ADDENDUM

Addendum: another explanation about the bounds of the Taylor exponent

Kumiko Tanaka-Ishii and Tatsuru Kobayashi
Research Center for Advanced Science and Technology, The University of Tokyo 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan
E-mail: kumiko@cl.rcast.u-tokyo.ac.jp and kobayashi@cl.rcast.u-tokyo.ac.jp

Taylor’s law describes the fluctuation characteristics underlying a complex system in which the variance of an event within a time span grows by a power law with respect to the mean. The previous paper, Taylor’s Law for Linguistic Sequences and Random Walk Models (Tanaka-Ishii and Kobayashi 2018), appeared in Journal of Physics Communications and described a new way to apply Taylor analysis to texts. The method was applied to over 1100 texts across 14 languages. The results showed how the Taylor exponents of natural-language written texts were consistently around 0.58, thus being universal.

Experimentally, the Taylor exponent $\alpha$ is known to take a value within the range of $0.5 \leq \alpha \leq 1.0$ across a wide variety of domains, including finance, meteorology, agriculture, and biology. The previous paper shows how this is the case for language.

The Taylor exponent is analytically proven to be 0.5 for an independent and identically distributed (i.i.d.) process. The paper also shows a case when 1.0 is reached. This Addendum provides two additional cases of rare word alignment for $\alpha = 0.5$ and $\alpha = 1.0$. These cases provide an understanding to interpret the value of the exponent of a real text.

Consider dividing a text of length $N$ into $Q$ segments of length $\Delta t$, i.e., $N = Q \Delta t$. Suppose that $Q$ is sufficiently large.

First of all, if a word only appears once in the entire text, then $\mu_1$ and $\sigma_1$ are calculated as follows.

$$\mu_1 = \frac{1}{Q},$$

$$\sigma_1 = \sqrt{\frac{1}{Q} \left( Q - 1 \left( \frac{1}{Q} \right)^2 + \left( 1 - \frac{1}{Q} \right)^2 \right)} = \frac{\sqrt{Q - 1}}{Q}.$$

For words that appear $n (\ll Q)$ times in a text, there are two extreme possibilities: when the $n$ words are all in the same segment, and on the other hand, when the $n$ words are all in different segments.

When the $n$ words all appear in the same segment, then $\sigma$ becomes the largest for this case, as follows.

$$\mu_n = \frac{n}{Q},$$

$$\sigma_{n,\text{max}} = \sqrt{\frac{1}{Q} \left( Q - 1 \left( \frac{n}{Q} \right)^2 + \left( n - \frac{n}{Q} \right)^2 \right)} = \frac{n\sqrt{Q - 1}}{Q}.$$

Because the extreme of such a case occurs when $n = 1$, $\alpha$ is 1, as follows.

$$\alpha_{\text{max}} = \frac{\log \sigma_{n,\text{max}} - \log \sigma_1}{\log \mu_n - \log \mu_1} = 1.$$

© 2019 The Author(s). Published by IOP Publishing Ltd
On the other hand, when the $n$ words all appear in different segments, then $\sigma$ becomes the smallest, as follows.

$$\mu_n = \frac{n}{Q},$$

$$\sigma_{n,\min} = \sqrt{\frac{1}{Q} \left( (Q - n) \left( \frac{n}{Q} \right)^2 + n \left( 1 - \frac{n}{Q} \right)^2 \right)}$$

$$= \frac{\sqrt{n(Q - n)}}{Q}.$$  

Because the extreme of such a case occurs when $n = 1$, $\alpha$ is 0.5, as follows.

$$\alpha_{\min} = \frac{\log \sigma_{n,\min} - \log \sigma_1}{\log \mu_n - \log \mu_1}$$

$$= \frac{\log \sqrt{\frac{Q - n}{Q - 1}}}{\log n}$$

$$\approx 0.5.$$

**ORCID iDs**

Kumiko Tanaka-Ishii @ https://orcid.org/0000-0003-1752-3951

**References**

Tanaka-Ishii K and Kobayashi T 2018 Taylor’s law for linguistic sequences and random walk models *Journal of Physics Communications* 2 2399–6528
Taylor’s law for linguistic sequences and random walk models

Kumiko Tanaka-Ishii and Tatsuru Kobayashi
Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan

E-mail: kumiko@cl.rcast.u-tokyo.ac.jp and kobayashi@cl.rcast.u-tokyo.ac.jp

Keywords: Taylor’s law, fluctuation analysis, random walk, Barabási-Albert network

Abstract

Taylor’s law describes the fluctuation characteristics underlying a complex system in which the variance of an event within a time span grows by a power law with respect to the mean. Although Taylor’s law has been applied in many natural and social systems, its application for language has been scarce. This article describes a new, natural way to apply Taylor analysis to texts. The method was applied to over 1100 texts across 14 languages and showed how the Taylor exponents of natural language written texts were consistently around 0.58, thus being universal. The exponents were also evaluated for other language-related data, such as speech corpora (0.63 for adult speech, 0.68 for child-directed speech), programming language sources (0.79), and music (0.79). The results show how the Taylor exponent serves to quantify the fundamental structural complexity underlying linguistic time series. To explain the nature of natural language sequences possessing such different degrees of fluctuation, we investigated various mathematical models that could produce a Taylor exponent similar to that of real data. While the majority of previous probabilistic sequential models could not produce a Taylor exponent larger than 0.50, the same as in the independent and identically distributed (i.i.d.) case, random walk sequences on complex networks could produce fluctuation. We show that among various possibilities, random walks on a Barabási-Albert (BA) graph with small mean degree could fulfill the scaling properties of Zipf’s law and the long-range correlation, in addition to having a Taylor’s law exponent larger than 0.5, thus giving a new perspective to reconsider the nature of language.

1. Introduction

Good quantification of the degree of fluctuation underlying human texts has been an issue in the statistical physics domain. Previous studies used analysis methods based on either the long-range correlation (LRC) (Altmann et al. 2009, Tanaka-Ishii and Bunde 2016, Lin and Tegmark 2016) or the variance (Ebeling and Pöeschel 1994). All these studies showed that language has power-law behavior due to underlying fluctuation, but none of them could quantify the exact degree or universality underlying language. One difficulty lies in the fact that language is not numerical, so the methods proposed in statistical mechanics, which were based on numerical data, require transformation of a linguistic sequence into a numerical sequence. One method proposed by Lin and Tegmark (2016) used mutual information and applies to text without such transformation, but it could not quantify the LRC underlying language (Takahashi and Tanaka-Ishii 2017). Hence, there is not only the problem of discovering fluctuation but also the question of the analysis method, i.e. whether it can quantify the underlying phenomena.

Taylor’s law characterizes how the standard deviation of the number of events for a given time and space grows with respect to the mean. Since the pioneering studies of this concept (Smith 1938, Taylor 1961), a substantial number of studies have applied Taylor’s law across various domains, including ecology, life science, physics, finance, and human dynamics, as well summarized in (Eisler et al. 2007). Despite such diverse application across domains, however, there has been little analysis based on Taylor’s law in studying natural language. The only application, to the best of our knowledge, was shown in (Gerlach and Altmann 2014), but they measured the mean and variance of the vocabulary size within a document. This approach essentially differs from the original concept of Taylor analysis, which counts the number of events of every kind and not the
number of kinds of events. Because of this, the theoretical background of Taylor’s law, such as that presented in (Eisler et al 2007), cannot be applied to interpret their results.

In contrast to that previous report, the present paper applies Taylor’s law in a novel way by examining the standard deviation of the number of occurrences of every word kind with respect to its mean within a given text length. While the Taylor exponent is analytically proven to be 0.5 for an independent and identically distributed (i.i.d.) process, our results show that the exponent is universally distributed around 0.58 for natural language written texts. Moreover, in verification across various kinds of human linguistic-related sequences, including the speech of infants and adults, programming language code, and music, the exponent could distinguish the kind of data. Previous versions of some of the content reported with this regard appeared originally in the proceedings of the Annual Conference for Computational Linguistics (Kobayashi and Tanaka-Ishii 2018). The content was updated to include results with additional datasets.

To understand the nature of these human linguistic sequences, we reconsidered various mathematical models that can produce a Taylor exponent larger than 0.5. While all the previous sequential models, including Markov and Simon models (Simon 1955), could not produce a Taylor exponent close to that of human language, some random walk models on a complex network graph could. Hence, we present the conditions for such models to fulfill the statistical laws of natural language.

2. The method

Given a set of elements $W$ (words), let $X = X_1, X_2, \ldots, X_N$ be a discrete time series of length $N$, where $X_j \in W$ for all $i = 1, 2, \ldots, N$, i.e. each $X_j$ represents a word. For a given segment length $\Delta t \in \mathbb{N}$ (a positive integer), a data sample $X$ is segmented by the length $\Delta t$. The number of occurrences of a specific word $w_k \in W$ is counted for every segment, and the mean $\mu_k$ and standard deviation $\sigma_k$ across segments are obtained. Doing this for all word kinds $w_0, \ldots, w_{|W|} \in W$ gives the distribution of $\sigma$ with respect to $\mu$, where $|W|$ indicates the number of elements of the set $W$. Following a previous work (Eisler et al 2007), in this article Taylor’s law is defined to hold when $\mu$ and $\sigma$ are correlated by a power law in the following way:

$$\sigma \propto \mu^\alpha.$$  \hspace{1cm} (1)

Experimentally, the Taylor exponent $\alpha$ is known to take a value within the range of $0.5 \leq \alpha \leq 1.0$ across a wide variety of domains as reported in (Eisler et al 2007), including finance, meteorology, agriculture, and biology.

Mathematically, it is analytically proven that $\alpha = 0.5$ for an i.i.d process, and thus obviously for a Poisson process. On the other hand, one case with $\alpha = 1.0$ occurs when all segments always contain the same proportions of the elements of $W$. For example, suppose that $W = \{a, b\}$. If $b$ always occurs twice as often as $a$ in all segments (e.g., three $a$ and six $b$ in one segment, two $a$ and four $b$ in another, etc.), then both the mean and standard deviation for $b$ are twice those for $a$, so the exponent is 1.0.

In a real text, this cannot occur for all $W$, so $\alpha < 1.0$ for natural language text. Nevertheless, for a subset of words in $W$, this could happen, especially for a regular grammatical sequence. For instance, consider a programming statement: while (i < 1000) do -. Here, the words while and do always occur once in this type of statement, whereas i always occurs twice. This example shows that the exponent indicates how consistently words depend on each other in $W$, i.e. how words co-occur systematically in a coherent manner, thus suggesting that the Taylor exponent is partly related to grammaticality.

To measure the Taylor exponent $\alpha$, the mean and standard deviation were computed for every word kind and then plotted in log-log coordinates. The number of points in this work was thus the number of different word kinds. We fitted the points to a linear function in log-log coordinates by the least-squares method and then estimated the Taylor exponent. The coefficient $\hat{\alpha}$ and exponent $\hat{\alpha}$ were then estimated as the following:

$$\hat{\alpha} = \frac{1}{|W|} \sum_{k=1}^{|W|} \left( \log \sigma_k - \log \mu_k \right)^2.$$  \hspace{1cm} (2)

This fit function could be a problem depending on the distribution of errors between the data points and the regression line. As seen later, the error distribution seems to differ with the kind of data: for a random source the error seems Gaussian, and so the above formula is relevant. On the other hand, for real data the distribution is biased. Changing the fit function according to the data source, however, would cause other essential problems for fair comparison. Here, because most empirical works on Taylor’s law used least-squares regression as reported in (Cohen and Xu 2015), this work also uses the above scheme, with the error defined as $e(\hat{\alpha}, \hat{\alpha})$.

2 In this work, words are not lemmatized, e.g., ‘say,’ ‘said,’ and ‘says’ are all considered different words. We chose this approach because the Taylor exponent considers systematic co-occurrence of words, and idiomatic phrases should thus be considered in their original forms.
3. Data

Our data consisted of natural language texts and language-related sequences. The details of the data are included in appendix A and can be roughly summarized as follows. The natural language written texts consisted of 1142 single-author long texts in various languages, taken from Project Gutenberg and other sources (long, single-author texts), and large-scale newspaper data in three languages (multiple-author texts). We also assembled other types of linguistic-related sequences. The enwiki8 100-MB dump dataset is tagged Wikipedia data. We also assembled speech data from both adults and children: the spoken record of Japan’s National Diet (political content), and the 10 longest child-directed speech utterances in CHILDES (Child Language Data Exchange System, a well-studied children’s speech archive), in 10 different languages. We also assembled four computer program sources and 12 music performance sources and preprocessed them to exclude annotations. For a reference, we took the text of Moby Dick and generated 10 different shuffled samples.

4. Results

Figure 1 shows a Taylor analysis of the distribution of words in Moby Dick. The horizontal axis indicates the mean frequency of a specific word within a segment of $\Delta t = 5620$ words, whereas the vertical axis indicates the standard deviation. The choice of $\Delta t$ is arbitrary but has to be sufficiently smaller than the document length. Among different values of $\Delta t$ taken from logarithmic bins, we adopted a maximum $\Delta t$ that could apply to all documents that we used. This resulted in a choice of $\Delta t$ around 5000 words, specifically $\Delta t \approx 10^3.75$.

In the figure, the first graph shows the complete Taylor’s law results for Moby Dick. The points were distributed around the regression line in a log-log graph, thus showing the power-law tendency. The exponent was $\alpha = 0.57$, with the standard deviation error of the estimate being $\pm 0.0009$. The tendency of the points seemed to change around $\mu = 1.0$, which might seem a crossover, but this is doubtful according to the rationale shown by the second and third graphs in figure 1. These graphs show the regression for only the ranges of $\mu < 1.0$ and $\mu \geq 1.0$, respectively. The $\alpha$ value did increase when $\mu$ got larger, but very slightly, suggesting that the $\mu - \sigma$ plot did form almost a straight power law.

As explained previously, the Taylor exponent indicates the degree of consistent co-occurrence among words. The value of 0.57 obtained here suggests that the words of natural language texts are not so strongly or consistently coherent with respect to each other. Nevertheless, the value is well above 0.5.

Although the overall global tendency in figure 1 followed a power law, many points deviated significantly from the regression line. The words with the greatest fluctuation were often keywords. For example, among words in Moby Dick with large $\mu$, those with the largest $\sigma$ included whale, captain, and Ahab, among the book’s most important keywords, whereas those with the smallest $\sigma$ included functional words such as to, that, and all.
The same graphs for other natural language texts (two literary texts in Chinese and French, two newspapers in English and Chinese) are shown in appendix B figure B1. As seen in those graphs, the natural language written texts had almost the same exponent of around 0.58. Analysis of different kinds of data showed how the Taylor exponent differed according to the data source. Another set of graphs is shown in figure B2 for samples from enwiki8 (tagged Wikipedia), the child-directed speech of Thomas (taken from CHILDES), a programming language dataset, and a music source. The distributions appeared different from those of the natural language written texts, and the exponents were significantly larger. This suggests that those datasets contained expressions with fixed forms much more frequently than did the natural language texts.

Figure 2 summarizes the overall picture among all data sources. Here, each plot represents one data source (text), and the median and quantiles of the Taylor exponents were calculated for the different kinds of data listed in table A1 of appendix A. The first plot shows the result for the shuffled data, with an exponent of 0.50. We will consider these randomized results more in the next section.

The remaining plots show the results for real data. For all the real data, not a single sample gave an exponent as low as 0.5. The exponents for the texts from Project Gutenberg ranged from 0.53 to 0.68. Appendix C includes a histogram (figure C1) of these texts with respect to the value of $\alpha$. The number of texts decreased significantly at a value of 0.63, showing that the distribution of the Taylor exponent was rather tight. The kinds of texts in Project Gutenberg at the upper limit of the exponent included structured texts of fixed style, such as lists of histories, and Bibles. As described in more detail in appendix C, a non-parametric test indicated that different languages were not distinguishable by the value of $\alpha$.

Turning to the last five columns of figure 2, representing enwiki8, speech data (the Diet’s spoken record and child-directed speech), programming language, and music data, the Taylor exponents clearly differed from those of the natural language written texts. The enwiki8 dataset showed a higher exponent of 0.63, because of the wiki tags. For the Diet spoken record, another histogram in appendix C (figure C2) shows how tightly the points were located around the mean. Most importantly, the speech data showed higher exponents than did the written texts, with 0.63 for the Diet record and 0.68 for the child-directed speech. Finally, the programming and musical sources produced exponents near 0.79, higher than for any of the natural language data.Overall, the results show the possibility of applying the Taylor exponent to quantify the structural complexity underlying language data. Grammatical complexity was formalized by Chomsky via his hierarchy (Chomsky 1956), which describes grammar through rewriting rules. The constraints placed on the rules distinguish four levels of grammar: regular, context-free, context-sensitive, and phrase structure. As indicated in (Badii and Politi 1997), however, this does not quantify complexity on a continuous scale. For example, we might want to quantify the complexity of child-directed speech as compared to that of adults, and this could be addressed in only a limited way through the Chomsky hierarchy. Another point is that the hierarchy is sentence-based and does not consider fluctuation in the kinds of words appearing.

The Taylor exponent depended on the two parameters of the experiments, namely, the segment size and data size. First, the exponent differed according to $\Delta t$, as shown in figure 3. For each kind of data shown in figure 2, the mean exponent was plotted for various $\Delta t$. As reported in (Eisler et al 2007), the exponent is known to grow...
when the segment size gets larger. The reason is that words occur in a bursty, clustered manner at all $\Delta t$ length scales: no matter how large the segment size becomes, a segment can include very many or few instances of a given word, leading to larger variance growth. This phenomenon suggests how word co-occurrences in natural language are self-similar across $\Delta t$.

The Taylor exponent was initially 0.5 when the segment size was very small. This can be analytically explained as follows (Eisler et al 2007). Consider the case of $\Delta t = 1$. Let $n$ be the frequency of a particular word in a segment. We then have $\langle n \rangle \ll 1.0$, because the possibility of a specific word appearing in a segment becomes very small. Because $\langle n^2 \rangle \approx 0$, $\sigma^2 = \langle n^2 \rangle - \langle n \rangle^2 \approx \langle n^2 \rangle$. Then, because $n = 1$ or 0 (with $\Delta t = 1$), $\langle n^2 \rangle = \langle n \rangle = \mu$. Thus, $\sigma^2 \approx \mu$.

The robustness of figure 2 can be understood by vertically considering figure 3, which includes a dashed vertical line at $\Delta t = 5260$ on the right side. The $\alpha$ value at this $\Delta t$ for every document kind gives the mean values shown in figure 2. When $\Delta t$ was varied, the mean $\alpha$ values would change. Because the plots in figure 3 for different data kinds do not intersect, however, for any $\Delta t$, $\alpha$ is able to distinguish the data kind.

As a summary, the importance of the $\alpha$ value is threefold, as follows, although $\alpha$ changes for a $\Delta t$:

- For any real document, $\alpha > 0.5$ for any $\Delta t > 1$.
- For a given $\Delta t$, $\alpha$ takes highly similar values for documents in a single category.
- $\alpha$ takes different values for documents in different categories.

In contrast, the exponent depended only slightly on the data size. Figure 4 shows the dependency for the two largest datasets used: the New York Times (NYT, 1.5 billion words) and Mainichi (24 years) newspapers. When the data size was increased, the exponent exhibited a slight tendency to decrease. For the NYT, the decrease seemed to have a lower limit, as the figure shows that the exponent stabilized at around $10^7$ words.

The reason for this decrease can be explained as follows. The Taylor exponent becomes larger when some words occur in a clustered manner. Making the text size larger increases the number of segments (since $\Delta t$ was fixed in this experiment). If the number of clusters does not increase as fast as the increase in the number of segments, then the number of clusters per segment becomes smaller, leading to a smaller exponent. In other words, the influence of each consecutive co-occurrence of a particular word decays slightly as the overall text size grows.

5. Analysis

We are interested in finding what kind of random sequence produces a Taylor exponent larger than 0.5, to better understand the nature of linguistic sequences.

Ways to produce random sequences that model language have been studied especially in the domain of language engineering. These include Markov models, Bayesian models, grammatical models, and neural models. Bayesian models have also been studied substantially in the domain of statistical mechanics and physics,
on the basis of the Simon process (Simon 1955). Previous works such as (Gerlach and Altmann 2013) explain statistical phenomena as extensions of the Simon process.

Our group (Takahashi and Tanaka-Ishii 2018) has also shown how all these models except neural models could not produce a Taylor exponent larger than 0.5. For example, figure 5 shows the Taylor analyses of texts generated by a first-order Markov model trained with the real Moby Dick and by the Simon process. The corresponding rows of table 1 list the details of the data. Both plots were tightly distributed around the regression line with an exponent of 0.5. (Takahashi and Tanaka-Ishii 2018) shows how the exponent cannot be improved by increasing the order of the model or by using advanced techniques.

That work also shows how neural network models can produce a Taylor exponent larger than 0.5. There are limitations, however, in the experimental approach. First, because neural models cannot handle a large vocabulary size, the experiments were all conducted by ignoring the word kind for the least frequent vocabulary. Hence, although the report shows how the results followed Zipf’s law in the region of frequent words, the most important tail part of Zipf’s law was not considered. Overall, the final Taylor analysis plots were tightly distributed around the regression line, similarly to figure 5. Second, and above all, there is currently no explanation of why neural models work, and we still do not understand the nature of processes producing a large Taylor exponent.

Apart from neural models, no other language models are known to produce a sequence with a Taylor exponent larger than 0.5. In this article, we report a study of another genre of model: a random walk on a graph, in which successive visits to nodes are considered to form a sequence. Intuitively, a random walk can leave a node and come back to it, thus leading to fluctuation. Previously, (Eisler et al 2007) showed that a kind of random walk model on a Barabási-Albert (BA) graph produces a Taylor exponent larger than 0.5, provided that the random walk meets a specific condition. They assumed multiple random walkers simultaneously, so their approach would not form a single sequence. They excluded the special case of a single walker, and the condition for the random walk to give a Taylor exponent larger than 0.5 is unknown in this case.

In the rest of this paper, we consider using a random walk on a graph to produce a sequence with a Taylor exponent larger than 0.5.
Table 1. Taylor exponent, Zipf’s law, and long-range correlation (LRC) results for real data and various language models. The models include both sequential ones (second block), based on a first-order Markov model (learned from Moby Dick) and the Simon process, and random walk models with various complex network graph structures (third block). The models listed in the second and third blocks are roughly described in the main text, with the details provided in appendix E. The last four columns in the table indicate the judgment of how well the sequence fulfilled each of the scaling laws. For Zipf’s law, if the distribution exhibited a power law and the exponent was $\xi \approx 1.0$, then the result was judged ‘Strong’; if the distribution exhibited a power law with a different exponent from $\xi = 1.0$, then the result was judged ‘Weak’ (the cutoff was $\xi < 0.9$); otherwise, the judgment was ‘NA.’ Including cases of exponential decay. For the LRC, if the points plotted for the autocorrelation function (CIs) in formula (1), for the intervals described in appendix D were aligned tightly around the regression line with no negative values before $s = 50$, then the result was judged ‘Strong’; if the plot merely presented some power-law tendency (with five values at maximum for our log-bin sampled $s < 50$), then the result was judged ‘Weak’; otherwise, the judgment was ‘NA.’ Finally, the column labeled ‘Max scale of $\mu$’ indicates the largest $\mu$ in the Taylor’s law graph as a power of 10. The table shows how strongly all scaling laws were fulfilled by real data (first block). The other models variously fulfilled some of the statistical laws, but only a limited random walk on a Barabási-Albert (BA) graph could fulfill all the laws. The red row indicates a case when all the laws were fulfilled, while the blue row indicates a case when only one law was weakly fulfilled. Both rows are for BA graphs. For these graphs, the max scale presented 3 because the length of walks was 1 million, much longer than Moby Dick (length is 25 thousand words):

|                  | Size (nodes) $n$ | Length $s$ | Other parameters | Zipf $\xi$ | LRC $\alpha$ | Max scale of $\mu$ |
|------------------|-----------------|------------|------------------|------------|--------------|------------------|
| Moby Dick        |                 |            |                  |            |              |                  |
| Original         | 20,472          | 254,654    | —                | Strong     | Strong       | 0.57             |
| Thomas (CHILDES)| 8,153           | 448,771    | —                | Strong     | Strong       | 0.66             |
| Mathews Passion  | 4,851           | 205,771    | —                | Strong     | Strong       | 0.79             |
|                  |                 |            |                  |            |              |                  |
| **Sequencial models** |                  |            |                  |            |              |                  |
| Moby Dick        |                 |            |                  |            |              |                  |
| First-order Markov | 16,975          | 300,000    | $a = 0.1$        | Strong     | NA           | 0.50             |
| Simon           | 99,833          | 1,000,000  | $a = 0.1$        | Strong     | Strong       | 0.50 $-1$ |
|                  |                 |            |                  |            |              |                  |
| **Graph models** |                  |            |                  |            |              |                  |
| Moby Dick        |                 |            |                  |            |              |                  |
| Random           | 20,471          | 250,001    | —                | Strong     | Weak         | 0.51             |
| Moby Dick        |                 |            |                  |            |              |                  |
| Preference       | 20,471          | 250,001    | —                | Strong     | NA           | 0.50             |
| Regular          |                 |            |                  |            |              |                  |
| Random           | $10^2$          | 1,000,000  | $d = 2$          | NA         | NA           | 0.66 $1$    |
| Random           | $10^2$          | 1,000,000  | $d = 4$          | NA         | NA           | $-1$         |
| $G_{0,m}$        | Random          | $10^3$     | $m = 10^3$       | NA         | Weak         | 0.76 $-1$    |
| $G_{0,m}$        | Preference      | $10^3$     | $m = 10^3$       | NA         | Weak         | 0.71 $0$     |
| $G_{0,p}$        | Random          | $10^3$     | $p = 0.001$      | NA         | NA           | 0.51 $-1$    |
| $G_{0,p}$        | Preference      | $10^3$     | $p = 0.001$      | NA         | NA           | 0.51 $-1$    |
| CWS              | Random          | $10^3$     | $k = 2, p = 0.2$ | NA         | Weak         | 0.76 $0$    |
| CWS              | Preference      | $10^3$     | $k = 2, p = 0.2$ | NA         | Strong       | 0.83 $0$    |
| NSW              | Random          | $10^3$     | $k = 2, p = 0.2$ | NA         | Weak         | 0.80 $-1$    |
| NSW              | Preference      | $10^3$     | $k = 2, p = 0.2$ | NA         | Weak         | 0.77 $-1$    |
| BA               | Random          | $10^3$     | $d = 2$          | NA         | Weak         | 0.88 $1$     |
| BA               | Preference      | $10^3$     | $d = 2$          | NA         | Weak         | 0.75 $1$     |
| BA               | Random          | $10^3$     | $d = 4$          | NA         | NA           | 0.63 $0$     |
| BA               | Preference      | $10^3$     | $d = 8$          | NA         | NA           | 0.56 $0$     |
| BA               | Preference      | $10^3$     | $d = 2$          | NA         | Strong       | 0.62 $3$    |
| BA               | Preference      | $10^3$     | $d = 4$          | NA         | Strong       | 0.56 $3$    |
| BA               | Preference      | $10^3$     | $d = 8$          | NA         | Strong       | 0.52 $2$    |

5.1. Other scaling properties of language
As will be seen, because of the random walk’s nature of going back and forth, a variety of graph settings can produce a Taylor exponent larger than 0.5. Human linguistic sequences, however, exhibit other properties such as the well-known Zipf’s law and other power laws. We are interested in random walk models that explain the nature of human linguistic sequences, and we therefore also consider other scaling laws. Precisely, we seek the condition for a random walk on a graph to exhibit not only Taylor’s law but also the following two additional scaling laws.

The first is Zipf’s law, defined as follows. Let $r$ be the rank of a particular word type and $f(r)$ be its frequency.

$$f(r) \propto r^{-\xi},$$

with $\xi \approx 1.0$. Heaps’ law is another scaling property related to vocabulary population. It shows how the vocabulary grows with the text size, forming a power-law distribution. Because there have been multiple debates on how Heaps’ law is mathematically related to Zipf’s law (van Leijenhorst and van der Weide 2005, Lu et al 2013), we only consider Zipf’s law here.

The second scaling property is the long-range correlation (LRC). A time series is said to be long-range correlated if a similarity function $C(s)$ for two of its subsequences separated by a distance $s$ follows a power law:
\[ C(s) \propto s^{-\gamma}, \quad s > 0, \quad 0 < \gamma < 1. \]  

(3)

The most commonly studied similarity function is the autocorrelation function (ACF). As the ACF is applicable only for numerical time series, application of this method to natural language text requires transforming the text into a numerical series. Recent methods do so by considering the intervals of word occurrences (Altmann et al 2009, Tanaka-Ishii and Bunde 2016). In this article, we apply the method in (Tanaka-Ishii and Bunde 2016), which is the only available method for quantifying the degree of the power law representing the LRC with respect to a linguistic sequence. The details of the method are described in appendix D.

Long-range correlation analysis has a strong relation with Taylor’s law, but the two methods capture different aspects underlying a sequence: LRC considers the self-resemblance, whereas fluctuation analysis methods, including Taylor’s law, consider the variance of the number of events. Both methods produce a power law when a sequence fluctuates. Because of this difference, however, there are cases when one scaling law is fulfilled but not the other. For example, we saw that the Simon process gives a Taylor exponent of 0.5, but it produces a strong LRC because it increases the number of new words towards the end, leading to a strong similarity almost independently of the distance \( s \) (table 1, second row in the second block). As will be seen, the opposite case of a high Taylor exponent but no LRC occurs for random walks on graphs. Real data fulfills both scaling properties (table 1, first block), and we thus seek a random sequence that also fulfills both, in addition to Zipf’s law.

5.2. Random walks on graph structures

Among various graph structures, in addition to a regular graph as a reference, we chose different structures from complex networks, as listed in the third block of table 1. The leftmost column shows the graph kind, as follows. The ‘Moby Dick’ graph was a graph obtained from the real data. The regular graphs had degree \( d \) for all nodes. The label \( G_{n,m} \) indicates a graph with \( n \) nodes and \( m \) edges, while \( G_{n,p} \) indicates a graph with \( n \) nodes and edge generation probability \( p \) for each sample. The label CWS indicates a small-world graph as defined by the Watts-Strogatz model (Watts and Strogatz 1998), connected with original degree \( k \) and edge replacement probability \( p \), while NWS indicates an alternative of the CWS graph as defined by (Newman and Watts 1999). Finally, the label BA indicates a Barabási-Albert graph (Barabasi and Reka 1999, Barabási 2016) with mean degree \( d \). Because the BA generative algorithm extends the graph by adding a node with at least one edge, the minimum mean degree \( d = 2 \).

Every graph was generated with \( n \) nodes, and \( s \) steps of a random walk were then conducted on the graph. Table 1 lists the results for 1,000,000 walks and 100,000 nodes (second and third columns) except for the Moby Dick graph. All the nodes were uniquely numbered. Every node received a random, unique ID number, and this numbering had no effect on the results for the scaling law analyses. When the walker visited a node, the number was output. This generated a number sequence, which was then subjected to scaling law analysis.

The first column of table 1 also indicates one of two different representative ways of walking, defined as follows.

- **Random** means that the walker chooses an edge uniformly and randomly, whereas

- **Preference** means that the walker chooses a neighbor node in proportion to the neighbor’s degree.

Other parameter values are listed in the fourth column. More detailed definitions and further explanation of every graph in the second and third blocks are included in appendix E.

The remaining columns indicate whether the random walk fulfilled each of the scaling laws. The values of the Taylor exponents are listed (seventh column), and the judgments with respect to Zipf’s law, LRC, and Taylor analysis were made using the arbitrary criteria cutoffs detailed in the caption of table 1.

In the third block of the table, the first two rows show how the random walks on the graph acquired from Moby Dick had a Taylor exponent of \( \alpha \approx 0.5 \), almost like an i.i.d. process. We saw in the second block that the first-order Markov model acquired from Moby Dick behaved similarly. These results indicate that Markov models acquired from human linguistic sequences cannot produce fluctuation.

In contrast, other random walks on different graphs could produce a Taylor exponent larger than 0.50, including even the regular graphs. In contrast, considering Zipf’s law and the LRC, only a few of the BA graphs could reproduce a sequence similar to the original Moby Dick. Moreover, comparing the LRC and \( \alpha \) columns shows that when \( \alpha \) was large, the LRC tended to hold. In the case of the original Moby Dick (first row), the LRC held strongly even though \( \alpha \) was not so large, a combination that could be considered to characterize natural language as fulfilling both laws. Appendix F shows the Zipf’s law, LRC, and Taylor exponent results for the two examples of random walks on BA graphs that best fulfilled the laws, which are highlighted in red and blue in table 1.
Given these results, we are especially interested in the random walks on BA graphs, because they apparently have a similar tendency to human linguistic sequences with respect to the scaling properties.

5.3. Conditions for BA graph to produce scaling properties

Next, the conditions for a BA graph to produce the three scaling properties were studied. The following three parameters are important: the mean degree of all nodes, \( d \); the total number of nodes in the graph, \( n \); and the number of walking steps, \( s \). Note that \( n \) corresponds to the vocabulary, whereas \( s \) corresponds to the document length, in a sequence.

The most important factor is \( d \). Figure 6 shows the value of \( \alpha \) with respect to \( d \) for the case of \( n = s \), with \( n \) being 1,000,000 or 100,000. The value of \( \alpha \) decreased rapidly with respect to \( d \). The random walk by randomly choosing edges gave higher values of \( \alpha \) than by using the Preference method. The LRC held strongly only when \( d \) was 2 or slightly larger, whereas Zipf’s law held only for the Preference case (the two plots at the bottom).

As seen in figure 6, when \( n \) equaled \( s \), the graph size made little difference in the results. Therefore, we varied the ratio of \( s \) and \( n \) and examined the resulting values of \( \alpha \). Figure 7 shows the dependency of \( \alpha \) on \( s/n \), with the number of walks increasing to the right. The figure shows the case when \( d = 2 \).

The results show that as more walks were made, \( \alpha \) tended to decrease. This trend was shared with the results shown in figure 4 for how \( \alpha \) slightly depended on the natural language text size. Similarly to figure 6, random walks by Preference presented far lower Taylor exponents as compared with those by the Random method. Zipf’s law and the LRC were strongly present for all points in the Preference case. On the other hand, for the Random case, the Zipf’s law’s global exponent was not as steep as \(-1.0\), although the LRC was consistently positive.
5.4. Discussion
We have seen that random walks on various graphs fulfill Taylor’s law with a larger exponent than 0.5. A random walk on a graph structure fluctuates, because locally connected nodes co-occur through random walks. Because of this, many different experimental settings fulfilled the condition of the Taylor exponent being larger than 0.5, even with regular graphs. Many of these settings did not, however, fulfill the other scaling properties of language. The search for conditions under which a random walk could fulfill Zipf’s law, the LRC, and Taylor’s law suggest reconsideration of the nature of language. Only a Barabási-Albert (BA) graph with small mean degree $d$ using Preference fulfilled all three scaling laws. Given that most previous language models do not fulfill the three scaling properties (Takahashi and Tanaka-Ishii 2018), our findings provide one possible alternative mathematical model.

Besides these good BA graph results, however, all random walks on a graph structure learned from Moby Dick (first row of the second block and first two rows of the third block of table 1) produced $\alpha = 0.5$. This suggests that linguistic sequences cannot be modeled by Markov models, which confirms both previous mathematical results (Lin and Tegmark 2016) and experimental results (Takahashi and Tanaka-Ishii 2018). The main reason is that the mean degree of the Markov models was large (above 10). These Markov graphs, however, form another BA graph, because the degree distribution is known to produce a power law. The closeness of the BA graphs and the Markov graphs generated from a real sequence raises a question of the relation between them.

In the language engineering domain, considering hidden states always contributes to increasing model quality by alleviating the low-frequency problem. For example, neural models, which are the sole model type to date that fulfills the three scaling laws (other than the random walk models discussed here), are also based on hidden states encoding the context. This fact suggests the idea that the nodes of a BA graph might not be direct words but could be some hidden states, and that these hidden states produce words in a probabilistic manner. This would alleviate the problem of a large mean degree (as semantically similar words would aggregate into single nodes).

As mentioned previously, neural models still have the problems of not being able to handle rare events. We consider this study on BA graphs to offer some prospect towards reconsidering the nature of language models. How exactly to reverse-engineer a BA graph language model from a sequence in the form of a language model remains for our future work.

6. Conclusion
We have shown how Taylor’s law holds universally for human linguistic sequences and considered the nature of linguistic sequences via random walk models. With our method, a non-numerical discrete sequence is divided into segments of a given length, and the mean and standard deviation of the frequency of every event (word kind) are measured. Taylor’s law is considered to hold when the standard deviation varies with the mean according to a power law, thus giving a value for the Taylor exponent.

Theoretically, an i.i.d. process has a Taylor exponent of 0.5, whereas larger exponents indicate sequences in which words co-occur systematically. Using over 1100 texts across 14 languages, we showed that natural language texts follow Taylor’s law, with the exponent distributed around 0.58. This value differed greatly from the exponents for other data sources: enwiki8 (tagged Wikipedia, 0.63), speech data (Diet record, 0.63; child-directed speech, 0.68), programming language data (0.79), and music (0.79). None of the real data exhibited an exponent of 0.5, but it consistently showed a degree of fluctuation distinguishing the data kind.

To explain the statistical mechanics of human linguistic sequences, we sought a mathematical model that could produce a Taylor exponent larger than 0.5, in addition to fulfilling Zipf’s law and the long-range correlation (LRC). Motivated by the greater likelihood of fluctuation in a random walk of a graph, we sought the conditions under which a random walk would fulfill Taylor’s law. We found that random walks on a graph constructed from a first-order Markov model of real data behaved almost like an i.i.d. process, whereas random walks on a Barabási-Albert (BA) graph with small mean degree could fulfill the scaling laws of natural language. Our findings provide another alternative perspective to consider the nature of human linguistic activities.

Acknowledgments
We thank the PRESTO and HITE projects of the Japan Science and Technology Agency for supporting our study.
Appendix A. Data in detail

Table A1 lists all the data used for the article. The data consisted of natural language texts and language-related sequences, listed as different blocks in the table.

The natural language texts consisted of 1142 single-author long texts (first block) extracted from Project Gutenberg and Aozora Bunko across 14 languages. These are the texts above a size threshold (1 megabyte) in the two archives. The second block lists individual samples taken from Project Gutenberg. The third block lists newspaper data taken from the Gigaword corpus, available from the Linguistic Data Consortium in English, Chinese, and other major languages.

Other types of linguistic sequences appear in the fourth block. The enwiki8 100-MB dump dataset consists of tag-annotated text from English Wikipedia. The speech data consisted of the spoken record of Japan’s National Diet, and the 10 longest child-directed speech utterances in the CHILDES (Child Language Data Exchange System) database, which was preprocessed by extracting only children’s utterances (MacWhinney 2000, Bol 1995, Lieven et al 2009, Rondal 1985, Behrens 2006, Gil and Tadmor 2007, Oshima-Takane et al 1995, Smoczynska 1985, Andelković et al 2001, Benedet et al 2004, Plunkett and Strömqvist 1992). Four program sources (in Lisp, Haskell, C++, and Python) were crawled from large representative archives, parsed, and stripped of natural language comments. Finally, 12 pieces of musical data (long symphonies and so forth) were transformed from MIDI into text with the software SMF2MML (available at http://shaw.la.coocan.jp/smf2mml/), with annotations removed.

Appendix B. Taylor’s law plots for different texts

Figure B1 shows typical distributions for natural language written texts, with two single-author texts in Chinese and French, and two multiple-author texts in English and Chinese. The segment size was $\Delta r = 5620$ words. The points at the upper right represent the most frequent words, whereas those at the lower left represent the least frequent. The exponent $\alpha$ was almost the same even though English, French, and Chinese are different languages using different kinds of script.

Figure B2 shows results for other kinds of data, listed in the fourth block of table A1. The segment size was again $\Delta r = 5620$ words. The exponent became larger as the kind of text differed more from written natural language.

Appendix C. Histograms of exponents for long texts from project Gutenberg and diet record

Figure C1 shows a histogram of the Taylor exponents for the long texts taken from Project Gutenberg. Most of the texts were in English, followed by French, and then other languages, as listed in table A1. The histogram shows how the Taylor exponent ranged mainly between 0.55–0.63, and natural language texts with an exponent larger than 0.63 were rare.

Whether $\alpha$ distinguishes languages is a difficult question. The histogram suggests that Chinese texts exhibited larger values than did texts in Indo-European languages. We conducted a statistical test to evaluate whether this difference was significant as compared to English. Since the numbers of texts were very different, we used the non-parametric statistical test of the Brunner–Munzel method, among various possible methods, to test a null hypothesis of whether $\alpha$ was equal for the two distributions (Brunner and Munzel 2000). The p-value for Chinese was $p = 1.24 \times 10^{-16}$, thus rejecting the null hypothesis at the significance level of 0.01. This confirms that $\alpha$ was generally larger for Chinese texts than for English texts. Similarly, the null hypothesis was rejected for Finnish and French, but it was accepted for German. It was also accepted, at the 0.01 significance level, for Japanese texts taken from Aozora Bunko (not shown in the histogram). Because the null hypothesis was accepted for Japanese despite its large linguistic difference from English, we could not conclude whether the Taylor exponent distinguishes languages. Overall, the exponent characterizes written texts in natural language but does not characterize languages.

Figure C2 shows a histogram of the Taylor exponents for the spoken record of Japan’s Diet (adult speech). The results show how the Taylor exponent ranged tightly around the mean.

Appendix D. Long-range correlation analysis method

A self-contained summary of the analysis scheme used for the long-range correlation (LRC) is provided here, and a detailed argument for the method is found in (Tanaka-Ishii and Bunde 2016). The method basically uses the autocorrelation function to quantify the LRC. Given a numerical sequence $R = r_1, r_2, \ldots$, 

Table A1. Summary of the data used in this article. For each dataset, the length is the total number of words, and the vocabulary is the number of different words.

| Texts | Language | $\alpha$ | Number of samples | Length | Vocabulary |
|-------|----------|---------|-------------------|--------|------------|
|       |          | mean    |                   | Mean   | Min        | Max        | Mean   | Min   | Max |
| English | 0.58     | 910     | 313 127.5         | 185939 | 2488933    | 17 236.7  | 7320  | 69811 |
| French  | 0.57     | 66      | 293 192.5         | 169415 | 1528177    | 22 097.4  | 14105 | 57192 |
| Finnish | 0.55     | 33      | 197 519.3         | 149488 | 396920     | 33 596.2  | 26274 | 47262 |
| Chinese | 0.61     | 32      | 629 916.8         | 315099 | 4145117    | 15 351.9  | 9132  | 60949 |
| Dutch   | 0.57     | 27      | 256 859.3         | 198924 | 435683     | 19 158.2  | 13879 | 31594 |
| German  | 0.59     | 20      | 236 174.0         | 184320 | 331321     | 24 241.3  | 11078 | 37227 |
| Gutenberg | 0.57   | 14      | 266 809.7         | 196961 | 369326     | 29 102.6  | 18640 | 45031 |
| Italian | 0.57     | 12      | 363 837.2         | 219787 | 903051     | 26 110.1  | 18110 | 36306 |
| Spanish | 0.58     | 10      | 159 969.2         | 119196 | 243953     | 22 804.7  | 15876 | 31385 |
| Greek   | 0.57     | 2       | 505 743.5         | 205228 | 806259     | 59 666.5  | 28738 | 90595 |
| Latin   | 0.56     | 1       | 261 382.0         | 261382 | 261382     | 24 718.0  | 24718 | 24718 |
| Portuguese | 0.56  | 1       | 198 303.0         | 198303 | 198303     | 38 383.0  | 38383 | 38383 |
| Hungarian | 0.57    | 1      | 208 455.0         | 208455 | 208455     | 26 334.0  | 26334 | 26334 |
| Tagalog | 0.59     | 13      | 616 676.2         | 105342 | 2951319    | 19 759.0  | 6619  | 49099 |
| Japanese | 0.59    | 1       | 254 654.0         | 254654 | 254654     | 20 472.0  | 20472 | 20472 |
| English | 0.57     | 1       | 701 255.0         | 701255 | 701255     | 18 450.0  | 18450 | 18450 |
| Chinese | 0.59     | 1       | 691 407.0         | 691407 | 691407     | 31 956.0  | 31956 | 31956 |
| French  | 0.57     | 1       | 22 679 512.0      | 2267912 | 2267912    | 137 466.0 | 137466 | 137466 |
| English | 0.58     | 1       | 1 528 137 194.0   | 1528137194 | 1528137194 | 3 155 494.0 | 3155494 | 3155494 |
| Chinese | 0.58     | 1       | 19 420 853.0      | 19420853 | 19420853   | 172 139.0 | 172139 | 172139 |
| Japanese | 0.56    | 24 (yrs) | 31 321 593.3     | 2448330 | 40270705   | 145 533.5 | 127289 | 169269 |
| enwiki8 | 0.63     | 1       | 14 647 847.0      | 14647847 | 14647847   | 1 430 790.0 | 1430790 | 1430790 |
| Diet record | 0.63 | 250     | 348 902.8         | 101441 | 1467337    | 9 511.4  | 4957  | 20429 |
| CHILDES | 0.68     | 10      | 193 434.0         | 48952  | 448772     | 9 907.0  | 5618  | 17892 |
| Programs | 0.79     | 4       | 34 161 017.8      | 3697198 | 68622161   | 838 906.8 | 127652 | 1545126 |
| Music   | 0.79     | 12      | 135 992.4         | 76628  | 215479     | 9 186.9  | 906   | 27042 |
Given a sequence of length $M$, let the mean and standard deviation be $\mu$ and $\sigma$, respectively. Consider the following autocorrelation function:

$$C(s) = \frac{1}{(M - s)\sigma^2} \sum_{i=1}^{M-s} (r_i - \mu)(r_{i+s} - \mu).$$

(4)

This is a fundamental function to measure the correlation, the similarity of two subsequences separated by distance $s$, it calculates the statistical covariance between the original sequence and a subsequence starting from the $s$th offset element, standardized by the original variance of $\sigma^2$. For every $s$, the value ranges between $-1.0$ and $1.0$, with $C(0) = 1.0$ by definition. For a simple random sequence, such as a random binary sequence, the function gives small values fluctuating around zero for any $s$, as the sequence has no correlation with itself. The sequence is judged to be long-range correlated when $C(s)$ decays by a power law, as specified in formula (3).

The essential problem lies in the fact that a language sequence is not numerical and thus must be transformed into some numerical sequence. The method of (Tanaka-Ishii and Bunde 2016) transforms a word

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figureB1}
\caption{Examples of Taylor’s law for natural language texts. Hong Lou Meng and Les Misérables are Chinese and French literary texts, respectively, appearing in the second block of table A1 as examples of single-author texts. The two newspapers are representative of multiple-author texts, in English and Chinese, respectively. Each point represents a kind of word. The values of $\sigma$ and $\mu$ for each word kind are plotted across texts within segments of size $\Delta t = 5620$. The Taylor exponents obtained by the least-squares method were all close to the mean value of 0.58.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figureB2}
\caption{Examples of Taylor’s law for the alternative datasets listed in table A1: enwiki8 (tag-annotated Wikipedia), Thomas (longest dataset in CHILDES), Lisp source code, and the music of Bach’s St Matthew Passion, all with $\Delta t = 5620$. These datasets exhibited far larger Taylor exponents than did typical natural language written texts.}
\end{figure}
sequence into a numerical sequence by using intervals of rare words. The following example demonstrates how this is done. Consider the target *Romeo* in the sequence ‘Oh Romeo Romeo wherefore art thou Romeo.’ *Romeo* has a one-word interval between its first and second occurrences, and the third *Romeo* occurs as the fourth word after the second *Romeo*. This gives the numerical sequence [1, 4] for this clause and the target word *Romeo*. The target does not have to be one word but can be any element in a set of words. For example, suppose that the target consisted of two words, the two rarest words in this clause: *Romeo*, and *wherefore*. Then, the interval sequence would be [1, 1, 3], because *wherefore* occurs right after the second *Romeo*, and the third *Romeo* occurs as the third word after *wherefore*. As rare words occur only in small numbers, consideration of multiple rare words serves to quantify their behavior as an accumulated tendency.

As a summary, the overall procedure is described as follows. Given a numerical sequence of length $M$, the number of intervals for one $N$th of (rare) words is $M_N \equiv M/N - 1$. For the resulting interval sequence

\[ \frac{M}{N} \equiv M/N - 1 \]

Figure C1. Histogram of Taylor exponents for long texts in Project Gutenberg (1129 texts). The mean value was 0.58. The legend indicates the languages, in frequency order. Each bar shows the number of texts with that value of $\alpha$. Because of the skew of languages in the original conception of Project Gutenberg, most of the texts are in English, shown in blue, whereas texts in other languages are shown in other colors. The histogram shows how the Taylor exponent ranged fairly tightly around the mean, and natural language texts with an exponent larger than 0.63 were rare.

Figure C2. Histogram of Taylor exponents for long texts taken from the Diet record (250 texts). The mean value was 0.63. Each bar shows the number of texts with that value of $\alpha$. The histogram shows how the Taylor exponent ranged fairly tightly around the mean, within a range of ±0.3.

3 Considering one $N$th of words as rare means that the average number of intervals is almost $N$, for any given total number of words $M$, as follows. One $N$th of the words means $M/N$ words. Then, for sufficiently large $M$, the mean interval length is $(M - 1) / (M/N - 1) \approx N$. 
The analysis included two types of models, namely, sequential and graph models. The second block of table 1 in the main text lists the sequential models, which gave a Taylor exponent of 0.5:

- **Moby Dick first-order Markov model** is a model that generates a word, after a preceding word, with a probability defined as the relative frequency of occurrence in the original *Moby Dick*. Note that this is also a random walk on a graph, where a node is a word, and an edge is formed between a word and its subsequent word. The edge is directed, and the random walk is made in proportion to the number of times that the two words occur successively.

- **Simon** uses the Simon process (Simon 1955) to generate a subsequent element, given the past sequence. The process starts with an element sequence. With constant probability \( a \), a new element is generated, whereas with probability \( 1 - a \), an element is sampled from the past sequence.

The third block lists the graph models, which often exhibited a Taylor exponent larger than 0.5:

- **Moby Dick graph models**: The first two rows lists graph models constructed from the original *Moby Dick*. A word forms a node, with an edge indicating that the two words occur successively. Unlike the first-order Markov model mentioned above, the edge is undirected, and the random walk is conducted by either the *Preference or Random* method (as described in the main text). The resulting Taylor exponents were 0.51 and 0.50, respectively, suggesting that random walks on a *Moby Dick* graph cannot produce a Taylor exponent as large as that of the original text.

- **Regular** graphs are those in which all nodes have degree \( d \).

- **\( G_{n,m} \)** is a random graph with \( n \) nodes and \( m \) edges.
\( G_{n,p} \) is a random graph with \( n \) nodes connected with probability \( p \).

- **CWS** is a graph based on the Watts-Strogatz model, with \( n \) nodes that are first connected to \( k \) nearest-neighbor nodes; then, with probability \( p \), each edge is reconnected to another node that is connected (Watts and Strogatz 1998).

- **NWS** is a graph, based on Newman’s variation of the WS graph, with \( n \) nodes that are first connected to \( k \) nearest-neighbor nodes; then, with probability \( p \), a new edge is added for each edge Newman and Watts (1999).

- **BA** is a Barabási-Albert graph with mean degree \( d \) (Barabasi and Reka 1999, Barabási 2016). The graph starts with one node having one edge, and nodes are generated successively by adding them to the preexisting graph by preferential attachment with \( d/2 \) edges. Because of this algorithm, the graph’s minimal degree is 2. When \( d \) is not an integer, edges are probabilistically chosen so that the mean degree becomes the specified value of \( d \). For example, consider the rows with \( d = 2.4 \) listed in table 1 of the main text. In that case, a node is added to the graph with either one or two edges so that the overall mean becomes \( d = 2.4 \).

**Appendix F. Zipf’s law, long-range correlation, and Taylor’s law of best models of table A1**

Figures F1 and F2 show the rank-frequency distribution, long-range correlation (LRC), and Taylor analysis calculated for the two best random walk sequences discussed in this article. Figure F1 is for a BA model using *Preference* walk with \( d = 2 \), which was the best model (listed in red in table 1). The sequence fulfilled all three properties in a reasonable way, although the LRC differed by having much faster decay of \( C(s) \). Figure F2 is for a BA model using *Preference* walk with \( d = 2.4 \), which was the second best model (listed in blue in table 1). The sequence fulfilled Zipf’s law and Taylor’s law but was barely long-range correlated, which led to the judgment of ‘Weak’.

**ORCID iDs**

Kumiko Tanaka-Ishii  https://orcid.org/0000-0003-1752-3951
References

Altmann E G, Pierrehumbert J B and Motter A E 2009 Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words PLOS ONE 4 1–7

Andelkovic D, Ševa N and Moskovlijević J 2001 Serbian Corpus of Early Child Language (Belgrade: Laboratory for Experimental Psychology, Faculty of Philosophy, and Department of General Linguistics, Faculty of Philology, University of Belgrade)

Badii R and Politi A 1997 Complexity: Hierarchical Structures and Scaling in Physics (Cambridge: Cambridge University Press)

Barabási A-L 2016 Network Science (Cambridge: Cambridge University Press)

Barabási A L and Reka A 1999 Emergence of scaling in random networks Science 286 509–612

Behrens H 2006 The input-output relationship in first language acquisition Language and Cognitive Processes 21 2–24

Benedet M, Cruz C, Carrasco M and Snow C 2004 Spanish BecaCESNo Corpus (Pittsburgh, PA: TalkBank)

Bo G W 1995 Implicational Scaling in Child Language Acquisition: The Order of Production of Dutch Verb Constructions (Amsterdam: Institute for General Linguistics) ch 3. Amsterdam Series in Child Language Development

Brunner T and Munzel U 2000 The nonparametric Behrens–Fisher problem: asymptotic theory and a small-sample approximation Biometrical Journal 42 17–25

Chomsky N 1956 Three models for the description of language IRE Transactions on Information Theory 2 113–24

Cohen J E and Xu M 2015 Random sampling of skewed distributions implies Taylor’s power law of fluctuation scaling Proc. Natl Acad. Sci. USA 112 7749–54

Ebeling W and Pöeschel T 1994 Entropy and long-range correlations in literary English Europhys. Letters 26 241–6

Eisler Z, Bartos I and Kertész J 2007 Fluctuation scaling in complex systems: Taylor’s law and beyond Adv. Phys. 89–142

Gerlach M and Altmann E G 2014 Scaling laws and fluctuations in the statistics of word frequencies New J. Phys. 16 113010

Gil D and Tadmor U 2007 The MPY-EVA Jakarta Child Language Database (Leipzig: A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University)

Kobayashi T and Tanaka-Ishii K 2018 Taylor law of human linguistic sequences Annual Conference of Association for Computational Linguistics pp 1138–48

Lennartz S and Bunde A 2009 Eliminating finite-size effects and detecting the amount of white noise in short records with long-term memory Phys. Rev. E 79 066101

Lieven E, Salomo D and Tomasello M 2009 Two-year-old children’s production of multiword utterances: a usage–based analysis Cogn. Linguist. 20 481–507

Lin H W and Tegmark M 2016 Critical behavior from deep dynamics: a hidden dimension in natural language Entropy arXiv:1606.06737v2 https://arxiv.org/abs/1606.06737

Lu L, Zhang Z, K and Zhou T 2013 Deviation of zipf’s and heaps’ laws in human languages with limited dictionary sizes Sci. Rep. MacWhinney B 2000 The CHILDES Project (New York: Psychology Press)

Newman M and Watts D J 1999 Renormalization graph analysis of the small–world network model Phys. Lett. A 263 341–6

Oshima–Takane Y, MacWhinney B, Sirai H, Miyata S and Naka N 1995 CHILDES Manual for Japanese (Montreal: McGill University)

Plunkett K and Strömqvist S 1992 The acquisition of scandinavian languages The Crosslinguistic Study of Language Acquisition ed D I Slobin vol 3 (Mahwah, NJ: Lawrence Erlbaum Associates) pp 457–556

Rondal J A 1985 Adult-Child Interaction and the Process of Language Acquisition (Westport: Praeger Publishers)

Simon H A 1955 On a class of skew distribution functions Biometrika 42 425–40

Smith H F 1938 An empirical law describing hetero-genitivity in the yields of agricultural crops J. Agriculture Sci. 28 1–23

Smoczynska M 1985 The acquisition of polish The Crosslinguistic Study of Language Acquisition ed D I Slobin (Mahwah: Lawrence Erlbaum Associates) pp 595–686

Takahashi S and Tanaka-Ishii K 2017 Do neural nets learn statistical laws behind natural language? PLOS One https://arxiv.org/abs/1707.04848 (https://doi.org/10.1371/journal.pone.0189326)

Takahashi S and Tanaka-Ishii K 2018 Assessing Language Models with Scaling Properties https://arxiv.org/abs/1804.08881

Tanaka-Ishii K and Bunde A 2016 Long-range memory in literary texts: on the universal clustering of the rare words PLOS One 12 e0189326

Taylor L R 1961 Aggregation, variance and the mean Nature 181 732–5

van Leijenhorst D C and van der Weide T P 2005 A formal derivation of Heaps’ law Inf. Sci. 170 263–72

Watts D J and Strogatz S H 1998 Collective dynamics of small-world networks Nature 393 440–2