Open Collaboration for Developing and Using Asian Language Treebank

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah (BPPT)
Aw Ai Ti, Sharifah Mahani Aljunied (I2R)
Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thái (IOIT/UET)
Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng (NIPTICT)
Khin Mar Soe, Khin Thandar Nwet (UCSY)
Masao Utiyama, Chenchen Ding (NICT)
Chai Wutiwiwatchai, Thepchai Supnithi, Pranchya Boonkwan (NECTEC) New!
Ria A. Sagum, Michael B. dela Fuente (PUP) New!

Current Status of Asian NLP resources

- No publicly available treebanks for most of Asian languages

→Development of Asian NLP is slow

→Difficult to compare research results among Asian NLP

Objective of 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 will release ALT with a

Creative Commons

Attribution-NonCommercial-ShareAlike

Benefits to ASEAN and beyond

- Intelligent ICT (IICT) needs NLP

→Web search, Speech-to-speech machine translation, IBM Watson, Text mining, Chat bots, and many more

- Without NLP, IICT does not work
- Without NPL resource, no NLP development
- ALT provides the core resource for Asian NLP
- ALT fosters the development of IICT in ASEAN and beyond

What will be the Asian Language Treebank (ALT)

20,000 English Wikinews sentences

Translated into

Indonesian
Japanese
Khmer
Malay
Myanmar
Vietnamese
Thai New!
Laos New!
Filipino New!

Annotated with Word segmentation, POS, Syntax, Word alignment

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

- 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 ក្នុងប៊ូលCនៃពិធីប្រកួតពានរង្វាន់ពិភពលោកនៃកីឡាបាល់ឱបឆ្នាំ2007 ដែលប្រព្រឹត្តនៅប៉ាសដេសប្រិន ក្រុងប៉ារីស ប៉ារាំង។
- Itali telah mengalahkan Portugal 31-5 dalam Pool C pada Piala Dunia Ragbi 2007 di Parc des Princes, Paris, Perancis.
- မြင်သစ်နိုင်ငံ ပါရီမြို့ ပါဒ့ကွ်စ် ပရင်စက် ဥ္ဒ ၂၀၀၇ခုနှစ် ရပ်ဘီ ကမ္ဘာ 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
- ອິຕາລີໄດ້ເສຍໃຫ້ປ໊ອກຕຸຍການ 31 ຕໍ່ 5 ໃນພູລ C ຂອງ ການແຂ່ງຂັນຣັກບີ້ລະດັບໂລກປີ 2007 ທີ່ ປາກເດແພຣ໌ງ ປາຣີ ປະເທດຝຣັ່ງ.
- 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.

Project Member Institutes (Languages)

- BPPT , Indonesia(Indonesian)
- I2R, Singapore (Malay)
- IOIT , UET, Vietnam (Vietnamese)
- NIPTICT, Cambodia (Khmer)
- UCSY, Myanmar (Myanmar)
- NICT , Japan (Japanese, English)

- Two new members below join to the project in FY 2017
- NECTEC, Thailand (Thai, Laos)
- PUP, Republic of the Philippines (Filipino)

Project Goal (Final outcome)

- NICT will develop and release the parallel corpus for ALT
- Each member institute shall develop and release ALT for each language
- Each member institute shall decide the amount of ALT, which will be developed and released by that institute
- ALT will be used for research and development on Asian NLP

Results of FY 2016 (Details are in Appendix)

- First meeting was hosted by NIPTICT
- Second meeting was hosted by BPPT
- Each member institute started developing each ALT
- **Four papers were published in international conferences**
- ALT resources are available at the project page

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

Indonesian ALT Project by BPPT

- Establish a cooperation with INACL (Indonesian Association for Computational Linguistic)
  - Develop Indonesian POS Tagset Standard
    - Based on MMTS (Multilingual Machine Translation System) Project
    - 13 main category and 26 sub category
  - Develop Indonesian Treebank Syntactic Tagset
    - Based on Penn Tree Bank
    - 9 syntactic tag
  - Compose the Guideline for Indonesian Treebank building
- Create tools
  - POS tagger based on new Indonesian POS tagset
  - Simple Syntax Tree Builder

Maly ALT Project by I2R

Major Work Done

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 using our defined tags. Manual verification of the tagged results completed.

English/Japanese ALT by NICT

- 20,000 sentences were tokenized and parsed
- The parsing style mostly follows the Penn Treebank style
- 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 。))

Khmer ALT by NIPTICT

- 20,106 sentences were automatically tokenized and tagged.
- 15,000 sentences were manually checked for tokenization and tagging.
- POS tagging follows the NOVA tagging system

អ៊ុតាលី/n បាន/o ឈ្នះ/v លើ/o ពំរទុយហ្គាល់/n 31-5/n ក្នុង/o ប៉ូល/n[n C/n]n នៃ/o ពិធី/n[n ប្រកួត/v]n ពាន/n[n រង្វាន់/n]n ពិភព/n[n លោក/n]n នៃ/o កីឡា/n[ក់ បាល់/n ឱប/v]n ឆ្នាំ/n[n 2007/1]n ដែល/n ប្រព្រឹត្ត/v នៅ/o ប៉ាស/n[n ឌេស/n ប្រ៊ីន/n]n ក្រុង/n[n ប៉ារីស/n]n បារាំង/n ។/.

Myanmar ALT by UCSY

- Works with intern in NICT
- 20,000 sentences were annotated
- Tokenized and POS-tagged by NOVA
- Released on ALT home page. **The largest Myanmar data in the world**

  - SNT.52859.60 The FBI has declined to release the details of the investigation.

    SNT.52859.60 အက်ဖ်ဘီအိုင် မှ စုံစမ်း စစ်ဆေး ခြင်း ၏ အသေးစိတ် အချက်အလက် များ သတင်းထုတ်ပြန် ခြင်း ကို ငြင်းဆို ခဲ့ သည် ။

    SNT.52859.60 အက်ဖ်ဘီအိုင် |n မှ |o- စုံစမ်း |v စစ်ဆေး |n[v ခြင်း |o-]n ၏ |o-အသေးစိတ် |o အချက်အလက် |n[n များ |o-]n သတင်း |n ထုတ်ပြန် |n[v ခြင်း|o-]n ကို |o- ငြင်းဆို |v[v ခဲ့ |o- သည် |o-]v ။ |.

Vietnamese ALT by IOIT & UET

Done for using Vietnamese tools to produce automatically:

- Word segmentation 9,000 sentences (manually revised)
- POS tagging 6,000 sentences (manually revised)
- Syntax Annotation 3,000 sentences (manually revised 1,000)

(Tag set summarization: POS tag set: 20 tags; syntax tag set: 18 phrasal and clausal tags, 9 functional tags, 8 adjunct tags, and 4 null element tags)

Progress and future plan

- 3rd meeting was held at Myanmar
- Steady development for the project goal
- Three papers have been published. One paper got the best paper award at PACLING 2017.

- **Corporation with U-STAR**
  - Khmer machine translation has been released to U-STAR
  - Parallel corpora of ALT has been released to U-STAR

Appendix

Indonesian ALT developed by BPPT

Indonesian ALT Project Preparation

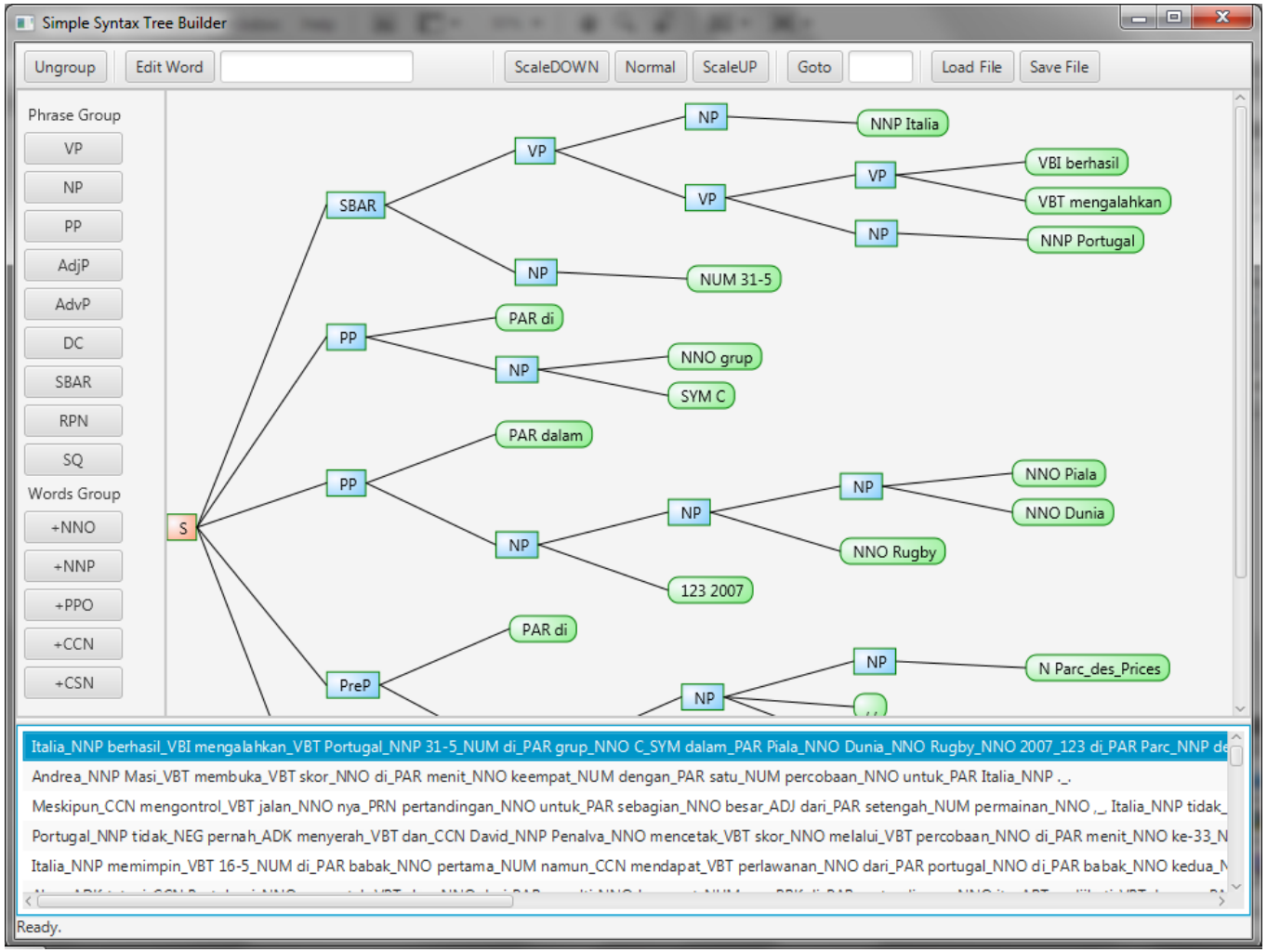
- Establish a cooperation with INACL (Indonesian Association for Computational Linguistic)
  - Develop Indonesian POS Tagset Standard
    - Based on MMTS (Multilingual Machine Translation System) Project
    - 13 main category and 26 sub category
  - Develop Indonesian Treebank Syntactic Tagset
    - Based on Penn Tree Bank
    - 9 syntactic tag
  - Compose the Guideline for Indonesian Treebank building
- Create tools
  - POS tagger based on new Indonesian POS tagset
  - Simple Syntax Tree Builder

Indonesian ALT Tree Building

- 20K sentences have been tagged automatically using new POS Tagger
- Distribute 20.000 sentences to INACL members
- INACL member universities that involved in the treebank development:
  - Indonesian University (Depok, West Java)
  - Diponegoro University (Semarang, Central Java)
  - UIN Sjarif Hidayatullah (Jakarta)
  - Bina Nusantara University (Jakarta)
  - Bandung Institute of Technology (Bandung, West Java)
  - Dian Nuswantoro University (Semarang, Central Java)
  - STT Nurul Fikri
  - Tanjungpura University (Pontianak, West Kalimantan)
  - Trisakti University (Jakarta)
  - UPN Veteran (Jakarta)
  - Telkom University (Bandung, West Java)

Indonesian Simple Syntax Tree Builder

- Java based
- Input from POS Tagger
- Function to revise the incorrect POS label
- Can handle idioms or compound words



Current Status and Future Plan for Indonesian ALT

- Current Status
  - Syntax tree of 10K sentences have been built.
  - Recheck by others for the correctness and will be completed by the end of 2017
  - Another 10K sentences will be completed by the end of 2018

(PRN Ini)(VP (VBL adalah)(NP (NP (NNO bagian)(PP (PPO dari)(NP (NNO rencana)(NP (NP (NUM Perdana)(NNO Menteri))(NP (NNP Tony_Blair)(CCN dan)(NNP Irish_Toaiseach_Bertie_Ahern)))))(PP (PPO untuk)(VP (VBT menyusun)(NP (NP (NNO devolusi)(NNO pemerintahan))(PP (PPO di)(NP (NNP Irlandia)(NNO Utara)))))))(AdvP (ADV setelah)(SBAR (NP (NNO parlemen)(PRK nya))(VP (VBP dibekukan)(PP (PPO pada)(NP (NP (NNO bulan)(NNO Oktober))(NUM 2002)))(PP (PPO tentang)(NP (NNO dugaan)(NP (NP (NUM lingkaran)(NNO mata-mata))(NNO IRA)))))))))(SYM .))

(PP (PPO Pada)(NP (NNO hari)(NNO Senin))(SYM ,))(SBAR (NP (NNP Ian_Paisley))(VP (VBT meminta)(NP (NP (NP (NNO perpanjangan)(NP (NNO batas)(NNO waktu)))(NP (NUM 24)(NNO November)))(NP (NNO selama)(NP (NUM 2)(NNO minggu)))))))(SYM .))

- Future Plan
  - Distribute the result to INACL members
  - Improve the English ←→Indonesian SMT System
  - Adding more sentences to the existing Indonesian ALT

Malay ALT developed by I2R

I2R Progress Report

Major Work Done

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 using our defined tags. Manual verification of the tagged results completed.

Malay Tagset

New 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

Work Plan FY2017

1. Study and propose the annotation scheme for Malay treebank

English/Japanese ALT developed by NICT

English ALT

- 20,000 sentences were tokenized and parsed
- The parsing style mostly follows the Penn Treebank style
- 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 .))

Japanese ALT

- 20,000 sentences were annotated
- Word segmentation and POS tagging following the IPA-DIC standard
- Parsing adapting the Penn Treebank style into Japanese
- Word alignment between Japanese and English
- These are available to the public from the Project Web site

(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 。))

Future Plan for English/Japanese ALT

- English/Japanese ALT has been completed
- English ALT is now annotated with Named Entity Tags
  - Person names
  - Company names
  - Date
  - ……
- Named Entities are very important for Knowledge Extraction
- Named Entities tags will be mapped to other languages

→ Every language will have named entity tags

Khmer ALT developed at NICT with NIPTICT

Annotation Guidelines for Khmer

- Released annotation guideline for Khmer on ALT home page:
  - http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/Khmer-annotation-guideline.pdf
- Updated along with the data preparing
- A temporary stable version

Khmer ALT data

- 20,106 sentences were automatically tokenized and tagged.
- 15,000 sentences were manually checked for tokenization and tagging.
- POS tagging follows the NOVA tagging system

អ៊ុតាលី/n បាន/o ឈ្នះ/v លើ/o ពំរទុយហ្គាល់/n 31-5/n ក្នុង/o ប៉ូល/n[n C/n]n នៃ/o ពិធី/n[n ប្រកួត/v]n ពាន/n[n រង្វាន់/n]n ពិភព/n[n លោក/n]n នៃ/o កីឡា/n[ក់ បាល់/n ឱប/v]n ឆ្នាំ/n[n 2007/1]n ដែល/n ប្រព្រឹត្ត/v នៅ/o ប៉ាស/n[n ឌេស/n ប្រីន/n]n ក្រុង/n[n ប៉ារីស/n]n បារាំង/n ។/.

Future Plan for Khmer ALT

- Complete the manual word tokenization and POS tagging.
- Process the word alignment between English and Khmer ALT data
- Write a guideline for Tree parsing

Myanmar ALT developed at NICT with UCSY

Myanmar ALT

- Works with intern in NICT
- 20,000 sentences were annotated
- Tokenized and POS-tagged by NOVA
- Released on ALT home page
  - http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/my-nova-170405.zip

Myanmar ALT : POS Tagging Example

- SNT.52859.60 The FBI has declined to release the details of the investigation.

SNT.52859.60 အက်ဖ်ဘီအိုင် မှ စုံစမ်း စစ်ဆေး ခြင်း ၏ အသေးစိတ် အချက်အလက် များ သတင်းထုတ်ပြန် ခြင်း ကို ငြင်းဆို ခဲ့ သည် ။

SNT.52859.60 အက်ဖ်ဘီအိုင် |n မှ |o- စုံစမ်း |v စစ်ဆေး |n[v ခြင်း |o-]n ၏ |o-အသေးစိတ် |o အချက်အလက် |n[n များ |o-]n သတင်း |n ထုတ်ပြန် |n[v ခြင်း|o-]n ကို |o- ငြင်းဆို |v[v ခဲ့ |o- သည် |o-]v ။ |.

Myanmar ALT : Annotation Guidelines

- Completed and Released on ALT home page
  - **http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/Myanmar-annotation-guideline.pdf**
  - **http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/Myanmar-annotation-guideline-supplemantary.pdf**

Current Work for Myanmar ALT

- Doing morphology analysis on Myanmar Language
- Next intern to NICT
  - for checking and enhancing in building syntactic tree

Future Plan for Myanmar ALT

- To refine word alignment
- To check and enhance syntactic trees
- To release the resources for research community

Report from Vietnam

Content

- Results obtained so far
- Final outcome, or the goal of this project in the end of FY 2018

Result obtained

1/ Done for using Vietnamese tools to produce automatically:

- Word segmentation 9,000 sentences (manually revised)
- POS tagging 6,000 sentences (manually revised)
- Syntax Annotation 3,000 sentences (manually revised 1,000)

(Tag set summarization: POS tag set: 20 tags; syntax tag set: 18 phrasal and clausal tags, 9 functional tags, 8 adjunct tags, and 4 null element tags)

Final outcome at the end of FY 2018

- Revise data that automatically produced
- Complete Word alignment with 3,000 sentences
- Publish one conference paper