Ling 235 Homework #2

Due Wednesday, January 26, 2005

1. A very famous early example of probabilities done over text was A. A. Markov's counts of consonants and vowels and their sequencing. Here's the data from A. A. Markov's count of exactly 20000 letters of the first part of A. Pushkin's novel *Eugène Onégin* – in Russian. We assume that he just ignored spaces.

   Markov defined two events, which he (unmnemonically) called:

   E: a vowel occurred
   F: a consonant occurred

   He then did counts to get the following probabilities, all as relative frequency estimates:

   p estimate of probability that the next letter is a vowel
   p_1 estimate of probability that the next letter is a vowel, given that the preceding letter was a vowel
   p_2 estimate of probability that the next letter is a vowel, given that the preceding letter was a consonant
   q estimate of probability that the next letter is a consonant
   q_1 estimate of probability that the next letter is a consonant, given that the preceding letter was a vowel
   q_2 estimate of probability that the next letter is a consonant, given that the preceding letter was a consonant

   In the text, Markov found 1104 instances of vowel-vowel sequences, while the total number of vowels in the text was 8638. Assume for convenience that we regard the text as circular, so that the very last letter counts as the preceding letter for the very first letter.

   Using this information:

   (a) Express the quantities p, p_1, p_2, q, q_1, and q_2 in a more mnemonic, intelligible notation (e.g., with events and conditional probabilities).

   (b) What is the value of p, p_1, and p_2? By how much do p_1 and p_2 differ? What about for q, q_1 and q_2?

   (c) What simple fact of language does the differences between p and p_1 and p_2 (or between q, q_1, and q_2) capture?

(d) Suppose we see a short Russian text with CV structure:

CVCCVCCVCV

i. What is its probability in a unigraph model (using just p and q)?
ii. What is its probability in a digraph model (using p_1, p_2, q_1, and q_2)?
iii. Assuming that this is a representative text of Russian, which appears to be a better model of Russian CV structure? Explain why.
iv. What is the likelihood ratio between the two models?

2. This problem uses the analysis of Weiner and Labov in "Constraints on the agentless passive".

Recall that a Varbul weight for a factor is a probability (between 0 and 1 inclusive), which is turned into an *odds* to work out the aggregate model prediction in a Varbul model (see the formula given on the bottom of page 39). If a probability/Varbrul weight is p then you can convert between the p and the odds o as follows:

$$o = \frac{p}{1-p} \qquad\qquad p = \frac{o}{o+1}$$

The odds of passivization are the product of the odds of the base passivization rate (p_0) and the odds in turn of each additional factor that is true for a data instance.

Using the Varbrul weights given in the "Full Analysis" column of Tables 10 and 11, calculate the estimated probability of passivization for an agentless transitive clause in the model they build if:

(a) the speaker is an adult working-class white male speaking carefully, there was no passive in the previous five clauses, and the logical object was the surface subject of the previous two clauses.

(b) the speaker is an adolescent working-class black male speaking carefully, there was no passive in the previous five clauses, and the last NP coreferential with the logical object was two clauses ago.

(c) Also calculate the probability of passivization for (a) and (b) in the model given in column 14 of Table 11, where parallel surface subject has been eliminated. Compare these probabilities with those for the "full analysis" model, and comment on the implications for the importance of varying factors in the model. In particular, why do the weights associated with Given status change the way they do when the parallel surface subject constraint is removed?

Be sure to show your work in the calculations above explicitly!