Ling 289 Homework #1

Due Wed 3 October 2007

These are meant to be not-too-difficult questions to help you understand important concepts. . . .

[Conceptual side issue: We're meant to have only covered probabilities, not statistics, so far. That's why statistics courses like to start with coins and dice, since they have natural probabilities. Unfortunately, things in language don't (AFAIK). In the questions below, we will use simple relative frequencies (that is, "maximum likelihood point estimates") to get from statistics to probabilities. But you should be aware that this is happening, and that there's a lot more possible subtlety in going between a frequency probability model and counts from a data sample.]

1. Redo and finish the tregex/R example with *that* deletion, and briefly write up the results.
2. Combinatorics and probability:
   (a) Assume that the only possible vowels in human languages are /a/, /e/, /i/, /o/, /u/, /ə/. With no other restrictions, how many 4 vowel systems are possible?
   (b) Assuming that each of those (4 vowel) inventories is equiprobable, what are the chances that a language will have /a/ in its inventory?
   (c) What are the chances that a 4 vowel language will have /a/ and /u/ in its inventory?
   (d) Really, things aren't equiprobable (and some languages have more than 6 vowels!). According to the UCLA Phonological Segment Inventory Database sample, $P(\text{lg has /i/}) = 0.87$. Assuming that the other vowels are equiprobable, and that their occurrence is independent of /i/ and each other, reanswer questions (b) and (c) (for 4 vowel languages).
3. Suppose (in a certain genre of text) the probability that a word is a noun is 0.4, and the probability that a word is a verb is 0.2. Suppose also that the probability that the word is of Latin origin is 0.3.
   (a) Given just the above information what are the bounds on the minimum and maximum possible probability of a random word being a latinate noun.
   (b) Assuming that part of speech and latinate origin are independent, what is the probability that a random word is a latinate noun.
   (c) Suppose the probability of latinate nouns is actually 0.15. What is the probability that a random word is a noun not of Latin origin?
4. Which of these are true?
   (a) $P(A|B,C) \leq P(A|C)$
   (b) $P(A,B|C) \leq P(A|C)$
   (c) $P(A \cup B) \geq P(A) + P(B) - 1$

5. 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 an event E
   p_1 estimate of probability that the next letter is an event E, given that the preceding letter was an event E
   p_2 estimate of probability that the next letter is an event E, given that the preceding letter was an event F
   q estimate of probability that the next letter is an event F
   q_1 estimate of probability that the next letter is an event F, given that the preceding letter was an event E
   q_2 estimate of probability that the next letter is an event F, given that the preceding letter was an event F

   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?