**Indonesian computational grammar**

A grammar (a set of syntactic and semantic production rules for strings in a formal language) written explicitly enough to be interpreted by a computer.

Why?

- Consistency of grammatical analysis
- Robustness
- Efficient processing
- Language learning

**HOW?**

- Syntax and semantics
- Linguistic theory
- Linguistic resources
- Component platform
- Computer

**Indonesian language**

- Austronesian > Malayo-Polynesian > ... > Malay
- Speakers: 43 mio (L1), 156 mio (L2) (2010 census)
- Diglossic
- Language resources: KBBI Indonesian dictionary fifth edition, Indonesian Reference Grammar (Sneddon et al., 2010), etc.
- Mildly agglutinative, SVO, N-Adj-Dem

**Linguistic theory**

- Head-Driven Phrase Structure Grammar (HP SG; Pollard and Sag, 1994; Sag et al., 2003)
- Minimal Recursion Semantics (MRS; Copestake et al., 2005)

---

**Phrase Structure Grammar**

1. A set of rules
   - S → NP VP
   - NP → N, NP → N D
   - VP → V, VP → V1 NP

2. A lexicon of parts-of-speech (POS) and words

**HPSG - Feature structures and unification**

(1)

```
<table>
<thead>
<tr>
<th>NP</th>
<th>VP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>anjing ini</td>
<td>menggonggong</td>
</tr>
</tbody>
</table>
```

(2)

```
<table>
<thead>
<tr>
<th>NP</th>
<th>VP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>ini</td>
<td>kucing mengjar</td>
</tr>
</tbody>
</table>
```

**Applications: Treebank & LTDB**

- A linguistically annotated corpus that includes some grammatical analysis beyond the POS level
- To find an example of a certain linguistic construction or a counter-example to a claim about syntactic structure, as a source of information about frequencies and co-occurrences, to build a statistical model, grammar development
- Semi-automatic (parse and select by hand)
- JATI: 1,253 items of KBBI Fifth Edition dictionary definition sentences related to food and drinks
- Cendana: 552 items of Traveloka (Indonesian online travel agent) customer service chat data
- Linguistic type database (Hashimoto et al., 2007)

**Grammar development cycle**

- Implement
- Parse
- Treebank
- Analyze

**Summary**

- INDRA is an open-source HPSG Indonesian computational grammar (MIT license)
- Phenomena covered: voice, copulas, full noun reduplication, noun & verb subcategorization etc.
- Homepage: http://main.delph-in.net/
- Applications: sentence parsing and generation, treebanking, linguistic type database etc.

**Acknowledgements**

I gratefully acknowledge the support of the European Regional Development Fund-Project “Spino- phone Borderlands - Interaction at the Edges” CZ.02.1.01/0.0/16.019/0000791.

**References**


---

**OPTIONS**

- Treebank
- Linguistic type database
- Computational grammar
- Text corpus

**Tools**

- Deep Linguistic Processing with HPSG - Initative (DELPH-IN) (www.delph-in.net)
- Open-source grammars:
  - ERC English grammar (Flickinger, 2000)
  - Jacy Japanese grammar (Sagel et al., 2016) etc.
- Tools:
  - The Linguistic Knowledge Builder (LKB) grammar development system (Copestake, 2002) etc.

---

David Moeljadi <davidmoeljadi@gmail.com> 11th International Austronesian and Papuan Languages and Linguistics Conference (APLL11) Leiden, 13–15 June 2019