context-free grammar $G = (N, \Sigma, S, P)$

- N finite set (nonterminals)
- Σ finite set (terminals) $N \cap \Sigma = \emptyset$
- $S \in N$ (initial nonterminal)
- P finite set (rules)

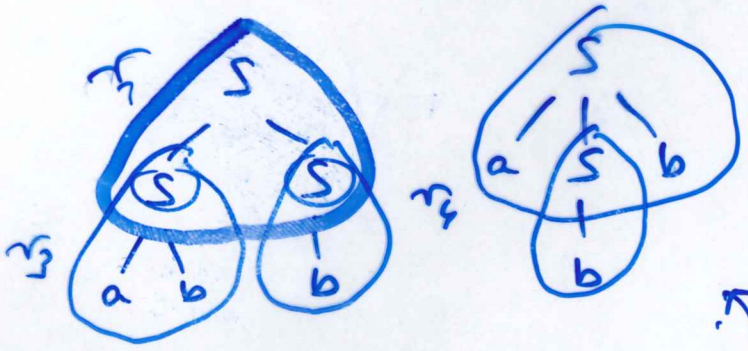
$A \to \alpha$ $A \in N$, $\alpha \in (N \cup \Sigma)^*$

Example:

r_1: $S \to SS$
r_2: $S \to aSb$
r_3: $S \to ab$
r_4: $S \to b$

$S \to A\underbrace{SB}_{S'}$
$S \Rightarrow AS' \Rightarrow ASB$

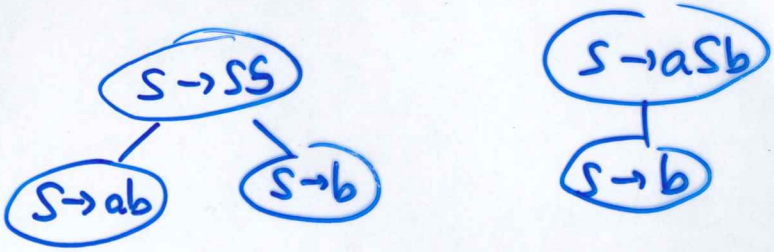
parse trees for abb:



$r_1 r_3 r_4$

bzw. $r_2 r_4$

abstract syntax trees:



probabilistic cfg (G, p)

- $G = (N, \Sigma, S, P)$ cfg
- $p: P \to [0,1]$

p is proper if $\forall A \in N: \sum_{r=(A \to \alpha) \text{ in } P} p(r) = 1$

probability of parse tree d: $d = r_1 \dots r_n$

if $d = r_1 \dots r_n$, then $P(r_1 \dots r_n) = p(r_1) \cdot \dots \cdot p(r_n)$

$P(r_1 \dots r_n \mid (G,p))$

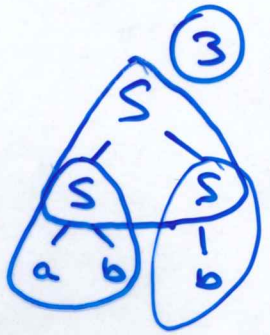
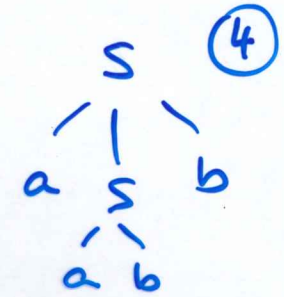
probability of sentence $w \in \Sigma^*$:

$$P(w \mid (G,p)) = \sum_{d \text{ parse tree of } w} P(d \mid (G,p))$$

$\approx$ max

grammar induction

given: corpus

$C = \{\!\{$  ,  ,  $\}\!\}$

multiset

readoff cfg: $S \to SS,\ S \to ab,\ S \to b,\ S \to aSb$

frequency analysis:

r	$S \to SS$	$S \to ab$	$S \to b$	$S \to aSb$
number of occ. of r	3	7	5	6
sum of number of occ. of all rules with lhs S	21			
p(r)	$\frac{3}{21} \approx 0.143$	$\frac{7}{21} \approx 0.33$	$\frac{5}{21} \approx 0.238$	$\frac{6}{21} \approx 0.286$

CYK - algorithm (Cocke - Younger - Kasami)

Chomsky NF

$S \to SS$

$S \to AS'$

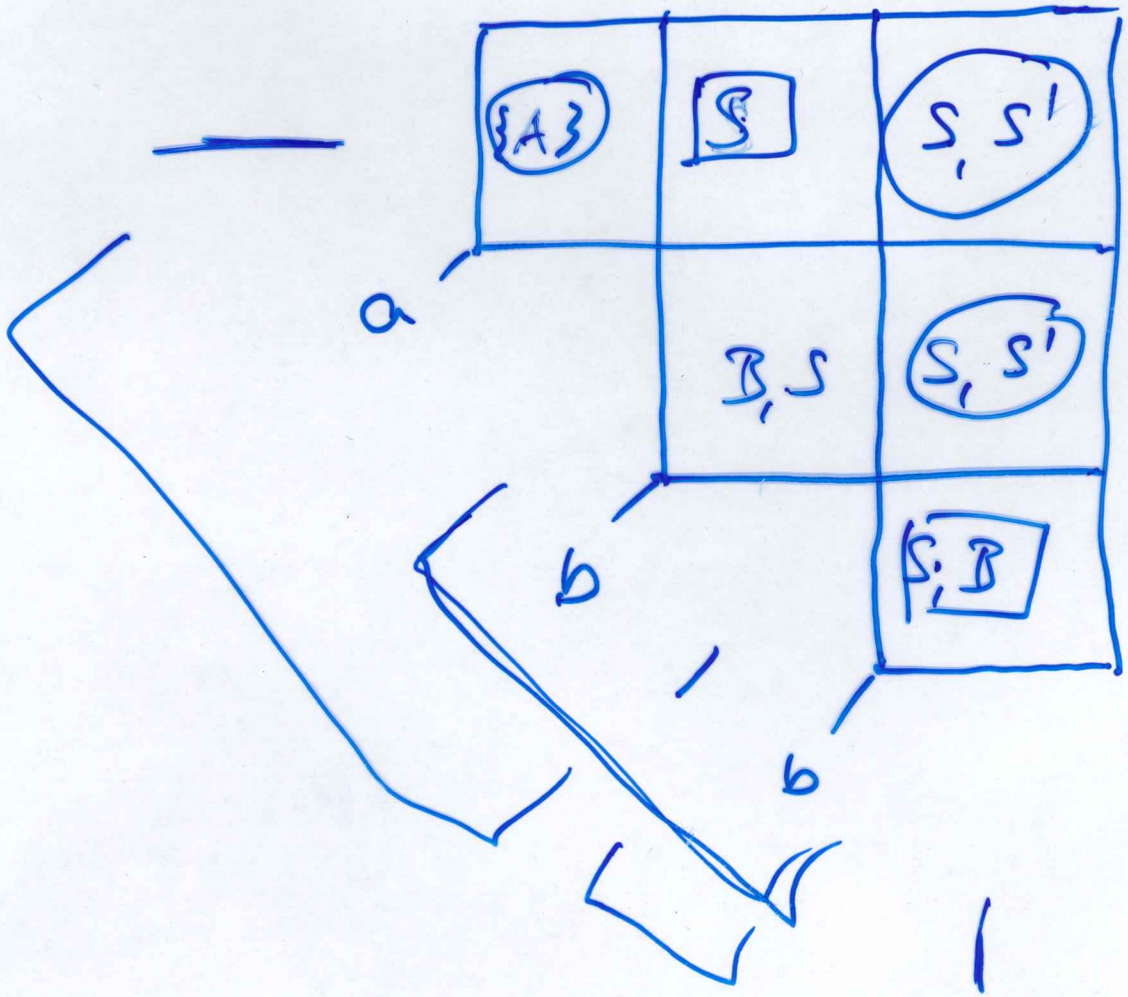
$S' \to SB$

$S \to AB$

$S \to b$

$A \to a$

$B \to b$



An adaptation of Knuth's algorithm [Knu77]

Require: PCFG $(\mathcal{G}, p)$ with $\mathcal{G} = (N, \Sigma, S, R)$ and $p(r) > 0$ for every r
Ensure: $\hat{d} = \arg\max_{d \in D_{\mathcal{G}}} P(d)$
▷ family $\delta = (\delta_A \mid A \in N)$ with $\delta_A \in (D_{\mathcal{G}}(A, \Sigma^*) \times [0,1]) \cup \{(\bot, 0)\}$, $U, Q \subseteq N$

1: **for all** $A \in N$ **do**
2: $\quad \delta_A \leftarrow (\bot, 0)$;
3: **for all** $r = (A \to u) \in R$ with $u \in \Sigma^*$ **do**
4: $\quad$ **if** $p(r) > (\delta_A)_2$ **then**
5: $\quad\quad \delta_A \leftarrow (r, p(r))$
6: $Q \leftarrow \emptyset$;
7: $U \leftarrow N$;
8: **while** $U \neq \emptyset$ **do**
9: $\quad A \leftarrow \arg\max_{B \in U} (\delta_B)_2$;
10: $\quad U \leftarrow U \setminus \{A\}$;
11: $\quad Q \leftarrow Q \cup \{A\}$;
12: $\quad$ **for all** $r = (B \to u_0 B_1 u_1 \dots B_k u_k) \in R$ **do**
13: $\quad\quad$ **if** $B \in U$ and $A \in \{B_1, \dots, B_k\} \subseteq Q$ **then**
14: $\quad\quad\quad d := r((\delta_{B_1})_1, \dots, (\delta_{B_k})_1)$; ▷ derivation viewed as abstract syntax tree
15: $\quad\quad\quad s := p(r) \cdot \prod_{i=1}^{k} (\delta_{B_i})_2$;
16: $\quad\quad\quad$ **if** $s \geq (\delta_B)_2$ **then**
17: $\quad\quad\quad\quad \delta_B \leftarrow (d, s)$
18: $\hat{d} \leftarrow (\delta_S)_1$