CS626: NLP, Speech and the Web

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

Lecture 17, 18: Probabilistic parsing; parser comparisons

24th and 27th September, 2012

(16th lecture was on UNL by Avishek)

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]]]]]]

[$_{.}$.]]]

Noisy Channel Modeling

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

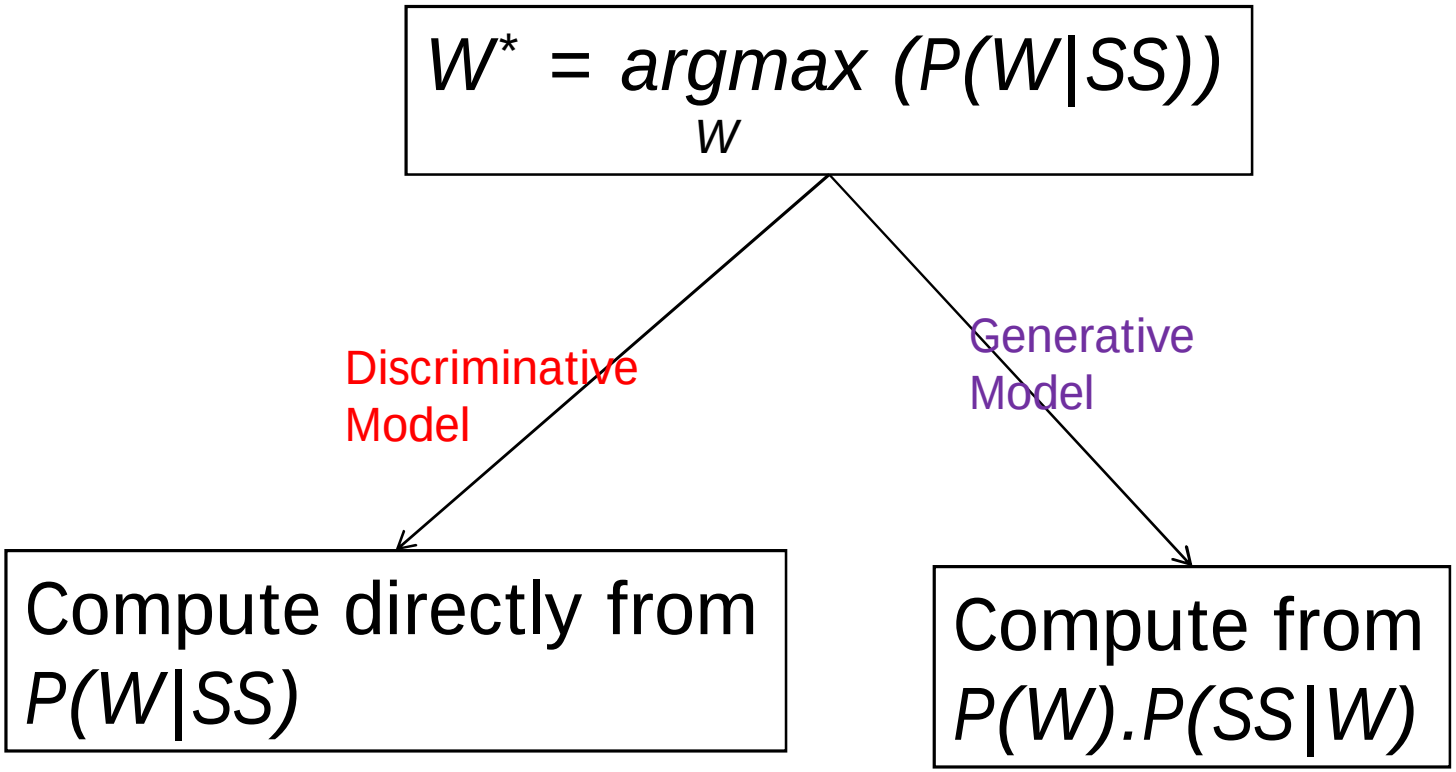
$$= \underset{T}{\operatorname{argmax}}\,[P(T).P(S|T)]$$

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

Corpus

- A collection of text called *corpus*, is used for collecting various language data
- With annotation: more information, but manual labor intensive
- Practice: *label automatically; correct manually*
- The famous *Brown Corpus* contains 1 million tagged words.
- **Switchboard:** very famous corpora 2400 conversations, 543 speakers, many US dialects, annotated with orthography and phonetics

Discriminative vs. Generative Model



Language Models

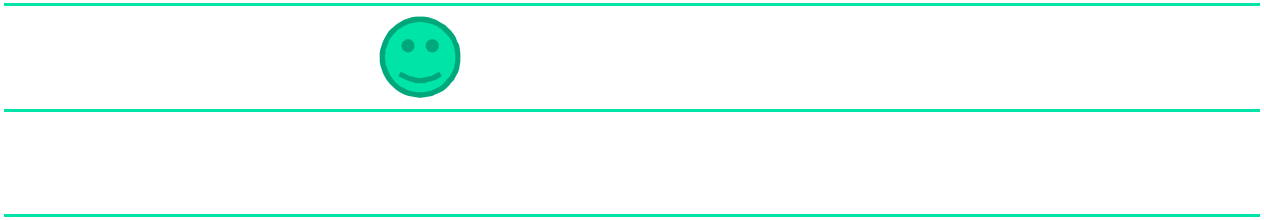
- N-grams: sequence of n consecutive words/characters
- Probabilistic / Stochastic Context Free Grammars:
  - Simple probabilistic models capable of handling recursion
  - A CFG with probabilities attached to rules
  - Rule probabilities → how likely is it that a particular rewrite rule is used?

PCFGs

- Why PCFGs?
  - Intuitive probabilistic models for tree-structured languages
  - Algorithms are extensions of HMM algorithms
  - Better than the n-gram model for language modeling.

Data driven parsing is tied up with what is seen in the training data

- *"Stand right walk left."* Often a message on the airport escalators.

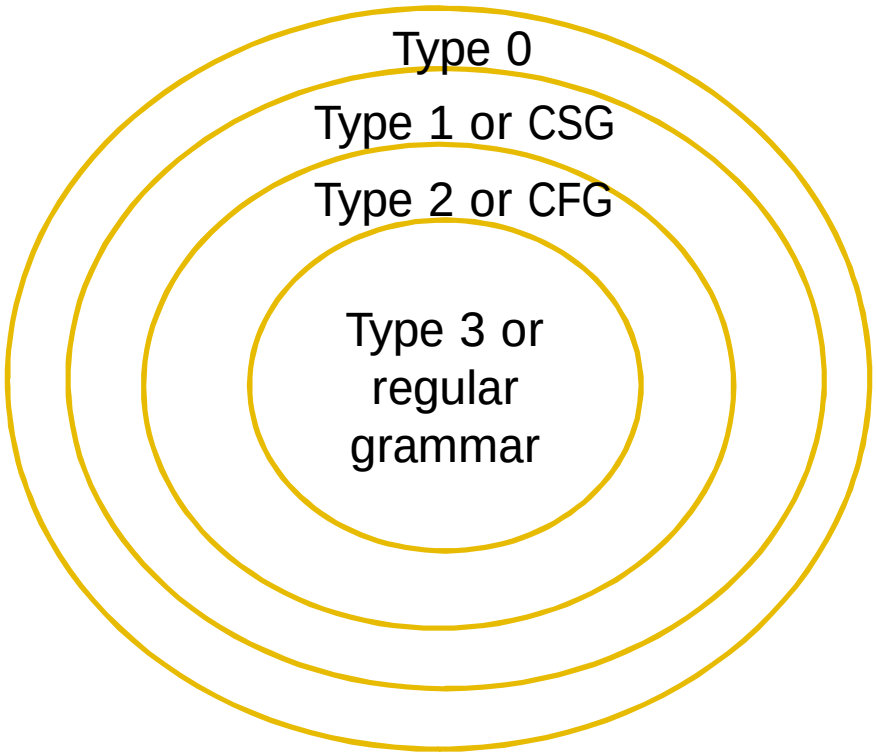


- "Walk left" can come from :- *"After climbing , we realized the walk left was still considerable."*
- "Stand right" can come from: - *"The stand right though it seemed at that time, does not stand the scrutiny now."*
- Then the given sentence "Stand right walk left" will never be parsed correctly.

Chomski Hierarchy

Properties of the hierarchy:-

- *The containment is proper.*
- *Grammar is not learnable even for the lowest type from positive examples alone.*



$$Type\,0 \supset Type\,1 \supset Type\,2 \supset Type\,3$$

Rationalism vs Empiricism

- Scientific paper : - *"The pendulum swings back"* by Ken Church.
- Every 20 years there is a shift from rationalism to empiricism.
- AI approaches are of two types:-
  - Rational (Theory and model driven e.g. Brill Tagger; humans decide the templates)
  - Empirical (Data driven e.g. HMM)
- Is it possible to classify probabilistic CFG into any of these categories?
  - Very difficult to answer.
  - Anything that comes purely from brain can be put into the category of "rationalism".
- There is no pure rationalism or pure empiricism

Laws of machine learning

- “Learning from vaccuum (zero knowledge) is impossible”.
- “Inductive bias decides what to learn”.
  - *In case of POS tagger , bias is that we assume the tagger to be a HMM model.*

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

$$\forall i \ \sum_{\mathrm{i}} \mathrm{P}(\mathrm{N}^i \rightarrow \zeta^j) = 1$$

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

P(NP $\rightarrow$ NN) = 0.5

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

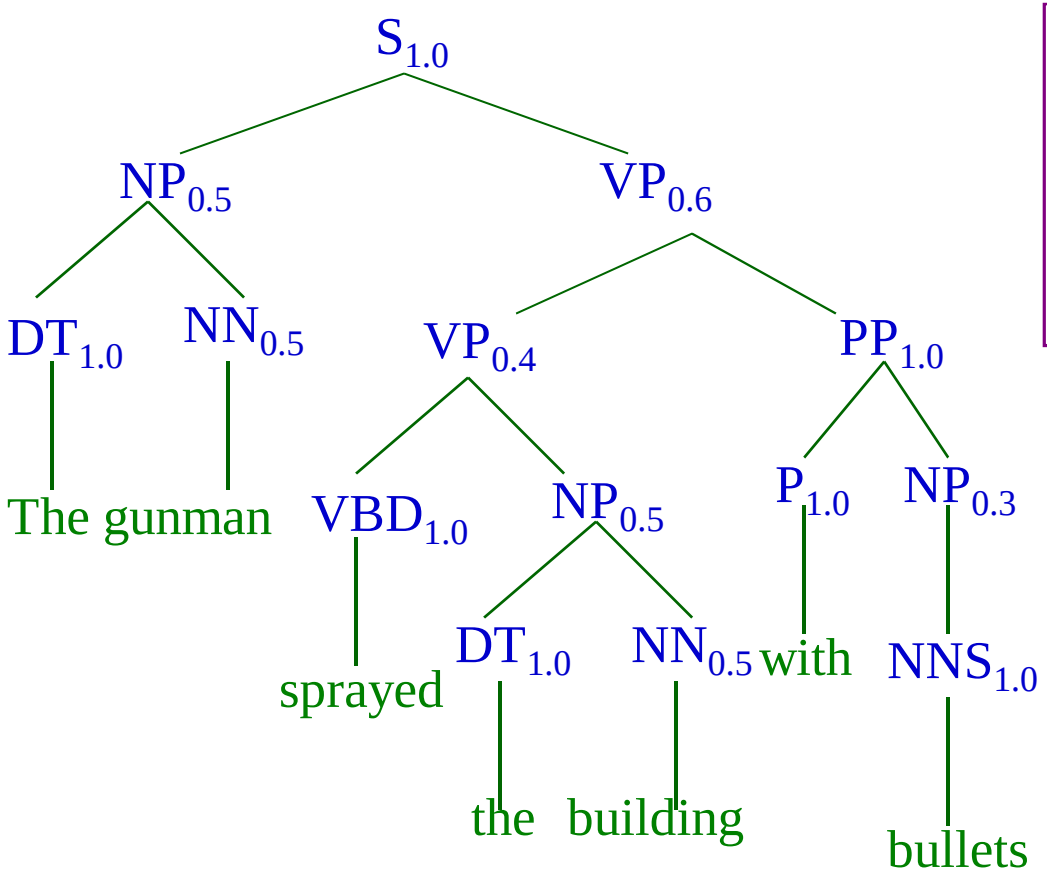
- P(NP $\rightarrow$ DT NN) = 0.2
  - Means 20 % of the training data parses use the rule NP $\rightarrow$ DT NN

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`

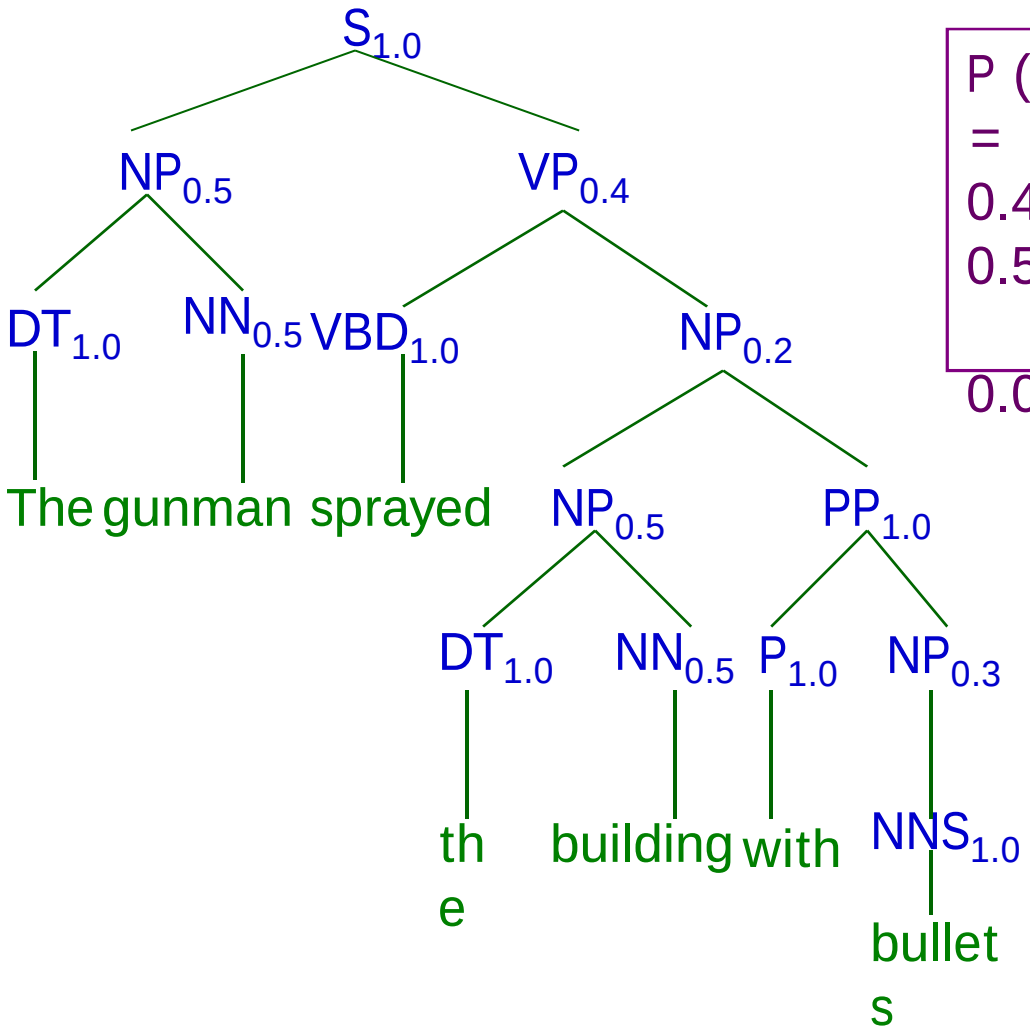
- The gunman sprayed the building with bullets.



$P(t_1)$ = 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.

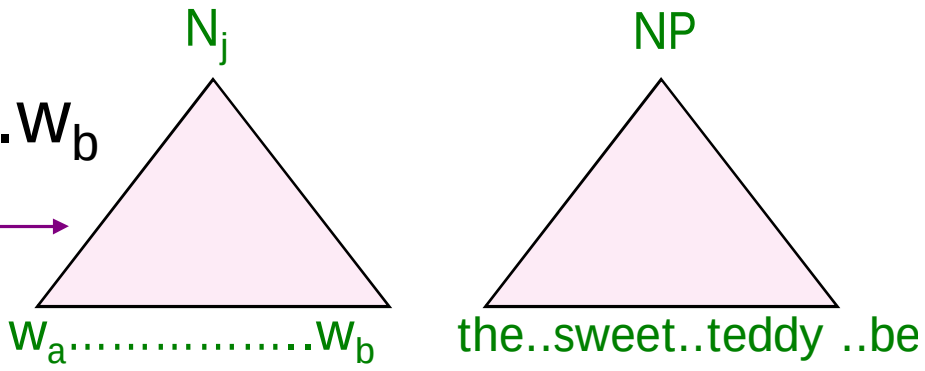


$P(t_2)$
$= 1.0 * 0.5 * 1.0 * 0.5 * 0.4 * 1.0 * 0.2 * 0.5 * 1.0 * 0.5 * 1.0 * 1.0 * 0.3 * 1.0$
$=$
0.0015

Probability of a sentence

- Notation :
  - w_{ab} – subsequence $w_a \ldots w_b$
  - N_j dominates $w_a \ldots w_b$

    or yield(N_j) = $w_a \ldots w_b$



- Probability of a sentence = $P(w_{1m})$

$$P(w_{1m}) = \sum_t P(w_{1m}, t)$$

→ Where t is a parse tree of the sentence

$$= \sum_t P(t) P(w_{1m} \mid t)$$

$$= \sum_{t:yield(t)=w_{1m}} P(t) \qquad \because P(w_{1m} \mid t) = 1$$

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

Assumptions of the PCFG model

- Place invariance :

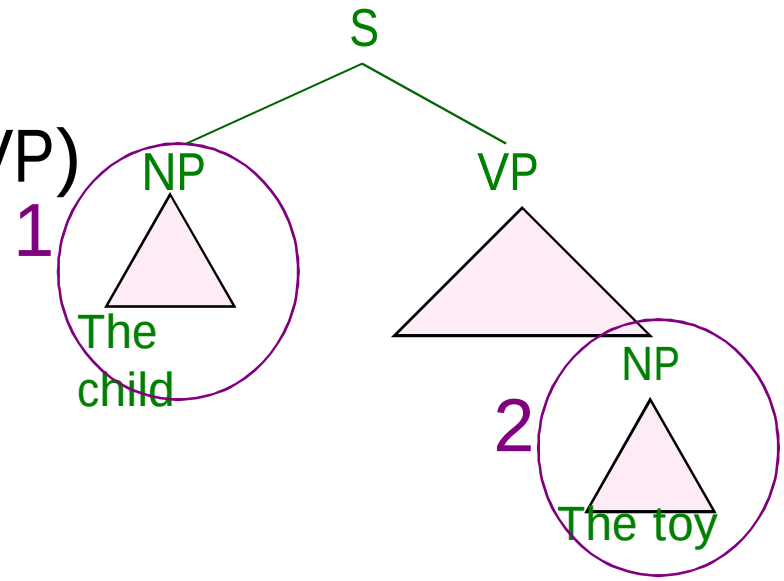
  P(NP $\rightarrow$ DT NN) is same in locations 1 and 2

- Context-free :

  P(NP $\rightarrow$ DT NN | anything outside “The child”)
  = P(NP $\rightarrow$ DT NN)

- Ancestor free : At 2,

  P(NP $\rightarrow$ DT NN|its ancestor is VP)
  = P(NP $\rightarrow$DT NN)

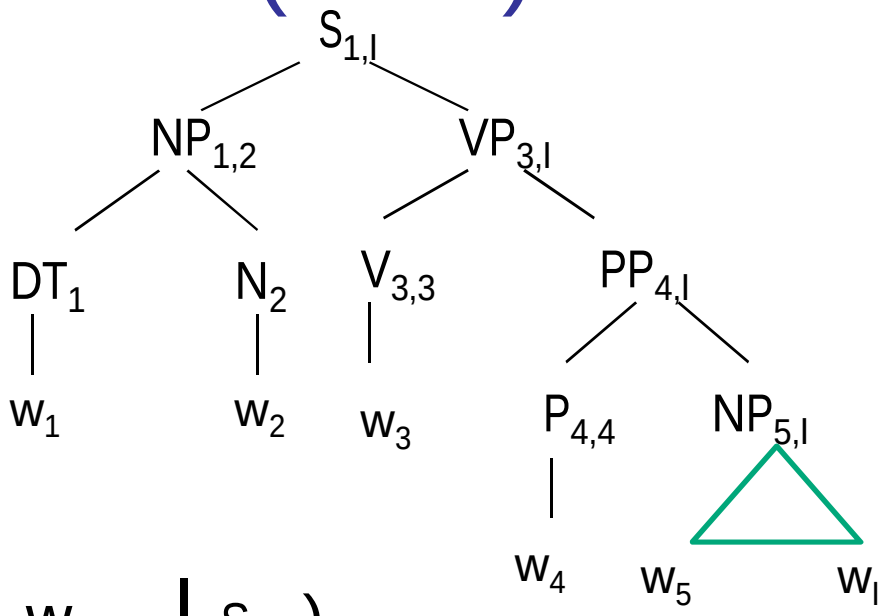


Probability of a parse tree

- *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 $S_{1,l}$ means that the start symbol S dominates the word sequence $W_{1,l}$
- P (t |s) approximately equals joint probability of constituent non-terminals dominating the sentence fragments (next slide)

Probability of a parse tree (cont.)



$P(t|s) = P(t \mid S_{1,l})$

$= P(NP_{1,2}, DT_{1,1}, w_1, N_{2,2}, w_2, VP_{3,l}, V_{3,3}, w_3, PP_{4,l}, P_{4,4}, w_4, NP_{5,l}, w_{5...l} \mid S_{1,l})$

$= P(NP_{1,2}, VP_{3,l} \mid S_{1,l}) * P(DT_{1,1}, N_{2,2} \mid NP_{1,2}) * D(w_1 \mid DT_{1,1}) * P(w_2 \mid N_{2,2}) * P(V_{3,3}, PP_{4,l} \mid VP_{3,l}) * P(w_3 \mid V_{3,3}) * P(P_{4,4}, NP_{5,l} \mid PP_{4,l}) * P(w_4 \mid P_{4,4}) * P(w_{5...l} \mid NP_{5,l})$

(Using Chain Rule, Context Freeness and Ancestor Freeness)

HMM ↔ PCFG

- O observed sequence ↔ w_{1m} sentence
- X state sequence ↔ t parse tree
- μ model ↔ G grammar

- Three fundamental questions

HMM ↔ PCFG

- How likely is a certain observation given the model? ↔ How likely is a sentence given the grammar?

$$P(O \mid \mu) \quad \leftrightarrow \quad P(w_{1m} \mid G)$$

- How to choose a state sequence which best explains the observations? ↔ How to choose a parse which best supports the sentence?

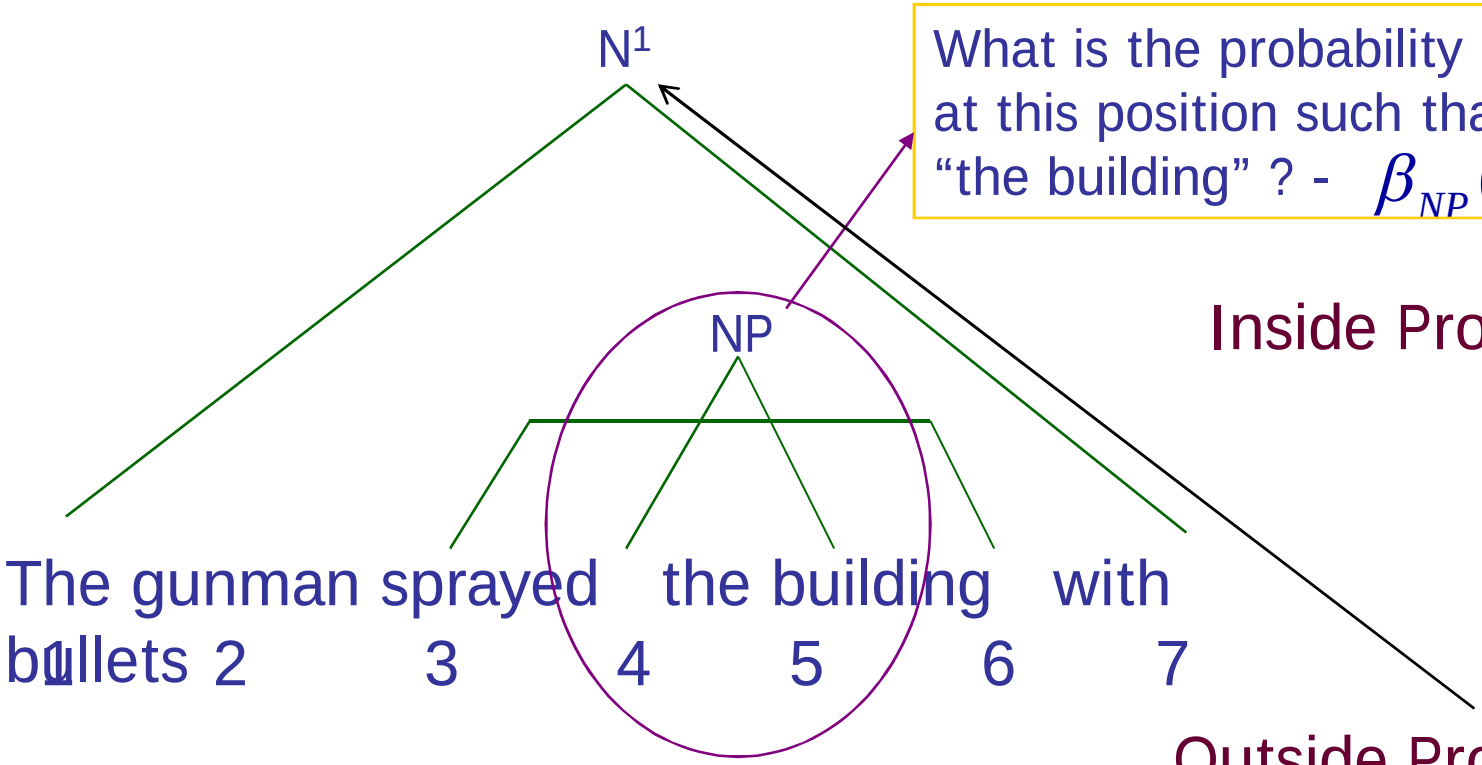
$$\arg\max_{X} P(X \mid O, \mu) \quad \leftrightarrow \arg\max_{t} P(t \mid w_{1m}, G)$$

HMM ↔ PCFG

- How to choose the model parameters that best explain the observed data? ↔ How to choose rule probabilities which maximize the probabilities of the observed sentences?

$$\arg\max_{\mu} P(O \mid \mu) \;\leftrightarrow\; \arg\max_{G} P(w_{1m} \mid G)$$

Interesting Probabilities



What is the probability of having a NP at this position such that it will derive "the building" ? - $\beta_{NP}(4,5)$

Inside Probabilities

Outside Probabilities

What is the probability of starting from N^1 and deriving "The gunman sprayed", a NP and "with bullets" ? - $\alpha_{NP}(4,5)$

Interesting Probabilities

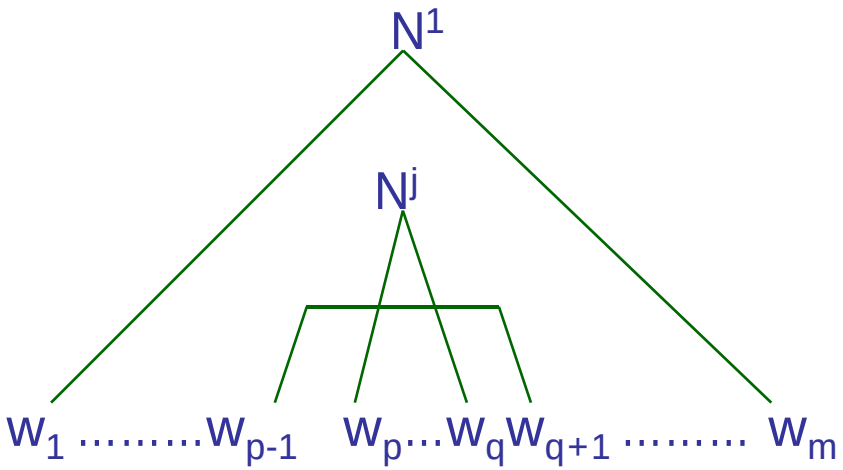
- Random variables to be considered
  - The non-terminal being expanded.
    E.g., NP
  - The word-span covered by the non-terminal.
    *E.g., (*4,5) refers to words “the building”

- While calculating probabilities, consider:
  - The rule to be used for expansion :
    E.g., NP $\rightarrow$ DT NN
  - The probabilities associated with the RHS non-terminals : *E.g.,* DT subtree’s inside/outside probabilities & NN subtree’s inside/outside probabilities

Outside Probability

- $\alpha_j(p,q)$: The probability of beginning with N^1 & generating the non-terminal N^j_{pq} and all words outside $w_p..w_q$

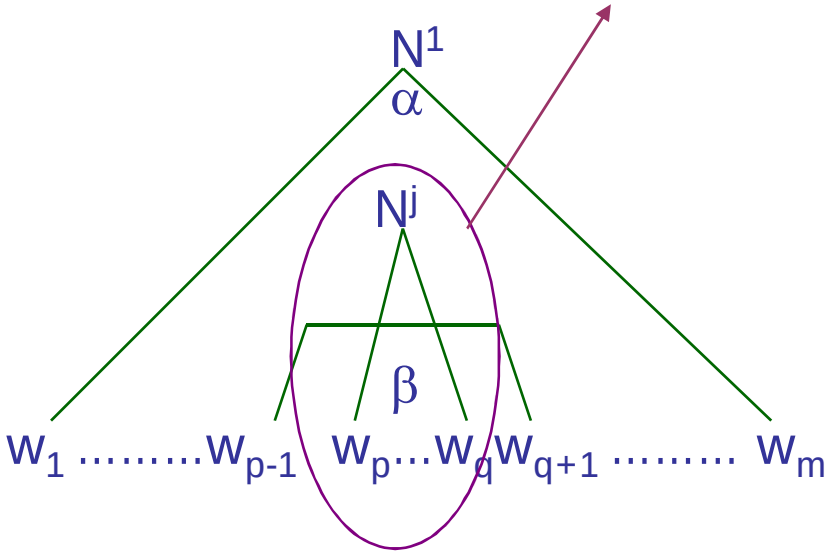
$$\alpha_j(p,q) = P(w_{1(p-1)}, N^j_{pq}, w_{(q+1)m} \mid G)$$



Inside Probabilities

- $\beta_j(p,q)$: The probability of generating the words $w_p..w_q$ starting with the non-terminal N^j_{pq}.

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

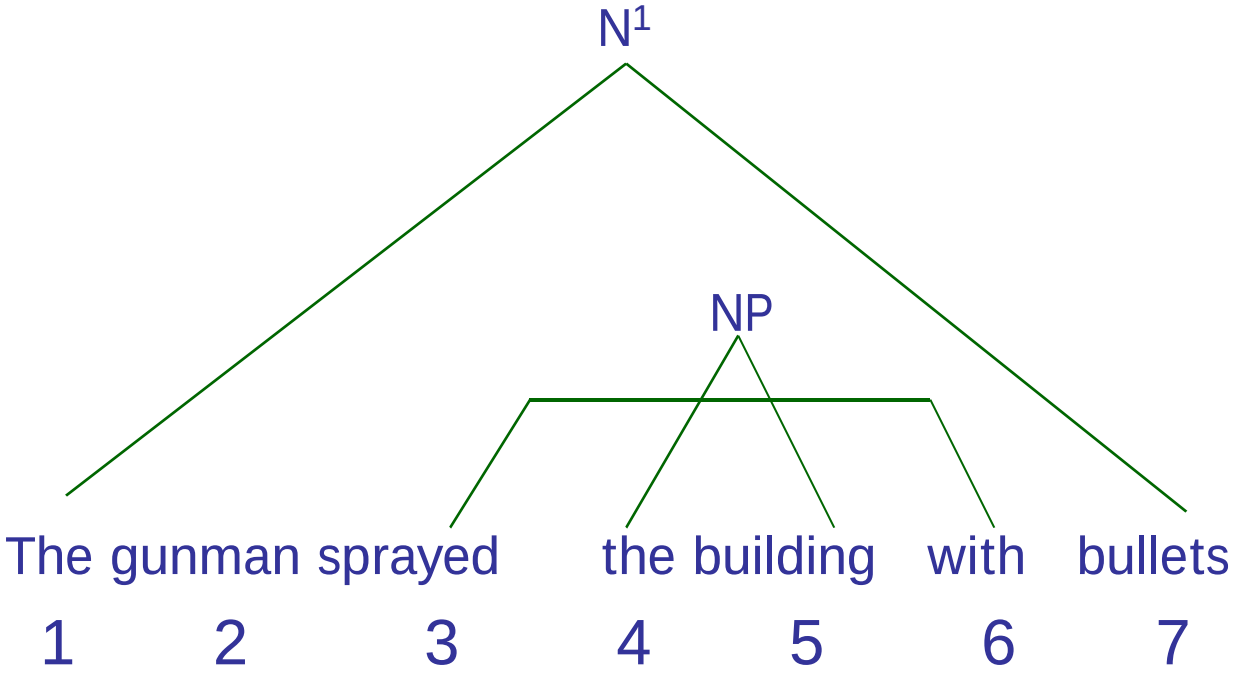


Outside & Inside Probabilities: example

$\alpha_{NP}(4,5)$ for "the building"

$= P(\text{The gunman sprayed}, NP_{4,5}, \text{with bullets} \mid G)$

$\beta_{NP}(4,5)$ for "the building" $= P(\text{the building} \mid NP_{4,5}, G)$



Parsers Comparison

(Charniack, Collins, Stanford, RASP)

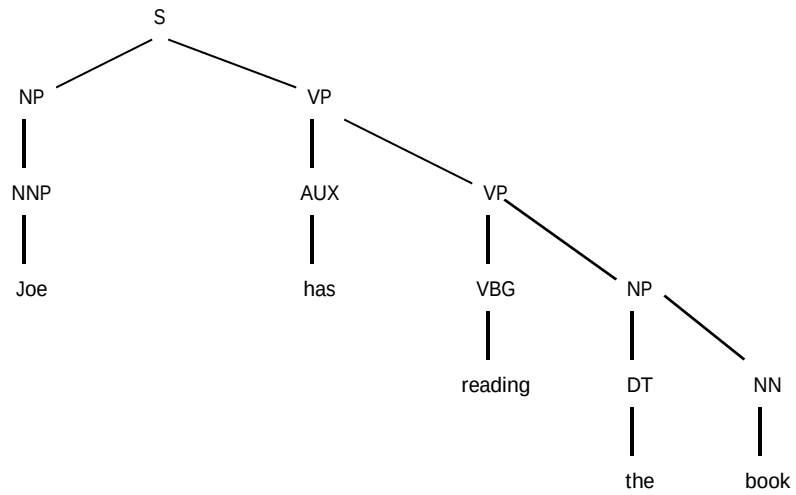
Study by many masters students, notably,

Avishek, Nikhilesh, Abhishek and *Harshada*

Parser comparison: Handling ungrammatical sentences

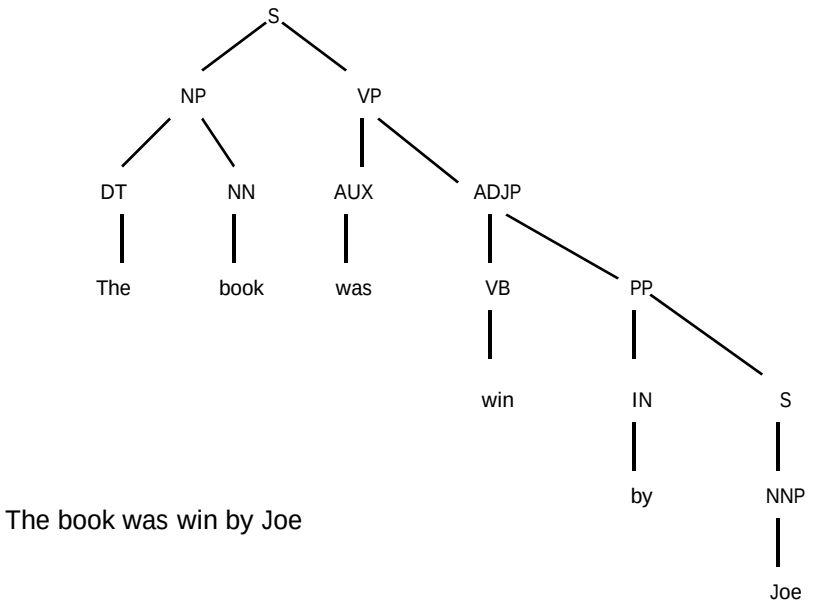
Charniak (ungrammatical 1)

- Here *has* is tagged as AUX



Joe has reading the book

Charniak (ungrammatical 2)

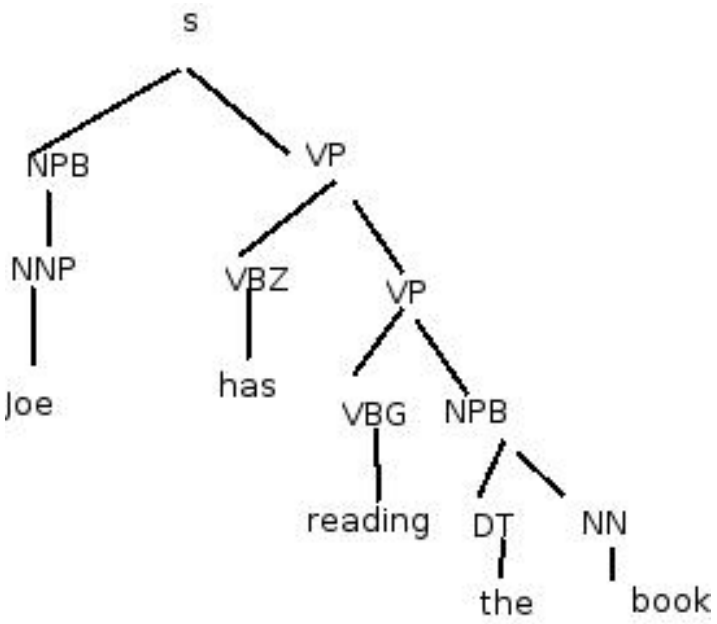


The book was win by Joe

- *Win* is treated as a verb and it does not make any difference whether it is in the present or the past tense

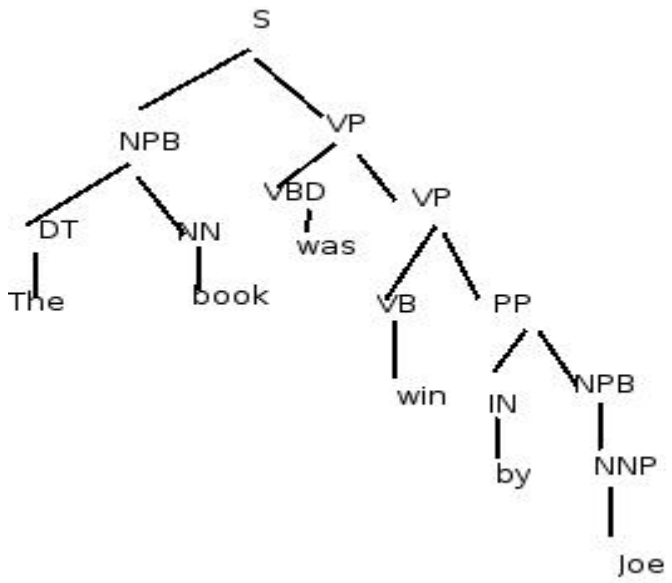
Collins (ungrammatical 1)

- *Has* should have been AUX.



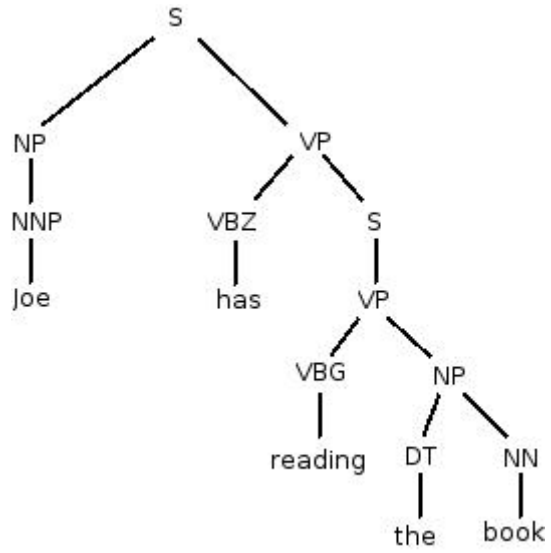
Collins (ungrammatical 2)

- Same as charniack



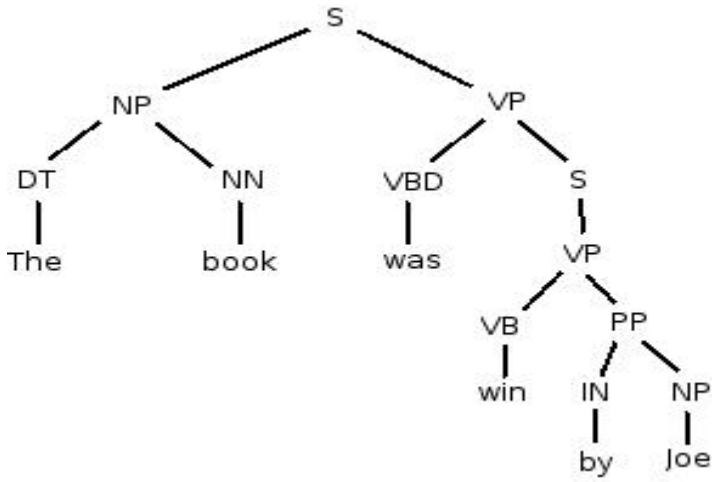
Stanford (ungrammatical 1)

- *has* is treated as VBZ and not AUX.



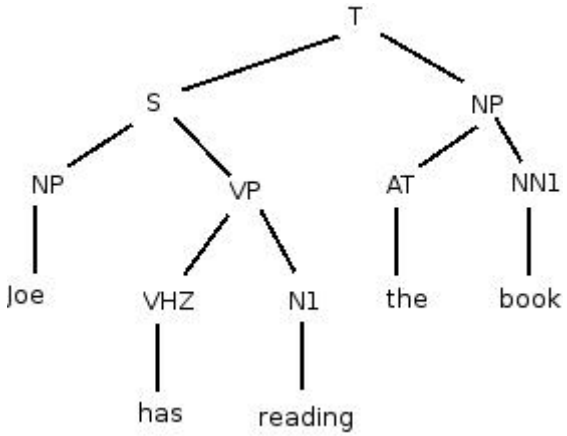
Stanford (ungrammatical 2)

- Same as Charniak



RASP (ungrammatical 1)

- Inaccurate tree



Observation

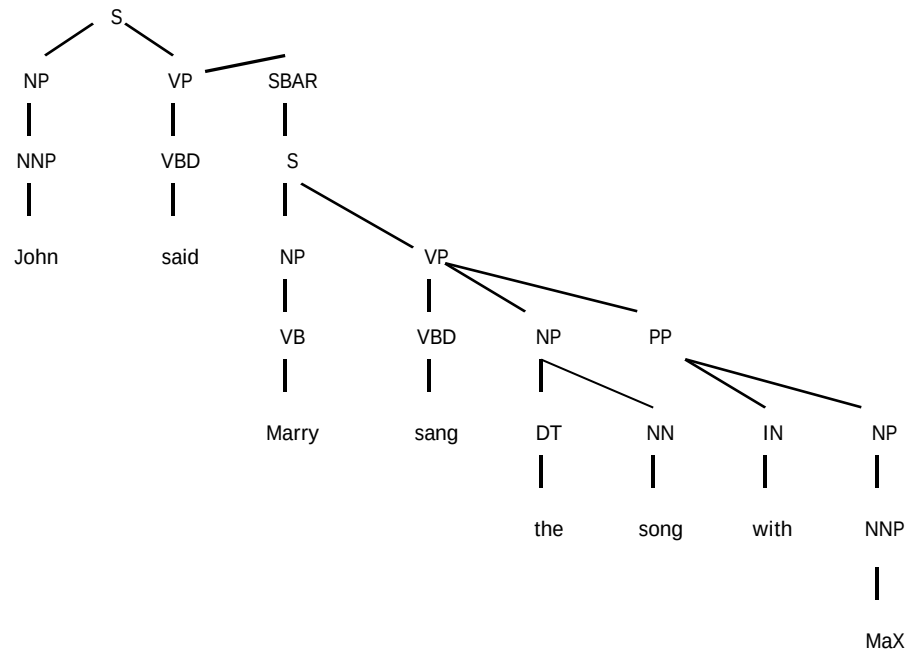
- For the sentence ‘Joe has reading the book’ Charniak performs the best; it is able to predict that the word ‘has’ in the sentence should actually be an AUX
- Though both the RASP and Collins can produce a parse tree, they both cannot predict that the sentence is not grammatically correct
- Stanford performs the worst, it inserts extra ‘S’ nodes into the parse tree.

Observation *(contd.)*

- For the sentence ‘The book was win by Joe’, all the parsers give the same parse structure which is correct.

Ranking in case of multiple parses

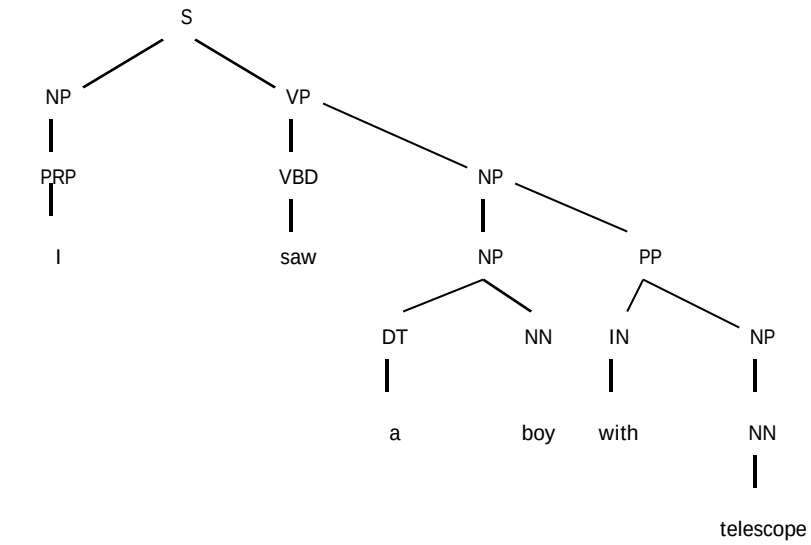
Charniak (Multiple Parses 1)



John said Marry sang the song with Max

- The parse produced is semantically correct

Charniak (Multiple Parses 2)

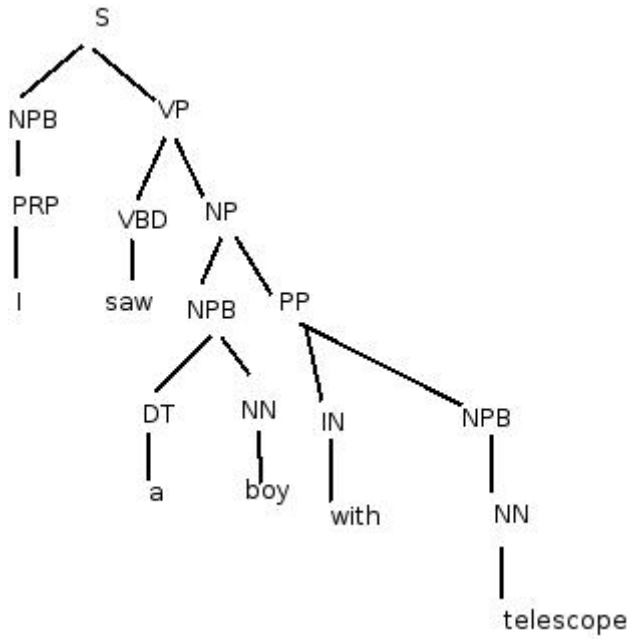


I saw a boy with telescope

- PP is attached to NP which is one of the correct meanings

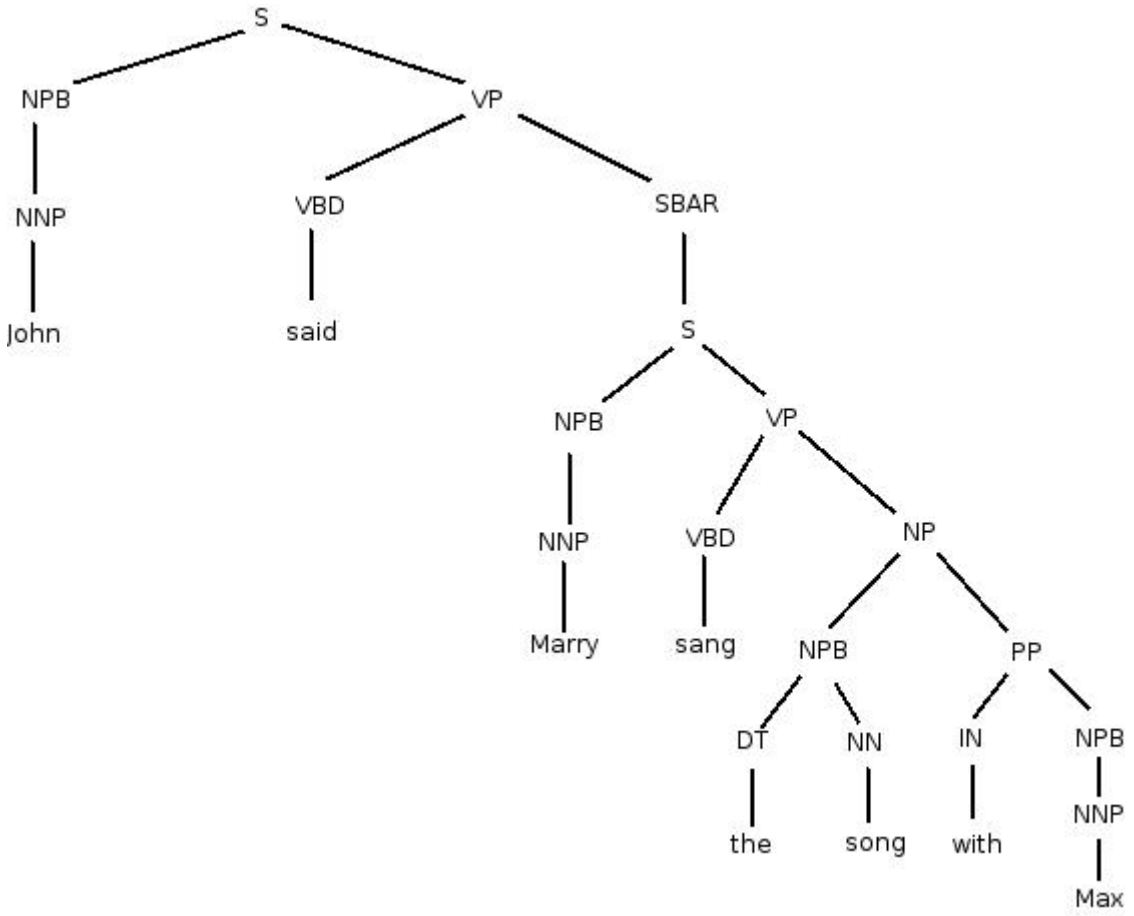
Collins (Multiple Parses 1)

- Same as Charniak.



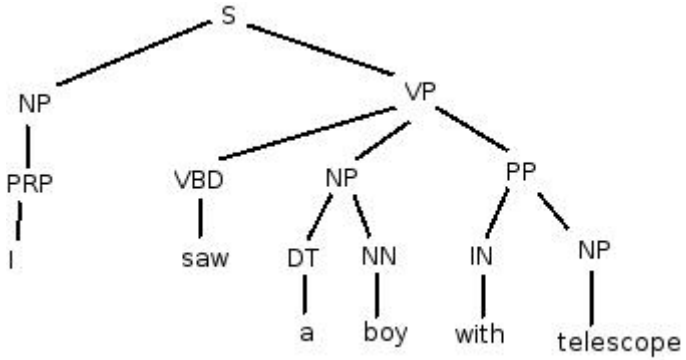
Collins (Multiple Parses 2)

- Same as Charniak



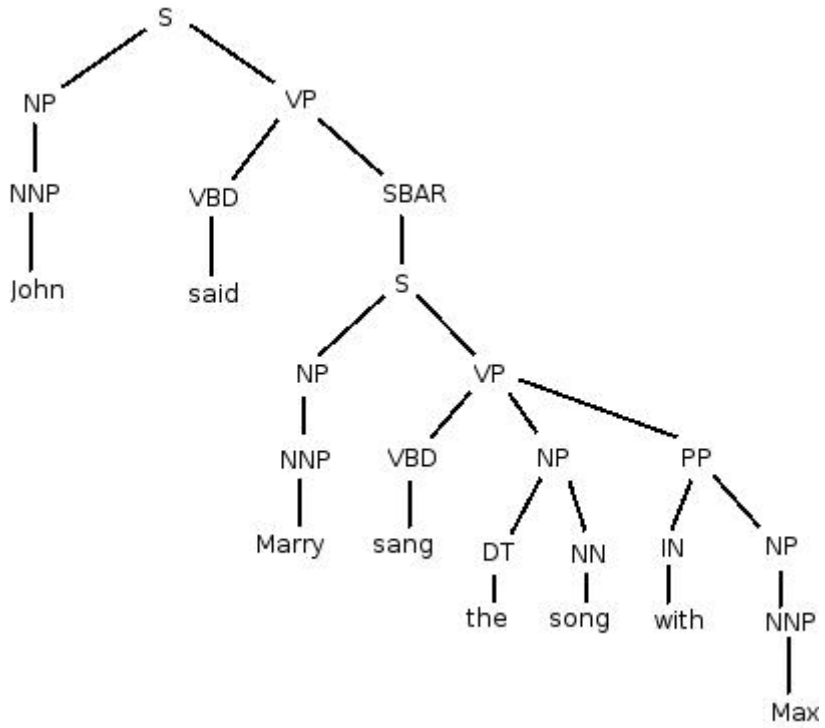
Stanford (Multiple Parses 1)

- PP is attached to VP which is one of the correct meanings possible



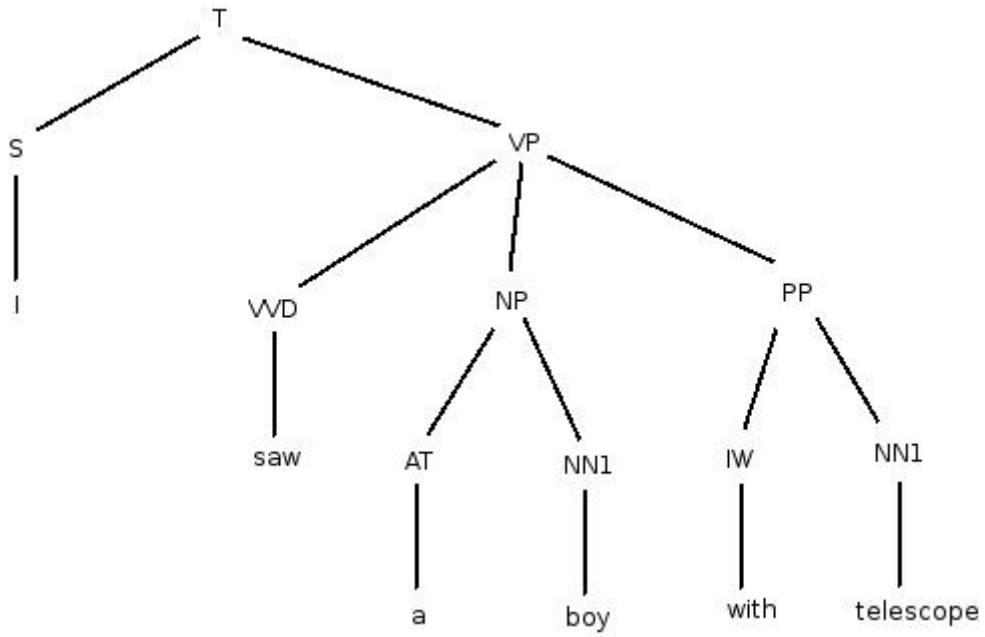
Stanford (Multiple Parses 2)

- Same as Charniak.



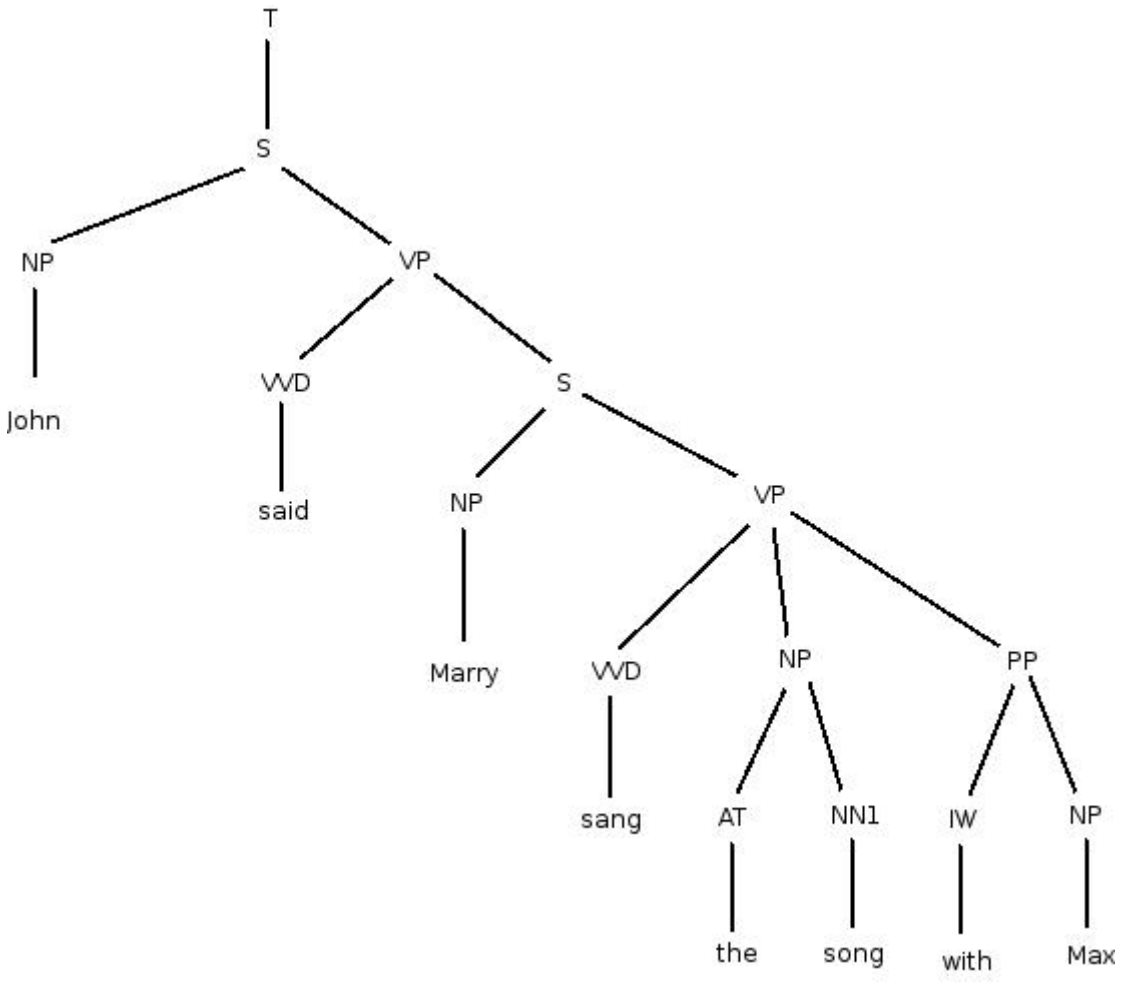
RASP (Multiple Parses 1)

- PP is attached to VP.



RASP (Multiple Parses 2)

- The change in the pos tags as compared to charniak is due to the different corpora but the parse trees are comparable.



Observation

- All of them create one of the correct parses whenever multiple parses are possible.

- All of them produce multiple parse trees and the best is displayed based on the type of the parser
  - Charniak — Probablistic Lexicalised Bottom-Up Chart Parser
  - Collins — Head-driven statistical Beam Search Parser
  - Stanford — Probalistic A* Parser
  - RASP — Probablistic GLR Parser

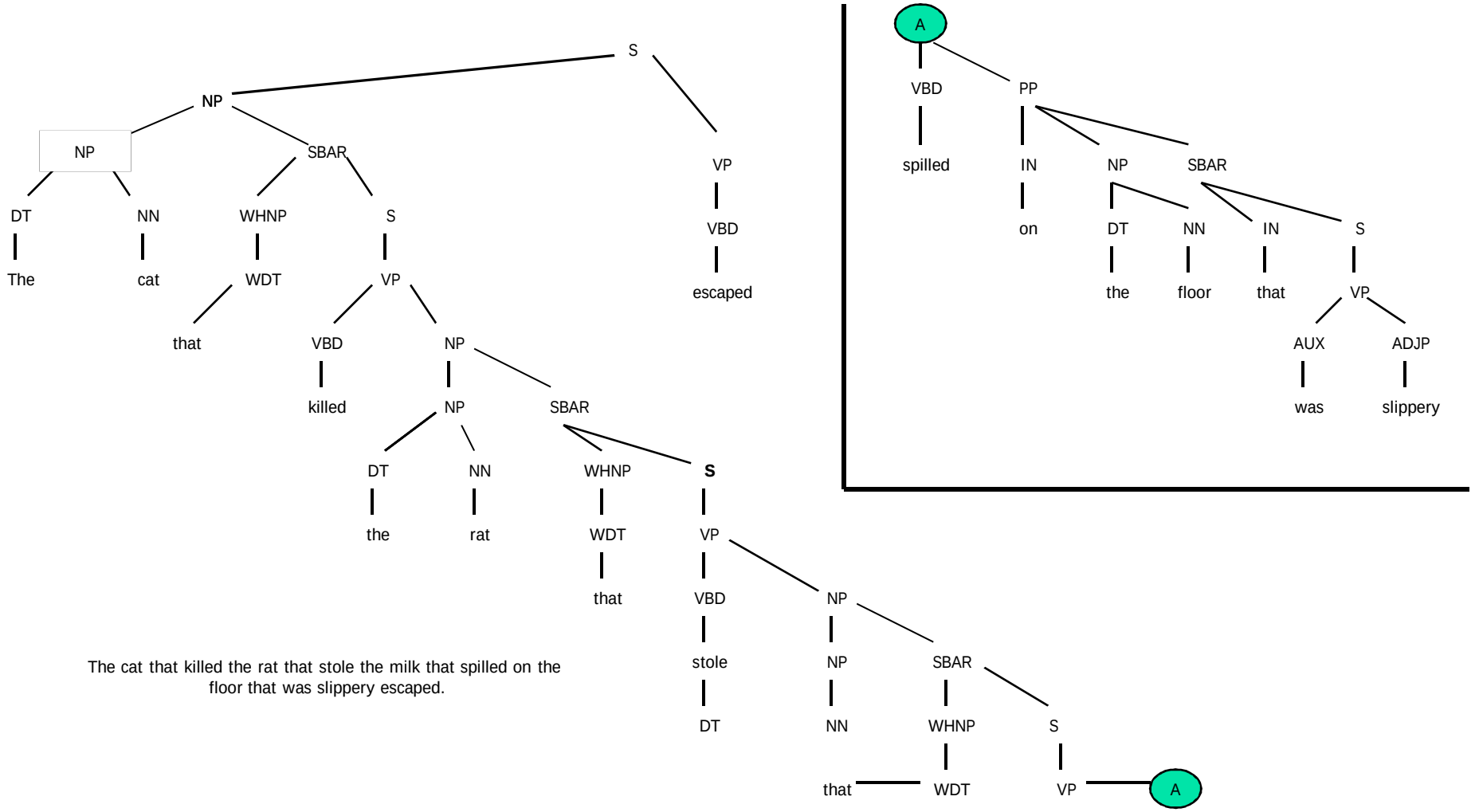
Time taken

- 54 instances of the sentence ‘This is just to check the time’ is used to check the time

- Time taken
  - Collins : 40s
  - Stanford : 14s
  - Charniak : 8s
  - RASP : 5s

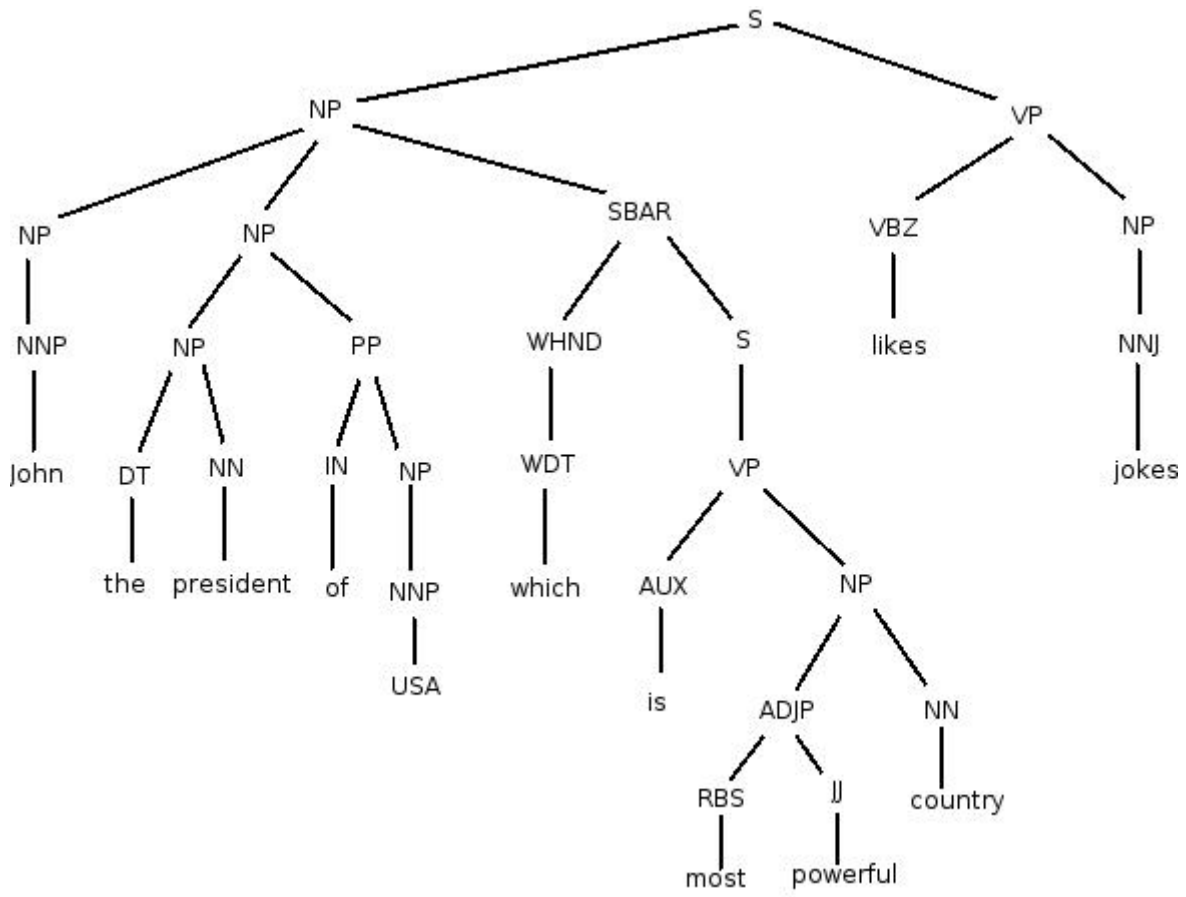
Embedding Handling

Charniak (Embedding 1)

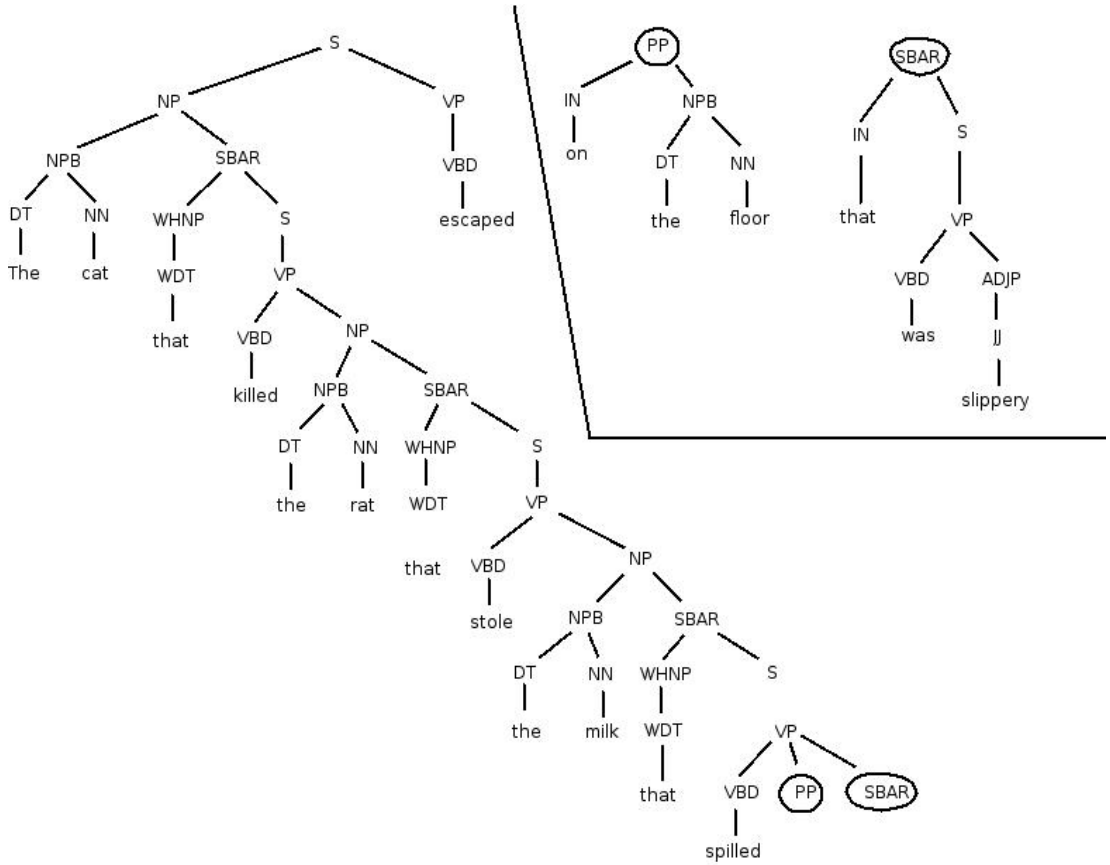


The cat that killed the rat that stole the milk that spilled on the floor that was slippery escaped.

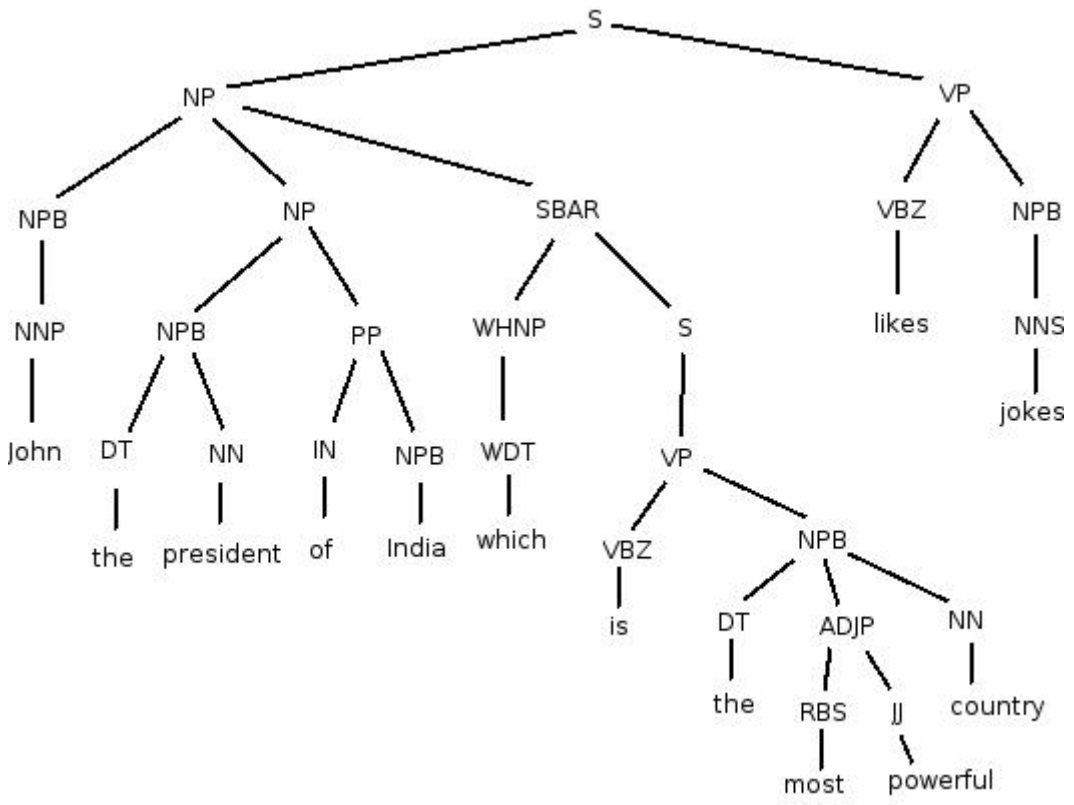
Charniak (Embedding 2)



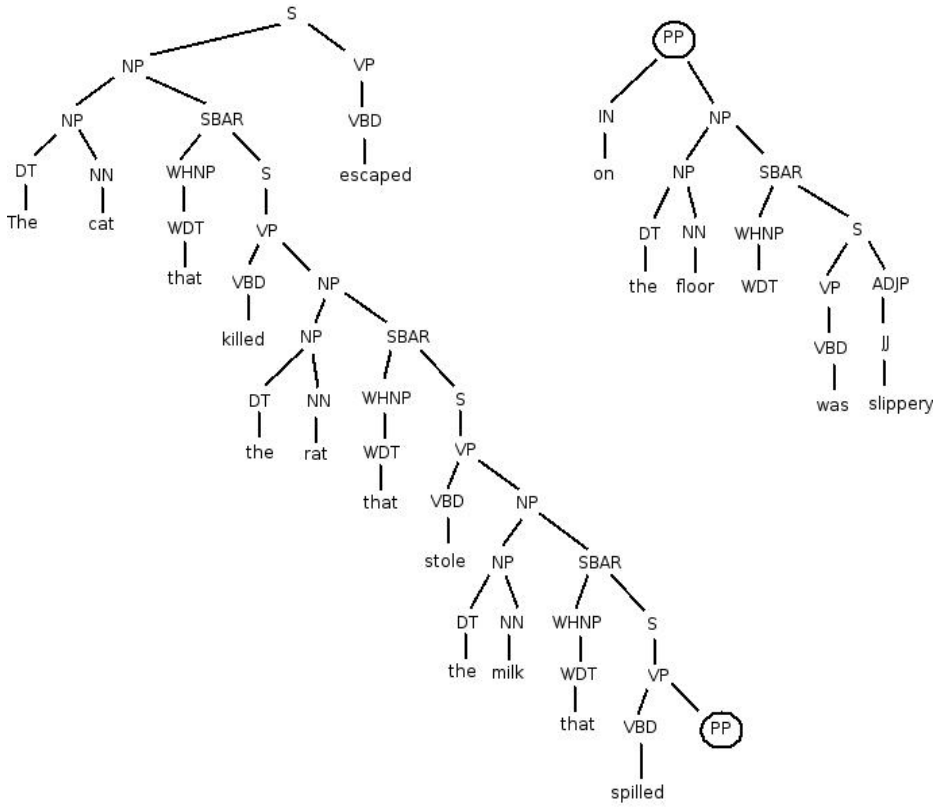
Collins (Embedding 1)



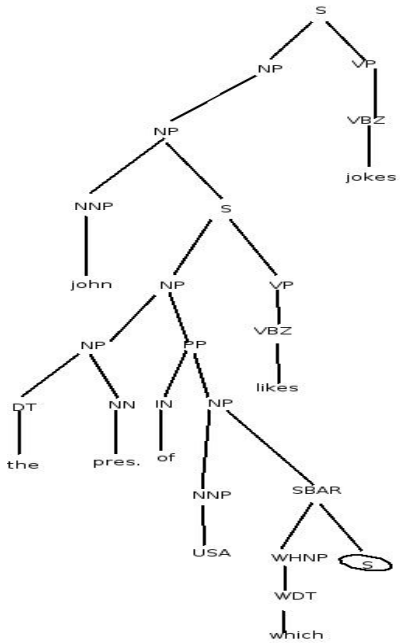
Collins (Embedding 2)

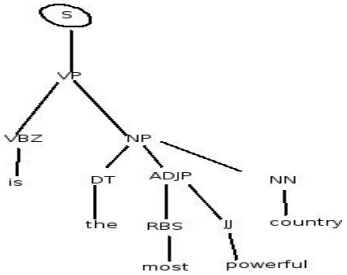


Stanford (Embedding 1)

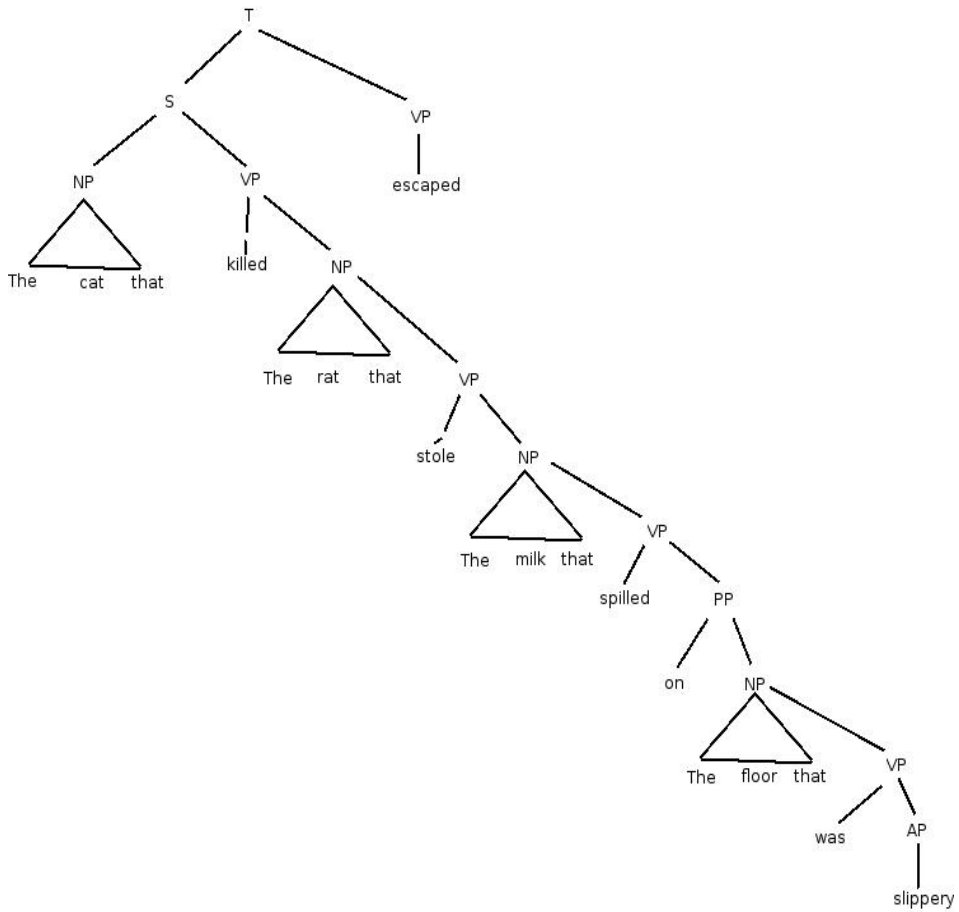


Stanford (Embedding 2)

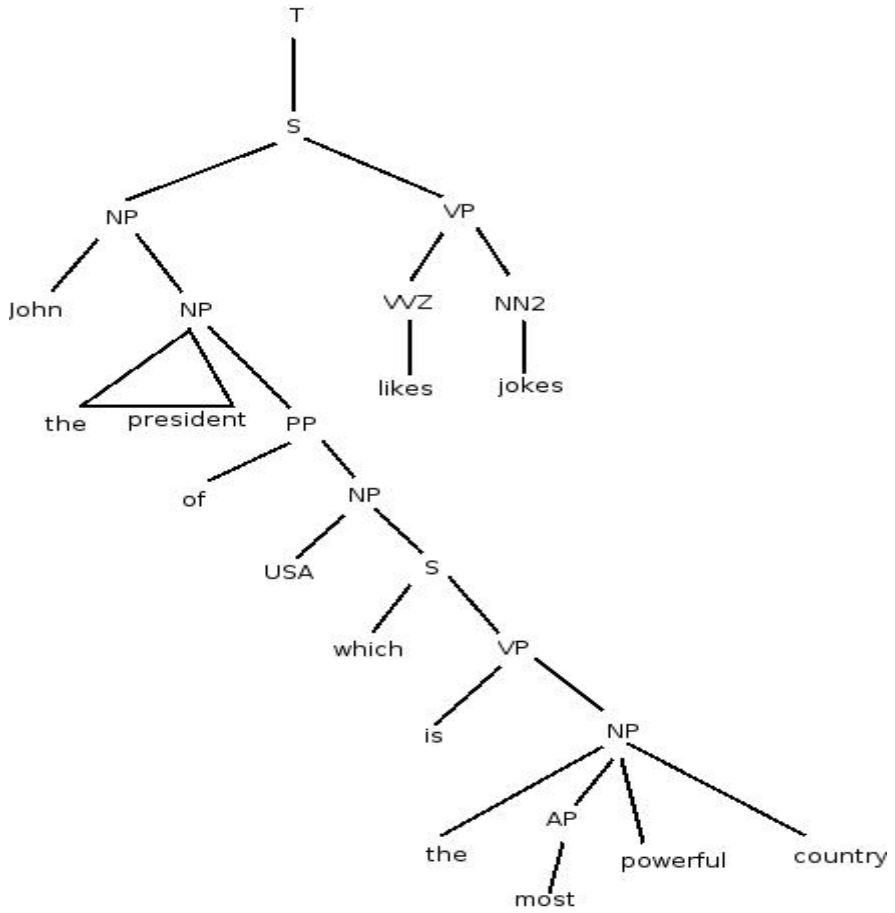




RASP (Embedding 1)



RASP (Embedding 2)



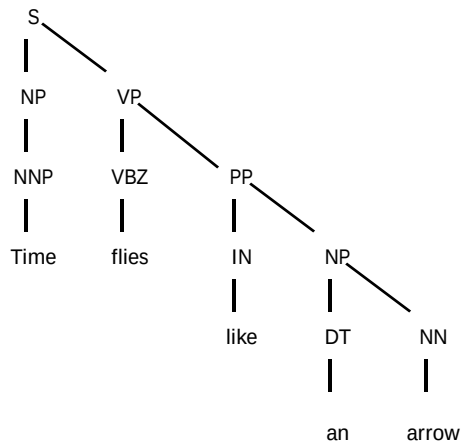
Observation

- For the sentence ‘The cat that killed the rat that stole the milk that spilled on the floor that was slippery escaped.’ all the parsers give the correct results.

- For the sentence ‘John the president of USA which is the most powerful country likes jokes’: RASP , Charniak and Collins give correct parse, *i.e.*, it attaches the verb phrase ‘likes jokes’ to the top NP ‘John’ .
- Stanford produces incorrect parse tree; attaches the VP ‘likes’ to wrong NP ‘the president of ...’

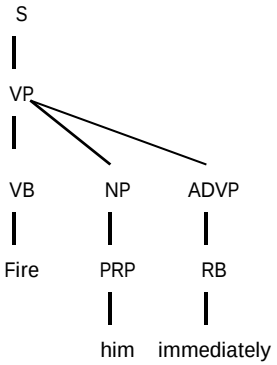
Handling multiple POS tags

Charniak (multiple pos 1)

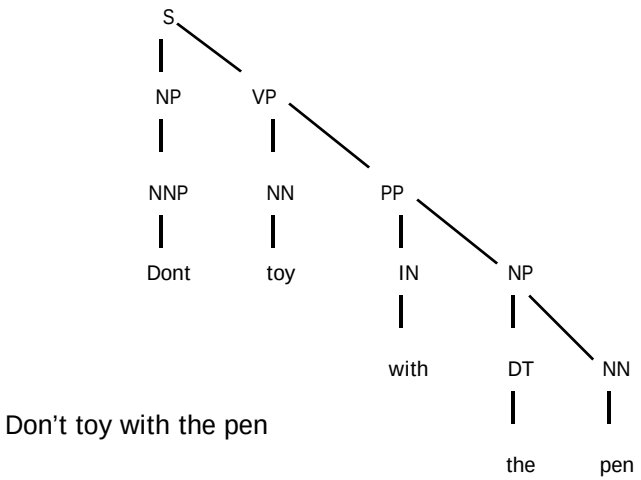


Time flies like an arrow

Fire him immediately

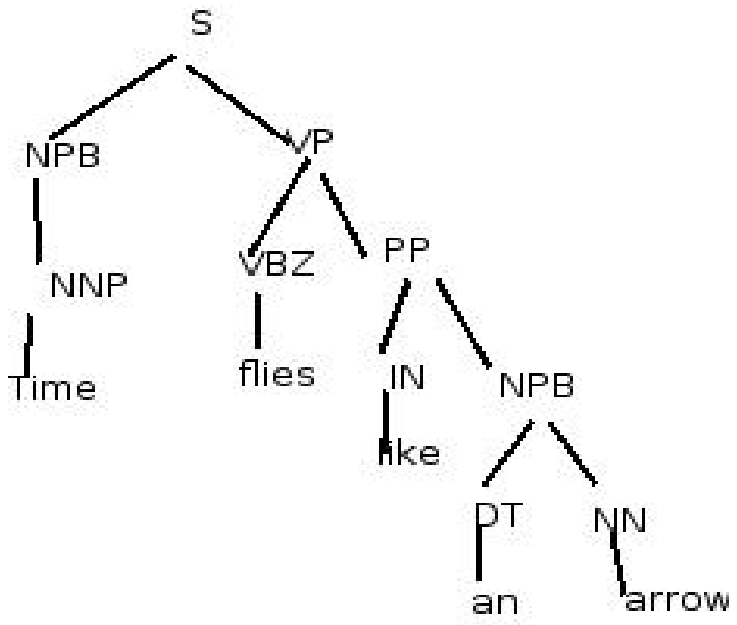


Charniak (multiple pos 2)

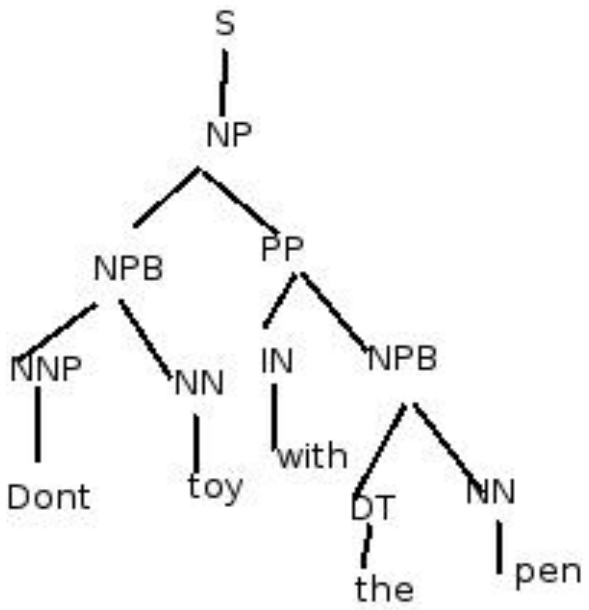


Don’t toy with the pen

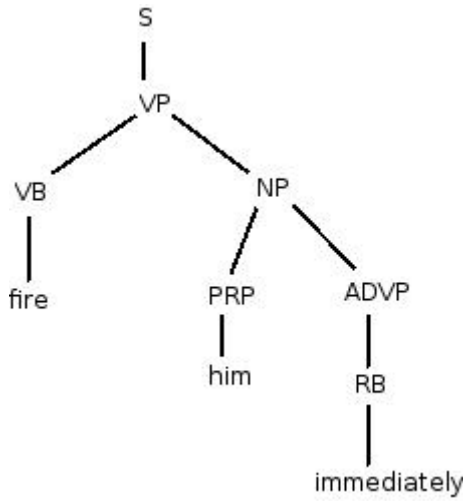
Collins (multiple pos 1)

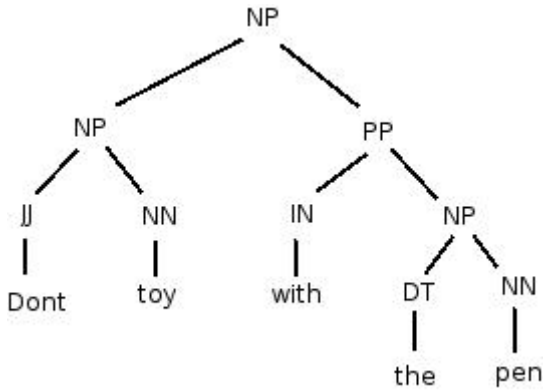


Collins (multiple pos 2)

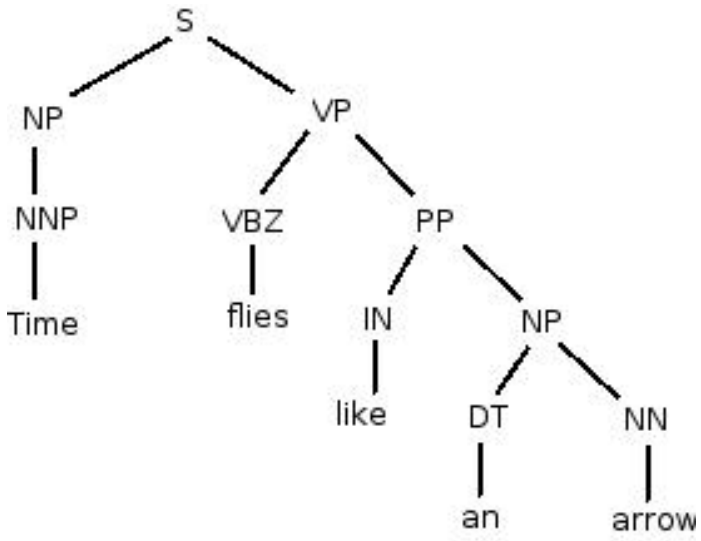


Stanford (multiple pos 1)

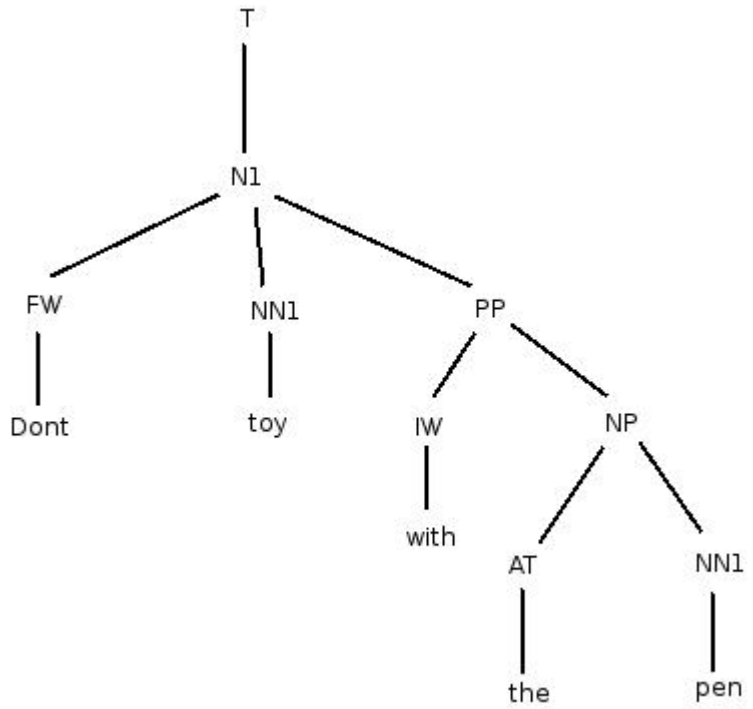




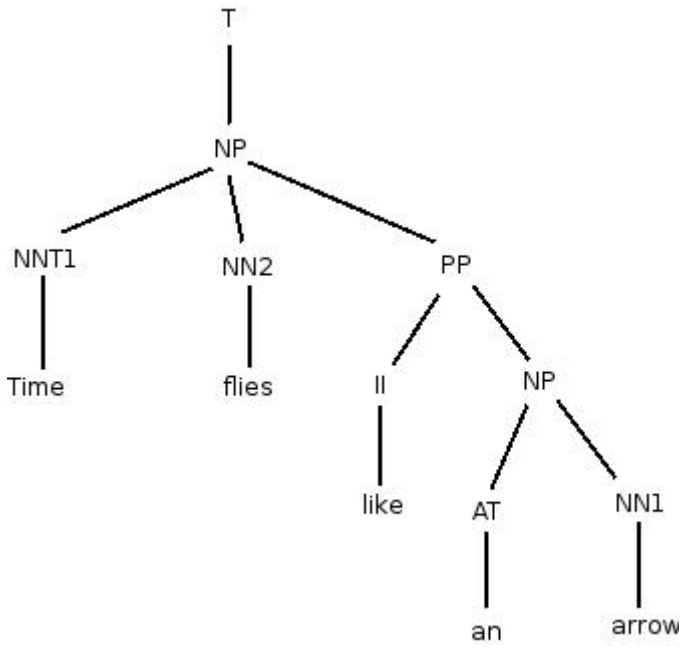
Stanford (multiple pos 2)



RASP (multiple pos 1)



RASP (multiple pos 2)

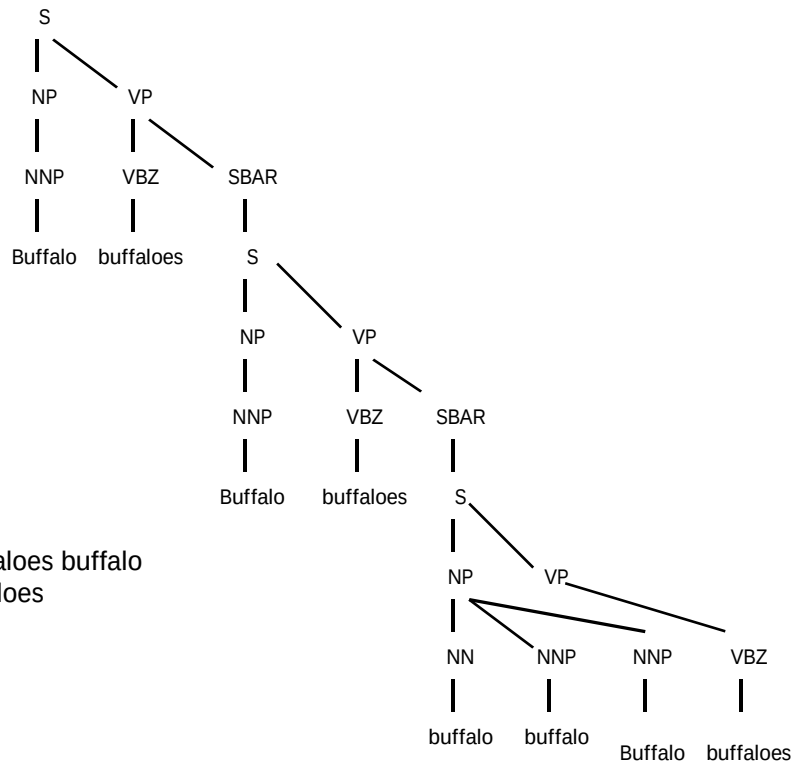


Observation

- All but RASP give comparable pos tags. In the sentence ‘Time flies like an arrrow’ RASP give flies as noun.
- In sentence ‘Don’t toy with the pen’, all parsers are tagging ‘toy’ as noun.

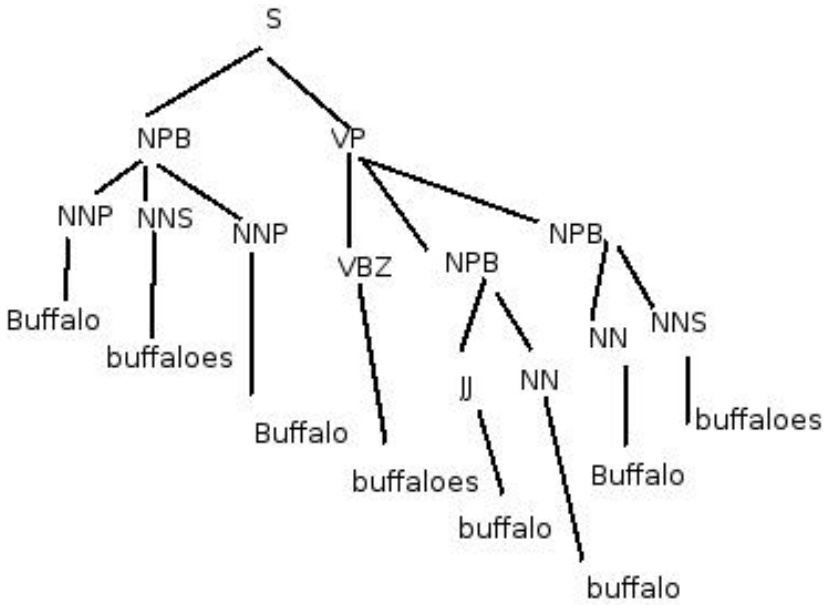
Repeated Word handling

Charniak

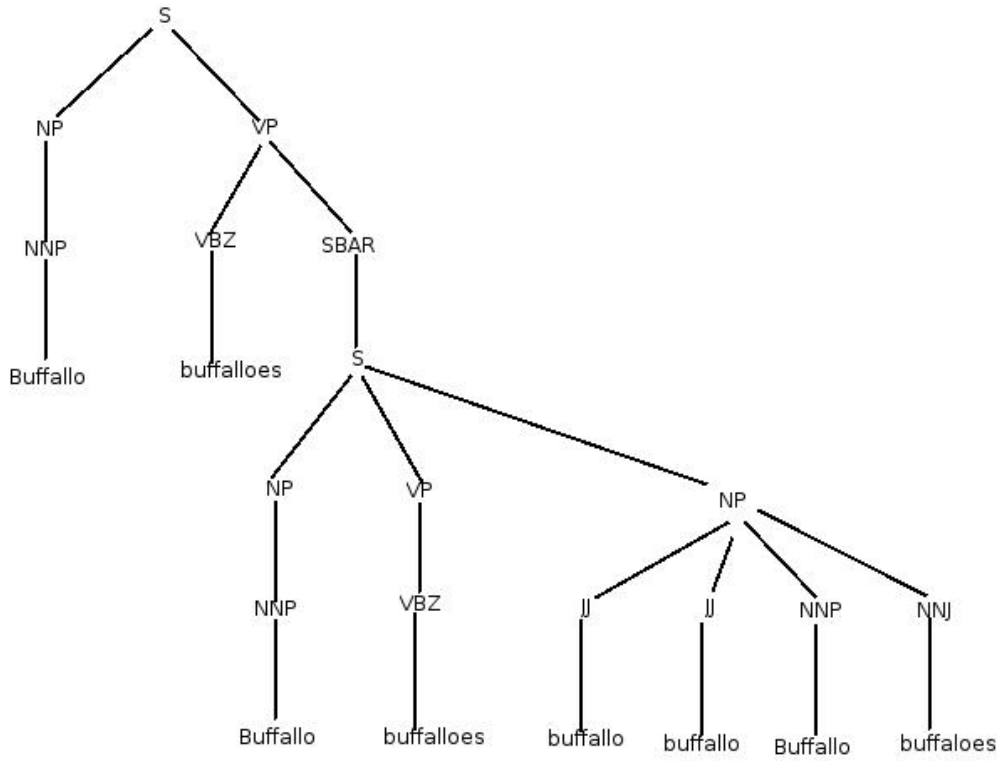


Buffalo buffaloes Buffalo buffaloes buffalo buffalo Buffalo buffaloes

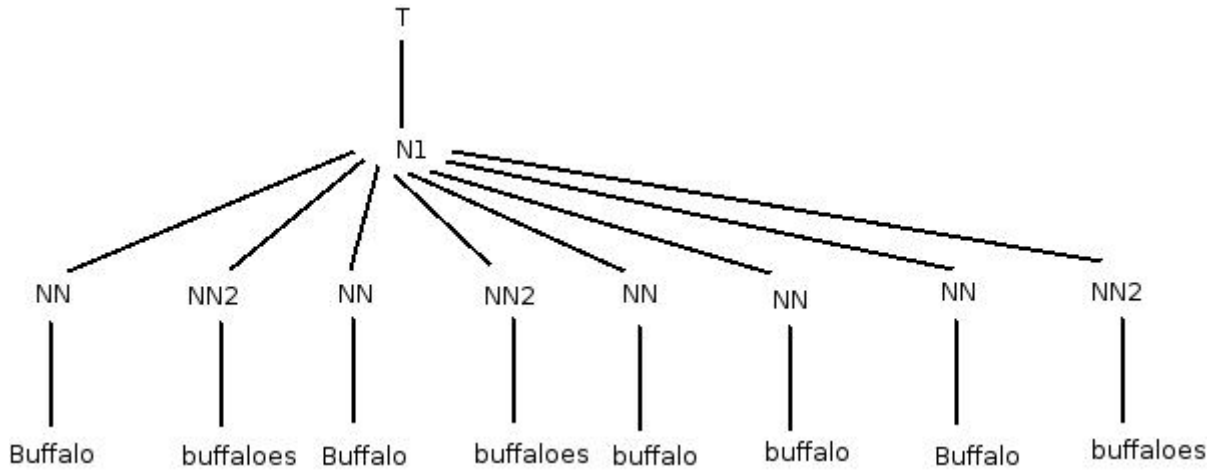
Collins


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

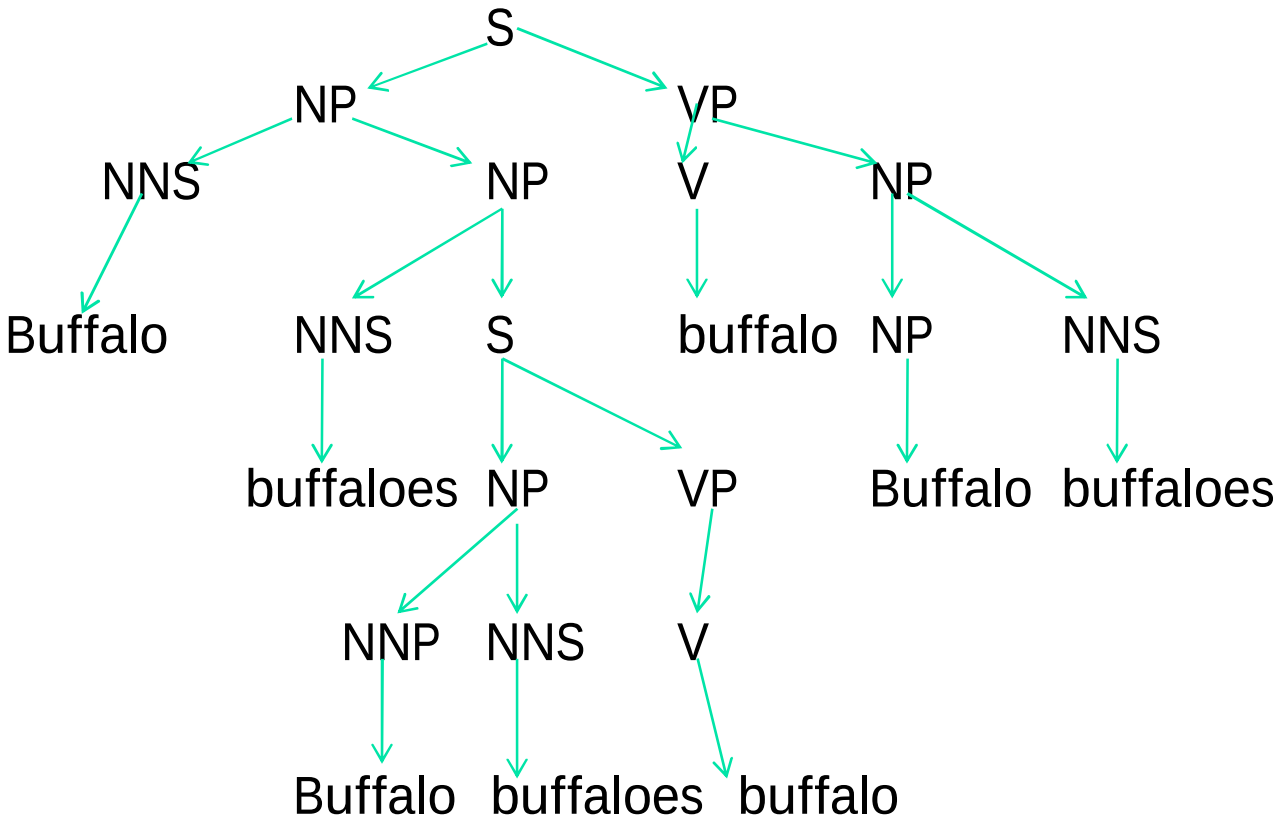
Stanford


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

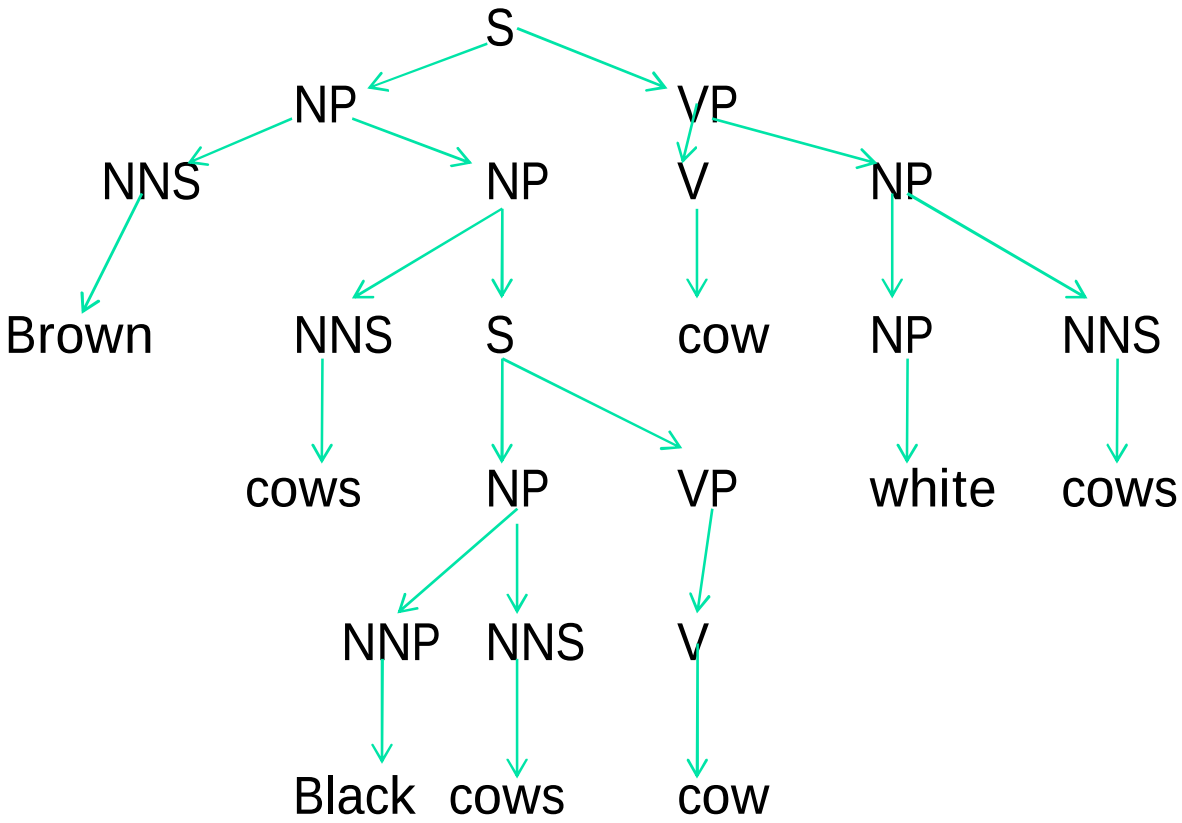
RASP


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

Correct parse



Another sentence of same structure



Observation

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

Long sentences

- Given a sentence of 394 words, only RASP was able to parse.

Lengthy sentence

- One day, Sam left his small, yellow home to head towards the meat-packing plant where he worked, a task which was never completed, as on his way, he tripped, fell, and went careening off of a cliff, landing on and destroying Max, who, incidentally, was also heading to his job at the meat-packing plant, though not the same plant at which Sam worked, which he would be heading to, if he had been aware that that the plant he was currently heading towards had been destroyed just this morning by a mysterious figure clad in black, who hailed from the small, remote country of France, and who took every opportunity he could to destroy small meat-packing plants, due to the fact that as a child, he was tormented, and frightened, and beaten savagely by a family of meat-packing plants who lived next door, and scarred his little mind to the point where he became a twisted and sadistic creature, capable of anything, but specifically capable of destroying meat-packing plants, which he did, and did quite often, much to the chagrin of the people who worked there, such as Max, who was not feeling quite so much chagrin as most others would feel at this point, because he was dead as a result of an individual named Sam, who worked at a competing meat-packing plant, which was no longer a competing plant, because the plant that it would be competing against was, as has already been mentioned, destroyed in, as has not quite yet been mentioned, a massive, mushroom cloud of an explosion, resulting from a heretofore unmentioned horse manure bomb manufactured from manure harvested from the farm of one farmer J. P. Harvenkirk, and more specifically harvested from a large, ungainly, incontinent horse named Seabiscuit, who really wasn't named Seabiscuit, but was actually named Harold, and it completely baffled him why anyone, particularly the author of a very long sentence, would call him Seabiscuit; actually, it didn't baffle him, as he was just a stupid, manure-making horse, who was incapable of cognitive thought for a variety of reasons, one of which was that he was a horse, and the other of which was that he was just knocked unconscious by a flying chunk of a meat-packing plant, which had been blown to pieces just a few moments ago by a shifty character from France.

Partial RASP Parse of the sentence

- (|One_MC1| |day_NNT1| |,_,| |Sam_NP1| |leave+ed_VVD| |his_APP$| |small_JJ| |,_,| |yellow_JJ| |home_NN1| |to_TO| |head_VV0| |towards_II| |the_AT| |meat-packing_JJ| |plant_NN1| |where_RRQ| |he_PPHS1| |work+ed_VVD| |,_,| |a_AT1| |task_NN1| |which_DDQ| |be+ed_VBDZ| |never_RR| |complete+ed_VVN| |,_,| |as_CSA| |on_II| |his_APP$| |way_NN1| |,_,| |he_PPHS1| |trip+ed_VVD| |,_,| |fall+ed_VVD| |,_,| |and_CC| |go+ed_VVD| |careen+ing_VVG| |off_RP| |of_IO| |a_AT1| |cliff_NN1| |,_,| |land+ing_VVG| |on_RP| |and_CC| |destroy+ing_VVG| |Max_NP1| |,_,| |who_PNQS| |,_,| |incidentally_RR| |,_,| |be+ed_VBDZ| |also_RR| |head+ing_VVG| |to_II| |his_APP$| |job_NN1| |at_II| |the_AT| |meat-packing_JB| |plant_NN1| |,_,| |though_CS| |not+_XX| |the_AT| |same_DA| |plant_NN1| |at_II| |which_DDQ| |Sam_NP1| |work+ed_VVD| |,_,| |which_DDQ| |he_PPHS1| |would_VM| |be_VB0| |head+ing_VVG| |to_II| |,_,| |if_CS| |he_PPHS1| |have+ed_VHD| |be+en_VBN| |aware_JJ| |that_CST| |that_CST| |the_AT| |plant_NN1| |he_PPHS1| |be+ed_VBDZ| |currently_RR| |head+ing_VVG| |towards_II| |have+ed_VHD| |be+en_VBN| |destroy+ed_VVN| |just_RR| |this_DD1| |morning_NNT1| |by_II| |a_AT1| |mysterious_JJ| |figure_NN1| |clothe+ed_VVN| |in_II| |black_JJ| |,_,| |who_PNQS| |hail+ed_VVD| |from_II| |the_AT| |small_JJ| |,_,| |remote_JJ| |country_NN1| |of_IO| |France_NP1| |,_,| |and_CC| |who_PNQS| |take+ed_VVD| |every_AT1| |opportunity_NN1| |he_PPHS1| |could_VM| |to_TO| |destroy_VV0| |small_JJ| |meat-packing_NN1| |plant+s_NN2| |,_,| |due_JJ| |to_II| |the_AT| |fact_NN1| |that_CST| |as_CSA| |a_AT1| |child_NN1| |,_,| |he_PPHS1| |be+ed_VBDZ| |torment+ed_VVN| |,_,| |and_CC| |frighten+ed_VVD| |,_,| |and_CC| |beat+en_VVN| |savagely_RR| |by_II| |a_AT1| |family_NN1| |of_IO| |meat-packing_JJ| |plant+s_NN2| |who_PNQS| |live+ed_VVD| |next_MD| |door_NN1| |,_,| |and_CC| |scar+ed_VVD| |his_APP$| |little_DD1| |mind_NN1| |to_II| |the_AT| |point_NNL1| |where_RRQ| |he_PPHS1| |become+ed_VVD| |a_AT1| |twist+ed_VVN| |and_CC| |sadistic_JJ| |creature_NN1| |,_,| |capable_JJ| |of_IO| |anything_PN1| |,_,| |but_CCB| |specifically_RR| |capable_JJ| |of_IO| |destroy+ing_VVG| |meat-packing_JJ| |plant+s_NN2| |,_,| |which_DDQ| |he_PPHS1| |do+ed_VDD| |,_,| |and_CC| |do+ed_VDD| |quite_RG| |often_RR| |,_,| |much_DA1| |to_II| |the_AT| |chagrin_NN1| |of_IO| |the_AT| |people_NN| |who_PNQS| |work+ed_VVD| |there_RL| |,_,| |such_DA| |as_CSA| |Max_NP1| |,_,| |who_PNQS| |be+ed_VBDZ| |not+_XX| |feel+ing_VVG| |quite_RG| |so_RG| |much_DA1| |chagrin_NN1| |as_CSA| |most_DAT| |other+s_NN2| |would_VM| |feel_VV0| |at_II| |this_DD1| |point_NNL1| |,_,| |because_CS| |he_PPHS1| |be+ed_VBDZ| |dead_JJ| |as_CSA| |a_AT1| |result_NN1| |of_IO| |an_AT1| |individual_NN1| |name+ed_VVN| |Sam_NP1| |,_,| |who_PNQS| |work+ed_VVD| |at_II| |a_AT1| |compete+ing_VVG| |meat-packing_JJ| |plant_NN1| |,_,| |which_DDQ| |be+ed_VBDZ| |no_AT| |longer_RRR| |a_AT1| |compete+ing_VVG| |plant_NN1| |,_,| |because_CS| |the_AT| |plant_NN1| |that_CST| |it_PPH1| |would_VM| |be_VB0| |compete+ing_VVG| |against_II| |be+ed_VBDZ| |,_,| |as_CSA| |have+s_VHZ| |already_RR| |be+en_VBN| |mention+ed_VVN| |,_,| |destroy+ed_VVN| |in_RP| |,_,| |as_CSA| |have+s_VHZ| |not+_XX| |quite_RG| |yet_RR| |be+en_VBN| |mention+ed_VVN| |,_,| |a_AT1| |massive_JJ| |,_,| |mushroom_NN1| |cloud_NN1| |of_IO| |an_AT1| |explosion_NN1| |,_,| |result+ing_VVG| |from_II| |a_AT1| |heretofore_RR| |unmentioned_JJ| |horse_NN1| |manure_NN1| |bomb_NN1| |manufacture+ed_VVN| |from_II| |manure_NN1| |harvest+ed_VVN| |from_II| |the_AT| |farm_NN1| |of_IO| |one_MC1| |farmer_NN1| J._NP1 P._NP1 Harvenkirk_NP1| |,_,| |and_CC| |more_DAR| |specifically_RR| |harvest+ed_VVN| |from_II| |a_AT1| |large_JJ| |,_,| |ungainly_JJ| |,_,| |incontinent_NN1| |horse_NN1| |name+ed_VVN| |Seabiscuit_NP1| |,_,| |who_PNQS| |really_RR| |be+ed_VBDZ| |not+_XX| |name+ed_VVN| |Seabiscuit_NP1| |,_,| |but_CCB| |be+ed_VBDZ| |actually_RR| |name+ed_VVN| |Harold_NP1| |,_,| |and_CC| |it_PPH1| |completely_RR| |baffle+ed_VVD| |he+_PPHO1| |why_RRQ| |anyone_PN1| |,_,| |particularly_RR| |the_AT| |author_NN1| |of_IO| |a_AT1| |very_RG| |long_JJ| |sentence_NN1| |,_,| |would_VM| |call_VV0| |he+_PPHO1| |Seabiscuit_NP1| |;_;| |actually_RR| |,_,| |it_PPH1| |do+ed_VDD| |not+_XX| |baffle_VV0| |he+_PPHO1| |,_,| |as_CSA| |he_PPHS1| |be+ed_VBDZ| |just_RR| |a_AT1| |stupid_JJ| |,_,| |manure-making_NN1| |horse_NN1| |,_,| |who_PNQS| |be+ed_VBDZ| |incapable_JJ| |of_IO| |cognitive_JJ| |thought_NN1| |for_IF| |a_AT1| |variety_NN1| |of_IO| |reason+s_NN2| |,_,| |one_MC1| |of_IO| |which_DDQ| |be+ed_VBDZ| |that_CST| |he_PPHS1| |be+ed_VBDZ| |a_AT1| |horse_NN1| |,_,| |and_CC| |the_AT| |other_JB| |of_IO| |which_DDQ| |be+ed_VBDZ| |that_CST| |he_PPHS1| |be+ed_VBDZ| |just_RR| |knock+ed_VVN| |unconscious_JJ| |by_II| |a_AT1| |flying_NN1| |chunk_NN1| |of_IO| |a_AT1| |meat-packing_JJ| |plant_NN1| |,_,| |which_DDQ| |have+ed_VHD| |be+en_VBN| |blow+en_VVN| |to_II| |piece+s_NN2| |just_RR| |a_AT1| |few_DA2| |moment+s_NNT2| |ago_RA| |by_II| |a_AT1| |shifty_JJ| |character_NN1| |from_II| France_NP1| ._.) -1 ; ()

An important parsing algo

Illustrating CYK [Cocke, Younger, Kashmi] Algo

- 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

CYK: 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.