Jeremy Pollock
MATH187
April 29, 2005

This paper tests the assertion of the following text concerning the characteristic distribution of letters in the English language:

> "Why would the frequency of the letter "e" be roughly the same in all English texts, independent of which century it was written, independent of the topic, independent of the author? First, it is isn't constant. There are variations from article to article, author to author, and as can be seen in the data in Figure 3.3.1, the frequency in a consistently edited source such as the New York Times still varies about 0.1%. That is, the last decimal place of the numbers in the table are meaningless.
>
> So why is there any regularity at all? The answer lies in the list of the 30 most frequent words: the, of, and, a, in, to, is, to, for, it, be, was, on, with, that, by, are, 's, this, from, which, at, not, or, an, he, but, has, will, I. With the exception of places 26 and 30, the words have no meaning. They are grammatical placeholders. Since a language (such as English) has only a few sentence types, and the sentence type is given by the placement of such "grammar words", one expects them to dominate. They dominate the list of frequent words, and their letters dominate the list of frequent letters."

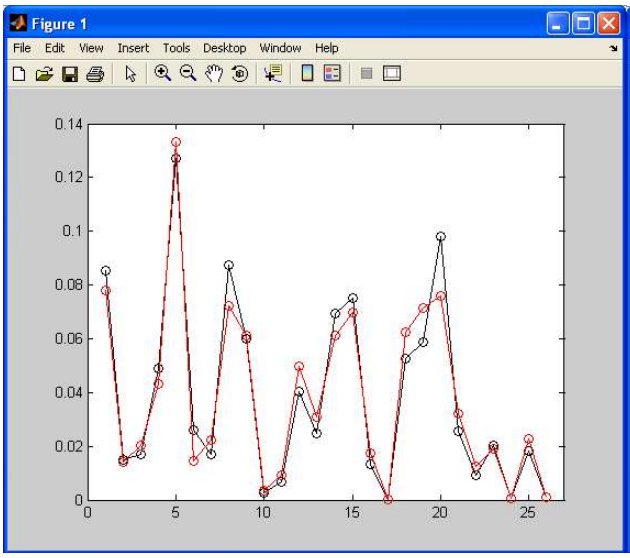
This is not an unreasonable assertion, that because each of the listed words are commonplace in the English language due to its grammatical structure, then the observed distribution of letters in English are related to, at least in part, the presence of these words. The following experiment tests this hypothesis.

The pseudocode of the attached Java program (`gwAnalysis.java`) follows. There are two frequency tables and also a list of the "grammar words" stored for comparison.

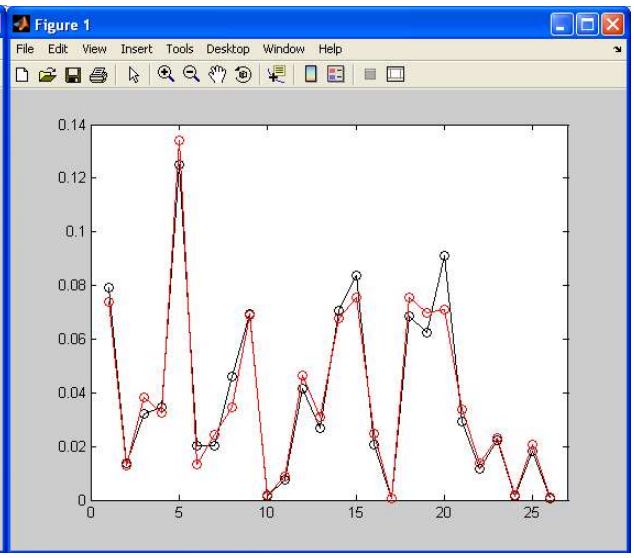
```
->given a file of plaintext,
->loop until there is no more text in the file
      ->read in a line of text from the file
      ->convert all characters in that line to lowercase and remove all
            formatting except spaces
      ->loop until there are no more words in that line
            ->read a word from the line of text
            ->check the current word against the grammar word list
            ->if the word is not a grammar word:
                  ->then, update both frequency tables
                  ->else, update only the first
            ->end if
      ->end loop
->end loop
->normalize the letter frequencies
->write the frequencies out to a Matlab script for easy visualization
```

When the program is done, the first frequency table has recorded information for all observed characters, and the second has only recorded information for characters that do not occur in grammar words. You can see in the resulting plots however, that there is little difference in the characteristic letter distribution.

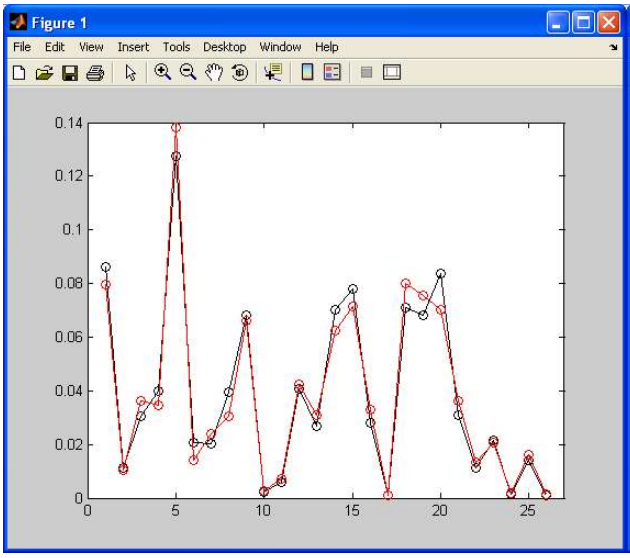
The results can be seen in the following images where the x-axis is the numerical index of each letter (where A=1, B=2, and so on) and the y-axis is the occurrence frequency. The distribution drawn in black is the complete letter frequency distribution. The red distribution is the character frequency of letters while ignoring grammar words.



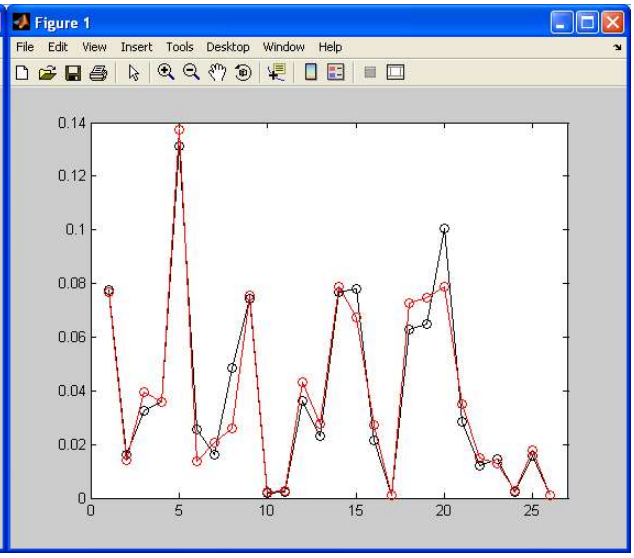
The King James Bible



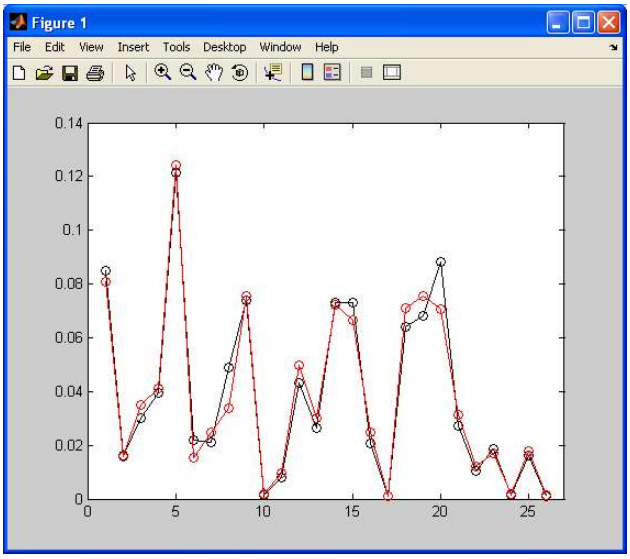
Bill Clinton's State of the Union



George Bush's State of the Union



Abraham Lincoln's State of the Union



463 Time articles from 1963

I tested a number of different texts, ranging from the State of the Union addresses given by Presidents Abraham Lincoln, Bill Clinton and George "Dubya" Bush, to a text file that included 463 distinct articles from Time magazine in 1963, to even a copy of the King James Bible, both the Old and New Testaments. Each of these texts illustrate that even with the so-called "grammar words" removed from the text, that the characteristic distribution of the letters was mostly preserved. The order of some of the most common letters was switched, but it's not true that the distribution was altogether flattened through the removal of these grammar words.

The assertion concerning "grammar words" is flawed because it attributes the observed distribution of the letters in the English language to the language's grammatical structure. Grammar governs the placement and behavior of words in English, but has little to do with the actual letters that make up English words, as this assertion seems to imply. I would instead contend that the rules or structure that governs the generation of "English-like" words is what determines this distribution.

That is, if I was to generate a list words whose constituent letters were randomly generated, most practiced English readers would struggle to read through this list. This is because there is an underlying structure for words in the English language that I believe we have an implicit understanding of. As motivation for this assertion, can you read the following words?

mave
sheed
dere
haid

None of these "words" can be found in an English dictionary, although they do appear to have an English construction. How about the following list?

rviz
wste
uaoi
bpct

My guess is that most practiced English readers either cannot read the second list or would struggle in doing so. The intuition that I am attempting to drive at is that there are rules that dictate the "Englishness" of a string of letters. Put another way, there are generative rules that define the structure of most English words.

For instance, quite a lot of words tend to start with a consonant, followed by a vowel, followed by another consonant (CVC). How many words can you think of that are generated with a CCC-based rule? VVV? A quick count over words in this document using regular expressions indicates that ~29.3% (282/960) of words start with the form CVC. 21.1% follow the form CCV. In comparison, there are thirteen total words that follow the either the CCC form or VVV form (1.4%). These numbers indicate that words in the English language are not random assemblies of English characters, but rather that they tend to follow a kind of structure that is common in the language.

This should also at least partly explain why vowels dominate the letter distribution... because there are so few of them relative to consonants, yet they play an integral role in constructing English words. On the other hand, it's not quite clear to me why, for instance, t is more common than c; there is nothing that I have mentioned that would explain this. However, I do believe that an argument based in phonemes or rules governing word generation can explain this.

Appendix:

The transcript of Clinton's State of the Union address is included in "su_clinton.txt." It can be found online at:
http://www.gutenberg.org/etext/5048

The transcript of Bush's State of the Union address is included in "su_georgeW.txt." It can be found online at:
http://www.gutenberg.org/etext/5049

The transcript of Lincoln's State of the Union address is included in "su_lincoln.txt." It can be found online at:
http://www.gutenberg.org/etext/5024

The electronic copy of The King James Bible, Old and New Testaments is included in "thebible_kjv10.txt." It can be found online at:
http://www.gutenberg.org/etext/10

The electronic copy of the 423 Time Magazine articles is included in "doc.edit." It can be found online at:
http://www.cs.utk.edu/~lsi/corpa.html