Algorithms for NLP (11-711)
Fall 2015

Introductory Lecture

Motivating the Course

What is NLP?

- 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?

Note

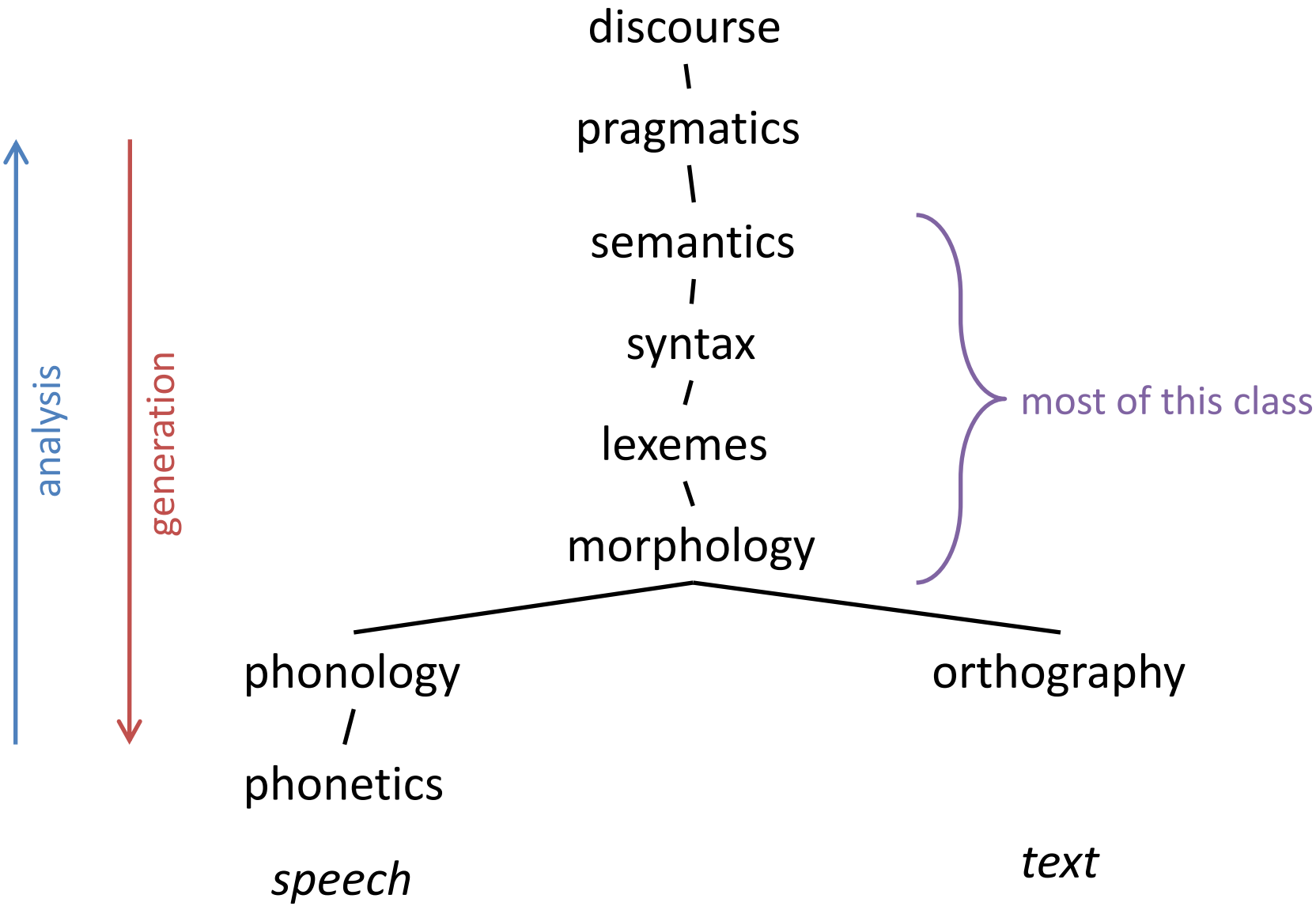
- Some people use “NLP” to mean all of language technologies.
- Some people use it only to refer to *analysis*.

Note 2

- “NLP” vs. “Computational Linguistics”

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

Levels of Linguistic Representation



Why It's Hard

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

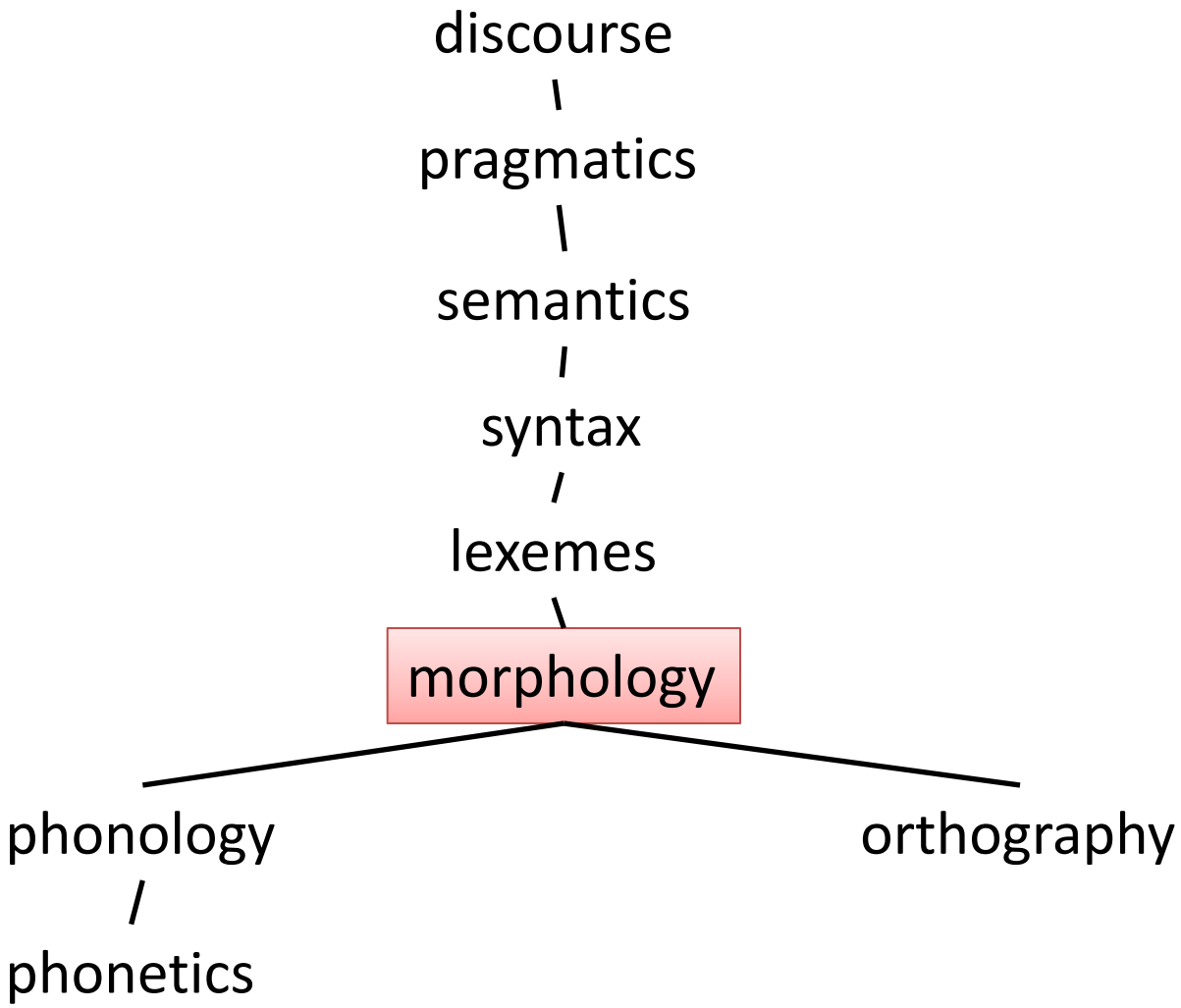
Complexity of Linguistic Representations

- 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.
- 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

Let's Examine Some of the Levels


discourse
pragmatics
semantics
syntax
lexemes
morphology
phonology
phonetics
orthography

Morphology

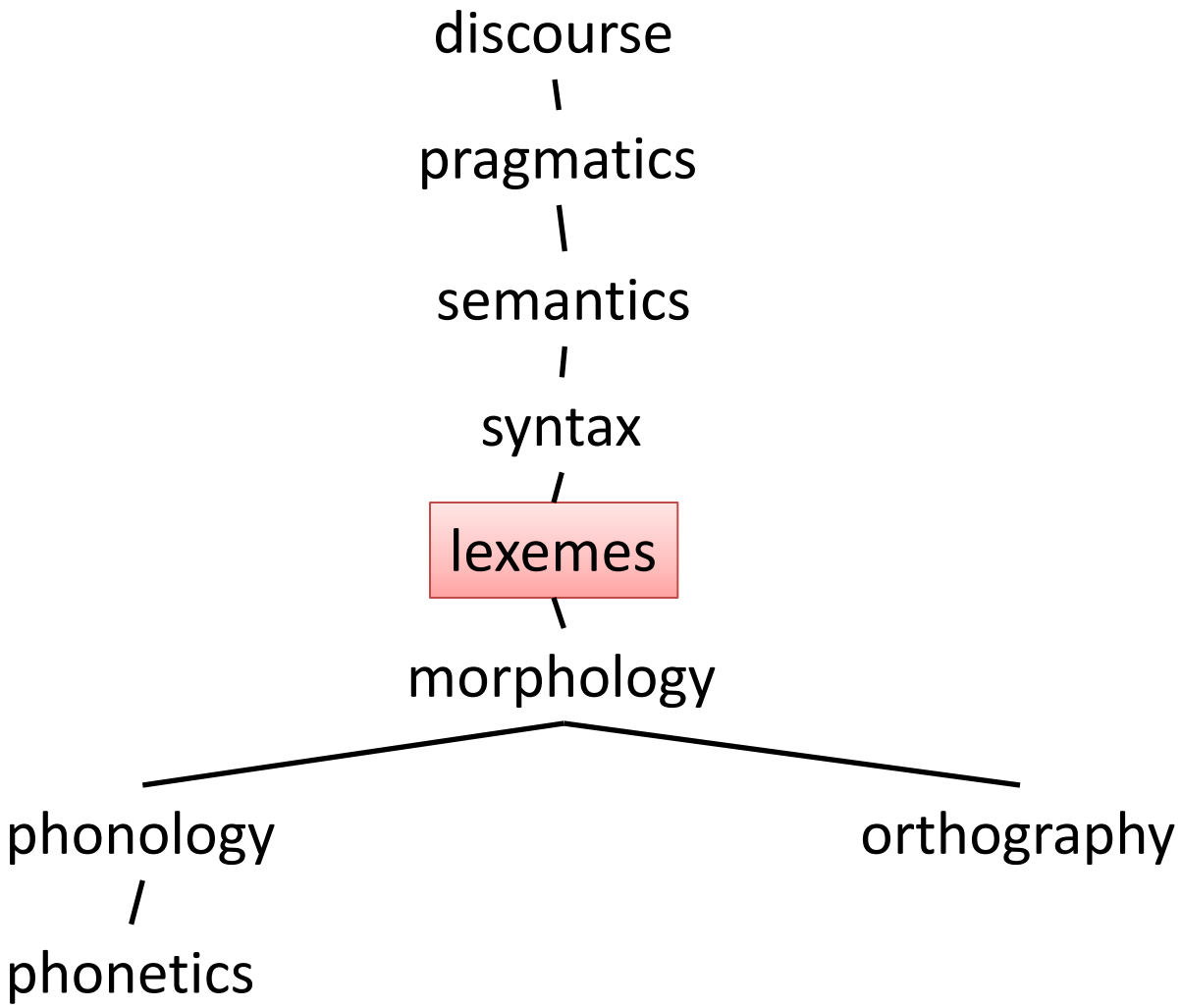
- Analysis of words into meaningful components
- Spectrum of complexity across languages
  - ***Analytic*** or ***Isolating*** languages (e.g., English, Chinese)
  - ***Synthetic*** languages (e.g., Finnish, Turkish, Hebrew)
- Examples

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"

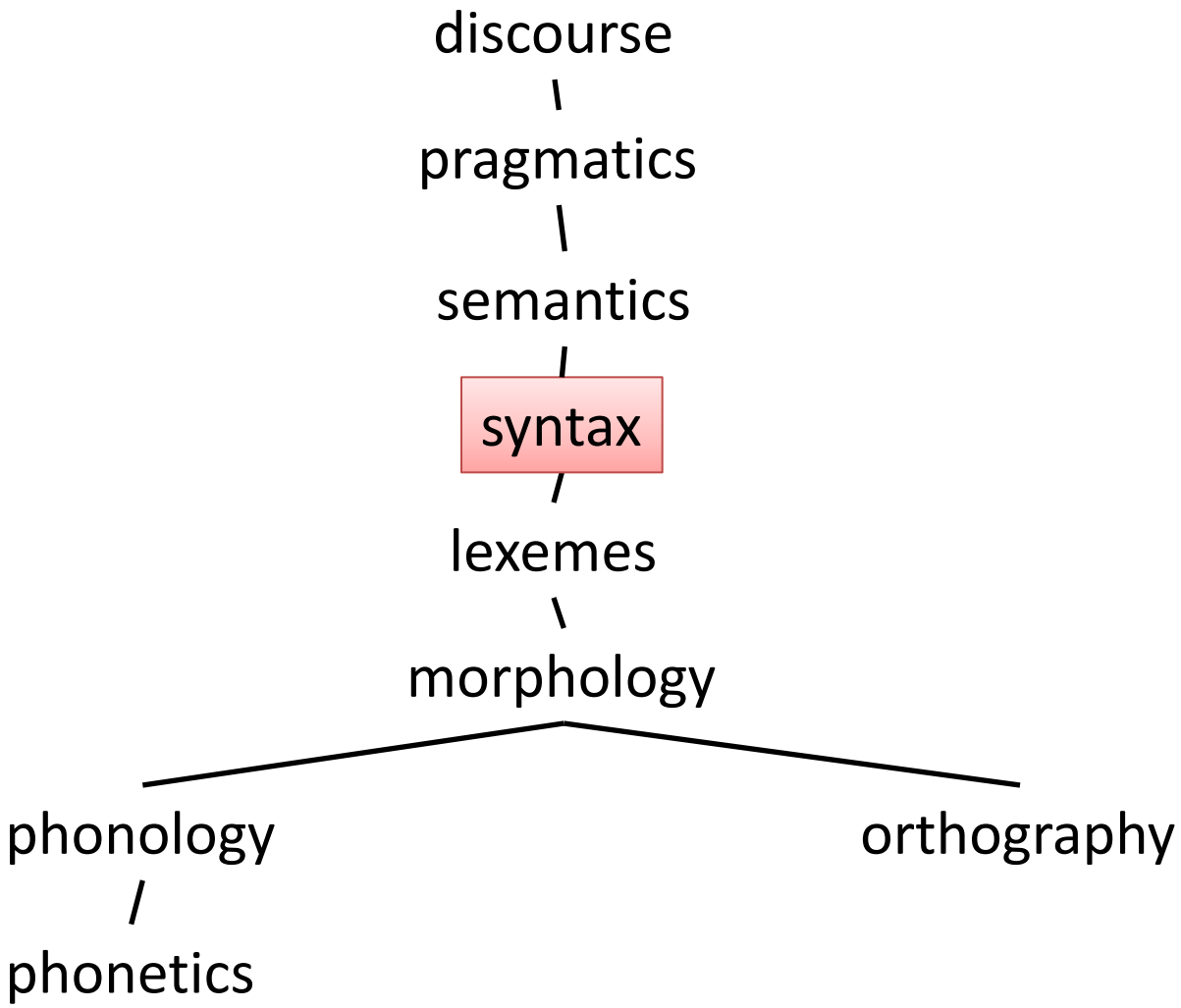
unfriend, Obamacare, Bill's


discourse
pragmatics
semantics
syntax
lexemes
morphology
phonology
phonetics
orthography

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
phonetics
orthography

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.

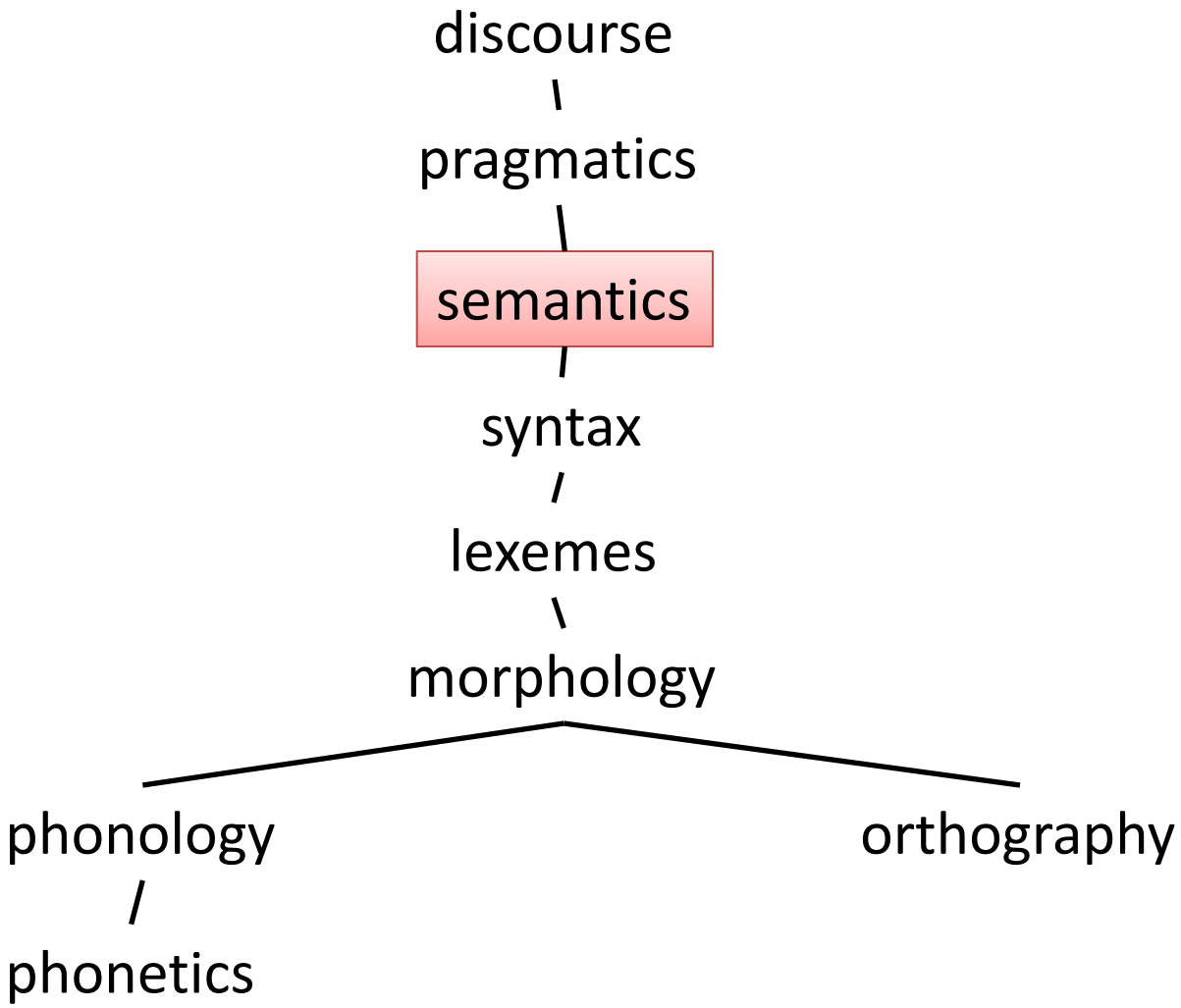
Some of the Possible Syntactic Analyses

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.

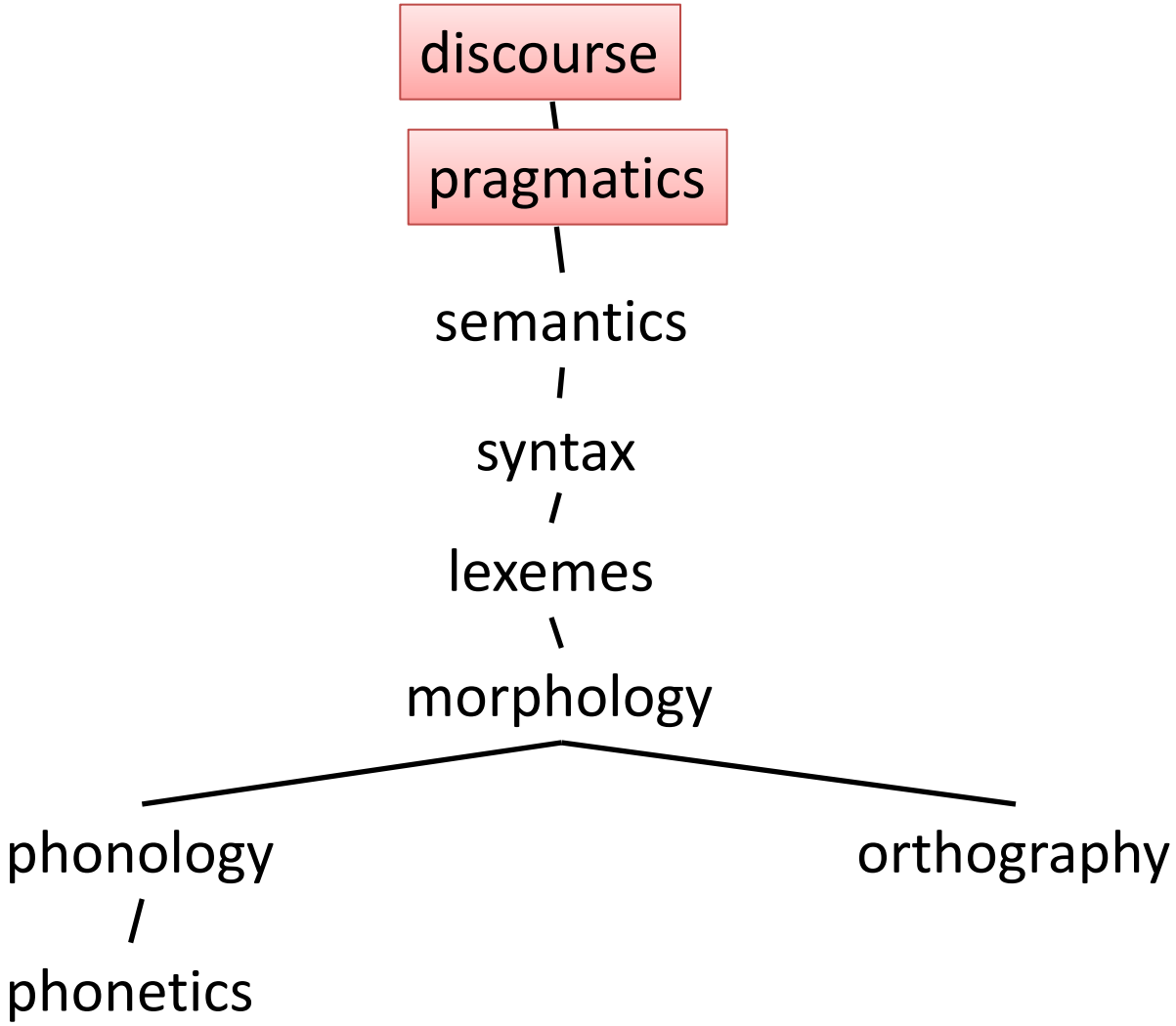
John saw the woman with the telescope wrapped in paper.


discourse
pragmatics
semantics
syntax
lexemes
morphology
phonology
phonetics
orthography

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:

  In this country a woman gives birth every fifteen minutes. –Groucho
- Going beyond specific domains is a goal of Artificial Intelligence


discourse
pragmatics
semantics
syntax
lexemes
morphology
phonology
phonetics
orthography

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?)

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 2015

- Computational linguistics (i.e., modeling the human capacity for language computationally)
- Information extraction, especially “open” IE
- Question answering (e.g., Watson, Siri)
- Machine translation
- Summarization
- Opinion and sentiment analysis
- Social media analysis

Course Scope

- 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*.
  - The skills you develop will apply to any subfield of NLP

Course Objectives

Algorithms for NLP is an introductory graduate-level course on the computational properties of natural languages and the fundamental algorithms for processing natural languages.

Objectives:

1. Develop a thorough understanding of the principles and formal methods used in the design and analysis of language processing algorithms.
2. Provide an in-depth presentation of the major algorithms used in NLP, including lexical, morphological, syntactic, and semantic analysis, with the primary focus on parsing algorithms and their analysis.

Introductions

Chris Dyer

Administrivia

Basic Information

- Instructors: (Chris Dyer, 5707 | Bob Frederking, 6515 | Miguel Ballesteros, 5413)
  - Office hours: by appointment
- TAs: (TBA1 | TBA2); Office hours: TBA
- Lecture: Tuesday and Thursday 1:30-2:50, GHC4307
- Recitation: Friday 1:30-2:20, DH2302
  - *Not* this week!
- http://demo.clab.cs.cmu.edu/fa2015-11711

What We're Going to Cover

1. Finite-state NLP
   - Formal (regular) language theory (5)
   - Finite-state methods in NLP (5)
2. Context-free NLP
   - Formal (context-free) language theory (2)
   - Parsing algorithms (4)
   - Dynamic programming and search (3)
3. Context-sensitive NLP and Semantics
   - Context-sensitive formalisms (2)
   - Semantic problems and representations (3)
4. Current NLP challenges and research (2)

Formal Background

1. Finite-state NLP
   - Formal (regular) language theory (5)
   - Finite-state methods in NLP (5)
2. Context-free NLP
   - Formal (context-free) language theory (2)
   - Parsing algorithms (4)
   - Dynamic programming and search (3)
3. Context-sensitive NLP and Semantics
   - Context-sensitive formalisms (2)
   - Semantic problems and representations (3)
4. Current NLP challenges and research (2)

Practical NLP Techniques

1. Finite-state NLP
   - Formal (regular) language theory (5)
   - Finite-state methods in NLP (5)
2. Context-free NLP
   - Formal (context-free) language theory (2)
   - Parsing algorithms (4)
   - Dynamic programming and search (3)
3. Context-sensitive NLP and Semantics
   - Context-sensitive formalisms (2)
   - Semantic problems and representations (3)
4. Current NLP challenges and research (2)

Course Philosophy

NLP is a very large field!

We aim to strike a balance between theory and practice, and between classic foundations and current applications.

But mind the gap.

Prerequisites and Corequisites

- Exposure to syntax and structure of natural language (or at least English)
- College-level course on algorithms
- College-level programming skills

The NLP Lab (11-712, offered in the spring) complements this course with further programming exercises.

Format

- Most material will come in the lectures.
  - Readings associated with each lecture will be found on the web page.
- About five assignments (35% of the grade), each taking about two weeks.
- Two exams: midterm (25%) and final (40%).

Books and Readings

- John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman, *Introduction to Automata Theory, Languages and Computation*. 2000, 2nd edition, chapters 1-7.
- Daniel Jurafsky and James H. Martin, *Speech and Language Processing*. 2008, 2nd edition, selected chapters.
- Noah A. Smith, *Linguistic Structure Prediction*. 2011, chapter 2. (Electronic version is available free through the CMU library.)
- Others as needed.

Electronic Communication

http://demo.clab.cs.cmu.edu/fa2015-11711

- Schedule, assignments, readings, lecture slides, additional handouts.
- Email the instructors:
  11711-fall15-instructors@lists.andrew.cmu.edu
- Subscribe to the course email list!

  http://lists.andrew.cmu.edu/11711-fall15

Electronic Communication

- Piazza?
  - Work it out with the TAs

Academic Integrity

- Please read the cheating policy carefully.
  - Sign the second page and turn it in.

Key things to remember:

- By default, all work must be done individually.
- Don't copy anyone else's work:
  - ***Includes previous years' solutions***
  - Includes materials from other courses (at CMU or elsewhere)
  - Includes publicly available materials
- Cite sources!
- Not sure? Ask instructors!

Academic Integrity

Severe actions will be taken against students that violate the policy, possibly resulting in course failure or dismissal from the program.