The global minima of the communicative energy of natural communication systems.

Ramon Ferrer i Cancho and Albert Díaz-Guilera

Departament de Física Fonamental, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain.

E-mail: `ramon.ferrericancho@gmail.com`

Abstract. Until recently, models of communication have explicitly or implicitly assumed that the goal of a communication system is just maximizing the information transfer between signals and 'meanings'. Recently, it has been argued that a natural communication system not only has to maximize this quantity but also has to minimize the entropy of signals, which is a measure of the cognitive cost of using a word. The interplay between these two factors, i.e. maximization of the information transfer and minimization of the entropy, has been addressed previously using a Monte Carlo minimization procedure at zero temperature. Here we derive analytically the globally optimal communication systems that result from the interaction between these factors. We discuss the implications of our results for previous studies within this framework. In particular we prove that the emergence of Zipf's law using a Monte Carlo technique at zero temperature in previous studies indicates that the system had not reached the global optimum.

Keywords: information theory, networks, least effort, cognitive cost, communication, language, Zipf's law

PACS numbers: 89.70.+c Information theory and communication theory, 87.53.Wz Monte Carlo applications

Submitted to: *Journal of Statistical Mechanics: Theory and Experiment*

1. Introduction

During the last years, the interest in the study of sound-meaning mappings from an analytical perspective has exploded (e.g. [1, 2, 3, 4, 5, 6, 7]). The majority of models study the evolution of sound-meaning mappings without worrying about the cognitive cost of using signals. It is known in psycholinguistics that the availability of a word is positively correlated with its frequency. Thus, the higher the frequency of a word, the lower its cost [8]. This phenomenon is known as the word frequency effect [9]. Imagine that we have a set of n signals $S = \{s_1, ..., s_i, ..., s_n\}$. In human language, the elements of S can be words. $H(S)$, the entropy of the set of signals S, has been proposed as a measure of cost of word use for both sender and receiver [10, 7]. By now, it is enough to know that $H(S)$ is a measure of disorder in the occurrence of signals, i.e. of how equally likely signals are. $H(S)$ takes its maximum value, $\log n$, when all signals are equally likely and takes its minimum value, 0 when only one signal has non-zero probability. When all signal are equally likely, we have the worst case for word availability because all words take the smallest frequency, i.e. $1/n$. When only one word is used (a single word has probability 1 and the rest have probability 0), we have the best case for word availability because one word has the the greatest availability and the rest are just simply not used. Independently, other entropies have been proposed for measuring the cost of linguistic units such as inflectional morphology [11] or words [8].

We assume that we have a general communication framework where signals are elicited by stimuli. The stimuli of our set of the signals S communicate about stimuli from a set of m stimuli $R = \{r_1, ..., r_j, ..., r_m\}$. In human language, the elements of R can be stimuli that elicit the words in S [12]. Stimuli could be objects or events. Animal behaviourists may prefer that R is the set of mental states triggering each signal.

A few of the large amount of the kind of studies mentioned above use the standard information theory framework, where the effectiveness of a communication system is measured using Shannon's information transfer. We define $I(S, R)$ as the Shannon information transfer between S and R. ‡ By now, it is enough to know that $I(S, R)$ is a non-negative function that measures the amount of information conveyed by signals in S about stimuli in R and vice versa [13].

A natural communication system must tend to maximize $I(S, R)$ to be communicatively effective and tend to reduce $H(S)$ due to word-frequency effects. A simple way of integrating this two communication factors is a linear combination though a single parameter λ that weight the contribution of each factor. This way, the function that a natural communication system should minimize can be written as

$$\Omega(\lambda) = -\lambda I(S, R) + (1 - \lambda)H(S). \tag{1}$$

The minimization of $\Omega(\lambda)$ has been studied numerically using a Monte Carlo algorithm at zero temperature in various models [14, 7]. The goal of the present article is studying analytically the global minima of $\Omega(\lambda)$ in these models for $\lambda \in [0, 1]$. In particular, this

‡ See Section 2 for a review of the definition of this standard information theory concept.

study aims to shed light on the nature of Zipf's law for word frequencies. Zipf's law states that the (relative) frequency of the i-th most frequent word in a text obeys [15]

$$P(i) \sim i^{-\alpha}, \tag{2}$$

where α is a constant, the so-called exponent of the law. In many real cases, $\alpha \approx 1$ although noticeable deviations from this value have been reported (see [16] for a review). Zipf's law for word frequencies has been obtained by minimizing $\Omega(\lambda)$ for a critical value of λ, λ^*, such that $\lambda^* \in [0, 1/2)$ using a Monte Carlo technique at zero temperature [14, 7]. We will show that Zipf's law indicates that the global minimum of $\Omega(\lambda)$ has not been reached.

The remainder of this article is organized as follows. Section 2 introduces the elementary entropies needed in this article and provides a general outline for studying the minima of $\Omega(\lambda)$. Section 3 introduces the family of models in which we will study the minima of $\Omega(\lambda)$. Section 4 gives the global minima of $\Omega(\lambda)$ for the two different models of the family mentioned before. Section 5 discusses the results with special emphasis on the implications for previous related work.

2. A quick review of information theory

We define $p(s_i)$ as the probability of s_i and $p(s_i|r_j)$ as the probability of producing s_i when r_j is given. We define $p(r_j)$ as the probability of r_j and $p(r_j|s_i)$ as the probability of interpreting r_j when s_i is given. The Shannon information transfer, $I(S, R)$, can be defined in two equivalent ways [13]. On the one hand,

$$I(S, R) = H(S) - H(S|R), \tag{3}$$

where

$$H(S) = -\sum_{i=1}^{n} p(s_i) \log p(s_i), \tag{4}$$

$$H(S|R) = \sum_{j=1}^{m} p(r_j) H(S|r_j), \tag{5}$$

and

$$H(S|r_j) = -\sum_{i=1}^{n} p(s_i|r_j) \log p(s_i|r_j). \tag{6}$$

On the other hand,

$$I(S, R) = H(R) - H(R|S), \tag{7}$$

where

$$H(R) = -\sum_{j=1}^{m} p(r_j) \log p(r_j), \tag{8}$$

$$H(R|S) = \sum_{i=1}^{n} p(s_i) H(R|s_i), \tag{9}$$

$$H(R|s_i) = -\sum_{j=1}^{m} p(r_j|s_i) \log p(r_j|s_i). \tag{10}$$

The model in [14] defines what constitutes an effort for the speaker and an effort for the hearer in $\Omega(\lambda)$. There, the function that a communication system has to minimize is

$$\Omega'(\lambda) = \lambda H(R|S) + (1-\lambda)H(S). \tag{11}$$

The minimization of $\Omega'(\lambda)$ is equivalent to the minimization of $\Omega(\lambda)$ when $H(R)$ is constant, which is the assumption of the model in [14]. To see this, we can write $\Omega(\lambda)$ as

$$\Omega(\lambda) = -\lambda H(R) + \lambda H(R|S) + (1-\lambda)H(S), \tag{12}$$

knowing $I(S,R) = H(R) - H(R|S)$.

It is argued in [14] that $H(R|S)$ is an effort for the hearer and $H(S)$ is an effort for the speaker. This issue needs to be clarified. $H(S)$ is both a source of effort for the speaker and the hearer because the word frequency effects concern both word production (e.g. through cues) [17, 18] and also recognition of spoken and written words [19, 20, 8]. For this reason, later articles referred to $H(S)$ as a measure of both effort for the speaker and the hearer [7, 10] although the confusion persists [21]. Besides $H(S)$, $H(S|R)$ within $I(S,R) = H(S) - H(S|R)$ is also a source of effort for the speaker. $H(S|R)$ is a measure of the effort of coding stimuli. Roughly speaking, $H(S|R)$ is a measure of the mean amount of candidate signals that the speaker has when a stimulus is given (recall Eq. 5). The less candidates there are, the easier the task of choosing a candidate signal. Besides $H(S)$, $H(R|S)$ within $I(S,R) = H(R) - H(R|S)$ is also a source of effort for the hearer. $H(R|S)$ is a measure of the effort of decoding signals. Roughly speaking again, $H(R|S)$ is a measure of amount of the mean amount of candidate stimuli that the hearer has when a signal is given (recall Eq. 9). The less candidates there are, the easier the task of interpreting the signal. In sum, there are actually two sources of effort for the speaker, i.e. $H(S)$ and $H(S|R)$, and two sources of effort for the hearer, i.e. $H(S)$ and $H(R|S)$ in our general definition of $\Omega(\lambda)$.

Now we focus on $\lambda \in [0,1]$ and aim to determine the kinds of minima that appear when $\Omega(\lambda)$ is minimized depending on λ. Here, by minima we mean the set of matrices of joint probability $p(s_i, r_j)$ such that $\Omega(\lambda)$ is a global minimum. Notice that once $p(s_i, r_j)$ is known for all signal-stimulus pairs, then we can obtain all the probabilities involved in the entropies needed for calculating $\Omega(\lambda)$. Recall that

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j), \tag{13}$$

$$p(r_j) = \sum_{i=1}^{n} p(s_i, r_j), \tag{14}$$

$p(s_i|r_j) = p(s_i, r_j)/p(r_j)$ and $p(r_j|s_i) = p(s_i, r_j)/p(s_i)$. Knowing that $I(S,R) = H(S) - H(S|R)$, we can write $\Omega(\lambda)$ in a more informative way

$$\Omega(\lambda) = (1-2\lambda)H(S) + \lambda H(S|R) \tag{15}$$

Using the previous equation, three different domains become obvious when minimizing $\Omega(\lambda)$:

(i) If $\lambda \in [0, 1/2)$, both $H(S)$ and $H(S|R)$ must be minimized. Since $H(S) \geq H(S|R)$ (equivalently, $I(S, R) \geq 0$ [22, 13]) and the minimum value of $H(S)$ and $H(S|R)$ is 0, it turns out that minimizing $H(S)$ implies minimizing $H(S|R)$. Thus, the minima of $\Omega(\lambda)$ when $\lambda \in [0, 1/2)$ are exactly the minima of just $H(S)$.

(ii) If $\lambda = 1/2$, only $H(S|R)$ has to be minimized.

(iii) If $\lambda \in (1/2, 1)$, $H(S)$ must be maximized and $H(S|R)$ must be minimized. The minima of $\Omega(\lambda)$ are the intersection of the minima of $H(S)$ and the minima of $H(S|R)$, if the intersection between minima is not empty (we will see that this is the case in the models studied here). It is easy to see that the minima of $\Omega(\lambda)$ when $\lambda \in (1/2, 1)$ are the maxima of $I(S, R) = H(S) - H(S|R)$.

In sum, the minima of $\Omega(\lambda)$ in the 1st, 2nd and 3rd domains are given by the minima of $H(S)$, $H(S|R)$ and the maxima of $I(S, R)$, respectively.

3. The family of models

In our general communication framework, links between signals and stimuli are defined by a binary matrix $A = \{a_{ij}\}$ where $a_{ij} = 1$ if s_i and r_j are linked and $a_{ij} = 0$ otherwise. A defines the structure of a communication system. i.e. the mapping of signals into stimuli. A matrix of this kind is the basis of different analytical [23, 24, 25, 5, 26, 27] and computational approaches [28, 29, 30, 7] to the evolution of language. We define the degree of s_i (i.e. the number of connections of s_i) as

$$\mu_i = \sum_{j=1}^{m} a_{ij}. \tag{16}$$

and the degree of r_j (i.e. the number of connections of r_j) as

$$\omega_j = \sum_{i=1}^{n} a_{ij}. \tag{17}$$

Here we focus on a family of probabilistic models that assumes that the probability that s_i is used for r_j is

$$p(s_i|r_j) = \frac{a_{ij}}{\omega_j}. \tag{18}$$

From Eq. 18 and the definition of conditional probability, we obtain

$$p(s_i, r_j) = p(s_i|r_j)p(r_j) = \frac{a_{ij}p(r_j)}{\omega_j} \tag{19}$$

and thus

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j) = \sum_{j=1}^{m} \frac{a_{ij}p(r_j)}{\omega_j}. \tag{20}$$

Applying the definition of conditional probability again we obtain

$$p(r_j|s_i) = \frac{p(s_i, r_j)}{p(s_i)} = \frac{a_{ij}p(r_j)}{\omega_j p(s_i)}. \tag{21}$$

Two models that stem from Eq. 18 are introduced in the next subsections.

3.1. Model A: $p(r_j) = \omega_j/M$

The models in [7, 16, 26, 31, 27, 5] assume that

$$p(r_j) = \frac{\omega_j}{M}, \tag{22}$$

where M is the total amount of connections, defined as

$$M = \sum_{j=1}^{m} \omega_j. \tag{23}$$

Assuming Eq. 22, Eqs. 19, 20 and 21 give, respectively,

$$p(s_i, r_j) = \frac{a_{ij}}{M}, \tag{24}$$

$$p(s_i) = \frac{\mu_i}{M}, \tag{25}$$

and

$$p(r_j|s_i) = \frac{a_{ij}}{\mu_i}. \tag{26}$$

3.2. Model B: $p(r_j) = 1/m$

The model in [14] assumes that $p(r_j)$ is independent of A and fixed a priori. Here we focus on a particular case: $p(r_j) = 1/m$. $p(r_j) = 1/m$ is chosen for various reasons: (a) simplicity (b) it is a sort of worst case for the occurrence of stimuli (the uncertainty about the stimulus that could appear next is maximum) and (c) as far as we know, this is the only assumption made by models assuming that $p(r_j)$ is fixed a priori (equally likely stimuli is the assumption explicitly made in the model in [14] and also implicitly made in the model in [1]; the latter is explained in Appendix D). Assuming $p(r_j) = 1/m$, Eqs. 19, 20 and 21 give, respectively,

$$p(s_i, r_j) = \frac{a_{ij}}{m\omega_j}, \tag{27}$$

$$p(s_i) = \frac{b_i}{m}, \tag{28}$$

and

$$p(r_j|s_i) = \frac{a_{ij}}{b_i\omega_j}, \tag{29}$$

where

$$b_i = \sum_{k=1}^{m} \frac{a_{ik}}{\omega_k}. \tag{30}$$

	Model A: $p(r_j) = \omega_j/M$	Model B: $p(r_j) = 1/m$ (with $\omega_j \geq 1$)
$H(S,R)$	$\log M$	$\frac{1}{m}\sum_{j=1}^{m} \frac{\log(m\omega_j)}{\omega_j}$
$H(R\|S)$	$\frac{1}{M}\sum_{i=1}^{n} \mu_i \log \mu_i$	$\frac{1}{m}\sum_{i=1}^{n} b_i H(R\|s_i)$
$H(R\|s_i)$	$\log \mu_i$	$\log b_i + \frac{1}{b_i}\sum_{j=1}^{m} \frac{a_{ij}}{\omega_j} \log \omega_j$
$H(S\|R)$	$\frac{1}{M}\sum_{j=1}^{m} \omega_j \log \omega_j$	$\frac{1}{m}\sum_{j=1}^{m} \log \omega_j$
$H(S\|r_j)$	$\log \omega_j$	$\log \omega_j$
$H(S)$	$H(S,R) - H(R\|S)$	$\log M - \frac{1}{M}\sum_{i=1}^{n} \mu_i \log \mu_i$
$H(R)$	$\log M - \frac{1}{M}\sum_{j=1}^{m} \omega_j \log \omega_j$	$\log m$

Table 1. Summary of results about the definition of various entropies for models A ($p(r_j) = \omega_j/M$) and B ($p(r_j) = 1/m$ with $\omega_j \geq 1$). S and R are, respectively, the set of signals and the set of stimuli. $H(S,R)$ is the joint entropy of S and R. $H(R|S)$ is the conditional entropy of R when S is known and $H(S|R)$ is the conditional entropy of S when R is known. $H(S)$ and $H(R)$ are, respectively, the entropy of S and R. $b_i = \sum_{k=1}^{m} a_{ik}/\omega_k$

3.3. Remarks about both models

With the probabilities of models A and B and the general definitions of the entropies (recall the beginning of Section 2) it is easy to calculate all the necessary entropies. See Table 1 for a summary of the specific form of the entropies that can be easily obtained after some algebra for each model.

It is important to notice that Eq. 18 is undetermined, i.e. $p(s_i|r_j) = 0/0$, when $\omega_j = 0$. The consequences of this indetermination depend on the kind of model. In practice, the indetermination has no consequence for the calculation of $I(S,R)$ and $H(S)$ when $p(r_j) \sim \omega_j$ (recall Table 1). In contrast, various technical problems arise when $p(r_j)$ is fixed a priori. For this reason, $\omega_j > 0$ was imposed in the model in [14].

4. The global minima of $\Omega(\lambda)$

Here we show the minima of $\Omega(\lambda)$ for the various domains of λ specified in Section 2. By minima we mean the set of matrices A for which $\Omega(\lambda)$ is minimum. For the sake of clarity, this section is essentially an enumeration of the minimum energy configurations for models A and B and the relevant domains of λ (the reader interested in more details is referred to Appendices A-C).

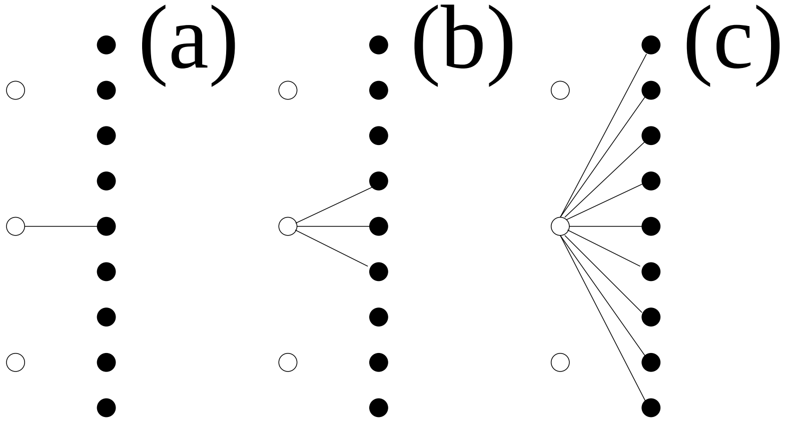

Figure 1. Some mappings between signals (white circles) and stimuli (black circles) that are minima of $H(S)$ and $H(S|R)$ with $n = 3$ signals and $m = 9$ stimuli. (a-c) are minima of model A while (c) is the only valid minima of model B.

4.1. The global minima of $H(S)$ ($\lambda \in [0, 1/2)$) §

The signal-stimulus mappings minimizing $H(S)$ for model A ($p(r_j) = \omega_j/M$) are those where

- All signals are unlinked except one.
- The only linked signal can have any degree (between 1 and m).

As for model B ($p(r_j) = 1/m$), the signal-stimulus mappings minimizing $H(S)$ are those where

- All signals are unlinked except one.
- The only linked signal must be connected to all stimuli.

Some signal-stimuli mappings minimizing $H(S)$ for model A are shown in Figure 1. As for model B, a minimal mapping is shown in Figure 1 (c). The mappings in Figs. 1 (a) and (b) are not minimal mappings of model B because they violate the constraint of not having disconnected signals. Notice that a system with the minimum $H(S)$ (i.e. $H(S) = 0$) cannot communicate using individual signals because the information transfer $I(S, R)$ is also zero (recall $I(S, R) = H(S) - H(S|R)$ and $I(S, R), H(S|R) \geq 0$ or see Appendix A for further details).

4.2. The global minima of $H(S|R)$ ($\lambda = 1/2$) ‖

The signal-stimulus mappings minimizing $H(S|R)$ for model A ($p(r_j) = \omega_j/M$) are the mappings in which stimuli can only be disconnected or have a single link. As for model B ($p(r_j) = 1/m$ with $\omega_j \geq 1$), the minimal mappings are those where all stimuli have only one link. Some signal-stimuli mappings minimizing $H(S|R)$ for model A are shown

§ See Appendix A for the details.

‖ See Appendix B for the details.

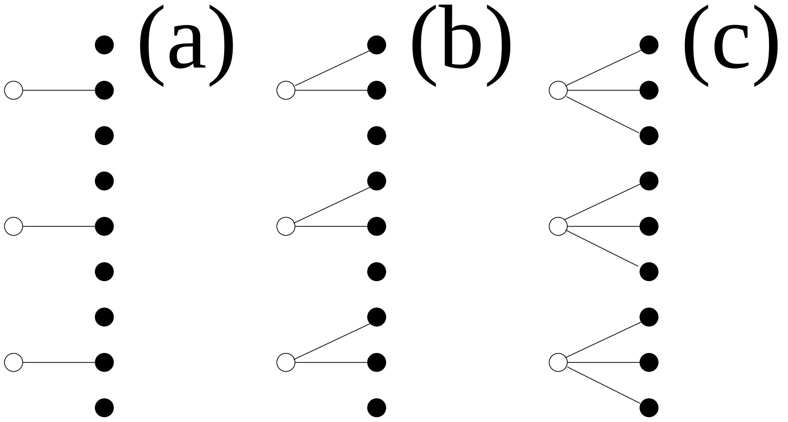

Figure 2. Some mappings between signals (white circles) and stimuli (black circles) that achieve maximum $I(S,R)$ with $n=3$ signals and $m=9$ stimuli. This mappings also achieve minimum $H(S|R)$.

in Figs. 1 and 2. As for model B, a minimal mapping is shown in Figure 1 (c) (the mappings in Figs. (a) and (b) are not valid minima of model B because they have disconnected stimuli).

4.3. The global minima of $I(S,R)$ ($\lambda \in (1/2,1]$) ¶

The signal-stimulus mappings maximizing $I(S,R)$ for model A are those in which

- All signals have the same amount of connections but are not disconnected.
- Stimuli have at most one link.

As for model B with $n \geq m$, the mappings maximizing $I(S,R)$ are those in which

- Signals have at most one link (there must be at least one link).
- There are no disconnected stimuli.

As for model B with $n \geq m$ and n/m is rational, the mapping maximizing $I(S,R)$ are those in which

(i) All signals have the same amount of connections.

(ii) All stimuli have one link.

In particular, the global minima are one-to-one mappings for models A and B when $n=m$ (Figure 3). Figure 2 shows examples of mappings between signals and stimuli that maximize $I(S,R)$ for model A ($p(r_j)=\omega_j/M$). As for model B ($p(r_j)=1/m$), a minimal mapping is shown in Figure 2 (c). Notice that $I(S,R)$ can be maximum even if signals have more than one connection. Examples of mappings between signals and stimuli maximizing $I(S,R)$ for $n \geq m$ can be obtained from Figure 2 and changing signals by stimuli and vice versa (exchanging white circles with black circles and vice versa).

¶ See Appendix C for the details.

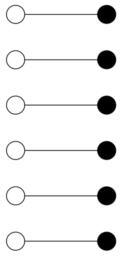

Figure 3. A one-to-one mapping between $n = 6$ signals (white circles) and $m = 6$ stimuli (black circles). This configuration achieves maximum $I(S, R)$.

5. Discussion

We have found that the global minimum of $\Omega(\lambda)$ are degenerate (in the physics sense) because there is more than one signal-stimulus mapping achieving the minimum energy. For instance, three different configurations with minimum energy for $\lambda \in [0, 1/2]$ are shown in Figure 1. Moreover, (c), for instance, can be transformed into a different mapping by swapping the central signal by the other signals while $\Omega(\lambda)$ remains the same.

Our formal approach to maximizing $I(S, R)$ has produced results that are against common intuitions about the effect of maximizing $I(S, R)$. We have seen that maximum $I(S, R)$ does not exclude the presence of ambiguous signals (signals with non-zero degree) when $n < m$ (recall Figure 2 B or C). In other words, maximizing the information transfer does not imply absence of signal ambiguity. Third, we have seen that making $H(S) = 0$ (one aspect of the cost of word use) and communication is a contradiction of terms in our models (recall that $I(S, R) = H(S) - H(S|R)$ and $I(S, R), H(S|R) \geq 0$ or see Appendix A for the details). Thus, it is impossible that word use is costless in our models.

Our study has implications for previous related work. Zipf's law for word frequencies had been obtained by minimizing $\Omega(\lambda)$ for a critical value of λ, λ^*, such that $\lambda^* \in [0, 1/2)$ using a Monte Carlo algorithm at zero temperature [14, 7]. The models in [14] and [7] reproduce Zipf's law (recall Eq. 2) with α close to 1 (for sufficiently large m). We have seen that the global minima of $\Omega(\lambda)$ for $\lambda \in [0, 1/2)$ give only one signal with non-zero probability, i.e. $\alpha \to \infty$. The analytical results of this article indicate that the finding of Zipf's law (with α close to 1) using a a Monte Carlo technique at zero temperature is not a global optimum. The absence of a temperature in these numerical minimization suggests that Zipf's law with a non-extremal exponent could be the consequence of local minima of $\Omega(\lambda)$. The fact that the Monte Carlo algorithm does not find the global optimum is not against the utility of this technique for understanding human language. Assuming that $\Omega(\lambda)$ is a psycholinguistically well-motivated function, reaching the global optimum ($H(S) = 0$) is problematic: communication is impossible because $H(S) = 0$ leads to $I(S, R) = 0$ as explained in this article. Thus, the need of communicating (the need of $I(S, R) > 0$) may be a serious obstacle for human language reaching the global optimum. Nonetheless, we do not mean that the reason why human language cannot apparently reach the global minimum is exactly the need of communication. For instance, the procedure that humans use for minimizing $\Omega(\lambda)$ may naturally prevent the system from reaching the global optimum, as suggested by the emergence of Zipf's law using the Monte Carlo technique.

Another implication of our study concerns a recent article where Solé and colleagues argue that the minimum cost of word use "is obtained when a single word refers to many objects" [21]. Put in our terms, they mean that the minimum signal entropy use is obtained when a single signal is connected with many stimuli. The problem is that Solé *et al.* are not covering all the configurations where the cost of communication is minimum. We have seen that a single signal connected with a few stimuli also achieves minimum $H(S)$ (recall Section 4) in model A. Eventually, a single signal with one connection (and the rest of the signals disconnected) still achieves the minimum cost of communication. If Solé *et al.* actually refer to the minimum cost of word use in model B (where disconnected stimuli are not allowed), we have seen in this case (Appendix A) that the minimum is not achieved when a single signal is connected with many stimuli but with exactly all stimuli.

There is another aspect of the model in [14] that needs to be reconsidered: the statement that animal communication systems (except human language) should behave according to $\lambda > \lambda^*$, which is equivalent to $\lambda \geq 1/2$ when looking for the global optima. The are two reasons for thinking this statement does not stand. First, the pioneering work by McCowan and collaborators [32, 33] showed that the vocalizations of dolphins and other species exhibit a frequency distribution consistent with Zipf's for word frequencies. Although these findings have been the subject of an open debate [34, 35], at present it cannot be categorically stated that the frequency distribution of others species is consistent with that of $\lambda \geq 1/2$, where all signals must be equally likely. Second, it is hard to imagine that the brains of other species do not need to worry about minimizing $H(S)$ due to cognitive pressures. The only way of getting rid of this cognitive pressures is, as argued in [14], having a small repertoire of signals. The point is: how small should it be in order to scape from this cognitive pressures?

In sum, we need to reflect about the models in [14, 7] to the light of the global minima and other aspects discussed in this article. One of the most important questions that the findings in this article raise is: assuming that the rationale behind $\Omega(\lambda)$ minimization is essentially correct, why do natural communications not reach the global minimum?

Acknowledgments

We are grateful to P. Cermeli and G. Zanzotto for helpful discussions. We thank F. Moscoso del Prado Martín for pointers to the literature on the cognitive cost of linguistic elements. This work was supported by the projects FIS2006-13321-C02 and BFM2003-08258-C02-02 of the Spanish Ministry of Education and Science. This work was funded by a Juan de la Cierva contract from the Spanish Ministry of Education and Science (RFC).

Appendix A. The minima of the entropy of signals

First, we study the consequences of minimum $H(S)$. We will show that systems that minimize $H(S)$ alone cannot communicate, more precisely, $H(S) = 0$ implies $I(S,R) = 0$. To see it, consider that the minimum value that $H(S)$ can take is 0 [13]. Knowing that $I(S,R) = H(S) - H(S|R)$ and $I(S,R), H(S), H(S|R) \geq 0$, it follows that $I(S,R) = 0$ when $H(S) = 0$.

We define n_+ as the number of signals such that $p(s_i) \neq 0$. We will show that $H(S)$ is minimum (i.e. $H(S) = 0$) if and only if $n_+ = 1$, i.e. only one signal s_h satisfies $p(s_h) = 1$ and the remaining signals have probability zero. Knowing

- $H(S) \geq 0$ [13],
- Equation 4,
- $-x \log x \geq 0$ if $x \in \{0, 1\}$,
- $x \log x = 0$ if and only if $x \in \{0, 1\}$,

it follows that the signal probabilities giving $H(S) = 0$ need $p(s_i) \in \{0, 1\}$ for each $1 \leq i \leq n$. Adding the constraint

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

the only signal probabilities giving $H(S) = 0$ turn out to be those where there is a single signal s_h that satisfies $p(s_h) > 0$ and the remaining signals have probability zero (i.e. $p(s_i) = 0$ for $i \neq h$), i.e. $n_+ = 1$

Second, we present the minima of $H(S)$ for models A and B together. We assume that $M \geq 1$ and both n and m are finite. We will show that A minimizes $H(S)$ if and only if there is a single linked signal (recall that model B adds a further constraint from its definition: unlinked stimuli are not allowed). To see it, we proceed in too steps. We will start by showing that that within this family of models, the only way a signal can have probability zero is by being disconnected ($p(s_i) = 0$ if and only if $\mu_i = 0$). As for model A (where $p(r_j)$ is no fixed a priori), we have that $p(s_i) = \mu_i/M$, hence $p(s_i) = 0$ if and only if $\mu_i = 0$. As for model B (where all stimuli are equally likely), we have that

$$p(s_i) = \sum_{j=1}^{m} \frac{a_{ij}}{\omega_j} p(r_j) = \frac{1}{m} \sum_{j=1}^{m} \frac{a_{ij}}{\omega_j}, \tag{A.2}$$

hence $p(s_i) = 0$ if and only if $\mu_i = 0$ again. Therefore, knowing that $H(S)$ is minimum (i.e. $H(S) = 0$) if and only if $n_+ = 1$ (see above), it follows for model A that the minima of $H(S)$ are achieved only when there is a single connected signal s_h (s_h can have any degree within $[1, m)$). As for model B, the constraint $\omega_j \geq 1$ implies that the minima of $H(S)$ are those where there is a single connected signal s_h, such that $\mu_h = m$.

Appendix B. The minima of the conditional entropy of signals

We assume that $M \geq 1$ and both n and m are finite. First, we will show that A minimizes $H(S|R)$ in model A ($p(r_j) = \omega_j/M$) if and only if stimuli have at most one link, i.e. $\omega_j \in \{0, 1\}$ for $1 \leq j \leq m$. To see it, consider that $H(S|R)$ can be written as (Table 1)

$$H(S|R) = \frac{1}{M} \sum_{j=1}^{m} \omega_j \log \omega_j \tag{B.1}$$

assuming that $p(r_j) = \omega_j/M$. Given Eq. B.1, $H(S|R) = 0$ if and only if $\omega_j \in \{0, 1\}$ for $1 \leq j \leq m$, as we wanted to prove.

Second, we will show that A minimizes $H(S|R)$ in model B ($p(r_j) = 1/m$ with $\omega_j \geq 1$) if and only if stimuli have one link, i.e. $\omega_j = 1$ for $1 \leq j \leq m$. To see it, consider that $H(S|R)$ can be written as (Table 1)

$$H(S|R) = \frac{1}{m} \sum_{j=1}^{m} \log \omega_j \tag{B.2}$$

assuming that $p(r_j) = 1/m$. Given Eq. B.2 and the initial assumption $\omega_j \geq 1$, $H(S|R) = 0$ if and only if $\omega_j = 1$ for $1 \leq j \leq m$, as we wanted to prove.

Appendix C. The maxima of information transfer

First, we will bound $I(S, R)$ above. It is easy to see that $I(S, R) \leq min(H(S), H(R))$. Knowing that [13]

- $I(S, R) = H(S) - H(S|R) = H(R) - H(R|S)$,
- $I(S, R) \geq 0$
- $H(S|R), H(R|S) \geq 0$

we obtain

$$I(S, R) \leq H(S) \tag{C.1}$$

from $I(S, R) = H(S) - H(S|R)$ and

$$I(S, R) \leq H(R) \tag{C.2}$$

from $I(S, R) = H(R) - H(R|S)$. Mixing Eq. C.1 and Eq. C.2 we obtain

$$I(S, R) \leq min(H(S), H(R)). \tag{C.3}$$

From the previous inequality it easily follows that $I(S, R) \leq \log min(n, m)$, knowing that $H(S) \leq n$ and $H(R) \leq \log m$ [13].

Second, we study the mappings of signals and stimuli maximizing $I(S, R)$ for the models A and B. We follow the same steps in both cases. We study the cases $n \leq m$ and then $n \geq m$ separately. We assume $M \geq 1$ and both n and m are finite.

Appendix C.1. Model A: stimulus probability proportional to stimulus degree

First, we consider the case $n \leq m$. We will show that A maximizes $I(S,R)$ if and only if

(i) All signals have the same amount of connections within a particular range, more precisely, $\mu_i = K_\mu$ with $1 \leq K_\mu \leq \lfloor m/n \rfloor$ for $1 \leq i \leq n$.

(ii) Stimuli have at most one link, i.e. $\omega_j \in \{0,1\}$ for $1 \leq j \leq m$.

To see it, consider that $n \leq m$ implies that $I(S,R)$ cannot exceed $\log n$ (recall $I(S,R) \leq \log min(n,m)$). Hence, $I(S,R)$ is maximized according to $I(S,R) = H(S) - H(S|R)$ when $H(S) = \log n$ and $H(S|R) = 0$, knowing $H(S) \leq n$ and $H(S|R) \geq 0$. On the one hand, we have seen in Appendix B that $H(S|R) = 0$ is achieved if and only if $\omega_j \in \{0,1\}$ for $1 \leq j \leq m$. Thus, $M \leq m$. On the other hand, $H(S) = \log n$ if and only if all signals are equally likely. Knowing that $p(s_i) = \mu_i/M$ (Eq. 25), all signals are equally likely if and only if $\mu_i = K_\mu$, where K_μ is a constant such that $K_\mu \in [1, m]$. Knowing that

$$\sum_{i=1}^{n} p(s_i) = 1 \tag{C.4}$$

and Eq. 20, we obtain

$$K_\mu \geq 1. \tag{C.5}$$

$\omega_j \in \{0,1\}$ for $1 \leq j \leq m$ gives $M \leq m$. Replacing $M = nK_\mu$ into $M \leq m$ we obtain $K_\mu \leq m/n$. Knowing that μ_i and therefore K_μ are natural numbers, a tighter upper bound for K_μ that still preserves $H(S) = \log n$ (and compatible with $H(S|R) = 0$) is given by $\lfloor m/n \rfloor$. Therefore, $1 \leq K_\mu \leq \lfloor m/n \rfloor$, as we wanted to prove.

Second, we consider the case $n \geq m$. We will show that A maximizes $I(S,R)$ if and only if

(i) All stimuli have the same amount of connections within a particular range, more precisely, $\omega_j = K_\omega$ with $1 \leq K_\omega \leq \lfloor n/m \rfloor$ for $1 \leq j \leq m$.

(ii) Signals have at most one link, i.e. $\mu_i \in \{0,1\}$ for $1 \leq j \leq n$.

The proof is analogous to that of the case $n \leq m$. If $m \geq n$ then the fact that $I(S,R) \leq \log min(n,m)$ implies that the maximum $I(S,R)$ cannot exceed $\log m$. Hence, $I(S,R)$ is maximized according to $I(S,R) = H(R) - H(R|S)$ when $H(R) = \log m$ and $H(R|S) = 0$, knowing $H(R) \leq m$ and $H(R|S) \geq 0$. On the one hand, $H(R|S)$ can be written as (recall Table 1)

$$H(R|S) = \frac{1}{M} \sum_{i=1}^{n} \mu_i \log \mu_i \tag{C.6}$$

assuming $p(r_j) = \omega_j/M$ (Eq. 22). Given Eq. C.6, $H(R|S) = 0$ if and only if $\mu_i \in \{0,1\}$ for $1 \leq i \leq m$. Thus, $M \leq n$. On the other hand, $H(R) = \log m$ if and only if all stimuli are equally likely. Given $p(r_j) = \omega_j/M$, all stimuli are equally likely if and only if $\omega_i = K_\omega$, where K_ω is a constant. Knowing that

$$\sum_{j=1}^{m} p(r_j) = 1 \tag{C.7}$$

and $p(r_j) = \omega_j/M$, we obtain

$$K_\omega \geq 1. \tag{C.8}$$

Replacing $M = mK_\omega$ into $M \leq n$ we obtain $K_\omega \leq n/m$. Knowing that ω_i and therefore K_ω are natural numbers, a tighter upper bound for K_ω that preserves $H(R) = \log m$ (and compatible with $H(R|S) = 0$) is given by $\lfloor n/m \rfloor$. Therefore, $1 \leq K_\omega \leq \lfloor n/m \rfloor$, as we wanted to prove.

Appendix C.2. Model B: stimulus probability fixed a priori

We define x mod y as the remainder of the division of x by y. First, we consider the case $n \leq m$. For simplicity, it is convenient to assume m mod $n = 0$ in for deriving the maxima when $n \leq m$. In this case, we will show that A maximizes $I(S, R)$ if and only if

(i) All signals have the same amount of connections, more precisely, $\mu_i = m/n$ for $1 \leq i \leq n$.

(ii) All stimuli have one link, i.e. $\omega_j = 1$ for $1 \leq j \leq m$.

To see it, remember that the maximum $I(S, R)$ cannot exceed $\log n$ when $n \leq m$. Hence, $I(S, R)$ is maximized according to $I(S, R) = H(S) - H(S|R)$ when $H(S) = \log n$ and $H(S|R) = 0$, knowing $H(S) \leq n$ and $H(S|R) \geq 0$. On the one hand, we have seen in Appendix B that $H(S|R) = 0$ if and only if stimuli have one link, i.e. $\omega_j = 1$ for $1 \leq j \leq m$. On the other hand, $H(S) = \log n$ if and only if all signals are equally likely. Knowing Eq. 20 and $\omega_j = 1$, all signals are equally likely if and only if

$$\sum_{j=1}^{m} \frac{a_{ij} p(r_j)}{\omega_j} = 1/n. \tag{C.9}$$

Replacing the assumption $p(r_j) = 1/m$ and the requirement $\omega_j = 1$ (imposed by $H(S|R) = 0$) into Eq. C.9, we obtain

$$\mu_i = m/n. \tag{C.10}$$

The assumption m mod $n = 0$ warrants that the quotient m/n provides a degree that is a natural number, as expected for μ_i, as we wanted to prove.

Second, we consider the case $n \geq m$. We will show that A maximizes $I(S, R)$ if and only if signals have at most one link, i.e. $\mu_i \in \{0, 1\}$ for $1 \leq j \leq n$. The proof is similar to that of the case $n \leq m$. If $n \geq m$ then the fact that $I(S, R) \leq \log min(n, m)$ implies that the maximum $I(S, R)$ cannot exceed $H(R) = \log m$. Hence, $I(S, R)$ is maximized according to $I(S, R) = H(R) - H(R|S)$ when $H(R) = \log m$ and $H(R|S) = 0$, knowing

$H(R) \leq m$ and $H(R|S) \geq 0$. On the one hand, we already have that $H(R) = \log m$ because $p(r_j) = 1/m$. On the other hand, $H(R|S)$ can be written as (recall Table 1)

$$H(R|S) = \frac{1}{M} \sum_{i=1}^{n} \mu_i \log \mu_i \tag{C.11}$$

assuming Eq. 22. Given Eq. C.11, $H(R|S) = 0$ if and only if $\mu_i \in \{0, 1\}$ for $1 \leq i \leq m$, as we wanted to proof.

Finally, we will show that $I(S, R)$ is maximum if and only if A defines a one-to-one mapping between signals and stimuli in both model A ($p(r_j) = \omega_j/M$) and model B ($p(r_j) = 1/m$) when $n = m$. To see it, consider that maximum $I(S, R)$ implies that the degree of each signal and each stimulus must be one when $n = m$ according to the results obtained within this section. For this reason, the mapping between signals and stimuli must be one-to-one, as we wanted to prove.

Appendix D. Implicit equally likely stimuli.

Here we show that the evolution of language model in [1] makes assumptions consistent with $p(r_j) = 1/m$ for each stimulus. In this model, each agent is endowed with a speaking matrix $P = \{p_{ji}\}$ and a listening matrix $Q = \{q_{ij}\}$. p_{ji} is the probability that the speaker of a conversation uses utterance i for referring to meaning j. q_{ij} is the probability that the hearer of a conversation understands meaning j after hearing utterance i. p_{ji} in this model is equivalent to our $p(s_i|r_j)$ whereas q_{ij} is equivalent to our $p(r_j|s_i)$. Our notation makes explicit that the speaking and hearing matrices contain conditional probabilities. First, we will show how the speaking and hearing matrices are coupled through the definition of conditional probability and then we will show that the coupling used in [1] is a special case of the former coupling assuming $p(r_j) = 1/m$.

If we start from $p(s_i|r_j)$, the definition of conditional probability gives

$$p(s_i, r_j) = p(s_i|r_j)p(r_j). \tag{D.1}$$

The definition of conditional probability also gives

$$p(r_j|s_i) = \frac{p(s_i, r_j)}{p(s_i)}. \tag{D.2}$$

Replacing Eq. D.1 into Eq. D.2, we obtain

$$p(r_j|s_i) = \frac{p(r_j)}{p(s_i)} p(s_i|r_j). \tag{D.3}$$

and

$$p(s_i|r_j) = \frac{p(s_i)}{p(r_j)} p(r_j|s_i). \tag{D.4}$$

In [1], the hearing matrix is calculated from the speaking matrix through the formula (see caption of Figure 2 in [1]):

$$q_{ij} = \frac{p_{ji}}{\sum_j p_{ji}}, \tag{D.5}$$

which can be written as

$$p(r_j|s_i) = \frac{p(s_i|r_j)}{\sum_{k=1}^{m} p(s_i|r_k)} \tag{D.6}$$

using our notation.

Now we will show that Eq. D.6 is a special case of the coupling in Eq. D.3. We have seen above that the coupling between speaking and hearing matrices involves an iterative application of the definition of conditional probability which is reminiscent of the chain rule for derivatives. Replacing Eq. D.1 into

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j) \tag{D.7}$$

we obtain

$$p(s_i) = \sum_{j=1}^{m} p(s_i|r_j)p(r_j). \tag{D.8}$$

Replacing the previous equation into Eq. D.3 we obtain

$$p(r_j|s_i) = \frac{p(r_j)}{\sum_{k=1}^{m} p(s_i|r_k)p(r_k)} p(s_i|r_j). \tag{D.9}$$

Eq. D.6 is obtained when $p(r_j) = 1/m$, that is, when all meanings are equally likely. The assumptions behind Eq. D.6 are not explained in [1].

References

[1] M. A. Nowak and D. C. Krakauer. The evolution of language. *Proc. Natl. Acad. Sci. USA*, 96:8028–8033, July 1999.

[2] M. A. Nowak, D. C. Krakauer, and A. Dress. An error limit for the evolution of language. *Proc. R. Soc. London B*, 266:2131–2136, 1999.

[3] M. A. Nowak, J. B. Plotkin, and V. A. Jansen. The evolution of syntactic communication. *Nature*, 404:495–498, 2000.

[4] J. B. Plotkin and M. A. Nowak. Major transitions in language evolution. *Entropy*, 3:227–246, 2001.

[5] R. Ferrer i Cancho. Decoding least effort and scaling in signal frequency distributions. *Physica A*, 345:275–284, 2005. doi:10.1016/j.physa.2004.06.158.

[6] N. Komarova and P. Niyogi. Optimizing the mutual intelligibility of linguistic agents in a shared world. *Artificial Intelligence*, 154:1–42, 2004.

[7] R. Ferrer i Cancho. Zipf's law from a communicative phase transition. *Eur. Phys. J. B*, 47:449–457, 2005.

[8] S. A. McDonald and R. C. Shillcock. Rethinking the word frequency effect: the neglected role of dsitributional information in lexical processing. *Language and Speech*, 44(3):295–323, 2001.

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

[10] R. Ferrer i Cancho. On the universality of zipf's law for word frequencies. In P. Grzybek and R. Köhler, editors, *Exact methods in the study of language and text. To honor Gabriel Altmann*, pages 131–140. Gruyter, Berlin, 2006.

[11] F. Moscoso del Prado Martín, Alexandar Kostić, and R. H. Baayen. Putting the bits together: an information theoretical perspective on morphological processing. *Cognition*, 94:1–18, 2004.

[12] F. Pulvermuller. Brain reflections of words and their meaning. *Trends in Cognitive Sciences*, 5(12):517–524, 2001.

[13] R. B. Ash. *Information Theory.* John Wiley & Sons, New York, 1965.

[14] R. Ferrer i Cancho and R. V. Solé. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. USA*, 100:788–791, 2003.

[15] G. K. Zipf. *Human behaviour and the principle of least effort. An introduction to human ecology.* Hafner reprint, New York, 1972. 1st edition: Cambridge, MA: Addison-Wesley, 1949.

[16] R. Ferrer i Cancho. The variation of Zipf's law in human language. *Eur. Phys. J. B*, 44:249–257, 2005.

[17] R. C. Oldfield and A. Wingfield. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, 17:273–281, 1965.

[18] A. S. Brown. A review of the tip-of-the-tongue experience. *Psychological Bulletin*, 109:204–223, 1991.

[19] S. Monsell. The nature and the locus of word frequency effects in reading. In D. Besner and G. W. Humphreys, editors, *Basic processes in reading: visual word recognition.* LEA, London, 1991.

[20] C. M. Connine, J. Mullennix, E. Shernoff, and J. Yelen. Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 16:1084–1096, 1990.

[21] R. V. Solé, B. Corominas Murtra, Sergi Valverde, and L. Steels. Language network: their structure, function and evolution. *Santa Fe Working paper 05-12-042*, 2005.

[22] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423623–656, 1948.

[23] D. Lewis. *Convention: a philosophical study.* Harvard University Press, Cambridge, MA, 1969.

[24] M. A. Nowak. Evolutionary biology of language. *Phil. Trans. R. Soc. Lond. B*, 355:1615–1622, 2000.

[25] N. Komarova and M. A. Nowak. The evolutionary dynamics of the lexical matrix. *Bulletin of Mathematical Biology*, 63:451–484, 2001.

[26] R. Ferrer i Cancho, O. Riordan, and B. Bollobás. The consequences of Zipf's law for syntax and symbolic reference. *Proc. R. Soc. Lond. Series B*, 272:561–565, 2005.

[27] R. Ferrer i Cancho. When language breaks into pieces. a conflict between communication through isolated signals and language. *Biosystems*, 84:242–253, 2006.

[28] J. Hurford. Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, 77:187–222, 1989. doi:10.1016/0024-3481(89)90015-6.

[29] L. Steels. Self-organizing vocabularies. In C. Langton, editor, *Proceedings of Alife V*, Nara Japan., 1996.

[30] L. Steels. Language games for autonomous robots. *IEEE Intelligent systems*, 16:16–22, 2001.

[31] R. Ferrer i Cancho. Hidden communication aspects inside the exponent of Zipf's law. *Glottometrics*, 11:96–117, 2005.

[32] B. McCowan, S. F. Hanser, and L. R. Doyle. Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Anim. Behav.*, 57:409–419, 1999.

[33] B. McCowan, L. R. Doyle, and S. F. Hanser. Using information theory to assess the diversity, complexity and development of communicative repertoires. *Journal of Comparative Psychology*, 116:166–172, 2002.

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

[35] B. McCowan, L. R. Doyle, J. M. Jenkins, and S. F. Hanser. The appropriate use of Zipf's law in animal communication studies. *Anim. Behav.*, 69:F1–F7, 2005.