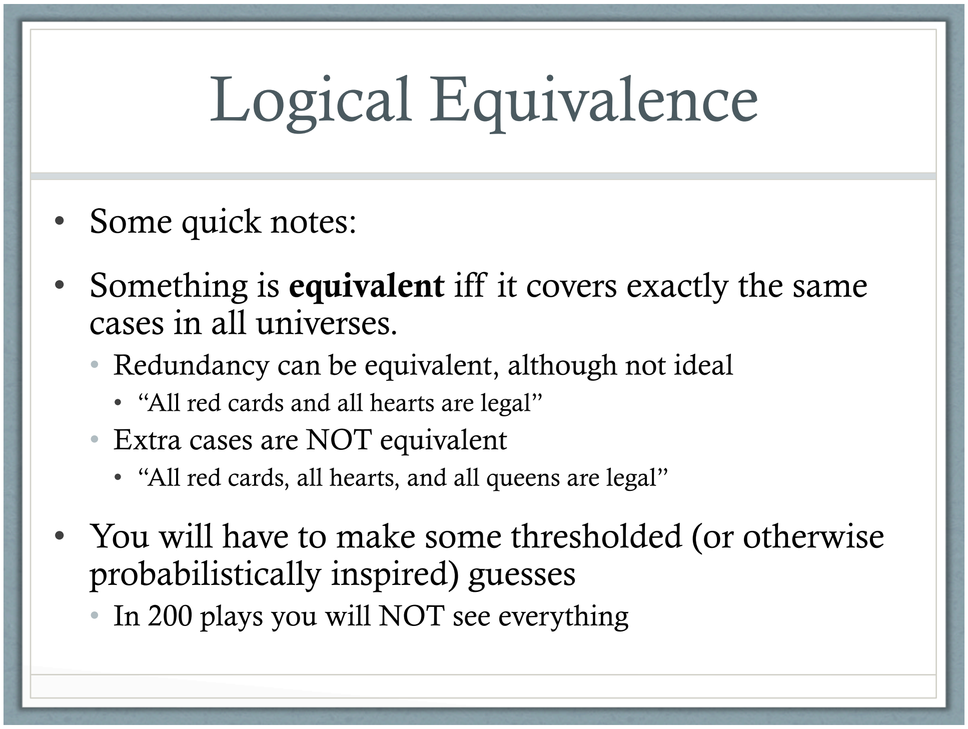
Natural Language from 20,000 feet

AI Class 28 (no reading)

Slides from Paula Matuszek and Mary-Angela Papalaskari, Villanova University, with thanks

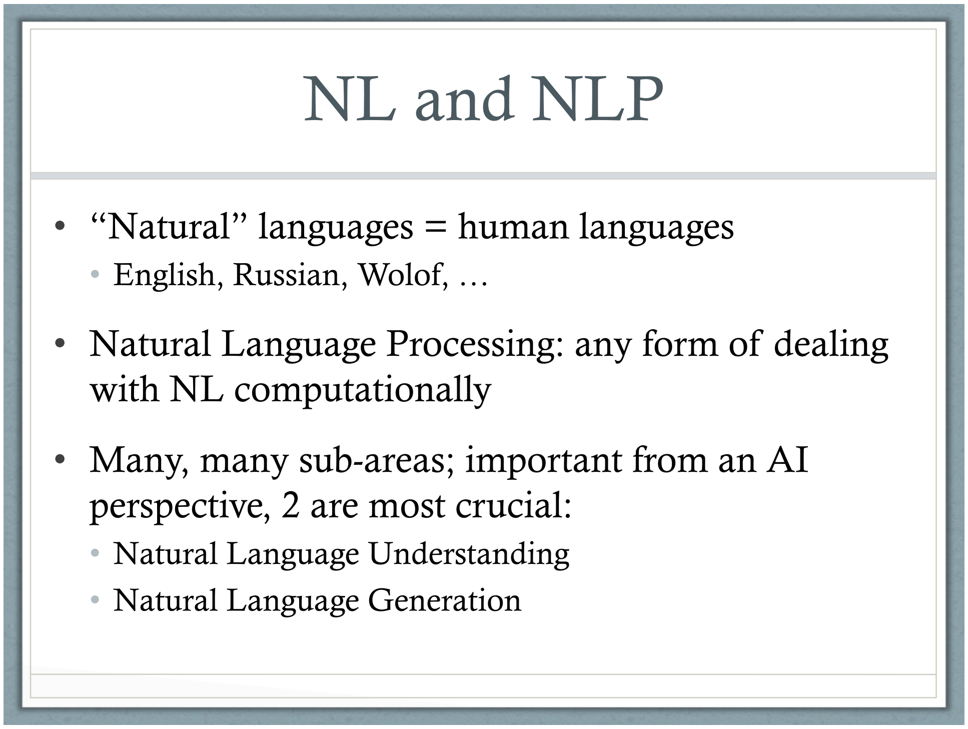
Bookkeeping

- Review Session: Friday, 12/16, 6pm-8pm
  - If you can’t make that time, see posted slides
- End of class: SECQ



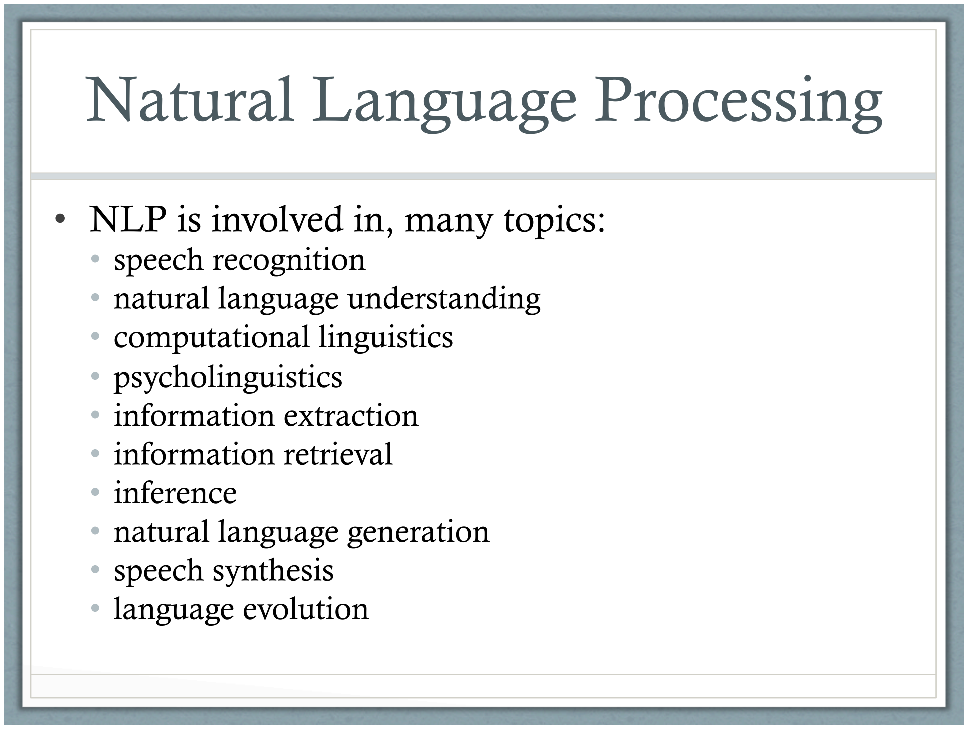
Logical Equivalence

- Some quick notes:
- Something is **equivalent** iff it covers exactly the same cases in all universes.
  - Redundancy can be equivalent, although not ideal
    - "All red cards and all hearts are legal"
  - Extra cases are NOT equivalent
    - "All red cards, all hearts, and all queens are legal"
- You will have to make some thresholded (or otherwise probabilistically inspired) guesses
  - In 200 plays you will NOT see everything



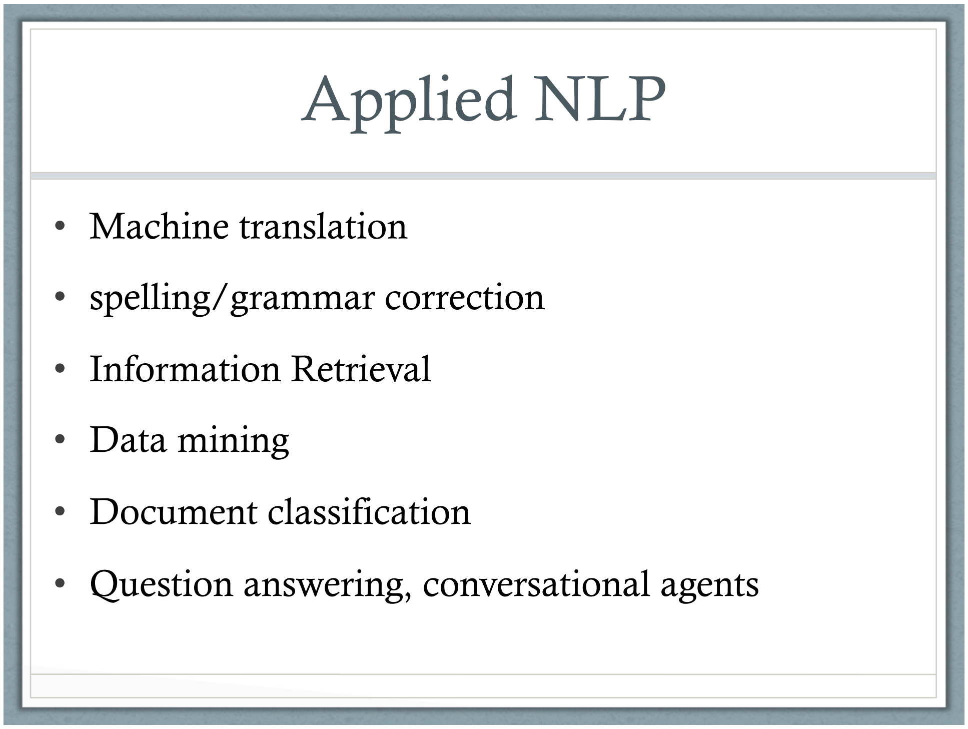
NL and NLP

- "Natural" languages = human languages
  - English, Russian, Wolof, …
- Natural Language Processing: any form of dealing with NL computationally
- Many, many sub-areas; important from an AI perspective, 2 are most crucial:
  - Natural Language Understanding
  - Natural Language Generation



Natural Language Processing

- NLP is involved in, many topics:
  - speech recognition
  - natural language understanding
  - computational linguistics
  - psycholinguistics
  - information extraction
  - information retrieval
  - inference
  - natural language generation
  - speech synthesis
  - language evolution



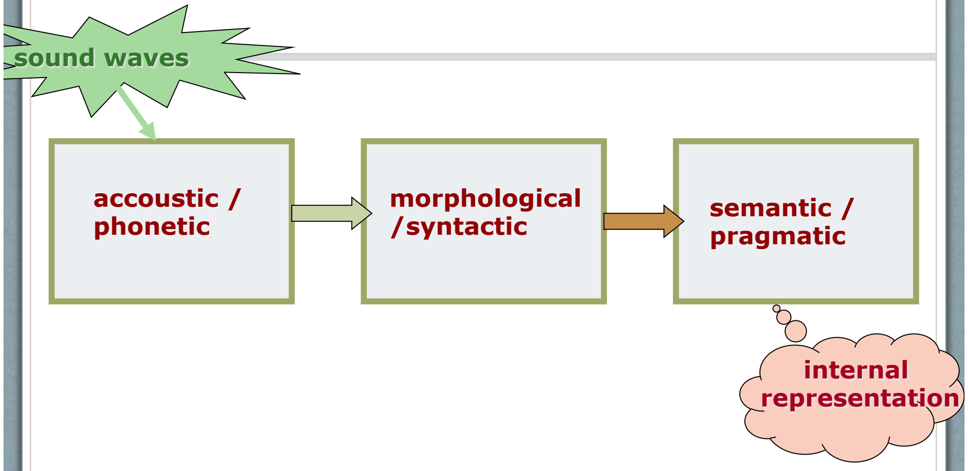
Applied NLP

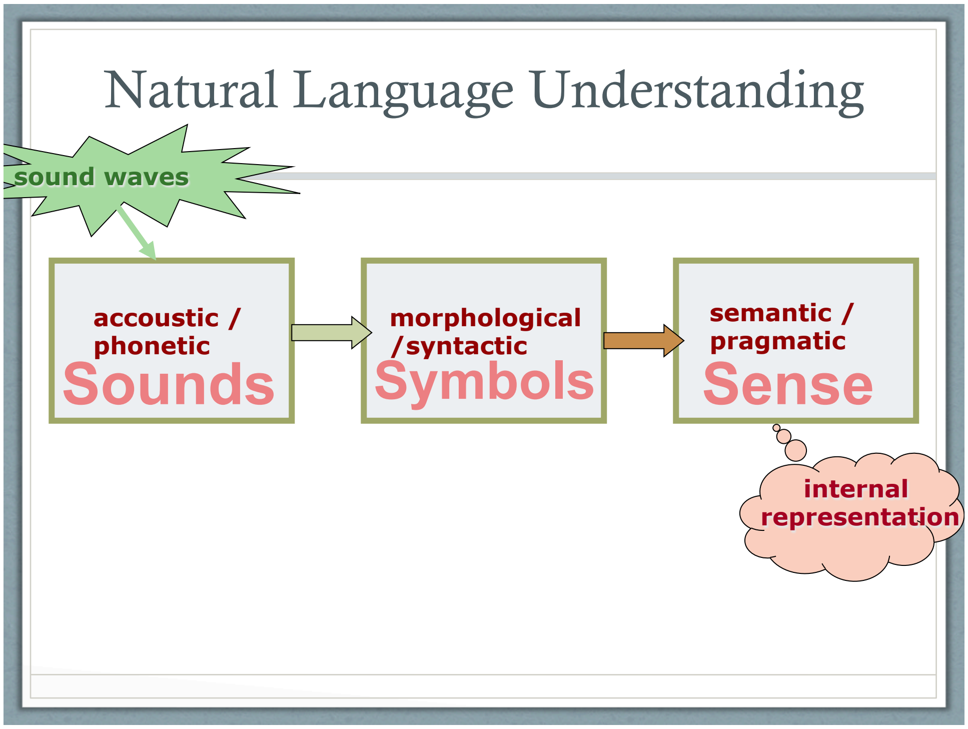
- Machine translation
- spelling/grammar correction
- Information Retrieval
- Data mining
- Document classification
- Question answering, conversational agents

You See It Daily

- Question answering: Siri, OK Google, Cortana
- spelling/grammar correction
- Automated response systems
- To get input for
  - Information Retrieval
  - Data mining
  - Document classification
- Machine translation

Natural Language Understanding





Natural Language Understanding

sound waves

accoustic / phonetic

Sounds

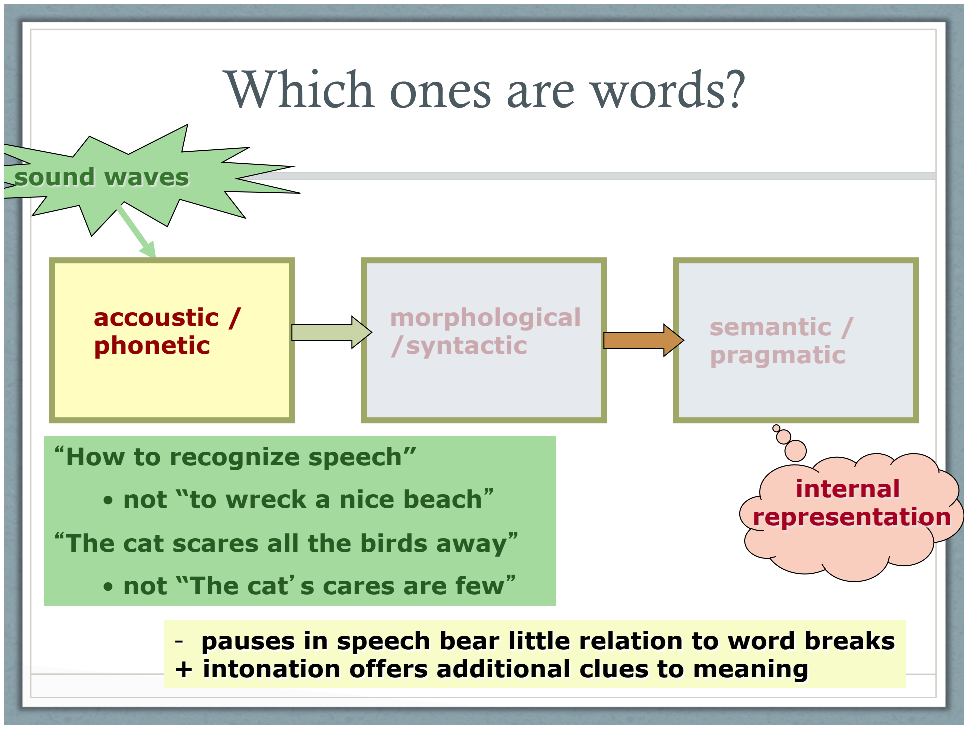
morphological / syntactic

Symbols

semantic / pragmatic

Sense

internal representation



Which ones are words?

sound waves

accoustic / phonetic

morphological / syntactic

semantic / pragmatic

internal representation

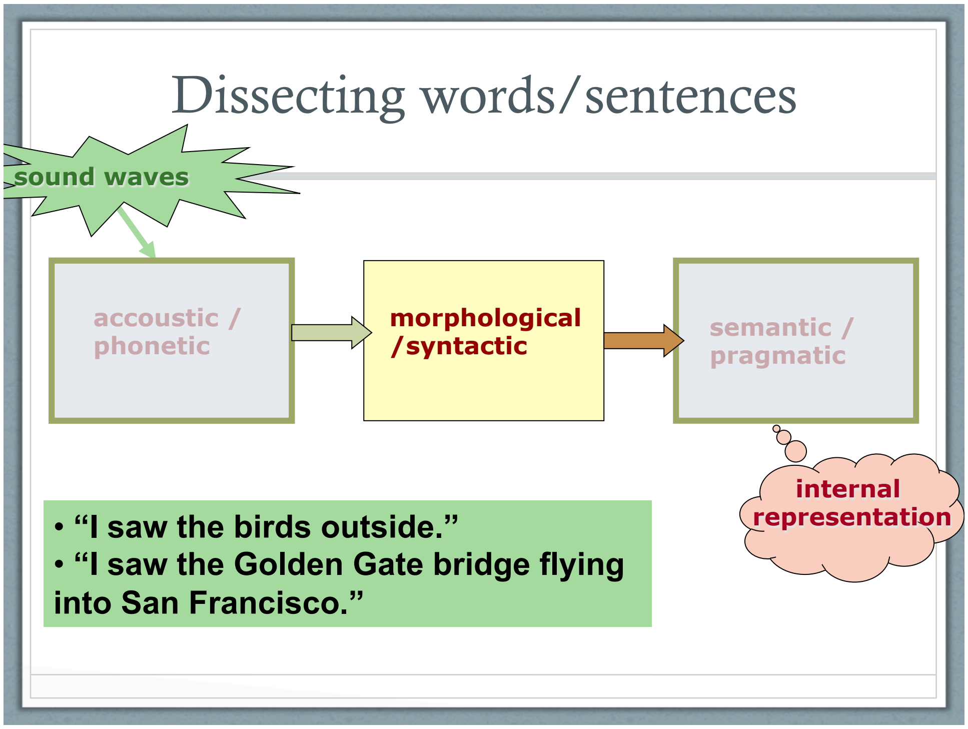
"How to recognize speech"

- not "to wreck a nice beach"

"The cat scares all the birds away"

- not "The cat' s cares are few"

- pauses in speech bear little relation to word breaks
+ intonation offers additional clues to meaning



Dissecting words/sentences

sound waves

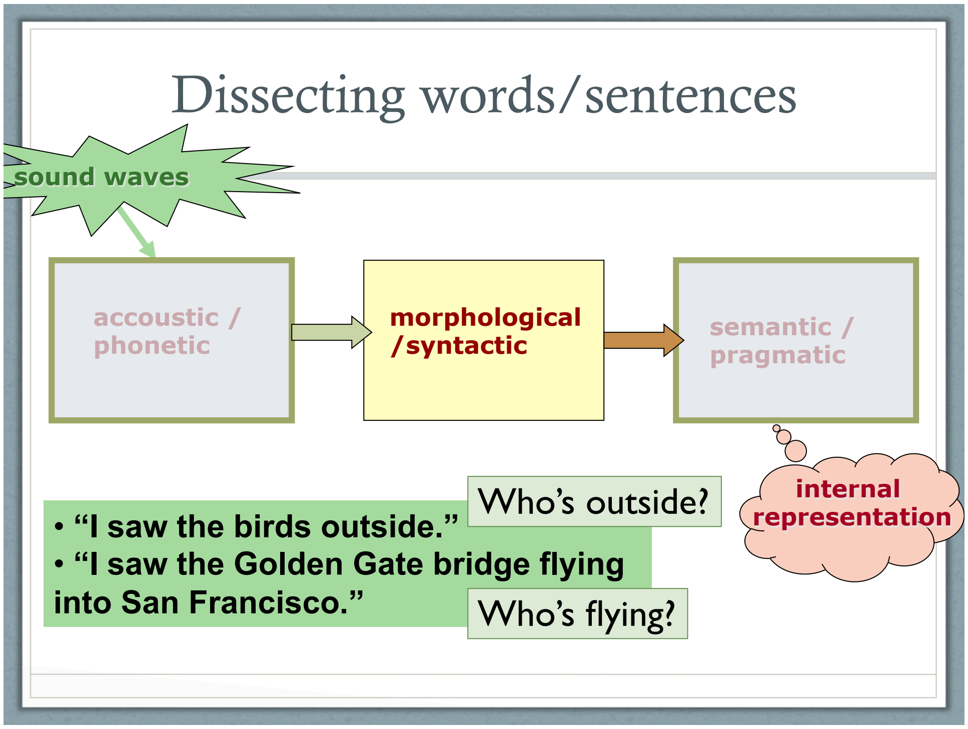
accoustic / phonetic

morphological /syntactic

semantic / pragmatic

internal representation

- "I saw the birds outside."
- "I saw the Golden Gate bridge flying into San Francisco."



Dissecting words/sentences

sound waves

accoustic / phonetic

morphological /syntactic

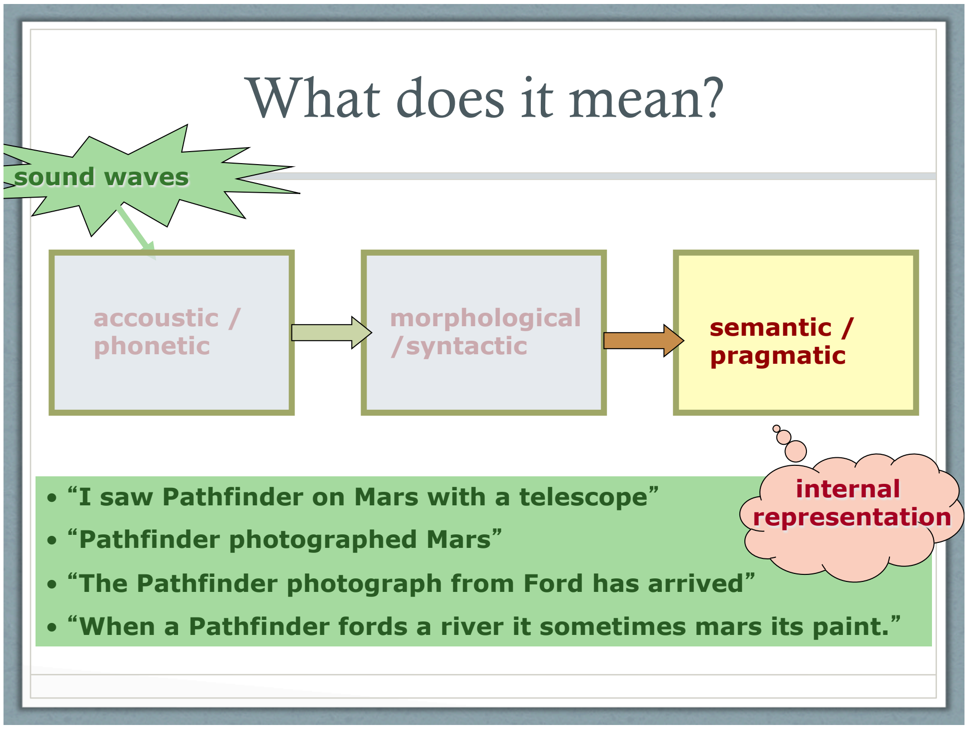
semantic / pragmatic

internal representation

Who's outside?

- "I saw the birds outside."
- "I saw the Golden Gate bridge flying into San Francisco."

Who's flying?



What does it mean?

sound waves

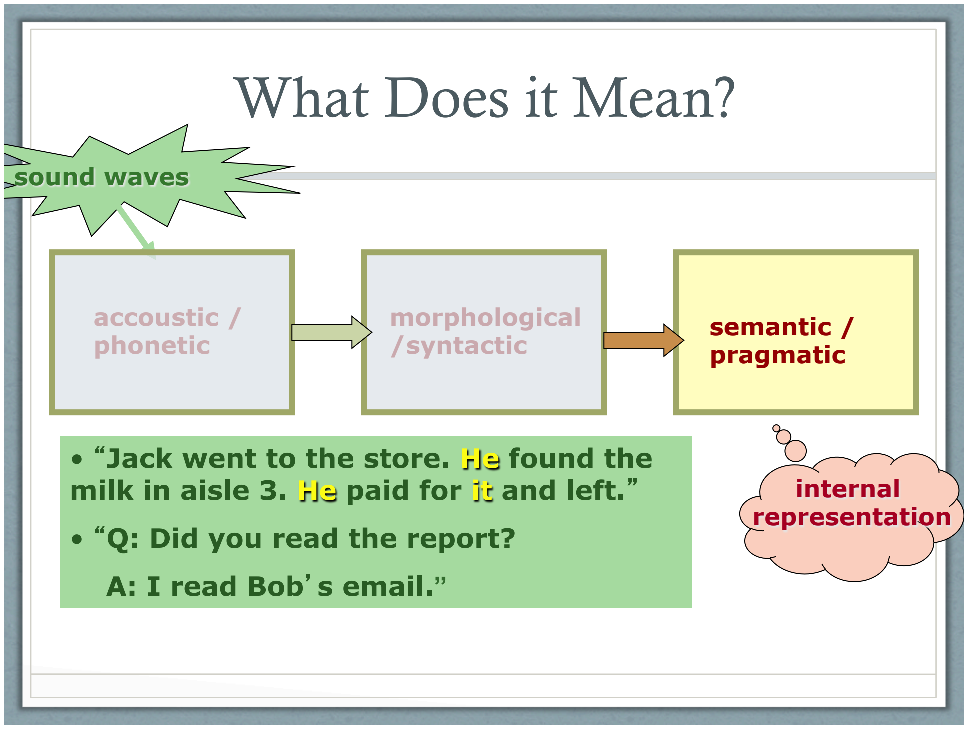
accoustic / phonetic

morphological /syntactic

semantic / pragmatic

internal representation

- “I saw Pathfinder on Mars with a telescope”
- “Pathfinder photographed Mars”
- “The Pathfinder photograph from Ford has arrived”
- “When a Pathfinder fords a river it sometimes mars its paint.”



What Does it Mean?

sound waves

accoustic / phonetic

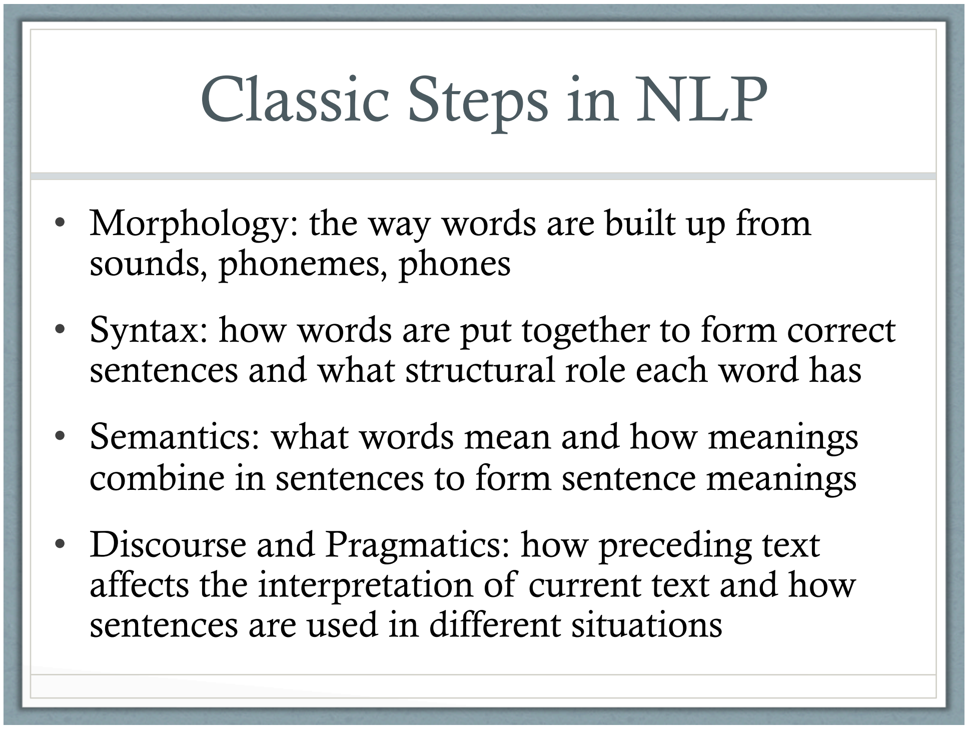
morphological /syntactic

semantic / pragmatic

internal representation

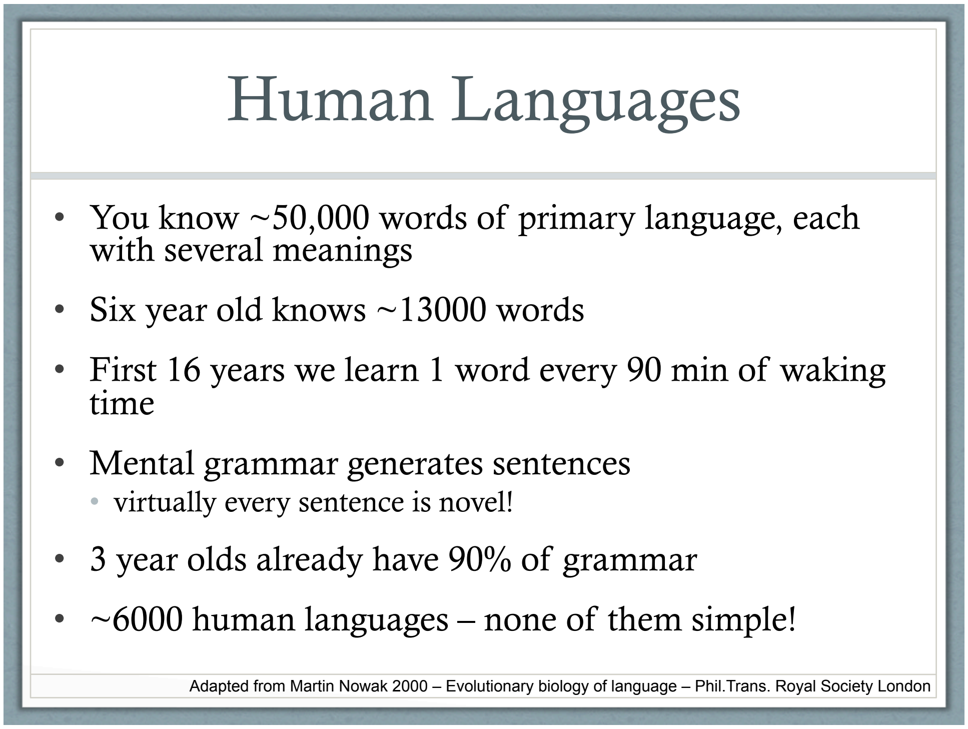
- “Jack went to the store. He found the milk in aisle 3. He paid for it and left.”
- “Q: Did you read the report?

  A: I read Bob’s email.”



Classic Steps in NLP

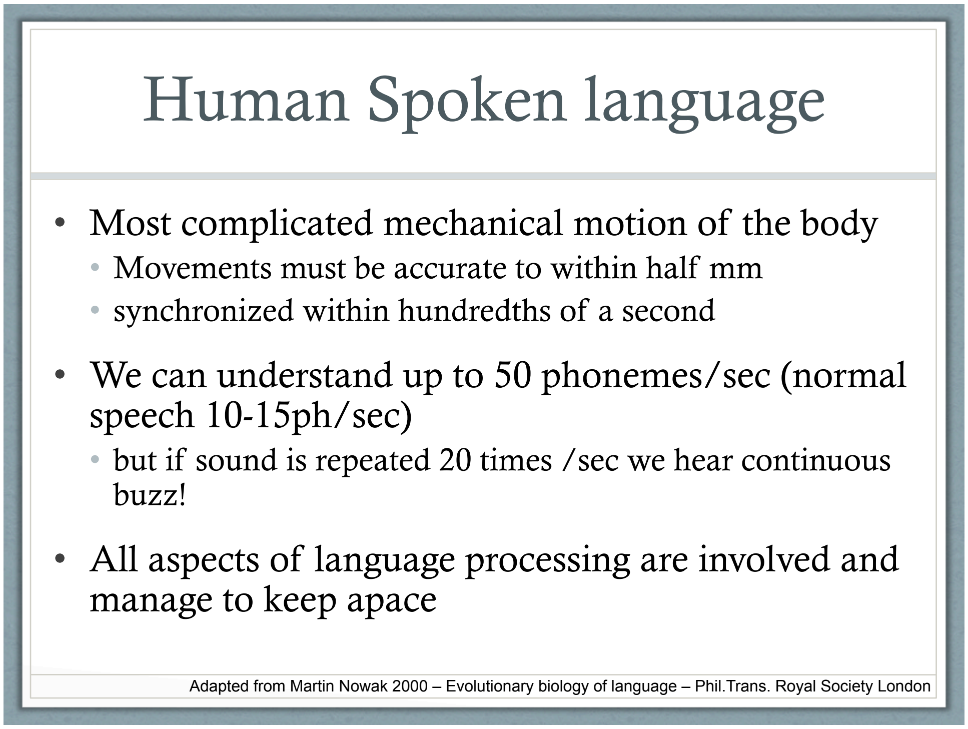
- Morphology: the way words are built up from sounds, phonemes, phones
- Syntax: how words are put together to form correct sentences and what structural role each word has
- Semantics: what words mean and how meanings combine in sentences to form sentence meanings
- Discourse and Pragmatics: how preceding text affects the interpretation of current text and how sentences are used in different situations



Human Languages

- You know ~50,000 words of primary language, each with several meanings
- Six year old knows ~13000 words
- First 16 years we learn 1 word every 90 min of waking time
- Mental grammar generates sentences
  - virtually every sentence is novel!
- 3 year olds already have 90% of grammar
- ~6000 human languages – none of them simple!

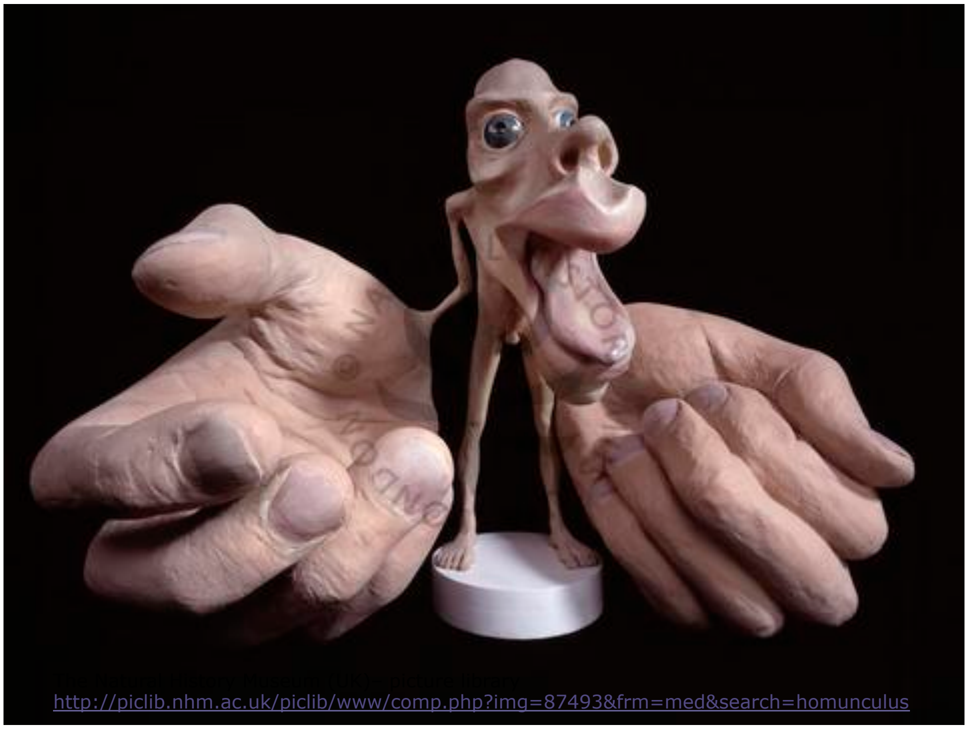
Adapted from Martin Nowak 2000 – Evolutionary biology of language – Phil.Trans. Royal Society London



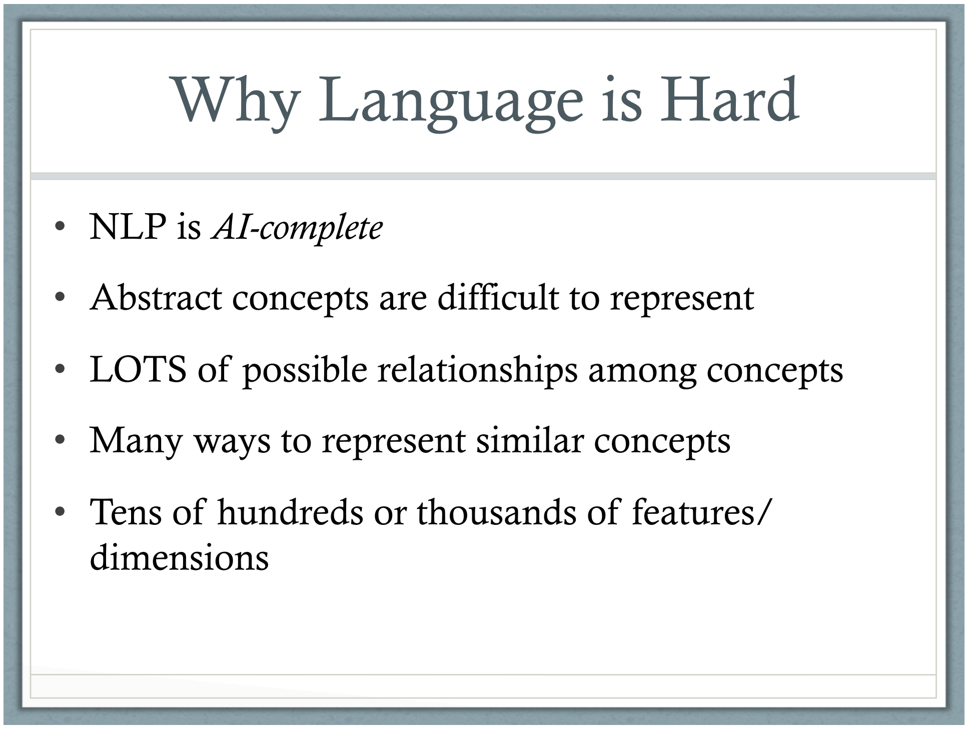
Human Spoken language

- Most complicated mechanical motion of the body
  - Movements must be accurate to within half mm
  - synchronized within hundredths of a second
- We can understand up to 50 phonemes/sec (normal speech 10-15ph/sec)
  - but if sound is repeated 20 times /sec we hear continuous buzz!
- All aspects of language processing are involved and manage to keep apace

Adapted from Martin Nowak 2000 – Evolutionary biology of language – Phil.Trans. Royal Society London

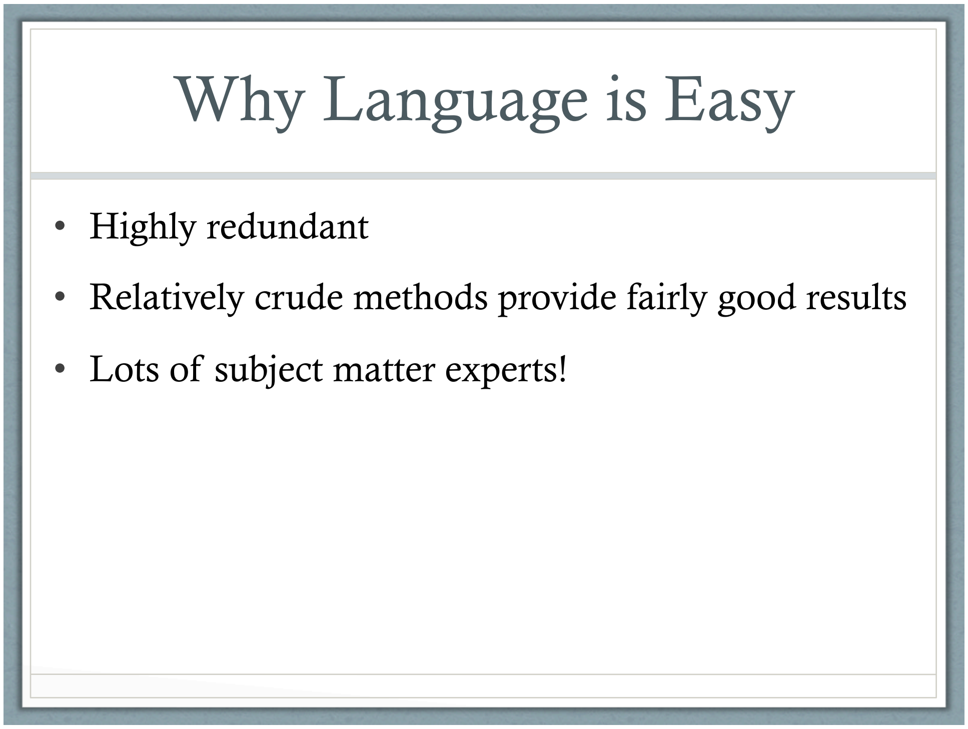


The Natural History Museum (UK)– picture library
http://piclib.nhm.ac.uk/piclib/www/comp.php?img=87493&frm=med&search=homunculus



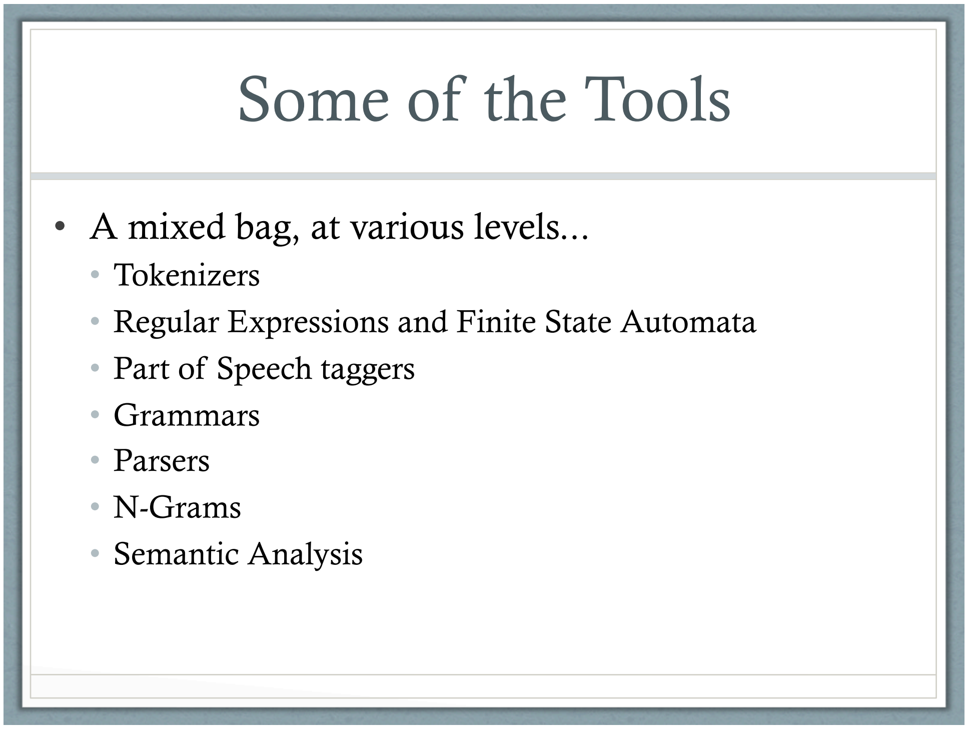
Why Language is Hard

- NLP is *AI-complete*
- Abstract concepts are difficult to represent
- LOTS of possible relationships among concepts
- Many ways to represent similar concepts
- Tens of hundreds or thousands of features/ dimensions



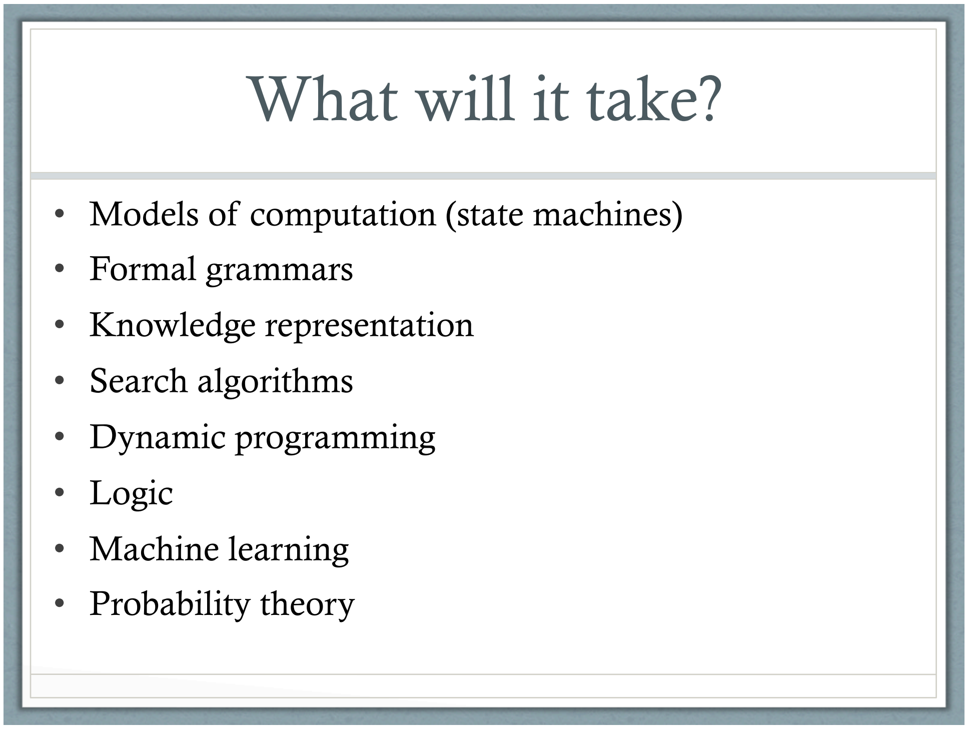
Why Language is Easy

- Highly redundant
- Relatively crude methods provide fairly good results
- Lots of subject matter experts!



Some of the Tools

- A mixed bag, at various levels...
  - Tokenizers
  - Regular Expressions and Finite State Automata
  - Part of Speech taggers
  - Grammars
  - Parsers
  - N-Grams
  - Semantic Analysis



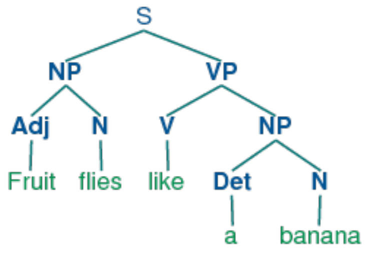
What will it take?

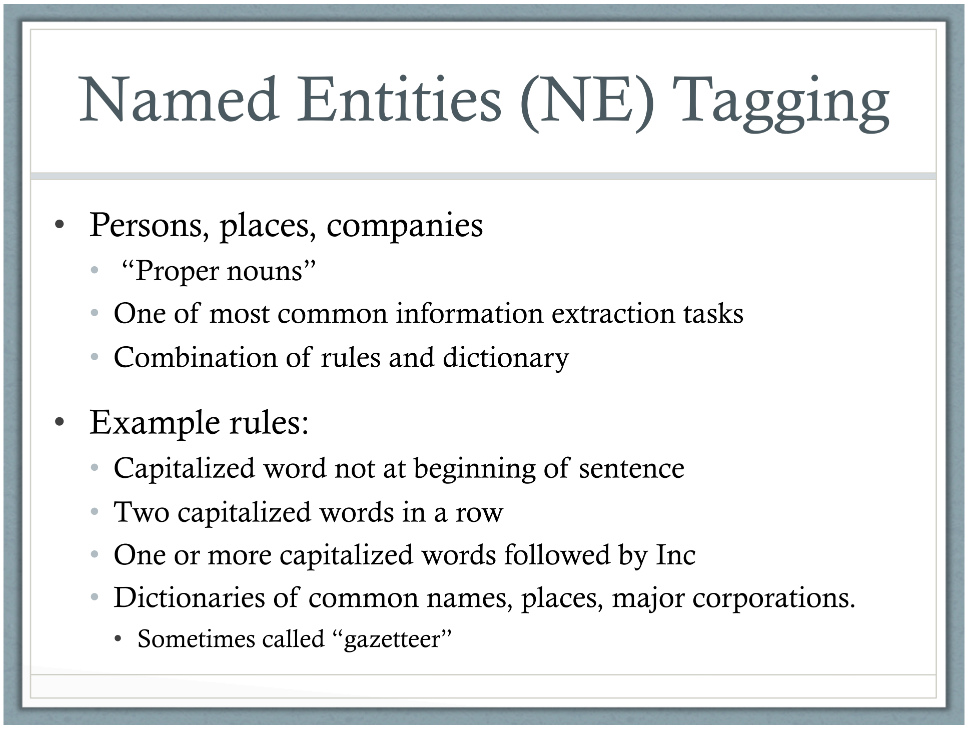
- Models of computation (state machines)
- Formal grammars
- Knowledge representation
- Search algorithms
- Dynamic programming
- Logic
- Machine learning
- Probability theory

A Few Key Problems and Tools

Parts of Speech Tagging

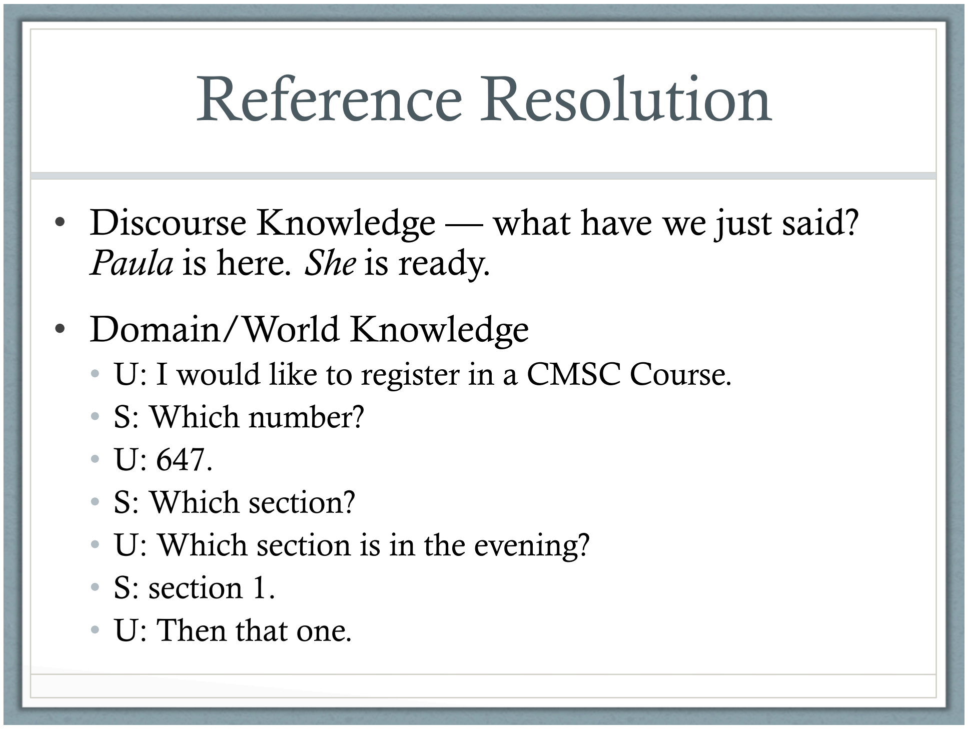
- Part-of-Speech (POS) taggers identify nouns, verbs, adjectives, noun phrases, etc.
- More recent work uses machine learning to create taggers from labeled examples





Named Entities (NE) Tagging

- Persons, places, companies
  - “Proper nouns”
  - One of most common information extraction tasks
  - Combination of rules and dictionary
- Example rules:
  - Capitalized word not at beginning of sentence
  - Two capitalized words in a row
  - One or more capitalized words followed by Inc
  - Dictionaries of common names, places, major corporations.
    - Sometimes called “gazetteer”



Reference Resolution

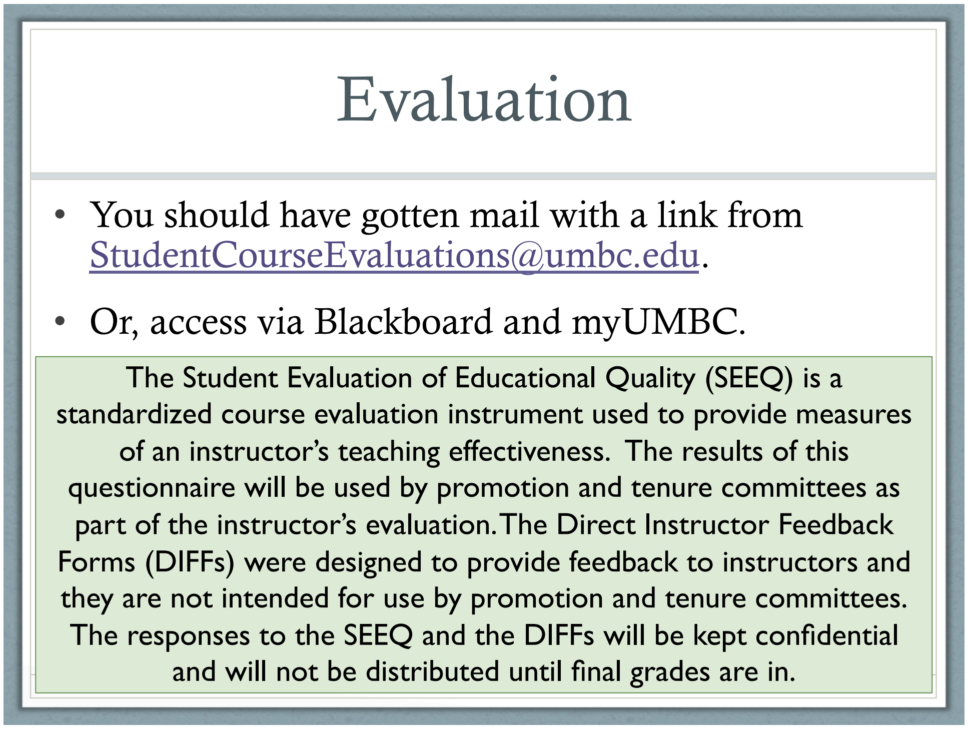
- Discourse Knowledge — what have we just said?
  Paula is here. *She* is ready.
- Domain/World Knowledge
  - U: I would like to register in a CMSC Course.
  - S: Which number?
  - U: 647.
  - S: Which section?
  - U: Which section is in the evening?
  - S: section 1.
  - U: Then that one.

Word Sense Resolution

- Many words have several meanings or **senses**
- We need to resolve which of the senses of an ambiguous word is invoked in a particular use of the word
- I made her duck. (meanings?)
- Again, discourse and world knowledge

Semantics

- What kinds of things can we not do well with the tools we have already looked at?
  - Retrieve information in response to unconstrained questions: e.g., travel planning
  - Accurate translations?
  - Play the “chooser” side of 20 Questions
  - Read a newspaper article and answer questions about it
- These tasks require that we also consider **semantics**: the meaning of our tokens and their sequences



Evaluation

- You should have gotten mail with a link from StudentCourseEvaluations@umbc.edu.
- Or, access via Blackboard and myUMBC.

The Student Evaluation of Educational Quality (SEEQ) is a standardized course evaluation instrument used to provide measures of an instructor's teaching effectiveness. The results of this questionnaire will be used by promotion and tenure committees as part of the instructor's evaluation.The Direct Instructor Feedback Forms (DIFFs) were designed to provide feedback to instructors and they are not intended for use by promotion and tenure committees. The responses to the SEEQ and the DIFFs will be kept confidential and will not be distributed until final grades are in.