Ideas for Final Papers

We have basically covered two topics in this course: Regression and Language Modeling.

Regression

Conduct an original regression analysis on a new dataset you construct or on an existing one.

1. Formulate a hypothesis to investigate. Build a dataset from a corpus, an experiment, or other material and see if the regression analysis supports the hypothesis.

2. Many regression analyses in experimental linguistics prior to 2005 or so did not make use of mixed effects modeling. Find such a dataset, run the mixed effects modeling regression and compare the results.

Language Modeling

Find a corpus you are interested in from the Linguistic Data Consortium Catalog. Alternatively you can use one or more of the datasets from the Pautomac or SPiCe challenges. For phonology, your corpus could be a list of word forms such as those found in a dictionary. Ask me for potential sources.

You can also generate artificial datasets from a known stochastic language (given by some probabilistic grammar) and see how well different methods can recover the known target. See Avcu et al 2017 or Nelson et al 2020 for examples in this vein.

Then using off the shelf software you can train and test at least two different LMs and compare the results.

1. n-gram models (with or without smoothing) using the openfst ngram library

2. HMMs using NLTK or some other software

3. Alegria using the GI toolkit

4. Factored deterministic regular stochastic languages (Huteng Dai has an implementation)

5. PCFGs using NLTK or other software

6. RNNs or LSTMs using pyTorch or Tensorflow, or some other NN toolkits

Other ideas

I am open to other ideas. Generally I am sympathetic to anything involving formal grammars (such as comparing NNs to formal grammars).