Natural Language Processing

11-411/11-611
Spring 2021
Alan W Black and David Mortensen

Introductory Lecture

What is NLP?

- Automating the analysis, generation, and acquisition of human (“natural”) language
  - **Analysis** (or “understanding” or “processing” …)
  - **Generation**
  - **Acquisition**

Note

- Some people use “NLP” to mean all of language technologies.
- Some people use it only to refer to *analysis*.

Why NLP? Web search!

- “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** with Google search!”
  - **Larry Page**, co-founder of **Google**
    - Google news release, 8/10/2015

Why NLP?

- Answer questions using the Web
- Translate documents from one language to another
- Do library research; summarize
- Manage messages intelligently
- Help make informed decisions
- Follow directions given by any user
- Fix your spelling or grammar
- Grade exams
- Write poems or novels
- Listen and give advice
- Estimate public opinion
- Read everything and make predictions
- Interactively help people learn
- Help disabled people
- Help refugees/disaster victims
- Document or reinvigorate indigenous languages

NLP Careers

- Industry
  - Educational technology
- Government
- Academia
- Humanitarian organizations

What about Ethics?

- Career choice isn’t just about money
  - Is what you are doing *bad* for humanity?
  - Is it good *enough* for humanity?
- Not just a question regarding government careers, or government funding, but…

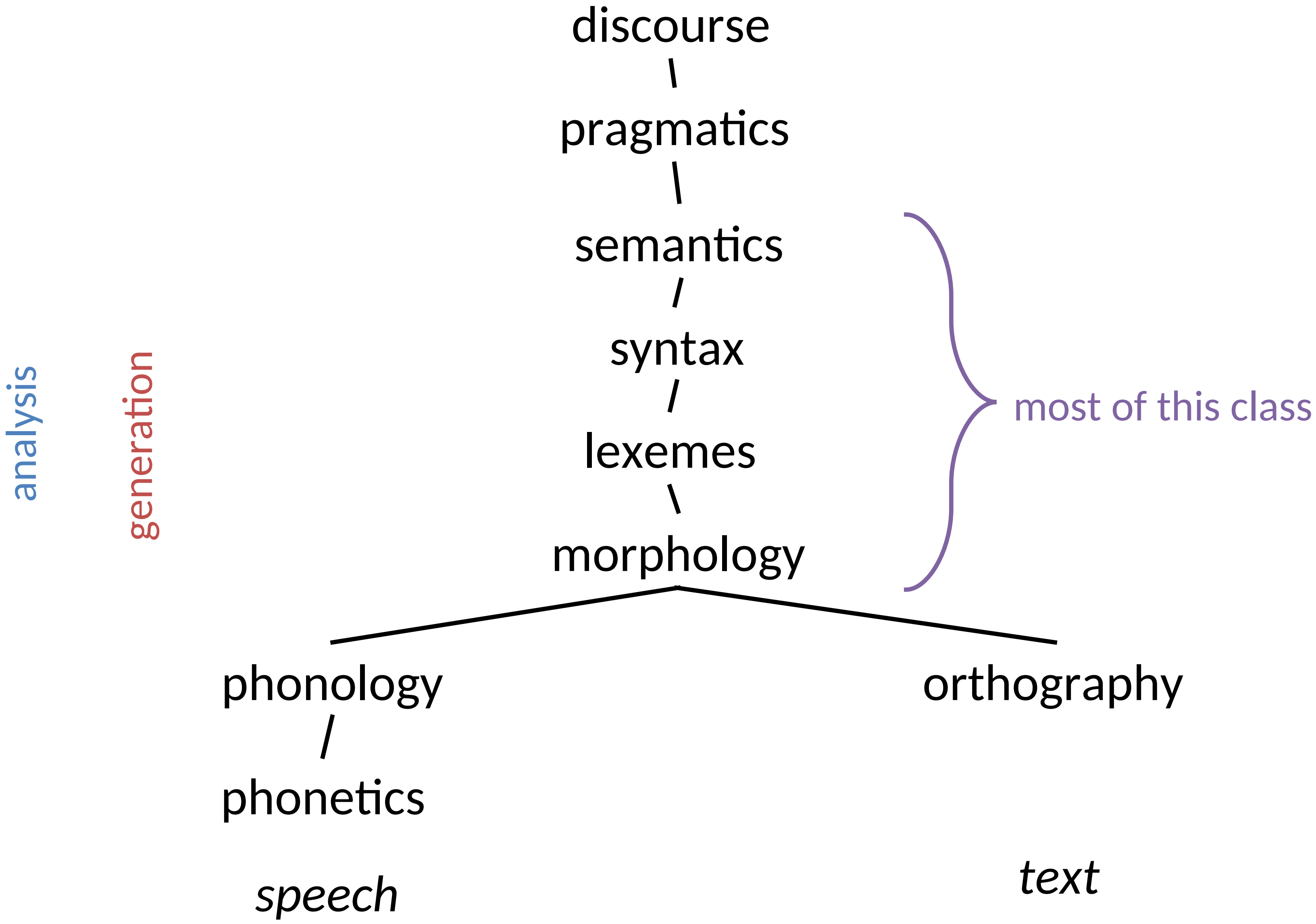
Work for the government?

Work for the government?

What is NLP? (more detail)

- Automating language analysis, generation, acquisition.
  - **Analysis** (or “understanding” or “processing” ...): input is language, output is some representation that supports useful action
  - **Generation**: input is that representation, output is language
  - **Acquisition**: obtaining the representation and necessary algorithms, from knowledge and data
- Representation?

Levels of Linguistic Representation



analysis

generation

discourse

pragmatics

semantics

syntax

lexemes

morphology

most of this class

phonology

phonetics

speech

orthography

text

Why It's Hard

1. The mappings between levels are extremely complex.
2. Appropriateness of a representation depends on the application.

Complexity of Linguistic Representations

- Input is likely to be noisy.
- Linguistic representations are *theorized* constructs; we cannot observe them directly.
- **Ambiguity**: each string may have many possible interpretations at every level. The correct resolution of the ambiguity will depend on the *intended meaning*, which is often inferable from context.
  - People are good at linguistic ambiguity resolution
  - Computers are not so good at it
    - How do we represent sets of possible alternatives?
    - How do we represent context?

Complexity of Linguistic Representations

- **Richness**: there are many ways to express the same meaning, and immeasurably many meanings to express. Lots of words/phrases.
- Each level *interacts* with the others.
- There is tremendous *diversity* in human languages.
  - Languages express the same kind of meaning in different ways
  - Some languages express some meanings more readily/often

We will study models

What is a Model?

- An abstract, theoretical, predictive construct. Includes:
  - a (partial) representation of the world
  - a method for creating or recognizing worlds
  - a system for reasoning about worlds
- NLP uses *many* tools for modeling.
- Surprisingly shallow models work fine for some applications.

Using NLP models/tools

- This course is meant to introduce some formal tools that will help you navigate the field of NLP.
- We focus on **formalisms** and **algorithms**.
  - This is not a comprehensive overview; it's a deep introduction to some key topics.
  - We'll focus mainly on *analysis* and mainly on *English text*.
  - The skills you develop will apply to any subfield of NLP

Applications: Challenges

- Application tasks evolve and are often hard to define formally.
- Objective evaluations of system performance are always up for debate
    - This holds for NL analysis as well as application tasks.
- Different applications may require different kinds of representations at different levels.

Key Applications in 2021

- Computational linguistics (i.e., modeling the human capacity for language computationally)
- Information extraction, especially “open” IE
- QA, chatbots (e.g., Siri, Alexa)
- Machine translation
- Summarization
- Opinion and sentiment analysis
- Social media analysis
- Fake News Recognition

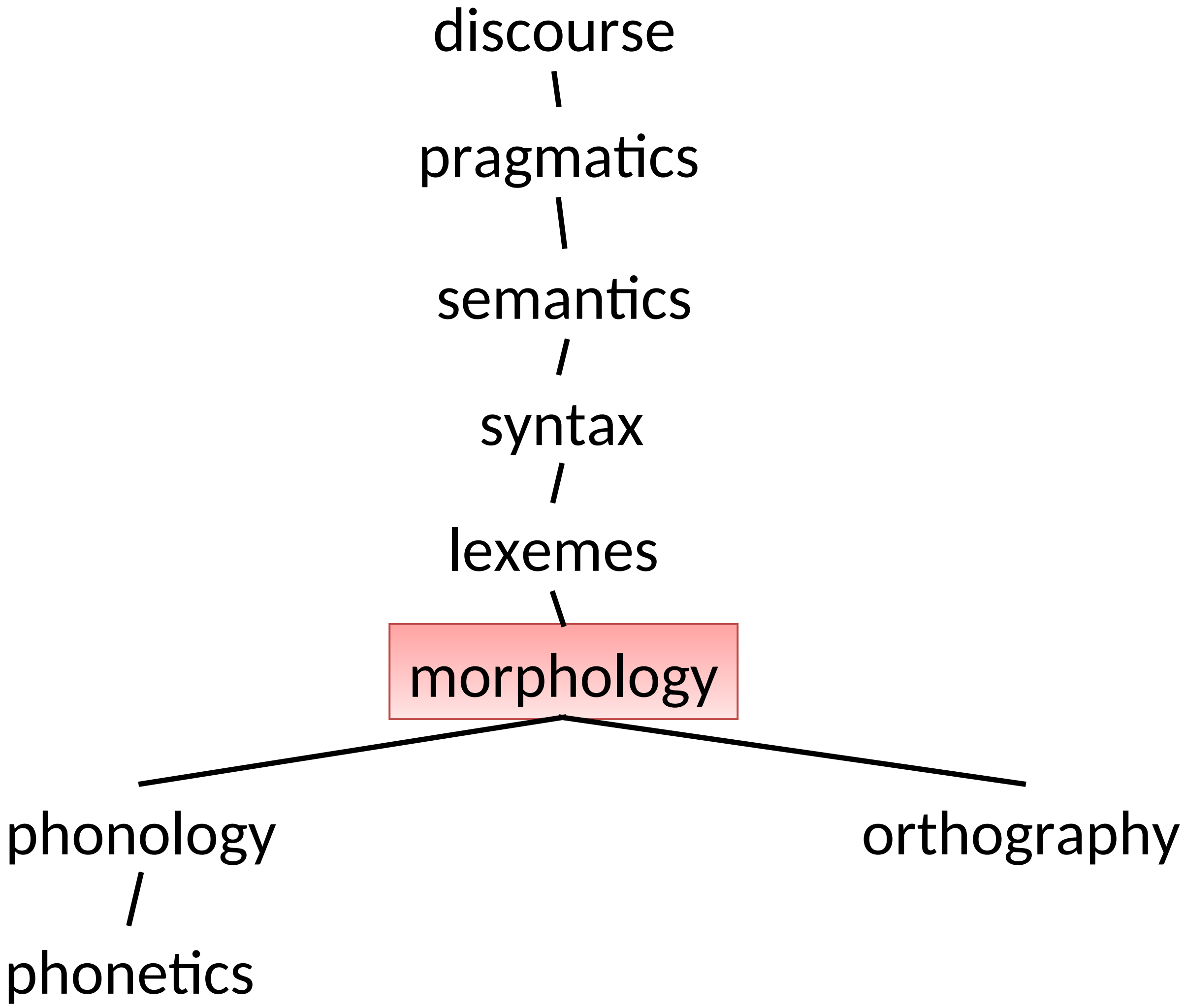
What about Brains?

“NLP” vs. “Computational Linguistics”

- “You have taken a beautiful living thing, killed it, and chopped it up into pieces.”
  - – paraphrase of student (different course)

- NLP is focused on the *technology* of processing language
- CL is focused on using technology to support/implement *linguistics*
- (Like “AI” vs. “cognitive science”)

Let's Examine Some of the Levels



discourse

pragmatics

semantics

syntax

lexemes

morphology

phonology

orthography

phonetics

Morphology

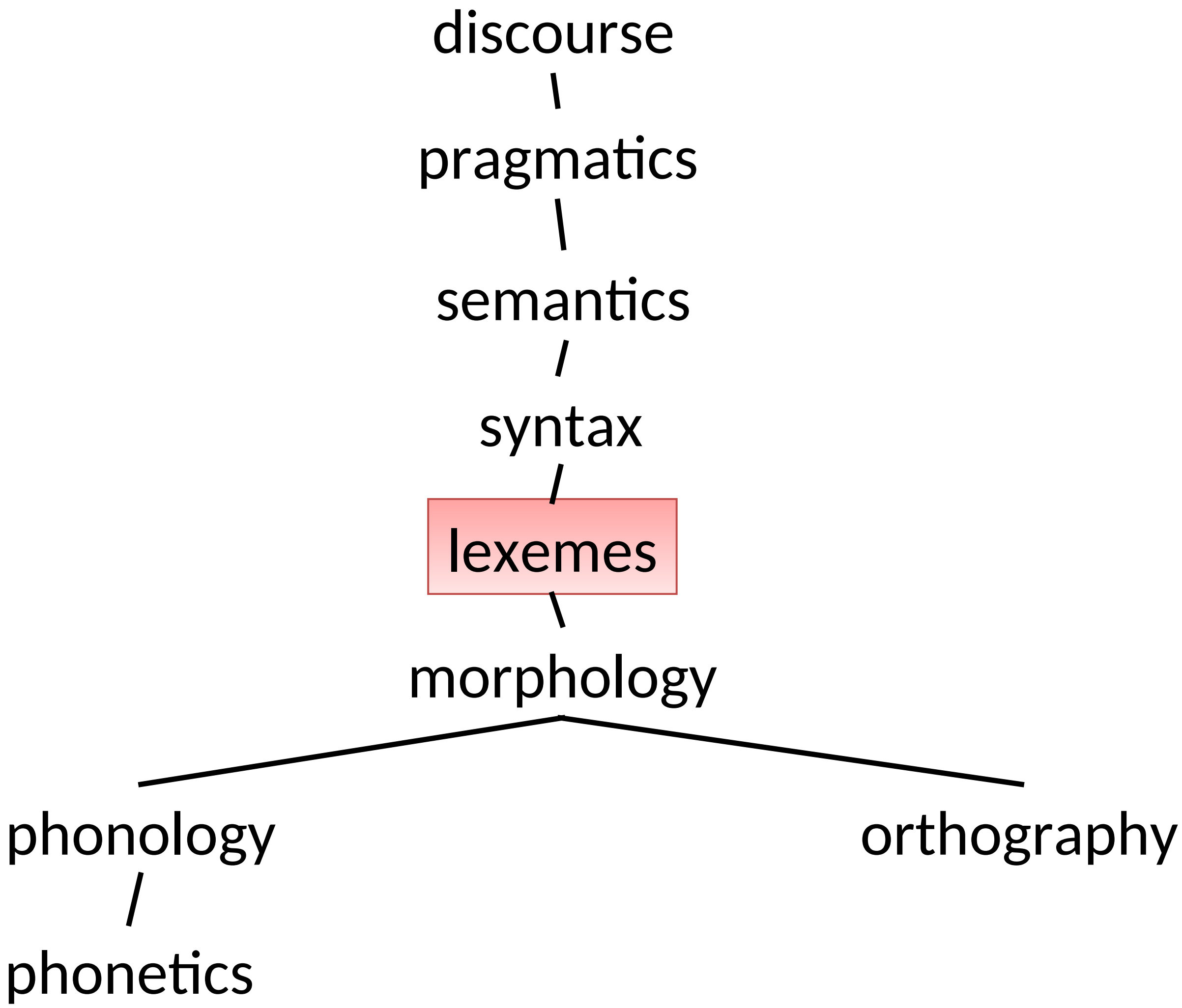
- Analysis of words into meaningful components
- Spectrum of complexity across languages
  - ***Analytic*** or ***Isolating*** languages (e.g., English, Chinese)
  - ***Synthetic*** languages (e.g., Finnish, Turkish, Hebrew)
- Examples

TIFGOSH ET HAYELED BAGAN
"you will meet the boy in the park"

Puedes dármelo
"You can give it to me"

uygarlaştıramadıklarımızdanmışsınızcasına
"(behaving) as if you are among those whom we could not civilize"

unfriend, Obamacare, Bill's



discourse

pragmatics

semantics

syntax

lexemes

morphology

phonology

orthography

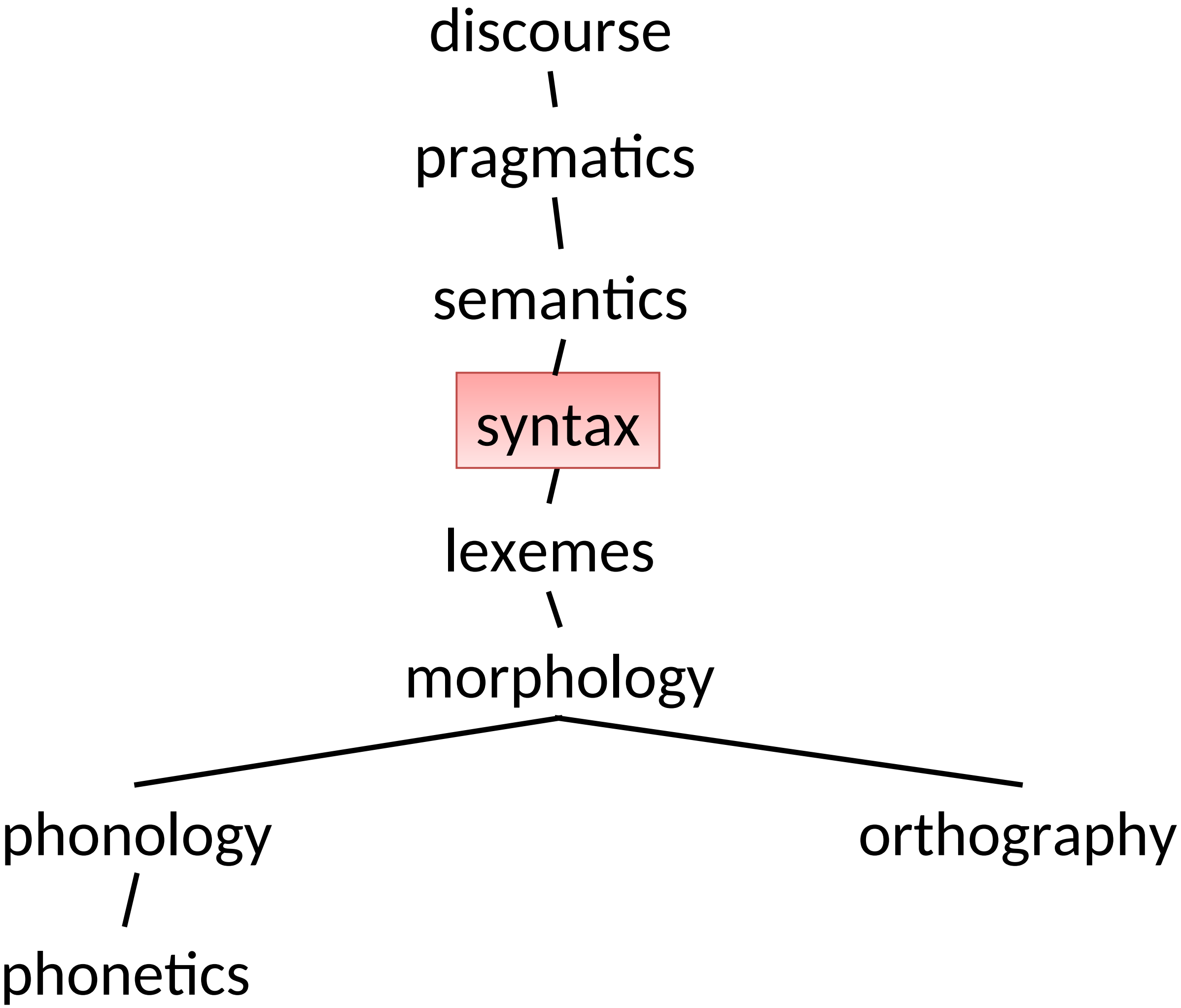
phonetics

第二阶段的奥运会体育比赛門票与残奥会开闭幕 式門票的预订工作已经结束，现在进入門票分配阶 段。在此期间，我们不再接受新的

Lexical Analysis

- Normalize and disambiguate words
- Words with multiple meanings: *bank*, *mean*
  - Extra challenge: domain-specific meanings
- Multi-word expressions

  make … decision, *take out*, *make up*, …
- For English, part-of-speech tagging is one very common kind of lexical analysis
  - Others: supersense tagging, various forms of word sense disambiguation, syntactic "supertags," …



discourse

pragmatics

semantics

syntax

lexemes

morphology

phonology

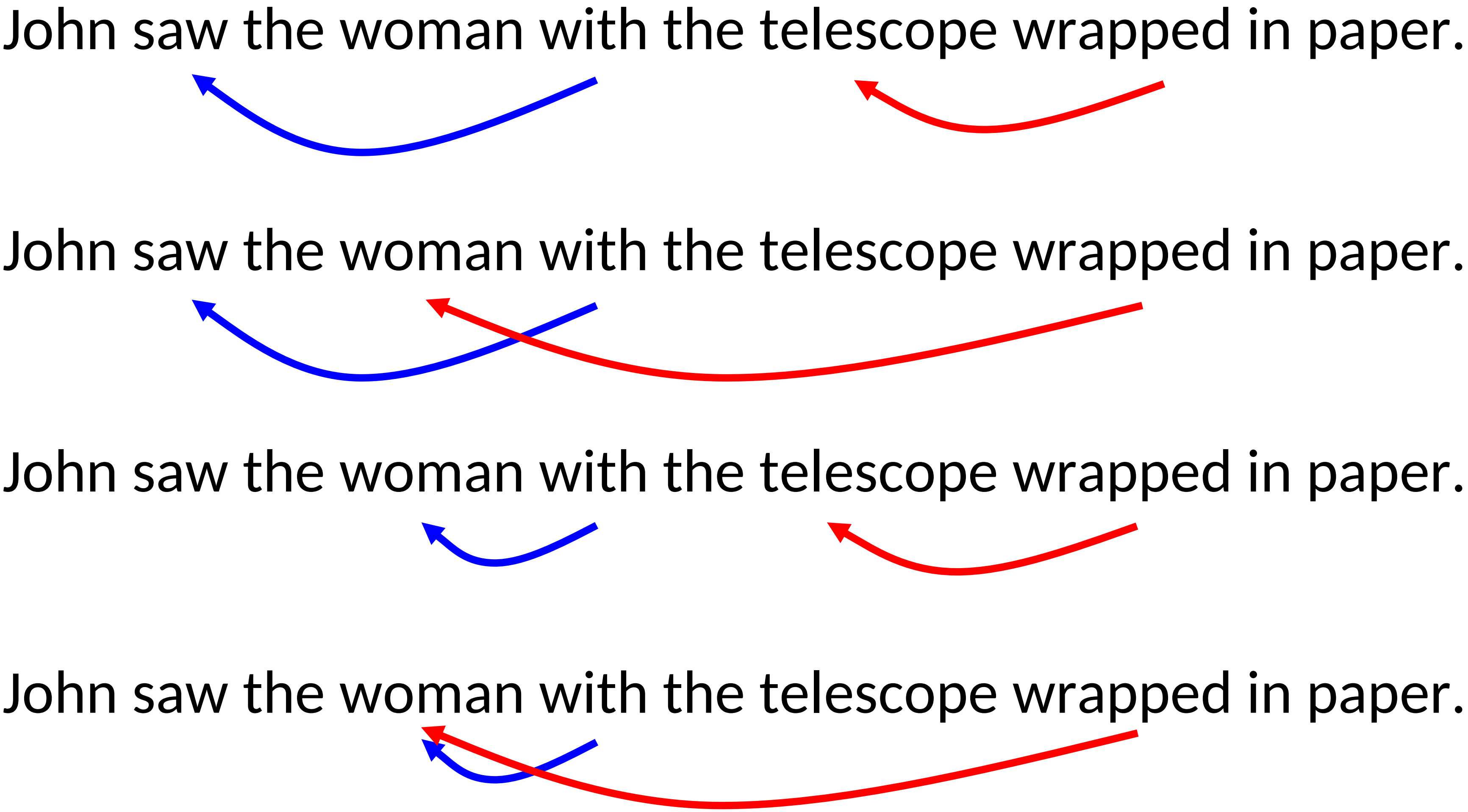
orthography

phonetics

Syntax

- Transform a sequence of symbols into a hierarchical or compositional structure.
- Closely related to linguistic theories about what makes some sentences well-formed and others not. For example:
  - ✓ I want a flight to Tokyo
  - ✓ I want to fly to Tokyo
  - ✓ I found a flight to Tokyo
  - ✷ I found to fly to Tokyo
- Ambiguities explode combinatorially
- Simple examples:
  Students hate annoying professors.
  John saw the woman with the telescope.
  John saw the woman with the telescope wrapped in paper.

Some of the Possible Syntactic Analyses

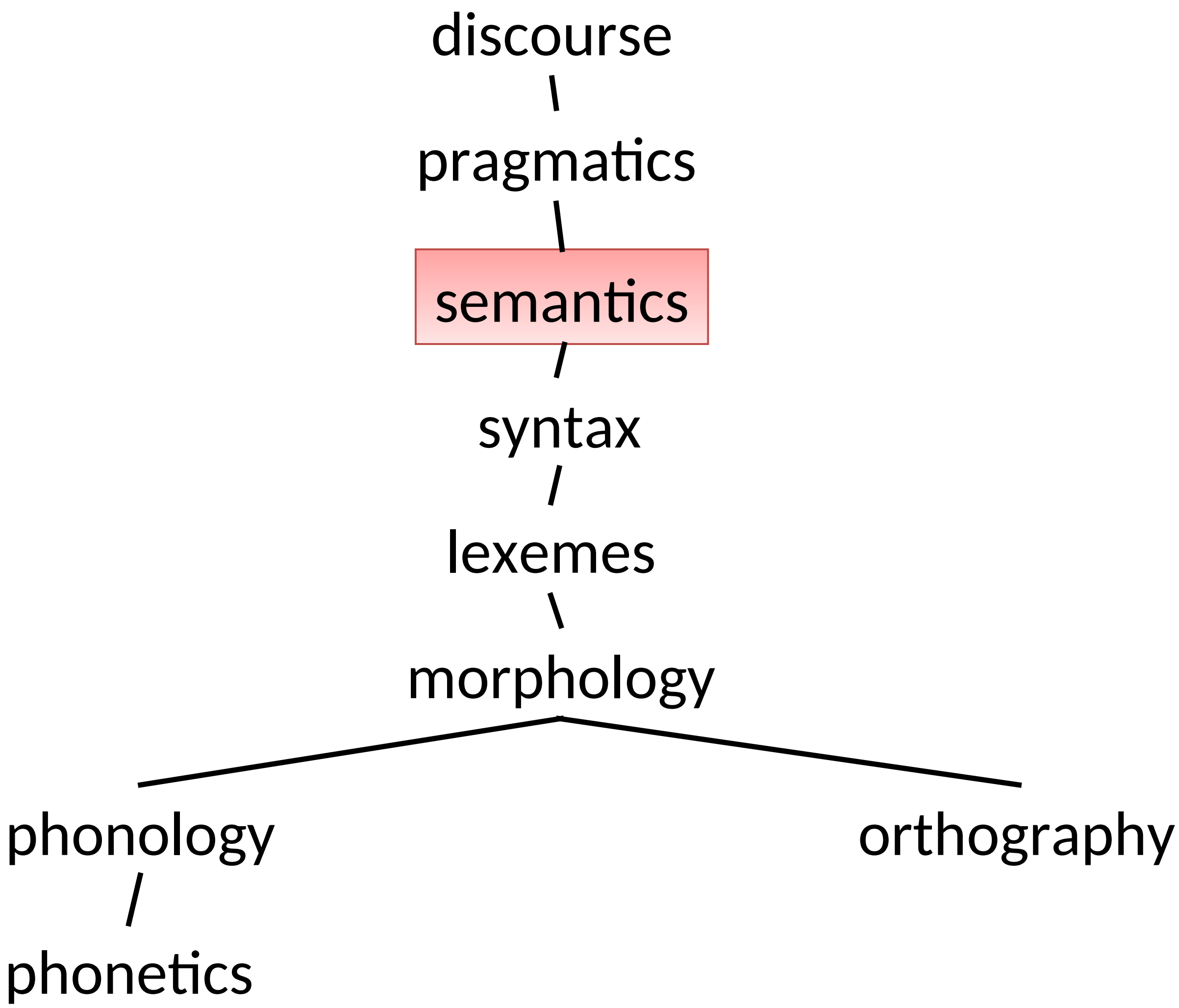


John saw the woman with the telescope wrapped in paper.

John saw the woman with the telescope wrapped in paper.

John saw the woman with the telescope wrapped in paper.

John saw the woman with the telescope wrapped in paper.



discourse

pragmatics

semantics

syntax

lexemes

morphology

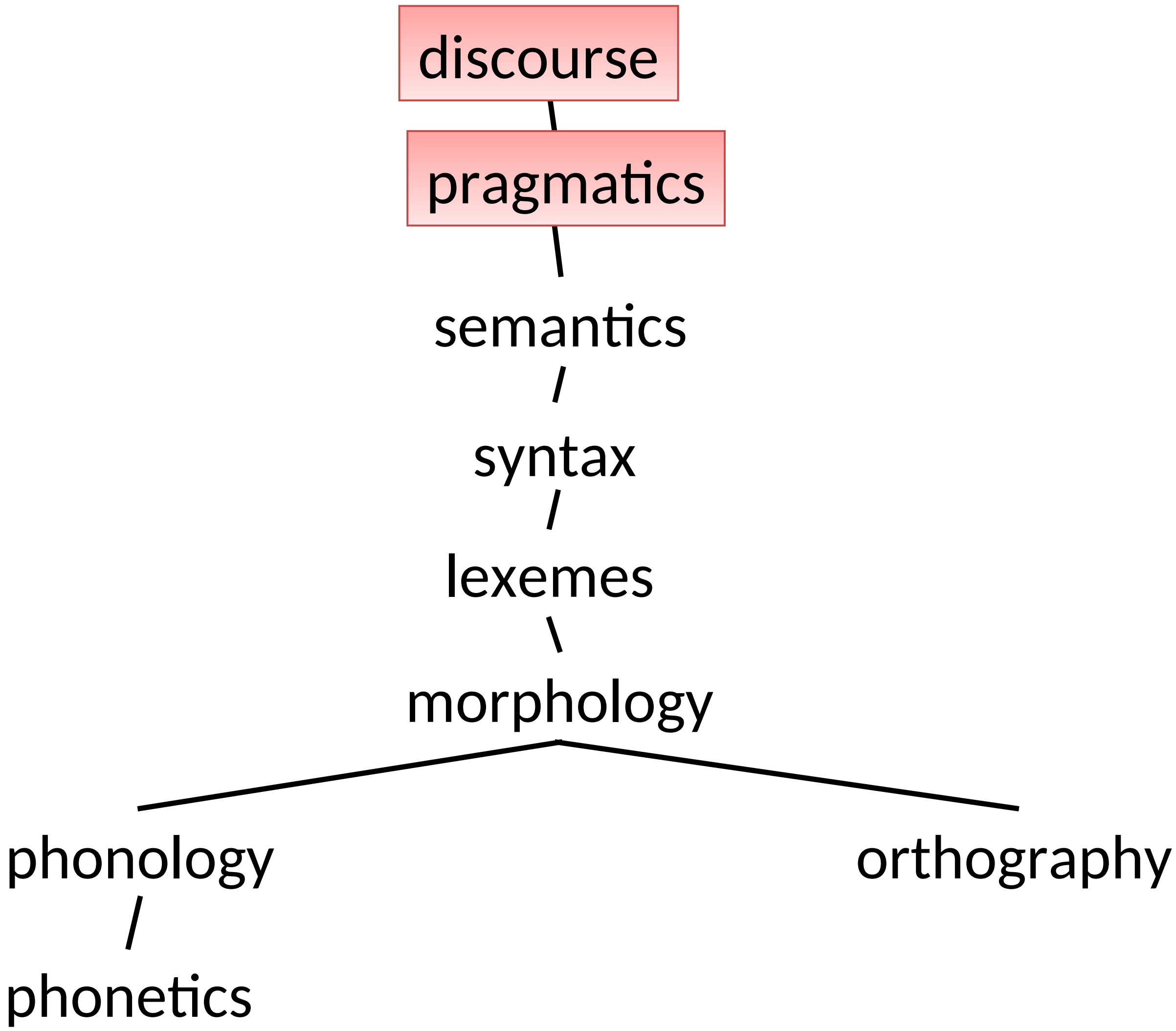
phonology

orthography

phonetics

Semantics

- Mapping of natural language sentences into domain representations.
  - E.g., a robot command language, a database query, or an expression in a formal logic.
- Scope ambiguities:
  - A seat is available to every customer
  - A web site is available to every customer
- Going beyond specific domains is a goal of Artificial Intelligence



discourse

pragmatics

semantics

syntax

lexemes

morphology

phonology

orthography

phonetics

Pragmatics, Discourse

- Pragmatics
  - Any *non-local* meaning phenomena

    “Can you pass the salt?”

    “Is he 21?” “Yes, he’s 25.”
- Discourse
  - Structures and effects in related sequences of sentences
  - Texts, dialogues, multi-party conversations

    “I said the ***black*** shoes.”

    “Oh, ***black***.” (Is that a sentence?)

Administrivia

- Web page: http://demo.clab.cs.cmu.edu/NLP
- Book: *Speech and Language Processing*, Jurafsky and Martin, 2nd ed. (plus)
- Instructors: Alan W Black, David Mortensen

Piazza for questions, announcements

Canvas for assignment submission

Other Policies

- Waitlist
- Readings
- Homework
  - Everything you submit must be your own work
  - Any outside resources (books, research papers, web sites, etc.) or collaboration (students, professors, etc.) must be explicitly **acknowledged**
- Project
  - Collaboration is **required** (teams of 4)
  - It's okay to use existing tools, but you must acknowledge them
  - Grade is mostly shared
  - Programming language is up to you, but the project must run gracefully on our server.
- Do people know Python?