Indonesian Resource Grammar (INDRA) Update

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

The 14th DELPH-IN Summit,
University of Chicago Center, Paris

18 June 2018

NANYANG TECHNOLOGICAL UNIVERSITY

INDonesian **R**esource gr**A**mmar (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)**
- Has a **treebank** called **JATI**, the text is from a subset of dictionary definition sentences: **Kamus Besar Bahasa Indonesia (KBBI) Fifth Edition**
- Indonesian POS Tagger (for unknown word handling), transfer grammars for machine translation
- `github.com/davidmoeljadi/INDRA` (MIT license)
- 2,057 types, 16,751 lexical items, 63 rules, 12 orules, 168 features (as of 23 Jan 2018)

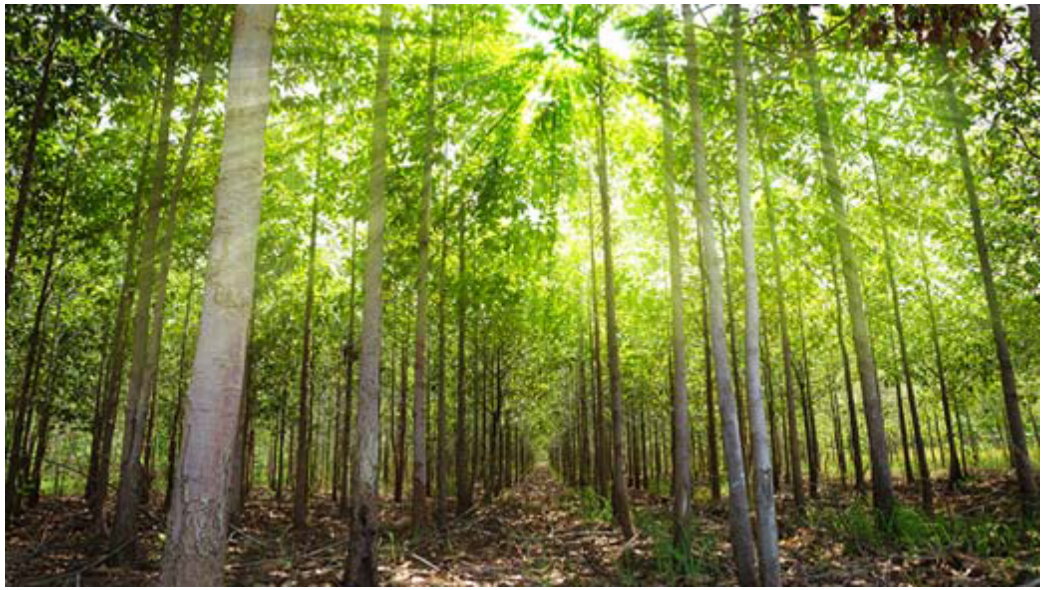
NANYANG TECHNOLOGICAL UNIVERSITY

Linguistic phenomena implemented in INDRA

- nouns
  - noun subcategorization
  - clitics
  - determiners
  - numerals and classifiers
  - reduplication
  - relative clause
- verbs
  - verb subcategorization
  - inflectional rules: active and passive voice
  - auxiliaries
- adjectives and prepositions
- copula constructions
- compounds

NANYANG TECHNOLOGICAL UNIVERSITY

JATI Treebank



- The Indonesian word for "teak", the national tree of Indonesia
- **J**... **A**... **T**reebank for **I**ndonesian ??
- 2,003 KBBI dictionary definition sentences related to food, drinks, spices, edible things were extracted and edited
  - total number of words: 23,129 words
  - shortest definition: 1 word
  - longest definition: 51 words
  - average: 11.5 words

NANYANG TECHNOLOGICAL UNIVERSITY

JATI Treebank

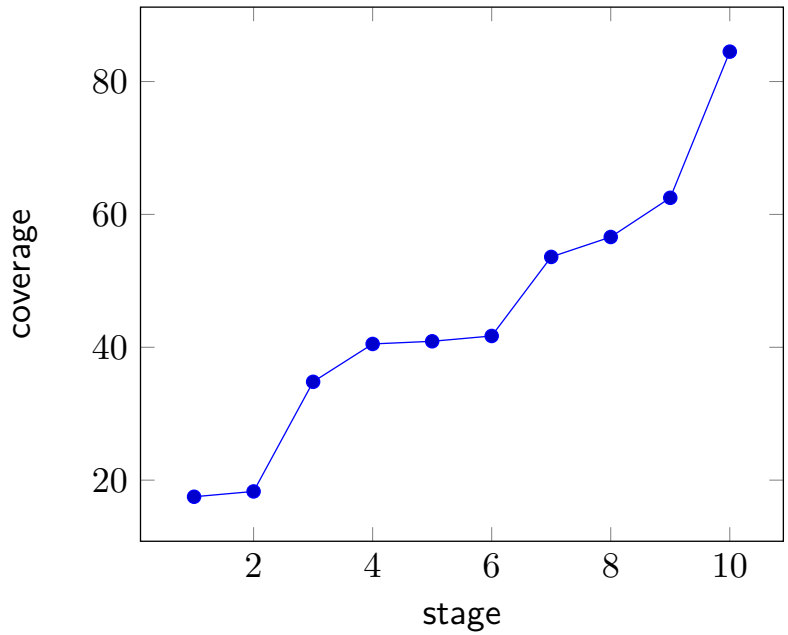


Figure: Evolution of coverage

- 3rd stage (18.3% → 34.8%): lexical acquisition (words)
- 10th stage (62.5% → 84.5%): adding homographs and compounds
- 62.6% (1,253 out of 2,003 sentences) have correct syntactic trees and semantics

NANYANG TECHNOLOGICAL UNIVERSITY

Evaluation

	Jun 16, 2016	Aug 7, 2017	Jan 23, 2018
MRS	65/172 (38%)	95/172 (55%)	122/172 (70.9%)
KBBI (JATI)	—	500/2004 (25%)	1692/2003 (84.4%)

More phenomena to be covered:

- possessor topic-comment relative clauses with many relative clauses
- equative, comparative, and superlative adjectives
- coordinate constructions with constituents having different POS
- and many more

NANYANG TECHNOLOGICAL UNIVERSITY

Some examples

(1) *buah* ***yang*** *mirip melon, berwarna jingga, kulit****nya*** *dipenuhi*
fruit REL resemble melon POSS-color orange peel=DEF PASS-fill
*tonjolan... dan daging buah****nya*** *lunak...*
bulge and flesh fruit=DEF soft
"a fruit which resembles a melon, has orange color, whose peel is filled with bulges... and whose flesh is soft..." (e99481m116467)

(2) *kentang* ***yang*** *daging****nya*** *padat*, *mengandung* *sedikit* *air*, *dan*
potato REL flesh=DEF solid ACT-contain little water and
digunakan *untuk...*
PASS-use for
"potato whose flesh is solid, contains little water, and is used to..."
(e39841m47126)

(Un)related to INDRA and JATI

- **MALINDO Morph**
  - a morphological dictionary and analyser for Malay/Indonesian
  - `https://github.com/matbahasa/MALINDO_Morph`
- **MALINDO Conc**
  - a new open online concordancer for Malay/Indonesian
  - `https://malindoconc.lagoinst.info/concordance/en/`
- **TUFS Asian Language Parallel Corpus (TALPCo)**
  - an open parallel corpus consisting of Japanese sentences and their translations into Burmese, Malay, Indonesian, and English
  - `https://github.com/matbahasa/TALPCo`

NANYANG TECHNOLOGICAL UNIVERSITY

Future plan/projects

- 30 Jun-2 Jul 2018: **HPSG 2018** (The 25th International Conference on Head-Driven Phrase Structure Grammar) in Tokyo, Japan
  - David Moeljadi and Francis Bond. *HPSG Analysis and Computational Implementation of Indonesian Passives*
  - David Moeljadi and Takayuki Kuribayashi. *Introduction and demo of Jacy: an implemented HPSG grammar of Japanese*
- 1-3 Aug 2018: **Seminar Leksikografi Indonesia** (Indonesian Lexicography Seminar) in Jakarta, Indonesia
- 28-31 Oct 2018: **Kongres Bahasa Indonesia XI** (The 11th Indonesian Language Congress) in Jakarta, Indonesia

NANYANG TECHNOLOGICAL UNIVERSITY

Future plan/projects

- **Automatic Methods for Detection of Morphological Classes in Under-resourced Languages** project
  - Dr. František Kratochvíl, Palacký University Olomouc
  - proposal submitted to the Czech Science Foundation
  - Malay/Indonesian grammar
  - deadline: April 2018 (submitted)
  - result: October 2018
  - start: January 2019 (for 3 years)

NANYANG TECHNOLOGICAL UNIVERSITY

Future plan/projects

- **Proyek Jawa Kuno** (Old Javanese project)
  - Dr. Arlo Griffiths, l'École française d'Extrême-Orient (EFEO)
  - application for the European Research Council (ERC) Advanced Grants
  - linguists, philologists, epigraphists and historians
  - Old Javanese grammar, dictionary, and corpus
    - ★ produce the first ever Diachronic Descriptive **Grammar** of Old Javanese
    - ★ create a greatly expanded and dynamic **Dictionary** of Old Javanese
    - ★ compile, (re-)edit, (re-)translate, (lexical and grammatical) tag primary sources/**corpora** (inscriptions, manuscripts) in XML files (TEI guidelines)
  - Fourth International Intensive course in Old Javanese, in Yogyakarta, Central Java, Indonesia, 15 - 29 July 2018
  - deadline: August 2018
  - result: April 2019
  - start: between September 2019 and January 2020 (for 5 years)

NANYANG TECHNOLOGICAL UNIVERSITY

Future plan/projects

- **Computer-aided language learning** project at Tokyo University of Foreign Studies (TUFS)
  - research proposal submitted to the Japan Society for the Promotion of Science (JSPS) Postdoctoral Fellowships for Research in Japan
  - INDRA for computer-aided Indonesian language learning
  - deadline: April 2018 (submitted)
  - result: mid August 2018
  - start: September 2018 (for 2 years)
- ...sending applications for other post-doc projects at other universities

NANYANG TECHNOLOGICAL UNIVERSITY

Thank you

Terima kasih