PLOS | ONE

RESEARCH ARTICLE

Zipf's Law: Balancing Signal Usage Cost and Communication Efficiency

Christoph Salge[1], Nihat Ay[2,3,4], Daniel Polani[1,5], Mikhail Prokopenko[5,6]*

1 Department of Computer Science, University of Hertfordshire, Hatfield, United Kingdom, **2** Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, **3** Santa Fe Institute, Santa Fe, United States of America, **4** Department of Mathematics and Computer Science, Leipzig University, Leipzig, Germany, **5** Complex Systems Research Group, Faculty of Engineering and IT, The University of Sydney, Sydney, Australia, **6** Department of Computing, Macquarie University, Sydney, Australia

* mikhail.prokopenko@sydney.edu.au

Citation: Salge C, Ay N, Polani D, Prokopenko M (2015) Zipf's Law: Balancing Signal Usage Cost and Communication Efficiency. PLoS ONE 10(10): e0139475. doi:10.1371/journal.pone.0139475

Editor: Neil R. Smalheiser, University of Illinois-Chicago, UNITED STATES

Received: January 18, 2015

Accepted: September 14, 2015

Published: October 1, 2015

Data Availability Statement: All relevant data are within the paper.

Funding: CS and DP were supported by the European Commission as part of the CORBYS (Cognitive Control Framework for Robotic Systems) project under contract FP7 ICT-270219.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

We propose a model that explains the reliable emergence of power laws (e.g., Zipf's law) during the development of different human languages. The model incorporates the principle of least effort in communications, minimizing a combination of the information-theoretic communication inefficiency and direct signal cost. We prove a general relationship, for all optimal languages, between the signal cost distribution and the resulting distribution of signals. Zipf's law then emerges for logarithmic signal cost distributions, which is the cost distribution expected for words constructed from letters or phonemes.

Introduction

Zipf's law [1] for natural languages states that the frequency $p(s)$ of a given word s in a large enough corpus of a (natural) language is inversely proportional to the word's frequency rank. Zipf's law postulates a power-law distribution for languages with a specific power law exponent β, so if s_t is the t-th most common word, then its frequency is proportional to

$$p(s_t) \sim \frac{1}{t^{\beta}}, \tag{1}$$

with $\beta \approx 1$. Empirical data suggests that the power law holds across a variety of natural languages [2], but the exponent β can vary, depending on the language and the context, with a usual value of $\beta \approx 2$ [3]. While the adherence to this "law" in different languages suggests a underlying common principle or mechanism, a generally accepted explanation for this phenomenon is still lacking [4].

Several papers [5–7] suggest that random texts already display a power law distribution sufficient to explain Zipf's law, but a detailed analysis [8] with different statistical tests rejects this hypothesis and argues, that there is a "meaningful" mechanism at play, which causes this distribution across different natural languages.

If we reject the idea that Zipfian distribution are produced as a result of a process that randomly produces words, then the next logical step is to ask what models can produce such distributions and agrees with our basic assumptions about language? Mandelbrot [9] models language as a process of producing symbols, where each different symbol (word) has a specific cost. He argues that this cost grows logarithmically for more expensive symbols. He then considers the information of this process, and proves that a Zipfian distribution of the symbols produces the maximal information per cost ratio. Similar, more recent models [10] prove that power laws result from minimizing a logarithmic cost functions while maximising a process's entropy (or self-information) [11]. But all these cost functions look at languages as a single random process, only optimizing the output distribution and ignoring any relationship between used words and intended meaning. This makes it a questionable model for human language (similar to the models with random text) as it does not account for communication efficiency, i.e., the model is not sensitive to how much information the words contain about the referenced concepts, nor does it offer any explanation on how certain words come to be assigned to certain meanings.

An alternative model by Cancho and Solé [12] follows the original idea of Zipf [1], by modelling the evolution of language based on the principle of least effort, where the assignment of words to concepts is optimized to minimize a weighted sum of speaker and listener effort. While simulations of the model produce distributions which qualitatively resemble power laws, a detailed mathematical investigation [4] reveals that the optimal solution of this model is, in fact, not following a power law; thus, the power law characteristics of the simulation results seems to be an artefact of the particular optimization model utilized.

Thus, to our knowledge, the question of how to achieve power laws in human language from the least effort principle is still not satisfactorily solved. Nevertheless, the idea from [1, 12] to explain power laws as the result of an evolutionary optimization process that minimizes some form of language usage cost remains attractive. In this vein, we present an alternative model for the least effort principle in language: we minimize a cost function consisting of communication inefficiency and an inherent cost for each signal (word). To avoid past pitfalls of statistical analysis when looking for power laws [13], we offer mathematical proof that any optimal solution for our cost function necessarily realizes a power law distribution, as long as the underlying cost function for the signals increases logarithmically (if the signals are ordered according to cost rank). The result generalizes beyond this as we can state a general relationship between the cost structure of the individual signals and the resulting optimal distribution of the language signals.

We should also point out that a power-law often is not the best fit to real data [14]. However, the motivation of our study differs from that of [14] which attempted to find a mechanism, i.e., Random Group Formation (RGF), that fits and, crucially, predicts the data very well —instead, we attempt to find a model formalizing the least effort principle as a mechanism generating power laws.

Another important consideration is that there in general may be multiple mechanisms generating power laws, and one cannot *post hoc* reconstruct necessarily which mechanism resulted in the observed power law. We believe, however, that it is nevertheless useful to develop a mathematically rigorous version of such a mechanism (i.e., the least effort principle) applicable to languages in particular, as it would provide additional explanatory capacity in analyzing structures and patterns observed in languages [15, 16].

The resulting insights may be of interest beyond the confines of power-law structures and offer an opportunity to study optimality conditions in other types of self-organizing coding systems, for instance in the case of the genetic code [17]. The suggested formalization covers a general class of optimal solutions balancing cost and efficiency, with power laws appearing as a

special case. Furthermore, the proposed derivation highlights a connection between scaling in languages and thermodynamics, as the scaling exponent of the resulting power law is given by the corresponding inverse temperature (which in general relates the information-theoretic or statistical-mechanical interpretation of a system through its entropy and the system's thermodynamics associated with its energy).

1 Model

We will use a model, similar to that used by Ferrer i Cancho and Solé [12], which considers languages as an assignment of symbols to objects, and then optimizes this assignment function in regard to some form of combined speaker and listener effort. The language emerging from our model is also based on the optimality principle of least effort in communication, but uses a different cost function.

The model has a set of n signals S and a set of m objects R. Signals are used to reference objects, and a language is defined by how the speaker assigns signals to objects, i.e. by the relation between signals and objects. The relation between S and R in this model can be expressed by a binary matrix A, where an element $a_{i,j} = 1$ if and only if signal s_i refers to object r_j.

This model allows one to represent both *polysemy* (that is, the capacity for a signal to have multiple meanings by referring to multiple objects), and *synonymy*, where multiple signals refer to the same object. The relevant probabilities are then defined as follows:

$$p(s_i|r_j) = \frac{a_{i,j}}{\omega_j} \tag{2}$$

where ω_j is the number of synonyms for object r_j, that is $\omega_j = \sum_i a_{i,j}$. Thus, the probability of using a synonym is equally distributed over all synonyms referring to a particular object. Importantly, it is also assumed that $p(r_j) = \frac{1}{m}$ is uniformly distributed over the objects, leading to a joint distribution:

$$p(s_i, r_j) = p(r_j)\ p(s_i|r_j) = \frac{a_{i,j}}{m\omega_j}\ . \tag{3}$$

In the previous model [12] each language has a cost based on a weighted combination of speaker and listener effort. The effort for the listener should be low if the received signal s_i leaves little ambiguity as to what object r_j is referenced, so there is little chance that the listener misunderstands what the speaker wanted to say. In the model of Ferrer i Cancho and Solé [12], the cost for listening to a specific signal s_i is expressed by the conditional entropy:

$$H_{R|s_i}(p) \equiv -\sum_{j=1}^{m} p(r_j|s_i)\log_m\ p(r_j|s_i)\ . \tag{4}$$

The overall effort for the listener is then dependent on the probability of each signal and the effort to decode it, that is

$$H_{R|S}(p) \equiv \sum_{i=1}^{n} p(s_i)H_{R|s_i}\ . \tag{5}$$

Ferrer i Cancho and Solé argue that the listener effort is minimal when this entropy is minimal, in which case there is a deterministic mapping between signals and objects.

The effort for the speaker is expressed by the entropy H_S, which is, as the term in Eq (5), bound between 0 and 1, via the log with respect to n:

$$H_S(p) \equiv -\sum_{i=1}^{n} p(s_i)\log_n p(s_i) \ . \tag{6}$$

Ferrer i Cancho and Solé then combine the listener's and speaker's efforts within the cost function Ω_λ as follows:

$$\Omega_\lambda = \lambda H_{R|S} + (1-\lambda)H_S \ , \tag{7}$$

with $0 \leq \lambda \leq 1$.

It can be shown that the cost function Ω_λ given by Eq (7) is a specific case of a more general *energy* function that a communication system must minimize [4, 18]

$$\Omega^0_\lambda = -\lambda I(S;R) + (1-\lambda)H_S \ , \tag{8}$$

where the mutual information $I(S;R) = H_R - H_{R|S}$ captures the communication efficiency, i.e. how much information the signals contain about the objects. This energy function better accounts for subtle communication efforts [19], since H_S is arguably both a source of effort for the speaker and the listener because the word frequency affects not only word production but also recognition of spoken and written words [16]. The component $I(S;R)$ also implicitly accounts for both $H_{S|R}$ (a measure of the speaker's effort of coding objects) and $H_{R|S}$ (i.e., a measure of the listener's effort of decoding signals). It is easy to see that

$$\Omega^0_\lambda = -\lambda H_R + \lambda H_{R|S} + (1-\lambda)H_S = -\lambda H_R + \Omega_\lambda \ , \tag{9}$$

and so when the entropy H_R is constant, e.g. under the uniformity condition $p(r_j) = \frac{1}{m}$, the more generic energy function Ω^0_λ reduces to the specific Ω_λ.

We propose instead another cost function that not only produces optimal languages exhibiting power laws, but also retains the clear intuition of generic energy functions which typically reflect the global quality of a solution. Firstly, we represent the communication inefficiency by the information distance, the Rokhlin metric, $H_{S|R} + H_{R|S}$ [20, 21]. This distance is often more sensitive than $-I(S;R)$ in measuring the "disagreements" between variables, especially in the case when one information source is contained within another [22].

Secondly, we define the signal usage effort by introducing an explicit cost function $c(s_i)$, which assigns each signal a specific cost. The signal usage cost for a language is then the weighted average of this signal specific cost:

$$\sum_{i=1}^{n} p(s_i)c(s_i) \ . \tag{10}$$

This is motivated by the basic idea that words have an intrinsic cost associated with using (speaking, writing, hearing, reading) them. To illustrate, a version of English where each use of the word "I" is replaced with "Antidisestablishmentarianism" and vice versa should not have the same signal usage cost as normal English. The optimal solution considering the signal usage cost alone would be to reference every object with the cheapest signal.

The overall cost function for a language Ω^c_λ is the energy function trading off the communicative inefficiency with the signal usage cost, with $0 < \lambda \leq 1$ trading off the efforts as follows:

$$\Omega^c_\lambda(p) = \lambda(H_{S|R}(p) + H_{R|S}(p)) + (1-\lambda)\sum_{i=1}^{n} p(s_i)c(s_i) \ , \tag{11}$$

where $p = p(s_i, r_j)$ is the joint probability. A language can be optimized for different values of λ, weighting the respective costs. The extreme case ($\lambda = 0$) with only the signal usage cost defining the energy function is excluded, while the opposite extreme ($\lambda = 1$) focusing on the communication inefficiency is considered. Following the principle of least effort, we aim to determine the properties of those languages that have minimal cost according to Ω^c_λ.

2 Results

First of all, we establish that all local minimizers, and hence all global minimizers, of the cost function (11) are solutions without synonyms. Formally, we obtain the following result.

Theorem 1. *Each local minimizer of the function*

$$\mathcal{C} \;\to\; \mathbb{R}, \qquad p \mapsto \Omega^c_\lambda(p),$$

where

$$\mathcal{C} \;:=\; \{p \in \mathcal{P}(S \times R) \;:\; p(r_j) \;=\; \textstyle\sum_i p(s_i, r_j) = \dfrac{1}{m} \;\; \textit{for all } j\} \;\;,$$

and $\Omega^c_\lambda(p)$ is specified by the Eq (11), $0 < \lambda \leq 1$, can be represented as a function f: $R \to S$ such that

$$p(s_i, r_j) = \begin{cases} 1/m & \text{if} \quad s_i = f(r_j); \\ 0 & \text{otherwise}. \end{cases} \tag{12}$$

The proof is given in Appendix 1. Note that each solution, i.e. each distribution p in expression (3), corresponds to a matrix A (henceforth called *minimizer matrix*) which is given in terms of function f as follows:

$$a_{i,j} = \begin{cases} 1 & \text{if} \quad s_i = f(r_j); \\ 0 & \text{otherwise}. \end{cases} \tag{13}$$

The main outcome of this observation is that the analytical minimization of the suggested cost function results in solutions without synonyms—since any function f precludes multiple signals s referring to the same object r. That is, each column in the minimizer matrix has precisely one non-zero element. Polysemy is allowed within the solutions.

We need the following lemma as an intermediate step towards deriving the analytical relationship between the specific word cost $c(s)$ and the resulting distribution $p(s)$.

Lemma 2. *For each solution p minimizing the function Ω^c_λ,*

$$H_{R|S} + \frac{1}{\log_n m} H_S = 1 \;\;. \tag{14}$$

The proof follows from the joint entropy representations

$$H_{S,R} = \frac{H_{R|S}}{1 + \log_m n} + \frac{H_S}{1 + \log_n m} \tag{15}$$

$$= \frac{H_{S|R}}{1 + \log_n m} + \frac{H_R}{1 + \log_m n} \;\;, \tag{16}$$

noting that for each minimal solution $H_{S|R} = 0$, while $H_R = 1$ under the uniformity constraint $p(r_j) = \frac{1}{m}$.

Corollary 3. *If* $n = m$, $H_{R|S} + H_S = 1$.

Using this lemma, and noting that each such solution represented as a function $f: R \to S$ has the property $H_{S|R} = 0$, we reduce the Eq (11) to

$$\Omega^c_\lambda(p) = \lambda\left(1 - \frac{1}{\log_n m} H_S(p)\right) + (1-\lambda)\sum p(s_i)c(s_i) \tag{17}$$

$$= \lambda + \frac{\lambda}{\log_n m}\sum p(s_i)\log_n p(s_i) + (1-\lambda)\sum p(s_i)c(s_i). \tag{18}$$

Varying with respect to $p(s_i)$, under the constraint $\sum p(s_i) = 1$, yields the extremality condition

$$\frac{\lambda}{\log_n m}\left(\log_n p(s_i) + 1\right) + (1-\lambda)c(s_i) - \kappa' = 0 \tag{19}$$

for some Lagrange multiplier κ'. The minimum is achieved when

$$p(s_i) = \kappa e^{-\beta c(s_i)} \ , \tag{20}$$

where

$$\beta = \frac{1-\lambda}{\lambda}\log_n m \ , \tag{21}$$

$$\kappa = \frac{1}{\sum e^{-\beta c(s_j)}} \ . \tag{22}$$

In addition, we require

$$c(s_i) = \ln m_i \tag{23}$$

for some integer m_i such that $\sum m_i = m$. The last condition ensures that the minimal solutions $p(s_i)$ correspond to functions $p(s_i, r_j)$ (i.e., minimizer matrices without synonyms). In other words, the marginal probability (20) without the condition (23) may not concur with the probability $p(s_i, r_j)$ that represents a minimizer matrix under the uniformity constraint $p(r_j) = \frac{1}{m}$.

Under the condition (23), we have $p(s_i) = \kappa m_i^{-\beta}$, while $\kappa = 1/\sum m_i^{-\beta}$. In general, one may relax the condition (23), specifying instead an upper-bounded error of approximating the minimal solution by any $p(s_i) = \kappa e^{-\beta c(s_i)}$ which would then allow for arbitrary cost functions $c(s)$.

Interestingly, the optimal marginal probability distribution (20) is the Gibbs measure with the energy $c(s_i)$, while the parameter β is, thermodynamically, the inverse temperature. It is well-known that the Gibbs measure is the unique measure maximizing the entropy for a given expected energy, and appears in many solutions outside of thermodynamics [23–25].

Let us now consider some special cases. For the case of equal effort, i.e. $\lambda = 0.5$, and $n = m$, the solution simplifies to $\beta = 1$ and $p(s_i) = \kappa m_i^{-1}$, where $\kappa = 1/\sum m_i^{-1}$.

Another important special case is given by the cost function $c(s_i) = \ln \rho_i/N$, where ρ_i is the rank of symbol s_i, and N is a normalization constant equal to $\frac{n(n+1)}{2m}$ (so that $\sum \rho_i/N = m$). In this case, the optimal solution is attained when

$$p(s_i) = \frac{\kappa N^\beta}{\rho_i^\beta} \tag{24}$$

with

$$\kappa = \frac{1}{N^{\beta} \sum \rho_j^{-\beta}} \quad . \tag{25}$$

This means that a power law with the exponent β, specified by Eq (21), is the optimal solution in regard to our cost function (11) if the signal usage cost increases logarithmically. In this case, the exponent β depends on the system's size (n and m) and the efforts' trade-off λ. Importantly, this derivation shows a connection between scaling in languages and thermodynamics: if the signal usage cost increases logarithmically, then the scaling exponent of the resulting power law is given by the corresponding inverse temperature.

Zipf's law (a power law with exponent $\beta = 1$) is then nothing but a special case for systems that satisfy $\log_n m = \frac{\lambda}{1-\lambda}$. For instance, for square matrices, Zipf's law results from the optimal languages which satisfy equal efforts, i.e., $\lambda = 0.5$. The importance of equal cost was emphasized in earlier works [4, 26]. The exponent defined by Eq (21) changes with the system size (n or m), and so the resulting power law "adapts" to linguistic dynamics and language evolution in general.

The assumption that the cost function is precisely logarithmic results in an exact power law. If, on the other hand, the cost function deviates from being precisely logarithmic, then the resulting dependency would only approximate a power law—this imprecision may in fact account for different degrees of success in fitting power laws to real data.

In summary, the derived relationship expresses the optimal probability $p(s)$ in terms of the usage cost $c(s)$, yielding Zipf's law when this cost is logarithmically distributed over the symbols.

3 Discussion

To explain the emergence of power laws for signal selection, we need to explain why the cost function of the signals would increase logarithmically, if the signals are ordered by their cost rank. This can be motivated, across a number of languages, by assuming that signals are in fact words, which are made up of letters from a finite alphabet; or in regard to spoken language, are made of from a finite set of phonemes. Compare [27], in which Nowak and Krakauer demonstrate how the error limits of communication with a finite list of phonemes can be overcome by combining phonemes into words.

Lets assume that each letter (or phoneme) has an inherent cost which is approximate to a unit letter cost. Furthermore, assume that the cost of a word roughly equals the sum of its letter costs. A language with an alphabet of size a then has a unique one letter words which the approximate cost of one, a^2 two letter words with an approximate cost of two, a^3 three letter words with a cost of three, etcetera. If we rank these words by their cost, then their cost will increase approximately logarithmically with their cost rank. To illustrate, Fig 1 is a plot of the 1000 cheapest unique words formed with a ten letter alphabet (with no word length restriction), where each letter has a random cost between 1.0 and 2.0. The first few words deviate from the logarithmic cost function, as their cost only depends on the letter cost itself, but the latter words closely follow a logarithmic function. A similar derivation of the logarithmic cost function from first principles can be found in the model of Mandelbrot [9].

This signal usage cost can be interpreted in different ways. In spoken language it might simply be the time needed to utter a word, which makes it a cost both for the listener and the speaker. In written language it might be the effort to write a word, or the bandwidth needed to transmit it, in which case it is a speaker cost. On the other hand, if one is reading a written text, then the length of the words might translate into "listener" cost again. In general, the average

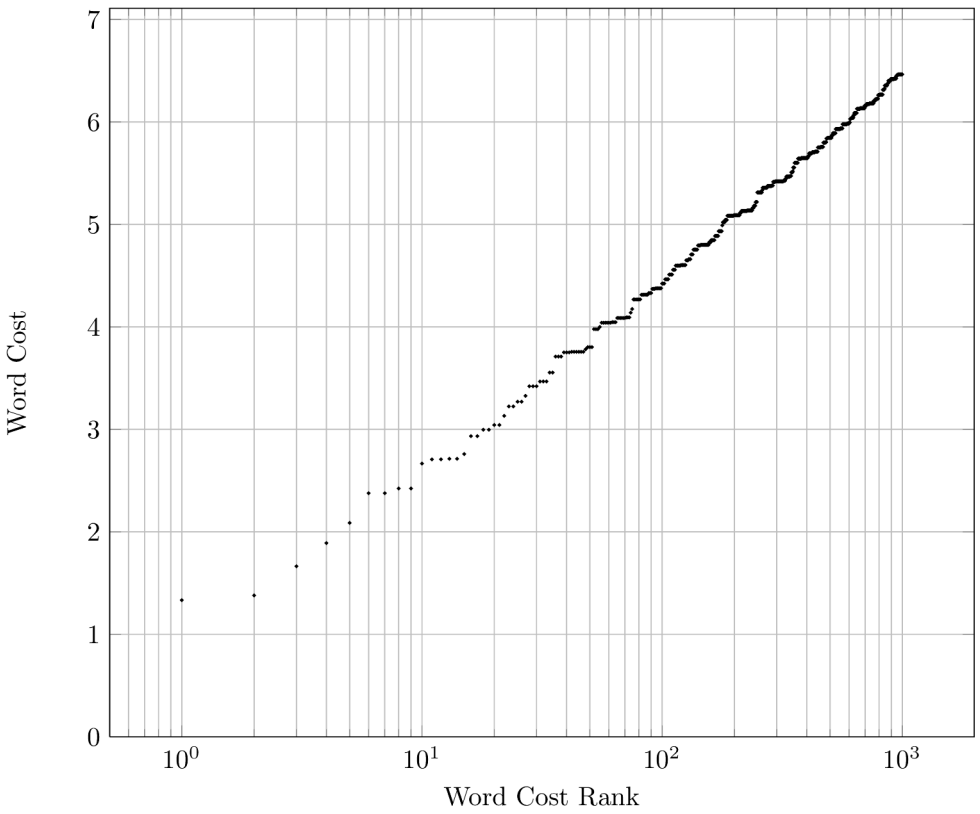

Fig 1. A log-plot of the 1000 cheapest words created from a 10 letter alphabet, ordered by their cost rank. Word cost is a sum of individual letter cost, and letter cost is between 1.0 and 2.0 units.

doi:10.1371/journal.pone.0139475.g001

signal usage cost corresponds to the effort of using a specific language to communicate for all involved parties. This differs from the original least effort idea, which balances listener and speaker effort [1]. In our model we balance the general effort of using the language with the communication efficiency, which creates a similar tension, as described in [12], between using a language that only uses one signal, and a language that references every object with its own signal. If only communication efficiency was relevant, then each object would have its own signal. Conversely, if only cost mattered, then all objects would be referenced by the same cheapest signal. Balancing these two components with a weighting factor λ yields power laws, where β varies with changes in the weighting factor. This is in contrast to the model in [12], where power laws were only found in a phase transition along the weighting factor. Also, in [3] Cancho discusses how some variants of language (military, children) have β values that deviate from the β value of their base language, which could indicate that the effort of language production or communication efficiency is weighted differently in these cases, resulting in different optimal solutions, which are power laws with other values for β.

We noted earlier that there are other options to produce power laws, which are insensitive to the relationship between objects and signals. Baek et al. [14] obtain a power law by minimizing the cost function $I_{cost} = -H_S + \langle \log s \rangle + \log N$, where $\langle \log s \rangle = \sum p(s_i) \log(s_i)$, and $\log(s_i)$ is interpreted as the logarithm of the index of s_i (specifically, its rank). Their argument that this cost function follows from a more general cost function $H_{R|S} = -I(S;R) + H_R$, where H_R is

constant, is undermined by their unconventional definition of conditional probability (cf. Appendix A [14]). Specifically, this probability is defined as $p(r \mid s) = \frac{\delta_{s'(r),s}}{sN(s)}$, where $N(s)$ is the number of objects to which signal s refers. This definition not only requires some additional assumptions in order to make $p(r|s)$ a conditional probability, but also implicitly embeds the "cost" of symbol s within the conditional probability $p(r|s)$, by dividing it by s. Thus, we are left with the cost function I_{cost} *per se*, not rigorously derived from a generic principle, and this cost function ignores joint probabilities and the communication efficiency in particular.

A very similar cost function was offered by Visser [10], who suggested to maximize H_S subject to a constraint $\langle \log s \rangle = \chi$, for some constant χ. Again, this maximization produces a power law, and again we may note that the cost function and the constraint used in the derivation do not capture communication efficiency or trade-offs between speaker and listener, omitting joint probabilities as well.

Finally, we would like to point out that the cost function $-H_S + \langle \log s \rangle$ is equivalent to the cost function $H_{R|S} - H_{S|R} + \langle \log s \rangle$, under constant H_R. This expression reveals another important drawback of minimizing $-H_S + \langle \log s \rangle$ directly: while minimizing $H_{R|S}$ reduces the ambiguity of polysemy, minimizing $-H_{S|R}$ explicitly "rewards" the ambiguity of synonyms. In other words, languages obtained by minimizing such a cost directly do exhibit a power law, but mostly at the expense of potentially unnecessary synonyms.

There may be a number of reasons for the avoidance of synonyms in real languages. While an analysis of synonymy dynamics in child languages or aphasiacs is outside of scope of this paper, it is worth pointing out that some studies have suggested that the learning of new words by children is driven by synonymy avoidance [28]. As the vocabulary and the word use are growing in children (with meaning overextensions decreasing over time), reducing the effort for the listener becomes more important [29]. Several principles underlying lexicon acquisition by children, identified by Clark [30], emphasize the dynamics of synonymy reduction. For example, the principle of *conventionality* and *contrast* ("speakers take every difference in form to mark a difference in meaning") combine in providing some precedence to semantic overlaps, leading children to eventually accept the parents' (more conventional) word for a semantically overlapping concept. The principle of *transparency* explains how a preference to use a more transparent word helps to reduce ambiguity in the lexicon. It has also been recently shown that the exponent of Zipf's law (when rank is the random variable) tends to decrease over time in children [31]. The study correlated this evolution of the exponent with the reduction of a simple indicator of syntactic complexity given by the mean length of utterances (MLU), and concluded that this supports the hypothesis that the inter-related exponent of Zipf's law and linguistic complexity tend to decrease in parallel.

Regarding synonyms it should also be noted, that while they exist, their number is usually comparatively low. If we are looking at a natural language, which might have ca. 100.000 words, we will not find a concept that has 95.000 synonyms. Most concepts have synonyms in the single digits, if they have any. The models that look at just the output distribution could produce languages with such an excessive number of synonyms. In our model the ideal solution has no synonyms, but the existing languages, which are constantly adapting, could be seen as close approximations, where out of 100.000 possible synonyms, most concepts have only very few synonyms, if any. As noted earlier, while precise logarithmic cost functions would produce perfect power-law distributions, natural languages do not fit Zipf's law exactly but only approximately.

These observations support our conjecture that, as languages mature, the communicative efficiency and the balance between speaker's and listener's efforts become a more significant driver, and so the simplistic cost function $-H_S + \langle \log s \rangle$ can no longer be justified.

In contrast, the cost function proposed in this paper $H_{R|S} + H_{S|R} + \langle \log s \rangle$ reduces to $-H_S + \langle \log s \rangle$ only *after* minimizing over the joint probabilities $p(s, r)$. Importantly, it captures communication (in)efficiency and average signal usage explicitly, balancing out different aspects of the communication trade-offs and representing the concept of least effort in a principled way. The resulting solutions do not contain synonyms, which disappear at the step of minimizing over $p(s, r)$, and so correspond to "perfect", maximally efficient and balanced, languages. The fact that even these languages exhibit power (Zipf's) laws is a manifestation of the continuity of scale-freedom in structuring of languages, along the refinement of cost functions representing the least effort principle: as long as the language develops closely to the optima of the prevailing cost function, power laws will be adaptively maintained.

In conclusion, our paper addresses the long-held conjecture that the principle of least effort provides a plausible mechanism for generating power laws. In deriving such a formalization, we interpret the effort in suitable information-theoretic terms and prove that its global minimum produces Zipf's law. Our formalization enables a derivation of languages which are optimal with respect to both the communication inefficiency and direct signal cost. The proposed combination of these two factors within a generic cost function is an intuitive and powerful method to capture the trade-offs intrinsic to least-effort communication.

4 Appendix

Theorem 1. *Each local minimizer of the function*

$$\mathcal{C} \;\to\; \mathbb{R}, \qquad p \mapsto \Omega^c_\lambda(p),$$

where

$$\mathcal{C} \;:=\; \{p \in \mathcal{P}(S \times R) \;:\; p(r_j) \;=\; \textstyle\sum_i p(s_i, r_j) = \dfrac{1}{m} \;\; \textit{for all } j\} \;,$$

and $\Omega^c_\lambda(p)$ *is specified by the* Eq (11), $0 < \lambda \leq 1$, *can be represented as a function* $f : R \to S$ *such that*

$$p(s_i, r_j) = \begin{cases} 1/m & \text{if} \quad s_i = f(r_j); \\ 0 & \text{otherwise}. \end{cases} \tag{26}$$

In order to prove this theorem, we establish a few preliminary propositions (these results are obtained by Nihat Ay).

4.1 Extreme points

The extreme points of $\mathcal{C}$ are specified by the following proposition.

Proposition 2. *The set* $\mathcal{C}$ *has the extreme points*

$$\mathrm{Ext}(\mathcal{C}) \;=\; \{p \in \mathcal{P}(S \times R) \;:\; p(s_i, r_j) = \frac{1}{m}\, \delta_{f(r_j)}(s_i) \;\} \;,$$

where f *is a function* $R \to S$.

Proof. Consider the convex set

$$\mathcal{T} \;=\; \{A = (a_{i|j})_{i,j} \in \mathbb{R}^{m \cdot n} \;:\; a_{i|j} \geq 0 \text{ for all } i, j,$$

$$\text{and } \textstyle\sum_i a_{i|j} = 1 \text{ for all } j\}$$

of transition matrices. The extreme points of $\mathcal{T}$ are given by functions $f : j \mapsto i$. More precisely,

each extreme point has the structure

$$a_{i|j} \;=\; \delta_{f(j)}(i) \;.$$

Now consider the map $\varphi : \mathcal{T} \to \mathcal{C}$ that maps each matrix $A = (a_{i|j})_{i,j}$ to the probability vector

$$p(s_i, r_j) \;:=\; \frac{1}{m}\, a_{i|j}, \qquad \text{for all } i, j.$$

This map is bijective and satisfies $\varphi((1-t)\,A + t\,B) = (1-t)\,\varphi(A) + t\,\varphi(B)$. Therefore, the extreme points of $\mathcal{C}$ can be identified with the extreme points of $\mathcal{T}$.

4.2 Concavity

Consider the set $S = \{s_1, \ldots, s_n\}$ of signals with n elements and the set $R = \{r_1, \ldots, r_m\}$ of m objects, and denote with $\mathcal{P}(S \times R)$ the set of all probability vectors $p(s_i, r_j)$, $1 \le i \le n$, $1 \le j \le m$. We define the following functions on $\mathcal{P}(S \times R)$:

$$H_{S|R}(p) \;:=\; -\sum_j p(r_j) \sum_i p(s_i|r_j) \log_n p(s_i|r_j) \;,$$

and

$$H_{R|S}(p) \;:=\; -\sum_i p(s_i) \sum_j p(r_j|s_i) \log_m p(r_j|s_i) \;.$$

Proposition 3. *All three functions $H_{R|S}$, $H_{S|R}$, and*

$$\langle c \rangle \;:\; p \mapsto \sum_i p(s_i)\, c(s_i)$$

that are involved in the definition of Ω^c_λ are concave in p. Furthermore, the restriction of $H_{S|R}$ to the set $\mathcal{C}$ is strictly concave.

Proof. The statements follow from well-known convexity properties of the entropy and the relative entropy.

(1) *Concavity of* $\mathrm{H_{R|S}}$: We rewrite the function $H_{R|S}$ as

$$\begin{aligned}
H_{R|S}(p) &= -\sum_i p(s_i) \sum_j p(r_j|s_i) \log_m p(r_j|s_i) \\
&= -\sum_{i,j} p(s_i, r_j) \log_m \frac{p(s_i, r_j)}{\sum_j p(s_i, r_j)} \\
&= -\sum_{i,j} p(s_i, r_j) \log_m \frac{p(s_i, r_j)}{m\, \frac{1}{m} \sum_j p(s_i, r_j)} \\
&= -\sum_{i,j} p(s_i, r_j) \log_m \frac{p(s_i, r_j)}{\frac{1}{m} \sum_j p(s_i, r_j)} \;+\; 1.
\end{aligned}$$

The concavity of $H_{R|S}$ now follows from the joint convexity of the relative entropy $(p, q) \mapsto D(p\|q) = \sum_{i,j} p(s_i, r_j) \log_m \frac{p(s_i, r_j)}{q(s_i, r_j)}$.

(2) *Concavity of* $H_{S|R}$: The concavity of $H_{S|R}$ follows by the same arguments as in (1). We now prove the strict concavity of its restriction to $\mathcal{C}$.

$$\begin{aligned} H_{S|R}(p) &= -\sum_j p(r_j) \sum_i p(s_i|r_j) \log_n p(s_i|r_j) \\ &= -\sum_{i,j} p(s_i, r_j) \log_n \frac{p(s_i, r_j)}{p(r_j)} \\ &= -\sum_{i,j} p(s_i, r_j) \log_n \frac{p(s_i, r_j)}{\frac{1}{m}} \\ &= -\sum_{i,j} p(s_i, r_j) \log_n p(s_i, r_j) \; - \; \log_n m. \end{aligned}$$

The strict concavity of $H_{R|S}$ now follows from the strict concavity of the Shannon entropy.

(2) *Concavity of* $\langle c \rangle$: This simply follows from the fact that $\langle c \rangle$ is an affine function and therefore concave and convex at the same time.

With a number $0 < \lambda \leq 1$, we now consider the function

$$\Omega^c_\lambda(p) \; = \; \lambda(H_{R|S}(p) + H_{S|R}(p)) + (1 - \lambda) \sum_i p(s_i)\, c(s_i) \; .$$

From Proposition 3, it immediately follows that Ω^c_λ also has corresponding concavity properties.

Corollary 4. *For* $0 \leq \lambda \leq 1$, *the function* Ω^c_λ *is concave in p, and, if* $\lambda > 0$, *its restriction to the convex set* $\mathcal{C}$ *is strictly concave.*

4.3 Minimizers

We have the following direct implication of Corollary 4.

Corollary 5. *Let* $0 < \lambda \leq 1$ *and let p be a local minimizer of the map*

$$\mathcal{C} \; \to \; \mathbb{R}, \qquad p \mapsto \Omega^c_\lambda(p).$$

Then p is an extreme point of $\mathcal{C}$.

Proof. This directly follows from the strict concavity of this function.

Together with Proposition 2, this implies Theorem 1, our main result on minimizers of the restriction of Ω^c_λ to the convex set $\mathcal{C}$.

We finish this analysis by addressing the problem of minimizing Ω^c_λ on a discrete set. In order to do so, consider the set of 0/1-matrices that have at least one "1"-entry in each column:

$$\mathcal{S} \; := \; \left\{ (a_{i,j}) \in \{0,1\}^{n \cdot m} \; : \; \sum_i a_{i,j} \geq 1 \text{ for all } j \right\}.$$

This set can naturally be embedded into the set $\mathcal{T}$, which we have considered in the proof of Proposition 2:

$$\imath : \;\; \mathcal{S} \hookrightarrow \mathcal{T}, \quad (a_{i,j})_{i,j} \mapsto a_{i|j} \; := \; \frac{a_{i,j}}{\sum_i a_{i,j}} \; .$$

Together with the map $\varphi : \mathcal{T} \to \mathcal{C}$ we have the injective composition $\varphi \circ \imath$. From Proposition 2 it follows that the extreme points of $\mathcal{C}$ are in the image of $\varphi \circ \imath$. Furthermore, Corollary 5 implies that all local, and therefore also all global, minimizers of Ω^c_λ are in the image of $\varphi \circ \imath$. The previous work of Ferrer i Cancho and Sole [12] refers to the minimization of a function on

the discrete set $\mathcal{S}$:

$$\tilde{\Omega}^c_\lambda := \Omega^c_\lambda \circ \varphi \circ \imath : \mathcal{S} \to \mathbb{R}.$$

It is not obvious how to relate local minimizers of this function, with an appropriate notion of locality in $\mathcal{S}$, to local minimizers of Ω^c_λ. However, we have the following obvious relation between global minimizers.

Corollary 6. *A point $p \in \mathcal{C}$ is a global minimizer of Ω^c_λ if and only if it is in the image of $\varphi \circ \imath$ and $(\varphi \circ \imath)^{-1}(p)$ globally minimizes $\tilde{\Omega}^c_\lambda$.*

Author Contributions

Conceived and designed the experiments: CS NA DP MP. Performed the experiments: CS. Analyzed the data: CS NA DP MP. Contributed reagents/materials/analysis tools: NA MP. Wrote the paper: CS NA DP MP.

References

1. Zipf GK (1949) Human behavior and the principle of least effort.
2. Balasubrahmanyan V, Naranan S (1996) Quantitative linguistics and complex system studies*. Journal of Quantitative Linguistics 3: 177–228. doi: 10.1080/09296179608599629
3. Ferrer i Cancho R (2005) The variation of Zipf's law in human language. The European Physical Journal B-Condensed Matter and Complex Systems 44: 249–257. doi: 10.1140/epjb/e2005-00121-8
4. Prokopenko M, Ay N, Obst O, Polani D (2010) Phase transitions in least-effort communications. Journal of Statistical Mechanics: Theory and Experiment 2010: P11025. doi: 10.1088/1742-5468/2010/11/P11025
5. Li W (1992) Random texts exhibit Zipf's-law-like word frequency distribution. Information Theory, IEEE Transactions on 38: 1842–1845. doi: 10.1109/18.165464
6. Miller GA (1957) Some effects of intermittent silence. The American Journal of Psychology 70: 311–314. doi: 10.2307/1419346 PMID: 13424784
7. Suzuki R, Buck JR, Tyack PL (2005) The use of Zipf's law in animal communication analysis. Animal Behaviour 69: F9–F17. doi: 10.1016/j.anbehav.2004.08.004
8. Ferrer i Cancho R, Elvevåg B (2010) Random texts do not exhibit the real Zipf's law-like rank distribution. PLoS One 5: e9411. doi: 10.1371/journal.pone.0009411 PMID: 20231884
9. Mandelbrot B (1953) An informational theory of the statistical structure of language. Communication theory 84: 486–502.
10. Visser M (2013) Zipf's law, power laws and maximum entropy. New Journal of Physics 15: 043021. doi: 10.1088/1367-2630/15/4/043021
11. Shannon CE (1948) A Mathematical Theory of Communication. Bell System Technical Journal 27: 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x
12. Ferrer i Cancho R, Solé RV (2003) Least effort and the origins of scaling in human language. Proceedings of the National Academy of Sciences 100: 788–791. doi: 10.1073/pnas.0335980100
13. Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. SIAM review 51: 661–703. doi: 10.1137/070710111
14. Baek SK, Bernhardsson S, Minnhagen P (2011) Zipf's law unzipped. New Journal of Physics 13: 043004. doi: 10.1088/1367-2630/13/4/043004
15. Ferrer i Cancho R (2005) Decoding least effort and scaling in signal frequency distributions. Physica A: Statistical Mechanics and its Applications 345: 275–284. doi: 10.1016/j.physa.2004.06.158
16. Ferrer i Cancho R (2014) Optimization models of natural communication. arXiv preprint arXiv:14122486.
17. Obst O, Polani D, Prokopenko M (2011) Origins of scaling in genetic code. In: Advances in Artificial Life. Darwin Meets von Neumann. Proceedings, 10th European Conference, ECAL 2009, Budapest, Hungary, September 13-16, 2009, Springer. pp. 85–93.
18. Ferrer i Cancho R (2005) Zipf's law from a communicative phase transition. The European Physical Journal B-Condensed Matter and Complex Systems 47: 449–457. doi: 10.1140/epjb/e2005-00340-y

19. Ferrer i Cancho R, Díaz-Guilera A (2007) The global minima of the communicative energy of natural communication systems. Journal of Statistical Mechanics: Theory and Experiment 2007: P06009.
20. Rokhlin VA (1967) Lectures on the entropy theory of measure-preserving transformations. Russian Mathematical Surveys 22: 1–52. doi: 10.1070/RM1967v022n05ABEH001224
21. Crutchfield JP (1990) Information and its Metric. In: Lam L, Morris HC, editors, Nonlinear Structures in Physical Systems—Pattern Formation, Chaos and Waves, Springer Verlag. pp. 119–130.
22. Prokopenko M, Polani D, Chadwick M (2009) Stigmergic gene transfer and emergence of universal coding. HFSP Journal 3: 317–327. doi: 10.2976/1.3175813 PMID: 20357889
23. Niven RK (2009) Steady state of a dissipative flow-controlled system and the maximum entropy production principle. Physical Review E 80: 021113. doi: 10.1103/PhysRevE.80.021113
24. Niven RK (2010) Minimization of a free-energy-like potential for non-equilibrium flow systems at steady state. Philosophical Transactions of the Royal Society B: Biological Sciences 365: 1323–1331. doi: 10.1098/rstb.2009.0296
25. Prokopenko M, Lizier JT, Obst O, Wang XR (2011) Relating Fisher information to order parameters. Physical Review E 84: 041116. doi: 10.1103/PhysRevE.84.041116
26. Dickman R, Moloney NR, Altmann EG (2012) Analysis of an information-theoretic model for communication. Journal of Statistical Mechanics: Theory and Experiment 2012: P12022. doi: 10.1088/1742-5468/2012/12/P12022
27. Nowak MA, Krakauer DC (1999) The evolution of language. Proceedings of the National Academy of Sciences 96: 8028–8033. doi: 10.1073/pnas.96.14.8028
28. Ferrer i Cancho R (2013) The optimality of attaching unlinked labels to unlinked meanings. arXiv preprint arXiv:13105884.
29. Saxton M (2010) Child language: Acquisition and development. Sage.
30. Clark EV (1995) The lexicon in acquisition, volume 65 of *Cambridge Studies in Linguistics*. Cambridge University Press.
31. Baixeries J, Elvevåg B, Ferrer i Cancho R (2013) The evolution of the exponent of zipf's law in language ontogeny. PloS ONE 8: e53227. doi: 10.1371/journal.pone.0053227 PMID: 23516390