Foundations of Language Science and Technology (FLST)

Lecture 6 (06.11.2008)

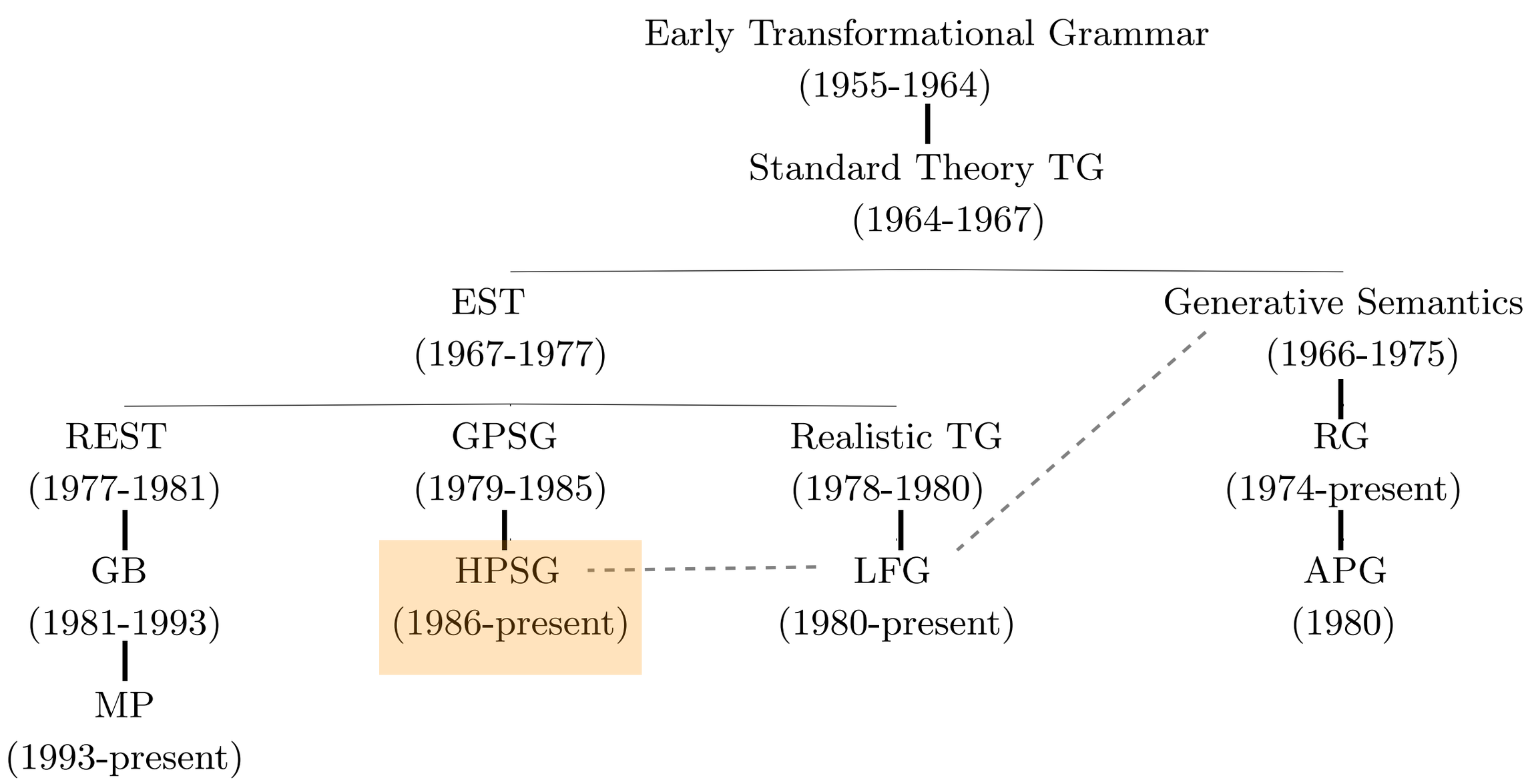
PD Dr.Valia Kordoni

Email: kordoni@coli.uni-sb.de

http://www.coli.uni-saarland.de/courses/FLST/2008/

Linguistic Foundations
Syntax

Family Tree of Syntactic Theories



Why Study Syntax?

- Why should linguists study syntax?
- Why should computational linguists study syntax?
- Should anyone else study syntax? Why?

Context-Free Grammar

- A quadruple: $< C, \Sigma, P, S >$
  - C: set of categories
  - $\sum$: set of terminals (vocabulary)
  - P: set of rewrite rules $\alpha \rightarrow \beta_1, \beta_2, \ldots, \beta_n$
  - S in C: start symbol
  - For each rule $a \rightarrow \beta_1, \beta_2, \ldots, \beta_n \in P$
    $a \in C;\ \beta_i \in C \cup \sigma;\ 1 \leq i \leq n$

A Toy Grammar

RULES

S ⟶ NP VP

NP ⟶ (D) A* N PP*

VP ⟶ V (NP) (PP)

PP ⟶ P NP

LEXICON

D: the, some

A: big, brown, old

N: birds, fleas, dog, hunter, I

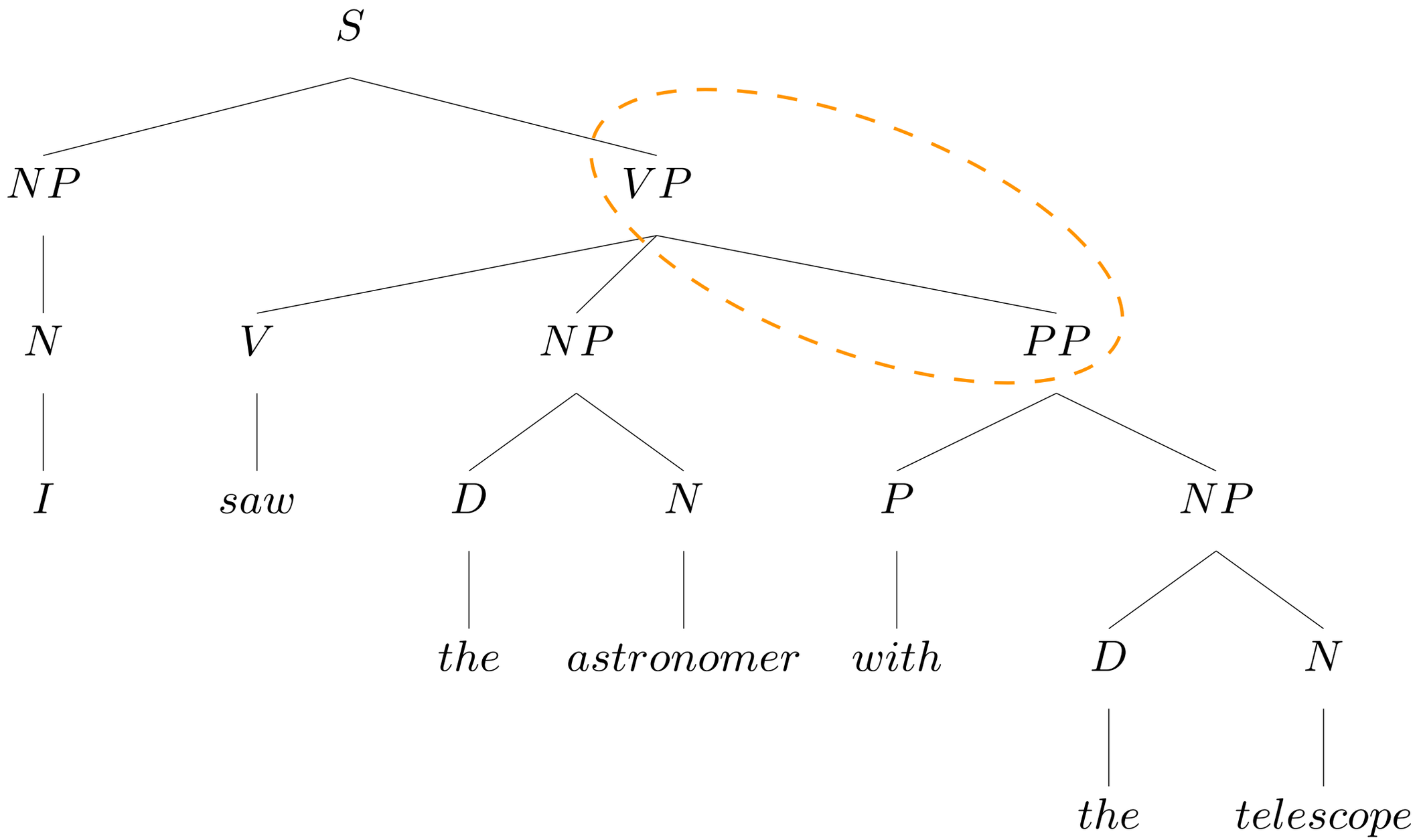
V: attack, ate, watched

P: for, beside, with

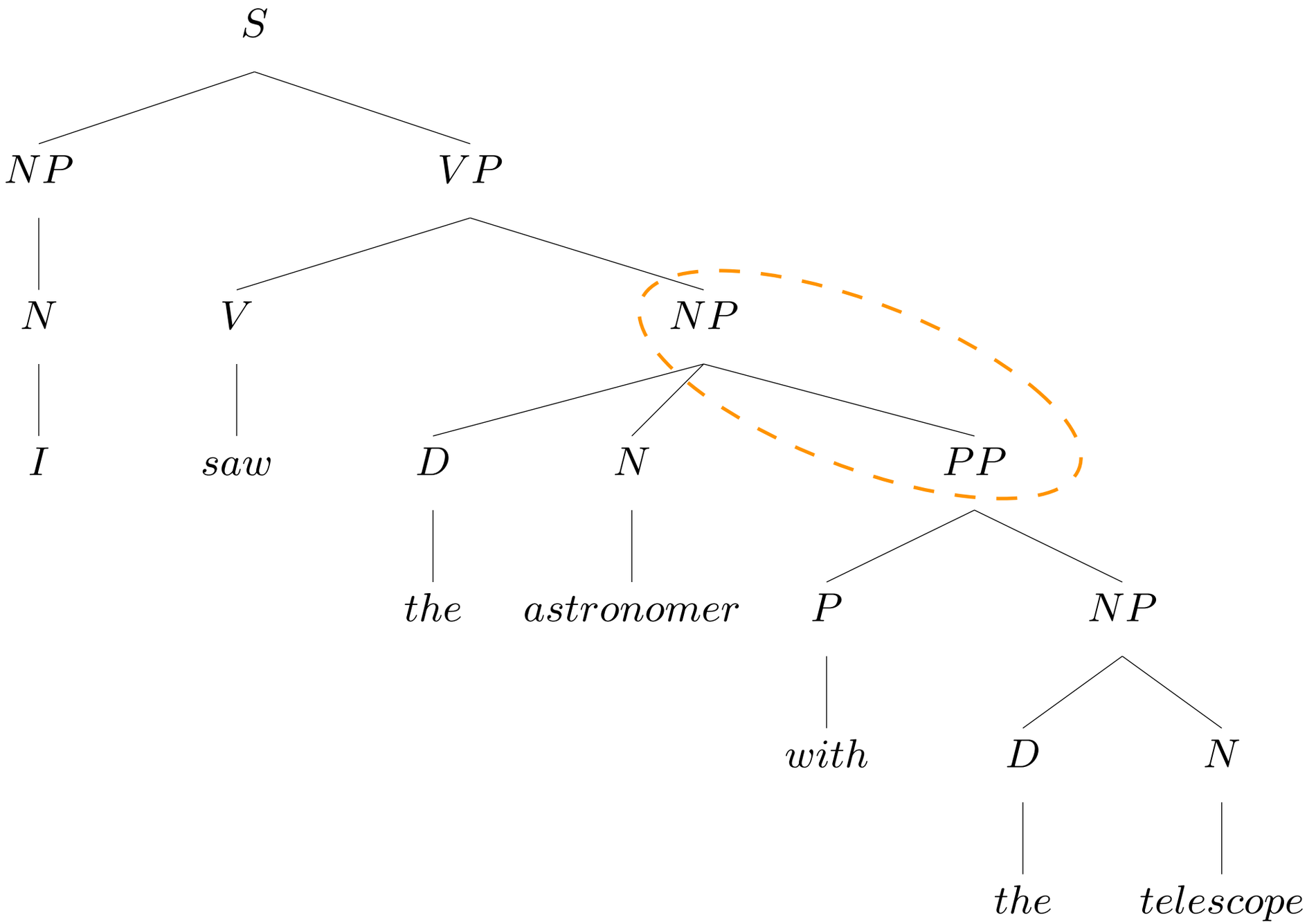
Structural Ambiguity

I saw the astronomer with the telescope.

Structure 1: PP under VP



Structure 1: PP under NP



Constituency Tests

- Recurrent Patterns

  The quick brown fox with the bushy tail jumped over the lazy brown dog with one ear.

- Coordination

  The quick brown fox with the bushy tail and the lazy brown dog with one ear are friends.

- Sentence-initial position

  The election of 2000, everyone will remember for a long time.

- Cleft sentences

  It was a book about syntax they were reading.

General Types of Constituency Tests

- Distributional
- Intonational
- Semantic
- Psycholinguistic

... but they don't always agree.

Central claims implicit in CFG formalism:

1. Parts of sentences (larger than single words) are linguistically significant units, i.e. phrases play a role in determining meaning, pronunciation, and/or the acceptability of sentences.

2. Phrases are contiguous portions of a sentence (no discontinuous constituents).

3. Two phrases are either disjoint or one fully contains the other (no partially overlapping constituents).

4. What a phrase can consist of depends only on what kind of a phrase it is (that is, the label on its top node), not on what appears around it.

- Claims 1-3 characterize what is called ‘phrase structure grammar’

- Claim 4 (that the internal structure of a phrase depends only on what type of phrase it is, not on where it appears) is what makes it ‘context-free’.

- There is another kind of phrase structure grammar called ‘context-sensitive grammar’ (CSG) that gives up 4. That is, it allows the applicability of a grammar rule to depend on what is in the neighboring environment. So rules can have the form $A \underset{\_}{\rightarrow} X$, in the context of Y_Z.

Possible Counterexamples

- To Claim 2 (no discontinuous constituents):

  A technician arrived who could solve the problem.

- To Claim 3 (no overlapping constituents):

  I read what was written about me.

- To Claim 4 (context independence):

  - *He arrives this morning.*
  - **He arrive this morning.*
  - **They arrives this morning.*
  - *They arrive this morning.*

A Trivial CFG

S $\longrightarrow$ NP VP

NP $\longrightarrow$ D N

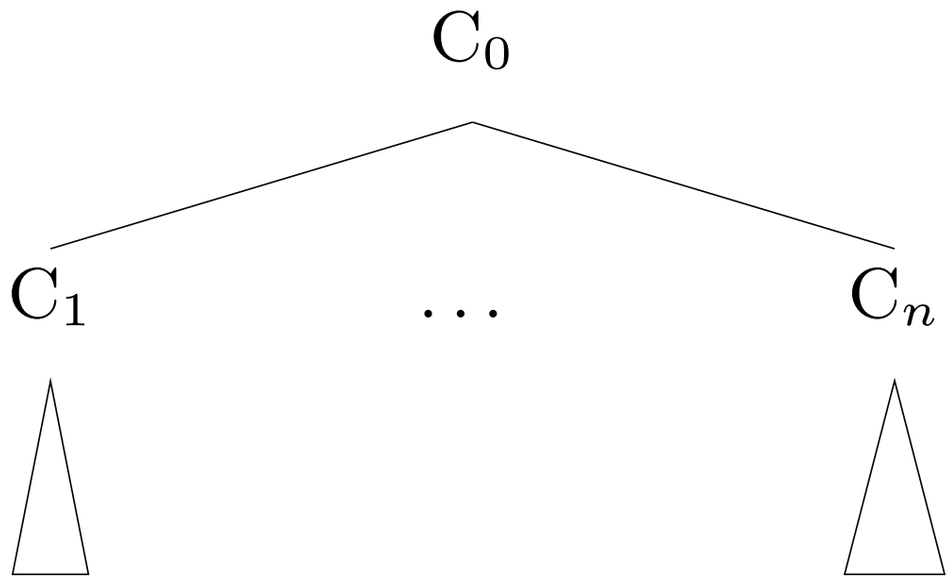
VP $\longrightarrow$ V NP

D: *the*

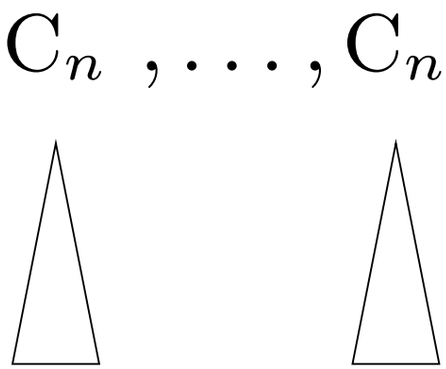
V: *chased*

N: *dog, cat*

Trees and Rules



is a well-formed nonlexical tree if (and only if)

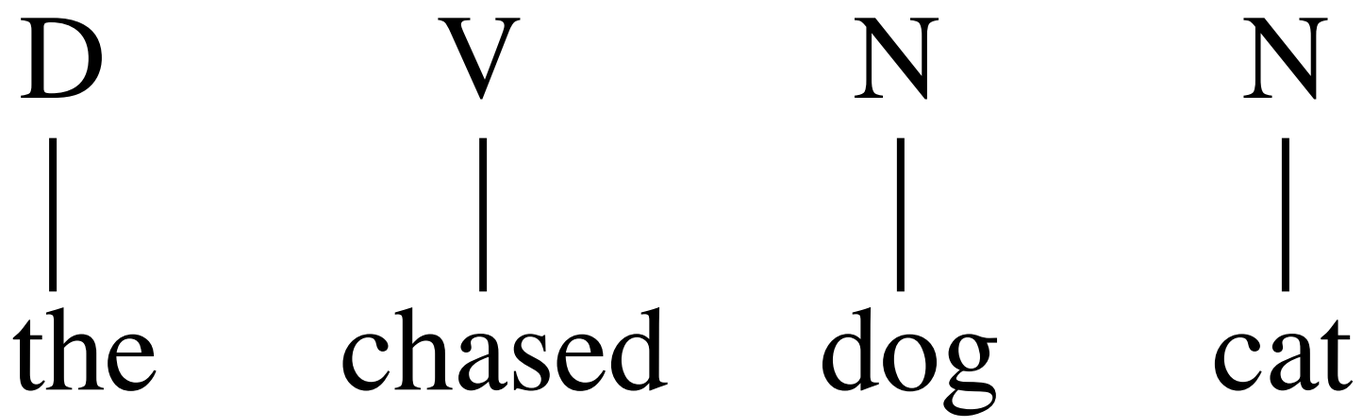


are well-formed trees, and

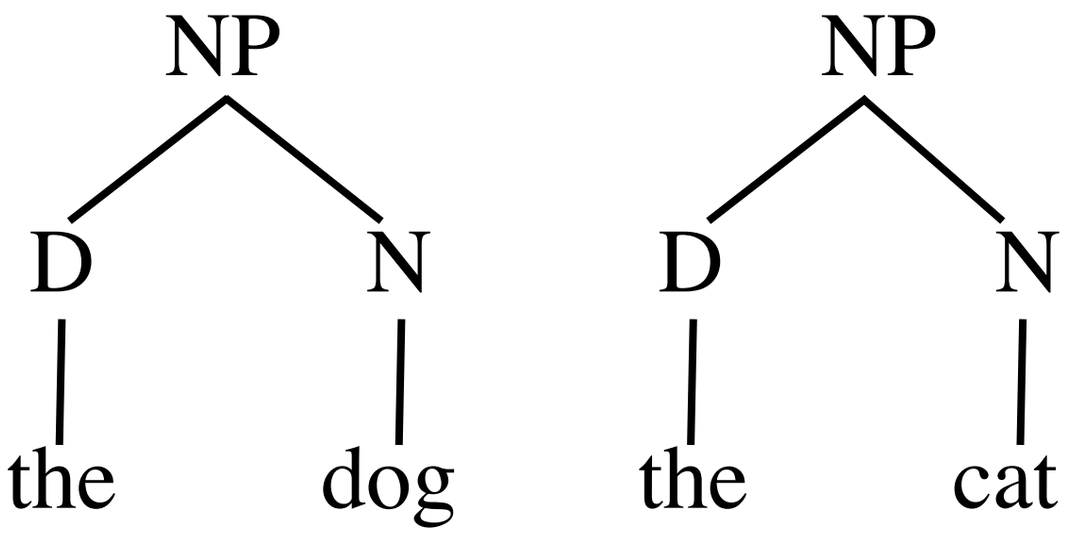
$C_0 \rightarrow C_1 \ldots Cn$ is a grammar rule.

Bottom-up Tree Construction

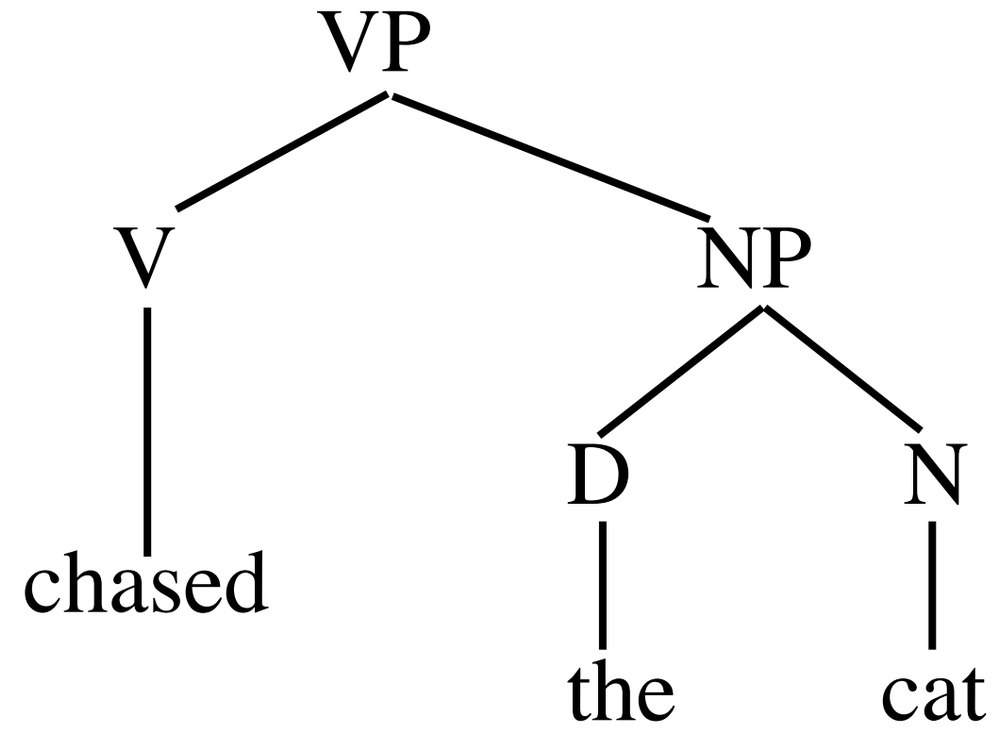
D: *the*
V: *chased*
N: *dog, cat*



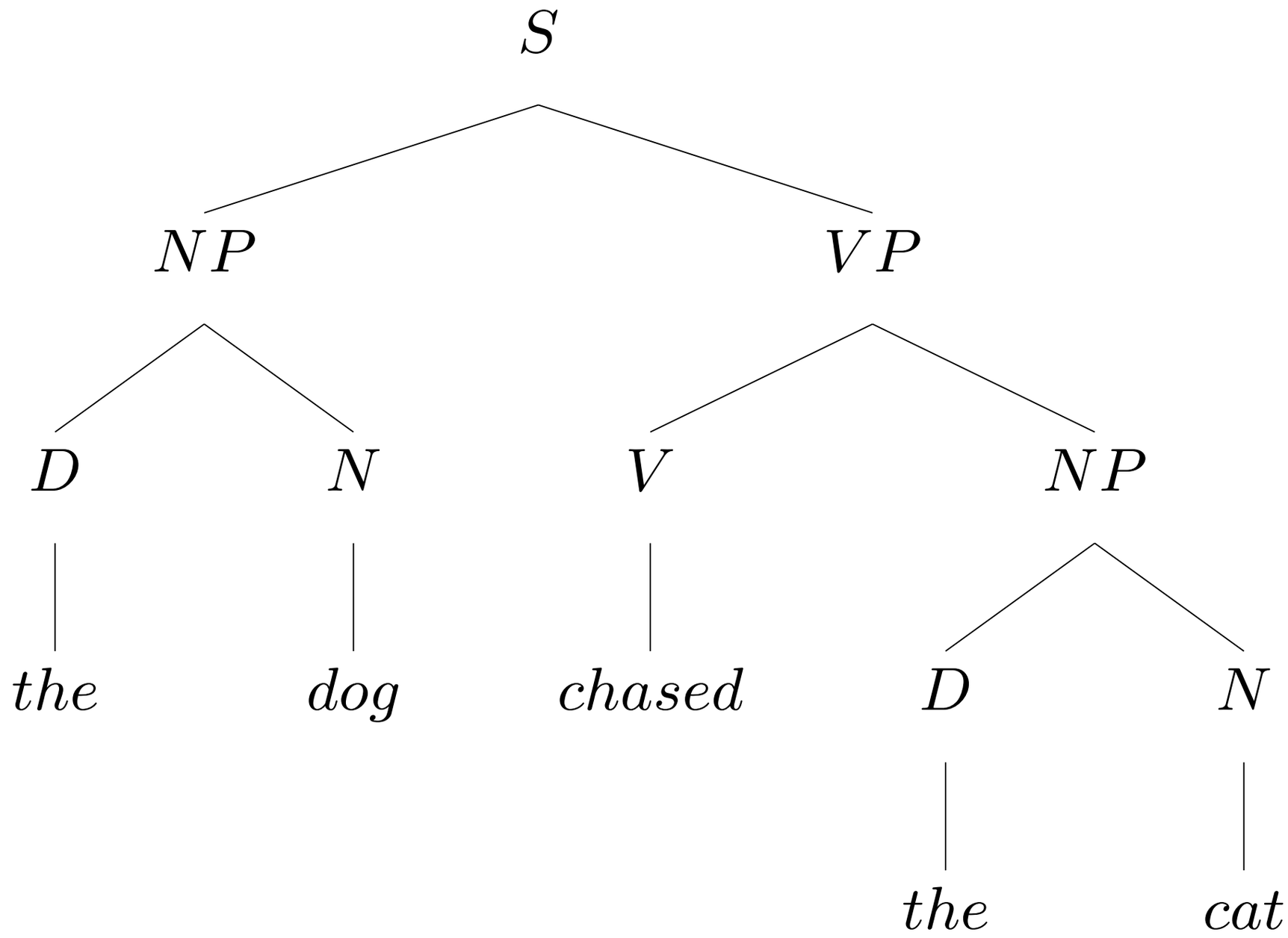
$$NP \longrightarrow D \;\; N$$



$$VP \longrightarrow V \;\; NP$$

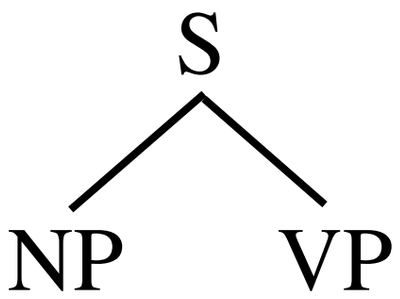


$$\mathrm{S} \longrightarrow \mathrm{NP}\ \ \mathrm{VP}$$

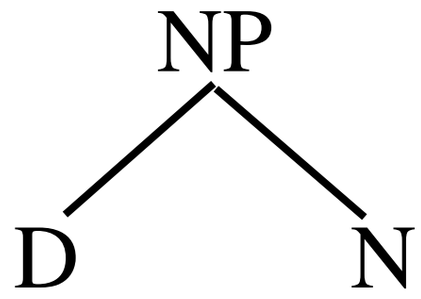


Top-down Tree Construction

$S \longrightarrow NP\ \ VP$

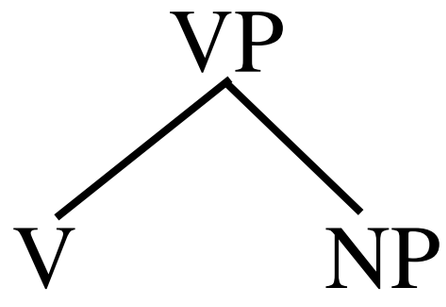


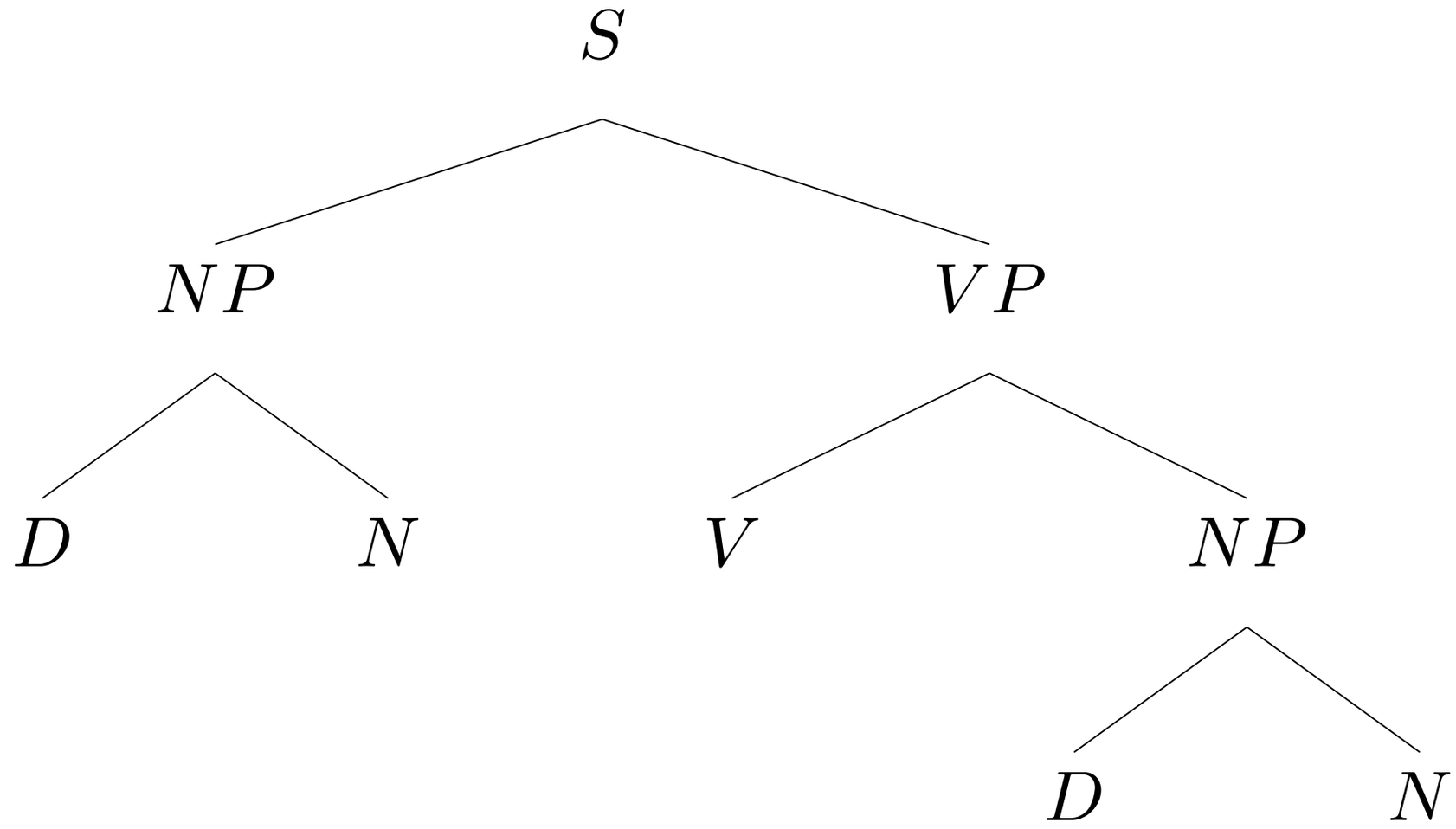
$NP \longrightarrow D\ \ N$

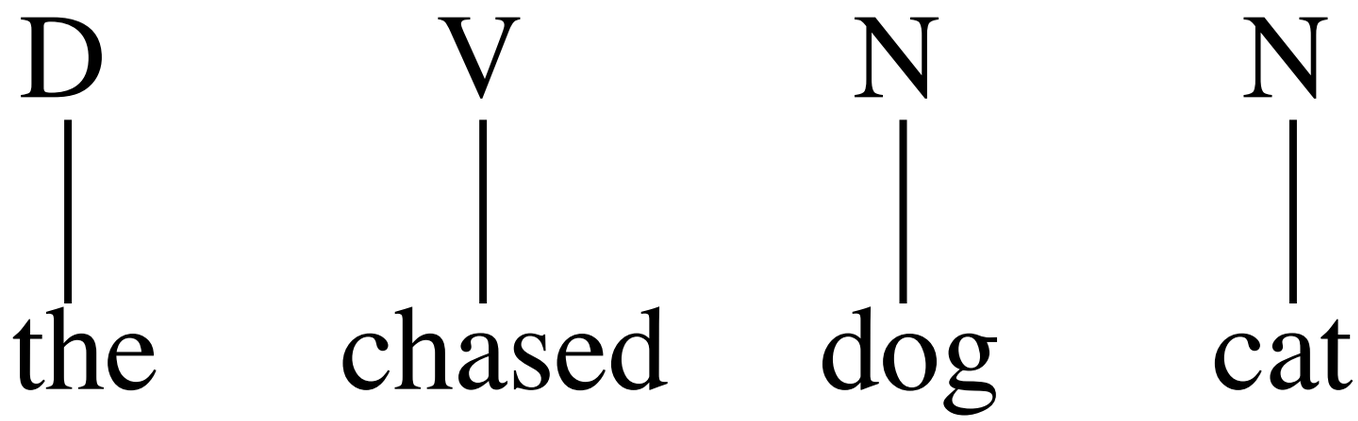


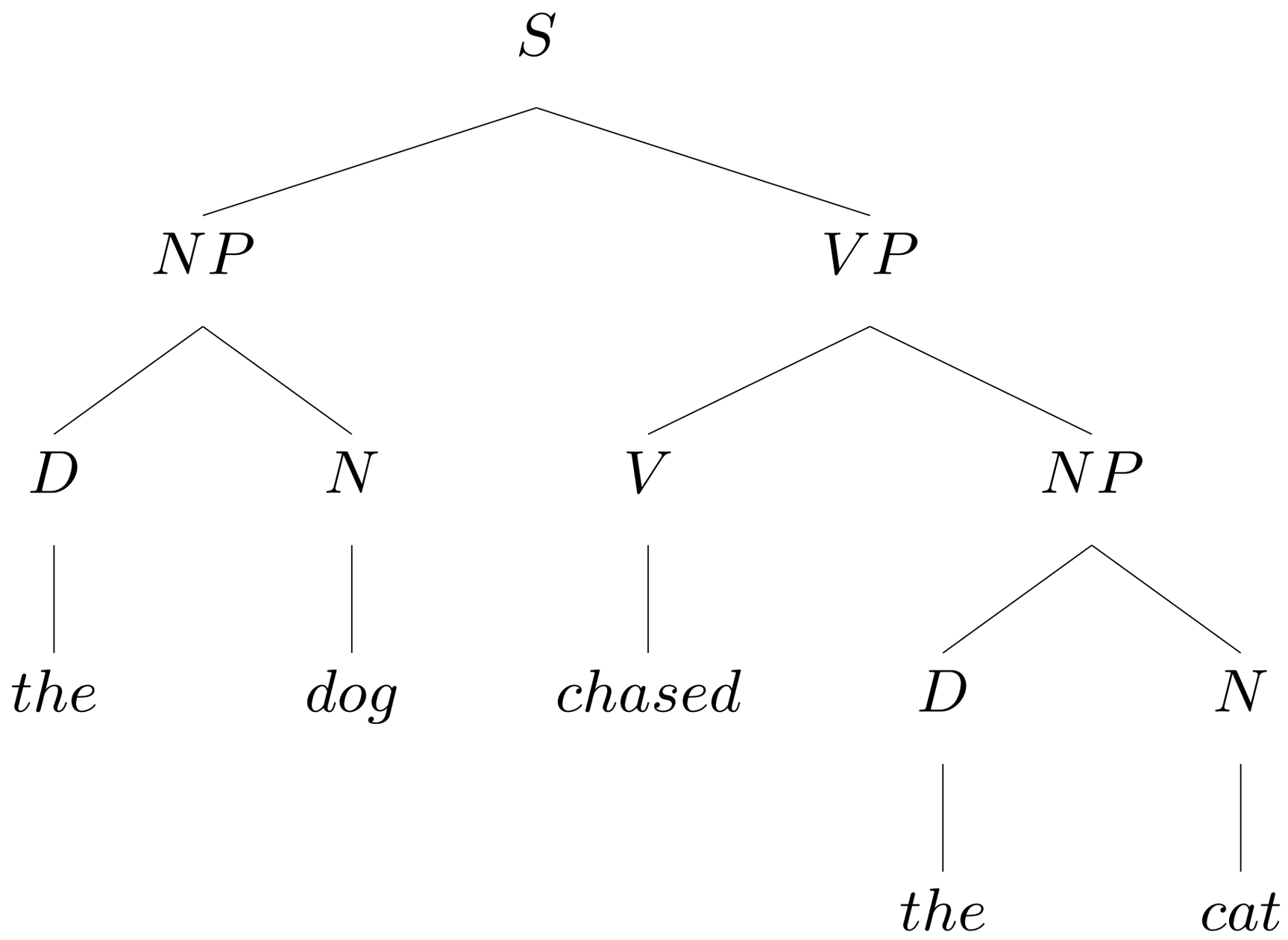
(twice)

$VP \longrightarrow V\ \ NP$




S
NP
VP
D
N
V
NP
D
N


D
V
N
N
the
chased
dog
cat


S
NP
VP
D
N
V
NP
the
dog
chased
D
N
the
cat

Bottom-up and top-down approaches are equivalent for CFG, but can differ for more complex types of grammars

Rules

S ⟶ A B

A ⟶ C D, in the environment __E.

B ⟶ E F, in the environment D__.

Lexicon

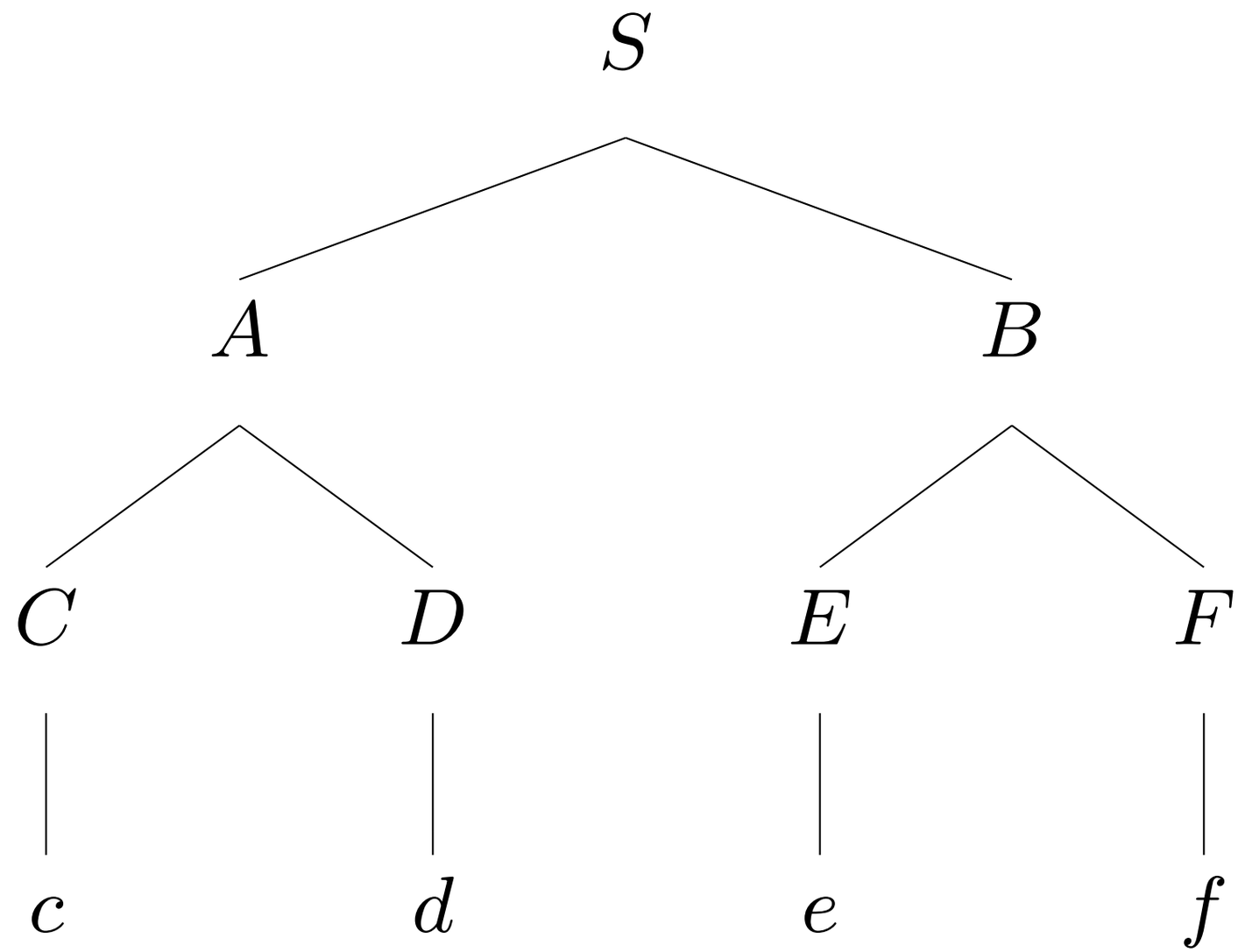
C: c

D: d

E: e

F: f

This tree is licensed bottom-up, but not top-down



Weaknesses of CFG

- It doesn't tell us what constitutes a linguistically natural rule

  VP → P NP

  NP → VP S

- Rules get very cumbersome once we try to deal with things like agreement and transitivity.

- It has been argued that certain languages (notably Swiss German and Bambara) contain constructions that are provably beyond the descriptive capacity of CFG.

On the other hand....

- It's a simple formalism that can generate infinite languages and assign linguistically plausible structures to them.
- Linguistic constructions that are beyond the descriptive power of CFG are rare.
- It's computationally tractable and techniques for processing CFGs are well understood.