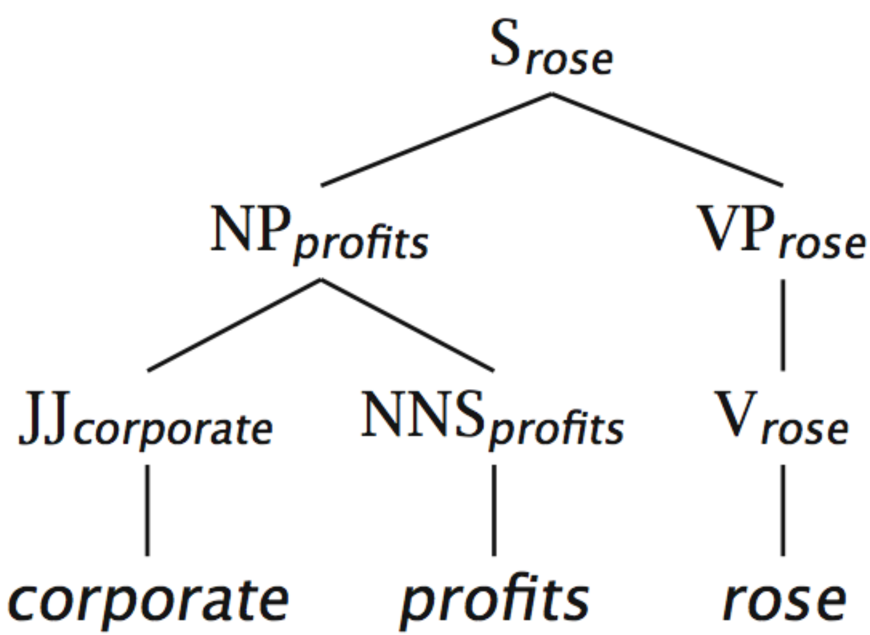
Lexicalization of PCFGs

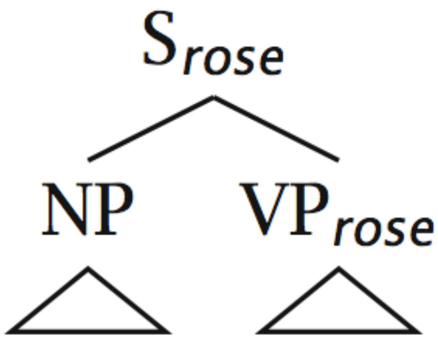
The model of Charniak (1997)

Charniak (1997)

- A very straightforward model of a lexicalized PCFG
- Probabilistic conditioning is "top-down" like a regular PCFG
  - But actual parsing is bottom-up, somewhat like the CKY algorithm we saw



Charniak (1997) example

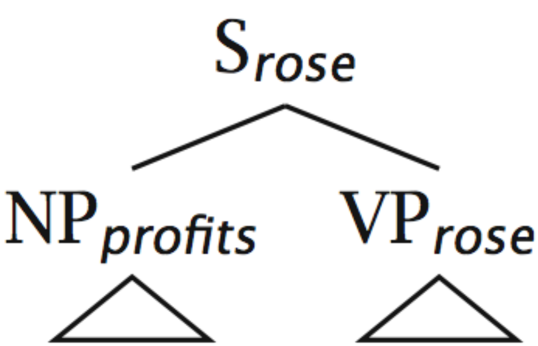


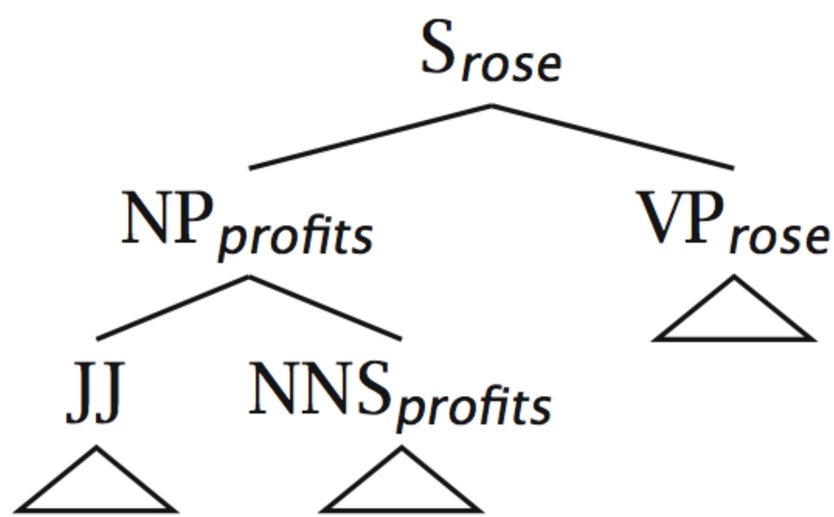
a. $h = \mathit{profits}$; c = NP

b. $ph = \mathit{rose}$; pc = S

c. $P(h|ph, c, pc)$

d. $P(r|h, c, pc)$





Lexicalization models argument selection by sharpening rule expansion probabilities

- The probability of different verbal complement frames (i.e., "subcategorizations") depends on the verb:

Local Tree	*come*	*take*	*think*	*want*
VP → V	9.5%	2.6%	4.6%	5.7%
VP → V NP	1.1%	32.1%	0.2%	13.9%
VP → V PP	34.5%	3.1%	7.1%	0.3%
VP → V SBAR	6.6%	0.3%	73.0%	0.2%
VP → V S	2.2%	1.3%	4.8%	70.8%
VP → V NP S	0.1%	5.7%	0.0%	0.3%
VP → V PRT NP	0.3%	5.8%	0.0%	0.0%
VP → V PRT PP	6.1%	1.5%	0.2%	0.0%

"monolexical" probabilities

Lexicalization sharpens probabilities: Predicting heads

"Bilexical probabilities"

- P(prices | n-plural) = .013
- P(prices | n-plural, NP) = .013
- P(prices | n-plural, NP, S) = .025
- P(prices | n-plural, NP, S, v-past) = .052
- P(prices | n-plural, NP, S, v-past, fell) = .146

Charniak (1997) linear interpolation/ shrinkage

$$\begin{aligned}\hat{P}(h|ph,c,pc) &= \lambda_1(e)P_{\text{MLE}}(h|ph,c,pc) \\ &\quad +\lambda_2(e)P_{\text{MLE}}(h|C(ph),c,pc) \\ &\quad +\lambda_3(e)P_{\text{MLE}}(h|c,pc) + \lambda_4(e)P_{\text{MLE}}(h|c)\end{aligned}$$

- $\lambda_i(e)$ is here a function of how much one would expect to see a certain occurrence, given the amount of training data, word counts, etc.
- $C(ph)$ is semantic class of parent headword
- Techniques like these for dealing with data sparseness are vital to successful model construction

Charniak (1997) shrinkage example

	$P(\text{prft}\|\text{rose}, \text{NP}, \text{S})$	$P(\text{corp}\|\text{prft}, \text{JJ}, \text{NP})$
$P(h\|ph, c, pc)$	0	0.245
$P(h\|C(ph), c, pc)$	0.00352	0.0150
$P(h\|c, pc)$	0.000627	0.00533
$P(h\|c)$	0.000557	0.00418

- Allows utilization of rich highly conditioned estimates, but smoothes when sufficient data is unavailable
- One can't just use MLEs: one commonly sees previously unseen events, which would have probability 0.

Lexicalization of PCFGs

The model of Charniak (1997)