Indian Institute of Technology Kharagpur
Department of Computer Science & Engineering

CS60075
Natural Language Processing
Autumn 2020

Module 5: Part A
Syntax

Sep 30 2020

What is Syntax?

- Structure of language
  - How words are arranged together and related to one another
  - Ordering words in sequences to express meanings for which no separate word exists.
- Goal of syntactic analysis
  - relate surface form to underlying structure, to support semantic analysis
- Syntactic representation
  - typically a tree structure

Regularities in language

- Word n-grams model **regularities in word sequences**
- Part-of-speech n-grams model **regularities in word category sequences.**
- Language has richer structure.
- Two views of linguistic structure:
  1. Constituency = phrase structure grammar = context-free grammars (CFGs)
  2. Dependency Structure
     - Dependency structure shows which words depend on (modify or are arguments of) which other words.

Phrase Structure Grammar

Phrase structure organizes words into nested constituents

- **Starting unit: words**

  the, cat, cuddly, by, door

- **Words combine into phrases**

  the cuddly cat, by the door

- **Phrases can combine into bigger phrases**

  the cuddly cat by the door

Phrase Structure Grammar

Phrase structure organizes words into nested constituents - Can represent the grammar with CFG rules

- **Starting unit: words** are given a category (part of speech = pos)

  the, cat, cuddly, by, door

  Det N Adj P N

- **Words combine into phrases** with categories

  the cuddly cat, by the door

  NP →Det Adj N PP →P NP

- **Phrases can combine into bigger phrases** recursively

  the cuddly cat by the door

  NP →NP PP

कार्यशाला	N
अनूठा	Adj

एक अनूठा संग्रहालय खुला रहता है

Dependency Structure

Dependency structure shows which words depend on (modify or are arguments of) which other words.

Put the book on the big table in the room next to the vase.

Why is Syntax Important?

- Many aspects of meaning can be learnt using the syntactic structure.
  - The NP preceding VP is likely the subject of the action.
  - The NP following the VP is likely the object of the action.
- Knowing basic units is helpful in modeling language.
  - You can use this to predict or complete the sentence.
  - Re-organize sentences or simplify them.

- Grammar checkers
- Question answering
- Information extraction
- Machine translation
- Semantic role labeling

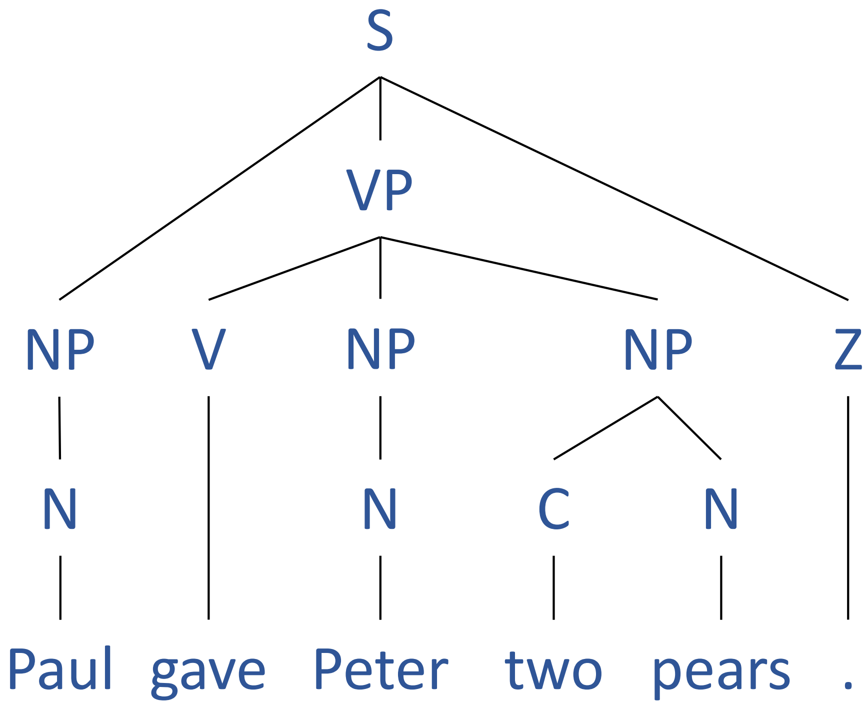
Syntactic Structure

- Different shapes in different theories
- Typically a tree
  - Phrasal (constituent) tree, parse tree
  - Dependency tree

Example of Constituent Tree

- Constituency: abstraction— groups of words behaving as a single units, or constituents

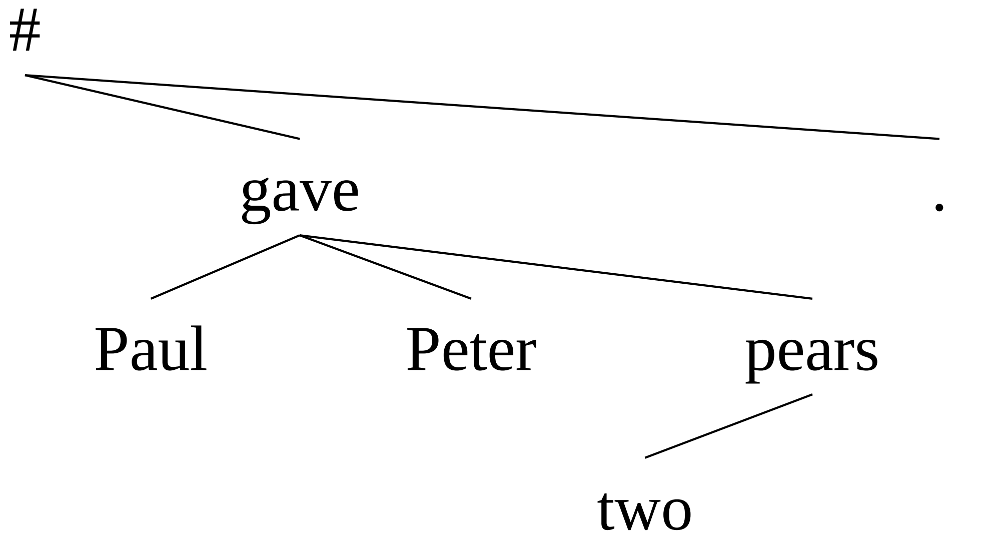
((Paul (gave Peter (two pears))) .)



Example of Dependency Tree

Paul gave Peter two pears

[#,0] ([gave,2] ([Paul,1], [Peter,3], [pears,5] ([two,4])), [.,6])

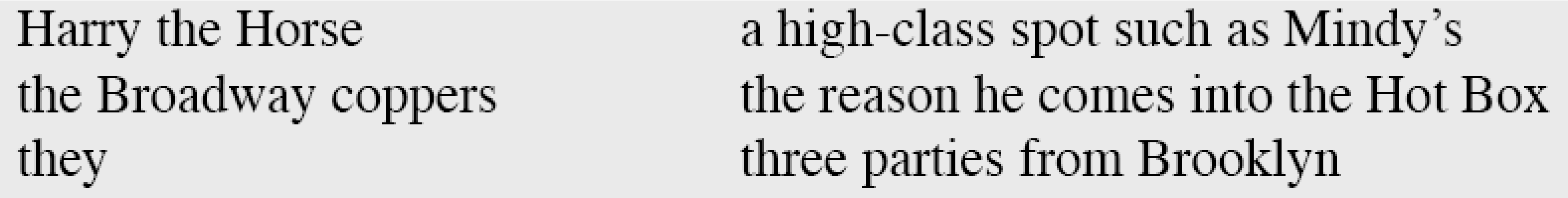


Constituency

- Basic idea: groups of words act as a single unit
- Constituents form coherent classes that behave similarly
  - With respect to their internal structure: e.g., at the core of a noun phrase is a noun
  - With respect to other constituents: e.g., noun phrases generally occur before verbs

Constituency: Example

- Noun phrases in English...



Harry the Horse
the Broadway coppers
they

a high-class spot such as Mindy's
the reason he comes into the Hot Box
three parties from Brooklyn

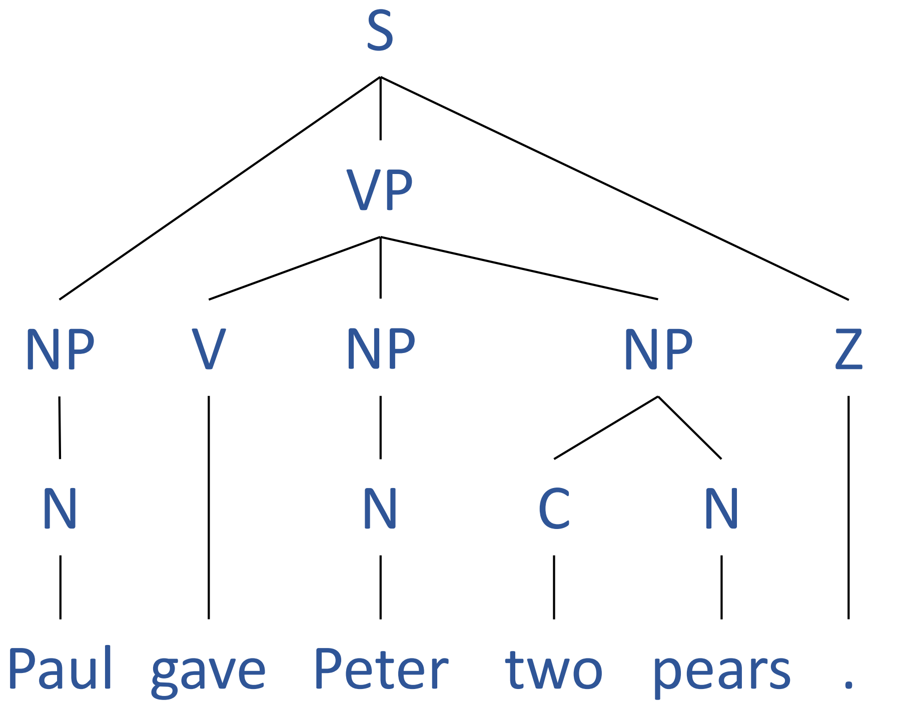
- They can all precede verbs
- They can all be preposed/postposed

Ram ke ve tiin kaale ghore
Mandir kaa udghaatan

Example of Constituent Tree

- Constituency: abstraction— groups of words behaving as a single units, or constituents

((Paul (gave Peter (two pears))) .)



Words and Phrases

- Word (token): smallest unit of the syntactic layer
  - grammatical (function) words
  - lexical (content) words
- Phrase
  - Sequence of immediate constituents (words or phrases).
- Phrase types by their main word—head
  - Noun phrase: *the new book of my grandpa*
  - Adjectival phrase: *brand new*
  - Adverbial phrase: *very well*
  - Prepositional phrase: *in the classroom*
  - Verb phrase: *to catch a ball*

Noun Phrase

- A noun or a (substantive) pronoun is the head.
  - *water*
  - *the book*
  - *new ideas*
  - *two millions of inhabitants*
  - *one small village*
  - *the greatest price movement in one year since the World War II*
  - *operating system that, regardless of all efforts by our admin, crashes just too often*
  - *he*
  - *whoever*

Evidence of Constituency

1. They can all appear in similar syntactic environments
   - NP before a verb

2. Preposed or Postposed constructions
   - The prepositional phrase can be placed in a number of different locations in the sentence
   - But the individual words in the phrase cannot.

Adjective Phrase

- An adjective or a determiner (attributive pronoun) is the head.
- Simple ADJPs are very frequent, complex ones are rare.
  - *old*
  - *very old*
  - *really very old*
  - *five times older than the oldest elephant in our ZOO*
  - *sure that he will arrive first*

Adverbial Phrases

- An adverb is the head.
  - *quickly*
  - *much more*
  - *how*
  - *louder than you can imagine*
  - *yesterday*

Prepositional (Postpositional) Phrase

- The preposition serves as head (because it determines the case of the rest of the phrase).
- Often have a function similar to adverbial phrases or noun phrases (object of a verb).
  - *in the city center*
  - *in God*
  - *around five o’clock*
  - *to a better future*
  - *up to a situation where neither of them could back out*
  - *with respect to his nonage*

Clause and Sentence

- Group of words with 1 predicate, e.g.:
  - *John loves Mary.*
  - *…that you are right.*
  - simple sentence or part of compound sentence
- Sentence
  - simple sentence or compound sentence
  - consists of one or more clauses
  - e.g. *John loves Mary.* or *“I realized that you were right.”*

Clause and Sentence

- Main clause
  - Independent of other clauses in the sentence
- Nested clause, relative clause
  - Depends on another clause, carries out a function in that clause (as a dependent phrase)
  - This is the man [that] I saw

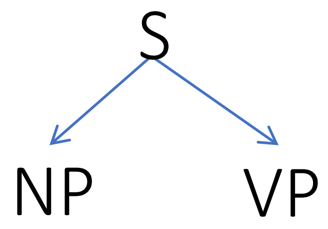
Sentence

  - Consists of one or more main clauses.
  - If there are more than one main clause then they are usually coordinated.

Formal Grammars of English

Context-free grammars (CFGs)

- Consist of
  - Rules
  - Terminals
  - Non-terminals
  - Start Symbol

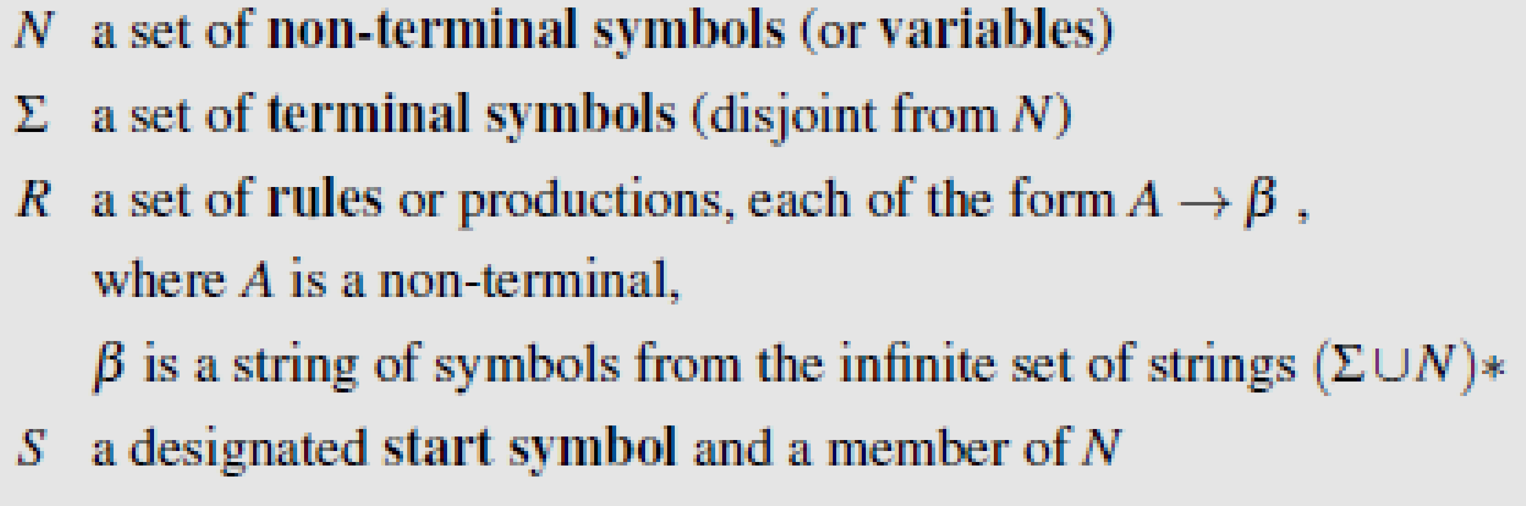


S

NP

VP

- Specifies a set of tree structures that capture constituency and ordering in language



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

Productions of CFG

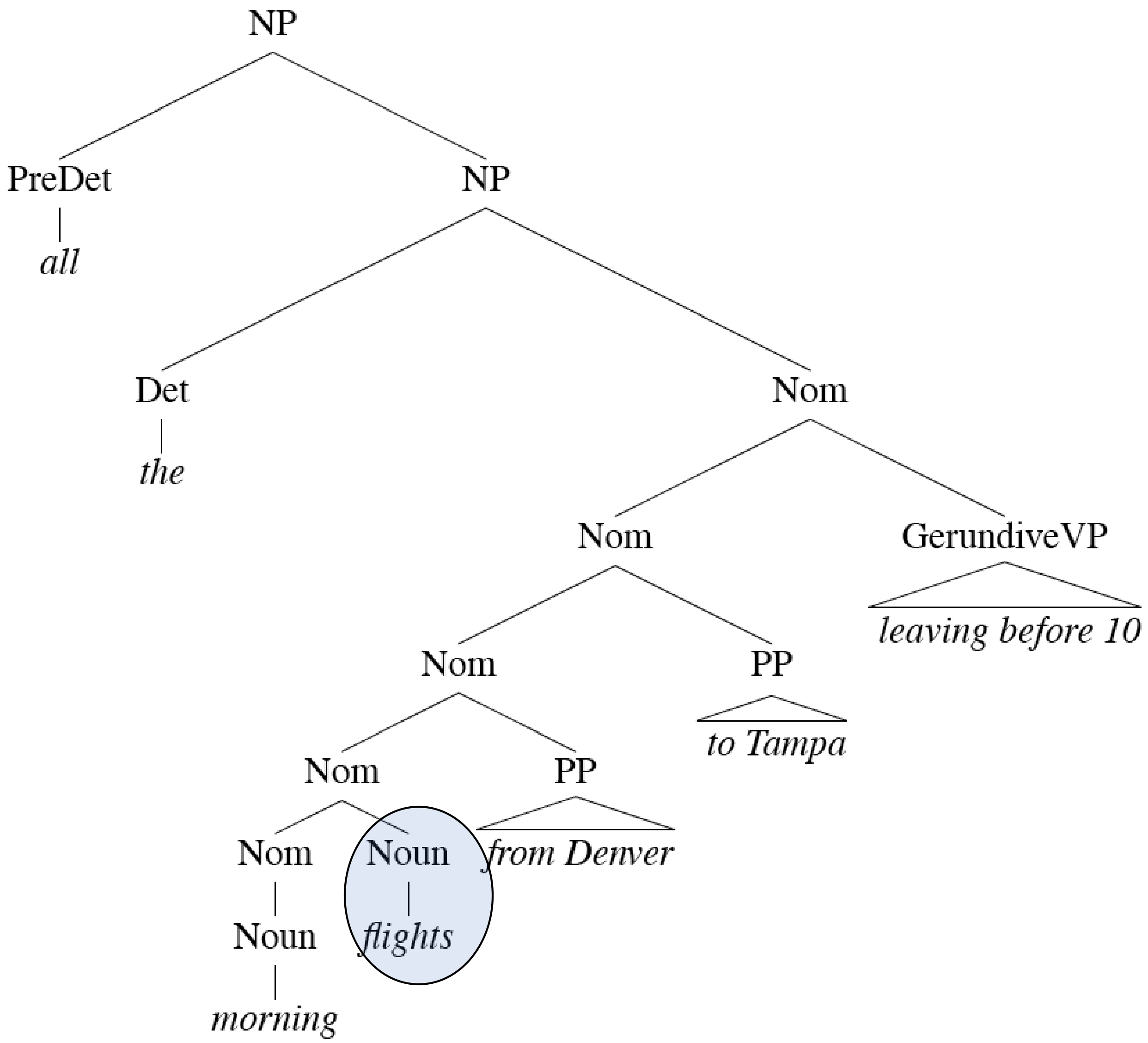
- A CFG can be thought of in two ways:
  - a device for generating sentences

    (Derivation)
  - a device for assigning a structure to a given sentence.

- Some rules for noun phrases:

$$
\begin{array}{rcl}
NP & \rightarrow & Det\ Nominal \\
NP & \rightarrow & ProperNoun \\
Nominal & \rightarrow & Noun \mid Nominal\ Noun
\end{array}
$$

Noun Phrases



Nominals

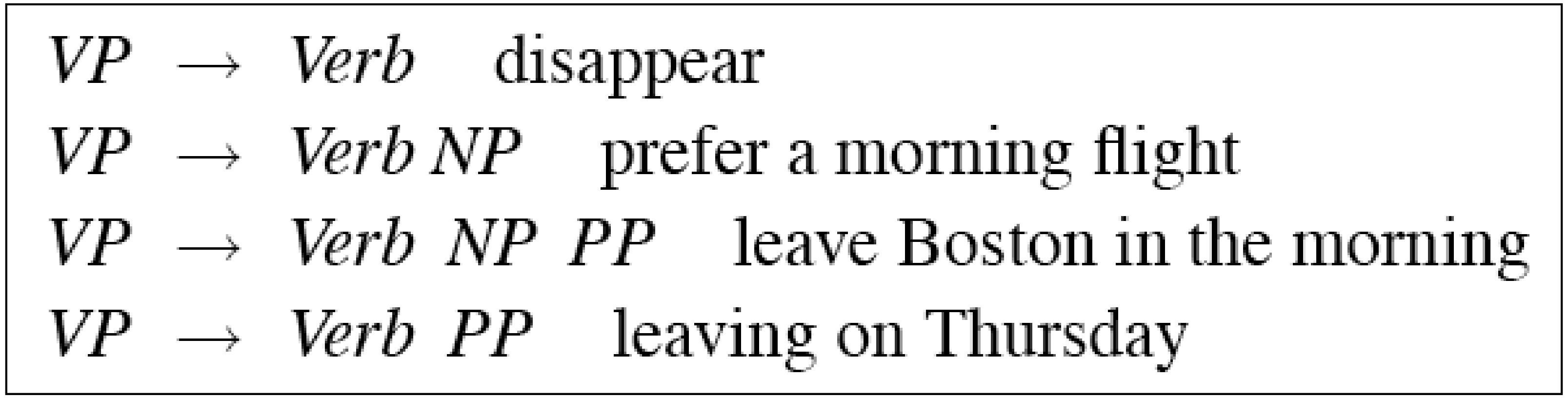
- Contain the head and any pre- and post- modifiers of the head.
    - Pre-
        - Quantifiers, cardinals, ordinals...
            - *Three cars*
        - Adjectives
            - *large cars*

Postmodifiers

- Three kinds
  - Prepositional phrases
    - *From Seattle*
  - Non-finite clauses
    - *Arriving before noon*
  - Relative clauses
    - *That serve breakfast*
- Same general (recursive) rules to handle these
  - *Nominal → Nominal PP*
  - *Nominal → Nominal GerundVP*
  - *Nominal → Nominal RelClause*

Verb Phrases

- English *VP*s consist of a verb (the head) along with 0 or more *following* constituents which we'll call *arguments*.



$VP \rightarrow Verb$		disappear
$VP \rightarrow Verb\ NP$		prefer a morning flight
$VP \rightarrow Verb\ NP\ PP$		leave Boston in the morning
$VP \rightarrow Verb\ PP$		leaving on Thursday

Subcategorization

- Even though there are many valid VP rules in English, not all verbs are allowed to participate in all those VP rules.
- We can *subcategorize* the verbs in a language according to the sets of VP rules that they participate in.
- This is just an elaboration on the traditional notion of transitive/intransitive.
- Modern grammars have many such classes

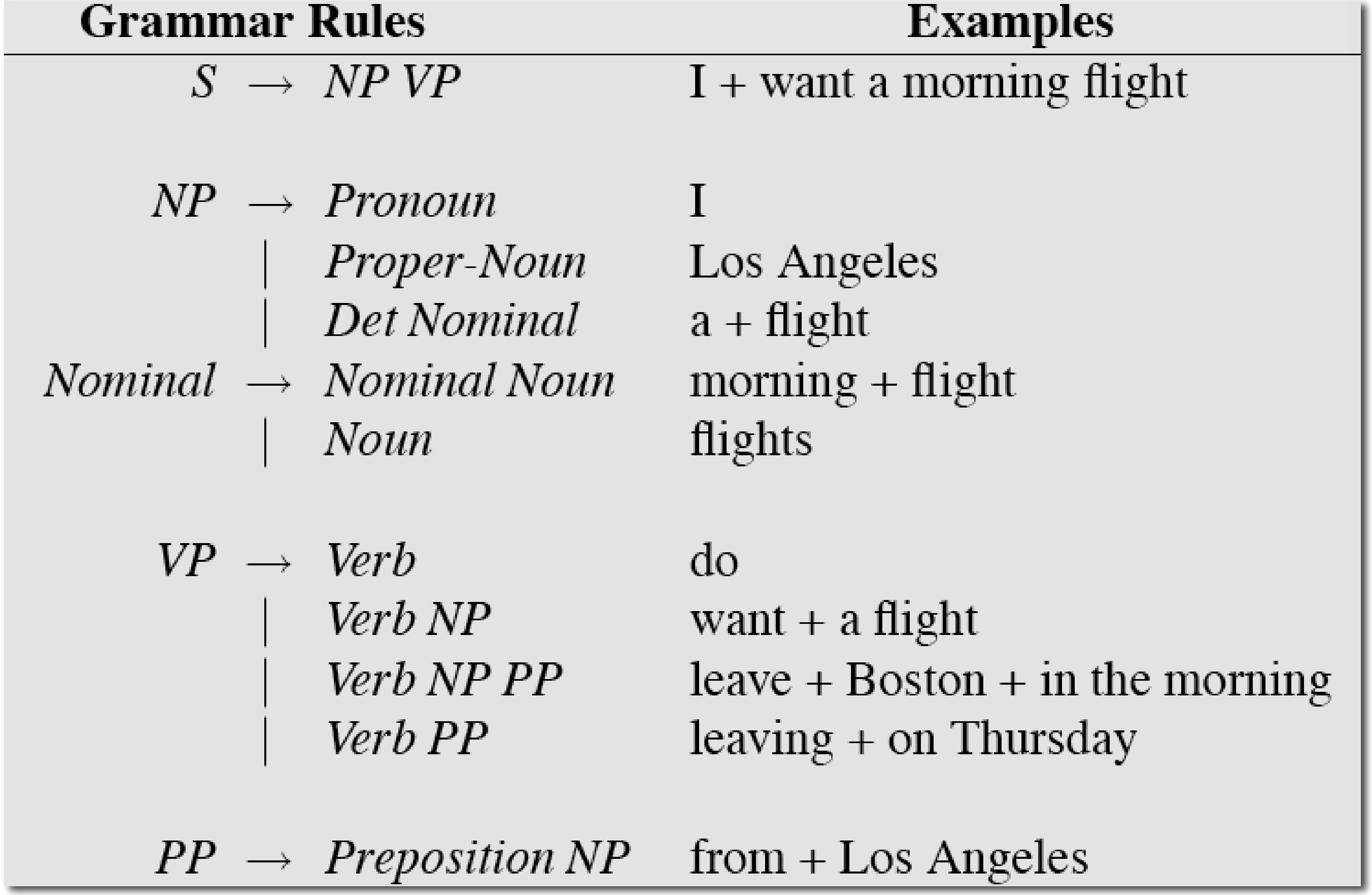
Subcategorization

- Sneeze: John sneezed
- Find: Please find [a flight to NY]$_{NP}$
- Give: Give [me]$_{NP}$[a cheaper fare]$_{NP}$
- Help: Can you help [me]$_{NP}$[with a flight]$_{PP}$
- Prefer: I prefer [to leave earlier]$_{TO\text{-}VP}$
- Told: I was told [United has a flight]$_{S}$
- …

Generative Grammar

- The use of formal languages to model Generative natural languages is called ***generative grammar*** since the language is defined by the set of possible sentences “generated” by the grammar.
- You can view these rules as either analysis or synthesis engines
  - Generate strings in the language
  - Reject strings not in the language
  - Assign structures (trees) to strings in the language

L0 Grammar



Grammar Rules			Examples
S	→	*NP VP*	I + want a morning flight
NP	→	*Pronoun*	I
	$\mid$	*Proper-Noun*	Los Angeles
	$\mid$	*Det Nominal*	a + flight
Nominal	→	*Nominal Noun*	morning + flight
	$\mid$	*Noun*	flights
VP	→	*Verb*	do
	$\mid$	*Verb NP*	want + a flight
	$\mid$	*Verb NP PP*	leave + Boston + in the morning
	$\mid$	*Verb PP*	leaving + on Thursday
PP	→	*Preposition NP*	from + Los Angeles

Sentence Types

- Declaratives: *A plane left.*

  S → NP VP

- Imperatives: *Leave!*

  S → VP

- Yes-No Questions: *Did the plane leave?*

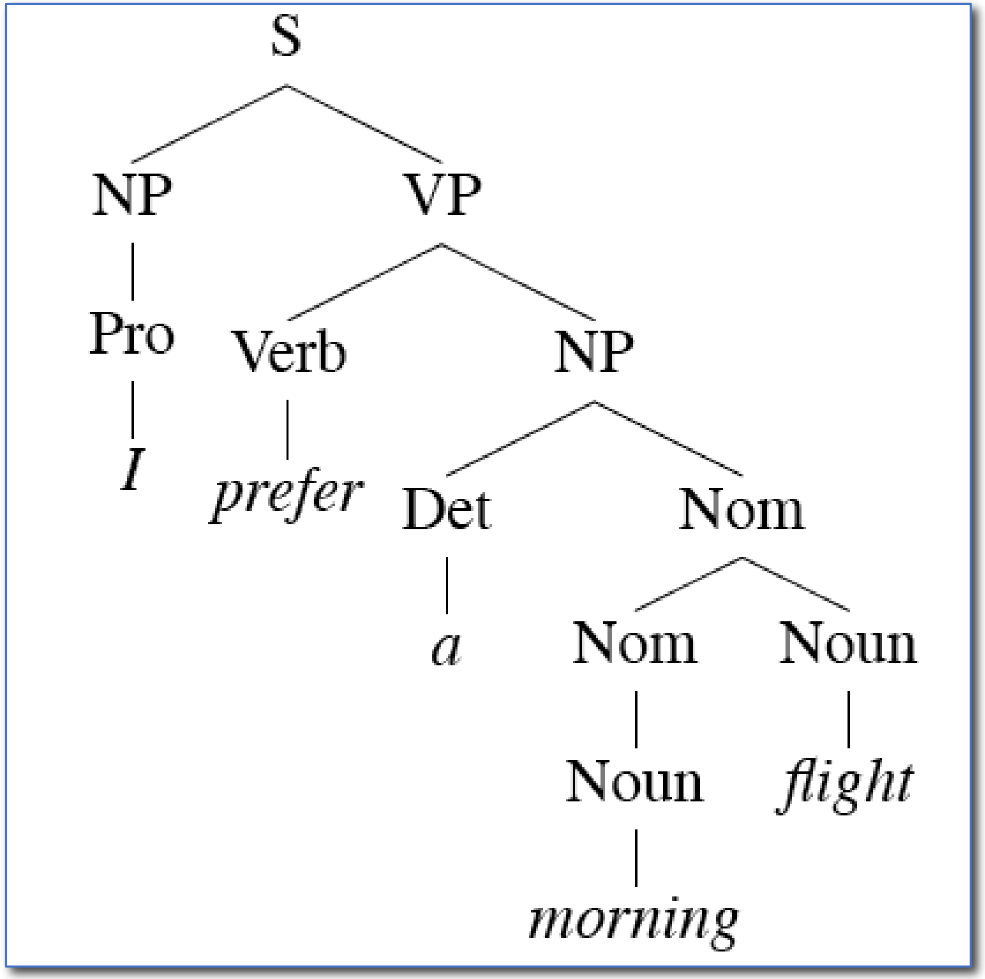
  S → Aux NP VP

- WH Questions: *When did the plane leave?*

  S → WH-NP Aux NP VP

Derivations

- A *derivation* is a sequence of rules applied to a string that *accounts* for that string
    - Covers all the elements in the string
    - Covers only the elements in the string



Parsing

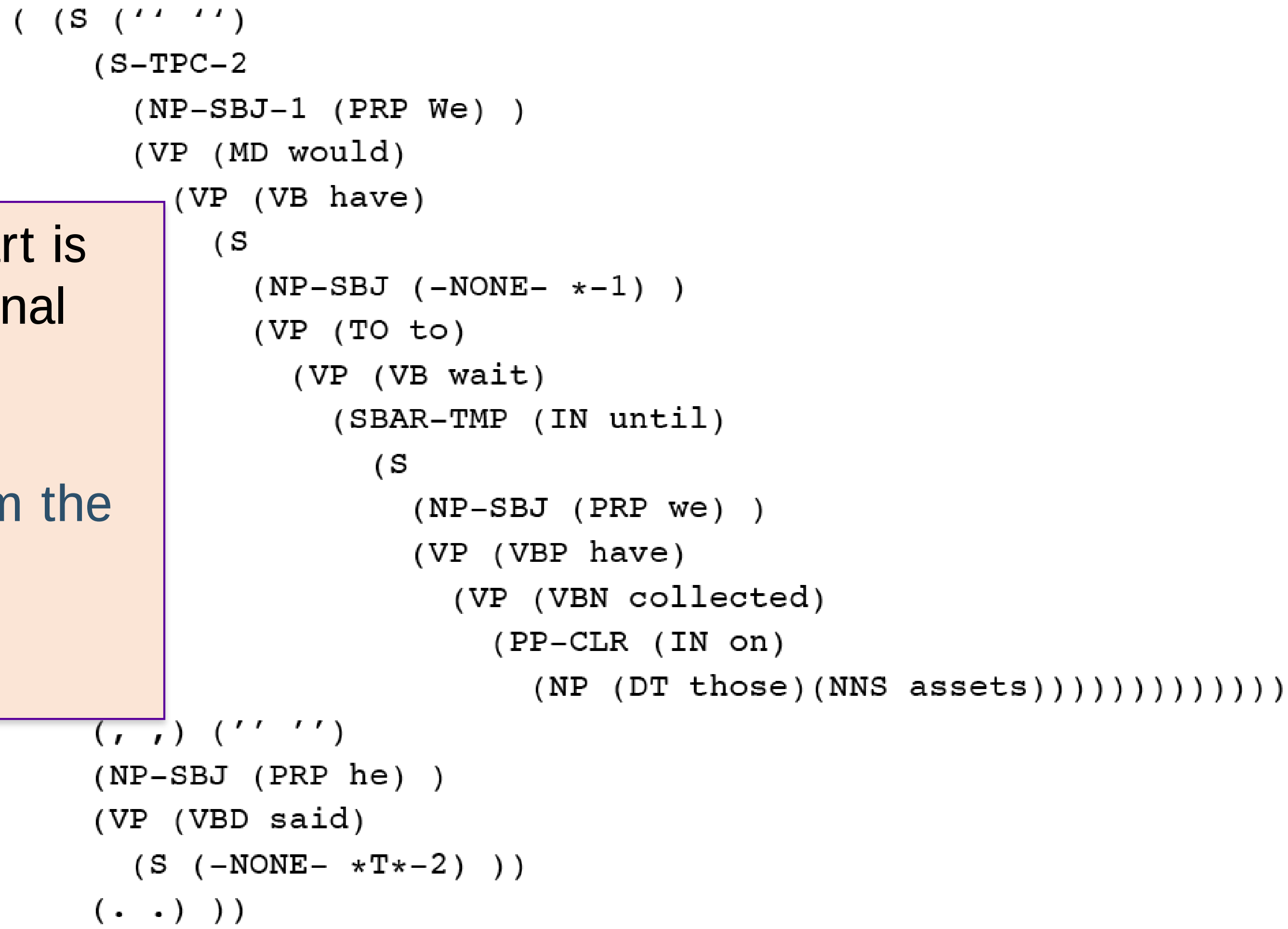
- Parsing is the process of taking a string and a grammar and returning parse tree(s) for that string

Treebank

- A syntactically annotated corpus where every sentence is paired with a corresponding tree.
- The Penn Treebank project
  - treebanks from the Brown, Switchboard, ATIS, and Wall Street Journal corpora of English
  - treebanks in Arabic and Chinese.
- Others
  - the Prague Dependency Treebank for Czech,
  - the Negra treebank for German, and
  - the Susanne treebank for English
  - Universal Dependencies Treebank

Penn Treebank

- Penn TreeBank is a widely used treebank.



```
( (S (‘‘ ‘‘)
    (S-TPC-2
      (NP-SBJ-1 (PRP We) )
      (VP (MD would)
        (VP (VB have)
          (S
            (NP-SBJ (-NONE- *-1) )
            (VP (TO to)
              (VP (VB wait)
                (SBAR-TMP (IN until)
                  (S
                    (NP-SBJ (PRP we) )
                    (VP (VBP have)
                      (VP (VBN collected)
                        (PP-CLR (IN on)
                          (NP (DT those)(NNS assets)))))))))))))
    (, ,) (’’ ’’)
    (NP-SBJ (PRP he) )
    (VP (VBD said)
      (S (-NONE- *T*-2) ))
    (. .) ))
```

Most well known part is the Wall Street Journal section of the Penn TreeBank.

- 1 M words from the 1987-1989 Wall Street Journal.

```
((S
   (NP-SBJ (DT That)
     (JJ cold) (, ,)
     (JJ empty) (NN sky) )
   (VP (VBD was)
     (ADJP-PRD (JJ full)
       (PP (IN of)
         (NP (NN fire)
           (CC and)
           (NN light) ))))
   (. .) ))
```

(a)

```
((S
   (NP-SBJ The/DT flight/NN )
   (VP should/MD
     (VP arrive/VB
       (PP-TMP at/IN
         (NP eleven/CD a.m/RB ))
       (NP-TMP tomorrow/NN )))))
```

(b)

Figure 11.7 Parsed sentences from the LDC Treebank3 version of the Brown (a) and ATIS (b) corpora.

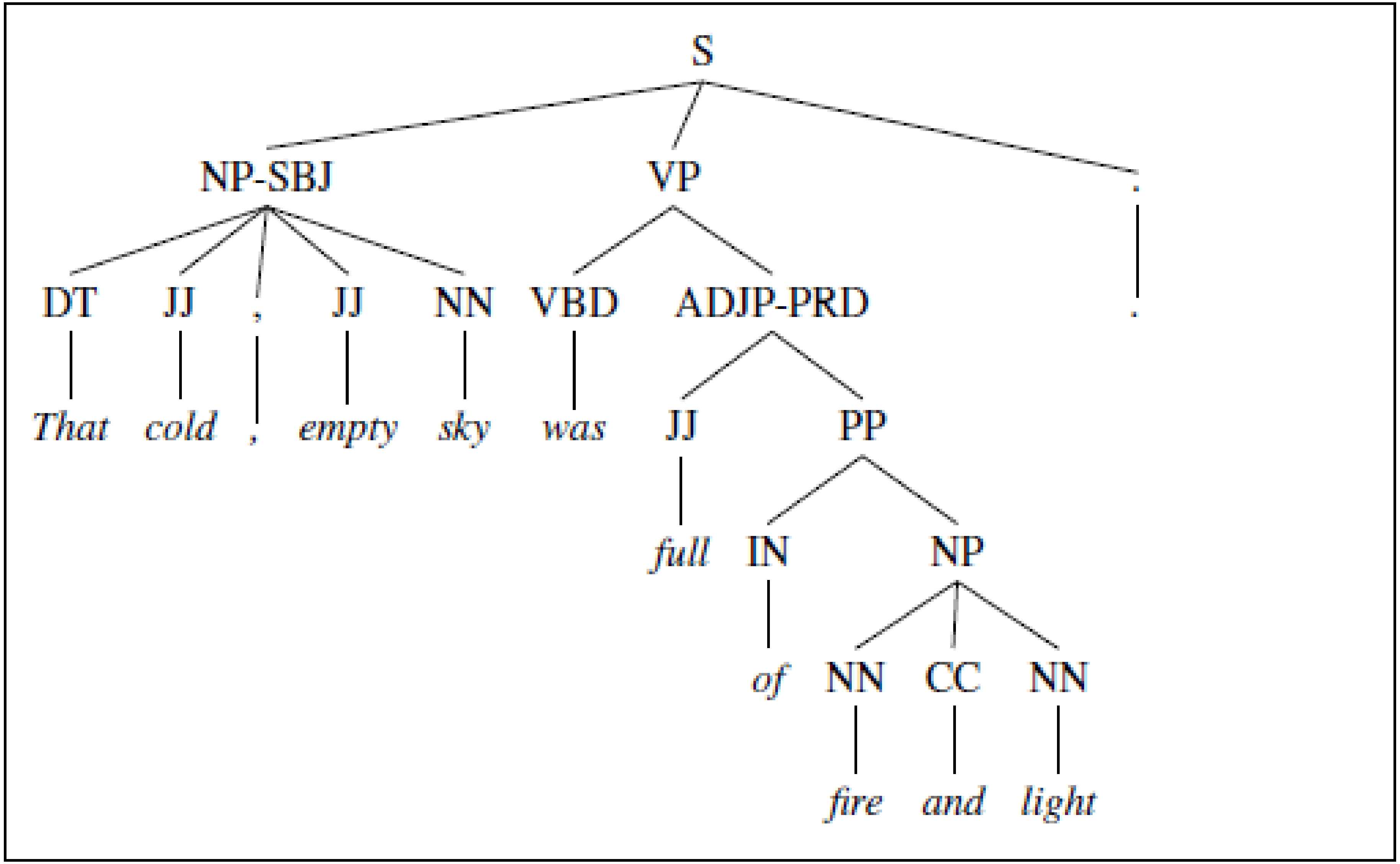


Figure 11.8 The tree corresponding to the Brown corpus sentence in the previous figure.

Treebanks as Grammars

- The sentences in a treebank implicitly constitute a grammar of the language represented by the corpus being annotated.

- Simply take the local rules that make up the sub-trees in all the trees in the collection and you have a grammar
    - The WSJ section gives us about 12k rules

Parsing

- Parsing with CFGs refers to the task of assigning proper trees to input strings
- Proper here means a tree that covers all and only the elements of the input and has an S at the top
- It doesn’t mean that the system can select the correct tree from among all the possible trees

Treebanks as Grammars

- The sentences in a treebank implicitly constitute a grammar of the language represented by the corpus being annotated.
- Simply take the local rules that make up the sub-trees in all the trees in the collection and you have a grammar
  - The WSJ section gives us about 12k rules if you do this
- Treebanks (and head-finding) are particularly critical to the development of statistical parsers