Zipfian Distribution of Words and Word Phrases in American English Speech

Bethany Bogensberger
Department of Computer Science
Gonzaga University, Spokane, WA

Abstract

Zipf's law states that given some written text, the frequency of any word is inversely proportional to its statistical rank. As applied to words this means the top ranking word, the one used most frequently, occurs twice as often as the second most used word, and three times as often as the third most used word. This distribution has been shown to hold for all languages examined. Zipf's Law has also been shown to hold for word sequences within a text. However, Zipf's law has not been shown to hold for the spoken word. Linguists have recently shown that speech and writing are structurally different. This paper examines a corpus of spoken language to see if Zipf's Law holds for speech.

1. Introduction

Zipf's Law is a power law which states that the "r^{th} most frequent word has a frequency f(r) that scales according to

$$f(r) \propto \frac{1}{r^{\alpha}}$$

for $\alpha \approx 1$. In this equation r is the "frequency rank" of any given word, and f(r) is the frequency of the word in a natural corpus" (Piantadosi 2014, 1). To process a corpus to see if it holds to Zipf's Law a graph of the log(r) vs log(f(r)) should produce a straight line with the slope close to -1. Looking at a graph of this log-log graph we can see how the languages analyzed follow this distribution.

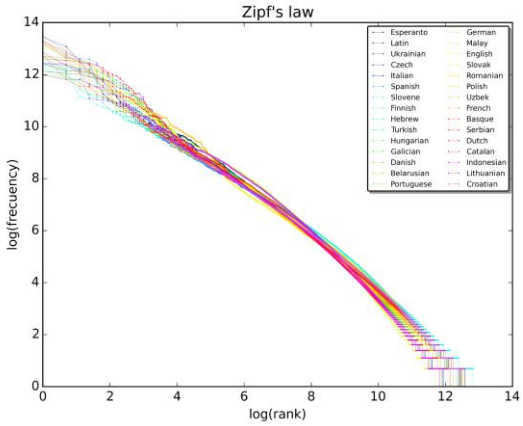

Figure 1 from Wikipedia entry, "Zipf's law" (2016)

The highst and lowest frequencies tend to deviate slightly from the line predicted by zipf's law. The separation stems from the different sizes of the corpora (Ha, et al. 2002).

Linguists have found that the written word differs from the spoken word. When looking at the tapes from the Watergate scandal, linguists spotted differences between the spoken word and written word. One of the key differences was that conversants spoke in "intonation units," rather than the complete sentences found in texts like this one (Tomasello, 2003). In writing, unlike speech, there is no opportunity to restart an utterance. These differences raise the question, does Zipf's law also hold to the spoken word?

2. Method

In order to analyze whether or not the spoken word holds to Zipf's law we looked at the Buckeye Corpus (Pitt, et al., 2007). This corpus is made up of forty interviews containing spontaneous speech. Using transcripts of these, we obtained roughly

300,000 words. Writing code in Python 2.7, we were able to count the occurrences of each word, as well as counting the occurrences of word sequences. The word sequences we examined were individual words (unigrams), two words (bigrams), three words (trigrams), and four words (quadgrams). With these numbers, we were able to calculate the frequencies, the log of the frequencies, and the log of the rankings. Using Excel, we produced graphs comparing the ideal Zipfian distribution line to the data gathered. These are shown in figures 2 through 5.

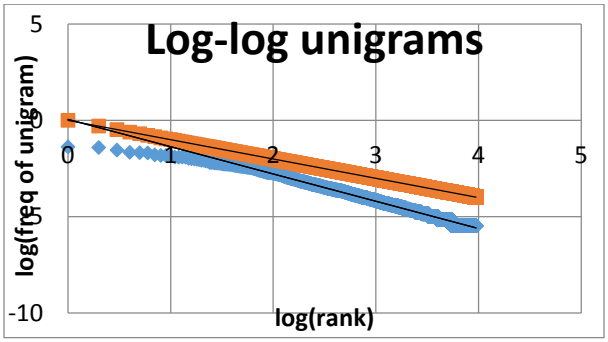

Figure 2 Unigrams

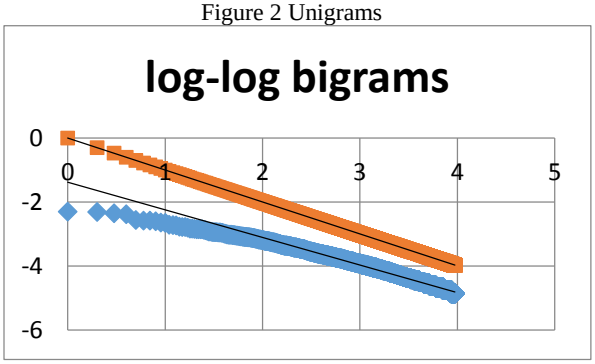

Figure 3 Bigrams

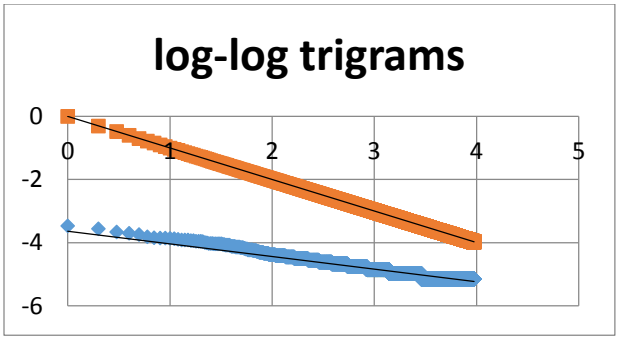

Figure 4 Trigrams

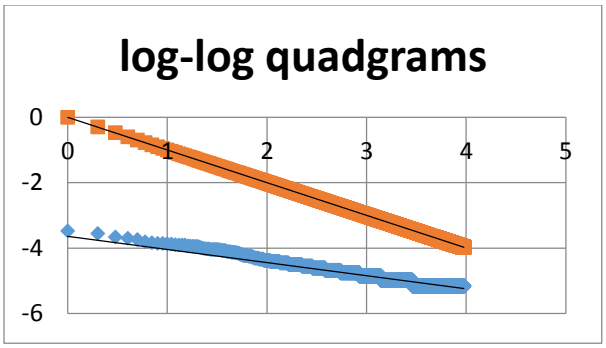

Figure 5 Quadgrams

In all of these graphs the orange line is the ideal Zipfian distribution, and the blue line is the distribution we observed in the Buckeye Corpus.

Looking at the graphs does not give a conclusive result, although they do suggest that spontaneous speech does not hold to Zipf's Law. Taking it a step further to know definitively, we did a goodness of fit test. We computed the expected value of a given rank and compared it to the actual value we observed.

The ideal Zipfian distribution

$$total = x + \frac{x}{2} + \frac{x}{3} + \cdots + \frac{x}{n}$$

can be used to find the expected value of a given rank. *Total* is the total number of occurrences of the ranks being looked at, *x* is the expected number of occurrences of the most frequent word, and *n* is the number of ranks. For example, suppose our words are *the, of, a*, and *to* with occurrences of 55, 30, 10, and 3 respectively. *total* would be 98 and we would solve for x. We can now use a chi-squared test to judge the goodness of fit. The first step is to calculate the chi value:

$$\sum_{i=1}^{n} \frac{(expected_i - observed_i)^2}{expected_i}$$

Here *expected* is the expected number of occurrences at rank *i*, *observed* is the number of occurrences actually observed at rank *i*, and *n* is again the number of ranks we are inspecting.

3. Results and Further Research

Using *pchisq()* in r-fiddle (r-fiddle, 2016), the chi-squared result came back as 0. This means that our observed distribution does not depend on the Zipfian distribution. In essence, there is no correlation between the two. Not only are speech and writing structured differently, but they also have completely different word and word sequence distributions.

Linguist have argued that speech and writing are different; this result offers further evidence for that argument. In the future we hope to replicate the results using larger corpora and corpora of other languages other than English. These results may be relevant to word prediction in text messaging systems, to offer a single example.

4. References

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., & Smith, F. J. (2002). Extension of Zipf's law to words and phrases. *Proceedings of the 19th International Conference on Computational Linguistics* -, 1-6. doi:10.3115/1072228.1072345

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review Psychon Bull Rev, 21*(5), 1112-1130. doi:10.3758/s13423-014-0585-6

R-Fiddle. (n.d.). Retrieved April 23, 2016, from http://www.r-fiddle.org/

Tomasello, M. (2003). *The New Psychology of Language: Volume 2: Cognitive and Functional Approaches To Language Structure*. S.l.: Lawrence Erlbaum Associates.

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E. (2007) Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Zipf's law. (2016, April 26). Retrieved May 10, 2016, from https://en.wikipedia.org/wiki/Zipf's_law