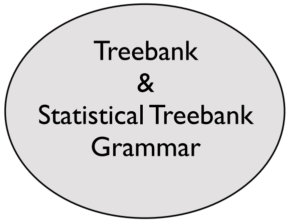
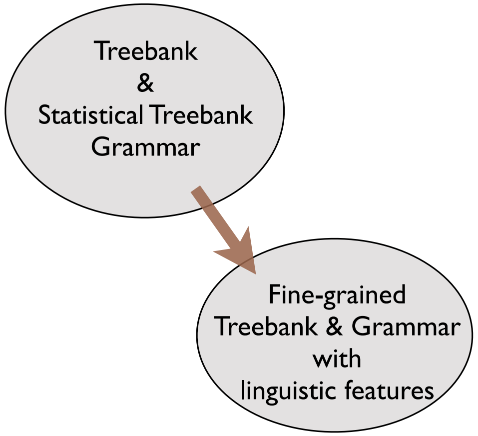
Fine-grained Valence acquisition for Treebank grammars from large corpora

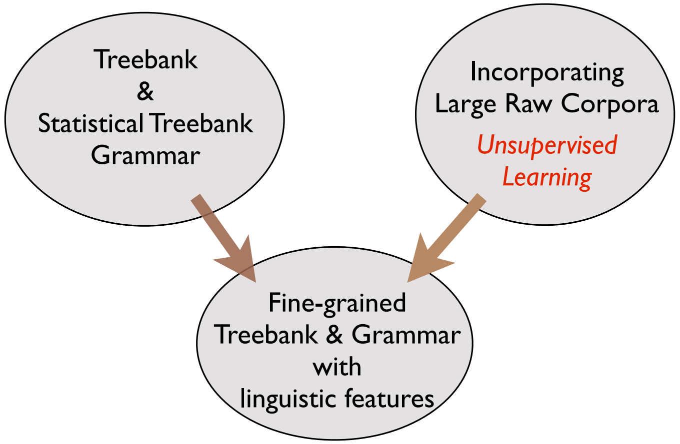
Tejaswini Deoskar
University of Edinburgh

Explorations in Syntactic Government and Subcategorisation
University of Cambridge
2 Sept. 2011

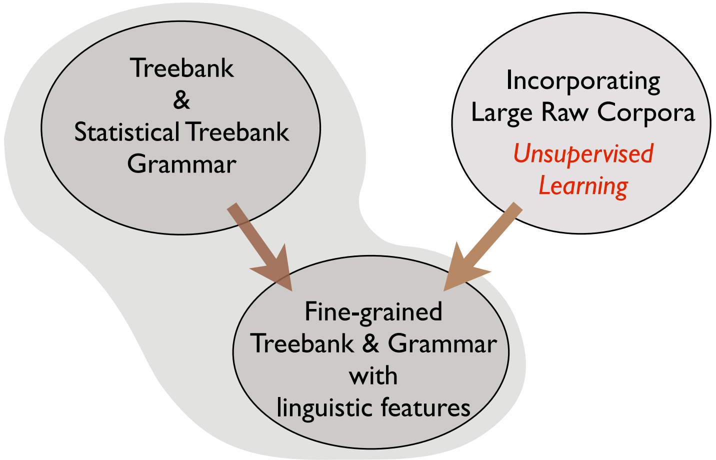




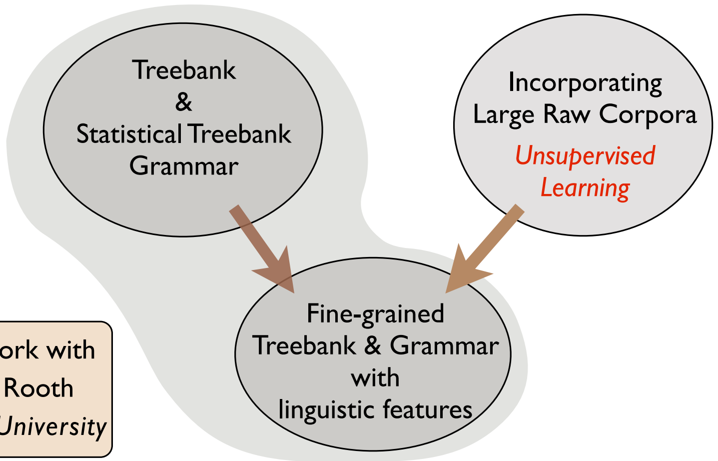
Fine-grained Valence acquisition for **treebank** grammars from **large corpora**



Fine-grained Valence acquisition for **treebank** grammars from **large corpora**

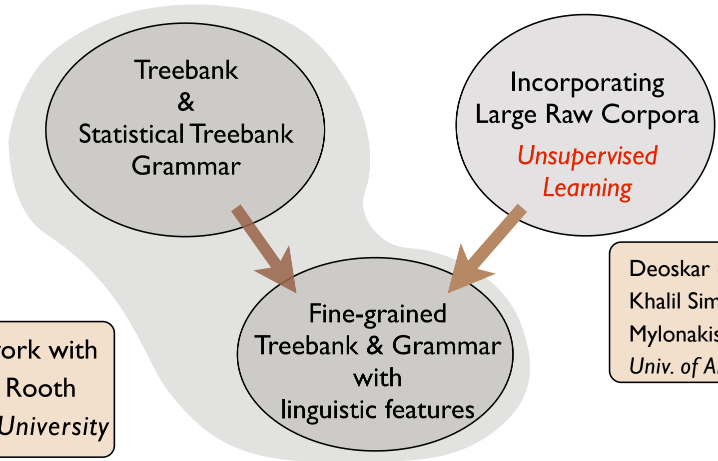


Fine-grained Valence acquisition for **treebank** grammars from **large corpora**



Joint work with
Mats Rooth
Cornell University

Fine-grained Valence acquisition for **treebank** grammars from **large corpora**



Joint work with
Mats Rooth
Cornell University

Deoskar (2009)
Khalil Sima'an, Markos Mylonakis
Univ. of Amsterdam

Outline

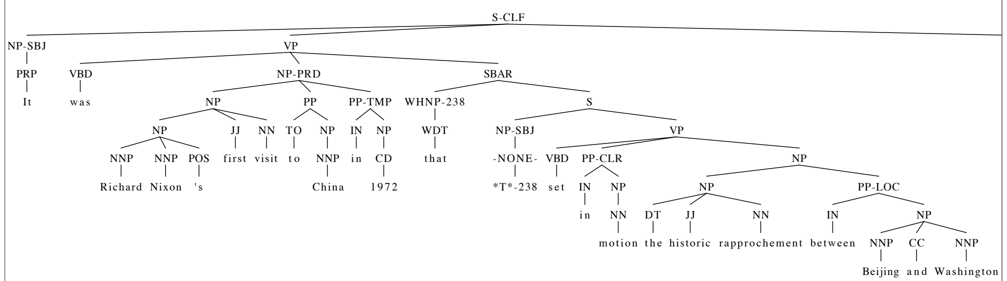
- Part I : A framework that will allow
  - addition of linguistically interesting features to existing treebank resulting in a more 'fine-grained' treebank
  - building statistical grammars parametrized on these features
- Part II : learn statistical tendencies of these features : connect to large amounts of data :
  - ➡ particularly relevant for phenomenon that is lexical in nature (e.g. valence)
  - ➡ evidence for these in treebank is sparse due to Zipfian distributions
- Evaluate utility of various features for learning

Example of Lexical Scarcity in Treebank data

- Penn Treebank (1 million sentences) contains about 7450 verb types (125,000 tokens)
  - ➡ 2830 have occurred only once (38% types)
  - ➡ 1034 have occurred twice (14% types)
- Thus not possible to obtain accurate statistical subcategorization tendencies for a large portion of lexicon

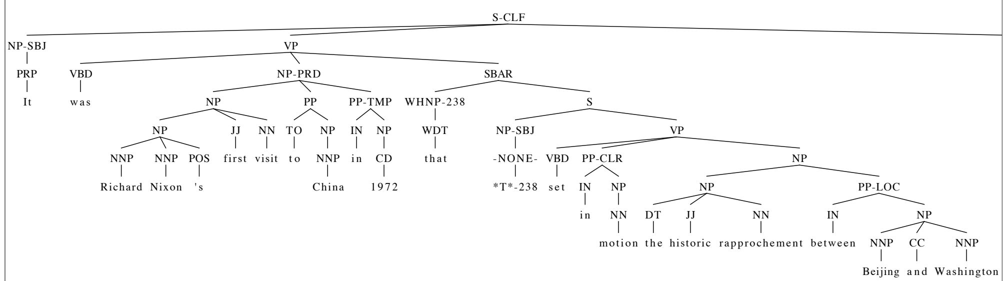
Treebanks

- Collections of sentences hand-annotated with linguistic structure
- Penn Treebank (Marcus et. al., 1993): 40, 000 Wall Street Journal sentences



Treebanks

- Treebank Grammar : extracted from a treebank
  - Both Symbolic and Probabilistic parts from Treebank
  - This talk : PCFGs (Probabilistic Context Free Grammars)



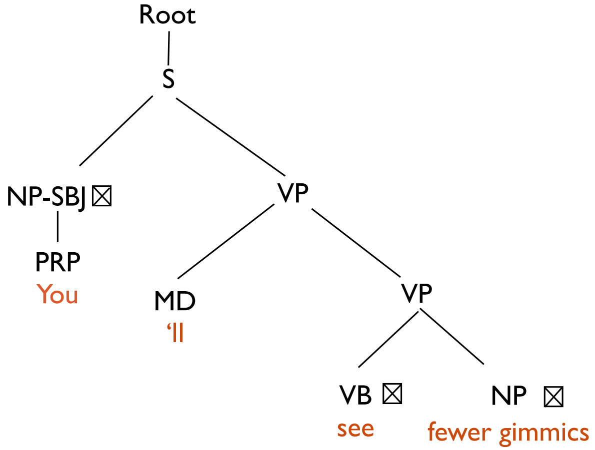
Treebanks

- Current treebanks contain coarse representations
  - Spurred research in statistical parsing
  - Allows for consistent and cheaper annotation
- Statistical grammars use coarse representations
  - statistics become very sparse if fine-grained
  - parsers even coarser than treebank
- For some aspects of linguistic research, and also high-end parsers
  - Fine-grained representations might be better
  - Overt representations of valence, agreement, and localising long-distance dependencies useful

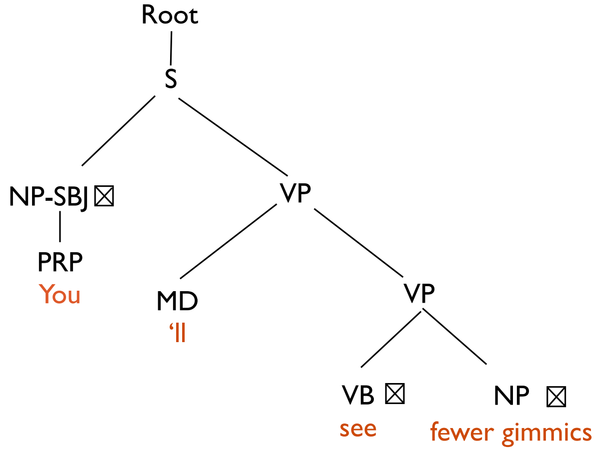
Outline of Methodology for Treebank refinement

1. Augment each node-label in tree with a feature-structure
   - ➡ feature-structures contain (typed) features with (atomic) values
2. values of features incorporated into node-label of tree
   - ➡ more fine-grained label

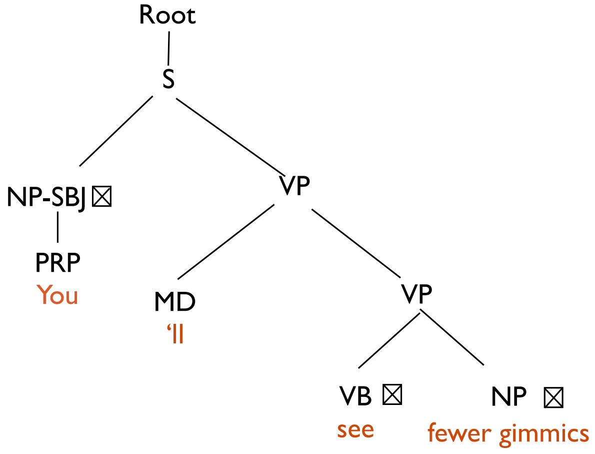
Augmenting Treebank Trees



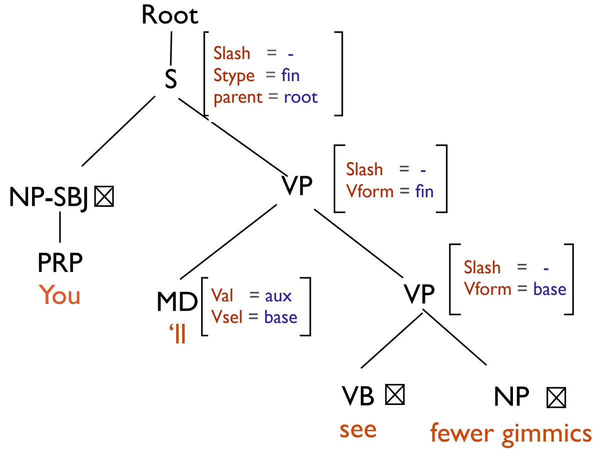
Augmenting Treebank Trees



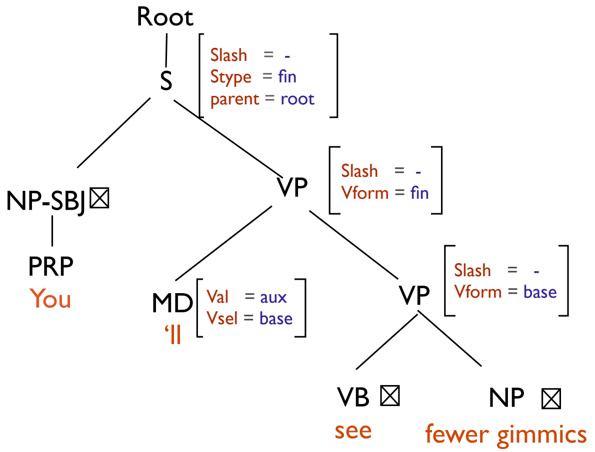
Step I : Tree augmented with feature-structures



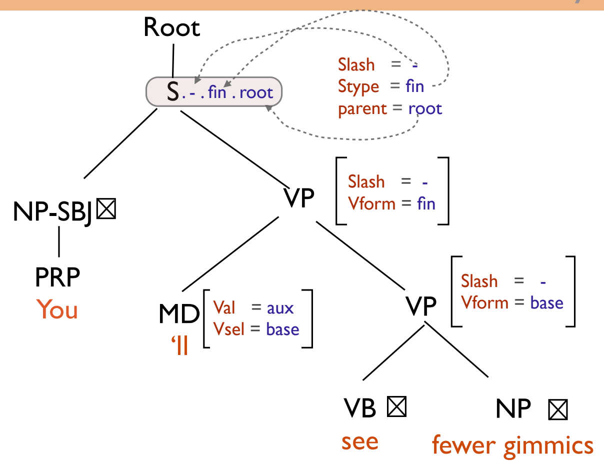
Step I : Tree augmented with feature-structures



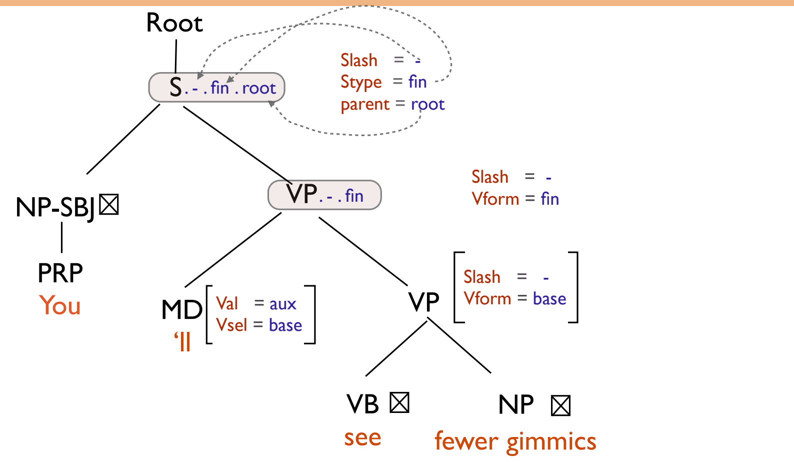
Step II : Convert features into context-free symbols



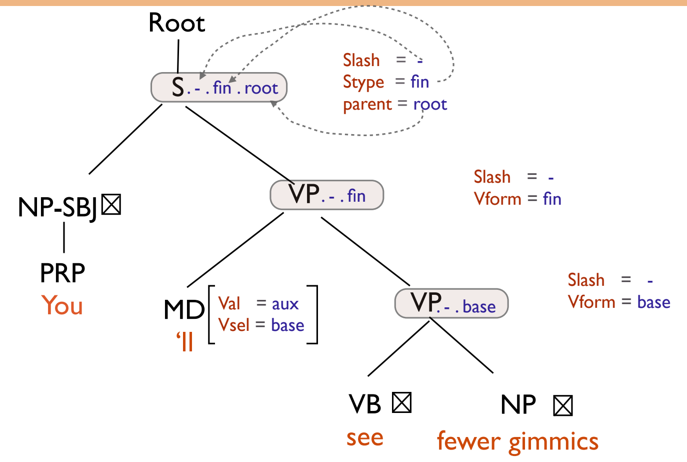
Step II : Convert features into context-free symbols



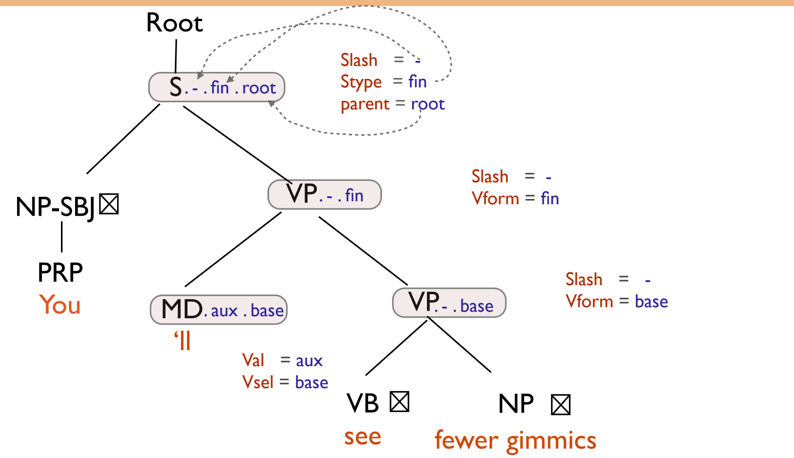
Step II : Convert features into context-free symbols



Step II : Convert features into context-free symbols



Step II : Convert features into context-free symbols



Implementation

- Parsing treebank trees with a Feature-constraint grammar
  - Details of implementation in Schmid (2000), Deoskar & Rooth (2008), Deoskar (2009)
- Highlights
  - ➡ Reusable software for constraint-solving, and PCFG compilation
  - ➡ Robust : In case of ambiguities, unit freq of tree split into fractions
- Effort required for grammar-development : Feature-constraint grammar
  - Intuitive for linguists
  - Difficult to manipulate existing parsers

PCFGs incorporating different features

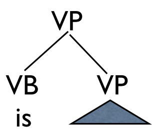
- For each category, stipulate the set of features to be incorporated into the PCFG.
  - allows PCFGs of various granularity to be built
  - empirically evaluate the utility of various features

Feature - Grammar development

- Example of Auxiliary construction
- Adding constraints requires checking treebank conventions

Feature - Grammar development

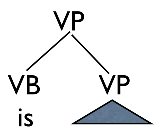
- Example of Auxiliary construction
- Adding constraints requires checking treebank conventions




```
VP { }    ->  VB { }   VP { }
```


Feature - Grammar development

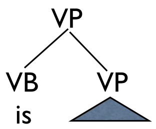
- Example of Auxiliary construction
- Adding constraints requires checking treebank conventions




```
VP { }    ->  VB { }   VP { }
VP { }    ->  VB { Val=aux } VP { }
```

Feature - Grammar development

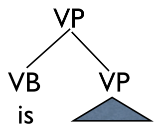
- Example of Auxiliary construction
- Adding constraints requires checking treebank conventions




```
VP { }      ->  VB { }    VP { }
VP { }      ->  VB { Val=aux } VP { }
VP {Vform = fin } -> VB { Val=aux } VP { }
```


Feature - Grammar development

- Example of Auxiliary construction
- Adding constraints requires checking treebank conventions

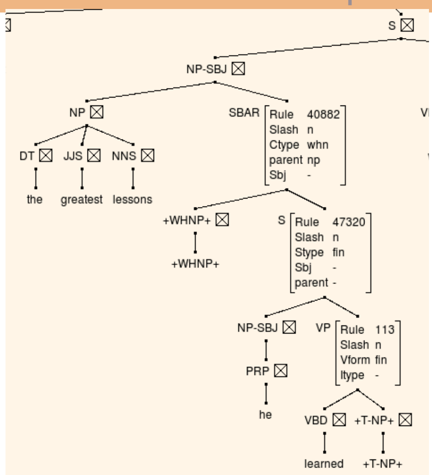



```
VP { }      ->  VB { }    VP { }
VP { }      ->  VB { Val=aux } VP { }
VP {Vform = fin } -> VB { Val=aux } VP { }
VP {Vform=fin; Slash= sl} -> VB {Val=aux} VP {Slash =sl}
```

GPSG-like

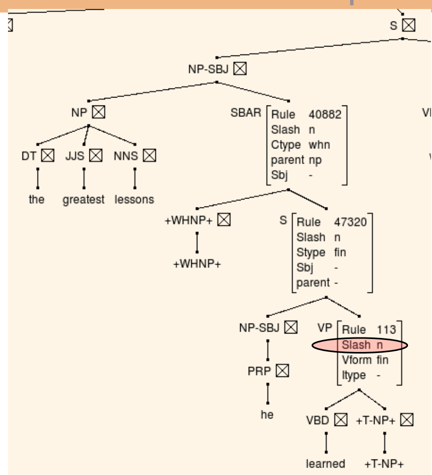
Valence

Slash feature for A-bar dependencies



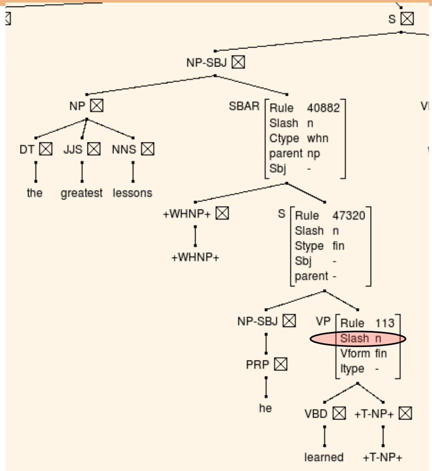
GPSG-like Slash feature for A-bar dependencies

Slash feature for A-bar dependencies



GPSG-like Slash feature for A-bar dependencies

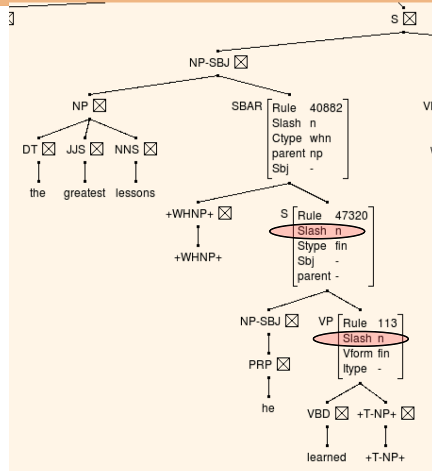
Slash feature for A-bar dependencies



GPSG-like Slash feature for A-bar dependencies

VP {Slash = n} -> VBD +T-NP+

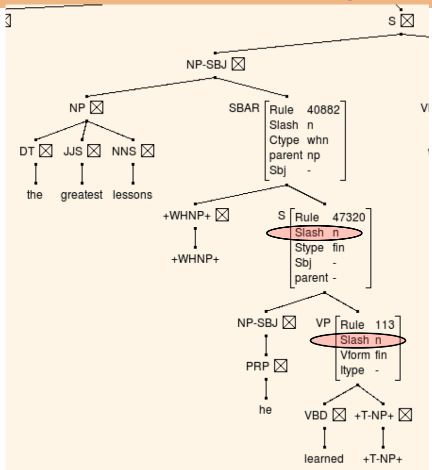
Slash feature for A-bar dependencies



GPSG-like Slash feature for A-bar dependencies

VP {Slash = n} -> VBD +T-NP+

Slash feature for A-bar dependencies

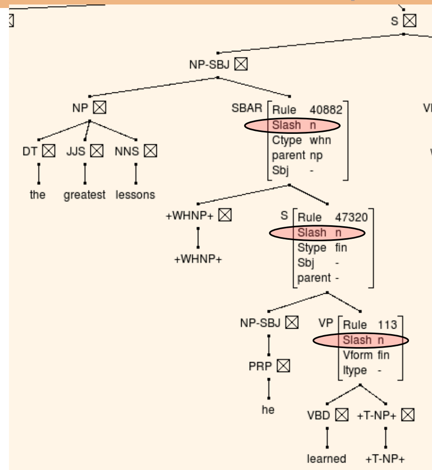


GPSG-like Slash feature for A-bar dependencies

S {Slash = sl} ->
    NP-SBJ VP {Slash = sl}

VP {Slash = n} -> VBD +T-NP+

Slash feature for A-bar dependencies

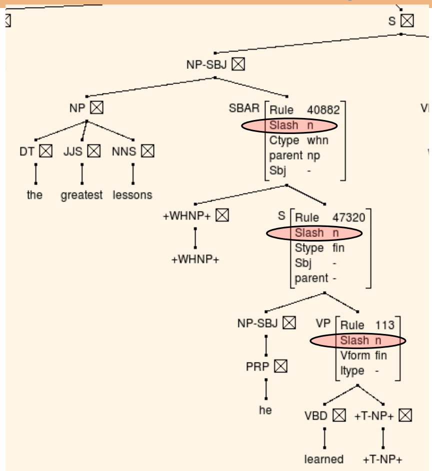


GPSG-like Slash feature for A-bar dependencies

S {Slash = sl} ->
    NP-SBJ VP {Slash = sl}

VP {Slash = n} -> VBD +T-NP+

Slash feature for A-bar dependencies



GPSG-like Slash feature for A-bar dependencies

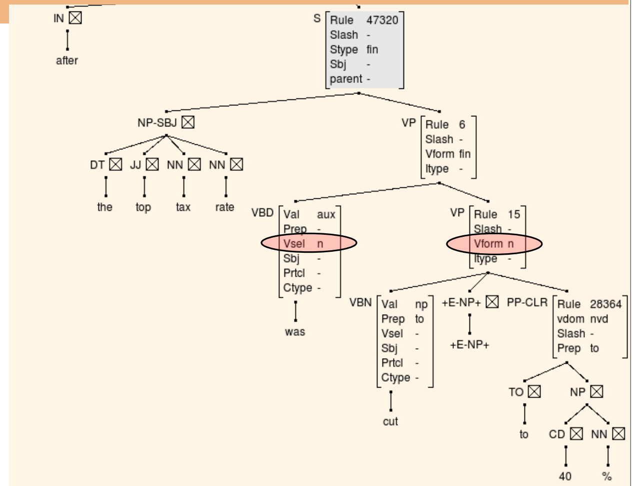
```
SBAR {Slash = sl} ->
        +WH-NP+  S{Slash = sl}

S {Slash = sl} ->
        NP-SBJ  VP{Slash = sl}

VP {Slash = n} ->  VBD +T-NP+
```


Passive

A-dependencies like passive and raising are effectively lexicalised

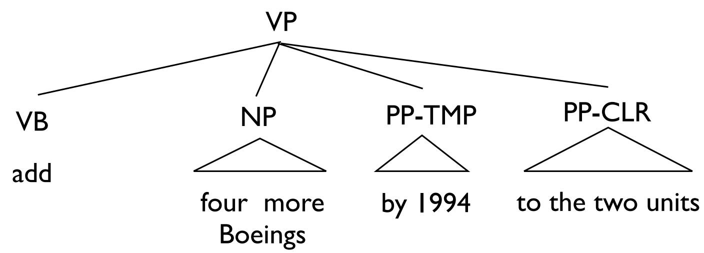


Lexical Features

For open class words

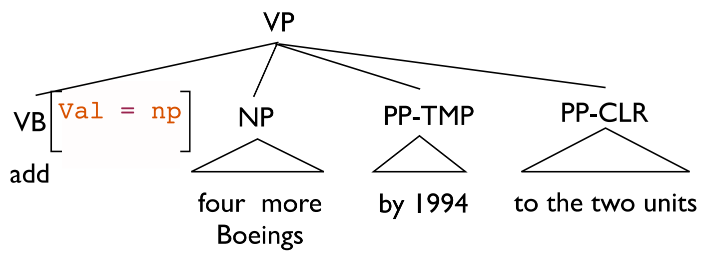
- add information about tree-shape onto pre-terminal label of word
  - ➡ For example: verbal valence
- could also be done for any lexico-syntactic dependencies other than valence
  - ➡ e.g. adverbial attachment to S, NP , VP nodes

Verb valence



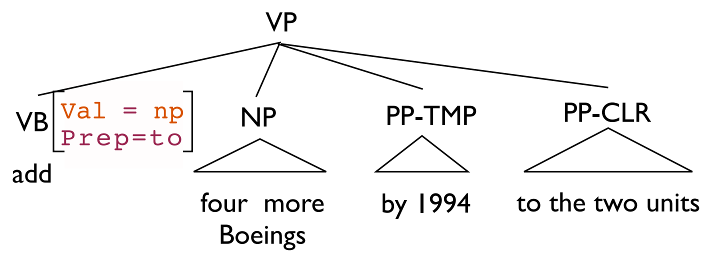
an NP- PP frame

Verb valence



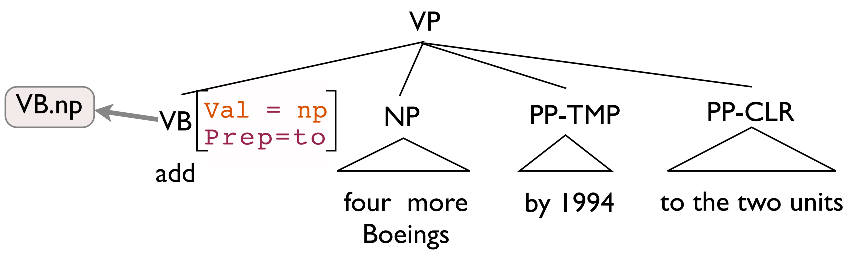
an NP- PP frame

Verb valence



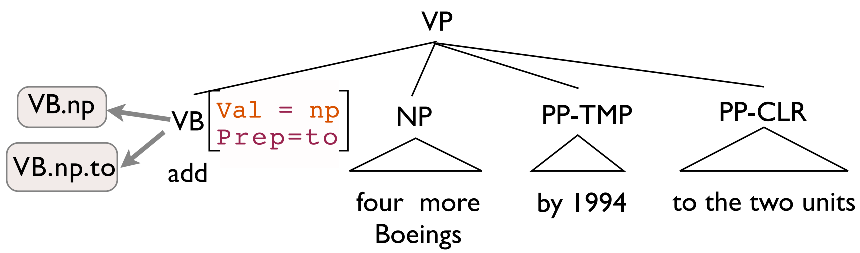
an NP- PP frame

Verb valence



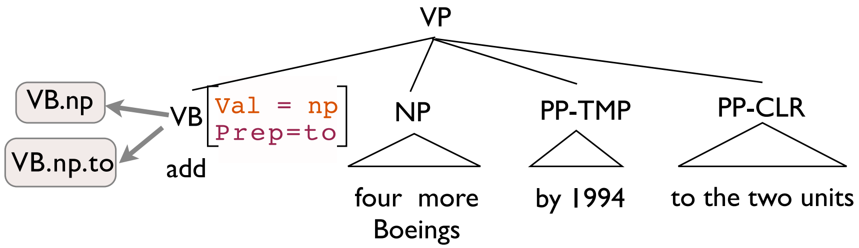
an NP- PP frame

Verb valence



an NP- PP frame

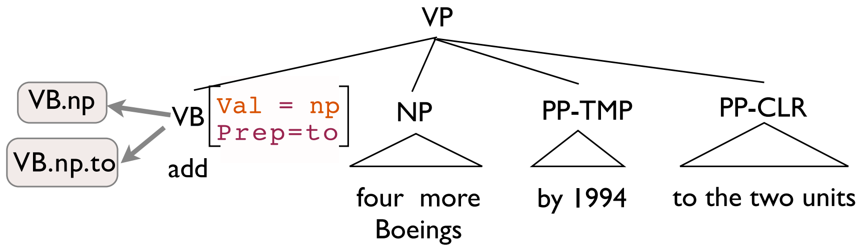
Verb valence



an NP- PP frame

31 basic frames

Verb valence



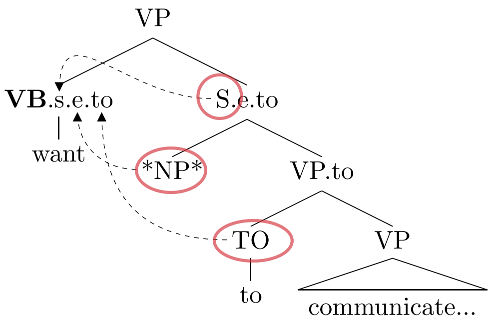
an NP- PP frame

31 basic frames

PRT	teamed up, moving in
PRT NP	taking on added risk, played himself out
PP SBAR	see to it that the kids don't play truant

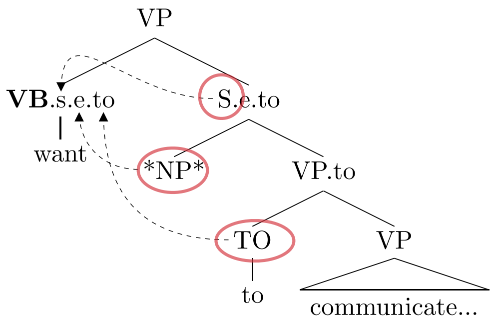
A control verb

- basic valence is sub-classified further for S complements



A control verb

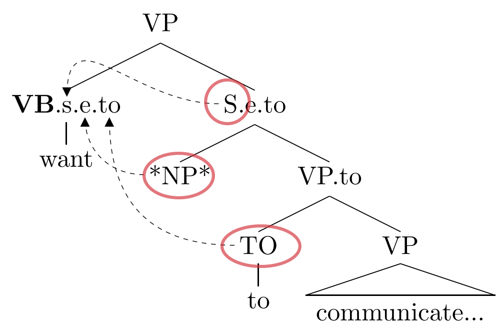
- basic valence is sub-classified further for S complements



Total 81 frames (without prepositions)

A control verb

- basic valence is sub-classified further for S complements



Total
81 frames
(without
prepositions)

set the economy moving again (non-empty subject and gerund)
wish to be a full time administrator (empty subject, predicative)
persuade consumers to pay more than $14...

Performance

- Treebank conversion

  Coverage: > 98.5 % of Treebank trees

  - Most ambiguities/failures due to remaining grammar bugs.

- PCFG
  - Labelled bracketing f-score: 86.8 % on Section 23 of the Penn Treebank
  - Competitive performance for English
    - Best results for empty category detection (84.1 %)

Lexical Entry for a verb in fine-grained PCFG

named	VBN 161	VBD 20

Original entry

New fine-grained entry

named	VBN.s.e.sc.-	118.0	VBN.n.-.-.-	20.0	VBN.np.-.-.to	15.0
	VBN.s.e.to.-	4.0	VBN.np.-.-.as	2.0	VBN.s.-.to.-	1.0
	VBN.np.-.-.for	1.0	VBD.s.-.sc.-	8.0	VBD.n.-.-.-	5.0
	VBD.np.-.-.as	4.0	VBD.np.-.-.to	1.0	VBD.s.-.to.-	2.0

Motivation for learning from unlabelled data

Most words have impoverished entries !!

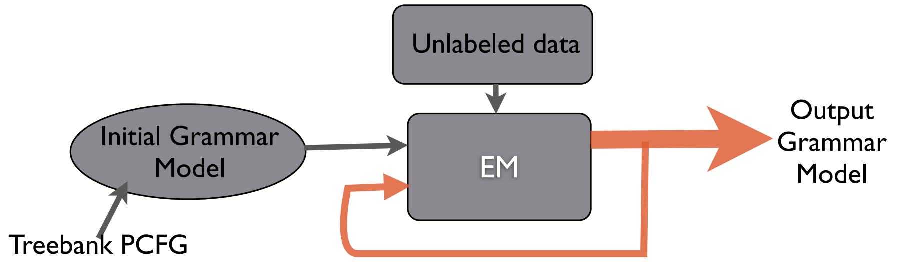
attaches	VBZ.np.-.- 1.0			
attack	NN 22	VBP.-.-.- 1.0	VB.n.-.- 3.0	VB.z.-.- 1.0
abandon	VBZ.n.-.- 2.0			
abate	VB.z.-.- 1.0			

Penn Treebank : 7450 verb types , 38% once, 14 % twice

Learning from unlabelled data

How?

- Expectation Maximization (EM)   (Dempster, et.al., 1977)
  - good mathematical properties, convergence



Learning from unlabelled data

Challenge : Unlabeled data tends to harm rather than help an already accurate model

- Constrain Unsupervised Model
  - Frequency transformations   Deoskar(2008, 2009)
  - N copies of Labelled data + unlabelled data   (To appear (2011), with Mylonakis, Sima'an)
    - More general method but worse results

Experimental Setup

- ~ 1 Million words from Penn Treebank
- 4, 8, 12, 16 Million words of unlabeled text (Wall Street Journal , sentence length < 25 words)
- Evaluations by parsing held-out sentences from the Penn Treebank
  - Task: assigning correct valence to verbs that are *unseen* in the labeled data.
    - 118 novel verb types, 1200 tokens
    - evaluated against the treebank tree

Valence Detection for Novel Verbs

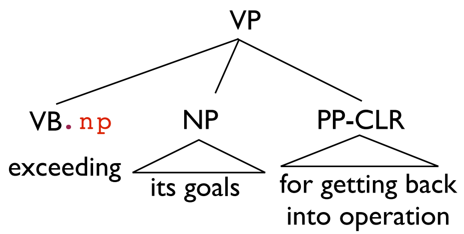
	4 M words	8 M words	12 M words	16 M words
Smoothed Treebank Model	29.86	29.86	29.86	29.86
Parsing Unlabeled Data	27.8	27.8	27.8	27.8
EM - based Method	27.08	25.89	25.18	24.7
% Error Reduction	9.31	12.76	15.67	17.5

No verb specific information

$p<0.0001$

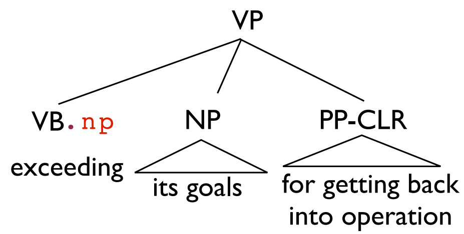
Valence Error Percentages for Novel Verbs

Improved PP attachment

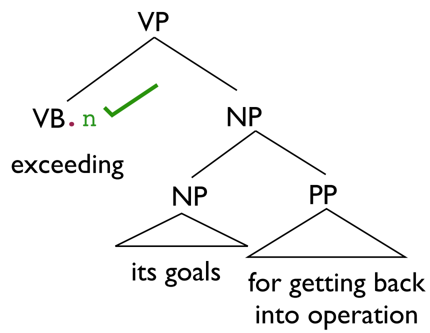


Wrong parse by Treebank grammar

Improved PP attachment

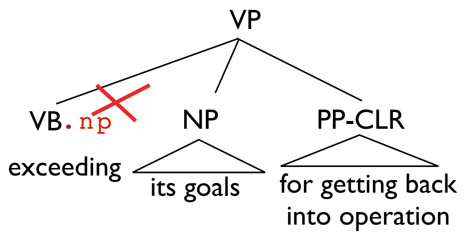


Wrong parse by Treebank grammar

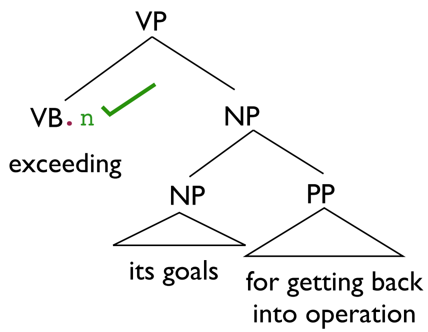


Correct parse by EM-trained grammar

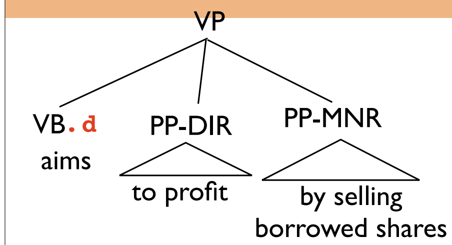
Improved PP attachment



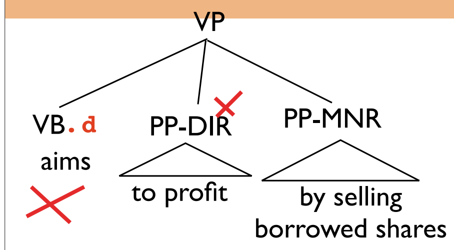
Wrong parse by Treebank grammar



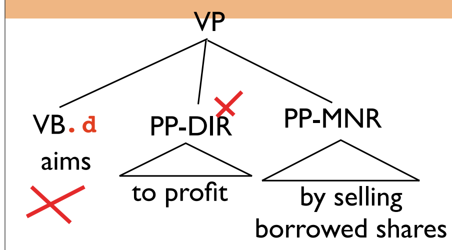
Correct parse by EM-trained grammar



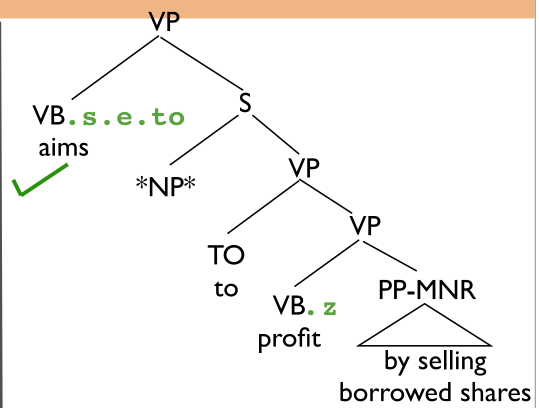
Wrong parse by Treebank grammar



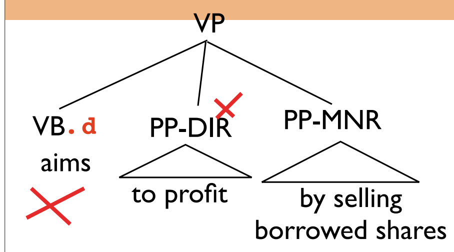
Wrong parse by Treebank grammar



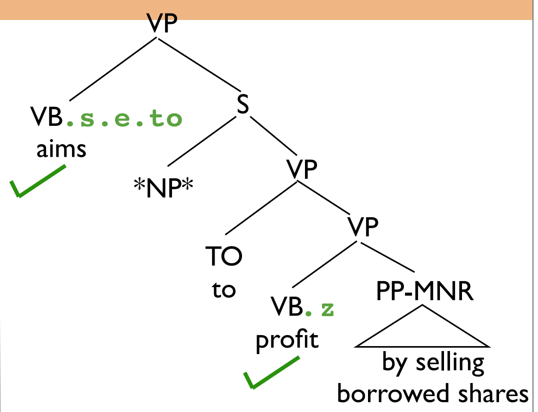
Wrong parse by Treebank grammar



Correct parse by EM-trained grammar

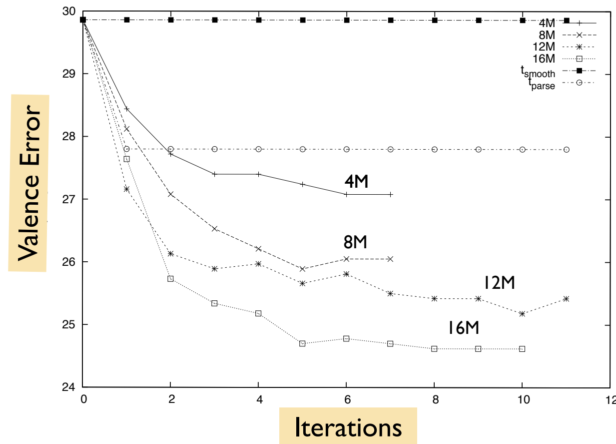


Wrong parse by Treebank grammar



Correct parse by EM-trained grammar

Learning Curves



Improvements in a variety of frame-types

Frame	% Error Reduction
transitive	21.52
intransitive	11.36
NP PP-CLR	7.14
PP-CLR	25
SBAR	0
s.e.to (control)	25
PRT NP	12.5
NP PP-DIR	14.28
NP NP	11.11

Other categories

- Improvements in Noun valence (but impoverished frames)
- Improvements in other lexico-syntactic dependencies: Adverb attachment to sentential, nominal, verbal nodes

Summary

Summary

- Framework
  - allows easy annotation of Treebank trees with feature-structures
  - compilation of PCFG grammars containing features
    - ➡ Effort required is in development of a feature-contraint grammar
    - ➡ PCFGs can be built containing various subsets of features

Summary

- Framework
  - allows easy annotation of Treebank trees with feature-structures
  - compilation of PCFG grammars containing features
    - Effort required is in development of a feature-contraint grammar
    - PCFGs can be built containing various subsets of features
- Connect to much larger data : Possible to improve the distributions of these features from unlabelled data (at least for some features, like valence)

Summary

- Framework
  - allows easy annotation of Treebank trees with feature-structures
  - compilation of PCFG grammars containing features
    - Effort required is in development of a feature-contraint grammar
    - PCFGs can be built containing various subsets of features
- Connect to much larger data : Possible to improve the distributions of these features from unlabelled data (at least for some features, like valence)
- Experiment with utility of various features for statistical grammar learning

Future Work

- Which features?
- Current grammar contains very few features: focus on features related to valence and constraining empty categories.
- Experiment with more features
  - Finer divisions of clausal valence: S and SBAR
- Fine-grained Treebank grammars for other languages.

Thank You!