CS 343: Artificial Intelligence
Natural Language Processing

Raymond J. Mooney

University of Texas at Austin

Natural Language Processing

- NLP is the branch of computer science focused on developing systems that allow computers to communicate with people using everyday language.
- Also called Computational Linguistics
  - Also concerns how computational methods can aid the understanding of human language

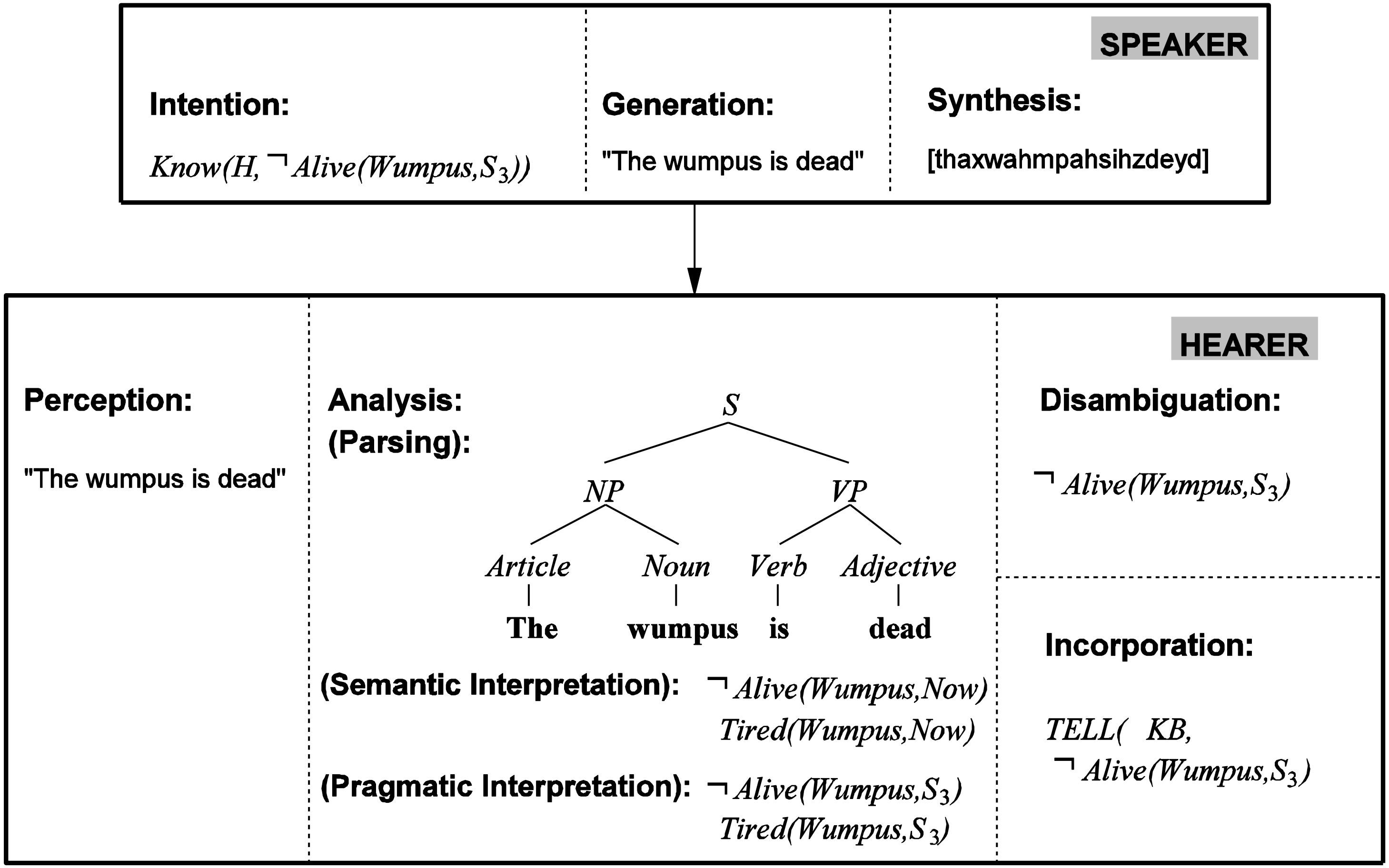
Communication

- The goal in the production and comprehension of natural language is communication.
- Communication for the speaker:
  - **Intention**: Decide when and what information should be transmitted (a.k.a. *strategic generation*). May require planning and reasoning about agents' goals and beliefs.
  - **Generation**: Translate the information to be communicated (in internal logical representation or "language of thought") into string of words in desired natural language (a.k.a. *tactical generation*).
  - **Synthesis**: Output the string in desired modality, text or speech.

Communication (cont)

- Communication for the hearer:
  - **Perception**: Map input modality to a string of words, e.g. *optical character recognition* (OCR) or *speech recognition*.
  - **Analysis**: Determine the information content of the string.
    - **Syntactic interpretation (parsing):** Find the correct parse tree showing the phrase structure of the string.
    - **Semantic Interpretation**: Extract the (literal) meaning of the string (*logical form*).
    - **Pragmatic Interpretation**: Consider effect of the overall context on altering the literal meaning of a sentence.
  - **Incorporation**: Decide whether or not to believe the content of the string and add it to the KB.

Communication (cont)



SPEAKER

Intention:

$Know(H, \neg Alive(Wumpus, S_3))$

Generation:

"The wumpus is dead"

Synthesis:

[thaxwahmpahsihzdeyd]

HEARER

Perception:

"The wumpus is dead"

Analysis:
(Parsing):

(Semantic Interpretation): $\neg Alive(Wumpus, Now)$
$Tired(Wumpus, Now)$

(Pragmatic Interpretation): $\neg Alive(Wumpus, S_3)$
$Tired(Wumpus, S_3)$

Disambiguation:

$\neg Alive(Wumpus, S_3)$

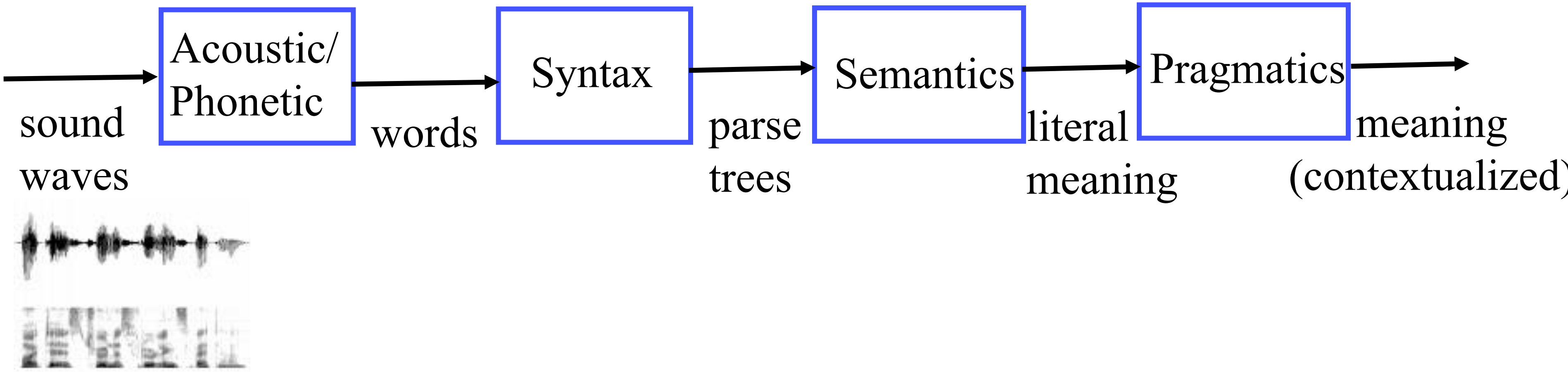
Incorporation:

$TELL(\ KB,$
$\neg Alive(Wumpus, S_3))$

Syntax, Semantic, Pragmatics

- Syntax concerns the proper ordering of words and its affect on meaning.
  - The dog bit the boy.
  - The boy bit the dog.
  - * Bit boy dog the the.
  - Colorless green ideas sleep furiously.
- Semantics concerns the (literal) meaning of words, phrases, and sentences.
  - "plant" as a photosynthetic organism
  - "plant" as a manufacturing facility
  - "plant" as the act of sowing
- Pragmatics concerns the overall communicative and social context and its effect on interpretation.
  - The ham sandwich wants another beer. (co-reference, anaphora)
  - John thinks vanilla. (ellipsis)

Modular Comprehension



sound waves

Acoustic/ Phonetic

words

Syntax

parse trees

Semantics

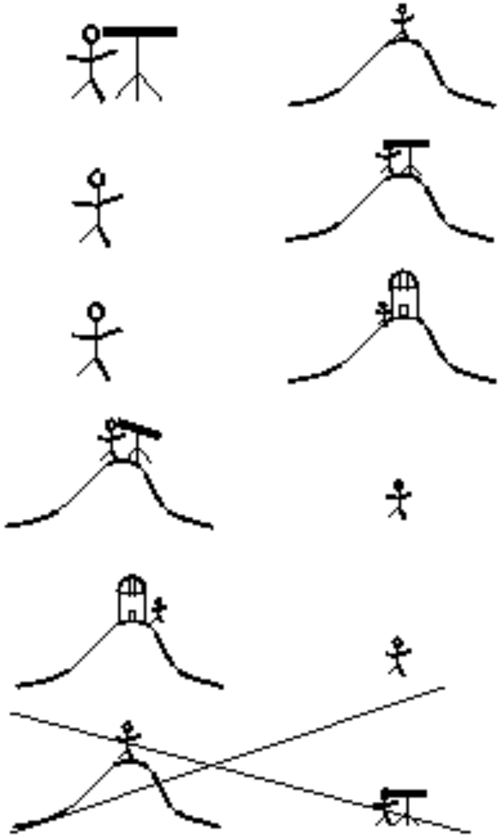
literal meaning

Pragmatics

meaning (contextualized)

Ambiguity

- Natural language is highly ambiguous and must be *disambiguated.*
  - I saw the man on the hill with a telescope.
  - I saw the Grand Canyon flying to LA.
  - Time flies like an arrow.
  - Horse flies like a sugar cube.
  - Time runners like a coach.
  - Time cars like a Porsche.



Ambiguity is Ubiquitous

- Speech Recognition
  - "recognize speech" vs. "wreck a nice beach"
  - "youth in Asia" vs. "euthanasia"
- Syntactic Analysis
  - "I ate spaghetti with chopsticks" vs. "I ate spaghetti with meatballs."
- Semantic Analysis
  - "The dog is in the pen." vs. "The ink is in the pen."
  - "I put the plant in the window" vs. "Ford put the plant in Mexico"
- Pragmatic Analysis
  - **From "The Pink Panther Strikes Again":**
  - **Clouseau**: Does your dog bite?
    Hotel Clerk: No.
    Clouseau: [*bowing down to pet the dog*] Nice doggie.
    [*Dog barks and bites Clouseau in the hand*]
    Clouseau: I thought you said your dog did not bite!
    Hotel Clerk: That is not my dog.

Ambiguity is Explosive

- Ambiguities compound to generate enormous numbers of possible interpretations.
- In English, a sentence ending in *n* prepositional phrases has *over* 2^n syntactic interpretations (cf. Catalan numbers).
  - "I saw the man with the telescope": 2 parses
  - "I saw the man on the hill with the telescope.": 5 parses
  - "I saw the man on the hill in Texas with the telescope": 14 parses
  - "I saw the man on the hill in Texas with the telescope at noon.": 42 parses
  - "I saw the man on the hill in Texas with the telescope at noon on Monday" 132 parses

Humor and Ambiguity

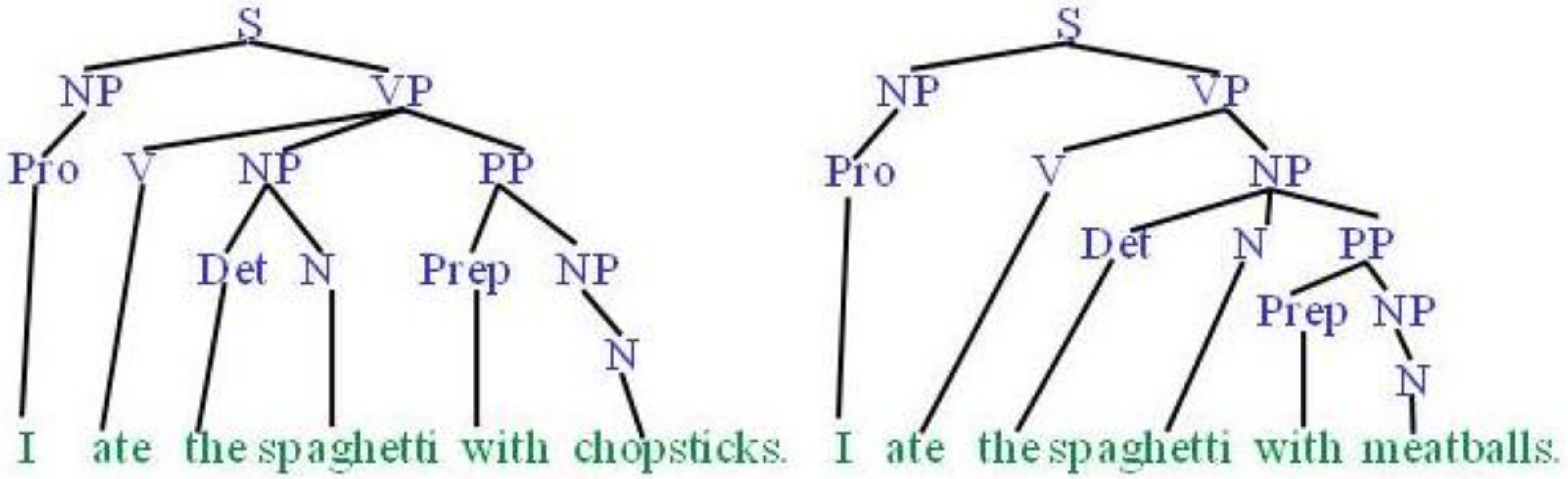
- Many jokes rely on the ambiguity of language:
  - Groucho Marx: One morning I shot an elephant in my pajamas. How he got into my pajamas, I'll never know.
  - She criticized my apartment, so I knocked her flat.
  - Noah took all of the animals on the ark in pairs. Except the worms, they came in apples.
  - Policeman to little boy: "We are looking for a thief with a bicycle." Little boy: "Wouldn't you be better using your eyes."
  - Why is the teacher wearing sun-glasses. Because the class is so bright.

Natural Languages vs. Computer Languages

- Ambiguity is the primary difference between natural and computer languages.
- Formal programming languages are designed to be unambiguous, i.e. they can be defined by a grammar that produces a unique parse for each sentence in the language.
- Programming languages are also designed for efficient (deterministic) parsing, i.e. they are deterministic context-free languages (DCLFs).
  - A sentence in a DCFL can be parsed in O(*n*) time where *n* is the length of the string.

Syntactic Parsing

- Produce the correct syntactic parse tree for a sentence.


S
NP
VP
Pro
V
NP
PP
Det
N
Prep
NP
N
I ate the spaghetti with chopsticks.
S
NP
VP
Pro
V
NP
Det
N
PP
Prep
NP
N
I ate the spaghetti with meatballs.

Context Free Grammars (CFG)

- *N* a set of ***non-terminal symbols*** (or ***variables***)
- Σ a set of ***terminal symbols*** (disjoint from *N*)
- *R* a set of ***productions*** or ***rules*** of the form A→β, where A is a non-terminal and β is a string of symbols from $(\Sigma \cup N)^*$
- S, a designated non-terminal called the ***start symbol***

Simple CFG for ATIS English

Grammar

S → NP VP

S → Aux NP VP

S → VP

NP → Pronoun

NP → Proper-Noun

NP → Det Nominal

Nominal → Noun

Nominal → Nominal Noun

Nominal → Nominal PP

VP → Verb

VP → Verb NP

VP → VP PP

PP → Prep NP

Lexicon

Det → the | a | that | this

Noun → book | flight | meal | money

Verb → book | include | prefer

Pronoun → I | he | she | me

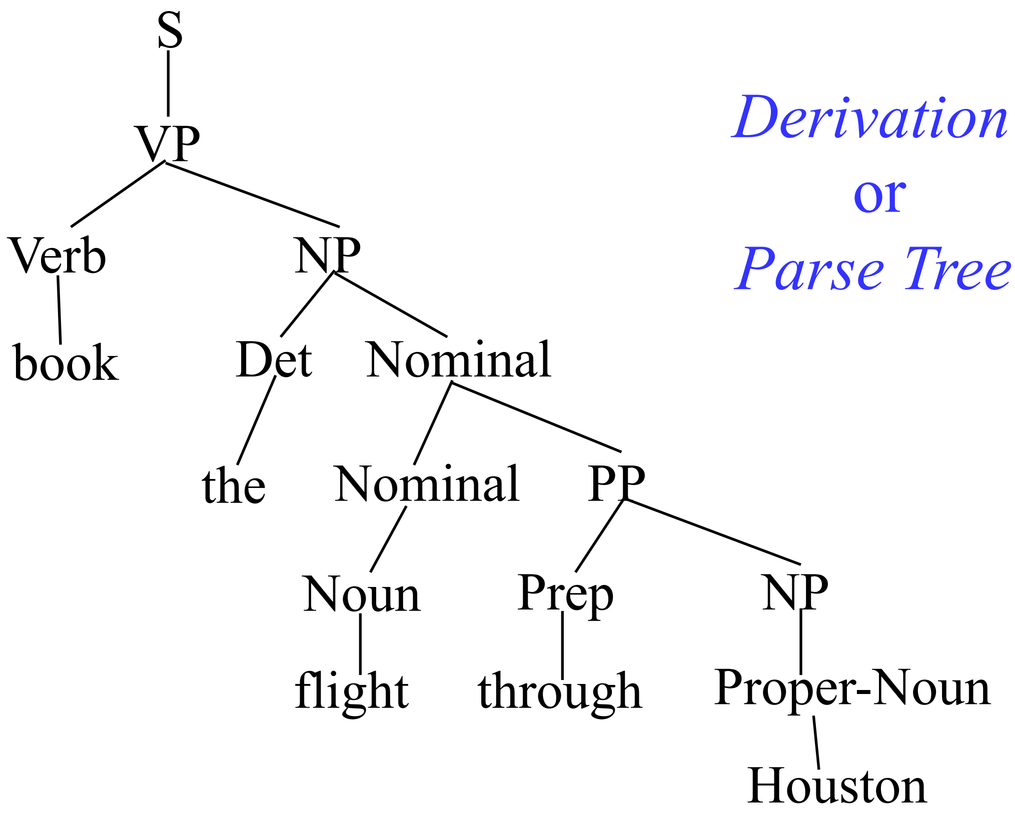
Proper-Noun → Houston | NWA

Aux → does

Prep → from | to | on | near | through

Sentence Generation

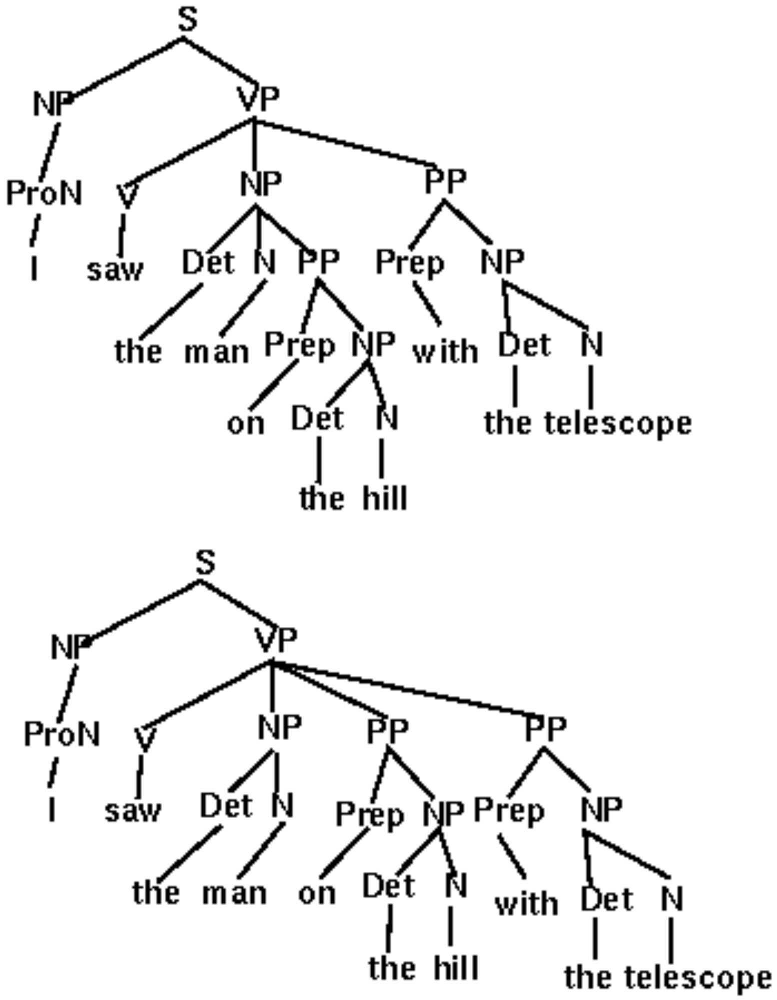
- Sentences are generated by recursively rewriting the start symbol using the productions until only terminals symbols remain.



Derivation
or
Parse Tree

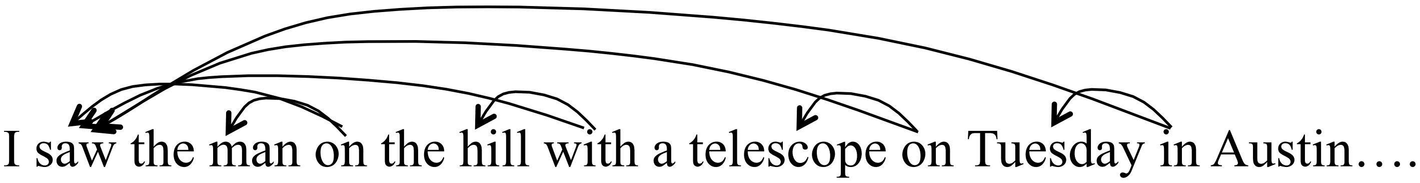
Parse Trees and Syntactic Ambiguity

- If a sentence has more than one possible derivation (parse tree) it is said to be *syntactically ambiguous*.



Prepositional Phrase Attachment Explosion

- A transitive English sentence ending in *m* prepositional phrases has *at least* 2^m parses.



I saw the man on the hill with a telescope on Tuesday in Austin….

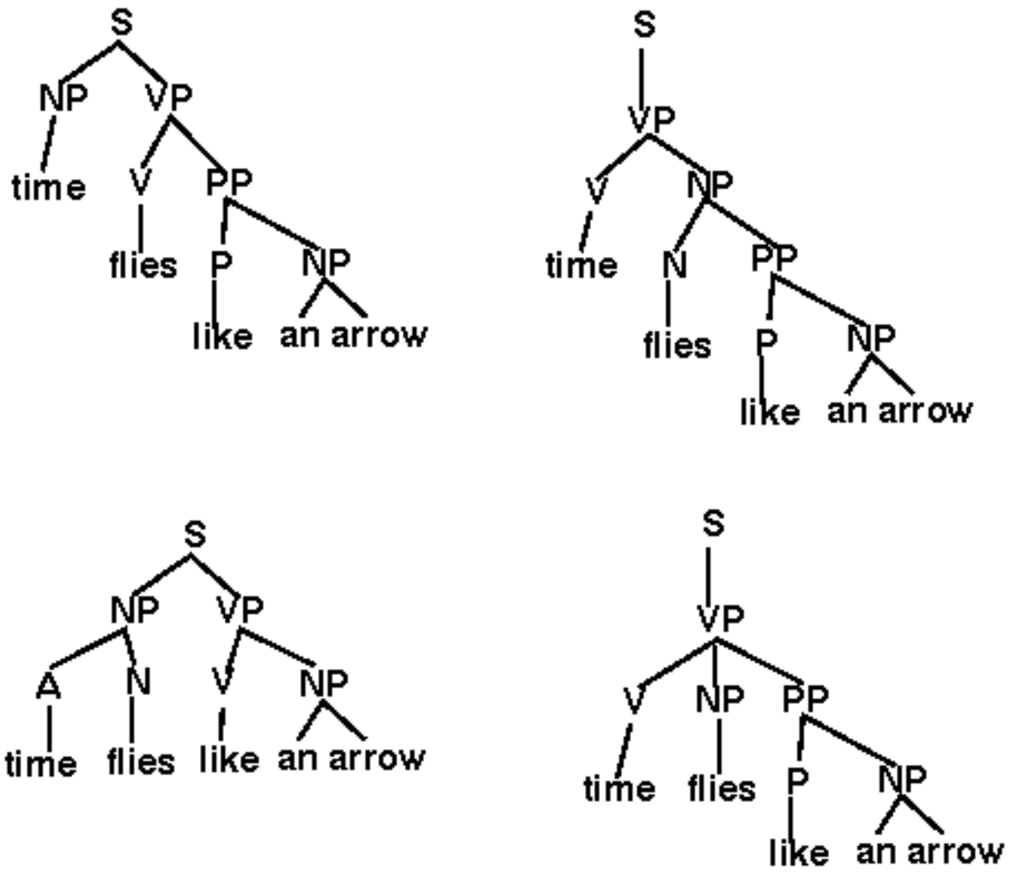
- The exact number of parses is given by the *Catalan numbers* (where n=m+1)

$$\binom{2n}{n} - \binom{2n}{n-1} \approx \frac{4^n}{n^{3/2}\sqrt{\pi}}$$

1, 2, 5, 14, 132, 429, 1430, 4862, 16796, ……

Spurious Ambiguity

- Most parse trees of most NL sentences make no sense.

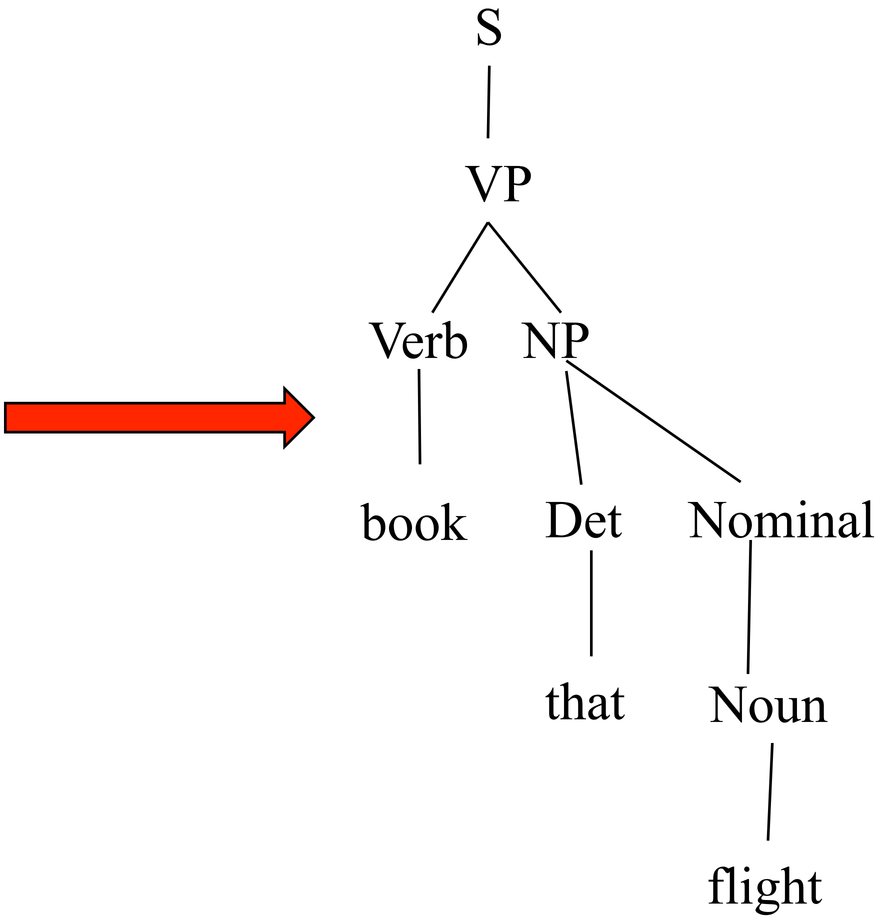


Parsing

- Given a string of non-terminals and a CFG, determine if the string can be generated by the CFG.
  - Also return a parse tree for the string
  - Also return all possible parse trees for the string
- Must search space of derivations for one that derives the given string.
  - **Top-Down Parsing**: Start searching space of derivations for the start symbol.
  - **Bottom-up Parsing**: Start search space of reverse deivations from the terminal symbols in the string.

Parsing Example

book that flight



S

VP

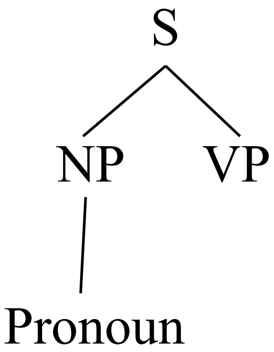
Verb NP

book Det Nominal

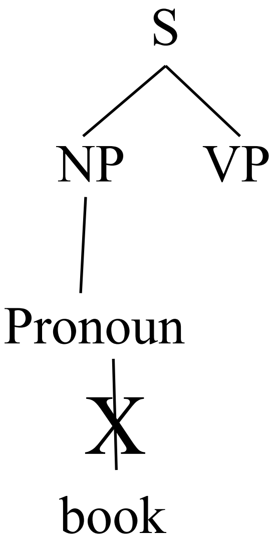
that Noun

flight

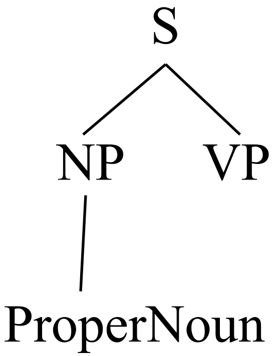
Top Down Parsing



Top Down Parsing



Top Down Parsing

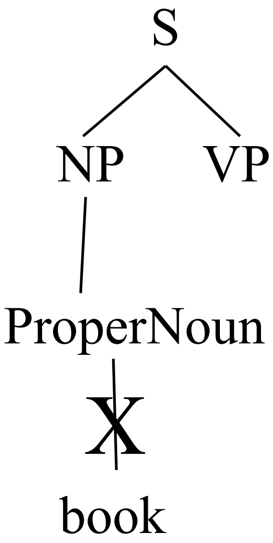


S

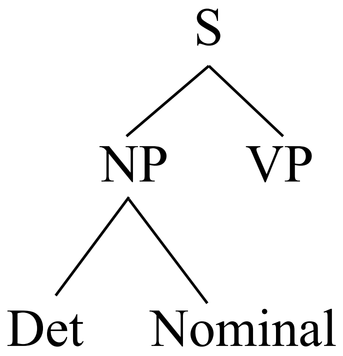
NP VP

ProperNoun

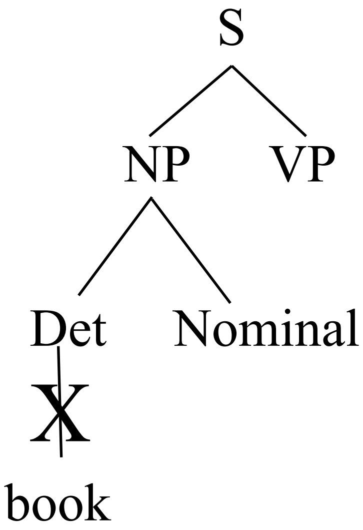
Top Down Parsing



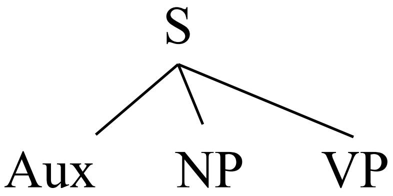
Top Down Parsing



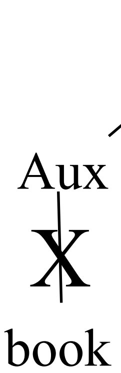
Top Down Parsing


S
NP
VP
Det
Nominal
X
book

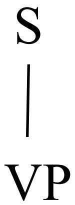
Top Down Parsing



Top Down Parsing



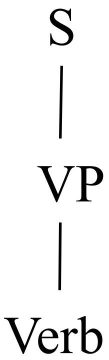
Top Down Parsing



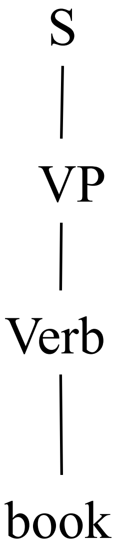
S

VP

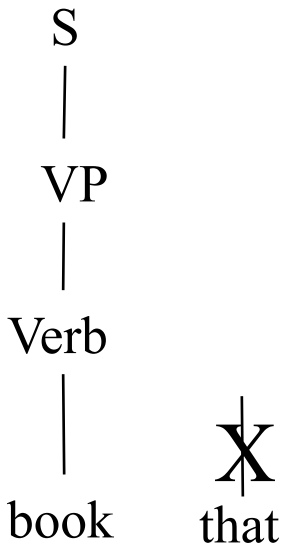
Top Down Parsing



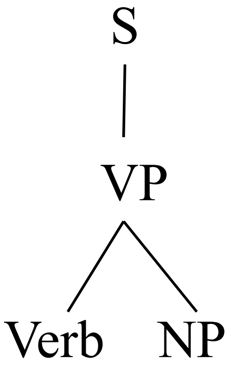
Top Down Parsing



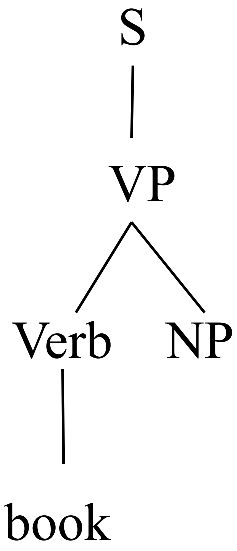
Top Down Parsing



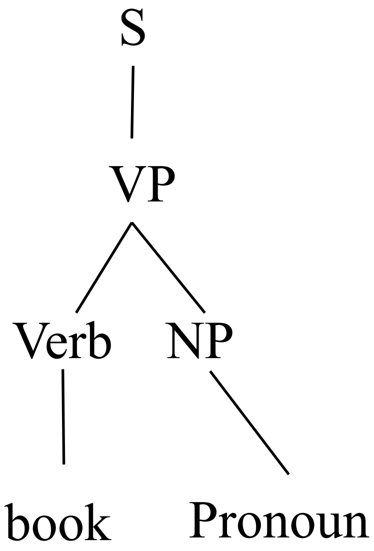
Top Down Parsing



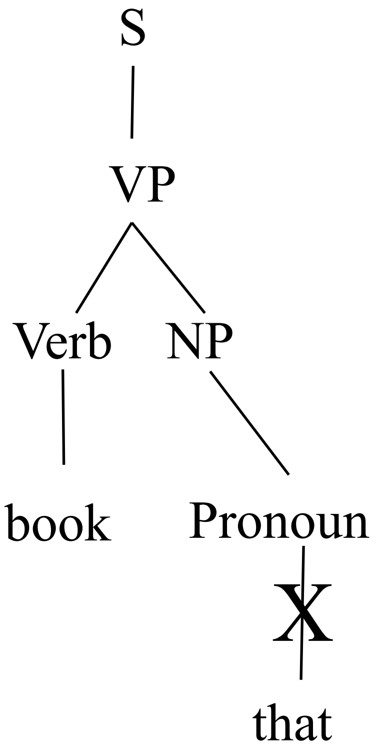
Top Down Parsing



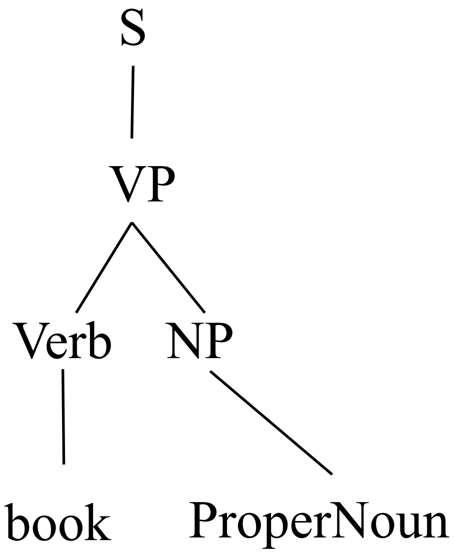
Top Down Parsing



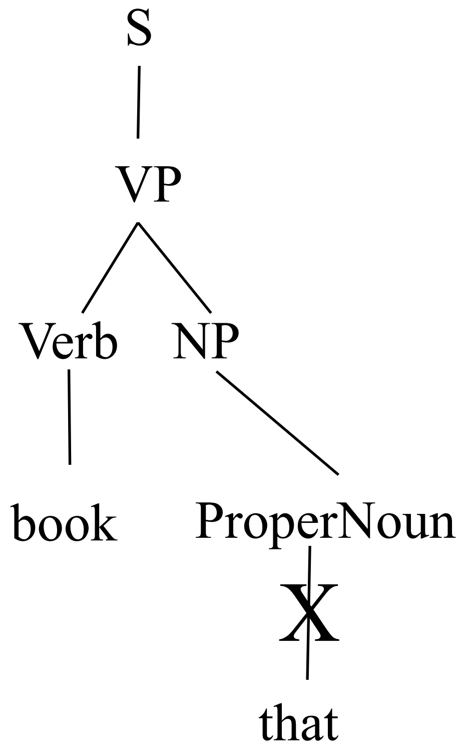
Top Down Parsing



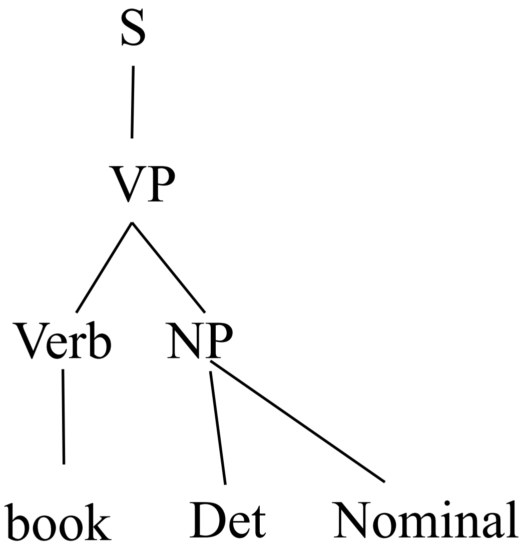
Top Down Parsing



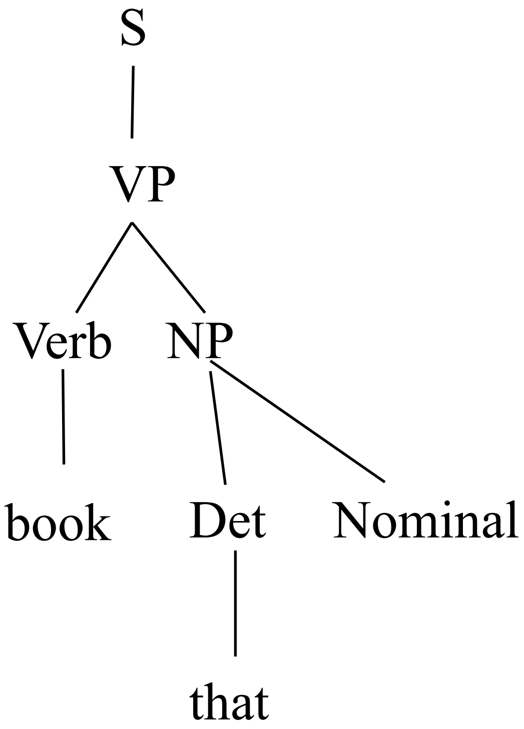
Top Down Parsing



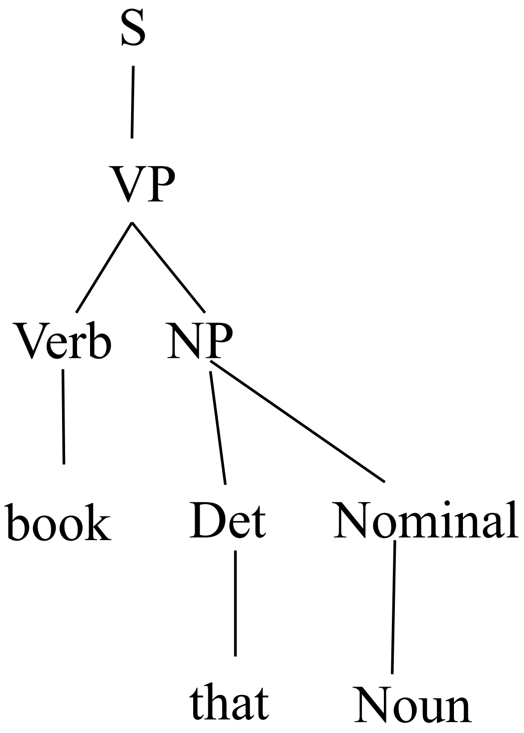
Top Down Parsing



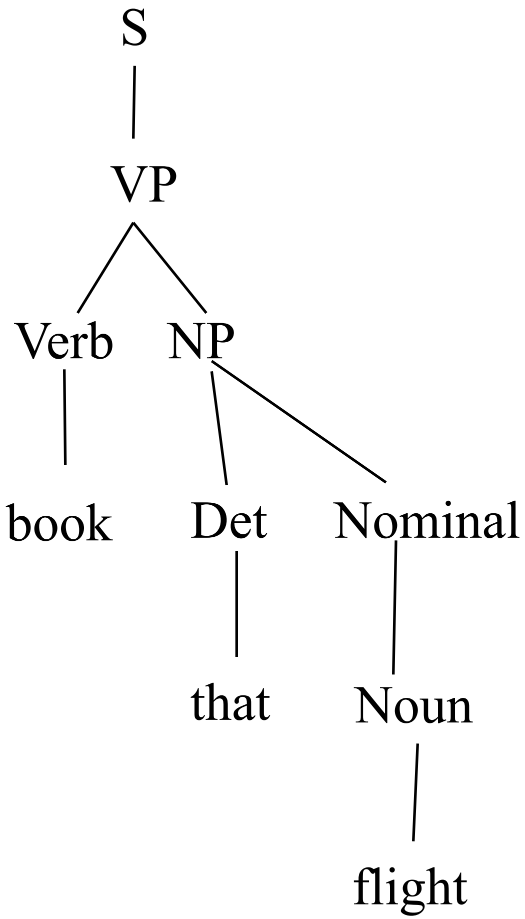
Top Down Parsing



Top Down Parsing



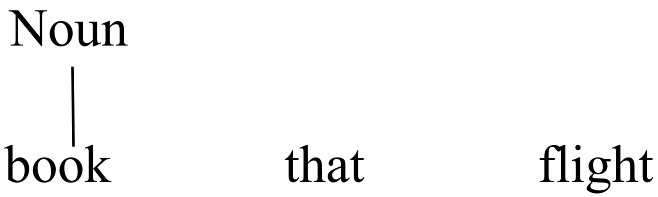
Top Down Parsing



Bottom Up Parsing

book that flight

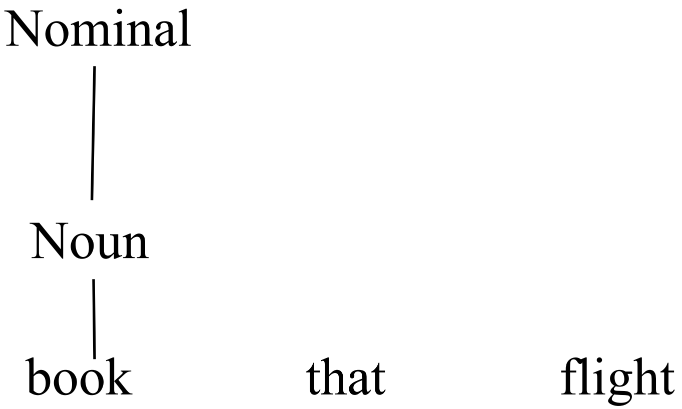
Bottom Up Parsing



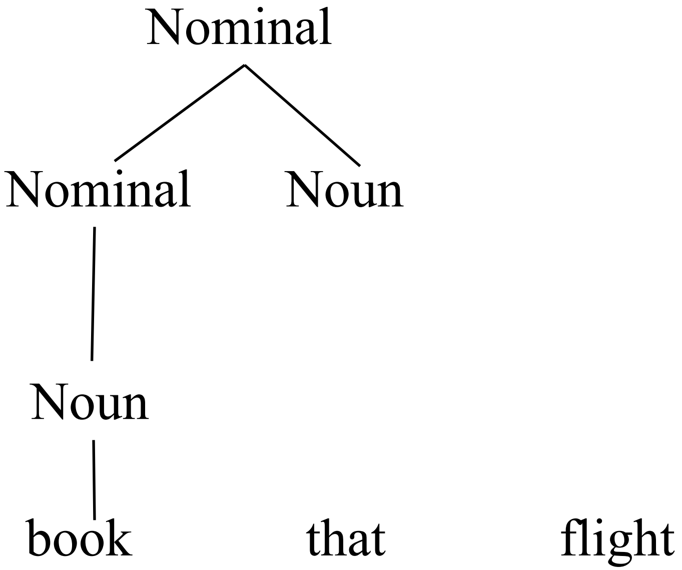
Noun

book that flight

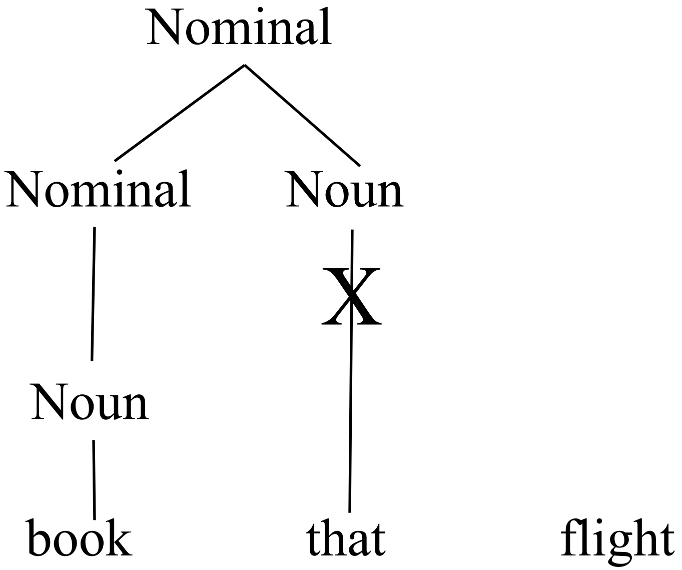
Bottom Up Parsing



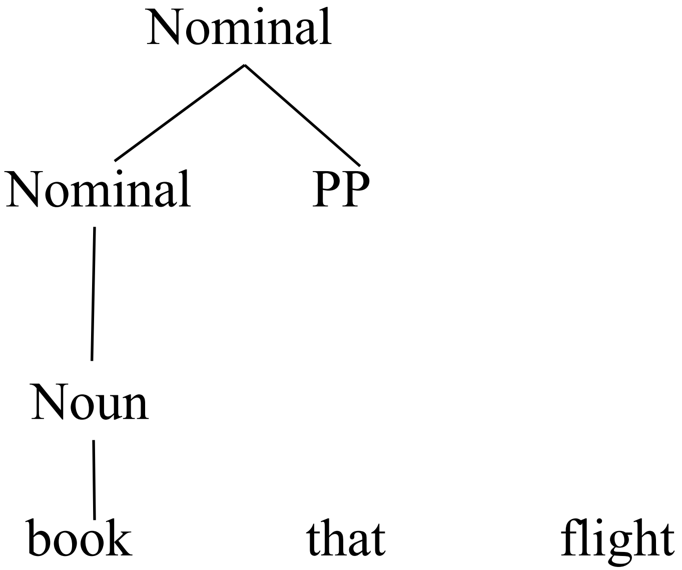
Bottom Up Parsing



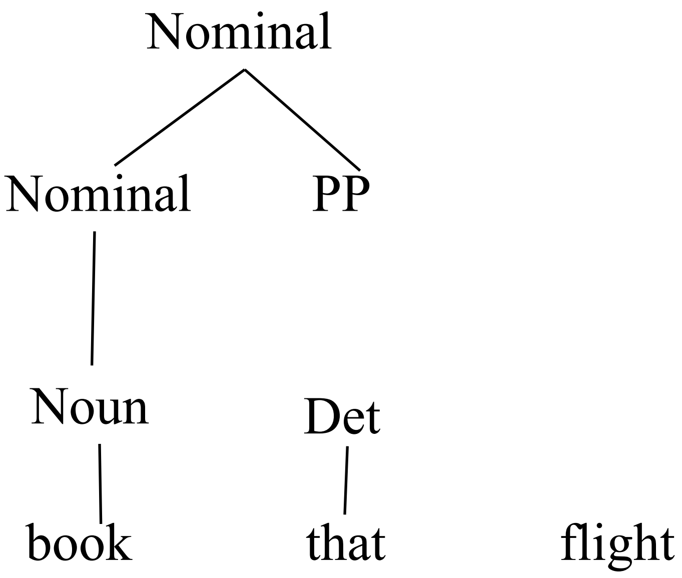
Bottom Up Parsing



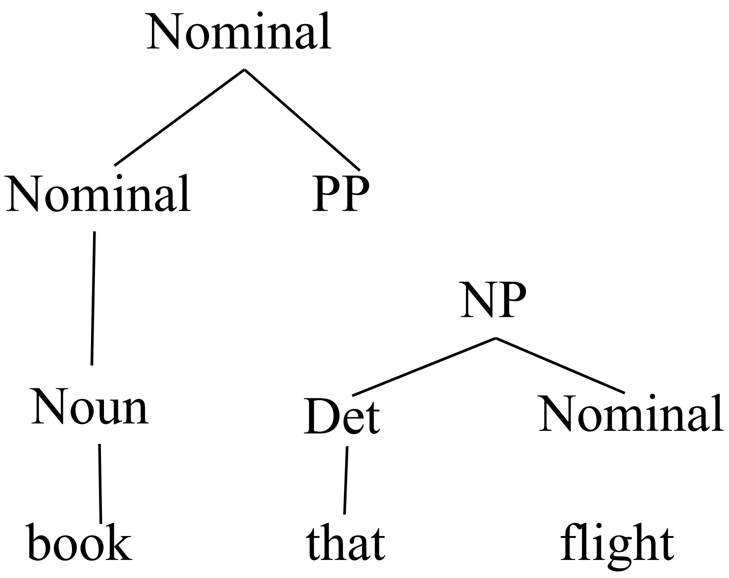
Bottom Up Parsing



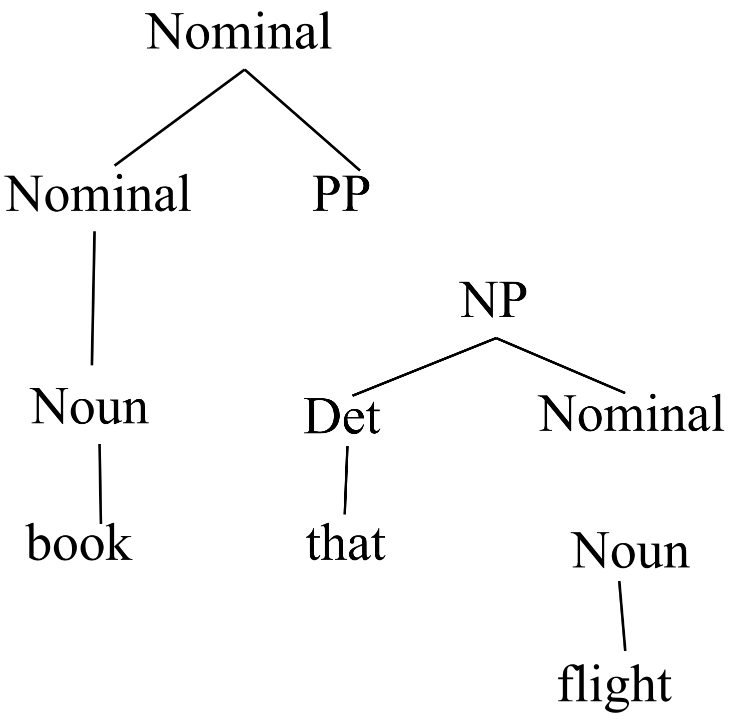
Bottom Up Parsing



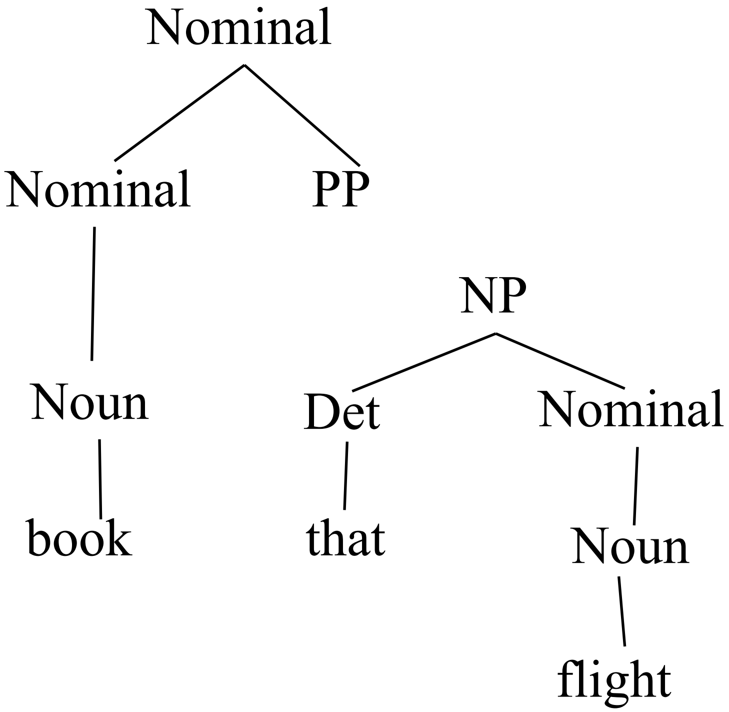
Bottom Up Parsing



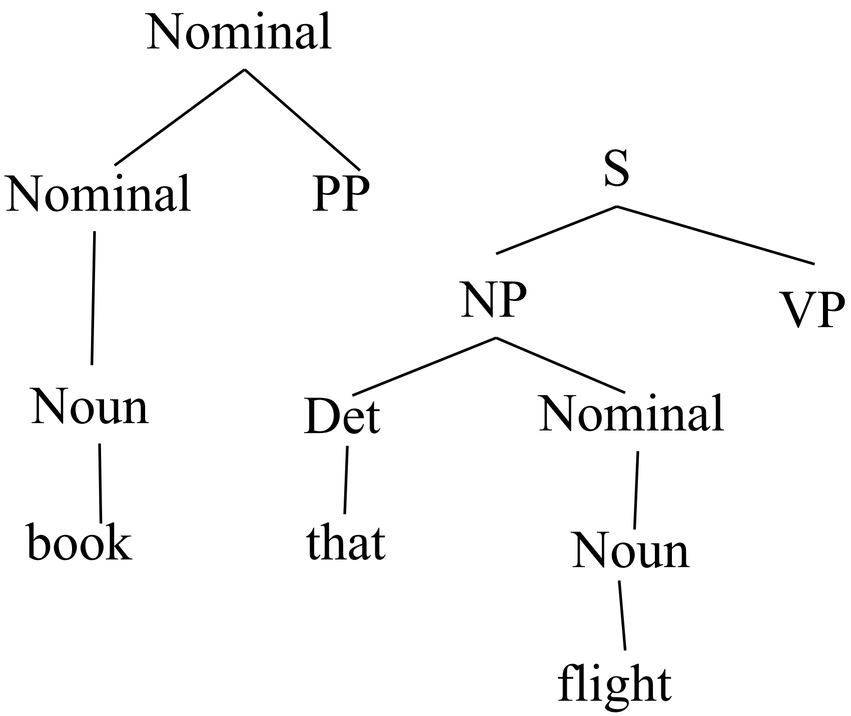
Bottom Up Parsing



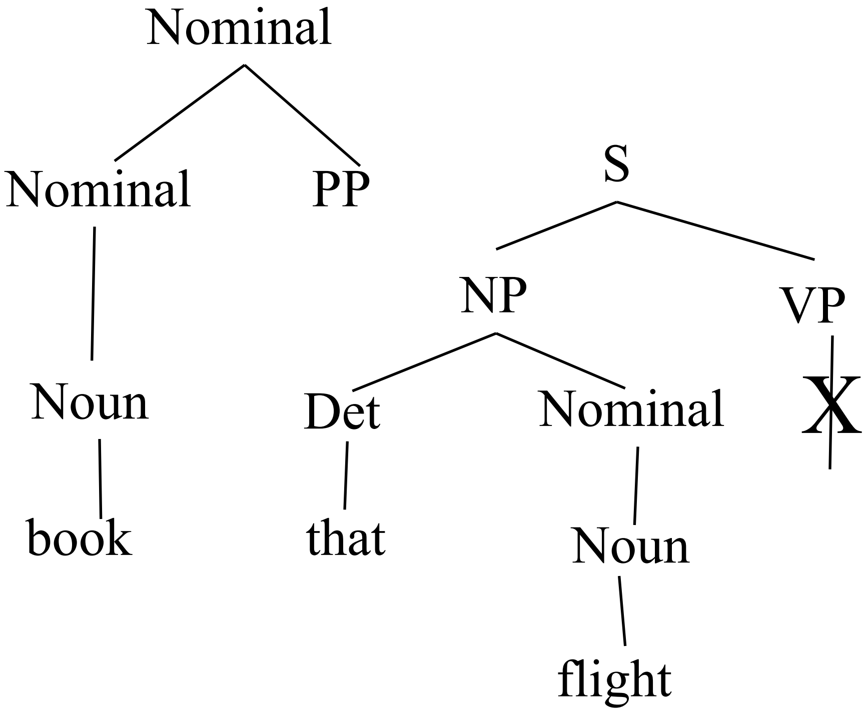
Bottom Up Parsing



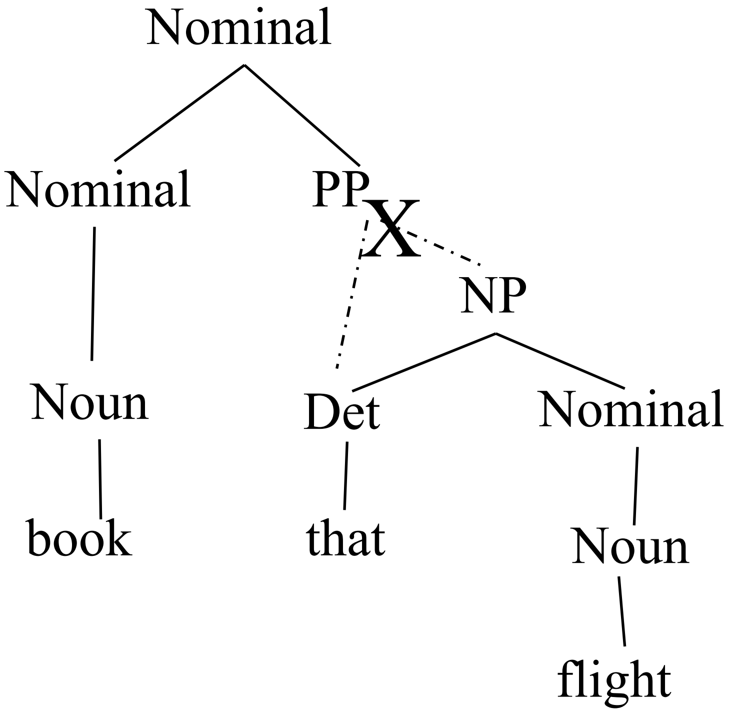
Bottom Up Parsing



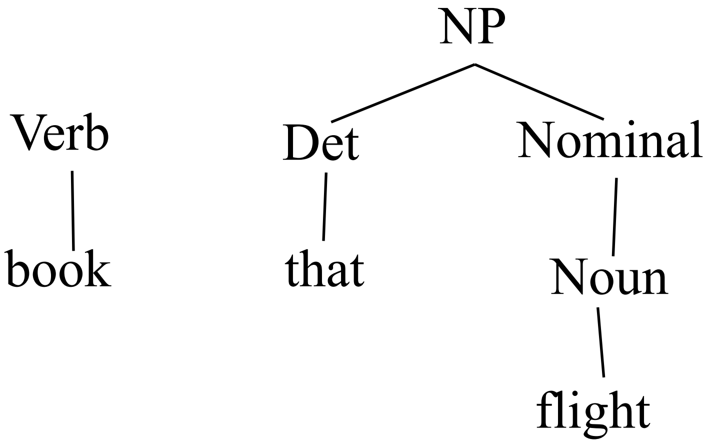
Bottom Up Parsing



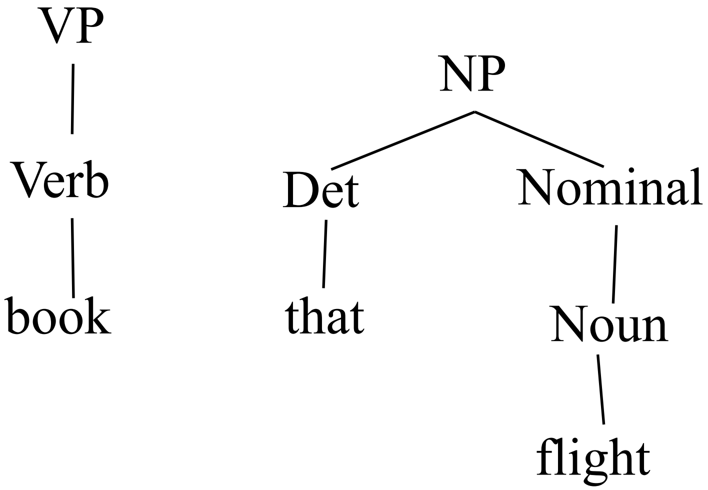
Bottom Up Parsing



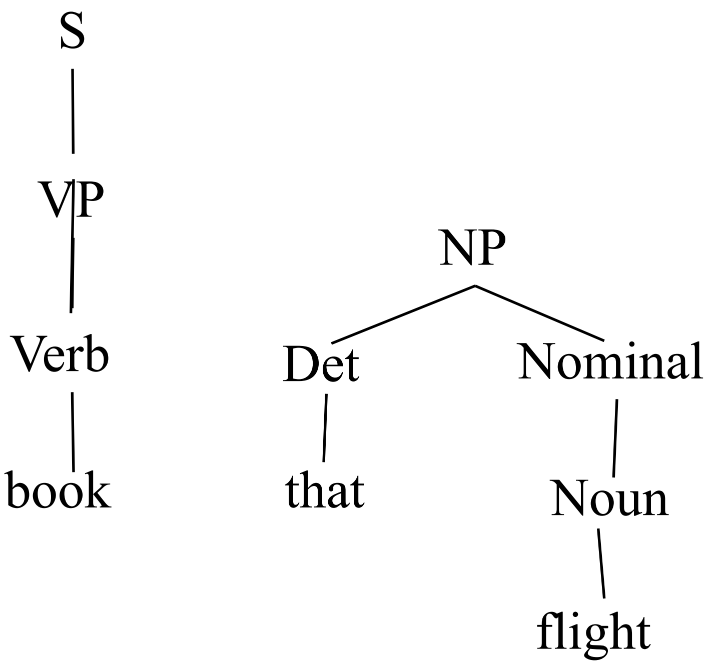
Bottom Up Parsing



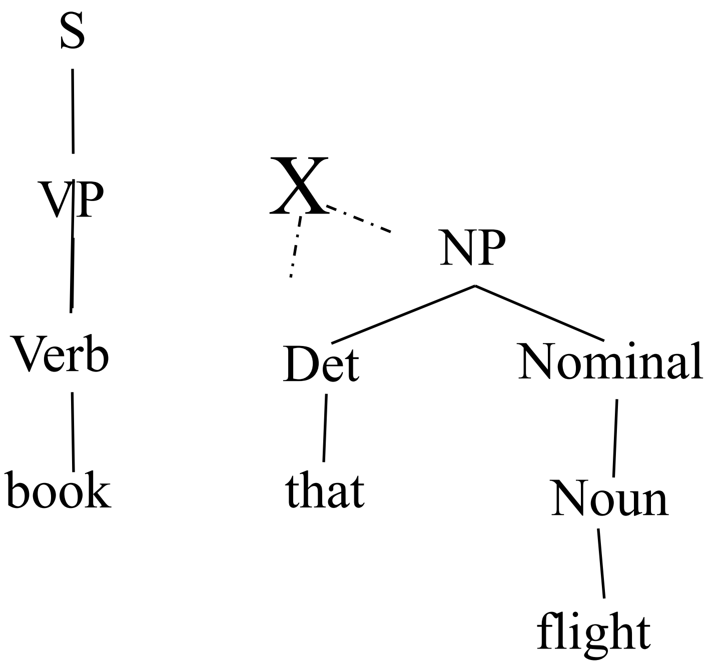
Bottom Up Parsing



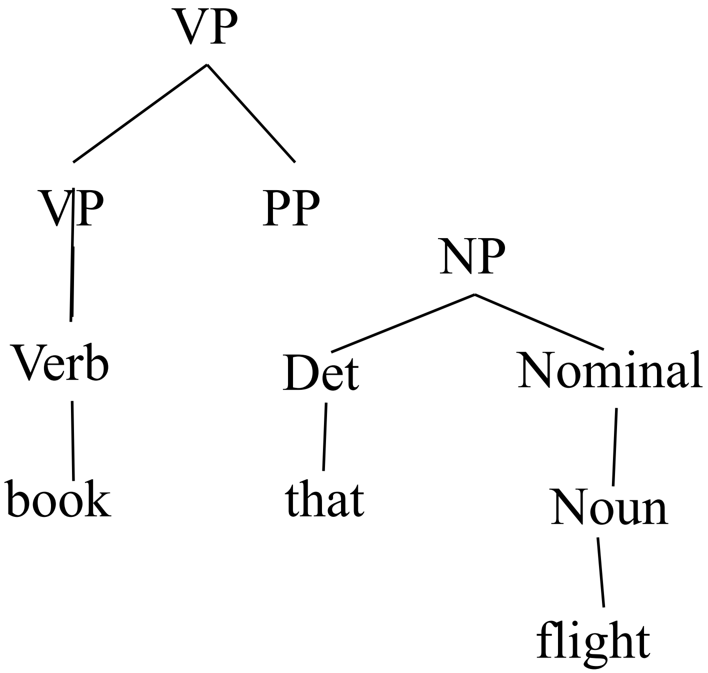
Bottom Up Parsing



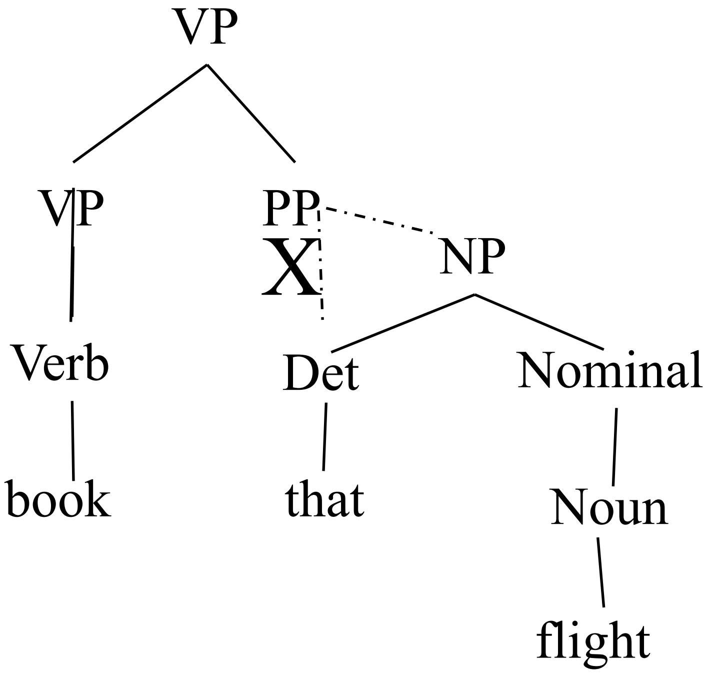
Bottom Up Parsing



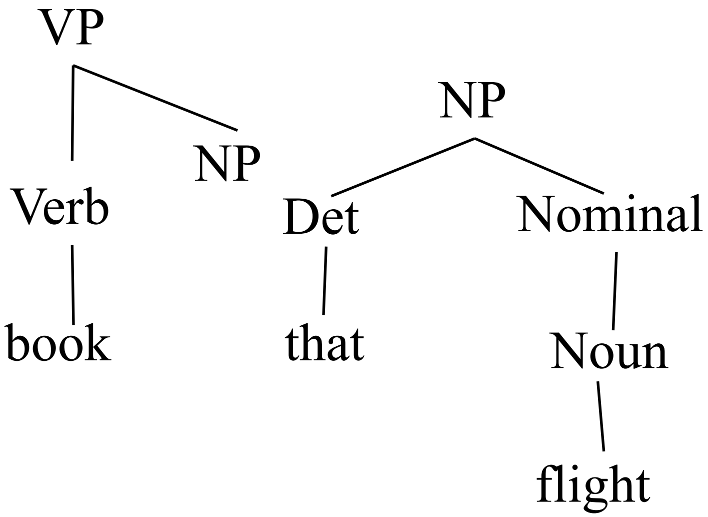
Bottom Up Parsing



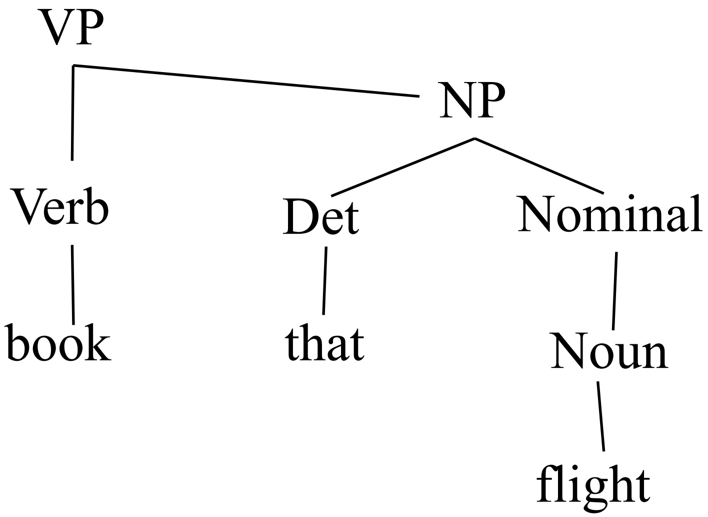
Bottom Up Parsing



Bottom Up Parsing



Bottom Up Parsing



VP

NP

Verb

Det

Nominal

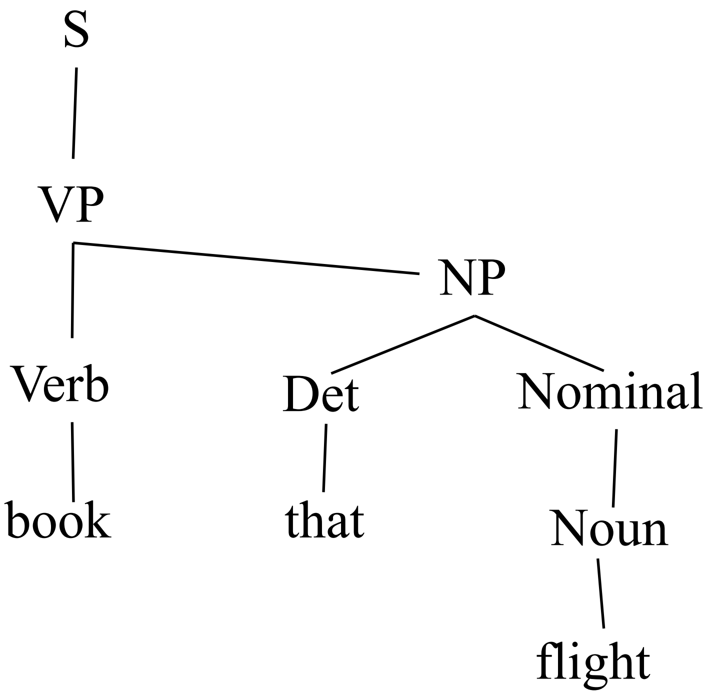
book

that

Noun

flight

Bottom Up Parsing



Top Down vs. Bottom Up

- Top down never explores options that will not lead to a full parse, but can explore many options that never connect to the actual sentence.
- Bottom up never explores options that do not connect to the actual sentence but can explore options that can never lead to a full parse.
- Relative amounts of wasted search depend on how much the grammar branches in each direction.

Syntactic Parsing & Ambiguity

- Just produces all possible parse trees.
- Does not address the important issue of ambiguity resolution.

Statistical Parsing

- Statistical parsing uses a probabilistic model of syntax in order to assign probabilities to each parse tree.
- Provides principled approach to resolving syntactic ambiguity.
- Allows supervised learning of parsers from tree-banks of parse trees provided by human linguists.
- Also allows unsupervised learning of parsers from unannotated text, but the accuracy of such parsers has been limited.

Probabilistic Context Free Grammar (PCFG)

- A PCFG is a probabilistic version of a CFG where each production has a probability.
- Probabilities of all productions rewriting a given non-terminal must add to 1, defining a distribution for each non-terminal.
- String generation is now probabilistic where production probabilities are used to non-deterministically select a production for rewriting a given non-terminal.

Simple PCFG for ATIS English

Grammar	Prob	
S → NP VP	0.8	
S → Aux NP VP	0.1	+ 1.0
S → VP	0.1	
NP → Pronoun	0.2	
NP → Proper-Noun	0.2	+ 1.0
NP → Det Nominal	0.6	
Nominal → Noun	0.3	
Nominal → Nominal Noun	0.2	+ 1.0
Nominal → Nominal PP	0.5	
VP → Verb	0.2	
VP → Verb NP	0.5	+ 1.0
VP → VP PP	0.3	
PP → Prep NP	1.0	

Lexicon

Det → the | a | that | this
0.6 0.2 0.1 0.1

Noun → book | flight | meal | money
0.1 0.5 0.2 0.2

Verb → book | include | prefer
0.5 0.2 0.3

Pronoun → I | he | she | me
0.5 0.1 0.1 0.3

Proper-Noun → Houston | NWA
0.8 0.2

Aux → does
1.0

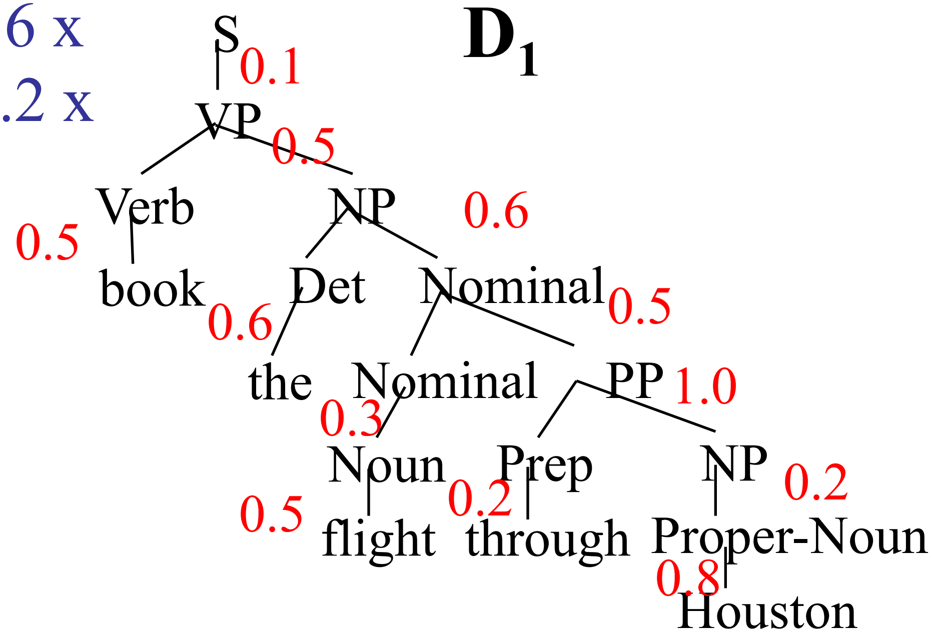
Prep → from | to | on | near | through
0.25 0.25 0.1 0.2 0.2

Sentence Probability

- Assume productions for each node are chosen independently.
- Probability of derivation is the product of the probabilities of its productions.

$$P(D_1) = 0.1 \times 0.5 \times 0.5 \times 0.6 \times 0.6 \times 0.5 \times 0.3 \times 1.0 \times 0.2 \times 0.2 \times 0.5 \times 0.8$$

$$= 0.0000216$$

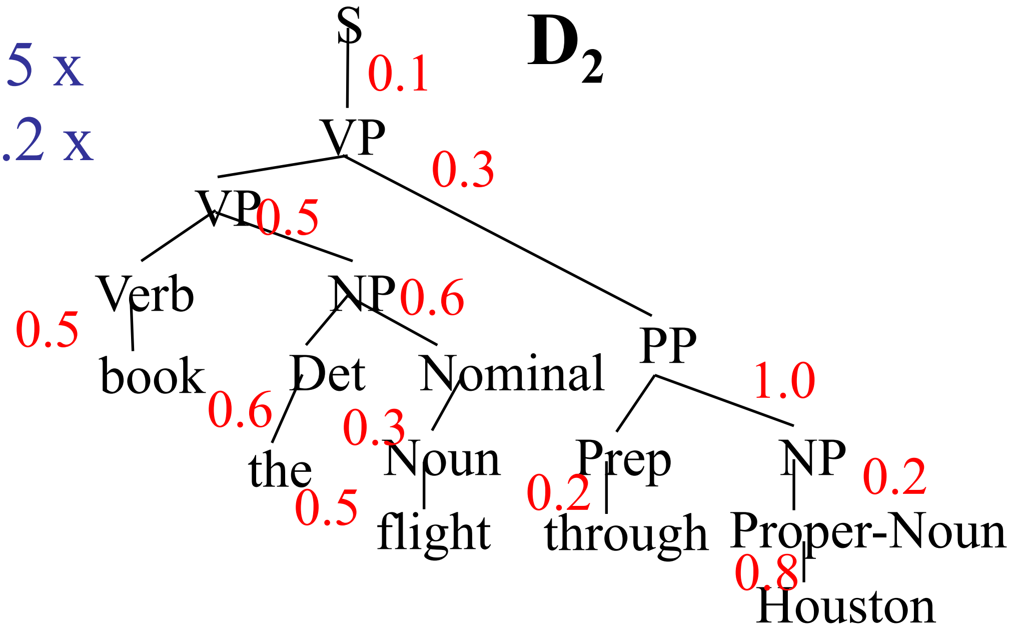


Syntactic Disambiguation

- Resolve ambiguity by picking most probable parse tree.

$$P(D_2) = 0.1 \times 0.3 \times 0.5 \times 0.6 \times 0.5 \times 0.6 \times 0.3 \times 1.0 \times 0.5 \times 0.2 \times 0.2 \times 0.8$$

$$= 0.00001296$$



Sentence Probability

- Probability of a sentence is the sum of the probabilities of all of its derivations.

P("book the flight through Houston") =

$P(D_1) + P(D_2) = 0.0000216 + 0.00001296$

$= 0.00003456$

Three Useful PCFG Tasks

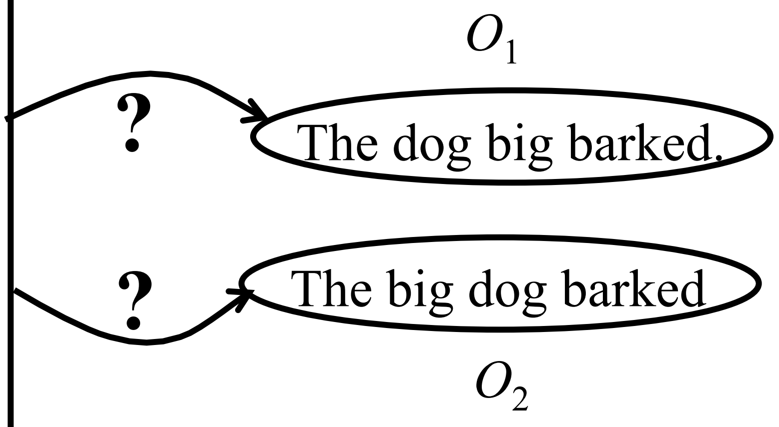
- Observation likelihood: To classify and order sentences.
- Most likely derivation: To determine the most likely parse tree for a sentence.
- Maximum likelihood training: To train a PCFG to fit empirical training data.

PCFG: Observation Likelihood

- What is the probability that a given string is produced by a given PCFG.
- Can use a PCFG as a language model to choose between alternative sentences for speech recognition or machine translation.

S → NP VP	0.9
S → VP	0.1
NP → Det A N	0.5
NP → NP PP	0.3
NP → PropN	0.2
A → ε	0.6
A → Adj A	0.4
PP → Prep NP	1.0
VP → V NP	0.7
VP → VP PP	0.3

English



O_1

The dog big barked.

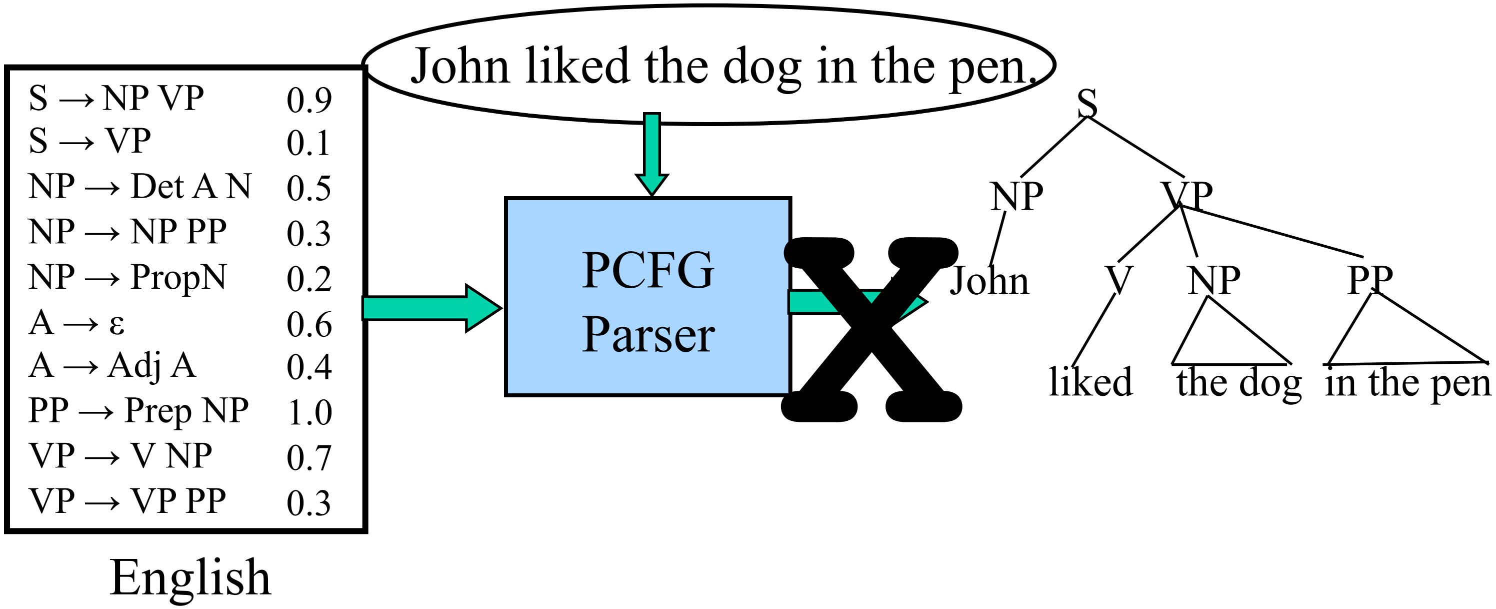
The big dog barked

O_2

P(O_2 | English) > P(O_1 | English) ?

PCFG: Most Likely Derivation

- What is the most probable derivation (parse tree) for a sentence.



S → NP VP	0.9
S → VP	0.1
NP → Det A N	0.5
NP → NP PP	0.3
NP → PropN	0.2
A → ε	0.6
A → Adj A	0.4
PP → Prep NP	1.0
VP → V NP	0.7
VP → VP PP	0.3

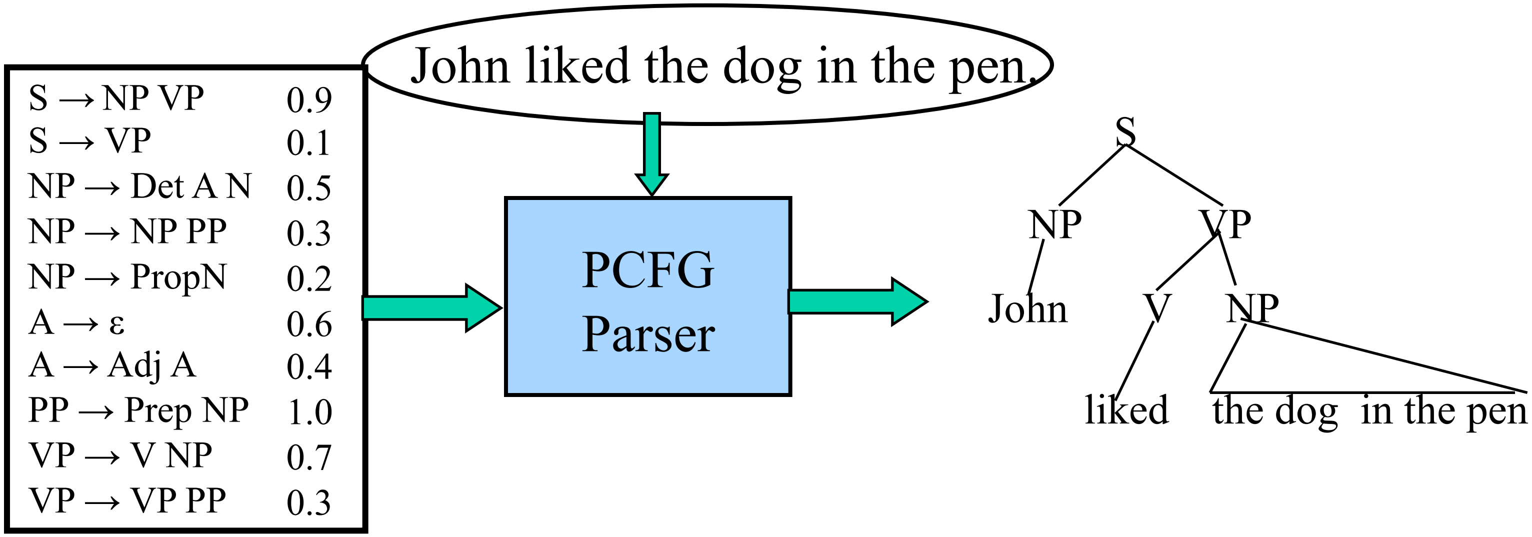
English

John liked the dog in the pen.

PCFG Parser

PCFG: Most Likely Derivation

- What is the most probable derivation (parse tree) for a sentence.



S → NP VP	0.9
S → VP	0.1
NP → Det A N	0.5
NP → NP PP	0.3
NP → PropN	0.2
A → ε	0.6
A → Adj A	0.4
PP → Prep NP	1.0
VP → V NP	0.7
VP → VP PP	0.3

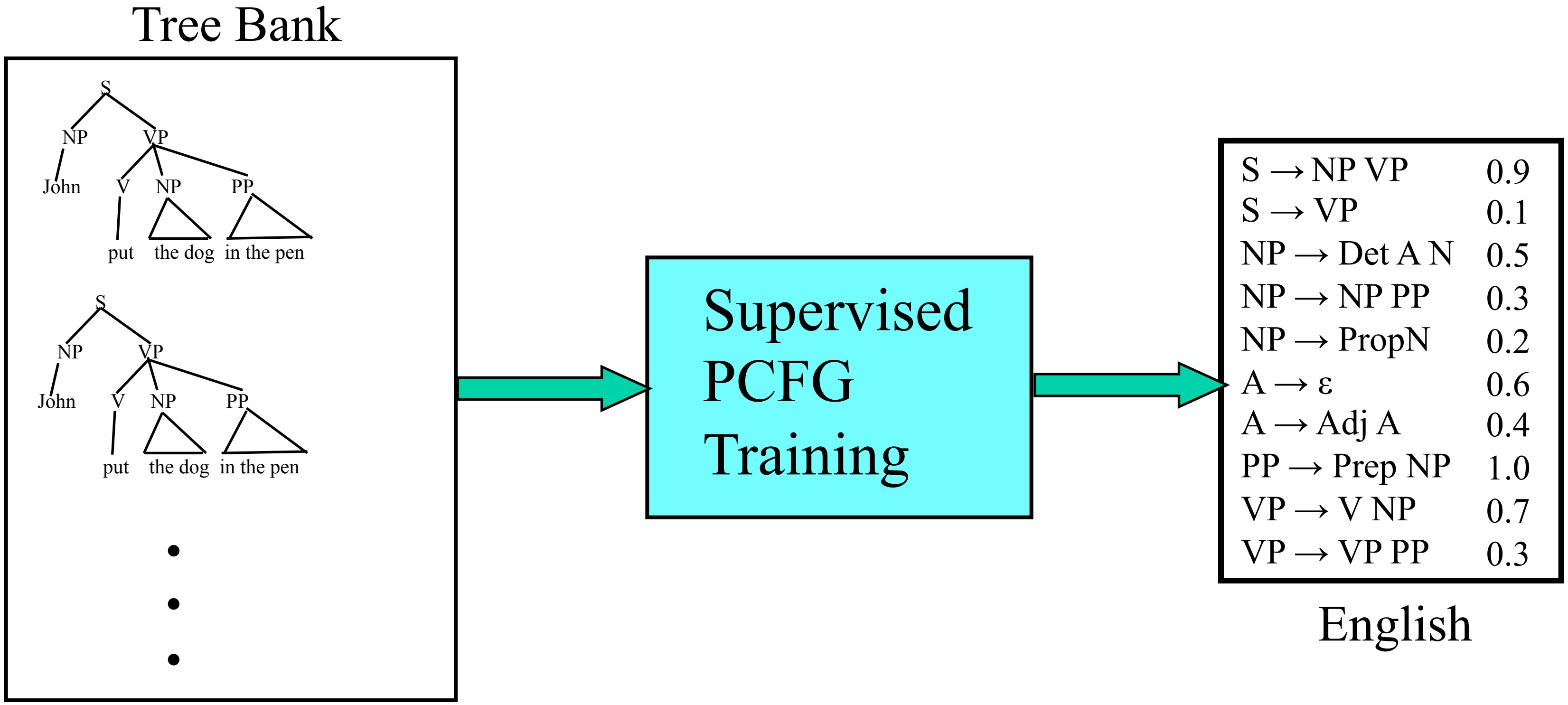
English

John liked the dog in the pen.

PCFG Parser

PCFG: Supervised Training

- If parse trees are provided for training sentences, a grammar and its parameters can be can all be estimated directly from counts accumulated from the tree-bank (with appropriate smoothing).



Supervised PCFG Training

S → NP VP	0.9
S → VP	0.1
NP → Det A N	0.5
NP → NP PP	0.3
NP → PropN	0.2
A → ε	0.6
A → Adj A	0.4
PP → Prep NP	1.0
VP → V NP	0.7
VP → VP PP	0.3

English

Estimating Production Probabilities

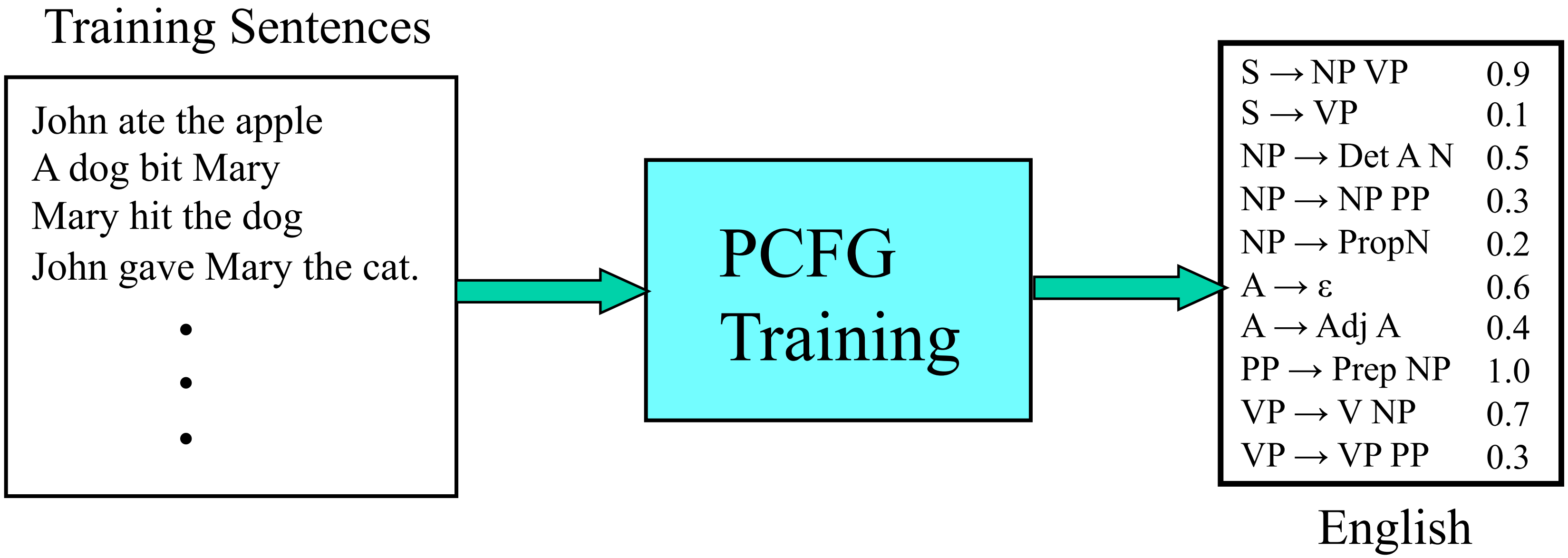
- Set of production rules can be taken directly from the set of rewrites in the treebank.
- Parameters can be directly estimated from frequency counts in the treebank.

$$P(\alpha \rightarrow \beta \mid \alpha) = \frac{\text{count}(\alpha \rightarrow \beta)}{\sum_{\gamma} \text{count}(\alpha \rightarrow \gamma)} = \frac{\text{count}(\alpha \rightarrow \beta)}{\text{count}(\alpha)}$$

PCFG: Maximum Likelihood Training

- Given a set of sentences, induce a grammar that maximizes the probability that this data was generated from this grammar.
- Assume the number of non-terminals in the grammar is specified.
- Only need to have an unannotated set of sequences generated from the model. Does not need correct parse trees for these sentences. In this sense, it is unsupervised.

PCFG: Maximum Likelihood Training



Training Sentences

John ate the apple
A dog bit Mary
Mary hit the dog
John gave Mary the cat.
•
•
•

PCFG Training

S → NP VP	0.9
S → VP	0.1
NP → Det A N	0.5
NP → NP PP	0.3
NP → PropN	0.2
A → ε	0.6
A → Adj A	0.4
PP → Prep NP	1.0
VP → V NP	0.7
VP → VP PP	0.3

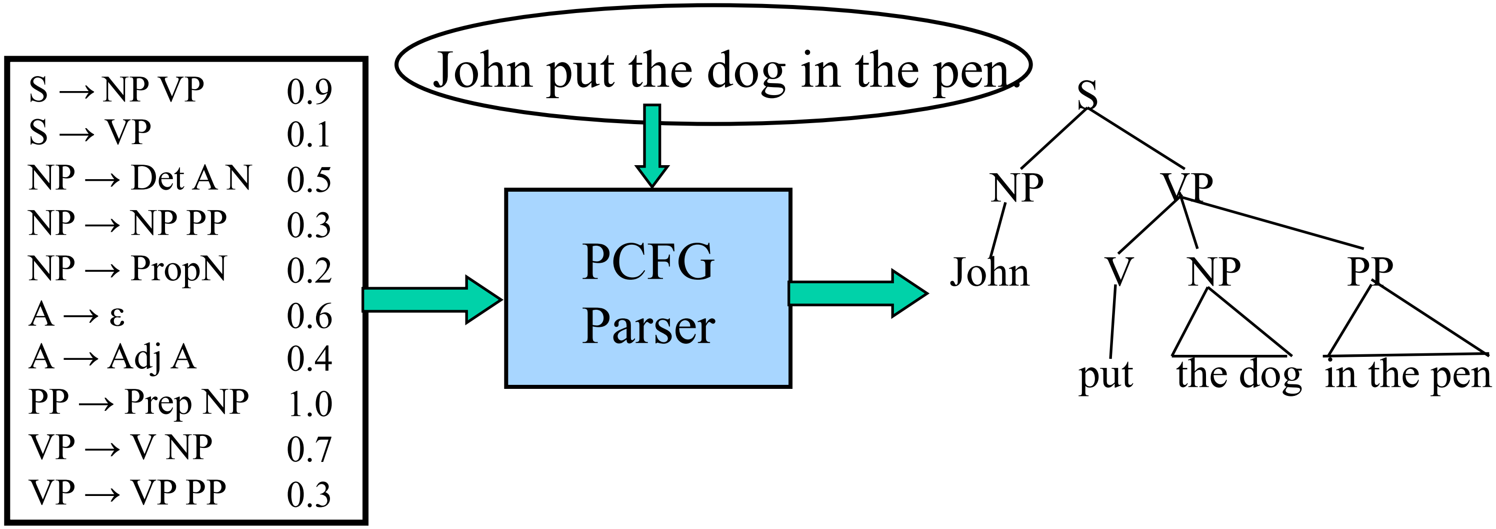
English

Vanilla PCFG Limitations

- Since probabilities of productions do not rely on specific words or concepts, only general structural disambiguation is possible (e.g. prefer to attach PPs to Nominals).
- Consequently, vanilla PCFGs cannot resolve syntactic ambiguities that require semantics to resolve, e.g. ate with fork vs. meatballs.
- In order to work well, PCFGs must be lexicalized, i.e. productions must be specialized to specific words by including their head-word in their LHS non-terminals (e.g. VP-ate).

Example of Importance of Lexicalization

- A general preference for attaching PPs to NPs rather than VPs can be learned by a vanilla PCFG.
- But the desired preference can depend on specific words.



S → NP VP	0.9
S → VP	0.1
NP → Det A N	0.5
NP → NP PP	0.3
NP → PropN	0.2
A → ε	0.6
A → Adj A	0.4
PP → Prep NP	1.0
VP → V NP	0.7
VP → VP PP	0.3

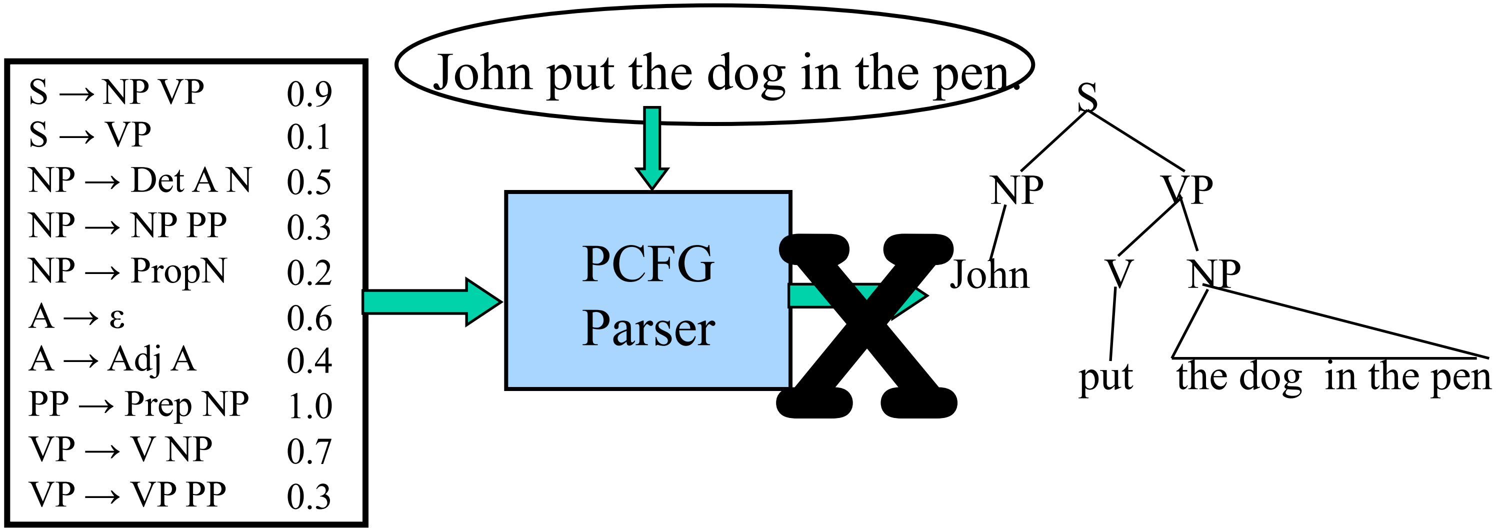
English

John put the dog in the pen.

PCFG Parser

Example of Importance of Lexicalization

- A general preference for attaching PPs to NPs rather than VPs can be learned by a vanilla PCFG.
- But the desired preference can depend on specific words.



S → NP VP	0.9
S → VP	0.1
NP → Det A N	0.5
NP → NP PP	0.3
NP → PropN	0.2
A → ε	0.6
A → Adj A	0.4
PP → Prep NP	1.0
VP → V NP	0.7
VP → VP PP	0.3

English

John put the dog in the pen.

PCFG Parser

Treebanks

- English Penn Treebank: Standard corpus for testing syntactic parsing consists of 1.2 M words of text from the Wall Street Journal (WSJ).
- Typical to train on about 40,000 parsed sentences and test on an additional standard disjoint test set of 2,416 sentences.
- Chinese Penn Treebank: 100K words from the Xinhua news service.
- Other corpora existing in many languages, see the Wikipedia article "Treebank"

First WSJ Sentence

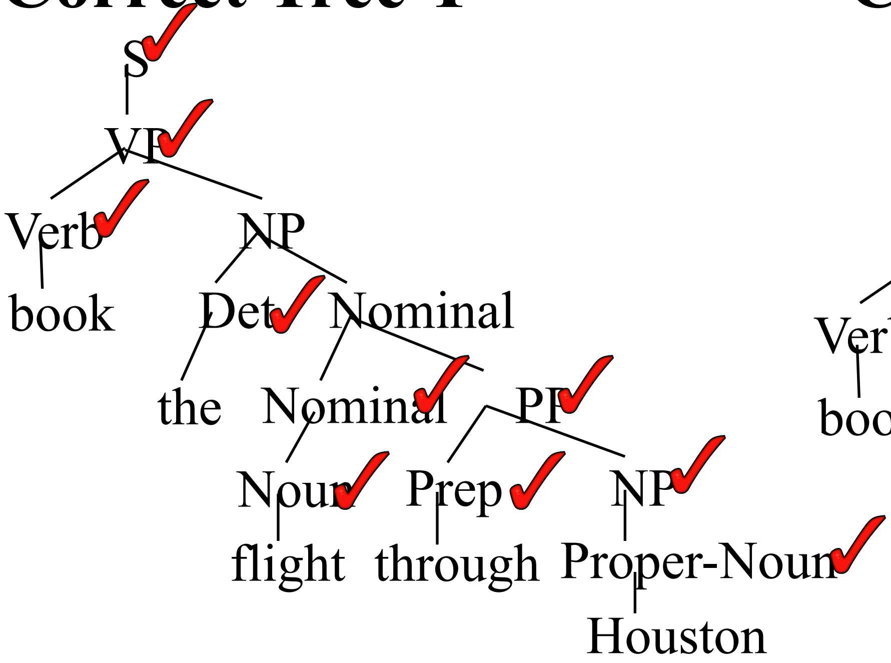
```
( (S
    (NP-SBJ
      (NP (NNP Pierre) (NNP Vinken) )
      (, ,)
      (ADJP
        (NP (CD 61) (NNS years) )
        (JJ old) )
      (, ,) )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board) )
        (PP-CLR (IN as)
          (NP (DT a) (JJ nonexecutive) (NN director) ))
        (NP-TMP (NNP Nov.) (CD 29) )))
    (. .) ))
```

Parsing Evaluation Metrics

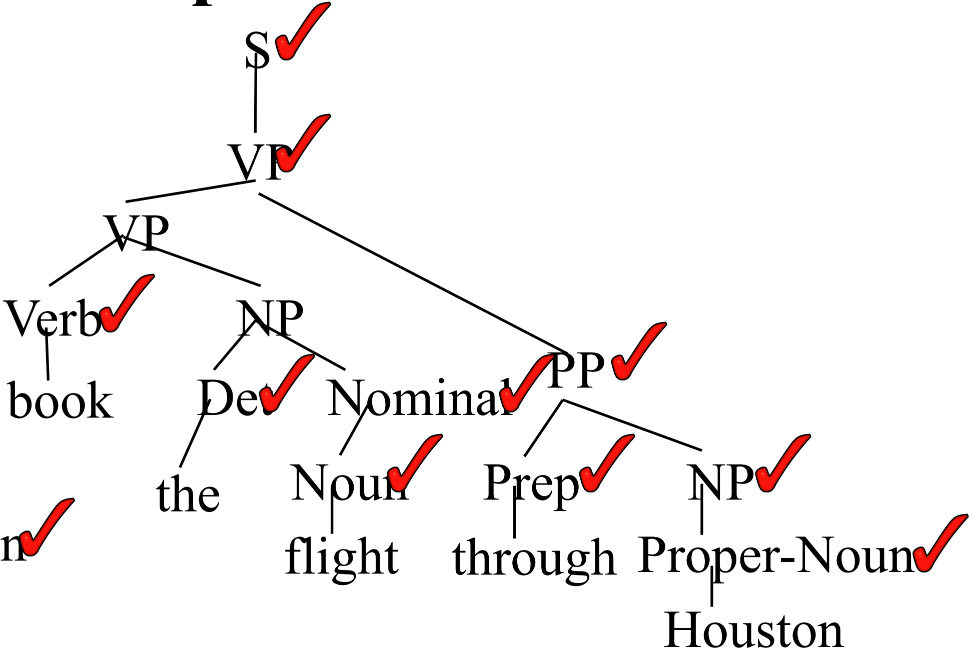
- PARSEVAL metrics measure the fraction of the constituents that match between the computed and human parse trees. If *P* is the system's parse tree and *T* is the human parse tree (the "gold standard"):
  - ***Recall*** = (# correct constituents in *P*) / (# constituents in *T*)
  - ***Precision*** = (# correct constituents in *P*) / (# constituents in *P*)
- ***Labeled Precision*** and ***labeled recall*** require getting the non-terminal label on the constituent node correct to count as correct.
- F_1 is the harmonic mean of precision and recall.

Computing Evaluation Metrics

Correct Tree T



Computed Tree P



# Constituents: 12

# Constituents: 12

# Correct Constituents: 10

Recall = 10/12= 83.3% Precision = 10/12=83.3% F_1 = 83.3%

Treebank Results

- Results of current state-of-the-art systems on the English Penn WSJ treebank are slightly greater than 90% labeled precision and recall.

Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings.
  - Ellen has a strong interest in computational linguistics.
  - Ellen pays a large amount of interest on her credit card.
- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

Ambiguity Resolution is Required for Translation

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
  - "John plays the guitar." → "John toca la guitarra."
  - "John plays soccer." → "John juega el fútbol."
- An apocryphal story is that an early MT system gave the following results when translating from English to Russian and then back to English:
  - "The spirit is willing but the flesh is weak." ⇒ "The liquor is good but the meat is spoiled."
  - "Out of sight, out of mind." ⇒ "Invisible idiot."

Word Sense Disambiguation (WSD) as Text Categorization

- Each sense of an ambiguous word is treated as a category.
  - "play" (verb)
    - play-game
    - play-instrument
    - play-role
  - "pen" (noun)
    - writing-instrument
    - enclosure
- Treat current sentence (or preceding and current sentence) as a document to be classified.
  - "play":
    - play-game: "John played soccer in the stadium on Friday."
    - play-instrument: "John played guitar in the band on Friday."
    - play-role: "John played Hamlet in the theater on Friday."
  - "pen":
    - writing-instrument: "John wrote the letter with a pen in New York."
    - enclosure: "John put the dog in the pen in New York."

Learning for WSD

- Assume part-of-speech (POS), e.g. noun, verb, adjective, for the target word is determined.
- Treat as a classification problem with the appropriate potential senses for the target word given its POS as the categories.
- Encode context using a set of features to be used for disambiguation.
- Train a classifier on labeled data encoded using these features.
- Use the trained classifier to disambiguate future instances of the target word given their contextual features.

WSD “line” Corpus

- 4,149 examples from newspaper articles containing the word “line.”
- Each instance of “line” labeled with one of 6 senses from WordNet.
- Each example includes a sentence containing “line” and the previous sentence for context.

Senses of “line”

- Product: “While he wouldn’t estimate the sale price, analysts have estimated that it would exceed $1 billion. Kraft also told analysts it plans to develop and test a line of refrigerated entrees and desserts, under the Chillery brand name.”
- Formation: “C-LD-R L-V-S V-NNA reads a sign in Caldor’s book department. The 1,000 or so people fighting for a place in line have no trouble filling in the blanks.”
- Text: “Newspaper editor Francis P. Church became famous for a 1897 editorial, addressed to a child, that included the line “Yes, Virginia, there is a Santa Clause.”
- Cord: “It is known as an aggressive, tenacious litigator. Richard D. Parsons, a partner at Patterson, Belknap, Webb and Tyler, likes the experience of opposing Sullivan & Cromwell to “having a thousand-pound tuna on the line.”
- Division: “Today, it is more vital than ever. In 1983, the act was entrenched in a new constitution, which established a tricameral parliament along racial lines, whith separate chambers for whites, coloreds and Asians but none for blacks.”
- Phone: “On the tape recording of Mrs. Guba's call to the 911 emergency line, played at the trial, the baby sitter is heard begging for an ambulance.”

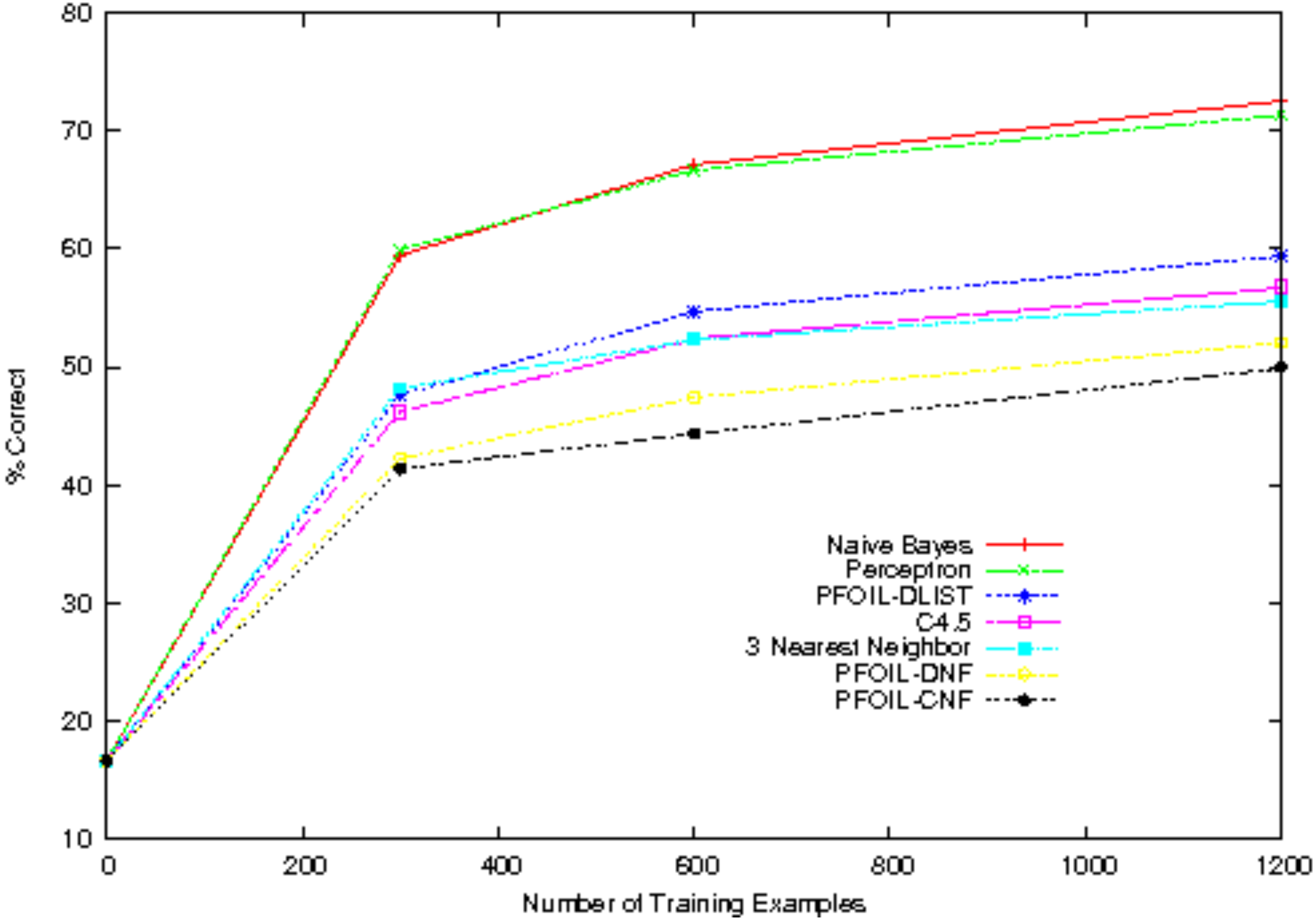
Experimental Data for WSD of "line"

- Sample equal number of examples of each sense to construct a corpus of 2,094.
- Represent as simple binary vectors of word occurrences in 2 sentence context.
  - Stop words eliminated
  - Stemmed to eliminate morphological variation
- Final examples represented with 2,859 binary word features.

Learning Algorithms

- Naïve Bayes
  - Binary features
- K Nearest Neighbor
  - Simple instance-based algorithm with k=3 and Hamming distance
- Perceptron
  - Simple neural-network algorithm.
- C4.5
  - State of the art decision-tree induction algorithm
- PFOIL-DNF
  - Simple logical rule learner for Disjunctive Normal Form
- PFOIL-CNF
  - Simple logical rule learner for Conjunctive Normal Form
- PFOIL-DLIST
  - Simple logical rule learner for decision-list of conjunctive rules

Learning Curves for WSD of "line"



Discussion of Learning Curves for WSD of “line”

- Naïve Bayes and Perceptron give the best results.
- Both use a weighted linear combination of evidence from many features.
- Symbolic systems that try to find a small set of relevant features tend to overfit the training data and are not as accurate.
- Nearest neighbor method that weights all features equally is also not as accurate.
- Of symbolic systems, decision lists work the best.

Other Syntactic Tasks

Word Segmentation

- Breaking a string of characters (graphemes) into a sequence of words.
- In some written languages (e.g. Chinese) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ( )]
- Examples from English URLs:
  - jumptheshark.com ⇒ jump the shark .com
  - myspace.com/pluckerswingbar

    ⇒ myspace .com pluckers wing bar

    ⊗⇒ myspace .com plucker swing bar

Morphological Analysis

- ***Morphology*** is the field of linguistics that studies the internal structure of words. (Wikipedia)
- A ***morpheme*** is the smallest linguistic unit that has semantic meaning (Wikipedia)
  - e.g. “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried ⇒ carry + ed (past tense)
  - independently ⇒ in + (depend + ent) + ly
  - Googlers ⇒ (Google + er) + s (plural)
  - unlockable ⇒ un + (lock + able) ?

    ⇒ (un + lock) + able ?

Part Of Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech.

I	ate	the	spaghetti	with	meatballs.
Pro	V	Det	N	Prep	N

John	saw	the	saw	and	decided	to	take	it	to	the	table.
PN	V	Det	N	Con	V	Part	V	Pro	Prep	Det	N

- Useful for subsequent syntactic parsing and word sense disambiguation.

Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.
  - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
  - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

Other Semantic Tasks

Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

  agent patient source destination instrument

  - John drove Mary from Austin to Dallas in his Toyota Prius.
  - The hammer broke the window.

- Also referred to a "case role analysis," "thematic analysis," and "shallow semantic parsing"

Semantic Parsing

- A ***semantic parser*** maps a natural-language sentence to a complete, detailed semantic representation (***logical form***).
- For many applications, the desired output is immediately executable by another program.
- Example: Mapping an English database query to Prolog:

  How many cities are there in the US?

  ```
  answer(A, count(B, (city(B), loc(B, C),
                      const(C, countryid(USA))),
                 A))
  ```

Textual Entailment

- Determine whether one natural language sentence entails (implies) another under an ordinary interpretation.

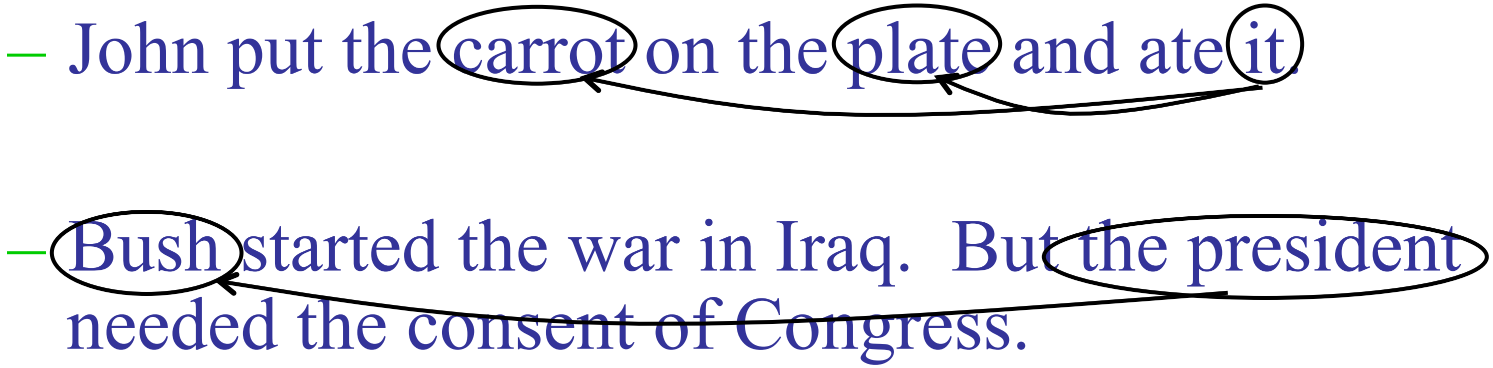
Textual Entailment Problems from PASCAL Challenge

TEXT	**HYPOTHESIS**	**ENTAIL MENT**
Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.	*Yahoo bought Overture.*	TRUE
Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.	*Microsoft bought Star Office.*	FALSE
The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.	*Israel was established in May 1971.*	FALSE
Since its formation in 1948, Israel fought many wars with neighboring Arab countries.	*Israel was established in 1948.*	TRUE

Pragmatics/Discourse Tasks

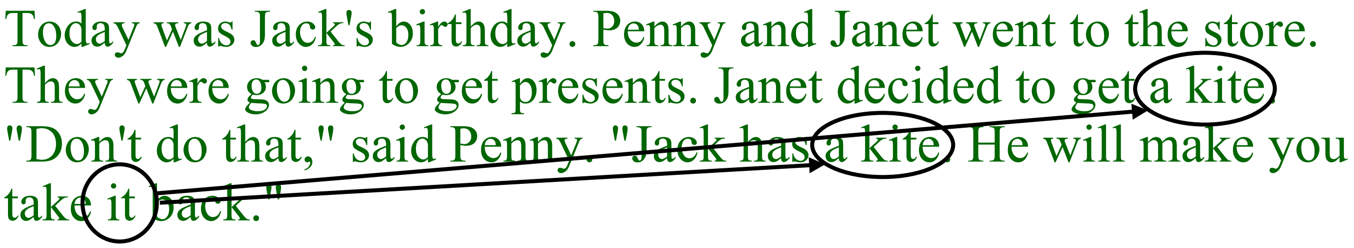
Anaphora Resolution/ Co-Reference

- Determine which phrases in a document refer to the same underlying entity.



  - John put the carrot on the plate and ate it.
  - Bush started the war in Iraq. But the president needed the consent of Congress.

- Some cases require difficult reasoning.



  - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Ellipsis Resolution

- Frequently words and phrases are omitted from sentences when they can be inferred from context.

"Wise men talk because they have something to say;
fools, because they have to say something.“ (Plato)

fools, talk because they have to say something.“ (Plato)

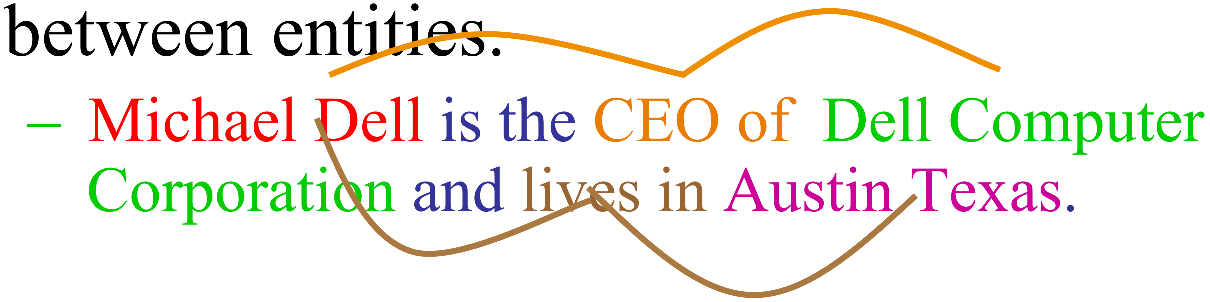
Other Tasks

Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is task of identifying names of people, places, organizations, etc. in text.

  people organizations places

  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
- Relation extraction identifies specific relations between entities.



  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

Question Answering

- Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the web).
  - When was Barack Obama born? (*factoid*)
    - August 4, 1961
  - Who was president when Barack Obama was born?
    - John F. Kennedy
  - How many presidents have there been since Barack Obama was born?
    - 9

Text Summarization

- Produce a short summary of a longer document or article.
  - Article: With a split decision in the final two primaries and a flurry of superdelegate endorsements, Sen. Barack Obama sealed the Democratic presidential nomination last night after a grueling and history-making campaign against Sen. Hillary Rodham Clinton that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against Sen. John McCain, the presumptive Republican nominee….
  - Summary: Senator Barack Obama was declared the presumptive Democratic presidential nominee.

Machine Translation (MT)

- Translate a sentence from one natural language to another.
  - Hasta la vista, bebé ⇒

    Until we see each other again, baby.

NLP Conclusions

- The need for disambiguation makes language understanding difficult.
- Levels of linguistic processing:
  - Syntax
  - Semantics
  - Pragmatics
- CFGs can be used to parse natural language but produce many spurious parses.
- Statistical learning methods can be used to:
  - Automatically learn grammars from (annotated) corpora.
  - Compute the most likely interpretation based on a learned statistical model.