Statistical laws in linguistics

Eduardo G. Altmann[1] and Martin Gerlach[1]

[1]*Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany*

Zipf's law is just one out of many universal laws proposed to describe statistical regularities in language. Here we review and critically discuss how these laws can be statistically interpreted, fitted, and tested (falsified). The modern availability of large databases of written text allows for tests with an unprecedent statistical accuracy and also a characterization of the fluctuations around the typical behavior. We find that fluctuations are usually much larger than expected based on simplifying statistical assumptions (e.g., independence and lack of correlations between observations). These simplifications appear also in usual statistical tests so that the large fluctuations can be erroneously interpreted as a falsification of the law. Instead, here we argue that linguistic laws are only meaningful (falsifiable) if accompanied by a model for which the fluctuations can be computed (e.g., a generative model of the text). The large fluctuations we report show that the constraints imposed by linguistic laws on the creativity process of text generation are not as tight as one could expect.

Contents

I. INTRODUCTION

"…'language in use' cannot be studied without statistics" Gustav Herdan (1964) [1]

In the past 100 years regularities in the frequency of text constituents have been summarized in the form of *linguistic laws*. For instance, Zipf's law states that the frequency f of the r-th most frequent word in a text is inversely proportional to its rank: $f \propto 1/r$ [2]. This and other less famous linguistic laws are one of the main objects of study of *quantitative linguistics* [3–8].

Linguistic laws have both theoretical and practical importance. They provide insights on the mechanisms of text (language, thought) production and are also crucial in applications of statistical natural language processing (e.g., information retrieval). Both the generative and data-analysis views of linguistic laws are increasingly important in modern applications. Data-mining algorithms profit from accurate estimations of the vocabulary size of a collection of texts (corpus), e.g., through Heaps' law discussed in the next section. Methods for the automatic generation of natural language can profit from knowing the linguistic laws underlying usual texts. For instance, linguistic laws may be included as (additional) constraints in the space of possible (Markov generated) texts [9] and can thus be considered as constrains to the creativity of authors.

Besides giving an overview on various examples of linguistic laws (Sec. II), in this paper we focus on their probabilistic interpretation (Sec. III), we discuss different statistical methods of data analysis (Sec. IV), and the possibilities of connecting different laws (Sec. V). The modern availability of large text databases allows for an improved view on linguistic laws that requires a careful discussion of their interpretation. Typically, more data confirms the observations motivating the laws – mostly based on visual inspection – but makes increasingly difficult for the laws to pass statistical tests designed to evaluate their validity. This leads to a seemingly contradictory situation: while the laws allow for an estimation of the general behavior (e.g., they are much better than alternative descriptions), they are strictly-speaking falsified. The aim of this contribution is to present this problem and discuss alternative interpretations of the results. We argue that the statistical analysis of linguistic laws often shows long-range correlations and large (topical) fluctuations. We conclude that null models accounting for these observations are often ignored yet crucial in the tests of the validity of linguistic laws.
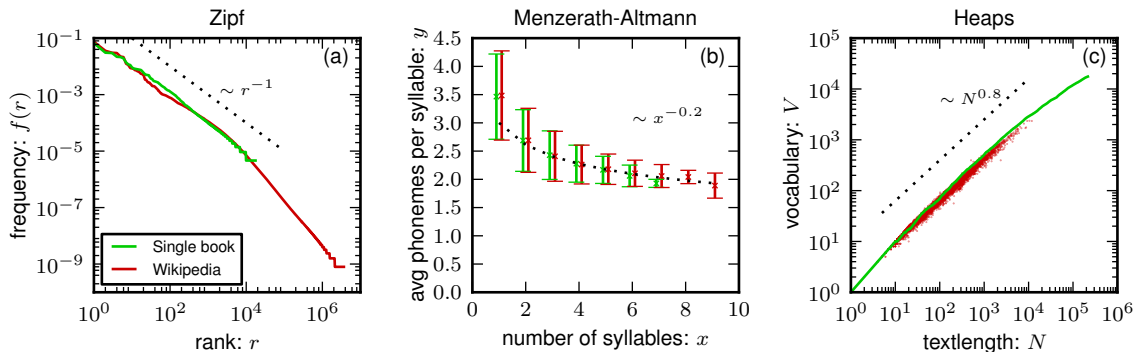
FIG. 1: Examples of linguistic laws: (a) Zipf, (b) Menzerath-Altmann, and (c) Heaps laws. Data from one book (green, Moby Dick by H. Melville) and for the English Wikipedia (red) are shown. Dotted (black) lines are the linguistic laws with arbitrary parameter, chosen for visual comparison (see Appendix for details).

II. EXAMPLES AND OBSERVATIONS

An insightful introduction to Linguistic Laws is given in Ref. [5] by Köhler, who distinguishes between three kinds of laws as follows:

1. "The first kind takes the form of probability distributions, i.e. it makes predictions about the number of units of a given property."

2. "The second kind of law is called the functional type, because these laws link two (or more) variables, i.e. properties."

3. "The third kind of law is the developmental one. Here, a property is related to time." (time may be measured in terms of text length)

We use the term linguistic law to denote quantitative relationships between measurements obtained in a written text or corpus, in contrast to syntactic rules and to phonetic and language-change laws (e.g., Grimm's law). We assume that the laws make statements about individual texts (corpus) and are exact in an appropriate limit (e.g., large corpus)[65]. Each law contains parameters which we denote by Greek letters α, β, γ, and often refer to the frequency $f(q)$ of a quantity q in the text (with $\sum_q f(q) = 1$). Probabilities are denoted by $P(q)$.

Next we discuss in detail one representative example of each of the three types of laws mentioned above: Zipf, Menzerath-Altmann, and Heaps laws, respectively, see Fig. 1.

A. Zipf's law

Zipf's law is the best known linguistic law (see, e.g., Ref. [11] for historical references). In an early and simple formulation, it states that if words (types) are ranked according to their frequency of appearance $r = 1, 2, \ldots, V$, the frequency $f(r)$ of the r-th word (type) scales with the rank as

$$f(r) = \frac{f(1)}{r}, \quad (1)$$

where $f(1)$ is the frequency of the most frequent word. The above expression cannot hold for large r because for any $f(1) > 0$, there is an r^* such that $\sum_{r=1}^{r^*} f(1)/r > 1$. Taking also into account that $f(1)$ may not be the best proportionality factor, a modern version of Zipf's law is

$$f(r) = \frac{\beta_Z}{r^{\alpha_Z}}, \quad (2)$$

with $\alpha_Z \geq 1$, see Fig. 1(a). The analogy with other processes showing fat-tailed distribution motivates the alternative formulation

$$P(f) = \frac{\beta_Z^{\dagger}}{f^{\alpha_Z^{\dagger}}}, \quad (3)$$

where $P(f)$ is the fraction of the total number of words (probability) that have frequency f. Formulations (2) and (3) can be mapped to each other with $\alpha^{\dagger} = 1 + 1/\alpha$ [11, 12, 36].

B. Menzerath-Altmann law

The Menzerath-Altmann law received considerable attention after the works of Gabriel Altmann [4–6, 16]. Menzerath's general (qualitative) statement originating from his observations about phonemes is that "the greater the whole the smaller its parts". The quantitative law intended to describe this observation is [16]

$$y = \alpha_M x^{\beta_M} e^{-\gamma_M x}, \quad (4)$$

where x measures the length of the whole and y the (average) size of the parts. One example [16] is obtained computing for each word w the number of syllables x_w and the number of phonemes z_w. The length of the word (the whole) is measured by the number of syllables x_w, while

TABLE I: List of linguistic laws.

Name of the law	Observables	Functional form	References
Zipf	f: freq. of word w; r: rank of w in f	$f(r) = \beta_Z r^{-\alpha_Z}$	[2, 10–15]
Menzerath-Altmann	x : length of the whole; y : size of the parts	$y = \alpha_M x^{\beta_M} e^{-\gamma_M x}$	[16, 17]
Heaps	V : number of words; N : database size	$V \sim N^{\alpha_H}$	[1, 18–23]
Recurrence	τ : distance between words	$P(\tau) \sim \exp(\alpha\tau)^{\beta}$	[2, 24, 25]
Long-range correlation	$C(\tau)$: autocorrelation at lag τ	$C(\tau) \sim \tau^{-\alpha}$	[26–28]
Entropy Scaling	H : Entropy of text with blocks of size n	$H \sim \alpha n^{\beta} + \gamma n$	[29, 30]
Information content	$I(l)$: Information of word with length l	$I(l) = \alpha + \beta l$	[2, 31]
Taylor's law	σ: standard deviation around the mean μ	$\sigma \sim \mu^{\alpha}$	[23]
Networks	Topology of lexical/semantic networks	various	[32–35]

the length of the parts is measured for each word as the average number of phonemes per syllable $y_w = z_w/x_w$. The comparison to the law is made by averaging y_w over all words w with $x_w = x$, see Fig. 1(b). The ideas of Menzerath-Altmann law and Eq. (4) have been extended and applied to a variety of problems, see Ref. [17] and references therein.

C. Heaps' law

Heaps' law states that the number of different words V (i.e., word types) scales with database size N measured in the total number of words (i.e., word tokens) as [1, 18]

$$V \sim N^{\alpha_H}. \qquad (5)$$

In Fig. 1(c) this relationship is shown in two different representations. For a single book, the value of N is increased from the first word (token) until the end of the book so that $V(N)$ draws a curve. For the English Wikipedia, each article is considered as a separate document for which V and N are computed and shown as dots.

The non-trivial regularities and the similarity between the two disparate databases found for the three cases analyzed in Fig. 1 strongly suggest that the three linguistic laws summarized above capture important properties of the structure of texts. Additional examples of linguistic laws are listed in Tab. I, see also the vast literature in quantitative linguistics [3–6]. The (qualitative) observations reported above motivate us to search for quantitative analysis that match the requirements of applications and the accuracy made possible through the use of large corpora. The natural questions that we would like to address here are: Are these laws true (compatible with the observations)? How to determine their parameters? How much fluctuations around them should be expected (allowed)? Are these laws related to each other? Before addressing these questions we discuss *how should one interpret linguistic laws*.

III. INTERPRETATION OF LINGUISTIC LAWS

In Chap. 26 *Text Laws* of Ref. [3], Hřebíček argues that

"...the notion *law* (in the narrower sense *scientific law*) in linguistics and especially in quantitative linguistics ... need not obtain some special comprehension different from its validity in other sciences. Probably, the best delimitation of this concept can be found in the works by the philosopher of scientific knowledge Karl Raimund Popper..."

This view is also emphasized by Köhler in Ref. [5], who distinguishes *laws* from *rules* and states that a *"significant difference is that rules can be violated - laws (in the scientific sense) cannot."*.

Such a straight-forward identification between linguistic and scientific laws masks the central role played by statistics (and probability theory) in the interpretation of linguistic laws. To see this, first notice that these laws do not directly affect the production of (grammatically and semantically) meaningful sentences because they typically involve scales much larger or shorter than a sentence. It is thus not difficult to be convinced that a creative and persistent daemon [66], trained in the techniques of *constrained writing* [37], can generate understandable and arbitrary long texts which deliberately violate any single law mentioned above. In a strict Popperian sense, a single of such demonic texts would be sufficient to falsify the proposed laws. Linguistic laws are thus different from syntactic rules and require a different interpretation than, e.g., the laws of classical Physics.

The central role of statistics in Quantitative Linguistics was emphasized by its founding father Gustav Herdan:

"The distinction between language laws in the conventional sense and statistical laws of language corresponds closely to that between the classical laws of nature, or the physical universe, and the statistical laws of modern physics." [1]

Altmann, when discussing Menzerath law [16], also emphasizes that "this law is a stochastic one", and Köhler [3]

refers to the concept of stochastic hypothesis. There are at least two instances in which a statistical interpretation should be included:

1. In the statement of the law, e.g., in Zipf's law the probability of finding a word with frequency f decays as $P(f) \sim f^{-\alpha_z^{\dagger}}$.

2. In the interpretation of the law as being *typical* in a collection of texts, e.g., in Heaps' law the vocabulary V of a (typical) text of size N is $V \sim N_H^{\alpha}$.

The demonic texts mentioned above would be considered *untypical* (or highly unlikely). Statistical laws in at least one of these senses are characteristic not only of modern Physics, as pointed out by Herdan, but also of different areas of natural and social sciences: Benford's law predicts the frequency of the first digit of numbers appearing in a corpus and the Gutenberg-Richter law determines the frequency of earthquakes of a given magnitude. The analysis of these laws, including possible refutations, have to be done through rigorous statistical methods, the subject of the next section. Important aspects of linguistic laws not discussed in detail in this Chapter include: (i) the universality and variability of parameters of linguistic laws (e.g., across different languages [20, 21, 35, 38, 40], as a function of size [39] and degree of mixture of the corpus [41], styles [27], and age of speakers [42]); and (ii) the relevance and origins of the laws. This second point was intensively debated for Zipf's law [8, 14, 43], with quantitative approaches based on stochastic processes – e.g., the Monkey typewriter model [12]), rich-get-richer mechanisms [10–12, 15, 21] – and on optimization principles – e.g., between speaker and hearer [2, 44, 45] or entropy maximization [46, 47].

IV. STATISTICAL ANALYSIS

In Sec II we argued in favor of linguistic laws by showing a graphical representation of the data (Fig. 1). The widespread availability of large databases and the applications of linguistic laws require and allow for a more rigorous statistical analysis of the results. To this end we assume the linguistic law can be translated in a precise mathematical statement about a curve or distribution. This distribution has a set of parameters and observations. Legitimate questions to be addressed are:

(1) **Fitting.** What are the best parameters of the law to describe a given data?

(2) **Model Comparison.** Is the law better than an alternative one?

(3) **Validity.** Is the law compatible with the observations?

These points are representative of statistical analysis performed more generally and should preceed any more fundamental discussion on the origin and importance of a specific law. Below we discuss in more details how each of the three points listed above has been and can be addressed in the case of linguistic laws.

A. Graphical approaches

Visual inspection and graphical approaches were the first type of analysis of linguistic laws and are still widely used. One simple and still very popular fitting approach is least squares (minimize the squared distance between data and models). Often this is done in combination with a transformation of variables that maps the law into a straight line (e.g., using logarithmic scales in the axis or taking the logarithm of the independent and dependent variable in the Zipf's and Heaps' laws). These transformations are important to visually detect patterns and are valuable part of any data analysis. However, they are not appropriate for a quantitative analysis of the data. The problem of fitting straight lines in log-log scale is that least-square fitting assumes an uncertainty (fluctuation) on each point that is independent, Gaussian distributed, and equal in size for all fitted points. These assumptions are usually not justified (see, e.g., Refs. [48, 49] for the case of fitting power-law distributions), while at the same time the uncertainties are modified through the transformation of variables (such as using the log scale). Furthermore, quantifying the goodness-of-fit by using the correlation-coefficient R^2 in these scales is insufficient to evaluate the validity of a given law. A high quality of the fit indicates a high correlation between data and model, but is unable to assign a probability for observations and thus it is not suited for a rigorous test of the law.

B. Likelihood methods

A central quantity in the statistical analysis of data is the likelihood $\mathcal{L}(\mathbf{x}; \vec{\alpha})$ that the data $\mathbf{x}$ was generated by the model (with a set of parameters α).

(1) Fitting When fitting a model (law) to data the approach is to tacitly assume its validity and then search for the best parameters to account for the data. It corresponds to a search in the (multidimensional) parameter space α of the law for the value $\hat{\alpha}$ that maximize $\mathcal{L}$.

In laws of the first kind – as listed in Sec. II – the quantity to be estimated from data is a probability distribution $P(\mathbf{x}; \alpha)$. The probability of an observation x_j is thus given by $P(x_j; \alpha)$. Assuming that all J observations are independent, the best parameter estimates $\hat{\alpha}$ are the values of α that maximize the log-likelihood

$$\log_e \mathcal{L} = \log P(x_1, x_2, \ldots x_J; \alpha) = \sum_{j=1}^{J} \log \tilde{P}(x_j; \alpha), \quad (6)$$

The need for Maximum Likelihood (ML) methods when fitting power-law distributions (such as Zipf's law) has been emphasized in many recent publications. We refer

to the review article Ref. [50] and references therein for more details, and to Ref. [51] for fitting truncated distributions (e.g., due to cut-offs).

In laws of the second and third kind – as listed in Sec. II – the quantity to be described y is a function $y = y_g(\mathbf{x}; \vec{\alpha})$. Fitting requires assumptions regarding the possible fluctuations in $y(\mathbf{x})$. One possibility is to assume Gaussian fluctuations with a standard deviation $\sigma(\mathbf{x})$. In this case, assuming again that the observations $\mathbf{x}$ are independent [52]

$$\log_e \mathcal{L} \sim -\sum_j \left(\frac{y(\mathbf{x}_j) - y_g(\mathbf{x}_j)}{\sigma(\mathbf{x}_j)} \right)^2, \qquad (7)$$

where the sum is over all observations j. The best estimated parameters $\hat{\vec{\alpha}}$ are obtained minimizing $\chi^2 = \sum_j (\frac{y(\mathbf{x}_j) - y_g(\mathbf{x}_j)}{\sigma(\mathbf{x}_j)})^2$, which maximizes (7). Least-squares fitting is equivalent to Maximum-Likelihood fitting only in the case of constant σ (independent of $\mathbf{x}$) [52].

(2) Model Comparison The comparison between two different functional forms of the law ($m1$ and $m2$) is done comparing their likelihoods, e.g., through the log-likelihood ratio $\log_e \mathcal{L}_{m1}/\mathcal{L}_{m2}$ [53]. A value $\log_e \mathcal{L}_{m1}/\mathcal{L}_{m2} = 1$ (-1) means it is $e^1 = 2.718\ldots$ times more (less) likely that the data was generated by function $m1$ than by function $m2$. If the two models have a different number of parameters, one can penalize the model with higher number of parameters using, e.g. the Akaike information criterion [54], or calculate the Bayes factor by averaging (in the space of parameters) over the full posterior distribution [55].

(3) Validity The probabilistic nature of linguistic laws requires statistical tests. One possible approach is to compute the probability (p-value) of having observations similar to the data from a null model compatible with the linguistic law (which is assumed to be true). A low p-value is a strong indication that the null model is violated and may be used to refute the law (e.g., if p-value< 0.01). Defining a measure of distance D between the data and the model, the p-value can be computed as the fraction of finite-size realizations of the model (assuming it is true) that show a distance $D' > D$. In the case of probability distributions – linguistic laws of the first kind in Sec. II – the distance D is usually taken to be the Kolmogorov-Smirnov distance (the largest distance between the empirical and fitted cumulative distributions). In the case of simple function – linguistic laws of the second and third kind in Sec. II – one can consider $D = \chi^2$.

Application: Menzerath-Altmann law. We applied the likelihood analysis summarized above to the case of the Menzerath-Altmann law introduced in Sec. II. Our critical assumption here is that the law is intended to describe the average number of phonemes per syllable, y, computed over many words w with the same number of syllables x. Assuming the words are independent of each other, the uncertainty in $y(x)$ is thus the standard error of the mean given by $\sigma_y(x) = \sigma_w(x)/\sqrt{N(x)}$, where $\sigma_w(x)$

is the (empirical) standard deviation over the words with x-syllables and $N(x)$ is the number of such words.

In Fig. 2 and Tab. II we report the fitting, model comparison, and validity analysis for the Menzerath-Altmann law – Eq. (4) – and three alternative functions with the same number of parameters. The results show that two of the three alternative functions (shifted power law and stretched exponential) provide a better description than the proposed law, which we can safely consider to be incompatible with the data (p-value$< 10^{-5}$). Considering the two databases, the stretched exponential distribution provides the best description and is not refuted. These results depend strongly on the procedure used to identify phonemes and syllables (see Appendix).

C. Critical discussion

In the next paragraphs we critically discuss the likelihood approach considering the example of Zipf's law.

Fitting as model comparison. In the beginning of this section we started with the distinction between fitting (i.e., fixing free parameters) and model comparison (i.e. choosing between different models). This division is didactic [50], but from a formal point of view both procedures correspond to hypothesis testing because the free parameters of one fitting model can be thought as a *continuous* parameterization of different models which should be compared and selected according to their likelihood [56]. This means that the points mentioned below apply equally well to both fitting and hypothesis testing (and, in most cases, also to test the validity of the models).

Fitting ranks. Power-law fitting recipes [50]– employed for linguistic [43] and non-linguistic problems – suggest to fit Zipf's law using the distribution of frequencies $P(f)$ given in Eq. (3). However, it is also possible to use the rank formulation (2) [21] because the frequency of ranks $f(r)$ is normalized $\sum_r f(r) = 1$ and can thus be interpreted as a probability distribution. However, a drawback in fitting $f(r)$ is that the process of ranking introduces a bias in the estimator [57, 58]. For instance, consider a finite sample from a true Zipf distribution containing ranks $r = 1, \ldots, \infty$. Because of statistical fluctuations, some of the rankings will be inverted (or absent) so that when we rank the words according to the observations obtain ranks different from the ones drawn. This effect introduces bias in our estimation of the parameters (overestimating the quality of the fit). The words affected by this bias are the ones with largest ranks, which contribute very little to the estimation of the parameters of Zipf's law (as discussed below). Therefore, we expect that this bias to become negligible for sufficiently large sample sizes.

Representation matters. Equivalent formulations of the linguistic laws lead to *different* statistical analysis and conclusions [57, 58]. One example of this point is the use of transformations before the fitting is performed, such as
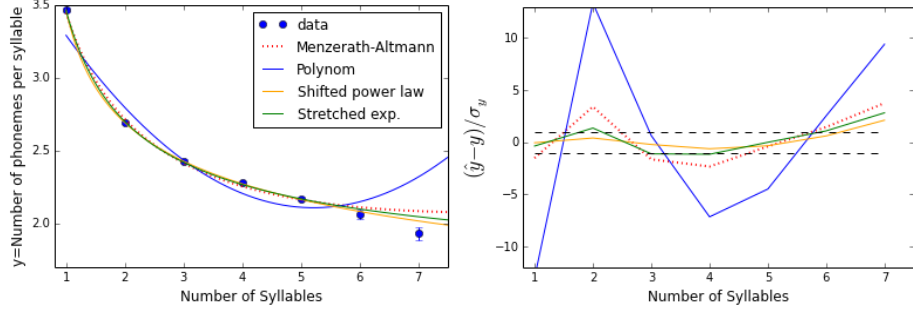
FIG. 2: Model comparison for the Menzerath-Altmann law. Data points are the average over all word (types) in a book (Moby Dick by H. Melville, as in Fig. 1). The curves show the best fits of the four alternative curves, as reported in Tab. II. Left plot: the data in the original scales, as in Fig. 1. Right plot: the distance between the curves and the points $(\hat{y} - y)/\sigma_y$, where the uncertainty σ_y is the standard error of the mean.

TABLE II: Likelihood analysis of the Menzerath-Altmann law and three alternative functions. The parameters $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ that maximize the likelihood $\mathcal{L}_m$ of model m were computed using the downhill simplex algorithm (using the Python library scipy). The reported p-value corresponds to the fraction of random realizations with a χ^2 larger than the observed χ^2. In each realization, one point $y^{\dagger}(x)$ was generated at each x from a Gaussian distribution centered at the model prediction $y_m(x)$ with a standard deviation $\sigma_y(x)$ given by the data. The best models and the results with $p > 0.01$ are shown in bold face.

	Menzerath-Altmann (MA) $\alpha x^{\beta} \exp{(-\gamma x)}$	Shifted power law $\alpha(x+\beta)^{\gamma}$	Stretched exp. $\alpha \exp{(\beta x)^{\gamma}}$	Polynom $\alpha + \beta x + \gamma x^2$
Results for one book (Moby Dick by H. Melville)				
$(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$	$(3.3, -0.12, -0.051)$	$(2.8, -0.65, -0.19)$	$(1.5, 1.4, -0.51)$	$(3.9, -0.69, 0.066)$
$\log_e \mathcal{L}_m/\mathcal{L}_{MA}$	0	**33**	25	-475
p-value	$< 10^{-5}$	**0.611**	**0.064**	$< 10^{-5}$
Results for English Wikipedia				
$(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$	$(3.2, -0.45, -0.064)$	$(2.8, -0.70, -0.18)$	$(1.6, 1.5, -0.60)$	$(3.8, -0.64, 0.061)$
$\log_e \mathcal{L}_i/\mathcal{L}_{MA}$	0	11	**49**	-1898
p-value	$< 10^{-5}$	2×10^{-5}	**0.93**	$< 10^{-5}$

the linear fit of Zipf's law in logarithmic scale discussed in Sec. IV A. The variables used to represent the linguistic law are also crucial when likelihood methods are used, as discussed above for the case of Zipf's law represented in $f(r)$ or $P(f)$. While asymptotically these formulations are equivalent, the likelihood computed in both cases is different. In the likelihood of $P(f)$, an observation corresponds to the frequency of a word *type*. This means that the most frequent words in the database count the same as words appearing only once (the hapax-legomenan). In practice, the part of the distribution that matters the most in the fitting (and in the likelihood) are the words with very few counts, which contribute very little to the total text. In the likelihood of $f(r)$ the observational quantity is the rank r of each occurrence of the word meaning that each word *token* counts the same. This means that the frequent words contribute more and the fitting of $f(r)$ is robust against rare words. Linear regression in log-log plot counts every point in the plot the same and, since there are more points for large r, low-frequency words dominate the fit. Using logarithmic binning, as suggested in Ref. [48], equalize the importance of words across $log(r)$. In summary, while fitting

a straight line in log-log scale using logarithmic binning gives the same value for words across the full spectrum (in a logarithmic scale), the statistical rigorous methods of Maximum Likelihood will be dominated either by the most frequent (in case of fitting in $f(r)$) or least frequent (in case of fitting in $P(f)$) words.

Beyond Zipf's law, the reasoning above shows that even if asymptotically (i.e. infinite data) different formulations of a law are equivalent, the representation in which we test the law matters because it assumes a sampling process of the data. This in turn leads to different results when applied to finite and often noisy data and has to be taken into account when interpreting the results.

Application: Fitting Zipf's law. In Fig. 3 and Tab. III we compare the different fitting methods described above. The visual agreement between data and the fitted curves reflects the different weights given by the methods to different regions of the distribution as discussed above (high-frequency words for $f(r)$ and low-frequency words for the other two cases). Not surprisingly, Tab. III shows that the estimated exponent α varies from method to method. This variation is larger than the variation
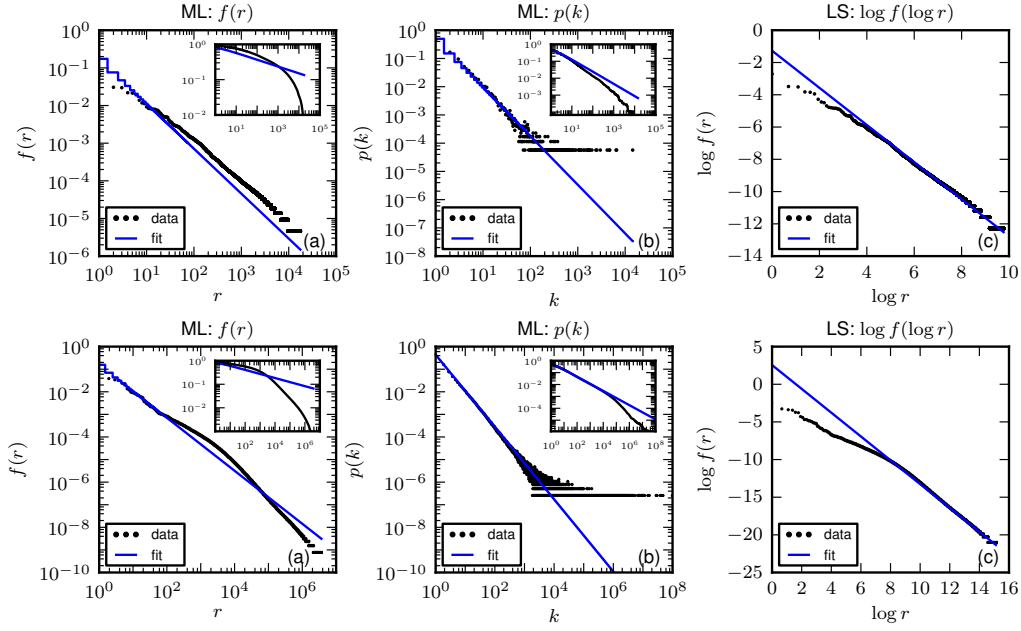
FIG. 3: Comparison of the Zipf's law obtained using three different fitting methods. Results are shown for one book (Moby Dick by H. Melville, top row) and for the complete English Wikipedia (bottom row). Data is fitted using Maximum Likelihood (ML) in the frequency rank $f(r)$ (left), ML in the frequency distribution $P(f) \sim p(k)$ (center), and least square (LS) in the $\log f$ vs. $\log r$ representation (right). Insets show the cumulative distributions. See Tab. III for the parameter $\hat{\alpha}$ and significance test of the fits. In the plot in the center, instead of $P(f)$ we use the distribution the unnormalized frequency $p(k)$ (i.e., k is the number of occurrences of a word in the database). For ML fits, we used a discrete power law in $f(r)$ and $p(k)$ with support in $[1, \infty)$ (exponents were obtained using the downhill simplex algorithm of the Python library scipy). For the LS fit, we used a continuous straight line in $\log f(\log r)$ [52].

TABLE III: Zipf's law exponent obtained using different fitting methods, see Fig. 3. In the fit of $P(f)$ (frequency) we obtain $\hat{\alpha}_Z^\dagger$ and calculate $\hat{\alpha}_Z = 1/(1 - \hat{\alpha}_Z^\dagger)$, see Eqs. (2)-(3). English version of the books were obtained from the Project Gutenberg, see Appendix.

| | Rank: $f(r)$ | | Frequency: $P(f)$ | | Linear: $\log f(\log r)$ | |
Book	$\hat{\alpha}_Z$	p-value	$\hat{\alpha}_Z$	p-value	$\hat{\alpha}_Z$	R^2
Alice's Adventures in Wonderland (L. Carroll)	1.22	$< 10^{-4}$	1.46	$< 10^{-4}$	1.21	0.97
The Voyage Of The Beagle (C. Darwin)	1.20	$< 10^{-4}$	1.59	$< 10^{-4}$	1.29	0.97
The Jungle (U. Sinclair)	1.21	$< 10^{-4}$	1.45	$< 10^{-4}$	1.22	0.98
Life On The Mississippi (M. Twain)	1.20	$< 10^{-4}$	1.38	$< 10^{-4}$	1.16	0.98
Moby Dick; or The Whale (H. Melville)	1.19	$< 10^{-4}$	1.38	$< 10^{-4}$	1.15	0.98
Pride and Prejudice (J. Austen)	1.21	$< 10^{-4}$	1.66	$< 10^{-4}$	1.35	0.98
Don Quixote (M. Cervantes)	1.21	$< 10^{-4}$	1.70	$< 10^{-4}$	1.38	0.98
The Adventures of Tom Sawyer (M. Twain)	1.21	$< 10^{-4}$	1.29	$< 10^{-4}$	1.12	0.98
Ulysses (J. Joyce)	1.18	$< 10^{-4}$	1.15	$< 10^{-4}$	1.03	0.97
War and Peace (L. Tolstoy)	1.20	$< 10^{-4}$	1.84	$< 10^{-4}$	1.44	0.97
English Wikipedia	1.17	$< 10^{-4}$	1.60	$< 10^{-4}$	1.58	0.99

across different databases. Large values of R^2 computed in the linear fit, usually interpreted as an indication of good fitting, are observed also when the p-value are very low.

Correlated samples The failure of passing significance tests for increasing data size is not surprising because any small deviation from the null model becomes statistically significant. A possible conclusion emerging from these

analysis is that power-law distributions are not as widely valid as previously claimed (see also Refs. [50, 59]), but often are better than alternative (simple) descriptions (see our previous publication Ref. [21] in which we consider two-parameter generalizations of Zipf's law). The main criticism we have on this widely used framework of analysis is that it ignores the presence of correlations in the data: the computation of the likelihood in Eq. (6)
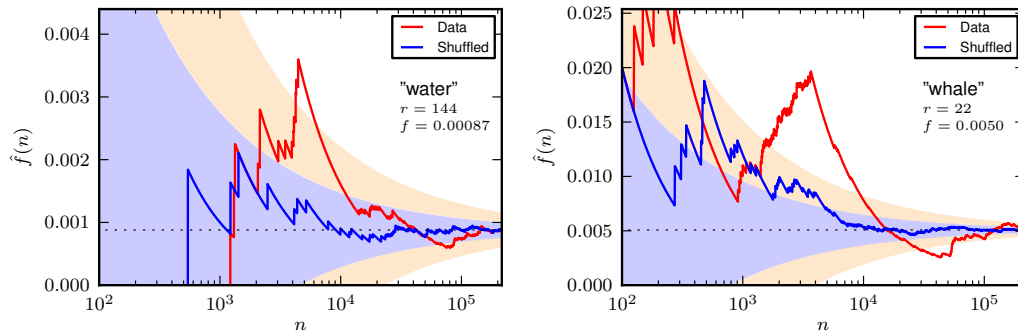
FIG. 4: Estimation of the frequency of a word in the first n word tokens of a book (Moby Dick by H. Melville). The red curve corresponds to the actual observation (word "water" in the left and word "whale" in the right) and the blue curve to the curve measured in a version of the book in which all word tokens were randomly shuffled. The shaded regions show the expected fluctuations ($\pm 2\sigma$) assuming that the probability of using the word is given by the frequency of the word in the whole book ($f(n = N)$ and that: (i) usage is random (blue region) – see also Ref. [7] or (ii) the time between successive usages of the word is drawn randomly from a stretched exponential distribution with exponent $\beta = 0.5$, as proposed in Ref. [24].

assumes independent observations. Furthermore, this assumption leads to an underestimation of the expected fluctuations (e.g. KS-distance) in the calculation of the p-value when assessing the validity of the law. It is thus unclear in which extent a negative result in the validity test (e.g., p-value$\ll 0.01$) is due to a failure of the proposed law or, instead, is due to the violation of the hypothesis of *independent* sampling. This hypothesis is known to be violated in texts [7, 25]: the sequence of words and letters are obviously related to each other. In Fig. 4 we show that these correlations affect the estimation of the frequency of individual words, which show fluctuations much larger than those expected not only based on the independent random usage of words (Poisson or bag of word models) but also in a null model in which burstiness is included [24, 25]. Altogether, this shows that the independence assumption – used to write the likelihood (6) – is strongly violated and affects both the analysis based on $f(r)$ (correlation throughout a book) and $P(f)$ (f can be thought as a finite size estimation as the ones shown in Fig. 4).

One approach to take into account correlations is to estimate a time for which two observations are independent, and then consider observations only after this time (a smaller effective sample size). Alternative approaches considered statistical tests for specific classes of stochastic processes (correlated in time) [60] or based on estimations of the correlation coming from the data [61]. The application of these methods to linguistic laws is not straightforward because these methods fail in cases in which no characteristic correlation time exist. Books show such long-range correlations [27], also in the position of individual words in books [26, 28], in agreement with the observations reported in Fig. 4. More generally, correlations lead to a slower convergence to asymptotic values and it is thus possible to create processes of text generation that comply to a linguistic law asymptotically but that (in finite samples) violate statistical tests

based on independent sampling. The problem affects also model comparison and fitting because these problems are also based on the likelihood (in these cases, correlation affects all models and therefore it is unclear the extent in which it impacts the choice of the best model).

V. RELATION BETWEEN LAWS

In view of the different laws proposed to describe text properties, a natural question is the relationship between them (e.g., whether one law can be derived from another or whether there are generative processes that account for more than one law simultaneously). For instance, Ref. [28] clarifies how the long-range correlation of texts is related to the skewed distribution of recurrence time between words [2, 24, 25] (a consequence of burstiness [8, 62]). Another well-known relation is the connection between Heaps' law and Zipf's law [19, 21–23, 39] (see Refs. [23, 29, 30, 62] for other examples). Here again the importance of fluctuations and an underlying null model is often neglected.

The need for a null model is evident if we consider a text in which all possible words appear once in the very beginning of the text, violating Heaps' law, even though their frequency over the full text is still compatible with Zipf's law. A typical null model is to consider that every word is used independently from the others with a probability equal to its global frequency. This probability is usually taken to be constant throughout the text (Poisson process), but alternative formulations considering time-dependent frequencies lead to similar results. For this generative model, Zipf's law (2) leads to a Heaps' law (5) with parameters $\alpha_H = 1/\alpha_Z$ [21]. Similar null models are implicitly or explicitly assumed in different derivations [19, 21–23, 39].

Figure 5 shows that the connection between Zipf's and Heaps' law using the independent usage of words fails to
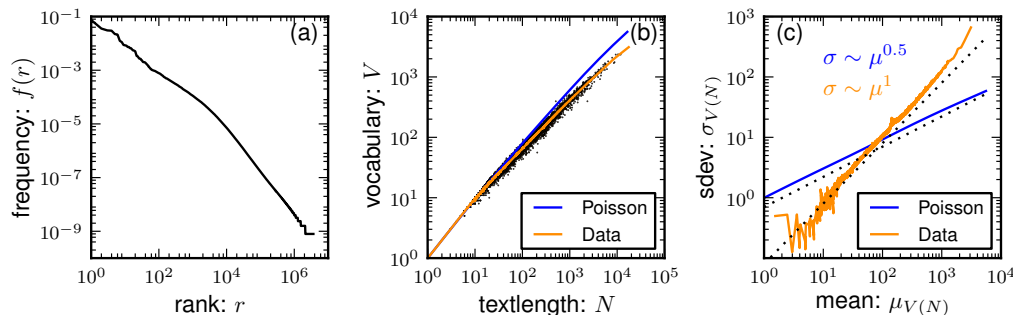
FIG. 5: Relation between Zipf's law and Heaps' law in the English Wikipedia. Fixing the rank-frequency distribution of the complete English Wikipedia – shown in panel (a) – and assuming each word to follow a Poisson process (i.e., to be used randomly) with fixed frequency $f(r)$, one obtains the blue curve for the Heaps' law in (b). Considering each Wikipedia article separately – as shown by black dots in (b) – we estimate in a moving window centered in N the average $\mu_{V(N)}$ and standard deviation $\sigma_{V(N)}$ over all articles in the window. The dependence of $\mu_V(N)$ on N is shown in (b) by a solid line. The dependence of $\sigma_V(N)$ on $\mu_V(N)$ is shown in (c) and reveals a different scaling than the one predicted by the Poisson model. Figure adapted from Ref. [23].

reproduce the fluctuations observed in data. In particular, the fluctuations around the average vocabulary size V predicted from Heaps' law scales linearly with N, and not as $\sqrt{N}$ as predicted by the independence assumption (through the central limit theorem). In Ref. [23] we have shown that this scaling – also known as Taylor's law [63] – is a result of correlations in the usage among different words induced by the existence of topical structures inside and across books.

VI. DISCUSSION

It is common to find claims that a particular linguistic law is valid in a language or corpus. A closer inspection for the statistical support of these claims is often disappointing. In this Chapter we performed a critical discussion of linguistic laws, the sense in which they can be considered valid, and the extent in which the evidence support its validity. We argued that linguistic laws have to be interpreted in a statistical sense. Therefore, model selection (also fitting) and the compatibility to data have to be performed computing statistical tests based on the likelihood (plausibility) of the observations. The statistical analysis is far from being free of choices, both in terms of the methods employed and also about additional assumptions not contained in the original law, as discussed below. The analysis we presented above is intended to show that these choices matter and should be carefully discussed. The picture that emerges from the straight applications of the statistical tests above is that: (i) the linguistic laws are often the best simple description of the data, but (ii) the data is not generated according to it so that in a strict sense the validity of the law is falsified. This interpretation suggests that linguistic laws are useful and capture some of the ingredients seen in language, but are unable to describe observations in full detail even in the limit of large texts (possibly because of

the existence of additional processes ignored by the law).

The main limitation of the methods we described, and thus of the conclusions summarized above, is that they were based not only on the statement of the law but also on the hypothesis that observations are independent and identically distributed. This hypothesis is known to be violated in almost all observations of written language. It is thus unclear in which extent the rejection of the null model (small p-value) can be considered a falsification of the linguistic law. On the one hand, this reasoning shows the limitation of the statistical methods and the necessity to apply and develop tests able to deal with (long-range) correlated data. On the other hand, it shows that the usual statements of linguistic laws are incomplete because they cannot be properly tested. A meaningful formulation of a linguistic law allows for the computation of the likelihood of the observations, e.g., it should be accompanied by a prediction of the fluctuations, a generative model for the relevant variables, or, ultimately, a model for the generation of texts. Such models are usually interpreted as an explanation of the origin of the laws [11, 12, 14] and are absent from the statement of the linguistic laws, despite the fact that Herdan already drew attention to this point [1]: *"The quantities which we call statistical laws being only expectations, they are subject to random fluctuations whose extent must be regarded as part of the statistical law."* In the same sense that a scientific law cannot be judged separated from a theory, linguistic laws are only fully defined once a generative process is given. The existence of long-range correlations, burstiness, and topical variations lead to strong fluctuations in the estimations of observables in texts, including the quantities described by linguistic laws.

Our findings have consequences to applications in information retrieval and text generation. For instance, our results show that strong fluctuations around specific laws are observed and that results obtained using the independence assumption (e.g., bag-of-words models) have

a limited applicability. Therefore, statistical laws should not be imposed too strictly in the generation of artificial texts or in the analysis of unknown databases. Large fluctuations are as much a characteristic of language as the laws themselves and therefore the creativity in the generation of texts is much larger than the one obtained if laws are imposed as strict constraints.

Finally we would like to mention that our conclusions apply also to other statistical laws beyond linguistic. Invariably, the increase of data size leads to a rejection of null-models, e.g. many recent works emphasize that claims of power-law distributions do not survive rigorous statistical tests [13, 50, 59]. However, the statistical tests employed in these references, and in most likelihood-based analysis, rely on the independence assumption of the observations (known to be violated in many of the treated cases). Nevertheless, we are not aware that this point has been critically discussed in the large number of publications on power-law fitting. The crucial role of mechanistic models in the fitting and statistical analysis of scaling laws was emphasized in Ref. [64] for urban-economic data.

Appendix

The books listed in Tab. III were obtained from Project Gutenberg (`http://www.gutenberg.org`). The books and data filtering are the same as the ones used in Ref. [28] (see the Supplementary information of that paper for further details). We removed capitalization and all symbols except the letters "a-z", the number "0-9", the apostrophe, and the blank space. A string of symbols between two consecutive blank spaces was considered to be a word.

The English Wikipedia data was obtained from Wikimedia dumps (`http://dumps.wikimedia.org/`). The filtering was the same as the one used in Ref. [23], in which we removed capitalization and kept only those words (i.e. sequences of symbols separated by blank space) which consisted exclusively of the letters "a-z" and the apostrophe.

The computation of Menzerath-Altmann law appearing in Figs. 1, 2, and Tab. II was done starting from the unique words (word type) in the database discussed in the previous paragraphs. For each word w we applied the following steps:

1. Lemmatize using the WordNetLemmatizer (`http://wordnet.princeton.edu` in the NLTK Python package `http://www.nltk.org/`) .

2. Count the number of syllables x_w based on the *Moby Hyphenation List by Grady Ward*, available at `http://www.gutenberg.org/ebooks/3204`

3. Count the number of phonemes z_w based on *The CMU Pronouncing Dictionary*, version 0.7b available at `www.speech.cs.cmu.edu/cgi-bin/cmudict`

For the book *Moby Dick* by H. Melville, this procedure allowed to compute x_w and z_w for $11,595$ words, 66% of the total number of words (before lemmatization). For the Wikipedia, we obtain $60,749$ words, 1.7% of the total number. The low success in Wikipedia is due to the size of the database (large number of rare words) and the results depend more strongly on the procedure described above than on the database itself.

[1] G. Herdan, Quantitative Linguistics (Butterworth Press, Oxford, 1964)

[2] G. K. Zipf, The Psycho-Biology of Language (Routledge, London, 1936). *Id.*, Human behavior and the principle of least effort (Addison-Wesley Press, Oxford, 1949).

[3] R. Köhler, G. Altmann, and R. G. Piotrowski (Eds.) Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An international Handbook. (=HSK27) (de Gruyter, Berlin, 2005).

[4] R. Köhler, G. Altmann, and P. Grzybek (Eds), Quantitative Linguistics, De Gruyer Mouton, `www.degruyter.com/view/serial/35295` Cited 6 Feb 2015.

[5] Glottopedia: the free Encyclopedia of Linguistics, `http://www.glottopedia.org/index.php/Laws` Cited 17 Dec 2014.

[6] Enciclopedia entry: *Laws in Quantitative linguistics*, `http://lql.uni-trier.de` Cited 3 Dec 2014

[7] R. Harald Baayen, Word Frequency Distributions, Kluwer Academic Pub., Dordrecht (2001).

[8] D. H. Zanette, Statistical Patterns in Written Language, arXiv:1412.3336 (2014).

[9] G. Barbieri, F. Pachet, P. Roy, and M. Degli Esposti, Markov Constraints for Generating Lyrics with Style, 20th European Conference on Artificial Inteligence – ECAI (IOS Press, Amsterdam, 2012)

[10] H.A. Simon, On a Class of Skew Distribution Functions, Biometrika 42, 425 (1955).

[11] M. Mitzenmacher, A Brief History of Generative Models for Power Law and Lognormal Distributions, Internet Mathematics 1, 226 (2004).

[12] M.E.J. Newman, Power Laws, Pareto Distributions and Zipfs law, Contemp. Phys. 46, 323 (2005).

[13] W. Li, Zipf's Law Everywhere, Glottometrics 5, 14 (2002).

[14] S.T. Piantadosi, Zipfs word frequency law in natural language: A critical review and future directions. Psychonomic bulletin & review, 21,1112 (2014).

[15] D. Zanette and M. Montemurro, Dynamics of Text Generation with Realistic Zipf's Distribution, J. Quant. Linguist. 12, 29 (2005).

[16] G. Altmann, Prolegomena to Menzerath's law, Glot-tometrika 2, 1 (1980).

[17] I. Cramer. The Parameters of the Altmann-Menzerath Law. J. Quant. Linguist. 12, 41 (2005).

[18] L. Egghe, Untangling Herdan's Law and Heaps' Law : Mathematical and Informetric Arguments, J. Am. Soc. Inf. Sci. Tec. 58,702 (2007).

[19] L. Lü, Z.-K. Zhang, T. Zhou, Zipf's Law leads to Heaps' law: Analyzing Their Relation in Finite-Size Systems, PLOS ONE 5, e14139 (2010).

[20] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc, Languages cool as they expand: Allometric scaling and the decreasing need for new words, Scientific Reports 2, 943 (2012)

[21] M. Gerlach and E. G. Altmann, Stochastic model for the vocabulary growth in natural languages, Physical Review X 3, 021006 (2013)

[22] F. Font-Clos, G. Boleda, A. Corral, A scaling law beyond Zipf's law and its relation to Heaps' law, New Journal of Physics, 2013.

[23] M Gerlach and EG Altmann, Scaling laws and fluctuations in the statistics of word frequencies, New Journal of Physics 16, 113010 (2014)

[24] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter. Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words. PlosOne. 4, e7678 (2009).

[25] J. Lijffijt, P. Papapetrou, K. Puolamäki, H. Mannila, Analyzing Word Frequencies in Large Text Corpora Using Inter-arrival Times and Bootstrapping, Chapter in Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science 6912, 341 (2011).

[26] F.J. Damerau, B. Mandelbrot, Tests of the degree of word clustering in samples of written English, Linguistics 102, pp. 58-72 (1973).

[27] A. Schenkel, J. Zhang, and Y. Zhang, Long range correlation in human writings, Fract. 1, 47 (1993).

[28] E. G. Altmann, G. Cristadoro, and M. Degli Esposti. On the origin of long-range correlations in texts. PNAS. 109, 11582 (2012).

[29] W. Ebeling and T. Pschel Europhys. Lett. 26 24 (1994).

[30] L. Debowski. On Hilberg's law and its links with Guiraud's law. J. Quant. Linguist.. 13, 81-109 (2006).

[31] S. T. Piantadosi, H. Tily, and E. Gibson, Word lengths are optimized for efficient communication, PNAS 108, 3526 (2011).

[32] R. V. Solé, B. Corominas-Murtra, S. Valverde, and L. Steels, Language Networks: their structure, function and evolution, Complexity 15, 20 (2009)

[33] M. Choudhury and A. Mukherjee: The Structure and Dynamics of Linguistic Networks, in Dynamics On and Of Complex Networks (Springer 2009) pp 145-166.

[34] A. Baronchelli, R. Ferrer-i-Cancho, R. Pastor-Satorras, N. Chater, M. H. Christiansen, Networks in Cognitive Science, Trends in Cognitive Sciences 17, 348 (2013).

[35] J. Cong and H. Liu, Approaching human language with complex networks, Physics of Life Reviews 11, 598 (2014).

[36] B. Mandelbrot, 1961 On the Theory of Word Frequencies and on Related Markovian Models of Discourse *Structure of Language and Its Mathematical Aspects: Proceedings of Symposia in Applied Mathematics Vol. XII* (Providence, RI: American Mathematical Society)

[37] Constrained Writing, in *Wikipedia*, `http://en.wikipedia.org/wiki/Constrained_writing` Cited 3 Dec 2014.

[38] D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira Jr, LdF Costa, Probing the Statistical Properties of Unknown Texts: Application to the Voynich Manuscript, PLOS ONE 8, e67310 (2013).

[39] S. Bernhardsson, L. E. C. da Rocha, P. Minnhagen, Size-dependent word frequencies and translational invariance of books, Physica A 389, 330 (2010).

[40] G. Febres, K. Jaffé, and C. Gershenson, Complexity Measurement of Natural and Artificial Languages, Complexity doi:10.1002/cplx.21529 (2014)

[41] J.R. Williams, J.P. Bagrow, C.M. Danforth, and P.S. Dodds, Text mixing shapes the anatomy of rank-frequency distributions: a modern zipfian mechanics for natural language, arXiv:1409.3870 (2014)

[42] J. Baixeries, B. Elvevag, R. Ferrer-i-Cancho, The Evolution of the Exponent of Zipf's Law in Language Ontogeny, PLOS ONE 8, e53227 (2013).

[43] G. Jäger, Power Laws and other heavy-tailed distribution in linguistic typology, Adv. Comp. Syst. 15, 1150019 (2012).

[44] B. Corominas-Murtra, J. Fortuny, and R.V. Sol, Emergence of Zipfs Law in the Evolution of Communication, Phys. Rev. E 83, 036115 (2011).

[45] R. Ferrer-i-Cancho, Optimization models of natural communication, arXiv:1412.2486 (2014).

[46] M. Marsili, I. Mastromatteo, and Y. Roudi, On sampling and modeling complex systems, J. of Stat. Mech. 2013, P09003 (2013).

[47] J. Peterson, P. D. Dixit, and K. Dill, A maximum entropy framework for nonexponential distributions, PNAS, 110, 20380 (2013)

[48] M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution. European Journal of Physics B. 41, 255-258 (2004).

[49] H. Bauke. Parameter estimation for power-law distributions by maximum likelihood methods. European Journal of Physics B. 58, 167-173 (2007).

[50] A. Clauset, C. R.Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. SIAM Review. 51, 661703 (2009).

[51] A. Deluca and A. Corral. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. Acta Geophysica. 61, 1351-1394 (2013).

[52] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning (Springer, New York, 2009).

[53] K. P. Burnham and D. R. Anderson. Model selection and multimodal inference: a practical information-theoretic approach. Spinger, New York (2002).

[54] H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 19, 716-723 (1974).

[55] R. E. Kass and A. E. Raftery. Bayes Factors. Journal of the American Statistical Association. 90, 773-795 (1995).

[56] E. T. Jaynes, Probability Theory: The Logic of Science (Oxford University Press, Oxford, 2003).

[57] R. Günther, L. Levitin, B. Schapiro, P. Wagner, Zipf 's law and the effect of ranking on probability distributions, Int. J. Theor. Phys. 35, 395 (1996).

[58] M. Cristelli, M. Batty, and L. Pietronero, There is more than a power law in Zipf, Scientific reports 2, 812. (2015).

[59] M.P.H. Stumpf and M. A. Porter. Critical Truths About Power Laws. Science, 335, 665666 (2012).

[60] M.S. Weiss. Modification of the Kolmogorov-Smirnov Statistic for Use with Correlated Data. Journal of the American Statistical Association. 73, 872-875 (1978).

[61] R. Chicheportiche and J.-P. Bouchaud. Goodness-of-Fit tests with Dependent Observations. Journal of Statistical Mechanics: Theory and Experiment. P09003 (2011).

[62] M.A. Serrano, A. Flammini, and F. Menczer, Modeling statistical properties of written text., PlOS ONE 4, e5372 (2009).

[63] Z. Eisler, I. Bartos, and J. Kertész. Fluctuation scaling in complex systems: Taylor's law and beyond. Advances in Physics. 57, 89-142 (2008).

[64] R. Louf and M. Barthelemy, Scaling: lost in the smog, Environment and Planning B: Planning and Design 41, 767 (2014).

[65] While some of the laws clearly intend to speak about the language as whole, in practice they are tested and motivated by observations in specific texts which are thus implicitly or explicitly assumed to reflect the language as a whole.

[66] A relative of Maxwell's Daemon known from Thermodynamics.