Comparing Hand-Crafted with Automatically Acquired English Constraint-Based Parsing Resources

Aoife Cahill*, Michael Burke*, Ruth O'Donovan*,
Stefan Riezler§, Josef van Genabith* and Andy Way*

* National Centre for Language Technology
Dublin City University
Dublin, Ireland
{acahill, mburke, rodonovan, josef, away}
@ computing.dcu.ie

§ Palo Alto Research Center
3333 Coyote Hill Rd.,
Palo Alto, CA 94304
riezler@parc.com

Outline

- Introduction
- Automatic Annotation Algorithm
- Parsing Experiments
  - Development phase
  - Evaluation against PARC 700 (King et al., 2003) and CB 500 (Carroll et al., 1998)
- Results
- Initial Analysis of results
- Demo
- Conclusion

Introduction

- Development of large-scale deep unification grammars by hand is time-consuming and expensive
- Most treebank-based parsing technology produces “shallow” grammars
- Can “deep”, probabilistic constraint-based grammars be aquired from treebanks?
  - *yes, if we have an f-structure annotated treebank at our disposal*
  - but, they don’t exist

Introduction

Penn-II Treebank (WSJ section): contains more than 1,000,000 words in 50,000 sentences and trees.

►But no f-structure information.

Manual f-structure annotation of grammar rules extracted is time consuming and expensive.

(Large number of CFG rule types - > 19,000 for Penn-II)

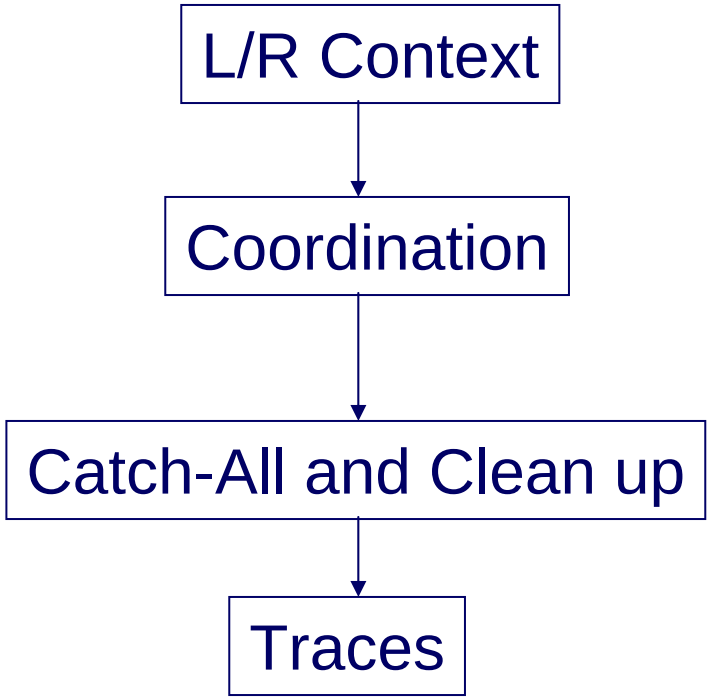
Contains (often) flat rules and some mistagging

Ambiguity of natural language makes our task harder.

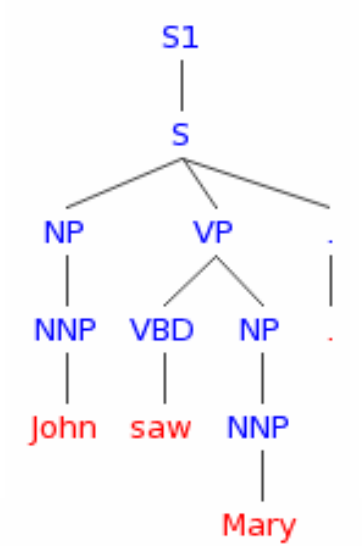
Annotation Algorithm for Penn-II

Annotation Algorithm exploits:

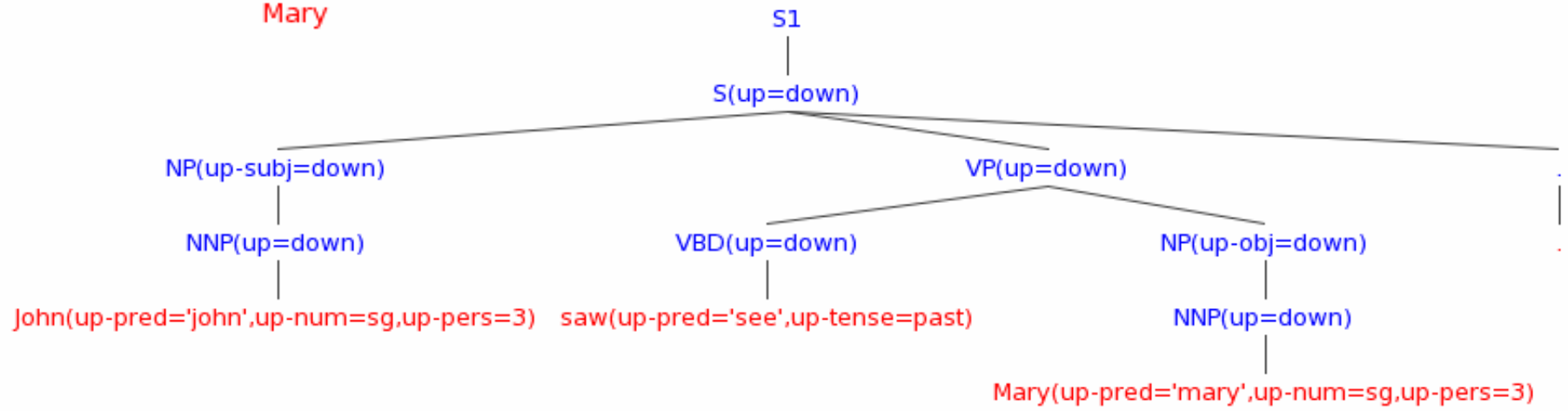
- Configurational information
- Categorial information
- Head information
- Penn functional tags
- Trace information



Automatic Annotation

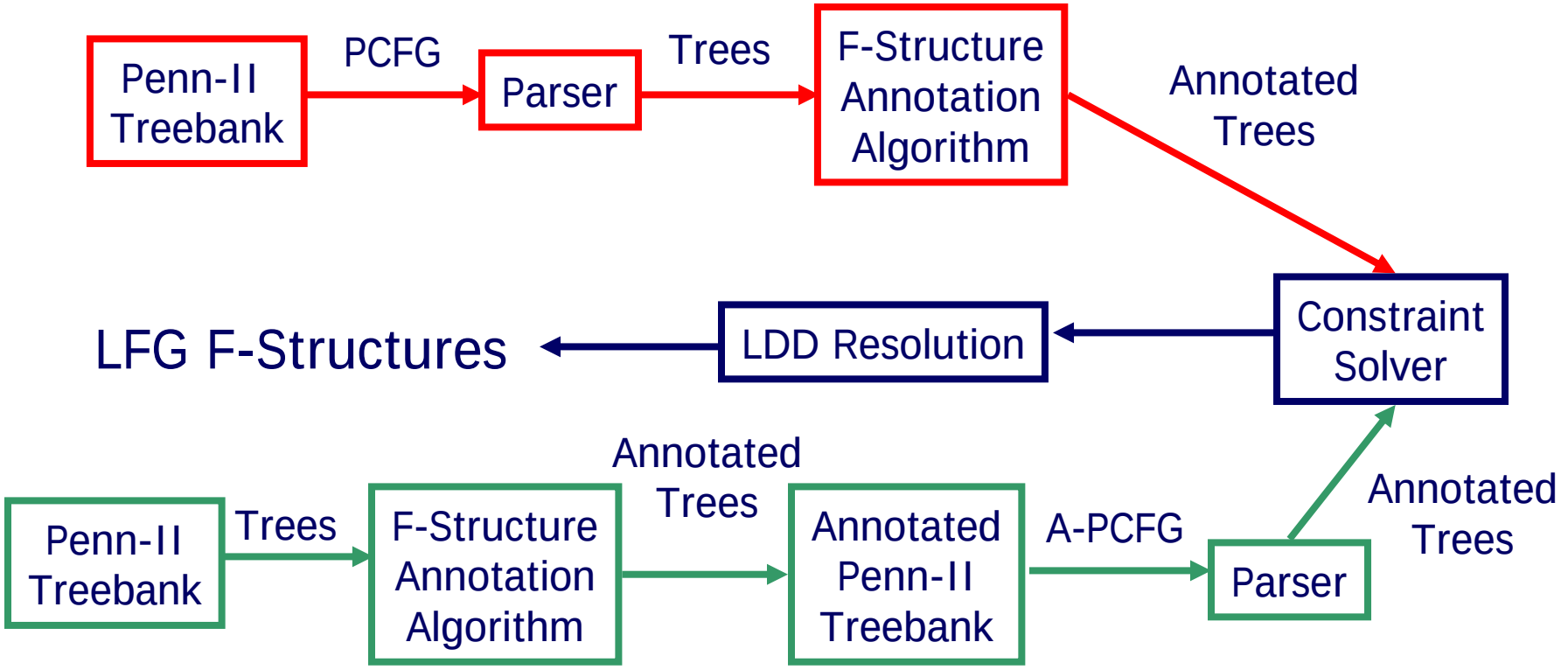


$$f_1:\begin{bmatrix} \text{PRED} & \text{`SEE}\langle(\uparrow\text{SUBJ})(\uparrow\text{OBJ})\rangle\text{'} \\ \text{SUBJ} & f_2:\begin{bmatrix} \text{PRED} & \text{`JOHN'} \\ \text{NUM} & \text{SG} \\ \text{PERS} & 3 \end{bmatrix} \\ \text{OBJ} & f_3:\begin{bmatrix} \text{PRED} & \text{`MARY'} \\ \text{NUM} & \text{SG} \\ \text{PERS} & 3 \end{bmatrix} \\ \text{TENSE} & \text{PAST} \end{bmatrix}$$



Parsing into LFG F-Structures

- PCFG-Based LFG approximations
- Two Parsing Architectures:
  (pipeline & integrated)



Experiment Motivation

- Automatically induced large-scale, robust PCFG-based LFG resources
- Compare with large-scale hand-crafted resources: XLE (Riezler et al., 2002; Kaplan et al., 2004) and RASP (Carroll and Briscoe, 2002)
- Evaluation in terms of dependencies (using XLE triples program and Carroll and Briscoe evaluation software)
- Long-term aim is to make our resources available to download
  - First step: online demo

Parsers

- PCFG
- P-PCFG
  - Parent transformation (Johnson, 1999)
- Collins (1999)
  - history-based parser
  - Model 3 (includes WHNP LDDs)
- Charniak (2000)
  - “maximum entropy inspired” parser
- Bikel (2002)
  - emulation of Collins Model 2
  - retrained to keep Penn functional tags

Development Data

- DCU 105
  - 105 sentences extracted randomly from Section 23 of the Penn II Treebank
  - Automatically annotated and hand corrected (After numerous iterations)
- WSJ 2416
  - Automatically annotate the *original* trees from Section 23 and use it as a gold standard
  - CCG-style evaluation
- Evaluation
  - XLE triples notation: *rel(arg,arg)*
  - E.g. *subj(see~1, John~0)*

Development Phase Results

- DCU 105

Parser	Preds-Only F-Score (%)	All GFs F-Score (%)
PCFG	74.00	84.02
P-PCFG	78.33	86.46
Collins Model 3	77.86	85.66
Charniak	80.50	86.75
Bikel M2 Emulation	79.73	86.80
Bikel Retrained	83.27	88.59

- WSJ 2416

Parser	Preds-Only F-Score (%)	All GFs F-Score (%)
PCFG	73.78	84.00
P-PCFG	77.95	86.19
Collins Model 3	80.10	86.82
Charniak	82.63	88.08
Bikel M2 Emulation	82.35	88.23
Bikel Retrained	84.38	89.14

Significance Testing

Approximate Randomization Test (Noreen, 1988)

	PCFG	P-PCFG	C3	CH	BK	BKR
PCFG						
P-PCFG	<0.0001					
C3	<0.0001	<0.0001				
CH	<0.0001	<0.0001	<0.0001		0.3513	
BK	<0.0001	<0.0001	<0.0001			
BKR	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	

Evaluation

Evaluation of Bikel Retrained Parser + Automatic Annotation Algorithm against **PARC 700** (King et al., 2003)

- 700 sentences randomly selected from Section 23 of the Penn II treebank
- Dependency triples
- Initially parsed with XLE English grammar and manually hand corrected

- Mapping to convert our DCU-style f-structures into PARC format (Burke et al., 2004)
- Test set split, 140 development, 560 test

PARC 700 Results

Dependency	% of total deps	F-Score (%)	
		BKR	XLE
adegree	6.17	81	82
adjunct	14.32	68	66
aquant	0.06	78	61
comp	1.23	80	74
conj	2.64	72	69
coord_form	1.20	83	90
det_form	4.61	97	91
focus	0.02	0	36
mod	2.74	75	67
num	19.82	91	89
number	1.42	88	83
number_type	2.10	95	86
obj	8.92	88	78
obj_theta	0.05	43	31
obl	0.83	56	69
obl_ag	0.22	80	76
obl_compar	0.07	38	56
passive	1.14	81	88

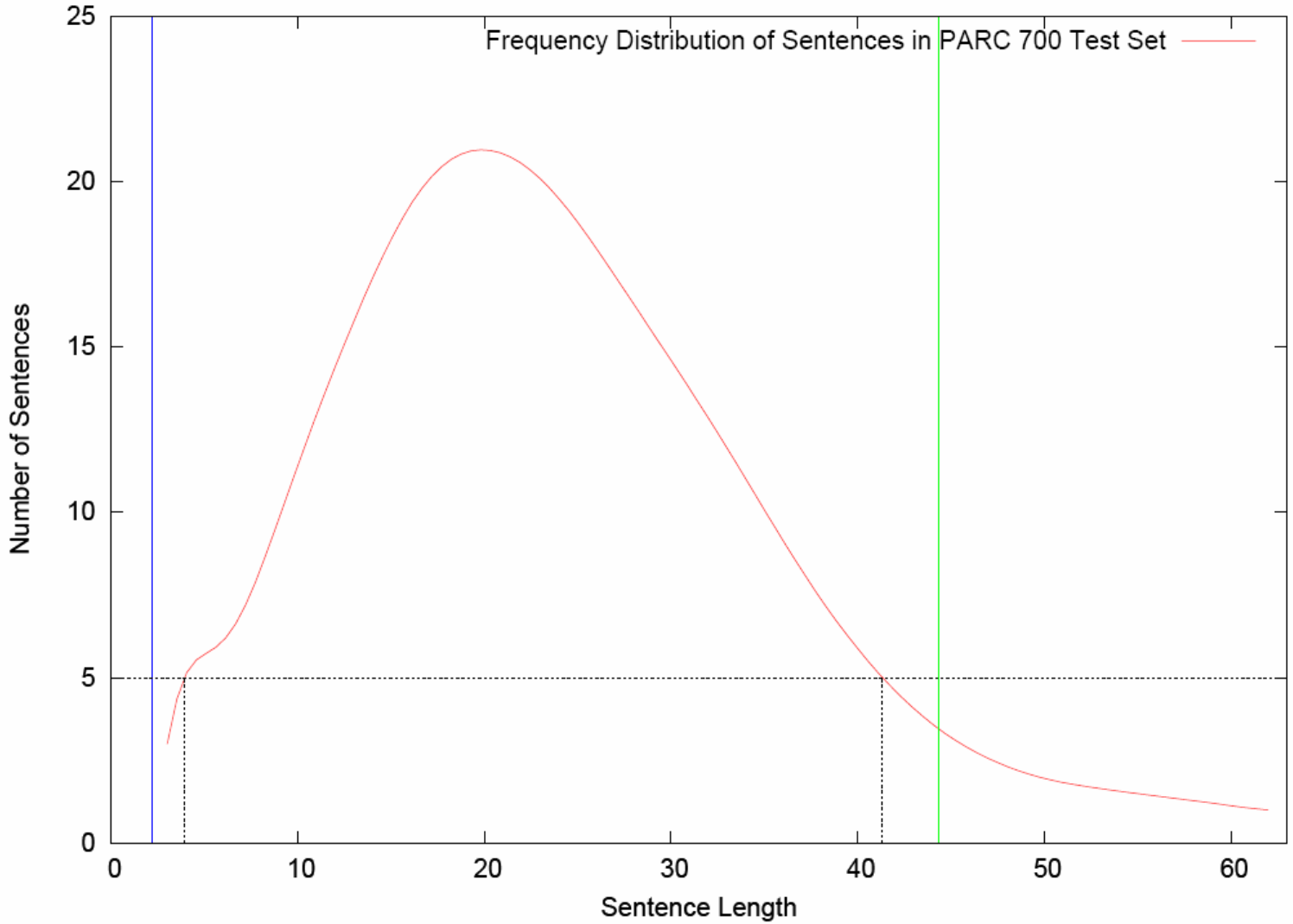
Dependency	% of total deps	F-Score (%)	
		BKR	XLE
pcase	0.25	77	68
perf	0.41	89	90
poss	0.98	89	80
precoord_form	0.03	0	91
prog	0.97	90	81
pron_form	2.54	92	94
pron_int	0.03	0	33
pron_rel	0.57	75	72
proper	3.56	84	93
prt_form	0.22	80	41
quant	0.34	77	80
stmt_type	5.23	88	80
subj	8.51	78	78
subord_form	0.93	47	42
tense	5.02	95	90
topic_rel	0.57	56	73
xcomp	2.29	80	78
Overall	100	83.08	80.55

PARC 700 Preds-Only Results

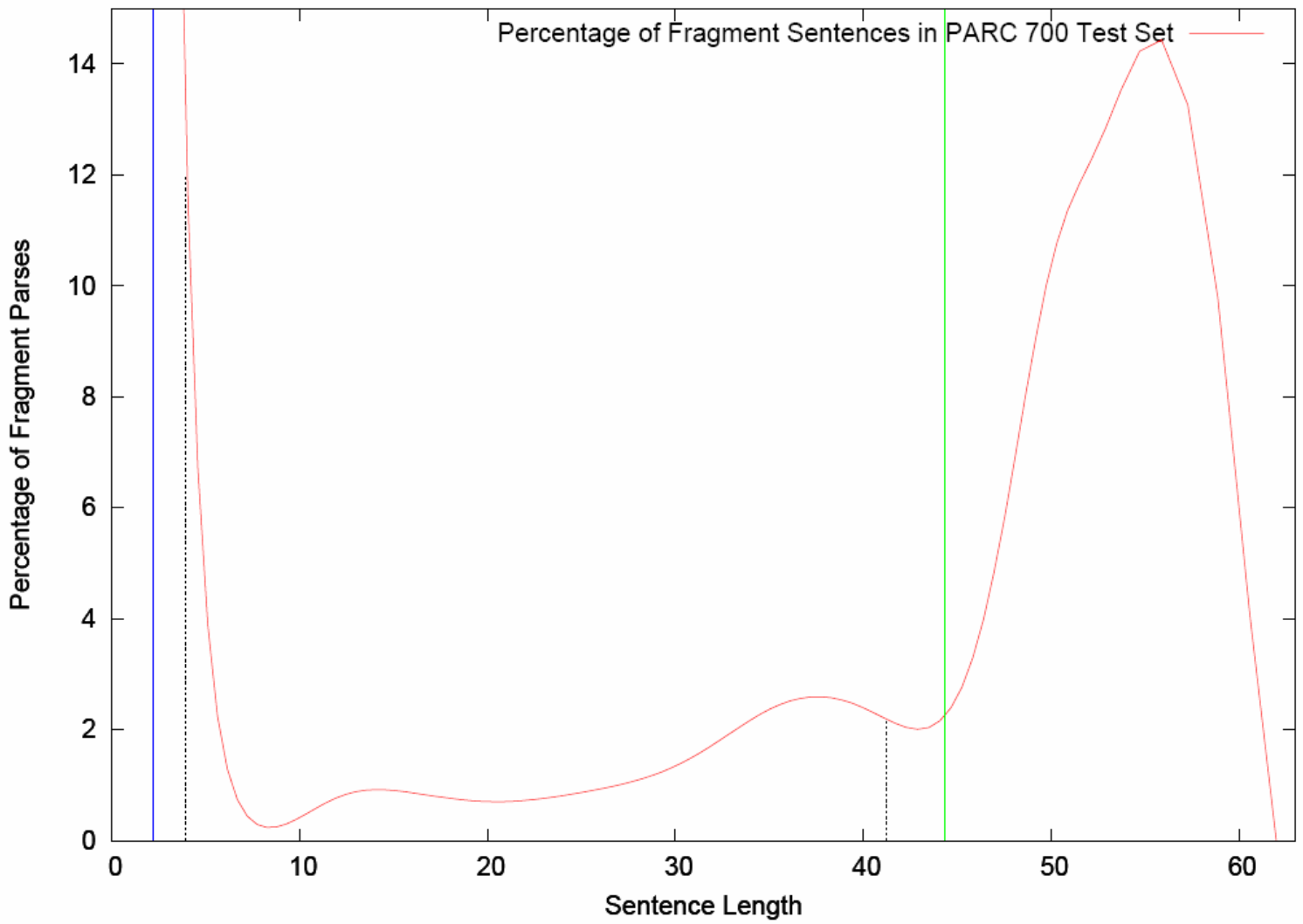
Dependency	% of total deps	F-Score (%)	
		BKR	XLE
adjunct	27.14	68	66
aquant	0.12	78	61
comp	2.33	80	74
conj	5.00	72	69
coord_form	2.28	83	90
det_form	8.74	97	91
focus	0.05	0	36
mod	5.19	75	67
number	2.69	88	83
obj	16.91	88	78
obj_theta	0.10	43	31
obl	1.57	56	69

Dependency	% of total deps	F-Score (%)	
		BKR	XLE
obl_ag	0.41	80	76
obl_compar	0.14	38	56
poss	1.86	89	80
pron_int	0.05	0	33
pron_rel	1.08	75	72
prt_form	0.42	80	41
quant	0.64	77	80
subj	16.12	78	78
subord_form	1.77	47	42
topic_rel	1.08	56	73
xcomp	4.33	80	78
Overall	100	77.68	74.31

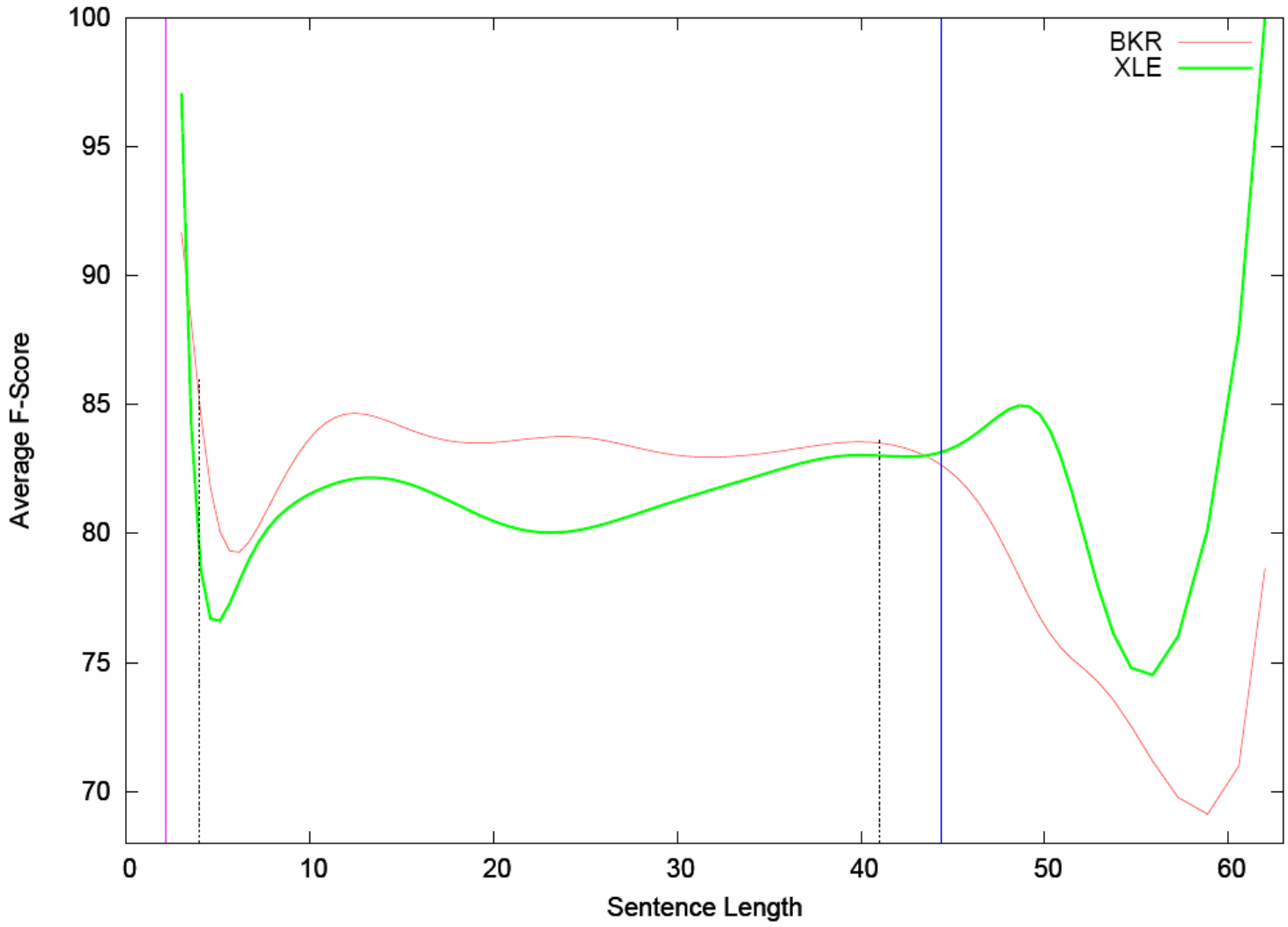
PARC 700 Evaluation



PARC 700 Evaluation



PARC 700 Evaluation



Evaluation

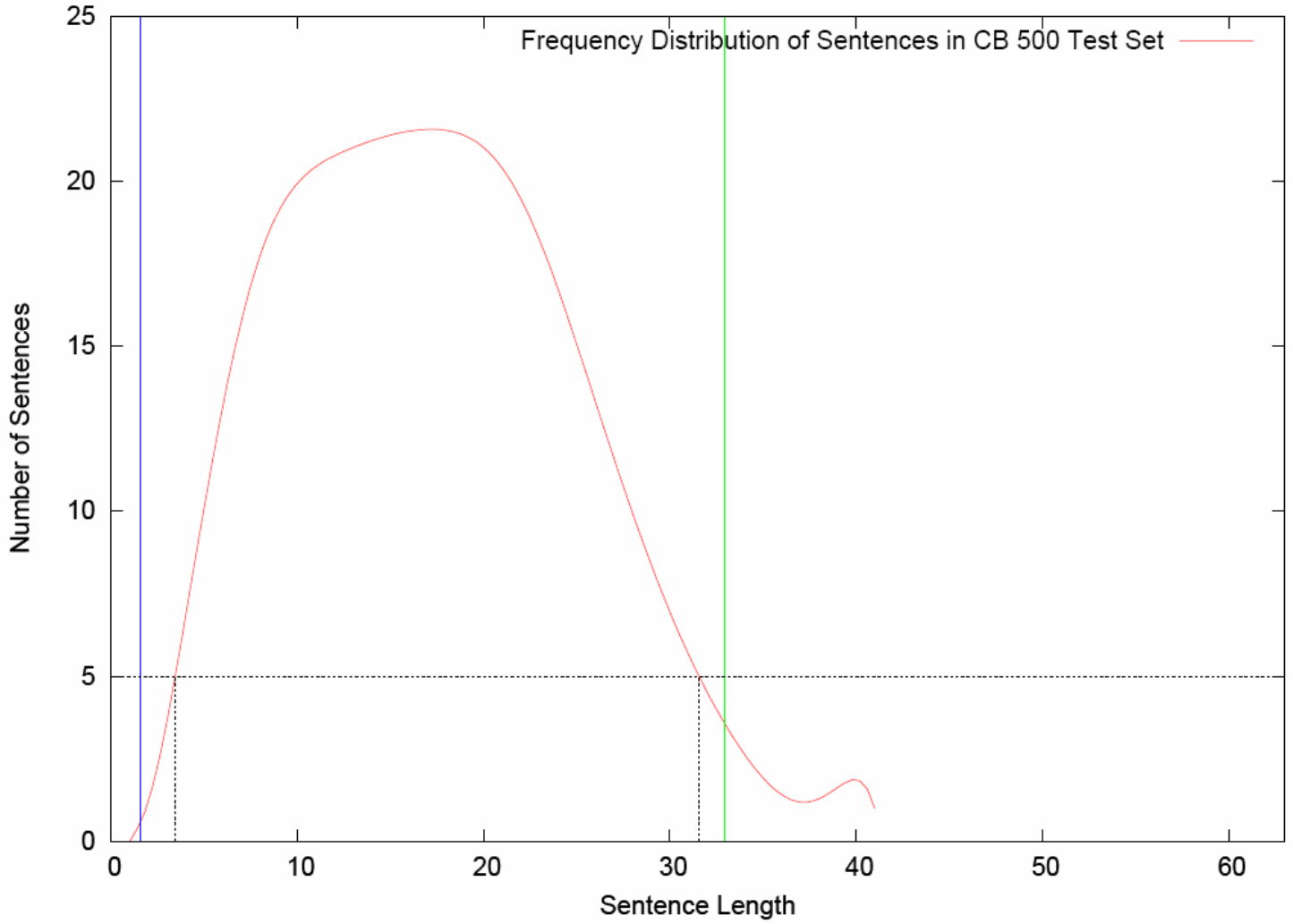
Evaluation of Bikel Retrained Parser + Automatic Annotation Algorithm against **CB 500** (Carroll et al., 1998)

- 500 sentences from SUSANNE corpus
- Manually annotated
- All sentences chosen were parsable by the RASP parser
- Capture dependency relations

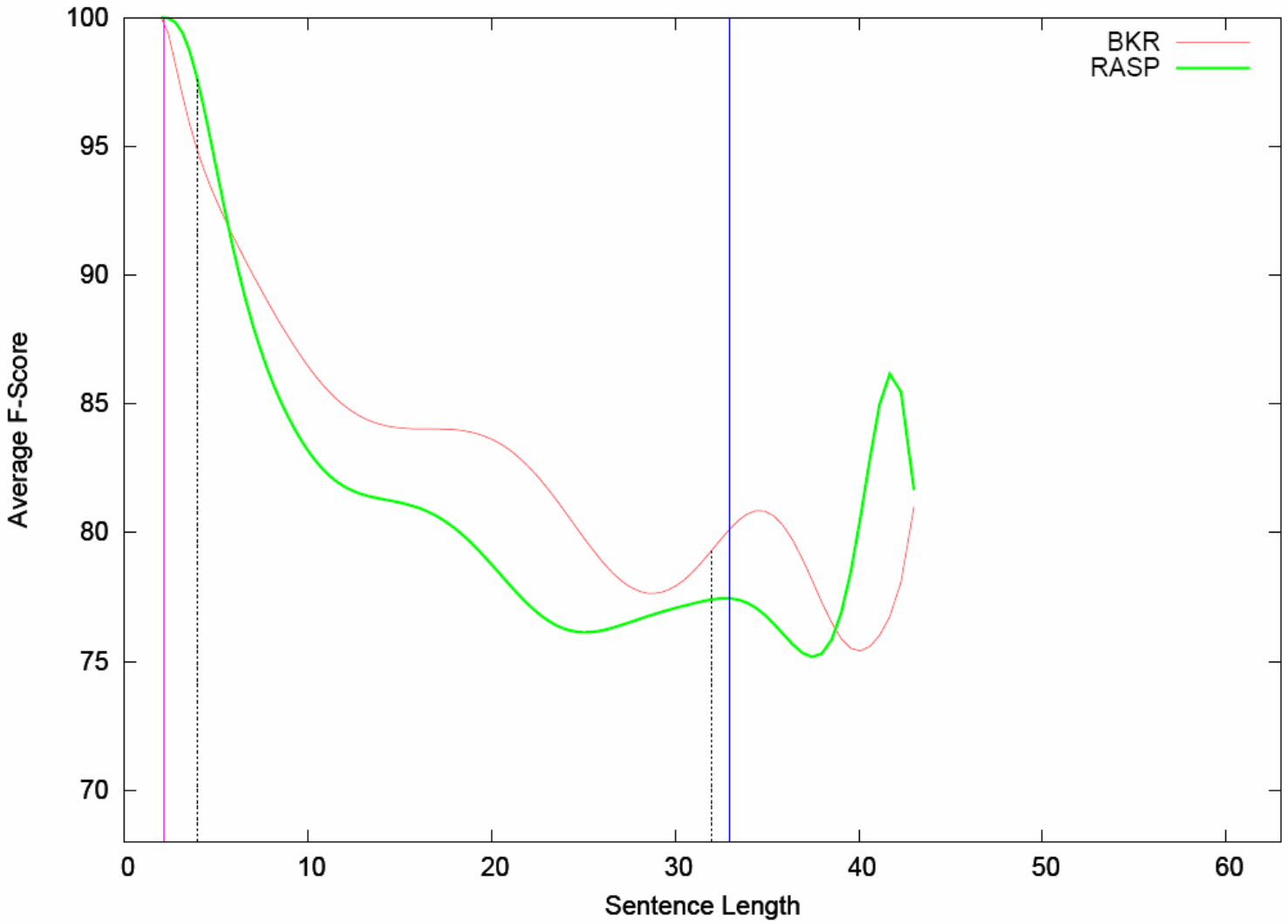
CB 500 Evaluation

Dependency	% total deps.	BKR	RASP
dependent	100	80.23	76.57
mod	48.96	80.3	75.29
ncmod	30.42	85.37	72.98
xmod	1.6	70.05	55.88
cmod	2.61	75.6	53.08
detmod	14.06	95.85	91.97
arg_mod	0.51	80	64.52
arg	43.7	78.28	77.57
subj	13.1	79.84	83.57
ncsubj	12.99	87.84	84.32
xsubj	0.06	0	88.89
csubj	0.04	0	22.22
subj_or_dobj	18.22	81.21	83.84
comp	12.38	76.73	71.87
obj	7.34	76.05	69.53
dobj	5.11	84.55	84.57
obj2	0.25	48	43.84
iobj	1.98	59.04	47.6
clausal	5.04	77.74	75.37
xcomp	4.03	80	84.11
ccomp	1.01	69.61	75.14
aux	4.76	94.94	88.27
conj	2.06	68.84	69.09

CB 500 Evaluation



CB 500 Evaluation



Demo

http://lfg-demo.computing.dcu.ie/lfgparser.html

Conclusions

- Successfully developed and implemented a methodology for automatically acquiring large-scale, probabilistic LFG approximations
- Statistically significantly better results than hand-crafted XLE and RASP resources
- Initial analysis of differences in results
- Online Demo
  - Soon to be released on LFG/HPSG/Linguist lists
  - Please email us with any feedback☺

References

Bikel, Dan. 2002. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In Proceedings of HLT 2002, pages 24–27, San Diego, CA.

Carroll, John and Edward Briscoe. 2002. High precision extraction of grammatical relations. In Proceedings of the 19th International Conference on Computational Linguistics (COLING), pages 134–140, Taipei, Taiwan

Carroll, John, Edward Briscoe, and Antonio Sanfilippo. 1998. Parser evaluation: A survey and new proposal. In Proceedings of the International Conference on Language Resources and Evaluation, pages 447–454, Granada, Spain.

Charniak, Eugene. 2000. A maximum entropy inspired parser. In Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000), pages 132–139, Seattle, WA.

References

Kaplan, Ron, Stefan Riezler, Tracy Holloway King, John T. Maxwell, Alexander Vasserman, and Richard Crouch. 2004. Speed and Accuracy in Shallow and Deep Stochastic Parsing. In Proceedings of the Human Language Technology Conference and the 4th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04), pages 97–104, Boston, MA.

King, Tracy Holloway, Richard Crouch, Stefan Riezler, Mary Dalrymple, and Ron Kaplan. 2003. The PARC 700 dependency bank. In Proceedings of the EACL03: 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), pages 1–8, Budapest, Hungary.

Riezler, Stefan, Tracy King, Ronald Kaplan, Richard Crouch, John T. Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL-02), pages 271–278, Philadelphia, PA.