, NLP is one of the core technologies in ICT and NLP is based on treebanks. Consequently, ALT is very important in the development of fundamental NLP tools, such as word segmenters, POS taggers, and syntax parsers [1][2][3][4].

In addition, because ALT is made from translated English Wikinews, it is now available from our project website. In contrast, usual treebanks are very hard to share, because the original texts in usual treebanks have strict copyrights, which do not let researchers share treebanks.

ALT has already been used by other researchers. For example, "The 5th Workshop on Asian Translation" (WAT 2018) used ALT project data in their translation task.

We elaborate how ALT is useful for each language below from the point of each institute.

- BPPT

BPPT will use the output and data of the project for improving the quality of English – Indonesian Machine Translation System. Currently the system is developed using phrase based stochastic method. INACL (Indonesian Association of Computational Linguistics) also will use ALT project data for research in the field of computational linguistics.

- I2R

ALT data is very useful for under-resourced NLP research & development as many languages under ALT do not have much open source data. The data helps to promote the research work in these languages.

- IOIT / VNU UET

The ALT data can be used for the Vietnamese Language and Speech Processing (VLSP) community. Currently, there are significant needs for this kind of data, not only for developing applications such as machine translation, question answering, etc. but also for teaching and studying AI, NLP and related subjects. The VLSP community is growing and having annual activities such as organizing workshops and evaluation campaigns, sharing corpora, etc.

- NIPTICT

Khmer is under resource language. This project provided potential and positive outcomes to the research and development of Khmer NLP. The ALT data will be used by NITPICT NLP research team to improve not only the Khmer MT but also other Khmer NLP and AI systems.

- UCSY

Since Myanmar language is a low resource language, building Myanmar ALT becomes very effective not only for Myanmar NLP researches, but also for improving academic part especially in AI field.

- NICT

NICT has already used the ALT project data for improving machine translation technology. In Japan, machine translation is very important because it is essential for improving communication all over the world.

- NECTEC

ALT Data is very useful for NECTEC. We plan to use the ALT data in Machine Translation project. We plan to do a hybrid SMT-NMT machine translation. This data will be useful for us to improve the accuracy of Machine translation.

- PUP Filipino-ALT

PUP-CCIS will be using the output of the project (project data) for improving the translation on Filipino language with other languages. At the moment, translation is one of the in-demand areas in the field of language processing the project data may be used in different researches and project not only in ASIA. There are companies who are developing NLP projects that needs to have this project's data and it would be convenient to have it available for use.

(5) Future Developments

The ALT project data is the only data that cover wide ASEAN languages. Consequently, NLP researchers working with ASEAN languages will use the ALT project data. The researchers can obtain the ALT project data from the project website.

ALT has already been used by other researchers. For example, "The 5th Workshop on Asian Translation" (WAT 2018) used ALT project data in their translation task. We expect more and more NLP researchers use ALT in the future.

iii) Social Contribution

The ALT project data has been used for NLP as evident from the list below.

1) Academic papers published

- Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Rapid Sun, Vichet Chea, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, Chenchen Ding. (2016) "Introduction of the Asian Language Treebank" Oriental COCOSDA.
- Chenchen Ding, Masao Utiyama, Eiichiro Sumita. (2016) Similar Southeast Asian Languages: Corpus-Based Case Study on Thai-Laotian and Malay-Indonesian. WAT.
- Gunarso Gunarso, Hammam Riza. (2016) An Overview of BPPT's Indonesian Language Resources. ALR12.

- Hsu Myat Mo , KhinThandar Nwet , Khin Mar Soe, "CRF-Based Named Entity Recognition for Myanmar Language", The Proceedings of the International Conference on Genetic and Evolutionary Computing (ICGEC2016), pages 204–211, Fuzhou, China, November 7-9 2016
- KhinThandar Nwet , Khin Mar Soe, "Myanmar-English Machine Translation Model", The Proceedings of the International Conference on Genetic and Evolutionary Computing (ICGEC2016), pages 195-203, Fuzhou, China, November 7-9 2016
- Hnin Thu Zar Aye, Chenchen Ding, Win Pa Pa, Khin Thandar Nwet, Masao Utiyama, Eiichiro Sumita, "English-to-Myanmar Statistical Machine Translation Using a Language Model on Part-of-Speech in Decoding", The Proceedings of 15th International Conference on Computer Application(ICCA2017), pages 409-414. Yangon, Myanmar, 16-17 February
- Chenchen Ding, Vichet Chea, Masao Utiyama, Eiichiro Sumita, Sethserey Sam and Sopheap Seng.(2017) Statistical Khmer Name Romanization. PACLING. (Best Paper Award)
- Chenchen Ding, Win Pa Pa, Masao Utiyama and Eiichiro Sumita. (2017) Burmese (Myanmar) Name Romanization: A Sub-Syllabic Segmentation Scheme for Statistical Solutions. PACLING
- Chenchen Ding, Masao Utiyama and Eiichiro Sumita. (2018) Simplified Abugidas. ACL
- Agung Santosa, Asril Jarin, Made Gunawan, Teduh Uliniansyah, Gunarso, Elvira Nurfadhilah, Lyla Ruslana, Fara Ayuningtyas, Harnum Annisa, and Hammam Riza (2018) "Utilizing Indonesian Data Resources for Text-to-Speech Using End-to-End Method" O-Cocosda.
- Yi Mon Shwe Sin, Khin Mar Soe. (2018) Large Scale Myanmar to English Neural Machine Translation System. IEEE-GCCE.
- Minh-Thuan Nguyen, Van-Tan Buiy, Huy-Hien Vuz, Phuong-Thai Nguyen, Chi-Mai Luong. (2018) Enhancing the quality of Phrase-table in Statistical Machine Translation for Less-Common and Low-Resource Languages. IALP.
- Hsu Myat Mo, Khin Mar Soe. (2019) Syllable-Based Neural Named Entity Recognition For Myanmar Language" in the International Journal on Natural Language Computing(IJNLC).

2) Report for international standardization

NA

3) Patent

- ChenChen Ding of NICT submitted a Japanese domestic patent using the ALT project data.

4) Exhibition of the application or system the project developed

- ALT project webpage
http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html

- VoiceTra developed by NICT uses part of the ALT project data.
http://voicetra.nict.go.jp/en/

- TexTra developed by NICT uses part of the ALT project data
https://mt-auto-minhon-mlt.ucri.jgn-x.jp/

- ChenChen Ding of NICT has released an application using the ALT project data
http://www2.nict.go.jp/astrec-att/member/ding/my-akkhara.html

iv) References

[1] Hammam Riza. (2015) Bootstrapping Asian Language Treebank using Indonesian Language Resource. ASEAN IVO Forum 2015.
[2] Masao Utiyama, Eiichiro Sumita. (2015) Open collaboration for developing and using Asian Language Treebank (ALT). ASEAN IVO Forum 2015.
[3] Khin Mar Soe. (2015) Myanmar NLP research and Usefulness of ALT data. ASEAN IVO Forum 2015.
[4] Tat Thang Vu, Chi Mai Luong. (2015) Current status of Vietnamese Treebank development, usefulness of collaboration with Asian Language Treebank. ASEAN IVO Forum 2015.
[5] Vichet Chea, Ye Kyaw Thu, Chenchen Ding, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. (2015) Khmer Word Segmentation Using Conditional Random Fields. In Proc. of Conference on Khmer Natural language Processing.
[6] Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch and Eiichiro Sumita. "Introducing the Asian Language Treebank (ALT)," LREC, 2016.
[7] Stanford Tokenizer. https://nlp.stanford.edu/software/tokenizer.shtml
[8] Modifications to BerkeleyParser 1.7.
http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html
[9] Mecab IPADIC. https://github.com/taku910/mecab/tree/master/mecab-ipadic