Ling 289 Homework #3

Due: Wed 17 Oct 2007

1. There's a small subfield of humanities computing that attempts to use statistics from texts for authorship attribution. Commonly this involves looking at the frequencies of fairly common function words or function word sequences, which are assumed to be fairly consistently used by an author across texts, but on which another author might have idiosyncratically different usage. This problem considers texts of Jane Austen. In particular, when she died, she left a partly finished novel (*Sanditon I*), which was completed by a fan, attempting to write in her style (*Sanditon II*), and the composite was then published. The table gives the relative frequency of the word *a* preceded by and not preceded by *such* (i.e., the latter is the sum of the counts for all other words *X a*, the word *and* followed by or not followed by *I* and the word *the* preceded by or not preceded by *on* in the two halves of *Sanditon* and two other Jane Austen novels. Was Austen consistent in these habits of style from one work to another? Did her imitator successfully copy these aspects of her style? What evidence can you draw in favor or against it being a successful copy. (You should be looking to use a significance test from class. . . .)

Word sequence	*Sense and Sensibility*	*Emma*	*Sanditon I*	*Sanditon II*
such a	14	16	8	2
¬*such a*	133	180	93	81
and I	12	14	12	1
and ¬*I*	241	285	139	153
on the	11	6	8	17
¬*on the*	259	265	221	204

2. In Wasow (1997), in the section on Collocations and HNPS, one finds the following data (graphed in Figure 12 of the paper):

	HNPS	not	
Transparent collocations	90	102	192
Non-collocations	59	329	388

(i) Confirm the figure given in the paper for the chi-square test. (ii) Calculate the odds ratio for heavy NP shift for transparent collocations versus non-collocations. That is, how many times larger are the odds of HNPS with a transparent collocation than with a non-collocation?

3. 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 → 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 → AdjP N in the grammar, even though sequences like A N could have been generated using the NP → 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?

The easiest way to do this problem is probably by adapting the spreadsheet I used in class.