The field of Natural Language Processing (NLP) and spoken language processing (SLP) has applications such as:

A. Machine Translation
B. Information Retrieval and Search
C. Information Filtering and Text Categorization
D. Information Extraction
E. Input Methods
F. Spell Checking
G. Dictation
H. Command Interfaces
I. Question-Answering Systems
J. Tutorial Systems
K. Other Dialog Systems

Course Goals:

• to learn some useful concepts, models, algorithms, and techniques

• to practice some of the techniques used in building natural language systems

• to introduce or reinforce basic knowledge of:
  – probability
  – English grammar
  – formal language and automata theory
  – human-computer interaction
– machine learning and AI
– simple data structures
– basic programming skills
– the engineering issues involved in building systems

• to appreciate the complexities of language

Coverage
This class will cover the basics of NLP, including:

- representations of syntactic structure: PSG, bracketing, dependency, deep case
- parsing: FSM, CFG, PCFG; chart, unification, Viterbi search
- models of meaning: logic-based, case frames, semantic networks, connectionist
- knowledge representation: semantic networks, vector spaces, database semantics
- techniques for modeling spelling and morphology:
- architectures for integration: pipeline, integrated, blackboard, Bayesian
- learning methods: unsupervised, clustering, perceptron, decision trees, EM
- performance evaluation: objective measures, usability metrics
- human language vs. computer language: properties, uses
- user needs: embedded NLP, rival interface technologies

Textbooks and Readings:
This class will use two textbooks.


SLP should be available in the bookstore. MMML will be xeroxed off and distributed somehow.

It is important that you read the assigned chapters before each class.

There will also be articles chosen to present classic issues, to illustrate NL systems, or to present recent research results.

For more background, you may want to refer to

• Natural Language Understanding, 2nd edition, by James Allen, Benjamin-Cummings 1995

• Foundations of Statistical Natural Language Processing Christopher Manning and Hinrich Schütze, MIT Press, 1999,

Assignments:
The assignments are also important. There will be several types of assignments:

• thought assignments
• observation assignments
• computer assignments

Graduate students will do two additional assignments:

• leading in-class discussion
• writing a research proposal

Most assignments may be done either individually or in pairs. Some assignments will be done partly in class.
Late assignments will be penalized.

Tests:
There will probably be two tests, tentatively February 11 and March 13.
There will be a final examination, tentatively 13:00–15:45, Thursday, May 8.

Grading: The weighting will be approximately: Final Exam 35%, Assignments 30%, Tests 25%, Quizzes 5%, and Class Participation 5%.

Office Hours:
Fridays 13:15–14:15 in my office, or by appointment, or whenever the door is open. Come with any questions, or just to chat.
Tentative Schedule of Readings and Assignments

a. **Introduction**
a1. Overview of NLP Applications
   
   Read SLP1: “Introduction”

a2. Overview of the Course

b. **Words**
b1. Review of Simple Finite State Models
   
   Read SLP2: “Regular Expressions and Automata”

b2. Finite State Transducers
   
   Read SLP3 “Morphology and Finite-State Transducers”

b3. Pronunciation
   
   Read SLP4 (except 4.4,4.5) “Computational Phonology and Text-to-Speech”

b4. Basic Recognition Algorithms
   
   Read SLP5 “Probabilistic Models of Pronunciation and Spelling”

b5. Language Modeling
   
   Read SLP6 “N-gram Models of Syntax”

b6. Input Methods
b7. Hidden Markov Models
   
   Read SLP7 “HMMs and Speech Recognition”

   Assignment: transcribe one minute of a conversation

6,7

Assignment: train a part-of-speech tagger for Spanish

9

c. **Syntax**
c1. Motivation
   
   Read MMML “Why We Ascribe Structures to Sentences” (Ch.7+6.7)

c2. Some Complexities of English
   
   Read SLP8: “Word Classes”

9

c3. Part-of-Speech Tagging
   
   Assignment: train a part-of-speech tagger for Spanish

10

c4. English Grammar
   
   Read SLP9: “Context Free Grammars”
c5. Context-Free Parsing
   Read SLP10 “Parsing with Context-Free Grammars”
   Assignment: parse by hand and introspect on how
   Assignment: improve and test a grammar

   Assignment: improve and test a grammar

c6. Probabilistic Parsing
   Read SLP12 “Lexicalized and Probabilistic Parsing”

   Assignment: improve and test a grammar

  d. Systems and Semantics

  d1. Classic NLP
     Read MMML “Five or Six Classic NLP Systems” (Ch.7+6.7)
     Read “Experience with the Evaluation of Natural Language Question Answerers” (Tennant 1979)

  d2. Disambiguation
     Read “Parsing, How to” (Charniak 1983)
     Read “Introduction to ... Word Sense Disambiguation” (Ide and Veronis 1998)

  d3. Information Retrieval, Web Search
     Read SLP17: “Word Sense Disambiguation and Information Retrieval”
     Read “Topics in Information Retrieval” (Manning and Schuetze 1999), pp 529-543, 554-556
     Assignment: index creation with perl

  d4. Text Categorization
     Read “Learning to Classify Text” (Mitchell 1997) pp 180–184
     Assignment: message classification

  d5. Information Extraction
     Read discussion article “Fastus: A Cascaded Finite-state Transducer for Extracting Information from Natural-Language Text” (Hobbs, Appelt et al 1997)

  d6. Template-Filling; Database Interfaces

  d7. The Dream of General-Purpose Meaning Understanding
     Read SLP14: “Representing Meaning”
     Read MMML “AI and Connectionist Models of Meaning and Knowledge”
     [ch8,9,13]
     Read discussion article “KBMT ...”

  e. Spoken Language Systems
e1. **Speech Recognition and Understanding**
   
   Read discussion article “Hidden Understanding Models of Natural Language”  
   (Miller, R. Bobrow et al 1994)

e2. **Applications for Spoken Language Systems**

e3. **Dialog Management**
   
   Read SLP 19: Dialogue and Conversational Agents  
   Assignment: dialog design using VoiceXML

e4. **Natural Language Generation**

e5. **Usability Issues in Spoken Language Interfaces**

   
   Read “A Simple Rule for the Cooperative Timing of Utterances in Spoken Dialog” (N. Ward 1997)

e7. **Non-Verbal Communication and Multi-Modal Systems**

f. **Machine Translation**
   
   Read SLP21 “Machine Translation”  
   Assignment: translate by hand and introspect on the process

   Read discussion article “Integrating Knowledge Bases and Statistics in MT”, Knight et al. 1994)
   Read discussion article “Automatic Acquisition of Hierarchical Transduction Models” (Alshawi et al. 1998)

h. **Review**