JMORF — Morpho-Syntax

First attempts at a theory of grammar

Francis Bond
Palacký University
`https://fcbond.github.io/`
`bond@ieee.org`

Lecture 1b
Location: SV 2.39

Overview

- Two Syntactic Theories that won't work
- Context Free Grammars
- Central claims of CFG

What makes a good model?

- **generative**: license all grammatical sentences and only them
  $\Rightarrow$ **precise**

- **explanatory**: can explain generalizations
  - *the cat chased the rat* $\sim$ *the rat was chased by the cat* (semantics)
  - phrases tend to act like one member of the phrase (headedness)
  - new information tends to come first/last (information theory)

- **concise**: the model is as simple as possible (elegant)
  $\Rightarrow$ **universal** (minimal stipulations)

- **tractable**: the model can be modeled computationally

Our models are normally imperfect:
we aim for iteratively improved approximations

Insufficient Theory #1

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

Insufficient Theory #2: Regular Expressions

(1) *the noisy dogs left*
D A N V

(2) *the noisy dogs chased the innocent cats*
D A N V D A N

➢ (D) A* N V ((D) A* N)

Regular expressions: a formal language for matching things.

Symbol	Matches
.	any single character
$*$	the preceding element zero or more times.
?	the preceding element zero or one time: OR just () = ()?.
$+$	the preceding element one or more times.
$\vert$	either the expression before or after the operator.

Context-Free Grammar

- A quadruple: $\langle C, V, P, S \rangle$

  C set of categories $(\alpha, \beta, \ldots)$
  V set of terminals (vocabulary)
  P set of rewrite rules $\alpha \rightarrow \beta_1, \beta_2, \ldots, \beta_n$
  S the start symbol $\mathbf{S} \in C$

- For each rule $\alpha \rightarrow \beta_1, \beta_2, \ldots, \beta_n \in P$
  - $\alpha \in C$
  - $\beta_i \in C \cup V; 1 \leq i \leq n$

A Toy Grammar

➢ RULES

S	$\rightarrow$	NP VP
NP	$\rightarrow$	(D) A* N PP*
VP	$\rightarrow$	V (NP) (PP)
PP	$\rightarrow$	P NP

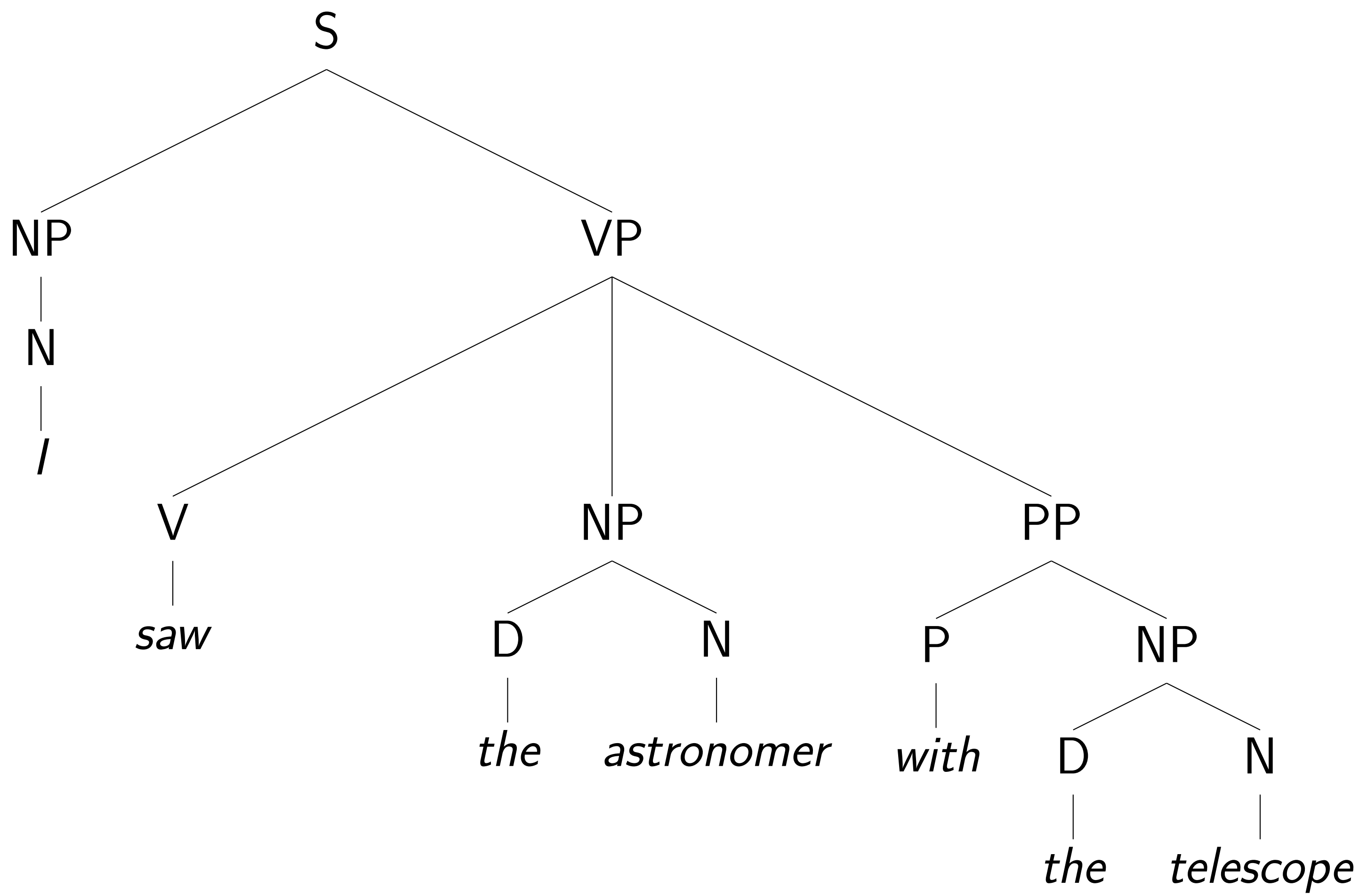
➢ VOCABULARY

D: the, some
A: big, brown, old
N: birds, fleas, dog, hunter, I
V: attack, ate, watched
P: for, beside, with

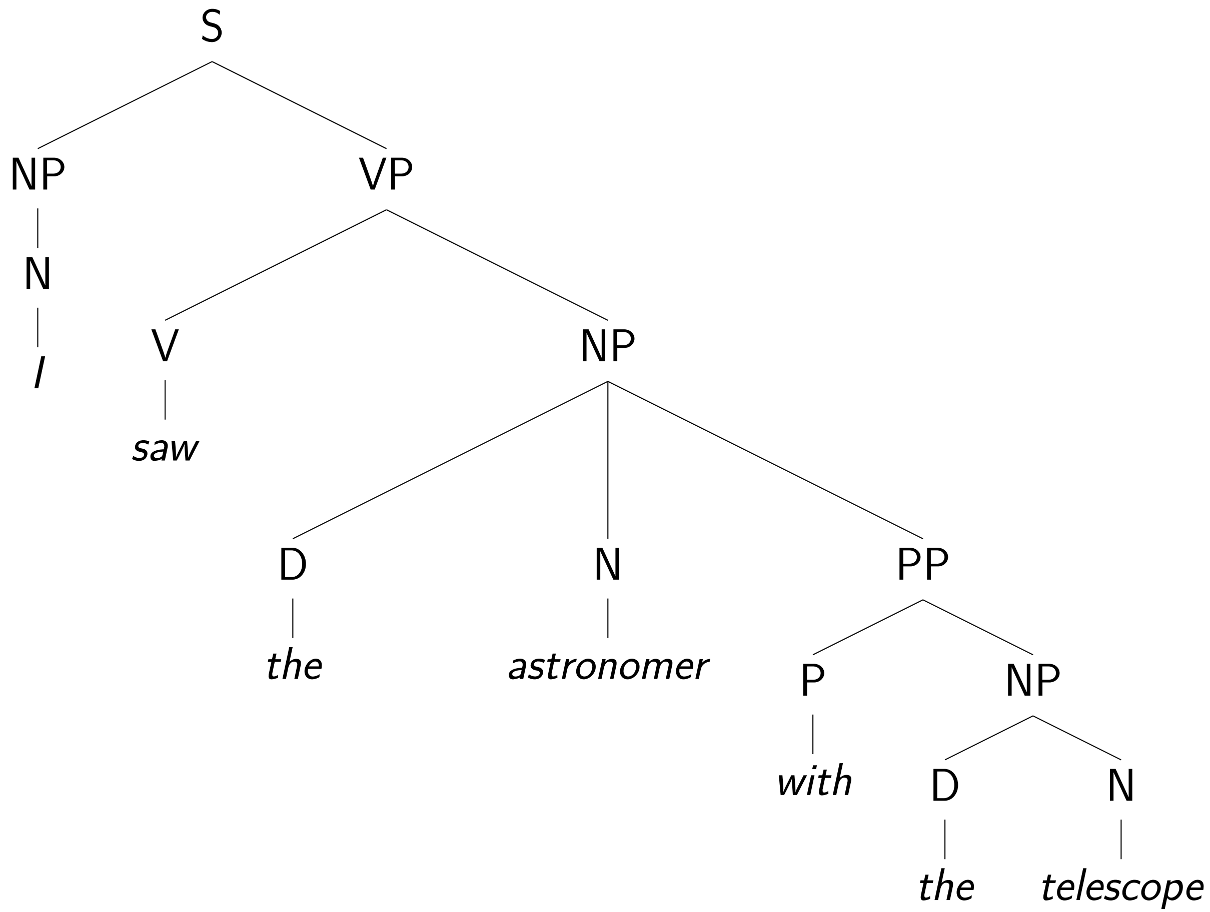
Structural Ambiguity

I saw the astronomer with the telescope.

Structure 1: PP under VP



Structure 2: PP under NP



Constituency Tests

- Recurrent Patterns

(3) *The quick brown fox with the bushy tail jumped over the lazy brown dog with one ear.*

- Coordination

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

- Sentence-initial position

(5) *The election of 2000, everyone will remember for a long time.*

- Cleft sentences

(6) *It was a book about syntax that 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**.

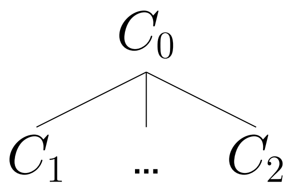
- **Context-Sensitive Grammar** (CSG) gives up 4. That is, it allows the applicability of a grammar rule to depend on what is in the neighboring environment. So rules can have the form:
  $A \rightarrow X$ in the context of α_β ($\alpha A\beta \rightarrow \alpha X\beta$)

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):

  (7) *He arrives this morning.*
  (8) **He arrive this morning.*
  (9) **They arrives this morning.*
  (10) *They arrive this morning.*

Trees and Rules

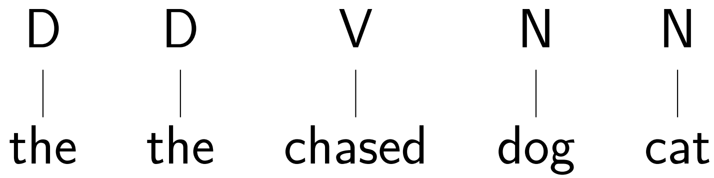


is a well-formed nonlexical tree if (and only if)

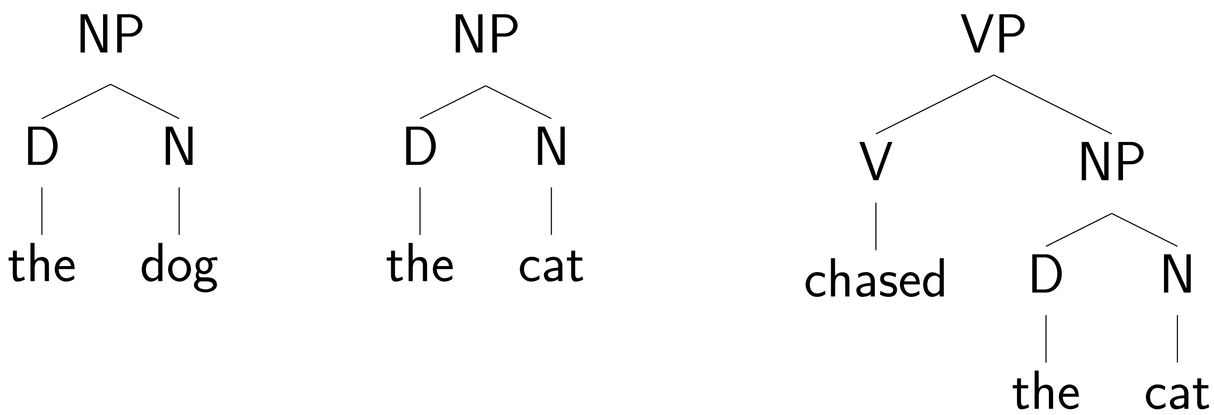
- $C_0, \ldots, C_n$ are well-formed trees
- $C_0 \to C_1 \ldots C_n$ is a grammar rule

Bottom-up Tree Construction

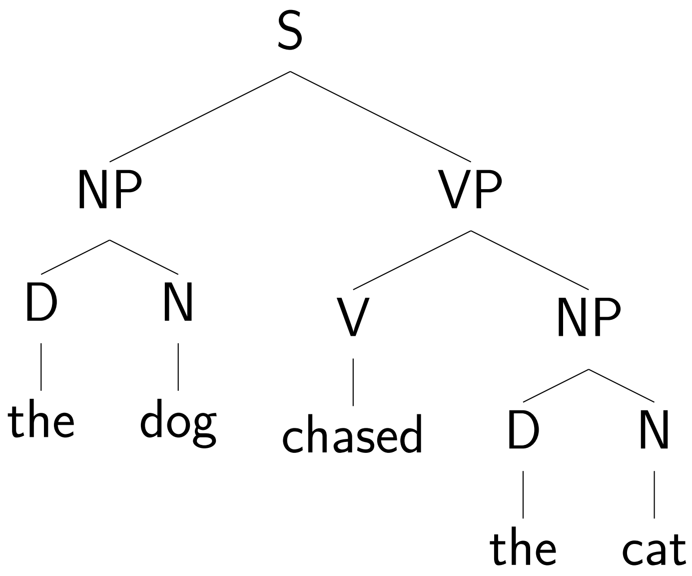
D: the
V: chased
N: dog, cat



$NP \rightarrow D\ N$ $VP \rightarrow V\ NP$

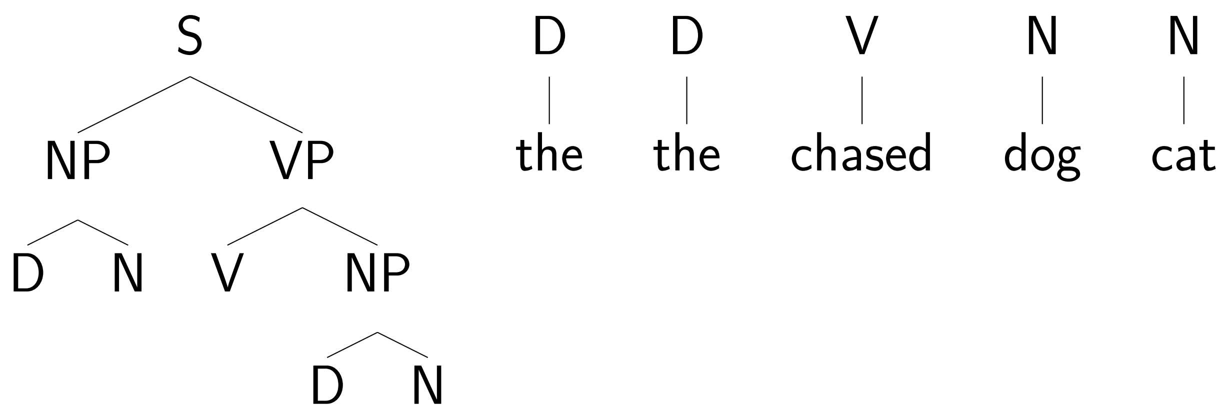


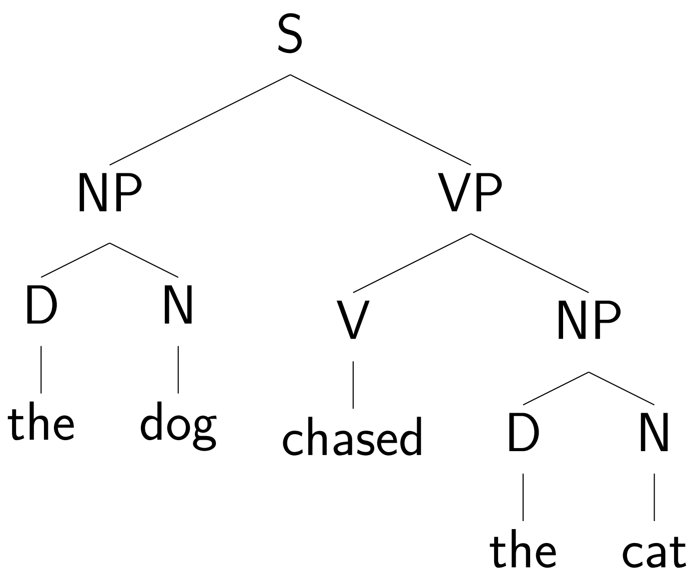
$S \rightarrow NP\ VP$



Top-down Tree Construction

$\mathsf{S} \rightarrow \mathsf{NP\ VP} \quad \mathsf{VP} \rightarrow \mathsf{V\ NP} \quad \mathsf{NP} \rightarrow \mathsf{D\ N} \quad \mathsf{NP} \rightarrow \mathsf{D\ N}$





- **Bottom-up**: string $\rightarrow$ tree
- **Top-down**: tree $\rightarrow$ string
- CFG is **declarative** so it is independent of order

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

Transitivity and Agreement

- Consider the following transitivity examples

  (11) *The bird arrives*
  (12) *The bird devours the worm*
  (13) **The bird arrives the worm*
  (14) **The bird devours*

- Consider the following agreement examples

  (15) *The bird sings*
  (16) *The birds sing*
  (17) **The bird sing*
  (18) **The birds sings*

- Can we deal with them with a CFG?

Summary

1. Fundamentals

2. Investigate

3. Find out some stuff

4. Break our theory

5. Try to fix it.

6. Break it again.

7. Lather, rinse, repeat: we'll do that until we run out of time.

Jorge Hankamer's outline of a syntax course, but it's pretty applicable to everything we do. More formally: Successive Approximation.

Chapter 2, Problem 1

RULES

S	$\rightarrow$	NP VP
NP	$\rightarrow$	(D) NOM
VP	$\rightarrow$	V (NP) (NP)
NOM	$\rightarrow$	N
NOM	$\rightarrow$	NOM PP
VP	$\rightarrow$	VP PP
PP	$\rightarrow$	P NP
X	$\rightarrow$	X+ CONJ X

VOCABULARY

D: a, the
N: cat, dog, hat, man, woman, roof
V: admired, disappeared, put, relied
P: in, on, with
CONJ: and, or

Chapter 2, Problem 1

A Make a well-formed English sentence unambiguous according to this grammar

B Make a well-formed English sentence ambiguous according to this grammar: draw trees

C Make a well-formed English sentence not licensed by this grammar (using V)

D Why is this (C) not licensed?

E Make a string licensed by this grammar that is not a well-formed English sentence

F How can we stop licensing the string in E (stop over-generating)

G How many strings does this grammar license?

H How many strings does this grammar license without conjunctions?

Shieber 1985

- Swiss German example:

(19) *...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:
  - *lönd* "let" governs case on *d'chind* "children"
  - *hälfe* "help" governs case on *Hans* "Hans"
  - *aastriiche* "paint" governs case on *huus* "house"

- This cannot be modeled in a context free language

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 therefore 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

- Prescriptive/descriptive grammar; Competence/performance
- Some history
- Why study syntax?
- Unsuccessful Attempts to model language
- Formal definition of CFG
  - Constituency, ambiguity, constituency tests
  - Central claims of CFG
  - Order independence
  - Weaknesses of CFG
- Next Week: Feature structures

Acknowledgments and References

- Course design and slides borrow heavily from Emily Bender's course: *Linguistics 566: Introduction to Syntax for Computational Linguistics* `http://courses.washington.edu/ling566`

- Thanks to Na-Rae Han for inspiration for the student policies (from *LING 2050 Special Topics in Linguistics: Corpus linguistics*, U Penn; adapted).

- Stuart M. Shieber. (1985) Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:333-343