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) → 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 ∼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 ∼25–54%.

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 ~10–25%);
- great var ency types;
- analysis ctic parsers;
- non-trivi c analysis;

→ very poo ll ~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;

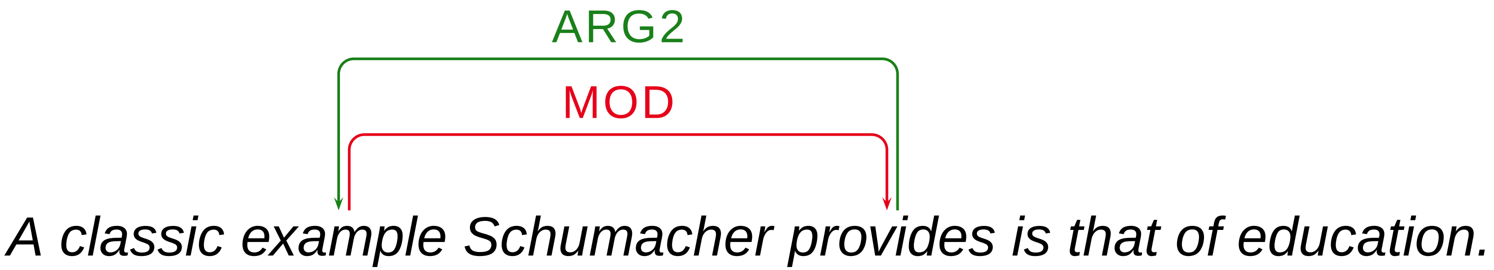
(3) dual-annotate and reconcile for ‘relevant’ dependencies;

(4) run seven off-the-shelf parsers on this data (the strings);

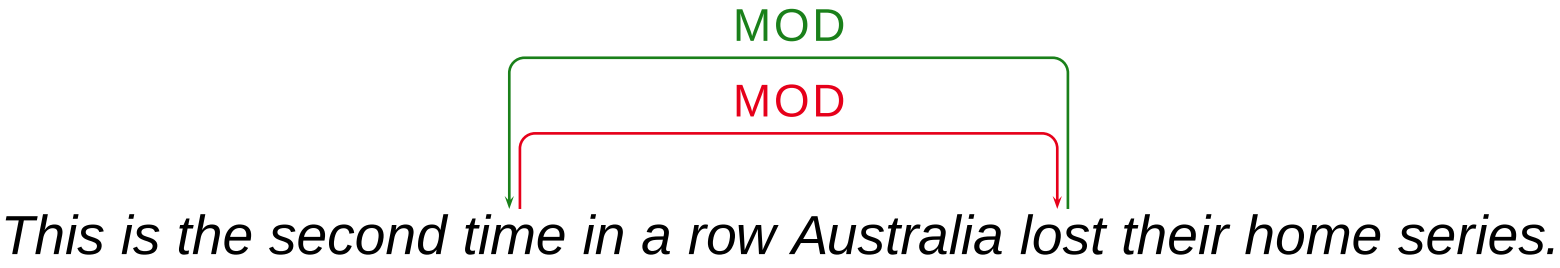
(5) design parser-specific patterns for automated evaluation;

(6) release annotated corpus, evaluation scripts, and results.

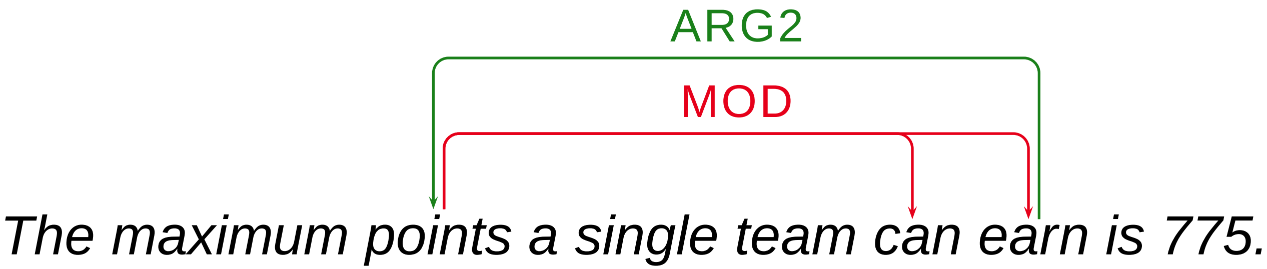
Phenomena (1/10): Bare Relatives (Non-Local)



A classic example Schumacher provides is that of education.

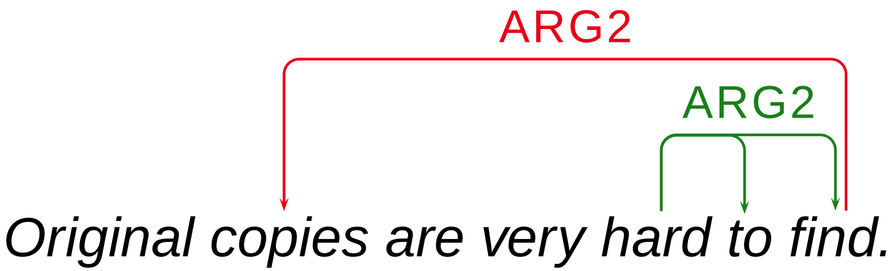


This is the second time in a row Australia lost their home series.



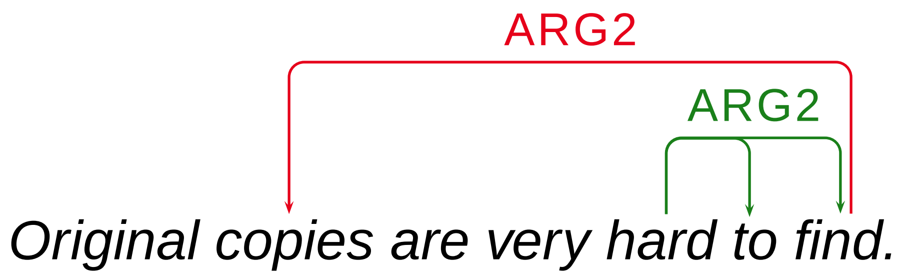
The maximum points a single team can earn is 775.

Phenomena (2/10): Tough Adjectives (Non-Local)



Original copies are very hard to find.

Phenomena (2/10): Tough Adjectives (Non-Local)

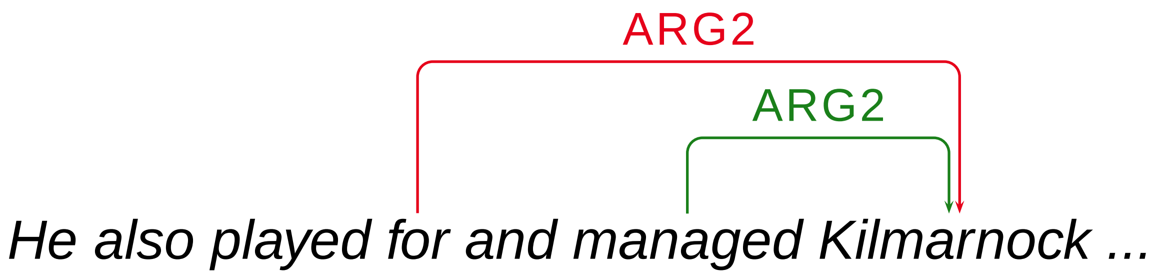


ARG2

ARG2

Original copies are very hard to find.

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

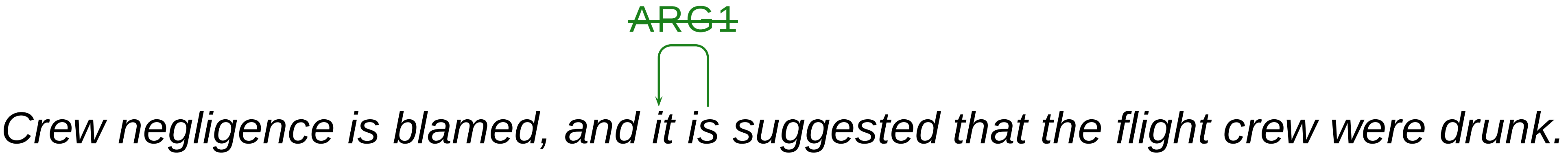


ARG2

ARG2

He also played for and managed Kilmarnock ...

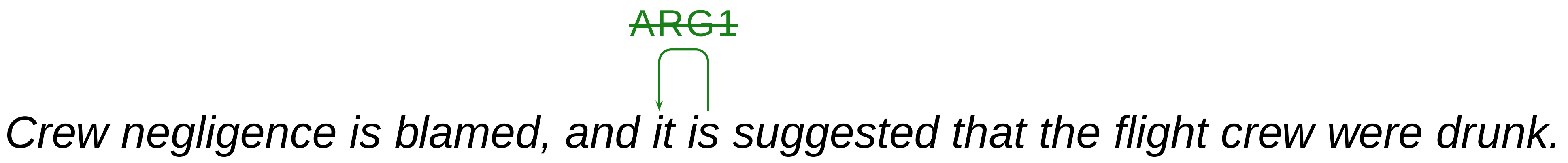
Phenomena (4/10): It Expletives (Non-Dependency)



~~ARG1~~

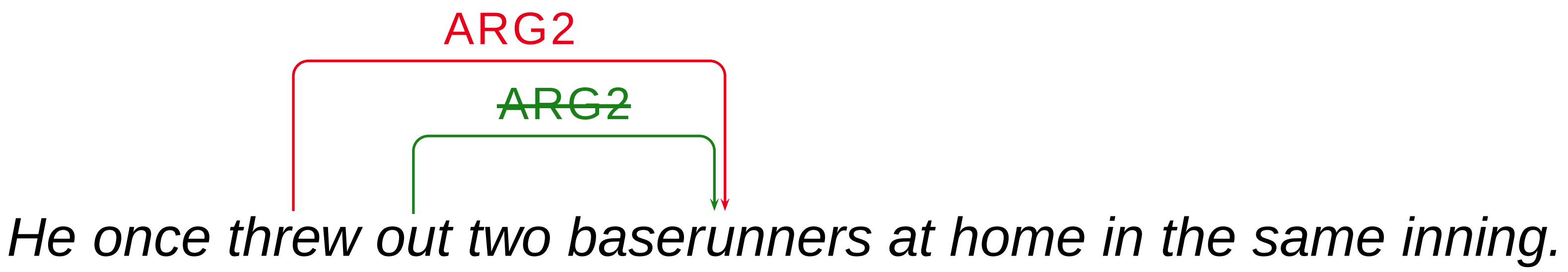
Crew negligence is blamed, and it is suggested that the flight crew were drunk.

Phenomena (4/10): It Expletives (Non-Dependency)



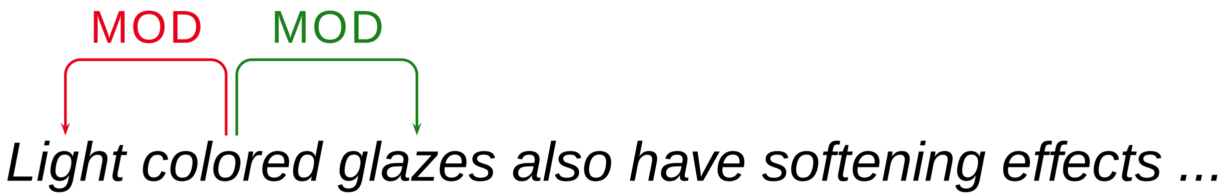
Crew negligence is blamed, and it is suggested that the flight crew were drunk.

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



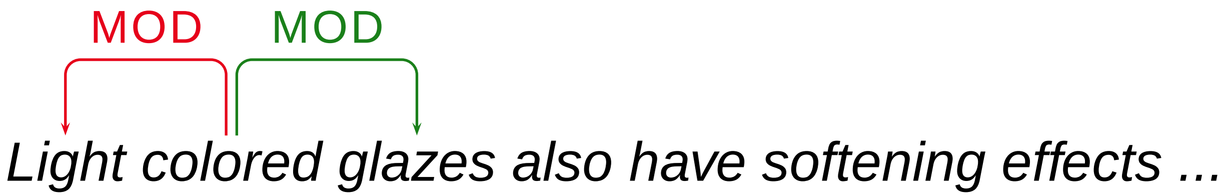
He once threw out two baserunners at home in the same inning.

Phenomena (6/10): Our Very Own ‘NED’ (Local)



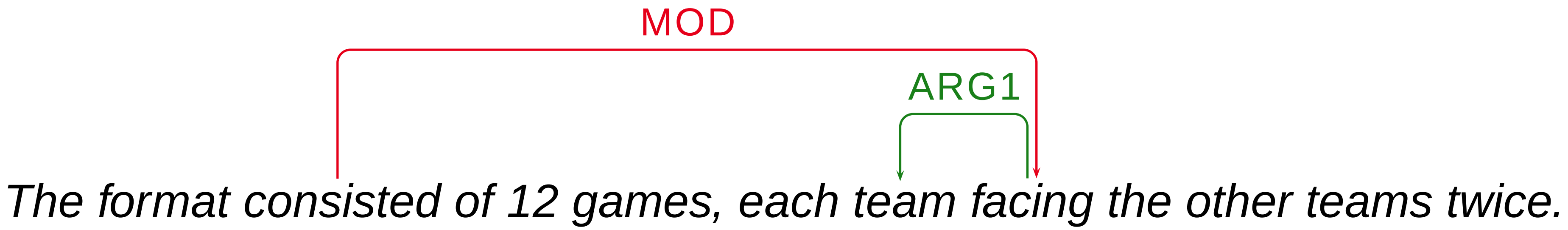
Light colored glazes also have softening effects ...

Phenomena (6/10): Our Very Own ‘NED’ (Local)



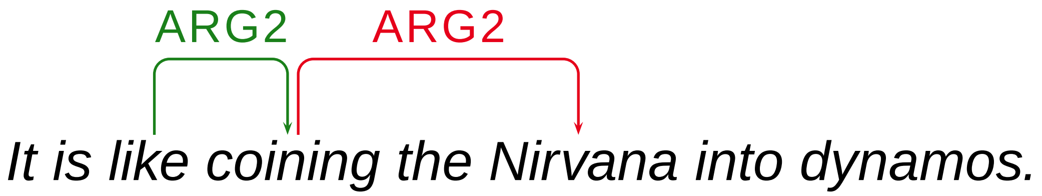
Light colored glazes also have softening effects ...

Phenomena (7/10): Absolutives (Local)



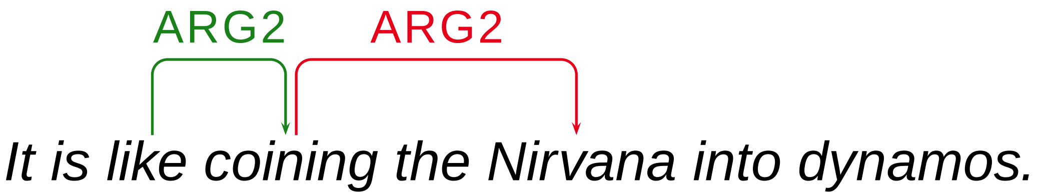
The format consisted of 12 games, each team facing the other teams twice.

Phenomena (8/10): Verbal Gerunds (Local)



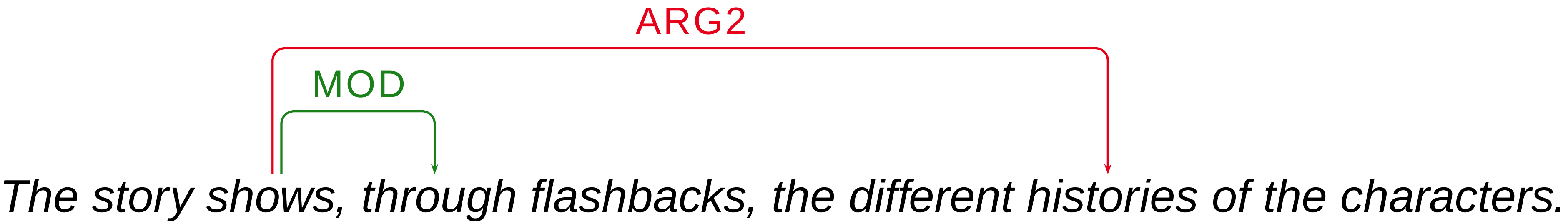
It is like coining the Nirvana into dynamos.

Phenomena (8/10): Verbal Gerunds (Local)


ARG2 ARG2

It is like coining the Nirvana into dynamos.

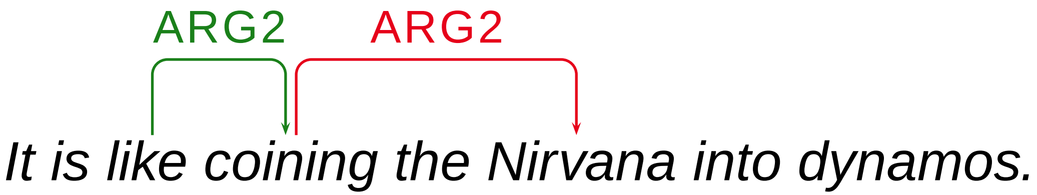
Phenomena (9/10): Interspersed Adjuncts (Local)


ARG2

MOD

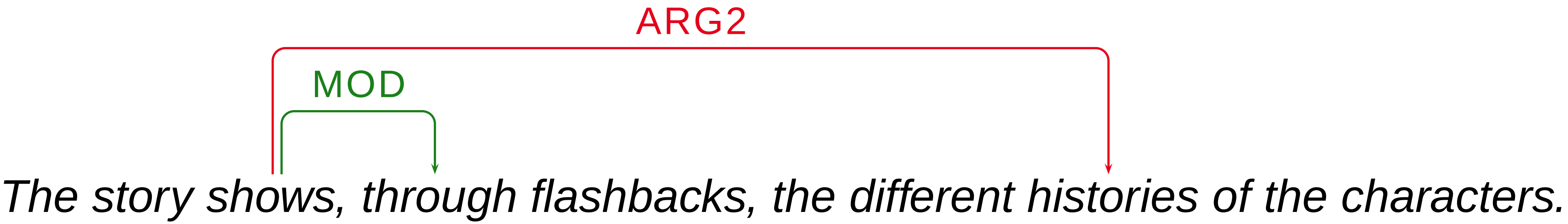
The story shows, through flashbacks, the different histories of the characters.

Phenomena (8/10): Verbal Gerunds (Local)



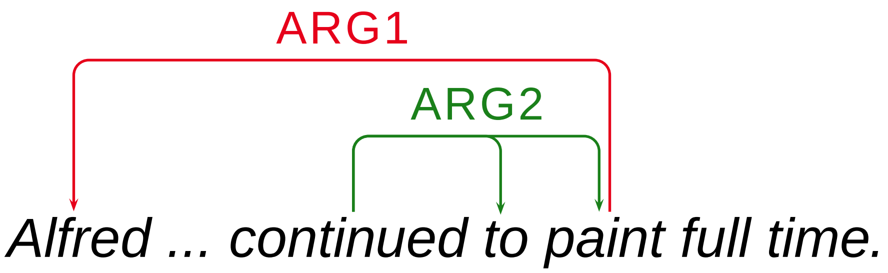
It is like coining the Nirvana into dynamos.

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



The story shows, through flashbacks, the different histories of the characters.

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



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;
- d plex.

→ 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;
- d plex.

→ *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;
- d plex.

→ *one thousand sentences (for our ten phenomena).*

Annotation and Reconciliation

- Specify target scheme; parallel annotation by two expert linguists;
- d;
- s.

coordination of heads or dependents multiplied out;

→ *2127 dependency triples (253 negative; 580 disjunctive).*

Example Annotations

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

<table>
<tr><th>Item ID</th><th>Type</th><th>Dependency</th></tr>
<tr><td>10110791002000</td><td>ABSOL</td><td>having|been|passed ARG act</td></tr>
<tr><td>10110791002000</td><td>ABSOL</td><td>withdrew MOD having|been|passed</td></tr>
<tr><td>10110791002000</td><td>ABSOL</td><td>carried+on MOD having|been|passed</td></tr>
</table>

Example Annotations

and *son.*

Disjunctive heads or dependents for:

auxiliaries and (some) modals;
complementizers (e.g. that);
multi-word proper names; and
genuine attachment ambiguity.

Item ID	Type	Dependency
1011079100200	ABSOL	having|been|passed ARG act
1011079100200	ABSOL	withdrew MOD having|been|passed
1011079100200	ABSOL	carried+on MOD having|been|passed

(Select) Phenomena Summaries and Locality

Type	Head		Dependent	Distance
BAREREL	gapped predicate	A|M	modified noun	3.0 (8)
	modified noun	M	head of relative	3.3 (8)
TOUGH	*tough* adjective	A	VP complement	1.7 (5)
	gapped predicate	A	subject of adjective	6.4 (21)
RNR	right conjunct	A	shared noun	2.8 (9)
	left conjunct	A	shared noun	6.1 (12)
ITEXPL	expletive predicate	¬A	*it*	1.2 (3)
ABSOL	absolutive predicate	A	subject of absolutive	1.7 (12)
	head of main clause	M	absolutive predicate	9.8 (26)
ARGADJ	head verb	M	interspersed adjunct	1.2 (7)
	head verb	A	displaced complement	5.9 (26)
CONTROL	‘upstairs’ verb	A	‘downstairs’ verb	2.4 (23)
	‘downstairs’ verb	A	shared complement	4.8 (17)

(Select) Phenomena Summaries and Locality

Type	Head		Dependent	Distance
BAREREL	gapped predicate	A|M	modified noun	3.0 (8)
	modified noun	M	head of relative	3.3 (8)
TOUGH	*tough* adjective	A	VP complement	~0.04 %
	gapped predicate	A	subject of adjective	
RNR	right conjunct	A	shared noun	2.8 (9)
	left conjunct	A	shared noun	6.1 (12)
ITEXPL	expletive predicate	¬A	*it*	1.2 (3)
ABSOL	absolutive predicate	A	subject of absolutive	1.7 (12)
	head of main clause	M	absolutive predicate	9.8 (26)
ARGADJ	head verb	M	interspersed adjunct	1.2 (7)
	head verb	A	displaced complement	5.9 (26)
CONTROL	‘upstairs’ verb	A	‘downstairs’ verb	~3.1 %
	‘downstairs’ verb	A	shared complement	

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	`/\(ncsubj \W*{W1}\W*_\d+ \W*{W2}\W*_\d+ _\)/`
	`/\(ncmod _ \W*{W2}\W*_\d+ \W*{W1}\W*_\d+\)/`
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)

<table>
<tr><td rowspan="2">ARG</td><td><code>/\(ncsubj \W*{W1}\W*_\d+ \W*{W2}\W*_\d+ _\)/</code></td></tr>
<tr><td><code>/\(ncmod _ \W*{W2}\W*_\d+ \W*{W1}\W*_\d+\)/</code></td></tr>
<tr><td>MOD</td><td><code>/\((c|nc|x)mod _ \W*{W1}\W*_\d+ \W*{W2}\W*_\d+\)/</code></td></tr>
</table>

- 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.

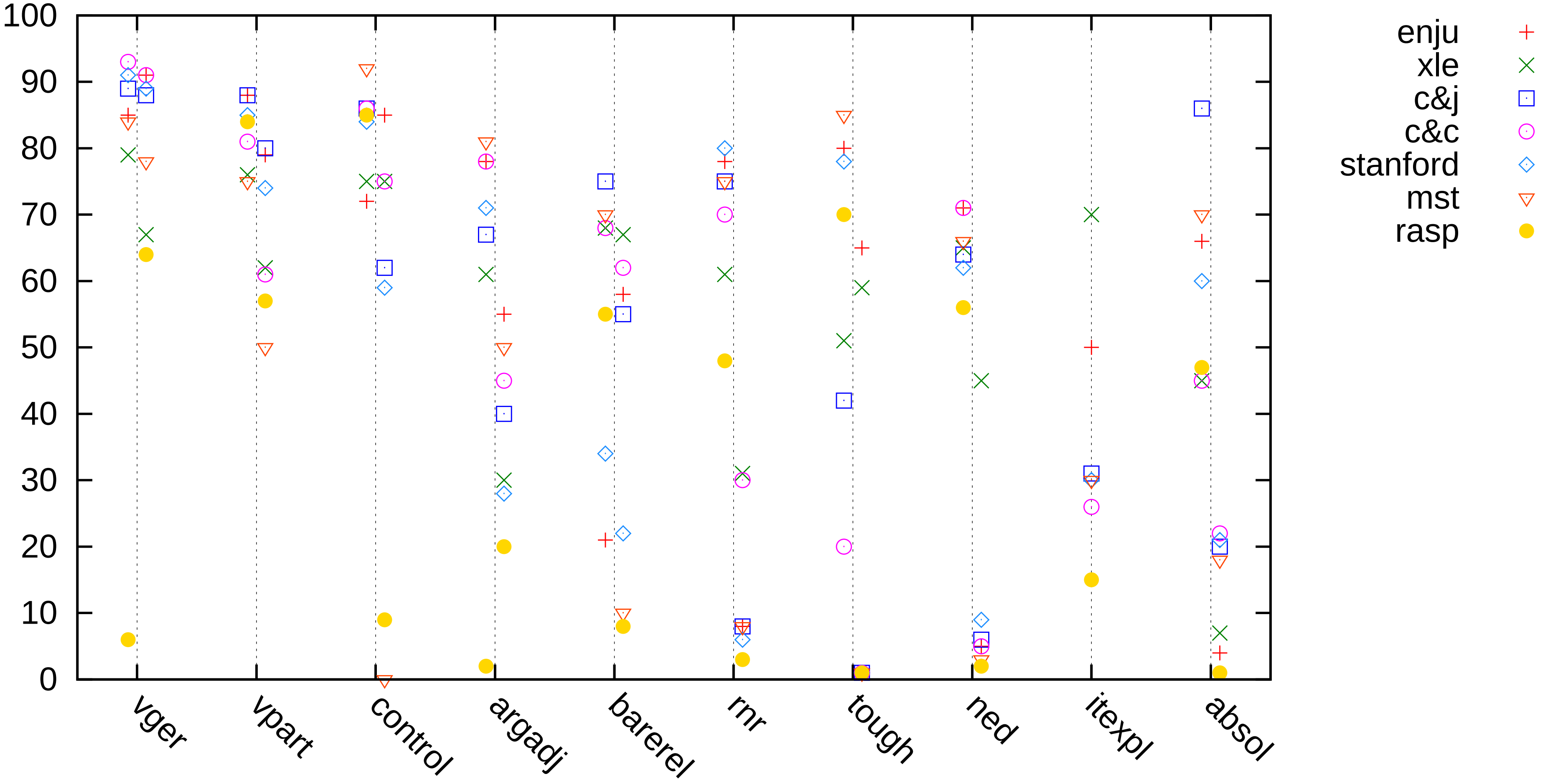
```
(xmod _ Act_1 passed_4)   (ncsubj passed_4 Act_1 _)
(ncmod _ withdrew,_9 Jessop_8)   (dobj year,_7 withdrew,_9)
(ncmod _ carried_12 on_13)   (ncsubj carried_12 Whitworth_11 _)
```

Absolutives (ABSOL)

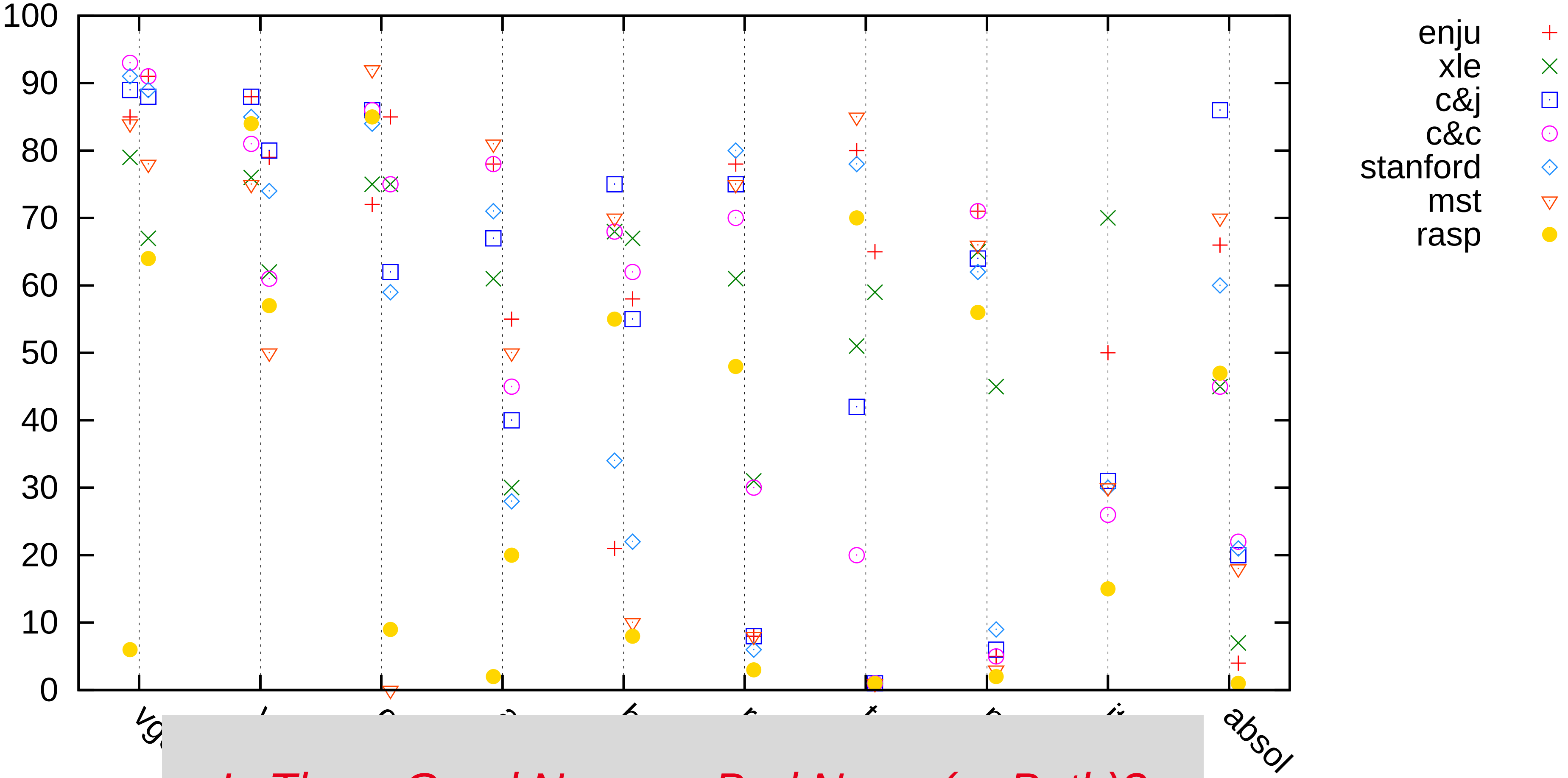
<table>
<tr><td>ARG</td><td><code>/\(ncsubj \W*{W1}\W*_\d+ \W*{W2}\W*_\d+ _\)/</code>
<code>/\(ncmod _ \W*{W2}\W*_\d+ \W*{W1}\W*_\d+\)/</code></td></tr>
<tr><td>MOD</td><td><code>/\((c|nc|x)mod _ \W*{W1}\W*_\d+ \W*{W2}\W*_\d+\)/</code></td></tr>
</table>

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

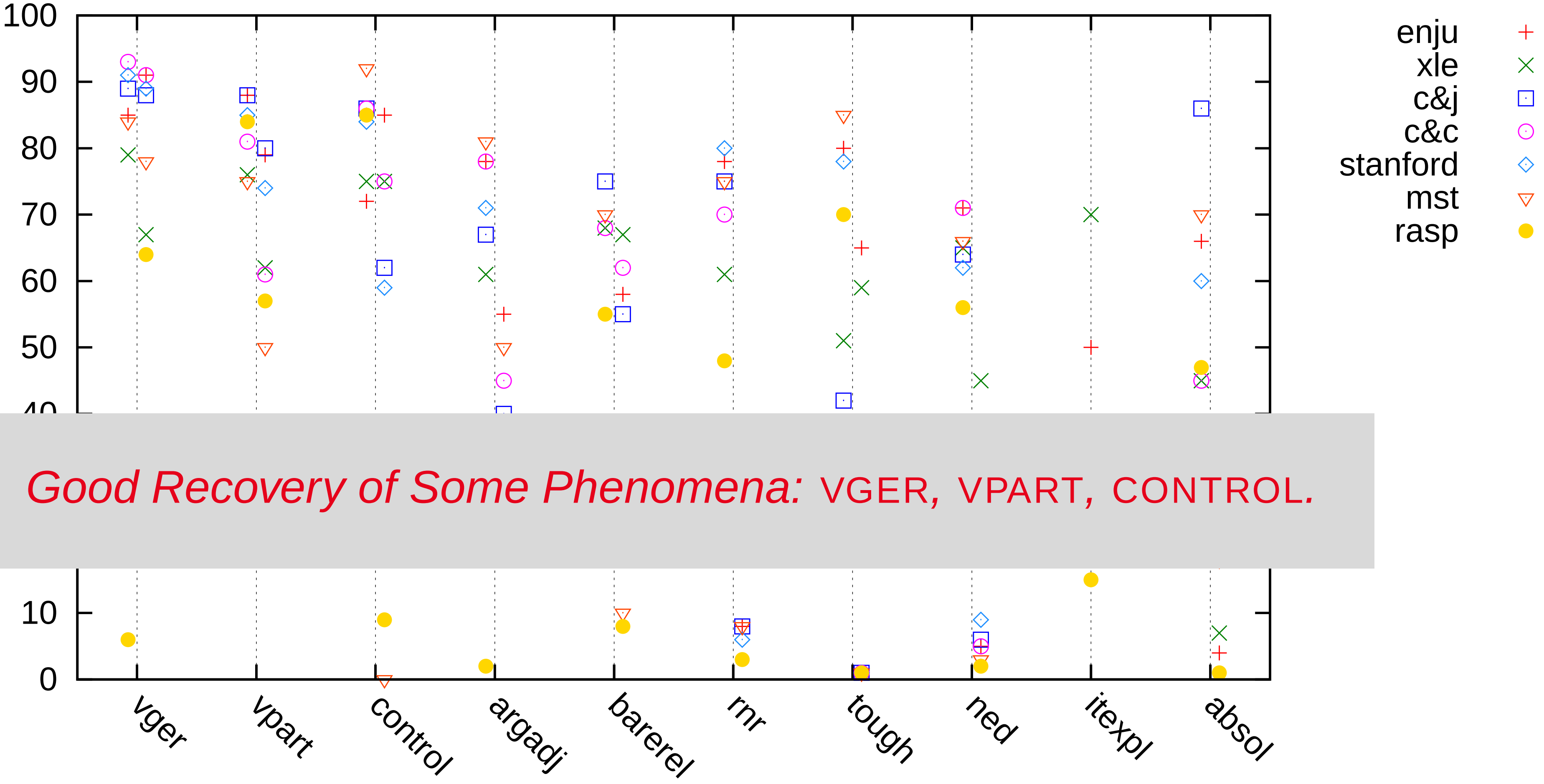


Results Summary: Per-Dependency Recall



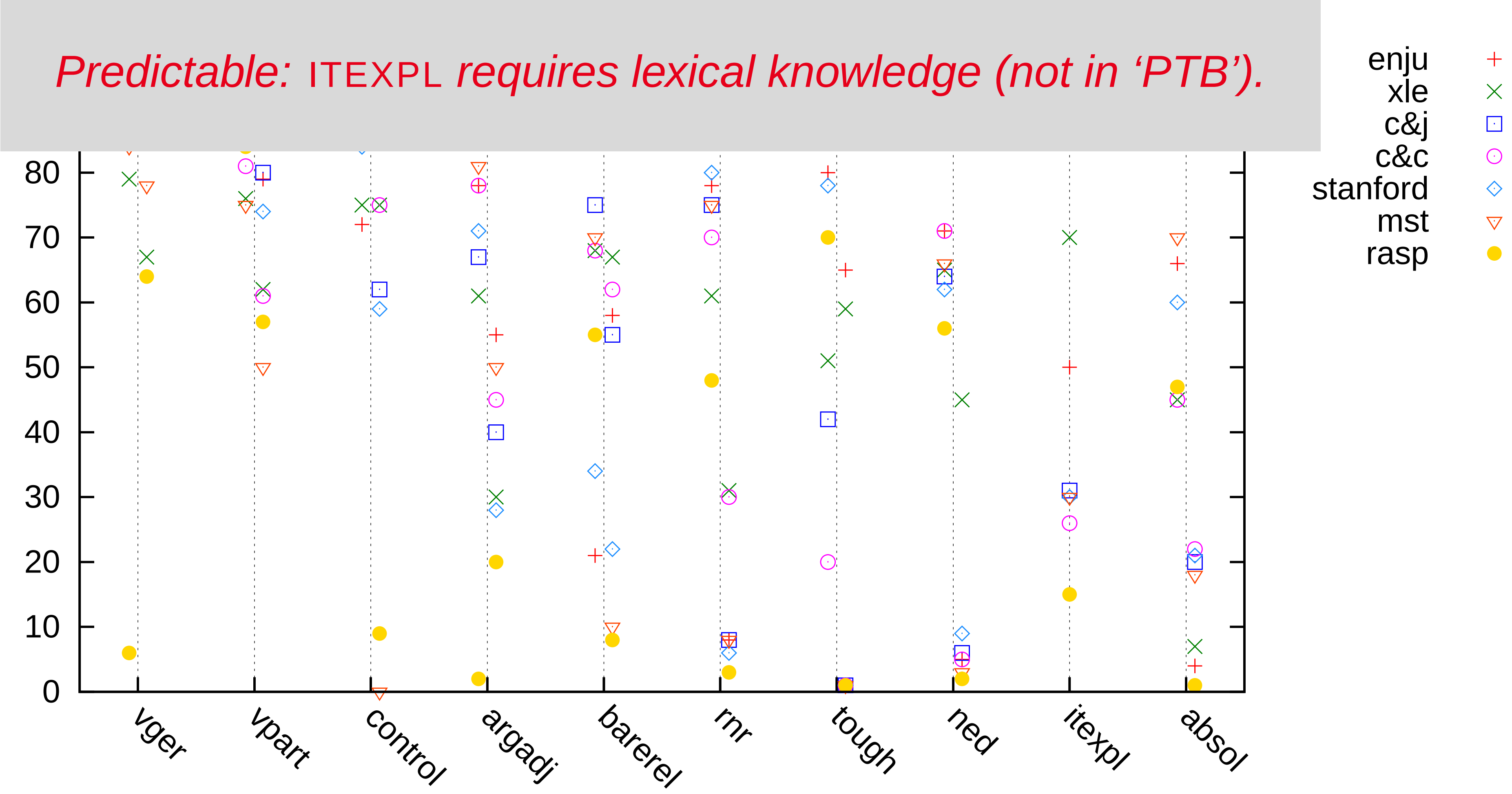
Is There Good News or Bad News (or Both)?

Results Summary: Per-Dependency Recall



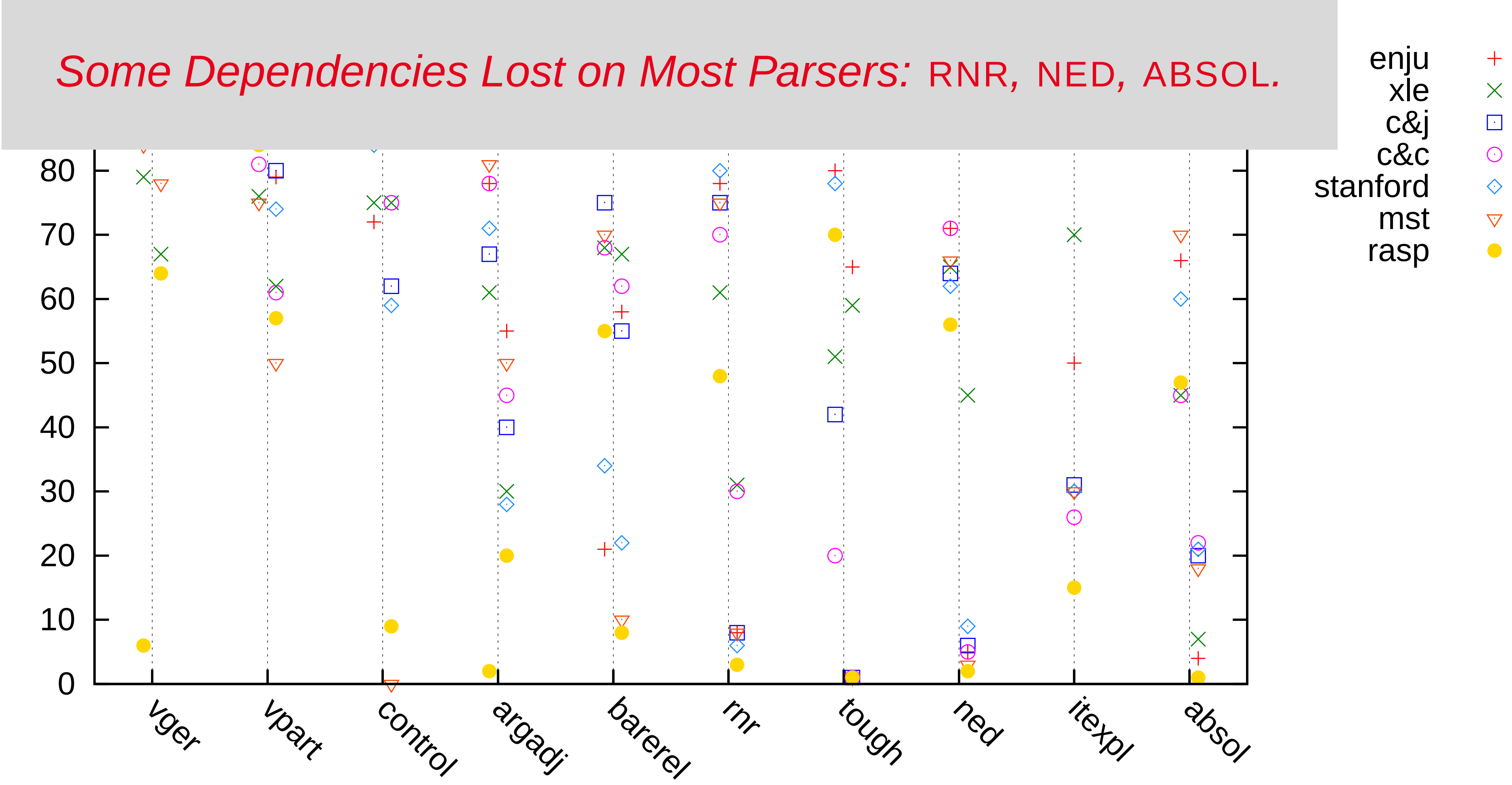
Good Recovery of Some Phenomena: VGER, VPART, CONTROL.

Results Summary: Per-Dependency Recall


Predictable: ITEXPL *requires lexical knowledge (not in 'PTB').*

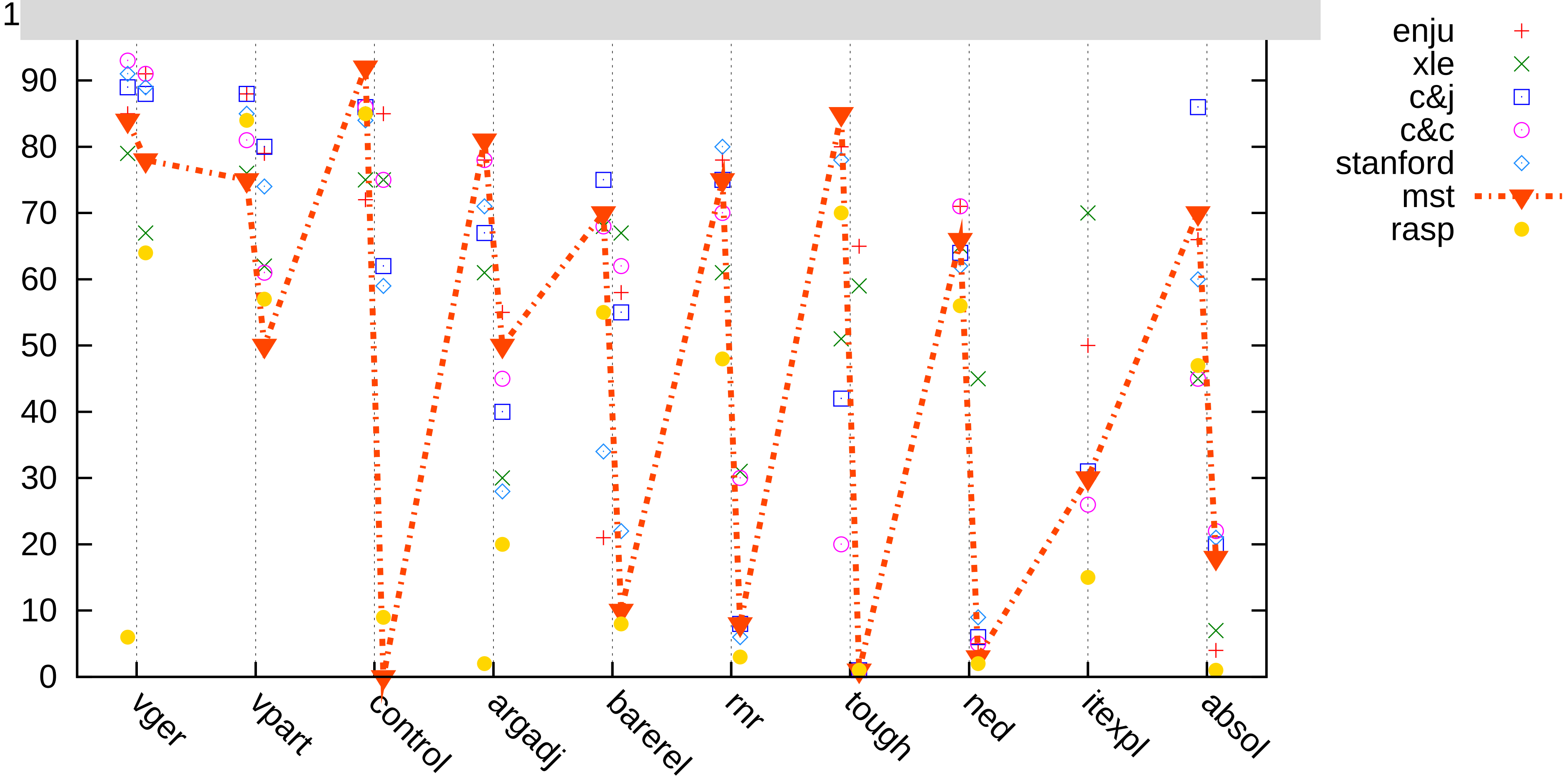
Results Summary: Per-Dependency Recall

Some Dependencies Lost on Most Parsers: RNR, NED, ABSOL.



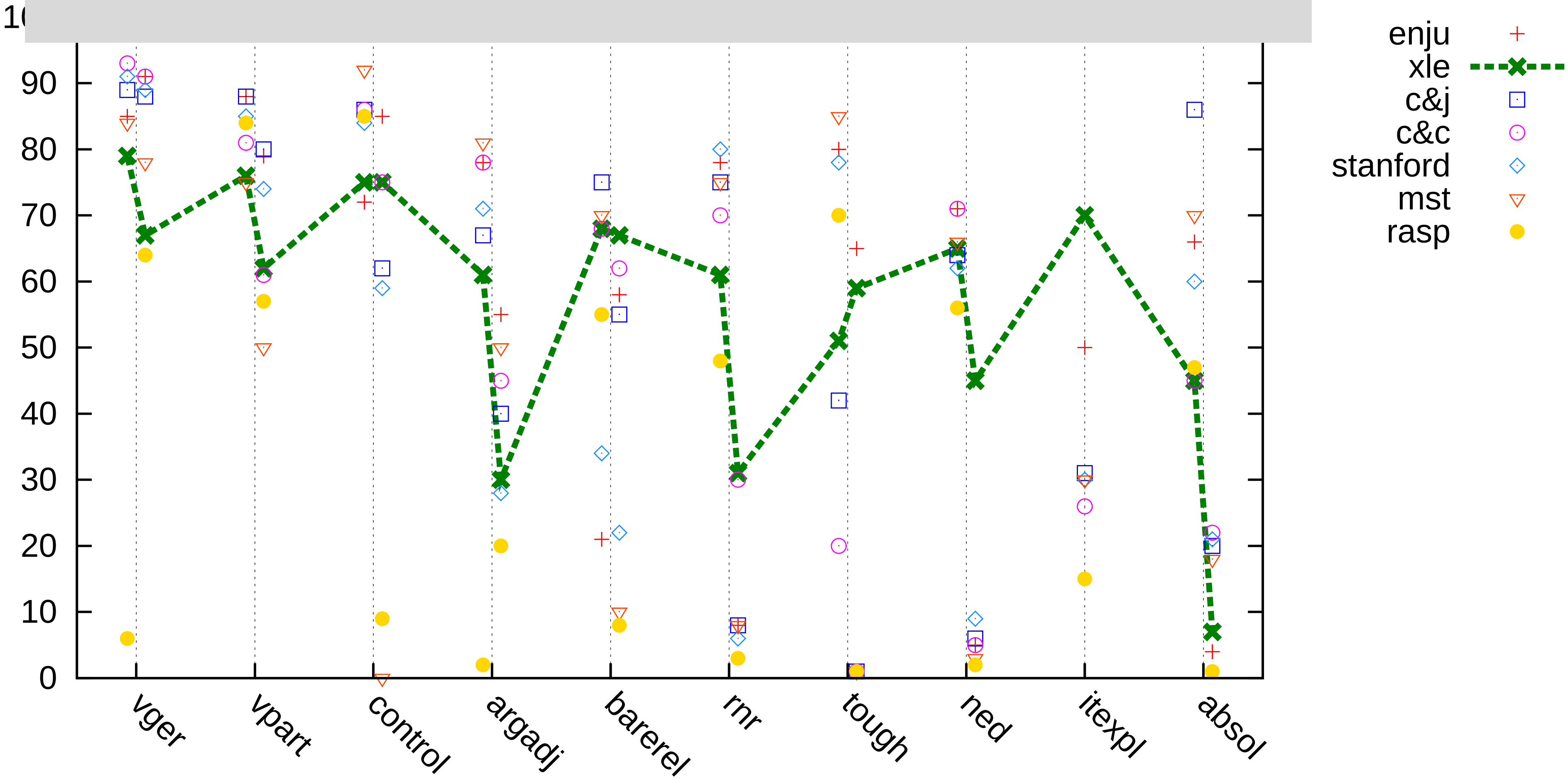
Cross-Phenomenon and -Dependency Variation (MST)

Great Variation Within Many Phenomena for Most Parsers.

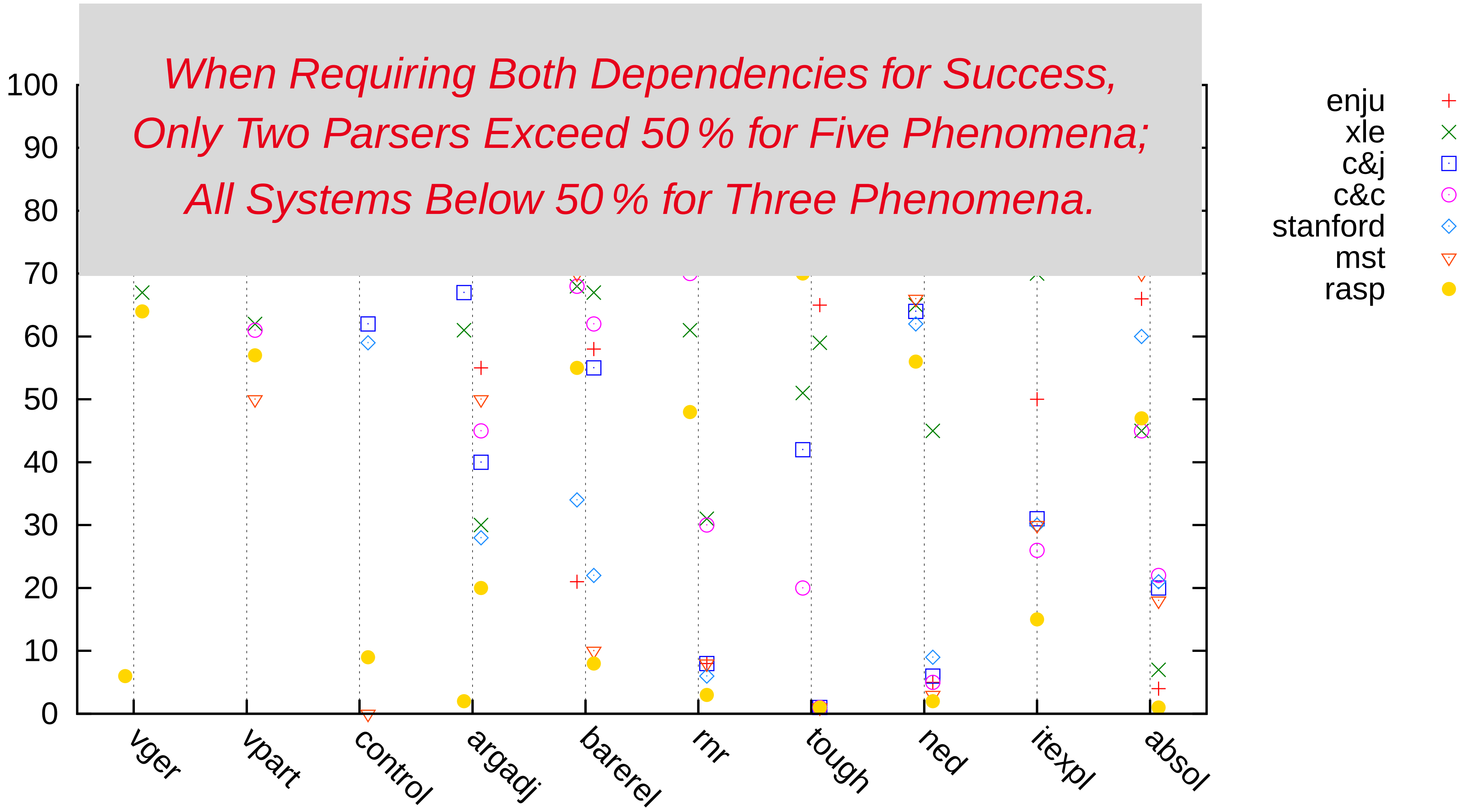


By Comparison: Grammar-Based Parsing (XLE)

With Some Exceptions, Comparatively Even Performance.

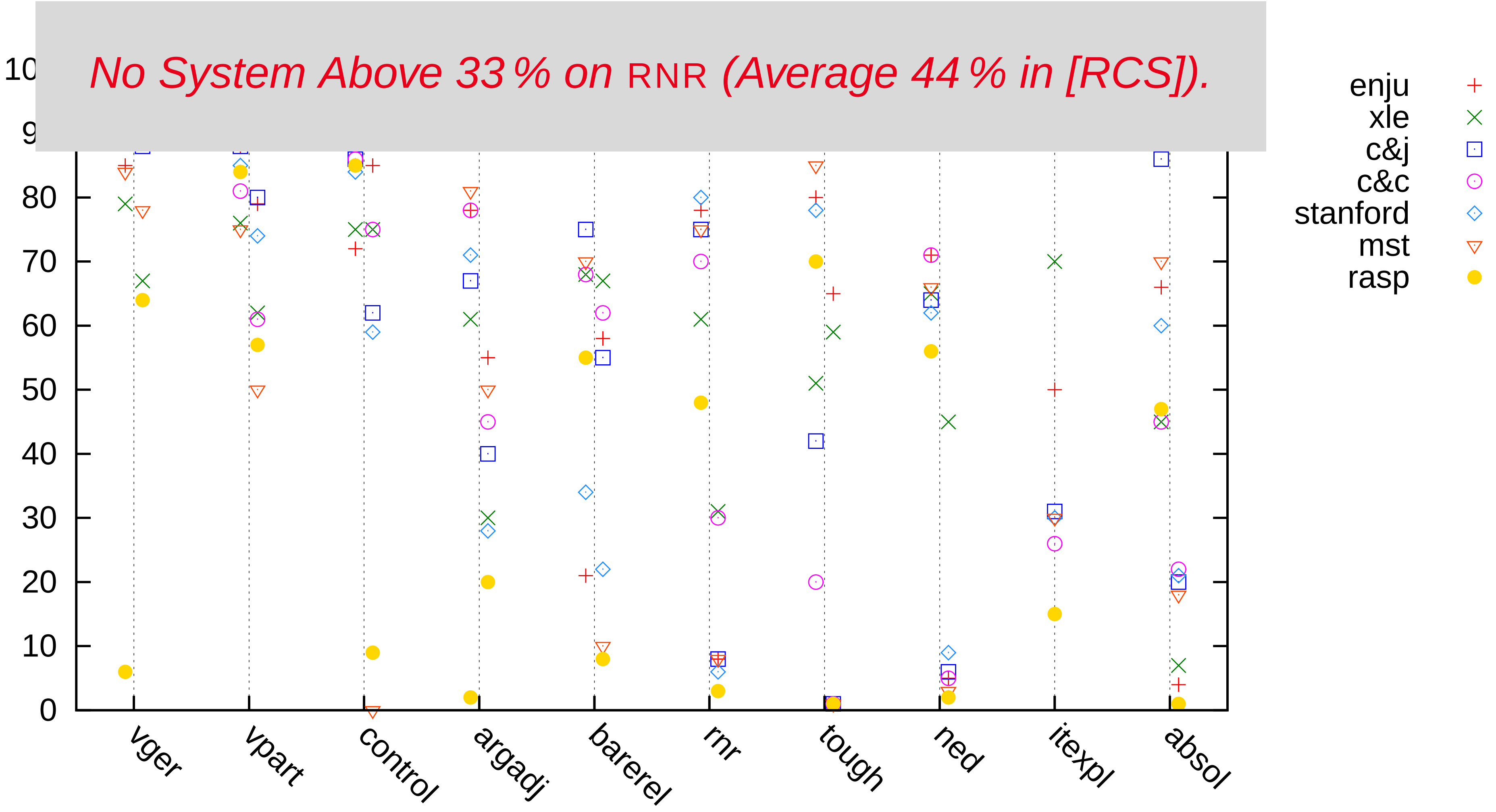


Results Summary: A Somewhat Grim Point of View



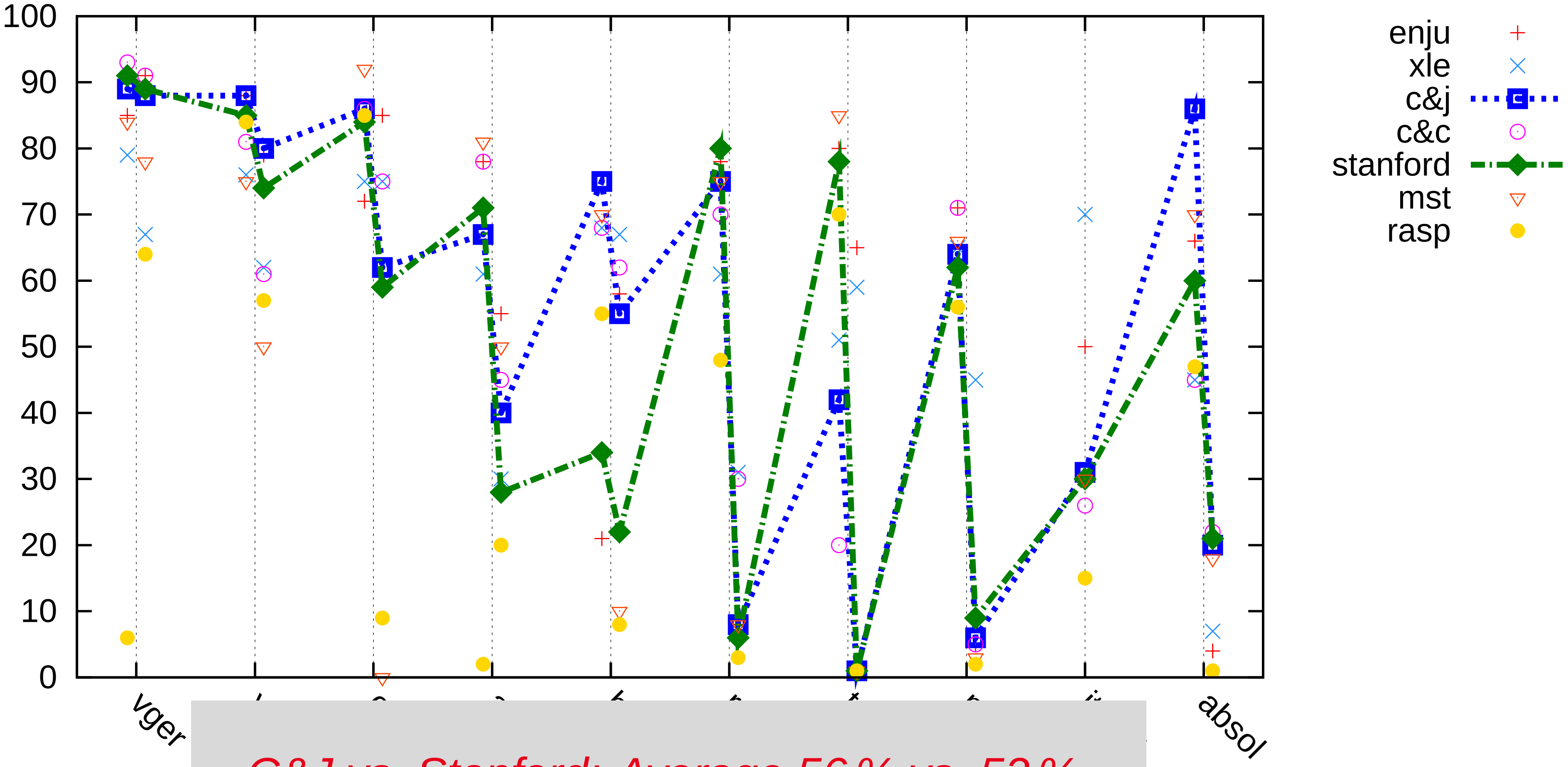
When Requiring Both Dependencies for Success,
Only Two Parsers Exceed 50 % for Five Phenomena;
All Systems Below 50 % for Three Phenomena.

Results Summary: A Somewhat Grim Point of View



No System Above 33 % on RNR *(Average 44 % in [RCS]).*

Results Summary: Pointwise Parser Comparison



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;

ded.

Background and download: `http://www.delph-in.net/ddec/`

Bibliography