Open Collaboration for Developing and Using Asian Language Treebank

Background :

NLP is one of the core technologies in ICT. This is because the contents of information are conveyed by natural languages. The state-of-the-art technologies in NLP are based on treebanks. A treebank is a linguistic knowledge representation of natural language texts. 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. In particular, there has been no publicly available POS-tagged and constituency tree corpora for most of Asian languages before our project.

Targets:

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.

Speaker:

ChenChen Ding, NICT on behalf of Masao Utiyama, NICT

Open Collaboration for Developing and Using Asian Language Treebank

Project Members :

Only team leaders are listed due to limited space.

Name	Position/Degree	Institution,Country
Hammam Riza	Deputy Chairman IT,Energy and Material / PhD	BPPT, Indonesia
Aw Ai Ti	Unit Head, Human Language Technologies / Ms.	I2R, Singapore
Luong Chi Mai	Assoc. Prof. / PhD	IOIT, Vietnam
Sethserey Sam	Vice President of Research / PhD	NIPTICT, Cambodia
Khin Mar Soe	Professor / PhD	UCSY, Myanmar
Masao Utiyama	Executive Researcher / PhD	NICT, Japan
Thepchai Supnithi	Research Team Leader/ Principal Researcher	NECTEC, Thailand
Ria A. Sagum	Assoc. Prof. / CCIS Faculty Researcher	PUP, Philippines

Project Duration :

36 months starting from April 1st, 2016

What is Asian Language Treebank

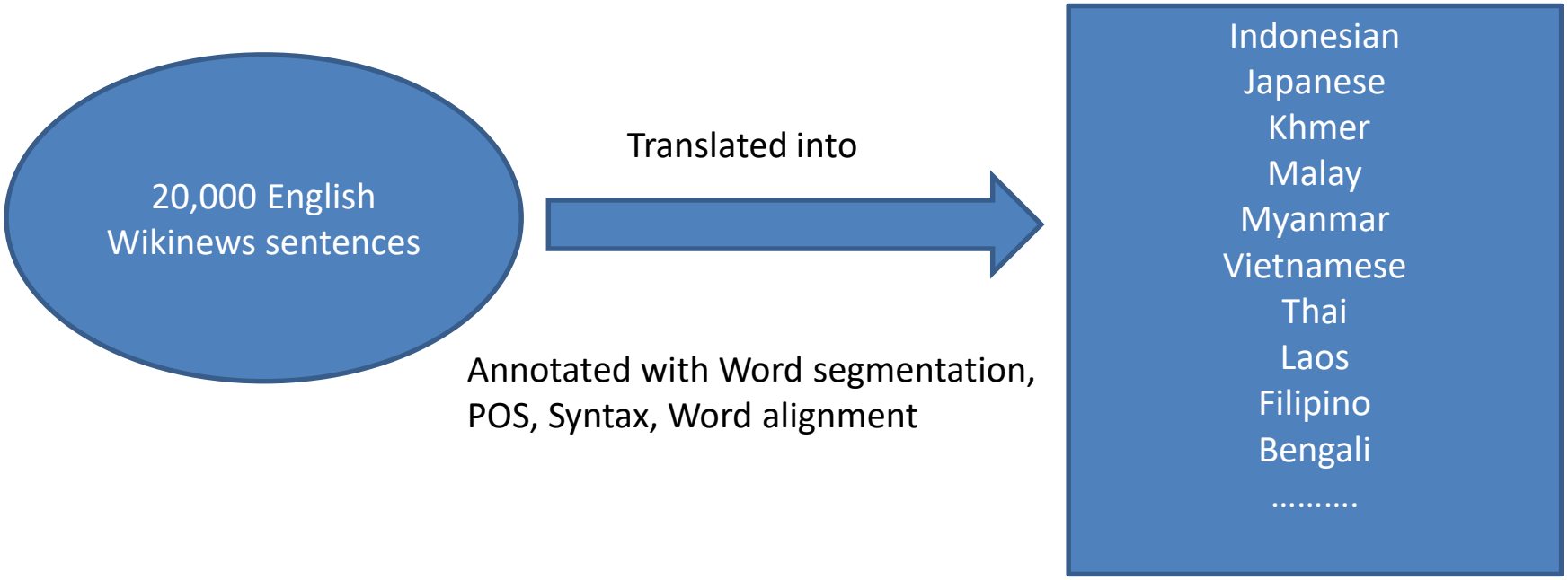
- Provide Asian Language Treebank for free for research

→Cover many under-resourced Asian languages

→Facilitate the rapid development of Asian NLP

→Provide the common ground for comparison/evaluation of Asian NLP

→We release ALT with a
Creative Commons Attribution-NonCommercial-ShareAlike



Samples (en, id, ja, km, ms, my, vi, th, fli)

- Italy have defeated Portugal 31-5 in Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France.
- Italia berhasil mengalahkan Portugal 31-5 di grup C dalam Piala Dunia Rugby 2007 di Parc des Princes, Paris, Perancis.
- フランスのパリ、パルク・デ・プランスで行われた2007年ラグビーワールドカップのプールCで、イタリアは31対5でポルトガルを下した。
- អ៊ីតាលីបានឈ្នះលើព័រទុយហ្គាល់ 31-5 ក្នុងប៉ូលCនៃពីធីប្រកួតពានរង្វាន់ពិភពលោកនៃកីឡាបាល់ឱបឆ្នាំ2007ដែលប្រព្រឹត្តនៅប៉ាសដេសប្រិន ក្រុងប៉ារីស បារាំង។
- Itali telah mengalahkan Portugal 31-5 dalam Pool C pada Piala Dunia Ragbi 2007 di Parc des Princes, Paris, Perancis.
- ပြင်သစ်နိုင်ငံ ပါရီမြို့ ပါဒ့ဂ်စ် ပရင့်စက် ၌ ၂၀၀၇ခုနှစ် ရပ်ဘီ ကမ္ဘာ့ ဖလား တွင် အီတလီ သည် ပေါ်တူဂီ ကို ၃၁-၅ ဂိုး ဖြင့် ရေကူးကန် စီ တွင် ရှုံးနိမ့်သွားပါသည် ။
- Ý đã đánh bại Bồ Đào Nha với tỉ số 31-5 ở Bảng C Giải vô địch Rugby thế giới 2007 tại Parc des Princes, Pari, Pháp.
- อิตาลีได้เอาชนะโปรตุเกสด้วยคะแนน31ต่อ5 ในกลุ่มC ของการแข่งขันรักบี้เวิลด์คัพปี2007 ที่สนามปาร์กเดแพร็งส์ ที่กรุงปารีส ประเทศฝรั่งเศส
- Natalo ng Italya ang Portugal sa puntos na 31-5 sa Grupong C noong 2007 sa Pandaigdigang laro ng Ragbi sa Parc des Princes, Paris, France.

Overview of R&D Results

- Each member institute have developed ALT for each languge
- Four Meetings hosted by NIPTICT, BBPT, UCSY, and NECTEC.
- ALT resources are available at the project page

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

- Corporation with U-STAR

  ALT parallel corpus has been used for U-STAR

  Khmer SMT has been released to U-STAR

- nine papers. One paper got a best paper award. Another paper was published at the most competitive international conference.
- Myanmar ALT was used in a machine translation workshop, improving MT over a well-known online machine translation service.
- Budgets: Most were for meeting. Others were for building ALT. Each member institute used their own budgets, too.

English/Japanese ALT by NICT

- 20,000 sentences were tokenized and parsed
- The parsing style mostly follows the Penn Treebank style
- Named Entity Tags based on OntoNotes also provided
- These are available to the public from the Project Web site

(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 .))

(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 2007) (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 。))

(S (S (BASENP (PDT All)(DT the)(NNS suspects))(VP (VBP are)(NP (BASENP (JJ male)(NE-NRP (NNPS Finns)))(VP (VBG residing)(PP (IN in)(BASENP (JJ southern)(NE-GPE (NNP Finland))))))))(PERIOD .))

Indonesian ALT Results by BPPT

2016

- Indonesian POS Tagger
- Simple Syntax Tree Builder
- Indonesian POS Tagset
- Indonesian Syntactic Tagset

2017

- Re-Translation EnWN to IdWN : 20 K sentences
- IdWN Tagging : 20 K Sentences
- Indonesian Syntax Tree : 20 K Sentences
- Word alignment between EnWN and IdWN using MGIZA automatically : 20 K Sentences

2018

- Re-Translation EnWN to IdWN : 10 K sentences
- Re-tagging sentences : 10 K Sentences
- Rebuild Tree : 10 K Sentences
- Re-Processing word alignment between EnWN and IdWN using MGIZA and Check manually : 10 k Sentences

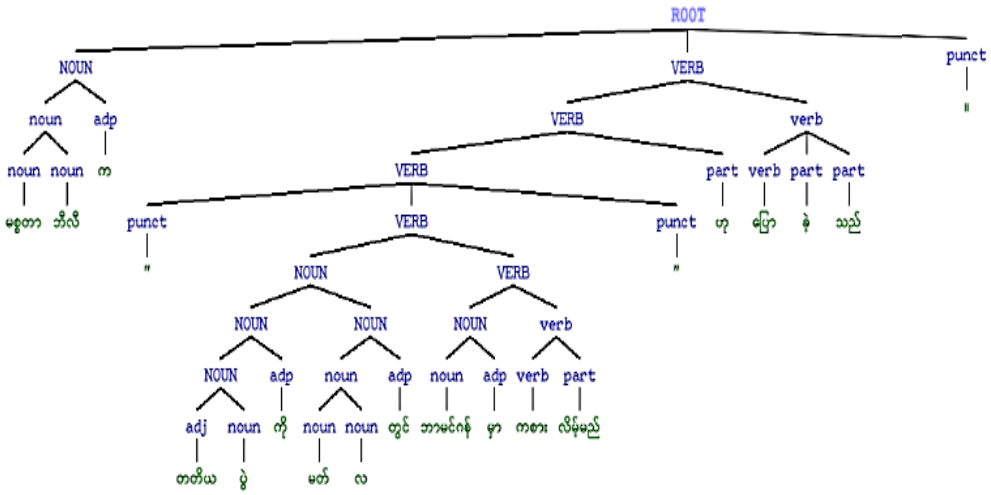
Malay ALT by I2R

1. Word segmentation and word alignment for 20k parallel sentences have been done automatically in August 2016.
2. The automatic aligned sentences are checked manually by linguists who know both languages.
3. The verified and aligned sentences are uploaded to ALT server in April 2017.
4. The 20k Malay sentences are automatically tagged and verified manually.
5. Propose the Malay dependency relations categories and annotation scheme.

	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

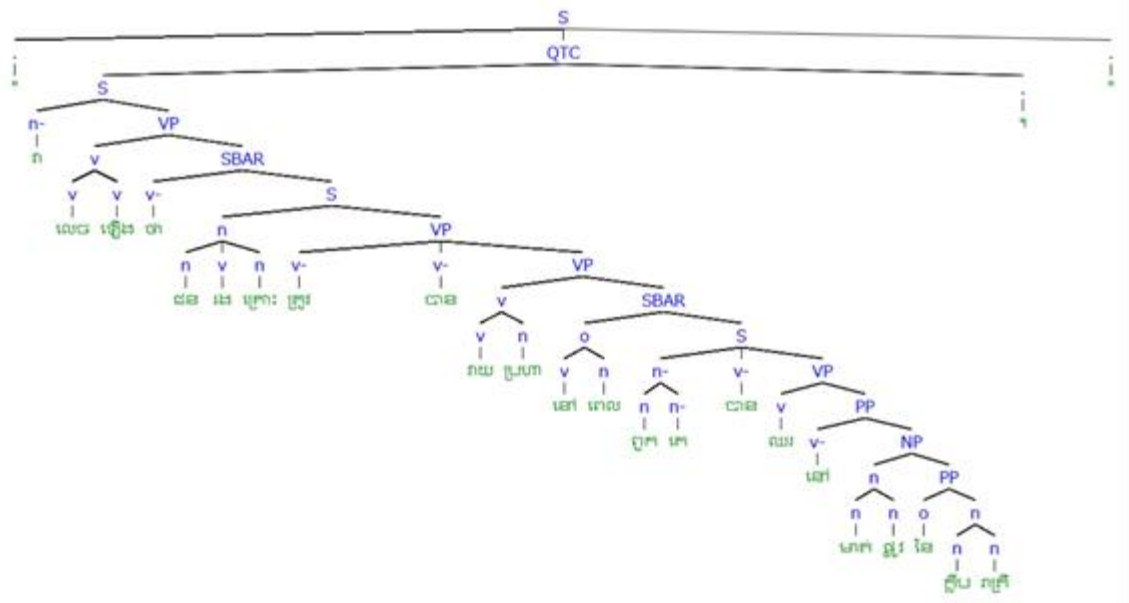
Myanmar ALT by UCSY with NICT

- Finished and released resources at the project Web site
  - Refined word alignment and syntactic tree
  - Improved tokenization and POS tagging
  - Get recommendation from Myanmar Language Commission



Khmer ALT by NIPTICT with NICT

- Annotation Guidelines for Khmer
  - Released on ALT home page
  - Updated along with the data preparing
  - A temporary stable version
- Final outcome of FY 2018
  - Word segmentation : 20,106 sentences
  - POS tagging : 20,106 sentences
  - Word alignment : 9,000 sentences



Vietnamese ALT by UET && IOIT

- Difficulties in Annotation
  - Isolating property
  - Training annotators
  - Long sentences
- Results
  - Word segmentation 10,000 sentences
  - POS tagging 7,000 sentences
  - Syntax annotation 4,000 sentences
  - Word alignment 3,000 sentences

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, a, 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)

Thai ALT by NECTEC & Filipino by PUP

Thai ALT syntax tags

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)

Filipino ALT syntax analysis

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(.)))

Scientific Contribution:

The presentations at International Conference:

No:	Paper title:	Author names	Affiliation	Conference name:	The date of the conference	The venue of the conference
	Statistical Khmer Name Romanization. PACLING	Chenchen Ding, Vichet Chea, Masao Utiyama, Eiichiro Sumita, Sethserey Sam and Sopheap Seng	National Institute of Information and Communications Technology, National Institute of Posts, Telecoms & ICT	International Conference of the Pacific Association for Computational Linguistics	16-18/08/2017	Yangon, Myanmar
	Burmese (Myanmar) Name Romanization: A Sub-Syllabic Segmentation Scheme for Statistical Solutions	Chenchen Ding, Win Pa Pa, Masao Utiyama and Eiichiro Sumita	National Institute of Information and Communications Technology, University of Computer Studies, Yangon	International Conference of the Pacific Association for Computational Linguistics	16-18/08/2017	Yangon, Myanmar

Scientific Contribution:

The presentations at International Conference:

No:	Paper title:	Author names	Affiliation	Conference name:	The date of the conference	The venue of the conference
	Simplified Abugidas	Chenchen Ding, Masao Utiyama and Eiichiro Sumita	National Institute of Information and Communications Technology	Annual Meeting of the Association for Computational Linguistics	15-20/07/2018	Melbourne, Australia
	MY-AKKHARA: A Romanization-based Burmese (Myanmar) Input Method	Chenchen Ding, Masao Utiyama and Eiichiro Sumita	National Institute of Information and Communications Technology	Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing	03-07/11/2019	Hong Kong, China

Scientific Contribution:

The presentations at International Conference:

No:	Paper title:	Author names	Affiliation	Conference name:	The date of the conference	The venue of the conference
	Overview of the 5th Workshop on Asian Translation	Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi	National Institute of Information and Communications Technology	Workshop on Asian Translation	03/12/2018	Hong Kong, China

Scientific Contribution:

The presentations at International Conference:

No:	Paper title:	Author names	Affiliation	Conference name:	The date of the conference	The venue of the conference
	Modifying NOVA-annotated Myanmar Data to Universal Part-of-Speech Tagset	Sann Su Su Yee, Chenchen Ding, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita	National Institute of Information and Communications Technology, University of Computer Studies, Yangon	International Conference on Computer Applications	27/02 – 01/03/2019	Yangon, Myanmar

Scientific Contribution:

The presentations at International Conference:

No:	Paper title:	Author names	Affiliation	Conference name:	The date of the conference	The venue of the conference
	Overview of the 6th Workshop on Asian Translation	Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi	National Institute of Information and Communications Technology	Workshop on Asian Translation	04/11/2018	Hong Kong, China

Scientific Contribution:

The Published Journal Papers:

No:	Paper title:	Author names	Affiliation	Journal name:	The publisher of the Journal	The volume number and Pages
	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	Transactions on Asian and Low-Resource Language Information Processing	Association for Computing Machinery	Vol. 18 Issue 2, Article No. 17
	Towards Burmese (Myanmar) Morphological Analysis: Syllable-based Tokenization and Part-of-Speech Tagging	Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita	National Institute of Information and Communications Technology	Transactions on Asian and Low-Resource Language Information Processing	Association for Computing Machinery	Vol. 19 Issue 1, Article No. 5

Societal Impact:

ALT is very important in the development of fundamental NLP tools, such as word segmenters, POS taggers, and syntax parsers.

ALT is now available from our project website, because ALT is made from translated English Wikinews,. In contrast, usual treebanks are very hard to share, because the original texts in usual treebanks have strict copyrights.

ALT has already been used by other researchers. For example, "The 5th and 6th Workshop on Asian Translation" (WAT 2018 and 2019) used ALT project data in their translation tasks. We elaborate how ALT is useful for each language below

- 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.

Societal Impact:

- 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.

Conclusion:

- Thank ASEAN IVO for the great opportunity to work together!
- The goal has been achieved.
- We have provided Asian Language Treebank for free for research
- Many research papers has been published
- ALT has been used for the development of machine translation technology

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

Future prospects:

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 and 6th Workshop on Asian Translation” (WAT 2018 and 2019) used ALT project data in their translation task. We expect more and more NLP researchers use ALT in the future.