Accurate Unlexicalized Parsing

by Dan Klein and Christopher D. Manning (ACL 2003)

Presented by Ulrich Germann

Background

- Naïve PCFGs tend to perform poorly, because their assumptions of context-freeness are too strong.
- Previous work (Magerman (1995), Collins(1996,1999), Charniak (1997,2000,2001)) relies on *lexicalization* of the PCFGs to improve performance.

Performance on sentences of up to 40 words

system	Precision	Recall	F_1
baseline naïve PCFG			72.6%
Magerman (1995)	84.9%	84.6%	84.7%
Collins (1996)	86.3%	85.8%	86.0%
Charniak (1997)	87.4%	87.5%	87.4%
Collins (1999)	88.7%	88.6%	88.6%
Charniak (2001)	90.1%	90.1%	90.1%

Beyond and Apart from Lexicalization

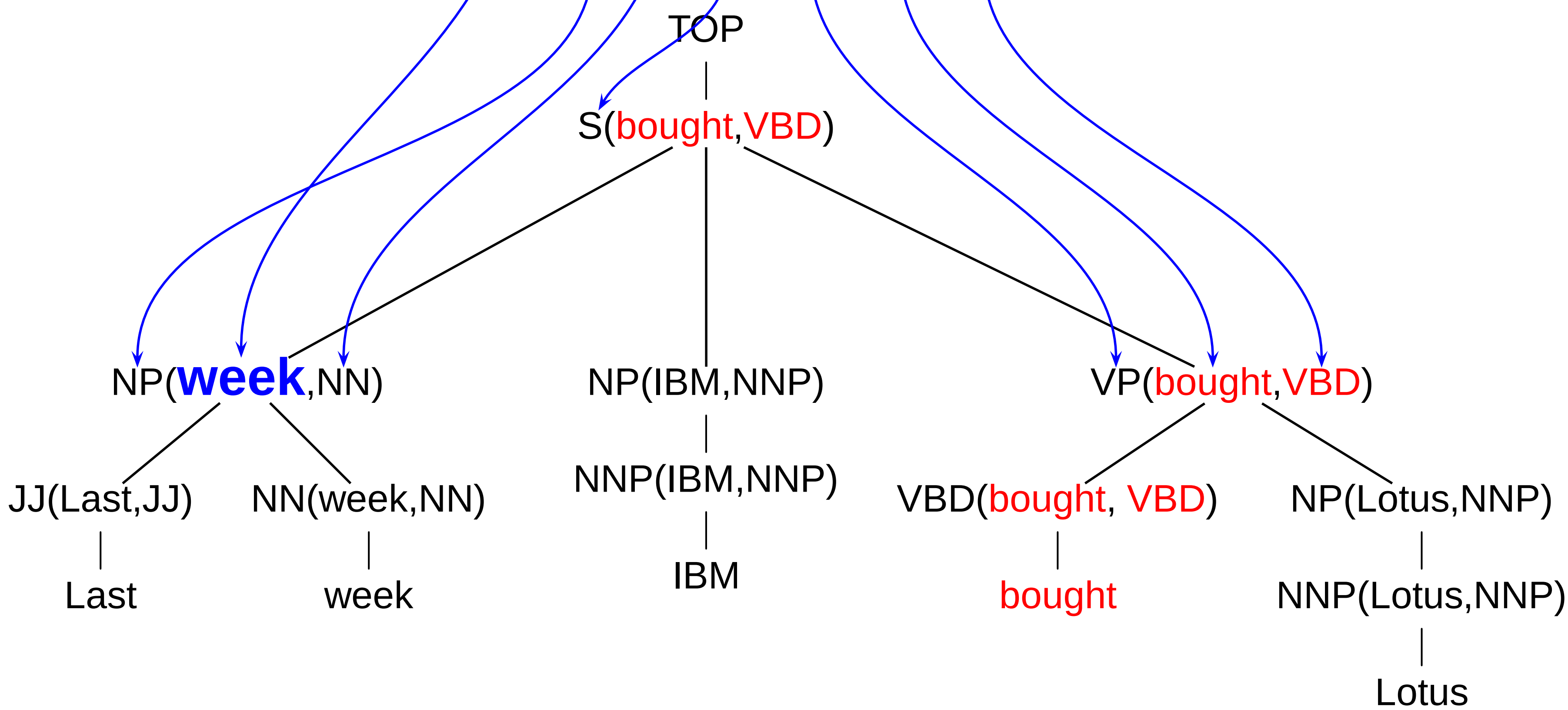
- Johnson (1998): Annotating each node with the category of its parent category boosts performance from 73.5% to 80.0% on sequences of POS tags.
- Charniak (2001) also considers parent annotation in a ME framework.
- Collins (1997,1999,2003) uses subcategorization information in his model 2.
- Gildea (2001) shows that removing *bilexical* probabilities from Collins's model 1 has only a very small negative effect on parsing quality.

Daniel Gildea's Experiment (EMNLP '01)

Collins's Model 1: $P(w_i, T_i, t_i | T_p, T_h, w_h, t_h, \Delta)$

$$= \boxed{P(w_i | T_i, t_i, T_p, T_h, w_h, t_h, \Delta)}$$

$$\times P(T_i, t_i | T_p, T_h, w_h, t_h, \Delta)$$



Daniel Gildea's Experiment (cont'd)

$$P(w_i|T_i, t_i, T_p, T_h, w_h, t_h, \Delta) \approx$$

$$\lambda_1 \bar{P}(w_i|T_i, t_i, T_p, T_h, w_h, t_h, \Delta)$$

$$+ \quad (1 - \lambda_1) \left(\lambda_2 \bar{P}(w_i|T_i, t_i, T_p, T_h, t_h, \Delta) + (1 - \lambda_2)\bar{P}(w_i|t_i)\right)$$

Daniel Gildea's Experiment (cont'd)

$$P(w_i|T_i, t_i, T_p, T_h, w_h, t_h, \Delta) \approx$$

$$\cancel{\lambda_1 \bar{P}(w_i|T_i, t_i, T_p, T_h, w_h, t_h, \Delta)}$$

$$+ \cancel{(1 - \lambda_1)} \left(\lambda_2 \bar{P}(w_i|T_i, t_i, T_p, T_h, t_h, \Delta) + (1 - \lambda_2)\bar{P}(w_i|t_i)\right)$$

		w/ bigrams		w/o bigrams	
training set	**test set**	**recall**	**prec.**	**recall**	**prec.**
WSJ	WJS	86.1	86.6	85.6	86.2
WSJ	Brown	80.3	81.0	80.3	81.0
Brown	Brown	83.6	84.6	83.5	84.4
WSJ+Brown	Brown	83.9	84.8	83.4	84.3
WSJ+Brown	WSJ	86.3	86.9	85.7	86.4

WSJ: $\sim$40k sentences/950k words; Brown: $\sim$ 22k sentences/413k words

What the Paper is About ...

How far can we get ***without*** *lexicalization?*

Why bother?

- improved baseline for unlexicalized probabilistic parsing
- insights
- smaller grammars that are easier to reason about
- faster parsing $O(n^3)$ with lower grammar constant

What's Wrong with Naïve PCFGs?

- Category symbols are too coarse; the probability distribution within the categories is not accounted for well.

  Example: A subject-NP is 8.7 times more likely than an object-NP to expand just as a pronoun.

- Training data is too sparse for accurate occurrence counts of rare rules.
  - probability of seen rare events is overestimated
  - probability of unseen rare events is underestimated

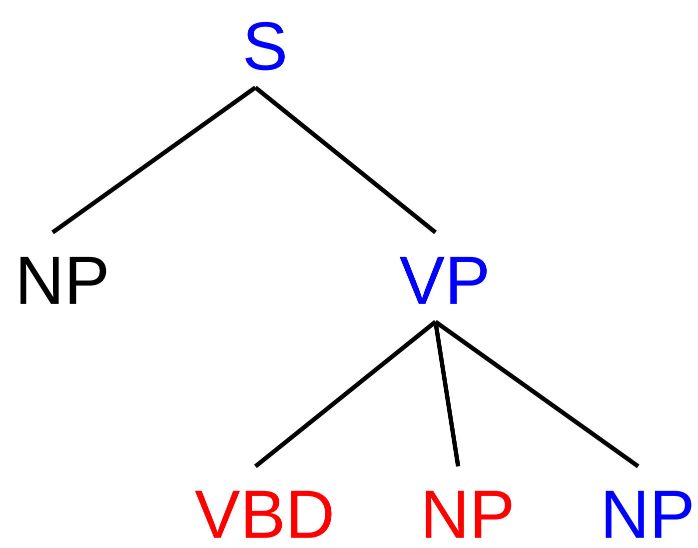
Klein & Manning’s Approach

- Vertical and horizontal “Markovization” of probabilistic estimates.
- Additional annotation of tags with information available from the trees.
- Linguistically (and empirically) motivated splitting of POS-level categories into subcategories.
- Selective splitting of categories based on information obtainable from the trees in the treebank.
- Expressly no smoothing except for POS tagging.

Markovization

Except for the root node, every node in a parse tree has

- a **vertical** history/context (parent, grandparent, etc.)
- a **horizontal** history/context



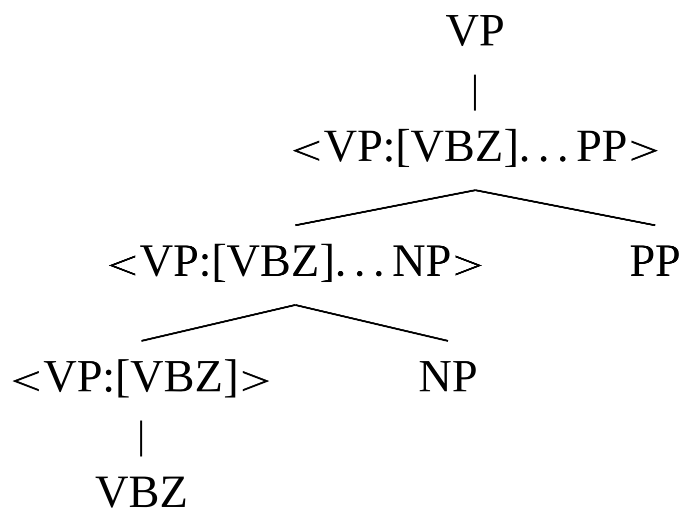
- Traditional PCFGs use the full horizontal context and a vertical context of 1.

Horizontal Markovization

- Also used by Collins (1997,1999).
- Always takes the head into account (not by definition, but as used by Collins and K&M).
- Markov assumption:

$$P(L_i|P,H,L_1,\ldots,L_{i-n+1},\ldots,L_{i-1}) \\ \qquad = P(L_i|P,H,L_{i-n+1},\ldots,L_{i-1})$$

$$P(R_i|P,H,R_1,\ldots,R_{i-n+1},\ldots,R_{i-1}) \\ \qquad = P(R_i|P,H,R_{i-n+1},\ldots,R_{i-1})$$



- Amounts to tree binarization:
  VP $\rightarrow$ VBZ NP PP PP

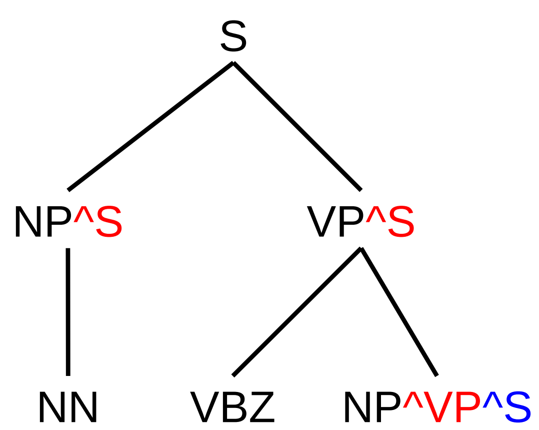
$$\begin{array}{llll} \Rightarrow & \langle\text{VP:[VBZ]}\rangle & \rightarrow & \text{VBZ} \\ & \langle\text{VP:[VBZ]} \ldots \text{NP}\rangle & \rightarrow & \langle\text{VP:[VBZ]}\rangle\ \text{NP} \\ & \langle\text{VP:[VBZ]} \ldots \text{PP}\rangle & \rightarrow & \langle\text{VP:[VBZ]} \ldots \text{NP}\rangle\ \text{PP} \end{array}$$

Vertical Markovization

- generalization of parent annotation

$$\begin{array}{lll c lll} S & \rightarrow & NP\ VP & & S & \rightarrow & NP\hat{}S\ VP\hat{}S \\ NP & \rightarrow & NN & \Rightarrow & NP\hat{}S & \rightarrow & NN \\ VP & \rightarrow & VBZ\ NP & & VP\hat{}S & \rightarrow & VBZ\ NP\hat{}VP \\ & & & & \ldots & & \end{array}$$

On a marginal note: K&M treat POS tags as terminals and discuss parent-annotation of POS-tags separately.



Markovization: Results

Vertical Order		Horizontal Markov Order $h = 0$	$h = 1$	$h \leq 2$	$h = 2$	$h = \infty$
$v = 1$	No annotation	71.27 (854)	72.5 (3119)	73.46 (3863)	72.96 (6207)	72.62 (9657)
$v \leq 2$	Sel. Parents	74.75 (2285)	77.42 (6564)	77.77 (7619)	77.50 (11398)	76.91 (14247)
$v = 2$	All Parents	74.68 (2984)	77.42 (7312)	77.81 (8367)	77.50 (12132)	76.81 (14666)
$v \leq 3$	Sel. GParents	76.50 (4943)	78.59 (12374)	79.07 (13627)	78.97 (19545)	78.54 (20123)
$v = 3$	All GParents	76.74 (7797)	79.18 (15740)	79.74 (16994)	79.07 (22886)	78.72 (22002)

Figure 2: Markovizations: F_1 and grammar size.

Figure from Klein & Manning (2003)

Markup of Unary Nodes

ˆU (external unary) “I am the only child.”

-U (internal unary) “I have only one child.”

- Roughly the same performance in isolation; in combination with other features “internal unary” is better.
- On the preterminal level (POS $\rightarrow$ word), external unary mark-up helps with
  - demonstratives (*that, this*) vs. articles (*a, the*) — both labeled as DT in Penn TreeBank
  - adverbs (e.g., *also* vs. *as well*).
- “Beyond these cases, unary tag marking was detrimental.”

Benefits of Unary Markup: Example

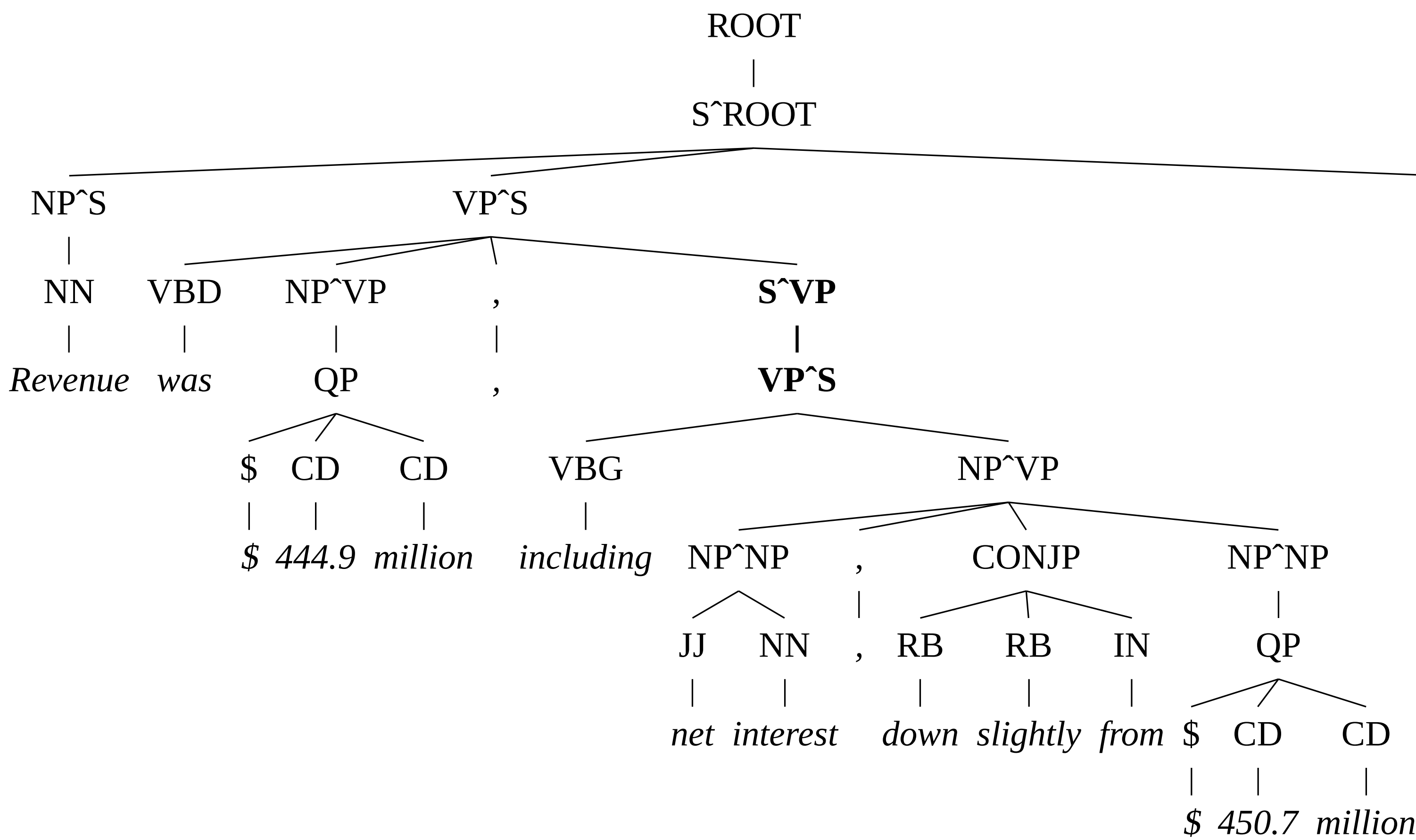


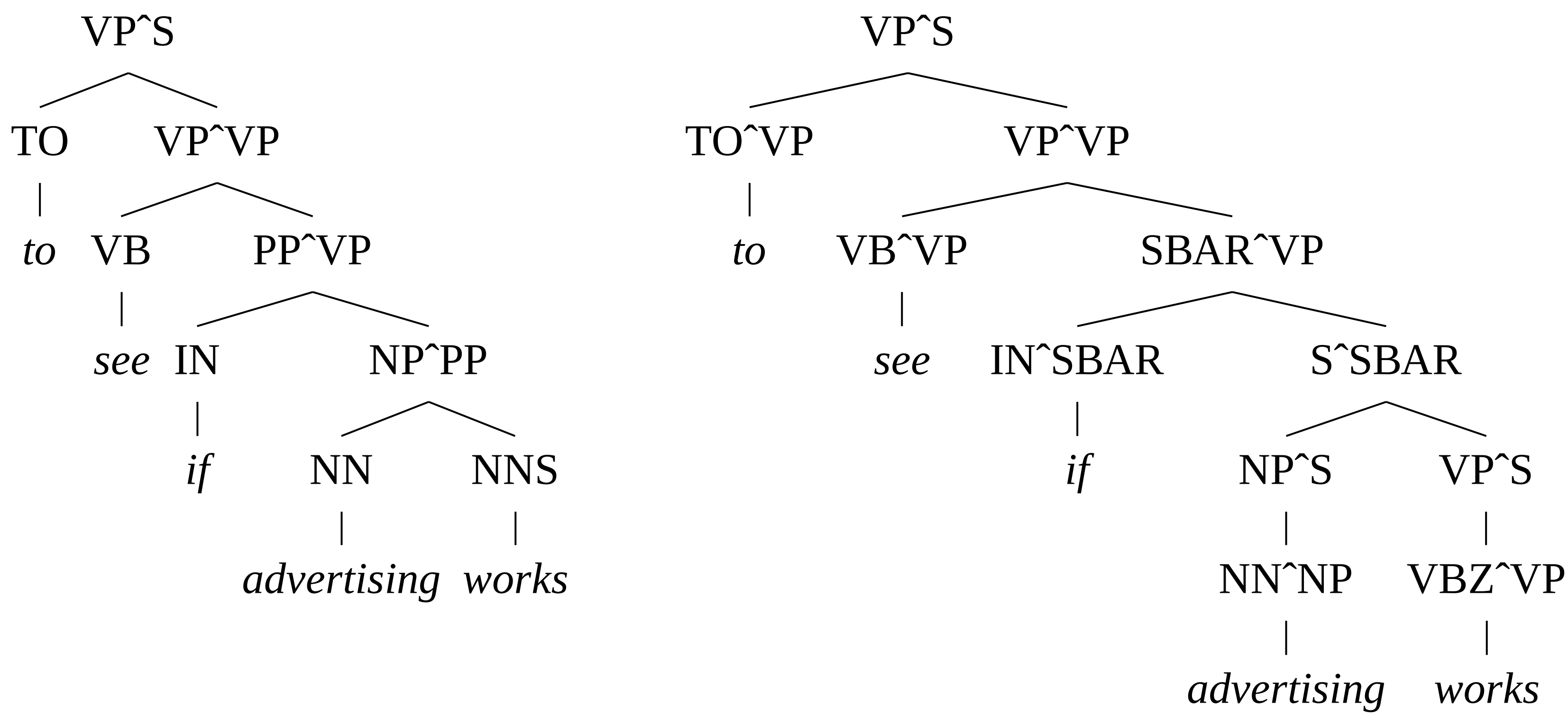
Figure 4: An error which can be resolved with the UNARY-INTERNAL annotation (incorrect baseline parse shown).

Figure from Klein & Manning (2003)

Tag Splitting

- Parent annotation also for preterminal tags.
- Splitting of IN tags into 6 linguistically motivated groups (prepositions vs. conjunctions vs. complementizers; noun-modifying vs. primarily verb-modifying prepositions (*of* vs. *as*)).
- Distinction between auxiliaries *have* and *be*.
- Special conjunction class containing *but/But* and &.
- % gets its own tag.

Benefits of TAG-PA/SPLIT-IN



(a) (b)

Figure 5: An error resolved with the TAG-PA annotation (of the IN tag): (a) the incorrect baseline parse and (b) the correct TAG-PA parse. SPLIT-IN also resolves this error.

Annotations already in the treebank

- generally hurt, with two exceptions
  - mark-up of temporal NPs (NP-TMP)
  - mark-up of sentences with a gap (GAPPED-S)

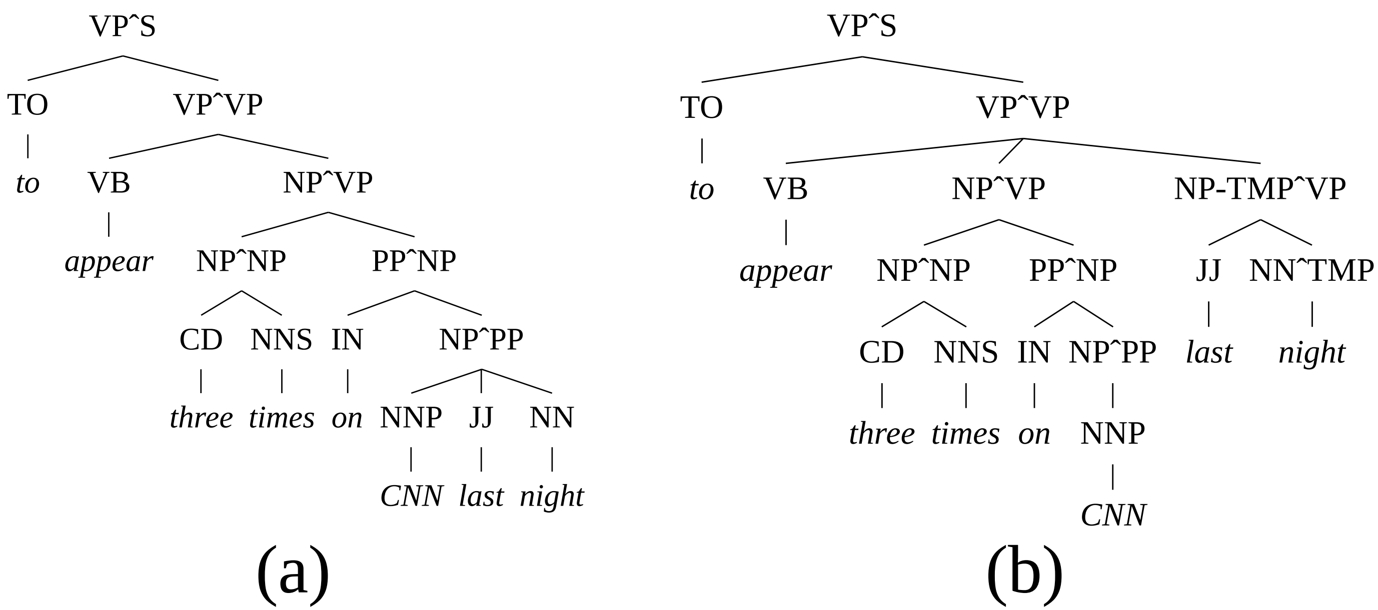


Figure 6: An error resolved with the TMP-NP annotation: (a) the incorrect baseline parse and (b) the correct TMP-NP parse.

Figure from Klein & Manning (2003)

Head Annotation

- propagates information from the head to the parent
- 2 mark-ups found particularly useful:
  - Mark-up of possessive NPs (POSS-NP).
  - Distinction between finite and non-finite VPs (SPLIT-VP).

Tackling Attachment Ambiguities

Three features found useful:

- mark-up of plain base NPs (NP $\rightarrow$ NN)
- mark-up of nodes that dominate a verb
- mark-up of NPs that contain another NP in their right periphery

Results

	Cumulative			Indiv.
Annotation	Size	F_1	Δ F_1	Δ F_1
Baseline ($v \leq 2$, $h \leq 2$)	7619	77.77	–	–
UNARY-INTERNAL	8065	78.32	0.55	0.55
UNARY-DT	8066	78.48	0.71	0.17
UNARY-RB	8069	78.86	1.09	0.43
TAG-PA	8520	80.62	2.85	2.52
SPLIT-IN	8541	81.19	3.42	2.12
SPLIT-AUX	9034	81.66	3.89	0.57
SPLIT-CC	9190	81.69	3.92	0.12
SPLIT-%	9255	81.81	4.04	0.15
TMP-NP	9594	82.25	4.48	1.07
GAPPED-S	9741	82.28	4.51	0.17
POSS-NP	9820	83.06	5.29	0.28
SPLIT-VP	10499	85.72	7.95	1.36
BASE-NP	11660	86.04	8.27	0.73
DOMINATES-V	14097	86.91	9.14	1.42
RIGHT-REC-NP	15276	87.04	9.27	1.94

Figure from Klein & Manning (2003)

Conclusions

- K&M significantly raise the baseline on unlexicalized parsing.
- Their work shows that one can recover from over-generalizations in the treebank ...
- ... and that it's worth the effort.
- Better modeling is based on linguistic analysis.
- Raises some interesting questions . . .

Questions

- What do the learning curves for unlexicalized vs. lexicalized parsing look like?
- How do the different parsers perform on out-of-domain data?
- What are the confidence intervals for the results?
- What dow the parsers still struggle with? (According to Collins (2003), coordination structures are a big problem.)