Natural Language Processing
IST 664/400 and CIS ~~563~~600/400

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

- 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

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
- Currently, NLP is providing these practical applications (yet still dreaming of the AI goals)

Need for NLP

- Huge amounts of data
  - Internet = at least 20 billions pages
  - Intranet
- Applications for processing large amounts of texts require NLP expertise

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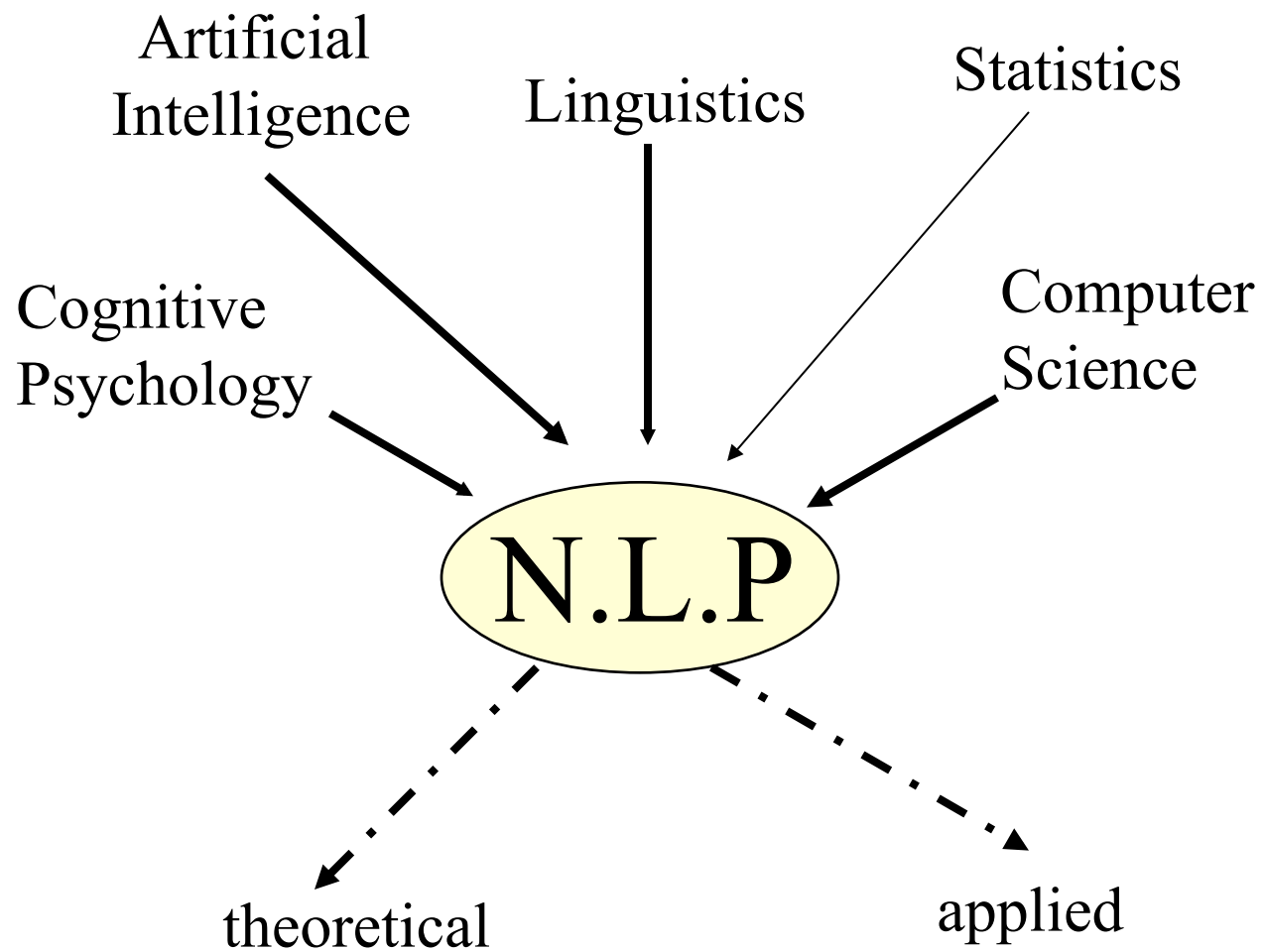   Extract useful information from
     resumes
Automatic summarization
     Condense 1 book into 1 page
Question answering
Knowledge acquisition
Text generations / dialogues

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
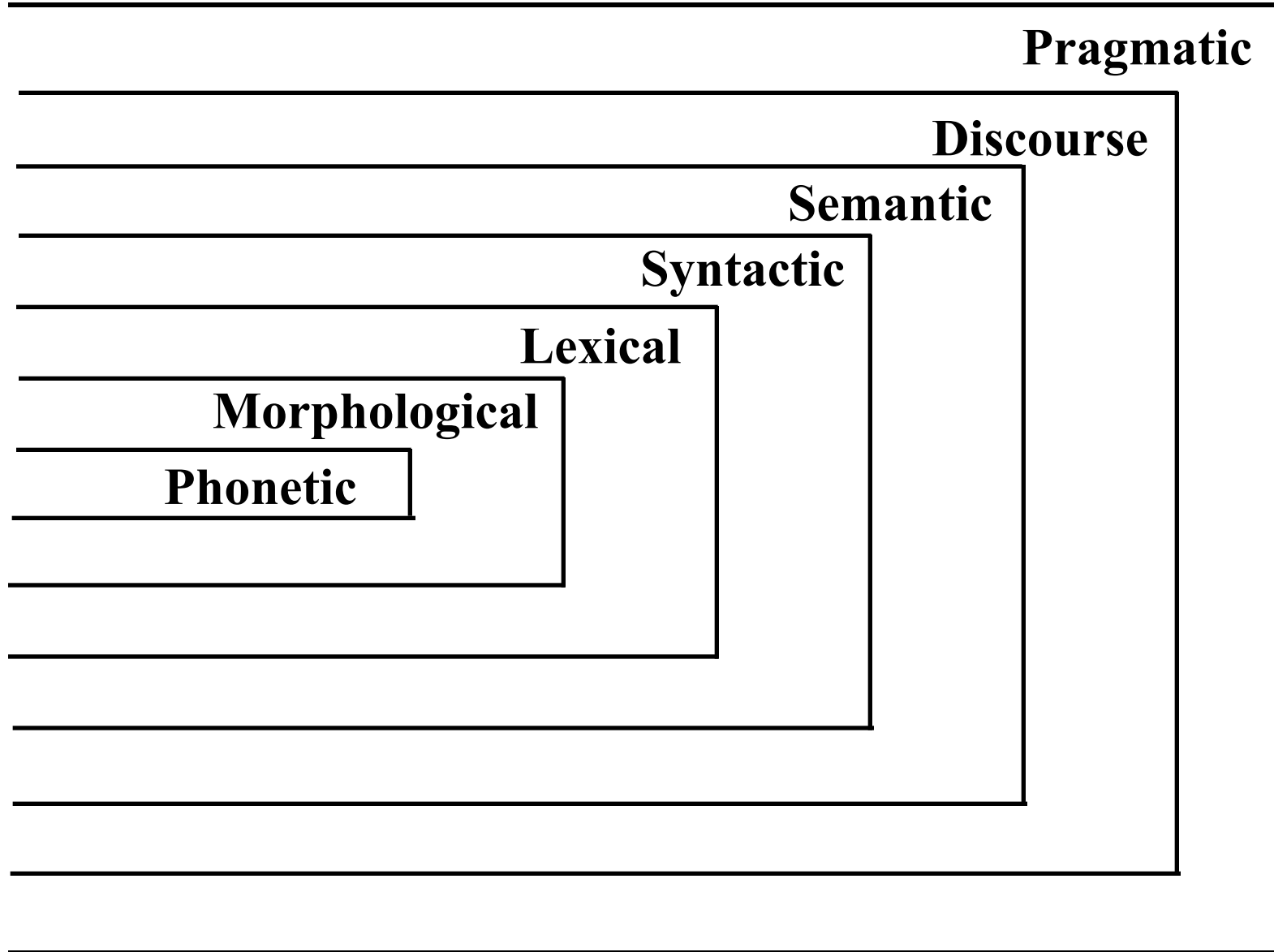 - frequencies, probabilities for detecting linguistic patterns

Two Sides of NLP: analysis and production

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 Production

Synchronic Model of Language

Pragmatic

Discourse

Semantic

Syntactic

Lexical

Morphological

Phonetic

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:

prefix $\longrightarrow$ *pre – registra – tion* $\longleftarrow$ suffix

stem/root
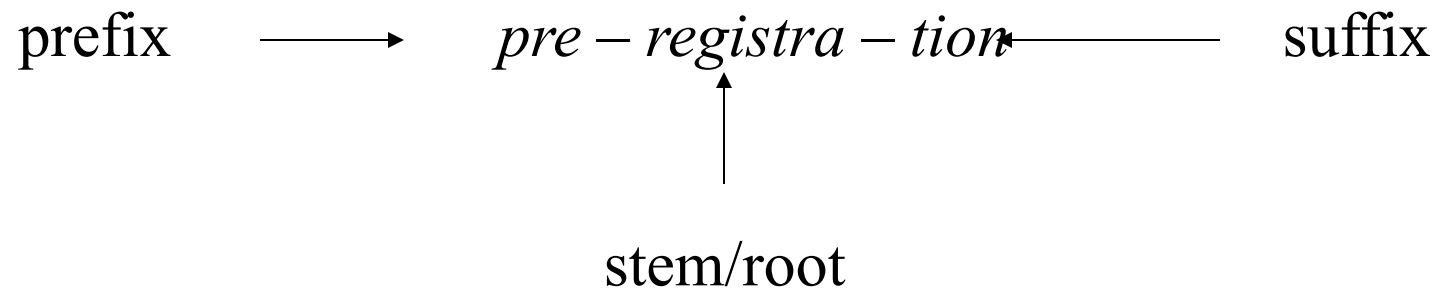
- What features do inflections reveal?

Verbs $\longrightarrow$ tense & number

Nouns $\longrightarrow$ single/plural

Adjectives $\longrightarrow$ 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

```
--------------------------------------------------------------------
(launch)

(a large, usually motor-driven boat used for carrying people
    on rivers, lakes harbors, etc.) →

((CLASS BOAT) (PROPERTIES (LARGE)
(PURPOSE
  (PREDICATION (CLASS CARRY) (OBJECT
    PEOPLE))))
--------------------------------------------------------------------
```

- requires a large, well-organized lexicon

Sample Lexical Entry Data Elements

Headword Pronunciation

  Part of Speech

  Inflection

  Inflectional Pronunciation
  Definition – Sense 1

  Usage
  Etymology

   Grammatical Code
   Idiom Definition
   National Varieties (Irish English)
   Region (South)
   Level & Attitude (Derogatory)
   Time & Frequency (Obsolete)
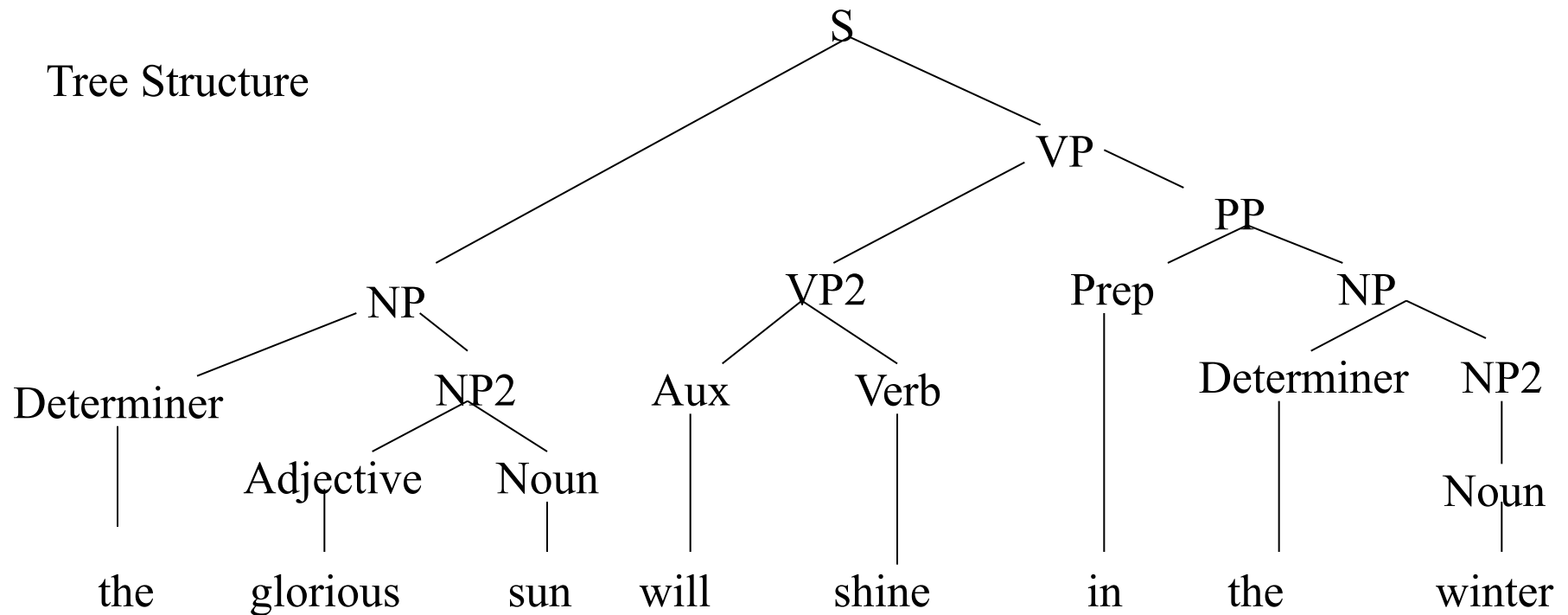   Semantic Restrictions (Animate, Liquid)
     On subject of verb
     On object of verb
   Subject Field (Agriculture, Economics)

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
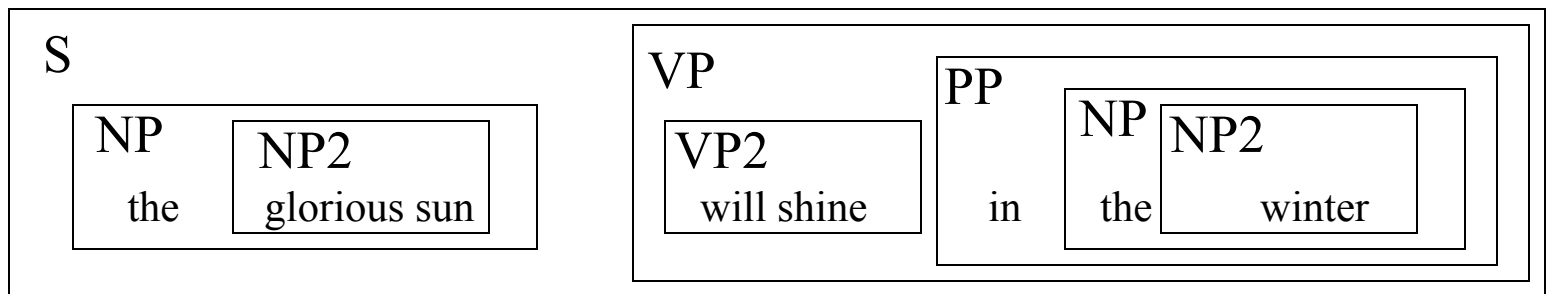
Tree Structure

```
                                        S
                                         \
                                          VP
                                           \
                                            PP
        NP              VP2          Prep      NP
 Determiner    NP2   Aux   Verb            Determiner  NP2
          Adjective  Noun                               Noun

    the    glorious   sun   will  shine      in    the    winter
```
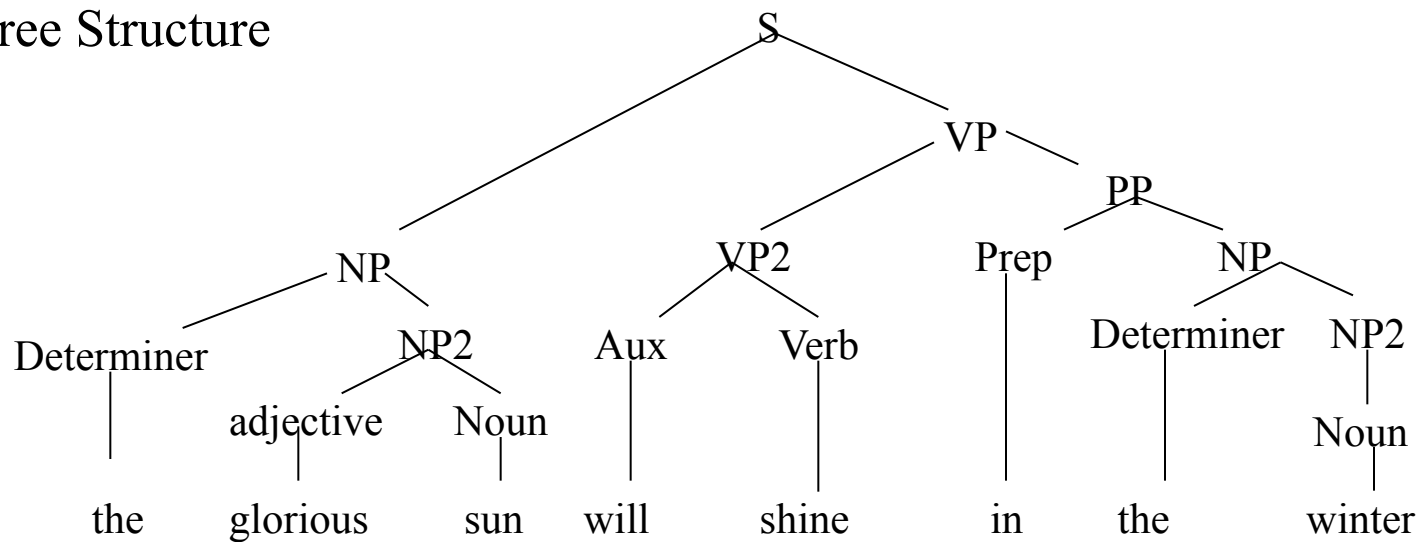
Bracketed text

[S [NP the [NP2 glorious sun]]

[VP [VP2 will shine] [PP in [NP the [NP2 winter]]]]]

Nested Boxes

S
NP
NP2
the   glorious sun

VP
VP2
will shine

PP
NP
NP2
in   the   winter

Tree Structure

S
VP
PP
NP
VP2
Prep
NP
Determiner
NP2
Aux   Verb
Determiner   NP2
adjective   Noun
Noun
the   glorious   sun   will   shine   in   the   winter

Sentence

Noun Phrase   Verb Phrase

Determiner   Noun   Verb   Noun Phrase

                                Determiner   Noun

The   cat   ate   the   mouse
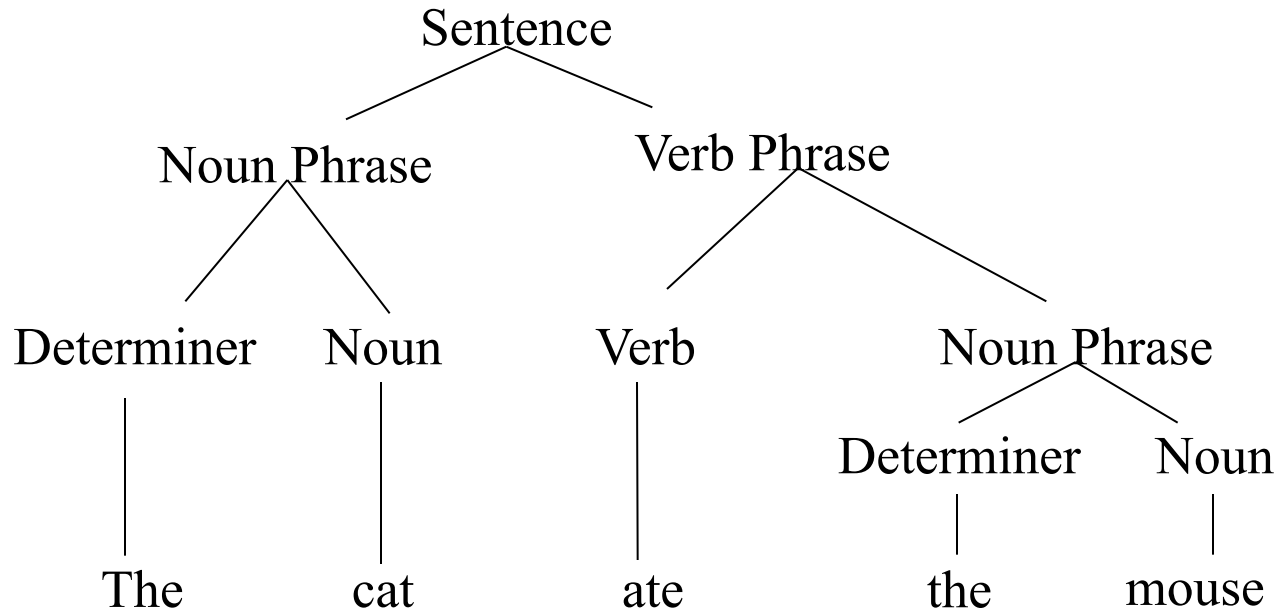
The phase structure rules underlying this analysis are as follows:

Sentence ⟶ Noun Phrase   Verb Phrase

Noun Phrase ⟶ Determiner   Noun

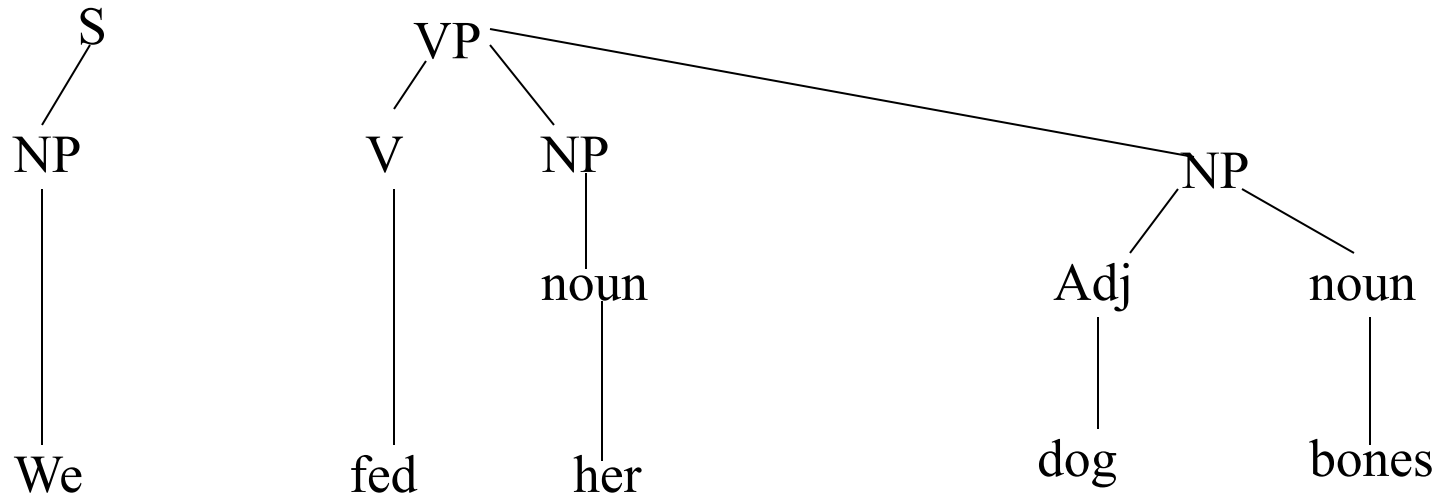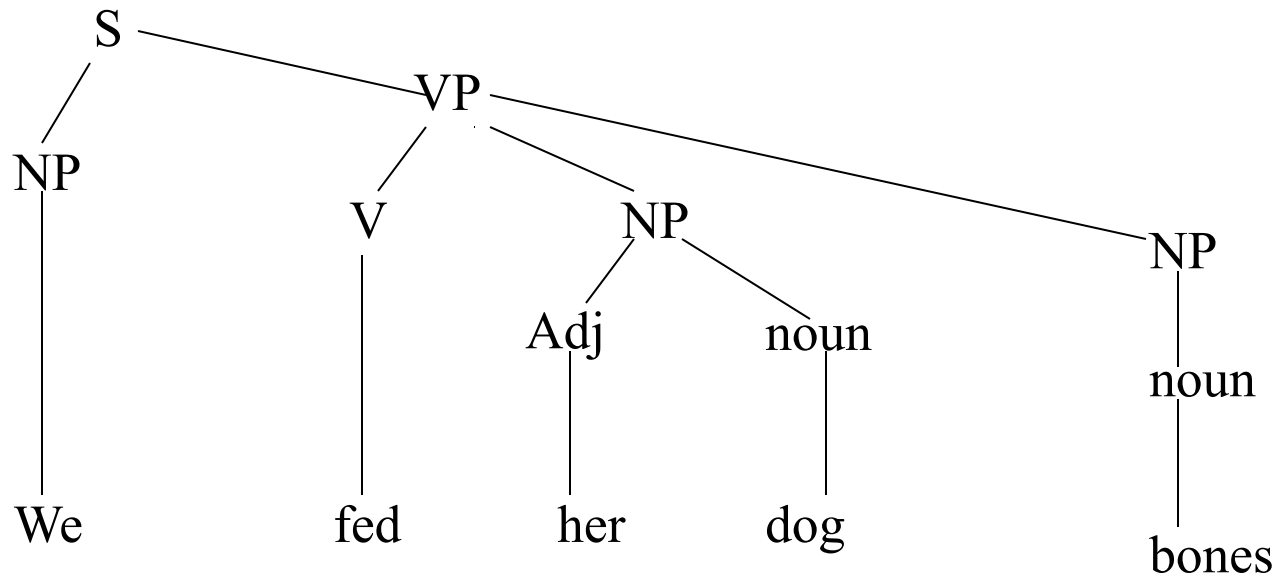Verb Phrase ⟶ Verb   Noun Phrase

Determiner = The

Noun = cat

Noun = mouse

Verb = ate   **Parsing a sentence using simple phrase structure rules**

We fed her dog bones

Tree 1:

S
- NP → We
- VP
  - V → fed
  - NP
    - Adj → her
    - noun → dog
  - NP
    - noun → bones

Tree 2:

S
- NP → We
- VP
  - V → fed
  - NP
    - noun → her
  - NP
    - Adj → dog
    - noun → 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.
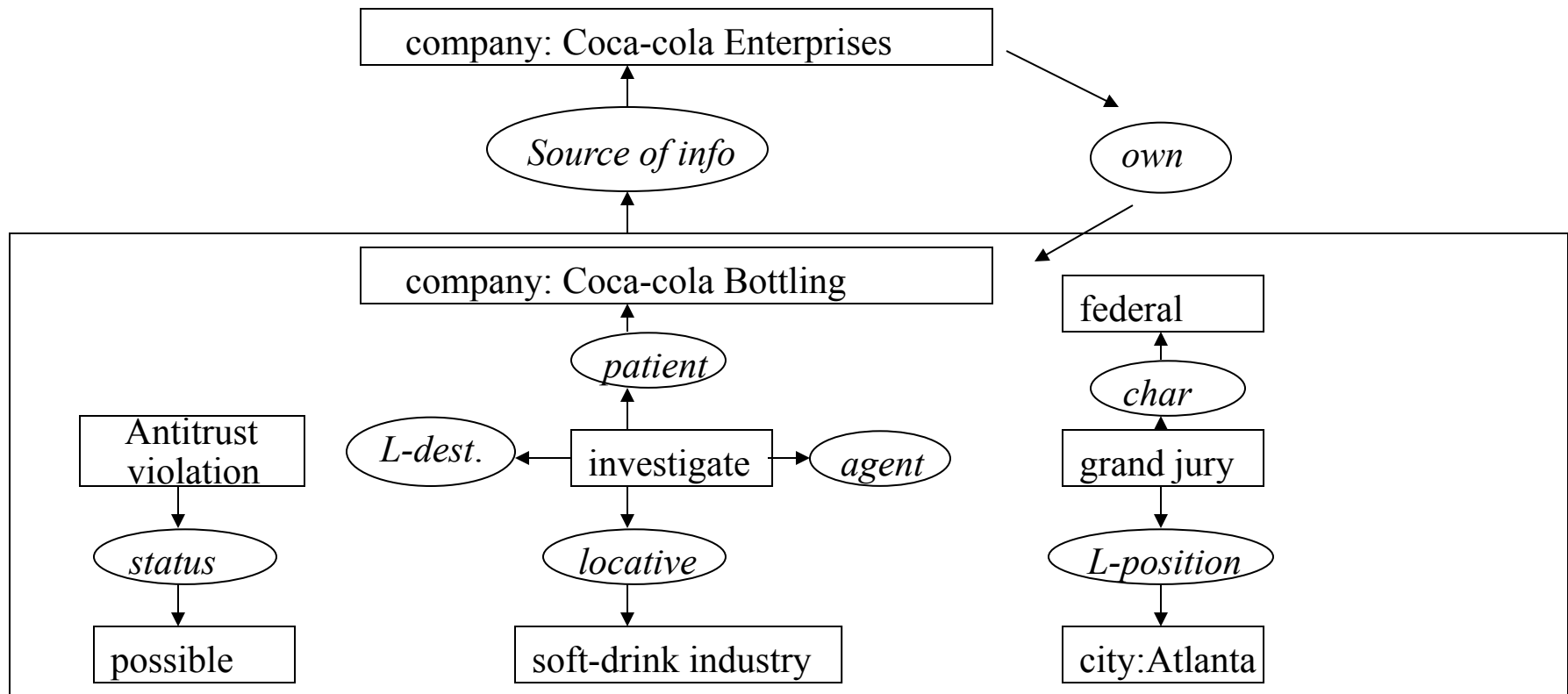
When Disney offered to pay Mr. Steinberg a premium for his shares, the New York investor didn't demand the company also pay a premium to other shareholders.

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

When [$_{payer}$ Disney] offered to [$_V$ pay] [$_{recipient}$ Mr. Steinberg] [$_{money}$ a premium] for [$_{commodity}$ 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.

company: Coca-cola Enterprises

Source of info

own

company: Coca-cola Bottling

patient

federal

char

Antitrust violation

L-dest.

investigate

agent

grand jury

status

locative

L-position

possible

soft-drink industry

city: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

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
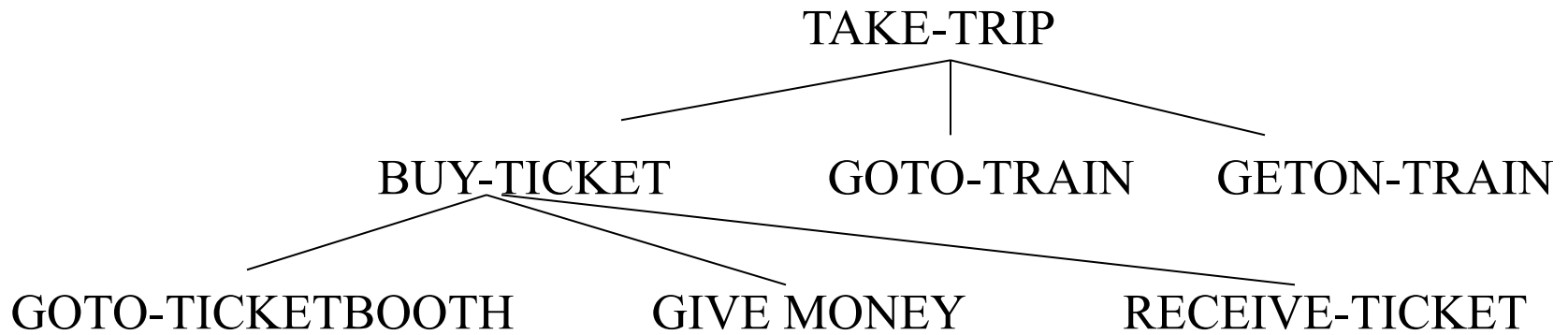
Sketch of a commonsense task plan to take a trip

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

Why is NLP so hard?

- **Seems pretty simple for humans**

    – Usually quite unaware of the complexity of the language tasks they perform so effortlessly

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**
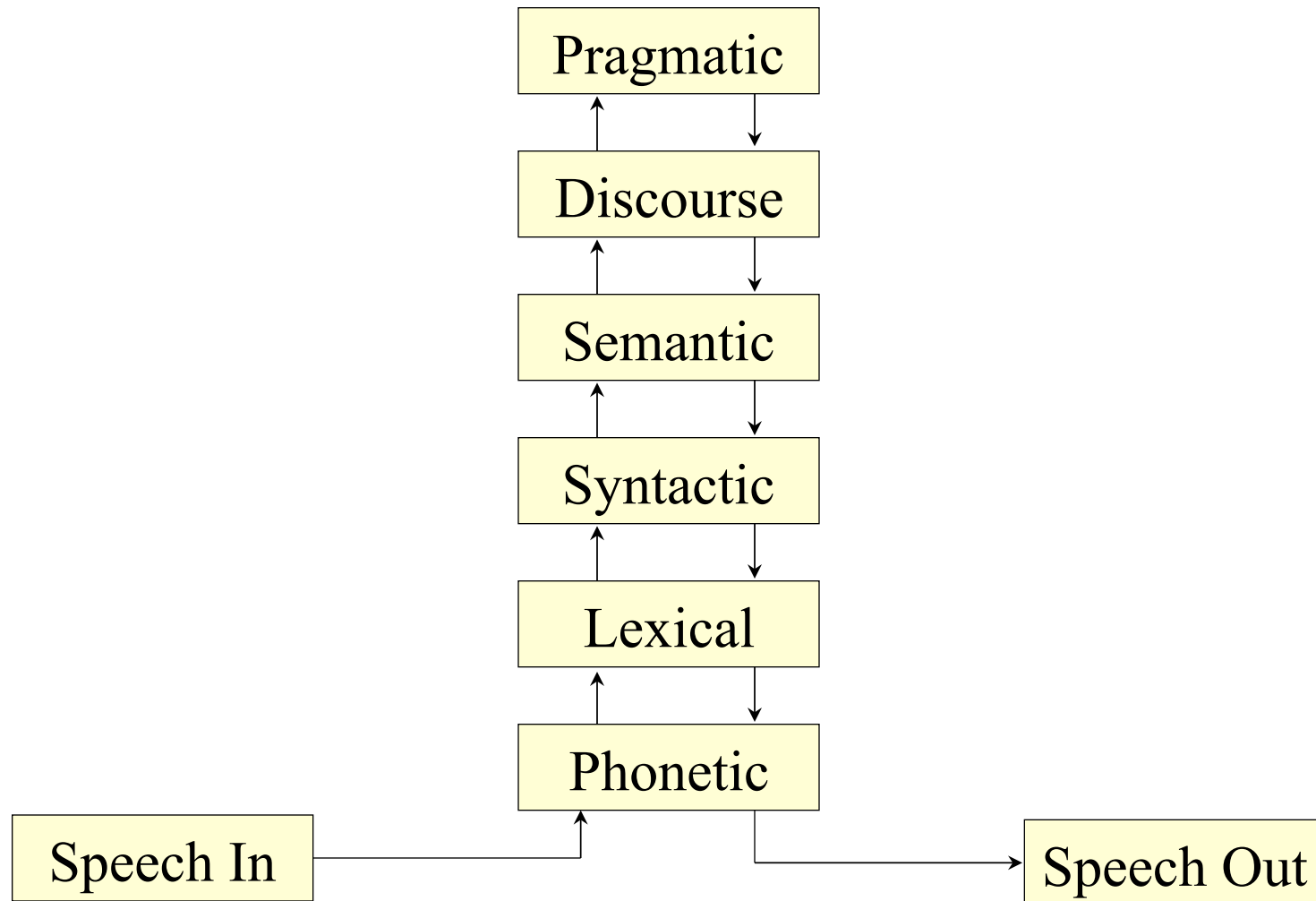 - *Mom said that when I visited Aunt Peggy in the hospital I should take her flowers.*

- **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

Pragmatic

Discourse

Semantic

Syntactic

Lexical

Phonetic

Speech In

Speech Out

NLP Application Areas

- <u>Machine Translation</u> – 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

- <u>Information Retrieval / Search Engines</u> – 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

- <u>Information Extraction / Text-mining</u> – populating a structured database with specific bits of information found in text
  - Competitive Intelligence analyses 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
- <u>Human-computer Interfaces</u> – interactive querying of databases
- <u>Summarization</u> – abstraction and condensation of text's major points
  - MEAD demo: http://tangra.si.umich.edu/clair/md/demo.cgi
  - Current systems select a set of significant sentences from the document as a summary

NLP Application Areas

- <u>Question & Answering Systems</u> – 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
  - Systems such as Start: http://start.csail.mit.edu/

- <u>Metadata Generation</u> – assignment of values for metadata elements in a particular standard, e.g. Dublin Core

Course Work Description

- Classroom sessions – every Tuesday
  - Lecture, In-class exercises, Discussions
- Lab sessions – every Thursday
  - Guided lab exercises, independent (group) exercises, resulting in discussion or discussion thread posting in the iLMS system
  - Use NLP programs provided in the Natural Language Toolkit (NLTK) to process text for analysis (see example)
- Two Homework Problems
  - Corpus Statistics
    - Use NLTK to collect frequencies of words and word pairs for text analysis; text may be collected
  - Entity Extraction (tentative)
    - Use NLTK for noun phrase analysis; use either for collecting entities from text (companies, people, etc.) or for topic detection

Course Work Description

- Final project
  - Choice of projects
    - Two (tentatively) will be defined for you, i.e. opinion or sentiment analysis,
      - Themes: business intelligence or artificial intelligence
    - or you may choose to define your own
    - Group work is encouraged but not required
- NLP applications
  - Groups will be assigned according to interest expressed in topics
  - Investigate the topic (read paper(s), search for examples, etc.) and give class presentation
- Undergraduates are required to do just one of the above