CS 674: Natural Language Processing
Spring 2002

Thackara said, "Look – I'm a postmodernist. Is all this grammar really necessary?" "James," di Giovanni said, "we're talking on this phone *in grammar*." –John Walsh, "A Legend in His Own Mind", *The New Yorker*, Dec. 22 1997

Time and Place Upson 207, MW 11:15–12:05.

Instructor Lillian Lee. Upson 4152, `llee@cs.cornell.edu`. Office hours: W 2-3:30, or by appt.

Course homepage `http://courses.cs.cornell.edu/cs674/2002SP/`. This page contains links to handouts, the reaction essay reading list, references mentioned in class, online resources, and the course newsgroup, `cornell.class.cs674`, where announcements will be posted. Feel free to post relevant news (talk announcements, questions about papers, and so on).

Reference Materials

- Allen, *Natural Language Understanding*, 2^{nd} ed., 1995.
- Cover and Thomas, *Elements of Information Theory*, 1991.
- Durbin, Eddy, Krogh, and Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, 1998.
- Jelinek, *Statistical Methods for Speech Recognition*, 1997.
- Jurafsky and Martin, *Speech and Language Processing*, 2000.
- Manning and Schütze, *Foundations of Statistical Natural Language Processing*, 1999.
- Conference and workshop proceedings: ACL, COLING, EMNLP, NAACL, WVLC.
- Journal: *Computational Linguistics*

All items are on reserve at the Carpenter Hall Engineering and Computer Science Library except for *Computational Linguistics* (Olin library) and the Durbin et al. text.

Prerequisites Elementary computer science background (ability to analyze the space and time complexity of an algorithm, for example), elementary knowledge of probability (e.g., what conditional probability is), familiarity with context-free grammars, and some degree of mathematical maturity. You may wish to skim Chapter 2 of Allen or Chapter 3 of Manning and Schütze for background linguistic knowledge.

CS474 (Introduction to Natural Language Processing) is not a formal prerequisite; one could take CS674 and CS474 concurrently, or even potentially out of sequence. CS474 emphasizes parsing; CS674 will emphasize statistical approaches in general, although non-empirical subjects will also be covered; parsing will not be treated in great detail. (See syllabus on reverse for more information.)

Grading Policy

- Reaction essays: 30%. Due on the following Mondays: February 4, 11, 18, 25, and March 4. Each reaction essay should be one to two pages, based on a paper from the on-line reading list at `http://courses.cs.cornell.edu/cs674/2002SP/readings.html` (a printout of this webpage, which contains more specific instructions, will be distributed on the first day of class). The intent is to help you find a project topic, to acquaint you with some recent papers and research areas in the field, and to give you practice in reading research papers quickly and effectively.

- Final project: 60%, divided between the proposal (due noon March 15), a preliminary literature survey (due April 8), the project presentation (tentatively scheduled for April 29 and May 1), and the final project write-up (due 5pm May 13). You are to complete an independent project on some topic in natural language processing. Both programming and non-programming projects, including (thoughtful) surveys of papers, are welcome; all must include a careful writeup. Projects that attempt novel algorithms, implementations, proofs, analyses, and so on are especially encouraged – NLP is a fairly easy field to start doing research in!
- Participation: 10%. This includes class and/or office hours.

Academic Integrity You are encouraged to discuss issues, papers, projects, etc. with others. However, the work you submit, if not otherwise attributed, must be your own. In all your written work, including your reaction essays, be sure to credit your sources correctly; *it is not acceptable to copy out sentences or phrases from papers without citation.* Violating the Code of Academic Integrity may result and has resulted in failing the course. When in doubt about an issue, ask beforehand!

Syllabus

Jan 21	Introduction.
Jan 23	Feature-based CFG's.
Jan 28-30	Tree adjoining grammars (TAGs).
Feb 4	Statistical NLP: models and methods. **Reaction essay due**.
Feb 6	The statistics of language: Zipf's law.
Feb 11	Hidden Markov models (HMMs) I. **Reaction essay due**.
Feb 13	HMMs II.
Feb 18	HMMs and selectional restrictions: Abney and Light 1999. **Reaction essay due**.
Feb 20	Smoothing.
Feb 25	Introduction to information theory. **Reaction essay due**.
Feb 27	Class-based smoothing I.
Mar 4	Class-based smoothing II. **Reaction essay due**.
Mar 6	Global discourse phenomena.
Mar 11	The Grosz and Sidner 1986 discourse model.
Mar 13	Centering.
Mar 15 (Friday)	**Project proposals due at noon**.
Mar 25-27	Semantics: scope and ellipsis.
Apr 1-3	Word sense disambiguation.
Apr 8	Machine translation: paradigms. **Project preliminary literature survey due.**
Apr 10	Statistical machine translation: IBM models.
Apr 15	Statistical summarization: Knight and Marcu 2000.
Apr 17	Machine learning of language: theoretical results.
Apr 22	Grammar induction: the EM algorithm.
Apr 24	Statistics vs. linguistics: a case study in POS tagging
Apr 29-May 1	**Project presentations.** (tentative)
May 13	**Projects due at 5pm.**