CS674, Spring 2002
Project Guidelines, 2/27/02

Project grades are based on the following four parts:

- Proposal: due March 13 in class, but by noon March 15 is also acceptable. (If you take the latter date, hand your proposal in to Cindy Robinson, Upson 4146; please notify her by Thursday March 14 that you are taking this option by sending email to `cindy@cs.cornell.edu`, CC'ing `llee@cs.cornell.edu`)
- Preliminary literature survey: due April 8
- Presentations: April 29 and May 1 (tentative)
- Final report: due May 13

Projects should be focused and include careful evaluation. For example:

- **Bad:** I implemented a neural net to assign part-of-speech tags.
- **Good:** I studied various factors impacting the performance of neural nets as part-of-speech taggers, including the number of nodes, the training algorithm used, and the amount of training data available. I discovered that the amount of training data was the key factor, corroborating the claims of X.
- **Bad:** I read six papers on Martian discourse. A summary of each paper follows.
- **Good:** Current theories of Martian discourse provide different accounts of the Double Anaphoric Missing Negative (DAMN). I considered six different mechanisms proposed for dealing with DAMN, and devised a uniform approach incorporating the good points of the aforementioned six methods.

Your *proposal* should be 2-3 pages, and must include a preliminary list of related references. You should describe what problem you intend to consider, what has been previously proposed (this can be a bit sketchy, since you probably haven't done extensive background reading yet), what evaluation you intend to do, and what schedule you will follow for accomplishing your goals. Programming project proposals should list the resources you will use; survey proposals should include the questions you will "ask" of each paper you read.

For programming projects, the following resources are available (but this is not an exclusive list): Brill's part-of-speech tagger, the Brown corpus (POS-tagged), the BLLIP treebank, the Penn treebank, MUC4 text (POS-tagged), English newswire, Mandarin newswire, and WordNet. You are responsible for making sure that you can acquire and/or build what other resources you need – this should be reflected in your proposal schedule. The `corpora` mailing list linked off the course homepage is a good place to find out what datasets are available elsewhere.

Your *preliminary literature survey* should be approximately five pages long plus bibliography, and should represent some of the major papers on the problem you are examining.

The *final report* should be about ten pages plus appendices for programming projects and about fifteen pages plus bibliography for surveys. For other types of projects, such as the presentation of a new theoretical result, fewer pages may be appropriate. In all cases, a thorough discussion of related work is expected.

Sources of ideas include the recommended texts and readings, the list of some previous projects (see below) and the numerous conference proceedings that are on reserve at the Engineering Library: a useful activity is to pick a few and scan through the abstracts. Once you have a preliminary idea, feel free to see me at office hours or make an appointment to talk about it.

Here are the titles of some previous CS674 projects. At least one has led to a journal submission.

- Chinese word segmentation — A better way? (survey + human validations)
- Text-translation alignment for Polish and English using the cognate method (experiment)
- Segmentation of Chinese free text (experiment)
- Using WordNet to solve word analogies (experiment)
- Automated learning of natural language grammars (experiment)
- "A tangled thicket of examples": Syntax and semantics in ellipsis resolution (survey)
- On a class of deterministic TAGs (theoretical exploration)
- Information extraction with negation elimination (experiment)
- Stylometric Clustering: A comparative analysis of data-driven and syntactic features (experiment)
- Learning programming language parsing rules (experiment)
- CFGs and matrices and TAGs, oh my! (theoretical exploration)
- An analysis of automatically constructed thesauri (survey)
- Anaphora resolution algorithms on noun phrases (experiment)
- Word sense disambiguation in Malay (experiment)
- The importance of different parts of speech to word sense disambiguation (project)
- Part of speech tagging: An experiment in extending the unsupervised learner of disambiguation rules
- A linguistics bias case-based learning approach to anaphora resolution (experiment)
- Word sense disambiguation with semantic networks: Evaluating Resnik's semantic similarity measure in the disambiguation of free text (experiment)
- An unsupervised grammar inducer and parser based on part of speech n-Grams (experiment)
- Tree adjoining and combinatory categorial grammars as natural language formalisms, or How I Learned to Stop Worrying and Love the Lambda Calculus (theoretical exploration)
- Introducing explicit syntactic/semantic representation into a statistical machine translation system: An investigation of the use of linguistic information to assist the Candide MT system (theoretical exploration)