Parser Evaluation over Local and Non-Local Deep Dependencies in a Large Corpus

Emily M. Bender♠, Dan Flickinger♥, Stephan Oepen♣, and Yi Zhang♦

♠Department of Linguistics, University of Washington
♥CSLI, Stanford University
♣Department of Informatics, Universitetet i Oslo
♦Deutsches Forschungszentrum für Künstliche Intelligenz
Motivation — Related Work

(*To what degree*) Is syntactic analysis a solved problem?

PTB$^{23}$ $F_1$: 0.84 (Magerman, 1994) $\rightarrow$ 0.92 (McClosky et al., 2006)
Motivation — Related Work

(To what degree) Is syntactic analysis a solved problem?

PTB$^{23}$ $F_1$: 0.84 (Magerman, 1994) → 0.92 (McClosky et al., 2006)

Rimell, Clark, & Steedman (2009) [RCS]

- single aggregate score mis-leading (sentence accuracy $\sim$10–25%);
- great variation across different phenomena and dependency types;
- analysis of non-local dependency recovery in five syntactic parsers;
- non-trivial frequency (in PTB); indicative of ‘full’ syntactic analysis;
  → very poor recovery of seven phenomena: average recall $\sim$25–54%. 

Parser Evaluation over Local and Non-Local Dependencies (2)
Motivation — Related Work

(To what degree) Is syntactic analysis a solved problem?

PTB$^{23}$ $F_1$: 0.84 (Magerman, 1994) → 0.92 (McClosky et al., 2006)

Rimell, Clark, & Steedman (2009) [RCS]

- single aggregate score misleading (sentence accuracy $\sim$10–25%);
- great variation across different phenomena and dependency types;
- analysis of non-local dependency recovery in five syntactic parsers;
- non-trivial frequency (in PTB); indicative of 'full' syntactic analysis;
  → very poor recovery of seven phenomena: average recall $\sim$25–54%.

- relatively narrow phenomenon range;
- no intra-phenomenon differentiation;
- not included a classic 'deep' parser;
- manual judgment of parser outputs.
Birds-Eye View on the Sequence of Events

(1) Select ten ‘hard’ syntactic phenomena, local and non-local;
(2) find 100 ‘suitable’ sentences per phenomenon in Wikipedia;
(3) dual-annotate and reconcile for ‘relevant’ dependencies;
(4) run seven off-the-shelf parsers on this data (the strings);
(5) design parser-specific patterns for automated evaluation;
(6) release annotated corpus, evaluation scripts, and results.
Phenomena (1/10): Bare Relatives (Non-Local)

A classic example Schumacher provides is that of education.

This is the second time in a row Australia lost their home series.

The maximum points a single team can earn is 775.
Phenomena (2/10): Tough Adjectives (Non-Local)

Original copies are very hard to find.
Phenomena (2/10): Tough Adjectives (Non-Local)

Original copies are very hard to find.

Phenomena (3/10): Right Node Raising (Non-Local)

He also played for and managed Kilmarnock ...
Phenomena (4/10): It Expletives (Non-Dependency)

Crew negligence is blamed, and it is suggested that the flight crew were drunk.
Phenomena (4/10): It Expletives (Non-Dependency)

Crew negligence is blamed, and it is suggested that the flight crew were drunk.

Phenomena (5/10): Verb–Particles (Non-Dependency)

He once threw out two baserunners at home in the same inning.
Phenomena (6/10): Our Very Own ‘NED’ (Local)

Light colored glazes also have softening effects ...
Phenomena (6/10): Our Very Own ‘NED’ (Local)

Light colored glazes also have softening effects ...

Phenomena (7/10): Absolutives (Local)

The format consisted of 12 games, each team facing the other teams twice.
Phenomena (8/10): Verbal Gerunds (Local)

*It is like coining the Nirvana into dynamos.*
Phenomena (8/10): Verbal Gerunds (Local)

\[
\text{It is like coining the Nirvana into dynamos.}
\]

Phenomena (9/10): Interspersed Adjuncts (Local)

\[
\text{The story shows, through flashbacks, the different histories of the characters.}
\]
Phenomena (8/10): Verbal Gerunds (Local)

\[ \text{It is like coining the Nirvana into dynamos.} \]

Phenomena (9/10): Interspersed Adjuncts (Local)

\[ \text{The story shows, through flashbacks, the different histories of the characters.} \]

Phenomena (10/10): Controlled Arguments (Local)

\[ \text{Alfred ... continued to paint full time.} \]
Data Preparation

Selection from English Wikipedia (‘WikiWoods’)

- Parsed with the ERG (Flickinger et al., 2010): 900 million tokens;
- indexed by HPSG constructions; random selection of candidates;
- dual-vetted: skip false positive, overly basic, and all too complex.
Data Preparation

Selection from English Wikipedia (‘WikiWoods’)

- Parsed with the ERG (Flickinger et al., 2010): 900 million tokens;
- indexed by HPSG constructions; random selection of candidates;
- dual-vetted: skip false positive, overly basic, and all too complex.

→ one thousand sentences (for our ten phenomena).
Data Preparation

Selection from English Wikipedia (‘WikiWoods’)

- Parsed with the ERG (Flickinger et al., 2010): 900 million tokens;
- indexed by HPSG constructions; random selection of candidates;
- dual-vetted: skip false positive, overly basic, and all too complex.

→ one thousand sentences (for our ten phenomena).

Annotation and Reconciliation

- Specify target scheme; parallel annotation by two expert linguists;
- initial agreement: 79% (full sentences); all mismatches reconciled;
- employ disjunctive heads or dependents for plausible alternatives.
Data Preparation

Selection from English Wikipedia (‘WikiWoods’)
- Parsed with the ERG (Flickinger et al., 2010): 900 million tokens;
- indexed by HPSG constructions; random selection of candidates;
- dual-vetted: skip false positive, overly trivial, and overly complex.

→ one thousand sentences (for our ten phenomena).

Annotation and Reconciliation
- Specify target scheme; parallel annotation by two expert linguists;
- initial agreement: 79% (full sentences); all mismatches reconciled;
- employ disjunctive heads or dependents for plausible alternatives.

coordination of heads or dependents multiplied out;
→ 2127 dependency triples (253 negative; 580 disjunctive).
The Act having been passed in that year, Jessop withdrew, and Whitworth carried on with the assistance of his son.

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Type</th>
<th>Dependency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1011079100200</td>
<td>ABSOL</td>
<td>having</td>
</tr>
<tr>
<td>1011079100200</td>
<td>ABSOL</td>
<td>withdrew MOD having</td>
</tr>
<tr>
<td>1011079100200</td>
<td>ABSOL</td>
<td>carried+on MOD having</td>
</tr>
</tbody>
</table>
The Act having been passed in that year, Jessop withdrew, and Whitworth carried on with the assistance of his son.

Disjunctive heads or dependents for:
- auxiliaries and (some) modals;
- complementizers (e.g. that);
- multi-word proper names; and
- genuine attachment ambiguity.

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Type</th>
<th>Dependency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1011079100200</td>
<td>ABSOL</td>
<td>having|been|passed ARG act</td>
</tr>
<tr>
<td>1011079100200</td>
<td>ABSOL</td>
<td>withdrew MOD having|been|passed</td>
</tr>
<tr>
<td>1011079100200</td>
<td>ABSOL</td>
<td>carried+on MOD having|been|passed</td>
</tr>
</tbody>
</table>
### (Select) Phenomena Summaries and Locality

<table>
<thead>
<tr>
<th>Type</th>
<th>Head</th>
<th>Dependent</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BAREREL</strong></td>
<td><code>gapped predicate</code></td>
<td>`A</td>
<td>M` modified noun</td>
</tr>
<tr>
<td></td>
<td><code>modified noun</code></td>
<td><code>M</code> head of relative</td>
<td>3.3 (8)</td>
</tr>
<tr>
<td><strong>TOUGH</strong></td>
<td><code>tough adjective</code></td>
<td><code>A</code> VP complement</td>
<td>1.7 (5)</td>
</tr>
<tr>
<td></td>
<td><code>gapped predicate</code></td>
<td><code>A</code> subject of adjective</td>
<td>6.4 (21)</td>
</tr>
<tr>
<td><strong>RNR</strong></td>
<td><code>right conjunct</code></td>
<td><code>A</code> shared noun</td>
<td>2.8 (9)</td>
</tr>
<tr>
<td></td>
<td><code>left conjunct</code></td>
<td><code>A</code> shared noun</td>
<td>6.1 (12)</td>
</tr>
<tr>
<td><strong>ITEXPL</strong></td>
<td>expletive predicate</td>
<td><code>¬A</code> <code>it</code></td>
<td>1.2 (3)</td>
</tr>
<tr>
<td><strong>ABSOL</strong></td>
<td><code>absolutive predicate</code></td>
<td><code>A</code> subject of absolutive</td>
<td>1.7 (12)</td>
</tr>
<tr>
<td></td>
<td><code>head of main clause</code></td>
<td><code>M</code> absolutive predicate</td>
<td>9.8 (26)</td>
</tr>
<tr>
<td><strong>ARGADJ</strong></td>
<td><code>head verb</code></td>
<td><code>M</code> interspersed adjunct</td>
<td>1.2 (7)</td>
</tr>
<tr>
<td></td>
<td><code>head verb</code></td>
<td><code>A</code> displaced complement</td>
<td>5.9 (26)</td>
</tr>
<tr>
<td><strong>CONTROL</strong></td>
<td><code>‘upstairs’ verb</code></td>
<td><code>A</code> ‘downstairs’ verb</td>
<td>2.4 (23)</td>
</tr>
<tr>
<td></td>
<td><code>‘downstairs’ verb</code></td>
<td><code>A</code> shared complement</td>
<td>4.8 (17)</td>
</tr>
</tbody>
</table>

Parser Evaluation over Local and Non-Local Dependencies (11)
## (Select) Phenomena Summaries and Locality

<table>
<thead>
<tr>
<th>Type</th>
<th>Head</th>
<th>Dependent</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BAREREL</strong></td>
<td>gapped predicate</td>
<td>modified noun</td>
<td>3.0 (8)</td>
</tr>
<tr>
<td></td>
<td>modified noun</td>
<td>head of relative</td>
<td>3.3 (8)</td>
</tr>
<tr>
<td><strong>TOUGH</strong></td>
<td><em>tough</em> adjective</td>
<td>VP complement</td>
<td>(~0.04%)</td>
</tr>
<tr>
<td></td>
<td>gapped predicate</td>
<td>subject of adjective</td>
<td></td>
</tr>
<tr>
<td><strong>RNR</strong></td>
<td>right conjunct</td>
<td>shared noun</td>
<td>2.8 (9)</td>
</tr>
<tr>
<td></td>
<td>left conjunct</td>
<td>shared noun</td>
<td>6.1 (12)</td>
</tr>
<tr>
<td><strong>ITEXPL</strong></td>
<td>expletive predicate</td>
<td><em>it</em></td>
<td>1.2 (3)</td>
</tr>
<tr>
<td><strong>ABSOL</strong></td>
<td>absolutive predicate</td>
<td>subject of absolutive</td>
<td>1.7 (12)</td>
</tr>
<tr>
<td></td>
<td>head of main clause</td>
<td>absolutive predicate</td>
<td>9.8 (26)</td>
</tr>
<tr>
<td><strong>ARGADJ</strong></td>
<td>head verb</td>
<td>interspersed adjunct</td>
<td>1.2 (7)</td>
</tr>
<tr>
<td></td>
<td>head verb</td>
<td>displaced complement</td>
<td>5.9 (26)</td>
</tr>
<tr>
<td><strong>CONTROL</strong></td>
<td>‘upstairs’ verb</td>
<td>‘downstairs’ verb</td>
<td>(\sim3.1%)</td>
</tr>
<tr>
<td></td>
<td>‘downstairs’ verb</td>
<td>shared complement</td>
<td></td>
</tr>
</tbody>
</table>
Participating Parsers

Trained ‘Directly’ on the (WSJ Portion of the) PTB

- **Stanford** (Klein & Manning, 2003) factored model; GR output;
- **C&J** (Charniak & Johnson, 2005) Stanford GR post-processor;
- **MST** (McDonald et al., 2005) second-order projective model.

Trained Indirectly on the (WSJ Portion of the) PTB

- **Enju** (Miyao et al., 2004) HPSG; predicate–argument outputs;
- **C&C** (Clark & Curran, 2007) CCG; grammatical relation outputs.

(Partly) Analytically Engineered

- **RASP** (Briscoe et al., 2006) PoS ‘tag sequence grammar’; GRs;
- **XLE** (Kaplan et al., 2004) hand-built LFG and lexicon; f-structures.
Operationalizing the Evaluation Process

The Act having been passed in that year, Jessop withdrew, and Whitworth carried on with the assistance of his son.

(xmod _ Act_1 passed_4) (ncsubj passed_4 Act_1 _)
(ncmod _ withdrew,_9 Jessop_8) (dobj year,_7 withdrew,_9)
(ncmod _ carried_12 on_13) (ncsubj carried_12 Whitworth_11 _)

Absolutives (ABSOL)

ARG /
(nsubj \W*{W1}\W*_\d+ \W*{W2}\W*_\d+ _)/
(nmod _ \W*{W2}\W*_\d+ \W*{W1}\W*_\d+\))/

MOD /
((c|nc|x)mod _ \W*{W1}\W*_\d+ \W*{W2}\W*_\d+\))/

• Phenomenon- and parser-specific patterns; avoid lexical information;
• annotation instantiates \{W1\} and \{W2\}; allow (non-contentful) variation.
Operationalizing the Evaluation Process

The Act having been passed in that year, Jessop withdrew, and Whitworth carried on with the assistance of his son.

Absolutives (ABSOL)

ARG

\[/(ncsubj \ W*{W1}\ W* \_d+ \ W*{W2}\ W* \_d+ \_)/\]

\[/(ncmod _ \ W*{W2}\ W* \_d+ \ W*{W1}\ W* \_d+)/\]

MOD

\[/((c|nc|x)mod _ \ W*{W1}\ W* \_d+ \ W*{W2}\ W* \_d+)/\]

- Phenomenon- and parser-specific patterns; avoid lexical information;
- annotation instantiates \{W1\} and \{W2\}; allow (non-contentful) variation.
Operationalizing the Evaluation Process

The Act having been passed in that year, Jessop withdrew, and Whitworth carried on with the assistance of his son.

(xmod _ Act_1 passed_4) (ncsubj passed_4 Act_1 _)
(ncmod _ withdrew,_9 Jessop_8) (dobj year,_7 withdrew,_9)
(ncmod _ carried_12 on_13) (ncsubj carried_12 Whitworth_11 _)

Absolutives (ABSOL)

ARG /
\( (\text{ncsubj } W^*\{W1}\ W^*\_d+ \ W^*\{W2}\ W^*_\_d+ \_)/
\( (\text{ncmod } _ W^*\{W2}\ W^*\_d+ \ W^*\{W1}\ W^*_\_d+\)/

MOD /
\( ((c|nc|x)mod _ W^*\{W1}\ W^*_\_d+ \ W^*\{W2}\ W^*_\_d+)/

In some regards akin to ‘interpretation’ by a back-end application;
→ 364 patterns (for 19 dependencies and six output formats).
Results Summary: Per-Dependency Recall

Parser Evaluation over Local and Non-Local Dependencies (14)

EMNLP — 28-JUL-11 (oe@ifi.uio.no)
Results Summary: Per-Dependency Recall

Is There Good News or Bad News (or Both)?

Parser Evaluation over Local and Non-Local Dependencies (14)
Results Summary: Per-Dependency Recall

Good Recovery of Some Phenomena: VGER, VPART, CONTROL.
Results Summary: Per-Dependency Recall

Predictable: ITEXPL requires lexical knowledge (not in ‘PTB’).
Parser Evaluation over Local and Non-Local Dependencies (14)
Cross-Phenomenon and -Dependency Variation (MST)

Great Variation Within Many Phenomena for Most Parsers.

Parser Evaluation over Local and Non-Local Dependencies (15)
By Comparison: Grammar-Based Parsing (XLE)

With Some Exceptions, Comparatively Even Performance.
Results Summary: A Somewhat Grim Point of View

When Requiring Both Dependencies for Success, Only Two Parsers Exceed 50% for Five Phenomena; All Systems Below 50% for Three Phenomena.
Results Summary: A Somewhat Grim Point of View

No System Above 33% on RNR (Average 44% in [RCS]).
Results Summary: Pointwise Parser Comparison

Parser Evaluation over Local and Non-Local Dependencies (18)

C&J vs. Stanford: Average 56% vs. 52%.
Discussion — Outlook

Some High-Level Observations

- Arguably, our dependencies (and more) play into ‘text understanding’;
- construction-specific evaluation yields in-depth, albeit partial picture;
- intra-phenomenon differentiation helps reveal incomplete analyses;
- automating pattern-based construction evaluation appears feasible;

Candidate Take-Home Lessons

- Search for better understanding of strong and weak points in parsers;
- work towards larger inventory of target dependencies and patterns;
  → linguistically richer and more diverse treebanks (or grammars) needed.
Discussion — Outlook

Some High-Level Observations

- Arguably, our dependencies (and more) play into ‘text understanding’;
- construction-specific evaluation yields in-depth, albeit partial picture;
- intra-phenomenon differentiation helps reveal incomplete analyses;
- automating pattern-based construction evaluation appears feasible;

Candidate Take-Home Lessons

- Search for better understanding of strong and weak points in parsers;
- work towards larger inventory of target dependencies and patterns;

Background and download: http://www.delph-in.net/ddec/
Parser Evaluation over Local and Non-Local Dependencies (20)