Indonesian Resource Grammar (INDRA)

David **Moeljadi**
Division of Linguistics and Multilingual Studies,
Nanyang Technological University, Singapore

The 11th DELPH-IN Summit,
Nanyang Technological University, Singapore

3 August 2015

Indonesian Resource Grammar (INDRA)

- The first broad-coverage, *open-source* **computational grammar** for Indonesian, modelled in **Head Driven Phrase Structure Grammar (HPSG)** and **Minimal Recursion Semantics (MRS)**
- Created and developed using tools from **Deep Linguistic Processing with HPSG Initiative (DELPH-IN)**
- Aims to parse and treebank Indonesian text in **the Nanyang Technological University — Multilingual Corpus (NTU-MC)**
- Will be applied to **machine translation**

Indonesian language

- Classification: Austronesian > ...> Western Malayo-Polynesian > ...> Malayic > Malay > Indonesian
- Alternate names: bahasa Indonesia
- Population: 43 million L1 speakers (2010 census), 156 million L2 speakers (2010 census)
- Language status: national language of Indonesia (1945 Constitution, Article 36)
- Dialects: over 80% lexical similarity with Standard Malay
- Writing: Latin script

Morphology and syntactic typology of Indonesian

- Morphological classification: mildly agglutinative
- Word order: SVO
- Position of negative word: S-Neg-V-O
- Order of Adj and Noun: N-Adj
- Order of Dem and Noun: N-Dem

Some Indonesian sentences

(1) ***X V-intransitive***
Adi tidur.
Adi sleep

"Adi sleeps."

(2) ***X V-transitive Y***
Adi mengejar Budi.
Adi ACT-chase Budi

"Adi chases Budi."

Previous work on Indonesian computational grammar

- No previous work done on a broad-coverage Indonesian HPSG grammar
- Much work has been done using Lexical Functional Grammar (LFG) (Kaplan and Bresnan, 1982)
  - Arka and Manning (2008) on active and passive voice
  - Arka (2000) on control constructions
- Arka (2012) and Mistica (2013) have worked on the computational grammar "IndoGram" which is a part of the ParGram (Sulger et al., 2013)
  - Has details of many phenomena

  but

  - Not *open-source*
  - Not very broad in its coverage
  - Does not produce MRS, so it cannot be easily incorporated into our machine translation system

Creation and development of INDRA

- Bootstrapped using The LinGO Grammar Matrix (Bender et al., 2010) (`http://www.delph-in.net/matrix/customize/matrix.cgi`)
  - Word order
  - Noun and verb subcategorization
  - Morphology
  - ...
- **Lexical acquisition**
- Additions and changes to Type Description Language (TDL) files
  - Pronouns, proper names and adjectives
  - **Decomposing words**
  - **Morphology**
  - ...
- Associated resources

Evaluation with MRS test-suite

- The original set of 107 sentences are in English, translated into many languages including Indonesian (172 sentences) (http://moin.delph-in.net/MatrixMrsTestSuiteIndonesian)
- 55 of 172 sentences (32%) can be parsed. INDRA is not currently able to parse the others.
- 15% more would be covered once passives and relative clauses were added

Associated resources

- Indonesian POS Tagger (Rashel et al., 2014) with ACE's YY-mode for unknown word handling
- Transfer grammar for machine translation

Future work

- Increase the coverage of (phenomena in) INDRA
    - Relative clauses
    - Numbers
    - Quantifiers
    - Classifiers
    - Copula constructions
    - Passive constructions
    - Topic-comment constructions
    - Particles
    - Interrogatives
    - Imperatives
- Simultaneously build up MT (learning **and** building rules)

INDRA Top page

`http://moin.delph-in.net/IndraTop`

- Specifications
- Test-suites
- Demo page

Acknowledgments

- Thanks to Michael Wayne Goodman for setting up the demo page, giving precious comments on the slides and sharing his knowledge about GitHub
- Thanks to Dan Flickinger for teaching us Full Forest Treebanker (FFTB)
- Thanks to Fam Rashel for helping us with POS Tagger
- Thanks to Lian Tze Lim for helping us improve Wordnet Bahasa
- This research was partly supported by the Singapore MOE ARF Tier 2 grant *That's what you meant: A Rich Representation for Manipulation of Meaning* (MOE ARC41/13) and by joint research with Fuji-Xerox Corporation on *Multilingual Semantic Analysis*

NANYANG
TECHNOLOGICAL
UNIVERSITY