Building an HPSG-based Indonesian Resource Grammar (INDRA)

David **Moeljadi**, Francis **Bond**, Sanghoun **Song**
{D001,fcbond,sanghoun}@ntu.edu.sg

Division of Linguistics and Multilingual Studies,
Nanyang Technological University
Singapore

30 July 2015

Why we need the Indonesian Resource Grammar (INDRA)?

- No broad-coverage, *open-source* computational grammar for Indonesian
- No robust Indonesian grammar modelled in **Head Driven Phrase Structure Grammar (HPSG)** and **Minimal Recursion Semantics (MRS)** framework
- No robust **rule-based machine translation** for Indonesian

Indonesian Resource Grammar (INDRA)

- The first broad-coverage, *open-source* **computational grammar** for Indonesian, modelled in **HPSG** and **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 Indonesian HPSG
- 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

DEep Linguistic Processing with HPSG - INitiative (DELPH-IN)

- Research collaboration between linguists and computer scientists adopting HPSG and MRS
- Builds and develops *open-source* grammar
  - English Resource Grammar (ERG)
  - Jacy (Japanese grammar)
  - ...
- Typed feature structures are defined using Type Description Language (TDL)
- Builds and develops *open-source* tools for grammar development
  - Grammar and lexicon development environment (LKB)
  - A web-based questionnaire for writing new grammars (The LinGO Grammar Matrix)
  - Efficient parsers/generators (ACE)
  - Dynamic treebanking (ITSDB, FFTB, ACE)
  - Machine Translation engine (LOGON, ACE)

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 TDL files
  - Pronouns, proper names and adjectives
  - **Decomposing words**
  - **Morphology**
  - ...
- Associated resources

Lexical acquisition

- Assumptions
  - Manually building a lexicon is labor-intensive and time-consuming
    - (Semi-)automatic lexical acquisition is vital
    - Wordnet Bahasa can be the lexical source
  - The number of arguments of verbs with similar meaning should be the same across languages
    - Verb subcategorization in ERG can be used
- Verbs in ERG
  - 345 verb types: intransitive, transitive, 'be'-type etc.
  - Top 11 most frequently used types in the corpus were chosen
    - Verb of motion (+PP): *go*, *come*
    - Intransitive: *occur*, *stand*
    - Verb with optional complementizer: *believe*, *know*
    - ...

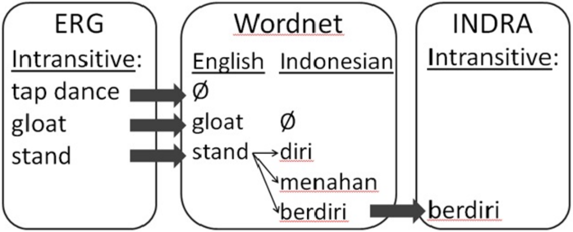
Wordnet verb frames for lexical acquisition

- Wordnet Bahasa
  - Groups nouns, verbs, adjectives and adverbs into sets of concepts or **synsets**
  - Verb frames or subcategorization for each verb

Synset	Definition	Verb frame
01168468-v	Take in solid food	8 Somebody —s something
01166351-v	Eat a meal, take a meal	2 Somebody —s
01157517-v	Use up (resources or materials)	11 Something —s something 8 Somebody —s something

Table: Three of 69 synsets of *makan* "eat" and their verb frames in Wordnet

Workflow of lexical acquisition and results



1. Check whether the verb is in Wordnet
2. Check whether the verb has Indonesian translation(s)
3. Check whether the verb has the correct verb frame(s)
4. Check manually the Indonesian translation(s)

Result: 939 subcategorized verbs and 6 rules were added

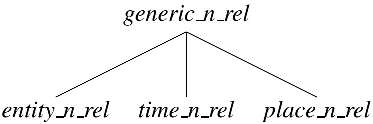
Decomposed words

- Assumption: pronouns can be decomposed across grammars (Seah and Bond, 2014)
  e.g. *sini* “here” –> *tempat* “place” + *ini* “this”

	proximal	medial	remote
Demonstratives	ini “this”	itu “that”	
Locatives	sini “here”	situ “there” (not far off)	sana “there” (far off)

Table: Demonstrative and locative pronouns in Indonesian

Type hierarchies for heads and demonstratives



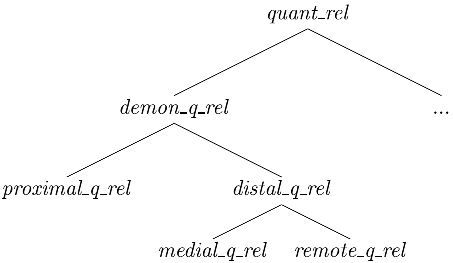
generic_n_rel

entity_n_rel

time_n_rel

place_n_rel

Figure: Type hierarchy for heads



quant_rel

demon_q_rel

...

proximal_q_rel

distal_q_rel

medial_q_rel

remote_q_rel

Figure: Type hierarchy for demonstratives

MRS representations of *di situ* “there”

$$\begin{bmatrix} \textit{mrs} & \\ \text{TOP} & \boxed{0} \\ \text{INDEX} & \boxed{2} \\ \text{RELS} & \left\langle \begin{bmatrix} \textit{\_di\_p\_rel} & \\ \text{LBL} & \boxed{1} \\ \text{ARG0} & \boxed{2} \\ \text{ARG1} & \boxed{3} \\ \text{ARG2} & \boxed{4} \end{bmatrix}, \begin{bmatrix} \textit{place\_n\_rel} & \\ \text{LBL} & \boxed{5} \\ \text{ARG0} & \boxed{4} \end{bmatrix}, \begin{bmatrix} \textit{medial\_q\_rel} & \\ \text{LBL} & \boxed{6} \\ \text{ARG0} & \boxed{4} \\ \text{RSTR} & \boxed{7} \\ \text{BODY} & \boxed{8} \end{bmatrix} \right\rangle \\ \text{HCONS} & \left\langle \begin{bmatrix} \textit{qeq} & \\ \text{HARG} & \boxed{0} \\ \text{LARG} & \boxed{1} \end{bmatrix}, \begin{bmatrix} \textit{qeq} & \\ \text{HARG} & \boxed{7} \\ \text{LARG} & \boxed{5} \end{bmatrix} \right\rangle \end{bmatrix}$$

Figure: MRS representation of *di situ* (lit. “at there”)

Morphology

Inflection with active prefix *meN-* and passive prefix *di-*

(3) a. ***X meN-kejar Y***
Adi mengejar Budi.
Adi ACT-chase Budi

"Adi chases Budi."

b. ***Y di-kejar X***, *X is a 3rd person pronoun or a noun*
Budi dikejar Adi.
Budi PASS-chase Adi

"Budi is chased by Adi."

c. ***Y X kejar***, *X is a pronoun or pronoun substitute*
Budi saya kejar.
Budi 1SG chase

"Budi is chased by me."

Morphology of *meN-*

A number of sound changes occur when *meN-* combines with bases

Base	meN-+base	meaning
***p**akai*	*me**m**akai*	use
***t**anam*	*me**n**anam*	plant
***k**ejar*	*me**ng**ejar*	chase
***pr**oses*	*me**mpr**oses*	process

Base	meN-+base	meaning
***b**eli*	*me**mb**eli*	buy
***d**apat*	*me**nd**apat*	get
***g**anti*	*me**ngg**anti*	replace
bom	*me**ngebom***	bomb

Morphology of *meN-*

Allomorph	Initial orthography of the base		Example
mem-	p	(L)	*me**m~~p~~**akai* “use”
	pl, pr, ps, pt, b, bl, br, f, fl, fr, v	(R)	*me**mb**eli* “buy”
men-	t	(L)	*me**n~~t~~**anam* “plant”
	tr, ts, d, dr, c, j, sl, sr, sy, sw, sp, st, sk, sm, sn, z	(R)	*me**nc**ari* “seek”
meny-	s	(L)	*me**ny~~s~~**ewa* “rent”
meng-	k	(L)	*me**ng~~k~~**irim* “send”
	kh, kl, kr, g, gl, gr, h, q, a, i, u, e, o	(R)	*me**ngg**anti* “replace”
me-	m, n, ny, ng, l, r, w, y	(R)	*me**l**empar* “throw”
menge-	(base with one syllable)		*me**ngecek*** “check”

Parse tree result

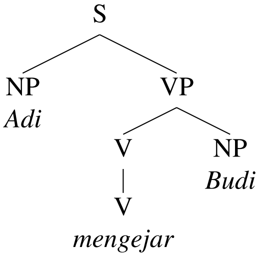


Figure: Parse tree of *Adi mengejar Budi* “Adi chases Budi”

MRS result

$$\begin{bmatrix} \textit{mrs} & \\ \text{TOP} & \boxed{0} \\ \text{INDEX} & \boxed{2} \\ \text{RELS} & \left\langle \begin{bmatrix} \textit{named\_rel} & \\ \text{LBL} & \boxed{4} \\ \text{CARG} & \textit{"adi"} \\ \text{ARG0} & \boxed{3} \end{bmatrix}, \begin{bmatrix} \textit{proper\_q\_rel} & \\ \text{LBL} & \boxed{6} \\ \text{ARG0} & \boxed{3} \\ \text{RSTR} & \boxed{7} \\ \text{BODY} & \boxed{8} \end{bmatrix}, \begin{bmatrix} \textit{\_kejar\_v\_rel} & \\ \text{LBL} & \boxed{1} \\ \text{ARG0} & \boxed{2} \\ \text{ARG1} & \boxed{3} \\ \text{ARG2} & \boxed{9} \end{bmatrix}, \begin{bmatrix} \textit{named\_rel} & \\ \text{LBL} & \boxed{10} \\ \text{CARG} & \textit{"budi"} \\ \text{ARG0} & \boxed{9} \end{bmatrix}, \begin{bmatrix} \textit{proper\_q\_rel} & \\ \text{LBL} & \boxed{12} \\ \text{ARG0} & \boxed{9} \\ \text{RSTR} & \boxed{13} \\ \text{BODY} & \boxed{14} \end{bmatrix} \right\rangle \\ \text{HCONS} & \left\langle \begin{bmatrix} \textit{qeq} & \\ \text{HARG} & \boxed{0} \\ \text{LARG} & \boxed{1} \end{bmatrix}, \begin{bmatrix} \textit{qeq} & \\ \text{HARG} & \boxed{7} \\ \text{LARG} & \boxed{4} \end{bmatrix}, \begin{bmatrix} \textit{qeq} & \\ \text{HARG} & \boxed{13} \\ \text{LARG} & \boxed{10} \end{bmatrix} \right\rangle \end{bmatrix}$$

Figure: MRS representation of *Adi mengejar Budi* “Adi chases Budi”

Evaluation with MRS test-suite

- MRS test-suite: a representative set of sentences designed to show some of the semantic phenomena
- 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

	results / items	coverage
before	52 / 172	30.2%
after	55 / 172	32.0%

Table: Comparison of coverage in MRS test-suite before and after lexical acquisition

Associated resources

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

Nanyang Technological University Multilingual Corpus (NTU-MC)

- Parallel corpus, sense-tagged using Wordnet (lexical database) (`http://compling.hss.ntu.edu.sg/ntumc/`)
- Indonesian text data contains 2,197 sentences from Singapore Tourism Board (STB) website (`http://www.yoursingapore.com`)
- Ongoing process of adding Sherlock Holmes short stories
- INDRA aims to parse at least 60% of the NTU-MC Indonesian text in 2.5 years

Future work

- Increase the coverage of (phenomena in) INDRA
- Simultaneously build up MT (learning **and** building rules)
- Lexical acquisition
  - Extract more words from various parts-of-speech
    Simultaneously add lexical types, rules and constraints
  - Improve Wordnet Bahasa
    Wordnet Bahasa is growing, so hopefully the semi-automatic methodology for lexical acquisition may give better results
- Decomposed words
  - Expand to other heads such as *time_n_rel* and *entity_n_rel*
- Morphology
  - Cover all the exceptions
  - Expand to other verb types such as ditransitives
  - Analyze and implement passive constructions

Future work

- Phenomena to be covered
  - Relative clauses
  - Numbers
  - Quantifiers
  - Classifiers
  - Copula constructions
  - Passive constructions
  - Topic-comment constructions
  - Particles
  - Interrogatives
  - Imperatives

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