On the shape of semantic space - what can we infer from large-scale statistical properties of texts?

Dániel CzégeL

with the supervision of Prof. Maxi San Miguel

Master’s Thesis

Master’s degree in Physics of Complex Systems

at the

UNIVERSITAT DE LES ILLES BALEARS

academic year 2015/16

February 12, 2017
Abstract

The large amount of digitized linguistic data opens up the unique possibility of using the methodology of complex systems to understand high-level human cognitive processes. Two such issues are (i) the way we categorize the continuous space of real-world features into discrete concepts, and (ii) the way we use language to copy a line a thought from one brain to another. In this work I address both questions by formulating a simple text generation model which reproduces the three major characteristic large-scale statistical laws of human language streams, namely Zipf’s law, Heaps’ law and Burstiness. Furthermore, the generation itself can be described as a random walk on a scale-free, highly clustered and low dimensional complex network, suggesting that this class of networks is appropriate as a minimal model of the semantic space. Entangling the global characteristics of the semantic space is an inevitable step towards analyzing texts as trajectories in such a space, with promising applications such as author or style identification, personal disorder diagnosis, or the evolution of cultural traits mirrored by text production characteristics.
Acknowledgements

I would like to express my gratitude to my supervisor, Maxi San Miguel, who gave useful suggestions from which I could benefit a lot, both regarding this project and research in general. Also I appreciate his openness for accepting my topic suggestion.

Some of the initial ideas came from the enlightening discussions with Pedro Parrado and Víctor Eguíluz.

Finally I would like to thank the opportunity of taking part in the high-quality master's program provided by IFISC, which I enjoyed a lot.
Contents

1 Introduction 4

2 Statistical laws of natural languages 7
   2.1 Zipf’s law 7
   2.2 Heaps’ law 8
   2.3 Burstiness 11

3 Models of text generation 15
   3.1 Zipf’s law and the principle of least effort 15
   3.2 The Yule-Simon process 15
   3.3 Sentence formation as a sample space reduction process 19

4 Text generation as Markovian dynamics on complex networks 20
   4.1 Correspondence between concepts of text statistics, random walks, and network topology 21
      4.1.1 Zipf’s law - stationary distribution - degree distribution 21
      4.1.2 Heaps’ law - cover time distribution - path lengths 21
      4.1.3 Burstiness - recurrence time distribution - path lengths 22
      4.1.4 First return time distribution: analytical approaches 23
      4.1.5 A summary of required network properties 26

5 Random walk on scale-free networks: numerical results 27
   5.1 Summary of the results 31

6 Discussion 31
1 Introduction

Language is one of the few phenomena that can be regarded as uniquely human [1,2], including our ability of creating and comprehending algebraic and recursive rules of grammar [3,4] as well as our many-level hierarchical organization of semantics where the lower level concepts are rooted in perceptual or motor skills [5,6]. One of the yet to be understood corresponding phenomenon is the way we categorize the continuous space of features of the world into discrete classes of concepts, which could reveal fundamental properties of how high level cognition works.

There have been many attempts to describe the organizational principles of words in our mind corresponding to these discrete concepts, the ones based on experimental data are mostly from the field of psycholinguistics [7,8]. Although these psycholinguistic models are motivated by experiments having quantitative results, like reaction times or error rates, the models themselves are mostly qualitative, which makes them hard to falsify. To illustrate these ideas, Figure 1 and 2 shows an example of an experimental method (lexical decision task to measure the semantic priming effect, see e.g., [9]) and of such a model (the Spreading Activation Model, [10]).

Recently, however, the explosion of the amount of digitalized linguistic data made it possible to address the question in a quantitative manner by the so-called distributional semantics models [12,13]. The basic idea behind these models is that two words are close to each other in a space if their neighborhood mostly consist of the same words. Figure 3 illustrates the process. This method locates the words in an ambient $N$-dimensional space, with $N$ being the number of distinct words in the text, which is enormous. Although one can use dimension reduction techniques such as singular value decomposition [14] to find the most important dimensions, still, the resulting distribution of words in possesses the artificial constraints of being embedded in a(n euclidean) space. Also one might wish to have a minimal model of the structure of the semantic space, for which a set of pairwise distances may not be appropriate. Instead, the theory of complex networks are more suitable as it offers both more flexible and more concise ways to (statistically) describe possible underlying structures.

Here, I would like to propose a new approach based only on large-scale statistical properties of large text databases (i.e., corpora). The idea is to find the simplest possible text generation model that i) reproduces the main observed statistical properties of real texts, ii) the generation itself can be considered as a continuous stochastic trajectory on some topological structure. To my best knowledge, this model is novel in two respects: it reproduces three of the most characteristic observed statistical regularities of real texts, namely, Zipf’s law, Heaps’ law, and Burstiness; and it relates these "laws" to topological properties, allowing us to make quantitative statements about the global structure of the semantic space. It is important to note,
Figure 1: Illustration of a lexical decision task to experimentally measure the priming effect of one word (here: bird, the priming word) to the other (here: eagle, the target word). The task of the subject is to decide whether the target word is an existing (i.e., correctly spelled) word or not. The hypothesis, which has been justified in many experimental conditions, is that people decide faster and have less error in case of "bird" → "eagle/aegle", than if there is no semantic connection between the words, e.g., "bird → desk/deks", even if the duration of showing the prime word is so short that the subject is not able to consciously perceive what the prime word is. Source of the figure: [11]
Figure 2: Illustration of the Spreading Activation Model of the mental lexicon. Adapted from [10].

Figure 3: Illustration of a distributional semantics method. Source: [15]
however, that the goal here is to understand what are the simplest possible mechanisms that can exhibit the above mentioned properties and not to capture cognitive processes in detail. Thus it does not aims to substitute previous (cognitive, linguistic) models, rather, it might serve as a fresh approach to complement those ideas for a better overall understanding.

The thesis is organized as follows. The statistical laws that are discussed throughout the work are introduced in section 2. Section 3 summarizes the most popular text generation models that are supposed to exhibit some of the observed statistical laws. Section 4 discusses the analogy between concepts of text generation and that of random walks on complex networks and it outlines the possible analytical approaches to compute relevant quantities describing random walks on networks. Section 5 describes the main novelty of the thesis, which is a text generation model based on random walks on structured scale-free networks, and shows numerical justification of the fact that it indeed reproduces the three main statistical laws discussed in section 2. Section 6 discusses the results in a general context and points at possible future directions.

2 Statistical laws of natural languages

There are quite a few statistical relationships measured in large corpora, taking different linguistic entities as units [10]. Since in this work we focus on the level of semantics, the ones that are of our interest are taking words (or, more preferably, lemmas) as units. Here I will analyze and model three of the most well-known and robust such laws, namely Zipf’s law, Heaps’ law and Burstiness, pertaining to the frequency of words, the rate of using new words, and the inter-event times of words in long texts, respectively.

2.1 Zipf’s law

The most well-known and robust statistical law regarding natural languages is Zipf’s law of word frequencies. As originally put by Zipf [17, 18], the relationship between a word’s frequency $f$ and its rank $r$ in a frequency order in a given text is

$$f(r) \propto r^{-z}$$

as illustrated in Figure 1. Equivalently, [1] can be phrased as

$$p(f) \propto f^{-z}$$

with $p(f)$ being the fraction of words appearing exactly $f$ times in the text.
The exponents $z_r$ and $z_f$ can be related by e.g., using the simple observation that the rank $r(f)$ of a word having frequency $f$ is just the number of words with higher frequency [19],

$$r(f) \propto \sum_{f' = f}^{\infty} p(f') \approx \int_{f}^{\infty} p(f') \, df' \propto f^{-z_f+1}$$  \hspace{1cm} (3)$$

which, compared with [1] yields

$$r(f) \propto f^{-1/z_r} \propto f^{1-z_f}$$  \hspace{1cm} (4)$$

consequently,

$$z_f = 1 + \frac{1}{z_r}$$  \hspace{1cm} (5)$$

Although the amount of linguistic data analyzed by Zipf (without using computer) was unsatisfactory to justify any statistical relationship in a robust way, later on, especially in the last two decades, Zipf’s law has been robustly observed [20, 16, 21, 22], also independently of modality (i.e., whether it is a written or spoken corpora), [23] and independently of the units chosen as words [24, 25], as illustrated in Figure 5.

Furthermore, in many corpora, taken words as units, $z_r \approx 1$, or equivalently, $z_f \approx 2$ [16, 21, 22, 24, 25].

### 2.2 Heaps’ law

Heaps’ law [27, 19, 16] refers to the sublinear growth of vocabulary size in the function of text length. Mathematically,

$$V(L) \propto L^h$$  \hspace{1cm} (6)$$

where $V(L)$ is the number of distinct words already used in the text of length $L$ (measured in the number of words), and $h < 1$ is the Heaps exponent. Figures [6] and [7] illustrates this relationship.

Interestingly, there is a relation between the Heaps and Zipf exponent in the limit of infinite text size, regardless of any structure (e.g., correlations) of the text.

By definition of the maximal frequency $f_{max}$,

$$\int_{f_{max}}^{\infty} p(f) \, df \propto V^{-1}$$  \hspace{1cm} (7)$$

Substituting Zipf’s law, $p(f) \propto f^{-z_f}$ and Heaps’ law $V = L^h$ yields

$$f_{max}^{-z_f+1} \propto V^{-1} = L^{-h}$$  \hspace{1cm} (8)$$
Figure 4: Frequency-rank plot of the words in the Brown corpus. Note that the Zipf’s law, \( f(r) \propto r^{-\alpha} \) with \( \alpha \approx 1 \), holds for multiple decades. Source: [26]

Figure 5: Illustration of Zipf’s law in Charles Dickens’s “Tale of Two Cities”, where the units are corresponding to different levels of linguistic organization. Note that the units here are defined algorithmically and only approximately mirror the actual linguistic units. Source: [24]
Figure 6: Number of different words (or, in other words, the vocabulary size) $V$ in the function of text length $L$, denoted here by $N$. The green line is corresponding to data from one given book (Moby Dick by H. Melville; the textlength should be understood as running textlength), whereas the red dots are corresponding to the articles from the English Wikipedia (here, textlengths are the actual textlengths). Source: [28]

Figure 7: Number of different words $V$ in the function of running text length $L$ for Don Quijote, David Copperfield, and Aeneid, as written in their original languages (Spanish, English and Latin, respectively). Source: [19]
Taking into account that the maximal frequency cannot be larger than the textlength,

\[ z_f \geq 1 + h \tag{9} \]

which is valid for any symbol (word) sequence that exhibits Zipf’s and Heaps’ law. Furthermore, if \( f_{\text{max}} \) scales as \( L \), which is highly sensible assumption (i.e., the relative frequency of the most frequent word is independent of the length of the text), we get an equality,

\[ z_f = 1 + h \tag{10} \]

This explicitly shows that there is no model which can reproduce Zipf’s and Heaps law with their exact exponents \( z_f = 2 \) and \( h \) significantly smaller than one in the limit of infinite text size. Consequently, any model that wants to reproduce these laws analytically has to take into account finite-size effects. The question now is if one is able to formulate a model where the finite-size effects are still dominant at large textlengths, e.g., \( L \approx 10^{10} \).

Lü et al. [29] investigated finite-size effects in Heaps’ law in this simple model (i.e., assuming Zipf’s law) both analytically and numerically, and found that indeed they are present even at relatively long texts, as it is shown in Figure 8. Another way to resolve this conflict would be to keep Heaps’ law with constant exponent \( h < 1 \) but introduce a cut-off (or multiple scaling regimes with different exponent) in Zipf’s law.

This phenomenon clearly hints at the fact that the usual procedure in statistical physics of creating a comprehensive theory at the infinite-size limit and then introducing consistent finite-size effects might fail as many relevant phenomena might disappear already at the infinite-size limit.

### 2.3 Burstiness

Another characteristic statistical feature of natural texts is the bursty appearance of words, reflecting the hierarchical topical large-scale organization of texts [30, 31, 32, 19, 33].

In general, burstiness refers to the fact that a given type of event takes place many times in a short interval, leaving large silent periods between those bursts. The two main characteristic statistical features of burstiness are long-range correlations and non-exponential recurrence time distributions [34]; here I will focus on first recurrence time distributions \( p_i(\tau) \) defined as the distribution of the number of words \( \tau \) between two subsequent appearance of word \( i \) in the text.

Interestingly, in case of various long enough continuous texts, \( p_i(\tau) \) can be well fitted by the two-parameter Weibull-distribution [31]

\[ p(\tau; \langle \tau \rangle, \beta) \propto \tau^{\beta - 1} e^{-\alpha \tau^\beta} \tag{11} \]
Figure 8: Finite size effects in systems exhibiting Zipf’s and Heaps’ law simultaneously. In this figure the Heaps exponent \( h \) is denoted by \( \lambda \) and the textlength \( L \) is denoted by \( t \). Source: [29]

with \( a(\langle \tau \rangle, \beta) = \left[ \frac{1}{\langle \tau \rangle} \Gamma \left( \frac{\beta+1}{\beta} \right) \right]^\beta \), which interpolates between the exponential distribution (\( \beta = 1 \)), corresponding to a homogeneous Poisson process (the given word appears in every slot with the same probability), and a power-law \( p(\tau) \propto \tau^{-1} (\beta \to 0) \), corresponding to a scale-invariant appearance of words with the fattest possible tail a normalized probability distribution can have. Thus, the burstiness of a given word can be nicely quantified by the \( \beta \) parameter of the fitted Weibull distribution\(^1\). Unfortunately this fitting method is not feasible if the frequency of a given word, and consequently, the number of data points, is low. One way to remedy this is to only take into account frequent words; however, according to Zipf’s law, this way we get rid of the vast majority of the vocabulary. On the other hand, an alternative scalar measure of burstiness from \( p(\tau) \), already used in this context by [30], is its relative standard deviation \(^2\):

\[
s = \frac{\sigma_\tau}{\langle \tau \rangle} = \sqrt{\frac{\langle \tau^2 \rangle - \langle \tau \rangle^2}{\langle \tau \rangle}}
\]  

(12)

as it is a dimensionless quantity and also \( s \equiv 1 \) for the exponential family\(^3\). As it has been emphasized, burstiness enlarges the inhomogeneity

\(^1\)The other parameter, \( \langle \tau \rangle \) is inversely proportional to the frequency of the word, consequently, the only parameter to be fitted is \( \beta \).

\(^2\)Also called the coefficient of variation

\(^3\)Note that the exponential inter-event time distribution corresponding to the most random arrangement (i.e., the given word appears at each position with the same probability, independently) is only recovered in the infinite text-size limit. Recently it was shown\(^3\) that if a word occurs \( f_t \) times in a text having length \( L \), the relative standard deviation of the recurrence time distribution corresponding to the most random configuration is
Figure 9: Visualization of the positions of the words *island*, *instinct* and *however* in Charles Darwin’s ‘On the Origin of Species’. Source: [19]

of the $\tau$ values, consequently, $s > 1$ ($s < 1$) indicate bursty (anti-bursty) behavior of the word.

Figure 9 and 10 illustrate the typical occurrence patterns of words in long texts.

The relation between the two above-mentioned measure of burstiness, $\beta$ and $s$, calculated simply from the mean and the variance of the Weibull distribution, is

$$s = \frac{\sqrt{\Gamma(1 + 2/\beta)} - [\Gamma(1 + 1/\beta)]^2}{\Gamma(1 + 1/\beta)}$$

plotted in Figure 11.

Burstiness and a related phenomena, the long-range correlation between the elements, are mostly explained by the hierarchical topical organization, characteristic to any text [36, 32, 33]. An explicit minimal model of this phenomenon might be a tree-graph of random variables, with a given level of dependence (measured as e.g., correlation or mutual information) between

$$s_0(f_i, L) = \sqrt{\frac{L}{(L+1)(L+2)}}$$

which indeed tends to 1 as $L, f_i \to \infty$ such that $L/f_i \to \infty$.

Thus, in finite-size systems, the authors recommend to use $\tilde{s} = s/s_0$ as a measure of burstiness. For the sake of simplicity, and also since we consider long texts in general, we will simply use $s$ in this work.

$^4$measured by e.g., the power-law decay of mutual information between the elements [39], the anomalous diffusion of the "cumulative" associated random walker of a word $i$, $X_i(t) = \sum_k \delta(t,t_k)$ with $t_k$ being the positions of word $i$ in the text [32], or the power-law decay of correlation of the position in the trajectory corresponding to the ongoing text in a semantic space constructed by distributional semantic methods [33].
Figure 10: Analysis of burstiness of different linguistic units in the novel “War and Peace” by L. Tolstoy. \( \hat{\gamma} \) is the finite-time estimate of \( \gamma \) in \( \sigma^2(t) \propto t^\gamma \), where \( \sigma^2(t) \) is the variance of the associated random walk of the process (see [32] for details); \( s = \frac{\sigma^2}{\langle \tau \rangle} \) is the relative standard deviation of \( p(\tau) \). Source: [32]

Figure 11: Log-linear plot of the relative standard deviation \( s \) and the parameter \( \beta \) of the Weibull distribution.
those variables that are connected by edges, yielding an exponential decay of dependence in the function of distance. However, the distance between two leaf nodes scales as the logarithm of the separation of the leafs placed next to each other (forming a symbol sequence). These two effects result in a power-law decay of dependence between leaf-nodes in the function of their separation [36].

Nevertheless, the only model that accounts for the form (11) of the inter-event time distribution does not use this idea; instead, it evokes survival analysis and treats the sequence of appearance times of a single word the result of a non-uniform Poisson process [31]. Interestingly, if the rate $\lambda_i(t)$ corresponding to word $i$ decays as a power-law since the last appearance of the word (which can be interpreted as a "forgetting rate" of the word), the inter-event times will follow the measured Weibull distribution.

3 Models of text generation

The observation of statistical laws in language streams induced the need for generative or non-generative models to explain them. The central, not yet resolved debate is how much assumption should be made from outside (e.g., explained by cognitive mechanisms), and how large part of these statistical laws are of purely mathematical nature. In the following, I outline three major approaches of modeling these laws.

3.1 Zipf’s law and the principle of least effort.

Zipf himself qualitatively explained the observed frequency distribution of words by an optimization process between the speaker and the listener [17]: they want to minimize the overall effort of communication. This means, according to Zipf, that the speaker would like to use as few number of different words as possible to convey the meaning, on the other hand, the listener needs as many words as possible to resolve ambiguities.

Interestingly, this argument can be phrased in a simple quantitative framework [37], which shows that at a special value of the weight ratio of the speakers and the listeners effort, there is a 'communicative phrase transition' from no information transmission to a language having one-to-one correspondence between words and meanings (see Figure 12). At this critical point, the frequency distribution of words display Zipf’s law, with exponent $z_r = 1$ which is in accordance with the exponent measured in real texts.

3.2 The Yule-Simon process

Besides being dynamic and also new words appear time to time, we have the intuition about text production that the words that have been used (and
Figure 12: Schematic representation of numerical results from the quantitative model for the principle of least effort applied to the evolution of language [37]. λ is the tunable parameter weighting the speakers effort and the listeners effort in the overall cost function. Mutual information refers to the mutual information between words and meaning, whereas the relative lexicon size is the ratio between the actually used words and the total number of possible words, see [19] for details. The inset shows the frequency-rank distribution of words at the critical point $\lambda = 0.41$. Source: [19].
consequently heard) more are used more later on. This motivates the idea of modelling text generation by a Yule-Simon process [38], as described below.

Let us first see what are the assumptions we need in order to obtain Zipf’s law with its exponent $z_f = 2$, and also whether it is possible to incorporate the decreasing rate of newly appearing elements (words), namely, Heaps’ law. In the following I generalize the continuous-time, mean-field approach used by Barabási and Albert in the context of scale-free networks [39], to give an approximate but simpler explanation of the formula derived by Montemurro and Zanette [40].

The model is the following. At every timestep $t = 1, 2, 3...$, corresponding to an uttered or written word, we either add a word with time-dependent probability $\alpha(t)$, or we use one of the already used ones with probability $1 - \alpha(t)$. Furthermore, in this latter case, the probability of choosing a word $i$ is proportional to its frequency $f_i$. What we are interested in is the resulting probability density $p(f)$ of word frequencies at large times (or equivalently, long texts).

The first step is to find how the frequency $f_i(t; \tau)$ of word $i$, introduced at time $\tau$, grows with time $t$. Here we neglect fluctuations, and approximate $f(t; \tau) - f(t - 1; \tau)$ by $\partial_t f(t; \tau)$ to obtain the corresponding differential equation,

$$\partial_t f(t; \tau) = (1 - \alpha(t)) \frac{f(t; \tau)}{t}$$

(14)

The rationale behind this is that at every timestep, the frequency of our word increases on average by $(1 - \alpha(t))$ times the probability of choosing our word from the already existing ones, namely, $\frac{f(t; \tau)}{t}$, with $t$ being the sum of the frequencies so far (as we write one word at a timestep).

Let us first outline the main steps of our approach. Once we have the time evolution of the frequency of the word which has been introduced at $\tau$, denoted by $f(t; \tau)$, we invert it to obtain $\tau(f, t)$. This inversion is the key step in order to obtain the probability density of word frequencies $p(f, t)$ at time $t$, since we can obtain $p(f, t)$ as the product of the rate $\alpha(t)$ by which new words are introduced, and the Jacobian of the transformation $\tau \rightarrow f$,

$$p(f, t) \propto \alpha \left[ \tau(f, t) \right] \left| \frac{\partial \tau(f, t)}{\partial f} \right|$$

(15)

This can be regarded as a simple transformation of random variables, the only difference is that the function $\alpha(t)$ needs not to be normalizable.

Now consider special choices of $\alpha(t)$, first a constant one

\footnote{I did not found this derivation in the literature, however, it might happen that it is actually already described given its simplicity and the amount of work focused on explaining power laws.}

\footnote{since either zero or one word is added at each timestep, $\tau$ also serves as a unique identifier of the introduced word}
\[ \alpha(t) \equiv \alpha \]  
(16)

In this case the solution of (14) with initial condition \( f(t = \tau) = 1 \) is

\[ f(t, \tau) = \left( \frac{t}{\tau} \right)^{1-\alpha} \]  
(17)

inverting this yields

\[ \tau(f, t) = tf^{\frac{1}{\alpha-1}} \]  
(18)

Since \( \alpha \) is constant here,

\[ p(f, t) \propto \left| \frac{\partial \tau}{\partial f} \right| \propto f^{\frac{1}{\alpha-1}-1} \]  
(19)

In particular, for the Barabási-Albert model \[39\], where \( \alpha = 1/2 \) we get \( p(f) \propto f^{-3} \). More interestingly from our point of view, Zipf’s law \( p(f) \propto f^{-z_f} \) with \( z_f \approx 2 \) can be obtained for \( \alpha \approx 0 \), that is, the rate of introducing new words is very small.

However, in reality, \( \alpha(t) \) is not constant; instead, it follows Heaps’ law,

\[ \alpha(t) = \alpha_0 t^{h-1} \]  
(20)

being the derivative of the vocabulary growth rate \( V(t) \propto t^h \), with \( 0 < h < 1 \).

Simple integration together with imposing the same initial condition as before, \( f(t = \tau) = 1 \), yields

\[ f(t, \tau) = \frac{t}{\tau} e^{-\alpha_0 \left( (t^{h-1} - \tau^{h-1}) \right)} \]  
(21)

Its inverse, \( \tau(f, t) \), turns out the be

\[ \tau(f, t) = -\alpha_0 \frac{1}{e^{\frac{1}{n}} \left( W(X) \right)^{\frac{1}{n}}} \]  
(22)

where \( X = -\alpha_0 e^{-\alpha_0 h^{-1}} \left( \frac{t}{\tau} \right)^{1-h} \) and \( W(z) \) is the principal part of the Lambert-W function, defined by \( z = W(ze^z) \).

By algebraic manipulations and differentiation,

\[ p(f, t) \propto \alpha(\tau) \left| \frac{\partial \tau}{\partial f} \right| \propto -\alpha_0 \frac{2^{\frac{1}{n}} f^{-1} \left( -W(X) \right)^{\frac{1}{n}}}{1 + W(X)} \]  
(23)

with \( X \) being the same as before.

\[ \frac{7}{7} \]since half of the newly introduced degrees belong to the new node and half of them belong to the old ones.
This seems complicated, however, in the limit of $\frac{f}{t} \ll 1$ (which can be realized for not-so frequent words and for long texts), $X$ is very small, and consequently, $W(X)$ can be approximated by its Taylor-series, $W(X) = X + \mathcal{O}(X^2)$, yielding

$$p(f, t >> 1) \propto f^{-(1+h)}$$  \hspace{1cm} (24)

and consequently

$$z_f = 1 + h$$ \hspace{1cm} (25)

The relation between the Zipf and Heaps exponent obtained from this model is the same as [10], the one we got at infinite system size solely from the simple scaling relations without assuming any dynamics. This validates the assumptions (mean-field, continuous time) we made. On the other hand, it shows again that the one has to take into account finite-size effect to come up with a model that reproduces both Zipf and Heaps law at the same time. However, analyzing the finite-size solution of (23), that is, the case when the frequency of any of the words is non-negligible compared to the length of the text, is beyond the scope of this work.

### 3.3 Sentence formation as a sample space reduction process

A recent model of sentence formation, exhibiting Zipf’s law with exponent $z_r = 1$, is based on the idea that although the selection of words is a stochastic process, it is constrained by the gradual reduction of sample space during sentence formation, as we as speakers narrow the listener’s possible interpretations of our utterance [41, 42]. The process is illustrated in Figure 13.

Just like the Yule-Simon model, this procedure, being general and robust, also has the potential to explain many natural phenomena where a system...
variable exhibits power-law distribution. The most general formulation, illustrated in Figure 14 is the following [42, 43]: Take any directed acyclic network, choose an initial node from a wide variety of "prior" distributions, and do a random walk on the nodes such that the walker follows the possible outgoing with uniform probability. As the process unfolds, the number of nodes that can be visited later on (i.e., the sample space volume) decreases, eventually each walker converging to the root node(s). Once this happens, we start the process again. It can be showed via analytical and numeric calculations, that if one starts from a prior distribution that is "widespread enough", the relative frequency $f$ of visiting a node will be inversely proportional to its rank in this frequency distribution, that is,

$$f(r) \propto r^{-z_r}$$

with $z_r = 1$, just as in case of the real texts. Also, this relationship is robust against noise being for example in the form of random directed cycles in the network. For details, see [43].

4 Text generation as Markovian dynamics on complex networks

In this section I propose a simple text generation model which is based on random walks on complex networks: a unique word is assigned to every node of a network, and at every discrete timestep, the word is written/uttered that is corresponding to the current position of the walker. In order to do so, first, I make correspondences between statistical topological properties of complex networks and the statistical features of simple Markovian dynamics (a random walk with uniform transition probabilities) on it. The goal is to investigate the question whether it is indeed possible to find any topology on which a random walk as text generation in the sense described above can
reproduce all of the main statistical laws of text at the same time (namely, Zipf’s law, Heap’s law and Burstiness), and if yes, what (statistical) classes of networks are appropriate. Such a process might serve as a minimal model of text generation.

4.1 Correspondence between concepts of text statistics, random walks, and network topology

The random walk process is the following: at every discrete timestep $t = 1, 2, 3, \ldots$ the walker steps to one of the neighboring nodes with equal probability. Since the network is unweighted and undirected, this corresponds to a Markov chain with transition probabilities $T(i \to j) = k_i^{-1}$, $k_i$ being the degree of node $i$, if $i$ is connected to $j$ and $T(i \to j) = 0$ otherwise. Also, for networks that are connected and not bipartite, the so obtained Markov chain is ergodic, a property which is assumed in the following.

4.1.1 Zipf’s law - stationary distribution - degree distribution

Probably the simplest and most well studied such correspondence is between the frequency distribution $f_i$ of the elements of the resulting sequence, the stationary probability distribution $p_i$ of the process and the transition probabilities $T(i \to j)$. By the definition of $p_i$,

$$
limit_{t \to \infty} f_i \propto p_i \quad (27)
$$

On the other hand, from the detailed balance condition with $T(i \to j) = k_i^{-1}$

$$
p_i \propto k_i \quad (28)
$$

Consequently, in a long run, the obtained sequence will reproduce Zipf’s law $p(f) \propto f^{-z_f}$ if the network is scale-free with exponent $z_f$, that is, if the degree distribution is $p(k) \propto k^{-z_f}$.

4.1.2 Heaps’ law - cover time distribution - path lengths

Heaps’s law, which is the number of distinct words $V(L)$ used in the function of textlength $L$ (measured also in words) can be rephrased in the language of random walks as the cover time distribution $N_{cov}(t)$, that is, the number of nodes visited in the function of the number of timesteps.

How can we relate $N_{cov}$ to network properties? One possible characterisation is through dimensionality: as it is shown in [41], for $d$-dimensional infinite regular hypercubic lattices, $N_{cov}(t) \propto t^{1/2}$ for $d = 1$, $N_{cov}(t) \propto t / \ln t$ for $d = 2$, and $N_{cov}(t) \propto t$ for $d > 2$.

In accordance with the previous argument, Almaas et al. [45] found that for the Watts-Strogatz (1-d) model, $N_{cov}(t) \propto t^{1/2}$ for $t << p^{-2}$ (where $p$ is
the rewiring probability) and $N_{cov}(t) \propto t$ for larger times when the walker starts to feel the effect of shortcuts.

Although dimensionality can be properly defined in case of complex networks, basically by measuring the scaling of the number of nodes in a given distance from a node, see e.g., [46], it is more instructive here to choose a related but simpler measure, namely, the average path length $L$, defined as the average of the length of the shortest path(s) between any node pairs.

The connection between the average path length scaling $L(N)$ (where $N$ is the number of nodes in the network) and the dimension $d$, at least in case of regular lattices, is $L \propto N^{1/d}$. On the other paradigmatic extreme, for Erdős-Rényi networks, $L \propto \ln N$ indicating that it is infinite dimensional.

Consequently, a conjecture is that the Heaps’s law $V(L) \propto L^h$ with exponent $h < 1$ is realized in low-dimensional networks, or correspondingly, networks with high average path length.

4.1.3 Burstiness - recurrence time distribution - path lengths

Inter-event time distribution of words can be translated as the recurrence time distribution (or first return time distribution) $p(\tau)$ of random walks. Here we have even less clue as according to my best knowledge no paper investigated this in complex networks. On the other hand, return time distributions, that is, probabilities describing the event that the random walker is in the starting node at time $t$, irrespectively of how many times it has already been there, has been studied, which can provide hints to $p(\tau)$ as well (the analytical connection between the two will be sketched below).

However, we might gain intuition from somewhat analogous systems. It is known that for a simple one dimensional random walk on a lattice, the recurrence time distribution is asymptotically $p(\tau) \propto \tau^{-3/2}$ [47].

On the other hand, in a complete network, where every node is connected to every other, every node is in unit distance from node $i$ (for all $i$), and also the degree of each node is the same, which implies that the probability $q$ of visiting node $i$ is the same at every time step and independent from the past, corresponding to a homogeneous (discrete) Poisson process, with $p(\tau)$ exhibiting a geometric distribution,

$$p(\tau) = (1 - q)^{\tau - 1}q$$

having mean $\langle \tau \rangle = q^{-1}$ and standard deviation $\sigma_\tau = q^{-1}\sqrt{(1 - q)} \approx q^{-1}$ for large networks (and consequently small $q = k^{-1} = 1/(N - 1)$). This implies that for a large full network,

$$s = \frac{\sigma_\tau}{\langle \tau \rangle} \approx 1$$

In order to intuitively interpolate between these extremes, let us do some basic calculations on an idealized model. This idealized model has infinite
nodes, an arbitrary degree distribution, and the following property: for every node \( i \), the number of nodes \( A_d \) being at distance \( d \) has a definite scaling behavior. We will consider here two cases: either \( A_d \propto e^{\mu d} \), with some \( \mu > 0 \), as in the case of the configuration model \([48]\), or \( A_d \propto d^{D-1} \) with some finite \( D \), which can be regarded as the dimension of the network.

Assuming that there is no correlation between the degree of the nodes and their distance from node \( i \), the random walk’s stationary probability \( P_d \) of being at distance \( d \) from node \( i \) is proportional to the number of nodes at distance \( d \), that is, \( A_d \propto P_d \). This can be realized such that the detailed balance condition holds if the transition probabilities \( T(d \rightarrow d') \) satisfy

\[
\frac{A_d}{A_{d'}} = \frac{T(d' \rightarrow d)}{T(d \rightarrow d')} := r
\]

if \( d' = d \pm 1 \) and zero otherwise. Now let us investigate the \( d >> 1 \) limit. If the scaling is exponential (making the network infinite-dimensional), then \( r = e^\mu \) is independent of \( d \); however if the network if finite-dimensional, \( r \rightarrow 1 \), makes the whole system very similar to the one-dimensional lattice at \( d >> 1 \).

Again, these arguments suggest that the appropriate network we should implement the random walk process on is a low-dimensional one and correspondingly, one that has large average path length.

### 4.1.4 First return time distribution: analytical approaches

In this subsection we briefly describe an analytical method for calculating the first return time distribution \( p(\tau) \) in any complex network. This procedure consist of two steps. First, one has to compute the (not first!) return time distribution, defined as

\[
p_0(t) = \frac{1}{N} \sum_i p(i, t|i, 0)
\]

where \( N \) is the number of nodes, and \( p(i, t|i, 0) \) is the probability that the walker is at node \( i \) at time \( t \), given that it was there at time \( 0 \), too.

Now, as described in \([48]\), \( p_0(t) \) can be calculated from the is the spectral density

\[
\rho(\lambda) = \frac{1}{N} \sum_i \delta(\lambda - \lambda_i)
\]

of the modified Laplacian

\[
L_{ij} = \delta_{ij} - \frac{A_{ij}}{k_j}
\]

of the network. Once \( p_0(t) \) is calculated, the second step of the procedure is to compute the first return time distribution \( p(\tau) \) from \( p(t) \).
Although in our case this analytic approach is not particularly helpful as it requires the knowledge of the spectral density $\rho(\lambda)$, the line of reasoning can be instructive whenever one wants to connect topological and dynamical properties of a network. Furthermore, averaging over network realizations in (32) and (33) makes this framework suitable for analytic calculations regarding whole statistical network classes.

After outlining of the reasoning above, here we provide a more step-by-step derivation of $p_0(t)$, based on [48]. For the sake of simplicity, here we use a continuous-time description of the random walk process, described by the master equation

$$\partial_t p_i(t) = \sum_j W(j \to i)p_j(t) - W(i \to j)p_i(t)$$  (35)

where $p_i(t)$ is the probability of the walker being at node $i$ at time $t$ and transition rates are

$$W(i \to j) = \frac{A_{ij}}{k_j}$$  (36)

which can be written, in case of a random walk on a network defined by its adjacency matrix $A$, by means of the modified Laplacian $L$ as

$$\partial_t p_i = -\sum_j L_{ij}p_j$$  (37)

In order to solve (37) with initial condition $p_j(0) = \delta_{ij}$ and average over starting node $i$, it is useful to work in "Laplace-space", that is, we first find the solution for the Laplace transform of $p(i,t|j,0)$

$$\tilde{p}_{ij}(s) = \int_0^\infty dt e^{-st} p(i,t|j,0)$$  (38)

and then use the inverse Laplace-transform to get the solution in real space. Substituting (38) to (37) yields the solution in Laplace space

$$\tilde{\rho}(s) = (sI + L)^{-1}$$  (39)

with $I$ being the identity matrix. Writing $\tilde{\rho}_0$ as a trace of $\tilde{\rho}$ allows us to connect $p_0$ with the spectrum of $L$

$$\tilde{\rho}_0 = \sum_i \tilde{\rho}_{ii}(s) = \text{Tr} \tilde{\rho} = \sum_i \frac{1}{s + \lambda_i}$$  (40)

which gives, after performing the inverse Laplace-transform, that $p_0$ is simply the Laplace transform of the spectral density $\rho$

$$p_0(t) = \frac{1}{N} \sum_i p(i,t|i,0) = \frac{1}{N} \sum_i e^{-\lambda t} = \int_0^\infty d\lambda \rho(\lambda) e^{-\lambda t}$$  (41)
Since $-L$ is a stochastic matrix in the sense that its non-diagonal elements are non-negative and $\sum_i L_{ij} = 0$, it has a zero eigenvalue corresponding to the stationary solution, and all the other eigenvalues are strictly negative. Consequently, the long-time behavior of $p_0(t)$ is determined by the asymptotic shape of $\rho(\lambda)$ with $\lambda \to 0$.

To sum up, if the asymptotic shape of the spectral density function $\rho(\lambda)$ of the modified Laplacian $L_{ij} = \delta_{ij} - \frac{A_{ij}}{k_j}$ is known, it is possible to analytically calculate the long-time behavior of the node-averaged return time distribution $p_0(t)$. Below we list some paradigmatic examples, following [48].

- For a $D$-dimensional regular lattice, asymptotically $\rho(\lambda) \sim \lambda^{D^2-1}$ [49], which yields $p_0(t) \sim t^{-D/2}$.

- For an Erdős-Rényi network, asymptotically $\rho(\lambda) \sim e^{-c\lambda - 1/2}$ [50], leading to $p_0(t) \sim e^{-\tilde{c}t^{1/3}}$, where $c$ and $\tilde{c}$ are constants determined by the topology of the specific network.

- For a degree-uncorrelated random scale-free network (exhibiting locally tree-like structure), having minimum degree $m = 1$ or $m = 2$, $\rho(\lambda) \sim \lambda^{-\xi(m)} e^{-a\lambda^{-1/2}}$ [51]. This implies $p_0(t) \sim t^{\eta(m)} e^{-bt^{1/3}}$ [51].

This shows that a finite-dimensional lattice with long average path-length yields a pure power-law return-time distribution $p_0(t)$, as opposed to the most important random network models (having path-lengths that scale as $\log N$) whose $p_0(t)$ exhibits stretched exponential cutoff.

However, we are not done yet: in order to connect this theory and our framework based on first return time distribution $p$ (as this is the one which has been extensively measured in texts [47], we need to compute the $p$ from $p_0$. Here we sketch a general way to do this, thoroughly described in [52]. Note that is a mean-field-like approximation of $p$ since in the definition of $p_0$ we already averaged over the nodes.

The key idea is that the return probability at time $t$ can be written by means of the probability of being back for the first time at time $\tau$ and then, independently from this, being back again at $t$ as a convolution

$$p_0(t) = \delta(t) + \int_0^\infty d\tau p(\tau)p_0(t-\tau)$$

where the $\delta(t)$ term is corresponding to the initial condition. Note that the interval $(t, \infty)$ gives no contribution to the integral, however, this way

---

8 still assuming that the process is ergodic which implies that the stationary solution is unique

9 $\xi = 9/10$ for $m = 1$ and $\xi = 4/3$ for $m = 2$

10 $\eta = -7/30$ for $m = 1$ and $\eta = 1/18$ for $m = 2$

11 One apparently needs shorter texts to investigate $p$ than to $p_0$, on the other hand, correlations between the subsequent inter-event times are being lost.
we can use an appropriate integral transform to turn the convolution into a product. Since the distributions are defined for non-negative values, this appropriate transform is the Laplace-transform $\mathcal{L}$,

$$\mathcal{L} p_0 = 1 + \mathcal{L} p \cdot \mathcal{L} p_0$$  \hspace{1cm} (43)

Consequently,

$$\mathcal{L} p = 1 - \frac{1}{\mathcal{L} p_0}$$  \hspace{1cm} (44)

and thus formally the way one obtains the first return time distribution $p$ from the return time distribution $p_0$ is

$$p = \mathcal{L}^{-1} \left( 1 - \frac{1}{\mathcal{L} p_0} \right)$$  \hspace{1cm} (45)

In sum, the analytic approach faces us with two main obstacles, one being the calculation of the spectral density $\rho$ of the modified Laplacian, and the other being the technical difficulty computing Laplace and inverse Laplace transforms in order to obtain $p(\tau)$ from $p_0(t)$.

### 4.1.5 A summary of required network properties

As a summary, Zipf’s law clearly requires the network to exhibit a power-law degree distribution, Heaps’ law and Burstiness only hints at the fact that the network we use should be one with low dimensionality/large average path length. This suggests that the most proper model for our goal is a simple growing network model proposed by Klemm and Eguiluz [53], called the structured scale free (SSF) network, which has a scale-free degree distribution, the average path length $L$ scales as the network size $N$, and has dimensionality close to 1. [54] [46] Also it exhibits high clustering $C \propto \text{const}$, as opposed to an Erdős-Rényi network having $C \propto N^{-1}$ and a Barabási-Albert network having $C \propto (\ln N)^2 N^{-1}$ [55]. The choice of the class of SSF networks might be nicely justified by following the analytical steps of calculating $p(\tau)$. However, at least until this point I could not find such an analytical description nor I found the parts in the literature. Consequently, I will follow a numerical approach here.

In order to investigate the effect of average path length and possibly that of clustering, I implemented an SSF network with edge rewiring, as described in [50], by which it is possible to sweep through a class of scale-free networks and see how the aforementioned quantities affect the statistical properties of random walks.
5 Random walk on scale-free networks: numerical results

The generation of the network on which I ran the random walk simulation consisted of two main steps:

1. Generation of a structured scale-free (SSF) network. Starting from an initial seed of $m$ fully connected nodes being in the 'active state', at every time step one new active node with $m$ edges is added. This new node connects to the active nodes. As a final action of the time step, we deactivate one node, being chosen from the active ones with probability $\Pi(k) \propto k^{-1}$, $k$ being the degree of the node. Note that at every time step one node is activated and one is deactivated, so the number of active nodes has a constant value $m$, in accordance with the fact that the new node enters the network with $m$ links.

Also, according to [53], the degree distribution $p_k$ of the so-obtained network in the limit of large network size is a scale-free one with exponent $\gamma = -2$, that is, $p_k \propto k^{-2}$ which is consistent with Zipf’s law; even the exponent agrees.

Also it has large average path length, namely, $L(N) \propto N$ [54], dimensionality close to 1 [46] and high clustering coefficient $C(N) \approx 5/6$ [54].

2. Rewiring of the edges.

After the SSF network has been generated, I implemented a simple edge rewiring procedure by which the degree distribution is preserved, but the topological properties approach the maximum entropy configuration (with the given degree distribution as a constraint). This edge rewiring process is the following: at every step, a edge pair has been chosen randomly (with uniform probability over the edges) such that they do not share any common nodes, and one end of both edges is "rewired" to the node which was the endpoint of the other edge initially, given that the resulted network still does not contain any multiple edges or self-loops. This process has been repeated until every edge has been rewired on average $p$ times. As $p \to \infty$ would yield the configuration model (that is, the maximum entropy model) with the prescribed degree distribution, by changing $p$, one can investigate the transition between the SSF network and the most random scale-free network (having the same degree distribution as the SSF had).

Once the network has been generated, I measured the following (already known) topological quantities on them:
• Average path length $L$,

$$L = \langle \ell \rangle_{\text{network realizations, node pairs}}$$  \hspace{1cm} (46)

where $\ell$ is the length of the shortest path(s) between the given node pair.

• Clustering coefficient $C$,

$$C = \langle c \rangle_{\text{network realizations, nodes}}$$  \hspace{1cm} (47)

where $c(i)$ is the local clustering coefficient of node $i$ measured as the fraction of links between $i$’s neighbors that are actually exist.

Then, I simulated a random walk process on the generated networks with transition probabilities $P(i \to j) = k_i^{-1}$ if $i$ and $j$ are connected and $P(i \to j) = 0$ otherwise. The quantities I measured to describe the random walk process are the following:

• Average relative standard deviation of the first recurrence time distribution,

$$S = \langle s \rangle_{\text{network realizations, nodes}}$$  \hspace{1cm} (48)

where $s = \frac{\sigma_\tau}{\langle \tau \rangle}$.

• Heaps plot,

$$\langle N_{\text{cov}}(t) \rangle_{\text{network realizations, random walk processes}}$$  \hspace{1cm} (49)

with $N_{\text{cov}}$ being the number of distinct nodes visited by the random walker by time $t$.

• Heaps exponent $\beta$, from fitting the $N_{\text{cov}}(t)$ function with a power-law

$$N_{\text{cov}} \propto t^\beta$$  \hspace{1cm} (50)

• Zipf plot,

$$\langle p(f) \rangle_{\text{network realizations}}$$  \hspace{1cm} (51)

where $p(f)$ is the relative frequency of nodes that are visited $f$ times during the random walk.

The results of the simulations are shown in case of $N = 10^3$, $m = 5$ in Figures 15, 16 and 17. The corresponding $S$ values are shown in Table 1.
Figure 15: Average relative standard deviation of first recurrence times $S$ and the fitted Heaps exponent $\beta = h$ in the function of edge rewiring probability $p$, compared to the already known behaviour of average path length $L$ and average clustering coefficient $C$. Since the changes in $S$ follow the changes in the average path length $L$ rather than that of the clustering coefficient $C$, we conclude that fact that this model reproduces Burstiness is the consequence of the long average path lengths, and not of the high clustering. Note that $S$, $L$ and $C$ are normalized to the value corresponding to $p = 0$. Also note the slight increase in the Heaps exponent $\beta$ from around 0.8 to 0.95 as $p$ changes from 0 to 1. All shown values are averaged over 10 network realizations with (in case of $S$ and $\beta$) a random walk process of $10^6$ steps on each network.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$10^{-6}$</th>
<th>$10^{-5}$</th>
<th>$10^{-4}$</th>
<th>$10^{-3}$</th>
<th>$10^{-2}$</th>
<th>$10^{-1}$</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>2.27</td>
<td>2.11</td>
<td>2.07</td>
<td>1.86</td>
<td>1.42</td>
<td>1.17</td>
<td>1.10</td>
</tr>
</tbody>
</table>

Table 1: Average relative standard deviation of first recurrence times $S$ for the different $p$ edge rewiring probability values. The values are corresponding to the datapoints of figure 15.
Figure 16: Histogram of node visiting frequencies $p(f)$, which is equivalent to the Zipf plot in terms of language laws. The solid line corresponds to $p(f) \propto f^{-2}$. Note that theoretically $p(f)$ is independent of the rewiring probability $p$ since the rewiring does not change the network’s degree distribution. The values are obtained by averaging over 100 network realizations with a random walk process of $10^6$ steps on each network.

Figure 17: Cover time distribution $N_{cov}(t)$, which is equivalent to the Heaps plot $V(L)$, with $p = 10^{-6}$ (left panel), $p = 1$ (right panel). The values are corresponding to averages over 100 network realizations. The dashed line shows $N_{cov}(t) = t$. 

30
5.1 Summary of the results

As Figure 15, 16 and 17 shows, if the edge rewiring probability $p$ (which can be interpreted as e.g., noise) is small enough, the generated text displays all of the three required statistical properties. The Zipf’s law $p(f) \propto f^{-z_f}$ holds with exponent $z_f$ close to 2, the Heaps’ law $V(L) \propto L^{-h}$ is also justified with a Heaps exponent $h \approx 0.8$ significantly smaller than one, in correspondence with the measured exponent values in real texts. The words in the text generated by the model at low $p$ also shows bursty appearance, as the average relative standard deviation of inter-event time distributions $S$ is much higher than 1, see Table 1.

Furthermore, the texts exhibits a change towards a non-bursty behaviour with $S \approx 1$ along with a too high Heaps exponent $h \approx 0.95$ around $p \approx 10^{-2}$. In particular, the $S$ burstiness value follows the average path length $L$ rather than the average clustering coefficient $C$, showing that if any of these two characteristic network properties is connected to burstiness, it is the average path length and not the clustering, as it has been proposed earlier in this work.

In the other extreme of the scale, when every link is rewired on average $p = 1$ times, we obtain scale-free network with small average path length and low clustering (similarly to the Barabási-Albert model [39]). The generated text corresponding to this class of networks still exhibits Zipf’s law (as the network is still scale-free), but it fails to reproduce Heaps’ law and the Burstiness property.

In sum, we have shown numerically that a random walk on an SSF network as text generation exhibits the three major statistical laws that has been robustly measured in language streams: Zipf’s law, Heaps’ law and Burstiness. On the other hand, if the specific properties of the SSF network, namely, its long average path length and large average clustering coefficient, are randomized out by an edge rewiring procedure, two out of the three laws are not reproduced any more: the Heaps exponent becomes unrealistically large and the walk becomes non-bursty.

6 Discussion

In this work, we formulated a text generation model that reproduces the three major statistical laws observed ubiquiously in human languages: Zipf’s law, Heaps’ law, and Burstiness. Apart from the fact that according to my best knowledge, this is the first such model, it has the value being very simple and thus might contribute to the understanding of the processes going on in the mind during text production.

Besides the theoretical question of how concepts self-organize in our mind, entangling the global characteristics of the semantic space is of practical importance as this is an inevitable step towards analyzing texts as
trajectories in such a space. This method might have multiple promising applications from author or style identification, personal disorder diagnosis, to the quantitative investigation of how higher-order cognitive processes such as consciousness evolved throughout the history of (literate) humankind [57, 58, 59].

A major step towards a more realistic model would be to incorporate knowledge about the semantic space itself, e.g., the hierarchical modular structure of it [51], along with generating texts that reproduce more intricate statistical properties of real texts, described e.g., in terms of correlations or information theoretic quantities [61, 32, 62, 36].

Ultimately, the main goal would be to integrate the knowledge we pile up in different disciplines. A first step towards this goal is to identify a (more) common conceptual ground of linguistics, neurolinguistics, psycholinguistics, computational linguistics and the effort made by people having background in complex systems/statistical physics, to be able to at least relate or compare the results obtained from as different methodologies as human experiments, brain imaging, data analysis, and mathematical modeling.
References


