

Project Leader: Masao Utiyama, NICT. Project members: Hammam Riza (BPPT), Aw Ai Ti (I2R), Luong Chi Mai (IOIT), Sethserey Sam (NIPTICT), Khin Mar Soe (UCSY), Thepchai Supnithi (NECTEC), Ria A. Sagum (PUP) (Only project leaders in each institute are listed because of many members)

Budget: 87,000 USD, Duration: 36 months from April 1<sup>st</sup>, 2016

Target of this project: NICT built an annotation server for developing Asian Language Treebank (ALT) using English Wikinews. 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.

Findings and Outcomes: Each member institute has built their ALT. The results are available from the project website:

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

The outcomes are annotation guidelines for targeting languages and annotated corpora.

Collaborations: 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).

Broader Impact and Future Developments: ALT is very important in the development of fundamental NLP tools, such as word segmenters, POS taggers, and syntax parsers. Many NLP researchers will use ALT for developing NLP and ICT technologies in ASEAN.

Social Contribution:

- \* 13 academic papers have been published.
- \* 1 patent has been submitted.
- \* The ALT data have been used for developing machine translation.
- \* The ALT data have been used in a MT workshop