Optimal coding and the origins of Zipfian laws

Ramon Ferrer-i-Cancho[1], **Christian Bentz**[2,3] **and Caio Seguin**[4]

[1] Complexity & Quantitative Linguistics Lab, LARCA Research Group
Departament de Ciències de la Computació,
Universitat Politècnica de Catalunya,
Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3.
08034 Barcelona, Catalonia (Spain)
[2] URPP Language and Space,
University of Zürich,
Freiestrasse 16, CH-8032 Zürich, Switzerland
[3] DFG Center for Advanced Studies "Words, Bones, Genes, Tools",
University of Tübingen,
Rümelinstrae 23, D-72070 Tübingen, Germany
[4] Melbourne Neuropsychiatry Centre
The University of Melbourne and Melbourne Health
Melbourne, VIC 3010, Australia.

E-mail: `rferrericancho@cs.upc.edu, chris@christianbentz.de, caioseguin@gmail.com`

Abstract. The problem of compression in standard information theory consists of assigning codes as short as possible to numbers. Here we consider the problem of optimal coding – under an arbitrary coding scheme – and show that it predicts Zipf's law of abbreviation, namely a tendency in natural languages for more frequent words to be shorter. We apply this result to investigate optimal coding also under so-called non-singular coding, a scheme where unique segmentation is not warranted but codes stand for a distinct number. Optimal non-singular coding predicts that the length of a word should grow approximately as the logarithm of its frequency rank, which is again consistent with Zipf's law of abbreviation. Optimal non-singular coding in combination with the maximum entropy principle also predicts Zipf's rank-frequency distribution. Furthermore, our findings on optimal non-singular coding challenge common beliefs about random typing. It turns out that random typing is in fact an optimal coding process, in stark contrast with the common assumption that it is detached from cost cutting considerations. Finally, we discuss the implications of optimal coding for the construction of a compact theory of Zipfian laws and other linguistic laws.

Keywords: optimal coding, maximum entropy principle, Zipf's law for word frequencies, Zipf's law of abbreviation

PACS numbers: 89.70.-a Information and communication theory
89.75.Da Systems obeying scaling laws
05.40.-a Fluctuation phenomena, random processes, noise, and Brownian motion

1. Introduction

Zipf's law of abbreviation states that more frequent words tend to be shorter [1]. Its widespread presence in human languages [2], and the growing evidence in other species [3, 4, 5, 6, 7, 8, 9], calls for a theoretical explanation. The law of abbreviation has been interpreted as a manifestation of compression [7], assigning strings as short as possible to represent information, a fundamental problem in information theory, and coding theory in particular [10]. Here we aim to investigate compression as a fundamental principle for the construction of a compact theory of linguistic patterns in natural communication systems [11]. We explore the relationship between compression and Zipf's law of abbreviation, as well as other regularities such as Zipf's law for word frequencies. The latter states that p_i, the probability of i-th most frequent word, follows [1],

$$p_i \approx i^{-\alpha}, \tag{1}$$

where α is the exponent (a parameter of the distribution) that is assumed to be about 1 [12]. Zipf referred to equation 1 as the *rank-frequency distribution* [1].

In standard information theory, codes are strings of symbols from a certain alphabet of size N which are used to represent discrete values from a set of V elements, e.g., natural numbers [13]. Suppose that the codes have minimum length l_{min} (with $l_{min} = 1$ by default). For example, if the alphabet is formed by letters a and b, the possible codes are

$$a, b, aa, ab, ba, bb, aaa, aab, aba, abb, baa, ... \tag{2}$$

As a set of discrete values one may have natural numbers,

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, ...$$

For simplicity, we assume that we wish to code for natural numbers from 1 to V. These numbers should be interpreted as what one wishes to code for or as indices or identifiers of what one actually wishes to code for. Therefore, if one wished to code for V different objects that are not numbers from 1 to V, one should label each object with a distinct number from 1 to V.

In that framework, the problem of compression consists of assigning codes to natural numbers from 1 to V in a way to minimize the mean length of the codes, defined as [10]

$$L = \sum_{i=1}^{V} p_i l_i, \tag{3}$$

where p_i is the probability of the i-th number and l_i is the length of its code in symbols. The standard problem of compression consists of minimizing L with the p_i's as a given, and under some coding scheme [10]. Roughly speaking, a coding scheme is a constraint on how to translate a number into a code in order to warrant successful decoding, namely retrieving the original number from the code from the receiver's perspective. In the examples of coding that will follow, we assume that one wishes to code numbers from 1 to 6 on strings from an alphabet of two letters a and b. Table 1 shows an example

Table 1. An example of optimal unconstrained coding of numbers from 1 to 6 on strings from an alphabet of two letters a and b.

Number	Code
1	a
2	a
3	a
4	b
5	b
6	b

Table 2. An example of non-singular coding of numbers from 1 to 6 on strings from an alphabet of two letters a and b.

Number	Code
1	aa
2	ab
3	a
4	b
5	ba
6	bb

of unconstrained coding (no scheme is used). The coding in that example is optimal because all strings have minimum length but it is not very useful because each string has three numbers as possible interpretations.

Table 2 shows an example of so-called *non-singular* coding, meaning that a unique code is assigned to each number. Thus, every code has only one possible interpretation. If we assigned the string aa to more than one number, the coding would not be non-singular. The example in Table 1 is not non-singular either. In the standard problem of compression, the alphabet is also a given. Therefore, L is minimized with N constant.

The problem of compression can be related to human languages in two ways: either we think of the numbers as representing word types (distinct words), or as representing meaning types (distinct meanings). In the former case, codes stand for distinct word types, in the latter case, they stand for distinct meanings. If numbers represent word types, then a typical application is to solve the problem of optimal *recoding*, namely reducing the length of words as much as possible without losing their distinctiveness. If we consider numbers to represent meaning types, then human languages do not perfectly fit the non-singular coding scheme due to polysemy (the same word types can have more than one meaning). However, non-singularity is convenient for language *a priori* because it reduces the cost of communication from the listeners perspective [1, 14] as well as the cost of vocabulary learning in children [15]. Optimization pressures in both ways – shortening of codes, on the one hand, and reducing polysemy (eventually leading to non-singular coding), on the other – are likely to coexist in real languages, as suggested by experiments [16]. See [11, Section 5.2] for a possible formalization based

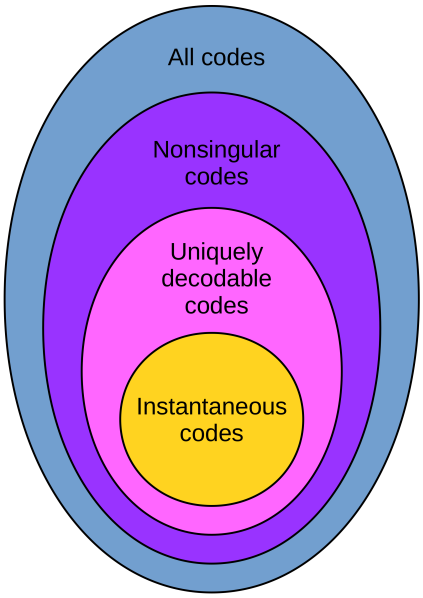

Figure 1. Classes of codes. Adapted from [10, p. 106]. Instantaneous codes, that are not described in the main text, are codes such that there is no string in the coding table that matches the begining of another string (totally or partially). An example of instantaneous code would be the binary representation of numbers from 1 to 6 in the examples of the article.

on a generalization of standard coding theory.

The information theory concepts introduced above have a direct correspondence with popular terms used in research on language optimization. The non-singular scheme implies least effort for the listener, in G. K. Zipf's terms [1]. In [16], Zipf's law of abbreviation was explained as the result of combining two pressures: *accuracy*, i.e. avoiding ambiguity, and *efficiency*, i.e. using word forms as short as possible. Communicating with maximum accuracy (no ambiguity) is equivalent to the non-singular scheme. Compression (the minimization of L) is equivalent to efficiency.

A further coding scheme, which is central to information theory, is *uniquely decodable* coding, namely, non-singular coding with unique segmentation. That is, when codes are concatenated without a separator, e.g., space, there should be only one way of breaking the sequence into codes. Uniquely decodable codes are hence a subset of non-singular codes (Fig. 1).

The coding in Table 2 is not uniquely decodable because the string *baba* can be interpreted as 4343, 55, etc. In contrast, Table 3 shows a coding that is uniquely decodable. The string *baba* can here only be interpreted as 12.

It is easy to see that written English, when written without spaces, is often not uniquely decodable. *together* can be read as both a single word and also *to get her* [18]. *Godisnowhere* illustrates the same problem: it can be read either as *God is nowhere* or as *God is now here*. Similar examples can be found in spoken English or other

Table 3. An example of uniquely decodable coding of numbers from 1 to 6 on strings from an alphabet of two letters a and b using Elias gamma encoding (a coding procedure where the code itself tells its length, turning segmentation straightforward [17, 199].

Number	Code
1	*b*
2	*aba*
3	*abb*
4	*aabaa*
5	*aabab*
6	*aabba*

Table 4. Optimal non-singular coding of numbers from 1 to 6 on strings consisting of symbols a and b. Notice that codes are assigned to frequency ranks.

Rank	Code
1	*a*
2	*b*
3	*aa*
4	*ab*
5	*ba*
6	*bb*

languages. However, unique decodability would be generally convenient for segmenting speech easily [19]. Again, unique decodability is a listener's requirement, who has to be able to retrieve the codes and the corresponding numbers when the codes are produced in a row (lacking spaces or silences in between them).

Suppose that we assign a frequency rank to each number (the most frequent number has rank 1, the 2nd most frequent number has rank 2, and so on). In his pioneering research, Mandelbrot considered the problem of compression implicitly, by assuming that word types are the numbers to code, and wrote that given any prescribed multiset of word probabilities, the average number of letters per word (L in our notation above) *is minimized if the list of words ranked by decreasing probability, coincides with the list of the V shortest letter sequences, ranked by increasing number of letters* (as in equation 2 for the case of only two letters) [20, 365]. In the language of information theory, he addressed the problem of compression under the scheme of optimal non-singular coding. To our knowledge, a formal proof of the optimality of his coding procedure is still lacking. In fact, information theoretic research has generally neglected the problem of optimal non-singular coding since then, and instead focused on uniquely decodable encoding. The reasons for this are three-fold:

- The primary target of standard information theory are artificial devices (not human brains or natural communication systems).
- The hard segmentation problem that non-singular coding implies if codes are concatenated without separators (word delimiters).

- The waste of time/space that is implied if separators are added to facilitate segmentation [10, p. 105].

These considerations may have prevented information theory from providing simple explanations to linguistic laws.

The remainder of the article is organized as follows. Section 2 presents a generalization of the problem of compression that predicts the law of abbreviation under an arbitrary coding scheme. This type of compression problem is used to prove that non-singular coding consists of assigning a string as short as possible (preserving non-singularity) to each number following frequency ranks in ascending order – as expected by Mandelbrot [20]. The coding in Table 4 satisfies this design, while that of Table 3 does not (in the latter, all codes are unnecessarily long from non-singular coding perspective except for rank 1). In case of optimal non-singular coding, Section 2 shows that l_i is an increasing logarithmic function of i, the frequency rank when $N > 1$, and a linear function of i when $N = 1$, giving an exact formula in both cases. These predictions are particular cases of Zipf's law of abbreviation.

The logarithmic relation between length and frequency rank that results from optimal non-singular coding is crucial: it provides a justification for the logarithmic constraint that is needed by the most parsimonious derivation of Zipf's rank-frequency distribution based on the maximum entropy principle [21]. For this reason, Section 3 revisits Mandelbrot's derivation of Zipf's distribution combining optimal non-singular coding, and the maximum entropy (maxent) principle [20]. This adds missing perspectives to his original analysis, and illustrates the predictive capacity of optimal non-singular coding with regards to linguistic laws. Although the distribution of word frequencies is power-law like, an exponential distribution is found for other linguistic units, e.g. part-of-speech tags [22, 116-122], colors [23], kinship terms [23] and verbal alternation classes [23]. Beyond texts, exponential distributions are found in first names in the census or social security records [23]. Non-singular coding and maxent can shed light on the emergence of these two types of distributions. In particular, Section 3 shows how the combination of the maximum entropy principle and optimal non-singular coding predicts two different distributions of ranks depending on the value of N. When $N > 1$, it predicts equation 1. When $N = 1$, it predicts an geometric distribution of ranks, namely,

$$p_i = q(1-q)^{i-1}, \tag{4}$$

where q is a parameter between 0 and 1.

Section 4 then challenges the long-standing believe that random typing constitutes evidence that Zipfian laws (Zipf's rank-frequency law and Zipf's law of abbreviation) can be derived without any optimization or cost-cutting consideration [24, 25, 16, 26]: random typing emerges as an optimal non-singular coding system in disguise. In addition, we investigate various properties of random typing, applying results on optimal coding from Section 2, and providing a simple analytical expression for the relationship

between the probability of a word and its rank – a result that Mandelbrot believed to be impossible to obtain [20].

Section 5 discusses the implications for empirical research on linguistic laws and how compression, optimal coding and maximum entropy can contribute to the construction of a general but compact theory of linguistic laws.

2. Optimal coding

Here we investigate a generalization of the problem of compression, where L (equation 3) is generalized as mean energetic cost, i.e.

$$\Lambda = \sum_{i=1}^{V} p_i \lambda_i, \tag{5}$$

and p_i and λ_i are, respectively, the probability and the energetic cost of the i-th type. Without any loss of generality, suppose that the types to be coded are sorted nonincreasingly, i.e.

$$p_1 \geq p_2 \geq ... \geq p_V, \tag{6}$$

Roughly speaking, a nonincreasing order is the outcome of sorting in decreasing order. We refer to it as nonincreasing instead of decreasing because, strictly, a decreasing order can only be obtained if all the values are distinct.

We assume that $\lambda_i = g(l_i)$, where g is a strictly monotonically increasing function of its length l_i. When $g(l_i) = l_i$ and l_i is the length in symbols of the alphabet, Λ becomes L (Equation 3), the mean code length of standard information theory [10]. The generalization follows from other research on the optimization of communication where the energetic cost of the distance between syntactically related words is assumed to be a strictly monotonically increasing function of that distance [27]. The goal of g is abstracting away from the translation of some magnitude (word length or distance between words) into a real energetic cost. Here we investigate the minimization of Λ when the p_i's are constant (given) as in the standard problem of compression, where the magnitudes are lengths of strings following a certain scheme [10].

2.1. Unconstrained optimal coding

The solution to the minimization of Λ when no further constraint is imposed is that all types have minimum length, i.e.

$$l_i = l_{min} \text{ for } i = 1, 2, ..., V. \tag{7}$$

Then Λ is minimized when $l_{min} = 0$, the smallest possible length, i.e. all types are assigned the empty string. Then the coding fails to be non-singular (for $V > 1$). If empty strings are not allowed then $l_{min} = 1$. In that case, optimal coding will produce codes that are not non-singular if $N < V$ (as in Table 1). One may get codes that are non-singular by increasing N. However, recall that N is constant in the standard problem of compression.

First, we will investigate the problem of compression (minimization of Λ) when the lengths replaced by positive real numbers belong to a given multiset. Second, we will apply the results to the problem of compression in the non-singular scheme (and the magnitudes are the lengths of strings that are non-singular).

2.2. Optimal coding with given magnitudes

Suppose that we wish to minimize Λ where the l_i's are taken from a multiset $\mathcal{L}$ of real positive values with $|\mathcal{L}| \geq V$. For instance, the values could be the length in symbols of the alphabet or the duration of the type. An assignment of elements of $\mathcal{L}$ to the l_i's consists of sorting the elements of $\mathcal{L}$ forming a sequence and assigning to each l_i the i-th element of the sequence. For an assignment, only the V first elements of the sequence matter. After an assignment, the l_i's define a subset of $\mathcal{L}$, i.e.

$$\{l_1, ..., l_i, ..., l_V\} \subseteq \mathcal{L}.$$

Therefore, $\mathcal{L}$ is a given in addition to the p_i's. $\mathcal{L}$ allows one to capture arbitrary constraints on word length, beyond the traditional coding schemes (e.g., non-singular coding or uniquely decodable encoding). Perceptibility and distinguishability factors may prevent the use of very short strings, even under a uniquely decodable scheme. Phonotactics (a branch of phonology) shows that not all possible combinations of phonemes are present in a language. Certain phonemes or combinations are harder (if not impossible) to articulate or perceive. See [28, chapters 3 and 4] for an overview of these concepts and constraints from linguistics.

This problem of compression is more general than the compression problem in standard information theory because:

- l_i is generalized as a magnitude, namely a positive real number. The strings, even when the magnitude is a length, are irrelevant.
- In case the magnitudes are string lengths, the non-singular coding scheme is obtained defining $\mathcal{L}$ as the lengths of all the different strings that can be formed. Similarly, in case of uniquely decodable coding, the string lengths have to allow one to find strings that produce them while preserving the constraints of the scheme.

These two generalization allow us to shed light on the origins of Zipf's law of abbreviation in human languages, where words do not match perfectly the constraints of traditional schemes, as well as in other species, where the coding scheme is unknown and the magnitude is measured as a time duration, namely a positive real value (e.g., [29, 9]). Moreover, it is conceivable that certain natural communication systems do not build signs by combining elementary units (such as phonemes or syllables as in human languages) – as assumed by standard information theory – but rather hollistically. Such cases could be implemented as strings of length 1 and their magnitude could be a real number indicating their expected duration.

When $|\mathcal{L}| = V$ there are as many different assignments as different sequences that can be produced from $\mathcal{L}$. When $|\mathcal{L}| \geq V$, the solution to the problem of compression

consists of finding Λ_{min}, the minimum value of Λ, and the assignments that achieve the minimum, over all the

$$\frac{|\mathcal{L}|!}{(|\mathcal{L}|-V)!}$$

assignments of elements of $\mathcal{L}$ to the l_i's. We will show that Λ_{min} is minimized exclusively by all the assignments from orderings of the elements of $\mathcal{L}$ such that the V first elements are the V smallest elements of $\mathcal{L}$ sorted in *nondecreasing* order (we refer to it as nondecreasing instead of increasing because, strictly, an increasing order can only be obtained if all the values are distinct). There is only one assignment if the values in $\mathcal{L}$ are distinct and $|\mathcal{L}| = V$.

Suppose that n_c is the number of concordant pairs of an assignment. (p_i, l_i) and (p_j, l_j) with $i \neq j$ are said to be concordant if

$$\text{sgn}(p_i - p_j) = \text{sgn}(l_i - l_j),$$

where sgn is the sign function, i.e.

$$\text{sgn}(x) = \begin{cases} \frac{x}{|x|} \text{ if } x \neq 0 \\ 0 \text{ if } x = 0. \end{cases}$$

The following lemma gives a crucial necessary condition of optimal configurations:

Lemma 1. *$\Lambda = \Lambda_{min}$ implies that the sequence l_1,...,l_i,...,l_V is sorted in nondecreasing order, i.e. $n_c = 0$ over*

$$(p_1, l_1), ..., (p_i, l_i), ...(p_V, l_V)$$

because the sequence p_1,...,p_i,...,p_V is sorted in nonincreasing order.

Proof. We will proof the contrapositive, namely that $n_c > 0$ implies $\Lambda > \Lambda_{min}$ adapting arguments in previous work [7]. Let the pair (p_i, l_i) and (p_j, l_j) be such that $1 \leq i \leq V$ and $i \neq j$ are concordant. Without any loss of generality suppose that $i < j$. Then $p_i > p_j$ by equation 6 (the case $p_i = p_j$ is excluded as the pair is concordant) and $l_i > l_j$ because the pair is concordant. If we swap l_i and l_j, then Λ will become

$$\begin{aligned} \Lambda' &= \Lambda - p_i\lambda_i - p_j\lambda_j + p_i\lambda_j + p_j\lambda_i \\ &= \Lambda + (p_i - p_j)(\lambda_j - \lambda_i) \end{aligned}$$

and then the difference between the final and the initial value of Λ becomes

$$\begin{aligned} \Delta &= \Lambda' - \Lambda \\ &= (p_i - p_j)(\lambda_j - \lambda_i). \end{aligned}$$

It is easy to see that $\Lambda > \Lambda_{min}$ as we wished because $\Delta < 0$. Recall that, in this context, $p_i > p_j$ and $l_i > l_j$ (as explained above) and that g is a strictly monotonically increasing function. □

An assignment stemming from sorting the V smallest elements of $\mathcal{L}$ in nondecreasing order (increasing order if the V smallest elements of $\mathcal{L}$ are distinct) is equivalent to one where $n_c = 0$. The following theorem expresses it formally:

Theorem 1. $\Lambda = \Lambda_{min}$ *if and only if two conditions are met*

1. l_1,...,l_i,...,l_V *are the* V *smallest elements of* $\mathcal{L}$.
2. *The sequence* l_1,...,l_i,...,l_V *is sorted in nondecreasing order, i.e.* $n_c = 0$ *over*

$$(p_1, l_1), ..., (p_i, l_i), ...(p_V, l_V)$$

because the sequence p_1,...,p_i,...,p_V *is sorted in nonincreasing order.*

Proof. We proceed proving each direction of the equivalence separately.

(i) $\Lambda = \Lambda_{min}$ implies conditions 1 and 2
We will prove the contrapositive, namely that the failure of condition 1 or 2 implies $\Lambda > \Lambda_{min}$.

(a) Suppose that condition 1 fails. Then there is an element l' in $\mathcal{L} \backslash \{l_1, ..., l_i, ..., l_V\}$ such that $l' < \max(l_1, ..., l_i, ..., l_V)$, where $\backslash$ is the multiset difference operator. Suppose that k is the index of a magnitude such that $1 \leq k \leq V$ and $l_k > l'$. Assigning l' to l_i, Λ will decrease strictly because $l_k > l'$. Thus, the original value of Λ satisfied $\Lambda > \Lambda_{min}$.

(b) Suppose that condition 2 fails. Then $\Lambda > \Lambda_{min}$ by the contrapositive of Lemma 1.

(ii) Conditions 1 and 2 imply $\Lambda = \Lambda_{min}$
We will show the contrapositive, namely that $\Lambda > \Lambda_{min}$ implies that condition 1 or 2 fails. $\Lambda > \Lambda_{min}$ can happen when condition 1 fails, as we have seen above. Suppose that condition 1 does not fail. Can we conclude that condition 2 fails? Let l_i^{min} and λ_i^{min} be the values of l_i and λ_i, respectively, in some minimum assignment, namely one yielding $\Lambda = \Lambda_{min}$. By Lemma 1, the sequence l_1^{min},...,l_i^{min},...,l_V^{min} is sorted in nondecreasing order. Notice that $\Lambda > \Lambda_{min}$ implies that the V smallest values of $\mathcal{L}$ are not identical (otherwise $\Lambda = \Lambda_{min}$ for any assignment satisfying condition 1). With this clarification in mind, it is easy to see that there must be some i such that $\lambda_i > \lambda_i^{min}$, or equivalently, $l_i > l_i^{min}$. If that did not happen, then one would have $\lambda_j \leq \lambda_j^{min}$ for each j such that $1 \leq j \leq V$ and then $\Lambda \leq \Lambda_{min}$, contradicting $\Lambda > \Lambda_{min}$. Crucially, such particular i prevents the $l_i's$ from having the non-decreasing order that is defined by the λ_i^{min}'s, leading to $n_c > 0$ as we wished.

□

The Kendall τ correlation between the p_i's and the l_i's is [30]

$$\tau(p_i, l_i) = \frac{n_c - n_d}{\binom{V}{2}},$$

where n_d is the number of discordant pairs. (p_i, l_i) and (p_j, l_j) with $i \neq j$ are said to be discordant if

$$\text{sgn}(p_i - p_j) = -\text{sgn}(l_i - l_j).$$

An implication of minimum Λ is that the $\tau(p_i, l_i)$ cannot be positive, namely $\tau(p_i, l_i) \leq 0$. In case of optimal coding, $n_c = 0$ and then

$$\tau(p_i, l_i) = -\frac{n_d}{\binom{V}{2}}$$

Since $n_d \geq 0$ one has $\tau(p_i, l_i) \leq 0$, with equality if and only if $n_d = 0$.

2.3. Optimal non-singular coding

Under the scheme of uniquely decodable codes, standard information theory tells us that the minimization of L leads to [10]

$$l_i \propto \lceil -\log_N p_i \rceil, \tag{8}$$

which is indeed a particular case of Zipf's law of abbreviation. This corresponds to the minimization of Λ with g as the identity function in our framework. Here we wish to minimize Λ with l_i as the length of the i-th most frequent type when only the p_i's are prescribed under the non-singular coding scheme (Fig. 1).

Under non-singular coding, the set of available strings consists of all the different strings of symbols that can be built with an alphabet of size N. There are N^l different strings of length l. Let S be the infinite sequence of these strings sorted by increasing length (the relative ordering of strings of the same length is arbitrary). If empty strings are not allowed, the strings in positions 1 to N have length 1, the strings in positions $N+1$ to $N+N^2$ have length 2, and so on as in 2 for $N = 2$.

Corollary 1. *Optimal non-singular coding consists of assigning the i-th string of S to the i-th most probable type for $1 \leq i \leq V$.*

Proof. We define $\mathcal{L}$ as the multiset of the lengths of all the available strings for non-singular coding (l appears N^l times in $\mathcal{L}$). As there is a one-to-one correspondence between an element of $\mathcal{L}$ and an available string, the application of theorem 1 with g as the identity function gives that

- The sequence $l_1, ..., l_i, ..., l_V$ contains the V smallest lengths, and then the codes are the shortest possible strings.
- $l_1, ..., l_i, ..., l_V$ is sorted in nondecreasing order, and then the i-th type is assigned the i-th shortest string.

□

2.4. Length as a function of frequency rank in optimal non-singular coding

We aim to derive the relationship between the rank of a type (defined according to its probability) and its length in case of optimal non-singular codes for $N \geq 1$. Suppose that p_i is the probability of the i-th most probable type and that l_i is its length.

The derivation is based on a generalization where the rank i is assigned the shortest possible string that has length l_{min} or greater. Then largest rank of types of length l is

$$i = \sum_{k=l_{min}}^{l} N^k.$$

When $N > 1$, we get

$$i = \frac{N^{l+1} - N^{l_{min}}}{N-1}$$

and equivalently

$$N^l = \frac{1}{N}[(N-1)i + N^{l_{min}}].$$

Taking logs on both sides of the equality, one obtains

$$l = \frac{\log\left(\frac{1}{N}[(N-1)i + N^{l_{min}}]\right)}{\log N}.$$

The result can be generalized to any rank of types of length l as

$$l = \left\lceil \frac{\log\left(\frac{1}{N}[(N-1)i + N^{l_{min}}]\right)}{\log N} \right\rceil. \tag{9}$$

Changing the base of the logarithm to N, one obtains

$$l = \left\lceil \log_N\left((1-1/N)i + N^{l_{min}-1}\right) \right\rceil.$$

Alternatively, equation 9 also yields

$$\begin{aligned} l &= \left\lceil \frac{\log[(N-1)i + N^{l_{min}}]}{\log N} - 1 \right\rceil. \\ &= \left\lceil \log_N[(N-1)i + N^{l_{min}}] \right\rceil - 1. \end{aligned}$$

The case $N = 1$ is trivial, one has $l = i + l_{min} - 1$. Therefore, the length of the i-th most probable type is

$$l_i = \begin{cases} \left\lceil \log_N\left((1-1/N)i + N^{l_{min}-1}\right) \right\rceil & \text{for } N > 1 \\ i + l_{min} - 1 & \text{for } N = 1. \end{cases} \tag{10}$$

We conclude that optimal coding with non-singular codes yields that the length of the i-th most probable type follows equation 10 with $l_{min} = 1$. When $N > 1$, one obtains

$$l_i = \lceil \log_N\left((1-1/N)i + 1\right) \rceil,$$

the same conclusion reached by [31] though lacking a detailed explanation.

3. The maximum entropy principle

Now we turn onto the question of making a safe prediction on the distribution of word ranks in case of optimal non-singular coding. The maximum entropy principle states that [32]

Out of all probability distributions consistent with a given set of constraints, the distribution with maximum uncertainty should be chosen.

The distribution of word frequencies has been derived via maximum entropy many times with similar if not identical methods [20, 33, 34, 35, 36, 37, 21]. Depending on the study, the target was Zipf's rank-frequency distribution (equation 1) [20, 34, 36] or its sister law with frequency as the random variable [33, 35], stating that the n_f, the number of words of frequency f, satisfies approximately

$$n_f \approx f^{-\beta}$$

with $\beta \approx 2$ [1, 38]. In some cases, maximum entropy is used as an explanation for the ubiquity of power-law like distributions, with Zipf's law for word frequencies or its sister as a particular case [37, 21]. For simplicity, here we revisit the essence of the principle focusing on how our results on optimal non-singular coding can be used to derive different rank distributions.

The maximum entropy principle allows one to obtain a distribution that maximizes the entropy of probability ranks, namely,

$$H = -\sum_{i=1}^{V} p_i \log p_i$$

under certain constraints on cost over the i's and a couple of elementary constraints on the p_i's, i.e. $p_i \geq 0$ and

$$\sum_{i=1}^{V} p_i = 1.$$

[39, 40]. For simplicity, we assume a single non-elementary cost constraint, namely L, as defined in Eq. 3. For simplicity, we assume that V is not finite. See [21] for an analysis of the case of more than one non-elementary constraint and a comparison of the finite versus infinite case. See [40] for some critical aspects of the traditional application of maximum entropy.

In our simple setup, the method leads to distributions of the form

$$p_i = \frac{e^{-\alpha l_i}}{Z}, \tag{11}$$

where α is a Lagrange multiplier and

$$Z = \sum_{j=1}^{\infty} e^{-\alpha l_j}$$

is the partition function. In case of optimal non-singular coding, we have two cases. If $N > 1$ then $l_i \approx \log_N i$ for sufficiently large N (equation 10), which transforms equation 11 into a zeta distribution, i.e.

$$p_i = \frac{1}{Z} i^{-\alpha}$$

while the partition function becomes

$$Z = \sum_{j=1}^{\infty} j^{-\alpha},$$

namely the Riemann zeta function. The zeta distribution is an approximation to Zipf's law for word frequencies.

When $N = 1$ then $l_i = i$ (equation 10 with $l_{min} = 1$), which transforms equation 11 into an exponential distribution of word frequencies, i.e.

$$p_i = \frac{1}{Z} e^{-\alpha i} \tag{12}$$

while

$$Z = \sum_{j=1}^{\infty} e^{-\alpha i}.$$

that matches the geometric distribution that is found for certain linguistic units [22, 23]. Although Equation 12 is for a discrete random variable, it has the form of the popular exponential distribution for continuous random variables. That equation actually matches the definition of the customary geometric distribution in equation 4. To see it, notice that Z is the summation of a geometric series where the first term a and the common factor r are the same, i.e. $a = r = e^{-\alpha}$. Therefore, assuming $|r| < 1$, i.e. $\alpha > 0$,

$$\begin{aligned} Z &= \frac{a}{1-r} \\ &= \frac{e^{-\alpha}}{1-e^{-\alpha}}. \end{aligned}$$

Then equation 12 can be rewritten equivalently as

$$p_i = \frac{1-e^{-\alpha}}{e^{-\alpha}} (e^{-\alpha})^i. \tag{13}$$

The substitution $q = 1 - e^{-\alpha}$ transforms equation 13 into the customary definition of a geometric distribution in equation 4 as we wished.

4. The optimality of random typing

The results on optimal coding above allow one to unveil the optimality of typing at random, assuming that the space bar is hit with a certain probability and that letters are equally likely [24]. It has been argued many times that random typing reproduces Zipf's rank-frequency distribution, e.g. [24, 41, 42, 43]. In particular, Miller concluded that the law *"can be derived from simple assumptions that do not strain ones credulity*

(unless the random placement of spaces seems incredible), without appeal to least effort, least cost, maximal information, or any other branch of the calculus of variations. The rule is a simple consequence of those intermittent silences which we imagine to exist between successive words." [24]. Similarly, W. Li argued that *"random typing shows that a random process can mimic a cost-cutting process, but not purposely."* [25]. A similar view is found in reviews of Zipf's law for word frequencies, where optimization and random typing are considered to be different mechanisms [44, 45]. The view of random typing as detached from cost reduction is also found in research on the origins of Zipf's law of abbreviation [16, 26]. Leaving aside the problem of the poor fit of random typing to their original target, i.e. the distribution of word frequencies [46, 47], these views are also problematic because random typing and least cost are not really independent issues. We will show it through the eye of the problem of compression.

The optimality of random typing can be seen in two ways. One through recoding, namely replacing each word it produces by another string so as to minimize L under the non-singular coding scheme. The other - indeed equivalent - consists of supposing that random typing is used to code for numbers whose probability matches that of the words produced by random typing. In both cases, we will show that the value of L of a random typing process cannot be reduced and thus it is optimal. Put differently, we will show that there is no non-singular coding system that can do it more efficiently (with a smaller L) than random typing.

It is easy to see that the strings that random typing produces are optimal according to Corollary 1. Recall that the probability of a "word" w of length l in random typing is [46, p. 838]

$$p_l(w) = \left(\frac{1-p_s}{N}\right)^l \frac{p_s}{(1-p_s)^{l_{min}}}, \tag{14}$$

where l is the length of w, p_s is the probability of producing the word delimiter (a whitespace), N is the size of the alphabet that the words consist of ($N > 0$) and l_{min} is the minimum word length ($l_{min} \geq 0$). Hereafter we assume for simplicity that $0 < p_s < 1$. If $p_s = 0$, strings never end. If $p_s = 1$, all the strings have length l_{min} and then random typing has to be analyzed following the arguments for unconstrained optimal coding in Section 2.1.

We will show that after sorting nondecreasingly all possible strings of length at least l_{min} that can be formed with N letters, the i-th most likely type of random typing receives the i-shortest string. First, equation 14 indicates that all words of the same length are equally likely and $p_{l+1}(w) \geq p_l(w)$ for $l \geq l_{min}$ because p_s, l_{min} and N are constants. Therefore, the ranks of words of length l are always larger than those of words of length $l+1$. Keeping this property in mind, words of the same length are assigned an arbitrary rank. Second, $p_l(w) > 0$ for all the N^l different words of length l that can be formed. Therefore, all available strings of a given length are used. The optimality of random typing for $N = 2$ and $l_{min} = 1$ can be checked easily in Table 5. The exact relationship between rank and length in random typing will be derived below.

Table 5. The probability (p_i), the length (l_i) of the i-th most frequent string (code) or random typing with $N = 2$ and $l_{min} = 1$. l_i is calculated via equation 10 with $N = 2$ and $l_{min} = 1$. p_i is calculated applying $l_{min} = 1$, $N = 2$ and l_i to equation 16.

Code	i	l_i	p_i
a	1	1	$p_s/2$
b	2	1	$p_s/2$
aa	3	2	$(1-p_s)p_s/4$
ab	4	2	$(1-p_s)p_s/4$
ba	5	2	$(1-p_s)p_s/4$
bb	6	2	$(1-p_s)p_s/4$
aaa	7	2	$(1-p_s)^2p_s/8$
...	...	...	...

Random typing satisfies a particular version of Zipf's law of abbreviation where the length of a word (l) is a linear function of its probability (p), i.e.

$$l = a \log p + b, \tag{15}$$

where a and b are constants ($a < 0$). Namely, the probability of a word is determined by its length (the characters constituting the words are irrelevant), equation 14 allows one to express l as a function of $p(w)$. Rearranging the terms of equation 14, taking logarithms, and replacing $p(w)$ by p, one recovers equation 15 with

$$a = \left(\log \frac{1-p_s}{N}\right)^{-1}$$

and

$$b = a \log \frac{(1-p_s)^{l_{min}}}{p_s}.$$

Does random typing also satisfy Zipf's law for word frequencies (equation 1)? Mandelbrot was aware "*that the relation between rank and probability is given by a step function*" for the random typing model we have considered here, but he argued that *"such a relation cannot be represented by any simple analytic expression"* [20, 364]. Knowing that random typing is optimal from the standpoint of non-singular coding it is actually possible to obtain a simple analytic expression for p_i, the probability that random typing produces a word of rank i. Replacing $p_l(w)$ by p_i and l by l_i, equation 14 becomes

$$p_i = \left(\frac{1-p_s}{N}\right)^{l_i} \frac{p_s}{(1-p_s)^{l_{min}}}, \tag{16}$$

where l_i the length of the word of rank i that is given by equation 10. To our knowledge, this is the first exact equation for p_i. In previous research, only approximate expressions for p_i have been given [24, 41, 20, 42]. Figure 2 shows p_i versus i for $N = 26$ and $p_s = 0.18$ (panel a). These are the parameters that Miller used in his classic article on random typing to mimic English [24]. The stepwise shape can be smoothed by introducing a bias towards certain letters (in the original setup, all letters are equally likely) [42, 47] or

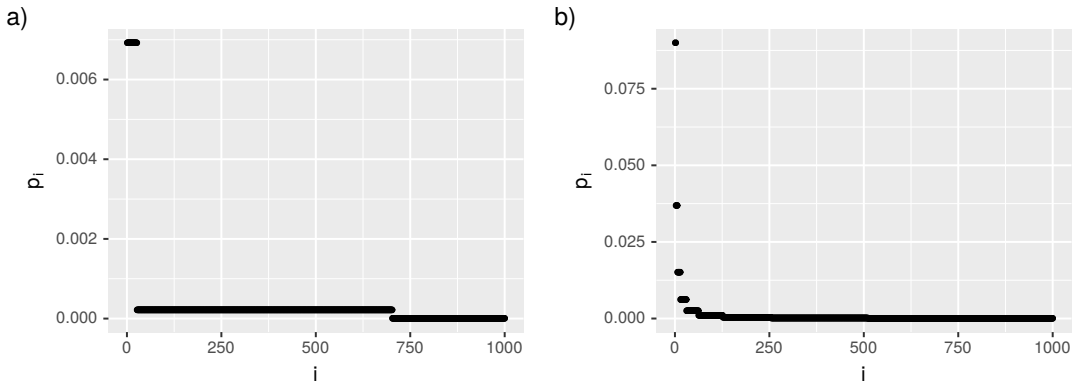

Figure 2. p_i the probability of a word of rank i produced by random typing with $p_s = 0.18$, $l_{min} = 1$ up to rank $i_{max} = 1000$ (in the standard random typing model, the maximum rank is infinite). a) $N = 26$. b) $N = 2$.

by reducing N as much as possible (reducing N is a particular case of bias that consists of turning 0 the probability of certain symbols). Figure 2 b) shows the smoothing effect of $N = 2$ (corresponding to the examples given in Table 5). Notice that N cannot be reduced further: we have shown above that $N = 1$ transforms the distribution of ranks of random typing into a geometric distribution.

5. Discussion

In his pioneering research, Zipf found a tendency of more frequent words to be shorter. He termed this observation the law of abbreviation [1]. However, he never proposed a functional dependency or mathematical model for the relationship between frequency and length.

Here we have filled a gap in standard information theory concerning optimal non-singular coding, that predicts $l_i \approx \log i$, where i is the probability rank. This result complements the well-known relationship $l_i \approx -\log p_i$ predicted by optimal uniquely decodable coding [10]. Derivations of a logarithmic relationship between the length of a word and its probability rank can be found in classic work [20, 48]. However, our derivation is novel in the sense of providing a general exact formula (not an approximation; covering $N \geq 1$ and $l_{min} \geq 0$) and involving optimal non-singular coding in the argument. It was clear to Mandelbrot that *"given any prescribed set of word probabilities, the average number of letters per words is minimized if the list of words, ranked by decreasing probability, coincides with the list of the V shortest latter sequences"* [20, 365] but he never provided an exact formula for the relationship between l_i and i as far as we know. Indeed, he actually thought it was impossible [20]. In addition, Rapoport did not take information theoretic optimality considerations into account and simply stated that *"we shall want the shortest words to be the most frequent"* [48, p. 9].

Traditionally, quantitative linguistics research has been based on the fit of power-law like models [49, 50]. Surprisingly, the predictions of information theory reviewed

above have largely been neglected. The problem concerns not only the relationship between length and frequency but also parallel quantitative linguistics research where frequency is replaced by the frequency rank (see [50, p. 274] and references therein). Some notable exceptions are discussed in the following.

In the work by Hammerl [51], both the relationship $l_i \approx \log p_i$ and $l_i \approx \log i$ are considered. He explains that Guiraud (in 1959) derived $l_i \approx \log i$ by *"purely combinatorial considerations, where all possible combinations of letters in the respective languages were allowed"* [51].‡ Unfortunately, we have not been able to find a proper reference to Guiraud's work of 1959. Therefore, we cannot tell if Guiraud was following some optimization hypothesis akin to optimal singular-coding or if he actually provided an exact formula like ours. Finally, the logarithmic relationship between the frequency of a word and its length in phonemes has also been inferred based on empirical data collected for overall eight languages (see Equation 11 in [52]). However, this particular study is bare of any mathematical/information theoretic considerations.

Besides historical considerations, our findings also have practical implications for empirical research on the law of abbreviation as an indication of optimal coding. First, it is usually assumed that a significant negative correlation between frequency and magnitude is needed for efficient coding [5, 29, 53, 9]. Our analyses indicate that a non-significant correlation can still be associated with efficient coding. For instance, we have seen that optimal coding with prescribed probabilities and magnitudes coming from some given multiset is equivalent to $\tau(p_i, l_i) \leq 0$. The same conclusion can be reached from optimal uniquely decodable coding, where all strings must have the same length when types are equally likely (recall equation 8). Therefore, the influence of compression could be wider than commonly believed. What cannot be attributed to compression is the significant positive correlation between frequency and magnitude that has been found in a subset of the repertoire of chimpanzee gestures (full body gestures) [9], and also in computer experiments with neural networks [26]. Importantly, this illustrates that compression – as reflected in the law of abbreviation – is not necessarily found in all communication systems, which undermines arguments that quantitative linguistic laws are unavoidable and hence "meaningless" [54].

Another argument along those lines is based on random typing: if random typing recreates Zipfian laws, then surely they are not an interesting subject of study [24]. However, surprisingly, random typing turns out to be an optimal encoding system. Thus, finding linguistic laws in random typing does not preclude that these laws can be explained by information theoretic principles. However, while we have unveiled the optimality of random typing, we emphasize that we have done it only from the perspective of optimal non-singular coding. The fact that random typing and optimization are not independent issues as commonly believed [24, 25, 16, 26], does not

‡ The German original reads "Guiraud (1959) hat aus rein kombinatorischen Überlegungen, wo alle möglichen Buchstabenkombinationen aus den Buchstaben der jeweiligen Sprache bei der Bildung von Wörtern zugelassen wurden [...] folgende Abhängigkeit [...] abgeleitet." This is followed by the formulars given above in the main text.

imply that random typing satisfies to a sufficient degree the optimization constraints imposed on natural languages.

We have investigated a problem of optimal coding where magnitudes stem from a given multiset of values. The problem is related to other mathematical problems outside coding theory. Notice that Λ can be seen as a scalar product of two vectors, i.e. $\vec{p} = \{p_1, ..., p_i, ..., p_V\}$ and $\vec{\lambda} = \{\lambda_1, ..., \lambda_i, ..., \lambda_V\}$ and L as a scalar product of $\vec{p}$ and $\vec{l} = \{l_1, ..., l_i, ..., l_V\}$. When $|\mathcal{L}| = V$ the problem is equivalent to minimizing the scalar (or dot) product of two vectors (of positive real values) over all the permutations of the content of each vector [55]. By the same token, the problem is equivalent to minimizing the Pearson correlation between $\vec{p}$ and $\vec{\lambda}$ when the content (but not the order) of each vector is preserved. Recall that the Pearson correlation between $\vec{p}$ and $\vec{\lambda}$ can be defined as [30]

$$r(\vec{p}, \vec{\lambda}) = \frac{\vec{p} \cdot \vec{\lambda} - \mu_p \mu_\lambda}{\sigma_p \sigma_\lambda} \tag{17}$$

where μ_x and σ_x are, respectively, the mean and the standard deviation of vector $\vec{x}$.

The link with Pearson correlation goes back to the original coding problem: such a correlation has been used to find a concordance with the law of abbreviation that is in turn interpreted as a sign of efficient coding [29]. According to equation 17, such a correlation turns out to be a linear transformation of the cost function. Put differently, minimizing Λ with prescribed $p_i's$ and with λ_i as the identity function (as it is customary in standard coding theory), is equivalent to minimizing the Pearson correlation at constant mean and standard deviation of both probabilities and magnitudes. Therefore, the Pearson correlation is a measure of the degree of optimization of a system when these means and standard deviations are constant (it is implicit that the standard deviations are not zero, otherwise the Pearson correlation is not defined).

We have seen that optimal non-singular coding predicts both a form of Zipf's law of abbreviation as well as a power-law distribution consistent with Zipf's law for word frequencies when combined with the maxent principle, revisiting an old argument by Mandelbrot [20]. The capacity of maxent to obtain Zipfian laws as well as the less popular exponential distribution of parts-of-speech [22] suggests that the principle should be considered as a critical component of a compact theory of linguistic patterns in general. For instance, $p(d)$, the probability that two syntactically related words are at distance d (in words), exhibits an exponential decay that has been derived with the help of a combination of maxent and a constraint on the average value of d [56].

The principle of maximum entropy used to derive Zipf's law for word frequencies ensures that *one is maximally uncertain about what one does not know* [32]. In the context of natural languages, a further justification of the use of the principle is that $I(S, R)$, the mutual information between words (S) and meanings (R) satisfies

$$I(S, R) \leq H(S), \tag{18}$$

where $H(S)$ is the entropy of words, namely the entropy of word probability ranks as defined above. The inequality in equation 18 follows from elementary information theory

[10], and has been applied to investigate the properties of dual optimization models of natural communication [57]. $I(S, R)$ is a measure of the capacity of words to convey meaning: maximizing $I(S, R)$ one promotes that words behave like meaning identifiers [11, Section 3]. Therefore, equation 18 suggests that the maximum entropy principle in the context of word entropy maximizes the potential of words to express meaning. The hypothesis of pressure to maximize $H(S)$ is supported by the skew towards the right that is found in the distribution of $H(S)$ in languages across the world [58].

The challenge of mathematical modelling is to find a compromise between parsimony and predictive power [59]. Concerns about parsimony are a recurrent theme when modelling Zipf's law for word frequencies [20, 21, 11]. As for maximum entropy models, it has been argued that Shannon entropy and a logarithmic constraint offer the simplest explanation for the origins of the law [21]. However, the argument is incomplete unless a justification for such a constraint is provided. Here we have shown how the logarithmic constraint follows from optimal non-singular coding. There are many possible explanations for the origins of Zipf's law based on maximum entropy [20, 33, 60, 34, 35, 36, 37, 21], and many more through other means [44, 45], but only compression can shed light on the origins of both Zipf's law for word frequencies and Zipf's law of abbreviation. The explanation of Zipf's law for word frequencies should not be separated from the explanation of other quantitative laws. Otherwise, the space of possible models is not sufficiently constrained [61], and the resulting "theory" is not a well organized theory but a patchwork of models [11].

Our theoretical framework is highly predictive in at least two senses. First, optimal coding predicts Zipf's law of abbreviation, but adherence to a traditional scheme (non-singular coding or uniquely decodable coding) is not necessary. It suffices to assume that the magnitudes come from some predefined multiset. Second, its applicability goes beyond laws from Zipf's classic work. It can also be applied to Menzerath's law, the tendency of constructs with more parts to be made of smaller parts, i.e. the tendency of words with more syllables to be made of shorter syllables [62]. Taking the number of parts of constructs as probabilities of types (p_i's) and the size of the parts as magnitudes (l_i's) and simply assuming that the number of parts are constant, Menzerath's law follows applying theorem 1 [63]. This allows one to put forward optimization as a possible hypothesis to explain the pervasiveness of the law in nature (e.g. [64, 65, 63]).

Acknowledgments

This article is dedicated to the memory of P. Grzybek (1957-2019) [50]. We thank A. Hernández-Fernández for many corrections and valuable comments and to L. Debowski for helping us to strengthen some of the mathematical proofs. We also thank N. Ay and M. Gustison for helpful discussions. RFC is supported by the grant TIN2017-89244-R from MINECO (Ministerio de Economia, Industria y Competitividad) and the recognition 2017SGR-856 (MACDA) from AGAUR (Generalitat de Catalunya). CB is supported by the DFG Center for Advanced Studies *Words, Bones, Genes,*

Tools at the University of Tübingen, and by the Swiss National Foundation Grant on "Non-randomness in Morphological Diversity: A Computational Approach Based on Multilingual Corpora" (SNF 176305) at the University of Zürich. CS is funded by a Melbourne Research Scholarship.

References

[1] G. K. Zipf. *Human behaviour and the principle of least effort.* Addison-Wesley, Cambridge (MA), USA, 1949.

[2] C. Bentz and R. Ferrer-i-Cancho. Zipf's law of abbreviation as a language universal. In Christian Bentz, Gerhard Jäger, and Igor Yanovich, editors, *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics.* University of Tübingen, 2016.

[3] M. S. Ficken, J. P. Hailman, and R. W. Ficken. A model of repetitive behaviour illustrated by chickadee calling. *Animal Behaviour*, 26(2):630–631, 1978.

[4] J. P. Hailman, M. S. Ficken, and R. W. Ficken. The 'chick-a-dee' calls of *Parus atricapillus*: a recombinant system of animal communication compared with written English. *Semiotica*, 56:121–224, 1985.

[5] R. Ferrer-i-Cancho and D. Lusseau. Efficient coding in dolphin surface behavioral patterns. *Complexity*, 14(5):23–25, 2009.

[6] R. Ferrer-i-Cancho and A. Hernández-Fernández. The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory*, 4(1), 2013.

[7] R. Ferrer-i-Cancho, A. Hernández-Fernández, D. Lusseau, G. Agoramoorthy, M. J. Hsu, and S. Semple. Compression as a universal principle of animal behavior. *Cognitive Science*, 37(8):1565–1578, 2013.

[8] B. Luo, T. Jiang, Y. Liu, J. Wang, A. Lin, X. Wei, and J. Feng. Brevity is prevalent in bat short-range communication. *Journal of Comparative Physiology A*, 199:325–333, 2013.

[9] R. Heesen, C. Hobaiter, R. Ferrer i Cancho, and S. Semple. Linguistic laws in chimpanzee gestural communication. *Proceedings of the Royal Society B: Biological Sciences*, 286:20182900, 2019.

[10] T. M. Cover and J. A. Thomas. *Elements of information theory.* Wiley, New York, 2006. 2nd edition.

[11] R. Ferrer-i-Cancho. Optimization models of natural communication. *Journal of Quantitative Linguistics*, 25:207–237, 2018.

[12] R. Ferrer-i-Cancho. The variation of Zipf's law in human language. *European Physical Journal B*, 44:249–257, 2005.

[13] M. Borda. *Fundamentals in information theory and coding.* Springer, Berlin, 1st edition, 2011.

[14] S. T. Piantadosi, H. T., and E. Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280 – 291, 2012.

[15] B. Casas, N. Català, R. Ferrer i Cancho, A. Hernández-Fernández, and J. Baixeries. The polysemy of the words that children learn over time. *Interaction Studies*, 19:389 – 426, 2018.

[16] J. Kanwal, K. Smith, J. Culbertson, and S. Kirby. Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165:45–52, 2017.

[17] P. Elias. Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, 21(2):194–203, 1975.

[18] B. McMillan. Two inequalities implied by unique decipherability. *IRE Transactions on Information Theory*, 2(4):115–116, 1956.

[19] Alexa R. Romberg and Jenny R. Saffran. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):906–914, 2010.

[20] B. Mandelbrot. Information theory and psycholinguistics: a theory of word frequencies. In P. F.

Lazarsfield and N. W. Henry, editors, *Readings in mathematical social sciences*, pages 151–168. MIT Press, Cambridge, 1966.

[21] M. Visser. Zipf's law, power laws and maximum entropy. *New Journal of Physics*, 15(4):043021, 2013.

[22] A. Tuzzi, I.-I. Popescu, and G. Altmann. *Quantitative analysis of Italian texts*, volume 6 of *Studies in Quantitative Linguistics*. RAM Verlag, Lüdenscheid, Germany, 2010.

[23] M. Ramscar. Source codes in human communication. *https://psyarxiv.com/e3hps*, 2019.

[24] G. A. Miller. Some effects of intermittent silence. *Am. J. Psychol.*, 70:311–314, 1957.

[25] W. Li. Comments to "Zipf's law and the structure and evolution of languages" A.A. Tsonis, C. Schultz, P.A. Tsonis, Complexity, 2(5). 12-13 (1997). *Complexity*, 3:9–10, 1998. Letters to the editor.

[26] R. Chaabouni, E. Kharitonov, E. Dupoux, and M. Baroni. Anti-efficient encoding in emergent communication. *arXiv:1905.12561*, 2019.

[27] R. Ferrer-i-Cancho. The placement of the head that minimizes online memory: a complex systems approach. *Language Dynamics and Change*, 5:114–137, 2015.

[28] A. Akmajian, R. A. Demers, A. K. Farmer, and R. M. Harnish. *Linguistics. An Introduction to Language and Communication*. MIT Press, 4th edition, 1995.

[29] S. Semple, M. J. Hsu, and G. Agoramoorthy. Efficiency of coding in macaque vocal communication. *Biology Letters*, 6:469–471, 2010.

[30] W. J. Conover. *Practical nonparametric statistics*. Wiley, New York, 1999. 3rd edition.

[31] M. Sudan. Transmission of information. *http://people.csail.mit.edu/madhu/ST06/scribe/L07_xshi_main.pdf*, 2006.

[32] H. K. Kesavan. *Jaynes' maximum entropy principle*, pages 1779–1782. Springer US, Boston, MA, 2009.

[33] S. Naranan and V. K. Balasubrahmanyan. Information theoretic models in statistical linguistics - Part I: A model for word frequencies. *Current Science*, 63:261–269, 1992.

[34] S. Naranan and V. K. Balasubrahmanyan. Information theoretic model for frequency distribution of words and speech sounds (phonemes) in language. *Journal of Scientific and Industrial Research*, 52:728–738, 1993.

[35] R. Ferrer-i-Cancho. Decoding least effort and scaling in signal frequency distributions. *Physica A*, 345:275–284, 2005.

[36] C.-S. Liu. Maximal non-symmetric entropy leads naturally to Zipf's law. *Fractals*, 16(01):99–101, 2008.

[37] S. K. Baek, S. Bernhardsson, and P. Minnhagen. Zipf's law unzipped. *New Journal of Physics*, 13(4):043004, 2011.

[38] I. Moreno-Sánchez, F. Font-Clos, and A. Corral. Large-scale analysis of Zipf's law in English. *PLoS ONE*, 11:1–19, 01 2016.

[39] J. N. Kapur and H. K. Kesavan. *Entropy Optimization Principles and Their Applications*, pages 3–20. Springer Netherlands, Dordrecht, 1992.

[40] P. Harremoës and F. Topsøe. Maximum entropy fundamentals. *Entropy*, 3(3):191–226, 2001.

[41] G. A. Miller and N. Chomsky. Finitary models of language users. In R. D. Luce, R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, volume 2, pages 419–491. Wiley, New York, 1963.

[42] W. Li. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE T. Inform. Theory*, 38(6):1842–1845, 1992.

[43] R. Suzuki, P. L. Tyack, and J. Buck. The use of Zipf's law in animal communication analysis. *Anim. Behav.*, 69:9–17, 2005.

[44] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1:226–251, 2003.

[45] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[46] R. Ferrer-i-Cancho and R. Gavaldà. The frequency spectrum of finite samples from the intermittent silence process. *Journal of the American Association for Information Science and Technology*, 60(4):837–843, 2009.

[47] R. Ferrer-i-Cancho and B. Elvevåg. Random texts do not exhibit the real Zipf's-law-like rank distribution. *PLoS ONE*, 5(4):e9411, 2009.

[48] A. Rapoport. Zipf's law re-visited. In H. Guiter and M. V. Arapov, editors, *Quantitative linguistis: Studies on Zipf's law*, pages 1–28. Studienverlag Dr. N. Brockmeyer, Bochum, 1982.

[49] B. Sigurd, M. Eeg-Olofsson, and J. van Weijer. Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica*, 58(1):37–52, 2004.

[50] U. Strauss, P. Grzybek, and G. Altmann. Word length and word frequency. In P. Grzybek, editor, *Contributions to the science of text and language*, pages 277–294. Springer, Dordrecht, 2007.

[51] R. Hammerl. Länge - Frequenz, Länge - Rangnummer: Überprüfung von zwei lexikalischen Modellen. *Glottometrika*, 12:124, 1990.

[52] H. Guiter. Les relationes frequence - longueur - sens des mots (langues romanes et anglais). In *XIV Congresso Internazionale di linguistica e filologia romanza*, pages 373–381, Napoli, 1974.

[53] B. M. Bezerra, A. Souto, A. N. Radford, and G. Jones. Brevity is not always a virtue in primate communication. *Biology letters*, 7(1):23–25, 2011.

[54] R. Ferrer-i-Cancho, N. Forns, A. Hernández-Fernández, G. Bel-Enguix, and J. Baixeries. The challenges of statistical patterns of language: the case of Menzerath's law in genomes. *Complexity*, 18(3):11–17, 2013.

[55] Aadam. Minimum dot product. *https://medium.com/competitive/minimum-dot-product-62daa5281ba6*, 2016.

[56] R. Ferrer-i-Cancho. Euclidean distance between syntactically linked words. *Physical Review E*, 70:056135, 2004.

[57] R. Ferrer-i-Cancho and A. Díaz-Guilera. The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics*, page P06009, 2007.

[58] C. Bentz, D. Alikaniotis, M. Cysouw, and R. Ferrer-i-Cancho. The entropy of wordslearnability and expressivity across more than 1000 languages. *Entropy*, 19(6), 2017.

[59] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference. A practical information-theoretic approach*. Springer, New York, 2nd edition, 2002.

[60] S. Naranan and V. K. Balasubrahmanyan. Information theoretic models in statistical linguistics - Part II: Word frequencies and hierarchical structure in language. *Current Science*, 63:297–306, 1992.

[61] M. P. H. Stumpf and M. A. Porter. Critical truths about power laws. *Science*, 335(6069):665–666, 2012.

[62] G. Altmann. Prolegomena to Menzerath's law. *Glottometrika*, 2:1–10, 1980.

[63] M. L. Gustison, S. Semple, R. Ferrer i Cancho, and T. Bergman. Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences USA*, 113:E2750–E2758, 2016.

[64] M. G. Boroda and G. Altmann. Menzerath's law in musical texts. *Musikometrika*, 3:1–13, 1991.

[65] K. Shahzad, J.E. Mittenthal, and G. Caetano-Anollés. The organization of domains in proteins obeys Menzerath-Altmann's law of language. *BMC Systems Biology*, 9:1–13, 2015.