Parsing, and Context-Free Grammars

Michael Collins, Columbia University

Overview
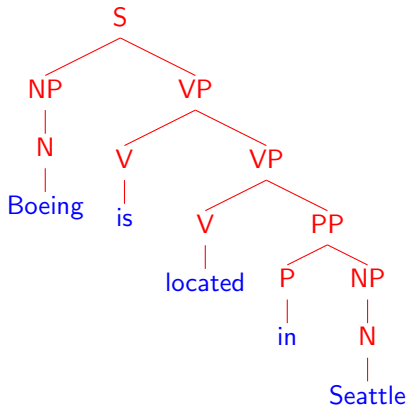
- An introduction to the parsing problem

- Context free grammars

- A brief(!) sketch of the syntax of English

- Examples of ambiguous structures

Parsing (Syntactic Structure)

Boeing is located in Seattle.

# Syntactic Formalisms

- Work in formal syntax goes back to Chomsky's PhD thesis in the 1950s

- Examples of current formalisms: minimalism, lexical functional grammar (LFG), head-driven phrase-structure grammar (HPSG), tree adjoining grammars (TAG), categorial grammars

Data for Parsing Experiments

- Penn WSJ Treebank = 50,000 sentences with associated trees

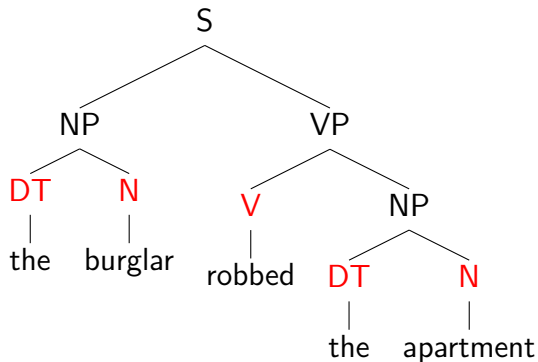- Usual set-up: 40,000 training sentences, 2400 test sentences

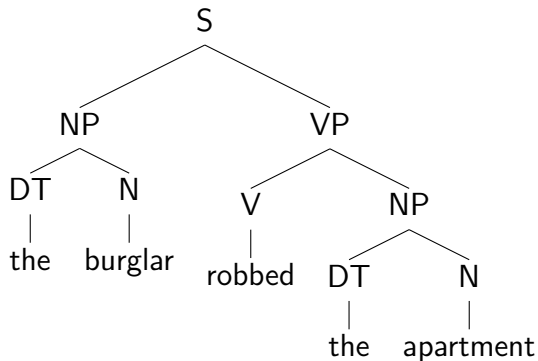**An example tree:**
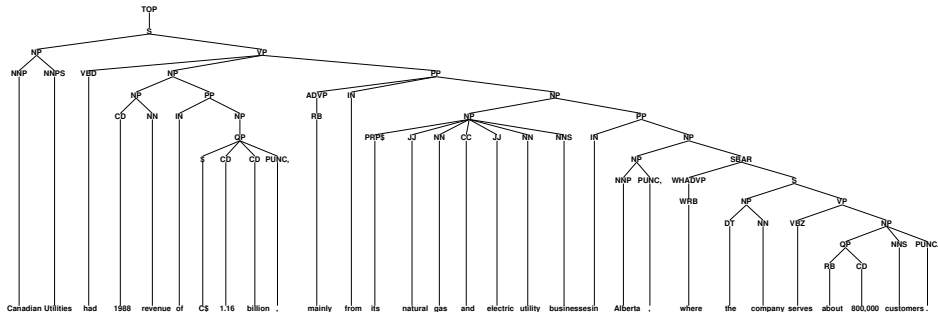
The Information Conveyed by Parse Trees

(1) Part of speech for each word
(N = noun, V = verb, DT = determiner)

The Information Conveyed by Parse Trees (continued)

(2) Phrases

```
                        S
              ┌─────────┴─────────┐
             NP                   VP
          ┌───┴───┐          ┌────┴────┐
         DT        N         V         NP
          │        │         │      ┌───┴────┐
         the    burglar    robbed  DT         N
                                    │         │
                                   the    apartment
```
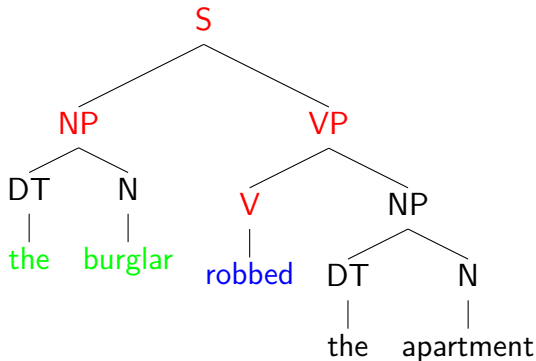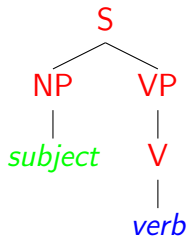
Noun Phrases (NP): "the burglar", "the apartment"
Verb Phrases (VP): "robbed the apartment"
Sentences (S):   "the burglar robbed the apartment"

The Information Conveyed by Parse Trees (continued)

(3) Useful Relationships

$\Rightarrow$ "the burglar" is the subject of "robbed"

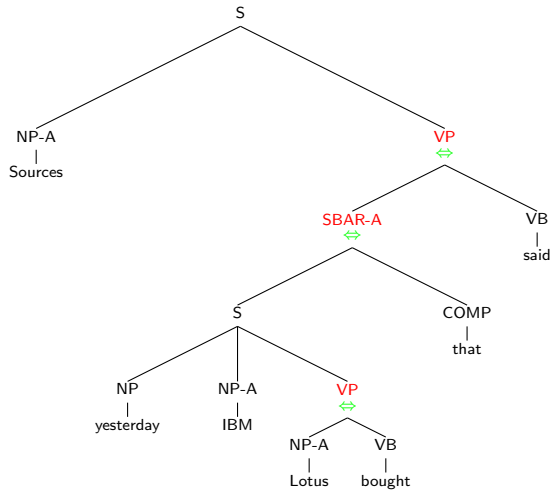# An Example Application: Machine Translation

- English word order is   *subject – verb – object*

- Japanese word order is   *subject – object – verb*

  English: IBM bought Lotus
  Japanese: *IBM Lotus bought*

  English: Sources said that IBM bought Lotus yesterday
  Japanese: *Sources yesterday IBM Lotus bought that said*

```
                                    S
                      ┌─────────────┴─────────────┐
                  NP-A                            VP
                    │                             ⇔
                 Sources              ┌───────────┴───────┐
                                   SBAR-A                 VB
                                     ⇔                     │
                         ┌───────────┴───────┐           said
                         S                  COMP
              ┌──────────┼──────────┐         │
             NP        NP-A        VP        that
              │          │          ⇔
          yesterday     IBM    ┌─────┴─────┐
                             NP-A         VB
                               │           │
                             Lotus      bought
```

Overview

- An introduction to the parsing problem

- Context free grammars

- A brief(!) sketch of the syntax of English

- Examples of ambiguous structures

Context-Free Grammars

Hopcroft and Ullman, 1979

A context free grammar $G = (N, \Sigma, R, S)$ where:

- N is a set of non-terminal symbols
- Σ is a set of terminal symbols
- R is a set of rules of the form $X \to Y_1 Y_2 \ldots Y_n$
  for $n \geq 0$, $X \in N$, $Y_i \in (N \cup \Sigma)$
- $S \in N$ is a distinguished start symbol

A Context-Free Grammar for English

$N = \{$S, NP, VP, PP, DT, Vi, Vt, NN, IN$\}$

$S = $ S

$\Sigma = \{$sleeps, saw, man, woman, telescope, the, with, in$\}$

$R = $

S	$\rightarrow$	NP	VP
VP	$\rightarrow$	Vi	
VP	$\rightarrow$	Vt	NP
VP	$\rightarrow$	VP	PP
NP	$\rightarrow$	DT	NN
NP	$\rightarrow$	NP	PP
PP	$\rightarrow$	IN	NP

Vi	$\rightarrow$	sleeps
Vt	$\rightarrow$	saw
NN	$\rightarrow$	man
NN	$\rightarrow$	woman
NN	$\rightarrow$	telescope
DT	$\rightarrow$	the
IN	$\rightarrow$	with
IN	$\rightarrow$	in

Note: S=sentence, VP=verb phrase, NP=noun phrase,
PP=prepositional phrase, DT=determiner, Vi=intransitive verb,
Vt=transitive verb, NN=noun, IN=preposition

Left-Most Derivations 最左派生

A left-most derivation is a sequence of strings $s_1 \ldots s_n$, where

- $s_1 = S$, the start symbol
- $s_n \in \Sigma^*$, i.e. s_n is made up of terminal symbols only
- Each s_i for $i = 2 \ldots n$ is derived from s_{i-1} by picking the left-most non-terminal X in s_{i-1} and replacing it by some β where $X \rightarrow \beta$ is a rule in R

For example: [S], [NP VP], [D N VP], [the N VP], [the man VP], [the man Vi], [the man sleeps]

Representation of a derivation as a tree:
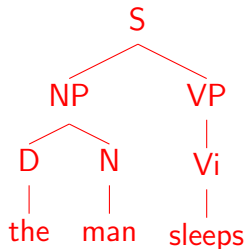
An Example

DERIVATION   RULES USED

S

An Example

DERIVATION   RULES USED

S

                            $S \rightarrow NP\ VP$

NP VP

An Example

DERIVATION	RULES USED
S	
NP VP	S → NP VP
DT N VP	NP → DT N

An Example

DERIVATION	RULES USED
S	S → NP VP
NP VP	NP → DT N
DT N VP	DT → the
the N VP	

An Example

DERIVATION	RULES USED
S	S → NP VP
NP VP	NP → DT N
DT N VP	DT → the
the N VP	N → dog
the dog VP	

An Example

DERIVATION	RULES USED
S	S → NP VP
NP VP	NP → DT N
DT N VP	DT → the
the N VP	N → dog
the dog VP	VP → VB
the dog VB	

An Example

DERIVATION	RULES USED
S	S → NP VP
NP VP	NP → DT N
DT N VP	DT → the
the N VP	N → dog
the dog VP	VP → VB
the dog VB	VB → laughs
the dog laughs	

An Example

DERIVATION	RULES USED
S	S → NP VP
NP VP	NP → DT N
DT N VP	DT → the
the N VP	N → dog
the dog VP	VP → VB
the dog VB	VB → laughs
the dog laughs	

```
          S
        /   \
      NP     VP
     /  \     |
    DT   N    VB
    |    |    |
   the  dog  laughs
```
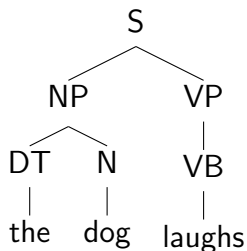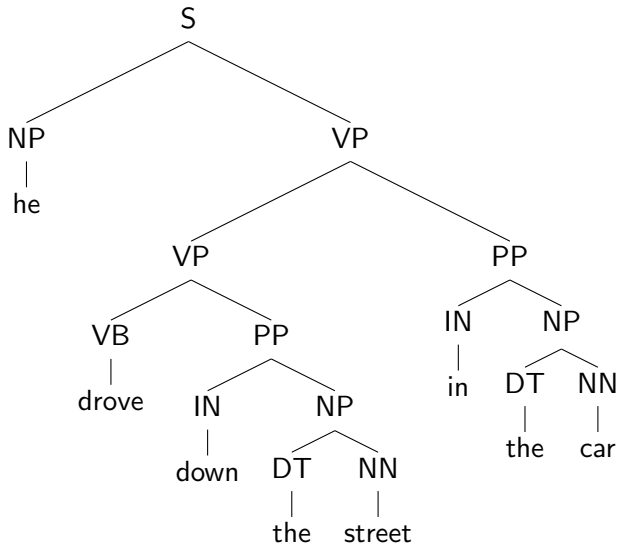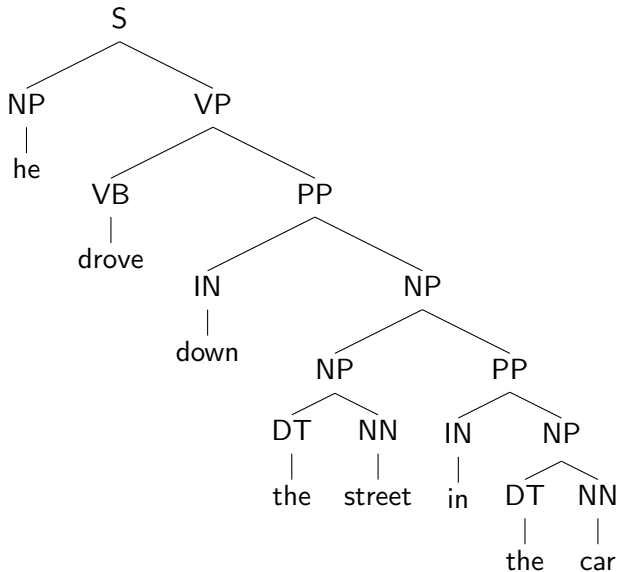
Properties of CFGs

- A CFG defines a set of possible derivations

- A string $s \in \Sigma^*$ is in the *language* defined by the CFG if there is at least one derivation that yields s

- Each string in the language generated by the CFG may have more than one derivation ("ambiguity")

An Example of Ambiguity

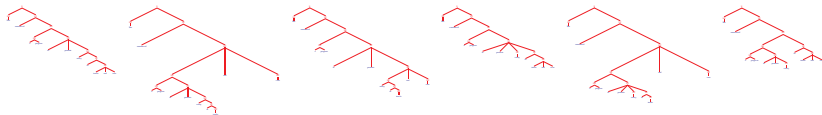# An Example of Ambiguity (continued)

The Problem with Parsing: Ambiguity

INPUT:
She announced a program to promote safety in trucks and vans

⇓

POSSIBLE OUTPUTS:

And there are more...

Overview

- An introduction to the parsing problem

- Context free grammars

- A brief(!) sketch of the syntax of English

- Examples of ambiguous structures

A
COMPREHENSIVE
GRAMMAR
OF THE
ENGLISH
LANGUAGE

Randolph Quirk
Sidney Greenbaum
Geoffrey Leech
Jan Svartvik
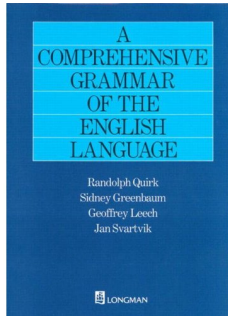
LONGMAN

Product Details (from Amazon)
Hardcover: 1779 pages
Publisher: Longman; 2nd Revised edition
Language: English
ISBN-10: 0582517346
ISBN-13: 978-0582517349
Product Dimensions: 8.4 x 2.4 x 10 inches
Shipping Weight: 4.6 pounds

A Brief Overview of English Syntax

Parts of Speech (tags from the Brown corpus):

- Nouns
  NN = singular noun   e.g., man, dog, park
  NNS = plural noun   e.g., telescopes, houses, buildings
  NNP = proper noun   e.g., Smith, Gates, IBM
  专有名词
- Determiners
  DT = determiner   e.g., the, a, some, every
- Adjectives
  JJ = adjective   e.g., red, green, large, idealistic

A Fragment of a Noun Phrase Grammar

N̄	⇒	NN
N̄	⇒	NN N̄
N̄	⇒	JJ N̄
N̄	⇒	N̄ N̄
NP	⇒	DT N̄

NN	⇒	box
NN	⇒	car
NN	⇒	mechanic
NN	⇒	pigeon
DT	⇒	the
DT	⇒	a

JJ	⇒	fast
JJ	⇒	metal
JJ	⇒	idealistic
JJ	⇒	clay

Prepositions, and Prepositional Phrases

- Prepositions
  IN = preposition   e.g., of, in, out, beside, as

An Extended Grammar

N̄	⇒	NN
N̄	⇒	NN N̄
N̄	⇒	JJ N̄
N̄	⇒	N̄ N̄
NP	⇒	DT N̄
PP	⇒	IN NP
N̄	⇒	N̄ PP

NN	⇒	box
NN	⇒	car
NN	⇒	mechanic
NN	⇒	pigeon
DT	⇒	the
DT	⇒	a

JJ	⇒	fast
JJ	⇒	metal
JJ	⇒	idealistic
JJ	⇒	clay
IN	⇒	in
IN	⇒	under
IN	⇒	of
IN	⇒	on
IN	⇒	with
IN	⇒	as

Generates:

in a box, under the box, the fast car mechanic under the pigeon in the box, . . .

An Extended Grammar

$$
\begin{array}{lcll}
\bar{N} & \Rightarrow & NN & \\
\bar{N} & \Rightarrow & NN & \bar{N} \\
\bar{N} & \Rightarrow & JJ & \bar{N} \\
\bar{N} & \Rightarrow & \bar{N} & \bar{N} \\
NP & \Rightarrow & DT & \bar{N} \\
\\
PP & \Rightarrow & IN & NP \\
\bar{N} & \Rightarrow & \bar{N} & PP \\
\end{array}
$$

Verbs, Verb Phrases, and Sentences

- ► Basic Verb Types
  Vi = Intransitive verb   e.g., sleeps, walks, laughs
  Vt = Transitive verb   e.g., sees, saw, likes
  Vd = Ditransitive verb   e.g., gave

- ► Basic VP Rules
  VP   →   Vi
  VP   →   Vt   NP
  VP   →   Vd   NP   NP

- ► Basic S Rule
  S   →   NP   VP

Examples of VP:
sleeps, walks, likes the mechanic, gave the mechanic the fast car
Examples of S:
the man sleeps, the dog walks, the dog gave the mechanic the fast car

PPs Modifying Verb Phrases

A new rule: VP → VP   PP

New examples of VP:
sleeps in the car, walks like the mechanic, gave the mechanic the fast car on Tuesday, ...

Complementizers, and SBARs

- Complementizers 标句词
  COMP = complementizer   e.g., that

- SBAR 非限定状态
  SBAR → COMP S

Examples:
that the man sleeps, that the mechanic saw the dog . . .

More Verbs

- New Verb Types
  V[5] e.g., said, reported
  V[6] e.g., told, informed
  V[7] e.g., bet

- New VP Rules
  VP → V[5] SBAR
  VP → V[6] NP   SBAR
  VP → V[7] NP   NP   SBAR

Examples of New VPs:

said that the man sleeps

told the dog that the mechanic likes the pigeon

bet the pigeon $50 that the mechanic owns a fast car

Coordination

- A New Part-of-Speech:
  CC = Coordinator   e.g., and, or, but
  并列连词
- New Rules

NP	$\rightarrow$	NP	CC	NP
$\bar{N}$	$\rightarrow$	$\bar{N}$	CC	$\bar{N}$
VP	$\rightarrow$	VP	CC	VP
S	$\rightarrow$	S	CC	S
SBAR	$\rightarrow$	SBAR	CC	SBAR

We've Only Scratched the Surface...

- ▶ Agreement

  The dogs laugh *vs.* The dog laughs

- ▶ Wh-movement

  The dog that the cat liked ___

- ▶ Active vs. passive

  The dog saw the cat *vs.*
  The cat was seen by the dog

- ▶ If you're interested in reading more:

  Syntactic Theory: A Formal Introduction, 2nd Edition. Ivan A. Sag, Thomas Wasow, and Emily M. Bender.

Overview

- An introduction to the parsing problem

- Context free grammars

- A brief(!) sketch of the syntax of English

- Examples of ambiguous structures

Sources of Ambiguity

- Part-of-Speech ambiguity
  NN → duck
  Vi → duck

```
                        S
              ┌─────────┴─────────┐
             NP                   VP
              │          ┌─────────┴─────────┐
              I         VP                   PP
                   ┌────┴────┐          ┌────┴────┐
                  Vi        PP         IN        NP
                   │     ┌───┴───┐      │     ┌───┴───┐
                 drove  IN      NP      in   DT      NN
                         │   ┌───┴───┐        │       │
                       down DT      NN       the     car
                             │       │
                            the     road
```

```
                        S
              ┌─────────┴─────────┐
             NP                   VP
              │           ┌───────┴────────┐
              I          Vi               PP
                          │         ┌──────┴──────┐
                        drove      IN            NP
                                    │      ┌──────┴──────┐
                                  down    NP            PP
                                      ┌────┴───┐    ┌────┴────┐
                                     DT       NN   IN        NP
                                      │        │    │    ┌────┴───┐
                                     the     road   in  DT       NN
                                                     │    │        │
                                                         the      car
```
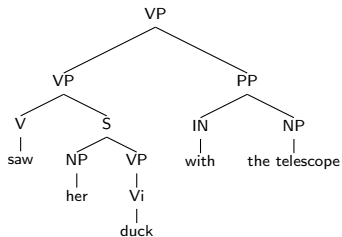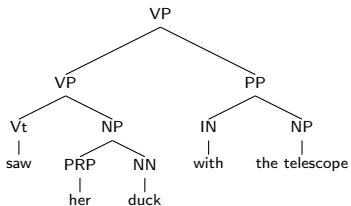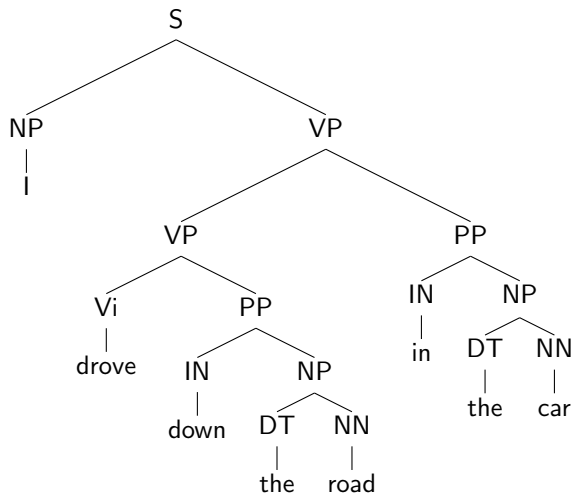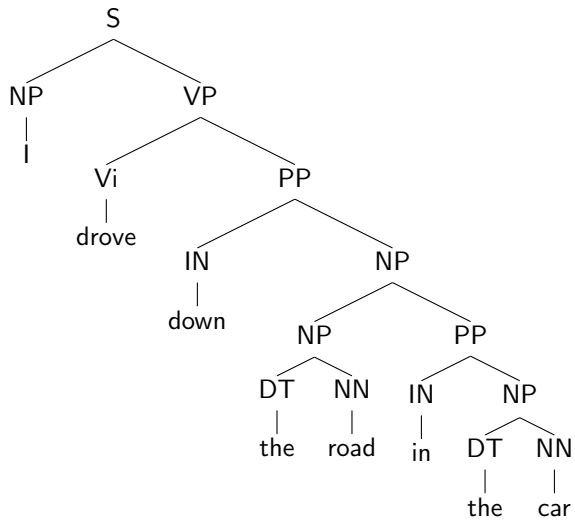
Two analyses for: John was believed to have been shot by Bill

Sources of Ambiguity: Noun Premodifiers

- Noun premodifiers: