Long Paper

Thermodynamics of words

Klaus M. Frahm[*,1], Dima L. Shepelyansky[**,*,1]

Univ Toulouse, CNRS, Laboratoire de Physique Théorique, Toulouse, France
dima@irsamc.ups-tlse.fr

We introduce the words thermalization hypothesis according to which in a literary work there are effective interactions between words that drive them to thermalization with their word frequencies distribution being the same as for the phenomenon of Rayleigh-Jeans (RJ) thermalization and condensation. In this concept the different words play a role of classical particles and their word frequencies, or related tokens, are analogous to certain filling factors over energy state levels. The thermalization process is characterized by the conservation of total system energy, or total number of words/tokens, and total probability norm, or total number of different words. We apply this approach to analysis of 16 literary works, including Bible, complete works of Shakespeare, and the American National Corpus. We show that for these cases the effective system energy is very low and thus there is the RJ condensation phenomenon with a big fraction of total probability concentrated at the lowest energy states corresponding to a big fraction of rare words appearing only once or twice in the whole text. Such Rayleigh-Jeans condensation is observed for light propagation in multimode optical fibers and other physical systems. We also trace analogy between thermalization of words and the recently proposed wealth thermalization for households in world countries. The economy concepts of Lorenz and Pareto curves, and also Zipf curve, are applied to the word frequency distributions of the selected 16 cases and it is shown that the thermalization theory gives a strikingly good description of real data.

1. Introduction

The statistical properties of distribution of word frequencies in literary works are actively studied by scientific community and are reported in various review articles and books (see e.g. (Baayen 2001; Popescu 2009; Piantadosi 2014) and Refs. therein). A significant part of research in devoted to the analysis of properties and origins of the Zipf frequency law in natural language (Zipf 1936, 1949). According to this law the word frequencies decay algebraically with an exponent $\alpha \approx 1$ at high rank values of words ordered by frequency. The detailed investigations of the Zipf law are described e.g. in (Newman 2005; Brysbaert and New 2009; Moreno-Sánchez, Font-Clos, and Corral 2016; Linders and Louwerse 2022) but still its deep origins remain under debate (Piantadosi 2014).

∗ Equal contribution
∗∗ Corresponding author

Action editor: {action editor name}. Submission received: DD Month YYYY; revised version received: DD Month YYYY; accepted for publication: DD Month YYYY.

© 2025 Association for Computational Linguistics

To understand the properties of the whole word frequency distribution in literary works we introduce the WOrds Thermalization Hypothesis (WOTH) according to which the semantic laws, meanings and literary work concepts creates links and interactions between words that drive the word frequency distribution to thermalization of words. These links and interactions between words describe the classical objects and thus the physical laws of classical mechanics (Landau and Lifshitz 1976; Zakharov, L'vov, and Falkovich 1992) conduct the whole system to a stationary thermal word frequencies distribution. According to the laws of physics this distribution is given by the Rayleigh-Jeans (RJ) thermalization and condensation phenomenon well studied in various physical systems. In physics the RJ distribution describes usually the classical particles or objects where the thermalization appears due to nonlinear interactions between particles or due to a system contact with an external thermal bath (see e.g. (Frahm and Shepelyansky 2023; Ermann, Chepelianskii, and Shepelyansky 2026)). It is also assumed that the system evolution conserves two quantities (integrals of motion) being the system energy and the probability norm (or number of particles).

In physical systems the RJ thermalization and condensation were found in numerical simulations and experiments with light propagation in multimode optical fibers (Connaughton et al. 2005; Krupa et al. 2016; Baudin et al. 2020; Ermann, Chepelianskii, and Shepelyansky 2026). The surprising feature of the RJ thermal distribution is that at low total system energy a large probability fraction (or number of particles) is concentrated at system low energy states and only a very small fraction is located at high energy states. In fact the RJ condensation belongs to a more generic phenomenon of statistical mechanics known as the constraint-driven condensation (Evans et al. 2010; Szavits-Nossan, Evans, and Majumdar 2014; Filiasi et al. 2014) that describes such diverse physical systems as jamming in traffic, coalescence in granular media, gelation in networks (Szavits-Nossan, Evans, and Majumdar 2014) and financial data analysis (Filiasi et al. 2014).

Recently it was shown that the RJ thermalization and condensation appear not only for the above physical systems but also for the wealth inequality in the world countries, stock exchange markets and the distribution of world Gross Domestic Product (GDP) over countries (Frahm and Shepelyansky 2026b; Frahm, Ermann, and Shepelyansky 2026). It was also shown that the nonlinear interactions between system agents in social stratification models on social networks also drive the system to the RJ thermalization and condensation (Frahm and Shepelyansky 2026b,a).

The striking feature for the wealth distribution in the world countries is the strong inequality of wealth of households as demonstrated e.g. in (Piketty 2014; Chancel et al. 2022; Roach et al. 2003). Thus, according to (Chancel et al. 2022) for the whole world 50% of the population owns only 2% of total wealth, while the top 10% owns 75% of total wealth, of which 38% is held by the top 1% alone. The emergence of the phase of very poor households and the oligarchic phase of very rich ones is naturally explained in the frame of the RJ thermalization and condensation of wealth. In the frame of the RJ thermal concept a wealth of a household is analogous to a state energy level of a physical system.

In this work we expend the analogy between system energy levels and wealth of households to the word frequency in literary works. All N different words of a given literary work are ordered by their frequency rank K increasing from most rare words to most frequent ones as it is the case of wealth of households (Piketty 2014; Chancel et al. 2022; Roach et al. 2003). In the modern Large Language Models (LLM) for Artificial Intelligence (AI) English words are related to tokens with on average $4/3$ token for a word (OpenAI 2026). The tokens have their price usually expressed in US dollars (USD)

and that a word frequency determines USD price of a given word in a literary work. Thus a word frequency of the whole text determines a USD price for a given word being related to an amount of work and energy related to this word in the whole text. Indeed, a word frequency is proportional of a work done by a text writer when writing the text of a book. Thus the word frequency determines a number of tokens in a text being analogous to a state level energy in physical system or wealth of a household. This analogy represents the main element of the WOTH approach.

To verify the validity of the WOTH concept we follow the approach used in (Frahm and Shepelyansky 2026b; Frahm, Ermann, and Shepelyansky 2026) for the wealth inequality in the world. In fact in economy studies there is a very useful description of wealth distribution based on the Lorenz curve (Lorenz 1905). Namely, the Lorenz curve (Lorenz 1905; Roach et al. 2003) represents the dependence of cumulated normalized wealth $0 \leq w \leq 1$ on the cumulated normalized fraction of population or households $0 \leq h \leq 1$. The perfect equipartition of wealth corresponds to the diagonal $w = h$ and the doubled area between diagonal and the Lorenz curve $w(h)$ gives the Gini coefficient $0 \leq G \leq 1$ (Gini 1914; Roach et al. 2003).

The values of Gini coefficients G for world countries in 2021 are presented in (Wikipedia and contributors 2026a) being in the range $0.59 < G < 0.90$ while for the whole world $G = 0.889$. High G values being close to unity correspond to a high inequality in a country with a large fraction of poor population and a tiny oligarchic population fraction that owns a large fraction of the total wealth (Piketty 2014; Chancel et al. 2022). The distribution of high revenues is usually described by the Pareto distribution curve (Pareto 1897; Yakovenko and Rosser 2009). From the above analogy the rare words corresponds to poor households and frequent words to rich households. Using this analogy between wealth of household and word frequency (or number of tokens) of a given word in a literary work we construct the Lorenz and Pareto curves for the word frequency for 10 literary works in English, including Bible and complete works of Shakespeare, and 5 works in French, German, Italian, Russian and Spanish. We also analyze the American National Corpus (ANC). For all these 16 case we show that the WOTH concept based on the RJ thermalization and condensation gives the strikingly good description of the Lorenz, Pareto and Zipf distributions (we show that the Zipf distribution is directly related to the Pareto one).

The paper is constructed as follows: data sets of literary works (books) are described in Section 2, RJ thermalization and condensation concept is introduced in Section 3, cotruction rules of Lorenz, Pareto and Zipf distributions are explained in Section 4, results are presented in Section 5 and teir discussion is given in Section 5. Appendix provides additional figures and certain analytical formula.

2. Data sets

In our study we consider 15 literary works of different authors and the ANC data set openly available at (Corpus 2026) (W0 with abbreviation ANC in Table 1)). Thus we study word frequency distributions for 16 cases numbered as W0 to W15. The texts of each literary work in English and 5 other languages (cases W0 to W15) are obtained from open sources on Internet (including such web sites as www.gutenberg.org; bdebooks.com; archive.org; royallib.com; bdebooks.com; oceanofpdf.com; standardebooks.org).

There are 10 works in English including:

W1) an Early Modern English translation of the Christian Bible commissioned in 1604 by King James VI and I and published in 1611 [cite:bible] and taken from the open source of the Project Gutenberg, with abbreviation Bible;

W2) The Complete Works of William Shakespeare by William Shakespeare taken from open source of the Project Gutenberg with abbreviation Shakespeare;

and 8 other works in English are

W3) *I, Robot* by Isaac Asimov with abbreviation Asimov;

W4) *The Complete Novels* of Jane Austen with abbreviation Austin;

W5) *David Copperfield* by Charles Dickens with abbreviation Dickens;

W6) *Moby Dick* by Herman Melville with abbreviation Melville;

W7) *1984* by George Orwell with abbreviation Orwell;

W8) *The Lord of the Rings, 1-3* by J.R.R.Tolkien with abbreviation Tolkien;

W9) *The Adventures of Tom Sawyer* by Mark Twain with abbreviation Twain;

W10) *Mrs. Dalloway* by Virginia Woolf with abbreviation Woolf;

The works in other languages are

W11) *Don Quixote* by Miguel de Cervantes in Spanish with abbreviation Cervantes;

W12) *La Divina Comedia* by Dante Alighieri in Italian with abbreviation Dante;

W13) *Faust Part I* by Johann Wolfgang von Goethe in German with abbreviation Goethe;

W14) *Les Miserables* by Victor Hugo in French with abbreviation Hugo;

W15) *Peace and War* by Lev Tolstay in Russian with abbreviation Tolstoy.

The total number of words N_w and the total number of different words N are given in Table 1 for each literary work. For each work the text in format epub or txt is analyzed with the help of Open AI Chart GPT giving the word frequency distribution with frequency of a given word f_m and word index m ordered by frequency. Following the analogy with the wealth inequality analysis (Frahm and Shepelyansky 2026b; Frahm, Ermann, and Shepelyansky 2026) the normalized values of m/N and f_m are considered as wealth $w_m = f_m$ of a household $h_m = m/N$ and the Lorenz, Pareto and Zipf curves are constructed from these values as described below in Section 4. All text of works W1 to W15 and their word frequencies are available on the web site relayed to this work at [cite:ourwebpage]. The ANC data set is available at (Corpus 2026).

3. RJ thermodynamics and condensation

Rayleigh-Jeans (RJ) thermalization appears in certain nonlinear oscillator systems with given oscillator energies or frequencies E_m (with $E_0 < E_1 < \ldots < E_{N-1}$). For example, in Refs. (Frahm and Shepelyansky 2026b, 2023; Ermann, Chepelianskii, and Shepelyansky 2026; Frahm, Ermann, and Shepelyansky 2026), a nonlinear oscillator system with complex amplitudes $C_m(t)$ of the form $i\frac{\partial C_m(t)}{\partial t} = E_m C_m(t) + \beta F_m(\{C_l(t)\})$ was considered where $F_m(\{C_l(t)\})$ is some complicated nonlinear function of all the amplitudes $C_l(t)$ obtained from some special potential (see, e.g., (Frahm and Shepelyansky 2026b) for details) and β is a parameter for the strength of the nonlinear perturbation. The particular form of $F_m(\{C_l(t)\})$ used in the above references ensured that in this nonlinear oscillator system there are two exact integrals of motion which are system energy $\mathcal{H} = \sum_m E_m |C_m(t)|^2 + \beta V(\{C_l(t)\})$ and probability norm $\eta = \sum_m |C_m(t)|^2$. Here $V(\{C_l(t)\})$ is the potential associated with the nonlinear term $F_m(\{C_l(t)\})$. Typically, once the system is thermalized (at long times) the contribution of the potential is small, $V \sim \beta/N$, and we have $\mathcal{H} \approx \sum_m E_m |C_m(t)|^2$.

This type of system corresponds to a special type of microcanonical ensemble with fixed energy $\mathcal{H} = E$ and the additional norm constraint that η is conserved (typically the value $\eta = 1$ is chosen). According to standard techniques of statistical physics (Landau and Lifshitz 1976), such a system can be replaced (for $N \gg 1$) by a grand canonical ensemble. In this approach, the amplitudes $C_m(t)$ have (in equilibrium) a Gaussian distribution with the RJ average

$$\rho_m = \langle |C_m(t)|^2 \rangle = \frac{T}{E_m - \mu} \tag{1}$$

where the temperature T and the chemical potential μ are determined to satisfy both constraints in average:

$$E = \langle \mathcal{H} \rangle = \sum_m E_m \rho_m \quad , \quad 1 = \langle \eta \rangle = \sum_m \rho_m \; . \tag{2}$$

Here ρ_m is an average probability norm at a state m with energy E_m. These equations allow a unique physical solution of μ (and T) where $\rho_m > 0$ for all modes and which can be computed by standard numerical techniques. Depending on $E < E_c = \sum_m E_m/N$ (or $E > E_c$), we have $\mu < E_0$ and $T > 0$ ($\mu > E_{N-1}$ and $T < 0$ respectively; see (Frahm and Shepelyansky 2026b) for details). In this work, we discuss only the physical case of positive temperature $T > 0$ and, actually, in nearly all cases related to real word distribution data, we are in the low temperature regime, corresponding to $E_0 < E \ll E_c$ with possible RJ condensation where a big part of probability will be concentrated on the lowest modes m. In the next section, the construction of the associated Lorenz (and other related) curves for such an RJ system is defined.

The striking feature of the RJ distribution (1) at low energies $E \ll E_c$ is a condensation of a high fraction of total probability η is the states in a vicinity of energy ground state. This phenomenon is found at the numerical simulations and experiments with multimode optical fibers as discussed in (Connaughton et al. 2005; Krupa et al. 2016; Baudin et al. 2020; Ermann, Chepelianskii, and Shepelyansky 2026).

4. Construction of Lorenz, Pareto and Zipf curves

The concept of Lorenz curves (Wikipedia and contributors 2026b) is widely used in economy science where it often shows for a given population the cumulated fraction w of total wealth as a function of the cumulated fraction of households h (the households or population are assumed to be ordered by the increasing wealth). The line $w = h$ corresponds to the case of perfect equipartion. It is useful to define the Gini coefficient $G \in [0, 1]$, being the doubled area between the equipartition line and the Lorenz curve. Thus the Gini coefficient characterizes the global inequality of the wealth distribution, i.e. large of values of G close to unity correspond to a strong inequality.

In this work, we construct two types of Lorenz curves, first for real word frequency distribution data and second for suitable RJ models with certain energy spectra (see below). For the real data of a certain literary work (or book), we consider N different words used in this book (or set of books, data sets) with word frequency values f_m being the number of times each word is used in this book. We assume these frequency values being ordered with increasing values, $f_0 \leq f_1 \leq \ldots f_{N-1}$, and we attribute to each word a wealth or token value $w_m = f_m - 1$ (reduced frequency) such that for words with minimal frequency $f_m = 1$ we have a vanishing wealth $w_m = 0$ (very "poor" words).

Then the associated (real data) Lorenz curve $(h(m), w(m))$ for $N+1$ data points $m = 0, 1, \ldots, N-1$ is simply constructed as $h(m) = m/N$ and $w(m) = \sum_{k=0}^{m-1} w_m/(N\langle w_m\rangle)$ where $\langle w_m\rangle = \sum_{k=0}^{N-1} w_k/N$ is the average wealth or word frequency with $w_m = f_m - 1$. Here $h(m)$ represents the cumulative fraction of households (different words) at index m and $w(m)$ is the cumulative normalized reduced frequency (wealth) fraction such that $h(0) = w(0) = 0$ and $h(N) = w(N) = 1$. In this construction, we assume that each word has the same uniform household fraction $1/N$.

For the RJ theory, we consider an ordered model spectrum E_k of N levels with $0 = E_0 < E_1 < \ldots < E_{N-1}$ for large values of N (e.g. $N = 10^4$, see also below). Here, each level corresponds to a given agent k with associated wealth E_k. We attribute to each agent a thermalized RJ probability according to (1) for a given energy value $E \in]E_0, E_{N-1}[$ which fixes the parameters T and μ as described in the last section. Then, we define the RJ Lorenz curve as $N+1$ points $(h(m), w(m))$ (for $m = 0, 1, \ldots, N$) with $h(m) = \sum_{k=0}^{m-1} \rho_k$ and $w(m) = \sum_{k=0}^{m-1} E_k\rho_k/E$, again satisfying $h(0) = w(0) = 1$ and $h(N) = w(N) = 1$ (in principle number of agents can be larger than number of different words N). This curve is invariant with respect to a simple energy rescaling $E_m \to \alpha E_m$, $E \to \alpha E$ (and similarly for μ and T). Therefore, the relevant parameter is the rescaled energy $\varepsilon = E/B$ where $B = E_{N-1}$ is the energy bandwidth of the spectrum. In this theoretical RJ approach, it is assumed that each word is represented by a certain number of thermalized agents distributed over the energy levels E_k being in the energy band $[0, B]$.

Concerning, the specific choice of the RJ energy spectrum, we consider two models. The first one is very simple with a constant density of energy states $\nu(k) = dk/dE_k = const$ so that $0 \leq E_k = k/N < 1$ for $k = 0, 1, 2, \ldots, (N-1)$. We call this model, the RJ standard (RJS) model that was analyzed in detail in (Ermann, Chepelianskii, and Shepelyansky 2026; Frahm and Shepelyansky 2026b). For sufficiently small values of $\varepsilon \ll 1$, we have RJ condensation in this model where a significant fraction of probability is concentrated on a small number of modes at minimal k values. This description provides Lorenz curves that usually qualitatively agree with Lorenz curves from real data.

However, to obtain a more exact agreement, we introduce the RJ extended (RJE) model with $E_k = (e^{ak/N} - 1)/(e^a - 1)$ and a being an additional parameter besides ε. The limit $a \to 0$ reproduces the RJS model with $E_k = k/N$ and for $a > 0$ the density of states $\nu(E_m)$ is decreasing at high energies as $\nu(E_k) = dk/dE_k = N(e^a - 1)/[a(1 + (e^a - 1)E_k]$. The global energy scale here is chosen such that $B = E_{N-1} \approx 1$ (such that $E \approx \varepsilon$). For the wealth inequality, the decay of ν at high energies is rather natural since there are fewer rich people at high wealth/energy values (or smaller number of words with very high frequency $f_m = w_m + 1$). Such a situation is also plausible for the word distribution in a data set where very high frequency words are rarer.

In this work, we consider only results for the RJE model, since the real word distribution data requires typically quite high values of $a \approx$ 5-10 and the RJS model (at $a = 0$) only gives a qualitative agreement with real data.

To compute the Lorenz curve from the RJ theory (1), it is possible to use the above procedure for large but finite values of N, e.g. $N = 10000$. However, here, we use the continuous limit $N \to \infty$ for which it is possible to replace the discrete sums for cumulated household and wealth fractions by continuous integrals and to work out closed analytic expressions (Frahm, Ermann, and Shepelyansky 2026) for different key

quantities. For example the analytical RJE Lorenz curve (at $N \to \infty$) is given by

$$w^{(\mathrm{RJE})}(h) = \left(1 - \frac{\mu}{\varepsilon}\right) s(h) + \frac{\mu h}{\varepsilon} \quad , \quad D(a) = e^a - 1 \quad , \tag{3}$$

$$s(h) = -\frac{1}{a} \ln \left\{ 1 + \frac{\mu D(a)}{1+\mu D(a)} \left[\exp\left(\frac{a(1+\mu D(a))h}{D(a)(\varepsilon - \mu)} \right) - 1 \right] \right\} \tag{4}$$

where μ is determined by the implicit equation:

$$1 = \frac{\varepsilon - \mu}{a(D^{-1}(a) + \mu)} \ln \left(\frac{e^{-a}(1-\mu)}{-\mu} \right) . \tag{5}$$

Note that the quantity $s(h)$ in (4) corresponds to the continuous index $s = m/N$ in the limit $N \to \infty$ which has a non-trivial dependence on the household fraction h.

These analytic RJE expressions agree very well the finite N RJE curves provided N is sufficiently large (identical on graphical precision for $N \geq 10^4$). We remind, that in the RJE approach with (very) large values of N, multiple agents (many values of the spectral index k) correspond to a given individual data point for real word reduced frequency values $w_m = f_m - 1$ (one fixed value of m).

To compare the RJE model with real data Lorenz curves, the parameter ε is fixed by the condition that the RJE and real data Lorenz curves have the same Gini coefficient G and the parameter a is determined by minimizing the geometrical curve distance between both curves (under the constraint of respecting the first condition). This procedure was already used in (Frahm and Shepelyansky 2026b; Frahm, Ermann, and Shepelyansky 2026).

Another useful charcteristics, often used in economy studies (see, e.g., (Pareto 1897; Yakovenko and Rosser 2009)), is the cumulative distribution function (CDF) $C(w_m)$, which represents the fraction of households (words) having a wealth $w_{m'}$ with $w_{m'} \geq w_m$. It can be directly obtained by drawing $C(w_m) = 1 - h(m)$ versus w_m (for both cases of real data and RJ models). Usually, this distribution is described by the Pareto distribution representing an algebraic decay law for $w_m \geq w_c$ with some critical w_c (and $C(w_m) = 1$ for $w_m < w_c$). In our work, we simply use the notation *Pareto distribution (or curve)* for the CDF $C(w_m)$ as such, also for more general cases even if there is no simple power law.

In contrast to the Lorenz curve, the Pareto curve is not scale independent and in order to compare real data with possible RJE curves, it is necessary to fix the global wealth scale either by the maximal or the average value of w_m. In this work, we show $C(w_m)$ as a function of $w_m/\langle w_m \rangle$ where for the RJE model, we simply have $\langle w_m \rangle = \varepsilon$ (since $B = 1$ and $E_m = w_m$). For real data the average value $\langle w_m \rangle$ is more reliable since the maximal value of w_m for a given data set may be subject to strong statistical fluctuations (note that $\max(w_m) = 1$ for the RJE model since $B = 1$).

The analytical expression for the RJE Pareto curve at $N \to \infty$ is given by (Frahm, Ermann, and Shepelyansky 2026):

$$C_{\mathrm{RJE}}(w_m) = 1 - \frac{\varepsilon - \mu}{a(D^{-1}(a) + \mu)} \ln \left(\frac{1 - \mu^{-1} w_m}{1 + D(a) w_m} \right) . \tag{6}$$

For $w_m \gg \max(-\mu, D^{-1}(a))$, one can apply an expansion in $1/w_m$ which gives the quite simple expression (Frahm, Ermann, and Shepelyansky 2026):

$$C_{\rm RJE}(w_m) \approx \frac{\varepsilon-\mu}{a}\left(\frac{1}{w_m}-1\right) = A_{\rm RJE}\left(\frac{w_m}{\langle w_m\rangle}\right)^{-1} - A_{\rm RJE}\langle w_m\rangle \quad , \tag{7}$$

$$\text{with} \quad A_{\rm RJE} = \frac{\varepsilon-\mu}{\varepsilon a} \approx \frac{1}{a} \tag{8}$$

since $\varepsilon = \langle w_m\rangle$ (for the RJE model) and typically $|\mu| = -\mu \ll \varepsilon$ (for all cases in this work; see also below). The expression (7) is valid in the full interval $w_m \in [w_{\rm min}, 1]$ where $w_{\rm min}$ is some value with $w_{\rm min} \gg \max(-\mu, D^{-1}(a)) \approx e^{-a}$ (since also $-\mu \ll e^{-a}$ in all cases here).

In the context of Linguistics, one typically considers the dependence of the word frequencies f_m on the rank K with $K = 1$ ($K = N$) for the word with maximal (minimal) value of f_m. For example in (Piantadosi 2014), power law fits of the type $f_m \sim K^{-\alpha}$ with some fit exponent α are performed. In order to provide a presentation closer to this convention, we define the quantity $Z = w_m/\langle w_m\rangle$ (analogous to f_m apart from rescaling) which we consider as a function of the rescaled rank K/N. We note that the rank index K is a cumulative index and hence it is related with the cumulative distribution $C(w)$. Thus Z is actually just the inverse Pareto function and the corresponding plots in double logarithmic representation are very similar to the Pareto curves but with the roles of x- and y-axis exchanged since we can identify $C(w_m) = K/N$ (the number of words having $w_{m'} \geq w_m$ is just the rank K). We call the curves for the dependence of Z on K/N *Zipf curves* in analogy to the Zipf law that is a decaying power law with exponent $\alpha = 1$, even if these curves are not exact power laws.

For the RJE model, we can in principle extract from (6) a rather complicated analytic expression for the exact RJE Zipf curve $Z_{\rm RJE}(K/N)$ which is just the inverse function of (6). However, for practical purposes the approximation (7) is sufficient and in Zipf curve representation it becomes

$$Z_{\rm RJE}(K/N) \approx \frac{A_{\rm RJE}}{A_{\rm RJE}\langle w_m\rangle + (K/N)} \, . \tag{9}$$

This expression is valid for $Z \gg e^{-a}/\varepsilon$ implying $K/N \ll \varepsilon e^a/a$.

In the next section, we present figures which compare real data and (precise) RJE Lorenz, Pareto and Zipf curves for our 16 data sets. Later, we also discuss these results and the RJE approximations (7) and (9) in the context of a possible power or Zipf law (for a certain interval of parameters).

5. Results

In Table 1, we provide for the 16 data sets (ANC and certain books of certain writers) the Gini coefficient G, the total number of words N_w in the data set ($N_w = \sum_m f_m = N + \sum_m w_m$), the number N of different words and certain wealth/word inequality indicators which can be obtained from the Lorenz curve. These are the word frequency (wealth) share of bottom 50% of words (households) and similarly the wealth share of top 10% (top 1%) of words (households). Note, that in certain cases the bottom 50% share is simply 0% since here the household fraction of (different) words $h_{\rm min}$ with vanishing wealth $w_m = f_m - 1 = 0$ (minimal frequency $f_m = 1$) is larger than 0.5. For

Table 1
Word frequency (wealth) inequality indicators for the ANC data base and 15 book writers. The columns correspond to name, Gini coefficient G, the total number of words N_w, number of different words N, frequency share (wealth) held by the bottom 50%, the top 10%, and the top 1% of words (households) for each writer/data base. Values of 0% for the bottom 50% of words indicate that more than 50% of words have a vanishing frequency/wealth value $w_m = 0$ (corresponding to $f_m = 1$, $w_m = f_m - 1$). The ANC data is taken from (Corpus 2026).

Name	G	N_w	N	**Bottom 50%**	**Top 10%**	**Top 1%**
ANC	0.967	18300430	293867	0.06%	95.6%	75.9%
Asimov	0.910	70704	7323	0.12%	85.9%	52.6%
Austin	0.937	722747	14923	0.35%	90.6%	64.1%
Bible	0.944	789668	12765	0.37%	92.0%	67.6%
Cervantes	0.944	377633	22705	0.07%	91.5%	68.4%
Dante	0.931	97209	13603	0%	88.9%	59.7%
Dickens	0.934	356285	14159	0.50%	90.3%	65.0%
Goethe	0.926	77007	13734	0%	87.6%	58.7%
Hugo	0.940	67100	12865	0%	90.3%	63.4%
Melville	0.921	215304	17406	0.50%	87.6%	62.7%
Orwell	0.915	103731	8964	0.43%	86.7%	58.2%
Shakespeare	0.941	909658	26866	0.33%	91.1%	67.7%
Tolkien	0.930	518645	13928	0.50%	89.8%	61.1%
Tolstoy	0.929	451496	50869	0%	88.6%	62.6%
Twain	0.915	71221	7305	0.17%	86.8%	55.4%
Woolf	0.918	63958	7055	0%	87.4%	56.4%

the other cases, this bottom 50% share is between 0.06% and 0.5%. The top 10% (1%) have typically 85-95% (52-75%). These values are even more extreme than those for the word inequality in 2021 which is characterized by a high Gini coefficient $G = 0.843$ (Chancel et al. 2022; Frahm and Shepelyansky 2026b) and where 50% of the world population owns only 2% of total wealth while 10% (1%) of the population owns 75% (38%) of total wealth.

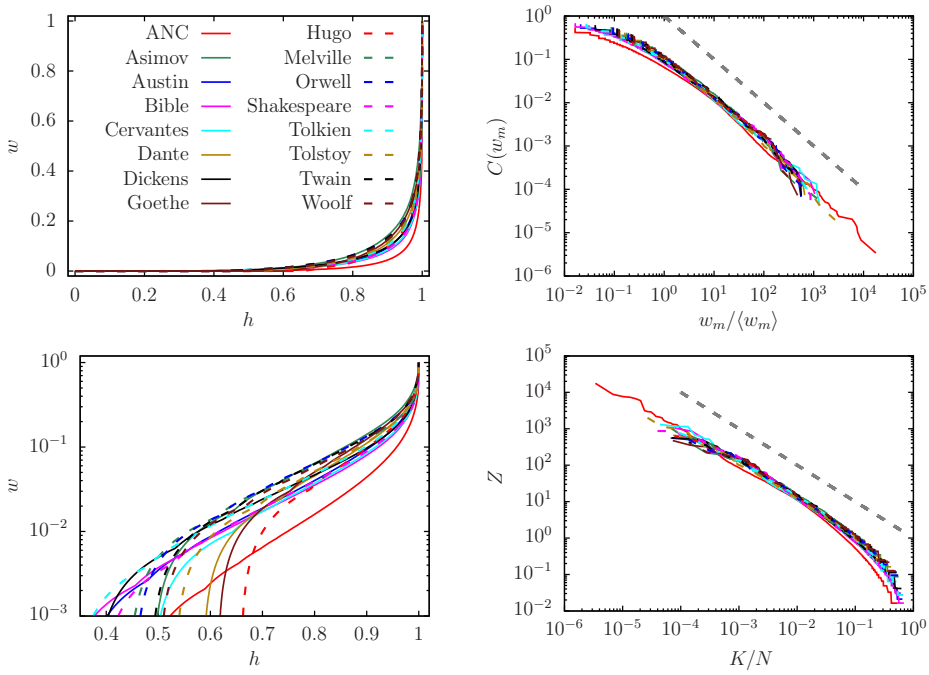


Figure 1
Top left: Lorenz curves for the 16 data sets of Table 1 (see text for details). The x-axis corresponds to the cumulated fraction of households/words (h) and the y-axis to the cumulated fraction of wealth/word frequency (w). The other panels show certain curves for the same data sets with same curve colors and types (dashed or not dashed). The Gini coefficients for all cases are given in Table 1. *Top right:* Pareto curves $C(w_m)$ for the same data where $C(w_m)$ represents the fraction of words m' with a word frequency $w_{m'} \geq w_m$ (analogous to wealth and with $w_m = f_m - 1$). The x-axis corresponds to values $w_m/\langle w_m \rangle$ where $\langle w_m \rangle = N_w/N - 1$ (see Table 1 for the values of N_w and N) is the average reduced name frequency. Both axis are in logarithmic scale and the thick dashed gray line shows the power law $C(w_m) = (w_m/\langle w_m \rangle)^{-1}$. *Bottom left:* Lorenz curves for the same data but in logarithmic scale for the cumulated wealth/frequency fraction w and a slightly zoomed range for the household variable $h \in [0.35, 1]$. Note that data points with $h < h_{\min}$, and therefore exactly $w(h) = 0$, are not visible due to the logarithmic scale in this panel (here $h_{\min}$ is the household fraction of words with vanishing wealth $w_m = f_m - 1 = 0$, i.e. minimal frequency $f_m = 1$). *Bottom right:* Zipf variable $Z = w_m/\langle w_m \rangle$ versus rescaled rank K/N in a double logarithmic scale and where the rank $K = 1$ ($K = N$) corresponds to the word m with maximal (minimal) value of w_m. In right panels the thick dashed gray line corresponds to the power law $Z = (K/N)^{-1}$. Note that K/N coincides with $C(w_m)$ and this panel is essentially identical to the top right panel but with the roles of x- and y-axis exchanged.

In Fig. 1 we present the Lorenz, Pareto and Zipf curves for all 16 data sets. The Lorenz curve presentation gives more stress to the cumulated word/wealth intermediate range $h = (m-1)/N$ below unity. All 16 cases have a big phase with the RJ condensation: for 50% of words the fraction of the whole word frequency or wealth w is below 0.5% being by a factor 4 or even more (see Table 1) smaller than for the world wealth inequality with $w = 2\%$. There is also a certain $?? \leq h_{min} \leq ??$ border below which $w = 0$ depending on a considered data set. In the RJ condensate phase the number of words is small and hence the fluctuations of Lorenz curves of 16 sets are significant. However, starting from $h > 0.65$ when $w > 0.01$ the Lorenz curves are more compact demonstrating a presence of global stationary steady-state distribution of cummulated word frequency (or wealth). There is a visible deviation from this stationary state for

the ANC case that we attribute to the fact that the ANC data is a composition of many literary works written by many different writers. Indeed, data of Table 1 shows that the books of given writers have their own Gini coefficients being in the narrow range $0.915 \leq 0.944$ while the ANC set has visibly higher $G = 0.967$. This can be a result of mixing of works/books of very different authors while each author is characterized by his/her own style of writing. We note that the dashed line in the right panels of Fig. 1 shows the usual Zipf and Pareto algebraic dependence $Z \sim 1/K$ that captures the global probability decay of data even if there is a certain curvature present in Zipf curves that we discuss below.

Indeed, below we show that each of 16 cases of Table 1 has different average value of frequency/wealth $\langle w_m \rangle$ being in the range $[4.22, 60.9]$??. If the distributions of word frequencies are rescaled with their own $\langle w_m \rangle$ values then the steady-state Pareto and Zipf curves, in right panels of Fig. 1, clearly show the presence of the universal stationary distribution identical for all 16 cases. We attribute this ubiquity of word frequency distribution to the universal feature of RJ thermalization and condensation based of statistical laws of physics.

Below we discuss the properties of Lorenz, Pareto and Zipf distributions for 3 literary works from Table 1 and compare them with the theory results of the RJE thermalization model. The comparison for other 13 cases are presented in Appendix Figs. A.1 – A.13.

As explained in the previous Section, we determine the optimal RJE parameters a, ε for the 16 data sets. The obtained values of the parameter a are rather high, typically $a =$ 5-11, the values of ε are quite small with $\varepsilon \sim 10^{-5}$-10^{-3} (specific values are given in the Figures below and in the Appendix). Thus all 16 cases are located in the RJ condensate phase. The related values of $-\mu$ are even smaller (typically $-\mu/\varepsilon \sim 10^{-3}$) indicating that all cases correspond to a strong RJ condensation. We note that it is possible to show from (5) that $0 < -\mu \ll \varepsilon$ if $\varepsilon \ll 1$ (see (Frahm, Ermann, and Shepelyansky 2026)). The two maximal values $a = 10.7$ and $a = 9.05$ are obtained for ANC or the Bible respectively.

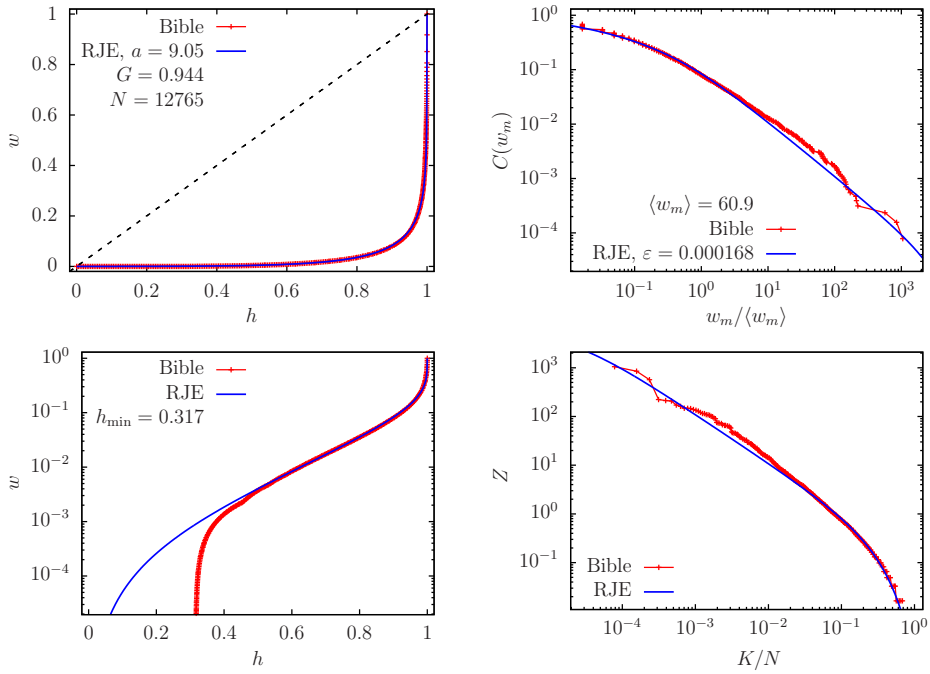


Figure 2
All panels are in the same style as the corresponding panels in Fig. 1 but only for two curves in each panel being the data curve for the King James Bible (red data points) and the corresponding optimal RJE curve shown in blue. Furthermore, in the top left panel for the Lorenz curves also the (dashed black) line of perfect equipartition $w = h$ is shown. The two top panels also give the values of G, N, the real data average $\langle w_m \rangle$ and the two optimal RJE parameters a and ε. The Lorenz curves in the bottom left panel are shown in logarithmic scale for the cumulated frequency/wealth fraction w and in full (normal) scale for the word/household variable $h \in [0, 1]$. Due to the logarithmic scale the (red) data points for $h < h_{\min}$ with $w(h) = 0$ are not visible and the range of w is chosen such that all other data points with $w(h) > 0$ are visible (the value of $h_{\min}$ is given in the panel).

In Fig. 2 we present the comparison of real data Lorenz, Pareto and Zipf curves for the case of Bible with the RJE theory results. There is the striking good agreement between data and the RJE theory. Certainly, there the RJE theory describes well the Lorenz curve for values $w > 10^{-3}$ while below this value there is a divergence between theory and data that we attribute to the low frequency statistics of rare words. For the Pareto and Zipf distributions in the right panels of Fig. 2 the RJE theory gives a strikingly good description of data on the whole scale of the distribution variations being 4 orders of magnitude. We stress that the Pareto and Zipf curves have well visible curvature and cannot be described by a simple algebraic decay with a certain exponent $\alpha \approx 1$. This curvature is well captured and described by the RJE theory.

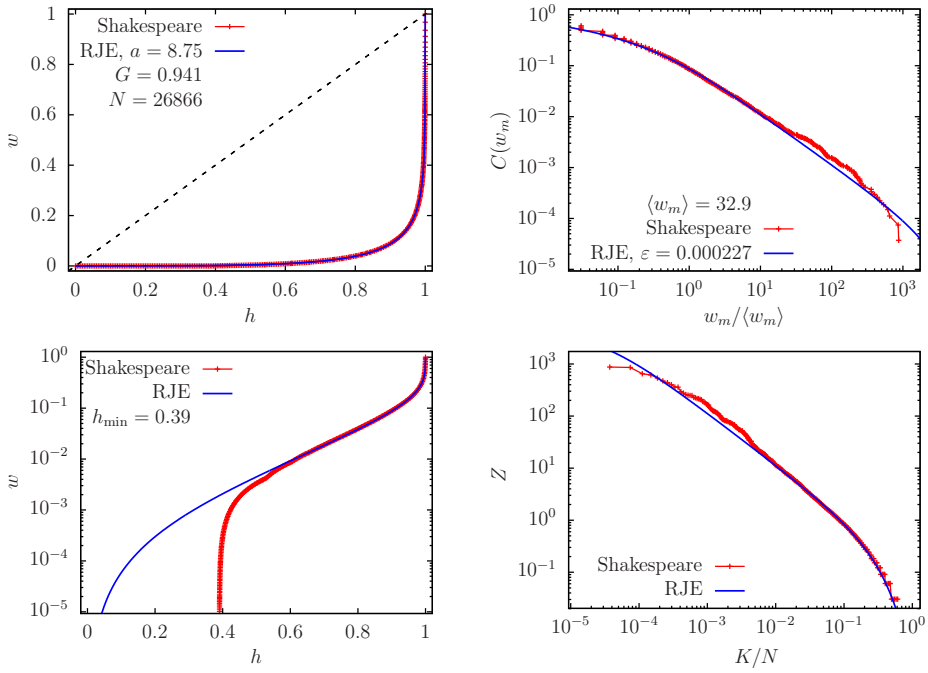


Figure 3
As Fig. 2 for Shakespeare.

For the case of Shakespeare the comparison of data and RJE theory is presented in Fig. 3 constructed in the same way as Fig. 2 for the Bible case. Here, as for the case of Bible, we also find very good agreement of the RJE theory with the real data. Of course, for the values of Lorenz curve with $w < 10^{-3}$ there is a difference between theory and real data which we attribute to low statistics of rare words with $f_m = 1$. The RJE theory well describes the variations of Pareto and Zipf probabilities in a range of 4 orders of magnitude.

In Fig. 4 we present the comparison of the RJE theory with the data of Hugo book *Les Miserables* in French. Again we find a very good agreement between data sets and the theory. In this case the statistics of words N_w is by a factor 10 smaller compared to the cases of Bible and Shakespeare (see Table 1). Due to that for the Lorenz curve of w values the agreement works only for $w > 0.01$. The RJE theory well describes the Pareto and Zips distribution variations on a scale of 3 orders of magnitude. The theory captures well the curved feature of these distributions.

In Fig. 5 we give the color view of Lorenz curves obtained within the RJE theory for all 16 literary works if Table 1. The theory gives almost exact values of cumulated wealth or rescaled word frequencies w for $w > 0.01$ for all 16 cases as it is shown in Figs. 2-4 and Appendix Figs. A.1-A.13. There are certain variations from one case to another correlated with the variations of Gini coefficient (see Table 1). At the same time the Pareto distributions $C(w_m)$ rescaled by their average values $\langle w_m \rangle$ give very good description of data with the universal RJE theory curve shape seen in Figs. 2-4 and Appendix Figs. A.1-A.13. We illustrate this universality of the RJE Pareto curves in Fig. 6 showing all 16 cases of Fig. 5 for the rescaled Pareto distributions $C(w)$. Indeed, the results of Fig. 6 show that, except of fluctuations, we have very similar Pareto distributions $C(w_m)$ vs. $w_m/\langle w_m \rangle$ for all 16 cases of Table 1.

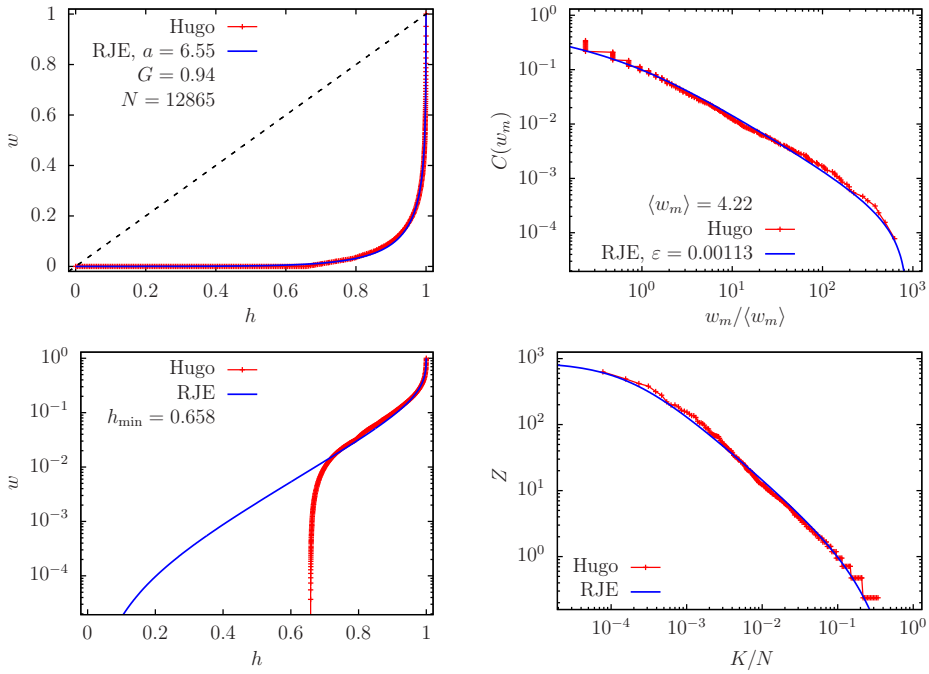


Figure 4
As Fig. 2 for Hugo.

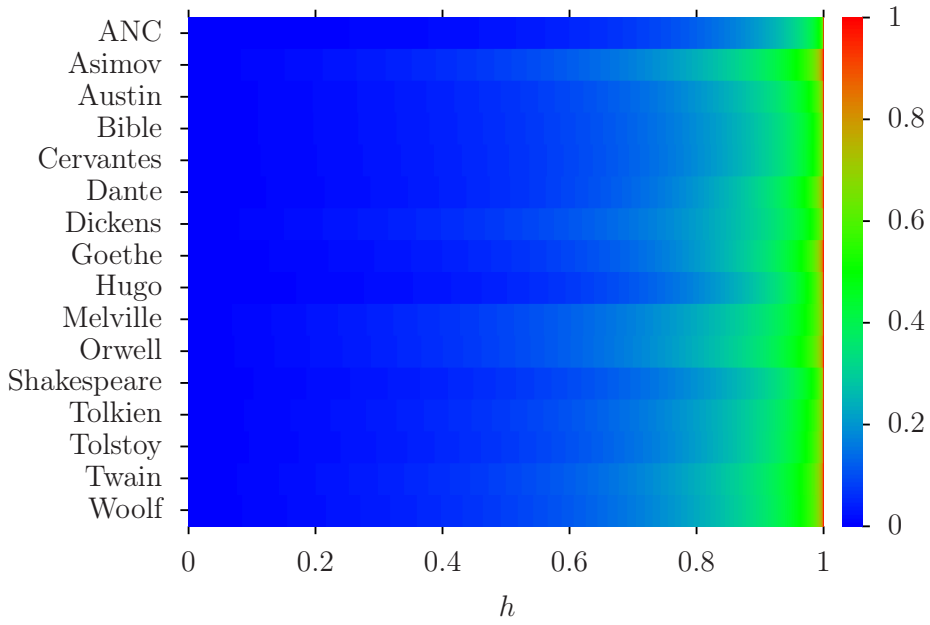


Figure 5
Density color plot of Lorenz curves for the 16 cases of Table 1 obtained from the RJE model with optimal RJE parameters for ε and a. The x-axis corresponds to the cumulated fraction of households/words (h) and the y-axis to data sets/writers (in alphabetical order). The color value corresponds to the cumulated fraction of wealth or rescaled word frequency (w). In order to enhance small values of w a nonlinear scale is chosen where the values of the color bar correspond to $w^{1/2}$. The RJE Lorenz curves here correspond to the optimal RJE fits shown as blues lines in the top left panels of Figs. 2-4 and Appendix Figs. A.1-A.13 (see also these figures for the values of ε and a).

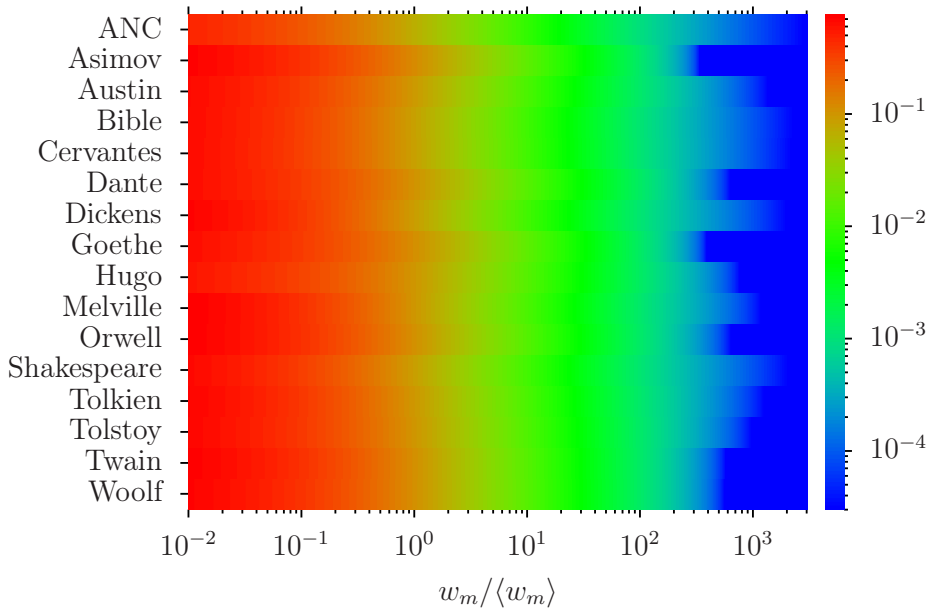


Figure 6
Density color plot of Pareto curves for the 16 cases of Table 1 obtained from the RJE model with optimal RJE parameters for ε and a. The x-axis corresponds to the rescaled wealth $w_m/\langle w_m \rangle$ (in logarithmic scale) and the y-axis to data sets/writers (in alphabetical order). The color value (in logarithmic scale) corresponds to the RJE value of the Pareto function $C_{\rm RJE}(w_m)$ given by Eq. (6). Values of $C_{\rm RJE}(w_m)$ below the minimal color value 3×10^{-5}, in particular if $C_{\rm RJE}(w_m) = 0$ for $w_m > w_{\rm max}$, are also shown by blue color for the minimal value. (Note for the RJE model, we have $w_{\rm max} = 1$, $\langle w_m \rangle = \varepsilon$ and $w_{\rm max}/\langle w_m \rangle = 1/\varepsilon$.) The RJE Pareto curves here correspond to the blues lines in the top right panels of Figs. 2-4 and A.1-A.13 (see also these figures for the values of ε and a).

In this Section we presented results for the cases of Bible, Shakespeare and Hugo (see Figs. 2-4). Other 13 cases are shown in Appendix Figs. A.1-A.13 being composed in the same way as Fig. 2. Thus the ANC case is presented in Fig. A.1; here the statistics of all words N_w is very high and the RJE theory describes well the Pareto and Zipf distributions on a scale of 5 orders of magnitude. The cases of Asimov, Austin, Cervantes, Dante, Dickens, Goethe, Melville, Orwell, Tolkien, Tolstoy, Twain and Woolf are presented in Appendix Figs. A.2-A.13 respectively. All these cases have a god agreement of data with the RJE theory.

We also point that at high values of parameter a and low values of ε the RJE theory for Pareto curve gives an approximate algebraic decay of probability with the exponent $\alpha = 1$ in a certain restricted range (Frahm, Ermann, and Shepelyansky 2026). Due to the direct relation between Pareto and Zips curves this is also valid for the Zipf distribution. We discuss this feature in Appendix with illustrations in Fig. A.14 and Table A.1.

6. Discussion

In this work we presented the concept of RJ thermalization and condensation of words in literary works (or books) of different writers and different languages. In this WOTH approach the different N words of a book are analogous to independent classical particles and also analogous to households in the wealth inequality problem. The inter-

actions between these words (or particles) appear due to semantic rules, meanings and writer concepts that generate links and nonlinear couplings between N words. These interactions are not known but as for a gas of classical particles in three dimensions almost any weak or moderate interaction between particles drives a system to the thermal state with kinetic energy per particles being $3kT/2$ where k and T are the Boltzmann constant and gas temperature.

In the frame of WOTH thermodynamic theory all words in a book are distributed over a certain states levels characterizing the word frequencies (and corresponding tokens). These levels are analogous to the energy levels in a physical system of classical particles, or to frequency or token levels for a book. This is also similar to the wealth state levels of households analyzed in (Frahm and Shepelyansky 2026b; Frahm, Ermann, and Shepelyansky 2026). We propose the RJE model of this energy (token) spectrum with a density of states decreasing at high word frequencies (or at high energies or high wealth revenues) characterized by a certain system parameter a. The thermal steady-state distribution of frequencies or tokens over these state levels is described by the classical RJ thermalization law (1) characterized by the rescaled total system energy ε.

Using these RJ thermalization theory parameters ε, a we compare the results of the RJE distribution of word frequencies of 16 literary works described in Section 2 and Table 1. We find that the RJE theory gives the strikingly good description of real data of Lorenz, Pareto and Zipf distributions for these 16 cases. This comparison shows that the rescaled effective system energy ε is very low and due to that there is the phenomenon of RJ condensation of a significant fraction of total probability (fraction of all different words N) at the lowest energy (token) level. This leads to a huge fraction of different words at lowest frequency $f_m = 1$ where there can be present about 30% or more of total number of words of a book. Also a 50% fraction of all different words contains only less than 0.5% of total number of book words N_w (see Table 1). This represents the phenomenon of the RJ condensation studied previously in physics of light propagation in multimode optical fibers (Connaughton et al. 2005; Krupa et al. 2016; Baudin et al. 2020) and for the wealth inequality in the world (Frahm and Shepelyansky 2026b; Frahm, Ermann, and Shepelyansky 2026) (and also constraint driven condensation in other physical systems (Evans et al. 2010; Szavits-Nossan, Evans, and Majumdar 2014; Filiasi et al. 2014)).

The WOTH theory gives a very good description of the word frequencies in 16 considered cases but one can object that there are two fit parameters in these RJE distributions. This is true even if it is better to have a good comparison of theory with real data even if there are two fit parameters. However, there are important conceptual features that in our opinion provide a strong support to the WOTH concept:

F1) the distribution of word frequencies is a stable steady-state being very similar for all 16 cases considered (see Fig. 1, Fig. 6, there is also a rather narrow variation interval for Gini coefficient in Table 1), this is in agreement with the thermodynamic stability of the RJ thermal distribution;

F2) a huge condensation of rare words at lowest word frequencies is in a qualitative and quantitative agreement with the RJ condensation phenomenon;

F3) The RJ type distributions determined by tthermalization and condensation appear in very different systems like wealth inequality of households in the world, stock market capitalization of companies, GDP distribution over world countries (Frahm and Shepelyansky 2026b; Frahm, Ermann, and Shepelyansky 2026) and word frequencies in books; this stress the universal features of the RJ thermal concept,

We note that recently the thermodynamic concepts start to be discussed in relations to LLM (see e.g. [cite:vetrov] and Refs. therein) and we hope that the thermodynamics of words established here will find useful applications there.

Appendix A:

In this appendix, we present additional figures in the same style of Fig. 2 for the other data sets not shown in the main part and also a discussion about the limits of a potential power law for the Zipf curves of real data.

1.1 Additional figures

Here, we present Figs. A.1-A.13 which are in the same style as Fig. 2 for the cases of ANC, Asimov, Austin, Cervantes, Dante, Dickens, Goethe, Melville, Orwell, Tolkien, Tolstoy, Twain, and Woolf. These figures are mostly discussed in the main part of the text. Here, we point that globally all real data curves (red data points) agree very well with the corresponding RJE theory curves (blue lines) for optimal parameters a and ε given in Figures for each data set.

For the case of Tolstoy in Fig. A.11, there are some modest deviations at the tail of the Pareto curve with $C(w) < 10^{-4}$ (initial part of the Zipf curve) between data and the RJE theory. We note in analogy with the WTH concept for the wealth inequality (Frahm and Shepelyansky 2026b; Frahm, Ermann, and Shepelyansky 2026) that this tail corresponds to the case of very rich households (a lot of tokens for these words). Such households or words of very high frequency are rare that leads to a lot statistics of such words at Pareto tail.

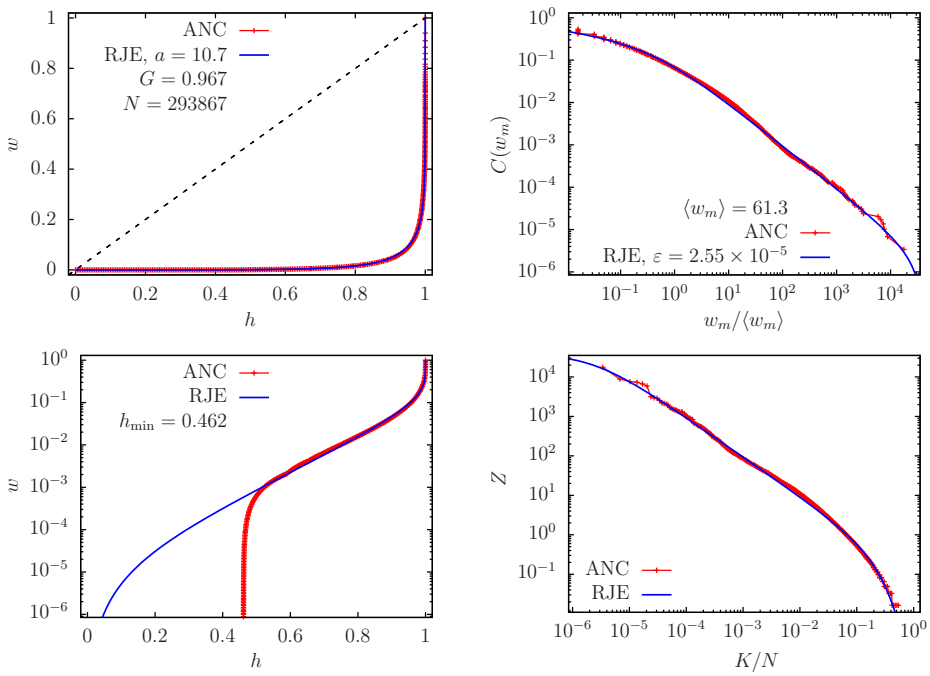


Figure A.1
As Fig. 2 for ANC.

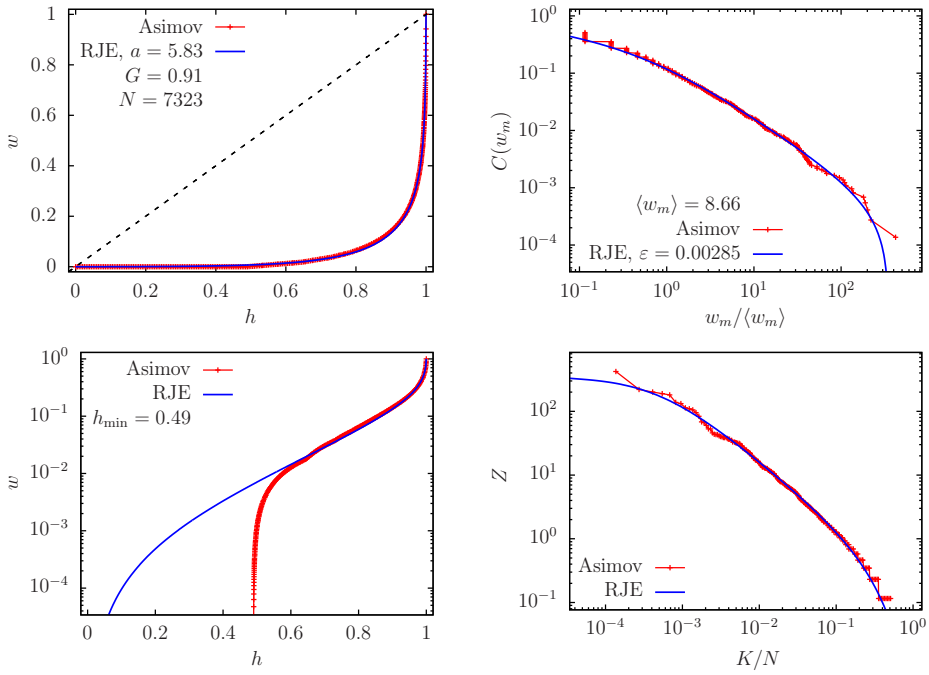


Figure A.2
As Fig. 2 for Asimov.

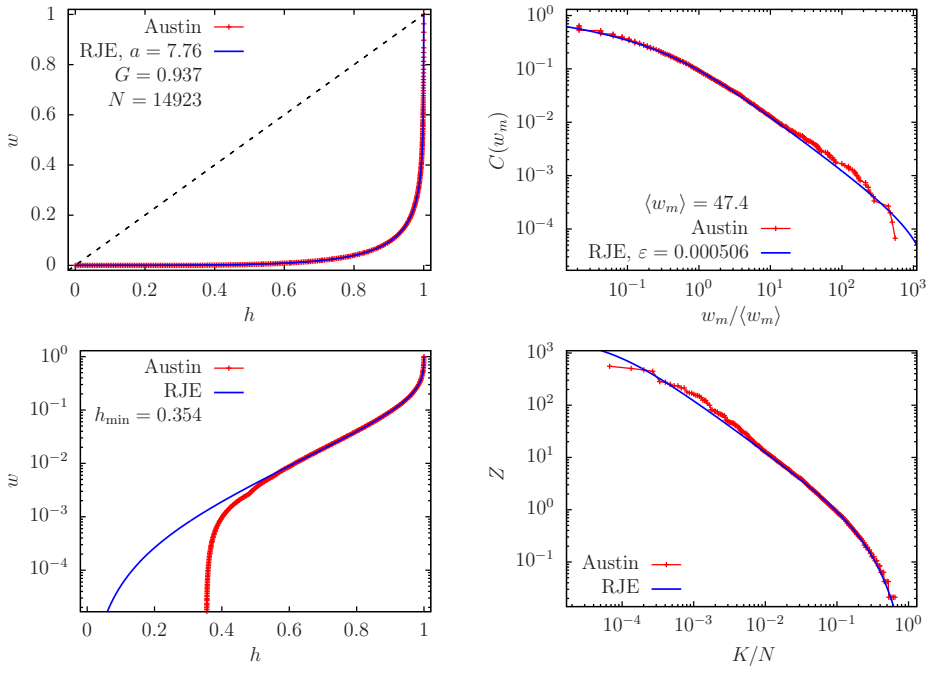


Figure A.3
As Fig. 2 for Austin.

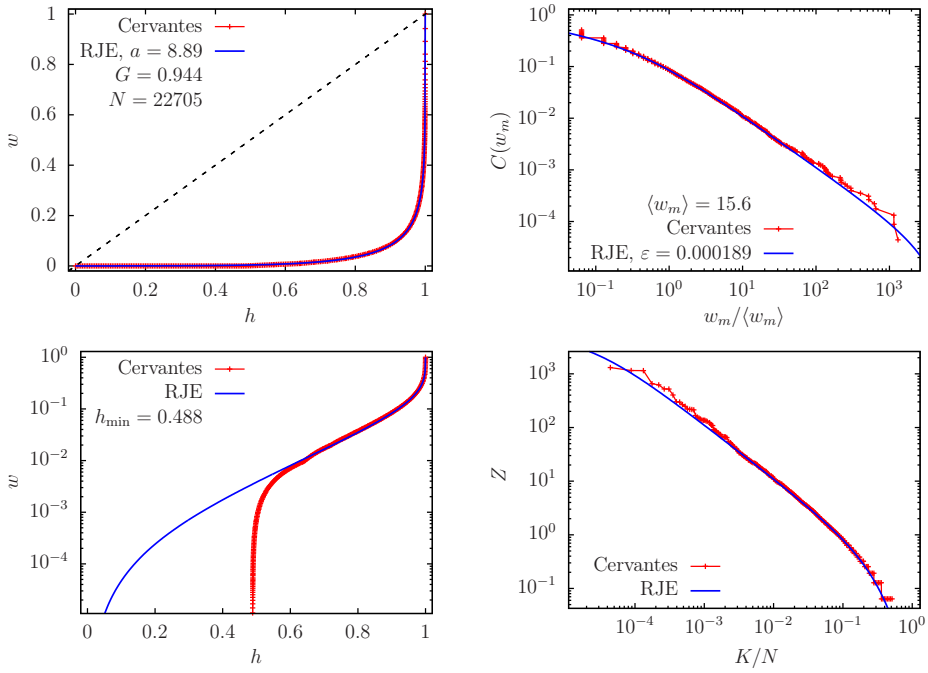


Figure A.4
As Fig. 2 for Cervantes.

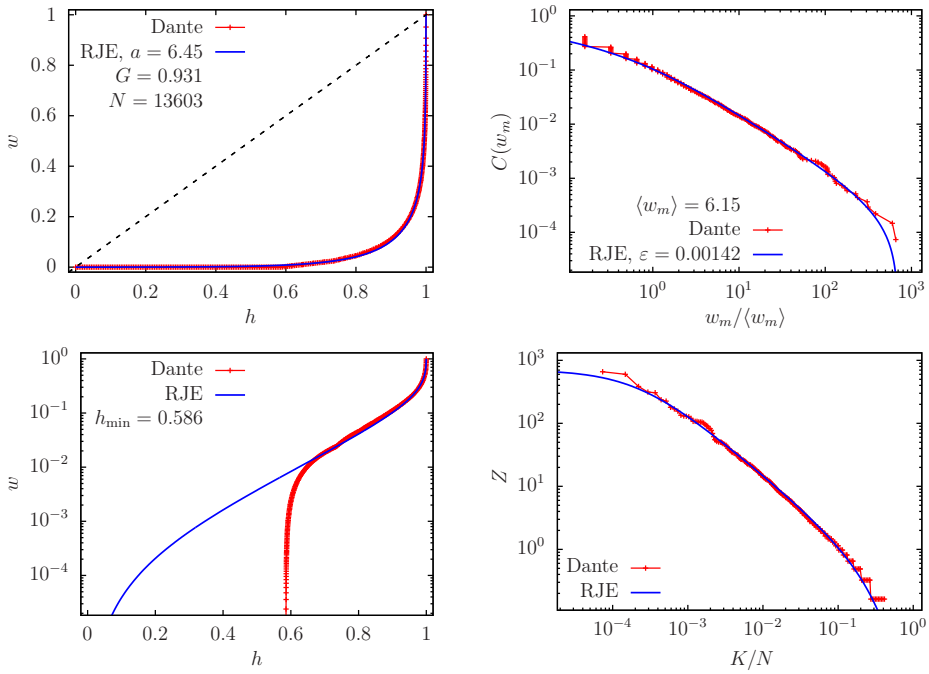


Figure A.5
As Fig. 2 for Dante.

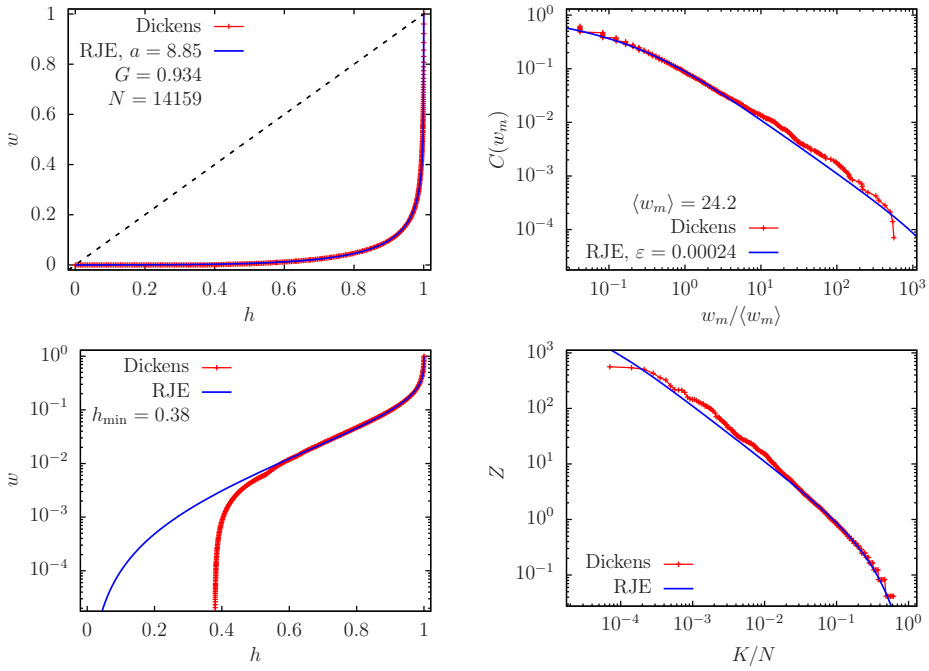


Figure A.6
As Fig. 2 for Dickens.

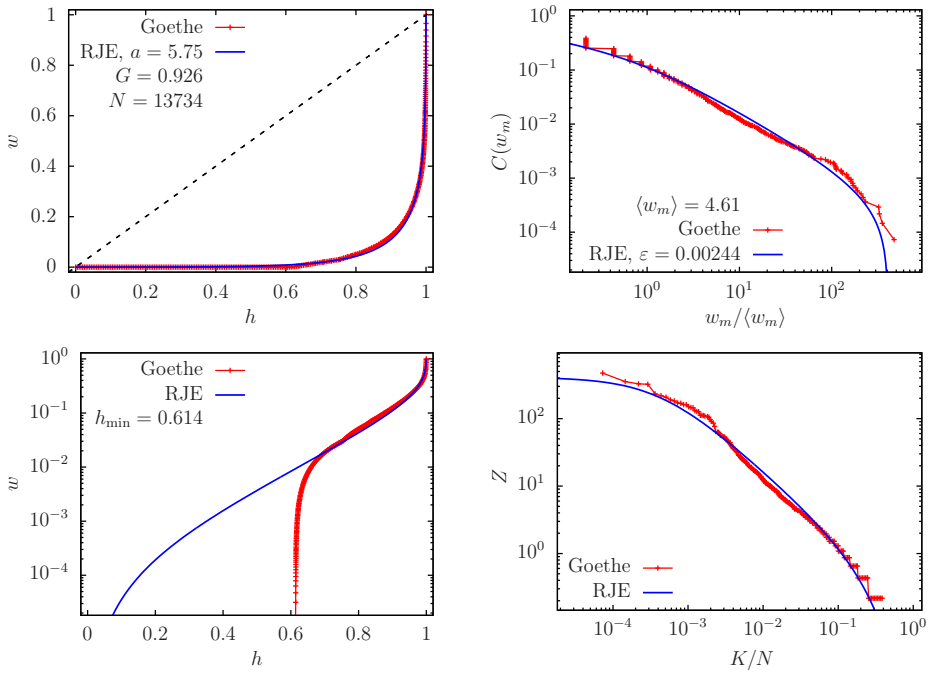


Figure A.7
As Fig. 2 for Goethe.

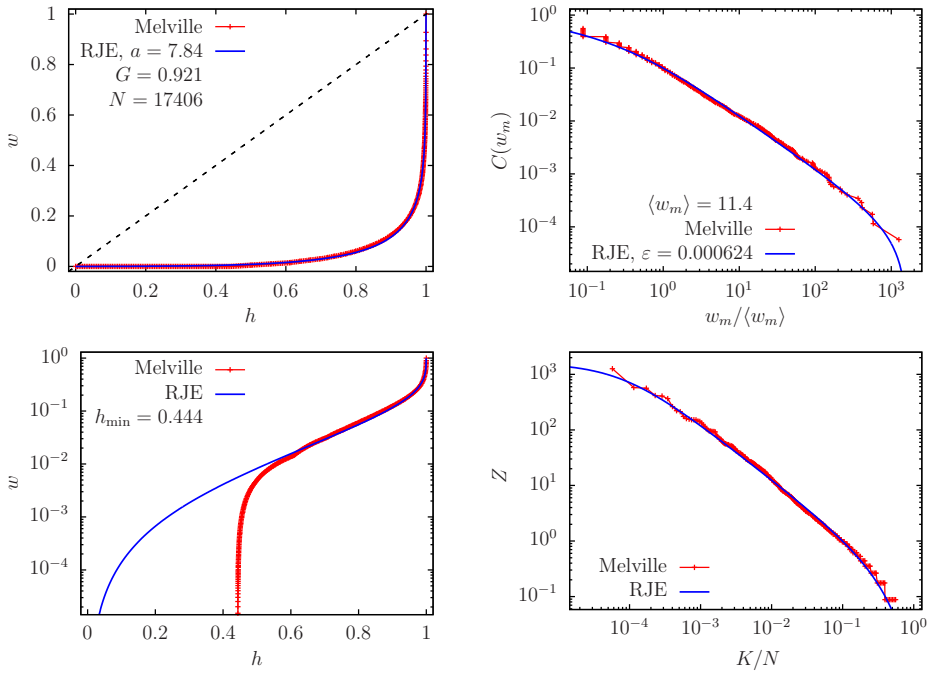


Figure A.8
As Fig. 2 for Melville.

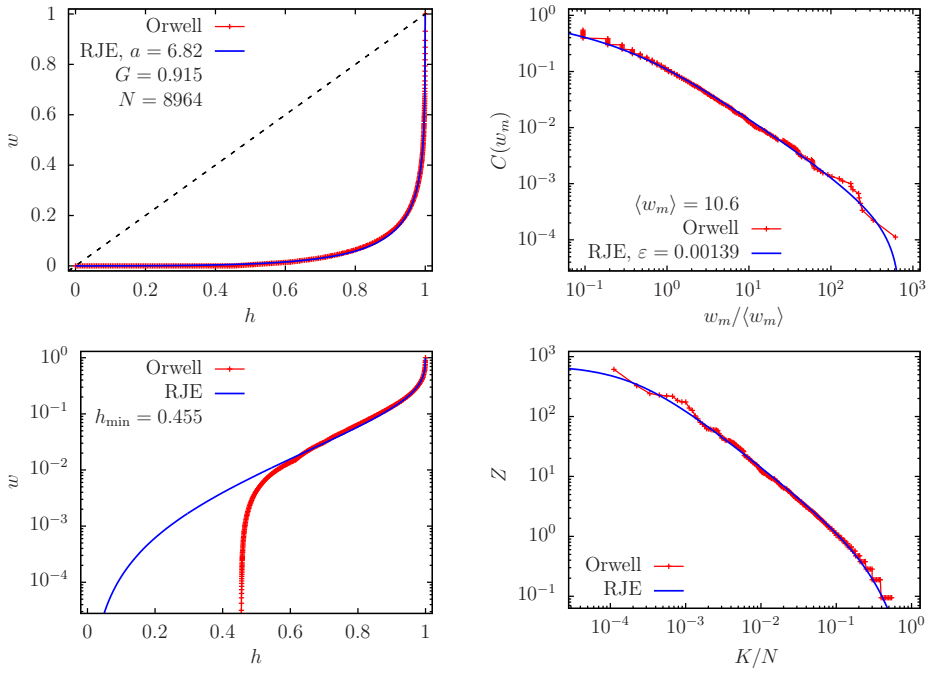


Figure A.9
As Fig. 2 for Orwell.

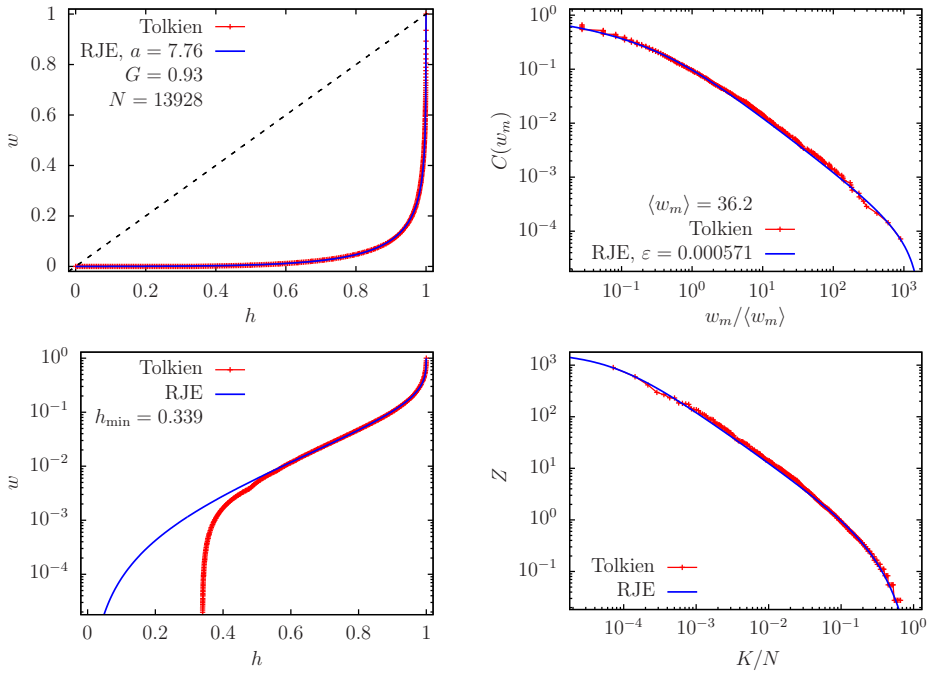


Figure A.10
As Fig. 2 for Tolkien.

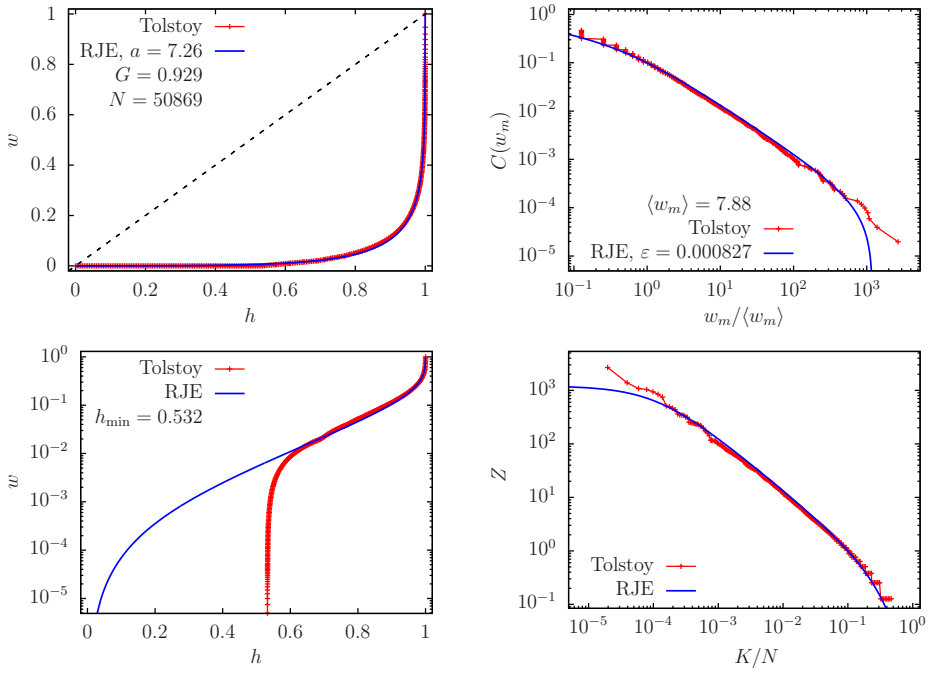


Figure A.11
As Fig. 2 for Tolstoy.

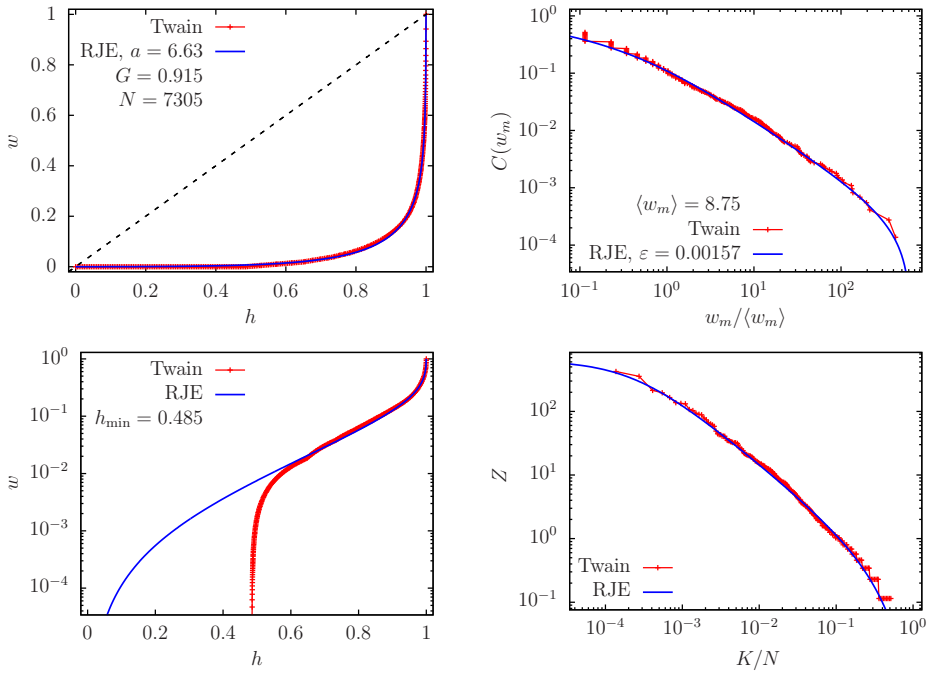


Figure A.12
As Fig. 2 for Twain.

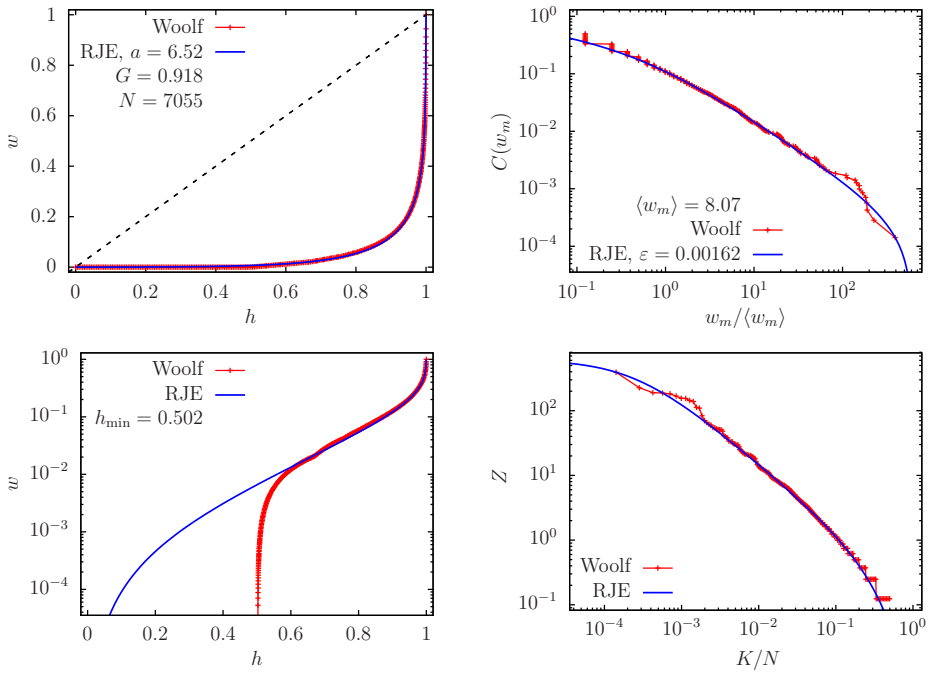


Figure A.13
As Fig. 2 for Woolf.

1.2 Discussion of the limits of the Zipf law

The analytical RJE approximation (7) for the Pareto function, valid for $w_m \gg e^{-a}$ (assuming $|\mu| \ll e^{-a}$ which is well verified for all cases here), resembles partly a power law behavior except for the constant offset that ensures that $C_{\rm RJE}(1) = 0$. If we assume in addition that also $w_m \ll 1$, we can neglect this offset which indeed gives the Zipf law, i.e. a decaying power law with exponent "$\alpha = 1$":

$$C_{\rm RJE}(w_m) \approx \left(\frac{\varepsilon - \mu}{a}\right) w_m^{-1} = A_{\rm RJE} \left(\frac{w_m}{\langle w_m \rangle}\right)^{-1} \quad \text{if} \quad e^{-a} \ll w_m \ll 1 \; . \tag{A.1}$$

To have a significant interval of validity for this expression, we need rather larger values of a (small values of $D^{-1}(a) \approx e^{-a}$). If this condition is verified the analytical Pareto curve of the RJE model is close to a decaying power law with exponent "$\alpha = 1$" for this interval.

Concerning, the Zipf expression (9), we can add the additional condition $K/N \gg A_{\rm RJE}\langle w_m \rangle \approx \varepsilon/a$, such that we can neglect the constant term in the denominator which also gives the Zipf law

$$Z \approx A_{\rm RJE} \left(\frac{K}{N}\right)^{-1} \quad \text{if} \quad \frac{\varepsilon}{a} \ll \frac{K}{N} \ll \frac{\varepsilon e^a}{a} \; . \tag{A.2}$$

This expression corresponds to the approximation (A.1) for the Pareto function which are both valid for some intermediate interval in either K or w_m.

Table A.1
Table of power law fit parameters for the ANC data base and 15 books of writers. The columns correspond to name, the RJE power law coefficient $A_{\rm RJE} = (\varepsilon - \mu)/(\varepsilon a) \approx 1/a$, the Zipf fit parameters $A_{\rm Zipf}$, α and the min/max values of $Z = w_m/\langle w_m \rangle$ for the used fit interval. These parameters correspond either to the RJE power law behavior $C_{\rm RJE}(w) \approx A_{\rm RJE}(w_m/\langle w_m \rangle)^{-1}$ or the Zipf power law fit $K/N = C_{\rm data}(w_m) = A_{\rm Zipf}(w_m/\langle w_m \rangle)^{-1/\alpha} = A_{\rm Zipf} Z^{-1/\alpha}$. The fit interval borders have been chosen as $Z_{\rm min} = 2e^{-a}/\varepsilon$ and $Z_{\rm max} = 0.1/\varepsilon$ and we expect the RJE power law behavior to be approximately valid for $Z_{\rm min} < Z < Z_{\rm max}$. The statistical fit errors of $A_{\rm Zipf}$ and α are well below 0.5% for all cases.

Name	$A_{\rm RJE}$	$A_{\rm Zipf}$	α	$Z_{\rm min}$	$Z_{\rm max}$
ANC	0.0933	0.0847	1.07	1.73	3.93×10^3
Asimov	0.172	0.129	1.09	2.06	35.1
Austin	0.129	0.101	1.12	1.69	198
Bible	0.111	0.0896	1.16	1.4	594
Cervantes	0.113	0.0917	1.09	1.45	530
Dante	0.155	0.113	1.08	2.23	70.6
Dickens	0.113	0.0894	1.17	1.2	417
Goethe	0.174	0.128	0.999	2.6	40.9
Hugo	0.153	0.103	1.1	2.54	88.5
Melville	0.128	0.0963	1.1	1.27	160
Orwell	0.147	0.115	1.06	1.57	71.9
Shakespeare	0.115	0.0877	1.14	1.4	441
Tolkien	0.129	0.107	1.11	1.49	175
Tolstoy	0.138	0.11	1.02	1.7	121
Twain	0.151	0.113	1.13	1.68	63.6
Woolf	0.154	0.123	1.07	1.82	61.7

To test more explicitly the validity of these two expressions for real data, or more generally their power law properties, we perform for each data set a power law fit $C_{\rm data}(w) = A_{\rm Zipf}(w_m/\langle w_m \rangle)^{-1/\alpha}$ for the fit interval $Z = w_m/\langle w_m \rangle \in [2e^{-a}/\varepsilon, 0.1/\varepsilon]$ where a and ε are the optimal RJE parameters for the used data set and $\langle w_m \rangle$ is real data average of $w_m = f_m - 1$ (see previous figures for specific values for each data set). Here, we choose the notation "$-1/\alpha$" for the fit exponent to be coherent with the notations of (Piantadosi 2014) where $Z \sim w_m \sim f_m \sim K^{-\alpha} \sim C(w)^{-\alpha}$. We expect, that the analytical RJE power law approximations (A.1) and (A.2) are roughly valid for the proposed fit interval above. Therefore, it is reasonable to use the same interval (in the rescaled variable $Z = w_m/\langle w_m \rangle$) to perform a power law fit for the real data. (To be more precise the power law fit is actually done as the linear fit $y(x) = b + px$ with $y = \ln(C_{\rm data}(w))$, $x = \ln(w_m/\langle w_m \rangle)$, $b = \ln(A_{\rm Zipf})$ and $p = -1/\alpha$.)

In Table A.1, we provide for each data set the RJE power law coefficient $A_{\rm RJE}$ defined in Eq. (8), the two fit parameters $A_{\rm Zipf}$, α and also the numerical values of the borders of the fit interval (in units of $\langle w_m \rangle$, i.e. for $Z = w_m/\langle w_m \rangle$). The obtained values of α are typically in the interval $[1, 1.17]$ with most values $\alpha \approx 1.1$ and minimal value $\alpha = 0.999$ for Goethe. The values of the coefficients $A_{\rm RJE} \approx 1/a$ and $A_{\rm Zipf}$ are quite close with roughly comparable values.

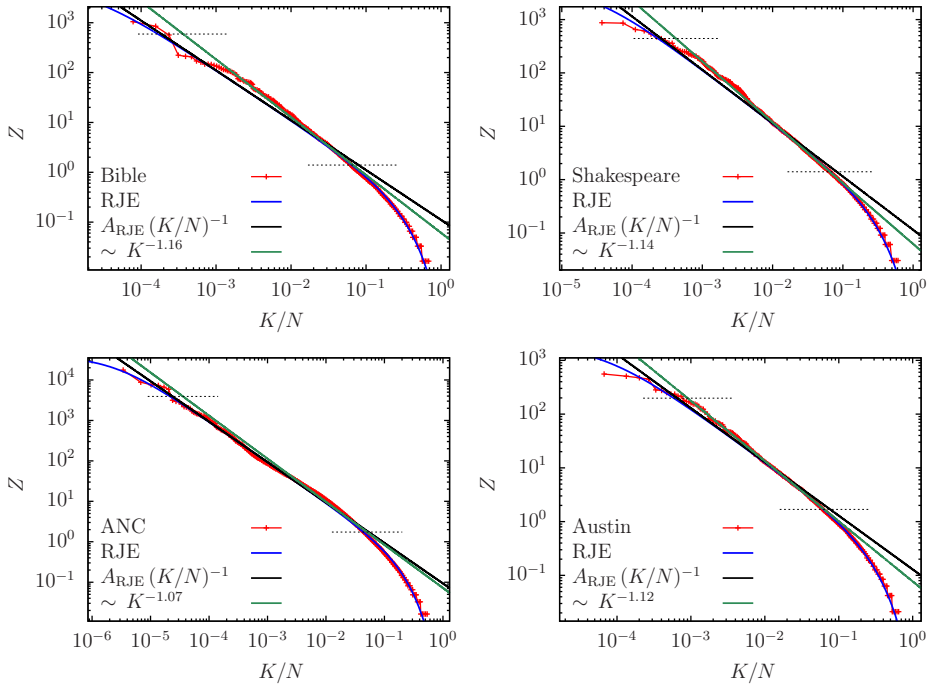


Figure A.14
Graphical illustration of the power law fits according to Table A.1 for the cases of Bible, Shakespeare, ANC and Austin. Shown are the Zipf curves of Z versus the rescaled rank K/N for the word data (red data points) and the precise optimal RJE curve (blue line). The two straight lines correspond to the analytical RJE power law approximation $Z = A_{\rm RJE}(K/N)^{-1}$ (black line) and the power law expression $Z = (K/(NA_{\rm Zipf}))^{-\alpha}$ (green line) obtained from the fit $K/N = A_{\rm Zipf} Z^{-1/\alpha}$. The thin horizontal black dotted lines indicate the values of $Z_{\rm min}$ and $Z_{\rm max}$ for the used fit interval given in Table A.1.

In Fig. A.14, we provide a graphical illustration of this power law fit for the Zipf curves of the data for the four examples of Bible, Shakespeare, ANC and Austin. As in the previous figures, the real word data corresponds to the red data points and the exact RJE curve (for optimal values a and ε for the given data set) to the blue line. The analytical RJE power law approximation (A.2 with exponent "-1" is shown as the black line and the above power law fit corresponds to the green line. In the figure panels, also the values of $Z_{\rm min} = 2e^{-a}/\varepsilon$ and $Z_{\rm max} = 0.1/\varepsilon$ for the used fit interval (in the Zipf variable Z) are marked as thin dotted black lines.

First, we observe that the blue line (exact RJE expression obtained as the inverse function of (6)) and black line (analytic RJE power law approximation (A.2)) are quite close for the values in the fit interval which is expected. There are slight deviations between them at the interval boundaries which is due to the fact that the chosen interval is slightly on the "larger" side and that the conditions of validity of (A.2) are not perfectly verified at its boundaries. The green line for the power law fit is rather close to the black line but there are some visible deviations and the slope of the green line corresponds to values of "$-\alpha$" where α is slightly larger than unity. The green fit line agrees typically very well with most of the (red) data points inside the fit interval except for a small number of data points at the top of the interval where the density of data points is also reduced and which have a smaller influence on the fit result. As already

observed in the previous figures, the blue RJE curve agrees globally quite well with the data nearly for all data points, also outside the fit interval and in particular in the range of large $K/N > 0.1$ where both data and precise RJE theory are very different from the power law behavior. Concerning the range of smallest values of $K \leq$ 5-10 (top frequent words with largest values of w_m), the data points are not very far from the black line for the cases of Bible and ANC while for Shakespeare and and Austin they deviate a bit more from both blue and black RJE curves (and even more from the green fit line). This region is statically not very reliable and concerns only a few data points. However, also here the blue RJE curve still provides the best agreement in comparison to the other two power law lines.

We mention, that the more refined approximation (9) (not shown in the figure), which is valid for $K/N \ll \varepsilon e^a/a$, agrees quite precisely with the exact blue RJE curve for smaller values of $K/N < 10^{-2}$ and with the black line for larger values of $K/N > 10^{-2}$.

We have also tested the case of a reduced fit interval, i.e. $Z_{\min} = 4e^a/\varepsilon$ and $Z_{\max} = 0.05/\varepsilon$ (factor 2 modification of both interval borders to the inside). In this case, the obtained fit values of A_{Zipf} and α are still quite similar. However, the values of α are generally a bit closer to unity, e.g. for the four cases of Fig. A.14, we have $\alpha = 1.12, 1.15, 0.993, 1.1$ instead of the values $\alpha = 1.16, 1.14, 1.07, 1.12$ in Fig. A.14 for the initial fit interval. For the reduced interval, the blue and green lines are typically a bit closer.

In summary, we see that the RJE Zipf curves exhibit for the cases of rather large values of a and small values of ε the Zipf power law in some intermediate interval, $\varepsilon/a \ll K/N \ll \varepsilon e^a/a$. which can be quite large. However, for the regions outside this interval there are significant deviations from this power law, and as we can see in Fig. A.14 and also in the previous figures, the RJE theory curves fit remarkably well the real word distribution data for most cases and the full range with only two free parameters being ε and a. We note, that typical power law type fits of (Piantadosi 2014) are done with at least two parameters being the exponent and the numerical prefactor. In certain cases, a third parameter β for a shifted power law, $f_m \sim Z \sim (\beta + K)^{-\alpha}$, is added which is similar to the offset term in either (7) or (9) but there this term is simply fixed by the theory and does not add a third parameter. We note that even with this extension, there are still significant deviations with the data (used in this work) for larger values of $K/N > 0.1$.

The above discussion focuses on the Zipf curves but the situation is very similar for Pareto curves where the roles of x- and y-axis exchanged. However, it is quite complicated to construct potential Lorenz curves from simple power law Pareto functions due to their singular behavior for small and large values of w_m and the necessary integral to compute the cumulated wealth variable. For the more refined shifted power law approximation (7) for the Pareto function, it is still possible to reconstruct the Lorenz curve, using a simple procedure given in (Frahm, Ermann, and Shepelyansky 2026), which gives the formula $w(h) \approx \ln(1 - h/(1 + C_2)) / \ln(C_2/(1 + C_2))$ with $C_2 = (\varepsilon - \mu)/a$. However, this expression does not agree well with neither the real data nor the precise RJE Lorenz curve (3) except for the very rich phase with $h \approx 1$. On the other hand, the precise RJE Lorenz curve typically fits the data Lorenz curve extremely well as can be seen in the above figures. The reason for this is that the Lorenz curve emphasizes the poor phase (large values of K) where the deviations between power law and data are strongest while the RJE theory agrees quite well with the data in this region.

Acknowledgments
The authors acknowledge support from NANOX N° ANR-17-EURE-0009 in the framework of the Programme Investissements d'Avenir (project MTDINA).

References

Baayen, R. 2001. *Word frequency distributions*, volume 1. Kluwer Academic Publishers.

Baudin, K., A. Fusaro, K. Krupa, J. Garnier, S. Rica, G. Millot, and A. Picozzi. 2020. Classical Rayleigh-Jeans condensation of light waves: observation and thermodynamic characterization. *Phys. Rev. Lett.*, 125:244101.

Brysbaert, Marc and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Chancel, L., T. Piketty, E. Saez, and G. et al Zucman. 2022. *World Inequality Report*. World Inequality Lab. Retrieved 2026-09-12. https://wir2022.wid.world.

Connaughton, C., C. Josserand, A. Picozzi, Y. Pomeau, and S. Rica. 2005. Condensation of classical nonlinear waves. *Phys. Rev. Lett.*, 95:263901.

Corpus, Americal National. 2026. *Open data for language research and education*. ANC. Retrieved: 2026-10-17 `https://anc.org/`.

Ermann, L., A. D. Chepelianskii, and D. L. Shepelyansky. 2026. Dynamical thermalization, Rayleigh-Jeans condensate, vortexes and wave collapse in quantum chaos fibers and fluid of light. *J. Phys. A: Math. Theor.*, 59:055702.

Evans, M. R., S. N. Majumdar, I. Pagonabarraga, and E. Trizac. 2010. Condensation transition in polydisperse hard rods. *J. Chem. Phys.*, 132:014102.

Filiasi, M., G. Livan, M. Marsili, M. Peressi, E. Vesselli, and E. Zarinelli. 2014. On the concentration of large deviations for fat tailed distributions, with application to financial data. *J. Stat. Mech.*, 2014:P09030.

Frahm, K. M., L. Ermann, and D. L. Shepelyansky. 2026. Thermodynamic description of wealth inequality in the world. *Complexities*, 2:20.

Frahm, K. M. and D. L. Shepelyansky. 2023. Nonlinear perturbation of Random Matrix Theory. *Phys. Rev. Lett.*, 131:077201.

Frahm, K. M. and D. L. Shepelyansky. 2026a. Dynamical thermalization and turbulence in social stratification models. *CHAOS*, 36:073117.

Frahm, K. M. and D. L. Shepelyansky. 2026b. Wealth thermalization hypothesis and social networks. *J. Stat. Phys.*, 193:64. arXiv:2506.17720 [cond-mat.stat-mech].

Gini, C. 1914. Sulla misura della concentrazione e della variabilit'a dei caratteri. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, 73:1203–1248. English translation in Metron - Int. J. Statistics, 2005, 63, 3–38, Retrieved 2026-09-23. `https://www.dss.uniroma1.it/RePec/mtn/articoli/2005-1-1.pdf`.

Krupa, K., A. Tonello, A. Barthelemy, B. M. Shalaby, A. Bendahmane, G. Millot, and S. Wabnitz. 2016. Observation of geometric parametric instability induced by the periodic spatial self-imaging of multimode waves. *Phys. Rev. Lett.*, 116:183901.

Landau, L. D. and E. M. Lifshitz. 1976. *Statistical mechanics*. Wiley, New York.

Linders, Guido M. and Max M. Louwerse. 2022. Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort. *Psychonomic Bulletin & Review*, 30(1):77–101.

Lorenz, M. O. 1905. Methods of measuring the concentration of wealth. *Quarterly Publications of the American Statistical Association*, 9:209–219. (New Series, No. 70); doi:10.1080/15225437.1905.10503443. Retrieved 2026-09-23.

Moreno-Sánchez, Isabel, Francesc Font-Clos, and Álvaro Corral. 2016. Large-Scale Analysis of Zipf's Law in English Texts. *PLOS ONE*, 11(1):e0147073.

Newman, M. E. J. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323.

OpenAI. 2026. *Understanding and counting tokens*. ChatGPT.

Pareto, V. 1897. Cours d'Économie Politique Professé a l'Université de Lausanne. *Lausanne Univ.*, 1896.

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Piketty, T. 2014. *Capital in the Twenty-First Century*. Belknap Press of. Harvard University Press, Cambridge, MA.

Popescu, I. I. 2009. *Word frequency studies*, volume 64. Walter de Gruyter, Berlin.

Roach, B., P. J. Rajkanikar, N. Goodwin, and J. Harris. 2003. *Social and Economic Inequality*. Economics in Context Initiative, Global Development Policy Center, Boston University, An ECI Teaching Module on Social and Environmental Issues. Retrieved 2026-09-23. https://www.bu.edu/eci/files/2023/05/Inequality-Module-2023.pdf.

Szavits-Nossan, J., M. R. Evans, and S. N. Majumdar. 2014. Constraint driven condensation in large fluctuations of linear statistics. *Phys. Rev. Lett.*, 112:020602.

Wikipedia and contributors. 2026a. *List of sovereign states by wealth inequality*. Wikipedia. Retrieved 2026-10-12. `https://en.wikipedia.org/wiki/List_of_sovereign_states_by_wealth_inequality`.

Wikipedia and contributors. 2026b. *Lorenz curve*. Wikipedia. Retrieved 2026-09-25. `https://en.wikipedia.org/wiki/Lorenz_curve`.

Yakovenko, V. M. and J. B. Rosser. 2009. Colloquium: Statistical mechanics of money, wealth, and income. *Rev. Mod. Phys.*, 81:1703.

Zakharov, V. E., V. S. L'vov, and G. Falkovich. 1992. *Kolmogorov spectra of turbulence*. Springer-Verlag, Berlin.

Zipf, G. 1936. *The Psychobiology of Language*. London: Routledge.

Zipf, G. 1949. *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.