Modeling Statistical Properties of Written Text

M. Ángeles Serrano[1]*, **Alessandro Flammini**[2], **Filippo Menczer**[2,3]

[1] Departament de Química Física, Universitat de Barcelona, Barcelona, Spain, [2] School of Informatics, Indiana University, Bloomington, Indiana, United States of America, [3] Complex Networks Lagrange Lab, ISI Foundation, Torino, Italy

Abstract

Written text is one of the fundamental manifestations of human language, and the study of its universal regularities can give clues about how our brains process information and how we, as a society, organize and share it. Among these regularities, only Zipf's law has been explored in depth. Other basic properties, such as the existence of bursts of rare words in specific documents, have only been studied independently of each other and mainly by descriptive models. As a consequence, there is a lack of understanding of linguistic processes as complex emergent phenomena. Beyond Zipf's law for word frequencies, here we focus on burstiness, Heaps' law describing the sublinear growth of vocabulary size with the length of a document, and the topicality of document collections, which encode correlations within and across documents absent in random null models. We introduce and validate a generative model that explains the simultaneous emergence of all these patterns from simple rules. As a result, we find a connection between the bursty nature of rare words and the topical organization of texts and identify dynamic word ranking and memory across documents as key mechanisms explaining the non trivial organization of written text. Our research can have broad implications and practical applications in computer science, cognitive science and linguistics.

Citation: Serrano MÁ, Flammini A, Menczer F (2009) Modeling Statistical Properties of Written Text. PLoS ONE 4(4): e5372. doi:10.1371/journal.pone.0005372

Editor: Enrico Scalas, University of East Piedmont, Italy

Received February 5, 2009; **Accepted** March 31, 2009; **Published** April 29, 2009

Funding: This work was supported in part by a Lagrange Senior Fellowship of the CRT Foundation to F.M., by the Institute for Scientific Interchange Foundation, the EPFL Laboratory of Statistical Biophysics, and the Indiana University School of Informatics; M. A. S. acknowledges support by DGES grant No. FIS2007-66485-C02-01 and the Ramón y Cajal program of the Spanish Ministry of Science. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: marian.serrano@ub.edu

These authors contributed equally to this work.

Introduction

The understanding of human language [1] requires an interdisciplinary approach and has broad conceptual and practical implications over a broad range of fields. Computer science, where natural language processing [2–4] seeks to model language computationally, and cognitive science, that tries to understand our intelligence with linguistics as one of its key contributing disciplines [5], are among the fields more directly involved.

Written text is a fundamental manifestation of human language. Nowadays, electronic and information technology media offer the opportunity to easily record and access huge amounts of documents that can be analyzed in quest for some of the signatures of human communication. As a first step, statistical patterns in written text can be detected as a trace of the mental processes we use in communication. It has been realized that various universal regularities characterize text from different domains and languages. The best-known is Zipf's law on the distribution of word frequencies [6–8], according to which the frequency of terms in a collection decreases inversely to the rank of the terms. Zipf's law has been found to apply to collections of written documents in virtually all languages. Other notable universal regularities of text are Heaps' law [9,10], according to which vocabulary size grows slowly with document size, i.e. as a sublinear function of the number of words; and the bursty nature of words [11–13], making a word more likely to reappear in a document if it has already appeared, compared to its overall frequency across the collection.

The structure of written text is key to a broad range of critical applications such as Web search [14,15] (and the booming business of online advertising), literature mining [16,17], topic detection [18,19], and security [20–22]. Thus, it is not surprising that researchers in linguistics, information and cognitive science, machine learning, and complex systems are coming together to model how universal text properties emerge. Different models have been proposed that are able to predict each of the universal properties outlined above. However, no single model of text generation explains all of them together. Furthermore, no model has been used to interpret or predict the empirical distributions of text similarity between documents in a collection [23,24].

In this paper, we present a model that generates collections of documents consistently with all of the above statistical features of textual corpora, and validate it against large and diverse Web datasets. We go beyond the global level of Zipf's law, which we take for granted, and focus on general correlation signatures within and across documents. These correlation patterns, manifesting themselves as burstiness and similarity, are destroyed when the words in a collection are reshuffled, even while the global word frequencies are preserved. Therefore the correlations are not simply explained by Zipf's law, and are directly related to the global organization and topicality of the corpora. The aim of our model is not to reproduce the microscopic patterns of occurrence of individual words, but rather to provide a stylized generative mechanism to interpret their emergence in statistical terms. Consequently, our main assumption is a global distribution of word probabilities; we do not need to fit a large number of parameters to the data, in contrast to parametric models proposed to describe the bursty nature or topicality of text [25–27]. In our model, each document is derived by a local ranking of dynamically reordered words, and different documents are related by sharing subsets of these rankings that represent emerging topics. Our analysis shows that the statistical structure of text collections, including their level of topicality, can be derived from such a simple ranking mechanism. Ranking is an alternative to preferential attachment for explaining scale invariance [28] and has been used to explain the emergent topology of complex information, technological, and social networks [29]. The present results suggest that it may also shed light on cognitive processes such as text generation and the collective mechanisms we use to organize and store information.

Results and Discussion

Empirical Observations

We have selected three very diverse public datasets, from topically focused to broad coverage, to illustrate the statistical regularities of text and validate our model. The first corpus is the Industry Sector database (IS), a collection of corporate Web pages organized into categories or sectors. The second dataset is a sample of the Open Directory (ODP), a collection of Web pages classified into a large hierarchical taxonomy by volunteer editors. The third corpus is a random sample of topic pages from the English Wikipedia (Wiki), a popular collaborative encyclopaedia that also is comprised of millions of online entries. (See Materials and Methods for details.)

We measured the statistical regularities mentioned above in our datasets and the empirical results are shown in Fig. 1. We stress that although our work focuses on collections of documents written in English, the regularities discussed here are universal and apply to documents written in virtually all languages. The distributions of document length for all three collections can be approximated by a lognormal with different first and second moment parameters [30] (see Web Datasets under Materials and Methods). Another universal property of written text is Zipf's law [6–8,31], according to which the global frequency f_g of terms in a collection decreases roughly inversely to their rank r: $f_g \sim 1/r$ or, in other words, the distribution of the frequency f_g is well approximated by a power law $P(f_g) \sim f_g^{-\alpha}$ with exponent around $\alpha \approx 2$. Zipf's law also applies to the datasets used here, as supported by a Kolmogorov-Smirnov goodness-of-fit test [32] (see Fig. 1a and its caption for details). Heaps' law [9,10] describes the sublinear growth of vocabulary size (number of unique words) w as a function of the size of a document (number of words) n (Fig. 1b). This feature has also been observed in different languages, and the behavior has been interpreted as a power law $w(n) \sim n^{\beta}$ with $\beta < 1$, although the exponent β between 0.4 and 0.6 is language dependent [33].

Burstiness is the tendency of some words to occur clustered together in individual documents, so that a term is more likely to reappear in a document where it has appeared before [11–13]. This property is more evident among rare words, which are more likely to be topical. Following Elkan [27], the bursty nature of words can be illustrated by dividing words into classes according to their global frequency (e.g., common vs. rare). For words in each class, we plot in Fig. 1c the probability $P(f_d)$ that these words occur with frequency f_d in single documents, averaged over all documents in the collection. We compare the distribution $P(f_d)$ of common and rare terms with those predicted by the null independence hypothesis. This reference model generates documents whose length is drawn from the lognormal distribution fitted to the empirical data (see Materials and Methods) by drawing words independently at random from the global Zipf frequency distribution (Fig. 1a). As compared to the reference of such a *Zipf model*, rare terms are much more likely to cluster in specific documents and not to appear evenly distributed in the collection, so that ordering principles beyond those responsible for Zipf's law have to be at play.

Another signature of text collections, which is more telling about topicality, is the distribution of lexical similarity across pairs of documents. In information retrieval and text mining, documents are typically represented as term vectors [15,34]. Each element of a vector represents the weight of the corresponding term in the document. There are various vector representations according to different weighting schemes. Here, we focus on the simplest scheme, in which a weight is simply the frequency of the term in the document. The similarity between two documents is given by the cosine between the two vectors: $s(p,q) = \sum_t w_{tp} w_{tq} \Big/ \sqrt{\sum_t w_{tp}^2 \cdot \sum_t w_{tq}^2}$ where w_{tp} is the weight of term t in document p. It has been observed that for documents sampled from the ODP, the distribution of cosine similarity based on term frequency vectors is concentrated around zero and decays in a roughly exponential fashion for $s > 0$ [23,24]. Figure 1d shows that different collections yield different similarity profiles, however they all tend to be more skewed toward small similarity values than predicted by the Zipf model.

Modeling how these properties emerge from simple rules is central to an understanding of human language and related cognitive processes. Our understanding, however, is far from definitive. First, the empirical observations are open to different interpretations. As an example, much has been written about the debate between Simon and Mandelbrot around different interpretations of Zipf's law (see www.nslij-genetics.org/wli/zipf for a historical review of the debate). Second, and perhaps more importantly, no single model of text generation explains all of the above observations simultaneously. Third, models at hand are usually based on descriptive methods that cannot explain linguistic processes as emergent phenomena.

In the remainder of this paper, we focus on burstiness and similarity distributions. Regarding similarity, little attention has been given to its empirical distribution and, to the best of our knowledge, no model has been put forth to explain its profile. Regarding text burstiness, on the other hand, several models have been proposed including the two-Poisson model [11], the Poisson zero-inflated mixture model [35], Katz' k-mixture model [12], and a gap-based variation of Bayes model [36]. Another line of generative models extends the simple multinomial family with increasingly complex views of topics. Examples include probabilistic latent semantic indexing [37], latent Dirichlet allocation (LDA) [25], and Pachinko allocation [38]. These models assume a set of topics, each typically described by a multinomial distribution over words. Each document is then generated from some mixture of these topics. In LDA, for example, the parameters of the mixture are drawn from a Dirichlet distribution, independently for each document. Repeatedly drawing a topic from the mixture first, and then drawing a term from the corresponding word distribution generate the words' sequence in a document. A variety of techniques have been developed to estimate from data the parameters that characterize the many distributions involved in the generative process [21,26,39]. Although the above models were mainly developed for subject classification, they have also been used to investigate burstiness since bursty words can characterize the topic of a document [27,40].

The very large numbers of free parameters associated with individual terms, topics, and/or their mixtures grant the above models great descriptive power. However, their cognitive plausi-

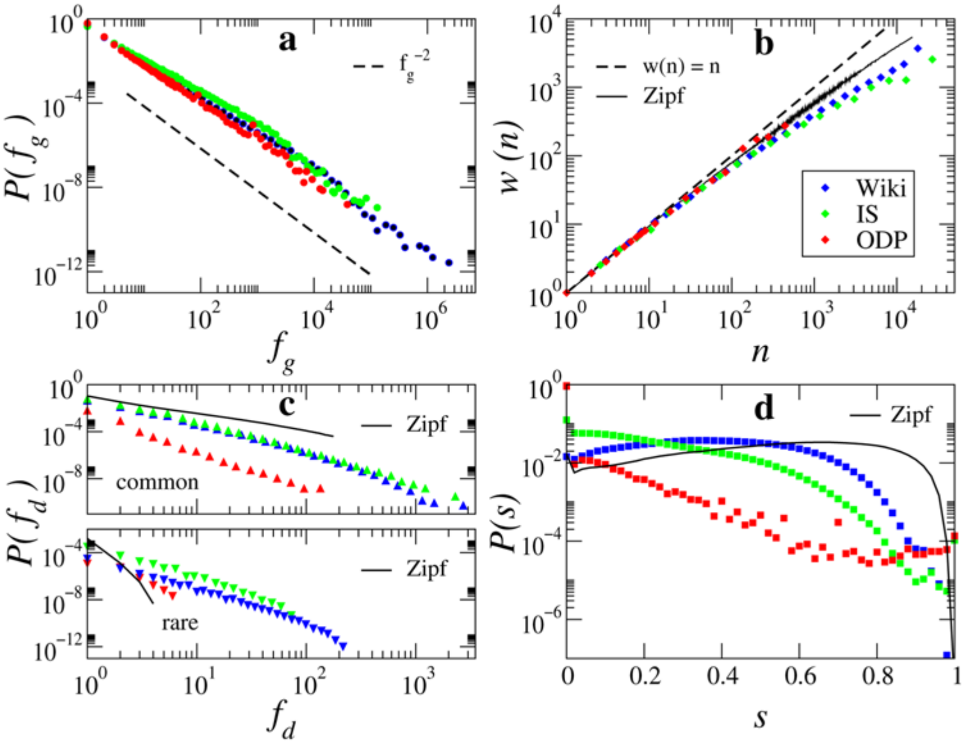


Figure 1. Regularities in textual data as observed in our three empirical datasets. (a) Zipf's Law: word counts are globally distributed according to a power law $P(f_g) \sim f_g^{-\alpha}$. The maximum likelihood estimates of the characteristic exponent α are 1.83 for Wikipedia, 1.78 for IS, and 1.88 for ODP. A Kolmogorov-Smirnov goodness-of-fit test [32] comparing the original data against 2500 synthetic datasets gives p-values for the maximum likelihood fits of 1 for Wikipedia and IS and 0.56 for ODP, all well above a conservative threshold of 0.1. This ensures that the power-law distribution is a plausible and indeed very good model candidate for the real distributions. (b) Heaps' law: as the number of words *n* in a document grows, the average vocabulary size (i.e. the number of distinct words) *w(n)* grows sublinearly with *n*. (c) Burstiness: fraction of documents *P(f_d)* containing f_d occurrences of common or rare terms. For each dataset, we label as "common" those terms that account for 71% of total word occurrences in the collection, while rare terms account for 8%. (d) Similarity: distribution of cosine similarity *s* across all pairs of documents, each represented as a term frequency vector. Also shown are *w(n)*, the distributions of f_d, and the distribution of *s* according to the Zipf null model (see text) corresponding to the IS dataset.

doi:10.1371/journal.pone.0005372.g001

bility is problematic. Our aim here is instead to produce a simpler, more plausible mechanism compatible with the high-level statistical regularities associated with *both* burstiness and similarity distributions, without regard for explicit topic modeling.

Model and Validation

Two basic mechanisms, reordering and memory, can explain burstiness and similarity consistently with Zipf's law. We show this by proposing a generative model that incorporates these processes to produce collections of documents characterized by the observed statistical regularities. Each document is derived by a local ranking of words that reorganizes according to the changing word frequencies as the document grows, and different documents are related by sharing subsets of these rankings that represent emerging topics. With just the main assumptions of the global distribution of word probabilities and document sizes and a single tunable parameter measuring the topicality of the collection, we are able to generate synthetic corpora that re-create faithfully the features of our Web datasets. Next, we describe two variations of the model, one without memory and the second with a memory mechanism that captures topicality.

Dynamic Ranking by Frequency. In our model, *D* documents are generated drawing word instances repeatedly with replacement from a vocabulary of *V* words. The document lengths in number of words are drawn from a lognormal distribution. The parameters *D*, *V*, and the maximum likelihood estimates of the lognormal mean and variance are derived empirically from each dataset (see Table 1 in Materials and Methods). We further assume that at any step of the generation process, word probabilities follow a Zipf distribution $P[r(t)] \propto r(t)^{-1}$ where *r(t)* is the rank of term *t*. (We also tested the model using the empirical distributions of document length and word frequency for each collection and the results are essentially the same.) However, rather than keeping a fixed ranking, we imagine that words are sorted dynamically during the generation of each document according to the number of times they have already occurred. Words and ranks are thus decoupled: at different times, a word can have different ranks and a position in the ranking can be occupied by different words. The idea is that as the topicality of a document emerges through its content, topical words will be more likely to reoccur within the same document. This idea is incorporated into the model as a frequency bias favoring words that occur early in the document.

In the first version of the model, each document is produced independently of each other. Before each new document is generated, words are sorted according to an initial global ranking, which remains fixed for all documents. This ranking r_0 is also used to break ties during the generation of documents, among words with the same occurrence counts. The algorithm corresponding to this dynamic ranking model is illustrated in Fig. 2 and detailed in Materials and Methods.

Table 1. Statistics for the different document collections.

Dataset	V	D	$<w>$	$<n>$	$\sigma^2(n)$	μ	σ^2
Wiki	588639	100000 (0)	160.44	373.86	457083	5.13	1.57
IS	47979	9556 (15)	124.26	313.46	566409	4.81	2.10
ODP	105692	107360 (32558)	8.88	10.34	345	1.93	1.39

V stands for vocabulary size, *D* for the number of documents containing at least one word (in parenthesis the number of empty documents in the collection), $<w>$ for the average size of documents in number of unique words, and $<n>$ and $\sigma^2(n)$ for the average and variance of document size in number of words. For each collection, the distribution of document size is approximately fitted by a lognormal with parameters μ and σ^2 (values shown are maximum likelihood estimates).

doi:10.1371/journal.pone.0005372.t001

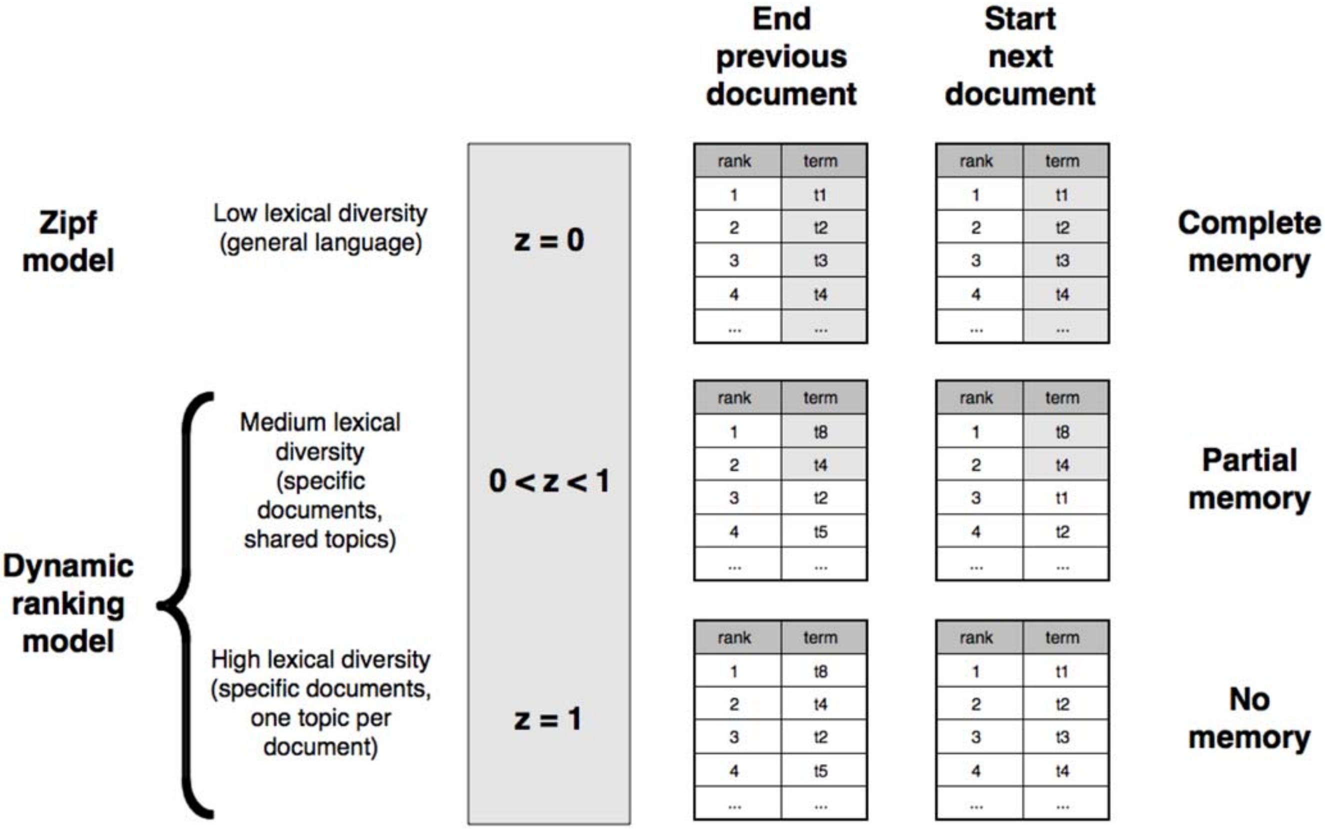
When a sufficiently large number of documents is generated, the measured frequency of a word *t* over the entire corpus approaches the Zipf distribution $P(t) \sim [r_0(t)]^{-1}$, ensuring the self consistency of the model. We numerically simulated the dynamic ranking model for each dataset. A direct comparison with the empirical burstiness curves shown in Fig. 1c can be found in Fig. 3a. The excellent agreement suggests that the dynamic ranking process is sufficient for producing the right amount of correlations inside documents needed to realistically account for the burstiness effect.

Heaps' law can be derived analytically from our model. The probability *P(w,n)* to find *w* distinct words in a document of size *n* satisfies the following discrete master equation:

$$P(w+1,n+1) = P(w+1,n)F(w+1) + P(w,n)[1-F(w)], \quad (1)$$

where $F(w) = \sum_{r=1}^{w} P(r)$, and *P(r)* is the Zipf probability associated with rank *r.*

There are two contributions to the probability to have *w+1* distinct words in a document of length *n+1*, represented by the two terms in the *r.h.s* of Eq. (1) above. Before adding the *(n+1)*th the document may already contain *w+1* distinct words, and such number remains the same if an already observed word is added. Since the *w+1* words that have been already observed occupy the first *w+1* position in the rank, one of them is observed with probability $F(w+1) = \sum_{r=1}^{w+1} P(r)$, therefore the first contribution ensues. The other possibility is that the document contains only *w* distinct words and that a previously unobserved word is added. For the same reasons presented above this happen with probability $\sum_{r=w+1}^{V} P(r) = 1 - F(w)$, and this accounts for the second contribution. To make progresses it is useful to write an equation for the expected number of distinct words. This can be done by multiplying both sides of Eq. (1) by *(w+1)* and summing over *w.* This leads to:


End previous document
Start next document
Zipf model
Low lexical diversity (general language)
z = 0
Complete memory
Dynamic ranking model
Medium lexical diversity (specific documents, shared topics)
0 < z < 1
Partial memory
High lexical diversity (specific documents, one topic per document)
z = 1
No memory

rank	term
1	t1
2	t2
3	t3
4	t4
...	...

rank	term
1	t1
2	t2
3	t3
4	t4
...	...

rank	term
1	t8
2	t4
3	t2
4	t5
...	...

rank	term
1	t8
2	t4
3	t1
4	t2
...	...

rank	term
1	t8
2	t4
3	t2
4	t5
...	...

rank	term
1	t1
2	t2
3	t3
4	t4
...	...

Figure 2. Illustration of the dynamic ranking model. The parameter *z* regulates the lexical diversity, or topicality of the collection. The extreme case *z = 0* is equivalent to the null Zipf model, where all documents are generated using the global word rank distribution. The opposite case *z = 1* is the first version of the dynamic ranking model, with no memory, in which each new document starts from the global word ranking r_0. Intermediate values of *z* represent the more general version of the dynamic ranking model, where correlations across documents are created by a partial memory of word ranks. A more detailed algorithmic description of the model can be found in Materials and Methods.

doi:10.1371/journal.pone.0005372.g002

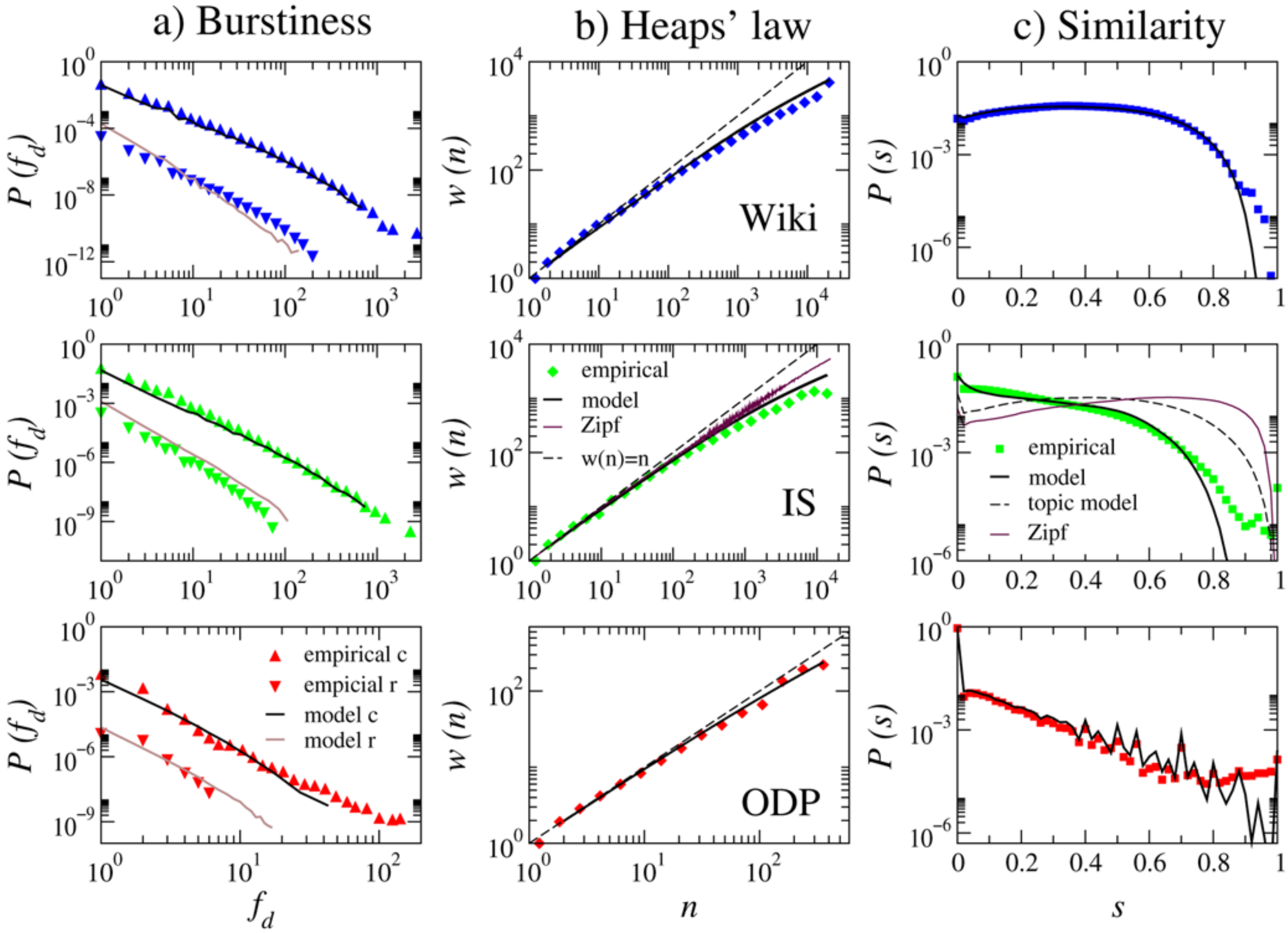


Figure 3. Model vs. empirical observations. The coefficient of determination R^2 is computed in all cases as an estimator of the goodness of fit between the simulation and the empirical measurements. (a) Comparison of burstiness curves produced by the dynamic ranking model with those from the empirical datasets. Common and rare words are defined in Fig. 1c. For all the comparisons, R^2 is larger than 0.99. (b) Comparison of Heaps' law curves produced by the dynamic ranking model with those from the empirical datasets. Simulations of the model provide the same predictions as numerical integration of the analytically derived equation using the empirical rank distributions (see text). For the IS dataset we also plot the result of the Zipf null model, which produces a sublinear *w(n)*, although less pronounced than our model. The ODP collection has short documents on average (cf. Table 1 in Materials and Methods), so Heaps' law is barely observable. For all the comparisons, R^2 is larger than 0.99. (c) Comparison between similarity distributions produced by the dynamic ranking model with memory, and those from the empirical datasets also shown in Fig. 1d. The parameter *z* controlling the topical memory is fitted to the data. The peak at $s = 0$ suggests that the most common case is always that of documents sharing very few or no common terms. The discordance for high similarity values is due to corpus artifacts such as mirrored pages, templates, and very short (one word) documents. The fluctuations in the curves for the ODP dataset are due to binning artifacts for short pages. Also shown is the prediction of the topic model for the IS dataset (see text). Finally, the R^2 statistic has a value 0.98 for Wikipedia, 0.94 for IS, and larger than 0.99 for ODP.
doi:10.1371/journal.pone.0005372.g003

$$\begin{aligned}&\sum P(w+1,n+1)(w+1) = \\ &\left\{\sum P(w+1,n)F(w+1)(w+1) - P(w,n)F(w)w\right\} \qquad (2)\\ &+\sum P(w,n)w + \sum P(w,n) - \sum P(w,n)F(w).\end{aligned}$$

To simplify notations we will use $E_n[f(w)] = \sum P(w,n)f(w)$ to indicate the expected value of a function *f(w)* at step n. Using the fact that $\sum P(w,n) = 1$, and that the term in curly brackets on the *r.h.s.* of Eq. (2) is null, one finds:

$$E_{n+1}[w] = E_n[w] + 1 - E_n[F(w)]. \qquad (3)$$

To further simplify notations, we pose $w(n) = E_n[w]$. To close Eq. (3) in terms of *w(n)* we neglect fluctuations and assume that the probability to observe *w* distinct words in a document of size *n* is strongly peaked around *w(n)*. Eq. (3) can then be rewritten as:

$$w(n+1) - w(n) = 1 - F(w(n)). \qquad (4)$$

It is convenient to take the continuous limit, replacing finite differences by derivative, and sums by integrals. One finally obtains:

$$\frac{dw(n)}{dn} = \int_w^V P(r)dr. \qquad (5)$$

Eq. (5) can be integrated numerically using the actual *P(r)* from the data. Alternatively, Eq. (5) can be solved analytically for special cases. Assuming a Zipf's law with a tail of the form $P(r) \sim r^{-\gamma}$ where $\gamma > 1$, the solution is $w(n) \sim n^{1/\gamma}$ and we recover Heaps' sublinear growth with $\beta \approx 1/\gamma$ for large *n*. According to the Yule-Simon model [41], which interprets Zipf's law through a preferential attachment process, the rank distribution should have a tail with exponent $\gamma > 1$. This is confirmed empirically in many English collections; for example our ODP and Wikipedia datasets yield Zipfian tails with γ between 3/2 and 2. Our model predicts that in these cases Heaps' growth should be well approximated by a power law with exponent β between 1/2 and 2/3, closely matching those reported for the English language [33]. Simulations using the empirically derived *P(r)* for each dataset display growth trends for large *n* that are in good agreement with the empirical behavior (Fig. 3b).

Topicality and Similarity. The agreement between empirical data and simulations of the model with respect to the similarity distributions gets worse for those datasets that are more topically focused. A new mechanism is needed to account for topical correlations between documents.

The model in the previous section generates collections of independent text documents, with specific but uncorrelated topics captured by the bursty terms. For each new document, the rank of each word *t* is initialized to its original value $r_0(t)$ so that each document has no bias toward any particular topic. The resulting synthetic corpora display broad coverage. However, real corpora may cover more or less specific topics. The stronger the semantic relationship between documents, the higher the likelihood they share common words. Such collection topicality needs to be taken into account to accurately reproduce the distribution of text similarity between documents.

To incorporate topical correlations into our model, we introduce a memory effect connecting word frequencies across different documents. Generative models with memory have already been proposed to explain Heaps' law [10]. In our algorithm (see Fig. 2 and Materials and Methods) we replace the initialization step so that a portion of the initial ranking of the terms in each document is inherited from the previously generated document. In particular, the counts of the r^* top-ranked words are preserved while all the others are reset to zero. The rank r^* is drawn from an exponential distribution $P(r^*) = z(1-z)^{r^*}$, where z is a probability parameter that models the lexical diversity of the collection and r^* has expected value $1/z-1$, which can be interpreted as the collection's shared topicality.

This variation of the model does not interfere with the reranking mechanism described in the previous section, so that the burstiness effect is preserved. The idea is to interpolate between two extreme cases. The case $z = 0$, in which counts are never reset, converges to the null Zipf model. All documents share the same general terms, modeling a collection of unspecific documents. Here we expect a high similarity in spite of the independence among documents, because the words in all documents are drawn from the identical Zipf distribution. The other extreme case, $z = 1$, reduces to the original model, where all the counts are always initialized to zero before starting a document. In this case, the bursty words are numerous but not the same across different documents, modeling a situation in which each document is very specific but there is no shared topic across documents. Intermediate cases $0 < z < 1$ allow us to model correlations across documents not only due to the common general terms, but also to topical (bursty) terms.

We simulated the dynamic ranking model with memory under the same conditions corresponding to our datasets, but additionally fitting the parameter z to match the empirical similarity distributions. The comparisons are shown in Fig. 3c. The similarity distribution for the ODP is best reproduced for $z = 1$, in accordance to the fact that this collection is overwhelmingly composed of very specific documents spanning all topics. In such a situation, the original model accurately reproduces the high diversity among document topics and there is no need for memory. In contrast, Wikipedia topic pages use a homogenous vocabulary due to their strict encyclopaedic style and the social consensus mechanism driving the generation of content. This is reflected in the value $z = 0.005$, corresponding to an average of $1/z = 200$ common words whose frequencies are correlated across successive pairs of documents. The industry sector dataset provides us with an intermediate case in which pages deal with more focused, but semantically related topics. The best fit of the similarity distribution is obtained for $z = 0.1$.

With the fitted values for the shared topicality parameter z, the agreement between model and empirical similarity data in Fig. 3c is excellent over a broad range of similarity values. To better illustrate the significance of this result, let us compare it with the prediction of a simple topic model. For this purpose we assume a priori knowledge of the set of topics to be used for generating the documents. The IS dataset lends itself to this analysis because the pages are classified into twelve disjoint industry sectors, which can naturally be interpreted as unmixed topics. For each topic *c*, we measured the frequency of each term *t* and used it as a probability $p(t|c)$ in a multinomial distribution. We generated the documents for each topic using the actual empirical values for the number of documents in the topic and the number of words in each document. As shown in Fig. 3c, the resulting similarity distribution is better than that of the Zipf model (where we assume a single global distribution), however the prediction is not nearly as good as that of our model.

Our model only requires a single free parameter z plus the global (Zipfian) distribution of word probabilities, which determines the initial ranking. Conversely, for the topic model we must have —or fit— the frequency distribution $p(t|c)$ over all terms for each topic, which implies an extraordinary increase in the number of free parameters since, apart from potential differences in the functional forms, each distribution would rank the terms in a different order.

Aside from complexity issues, the ability to recover similarities suggests that the dynamic ranking model, though not as well informed as the topic model on the distributions of the specific topics, better captures word correlations. Topics emerge as a consequence of the correlations between bursty terms across documents as determined by z, but it is not necessary to predefine the number of topics or their distributions.

Conclusion

Our results show that key regularities of written text beyond Zipf's law, namely burstiness, topicality and their interrelation, can be accounted for on the basis of two simple mechanisms, namely frequency ranking with dynamic reordering and memory across documents, and can be modeled with an essentially parameter-free algorithm. The rank based approach is in line with other recent models in which ranking has been used to explain the emergent topology of complex information, technological, and social networks [29]. It is not the first time that a generative model for text has walked parallel paths with models of network growth. A remarkable example is the Yule-Simon model for text generation [41] that was later rediscovered in the context of citation analysis [42], and has recently found broad popularity in the complex networks literature [43].

Our approach applies to datasets where the temporal sequence of documents is not important, but burstiness has also been studied in contexts where time is a critical component [13,44], and even in human languages evolution [45]. Further investigations in relation to topicality could attempt to explicitly demonstrate the role of the topicality correlation parameter by looking at the hierarchical structure of content classifications. Subsets of increasingly specific topics of the whole collection could be extracted to study how the parameter z changes and how it is related to external categorizations. The proposed model can also be used to study the co-evolution of content and citation structure in the scientific literature, social media such as the Wikipedia, and the Web at large [10,23,46,47].

From a broader perspective, it seems natural that models of text generation should be based on similar cognitive mechanisms as models of human text processing since text production is a translation of semantic concepts in the brain into external lexical representations. Indeed, our model's connection between frequency ranking and burstiness of words provides a way to relate two key mechanisms adopted in modeling how humans process the lexicon: rank frequency [48] and context diversity [49]. The latter, measured by the number of documents that contain a word, is related to burstiness since, given a term's overall collection frequency, higher burstiness implies lower context diversity. While tracking frequencies is a significant cognitive burden, our model suggests that simply recognizing that a term occurs more often than another in the first few lines of a document would suffice for detecting bursty words from their ranking and consequently the topic of the text.

In summary, a picture of how language structure and topicality emerge in written text as complex phenomena can shed light into the collective cognitive processes we use to organize and store information, and find broad practical applications, for instance, in topic detection, literature analysis, and Web mining.

Materials and Methods

Web Datasets

We use three different datasets. The Industry Sector database is a collection of almost 10,000 corporate Web pages organized into 12 categories or sectors. The second dataset is a sample of the Open Directory Project, a collection of Web pages classified into a large hierarchical taxonomy by volunteer editors (*dmoz.org*). While the full ODP includes millions of pages, our collection comprises of approximately 150,000 pages, sampled uniformly from all top-level categories. The third corpus is a random sample of 100,000 topic pages from the English Wikipedia, a popular collaborative encyclopedia that also is comprised of millions of online entries (*en.wikipedia.org*).

These English text collections are derived from public data and are publicly available (the IS dataset is available at www.cs.umass.edu/~mccallum/code-data.html, the ODP and Wikipedia corpora are available upon request); have been used in several previous studies, allowing a cross check of our results; and are large enough for our purposes without being computationally unmanageable. The datasets are however very diverse in a number of ways. The IS corpus is relatively small and topically focused, while ODP and Wikipedia are larger and have broader coverage, as reflected in their vocabulary sizes. IS documents represent corporate content, while many Web pages in the ODP collection are individually authored. Wikipedia topics are collaboratively edited and thus represent the consensus of a community.

The distributions of document length for all three collections can be approximated by lognormals shown in Fig. 4, with different first and second moment parameters. The values shown in Table 1 summarize the main statistical features of the three collections (lognormal parameters are the maximum likelihood estimates).

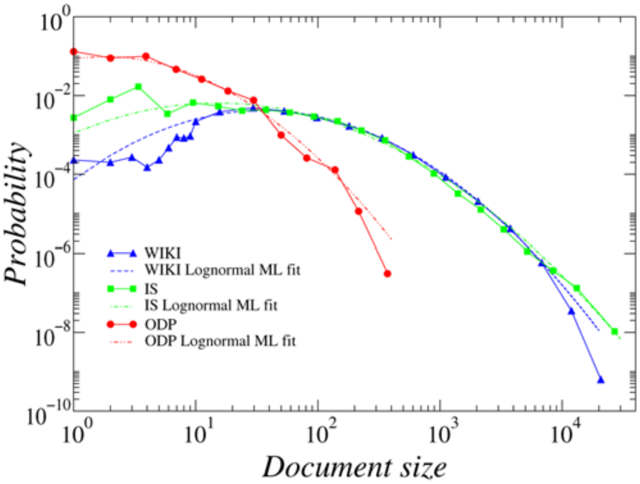


Figure 4. Distributions of documents' length for all three collections. Each distribution can be approximated by a lognormal, with different first and second moment parameters obtained by maximum likelihood (ML) (see Table 1).
doi:10.1371/journal.pone.0005372.g004

Before our analysis, all documents in each collection have been parsed to extract the text (removing HTML markup) and syntactic variations of words have been conflated using standard stemming techniques [50].

Algorithm

The following algorithm implements the dynamic ranking model:

```
Vocabulary: $t \in \{1,\ldots,V\}$
Initial ranking: $\forall t : r_0(t) = t$
Repeat until $D$ documents are generated:
  Initialize term counts to $\forall t : c(t) = 0$ (*)
  Draw $L$ from lognormal $(\mu,\sigma^2)$
  Repeat until $L$ terms are generated:
    Sort terms to obtain new rank $r(t)$ according to $c(t)$
      (break ties by $r_0$)
    Select term $t$ with probability $P(t) \propto r(t)^{-1}$
    Add t to current document
    $c(t) \leftarrow c(t) + 1$
  End of document
End of collection
```

The document initialization step (line marked with an asterisk in above pseudocode) is altered in the more general, memory version of the model (see main text). In particular we set to zero the counts $c(t)$ not of all terms, but only of terms t such that $r(t) \geq r^*$. The rank r^* is drawn from an exponential distribution $P(r^*) = z(1-z)^{r^*}$ with expected value $1/z$-1, as discussed in the main text. In simpler terms, the counts of the r^* top-ranked words are preserved while all the others are reset to zero.

Algorithmically, terms are sorted by counts so that the top-ranked term t ($r(t) = 1$) has the highest $c(t)$. We iterate over the ranks r, flipping a biased coin for each term. As long as the coin returns false (probability 1-z), we preserve $c(t(r))$. As soon as the coin returns true (probability z), say for the term $t(r^*)$, we reset all the counts for this and the following terms: $\forall r > r^*$ $c(t(r)) = 0$.

The special case $z = 1$ reverts to the original, memory-less model; all counts are reset to zero and each document restarts from the global Zipfian ranking r_0. The special case $z = 0$ is equivalent to the Zipf null model as the term counts are never reset and thus rapidly converge to the global Zipfian frequency distribution.

Acknowledgments

We thank Charles Elkan, Rich Shiffrin, Michael Jones, Alessandro Vespignani, Dragomir Radev, Vittorio Loreto, Ciro Cattuto, and Marián Boguñá for useful discussions. Jacob Ratkiewicz provided assistance in gathering and analyzing the Wikipedia dataset.

Author Contributions

Designed research, performed research, analyzed data, and wrote the paper: MAS AF FM.

References

1. Hauser MD, Chomsky N, Fitch WT (2002) The faculty of language: What is it, who has it, and how did it evolve? Science 298: 1569–1579.
2. Joshi AK (1991) Natural language processing. Science 253: 1242–1249.
3. Manning C, Schütze H (1999) Foundations of statistical natural language processing. Cambridge: MIT press.
4. Nowak MA, Komarova NL, Niyogi P (2002) Computational and evolutionary aspects of Language. Nature 417: 611–617.
5. Chomsky N (2006) Language and Mind. Cambridge: Cambridge University Press, 3rd ed.
6. Zipf GK (1949) Human Behaviour and the Principle of Least Effort. Cambridge: Addison-Wesley.
7. Baayen RH (2001) Word Frequency Distributions. Dordrecht, Boston, London: Kluwer Academic Publishers.
8. Saichev A, Malevergne Y, Sornette D (2008) Theory of Zipf's Law and of General Power Law Distributions with Gibrat's Law of Proportional Growth, Lecture Notes in Economics and Mathematical Systems. Berlin, Heidelberg, New York: Springer.
9. Heaps HS (1978) Information Retrieval: Computational and Theoretical Aspects. Orlando: Academic Press.
10. Cattuto C, Loreto V, Pietronero L (2007) Semiotic dynamics and collaborative tagging. Proc Natl Acad Sci U S A 104: 1461–1464.
11. Church KW, Gale WA (1995) Poisson mixtures. Natural Language Engineering. pp 163–190.
12. Katz SM (1996) Distribution of content words and phrases in text and language modelling. Natural Language Engineering 2: 15–59.
13. Kleinberg J (2002) Bursty and hierarchical structure in streams. In: Proc 8th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining.
14. Chakrabarti S (2003) Mining the Web: Discovering knowledge from hypertext data. San Francisco: Morgan Kaufmann.
15. Liu B (2007) Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Berlin, Heidelberg, New York: Springer.
16. Ananiadou S, Mcnaught J, eds (2005) Text Mining for Biology And Biomedicine. Norwood: Artech House Publishers.
17. Feldman R, Sanger J (2006) The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge: Cambridge University Press.
18. Allan J, Papka R, Lavrenko V (1998) On-line new event detection and tracking. In: Proc SIGIR. pp 37–45.
19. Yang Y, Pierce T, Carbonell JG (1998) A study of retrospective and on-line event detection. In: Proc SIGIR. pp 28–36.
20. Chen H (2006) Intelligence and Security Informatics for International Security Information Sharing and Data Mining. Berlin, Heidelberg, New York: Springer.
21. Newman D, Chemudugunta C, Smyth P, Steyvers M (2006) Analyzing entities and topics in news articles using statistical topic models. Lecture Notes in Computer Science (Intelligence and Security Informatics) 3975: 93–104.
22. Pennebaker J, Chung C (2008) Computerized text analysis of al-qaeda transcripts. In: Krippendor K, Bock M, eds (2008) A content analysis reader. Thousand Oaks: Sage.
23. Menczer F (2004) The evolution of document networks. Proc Natl Acad Sci U S A 101: 5261–5265.
24. Menczer F (2002) Growing and navigating the small world Web by local content . Proc Natl Acad Sci U S A 99: 14014–14019.
25. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of Machine Learning Research 3: 993–1022.
26. Griffiths T, Steyvers (2004) Finding scientific topics. Proc Natl Acad Sci U S A. pp 5228–5235.
27. Elkan C (2006) Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution. In: Proc 23rd Intl Conf on Machine Learning (ICML).
28. Goh K-I, Kahng B, Kim D (2001) Universal behavior of load distribution in scale-free networks. Phys Rev Lett 87: 278701.
29. Fortunato S, Flammini A, Menczer F (2006) Scale-free network growth by ranking. Phys Rev Lett 96: 218701.
30. Dolby JL (1971) Programming languages in mechanized documentation. Journal of Documentation 27: 136–155.
31. Maslov VP, Maslova TV (2006) On zipf's law and rank distributions in linguistics and semiotics. Mathematical Notes 80: 679–691.
32. Clauset A, Shalizi R, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Review, to appear.
33. Baeza-Yates R, Ribeiro-Neto B (1999) Modern Information Retrieval. Wokingham: Addison-Wesley.
34. Salton G, McGill M (1983) An Introduction to Modern Information Retrieval. New York: McGraw-Hill.
35. Jansche M (2003) Parametric models of linguistic count data. In: Proc 41st Annual Meeting of the Association for Computational Linguistics. pp 288–295.
36. Sarkar A, Garthwaite P, De Roeck A (2005) A Bayesian mixture model for term reoccurrence and burstiness. In: Proc 9th Conference on Computational Natural Language Learning. pp 48–55.
37. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proc 22th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval. pp 50–57.
38. Li W, McCallum A (2006) Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proc 23rd Intl Conf on Machine Learning (ICML). pp 577–584.
39. Griffiths TL, Steyvers M, Blei DM, Tenenbaum JB (2005) Integrating topics and Syntax. In: Advances in Neural Information Processing Systems. Cambridge: MIT Press. vol. 17. pp 537–544.
40. Madsen R, Kauchak D, Elkan C (2005) Modeling word burstiness using the dirichlet distribution. In Proc 22nd Intl Conf on Machine Learning (ICML). pp 545–552.
41. Simon HA (1955) On a class of skew distribution functions. Biometrika 42: 425–440.
42. de Solla Price D (1976) A general theory of bibliometric and other cumulative advantage processes. J Amer Soc Inform Sci 27: 292–306.
43. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Reviews of Modern Physics 74: 47–97.
44. Barabási A-L (2005) The origin of bursts and heavy tails in human dynamics. Nature 435: 207–211.
45. Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M (2008) Languages evolve in punctuational bursts. Science 319: 588.
46. Kleinberg J (2004) Analysing the scientific literature in its online context. Nature Web Focus on Access to the Literature.
47. Alvarez-Lacalle E, Dorow B, Eckmann J-P, Moses E (2006) Hierarchical structures induce long-range dynamical correlations in written texts. Proc Natl Acad Sci U S A 103: 7956–7961.
48. Murray WS, Forster KI (2004) Seriel mechanisms in lexical access: The Rank Hypothesis Psychological Review 111: 721–756.
49. Adelman JS, Brown GDA, Quesada JF (2006) Contextual diversity, not word frequency, determines word naming and lexical decision times. Psychological Science 17: 814–823.
50. Porter M (1980) An algorithm for suffix stripping. Program 14: 130–137.