Grammar Update for Indonesian Resource Grammar (INDRA)

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

The 13th DELPH-IN Summit,
University of Oslo

7 August 2017

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 build a **treebank** called **JATI**, the text is from ~~the Nanyang Technological University — Multilingual Corpus (NTU-MC)~~ a subset of dictionary definition sentences: **Kamus Besar Bahasa Indonesia (KBBI) Fifth Edition**
- Will be applied to machine translation in the future
- 1,885 types, 15,592 lexical items, 38 rules, 8 orules, 47 instances, 168 features (as of 7 August 2017)

NANYANG TECHNOLOGICAL UNIVERSITY

Linguistic phenomena implemented in INDRA

2016:

- Noun and adjective reduplication
- Predicative and attributive adjective
- (Zero) copula constructions
- Negation word *bukan*
- Existential *ada* "there is/are"
- Noun noun compound

2017 (until July):

- NP fragment
- Numbers and optional classifiers
- Serial verb constructions
- Passive voice (type 1)
- Relative clause with *yang* "REL"

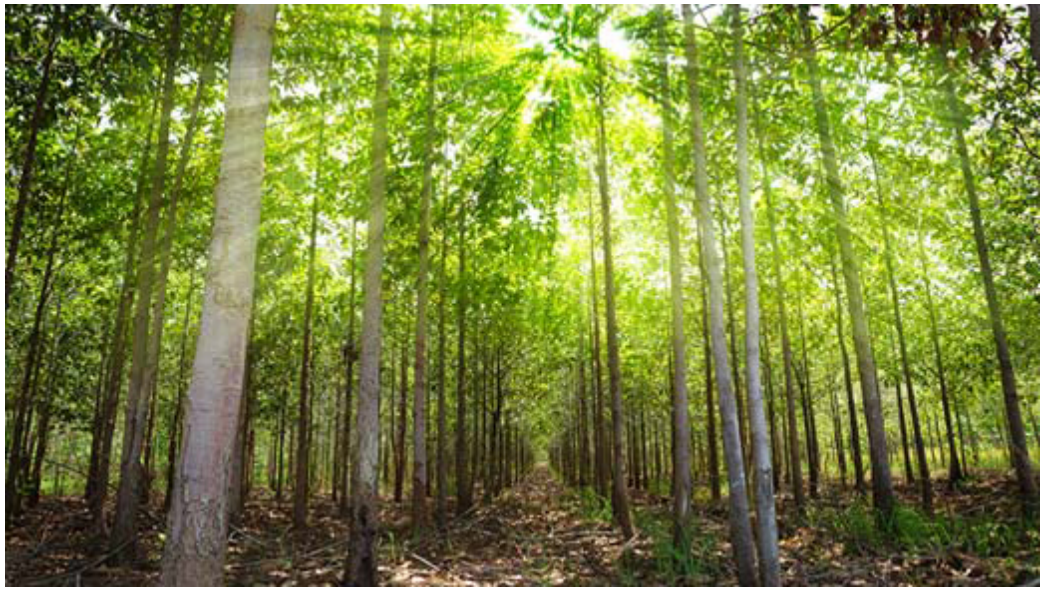
NANYANG TECHNOLOGICAL UNIVERSITY

Kamus Besar Bahasa Indonesia (KBBI)

- The official and the most comprehensive dictionary for the Indonesian language, published by Badan Pengembangan dan Pembinaan Bahasa (The Language Development and Cultivation Agency)
- Last year (2016), we got a request to make a database for the fourth edition from Word and Excel files and to add more entries for the fifth edition (also to make the online version and the Android and iOS mobile applications)
- 108,238 entries, 126,648 definitions, 29,261 examples (as of 6 August 2017)

NANYANG TECHNOLOGICAL UNIVERSITY

JATI Treebank



- The Indonesian word for "teak", the national tree of Indonesia
- 2,004 KBBI definition sentences related to food, drinks, spices, edible things were extracted and edited
  - shortest definition: 1 word
  - longest definition: 50 words
  - average: 11.7 words
- Work in progress: adding new rules (linguistic phenomena), lexical types, and lexical items (lexical acquisition)

NANYANG TECHNOLOGICAL UNIVERSITY

Evaluation

	Jun 16, 2016	Aug 7, 2017
MRS test-suite	65/172 (38%)	95/172 (55%)
KBBI test-suite (JATI)	—	500/2004 (25%)

* Lexical acquisition
* More phenomena to be covered:
  - relative clause **NP + *yang* + N=*nya***
    "NP whose N, NP of which the N"
  - [**ber-** "have/possess" **+ NP**]$_{\text{AP}}$
  - and many more

Some examples

(1) *acar* ***yang*** *bumbu**nya*** *diberi* *kunyit* *sehingga* ***ber**warna*
pickle REL spice=DEF PASS-give turmeric so.that POSS-color
kuning
yellow
lit. "pickles of which the spices are given turmeric so that having yellow color"

(2) *makanan* ***ber**gizi*
food POSS-nutrient
(1) S: "food has nutrients"
(2) NP: "food having nutrients, nutritious food"

(3) *makanan* *yang* ***ber**gizi,* *rendah* *kalori,* *lemak,* *dan* *gula*
food REL POSS-nutrient low calorie fat and sugar
lit. "food which has nutrients, low in calories, fat, and sugar"

Thank you

Terima kasih

NANYANG TECHNOLOGICAL UNIVERSITY