Languages cool as they expand: Allometric scaling and the decreasing need for new words

Alexander M. Petersen,[1] Joel Tenenbaum,[2] Shlomo Havlin,[3] H. Eugene Stanley,[2] and Matjaž Perc[4]

[1]*Laboratory for the Analysis of Complex Economic Systems, IMT Institute for Advanced Studies Lucca, 55100 Lucca, Italy*
[2]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA*
[3]*Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel*
[4]*Department of Physics, Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia*

(Dated: July 9, 2012)

We study language evolution by analyzing the word frequencies of millions of distinct words in seven languages recorded in books from the past two centuries. For all languages and time spans we confirm that two scaling regimes characterize the word frequency distributions, with the more common words in each language obeying the Zipf law. We measure the allometric scaling relation between corpus size and vocabulary size, confirming recent theoretical predictions that relate the Heaps law to the Zipf law. We measure a decreasing trend in the annual growth fluctuations of word use with increasing corpus size suggesting that the rate of linguistic evolution decreases as the language expands, implying that new words have increasing marginal returns, and that languages can be said to "cool by expansion." Counteracting this cooling are periods of political conflict which are not only characterized by decreases in literary productivity but also by a globalized media focus which may increase the mobility of concepts and words across political borders.

The annals of written language housed in libraries around the world serve as an immense "crowd-sourced" historical record that traces humanity further back than the limits of oral history. Google Inc.'s massive book digitization project presents this collection of written language to the public in the form of the *Google Books Ngram Viewer* application [1]. Approximately 4% of all books ever published have been scanned, making available over 10^7 cultural word trajectories that archive the dynamics of word use in seven different languages over a period of more than two centuries. It is the availability of such vast amounts of digitized data, sometimes called "Big Data", paired with interdisciplinary research efforts, that fuels much current progress in both social and natural sciences [2–4] and makes possible knowledge about knowledge, or "metaknowledge" [5].

The application of Google's high-throughput data collection and analysis to the study of human culture was recently termed "culturomics" by Michel et al. [6], who demonstrated that case studies of individual words provide insight into such aspects of our culture as linguistics, new technologies, and epidemiology. By performing a large scale analysis of the same data using methods of statistical physics and concepts from economics, Petersen et al. [7] found language-independent selection laws that govern fluctuations in word use and determine the extent of cultural memory. The latter was also investigated by Gao et al. [8], who found that words describing social phenomena tend to have different long-range correlations than words describing natural phenomena. Here we show that the allometric scaling properties of language imply that new words have an *increasing* marginal return, and that languages can be said to "cool by expansion."

We analyze the macroscopic scaling patterns that characterize word use frequency and the growth of a corpora in a large body (or "corpus") of text. The Zipf law [9–15], quantifying the distribution of word frequencies, and the Heaps law [11, 16–18], relating the size of a corpus to the vocabulary size of that corpus, are commonly used to demonstrate how aspects of the complexity of language can be captured by remarkably simple statistical patterns. While these laws have been exhaustively tested on relatively small snapshots of empirical data aggregated over short periods of time, here we test the validity of these laws on extremely large corpora aggregated at the 1-year time resolution.

Interestingly, we observe two scaling regimes in the probability density functions of word usage, with the Zipf law holding only for the set of more frequently used words, referred to as the "kernel lexicon" [12]. The word frequency distribution is different for the more rarely used words, suggesting that rare words belong to a distinct class. Heaps observed that the number of words in a vocabulary size exhibits sub-linear growth with document size [16] and postulated what we will describe here as an increasing "marginal utility" of new words (defined in Eq. (7)). However, even as new words may start in relative obscurity, their importance can be underappreciated by their initial frequency. A recent study [19] indicates that word niche can be just as much an essential factor in modeling word use dynamics. New niche words, for example, are anything but "marginal" - they are core words. This is particularly the case in online communities in which individuals strive to distinguish themselves on short timescales by developing stylistic jargon.

A corpora can be viewed as a collective voice, with aggregate patterns that don't necessarily reflect the microscopic individuals. At this microscopic level is the

"metabook" concept [17, 18], according to which word-frequency structures are author-specific, i.e. that the word-frequency characteristics of a random excerpt from a compilation of everything that a specific author could ever conceivably write (his/her "metabook") would accurately match those of the author's actual writings. The question is whether a compilation of all metabooks would still conform to the Zipf law and the Heaps law. The immense size and time span of the *Google n-gram* dataset allows us to examine this question in detail.

Results

Longitudinal analysis of written language. Allometric scaling analysis [20] characterizes the role of system size on system dynamics, and has been applied to systems as diverse as the metabolic rate of mitochondria [21] and city growth [22–28].

City growth in particular shares two common features with the growth of written text: (i) the Zipf law is able to describe the distribution of city sizes regardless of country or the time stamp on the data [25], and (ii) city growth has inherent constraints due to geography, changing labor markets and their effects on opportunities for innovation and wealth creation [26, 27], just as vocabulary growth is constrained by human brain capacity and the varying utilities of new words across users [12].

We construct a word counting framework by first defining the quantity $u_i(t)$ as the number of times word i is used in year t. Since the number of books and the number of distinct words grow dramatically over time, we define the *relative* word use, $f_i(t)$, as the fraction of the total body of text occupied by word i in the same year

$$f_i(t) \equiv u_i(t)/N_u(t), \tag{1}$$

where the quantity $N_u(t) \equiv \sum_{i=1}^{N_w(t)} u_i(t)$ is the total number of indistinct word uses (i.e. the size of the body of text) while $N_w(t)$ is the total number of distinct words digitized from books printed in year t (i.e. the vocabulary size). Both N_w and N_u are generally increasing over time.

The Zipf law and the two scaling regimes. Zipf investigated a number of bodies of literature and observed that the frequency of any given word is roughly inversely proportional to its rank [9], with the frequency of the z-ranked word given by the relation

$$f(z) \sim z^{-\zeta}, \tag{2}$$

with a scaling exponent $\zeta \approx 1$. This empirical law has been confirmed for a broad range of data, ranging from income rankings, city populations, and the varying sizes of avalanches, forest fires [29] and firm size [30] to the linguistic features of nonconding DNA [31]. The Zipf law can be derived through the "principle of least effort," which minimizes the communication noise between speakers (writers) and listeners (readers) [14]. The Zipf law has been found to hold for a large dataset of English text [12], but interestingly deviates from the schizophrenic lexicon [13]. Here, we also find statistical regularity in the distribution of relative word use for 11 different datasets, each comprising more than half a million distinct words taken from millions of books [6].

Figure 1 shows the probability density functions $P(f)$ resulting from data aggregated over all the years (A,B) as well as for the year $t = 2000$ alone (C,D). Regardless of the language and the considered time span, the probability density functions are characterized by a striking two-regime scaling, which was first noted by Ferrer i Cancho and Solé [12], and can be quantified as

$$P(f) \sim \begin{cases} f^{-\alpha_-}, & \text{if } f < f_\times \text{ ["unlimited lexicon"]} \\ f^{-\alpha_+}, & \text{if } f > f_\times \text{ ["kernel lexicon"]} . \end{cases} \tag{3}$$

These two regimes, designated "kernel lexicon" and "unlimited lexicon," are thought to reflect the cognitive constraints of the brain's finite vocabulary [12]. The specialized words found in the unlimited lexicon are not universally shared and are used less frequently than the words in the kernel lexicon. This is reflected in the kink in the probability density functions and gives rise to the two-scaling regime shown in Fig. 1.

The exponents α_+ and the corresponding rank-frequency scaling exponent ζ are related asymptotically by [12]

$$\alpha_+ \approx 1 + 1/\zeta, \tag{4}$$

with no analogous relationship for the unlimited lexicon value ζ_-. Table I lists the average α_+ and α_- values calculated by aggregating $\alpha_\pm$ values for each year using Hill's maximum likelihood estimator for the power-law distribution [32]. We characterize the two scaling regimes using a crossover region around $f_\times \approx 10^{-5}$ to distinguish between α_- and α_+: (i) $10^{-8} \leq f \leq 10^{-6}$ corresponds to α_- and (ii) $10^{-4} \leq f \leq 10^{-1}$ corresponds to α_+. For words (1-grams [6]) that satisfy $f \gtrsim f_\times$, hence making up the kernel lexicon, we verify the Zipf scaling law $\zeta \approx 1$ (corresponding to $\alpha \approx 2$) for all corpora analyzed. For the unlimited lexicon regime $f \lesssim f_\times$, however, the Zipf law is not obeyed, as we find $\alpha_- \approx 1.7$. Note that α_- is significantly smaller in the Hebrew, Chinese, and the Russian corpora, which suggests that a more generalized version of the Zipf law [12] may need to be slightly language-dependent, especially when taking into account the usage of specialized words from the unlimited lexicon.

The Heaps law and the increasing marginal returns of new words. Heaps observed that vocabulary size, i.e. the number of distinct words, exhibits a sub-linear growth with document size [16]. Entitled "the Heaps law", this observation has important implications for the "return on investment" of a new word as it is established and becomes disseminated throughout the lit-

erature of a given language. As a proxy for this return, Heaps studied how often new words are invoked in lieu of preexisting competitors and examined the linguistic value of new words and ideas by analyzing the relation between the total number of words printed in a body of text N_u, and the number of these which are distinct N_w, i.e. the vocabulary size [16]. The marginal returns of new words, $\partial N_u/\partial N_w$ quantifies the impact of the addition of a single word to the vocabulary of a corpus on the aggregate output (corpus size).

For individual books, the empirically-observed scaling relation between N_u and N_w obeys

$$N_w \sim (N_u)^b, \tag{5}$$

with $b < 1$, with Eq. (5) referred to as "the Heaps law" [16]. Using a stochastic model for the growth of book vocabulary size as a function of book size, Serrano et al. [11] recently proposed that $b = 1/\alpha$, where α is the scaling exponent in the probability density function $P(f)$ of relative word use,

$$P(f) \sim f^{-\alpha}. \tag{6}$$

Figure 2 confirms a sub-linear scaling ($b < 1$) between N_u and N_w for each corpora analyzed. Interestingly, Chinese and Russian display two Heaps scaling regions, as depicted in Fig. 2. These results show how the marginal returns of new words are directly related to the distribution of relative word use, given by

$$\frac{\partial N_u}{\partial N_w} \sim (N_w)^{\alpha-1}, \tag{7}$$

which is an increasing function of N_w for $\alpha > 1$. Thus, the relative increase in the induced volume of written languages is larger for new words than for old words. This is likely due to the fact that new words are typically technical in nature, requiring additional explanations that put the word into context with pre-existing words. Specifically, a new word requires the additional use of more preexisting words resulting from (i) the proper explanation of the new word using existing technical terms, and (ii) the grammatical infrastructure underlying efficient communication. Hence, there are large spillovers in the size of the written corpus that follow from the intricate dependency structure of language which forms a complex network of words that serve various grammatical roles [33, 34].

In order to investigate the role of rare and new words, we calculate N_u and N_w using only common words that satisfy the word usage criteria $u_i(t) > U_c$. Because a word in a given year can not paper with a frequency less than $1/N_u$, we use a word use threshold U_c and not a word frequency threshold f_c, since pruning with a frequency threshold discriminates in its treatment of small corpora. This pruning also serves to progressively remove more and more rare words that can spuriously arise from Optical Character Recognition (OCR) errors in the digitization process, as well as from intrinsic spelling errors and orthographic spelling variations.

Figures 3 and 4 show the dependence of N_u and N_w for the English corpus on the exclusion of low-frequency words using a variable cutoff $U_c = 2^n$ with $n = 0 \ldots 11$. As U_c increases the Heaps scaling exponent increases from $b \approx 0.5$, approaching $b \approx 1$, indicating that core words are structurally integrated into language as a proportional background.

Table I summarizes the b values pertaining to different corpora, obtained by means of ordinary least squares regression of $N_u(t)$ versus $N_w(t)$ dependence for $U_c = 0$. Comparing the scaling exponent $\alpha_+ \approx 2$ [Eq. (6)] calculated from $P(f)$ in Fig. 1, we confirm the relation $b = 1/\alpha_+$ proposed by Serrano et al. [11], since $b \approx 0.5$ for all languages analyzed. This simple scaling relation highlights the underlying structure of language, which forms a dependency network between the kernel lexicon and the unlimited lexicon. Table I lists the average $\langle \alpha_\pm(t) \rangle$ calculated from annual estimates of $\alpha_\pm$. Moreover, the allometric scaling $\partial N_w/\partial N_u \sim (N_w)^{1-\alpha}$ shows a *decreasing marginal need* for additional words. Because we get more and more "mileage" out of new words in an already large language, additional words are needed less and less.

Corpora size and word-use fluctuations. Lastly, it is instructive to examine how vocabulary size N_w and the overall size of the corpora N_u affect fluctuations in word use. Figure 5 shows how N_w and N_u vary over time over the past two centuries. Note that, apart from the periods during the two World Wars, the number of words printed, which we will refer to as the "literary productivity", has been increasing over time. The number of distinct words has been rising as well, although for certain languages, e.g. Russian and Hebrew, vocabulary appears to saturate. Note also that the downturn in productivity during adverse conditions was seldom accompanied by a smaller vocabulary size. Indeed, the size of the French vocabulary spiked during World War II, although there was a sharp decrease in literary productivity.

To investigate the role of fluctuations, we focus on the logarithmic growth rate, commonly used in finance and economics

$$r_i(t) \equiv \ln f_i(t + \Delta t) - \ln f_i(t) = \ln \Big[\frac{f_i(t + \Delta t)}{f_i(t)} \Big], \tag{8}$$

to measure the relative growth of word use over 1-year periods, $\Delta t \equiv 1$ year. Recent quantitative analysis on the distribution $P(r)$ of word use growth rates $r_i(t)$ indicates that word usage increases and decreases by larger amounts than would be expected by null models for language evolution [7].

Figure 6 shows $\sigma_r(t)$, the standard deviation of $r_i(t)$ calculated across all words, which is an aggregate measure for the "temperature" (strength of fluctuations) within a given written corpora. Visual inspection sug-

gests a general decrease in $\sigma_r(t)$ over time, marked by sudden increases during times of political conflict. Hence, the persistent increase in the volume of written language is correlated with a persistent downward trend in $\sigma_r(t)$: as a language grows and matures it also "cools off".

Figure 7 indicates that for large $N_u(t)$, each language, excepting Chinese, is characterized by a scaling relation

$$\sigma_r(t) \sim N_u(t|f_c)^{-\beta} \ , \tag{9}$$

with f_c giving the threshold for word inclusion as described in Table I, and language-dependent scaling exponent $\beta \approx 0.12 - -0.29$ quantifying how the increase in corpus size relates to a decrease in word-use fluctuations. This size-variance relation is analogous to the decreasing growth rate volatility observed as complex economic entities (i.e. firms or countries) increase in size [35–38]. Furthermore, this size-variance relation was was also demonstrated at the scale of individual words using the same *Google n-gram* dataset [7].

Interestingly, this decreasing fluctuation scale is counteracted by the influx of new words which have growth-spurts around 30-50 years following their birth in the written corpora [7]. Moreover, the fluctuation scale $\sigma_r(t)$ is positively influenced by adverse conditions such as wars and revolutions. Although literary productivity falls, new words may emerge more frequently due to globalization effects. The decrease in $N_u(t)$ may decrease the level of competition between old words and new words, allowing new words to rise in use. This may be the case for Chinese, where the accelerated production of new words [note the scale of the vertical axis in Fig. 5(B)] seems to offset the increase in literary productivity, leaving $\sigma_r(t)$ approximately constant over $N_u(t|f_c)$ (see Fig. 7).

Discussion

A coevolutionary description of language and culture requires many factors and much consideration [39, 40]. While scientific and technological advances are largely responsible for written language growth as well as the birth of many new words [7], socio-political factors also play a strong role. For example, the sexual revolution of the 1960s triggered the sudden emergence of the words "girlfriend" and "boyfriend" in the English corpora [1], illustrating the evolving culture of romantic courting. Such technological and socio-political perturbations require case-by-case analysis for any deeper understanding, as demonstrated comprehensively by Michel et al. [6].

Here we analyzed the macroscopic properties of written language using the *Google Books* database [1]. We find that the word frequency distribution $P(f)$ is characterized by two scaling regimes. While frequently used words that constitute the kernel lexicon follow the Zipf law, the distribution has a less-steep scaling regime quantifying the rarer words constituting the *unlimited lexicon*. Our result is robust across languages as well as across other data subsets, thus extending the validity of the seminal observation by Ferrer i Cancho and Solé [12], who first reported it for a large body of English text. The kink in the slope preceding the entry into the unlimited lexicon is a likely consequence of the limits of human mental ability that force the individual to optimize the usage of frequently used words and forget specialized words that are seldom used. This hypothesis agrees with the "principle of least effort" that minimizes communication noise between speakers (writers) and listeners (readers), which in turn may lead to the emergence of the Zipf law [14].

By analyzing the dependence of vocabulary growth on corpus growth, we have also validated the Heaps law for extremely large written corpora spanning millions of authors and their "metabooks" [17]. Using words in the frequency range $0 < f < 10^{-9}$ we found agreement between the Zipf exponent $\alpha \approx 2$ and the Heaps exponent $b \approx 0.5$, confirming the theoretical prediction $\alpha = 1/b$ by Serrano et al. [11]. However, we find that the exclusion of extremely rare words has a strong affect on the b value, which approaches unity as U_c increases (Figs. 3 and 4).

The economies of scale ($b < 1$) indicates that there is an *increasing marginal return* for new words, or alternatively, a *decreasing marginal need* for new words, as evidenced by allometric scaling. This can intuitively be understood in terms of the increasing complexities and combinations of words that become available as more words are added to a language, lessening the need for lexical expansion. However, a relationship between new words and existing words is retained. Every introduction of a word, from an informal setting (e.g. an expository text) to a formal setting (e.g. a dictionary) is yet another chance for the more common describing words to play out their respective frequencies, underscoring the hierarchy of words. This can be demonstrated quite instructively from Eq. (7) which implies that for $b = 1/2$ that $\frac{\partial N_u}{\partial N_w} \propto N_w$, meaning that it requires a quantity proportional to the vocabulary size N_w to introduce a new word, or alternatively, that a quantity proportional to N_w necessarily results from the addition.

Though new words are needed less and less, the expansion of language continues, doing so with marked characteristics. Taking the growth rate fluctuations of word use to be a kind of temperature, we note that like an ideal gas, most languages "cool" when they expand. The fact that the relationship between the temperature and corpus volume is a power law, one may, loosely speaking, liken language growth to the expansion of a gas or the growth of a company [35–38]. In contrast to the static laws of Zipf and Heaps, we note that this finding is of a dynamical nature.

Other aspects of language growth may also be understood in terms of expansion of a gas. Since larger literary productivity imposes a downward trend on growth rate fluctuations, productivity itself can be thought of as a kind of inverse pressure in that highly productive years are observed to "cool" a language off. Also, it is during the "high-pressure" low productivity years that new

words tend to emerge more frequently.

Interestingly, the appearance of new words is more like gas condensation, tending to cancel the cooling brought on by language expansion. These two effects, corpus expansion and new word "condensation," therefore act against each other. The Chinese language appears to be the most affected by the latter, likely the result of globalization, wherein the counter effects render the growth fluctuations unaffected by corpus size. For other corpora, however, we calculate a size-variance scaling exponent $\beta \approx 0.2$.

In the context of allometric relations, Bettencourt et al. [26] note that the scaling relations describing the dynamics of cities show an *increase* in the characteristic pace of life as system size grows, whereas those found in biological systems show *decrease* in characteristic rates as the system size grows. Since the languages we analyzed tend to "cool" as they expand, there may be deep-rooted parallels with biological systems based on principles of efficiency [14]. Languages, like biological systems demonstrate economies of scale ($b < 1$) manifesting from a complex dependency structure that mimics a hierarchical "circulatory system" required by the organization of language [33, 41–45] and the limits of the efficiency of the speakers/writers who employ the words [17, 19, 46].

[1] Google Books Ngram Viewer, `http://books.google.com/ngrams`.

[2] Lazer, D., et al. Life in the network: the coming age of computational social science. *Science* **323**, 721–723 (2009).

[3] Barabási, A. L. The network takeover. *Nature Physics* **8**, 14–16 (2012).

[4] Vespignani, A. Modelling dynamical processes in complex socio-technical systems. *Nature Physics* **8**, 32–39 (2012).

[5] Evans, J. A. and Foster, J. G. Metaknowledge. *Science* **331**, 721–725 (2011).

[6] Michel, J. B., et al. Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182 (2011).

[7] Petersen, A. M., Tenenbaum, J., Havlin, S., and Stanley, H. E. Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports* **2**, 313 (2012).

[8] Gao, J., Hu, J., Mao, X., and Perc, M. Culturomics meets random fractal theory: Insights into long-range correlations of social and natural phenomena over the past two centuries. *J. R. Soc. Interface***9** 1956–1964 (2012).

[9] Zipf, G. K. *Human Behavior and the Principle of Least-Effort: An Introduction to Human Ecology.* Addison-Wesley, Cambridge, MA, (1949).

[10] Tsonis, A. A., Schultz, C., and Tsonis, P. A. Zipf's law and the structure and evolution of languages. *Complexity* **3**, 12–13 (1997).

[11] Serrano, M. Á., Flammini, A., and Menczer, F. Modeling statistical properties of written text. *PLoS ONE* **4**, e5372 (2009).

[12] Ferrer i Cancho, R. and Solé, R. V. Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics* **8**, 165–173 (2001).

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

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

[15] Baek, S. K., Bernhardsson, S., and Minnhagen, P. Zipf's law unzipped. *New J. Phys.* **13**, 043004 (2011).

[16] Heaps, H. S. *Information Retrieval: Computational and Theoretical Aspects.* Academic Press, New York, (1978).

[17] Bernhardsson, S., Correa da Rocha, L. E., and Minnhagen, P. The meta book and size-dependent properties of written language. *New J. Phys.* **11**, 123015 (2009).

[18] Bernhardsson, S., Correa da Rocha, L. E., and Minnhagen, P. Size-dependent word frequencies and translational invariance of books. *Physica A* **389**, 330–341 (2010).

[19] Altmann, E. G., Pierrehumbert, J. B., and Motter, A. E. Niche as a determinant of word fate in online groups. *PLoS ONE* **6**, e19009 (2011).

[20] Kleiber, M. Body size and metabolism. *Hilgardia* **6**, 315–351 (1932).

[21] West, G. B. Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. *Proc. Natl. Acad. Sci. USA* **98**, 2473–2478 (2002).

[22] Makse, H. A., Havlin, S., and Stanley, H. E. Modelling urban growth patterns. *Nature* **377**, 608–612 (1995).

[23] Makse, H. A., Andrade Jr., J. S., Batty, M., Havlin, S., and Stanley, H. E. Modeling urban growth patterns with correlated percolation. *Phys. Rev. E* **58**, 7054–7062 (1998).

[24] Rozenfeld, H. D., Rybski, D., Andrade Jr., J. S., Batty, M., Stanley, H. E., and Makse, H. A. Laws of population growth. *Proc. Natl. Acad. Sci. USA* **48**, 18702–18707 (2008).

[25] Gabaix, X. Zipf's law for cities: An explanation. *Quarterly Journal of Economics* **114**, 739–767 (1999).

[26] Bettencourt, L. M. A., Lobo, J., Helbing, D., Kuhnert, C., and West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci. USA* **104**, 7301–7306 (2007).

[27] Batty, M. The size, scale, and shape of cities. *Science* **319**, 769–771 (2008).

[28] Rozenfeld, H. D., Rybski, D., Gabaix, X., and Makse, H. A. The area and population of cities: New insights from a different perspective on cities. *American Economic Review* **101**, 2205–2225 (2011).

[29] Newman, M. E. J. Power laws, pareto distributions and Zipf's law. *Contemporary Phys.* **46**, 323–351 (2005).

[30] Stanley, M. H. R., Buldyrev, S. V., Havlin, S., Mantegna, R., Salinger, M., and Stanley, H. E. Zipf plots and the size distribution of firms. *Econ. Lett.* **49**, 453–457 (1995).

[31] Mantegna, R. N., et al. Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Phys. Rev. E* **52**, 2939–2950 (1995).

[32] Clauset, A., Shalizi, C. R., and Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
[33] Steyvers, M. and Tenenbaum, J. B. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cogn. Sci.* **29**, 41–78 (2005).
[34] Markosova, M. Network model of human language. *Physica A* **387**, 661–666 (2008).
[35] Amaral, L. A. N., et al. Scaling Behavior in Economics: I. Empirical Results for Company Growth. *J. Phys. I France* **7**, 621–633 (1997).
[36] Amaral, L. A. N., et al. Power Law Scaling for a System of Interacting Units with Complex Internal Structure. *Phys. Rev. Lett.* **80**, 1385–1388 (1998).
[37] Fu, D., et al. The growth of business firms: Theoretical framework and empirical evidence. *Proc. Natl. Acad. Sci. USA* **102**, 18801–18806 (2005).
[38] Riccaboni, M., Pammolli, F., Buldyrev, S. V., Ponta, L., and Stanley, H. E. The size variance relationship of business firm growth rates. *Proc. Natl. Acad. Sci. USA* **105**, 19595–19600 (2008).
[39] Mufwene, S. *The Ecology of Language Evolution.* Cambridge Univ. Press, Cambridge, UK, (2001).
[40] Mufwene, S. *Language Evolution: Contact, Competition and Change.* Continuum International Publishing Group, New York, NY, (2008).
[41] Sigman, M. and Cecchi, G. A. Global organization of the wordnet lexicon. *Proc. Natl. Acad. Sci. USA* **99**, 1742–1747 (2002).
[42] Alvarez-Lacalle, E., Dorow, B., Eckmann, J.-P., and Moses, E. Hierarchical structures induce long-range dynamical correlations in written texts. *Proc. Natl. Acad. Sci. USA* **103**, 7956–7961 (2006).
[43] Altmann, E. A., Cristadoro, G., and Esposti, M. D. On the origin of long-range correlations in texts. *Proc. Natl. Acad. Sci. USA* , In press (2012). doi:10.1073/pnas.1117723109
[44] Montemurro, M. A. and Pury, P. A. Long-range fractal correlations in literary corpora. *Fractals* **10**, 451–461 (2002).
[45] Corral, A., Ferrer i Cancho, R., and Díaz-Guilera, A. Universal complex structures in written language. *arXiv*:0901.2924 (2009).
[46] Altmann, E. G., Pierrehumbert, J. B., and Motter, A. E. Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE* **4**, e7678 (2009).

Acknowledgments

AMP acknowledges support from the IMT Lucca Foundation. JT, SH and HES acknowledge support from the DTRA, ONR, the European EPIWORK and LINC projects, and the Israel Science Foundation. MP acknowledges support from the Slovenian Research Agency.

Author Contributions

A. M. P., J. T., S. H., H. E. S. & M. P. designed research, performed research, wrote, reviewed and approved the manuscript. A. M. P. and J. T. performed the numerical and statistical analysis of the data.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

TABLE I: **Summary of the scaling exponents characterizing the Zipf law and the Heaps law.** The "unlimited lexicon" scaling exponent $\alpha_-(t)$ is calculated for $10^{-8} < f < 10^{-6}$ and the "kernel lexicon" exponent $\alpha_+(t)$ is calculated for $10^{-4} < f < 10^{-1}$ using the maximum likelihood estimator method for each year. The average and standard deviation ($\langle \cdots \rangle \pm \sigma$) listed are computed using the $\alpha_+(t)$ and $\alpha_-(t)$ values over the 209-year period 1800–2008 (except for Chinese, which corresponds to 1950–2008). The last column lists the Zipf scaling exponent calculated as $\zeta = 1/\Big(\langle \alpha_+ \rangle - 1\Big)$. To calculate $\sigma_r(t)$ (see Figs. 6 and 7) we use only the relatively common words that meet the criterion that their average word use $\langle f_i \rangle$ over the entire word history is larger than a threshold f_c, defined in the first column for each corpus. The b values shown are calculated using all words ($U_c = 0$).

Corpus	The Heaps and the Zipf law parameters				
(1-grams)	f_c	b	$\langle \alpha_- \rangle$	$\langle \alpha_+ \rangle$	ζ
Chinese	1×10^{-8}	0.77 ± 0.02	1.49 ± 0.15	1.91 ± 0.04	1.10 ± 0.05
English	5×10^{-8}	0.54 ± 0.01	1.73 ± 0.05	2.04 ± 0.06	0.96 ± 0.06
English fiction	1×10^{-7}	0.49 ± 0.01	1.68 ± 0.10	1.97 ± 0.04	1.03 ± 0.04
English GB	1×10^{-7}	0.44 ± 0.01	1.71 ± 0.07	2.02 ± 0.05	0.98 ± 0.05
English US	1×10^{-7}	0.51 ± 0.01	1.70 ± 0.08	2.03 ± 0.06	0.97 ± 0.06
English 1M	1×10^{-7}	0.53 ± 0.01	1.71 ± 0.04	2.04 ± 0.06	0.96 ± 0.06
French	1×10^{-7}	0.52 ± 0.01	1.69 ± 0.06	1.98 ± 0.04	1.02 ± 0.04
German	1×10^{-7}	0.60 ± 0.01	1.63 ± 0.16	2.02 ± 0.03	0.98 ± 0.03
Hebrew	5×10^{-7}	0.47 ± 0.01	1.34 ± 0.09	2.06 ± 0.05	0.94 ± 0.05
Russian	5×10^{-7}	0.65 ± 0.01	1.55 ± 0.17	2.04 ± 0.06	0.96 ± 0.06
Spanish	1×10^{-7}	0.51 ± 0.01	1.61 ± 0.15	2.07 ± 0.04	0.93 ± 0.04

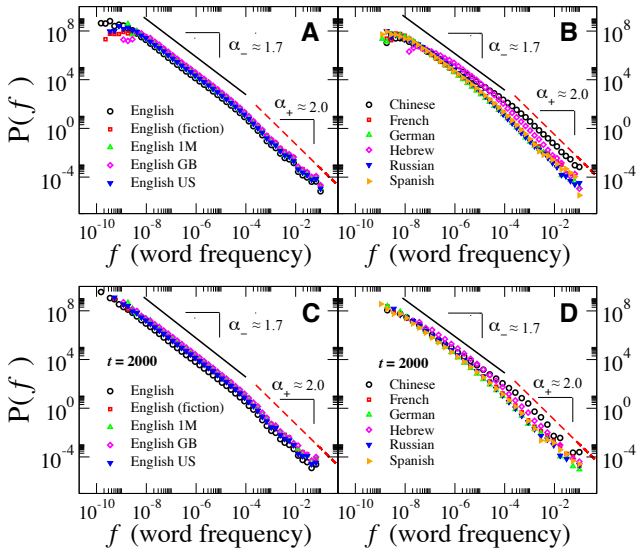

FIG. 1: **Two-regime scaling distribution of word frequency.** The kink in the probability density functions $P(f)$ occurs around $f_\times \approx 10^{-5}$ for each corpora analyzed (see legend). (A,B) Data from all years are aggregated into a single distribution. (C,D) $P(f)$ comprising data from only year $t = 2000$ providing evidence that the distribution is stable even over shorter time frames and likely emerges in corpora that are sufficiently large to be comprehensive of the language studied. For details concerning the scaling exponents we refer to Table I and the main text.

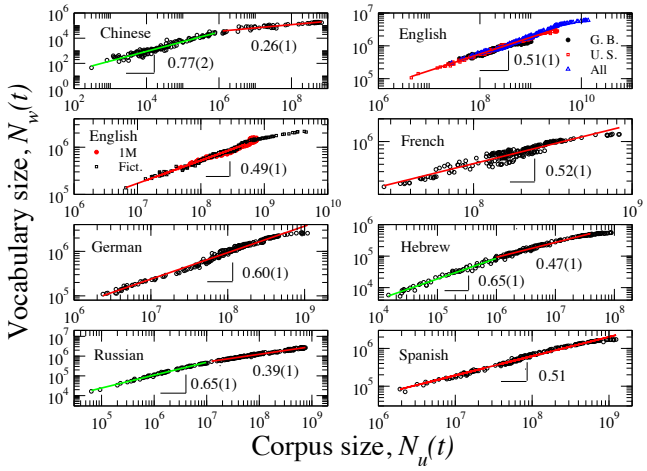

FIG. 2: **Increasing marginal returns of new words.** Scatter plots of the output corpora size N_u given the empirical vocabulary size N_w using data for the 209-year period 1800–2008. See Table I for all b values.

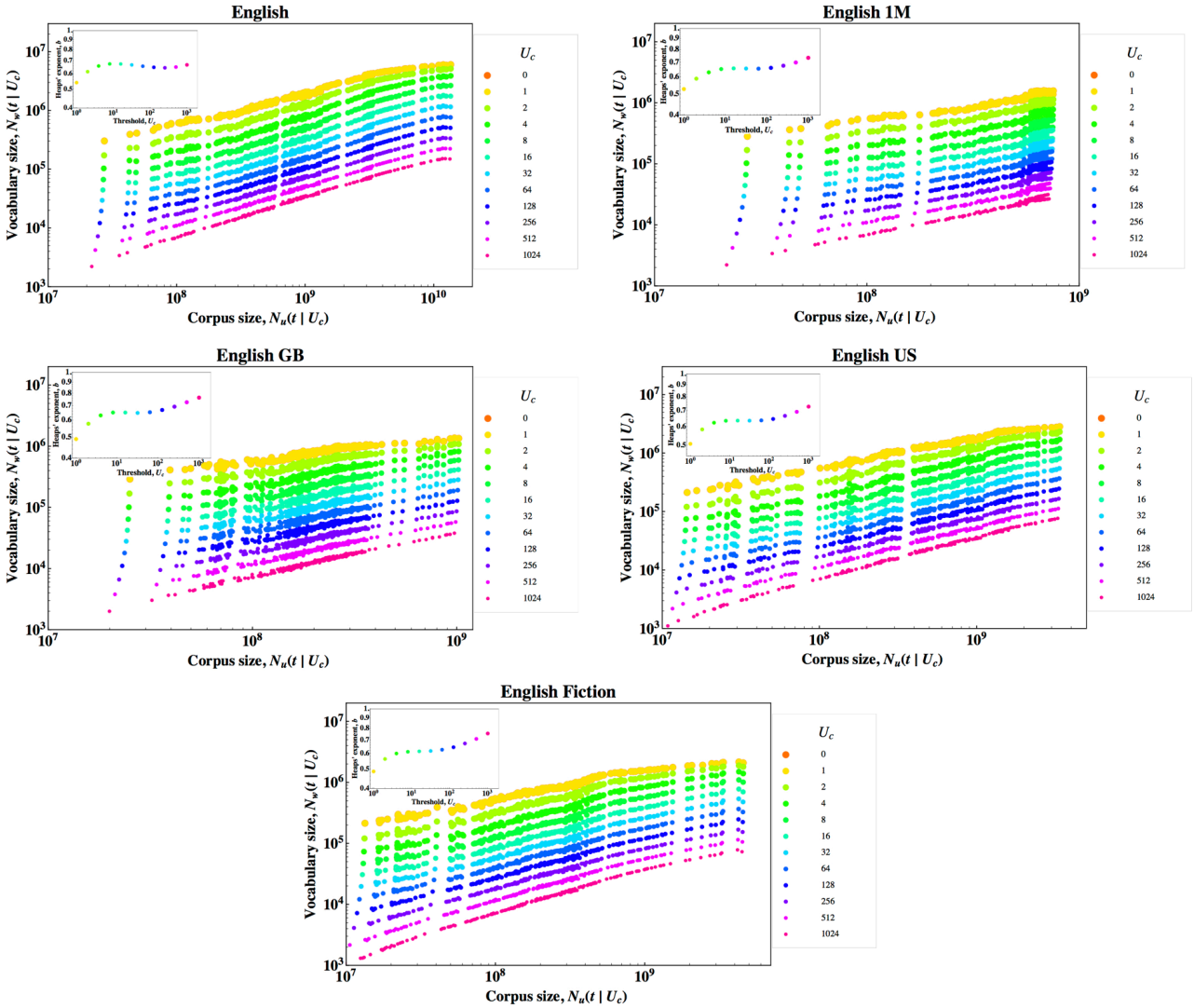

FIG. 3: **Allometric scaling of language.** Heaps law with varying inclusion for 5 English corpora. The Heaps scaling exponent b depends on the extent of the inclusion of the rarest words. Depicted are the scaling relations between the corpus size N_u and the vocabulary size N_w for the English corpora, obtained by using only words with $u_i(t) > U_c$. (Panel Inset) We find that b increases as we increasingly prune the corpora of extremely rare words, indicating the structural importance of the most frequent words which are used in the introduction of new and rare words.

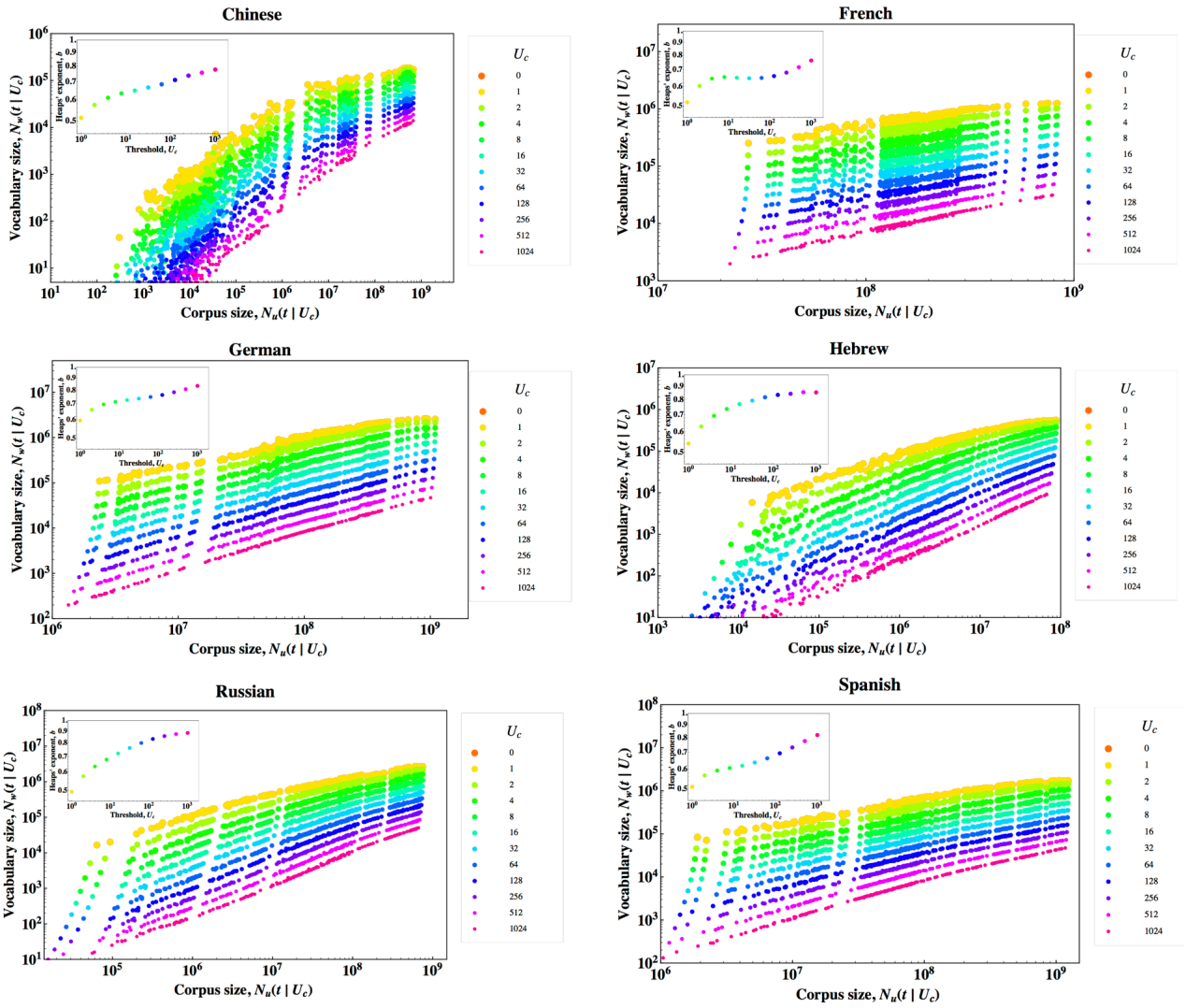

FIG. 4: **Allometric scaling of language.** Heaps law with variable inclusion for the Chinese, French, German, Hebrew, Russian, and Spanish corpora. The Heaps scaling exponent b depends on the extent of the inclusion of the rarest words. Depicted are the scaling relations between the corpus size N_u and the vocabulary size N_w for the English corpora, obtained by using only words with $u_i(t) > U_c$. (Panel Inset) We find that b increases as we prune the corpora of extremely rare words, indicating the structural importance of the most frequent words which are used more times per appearance.

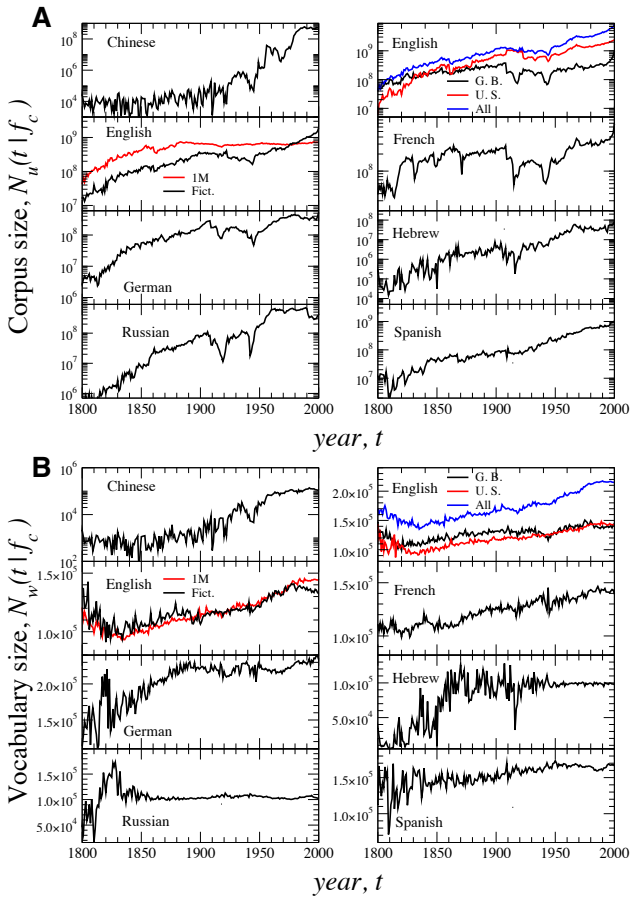

FIG. 5: **Literary productivity and vocabulary size in the examined dataset over the past two centuries.** (A) Total size of the different corpora $N_u(t|f_c)$ over time, calculated by using words that satisfy $f > f_c$, where f_c in Table I. (B) Size of the written vocabulary $N_w(t|f_c)$ over time, calculated under the same conditions as (A).

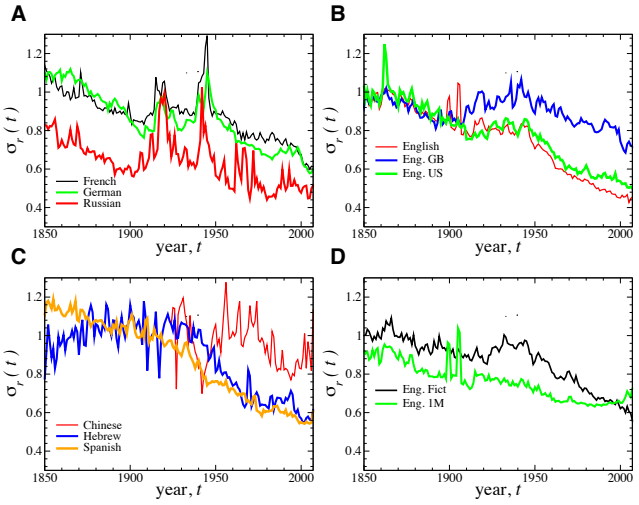

FIG. 6: **Non-stationarity in the characteristic growth fluctuation of word use.** The standard deviation $\sigma_r(t)$ of the logarithmic growth rate $r_i(t)$ is presented for all examined corpora. There is an overall decreasing trend arising from the increasing size of the corpora, as depicted in Fig. 5(A). On the other hand, the steady production of new words, as depicted in Fig. 5(B) counteracts this effect. We calculate $\sigma_r(t)$ using the relatively common words that meet the criterion that their average word use $\langle f_i \rangle$ over the entire word history is larger than a threshold f_c (see Table I).

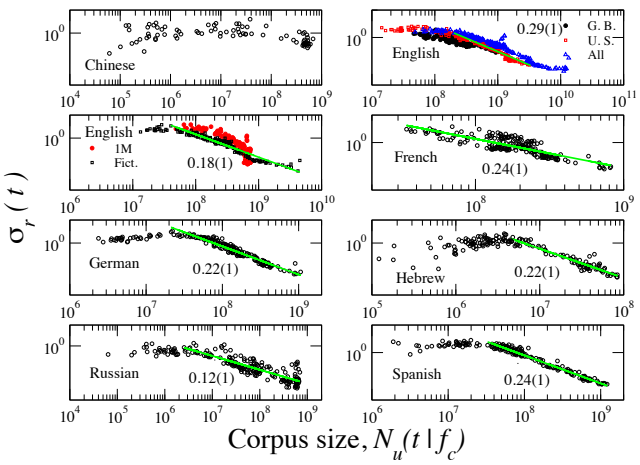

FIG. 7: **Growth fluctuation of word use scale with the size of the corpora.** Depicted is the quantitative relation in Eq.(9) between $\sigma_r(t)$ and the corpus size $N_u(t|f_c)$ obtained using f_c listed in Table I and with $\Delta t = 1$ year. Except for the Chinese language, all languages exhibit satisfactory scaling over several orders of magnitude in $N_u(t|f_c)$. We show the language-dependent scaling value $\beta \approx 0.12 - 0.29$ in each panel. For each language we show the value of the ordinary least squares best-fit β value with the standard error in parentheses.