

A decorative border in the top-left corner of the slide, composed of a grid of small orange squares and rectangles, creating a geometric pattern.

Ling 566

Oct 1, 2024

Context-Free Grammar

Overview

- Two insufficient theories
- Formal definition of CFG
- Constituency, ambiguity, constituency tests
- Central claims of CFG
- Weaknesses of CFG
- Reading questions

Insufficient Theory #1

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

An example hypothetical language (p.22)

- Some sentences go on and on
- *Sentences some go on and on
- *Some sentences go on and on and on
- Sentences some go on and on and on
- Some sentences go on and on and on and on
- *Sentences some go on and on and on and on
on

Insufficient Theory #2: FSMs

- the noisy dogs left

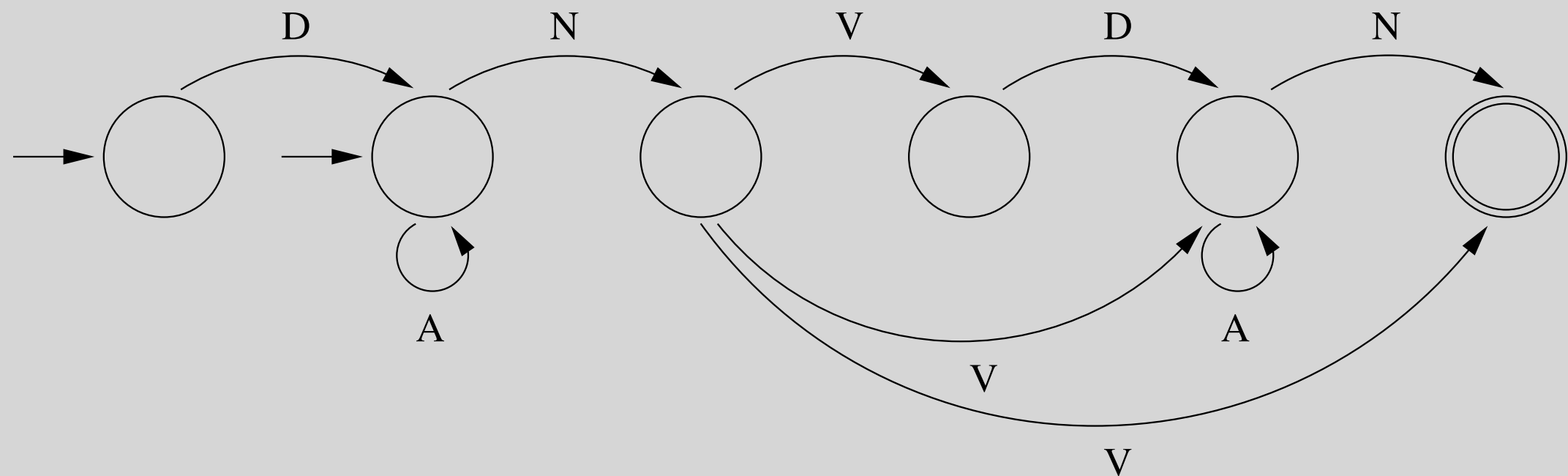
D A N V

- the noisy dogs chased the innocent cats

D A N V D A N

- $a^* = \{\emptyset, a, aa, aaa, aaaa, \dots\}$
- $a^+ = \{a, aa, aaa, aaaa, \dots\}$
- $(D) A^* N V ((D) A^* N)$

A Finite State Machine



Insufficient Theory #2: FSMs

- What's wrong with this?
- What can't it model?

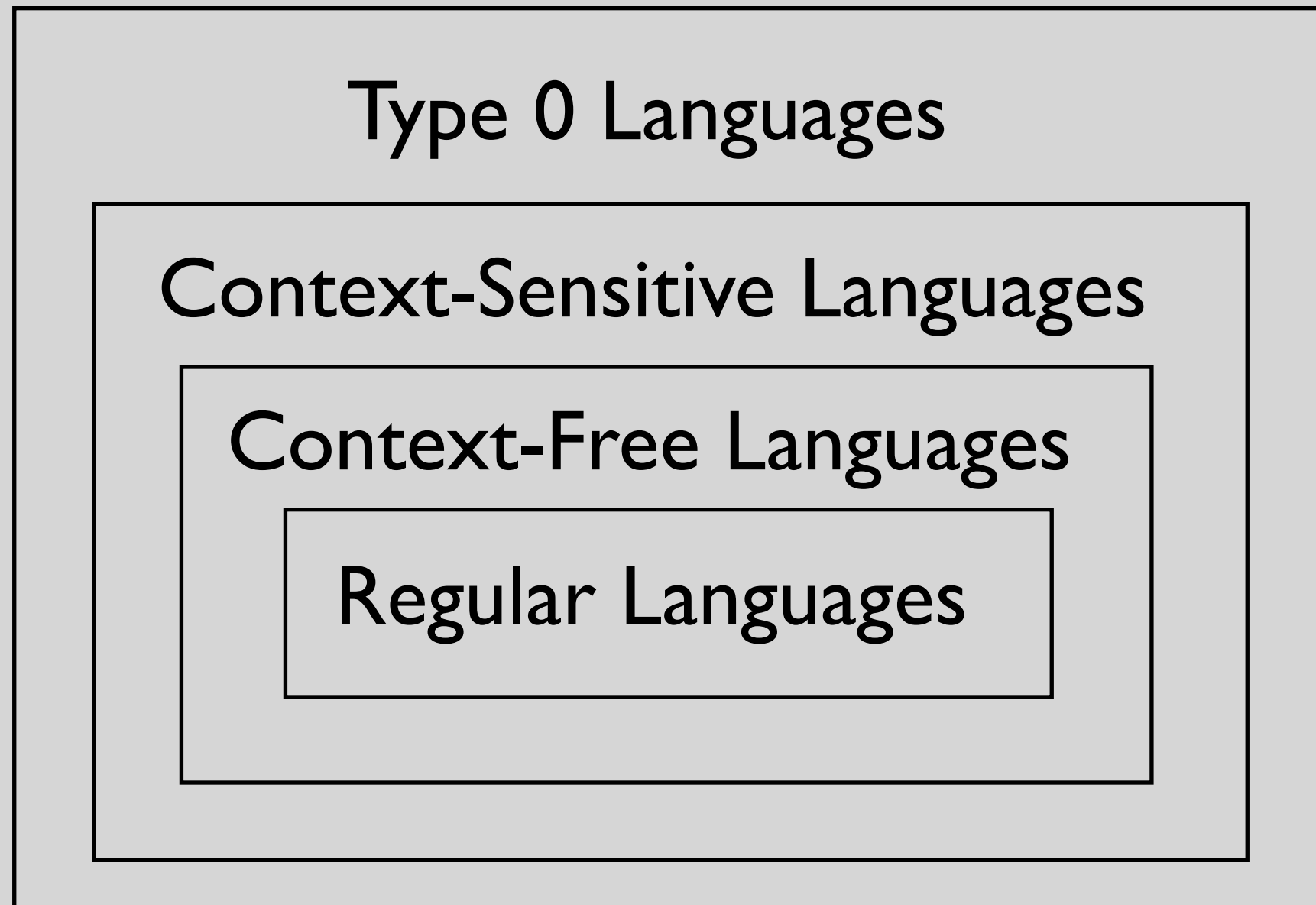
What does a theory do?

- 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

Chomsky Hierarchy



Context-Free Grammar

- A quadruple: $\langle C, \Sigma, P, S \rangle$
 - C : set of categories
 - Σ : 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

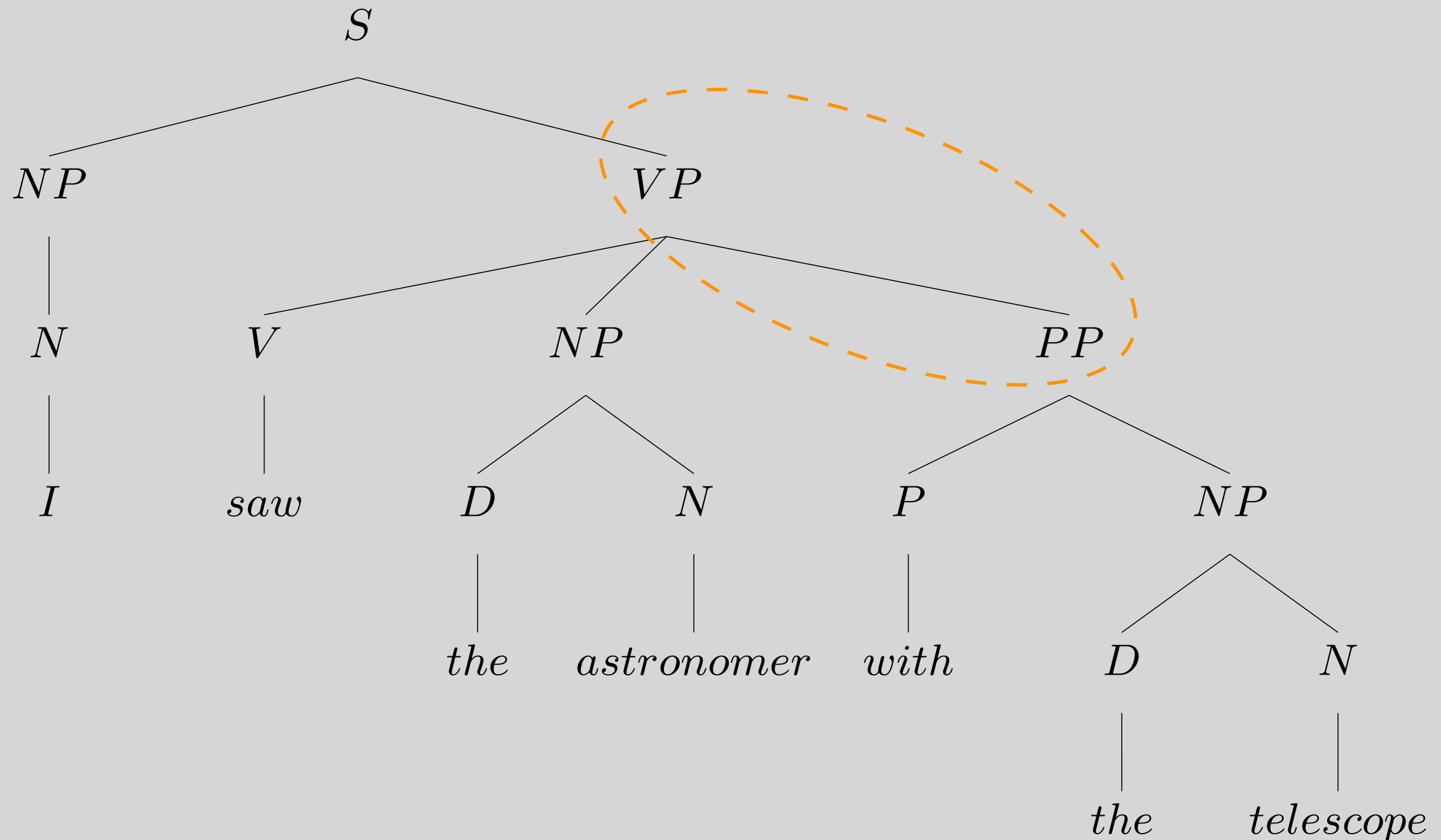
V: attack, ate, watched

P: for, beside, with

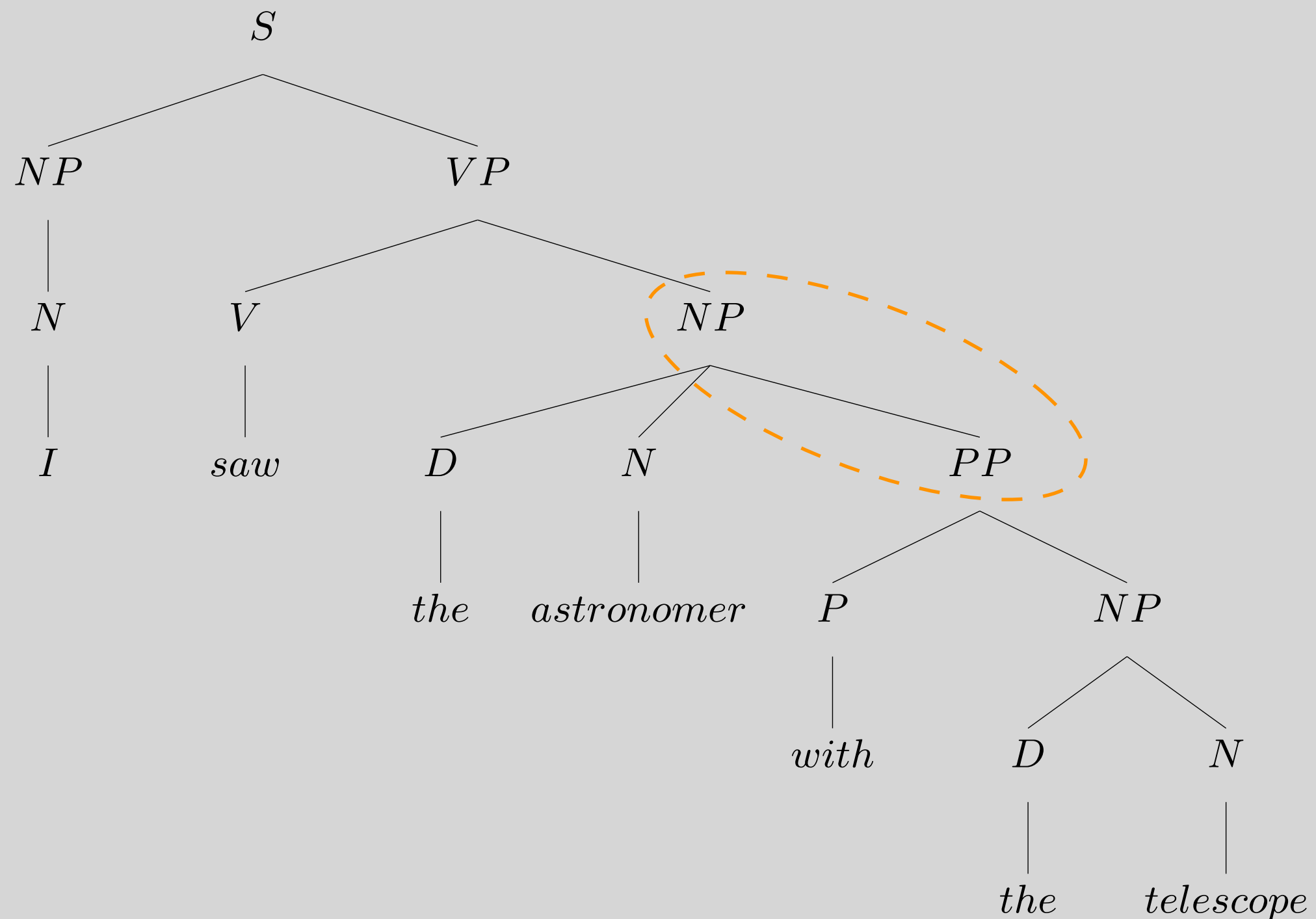
Structural Ambiguity

I saw the astronomer with the telescope.

Structure 1: PP under VP



Structure 2: 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

$S \rightarrow NP \ VP$

$NP \rightarrow D \ N$

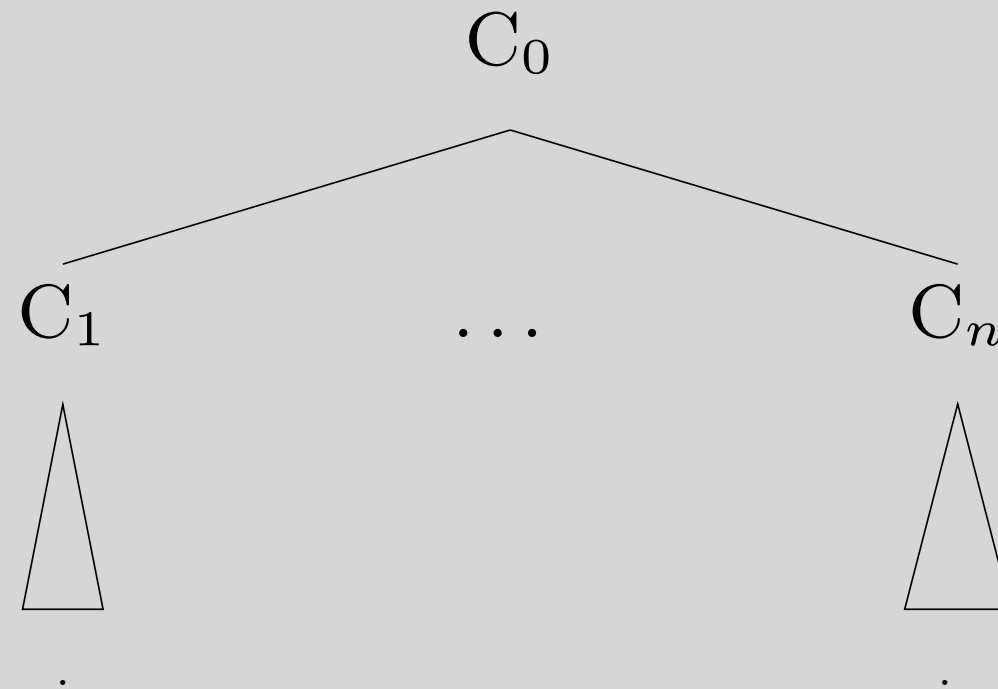
$VP \rightarrow V \ NP$

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

Two identical triangles, one under C_1 and one under C_n .

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

Bottom-up Tree Construction

D: *the*

V: *chased*

N: *dog, cat*

D

|

the

V

|

chased

N

|

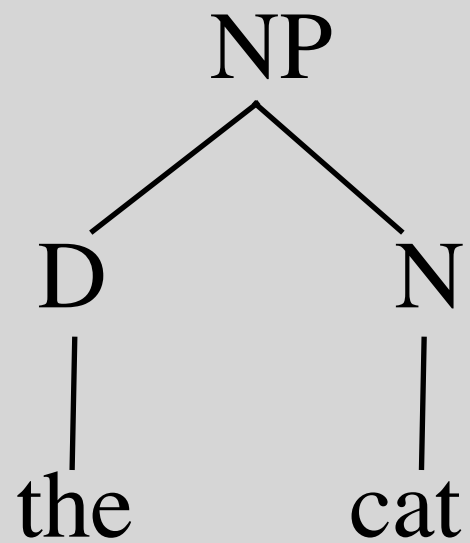
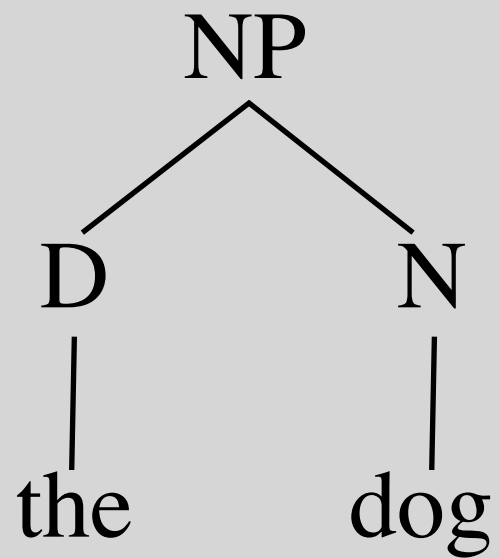
dog

N

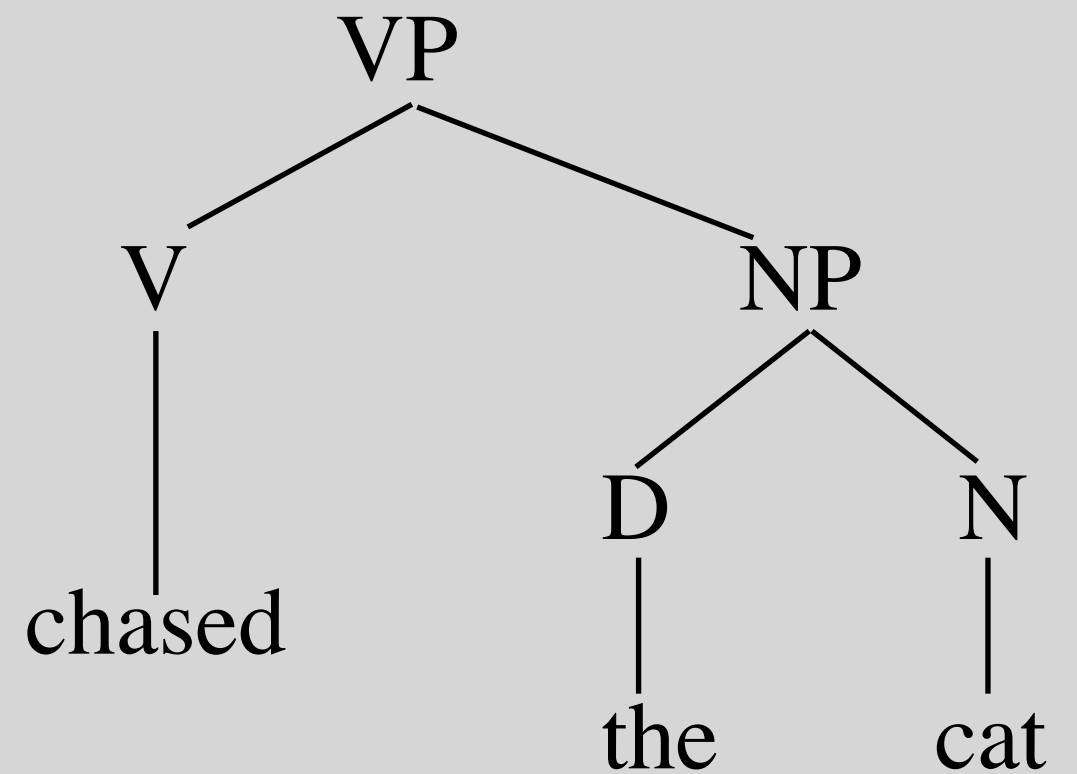
|

cat

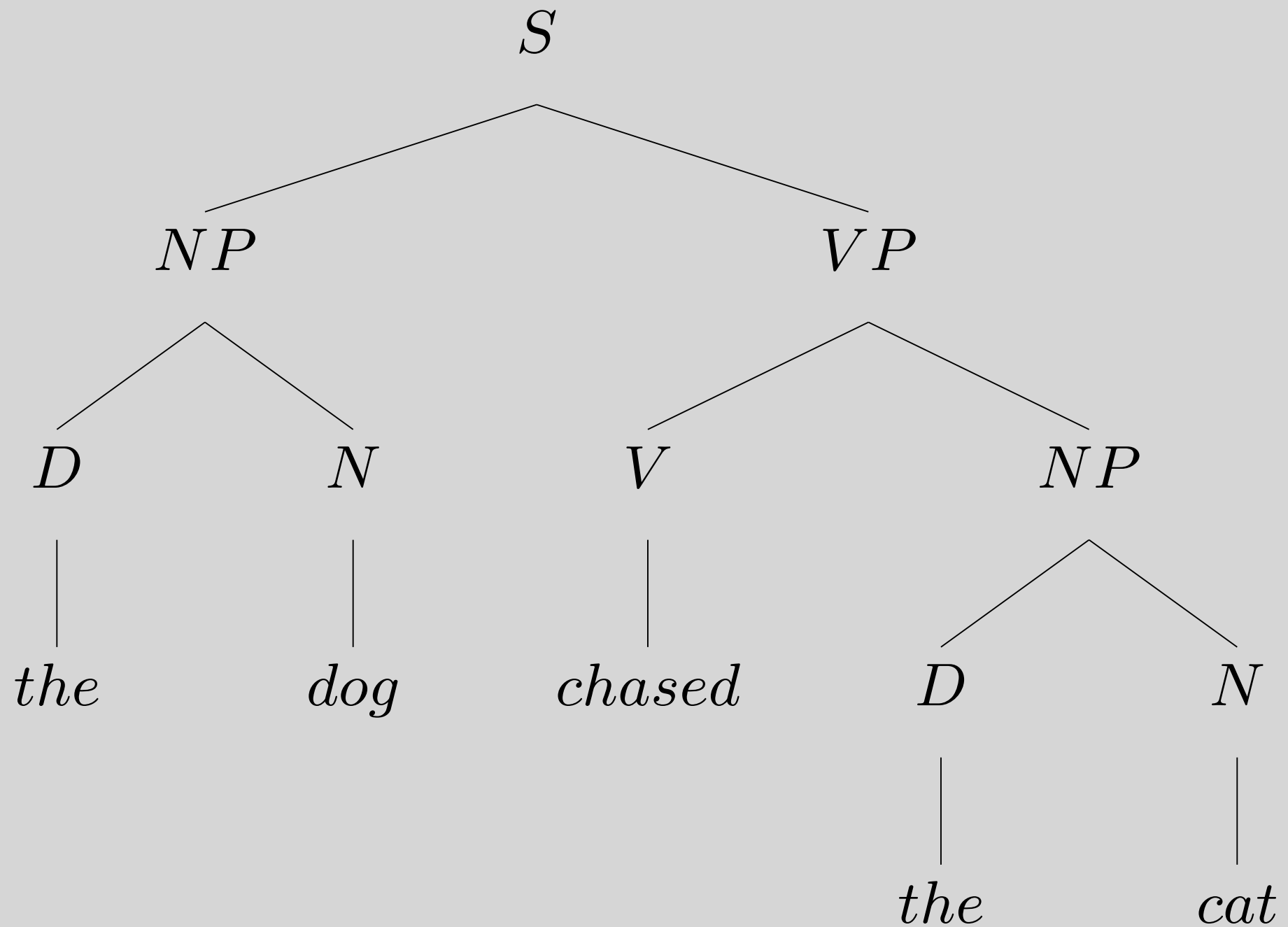
$NP \longrightarrow D \ N$



$VP \longrightarrow V \ NP$

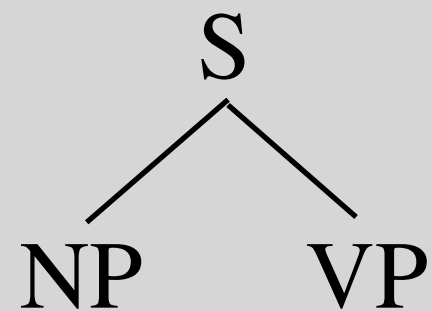


$S \rightarrow NP VP$

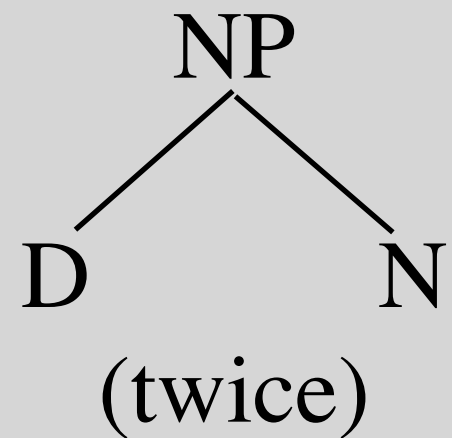


Top-down Tree Construction

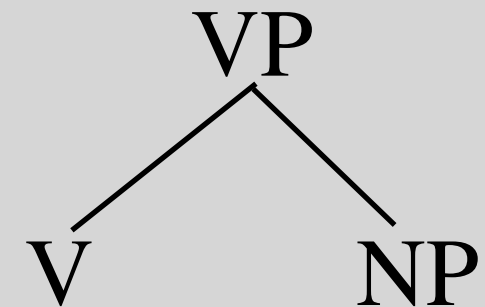
$S \longrightarrow NP \ VP$

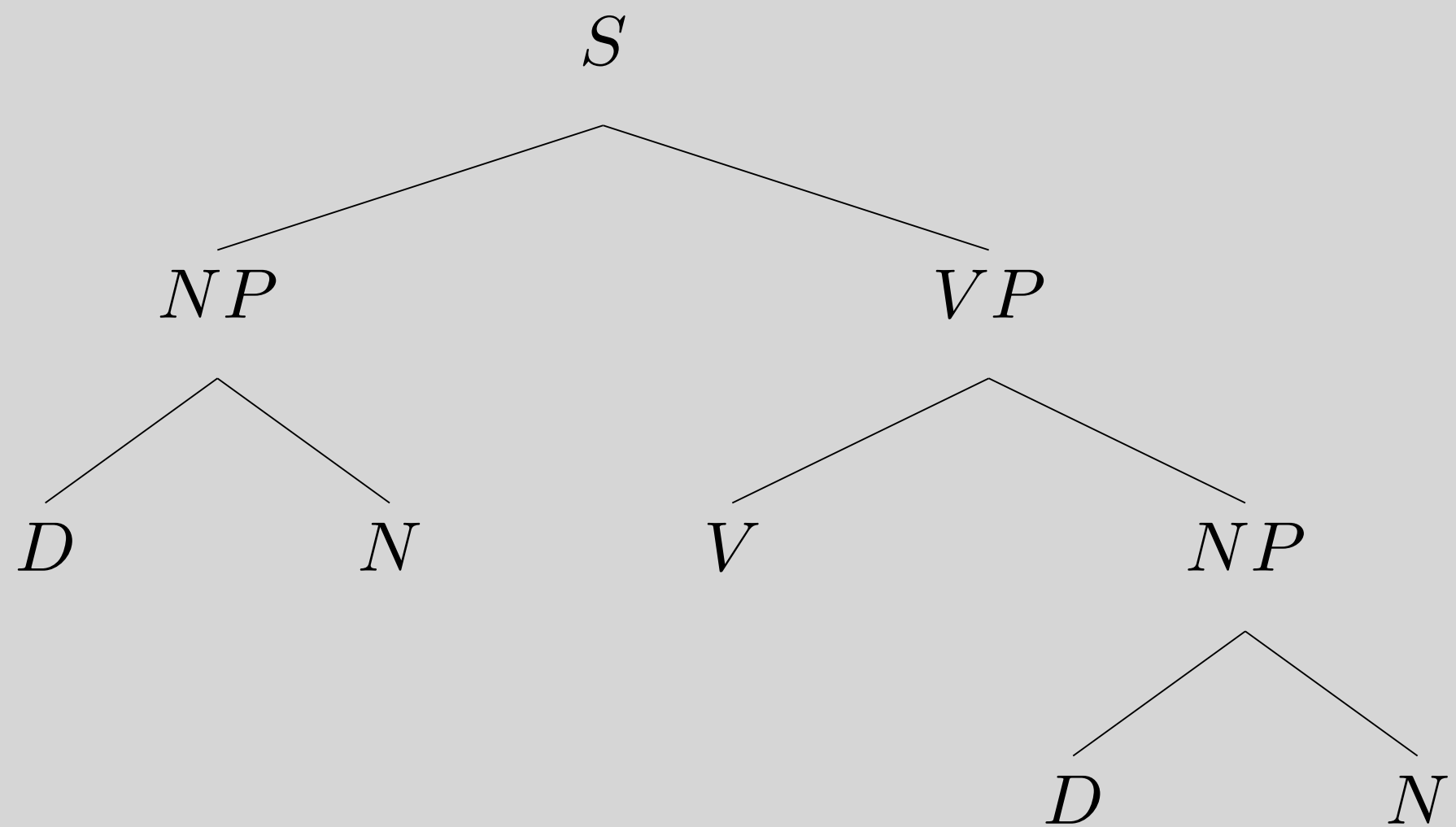


$NP \longrightarrow D \ N$



$VP \longrightarrow V \ NP$



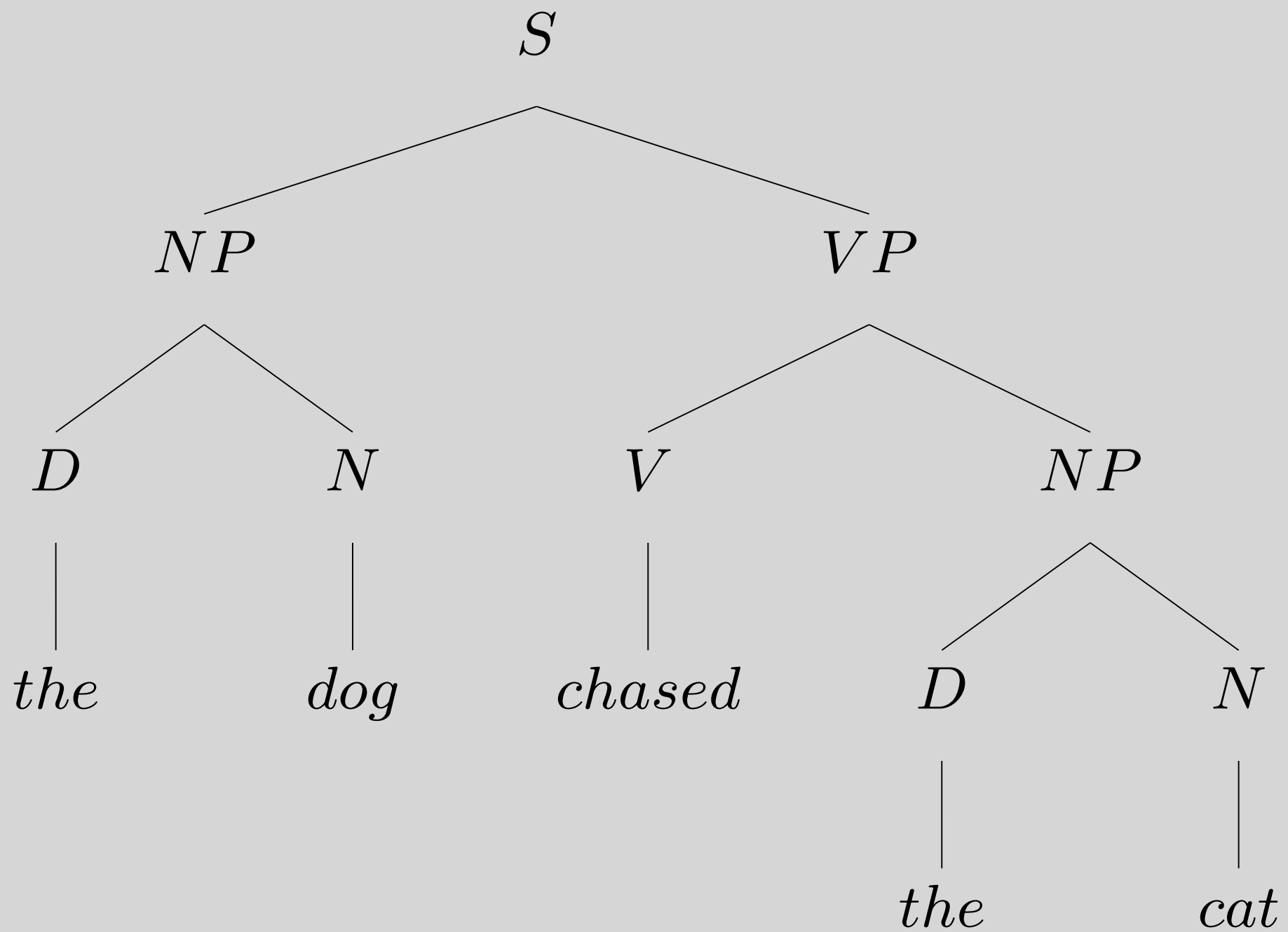


D
|
the

V
|
chased

N
|
dog

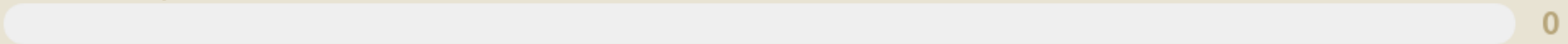
N
|
cat



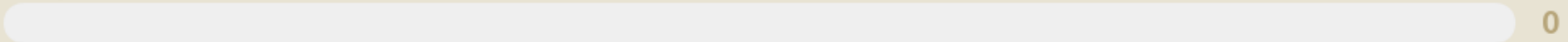
For parsing, which style feels most intuitive to you?



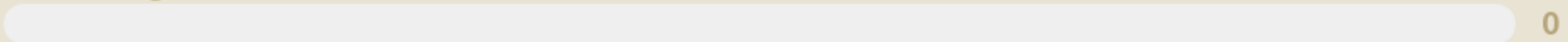
Bottom up



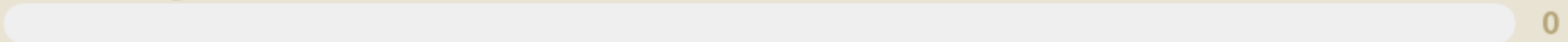
Top down



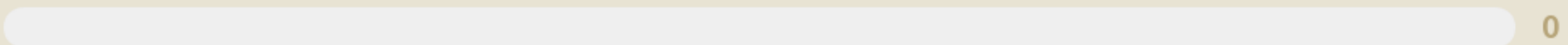
Left to right



Something else



None of the above



Weaknesses of CFG (w/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.

Agreement & Transitivity

S	→	NP-SG VP-SG	VP-SG	→	IV-SG
S	→	NP-PL VP-PL	VP-PL	→	IV-PL
NP-SG	→	(D) NOM-SG	VP-SG	→	TV-SG NP
NP-PL	→	(D) NOM-PL	VP-PL	→	TV-PL NP
NOM-SG	→	NOM-SG PP	VP-SG	→	DTV-SG NP NP
NOM-PL	→	NOM-PL PP	VP-PL	→	DTV-PL NP NP
NOM-SG	→	N-SG	VP-SG	→	CCV-SG S
NOM-PL	→	N-PL	VP-PL	→	CCV-PL S
NP	→	NP-SG	VP-SG	→	VP-SG PP
NP	→	NP-PL	VP-PL	→	VP-PL PP
	...			...	

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 $f(\text{SG})$:

$f(\text{d'chind})$	$=$	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

- 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. w/reg languages
- $\therefore f(\text{SG})$ (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 SG is not weakly context-free and a fortiori not strongly context-free.

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.

Overview

- Two insufficient theories
- Formal definition of CFG
- Constituency, ambiguity, constituency tests
- Central claims of CFG
- Weaknesses of CFG
- Reading questions

RQs: Regex

- ARTICLE NOUN VERB (ARTICLE NOUN)|ADJECTIVE. Why is ADJECTIVE not also in parentheses?

RQs: Parts of speech

- What would a distributional definition of verbs look like? Can't be as simple as "follows determiners" like for nouns.

RQs: Constituency

- Will there be more accurate and consistent heuristics for diagnosing constituents presented in the book later on? I think it frustrated me a bit that you can't strictly rely on coordinate conjunction or the environment diagnostics to decide if something is a constituent or not, and I was wondering if there is a more concrete definition/diagnostics for identifying a constituent in English, or is the idea of a constituent itself somehow flawed and that is why we can't define a more exhaustive universal rule for identifying a constituent?

RQs: Coordination

- In the coordination subsection in 2.4.1, we posit the rules $NP \rightarrow NP + \text{CONJ } NP$ and $S \rightarrow S + \text{CONJ } S$ and from just these two rules, we collapse into the generalized rule $X \rightarrow X + \text{CONJ } X$. The text goes on to say this general rule can be used (but not infallibly) as a constituency test. What are the guidelines to follow when proposing a generalized rule like this? Are two identified rules always sufficient? How would an exception to the generalized rule change how we approach this?

RQs: NOM

- Is NOM the same as N'?
- Why isn't there something corresponding to V'?

RQs: NOM

- Why can't we just collapse the rules containing NOM in (23) to $NP \rightarrow (D) N (PP+)$?
- (See poll next slide)



Which rules (w/PP->P NP and coord rule) license "no painting by A or drawing by B was displayed"?



NP -> NP PP

0

NP -> D NOM; NOM -> NOM PP; NOM -> N

0

NP -> (D) N (PP)

0

None of the above

0

RQs: Verb subcategories

- I am curious about intransitive, transitive, and ditransitive verbs. In order to have a functioning grammar and lexicon, does one require every verb to be catalogued and categorized in a corpus by whether they are IV, TV, or DTV? What about verbs that can be in multiple categories depending on how they are used?

RQs: Transformational Grammar

- Are the types of transformations finite, or is there potential for an infinite variety? If the list of transformation types grows, could that introduce new complexities? Would transformation types vary between languages, and how do we map out the interaction of rules, such as when one transformation (like passivization) requires another (such as agreement)?

RQs: Transformational Grammar

- Transformative & generative grammars: The book mentions "generative grammars" multiple times (for example, in section 2.8). An ideal grammar would be able to both parse and generate grammatical sentences, so why do we call it "generative"? Also, is transformative grammar connected to Chomsky's idea of "deep structure"? Would transformative grammars still be useful in describing languages with little inflection, such as Chinese?

RQs: Jurisdiction of syntax

- Chapter 2 mentions "Of course, there may be other interacting factors that cause grammatical sentences to sound less than fully acceptable": this makes me ask the question of whether and how you know an issue you are dealing with is a syntactic one and not something else, like a semantic or pragmatic issue.

RQs: Data handling

- How much should a linguist rely on their own intuition in research? Can they safely claim they are giving examples of English, or is it better to state that they are studying the version of English (or whatever native/near-native language of concern) from their I-language (or the version of I-language that is mutually comprehensible with language of other speakers conventionally named English as a whole)? Or should we always try to approximate the "E-language" by consulting several specialists, hopefully from much different backgrounds (albeit still belonging to the English-speaking environment), and honestly list the degrees of acceptance (see below) on each example for each participant (if any)?

RQs: Psycholinguistics

- To better understand the history of CFGs in grammar theory, was the main idea that humans also keep a hierarchical (maybe tree like) structure of view of received language? Is the tree like structure more of loose inspiration of mental activity (like neural nets are only loosely based on biological neurons) or was it supposed to be a strong assertion?

RQs: Psycholinguistics

- Section 2.9 brings up that the object of our study is not purely a matter of individual psychology, nor is it exclusively a social phenomenon. I am curious about what exactly are the objects of our syntactic theory trying to model? If they are neither mental constructs in our minds nor a simple set of conventions, what exactly are the objects under investigation?