CS626: Speech, Natural Language Processing and the Web

Discriminative PoS Tagging; start of parsing

Pushpak Bhattacharyya

Computer Science and Engineering Department

IIT Bombay

Week 5 of 22nd August, 2022

Discriminative Labelling

Motivation

- HMM based POS tagging cannot handle "free word order" and "agglutination" well
- If *adjective after noun* is equally likely as *adjective before noun*, the transition probability is no better than uniform probability which has high entropy and is uninformative.
- When the words are long strings of many morphemes, POS tagging w/o morph features is highly inaccuarte.

Variability in word order: problem for generative model

Both SOV and SVO!

Agglutination: problem for generative model

- *istahtaisinkohan "I wonder if I should sit down for a while"*
  - *ist: "sit", verb stem*
  - *ahta: verb derivation morpheme, "to do something for a while"*
  - *isi: conditional affix*
  - *n: first-person singular suffix*
  - *ko: question particle*
  - *han: a particle for things like reminder (with declaratives) or "softening" (with questions and imperatives)*

Agglutination in Manipuri

- Words in Manipuri can consists of ten or more morphemes
- *pusinhanjaramgadabanidako* ("I wish (I) myself would have caused to bring in the article")
- *pu-sin-han-ja-ram-ga-da-ba-ni-da-ko*

Modelling in Discriminative POS Tagging

- *T** is the best possible tag sequence
- Summation dropped, because given *W* and feature engineering, *F* is unique; also *P(F|T)=1*
- The final independence assumption is that the tag at any position *i* depends only on the feature vector at that position

$$T^* = \arg\max_T P(T \mid W)$$

$$P(T \mid W) = \sum_F P(T, F \mid W) = P(T, F \mid W)$$

$$= P(F \mid W).P(T \mid F, W)$$

$$= 1.P(T \mid F) = P(T \mid F)$$

$$P(T \mid F) = \prod_{i=0}^{n+1} [P(t_i \mid F_i)]$$

Feature Engineering

- Running example: ***^ brown foxes jumped over the fence .***
- *A. Word-based features*

  f_{21} – dictionary index of the current word ('foxes'): integer

  f_{22} – -do- of the previous word ('brown'): integer

  f_{23} – -do- of the next word ('jumped'): integer

B. Part of Speech (POS) tag-based feature

  f_{24} – index of POS of previous word (here JJ): integer

Feature engineering cntd.

- ***^ brown foxes jumped over the fence .***

- *C. Morphology-based features*
  - f_{25}– does the current word (‘foxes’) have a noun suffix, like ‘s’, ‘es’, ‘ies’, etc.: 1/0- here the value is
  - f_{26}– does the current word (‘foxes’) have a verbal suffix, like ‘d’, ‘ed’, ‘t’, etc.: 1/0- 0
  - f_{27} and f_{28} for ‘brown’ like for ‘foxes
  - f_{29} and $f_{2,10}$ for ‘jumped’ like for ‘foxes; here $f_{2,10}$ is 1 (jumped has ‘ed’ as suffix)

An Aside: word vectors

- These features are opaquely represented in word vectors created from huge corpora

- Word vectors are vectors of numbers representing words

- It is not possible to tell which component in the word vector does what

Modelling Equations

W: ^ $w_0\, w_1\, w_2 \ldots w_{n-2}\, w_{n-1}\, w_n$. ***T***: ^ $t_0\, t_1\, t_2 \ldots t_{n-2}\, t_{n-1}\, t_n$.

$$P(T) = \prod_{i=0}^{n+1} [P(t_i \mid F_i)]$$

$$P(t_i = t \mid F_i) = \frac{e^{\sum_{j=1.k} \lambda_j f_{ij}}}{\sum_{t' \in S} e^{\sum_{j=1.k} \lambda_j f_{ij}(t')}}$$

Maximum Entropy Markov Model (MEMM)

S: set of tags.

The sequence probability of a tag sequence *T* is the product of $P(t_i|F_i)$, *i* varying over the positions.

Beam Search Based Decoding

- ***^ The brown foxes jumped .***
- Let us assume the following tags for the purpose of the discussion:
  - D- determiner like ‘the’
  - A- adjective like ‘brown’
  - N- noun like ‘foxes’, ‘fence’
  - V- verb like ‘jumped’
- Let the decoder start at the state ‘^’ which denotes start of the sentence.

Step-1

- **^ *The brown foxes jumped .***
- The word '*the*' is encountered. First there are 4 next states possible corresponding to 4 tags, giving rise to 4 possible paths:
- ^ *D* -P_1
- ^ *A* -P_2
- ^ *N* -P_3
- ^ *V* -P_4

Commit to Beam Width

- Beam width is an integer which denotes how many of the possibilities should be kept *open*.
- Let the beam width be 2.
  - **This means that out of all the paths obtained so far we retain only the top 2 in terms of their probability scores.**
- We will assume that the actual linguistically viable sub-sequence appears amongst the top two choices.
  - ‘The’ is a determiner and we get the two highest probability paths for “^ The” as P_1 and P_3.

Step-2

- **^ *The* *brown* *foxes jumped .***
- ‘*brown*’ is the next word. P_1 and P_3 are extended as
- ^ *D D* -P_{11}
- ^ *D A* -P_{12}
- ^ *D N* -P_{13}
- ^ *D V* -P_{14}
- ^ *N D* -P_{31}
- ^ *N A* -P_{32}
- ^ *N N* -P_{33}
- ^ *N V* -P_{34}

Retain two paths

- Keep two possibilities corresponding to correct/almost-correct sub-sequences. '*brown*' is an adjective, but can be noun too (e.g., "*the brown of his eyes*").

^ *D A* -P_{12}

^ *D N* -P_{13}

Step-3

- ***^ The brown foxes jumped .***
- Can be both noun and verb (verb: "*he was foxed by their guile*").
- From P_{12} and P_{13}, we will get 8 paths, but retain only two, as per the beam width.
- We assume only the paths coming from P_{12} survive with 'A' and 'N' extending the paths:

^ *D A A*	-P_{122} (this is a wrong path!)
^ *D A N*	-P_{123}

Step-4

- ***^ The brown foxes jumped .***

- Can be both a past participial adjective (“*the halted train*”) and a verb.

- Retaining only two top probability paths we get

  $\hat{}$ *D A N A* $\quad\quad$ $-P_{1232}$

  $\hat{}$ *D A N V* $\quad\quad$ $-P_{1234}$

Step-5

- ***^ The brown foxes jumped .***

- Can be both a past participial adjective ("*the halted train*") and a verb.

- Retaining only two top probability paths we get

  ^ *D A N A* -P_{1232}

  ^ *D A N V* -P_{1234}

Step-6: Final Step

- **^ *The brown foxes jumped .***

- On encountering dot, the beam search stops.
- We assume we get the correct path probabilistically in the beam (width 2)

  ^ *D A N V.*

How to fix the beam width (1/2)

- English POS tagging with Penn POS tag set: approximately 40 tags
- Fine categories like NNS for plural NNP for proper noun, VAUX for auxiliary verb, VBD for past tense verb and so on.
- A word can have on an average at most 3 POSs recorded in the dictionary.

How to fix the beam width (2/2)

- Allow for 4 finer category POSs under each category and with support from a lexicon that records the broad category POSs,

- A practical beam width for POS tagging for English using Penn tagset could be 12 (=3 X 4). (think and justify)

Start of Parsing

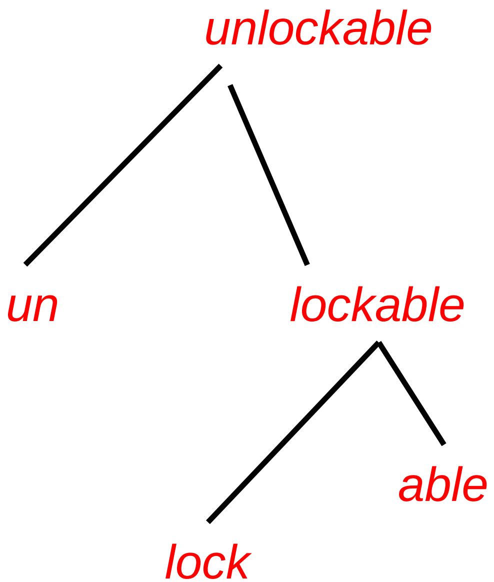
Strong evidence of existence of structure is *structural ambiguity*

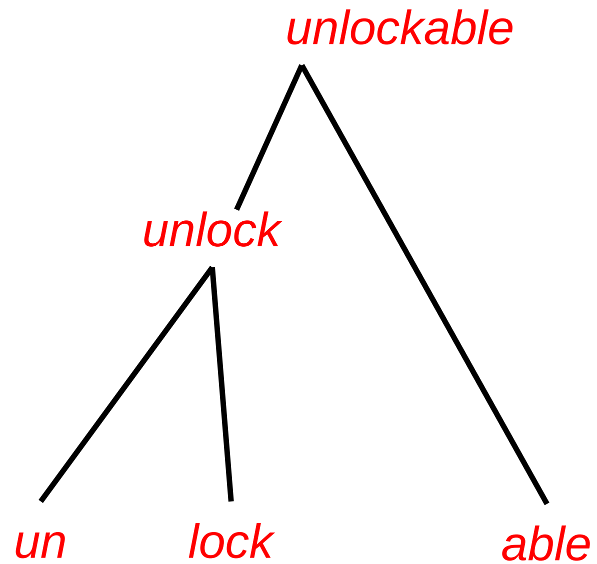
- Parsing of “Unlockable”
  - *Un + lockable*

    - *lock + able*
    - Meaning: something that cannot be locked (*this gate is unlockable- open and cannot be locked*)
  - *Unlock + able*

    - *un + lock*
    - Meaning: something that can be unlocked (*this gate is unlockable- shut with a lock and can be unlocked*)

Tree Structure: Morphotactics of "Unlockable": two structures

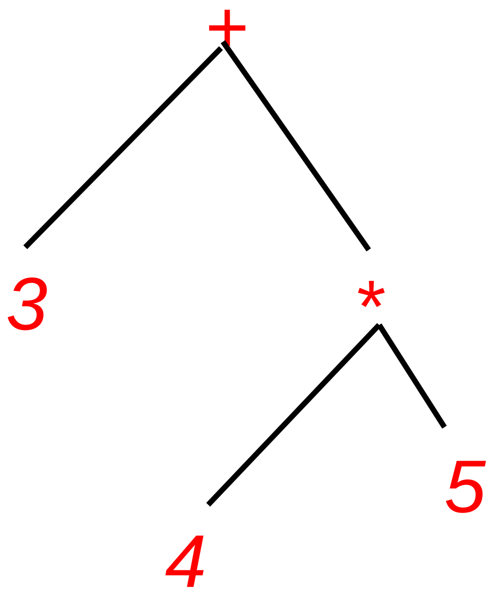




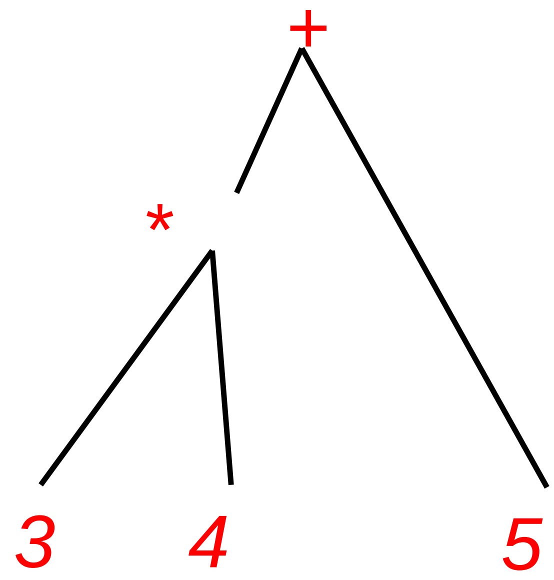
something that cannot be locked ("*this gate is unlockable*"- open and cannot be locked)

Something that can be unlocked ("*this gate is unlockable*"- shut with a lock, but can be unlocked)

Trees frequently used: Arithmetic Expressions



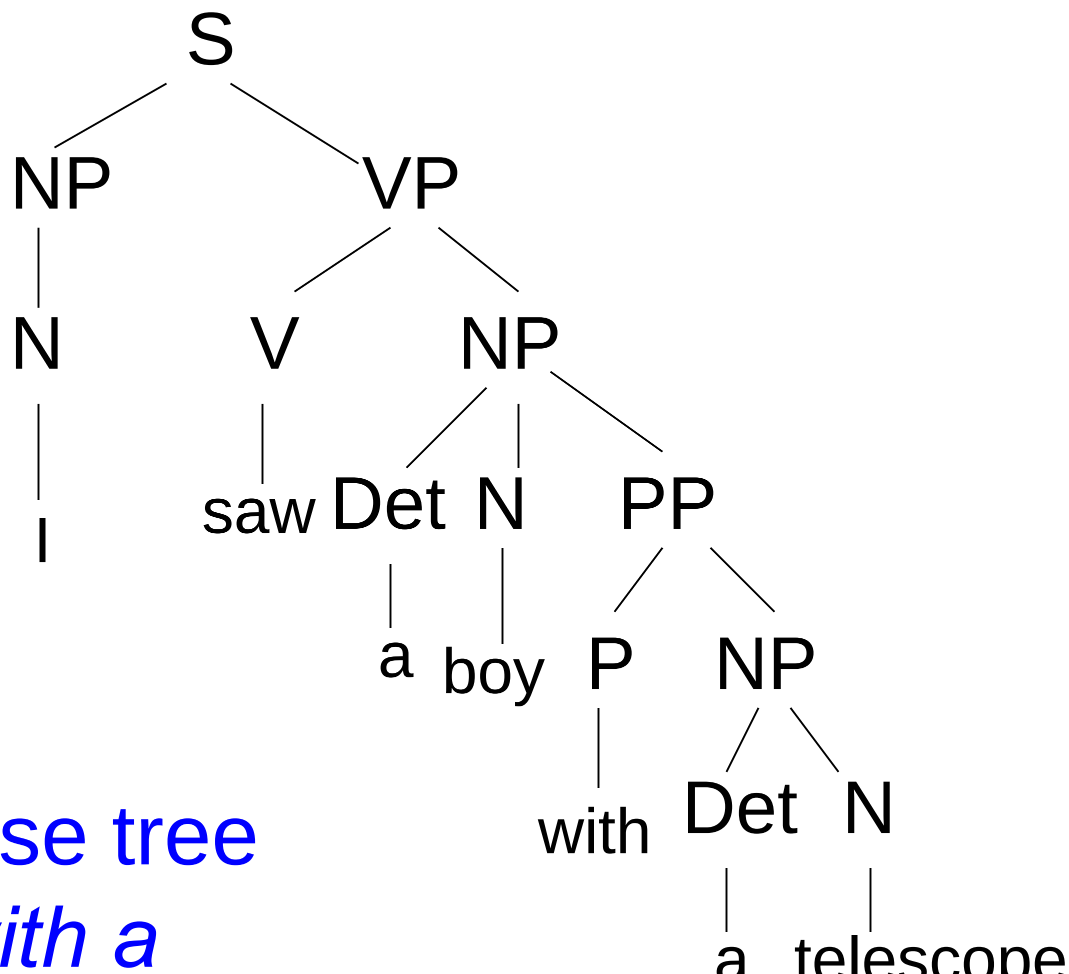
3+4*5=23



3*4+5=17

Trees are devices to represent constituents and their interactions to form bigger constituents

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

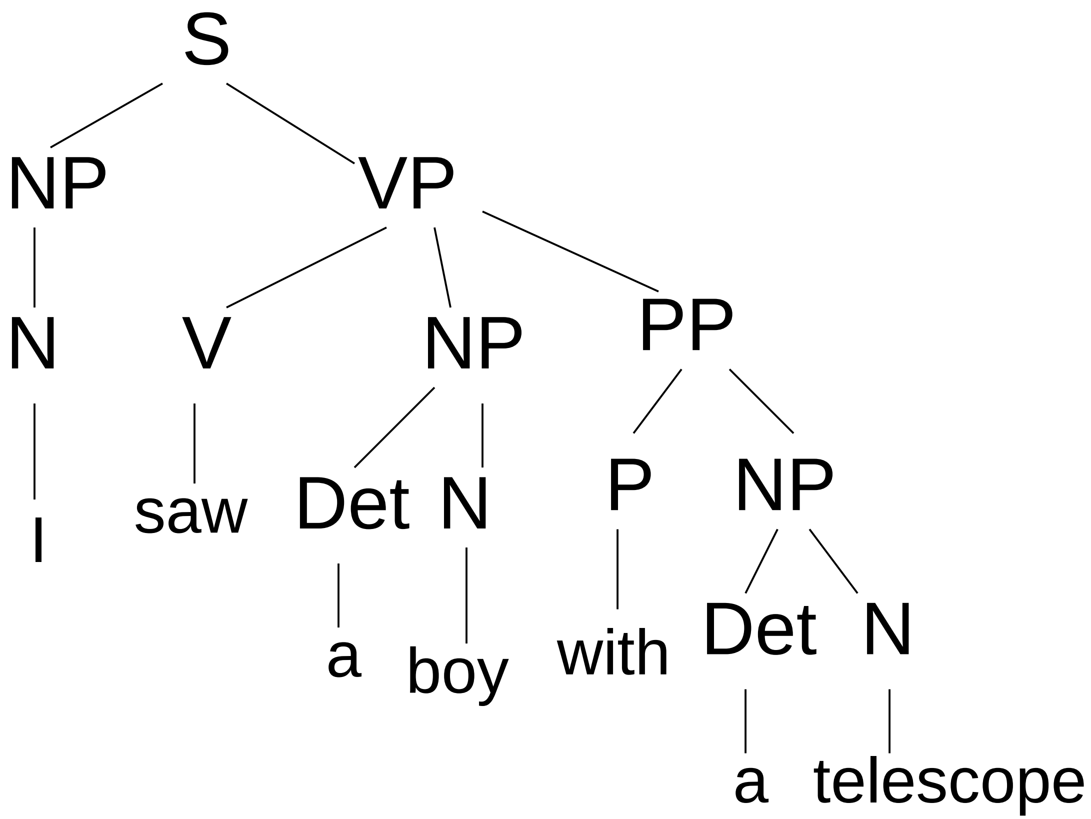


Constituency parse tree of *"I saw a boy with a telescope"*

Foundation of Constituency Tree

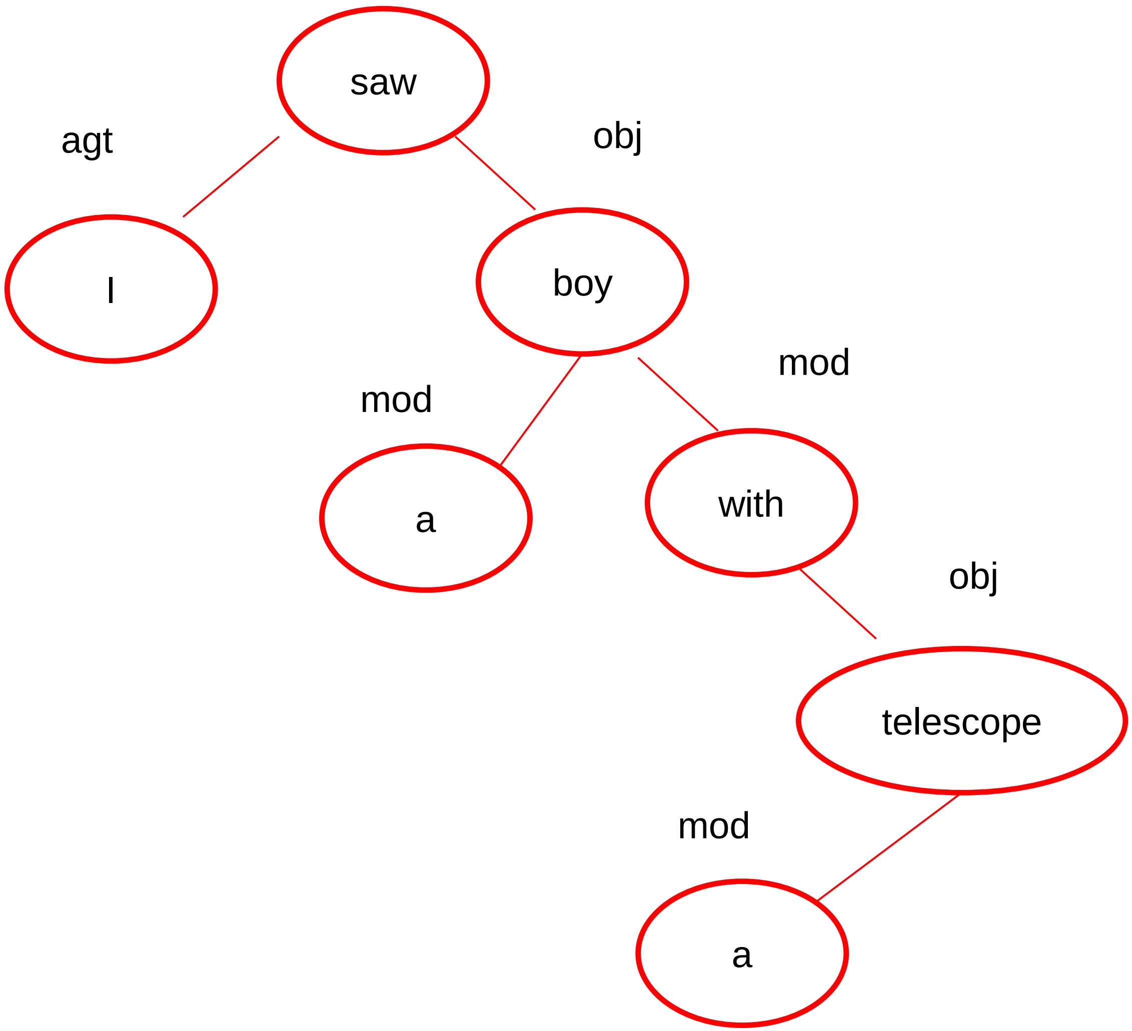
- Parent child relation means parent is constituted of child(ren)
- If there are multiple children, i.e., multiple constituents, one of them is the head and others are modifiers
- Thus given VP→ V NP, VP is constituted of V and NP
- V is the head and NP the modifier for the VP

Constituency Parse Tree -2

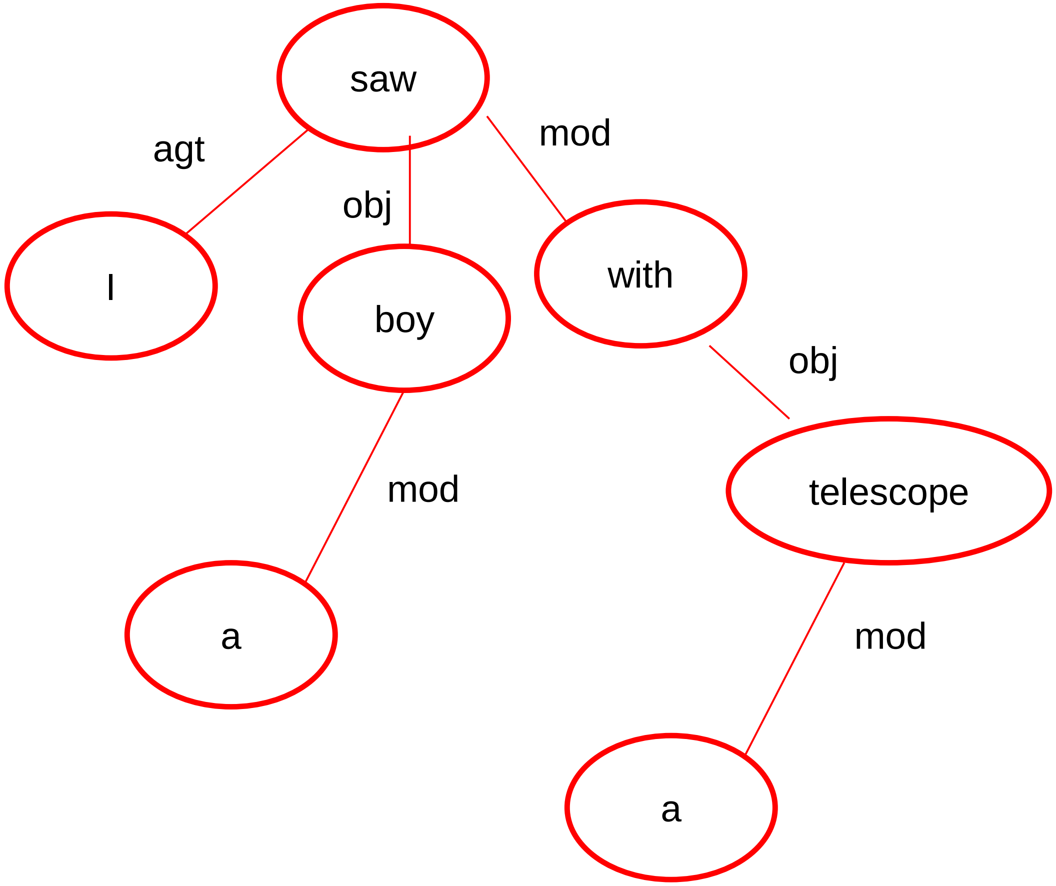


I saw a boy with a telescope

Dependency Parse Tree - 1



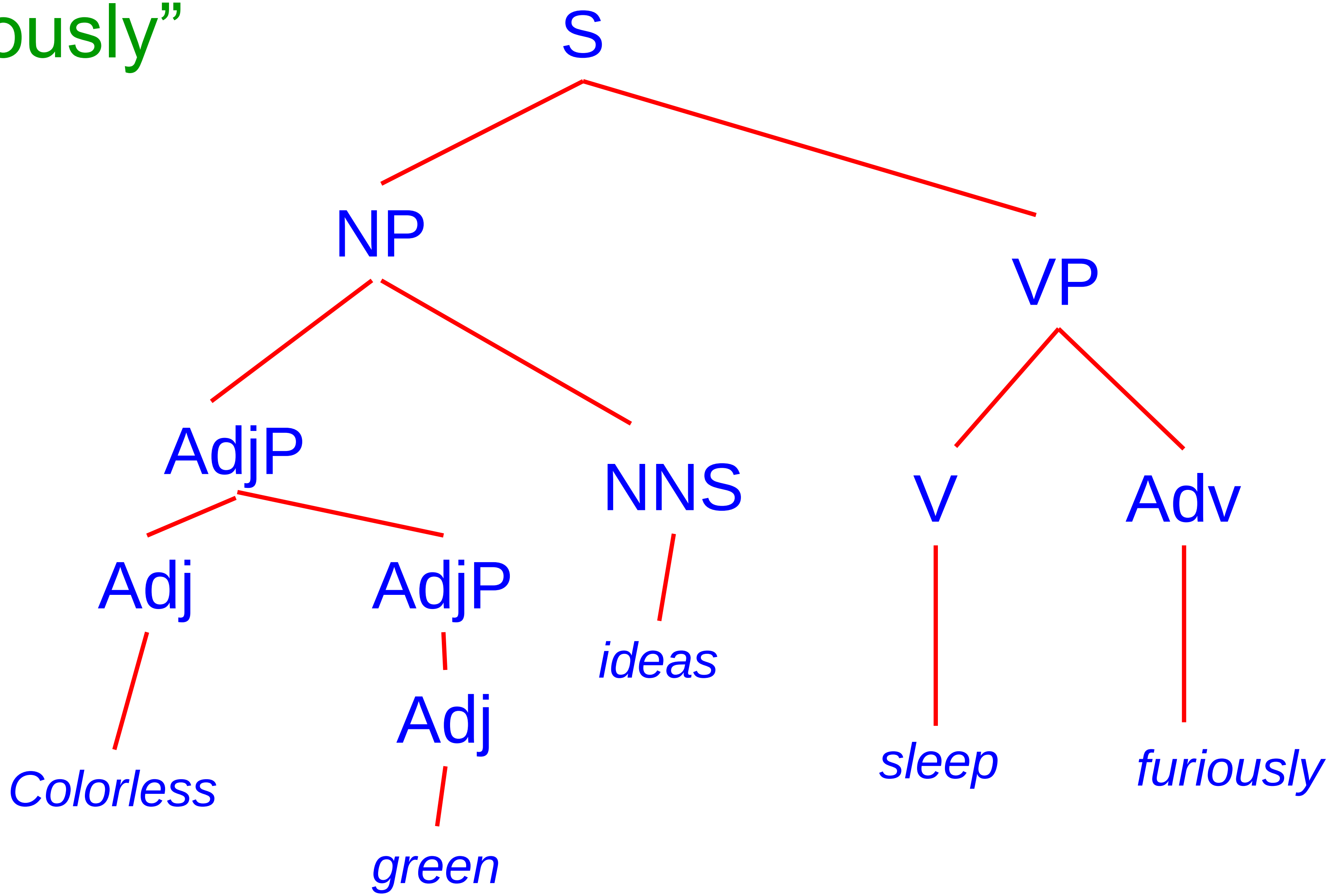
Dependency Parse Tree - 2



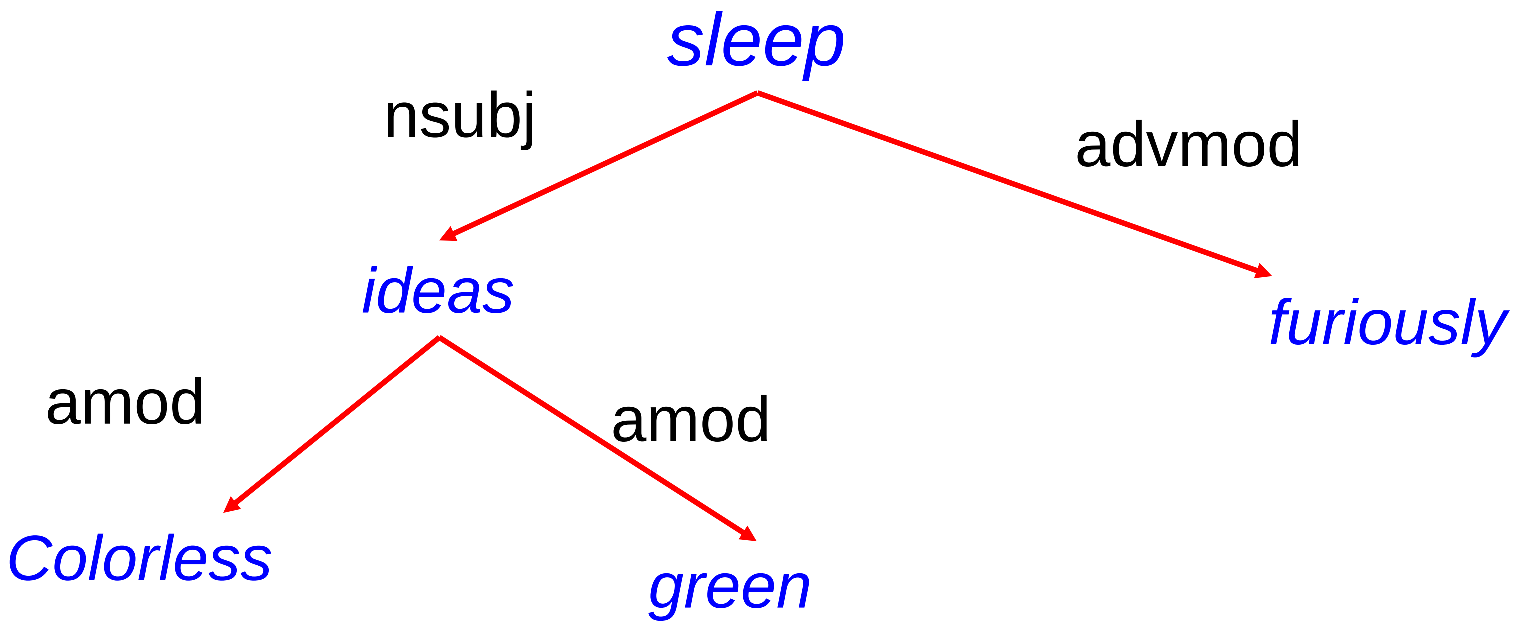
Foundation of Dependency Tree (DT)

- 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, $Jack_{nsubj}$ $gave_{mv}$ a $book_{dobj}$ to $Jill_{iobj}$.

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

Broca’s Area: wikipedia

- “Patients with lesions in Broca's area exhibit agrammatical speech production; they also show inability to use syntactic information to determine the meaning of sentences.”

Broca's area damage (1/2) (Wikipedia)

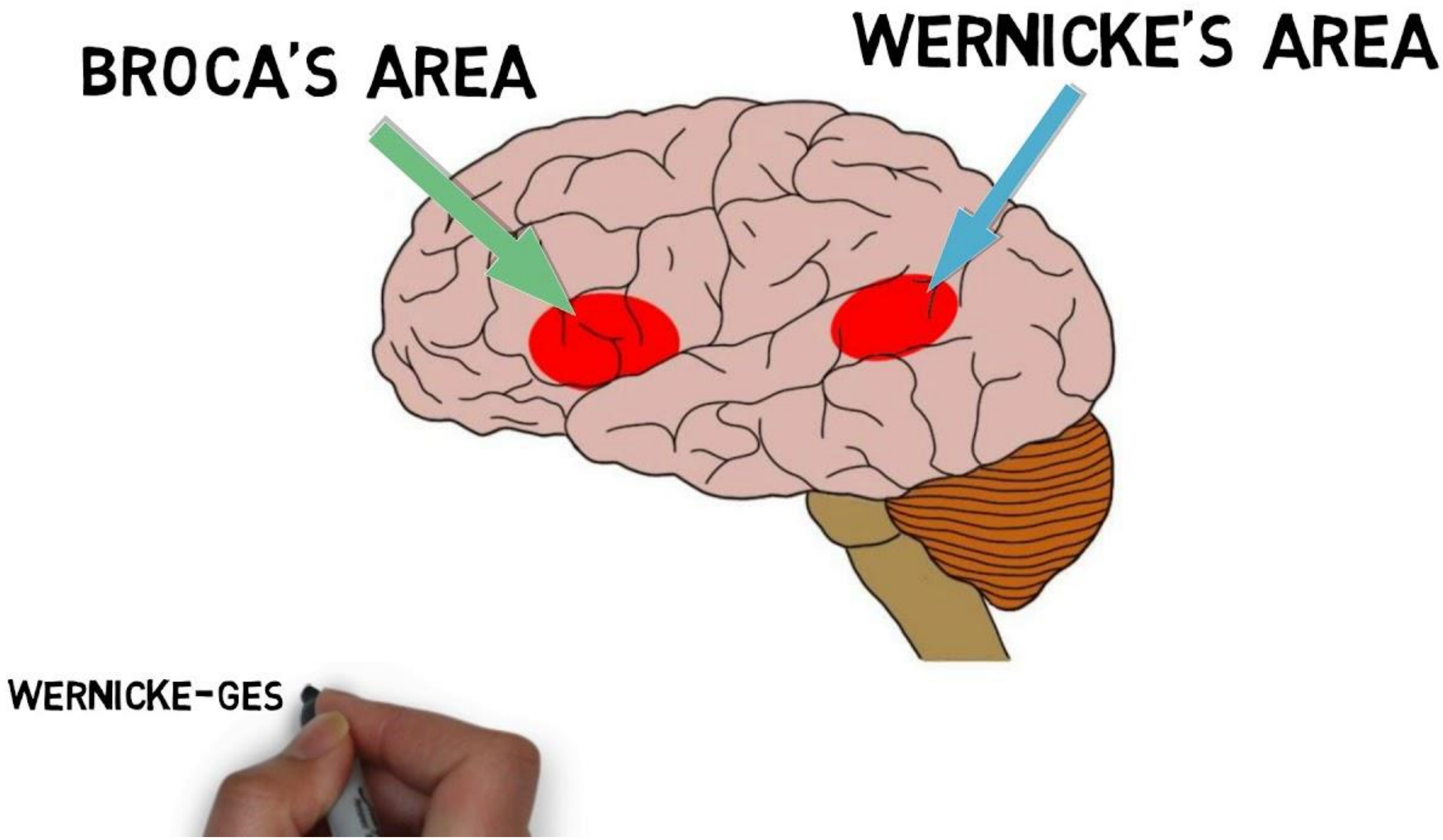
- Agrammatical speech production
- Disconnected speech
- Mainly loss of function words and affixes
- Uses **telegraphic** speech
  - *"I eat rice spoon" → I eat rice with spoon*
  - *"Many boy play football yesterday" → Many boys played football yesterday*

Broca’s Area Damage (2/2)

- Cannot use syntax to arrive at precise meaning

- No difference between
  - *Visit to the President vs.*
  - *Visit by the President*

Wernicke’s Area (Wikipedia)



Wernicke’s Area Damage

- “Damage caused to Wernicke's area results in receptive, fluent aphasia. This means that the person with aphasia will be able to fluently connect words, but the phrases will lack meaning”.
  - *The pink elephant sang and the blue stone danced*

Grammar Rules

A segment of English

- S → NP VP
- NP → DT NN
- NP → NNS
- NP → NP PP
- PP → P NP
- VP → VP PP
- VP → VBD NP

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

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** *.*

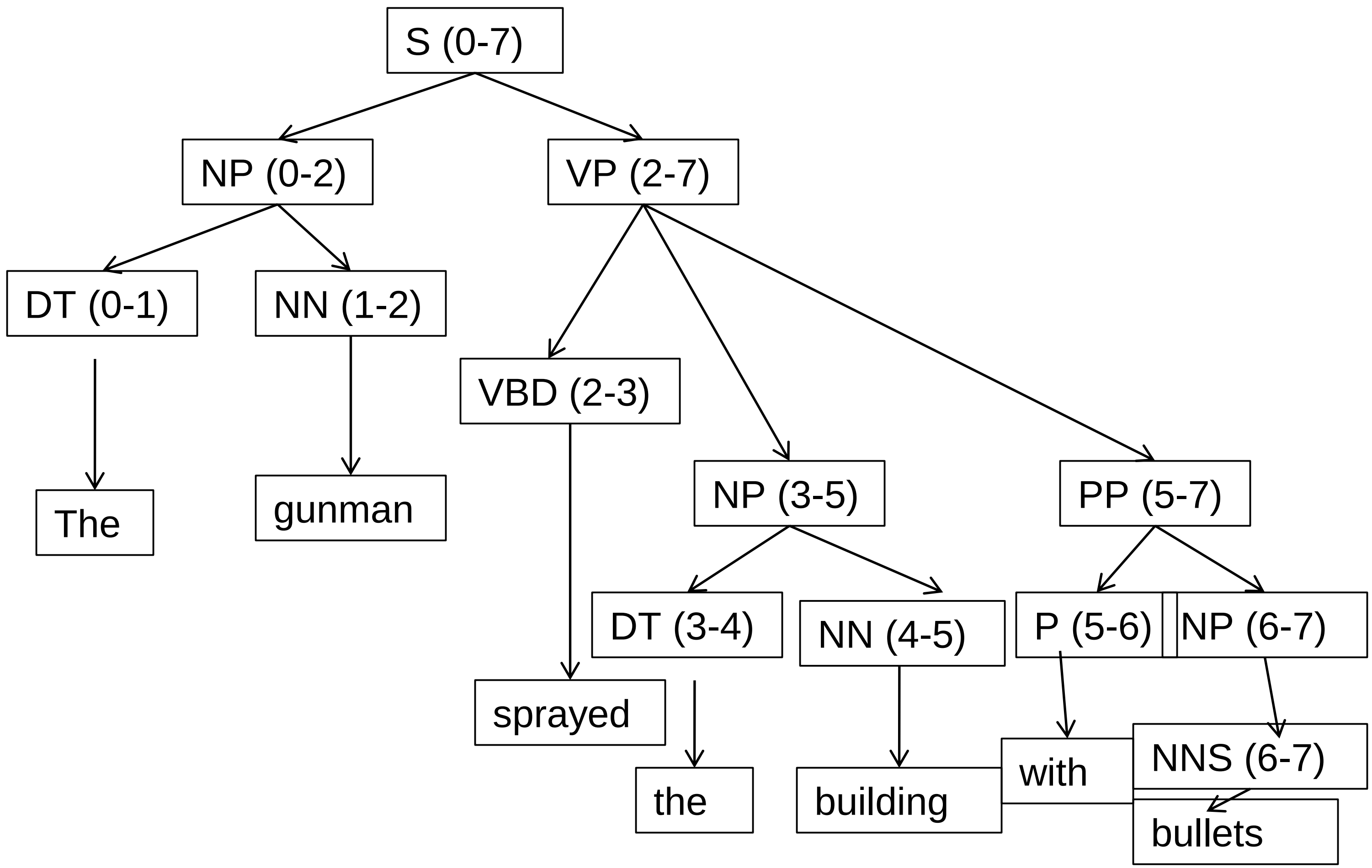
To From	1	2	3	4	5	6	7
0	DT	NP	--------	---------	S	---------	S
1	-------	NN	--------	---------	---------	---------	---------
2	-------	---------	VBD	---------	VP	---------	VP
3	-------	---------	--------	DT	NP	---------	NP
4	--------	---------	--------	---------	NN	---------	---------
5	--------	---------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

CYK: Extracting the Parse Tree

- The parse tree is obtained by keeping back pointers.

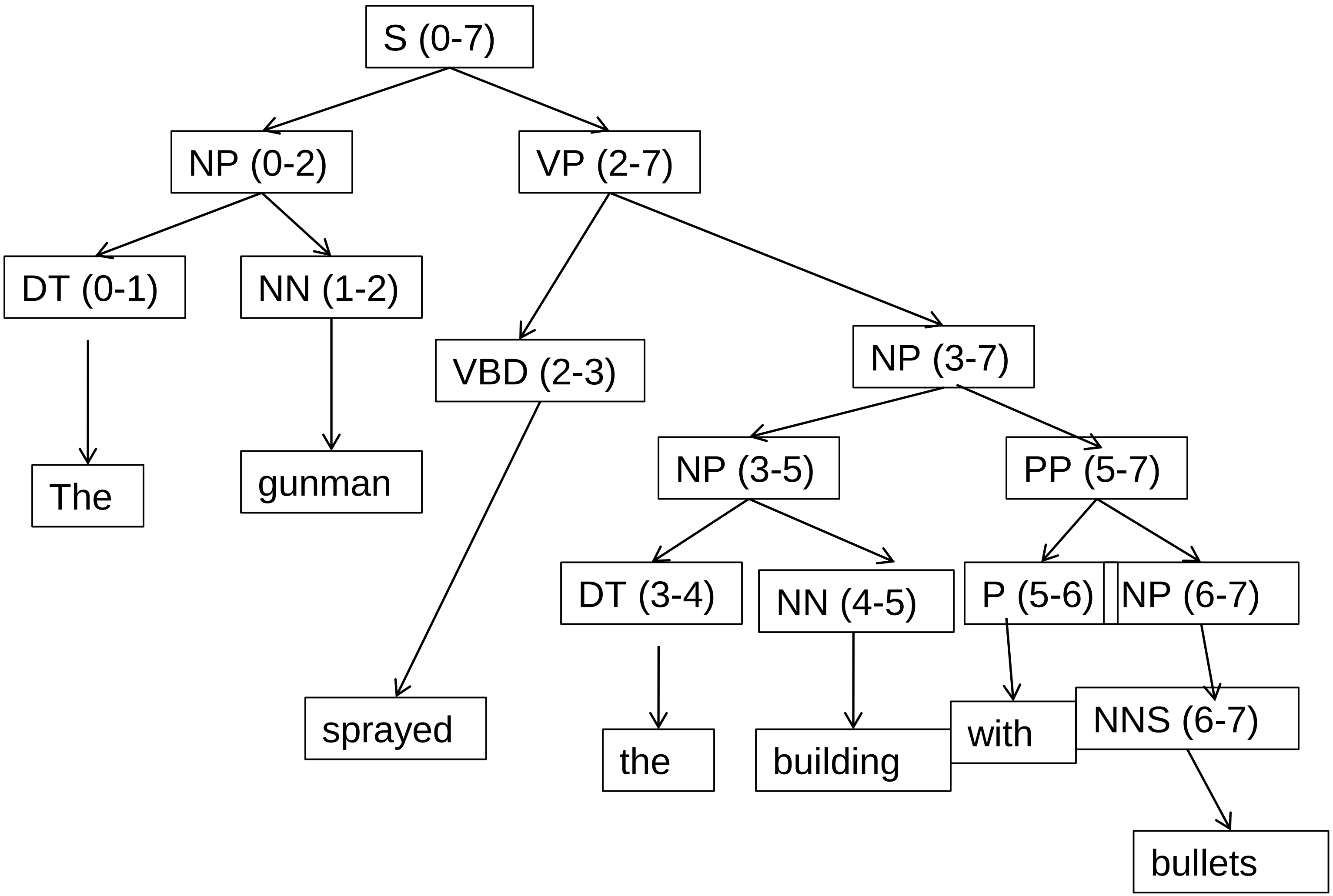
Parse Tree #1

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.

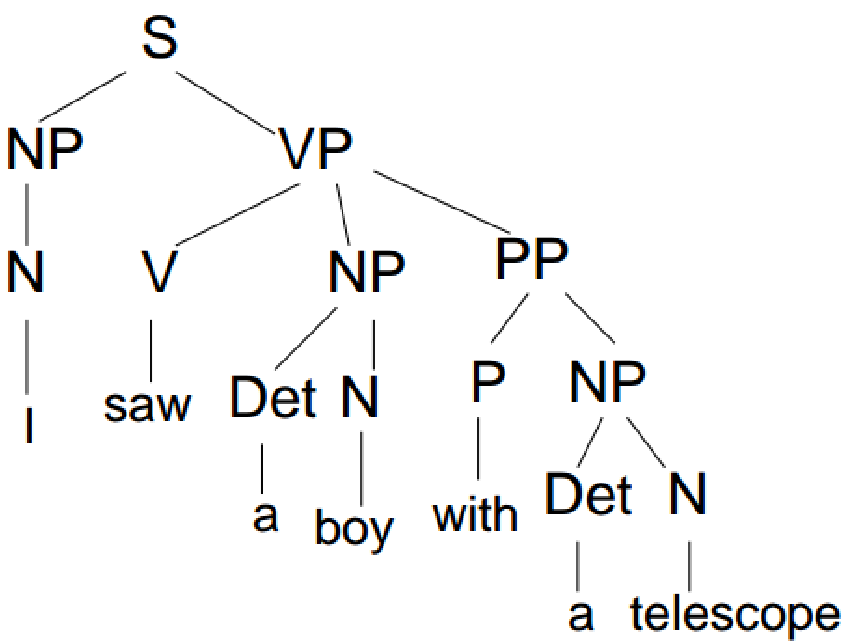


Notion of Domination

- 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

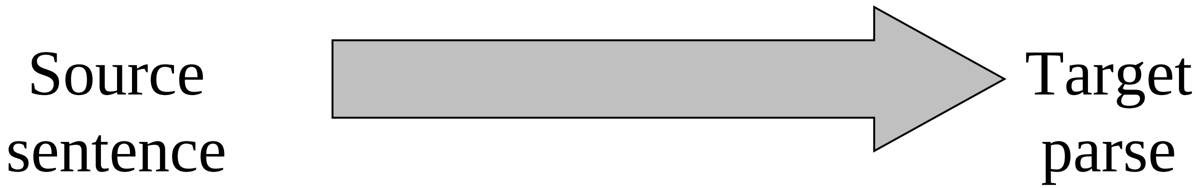


- Dominations
  - NP dominates “a telescope”
  - VP dominates “saw a boy with a telescope
  - S dominates the whole sentence
- Domination is composed of many sub-domination.

- I saw a boy with a telescope
  - Meaning: I used the telescope to see the boy

Probabilistic parsing

Noisy Channel Modeling



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

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

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

Example of Sentence labeling: Parsing

$[_{S1}[_{S}[_{S}[_{VP}[_{VB}$Come$][_{NP}[_{NNP}$July$]]]]]$

$[_{,}$,$]$

$[_{CC}$ and$]$

$[_{S}$ $[_{NP}$ $[_{DT}$ the$]$ $[_{JJ}$ IIT$]$ $[_{NN}$ campus$]]$

$[_{VP}$ $[_{AUX}$ is$]$

$[_{ADJP}$ $[_{JJ}$ abuzz$]$

$[_{PP}[_{IN}$ with$]$

$[_{NP}[_{ADJP}$ $[_{JJ}$ new$]$ $[_{CC}$ and$]$ $[$ $_{VBG}$ returning$]]$

$[_{NNS}$ students$]]]]]]$

$[_{.}$.$]]]$

Formal Definition of PCFG

- A PCFG consists of
  - 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 N^1
  - A set of rules $\{N^i \rightarrow \zeta^j\}$, where ζ^j is a sequence of terminals & non-terminals

    NP $\rightarrow$ DT NN
  - A corresponding set of rule probabilities

Rule Probabilities

- Rule probabilities are such that 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

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

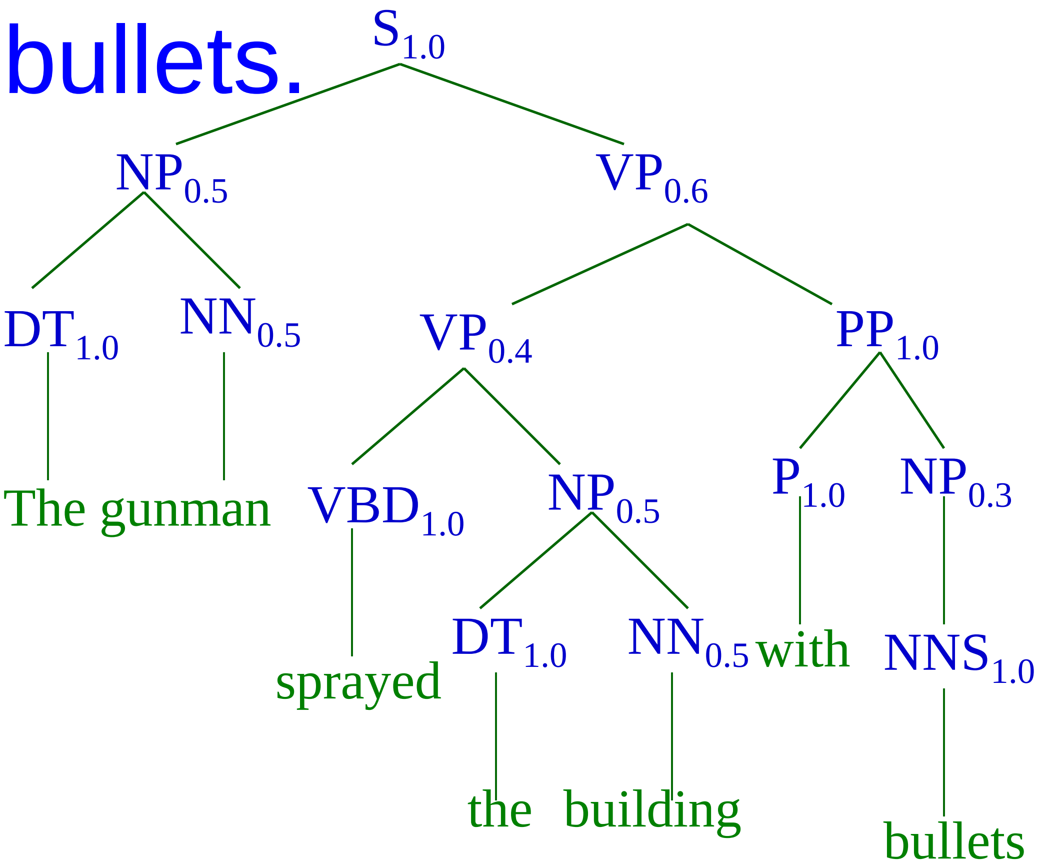
Probabilistic Context Free Grammars

- S → NP VP 1.0
- NP → DT NN 0.5
- NP → NNS 0.3
- NP → NP PP 0.2
- PP → P NP 1.0
- VP → VP PP 0.6
- VP → VBD NP 0.4

- DT → the 1.0
- NN → gunman 0.5
- NN → building 0.5
- VBD → sprayed 1.0
- NNS → bullets 1.0

Example Parse t_1`

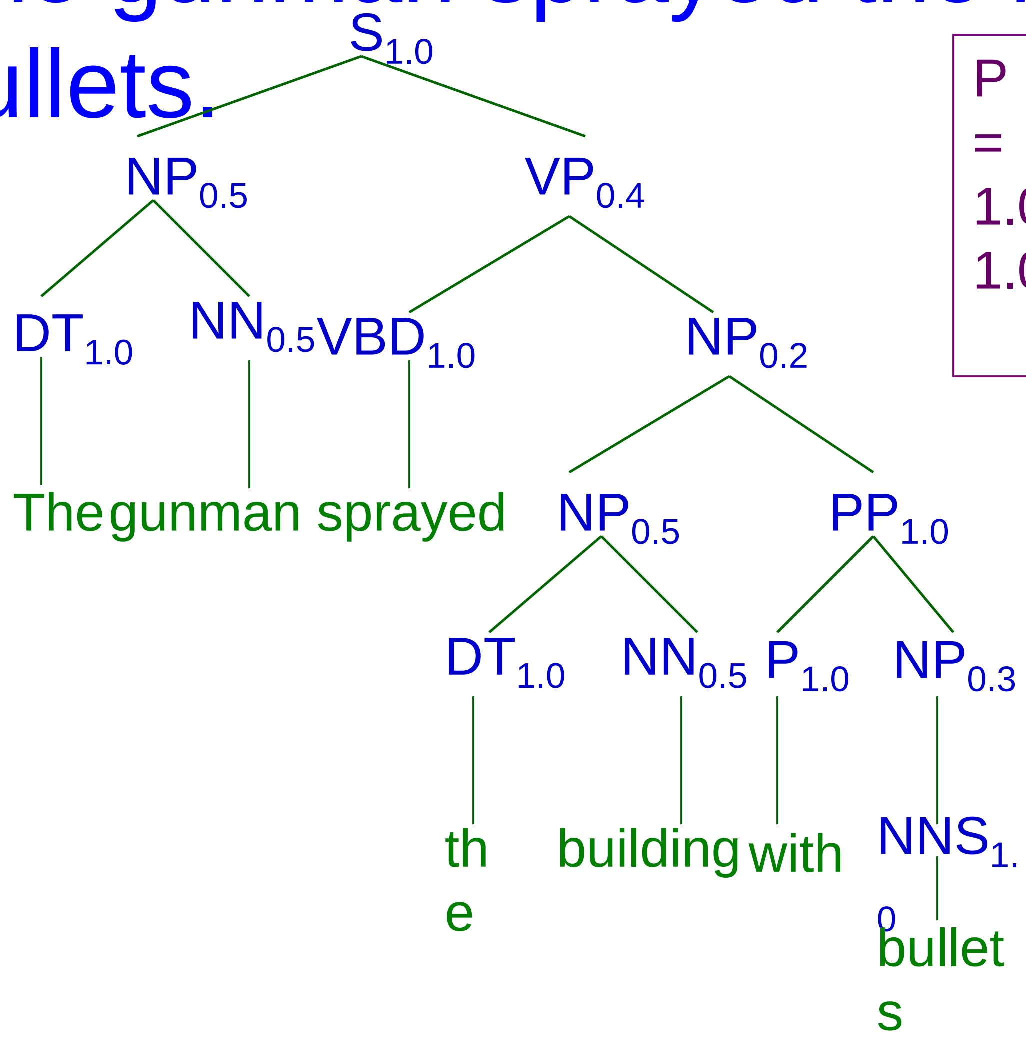
- The gunman sprayed the building with bullets.



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

Another Parse t_2

- The gunman sprayed the building with bullets.

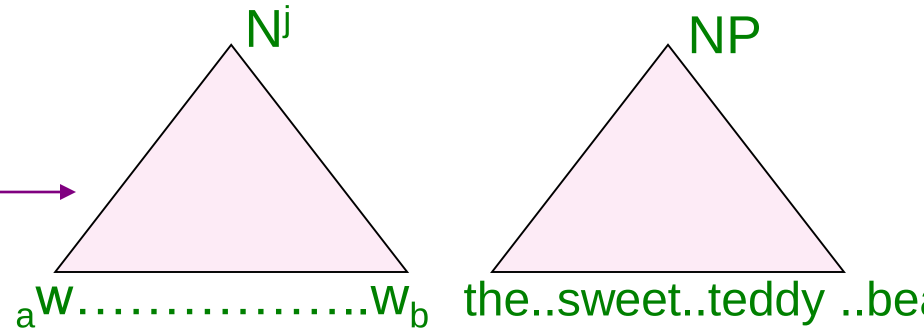


P (t_2)
= 1.0 * 0.5 * 1.0 * 0.5 * 0.4 * 1.0 * 0.2 * 0.5 * 1.0 * 0.5 * 1.0 * 1.0 * 0.3 * 1.0
= 0.0015

Probability of a sentence

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 = $P(w_{0,l})$ *(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 :

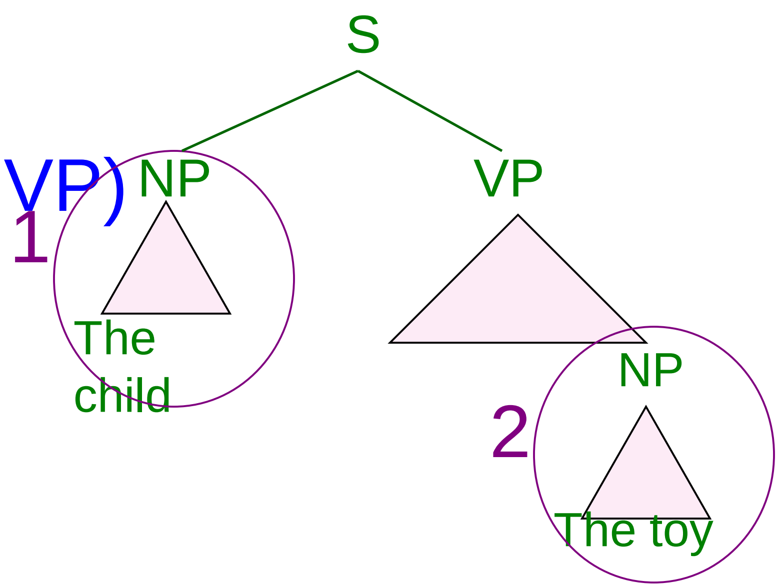
  P(NP → DT NN) is same in locations 1 and 2

- Context-free :

  P(NP → DT NN | anything outside "The child")
  = P(NP → DT NN)

- Ancestor free : At 2,

  P(NP → DT NN|its ancestor is VP)
  = P(NP →DT NN)

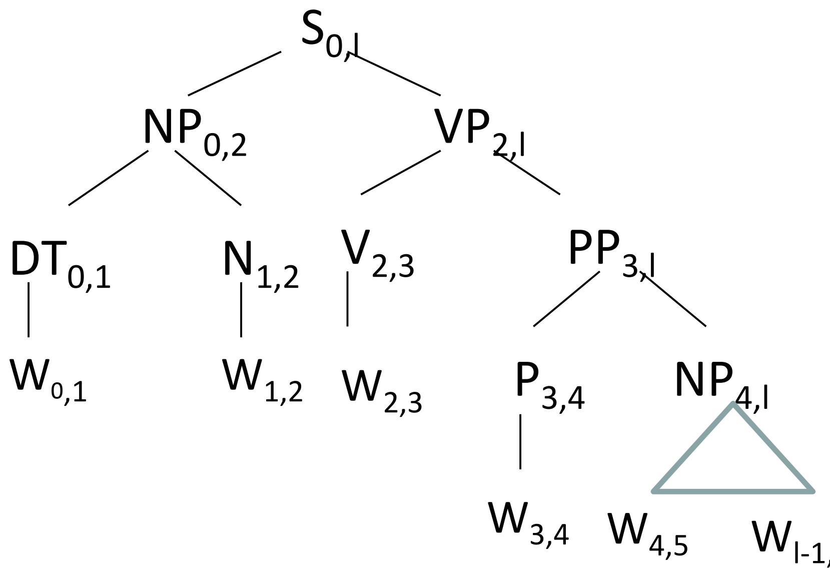


Probability of a parse tree

- *Domination* :We say N_j dominates from *k* to *l*, symbolized as $N^j_{k,l}$, if $W_{k,l}$ is derived from N_j
- P (tree |sentence) = P (tree | $S_{1,l}$)

  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)

Probability of a parse tree (cont.)



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

$= P(NP_{0,2}, DT_{0,1}, w_{0,1}, N_{1,2}, w_{1,2}, VP_{2,l}, V_{2,3}, w_{2,3},$
$PP_{3,l}, P_{3,4}, w_{3,4}\ NP_{4,l}, w_{4,l} \mid S_{0,l})$

$= P(NP_{0,2}, VP_{2,l} \mid S_{0,l}) * P(DT_{0,1}, N_{1,2} \mid NP_{0,2}) *$
$P(w_{0,1} \mid DT_{0,1}) * P(w_{1,2} \mid N_{1,2}) * P(V_{2,3}, PP_{3,l} \mid VP_{2,l}) *$
$P(w_{2,3} \mid V_{2,3}) * P(P_{3,4}, NP_{4,l} \mid PP_{3,l}) * P(w_{3,4} | P_{3,4}) *$
$P(w_{4,l} \mid NP_{4,l})$

(Using Chain Rule, Context Freeness and Ancestor Freeness)