

CS626: Speech, Natural Language Processing and the Web

Parsing

Pushpak Bhattacharyya
Computer Science and Engineering
Department
IIT Bombay

Week 6 of 29th August, 2022

Algorithmics of Parsing

Problem Statement

INPUT: (a) grammar rules, (b) input sentence

OUTPUT: Parse Tree
(Constituency/Dependency)

Top Down

- Start with the *S* symbol and draw its children: say, *NP* and *VP*, assuming the input to be a declarative sentence.
- Now the subtrees under *NP*, followed by that under the *VP* are developed.
- For example, $NP \rightarrow DT\ NN$ could be applied.
- After this, only POS tags will need to be resolved. *DT* will absorb, say, the word '*the*' in the input and *NN*, '*man*'.
- This will complete constructing the *NP* subtree.
- Similarly, *VP* subtree also will be constructed.

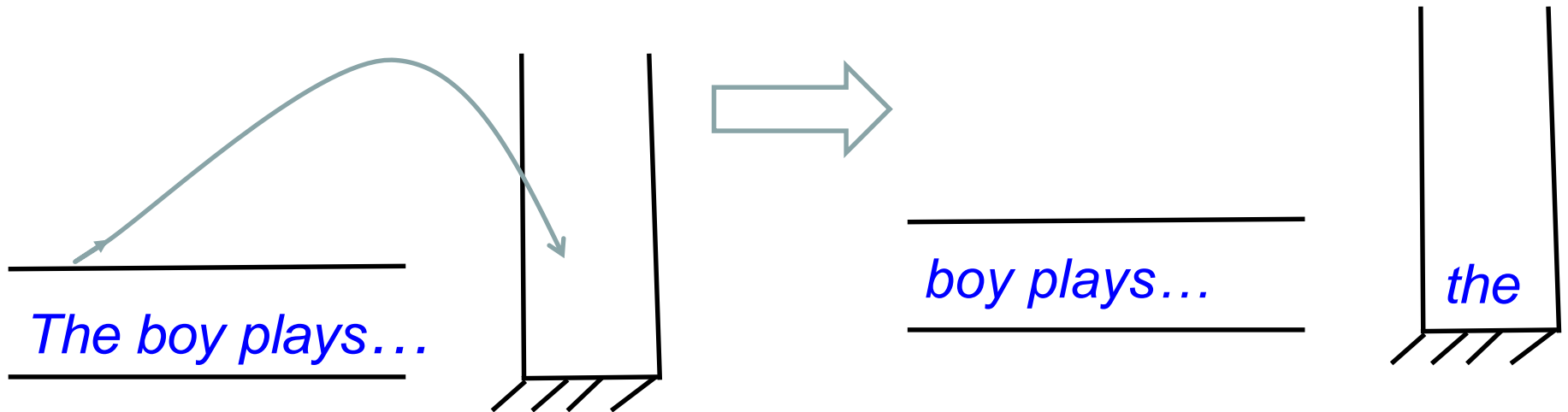
Bottom Up

- The words are resolved to their POS tags.
- Then POS tags are combined by constituency rules, e.g., $NP \rightarrow DT\ NN$. Generated non terminals are then attempted to be combined.
- For example, after generating JJP , NP they are combined to form a bigger NP , by applying $NP \rightarrow JJP\ NP$.

Main Operations

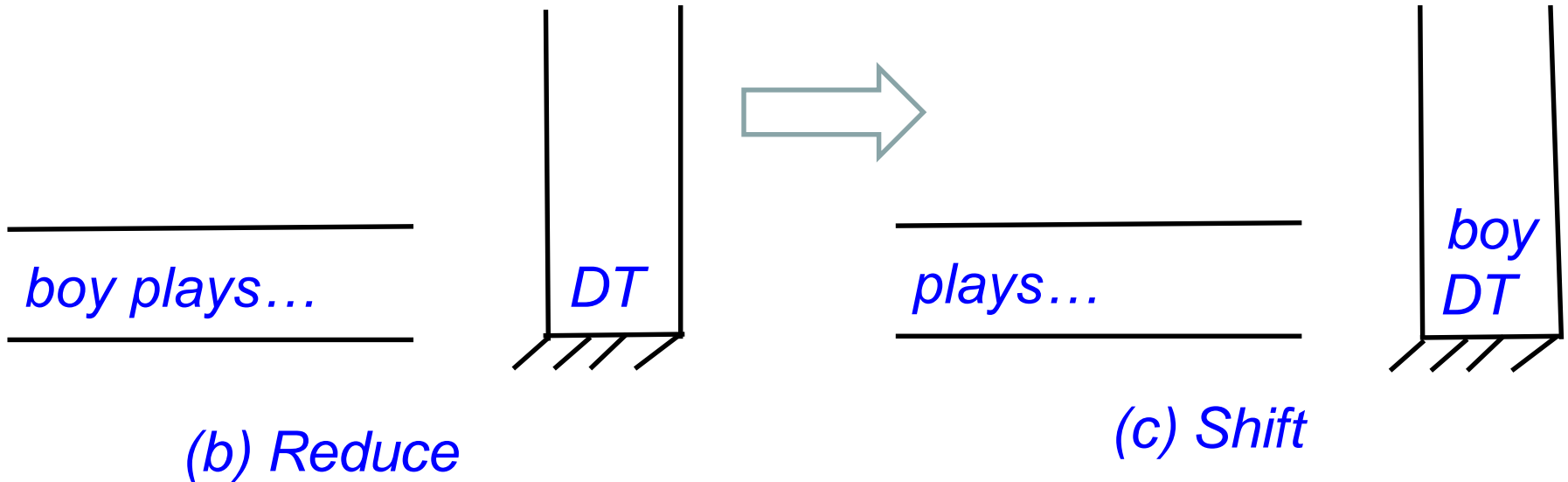
- Doing a left to right scan of the input sentence
- At every word, deciding if the word should (a) create a new constituent or (b) wait until more words get a look-in to create a constituent, and
- On creation of a new constituent, examining if the new constituent can be merged with an adjacent one to form a bigger constituent.

Shift Reduce (1/3)

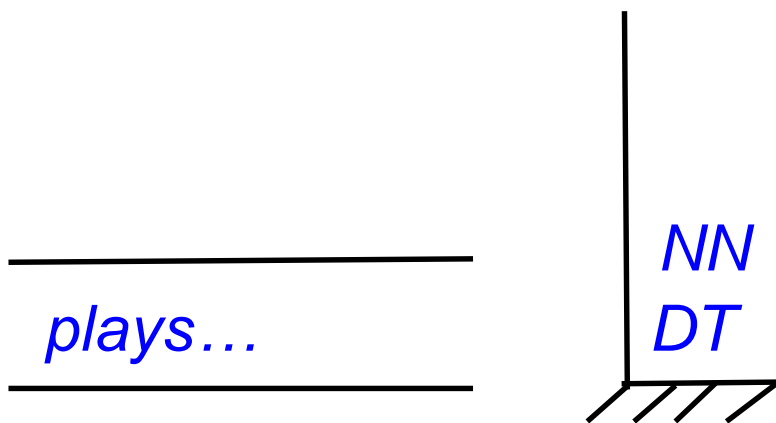


(a) Shift

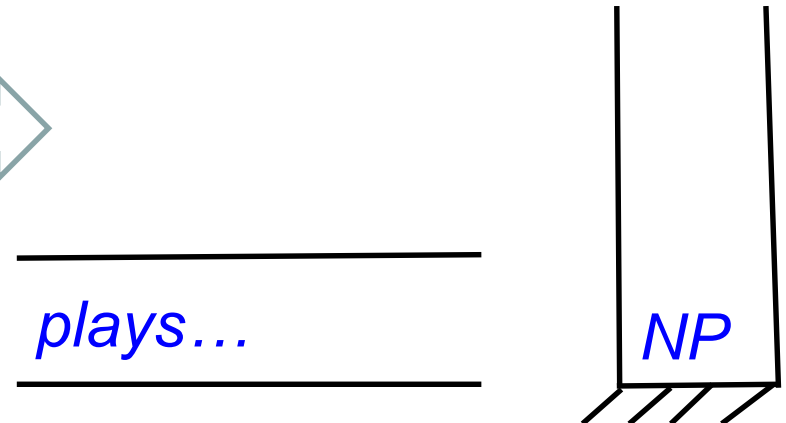
Shift Reduce (2/3)



Shift Reduce (3/3)



(d) Reduce



(e) Reduce

Grammar Rules

A segment of English

- $S \rightarrow NP VP$
- $NP \rightarrow DT NN$
- $NP \rightarrow NNS$
- $NP \rightarrow NP PP$
- $PP \rightarrow P NP$
- $VP \rightarrow VP PP$
- $VP \rightarrow VBD NP$
- $DT \rightarrow \text{the}$
- $NN \rightarrow \text{gunman}$
- $NN \rightarrow \text{building}$
- $VBD \rightarrow \text{sprayed}$
- $NNS \rightarrow \text{bullets}$

GENERATIVE GRAMMAR, due
to Noam Chomsky

Foundational Question

- Grammar rules are context free grammar (CFG) rules
- Is CFG enough powerful to capture language?
- CFG cannot accept/generate $a^n b^n c^n$
- Corresponding language phenomenon: Jack, Mykel and Mohan play tennis, soccer and cricket respectively.

CYK Parsing: Start with (0,1)

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------|-------|-------|-------|-------|-------|-------|---|
| 0 | DT | | | | | | |
| 1 | ----- | | | | | | |
| 2 | ----- | ----- | | | | | |
| 3 | ----- | ----- | ----- | | | | |
| 4 | ----- | ----- | ----- | ----- | | | |
| 5 | ----- | ----- | ----- | ----- | ----- | | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |

CYK: Keep filling diagonals

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------|-------|-------|-------|-------|-------|-------|---|
| 0 | DT | | | | | | |
| 1 → | ----- | NN | | | | | |
| 2 ↓ | ----- | ----- | | | | | |
| 3 | ----- | ----- | ----- | | | | |
| 4 | ----- | ----- | ----- | ----- | | | |
| 5 | ----- | ----- | ----- | ----- | ----- | | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |



CYK: Try getting higher level structures

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------|-------|-------|-------|-------|-------|-------|---|
| 0 | DT | NP | | | | | |
| 1 → | ----- | NN | | | | | |
| 2 ↓ | ----- | ----- | | | | | |
| 3 | ----- | ----- | ----- | | | | |
| 4 | ----- | ----- | ----- | ----- | | | |
| 5 | ----- | ----- | ----- | ----- | ----- | | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |

CYK: Diagonal continues

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|-------|-------|-------|-------|-------|-------|---|
| 0 | DT | NP | | | | | |
| 1  | ----- | NN | | | | | |
| 2  | ----- | ----- | VBD | | | | |
| 3 | ----- | ----- | ----- | | | | |
| 4 | ----- | ----- | ----- | ----- | | | |
| 5 | ----- | ----- | ----- | ----- | ----- | | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |



CYK (cont...)

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------|-------|-------|-------|-------|-------|-------|---|
| 0 → | DT | NP | ----- | | | | |
| 1 ↓ | ----- | NN | ----- | | | | |
| 2 | ----- | ----- | VBD | | | | |
| 3 | ----- | ----- | ----- | | | | |
| 4 | ----- | ----- | ----- | ----- | | | |
| 5 | ----- | ----- | ----- | ----- | ----- | | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |

CYK (cont...)

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|-------|-------|-------|-------|-------|-------|---|
| 0  | DT | NP | ----- | | | | |
| 1 | ----- | NN | ----- | | | | |
| 2  | ----- | ----- | VBD | | | | |
| 3 | ----- | ----- | ----- | DT | | | |
| 4 | ----- | ----- | ----- | ----- | | | |
| 5 | ----- | ----- | ----- | ----- | ----- | | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |



CYK (cont...)

0 The 1 gunman 2 sprayed 3 the 4 building 5 with 6 bullets 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------|-------|-------|-------|-------|-------|-------|---|
| 0 → | DT | NP | ----- | ----- | | | |
| 1 ↓ | ----- | NN | ----- | ----- | | | |
| 2 | ----- | ----- | VBD | ----- | | | |
| 3 | ----- | ----- | ----- | DT | | | |
| 4 | ----- | ----- | ----- | ----- | NN | | |
| 5 | ----- | ----- | ----- | ----- | ----- | | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |

CYK: starts filling the 5th column

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|-------|-------|-------|-------|-------|-------|---|
| 0 | DT | NP | ----- | ----- | | | |
| 1  | ----- | NN | ----- | ----- | | | |
| 2  | ----- | ----- | VBD | ----- | | | |
| 3 | ----- | ----- | ----- | DT | NP | | |
| 4 | ----- | ----- | ----- | ----- | NN | | |
| 5 | ----- | ----- | ----- | ----- | ----- | | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |

CYK (cont...)

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------|-------|-------|-------|-------|-------|-------|---|
| 0 | DT | NP | ----- | ----- | | | |
| 1 | ----- | NN | ----- | ----- | | | |
| 2 | ----- | ----- | VBD | ----- | VP | | |
| 3 | ----- | ----- | ----- | DT | NP | | |
| 4 | ----- | ----- | ----- | ----- | NN | | |
| 5 | ----- | ----- | ----- | ----- | ----- | | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |

CYK (cont...)

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------|-------|-------|-------|-------|-------|-------|---|
| 0 | DT | NP | ----- | ----- | | | |
| 1 | ----- | NN | ----- | ----- | ----- | | |
| 2 | ----- | ----- | VBD | ----- | VP | | |
| 3 | ----- | ----- | ----- | DT | NP | | |
| 4 | ----- | ----- | ----- | ----- | NN | | |
| 5 | ----- | ----- | ----- | ----- | ----- | | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |

CYK: S found, but NO termination!

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------|-------|-------|-------|-------|-------|-------|---|
| 0 | DT | NP | ----- | ----- | S | | |
| 1 | ----- | NN | ----- | ----- | ----- | | |
| 2 | ----- | ----- | VBD | ----- | VP | | |
| 3 | ----- | ----- | ----- | DT | NP | | |
| 4 | ----- | ----- | ----- | ----- | NN | | |
| 5 | ----- | ----- | ----- | ----- | ----- | | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |



CYK (cont...)

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|------------|-------|-------|-------|-------|-------|-------|---|
| 0 | DT | NP | ----- | ----- | S | | |
| 1 | ----- | NN | ----- | ----- | ----- | | |
| 2 | ----- | ----- | VBD | ----- | VP | | |
| 3 | ----- | ----- | ----- | DT | NP | | |
| 4 | ----- | ----- | ----- | ----- | NN | | |
| 5 | ----- | ----- | ----- | ----- | ----- | P | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |

CYK (cont...)

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.

| To From | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|-------|-------|-------|-------|-------|-------|---|
| 0 | DT | NP | ----- | ----- | S | ----- | |
| 1  | ----- | NN | ----- | ----- | ----- | ----- | |
| 2 | ----- | ----- | VBD | ----- | VP | ----- | |
| 3  | ----- | ----- | ----- | DT | NP | ----- | |
| 4 | ----- | ----- | ----- | ----- | NN | ----- | |
| 5 | ----- | ----- | ----- | ----- | ----- | P | |
| 6 | ----- | ----- | ----- | ----- | ----- | ----- | |

[illegible]

CYK (cont...)

0 The 1 gunman 2 sprayed 3 the 4 building 5 with 6 bullets 7.

[illegible]

CYK (cont...)

0 The 1 gunman 2 sprayed 3 the 4 building 5 with 6 bullets 7.

[illegible]

[illegible]

[illegible]

0 The 1 gunman 2 sprayed 3 the 4 building 5 with 6 bullets 7.

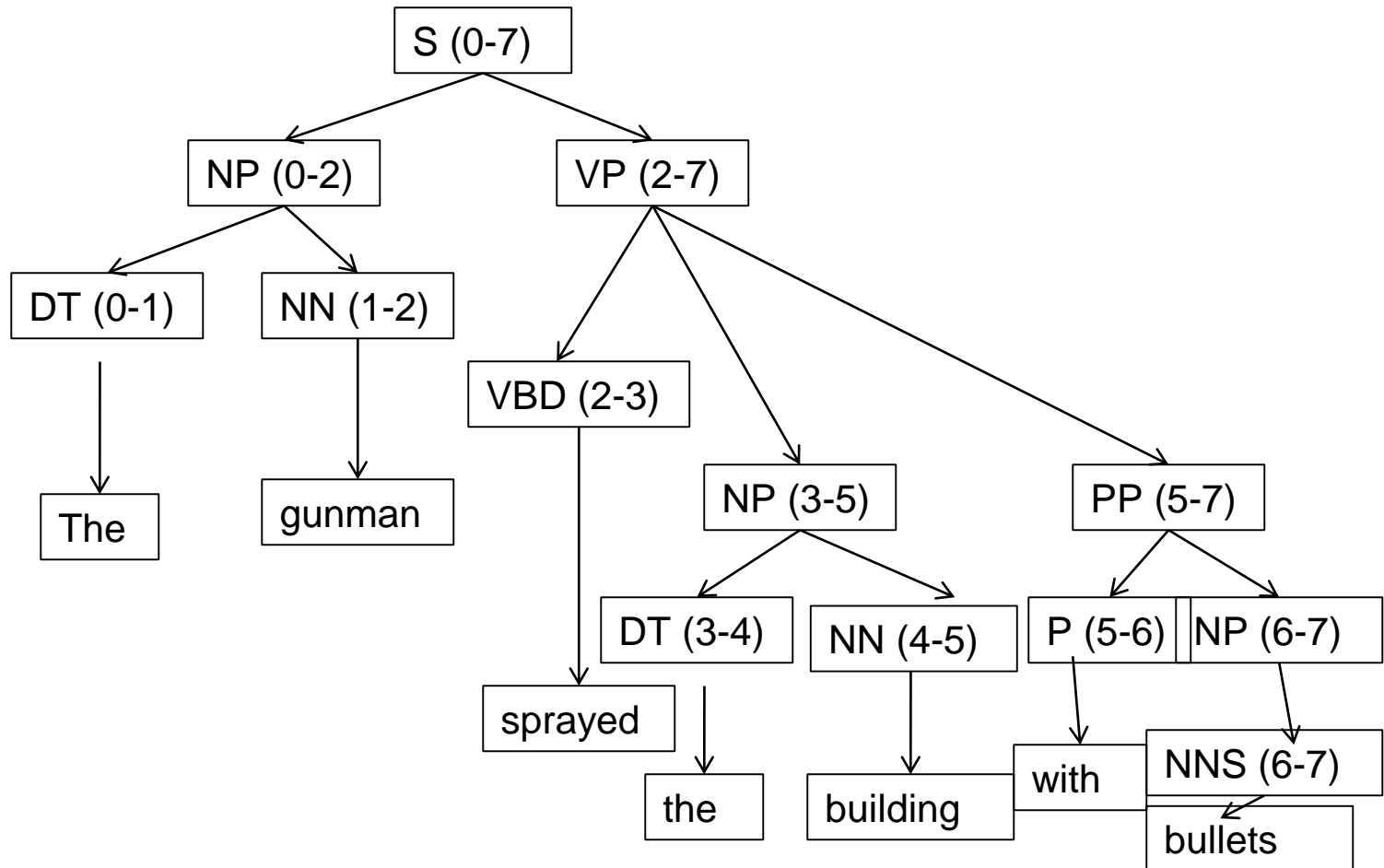
[illegible]

CYK: Extracting the Parse Tree

- The parse tree is obtained by keeping back pointers.

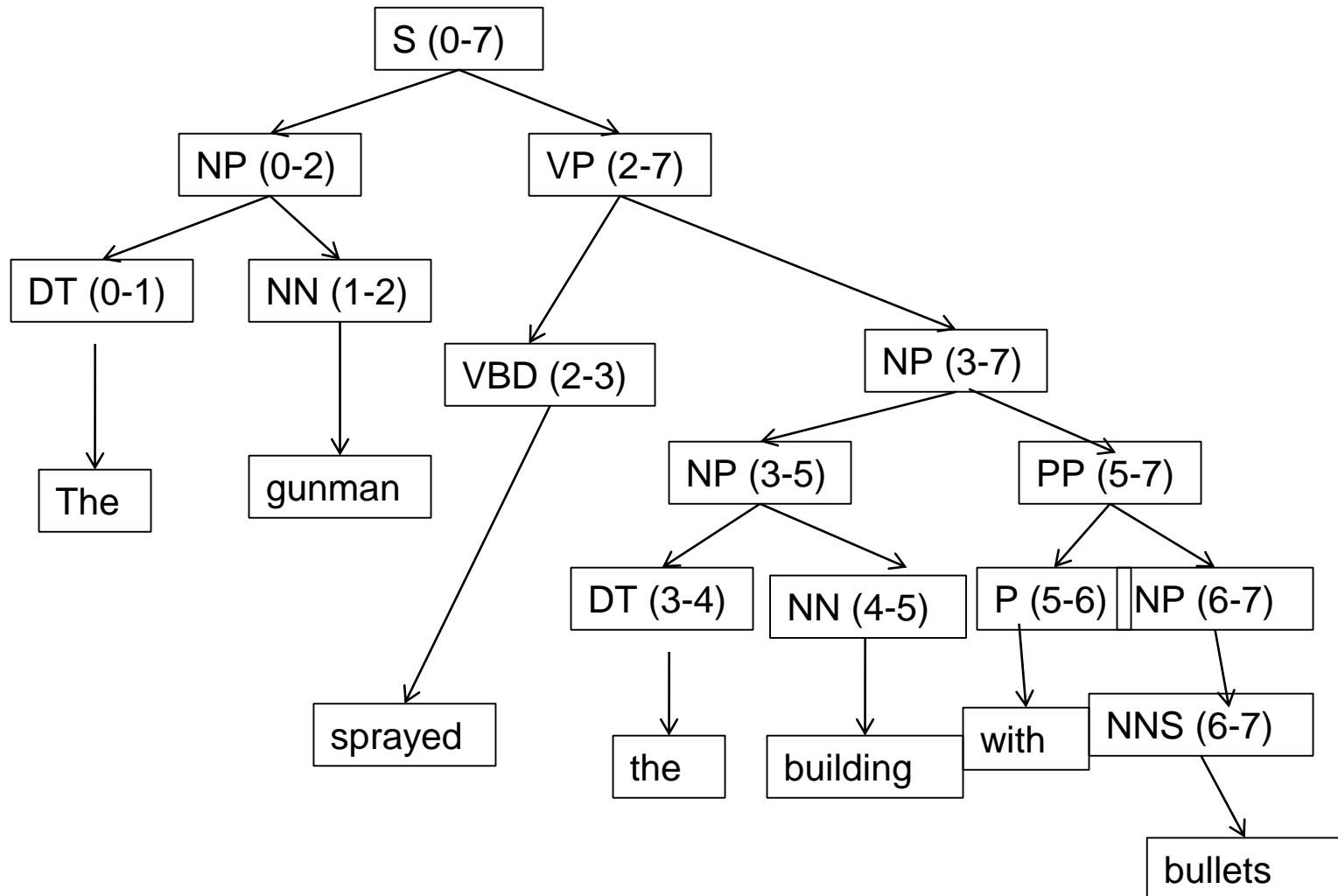
Parse Tree #1

0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.



Parse Tree #2

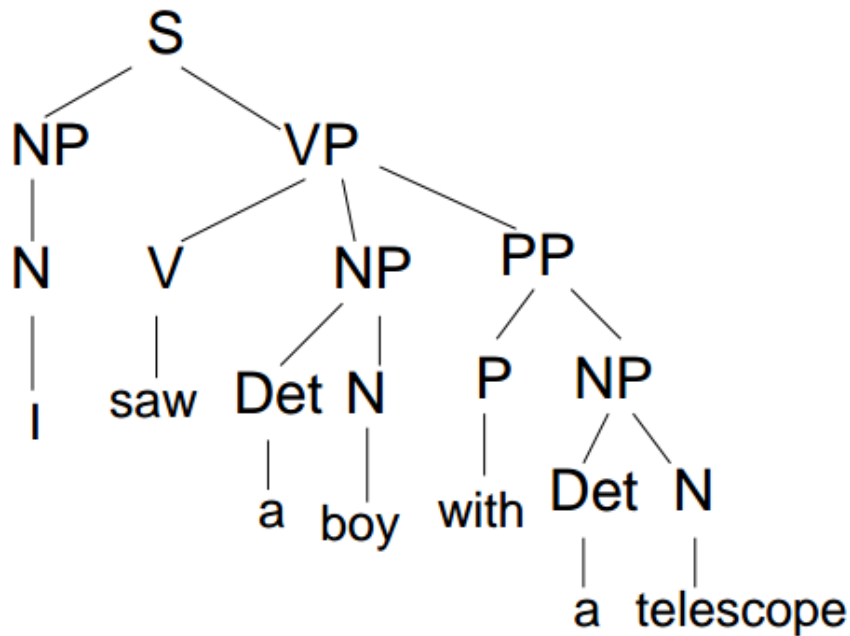
0 *The* 1 *gunman* 2 *sprayed* 3 *the* 4 *building* 5 *with* 6 *bullets* 7.



Notion of Domination

- A sentence is dominated by the symbol S through domination of segments by phrases
- Analogy
 - The capital of a country dominates the whole country.
 - The capital of a state dominates the whole state.
 - The district headquarter dominates the district.

Domination: Example



- Dominations
 - NP dominates “a telescope”
 - VP dominates “saw a boy with a telescope
 - S dominates the whole sentence
- Domination is composed of many sub-domination.
- I saw a boy with a telescope
 - Meaning: I used the telescope to see the boy

Probabilistic parsing

Main source:

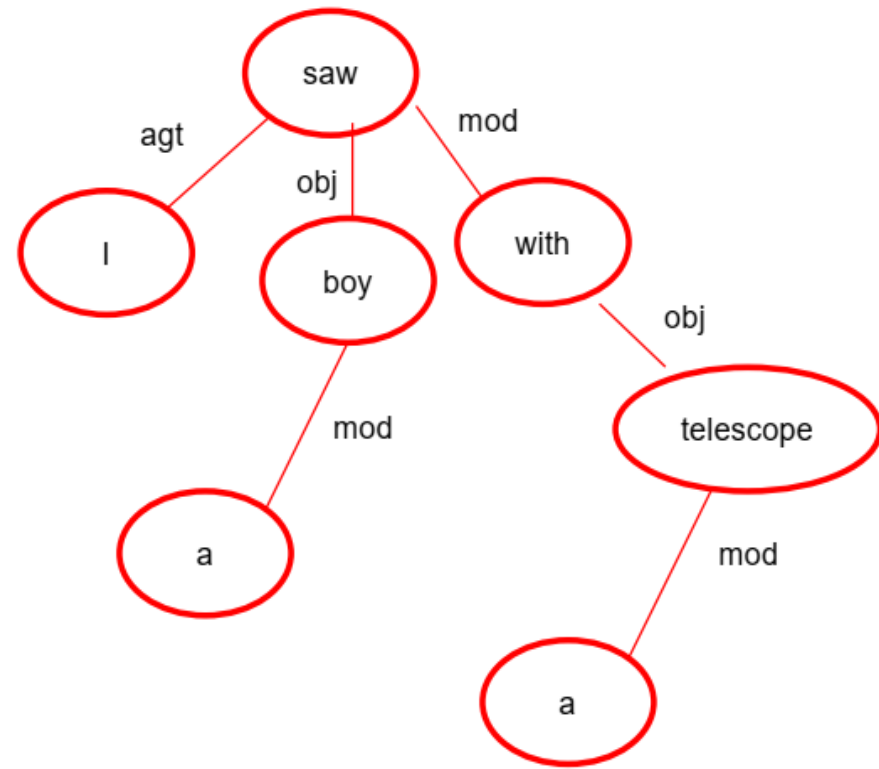
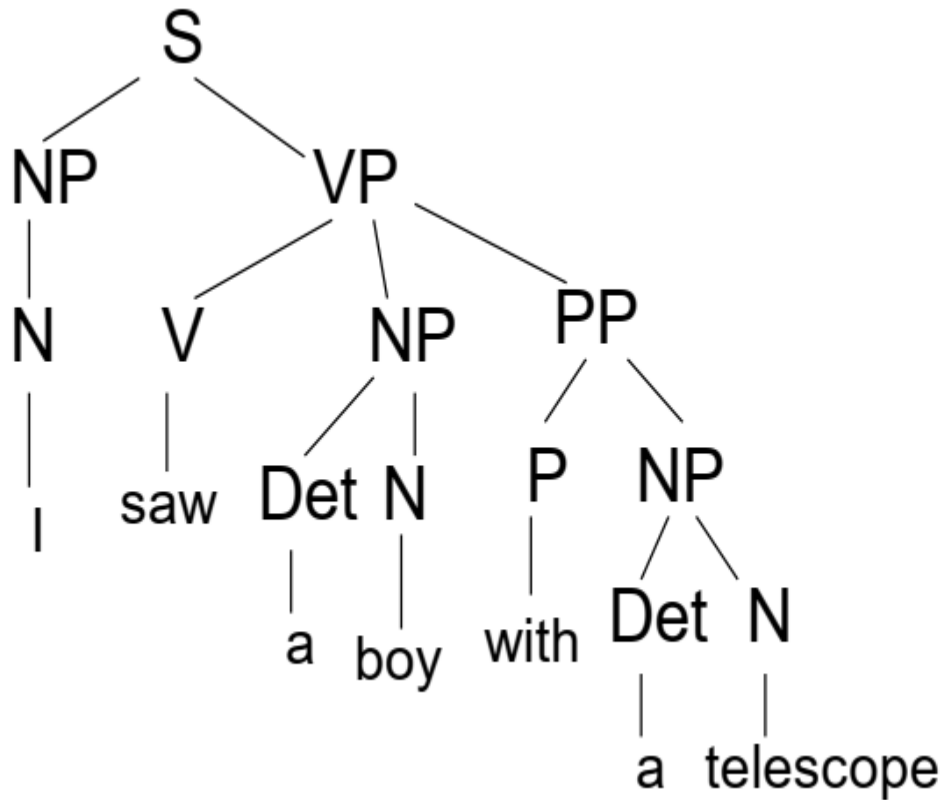
Christopher Manning and Heinrich Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

Noisy Channel Modeling



$$\begin{aligned}
 T^* &= \underset{T}{\operatorname{argmax}} [P(T|S)] \\
 &= \underset{T}{\operatorname{argmax}} [P(T).P(S|T)] \\
 &= \underset{T}{\operatorname{argmax}} [P(T)], \text{ since given the parse the sentence is completely determined and } P(S|T)=1
 \end{aligned}$$

“I saw...”: CP and DP #1

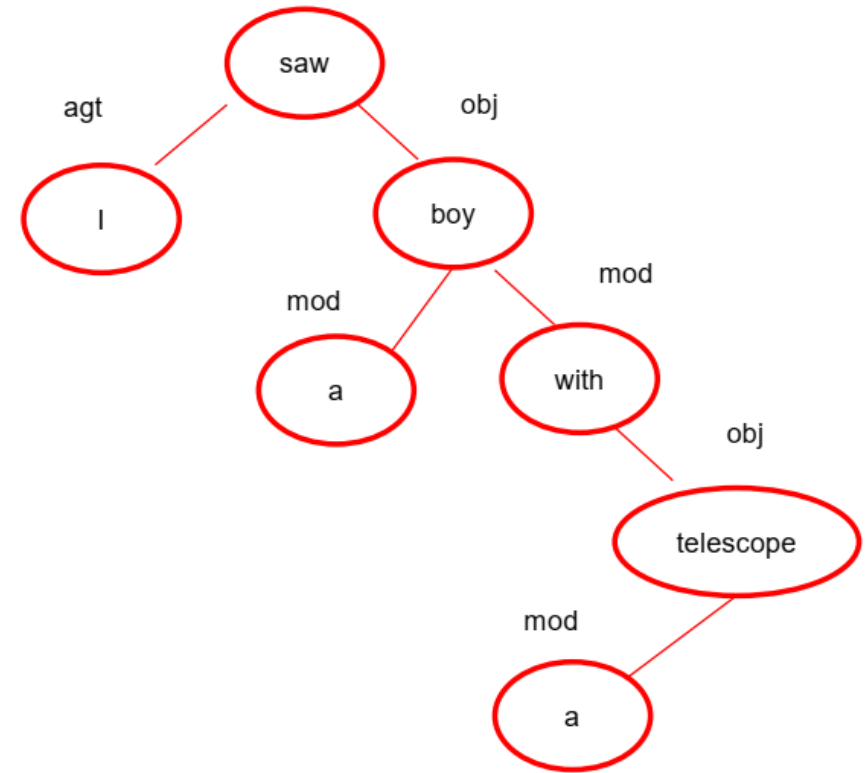
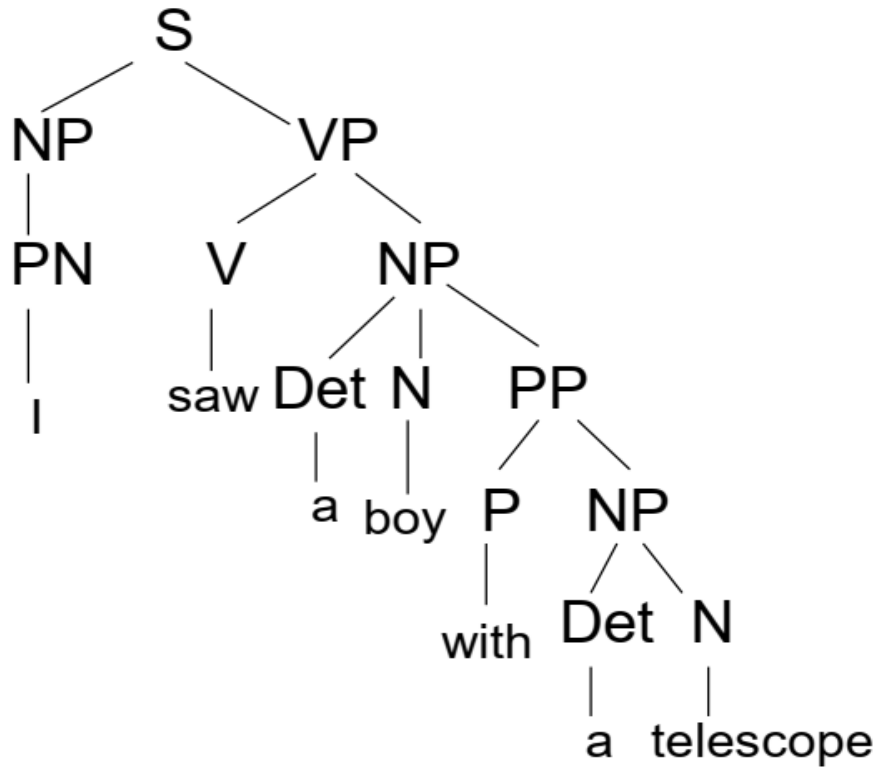


Bracketed Structure #1

Parse #1 (meaning: I have the telescope)

[
 [I]_{NP}
 [
 [saw]_{VBD}
 [the boy]_{NP}
 [with [a telescope]_{NP}]_{PP}
]_{VP}
]_S

“I saw...”: CP and DP #2



Bracketed Structure #2

Parse #2 (meaning: the boy has the telescope)

[
 [I]_{NP}
 [
 [saw]_{VBD}
 [
 [the boy]_{NP}
 [with [a telescope]_{NP}]_{PP}
]_{NP}
]_{VP}
]_S

Formal Definition of PCFG

- A set of terminals $\{w_k\}$, $k = 1, \dots, V$
 $\{w_k\} = \{ \text{child, teddy, bear, played...} \}$
- A set of non-terminals $\{N^i\}$, $i = 1, \dots, n$
 $\{N^i\} = \{ \text{NP, VP, DT...} \}$
- A designated start symbol S (sometimes given the symbol N^1)
- A set of rules $\{N^i \rightarrow \zeta^j\}$, where ζ^j is a sequence of terminals & non-terminals
e.g., $\text{NP} \rightarrow \text{DT NN}$
- A corresponding set of rule probabilities

Rule Probabilities

- Rule probabilities are such that for the same non terminal all production rules sum to 1.

$$E.g., P(NP \rightarrow DT NN) = 0.2$$

$$P(NP \rightarrow NNS) = 0.5$$

$$P(NP \rightarrow NP PP) = 0.3$$

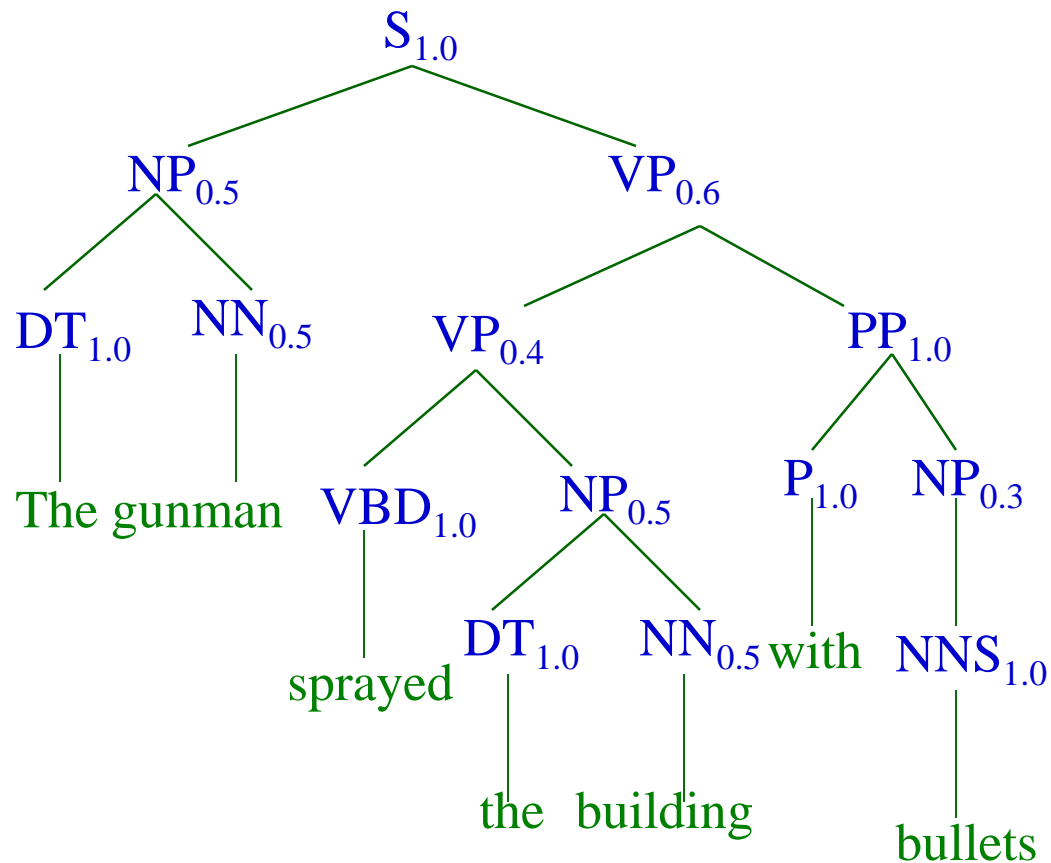
- Meaning of $P(NP \rightarrow DT NN) = 0.2$, 20% of the training data parses use the rule $NP \rightarrow DT NN$

Probabilistic Context Free Grammars

- $S \rightarrow NP VP$ 1.0
- $NP \rightarrow DT NN$ 0.5
- $NP \rightarrow NNS$ 0.3
- $NP \rightarrow NP PP$ 0.2
- $PP \rightarrow P NP$ 1.0
- $VP \rightarrow VP PP$ 0.6
- $VP \rightarrow VBD NP$ 0.4
- $DT \rightarrow the$ 1.0
- $NN \rightarrow gunman$ 0.5
- $NN \rightarrow building$ 0.5
- $VBD \rightarrow sprayed$ 1.0
- $NNS \rightarrow bullets$ 1.0

Example Parse t_1

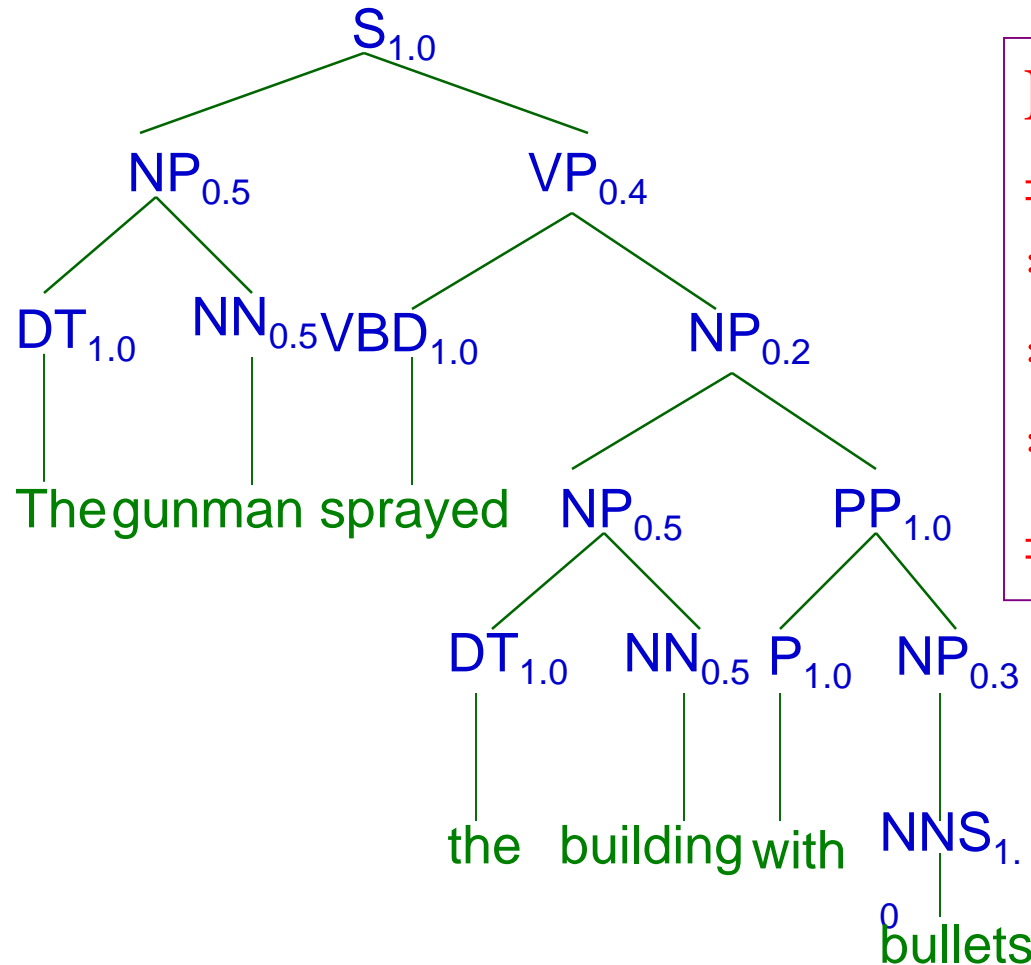
The gunman sprayed the building with bullets.



$$\begin{aligned}
 P(t_1) &= 1.0 * 0.5 * 1.0 \\
 &* 0.5 * 0.6 * 0.4 * 1.0 \\
 &* 0.5 * 1.0 * 0.5 * 1.0 \\
 &* 1.0 * 0.3 * 1.0 \\
 &= 0.00225
 \end{aligned}$$

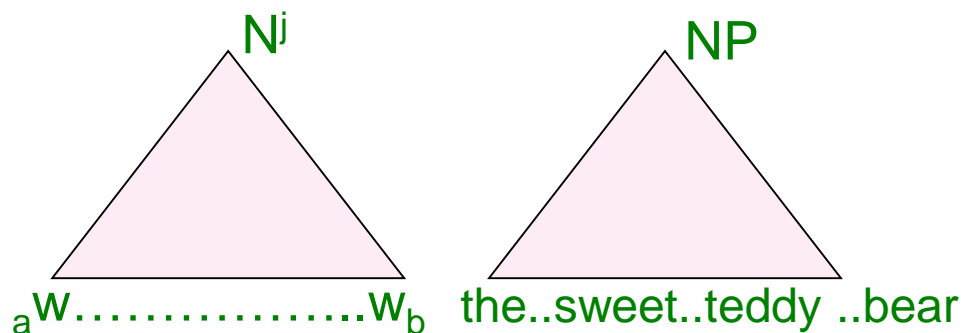
Another Parse t_2

The gunman sprayed the building with bullets.



$$\begin{aligned}
 P(t_2) &= 1.0 * 0.5 * 1.0 * 0.5 \\
 &\quad * 0.4 * 1.0 * 0.2 * 0.5 \\
 &\quad * 1.0 * 0.5 * 1.0 * 1.0 \\
 &\quad * 0.3 * 1.0 \\
 &= 0.0015
 \end{aligned}$$

Probability of a sentence (1/2)



Notation: (a, b etc. are BETWEEN-word indices)

- w_{ab} – subsequence $_a w \dots w_b$
- N_j dominates $_a w \dots w_b$
or $\text{yield}(N_j) = _a w \dots w_b$

Probability of a sentence (2/2)

Probability of a sentence = $P(w_{0,l})$

(0 is the index before the first word and l the index after the last word. All other indices are between words)

$$=\sum_t (P(w_{0,l} | t))$$

$$=\sum_t (P(t) \cdot P(w_{0,l} | t))$$

$$=\sum_t P(t) \cdot 1$$

where t is a parse tree of the sentence

If t is a parse tree for the sentence $w_{0,l}$, this will be 1 !!

Assumptions of the PCFG model

- **Place invariance:**

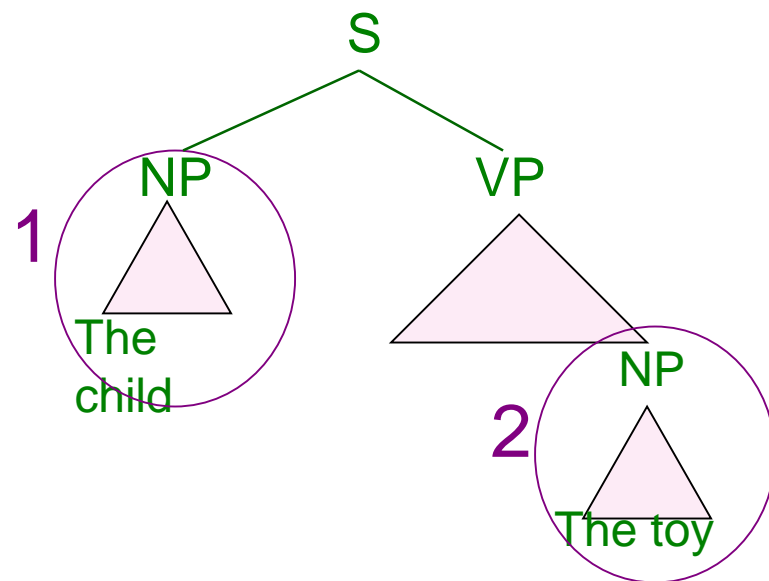
$P(\text{NP} \rightarrow \text{DT NN})$ is same
independent of location in the
tree

- **Context-free:**

$P(\text{NP} \rightarrow \text{DT NN} \mid \text{sisters of NP})$
 $= P(\text{NP} \rightarrow \text{DT NN})$

- **Ancestor free:**

$P(\text{NP} \rightarrow \text{DT NN} \mid \text{its}$
 $\text{ancestors})$
 $= P(\text{NP} \rightarrow \text{DT NN})$



Probability of a parse tree

Domination: we say the non-terminal N^j dominates from between-word indices k to l , symbolized as $N^j_{k,l}$, if $w_{k,l}$ is derived from N^j

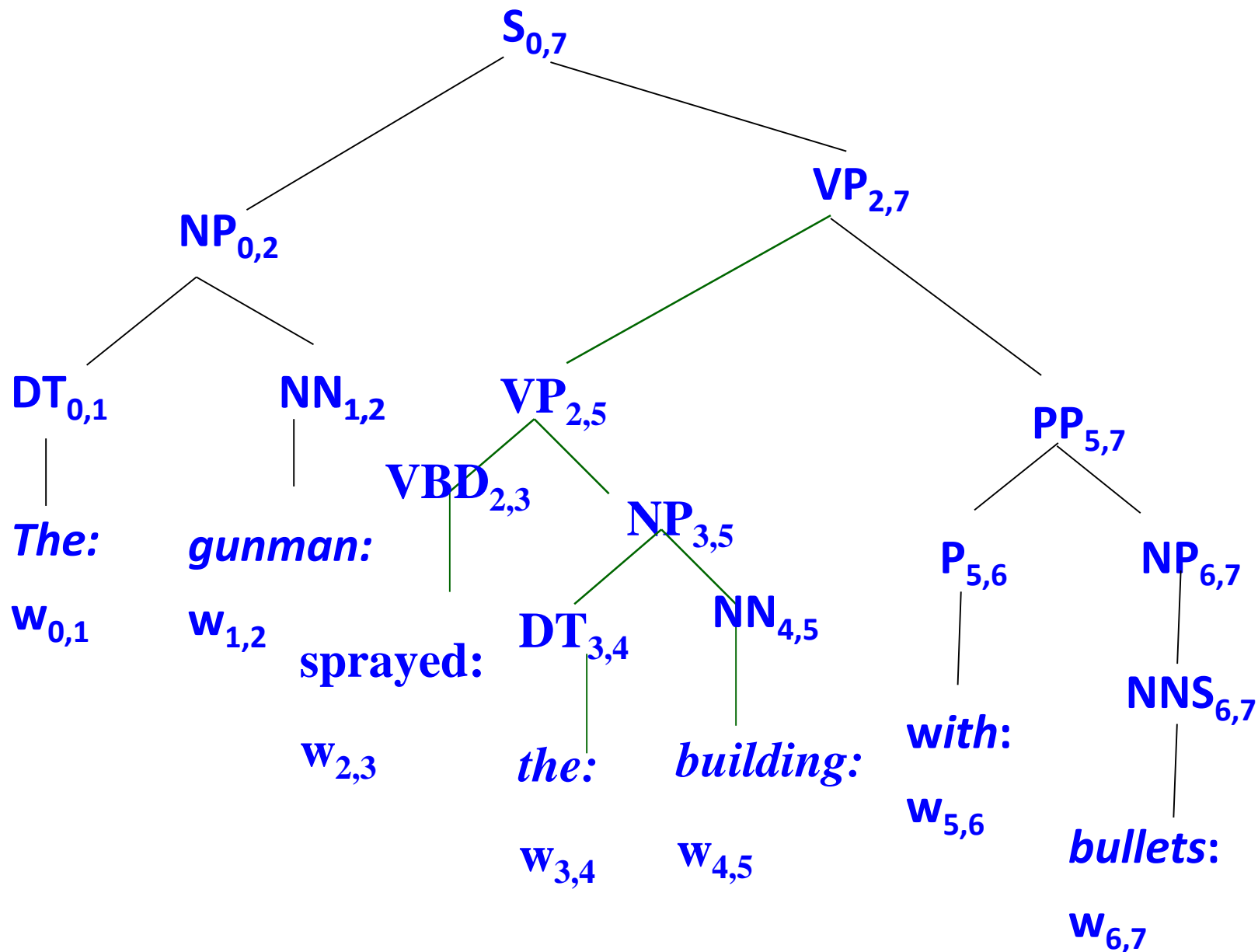
$P(\text{tree} / \text{sentence}) = P(\text{tree} / S_{0,l})$, where $S_{0,l}$ means that the start symbol S dominates the word sequence $w_{0,l}$

$P(t/s)$ approximately equals joint probability of constituent non-terminals dominating the sentence fragments (next slide)

Indexed sentence

₀The ₁ gunman ₂ sprayed ₃ the ₄
₄building ₅ with ₆ bullets ₇ . ₈

Probability of a parse tree



Probability of a parse tree (cont.)

$$P(t|s) = P(t \mid S_{0,7})$$

= P

(NP_{0,2}, DT_{0,1}, “*the*”:w_{0,1}, NN_{1,2}, “*gunman*”:w_{1,2},

VP_{2,7}, VP_{2,5}, VBD_{2,3}, “*sprayed*”:w_{2,3},

NP_{3,5}, DT_{3,4}, “*the*”:w_{3,4}, NN_{4,5}, “*building*”:w_{4,5},

PP_{5,7}, P_{5,6}, “*with*”:w_{5,6}, NP_{6,7}, NNS_{6,7}, “*bullets*”:w_{6,7}

|S_{0,7})

Probability of a parse tree (cont.)

$$\begin{aligned}
 &= P (NP_{0,2} , VP_{2,7} \mid S_{0,7}) * P(DT_{0,1} , NN_{1,2} \mid NP_{0,2}, VP_{2,7}, \\
 &\quad S_{0,7}) * \dots \\
 &= P (NP_{0,2} , VP_{2,7} \mid S_{0,7}) * P(DT_{0,1} , NN_{1,2} \mid NP_{0,2}) * \\
 &\quad \dots
 \end{aligned}$$

(Using Chain Rule, Context Freeness and Ancestor Freeness- $VP_{2,7}$ is $NP_{0,2}$'s sister and S its ancestor)