ANTONIS ANASTASOPOULOS

## CS499 INTRODUCTION TO NLP

# NATURAL LANGUAGE

https://cs.gmu.edu/~antonis/course/cs499-spring21/

# STRUCTURE OF THIS LECTURE

1 Introduction

2 Language

3 What is NLP

4 Course Logistics

# THIS COURSE IS NEW!

We are making the slides and developing the course (largely) from scratch

- Please give us feedback!

# HOW CAN YOU HELP

Help us make the lectures better!

- email us if you spot a typo
- slides and video will be posted (with typo fixes) after the lecture
- please email us or post on Piazza
- Most important: participate!

# WEBSITES

Course Website: https://cs.gmu.edu/~antonis/course/cs499-spring21/syllabus/

- will be regularly updated with important information

Piazza:

- not required, but highly encouraged
- group discussion can always help better understand the course material

# ABOUT ME

My name is Antonis

- pronounced A-**do**-nis
- no need for titles, Antonis is fine (or Antoni if you want to follow Greek inflection rules)

I do research in NLP at GMU

BSc/MSc from National Technical University of Athens
PhD from Notre Dame
Postdoc at Languages Technologies Institute at Carnegie Mellon University

# TEACHING ASSISTANT

Mahfuz Alam

- PhD at GMU
- Research on Machine Translation

Office Hours: Fridays, 3-4pm

# WHY NLP

“

WE LIKED THE NAME “ALPHABET” BECAUSE IT MEANS A COLLECTION OF LETTERS THAT REPRESENT LANGUAGE, ONE OF HUMANITY’S MOST IMPORTANT INNOVATIONS, AND IS THE CORE OF HOW WE INDEX GOOGLE SEARCH

*— Larry Page, co-founder of Google*

”

# NLP IS EVERYWHERE!

The Association of Computational Linguistics (ACL) was founded in 1962

In the 1970s, the conferences had < 100 participants

EMNLP 2019 had > 3000 participants

NLP is the backbone of many major companies

“On May 19 remind me it's dad's birthday”

Here's your reminder for May 19, 2012 at 9 am:

Dad's birthday

san f

san francisco weather
san francisco
san francisco giants
san fernando valley
san francisco state university
san francisco hotels
san francisco 49ers
san fernando
san fernando mission
san francisco zip code

# WHAT DO YOU THINK OF WHEN YOU THINK OF NLP?

# WHAT CAN YOU DO WITH NLP

Answer Questions using the Web

Translate from one language to another

Manage messages intelligently

Understand and follow directions

Fix spelling and/or grammar

Write poems

Grade exams

Read all scientific articles and discover new knowledge

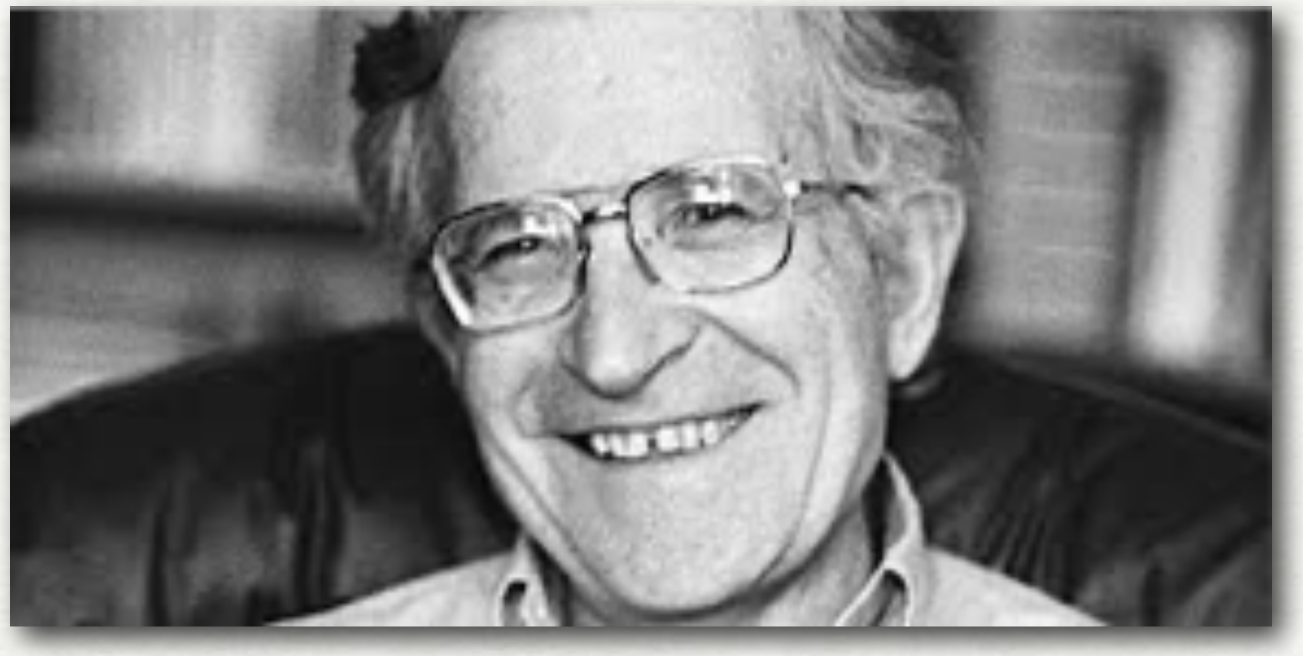
Help under-served and vulnerable populations (refugees, disabled)

Study and document/reinvigorate indigenous languages

# STATISTICAL NLP

In the 1990s, the field switched from intuition-driven to data-driven…

Noam Chomsky



“But it must be recognized that the notion ‘probability of a sentence’ is an entirely useless one, under any known interpretation of the term” (1969, p57)

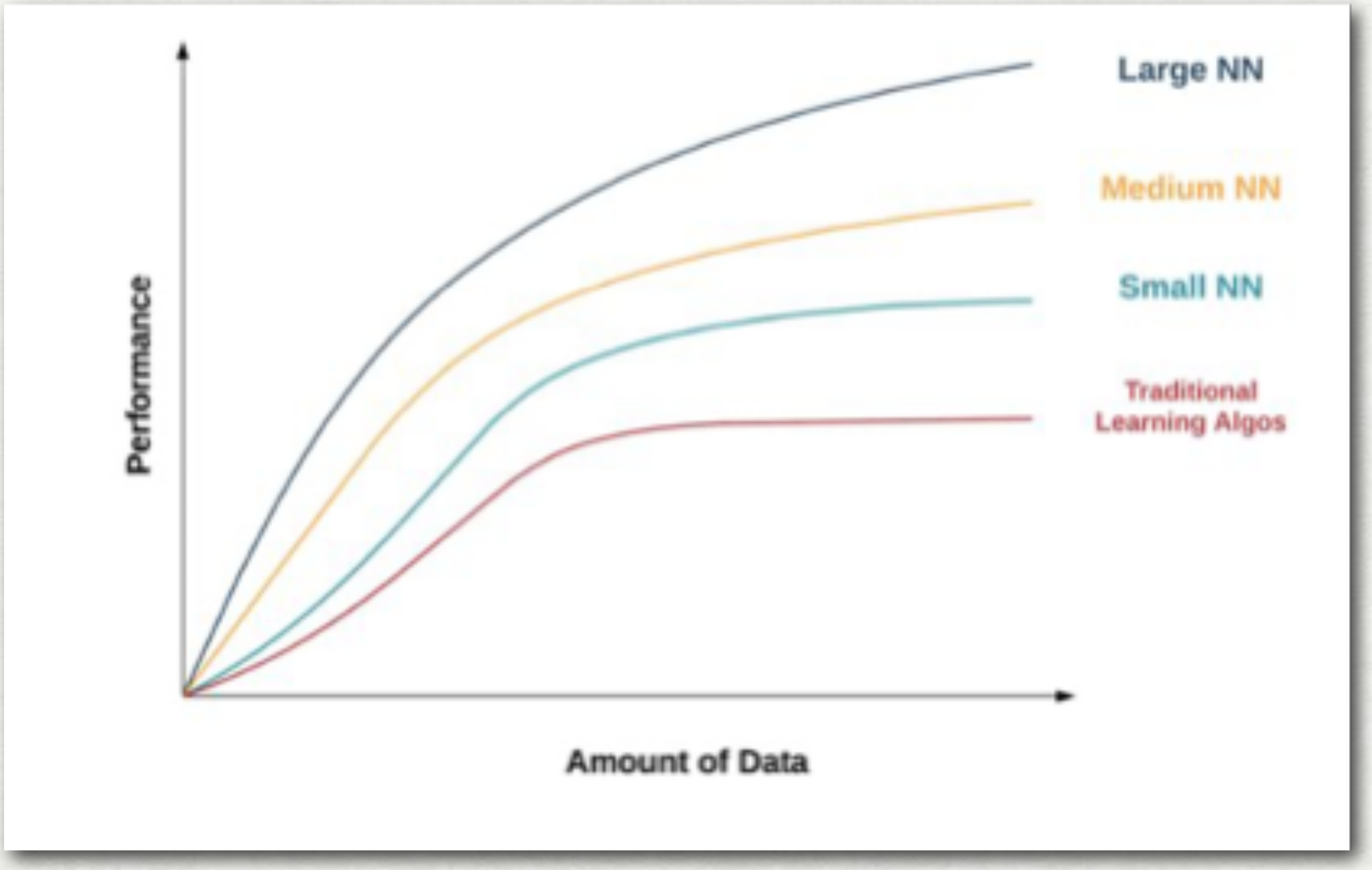
Fredrick Jelinek

“Every time I fire a linguist, the performance of my speech recognizer goes up” (1988)

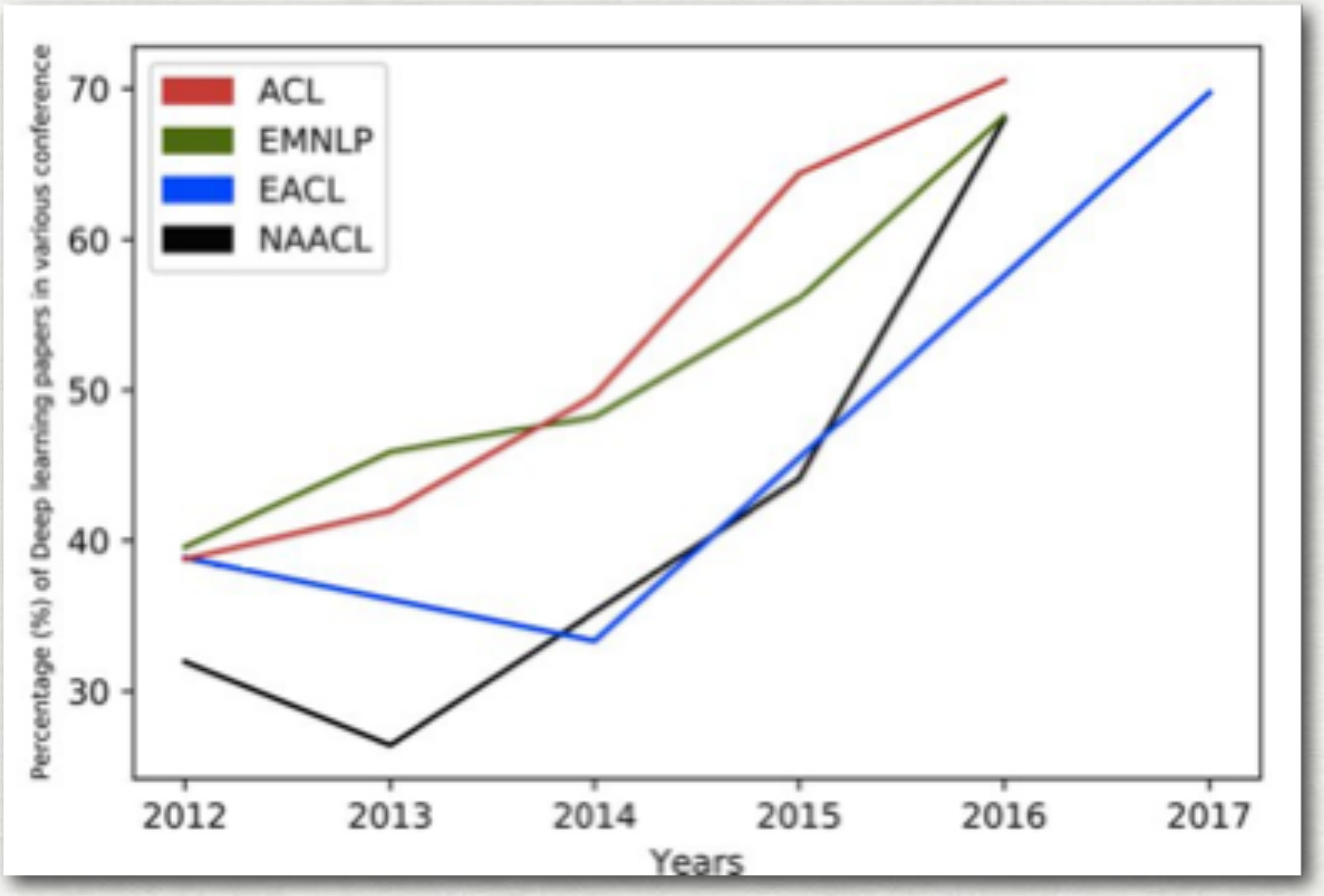
# DEEP LEARNING FOR NLP

…and after 2010 we are training bigger models on more data (using neural networks on GPUs).

More Data, Better Performance



Dominates top Venues



# WHAT ABOUT LINGUISTICS?

Does Language have inherent structure? How is it structured?

Natural language is extremely complex — have you been exposed to a formal description of it?

Other formal models for complex natural phenomena you have already studied:

- falling objects (Newton’s laws)
- electromagnetism (Maxwell’s equations)
- evolution (Darwin’s theory)

Linguistics is the ***scientific*** study of language

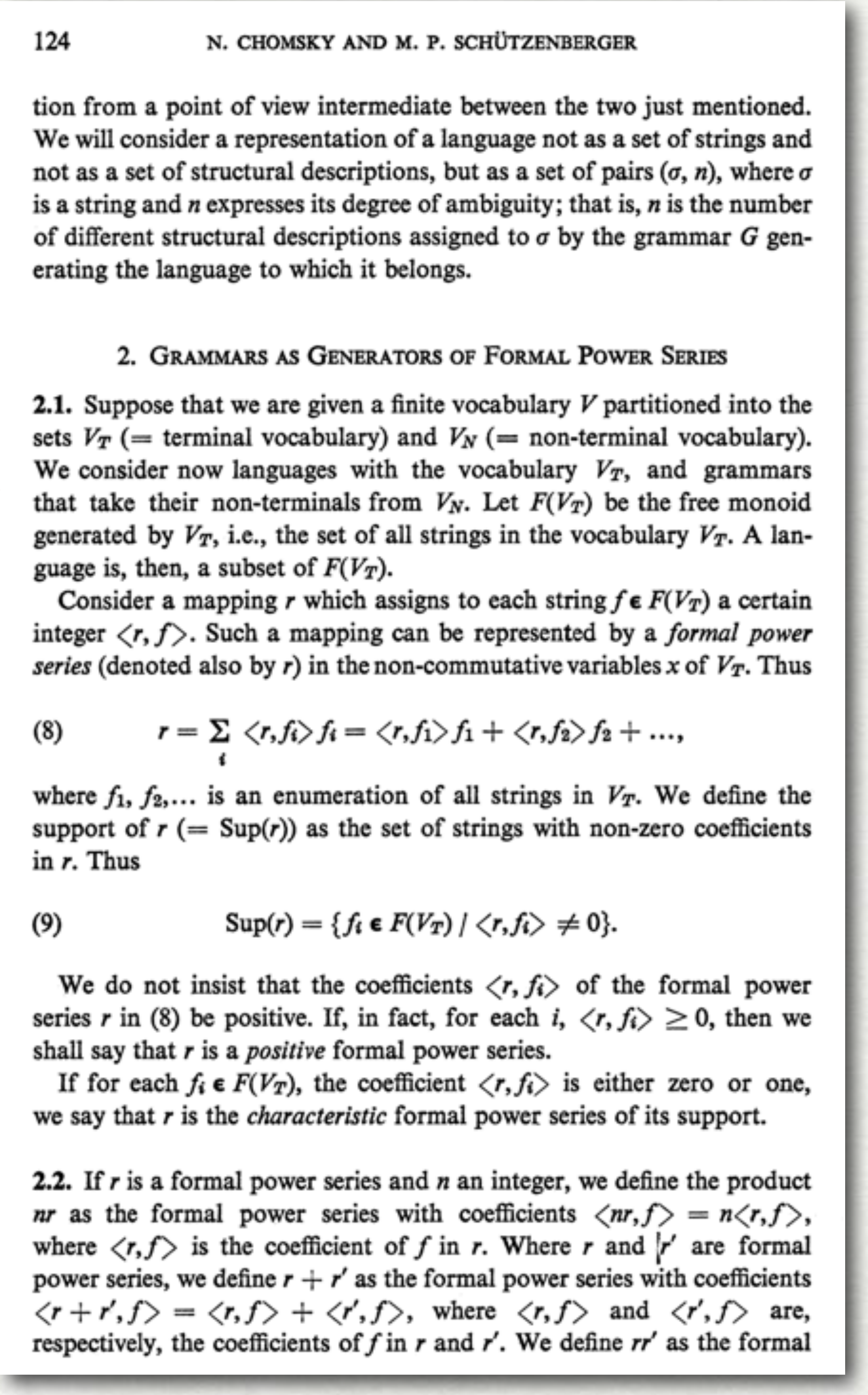
# WHAT ABOUT LINGUISTICS?

Traditionally, Linguistics was classified in the Humanities

But, it is a SCIENCE.

Have you thought about mathematically modeling language?

link



124 N. CHOMSKY AND M. P. SCHÜTZENBERGER

tion from a point of view intermediate between the two just mentioned. We will consider a representation of a language not as a set of strings and not as a set of structural descriptions, but as a set of pairs $(\sigma, n)$, where $\sigma$ is a string and $n$ expresses its degree of ambiguity; that is, $n$ is the number of different structural descriptions assigned to $\sigma$ by the grammar $G$ generating the language to which it belongs.

2. Grammars as Generators of Formal Power Series

**2.1.** Suppose that we are given a finite vocabulary $V$ partitioned into the sets $V_T$ (= terminal vocabulary) and $V_N$ (= non-terminal vocabulary). We consider now languages with the vocabulary $V_T$, and grammars that take their non-terminals from $V_N$. Let $F(V_T)$ be the free monoid generated by $V_T$, i.e., the set of all strings in the vocabulary $V_T$. A language is, then, a subset of $F(V_T)$.

Consider a mapping $r$ which assigns to each string $f \in F(V_T)$ a certain integer $\langle r, f \rangle$. Such a mapping can be represented by a *formal power series* (denoted also by $r$) in the non-commutative variables $x$ of $V_T$. Thus

$$(8) \qquad r = \sum_i \langle r, f_i \rangle f_i = \langle r, f_1 \rangle f_1 + \langle r, f_2 \rangle f_2 + \ldots,$$

where $f_1, f_2, \ldots$ is an enumeration of all strings in $V_T$. We define the support of $r$ (= Sup($r$)) as the set of strings with non-zero coefficients in $r$. Thus

$$(9) \qquad \mathrm{Sup}(r) = \{ f_i \in F(V_T) \,/\, \langle r, f_i \rangle \neq 0 \}.$$

We do not insist that the coefficients $\langle r, f_i \rangle$ of the formal power series $r$ in (8) be positive. If, in fact, for each $i$, $\langle r, f_i \rangle \geq 0$, then we shall say that $r$ is a *positive* formal power series.

If for each $f_i \in F(V_T)$, the coefficient $\langle r, f_i \rangle$ is either zero or one, we say that $r$ is the *characteristic* formal power series of its support.

**2.2.** If $r$ is a formal power series and $n$ an integer, we define the product $nr$ as the formal power series with coefficients $\langle nr, f \rangle = n\langle r, f \rangle$, where $\langle r, f \rangle$ is the coefficient of $f$ in $r$. Where $r$ and $r'$ are formal power series, we define $r + r'$ as the formal power series with coefficients $\langle r + r', f \rangle = \langle r, f \rangle + \langle r', f \rangle$, where $\langle r, f \rangle$ and $\langle r', f \rangle$ are, respectively, the coefficients of $f$ in $r$ and $r'$. We define $rr'$ as the formal

# FOR EXAMPLE: GRAMMATICALITY

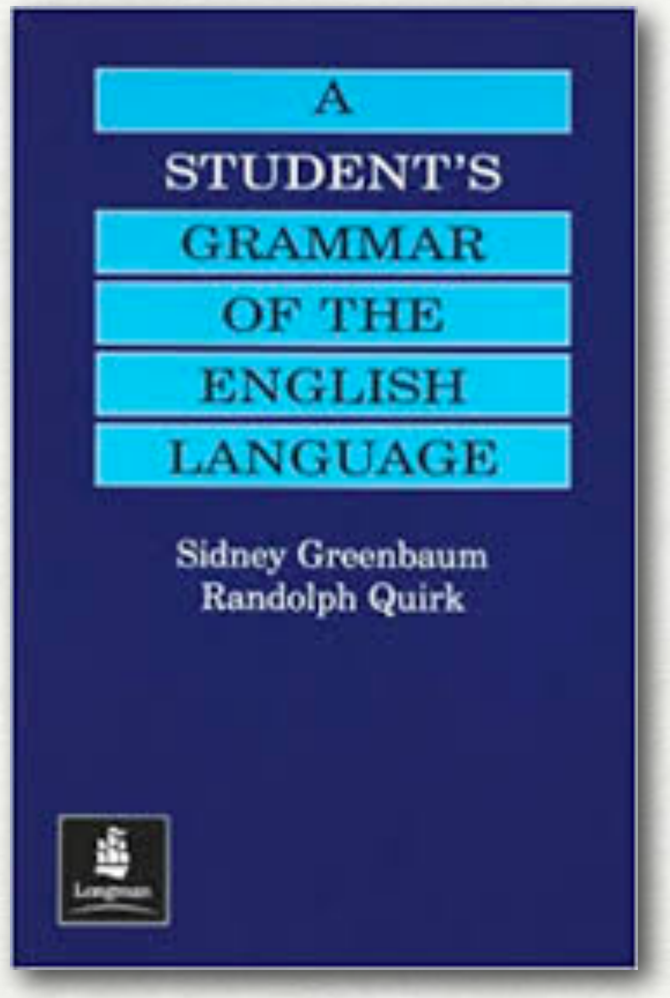
**Fact:** some sentences are grammatical, some are not
*(note: might depend on dialect/speaker)*

Humans tend to have strong (binary) judgements

Jane went to the store.

store to Jane went the.

Jane went store.



"But we learned grammar at school!"

**Perscriptivism**

- you focus on avoiding "common mistakes"
- forced to obey (arbitrary?) rules
- e.g. don't end a sentence in a preposition

# THE SET OF GRAMMATICAL SENTENCES

Based on a finite lexicon, the set is **infinite.**

Non-regular (show with pumping lemma)

Why? **Recursion**

$(NP)^n\ (VP)^{n-1}$ likes tuna fish

[The cat [likes tuna fish]]

The cat likes tuna fish

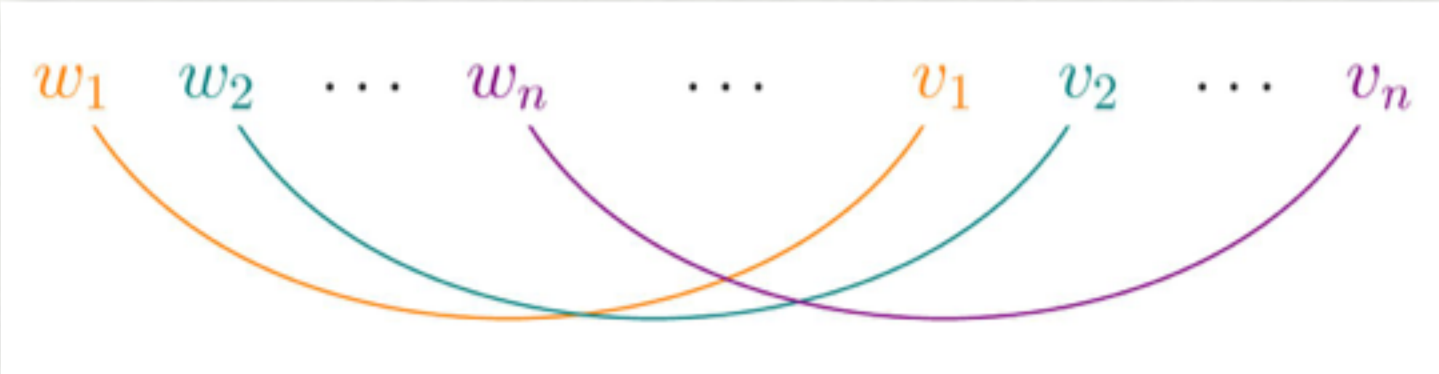
[The cat the dog [chased [likes tuna fish]]]

The cat the dog chased likes tuna fish

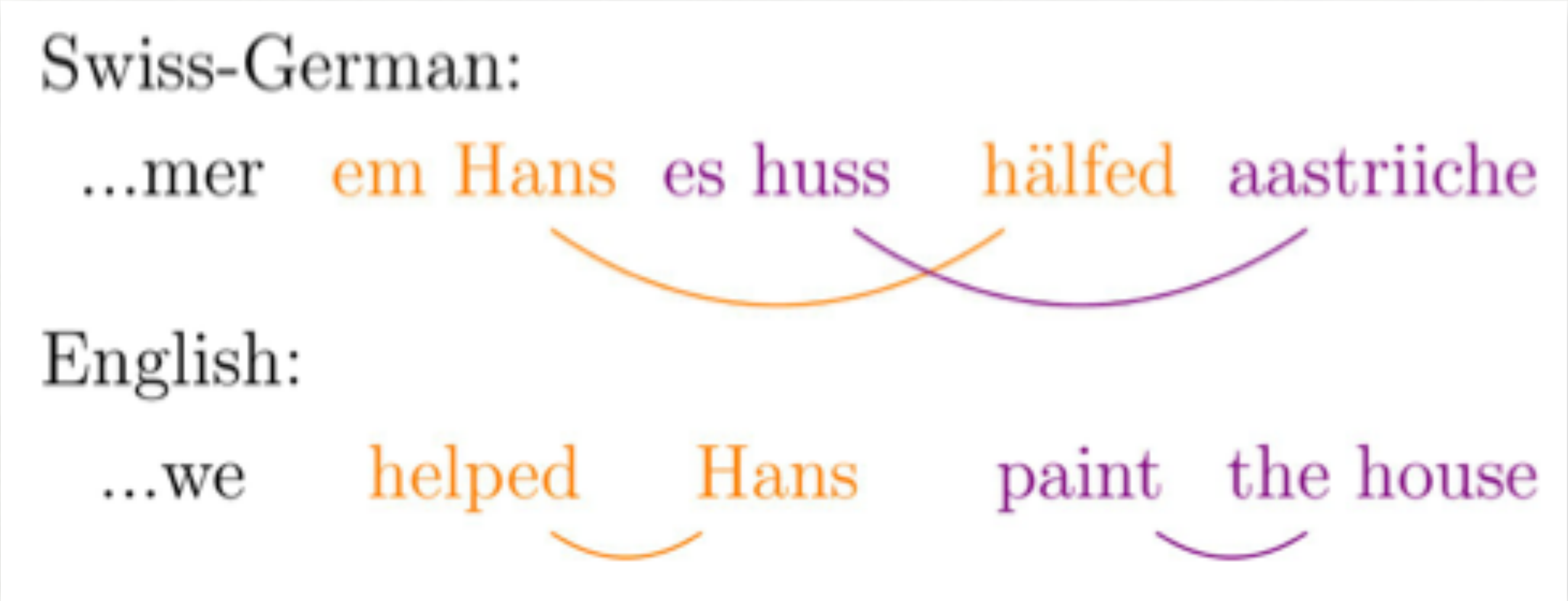
[The cat the dog the rabbit [bit [chased [likes tuna fish]]]]

The cat the dog the rabbit bit chased likes tuna fish

**Note:** Natural  Language is **not** context-free



Cross-serial dependencies are not context-free (link)



Swiss-German:

…mer em Hans es huss hälfed aastriiche

English:

…we helped Hans paint the house

# SIDE NOTE: HOW COMPLEX IS NATURAL LANGUAGE

Many suspect natural language is *mildly* context sensitive

Polynomial time recognition algorithm

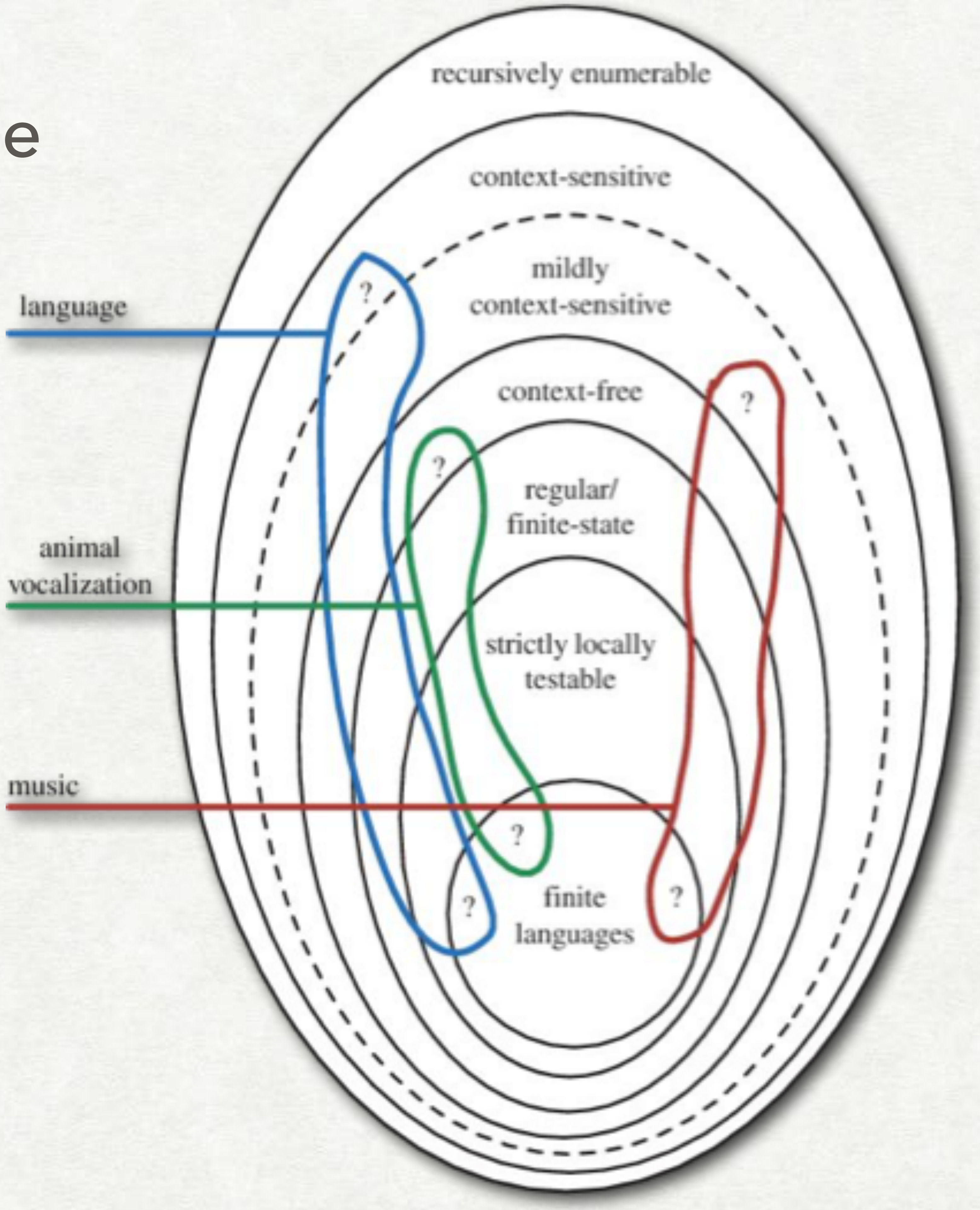
(context sensitive language generally require exponential time)

Existing formalisms:

tree-adjoining grammar — parsable in $O(n^6)$

combinatory categorial grammar — also $O(n^6)$

Morphology (word building) is speculated to be regular



Rohrmeier (2015)

# SYNTAX

Which sentences are well formed? (Grammaticality problem)

Formal Language Theory

has to prove the adequacy of the formalisms in modeling known syntactic phenomena, and prove properties of formalisms

Also, complexity

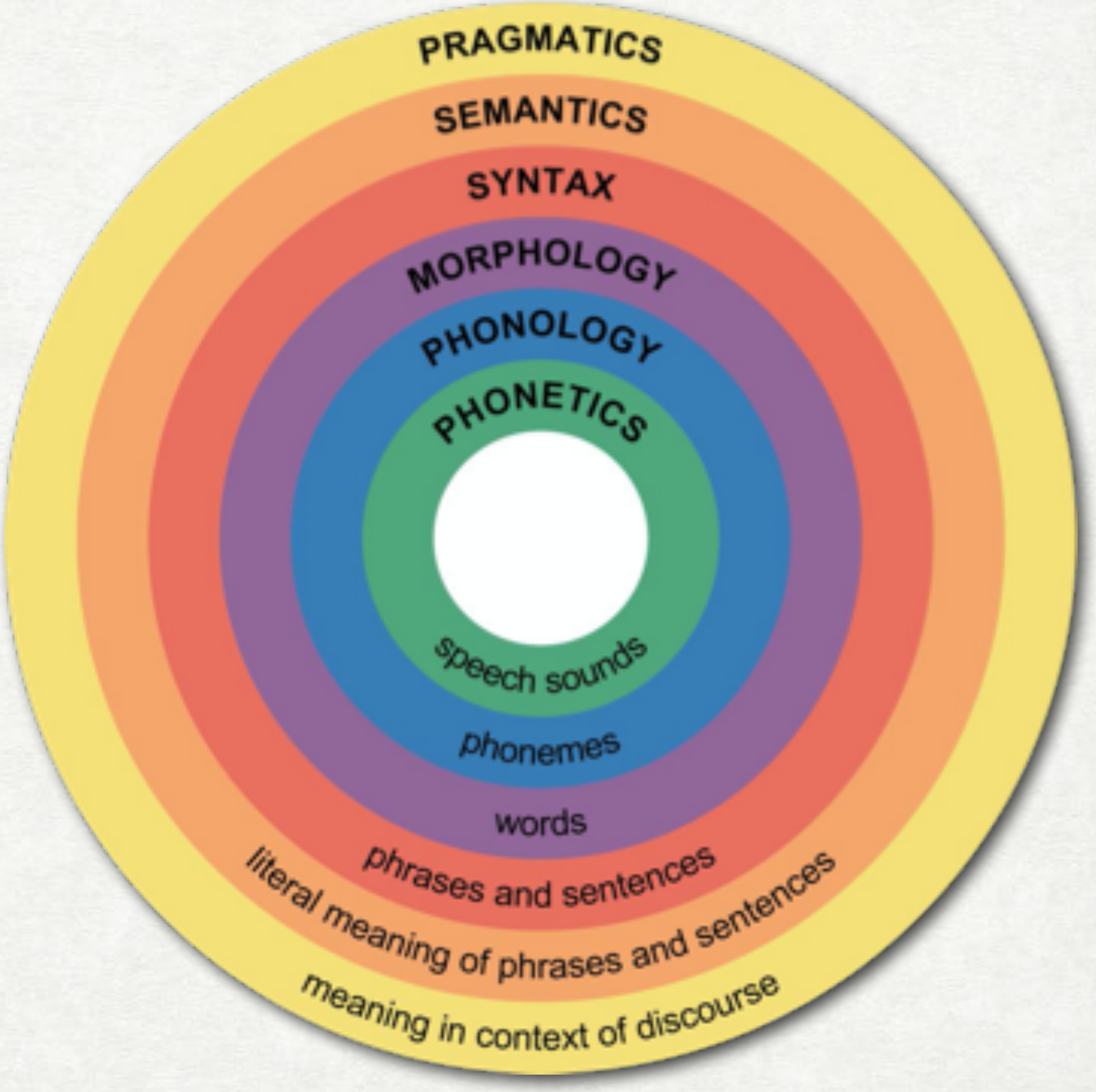
We need the simplest formalism possible

# LINGUISTICS

Linguistics is more than syntax!!

Linguistics studies all aspects of language

**Phonetics and Phonology**: sounds
**Morphology**: meaningful components of words
**Syntax**: relationships between words
**Semantics**: meaning
**Pragmatics:** meaning + intention
**Discourse:** go beyond single utterances



# NLP IS NOT LINGUISTICS

**Automate the analysis, generation, and acquisition of natural (i.e. human) language**

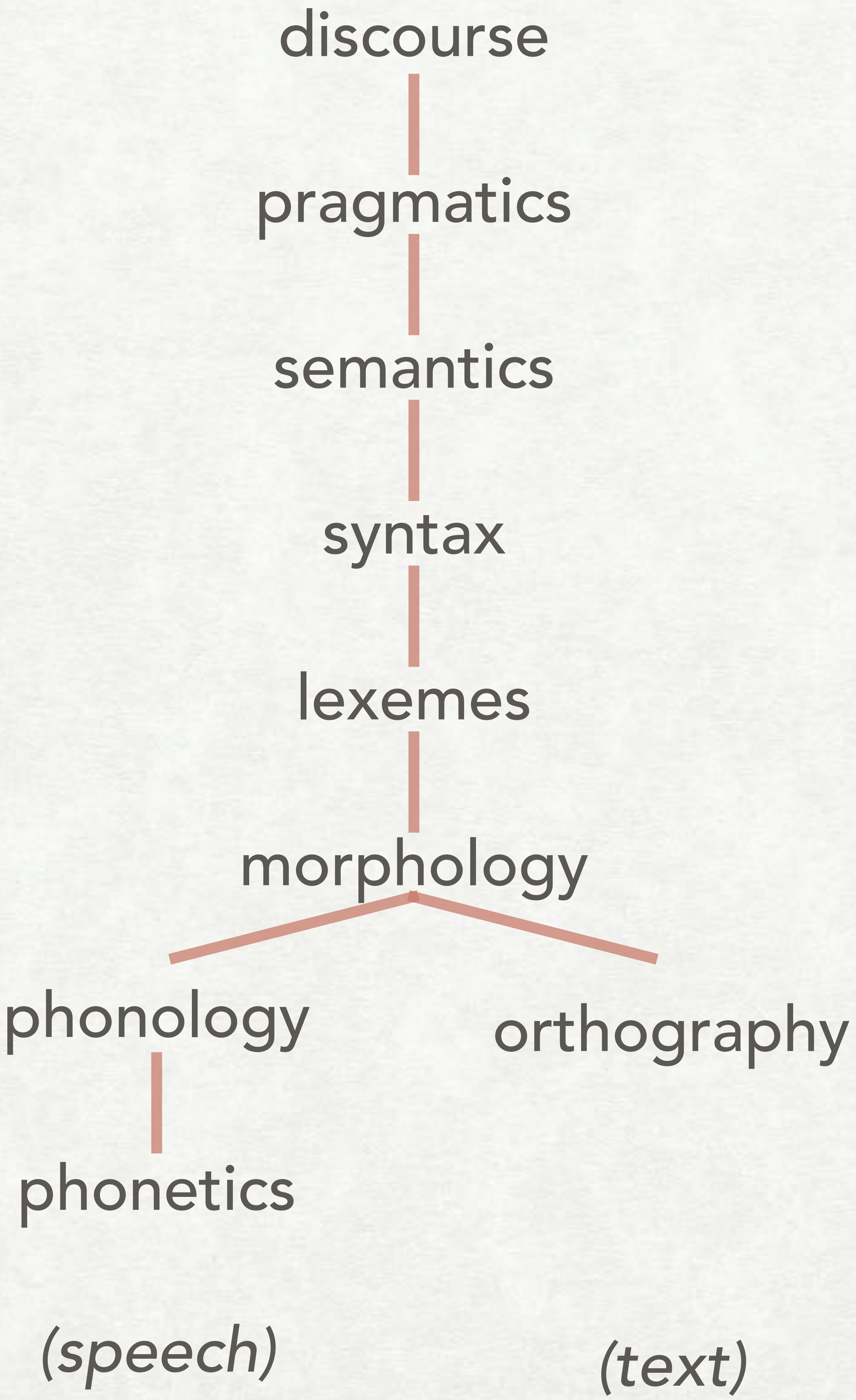
**Analysis/Understanding**: input is language, output is a representation

**Generation**: input is representation, output is language

**Acquisition**: obtain the representation and necessary algorithms from data

Our goal is to engineer systems to **solve a problem**.
*Note: this does not mean that the best solution is a machine learning (statistical) solution!*

# LEVELS OF REPRESENTATION



The mappings between level are extremely complex!

Different applications will require different representations:

- vector representations (embeddings) *[lectures 6, 7, +]*
- linguistic structure (e.g. parse) *[lectures 12-15]*
- “meaning” (e.g. AMR) *[lecture 19]*

# REPRESENTATIONS AND AMBIGUITY

There are myriad ways to express the same meaning, and there are immeasurable many meanings.

"Hello" — A greeting with an enquire about health or well-being

Mistress, what cheer?

How dost thou, sweet lord?

How, sweet Queen!

How do you do, pretty lady?

How fares my Kate?

Well be with you, gentlemen

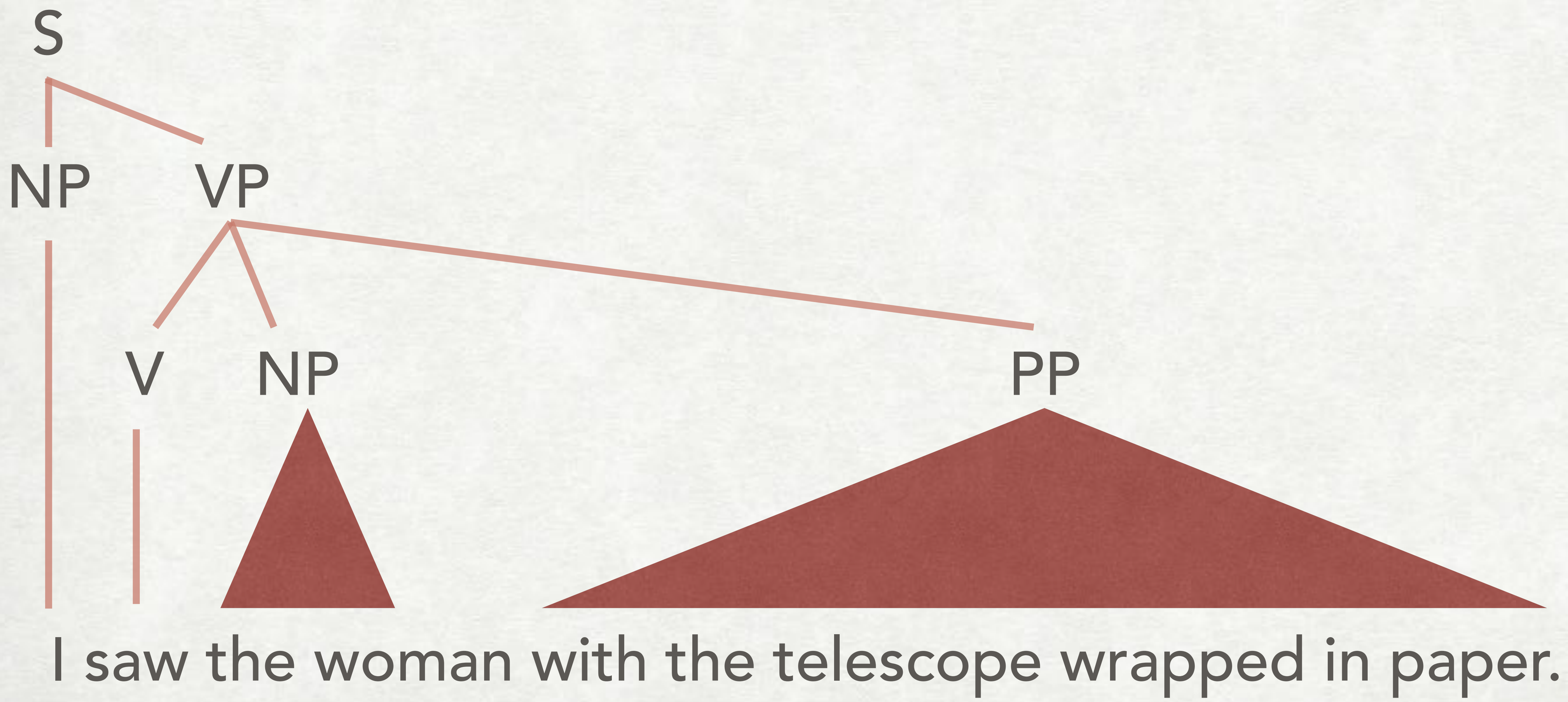
[source]

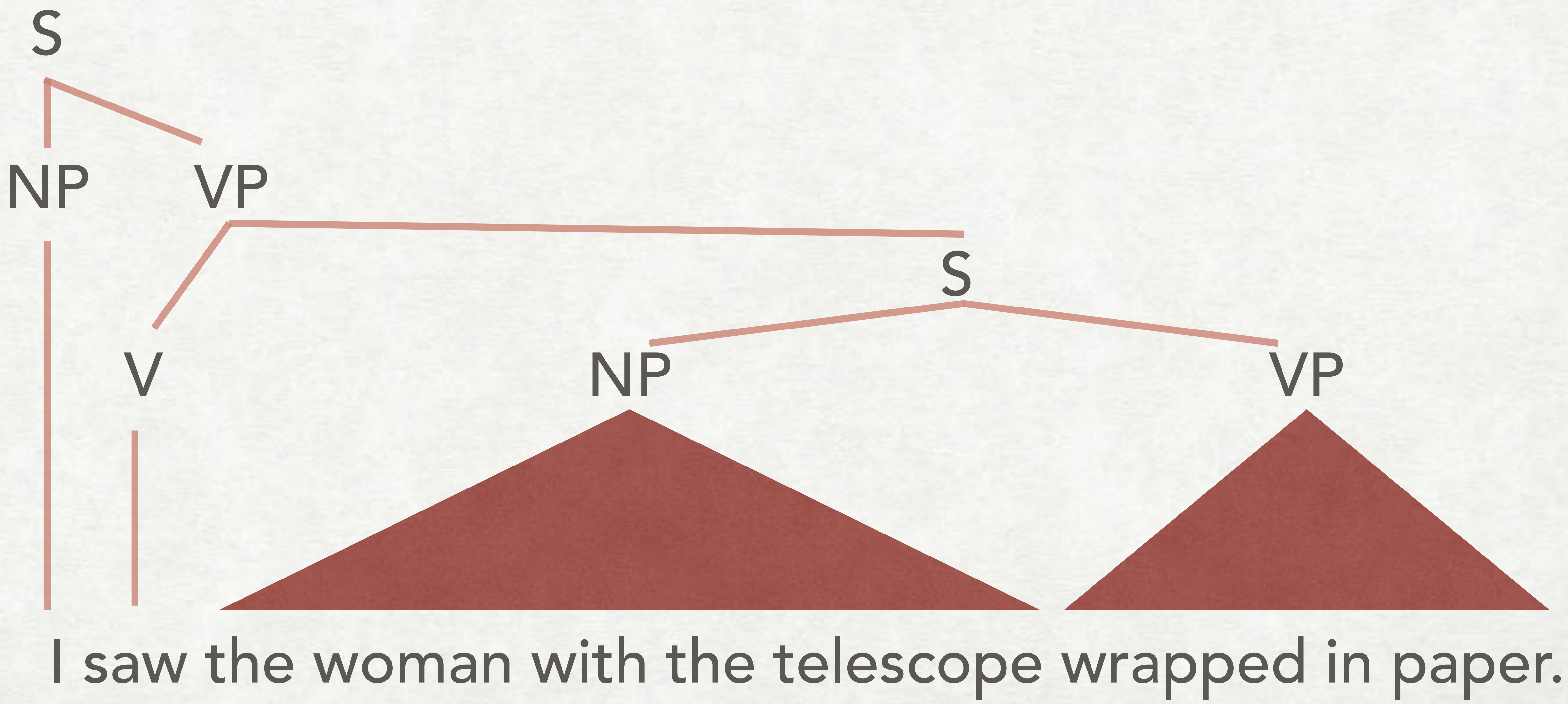
# REPRESENTATIONS AND AMBIGUITY

A string can have many possible interpretations in different contexts

I saw the woman with the telescope wrapped in paper.

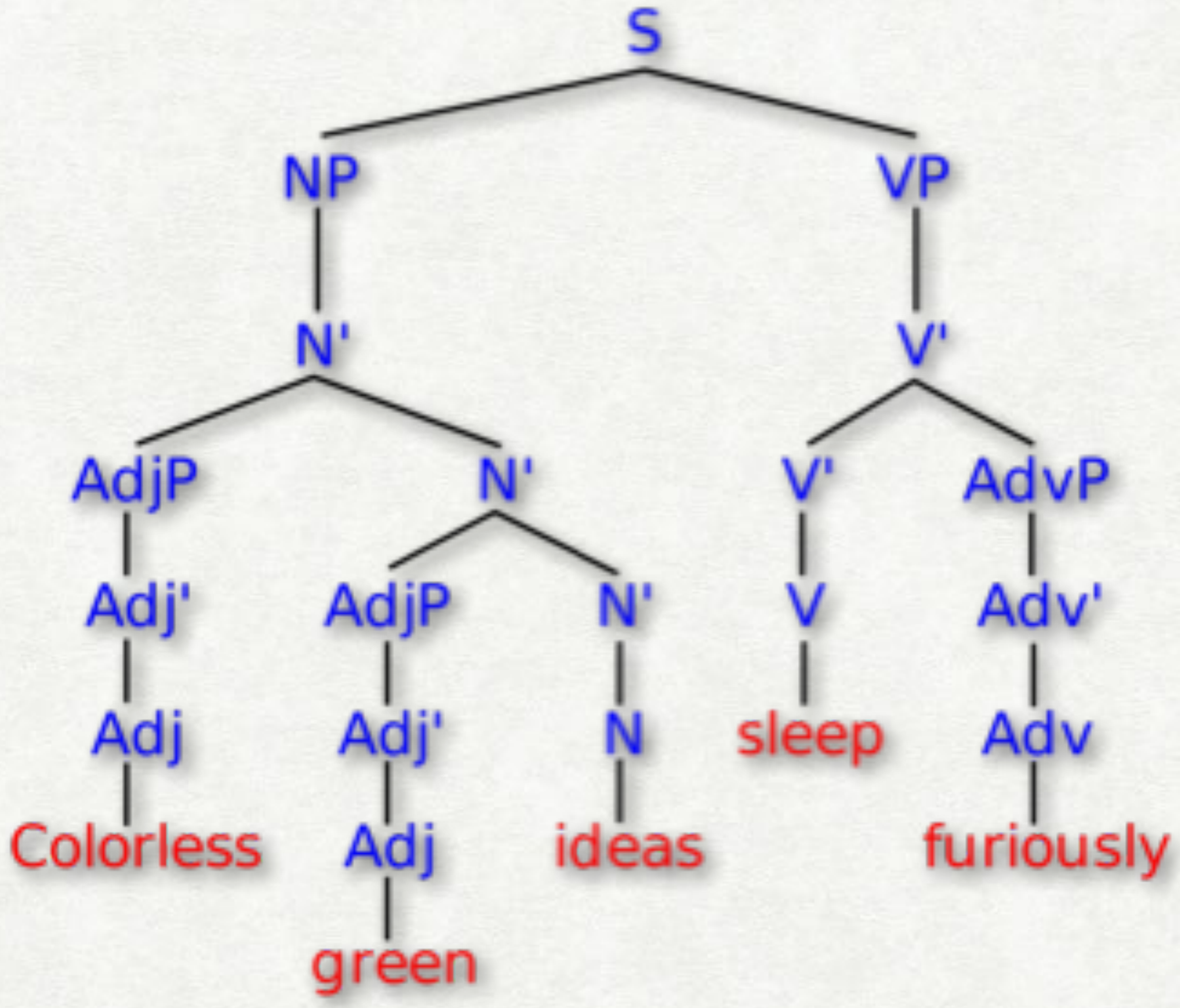
- Who has the telescope?
- Who/What is wrapped in paper?
- An event of perception or a questionable attempt at assault?





# SYNTAX VS SEMANTICS

Colorless green ideas sleep furiously



# NLP IS HARD!

- Natural language is complex!
    - Ambiguity
    - Linguistic Diversity
- Different tasks require different representations
- Any representation is a *theorized* construct (we do not observe it directly) that involves bias in the associated method.
- Many sources of variation and noise in linguistic input

# NLP VS COMPUTATIONAL LINGUISTICS

NLP focuses on the technology of processing language (to achieve a goal).

CL focuses on using technology to support/implement/supplement linguistics.

# NLP VS MACHINE LEARNING

NLP is not a subfield of machine learning!

Overlap: contemporary NLP uses a subset of ML methods.

- strings, unlike image or audio data, are discrete
- data are sequential ***and*** hierarchical

There exist some very useful and successful non-statistical techniques

- finite-state transducers for spell checking
- rule-based syntactic parsers

# MODELS

What is a model?

An abstract, theoretical, predictive construct.

- requires a (partial) representation of the world
- a method to create or recognize worlds
- a system for reasoning about worlds

This course will focus on formalisms and algorithms:
tools we can use to work with language data.
We'll also talk about state-of-the-art neural approaches.

# COURSE LOGISTICS

# LOGISTICS

Meeting times:

- Lectures: Tue, Thu 12-1:10

Main Reading

- Speech and Language Processing (2nd edition) — Yurafsky and Martin
  https://www.cs.colorado.edu/~martin/slp2.html
  Third edition (draft) is freely available here.

- Extra: Introduction to Natural Language Processing (Eisenstein)
  https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

Piazza: https://piazza.com/class/kkaenv2ty7x4tr

Website: https://cs.gmu.edu/~antonis/course/cs499-spring21/syllabus/

# GRADING

### Option 1

Homeworks (40%)

Group Project (30%)

Final Exam (30%)

### Option 2

Homeworks (50%)

Group Project (50%)

# HOMEWORK

Everything you submit must be your own work.

Any outside resources (books, research papers, websites, etc) or collaboration (students, professors) must be explicitly **acknowledged**.

Typically, a homework package will include a PDF with instructions and some data/code. You will have to submit a .zip file with a report and the code you wrote to create the answers.

- We WILL run your code on the data

https://cs.gmu.edu/~antonis/course/cs499-spring21/homework/

# PROJECT

Develop an application of NLP on a topic of interest to you.

You may work individually or in groups of two (each person should contribute equally)

Deliverables:

https://cs.gmu.edu/~antonis/course/cs499-spring21/project/

- Idea (up to 1 page)
- Baseline (up to 1 page + code)
- Presentation (slides + 5 minute YouTube video or in-class presentation)
- Final Report (2-4 pages per student)

[All .pdf files should use LaTeX and the ACL-style guide.]

# POLL TIME

Poll on neural network experience.

Poll on regular expressions.

Poll on programming languages.

Poll on LaTeX use.

Poll on exam or no-exam

# MORE READINGS

Finding a voice, Lane Green, *The Economist, 2017/05/01.*

AI’s Language Problem, Will Knight, *MIT Technology Review, 2016/08/09.*

# NEXT CLASS PREVIEW

Probability Preliminaries

Regular Expressions

Working with Text

Neural Network Basics