CA446: Statistical Machine Translation
Exercises
Thursday, 6th March 2014

1. Given the following frequency counts

   - the green 1748
   - the green paper 801
   - the green group 640
   - the green light 110
   - the green party 27
   - the green ecu 21

   Calculate $p(light|the\ green)$

2. Consider the following data:

   -

Count	Count of counts
1	5000
2	1600
3	800
4	500
5	300

   -

Count	Bigram
4	beer drinker
4	beer lover
2	beer glass

   Readjust the three bigram counts using Good-Turing smoothing.
   $r* = (r+1) * N_{r+1}/N_r$

3. Consider the formula for the probability of a sentence:
   $p(w_1...w_n) = p(w_1)p(w_2|w_1)...p(w_n|w_1...w_{n-1})$

   (a) What are the third and fourth terms on the right hand side of the = sign?

   (b) Reformulate the above model as a unigram, bigram and trigram model.