# Zipf's law holds for phrases, not words

Jake Ryland Williams,[1, *] Paul R. Lessard,[2, †] Suma Desu,[3, ‡] Eric Clark,[1, §] James P. Bagrow,[4, 5, ¶] Christopher M. Danforth,[1, **] and Peter Sheridan Dodds[1, ††]

[1] *Department of Mathematics & Statistics, Vermont Complex Systems Center, Computational Story Lab, & the Vermont Advanced Computing Core, The University of Vermont, Burlington, VT 05401.*
[2] *Department of Mathematics, University of Colorado, Boulder CO, 80309*
[3] *Center for Computational Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139*
[4] *Computational Story Lab, Vermont Advanced Computing Core, & the Department of Mathematics and Statistics, University of Vermont, Burlington, VT, 05401*
[5] *Vermont Complex Systems Center, University of Vermont, Burlington, VT, 05401*

(Dated: June 23, 2014)

Over the last century, the elements of many disparate systems have been found to approximately follow Zipf's law—that element size is inversely proportional to element size rank [1, 2]—from city populations [2–4], to firm sizes [5], and family names [6]. But with Zipf's law being originally and most famously observed for word frequency [2], it is surprisingly limited in its applicability to human language, holding only over a few orders of magnitude before hitting a clear break in scaling [7–9]. Here, building on the simple observation that a mixture of words and phrases comprise coherent units of meaning in language, we show empirically that Zipf's law for English phrases extends over seven to nine orders of rank magnitude rather than typically two to three for words alone. In doing so, we develop a simple, principled, and scalable method of random phrase partitioning, which crucially opens up a rich frontier of rigorous text analysis via a rank ordering of mixed length phrases rather than words.

Starting with Mandelbrot's optimality argument [10], and the dynamically growing, rich-get-richer model of Simon [3], contentious debates over theoretical mechanisms leading to Zipf's law have continued until the present [11–14]. Persistent claims of uninteresting randomness underlying Zipf's law [11] have been successfully challenged [12], and in non-linguistic systems, good evidence supports Simon's model [3, 15, 16] which has been found to be the basis of scale-free networks [17, 18].

For language, the vast majority of arguments have focused on the frequency of an individual word which we contend here is fundamentally the wrong unit of analysis. Words are an evident building block of language, and we are naturally drawn to simple counting as a primary means of analysis (the earliest examples are Biblical corcordances, dating to the 13th Century). And while we have defined morphemes as the most basic meaningful 'atoms' of language, the meaningful 'molecules' of language are clearly a mixture of individual words and phrases. The identification of meaningful phrases, or multi-word expressions, in natural language poses one of the largest obstacles to accurate machine translation [19]. In reading the phrases 'New York City' or 'Star Wars', we effortlessly take them as irreducible constructions, different from the transparent sum of their parts. Indeed, it is only with some difficulty that we actively parse highly common phrases and consider their individuals words.

While partitioning a text into words is straightforward computationally, partitioning into meaningful phrases would appear to require a next level of sophistication requiring online human analysis. But in order to contend with the increasingly very large sizes and rapid delivery rates of important text corpora—such as news and social media—we are obliged to find a simple, necessarily linguistically naive, yet effective method.

A natural possibility is to in some way capitalize on $n$-grams, which are a now common and fast approach for partitioning a text, especially when the original text corpora is not open such as for the Google Books project [20]. Unfortunately, $n$-grams fail on several fronts: longer phrases are omitted with typically $n \leq 5$, and $n$-grams overlap, obscuring true partitioning. Consequently, and crucially, we are unable to properly assign frequency of usage weights to $n$-grams combined across all values of $n$, rendering ranking impossible.

Here, we introduce 'random partitioning', a method that is fast, intelligible, scalable, and sensibly preserves counts: i.e., the sum of the weights all partitioned phrases should be equal to the total number of words in a text. As we show, our method immediately yields the profound basic science result that phrases of mixed lengths, rather than individual words, obey Zipf's law, indicating the method can serve as a more profitable approach to general text analysis. To explore a lower level of language, we also use random partitioning to construct graphemes by breaking words into letter sequences. In the remainder of the paper, we first describe random partitioning and then present results for a range of texts.

To begin our random partitioning process, we break a given text $T$ into clauses, as demarcated by standard punctuation (other defensible schemes for obtaining clauses may also be used), and define the total weight of a given clause $\sigma$ as its word count, $n_\sigma$. We randomly partition $\sigma$ into phrases by cutting $\sigma$ at each of its $n_\sigma - 1$

Typeset by REVTEX

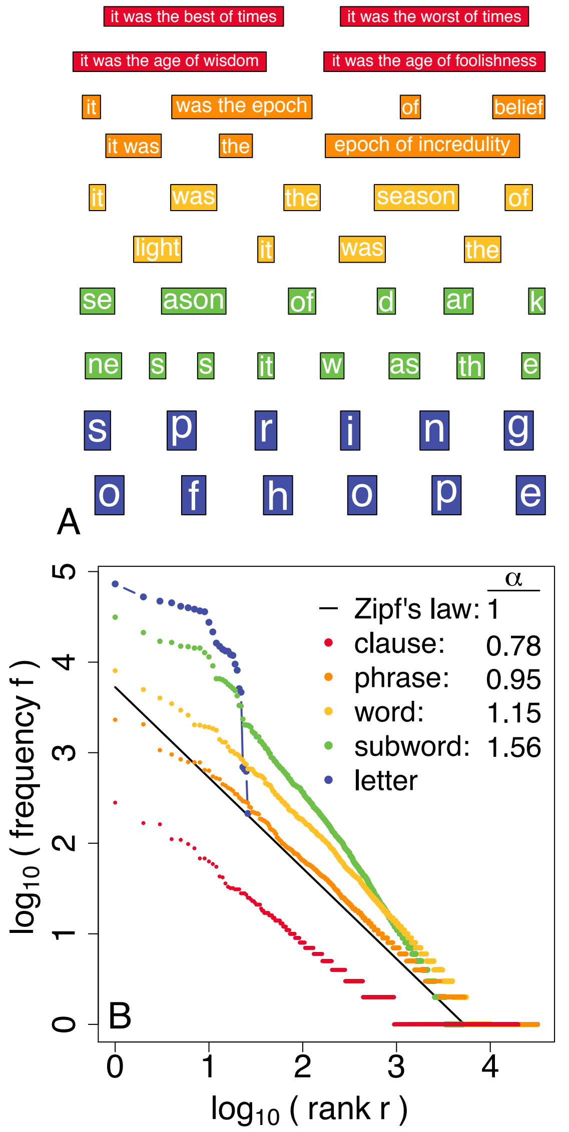

FIG. 1. **A.** Partition examples for the start of Charles Dickens's "Tale of Two Cities" at five distinct levels: clauses (red), pure random partitioning phrases ($q = \frac{1}{2}$, orange), words (yellow), pure random partitioning graphemes ($q = \frac{1}{2}$, green), and letters (blue). The specific phrases and graphemes shown are for one realization of pure random partitioning. **B.** Zipf distributions for the five kinds of partitions along with estimates of the Zipf exponent $\alpha$ when scaling is observed. See Ref. [21] and supplementary material for measurement details.

internal word-word boundaries with probability $q$. There are $2^{n_\sigma-1}$ possible partitions of $\sigma$ involving $\frac{1}{2}n_\sigma(n_\sigma+1)$ potential phrases. For each location-specific, contiguous phrase $\tau$, the probability a randomly chosen partition of $\sigma$ will include $\tau$ follows from successful cutting at the ends of $\tau$ and failures within, accommodating for $\tau$ reaching one or both ends of $\sigma$, i.e.,

$$P_q(\tau\,|\,\sigma) = q^{2-b}(1-q)^{n_\tau-1}, \tag{1}$$

where $b$ is the number of the clause's boundaries shared by $\tau$ and $\sigma$, and $n_\tau$ is the number of words in $\tau$. In summing over all clauses within a text $T$, we interpret the 'expected frequency' of $\tau$ within $T$, our generalization of standard word count, as

$$f_{q,\mathrm{exp}}(\tau\,|\,T) = \sum_{\sigma\in T} P_q(\tau\,|\,\sigma). \tag{2}$$

Departing from normal word counts, we may now have $f_{q,\mathrm{exp}} \ll 1$. When weighted by phrase length, these expected frequencies sum to $n_T$, the number of words in the text:

$$\sum_{\tau} n_\tau f_{q,\mathrm{exp}}(\tau\,|\,T) = n_T, \tag{3}$$

where $\tau$ ranges over all phrases appearing in $T$.

Here, we focus on three natural kinds of partitions: $q$=0: clauses are partitioned only as clauses themselves; $q$=$\frac{1}{2}$: what we call 'pure random partitioning'—all partitions of a clause are equally likely; $q$=1: clauses are partitioned into words.

In carrying out pure random partitioning ($q$=$\frac{1}{2}$), which we will show has the many desirable properties we seek, we are assuming all partitions are equally likely, reminiscent of equipartitioning used in statistical mechanics [22]. Extending the analogy, we can view $q$=0 as a zero temperature limit, and $q$=1 as an infinite temperature one.

For a simple example of the $q$=$\frac{1}{2}$ case, consider the punctuation-sensitive clause $\sigma$ = "let's eat Grandma". The four possible random partitions are "let's | eat | Grandma", "let's | eat Grandma", "let's eat | Grandma", and "let's eat Grandma", each occurring with probability $\frac{1}{4}$. These partitions give rise to six phrases which, along with their weighted effective frequencies, are: "let's" ($1\times\frac{1}{2}$), "eat" ($1\times\frac{1}{4}$), "Grandma" ($1\times\frac{1}{2}$), "let's eat" ($2\times\frac{1}{4}$), "eat Grandma" ($2\times\frac{1}{4}$), and "let's eat Grandma" ($3\times\frac{1}{4}$). As required, the sum of all weighted expected frequencies match the clause's length, $n_\sigma = 3$, to which each individual word contributes a unit weight. As an anchor for $f_{\frac{1}{2},\mathrm{exp}}$, we note that words that appear once within a text—hapax legomena—will have $f_{q,\mathrm{exp}} \in \{\frac{1}{4}, \frac{1}{2}, 1\}$ (depending on clause boundaries), on the order of 1 as per standard word partitioning.

We will first examine the results of applying these partition processes once ('one-off'), at the level of clauses and words, and then move to considering ensembles of all possible partitions.

In Fig. 1, we show results for one-off partitions of Charles Dickens' "Tale of Two Cities" and summarize findings for a collection of other works of literature [23]. In Fig. 1A, we give example partitions at the scales of clauses (red), pure random partition phrases (orange), words (yellow), pure random partition graphemes (green), and letters (blue). In Fig. 1B, we show Zipf distributions for all five partitioning scales.

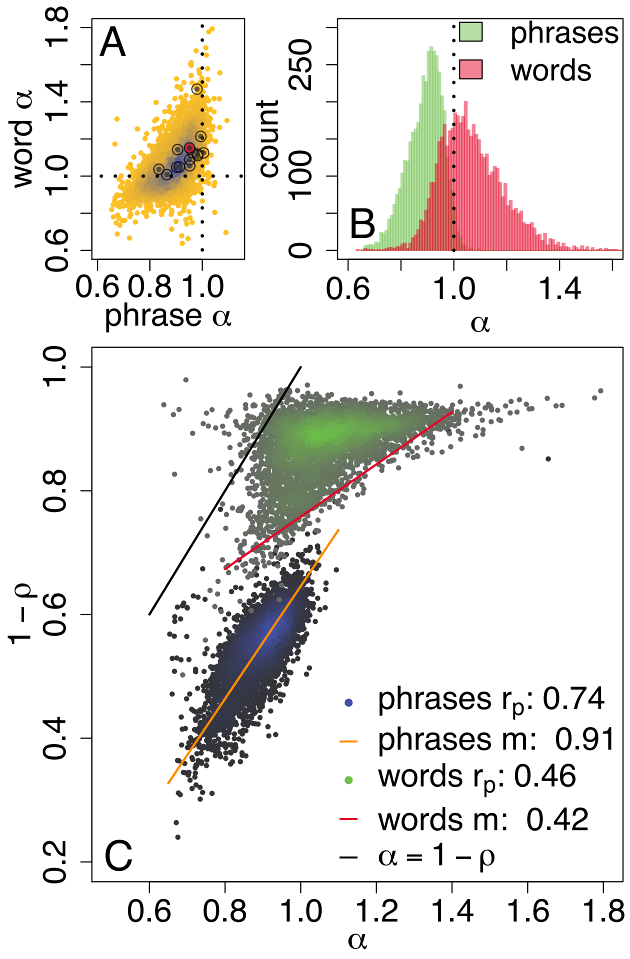

FIG. 2. **A.** Density plot showing $\alpha$ for phrase and word Zipf distributions for around 4000 works of literature. We indicate "Tale of Two Cities" by the red circle, and with black circles, we represent measurements for 14 other works of literature analyzed further in the supplementary material. **B.** Histograms of $\alpha$ for the same set of books (marginal distributions for **A**). **C.** Test of Simon's model analytical connection $\alpha = 1 - \rho$, where $\alpha$ is the Zipf exponent and $\rho$ is the rate at which new terms (e.g., graphemes, words, phrases) are introduced throughout a text. We estimate $\rho$ as the number of different words normalized by the total word volume. For both words and phrases, we compute linear fits using Reduced Major Axis (RMA) regression [24] to obtain slope $m$, along with the Pearson correlation coefficient $r_{\mathrm{p}}$. Words (green) do not exhibit a simple linear relationship whereas phrases do (blue), albeit clearly below the $\alpha = 1 - \rho$ line in black.

We see that clauses and pure random partitioning phrases both adhere well to the pure form of $f \propto r^{-\alpha}$ where $r$ is rank. For clauses, $q$=0, we find $\alpha \simeq 0.78$ and for $q=\frac{1}{2}$, $\alpha \simeq 0.98$ (see supplementary material for measurement details and for examples of other works of literature). The quality of scaling degrades as we move down to words and graphemes with the appearance of scaling breaks. Scaling vanishes entirely at the level of letters.

In Fig. 2A, we compare the Zipf exponent $\alpha$ for words and $q=\frac{1}{2}$ partitioning for around 4000 works of literature. We plot the corresponding marginal distributions in Fig. 2B. We see that clearly $\alpha \lesssim 1$ for $q=\frac{1}{2}$ phrases, while for words, there is a strong positive skew with the majority of values of $\alpha > 1$. These steep scalings for words (and graphemes), $\alpha > 1$, are not dynamically accessible for Simon's model [13].

Leaving aside this non-physicality of Zipf distributions for words and concerns about breaks in scaling, we recall that Simon's model connects the rate at which new terms are introduced, $\rho$ to $\alpha$ as simply $1 - \rho = \alpha$ [3]. We ask how well works of literature conform to this connection in Fig. 2C. We find that words (green dots) do not demonstrate any semblance of a linear relationship, whereas phrases (blue dots) exhibit a clear, if approximate linear connection between $1 - \rho$ and $\alpha$.

Despite this linearity, we see a pure Simon model fails to accurately predict the phrase distribution exponent $\alpha$. This is not surprising, as when $\rho \to 0$, an immediate adherence to the rich-get-richer mechanism produces a transient behavior in which the first few (largest-count) word varieties exist out of proportion to the eventual scaling. Because a pure Simon distribution preserves $\alpha = 1 - \rho$, we expect that a true, power-law consistently makes the underestimate $\alpha > 1 - \rho$.

Inspired by our results for one-off partitions of texts, we now consider ensembles of pure random partitioning for larger texts. In Fig. 3, we show Zipf distributions of expected frequency $f_{\frac{1}{2},\mathrm{exp}}$ for $q=\frac{1}{2}$ phrases for four large-scale corpora: English Wikipedia (EW), the New York Times (NYT), Twitter (TW), and music lyrics (ML). We color the main curves according to the length of a phrase for each rank. We include word-level Zipf distributions ($q$=1) for each text along with idealized Zipf distributions for reference ($\alpha$=1).

We observe an approximate $\alpha$=1 scaling for the expected frequency of phrases extending out over a remarkable 7–9 orders of magnitude. (Others have noted similar results by simply combining frequency distributions of n-grams, though such approaches are unprincipled as they overcount words [25].) In terms of $f_{\frac{1}{2},\mathrm{exp}}$, the scaling range persists down to $10^{-2}$, beyond the happax legomena mark of around $10^0$. Such robust scaling is in stark contrast to the very limited scaling of word frequencies (gray curves). For pure word partitioning, $f_{1,\mathrm{exp}}$, we see two rough scaling regimes exhibited by each corpus, with scaling at best extending over three orders of magnitude, and typically only two.

For all four corpora, random partitioning gives rise to a gradual interweaving of different length phrases when moving up through rank $r$. Single words remain the most frequent (purple), typically beginning to blend with two word phrases (blue) by rank $r = 100$. After the appearance of phrases of length around 10–20, depend-

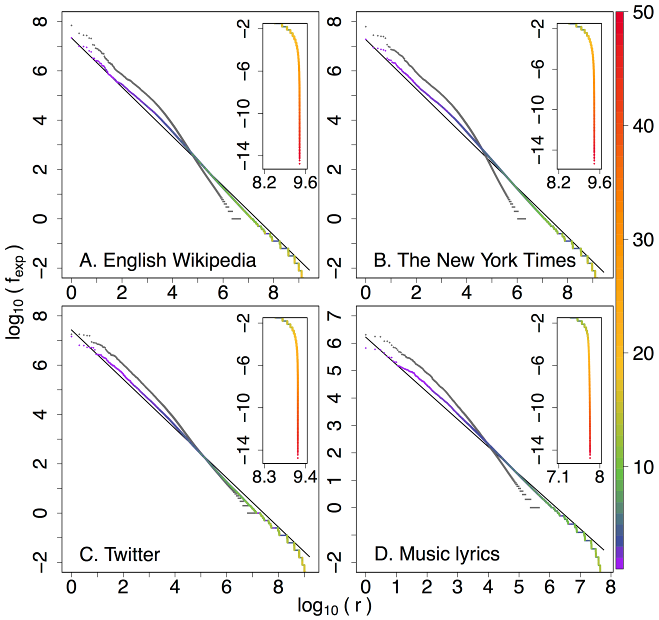

FIG. 3. Random partitioning distributions for the four large corpora: (A) Wikipedia (2009); (B) The New York Times (1987–2007); (C) Twitter (2009); and (D) Music Lyrics (1960–2007). Top right insets show the long tails of random partitioning distributions, and the colors represent phrase length as indicated by the color bar. See main text and Tabs. S1–S4 for example phrases.

ing on the corpus, we see the phrase rank distributions fall off sharply, due to long clauses that share little overlap (upper right insets).

In the supplementary material, we provide structured tables of example phrases extracted by pure random partitioning for all four corpora (Tabs. S1–S4), along with complete phrase data sets. As with standard $n$-grams, the texture of each corpora is quickly revealed by examining phrases of length 3, 4, and 5. For example, the second most common phrases of length 5 for the four corpora are routinized phrases: “the average household size was” (EW), “because of an editing error” (NYT), “i uploaded a youtube video” (TW), and “na na na na na” (ML). By design, random partitioning allows us to quantitatively compare and sort phrases of different lengths. For music lyrics, “la la la la la” has an expected frequency similar to “i don’t know why”, “just want to”, “we’ll have”, and “whatchu” (see Tab. S4), while for the New York Times, “the new york stock exchange” is comparable to “believed to have” (see Tab. S2).

The phrases and their effective frequencies produced by our pure random partitioning method may serve as input to a range of higher order analyses, For example, information theoretic work may be readily carried out, context models may be built around phrase adjacency using insertion and deletion, and specific, sentence-level partitions may be realized from probabilistic partitions.

While we expect that other principled, more sophisticated approaches to partitioning texts into rankable mixed phrases should produce Zipf’s law spanning similar or more orders of magnitude in rank, we believe random partitioning—through its transparency, simplicity, and scalability—will prove to be a powerful method for exploring and understanding large-scale texts.

To conclude, our results reclaim Zipf’s law for language ($\alpha$=1), demonstrating that the unit of analysis must be phrases and not words, and call in to question all past and present word-based studies. In considering entropy, our findings moreover imply that phrases of natural language are capable of describing an arbitrarily large concept space with unbounded information content.

The authors are grateful for the computational resources provided by the Vermont Advanced Computing Core which was supported by NASA (NNX 08A096G). CMD was supported by NSF grant DMS-0940271; PSD was supported by NSF CAREER Award #0846668.

[*] jake.williams@uvm.edu
[†] paul.lessard@boulder.edu
[‡] sdesu@mit.edu
[§] eric.clark@uvm.edu
[¶] james.bagrow@uvm.edu
[**] chris.danforth@uvm.edu
[††] peter.dodds@uvm.edu

[1] G. K. Zipf, *The Psycho-Biology of Language*, patterns (Houghton-Mifflin, New York, NY, 1935).

[2] G. K. Zipf, *Human Behaviour and the Principle of Least-Effort*, patterns (Addison-Wesley, Cambridge, MA, 1949).

[3] H. A. Simon, Biometrika **42**, 425 (1955).

[4] M. Batty, Science Magazine **319**, 769 (2008).

[5] R. Axtell, Science **293**, 1818 (2001).

[6] D. H. Zanette and S. C. Manrubia, Physica A **295**, 1 (2001).

[7] R. Ferrer i Cancho, C. Janssen, and R. V. Solé, Phys. Rev. E **64**, 046119 (2001).

[8] L. Q. Ha, P. Hanna, J. Ming, and F. J. Smith, Artif. Intell. Rev. **32**, 101 (2009).

[9] M. Gerlach and E. G. Altmann, Phys. Rev. X **3**, 021006 (2013).

[10] B. B. Mandelbrot, in *Communication Theory*, edited by W. Jackson (Butterworth, Woburn, MA, 1953) pp. 486–502.

[11] G. A. Miller, American Journal of Psychology **70**, 311 (1957).

[12] R. Ferrer-i Cancho and B. Elvevåg, PLoS ONE **5**, e9411 (2010).

[13] R. M. D’Souza, C. Borgs, J. T. Chayes, N. Berger, and R. D. Kleinberg, Proc. Natl. Acad. Sci. **104**, 6112 (2007).

[14] B. Coromina-Murtra and R. Solé, Phsyical Revew E **82**, 011102 (2010).

[15] S. Bornholdt and H. Ebel, Phys. Rev. E **64**, 035104(R) (2001).

[16] T. Maillart, D. Sornette, S. Spaeth, and G. von Krogh,

Phys. Rev. Lett. **101**, 218701 (2008).

[17] D. J. d. S. Price, J. Amer. Soc. Inform. Sci. **27**, 292 (1976).

[18] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[19] I. A. Sag, T. Baldwin, F. Bond, A. A. Copestake, and D. Flickinger, in *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02 (Springer-Verlag, London, UK, 2002) pp. 1–15.

[20] Google Labs ngram viewer. Availabe at `http://ngrams.googlelabs.com/`.

[21] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Review **51**, 661 (2009).

[22] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group*, general, Frontiers in Physics, Vol. 85 (Addison-Wesley, Reading, Massachusetts, 1992).

[23] The Gutenberg Project: `http://www.gutenberg.org`; accessed November 15, 2013.

[24] J. M. V. Rayner, J. Zool. Lond. (A) **206**, 415 (1985).

[25] L. Q. Ha, E. I. Sicilia-Garcia, J. Ming, and F. J. Smith, in *Proceedings of the 19th International Conference on Computational Linguistics (COLING)* (2002) pp. 315–320.