ALT (Asian Language Treebank) Project
- Indonesian Participation -

ASEAN IVO Forum
Hanoi, November 24, 2016

ALT Project

The ALT project aims to advance the state-of-the-art Asian natural language processing (NLP) techniques through the open collaboration for developing and using ALT.

The project is a joint effort of the six institutes, BPPT, I2R, IOIT, NIPTICT, UCSY, and NICT, for making a parallel treebank for seven languages: English, Indonesian, Japanese, Khmer, Malay, Myanmar, and Vietnamese.

Indonesian (BPPT) Roles

- NICT selected 20K sentences from English Wikinews (EnWN)
- NICT provides translation texts in Indonesian language (bahasa Indonesia)
    - Need some minor correction (BPPT will do the correction)
- Create word-aligned parallel translation corpus for EnWN ↔ bahasa Indonesia
- Create language treebank for Indonesian translations

Tools and Resources

- NICT
  - ALT server with following functions:
    - Translation
    - Alignment
    - Tagging
    - Tree Building

- BPPT
  - POS tagger
  - English – Indonesian parallel corpus: 250.000 sentences

ALT Server

- Data
  - The original English sentences and corresponding translations
  - Results of following tasks:
    - Translation
    - Alignment
    - Tagging
    - Tree building
- Tasks are done remotely
  - Two kinds of account: Manager and Translator

- BPPT is planning to develop tools that more suitable for processing bahasa Indonesia

Characteristics of bahasa Indonesia

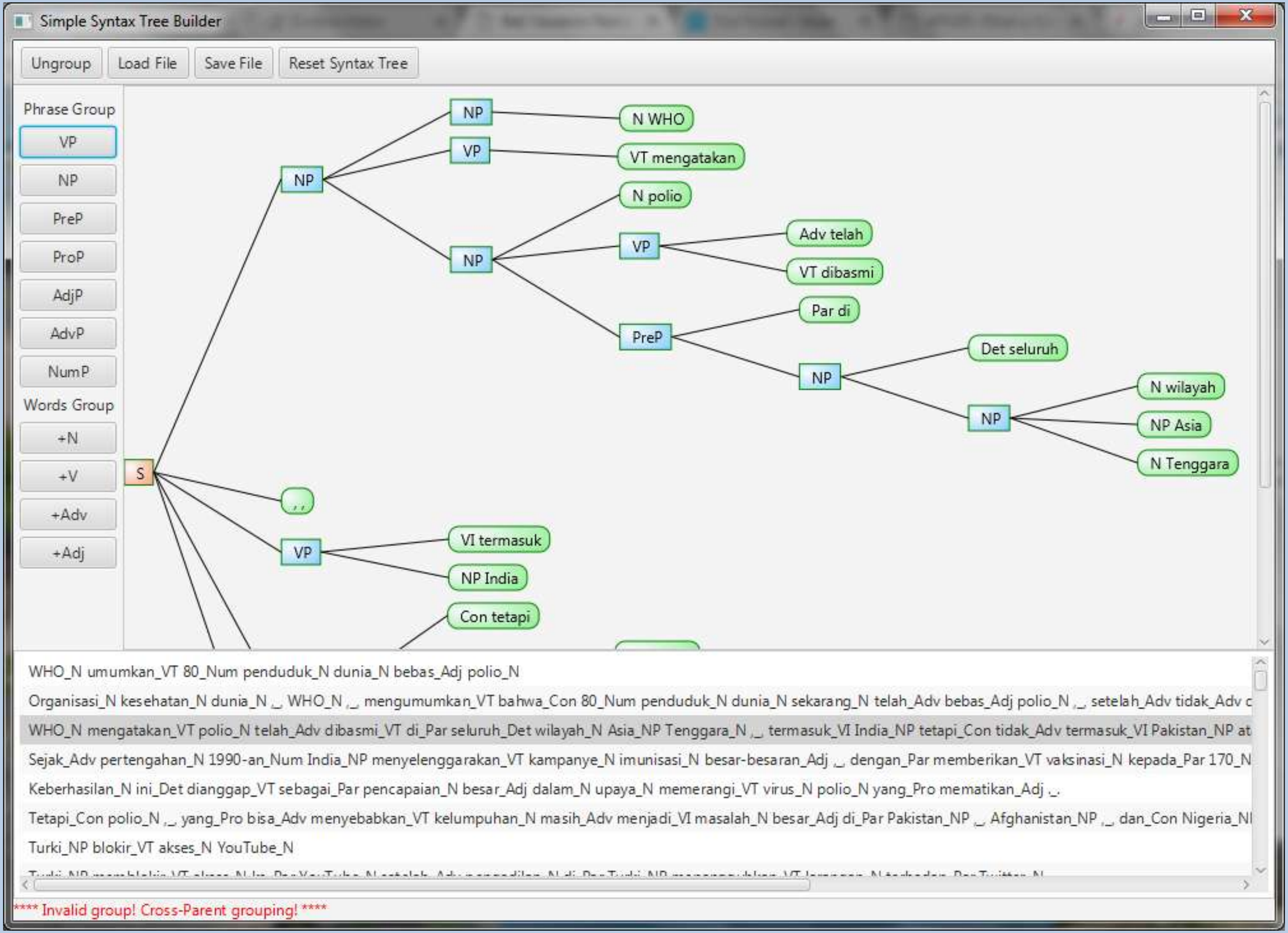
(in view of syntactical and semantical analysis)

- Agglutinative language; using complex affixes to create derivatives
- No tenses; using functional words to indicate time of events
- Using hyphen to create reduplications with many functions such as:
  - expressing plurality; exp.: *buku-buku* (books)
  - qualitative intensity; exp.: *bagus-bagus* (all is good)
  - quantitative intensity; exp.: *beratus-ratus* (hundreds of)
  - frequentative intensity; exp.: *pergi-pergi* (going out frequently),
  - etc.
- Unique abbreviations and idioms ← word and sentence alignment issue

Indonesian Tools

- Alignment tool
  - Utilize mgiza or Giza++
  - Trained using 250.000 English – Indonesian parallel corpus
- New POS tagger
  - Use Apache OpenNLP
  - 1000 tagged sentences for training
- Syntax Tree Builder
  - Java-based
  - Input from POS tagger
  - Handles idioms or compound words

Syntax Tree Builder



WHO_N umumkan_VT 80_Num penduduk_N dunia_N bebas_Adj polio_N

Organisasi_N kesehatan_N dunia_N ,_ WHO_N ,_ mengumumkan_VT bahwa_Con 80_Num penduduk_N dunia_N sekarang_N telah_Adv bebas_Adj polio_N ,_ setelah_Adv tidak_Adv c

WHO_N mengatakan_VT polio_N telah_Adv dibasmi_VT di_Par seluruh_Det wilayah_N Asia_NP Tenggara_N ,_ termasuk_VI India_NP tetapi_Con tidak_Adv termasuk_VI Pakistan_NP at

Sejak_Adv pertengahan_N 1990-an_Num India_NP menyelenggarakan_VT kampanye_N imunisasi_N besar-besaran_Adj ,_ dengan_Par memberikan_VT vaksinasi_N kepada_Par 170_N

Keberhasilan_N ini_Det dianggap_VT sebagai_Par pencapaian_N besar_Adj dalam_N upaya_N memerangi_VT virus_N polio_N yang_Pro mematikan_Adj ._.

Tetapi_Con polio_N ,_ yang_Pro bisa_Adv menyebabkan_VT kelumpuhan_N masih_Adv menjadi_VI masalah_N besar_Adj di_Par Pakistan_NP ,_ Afghanistan_NP ,_ dan_Con Nigeria_NI

Turki_NP blokir_VT akses_N YouTube_N

Turki_NP memblokir_VT akses_N ke_Par YouTube_N setelah_Adv pengadilan_N di_Par Turki_NP menangguhkan_VT larangan_N terhadap_Par Twitter_N

**** Invalid group! Cross-Parent grouping! ****

Future Plan

- Allocate budget for tree building in 2017 fiscal year
  - Approximately for 10.000 sentences
  - Will be offered to Indonesian Association for Computational Linguistics (INACL)
    - Community of researchers, developers, and observers in computational linguistics
    - Members are from universities, government agencies, and private companies
- Improving the tools for tagging and tree building
  - Adjust the output according to the standard ALT XML

Thank You