Natural Language Processing
IST 664/400 and CIS 668/468

Nancy McCracken
using materials developed in previous courses
by Liz Liddy and others

Natural Language Processing (NLP)

- A range of computational techniques
- for analyzing and representing naturally occurring texts
- at one or more levels of linguistic analysis
- for the purpose of achieving human-like language processing
- for a range of particular tasks or applications.

- Compuational Linguistics – doing linguistics on computers
  - Closely related, often treated as synonymous with NLP

Natural Language as the User Interface

- Goal is complete natural language understanding
  - Enables computers to interact with humans with natural language
    - Vision of future with HAL in 2001

      Dave: "Open the pod bay doors, HAL."
      HAL: "I'm sorry Dave. I'm afraid I can't do that."

- Current approach is to craft human/computer interfaces that are in terms that the computer can understand
    - XML, drop down boxes, other forms of knowledge representation …
    - cleverness is supplied by the human
- Nascent natural language interfaces are being deployed

Where is NLP now?

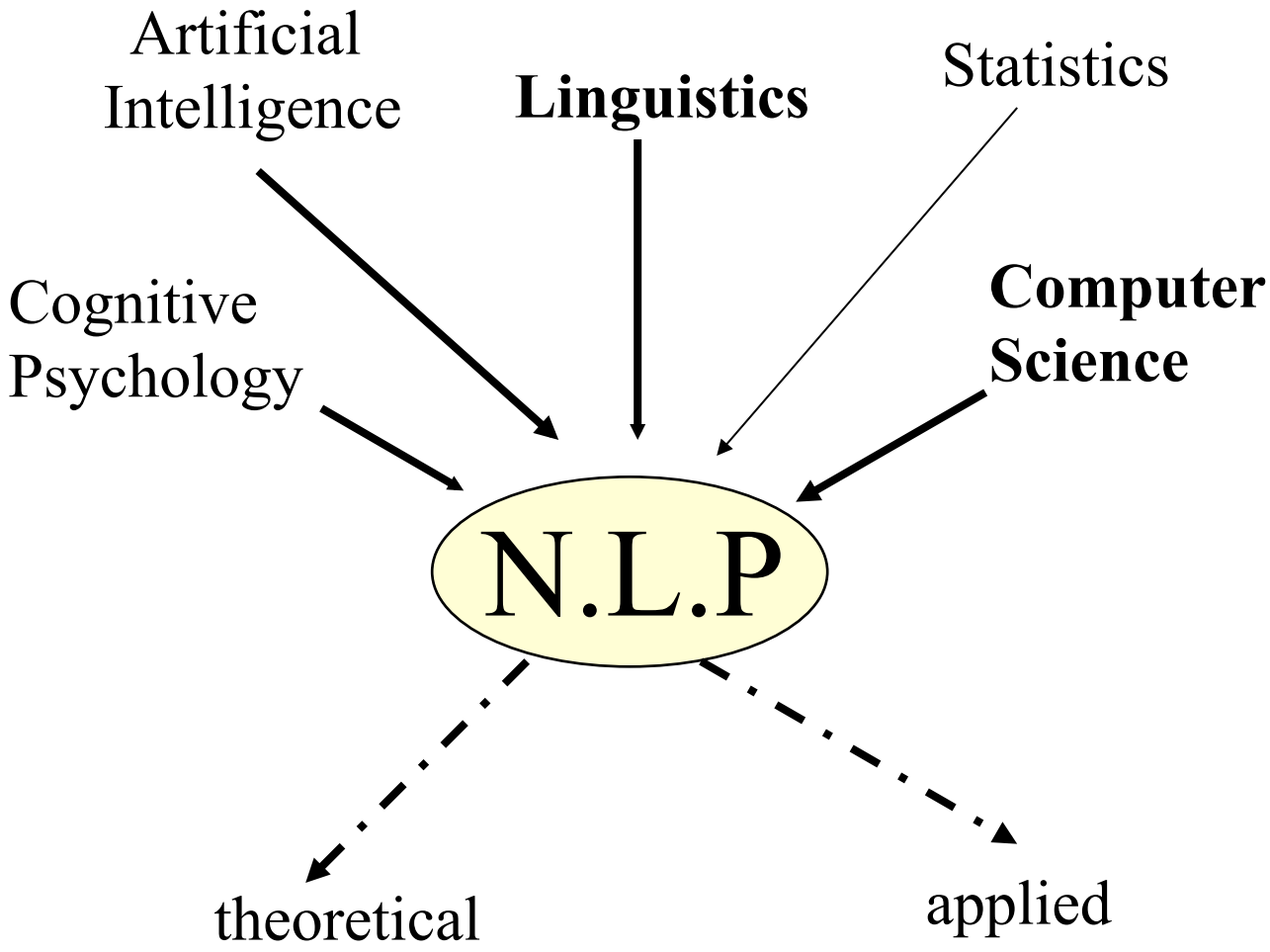
- Goals can be far-reaching
  - True text understanding
  - Reasoning about knowledge in text
  - Real-time participation in spoken dialogs
- Or very down-to-earth
  - Finding the price of products on the web
  - Context-sensitive spell-checking
  - Analyzing authorship or opinions statistically
  - Extracting facts or relations from documents
  - Remembering previous searches and contexts to guide future interactions
- Currently, NLP is providing these practical applications (yet still dreaming of the AI goals)

Need for NLP

- Huge amounts of data
  - Internet
  - Intranet
- Applications for processing large amounts of texts require NLP expertise

- Data Science/Text Mining

Classify text into categories
Index and search large texts
Automatic translation of web documents in different languages
Speech understanding
Understand phone conversations
Information extraction
Extract useful information from resumes
Automatic summarization
Condense 1 book into 1 page
Daily news summaries
Question answering
Knowledge acquisition
Text generations / dialogues


Artificial Intelligence
Linguistics
Statistics
Cognitive Psychology
Computer Science
N.L.P
theoretical
applied

Natural Language Processing's Mixed Lineage

- Linguistics
  - concerned with formal, structural models of language
  - goal is the discovery of language universals
  - not concerned with computational effectiveness of their models
- Computer Science
  - concerned with developing internal representations of data
  - emphasis on efficient processing of these structures

Natural Language Processing's Mixed Lineage

- Cognitive Psychology
  - concerned with modeling the use of language in a psychologically plausible way
  - language as a vehicle for studying human cognition
- Artificial Intelligence
  - interested in development of a computational theory of human language capacity and processing
- Statistics
  - frequencies, probabilities for detecting linguistic patterns

Two Sides of NLP: analysis and generation

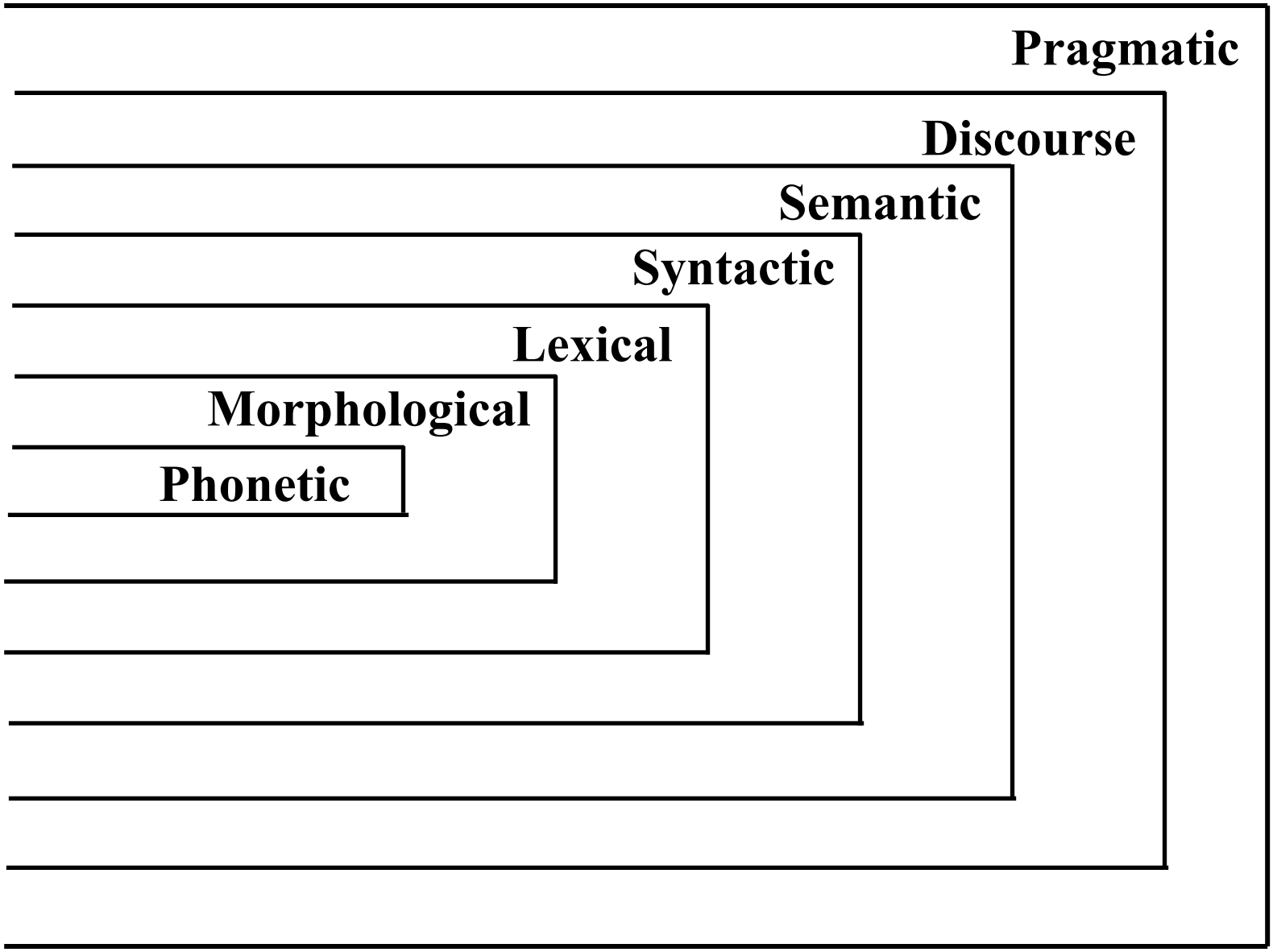
1. paraphrase an input text
2. translate it to another language
3. answer questions about it
4. draw inferences from it

Natural Language Processing

Language Analysis* Language Generation

*Main emphasis in this course

Synchronic Model of Language



Why is NLP so hard?

- **Seems pretty simple for humans**
  - Usually quite unaware of the complexity of the language tasks they perform so effortlessly
- Some reasons are
  - Ambiguity
  - Subleties of meaning
    - Irony, sarcasm, humor, metaphor

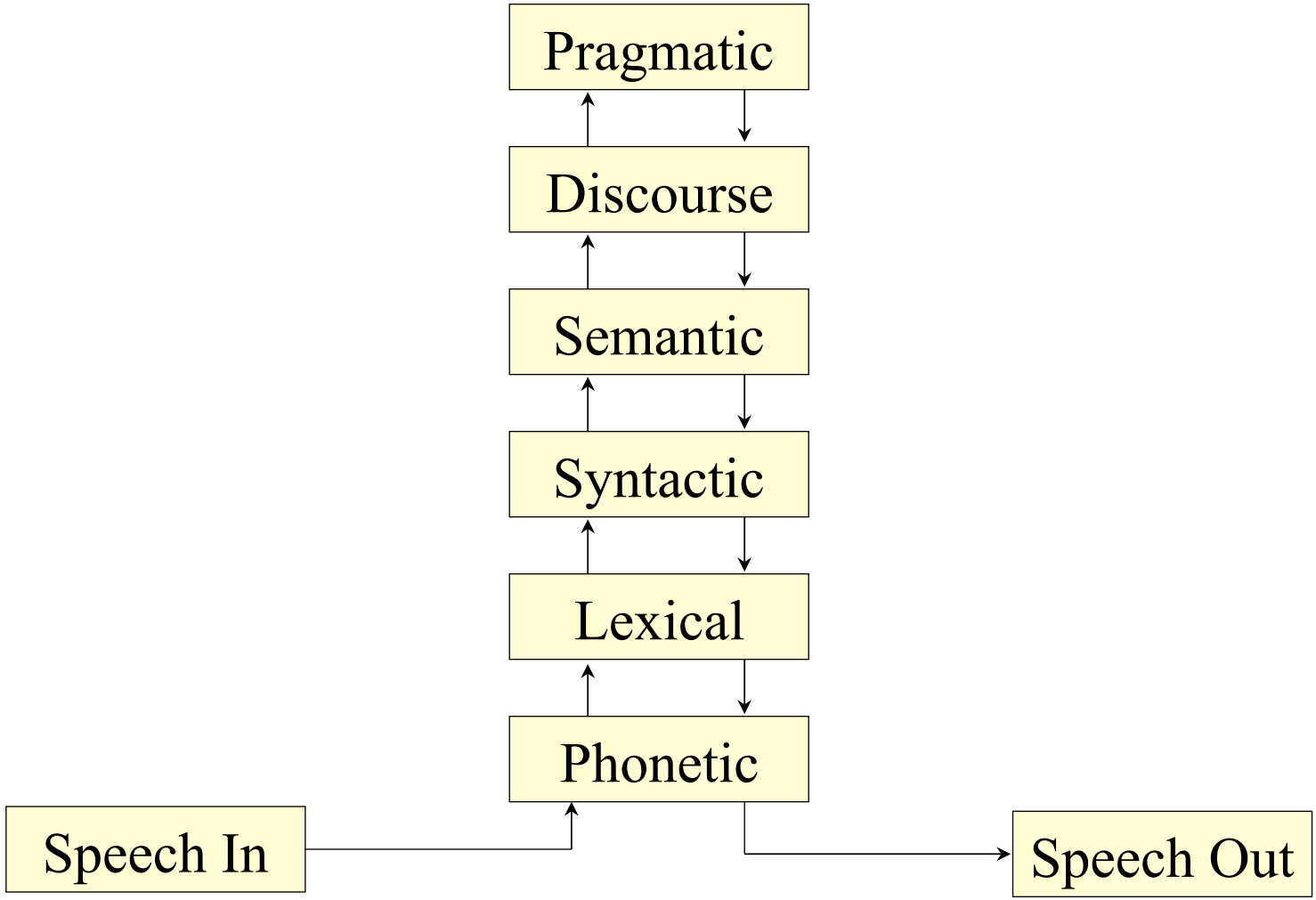
Ambiguous Newspaper Headlines

- Ban on Nude Dancing on Governor's Desk
- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Local High School Dropouts Cut in Half
- Red Tape Holds Up New Bridges
- Clinton Wins on Budget, but More Lies Ahead
- Hospitals are Sued by 7 Foot Doctors
- Kids Make Nutritious Snacks
  - Examples collected by Chris Manning

Ambiguity at many levels

- **Word sense ambiguity**
  - *I need some information on getting rid of moles.*
- **Structural ambiguity**
  - *Visiting relatives can be a nuisance.*
  - *He was shot by the man from Moscow.*
- **Semantic ambiguity**
  - *Mom said that when I visited Aunt Peggy in the hospital I should take her flowers.*
- **Referential ambiguity**
  - *Take Michael to the doctor. Tell him what happened.*
- **Literal ambiguity**
  - *Do you know what time it is?*

A Linear Model of Language Processing



NLP Application Areas

- Machine Translation – conversion of text from one language to another
  - See Yahoo Babelfish
  - MT techniques use context , not just word for word substitution
  - Often statistically based patterns of word usage and context
  - Usefulness of Parallel Corpora

- Information Retrieval / Search Engines – provision of documents containing requested information
  - Google, many other search engines
  - Use lowest levels of NLP to stem words, find phrases for indexing documents
  - Users conform to keyword query restriction, instead of natural language queries

NLP Application Areas

- Information Extraction / Text-mining – populating a structured database with specific bits of information found in text
  - Competitive Intelligence analyzes news text and web blogs for
    - Names of people, companies and other entities
    - Relations between them, e.g. corporate roles, or events such as mergers
- Human-computer Interfaces – interactive querying of databases
- Summarization – abstraction and condensation of text's major points
  - Current systems select a set of significant sentences from the document as a summary
  - Example summaries: http://www.tnewfields.info/Articles/sum1.htm

NLP Application Areas

- Metadata Generation – assignment of values for metadata elements in a particular standard, e.g. Dublin Core
- Question & Answering Systems – focused information provision
  - Identify question focus as desired information
  - Must be able to handle many different phrasings of desired answer and to provide justification
    Q: *What year did Marco Polo travel to Asia?*
    A: *Marco polo divulged the truth after returning in 1292 from his travels, which included several months on Sumatra.*
  - Web sites like ask.com
  - And . . .

NLP Application Areas

- Question & Answering Systems – Watson
  - IBM's question answering system trained to play Jeopardy
  - Extensive development of NLP techniques

Course Work Description

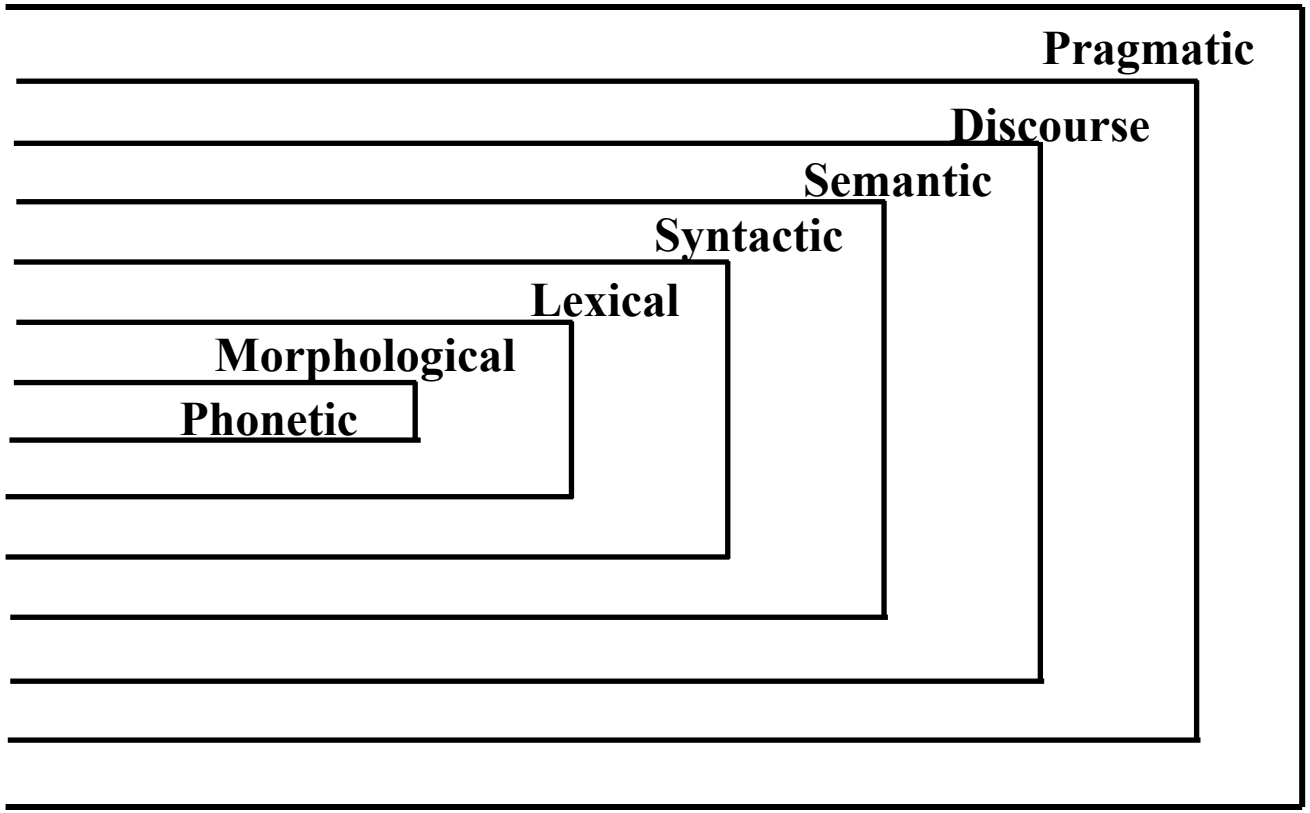
- Classroom sessions – every Monday
  - Lecture, In-class exercises, Discussions
- Lab sessions – every Wednesday
  - Guided lab exercises, independent (group) exercises, resulting in discussion or discussion thread posting in the Blackboard system
  - Use NLP programs provided in the Natural Language Toolkit (NLTK) to process text for analysis
- Three Homework Problems
  - Corpus Statistics
    - Use NLTK to collect frequencies of words and word pairs for text analysis; text may be collected
  - Regular Expressions for obfuscated email addresses
  - Probabilistic CFG

Course Work Description

- Final project
  - Basic part of project will be a sentiment classification task
  - And choice of advanced projects
    - Further work on classification
    - Other computational task such as lexical chains
    - Semester long annotation task – labeling emotion in tweets
    - or you may choose to define your own
    - Group work is encouraged but not required
- NLP applications
  - Graduates
    - Groups will be assigned according to interest expressed in topics
    - Investigate the topic (read paper(s), search for examples, etc.) and write up as a presentation
  - Undergraduates read presentations and write short report

Levels of Language Analysis

- Use the synchronic model to guide computational techniques to analyze text (as much as possible)



Synchronic Model of Language

- The more exterior the level of language processing:
  - The larger the unit of analysis
    - phoneme-> morpheme -> word -> sentence -> text -> world
    - The less precise the language phenomena
  - The more free choice & variability
    - less rule-oriented, more exceptions
    - just regularities
  - The more levels it presumes a knowledge of or reliance on
  - Theories used to explain the data move more into the areas of cognitive psychology and AI
- Lower levels of the model have been more thoroughly investigated and incorporated into NLP systems

Speech Processing

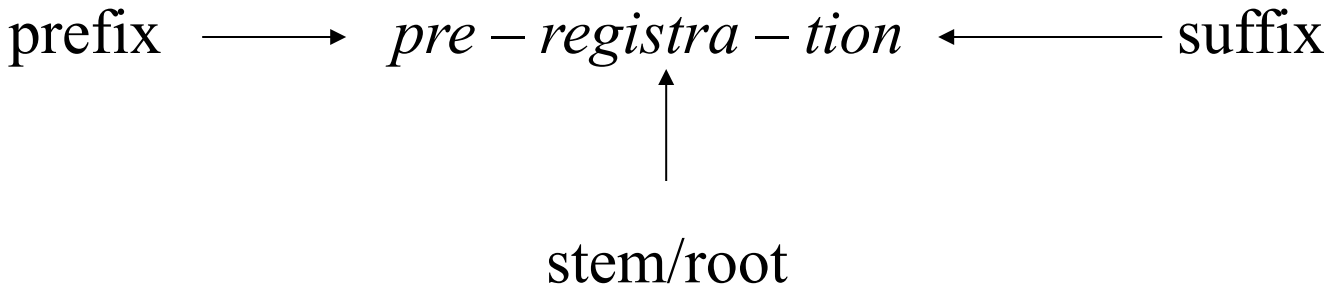
- Interpretation of speech sounds within & across words
- sound waves are analyzed and encoded into a digitized signal

Rules used in Phonological Analysis

1. Phonetic rules – sounds within words
2. Phonemic rules – variations of pronunciation when words are spoken together
3. Prosodic rules – fluctuation in stress and intonation across a sentence

Morphological Analysis

- deals with the componential nature of lexical entities:



- What features do inflections reveal?

Verbs → tense & number

Nouns → single/plural

Adjectives → comparison features

Lexical

1. Part-of-speech (POS) tagging

03/14/1999 (AFP)… the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

… the|DT extremist|JJ Harkatul_Jihad|NP group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NP dissident| NN Osama_bin_Laden|NP …

2. Productive rules which explain how new words are formed

highchair

egghead

Word Level Meaning

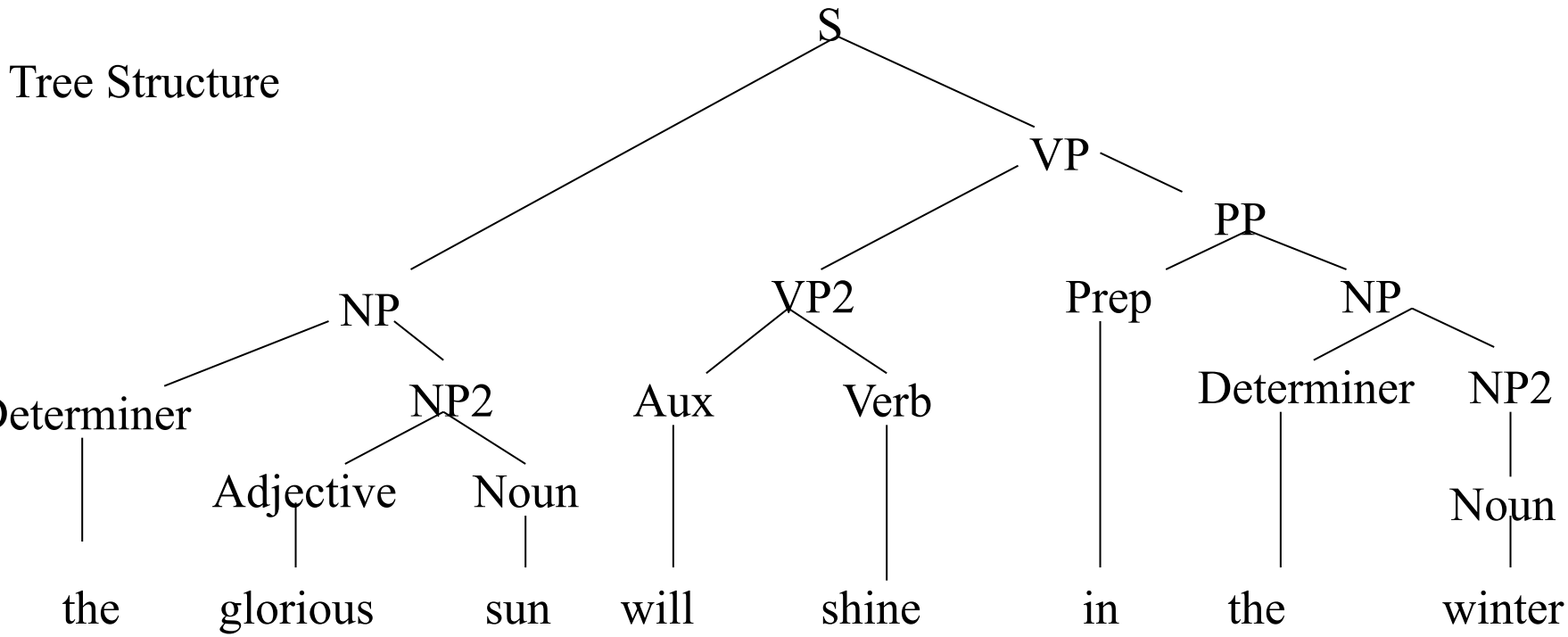
- Usually given by online lexicon such as WordNet
- Word with senses
  - Example: launch
- Definitions
  - Noun sense 1: a large, usually motor-driven boat used for carrying people on rivers, lakes harbors, etc.
  - Verb sense 1: set up or found
- Synonyms
  - Verb sense 1: establish, set up, found

Syntactic Analysis

- analyzing of words in a sentence so as to uncover the grammatical structure of the sentence

- requires both a grammar and a parser

- produces a de-linearized representation of a sentence which reveals dependency relationships between words

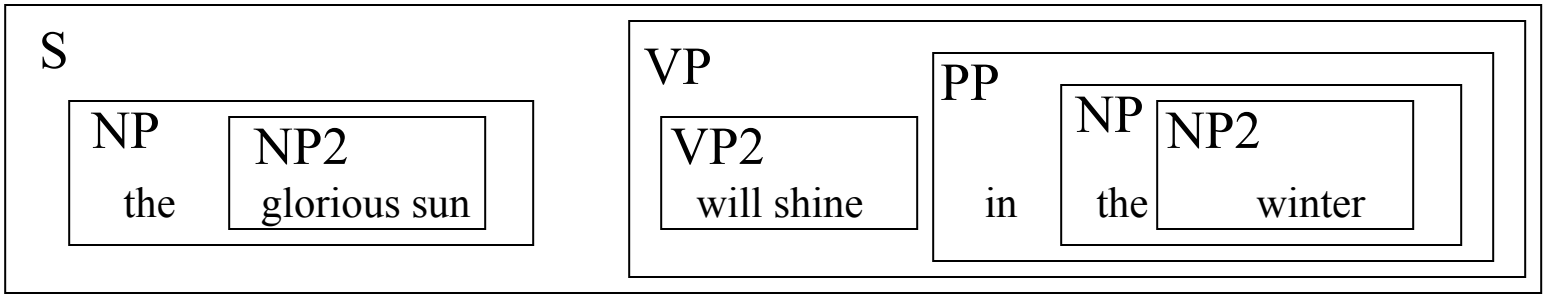


Bracketed text

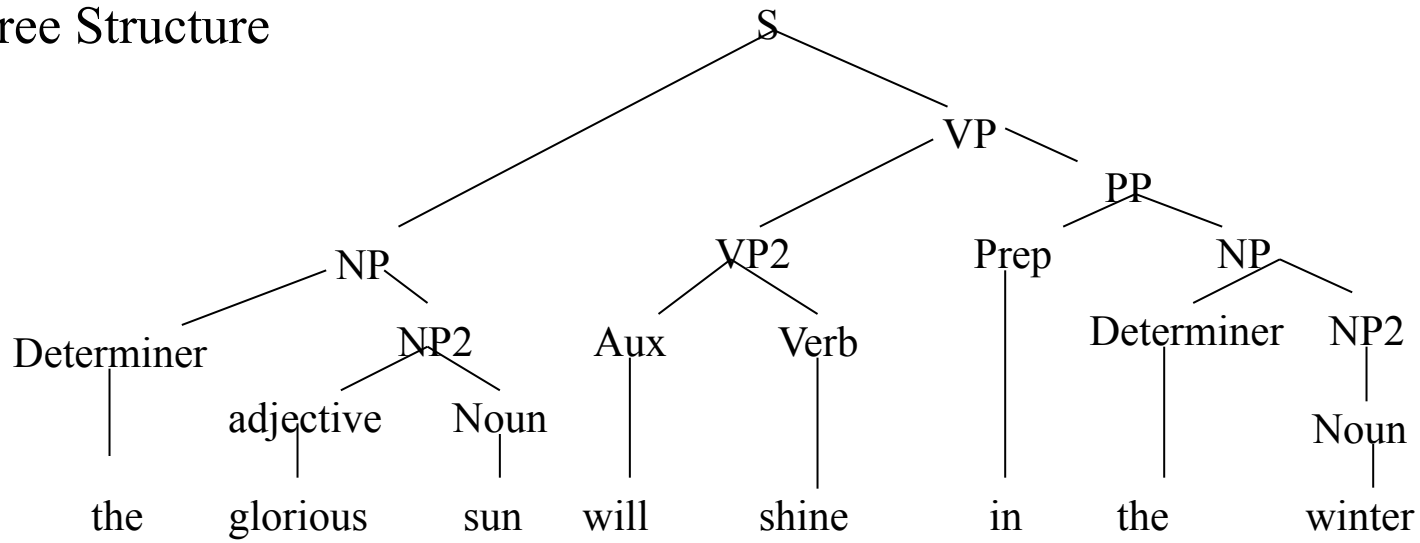
[$_{S}$ [$_{NP}$ the [$_{NP2}$ glorious sun]]

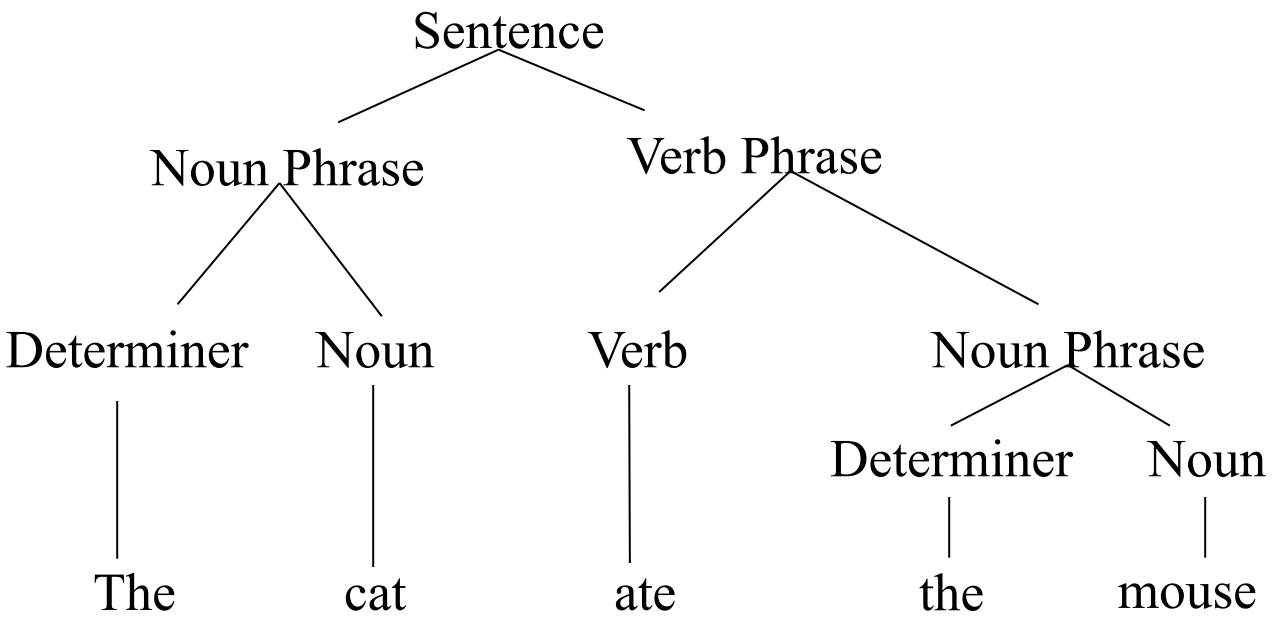
[$_{VP}$ [$_{VP2}$ will shine] [$_{PP}$ in [$_{NP}$ the [$_{NP2}$ winter]]]]]

Nested Boxes



Tree Structure





The phase structure rules underlying this analysis are as follows:

Sentence → Noun Phrase Verb Phrase

Noun Phrase → Determiner Noun

Verb Phrase → Verb Noun Phrase

Determiner = The

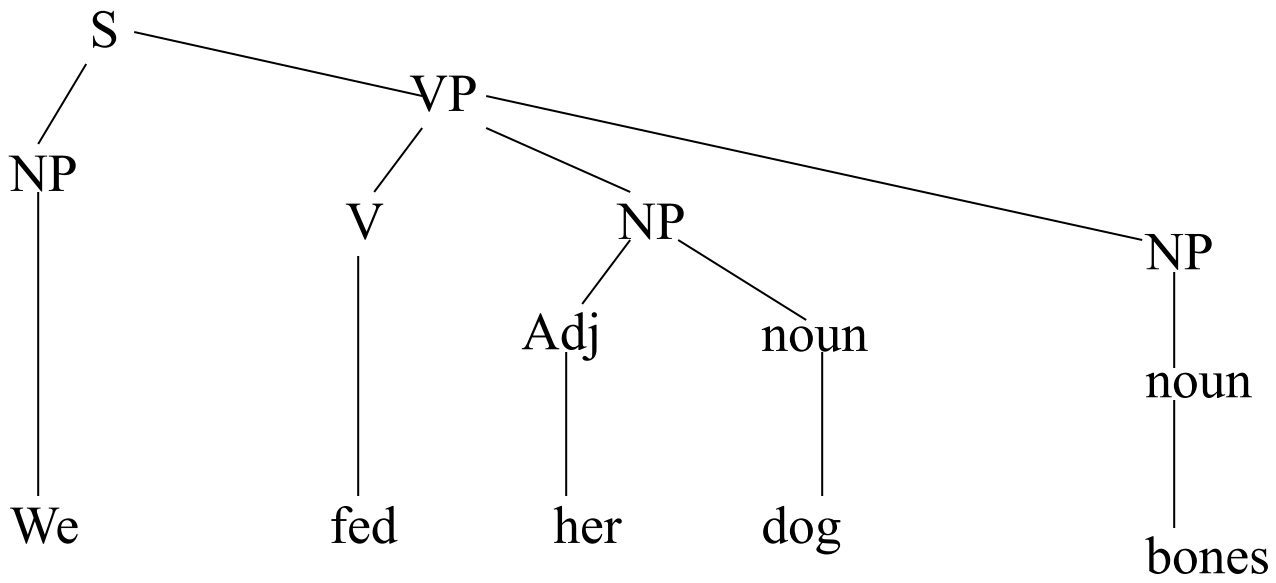
Noun = cat

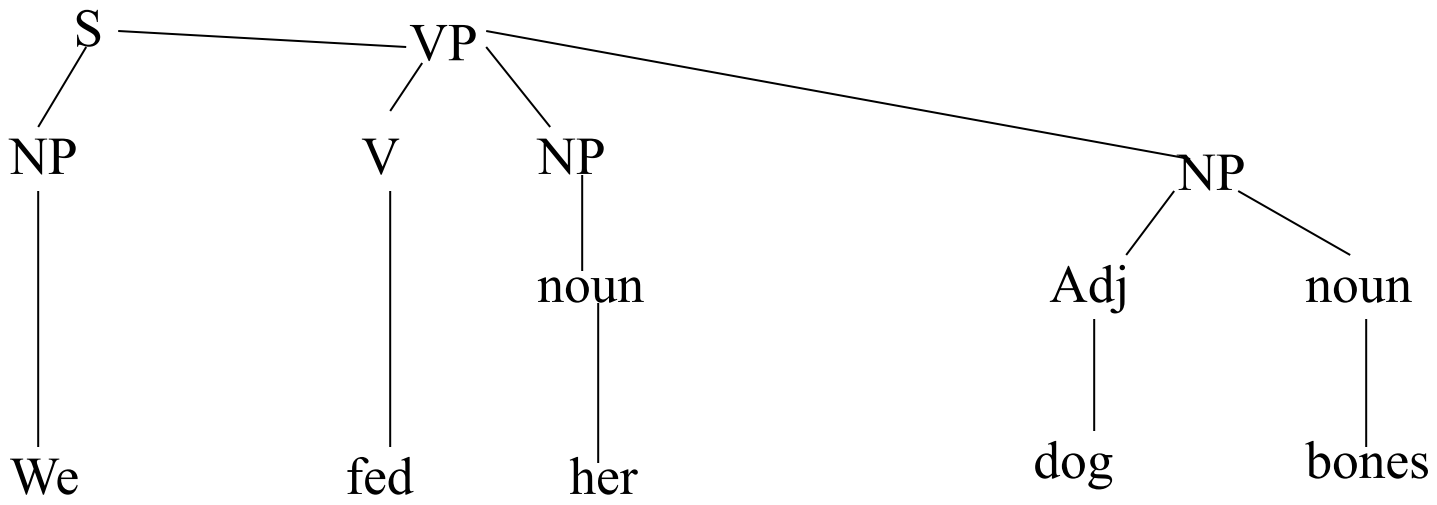
Noun = mouse

Verb = ate

Parsing a sentence using simple phrase structure rules

Syntactic Ambiguity: We fed her dog bones





Semantics

- Determining possible meanings of a sentence
  - Interactions among words affect lexico-semantic interpretation
- Capturing meaning of a sentence in a knowledge representation formalism

Semantic Role Labeling (SRL) Problem

- In a sentence, a verb and its semantic roles form a proposition; the verb can be called the predicate and the roles are known as arguments.
- Given a target verb, the Semantic Role Labeling task is to identify and label each semantic role present in the sentence.

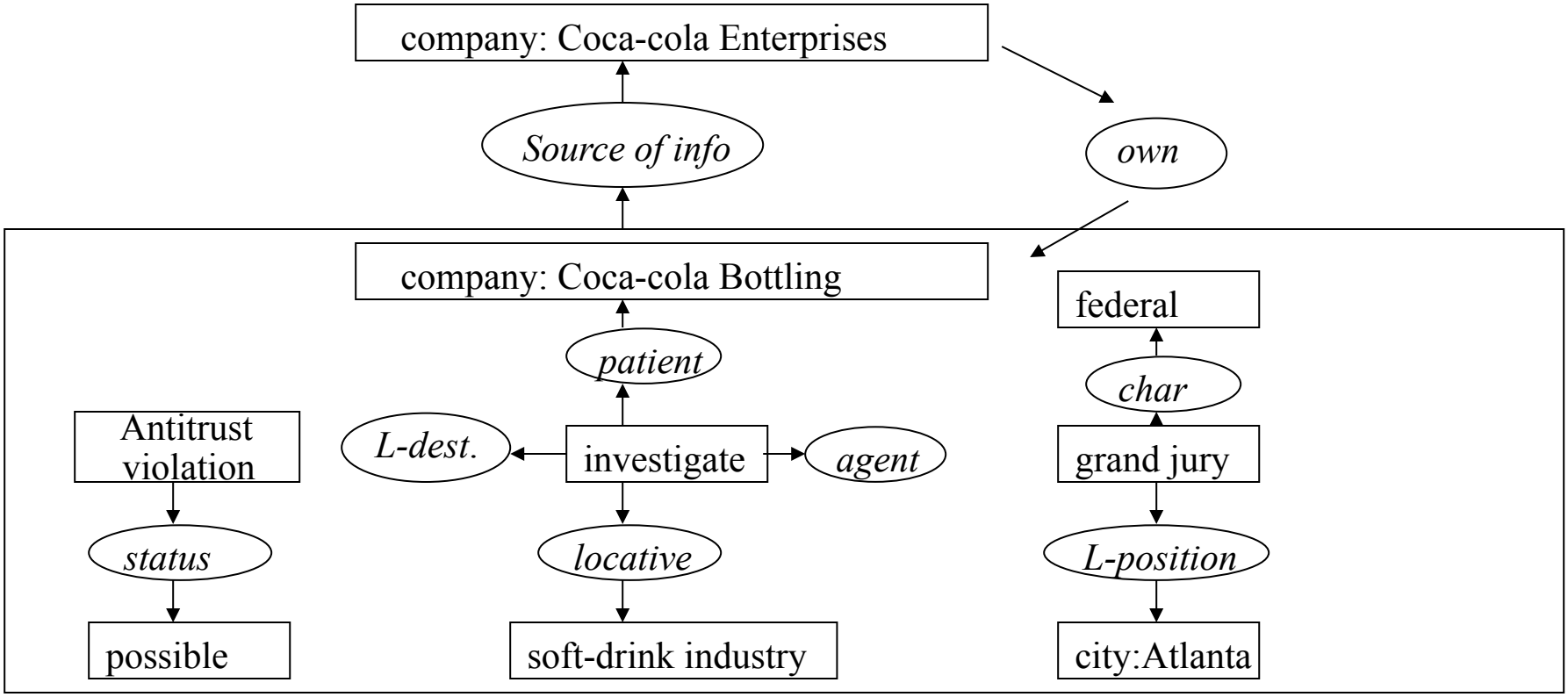
When Disney offered to pay Mr. Steinberg a premium for his shares, the New York investor didn't demand the company also pay a premium to other shareholders.

Example roles for the verb "pay", using roles more specific than theta roles:

When [$_{\text{payer}}$ Disney] offered to [$_{\text{V}}$ pay] [$_{\text{recipient}}$ Mr. Steinberg] [$_{\text{money}}$ a premium] for [$_{\text{commodity}}$ his shares], the New York investor …

Semantic Relation Extraction

Coca-Cola Enterprises, Inc. said its Atlanta Coca-Cola Bottling Co. unit is a target of an investigation into alleged antitrust violations in the soft-drink industry by a federal grand jury in Atlanta.



Discourse

- determining meaning in texts longer than a sentence

- making connections between component sentences
  - multi-sentence texts are not just concatenated sentences to be interpreted singly
  - Documents may have distinct patterns in different sections: introduction, conclusions, methodology, etc.
  - Text in dialogs has distinct forms according to position in the dialog

- interpretation of later-mentioned entities depends on interpretation of earlier-mentioned entities – ‘anaphora’

Anaphora (coreference) resolution

- Excerpt from story by Farhad Manjoo of Slate "Siri vs. Google"

> "Google Voice Search isn't close to realizing that vision, but it's not impossibly far off either. Huffman points out that Google's app can already hold very small conversations. It understands pronouns, so if you ask, "Who is Barack Obama?" and then ask, "Who is his wife?", it knows that his refers to Obama. And most important, it gives you the correct answer.
>
> I just tried the same set of queries with Siri. First, she correctly identified the president. But when I asked, "Who is his wife?" she shot back, "What is your wife's name?" That's not what I asked. Actually, it's really, really far off. And there aren't any signs that Apple's voice assistant is going to get much closer any time soon."

Anaphora (coreference) resolution

The city councilors refused the demonstrators a permit because **they** feared violence.

The city councilors refused the demonstrators a permit because **they** advocated revolution.

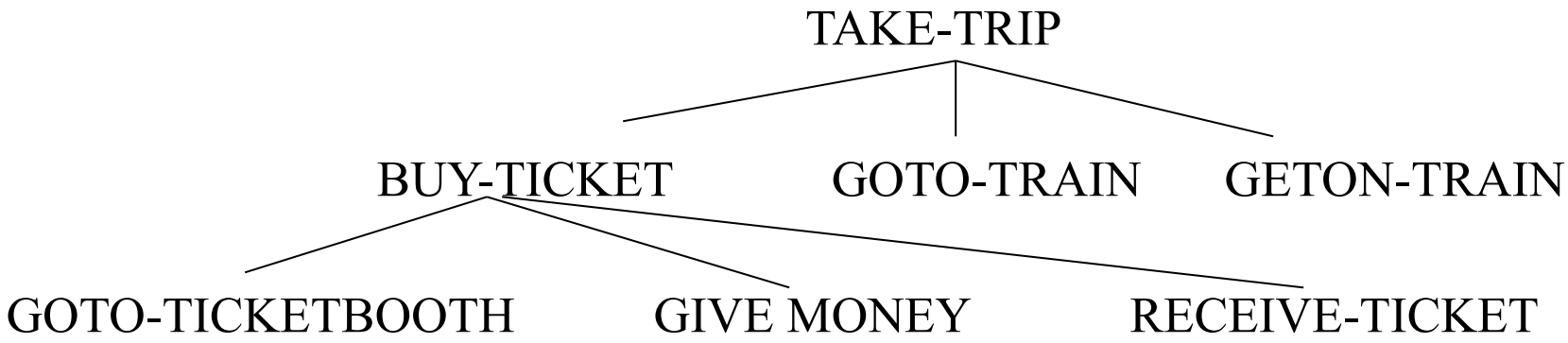
Pragmatics

- The purposeful use of language in situations
  - A functional perspective

- Those aspects of language which require context for understanding

- Goal is to explain how extra meaning is *read into* texts without actually being encoded in them

- Requires much world knowledge
  - Understanding of intentions / plans / goals



Sketch of a commonsense task plan to take a trip

Techniques for NLP Analysis

- Corpus Statistics
  - Frequencies of words
  - Frequencies of word pairs, using co-occurrence or semantic measures

- Classification or other Machine Learning
  - Use NLP to produce features, also known as attributes, of the text
  - Classify the text according to a set of labels
    - Classify customer reviews as positive or negative
    - Classify news articles according to topic