Building JATI: A Treebank for Indonesian

David **Moeljadi**

Nanyang Technological University, Singapore

The 4th Atma Jaya Conference on Corpus Studies (ConCorps 2017),
Atma Jaya Catholic University, Jakarta

21 July 2017

Outline

1. What is a treebank?
2. Indonesian treebanks
3. The corpus: Kamus Besar Bahasa Indonesia (KBBI)
4. The parser: Indonesian Resource Grammar (INDRA)
5. Treebank development
6. Summary and future work

Treebank

- A treebank is a linguistically annotated corpus that includes some grammatical analysis beyond the part-of-speech level [8]
- Usages:
  - empirical linguistic research, as well as Natural Language Processing (NLP)
  - enables more precise queries
  - in qualitative research, such as finding an example of a certain linguistic construction or a counter-example to a claim about syntactic structure
  - in quantitative research, as a source of information about frequencies and co-occurrences
  - building statistical model, robust broad-coverage parsing
  - developing a broad-coverage grammar, test the grammar

Motivation

- We want to understand natural language
  - it is interesting in and of itself
  - it offers a view into human cognition
  - much knowledge is encoded in natural language
  - we want to make computers understand
- What does it mean for a machine to understand?
  - The system analyses text and grows clever
    - ★ it increase the lexicon
    - ★ it builds up the ontology
    - ★ it changes the stochastic model

Indonesian treebanks

- The Indonesian Dependency Treebank developed by Charles University in Prague [5]
- The Indonesian Treebank developed by the Faculty of Computer Science of University of Indonesia [4]
- The Indonesian Treebank in the Asian Language Treebank (ALT), built by the Agency for the Assessment and Application of Technology (BPPT) [13]
- the Indonesian Treebank in the ParGram Parallel Treebank (ParGramBank), based on LFG "IndoGram" [15]

Other treebanks

- Penn Treebank
- The LinGO Redwoods Treebank of English [11]
- Hinoki [2]

JATI Overview

- Based on an HPSG grammar of Indonesian: Indonesian Resource Grammar (INDRA) [6]
  We want to develop a broad-coverage grammar together with the treebank. Treebanking allows us to immediately identify problems in the grammar and improving the grammar directly improves the quality of the treebank [9]
- Parsing (a subset of) dictionary definition sentences: KBBI Fifth Edition [1]
- Creating a corpus that can be studied: JATI

The corpus: Kamus Besar Bahasa Indonesia (KBBI)

- The fifth edition of KBBI [1], published by Badan Pengembangan dan Pembinaan Bahasa
- The KBBI database, a machine-tractable dictionary [7]
- 108,240 entries, 126,643 definitions, 29,260 examples (as of 15 June 2017)

KBBI definition sentences

Definitions related to food, drinks, spices, edible things are extracted and edited

Before	After
minuman keras *yg* dibuat *dr* nira *yg* telah disuling	minuman keras *yang* dibuat *dari* nira *yang* telah disuling
kue kering, dibuat *dr* sagu dan dibungkus *dng* daun nipah	kue kering *yang* dibuat *dari* sagu dan dibungkus *dengan* daun nipah
makanan *terbuat dr* daging, udang, ikan *yg* dicincang	makanan *yang dibuat dari* daging, udang, *atau* ikan *yang* dicincang

- Shorter, compared with other commonly used text for corpora, such as newspaper text
- Contain more fragments, especially noun phrases
- Valid examples of naturally occurring texts

The parser: Indonesian Resource Grammar (INDRA)

- open-source Indonesian computational grammar [6]
  `https://github.com/davidmoeljadi/INDRA`
- parse and generate Indonesian text
- open-source tools in Deep Linguistic Processing with HPSG Initiative (DELPH-IN)
  - Documentation (`http://moin.delph-in.net/IndraTop`)
  - ITSDB or [incr tsdb()] [10]
  - Full Forest Treebanker (FFTB) [12]
- theoretical framework of Head Driven Phrase Structure Grammar (HPSG) [14]
- Minimal Recursion Semantics (MRS) [3]
- 1,885 types, 15,099 lexical items, 38 rules (as of 15 June 2017)

Choosing a Grammar

HPSG is chosen for the following reasons:

- Serious attempt to cover linguistic phenomena both core and periphery
- unification- and constraint-based context free grammar (phrase structure grammar)
  - consists of a set of rules and a lexicon of symbols (parts-of-speech) and words, surface oriented (no additional abstract structures)
- Integration of syntax and semantics (mono-stratal)
  we are most interested in semantics
  - tractable representation: MRS
- A vibrant research community
  - well developed open source tools
  - integration with shallow processing

Open Resources: DELPH-IN

Deep Linguistic Processing with HPSG Initiative

- Grammars: English (ERG), Japanese (JACY), Chinese (Zhong), Indonesian (INDRA), ...
- Development Environment: Linguistic Knowledge Builder (LKB)
- Processor: Answer Constraint Engine (ACE)
- Test Environment: ITSDB or [incr tsdb()]
- Treebanking tools: FFTB
- Machine Translation: LOGON

Approaches to Treebanking

- Manual Annotation
- Semi-Automatic
  - **Parse and repair by hand**: Penn WSJ, Kyoto Corpus

    ↑ 100% cover, reasonably fast

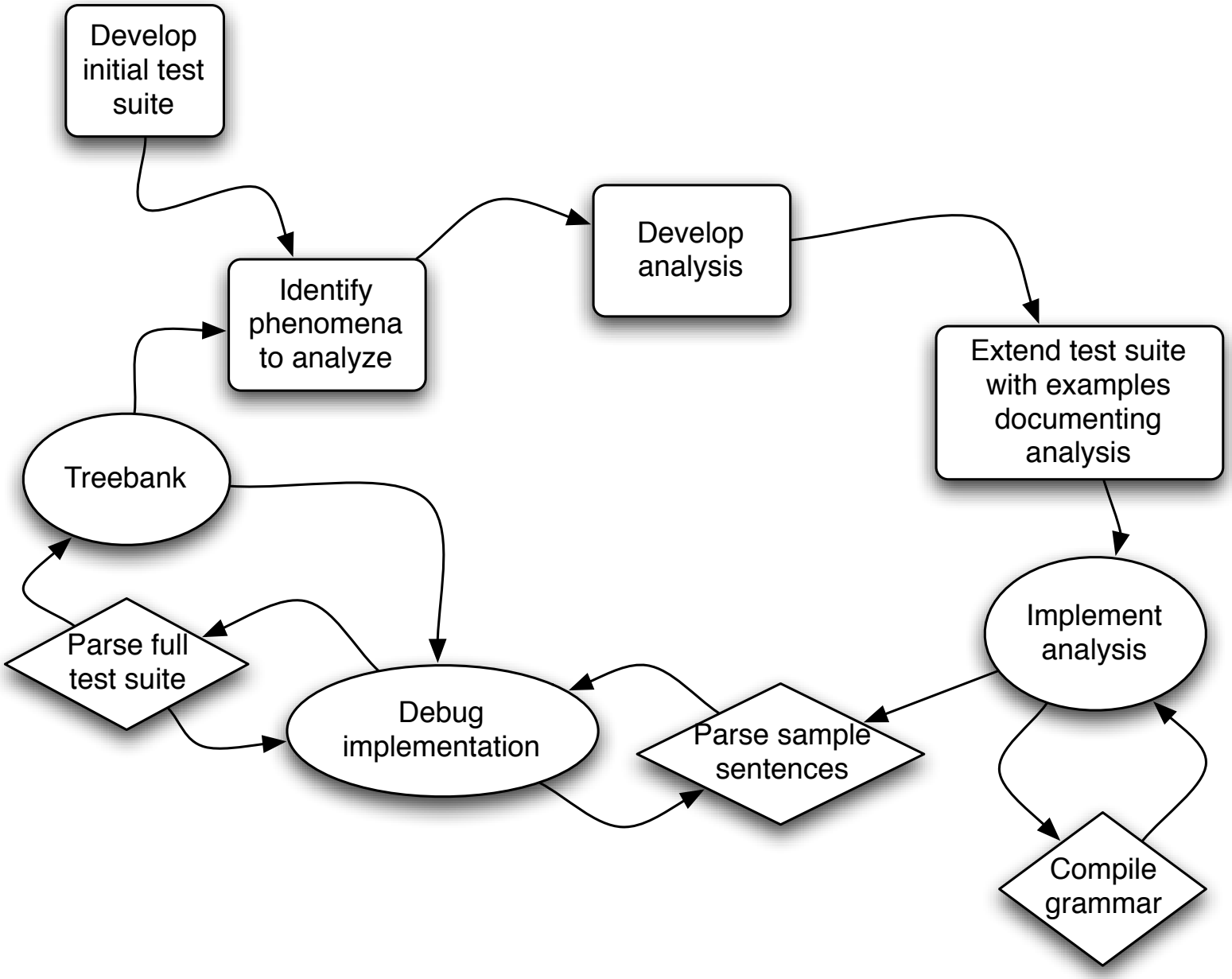
    ↓ Often inconsistent, Hard to update,
    Simple grammars only (prop-bank is separate)
  - **Parse and select by hand**: Redwoods, Hinoki, **JATI**

    ↑ All parses grammatical, Feedback to grammar,
    Consistent
    Both syntax and semantics, Easy to update

    ↓ Cover restricted by grammar

    ★ **Discriminant-based treebanking**: select or reject discriminants until one parse remains

Grammar development



Summary and future work

- Refining the analyses
  - Improving INDRA by adding new rules and lexical types
- Automate analysis
  - parse ranking
- Expanding the system
  - Adding non-familiar words (lexical acquisition)
  - Dynamic handling of unknown words

Long Term Goals

- Make text understanding available to everyone
  - Machine translation
  - Question answering
  - Speech recognition
  - Man-machine interfaces
- Link words to meanings for all languages

Acknowledgments

- Thanks to Francis Bond for his inspiration and advice to build JATI
- Thanks to Dora Amalia who gave permission to use a part of the fifth edition of KBBI data
- Some slides use material from:
  - "The Hinoki Treebank: Toward Text Understanding" by Francis Bond, Sanae Fujita, Chikara Hashimoto, Shigeko Noriyama, Eric Nichols, Takaaki Tanaka, and Hiromi Nakaiwa
  - "Treebanking an Open Forest: The Tanaka Corpus" by Francis Bond and Takayuki Kuribayashi

References I

Dora Amalia, ed. *Kamus Besar Bahasa Indonesia*. 5th ed. Jakarta: Badan Pengembangan dan Pembinaan Bahasa, 2016. ISBN: 9786024371715.

Francis Bond et al. "The Hinoki Treebank: A Treebank for Text Understanding". In: *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*. Springer Verlag Lecture Notes in Computer Science, 2004, pp. 158–167.

Ann Copestake et al. "Minimal Recursion Semantics: An Introduction". In: *Research on Language and Computation* 3.4 (2005), pp. 281–332.

Arawinda Dinakaramani et al. "Developing (and Utilizing) an Indonesian Treebank". In: *The Second Wordnet Bahasa Workshop*. Nanyang Technological University, Singapore, Jan. 2016.

References II

Nathan Green, Septina Dian Larasati, and Zdeněk Žabokrtský. "Indonesian Dependency Treebank: Annotation and Parsing". In: *26th Pacific Asia Conference on Language, Information and Computation*. 2012, pp. 137–145.

David Moeljadi, Francis Bond, and Sanghoun Song. "Building an HPSG-based Indonesian Resource Grammar (INDRA)". In: *Proceedings of the GEAF Workshop, ACL 2015*. 2015, pp. 9–16. URL: http://aclweb.org/anthology/W/W15/W15-3302.pdf.

David Moeljadi, Ian Kamajaya, and Dora Amalia. "Building the Kamus Besar Bahasa Indonesia (KBBI) Database and Its Applications". In: *Proceedings of the 11th International Conference of the Asian Association for Lexicography*. Ed. by Hai Xu. the Asian Association for Lexicography. Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, 2017, pp. 64–80.

References III

Joakim Nivre. "Treebanks". In: *Corpus Linguistics: An International Handbook*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 1. Berlin: Walter de Gruyter, 2008. Chap. 13, pp. 225–241.

Stephan Oepen, Dan Flickinger, and Francis Bond. "Towards Holistic Grammar Engineering and Testing–Grafting Treebank Maintenance into the grammar revision cycle". In: *Beyond Shallow Analyses–Formalisms and Statistical Modelling for Deep Analysis (Workshop at IJCNLP-2004)*. Hainan Island, 2004.

Stephan Oepen and Daniel Flickinger. "Towards systematic grammar profiling: Test suite technology ten years after". In: *Journal of Computer Speech and Language*. Vol. 12. 4. 1998, pp. 411–436.

Stephan Oepen et al. "LinGO Redwoods: A Rich and Dynamic Treebank for HPSG". In: *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*. Sozopol, Bulgaria, 2002.

References IV

Woodley Packard. *FFTB: the full forest treebanker*. Dec. 2014. URL: http://moin.delph-in.net/FftbTop (visited on 04/24/2015).

Hammam Riza et al. "Introduction of the Asian Language Treebank". In: *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)*. IEEE. 2016, pp. 1–6.

Ivan A. Sag, Thomas Wasow, and Emily M. Bender. *Syntactic Theory: A Formal Introduction*. 2nd ed. Stanford: CSLI Publications, 2003.

Sebastian Sulger et al. "ParGramBank: The ParGram Parallel Treebank." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. Sofia, Bulgaria, Aug. 2013, pp. 550–560.

Thank you

te.ri.ma ka.sih *n* rasa syukur;

ber.te.ri.ma ka.sih *v* mengucap syukur; melahirkan rasa syukur atau membalas budi setelah menerima kebaikan dsb