Natural Language Processing
IST 664

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.

- Computational 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: A Space Odyssey

      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?

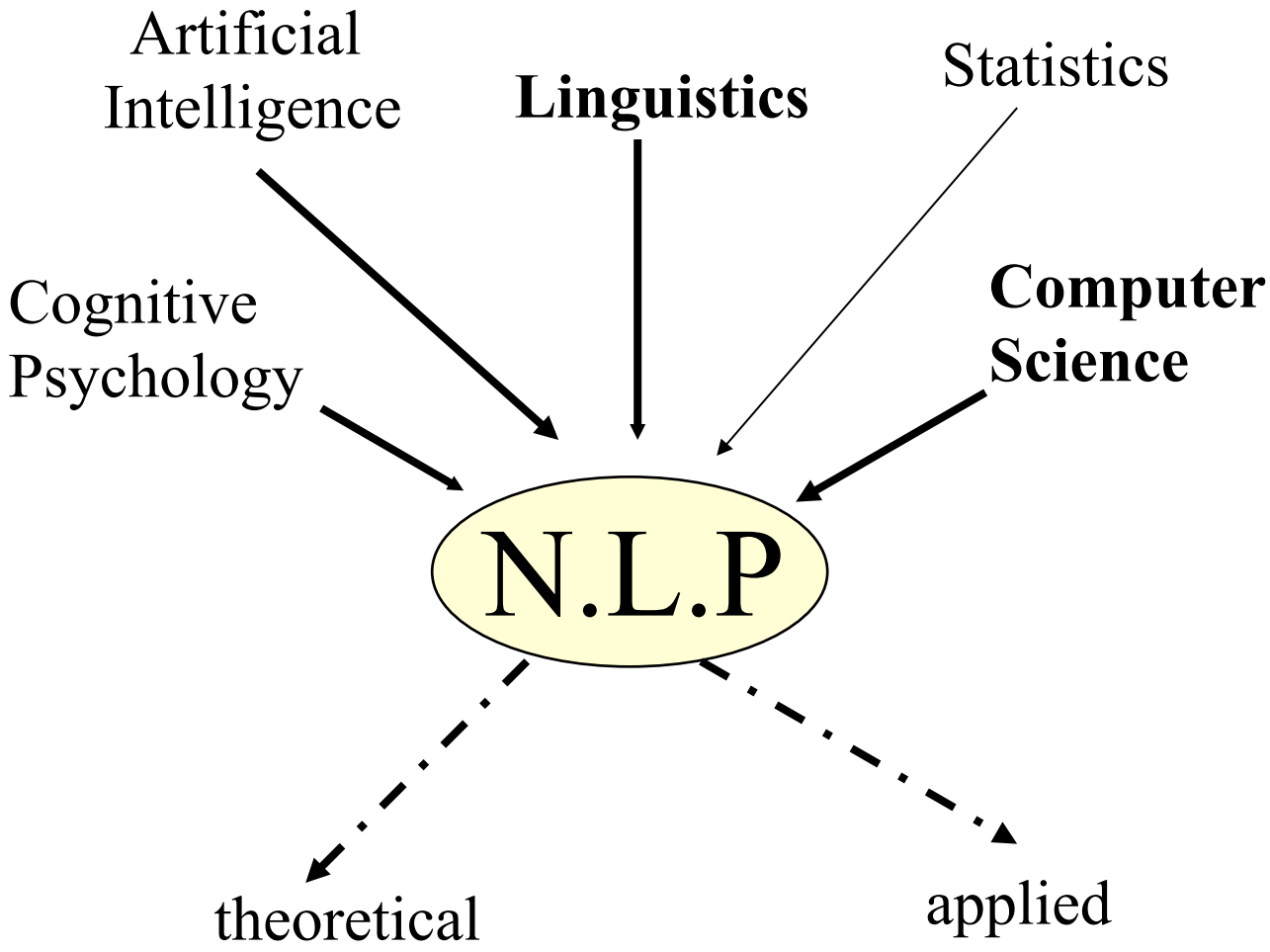
- Goals can be far-reaching
  - True text understanding
  - Reasoning about knowledge in text
  - Real-time participation in spoken dialogs
- Or very down-to-earth
  - Finding the price of products on the web
  - Context-sensitive spell-checking
  - Analyzing authorship or opinions statistically
  - 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

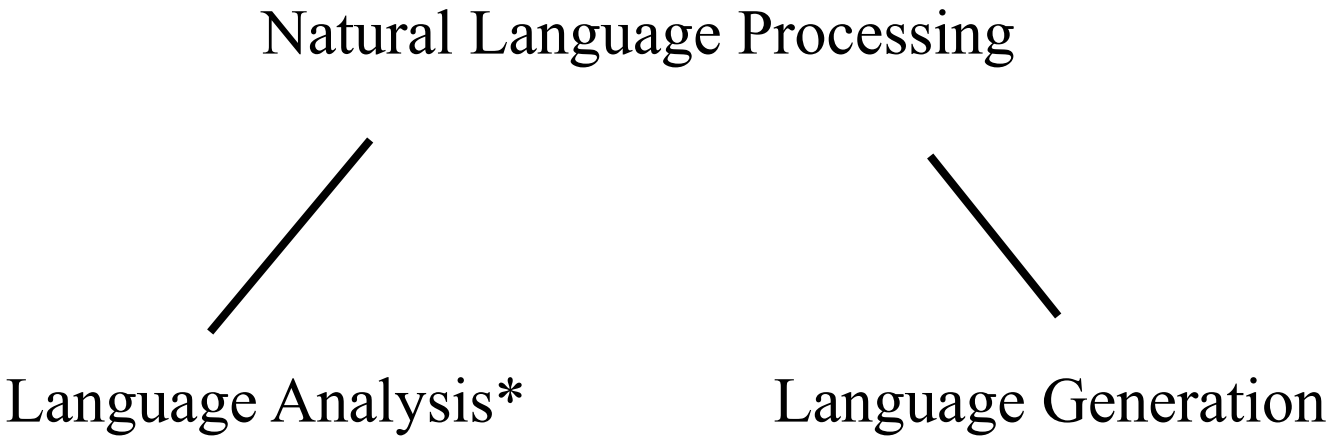
- Linguistics
  - concerned with formal, structural models of language
  - goal is the discovery of language universals
  - not concerned with computational effectiveness of their models
- Computer Science
  - concerned with developing internal representations of data
  - emphasis on efficient processing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

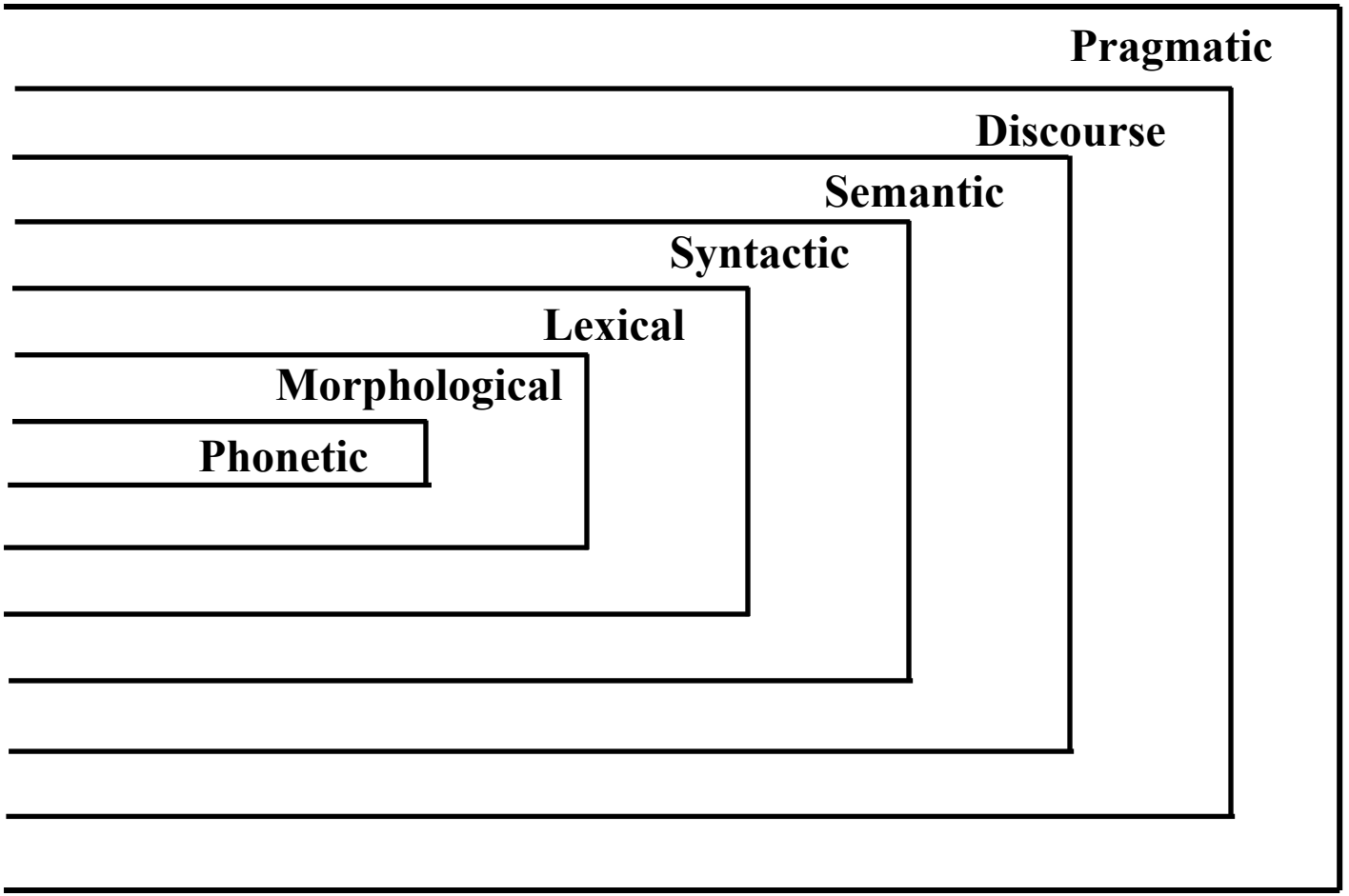
1. paraphrase an input text
2. translate it to another language or representation
3. answer questions about it
4. draw inferences from it
5. phrase the results in natural language



*Main emphasis in this course

Synchronic Model of Language

- The synchronic model postulates levels of language to understand the use of language at this point in time



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

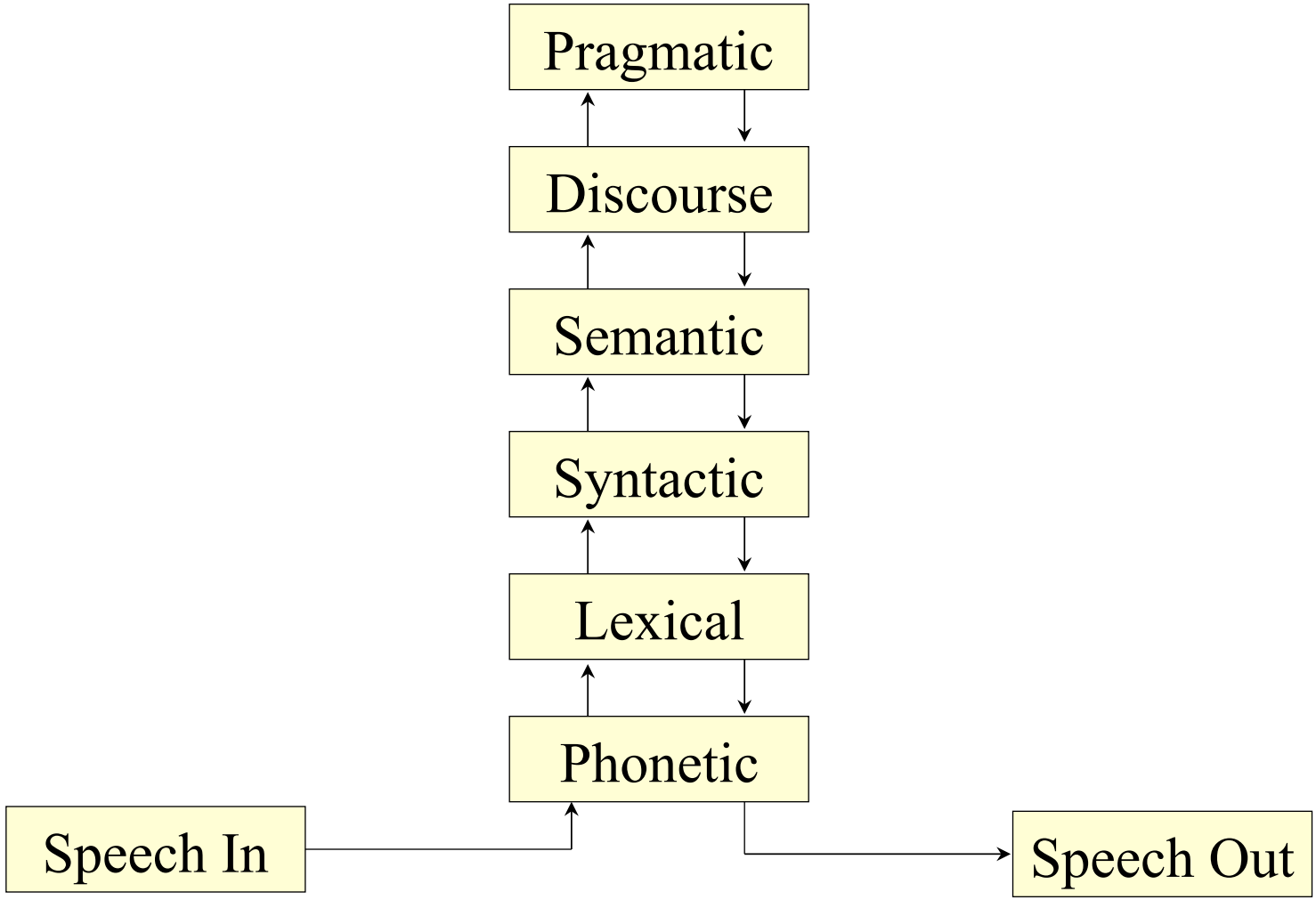
Ambiguous Newspaper Headlines

- 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
  - Google, Yahoo and Bing all have language translators
  - 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 summarizers: http://textsummarization.net/text-summarizer
    http://www.splitbrain.org/services/ots

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
    Question: *What year did Marco Polo travel to Asia?*
    Find the answer in text such as: *Marco polo divulged the truth after returning in 1292 from his travels, which included several months on Sumatra.*
  - Web sites like ask.com

NLP Application Areas

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