

Ling 566  
Oct 1, 2013  
Context-Free Grammar

# Overview

- Failed attempts
- Formal definition of CFG
- Constituency, ambiguity, constituency tests
- Central claims of CFG
- Order independence
- Weaknesses of CFG
- Reading questions
- If time: Work through Chapter 2, Problem 1

# Insufficient Theory #1

- A grammar is simply a list of sentences.
- What's wrong with this?

# Insufficient Theory #2: FSMs

- the noisy dogs left

D A N V

- the noisy dogs chased the innocent cats

D A N V D A N

- $a^* = \{\emptyset, a, aa, aaa, aaaa, \dots\}$

- $a^+ = \{a, aa, aaa, aaaa, \dots\}$

- $(D) A^* N V ((D) A^* N)$

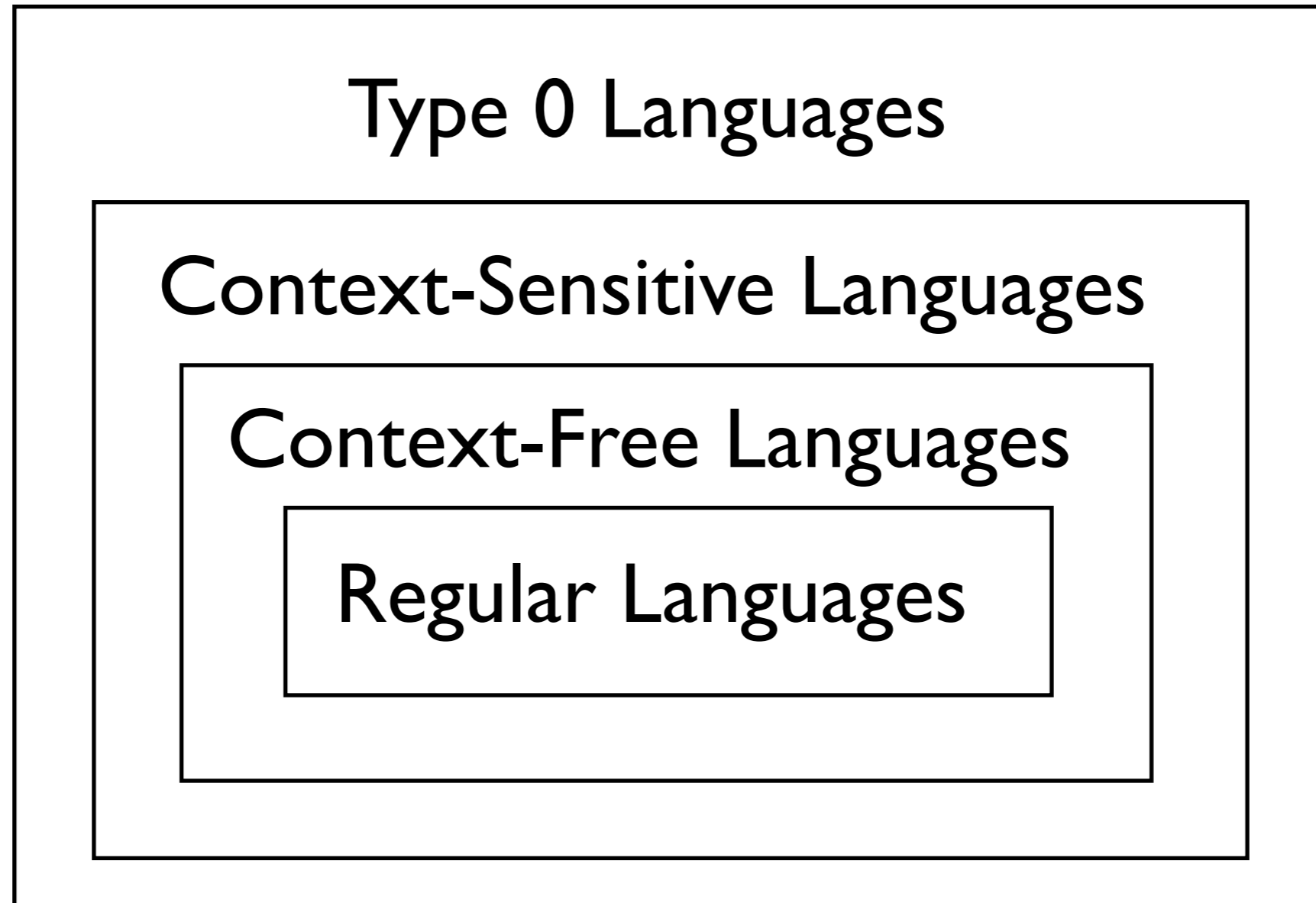
# What does a theory do?

- Monolingual
  - Model grammaticality/acceptability
  - Model relationships between sentences (internal structure)
- Multilingual
  - Model relationships between languages
  - Capture generalizations about possible languages

# Summary

- Grammars as lists of sentences:
  - Runs afoul of creativity of language
- Grammars as finite-state machines:
  - No representation of structural ambiguity
  - Misses generalizations about structure
  - (Not formally powerful enough)
- Next attempt: Context-free grammar (CFG)

# Chomsky Hierarchy



# Context-Free Grammar

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



# A Toy Grammar

## RULES

$S \longrightarrow NP VP$

$NP \longrightarrow (D) A^* N PP^*$

$VP \longrightarrow V (NP) (PP)$

$PP \longrightarrow P NP$

## LEXICON

D: the, some

A: big, brown, old

N: birds, fleas, dog, hunter, I

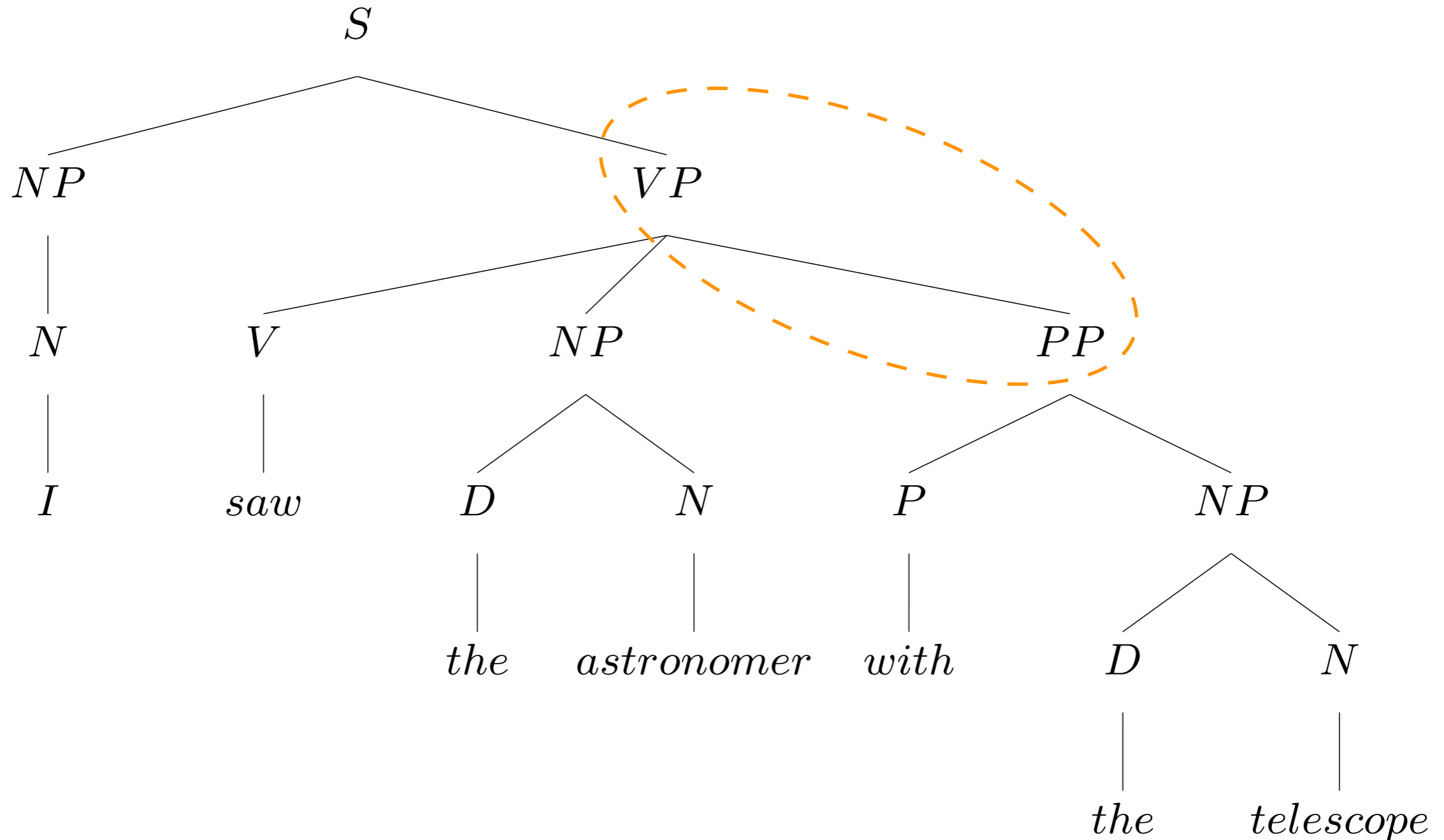
V: attack, ate, watched

P: for, beside, with

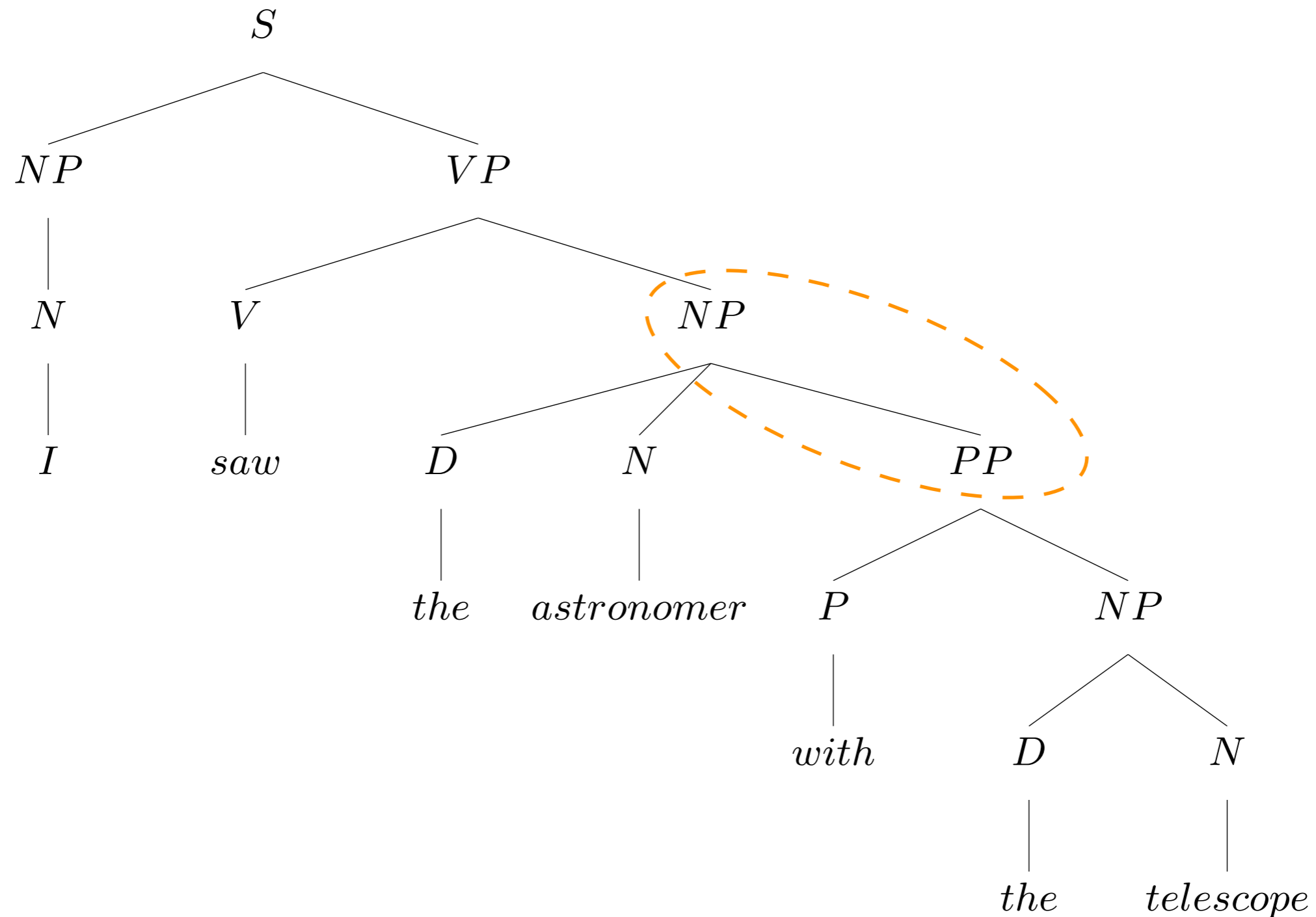
# Structural Ambiguity

I saw the astronomer with the telescope.

# Structure 1: PP under VP



# Structure 1: PP under NP



# Constituents

- How do constituents help us? (What's the point?)
- What aspect of the grammar determines which words will be modeled as a constituent?
- How do we tell which words to group together into a constituent?
- What does the model claim or predict by grouping words together into a constituent?

# 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  $A \rightarrow X$ , 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

$S \rightarrow NP VP$

$NP \rightarrow D N$

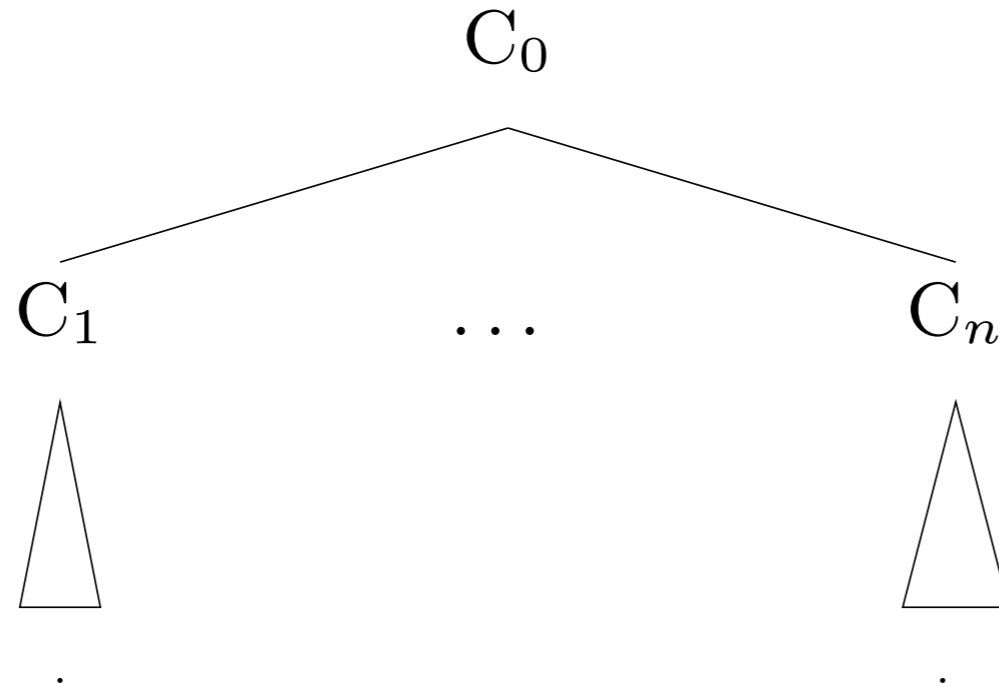
$VP \rightarrow V NP$

D: *the*

V: *chased*

N: *dog, cat*

# Trees and Rules



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

$C_1$  , ... ,  $C_n$  are well-formed trees, and



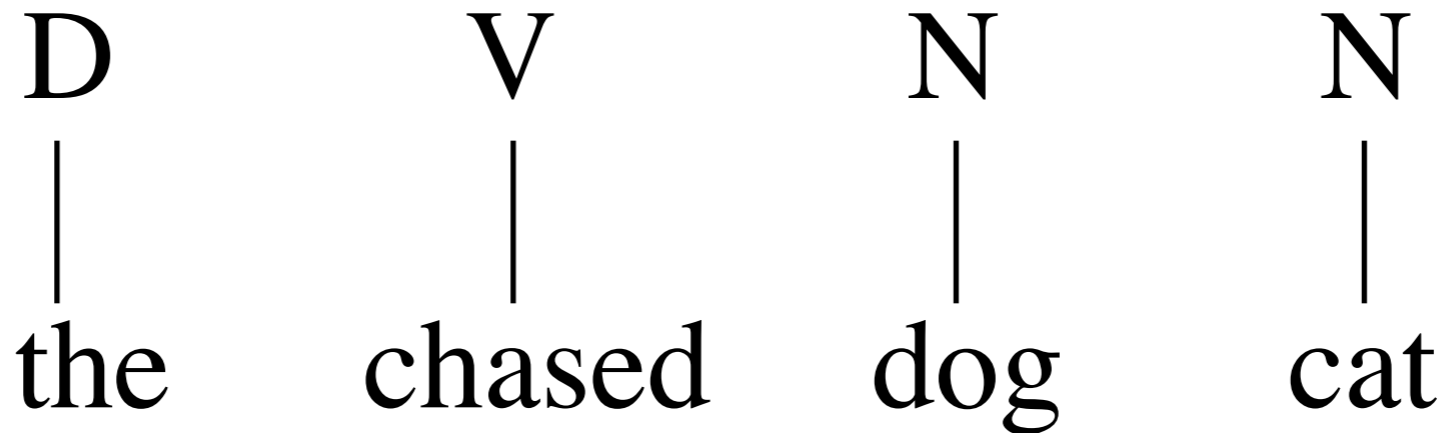
$C_0 \rightarrow C_1 \dots C_n$  is a grammar rule.

# Bottom-up Tree Construction

D: *the*

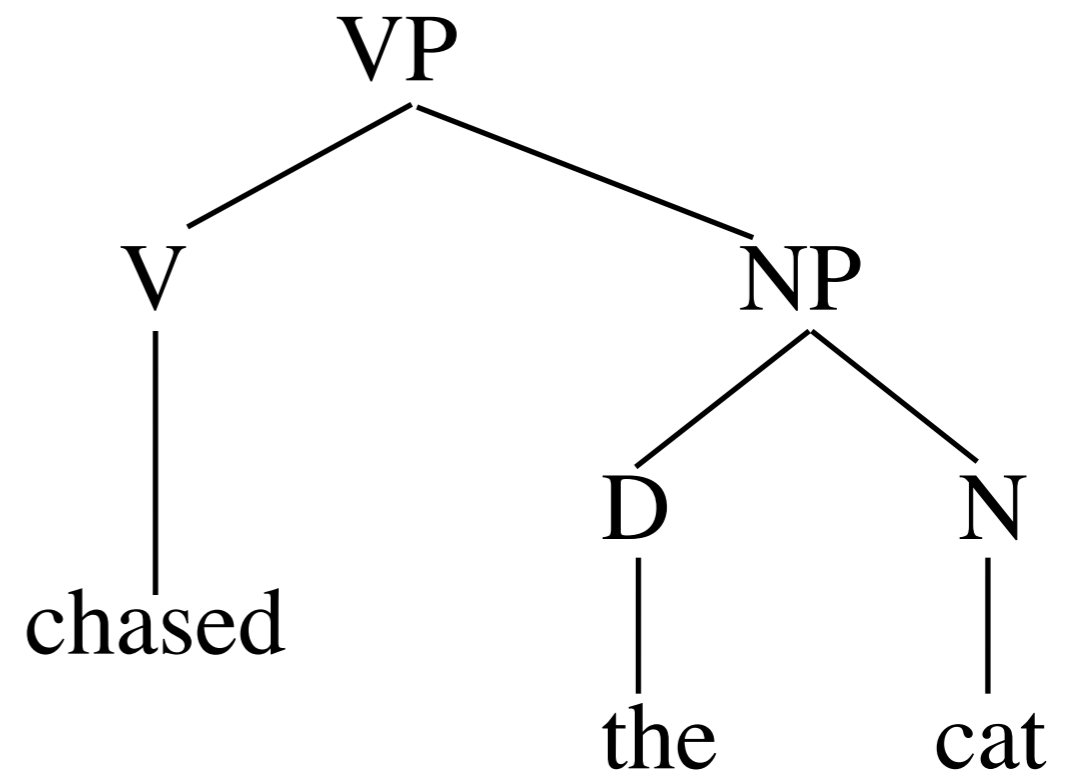
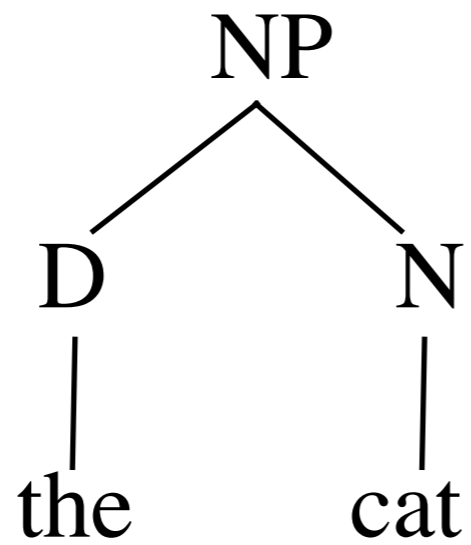
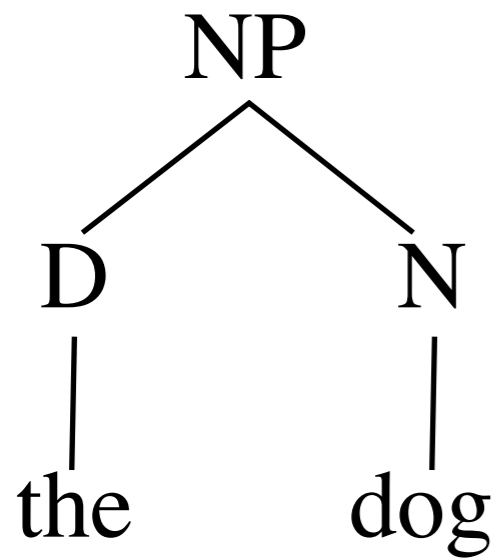
V: *chased*

N: *dog, cat*

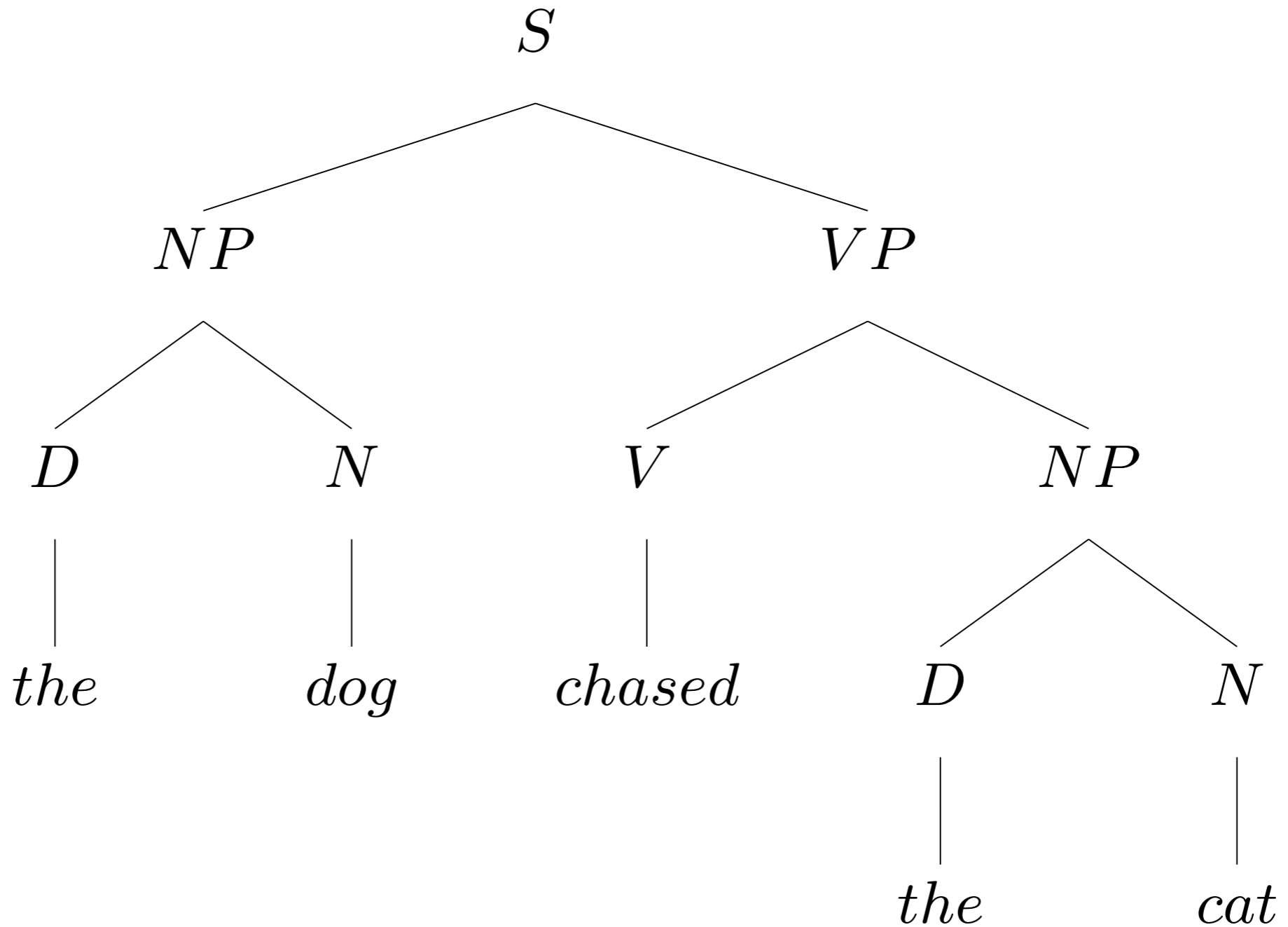


NP  $\rightarrow$  D N

VP  $\rightarrow$  V NP

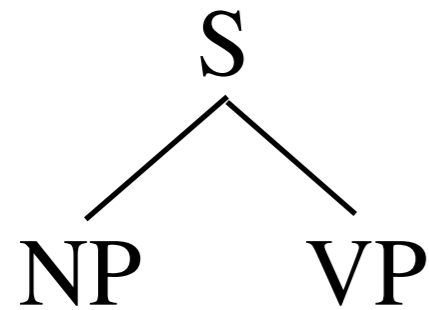


$S \rightarrow NP VP$

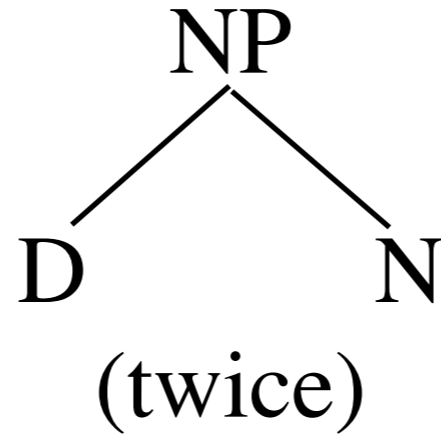


# Top-down Tree Construction

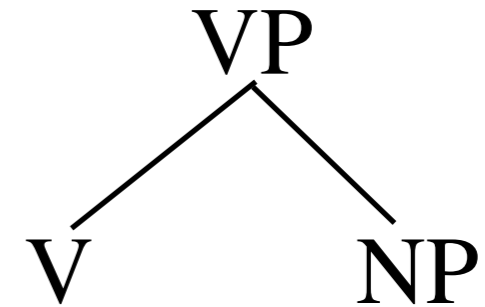
$S \longrightarrow NP VP$



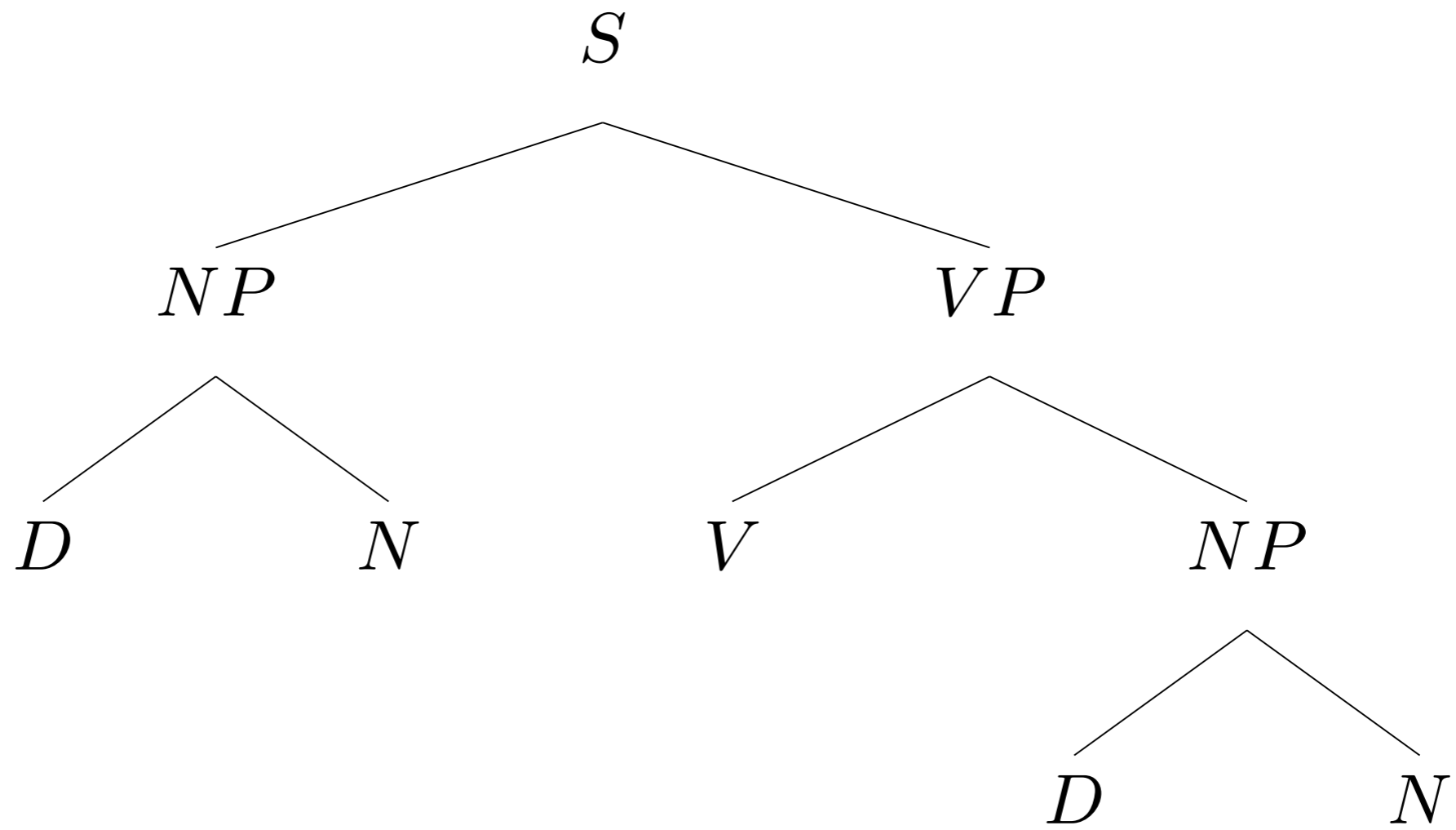
$NP \longrightarrow D N$



$VP \longrightarrow V NP$





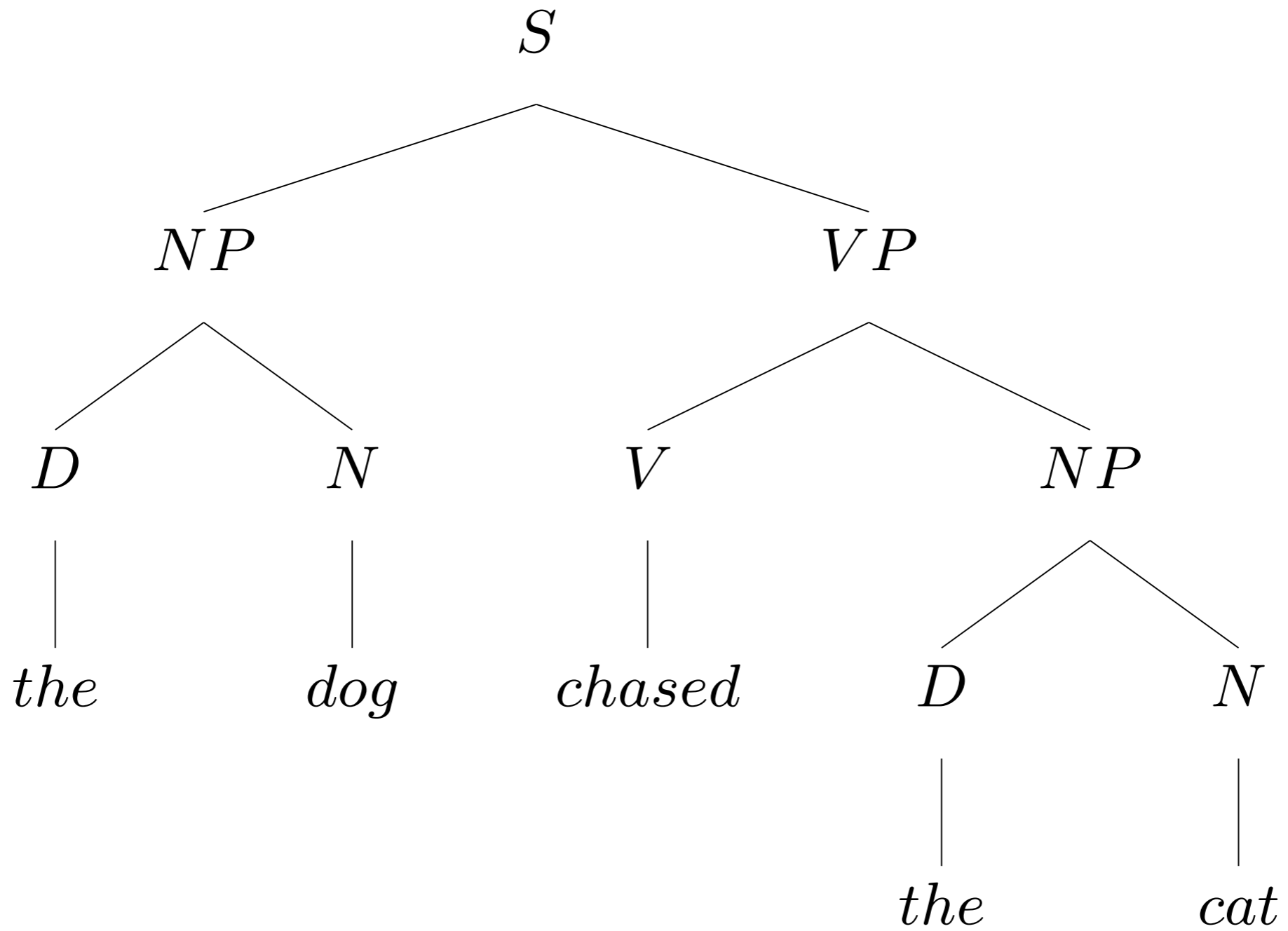


D  
|  
the

V  
|  
chased

N  
|  
dog

N  
|  
cat



# Weaknesses of CFG (atomic node labels)

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

$$VP \rightarrow P NP$$
$$NP \rightarrow VP S$$

- 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.

# So.....

- CFG has been the starting point for most types of generative grammar.
- The theory we develop in this course is an extension of CFG.

# Reading Questions

- What's up with NOM? How is it different from NP? Why do we need it? Why can't we simply branch NPs into two groups, one that includes determiners and one that doesn't?
- Can two different CFG trees have the same semantics?
- How are we going to do semantics without LF?

# Reading Questions

- Why is headedness a problem for CFGs?  
"the formalism of CFG, in and of itself, treats category names as arbitrary"
- Will the CFG grammar we are developing apply to all languages? What about languages like Dyirbal that have such free word order that it is difficult to posit VP and NP clauses?



# Reading Questions

- Is it better to test a grammar by sampling the sentences it generates and gauging their acceptability, or by running it against test suites already annotated for acceptability?
- Does the "context" in context-free mean the dialogue context? How can we talk about the meaning of sentences without context?

# Reading Questions

- Is natural language really not context-free?

# Shieber 1985

- Swiss German example:

... mer d'chind                      em Hans   es huus                      lönd hälfe aastriiche

... we   the children-ACC   Hans-DAT   the hous-ACC   let   help   paint

... we let the children help Hans paint the house

- Cross-serial dependency:

- *let* governs case on *children*

- *help* governs case on *Hans*

- *paint* governs case on *house*

# Shieber 1985

- Define a new language  $f(\text{SG})$ :

$$\begin{array}{llll} f(\text{d'chind}) & = & a & f(\text{Jan säit das mer}) & = & w \\ f(\text{em Hans}) & = & b & f(\text{es huus}) & = & x \\ f(\text{lönde}) & = & c & f(\text{aastriiche}) & = & y \\ f(\text{hälfe}) & = & d & f([\text{other}]) & = & z \end{array}$$

- Let  $r$  be the regular language  $wa^*b^*xc^*d^*y$
- $f(\text{SG}) \cap r = wa^mb^nc^md^ny$
- $wa^mb^nc^md^ny$  is not context free.
- But context free languages are closed under intersection.
- $\therefore f(\text{SG})$  (and by extension Swiss German) must not be context free.

# Strongly/weakly CF

- A language is *weakly* context-free if the set of strings in the language can be generated by a CFG.
- A language is *strongly* context-free if the CFG furthermore assigns the correct structures to the strings.
- Shieber's argument is that SW is not *weakly* context-free and *a fortiori* not *strongly* context-free.
- Bresnan et al (1983) had already argued that Dutch is *strongly* not context-free, but the argument was dependent on linguistic analyses.

# Overview

- Failed attempts
- Formal definition of CFG
- Constituency, ambiguity, constituency tests
- Central claims of CFG
- Order independence
- Weaknesses of CFG
- Next time: Feature structures