

# Open Collaboration for Developing and Using Asian Language Treebank (ALT)

**BPPT** Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah

**I2R** Aw Ai Ti, Sharifah Mahani Aljunied

**IOIT** Luong Chi Mai, Vu Tat Thang, **UET** Nguyen Phuong Thái

**NIPTICT** Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng

**UCSY** Khin Mar Soe, Khin Thandar Nwet

**NICT** Masao Utiyama, Chenchen Ding

# ALT 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 NLP resource, no NLP development
- ALT provides the core resource for Asian NLP
- ALT fosters the development of IICT in ASEAN and beyond

# Background and Objective of ALT

- No publicly available treebanks for most of Asian languages  
→ Development of Asian NLP is slow
- ALT covers many under-resourced Asian languages  
→ Facilitate the rapid development of Asian NLP
- We will release ALT with a  
**Creative Commons  
Attribution-NonCommercial-ShareAlike License**

# What will be the Asian Language Treebank (ALT)

Annotated with Word segmentation,  
POS, Syntax, Word alignment

20,000  
English  
Wikinews  
sentences

Translated into



Indonesian(BPPT)  
Japanese(NICT)  
Khmer(NIPTICT)  
Malay(I2R)  
Myanmar(UCSY)  
Vietnamese(IOIT)

Thai  
Laos  
Filipino

.....

Every institute works equally for  
annotating the language



# Established results

- Parallel Corpus is available on 10 languages:

English <-> Indonesian, Japanese, Khmer, Malay, Myanmar, Vietnamese, Thai, Laos, Filipino

- Open source software

- Fast and Robust BerkeleyParser

- CRFSegmenter for Word segmentation and POS tagging

- Available at the project page:

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

# Minimal amount of data annotated within the next fiscal year for each language

- Word segmentation 9,000 sentences
- POS tagging 6,000 sentences
- Syntax Annotation 3,000 sentences
- Word alignment 1,500 sentences

Those are useful for evaluation for Asian NLP.

# Corporation with ALT and U-STAR

- U-STAR project uses ALT for the **evaluation** of their results
- NICT will provide **Khmer MT** to U-STAR, based on the development of **Khmer ALT** with NIPTICT