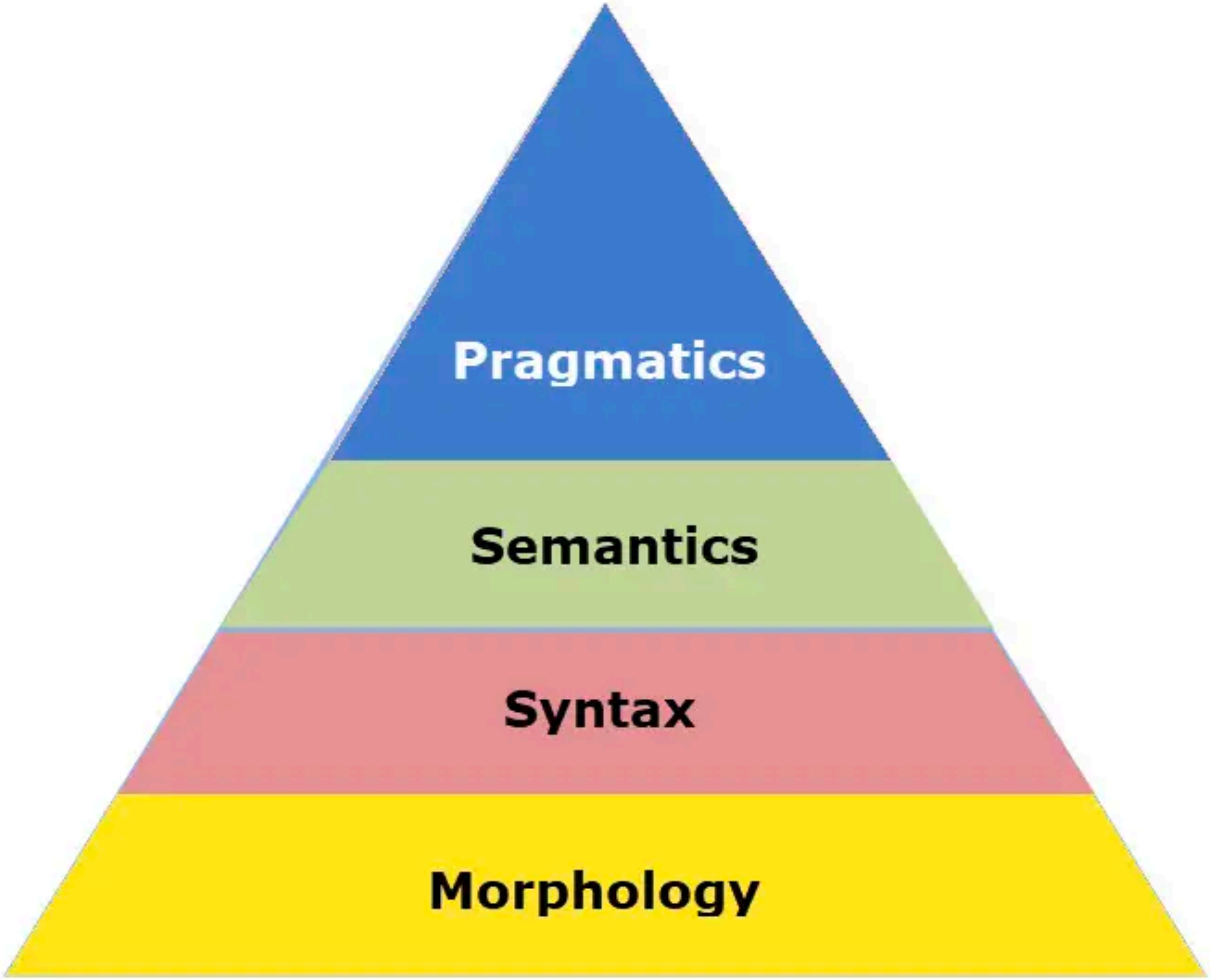
Lecture 11

Syntax - Structure of sentences

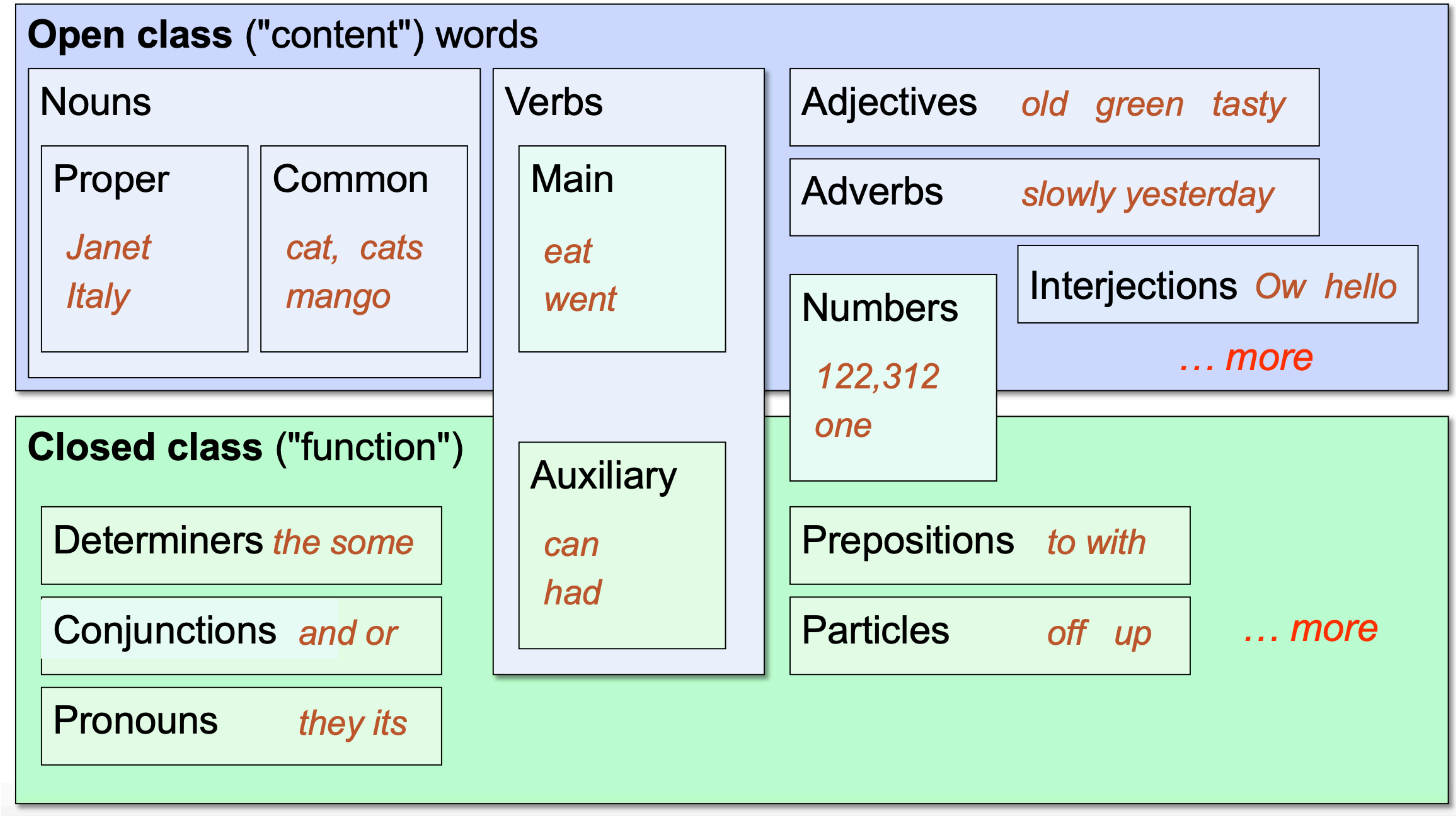
Zhizheng Wu

Agenda

- Recap
- Concept of syntax and constituency
- Context-free grammar
- Cocke-Kasami-Younger (CKY) algorithm



Natural Language Processing Pyramid


Open class ("content") words
Nouns
Proper
Janet
Italy
Common
cat, cats
mango
Verbs
Main
eat
went
Auxiliary
can
had
Adjectives old green tasty
Adverbs slowly yesterday
Interjections Ow hello
Numbers
122,312
one
… more
Closed class ("function")
Determiners the some
Conjunctions and or
Pronouns they its
Prepositions to with
Particles off up
… more

Part-of-speech tagging is a disambiguation process

Verb or Noun? ↑ (reading)

Verb or Noun? ↑ (Reading)

She is **reading** a book about **Reading**

One morning I shot an elephant in my pajamas

https://www.youtube.com/watch?v=NfN_gcjGoJo

One morning I shot an elephant <u>in my pajamas</u>

How he got into my pajamas I don’t know

Syntax is not Morphology

- Morphology deals with the internal structure of words
- Syntax deals with combinations of words

- Morphology is usually irregular
- Syntax has its irregularities, but it is usually regular
  - Syntax is mostly made up of general rules that apply across-the-board

Constituency

- One way of viewing the structure of a sentence is as a collection of nested constituents
- Constituent: a group of neighboring words relate more closely to one another than to other words in the sentence
- Constituents larger than a word are called phrases
  - Noun phrases
  - Prepositional phrases
  - Verb phrases
- Phrases can contain other phrases

Noun phrase (NP)

- a phrase that has a noun or pronoun as its head or performs the same grammatical function as a noun

  - The elephant arrived
  - It arrived.
  - Elephants arrived.
  - The big pretty elephant arrived.
  - The elephant she loves arrived.

Prepositional phrase (PP)

- I arrived on Tuesday.
- I arrived in March.
- I arrived under the leaking roof.

- Every prepositional phrase contains a noun phrase

Verb phrase

- A verb phrase in English consists of a verb followed by assorted other things
  - VP → Verb NP
    - I prefer an afternoon lecture
  - VP → Verb NP PP
    - have a lecture in the afternoon
  - VP → Verb PP
    - Teaching on Tuesday

Is a string constituent?

- Substitution test
  - Can the string be replaced by a single word?
- Movement test
  - Can the string be moved around in the sentence?
- Answer test
  - Can the string be the answer to a question?

He talks [in class]

He talks there

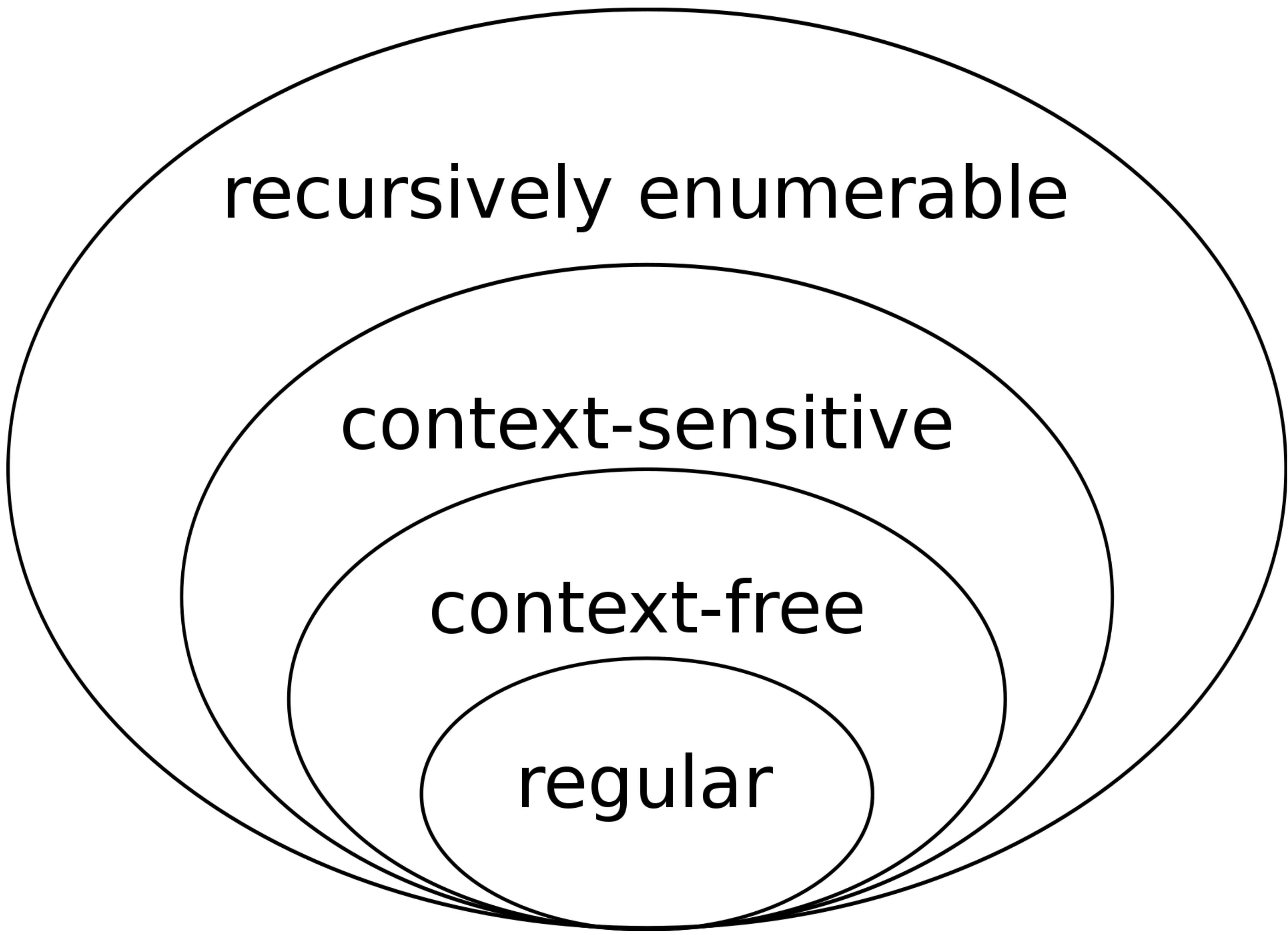
[In class], he talks

Where does he talk?

[In class]

Chomsky hierarchy

- Type-0 grammars include all formal grammars
- Type-1 grammars generate context-sensitive languages
- Type-2 grammars generate the context-free languages
- Type-3 grammars generate the regular languages, which can be described using regular expressions



Context-free grammar

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

Rules or productions

- Context-free
  - production rules are independent of the context
  - There is no context in the left hand side (LHS) of rules

Grammar Rules		
S	→	*NP VP*
NP	→	*Pronoun*
	|	*Proper-Noun*
	|	*Det Nominal*
Nominal	→	*Nominal Noun*
	|	*Noun*
VP	→	*Verb*
	|	*Verb NP*
	|	*Verb NP PP*
	|	*Verb PP*
PP	→	*Preposition NP*

Terminal vs Non-terminal

- **Terminal**: The symbols that that correspond to words in the language

- **Non-terminal**: The symbols that express abstractions over these terminals

VP → Verb NP

Non-terminal

Noun → lecture

Non-terminal Terminal

Verb → teaching

Non-terminal Terminal

Lexicon: Terminal vs Non-terminal

$$
\begin{aligned}
\textit{Noun} &\rightarrow \textit{flights} \mid \textit{flight} \mid \textit{breeze} \mid \textit{trip} \mid \textit{morning} \\
\textit{Verb} &\rightarrow \textit{is} \mid \textit{prefer} \mid \textit{like} \mid \textit{need} \mid \textit{want} \mid \textit{fly} \mid \textit{do} \\
\textit{Adjective} &\rightarrow \textit{cheapest} \mid \textit{non-stop} \mid \textit{first} \mid \textit{latest} \\
&\quad \mid \textit{other} \mid \textit{direct} \\
\textit{Pronoun} &\rightarrow \textit{me} \mid \textit{I} \mid \textit{you} \mid \textit{it} \\
\textit{Proper-Noun} &\rightarrow \textit{Alaska} \mid \textit{Baltimore} \mid \textit{Los Angeles} \\
&\quad \mid \textit{Chicago} \mid \textit{United} \mid \textit{American} \\
\textit{Determiner} &\rightarrow \textit{the} \mid \textit{a} \mid \textit{an} \mid \textit{this} \mid \textit{these} \mid \textit{that} \\
\textit{Preposition} &\rightarrow \textit{from} \mid \textit{to} \mid \textit{on} \mid \textit{near} \mid \textit{in} \\
\textit{Conjunction} &\rightarrow \textit{and} \mid \textit{or} \mid \textit{but}
\end{aligned}
$$

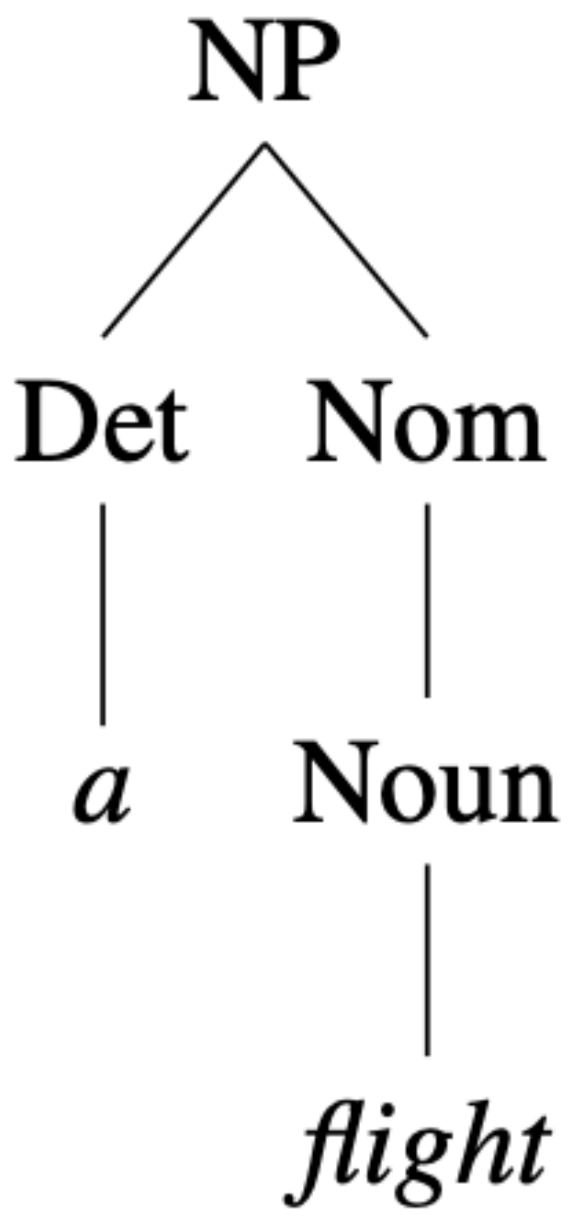
S: Start symbol

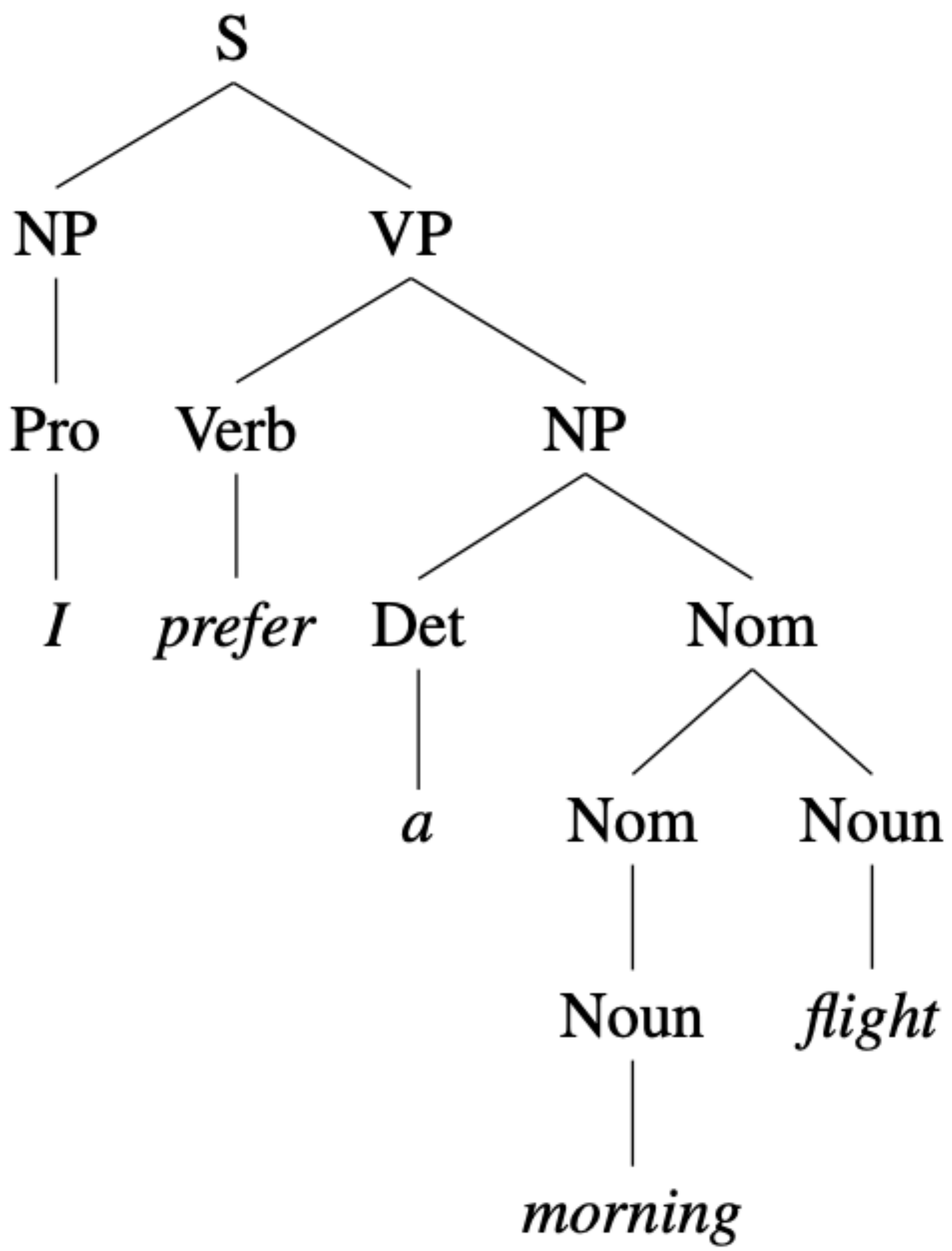
- The formal language defined by a CFG is the set of strings that are derivable from the designated start symbol
- Each grammar must have one designated start symbol
- S is usually interpreted as the “sentence” node

$$S \rightarrow NP\ VP$$

I prefer an afternoon lecture

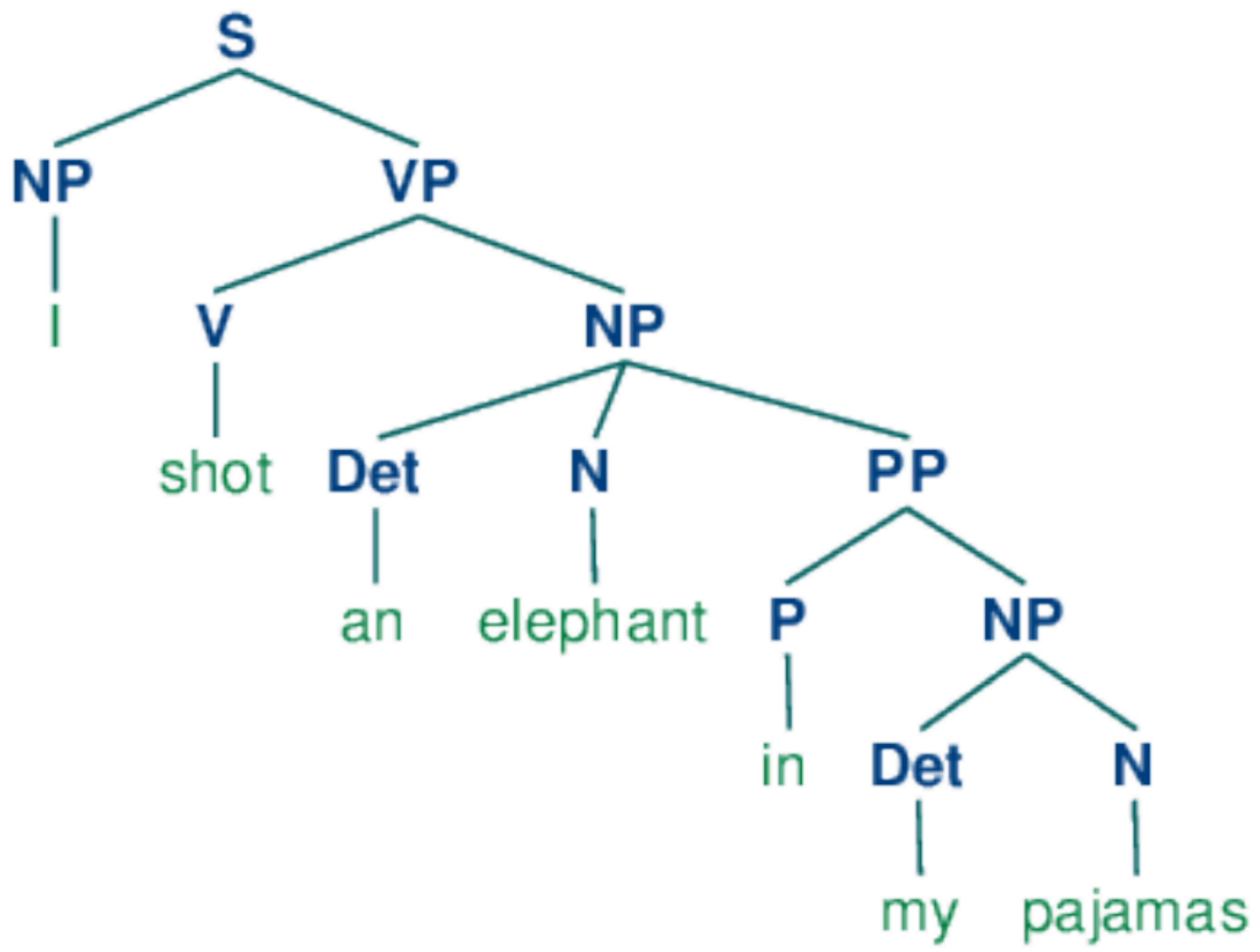
A (Constituency) Parse Tree

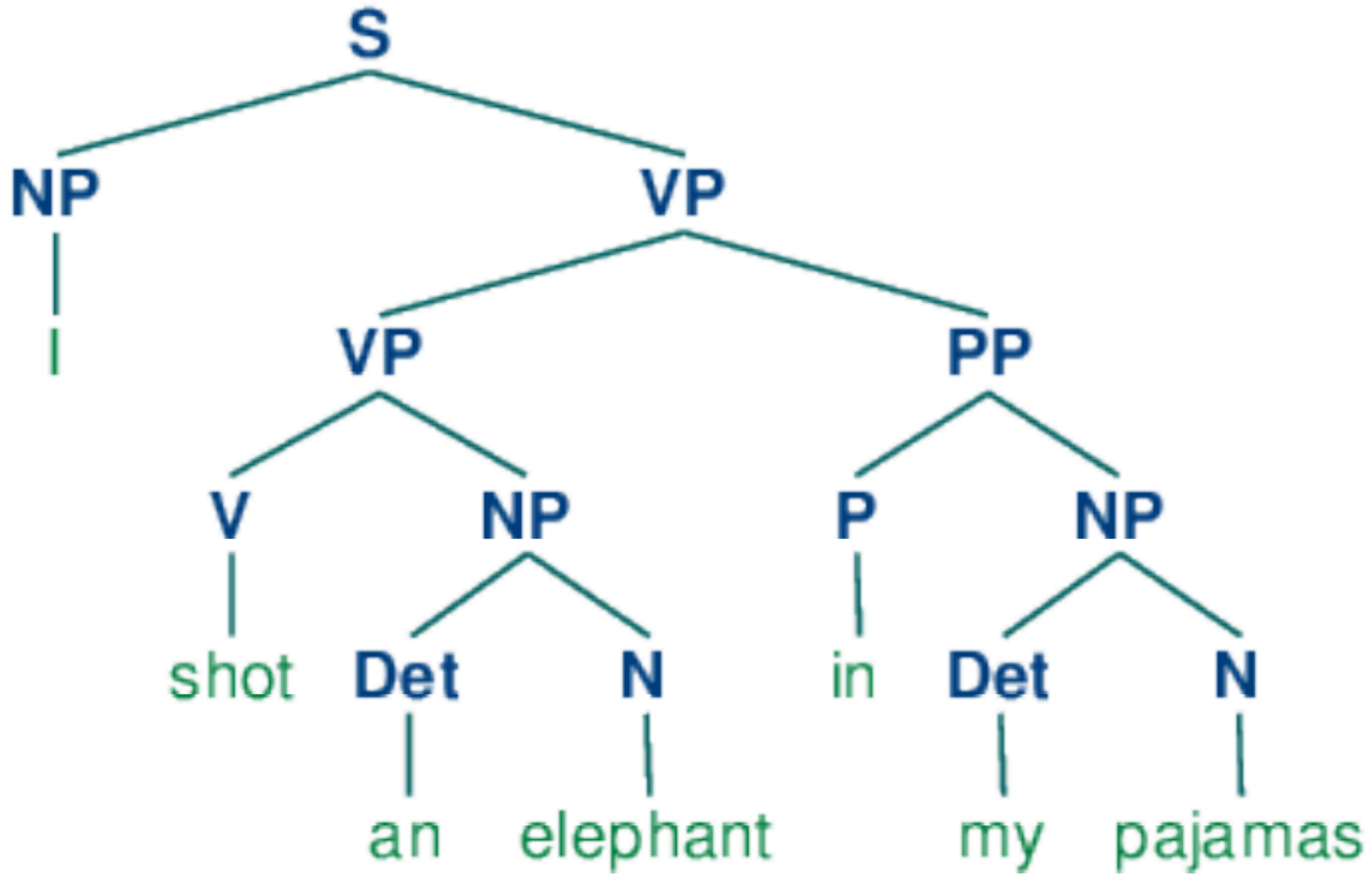




Ambiguity

- Structural ambiguity occurs when the grammar can assign more than one parse to a sentence





Cocke-Kasami-Younger (CKY) algorithm

- Bottom-up parsing
  - Start with words

- Dynamic programming
  - save the results in a table/chart
  - re-use these results in finding larger constituents

- Presumes a CFG in Chomsky Normal Form

Chomsky Normal Form

N	Finite set of non-terminal symbols	NP, VP, S
Σ	Finite alphabet of terminal symbols	the, dog, a
R	Set of production rules, each $A \rightarrow \beta$ β = single terminal (from Σ) or two non-terminals (from *N*)	$S \rightarrow NP\ VP$ $Noun \rightarrow dog$
S	Start symbol	

Chomsky Normal Form (CNF)

- Any CFG can be converted into weakly equivalent CNF
- In CNF, each non-terminal generates two non-terminals

$$A \rightarrow B\ C\ \gamma$$

$$\begin{aligned} A &\rightarrow X1\ \gamma \\ X1 &\rightarrow B\ C \end{aligned}$$

$S \rightarrow Aux\ NP\ VP$	$S \rightarrow X1\ VP$ $X1 \rightarrow Aux\ NP$

Chomsky Normal Form (CNF)

- Left hand side (LHS) rules
  - LHS will have non-terminals
- Right hand side (RHS) rules
  - Two non-terminals
  - One terminal

VP → Verb NP PP ✗

VP → Verb NP ✔

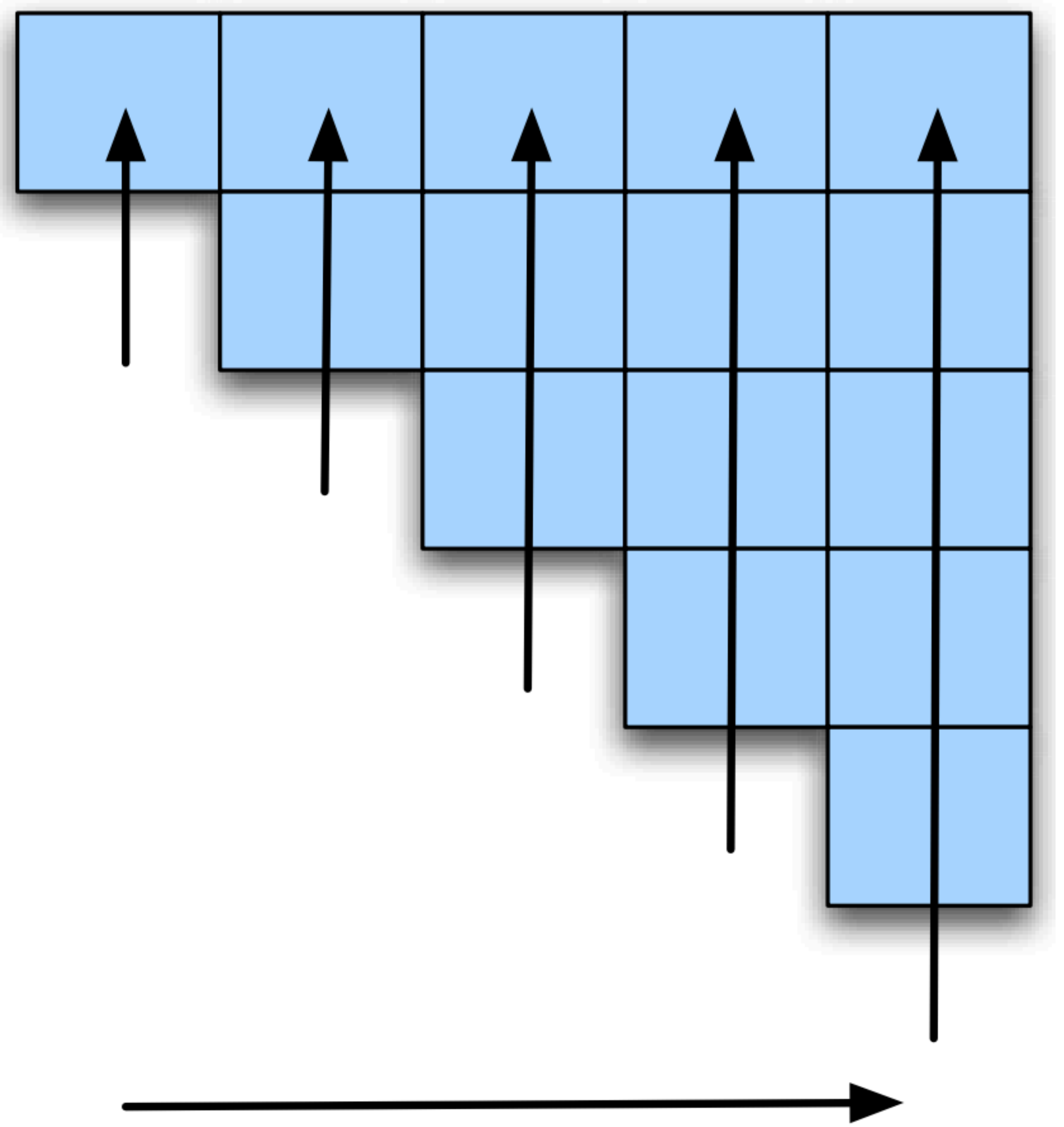
VP → teaching NP ✗

VP → eat ✔

$\mathscr{L}_1$ **Grammar**	$\mathscr{L}_1$ **in CNF**
$S \rightarrow NP\ VP$	$S \rightarrow NP\ VP$
$S \rightarrow Aux\ NP\ VP$	$S \rightarrow X1\ VP$
	$X1 \rightarrow Aux\ NP$
$S \rightarrow VP$	$S \rightarrow book \mid include \mid prefer$
	$S \rightarrow Verb\ NP$
	$S \rightarrow X2\ PP$
	$S \rightarrow Verb\ PP$
	$S \rightarrow VP\ PP$
$NP \rightarrow Pronoun$	$NP \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$NP \rightarrow TWA \mid Houston$
$NP \rightarrow Det\ Nominal$	$NP \rightarrow Det\ Nominal$
$Nominal \rightarrow Noun$	$Nominal \rightarrow book \mid flight \mid meal \mid money$
$Nominal \rightarrow Nominal\ Noun$	$Nominal \rightarrow Nominal\ Noun$
$Nominal \rightarrow Nominal\ PP$	$Nominal \rightarrow Nominal\ PP$
$VP \rightarrow Verb$	$VP \rightarrow book \mid include \mid prefer$
$VP \rightarrow Verb\ NP$	$VP \rightarrow Verb\ NP$
$VP \rightarrow Verb\ NP\ PP$	$VP \rightarrow X2\ PP$
	$X2 \rightarrow Verb\ NP$
$VP \rightarrow Verb\ PP$	$VP \rightarrow Verb\ PP$
$VP \rightarrow VP\ PP$	$VP \rightarrow VP\ PP$
$PP \rightarrow Preposition\ NP$	$PP \rightarrow Preposition\ NP$

CKY algorithm

- Fills the upper-triangular matrix a column at a time
  - From left to right
  - From bottom to top
- This scheme guarantees that at each point in time we have all the information we need



CKY algorithm: a toy example

Rules

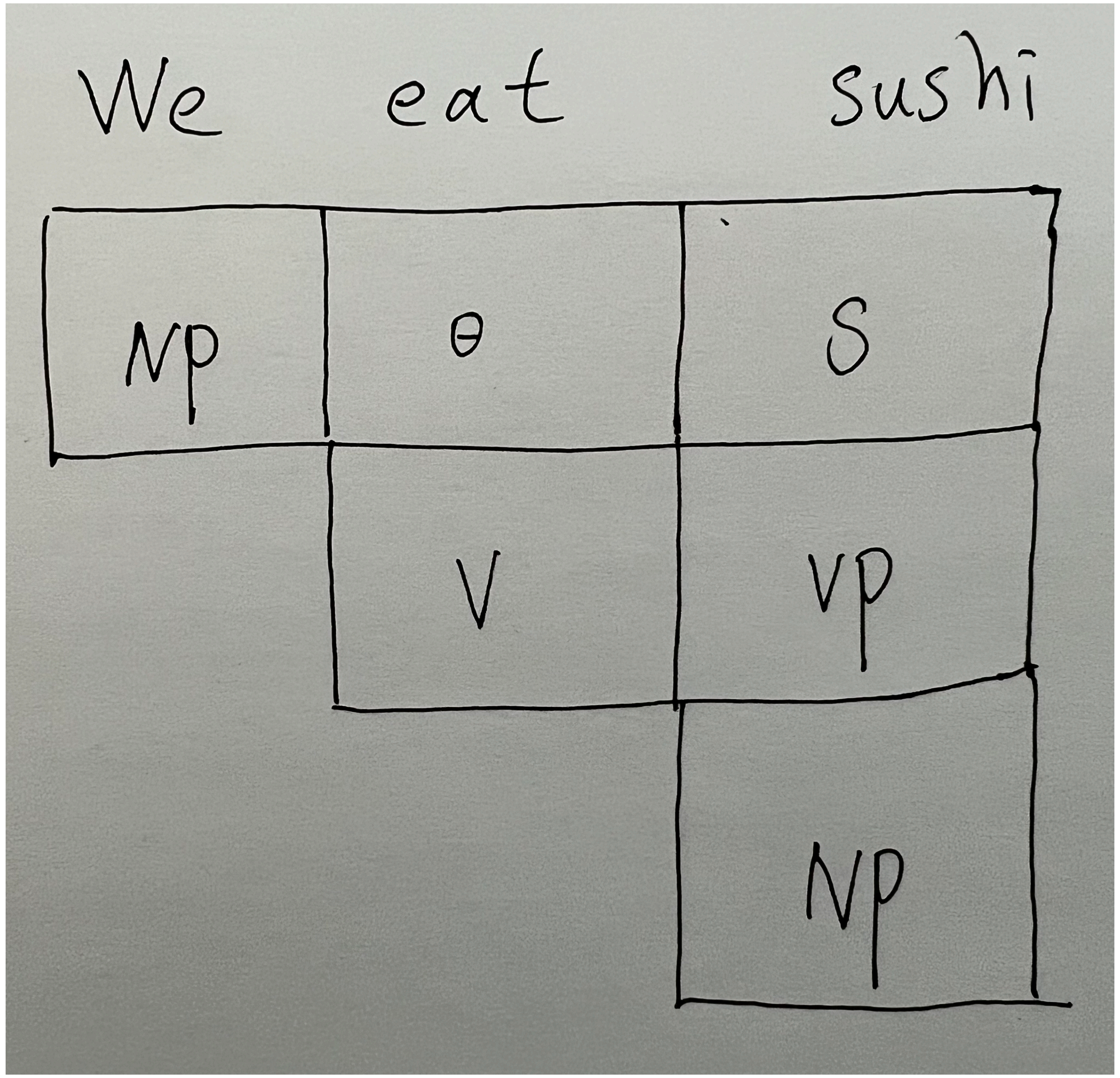
S → NP VP

VP → V NP

V → eat

NP → we

NP → sushi



CKY algorithm

function CKY-PARSE(*words, grammar*) **returns** *table*

for $j \leftarrow$ **from** 1 **to** LENGTH(*words*) **do**
    for all $\{A \mid A \rightarrow words[j] \in grammar\}$
        $table[j-1, j] \leftarrow table[j-1, j] \cup A$
    for $i \leftarrow$ **from** $j-2$ **down to** 0 **do**
        for $k \leftarrow i+1$ **to** $j-1$ **do**
            for all $\{A \mid A \rightarrow BC \in grammar$ **and** $B \in table[i,k]$ **and** $C \in table[k,j]\}$
                $table[i,j] \leftarrow table[i,j] \cup A$

https://web.stanford.edu/~jurafsky/slp3/17.pdf

CKY Exa

Noun → *flights* | *flight* | *breeze* | *trip* | *morning*
Verb → *is* | *prefer* | *like* | *need* | *want* | *fly* | *do*
Adjective → *cheapest* | *non-stop* | *first* | *latest* | *other* | *direct*
Pronoun → *me* | *I* | *you* | *it*
Proper-Noun → *Alaska* | *Baltimore* | *Los Angeles* | *Chicago* | *United* | *American*
Determiner → *the* | *a* | *an* | *this* | *these* | *that*
Preposition → *from* | *to* | *on* | *near* | *in*
Conjunction → *and* | *or* | *but*

Book	*the*	*flight*	*through*	*Houston*
S, VP, Verb, Nominal, Noun [0,1]	[0,2]	S,VP,X2 [0,3]	[0,4]	S,VP,X2 [0,5]
	Det [1,2]	NP [1,3]	[1,4]	NP [1,5]
		Nominal, Noun [2,3]	[2,4]	Nominal [2,5]
			Prep [3,4]	PP [3,5]
				NP, Proper-Noun [4,5]

$\mathscr{L}_1$ in CNF

S → *NP VP*
S → *X1 VP*
X1 → *Aux NP*
S → *book* | *include* | *prefer*
S → *Verb NP*
S → *X2 PP*
S → *Verb PP*
S → *VP PP*
NP → *I* | *she* | *me*
NP → *TWA* | *Houston*
NP → *Det Nominal*
Nominal → *book* | *flight* | *meal* | *money*
Nominal → *Nominal Noun*
Nominal → *Nominal PP*
VP → *book* | *include* | *prefer*
VP → *Verb NP*
VP → *X2 PP*
X2 → *Verb NP*
VP → *Verb PP*
VP → *VP PP*
PP → *Preposition NP*

ps://web.stanford.edu/~jurafsky/slp3/17.pdf

Exercise

- S ▸ NP VP
- PP ▸ IN NP
- NP ▸ DET NP
- NP ▸ NP PP
- VP ▸ VBD NP
- VP ▸ VP PP
- NP ▸ PRP$ NP

- DET ▸ "an"
- VBD ▸ "shot"
- NP ▸ "pajamas"
- NP ▸ "elephant"
- NP ▸ "I"
- PRP ▸ "I"
- IN ▸ "in"
- PRP$▸ "my"

I shot an elephant in my pajamas

Exercise

I	shot	an	elephant	in	my	pajamas
NP, PRP [0,1]						
	VBD [1,2]					
		DT [2,3]				
			NP, NN [3,4]			
				IN [4,5]		
					PRP$ [5,6]	
						NNS [6,7]

Summary

- Concept of syntax and constituency
  - Syntax deals with combinations of words
- Context-free grammar
  - production rules are independent of the context
- Cocke-Kasami-Younger (CKY) algorithm
  - Bottom-up parsing - start with words
  - Dynamic programming
  - Presumes a CFG in Chomsky Normal Form

Reading

- Chapter 17: Context-Free Grammars and Constituency Parsing
- https://web.stanford.edu/~jurafsky/slp3/17.pdf