Sometimes, you have to parse…

MAS.S60

Rob Speer

Catherine Havasi

Say wha?

- Nemo was eaten by a squid!
- Jules wrote that Nemo was eaten by a squid.
- Ben said Jules wrote that Nemo was eaten by a squid.
- I think Ben said Jules wrote that Nemo was eaten by a bear.

In after-years he liked to think that he had been in Very Great Danger during the Terrible Flood, but the only danger he had really been in was the last half-hour of his imprisonment, when Owl, who had just flown up, sat on a branch of his tree to comfort him, and told him a very long story about an aunt who had once laid a seagull's egg by mistake, and the story went on and on, rather like this sentence, until Piglet who was listening out of his window without much hope, went to sleep quietly and naturally, slipping slowly out of the window towards the water until he was only hanging on by his toes, at which moment, luckily, a sudden loud squawk from Owl, which was really part of the story, being what his aunt said, woke the Piglet up and just gave him time to jerk himself back into safety and say, "How interesting, and did she?" when — well, you can imagine his joy when at last he saw the good ship, Brain of Pooh (Captain, C. Robin; 1st Mate, P. Bear) coming over the sea to rescue him...

S but S when S

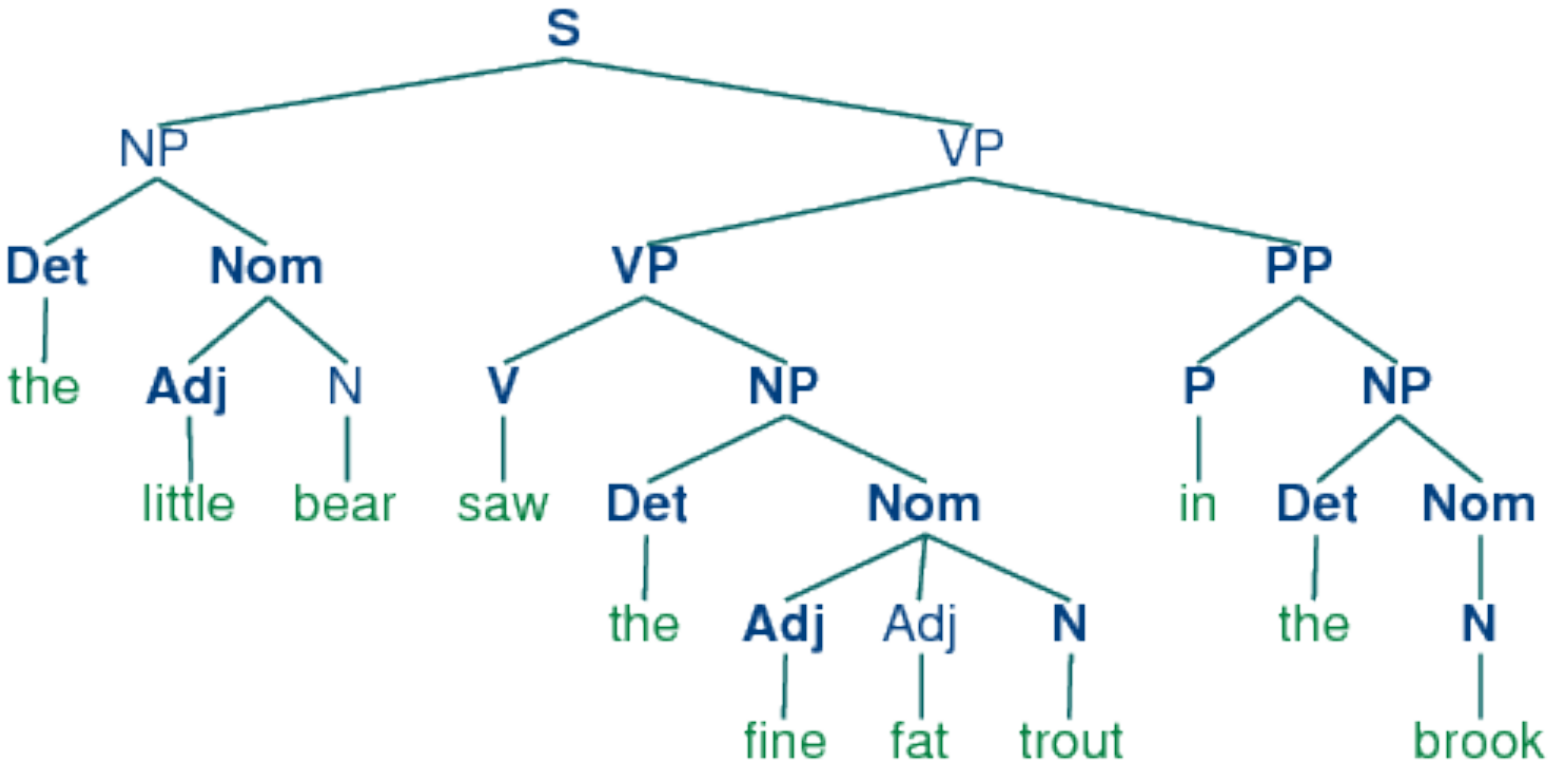
Beyond N-Grams

the	little	bear	saw	the	fine	fat	trout	in	the	brook
the	bear		saw	the	trout			in	it	
He			saw	it				there		
He			ran					there		
He			ran							

Add POS

Det the	Adj little	N bear	V saw	Det the	Adj fine	Adj fat	N trout	P in	Det the	N brook
Det the	Nom bear		V saw	Det the	Nom trout			P in	NP it	
NP He			V saw	NP it				PP there		
NP He			VP ran					PP there		
NP He			VP ran							

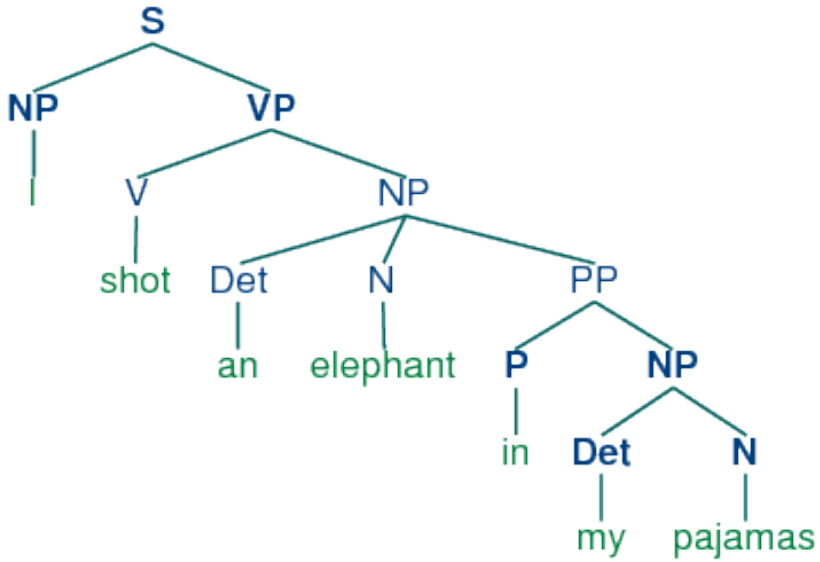
Forest for the Trees

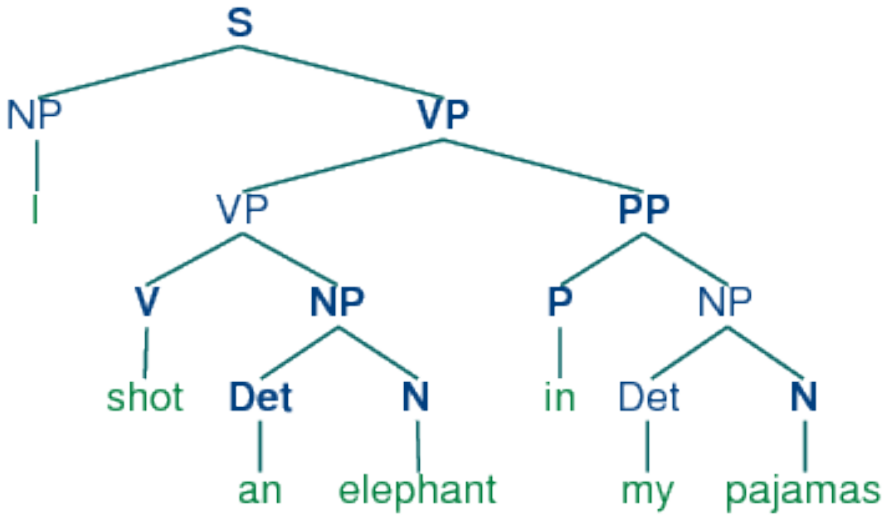


While hunting in Africa, I shot an elephant in my pajamas. How an elephant got into my pajamas I'll never know.

“Fighting elephants could be dangerous.”

Ambiguity


S
NP
VP
I
V
NP
shot
Det
N
PP
an
elephant
P
NP
in
Det
N
my
pajamas


S
NP
VP
I
VP
PP
V
NP
P
NP
shot
Det
N
in
Det
N
an
elephant
my
pajamas

Chomsky Normal Form

- A Context-Free Grammar
- A series of rules
  - Branches (A ->BC)
  - Leaves (B-> q)
  - Empty (S ->)

Grammar for Elephants

```
>>> groucho_grammar = nltk.parse_cfg("""
... S -> NP VP
... PP -> P NP
... NP -> Det N | Det N PP | 'I'
... VP -> V NP | VP PP
... Det -> 'an' | 'my'
... N -> 'elephant' | 'pajamas'
... V -> 'shot'
... P -> 'in'
... """)
```

Parsing for Elephants

```
>>> sent = ['I', 'shot', 'an', 'elephant', 'in', 'my', 'pajamas']
>>> parser = nltk.ChartParser(groucho_grammar)
>>> trees = parser.nbest_parse(sent)
>>> for tree in trees:
...     print tree
...
(S
  (NP I)
  (VP
    (V shot)
    (NP (Det an) (N elephant) (PP (P in) (NP (Det my) (N pajamas))))))
(S
  (NP I)
  (VP
    (VP (V shot) (NP (Det an) (N elephant)))
    (PP (P in) (NP (Det my) (N pajamas)))))
```

Whiteboard: Converting to CNF

Interlude: CNF in Python

What is a parser?

- A parser takes a grammar and maps a sentence according to that grammar.
- A grammar is a set of syntax rules
- It can be learned or coded

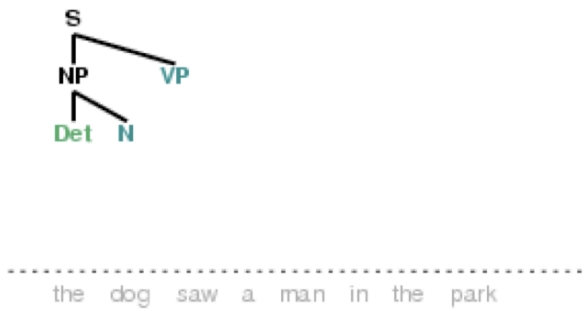
Recursive Descent Parsing

- Start with a top level goal and work **top down** to find the sentence
- Start with an S
- Get to subgoals to final leaf subgoals
- The parse tree grows downward

Recursive Descent Parsing

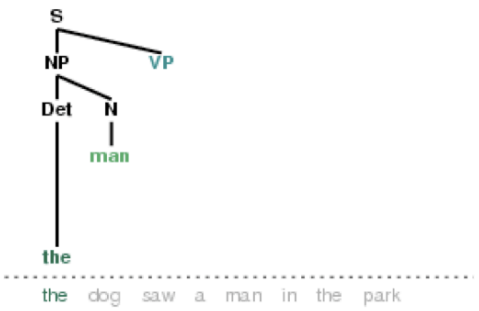
1. Initial stage

2. Second production

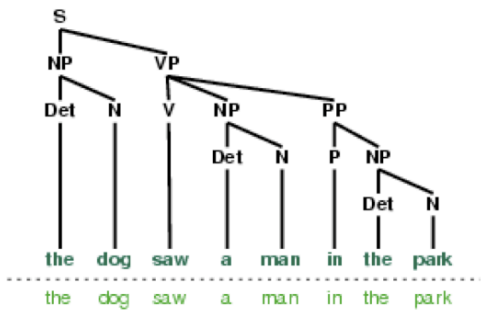


3. Matching *the*

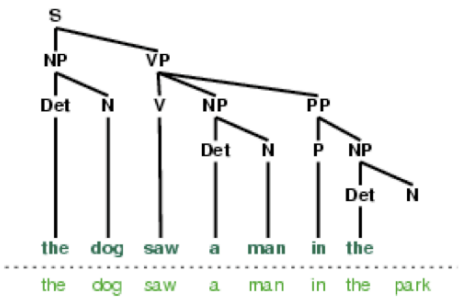
4. Cannot match *man*



5. Completed parse



6. Backtracking



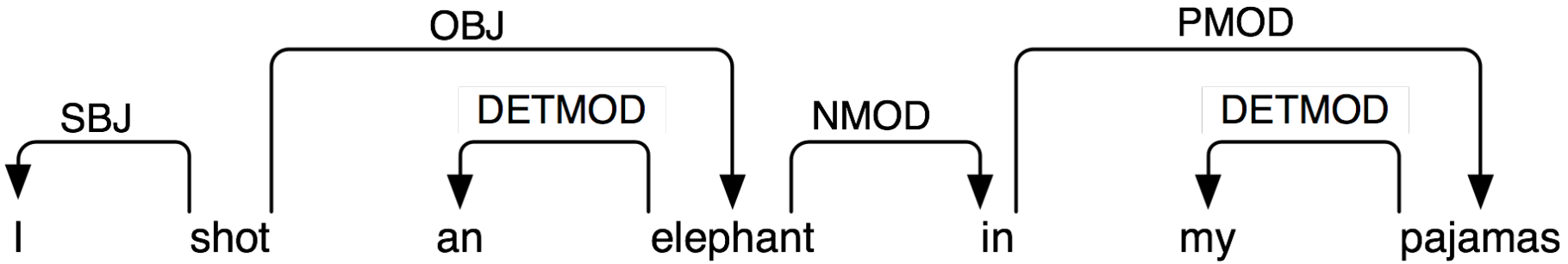
Some problems

- Left-recursive rules (*NP -> NP PP*) send the parser into an infinite loop
- Extra stuff is discarded when constituents are tossed
- Top down parsing is unnecessarily predictive

Bottom Up Parsing

- Start with the words and then build parse trees.
- Shift-reduce parser
- Problems:
  - End up with small forests but no tree
  - Still seems inefficient

A Different Way



- A dependency parser makes a graph
- Focuses on how words depend on each other