Parsing experiments with DeepBank

DELPH-IN summit
St. Wendel
31.07.2013

Angelina Ivanova
University of Oslo

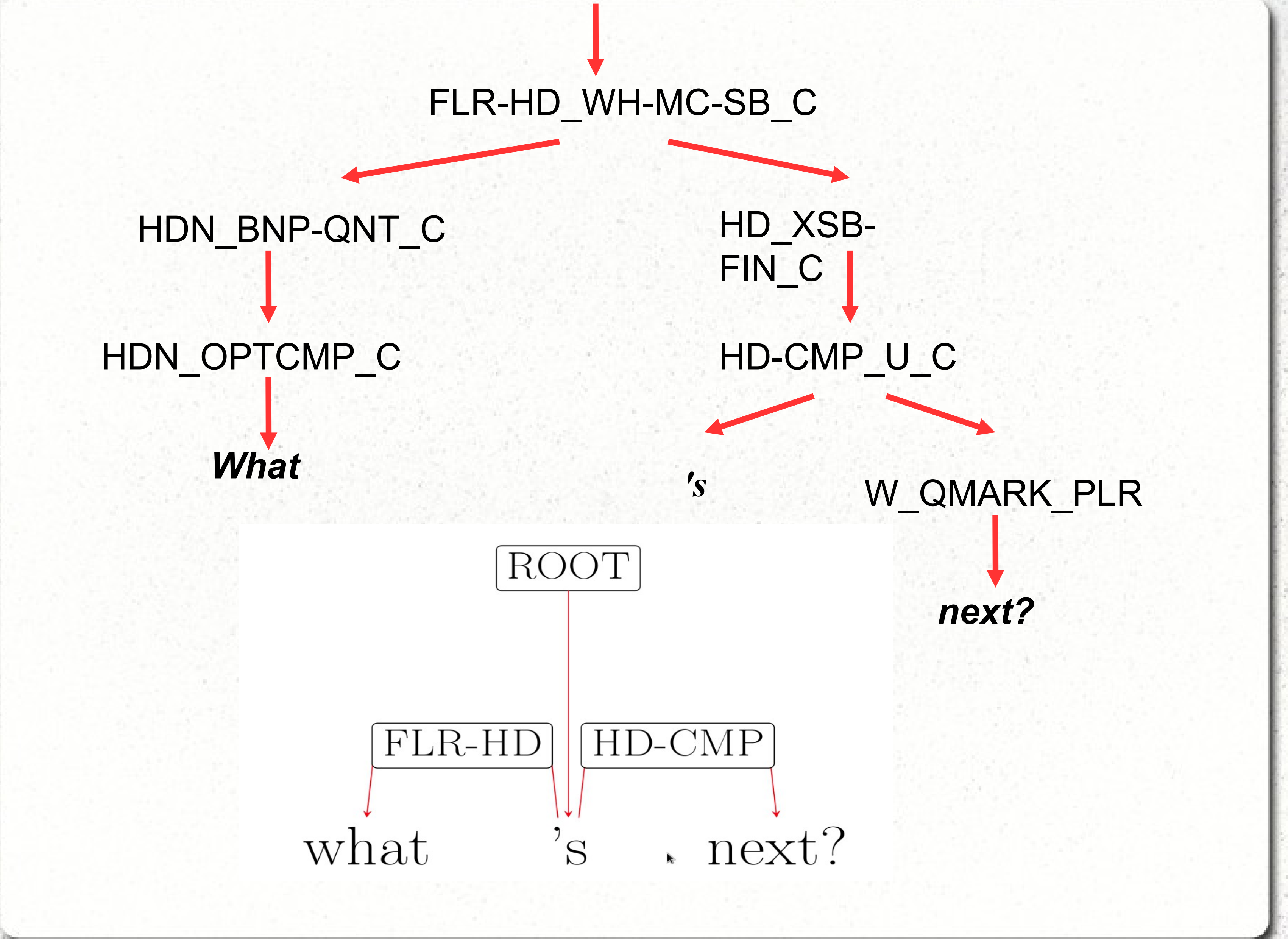
PLAN

- Dependency parsing with ERG Derivation Tree and two other annotation formats

  Angelina Ivanova, Stephan Oepen and Lilja Øvrelid. Survey on parsing three dependency representations for English. ACL Student Research Workshop 2013

- Comparison of PET to PCFG and dependency parsers

DELPH-IN Syntactic Derivation Tree (DT)



FLR-HD_WH-MC-SB_C

HDN_BNP-QNT_C

HD_XSB-FIN_C

HDN_OPTCMP_C

HD-CMP_U_C

What

's

W_QMARK_PLR

next?

Research questions:

Do Malt, MST and the Bohnet and Nivre (2012) parsers perform similarly on the Stanford Basic, CoNLL and DT?

What is the effect of using gold-standard supertags, PTB tags and combination of both for parsing?

Which linguistic structures are the hardest to parse?

Does more data help to improve the results?

Data size

	Resource	Filtered # of sent.	# of sent. used in experiment
Train	PTB, DeepBank 0-19	1.33%	33,334
Development	PTB, DeepBank 20	1.22%	1,700
Test	PTB,DeepBank 21	1.77%	1,389

Number of dependency labels: 64

Size of the tag sets

PTB tags	48
supertags	1,091

Evaluation measures: Labeled Attachment Score, Unlabeled Attachment Score, Label Accuracy

Parser Settings

Malt parser

- Tuned with MaltOptimizer
  - Algorithm: stackproj
  - Learner: liblinear

Bohnet and Nivre (2012)

- Beam: 80

Other settings are default

MST

- Default configuration

Part-of-speech tags

Similarity

Lexical category

Differences

PTB:
1) syntactic function
2) morphology

Supertags:
1) valency
2) annotations

Dependency schemes

- Stanford Basic
- CoNLL Syntactic Dependencies
- DELPH-IN Syntactic Derivation Tree

Angelina Ivanova, Stephan Oepen, Lilja Øvrelid and Dan Flickinger (2012). Who Did What to Whom? A Contrastive Study of Syntacto-Semantic Dependencies. In *Proceedings of the Sixth Linguistic Annotation Workshop*. Association for Computational Linguistics.

	Gold PTB tags		Predicted PTB tags
	Malt	MST	Bohnet and Nivre (2012)
Stanford Basic	**89.58**	88.94	90.43
CoNLL	88.70	**89.13**	**90.53**
DT	87.19	88.16	90.48

Evaluation measure: labelled accuracy score (LAS)

	Gold supertags		Predicted supertags
	Malt	MST	Bohnet and Nivre (2012)
Stanford Basic	88.53	86.10	86.64
CoNLL	88.68	87.83	87.69
DT	**90.65**	**90.74**	**87.74**

Evaluation measure: labelled accuracy score (LAS)

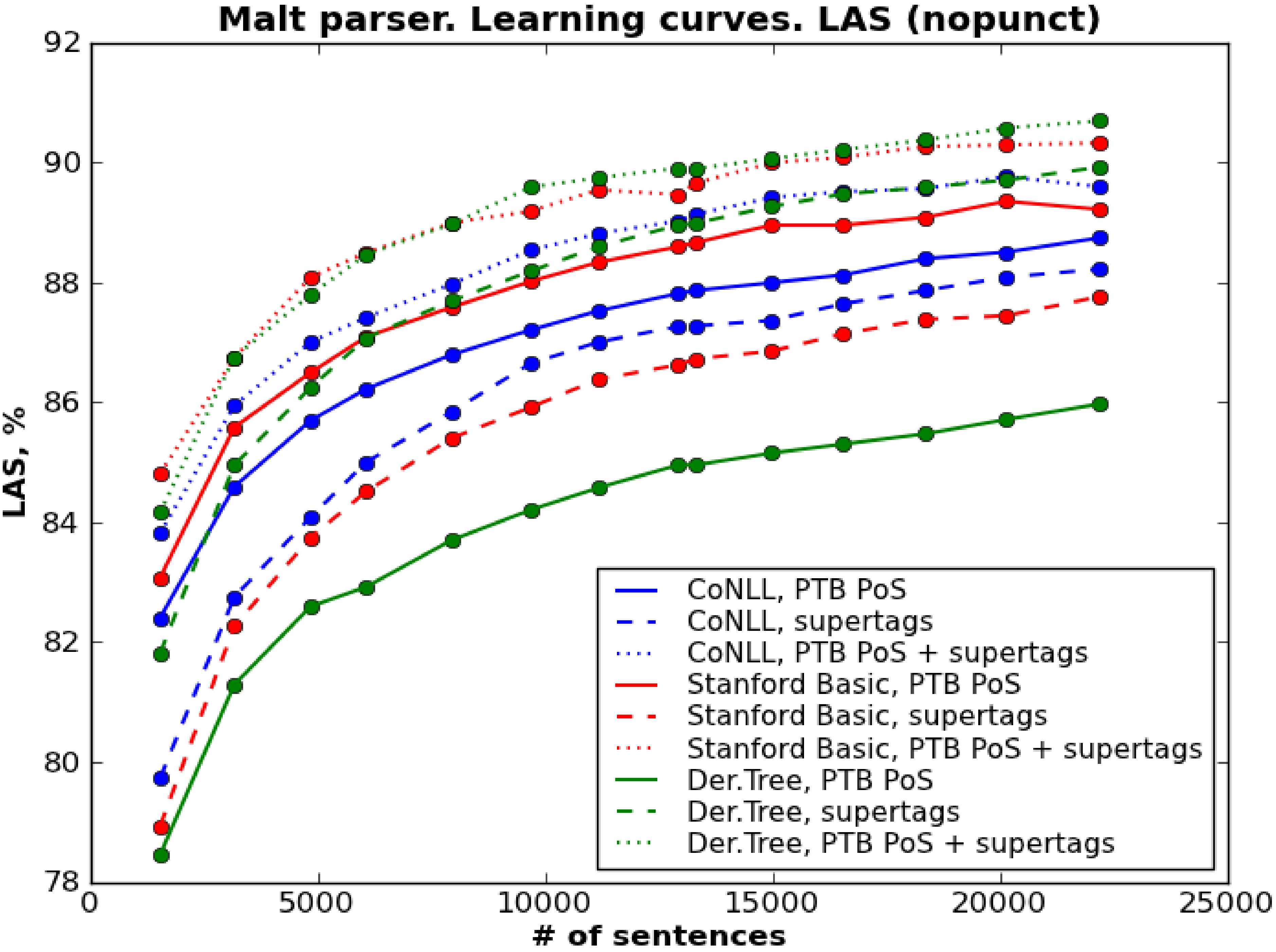
	Gold PTB tags + gold supertags	
	Malt	MST
Stanford Basic	90.79	89.63
CoNLL	89.97	89.72
DT	**91.40**	**92.43**

	Predicted PTB tags + gold supertags	Predicted supertags + gold PTB tags
	Bohnet and Nivre (2012)	
Stanford Basic	91.20	88.18
CoNLL	91.07	**89.05**
DT	**92.88**	88.44

Evaluation measure: labelled accuracy score (LAS)

Running time

Malt: minutes
MST: hours
Bohnet and Nivre (2012): minutes


Malt parser. Learning curves. LAS (nopunct)
LAS, %
# of sentences
92
90
88
86
84
82
80
78
0
5000
10000
15000
20000
25000
CoNLL, PTB PoS
CoNLL, supertags
CoNLL, PTB PoS + supertags
Stanford Basic, PTB PoS
Stanford Basic, supertags
Stanford Basic, PTB PoS + supertags
Der.Tree, PTB PoS
Der.Tree, supertags
Der.Tree, PTB PoS + supertags

From 16 to 22 sections of DEEPBANK

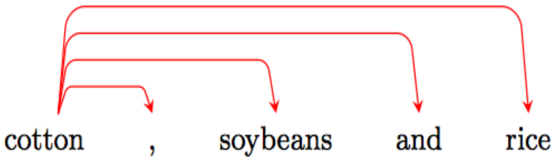
For MST and the Bohnet and Nivre (2012) **parser results improved significantly on DT scheme** in all the configurations

On SB and CD the changes are significant only in some configurations.

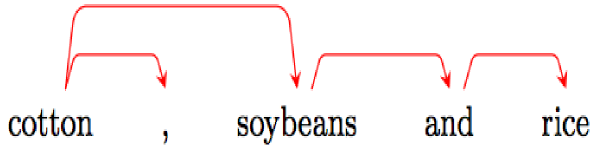
MST had significant increase of accuracy in more setups than Malt.

Error analysis

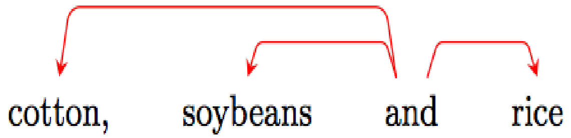
1) Coordination



Stanford Basic



CoNLL



DT

The error rate for the coordinating conjunction is not so high for CoNLL and Stanford Basic (2% and 1% correspondingly) while for the DT the error rate on the CPOSTAGS is especially high (26%).
It means there are many **errors on incoming arcs of CC** in DT.

On outgoing arcs parsers also make more mistakes on DT than on SB and CD.

Error analysis

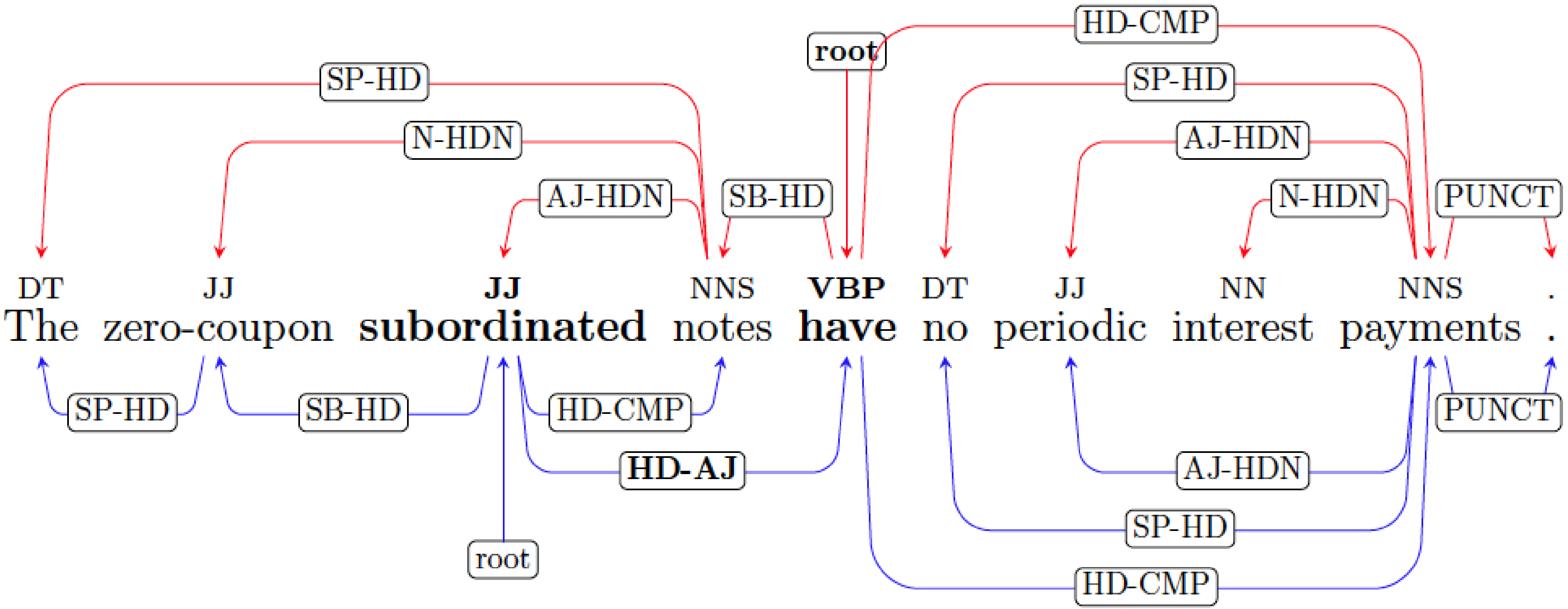
a) The fight is putting a tight squeeze on profits of many, threatening to drive the smallest ones out of business and straining relations between the **national** fast-food **chains** and their **franchisees**.

b) Proceeds from the sale will be used for re-modeling and re-forbishing projects, as well as for the **planned** MGM Grand **hotel/casino** and theme **park**.

Error analysis

2) Verbs

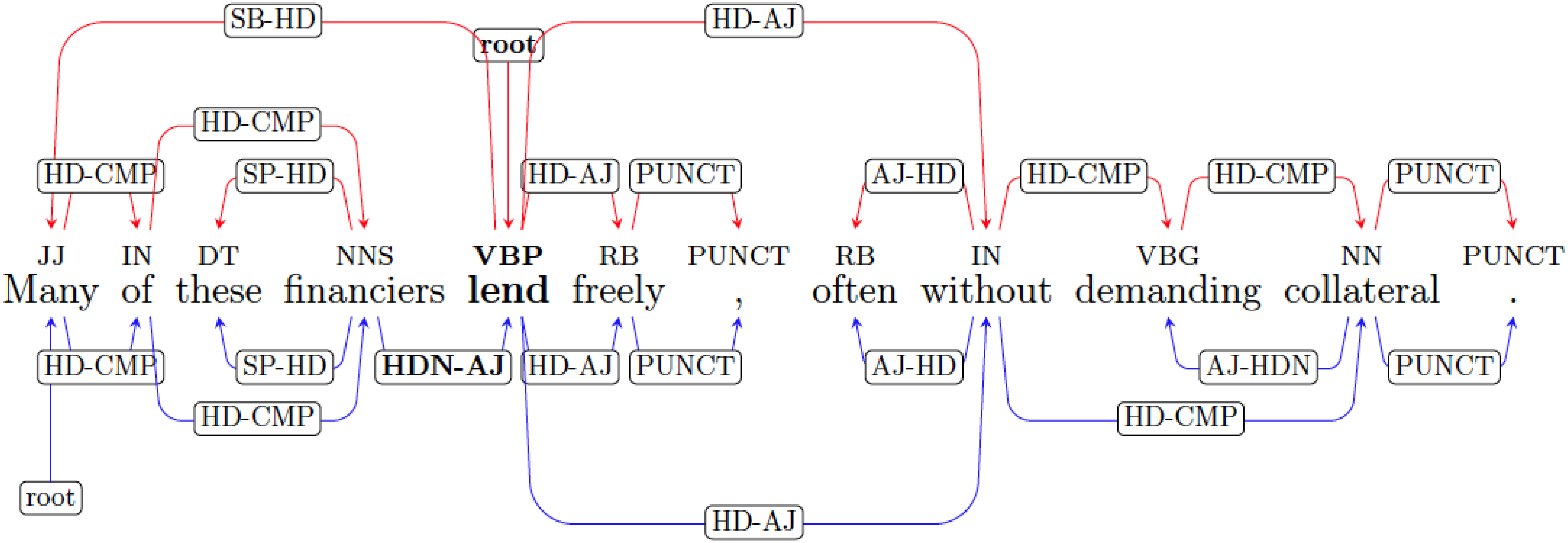
Most common errors with verbs:
1) Some nouns and adjectives are incorrectly assigned root role and this error destructs the rest of the dependency graph. It happened for the nouns and adjectives that have corresponding homonym verbs.



Error analysis

2) Verbs

Most common errors with verbs:
2) In many cases VBP errors are related to the root of the sentence. It is either treated as complement or adjunct instead of having a root status or vice versa.

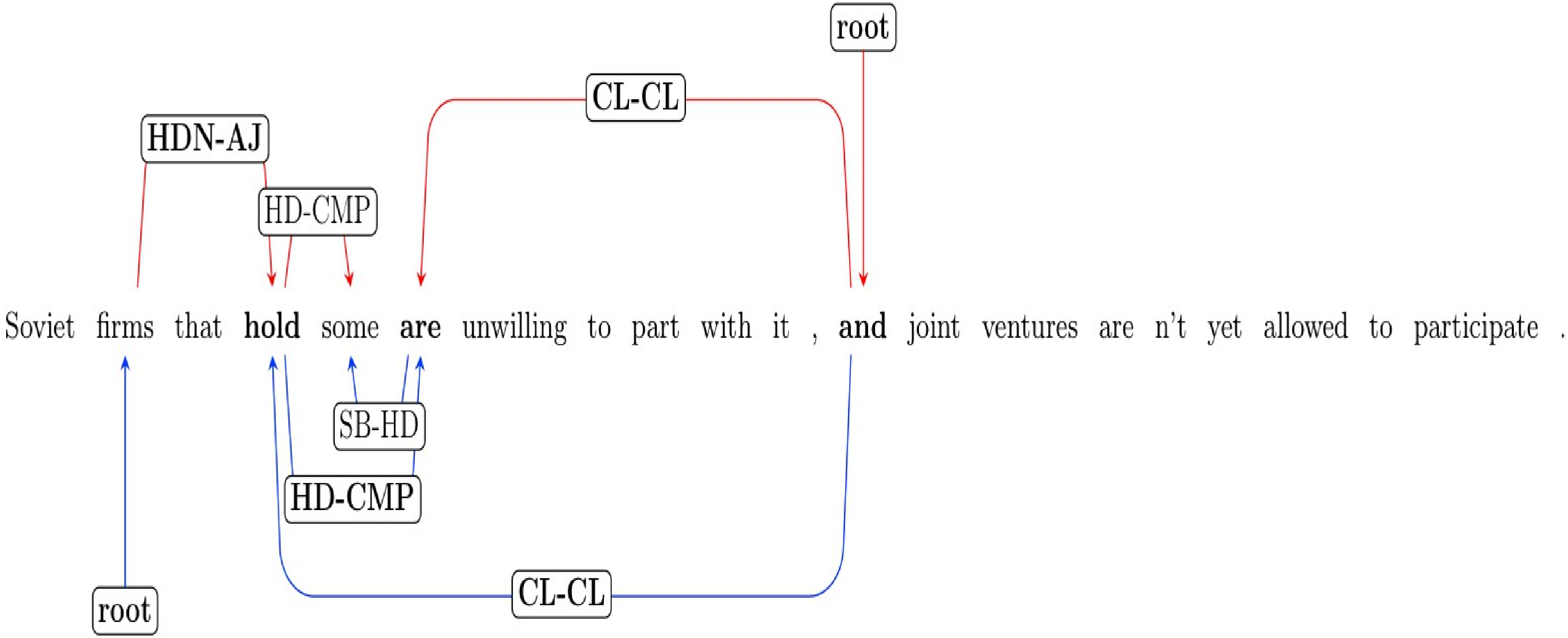


Error analysis

2) Verbs

Most common errors with verbs:

3) Erros with VBP mostly occur in the complex sentences that contain many verbs. Sentences with coordinating conjunction are particularly difficult for the correct attachment and labeling of the VBP.


root
CL-CL
HDN-AJ
HD-CMP
Soviet firms that hold some are unwilling to part with it , and joint ventures are n't yet allowed to participate .
SB-HD
HD-CMP
CL-CL
root

Parsers have best results on DT for the closed world classes:

- punctuation
- possessives
- determiners
- personal pronouns

Summary:

Parsers:

- the Bohnet and Nivre (2012) parser performs better than Malt and MST

Formats:

- DT appears to be a more difficult target dependency representation than SB and CD
- CD and DT are structurally closer to each other than SB and DT (Jaccard similarity), but it does not affect parsing results
- **Stanford Basic** is good for learning **dependency labels**
- **CoNLL** is good for learning **graph structure**

PoS Tags:

Stanford Basic and CoNLL are more correlated with PTB PoS tags;
DT is more correlated with supertags
Combination of PTB tags and supertags is beneficial for parsing Stanford, CoNLL and DT and parser have similar results for all 3 schemes in these configurations

Research questions:

- Do PCFG and dependency parsers perform better than PET on DeepBank data?
- Is PET less domain-sensitive than PCFG and dependency parsers ?

Related work:
Fowler and Penn (2010)
Plank and van Noord (2010)

Experiment setup:

- Format DeepBank derivation trees
    - into DT for the Bohnet and Nivre (2012) parser
    - into context-free trees for Berkeley parser
- Train PET, Berkeley and the Bohnet and Nivre (2012) parsers
- Test on non-annotated text with gold standard tokenization
- Convert outputs to DT
- Evaluate results with eval.pl

Data

Training set

sections 0-20 of **DeepBank** – 35,504 sentences

Test set

WSJ – Section 21 of DeepBank - 1,392 sent.

CB – Cathedral and Bazar (essay) – 598 sent.

SC01 – Part of the SemCore Corpus – 855 sent.

VM32 – Transcribed spoken dialogue (VerbMobil) – 949 sent.

ERG tokenization, supertags

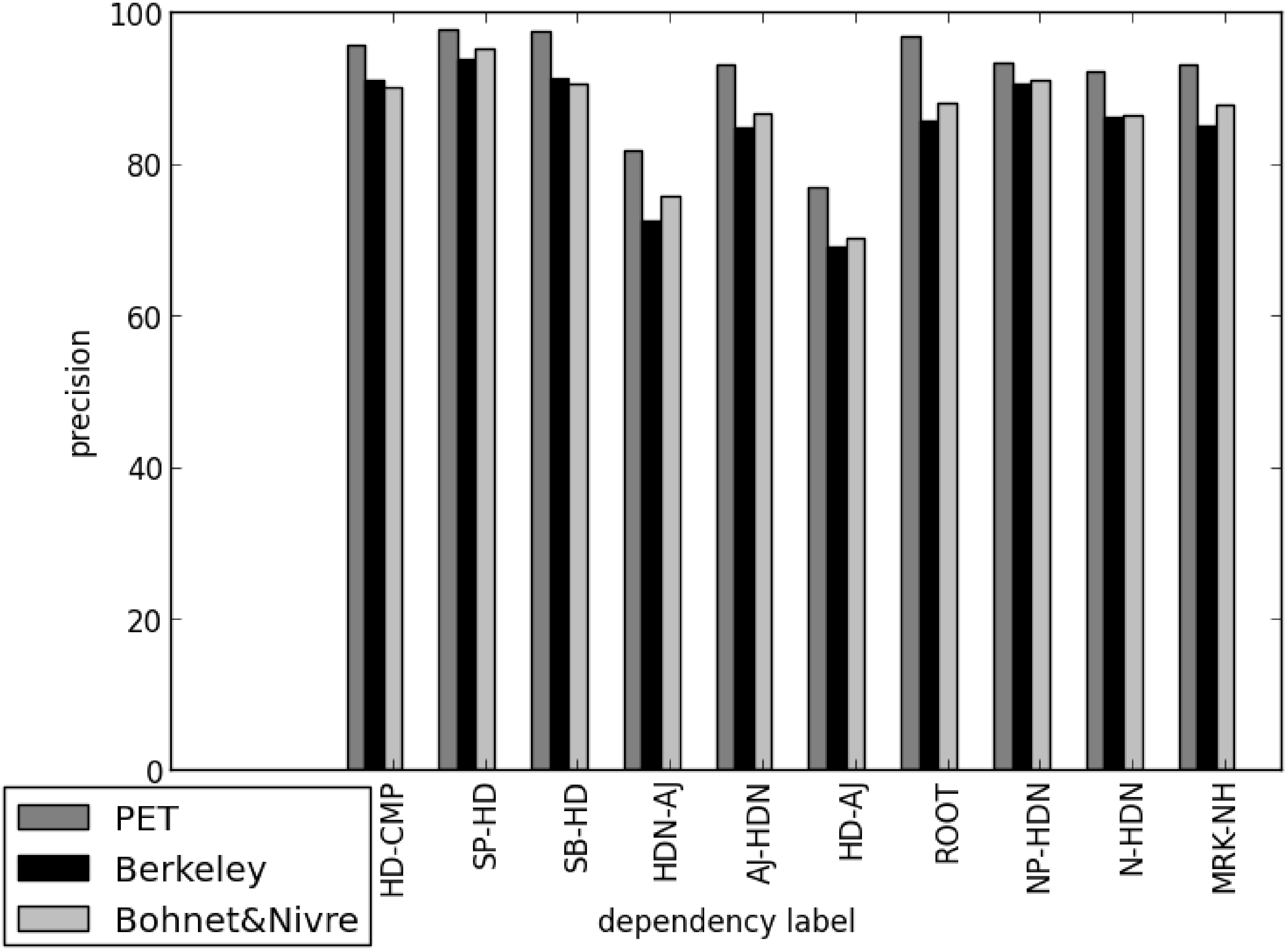
Data	PET	Bohnet and Nivre (2012)	Berkeley
wsj	86.85 / **93.09**	87.22	86.74
cb	**90.95**	78.32	79.46
sc01	**90.83**	77.47	80.48
vm32	**90.74**	76.96	77.97

Evaluation measure: labelled accuracy score (LAS)

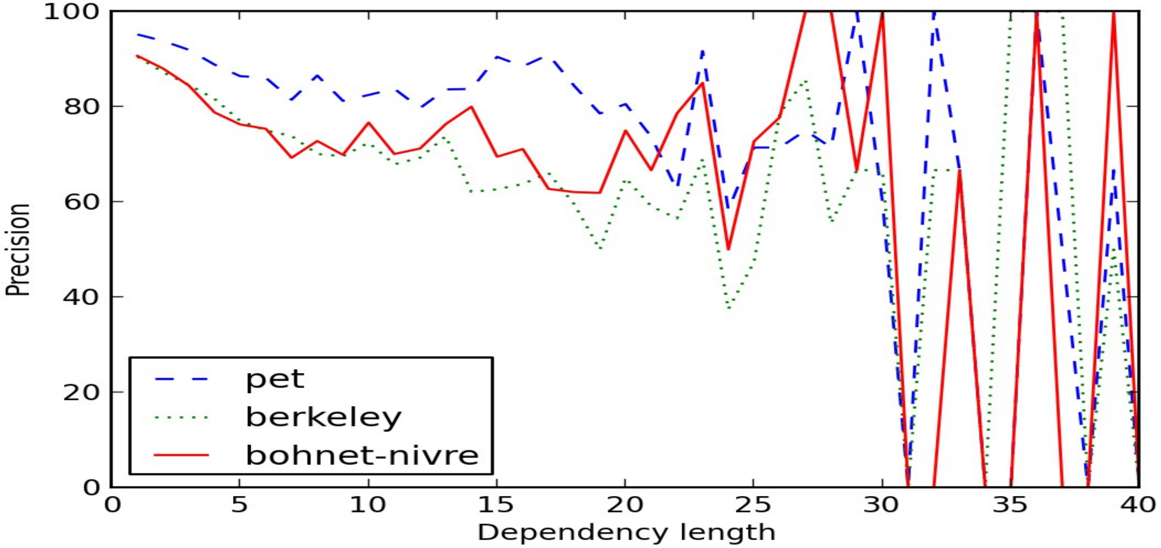
Coverage

Data	PET	Bohnet and Nivre (2012)	Berkeley
wsj	99.15%	**100%**	99.29%
cb	85.80%	**100%**	99.15%
sc01	92.00%	**100%**	99.15%
vm32	94.48%	**100%**	99.01%

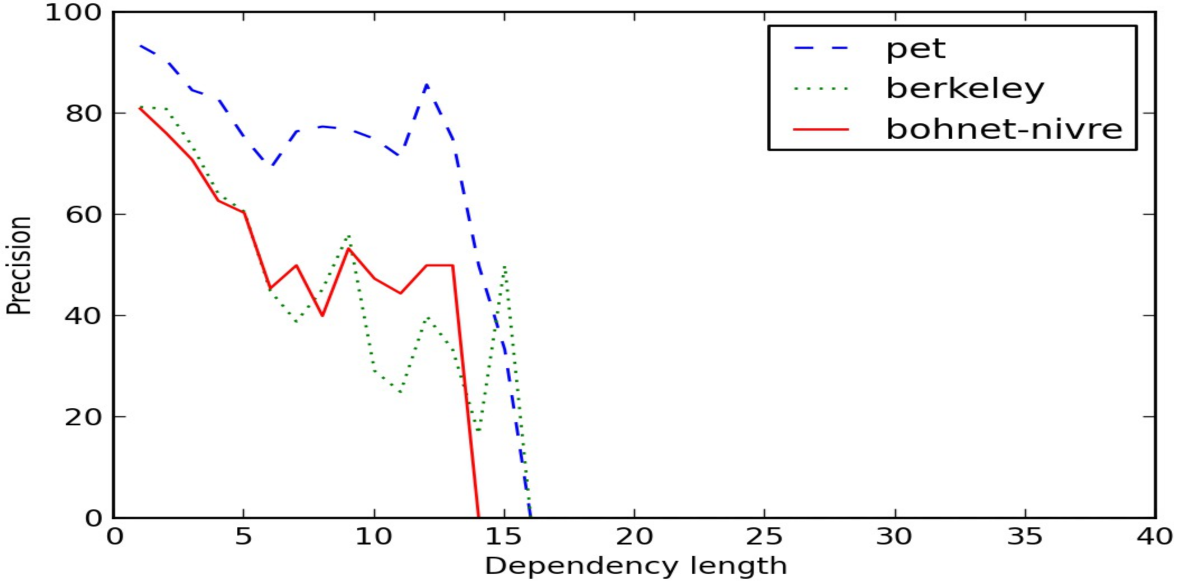
Error analysis: Dependency labels



Error analysis: Dependency length



WSJ



VM32

Error analysis: LAS over POS

- PET has better LAS over PoS than other parsers
- The Bohnet and Nivre (2012) parser has the lowest dependency label recall relative to gold dependency length on all the domains
- Coordination and constructions with prepositions are the hardest for parsers to analyze on all the domains.
- Determiners, complimentizers, verbs and nouns are among the easiest to analyze.

Error analysis: 10 best LAS over supertags

- PET the highest 10 best LAS over supertags
- The Bohnet and Nivre (2012) has the lowest 10 best LAS over supertags (except wsj domain).
- The drop of top accuracies due to the domain shifts is bigger for Berkeley and the Bohnet and Nivre (2012) parsers than for PET

Error analysis: 10 highest error rates over supertags

Error rates for PET are the lowest.

The total error rate is higher for Berkeley than for the Bohnet and Nivre (2012) parser on **wsj** domain and vice versa on the **sc01** domain.

Berkeley and the **Bohnet and Nivre (2012)** parser have the **same total error rate** on **cb** and **vm32** domains.

All parsers are error-prone on:

1) coordination conjunction ***and*** (``c_xp_and_le'')

2) ***for***, ***from*** and ***with*** (``p_np_i_le'')

3) ***in***, ***on***, ***at*** (``p_np_i-reg_le'' for spartial relationships, ``p_np_i-tmp_le'' for temporal relationships).

Conclusions:

- Contrary to expectations PET had higher LAS and UAS scores than Berkeley and the Bohnet and Nivre (2012) parsers

- Results give support to the statement that parsing with HPSG grammars is on the level of state-of-the art statistical parsing

- If we want to do parser combination

Malt, MST + PET

Or

Berkeley + PET

we can expect higher efficiency, but probably we won’t benefit in accuracy

Conclusions:

- All parsers experience the drop of accuracy due to domain shifts
- The more new domain differs from the training domain, the larger is the drop of accuracy for all parsers.
- PET is least sensitive to domain shifts in terms of accuracy, but the most sensitive in terms of coverage
- Bohnet and Nivre (2012) has 100% coverage on all the domains
- Berkeley is affected by domain shift both in terms of accuracy and coverage.

Thank you!