Log-log Convexity of Type-Token Growth in Zipf's Systems

[Accepted for publication in Physical Review Letters: 05 May 2015]

Francesc Font-Clos[1,2] and Álvaro Corral[1,2]

[1] *Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, E-08193 Barcelona, Spain.*
[2] *Departament de Matemàtiques, Facultat de Ciències, Universitat Autònoma de Barcelona, E-08193 Barcelona, Spain*

It is traditionally assumed that Zipf's law implies the power-law growth of the number of different elements with the total number of elements in a system - the so-called Heaps' law. We show that a careful definition of Zipf's law leads to the violation of Heaps' law in random systems, with growth curves that have a convex shape in log-log scale. These curves fulfil universal data collapses that only depend on the value of the Zipf's exponent. We observe that real books behave very much in the same way as random systems, despite the presence of burstiness in word occurrence. We advance an explanation for this unexpected correspondence.

PACS numbers:

A great number of systems in social science, economy, cognitive science, biology, and technology have been proposed to follow Zipf's law [1–6]. All of them have in common that they are composed by some "elementary" units, which we will call tokens, and that these tokens can be grouped into larger, concrete or abstract entities, called types. For instance, if the system is the population of a country, the tokens are its citizens, which can be grouped into different concrete types given by the cities where they live [7]. If the system is a text, each appearance of a word is a token, associated to the abstract type given by the word itself [8]. Zipf's law deals with how tokens are distributed into types, and can be formulated in two different ways, which are generally considered as equivalent [1, 3, 8, 9].

The first one is obtained when the number of tokens associated to each type are counted and the types are ranked in decreasing order of counts. We call this the rank-count representation. If a (decreasing) power law holds between the number of tokens of each type and the rank of the type, with an exponent close to one, this indicates the fulfilment of Zipf's law. An alternative version of the law arises when a second statistics is performed, considering the number of types that have the same number of counts; as the counts play the role of the random variable what one gets is the distribution of counts. If a (decreasing) power law is obtained, with an exponent around two, one gets a different formulation of Zipf's law, in principle.

However, in general, the fulfilment of Zipf's law has not been tested with rigorous statistical methods [2, 10]; rather, researchers have become satisfied with just qualitative resemblances between empirical data and power laws. In part, this can be justified by the difficulties of obtaining clear statistics from the rank-count representation, in particular for high ranks (that is, for rare types), and also by poor methods of estimation of probability distributions [2]. Despite the lack of unambiguous empirical support, from the theoretical point of view the search for explanations of Zipf's law has been extensive, but without a clearly accepted preferred mechanism [1, 11–14].

The presence of temporal order is an important feature in many Zipf-like systems, but this is not captured in Zipf's law. Indeed, this law only provides a static picture of the system (as the law is not altered under re-ordering of the data). In contrast, a suitable statistic that can unveil some of the dynamics is the type-token growth curve, which counts the total number of types, v, as a function of the total number of tokens, ℓ, as a system evolves, i.e., as citizens are born or a text is being read. Note that ℓ is a measure of system size (as system grows) and v is a measure of richness or diversity of types (with the symbol v borrowed from linguistics, where it stands for the size of the vocabulary).

It has long been assumed that Zipf's law implies also a power law for the type-token growth curve, i.e.,

$$v(\ell) \propto \ell^{\alpha}, \quad (1)$$

with exponent α smaller than one, and this is referred to as Heaps' law in general or Herdan's law in quantitative linguistics [15–17]. Indeed, Mandelbrot [18] and the authors of Ref. [19] obtain Heaps' law when tokens are drawn independently from a Zipf's system. Baeza-Yates and Navarro [16] argue that, if both Zipf's law and Heaps' law are fulfilled, their exponents are connected. A similar demonstration, using a different scaling of the variables, is found in Ref. [20], and with some finite-size corrections in Ref. [9]. Other authors have been able to derive Heaps' law from Zipf's law using a master equation [21] or pure scaling arguments [22]. Alternatives to Heaps' formula are listed in Ref. [23], but without a theoretical justification.

However, even simple visual inspection of the log-log plot of empirical type-token growth curves shows that Heaps' law is not even a rough approximation of the reality. On the contrary, a clear convexity (as seen from above) is apparent in most of the plots (see, for instance, some of the figures in [9, 24, 25]). This has been attributed to the fact that the asymptotic regime is not reached or to the effects of the exhaustion of the number of different types [26]. Nevertheless, the effect persists in very large systems, composed by many millions of tokens, and where the finiteness of the number of available types is questionable [22].

In the few reported cases where there seems to be a true power-law relation between number of tokens and number of types, as in Ref. [20], this turns out to come from a related but distinct statistics. Instead of considering the type-token growth curve in a single, growing system ($v(\ell)$ for $\ell = 1 \dots L$), one can look for the total type-token relationship in a collection or ensemble of $\mathcal{N}$ systems (V_j versus L_j, for $j = 1 \dots \mathcal{N}$, with $V_j = v(L_j)$), see also Refs. [21, 27–29]. We are, in contrast, interested in the type-token relation of a single growing system.

The fact that Heaps' law is so clearly violated for the type-token growth, given that this law follows directly from Zipf's law, casts doubts on the very validity of the latter law. But one may notice that, although the two versions of Zipf's law mentioned above are usually considered as equivalent, they are only asymptotically equivalent for high values of the count of tokens (i.e., for low ranks) [15, 17, 18]. However, the type-token growth curve emerges mainly from the statistics of the rarest types (i.e., the types with $n = 1$, for each value of ℓ), as it is only when a type appears for the first time that it contributes to the growth curve [22], and these are precisely the types for which the usual description in terms of the rank-count representation becomes problematic. So, the election of which is the form of Zipf's law that one considers to hold true becomes crucial for the derivation of the type-token growth curve and the fulfilment of Heaps' law or not.

Although most previous research has focused in Zipf's law in the rank-count representation, *i.e.*, the first version mentioned above, we argue that it is the second version of the law, that of the distribution of counts, the one that becomes relevant to describe the real type-token growth curve, at least in the case of written texts. Indeed, let us notice that the previously mentioned derivations of Heaps' law were all based on the rank-count representation [9, 16, 18–22]; therefore, the violation of Heaps' law for real systems invalidates the (exact) fulfilment of Zipf's law for the rank-count representation.

In contrast, when the viewpoint of Zipf's law for the distribution of counts is adopted, we prove that Heaps' law cannot be sustained for random systems and we derive an alternative law, which leads to "universal-like" shapes of the rescaled type-token growth curves, with the only dependence on the value of the Zipf's exponent. Quite unexpectedly, our prediction for random uncorrelated systems holds very well also for real texts. We are able to explain this effect despite the significant clustering or burstiness of word occurrences [30, 31], due to the singular role that the first appearance of a type plays in the type-token growth curve, in contrast to subsequent appearances.

Let us consider a Zipf's system of total size L, and a particular type with overall number of counts n; this means that the complete system contains n tokens of that type (and then L is the sum of counts of all types, $L = \sum_i n_i$). In fact, Zipf's law tells us that there can be many types with the same counts n, and we denote this number as $N_L(n)$. Quantitatively, in terms of the distribution of counts, Zipf's law reads

$$N_L(n) \propto \frac{1}{n^\gamma}, \tag{2}$$

for $n = 1, 2, \dots$ with the exponent γ close to 2. Note that $N_L(n)$ is identical, except for normalisation, to the probability mass function of the number of counts.

For a part of the system of size ℓ, with $\ell \leq L$, the number of types with k counts will be $N_\ell(k)$. The dependence of this quantity with the global $N_L(n)$ will be computed for a random system, which is understood as a sequence of tokens where these are taken at random from some underlying distribution. The $N_L(n)$ words with number of counts n in the whole system will lead, on average, to $N_L(n)h_{k,n}$ types with counts k in the subset, with $k \leq n$ and $h_{k,n}$ given by the hypergeometric distribution,

$$h_{k,n} = \frac{\binom{n}{k}\binom{L-n}{\ell-k}}{\binom{L}{\ell}}. \tag{3}$$

This is the probability to get k instances of a certain type when drawing, without replacement, ℓ tokens from a total population of L tokens of which there are n tokens of the desired type. The dependence of $h_{k,n}$ on ℓ and L is not explicit, to simplify the notation. The average number of types with k counts in the subset of size ℓ will result from the sum of $N_L(n)h_{k,n}$ for all $n \geq k$, i.e.,

$$N_\ell(k) = \sum_{n \geq k} N_L(n) h_{k,n}. \tag{4}$$

We will use this relationship between $N_\ell(k)$ and $N_L(n)$ to derive the type-token growth curve. For a subset of size ℓ we will have that, out of the total V types, $v(\ell)$ will be present whereas $N_\ell(0)$ will not have appeared (and so, their number of counts will be $k = 0$); therefore, $v(\ell) = V - N_\ell(0)$, and substituting Eq. (4) for $k = 0$ and using that $N_L(0) = 0$, then,

$$v(\ell) = V - \sum_{n \geq 1} N_L(n) h_{0,n}. \tag{5}$$

This formula relates the type-token growth curve with the distribution of counts in a random system, where it is exact, if we interpret $v(\ell)$ as an average over the

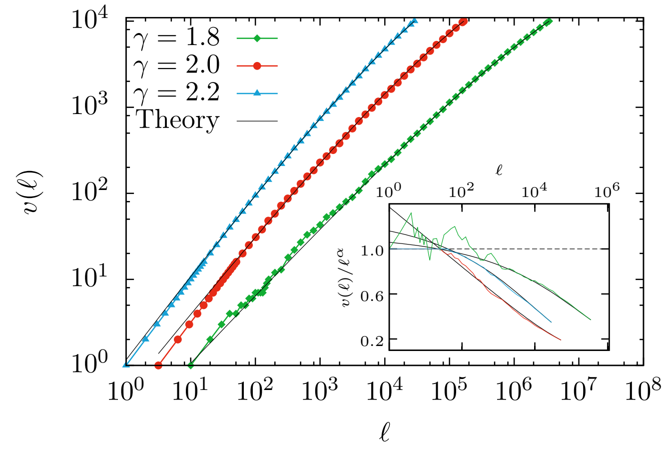

FIG. 1: **Main:** Type-token growth curve $v(\ell)$ for three random systems with number of counts drawn from a discrete power-law distribution $N_L(n) \propto n^{-\gamma}$, and $\gamma = 1.8$ (green diamonds), 2.0 (red circles) and 2.2 (blue triangles). The black lines correspond to our theoretical predictions, Eq. (9) for $\gamma \geq 2$ and Eq. (10) for $\gamma < 2$ (plotted with the help of the GSL libraries). No average over the reshuffling procedure is performed. Curves are consecutively shifted by a factor of $\sqrt{10}$ in the x-axis. **Inset:** The ratio $v(\ell)/\ell^\alpha$ is displayed, with $\alpha = \min\{1, \gamma - 1\}$, showing that an approximation of the form $v(\ell) \propto \ell^\alpha$ is too crude.

random ensemble.

We now show that a power-law distribution of type counts does not lead to a power law in the type-token growth curve, in other words, Zipf's law for the distribution of counts does not lead to Heaps' law, in the case of a random system. Assuming that $n \ll L$, the "zero-success" probability $h_{0,n}$ can be approximated as follows (see SI for details),

$$h_{0,n} = \frac{\binom{L-n}{\ell}}{\binom{L}{\ell}} = \simeq \left(1 - \frac{\ell}{L}\right)^n, \tag{6}$$

which in practice holds for all types; in fact, the smallest number of counts, for which the approximation is better, give the largest contribution to Eq. (5), due to the power-law form of $N_L(n)$. This is given, taking into account a normalisation constant A, by

$$N_L(n) = V\frac{A}{n^\gamma}, \tag{7}$$

for $n = 1, 2, \ldots$ (and zero otherwise), with $\sum_{n\geq1} N_L(n) = V$. Let us substitute the previous expressions for $h_{0,n}$ and $N_L(n)$ into Eq. (5), then

$$v(\ell) \simeq V\left(1 - A\sum_{n\geq1}\frac{(1-\ell/L)^n}{n^\gamma}\right). \tag{8}$$

Although there exists a maximum number of counts $n_{\max}$ beyond which $N_L(n) = 0$, as a first approximation the sum can be safely extended up to infinity, and hence we reach the following expression:

$$v(\ell) \simeq V\left(1 - \frac{\mathrm{Li}_\gamma(1-\ell/L)}{\zeta(\gamma)}\right), \tag{9}$$

where we have made use of the polylogarithm function, $\mathrm{Li}_\gamma(z) = \sum_{n=1}^{\infty} z^n/n^\gamma$, defined for $|z| < 1$, and of the fact that the normalisation of Zipf's law is given by $A = 1/\zeta(\gamma)$, with $\zeta(\gamma)$ the Riemann zeta function, $\zeta(\gamma) = \mathrm{Li}_\gamma(1)$. Notice that, for random systems with fixed γ, Eq. (9) yields a "universal" scaling relationship between the number of types $v(\ell)$, if expressed in units of the total number of types V, and the text position ℓ expressed in units of the total size L.
In fact, Eq. (9) can lead to an overestimation of $v(\ell)$ due to finite-size effects, but this is rarely noticeable in practice. If one wants a more precise version of Eq. (9), then, going back to Eq. (8) and limiting the sum up to $n_{\max}$ gives, after some algebra (see SI for details),

$$v(\ell) = V\left(1 - \frac{\mathrm{Li}_\gamma(q) - q^{n_{\max}+1}\Phi(q,\gamma,n_{\max}+1)}{\zeta(\gamma) - \Phi(1,\gamma,n_{\max}+1)}\right), \tag{10}$$

with $q = 1 - \ell/L$, and $\Phi(z,\gamma,a) = \sum_{n=0}^{\infty}\frac{z^n}{(a+n)^\gamma}$, $|z| < 1; a \neq 0, -1, \ldots$ the Lerch transcendent. Obviously, Eq. (10) gives better results at the cost of using an additional parameter, $n_{\max}$. As a rule of thumb, it appears to be worth the cost in cases where $\gamma < 2$, $\ell \ll L$ and L is not too large. In most practical cases Eq. (9) gives an excellent approximation; nevertheless, we include its more refined version, Eq. (10), for the sake of completeness.
In order to test these predictions, we simulate a random Zipf's system as follows: Let us draw $V = 10^4$ random numbers $n_1, n_2, \ldots n_V$, from the discrete probability distribution $N_L(n)/V = n^{-\gamma}/\zeta(\gamma)$, with $\gamma = 1.8, 2.0$ and 2.2. Each of these V values of n represents a type, with a number of counts given by the value of n. For each type $i = 1, \ldots, V$, we create then n_i copies (tokens) of its associated type, and make a list with all of them,

$$\underbrace{1,\ldots,1}_{n_1},\underbrace{2,\ldots,2}_{n_2},\ldots,\underbrace{V,\ldots,V}_{n_V}. \tag{11}$$

Then, the list is reshuffled in order to create a random system, of size $L = n_1 + n_2 + \cdots + n_V$. Figure 1 shows the resulting type-token growth together with the approximation given either by Eq. (9), which only depends on γ, or by Eq. (10), which depends on γ and $n_{\max}$. The agreement is nearly perfect, except for very small ℓ.
So far we have shown that Eqs. (9) and (10) capture very accurately the type-token growth curve for synthetic systems that have a perfect power-law distribution of counts but are completely random. Real systems, however, can have richer structures beyond the distribution of counts [30–32] and so one wonders if our derivations can provide acceptable predictions for them. In the following, we

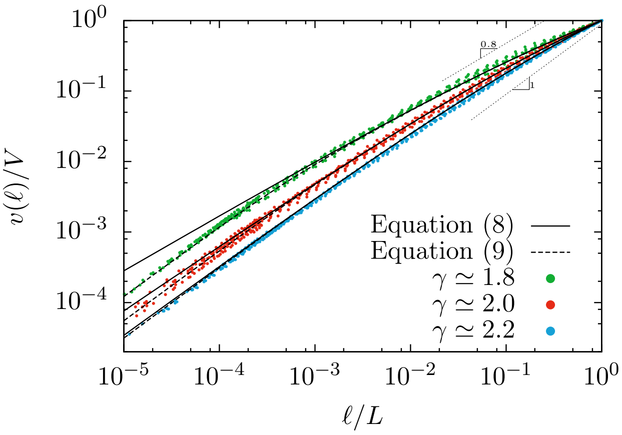

FIG. 2: The rescaled vocabulary-growth curve of 28 books from the PG database with exponents $\gamma = \{1.8, 2.0, 2.2\} \pm 0.01$ fitted for $n \geq 1$ or $n \geq 2$. The values of L and V range from $27,873$ to $146,845$ and from $5,639$ to $30,912$ respectively. As it is apparent, all books with the same exponent collapse into a single curve, which Eqs. (9) and (10) accurately capture. For the case of Eq. (10), we have used a value of $n_{\max}/L = 0.05$. The dotted straight lines (shifted for clarity) indicate the behaviour predicted by Heaps' law.

show that this is indeed the case when the system considered is that of natural language, and provide a qualitative explanation of this remarkable fact.

We analyse books from the *Project Gutenberg* (PG) database [33], selecting those whose distribution of frequencies $N_L(n)$ is statistically compatible with a pure, discrete power-law distribution. We fit the γ exponent with rigorous methods, see Refs. [34, 35]. In analogy with the previous section, we plot in Fig. 2 $v(\ell)/V$ versus ℓ/L for a total of 28 books for which $\gamma = 1.8, 2.0$, or 2.2. Books with the same Zipf's exponent collapse between them and into the corresponding theoretical curves, Eqs. (9) and (10). This is rather noticeable, as it points to the idea that the vocabulary-growth curve is unaffected by clustering, correlations, or by syntactic or discursive constraints. In other words, the vocabulary-growth curve of a real book fulfilling Zipf's law as given by Eq. (2) is not a power law but can be predicted using only its associated Zipf's exponent.

In order to understand why a prediction that heavily depends on the randomness hypothesis works so well for real books, we analyse the inter-occurrence-distance distribution of words. Given a word (type) with frequency n, we define its k-th inter-occurrence distance τ_k as the number of words (tokens) between its $k-1$-th and k-th appearances, plus one; i.e.,

$$\tau_k = \ell_k - \ell_{k-1} \tag{12}$$

(with ℓ_k the position of its k-th appearance and $k \leq n$). For the case of $k = 1$, we compute the number of words from the beginning of the text up to the first appearance, i.e., $\tau_1 = \ell_1$. If real books were completely random, then τ_k would be roughly exponentially distributed, and the rescaled distances

$$\hat{\tau}_k = \frac{\tau_k}{\langle \tau_k \rangle} \tag{13}$$

would be, for any value of n, exponentially distributed with parameter 1. Deviations from an exponential distribution for inter-occurrence distances in real books are well-known when all $k > 1$ are considered together, and constitute the so-called clustering or burstiness effect: instances of a given word tend to appear grouped together in the book, forming clusters and hence both very short and very long inter-occurrence distances are much more common than what an exponential distribution predicts [30, 31].

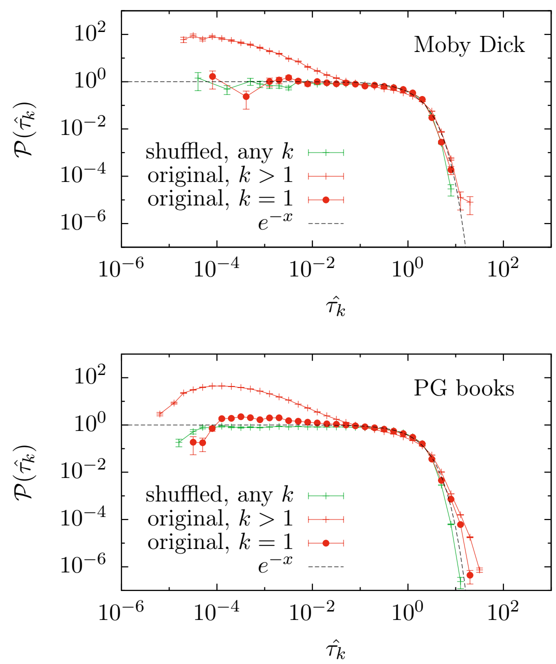

FIG. 3: Distribution of the rescaled inter-occurrence distances $\hat{\tau}_k$, see Eq. (13). The scale parameter $\langle \tau_k \rangle$ was computed from the data of each book (types with $n = 1$ or with $N(n) = 1$ were not included in the analysis). The original books (red) display clear deviations from an exponential distribution for $k > 1$, but not for $k = 1$. Shuffled versions of the books (green) do not show, as expected, any clustering effect, and hence their rescaled inter-occurrence distances are roughly exponentially distributed. **Top**: The book *Moby Dick*, by Herman Melville, as an illustrative example. **Bottom**: The one hundred longest books in the PG database.

In contrast to previous works [30, 31], our analysis introduces the distinction between $k > 1$ and $k = 1$. Note that for what concerns the vocabulary-growth curve, all that matters is $k = 1$, as it is only the first appearance of each word that adds to the vocabulary. Figure 3 shows the (estimated) probability mass function $\mathcal{P}(\hat{\tau}_k)$ of the rescaled inter-occurrence distance for the book *Moby Dick* as an example (top), and for the one hundred longest books in the PG database (bottom). As it is apparent, for $k > 1$, the distributions of distances are not exponentially distributed, and we recover a trace of the clustering effect; however the case $k = 1$ displays a clearly different shape, much more close to an exponential distribution. This explains, at a qualitative level, why our derivations, based on a randomness assumption, continue to work in the case of real books that display clustering effects.

In conclusion, we have shown that Eqs. (9) and (10), which are not power laws but contain the polylogarithm function and the Lerch transcendent, provide a continuum of universality classes for type-token growth, depending only on the value of Zipf's exponent for the distribution of counts. We have verified our results both on synthetic random systems and on real books, showing that despite correlations or clustering effects, they remain valid as long as Zipf's law is fulfilled for the distribution of counts. Our results open the door to investigations in other contexts beyond linguistics, where the validity of Heaps' law could be questioned in a similar manner.

Acknowledgements. The authors appreciate comments from E. Beltran-Saez and M. Gerlach. A. C. has enjoyed a long-term collaboration with G. Boleda and R. Ferrer-i-Cancho. Research projects in which this work is included are FIS2012-31324, from Spanish MINECO, and 2014SGR-1307 and 2012FI-B-00422, from AGAUR.

[1] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Cont. Phys.*, 46:323 –351, 2005.
[2] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51:661–703, 2009.
[3] L. A. Adamic and B. A. Huberman. Zipf's law and the Internet. *Glottometrics*, 3:143–150, 2002.
[4] C. Furusawa and K. Kaneko. Zipf's law in gene expression. *Phys. Rev. Lett.*, 90:088102, 2003.
[5] R. L. Axtell. Zipf distribution of U.S. firm sizes. *Science*, 293:1818–1820, 2001.
[6] J. Serrà, A. Corral, M. Boguñá, M. Haro, and J. Ll. Arcos. Measuring the evolution of contemporary western popular music. *Sci. Rep.*, 2:521, 2012.
[7] Y. Malevergne, V. Pisarenko, and D. Sornette. Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. *Phys. Rev. E*, 83:036111, 2011.
[8] D. Zanette. *Statistical Patterns in Written Language.* 2012.
[9] L. Lü, Z.-K. Zhang, and T. Zhou. Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE*, 5(12):e14139, 12 2010.
[10] A. Corral, G. Boleda, and R. Ferrer-i-Cancho. Zipf's law for word frequencies: word forms versus lemmas in long texts. *PLoS ONE*, (submitted), 2015.
[11] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Math.*, 1 (2):226–251, 2004.
[12] R. Ferrer i Cancho and R. V. Solé. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. U.S.A.*, 100:788–791, 2003.
[13] A. Saichev, Y. Malevergne, and D. Sornette. *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. Springer Verlag, Berlin, 2009.
[14] B. Corominas-Murtra, R. Hanel, and S. Thurner. Understanding scaling through history-dependent processes with collapsing sample space. *Proc. Natl. Acad. Sci. USA*, in press, 2015.
[15] H. S. Heaps. *Information retrieval: computational and theoretical aspects.* Academic Press, 1978.
[16] R. Baeza-Yates and G. Navarro. Block addressing indices for approximate text retrieval. *J. Am. Soc. Inform. Sci.*, 51(1):69–82, 2000.
[17] H. Baayen. *Word Frequency Distributions.* Kluwer, Dordrecht, 2001.
[18] B. Mandelbrot. On the theory of word frequencies and on related Markovian models of discourse. In R. Jakobson, editor, *Structure of Language and its Mathematical Aspects*, pages 190–219. American Mathematical Society, Providence, RI, 1961.
[19] D.C. van Leijenhorst and Th.P. van der Weide. A formal derivation of Heaps' law. *Inform. Sciences*, 170:263 – 272, 2005.
[20] A. Kornai. How many words are there? *Glottom.*, 2:61–86, 2002.
[21] M. A. Serrano, A. Flammini, and F. Menczer. Modeling statistical properties of written text. *PLoS ONE*, 4(4):e5372, 2009.
[22] F. Font-Clos, G. Boleda, and A. Corral. A scaling law beyond Zipf's law and its relation with Heaps' law. *New J. Phys.*, 15:093033, 2013.
[23] G. Wimmer and G. Altmann. On vocabulary richness. *J. Quant. Linguist.*, 6:1–9, 1999.
[24] Y. Sano, H. Takayasu, and M. Takayasu. Zipf's law and Heaps' law can predict the size of potential words. *Prog. Theor. Phys. Supp.*, 194:202–209, 2012.
[25] S. Bernhardsson, S. K. Baek, and P. Minnhagen. A paradoxical property of the monkey book. *J. Stat. Mech.*, 2011(07):P07013, 2011.
[26] L. Lü, Z.-K. Zhang, and T. Zhou. Deviation of Zipf's and Heaps' Laws in human languages with limited dictionary sizes. *Sci. Rep.*, 3:1–7, 2013.
[27] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Sci. Rep.*, 2:943, 2012.
[28] M. Gerlach and E. G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3:021006, 2013.
[29] M. Gerlach and E. G. Altmann. Scaling laws and fluctuations in the statistics of word frequencies. *New J. Phys.*, 16(11):113010, 2014.
[30] A. Corral, R. Ferrer-i-Cancho, and A. Díaz-Guilera. Universal complex structures in written language. *http://arxiv.org*, 0901.2924, 2009.
[31] E. G. Altmann, J. B. Pierrehumbert, and A. E. Mot-

ter. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *ArXiv*, 0901.2349v1, 2009.

[32] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1999.

[33] http://www.gutenberg.org/.

[34] A. Deluca and A. Corral. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophys.*, 61:1351–1394, 2013.

[35] A. Corral, A. Deluca, and R. Ferrer-i-Cancho. A practical recipe to fit discrete power-law distributions. *ArXiv*, 1209:1270, 2012.

Supplementary Information:
Log-log Convexity of Type-Token Growth in Zipf's Systems

Francesc Font-Clos[1,2] and Álvaro Corral[1,2]

[1] *Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, E-08193 Barcelona, Spain.*
[2] *Departament de Matemàtiques, Facultat de Ciències, Universitat Autònoma de Barcelona, E-08193 Barcelona, Spain*

This document contains Supplementary Information (SI) for the article entitled "Log-log Convexity of Type-Token Growth in Zipf's Systems". It consists of two sections that give details of derivations and algebraic manipulations used to reach equations (6) and (10) in the main text.

DERIVATION OF THE APPROXIMATION FOR THE "ZERO-SUCCESS" PROBABILITY $h_{0,n}$

Equation (6) in the main text gives an approximation for the zero-success probability $h_{0,n}$.

$$h_{0,n} = \frac{\binom{L-n}{\ell}}{\binom{L}{\ell}} \simeq \left(1 - \frac{\ell}{L}\right)^n, \tag{S1}$$

While the derivation is fairly elementary, some caution must be taken in order to get the desired result. Note that other approximations are possible, but they are not useful for our purposes.

The first equal sign in Eq. (S1) is just the definition of $h_{0,n}$, i.e. the probability mass function of a hypergeometric random variable that takes value $k = 0$ (number of successes), with parameters ℓ (number of draws), L (total population) and n (number of successes in the population). To get to the desired result, we write the binomial coefficients in terms of factorials, cancel out one $\ell!$ term, and regroup the rest of terms as follows:

$$\frac{\binom{L-n}{\ell}}{\binom{L}{\ell}} = \left[\frac{(L-n)!}{(L-n-\ell)!\ \ell!}\right] \Big/ \left[\frac{L!}{(L-\ell)!\ \ell!}\right] \tag{S2}$$

$$= \frac{(L-n)!}{L!} \times \frac{(L-\ell)!}{(L-\ell-n)!}. \tag{S3}$$

Notice that each fraction in equation (S3) has n terms, so that

$$\frac{\binom{L-n}{\ell}}{\binom{L}{\ell}} = \frac{(L-\ell)(L-\ell-1)\dots(L-\ell-n+1)}{L(L-1)\dots(L-n+1),} \tag{S4}$$

$$= \prod_{j=0}^{n-1}\left(\frac{L-\ell-j}{L-j}\right). \tag{S5}$$

The first term, $j = 0$, is equal to $(1 - \ell/L)$. If n is small compared to L, then $L - j \simeq L$ for all $j = 0, \dots, n-1$, and each of the n terms in the above product can be approximated by the first one. The result then follows easily, as

$$h_{0,n} = \frac{\binom{L-n}{\ell}}{\binom{L}{\ell}} = \prod_{j=0}^{n-1}\left(\frac{L-\ell-j}{L-j}\right) \simeq \prod_{j=0}^{n-1}\left(\frac{L-\ell}{L}\right) = \left(1-\frac{\ell}{L}\right)^n. \tag{S6}$$

It is important to bear in mind that the only assumption used was $L - n \simeq L$, or equivalently, $n \ll L$, but nothing was assumed regarding the ratio ℓ/L, and thus the approximation should work equally well for any ℓ, once L and n are fixed. In practice, the most common type, with $n_{\max}$ counts, will give an upper bound to the error that this approximation introduces. In the case of written books, $n_{\max}/L \simeq 0.05$ is a typical value for many languages, so that even in the worst case, the approximation is already quite good. In addition, most types have low counts (this is in essence Zipf's law), and hence the sum of equation (8) in the main text will be dominated by terms where the approximation is very good.

DERIVATION OF THE TYPE-TOKEN GROWTH CURVE $v(\ell)$ WHEN $n_{\max}$ IS TAKEN INTO ACCOUNT

Equation (10) in the main text offers a more refined version of the type-token growth curve, equation (8), by limiting the sum up to a given maximum value of the number of counts $n_{\max}$. Let us denote $q = 1 - \ell/L$ and introduce the Lerch transcendent $\Phi(z, \gamma, a)$, defined as

$$\Phi(z, \gamma, a) = \sum_{n=0}^{\infty} \frac{z^n}{(a+n)^\gamma}, \quad |z| < 1; a \neq 0, -1, \dots \; . \tag{S7}$$

We first perform the sum $\sum_{n=1}^{n_{\max}} q^n/n^\gamma$ as follows:

$$\sum_{n=1}^{n_{\max}} \frac{q^n}{n^\gamma} = \sum_{n=1}^{\infty} \frac{q^n}{n^\gamma} - \sum_{n=n_{\max}+1}^{\infty} \frac{q^n}{n^\gamma} \tag{S8}$$

$$= \mathrm{Li}_\gamma(q) - q^{n_{\max}+1} \sum_{n=n_{\max}+1}^{\infty} \frac{q^{n-n_{\max}-1}}{n^\gamma}, \tag{S9}$$

and defining $m = n - n_{\max} - 1$,

$$\sum_{n=1}^{n_{\max}} \frac{q^n}{n^\gamma} = \mathrm{Li}_\gamma(q) - q^{n_{\max}+1} \sum_{m=0}^{\infty} \frac{q^m}{(m + n_{\max} + 1)^\gamma} \tag{S10}$$

$$= \mathrm{Li}_\gamma(q) - q^{n_{\max}+1} \Phi(q, \gamma, n_{\max} + 1). \tag{S11}$$

The normalization constant can be similarly computed,

$$A^{-1} = \sum_{n=1}^{n_{\max}} \frac{1}{n^\gamma} = \sum_{n=1}^{\infty} \frac{1}{n^\gamma} - \sum_{n=n_{\max}+1}^{\infty} \frac{1}{n^\gamma} \tag{S12}$$

$$= \zeta(\gamma) - \Phi(1, \gamma, n_{\max} + 1), \tag{S13}$$

and the result in the main text follows:

$$v(\ell) = V \left(1 - A \sum_{n=1}^{n_{\max}} \frac{(1 - \ell/L)^n}{n^\gamma} \right) \tag{S14}$$

$$= V \left(1 - \frac{\mathrm{Li}_\gamma(q) - q^{n_{\max}+1} \Phi(q, \gamma, n_{\max} + 1)}{\zeta(\gamma) - \Phi(1, \gamma, n_{\max} + 1)} \right). \tag{S15}$$

As discussed in the main text, this improved version of the type-token growth curve makes use of an additional parameter, $n_{\max}$, and so it is not surprising that it gives better results. However, in practice only for $\gamma < 2$ is this usually noticeable. This is related to the tail of the distribution of counts, as for fixed $n_{\max}, \ell/L$ the overall weight given to the types with $n > n_{\max}$ [1] increases if γ is decreased. In any case, it is worth noting that (i) the additional "parameter" $n_{\max}$ is usually known in practical cases, so that it does not need to be fitted, and (ii) the polylogarithm function, Riemman's zeta function and the Lerch transcendent are functions with well-studied properties and usually can be found in most numerical packages.

---

[1] Obviously, these types are not present in the system, but the simple version of the type-token curve, Eq. (9) in the main text, includes them in the sum.