Algorithms for NLP

Parsing VI

Taylor Berg-Kirkpatrick – CMU

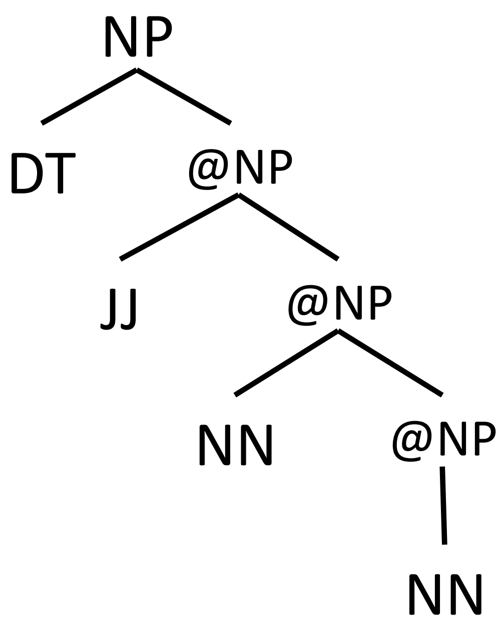
Slides: Dan Klein – UC Berkeley

P1 Shout-outs

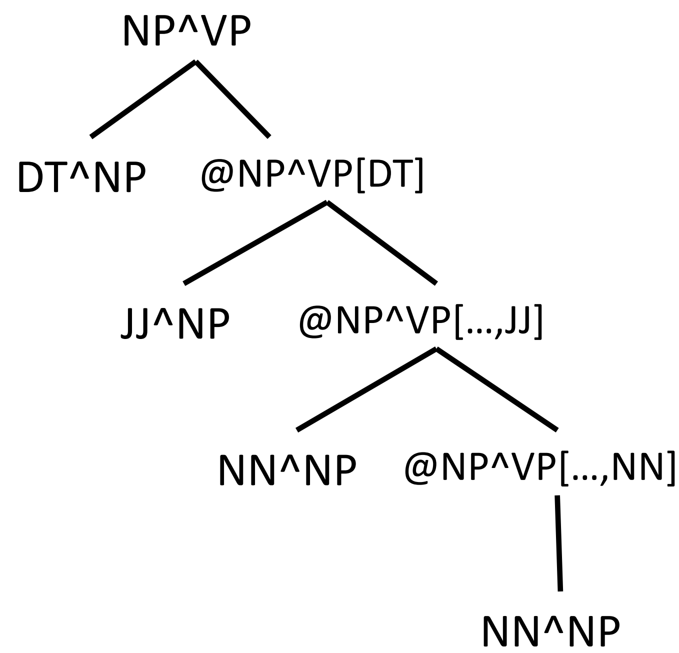
- **Saksham Singhal** -- implemented pseudo-tries. Used implicit caching (stored the most frequent n-grams on top of hash tables) and explicit caching.
- **Soumya Wadhwa, Tejas Nama** -- approximated by ignoring all trigrams with count 1. That dropped BLEU score by less than 0.1 only but freed half the memory!
- **Craig Stewart** -- rehash annealing idea. Made resizing factor and load factor change with every rehash to converge to 0.9 load factor to minimize wasted space.
- **Griffin Thomas Adams** -- Built a "waterfall" tiered cache system
- **Dean Alderucci** -- Built a class to pack data types of arbitrary size into an array of longs. Built a custom implementation of log that ran faster.
- **Robin Jonathan Algayres** – Context trie!
- **Raghuram Mandyam Annasamy** -- Used database inspired sharding technique on keys
- **Xianyang Chen** -- Compressed hash table and did smarter binary search by indexing chunks with the same last word
- **Aldrian Obaja** -- Implemented NestedMap, achieving 792 MB of memory.
- **Other things many people did** -- LRU caching, packing multiple values (counts and context fertilites) into a single long, binary search instead of hash table.

Grammar Projections

Coarse Grammar



Fine Grammar



NP → DT @NP

NP^VP → DT^NP @NP^VP[DT]

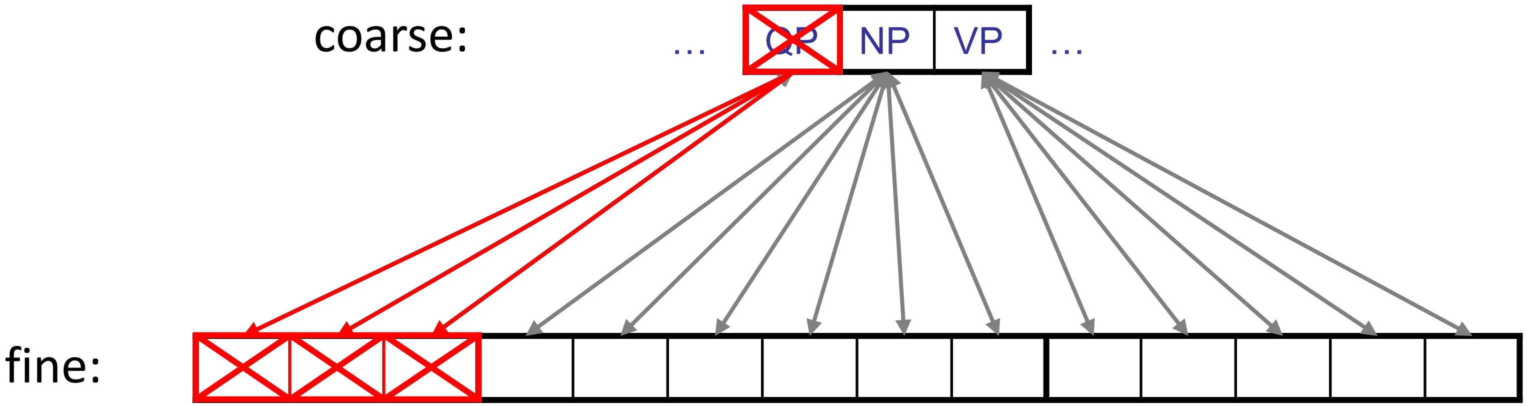
Note: X-Bar Grammars are projections with rules like XP → Y @X or XP → @X Y or @X → X

Efficient Parsing for Structural Annotation

Coarse-to-Fine Pruning

$$P(X|i, j, S) \quad < \quad \textbf{\textit{threshold}}$$

E.g. consider the span 5 to 12:

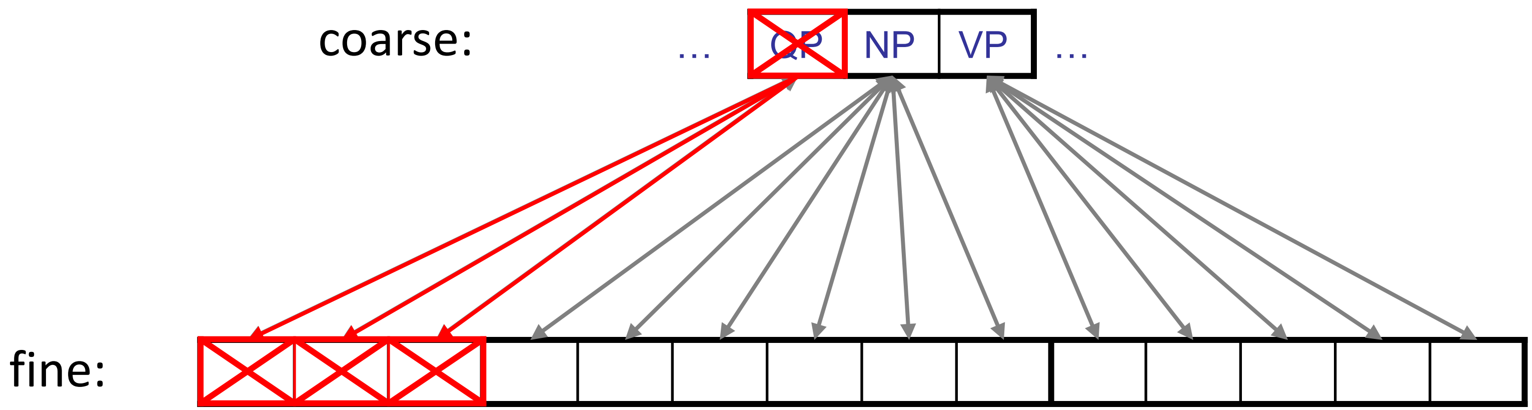


Coarse-to-Fine Pruning

For each coarse chart item $X[i,j]$, compute posterior probability:

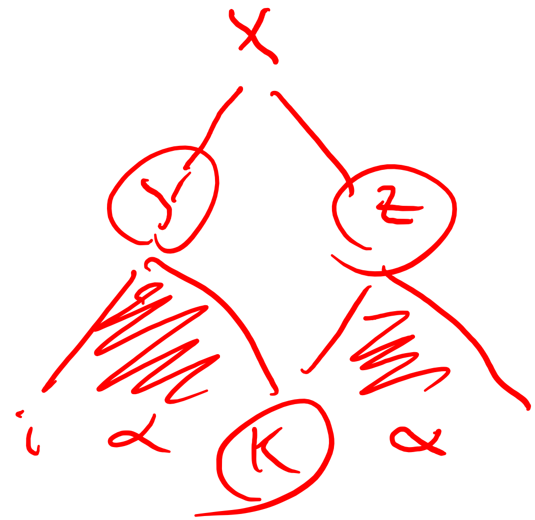
$$\frac{\alpha(X,i,j)\cdot\beta(X,i,j)}{\alpha(\text{root},0,n)} \quad < \quad \textbf{\textit{threshold}}$$

E.g. consider the span 5 to 12:



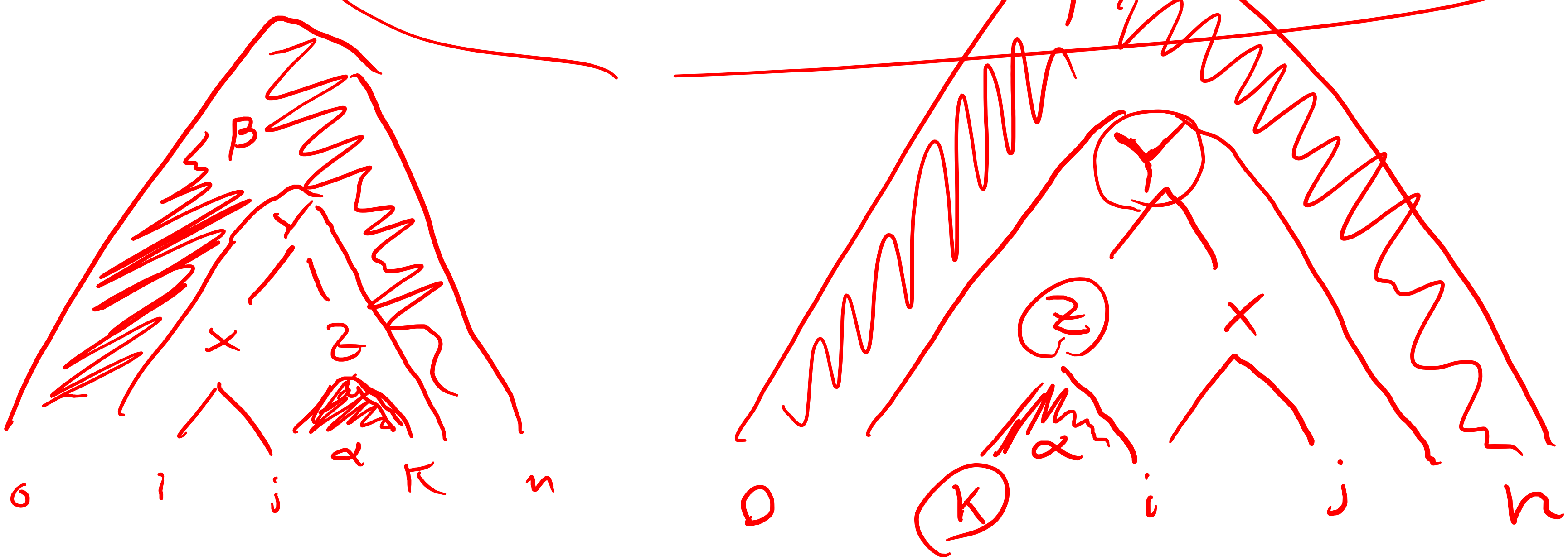
Computing Marginals

$$\alpha(X,i,j) = \overset{\max}{\cancel{\sum_{X\to YZ}}}\ \overset{\max}{\cancel{\sum_{k\in(i,j)}}} P(X\to YZ)\,\alpha(Y,i,k)\,\alpha(Z,k,j)$$



Computing Marginals

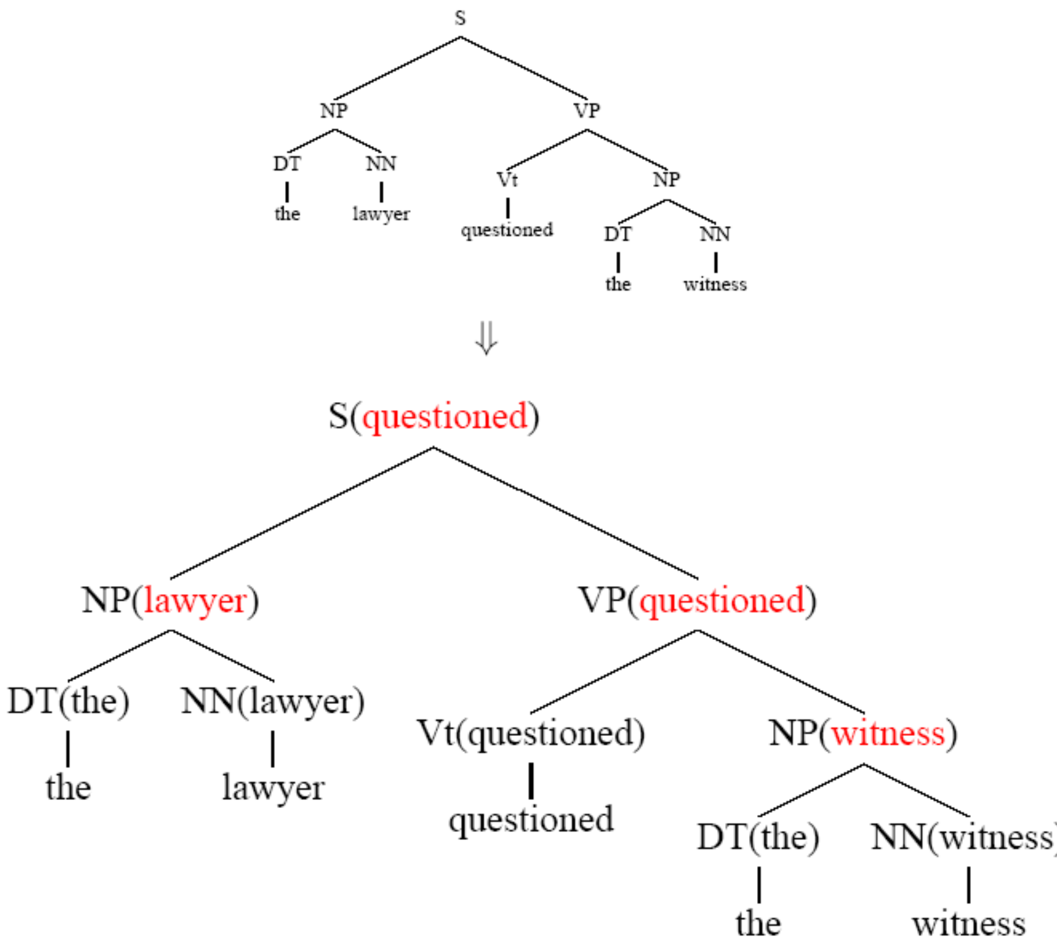
$$\beta(X,i,j) = \sum_{Y \to ZX} \sum_{k \in [0,i)} P(Y \to ZX)\beta(Y,k,j)\alpha(\overset{Z}{\cancel{R}},k,i)$$
$$+ \sum_{Y \to XZ} \sum_{k \in (j,n]} P(Y \to XZ)\beta(Y,i,k)\alpha(Z,j,k)$$



Efficient Parsing for Lexical Grammars

Lexicalized Trees

- Add “head words” to each phrasal node
  - Syntactic vs. semantic heads
  - Headship not in (most) treebanks
  - Usually *use head rules*, e.g.:
    - NP:
      - Take leftmost NP
      - Take rightmost N*
      - Take rightmost JJ
      - Take right child
    - VP:
      - Take leftmost VB*
      - Take leftmost VP
      - Take left child



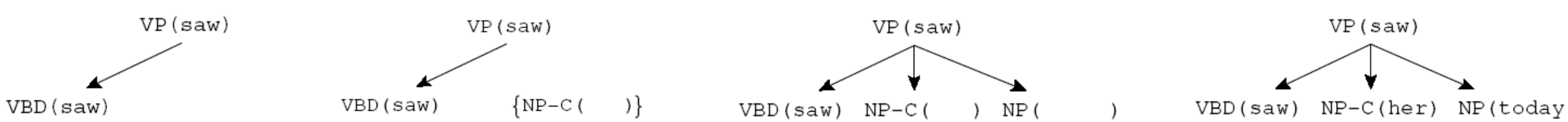
Lexicalized PCFGs?

- Problem: we now have to estimate probabilities like

```
VP(saw) -> VBD(saw) NP-C(her) NP(today)
```

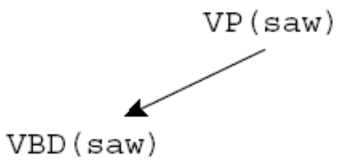
- Never going to get these atomically off of a treebank

- Solution: break up derivation into smaller steps

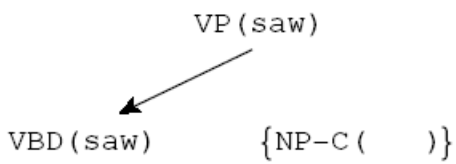


Lexical Derivation Steps

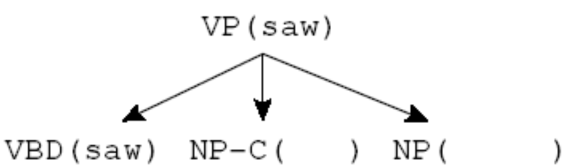
- A derivation of a local tree [Collins 99]



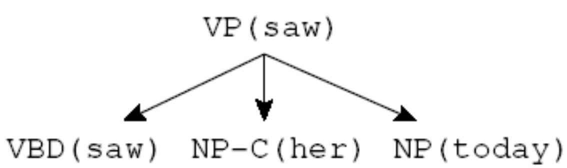
Choose a head tag and word



Choose a complement bag



Generate children (incl. adjuncts)

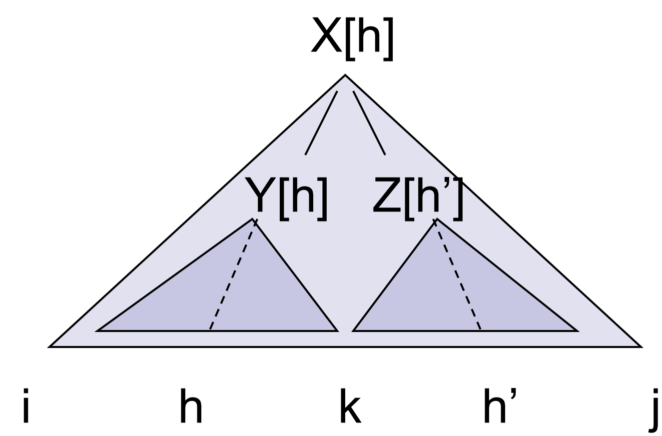


Recursively derive children

Lexicalized CKY

(VP->VBD...NP •)[saw]

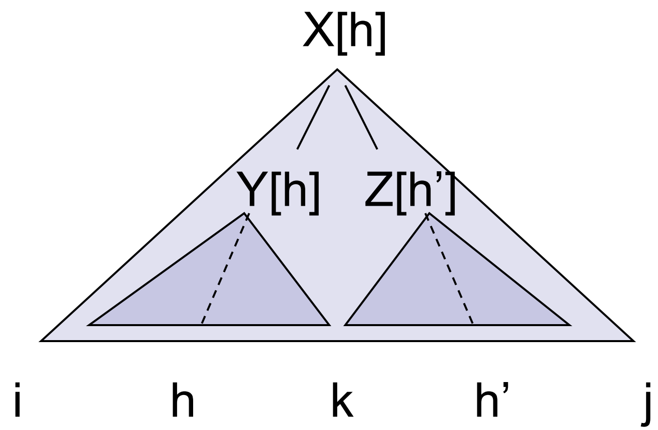
(VP->VBD •)[saw]   NP[her]

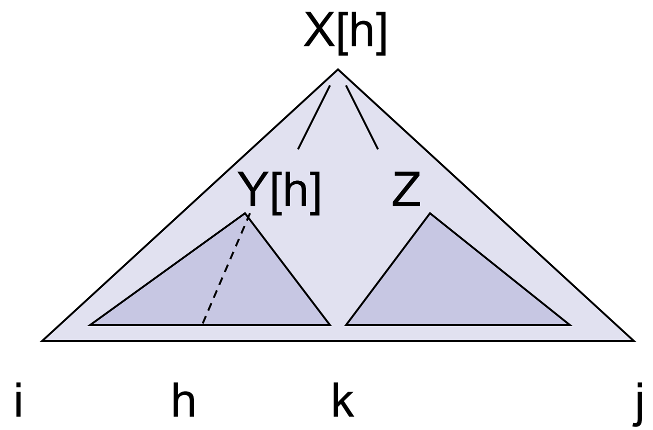



```
bestScore(X,i,j,h)
  if (j = i+1)
    return tagScore(X,s[i])
  else
    return
      max   max         score(X[h]->Y[h] Z[h']) *
          k,h',X->YZ
                        bestScore(Y,i,k,h) *
                        bestScore(Z,k,j,h')
            max         score(X[h]->Y[h'] Z[h]) *
          k,h',X->YZ
                        bestScore(Y,i,k,h') *
                        bestScore(Z,k,j,h)
```

Quartic Parsing

- Turns out, you can do (a little) better [Eisner 99]

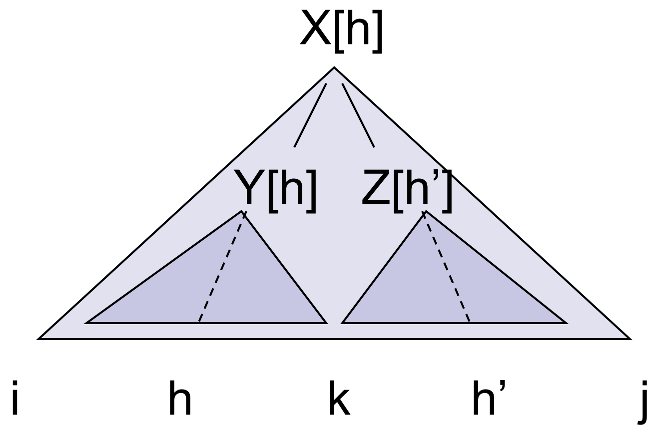




- Gives an $O(n^4)$ algorithm
- Still prohibitive in practice if not pruned

Pruning with Beams

- The Collins parser prunes with per-cell beams [Collins 99]
  - Essentially, run the $O(n^5)$ CKY
  - Remember only a few hypotheses for each span <i,j>.
  - If we keep K hypotheses at each span, then we do at most $O(nK^2)$ work per span (why?)
  - Keeps things more or less cubic (and in practice is more like linear!)



- Also: certain spans are forbidden entirely on the basis of punctuation (crucial for speed)

Pruning with a PCFG

- The Charniak parser prunes using a two-pass, coarse-to-fine approach [Charniak 97+]
  - First, parse with the base grammar
  - For each X:[i,j] calculate P(X|i,j,s)
    - This isn't trivial, and there are clever speed ups
  - Second, do the full $O(n^5)$ CKY
    - Skip any X :[i,j] which had low (say, < 0.0001) posterior
  - Avoids almost all work in the second phase!

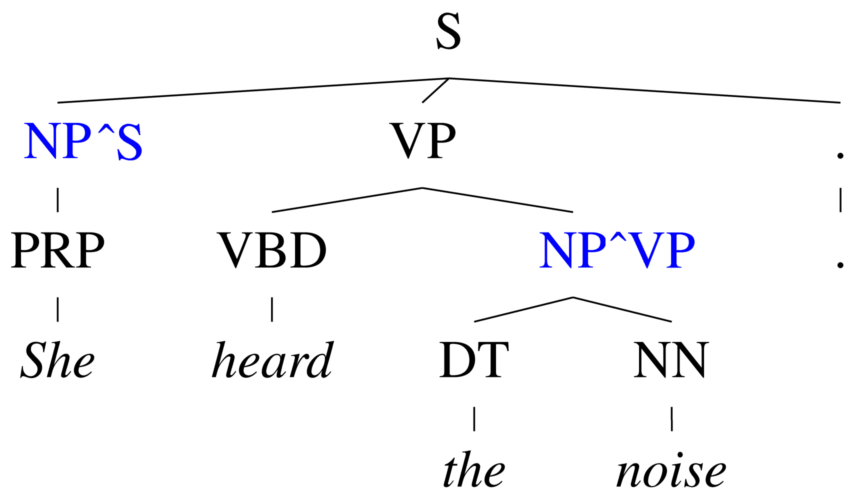
- Charniak et al 06: can use more passes
- Petrov et al 07: can use many more passes

Results

- Some results
  - Collins 99 – 88.6 F1 (generative lexical)
  - Charniak and Johnson 05 – 89.7 / 91.3 F1 (generative lexical / reranked)
  - Petrov et al 06 – 90.7 F1 (generative unlexical)
  - McClosky et al 06 – 92.1 F1 (gen + rerank + self-train)

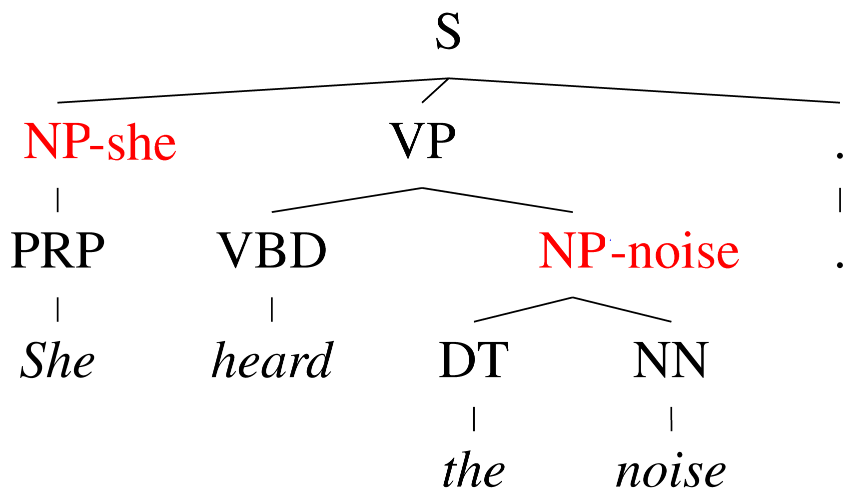
Latent Variable PCFGs

The Game of Designing a Grammar



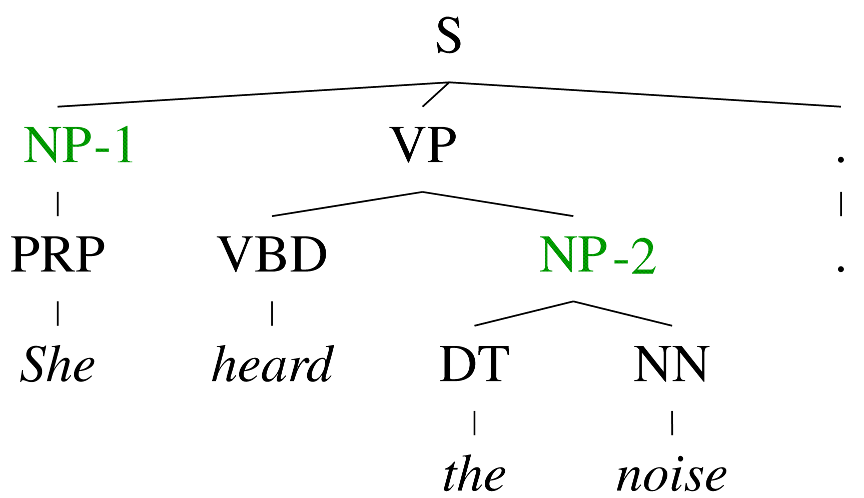
- Annotation refines base treebank symbols to improve statistical fit of the grammar
  - Parent annotation [Johnson ’98]

The Game of Designing a Grammar



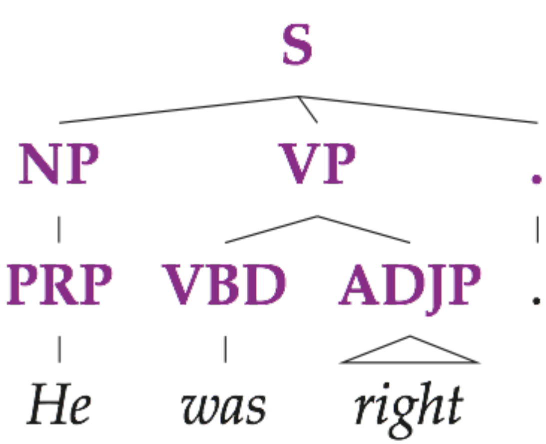
- Annotation refines base treebank symbols to improve statistical fit of the grammar
  - Parent annotation [Johnson '98]
  - Head lexicalization [Collins '99, Charniak '00]

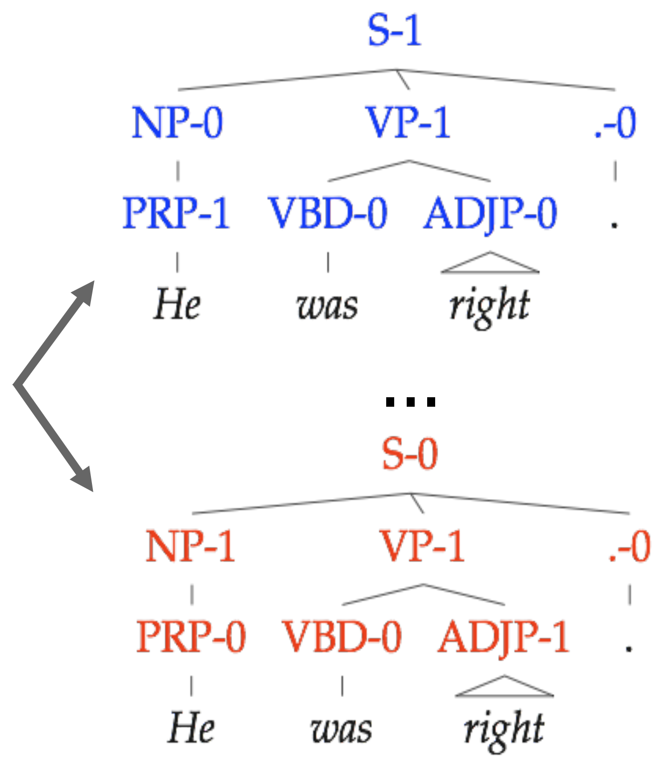
The Game of Designing a Grammar



- Annotation refines base treebank symbols to improve statistical fit of the grammar
  - Parent annotation [Johnson ’98]
  - Head lexicalization [Collins ’99, Charniak ’00]
  - Automatic clustering?

Latent Variable Grammars





Grammar G

$S_0 \rightarrow NP_0\ VP_0$?
$S_0 \rightarrow NP_1\ VP_0$?
$S_0 \rightarrow NP_0\ VP_1$?
$S_0 \rightarrow NP_1\ VP_1$?
$S_1 \rightarrow NP_0\ VP_0$?
...
$S_1 \rightarrow NP_1\ VP_1$?
...
$NP_0 \rightarrow PRP_0$?
$NP_0 \rightarrow PRP_1$?
...

Lexicon

$PRP_0 \rightarrow$ She ?
$PRP_1 \rightarrow$ She ?
...
$VBD_0 \rightarrow$ was ?
$VBD_1 \rightarrow$ was ?
$VBD_2 \rightarrow$ was ?
...

Parse Tree T
Sentence w

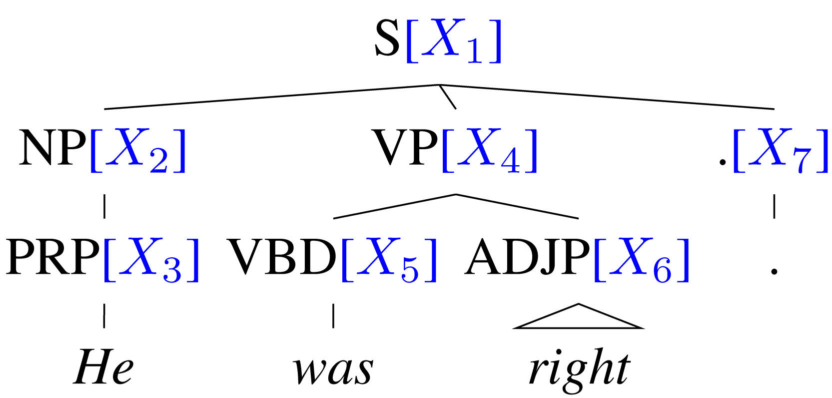
Derivations $t : T$

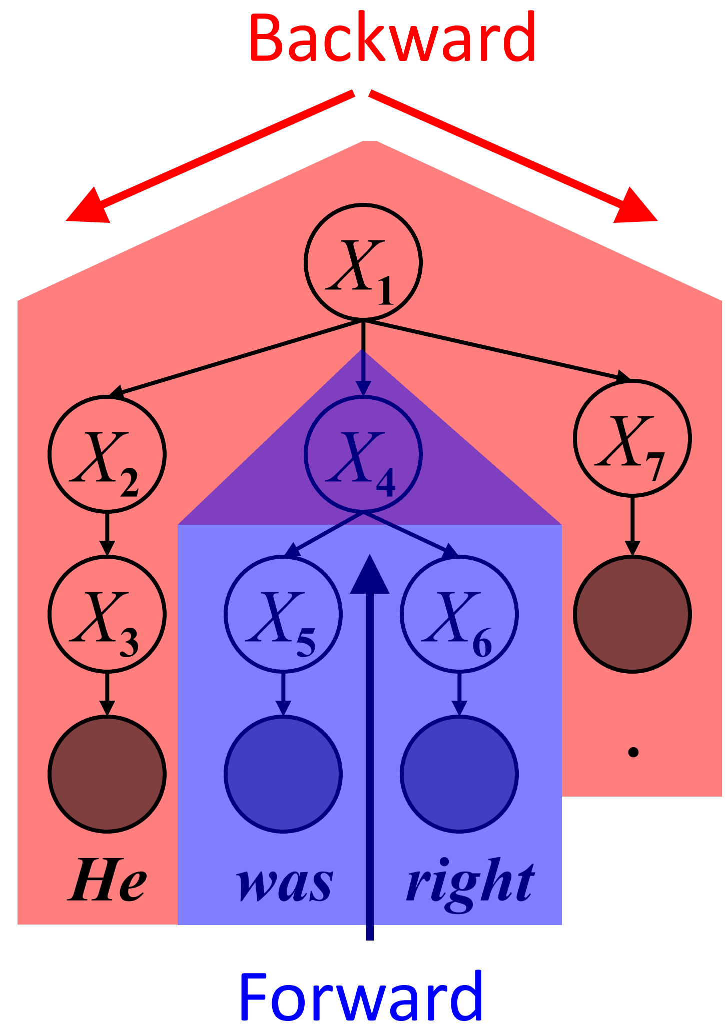
Parameters θ

Learning Latent Annotations

EM algorithm:

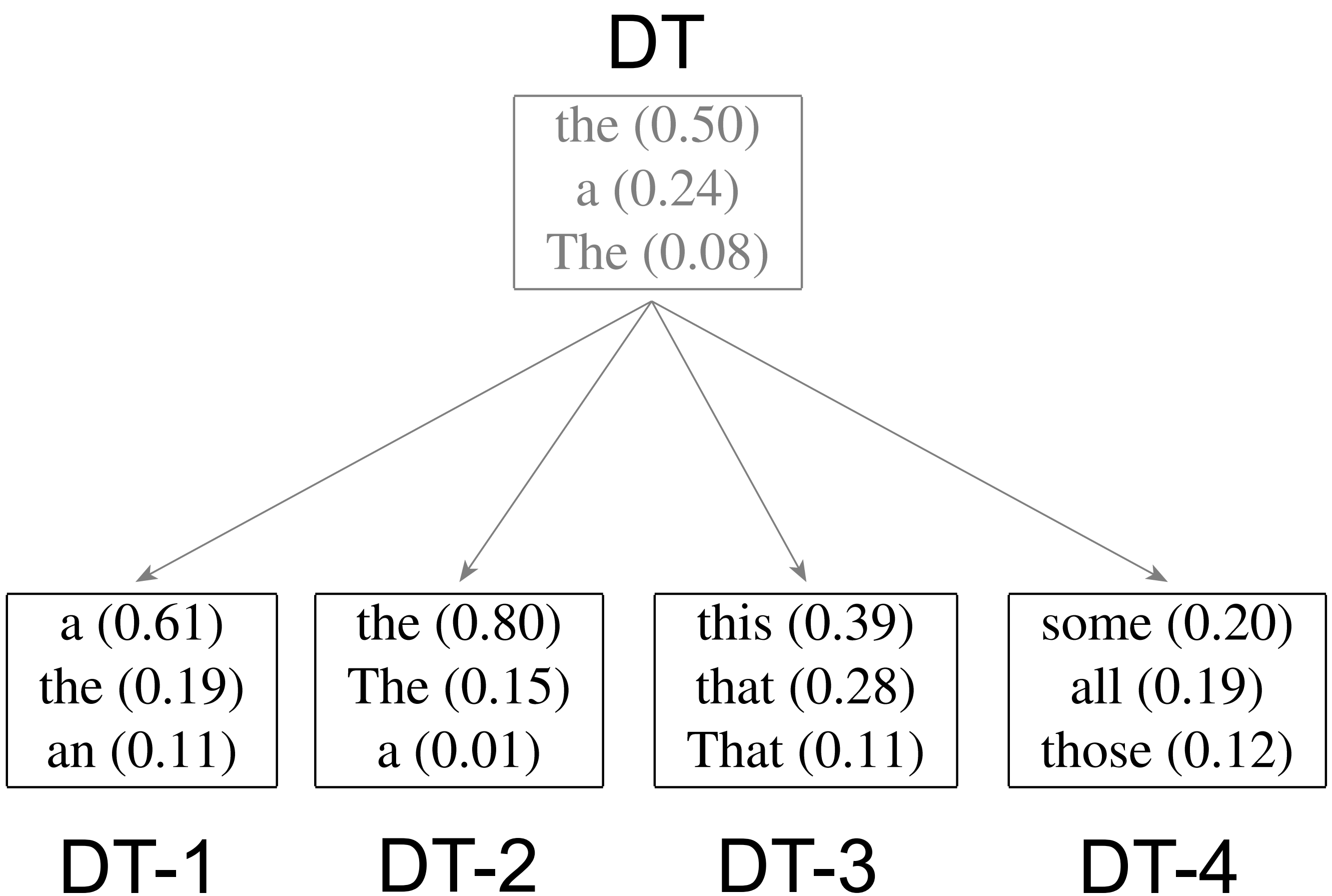
- Brackets are known
- Base categories are known
- Only induce subcategories



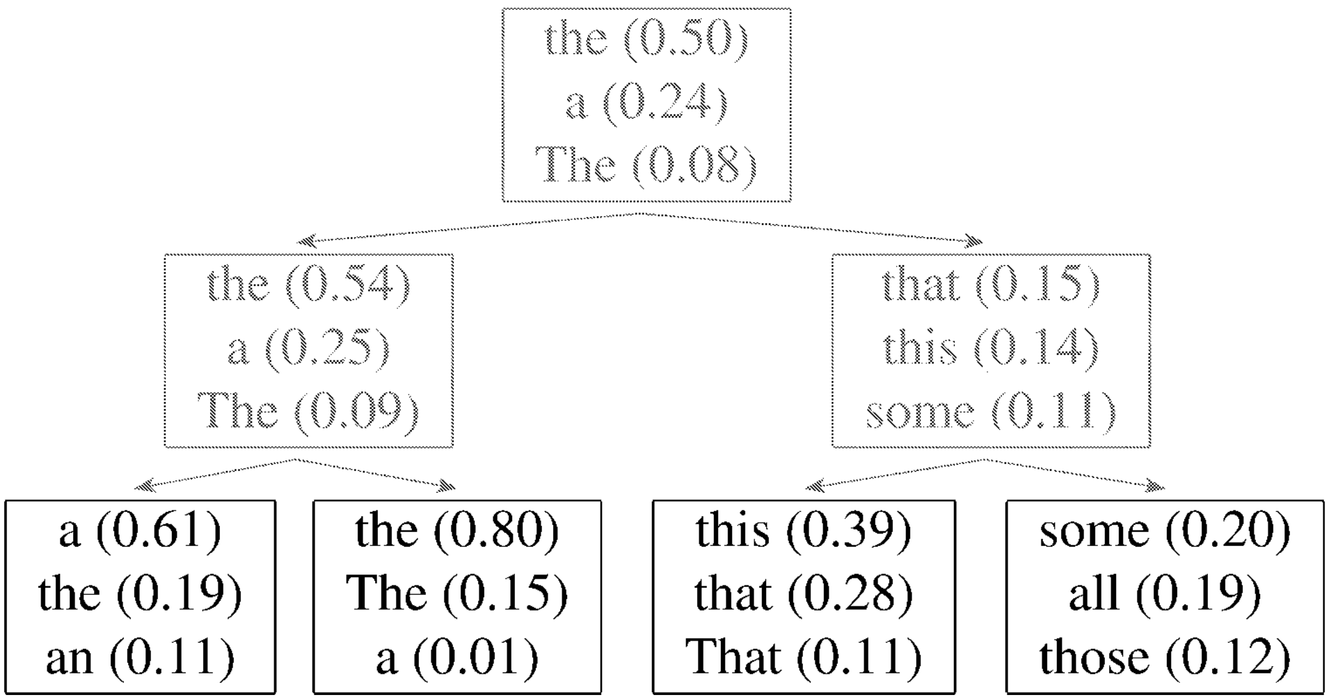


Just like Forward-Backward for HMMs.

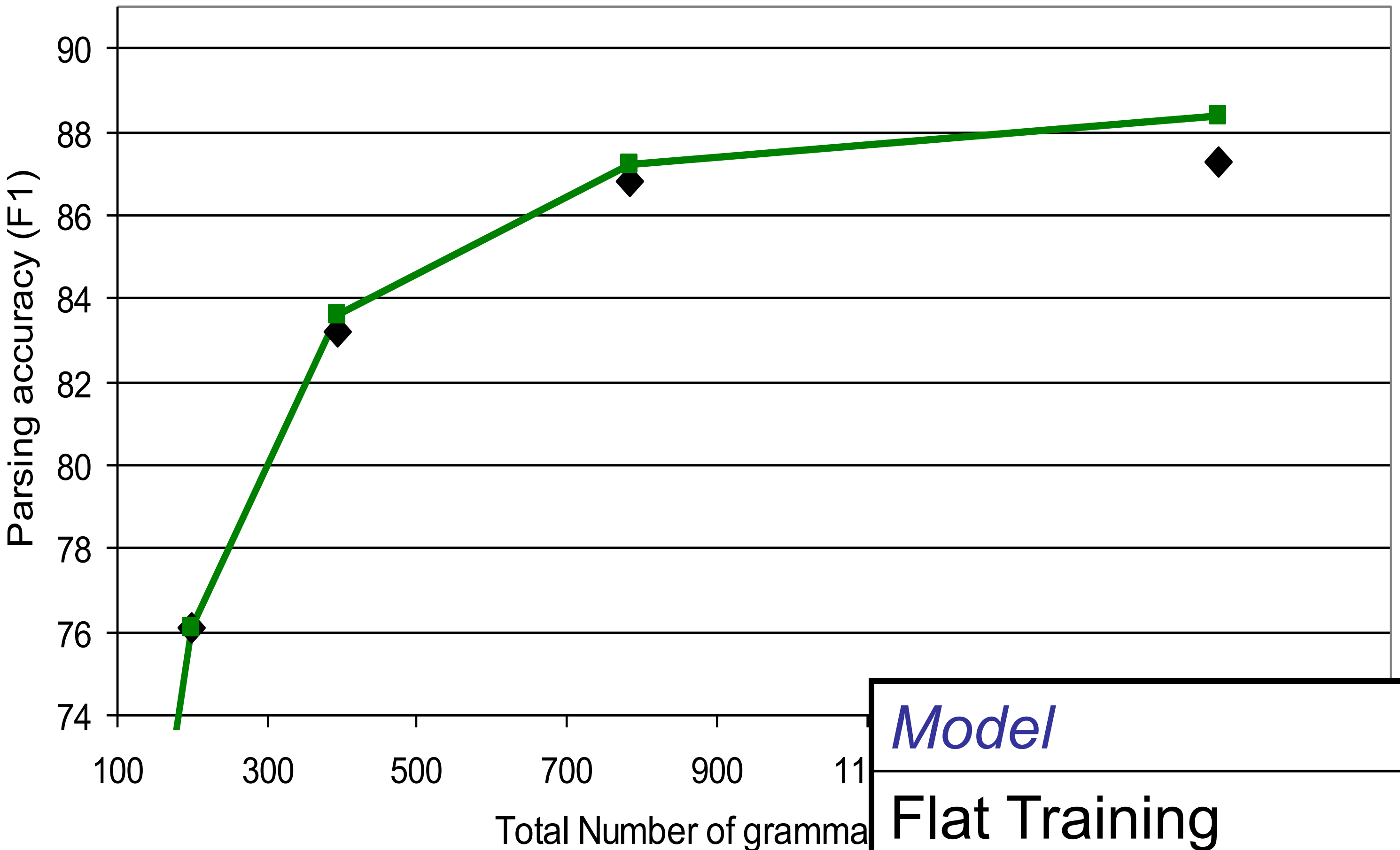
Refinement of the DT tag



Hierarchical refinement



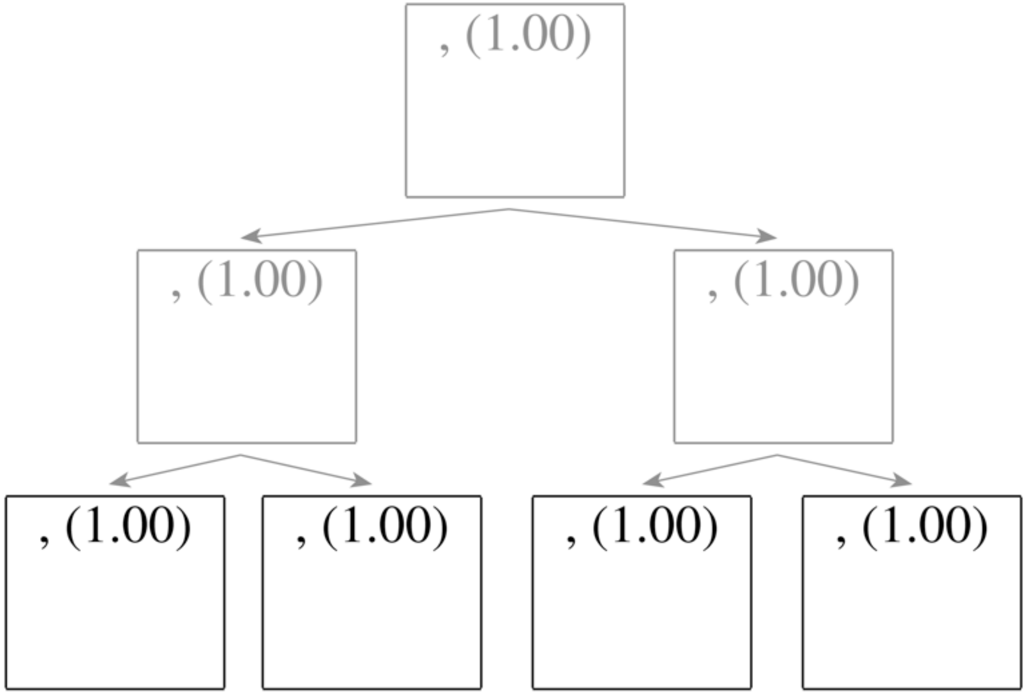
Hierarchical Estimation Results



Model	*F1*
Flat Training	87.3
Hierarchical Training	88.4

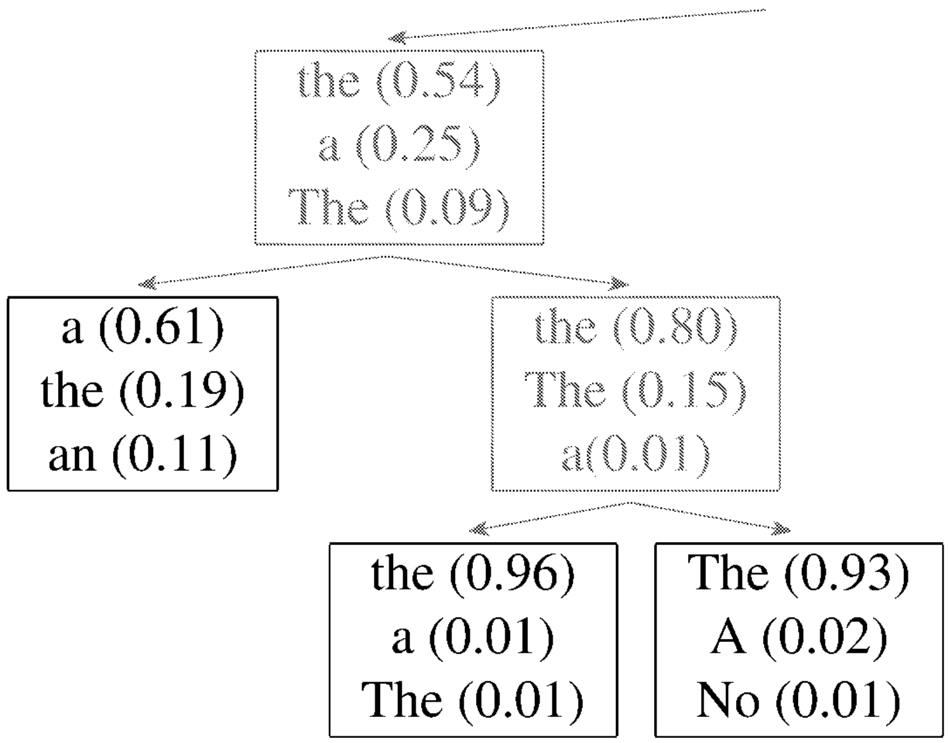
Refinement of the , tag

- Splitting all categories equally is wasteful:

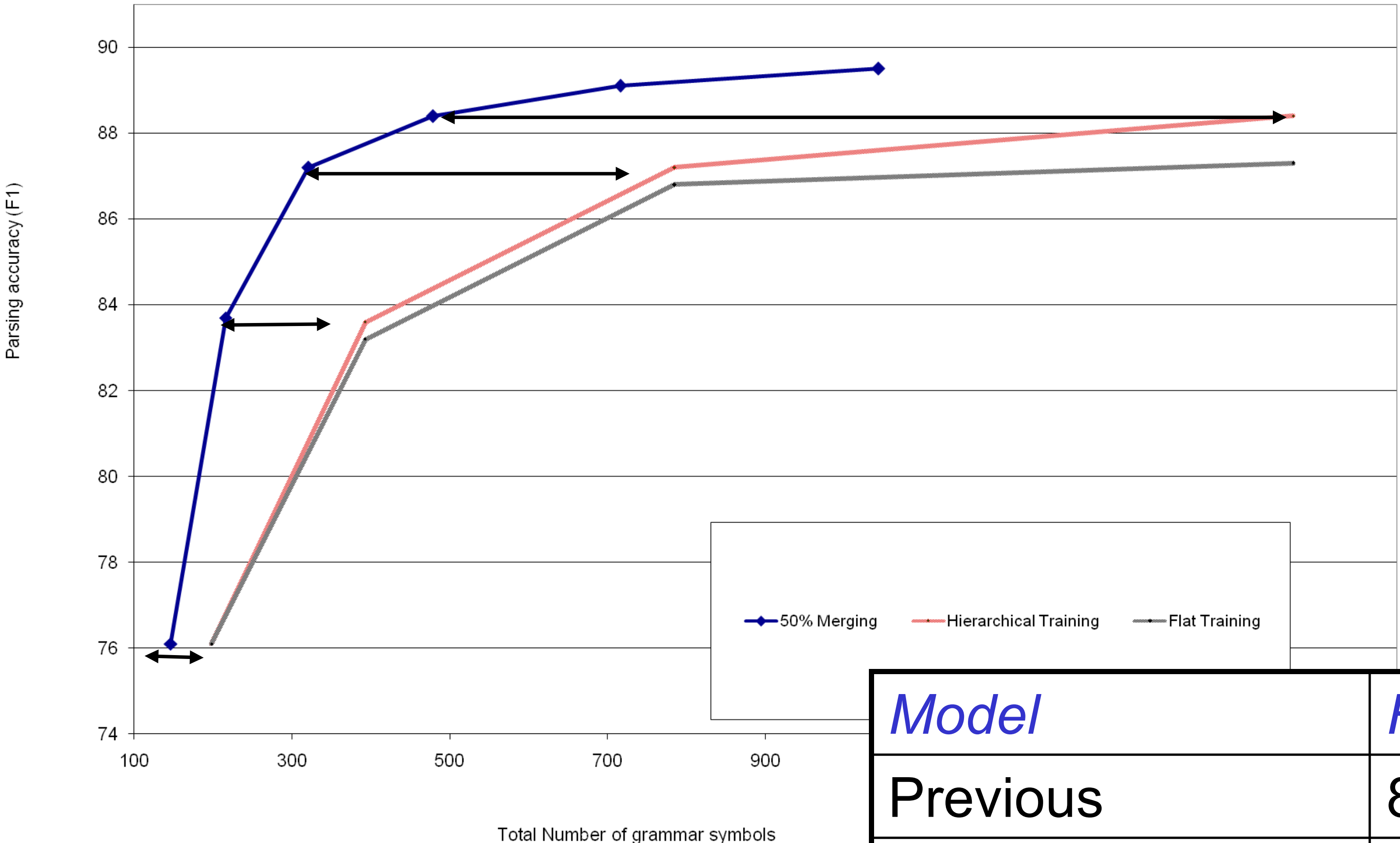


Adaptive Splitting

- Want to split complex categories more
- Idea: split everything, roll back splits which were least useful

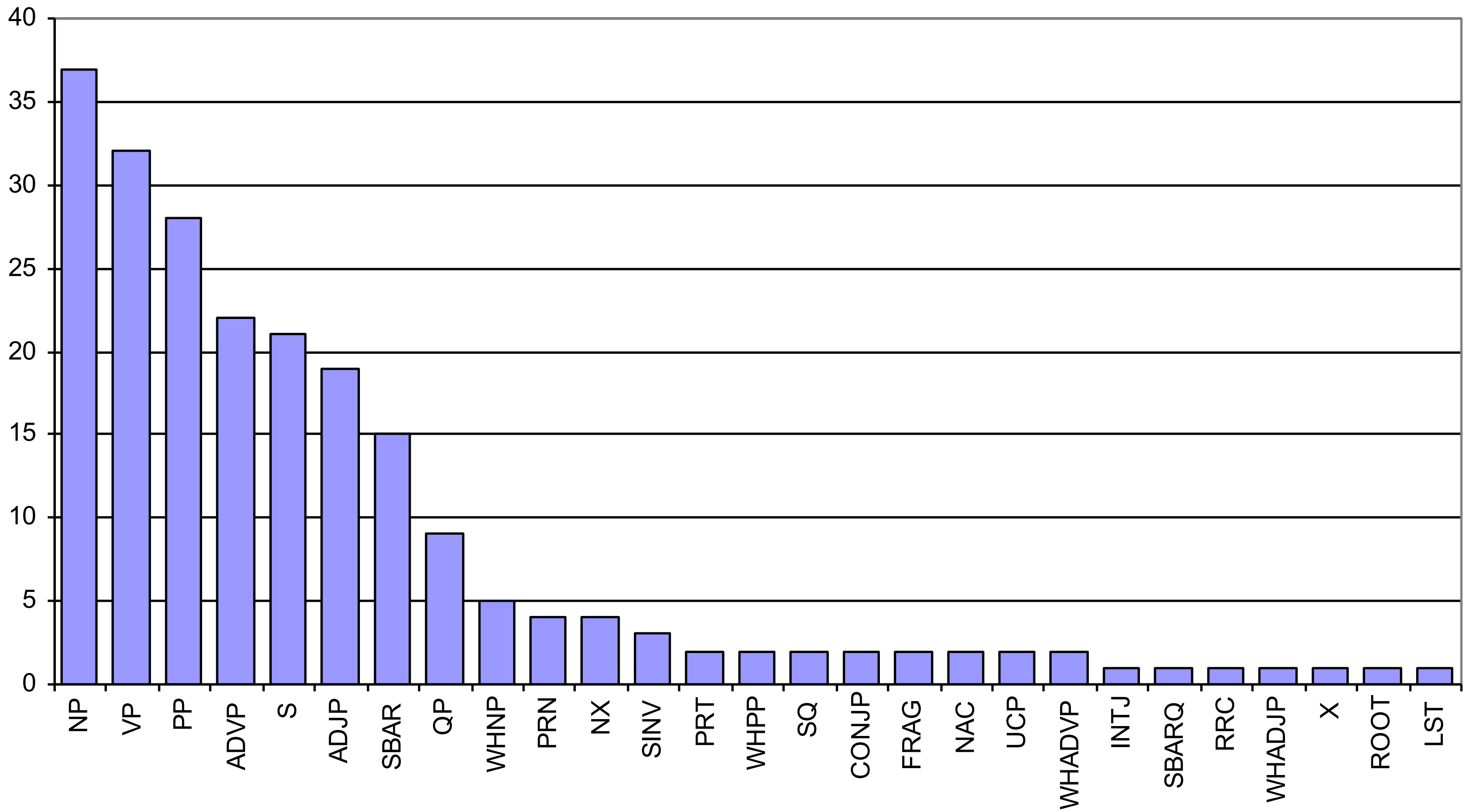


Adaptive Splitting Results

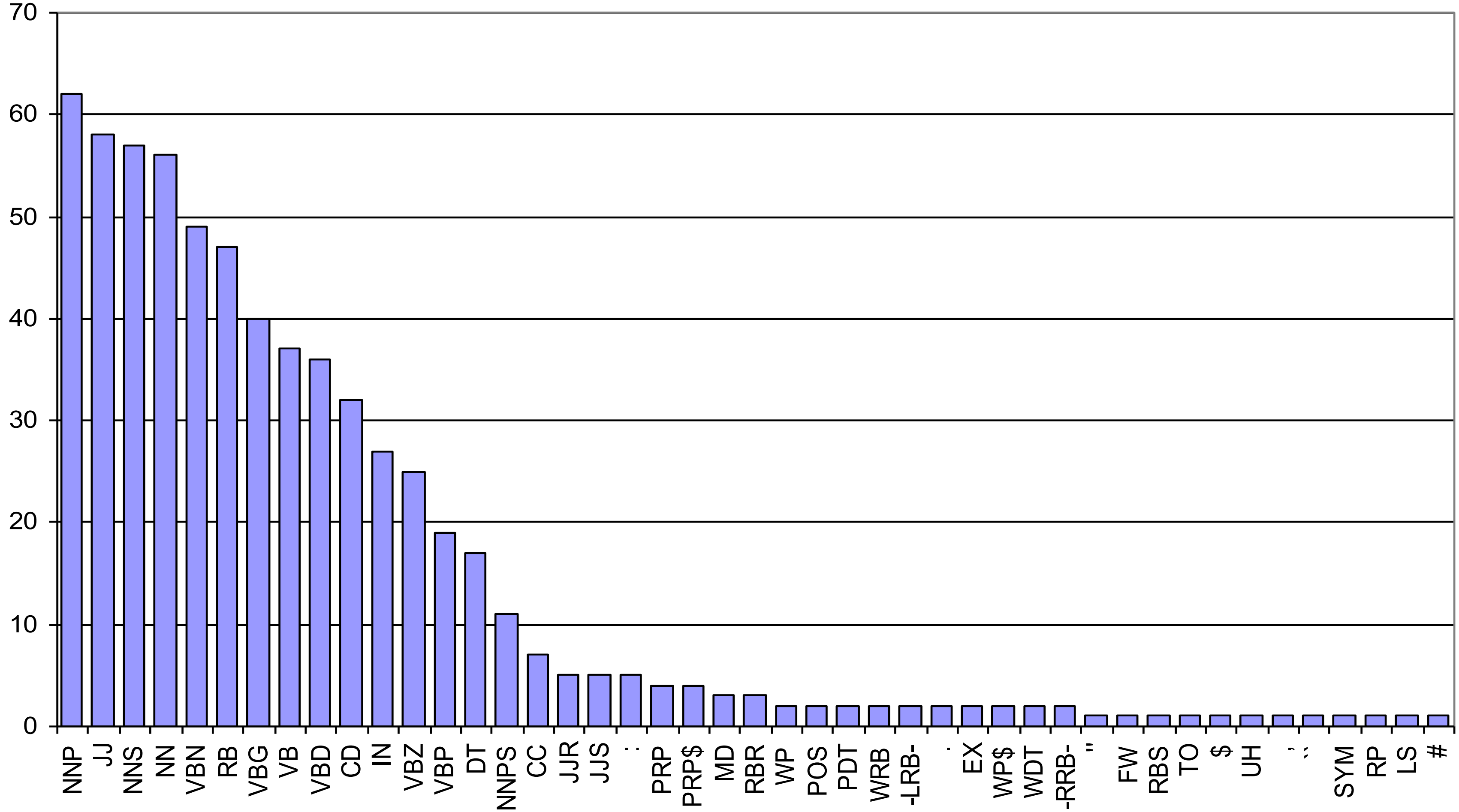


Model	*F1*
Previous	88.4
With 50% Merging	89.5

Number of Phrasal Subcategories



Number of Lexical Subcategories



Learned Splits

- Proper Nouns (NNP):

NNP-14	Oct.	Nov.	Sept.
NNP-12	John	Robert	James
NNP-2	J.	E.	L.
NNP-1	Bush	Noriega	Peters
NNP-15	New	San	Wall
NNP-3	York	Francisco	Street

- Personal pronouns (PRP):

PRP-0	It	He	I
PRP-1	it	he	they
PRP-2	it	them	him

Learned Splits

- Relative adverbs (RBR):

RBR-0	further	lower	higher
RBR-1	more	less	More
RBR-2	earlier	Earlier	later

- Cardinal Numbers (CD):

CD-7	one	two	Three
CD-4	1989	1990	1988
CD-11	million	billion	trillion
CD-0	1	50	100
CD-3	1	30	31
CD-9	78	58	34

Final Results (Accuracy)

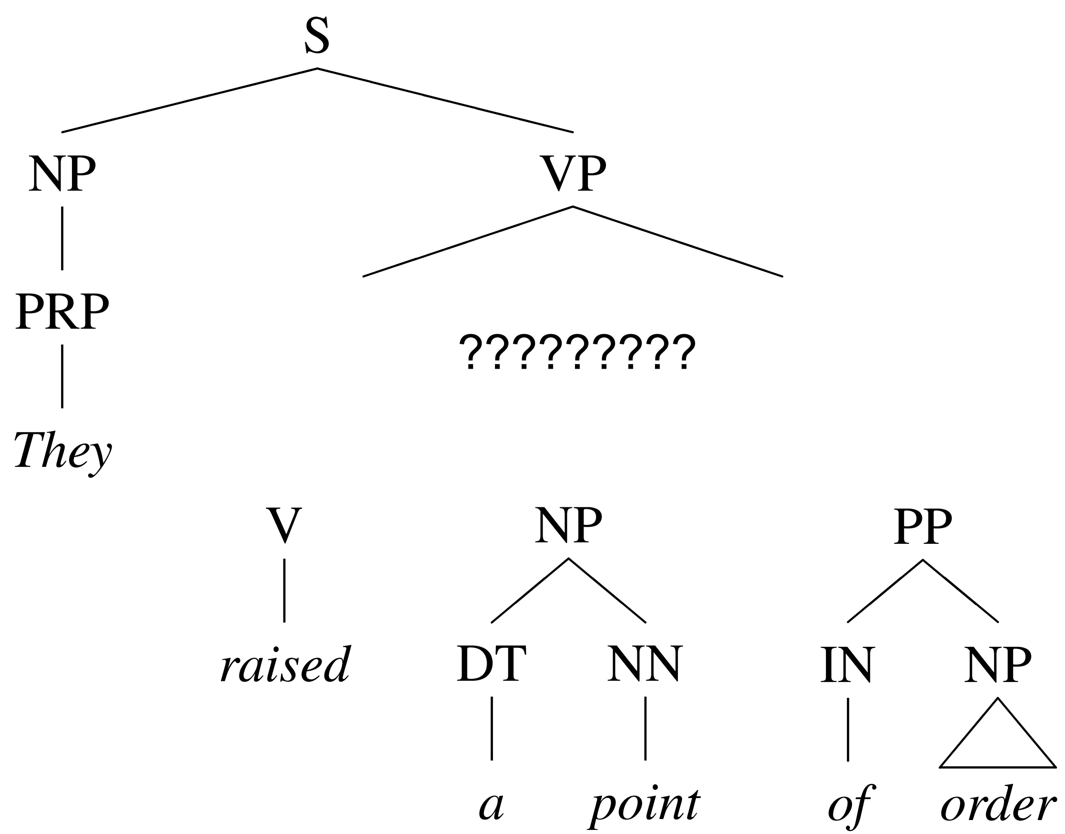
		≤ 40 words F1	all F1
ENG	Charniak&Johnson ‘05 (generative)	90.1	89.6
	Split / Merge	**90.6**	**90.1**
GER	Dubey ‘05	76.3	-
	Split / Merge	**80.8**	**80.1**
CHN	Chiang et al. ‘02	80.0	76.6
	Split / Merge	**86.3**	**83.4**

Still higher numbers from reranking / self-training methods

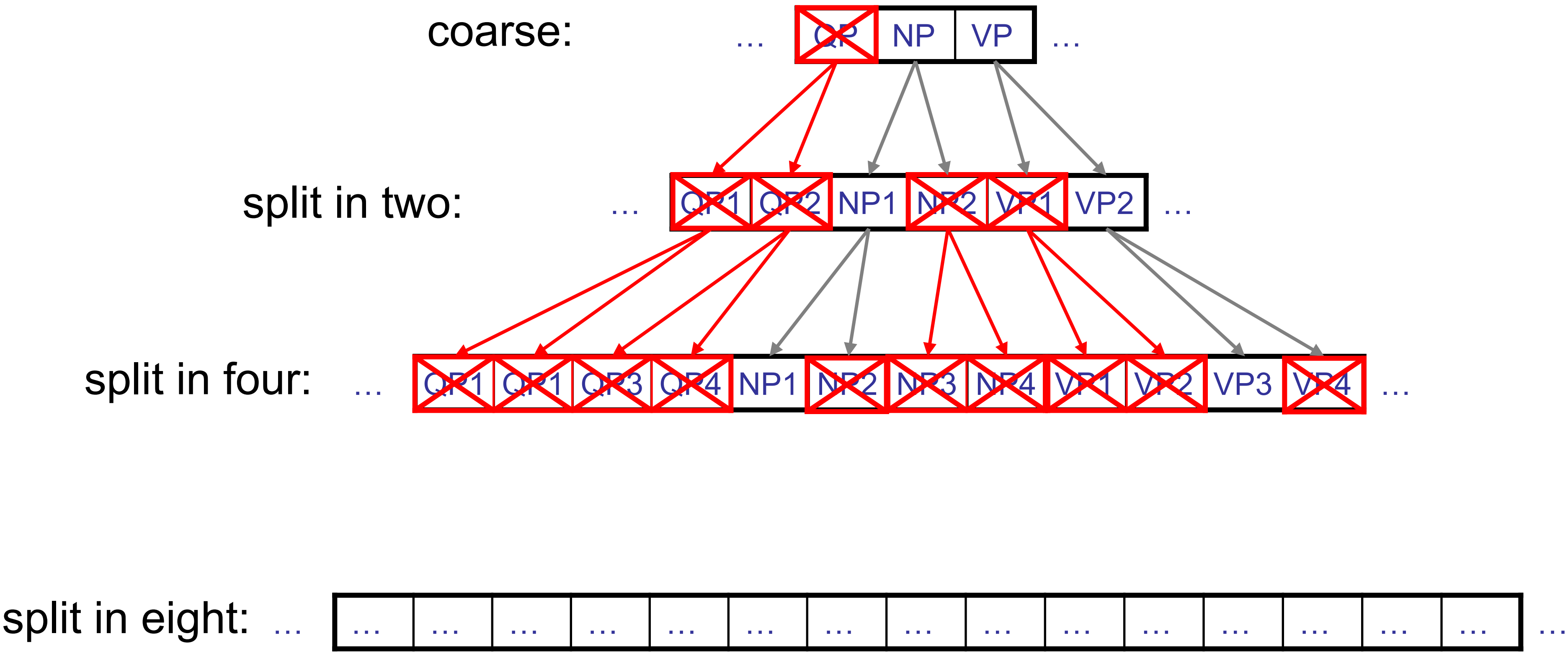
Efficient Parsing for Hierarchical Grammars

Coarse-to-Fine Inference

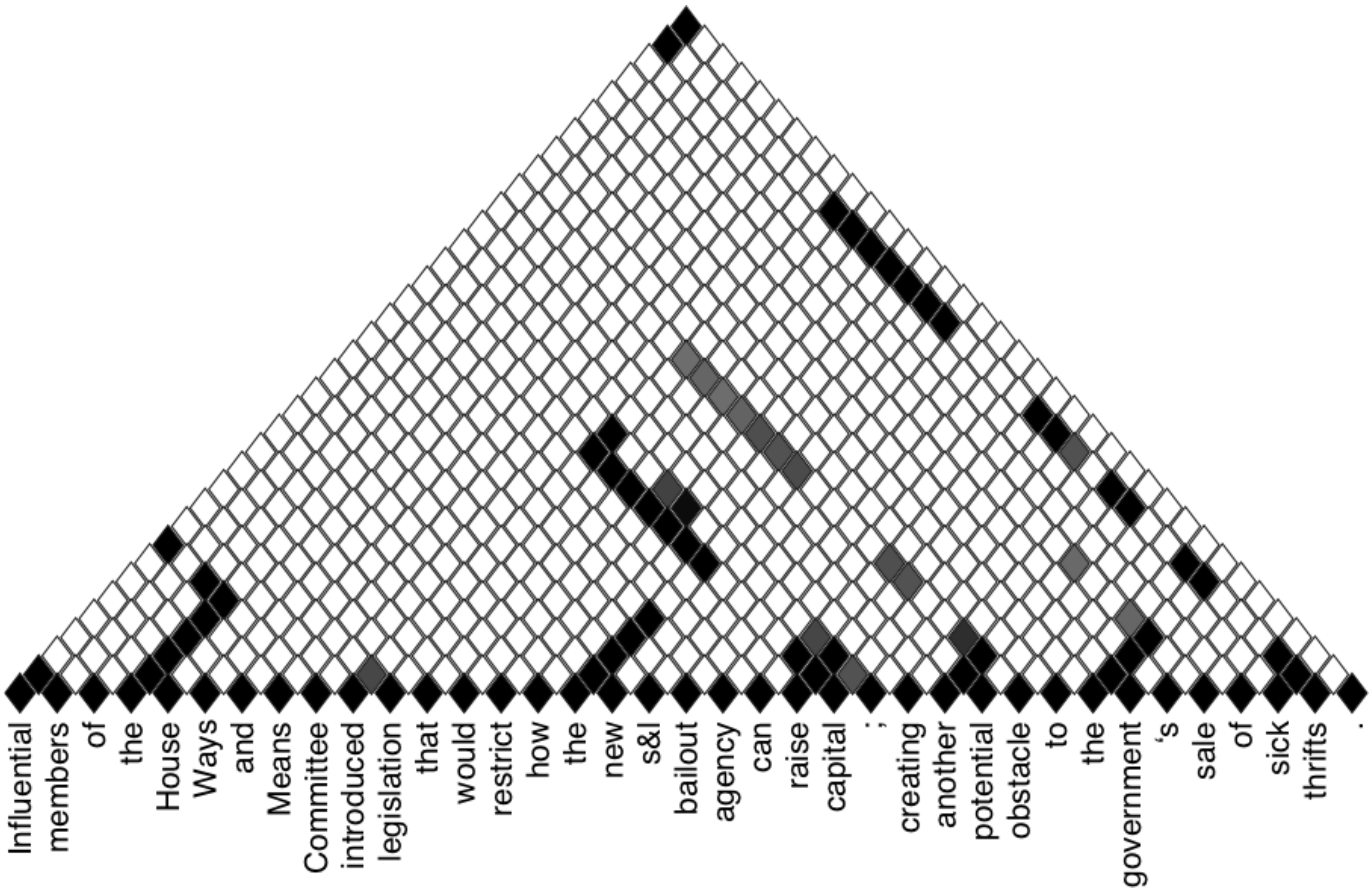
- Example: PP attachment



Hierarchical Pruning



Bracket Posteriors



1621 min

111 min

35 min

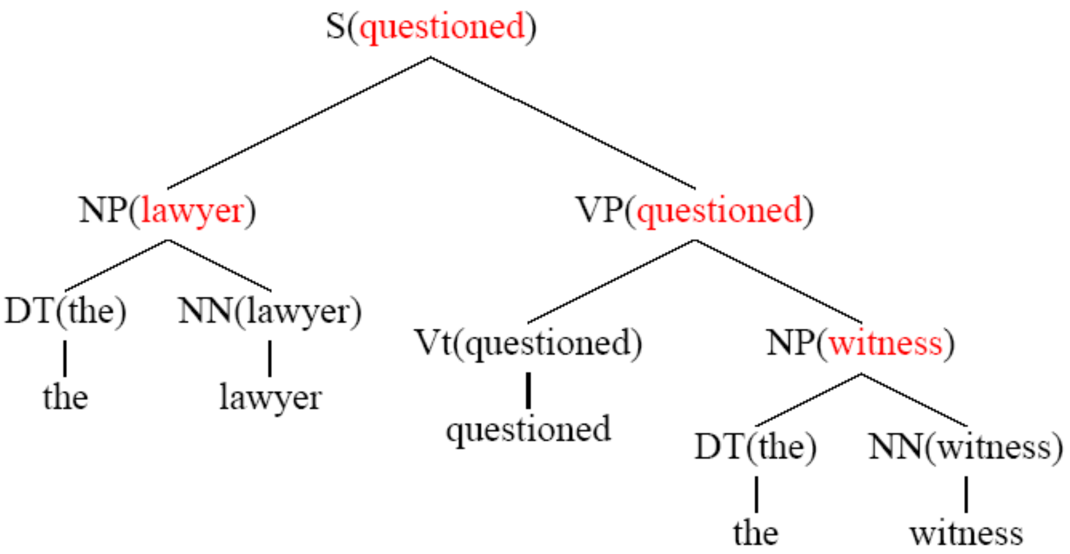
15 min

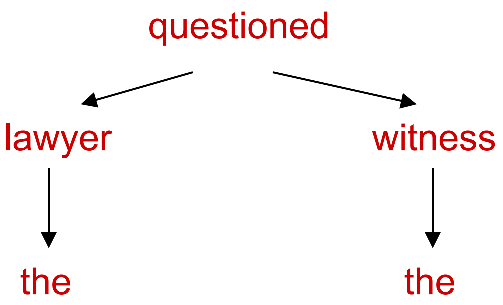
(no search error)

Other Syntactic Models

Dependency Parsing

- Lexicalized parsers can be seen as producing *dependency trees*

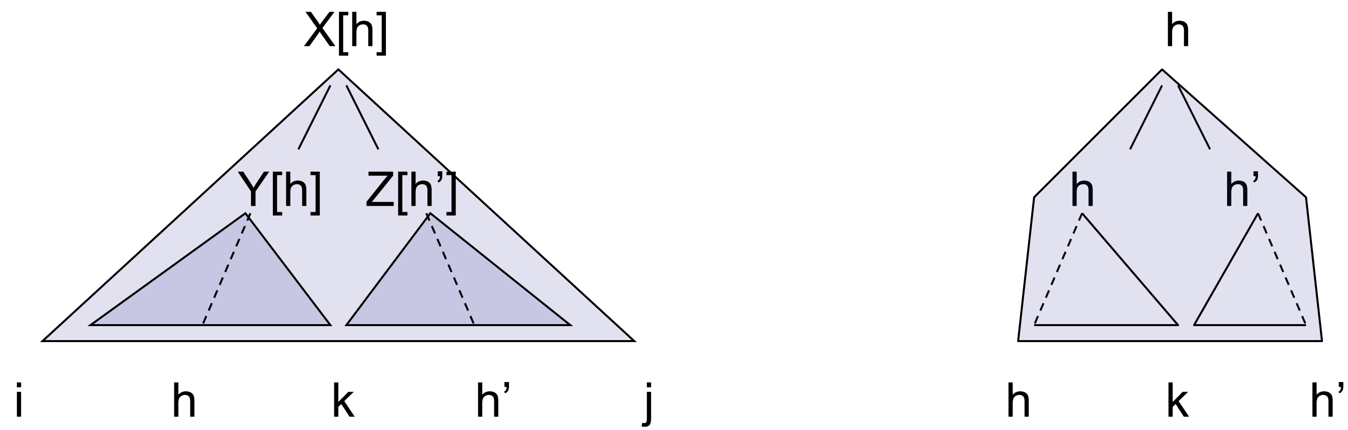




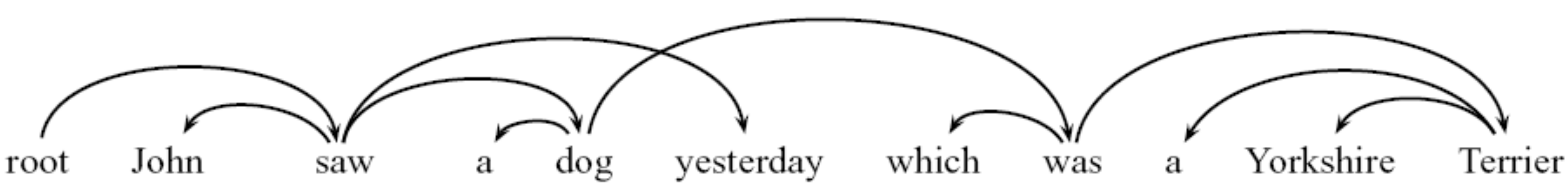
- Each local binary tree corresponds to an attachment in the dependency graph

Dependency Parsing

- Pure dependency parsing is only cubic [Eisner 99]

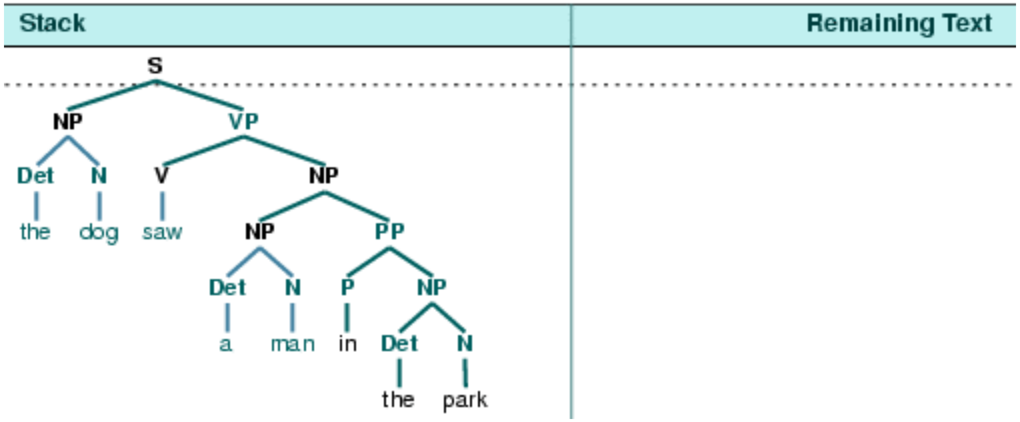


- Some work on *non-projective* dependencies
  - Common in, e.g. Czech parsing
  - Can do with MST algorithms [McDonald and Pereira 05]



Shift-Reduce Parsers

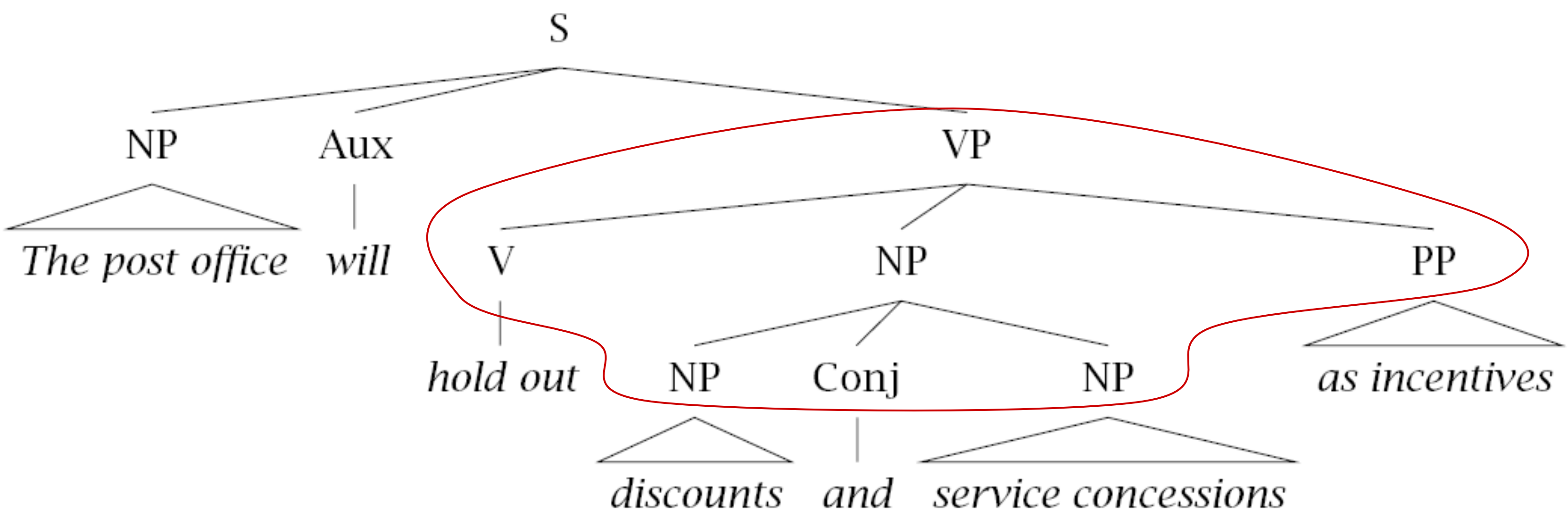
- Another way to derive a tree:



- Parsing
  - No useful dynamic programming search
  - Can still use beam search [Ratnaparkhi 97]

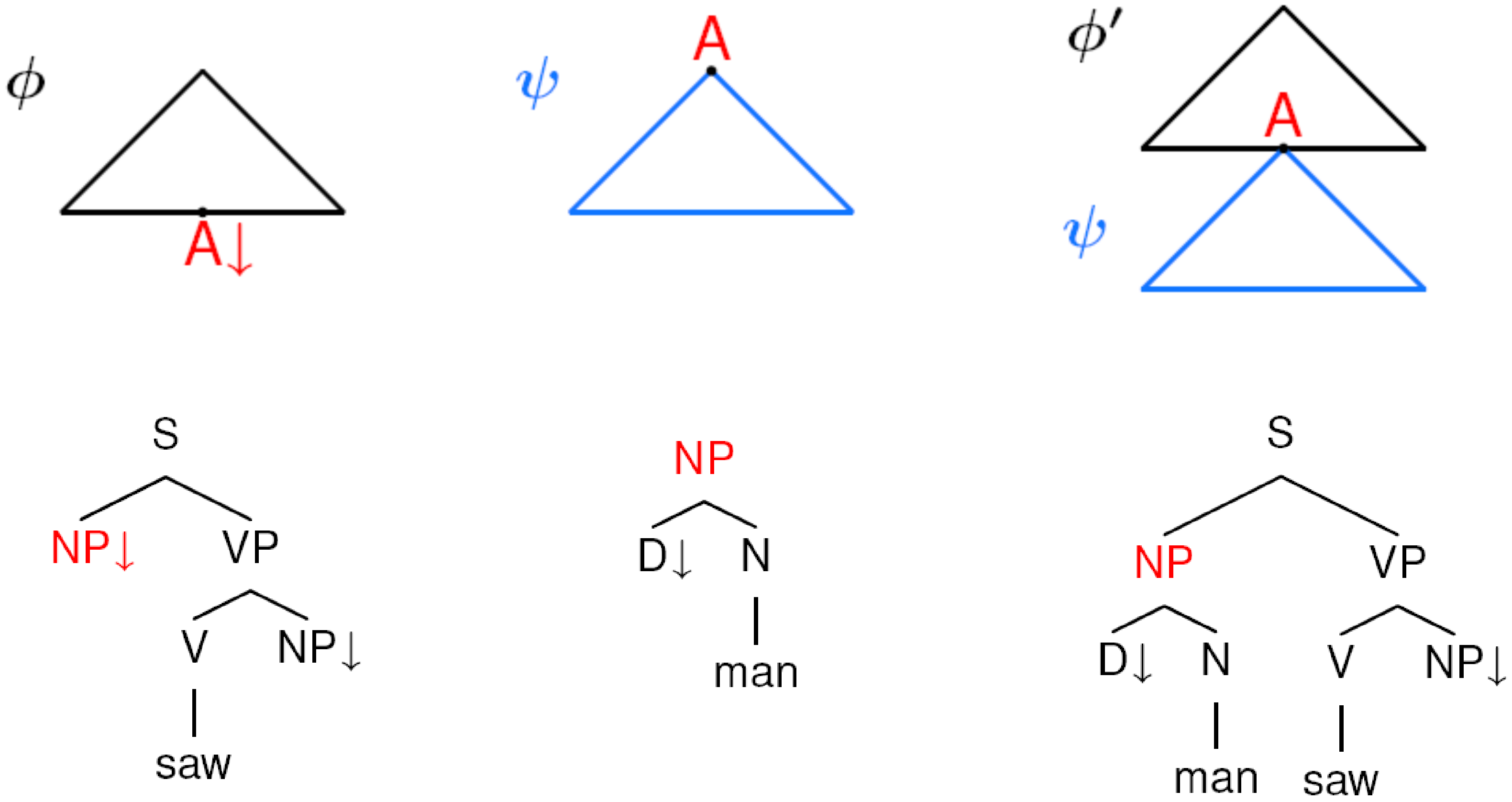
Tree Insertion Grammars

- Rewrite large (possibly lexicalized) subtrees in a single step



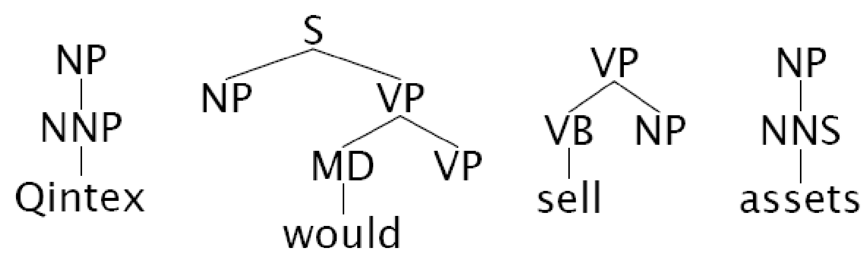
- Formally, a *tree-insertion grammar*
- Derivational ambiguity whether subtrees were generated atomically or compositionally
- Most probable *parse* is NP-complete

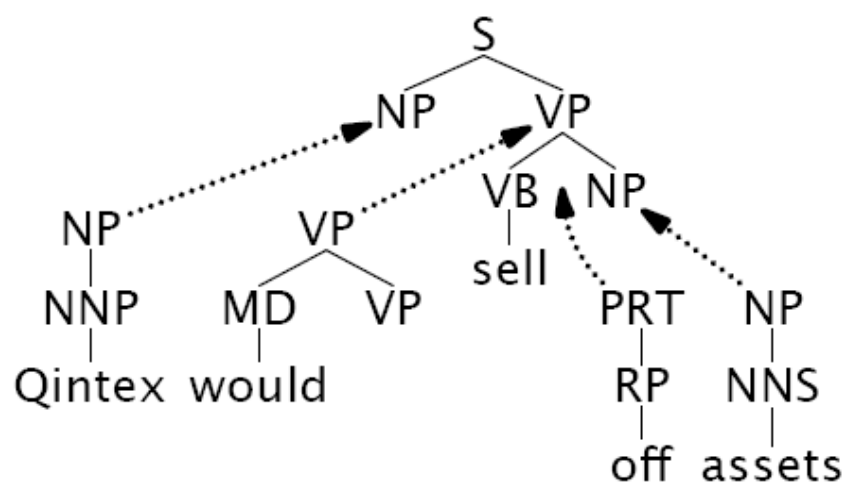
TIG: Insertion



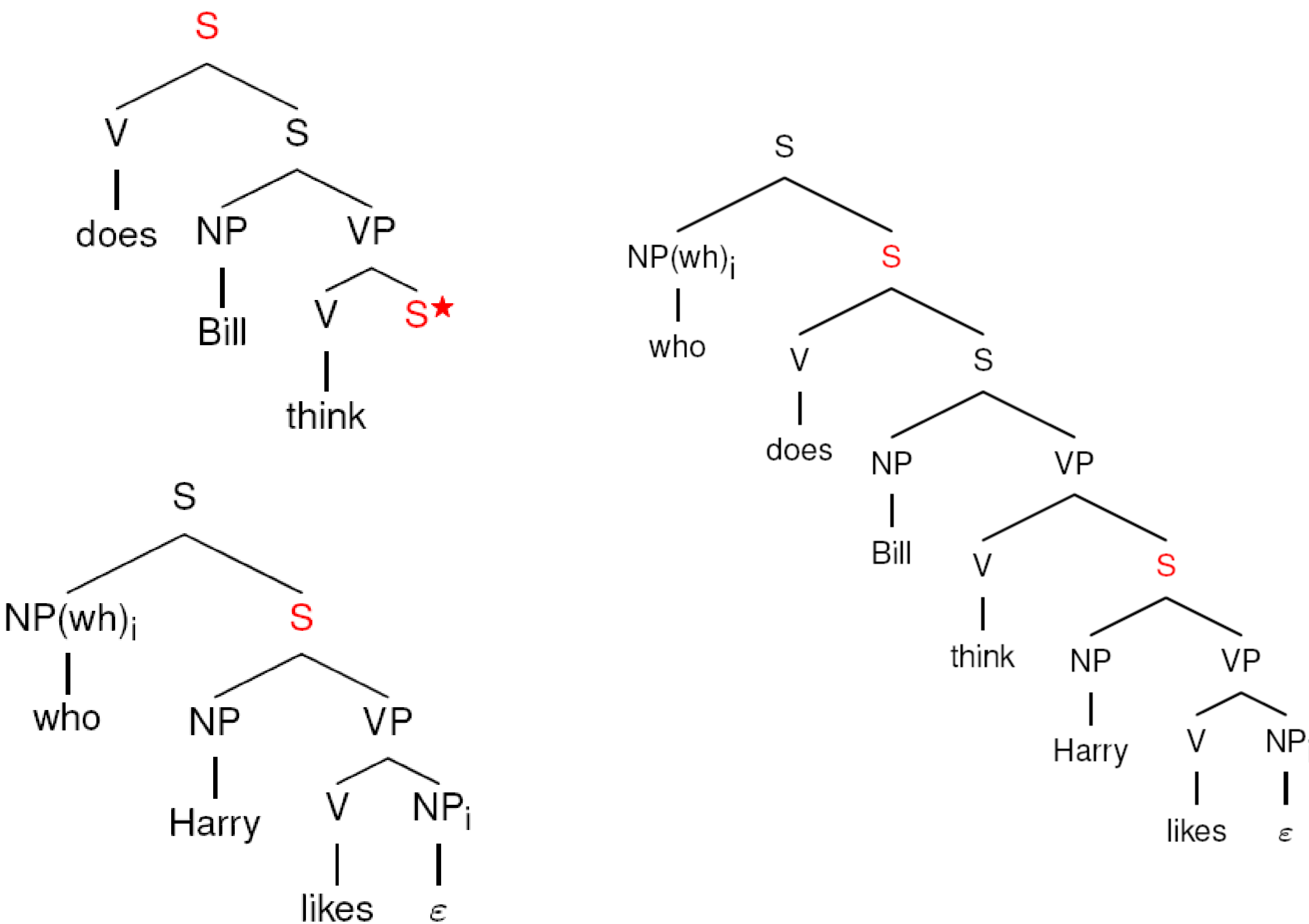
Tree-adjoining grammars

- Start with *local trees*
- Can insert structure with *adjunction* operators
- Mildly context-sensitive
- Models long-distance dependencies naturally
- … as well as other weird stuff that CFGs don't capture well (e.g. cross-serial dependencies)





TAG: Long Distance



CCG Parsing

- Combinatory Categorial Grammar
  - Fully (mono-) lexicalized grammar
  - Categories encode argument sequences
  - Very closely related to the lambda calculus (more later)
  - Can have spurious ambiguities (why?)

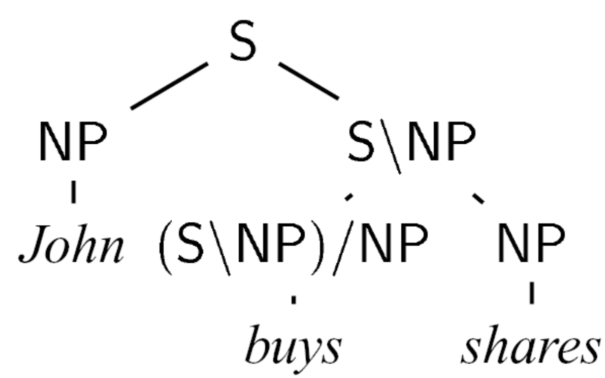
$$\textit{John} \vdash \mathsf{NP}$$

$$\textit{shares} \vdash \mathsf{NP}$$

$$\textit{buys} \vdash (\mathsf{S}\backslash\mathsf{NP})/\mathsf{NP}$$

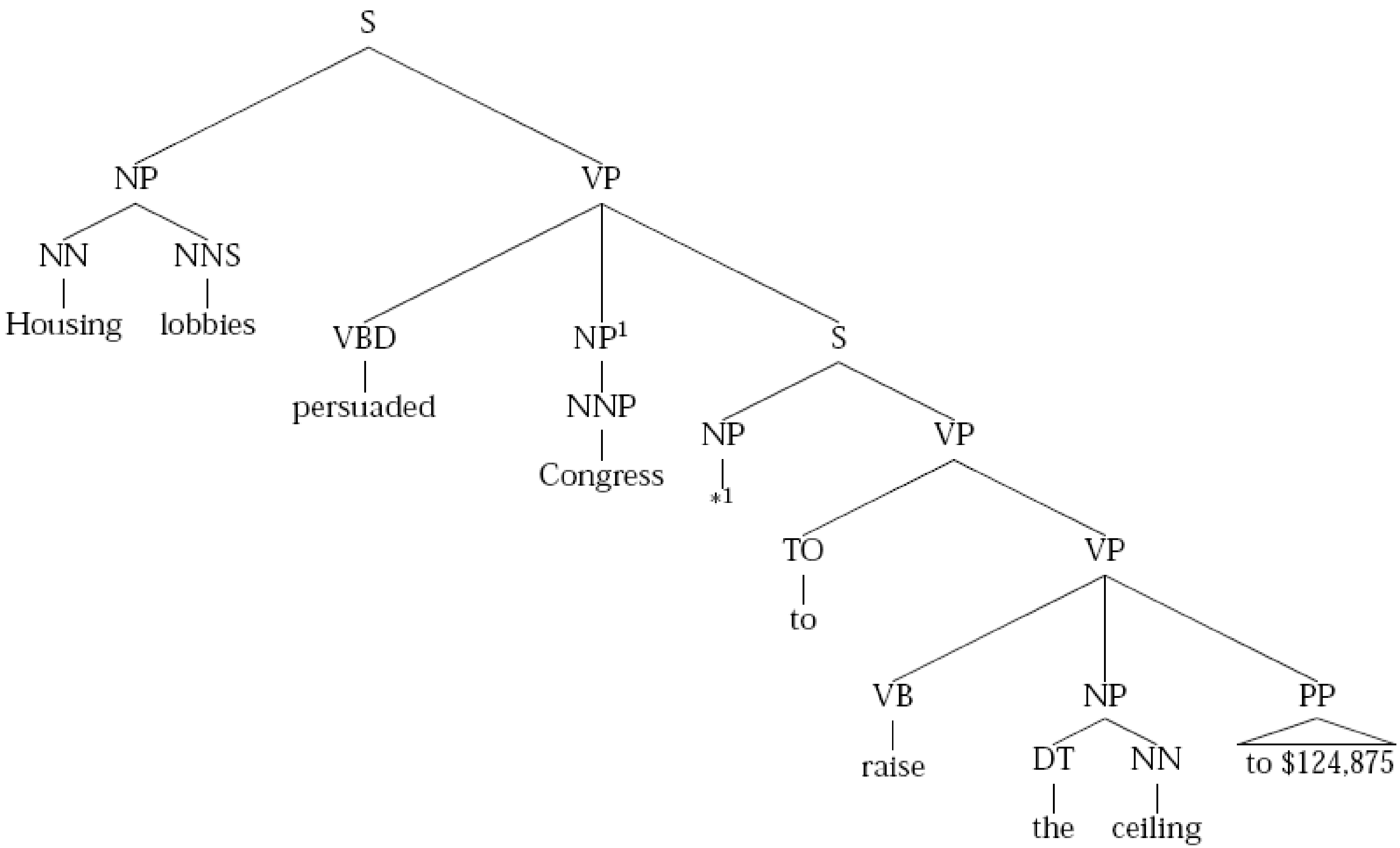
$$\textit{sleeps} \vdash \mathsf{S}\backslash\mathsf{NP}$$

$$\textit{well} \vdash (\mathsf{S}\backslash\mathsf{NP})\backslash(\mathsf{S}\backslash\mathsf{NP})$$



Empty Elements

Empty Elements

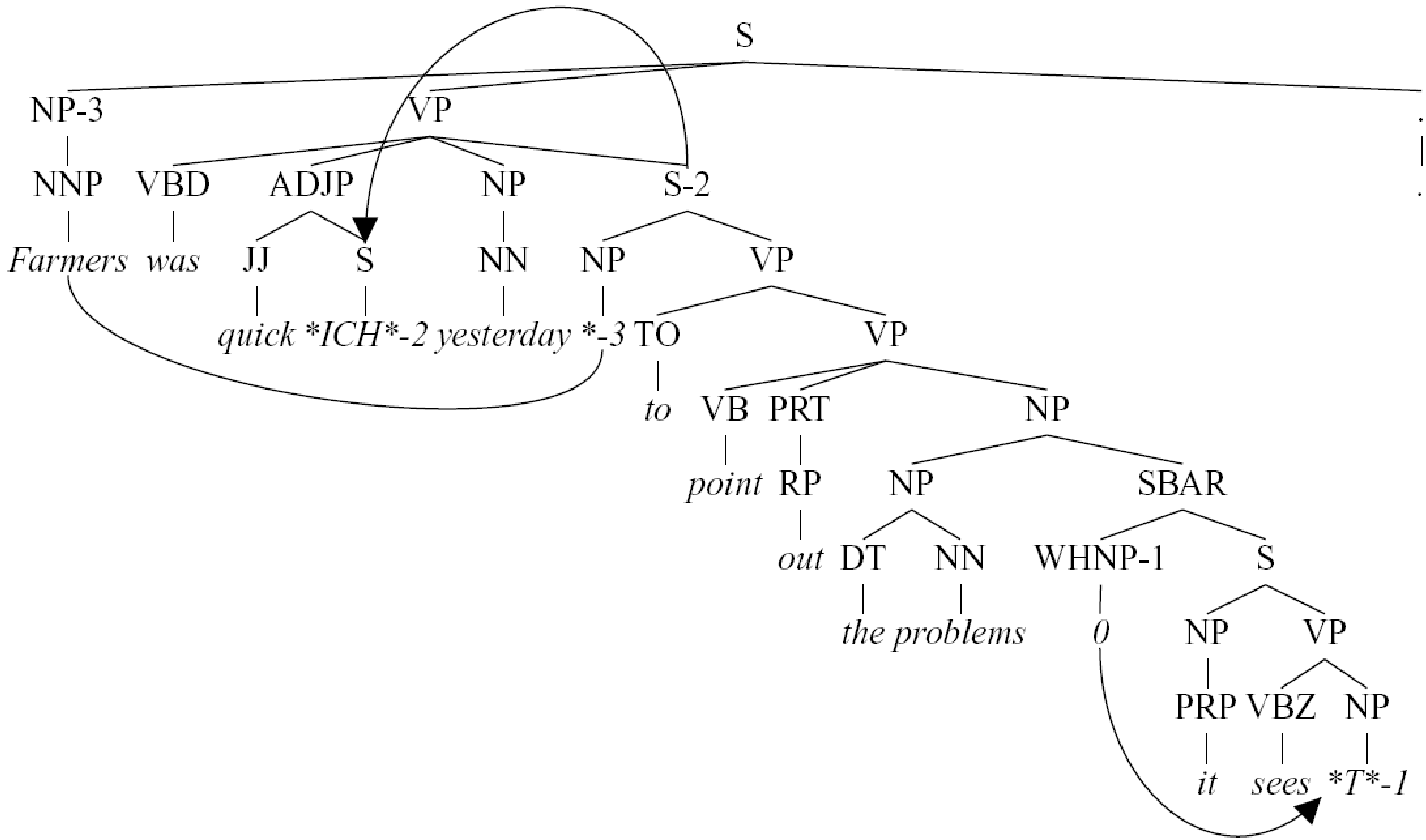


Empty Elements

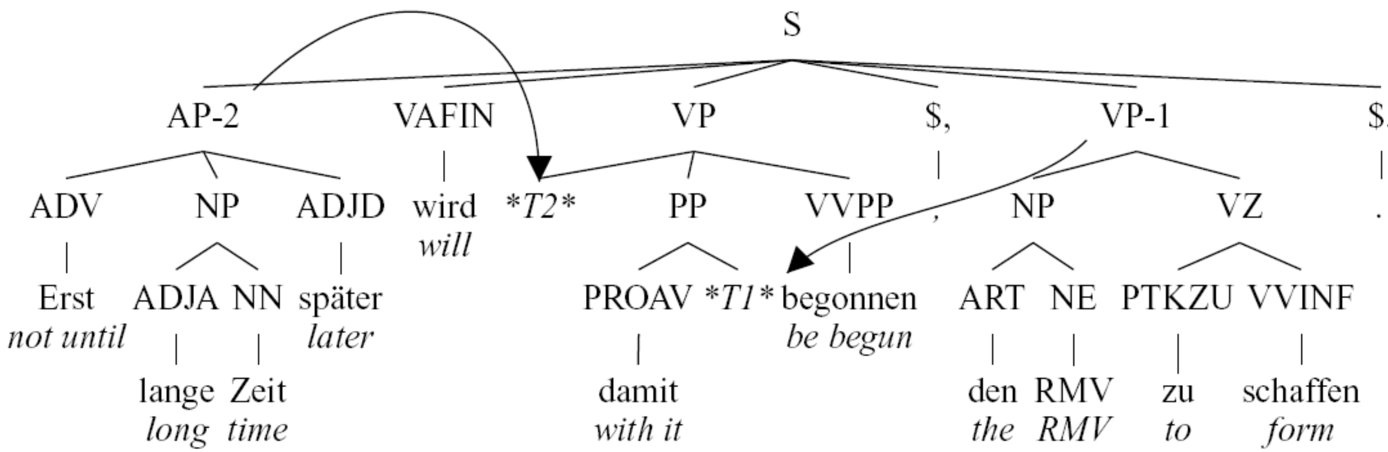
- In the PTB, three kinds of empty elements:
  - Null items (usually complementizers)
  - Dislocation (WH-traces, topicalization, relative clause and heavy NP extraposition)
  - Control (raising, passives, control, shared argumentation)

- Need to reconstruct these (and resolve any indexation)

Example: English

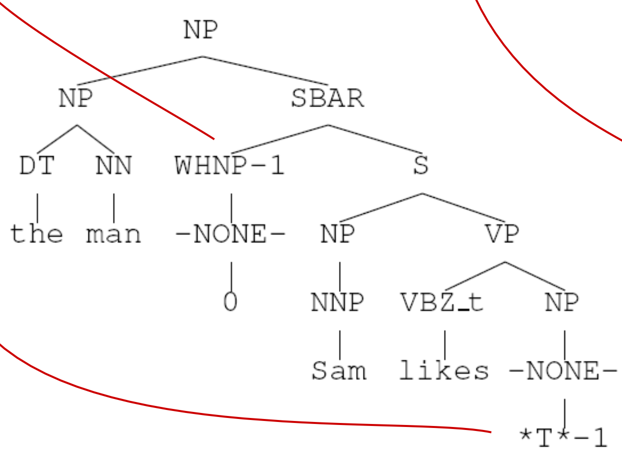


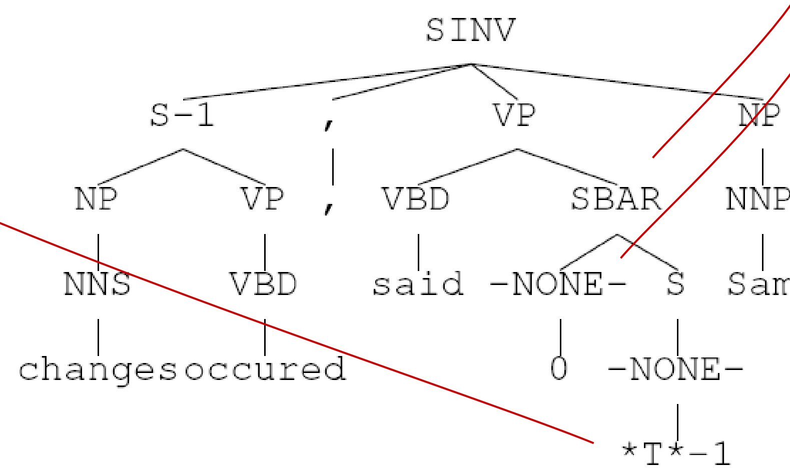
Example: German



Types of Empties

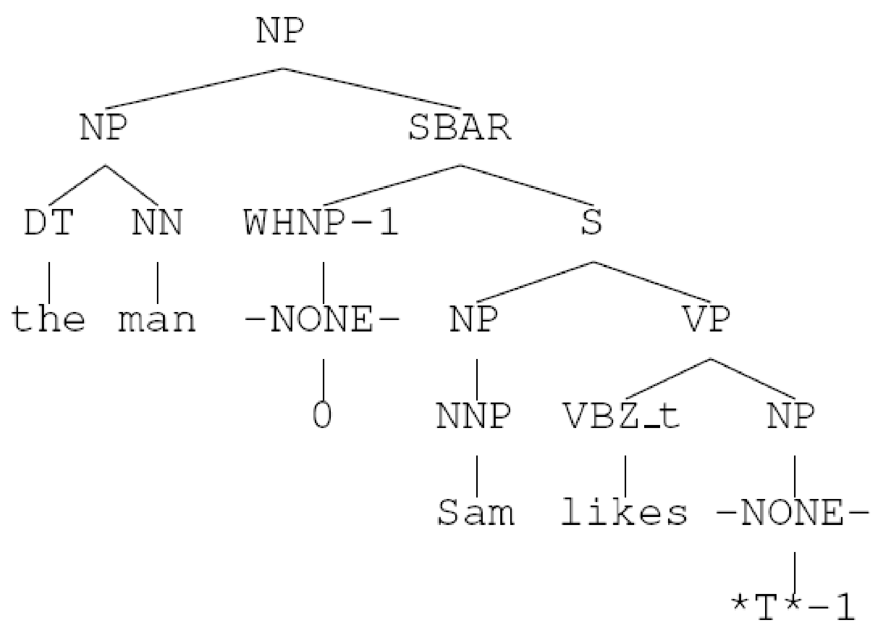
Antecedent	POS	Label	Count	Description
NP	NP	*	18,334	NP trace (e.g., *Sam was seen* *)
	NP	*	9,812	NP PRO (e.g., * *to sleep is nice*)
WHNP	NP	*T*	8,620	WH trace (e.g., *the woman who you saw *T**)
		U	7,478	Empty units (e.g., *$ 25 *U**)
		0	5,635	Empty complementizers (e.g., *Sam said 0 Sasha snores*)
S	S	*T*	4,063	Moved clauses (e.g., *Sam had to go, Sasha explained *T**)
WHADVP	ADVP	*T*	2,492	WH-trace (e.g., *Sam explained how to leave *T**)
	SBAR		2,033	Empty clauses (e.g., *Sam had to go, Sasha explained (SBAR)*)
	WHNP	0	1,759	Empty relative pronouns (e.g., *the woman 0 we saw*)
	WHADVP	0	575	Empty relative pronouns (e.g., *no reason 0 to leave*)

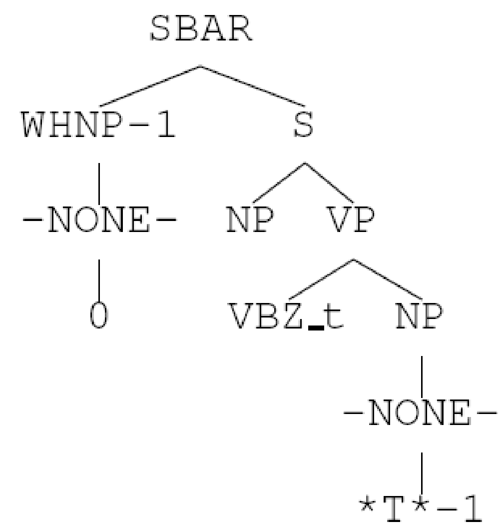


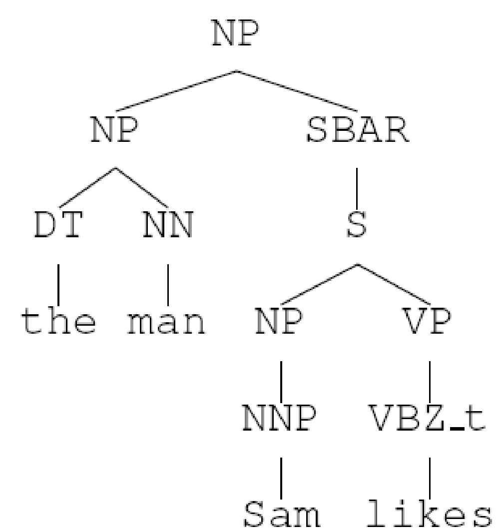


A Pattern-Matching Approach

- [Johnson 02]







Pattern-Matching Details

- Something like transformation-based learning
- Extract patterns
  - Details: transitive verb marking, auxiliaries
  - Details: legal subtrees
- Rank patterns
  - Pruning ranking: by correct / match rate
  - Application priority: by depth
- Pre-order traversal
- Greedy match

Top Patterns Extracted

Count	Match	Pattern
5816	6223	(S (NP (-NONE- *)) VP)
5605	7895	(SBAR (-NONE- 0) S)
5312	5338	(SBAR WHNP-1 (S (NP (-NONE- *T*-1)) VP))
4434	5217	(NP QP (-NONE- *U*))
1682	1682	(NP $ CD (-NONE- *U*))
1327	1593	(VP VBN_t (NP (-NONE- *)) PP)
700	700	(ADJP QP (-NONE- *U*))
662	1219	(SBAR (WHNP-1 (-NONE- 0)) (S (NP (-NONE- *T*-1)) VP))
618	635	(S S-1 , NP (VP VBD (SBAR (-NONE- 0) (S (-NONE- *T*-1)))) .)
499	512	(SINV `` S-1 , '' (VP VBZ (S (-NONE- *T*-1))) NP .)
361	369	(SINV `` S-1 , '' (VP VBD (S (-NONE- *T*-1))) NP .)
352	320	(S NP-1 (VP VBZ (S (NP (-NONE- *-1)) VP)))
346	273	(S NP-1 (VP AUX (VP VBN_t (NP (-NONE- *-1)) PP)))
322	467	(VP VBD_t (NP (-NONE- *)) PP)
269	275	(S `` S-1 , '' NP (VP VBD (S (-NONE- *T*-1))) .)

Results

Empty node		Section 23			Parser output		
POS	Label	P	R	f	P	R	f
(Overall)		0.93	0.83	0.88	0.85	0.74	0.79
NP	*	0.95	0.87	0.91	0.86	0.79	0.82
NP	*T*	0.93	0.88	0.91	0.85	0.77	0.81
	0	0.94	0.99	0.96	0.86	0.89	0.88
	U	0.92	0.98	0.95	0.87	0.96	0.92
S	*T*	0.98	0.83	0.90	0.97	0.81	0.88
ADVP	*T*	0.91	0.52	0.66	0.84	0.42	0.56
SBAR		0.90	0.63	0.74	0.88	0.58	0.70
WHNP	0	0.75	0.79	0.77	0.48	0.46	0.47

Semantic Roles

Semantic Role Labeling (SRL)

- Characterize clauses as *relations* with *roles*:

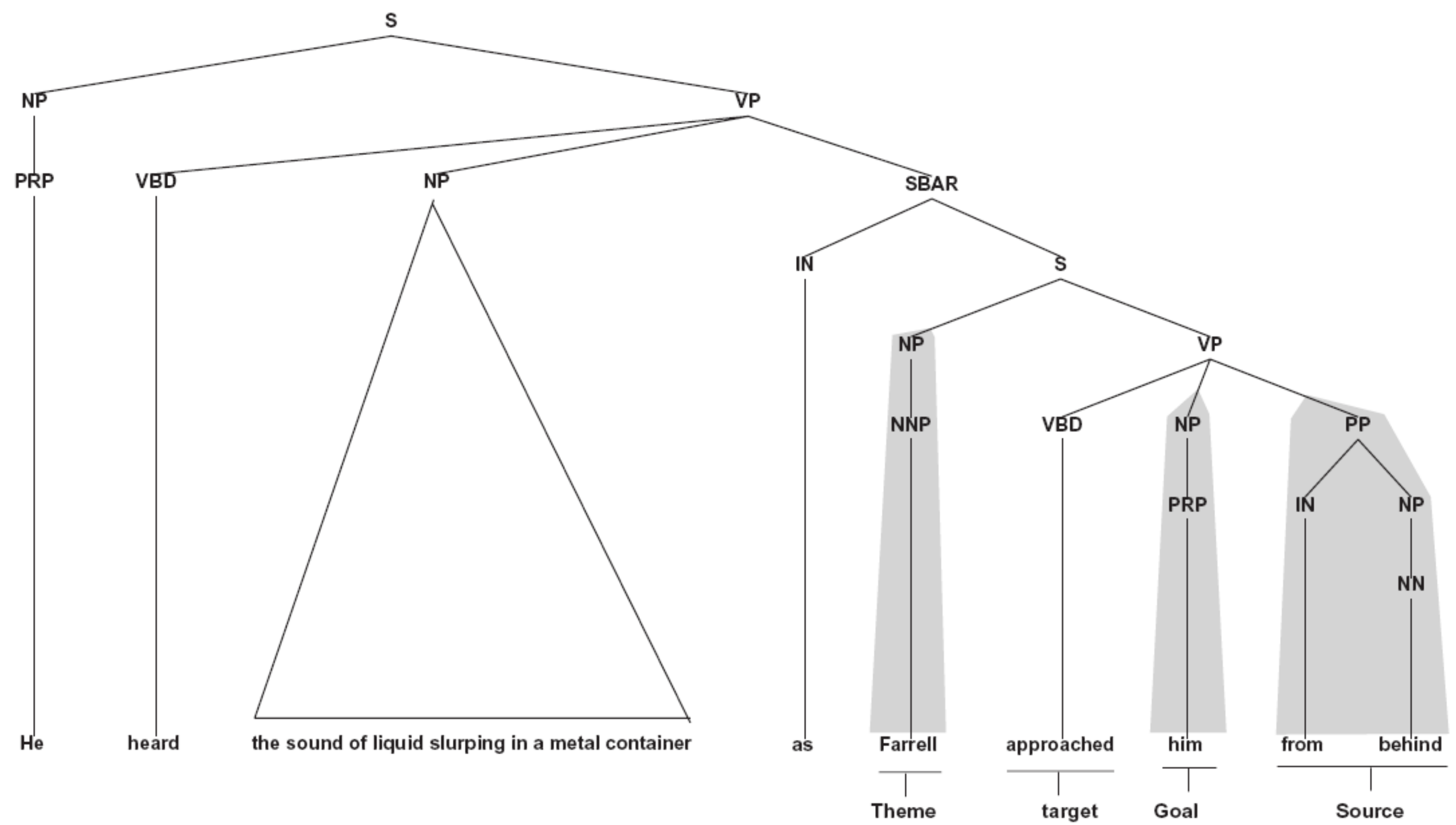
[$_{Judge}$ She] **blames** [$_{Evaluee}$ the Government] [$_{Reason}$ for failing to do enough to help] .

Holman would characterise this as **blaming** [$_{Evaluee}$ the poor] .

The letter quotes Black as saying that [$_{Judge}$ white and Navajo ranchers] misrepresent their livestock losses and **blame** [$_{Reason}$ everything] [$_{Evaluee}$ on coyotes] .

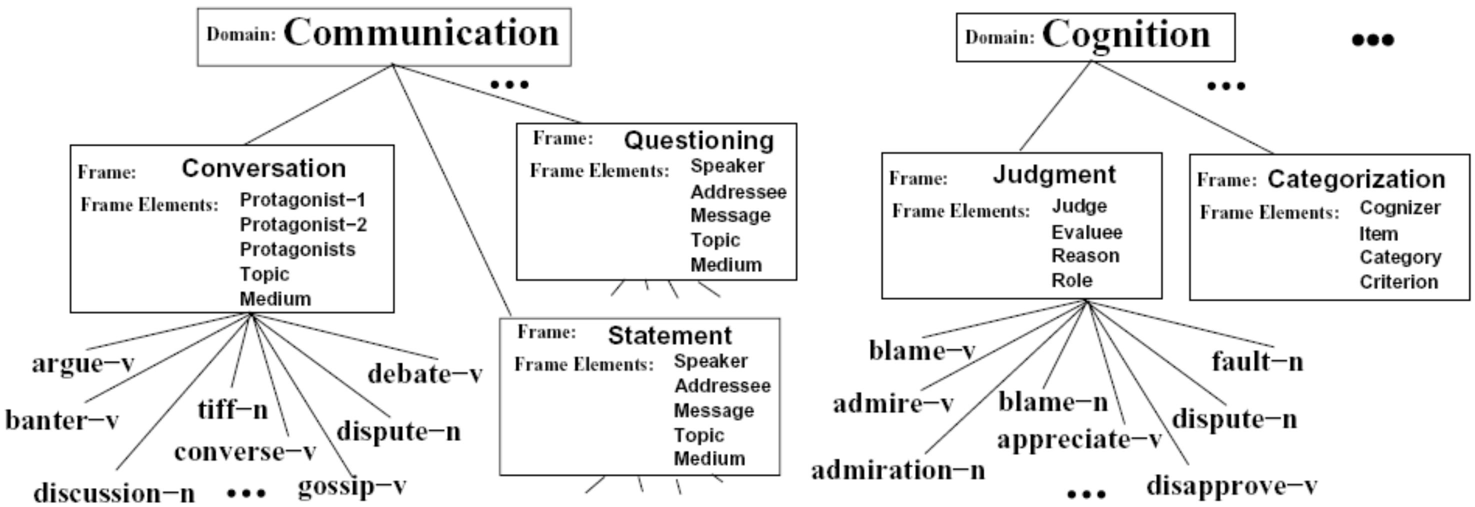
- Says more than which NP is the subject (but not much more):
- Relations like *subject* are syntactic, relations like *agent* or *message* are semantic
- Typical pipeline:
  - Parse, then label roles
  - Almost all errors locked in by parser
  - Really, SRL is quite a lot easier than parsing

SRL Example



S
NP
VP
PRP
VBD
NP
SBAR
IN
S
NP
VP
NNP
VBD
NP
PP
PRP
IN
NP
NN
He
heard
the sound of liquid slurping in a metal container
as
Farrell
approached
him
from
behind
Theme
target
Goal
Source

PropBank / FrameNet



- FrameNet: roles shared between verbs
- PropBank: each verb has its own roles
- PropBank more used, because it's layered over the treebank (and so has greater coverage, plus parses)
- Note: some linguistic theories postulate fewer roles than FrameNet (e.g. 5-20 total: agent, patient, instrument, etc.)

PropBank Example

fall.01		sense: move downward	
	roles:	Arg1:	thing falling
		Arg2:	extent, distance fallen
		Arg3:	start point
		Arg4:	end point

Sales fell to $251.2 million from $278.7 million.

arg1:	Sales
rel:	fell
arg4:	to $251.2 million
arg3:	from $278.7 million

PropBank Example

rotate.02		sense: shift from one thing to another	
	roles:	Arg0:	causer of shift
		Arg1:	thing being changed
		Arg2:	old thing
		Arg3:	new thing

Many of Wednesday's winners were losers yesterday as investors quickly took profits and rotated their buying to other issues, traders said. (wsj_1723)

arg0:	investors
rel:	rotated
arg1:	their buying
arg3:	to other issues

PropBank Example

aim.01 sense: intend, plan
roles: Arg0: aimer, planner
Arg1: plan, intent

The Central Council of Church Bell Ringers aims *trace* to improve relations with vicars. (wsj_0089)

arg0: The Central Council of Church Bell Ringers
rel: aims
arg1: *trace* to improve relations with vicars

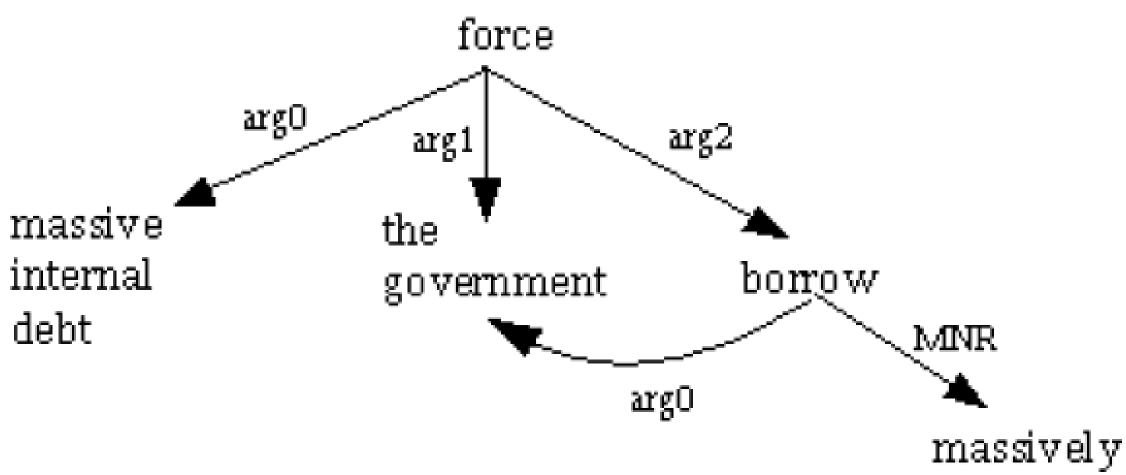
aim.02 sense: point (weapon) at
roles: Arg0: aimer
Arg1: weapon, etc.
Arg2: target

Banks have been aiming packages at the elderly.

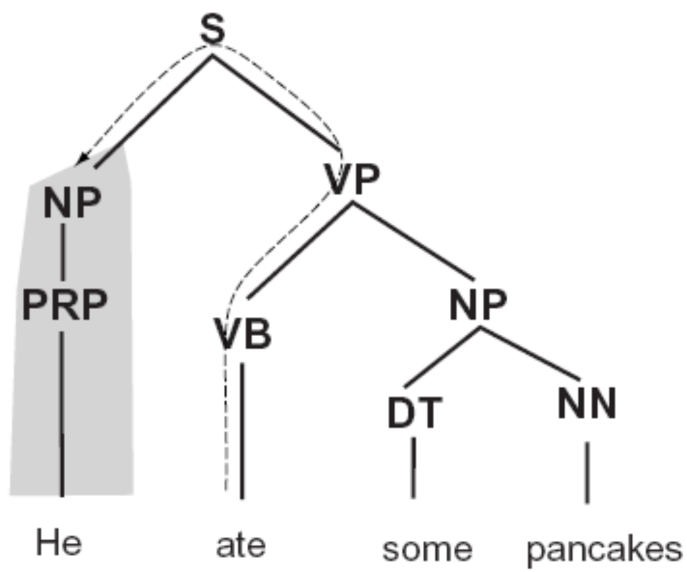
arg0: Banks
rel: aiming
arg1: packages
arg2: at the elderly

Shared Arguments

```
(NP-SBJ (JJ massive) (JJ internal) (NN debt) )
  (VP (VBZ has)
    (VP (VBN forced)
      (S
        (NP-SBJ-1 (DT the) (NN government) )
        (VP
          (VP (TO to)
            (VP (VB borrow)
              (ADVP-MNR (RB massively) )...
```



Path Features



Path	*Description*
VB↑VP↓PP	PP argument/adjunct
VB↑VP↑S↓NP	subject
VB↑VP↓NP	object
VB↑VP↑VP↑S↓NP	subject (embedded VP)
VB↑VP↓ADVP	adverbial adjunct
NN↑NP↑NP↓PP	prepositional complement of noun

Results

- Features:
  - Path from target to filler
  - Filler's syntactic type, headword, case
  - Target's identity
  - Sentence voice, etc.
  - Lots of other second-order features

- Gold vs parsed source trees
  - SRL is fairly easy on gold trees

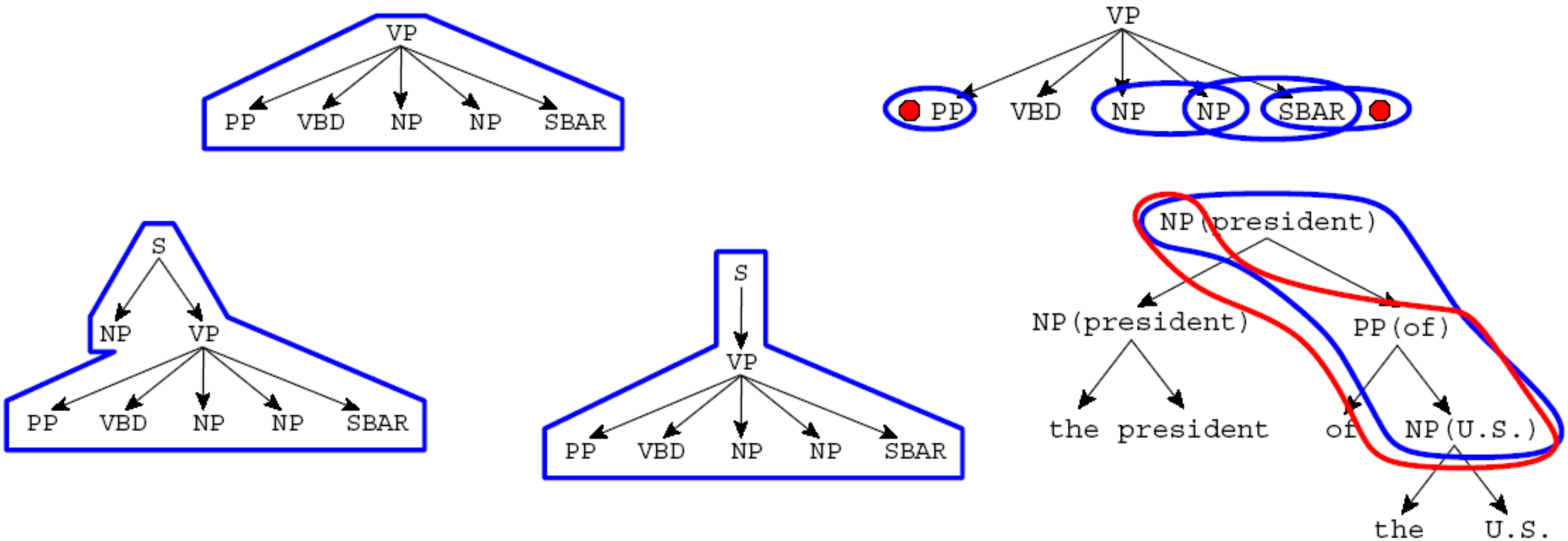
CORE		ARGM	
F1	Acc.	F1	Acc.
92.2	80.7	89.9	71.8

  - Harder on automatic parses

CORE		ARGM	
F1	Acc.	F1	Acc.
84.1	66.5	81.4	55.6

Parse Reranking

- Assume the number of parses is very small
- We can represent each parse T as a feature vector $\varphi(T)$
  - Typically, all local rules are features
  - Also non-local features, like how right-branching the overall tree is
  - [Charniak and Johnson 05] gives a rich set of features



K-Best Parsing

[Huang and Chiang 05, Pauls, Klein, Quirk 10]