The Origins of Zipf's Meaning-Frequency Law

Ramon Ferrer-i-Cancho (iD)

Complexity and 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. E-mail: rferrericancho@cs.upc.edu

Michael S. Vitevitch (iD)

Spoken Language Laboratory, Department of Psychology, University of Kansas, 1415 Jayhawk Blvd., Lawrence, KS 66045, USA. E-mail: mvitevit@ku.edu

In his pioneering research, G.K. Zipf observed that more frequent words tend to have more meanings, and showed that the number of meanings of a word grows as the square root of its frequency. He derived this relationship from two assumptions: that words follow Zipf's law for word frequencies (a power law dependency between frequency and rank) and Zipf's law of meaning distribution (a power law dependency between number of meanings and rank). Here we show that a single assumption on the joint probability of a word and a meaning suffices to infer Zipf's meaning-frequency law or relaxed versions. Interestingly, this assumption can be justified as the outcome of a biased random walk in the process of mental exploration.

Introduction

G.K. Zipf (1949) investigated many statistical regularities of language. Some of them have been investigated intensively, such as Zipf's law for word frequencies (Fedorowicz, 1982; Ferrer-i-Cancho, & Gavaldà, 2009; Font-Clos, Boleda, & Corral, 2013; Ferrer-i-Cancho, 2016a) or Zipf's law of abbreviation (Strauss, Grzybek, & Altmann, 2006; Ferrer-i-Cancho et al., 2013). Some others, such as Zipf's law of meaning distribution, have received less attention. In his pioneering research, Zipf (1945) found that more frequent words tend to have more meanings. The functional dependency between μ, the number of meanings of a word, and f, the frequency of a word, has been approximated with

(Ilgen & Karaoglan, 2007; Baayen & Moscoso del Prado Martín, 2005; Zipf, 1945)

$$\mu \propto f^{\delta}, \tag{1}$$

where δ is a constant such that $\delta \approx 1/2$. Equation 1 defines Zipf's meaning-frequency law. Equivalently, the meaning-frequency law can be defined as

$$f \propto \mu^{1/\delta}. \tag{2}$$

Zipf derived the meaning-frequency law assuming two laws, the popular Zipf's law for word frequencies and the law of meaning distribution. On the one hand, Zipf's law for word frequencies states that the relationship between the frequency of a word and its rank (the most frequent word has rank $i = 1$, the second most frequent word has rank $i = 2$ and so on) as

$$f \propto i^{-\alpha}, \tag{3}$$

where $\alpha \approx 1$ is a constant (Zipf, 1945, 1949). On the other hand, the law of meaning distribution (Zipf, 1945, 1949) states that

$$\mu \propto i^{-\gamma}, \tag{4}$$

where $\gamma \approx 1/2$. Notice that i is still the rank of a word according to its frequency. The constants α, γ, and δ can be estimated applying some regression method as in Zipf's pioneering research (Zipf, 1945, 1949).

Sometimes, power-laws such as those described in Equations (1–4) are defined using asymptotic notation. For instance, random typing yields $f = \Theta(i^{-\alpha})$, that is, for sufficiently large i, one has (Conrad & Mitzenmacher, 2004)

$$a_1 i^{-\alpha} \leq f \leq a_2 i^{-\alpha}, \tag{5}$$

where a_1 and a_2 are constants such that $a_1 \leq a_2$. Equation 5 can be seen as a relaxation of Equation 3. Similarly, Heaps'

law on the relationship between V, the number of types, as a function of T, the number of tokens, is defined as $V = \Theta(T^\beta)$ with $0 < \beta < 1$ (Baeza-Yates & Navarro, 2000), a relaxed version of $V \propto T^\beta$ (see Font-Clos & Corral, 2015, for a critical revision of the power law model for Heaps' law).

The meaning-frequency law (Equation 1) and the law of meaning distribution (Equation 4) predict the number of meanings of a word using different variables as predictors. The meaning-frequency law has been confirmed empirically in various languages: directly through Equation 1 in Dutch and English (Baayen & Moscoso del Prado Martín, 2005) or indirectly through Equation 4 and the assumption of Zipf's law (Zipf, 1945; Ilgen & Karaoglan, 2007) in Turkish and English. Qualitatively, the meaning-frequency law defines a positive correlation between frequency and the number of meanings. Using a proxy of word meaning, the qualitative version of the law has been found in dolphin whistles (Ferrer-i-Cancho & McCowan, 2009) and in chimpanzee gestures (Hobaiter & Byrne, 2014). Thus, the law is a candidate for a universal property of communication.

Zipf (1945) argued that Equation 1 with $\delta = 1/2$ follows from Equation 3 with $\alpha = 1$ and Equation 4 with $\gamma = 1/2$. Indeed, it has been proven that Equation 1 with $\delta = \gamma/\alpha$ follows from Equations 1 and 4 (Ferrer-i-Cancho, 2016b). Here we consider alternative derivations of Equation 1 or relaxed versions of Equation 1 from the assumption of a biased random walk (Sinatra, Gómez-Gardeñes, Lambiotte, Nocosia, & Latora, 2011; Gómez-Gardeñes & Latora, 2008) over word-meaning associations. The remainder of the article is organized as follows.

First, we will present the mathematical framework.

Second, we will present a minimalist derivation of the meaning-frequency law (Equation 1) with $\delta = 1/2$ law that is based on just one assumption on the joint probability of a word and a meaning. Suppose a word s that is connected to μ meanings and a meaning r that is connected to ω words. Assuming that the joint probability of s and r is proportional to μ if s and r are connected and zero otherwise, it suffices to obtain Equation 1 with $\delta = 1/2$. A problem of the argument is that the definition is somewhat arbitrary and theoretically superficial.

Third, we will replace this simplistic assumption by a more fundamental assumption, namely, that the joint probability of s and r is proportional to $\mu\omega$ if s and r are connected and zero otherwise. This assumption is a more elegant solution for two reasons: it corrects the arbitrariness of the assumption of the minimalist derivation, fits into standard network theory, and it can be embedded into a general theory of communication. From this deeper assumption we derive the meaning-frequency law following three major paths. The first path consists of assuming the principle of contrast (Clark, 1987) or the principle of no synonymy (Goldberg, 1995, p. 67), namely, $\omega \leq 1$ for all words, which leads exactly to Equation 1. The second path consists of assuming that meaning degrees are mean independent of word degrees, which leads to a mirror of the meaning-frequency law (Equation 2)

$$E[p|\mu] \propto \mu^{1/\delta}, \tag{6}$$

where $E[p|\mu]$ is the expectation of p, the probability of a word, knowing that its degree is μ. Notice that p is linked with f as $p \approx f/L$, where L is the length of a text in tokens, that is, total number of tokens (the sum of all type frequencies; Moreno-Sànchez, Font-Clos, & Corral, 2016). The third path makes no assumption to obtain a relaxed version of the meaning-frequency law, namely, the number of meanings is bounded above and below by two power-laws over f, that is

$$b_1 f^\delta \leq \mu \leq b_2 f^\delta,$$

where b_1 and b_2 are constants such that $b_1 \leq b_2$. The result can be summarized as

$$\mu = \Theta(f^\delta). \tag{7}$$

Put together, these three paths strongly suggest that languages are channeled to reproduce Zipf's meaning-frequency law.

Fourth, we will review a family of optimization models of communication that was put forward to investigate the origins of Zipf's law for word frequencies (Ferrer-i-Cancho & Díaz-Guilera, 2007) but that has recently been used to shed light on patterns of vocabulary learning and the mapping of words into meanings (Ferrer-i-Cancho, 2016c). Interestingly, models from that family give Equation 1 with $\delta = 1$ (Ferrer-i-Cancho, 2016b). Crucially, however, the true exponent is $\delta \approx 1/2$ (Zipf, 1945; Ilgen & Karaoglan, 2007). The mismatch should not be surprising. Imagine that a speaker has to choose a word for a certain meaning. Those models assume that given a meaning, all the words with that meaning are equally likely (Ferrer-i-Cancho & Díaz-Guilera, 2007). However, this simple assumption is not supported by psycholinguistic research (Snodgrass & Vanderwart, 1980). We will show how to modify their definition so that the words that are used for a certain meaning do not need to be equally likely and one can obtain Equation 1 with $\delta = 1/2$ or relaxed versions. Finally, we will discuss the results, highlighting the connection with biased random walks, and indicate directions for future research.

A Mathematical Framework

As in the family of models of communication above, we assume a repertoire of n words, $s_1, \ldots, s_i, \ldots s_n$ and a repertoire of m meanings, $r_1, \ldots, r_i, \ldots, r_m$. Words and meanings are associated through an $n \times m$ adjacency matrix $A = \{a_{ij}\}$: $a_{ij} = 1$ if s_i and r_j are associated ($a_{ij} = 0$ otherwise). A defines the edges of an undirected bipartite network of word-meaning associations. The degree of the i-th word is

$$\mu_i = \sum_{j=1}^{m} a_{ij} \tag{8}$$

while the degree of the j-th meaning is

$$\omega_j = \sum_{i=1}^{n} a_{ij}. \tag{9}$$

In human language, the relationship between sound and meaning has been argued to be arbitrary to a large extent (Saussure, 1916; Hockett, 1966; Pinker, 1999). That is, there is no intrinsic relationship between the word form and its meaning. For example the word "car" is nothing like an actual automobile. An obvious exception are onomatopoeias, which are relatively rare in language. However, despite the immense flexibility of the world's languages, some sound-meaning associations are preferred by culturally, historically, and geographically diverse human groups (Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016). The framework above is agnostic concerning the type of association between sound and meaning. By doing that, we are borrowing the abstract perspective of network theory, that is a priori neutral concerning the nature or the origins of the edges (Newman, 2010; Barthélemy, 2011). Our framework could be generalized to accommodate Peirce's classic types of reference, that is, iconic, indexical, and symbolic (Deacon, 1997), or the state-of-the-art on the iconicity/systematicity distinction (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015). A crucial reason to remain neutral is that the distinctions above were not made when defining the laws of meaning that are the target of this article.

The framework allows one to model lexical ambiguity: a lexically ambiguous word is a word such that its degree is greater than one. Although the model starts from a flat hierarchy of concepts (by default all concepts have the same degree of generality), a word with an abstract meaning could be approximated either as a word linked to a single abstract concept or as a word linked to the multiple specific meanings it covers (Ferrer-i-Cancho, 2016c). As for the latter approach, the word for vehicle would be linked to the meanings for car, bike, ship, airplane, etc.

Suppose that $p(s_i, r_j)$ is the joint probability of the unordered pair formed by s_i and r_j. By definition,

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p(s_i, r_j) = 1. \tag{10}$$

The probability of s_i is

$$p(s_i) = \sum_{j=1}^{m} p(s_i, r_j) \tag{11}$$

and the probability of r_i is

$$p(r_j) = \sum_{i=1}^{n} p(s_i, r_j). \tag{12}$$

Our model shares the assumptions of distributional semantics that the meaning of a word is represented as a vector of

the weights of different concepts for that word (Lund & Burgess, 1996). In our framework, the meaning of the word s_i is represented by the m-dimensional vector

$$\{p(s_i, r_1), \ldots, p(s_i, r_j), \ldots, p(s_i, r_m)\}$$

The joint probabilities $p(s_i, r_j)$ for all words and meanings defines a weighted matrix of the same size of A. In the coming sections, we will derive the meaning frequency-law defining $p(s_i, r_j)$ as a function of A. Put differently, we will derive the law from a weighted undirected bipartite graph that is built from the unweighted undirected graph defined by A. This organization in two graphs (one unweighted and the other weighted) instead of a single weighted graph is borrowed from successful models of communication (Ferrer-i-Cancho, 2016c) and allows one to apply the theory of random walks (Sinatra, Gómez-Gardeñes, Lambiotte, Nocosia, & Latora, 2011; Gómez-Gardeñes & Latora, 2008), as we will see later on.

A Minimalist Derivation of the Law

The law of meaning distribution can be derived by making just one rather simple assumption, that is

$$p(s_i, r_j) \propto a_{ij} \mu_i, \tag{13}$$

Applying Equation 10, one obtains

$$p(s_i, r_j) = c a_{ij} \mu_i, \tag{14}$$

where c is a normalization constant defined as

$$c = \frac{1}{\displaystyle\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} \mu_i} = \frac{1}{\displaystyle\sum_{i=1}^{n} \mu_i \sum_{j=1}^{m} a_{ij}} = \frac{1}{\displaystyle\sum_{i=1}^{n} \mu_i^2}.$$

Notice that c is not a parameter and its value is determined by the definition of probability in Equation 10. Applying Equation 11 to Equation 14 gives

$$p(s_i) = c \mu_i^2,$$

Namely, Equation 1 with $\delta = 1/2$. Our derivation of the strong and relaxed version of the meaning-frequency law is simpler than that of Zipf's in the sense that it requires assuming a smaller number of equations (we are assuming only Equation 13, while Zipf assumed Equations 3 and 4). However, the challenge of our approach is the justification of Equation 13.

A Theoretical Derivation of the Law

The definition of $p(s_i, r_j)$ in Equation 13 suffices as a model but not for the construction of a real theory of language. Equation 13 is simple but somewhat arbitrary: the degree of the word, μ_i, contributes raised to 1 but the degree of the meaning ω_j has no direct contribution, or one may say

that it contributes raised to 0. Therefore, a less arbitrary equation would be

$$p(s_i, r_j) \propto a_{ij}(\mu_i \omega_j)^{\phi}, \qquad (15)$$

where ϕ is a positive parameter ($\phi \geq 0$). Applying Equation 10 to Equation 15, one obtains

$$p(s_i, r_j) = c a_{ij}(\mu_i \omega_j)^{\phi} \qquad (16)$$

with $c = 1/M$

and

$$M = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}(\mu_i \omega_j)^{\phi} = \sum_{i=1}^{n} \mu_i^{\phi} \sum_{j=1}^{m} a_{ij} \omega_j^{\phi}. \qquad (17)$$

Notice that ϕ is the only parameter of the model given n and m. Applying Equation 11 to Equation 16, one obtains

$$p(s_i) = c \mu_i^{\phi} \sum_{j=1}^{m} a_{ij} \omega_j^{\phi}. \qquad (18)$$

Equation 15 is theoretically appealing for various reasons. If $p(s_i, r_j)$ is regarded as the weight of the association between s_i and r_j, it defines the general form of the relationship between the weight of an edge and the product of the degrees of vertices at both ends that is found in real networks (Barrat, Barthélemy, Pastor-Satorras, & Vespignani, 2004). For this reason, a unipartite version of Equation 15 is assumed to study dynamics on networks (Baronchelli, Castellano, & Pastor-Satorras, 2011). When $\phi = 0$, it matches the definition of models about the origins of Zipf's law for word frequencies (Ferrer-i-Cancho, 2005b), the variation of the exponent of the law (Ferrer-i-Cancho, 2005a, 2006) and vocabulary learning (Ferrer-i-Cancho, 2016c). When $\phi = 1$, it defines an approximation to the stationary probability of observing a transition involving s_i and r_j in a random walk on a network that is biased to maximize the entropy rate of the walks (Appendix A1), thus suggesting that the meaning-frequency law could be a manifestation of a particular random walk process on semantic memory.

Two equivalent linguistic principles, the principle of contrast (Clark, 1987) and the principle of no synonymy (Goldberg, 1995, p. 67) can be implemented in our model as $\omega_j \in \{0, 1\}$. From an algebraic standpoint, the condition $\omega_j \in \{0, 1\}$ is equivalent to orthogonality of the word vectors of the matrix A. If A_{i*} indicates the row vector of A for the i-th word, A_{i*} and A_{k*} are orthogonal if and only if $A_{i*} \cdot A_{k*} = 0$, where the dot indicates the scalar product of two vectors. To simplify matters, we assume that there is no row vector of A that equals $\vec{0}$, a vector that has 0 in all components. So far, we have used μ_i and ω_j to refer, respectively, to the degree of the i-th word and the j-th meaning. We define μ_i^e and ω_i^e as the degree of the word and the degree of the meaning of the i-th edge. μ_i^e and ω_i^e are

components of the vectors $\vec{\mu}_i^e$ and $\vec{\omega}_i^e$, respectively. We have $\vec{\mu}_i^e \cdot \vec{\omega}_i^e > 0$ because $\mu_i^e, \omega_i^e \geq 1$ by definition. A deeper insight can be obtained with the concept of remaining degree, the degree at one end of the edge after subtracting the unit contribution of the edge (Newman, 2002). The vectors of remaining degrees are then

$$\vec{\mu}_i^{\prime e} = \vec{\mu}_i^e - \vec{1}$$
$$\vec{\omega}_i^{\prime e} = \vec{\omega}_i^e - \vec{1}.$$

The condition $\omega_j \in \{0, 1\}$ is equivalent to $\vec{\omega}_i^{\prime e} = \vec{0}$. $\omega_j \in \{0, 1\}$ leads to $\vec{\mu}_i^{\prime e} \vec{\omega}_i^{\prime e} = 0$ but trivially for being $\vec{\omega}_i^{\prime e}$ null.

The assumption $\phi = 1$ and $\omega_j \in \{0, 1\}$ (orthogonality of row vectors of A), transforms Equation 18 into Equation 13 because $a_{ij} = 0$ and $\omega_j = 0$ are equivalent when ω_j does not exceed 1. In general, Equation 18 combined with the principle of contrast gives

$$p(s_i) = c \mu_i^{\phi} \sum_{j=1}^{m} a_{ij}$$
$$= c \mu_i^{\phi+1}$$

and finally Equation 1 with

$$\delta = \frac{1}{\phi + 1}.$$

When $\phi = 1$, we get $\delta = 1/2$ again. Interestingly, the principle of contrast follows from the principle of mutual information maximization, a more fundamental principle that allows one to predict vocabulary learning in children and that can be combined with the principle of entropy minimization to predict Zipf's law for word frequencies (Ferrer-i-Cancho, 2016c). With Equation 15, we follow Bunge (2013, pp. 32–33) preventing scientific knowledge from becoming "an aggregation of disconnected information" and aspiring to build a "system of ideas that are logically connected among themselves."

It is possible to obtain a relaxed meaning frequency-law under more general conditions. In particular, we would like to get rid of the heavy constraint that meaning degrees cannot exceed 1. Suppose that d is a constant such that $0 < d \leq n$. Some obvious but not very general conditions are $\omega_j \in \{0, d\}$ for all j or $\omega_j = d$ for all j. It is easy to see that they lead again to Equation 13 when $\phi = 1$. A more general condition can be defined as follows. First, we define $E[\omega^{\phi}|\mu]$ as the conditional expectation of ω^{ϕ} given μ for an edge. Here μ and ω are the degrees at both ends of an edge. Then suppose that A is given and that ω^{ϕ} is mean independent of μ, namely, $E[\omega^{\phi}|\mu] = E[\omega^{\phi}]$ for each value of μ (Kolmogorov, 1956; Poirier, 1995), then the expectation of $p(s_i)$ (as defined in Equation 18) knowing μ_i is

$$E[p(s_i)|\mu_i] = E\left[c \mu_i^{\phi} \sum_{j=1}^{m} a_{ij} \omega_j^{\phi} \middle| \mu_i \right]$$

$$= c\mu_i^\phi E\left[\sum_{j=1}^{m} a_{ij}\omega_j^\phi \,\Big|\, \mu_i\right]$$

$$= c\mu_i^\phi \sum_{j=1}^{m} a_{ij} E\left[\omega_j^\phi \,\Big|\, \mu_i\right]$$

$$= c\mu_i^\phi \sum_{j=1}^{m} a_{ij} E\left[\omega_j^\phi\right]$$

$$= cE[\omega_j]\mu_i^{\phi+1},$$

which can be seen as a regression model (Ritz & Streibig, 2008) for the meaning-frequency law (Equation 1) with word degree as predictor. Notice that mean independence is a more general condition than mutual or statistical independence but a particular case of uncorrelation (Ferrer-i-Cancho, Hernández-Fernández, Baixeries, Debowski, & Mačutek, 2014).

So far, we have seen ways of obtaining the meaning-frequency law from Equation 15 making further assumptions. It is possible to obtain a relaxed version of the meaning-frequency law making no additional assumption (except Equation 15 or the biased random walk that justifies it). Equation 18 can be expressed as

$$p(s_i) = \mu_i^\phi \sum_{j=1}^{m} a_{ij} T_j \qquad (19)$$

with

$$T_j = c\omega_j^\phi.$$

Assuming that

$$T_{\min} \le T_j \le T_{\max},$$

Equation 19 leads to

$$T_{\min}\mu_i^{\phi+1} \le p(s_i) \le T_{\max}\mu_i^{\phi+1} \qquad (20)$$

or equivalently

$$\frac{1}{T_{\min}} p(s_i)^{\frac{1}{\phi+1}} \le \mu_i \le \frac{1}{T_{\max}} p(s_i)^{\frac{1}{\phi+1}}.$$

Recalling $p \approx f/L$, these results can be summarized using asymptotic notation as $f = \Theta(\mu^{\phi+1})$ or $\mu = \Theta(f^{1/(\phi+1)})$. The power of the bounds above depends on the gap between $T_{\min}$ and $T_{\max}$. The gap can be measured with the ratio

$$\frac{T_{\max}}{T_{\min}} = \frac{\omega_{\max}^\phi}{\omega_{\min}^\phi},$$

where $\omega_{\min}$ and $\omega_{\max}$ are the minimum and the maximum meaning degree, respectively. The principle of mutual information maximization between words and meanings, a general principle of communication (Ferrer-i-Cancho, 2016c), puts pressure for concordance with the meaning-frequency law. To see it, we consider two cases: $n \le m$ and $m \le n$. When $n \le m$, its maximization predicts $\omega_j \le 1$ (Appendix A2). As unlinked meanings are irrelevant (they do not belong to the support set), we have $\omega_{\min} = 1$. As pressure for mutual information maximization increases, ω_{max} tends to 1 and thus $T_{\max}/T_{\min}$ tends to 1. Put differently, the gap between the upper and the lower bound in Equation 20 reduces as pressure for mutual information maximization increases. When $n \ge m$, mutual information maximization predicts that $\omega_j = d$, where d is an integer such that $d \in [1, \lfloor n/m \rfloor]$ (Appendix A2). We have seen above that one obtains the meaning-frequency law (Equation 1) immediately from Equation 15 when ω_j is constant. We conclude that the chance of observing the meaning-frequency law increases as pressure for mutual information maximization increases.

A Family of Optimization Models of Communication

Here we revisit a family of optimization models of communication (Ferrer-i-Cancho & Díaz-Guilera, 2007) in the light of the results of the previous sections. These models share the assumption that the probability that a word s_i is employed to refer to meaning r_j is proportional to a_{ij}, that is

$$p(s_i|r_j) \propto a_{ij}, \qquad (21)$$

Applying

$$\sum_{i=1}^{n} p(s_i|r_j) = 1$$

to Equation 21, we obtain

$$p(s_i|r_j) = \frac{a_{ij}}{\omega_j}. \qquad (22)$$

We adopt the convention $p(s_i|r_j) = 0$ when $\omega_j = 0$.

Equation 22 defines the probability of transition of a standard (unbiased) random walk to a word (Noh & Rieger, 2004), that is, given a meaning, all related words are equally likely. This is unrealistic in light of picture-naming norms (Snodgrass & Vanderwart, 1980; Duñabeitia et al., 2017). Consider the picture-naming norms compiled by Snodgrass and Vanderwart (1980), who simply asked participants to name 260 black-and-white line drawings of common objects. For some objects (for example, balloon, banana, sock, star) there was 100% agreement among the participants for the word used to name the pictured object. However, for other objects there was considerable variability in the word used to name the pictured object. Important for the present argument, the other words that were used in such cases were not selected with equal likelihood. For example, the picture of a *wineglass* had 50% agreement, with the

word *glass* (36% of the responses) and the word *goblet* (14% of the responses) also being used to name the object, showing that all the words that could be used for a given meaning are not equally likely. Although subjects tend to provide more specific responses when the concept is presented in textual form with respect to a visual form presentation (Tversky & Hemenway, 1983), we used the visual case simply to challenge the assumption of an unbiased random walk in general and justify a more realistic approach.

In contrast to Equation 22, the fundamental assumption in Equation 15 leads to

$$p(s_i|r_j) = \frac{a_{ij}\mu_i^{\phi}}{\sum_k a_{kj}\mu_k^{\phi}}, \tag{23}$$

Namely, the transition probabilities of a biased random walk when $\phi > 0$ (Sinatra, Gómez-Gardeñes, Lambiotte, Nocosia, & Latora, 2011; Gómez-Gardeñes & Latora, 2008). To see it, notice that the combination of Equation 12 and 16 produces

$$p(r_j) = \sum_{i=1}^{n} p(s_i, r_j) = c\omega_j^{\phi} \sum_{i=1}^{n} a_{ij}\mu_i^{\phi}. \tag{24}$$

Recalling the definition of conditional probability

$$p(s_i|r_j) = \frac{p(s_i, r_j)}{p(r_j)}$$

and applying Equation 16 again, one obtains Equation 23.

Recalling the definition of ω_j in Equation 9, it is easy to realize that Equation 22 is a particular case of Equation 23 with $\phi = 0$. While the family of models above stems from a concrete definition of a conditional probability, that is, $p(s_i|r_j)$ in Equation 22, the general model that we have presented in this article is specified by a definition of the joint probability, that is, $p(s_i, r_j)$ in Equation 15.

Models within that family are generated through

$$p(s_i, r_j) = p(s_i|r_j)p(r_j), \tag{25}$$

assuming an unbiased random walk from a meaning to a word (Equation 22) and making different assumptions on $p(r_j)$.

If one assumes that all meanings are equally likely ($p(r_j) = 1/m$ with $\omega_j \geq 1$) one obtains the first model (Ferrer-i-Cancho & Solé, 2003). If one assumes that the probability of a meaning is proportional to its degree ($p(r_j) \propto \omega_j$) one obtains the second model (Ferrer-i-Cancho, 2005b). While in the second model $p(r_j|s_i)$ defines an unbiased random walk from s_i to r_j (all r_j's connected to s_i are equally likely), this is not necessarily the case for the first model (Ferrer-i-Cancho & Díaz-Guilera, 2007). Therefore, the second model defines a pure unbiased random walk while the first model is unbiased from meaning to words but biased from words to meanings.

Now we will introduce a generalized version of the family of models above consisting of replacing Equation 22 by Equation 23 and generating the corresponding variants of the first and the second model applying the same procedure as in the original family, namely, via Equation 25. Notice that Equation 23 defines the probability of reaching s_i from r_j in a biased random walk when $\phi > 0$.

Concerning the first model, suppose that the probabilities of the meanings are given a priori (they are independent from the A matrix), for example, all meanings are equally likely. Then it is easy to show that the model yields a relaxed version of the meaning-frequency law, namely, $\mu_i = \Theta(p(s_i)^{\delta})$, the number of meanings is bounded above and below by two power-laws (Appendix A3)

$$b_1 p(s_i)^{\delta} \leq \mu_i \leq b_2 p(s_i)^{\delta}, \tag{26}$$

where b_1 and b_2 are constants ($b_1 \leq b_2$) and $\delta = 1/(\phi+1)$. Equation 26 defines nontrivial bounds when $\delta \neq 1$ (Appendix A3). The case $\delta = 1$ matches an optimization model of Zipf's law for word frequencies (Ferrer-i-Cancho, 2005b, 2016b).

To generate a variant of the second model, recall that Equation 23 comes from Equation 15. Equations 12 and 16 produce Equation 24. This variant of the second model derives all probability definitions from Equation 15. We have shown above that this variant is able to generate the meaning-frequency law.

Discussion

We have seen that it is possible to obtain the meaning-frequency law (Equations 1 and 2) from Equation 15 making certain assumptions. We have also seen that a relaxed version of the law (Equation 7) can be obtained from Equation 15 without making any further assumption. Our findings suggest that word probabilities are channeled somehow to manifest the meaning-frequency law. We have seen that the principle of mutual information maximization contributes to the emergence of the law. Our derivation is theoretically appealing for various reasons. First, it is more parsimonious than Zipf's concerning the number of equations that are assumed (we only need Equation 15 while Zipf involved Equations 3 and 4). Second, it can help a family of optimization models of language to reproduce the meaning-frequency law.

Therefore, a crucial assumption is Equation 16, that we have justified as the outcome of a random walk that is biased to maximize the entropy rate of the paths (Appendix A1). A random walk is the correlate in network theory of the concept of mental exploration (navigation without a target or nonstrategic search) in cognitive science and related fields (Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chatter, & Christiansen, 2013). Semantic memory processes can be usefully theorized as searches over a network (Thompson & Kello, 2014; Abbott, Austerweil, & Griffiths, 2015) or some semantic space (Smith, Huber, & Vul, 2013). These

approaches support the hypothesis of a Markov chain process for memory search (Bourgin, Abbott, Griffiths, & Vul, 2014), provide a deeper understanding of creativity (Kenett & Austerweil, 2016), and help to develop efficient navigation strategies (Capitán et al., 2012).

A random walk in a unipartite word network of word–word associations has been argued to underlie Zipf's law for word frequencies (Allegrini, Gricolini, & Palatella, 2004). Here we contribute with a new hypothesis linking random walks with a linguistic law: that the meaning-frequency law would be an epiphenomenon of a biased random walk over a bipartite network of word-meaning associations in the process of mental exploration. The bias consists of exploiting local information, namely, the degrees of first neighbors (Sinatra et al., 2011). Transitions to nodes with a higher degree are preferred. Our model shows that it is possible to approximate the optimal solution to a problem (maximizing the entropy rate of the paths) following an apparently non-strategic search (Hills, Jones, & Todd, 2012; Abbott et al., 2015).

The probability of a word in Equation 18 defines the probability that a random walker visits the word in the long run. This probability is what the PageRank algorithm estimates in the context of a standard (nonbiased) random walk (Page, Brin, Motwani, & Winograd, 1998). The assumption of a random walk with the particular bias above could help to improve random walk/PageRank methods to predict the prominence in memory of a word (Griffiths, Steyvers, & Firl, 2007) or the importance of a tag (Jäschke, Marinho, Hotho, Schmidt-Thieme, & Stumme, 2007). A virtue of our biased random walk is that it predicts an uneven conditional probability of a word given a meaning (Equation 23) as happens in real language (Snodgrass & Vanderwart, 1980). A standard (uniform) random walk cannot explain this fact and for that reason the family of optimization models of language revisited above fails to reproduce the meaning-frequency law with $\delta = 1/2$.

Although biased random walks have already been used to solve information retrieval problems (see Duarte Torres, Hiemstra, Weber, & Pavel, 2014, and references therein), a bias based on the degree of neighbors has not been considered as far as we know. We hope that our results stimulate further research on linguistics laws and biased random walks in the information sciences. Specifically, we hope that our article becomes the fuel of future empirical research.

Acknowledgments

Appendix A1: Random Walks

We will show that Equation 15 defines the probability of observing a transition between s_i and r_j in any direction in a biased random walk. We will proceed in two steps. First, we will summarize some general results on biased random walks on unipartite networks and then we will adapt them to bipartite networks.

Suppose a unipartite network of n nodes that is defined by an $n \times n$ adjacency matrix $B = \{b_{ij}\}$ such that $b_{ij} = 1$ if the i-th and the j-th node are connected and $b_{ij} = 0$ otherwise. Let k_i be the degree of the i-th node, namely,

$$k_i = \sum_{j=1}^{n} b_{ij}.$$

Suppose a random walk over the vertices of a network where $p(j|i)$ is the probability of jumping from i to j. A first order approximation to the $p(j|i)$ that maximizes the entropy rate is (Sinatra et al., 2011)

$$p(j|i) = \frac{b_{ij}k_j}{\sum_{l=1}^{n} b_{il}k_l}. \tag{27}$$

We choose a generalization (Gómez-Gardeñes & Latora, 2008)

$$p(j|i) = \frac{b_{ij}k_j^{\phi}}{\sum_{l=1}^{n} b_{il}k_l^{\phi}}, \tag{28}$$

that gives Equation 27 with $\phi = 1$. The stationary probability of visiting the i-th vertex in the biased random walk defined by Equation 28 is (Gómez-Gardeñes & Latora, 2008)

$$p(i) = \frac{k_i^{\phi}c_i}{T}, \tag{29}$$

where

$$c_i = \sum_{j=1}^{n} b_{ij}k_j^{\phi} \tag{30}$$

and

$$T = \sum_{i=1}^{n} c_i k_i^{\phi}. \tag{31}$$

Now we adapt the results above to a bipartite graph of word-meaning associations. As the graph is bipartite, the random walker will be alternating between words and meanings. The probability that the vertex visited is a

word is 1/2 (the same probability for a meaning). Suppose that there are n words and m meanings. Recall that the bipartite network of word-meaning associations is defined by an $n \times m$ adjacency matrix $A = \{a_{ij}\}$ such that $a_{ij} = 1$ if the i-th word and the j-th meaning are connected and $a_{ij} = 1$ otherwise. μ_i is the degree of the i-th word is (Equation 8) whereas ω_j is the degree of the j-th meaning (Equation 9). The probability of jumping from r_j to s_i becomes (recall Equation 28)

$$p_v(s_i | r_j) = \frac{a_{ij} \mu_i^{\phi}}{\displaystyle\sum_{l=1}^{n} a_{lj} \mu_l^{\phi}}.$$

The probability of jumping from s_i to r_j is

$$p_v(r_j | s_i) = \frac{a_{ij} \omega_j^{\phi}}{\displaystyle\sum_{l=1}^{m} a_{il} \omega_l^{\phi}}. \tag{32}$$

The stationary probability of visiting the word s_i becomes (recall Equations 29 and 30)

$$p_v(s_i) = \frac{\mu_i^{\phi} \displaystyle\sum_{j=1}^{m} a_{ij} \omega_j^{\phi}}{M_v}, \tag{33}$$

where M_v corresponds to T in Equation 29. Adapting Equations 31 and 30, one obtains

$$M_v = \sum_{i=1}^{n} \mu_i^{\phi} \sum_{j=1}^{m} a_{ij} \omega_j^{\phi} + \sum_{j=1}^{m} \omega_j^{\phi} \sum_{i=1}^{n} a_{ij} \mu_i^{\phi}$$

$$= 2M,$$

where M is defined as in Equation 17. Applying Equation 33, it is easy to see that

$$\sum_{i=1}^{n} p_v(s_i) = \frac{1}{2M} \sum_{i=1}^{n} \mu_i^{\phi} \sum_{j=1}^{m} a_{ij} \omega_j^{\phi}$$

$$= \frac{1}{2}.$$

as expected.

The combination of Equations 32 and 33 allows one to derive the probability of observing the transition from s_i to r_j as

$$p_v(s_i \to r_j) = p_v(r_j | s_i) p_v(s_i)$$

$$= c_v a_{ij} (\mu_i \omega_j)^{\phi},$$

where $c_v = 1/(2M)$. Similarly, the probability of observing the transition from r_j to s_i is

$$p_v(s_i \leftarrow r_j) = p_v(s_i | r_j) p_v(r_j)$$

$$= c_v a_{ij} (\mu_i \omega_j)^{\phi}.$$

Therefore, the stationary probability of observing a transition between s_i and r_j in any direction (from s_i to r_j or from r_j to s_i) is

$$p(s_i, r_j) = p_v(s_i \to r_j) + p_v(s_i \leftarrow r_j)$$

$$= 2c_v a_{ij} (\mu_i \omega_j)^{\phi}.$$

$$= c a_{ij} (\mu_i \omega_j)^{\phi}.$$

with $c = 1/M$, as we wanted to show.

Finally, we will link $p(s_i)$, the probability of a word that is used in the main text to derive the meaning-frequency law, with $p_v(s_i)$. Notice that $p(s_i) = p_v(s_i | S)$, the latter being the probability of visiting vertex s_i knowing that it belongs to the partition S, the partition of words. Since the graph is bipartite, $p_v(S)$, the probability that the random walk is visiting a vertex of partition S, is 1/2. The joint probability of visiting vertex s_i and that the vertex belongs to S is

$$p_v(s_i, S) = p_v(S | s_i) p_v(s_i)$$

$$= p_v(s_i).$$

Therefore,

$$p(s_i) = p_v(s_i | S)$$

$$= \frac{p_v(s_i, S)}{p_v(S)}$$

$$= 2 p_v(s_i).$$

Then $p(s_i)$ is the stationary probability of visiting s_i in a biased random walk knowing that the vertex is in S.

Appendix A2: Mutual Information Maximization

Suppose that $I(S, R)$ is the mutual information between words (S) and meanings (R), that can be defined as

$$I(S, R) = H(S) - H(S | R), \tag{34}$$

where $H(S)$ is the entropy of words and $H(S | R)$ is the conditional entropy of words given meanings. For the case $\phi = 0$, the configurations that maximize $I(S, R)$ when $n \le m$ are defined by two conditions (Ferrer-i-Cancho, 2016c)

1. $\mu_i = d$ with $d \in [1, \lfloor m/n \rfloor]$ for $i = 1, 2, \ldots, n$.
2. $\omega_j \in \{0, 1\}$ for $j = 1, 2, \ldots, m$.

When $n \ge m$, those configurations are the symmetric, that is (Ferrer-i-Cancho, 2016c)

1. $\omega_j = d$ with $d \in [1, \lfloor n/m \rfloor]$ for $j = 1, 2, \ldots, m$.
2. $\mu_i \in \{0, 1\}$ for $i = 1, 2, \ldots, n$.

Here we will show that the configurations that maximize $I(S, R)$ are the same as in the case $\phi=0$ when ϕ is a positive and finite real number ($\phi \geq 0$). By symmetry, it suffices to show it for the case $n \leq m$. We will proceed in three steps. First, deriving the configurations minimizing $H(S|R)$. Second, showing that the configurations above yield maximum $I(S, R)$. Third, showing that they are the only configurations.

Step 1: Recall that

$$H(S|R)=E[H(S|r_j)]$$
$$=\sum_{j=1}^{m} p(r_j)H(S|r_j)$$

where $H(S|r_j)$ is the conditional entropy of words given the meaning r_j. Equation 24 implies that $p(r_j) \neq 0$ is equivalent to $w_j > 0$ and then

$$H(S|R)= \sum_{\substack{j=1 \\ p(r_j) \neq 0}}^{m} p(r_j)H(S|r_j)$$
$$= \sum_{\substack{j=1 \\ w_j > 0}}^{m} p(r_j)H(S|r_j).$$

Knowing that

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

it is easy to see that $H(S|r_j)=0$ when $p(s_i|r_j) \in \{0, 1\}$ for $i=1, 2, \ldots, n$: $0\log 0=0$ by continuity since $x\log x \to 0$ as $x \to 0$ (Cover & Thomas, 2006, p. 14) and obviously $1\log 1=0$. Equation 23 implies that $p(s_i|r_j)=1$ is equivalent to s_i being the only neighbor of r_j, that is, $\omega_j=1$. Therefore, $H(S|R)=0$ implies $\omega_j \leq 1$ for $j=1, 2, \ldots, m$.

Step 2: Notice that the second condition of the case $n \leq m$ above implies $H(S|R)=0$ (recall Step 1). The second condition transforms Equation 17 into

$$M=\sum_{i=1}^{n} \mu_i^{\phi} \sum_{j=1}^{m} a_{ij}=\sum_{i=1}^{n} \mu_i^{\phi+1}$$

and Equation 18 into

$$p(s_i)=c\mu_i^{\phi} \sum_{j=1}^{m} a_{ij}=c\mu_i^{\phi+1}. \tag{35}$$

Adding the first condition, one obtains

$$M=\sum_{i=1}^{n} d^{\phi+1}=nd^{\phi+1}$$

$$p(s_i)=c\mu_i^{\phi+1}=\frac{1}{M}d^{\phi+1}=\frac{1}{n}.$$

and then $H(S)=\log n$ (as all words are equally likely). Thus, $H(S)$ is taking its maximum possible value, whereas $H(S|R)$ is taking its minimum value. As $I(S, R)=H(S)-H(S|R)$, it follows that $I(S, R)$ is maximum.

Step 3: Notice that
- If the second condition fails, then $H(S|R) > 0$ and thus $I(S, R) < \log n$ even if $H(S)$ is maximum because of Equation 34. Thus, the second condition is required to maximize $I(S, R)$.
- If the first condition fails (while the second condition holds), then words are not equally likely as the probability of a word is proportional to a power of its degree (Equation 35). Then one has that $H(S) < \log n$ and it follows that $I(S, R)$ is not maximum because $I(S, R) \leq H(S)$.

Appendix A3: New Models

Combining Equations 11 and 23, one obtains

$$p(s_i)=\sum_{j=1}^{m} p(s_i, r_j)=\sum_{j=1}^{m} p(s_i|r_j)p(r_j)=\mu_i^{\phi}\sum_{j=1}^{m} a_{ij}T_j \tag{36}$$

with

$$T_j=\frac{p(r_j)}{\sum_{i=1}^{n} a_{ij}\mu_i^{\phi}}.$$

Suppose that

$$T_{\min} \leq T_j \leq T_{\max}$$

when $\omega_j > 0$. Equation 36 leads to

$$\mu_i^{\phi}\sum_{j=1}^{m} a_{ij}T_{\min} \leq p(s_i) \leq \mu_i^{\phi}\sum_{j=1}^{m} a_{ij}T_{\max}$$

and finally

$$T_{\min}\mu_i^{\phi+1} \leq p(s_i) \leq T_{\max}\mu_i^{\phi+1} \tag{37}$$

recalling the definition of μ_i in Equation 8. Equivalently,

$$T_{\max}^{-\delta}p(s_i)^{\delta} \leq \mu_i \leq T_{\min}^{-\delta}p(s_i)^{\delta}, \tag{38}$$

with

$$\delta=\frac{1}{\phi+1},$$

Namely, a relaxed version of the meaning-frequency law when $\phi=1$.

Notice that Equations 37 and 38 define nontrivial bounds in the sense that they are not expected from bounds on join-probability alone. If the range of variation of $p(s_i, r_j)$ satisfies

$$\pi_{\min} \leq p(s_i, r_j) \leq \pi_{\max}$$

when $p(s_i, r_j) > 0$, then Equation 11 gives

$$\sum_{j=1}^{m} a_{ij}\pi_{\min} \leq p(s_i) \leq \sum_{j=1}^{m} a_{ij}\pi_{\max}$$

and then

$$\pi_{\min}\mu_i \leq p(s_i) \leq \pi_{\max}\mu_i.$$

Therefore, the finding that

$$b_1 p(s_i)^{\delta} \leq \mu_i \leq b_2 p(s_i)^{\delta},$$

where b_1 and b_2 are constants is trivial when $\delta = 1$.

References

Abbott, J.T., Austerweil, J.L., & Griffiths, T. (2015). Random walks on semantic networks can resemble optimal foraging. Psychological Science, 122, 558–569.

Allegrini, P., Gricolini, P., & Palatella, L. (2004). Intermittency and scale-free networks: a dynamical model for human language complexity. Chaos, solitons and fractals, 20, 95–105.

Baayen, H., & Moscoso del Prado Martín, F. (2005). Semantic density and past-tense formation in three Germanic languages. Language, 81, 666–698.

Baeza-Yates, R., & Navarro, G. (2000). Block addresing indices for approximate text retrieval. Journal of the American Society for Information Science, 51(1), 69–82.

Baronchelli, A., Castellano, C., & Pastor-Satorras, R. (2011). Voter models on weighted networks. Physical Review E, 83, 066117.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chatter, N., & Christiansen, M. (2013). Networks in cognitive science. Trends in Cognitive Sciences, 17, 348–360.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. Proceedings of the National Academy of the Sciences USA, 101(11), 3747–3752.

Barthélemy, M. (2011). Spatial networks. Physics Reports, 499(1), 1–101.

Blasi, D.E., Wichmann, S., Hammarström, H., Stadler, P., & Christiansen, M. (2016). Sound-meaning association biases evidenced across thousands of languages. Proceedings of the National Academy of Sciences, 113(39), 10818–10823.

Bourgin, D.D., Abbott, J.T., Griffiths, T.L., Smith, K.A., & Vul, E. (2014). Empirical evidence for Markov Chain Monte Carlo in memory search. In Proceedings of the 36th Annual Meeting of the Cognitive Science Society (p. 224–229).

Bunge, M. (2013). La ciencia. su método y su filosofía. Pamplona, Spain: Laetoli.

Capitán, J.A., et al. (2012, 08). Local-based semantic navigation on a networked representation of information. PLoS One, 7(8), 1–10.

Clark, E. (1987). The principle of contrast: a constraint on language acquisition. In B. MacWhinney (Ed.), Mechanisms of language acquisition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Conrad, B., & Mitzenmacher, M. (2004). Power laws for monkeys typing randomly: the case of unequal probabilities. IEEE Transactions on Information Theory, 50(7), 1403–1414.

Cover, T.M., & Thomas, J.A. (2006). Elements of information theory. New York: Wiley (2nd ed.).

Deacon, T.W. (1997). The symbolic species: the co-evolution of language and the brain. New York: W.W. Norton.

Dingemanse, M., Blasi, D.E., Lupyan, G., Christiansen, M.H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. Trends in Cognitive Sciences, 19(10), 603–615.

Duarte Torres, S., Hiemstra, D., Weber, I., & Pavel, S. (2014). Query recommendation in the information domain of children. Journal of the Association for Information Science and Technology, 65(7), 1368–1384.

Duñabeitia, J.A., et al. (2017). MultiPic: A standardized set of 750 drawings with norms for six European languages. The Quarterly Journal of Experimental Psychology, in press.

Fedorowicz, J. (1982). The theoretical foundation of Zipf's law and its application to the Bibliographic Database Environment. Journal of the Association for Information Science and Technology, 33, 285–293.

Ferrer-i-Cancho, R. (2005a). The variation of Zipf's law in human language. European Physical Journal B, 44, 249–257.

Ferrer-i-Cancho, R. (2005b). Zipf's law from a communicative phase transition. European Physical Journal B, 47, 449–457.

Ferrer-i-Cancho, R. (2006). When language breaks into pieces. A conflict between communication through isolated signals and language. Biosystems, 84, 242–253.

Ferrer-i-Cancho, R. (2016a). Compression and the origins of Zipf's law for word frequencies. Complexity, 21, 409–411.

Ferrer-i-Cancho, R. (2016b). The meaning-frequency law in Zipfian optimization models of communication. Glottometrics, 35, 28–37.

Ferrer-i-Cancho, R. (2016c). The optimality of attaching unlinked labels to unlinked meanings. Glottometrics, 36, 1–16.

Ferrer-i-Cancho, R., & Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. Journal of Statistical Mechanics, P06009.

Ferrer-i-Cancho, R., & Gavaldà, R. (2009). 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.

Ferrer-i-Cancho, R., Hernández-Fernández, A., Baixeries, J., Debowski, Ł., & Mačutek, J. (2014). When is Menzerath-Altmann law mathematically trivial? A new approach. Statistical Applications in Genetics and Molecular Biology, 13, 633–644.

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

Ferrer-i-Cancho, R., & McCowan, B. (2009). A law of word meaning in dolphin whistle types. Entropy, 11(4), 688–701.

Ferrer-i-Cancho, R., & Solé, R.V. (2003). Least effort and the origins of scaling in human language. Proceedings of the National Academy of Sciences USA, 100, 788–791.

Font-Clos, F., Boleda, G., & Corral, A. (2013). A scaling law beyond Zipf's law and its relation to Heaps' law. New Journal of Physics, 15, 093033.

Font-Clos, F., & Corral, A. (2015). Log-log convexity of type-token growth in Zipf's systems. Physical Review Letters, 114, 238701.

Goldberg, A. (1995). Constructions: a construction grammar approach to argument structure. Chicago: Chicago University Press.

Gómez-Gardeñes, J., & Latora, V. (2008). Entropy rate of difussion process on complex networks. Physical Review E, 78, 065102(R).

Griffiths, T., Steyvers, M., & Firl, A. (2007). Google and the mind. Predicting fluency with PageRank. Psychological Science, 18, 1069–1076.

Hills, T., Jones, M., & Todd, P. (2012). Optimal foraging in semantic memory. Psychological Science, 119, 431–440.

Hobaiter, C., & Byrne, R.W. (2014). The meanings of chimpanzee gestures. Current Biology, 24, 1596–1600.

Hockett, C.F. (1966). The problem of universals in language. In Universals of language (p. 1–29). Cambridge, MA: MIT Press.

Ilgen, B., & Karaoglan, B. (2007). Investigation of Zipf's "law-of-meaning" on Turkish corpora. In 22nd International Symposium on Computer and Information Sciences (ISCIS 2007) (p. 1–6).

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., & Stumme, G. (2007). Tag recommendations in folksonomies. In J.N. Kok, J. Koronacki, R.L. de Mantaras, S. Matwin, D. Mladenič, & A. Skowron

(Eds.), Knowledge discovery in databases: PKDD 2007: 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17–21, 2007. Proceedings (pp. 506–514). Berlin, Heidelberg: Springer.

Kenett, Y., & Austerweil, J. (2016). Examining search processes in low and high creative individuals with random walks. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (p. 313–318).

Kolmogorov, A.N. (1956). Foundations of the theory of probability (2nd ed.). New York: Chelsea Publishing Company.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, and Computers, 28(2), 203–208.

Moreno-Sánchez, I., Font-Clos, F., & Corral, A. (2016, 01). Large-scale analysis of Zipf's law in English texts. PLoS One, 11(1), 1–19.

Newman, M.E.J. (2002). Assortative mixing in networks. Physical Review Letters, 89, 208701.

Newman, M.E.J. (2010). Networks. An introduction. Oxford, UK: Oxford University Press.

Noh, J.D., & Rieger, H. (2004). Random walks on complex networks. Physical Review Letters, 92, 118701.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: bringing order to the web (Tech. Rep.). Stanford, CA: Stanford Digital Library Technologies Project.

Pinker, S. (1999). Words and rules: The ingredients of language. New York: Perseus Books.

Poirier, D.J. (1995). Intermediate statistics and econometrics: A comparative approach. Cambridge, MA: MIT Press.

Ritz, C., & Streibig, J.C. (2008). Nonlinear regression with R. New York: Springer.

Saussure, F. (1916). Cours de linguistique générale (C. Bally, A. Sechehaye, & A. Riedlinger, Eds.). Lausanne and Paris: Payot.

Sinatra, R., Gómez-Gardeñes, J., Lambiotte, R., Nocosia, V., & Latora, V. (2011). Maximal-entropy random walks in complex networks with limited information. Physical Review E, 83, 030103(R).

Smith, K.A., Huber, D.E., & Vul, E. (2013). Multiply-constrained semantic search in the Remote Associates Test. Cognition, 128(1), 64–75.

Snodgrass, J.G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory, 6, 174–215.

Strauss, U., Grzybek, P., & Altmann, G. (2006). Word length and word frequency. In P. Grzybek (Ed.), Contributions to the science of text and language: Text, speech and language technology (Vol. 31, p. 277–294). Berlin: Springer.

Thompson, G., & Kello, C. (2014). Walking across Wikipedia: a scale-free network model of semantic memory retrieval. Frontiers in Psychology, 5, 86.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. Cognitive Psychology, 15, 121–149.

Zipf, G.K. (1945). The meaning-frequency relationship of words. Journal of General Psychology, 33, 251–266.

Zipf, G.K. (1949). Human behaviour and the principle of least effort. Cambridge, MA: Addison-Wesley.