Jacy: an implemented HPSG grammar of Japanese

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

The 25th International Conference on Head-Driven Phrase Structure Grammar
University of Tokyo, Komaba Campus

2 July 2018
Jacy demo: Outline

1. Introduction
   - Motivation
   - History and applications
   - Deep Linguistic Processing with HPSG Initiative (DELPH-IN)
   - Grammar engineering
   - The current state
     - Covered phenomena
     - Coverage and evaluation
     - Corpus/Treebank

2. Phenomena *DEMO
   - Argument scrambling and omission
   - -reru / -rare ru verbal endings

3. Treebanking *DEMO

4. Japanese-English machine translation *DEMO

5. Conclusions and future work
Motivation

- **Applications** that rely on deep linguistic processing, such as message extraction systems, machine translation and dialogue understanding systems are becoming feasible.

- **Requirement** for rich and highly precise information, well-defined output structures.

- **Requirement** for robustness: wide coverage, large and extensible lexica, interfaces to preprocessing.

- **Requirement** for extensibility to multiple languages.

- **Requirement** for efficient processing.

- The JACY Japanese HPSG has been developed for and used in real-world applications that require the handling of peripheral phenomena.
History of the JACY grammar: Project context

- **1998-2000**
  - **Verbmobil**: Machine translation of application-oriented spoken dialogues (http://verbmobil.dfki.de/)
- **2001-2002**
  - Co-operation with YY Technologies (CA, USA): Automatic email response
    (Co-operation with Stephan Oepen, Ulrich Callmeier, Monique Sugimoto, Atsuko Shimada, Dan Flickinger)
    (http://www.dfki.de/~siegel/jacy/jacy.html)
- **2002-2004**
  - EU project **DeepThought**: Hybrid and shallow methods for knowledge-intensive information extraction
    (http://www.project-deepthought.net)
- Lexeed project at Nippon Telegraph and Telephone Corporation: Ontology extraction, **Hinoki** treebank
- Japanese-English machine translation project with the LOGON initiative: open-source semantic transfer-based machine translation — **JaEn**
Deep Linguistic Processing with HPSG Initiative (DELPH-IN)

- a research collaboration between linguists and computer scientists
- builds and develops open source grammar, tools for grammar development and NLP applications using HPSG and MRS
  - Head-Driven Phrase Structure Grammar (**HPSG**; Pollard and Ivan A Sag, 1994; Ivan A. Sag, Wasow, and Emily M. Bender, 2003): feature structures, type hierarchy, efficient processing
  - Minimal Recursion Semantics (**MRS**; Copestake et al., 2005): flat semantic formalism, works well with typed feature structures, structures are underspecified for scopal information (compact representation of ambiguities)
- 18-22 June 2018: The 14th **Annual DELPH-IN Summit**, hosted by Berthold Crysmann (Laboratoire de linguistique formelle, CNRS & U Paris Diderot)
- **wiki** page: http://moin.delph-in.net/FrontPage
- **DELPH-IN discourse** (Q&A): https://delphinqa.ling.washington.edu/
The Development Tools

- The Linguistic Knowledge Builder (LKB) (Copestake, 2002): grammar development system
- Platform for Experimentation with efficient HPSG processing Techniques (PET) (Callmeier, 2000): a very efficient HPSG parser, for processing
- Answer Constraint Engine (ACE) (Packard, 2013): an efficient processor for DELPH-IN HPSG grammars
- ITSDB or [incr tsdb()] (pronounced tee ess dee bee plus plus) (Oepen and Daniel Flickinger, 1998): a tool for testing, profiling the performance of the grammar (analyzing the coverage and performance), tracking changes, and annotating treebanks
- Full Forest Treebanker (FFTB) (Packard, 2014): a treebanking tool for DELPH-IN grammars, allowing the selection of an arbitrary tree from the “full forest” without enumerating/unpacking all analyses in the parsing stage
Multilingual grammar development

- **English Resource Grammar (ERG)** (Dan Flickinger, 2000; Dan Flickinger, 2011)
- **Jacy** (Siegel, Emily M Bender, and Bond, 2016)
- **Zhong** (Fan, Song, and Bond, 2015), for Chinese languages (Mandarin, Cantonese, ...)
- **Indonesian Resource Grammar (INDRA)** (Moeljadi, Bond, and Song, 2015), for Indonesian
- ...
- **The LinGO Grammar Matrix** (Emily M. Bender, Dan Flickinger, and Oepen, 2002) (Emily M. Bender, Drellishak, et al., 2010): a web-based questionnaire for writing new DELPH-IN grammars
Other tools

- **delphin-viz**: DELPH-IN data structure visualizations and demo interface
  
  http://delph-in.github.io/delphin-viz/demo/

- **Demophin**: a DELPH-IN web demo
  
  http://chimpanzee.ling.washington.edu/demophin/jacy/

- **PyDelphin**: a set of Python libraries for the processing of DELPH-IN data
  
  https://github.com/delph-in/pydelphin

- **typediff**: a tool to investigate and compare phenomena in one grammar (e.g. JACY) with those in other DELPH-IN grammars (e.g. ERG)
  
  https://github.com/ned2/typediff

- **Linguistic Type Data-Base (LTDB)**: a documentation containing linguistic description of lexical types, usage examples and distribution based on the grammar and treebanks, typed feature structure definitions of the lexical types
  
  https://github.com/fcbond/ltdb
  
  http://compling.hss.ntu.edu.sg/ltdb/Jacy_1301/
Grammar engineering

Figure: Grammar Development Cycle (Emily M. Bender, Dan Flickinger, and Oepen, 2011)
Grammar engineering

- Grammar engineering courses:
  http://moin.delph-in.net/TeachingCourses

- Grammar engineering FAQ:
  http://moin.delph-in.net/GrammarEngineeringFaq

- Feature Geometry FAQ:
  http://moin.delph-in.net/GeFaqFeatureGeometry (see also the cheat sheet)
Installation

- Install subversion
  sudo apt install subversion
- Install logon (see LogonInstallation page)
  svn checkout http://svn.emmtee.net/trunk logon
- Install Emacs
  sudo apt install emacs
- Install git
  sudo apt install git
- Install JACY
  git clone https://github.com/delph-in/jacy.git
- Install ACE
  http://sweaglesw.org/linguistics/ace/
The current state: grammar size

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rules</td>
<td>27</td>
<td>50</td>
<td>51</td>
<td>54</td>
<td>47</td>
<td>81</td>
<td>86</td>
<td>137</td>
</tr>
<tr>
<td>Lexemes</td>
<td>3,399</td>
<td>5,369</td>
<td>5,681</td>
<td>5,147</td>
<td>35,220</td>
<td>30,898</td>
<td>56,944</td>
<td>56,914</td>
</tr>
<tr>
<td>Types</td>
<td>1,246</td>
<td>1,709</td>
<td>1,736</td>
<td>1,889</td>
<td>2,204</td>
<td>2,185</td>
<td>2,324</td>
<td>2,473</td>
</tr>
</tbody>
</table>

**Table**: Change in grammar size over time
Covered phenomena

- Verbs and adjectives
  - Inflectional and derivational rules
  - Auxiliary constructions
  - Passive constructions
  - Causative

- Nominal structures
  - Names and named entities
  - Pronouns (demonstrative, locative, personal, reflexive)
  - Nominalizers
  - Temporal nouns
  - Noun modification (relative clause)
  - Numeral classifiers

- Particles
- Adverbs
- Interrogatives
- Demonstratives
- Honorifics
A **test suite** is a curated collection of test items (sometimes including both grammatical and ungrammatical examples) meant to test specific properties of a grammar.

- **‘mrs’**: a small set of sentences, originally in English, that are meant to cover some of the basic semantic phenomena (argument structure, quantification, negation, modification etc.)
  
  [http://moin.delph-in.net/MatrixMrsTestSuite](http://moin.delph-in.net/MatrixMrsTestSuite)

- **‘vanilla’**: a collection of phenomena that are specific to Japanese

- etc.

<table>
<thead>
<tr>
<th>Type</th>
<th>Test Suite</th>
<th># Sents</th>
<th># Sents</th>
<th>Cover (%)</th>
<th># Sents</th>
<th>Cover (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional</td>
<td>mrs</td>
<td>135</td>
<td>126</td>
<td>93</td>
<td>127</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>vanilla</td>
<td>120</td>
<td>105</td>
<td>87</td>
<td>105</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>kinou1</td>
<td>1500</td>
<td>1321</td>
<td>88</td>
<td>1328</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>kinou2</td>
<td>1099</td>
<td>918</td>
<td>83</td>
<td>940</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>kinou3</td>
<td>1116</td>
<td>866</td>
<td>77</td>
<td>883</td>
<td>79</td>
</tr>
<tr>
<td>Natural</td>
<td>tanaka/tc-003</td>
<td>1500</td>
<td>1145</td>
<td>76</td>
<td>1172</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>tanaka/tc-004</td>
<td>1500</td>
<td>1136</td>
<td>75</td>
<td>1173</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>tanaka/tc-005</td>
<td>1500</td>
<td>1114</td>
<td>74</td>
<td>1145</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>haikingu</td>
<td>104</td>
<td>34</td>
<td>32</td>
<td>66</td>
<td>63</td>
</tr>
</tbody>
</table>
The Hinoki Treebank

- The Lexeed corpus
  - at Nippon Telegraph and Telephone Corporation (NTT)
  - 53,600 dictionary definition sentences and 36,000 example sentences

- The Tanaka corpus
  - at the Japanese National Institute of Information and Communications Technologies (NICT)
  - 15,000 example sentences

**Table:** Hinoki manual annotation result

<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Single Good Tree</td>
<td>7,809</td>
<td>52.1</td>
</tr>
<tr>
<td>Multiple Good Trees</td>
<td>679</td>
<td>4.5</td>
</tr>
<tr>
<td>Bad No Good Trees</td>
<td>1,604</td>
<td>10.7</td>
</tr>
<tr>
<td>No Parse Found</td>
<td>2,826</td>
<td>18.8</td>
</tr>
<tr>
<td>Resource Limitation</td>
<td>2,082</td>
<td>14.0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>15,000</td>
<td>100</td>
</tr>
</tbody>
</table>
JACY: a Japanese open-source HPSG

- JACY is an open-source HPSG grammar for Japanese (MIT license)
- probably the most distributed grammar development, developed by researchers in different continents (unlike ERG)
- JACY homepage: 
  http://moin.delph-in.net/JacyTop
- Grammar sources (MIT license): 
  https://github.com/delph-in/jacy
- On-line documentation, linguistic type database (LTDB): 
  http://compling.hss.ntu.edu.sg/ltdb/Jacy_1301/
- Demo page: 
  http://delph-in.github.io/delphin-viz/demo 
  http://chimpanzee.ling.washington.edu/demophin/jacy/
- DELPH-IN mailing list to ask questions 
  https://delphinqa.ling.washington.edu/
Some Japanese phenomena in JACY

- Argument scrambling and omission
- `-reru` / `-rareru` verbal endings
- ...
Verbal arguments scramble

Argument order is free, but arguments can not appear after the verb

(1) フランシス が 田中 に ボール を 渡す
Furanshisu ga Tanaka ni bo-ru wo watasu
"Francis hands Tanaka a ball"

(2) 田中 に フランシス が ボール を 渡す
Tanaka ni Furanshisu ga bo-ru wo watasu

(3) ボール を 田中 に フランシス が 渡す
bo-ru wo tanaka ni Furanshisu ga watasu

(4) *フランシス が 渡す 田中 に ボール を
Furanshisu ga watasu Tanaka ni bo-ru wo

Verbal arguments omission

Verbal arguments are frequently omitted even if it is the subject

(5) フランシス が ボール を 渡す
Furanshisu ga bo-ru wo watasu
Francis NOM ball ACC hand
“Francis hands a ball”

(6) 田中 に フランシス が 渡す
Tanaka ni Furanshisu ga watasu
Tanaka DAT Francis NOM hand
“Francis hands to Tanaka”

(7) 田中 に ボール を 渡す
Tanaka ni bo-ru wo watasu
Tanaka DAT ball ACC hand
“Hand Tanaka a ball”
The verbal endings れる (reru) and られる (rareru) can be used for:

- passive
  - simple
  - adversative
- honorification
- potential
Indicative vs Simple passive

Simple passive is available for transitive/ditransitive verbs and promotes an object to the subject.

(10) 田中 が ご飯 を 食べ た
Tanaka ga gohan wo tabe ta
Tanaka NOM gohan ACC eat PAST
“Tanaka ate the rice”

(11) ご飯 が 田中 に 食べ られ た
gohan ga Tanaka ni tabe rare ta
Tanaka NOM gohan DAT eat PASS PAST
“the rice was eaten by Tanaka”
(2) Adversative passive

The passive forms of intransitive verbs and transitive verbs and almost always indicates the event is unfavorable for the subject

(12) 子供　が　親　に　死なれ　た
kodomo ga oya ni shina re ta
child NOM parent DAT die PASS PAST
passive expression for “the child lost his parent”

(13) フランシス　が　ご飯　を　田中　に　食べられ　た
Furanshisu ga gohan wo Tanaka ni tabe rare ta
Francis NOM gohan ACC Tanaka DAT eat PASS PAST
“Francis’s rice was eaten by Tanaka”
(3) Honorification

(14) 先生 が ご飯 を 食べ られ た
sensei ga gohan wo tabe rarer ta
teacher NOM rice ACC eat HON PAST
“The teacher ate the rice”
(15) 彼 が  ドリアン を 食べ られる
kare ga  dorian wo  tabe rareru
3SG NOM durian ACC eat POT
“He can eat durian”
A *treebank* is a syntactically annotated corpus of sentences with parse trees

Full Forest Treebanker (FFTB) (Packard, 2014): a tool for treebanking with DELPH-IN grammars that allows the users to select manually a tree from the “full forest” of possible trees without listing or specifying all analyses in the parsing stage and store it into database for statistical ranking of candidate parses, transfers, and translations

- grammar-based corpus annotation
- test-suite format:
  - [http://compling.hss.ntu.edu.sg/courses/hg7021/testsuites.html](http://compling.hss.ntu.edu.sg/courses/hg7021/testsuites.html)
- DEMO: FFTB with ‘mrs’ test-suite
Japanese-English machine translation

- Semantic-transfer-based Japanese-to-English machine translation system, built using the LOGON infrastructure
  https://github.com/delph-in/JaEn
- The system consists of the two HPSG grammars and one transfer grammar
  ▶ JACY used to parse the Japanese input
  ▶ ERG used for the generation of the English output
  ▶ transfer grammar which transfers the MRS representation produced by JACY into an MRS representation that ERG can generate from
Figure: Architecture of the JaEn MT system.
(16) 雨 が 降る
ame ga furu
rain NOM fall
“It rains.”

(17) 雨 が 降った
ame ga furu ta
rain NOM fall PAST
“It rained.”

(18) 日本 の ケーキ が あった
nihon no keeki ga ar ta
Japan ADN cake NOM exist PAST
“There was/were Japanese cake(s).”
Conclusions and Future Work

**JACY**
- a broad-coverage Japanese computational grammar
- uses the framework of Head-driven Phrase Structure Grammar (HPSG) with Minimal Recursion Semantics (MRS)
- encodes precise morphological, syntactic, semantic, and pragmatic information in feature structures
- has been developed within many different research projects
- is being developed in a multilingual context, where much value is placed on parallel and consistent semantic representations

**Future Work**
- will be further adapted to other domains: the newspapers (including the grammar of headline text) and general text such as Wikipedia
- revise analyses
- integration with Japanese Wordnet
- update the treebank
Acknowledgments

- Some slides borrow from Melanie Siegel’s presentation slides (http://www.delph-in.net/jacy/jacy.pdf)
(19) a. ありがとうございます
arigatou gozai masu
“Thank you”

b. UTT
    IDIOM
ありがとうございます

\[
\begin{array}{|c|}
\hline
mrs \\
TOP & 0 h \\
INDEX & 2 i \\
\hline
\end{array}
\]
\[
\begin{array}{|c|c|}
\hline
\text{RELs} & \left[ \begin{array}{l}
\text{discourse}_x_{\text{rel}} \\
\text{LBL} & 4 h \\
\text{ARG0} & 5 e \\
\text{L-HNDL} & 6 h \\
\text{R-HNDL} & 7 h \\
\hline
\end{array} \right], \\
\text{HCONS} & \left[ \begin{array}{l}
\text{qeq} \\
\text{HARG} & 0 h \\
\text{LARG} & 1 h \\
\hline
\end{array} \right] \\
\hline
\end{array}
\]
\[
\left[ \begin{array}{c}
\text{discourse}_x \quad \text{doumoarigatougozaimasu}_x_{\text{rel}} \\
\text{LBL} & 6 h \\
\text{ARG0} & 8 e \\
\hline
\end{array} \right]
\]
\[
\begin{array}{|c|}
\hline
\text{L-HNDL/HEQ} \\
\hline
\end{array}
\]

Moeljadi and Kuribayashi (LMS, NTU)
2 July 2018 32 / 35


Woodley Packard. ACE, the Answer Constraint Engine. 2013. URL: http://sweaglesw.org/linguistics/ace/ (visited on 04/21/2015).

