# LSA.308 homework #4

Due: Tuesday 24 July, 2007

The Goldwater, Griffiths, and Johnson (2006, 2007) unigram model involves the following parameters (collectively termed $\theta$):

- $h$ The parameter defining the geometric probability distribution over utterance length $P(L)$ (so that $P(L) = (1-h)^{L-1}h$)[1]
- $\alpha$ The concentration parameter defining the probability of the next word being novel (see GGJ 2007, page 5)
- $p_\#$ The parameter defining the geometric probability distribution over word length
- $V$ The number of phonemes in the language

The utterance `ba.di.ba` has the likelihood

$$P(\texttt{ba.di.ba}|\theta) = \overbrace{(1-h)^2 h}^{\text{P(L=3)}} \overbrace{\frac{\alpha}{\alpha}\frac{\alpha}{1+\alpha}\frac{2}{2+\alpha}}^{P(\text{new,new,old|L})} \overbrace{(1-p_\#)p_\#\frac{1}{V^2}}^{P(w_1|\text{new})} \overbrace{(1-p_\#)p_\#\frac{1}{V^2}}^{P(w_2|\text{new})} \overbrace{\frac{1}{2}}^{P(w_3|\text{old})}$$

(Note that this ignores the possibility of having multiple distinct lexical entries with the same phonemic form `ba`—this oversimplification is OK for the purposes of this homework.)

1. Calculate the likelihood of the utterance `ba.diba` and use it to calculate the likelihood ratio

$$\frac{P(\texttt{ba.diba})}{P(\texttt{ba.di.ba})}$$

   (we did this in class). You can think of this likelihood ratio as a posterior belief ratio for two alternative lexicons—{`ba,di`} and {`ba,diba`}.

2. Imagine that the (unsegmented) utterance were extended to become `badibaba`. What are the likelihoods of the segmented utterances `ba.di.ba.ba` and `ba.diba.ba`? What is the effect of adding this additional `ba` on the likelihood ratio for the two lexicons listed in problem 1?

[1] Actually, the GGJ model also puts a probability distribution over $h$, but we will ignore this detail here.