Ling 236 Homework #4

Due 13 February 2002

1. Consider the grammar for Adam I in the Suppes (1970) article. Suppes notes that the major reason for the only limitedly good fit of the grammar to the model is from the use of the NP $\rightarrow$ NP NP rule, which ends up badly underestimating the number of times you would see N N (see Table 1). However, this aspect of the grammar can be changed. Note that Suppes includes a special rule NP $\rightarrow$ AdjP N in the grammar, even though sequences like A N could have been generated using the NP $\rightarrow$ NP NP rule. Try making a change to the grammar that will improve its estimates (it isn't important that you succeed, providing that you do the calculations below correctly and present the results). Work out maximum likelihood estimates for the rules in your new grammar, and then the predicted ('theoretical') frequencies of each form (recall that there are 2434 total noun phrases in the corpus in Table I). This isn't quite as difficult as Suppes makes it look! Adopt the same simplifying assumption that he did under which each terminal sequence of parts of speech is given its 'simplest' analysis under the grammar. You should then be able to give an analysis to each string, and to count how often NP and any other nonterminals you use (i.e., also AdjP in Suppes' grammar) appear in the grammar, and how often they are rewritten in different ways. This will give maximimum likelihood estimates for the rules, and will allow you to calculate predicted frequencies for different strings of parts of speech over a corpus this size.

   (a) Give your grammar as a simple probabilistic CFG.

   (b) Show a table corresponding to Table I for your data.

   (c) Work out the goodness of fit of the grammar to the data using a chi-square test. (To work out the number of degrees of freedom to use, pay attention in class, and/or read carefully p. 112. The number of degrees of freedom is the number of cells in the table minus the number of parameters set from the data minus 1.) Is your grammar better than Suppes' grammar?

   (d) How much of the probability mass of the grammar is given to strings that were not observed at all in the data of Table I?

2. Load the catheter distance data from:

   http://nlp.stanford.edu/~manning/courses/ling236/handouts/ling236-statnotes5.html

   Do a simple linear regression between weight and catheter distance, and report the fitted values for slope, intercept, and the regression coefficient.

3. Keller provides the following results on mean acceptability with multiple constraint violations:

	Mean acceptability	
Number of violations	Soft constraint	Hard constraint
0	0.0865	0.0382
1	0.0630	−0.2830
2	−0.0146	−0.3814
3	−0.0628	−0.4486

Fit a simple linear regression line (separately) to both the soft and hard constraint relative acceptibilities as constraint violations increase. Give the fitted values for the slope and intercept, and the regression coefficient. Do the relationships appear to be linear? Comment (informally) on how reasonable a thing to do this attempt at linear regression modeling is.