Statistical patterns in written language

Lukas Münzel

Research Science Institute
June 28 - July 1, 2021

1 Summary

Humans have been writing down their ideas and thoughts for thousands of years, with the first accounts of written language being as old as civilization ifself. [1] With time, different styles of writing and different languages were developed. This minipaper analyzes the patterns resulting from these different writing styles. Specifically, we show that the length and distribution of words don't change between different authors and languages.

Most importantly we came to the conclusion that authors using more words per sentence tend to also use longer words.

2 Abstract

Our work shows that the distributions of the word length, word occurrences and sentence lengths don't change their shape, independent of the author. Nevertheless, all distributions show statistically significant differences with two sample t-test for different books. Furthermore, a Pearson's correlation test revealed a statistically significant positive correlation between the amount of words used per sentence and the amount of characters per word.

3 Introduction

Counting words, word lengths and the amount of word per sentence are tasks perfecttly suited for computers. Since automatizing the process of acquiring data big enough for statistical analysis is necessary to keep the use of resources reasonably low, no other metrics were analyzed.

To find trends more than 20'000 books were analyzed, for the comparison between individual distributions seven books were handpicked.

4 Methods

4.1 Computer science methods

The books *A brief history of time* by Stephen Hawking, *2001: A Space Odyssey* by Douglas Adams, *Homo Faber* by Max Frisch, *A Tale of Two Cities* by Charles Dickens as well as the origin, English and French translations of Albert Camus's *The plague* were used for the analysis of the individual distributions. Furthermore, 23'175 books from a standardized and cleaned up version of Project Gutenberg [2] were used to analyse relationships between two different metrics.

As visualized in figure 1, the PDF's downloaded for the analysis of individual books were converted to a suitable raw text format using the standard Linux command *pdftotext*. All punctuation except for the points for the sentence length analysis was removed in a next step using regular expressions. The modern and fast command line tool *sd* was used for this purpose. In a final step the metrics of interest were then visualized using the Python libraries *matplotlib* and *seaborn*.

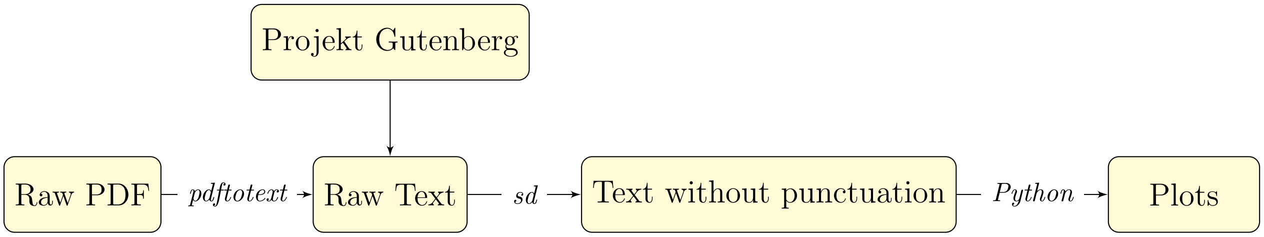


Figure 1: Data pipeline for the analysis

4.2 Statistical methods

To measure statistical significance, a two sample t-test was carried out for all pairs of the books selected by hand. The value of t for the two sample t-test is defined as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

And for the the Pearson's correlation

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

Where

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

5 Results

Table 1: P-values for the t-test between each pair of distributions of word lengths

	The plague	La peste	Die Pest
The plague	1.0	$< 10^{-100}$	$< 10^{-100}$
The plague	$< 10^{-100}$	1.0	$< 10^{-100}$
La peste	$< 10^{-100}$	$< 10^{-100}$	1.0

As shown in table 1, all differences between the medians of the distributions of world lengths are statistically significant. Similar results can be observed for all pairs of of books for all tested language metrics. Nevertheless, the shapes of the distribution for the same metric remain the same, as observable in the figures in the appendix.

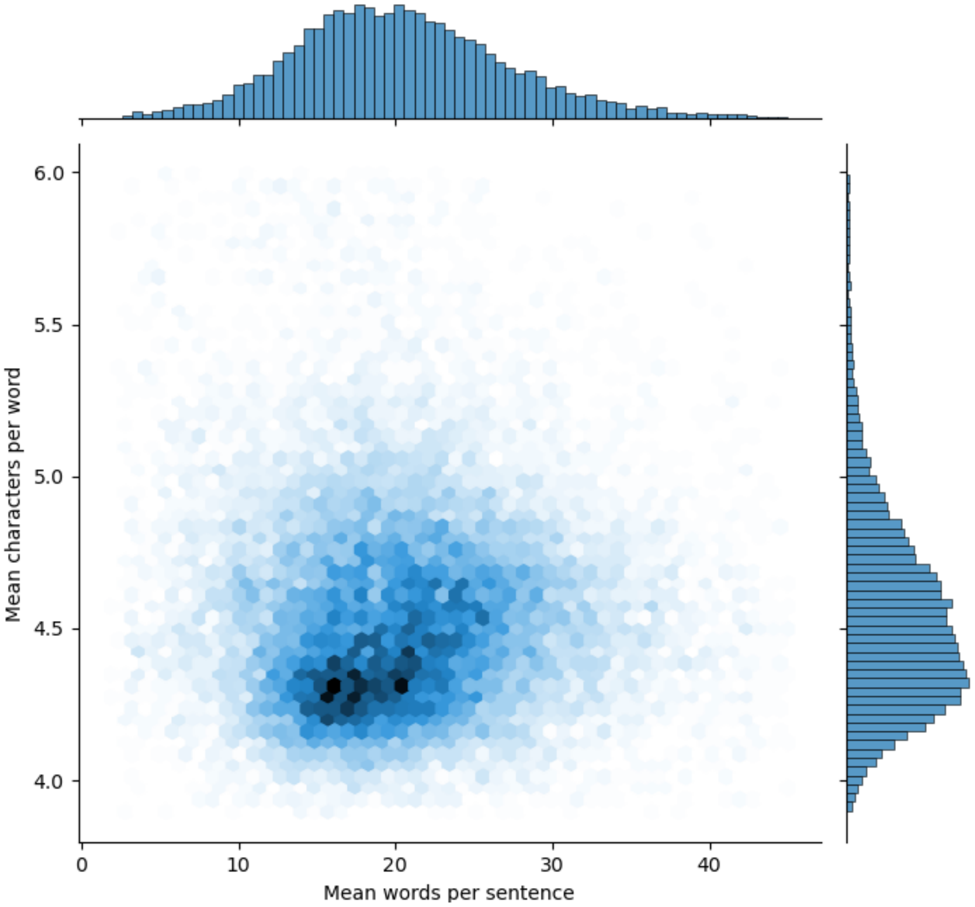


Figure 2: Joint distribution plot where darkness indicates the amount of books with marginal distributions on the sides

The correlation between the mean words per sentence and the mean characters per word can be observed in figure 2. Statistical analysis as explained in section four yields a p-value of 4.04%.

6 Conclusion

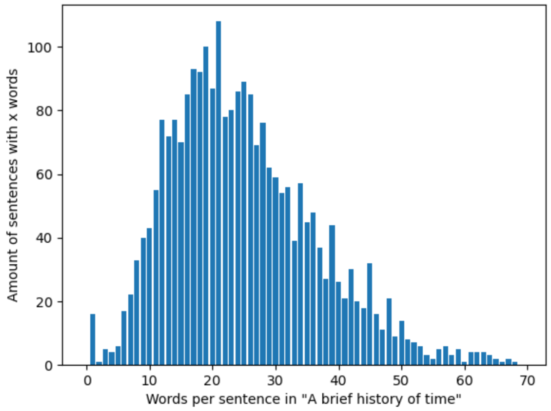
The p-value for the correlation between the mean words per sentence and the mean characters per word is below 5% and thus statistically significant. This means that authors who prefer to write longer sentences also tend to use longer words.

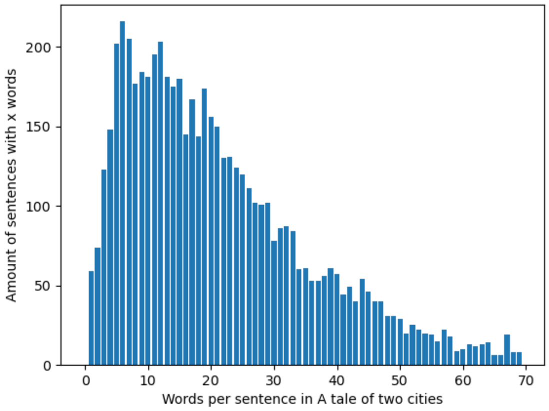
The similarity between the distributions independent of author or language suggests that the observed patterns are fundamental to language itself. Explanations for this include that the observed patterns emerge naturally from randomness. [3]

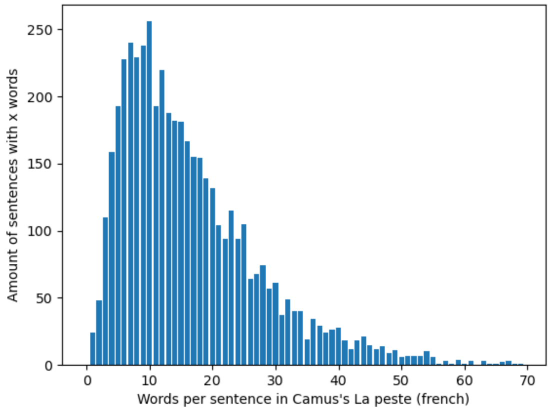
7 Acknowledgements

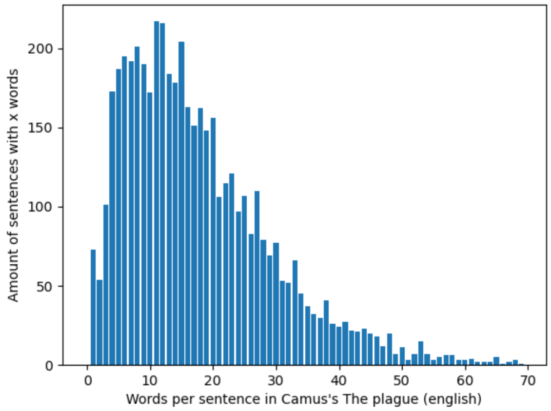
I would like to thank the Center for Excellence in Education and all of its employees, helpers and donors, with very special thanks to those making the Research Science Institute 2021 possible. Specifically, I would like to thank Mr. Peter Gaydarov, my tutor for this project. My gratitude also goes to Angelin Mathew, who is my counselor for the RSI.

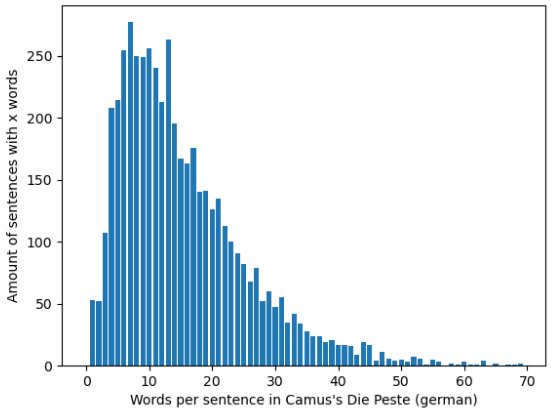
8 Appendix

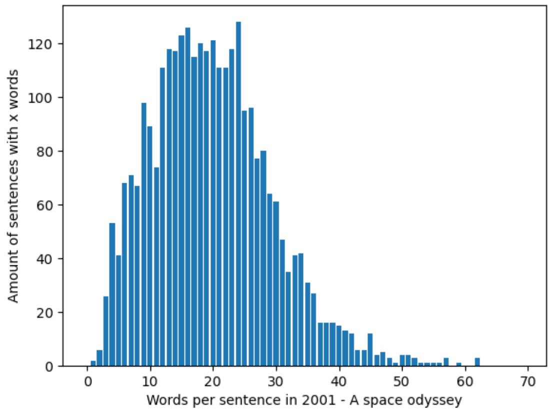


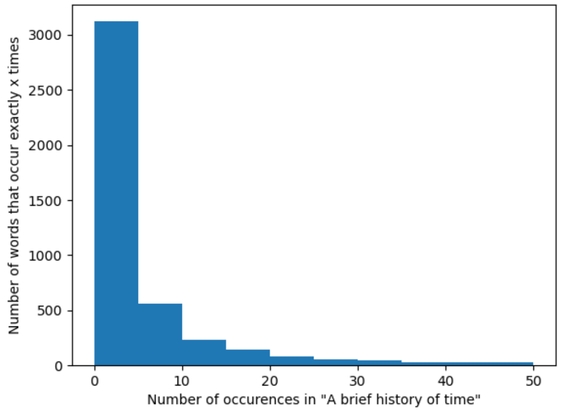


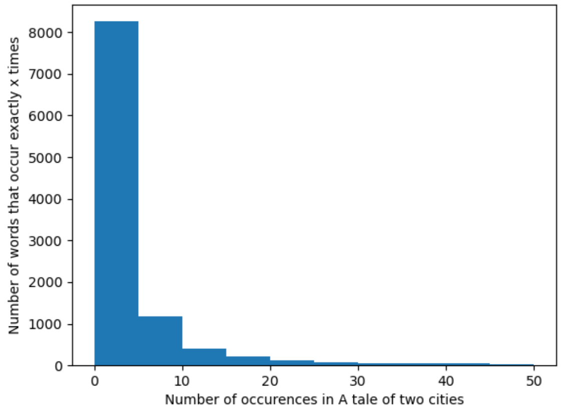

Amount of sentences with x words
250
200
150
100
50
0
0
10
20
30
40
50
60
70
Words per sentence in Camus's La peste (french)

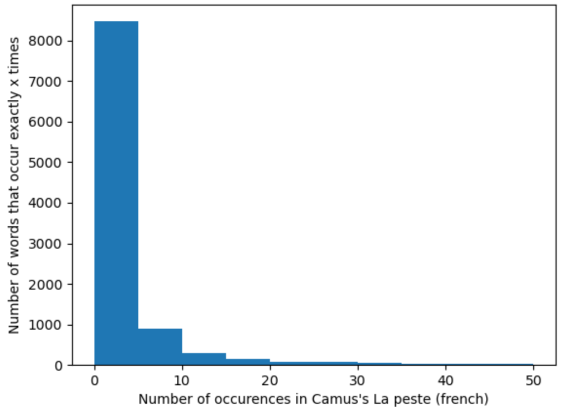

Amount of sentences with x words
200
150
100
50
0
0
10
20
30
40
50
60
70
Words per sentence in Camus's The plague (english)

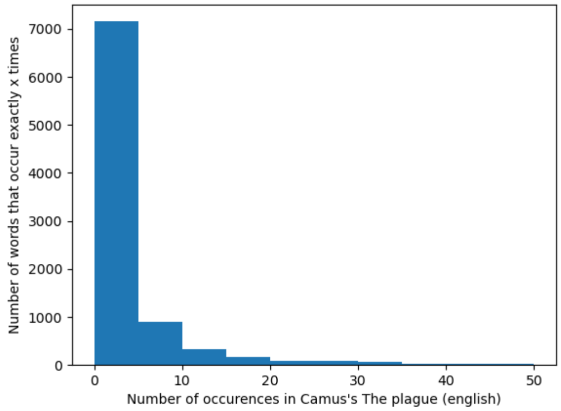

Amount of sentences with x words
250
200
150
100
50
0
0
10
20
30
40
50
60
70
Words per sentence in Camus's Die Peste (german)

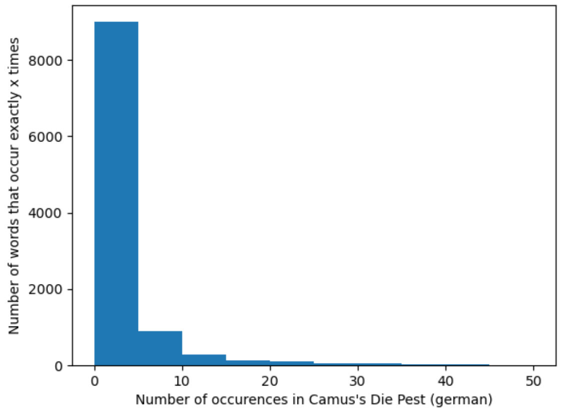

Amount of sentences with x words
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
Words per sentence in 2001 - A space odyssey

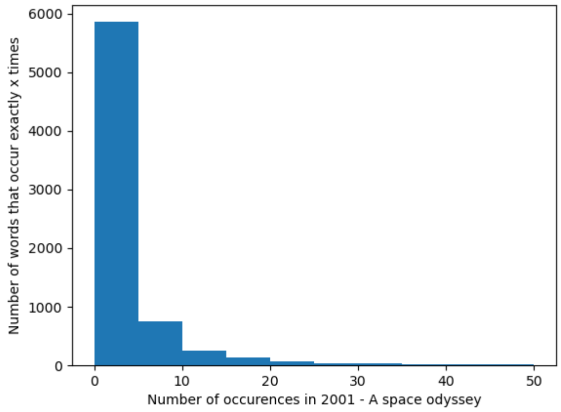

Number of words that occur exactly x times
3000
2500
2000
1500
1000
500
0
0
10
20
30
40
50
Number of occurences in "A brief history of time"

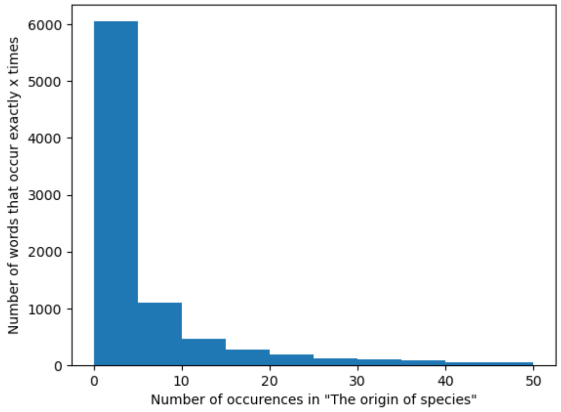

Number of words that occur exactly x times
8000
7000
6000
5000
4000
3000
2000
1000
0
0
10
20
30
40
50
Number of occurences in A tale of two cities

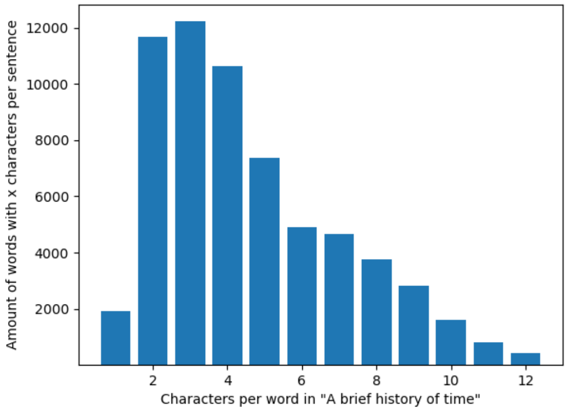

Number of words that occur exactly x times
0
1000
2000
3000
4000
5000
6000
7000
8000
0
10
20
30
40
50
Number of occurences in Camus's La peste (french)

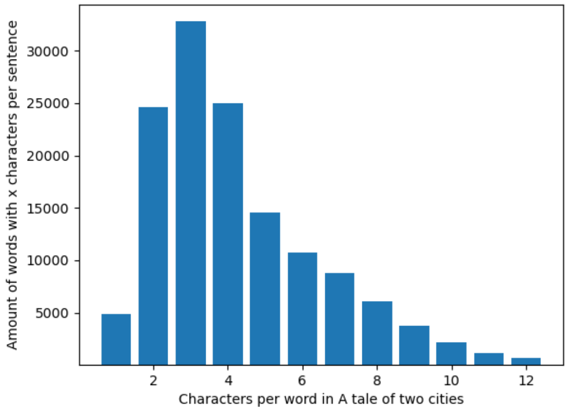

Number of words that occur exactly x times
0
1000
2000
3000
4000
5000
6000
7000
0
10
20
30
40
50
Number of occurences in Camus's The plague (english)

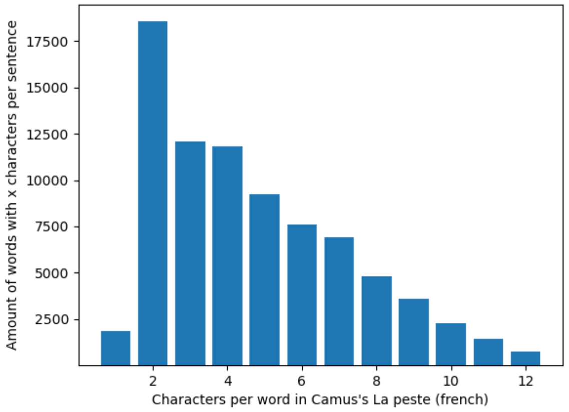

Number of words that occur exactly x times
0
2000
4000
6000
8000
0
10
20
30
40
50
Number of occurences in Camus's Die Pest (german)

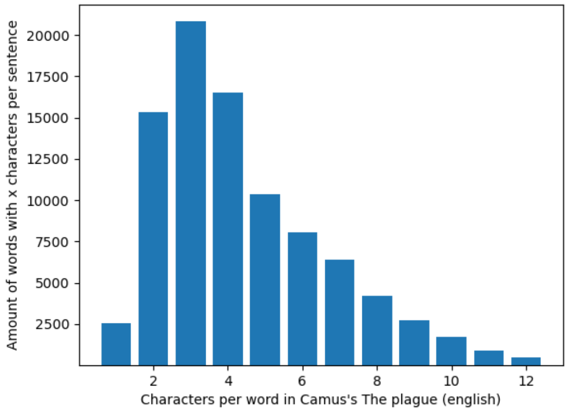

Number of words that occur exactly x times
0
1000
2000
3000
4000
5000
6000
0
10
20
30
40
50
Number of occurences in 2001 - A space odyssey

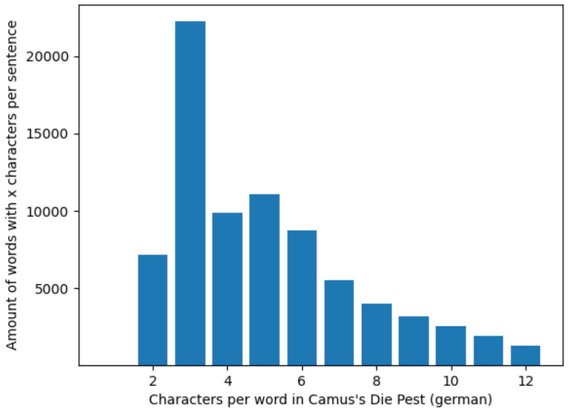

Number of words that occur exactly x times
0
1000
2000
3000
4000
5000
6000
0
10
20
30
40
50
Number of occurences in "The origin of species"

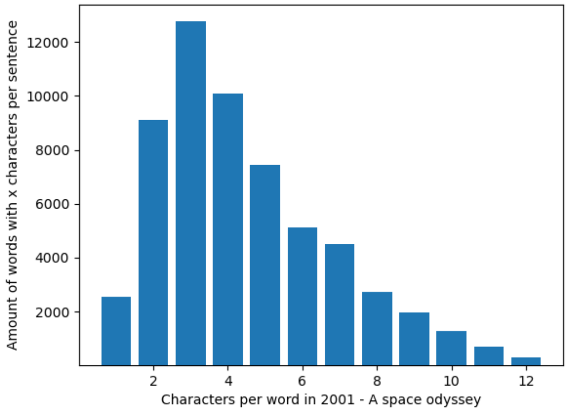

Amount of words with x characters per sentence
12000
10000
8000
6000
4000
2000
2
4
6
8
10
12
Characters per word in "A brief history of time"


Amount of words with x characters per sentence
30000
25000
20000
15000
10000
5000
2
4
6
8
10
12
Characters per word in A tale of two cities


Amount of words with x characters per sentence
17500
15000
12500
10000
7500
5000
2500
2
4
6
8
10
12
Characters per word in Camus's La peste (french)


Amount of words with x characters per sentence
20000
17500
15000
12500
10000
7500
5000
2500
2
4
6
8
10
12
Characters per word in Camus's The plague (english)


Amount of words with x characters per sentence
20000
15000
10000
5000
2
4
6
8
10
12
Characters per word in Camus's Die Pest (german)


Amount of words with x characters per sentence
12000
10000
8000
6000
4000
2000
2
4
6
8
10
12
Characters per word in 2001 - A space odyssey

References

[1] Wikipedia. History of writing. https://en.wikipedia.org/wiki/History_of_writing, 2021.

[2] M. Gerlach and F. Font-Clos. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics, 2018.

[3] Vscauce. The zipf mystery. https://www.youtube.com/watch?v=fCn8zs912OE, 2015.