Jack McGreevy
Professor Kraus
LING 144

Project proposal

Introduction:
The question we are investigating in our project is whether lexical richness in prepared speech is greater than that of spontaneous speech. We will be using English data to investigate this question. The two corpora we will be using are the Switchboard Corpus and the C-Span Inaugural Address Corpus, for spontaneous and prepared speech respectively. We will be using the Type/Token Ratio (TTR) and word frequency statistics to measure lexical richness.

Background:
While much work has been done using the TTR as a measure of lexical richness, there has been no major work comparing TTRs of spontaneous vs prepared speech. We hypothesize that the TTR will be higher in prepared speech under the assumption that deliberate planning of speech enables speakers to exceed their normal vocabulary size.

Methods:
We plan to collect samples from the Switchboard Corpus and C-Span Inaugural Speech Corpus. Ideally, we would use longer conversations from the Switchboard Corpus in order to match the lengthier samples from the C-Span Corpus. The language variety is English for both corpora, though the context of use is markedly different between the two. For the Switchboard Corpus, participants were given topics to discuss with an interlocutor over the phone. Meanwhile, the C-Span Corpus contains inaugural speeches given by U.S. presidents. The incongruence in the contexts of use does spell some problems for our investigation, because conversations on a prompted topic are perhaps not the most representative example of spontaneous speech, just as presidential speeches are not the most representative example of prepared speech. That being said, comparing the lexical richness found in these two corpora would provide some preliminary data for the comparison of spontaneous vs prepared speech. Our dataset will contain long chains of sentences to analyze, broken across two interlocutors in the case of the Switchboard Corpus. Finding the TTR will be simple enough using the nltk module. We can save the data for each corpus into a string and then use the lower() and word_tokenize() method to tokenize the data. The token count will be the length of that new string. Then a regular expression can be compiled to search for the number of words in the string; this will be our type count.
One additional measure of lexical richness is the relative frequency of the words employed. Greater lexical richness should correspond to a higher number of low-frequency words. It's not clear yet how we'll sample our datasets, but we intend to run statistics on the word frequencies found in both corpora.