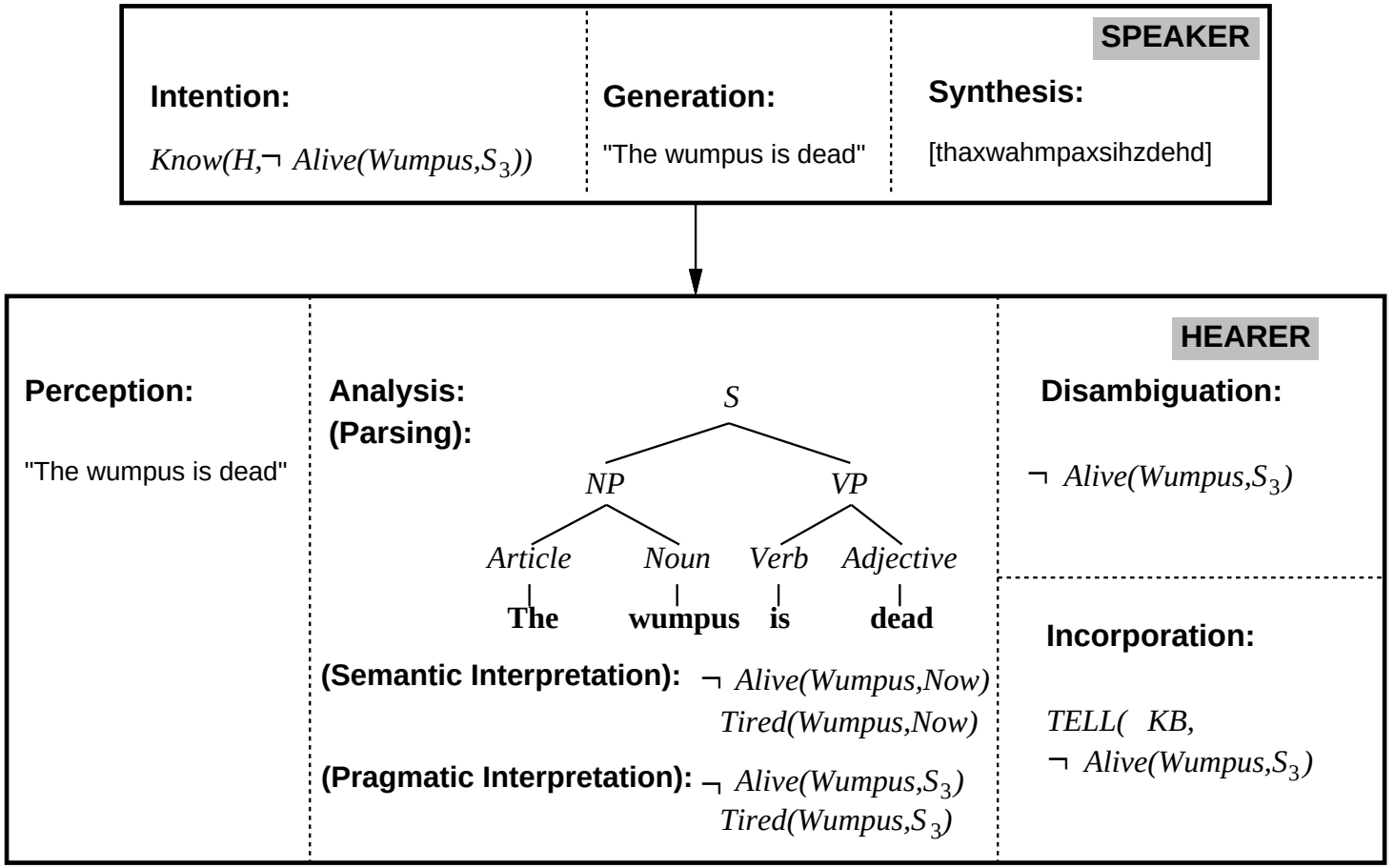
Language

Chapter 22

Outline

- Communication
- Grammar
- Syntactic analysis
- Problems
- Grammar induction

Stages in communication: example



Grammar

Grammar specifies the compositional structure of complex messages

A formal language is a set of strings of terminal symbols

Each string in the language can be analyzed/generated by the grammar

The grammar is a set of rewrite rules, e.g.,

$$S \rightarrow NP\ VP$$
$$Article \rightarrow \textbf{\textit{the}} \mid \textbf{\textit{a}} \mid \textbf{\textit{an}} \mid \ldots$$

Here S is the sentence symbol, NP and VP are nonterminals

Grammar types

Regular: $nonterminal \rightarrow \boldsymbol{terminal}[nonterminal]$

$S \rightarrow \boldsymbol{a}S$
$S \rightarrow \Lambda$

Context-free: $nonterminal \rightarrow anything$

$S \rightarrow \boldsymbol{a}S\boldsymbol{b}$

Context-sensitive: a single nonterminal in a string is replaced by a string

$ASB \rightarrow AA\boldsymbol{a}BB$

Recursively enumerable: no constraints

Natural languages probably context-free, parsable in real time!

Wumpus lexicon

$$
\begin{aligned}
\textit{Noun} &\rightarrow \textbf{\textit{stench}} \mid \textbf{\textit{breeze}} \mid \textbf{\textit{glitter}} \mid \textbf{\textit{nothing}} \\
&\quad \mid \textbf{\textit{wumpus}} \mid \textbf{\textit{pit}} \mid \textbf{\textit{pits}} \mid \textbf{\textit{gold}} \mid \textbf{\textit{east}} \mid \ldots \\
\textit{Verb} &\rightarrow \textbf{\textit{is}} \mid \textbf{\textit{see}} \mid \textbf{\textit{smell}} \mid \textbf{\textit{shoot}} \mid \textbf{\textit{feel}} \mid \textbf{\textit{stinks}} \\
&\quad \mid \textbf{\textit{go}} \mid \textbf{\textit{grab}} \mid \textbf{\textit{carry}} \mid \textbf{\textit{kill}} \mid \textbf{\textit{turn}} \mid \ldots \\
\textit{Adjective} &\rightarrow \textbf{\textit{right}} \mid \textbf{\textit{left}} \mid \textbf{\textit{east}} \mid \textbf{\textit{south}} \mid \textbf{\textit{back}} \mid \textbf{\textit{smelly}} \mid \ldots \\
\textit{Adverb} &\rightarrow \textbf{\textit{here}} \mid \textbf{\textit{there}} \mid \textbf{\textit{nearby}} \mid \textbf{\textit{ahead}} \\
&\quad \mid \textbf{\textit{right}} \mid \textbf{\textit{left}} \mid \textbf{\textit{east}} \mid \textbf{\textit{south}} \mid \textbf{\textit{back}} \mid \ldots \\
\textit{Pronoun} &\rightarrow \textbf{\textit{me}} \mid \textbf{\textit{you}} \mid \textbf{\textit{I}} \mid \textbf{\textit{it}} \mid \ldots \\
\textit{Name} &\rightarrow \textbf{\textit{John}} \mid \textbf{\textit{Mary}} \mid \textbf{\textit{Waikato}} \mid \textbf{\textit{Otago}} \mid \ldots \\
\textit{Article} &\rightarrow \textbf{\textit{the}} \mid \textbf{\textit{a}} \mid \textbf{\textit{an}} \mid \ldots \\
\textit{Preposition} &\rightarrow \textbf{\textit{to}} \mid \textbf{\textit{in}} \mid \textbf{\textit{on}} \mid \textbf{\textit{near}} \mid \ldots \\
\textit{Conjunction} &\rightarrow \textbf{\textit{and}} \mid \textbf{\textit{or}} \mid \textbf{\textit{but}} \mid \ldots \\
\textit{Digit} &\rightarrow \mathbf{0} \mid \mathbf{1} \mid \mathbf{2} \mid \mathbf{3} \mid \mathbf{4} \mid \mathbf{5} \mid \mathbf{6} \mid \mathbf{7} \mid \mathbf{8} \mid \mathbf{9}
\end{aligned}
$$

Divided into closed and open classes

Wumpus grammar

$S \rightarrow$	$NP\ VP$	I + feel a breeze
$\mid$	$S\ Conjunction\ S$	I feel a breeze + and + I smell a wumpus
$NP \rightarrow$	$Pronoun$	I
$\mid$	$Noun$	pits
$\mid$	$Article\ Noun$	the + wumpus
$\mid$	$Digit\ Digit$	3 4
$\mid$	$NP\ PP$	the wumpus + to the east
$\mid$	$NP\ RelClause$	the wumpus + that is smelly
$VP \rightarrow$	$Verb$	stinks
$\mid$	$VP\ NP$	feel + a breeze
$\mid$	$VP\ Adjective$	is + smelly
$\mid$	$VP\ PP$	turn + to the east
$\mid$	$VP\ Adverb$	go + ahead
$PP \rightarrow$	$Preposition\ NP$	to + the east
$RelClause \rightarrow$	$\boldsymbol{that}\ VP$	that + is smelly

Parse trees

Exhibit the grammatical structure of a sentence

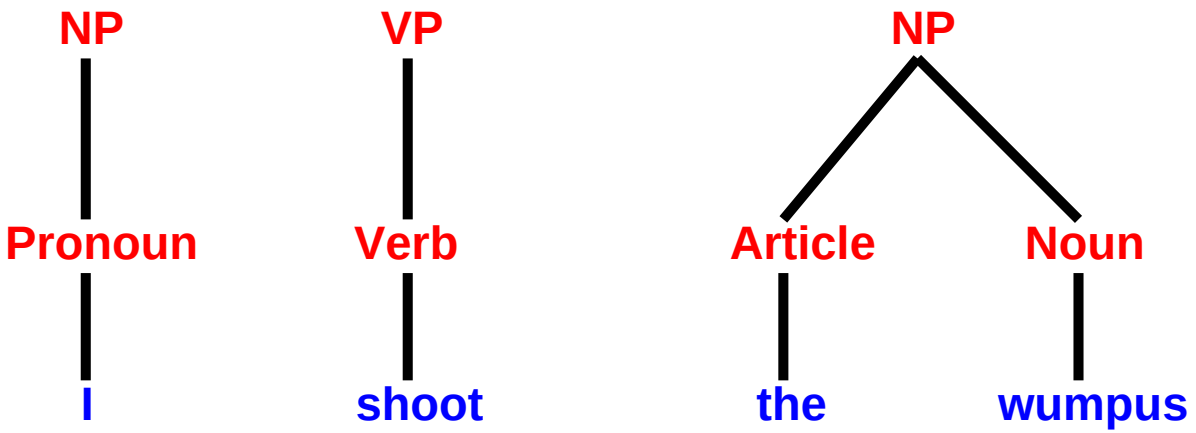
I **shoot** **the** **wumpus**

Parse trees

Exhibit the grammatical structure of a sentence

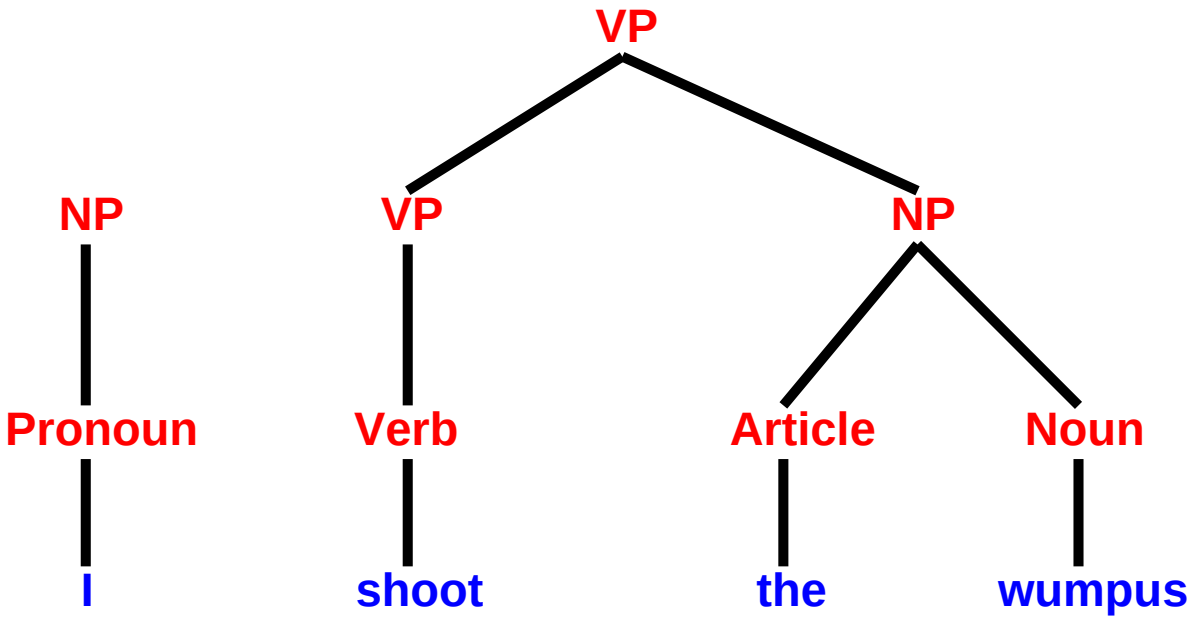
Parse trees

Exhibit the grammatical structure of a sentence



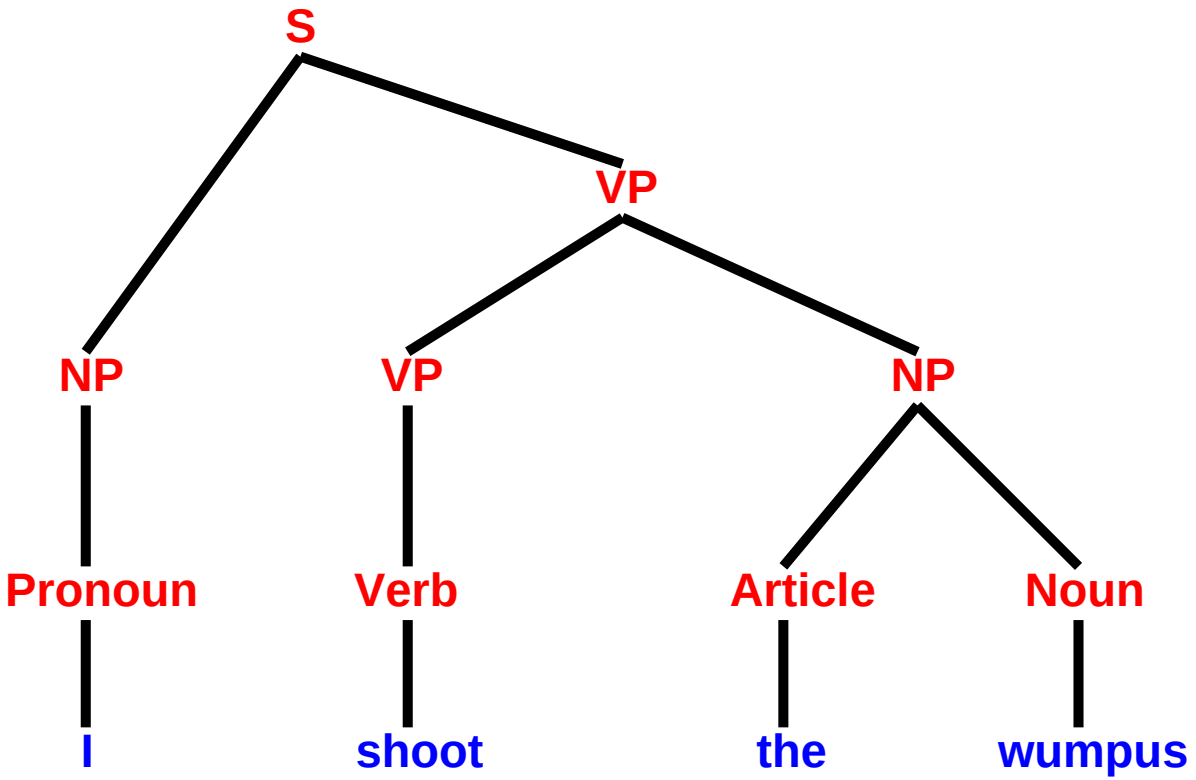
Parse trees

Exhibit the grammatical structure of a sentence



Parse trees

Exhibit the grammatical structure of a sentence



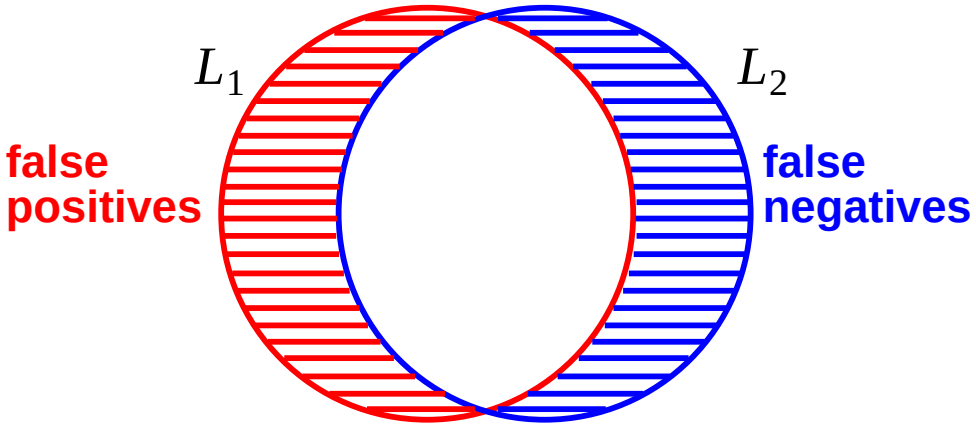
Context-free parsing

Bottom-up parsing works by replacing any substring that matches RHS of a rule with the rule's LHS

Efficient algorithms (e.g., chart parsing, Section 22.3) $O(n^3)$ for context-free, run at several thousand words/sec for real grammars

Language learning

Formal language L_1 may differ from natural language L_2



Adjusting L_1 to agree with L_2 is a learning problem

Real grammars 10–500 pages, insufficient even for "proper" English

Probablistic grammars are an alternative

Real language

Real human languages provide many problems for NLP:

- ambiguity
- anaphora
- indexicality
- metonymy
- metaphor

etc.

Ambiguity

Squad helps dog bite victim

Helicopter powered by human flies

American pushes bottle up Germans

Anaphora

Using pronouns to refer back to entities already introduced in the text

After Mary proposed to John, **they** found a preacher and got married.

For the honeymoon, **they** went to Hawaii

Mary saw a ring through the window and asked John for **it**

Mary threw a rock at the window and broke **it**

Indexicality

Indexical sentences refer to utterance situation (place, time, etc.)

I am over **here**

Why did **you** do **that**?

Metonymy

Using one noun phrase to stand for another

I've read **Shakespeare**

Chrysler announced record profits

The **ham sandwich** on Table 4 wants another beer

Metaphor

"Non-literal" usage of words and phrases, often systematic:

I've tried killing the process but it won't die. Its parent keeps it alive.

A linear-time algorithm for grammar learning

SEQUITUR algorithm developed by Nevill-Manning and Witten at Waikato: learns grammar that generates only one sentence, namely the input text

Basic idea: good grammar is compact grammar (i.e., grammar should compress the input text)

Scans input from left to right, one symbol at a time, building up a grammar for the text seen so far.

Grammar is built by enforcing the following two constraints:

No pair of adjacent symbols must appear more than once in the grammar.

Every rule in the grammar must be used at least twice (apart from the start rule).

SEQUITUR **example**

	Input	Grammar	Comments
1	a	$S \rightarrow a$	
2	ab	$S \rightarrow ab$	
3	abc	$S \rightarrow abc$	
4	$abcd$	$S \rightarrow abcd$	
5	$abcdb$	$S \rightarrow abcdb$	
6	$abcdbc$	$S \rightarrow abcdbc$	bc twice
		$S \rightarrow aAdA; A \rightarrow bc$	
7	$abcdbca$	$S \rightarrow aAdAa; A \rightarrow bc$	
8	$abcdbcab$	$S \rightarrow aAdAab; A \rightarrow bc$	
9	$abcdbcabc$	$S \rightarrow aAdAabc; A \rightarrow bc$	bc twice
		$S \rightarrow aAdAaA; A \rightarrow bc$	aA twice
		$S \rightarrow BdAB; A \rightarrow bc; B \rightarrow aA$	
10	$abcdbcabcd$	$S \rightarrow BdABd; A \rightarrow bc; B \rightarrow aA$	Bd twice
		$S \rightarrow CAC; A \rightarrow bc; B \rightarrow aA; C \rightarrow Bd$	B only once
		$S \rightarrow CAC; A \rightarrow bc; C \rightarrow aAd$	