# HG4041 Theories of Grammar

# Introduction, Organization
# First attempts at a theory of grammar

Francis Bond
**Division of Linguistics and Multilingual Studies**
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 1
Location: HSS SR3

# Overview

- Syllabus; Administrivia
- Prescriptive/descriptive grammar; Competence/performance
- Some history
- Why study syntax?
- Two theories that won't work
- Context Free Grammars
- Central claims of CFG

# Administrivia

**Coordinator** Francis Bond <bond@ieee.org> !<fcbond@ntu.edu.sg>

**Seminar** Tuesday 14:30–18:30 (HSS SR3)

# 100% Continuous Assessment

- Mid-term (20%)
- Final (20%)
- Group Project: Presentation (20%)
  - Give a precise and explicit model of some phenomenon not covered in class
  - The talk must motivate the choice of phenomenon
  - You need only cover existing work
  - In-class presentation with slides or handouts, not to exceed 17 minutes (12 presentation, 5 QA)
  - You should choose something relevant to your final project if possible

- Individual Project (40%)
  - Give a precise and explicit model for some phenomenon not covered in class
    - You should give attested and constructed examples
    - You should clearly indicate what you can and can't explain
    - It is expected that you can not explain everything perfectly
    - Your model should make clear predictions
  - The paper must motivate the choice of phenomenon
  - You should cover relevant existing work and add something new
  - LMS format, not to exceed 12 pages

# Guidelines for Written Work in LMS

- All assignments must follow the *Guidelines to Submitting Written Work for the Division of Linguistics and Multilingual Studies*
  - You can get it from: `http://linguistics.hss.ntu.edu.sg/CurrentStudents/Pages/Resources.aspx`
  - Useful advice on citation, transcription, formatting
  - I also recommend my own *(Computational) Linguistics Style Guide*: `www3.ntu.edu.sg/home/fcbond/data/ling-style.pdf`
  - Proper citation is important
    — failure to cite is plagiarism — **fail subject**
    See the NTU code of academic integrity
    `http://academicintegrity.ntu.edu.sg/`

# What do you learn?

On completion of this module, students should be able to:

- Recognize certain classes of syntactic phenomena
- Build analyses of those phenomena in a precise framework
- Apply the process of building a formalized analysis to test linguistic hypotheses

# Schedule

| Lec. | Topic | | Reading | Problems |
|---|---|---|---|---|
| 1 | Introduction (HPSG) | | SWB 1–2 | 1:1 |
| 2 | Feature Structures | | SWB 3 | 3:1, 3 |
| 3 | Complex Feature Values | | SWB 4 | 4:1, 5, 6 |
| 4 | Semantics | | SWB 5–6 | 5:1; 6:1, 3, 4, 5 |
| 5 | Binding | | SWB 7 | 7:1, 2 |
| 6 | The Structure of the Lexicon | **Mid-term** | SWB 8 | 8:1, 2, 6 |
| 7 | Realistic Grammar | | SWB 9 | 9:1 |
| 8 | Passive | | SWB 10 | 10: 1, 3 |
| 9 | Dummies and Idioms | | SWB 11 | 11:1, 3, 4 |
| 10 | Raising and Control | | SWB 12 | 12:1, 2, 4, 6 |
| 11 | Long Distance Dependencies | **Final** | SWB 14 | 14: 1, 2, 3 |
| 12 | Wrap-up | **Project Presentations** | SWB 16 | |
| | **Research Paper** due two weeks after presentations | | | |

# Textbook and Readings

- Textbooks
  - Sag, Wasow and Bender 2003 *Syntactic Theory: A Formal Introduction* 2nd ed. CSLI (**required**)
  - You should read all chapters assigned before class.
  - Ideas from the book will be pursued in parallel with the topics given above.

# Student Responsibilities

By remaining in this class, the student agrees to:

1. Make a good-faith effort to learn and enjoy the material.
2. Read assigned texts and participate in class discussions and activities.
3. Submit assignments on time.
4. Attend class at all times, barring special circumstances (see below).
5. Get help early: approach us when you first have trouble understanding a concept or homework problem rather than complaining about a lack of understanding afterward.
6. Treat other students with respect in all class-related activities, including on-line discussions.

# Attendance

1. You are expected to attend all classes.

2. Be on time - lateness is disruptive to your own and others' learning.

3. Valid reasons for missing class include the following:

   (a) A medical emergency (including mental health emergencies)
   (b) A family emergency (death, birth, natural disaster, etc).

   You must provide documentation to me and the student office.

4. There will be significant material covered in class that is not in your readings. You cannot expect to do well without coming to class.

5. If you miss a class, it is your responsibility to get the notes, any handouts you missed, schedule changes, etc. from a classmate.

# Remediation and Academic Integrity

1. No late work will be accepted, except in the case of a documented excuse.

2. For planned, justified, absences on class days or days on which assignments are due, advance notice must be provided.

3. Cheating will not be tolerated. Violations, including plagiarism, will be seriously dealt with, and could result in **a failing grade for the entire course**.

4. For all other issues of academic integrity, refer to the University Honour Code:
   `http://academicintegrity.ntu.edu.sg/`

5. As always, use your common sense and conscience.

# The winning strategy

- Read the books before class (and after again, if necessary)
- Work together: make study groups
- Homework: Discuss as much as you want, write up your own answers
- Exams: No discussion
- Ask questions . . . early and often!

# Resources

- Glossary at back of textbook
- Grammar summaries and Appendix A
- Answers to exercises at back of book
- Each other, grad-students, office hours, . . .
- Online:
  - HPSG at Stanford: `http://hpsg.stanford.edu/`
  - English Resource Grammar: `http://erg.delph-in.net/logon`
  - Wikipedia page has lots of links

# Two Conceptions of Grammar

- PRESCRIPTIVE
  - Rules against certain usages. Few if any rules for what is allowed
  - Proscribed forms generally in use
  - Explicitly normative enterprise
- DESCRIPTIVE
  - Rules characterizing what people do say
  - Goal to characterize all and only what speakers find acceptable
  - Tries to be scientific

# Uses of Grammar

- PRESCRIPTIVE
  - Identify speaker's socioeconomic class & education level
  - Identify level of formality of a particular usage
- DESCRIPTIVE
  - Understand how people produce & understand language
  - Identify similarities & differences across languages
  - Development of language technologies

# Prescriptive grammar

- Examples of silly prescriptive rules?
- Examples of useful prescriptive rules?
- Some applications which might need to encode prescriptive rules?

# Fill in the blanks:

*he/his, they/their*, or something else?

(1) Everyone insisted that _____ record was unblemished.
(2) Everyone drives _____ own car to work.
(3) Everyone was happy because _____ passed the test.
(4) Everyone left the room, didn't _____?
(5) Everyone left early. _____ seemed happy to get home.

# Descriptive Grammar: an example

(6) F_____ yourself!
(7) Go f_____ yourself!
(8) F_____ you!
(9) *Go f_____ you!

- Who taught you this?
- How did you learn it?

# Kinds of Things We'll Worry About

- Where to use reflexives (e.g. *myself*) vs. ordinary pronouns (*I, me*)
- Agreement (e.g. *We sing* vs. **We sings*)
- Word order (e.g. **Sing we*)
- Case (e.g. **Us sing*)
- Coordinate conjunction (e.g. *We sing and dance*)
- How to form questions, imperatives, negatives, . . .

. . . and much more

# Competence vs. Performance

- The Distinction
  - Competence - knowledge of language
  - Performance - how the knowledge is used
- Examples

(10) That Sandy left bothered me.

(11) That that Sandy left bothered me bothered Kim

(12) That that that Sandy left bothered me bothered Kim bothered Bo

(13) The horse raced past the barn fell

# Competence v. Performance

(14) You are what you eat

(15) You are what what you eat eats, too

(16) You are what what what you eat eats eats, too

# Acceptability vs. grammaticality

- A sentence is **acceptable** if native speakers say it sounds good.
- A sentence is **grammatical** (with respect to a particular grammar) if the grammar licenses it.
- Linguists are sometimes sloppy about the difference.
- Some people argue that it should be modeled probabilistically rather than as a binary distinction

# Some History

- Writings on grammar go back at least 3000 years
- Until 200 years ago, almost all of it was prescriptive
- Until 50 years ago, most linguistic work concerned sound systems (phonology), word structure (morphology), and the historical relationships among languages

# The Generative Revolution

- Noam Chomsky's work in the 1950s radically changed linguistics, making syntax central.
- Chomsky has been the dominant figure in linguistics ever since.
- The theory we will develop (HPSG) is in the tradition started by Chomsky, but diverges from his work in many ways.

# Main Tenets of Generative Grammar

- Grammars should be formulated precisely and explicitly.
- Languages are infinite, so grammars must be tested against invented data, not just attested examples.
- The theory of grammar is a theory of human linguistic abilities.

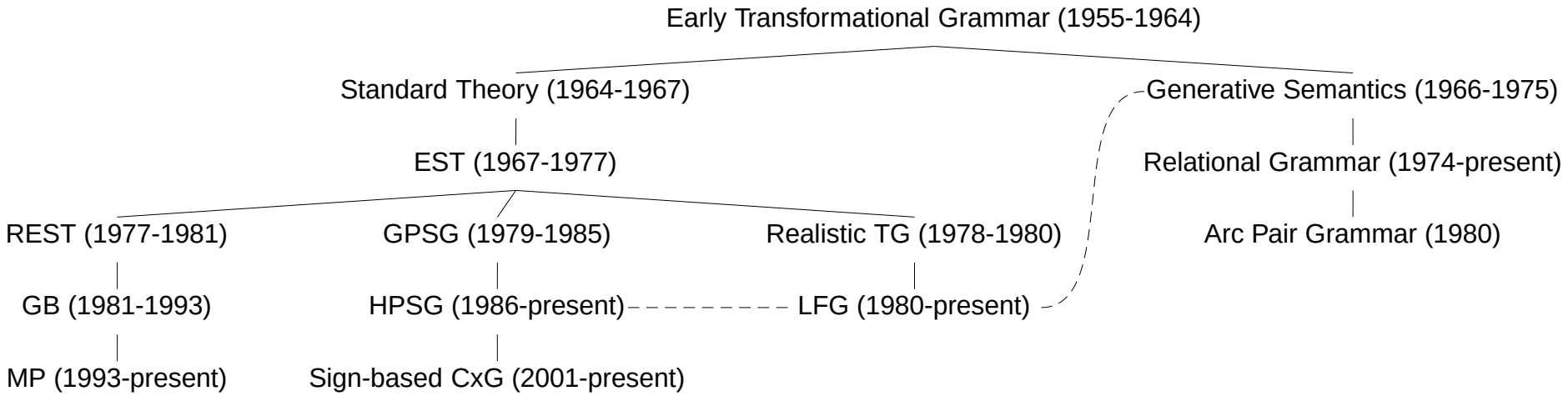
# What does a theory do?

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

# Some of Chomsky's Controversial Claims

- The superficial diversity of human languages masks their underlying similarity.
- All languages are fundamentally alike because linguistic knowledge is largely innate.
- The central problem for linguistics is explaining how children can learn language so quickly and easily.

# Family Tree of Generative Syntactic Theories




- Many Other Theories
  - Dependency Grammar
  - Combinatory Categorical Grammar
  - Optimality Theory
  - Tree Adjoining Grammar

# Why Study Syntax?

- Why should linguists study syntax?
- Why should computational linguists study syntax?
- Should anyone else study syntax? Why?
- Why are you studying syntax?

# 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

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

(18) *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. |
| \| | 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 |

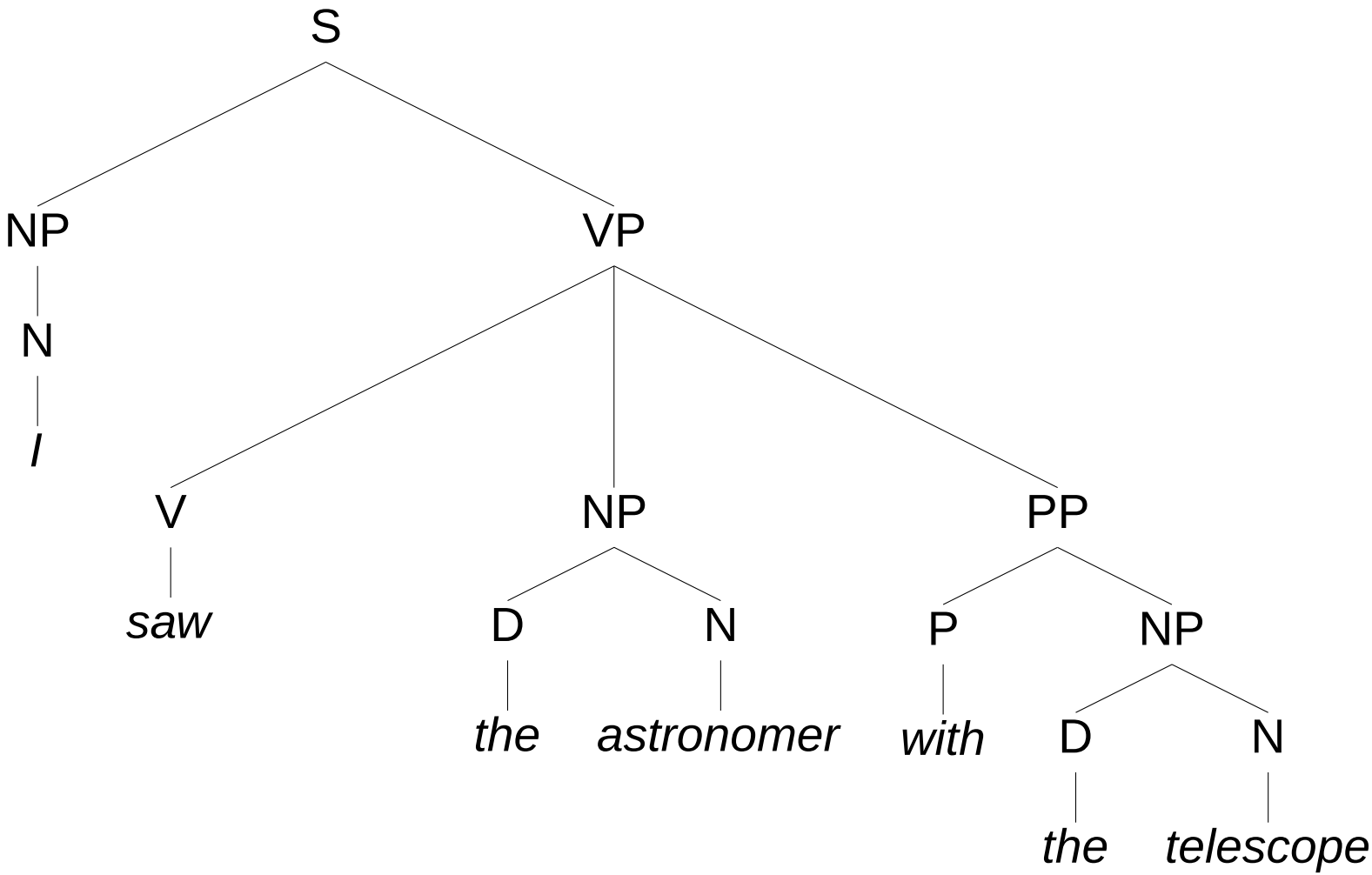
- VOCABULARY

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

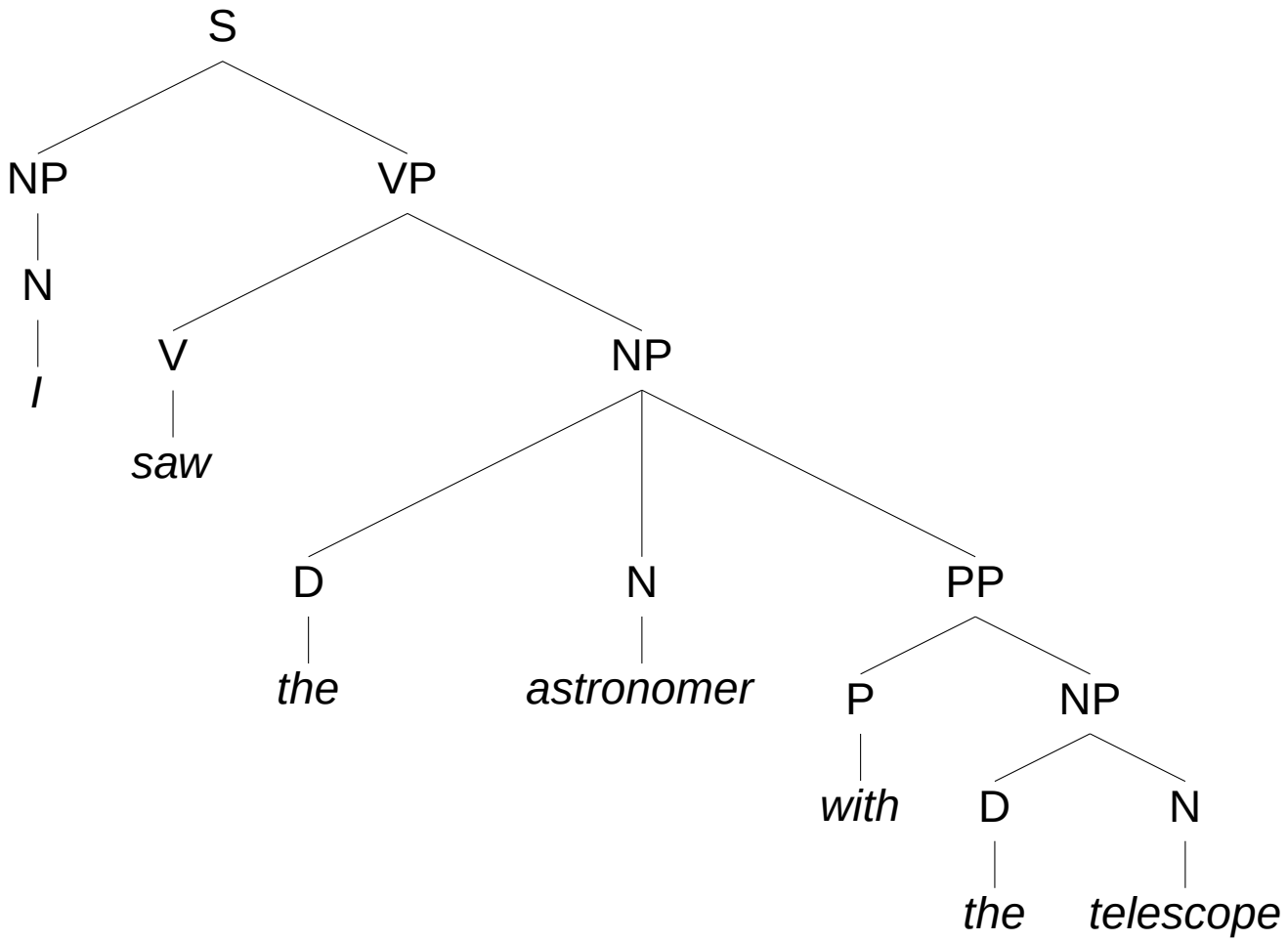
# Structural Ambiguity

I saw the astronomer with the telescope.

# Structure 1: PP under VP



# Structure 2: PP under NP



# Constituency Tests

- Recurrent Patterns
  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 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 $\alpha\_\beta$ ($\alpha A \beta \rightarrow \alpha X \beta$)

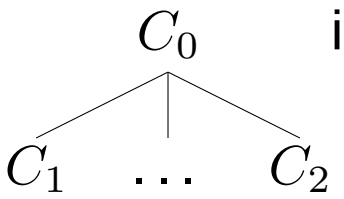
# Possible Counterexamples

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

# Trees and Rules

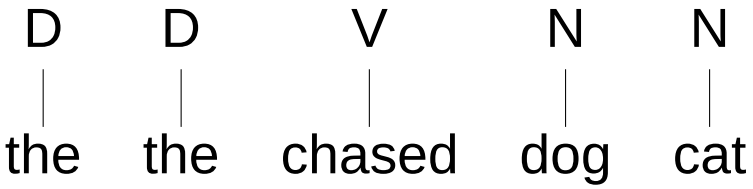



is a well-formed nonlexical tree if (and only if)

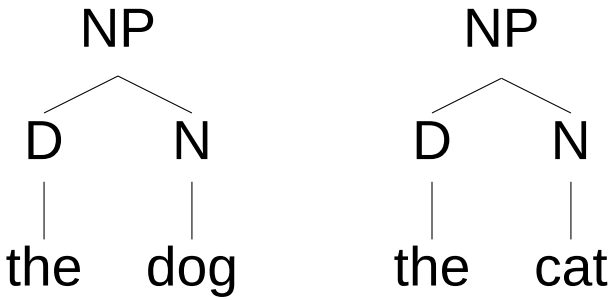
- $C_0, \ldots, C_n$ are well-formed trees
- $C_0 \rightarrow C_1 \ldots C_n$ is a grammar rule

# Bottom-up Tree Construction

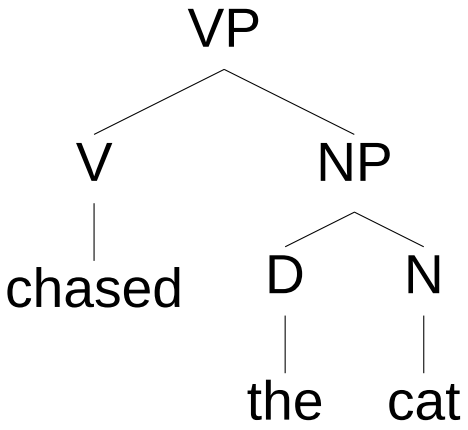
D: the
V: chased
N: dog, cat



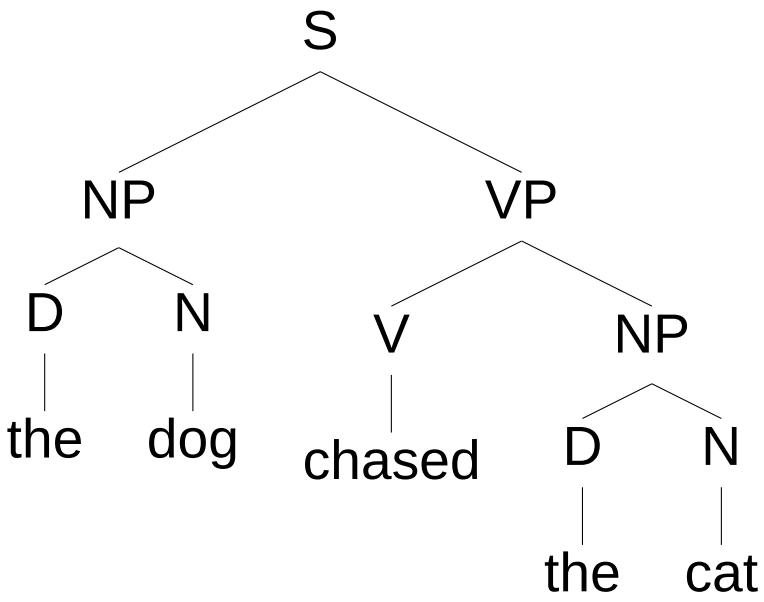

$NP \rightarrow D\ N$



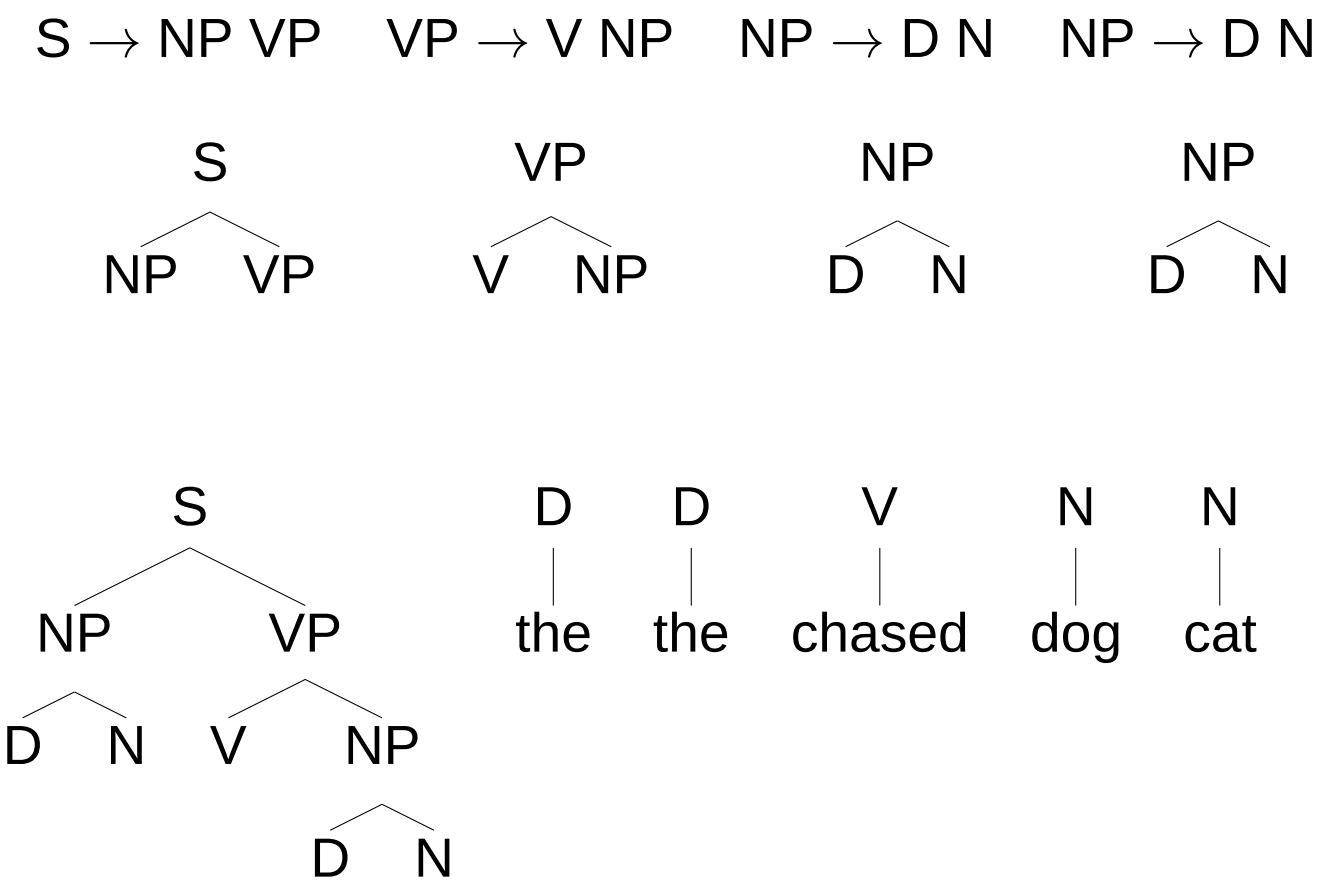

$VP \rightarrow V\ NP$

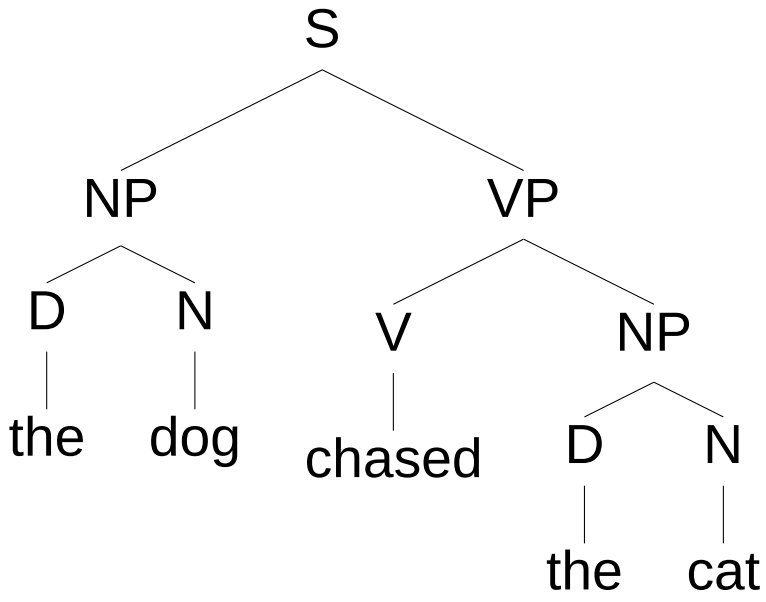


$S \rightarrow NP\ VP$



# Top-down Tree Construction






- **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 agreement examples

  (19) The bird sings
  (20) The birds sing
  (21) *The bird sing
  (22) * The birds sings

- Consider the following transitivity examples

  (23) The bird arrives
  (24) The bird devours the worm
  (25) *The bird arrives the worm
  (26) *The bird devours

- Can we deal with them with a CFG?

# 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 Well-formed English sentence unambiguous according to this grammar

B Well-formed English sentence ambiguous according to this grammar: draw trees

C Well-formed English sentence not licensed by this grammar (using $V$)

D Why is this not licensed?

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

F How can we license it (without 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:

(27) *. . . 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