1. Parsing the Web: Large-Scale Syntactic Processing

Project Leader: **Dr. Stephen Clark** (Cambridge University, UK)

Want to build the fastest linguistically-motivated parser in the world? Want to process billions of words of text, and solve fundamental language processing tasks? Interested in large-scale and distributed computing? If you've answered yes to most or all of these questions then this is the project for you! Check out the parser and try an online demo here: http://svn.ask.it.usyd.edu.au/trac/candc/wiki. We plan to adapt this parser to web data and use it to analyze Wikipedia, in the process producing the parser of choice for anyone needing a syntactic analysis of text. Who might want to use such a parser? A good example is any search engine company interested in "Semantic Search", for example Microsoft and Powerset.

2. Low Development Cost, High Quality Speech Recognition for New Languages and Domains

Project Leader: **Dr. Dan Povey** (Microsoft Research, U.S.A.)

This project involves applying newly developed techniques based on factor analysis to speech recognition. These techniques are derived from speaker identification technology, and are based on expressing the models for a particular speaker and speech sound as a linear combination of factors specific to the speaker and the speech sound. The framework allows us to specify the speech models much more compactly. We intend to apply it to recognizing speech from languages where the amount of transcribed speech data is relatively small. Based on very positive initial results, we anticipate that this project will have a big impact on the speech recognition community. Relevant skills include linear algebra, C/C++, Unix shell, and statistical modeling. Already committed team members include researchers working in the USA, UK, Czech Republic and Canada.

3. Unsupervised Acquisition of Lexical Knowledge from N-Grams

Project Leader: **Dr. Dekang Lin** (Google Research, U.S.A.)

The overall performance of machine-learned NLP systems is often ultimately determined by the size of the training data rather than the learning algorithms themselves. The web undoubtedly offers the largest textual data set. Previous researches that use the web as the corpus have mostly relied on search engines to obtain the frequency counts and/or contexts of given phrases. Unfortunately, this is hopelessly inefficient when building large-scale lexical resources.

We propose to build a system for acquiring lexical knowledge from ngram counts of the web data. Since multiple occurrences of the same string are collapsed to a single one, the ngram data is considerably smaller than the original text. Since most lexical learning algorithms only collect data from small windows of text anyway, the ngram data can provide the necessary statistics needed for the learning tasks in a much more compact and efficient fashion. The students participating in the workshop will gain experience in advanced development of search and parallel processing and develop skills for data analysis and exploration.