Word Frequency Patterns and Applications of Zipf's Law

Abby Sussman
College of William & Mary

March 26, 2024

Abstract

This paper uses a Python program to model the word frequency of various English texts. The analysis reveals patterns suggesting a logarithmic relationship, which prompts exploration into Zipf's law. This law states that there is an inverse relationship between word frequency and rank. Many sets of data, including AI-generated texts and YouTube views are explored in this work. While some domains adhere to Zipf's Law, others do not. Ultimately, this paper contributes to the understanding of word frequency dynamics and provides further applications of Zipf's law.

1 Introduction

The be to of and a in that have I. These are the most common written words in the English Language.[1] But is this the case for every text? This question sparked my interest in word frequency. I do not have a background in linguistics, so I had to start by learning the basic terms.

2 Definitions

What is a word? A word is a "speech sound or series of speech sounds that symbolizes and communicates a meaning usually without being divisible into smaller units capable of independent use".[2] As you can imagine, different types of words are important to define.

First, we have stop words: common words that do not add substantial meaning to a sentence. For example, the list 'a', 'an', 'and', 'as', 'at', 'by', 'for', 'from', 'if', 'in', 'of', 'on', 'or', 'the', 'to', and 'with' is commonly used for natural language processing, as these words add little meaning to a sentence. The use of stop-word lists began in 1959 and gained popularity among linguists during the 1960s. For example, when generating the index for the 1964 book *Geoscience Abstracts*, computer scientists decided to ignore words like 'approximate', 'etc.', 'geologic', 'rocks', and 'science' because they were frequently used and did not contribute to the meaning of the text.[3]

Second, are tokens and lemmas, which provide different ways to categorize words for analyses. Tokens are individual words. Consider the example 'I was running while he ran': this sentence has six tokens. Words can also be analyzed using lemmas. In linguistics, a lemma has a different definition than it does in mathematics. Here, it means "the canonical definition of a set of word forms".[4] Words like 'speak', 'speaking', and 'spoke' all have the same lemma 'speak'. Therefore, the same example sentence 'I was running while he ran' only has five lemmas because 'running' and 'ran' have the lemma 'run'.

3 Approach and Process

My main motivation for this investigation was to determine if word frequency followed a certain pattern or distribution. Although stop words are commonly removed in natural language processing, since the nature of this question involves the frequency of individual words, I will be counting every word and not omitting any stop words. I wrote a Python program that counts the frequency of each token in a given text and then graphs the ten most commonly used words. It also finds the total amount of tokens and unique words; this aspect will be useful for something I'll discuss later in this paper. A pseudocode of my program is as follows:

```
open .txt file
create empty dictionary
for each line in text:
    make line lowercase
    get rid of punctuation
    split into words
for every word:
    if word is in dictionary
        add 1 to value
    else
        add word to dictionary with a value of 1
sort dictionary from most to least used
print the top ten words
graph the top ten words
print the total amount of words (tokens)
print the number of unique words
print the frequency of the top 20% of words (⋆)
```

(⋆): this line becomes important later

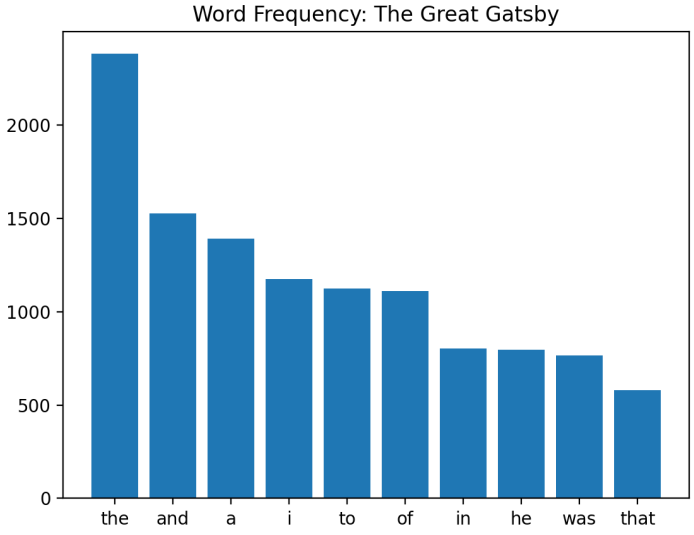
The next step involved selected texts to analyze. I first chose *The Great Gatsby* because it is a classic novel that has been widely read and studied.[5] While researching this topic, I noticed that *Moby-Dick* was a reoccurring source, so I decided to analyze that text myself.[6]

For the last piece, I wanted to do something relatively short to see if the resulting patterns were consistent across different word counts. I therefore chose *The Gettysburg Address*.[7]

4 Results

The Great Gatsby results are in Table 1.
The *Moby-Dick* results are in Table 2.
The Gettysburg Address results are in Table 3.

I noticed that the word frequency in the books had what seemed to be a logarithmic relationship. I then started to research laws and theorems about the frequency of words and came across Zipf's Law.



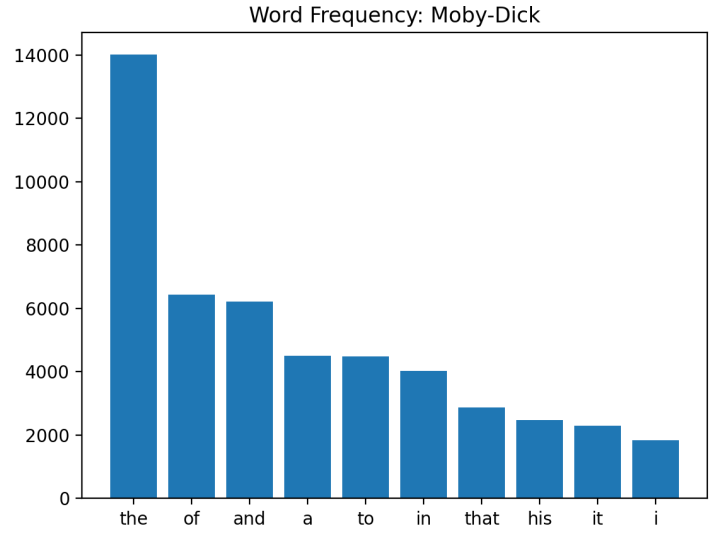
the: 2,380	
and: 1,526	
a: 1,392	
i: 1,176	
to: 1,125	
of: 1,109	
in: 801	
he: 795	
was: 764	
that: 581	
48,346 tokens with 6,408 unique words	

Table 1: The Great Gatsby

5 Zipf's Law

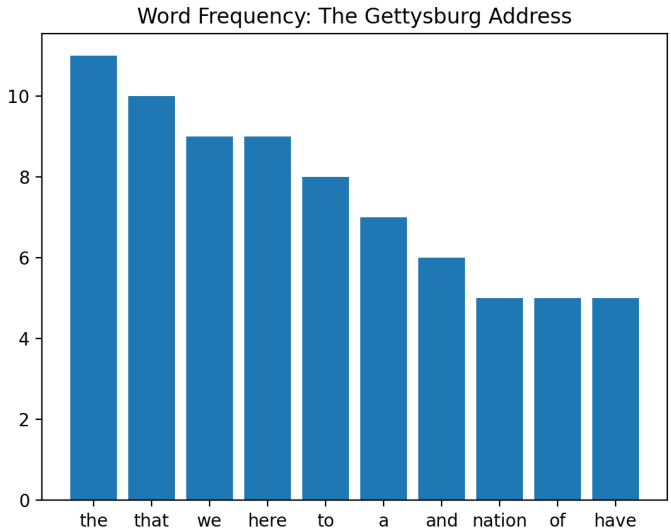
George Kingsley Zipf (1902-50) was an American linguist and philologist who specialized in statistical occurrences in languages. His most notable contribution is Zipf's law, which states that only a few words are used very often and most are used rarely.[8] More specifically, he says that the relative frequency of a word is inversely proportional to its rank. So,

$$\text{word frequency} \propto \frac{1}{\text{word rank}}$$



the: 14,005	
of: 6,435	
and: 6,222	
a: 4,512	
to: 4,485	
in: 4,033	
that: 2,864	
his: 2,479	
it: 2,287	
i: 1,835	
208,459 tokens with 21,024 unique words	

Table 2: Moby-Dick



the: 11	
that: 10	
we: 9	
here: 9	
to: 8	
a: 7	
and: 6	
nation: 5	
of: 5	
have: 5	
262 tokens with 139 unique words	

Table 3: The Gettysburg Address

In other words, the second most used word is used half as much as the first, the third word used a third as much as the first, the fourth a fourth as much, etc. and is written as:[9]

$$f(r) \propto \frac{1}{r^{\alpha}} \text{ , with } \alpha \approx 1 \text{ and } r = \text{rank} \tag{1}$$

We can verify this relationship by using a much bigger sample than I did. The Royal Statistical Society gathered a one-million-word collection of English texts and plotted the top one thousand words and their frequencies (Figure 1)[10].

The inverse relationship becomes even clearer when plotted logarithmically (Figure 2)[11]. Linear regression gives us a slope of $-0.98 \approx -1$. This means that Zipf's law follows a power-law distribution, which further confirms Zipf's observation in (1).

Referring back to Tables 1,2,3, you can see that *The Great Gatsby* and *Moby-Dick* approximately follow a power-law distribution, while *The Gettysburg Address* does not.

5.1 Pareto Principle

The Pareto Principle, also called the 80/20 rule, is an example of a power-law commonly used in business that states that 20% of the causes account for 80% of results. Another example of it can be seen in the distribution of wealth, saying that 20% of the population owns 80% of the wealth.[12] I wanted to see if this was true for word frequency: are 20% of words used 80% of the time? The final line in the pseudocode, ($\star$) gives us the following results:

The Great Gatsby : 20% of words are used 85% of the time
Moby-Dick : 20% of words are used 87% of the time
The Gettysburg Address : 20% of words are used 49% of the time

Again, *The Great Gatsby* and *Moby-Dick* follow the Pareto Principle, while *The Gettysburg Address* does not. This leads me to believe that the rules discussed so far start to appear when you reach a certain word count. Figuring out the critical point is beyond the scope of this paper, but nonetheless is an interesting point to mention.

6 Other Applications

6.1 AI-Generated Text

Does AI-generated text follow Zipf's law? I gave ChatGPT three prompts and asked it to write three stories. The first story is about a family going to an amusement park (Table 4), the second story is a murder mystery (Table 5), and the third is a period piece romance (Table 6)[13]. As you can see, ChatGPT uses 'the' at a much higher rate than the human-written texts we previously analyzed. (So, if you were going to try to publish a book written by ChatGPT, maybe get rid of half the 'the's first.) Additionally, the AI-generated text does not follow Zipf's law or the Pareto principle. The third story seems slightly more promising but does not retain the power-law distribution long enough to confidently say it follows the patterns of 'normal' text.

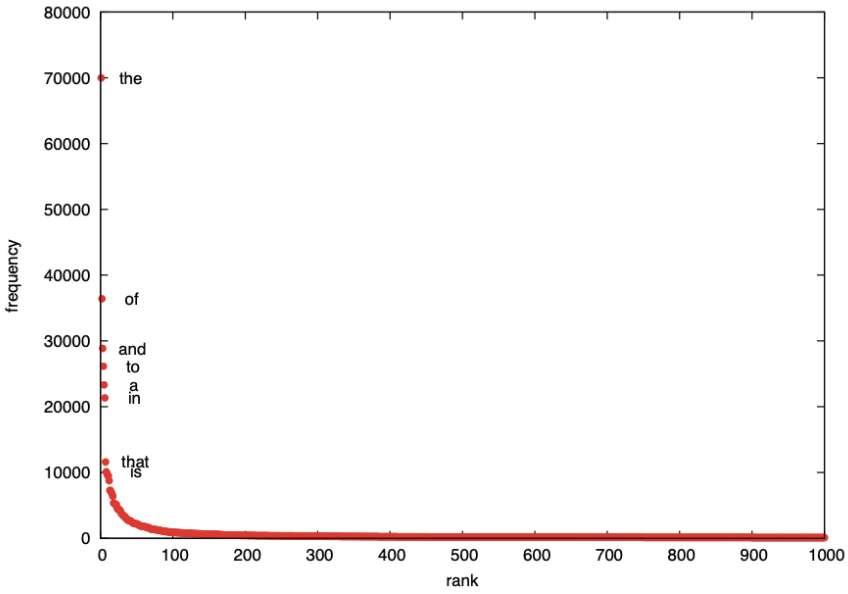


Figure 1: The word 'the' occurs 70,000 times and 'of' occurs 36,000 times

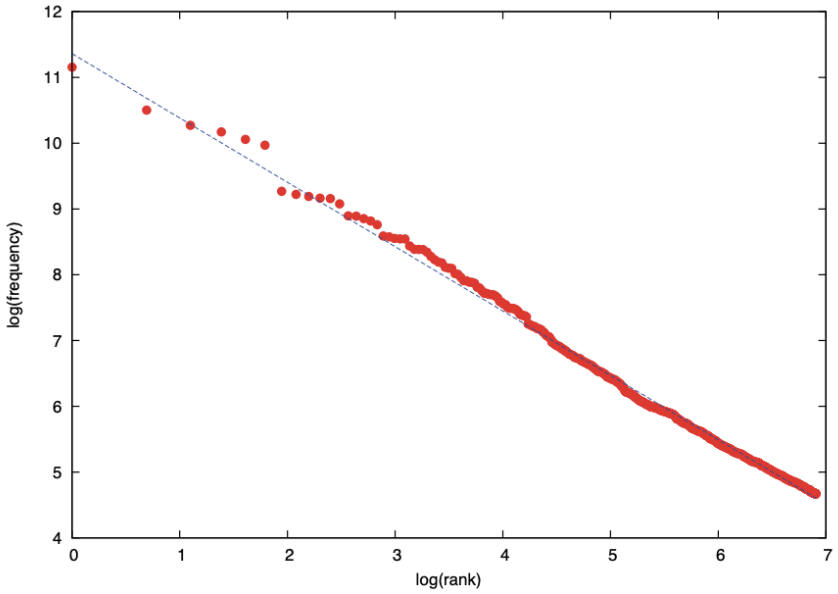
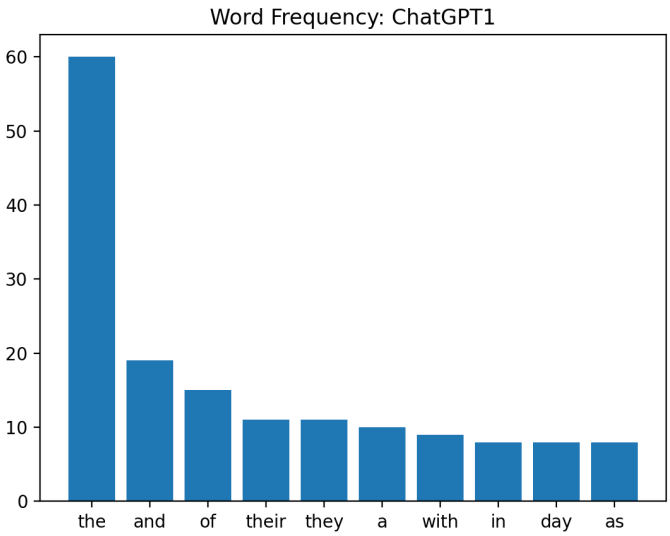
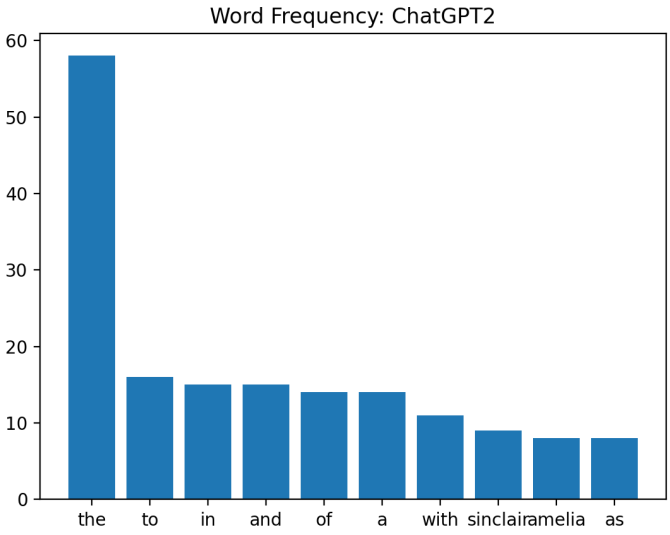


Figure 2: The log-log plot of Figure 1



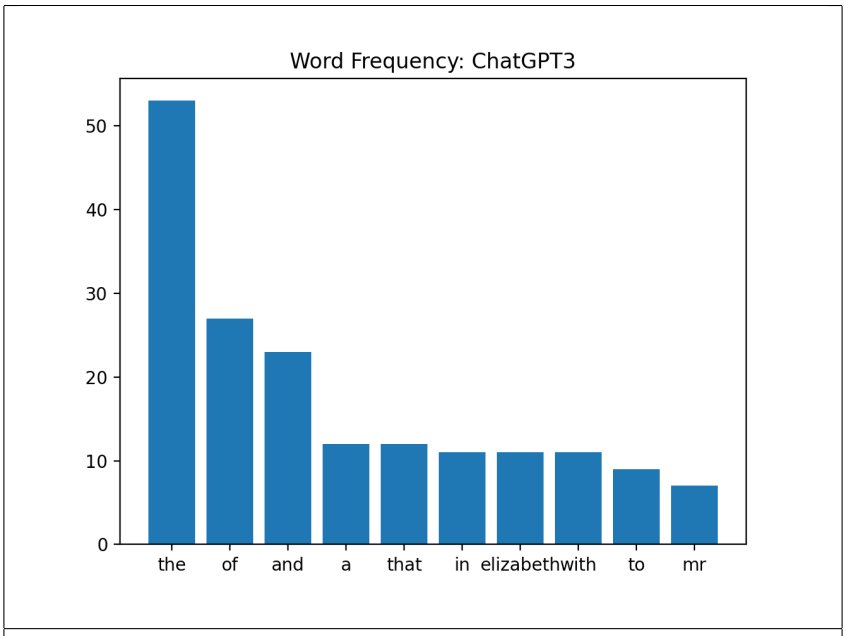
449 tokens with 230 unique words
20% of words are used 58% of the time

Table 4: First ChatGPT text



526 tokens with 298 unique words
20% of words are used 54% of the time

Table 5: Second ChatGPT text



523 tokens with 263 unique words
20% of words are used 57% of the time

Table 6: Third ChatGPT text

6.2 City Population

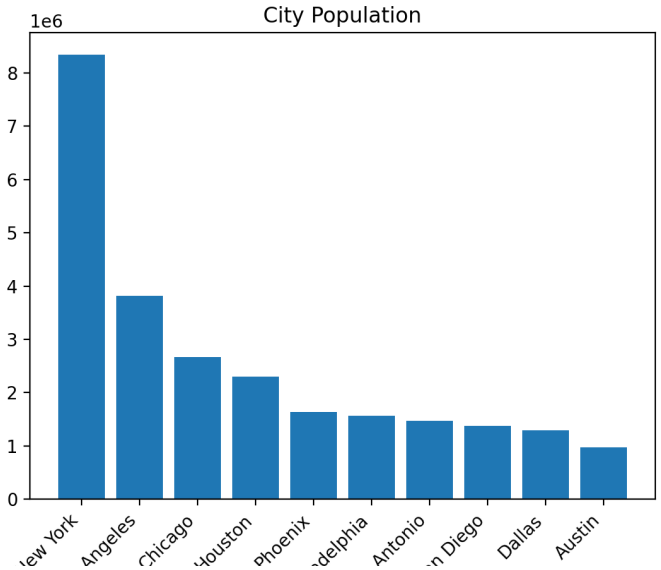
A strong example of Zipf's law can be found by measuring city population. In 2000, the size of United States cities with a population over 10,000 followed Zipf's law.[14] To see if this is still the case, I ran more current data from 2022 through my program (Table 7).[15]

Looking at the graph, you can see that Zipf's law still holds. To confirm this notion, we can see that half the population of New York is about the same size as the population of Los Angeles, a third of New York is about the same size as Chicago, a fourth is about the same size as Houston, etc. It gets less accurate as you go down the list, but it is clear that Zipf's law is present.

6.3 Youtube Views

A 2006 research paper found that the web hits received by servers on a single day followed Zipf's law, meaning the second-most viewed website was viewed half as much as the first, etc.[16] This made me wonder if other parts of the internet had similar distributions. I have always been a big YouTube video watcher, so I looked at the top ten most viewed YouTube videos and ran them through my program.[17] The results are in Table 8.

While the first two videos almost match Zipf's law, unfortunately, the pattern falls apart after that. Perhaps, as we get further into our digital age, we will start to see a power-law distributing form.



New York: 8,335,897

Los Angeles: 3,822,238

Chicago: 2,665,039

Houston: 2,302,878

Phoenix: 1,644,409

Philadelphia: 1,567,258

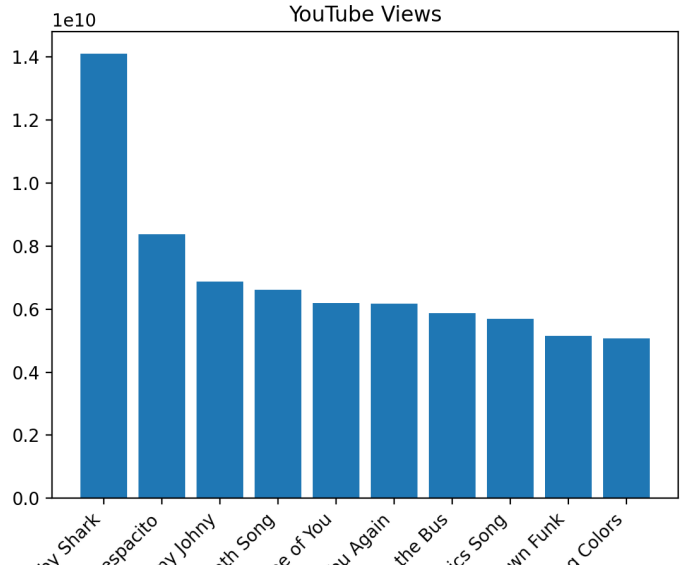
San Antonio: 1,472,909

San Diego: 1,381,162

Dallas: 1,299,544

Austin: 974,447

Table 7: City population



'Baby Shark Dance': 14.09 billion

'Despacito': 8.38

'Johny Johny': 6.87

'Bath Song': 6.62

'Shape of You': 6.20

'See You Again': 6.17

'Wheels on the Bus': 5.88

'Phonics Song': 5.70

'Uptown Funk': 5.15

'Learning Colors': 5.07

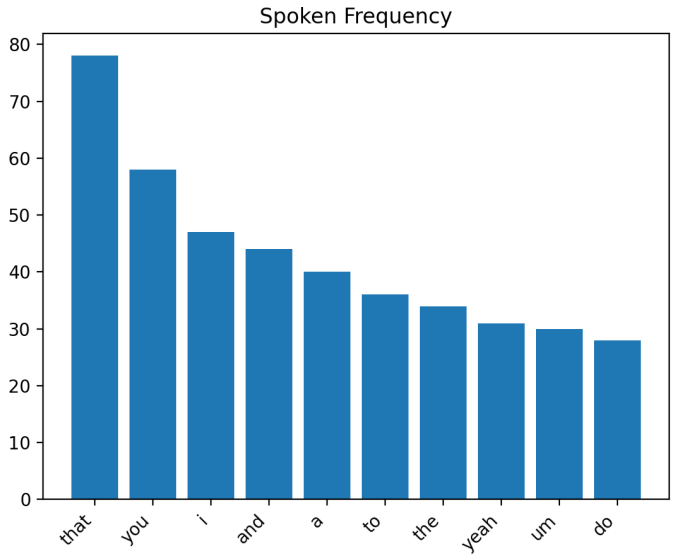
Table 8: YouTube views

6.4 Alien Language

Zipf's law holds crosslinguistically in other languages such as Spanish, French, and Russian.[18] But what about non-human languages? A team of scientists was able to decode bottlenose dolphin whistles and discovered they follow Zipf's law too: certain clicks and different durations of whistles are used more frequently than others. Laurence Doyle, part of that team of scientists, is determined to uncover "extraterrestrial blah blah" by listening to space radio.[19] If there are other beings somewhere in space with vocal language, we may be able to organize it into meaningful bits depending on their frequency.

7 Presentation Comments

The most popular comment I received after my presentation was whether spoken language also followed Zipf's law. At the time, I chose not to include spoken language because it drastically varies from person to person. For example, if I were to record myself speaking, the word 'like' would appear far more often than in a recording of my father. Furthermore, there is not a definitive list of the most frequently spoken English words like there is for written English. However, I was still curious about this and found a full-verbatim transcript of a personal interview that includes all utterances of 'uh', 'um', 'mm-hmm', etc.[20] The results are in Table 9. An interesting result is that 'um' made the top ten.



that: 78	
you: 58	
i: 47	
and: 44	
a: 40	
to: 36	
the: 34	
yeah: 31	
um: 30	
do: 28	
1,651 tokens with 383 unique words	
20% of words are used 69% of the time	

Table 9: Spoken frequency

There were great suggestions for Python language tools like SpaCy and NLTK. I have a beginner's understanding of Python and wrote my program the best I could. In the future, I will look into these tools and see what else they can do for further analysis.

Zip's law is a probability distribution. It has a probability mass function, a communicative density function, a moment generating function, a mean, etc. These were things mentioned in BlackBoard responses, but it wasn't something I felt confident enough in to go over myself. If my subsequent paper expands on this research, I'll be sure to include more statistical explanations.

8 Conclusion

This paper explored the word frequency distribution in written texts. Analysis using a Python program revealed that Zipf's law, which states there is an inverse relationship between word frequency and rank, was observed. Additionally, the texts that followed Zipf's law also aligned with the Pareto Principle, or the 80/20 rule, where 20% of words accounted for about 80% of usage. This was not the case, however, for the shorter and AI-generated texts. We also took this concept to other fields to see if the same pattern surfaced. City population and dolphin communication were shown to obey Zipf's law, while YouTube views did not. Overall, this investigation sheds light on the intriguing patterns underlying word usage while showcasing the relevance of Zipf's law and its potential implications across different domains.

Notes

[1]Wikimedia, "Most Common Words in English".
[2]Merriam-Webster, "Word Definition & Meaning".
[3]Rosenberg, "Stop, Words".
[4]Wikimedia, "Lemma(morphology)".
[5]Fitzgerald, "The Great Gatsby".
[6]Melville. "Moby-Dick".
[7]"The Gettysburg Address".
[8]Wikimedia, "George Kingsley Zipf".
[9]Piantadosi, "Zipf's Word Frequency Law in Natural Language"
[10]The Royal Statistical Society. "Who's afraid of George Kingsley Zipf?"
[11]Ibid.
[12]Newman, "Power Laws, Pareto Distributions and Zipf's Law".
[13]OpenAI. "ChatGPT".
[14]Newman, "Power Laws, Pareto Distributions and Zipf's Law".
[15]Wikimedia, "List of United States Cities by Population".
[16]Newman, "Power Laws, Pareto Distributions and Zipf's Law".
[17]Wikimedia, "List of Most-Viewed YouTube Videos".
[18]Wikimedia, "George Kingsley Zipf".
[19]Doyle, "Listening for extraterrestrial Blah Blah".
[20]"GoTranscript".

9 References

Doyle, L. R. (2016, December 9). Listening for extraterrestrial blah blah. Nautilus. https://nautil.us/listening-for-extraterrestrial-blah-blah-236287/

Fitzgerald. (2024, February 2). The Great Gatsby by F. Scott Fitzgerald. Project Gutenberg. https://www.gutenberg.org/ebooks/64317

The Gettysburg Address. (n.d.). https://rmc.library.cornell.edu/gettysburg/good_cause/transcript.htm

Inspect our transcription samples via gotranscript's onsite editor. (n.d.). GoTranscript. https://gotranscript.com/full-verbatim-sample

Melville. (2021, August 18). Moby Dick; or, the whale by Herman Melville. Project Gutenberg. https://www.gutenberg.org/ebooks/2701

Merriam-Webster. (n.d.). Word definition & meaning. Merriam-Webster. https://www.merriam-webster.com/dictionary/word

Newman, M. E. J. (2006). (publication). Power laws, Pareto distributions and Zipf's law.

OpenAI. (2024). ChatGPT (3.5) [Large language model]. https://chat.openai.com

Piantadosi, S. T. (2014, October). Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic bulletin & review. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4176592/

Rosenberg, D. (2014) Stop, Words. Representations, vol. 127, no. 1, 2014, pp. 83–92. JSTOR, https://doi.org/10.1525/rep.2014.127.1.83.

The Royal Statistical Society. (2013, December). Who's afraid of George Kingsley Zipf? https://www.ling.upenn.edu/∼ycharles/sign708.pdf

Wikimedia Foundation. (2023a, September 11). Lemma (morphology). Wikipedia. https://en.wikipedia.org/wiki/Lemma_(morphology)

Wikimedia Foundation. (2023b, November 9). George Kingsley Zipf. Wikipedia. https://en.wikipedia.org/wiki/George_Kingsley_Zipf#cite_note-1
Wikimedia Foundation. (2024a, January 21). Most common words in English. Wikipedia. https://en.wikipedia.org/wiki/Most_common_words_in_English
Wikimedia Foundation. (2024b, February 14). List of United States cities by population. Wikipedia. https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population
Wikimedia Foundation. (2024c, February 19). List of most-viewed YouTube videos. Wikipedia. https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos