Comparative Statistical Analysis of Word Frequencies in Human-Written and AI-Generated Texts

Anna Kudryavtseva[1*] (0009-0004-8577-9929), Artyom Kovalevskii[1,2,3] (0000-0001-5808-3134)

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Novosibirsk State Technical University, Novosibirsk, Russia
[3] Sobolev Institute of Mathematics, Novosibirsk, Russia
[*] Corresponding author's email: a.kudryavtseva@g.nsu.ru

DOI: https://doi.org/10.53482/2025_58_423

ABSTRACT

We classify texts using relative word frequencies. The task is to distinguish human-written texts from those generated by a computer using modern algorithms. We study two essay datasets, each containing an equal number of human-written and computer-generated essays. Studying Zipf diagrams shows that the generated texts have a significantly smaller vocabulary compared to human ones. However, the relative frequency of rare words (not included in the 1000 most common) does not allow us to confidently classify the texts. As additional features, we used the relative frequencies of the four most frequent words, as well as the ratio of the number of hapax legomena to the total number of different words. This feature allows to significantly improve the classification. Using these six features allows us to fairly confidently determine whether the text is computer-generated.

Keywords: Large Language Model, Zipf's Law, rare words.

1 Introduction

The advent of Large Language Models (LLMs) such as GPT-3/4 has opened entirely new possibilities in Artificial Intelligence (AI) text generation and radically changed the content of research in the field of AI.

Among the new research challenges arising from this event, one of the main ones is the problem of detecting texts created by AI, which is topical in various fields — from school and university education to information security. Intensive research in this direction is underway. In particular, such AI detection software tools as GPTZero (2023) and ZeroGPT (2024) have become widely known. Unfortunately, detecting LLM-generated texts is an intricate challenge, and until now the reliability of such software is debatable. In particular, in a study conducted by Weber-Wulff et al. (2023), researchers evaluated 14 detection tools, including GPTZero, and found that "all scored below 80% precision and only 5 above 70%."

A comprehensive review of different methods for the detection of AI-generated text is given by Wu et al. (2024). In this review, the detector techniques are divided into a few groups: watermarking techniques, statistics-based detectors, neural-based detectors, and human-assisted methods.

Recently, it was announced that a method had been developed to recognize machine-generated texts with a high degree of reliability (Hans et al., 2024). It is claimed that over a wide range of document types the method, called *Binoculars*, detects over 90% of generated samples from ChatGPT (and other LLMs) at a false positive rate of 0.01%, despite not being trained on any ChatGPT data. *Binoculars* belongs to neural-based detectors, it uses two LLMs, one is an "observer" LLM and another is a "performer" LLM.

One of possible approaches to distinguishing between human- and machine-generated texts can be based on a statistical analysis of the text vocabulary, in particular on investigating the usage of rarest and most frequent words.

Zipf (1949) and Mandelbrot (1965) showed that human texts approximately follow a power law of decreasing frequencies

$$f_r \simeq \frac{c}{(r+b)^a}, \tag{1}$$

where f_r is the relative frequency of word with rank r,

a is the Zipf exponent,

b is the Mandelbrot shift,

c is the normalizing constant.

However, Mandelbrot demonstrated that texts generated by a simple random algorithm also satisfy this law.

The distinction between human and machine texts may be found in the parameters of the law. It is known that these parameters vary over a fairly wide range, depending on the author, and are not constant for the entire language.

Piantadosi (2014) analyzed deviations of human language in the frequency distribution from the Zipf — Mandelbrot law and concluded that human language has a highly complex, reliable structure in the frequency distribution over and above this classic law.

Santis et al. (2024) studied the frequency distribution of words in novels and in texts generated by computer algorithms, but did not find a universal criterion for distinguishing them: "We have planned to go in depth on these interesting questions while maintaining the general claim that concerns the characterization of texts generated by machines with respect to some methodologies made available by the complexity sciences."

Two companion papers Abebe et al. (2022) and Abebe et al. (2023) explore the potential of the Heaps diagram (a process of counts of different words in a text) to analyze text homogeneity and find places where two different texts connect.

In addition to Zipf's law, studies highlight how temporal and structural factors shape word distributions. Altmann et al. (2009) demonstrate that word usage exhibits bursty patterns — clusters of high frequency followed by lulls — deviating from Poisson randomness and aligning with stretched exponential models. This variability is context-dependent, reflecting semantic and pragmatic influences.

Further complexity arises from the interplay of word properties and sentence structure. Popescu et al. (2009) reveals that relative word frequencies correlate with inherent linguistic features: shorter words and polysemous terms (e.g., "run") tend to occur more frequently, while morphological complexity reduces usage rates. Critically, positional dynamics in sentences also govern frequency—low-frequency words disproportionately occupy informationally salient positions, such as sentence-final slots, due to their role in conveying new or emphatic content.

Beyond these intrinsic and syntactic factors, variability across texts introduces additional stochasticity: Gerlach and Altmann (2014) demonstrate that vocabulary size exhibits Taylor's law (Taylor, 1961), where fluctuations in word diversity persist even for long texts, scaling linearly with the mean due to topic-driven heterogeneity. This quenched disorder — rooted in topical variations rather than pure randomness — renders vocabulary growth non-self-averaging, meaning lexical richness cannot be disentangled from contextual or discourse-level shifts.

These findings underscore that word frequency is not merely a function of statistical ubiquity but is mediated by syntactic roles, semantic richness, discourse structure, and systemic variability across textual domains.

Thus, while Zipf's law describes the global distribution of word frequencies, the interplay of burstiness, lexical properties, and positional constraints reveals finer-grained linguistic mechanisms that transcend frequency alone.

In the present paper, we study the statistical characteristics of AI-generated texts and compare them with those of human-written essays. For this purpose, two datasets are analyzed.

The human essays of the first dataset are taken from a project by Morgan (2012) aimed at developing an automated scoring algorithm for student-written essays. They are available from Kaggle (https://www.kaggle.com), a data science competition platform, and contain responses to a single prompt written by students.

The essays are analyzed and compared along with the essays generated by an LLM from the same prompt.

For this purpose, we used one of the most powerful freely available LLM (NousResearch, 2023). The generated essays can be found on the GitHub page https://github.com/kudrann/ai-human-data.

The second dataset is collected by Verma et al. (2023) who have been developing *Ghostbuster*, a system for detection of AI-generated texts. The dataset includes high school and university level essays taken from the IvyPanda web site (https://ivypanda.com/essays/) as well as LLM-generated essays prepared by Ghostbuster developers. They used ChatGPT to first generate a prompt corresponding to each human essay and then generate a corresponding essay that responds to that prompt. The full dataset can be found on their Github page https://github.com/vivek3141/ghostbuster-data.

2 Methodology

It is well known that in many natural languages the frequency of a word f is roughly inversely proportional to its number (rank) r in the list of the most frequent words, $f \sim 1/r$. This empirical relation is known as Zipf's law (Zipf, 1949). In fact, in many cases a generalized version of this relation known as the Zipf — Mandelbrot (ZM) law (1) works better (Mandelbrot, 1965). As an example, the frequency of words in the classic novel "Dracula" by Bram Stoker is shown in Figure 1 using the log-log scale.

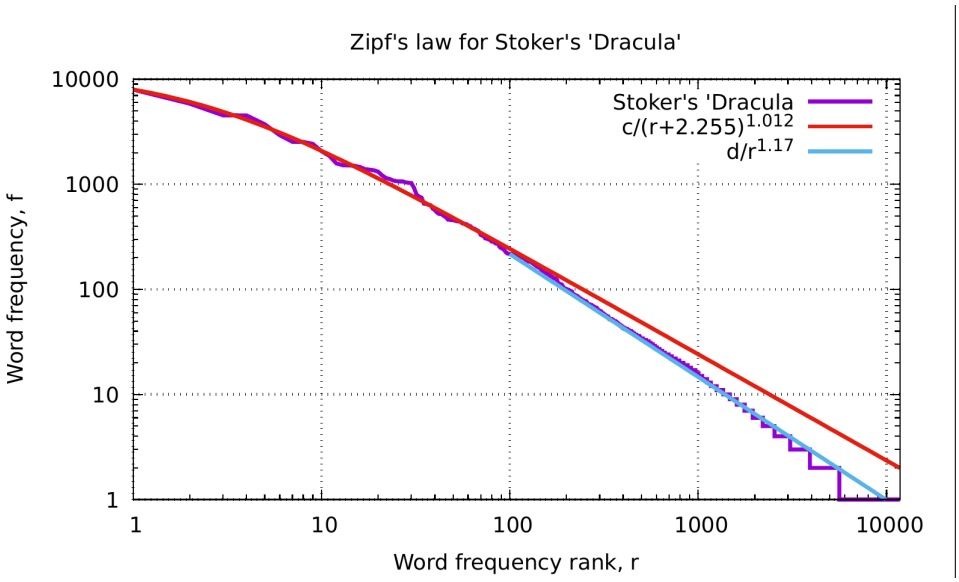

Figure 1: Word frequency distribution in the novel "Dracula" by B. Stoker.

The frequency distribution of the words is seen to be in close agreement with the ZM law at $a = 1.012$ and $b = 2.255$ for words whose rank is less than approximately 110. At the same time, the distribution of rare words deflects noticeably and cannot be described by the function with the same values of a and b. In fact, it can be better fitted by the function $\sim 1/r^a$ with $a = 1.17$ (Figure 1).

It can be expected that the distribution of rare words is specific to different authors and may be considered as an important characteristics of an author's style. In particular, one can assume that human-written and AI-generated texts can differ in statistical properties of distribution for rare words. So, we pay special attention to analyzing their usage in the essays.

For text analysis, a Python code was written using the text processing library *collections*. After preprocessing (removing punctuation and capitalization, splitting into separate words), it allows us to construct word frequency distributions, determine the parameters of the distributions, study the scatter in the frequency of rare words, and so on. The results of its work are presented below.

3 Results

3.1 The First Dataset

The essays in the dataset of Morgan (2012) were written as a response to the following prompt.

> *More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends. Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.*

800 essays are selected from this database for statistical analysis. They are compared with the same number of essays generated by *Nous-Hermes-Llama2* (NousResearch, 2023), one of the most powerful freely available LLM, containing 13 billion parameters. Using the LMStudio application (LMStudio, 2024), the model quantized to 8 bits was installed on a compute cluster with 8 Nvidia GeForce GTX 1080 graphics processing units (GPUs) and 11.264 GB of video memory on each GPU. 800 essays were generated with the temperature value (a parameter that determines the degree of difference between the generated essays) $T = 0.7$. The average length of essay is 284 words, the average generation time is 21 s.

An example of the generated essay is given below:

> *Dear Editor,*
>
> *I am writing this letter to express my thoughts on the impact of computers on society.*

As technology advances and more people become reliant on computers, it is essential to consider both the benefits and drawbacks of this development.

On one hand, computers have undoubtedly made our lives easier in many ways. They provide access to a wealth of information, allowing us to learn about any topic instantly, communicate with people across the globe, and perform tasks more efficiently. In addition, they help develop important skills such as hand-eye coordination and problem-solving.

However, there are also concerns that excessive computer use can lead to negative consequences. People may spend too much time in front of screens, neglecting their physical health, social interactions, and relationships with family and friends. Moreover, the widespread use of computers has led to job losses in some sectors, causing economic hardships for many individuals.

In conclusion, while computers have revolutionized our lives in numerous ways, it is crucial that we strike a balance between embracing technology and maintaining our physical, mental, and social well-being. By being mindful of the potential drawbacks and taking steps to mitigate them, we can ensure that computers continue to benefit society positively.

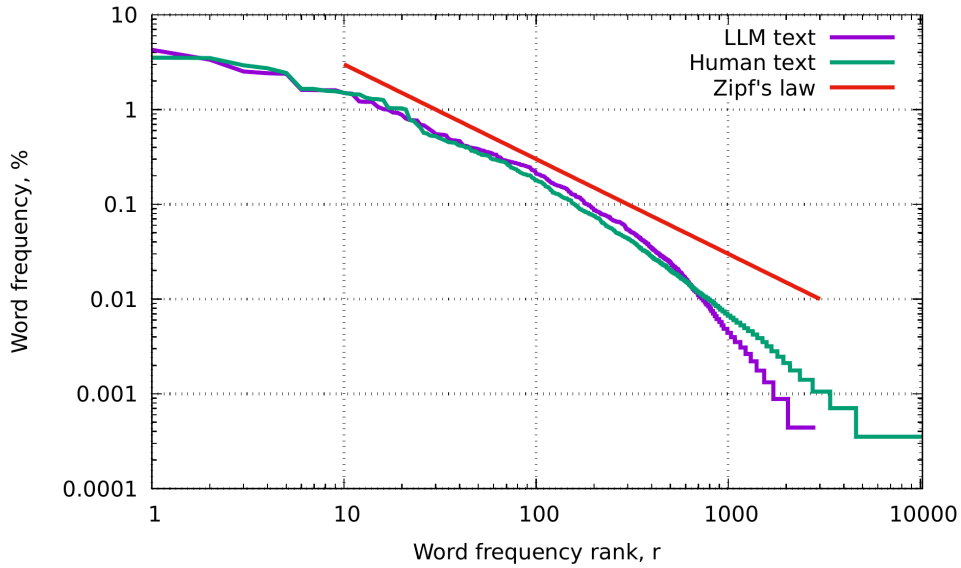

Figure 2: Word frequency distributions in the human-written and AI-generated essays. Dataset #1.

The word frequency distributions for the human-written and AI-generated essays are compared in Figure 2.

The word frequencies of both human-written and AI-generated essays deviate significantly from Zipf's law, especially if one looks at the "tails" of the distributions.

It is worth noting that there are also some distinctions in the list of the most frequent words — see Table 1.

Table 1: The most frequent words in human-written and AI-generated essays. Dataset #1.

Rank	1	2	3	4	5	6	7	8	9	10
					Humans					
Word	the	to	and	you	are	computers	on	people	of	that
Frequency	10029	9934	8308	7755	6916	4695	4682	4478	4419	4268
Percentage	3.53	3.50	2.93	2.73	2.44	1.65	1.65	1.58	1.56	1.50
					LLM					
Word	and	to	the	of	computers	have	that	in	on	with
Frequency	9707	7631	5730	5482	5418	3670	3642	3638	3632	3398
Percentage	4.27	3.36	2.52	2.41	2.38	1.62	1.60	1.60	1.60	1.50

In order to find the best fits of these distributions to the ZM law (1) we estimate the a and b constants by solving a nonlinear least-squares problem with the Levenberg—Marquardt (damped least-squares) algorithm (Gill et al., 1981, pp. 136-137). As a result, the constants are found to be $a = 1.235$, $b = 7.551$ for the human-written essays and $a = 1.035$, $b = 4.17$ for the AI-generated ones — see Figure 3 and Figure 4.

Thus, the parameters of the ZM fittings differ noticeably for two distributions. Moreover, in both cases the distributions are in reasonable agreement with the ZM law for word ranks $r \lesssim 300$, though there is some visible deviation from the ZM law in the range $50 < r < 200$ for the LLM-generated essays.

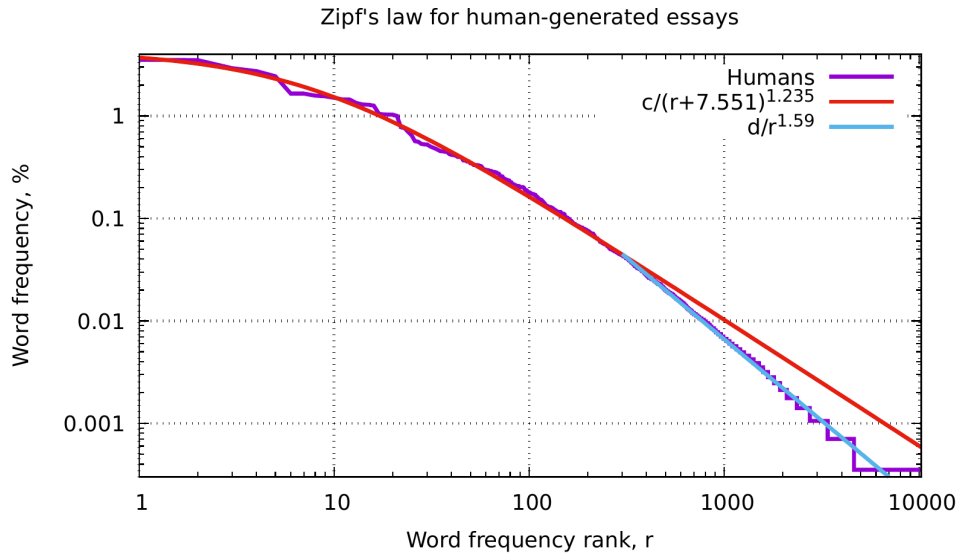

Figure 3: Word frequency distribution in the human-written essays compared with the ZM law. Dataset #1.

As concerns rare words, both distributions decay much faster than their calculated ZM fittings. The

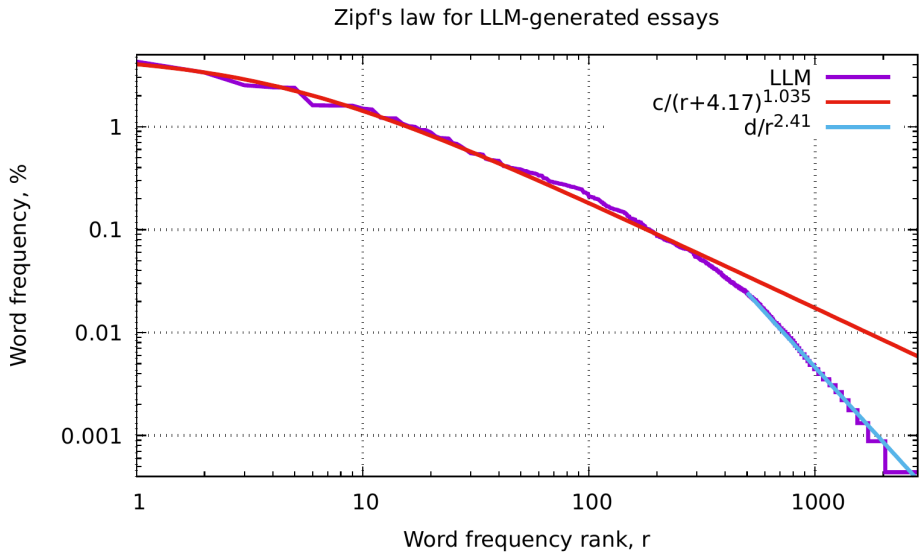

Figure 4: Word frequency distribution in the AI-generated essays compared with the ZM law. Dataset #1.

distribution tails can be better fitted by d/r^a functions with different values of a. As is seen in Figure 3 and Figure 4, the distribution for human-written essays at $r \geq 300$ are fitted well with $a = 1.59$ while the distribution for AI-generated essays at $r \geq 500$ better corresponds to $a = 2.41$. The exponents in the power law are significantly different. Also, it is seen that the transition to the power-law distributions happens for the AI-generated essays at a noticeably larger value of r than it does for the human-written ones ($r = 500$ instead of $r = 300$).

Peculiarities in the distribution tails prompted us to take a closer look at uncommon words occurring in the essays. In Figure 5 the proportion of uncommon ($r > 1000$) words is shown as a function of the essay number. It can be concluded that the average proportion of uncommon words in the human-written essays is much higher than in the AI-generated ones. Additionally, there are some essays composed by students in which the proportion is very high.

Our hypothesis is that the proportion of rare words remains stable for a homogeneous text of one author, but varies significantly between authors. The correlation coefficient between the proportion of rare words and the length of the essay in words is $corr = 0.64$. This confirms the difference between authors and also indicates that authors with a richer vocabulary write longer texts on average. The dependence of the proportion of rare words on errors and typos is analyzed in detail below.

In Table 2 the maximum proportion of uncommon words, their average proportion and the standard deviation are given for both sets of essays. One human-written essay contains 57.6% of uncommon words!

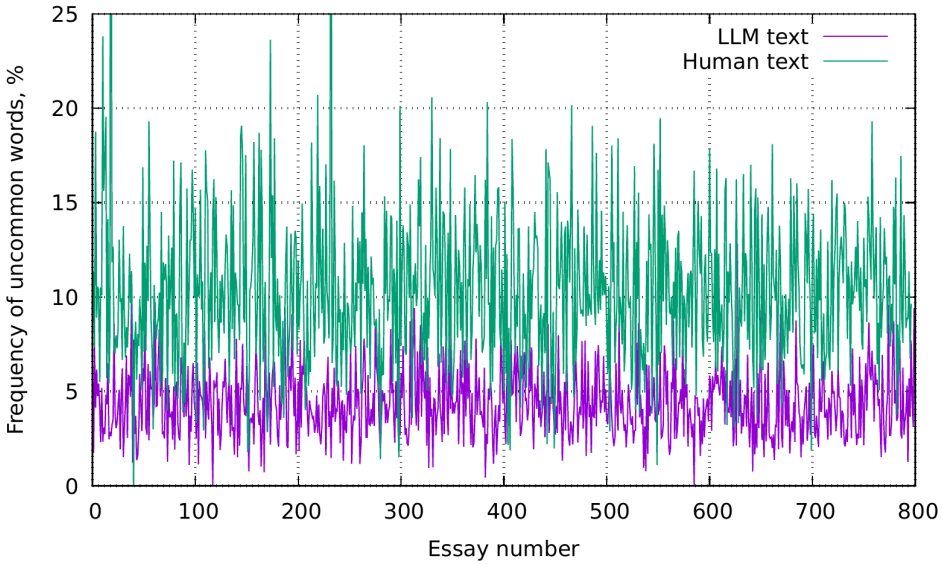

Figure 5: Frequency of uncommon words ($r > 1000$) in essays. Dataset #1.

Table 2: The proportion of uncommon words ($r > 1000$) in human-written and AI-generated essays. Dataset #1.

	Max. fraction	Mean value	Standard deviation
LLM	12.8%	4.4%	1.9%
Humans	57.6%	10.4%	4.3%

The "record-breaking" essay with 57.6 % of words not in the top 1000 looks like as follows.

> *I aegre waf the evansmant ov tnachnololage. The evansmant ov tnachnolige is being to halp fined a kohar froi alnsas. Tnanchnololage waf ont ot we wod not go to the moon. Tnachnologe evans as we maech at. The people are in tnacholege to the frchr fror the good ov live. Famas invanyor ues tnacholage leki lena orde dvanse and his fling mashine. Tnachologe is the grat.*

Spelling errors make this text virtually incomprehensible.

As can be seen from Fig. 5 and Table 2, the average proportion of uncommon words in the human-written essays is about 10%. There are some essays in which the proportion is noticeably higher, but the number of such essays does not seem high. To investigate the relationship between the number of uncommon words and that of orthographical mistakes, we analyzed 20 randomly chosen essays. There are mistakes and typos in all 20 essays. In 17 of them, their proportion does not exceed 4 %, there is also one essay each with 8, 9 and 11 % of mistakes and typos. The correlation between the number of words in an essay and the percentage of mistakes is weakly negative ($corr = -0.33$) and insignificant (p-value, statistical significance is $p = 0.16$). It is expected as poorly proficient students write shorter essays. The correlation between the percentages of mistakes and uncommon words is weakly positive ($corr = 0.30$)

and insignificant ($p = 0.20$). Thus, mistakes and typos contribute to the frequency of rare words, but their contribution is not decisive.

Thus, it can be concluded that the differences in statistical characteristics of human-written and AI-generated essays are caused, at least partially, by spelling errors inherent in humans.

3.2 Classifications of Texts of the First Dataset

We classify texts using C-Support Vector Classification (Pedregosa et al., 2011) with the parameter kernel='linear', see https://scikit-learn.org/dev/modules/generated/sklearn.svm.SVC.html.

Firstly, we use only one feature, namely x_1, which represents the fraction of uncommon words in the dataset ($r > 1000$). The model uses 75 % of the data for training and other 25 % for testing. The corresponding parts of the sets of human and AI-generated texts are selected at random.

The AI-generated texts are designated as the positive class, while the human texts are designated as the negative class. Thus, True Positive (TP) denotes the number of AI-generated texts correctly classified as AI-generated, False Positives (FP) — human-written texts incorrectly classified as AI-generated, True Negatives (TN) — human-written texts correctly classified as human-written and . False Negatives (FN) — AI-generated texts incorrectly classified as human-written.

The test is based on the pre-trained set and the test sample, which comprises 200 human texts and 200 AI-generated texts. In this test,

TP = 176, FP = 40, TN = 158, FN = 26, so that the accuracy = 0.835.

In order to obtain a more accurate classification, we use additional features of texts:

x_2 is the percentage of word "the",

x_3 is the percentage of word "and",

x_4 is the percentage of word "you",

x_5 is the percentage of word "are",

x_6 is the proportion of hapax legomena.

The features x_2, x_3, x_4, x_5 are selected on the base of Table 1 as words with the greatest differences in percentages.

The choice of feature x_6 is based on Figure 2. The last step (horizontal segment) of the relative frequency graph corresponds to hapax legomena. This step is significantly shorter in the set of AI-generated texts than in the human ones.

The Zipf parameter can be estimated by the inverse value, i.e. by dividing the number of different words by the number of hapax legomena. This estimate was proposed within the framework of the elementary probability model in Ohannessian and Dahleh (2012), its properties are studied in Chebunin and Kovalevskii (2019). In particular, the corresponding statistical test allows us to study the significance of differences in the number of hapax legomena. The correspondence of texts to the elementary probabilistic Zipf's model from the point of view of this statistics was studied in Fayzullaev and Kovalevskii (2024). Davis (2018) proposed and investigated an interesting and very precise relationship between the number of different words and the number of hapax legomena. Another interesting model for the number of hapax legomena was formulated by Milička (2009).

Using these 6 features, we have under the same approach for the same training and test sets of texts:

TP = 201, FP = 5, TN = 193, FN = 1, so we have 6 mistakes overall, and the accuracy = 0.985.

Our optimal linear classifier produces the following weights for the features (Table 3).

Table 3: Optimal linear classifier for dataset #1.

Feature	x_1	x_2	x_3	x_4	x_5	x_6
Importance	0.048	0.086	0.070	0.470	0.206	0.120

3.3 The Second Dataset

The analyzed texts consist of 1000 essays written by students and 1000 texts of approximately the same length generated by ChatGPT using prompts extracted from the students' essays. The word frequency distributions for the human-written and AI-generated essays are shown in Figure 6.

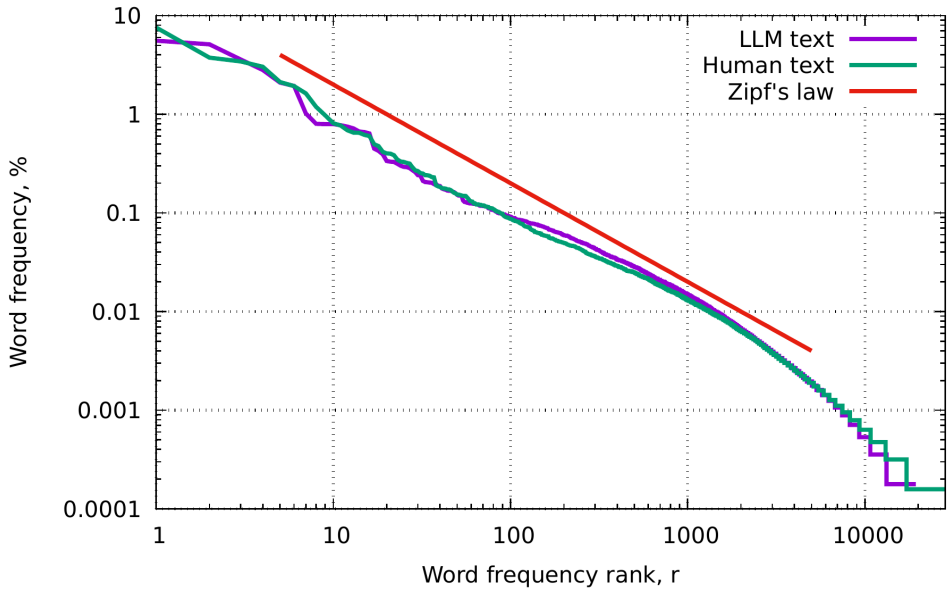

Figure 6: Word frequency distributions in the human-written and AI-generated essays. Dataset #2.

It can be seen that both distributions follow Zipf's law (not very precisely) up to $r \simeq 80 \div 100$. Their shapes for rarer words are very similar but clearly do not match the power-law distribution. It is worth noting that the distributions are much closer to each other than it was for the first dataset.

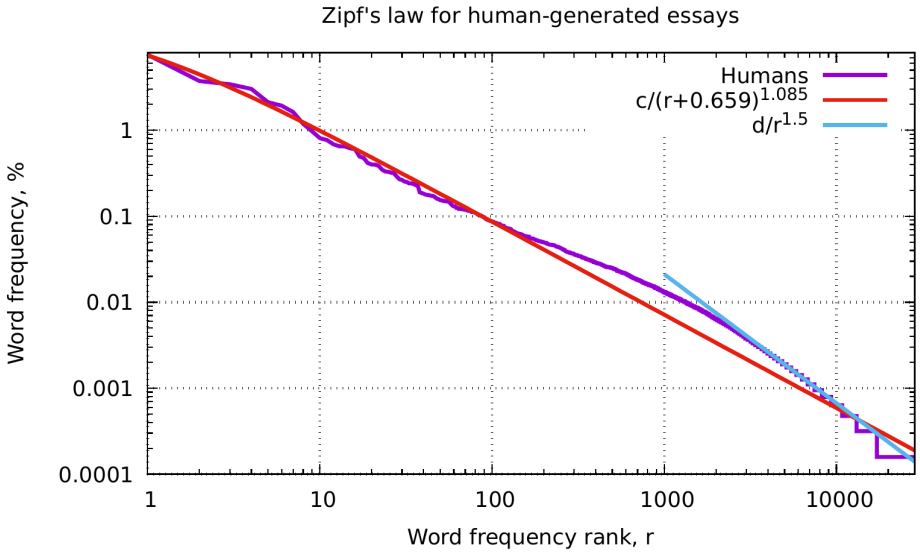

Figure 7: Word frequency distribution in the human-written essays compared with the ZM law. Dataset #2.

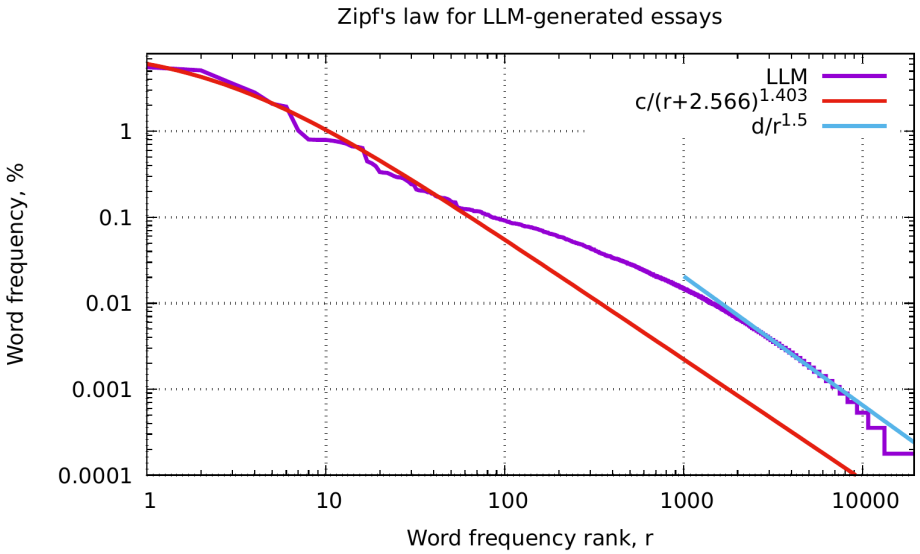

Figure 8: Word frequency distribution in the AI-generated essays compared with the ZM law. Dataset #2.

Least-square fitting of the word frequency distributions to the ZM law has been performed and the resulted best fits are compared with the distributions themselves for human-written and AI-generated essays in Figs. 7 and 8, respectively. The best fitting parameters are $a = 1.085$, $b = 0.659$ for the former curve and $a = 1.403$, $b = 2.566$ for the latter.

As can be seen, the deviations from the ZM law are particularly large for rare words, at $r > 150$ for the human-written essays and at $r > 60$ for the AI-generated ones. This is because the least-square fitting procedure primarily seeks to reduce errors at small values of r for which word frequencies are high. At the same time, the word frequency distributions for rare words follow the power law with the same exponent $d = 1.5$, much steeper than Zipf's law.

Frequencies of uncommon words (Figure 9) show that, for the second dataset, there are no such pronounced difference in their averaged and maximum fractions between human-written and AI-generated essays.

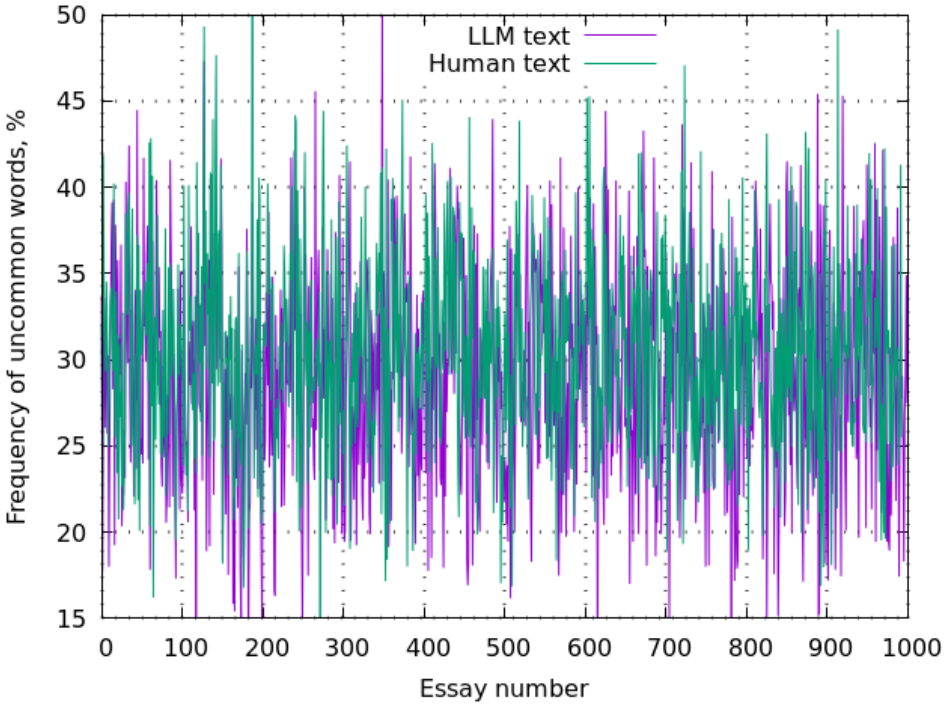

Figure 9: Frequency of uncommon words ($r > 1000$) in essays. Dataset #2.

The maximum and averaged fractions of uncommon words in human-written and AI-generated essays as well as the values of standard deviation are given in Table 4.

Table 4: The fraction of uncommon words ($r > 1000$) in human-written and AI-generated essays. Dataset #2.

	Max. fraction	Mean value	Standard deviation
LLM	50.7%	29%	6.2%
Humans	58.6%	30.4%	5.8%

The text classification based on the same one and six features as above has also been performed for the second dataset. We have for one feature:

TP = 127, FP = 119, TN = 129, FN = 124, the accuracy = 0.513, there are many mistakes.

For six features:

TP = 214, FP = 32, TN = 216, FN = 37, the accuracy = 0.862.

Overall, one can see that, in the second dataset, computer and human texts are not so easily distinguished. Using one feature leads to a large number of errors, and increasing the number of features allows us to significantly improve the accuracy.

The optimal linear classifier is shown in Table 5.

Table 5: Optimal linear classifier for dataset #2.

Feature	x_1	x_2	x_3	x_4	x_5	x_6
Importance	0.151	0.102	0.254	0.014	0.121	0.358

4 Conclusion

Computer-aided text generation is becoming increasingly common in essay writing. The present study contributes to the recognition of computer-aided text generation. This study is based on relative word frequencies and allows for the combination of the proposed methods with other methods for recognizing computer-aided text generation. A Python code has been developed for analyzing statistical features of word usage in different texts using the well-known *collections* library. Analysis of the entire text array reveals significant differences in the relative frequencies of the most common words, as well as in the total vocabulary size.

The relative frequency of rare words alone is not sufficient for confident recognition. A more accurate algorithm uses, in addition, the relative frequencies of the four most common words, as well as the ratio of hapax legomena to the total number of different words.

Dataset 1 shows a classification accuracy of 0.835 using only the relative frequency of rare words (r > 1000), and an accuracy of 0.985 using six features, with the greatest contribution to the classification coming from the personal pronoun “you”, the verb “are” and the proportion of hapax legomena. Dataset 2 shows an accuracy of 0.513 when using only the first feature, and an accuracy of 0.862 when using six features. Here, the largest weights are given to the proportion of hapax legomena, the relative frequency of the conjunction “and” and the relative frequency of rare words.

It is worth noting that, as follows from the investigation of Dataset 1, typos and orthographical mistakes common in human-written text can, to some extent, contribute to statistical features of word frequency distributions. Thus, a study comparing AI-generated texts with those written by real humans but contain no errors – whether originally or after correction – could shed even more light on the topic under study. However, it would require the use of an automatic spell-checking tool or the compilation of mistake-free

essay databases.

Acknowledgments

We are grateful to an anonymous reviewer for his remarks that helped us improve the paper. The work is supported by Program of Fundamental Scientific Research of the SB RAS, project FWNF-2022-0010.

References

Abebe, B., Chebunin, M., Kovalevskii, A. (2023). Text segmentation via processes that count the number of different words forward and backward. *Journal of Quantitative Linguistics*, *31*(1), pp. 1–18. https://doi.org/DOI: 10.1080/09296174.2023.2275342

Abebe, B., Chebunin, M., Kovalevskii, A., Zakrevskaya, N. (2022). Statistical tests for text homogeneity: Using forward and backward processes of numbers of different words. *Glottometrics*, *53*(1), pp. 42–58. https://doi.org/10. 53482/2022_53_401

Altmann, E. G., Pierrehumbert, J. B., Motter, A. E. (2009). Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words. *PLOS One*, *4*(11), e7678. https://doi.org/10.1371/journal.pone.0007678

Chebunin, M., Kovalevskii, A. (2019). Asymptotically Normal Estimators for Zipf's Law. *Sankhya A*, *81*(2), pp. 482–492. https://doi.org/10.1007/s13171-018-0135-9

Davis, V. (2018). Types, Tokens, and Hapaxes: A New Heap's Law. *Glottotheory*, *9*(2), pp. 113–129. https://doi.org/ 10.1515/glot-2018-0014

Fayzullaev, S., Kovalevskii, A. (2024). Hapax legomena via stochastic processes. *Glottometrics*, *56*, pp. 22–39. https://doi.org/10.53482/2024_56_415

Gerlach, M., Altmann, E. G. (2014). Scaling laws and fluctuations in the statistics of word frequencies. *New Journal of Physics*, *16*, 113010. https://doi.org/10.1088/1367-2630/16/11/113010

Gill, P. E., Murray, W., Wright, M. H. (1981). *Practical Optimization*. Academic Press.

GPTZero. (2023). *The Global Standard for AI Detection: Humans Deserve the Truth*. Retrieved October 10, 2024, from https://gptzero.me

Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., Goldstein, T. (2024). *Spotting LLMs with Binoculars: Zero-Shot Detection of Machine-Generated Text*. Retrieved October 10, 2024, from https://arxiv.org/pdf/2401.12070

LMStudio. (2024). *LM Studio. Discover, Download, and Run Local LLMs*. Retrieved October 10, 2024, from https://lmstudio.ai

Mandelbrot, B. (1965). Information Theory and Psycholinguistics. In: Wolman, B. B., Nagel, E. (Eds.). *Scientific Psychology*, pp. 550–562. Basic Books.

Milička, J. (2009). Type-token & hapax-token relation: A combinatorial model. *Glottotheory*, *2*(1), pp. 99–110. https://doi.org/10.1515/glot-2009-0009

Morgan, J. (2012). *The Hewlett Foundation: Automated Essay Scoring*. Retrieved October 10, 2024, from https: //www.kaggle.com/competitions/asap-aes/data

NousResearch. (2023). *Model Card: Nous-Hermes-Llama2-13b*. Retrieved October 10, 2024, from https : / / huggingface.co/NousResearch/Nous-Hermes-Llama2-13b

Ohannessian, M. I., Dahleh, M. A. (2012). Rare probability estimation under regularly varying heavy tails. In: Mannor, S., Srebro, N., Williamson, R. C. (Eds.). *Proceedings of the 25th annual conference on learning theory*, pp. 21.1–21.24, Vol. 23). PMLR.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, pp. 2825–2830.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic bulletin & review*, *25*(5), pp. 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Popescu, I.-I., Grzybek, G. A. P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M. N. (2009). *Word Frequency Studies* (1st ed.). Mouton de Gruyter.

Santis, E. D., Martino, A., Rizzi, A. (2024). Human Versus Machine Intelligence: Assessing Natural Language Generation Models Through Complex Systems Theory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(7), pp. 4812–4829. https://doi.org/10.1109/TPAMI.2024.3358168

Taylor, L. M. (1961). Aggregation, variance and the mean. *Nature*, *189*, pp. 732–735. https://doi.org/https://doi.org/ 10.1038/189732a0

Verma, V., Fleisig, E., Tomlin, N., Klein, D. (2023). *Ghostbuster: Detecting Text Ghostwritten by Large Language Models*. Retrieved February 16, 2025, from https://arxiv.org/pdf/2305.15047

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., Waddington, L. (2023). Testing of Detection Tools for AI-generated Text. *International Journal for Educational Integrity*, *19*(1), 26. https://doi.org/10.1007/s40979-023-00146-z

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., Chao, L. S. (2024). *A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions*. Retrieved October 10, 2024, from https://arxiv.org/pdf/2310.14724

ZeroGPT. (2024). *Trusted GPT-4, ChatGPT and AI Detector tool by ZeroGPT*. Retrieved October 10, 2024, from https://www.zerogpt.com

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.