CS626: Speech, NLP and Web

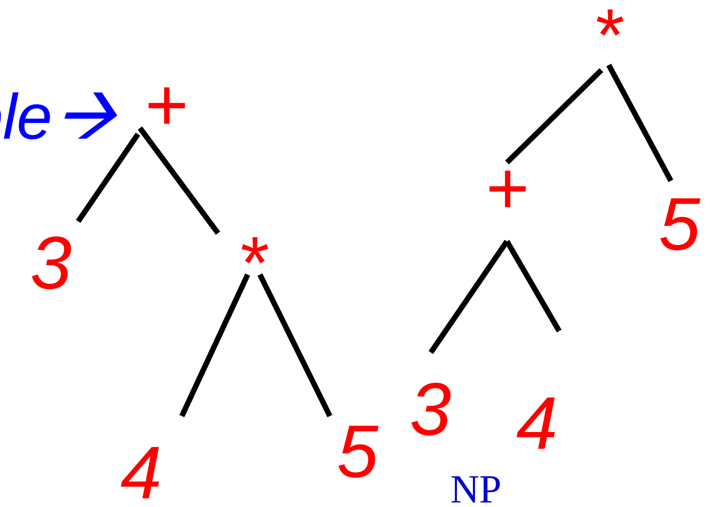
Parsing cntd

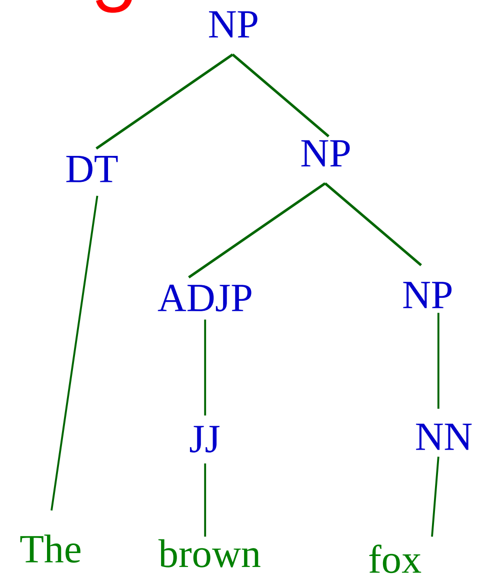
Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

Week 6 of 2nd September, 2024

1-slide recap of week of 19th Aug

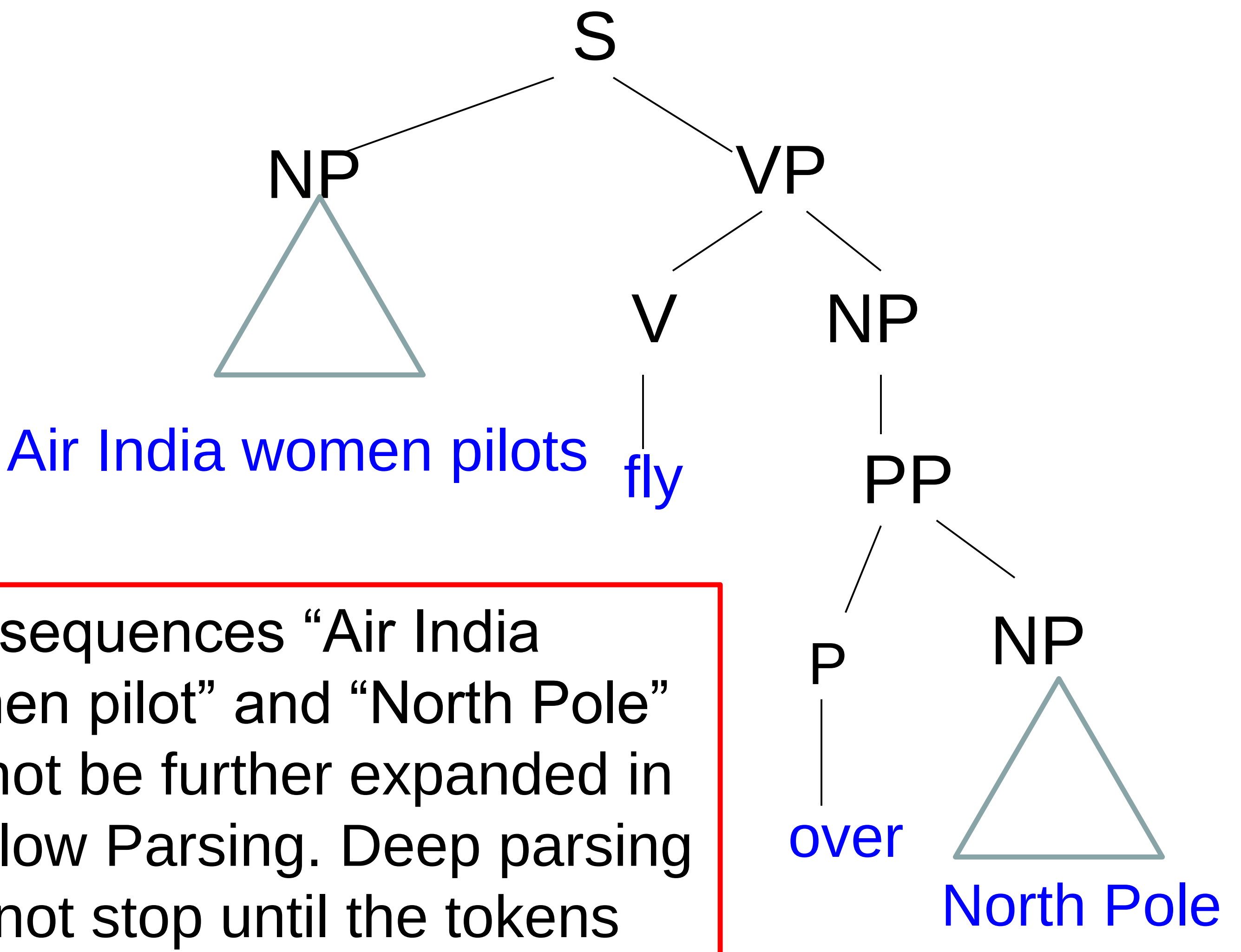
- Evidence of deep structure: *unlockable* → *un+lockable* or *unlock+able*
- **Structural ambiguity**
- Two kinds of parsing: CP and DP
- POS tagging facilitates chunking and parsing: short phrases and deep trees
- Parsing is important: e.g., aspect based SA
- Generative grammar, CFG
  - S→NP VP; NP→ NP PP
- Algorithmics of parsing: top down (TD), bottom up (BU), TDBU, CYK
- BI notation- vimp for NLP





Observations on Shallow and Deep Parsing

"Air India women pilots fly over North Pole"



The sequences "Air India women pilot" and "North Pole" will not be further expanded in Shallow Parsing. Deep parsing Will not stop until the tokens (terminals are reached)

BIO for Chunk (non recursive phrase)

Air/B India/I women/I pilots/I fly/O over/O North/B Pole/I./O

Target phrases:

Air India women pilots

North Pole

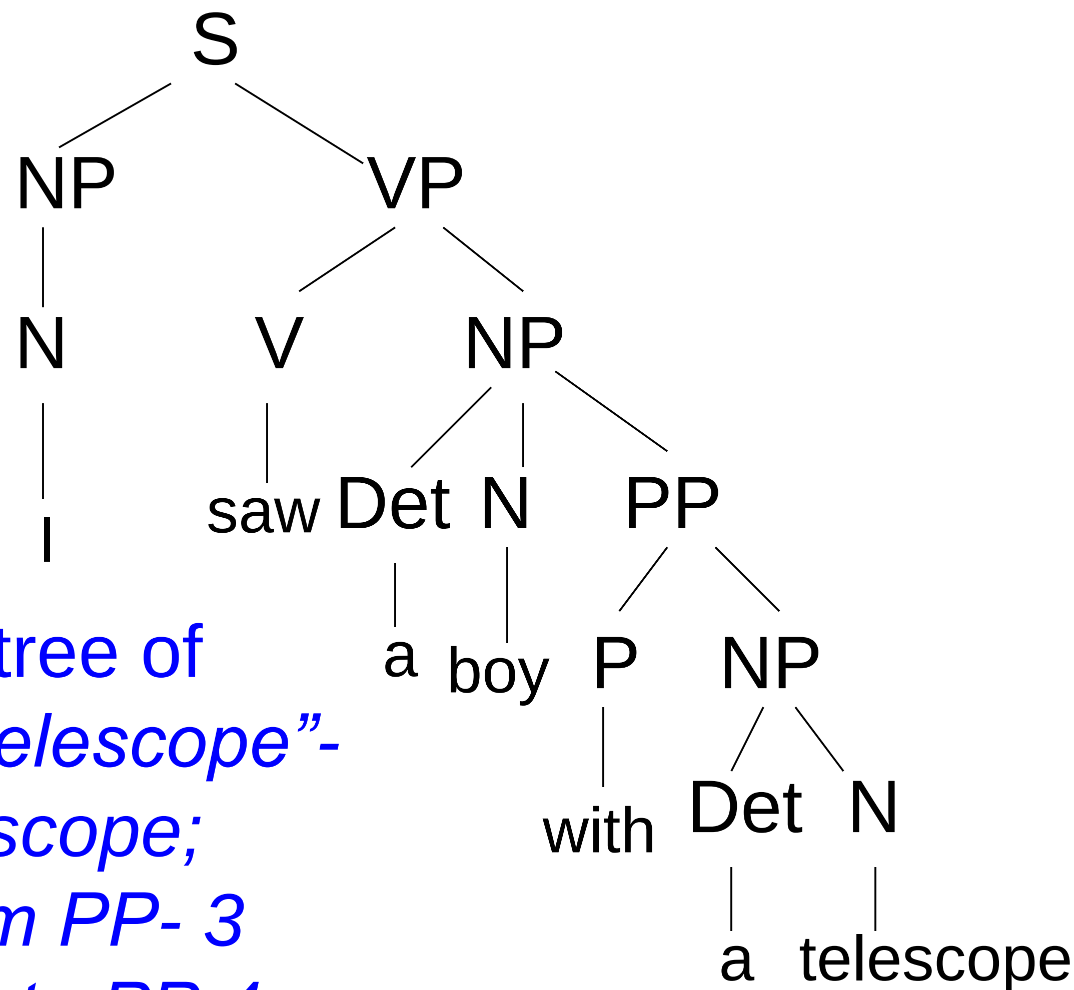
Chunk Extraction: ***local*** information adequate

- Use *argmax* computation

- Produce labels at positions

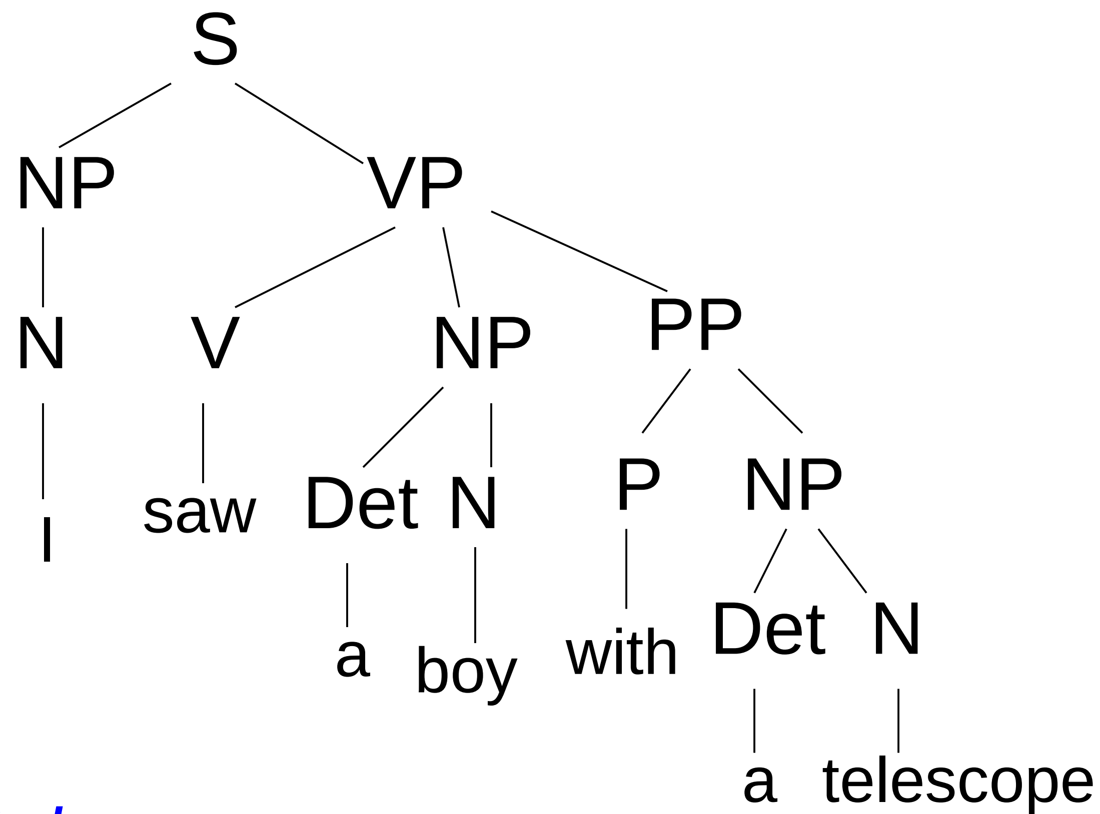
- Use features ON and AROUND positions

What is the proof that there is underlying structure? **Structural Ambiguity**



Constituency parse tree of *"I saw a boy with a telescope"- the boy has the telescope; distance of 'boy' from PP- 3 arcs, that from 'saw' to PP 4 arcs*

Constituency Parse Tree -2



I saw a boy with a telescope-

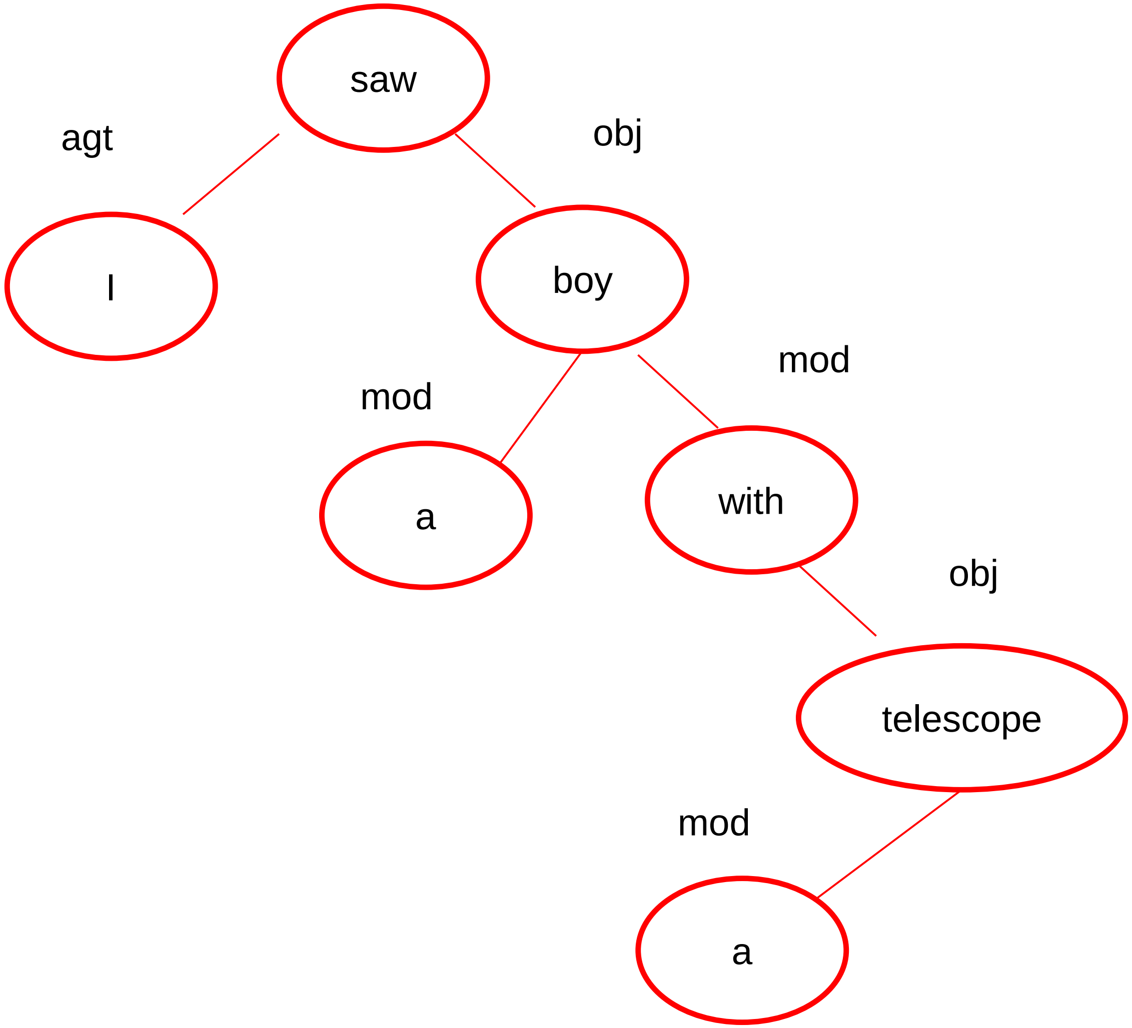
I have the telescope; *distance of 'boy' from PP- 4 arcs, that from 'saw' to PP 3 arcs*

SANNIDHI Principle (proximity)

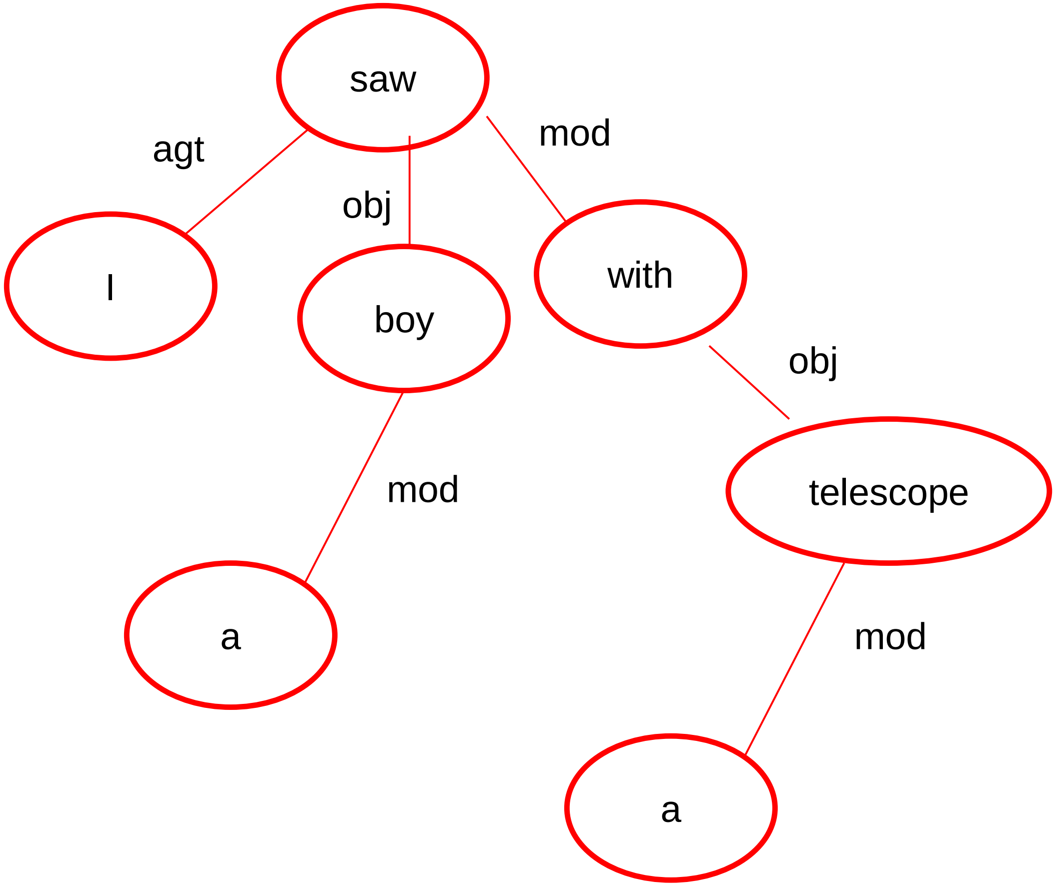
Foundation of Constituency Tree

- Parent child relation means parent is constituted of child(ren)
- If there are multiple children, i.e., multiple constituents, one of them is the head and others are modifiers
- Thus given VP→ V NP, VP is constituted of V and NP
- V is the head and NP the modifier for the VP

Dependency Parse Tree - 1



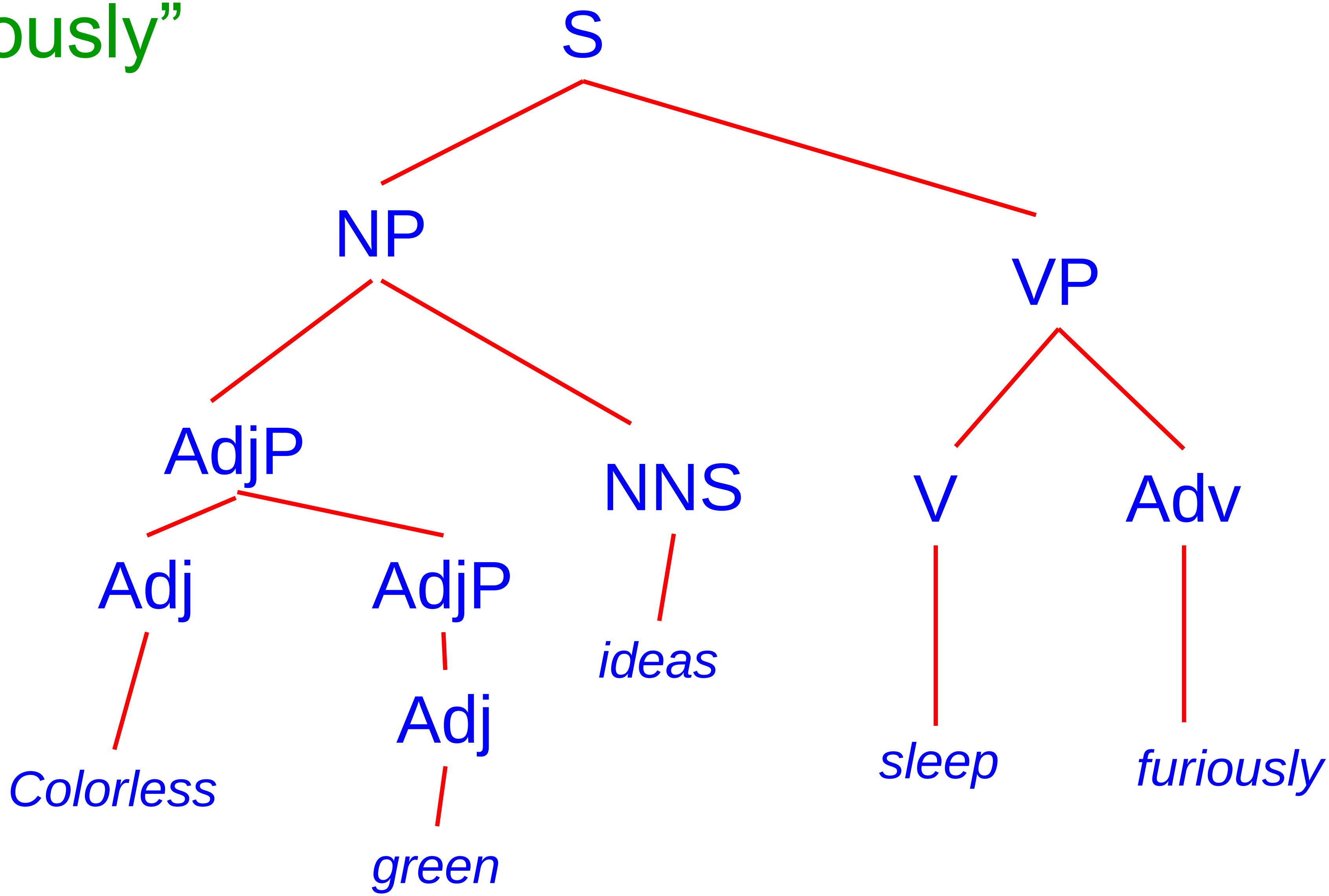
Dependency Parse Tree - 2



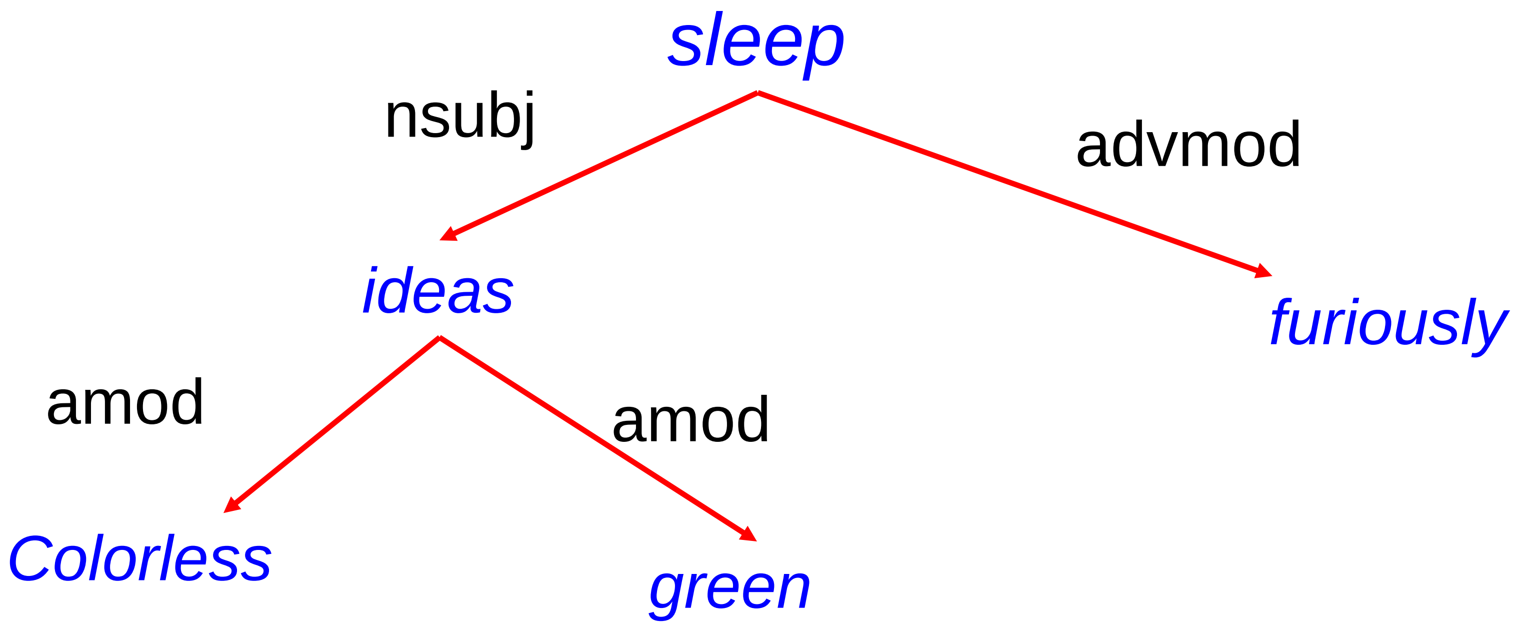
Foundation of Dependency Tree (DT)

- Parent child relation is head-modifier
- Labelled DT: the head-modifier relation is further specified with the type, e.g., *nsubj* meaning nominal subject, *dobj* meaning direct object and *iobj* meaning indirect object of the main verb (*mv*).
- E,g, $\textit{Jack}_{nsubj}$ $\textit{gave}_{mv}$ *a* $\textit{book}_{dobj}$ *to* $\textit{Jill}_{iobj}$.

Constituency parse tree of a famous sentence: "Colorless green ideas sleep furiously"



Dependency parse tree of “Colorless…”: Head Modifier Relations



Syntax-Semantics

- Syntax and semantics influence each other
- However, they can be independent too- as in the “*colourless green ideas*…” sentence
- Consistent with neurolinguistics- Broca’s and Wernicke’s areas

Grammar and Parsing Algorithms

A simplified grammar

- $S \rightarrow NP\ VP$
- $NP \rightarrow DT\ N\ |\ N$
- $VP \rightarrow V\ ADV\ |\ V$

- The above captures declarative sentences
- 4 kinds of sentences as per traditional grammar
  - Declarative (Sun rises in the east)
  - Interrogative (Does sun rise in the east?)
  - Imperative (Rise in the east please)
  - Exclamatory (Oh, sun rises in the east!)

Combining top-down and bottom-up strategies

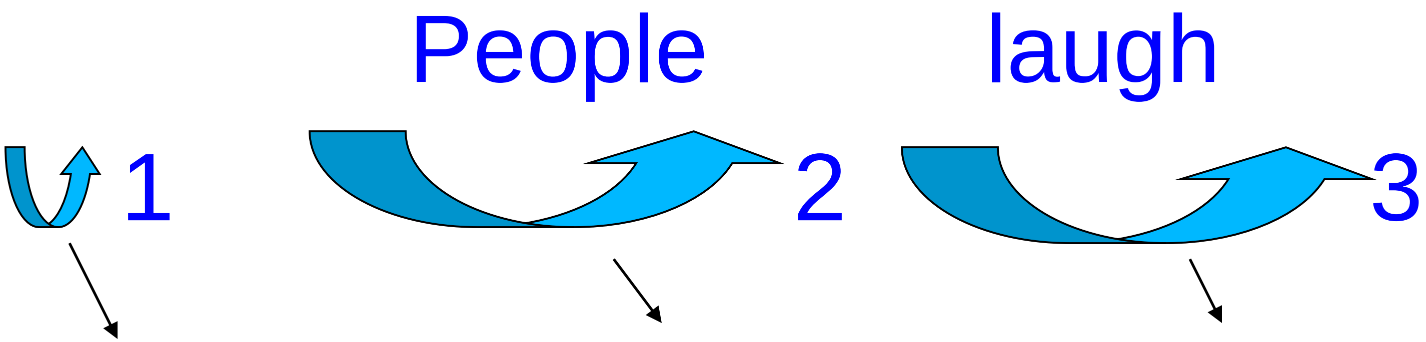
Top-Down Bottom-Up Chart Parsing

- Combines advantages of top-down & bottom-up parsing.
- Does not work in case of left recursion.
  - *e.g.* – “People laugh”
    - People – noun, verb
    - Laugh – noun, verb
  - Grammar – $S \rightarrow NP\ VP$

    $NP \rightarrow DT\ N \mid N$

    $VP \rightarrow V\ ADV \mid V$

Transitive Closure



S →•NP VP

NP →•DT N

NP →•N

NP →N•

S → NP•VP

VP →•V ADV

VP →•V

VP → V •

S → NP VP •

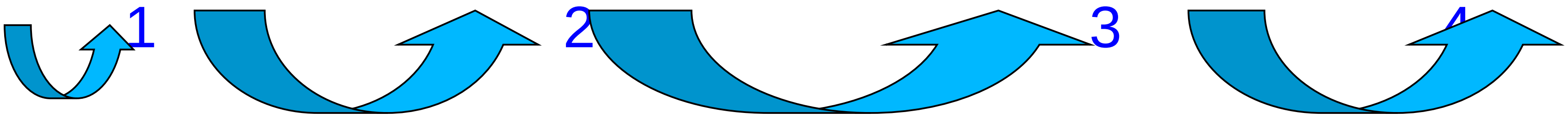
success

Arcs in Parsing

- Each arc represents a chart which records
  - Completed work (left of •)
  - Expected work (right of •)

Example

People laugh loudly



S $\rightarrow \bullet$ NP VP	NP $\rightarrow$ N$\bullet$	VP $\rightarrow$ V$\bullet$	VP $\rightarrow$ V ADV$\bullet$
NP $\rightarrow \bullet$ DT N	S $\rightarrow$ NP$\bullet$VP	VP $\rightarrow$ V$\bullet$ADV	S $\rightarrow$ NP VP$\bullet$
NP $\rightarrow \bullet$ N	VP $\rightarrow$ $\bullet$V ADV	S $\rightarrow$ NP VP$\bullet$	
	VP $\rightarrow$ $\bullet$V		

An important parsing algorithm

CYK Parsing

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 → the
- NN → gunman
- NN → building
- VBD → sprayed
- NNS → 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	--------	---------	--------	---------	---------	---------	

CYK: Control moves to last 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**	**--------**	**---------**	**S**	**---------**	
1	**-------**	**NN**	**--------**	**---------**	**---------**	**---------**	
2	**-------**	**---------**	**VBD**	**---------**	**VP**	**---------**	
3	**-------**	**---------**	**--------**	**DT**	**NP**	**---------**	
4	**--------**	**---------**	**--------**	**---------**	**NN**	**---------**	
5	**--------**	**---------**	**--------**	**---------**	**---------**	**P**	
6	**--------**	**---------**	**--------**	**---------**	**---------**	**---------**	**NP** **NNS**

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**	**PP**
6	**--------**	**---------**	**--------**	**---------**	**---------**	**---------**	**NP** **NNS**

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	---------	NP
4	--------	---------	--------	---------	NN	---------	---------
5	--------	---------	--------	---------	---------	P	PP
6	--------	---------	--------	---------	---------	---------	NP NNS

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	---------	VP
3	-------	---------	--------	DT	NP	---------	NP
4	--------	---------	--------	---------	NN	---------	---------
5	--------	---------	--------	---------	---------	P	PP
6	--------	---------	--------	---------	---------	---------	NP NNS

CYK: filling the last 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	--------	---------	S	---------	
1	-------	NN	--------	---------	---------	---------	---------
2	-------	---------	VBD	---------	VP	---------	VP
3	-------	---------	--------	DT	NP	---------	NP
4	--------	---------	--------	---------	NN	---------	---------
5	--------	---------	--------	---------	---------	P	PP
6	--------	---------	--------	---------	---------	---------	NP NNS

CYK: terminates with S in (0,7)

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

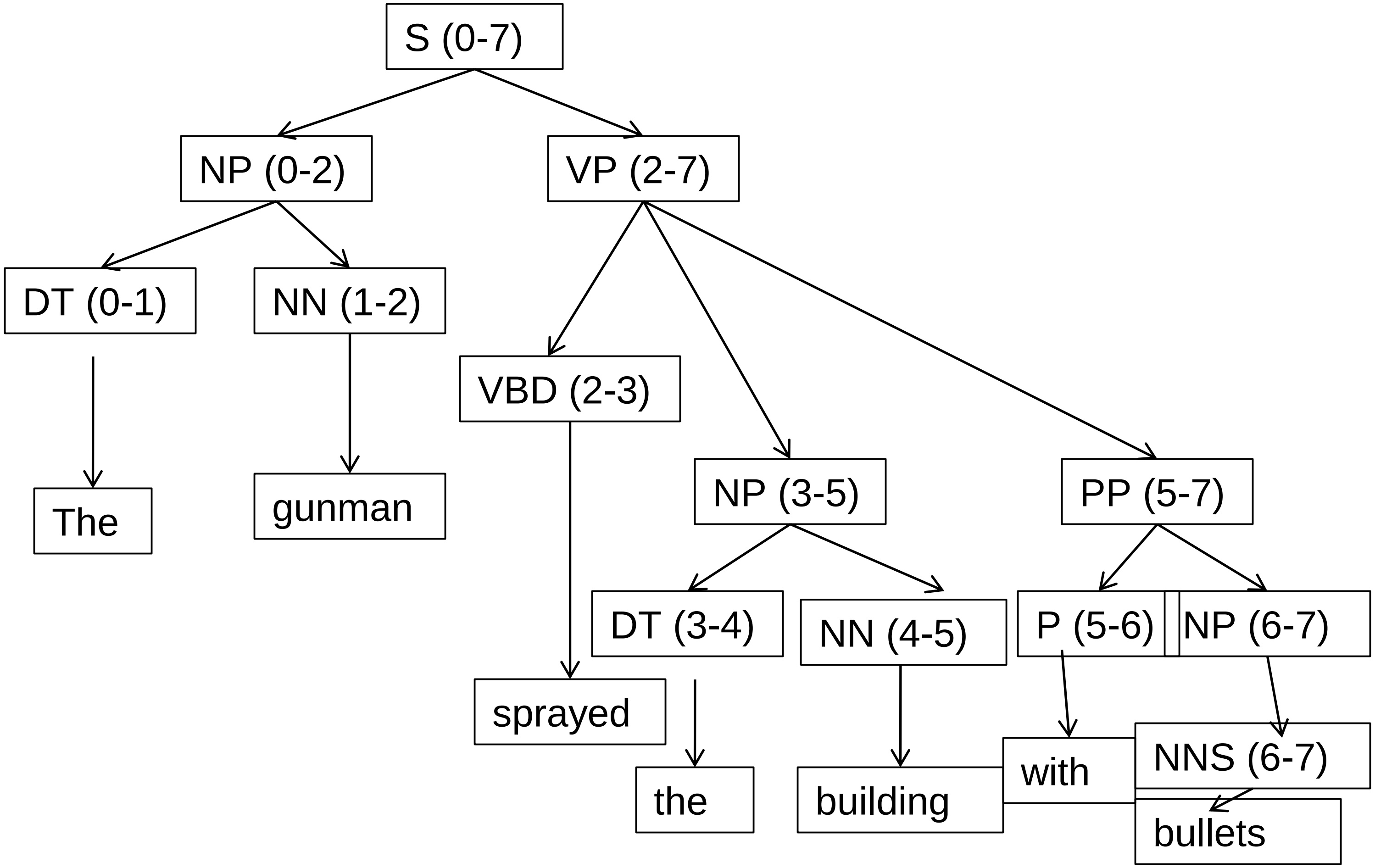
To From	1	2	3	4	5	6	7
0	DT	NP	--------	---------	S	---------	S
1	-------	NN	--------	---------	---------	---------	---------
2	-------	---------	VBD	---------	VP	---------	VP
3	-------	---------	--------	DT	NP	---------	NP
4	--------	---------	--------	---------	NN	---------	---------
5	--------	---------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

CYK: Extracting the Parse Tree

- The parse tree is obtained by keeping back pointers.

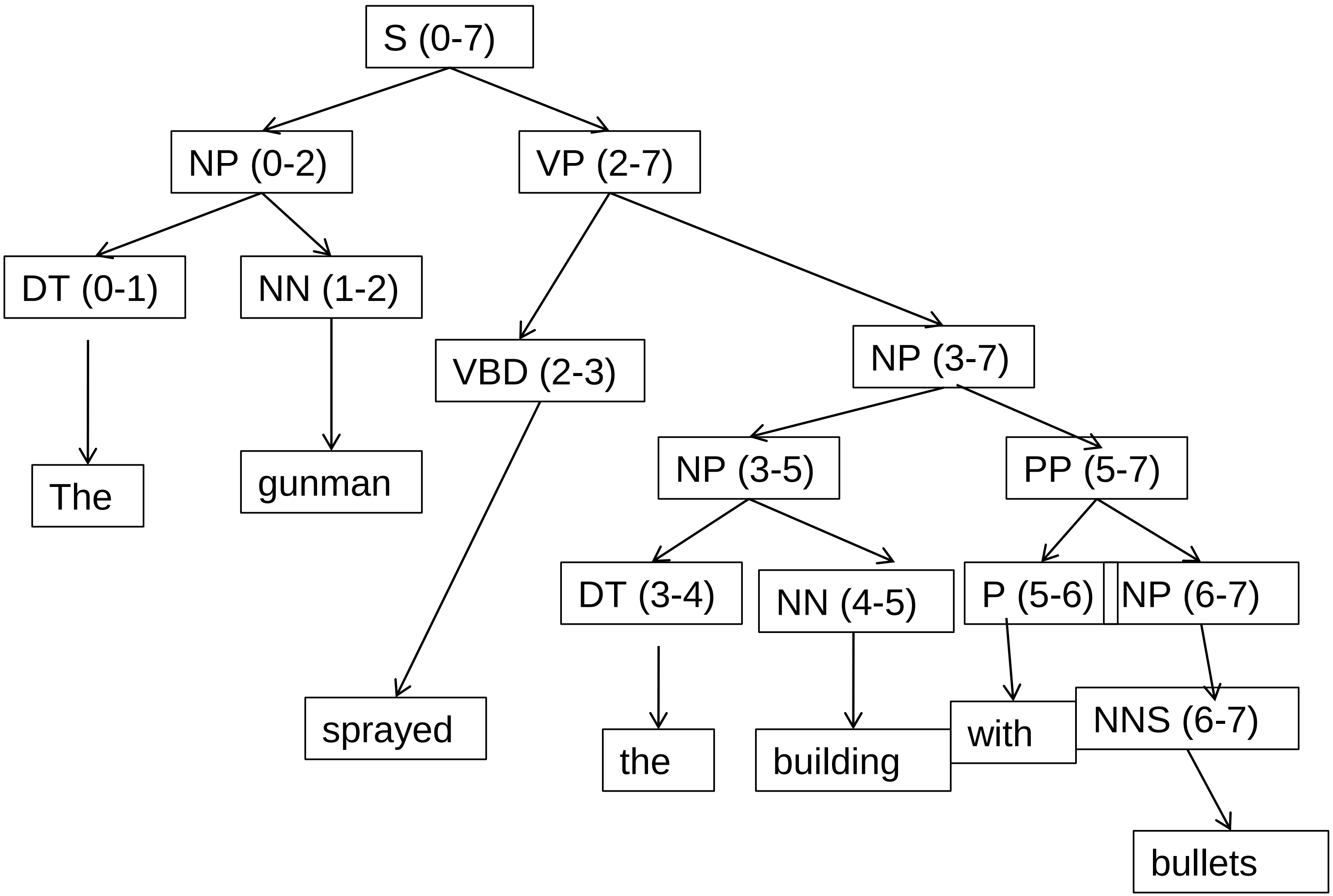
Parse Tree #1

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*.*

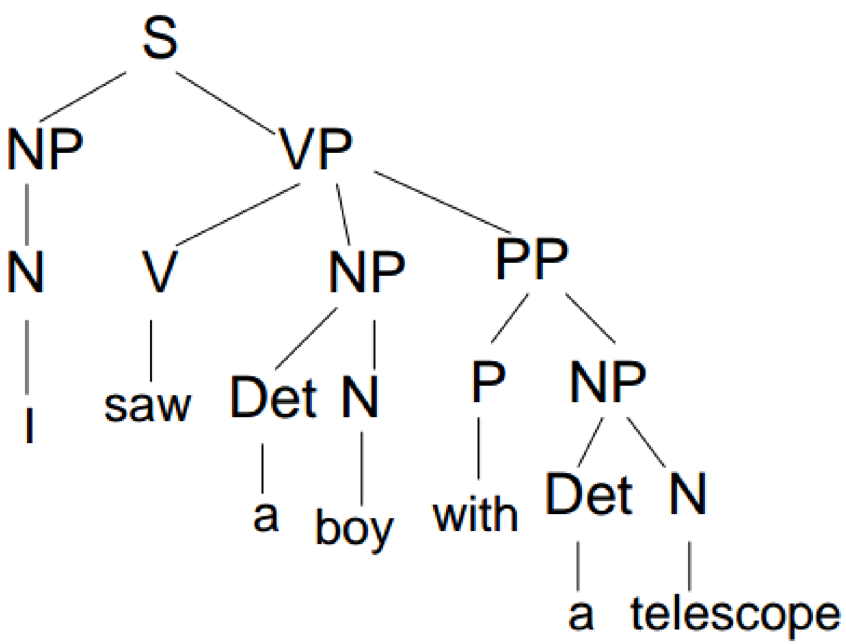


Notion of Domination

- 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

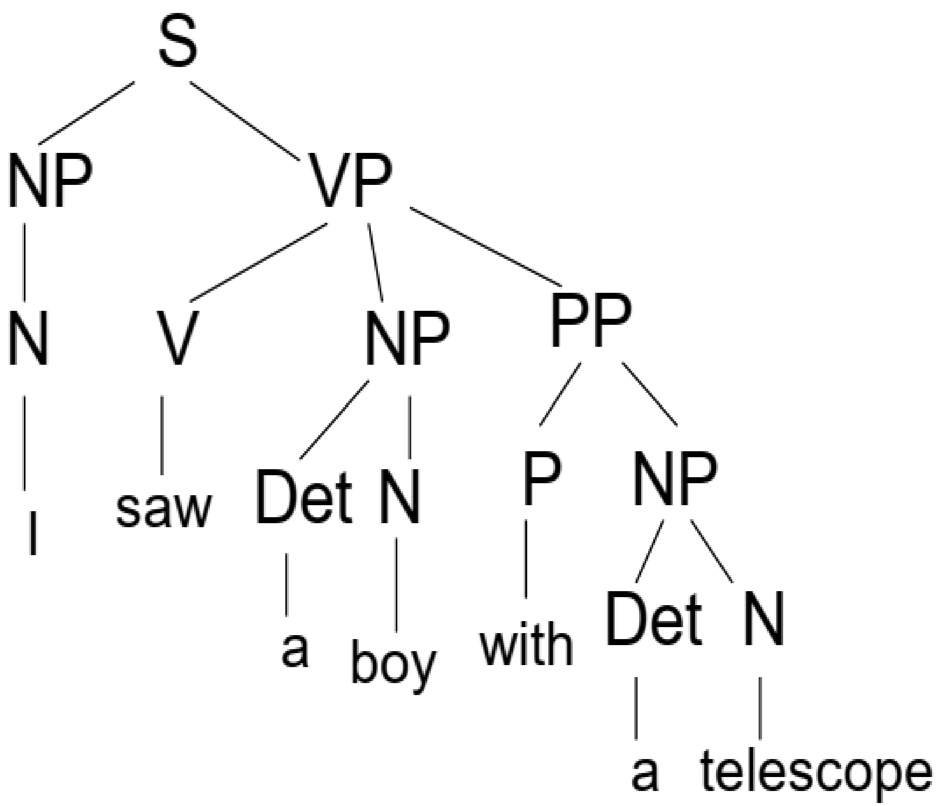
Source sentence → Target parse

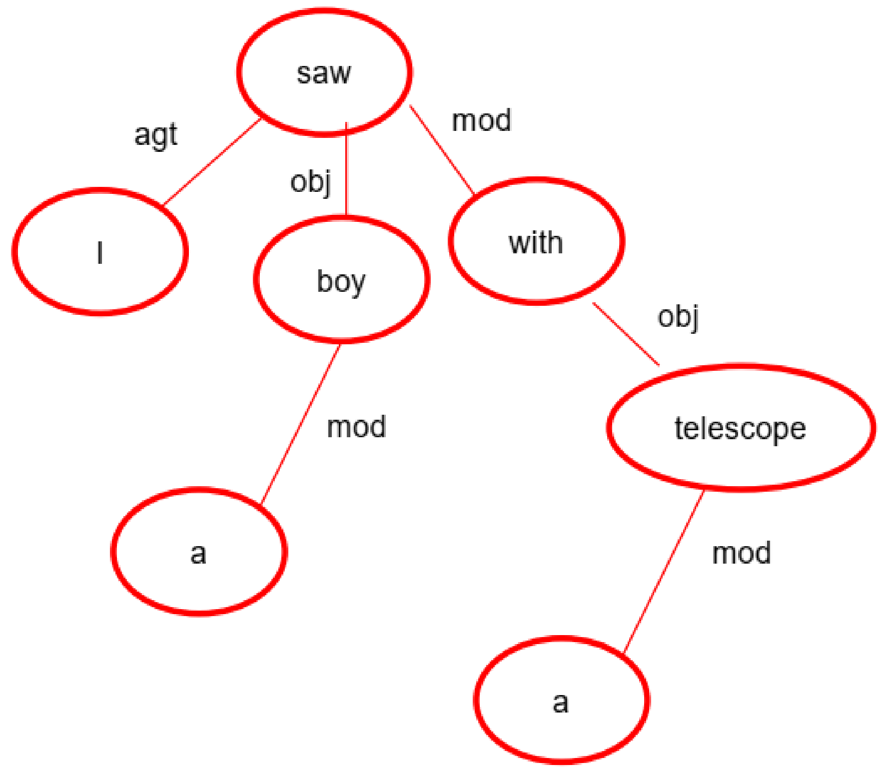
$$T^* = \underset{T}{\operatorname{argmax}}\,[P(T|S)]$$

$$= \underset{T}{\operatorname{argmax}}\,[P(T).P(S|T)]$$

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

“I saw…”: CP and DP #1





Bracketed Structure #1

Parse #1 (meaning: I have the telescope)

[

[I]$_{NP}$

[

[saw]$_{VBD}$

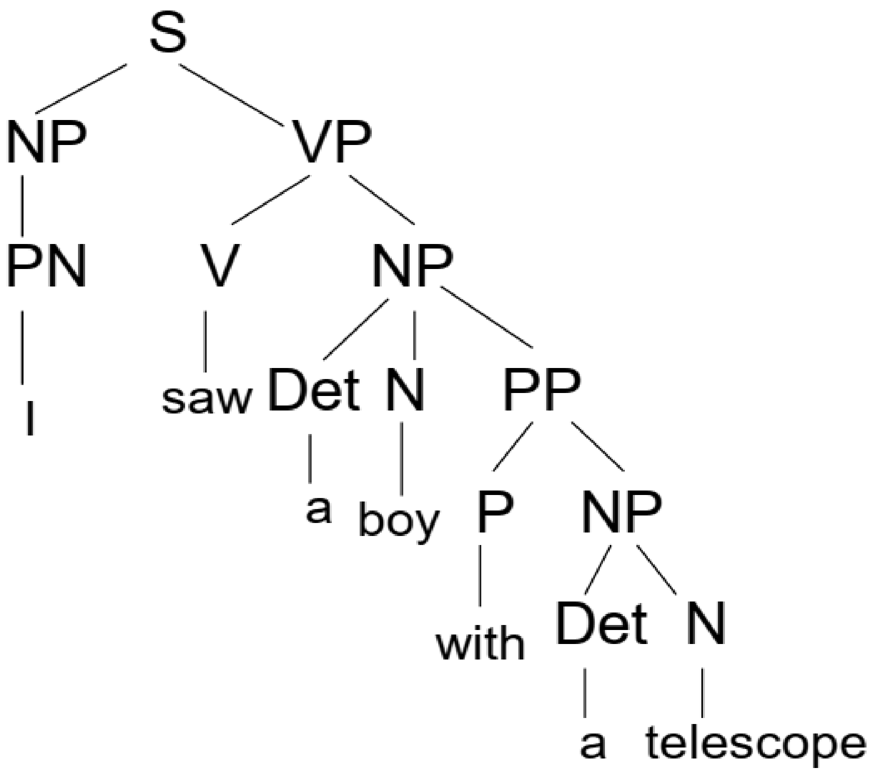
[the boy]$_{NP}$

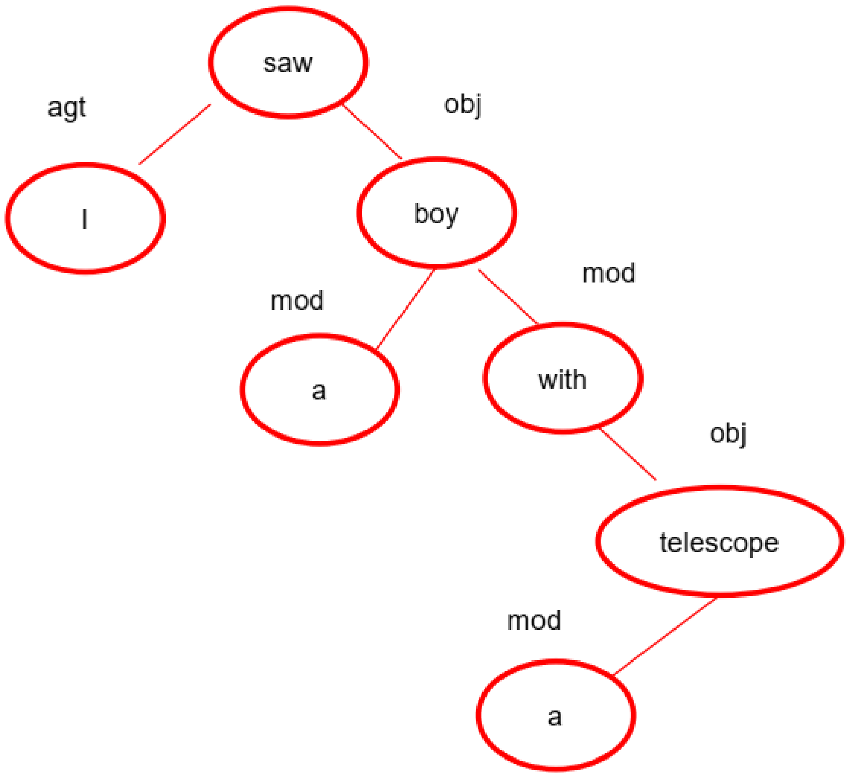
[with [a telescope]$_{NP}$]$_{PP}$

]$_{VP}$

]$_{S}$

“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,….,V

  $\{w_k\}$ = { child, teddy, bear, played…}

- A set of non-terminals $\{N^i\}$, i = 1,…,n

  $\{N^i\}$ = { 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., NP $\rightarrow$ DT NN

- A corresponding set of rule probabilities

Rule Probabilities

- Rule probabilities are such that for for the same non terminal all production rules sum to1.

  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

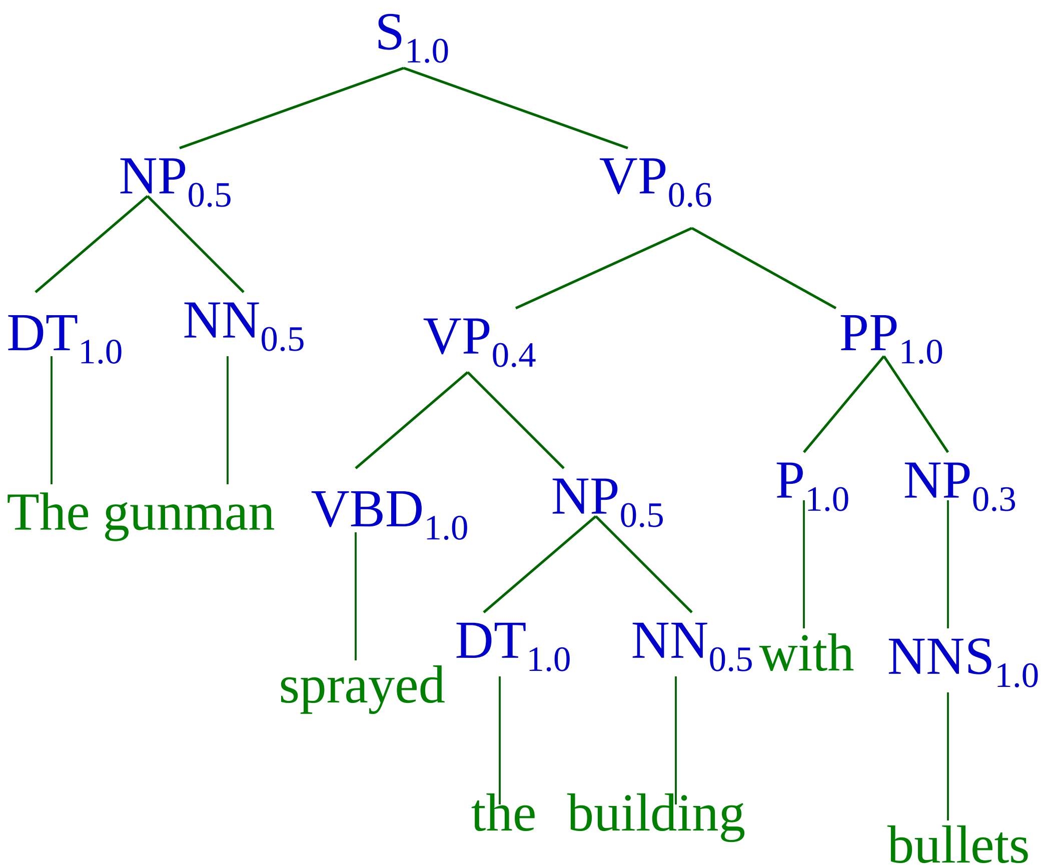
Probabilistic Context Free Grammars

- 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`

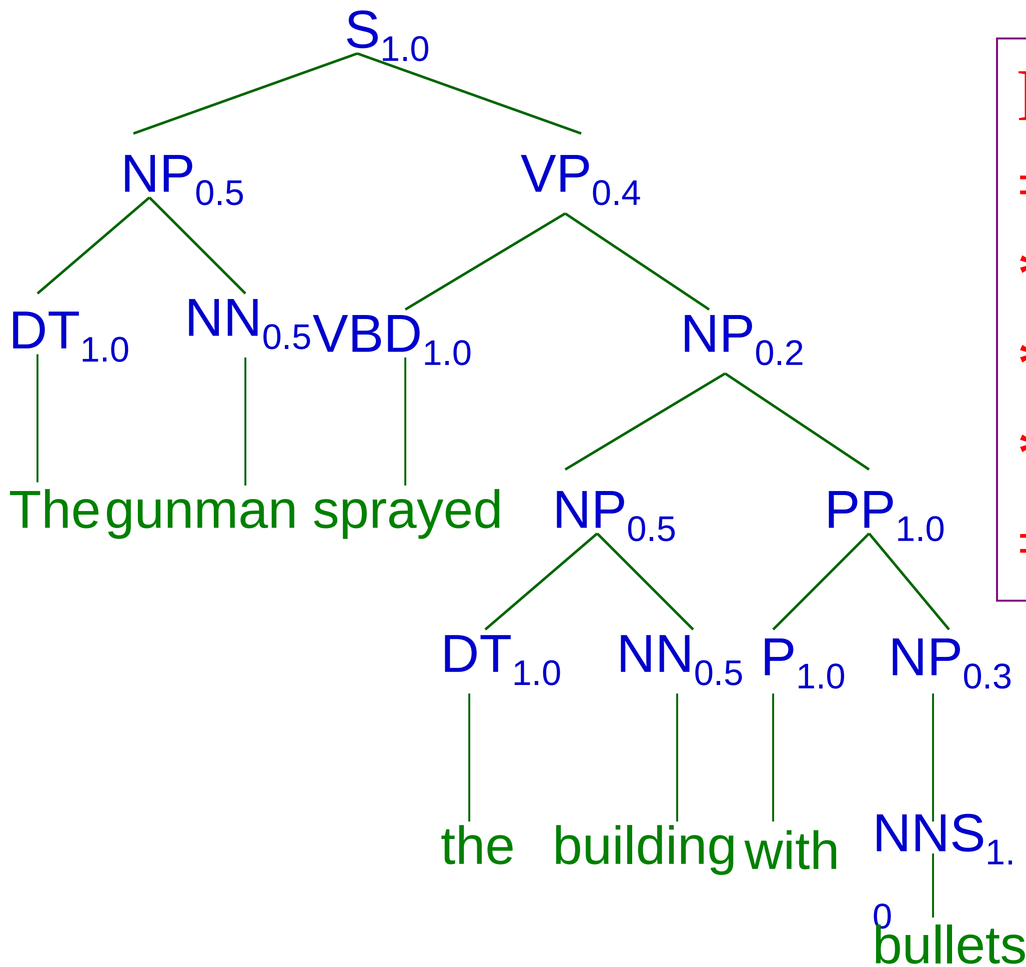
The gunman sprayed the building with bullets.



$P(t_1) = 1.0 * 0.5 * 1.0$
$* 0.5 * 0.6 * 0.4 * 1.0$
$* 0.5 * 1.0 * 0.5 * 1.0$
$* 1.0 * 0.3 * 1.0$
$= 0.00225$

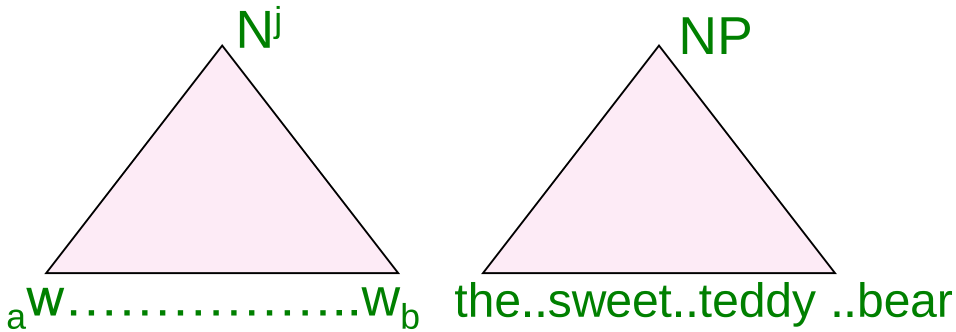
Another Parse t_2

The gunman sprayed the building with bullets.



$P(t_2)$
$= 1.0 * 0.5 * 1.0 * 0.5$
$* 0.4 * 1.0 * 0.2 * 0.5$
$* 1.0 * 0.5 * 1.0 * 1.0$
$* 0.3 * 1.0$
$= 0.0015$

Probability of a sentence (1/2)



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

– w_{ab}– subsequence $_aw....w_b$
– N^j dominates $_aw....w_b$
 or yield(N^j) = $_aw....w_b$

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)

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

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

$$=\Sigma_t P(t).1$$

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:

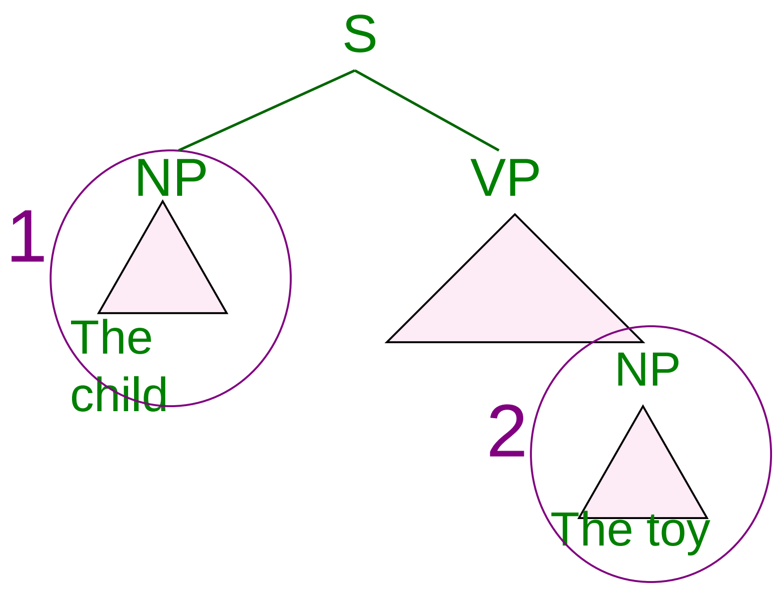
  P(NP → DT NN) is same independent of location in the tree

- Context-free:

  P(NP → DT NN | sisters of NP) = P(NP → DT NN)

- Ancestor free:

  P(NP → DT NN| its ancestors) = P(NP →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 (tree |sentence)= P (tree | $S_{0,l}$), where $S_{0,l}$ means that the start symbol S dominates the word sequence $w_{0,l}$

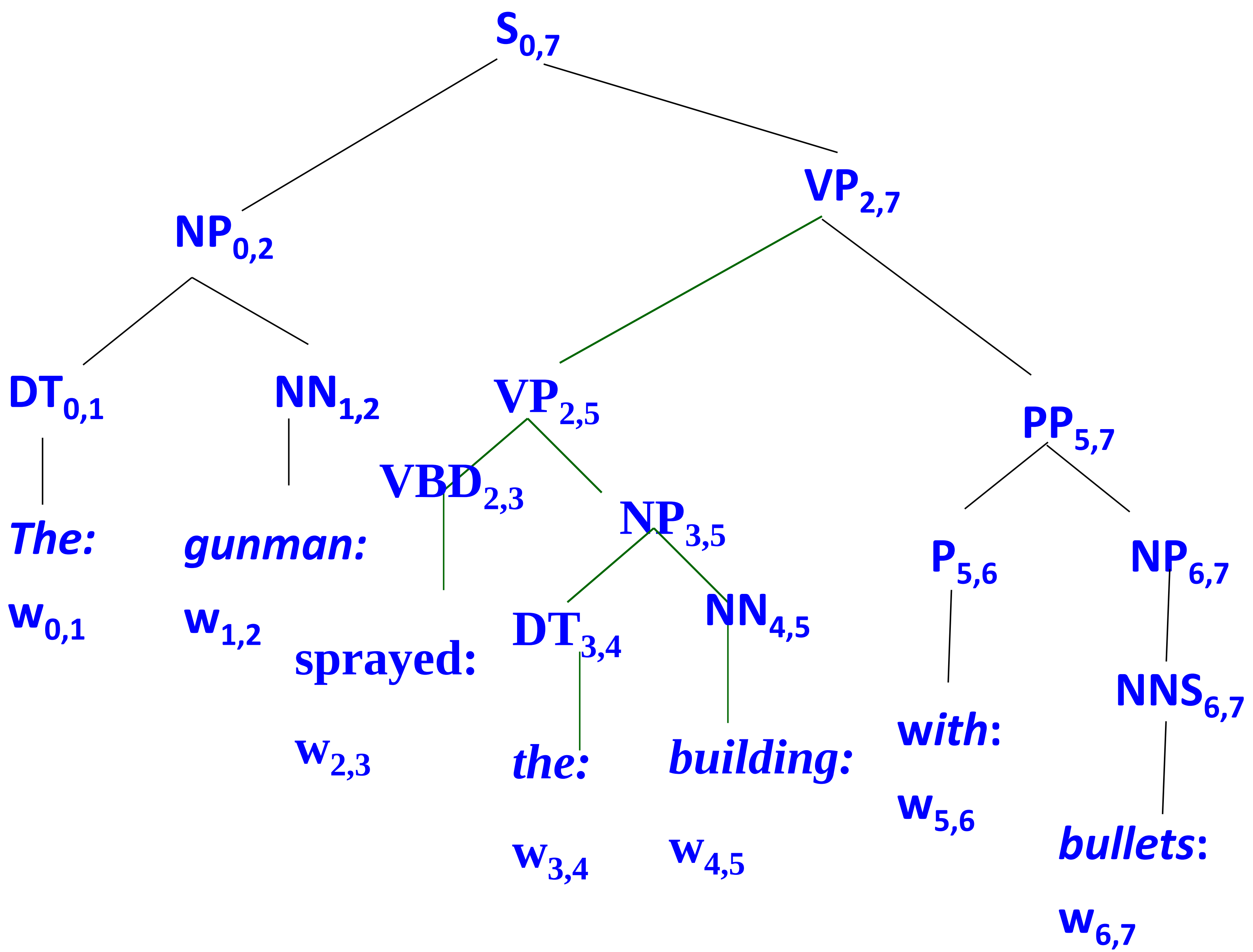
P(t|s) approximately equals joint probability of constituent non-terminals dominating the sentence fragments (next slide)

Indexed sentence

$_0$The $_1$ gunman $_2$ sprayed $_3$ the $_4$

$_4$building $_5$ 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.)

$= P(NP_{0,2}, VP_{2,7} \mid S_{0,7}) * P(DT_{0,1}, NN_{1,2} \mid NP_{0,2}, VP_{2,7}, S_{0,7}) * \ldots.$

$= P(NP_{0,2}, VP_{2,7} \mid S_{0,7}) * P(DT_{0,1}, NN_{1,2} \mid NP_{0,2}) * \ldots.$

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