First of all, **I prefer if you work in small groups of 2-3.** You learn more that way and usually wind up with better work. Just give me a brief statement of who did what, together with the final writeup.

Most students like to do some kind of dataset analysis for their final projects. I encourage this. There are three situations here:

1. **You have your own dataset.** You define your own analysis and carry it out.

2. **You don’t have your own dataset, but you are planning to collect some sort of data soon and want practice analyzing it.** In this situation, you can lay out the type of data you might encounter, and generate artificial data of this sort. Then you can look at the consequences of analyzing it in several different ways.

3. **You’d like to analyze someone else’s dataset.** This is totally fine. There are lots of datasets to analyze; here are a few:

   - The WARLPRI dataset in LANGUAGE (word order, case marking, and animacy)
   - The DATIVE dataset in LANGUAGE (English word order, phrasal weight, animacy, definiteness, lots of other things)
   - The ENGLISH dataset in LANGUAGE (word-recognition reaction times as a function of lots of stuff)
   - The Peterson & Barney vowel formant phonetic data
   - The “needs doing/needs to be done” dataset from Doyle & Levy 2008

Here are some examples of the kinds of final projects that people have done in the past/are good candidate ideas:
• Analyze binary or quantitative speaker felicity/grammaticality judgments for some set of constructions. Use ANOVAs and/or mixed-effects models to investigate.

• Construct models of speaker choice in some word order variation/similar study, investigating the effects of different variables. Mixed-effects logit models are a good candidate.

• What factors affect reaction times in some type of lexical processing/sentence processing? Use ANOVAs and/or mixed effects models to investigate.

• Can we gain anything from using non-Gaussian error distributions to analyze reaction-time data in language processing? Use generalized linear models (perhaps mixed-effects) to investigate.

• Compare Bayesian and frequentist inference procedures for some empirical domain of language.

• Investigate the relationship between frequency/predictability and word length in a corpus (shades of Zipf’s law).

• Couch some mini-problem in language acquisition in terms of Bayesian inference (hypothesis testing, model selection, prediction, and/or parameter estimation), and construct a model of it. Discuss the effect of different priors on the inferences that result.

I don’t have strict length requirements, but a good project will tend to be between 4 and 8 pages of text, plus code & graphs. Make sure you give me enough plain-English context that I understand the problem at a conceptual level.