

Article

Some Properties of Zipf's Law and Applications

Speranta Cecilia Bolea ^{1,*}, Mironela Pirnau ², Silviu-Ioan Bejinariu ¹, Vasile Apopei ¹, Daniela Gifu ^{1,3}
and Horia-Nicolai Teodorescu ^{1,4,*}

¹ Institute of Computer Science, Romanian Academy Iași Branch, 700481 Iași, Romania; silviu.bejinariu@iit.academiaromana-is.ro (S.-I.B.); vasile.apopei@iit.academiaromana-is.ro (V.A.); daniela.gifu@iit.academiaromana-is.ro (D.G.)

² Faculty of Informatics, Titu Maiorescu University, 040441 Bucharest, Romania; mironela.pirna@prof.utm.ro

³ Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iași, 700483 Iași, Romania

⁴ Department Applied Electronics, Technical University of Iași, 700050 Iași, Romania

* Correspondence: cecilia.bolea@iit.academiaromana-is.ro (S.C.B.); hteodor@etti.tuiasi.ro (H.-N.T.)

Abstract: The article extends the theoretical and applicative analysis of Zipf's law. We are concerned with a set of properties of Zipf's law that derive directly from the power law expression and from the discrete nature of the objects to which the law is applied, when the objects are words, lemmas, and the like. We also search for variations of Zipf's law that can help explain the noisy results empirically reported in the literature and the departures of the empirically obtained nonlinear graph from the theoretical linear one, with the variants analyzed differing from Mandelbrot and lognormal distributions. A problem of interest that we deal with is that of mixtures of populations obeying Zipf's law. The last problem has relevance in the analysis of texts with words with various etymologies. Computational aspects are also addressed.

Keywords: Zipf's law; power law; distribution; number theory; speech processing; natural language processing (NLP); etymology

MSC: Zentralblatt 11M; 60E99; 68T50; 94A16 ACM Classes G.2 and G.3; ACM Subject Class I.2.7; ACM Subject Class C.2



Citation: Bolea, S.C.; Pirnau, M.; Bejinariu, S.-I.; Apopei, V.; Gifu, D.; Teodorescu, H.-N. Some Properties of Zipf's Law and Applications. *Axioms* **2024**, *13*, 146. <https://doi.org/10.3390/axioms13030146>

Academic Editor: Hari Mohan Srivastava

Received: 24 January 2024

Revised: 17 February 2024

Accepted: 19 February 2024

Published: 23 February 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Whenever one deals with a process involving randomness (stochastic process), it is of interest to empirically determine the frequencies of the various occurring events. The ratios of these frequencies divided by the total number of occurrences are named relative frequencies; the frequencies provide an estimation of the probabilities of the events, which, together, compose the distribution of probabilities of the process. George K. Zipf (1902–1950), a Harvard graduate of 1923, determined the frequencies of the words in large collections of texts and found that, after arranging the words w , according to their relative frequencies, f_w , from the highest to the lowest, the relative frequencies were approximatively proportional to the inverse of word ranks, r_w , $f(r_w) \approx A/r_w$, where A was a constant depending on the number of words in the analyzed text. This became known as Zipf's law, and the probability distribution $p(w_r) = A/r_w$ is known as the Zipf's distribution. Zipf's law is a member of a larger family of distributions, frequently named the power law distribution, defined by $p(r) = A/r^a$, $a > 0$. Often, especially when a power law is determined for collections of texts, $p(r) = A/r^a$ is still referred to as Zipf's law. It is common to take the logarithm of the relation defining Zipf or power laws, $\ln(p(r)) = \ln A - a \ln(r)$; this is the equation of a line $y = c + ax$, where $y = \ln(p(r))$, $c = \ln A$, and $x = \ln r$. Zipf went on to find that the populations of the towns and cities in the US also obey a power law. Zipf, who had a professorship at Harvard until his death, tried to construct an explanation (i.e., generative model) for this specific distribution

occurring in many social processes, including language. One of his last works, “relative frequency and dynamic equilibrium in phonology and morphology”, presented in 1949 at a conference in Paris, France, was devoted to the subject. However, we still do not have a convincing model that explains Zipf’s law in linguistics. The purpose of this study is to reveal and explain some elementary properties of Zipf’s law.

There are innumerable papers on Zipf’s law dealing with theoretical aspects and applications. We analyze some elementary consequences of the power law (Zipf’s distribution) in the context of populations that typically satisfy the condition that the elements of the population have a feature expressed by a natural number, and thus they can be ranked. Examples include the set of cities in a country, where the feature is the number of people living in the city, a set of papers, with the feature defined by the number of citations, and the set of words (or lemmas) in a text, with the feature expressed by the number of apparitions of the word. A minor difficulty is that two elements in a set may have equal values of the feature, for example, two words may have the same number of occurrences. This is formally solved by randomly assigning successive ranks to those elements; alternatively, one can assign fractional ranks, but this is a departure from the intuitive concept of rank.

2. Related Work

Zipf investigated the frequency of words in larger texts and found that their frequency of apparition, $p(w)$, where w denotes words, can be well described by the law $p(w) \approx \frac{A}{r(w)}$, where A is a constant and $r(w)$ is the rank of the word w . Therefore, on a double-logarithmic graph, the law is represented by a line with slope -1 , $\log p(w) = \log A - \log(r(w))$. Further studies have shown that a better empirical law for word frequencies and for other applications would be $p(w) \approx \frac{A}{r^a(w)}$, where $a \in \mathbf{R}$, $a > 0$, with a typically close to 1, but in some cases taking values in the interval $[1, 2]$. The last case corresponds to the Pareto law. Zipf also studied the distribution of cities ranked by their population. By counting the number of cities $n(S)$ with a population (size) larger than S , Zipf derived that $n(S) \sim 1/S$; see [1].

Since Zipf’s publications, many authors and studies have contributed to enlarging the number of applications and to find explanations and models for Zipf’s law. In one research direction, generalizations of Zipf’s law were proposed; the most notable ones are the Zipf–Mandelbrot law, $p(r) = \frac{A}{(r+b)^a}$, $b \in \mathbf{R}$, $b > 0$, and the lognormal distribution, which may better fit some applications [2,3]. Another direction of research on Zipf’s law is explanatory and regards economic, social, linguistic, and psychologic models that can generate Zipf-like distributions [1,4].

The most prolific research direction regarding Zipf’s law has been the applicative one: finding applications verifying the laws with a reasonable approximation and the range of parameters that determine a good fit. Application domains comprise linguistics [5–7], medicine [8,9], arts [10,11], computer science [12,13], communications [14], signal processing [15–17], economics [4], including firm growth and death [18,19], demography [1,20], patents [21,22], tourism [23], and speech [24].

3. Zipf’s Law over Finite Multisets—Properties

When applying Zipf’s law to texts or other similar applications, we have to observe specific properties of the investigated objects. While this is problematic from a linguistic point of view, considering that Zipf’s law is equally valid for lemmas and words, for brevity, we use the term “word” to designate either a word or its lemma. The discussion uses the strong assumption that word occurrences are independent. Also, we assume that:

- (i). the objects in the population belong to a discrete, finite multiset of prototypes (individual words or lemmas), with the number of replications of the same type of object (multiplicity) not limited in the population; the set of prototypes is denoted by V and the ordinal by $|V|$; we assume V is the vocabulary of a language, or at least a large part of it;

- (ii). the ordinal of the set of prototypes is much smaller than the number N of elements in the population (text); $|V| \ll N$.

In the context of natural language processing (NLP), the set of prototypes is the set of words or lemmas of the vocabulary of a language, the objects are words (or lemmas), denoted by w , and the population typically represents a large corpus (a text, in general terms). The situation described corresponds to the case of “Big Data” (vocabularies of tens or hundreds of thousands of words), but the content of this article may also be applied to a small population, such as a novel; in the last case, the vocabulary is specific to the author or is specific to that novel. Even in these particular cases, the vocabulary is much smaller than the population; therefore, the discussion in this article still applies.

Because of (ii), at least some words are repeated a number of times. We denote by $n_w \in \mathbf{N}$ the multiplicity of the prototype (word) w . The numbers of occurrences of some words may be equal $n_{w_1} = n_{w_2}$. The fact that $n_w \in \mathbf{N}$ is a departure from the “pure” Zipf’s law; some of its consequences are described below.

The property of $n_w \in \mathbf{N}$ combined with $p(w) = \frac{A}{r(w)}$ leads to an integer close to $\frac{A}{r(w)}$ [5,25], which may be $n_w = \left\lfloor \frac{A}{r(w)} \right\rfloor$, or $n_w = \left\lfloor \frac{A}{r(w)} \right\rfloor$, or $n_w = \left\lceil \frac{A}{r(w)} \right\rceil$, where the symbols $\lfloor \cdot \rfloor$, $\lfloor \cdot \rfloor$, and $\lceil \cdot \rceil$ represent the integer part, floor, and ceiling functions, respectively. The choice from the three variants is somewhat arbitrary, but we are sure that, on average, $n_w \geq \left\lfloor \frac{A}{r(w)} \right\rfloor$. This is valid when we do not account for any randomness. The expression “on average” means that, over a large number of texts (corpora) with the same distribution of words, the expected frequency of the word is at least $\left\lfloor \frac{A}{r(w)} \right\rfloor$. The difference between the values $\frac{A}{r(w)}$ and $\left\lfloor \frac{A}{r(w)} \right\rfloor$ can be seen as the equivalent of a noise with two values, $\frac{A}{r(w)} - \left\lfloor \frac{A}{r(w)} \right\rfloor \in [0, 1)$, $\left\lceil \frac{A}{r(w)} \right\rceil - \left\lfloor \frac{A}{r(w)} \right\rfloor \in \{0, 1\}$.

The constant A is derived from the following condition:

$$N = \sum_{r=1}^{r_M} \frac{A}{r^a} = A \cdot H_G(a, r_M), \quad (1)$$

where N is the total number of words in the text, r_M is the maximal rank, and $H_G(a, r_M)$ is the generalized harmonic number, $H_G(a, r_M) = \sum_{r=1}^{r_M} \frac{A}{r^a}$, with $H_G(a, \infty) = \zeta(a)$; ζ denotes the Riemann zeta function. Therefore:

$$n_r = \frac{N}{H_G(a, r_M)} \cdot \frac{1}{r^a}. \quad (2)$$

Because the multiplicity of any word in a text is either 0 or a natural number, the above has to be corrected, for the application discussed, as:

$$n_r = \left\lfloor \frac{N}{H_G(a, r_M)} \cdot \frac{1}{r^a} \right\rfloor. \quad (3)$$

This establishes a condition for the maximal rank, r_M :

$$r_M = \operatorname{argmax}_r \left(\left\lfloor \frac{N}{H_G(a, r_M)} \cdot \frac{1}{r^a} \right\rfloor = 1 \right). \quad (4)$$

Therefore, when N and a are known, r_M is uniquely determined, as intuitively expected. This helps us to estimate one parameter of a text, r_M , when its size N is known and the standard Zipf’s law ($a = 1$) is assumed. Considerable departures from the theoretical value of r_M empirically found signal that standard Zipf’s law is not a good model for the text. For $a = 1$:

$$r_M = \operatorname{argmax}_r \left(\left\lfloor \frac{N}{H(r_M)} \cdot \frac{1}{r} \right\rfloor = 1 \right) \quad (5)$$

where $H(r_M)$ is the harmonic number. According to Taisinger's result cited by Sanna [26], $H(q \in N) \notin N$ for $q > 1$; therefore, $\left\lfloor \frac{N}{H(r_M)} \cdot \frac{1}{r} \right\rfloor \neq \frac{N}{H(r_M)} \cdot \frac{1}{r}$ for any $r > 1$; therefore, we were compelled to use approximations for finding r_M . For large enough N values, i.e., for large texts (corpora) with at least tens of thousands of words:

$$H(r_M) \approx \gamma + \ln r_M. \quad (6)$$

Above, γ is the Euler–Mascheroni constant, $\gamma \approx 0.577 \dots$. Then:

$$\frac{N}{\gamma + \ln r_M} \approx r_M, \quad (7)$$

which solved by approximation methods, provides the estimation of r_M . For $a < 1$, one can use approximations of $\zeta(a)$ for determining an approximate value of r_M . For $a = 2$ and $a = 3$, $H_G(a, \infty)$ has a known formula; for several fractional values of $0 < a < 1$, the approximations of $H_G(a, q)$ in terms of $\ln q$ are well known, allowing us to efficiently find r_M for texts of specified sizes.

The finite vocabulary V leads to the repetitions of all words in a text when the text is large enough. The minimal text size that guarantees that all the words are repeated is given by:

$$N > \sum_{r=1}^{|V|} \frac{A}{r^a}. \quad (8)$$

The multiplicity of an element (word) is a natural number and the finite nature of V is the specific aspect of the end part of the graph of Zipf's law. This property is well known [5]; the end part of the graph was named "broom-like tail" [25] because of the shape similarity. (The results made public by HNT as [25] have been derived before the author became aware of the book [5], a reference that is much less cited than it deserves.) First, notice that there may be a rank, denoted by r_m such that $n_{r_{m-1}} < n_{r_m} = n_{r_{m+1}} < n_{r_{m+2}}$, due to the equality of the predicted number of multiplicities when forcing to natural numbers the number of elements for a specified couple of successive ranks:

$$\left\lfloor \frac{A}{r_m^a} \right\rfloor = \left\lfloor \frac{A}{r_{m+1}^a} \right\rfloor = n_{r_m}. \quad (9)$$

Denoted by r_m , the smallest rank where multiplicity (i.e., count at least 2) occurs, the above is equivalent to:

$$\left\lfloor \frac{A}{r_{m-1}^a} \right\rfloor < \left\lfloor \frac{A}{r_m^a} \right\rfloor = \left\lfloor \frac{A}{r_{m+1}^a} \right\rfloor. \quad (10)$$

The ranks are successive natural numbers, $r_{h-1} + 1 = r_h$, $r_h + 1 = r_{h+1}$; therefore:

$$\left\lfloor \frac{A}{r_{h-1}^a} \right\rfloor < \left\lfloor \frac{A}{(r_{h-1} + 1)^a} \right\rfloor = \left\lfloor \frac{A}{(r_{h-1} + 2)^a} \right\rfloor. \quad (11)$$

The equality of the values of the floor function of the two successive elements implies that:

$$\frac{A}{(r_{h-1} + 1)^a} - \frac{A}{(r_{h-1} + 2)^a} < 1 \quad (12)$$

or:

$$A < (r_{h-1} + 1)^a (r_{h-1} + 2)^a. \quad (13)$$

When $a = 1$, this condition leads to:

$$A < (r_{h-1} + 1)(r_{h-1} + 2), \quad (14)$$

which is easy to solve. For a multiplicity of order, μ , the condition is:

$$\left\lfloor \frac{A}{r_h^a} \right\rfloor < \left\lfloor \frac{A}{r_{h+1}^a} \right\rfloor = \left\lfloor \frac{A}{r_{h+2}^a} \right\rfloor = \dots = \left\lfloor \frac{A}{r_{h+\mu}^a} \right\rfloor < \left\lfloor \frac{A}{r_{h+\mu+1}^a} \right\rfloor. \quad (15)$$

Consequently,

$$\left\lfloor \frac{A}{r_{h+1}^a} \right\rfloor - \left\lfloor \frac{A}{r_{h+\mu}^a} \right\rfloor < 1, \quad (16)$$

and, for $a = 1$, a simple condition easy to solve either in terms of μ or r_h is:

$$A\mu < (r_{h-1} + 1)(r_{h-1} + \mu). \quad (17)$$

As can be expected, the multiplicity increases with the rank (see Figure 1). The above discussion allows us to clarify how the tail of the Zipf's law graph splits and shows a "broom-like" shape; in addition, the discussion shows that this specific shape of the tail is not due to some statistical noise, as suggested by [27] ("the statistical noise, ... is always present, but is most obvious in the tails of discrete power law distributions").

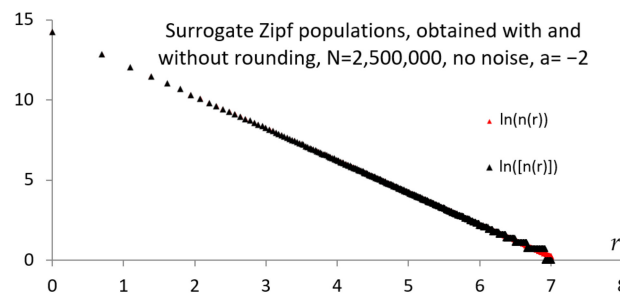


Figure 1. Standard (no noise in the exponent) Zipf's law, for the case of no rounding and with rounding of the multiplicity. Notice the "broom-end" effect of the floor function. (Surrogate data, exponent -2 . Variant of Figure 1 in [25]).

Determining the multiplicities for the highest ranks is significant in linguistics, especially in stylometry, where the number of hapaxes in a text are typically determined and may help differentiate authors when texts of similar sizes are compared.

It may be interesting to determine for a text of N words the total number of ranks r , with equal counts (multiplicities). Denote by $v(\mu)$ the number of ranks with the same multiplicity (count) μ ; then:

$$\zeta(N) = \sum_{r=1}^{r_M} \mu(r). \quad (18)$$

Next, determine the maximal and minimal ranks that have the same count μ :

$$h_2(\mu) = \arg \max_h \left(\left\lfloor \frac{A}{r_h^a} \right\rfloor = \mu \right), \quad h_1(\mu) = \arg \min_h \left(\left\lfloor \frac{A}{r_h^a} \right\rfloor = \mu \right). \quad (19)$$

Then, $v(\mu) = h_2(\mu) - h_1(\mu) + 1$. Notice that $v(\mu)$, $h_2(\mu)$, $h_1(\mu)$ depend on N and on a ; thus, a more appropriate notation is $v(N, a, \mu)$. We were interested in the distribution of $v(N, a, \mu)$. Based on the simulations (see Figure 2a,b), we conjectured that the distribution was a power law with the power depending on N, a . Simulations such as those presented in Figure 2 may help predict the number of expected hapaxes and dis- and tris-legomena in a text and determine if, from this point of view, the text style corresponds to power law distribution of the words.

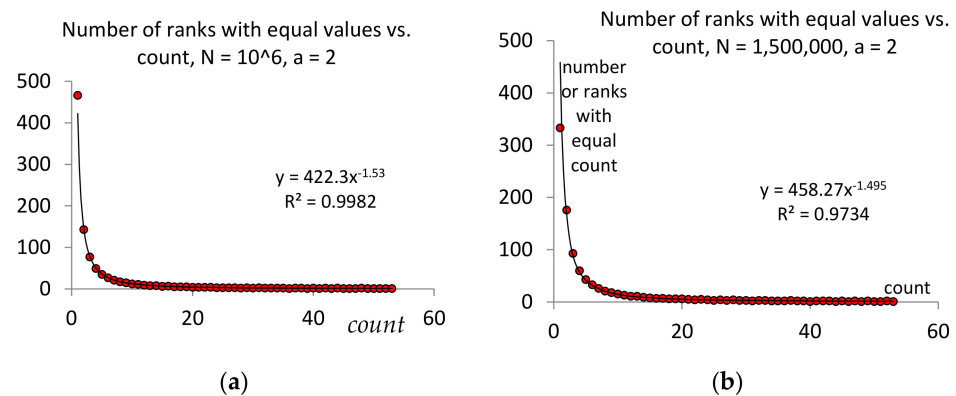


Figure 2. Examples of variations in $\nu(N, a, \mu)$ for $a = 2$ and $N = 10^6$ (a), $N = 1.5 \times 10^6$ (b).

4. Power Laws with Variable or Noisy Exponents

Empirical data often produce graphs that differ from a typical Zipf one. As previously mentioned, several variants of the power law were proposed for explaining the data, including Mandelbrot and lognormal distributions. However, a good formal approximation of the empirical data is not enough if one cannot find a theoretical (generative) model that explains the distribution production by the application at hand [28,29]. In addition, one should validate the matching of the data to the specific distribution used to model them, but distribution tests are seldom reported. Many distributions can produce graphs that apparently resemble Zipf, Mandelbrot, or lognormal laws, where the second part of the graph, for large ranks, seems to be a line with a slope unlike that of the part of the graph. When the second part of the graph is slightly different from a line, one may need to find a better matching distribution. For example, a family of distributions of probabilities $p(r)$, with variable power, where the power depends on the rank,

$$p(r) = K/r^{a+br} \quad (20)$$

has graphs, such as those in Figure 3. These distributions are candidates for modeling data when the second part of the graph is not a line.

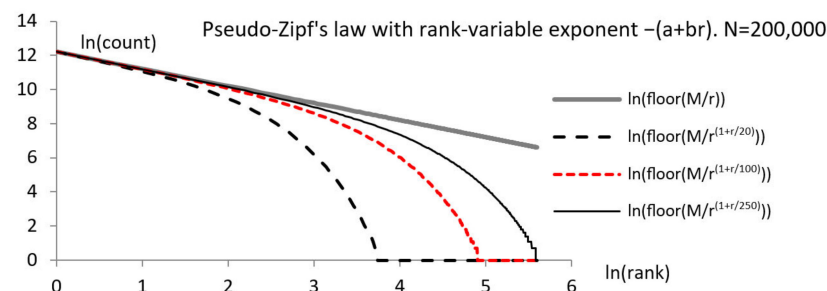


Figure 3. A family of power distributions with the power depending on the rank. (Surrogate data. Variant of Figure 4 in [25]).

While the noise is not required for explaining the staircase aspect of the end of the graph for Zipf's law, as already mentioned, noise can affect the data, making it difficult to find the power, a of the power distribution. The typical solution is to consider an additive noise, $n(r) = \frac{A}{r^a} + \varepsilon$, and to determine the best line matching the data. However, this procedure does not explain heteroscedasticity as, sometimes, it is observed and manifested by an increase in the variance with r .

In cases of corpora that include different oeuvres of an author, one frequently finds that a specific word changes the rank, depending on the texts in the corpus. This can be interpreted as an uncertain rank, $+\varepsilon$, or as an uncertain power. The last interpretation leads to the model $p(r) = A/r^{a+\varepsilon}$, which exhibits heteroscedasticity (see Figure 4).

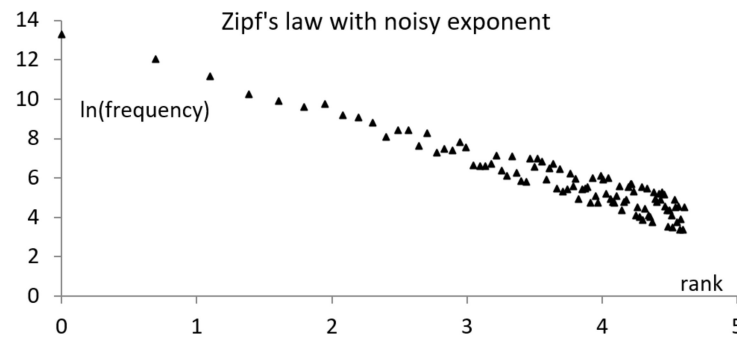


Figure 4. A rank distribution where the power is affected by noise, $a = 2 + \varepsilon$, explains heteroscedasticity.

In the next two sections, we applied the properties mentioned above determined for empirical cases.

5. Mixtures of Populations with Power Laws and Etymological Populations

We discussed the case of populations exhibiting a low power distribution and that consisted of subpopulations that also followed a power law. The topic is important in linguistics when one studies a text where the words fall into several categories, for example, neologisms, common words, and archaisms, or words of various etymologies. Another case is that of collections of texts where some of the texts carry negative sentiments, positive sentiments, or are sentiment neutral. We were not aware of any similar applicative treatment, but the main ideas originated from a similar analysis [30] applied to multinational companies and their subsidiaries [31].

Assume a population P , with the rank-frequency law according to Zipf, $p(r) = A/r$, and a partition into two sub-populations (sub-multisets), $P_1, P_2, P_1 \cap P_2 = \emptyset, P_1 \cup P_2 = P$. Furthermore, assuming that P_1 also follows Zipf's law, $p(r') = A_1/r'$, where r' represents the ranks in P_1 , we are interested in the consequences of this situation where a multiset and a subset of it obey Zipf's law. An empirical example is a large collection of tweets and the subset of tweets that feature negative sentiments and follow Zipf's distribution, as shown by [32].

Assume that $r'_1 = r_h$ is the lowest rank of an element from P_1 and $r'_2 = r_j$; then, for some ρ and A_1 values, $\frac{A}{r_h} = \frac{A_1}{1}, \frac{A}{r_j} = \frac{A_1}{2}, \frac{A}{r_k} = \frac{A_1}{3}$, etc. Then, $r_h = \frac{A}{A_1} = r'_1, r_j = \frac{2A}{A_1} = r'_2, r_k = \frac{3A}{A_1}$, etc., showing that, given that P has a Zipf distribution, P_1 also has a Zipf's distribution, only by selecting the elements of P_1 with ranks as natural multiples of $\frac{A}{A_1}$. When $\frac{A}{A_1}$ is not a natural number, it has to be rounded up, for the rank remains an integer; but, we skipped this detail. Assume that $A \gg A_1 = \frac{A}{r_h}$, that is, the subpopulation is a small part of the initial population, and denoting $C = \frac{A}{A_1}$, C satisfies the conditions $C \gg 1, \lfloor C \rfloor \approx C$. Using this approximation, the elements of the subpopulation have the frequencies of $p(r') = \frac{A}{C} \frac{1}{r'}$.

We are still left with an unknown number of elements in the subpopulation, which is:

$$N_1 = \frac{A}{r_h} + \frac{A}{r_{h+C}} + \frac{A}{r_{h+2C}} + \dots + \frac{A}{r_{h+KC}} \quad (21)$$

where $h + KC \leq r_M, K = \left\lfloor \frac{r_M - h}{C} \right\rfloor$. The subpopulation is small compared with the population when $N_1 \ll N$, which requires that r_h , the index where the subpopulation starts, is large.

We name "perfect population" a multiset T , which is a finite union of finite, disjointed multisets, each of them including a single type of object, with the number of elements in the subsets T_h satisfying the condition $|T_h| = n_h = n_1/h$ for all $h = 1, 2, \dots, h_M$. As the union

includes a finite number of finite multisets, the perfect population $T = \bigcup_{h=1}^{h_M} T_h$ is finite. Because $n_h \in \mathbf{N}$, n_1 should be divisible by all integers h from 1 to h_M ; therefore, necessarily,

$$n_1 = k \cdot \text{LCM}(2, 3, \dots, h_M) \quad (22)$$

where LCM denotes the least common multiple and k is an integer. In particular,

$$n_1 = k \cdot 1 \cdot 2 \cdot 3 \cdot \dots \cdot h_M = k \prod_{j=1}^{h_M} j, \quad n_h = k \prod_{\substack{j=1 \\ j \neq h}}^{h_M} j. \quad (23)$$

The multiset T is a model of an ideal Zipf population restricted to a finite set of ranks, with a finite number of elements for each rank.

Consider a subset T_q of a perfect population, T as above, T_q is composed of all the subsets: T_{qs} , $q, s \in \mathbf{N}$, $1 < q < h_M$, $s = 1, 2, \dots, \lfloor h_M/q \rfloor$, $qs \leq h_M$. Denote $n'_1 = n_q = n_1/q$, $n'_2 = n_{2q} = n_1/2q$, etc. Then, $n'_2 = n_{2q} = \frac{n_1}{2q} = n'_1/2$, $n'_3 = n_{3q} = n_1/3q = n'_1/3$, etc. Therefore, T_q is also a perfect population. This shows that a Zipf's distribution restricted to a finite set of objects does include at least one subpopulation that has a finite Zipf's distribution. In fact, we can choose any q to build such a Zipf subset. The next question is if one can build two perfect subpopulations of a perfect population that are disjointed. The answer is affirmative: one needs to choose two prime numbers, $1 < q_1, q_2 < r_M$, such that $q_1 \cdot q_2 > r_M$. The last condition is required by the condition that no common multiple index is present; if true, the two subpopulations are disjointed. The condition $q_1 \cdot q_2 > r_M$ indicates that the two prime numbers should be large enough. There are several unanswered questions, including the following: Considering a perfect population *maximal* r_M , and a minimal number of ranks in every subpopulation, what is the maximal number of disjointed perfect subpopulations? It is clear that we can construct several disjointed subpopulations—possibly with a single rank—at least equal to the number of prime numbers between $\sqrt{r_M}$ and r_M .

The implications of the above presentation are discussed in the next section.

6. Applications, Discussion, and Conclusions

In relation with social networks, particularly Twitter (currently X), Zipf's law has been used to find suspicious patterns of communication [33] and for sentiment analysis [32]. Suggestions that words follow Zipf's law, when words are used on social networks for subjects related to various events, have been made in studies such as [34–37]. We provide an example in Figure 5, which shows Zipf's law in tweets with negative sentiment, where the tweets refer to the energy crisis in Europe in 2021–2022. Full explanations are given in [38]. The example in Figure 5 parallels the results obtained by [32].

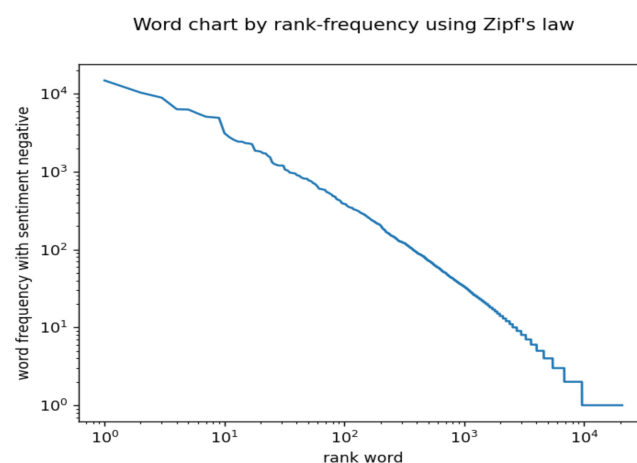


Figure 5. Zipf's law in tweets related to an energy crisis. See the text for explanations.

As large collections of messages on social media exhibit Zipf's law, the fact that subcollections of messages related to specific events also exhibit an approximate power law requires an explanation. In the same direction, the fact that subcollections of messages that are selected in the criterion of exhibited sentiment may be surprising. The above discussion explains why one can find subpopulations that have a distribution close to Zipf's one. However, the population, $P \setminus P_1$ is no longer a Zipf's population; yet, when P_1 is small enough, the $P \setminus P_1$ distribution can reasonably be approximated to Zipf's law, as in [36] and [32]. Yet, for the data collection presented in [38], the words related to positive or neutral sentiment have a much greater deviation from a straight line than the negative words.

An interesting partition of a text is obtained when considering the subsets of words with different etymologies. We have analyzed (in [39] and in the work reported here) several texts in Romanian, also applying the etymological criterion. The etymological analysis method is described in Appendix A. Similar to English, which has words from Latin, Old German, Old Norse, French, and others, the Romanian language includes words from Latin, French, Old Slavic, Old Greek, Turkish, Hungarian, etc. There is no linguistically known reason to us why the distribution of words with a specific etymology should obey Zipf's law. Yet, we found that words with certain etymologies followed Zipf's law closely enough. Figure 6 presents the segment of the graph for larger ranks for words from Old Slavic; the power of the best line fit was -1.071 , and the match to the line was almost perfect ($R^2 = 0.99$). As is well known, the first part of the graph has a typically less good fit to a line; this occurs also for the data partly presented in Figure 6. In conclusion, this study clarified some of the properties of Zipf's law and presented several new aspects of Zipf's law that pertained to the number theory and set theory.

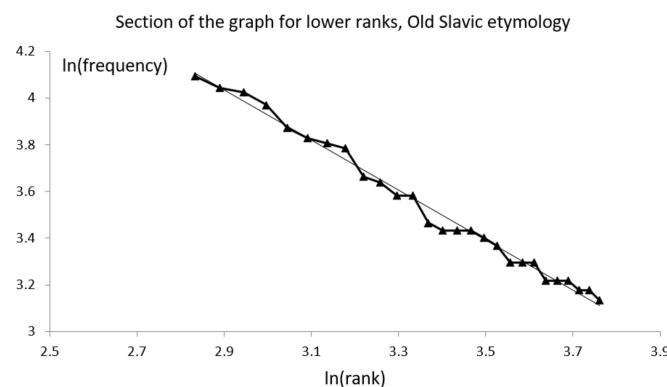


Figure 6. Zipf's law for the words with an Old Slavic etymology in the analyzed oeuvres. (See the text for details).

While the explanation of the possibility of Zipf's law in subsets of a multiset exhibiting Zipf's law was proved and discussed in Section 4, the language mechanisms explaining this surprising behavior remain to be provided by linguists. Extensions to Mandelbrot–Zipf's and Benford's laws [40] could be a natural way to further this study. We can conclude that the analysis of Zipf's and similar power laws are still a productive field of research, with interesting consequences to be unveiled. Although no explanation exists for why certain phenomena obey Zipf's law, notable attempts have been made, for example, by Gabaix [1]; yet, efforts should continue in the future, because few extant explanations are satisfactory, if at all.

Author Contributions: Conceptualization, H.-N.T.; methodology, H.-N.T., S.C.B. and D.G.; software, S.C.B., M.P. and S.-I.B.; validation, V.A., S.-I.B. and D.G.; formal analysis, H.-N.T.; investigation and data curation, S.C.B. and M.P.; writing—original draft preparation, H.-N.T.; writing—review and editing, H.-N.T., S.C.B., M.P. and D.G.; software application and etymology analysis, S.C.B.; final review of the manuscript: H.-N.T., S.C.B., M.P., S.-I.B., D.G. and V.A. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The texts analyzed for the section on etymology are available online. Part of the results have been made public in [39]. The results of the parsing and etymology extraction are available on demand from the first corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

In the automatic analysis of the etymologies of lemmas in the analyzed texts (self-biographies), we used a program written by us (S.C.B.) and used an online etymological dictionary (Dexonline) [41] and the related scraper Dexonline-Scraper package [42]. The texts analyzed are [43–46]. The software application has the structure described below.

We read the input file, found all the unique words/lemmas, and then obtained the HTML pages for each one from the etymology dictionary Dexonline [41]. All those pages were saved in the “input/pages” path. The download was performed once for an entry file. Also, if we ran this multiple times, or if we analyzed other literary works, only the pages that had not yet been downloaded would have been downloaded.

The main program read the previously downloaded data, and then extracted the language of origin. It went through the following steps:

The Dexonline-Scraper package [42] already provided us with the lemma, so we verified if that matched with our word/lemma. We manually verified and solved mismatches.

Step 3: find the entries with a specified part of speech (POS) role. After only having entries that matched the lemma, we filtered those that also matched the POS. This was performed with the Dexonline-Scraper package [42], which gave us the POS inside a “type” attribute. There was mapping between POS types and the results from Dexonline [41], which helped us with the matching.

The entries that matched the word were stored in a result structure and we continue processing the rest of the words.

The results were in the form:

cap,	Nc,	178,	1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
veni,	Vm,	163,	1,0
general,	A,	27,	1,0
mirat,	A,	4,	0,0
general,	Nc,	4,	0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0

where the interpretation is: the first position is the lemma; the second position is the part of speech (POS) as given by the parser; the third position is the count (number of occurrences in the text); and all other positions are either 0, meaning that the etymology is not the one provided in the list of etymologies, or 1, meaning a valid etymology. Multiple etymologies are possible when a word is composed of parts with different etymologies, or when the etymologic dictionary provides multiple etymologies. The list of etymologies is the following: Latin, unknown, Italian, German, French, English, Albanian, Old Slavic, regressive derivative, Turkish, Bulgarian, Hungarian, Russian, Serbian or Croatian, neo-Greek, Ukrainian, locution, Greek (modern), Old Greek, Polish, onomatopoeia, Romani (whatever dialect), archaism, Czech, German Sachs dialect in Romania, Romanian, and

Spanish (in this order). Only the most frequent etymologies were investigated and only the first three of them had distributions that were similar to Zipf's distribution.

References

- Gabaix, X. Zipf's Law for Cities: An Explanation. *Q. J. Econ.* **1999**, *114*, 739–767. Available online: <http://www.jstor.org/stable/2586883> (accessed on 20 January 2024). [CrossRef]
- Kondo, I.O.; Lewis, L.T.; Stella, A. Heavy tailed but not Zipf: Firm and establishment size in the United States. *J. Appl. Econom.* **2023**, *38*, 767–785. [CrossRef]
- Fazio, G.; Modica, M. Pareto or log-normal? Best fit and truncation in the distribution of all cities. *J. Reg. Sci.* **2015**, *55*, 736–756. [CrossRef]
- Gabaix, X. Power laws in economics: An introduction. *J. Econ. Perspect.* **2016**, *30*, 185–206. [CrossRef]
- Baayen, R.H. *Word Frequency Distributions, Chapter 1*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2001.
- Corral, Á.; Boleda, G.; Ferrer-i-Cancho, R. Zipf's Law for Word Frequencies: Word Forms versus Lemmas in Long Texts. *PLoS ONE* **2015**, *10*, e0129031. [CrossRef] [PubMed]
- Ferrer i Cancho, R. The variation of Zipf's law in human language. *Eur. Phys. J. B-Condens. Matter Complex Syst.* **2005**, *44*, 249–257. [CrossRef]
- Lin, J.H.; Lee, W.C. Complementary Log Regression for Sufficient-Cause Modeling of Epidemiologic Data. *Sci. Rep.* **2016**, *6*, 39023. [CrossRef] [PubMed]
- Furusawa, C.; Kaneko, K. Zipf's law in gene expression. *Phys. Rev. Lett.* **2003**, *90*, 088102. [CrossRef]
- Zanette, D.H. Zipf's law and the creation of musical context. *Music. Sci.* **2006**, *10*, 3–18. [CrossRef]
- Manaris, B.; Purewal, T.; McCormick, C. Progress towards recognizing and classifying beautiful music with computers—MIDI-encoded music and the Zipf-Mandelbrot law. In Proceedings of the IEEE SoutheastCon 2002 (Cat. No.02CH37283), Columbia, SC, USA, 5–7 April 2002; pp. 52–57. [CrossRef]
- Sharma, S.; Pendharkar, P.C. On the analysis of power law distribution in software component sizes. *J. Softw. Evol. Proc.* **2022**, *34*, e2417. [CrossRef]
- Wang, D.; Cheng, H.; Wang, P.; Huang, X.; Jian, G. Zipf's law in passwords. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 2776–2791. [CrossRef]
- Corominas-Murtra, B.; Fortuny, J.; Solé, R.V. Emergence of Zipf's law in the evolution of communication. *Phys. Rev. E* **2011**, *83*, 036115. [CrossRef] [PubMed]
- Dellandrea, E.; Makris, P.; Vincent, N.; Boiron, M. A medical acoustic signal analysis method based on Zipf law. In Proceedings of the 14th International Conference on Digital Signal Processing, DSP 2002 (Cat. No.02TH8628), Santorini, Greece, 1–3 July 2002; Volume 2, pp. 615–618. [CrossRef]
- Vincent, N.; Makris, P.; Brodier, J. Compressed image quality and Zipf law. In Proceedings of the WCC 2000—ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000, Beijing, China, 21–25 August 2000; Volume 2, pp. 1077–1084. [CrossRef]
- Adamic, L.A.; Huberman, B.A. Zipf's law and the Internet. *Glottometrics* **2002**, *3*, 143–150.
- Fujiwara, Y. Zipf Law in Firms Bankruptcy. *Phys. A Stat. Mech. Its Appl.* **2004**, *337*, 219–230. [CrossRef]
- Fujiwara, Y.; Di Guilmi, C.; Aoyama, H.; Gallegati, M.; Souma, W. Do Pareto-Zipf and Gibrat laws hold true? An analysis with European Firms. *Phys. A Stat. Mech. Its Appl.* **2004**, *335*, 197–216. [CrossRef]
- Jiang, B.; Jia, T. Zipf's law for all the natural cities in the United States: A geospatial perspective. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1269–1281. [CrossRef]
- Teodorescu, M.H.M. Machine Learning Methods for Strategy Research. Report number 18-011. In *Harvard Business School Research Paper Series*; Harvard Business School: Boston, MA, USA, 2017. [CrossRef]
- O'Neale, D.R.J.; Hendy, S.C. Power Law Distributions of Patents as Indicators of Innovation. *PLoS ONE* **2012**, *7*, e49501. [CrossRef]
- Blackwell, C.; Pan, B.; Li, X.; Smith, W. *Power Laws in Tourist Flows*; Travel and Tourism Research Association: Advancing Tourism Research Globally: Whitehall, MI, USA, 2011; p. 63. Available online: <https://scholarworks.umass.edu/ttra/2011/Oral/63> (accessed on 15 October 2023).
- Torre, I.G.; Luque, B.; Lacasa, L.; Kello, C.T.; Hernández-Fernández, A. On the physical origin of linguistic laws and lognormality in speech. *R. Soc. Open Sci.* **2019**, *6*, 191023. [CrossRef]
- Teodorescu, H.-N. Big Data and Large Numbers: Interpreting Zipf's Law. *arXiv* **2023**, arXiv:2305.02687. [CrossRef]
- Sanna, C. On the p-adic Valuation of Harmonic Numbers. *J. Number Theory* **2016**, *166*, 41–46. [CrossRef]
- Milojević, S. Power law distributions in information science: Making the case for logarithmic binning. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 2417–2425. [CrossRef]
- Mitzenmacher, M. New Directions for Power Law Research. Radcliffe.ppt. Harvard University. Available online: <https://www.eecs.harvard.edu/~michaelm/TALKS/Radcliffe.pdf> (accessed on 15 October 2023).
- Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet Math.* **2004**, *1*, 226–251. Available online: <https://dash.harvard.edu/bitstream/handle/1/24828534/tr-08-01.pdf?sequence=1> (accessed on 15 October 2023). [CrossRef]
- Teodorescu, M.H.M.; (Carroll School of Management, Boston College, Boston, MA, USA). Personal communication, 2023.

31. Teodorescu, M.H.M.; Choudhury, P.; Khanna, T. Role of context in knowledge flows: Host country versus headquarters as sources of MNC subsidiary knowledge inheritance. *Glob. Strategy J.* **2022**, *12*, 658–678. [CrossRef]
32. Zhang, L.; Dong, W.; Mu, X. Analysing the features of negative sentiment tweets. *Electron. Libr.* **2018**, *36*, 782–799. [CrossRef]
33. Sarna, G.; Bhatia, M.P. Identification of suspicious patterns in social network using Zipf's law. In Proceedings of the International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 12–13 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 957–962.
34. Thurner, S.; Szell, M.; Sinatra, R. Emergence of good conduct, scaling and Zipf laws in human behavioral sequences in an online world. *PLoS ONE* **2012**, *7*, e29796. [CrossRef]
35. Teodorescu, H.-N.L.; Bolea, S.C. On the algorithmic role of synonyms and keywords in analytics for catastrophic events. In Proceedings of the 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI 2016), Ploiesti, Romania, 30 June–2 July 2016; pp. 1–6. [CrossRef]
36. Pirnau, M. Analysis of data volumes circulating in SNs after the occurrence of an earthquake. *ROMJIST* **2017**, *20*, 286–298.
37. Teodorescu, H.N.; Pirnau, M. Twitter's Mirroring of the 2022 Energy Crisis: What It Teaches Decision-Makers—A Preliminary Study. *Rom. J. Inf. Sci. Technol.* **2023**, *26*, 312–322. [CrossRef]
38. Pirnau, M.; Priescu, I.; Joita, D.; Priescu, C.M. Analysis of the Energy Crisis in the Content of Users' Posts on Twitter. In Proceedings of the 17th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, Romania, 9–10 June 2023; pp. 1–4. [CrossRef]
39. Teodorescu, H.-N.; Bolea, C.S. A Comparative Lexical Analysis of Three Romanian Works—The Etymological Metalepsis Role and Etymological Indices. *Rom. J. Inf. Sci. Technol.* **2022**, *25*, 275–289.
40. Beretta, F.; Dimino, J.; Fang, W.; Martinez, T.C.; Miller, S.J.; Stoll, D. On Benford's Law and the Coefficients of the Riemann Mapping Function for the Exterior of the Mandelbrot Set. *Fractal Fract.* **2022**, *6*, 534. [CrossRef]
41. Dexonline. Available online: <https://dexonline.ro/> (accessed on 15 October 2023).
42. Dexonline-Scraper. MIT License. Available online: <https://github.com/vxern/dexonline-scraper> (accessed on 20 January 2024).
43. Teodoreanu, I.; La Medeleni, R. *Volumul I, Hotarul Nestatornic*; Editura "Cartea Românească": Bucharest, Romania, 1925.
44. Teodoreanu, I.; La Medeleni, R. *Volumul III, Între Vânturi*; Editura "Cartea Românească": Bucharest, Romania, 1927.
45. Averescu, A. *Notițe Zilnice din Războiul (1916–1918)*; Editura "Cultura Națională București": Bucharest, Romania, 1935.
46. Iorga, N. *Supt Trei Regi, Istorie a Unei Lupte Pentru un Ideal Moral și Național*; Ediția a II-a, București: Bucharest, Romania, 1932.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.