The Random Sentence Generator (RSG) is written in RSG.java. Running the main method will produce 50 random sentences for each of the 6 language models (unigram, bigram, trigram; and downspeak, upspeak for each).

RSG.java requires the 6 .txt files in the out directory. These files contain, for each of the language models, a list of all n-grams and the number of times each one appears in the training corpus. The information for each n-gram is given on separate line in the format <n-gram> <count>.

It is best to describe how the RSG uses the unigram models first, since the use of the bigram and trigram models builds upon it. First, the file containing the unigrams and their counts is read in. The values are inserted to a Java ArrayList of Strings. Each unique unigram is inserted into the ArrayList an amount of times equal to its count. For example, if our input file lists:

hello 3

there 1

my 2

friend 1

the resulting ArrayList will be:

["hello", "hello", "hello", "there", "my", "my", "friend"]

This approach was taken in order to intuitively implement distributed probability. Since ArrayList is an indexed data structure, picking a random index results in selecting a random word. The more times that word appears in the ArrayList, the more likely it is to be selected. Although doing this increases the running time, the extent to which it does is fairly negligible. Speed was not of concern for this project and the sentences will print out in a matter of seconds, even though the input files range from about 8,000 to 65,000 lines in length.

For the unigram model, N - 0 here is equal to 0, so each word is picked without consideration of the previous word. A String randomSentence is initialized with the start-of-sentence tag <s> and has each new random word appended to it. It is returned once the end-of-sentence tag </s> has been appended or SENTENCE\_MAX has been reached.

Generating sentences with the bigram models requires the consideration of the previous word. The RSG looks at the previous word and then reads in the bigram input file, adding to an ArrayList only the bigrams that begin with the previous word. This is done in the same fashion that the unigrams were stored in (with each bigram being inserted as many times as its respective count, so that distributed probability is applied). Again, this is done repeatedly until the sentence comes to an end.

For the trigram models, the previous two words of the sentence being generated need to be considered. The first iteration is done exactly the same as with the bigram models because at the beginning of the sentence, there is only one previous word (the start-of-sentence tag <s>). After that, new random words are picked just as they were using the bigram models but instead, we look at the previous two words. That is to say, each new word selected is the third word in the randomly selected trigram whose first two words are the previous bigram.

Example sentences:

Unigram downspeak:

<s> for question expect exception p&l 2001-09-01 it come manager physical look enough he . Leslie target in </s>

<s> any market . password . in the testing in regulation and a this </s><s> please ? on rely - discuss and attach from cost </s>

<s> I fix for 2,372,511,738 . indexing learn Daily . if Mike </s>

<s> </s>

Unigram upspeak:

<s> appear you . will and the to call ext match Engineering Skarness Tom way call in move </s>

<s> can document I ?? a should call of the 713-853-0329 on breath . be </s>

<s> Plaza Lynn question contract and Monday </s><s> 719-577-5769 which so , gallery you Steve I \ message \ Time get Risk Ladies presentation for and </s>

<s> </s>

Bigram downspeak:

<s> for next week start Mid East Power </s><s> no make sure that Kenneth Lovejoy have overall responsibility roll over Pease St. to use only if you , and b confirm as with yourselve </s><s> I 's credit term this print - there be a be handle in late intelligence support from a and the spreadsheet . </s><s> Louise kitchen ' and restriction . </s><s> I think the scope of the guy get the attorney who complain or to I just go target . </s>

Bigram upspeak:

<s> with the be out of the information . </s><s> Vince , and first item to I know the please contact </s><s> thank you to meet each month be only to the select to have comment </s><s> 713-853-6833 </s><s> Shall we , but I at be change that the golf or clickpaper-related counterparty for the incremental fee a Tentative Business school </s>

Trigram downspeak:

<s> I want to make payment directly to we , still we be take at this time of need . </s><s> you have any question . </s><s> feedback from customer and post on the actual date of the email I send a email to Kim Hillis be make progress , and wptf be go -rrb- . </s><s> standardize market , and be not move as quickly due to operational constraint on the website . </s><s> now let 's make it a 2:30 </s>

Trigram upspeak:

<s> Tuesday , July 12 , 2001 </s><s> Rick will hold a conference call schedule for Thursday 's call with John Lavorato and Louise , </s><s> I be not correct in TMS do not focus on achieve the objective of Enron . </s><s> sorry for the extra time for the last transaction that John have execute be Enron 's message and purpose . </s><s> Honey Baked ham </s>

The sentences generated using the unigram models look truly random and are notably less intelligible than those generated by the other models. Empty sentences ("<s> </s>") were quite common, since the counts for the start- and end-of-sentence tags were relatively high. It was also more common in the unigram models for punctuation such as periods to appear in the middle of sentences.

The bigram models provided slightly more intelligible sentences. Familiar phrases such as "thank you" "make sure" made appearances, as well as syntactically correct phrases such as "please" followed by a verb. Additionally, a distinction began to appear between upspeak and downspeak. Upspeak seemed to contain more interrogative language such as "please" and "shall we" than the downspeak sentences did.

The trigram models produced more syntactically as well as semantically correct sentences than the other models. Longer phrases with discernible meaning appeared, such as "Rick will hold a conference call," "sorry for the extra time," and "I want to make payment directly." Overall, there was a direct relationship between N and the quality of the sentences in terms of syntax and comprehensibility.