Psych156A/ Ling150
Homework 4: Words and Categories

1) Sigmund von Hacklestein among the Guins
(a) Here are some words Sigmund found in his initial survey of the Guin language that begin with "pen":

pengo, pentumbry, pentabula, penkist, pennor, pentale, penlig, pencrom

What is the transitional probability of the word "penlig", given this data? What about for the syllable sequence "penta"? You may write your answer as a ratio (ex: 2/3) rather than a decimal (ex: 0.66666…). Make sure you show how you derived this transitional probability. [4 pts]

(b) Sigmund has just heard about the idea of algebraic learning for word segmentation, and wants to see how it would work on his Guin data. He also heard about the Unique Stress Constraint, and that it may be helpful for segmenting words.

(i) Briefly explain to Sigmund what the Unique Stress Constraint states about words. [1 pt]

(ii) Here are some Guin words Sigmund has:
péngo, áz, péncrom, penlíg (Note: pén = the "pen" syllable has stress, etc.)

Between what syllables would an algebraic learner with knowledge of the Unique Stress Constraint posit word boundaries in the following syllable sequence? (You may feel free to justify your answer with an explanation, if you would like to receive partial credit in case your answer is incorrect.) [2 pts]

áz pén go zú má

(c) Sigmund decided to try out the transitional probability learning strategy on some of his Guin data, just to see how it does. He found that this strategy was able to identify 100 words from a particular stream of Guin speech. When he asked his Guin friends about these words, he found out that only 40 out of the 100 words identified were actually real Guin words. In addition, his Guin friends told him that he should have found 200 *more* real Guin words than he actually did identify with this strategy.

(i) How many words did the transitional probability strategy correctly identify as words? [1 pt]

(ii) How many Guin words total should Sigmund have found? [1 pt]

(iii) Use your answers above to calculate the precision and recall for Sigmund's transitional probability approach on that set of Guin data. (You can just give the ratios, rather than working out the exact decimal answer (ex: 2/3, rather than 0.6666…)). Be sure to show how you calculated them. [4 pts]

(d) Sigmund was also impressed by the frequent frames approach to word categorization that Mintz (2003) used. He decided to try it out with a stream of data he has from Guin. Here is the data stream, segmented into words:

"az pengo mu pino var peri lootem berg az bleeter mu pino fez trutti peri mono berg az bering mu blorno set beegle trem az porto mu pino var az felgo mu peri freggo berg"

(i) Based on this data, what are the two most frequent frames in the data? [2 pts]
(ii) What words would each of the two frames you identified in (d1) cluster together? [4 pts]

2) Sigmund was very impressed by Saffran et al. (1996)'s experiment with 8 month olds children. But he's a little unclear why it was necessary to test children on real words vs. part-words in the second experiment in order to show that children can segment words based on their transitional probabilities. Explain to Sigmund why this second experiment was necessary. (You may want to consider how children could have performed as they did in the first experiment without being able to segment words based on the transitional probabilities between syllables of those words.) [3 pts]

3) Sigmund thought transitional probabilities sounded like a great idea for word segmentation, and can't believe that this strategy failed to do well on realistic child-directed speech data.
(a) What property of the data set caused the model to fail in Gambell & Yang's (2006) study? [1 pt]
(b) Why did this property cause the transitional probability model to fail? (You may want to consider how a transitional probability learner decides that there should be a word boundary between two syllables. Also, remember that an example will help make your point clearer.) [3 pts]

4) Sigmund liked the idea that a word's category (like noun or verb) is really just a description of the way that word can be used in the language. Given the following contexts, tell Sigmund whether you think the novel word in each example is a noun (like "goblin"), verb (like "sing"), adjective (like "hot"), or adverb (like "dreamily"). Be sure to explain why you think so. (You might find it helpful to substitute words you know in place of the novel words, and see which ones fit best.). [2 pts each]

(i) Is Sir Didymus REKKing the Bog of Eternal Stench? (REKK = ???)
(ii) That's the NOOGest bog I've ever seen. (NOOG = ???)
(iii) They had to beware the BOFT's tricks. (BOFT = ???)