Rethinking the Relationship between the Power Law and Hierarchical Structures

Kai Nakaishi[1,2,*] **Ryo Yoshida**[3] **Kohei Kajikawa**[4,*] **Koji Hukuahima**[3] **Yohei Oseki**[3,5]
[1]RIKEN, Japan [2]National Institute for Japanese Language and Linguistics, Japan
[3]The University of Tokyo, Japan [4]Georgetown University, USA
[5]National Institute of Informatics, Japan
kai.nakaishi@riken.jp
{yoshiryo0617, k-hukushima, oseki}@g.ecc.u-tokyo.ac.jp
kk1571@georgetown.edu

Abstract

Statistical analysis of corpora provides an approach to quantitatively investigate natural languages. This approach has revealed that several power laws consistently emerge across different corpora and languages, suggesting universal mechanisms underlying languages. In particular, the power-law decay of correlations has been interpreted as evidence of underlying hierarchical structures in syntax, semantics, and discourse. This perspective has also been extended beyond corpora produced by human adults, including child speech, birdsong, and chimpanzee action sequences. However, the argument supporting this interpretation has not been empirically tested in natural languages. To address this gap, the present study examines the validity of the argument for syntactic structures. Specifically, we test whether the statistical properties of parse trees align with the assumptions in the argument. Using English and Japanese corpora, we analyze the mutual information, deviations from probabilistic context-free grammars (PCFGs), and other properties in natural language parse trees, as well as in the PCFG that approximates these parse trees. Our results indicate that the assumptions do not hold for syntactic structures and that it is difficult to apply the proposed argument not only to sentences by human adults but also to other domains, highlighting the need to reconsider the relationship between the power law and hierarchical structures.

1 Introduction

Natural languages exhibit characteristic statistical properties. Quantifying these properties through corpus-based analyses is crucial for understanding natural languages. Previous research has extensively explored this approach, uncovering various power laws, such as Zipf's law (Zipf, 1949) and Heaps' law (Heaps, 1978), that consistently appear across different corpora and languages. These findings suggest the existence of universal mechanisms underlying these power laws.

The present study focuses on the power-law decay of correlation. Previous studies have measured correlations between elements such as phones (Sainburg et al., 2019), characters (Li, 1989a; Ebeling and Pöschel, 1994; Ebeling and Neiman, 1995; Lin and Tegmark, 2017; Shen, 2019), and words (Tanaka-Ishii and Bunde, 2016; Takahashi and Tanaka-Ishii, 2017, 2019; Mikhaylovskiy and Churilov, 2023; Nakaishi et al., 2024), using metrics such as correlation functions and mutual information (MI), across various corpora. These studies have consistently found that the correlation decays according to a power-law function of the distance between elements.

The power-law decay is qualitatively slower than the exponential decay, which is typically observed when interactions between elements are local, as in Markov chains. Therefore, the power-law decay of correlation in natural languages indicates that distant elements remain strongly correlated. This observation is consistent with linguistic intuition. For example, syntactic dependencies can, in principle, span arbitrarily long distances through repeated center-embedding, and a word introduced early in a context may still influence interpretation much later.

Moreover, the power-law decay of correlation has been interpreted as evidence for underlying hierarchical structures. In natural languages, previous research has linked this power-law behavior to linguistic hierarchies in syntax, semantics, and discourse (Alvarez-Lacalle et al., 2006; Altmann et al., 2012; Lin and Tegmark, 2017). This interpretation has also been extended beyond speech and written texts by human adults. Simi-

*This work was conducted while KN and KK were at the University of Tokyo.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1917–1934, 2026. https://doi.org/10.1162/TACL.a.785
Action Editor: Carlos Gómez-Rodríguez. Submission batch: 11/2025; Revision batch: 2/2026; Published 8/2026.

lar power-law behavior has been reported in child speech (Sainburg et al., 2022), birdsong (Sainburg et al., 2019; Sainburg and Gentner, 2021; Youngblood, 2024), and chimpanzee action sequences (Howard-Spink et al., 2024), leading to the claim that these may also possess some form of underlying hierarchical structures, which need not correspond to linguistic hierarchies such as syntax and semantics.

This interpretation is justified by Lin and Tegmark (2017). They showed that the decay of correlation follows a power law in sequences generated by probabilistic context-free grammars (PCFGs; Jelinek et al., 1992), which are a simple model for generating hierarchical tree structures. To be more precise, they proved that in PCFGs, the correlation decays exponentially in sufficiently large structures. In addition, they assumed that the sequential distance grows exponentially with the structural distance. These two give rise to the power-law decay of correlation in sequences.

However, several assumptions underlying this argument seem implausible in light of well-established linguistic facts. Yet, previous studies have not tested the validity of these assumptions quantitatively using empirical data.

To address this gap, we test whether these assumptions hold for hierarchical structures in natural language syntax, i.e., parse trees. Syntactic structures have been well formalized, resulting in the availability of large-scale treebanks. This enables statistical analyses to test these assumptions.

Specifically, we formalize three key questions essential for examining the validity of the assumptions: (i) Does the correlation decay exponentially in syntactic structures?; (ii) Does the sequential distance grow exponentially with the structural distance?; (iii) Do the statistical properties of syntactic structures deviate significantly from those of PCFGs? To answer these questions, we analyze English and Japanese treebanks.

Our findings are negative for all three questions, indicating that the assumptions in the proposed argument do not hold for syntactic structures in natural languages. Therefore, it is necessary to reconsider the relationship between the two facts, the power-law decay of correlation and the hierarchical syntactic structures in natural languages.

In addition, we perform the same analyses on trees generated by the PCFG that approximates parse trees, to highlight differences between natural language parse trees and trees generated by the PCFG. These analyses further show that trees generated by the PCFG are too small to exhibit the asymptotic behavior assumed in the argument of Lin and Tegmark (2017). This observation raises doubts about the applicability of their argument not only to single sentences by adults but also to other sequences whose lengths are comparable to those of single sentences, such as child speech, birdsong, and chimpanzee action sequences.

While the present study focuses on the hierarchical syntactic structures of individual sentences, the structures of entire texts consisting of multiple sentences, such as discourse structures, also appear to be hierarchical. The argument by Lin and Tegmark (2017), if applied to these structures, might account for the power-law decay of correlation in natural languages.

It would therefore be fruitful to investigate the relationship between the power-law behavior and hierarchical structures at the discourse level. Although previous studies have proposed formal frameworks for discourse annotation (Mann and Thompson, 1988; Prasad et al., 2014; Zeldes et al., 2025), the size of existing corpora remain too small to examine the validity of the argument. Developing sufficiently large-scale discourse corpora is an important next step for future research.

Another direction is to explore alternative mechanisms that may give rise to power-law behavior in natural languages. One such mechanism is *critical phenomena* in the statistical physics literature.

2 Background

2.1 Power-law Decay of Correlation

Previous research has shown that the correlation in natural languages decays according to a power-law function of distance. This phenomenon has been observed across various corpora and languages (Li, 1989a; Ebeling and Pöschel, 1994; Ebeling and Neiman, 1995; Tanaka-Ishii and Bunde, 2016; Lin and Tegmark, 2017; Takahashi and Tanaka-Ishii, 2017; Shen, 2019; Takahashi and Tanaka-Ishii, 2019; Sainburg et al., 2019; Mikhaylovskiy and Churilov, 2023; Nakaishi et al., 2024).

The significance of the power-law decay of correlation is clear in comparison with the exponential decay. For example, in signals generated sequentially by a Markov chain, the correlation C

is asymptotically exponential with the distance r, i.e., $C \sim \exp(-\lambda r)$ (Li, 1990; Lin and Tegmark, 2017). In this case, if the distance is sufficiently larger than $\xi = 1/\lambda$, the correlation is negligible. Thus, ξ, which is called the *correlation length* in statistical physics (Stanley, 1971), means the scale within which the correlation is significant.

In contrast, the power-law decay is expressed as $C \sim r^{-\alpha}$. Although the right-hand side resembles a polynomial form, the exponent α is a real number rather than a natural number. This decay is slower than the exponential one for any correlation length ξ, meaning that the correlation is significant for any long distances. In other words, the scale of correlation diverges.

Moreover, power-law functions are scale-invariant. For the scale transformation $r' = kr$, we can rescale the correlation by $C' = k^{-\alpha}C$ so that the relation remains the same, $C' \sim r'^{-\alpha}$. This property also implies the divergence of the scale.

Therefore, the power-law decay of correlation observed in various corpora means that the scale diverges in natural languages. Intuitively, this implies that a local change can propagate and affect the global structure of a text.

2.2 Power Law and Hierarchical Structures

In studies of human language, this power-law behavior has been interpreted as evidence of underlying hierarchical structures. This interpretation is based on the work of Lin and Tegmark (2017), which shows that the decay of correlation follows a power law in sequences generated by PCFGs.

PCFGs generate tree structures in a top-down manner, from the root to the leaves iteratively, where the category of a parent node determines the categories of its children. The leaves correspond to a surface sequence, while the whole tree represents the underlying hierarchical structure.

PCFGs are a probabilistic extension of context-free grammars (Chomsky, 1956), which were originally introduced to describe the nested structures, such as the *if... then* construction. PCFGs have also been used to model syntactic structures (Charniak, 1997), due to their simplicity and ability to capture fundamental aspects of hierarchical syntactic structures.

We provide an overview of the discussion in Lin and Tegmark (2017). To understand their discussion, it is useful to distinguish between *sequential* and *structural* distances, which we refer to as r_{seq} and r_{str}, respectively. The former is defined along the sequence. The latter is the distance in the hierarchical structure behind the sequence, as shown in Figure 1.

Lin and Tegmark (2017) have proven that in PCFGs, the correlation decays exponentially with the structural distance, $C \sim \exp(-\lambda r_{\text{str}})$, at sufficiently large distances. Note that the decay may follow non-exponential forms at shorter distances. They also assumed that generated trees are balanced, i.e., the trees have neither left- nor right-branching bias, as in Figure 1 (a). Letting the depth of the two nodes from the common ancestor be d, the structural distance is roughly $r_{\text{str}} \sim 2d$, and the sequential distance is exponential with d, $r_{\text{seq}} \sim q^d$, where q is the typical number of children of each node. Thus, the sequential distance grows exponentially with the structural distance, $r_{\text{seq}} \sim \exp(\mu r_{\text{str}})$, where $\mu = (\ln q)/2$. Combining $C \sim \exp(-\lambda r_{\text{str}})$ and $r_{\text{seq}} \sim \exp(\mu r_{\text{str}})$ leads to the power-law decay of correlation in sequences, $C \sim r_{\text{seq}}^{-\alpha}$, with $\alpha = \lambda/\mu$.

This discussion demonstrates that the correlation in sequences decays according to a power law if the sequences have hierarchical structures that satisfy the two properties: First, the correlation decays exponentially with the structural distance, $C \sim \exp(-\lambda r_{\text{str}})$; second, the structures are balanced, so that the growth of sequential distance is exponential with the structural distance, $r_{\text{seq}} \sim \exp(\mu r_{\text{str}})$.

The present study aims to examine the validity of this proposed argument. Throughout this study, we focus on syntactic structures. It is widely accepted in linguistics that these structures are hierarchical, with well-established formal descriptions. Using large-scale datasets based on this formalism, we can test if the proposed argument holds for syntactic structures.

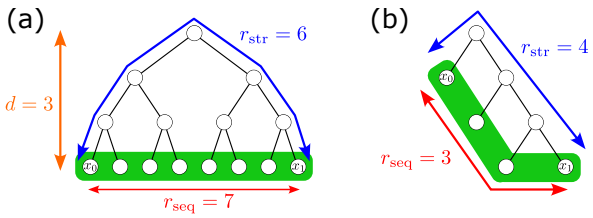

Figure 1: Sequential and structural distances. (a) If trees are balanced, the sequential distance grows exponentially with the structural distance, $r_{\text{seq}} \sim \exp(\mu r_{\text{str}})$. (b) If trees are strongly biased, the growth is slower. For example, the relation is linear, $r_{\text{seq}} \sim r_{\text{str}}$.

It is worth noting that, in principle, PCFGs can be models not only for syntactic structures but also for more general hierarchical structures. Indeed, PCFGs have been applied to represent hierarchical structures in diverse domains (Ellis et al., 2015; Worth and Stepney, 2005; Gilbert and Conklin, 2007; Knudsen and Hein, 1999; Harlow et al., 2012; Li, 1989b, 1991; Tano et al., 2020; Worth and Stepney, 2021).

From this perspective, the argument by Lin and Tegmark (2017) may also apply to other types of linguistic structures, such as semantic and discourse ones. For example, entire texts composed of many sentences intuitively appear to have hierarchical structures defined by paragraphs, sections, chapters, and so on. This hierarchy may explain why the decay of correlation follows a power law beyond individual sentences, at the level of entire texts.

However, it is difficult to empirically test whether the argument holds for multi-sentence texts. Although previous studies have proposed formal frameworks for annotating discourse structures (Mann and Thompson, 1988; Prasad et al., 2014; Zeldes et al., 2025), existing discourse corpora are not yet large enough to reliably estimate the statistical quantities introduced in Section 3, which we use to examine the validity of the assumptions in Lin and Tegmark (2017).

2.3 Rethinking Previous Research

To test whether the argument by Lin and Tegmark (2017) holds for syntactic structures in natural languages, several questions must be addressed.

First, do syntactic structures in natural languages actually exhibit the behaviors assumed in their argument? More specifically:

i) Does the correlation in syntactic structures decay exponentially with the structural distance?

ii) Are syntactic structures sufficiently balanced such that the sequential distance grows exponentially with the structural distance?

In particular, the second one appears to contradict the well-known fact that syntactic structures typically exhibit a left- or right-branching bias (Hawkins, 1990; Dryer, 1992) (for instance, see Figure 2). In such cases, the growth can be slower than exponential. For example, Figure 1 (b) illustrates a case where the growth is linear, $r_{\text{seq}} \sim r_{\text{str}}$.

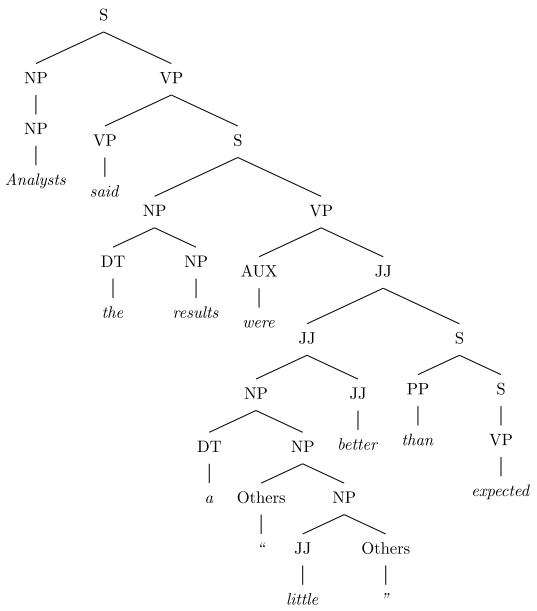

Figure 2: Example of a parse tree from BLLIP, preprocessed as described in Section 3.1. This tree has a strong right-branching bias.

In addition, Lin and Tegmark (2017) use PCFGs as models of hierarchical structures in natural languages. However, the expressive power of PCFGs is severely limited. In this model, the distribution of a subtree depends only on its root, a property we refer to as the *context-free independence*. Previous studies have suggested that relaxing this property is necessary for modeling syntactic structures (Collins, 2003; Johnson et al., 2006; O'Donnell et al., 2009). Therefore, the last question is:

iii) Do the statistical properties of syntactic structures deviate significantly from those of PCFGs?

These three questions remain unresolved. First, although extensive research has measured correlations in texts and speech, the correlation in the underlying structures have not been quantified. The answer to question (i) remains elusive. Second, previous studies have not quantitatively examined how the branching bias affects the relationship between structural and sequential distances, which is crucial for addressing question (ii). Finally, it is well known that PCFGs cannot adequately model natural language syntax without being augmented to relax the context-free independence. However, to answer question (iii) clearly, we need to systematically quantify the differences in the statis-

tical properties between natural language syntax and PCFGs.

In this study, we perform statistical analyses on parse trees to address these questions, one by one.

3 Methods

3.1 Datasets and Preprocessing

We primarily use the Brown Laboratory for Linguistic Information Processing 1987-89 Corpus Release 1 (BLLIP; Charniak et al., 2000). Its large scale enables precise statistical analyses.

To verify the robustness of our results, we also conduct the same analyses on two additional corpora. The first is WikiText (Merity et al., 2017), an English corpus from a different domain than BLLIP. The second is the NINJAL Parsed Corpus of Modern Japanese (NPCMJ; National Institute for Japanese Language and Linguistics, 2016), a treebank of Japanese, which is typologically distinct from English.

BLLIP BLLIP (Charniak et al., 2000) is a large-scale English treebank that provides Penn Treebank (Marcus et al., 1993)-style syntactic annotations for three years' worth of Wall Street Journal articles. This treebank contains approximately 30 million words and one million sentences in total. We remove the phonological null elements, such as traces of movement and ellipsis, to ensure alignment with WikiText-103 (Merity et al., 2017), which does not include traces.

The number of child nodes differs across nodes in the dataset, making it difficult to align nodes across different parse trees. To address this issue, we binarize the trees by recursively grouping child nodes from right to left, yielding right-branching structures. As the binarization procedure introduces additional unary branches, we collapse all unary branches. The binarization and unary-collapsing steps were implemented using the `Tree.chomsky_normal_form` and `Tree.collapse_unary` methods in the NLTK library (Bird and Loper, 2004). To provide intuition about how much these procedures affect tree shapes, we report depth, branching factor, and the proportion of unary nodes of the original trees in Appendix A.

In addition, we reduce the number of categories, since a larger category set requires more data for reliable statistical estimation. In particular, as categories become more fine-grained, correlations decrease and become difficult to distinguish from noise. Coarse-graining categories is crucial for capturing correlations.

Specifically, we replace part-of-speech (POS) and phrasal tags with a smaller set of categories. We refer to the Universal Dependencies (de Marneffe et al., 2021), which contains fewer types of tags than the Penn Treebank-style corpus, and group tags that tend to have similar types of dependents and heads. The eleven categories used in this study are listed in Appendix B. We apply the same categories to both POS and phrasal tags, as Lin and Tegmark (2017) treat leaves and internal nodes in the same manner. An example of a preprocessed parse tree is shown in Figure 2.

The statistical quantities analyzed here are defined for the distribution of trees and estimated from the empirical dataset. Because parse trees depend on the chosen annotation scheme, these quantities also depend on the scheme. Nevertheless, we expect that the main results, such as whether the decay is exponential or power-law, will remain unchanged unless the annotation scheme is too fine-grained to capture correlations.

To verify this expectation, we also perform the same analyses under two alternative settings: the unbinarized setting and the phrasal-tag-only setting, where we distinguish phrasal tags from POS tags and focus only on the former. Appendix C shows that the results for both settings are qualitatively similar to those in the main text, that is, the answers to questions (i), (ii), and (iii) are negative.

Because the argument of Lin and Tegmark (2017) is based on simple PCFGs, the coarse-grained annotation scheme considered here is sufficient to test the validity of their argument. At the same time, statistical analyses under more refined annotation schemes would be valuable not only for assessing the robustness of our results but also from a linguistic perspective. However, since finer-grained annotation schemes require more data for reliable estimation of statistical quantities, analyses under such schemes would require the construction of annotated corpora much larger than BLLIP. We leave this direction for future work.

WikiText To examine whether our conclusions hold beyond BLLIP, we perform the same analyses on WikiText-103 (`wikitext-103-raw-v1` split) (Merity et al., 2017), which consists of English Wikipedia articles that preserves the orig-

inal Wikipedia markup and contains over 100 million tokens and approximately 3.9 million sentences. We implement a rule-based detokenizer and segment sentences with a SPACY `sentencizer` (Honnibal et al., 2020). Then, we extract sentences up to 40 words (resulting in approximately 3.4 million sentences) and apply the Berkeley Neural Parser (`benepar_en3` model) (Kitaev and Klein, 2018; Kitaev et al., 2019) to them to obtain Penn Treebank-style constituency trees. Finally, we binarize the trees and reduce the number of categories using the same method applied to BLLIP.

NPCMJ We also analyze the Japanese manually-annotated treebank, NPCMJ (National Institute for Japanese Language and Linguistics, 2016). It contains more than 1.4 million tokens and approximately 90,000 sentences. We first remove all of the phonologically null tokens (null pronouns, traces, etc.) and extract sentences of up to 40 words, resulting in approximately 82,000 sentences. We then binarize all parse trees and reduce the number of categories by systematically rewriting the POS and phrasal tags using a rule-based mapping scheme listed in Appendix B. Since Japanese is a head-final langauge, all trees are binarized in a left-branching direction, in contrast to the right-branching binarization applied to English.

As with BLLIP, Appendix A reports depth, branching factor, and the proportion of unary nodes for the original trees in NPCMJ, serving as a reference for assessing both the effect of preprocessing and the differences between English and Japanese. In addition, Appendix C reports results for unbinarized trees. The results are qualitatively similar to those for binarized trees, as in BLLIP, supporting the robustness of our conclusions.

3.2 Mutual Information

As a metric for correlation, we use the mutual information (MI) as in Lin and Tegmark (2017). The MI between two tags is defined by

$$I = \sum_{x_0, x_1} P(x_0, x_1) \ln \frac{P(x_0, x_1)}{P(x_0)P(x_1)}, \quad (1)$$

where $P(x_0)$ and $P(x_1)$ are the marginal probabilities of the first and second tags, and $P(x_0, x_1)$ is their joint probability. The MI can be decomposed into

$$I = S_0 + S_1 - S_{01}, \quad (2)$$

where S_0, S_1, and S_{01} are the Shannon entropy, defined by $S_0 = -\sum_{x_0} P(x_0) \ln P(x_0)$, $S_1 = -\sum_{x_1} P(x_1) \ln P(x_1)$, and $S_{01} = -\sum_{x_0, x_1} P(x_0, x_1) \ln P(x_0, x_1)$. By estimating S_0, S_1, and S_{01}, we can estimate the MI.

A naive estimate can be obtained by substituting the empirical distributions $\hat{P}(x_0)$, $\hat{P}(x_1)$, and $\hat{P}(x_0, x_1)$ into $P(x_0)$, $P(x_1)$, and $P(x_0, x_1)$ in the definitions of entropy. In this case, the estimates of entropy and MI are biased due to finite samples (Li, 1990; Arora et al., 2022).

To reduce this bias, we employ the more sophisticated estimator proposed by Grassberger (2003). In this method, the entropy S_0 is estimated by

$$\hat{S}_0 = \psi(N) - \frac{1}{N} \sum_{x_0} n_{x_0} \psi(n_{x_0}), \quad (3)$$

where ψ is the digamma function, n_{x_0} is the number of samples such that the former tag is x_0, and $N = \sum_{x_0} n_{x_0}$ is the total number of samples. S_1 and S_{01} are estimated in a similar manner. Grassberger (2003) has proven that the bias of this estimator rapidly converges to zero as the number of samples, N, increases. Therefore, if the estimated value saturates with respect to N, the finite-sample bias should be negligible.

To estimate the MI in sequences, we sample N_{data} pairs of POS nodes separated by a sequential distance r_{seq}, from the dataset. For the MI in structures, we sample N_{data} pairs of nodes at a fixed structural distance r_{str}, where the nodes are either POS or phrasal. We treat ordered pairs as distinct to avoid arbitrarily selecting one of the two. For example, we count pairs (AUX, PP) and (PP, AUX), corresponding to *were* and *than*, in Figure 2 as distinct to each other if both pairs are sampled.

In the following analyses, we compute the MI estimates while varying the number of samples, N_{data}. In our results, the estimates saturate as N_{data} increases, indicating that the bias is sufficiently small.

To further support our claim, we also report the results obtained with six alternative entropy estimators (Madow, 1948; Miller, 1955; Zahl, 1977; Horvitz and Thompson, 1952; Chao and Shen, 2003; Wolpert and Wolf, 1995; Nemenman et al., 2001) in Appendix D. The results are consistent with those reported in the main text, except for the

estimators that were previously reported to exhibit larger errors (Arora et al., 2022).

3.3 Growth of Sequential Distance

To address question (ii), we investigate how the sequential distance depends on the structural distance. Specifically, for each r_{seq} and r_{str}, we count the number of POS-tagged node pairs over the whole treebank whose sequential and structural distances are r_{seq} and r_{str}, respectively. We also compute the average of sequential distance for each structural distance r_{str}.

3.4 Context-free Independence Breaking

To address question (iii), it is necessary to quantify the extent to which syntactic structures deviate from PCFGs. We employ a metric for this quantification, called *context-free independence breaking* (CFIB; Nakaishi and Hukushima, 2024).

The essential property of PCFGs, which we refer to as the *context-free independence*, is that the distribution of categories of each node depends only on its parent. Consequently, given the categories of two nodes, their children are mutually independent.

Therefore, we can quantify the deviation of the distribution of trees from PCFGs by the MI between two pairs under the condition that the categories of their parents are fixed, as presented in Figure 3. Specifically, provided that the categories of two nodes are x_0 and x_1, the CFIB is defined as

$$J_{x_0x_1} = \sum_{y_0,z_0,z_0,z_1} P(y_0,z_0,y_1,z_1|x_0,x_1) \times \ln \frac{P(y_0,z_0,y_1,z_1|x_0,x_1)}{P(y_0,z_0|x_0,x_1)P(y_1,z_1|x_0,x_1)}. \quad (4)$$

Here, $P(y_0,z_0|x_0,x_1)$ denotes the conditional probability that the left and right children of the node with label x_1 are y_0 and z_0, respectively. Similarly, $P(y_1,z_1|x_0,x_1)$ refers to those for the node with label x_1. $P(y_0,z_0,y_1,z_1|x_0,x_1)$ is the conditional joint probability of both pairs of children. By definition, the CFIB is non-negative for any tree structures and zero for trees generated by PCFGs.

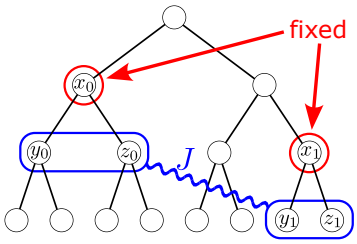

Figure 3: CFIB J is the MI between the children of two nodes whose categories are fixed.

The CFIB is defined for the distribution of trees, as is the MI. We can estimate this metric from the dataset of trees, similarly to the estimation of the MI. We sample N_{data} pairs of nodes that are tagged as x_0 and x_1, with the structural distance of r_{str}. Then, we compute the empirical distributions of tags on the children of the nodes. From these distributions, we estimate J in the same procedure explained in Section 3.2. We also distinguish between pairs and their permutations, as in the estimation of MI.

We verify the accuracy of the estimates for the CFIB by varying N_{data}, as we did for the MI. Appendix D also presents the results obtained with alternative entropy estimators, further supporting the validity of our estimates.

3.5 PCFG Approximation

We also construct a PCFG that approximates the distribution of natural language parse trees. This allows for a more detailed comparison between syntactic structures and PCFGs. The PCFG is constructed using the maximum likelihood estimation, which is equivalent to simply counting the frequency of each production rule in the dataset. Using the PCFG derived from the preprocessed dataset from BLLIP, we generate trees 1.6×10^7 times, with up to 100 iterations per generation. This procedure yields approximately 13 million terminated trees. We then apply the same analyses to these generated trees as to parse trees.

3.6 Fitting

We characterize the decay of correlation by fitting the data with exponential and power-law functions. Our goal is not to fit the data as accurately as possible or to estimate parameter values, but to test whether the observed behavior is closer to a power law or to an exponential form. For this purpose, using simpler models with fewer parameters is more appropriate, as it avoids issues such as overfitting. Accordingly, we use the two simple models, $A\exp(-\lambda r)$ and $Br^{-\alpha}$.

To perform the fitting, we apply the least squares method after log-scaling the MI and CFIB values, because fitting with a linear scale puts too much weight on data points at shorter distances.

In model selection, model complexity should be taken into account, because goodness of fit, such as likelihood or squared error, trivially improves as the number of parameters increases. However, in our case, both models have two parameters. Therefore, it is sufficient to compare goodness of fit. Specifically, we use the reduced chi-square statistic, χ^2_ν. Under the commonly used assumption of Gaussian noise, comparing this statistic is equivalent to using standard model selection criteria such as the Akaike information criterion (Akaike, 1974) and the Bayesian information criterion (Schwarz, 1978).

We also fit the growth of sequential distance using an exponential function, $C\exp(\mu r)$, and a power-law function, Dr^β, without log-scaling the values.

For fitting, we use the `lmfit` Python package (Newville et al., 2014). In the following results, we present the fitted parameters A, λ, and others along with the reduced chi-square statistic, χ^2_ν.

3.7 Synthetic Data

To demonstrate that our method can reliably estimate MI and distinguish between exponential and power-law decays, we apply it to mathematical models in which MI can be computed exactly. We generate random variables X and Y with joint probabilities defined as $P(X=0,Y=0) = P(X=1,Y=1) = (1+\delta)/4$ and $P(X=0,Y=1) = P(X=1,Y=0) = (1-\delta)/4$. The MI between X and Y is $I = \frac{1+\delta}{2}\ln(1+\delta) + \frac{1-\delta}{2}\ln(1-\delta)$, which is $\delta^2/2 + \mathcal{O}(\delta^3)$ as $\delta \to 0$. If $\delta = \exp(-\lambda r/2)$, the MI decays asymptotically as $I \sim \exp(-\lambda r)$. If $\delta = r^{-\alpha/2}$, the decay follows a power law, $I \sim r^{-\alpha}$. We refer to these as the exponential and power-law models, respectively.

We sample X and Y from the exponential model with $\lambda = 0.1$ and from the power-law model with $\alpha = 2$. We then estimate the MI and fit the resulting values. The results are shown in Figure 4. In both cases, the estimates converge to the exact values as the sample size N_{data} increases. In addition, χ^2_ν is significantly smaller when the fitting function matches the model. Our method successfully discriminates between the exponential and power-law models.

4 Results

4.1 Mutual Information in Sequences

First, we confirm that the correlation in sequences decays according to a power law. Figure 5 shows the MI I_{POS} between POS tags as a function of the sequential distance r_{seq} for different data sizes N_{data}. The estimated values converge as N_{data} increases, indicating that the estimation bias is sufficiently small.

The decay of I_{POS} is better fitted by a power-law function ($\chi^2_\nu = 2.6\times10^{-1}$) than by an exponential one ($\chi^2_\nu = 9.8\times10^{-1}$). This result indicates that the correlation between POS tags follows a power law, consistent with previous research.

4.2 Mutual Information in Structures

We now address question (i): whether the exponential decay of correlation with the structural distance holds in natural language syntactic structures. The estimated MI between POS or phrasal tags as a function of the structural distance is shown in Figure 6. As in the previous section, the convergence of the estimated values with increasing N_{data} indicates that the bias is negligible.

Model fitting reveals that the correlation decays according to a power law ($\chi^2_\nu = 1.1\times10^{-1}$) rather than an exponential form ($\chi^2_\nu = 1.3$) with respect to the structural distance. This result indicates that syntactic structures do not meet the condition necessary for the proposed argument to hold.

4.3 Growth of Sequential Distance

Question (ii) asks whether the sequential distance grows exponentially with increasing structural distance. To examine this, we count the number of POS tag pairs such that the structural and sequential distances are r_{str} and r_{seq}, respectively. The results are shown in Figure 7. We also compute the average sequential distance for each structural distance, represented by black circles, and fit the results using exponential and power-law functions.

The fitting reveals that the growth follows a power law ($\chi^2_\nu = 1.4\times10$) rather than an exponential form ($\chi^2_\nu = 1.4\times10^{21}$), although the trend becomes less clear for larger structural distances $r_{\text{str}} \gtrsim 50$ because of data sparsity. Notably, the exponent $\beta = 0.72$ is smaller than 1, indicating that the growth is slower than linear. As demonstrated in Figure 1 (b), this slow growth quantitatively reflects a strong branching bias in syntactic structures. These findings clearly show that the second

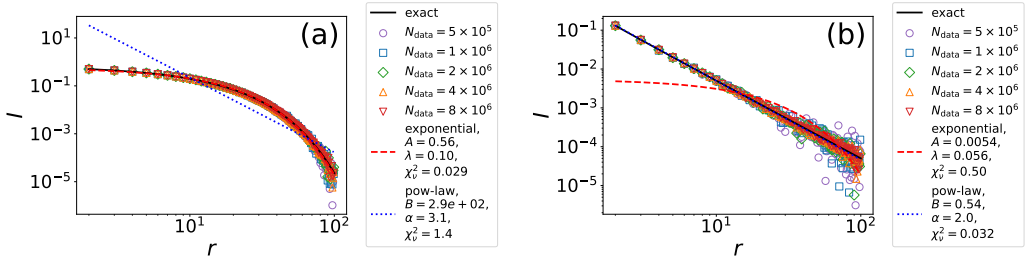

Figure 4: Estimation and fitting of the MI for (a) the exponential model with $\lambda = 0.1$ and (b) the power-law model with $\alpha = 2$. Fitting was performed for $N_{\text{data}} = 8 \times 10^6$.

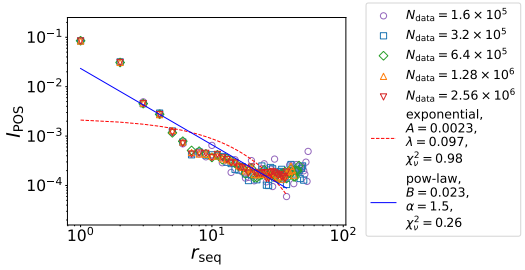

Figure 5: MI I_{POS} between POS tags as a function of the sequential distance r_{seq}. The fitted exponential and power-law decays for $N_{\text{data}} = 2.56 \times 10^6$ are also presented.

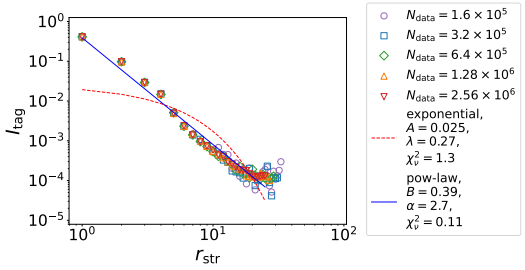

Figure 6: MI I_{tag} between POS or phrasal tags as a function of the structural distance r_{str}. The fitted exponential and power-law decays for $N_{\text{data}} = 2.56 \times 10^6$ are also presented.

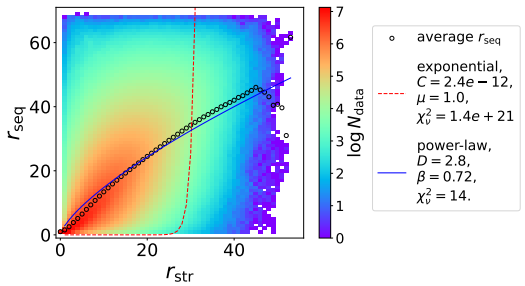

Figure 7: Number of POS tag pairs such that the structural and sequential distances are r_{str} and r_{seq}, respectively. Black circles represent the average sequential distance for each structural distance.

condition does not hold for syntactic structures.

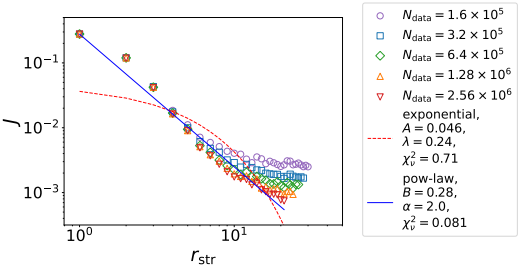

Figure 8: CFIB J for $(x_0, x_1) = (\text{NP}, \text{NP})$ as a function of the structural distance r_{str}. The fitted exponential and power-law decays for $N_{\text{data}} = 2.56 \times 10^6$ are also presented.

4.4 Context-free Independence Breaking

Question (iii) concerns whether the statistical properties of syntactic structures deviate significantly from those of PCFGs. We measure the CFIB J as a metric of this deviation. Since reliable estimation of this metric requires a sufficiently large number of samples, we focus on the three most frequent cases, namely the fixed node pairs $(x_0, x_1) = (\text{NP}, \text{NP})$, (NP, VP), and (VP, VP).

The result for (NP, NP) is shown in Figure 8. The convergence of the estimates with increasing N_{data} confirms that the estimation bias is negligible. The CFIB values are clearly positive, meaning that syntactic structures do not have the essential property of PCFGs. Furthermore, the metric decays according to a power law ($\chi^2_\nu = 8.1 \times 10^{-2}$) rather than an exponential form ($\chi^2_\nu = 7.1 \times 10^{-1}$) with respect to the structural distance, suggesting that the context-free independence is significantly broken even at large distances. These results indicate that the answer to question (iii) is negative.

This finding is also consistent with previous studies that developed PCFG-based parsers achieving high performance (Collins, 2003; Johnson et al., 2006; O'Donnell et al., 2009). These studies augmented PCFGs to relax the context-

free independence implicitly or explicitly.

As shown in Appendix E, similar behaviors are observed for other frequent node pairs $(x_0, x_1) = (\mathrm{NP}, \mathrm{VP})$ and $(x_0, x_1) = (\mathrm{VP}, \mathrm{VP})$, suggesting that the observed deviation from PCFGs is not specific to a particular syntactic category pair.

4.5 Other Corpora

We examined whether the three assumptions in the argument of Lin and Tegmark (2017) hold for BLLIP. The results show that all of these assumptions are violated. The argument cannot account for the statistical properties of parse trees in this corpus. To test whether this conclusion is specific to BLLIP, we conduct the same analyses on WikiText and NPCMJ.

Figure 9 summarizes the results. For both corpora, the statistical properties are similar to those observed for BLLIP: the MI decays according to a power law in both sequences and structures; the growth of sequential distance is slower than linear; and the CFIB also decays following a power law. These behaviors are supported by the reduced chi-square statistics, which are smaller for power-law fitting than for exponential fitting across all statistical quantities.

The results also exhibit slight differences from those for BLLIP. I_{POS} and I_{tag} for WikiText, as well as I_{POS} for NPCMJ, increase at large distances. This effect may be attributable to structural differences in unusually long sentences or to potential parsing errors in such cases. However, we cannot determine the significance of this behavior because the datasets contain too few long sentences.

Nonetheless, the results clearly support that the assumptions proposed by Lin and Tegmark (2017) are violated not only for BLLIP, but also for WikiText and for NPCMJ. These results imply the robustness of our conclusion, given that Japanese is typologically distinct from English, exhibiting head-final, left-branching syntactic structures.

4.6 PCFG Approximation

To further elucidate the difference between syntactic structures and PCFGs, we apply the same analyses to the PCFG constructed using maximum likelihood estimation on BLLIP. The results are summarized in Figure 10.

Mutual Information in Sequences Figure 10 (a) shows the MI in sequences. At short sequential distances, $r_{\mathrm{seq}} \lesssim 10$, the MI decays according to a power law, although the decay rate does not clearly match that observed in parse trees. At larger distances, $r_{\mathrm{seq}} \gtrsim 10$, the MI increases substantially, in contrast to the trend in parse trees.

Mutual Information in Structures The MI in structures, shown in Figure 10 (b), exhibits a power-law decay similar to that in syntactic structures, although the decay rate differs. This result does not contradict the properties of PCFGs proven by Lin and Tegmark (2017). While the MI in structures generated by PCFGs asymptotically follows an exponential decay at sufficiently large distances, the MI can exhibit different behavior at shorter distances. The non-exponential behavior of I_{tag} suggests that for individual trees, the structural distances are too short for the asymptotic behavior to emerge.

This pre-asymptotic behavior in the PCFG constructed from parse trees raises the possibility that the statistical behavior of parse trees themselves is also pre-asymptotic. If so, it also calls into question the applicability of the argument of Lin and Tegmark (2017) to natural language parse trees, since the argument relies on asymptotic behavior.

This observation also has an important implication about the applicability of their argument to domains beyond texts and speech by human adults. As discussed in Section 1, several studies have applied their argument to child speech and animal behavior, interpreting the power-law decay of correlations in these domains as evidence for the existence of some form of underlying hierarchical structures (Sainburg et al., 2019; Sainburg and Gentner, 2021; Sainburg et al., 2022; Howard-Spink et al., 2024; Youngblood, 2024).

However, in these domains, the sequence lengths are typically comparable to those in the PCFG studied here. In such cases, it is difficult to interpret the power-law decay of correlation as evidence for hierarchical structures, since the asymptotic behavior is unlikely to emerge.

Growth of Sequential Distance In Figure 10 (c), the average sequential distance grows more slowly than exponentially and even than linearly, indicating that the generated trees are strongly biased. While this behavior is qualitatively similar to that in parse trees, the estimated exponent $\beta = 0.41$ significantly differs from that

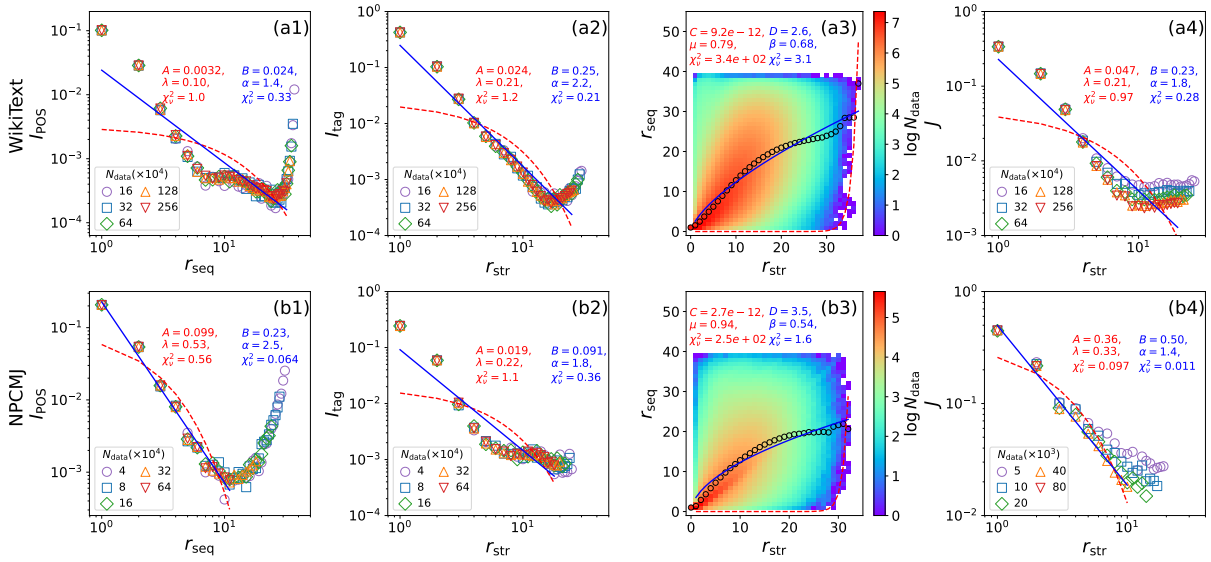

Figure 9: Statistical properties of syntactic structures in WikiText and NPCMJ. (a1–4) Results for WikiText: (a1) I_{POS} as a function of r_{seq}; (a2) I_{tag} as a function of r_{str}; (a3) relationship between r_{seq} and r_{str}; and (a4) CFIB J for $(x_0, x_1) = (\mathrm{NP}, \mathrm{NP})$ as a function of r_{str}. Fitting was performed with $N_{\mathrm{data}} = 2.56 \times 10^6$. (b1–b4) Corresponding results for NPCMJ. Fitting was performed with $N_{\mathrm{data}} = 6.4 \times 10^5$ for I_{POS} and I_{tag}, and with $N_{\mathrm{data}} = 4.0 \times 10^4$ for J.

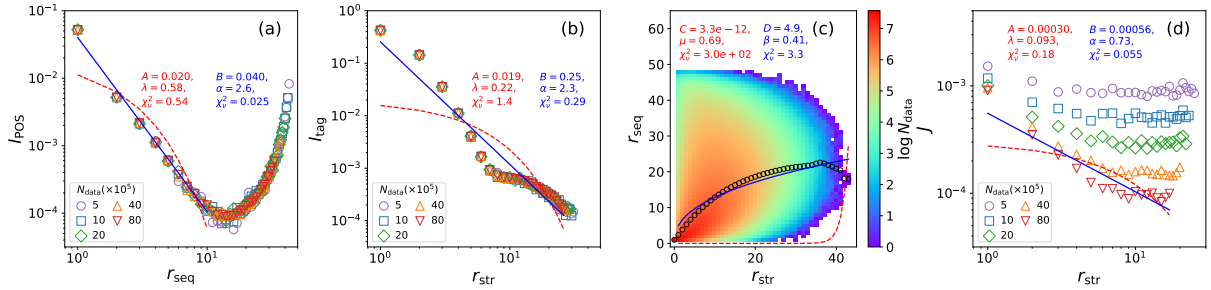

Figure 10: Statistical properties of trees generated by the PCFG obtained from the maximum likelihood estimation of syntactic structures. (a) MI I_{POS} between POS tags. (b) MI I_{tag} between POS or phrasal tags. (c) Distribution and average of sequential distances. (d) CFIB J for $(x_0, x_1) = (\mathrm{NP}, \mathrm{NP})$. Fitting was performed for $N_{\mathrm{data}} = 8 \times 10^6$ in (a) and (b), and for $r_{\mathrm{str}} \leq 10$ in (a).

for the parse trees, where $\beta = 0.72$.

Context-free Independence Breaking The most striking difference between parse trees and trees generated by the PCFG appears in the CFIB metric. The true value of this metric is zero for PCFGs by definition. If the estimates converge to this true value with increasing data size, the corresponding values in the log-scale plot continue to decrease, approaching $-\infty$, since the logarithm of zero is $-\infty$.

In Figure 10 (d), the estimates of CFIB is much smaller than those for syntactic structures in Figure 8. Furthermore, the estimates continue to decrease in the log-scale plot as N_{data} increases, which is the expected behavior. These results are in contrast to those for parse trees.

5 Discussion and Conclusion

In natural languages, the correlation decays as a power law with respect to the sequential distance, as reported in numerous previous studies. At the same time, linguists generally agree that natural languages exhibit hierarchical syntactic structures. Although both phenomena are well established, their relationship remains elusive.

The present study examines this relationship, specifically the argument proposed by Lin and Tegmark (2017). They attributed the power-law decay of correlation in natural languages to underlying hierarchical structures, which has since inspired numerous studies on both human languages

and animal behavior. However, their argument relies on several assumptions that are questionable from a linguistic perspective. To examine the validity of the assumptions for syntactic structures, we test the three key questions through statistical analyses of parse trees in natural languages.

The analyses yield negative answers to all three questions: (i) The correlation in syntactic structures decays according to a power law with respect to the structural distance; (ii) The growth of the sequential distance is slower than linear with the structural distance, due to the strong branching bias in syntactic structures; (iii) The statistical properties of syntactic structures significantly differ from those of PCFGs. These results reveal that the relationship between the power-law decay of correlation and syntactic structures differs from that proposed by Lin and Tegmark (2017).

We obtained similar results under various settings: for the English treebank BLLIP with several different schemes, for another English corpus, WikiText, and for the Japanese treebank NPCMJ. To further assess the robustness of our conclusions, it is important to examine statistical behavior under additional annotation schemes, domains, and languages. In particular, revealing the extent to which these behaviors are shared across languages is important from a linguistic perspective.

We also analyzed the PCFG obtained from the maximum likelihood estimation of BLLIP. The statistical properties of the PCFG clearly differ from those in parse trees. Furthermore, this analysis demonstrates that the asymptotic behavior of the PCFG does not emerge within individual sequences. This implies the difficulty of applying the proposed argument not only to sentences produced by human adults but also to child speech and animal behavior, where sequence lengths are comparable to those for the PCFG.

In addition, the statistical differences between natural language parse trees and trees generated by PCFGs may have implications for theoretical research on large language models. Some studies in this field analyze models trained on synthetic data, in which PCFGs are sometimes used to generate the data (Allen-Zhu and Li, 2023; Cagnetta and Wyart, 2024). However, the model's ability to learn PCFGs will not account for its ability to learn syntactic structures of natural languages, although some studies employ PCFGs as models of more abstract structures than syntax.

It is noteworthy that the pre-asymptotic behavior in the PCFG constructed from parse trees does not serve for a definitive evidence against the validity of the argument by Lin and Tegmark (2017), because this behavior does not necessarily mean that parse trees themselves also exhibit pre-asymptotic behavior. Our main evidence instead come from the direct analyses of parse trees in Sections 4.1–4.5, which clearly show that the statistical properties of natural language parse trees for their typical sizes differ from those assumed in the proposed argument.

Currently, we cannot rule out the possibility that the MI in natural language syntactic structures follows an exponential decay at structural distances much larger than those observed in this study. However, even if this were the case, it would not explain the power-law decay of correlation in natural language sequences, because such large parse trees would be too rare to have a significant influence on the statistical properties of the corpus. Moreover, as long as syntactic structures are biased, the sequential distance grows more slowly than exponentially, resulting in a decay of MI in sequences that is faster than a power law. Therefore, statistical analyses of larger parse trees are unlikely to account for the power-law decay of correlation observed in natural languages.

A more promising direction is to consider structures larger than individual sentences, namely, the hierarchical organization of entire texts composed of multiple sentences. The argument by Lin and Tegmark (2017) may account for the power-law decay of correlation observed across entire texts rather than within single sentences. As a first step in this direction, it will be necessary to develop sufficiently large-scale corpora annotated with discourse information to estimate statistical quantities such as I_{POS}, I_{tag}, and J.

Another direction is to explore approaches different from that of Lin and Tegmark (2017). One alternative is *critical phenomena* in statistical physics. These refer to the phenomena that statistical quantities exhibit power-law behaviors near *phase transition* points, at which statistical properties of a system change qualitatively. Indeed, as the temperature parameter of large language models varies, generated texts exhibit critical phenomena, sharing statistical properties with natural languages, including the power-law decay of correlation (Nakaishi et al., 2024).

Acknowledgments

We thank the action editor, Carlos Gómez-Rodríguez, and the anonymous reviewers for their constructive and helpful feedback. We also thank Taiga Ishii, Shinnosuke Isono, Sho Yokoi, and Keisuke Sakaguchi for valuable discussions and suggestions. This work was supported by JSPS KAKENHI Grant Nos. 23H01095, 23KJ0622, 24H00087, and 25K24434; JST, PRESTO Grant No. JPMJPR21C2; and the World-Leading Innovative Graduate Study Program for Advanced Basic Science Course at the University of Tokyo.

References

Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 1, learning hierarchical language structures. *Available at SSRN*.

Eduardo G. Altmann, Giampaolo Cristadoro, and Mirko Degli Esposti. 2012. On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, 109(29):11582–11587.

Enrique Alvarez-Lacalle, Beate Dorow, Jean-Pierre Eckmann, and Elisha Moses. 2006. Hierarchical structures induce long-range dynamical correlations in written texts. *Proceedings of the National Academy of Sciences*, 103(21):7956–7961.

Aryaman Arora, Clara Meister, and Ryan Cotterell. 2022. Estimating the entropy of linguistic distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Dublin, Ireland. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Francesco Cagnetta and Matthieu Wyart. 2024. Towards a theory of how the structure of language is acquired by deep neural networks. In *Advances in Neural Information Processing Systems 37*.

Anne Chao and Tsung-Jen Shen. 2003. Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. *Environmental and ecological statistics*, 10(4):429–443.

Eugene Charniak. 1997. Statistical techniques for natural language parsing. *The AI Magazine*, 18:33–44.

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 2000. BLLIP 1987-89 WSJCorpus Release 1 LDC2000T43. *Linguistic Data Consortium*.

Noam Chomsky. 1956. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637.

Matthew S. Dryer. 1992. The greenbergian word order correlations. *Language*, 68(1):81–138.

Werner Ebeling and Alexander Neiman. 1995. Long-range correlations between letters and sentences in texts. *Physica A*, 215(3):233–241.

Werner Ebeling and Thorsten Pöschel. 1994. Entropy and Long-Range correlations in literary english. *Europhysics Letters*, 26(4):241–246.

Kevin Ellis, Armando Solar-Lezama, and Josh Tenenbaum. 2015. Unsupervised learning by program synthesis. In *Advances in Neural Information Processing Systems 28*.

Édouard Gilbert and Darrell Conklin. 2007. A probabilistic context-free grammar for melodic reduction? In *Proceedings of the International Workshop on Artificial Intelligence and Music, 20th International Joint Conference on Artificial Intelligence*.

Peter Grassberger. 2003. Entropy estimates from insufficient samplings. arXiv:physics/0307138v2.

Daniel Harlow, Stephen H. Shenker, Douglas Stanford, and Leonard Susskind. 2012. Tree-like structure of eternal inflation: A solvable model. *Physical Review D*, 85(6):063516.

John A. Hawkins. 1990. A parsing theory of word order universals. *Linguistic inquiry*, 21(2):223–261.

Harold Stanley Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Daniel G. Horvitz and Donovan J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Elliot Howard-Spink, Misato Hayashi, Tetsuro Matsuzawa, Daniel Schofield, Thibaud Gruber, and Dora Biro. 2024. Nonadjacent dependencies and sequential structure of chimpanzee action during a natural tool-use task. *PeerJ*, 12:e18484.

Filip Jelinek, John D. Lafferty, and Robert L. Mercer. 1992. Basic methods of probabilistic context free grammars. In *Speech Recognition and Understanding*.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *Advances in Neural Information Processing Systems 19*.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Bjarne Knudsen and Jotun Hein. 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–454.

Wentian Li. 1989a. Mutual information functions of natural language texts. Technical Report 89-10-008, Santa Fe Institute. Accessed 2026-03-02.

Wentian Li. 1989b. Spatial 1/f spectra in open dynamical systems. *Europhysics Letters*, 10(5):395.

Wentian Li. 1990. Mutual information functions versus correlation functions. *Journal of statistical physics*, 60:823–837.

Wentian Li. 1991. Expansion-modification systems: a model for spatial 1/f spectra. *Physical Review A*, 43(10):5240.

Henry W. Lin and Max Tegmark. 2017. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299.

William G. Madow. 1948. On the limiting distributions of estimates based on samples from finite universes. *The Annals of Mathematical Statistics*, pages 535–545.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Nikolay Mikhaylovskiy and Ilya Churilov. 2023. Autocorrelations decay in texts and applicability limits of language models. arXiv:2305.06615v1.

George Miller. 1955. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*.

Kai Nakaishi and Koji Hukushima. 2024. Statistical properties of probabilistic context-sensitive grammars. *Physical Review Research*, 6(3):033216.

Kai Nakaishi, Yoshihiko Nishikawa, and Koji Hukushima. 2024. Critical phase transition in large language models. arXiv:2406.05335v2.

National Institute for Japanese Language and Linguistics. 2016. Ninjal parsed corpus of modern japanese. Version 1.0. Available at https://npcmj.ninjal.ac.jp/interfaces/ (accessed 2025-10-29). Licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Ilya Nemenman, Fariel Shafee, and William Bialek. 2001. Entropy and inference, revisited. *Advances in Neural Information Processing Systems 14*.

Matthew Newville, Till Stensitzki, Daniel B. Allen, and Antonino Ingargiola. 2014. Lmfit: Non-linear least-square minimization and curve-fitting for python.

Timothy J. O'Donnell, Joshua B. Tenenbaum, and Noah D. Goodman. 2009. Fragment grammars: Exploring computation and reuse in language. Technical Report MIT-CSAIL-TR-2009-013, Massachusetts Institute of Technology. Accessed 2026-03-02.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse TreeBank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.

Tim Sainburg and Timothy Q. Gentner. 2021. Toward a computational neuroethology of vocal communication: from bioacoustics to neurophysiology, emerging tools and future directions. *Frontiers in Behavioral Neuroscience*, 15:811737.

Tim Sainburg, Anna Mai, and Timothy Q. Gentner. 2022. Long-range sequential dependencies precede complex syntactic production in language acquisition. *Proceedings of the Royal Society B*, 289(1970):20212657.

Tim Sainburg, Brad Theilman, Marvin Thielk, and Timothy Q. Gentner. 2019. Parallels in the sequential organization of birdsong and human speech. *Nature Communications*, 10(1):3636.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Huitao Shen. 2019. Mutual information scaling and expressive power of sequence models. arXiv:1905.04271v1.

Harry E. Stanley. 1971. *Introduction to Phase Transitions and Critical Phenomena*. Oxford University Press.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2017. Do neural nets learn statistical laws behind natural language? *PloS one*, 12(12):e0189326.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2019. Evaluating computational language models with scaling properties of natural language. *Computational Linguistics*, 45(3):481–513.

Kumiko Tanaka-Ishii and Armin Bunde. 2016. Long-range memory in literary texts: On the universal clustering of the rare words. *PLoS One*, 11(11):e0164658.

Pablo Tano, Sergio Romano, Mariano Sigman, Alejo Salles, and Santiago Figueira. 2020. Towards a more flexible language of thought: Bayesian grammar updates after each concept exposure. *Physical Review E*, 101(4):042128.

David H. Wolpert and David R. Wolf. 1995. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841.

Peter Worth and Susan Stepney. 2005. Growing music: Musical interpretations of l-systems. In *Applications of Evolutionary Computing*, pages 545–550.

Peter Worth and Susan Stepney. 2021. Recursive bayesian networks: Generalising and unifying probabilistic context-free grammars and dynamic bayesian networks. In *Advances in Neural Information Processing Systems 34*, pages 4370–4383.

Mason Youngblood. 2024. Language-like efficiency and structure in house finch song. *Proceedings of the Royal Society B*, 291(2020):20240250.

Samuel Zahl. 1977. Jackknifing an index of diversity. *Ecology*, 58(4):907–913.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, 51(1):23–72.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press.

A Tree Statistics

To provide intuition about both the effect of preprocessing on tree shapes and the differences between English and Japanese, we compute basic tree-shape statistics on the original parse trees in BLLIP and NPCMJ. Specifically, we consider the following statistics:

Depth. The maximum path length from the root to any node in a tree, averaged over all parse trees.

Branching Factor. The number of children of an internal node, averaged over all internal nodes. This value ranges from 1 upward and equals 2 for strictly binary-branching trees without unary nodes.

Proportion of Unary Nodes. An internal node is *unary* if it has exactly one child. We report the proportion of unary nodes, defined as the number of unary internal nodes divided by the total number of internal nodes, excluding POS-to-terminal productions.

As shown in Table 1, the branching factor and the proportion of unary nodes are comparable between BLLIP and NPCMJ. In contrast, the depth for NPCMJ is smaller than that for BLLIP, suggesting that parse trees in NPCMJ tend to be flatter.

B Tag Reduction

To reduce the number of categories, we grouped tags with similar functions into the same category. The eleven categories used for BLLIP and WikiText and the seven categories for NPCMJ are shown in Tables 2 and 3, respectively.

C Other Schemes

In the main text, we analyzed parse trees after applying two preprocessing steps. First, we binarized the parse trees. Second, we reduced the number of tag types by grouping similar tags, where we use the same reduced categories for both POS and phrasal tags.

Although the statistical quantities depend on the specific preprocessing scheme, we expect that these details do not affect our conclusions, such as whether the MI decays according to a power law or an exponential function, and whether the CFIB is significantly positive.

To examine this expectation, we perform additional analyses on BLLIP under two alternative preprocessing schemes: the unbinarized setting and the phrasal-tag-only setting, in which POS and phrasal tags are distinguished and only the latter are analyzed.

In addition, we analyze unbinarized trees in NPCMJ. Because the original trees in NPCMJ are flatter than those in BLLIP, as shown in Appendix A, the effect of binarization could be larger.

Figures 11 (a1) and (a2) show the results for BLLIP under the unbinarized setting. Because binarization does not affect the MI in sequences and the dataset is too limited to reliably estimate the CFIB for unbinarized trees, we present only the MI in structures and the sequential distance as a function of the structural distance. For both quantities, the results are qualitatively similar to those for the binarized setting in the main text: the decay of I_{tag} follows a power-law function rather than an exponential function, and the growth of r_{seq} is slower than exponential.

In the phrasal-tag-only setting, the MI between POS tags and the sequential distance are not defined. We therefore estimate the MI between phrasal tags and the CFIB, presented in Figures 11 (b1) and (b2), respectively. Both quantities exhibit power-law decay, consistent with the results in the main text.

The MI in structures and the sequential distance for unbinarized trees in NPCMJ are presented in Figures 11 (c1) and (c2). Again, the results are similar to those in other cases. Both quantities display power-law dependence on the structural distance rather than exponential behavior.

Corpus	Depth	Branching	Prop. Unary
BLLIP	$11.598(\pm 4.354)$	$1.527(\pm 0.904)$	0.222
NPCMJ	$6.642(\pm 3.336)$	$1.547(\pm 1.282)$	0.271

Table 1: Tree-shape statistics computed on the original parse trees in BLLIP and NPCMJ. The depth and the branching factor are reported as mean $\pm$ standard deviation across sentences.

Tags	Part-of-Speech Tags and Phrase Labels
NP	NP NN NNP NNS CD PRP TMP POS QP NNPS EX NX FW
VP	VP VBD VB VBN VBG VBZ VBP TO
S	S S1 SBAR SINV FRAG SQ SBARQ RRC
PP	IN PP RP PRT
DT	DT PRP$ PDT
JJ	JJ ADJP JJR JJS NAC ADJ
Others	, . ” “ $: PRN -RRB- -LRB- # X INTJ UH SYM LS LST HYPH NFP
AUX	AUX MD AUXG
RB	RB ADVP RBR RBS
CC	CC UCP CONJP
WH	WHNP WDT WP WHADVP WRB WHPP WP$ WHADJP NML

Table 2: Eleven tags used for BLLIP and WikiText and the corresponding part-of-speech tags and phrase labels.

Tags	Part-of-Speech Tags and Phrase Labels
NP	NP NUMCLP PRN NML PNLP NUMCLPSYM N NPR NUM CL ADJN PRO ADJI D Q FN WPRO PNL PRN WNUM WD FW
IP	IP VB AX AXD VB0 VB2 PASS PASS2 MD
PP	PP P
ADVP	ADVP ADV NEG WADV
CONJP	CONJP CONJ
PP	PP P
CP	CP FRAG INTJP FS INTJ
Others	LST META LS PU PUL PUR SYM PUQ COMMENT

Table 3: Seven tags used for NPCMJ and the corresponding part-of-speech tags and phrase labels.

Across all alternative preprocessing schemes, our main conclusion holds: the assumptions in the argument by Lin and Tegmark (2017) are not satisfied. These results indicate that our findings are robust to the details of preprocessing.

D Alternative Estimators

In the main text, we computed the MI using the estimator of Grassberger (2003) (GR), whose bias decreases rapidly. The estimated values saturate as N_{data} increases, except for the CFIB in the PCFG. The estimates of the CFIB in the PCFG continue to decrease in the log-scale plot as N_{data} increases, which is reasonable because the true value is zero. These results show that we have properly estimated the statistical quantities.

To further support the reliability of our estimates, we compute I_{POS}, I_{tag}, and J using six alternative estimators proposed by Chao and Shen (2003) (CS), Horvitz and Thompson (1952) (HT), (Miller, 1955) and Madow (1948) (MM), Nemenman et al. (2001) (NSB), Wolpert and Wolf (1995) (WW) and Zahl (1977) (ZA). For these computations, we use the Python code `entropy-estimation` implemented by Arora et al. (2022).

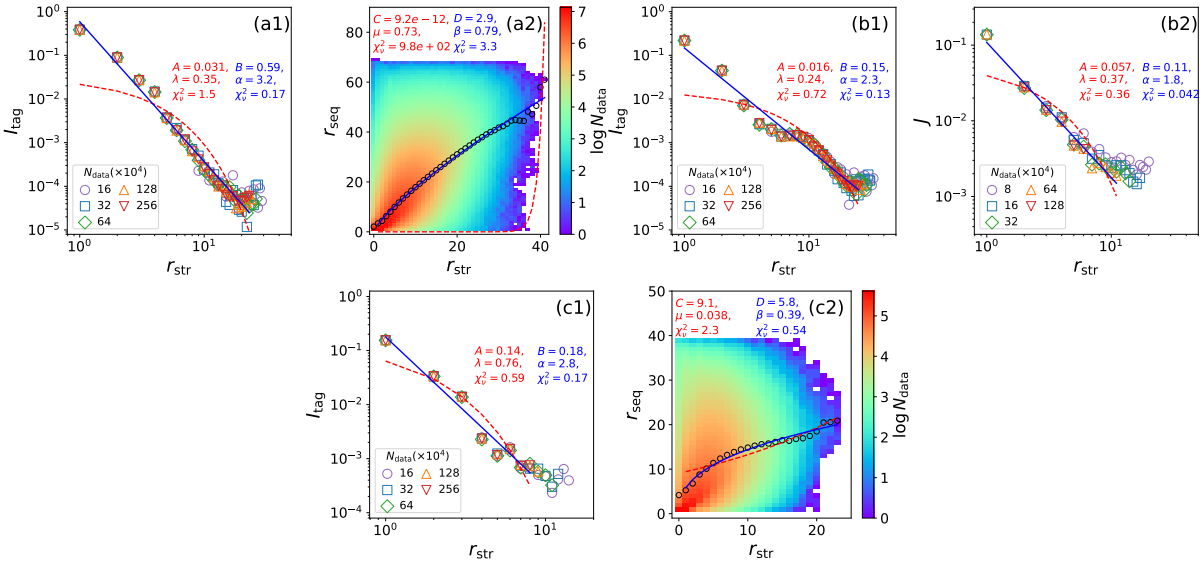

Figure 11: (a1, 2) Results for BLLIP under the unbinarized setting: (a1) MI I_{tag} between POS or phrasal tags, where fitting was performed for $N_{\text{data}} = 2.56 \times 10^6$; (a2) distribution and average of sequential distances. (b1, 2) Results for BLLIP under the phrasal-tag-only setting: (b1) MI I_{tag} between phrasal tags, where fitting was performed for $N_{\text{data}} = 2.56 \times 10^6$; (a2) CFIB J, where fitting was pwerformed for $N_{\text{data}} = 6.4 \times 10^5$. (c1, 2) Results for NPCMJ under the unbinarized setting: (a1) MI I_{tag} between POS or phrasal tags, where fitting was performed for $N_{\text{data}} = 2.56 \times 10^6$; (a2) distribution and average of sequential distances.

Figure 12 presents the results for BLLIP computed with the six alternative estimators, together with those obtained using the GR estimator, which are the same as in the main text. The estimates agree with ours in most cases. Figures 12 (b) and (c) show that the WW estimates for I_{tag} and J at large distances exhibit clear bias, producing some negative estimates even though the true MI and CFIB are always non-negative. The CS and HT estimates for J also exhibit similar behaviors. These results are consistent with Arora et al. (2022), which reports that these three estimators yield larger errors than the others when the sample size is large. From these results, we can confirm that our estimates in the main text are reliable.

Figure 13 shows the results for the same quantities for the PCFG obtained from the maximum likelihood estimation of BLLIP. The estimates of I_{POS} and I_{tag}, shown in Figures 13 (a) and (b), are consistent across all estimators. This again supports the reliability of the results in the main text.

We also show the estimates of the CFIB J for the PCFG obtained with the MM, NSB, and ZA estimators in Figures 13 (c), (d), and (e), respectively. The MM and ZA estimates continue to decrease in the log-scale plot as N_{data} increases, similar to the GR estimates in the main text. This means that these two estimates properly converge to zero, the true value of the CFIB in PCFGs. The convergence of the NSB estimates is rather unclear. We do not present results with CS, HT, and WW, because most of their estimates are negative and cannot be plotted on a log scale.

In summary, MM, ZA, and GR, which is used in the main text, perform well for all statistical quantities. NSB generally performs well, yet its convergence to the true value is unclear for the CFIB in the PCFG. The other estimators, CS, HT, and WW, sometimes exhibit clear bias, implying that they should be avoided when the sample size is large.

E Context-free Independence Breaking

In Section 4.4, we presented the results of CFIB for BLLIP when the two nodes are fixed as $(x_0, x_1) = (\text{NP}, \text{NP})$. In this section, we present the results with $(x_0, x_1) = (\text{NP}, \text{VP})$ and (VP, VP) to examine how the metric depends on the fixed nodes.

From the results shown in Figure 14, we can observe a power-law decay similar to that with $(x_0, x_1) = (\text{NP}, \text{NP})$, although the power law exponents may be different. This implies that the CFIB in syntactic structures follows a power-law decay, irrespective of the fixed nodes.

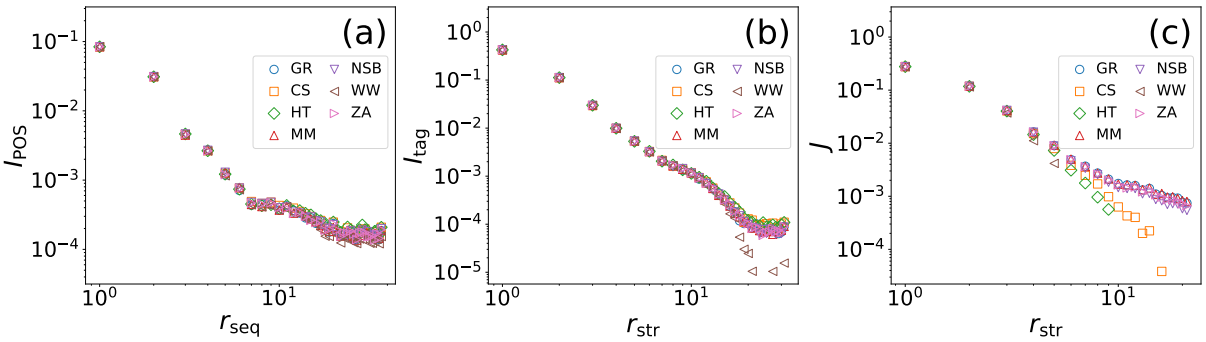

Figure 12: Results for BLLIP obtained with different estimators. (a) MI I_{POS} between POS tags. (b) MI I_{tag} between POS or phrasal tags. (c) CFIB J for $(x_0, x_1) = (\mathrm{NP}, \mathrm{NP})$. In all cases, $N_{\mathrm{data}} = 2.56 \times 10^6$.

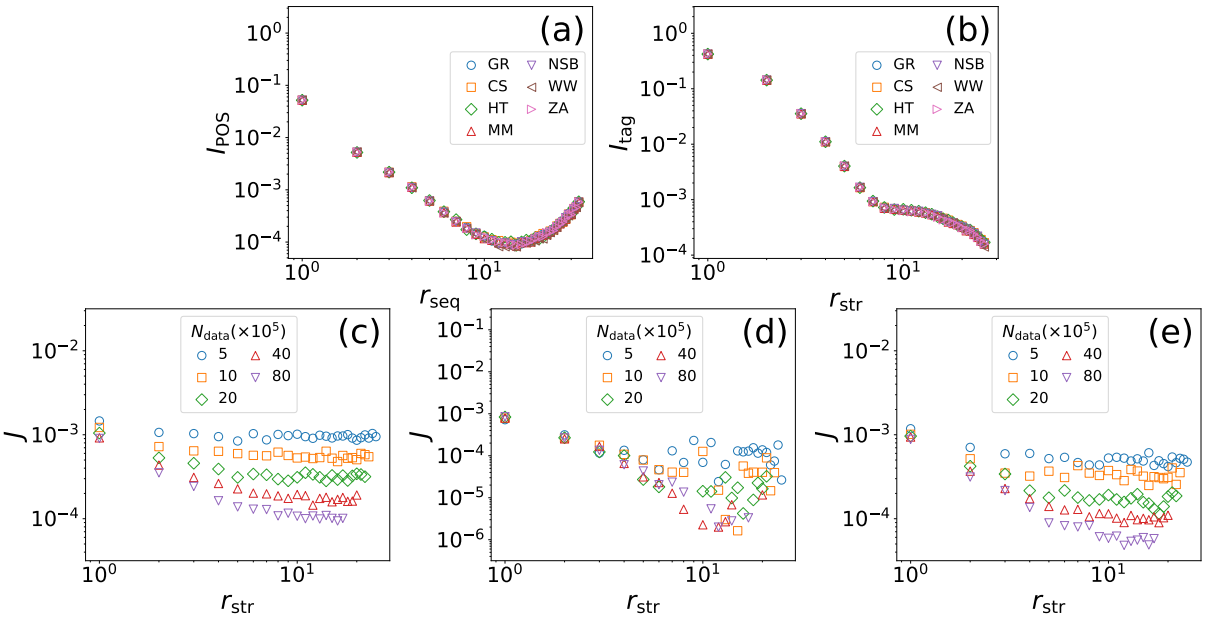

Figure 13: Results for the PCFG obtained with different estimators. (a) MI I_{POS} between POS tags. (b) MI I_{tag} between POS or phrasal tags. In both (a) and (b), $N_{\mathrm{data}} = 8 \times 10^6$. (c-e) CFIB J for $(x_0, x_1) = (\mathrm{NP}, \mathrm{NP})$ obtained with (c) MM, and (d) NSB, and (e) ZA.

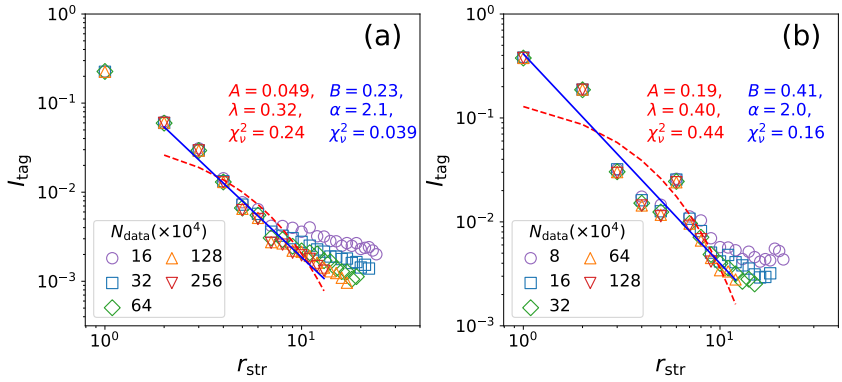

Figure 14: CFIB J for BLLIP with (a) $(x_0, x_1) = (\mathrm{NP}, \mathrm{VP})$ and (b) $(x_0, x_1) = (\mathrm{VP}, \mathrm{VP})$. Fitting was performed for $N_{\mathrm{data}} = 2.56 \times 10^6$ in (a) and $N_{\mathrm{data}} = 6.4 \times 10^5$ in (b).