Grammar Update for Indonesian Resource Grammar (INDRA)

David **Moeljadi**
Francis **Bond**, Sanghoun **Song**, Luis **Morgado da Costa**
and many more
Division of Linguistics and Multilingual Studies,
Nanyang Technological University, Singapore

The 12th DELPH-IN Summit,
Stanford University

16 June 2016

NANYANG TECHNOLOGICAL UNIVERSITY

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**

NANYANG TECHNOLOGICAL UNIVERSITY

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

NANYANG TECHNOLOGICAL UNIVERSITY

INDRA Top page

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

- Specifications
- Test-suites
- Demo page
  `http://chimpanzee.ling.washington.edu/demophin/indra`

NANYANG TECHNOLOGICAL UNIVERSITY

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

NANYANG TECHNOLOGICAL UNIVERSITY

Indonesian Morphology and Syntax

- 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
- **Reduplication**
- **(Zero) copula constructions**

- **Lexical Acquisition**

NANYANG TECHNOLOGICAL UNIVERSITY

Noun and Adjective Reduplication

- Reduplicated forms can have unreduplicated counterparts
  batu "stone(s)" > +REDUP > *batu-batu* "stones"
  mata "eye(s)" > +REDUP > *mata-mata* "eyes"
- Reduplicated forms can have no unreduplicated counterparts
  mata-mata "spy, spies"
  **mata-mata-mata-mata* FOR "spies"
- The adjective reduplication occurs when the noun it describes is plural
- See `http://moin.delph-in.net/LADIndonesianMorphology`

NANYANG TECHNOLOGICAL UNIVERSITY

(Zero) Copula Constructions

- Our analyses of Indonesian copula clauses are similar to Arka(2013)'s LFG analysis but cover more copula verbs with a refined type hierarchy

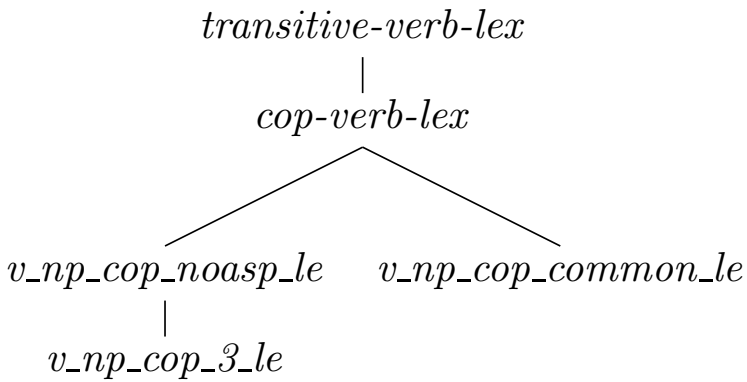

Figure: Type hierarchy of Indonesian copula verbs

- Our analysis also correspond to 'Constructional analysis II' in Bender (2001)
- Because of differences in syntactic structure, the constructional analysis which does not work for African American Vernacular English (AAVE), can be implemented for Indonesian.

NANYANG TECHNOLOGICAL UNIVERSITY

Lexical Acquisition

- 3,813 lexical items from NTU-MC have been added
  - 1,235 lex items (as of July 6, 2015) $\rightarrow$ 5,048 lex items (as of June 16, 2016)
  - Proper names such as Sentosa, Jurong, Din Tai Fung, etc. were not added
- plan to add more lexical items from The Great Dictionary of the Indonesian Language (*Kamus Besar Bahasa Indonesia* or KBBI), 4th edition, the official dictionary of the Indonesian language
  - we got a request to make a database for it

NANYANG TECHNOLOGICAL UNIVERSITY

Evaluation

	as of July 6, 2015	as of June 16, 2016
MRS test-suite	55/172 (32%)	65/172 (38%)
NTU-MC test-suite	1/2,197	8/2,197

- More phenomena to be covered: relative clauses, passives, zero-derivation (verbs-nouns)
- Lexical items from KBBI
- YY-mode in Demophin

NANYANG TECHNOLOGICAL UNIVERSITY

Acknowledgments

- Thanks to Michael Wayne Goodman for setting up the demo page
- 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
- Thanks to Dora Amalia from Badan Bahasa for sharing KBBI data
- We have benefited from VLAD discussion
- 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