CS626: Speech, NLP and Web

Probabilistic Parsing cntd, Dependency Parsing

Pushpak Bhattacharyya

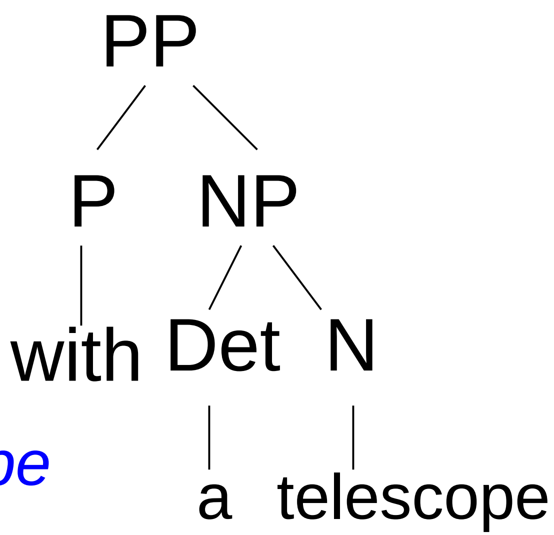
Computer Science and Engineering Department

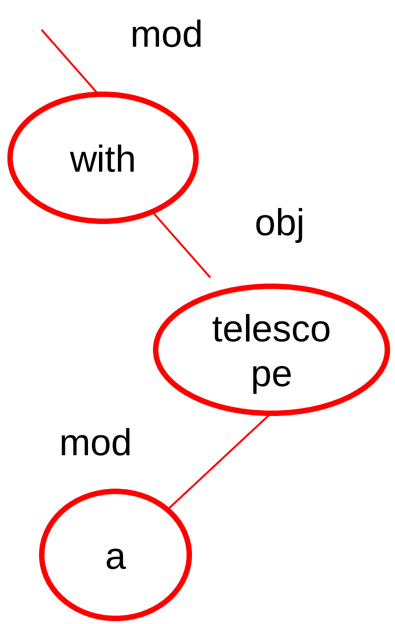
IIT Bombay

Week 7 of 9th September, 2024

1-slide recap of week of 2nd Sep

- CP: meaning of parent-child and sibling relationships- constituent; head-modifier
- Pre-modifier and post-modifier
- DP: Head-modifier expressed directly
- Ambiguity resolution by proximity: *telescope* example
- TD, BU, TDBY, CYK Parsing
- Notion of Domination: *X[p:q], the phrase/POS X dominates (generates) text segment from position p to position q*
- Probability of a CFG rule P(A→ B C) means P(B C|A)
- **Independence** of position, context and ancestry





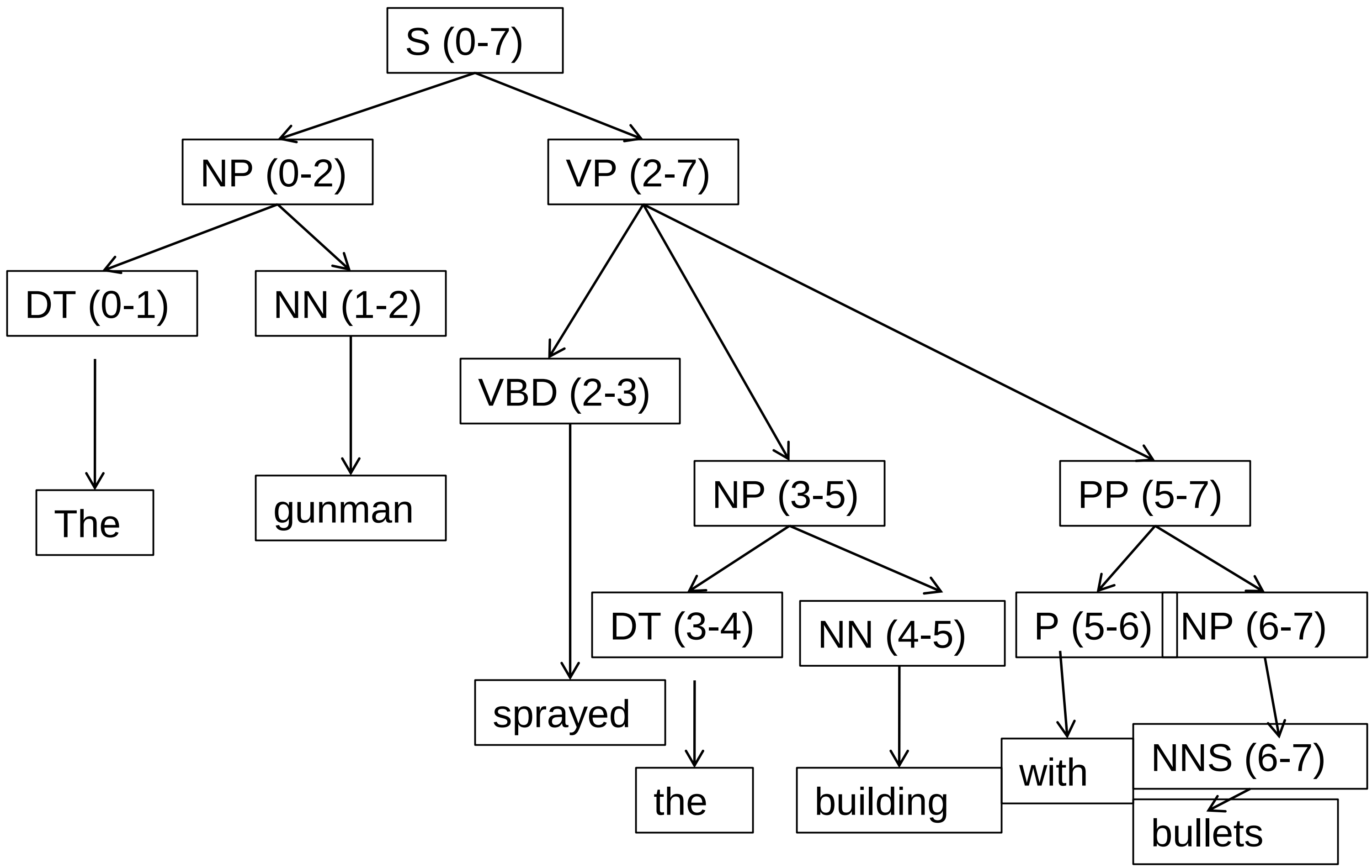
Notion of Domination

- 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.

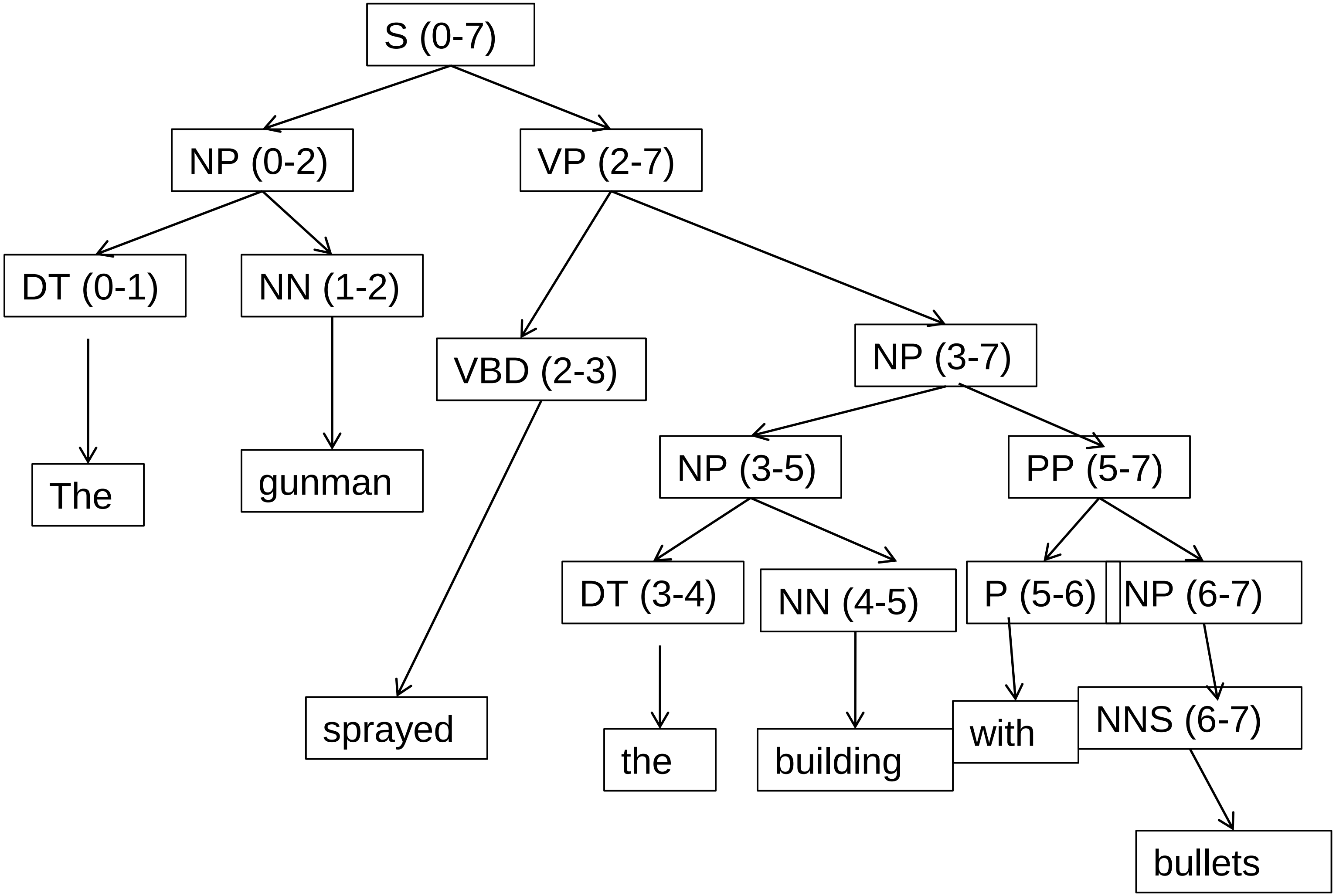
Parse Tree #1

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

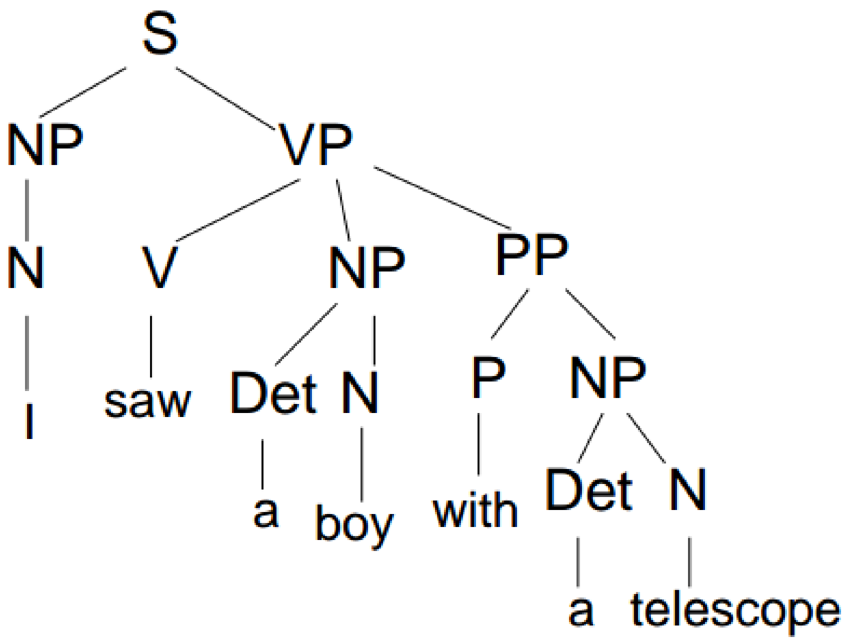


Parse Tree #2

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



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:

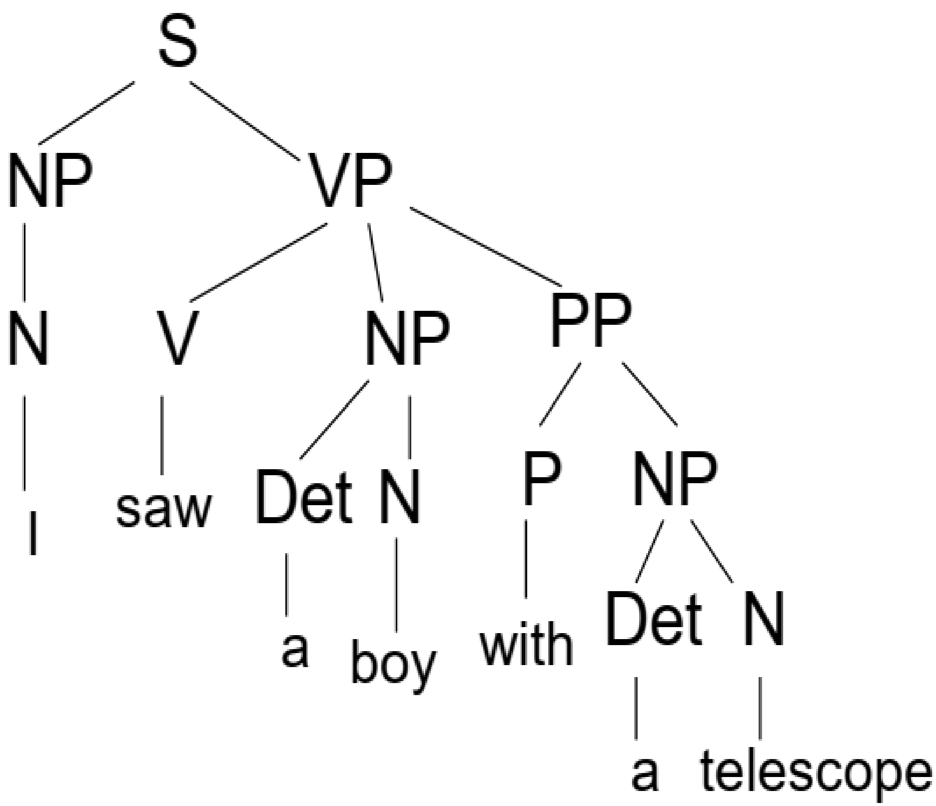
Christopher Manning and Heinrich Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

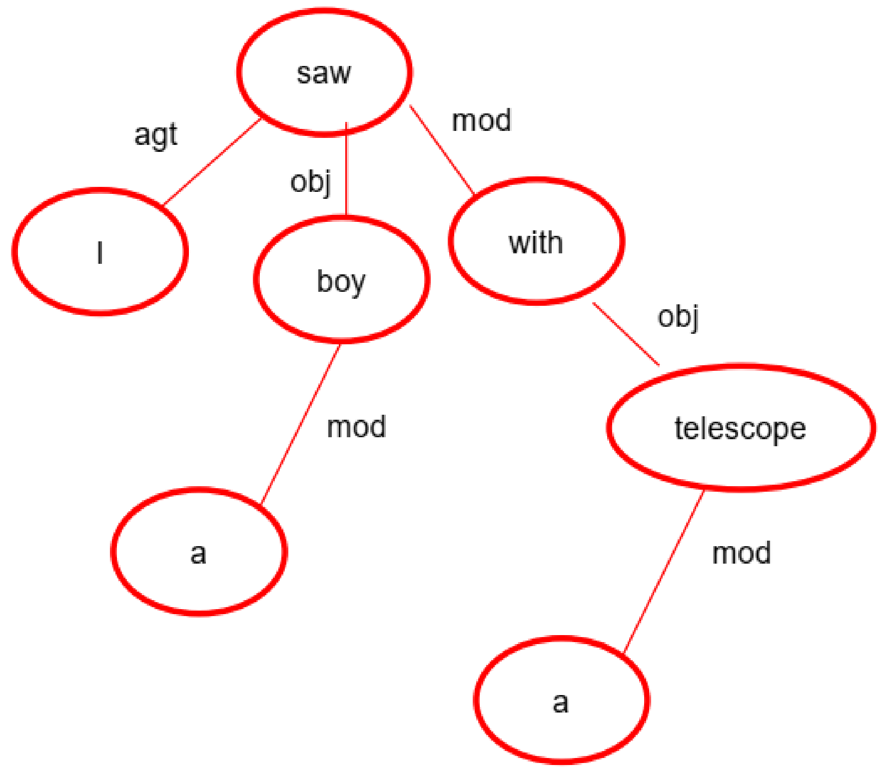
Noisy Channel Modeling

Source sentence $\longrightarrow$ Target parse

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

“I saw…”: CP and DP #1





Bracketed Structure #1

Parse #1 (meaning: I have the telescope)

[

$\quad [\text{I}]_{NP}$

$\quad [$

$\qquad [\text{saw}]_{VBD}$

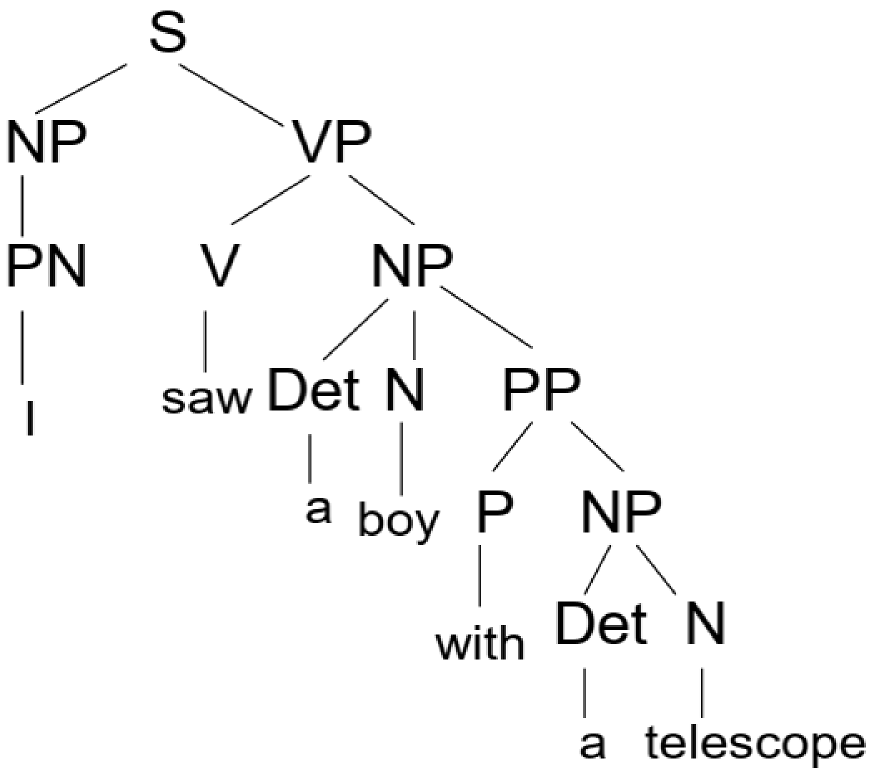
$\qquad [\text{the boy}]_{NP}$

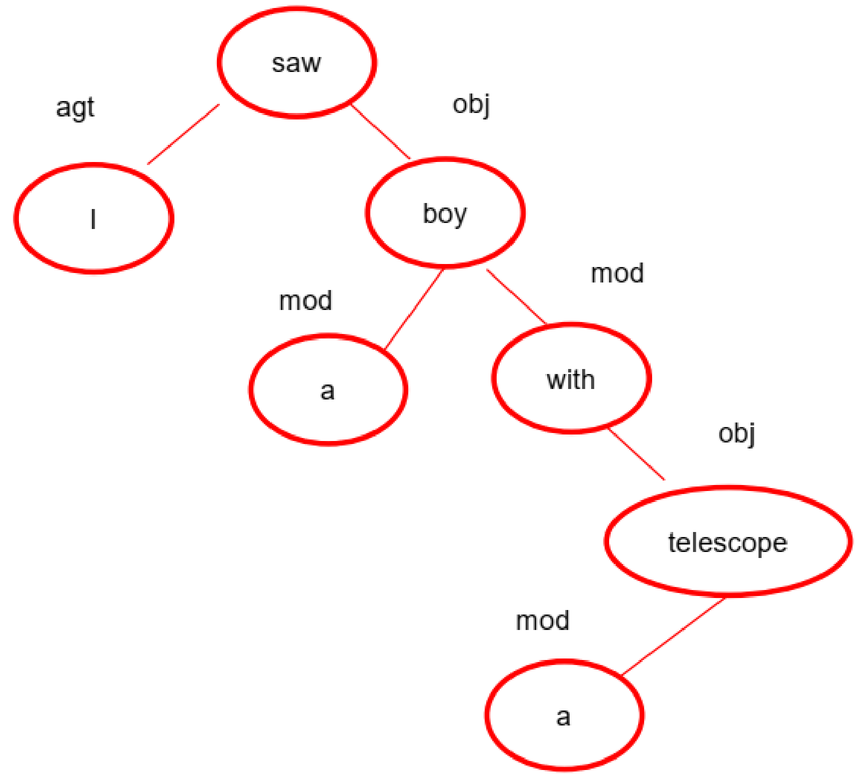
$\qquad [\text{with } [\text{a telescope}]_{NP}]_{PP}$

$\quad]_{VP}$

$]_{S}$

“I saw…”: CP and DP #2





Bracketed Structure #2

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

[

$\quad$ $[\text{I}]_{NP}$

$\quad$ [

$\quad\quad$ $[\text{saw}]_{VBD}$

$\quad\quad$ [

$\quad\quad\quad$ $[\text{the boy}]_{NP}$

$\quad\quad\quad$ $[\text{with } [\text{a telescope}]_{NP}]_{PP}$

$\quad\quad$ $]_{NP}$

$\quad$ $]_{VP}$

$]_{S}$

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

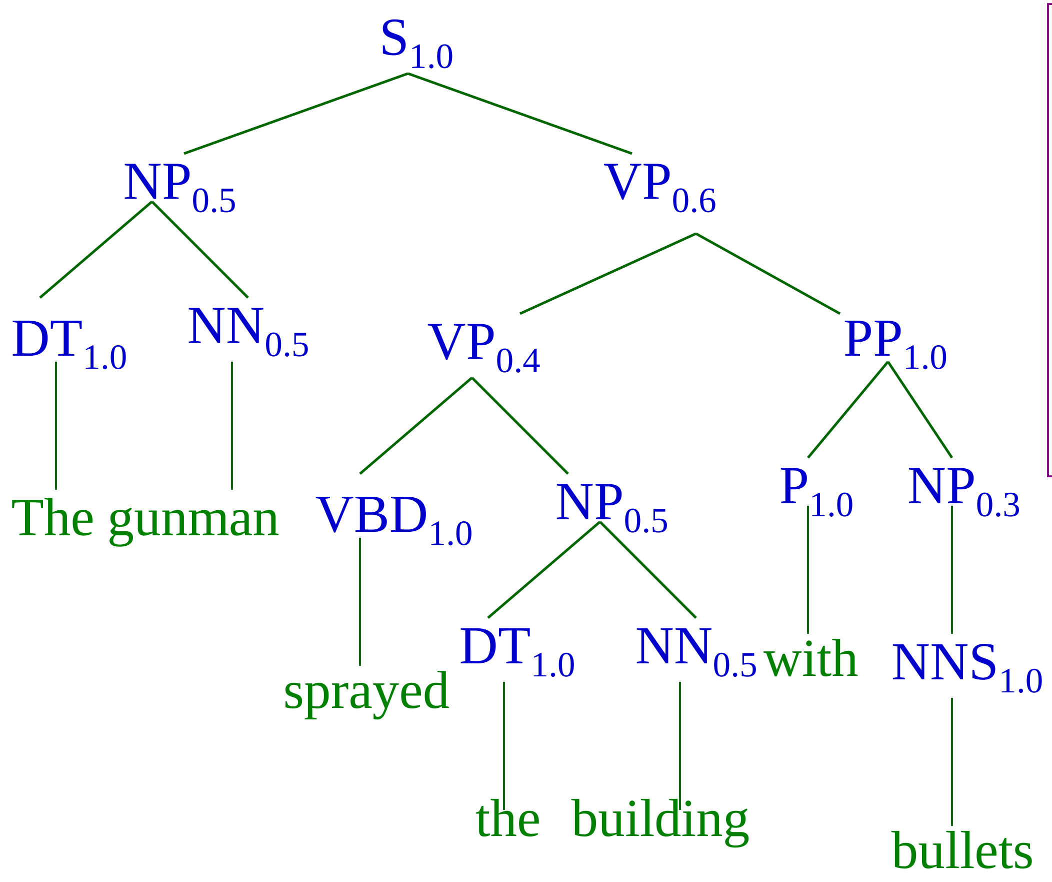
Probabilistic Context Free Grammars

- S → 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.

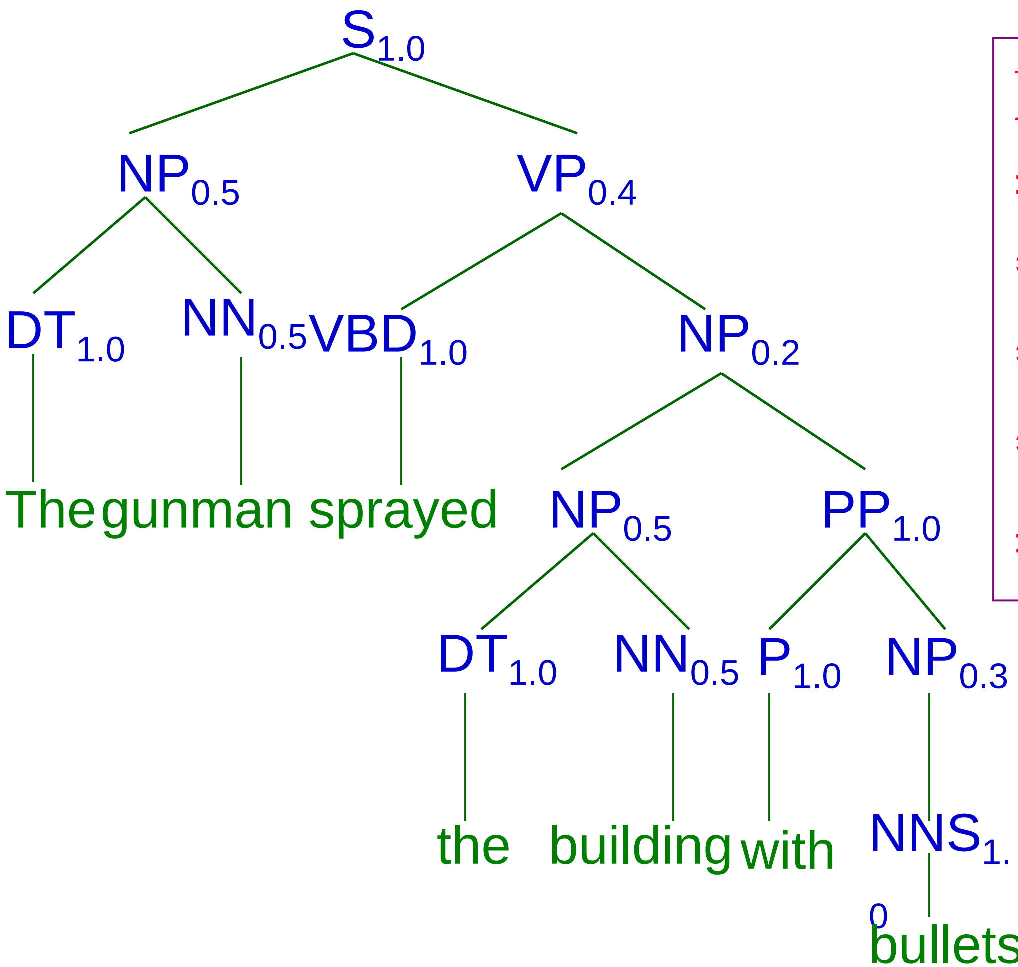


$P(t_1) = 1.0 * 0.5 * 1.0 * 0.5 * 0.6 * 0.4 * 1.0 * 0.5 * 1.0 * 0.5 * 1.0 * 1.0 * 0.3 * 1.0$

$= 0.00225$

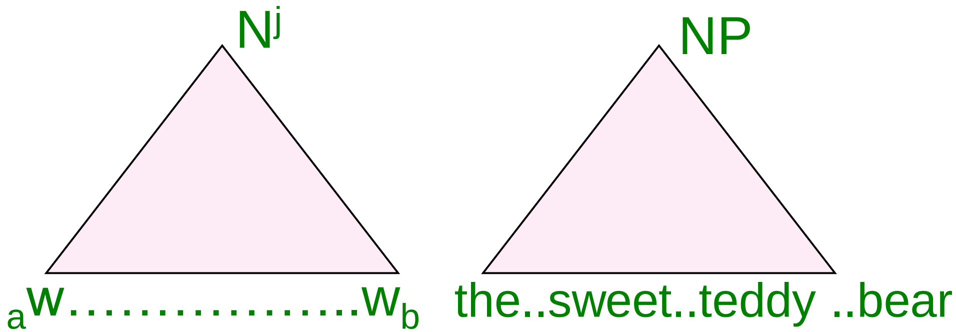
Another Parse t_2

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 (1/2)



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

– w_{ab}– subsequence $_a w \ldots. w_b$

– N^j dominates $_a w \ldots. w_b$

or yield(N^j) = $_a w \ldots. w_b$

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:

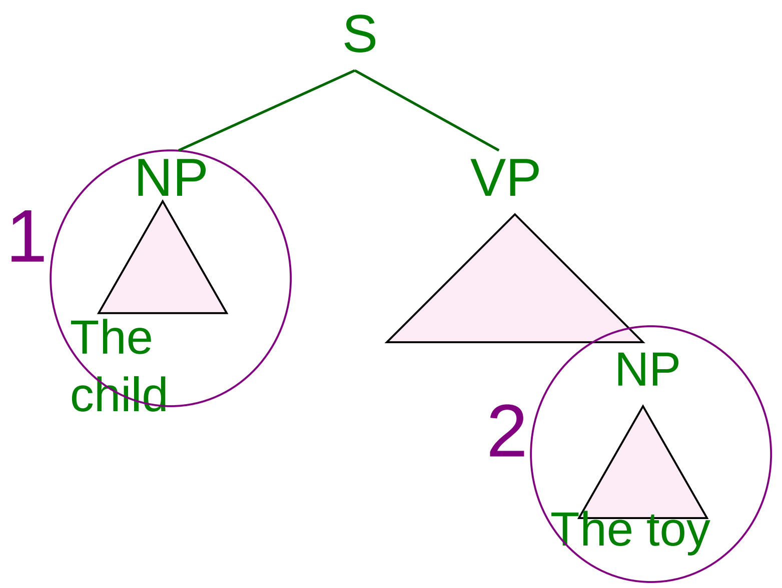
  P(NP → DT NN) is same independent of location in the tree

- 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 $w_{0,l}$

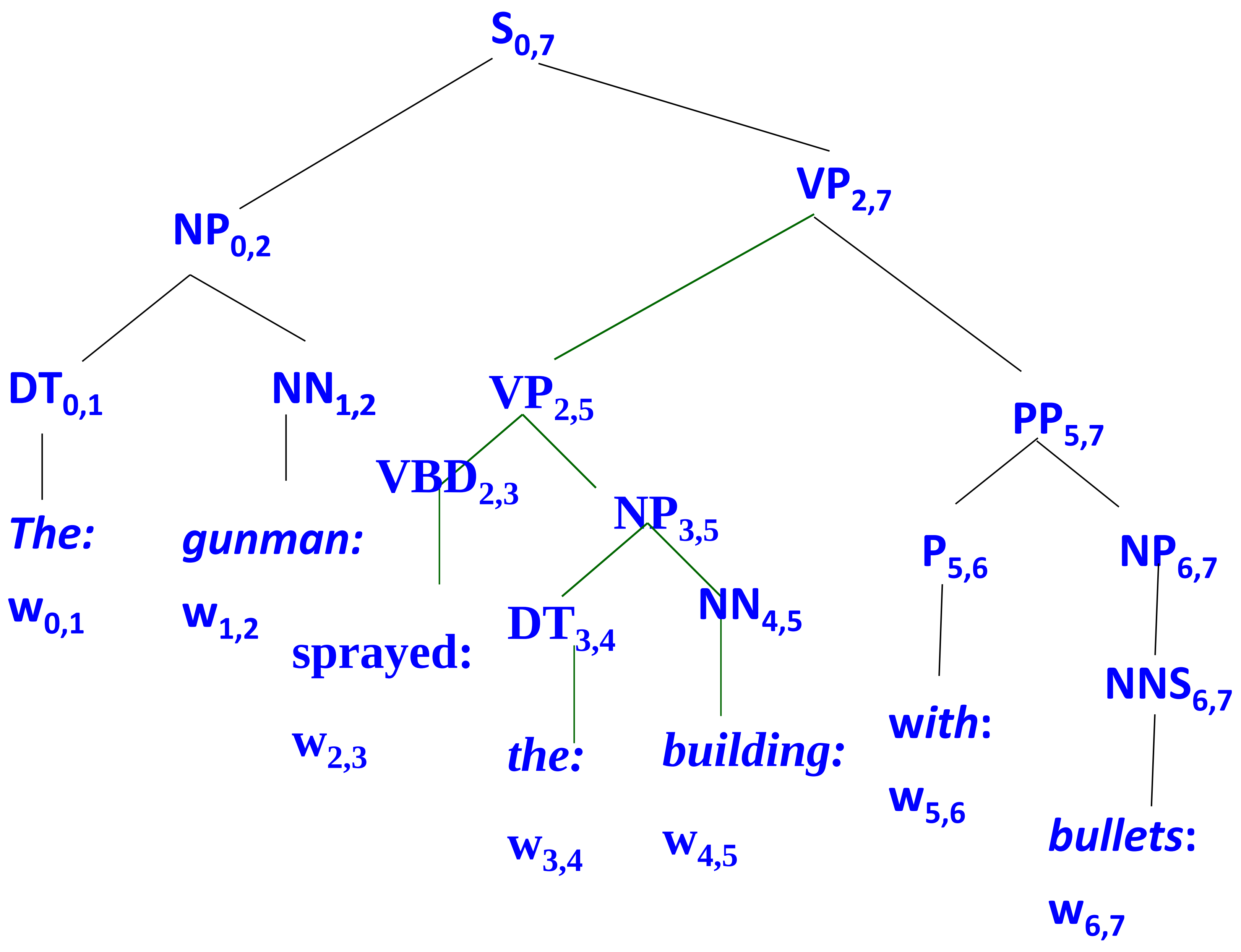
P(t|s) approximately equals joint probability of constituent non-terminals dominating the sentence fragments (next slide)

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)

Illustration with “The man saw the boy with a telescope”

From Bhattacharyya and Joshi. 2023. *Natural Language Processing.* Wiley.

CFG: $_0$ *the* $_1$ *man* $_2$ *saw* $_3$ *the* $_4$ *boy* $_5$ *with* $_6$ *a* $_7$ *telescope* $_8$

Syntax	**Lexicon**
S → NP VP	DT → *the*
NP → DT NN	DT → *a*
NP → NP PP	NN → *man*
VP → VBD	NN → *boy*
VP → VP NP	NN → *telescope*
VP → VP PP	VBD → *saw*
PP → P NP	P → *with*

PCFG: CYK: $_0$ *the* $_1$ *man* $_2$ *saw* $_3$ *the* $_4$ *boy* $_5$ *with* $_6$ *a* $_7$ *telescope* $_8$

Syntax		
S→	NP VP	1.0
NP→	DT NN	0.5
NP→	NP PP	0.5
VP→	VBD	0.2
VP→	VP NP	0.5
VP→	VP PP	0.3
PP→	P NP	1.0

Lexicon		
DT→	*the*	0.5
DT→	*a*	0.5
NN→	*man*	0.33
NN→	*boy*	0.33
NN→	*telescope*	0.33
VBD→	*saw*	1.0
P→	*with*	1.0

$_0$ *the* $_1$ *man* $_2$ *saw* $_3$ *the* $_4$ *boy* $_5$ *with* $_6$ *a* $_7$ *telescope* $_8$

Positions (row-col)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0	*the*$_{01}$ DT_{01}	NP_{02}	S_{03}	--	S_{05}	--	--	S_{08}
1		*man*$_{12}$ NN_{12}	--	--	--	--	--	--
2			*saw*$_{23}$ VBD_{23} VP_{23}	--	VP_{25}	--	--	VP_{28} (VP_{25} PP_{58}/VP_{23} NP_{38})
3				*the*$_{34}$ DT_{34}	NP_{35}	--	--	NP_{38}
4					*boy*$_{45}$ NN_{45}	--	--	--
5						*with*$_{56}$ P_{56}	--	PP_{58}
6							a_{67} DT_{67}	NP_{68}
7								*telescope*$_{78}$ NN_{78}

$_0$ *the* $_1$ *man* $_2$ *saw* $_3$ *the* $_4$ *boy* $_5$ *with* $_6$ *a* $_7$ *telescope* $_8$

Positions (row-col)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0	*the*$_{01}$ DT_{01} 0.5	NP_{02} 0.0825	S_{03} 1.0*0.08*0.2=0.0165	--	S_{05}	--	--	S_{08} 0.0825*0.0002/0.825*0.004=0.0000165/0.00033
1		*man*$_{12}$ NN_{12} 0.33	--	--	--	--	--	--
2			*saw*$_{23}$ VBD_{23} VP_{23} 1.0*0.2=0.2	--	VP_{25} 0.5*0.2*0.0825 =0.00825	--	--	VP_{28} VP_{25} PP_{58}/VP_{23} NP_{38} 0.3*0.00825*0.0825/0.5*0.2*0.04125=0.0002/0.004
3				*the*$_{34}$ DT_{34} 0.5	NP_{35} 0.0825	--	--	NP_{38} 0.5*0.0825=0.04125
4					*boy*$_{45}$ NN_{45} 0.33	--	--	--
5						*with*$_{56}$ P_{56} 1.0	--	PP_{58} 1.0*0.0.0825=0.0825
6							*a*$_{67}$ DT_{67} 0.5	NP_{68} 0.5*0.165=0.0825
7								*telescope*$_{78}$ NN_{78} 0.33

Probability Computations

$P(NP_{02}) = P(NP \rightarrow DT\ NN) \times P(DT_{01}) \times P(NN_{12}) = 0.5 \times 0.5 \times 0.33$

$= 0.0825$

$P(VP_{28}) = P(VP \rightarrow VP\ PP) \times P(VP_{25}) \times P(PP_{58})$ or

$= P(VP \rightarrow VP\ NP) \times P(VP_{23}) \times P(NP_{38})$

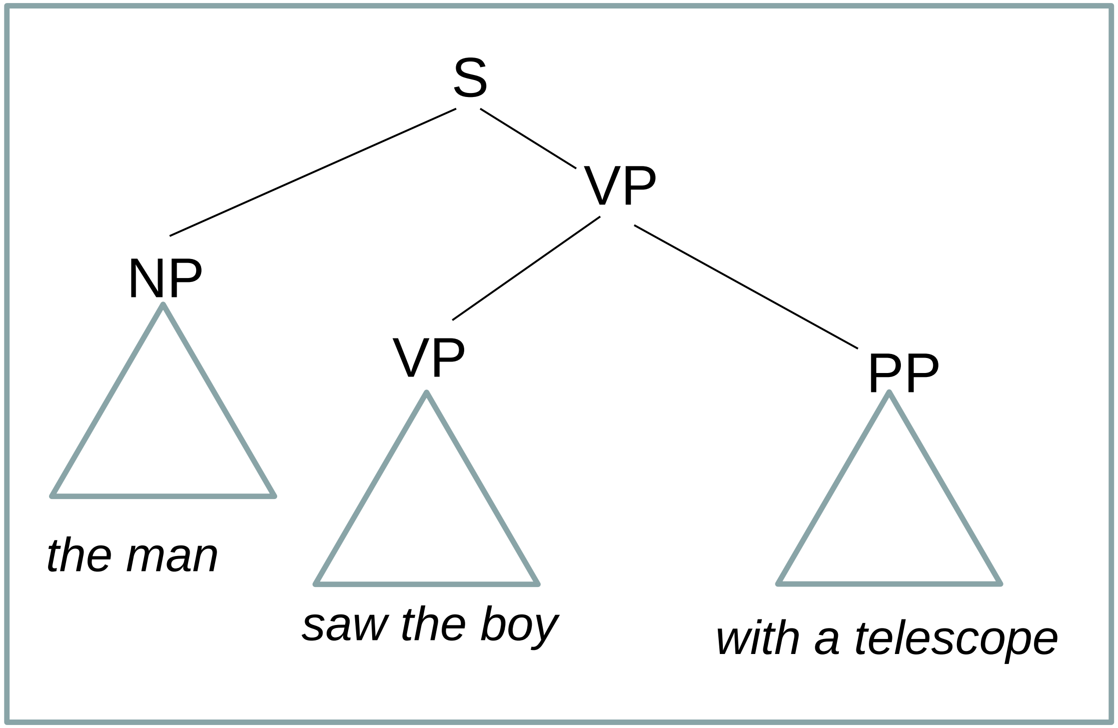
Now, $P(VP_{25}) = P(VP \rightarrow VP\ NP) \times P(VP_{23}) \times P(NP_{35}) = 0.5 \times P(VP_{23}) \times P(NP_{35})$

But, $P(VP_{23}) = P(VP \rightarrow VBD) \times P(VBD_{23}) = 0.2 \times P(VBD \rightarrow saw) = 0.2 \times 1.0 = 0.2$

$P(NP_{35}) = P(NP \rightarrow^* \text{'}the\ boy\text{'}) = 0.0825$ ($\rightarrow^*$ means chain of rule application)

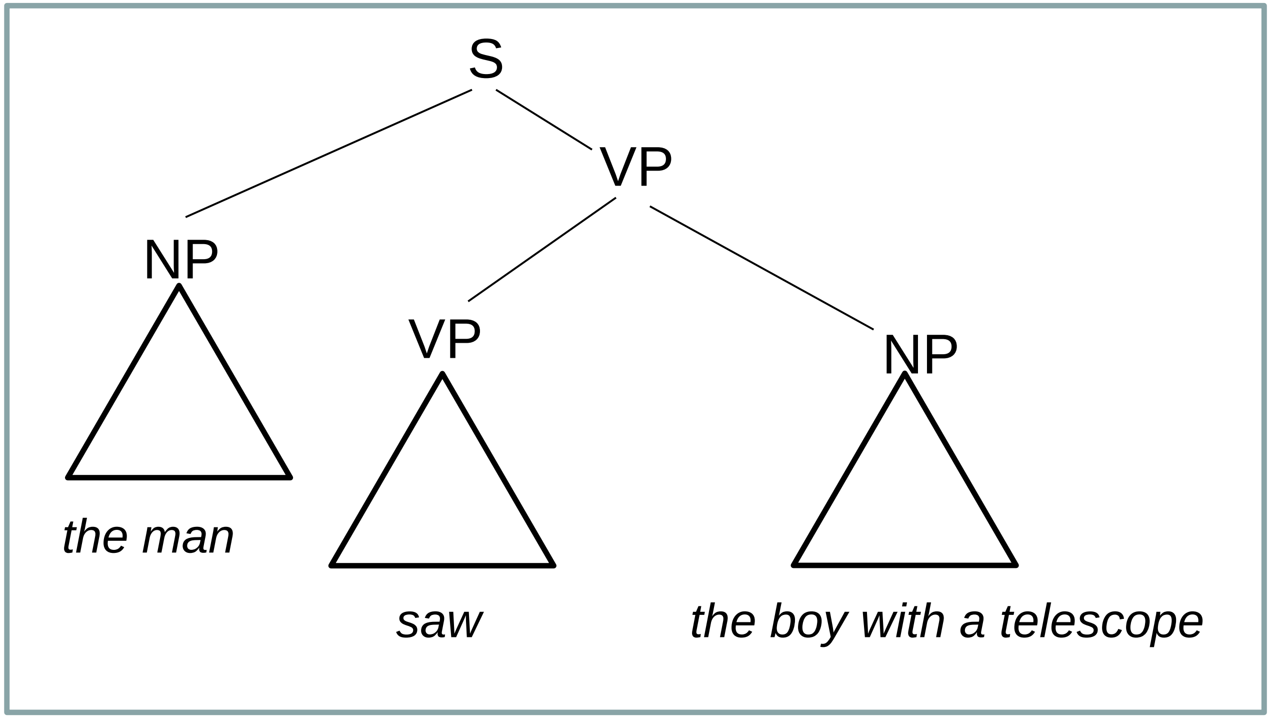
Positions (row-col)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0	*the*$_{01}$ DT_{01} 0.5	NP_{02} 0.0825	S_{03} 1.0*0.0825*0.2=0.0165	--	S_{05} 1.0*0.0825*0.00825=0.0007	--	--	S_{08} 0.0825*0.0002,0.0825*0.00034=0.000017, 0.000028
1		*man*$_{12}$ NN_{12} 0.33	--	--	--	--	--	--
2			*saw*$_{23}$ VBD_{23} VP_{23} 1.0*0.2=0.2	--	VP_{25} 0.5*0.2*0.0825= 0.00825	--	--	VP_{28} VP_{25} PP_{58}/VP_{23} NP_{38} 0.3*0.00825*0.0825,0.5*0.2*0.0034=0.0002/0.00034
3				*the*$_{34}$ DT_{34} 0.5	NP_{35} 0.0825	--	--	NP_{38} 0.5*0.0825*0.0825=0.0034
4					*boy*$_{45}$ NN_{45} 0.33	--	--	--
5						*with*$_{56}$ P_{56} 1.0	--	PP_{58} 1.0*0.0.0825=0.0825
6							*a*$_{67}$ DT_{67} 0.5	NP_{68} 0.5*0.165=0.0825
7								*telescope*$_{78}$ NN_{78} 0.33

Parse-1



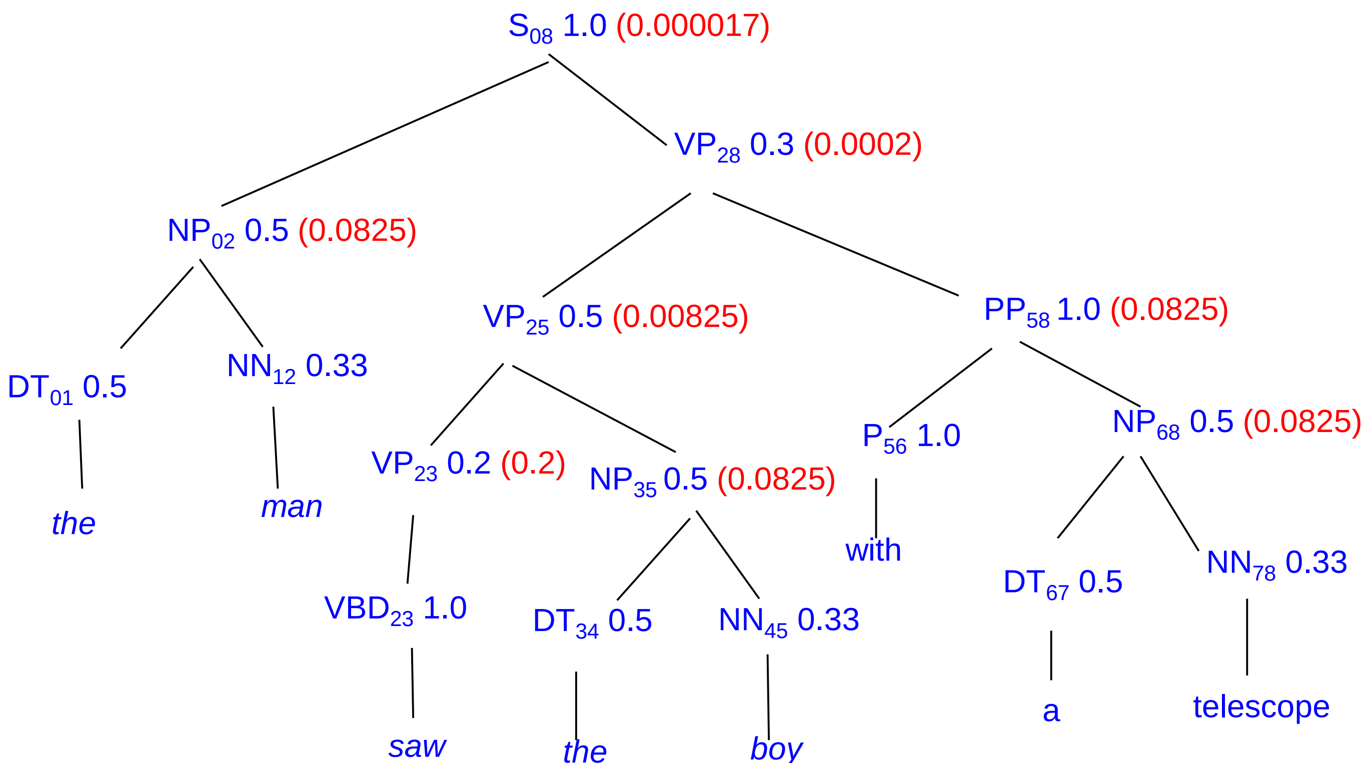
Constituency parse tree of *"The man saw the boy with a telescope"*

Parse-2



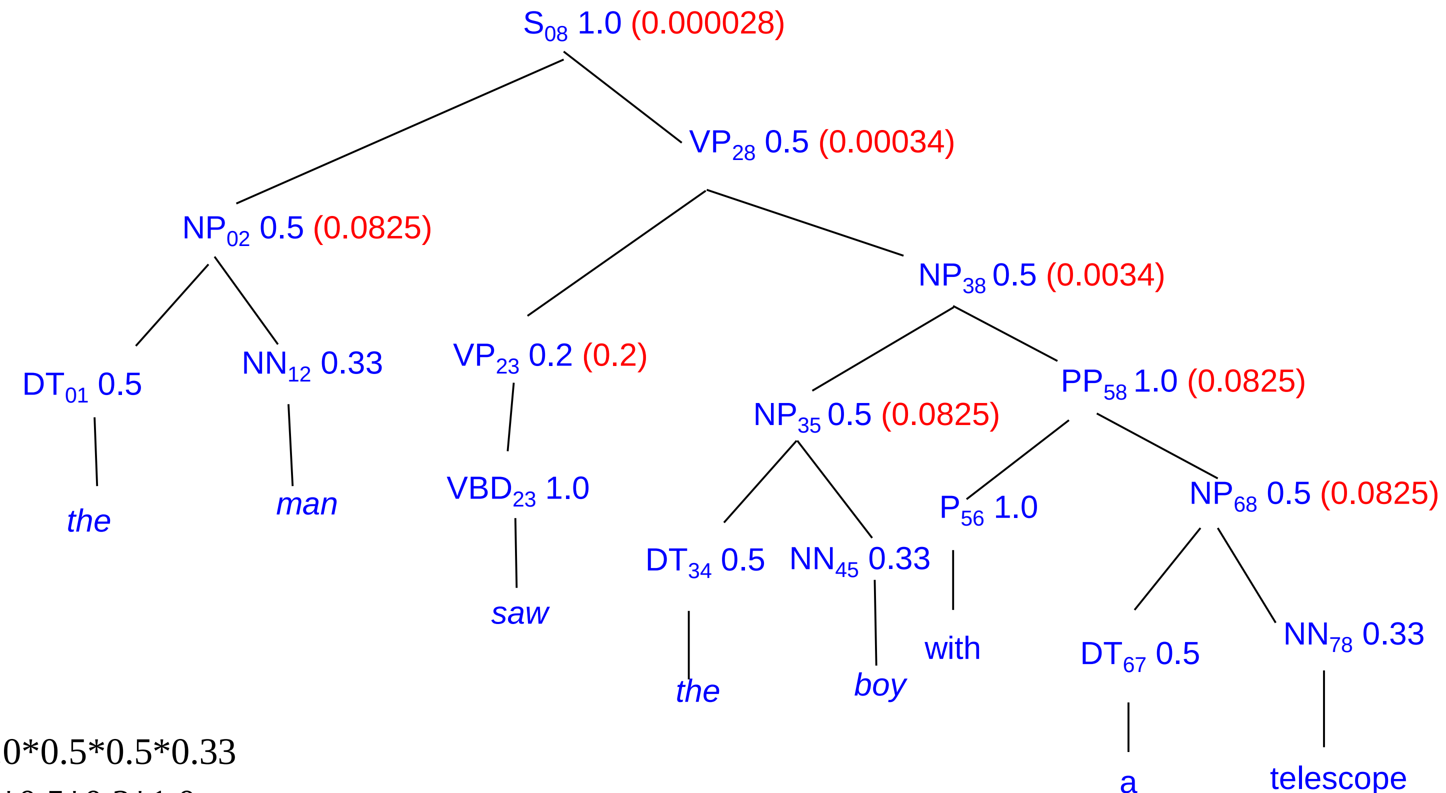
Constituency parse tree of *"The man saw the boy with a telescope"*

Detailed constituency tree of "*the man…*": Parse Tree-1 (man has telescope) with probability



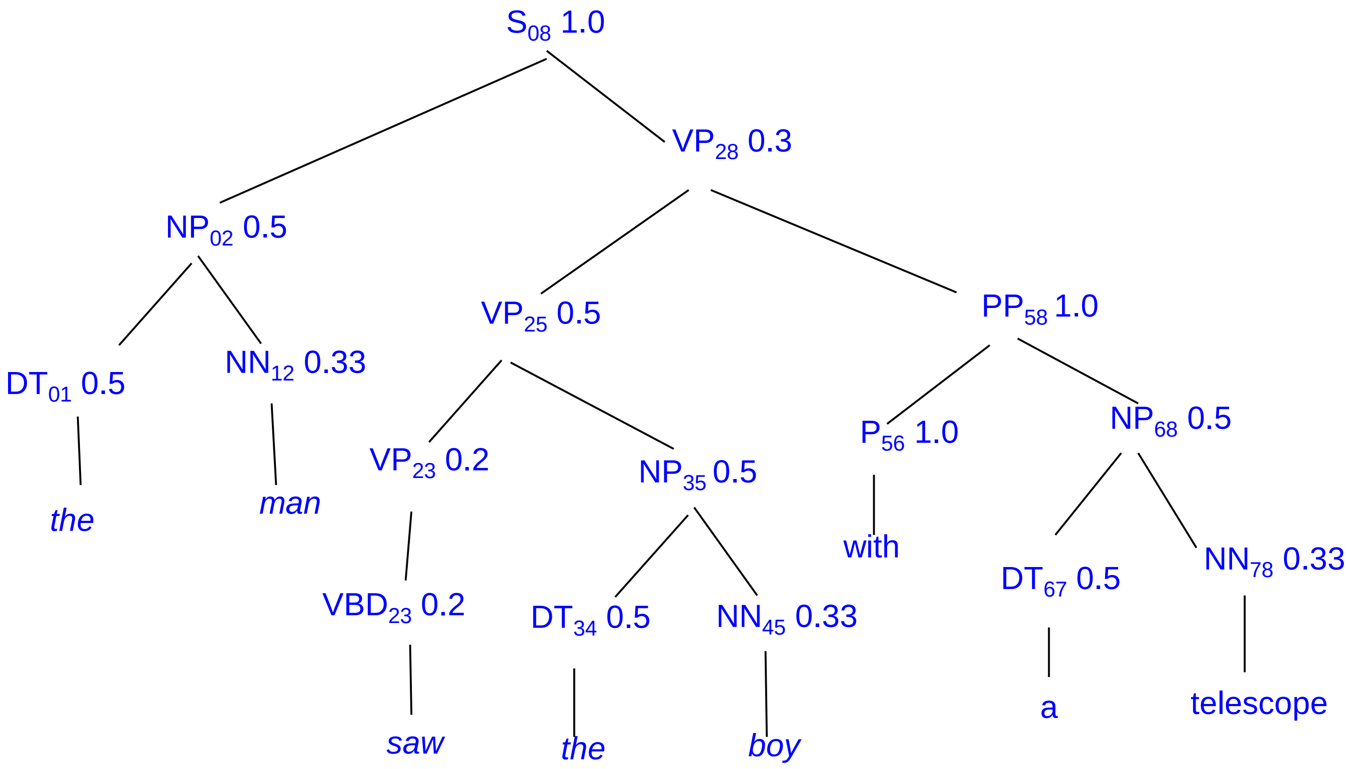
$P(tree\text{-}1)$ = 1.0*0.5*0.5* 0.33
*0.3*0.5*0.2*1.0
* 0.5*0.5
* 0.33
*1.0*1.0
* 0.5*0.5
* 0.33
= $1.7*10^{-5}$

Detailed constituency tree of "*the man…*": Parse Tree-2 (boy has telescope)

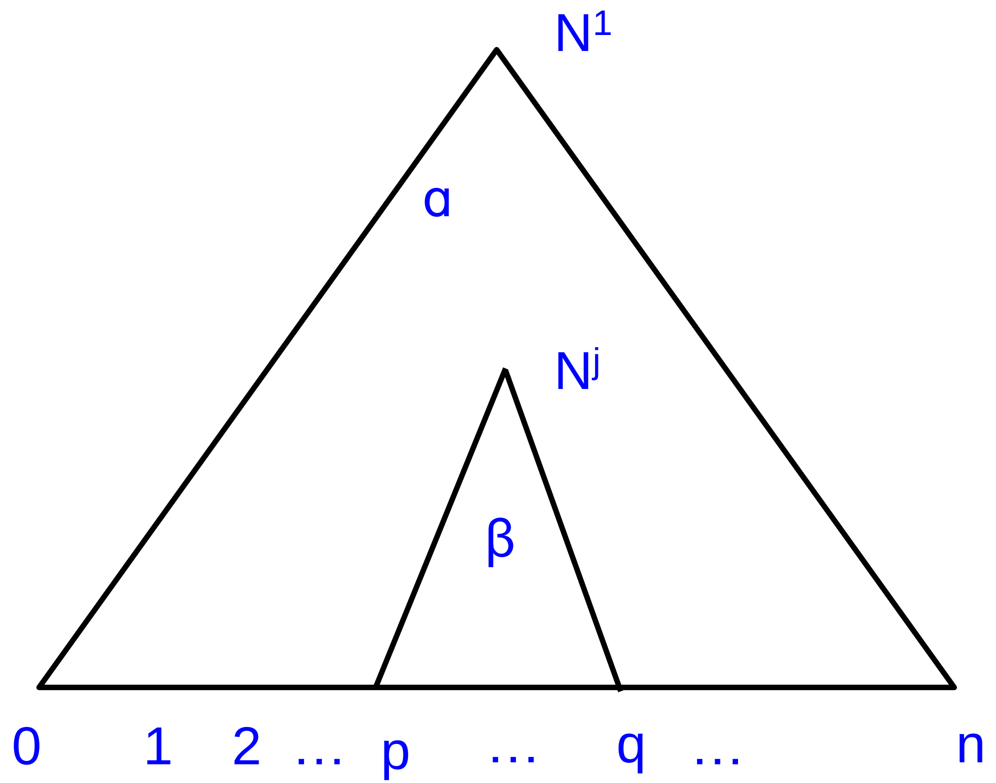


$P(tree\text{-}2)$ =1.0*0.5*0.5*0.33
*0.5*0.2*1.0
*0.5*0.5*0.5
*0.33
*1.0*1.0
*0.5*0.5
*0.33
$= 2.8\times10^{-5}$

Detailed constituency tree of "*the man…*": Parse Tree-2 (boy has telescope)



Alpha and beta probabilities



$\beta_j(p,q)$

$$\beta_j(p,q) = P(W_{pq} \mid N_{pq}^j)$$

$\beta_j(p,q)$ expansion

$$\beta_j(p,q) = P(W_{p-q} \mid N^j_{pq})$$

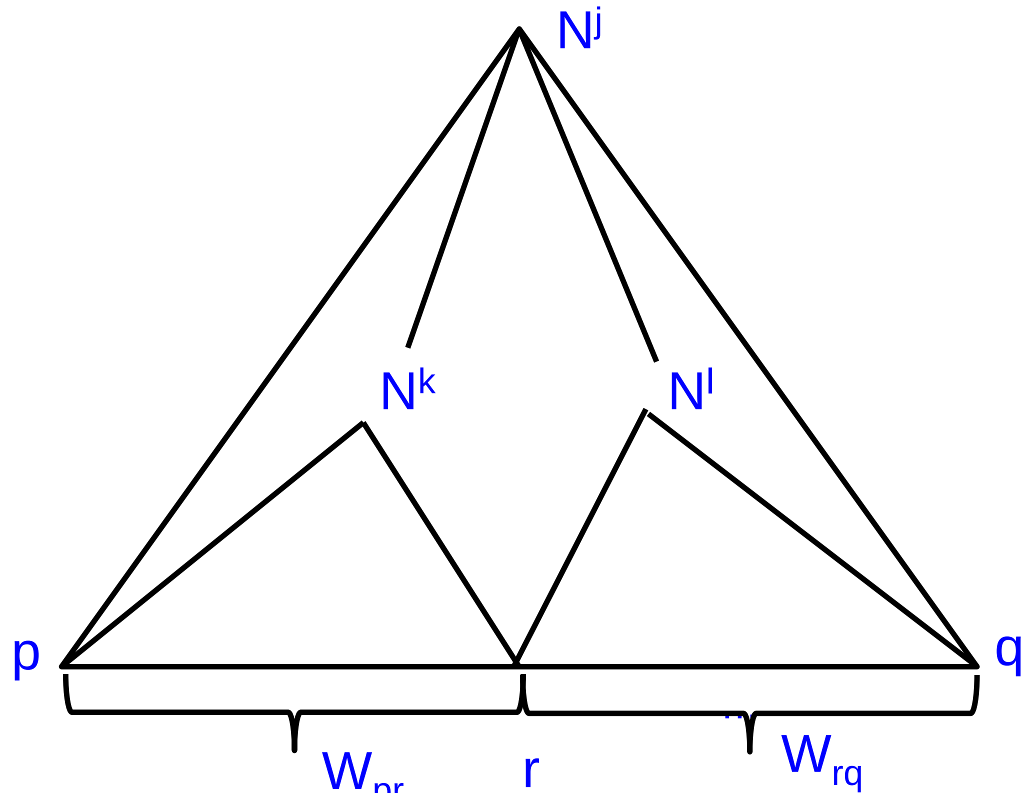
$$= \sum_{k,r,l} P(N^j \to N^k\ N^l).P(W_{p-r} \mid N^k_{pr}).P(W_{r-q} \mid N^l_{rq})$$

$$= \sum_{k,r,l} P(N^j \to N^k\ N^l).\beta_k(p,r).\beta_l(r,q)$$

$\beta_j(k,k+1)$ expansion

$$\beta_j(k,k+1) = P(w_{k,k+1} \mid N^j_{k,k+1})$$
$$= P(N^j \rightarrow w)$$

N^j, N^k, N^l



delta probability

$$\delta_i(p,q) = \max_{j,r,k} P(N^i \to N^j\ N^k).\delta_j(p,r).\delta_k(r,q)$$

Illustration of Parse Tree finding algo

$_0$ *the* $_1$ *man* $_2$ *saw* $_3$ *the* $_4$ *boy* $_5$ *with* $_6$ *a* $_7$ *telescope* $_8$

Positions (row-col)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0	*the*$_{01}$ DT$_{01}$ 0.5	NP$_{02}$ 0.0825	S$_{03}$ 1.0*0.08*0.2=0.0165	--	S$_{05}$	--	--	S$_{08}$ 0.0825*0.0002/0.825*0.004=0.0000165/0.00033
1		*man*$_{12}$ NN$_{12}$ 0.33	--	--	--	--	--	--
2			*saw*$_{23}$ VBD$_{23}$ VP$_{23}$ 1.0*0.2=0.2	--	VP$_{25}$ 0.5*0.2*0.0825 =0.00825	--	--	VP$_{28}$ VP$_{25}$ PP$_{58}$/VP$_{23}$ NP$_{38}$ 0.3*0.00825*0.0825/0.5*0.2*0.04125=0.0002/0.004
3				*the*$_{34}$ DT$_{34}$ 0.5	NP$_{35}$ 0.0825	--	--	NP$_{38}$ 0.5*0.0825=0.04125
4					*boy*$_{45}$ NN$_{45}$ 0.33	--	--	--
5						*with*$_{56}$ P$_{56}$ 1.0	--	PP$_{58}$ 1.0*0.0.0825=0.0825
6							*a*$_{67}$ DT$_{67}$ 0.5	NP$_{68}$ 0.5*0.165=0.0825
7								*telescope*$_{78}$ NN$_{78}$ 0.33

Illustration of the algo for finding the parse tree

- ‘the’ is reduced to *DT*; this is recorded in cell <0, 1> along with the probability 0.5
- Similarly for ‘man’ → *NN* along with the probability 0.33 at cell <1, 2>
- Cell <0, 2> gets NP_{02} along with the probability 0.0825

Illustration of the algo for finding the parse tree

- This process continues normally, until we come to cell <2, 8>; VP28 has two parse trees one through *VP* → *VP PP* with probability 0.0002 and the other through *VP* → *VP NP* with probability 0.00034
- Here we choose the parse tree through *VP NP* which has the higher probability
- The process terminates with parse tree rooted at *S*, having probability 0.000028 (*PP* attached to *NP*, indicating the boy has the telescope)

Efficiency considerations

Catalan number

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \prod_{k=2}^{n} \frac{k}{n+k}, \quad n \geq 0$$

which is exponential in n.

- It can be shown that the number of full binary trees with n + 1 leaves, or, equivalently, with a total of n internal nodes is C_n.
- Binary trees are obtained from X-bar grammars.
- They are also obtained when we apply CYK parsing since the grammar is in Chomsky normal form (CNF).
- Thus, efficiency concern is a real one for parsing

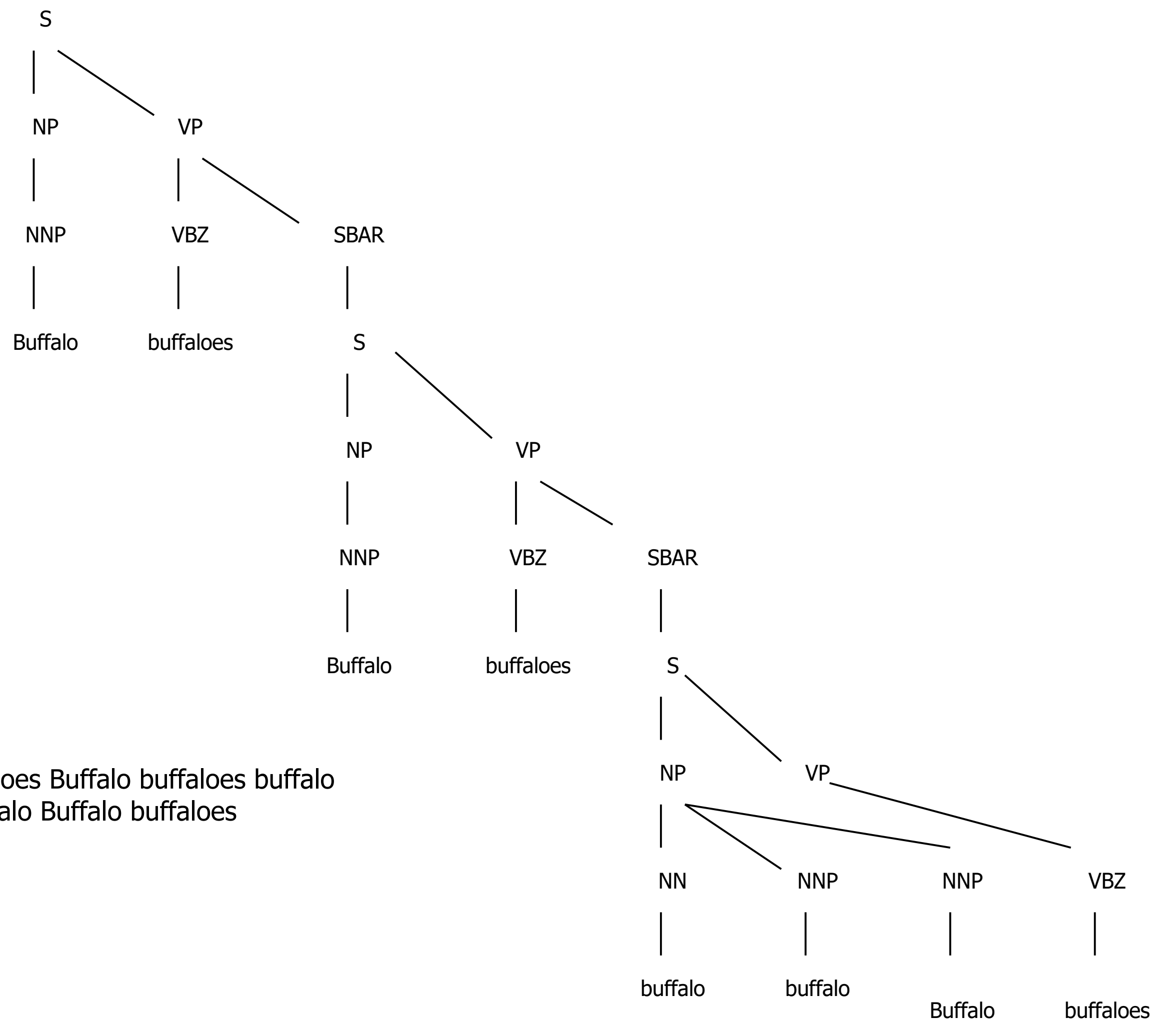
Stress Test for Parsing:
A very difficult parsing situation!

Repeated Word handling

Sentence on Buffaloes!

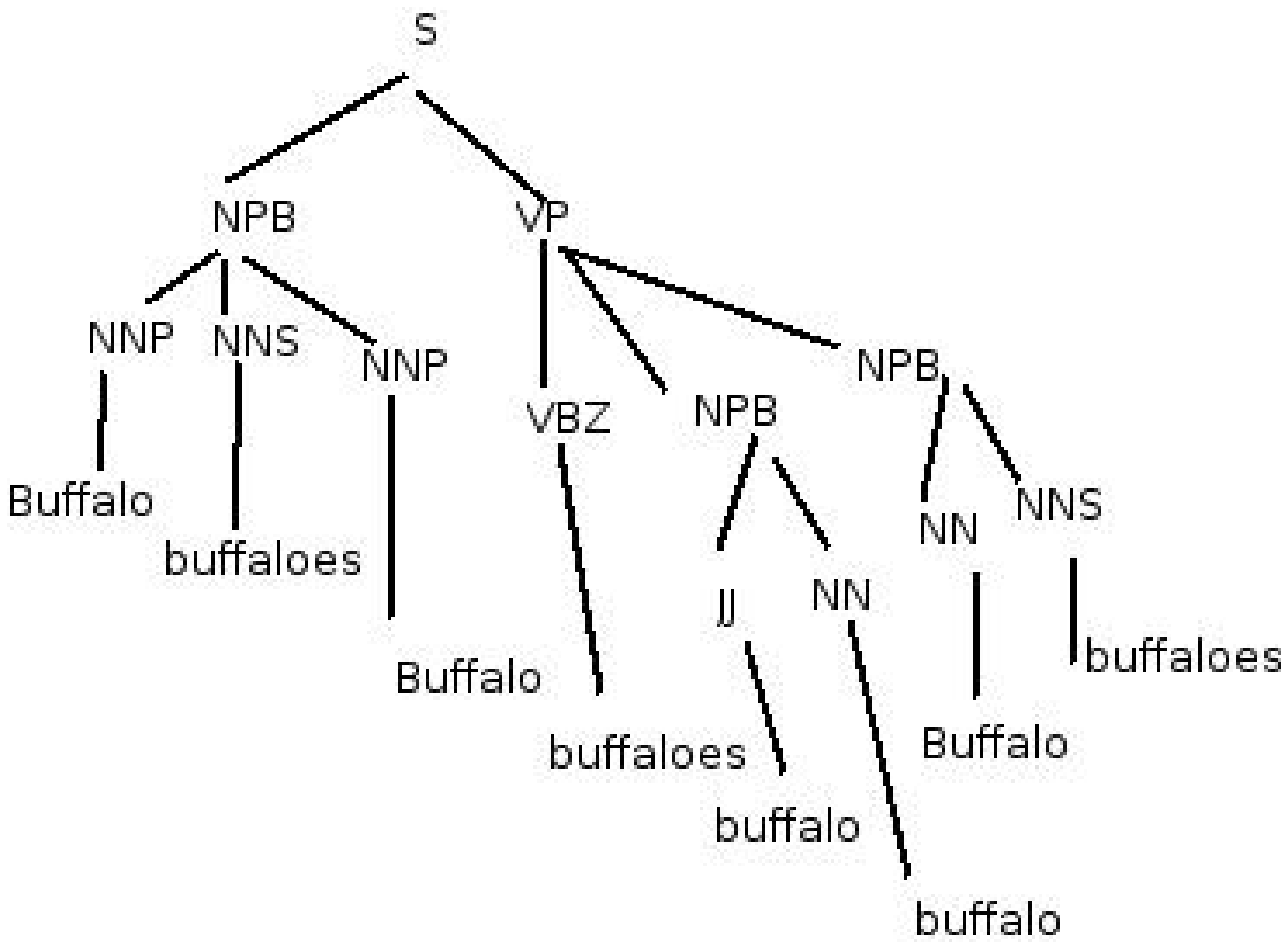
Buffaloe buffaloes Buffaloe buffaloes buffaloe buffaloe Buffaloe buffaloes

Charniak

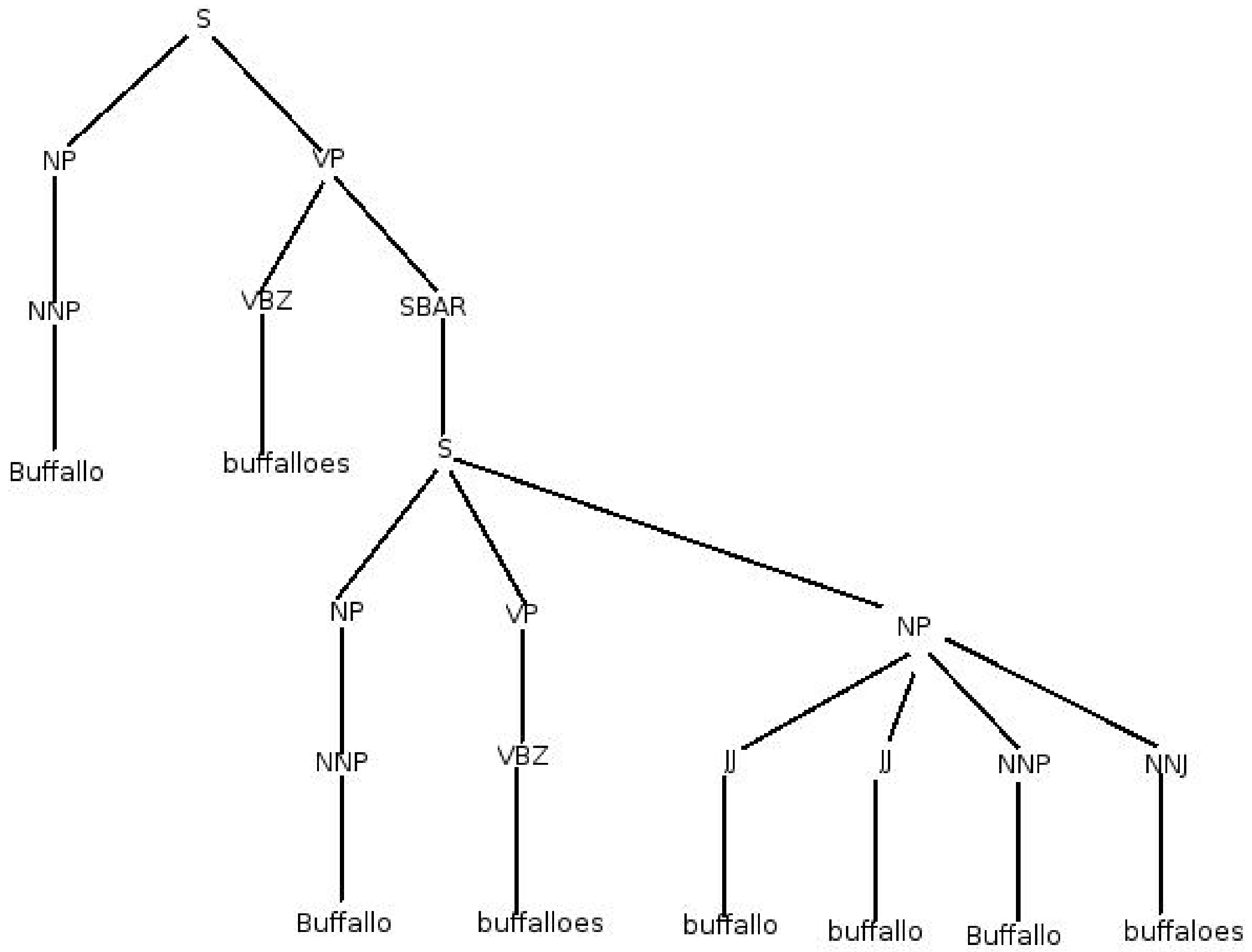


Buffalo buffaloes Buffalo buffaloes buffalo buffalo Buffalo buffaloes

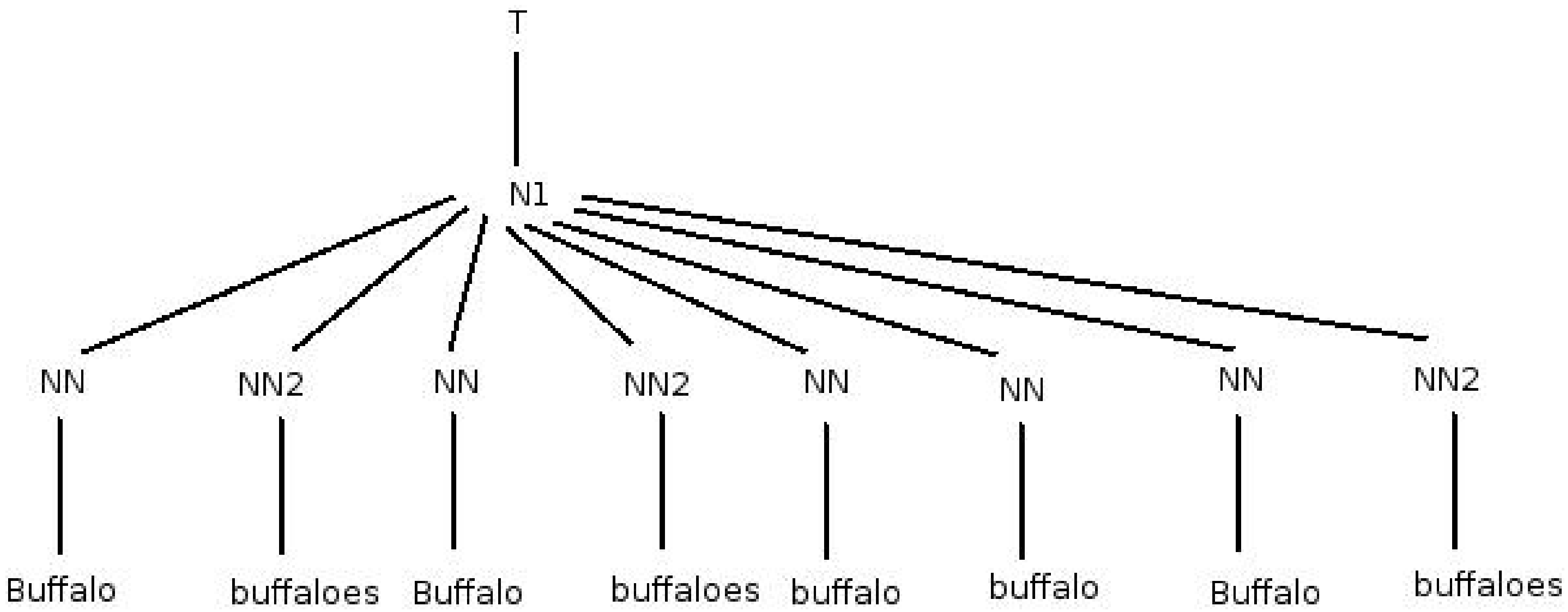
Collins


S
NPB
VP
NNP
NNS
NNP
VBZ
NPB
NPB
Buffalo
buffaloes
Buffalo
buffaloes
JJ
NN
NN
NNS
buffalo
buffalo
Buffalo
buffaloes

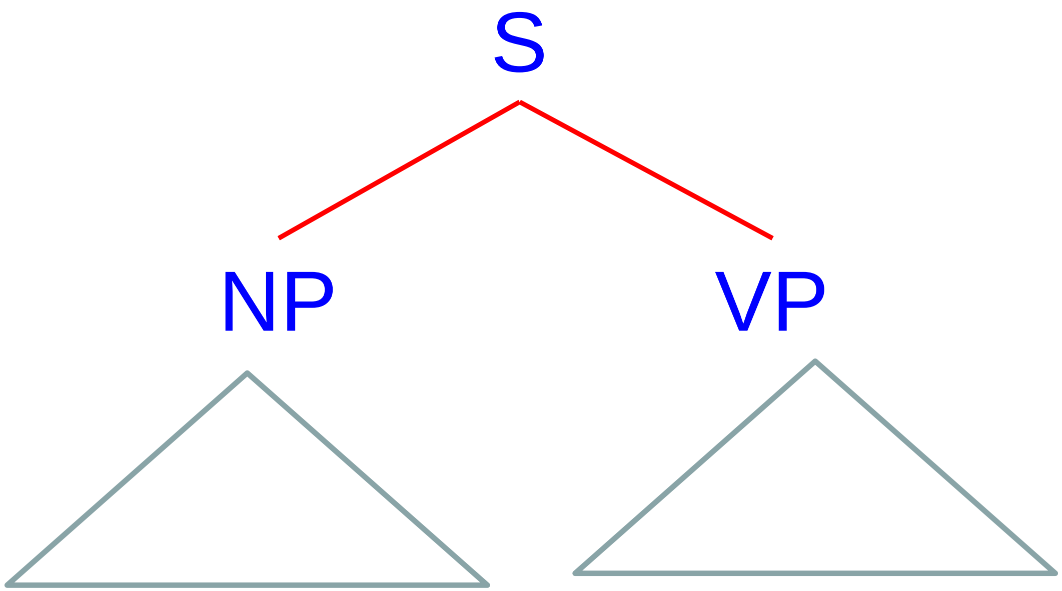
Stanford


S
NP
VP
NNP
VBZ
SBAR
Buffallo
buffalloes
S
NP
VP
NP
NNP
VBZ
JJ
JJ
NNP
NNJ
Buffallo
buffalloes
buffallo
buffallo
Buffallo
buffaloes

RASP


T
N1
NN
NN2
NN
NN2
NN
NN
NN
NN2
Buffalo
buffaloes
Buffalo
buffaloes
buffalo
buffalo
Buffalo
buffaloes

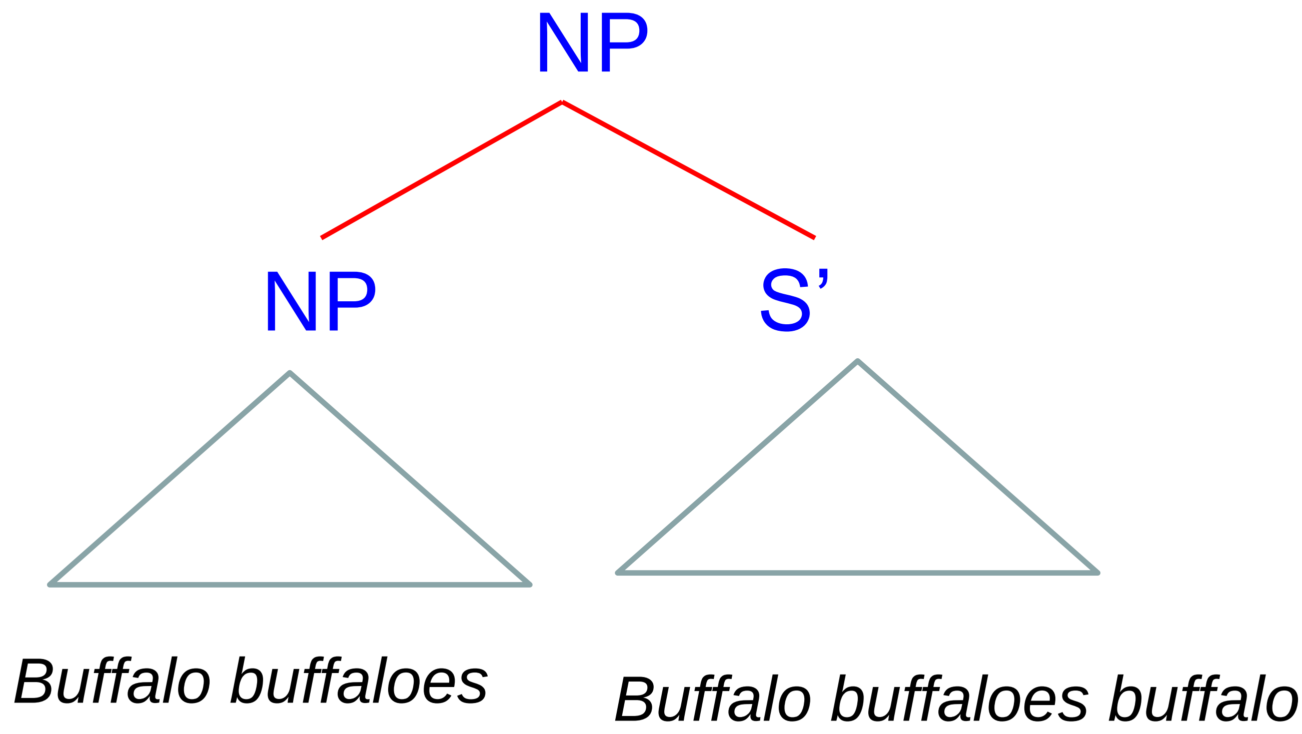
S: Buffalo buffaloes Buffalo buffaloes buffalo buffalo Buffalo buffaloes



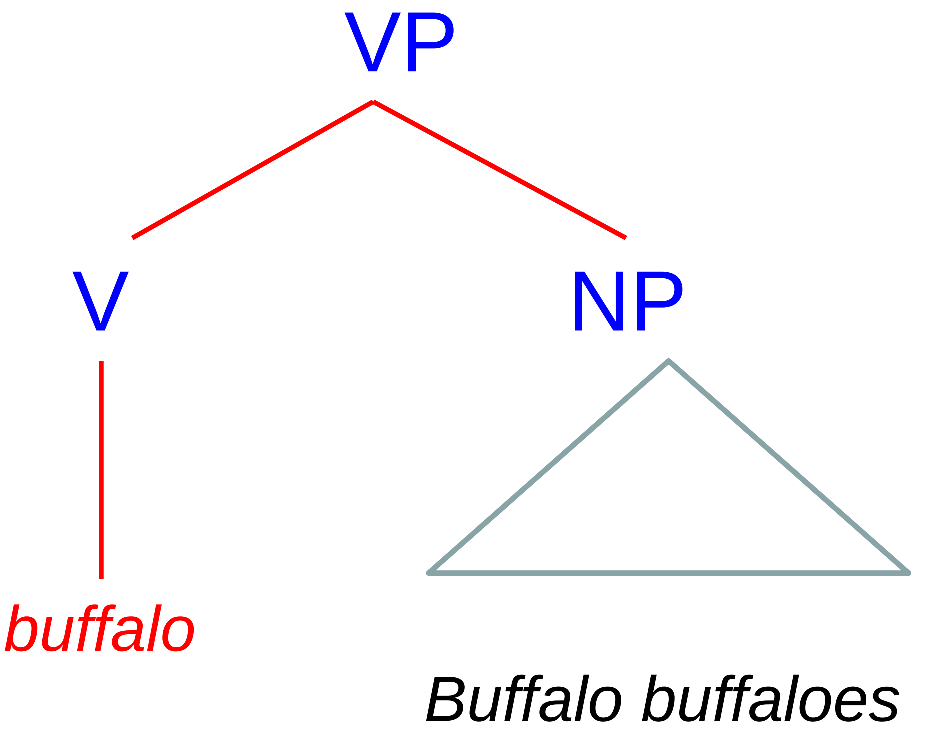
Buffalo buffaloes Buffalo buffaloes buffalo

buffalo Buffalo buffaloes

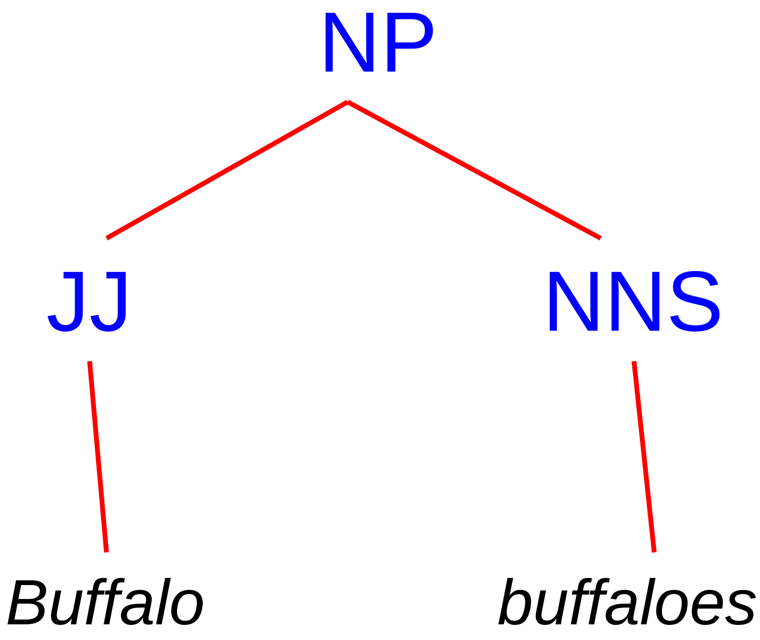
NP: Buffalo buffaloes Buffalo buffaloes buffalo



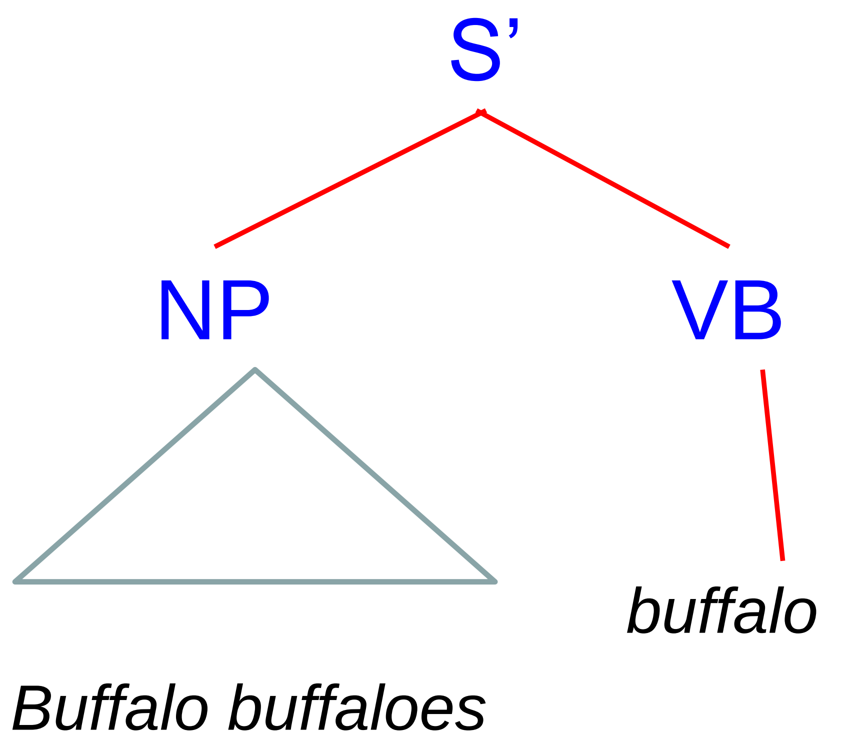
VP: buffalo Buffalo buffaloes



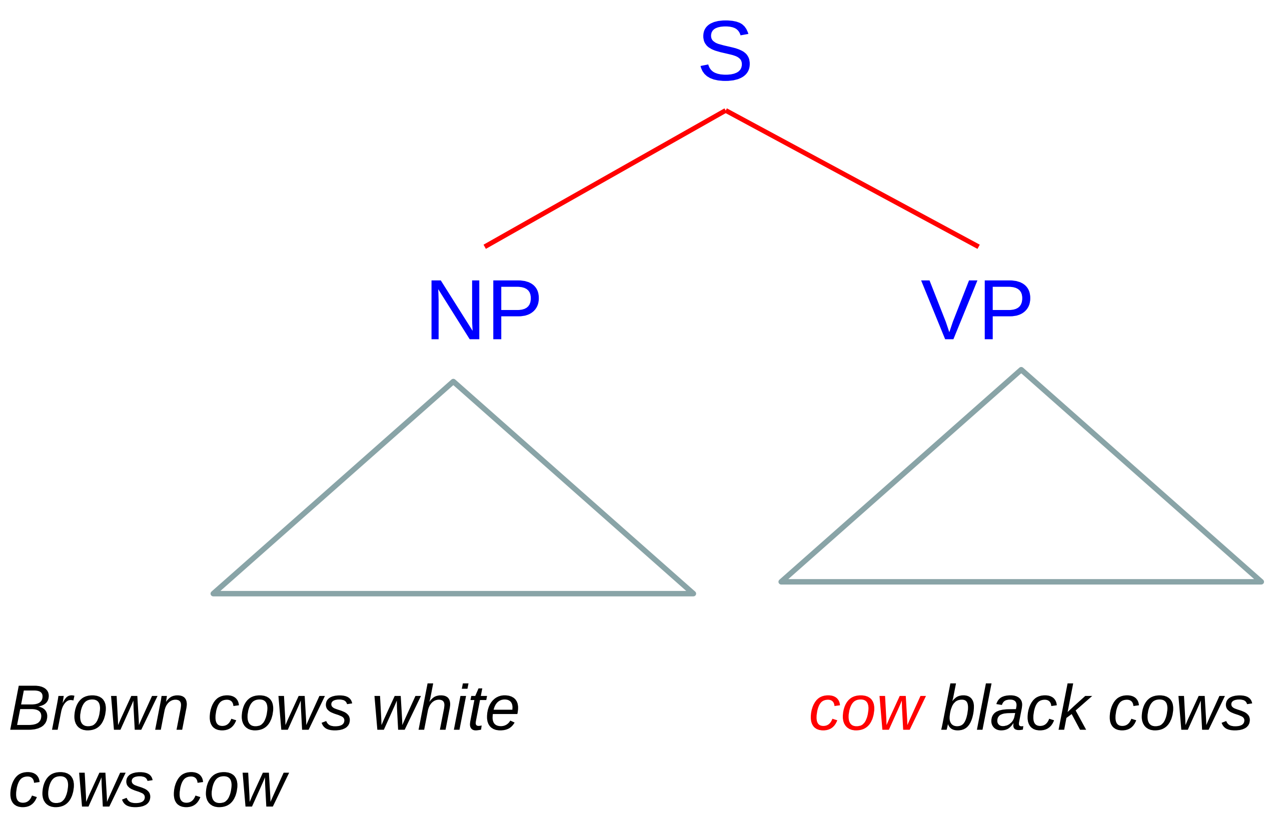
NP: Buffalo buffaloes



S’: Buffalo buffaloes buffalo



Another similar sentence: Brown cows white cows cow cow black cows

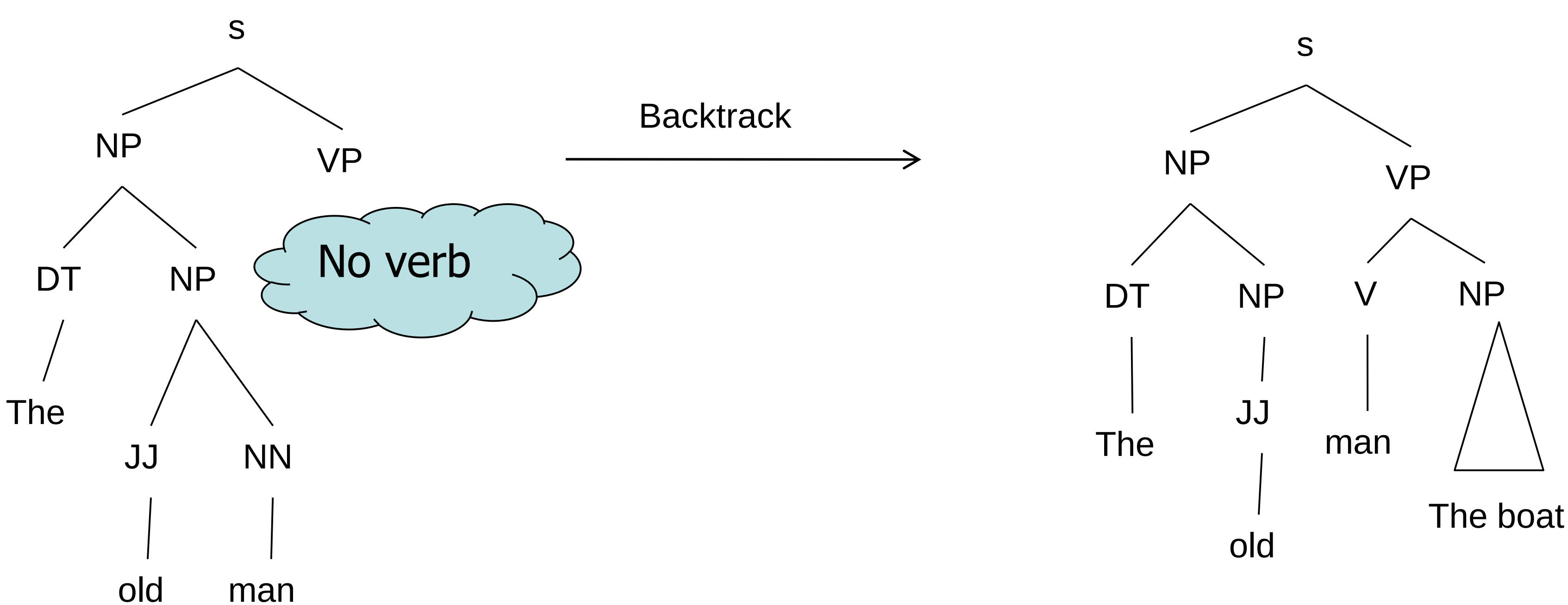


Observation

- Collins and Charniak come close to producing the correct parse.
- RASP tags all the words as nouns.

Another phenomenon: Garden pathing

e.g. The old man the boat.



Another example: The horse raced past the garden fell.

Expectation Maximization

From

Pushpak Bhattacharyya, *Machine Translation*, CRC Press, 2015

Maximum Likelihood of Observations

- ***Situation 1: Throw of a Single Coin***
- The parameter is the probability *p* of getting heads in a single toss. Let *N* be the number of tosses. Then the observation *X* and the data or observation likelihood *D* respectively are:

$$X :< x_1, x_2, x_3, \ldots, x_{N-1}, x_N >$$

$$D = \prod_{i=1}^{N} p^{x_i} (1-p)^{1-x_i}, \text{ s.t. } \mathrm{x_i} = 1 \text{ or } 0, \text{ and } 0 \leq \mathrm{p} \leq 1$$

where x_i is an indicator variable assuming values 1 or 0 depending on the *ith* observation being heads or tail. Since there are *N* identically and independently distributed (*i.i.d.*) observations, *D* is the product of probabilities of individual observations each of which is a Bernoulli trial.

Single coin

Since exponents are difficult to manipulate mathematically, we take log of *D*, also called log likelihood of data, and maximize with regard to *p*. This yields

$$p = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{M}{N}; \; M = \# Heads, N = \# tosses$$

Throw of 2 coins

- Three parameters: probabilities p_1 and p_2 of heads of the two coins and the probability p of choosing the first coin (automatically, $1\text{-}p$ is the probability of choosing the second coin).
- N tosses and observations of heads and tails. Only, we do not know which observation comes from which coin.
- Indicator variable z_i is introduced to capture coin choice ($z_i=1$ if coin 1 is chosen, else 0). This variable is hidden, *i.e.*, we do not know its values.
- However, without it the likelihood expression would have been very cumbersome.

Data Likelihood

Data Likelihood,

$D \quad = P_{<p1,p2,p>}(X) = P_\theta(X),\ \theta = <p, p_1, p_2>$

$\quad\; = \Sigma_Z P_\theta(X,Z))$

$$X : < x_1, x_2, x_3, \ldots, x_{N-1}, x_N >$$

$$Z : < z_1, z_2, z_3, \ldots, z_{N-1}, z_N >$$

$$P_\theta(X, Z) = \prod_{i=1}^{N} \left[\left(p p_1^{x_i} (1 - p_1)^{1-x_i} \right)^{z_i} \left((1-p) p_2^{x_i} (1 - p_2)^{1-x_i} \right)^{1-z_i} \right],$$

$$\mathrm{s.t.}\ \mathrm{z_i}, \mathrm{x_i} = 1 \text{ or } 0, \text{ and } 0 \le \mathrm{p}, \mathrm{p_1}, \mathrm{p_2} \le 1$$

Invoke Jensen Inequality

We would like to work with *log*$P_\theta(X)$. However, there will be a Σ inside *log*. Fortunately, *log* is a concave function, so that

$$\log\left(\sum_{i=1}^{K} \lambda_i y_i\right) \geq \left(\sum_{i=1}^{K} \lambda_i \log(y_i)\right); \sum_{i=1}^{K} \lambda_i = 1$$

Log likelihood of Data

$LL(D)$= log likelihood of data

$$= log(P_\theta(X)) = log(\Sigma_Z P_\theta(X,Z))$$

$$= log[\Sigma_Z \lambda_Z (P_\theta(X,Z)/\lambda_Z)]; \Sigma_Z \lambda_Z = 1$$

$$>= \Sigma_Z[\lambda_Z log[(P_\theta(X,Z)/\lambda_Z)]$$

After a number of intricate mathematical steps

$LL(D) >= E_{Z|X,\theta} log(P_\theta(X,Z))$, where $E(.)$ is the expectation function; note that the expectation is conditional on X.

Expectation of log likelihood

$$E_{Z|X}[\log(P_\theta(X,Z)]$$

$$= E_{Z|X}\left[\log\prod_{i=1}^{N}\left[\left(pp_1{}^{x_i}(1-p_1)^{1-x_i}\right)^{z_i}\left((1-p)p_2{}^{x_i}(1-p_2)^{1-x_i}\right)^{1-z_i}\right],\right]$$

$$= E_{Z|X}\left[\begin{array}{l}\displaystyle\sum_{i=}^{N} z_i\left(\log p + x_i\log p_1 + (1-x_i)\log\left(1-p_1\right)\right) + \\ (1-z_i)\left(\log(1-p) + x_i\log p_2 + (1-x_i)\log\left(1-p_2\right)\right)\end{array}\right]$$

$$= \sum_{i=1}^{N}\left[\begin{array}{l}E(z_i\,|\,x_i)\left(\log p + x_i\log p_1 + (1-x_i)\log\left(1-p_1\right)\right) + \\ (1-E(z_i\,|\,x_i))\left(\log(1-p) + x_i\log p_2 + (1-x_i)\log\left(1-p_2\right)\right)\end{array}\right]$$

$$\mathrm{s.t.}\ \mathrm{z_i}, \mathrm{x_i} = 1 \text{ or } 0, \text{ and } 0 \le \mathrm{p}, \mathrm{p_1}, \mathrm{p_2} \le 1$$

Derivation of E and M steps for 2 coin problem (1/2)- M step

Take partial derivative of $E_{Z|X,\theta}(.)$ (prev. slide) wrt p, p_1, p_2 and equate to 0.

$$p = \frac{\sum_{i=1}^{N} E(z_i \mid x_i)}{N}$$

$$p_1 = \frac{\sum_{i=1}^{N} E(z_i \mid x_i) x_i}{\sum_{i=1}^{N} E(z_i \mid x_i)}$$

$$p_2 = \frac{M - \sum_{i=1}^{N} E(z_i \mid x_i) x_i}{N - \sum_{i=1}^{N} E(z_i \mid x_i)} ; M = \# Heads, N = \# tosses$$

Derivation of E and M steps for 2 coin problem (2/2)- E step

$E(z_i|x_i)= 1.P(z_i=1|x_i)+0.P(z_i=0|x_i)$

$=P(z_i=1|x_i)$

$$P(z_i = 1 \mid x_i) = \frac{P(z_i = 1, x_i)}{P(x_i)}$$

$$= \frac{pp_1^{x_i}(1-p_1)^{1-x_i}}{P(x_i, z_i = 1) + P(x_i, z_i = 0)}$$

$$= \frac{pp_1^{x_i}(1-p_1)^{1-x_i}}{pp_1^{x_i}(1-p_1)^{1-x_i} + (1-p)p_2^{x_i}(1-p_2)^{1-x_i}}$$