Measuring Grammatical Diversity from Small Corpora: Derivational Entropy Rates, Mean Length of Utterances, and Annotation Invariance

Fermín Moscoso del Prado Martín
Department of Computer Science and Technology & Jesus College
University of Cambridge, UK
fm611@cst.cam.ac.uk

In many fields, such as language acquisition, neuropsychology of language, the study of aging, and historical linguistics, corpora are used for estimating the diversity of grammatical structures that are produced during a period by an individual, community, or type of speakers. In these cases, treebanks are taken as representative samples of the syntactic structures that might be encountered. Generalizing the potential syntactic diversity from the structures documented in a small corpus requires careful extrapolation whose accuracy is constrained by the limited size of representative sub-corpora. In this article, I demonstrate—both theoretically and empirically—that a grammar's derivational entropy and the mean length of the utterances (MLU) it generates are fundamentally linked, giving rise to a new measure, the derivational entropy rate. The mean length of utterances becomes the most practical index of syntactic complexity; I demonstrate that MLU is not a mere proxy, but a fundamental measure of syntactic diversity. In combination with the new derivational entropy rate measure, it provides a theory-free assessment of grammatical complexity. The derivational entropy rate indexes the rate at which different grammatical annotation frameworks determine the grammatical complexity of treebanks. I evaluate the Smoothed Induced Treebank Entropy (SITE) as a tool for estimating these measures accurately, even from very small treebanks. I conclude by discussing important implications of these results for both NLP and human language processing.

1. Introduction

Estimating the diversity of the grammatical structures—sometimes referred to as **syntactic complexity**—produced by an individual or group thereof is important in multiple areas of linguistics and psychology. Frequent applications include—among others, drawing inferences on the process of language acquisition by children (e.g., Theakston et al. 2004) or second language learners (e.g., Norris and Ortega 2009; Jiang, Bi, and Liu 2019), assessing the severity and progress of neuropsychological disorders (e.g., Roark, Mitchell, and Hollingshead 2007; Pakhomov et al. 2011; Rezaii et al. 2022), or

Action Editor: Liang Huang. Submission received: 12 December 2024; revised version received: 4 April 2025; accepted for publication: 18 April 2025.

https://doi.org/10.1162/COLI.a.15

investigating the changes in a language along its history (e.g., Moscoso del Prado Martín 2014) or along an individual's lifetime (e.g., Agmon et al. 2024). Typically, researchers obtain relevant samples of language (recorded and transcribed conversations, writings, etc.) which are then syntactically annotated, either semi-automatically using parsers, or manually by trained linguists. The syntactic annotations used in these resources are most frequently in the form of context-free derivation trees (e.g., Marcus et al. 1995) or, more recently, dependency graphs (e.g., de Marneffe et al. 2021).

Generally speaking, the number of distinct syntactic structures that an individual might be able to produce is, according to some (see Pullum and Scholz 2010, for discussion), not finite, and certainly extremely large. Researchers therefore need to generalize how much actual grammatical knowledge is represented by a collection of observed syntactic structures. Multiple techniques have been developed in the literature to try to measure the amount of syntactic knowledge represented by a corpus. On the one hand, some researchers have measured the syntax "by proxy," using indices that would correlate with the diversity of syntactic structures. The most common of such indices are the **mean length of utterance** in words (**MLU**; Nice 1925) or morphemes (Brown 1973). On the other hand, some have turned to measures of structural properties argued to be costly by specific theories of syntactic processing (e.g., Yngve 1960; Frazier 1985), or the mean/total number of words intervening between any two syntactic dependents (Lin 1996; Gibson 1998; Liu 2008). Both of these types of approaches fall short of looking at the actual diversity of syntactic structures. Finally, a third group have developed indices counting the presence/absence or the number of times that certain grammatical phenomena argued to be of importance are encountered (e.g., subordination, center-embedding; Scarborough 1990; Agmon et al. 2024), assigning some specific diversity values to each occurrence of each phenomenon. Notice, however, that choosing some specific syntactic constructions as the keystones of diversity makes such measures difficult to apply in languages that do not use such mechanisms, and indeed, none of the proposed constructions can be taken as "language universals" (Evans and Levinson 2009). Even disregarding the heavy annotation burden that such approach imposes, relying on the occurrence of specific constructions for measuring diversity will render the results incomparable across languages. In summary, each of these approaches to measuring diversity presents a number of advantages and inconveniences (see, Cheung and Kemper 1992; Agmon et al. 2024, for overviews of different measures). All these approaches share an ad hoc nature very specific to certain areas of linguistic research, and even to individual languages.

In contrast, some researchers have turned to more natural, standard information theorical measures for the diversity of syntactic structures. The entropy (Shannon 1948) of a population is the measure of diversity most commonly used across a variety of disciplines, such as economics (e.g., Attaran and Zwick 1987), Ecology (e.g., Gotelli and Chao 2013), or genetics (e.g., Poon et al. 2016), and is also used for measuring other aspects in linguistics, such as vocabulary richness (Moscoso del Prado Martín 2014), the size of derivational (Moscoso del Prado Martín, Kostić, and Baayen 2004) or inflectional (Baayen and Moscoso del Prado Martín 2005) morphological paradigms, or quantifying the degree of syntactic word order variations (Futtrell, Mahowald, and Gibson 2015; Levshina 2019) or language sequence predictability (Roark, Mitchell, and Hollingshead 2007; Sy et al. 2023). In contrast with those applications, where diversity is related to the distribution of a finite—even if perhaps very large—number of alternatives (words, species, etc.), syntactic constructions cannot be thought of in terms of their raw number: Even extremely simple grammatical structures can give rise to an infinite number of alternatives. Perhaps the most widely used method for representing syntactic structures

and their probabilistic properties are **Probabilistic Context-Free Grammars** (**PCFGs**; Booth 1969; Booth and Thompson 1973). Given a PCFG that has been induced from a treebank, it is known that the entropy of the possibly infinite set of syntactic structures it generates is finite and can be determined in closed form (Grenander 1967, 1976; Miller and O'Sullivan 1992). This **derivational entropy** has been shown to account for the varying levels of difficulty that people encounter when comprehending language, correlating with experimental measurements of sentence reading times (Hale 2003; Linzen and Jaeger 2014), human memory accuracy for sentences (Hale 2006), the complexity of the syntax used by adults aging (Moscoso del Prado Martín 2017), or across historical periods (Moscoso del Prado Martín 2014), and (within computational linguistics proper) the degree of difficulty one encounters when parsing different languages (Corazza, Lavelli, and Satta 2013).

Researchers often need to generalize the observed syntactic structures, not only those that are explicitly documented in the sample, but also those that could eventually be produced by the individual or population being studied. In these cases, the PCFG is not available a priori, but it is still necessary to draw conclusions beyond the sentences that were directly observed. A powerful tool is to then induce a PCFG from the observed treebank, and then estimate its entropy as a measure of its diversity, and by extension of the diversity of the structures that *could* have been encountered. Such a generalization step comes with its own problems. Typically, a PCFG is induced from a treebank using the method of maximum likelihood (ML). This estimation approach is known to be consistent: With increasing corpus sizes, the estimated rule probabilities converge to their true values with probability 1 (Sánchez and Benedí 1997; Chi and Geman 1998). In turn, the computation of the PCFG entropy from the rule probabilities is exact (Grenander 1967, 1976; Miller and O'Sullivan 1992), and therefore the entropy estimator based on the induced grammar is the ML estimate of the entropy and must also be consistent, eventually converging to true PCFG entropy value. The problem arises because, even if ultimately converging, ML entropy estimates are known to be very biased (Miller 1955): ML consistently underestimates true entropy values. The degree of underestimation correlates positively with two factors: (a) the size of the sample on which the estimates were computed, and (b) the number of possible values of the discrete random variable the entropy of whose distribution is being estimated. The second of these problems could not actually get worse in the case of PCFGs; the number of possible alternatives (the syntactic structures) is, in most cases, infinite. This is further aggravated by the Zipfian power-law nature of their distribution, by which some structures will only be observed extremely rarely. One can therefore expect the degree of underestimation to be rather substantial. The presence of this bias has even motivated some researchers to outright avoid using entropy-based measures of syntactic diversity (e.g., Jing, Widmer, and Bickel 2021). In many applications this problem is aggravated by the sample size (i.e., the size of the treebank used for entropy estimation) being very small, and often the sample size being itself related to the very properties of the individual or population that one is interested in studying. For instance, when studying child language acquisition, or language changes during aging, the available samples of utterances produced by an individual will be necessarily small. First, it is very costly to collect and annotate such samples, and that requires a certain degree of intrusion into the individuals' lives. Second, some individuals (e.g., younger children, older men) tend to talk less than others. This introduces a systematic relation between a person's age and the sample size available (Spokoyny, Irvin, and Moscoso del Prado Martín 2016; Moscoso del Prado Martín 2017). If one wants to draw inferences on the change in syntactic abilities with age, then when using ML estimates of derivational entropy one will encounter a

confound between changes in actual syntactic diversity, and mere changes in sample size (i.e., being more or less talkative). Such systematic relations with the amount of language produced will also arise when one is studying other populations, such as people suffering from neuropsychological disorders.

In this article, I demonstrate how the MLUs and the derivational entropies inferred from treebanks are fundamentally related through the concept of derivational entropy rate, and this has far reaching implications for measuring syntactic diversity from empirical samples. In what follows, I begin by establishing the notation and outlining some basic concepts in formal language theory and information theory (Section 2). This is followed, in Section 3, by developing the theoretical link between the mean length of utterances and the derivational entropy of a grammar, introducing the concept of derivational entropy rate. Here, I put forward two hypotheses about this relationship and its implications. Section 4 introduces a method for derivational entropy estimation from treebanks: the Smoothed Induced Treebank Entropy (SITE). This measure, first used in Moscoso del Prado Martín (2017), builds on the hybrid approach suggested by Moscoso del Prado Martín (2014), and combines it with a better method for bias correction in small datasets (Chao, Wang, and Jost 2013). In several corpus analyses (on both context-free and dependency treebanks), described in Section 5, I demonstrate the accuracy and the limits of SITE for estimating derivational entropies, and I test the theoretical predictions of the relationship between mean length of utterance and derivational entropy. Finally, I conclude in Section 6 by discussing the value and further implications of these results.

2. Notation and Basic Concepts

2.1 Probabilistic Contex-Free Grammars

A PCFG (Booth 1969; Booth and Thompson 1973) is a quadruple $G = (T, N, S, R)$. In this notation, $T = \{a_1, \ldots, a_{\|T\|}\}$ is a finite set of **terminal symbols**, often referred to as the **alphabet** on which the grammar is defined (here the operator $\|\cdot\|$ denotes the cardinality of a set). $N = \{A_1, \ldots, A_{\|N\|}\}$ is a finite set of **non-terminal symbols**, from which $S \in N$ is the **starting symbol**, that is, the non-terminal that serves as the root of all parse trees generated by the grammar. Finally, R is a set of rules $R = \{r_1, \ldots, r_{\|R\|}\}$ of the form $r \equiv p : A \rightarrow \alpha$, with the left-hand side of the arrow being a non-terminal symbol ($A \in N$), its right-hand side being a string of terminal and non-terminal symbols ($\alpha \in [N \cup T]^*$), and $0 \leq p \leq 1$ denoting a probability value for the rule. Each of these rules states that, in a string of terminals and/or non-terminals, rewriting the non-terminal A with the string α is an operation allowed by the grammar occurring with probability p.

For each non-terminal symbol A_i, $R_i \subseteq R$ denotes the subset of R consisting of the $\|R_i\|$ rules that have A_i in their left-hand side. For each non-terminal symbol, I assume a rule ordering such that the set of its rules can be denoted as $R_i = \{r_{i,1}, \ldots, r_{i,\|R_i\|}\}$. In this way $r_{i,j}$ always refers to the j-th rule having the non-terminal A_i as its left-hand side, and $p_{i,j}$ refers to the probability associated with that rule. The probabilities for the rules expanding each non-terminal symbol in a PCFG need to be proper, in the sense that they are normalized such that

$$\sum_{r_{i,j} \in R_i} p_{i,j} = 1, \qquad \text{for all } 1 \leq i \leq \|N\| \tag{1}$$

The PCFG generates derivation parse trees through recursive application of its production rules, beginning with the starting symbol S until one obtains a sequence composed only of terminal symbols. A PCFG can generate in this way a possibly infinite set of trees $\mathcal{T}[G] = \{t_1, t_2, t_3, \ldots\}$. The probabilities associated with the rules also associate a probability value to each parse tree; this is given by

$$\mathrm{p}_G(t) = \prod_{r_i \in R} p_i^{f(r_i;t)} \tag{2}$$

where $f(r_i; t)$ denotes the number of applications of rule r_i involved in the construction of the tree t. The PCFG itself also needs to be **proper**, in the sense that

$$\sum_{t \in \mathcal{T}(G)} \mathrm{p}_G(t) = 1 \tag{3}$$

In fact, it is possible to create improper PCFGs, but any PCFG that is induced from a treebank by ML will always be proper (Chi 1999; Chi and Geman 1998; Sánchez and Benedí 1997).

2.2 Maximum-Likelihood Induction of a PCFG from a Treebank

Given a treebank (a collection of parse trees), PCFG induction involves inferring the PCFG from which the trees in the treebank were sampled. A simple method for performing this inference is the ML method: From each of the internal nodes in a sample of trees, one can infer a context-free derivation rule. The internal nodes in the trees are considered non-terminal symbols, and the leaf nodes in the tree are taken as terminal symbols. Then, if an internal node is labeled A, and has as its immediate descendants an ordered list of nodes (leaves or otherwise) $x_1, \ldots, x_k$, with $k \geq 1$, one can infer that the grammar contains the context-free rule $A \rightarrow x_1 \ldots x_k$, as is illustrated in the example of Figure 1. If all trees have the same symbol A as their root, then A is taken to be the starting symbol of the PCFG. If, on the other hand, the trees have different root nodes, one can add a new additional non-terminal symbol A_0 as the starting symbol of the PCFG, and also include a context-free rule $A_0 \rightarrow A_i$ for each A_i that appears as the root of a tree in the treebank. Finally, the rule probabilities are induced by simple ML. If $f[A \rightarrow \alpha; T]$ denotes the frequency with which the rule $A \rightarrow \alpha$ has been observed in treebank T, and $f[A; T]$ is the frequency with which the non-terminal A was encountered, the ML estimate of probability associated with that rule is given by its relative frequency,

$$p \approx \frac{f[A \rightarrow \alpha; T]}{f[A; T]} \tag{4}$$

It is known that the PCFG induced by this method is consistent; for an increasingly large treebank, with probability 1, it converges to the true PCFG (Sánchez and Benedí 1997; Chi and Geman 1998) generating it (provided that such one does exist).

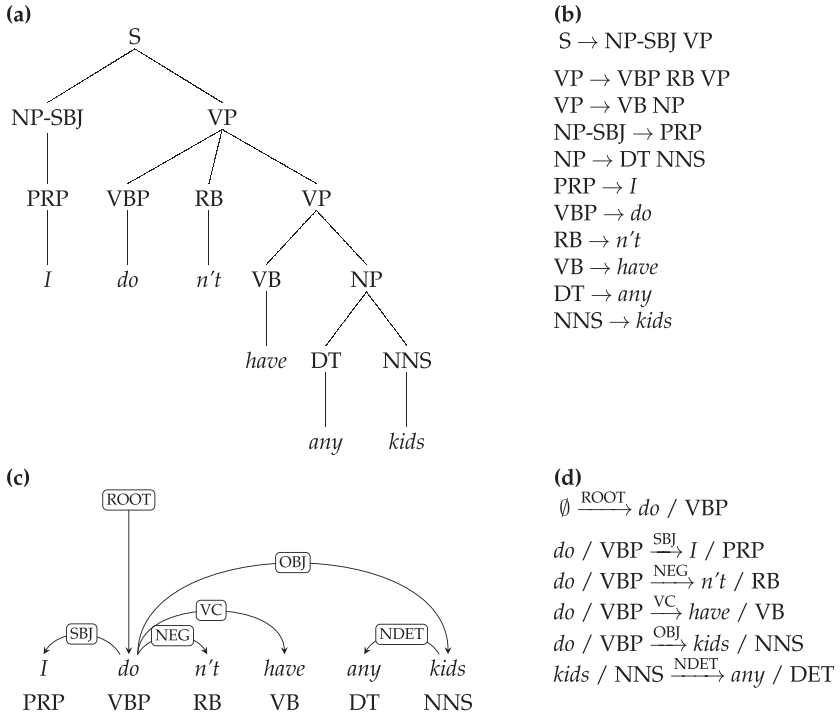

Figure 1
(a) Example of a syntactic derivation tree as can be contained in a treebank (example taken from the Penn Treebank (Marcus et al. 1995)). (b) Context-free derivation rules induced from the tree in (a). (c) Example of a dependency tree encoding the syntactic structure of the example in (a–b). (d) Dependency relations induced from the dependency graph in (c). The capitalized symbols are non-terminals and the italicized ones are terminals.

2.3 Derivational Entropy of a PCFG

The entropy (Shannon 1948) of the parse trees generated by a grammar, PCFG or otherwise, is given by the usual equation,

$$H[G] = -\sum_{t \in \mathcal{T}[G]} \mathrm{p}_G[t] \log \mathrm{p}_G[t] \tag{5}$$

and is referred to as the **derivational entropy** of the PCFG (Grenander 1967, 1976). Notice that, despite being a sum over a potentially infinite set of trees, the magnitude of the entropy remains necessarily finite for the types of grammars that can be inferred from a treebank by ML (Sánchez and Benedí 1997; Chi and Geman 1998). It is therefore a very useful tool for comparing the degree of productivity of two grammars, even when they both potentially produce an infinite number of trees.

PCFGs can potentially generate an infinite number of trees. Therefore, using Equation 5 directly is most often impractical as it would involve a sum with an infinite number of terms. However, the exact value of $H[G]$ can be computed exactly from a PCFG's production rules (Grenander 1967, 1976; Miller and O'Sullivan 1992) without

having to resort to the sum in Equation 5. For this it is necessary to introduce the concepts of the characteristic matrix of a PCFG, and its expansion entropy vectors.

The **characteristic matrix** of a PCFG is an $\|N\| \times \|N\|$ square matrix $\mathbf{M}_G$ whose i-th row – j-th column element $m_{i,j}$ denotes the average number of non-terminals $A_j \in N$ that will be produced in a single expansion of the non-terminal $A_i \in N$ using the rules of the grammar. More formally,

$$\mathbf{M}_G = (m_{i,j}), \quad m_{i,j} = \sum_{r_{i,k} \in R_i} p_{i,k} f(A_j; r_{i,k}) \tag{6}$$

where $f(A_j; r_{i,k})$ denotes the number of times that A_j occurs in the right-hand side of the rule $r_{i,k}$. In addition, for each non-terminal $A_i \in N$ we define the entropy of the alternative production rules that can expand it,

$$h_0[A_i] = -\sum_{r_{i,j} \in R_i} p_{i,j} \log p_{i,j} \tag{7}$$

to which I will refer to as the **local expansion entropy** of symbol A_i. If $\mathbf{h}_0$ is the column vector whose components are the $h_0[A_i]$ defined above (i.e., the local expansion entropy vector), and $\mathbf{I}$ is the identity matrix of dimension $\|N\|$, the derivational entropies vector

$$\mathbf{H} = (\mathbf{I} - \mathbf{M}_G)^{-1} \cdot \mathbf{h}_0 \tag{8}$$

has as its components the entropies of the subtrees generated by the grammar, starting with each of its non-terminal symbols (see Appendix A for derivation). For such entropies to exist, the inverse matrix $(\mathbf{I} - \mathbf{M}_G)^{-1}$ has to exist as well. Formally, such inverse will always exist for characteristic matrices whose *spectral radius* is smaller than one, $\rho(\mathbf{M}_G) < 1$. The spectral radius of a matrix is given by the largest absolute value of its eigenvalues. Intuitively, it indicates whether the transformation denoted by the matrix is "expanding" or "contracting." In the former case ($\rho(\mathbf{M}_G) < 1$), repeated applications of the transformation end up converging to zero, whereas in the latter case ($\rho(\mathbf{M}_G) > 1$) each transformation inflates the values, which therefore diverge. More formally, if the spectral radius is smaller than one, the power series formed my multiplying the matrix by itself repeatedly converges to zero

$$\lim_{k \to \infty} \mathbf{M}_G^k = \mathbf{0} \tag{9}$$

In terms of the grammar, convergence of the power series indicates that the average length of the strings it generates is finite. Fortunately, it is known that for any PCFG that is proper in the sense of Equation 3, $\rho(\mathbf{M}_G) < 1$, and all PCFGs induced from a treebank by ML satisfy this condition (Sánchez and Benedí 1997; Chi and Geman 1998; Chi 1999). Therefore, if $S = A_k$ is the starting symbol of the grammar, the entropy of Equation 5 is given by the k-th element of its derivational entropies vector $\mathbf{h}$,

$$H[G] = H_k \tag{10}$$

2.4 Probabilistic Dependency Grammars

Context-Free Grammars (CFGs; Chomsky 1956, of which PCFGs are a probabilistic extension) build syntax on the concept of **constituency**: The observation that blocks within sentences can be interchanged with other blocks, with the non-terminal internal nodes of the parse trees representing such blocks. An alternative tradition for the description of syntactic structures is that of **Dependency Grammar** (DG; Tesnière 1959), which relies instead on pairwise dependencies or directed links between words. In this framework, a sentence's syntactic structure is represented by a connected directed graph—which, in most conventions is further restricted to be a directed tree—with labeled nodes corresponding to either words or broader lexical categories, and vertices denoting their pairwise relations (often also labeled to specify the nature of the relation). These graphs are referred to as dependency graphs. For instance, the syntactic structure of Figure 1a, could also be represented as the dependency tree in Figure 1c. In DGs, instead of the context-free productions, the rules of the grammar consist of a list of pairwise directed dependency relations that are licensed by the grammar, such as those in Figure 1d. Depending on the particular formalism, the relations might be specified between lexical categories and/or, in fully lexicalized models, between individual words. As with the PCFGs, DGs can include probabilities in the dependencies themselves, and are then referred to as Probabilistic Dependency Grammars (PDGs; see, e.g., Eisner 1996; McDonald et al. 2005, for some probability models that can be applied to define a PDG).

DGs and CFGs are weakly equivalent (Gaifman 1965; Abney 1995) in the sense that they both characterize the set of context-free languages in Chomsky's traditional hierarchy (Chomsky 1956). Their relation is, however, asymmetric. Any DG can be converted to a strongly equivalent CFG (i.e., with a one-to-one mapping between trees and dependency graphs), but the reverse only holds for a restricted subset of CFGs. For instance, the dependency trees in Figure 1c can be represented, with or without the dependency labels, by the equivalent parse trees in Figure 2. I will make use of this property in order to estimate the syntactic diversity of PDGs.

2.5 Methods for Correcting Entropy Estimation Bias

ML estimators of entropy from a sample are long known to be biased (Miller 1955): They will underestimate—sometimes quite severely—the true value of the entropy of the process that generated the sample. There exists a broad literature on how to correct the bias of entropy estimates obtained from samples, even when these are rather small (Miller 1955; Nemenman, Shafee, and Bialek 2002; Chao and Shen 2003; Paninski 2003; Nemenman, Bialek, and de Ruyter van Steveninck 2004; Vu, Yu, and Kass 2007; Hausser and Strimmer 2009; Chao, Wang, and Jost 2013). Among these methods, one can distinguish between those that correct entropy estimations among a set of alternatives whose cardinality is either known a priori, or safely assumed to be very low (Miller 1955; Nemenman, Shafee, and Bialek 2002; Hausser and Strimmer 2009), and those that enable corrections for a very large, unknown, and potentially infinite number of alternatives (Paninski 2003; Nemenman, Bialek, and de Ruyter van Steveninck 2004; Chao and Shen 2003). Comparisons on the accuracy of the estimators in this latter group (Vu, Yu, and Kass 2007) reveal that the best performing estimator is the Coverage-Adjusted Estimator (CAE; Chao and Shen 2003). More recently, the same group that developed the CAE introduced a new estimator: the CWJ estimator (Chao, Wang, and Jost 2013). This estimator is shown to be less biased and faster converging than the CAE estimator. See Appendix B for details on these estimators.

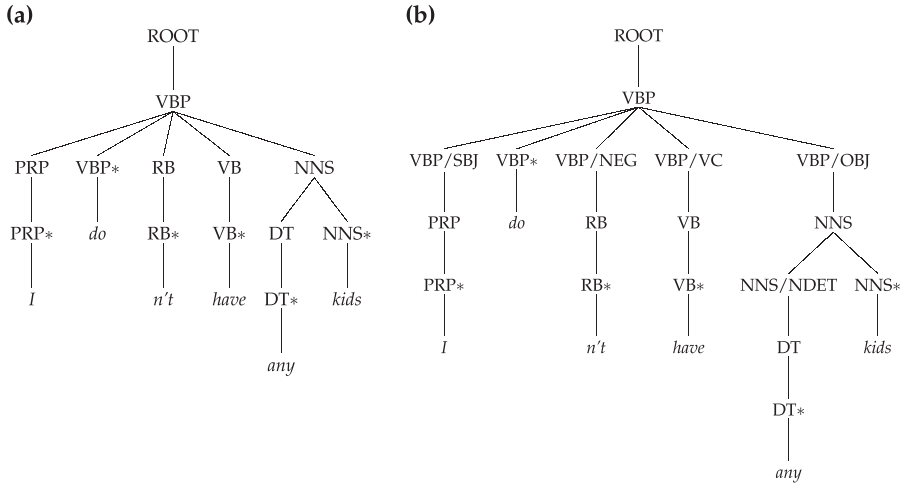

Figure 2
Representations of the dependency graph in Figure 1c as context-free derivation trees: (a) Omitting the dependency labels. (b) Taking dependency labels into account by means of additional non-terminals.

3. The Relation Between MLU and Derivational Entropy, and its Implications

3.1 The Crucial Role of MLU

The average length in words of utterances in a corpus, the MLU, is by far the oldest measure of syntactic complexity used by researchers. Its first usage, in language acquisition research, dates back to almost a full century ago (Nice 1925). Still today, it is the most commonly used measure in several fields, including language acquisition, second language learning, and aging research. Despite—or, perhaps, precisely because of—its simplicity, this measure has some very desirable properties. First, it is very simple to compute as an empirical average, even from unlabeled corpora, and its ML estimator is both consistent and unbiased. Furthermore, it exhibits an extremely fast convergence to its true value, so estimates based on very few samples are already quite reliable (Casby 2011).

Second, the measure is relatively theory-free. It provides an indicator of grammatical complexity that is not bound to any specific theory of syntax. There has been some debate in the literature about the correct units that one should use for measuring MLU, whether one should count words, as in the original measure (Nice 1925), or one should count morphemes instead (Brown 1973). In reality, however, the unit of measurement for MLU does not really matter much; for instance, Parker and Brorson (2005) find both measures to be almost perfectly correlated (Pearson's $r = .99$).

MLU is considered a "proxy" measure, in the sense that rather than measuring diversity itself, it measures a quantity that correlates well with it. Some argue that length and syntactic diversity are different things (e.g., Crystal 1974; Klee and Fitzgerald 1985), and one can find individual constructions where the two are decoupled. However possible, in actual datasets such occurrences are relatively rare in comparison with the number of cases where MLU and diversity are actually related. At the macroscopic scale, MLU actually correlates very well with many measures of syntactic diversity (e.g.,

Scarborough 1990; Scarborough et al. 1991; Cheung and Kemper 1992; Norris and Ortega 2009; Agmon et al. 2024). MLU and derivational entropy are in fact very closely related. As a collateral observation to the main research question, Moscoso del Prado Martín (2017, Figure 3 on p. 959) reports that, across the dialogues in the Switchboard I Corpus (Godfrey, Holliman, and McDaniel 1992), the relation between derivational entropy and MLU is almost perfectly linear, making both measures virtually indistinguishable (i.e., Pearson's $r = .98$).

3.2 Within a PCFG, MLU and Derivational Entropy Must Be Directly Proportional

Such a close relationship between derivational entropy and MLU boils down to a fundamental property of PCFGs. To see this, let $\boldsymbol{\ell}$ be a vector each of whose $\|N\|$ components is the average length of the strings of terminal symbols that can be directly derived from one of the grammar nonterminals. Then, the vector of average length is known (Hutchins 1972; Wetherell 1980; see also Appendix A) to be

$$\boldsymbol{\ell} = (\mathbf{I} - \mathbf{M}_G)^{-1} \cdot \boldsymbol{\ell_0} \tag{11}$$

where $\boldsymbol{\ell_0}$ is a vector whose i-th component is the average number of terminal symbols introduced by the rules that directly expand the i-th non-terminal symbol:

$$\ell_0[i] = \sum_{r \in R_i} \mathrm{p}[r] \sum_{c \in T} f[c; \mathrm{rhs}[r]] \tag{12}$$

with R_i denoting the set of rules expanding non-terminal symbol A_i, rhs$[r]$ the right-hand side of rule r, p$[r]$ the probability associated with that rule in the PCFG, and $f[c; s]$ the number of times terminal symbol c occurs in string $s \in (T \cup N)^*$. Notice already that Equation 11 is identical to Equation 8 giving a PCFG's derivational entropy, just replacing the entropy vectors $\mathbf{h}$ and $\mathbf{h_0}$ with $\boldsymbol{\ell}$ and $\boldsymbol{\ell_0}$, respectively.

In Equations 8 and 11, the i-th row vector of matrix $(\mathbf{I} - \mathbf{M}_G)^{-1}$, which I will abbreviate to $\mathbf{v}_i$, determines both the length of the strings generated by non-terminal A_i and the entropy of the subtrees it generates. With this notation, the entropy and MLU generated by a grammar are given by plain dot products:

$$H[A_i] = H[i] = \mathbf{v}_i \cdot \mathbf{h_0} = |\mathbf{v}_i|\,|\mathbf{h_0}|\,\cos[\mathbf{v}_i, \mathbf{h_0}] \tag{13}$$

$$\mathrm{MLU}[A_i] = \ell[i] = \mathbf{v}_i \cdot \boldsymbol{\ell_0} = |\mathbf{v}_i|\,|\boldsymbol{\ell_0}|\,\cos[\mathbf{v}_i, \boldsymbol{\ell_0}] \tag{14}$$

where $|\mathbf{x}|$ denotes the norm of vector $\mathbf{x}$. This implies that, internally, the entropies and MLUs of the symbols of any PCFG *must be strongly positively correlated*.

To see this last point, assume we have two vectors $\mathbf{h_0}$ and $\boldsymbol{\ell_0}$ and a characteristic matrix $\mathbf{M}$ that have been induced from a cropus. What relation should we expect between the pairs of $H[A]$ and MLU$[A]$ across the grammar? Notice that, within a PCFG, the sole sources of variability for both the entropy and the MLU of any symbol A are the variabilities of the vector norms $\|\mathbf{v_i}\|$ (i.e., the norms of the rows of the matrix $(\mathbf{I} - \mathbf{M}_G)^{-1}$), and the cosines of the angles that these vectors form with $\mathbf{h_0}$ and $\boldsymbol{\ell_0}$ (whose norms $|\mathbf{h_0}|$ and $|\boldsymbol{\ell_0}|$ remain constant across the non-terminals in the grammar). Furthermore, all vectors $\mathbf{h_0}$, $\boldsymbol{\ell_0}$, all $\mathbf{v}_i$, and—as a consequence—both the cosines are non-negative. Such non-negativity is certain because, for all grammars induced from a treebank by ML, $\mathbf{M}_G$ is non-singular and its largest eigenvalue is positive and smaller

than one, $\rho(\mathbf{M}_G) < 1$ (Sánchez and Benedí 1997; Chi and Geman 1998; Chi 1999). In turn, this entails that $(\mathbf{I} - \mathbf{M}_G)$ is an M-matrix (Ostrowski 1937), whose inverse is always non-negative. The cosine between randomly sampled non-negative vectors, for large dimensionalities (in our case the dimensionality is the cardinality of the non-terminal alphabet $\|N\|$) is expected to be very strongly concentrated, following an approximately normal distribution with mean $2/\pi$ and standard deviation $4/(3\pi\sqrt{\|N\|})$ (see Appendix C). Notice that the standard deviation of the cosine is inversely proportional to the square root of the cardinality of the non-terminal alphabet. This implies that the standard deviation vanishes as $\|N\|$ grows, while the mean remains constant, making the coefficient of variation also vanish with $\|N\|$,

$$\lim_{\|N\|\to\infty} \frac{4/(3\pi\sqrt{\|N\|})}{2/\pi} = \lim_{\|N\|\to\infty} \frac{2}{3\sqrt{\|N\|}} = 0 \tag{15}$$

Even for relatively small values of $\|N\|$, the variances of the cosines are expected to be orders of magnitude smaller than their mean, and also insignificant relative to the variability of the $|\mathbf{v_i}|$. Therefore, for all practical purposes, each of the cosines can be approximated as a constant with value $2/\pi$, leaving the norms $|\mathbf{v_i}|$ as the single and shared source of effective variability for both measures. Under this small approximation, the expected values of the relevant variances become

$$\mathrm{Var}\left[|\mathbf{v_i}|\right] = \sigma^2_{|\mathbf{v}|}; \quad \mathrm{Var}\left[H[A]\right] \approx \tfrac{2}{\pi}\,|\mathbf{h_0}|\,\sigma^2_{|\mathbf{v}|}; \quad \mathrm{Var}\left[\mathrm{MLU}[A]\right] \approx \tfrac{2}{\pi}\,|\boldsymbol{\ell_0}|\,\sigma^2_{|\mathbf{v}|} \tag{16}$$

and the respective covariances are

$$\begin{aligned}
\mathrm{Cov}\left[|\mathbf{v_i}|, H[A]\right] &\approx \sigma^2_{|\mathbf{v}|}\sqrt{\tfrac{2}{\pi}\,|\mathbf{h_0}|} \\
\mathrm{Cov}\left[|\mathbf{v_i}|, \mathrm{MLU}[A]\right] &\approx \sigma^2_{|\mathbf{v}|}\sqrt{\tfrac{2}{\pi}\,|\boldsymbol{\ell_0}|} \\
\mathrm{Cov}\left[H(A), \mathrm{MLU}[A]\right] &\approx \sigma^2_{|\mathbf{v}|}\tfrac{2}{\pi}\sqrt{|\mathbf{h_0}|\,|\boldsymbol{\ell_0}|}
\end{aligned} \tag{17}$$

Using the variances and covariances above, we see that we should expect all three measures $H(A)$, MLU$[A]$, and $|\mathbf{v}|$ to approximate a perfect Pearson correlation within a grammar. Crucially,

$$r\left[H(A), \mathrm{MLU}[A]\right] = \frac{\mathrm{Cov}\left[H(A), \mathrm{MLU}[A]\right]}{\sqrt{\mathrm{Var}\left[H[A]\right]}\sqrt{\mathrm{Var}\left[\mathrm{MLU}[A]\right]}} \approx \frac{\sigma^2_{|\mathbf{v}|}\,\frac{2}{\pi}\sqrt{|\mathbf{h_0}|\,|\boldsymbol{\ell_0}|}}{\sqrt{\frac{2}{\pi}\,|\mathbf{h_0}|\,\sigma^2_{|\mathbf{v}|}}\sqrt{\frac{2}{\pi}\,|\boldsymbol{\ell_0}|\,\sigma^2_{|\mathbf{v}|}}} = 1 \tag{18}$$

Furthermore, as a result of this correlation, dividing both components of Equations 13 and 14:

$$\frac{H[A_i]}{\mathrm{MLU}[A_i]} = \frac{|\mathbf{h_0}|}{|\boldsymbol{\ell_0}|}\,\frac{\cos[\mathbf{v_i}, \mathbf{h_0}]}{\cos[\mathbf{v_i}, \boldsymbol{\ell_0}]} = \alpha_i \geq 0 \tag{19}$$

Although, as discussed above, both of those cosines are expected to be strongly clustered around their mean, in this case they are the sole sources of variability in the data. Unlike in the correlation case, their variability will not be swamped by the variability of

the vector norms, and therefore, they should not be ignored. However, as is discussed in Appendix D, for the moderate or large values of $\|N\|$ that are typically found in real-life PCFGs, the ratio between both cosines will be very well approximated by a normal distribution, with mean 1, and standard deviation $1/(2\sqrt{\pi\,\|N\|})$. As a consequence, the ratios α_i will themselves be normally distributed

$$\frac{H[A_i]}{\mathrm{MLU}[A_i]} = \frac{|\mathbf{h_0}|}{|\boldsymbol{\ell_0}|}\,\frac{\cos[\mathbf{v_i}, \mathbf{h_0}]}{\cos[\mathbf{v_i}, \boldsymbol{\ell_0}]} \sim \mathrm{N}\left(\frac{|\mathbf{h_0}|}{|\boldsymbol{\ell_0}|}, \frac{1}{2}\sqrt{\frac{\pi\,|\mathbf{h_0}|}{\|N\|\,|\boldsymbol{\ell_0}|}}\right) \tag{20}$$

To appreciate the degree to which these ratios are expected to be concentrated for real-sized PCFG, if one considers a very small grammar with just 46 non-terminals (as the dependency-induced PCFG of Section 5.1.2), this would entail that the range between the expected maximum value of α_i and its minimum value would be about 16% of the average value (i.e., the difference between the $\|N\|/(\|N\|+1)$ and $1/(\|N\|+1)$ quantiles of the normal distribution). For a more typical, Penn Treebank–type, grammar with 662 non-terminals (as the grammar in Section 5.1.1), this range reduces to a mere 6% difference between the maximum and the minimum ratio among the 662 non-terminals. This small variability is consistent with the predicted correlation approaching 1.0. For instance, even for the smallest value of $\|N\| = 46$, the correlation between two sets of random variables related through the ratio in Equation 20 is typically $r > .9$, and the correlation increases for larger $\|N\|$ values, quickly getting to $r > .99$ for the usual large grammar sizes.

Evidently, all the above predictions are made under the assumption that the vectors $\mathbf{v_i}$ are chosen fully *at random*, which is obviously not the case in *real* PCFGs. The point being that, even if one used fully random vectors, one would expect an almost perfect internal correlation between the $H(A_i)$ and $\mathrm{MLU}[A_i]$ values of a PCFG. The baseline is therefore that their correlation should be strong. *If anything, we would expect the correlation in a real PCFG to be stronger*.

In sum, the MLUs and derivational entropies of the non-terminals of a PCFG, as long as the PCFG is proper, are the result of a non-negative linear transformation. Therefore, one should expect a *grammar-internal* correlation such that

$$H[A_i] = \alpha_i\,\mathrm{MLU}[A_i] \tag{21}$$

with a value of α_i very concentrated around the ratio $|\mathbf{h_0}|/|\boldsymbol{\ell_0}|$ (and this concentration increases with $\|N\|$).

3.3 The Derivational Entropy Rate

The proportionality in Equation 21 suggests that one can define the **derivational entropy rate** of a PCFG as the average amount of derivational entropy corresponding to each unit of string length in the grammar,

$$h[G] = \frac{H[G]}{\mathrm{MLU}[G]} = \alpha > 0 \tag{22}$$

In other words, we can consider changing from units of string length (i.e., terminal symbols) into derivational entropy units (bits, nats, etc.) as a mere change of measurement units.

The observation of Moscoso del Prado Martín (2017) goes one step further than mere proportionality. The telephone conversations contained in the Switchboard I Corpus are diverse with respect to the topic of conversation, the sociocultural origins, ages, and sex of the speakers. Nevertheless, the very strong correlation between the MLU and derivational entropies of the files would suggest that *the derivational entropy rate is roughly constant across all conversations in the corpus*. What all these conversations share is that they all happened in (different sub-dialects of) spoken American English and that they were all parsed using the same grammatical annotation convention, that of the Penn Treebank (Marcus et al. 1995). This motivates putting forward two hypotheses of decreasing generality:

Hypothesis 1: The derivational entropy rate of a PCFG is constant across PCFGs representing the same or closely related languages.

Hypothesis 2: The derivational entropy rate of a PCFG is constant across PCFGs representing the same or closely related languages, annotated using the same grammatical convention.

The strong hypothesis (Hypothesis 1) seems unlikely to hold. One can think that even for the same actual language sample, one can use different grammatical formalisms for annotating it. Importantly, at least in principle, different formalisms would often result in more or less generalization of possible structures from the grammar. On the other hand, Hypothesis 2 could be plausible, but might also not hold when the probability distributions of the induced grammars differ too strongly, as one may expect if the language samples came from very different registers or historical periods. In Section 5, I will test the validity of these hypotheses.

3.4 Relativity of Derivational Entropy

Consider now two PCFGs on the same terminal alphabet, $G_1 = (T, N_1 = \{A_1, \ldots, A_{\|N_1\|}\}, A_1, R_1)$ and $G_1 = (T, N_2 = \{B_1, \ldots, B_{\|N_2\|}\}, B_1, R_2)$ that generate the same probabilistic language ($\mathcal{L}(G_1) = \mathcal{L}(G_2)$) or, that at least generate strings of the same average length. These two PCFGs can, for instance, be ML-induced grammars from a context-tree treebank and a dependency treebank annotating the same corpus. MLU is the same for both grammars:

$$\mathrm{MLU}[A_1] = \mathrm{MLU}[B_1] \tag{23}$$

Both PCFGs are subject to the grammar internal correlation of Equation 21,

$$\begin{aligned} H[A_i] &= \alpha_1 \mathrm{MLU}[A_i] \\ H[B_i] &= \alpha_2 \mathrm{MLU}[B_i] \end{aligned} \tag{24}$$

with α_1, $\alpha_2 > 0$. Combining Equation 23 and Equation 24, one can express the *external* relationship between the derivational entropies of both grammars in terms of the parameters of their internal entropy-length relations,

$$H[G_2] = \frac{\alpha_2}{\alpha_1} H[G_1] = \beta_{G_1,G_2} H[G_1], \qquad \beta_{G_1,G_2} \geq 0 \tag{25}$$

Hence, we can expect that the entropies of grammars describing the same languages will be directly proportional to each other (without an intercept).

Suppose we have multiple corpora of the same language annotated according to two different grammatical theories. The actual entropy values of the corpora will change depending on which of the two grammatical theories was used for parsing. However, as long as the grammatical theories are sensible, it should be expected that—in the vast majority of cases—different grammatical structures will in most cases receive different parses in either theory, and identical grammatical structures will in most cases remain identical in both theories. This implies that the relative values of the entropies for different corpora should be strongly correlated across the two theories. For this to be true, the value of parameter β needs to be roughly constant across pairs of treebanks parsed according to two different conventions. Interestingly, this is exactly what one would predict if Hypothesis 2 above were true: The derivational entropy rates (α_1 and α_2) would be constant for a particular grammatical annotation convention, and so would β. In other words, one would expect that grammatical entropy estimates from different corpora are, *in relative terms*, independent of the grammatical convention used in building the treebanks: If corpus A has a greater diversity than corpus B according to grammatical theory 1, one should expect that Hypothesis 2 implies that the diversity of corpus A is also greater than that of B. On the other hand, Hypothesis 1 would have even stronger implications; the derivational entropy rates in both corpora should be identical (because they are dealing with the same language) and hence the derivational entropies should also be identical. These implications will be tested in Section 5.

4. Measuring Derivational Entropy from Small Treebanks

4.1 Generalizing from Induced PCFGs

If one's goal is to estimate the derivational entropy of the grammar from which a particular corpus was sampled, rather than considering just the directly observed syntactic trees, one should first generalize over the structures observed in the corpus. The goal is to account not just for the directly observed structures, but also for those that could have occurred. A way to achieve this is by inducing a grammar, and then estimating the entropy from the induced grammar, rather than from the original treebank. Approaches similar to this have been followed by several studies (Hale 2003, 2006; Corazza, Lavelli, and Satta 2013; Linzen and Jaeger 2014; Moscoso del Prado Martín 2014, 2017). The additional grammar induction step ensures that the internal structure of the parses is explicitly considered: Evidently, the induced grammar accounts for all the syntactic parses that have been directly observed. In addition to these, it also takes into account infinitely many other parses that, despite not having been actually found on the treebank, could be built by recombination of the partial structures observed in the corpus:

$$\underbrace{t_1, t_2, \ldots, t_{\|T\|}}_{\text{Observed structures}}, \underbrace{t_{\|T\|+1}, t_{\|T\|+2}, t_{\|T\|+3}, \ldots}_{\text{Generalized plausible structures}} \tag{26}$$

By generalizing, the grammar is very likely to also generate many structures that, on human inspection, can be determined to be actually impossible. This last point is less of an actual problem than one might think. What we are interested in measuring is the degree to which the observed parses provide evidence for internal structure. The over-productivity of the induced grammars, from this perspective, is just a tool for

measuring internal structure at a macroscopic scale, rather than an explicit model for the precise structures that an individual might actually produce. Evidently, the latter goal would require us to model well beyond syntax, including explicit models of semantics, pragmatics, and all kinds of socio-cultural aspects. In short, we are concerned with measuring syntactic diversity on a macroscopic scale, rather than identifying individual syntactic structures at a microscopic one.

4.2 ML Estimator

The simplest way for inducing a grammar from a treebank is using the ML method described in Section 2.2. Once such a grammar has been induced, there are some options for how to estimate its entropy. The most evident one would be to just compute the ML approximations of the characteristic matrix ($\mathbf{M}_G^{\mathrm{ML}}$) and local expansion entropy vector ($\mathbf{h}_\mathbf{0}^{\mathrm{ML}}$) directly from the induced grammar, so that the ML estimator for the entropy can be computed by simply applying Equation 8 to compute the ML estimate of the derivational entropies vector:

$$\mathbf{H}^{\mathrm{ML}} = (\mathbf{I} - \mathbf{M}_G^{\mathrm{ML}})^{-1} \cdot \mathbf{h}_\mathbf{0}^{\mathrm{ML}} \tag{27}$$

whose k-th component (if A_k is the root symbol of the PCFG) is the entropy estimator:

$$H^{\mathrm{ML}}[G] = H_k^{\mathrm{ML}} \tag{28}$$

This was the method used for computing derivational entropies by several previous studies (Hale 2003, 2006; Linzen and Jaeger 2014).

4.3 ML-Monte Carlo Estimator

An interesting alternative to this ML method was used by Corazza, Lavelli, and Satta (2013). Using a sampling approach to ML estimation, they take the true probability of the trees ($\mathrm{p}_G[t]$) to be unknown, but approximated by the probability of the tree using the grammar induced by ML from a training subset of the treebank ($\mathrm{p}_G^{\mathrm{ML}}[t]$). An upper bound for the true entropy of the underlying grammar would be provided by the cross-entropy (Shannon 1948) between both distributions:

$$H[G\|\mathrm{ML}] = -\sum_{t \in \mathcal{T}(G)} \mathrm{p}_G[t] \log \mathrm{p}_G^{\mathrm{ML}}[t] \geq H[G] \tag{29}$$

As the true tree probabilities (p_G) are unknown, Corazza, Lavelli, and Satta (2013) take advantage that, by the Shannon-McMillan-Breiman theorem (Breiman 1957/1960; Chung 1961), the treebank itself can be considered a sample from the true p_G that will converge with a sufficiently large corpus. Therefore, they propose using a Monte Carlo estimator for Equation 29, by sampling trees from a testing subset of the treebank ($\mathrm{p}_T[t]$) as an approximation of the true $\mathrm{p}_G[t]$, and hence one can use this testing subsample for computing the estimated cross-entropy:

$$\hat{H}[G\|\mathrm{ML}] = -\sum_{t \in T} \mathrm{p}_T[t] \log \mathrm{p}_G^{\mathrm{ML}}[t] = -\sum_{t \in T} \frac{f[t;T]}{\|T\|} \log \mathrm{p}_G^{\mathrm{ML}}[t] \tag{30}$$

This is indeed a consistent estimator for the cross-entropy in the sense that, for an infinite corpus size it would converge on the true cross-entropy of the system. Furthermore, in such a limit condition, the cross-entropy itself would actually converge on the true value of the entropy (i.e., it would not even be an upper bound, but an exact estimator), because both probability estimators p_T and $\mathrm{p}_G^{\mathrm{ML}}$ would converge on p_G given an infinitely large treebank. It is, however, incorrect that the estimated $\hat{H}[G, \mathrm{ML}]$ would be an upper bound of the true entropy value. Rather than a cross-entropy—which would indeed be an upper bound—this is a sample-based ML estimator of the cross-entropy, and it is subject to exactly the same estimation bias (Miller 1955) as the entropy: It will be an underestimation. Furthermore, note that both p_T and $\mathrm{p}_G^{\mathrm{ML}}$ are ML approximations of p_G taken from two subsamples from the same population. One therefore should expect that $\mathrm{p}_T \propto \mathrm{p}_G^{\mathrm{ML}}$, and as a direct consequence, the cross-entropy estimator is in fact a Monte Carlo estimator of the ML entropy:

$$\hat{H}[G\|\mathrm{ML}] = H^{\mathrm{MC}}[G] \approx H^{\mathrm{ML}}[G] \tag{31}$$

Indeed, as we will see in the corpus analyses below, the estimators in Equation 28 and Equation 30 are virtually indistinguishable, and are both underestimations of the true entropy.

4.4 Smoothed Induced Treebank Entropy (SITE)

The crucial insight of the SITE measure is that the estimate of the PCFG entropy can be decomposed into $\|N\|$ local expansion entropy estimates. When inducing a grammar by ML from a treebank T, both the characteristic matrix ($\mathbf{M}_G^{\mathrm{ML}}$) and the local expansion entropies ($\mathbf{h_0}^{\mathrm{ML}}$) are ML estimates of their true values. On the one hand, the elements $m_{i,j}$ of $\mathbf{M}_G^{\mathrm{ML}}$ are counts of the average number of instances of non-terminal A_i that result from expanding an instance of non-terminal A_i. These counts should converge very fast; they are plain frequency estimates that will be unbiased even when estimated from quite small treebanks, and therefore do not need correction. Furthermore, by not modifying the characteristic matrix, one ensures that its spectral radius remains bounded by one as the result of ML PCFG induction always is (Sánchez and Benedí 1997; Chi and Geman 1998; Chi 1999), and hence the derivational entropies will also converge after correction (Grenander 1967, 1976). On the other hand, the expression for the local expansion entropies (Equation 7) is just the classical entropy equation. The $h_{T,0}[A_i]$ values are ML estimates of the uncertainty on the different ways in which symbol A_i can be expanded. As with any other ML entropy estimate, these should be expected to be negatively biased. This bias, however, is of the precise type that can be corrected with the sort of estimators given in Section 2.5.

Moscoso del Prado Martín (2014) used such an approach. In order to estimate the entropy of the grammar from which a treebank was sampled, PCFGs were induced from the treebanks, and then the plain ML estimates of the characteristic matrices were used, together with the CAE estimates of the local expansion entropy vectors for computing the derivational entropy of the grammars (comparing samples originating in different historical periods). Moscoso del Prado Martín (2017) also followed this approach, but using the CWJ estimator instead of CAE for smoothing the local expansion entropies. As I will demonstrate with corpus analyses, SITE with CWJ local entropy smoothing achieves extremely fast, unbiased convergence to the true values of the derivational entropy associated with a treebank.

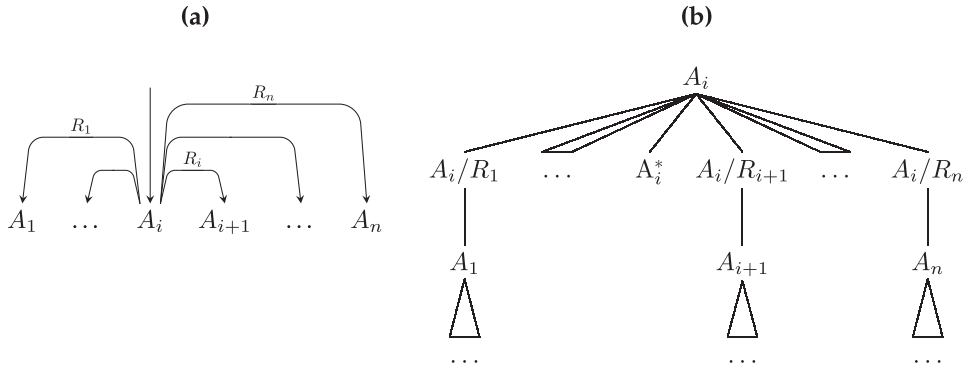

Figure 3
Expansion of node A_i with the relations depicted in (a) into the corresponding part context-free derivation tree in (b).

4.5 SITE on Dependency Treebanks

SITE can also be computed to dependency treebanks. To do this, one just needs an additional preprocessing step in which all dependency graphs in the treebank are converted to equivalent context-free derivation trees, as is illustrated in the change from the dependency structure in Figure 1c into the derivation tree in Figure 2b.

The procedure for converting a dependency graph into a derivation tree is a variation of the procedure described by Gaifman (1965), with the addition of nodes for representing the dependency labels. Variants of this procedure have also been used in recent studies (e.g., Rezaii et al. 2022). For a dependency graph with n nodes $A_1, \ldots, A_n$, and $n-1$ labeled relations R_1, R_2, …, R_{n-1}, one creates a context-free derivation tree with $2n$ internal nodes. Of these, n corresponds to the nodes in the original dependency graph (labeled $A_1, \ldots, A_n$), one is a root node A_0, and $n-1$ corresponds to the relations (labeled A_i/R_j, denoting a relation with label R_j going from node A_i to node A_j with label R_j). In addition, the tree has n leaves (labeled $\mathrm{A}_1^*, \ldots, \mathrm{A}_n^*$). The root node of the dependency graph is made to hang from the tree root A_0. Then for each node in the dependency graph, a corresponding context free branching is constructed, by making all the relations originating from it into daughters in the context-free graph, from which the corresponding nodes in the relation are made to hang. In addition, each internal node A_i has a leaf daughter A_i^*, positioned in its corresponding order in the original dependency graph. The process, illustrated in Figure 3, is then recursively repeated for each of the symbols that depend on A_i. For projective dependency trees, this procedure is a reversible one in the physical sense: The derivation trees constructed by this method preserve all information that was encoded in the original dependency structure. After applying this preprocessing to the corpus, a PCFG can be induced by the usual method, and its entropy can then be estimated by SITE.

5. Corpus Analyses

5.1 Accuracy of the SITE Method for Corpora with Known Entropy

In order to assess the accuracy and bias reduction of the SITE method in estimating the entropy of realistic treebanks with known true entropy value, I used the Wall Street

Journal subsample of the Penn TreeBank (Marcus et al. 1995) distributed with Python's Natural Language Toolkit (NLTK v3.9.1; Bird, Klein, and Loper 2009). This sample includes 3,914 parsed sentences. Importantly, the same sentences are available both as context-free derivation trees, and as dependency graphs. I induced a PCFG grammar from each version of the treebank. These grammars were used as the true underlying grammars, from which the task was estimating the true value of the derivational entropy.

This known entropy case is useful when evaluating the properties or ambiguity of existing grammars, parsing, and generation systems. In Section 5.2 I will study the more common case, where the actual true grammars are unknown, and so are their true derivational entropy values.

5.1.1 Context-Free Treebank Case. The context-free trees were pre-terminalized by removing the leaves of the original trees (corresponding to the actual words), so that the leaves of the transformed tree were part-of-speech (POS) tags. The PCFG induced by ML from this corpus contains a total of 8,009 context-free production rules using an alphabet of 662 non-terminal symbols, and it has a true entropy value of 103.445 bits (i.e., the equivalent of generating $1.38 \cdot 10^{31}$ distinct equiprobable derivation trees).

From the induced PCFG, I sampled artificial corpora of 24 increasing sizes, from 1 to 15,000 sentences (this number was chosen as it was found that this is how much the ML method required for converging on the true entropy value for this dataset), with the subsample sizes evenly spaced on a logarithmic scale (i.e., sample sizes of 1, 2, 3, 5, 7, 11, 17, 25, 37, 55, 82, 122, 183, 273, 407, 608, 908, 1,355, 2,023, 3,020, 4,509, 6,731, 10,048, and 15,000 sentences). From each of the generated corpora, I induced a PCFG and estimated their entropies using the plain ML (as in Hale 2003, 2006; Linzen and Jaeger 2014), and Monte Carlo ML (as in Corazza, Lavelli, and Satta 2013), and the SITE methods, using either the CAE (as in Moscoso del Prado Martín 2014) or CWJ (as in Moscoso del Prado Martín 2017) entropy smoothers from each of the samples. This process was repeated one hundred times.

Figure 4 plots the resulting entropy estimates. Panel (a) compares the accuracy of the three estimation methods for increasing sample sizes. Notice that both versions of the SITE method converge must faster than the ML estimates (which are indistinguishable from each other). Both SITE estimates approach the true value with just 122 sentences, and with just 55 sentences the true entropy value was within the 95% confidence interval for the mean estimate, something that took about 15,000 sentences to achieve with either of the ML methods (which are indistinguishable). In more extreme terms, the ML estimates with over one hundred sentences are about as biased as is the SITE estimate with a single sentence. In terms of overall error, the CS and CWJ methods achieve very similar performance. However, the CS-based method slightly overshoots the entropy estimates, and in practice is as biased (but in the opposite direction) as the ML methods are for samples of over 3,020 sentences. In contrast, the SITE method using the CWJ estimates completely eliminates the bias from 120 sentences onwards. Panel (b) illustrates how the ML methods do not achieve true convergence until every non-terminal symbol and production rule from the true grammar have been explicitly observed in the corpus.

5.1.2 Dependency Treebank Case. Each of the dependency graphs in the corpus was converted to a context-free derivation tree (the dependencies were considered as relating POS tags rather than words). From the resulting derivation trees, I induced a PCFG by ML, obtaining a PCFG with 8,104 context-free production rules using an alphabet

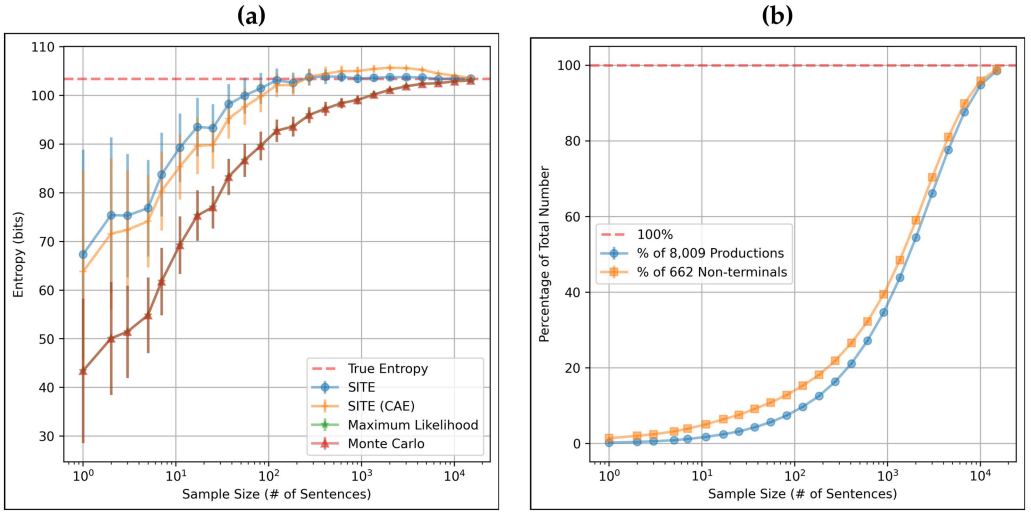

Figure 4
Results of Corpus Analysis 5.1.1. Error bars are 95% confidence intervals for the mean. Notice the horizontal logarithmic scales. (a) Estimated derivational entropies according to the four estimators as a function of the sample size. (b) Percentage of rules and non-terminal symbols from the original grammar that are documented as a function of sample size (the 95% C.I.s are imperceptible).

of 46 non-terminal symbols, and a true entropy value of 78.36 bits (i.e., the equivalent of generating $3.88 \cdot 10^{23}$ distinct equiprobable dependency trees; notice that the dependency paradigm is, for equivalent corpora, substantially more restrictive in its generalization than the context-free paradigm). As was done in Corpus Analysis 5.1.1, from the estimated PCFG, I sampled artificial corpora of the same 24 increasing sizes. The entropies of the induced PCFG were estimated using both ML methods, and two variants of SITE. This process was repeated one hundred times.

Figure 5 plots the resulting entropy estimates. Panel (a) compares the accuracy of the three estimation methods for increasing sample sizes. Notice that both versions of the SITE method converge much faster than the plain ML estimate, both approaching the true value with just 608 sentences, something that—as before—took about 15,000 sentences to achieve with the plain ML methods. In terms of overall error, the CS and CWJ methods once more achieve very similar performances, with the CS method again slightly overestimating the entropies. The SITE method using the CWJ estimates are very accurate from about 407 sentences, and it completely removes the bias from 908 sentences onwards. As before, Panel (b) shows that the ML methods do not achieve convergence until all rules and non-terminals of the grammar have been directly observed.

5.1.3 MLU-Derivational Entropy Correlations. In order to investigate whether the derivational entropy rate is relatively constant for a given grammatical theory, I compare the entropy estimates for the context-free and dependency versions of the samples used in Corpus Analyses 5.1.1 and 5.1.2 . These corpora are subdivided into 199 very small files, ranging in size from just a single sentence, to 185 of them (mean size 19.67 ± 1.64 sentences/file). For each of these files, I estimate the MLU, and derivational entropy

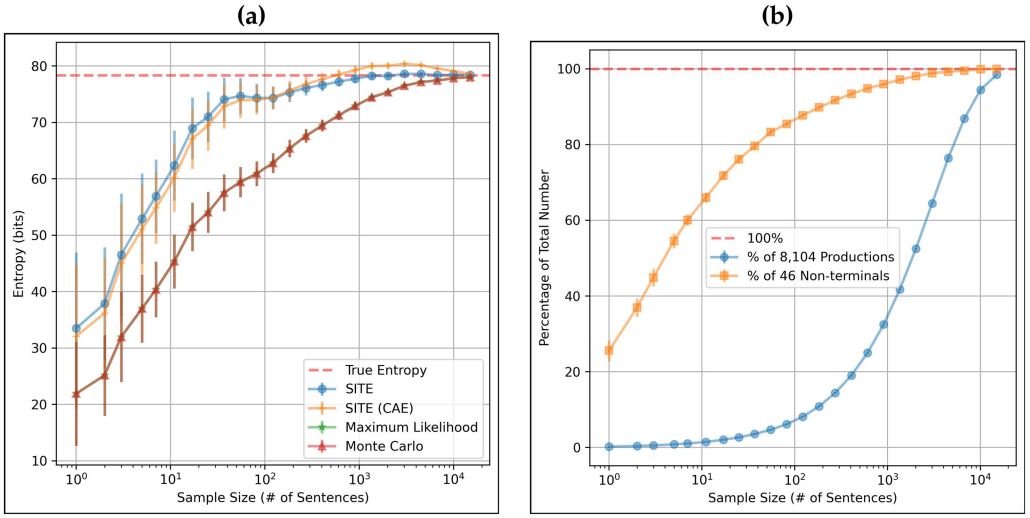

Figure 5
Results of Corpus Analysis 5.1.2. Error bars are 95% confidence intervals for the mean. Notice the horizontal logarithmic scales. (a) Estimated derivational entropies according to the four estimators as a function of the sample size. (b) Percentage of rules and non-terminal symbols from the original grammar that are documented as a function of sample size (the 95% C.I.s are imperceptible).

using SITE (based on pre-terminalized parse trees and dependency graphs), according to both versions of the corpora.

Panel (a) in Figure 6 plots the derivational entropies for each subcorpus as a function of their MLUs, for both the context-free (blue dots) and dependency (red dots) versions. There are clear correlations between the measures (Pearson's $r = .70, p < .0001$ for context-free versions, and Pearson's $r = .63, p < .0001$ for the dependency versions). Furthermore, both regression lines (dashed lines in the plot) converge almost exactly at the origin of coordinates, indicating the absence of any significant intercept (i.e., the regressions were fitted allowing for an intercept to be present, but they were insignificant; $1.29 \pm 3.81, t[197] = .34, p = .73$ for context-free and $1.18 \pm 3.29, t[197] = .359, p = .72$ for dependencies). This is in full agreement with the prediction of Equation 22, of MLU and derivational entropy being directly proportional. From the same equation, it follows that the regression slopes estimate the derivational entropy rates. If we refit the regressions to exclude the non-significant intercepts, the estimated derivational entropy rates are $h[G] \approx 2.05 \pm .03$ bits/symbol ($t[198] = 70.88, p < .0001$) for the context-free paradigm treebank, and $h[G] \approx 1.44 \pm .02$ bits/symbol ($t[198] = 57.84, p < .0001$) for the dependency treebank.

Although the regressions are exactly as we predicted, the overall correlation coefficients are nowhere as strong as the Pearson's $r = .98$ reported by Moscoso del Prado Martín (2017). Notice, however, that for such small corpora (the majority of which were well below one hundred sentences) even the fast-converging SITE is most likely severely underestimating the derivational entropies (the MLUs are, on the other hand, unbiased estimators). As illustrated by Figures 4 and 5, for very small corpora there is an approximately logarithmic relation between corpus size and the derivational entropy estimates, even after applying the SITE corrections. We can therefore correct the derivational entropy estimates by residualization: I take the residuals from a linear

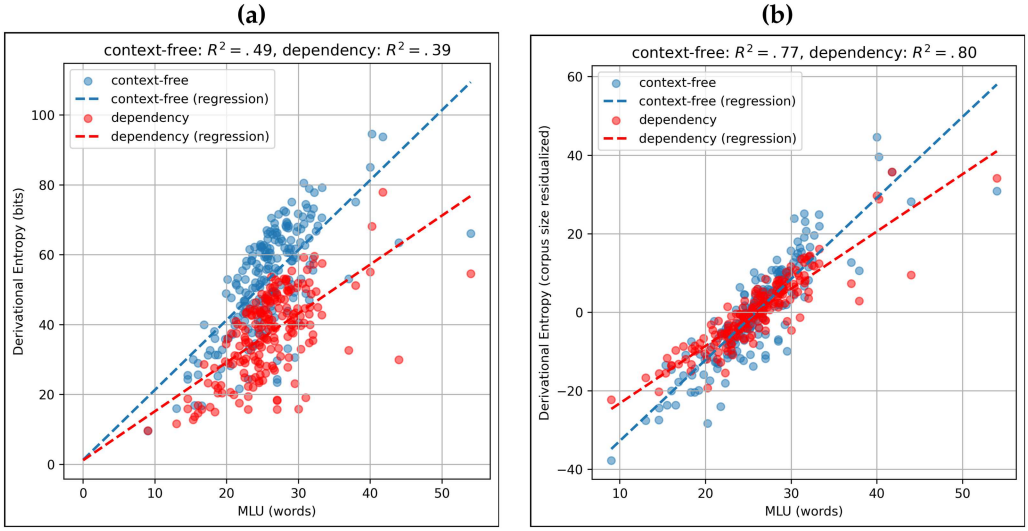

Figure 6
Results of Corpus Analysis 5.1.3. Each point plots the derivational estimates of a file as a function of the MLU, in the context-free (blue) and dependency (red) versions of the treebank. The dashed lines plot linear regressions. (a) Relationship between MLU and the raw derivational entropies. (b) Relationship between the MLU and the derivational entropies after residualizing log corpus size.

regression predicting the value of the derivational entropy as a function of the log corpus size. After this residualization, there was no additional non-linear monotonic relation between corpus size and either of the two residualized entropies, as revealed by the non-significant rank correlations (Spearman's $\rho = .05, p = .52$ for the context-free entropy, and Spearman's $\rho = .05, p = .44$ for the dependency entropy). Panel (b) in Figure 6 shows how, when the estimation biases are removed by residualization, the correlations between MLU and derivational entropy indeed become much stronger (Pearson's $r = .88, p < .0001$ for the context-free corpora, and Pearson's $r = .89, p < .0001$ for the dependency corpora), much more in line with the almost perfect relationship reported by Moscoso del Prado Martín (2017). It appears that the derivational entropy rates are roughly constant across the 199 subcorpora, depending just on which paradigm was used for annotating the parses. The difference in the value of the derivational entropy rate between the DG and CFG versions of the same corpus falsifies Hypothesis 1 of Section 3.3, that the constancy of the derivational entropy rate is a consequence of texts being in the same language. Even for identical texts, these entropy rates may change as a result of using different conventions for corpus annotation, as are in this case the CFG and DG syntactic annotation paradigms. These results, however, remain consistent with Hypothesis 2, that the derivational entropy rate remain constant within a language and annotation convention.

5.1.4 Relative Independence of Grammatical Formalism. The previous analyses found the derivational entropy rates to be roughly constant within a grammatical paradigm, consistent with Hypothesis 2 of Section 3.3. As argued in Section 3.4, this in turn predicts that the derivational entropies are indeed relatively independent of the grammatical formalism used for the treebank. In Equation 25, the dependency-based derivational

entropies should be directly proportional to the context-free ones by a factor equal to the ratio of their derivational entropy rates. Taking the values estimated in the previous analysis, the ratio between the two derivational entropy rates is expected to be:

$$\beta \approx \frac{1.44}{2.05} = .70 \tag{32}$$

In other words, the derivational entropy of the dependency annotated corpus should be about 70% of the derivational entropy of the same corpus annotated using the context-free paradigm.

Panel (a) in Figure 7 compares the SITE entropy estimates for each corpus file according to the two grammatical theories (CFG or DG). As was expected, although the individual entropy values are different from one formalism to the other, their relative values are very strongly correlated (Pearson's $r = .98, p < .0001$). Furthermore, the line plotting the regression between both derivational entropies once again seems to go almost exactly through the origin of coordinates. This is confirmed by the lack of a significant intercept in such regression ($-1.69 \pm 1.09, t[197] = -1.56, p = .12$). Refitting the regression to discard this non-significant intercept, one obtains an estimate of the relationship between the derivation entropies of $\beta \approx .71 \pm .01$ ($t[198] = 130.78, p < .0001$), almost exactly what was predicted above. In other words, just knowing the derivational entropy rates in both paradigms we could account for 96% of the variance of the derivational entropy rates of the dependency corpus using those from the context-free corpus (or vice versa).

It could be argued that such a strong correlation could arise merely from the differences in size of the subcorpora, as the smaller subcorpora will be subject to a

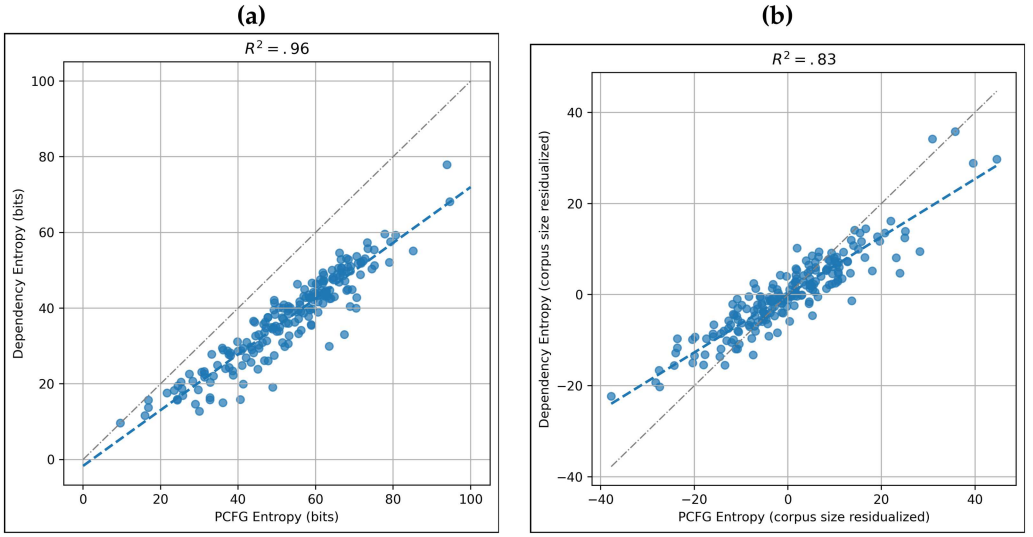

Figure 7
Results of Corpus Analysis 5.1.4. Each point plots the derivational entropy estimates of a file according to the context-free (horizontal axis) and dependency (vertical axis) versions. The dashed blue lines plot linear regressions, and black dashed lines plot the identity. (a) Relationship between the raw derivational entropies. (b) Relationship between the derivational entropies after residualizing log corpus size.

larger negative bias than the larger ones (i.e., in contrast with the previous section, the bias here acts in our favor). Indeed, both the context-free derivational entropy and the dependency-based one showed robust correlations with the logarithm of the corpus size (Pearson's $r = .56, p < .0001$, and $r = .69, p < .0001$, respectively). These correlations are, however, substantially weaker than that between the derivational entropies themselves. This makes it very unlikely that corpus size would account for the relationship between the two entropies. Nevertheless, to ensure that the relationship with corpus size did not confound the relations between the two entropies, I compared the corpus size residualized versions of the derivational entropies (see Analysis 5.1.3). Panel (b) in Figure 7 plots the correlation between the residualized derivational entropies. Their correlation is a conservative estimate of the true relationship between the two entropies (i.e., by attributing to corpus size the maximum possible amount of shared variance), and it is still extremely strong (Pearson's $r = .91, p < .0001$).

These analyses illustrate how, even if the generalization step might consider structures that would never be generated, the estimates of diversity are, in relative terms, not crucially dependent on this. The relative values of syntactic diversity are independent of whether one used CFGs or DGs to construct the treebanks. Still, one could argue that all samples used in these analyses were extremely similar, in that they were all extracts from The Wall Street Journal, dealing with similar topics in a more or less uniform language, hence the constancy of the derivational entropy rates and its consequences. In the following analysis we will investigate whether such relations hold also for extremely heterogeneous corpora.

5.2 A Large Heterogeneous Corpus with Unknown True Entropy

Above, it was demonstrated that the SITE method is able to correctly estimate the entropy of the grammar that generated a treebank. In most real world situations, not only is the true value of the entropy unknown, but the assumption that a large treebank was generated from a single grammar is quite probably untrue. Language is dynamic, and the grammar that generated a part of a corpus possibly differs from the grammar that generated other parts of it. This is most salient in diachronic treebanks that contain samples obtained from different speakers at very different historical periods. I use the Icelandic Parsed Historical Corpus (IcePaHC v2024.03; Rögnvaldsson et al. 2012) to investigate this situation. This treebank contains a longitudinal sample of Icelandic language spanning the history of the language: from the Old West Norse of the 12th century, to the contemporary Icelandic of the 21st century. The corpus contains a total of 73,051 sentences, parsed into context-free derivation trees by human annotators. The corpus is split into 61 files, each of which is a coherent text written by a single author. In addition, this corpus is also available annotated with a dependency grammar (Arnardóttir et al. 2020), as part of the Universal Dependencies project (UD; de Marneffe et al. 2021). Interestingly, although the original IcePaHC and UD versions are parses of the same texts, they exhibit substantial differences in what is considered a word and what counts as a sentence. For instance, the UD version contains a total of 44,029 sentences, about 40% fewer than the original IcePaHC. This is the result of the UD annotation conventions requiring that several "sentences" in the original IcePaHC are in fact taken as clauses within the same sentence in the UD annotation convention (Hinrik Hafsteinsson, personal communication).[1] It is interesting to investigate how the

1 26 November 2024.

different by-design definition of what counts as a word and what counts as a sentence would influence estimations of the derivational entropy and derivational entropy rate.

5.2.1 Divergence of Non-stationary and Heterogeneous Samples. As mentioned above, it would be quite unrealistic to assume that all, or even a large proportion, of the texts in IcePaHC originate from a single grammar. Therefore, one should not expect to find any single meaningful value for the derivational entropy of such (likely non-existent) grammar. To illustrate this point, I considered the corpus in an incremental manner, adding the trees in each file of the corpus one file at a time in chronological order. At each step, the pre-terminalized trees from a file were added to the corpus, and I estimated the derivational entropy (using SITE) of the process generating the corpus by inducing a grammar from the pre-terminalized trees.

The blue line in Figure 8 plots the resulting entropy estimates. The first noteworthy observation is that, instead of the roughly logarithmic, monotonically increasing patterns encountered in Corpus Analyses 5.1.1 and 5.1.2, the estimate of derivational entropy undergoes strong oscillations along the corpus. This non-monotonic pattern indicates a lack of convergence of the derivational entropy estimate for this dataset. Before assessing convergence, however, one needs to address the issue of **non-stationarity**. The analyses in Sections 5.1.1 and 5.1.2 contained sentences that had been sampled from the same grammar, and hence their statistical properties remained constant across the whole corpus. In contrast, in the current case, each file corresponds to a different author, in a different genre, written at very different historical periods. Hence, it is naïve to assume that the statistical properties remain in any way constant across the files. Such temporal variability of the sample's statistical properties is a textbook example of a non-stationary sequence. Among other problems, information-theoretical measures such as the entropy are meaningless in non-stationary situations.

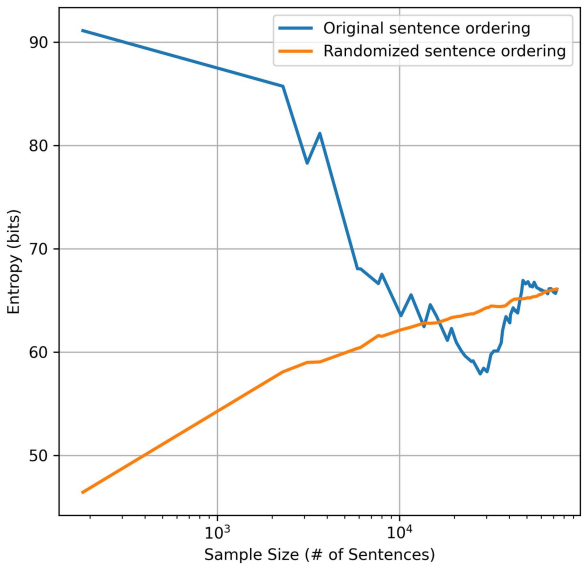

Figure 8
Results of Corpus Analysis 5.2.1. Convergence of the derivational entropies on IcePaHC sentence ordering across all its files. The line colors denote whether the original (blue line), or randomized sentence ordering (orange line) was used.

One can force stationarity onto this corpus by randomizing the order of the sentences across all the corpus files. Such transformation has the effect of making the statistical properties of the sentences constant across the corpus, and hence truly stationary. To illustrate the effect of such manipulation, I randomized the order of all corpus sentences, and divided the corpus into 61 chunks, each corresponding in number of sentences to one of the original corpus files. The orange line in Figure 8 plots the evolution of SITE estimates of derivational entropy as a function of corpus size on the randomized order corpus. Notice that, as expected, the randomization has brought back the monotonic pattern, which could lead to convergence. Note also that the end point for the randomized order estimates is 66.08 bits, identical to that obtained in the original order. The reason is that the entropy estimates themselves do not consider the order in which the sample was obtained. However, even after 73,051 sentences, SITE has not converged; this contrasts with the barely one hundred sentences that sufficed to achieve convergence on the context-free Corpus Analysis 5.1.1. Furthermore, considering the logarithmic horizontal scale of the graph, convergence—if at all achievable, of which I am skeptical—would require a corpus orders of magnitude larger, going into the hundreds of thousands or even millions of sentences. This indicates how, even with forced stationarity, the estimation methods detect the extreme level of variability stemming from the corpus heterogeneity, making it impossible to obtain a coherent entropy estimate.

5.2.2 Convergence of Stationary Homogeneous Subsamples. In contrast with the corpus as a whole, the individual files are homogeneous, corresponding to texts generated by a single speaker, at a single historical point, and in a given genre. In this case, it is quite plausible that the parses originate from one specific grammar, whose derivational entropy could indeed be estimated. In order to assess the convergence of the derivational entropy estimates for each of the subcorpora, I randomized the order of the sentences within each file. Each of the randomized-order files was subdivided into ten parts of equal length, and the SITE derivational entropies were computed incrementally for each of the files, adding one tenth of the file at a time.

As in the previous section, the order randomization forces stationarity within the individual files. This randomization implies discarding the discourse aspects of grammar use. Discourse makes samples of human language inherently non-stationary. For instance, if one simply looked at MLU, one would find that it tends to increase along a text. Our focus, however, is just syntax, narrowly defined in the classical sense as not extending beyond the sentence (see Du Bois [2014] for views that extend syntax beyond the individual sentence). By forcing stationarity we are discarding any supersentential context. This enables us to gauge the extent to which convergence for a single underlying source of syntactic rules can be achieved. Notice that trying to compute any entropy measure at a level higher than the syntax is very problematic; entropy is not even defined for non-stationary sequences.

Figure 9 plots the evolution of the SITE derivational entropy estimates. Except perhaps for some of the shortest files, they all seem to either have converged, or be very close to it. In contrast with the large heteregeneous corpus as a whole, the individual files are sufficiently homogeneous to be considered as originating each from a single grammar, whose derivational entropy can indeed be estimated.

5.2.3 Constancy of Derivational Entropy Rates across Heterogeneous Corpora. As seen above, each of the individual files in IcePaHC plausibly originates from a single grammar, whose derivational entropy converges. The question is whether the apparently constant

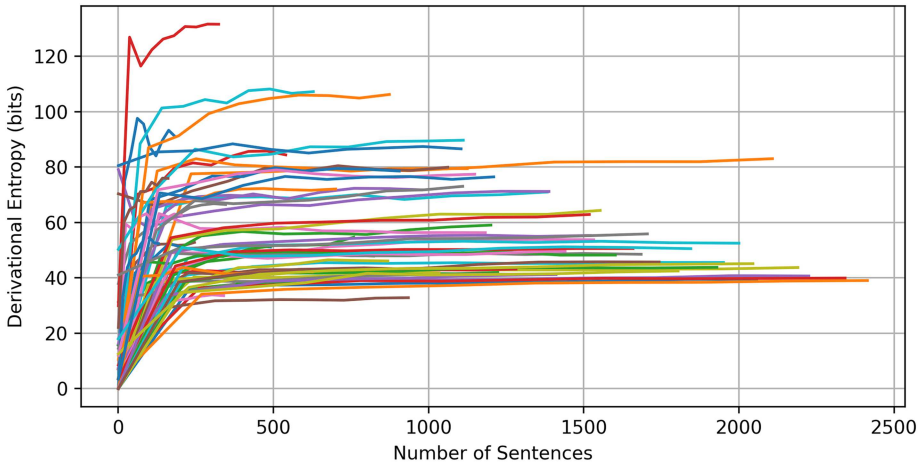

Figure 9
Results of Corpus Analysis 5.2.2. Convergenge of the SITE derivational entropy estimates for each of the individual documents in IcePaHC (each line corresponds to a single document).

derivational entropy rate will hold as well for these very heterogeneous samples. After all, it is indeed possible that the roughly constant derivational entropy rate found in Section 5.1.3 was the result of the Wall Street Journal samples used being extremely homogeneous.

The availability of the UD version of this treebank enables further exploration of the effects of heterogeneity. Although the actual texts of the files in the corpora are the same, the units of syntactic analysis—words and sentences—are defined using different criteria in these two versions. Panel (a) in Figure 10 compares the MLUs of the original IcePaHC files with those of their corresponding UD-parsed versions. As one would expect, combining multiple sentences from IcePaHC into single multi-clause sentences in the UD version results in the MLUs being substantially larger in the latter than in the former. As a result of the difference in criteria, the UD corpus has fewer sentences, and these sentences are longer. Some correlation between the MLUs across both versions does remain, but it only accounts for less than half of the MLU variance.

The derivational entropies were estimated using SITE for each of the files in each of the two corpora. The blue dots in panel (b) in Figure 10 plot the derivational entropies as a function of the corresponding MLU for the original CFG version. The correlation is remarkably strong (Pearson's $r = .99, p < .0001$). Following the predictions, the regression (blue dashed line) hits almost exactly on the origin of coordinates, reflecting the absence of a significant intercept term ($-1.63 \pm 1.47, t[59] = -1.115, p = .27$). Excluding the intercept, the slope of this regression gives us a derivational entropy rate of $h[G] \approx 3.56 \pm .02$ bits/symbol ($t[60] = 139.60, p < .0001$). Turning to the UD version of the corpus files, the correlation between their MLUs and derivational entropies (red dots in Figure 9b) was, as for the original IcePaHC, almost perfect (Pearson's $r = .99, p < .0001$). As before, the intercept term in the regression (red dashed line in the figure) was insignificant ($1.69 \pm 1.58, t[59] = 1.07, p = 0.29$), and the estimated derivational entropy rate was $h[G] \approx 3.59 \pm .02$ bits/symbol ($t[60] = 189.67, p < .0001$) after removing the intercept term. The degree of overlap between the estimates based on the two versions of the corpora is striking. Their derivational entropy rates are not just clearly constant within the files in each corpus, but they are also constant between

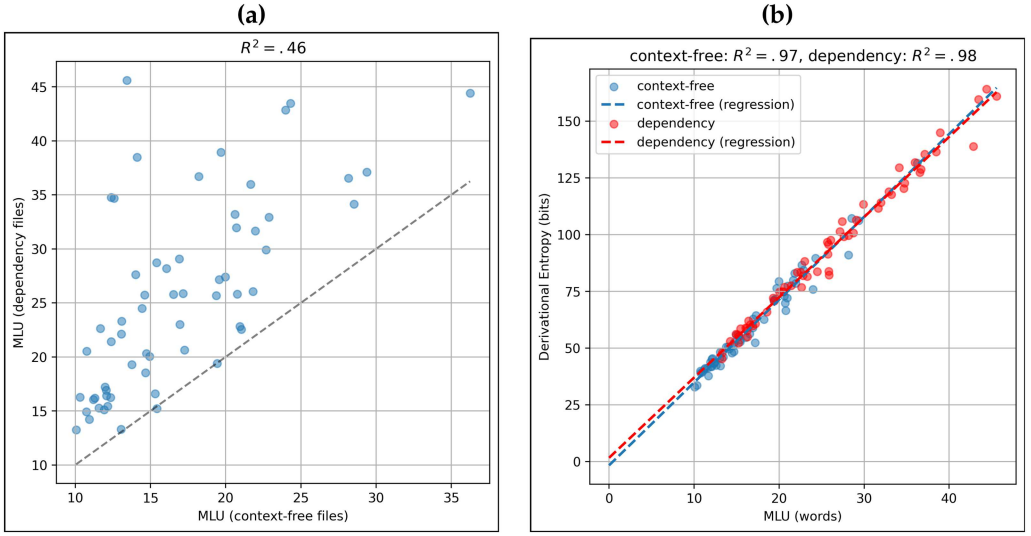

Figure 10
Results of Corpus Analysis 5.2.3. (a) Relationship between the MLUs of each of the 61 individual files of IcePaHC, comparing the CFG and UD versions of the corpus. The dashed line is the identity. (b) Relationship between the MLU and the derivational entropies for each file in the original IcePaHC (blue dots and regression line) and the UD version of the corpus (red dots and regression line).

the two corpora, their difference only noticeable on their second decimal digits, and their standard errors overlapping. From this, if we had two files with equivalent MLU, one from the CFG and another from the UD versions of the corpora, we can predict that their regression would have a slope of almost exactly one (i.e., 1.01 or .99 depending on the directionality in which one wants to predict), indicating that their derivational entropies would be, for all practical purposes, identical.

It appears, therefore, that the derivational entropy rate stays constant within a given annotation scheme, which is consistent with Hypothesis 2 of Section 3.3. This is found in both of the corpora described here, as well as in the two corpora analyzed in Section 5.1, and in the Switchboard I corpus studied in Moscoso del Prado Martín (2017). That the derivational entropy rate remains constant across the extremely diverse samples contained in IcePaHC is remarkable. These samples come from very different historical periods (ranging across nine centuries and three distinct stages of the Icelandic language) and also very different registers (spanning viking sagas, religious texts and sermons, natural sciences texts, etc.). Such heterogeneity is bound to induce rather different distributions of rule probabilities, which nevertheless result in the very same derivational entropy rate. Moreover, the degree of overlap found between the two versions of IcePaHC cannot be considered a mere coincidence. In fact, the two versions of IcePaHC were not separately annotated according to different guidelines. Rather, an automatic conversion software (*UDConverter*; Arnardóttir et al. 2020) was developed ad hoc for creating the dependency version of the treebank automatically from the context-free original. That the change of grammatical paradigm was performed without any loss of information (i.e., keeping the derivational entropy rate constant), bears witness to the quality of this conversion tool.

6. General Discussion

6.1 Convergence of SITE

I have shown that SITE provides an accurate estimation for the derivational entropy of the grammar from which a corpus has been sampled. This confirms the validity of the methods originally introduced by Moscoso del Prado Martín (2014, 2017), providing an explicit account of the circumstances under which the estimator can or cannot converge, as well as extending the methods to dependency treebanks. The analyses in Section 5.1 demonstrated that, even for large, realistic grammars—either context-free or dependency—the derivational entropy converges very fast to its true value, needing just over one hundred sentences in the context-free case, and closer to one thousand in the dependency case. In both cases, this is a substantial improvement over ML methods, which would only approach convergence with more than an order of magnitude additional sentences. This makes SITE suitable for investigating diversity in the small samples often available in many fields.

When one assumes a sample originates from a probabilistic grammar, of whichever type, one is stating that there exists a particular stable probability distribution for the syntactic regularities found in the sample. When the sample comes from heterogeneous sources, however, such assumption is not realistic. As seen in the analysis in Section 5.2.1, in high heterogeneity situations, SITE fails to converge, even with samples of tens of thousands of sentences. The lack of convergence remains even after ensuring stationarity through sentence order randomization. In itself, the lack of convergence of SITE becomes a useful tool for assessing the homogeneity of treebank samples, prior to any additional inference one intends to make on the grammar(s) that might have generated them. Note however, that when much smaller, but homogeneous, samples from the same corpus are considered separately (Section 5.2.2), convergence is indeed achieved quite quickly.

6.2 MLU is the Primary Measure of Syntactic Diversity ...

It is not without some irony that, after using substantial theoretical apparatus for developing a sophisticated, information theoretical measure of syntactic diversity, good, old, and humble MLU (Nice 1925) comes out not just unscathed, but rather reinforced on both theoretical and empirical grounds. I have demonstrated that MLU is more than just a "proxy" measure of syntactic diversity. Instead its value is *fundamentally* linked to that of the derivational entropy. It is, therefore, an explicit measure of syntactic diversity insomuch as the derivational entropy is. The strong intercorrelation between both measures is theoretically predictable to be a linear relationship without an intercept, that is, a plain direct proportionality (Section 3.1). The corpus analyses (Sections 5.1.3 and 5.2.3) confirmed this relationship, replicating the almost perfect correlation between MLU and derivational entropy that had been previously observed (Moscoso del Prado Martín 2017) in additional corpora.

In light of these findings, MLU appears to be the most suitable measure of syntactic diversity in most contexts. Sure enough, one can easily find *individual* examples where an increase in sentence length does not require a more complex grammar. However, such increases are not tenable in macroscopic terms. For grammars that are actually productive, those cases become mere anecdotes swamped by the strong correlation. That a substantial realistic sample (i.e., not a set of cherry picked sentences) in one language has a higher MLU than another, implies beyond doubt a higher grammatical

diversity. As discussed earlier, MLU has repeatedly proven its worth in investigating multiple aspects of language, from acquisition, to aging, and disease. It is by far the simplest measure to compute, requiring very little labeling of corpora, and few—if any—theoretical commitments. It is also unbiased, and it converges on rather small samples (Casby 2011). Although the units of measurement of MLU are in general not crucial (Parker and Brorson 2005), one might want to change the unit of measurement of MLU depending on the typological properties of the languages under study. For instance, in polysynthetic languages, where the definition of what counts as "one word" is less than clear, a morpheme-based or even syllable-based count would be a more natural choice (see, e.g., Allen and Dench 2015, for Inuktitut). Whatever the units used, *MLU directly measures derivational entropy* in those units. These units map linearly into classical information theoretical units by a factor corresponding to the derivational entropy rate.

6.3 . . . Complemented by the Derivational Entropy Rate

I have introduced a new measure, the derivational entropy rate, that provides a natural complement to MLU. In domains where the actual value of the derivational entropy is required explicitly, the derivational entropy rate becomes of crucial importance. As seen in the corpus analyses, and also in the results of Moscoso del Prado Martín (2017), this measure is constant for corpora of a single language annotated using the same guidelines, supporting Hypothesis 2 of Section 3.3 (and falsifying Hypothesis 1). Therefore, for estimating the actual derivational entropy of a group of texts within a language, it suffices with estimating it using SITE for a single annotated text (ensuring its convergence). From this text, the derivational entropy ratio can be computed as a simple division. That single derivational entropy rate can then be used in conjunction with the MLUs to accurately reconstruct the derivational entropies of the remaining texts. These additional texts do not even need to be syntactically annotated; it is sufficient to *assume* they would be annotated *using the same convention*. This approach is crucial when comparing the complexity of samples in different languages; as the correlation between length and derivational entropy is expected to vary across languages, one needs to choose (theoretically motivated) syntactic annotation schemes for each language that can be considered comparable. Then, the derivational entropies of the text can be explicitly compared across languages.

Another domain that requires the actual explicit derivational entropy rate values is for the estimation of grammar ambiguity and expected parsing difficulty (Corazza, Lavelli, and Satta 2013). The derivational entropy rate of a grammar is the rate at which the diversity of the structures of the grammar grows when adding one symbol to a sequence. Derivational entropy rate is closely related to the **entropy rate** (Shannon 1948) of the stochastic language generated by the grammar's closure (i.e., the infinite process generated by concatenating strings randomly sampled from the grammar; Soule 1974). In PCFGs, the entropy rate is given by

$$h_s[G] = \frac{H_s[G]}{\mathrm{MLU}[G]} \geq 0 \tag{33}$$

where $H_s[G]$ is the **sentential entropy** of the grammar, that is, the entropy of the distinct strings (as opposed to the distinct trees) that the grammar generates (see Kuich 1970 for properties). A grammar's sentential entropy is upper-bounded by the derivational

entropy ($H_s[G] \leq H[G]$), with equality if and only if the grammar is unambiguous. The difference

$$R[G] = H[G] - H_s[G] \geq 0 \tag{34}$$

is referred to as the **equivocation** of the grammar; it measures the degree to which a grammar is ambiguous (Soule 1974). By parallelism, we can define the **equivocation rate** of a grammar,

$$r[G] = \frac{R[G]}{\text{MLU}[G]} = h[G] - h_s[G] \geq 0 \tag{35}$$

which is zero if and only if the grammar is unambiguous. Estimation of $r[G]$ from a corpus is possible given an adequate estimate of the entropy rate $h_s[G]$ (which cannot be computed exactly from a grammar). Whenever the objective is to compare the parseability of multiple corpora that use the same annotation scheme, one only needs to estimate the derivational entropy rate for one of the corpora (it will be the same for all of them), and then individually estimate the entropy rates $h_s[G]$, greatly simplifying algorithms such as that proposed by Corazza, Lavelli, and Satta (2013). Notice that, if, as hypothesized above, the derivational entropy rate $h[G]$ is fully determined by a specific syntactic annotation convention, such convention also sets an upper bound for the degree of ambiguity that might be found.

This has one important consequence: Within a given annotation convention, the only ways to reduce the syntactic ambiguity would be either using shorter sentences (decreasing the MLU), or—counterintuitively—*increasing* the sentential entropy rate ($h_s[G]$). In other words, the only way of reducing the syntactic ambiguity of sentences of a given length would be to increase the unpredictability of the sequence of symbols; unpredictable sequences of words should be syntactically less ambiguous than predictable ones. This has important implications both for natural language processing (e.g., Corazza, Lavelli, and Satta 2013), and for human language comprehension and production (e.g., Hale 2003, 2006; Linzen and Jaeger 2014; Sy et al. 2023) that are well worth studying further.

6.4 A Conjecture: Consistent Families of Grammars and Annotation Invariance

Above, I have found support for Hypothesis 2 of Section 3.3. The derivational entropy rate is constant across diverse subcorpora in four different corpora: the CFG and DG versions of the Penn Treebank subsample, the original IcePaHC, and its UD version. In addition, this constancy was also found across the conversations in the Switchboard I corpus (Moscoso del Prado Martín 2017). In summary, I have repeatedly found this in all corpora I have investigated, including some additional ones not reported here for the sake of brevity.

When researchers define the convention that will be used for syntactically annotating a treebank, they are implicitly defining a family of possible grammars (and discarding many others). For instance, in the case of Icelandic, the guidelines that were given to the individual IcePaHC annotators defined: an alphabet (composed of the allowed pre-terminal symbols), a set of non-terminal symbols N, a root symbol, and a set of *possible rules*. It is not that the guidelines explicity define which individual rules are "legal," but they constrain which types of rules can or cannot be applied, and how specific grammatical phenomena should be annotated.

Any PCFG can be regarded as the combination of a **skeleton**, Skel[G], an ordinary (non-probabilistic) context-free grammar G with rules of the form $r \equiv A \rightarrow \alpha$, $A \in N$, $(\alpha \in T \cup N)^*$, combined with a **distribution**, Dist[G], a numerical vector containing the probabilities associated with the individual rules (Soule 1974). Two PCFGs are said to be **similar** if they share the same skeleton, differing only in their distribution (i.e., they have the same terminal and non-terminal alphabets, and the same rules). For any PCFG we can define the set (Soule 1974),

$$\mathbf{D}_f[G] = \{Dist[G'] \,|\, G' \text{ is similar to } G \text{ and } \rho(\mathbf{M}_{G'}) < 1\} \tag{36}$$

where $\rho(\mathbf{M}_{G'})$ denotes the spectral radius (i.e., the largest absolute value of any eigenvalue of $\mathbf{M}_G$) of the characteristic matrix of G', which by being bounded by 1, guarantees that both MLU and derivational entropies will be finite (Grenander 1967, 1976; Hutchins 1972; Soule 1974).

One can take the rule constraints in an annotation guideline as defining a, possibly infinite, set of grammar rules, which can be taken as a skeleton Skel[G] for a finite family of grammars $\mathbf{D}_f[G]$. Given a grammar skeleton Skel[G], I define a **consistent family of grammars** (with tolerance ε) as

$$C[G, \varepsilon] = \{\text{Dist}[G'] \,|\, \text{Dist}[G'] \in \mathbf{D}_f[G] \text{ and } \alpha - \varepsilon \leq h[G'] \leq \alpha + \varepsilon\} \tag{37}$$

where $\alpha \gg \varepsilon > 0$ are fixed parameters. I will call a set of guidelines for syntactically annotating a corpus **ε-consistent** if it defines a consistent family of grammars for some value of α. I conjecture that adequate corpus annotation schemes and guidelines are consistent in this sense.

When one has two alternative consistent annotation schemes C_1 and C_2, one can compare the values of their derivational entropy rates α_1 and α_2. If $\alpha_1 > \alpha_2$, that is, sentences annotated using C_1 have, on average, higher derivational entropies than the same sentences annotated according to C_2, then this implies that annotating a sentence using C_2 is an irreversible process with respect to its annotation in C_1, that is to say, this process loses information about possible additional parses of the same sentence. On the other hand, by its effect in reducing the equivocation rate, the syntactic ambiguity will be smaller for sentences parsed using C_2 than it would have been with C_1, and this implies that parsing using C_2 will be easier and more accurate than using C_1 (Corazza, Lavelli, and Satta 2013).

Importantly, when several linguistic samples or corpora are annotated with two different ε-consistent annotation schemes, their derivational entropies will exhibit *annotation invariance*: The specific annotation scheme used is irrelevant (up to a precision dictated by ε) for the relative values of the derivational entropies, namely, $H[G_1] \approx \alpha_1/\alpha_2 H[G_2]$. In other words, when comparing the syntactic diversities of samples within a consistent scheme, the choice of specific scheme will not affect the results. In relative terms, syntactic diversity is an inherent property of the linguistic samples, not of the specific theoretical choice of syntactic representation.

Appendix A. Exact Values of the Derivational Entropy and MLU of a PCFG

The concept of derivational entropy, and its exact calculation is originally due to Grenander (1967, Theorem 4.2), which is difficult to access today, but is still accessible in Grenander (1976, Theorem 10.6, pp. 88–90).

Consider $A_i \in N$ a non-terminal symbol of a PCFG (T, N, S, R). Let $R_i \subset R$ denote the subset of the production rules that expand the non-terminal A_i,

$$R_i = \{r_{i,1} \equiv p_{i,1} : A_i \rightarrow \alpha_{i,1}, \ldots, r_{i,\|R_i\|} \equiv p_{i,\|R_i\|} : A_i \rightarrow \alpha_{i,\|R_i\|}\},\ \alpha_{i,j} \in (T \cup N)^* \quad \text{(A.1)}$$

It is obvious that the derivational entropy produced by symbol A_i must equal that produced by the rules that can expand it. In turn, by the decompositional properties of entropy (Shannon 1948), there are two additive components to the entropy generated by A_i:

(i) The entropy of the choice of rules that can expand A_i, which is fully determined by the rule probabilities

$$h_0[A_i] = -\sum_{j=1}^{\|R_i\|} p_{i,j} \log p_{i,j} \quad \text{(A.2)}$$

This is what I refer to as the *local expansion entropy* in the main text.

(ii) The context-free property forces that the derivational entropy of a non-terminal remains constant independently of the context in which it is applied. Therefore, for each rule $A_i \rightarrow \alpha_{i,j}$, we obtain a term

$$B_{i,j} = \sum_{A_k \in N} f(A_k; \alpha_{i,j}) H[A_k] \quad \text{(A.3)}$$

where $f(s; \alpha)$ denotes the number of times non-terminal symbol $s \in N$ occurs in the string $\alpha \in (T \cup N)^*$. In addition, the contribution of each term $B_{i,j}$ needs to be weighted by the probability of the corresponding rule ($p_{i,j}$).

Combining both sources of derivational entropy above, for each non-terminal in the grammar, we obtain an equation of the form

$$H[A_i] = h_0[A_i] + \sum_{r_{i,j} \in R_i} p_{i,j} B_{i,j} \quad \text{(A.4)}$$

In Equation A.3, each term $B_{i,j}$ is a linear combination of terms $H[A_k]$, with a coefficient for each non-terminal symbol in N. We can therefore rearrange Equation A.4 as

$$H[A_i] = h_0[A_i] + \sum_{A_j \in N} \underbrace{\left(\sum_{r_{i,k} \in R_i} p_{i,k} f(A_j; \alpha_{i,k}) \right)}_{m_{i,j}} H[A_j] \quad \text{(A.5)}$$

so that the coefficients $m_{i,j}$ are exactly those of the characteristic matrix of the PCFG ($\mathbf{M_G}$). Writing the resulting system of equations in matrix form

$$\mathbf{H} = \mathbf{h_0} + \mathbf{M_G} \cdot \mathbf{H} \tag{A.6}$$

and then rearranging

$$\mathbf{h_0} = (\mathbf{I} - \mathbf{M_G}) \cdot \mathbf{H} \tag{A.7}$$

we obtain a system of linear equation whose possible solution must satisfy

$$\mathbf{H} = (\mathbf{I} - \mathbf{M_G})^{-1} \cdot \mathbf{h_0} \tag{A.8}$$

provided that the inverse exists, which will be the case whenever the spectral radius is smaller than one, $\rho(\mathbf{M}_G) < 1$, (Grenander 1976).

The derivation of the MLU is identical to the above, with the change that the expansion component above should correspond to the expected number of terminal symbols directly generated by each rule

$$\ell_0[A_i] = \sum_{r_{i,j} \in R_i} p_{i,j} \sum_{s \in T} f(s; \alpha_{i,j}) \tag{A.9}$$

Bearing this in mind, following the same reasoning as we did for the derivational entropies, one gets to

$$\boldsymbol{\ell} = (\mathbf{I} - \mathbf{M_G})^{-1} \cdot \boldsymbol{\ell_0} \tag{A.10}$$

Alternative derivations of the equations above, based on the convergence of matrix series, can be found in Hutchins (1972, for MLU) and Wetherell (1980, for derivational entropy).

Appendix B. Bias-reduced Entropy Estimators

The Coverage-Adjusted Entropy Estimator (CAE; Chao and Shen 2003) directly addresses the two sources of bias in ML entropy estimates: On the one hand, when the probabilities used in the definition of entropy are ML estimates, these are themselves positively biased. The relative frequencies of events only take into account those events that are actually attested in the sample. For an event observed f_i times on a sample of size n, its ML probability estimate is given by $\hat{p}_i = f_i/n$, so that the sum of all $\hat{p}_i$ over the V different types attested is one:

$$\sum_{i=1}^{V} f_i = n, \qquad \sum_{i=1}^{V} \hat{p}_i = \sum_{i=1}^{V} \frac{f_i}{n} = 1 \tag{B.1}$$

Of course, this expression distributes among the V observed types all the probability mass that actually corresponds to those types that are not impossible, but is not documented in the sample. It is long known that the ML probability estimates can be

optimally corrected using the Good-Turing (GT; Good 1953) estimates, which are given by[2]

$$\tilde{p}_i = \left(1 - \frac{f_1}{n}\right) \hat{p}_i \tag{B.2}$$

where f_1 denotes the number of types that were observed exactly once in the sample (i.e., the number *hapax legomena*). The GT estimates reduce the value of the ML estimates, in turn increasing the value of the ML entropy estimate, but still underestimating the true entropy values. If H denotes the true entropy value, and $\mathbb{E}\left[H_{\text{ML}}\right]$ and $\mathbb{E}\left[H_{\text{GT}}\right]$ are, respectively, the expected values of the ML and GT estimates (which are obtained by plugging the corresponding probability estimate into the classical entropy equation):

$$\mathbb{E}\left[H_{\text{ML}}\right] \leq \mathbb{E}\left[H_{\text{GT}}\right] \leq H \tag{B.3}$$

Although both estimators are consistent, converging to the true H value with probability 1 for infinite sample sizes, they are both negatively biased.

The other source of bias, shared by the ML and CAE estimators, are the missing terms (corresponding to unseen, but possible, events) in the classical entropy equation. CAE also accounts for this by, on top of using GT-corrected probability estimates, changing the entropy equation itself,

$$H_{\text{CAE}} = -\sum_{i=1}^{k} \frac{\tilde{p}_i \log(\tilde{p}_i)}{1 - (1 - \tilde{p}_i)^n} \tag{B.4}$$

Notice that the numerators of this expression are the usual terms in the entropy equation. The denominators are an additional correction to account for the missing terms. Using this equation has been shown to produce entropy estimates that are less biased and faster converging that both ML and GT estimates (and in fact many other methods [see Vu, Yu, and Kass 2007]).

Chao and her colleagues have more recently proposed a new improved estimator (CWJ; Chao, Wang, and Jost 2013). This estimator exploits properties of the accumulation curve of the number of species observed in an ecosystem (i.e., the species accumulation curve). It is given by

$$H_{\text{CWJ}} = \sum_{1 \leq F_i \leq n-1} \left[\frac{F_i}{n}\left(\sum_{k=F_i}^{n-1} \frac{1}{k}\right)\right] - \frac{f_1}{n}(1 - A)^{1-n}\left[\log(A) + \sum_{r=1}^{n-1} \frac{1}{r}(1 - A)^r\right] \tag{B.5}$$

where F_i are the frequencies observed, and

$$A = \begin{cases} \frac{2f_2}{(n-1)f_1 + 2f_2} & \text{if } f_2 > 0 \\ \frac{2}{(n-1)(f_1 - 1) + 2} & \text{if } f_2 = 0, f_1 > 0 \\ 1 & \text{if } f_1 = f_2 = 0 \end{cases} \tag{B.6}$$

2 This is a simplified version of the full Good-Turing estimator, which is found to be sufficient for entropy estimation (Chao and Shen 2003).

with f_1 and f_2 being the number of types that where encountered exactly once or twice respectively (i.e., the numbers of *hapax legomena* and *dis legomena*).

Appendix C. Distribution of the Cosines Between Non-negative Vectors

Here, I provide an outline of the theoretical derivation of the mean and variance of the cosine between non-negative vectors, as well as the convergence of their distribution to a normal distribution, together with a simulation ascertaining the validity of approximations. For a detailed derivation of the equations in this appendix, stepwise solutions of the relevant integrals, and so forth, see Moscoso del Prado Martín (2025).

Consider two randomly chosen n-dimensional vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_+$. As we are interested in the possible angles between the vectors, we should assume that the angles formed by the vectors with the coordinate axes are uniformly distributed. This is equivalent to considering vectors lying on the non-negative orthant of the zero-centred unit n-sphere (Marsaglia 1972). Under this assumption, it is easy to show that each component x_i or y_j of the vectors is distributed according to the density function

$$f(x_i; n) = \frac{2}{\mathrm{B}\left(\frac{1}{2}, \frac{n-1}{2}\right)} (1 - x_i^2)^{\frac{n-3}{2}} \tag{C.1}$$

where B denotes the Beta function

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} \mathrm{d}t \tag{C.2}$$

The components of vectors lying on the unit n-sphere are direction cosines, and as a result, the angle between the two vectors is given by the sum of the products of their components

$$\cos(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i \, y_i \tag{C.3}$$

Bearing in mind that all components of both vectors are identically distributed following Equation C.1, and $\mathbf{x}$ and $\mathbf{y}$ are independent of each other, one can predict the mean value of the cosine to be exactly

$$\begin{aligned}
\mathbb{E}[\cos(\mathbf{x}, \mathbf{y})] &= \mathbb{E}\left[\sum_{i=1}^{n} x_i \, y_i\right] = n \, \mathbb{E}\left[x_1 \, y_1\right] \\
&= n \, \frac{4}{\mathrm{B}^2\left(\frac{1}{2}, \frac{n-1}{2}\right)} \int_0^1 \int_0^1 (1 - x_1^2)^{\frac{n-3}{2}} (1 - y_1^2)^{\frac{n-3}{2}} x_1 y_1 \, \mathrm{d}x_1 \, \mathrm{d}y_1 \\
&= \frac{4n}{(n-1)^2 \mathrm{B}^2\left(\frac{1}{2}, \frac{n-1}{2}\right)}
\end{aligned} \tag{C.4}$$

Importantly, Equation C.4 converges very rapidly on $2/\pi$. To see this, one can use Stirling's approximation for the Beta function

$$\lim_{n\to\infty} \frac{4n}{(n-1)^2 \mathrm{B}^2\left(\frac{1}{2}, \frac{n-1}{2}\right)} \approx \lim_{n\to\infty} \frac{2n}{\pi(n-1)} = \frac{2}{\pi} \tag{C.5}$$

That is, the expected value of the cosine between non-negative vectors is a positive constant, and—for moderate to large n—independent of the dimensionality.

In addition, the variance of the cosine is given exactly by

$$\mathrm{Var}[\cos(\mathbf{x}, \mathbf{y})] = \frac{n+1}{n} \mathrm{I}_{\frac{1}{2}}\left(\frac{n-1}{2}, \frac{n+3}{2}\right) - 2\left(\mathrm{I}_{\frac{1}{2}}\left(\frac{n-1}{2}, \frac{n+1}{2}\right)\right)^2 \tag{C.6}$$

where I_x denotes the regularized incomplete Beta function, for $0 \leq x \leq 1$,

$$\mathrm{I}_x(\alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\, \mathrm{d}t \tag{C.7}$$

For moderate to large values of n, Equation C.6 is very well approximated by

$$\mathrm{Var}[\cos(\mathbf{x}, \mathbf{y})] \approx \left(\frac{4}{3\pi\sqrt{n}}\right)^2 \tag{C.8}$$

indicating that the value of the cosine is expected to be increasingly clustered around the mean with increasing dimensionality.

Finally, as seen in Equation C.3, the cosine itself is ultimately the result of a sum of weakly correlated terms, with the number of terms being the dimensionality of the space. Therefore, by the Central Limit Theorem one should expect it to ultimately converge to a normal distribution such that

$$\cos(\mathbf{x}, \mathbf{y}) \xrightarrow{D} \mathrm{N}\left(\frac{2}{\pi}, \frac{4}{3\pi\sqrt{n}}\right) \tag{C.9}$$

where $\xrightarrow{D}$ denotes convergence in distribution.

To investigate for what range of dimensionality the approximations above are valid, I performed a simulation study. For each of 21 increasing dimensionalities (n= [2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30,40, 80, 160, 320, 640, 1,280, 2,560, 5,120, 10,240, 20,480]), I randomly sampled 1,000 pairs of vectors $\mathbf{x_i}, \mathbf{y_i} \in \mathbb{R}^n_+$, distributed according to a multi-dimensional normal $\mathrm{N}(\mathbf{0}, \mathbf{I_n})$ truncated to the non-negative orthant, which were then projected onto the surface of the unit radius, $\mathbf{0}$-centered n-sphere. For each pair, I computed the cosines,

$$\cos(\mathbf{x_i}, \mathbf{y_i}) = \mathbf{x_i} \cdot \mathbf{y_i} \tag{C.10}$$

and estimated the corresponding means and standard deviations. For each dimensionality value, I also computed an omnibus test for deviation from normality (D'Agostino and Pearson 1973), and corrected the 21 resulting p-values controlling for multiple comparisons using the False Discovery Rate (Benjamini and Hochberg 1995). As expected,

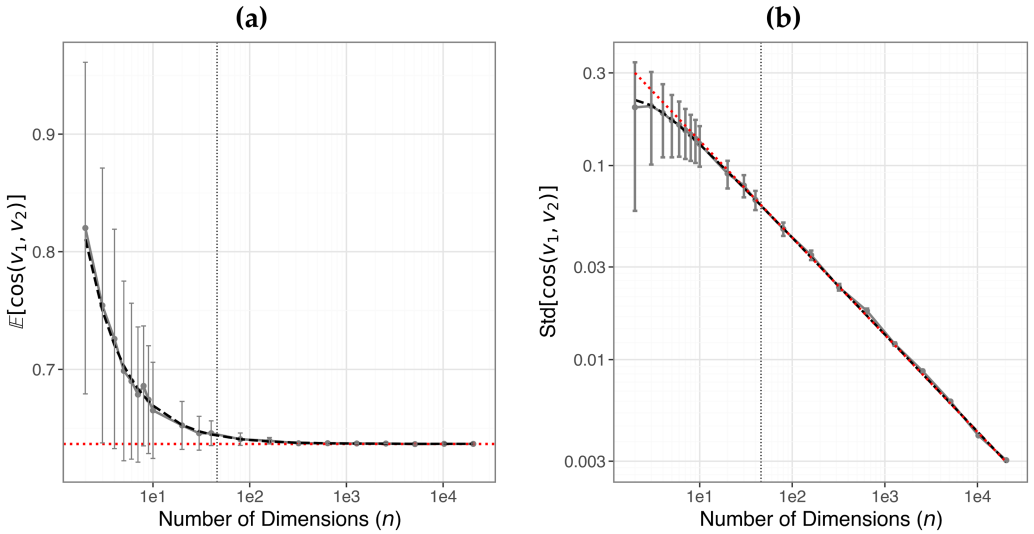

Figure C1
Results of simulation in Appendix B. (a) Accuracy of the mean cosine. The grey dots and curve plot the simulation result (error bars are standard errors). The dashed black line is exact theoretically predicted value (Equation C.4), and the red dotted line plots the predicted asymptote at $2/\pi$. The vertical dashed line indicates the size of the smallest grammars in our studies. Note the logarithmic horizontal axis. (b) Accuracy of the standard deviation of the cosine. The gray dots and curve plot the simulation result (error bars are standard errors). The dashed black line is the theoretically predicted standard deviation (Equation C.6). The red dotted line plots the approximated value using Equation C.8. Note the logarithmic axes. The vertical dashed lines in both subplots indicate the size of the smallest grammars in our studies.

I found that from dimensionalities of about $n = 30$ or larger, the distribution of the cosines could not be statistically distinguished from a normal distribution.

Figure C1 plots the results of the simulated mean cosines (panel a) and standard deviations (panel b). Notice that, in both cases, the measures settle very quickly on their predicted asymptotic values, denoted by the red dotted lines (i.e., Equation C.5 and Equation C.8, respectively). Together with the absent significant deviation from normality, this ensures that the normal approximation in Equation C.9 is indeed adequate for the dimensionality sizes in realistic grammars (indicated by the vertical dotted lines in both panels).

Appendix D. Distribution of the Ratio of Cosines Between Non-negative Vectors

Consider three randomly chosen vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n_+$. The previous appendix showed how, for moderate or large values of n, the cosines of these vectors will be approximately normally distributed, with a constant mean $2/\pi$, and a standard deviation of about $4/(3\pi\sqrt{n})$,

$$\cos(\mathbf{x}, \mathbf{y}), \cos(\mathbf{x}, \mathbf{z}) \xrightarrow{D} \mathrm{N}\left(\frac{2}{\pi}, \frac{4}{3\pi\sqrt{n}}\right) \tag{D.1}$$

The ratio between two normally distributed variables follows Fieller's Normal Ratio Distribution (Fieller 1932). This is a very complex four-parameter distribution, whose

moments are given (when they exist) by Craig (1928). Fortunately, however, the distribution is very well approximated by a normal distribution in the special case where the coefficient of variation of the denominator is smaller than .4 (Hinkley 1969) or—more strictly—.22 (Moscoso del Prado Martín 2008). Specifically, if $A \sim N(\mu_1, \sigma_1)$ and $B \sim N(\mu_2, \sigma_2)$, irrespective of the degree of correlation between A and B, if $\sigma_2/|\mu_2| < .22$, it is approximately true that

$$\frac{A}{B} \sim N\left(\frac{\mu_1}{\mu_2}, \frac{\sigma_1}{|\mu_2|}\right) \tag{D.2}$$

It is easy to see that the coefficient of variation in Equation D.1 is smaller than .22 for any value of n. Therefore, combining Equations D.1 and D.2, for moderate and large values of n we can approximate

$$\frac{\cos(\mathbf{x}, \mathbf{y})}{\cos(\mathbf{x}, \mathbf{z})} \sim N\left(1, \frac{2}{3\sqrt{n}}\right) \tag{D.3}$$

To investigate to what degree, and for what range of dimensionality, the approximation above is valid I performed a simulation study. For each of 21 increasing dimensionalities (n= [2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 80, 160, 320, 640, 1,280, 2,560, 5,120, 10,240, 20,480]), I randomly sampled 1,000 triplets of vectors $\mathbf{x_i}, \mathbf{y_i}, \mathbf{z_i} \in \mathbb{R}^n_+$, distributed

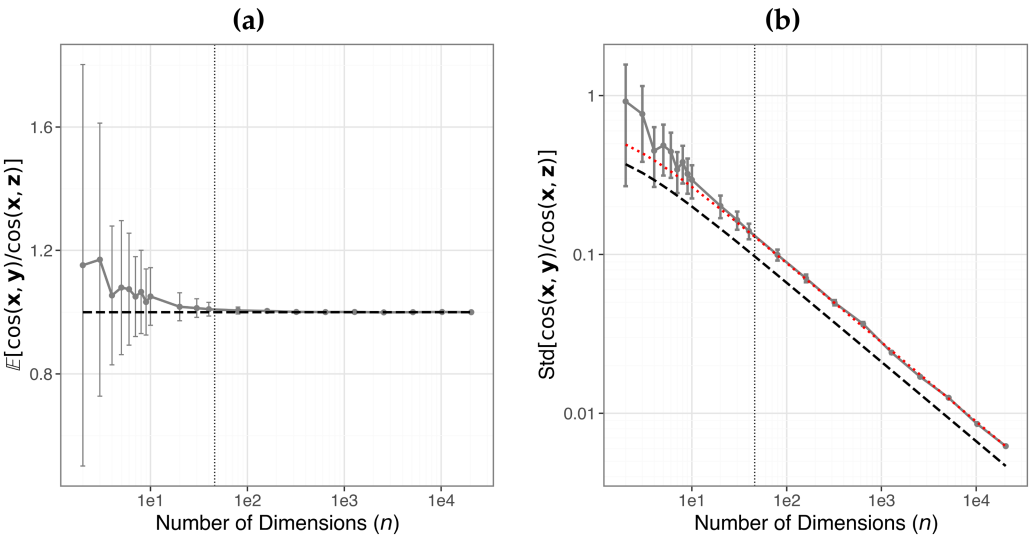

Figure D1
Results of simulation in Appendix C. (a) Accuracy of the mean ratio between cosines. The grey dots and curve plot the simulation result (error bars are standard errors). The dashed black line is the theoretically predicted mean of one. The vertical dashed line indicates the size of the smallest grammars in our studies. Note the logarithmic horizontal axis. (b) Accuracy of the standard deviation of the ratio. The grey dots and curve plot the simulation result (error bars are standard errors). The dashed black line is the theoretically approximated standard deviation (Equation D.3). The red dotted line plots the effect of introducing the ad hoc correction factor of $3\sqrt{\pi}/4$ (Equation D.5). Note the logarithmic axes. The vertical dashed lines in both subplots indicate the size of the smallest grammars in our studies.

according to a multi-dimensional normal $N(\mathbf{0}, \mathbf{I_n})$ truncated to the non-negative orthant, which were then projected onto the surface of the unit radius, **0**-centered n-sphere. These vectors can be taken as a uniform sample of the possible angles in the non-negative orthant (Marsaglia 1972). For each triplet, I computed the cosine ratio

$$k_i = \frac{\cos(\mathbf{x_i}, \mathbf{y_i})}{\cos(\mathbf{x_i}, \mathbf{z_i})} = \frac{\mathbf{x_i} \cdot \mathbf{y_i}}{\mathbf{x_i} \cdot \mathbf{z_i}} \tag{D.4}$$

The results of the simulation are plotted in Figure D1. Notice that, as predicted theoretically, the mean ratio of the cosines (panel a) is almost exactly one even for moderate values of n (and always within the standard error). Similarly, the standard deviation of the cosine ratios is quite accurately modeled by the theoretical prediction (dashed line). However, there is a very slight underestimation of the standard deviation. Crucially, it forming a parallel line on the logarithmic axes reveals that this error is a constant factor. I have found that the value of this constant is approximately $3\sqrt{\pi}/4$. If we multiply our theoretical prediction by this ad hoc factor, we obtain a basically exact prediction of the standard deviation (dotted line in the plot). Taking this correcting factor into account, Equation D.3 becomes

$$\frac{\cos(\mathbf{x}, \mathbf{y})}{\cos(\mathbf{x}, \mathbf{z})} \xrightarrow{D} N\left(1, \frac{1}{2}\sqrt{\frac{\pi}{n}}\right) \tag{D.5}$$

which is an extremely accurate description of the distribution of the ratio of the cosines. In sum, the ratio of the cosines between two non-negative vectors is expected to be normally distributed and very strongly concentrated around one.

Acknowledgments

I am indebted to Prof. John W. DuBois, Dr. Enrique Amigó Cabrera, Suchir Salham, and Paul Siewert for proofreading earlier versions of this article, comments, and discussion of the ideas, and to Dr. Hinrik Hafsteinsson for resolving doubts on the differences in sentence segmentation between the original and UD versions of IcePaHC. In addition, I am indebted to Prof. Liang Huang and three anonymous reviewers for extensive suggestions that have greatly improved this article.

References

Abney, Steven. 1995. Dependency grammars and context-free grammars. Presented at Annual Meeting of the Linguistic Society of America, January 1995.

Agmon, Galit, Sameer Pradhan, Sharon Ash, Naomi Nevler, Mark Liberman, Murray Grossman, and Sunghye Cho. 2024. Automated measures of syntactic complexity in natural speech production: Older and younger adults as a case study. *Journal of Speech, Language, and Hearing Research*, 67(2):545–561. https://doi.org/10.1044/2023_JSLHR-23-00009, PubMed: 38215342

Allen, Shanley E. M. and Catherine Dench. 2015. Calculating mean length of utterance for eastern Canadian Inuktitut. *First Language*, 35:377–406. https://doi.org/10.1177/0142723715596648

Arnardóttir, Þórunn, Hinrik Hafsteinsson, Einar Freyr Sigurðsson, Kristín Bjarnadóttir, Anton Karl Ingason, Hildur Jónsdóttir, and Steinþór Steingrímsson. 2020. A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 16–25.

Attaran, Mohsen and Martin Zwick. 1987. Entropy and other measures of industrial diversification. *Quarterly Journal of Bussiness and Economics*, 26:17–34.

Baayen, R. Harald and Fermín Moscoso del Prado Martín. 2005. Semantic density and past-tense formation in three Germanic languages. *Language*, 81:666–698. https://doi.org/10.1353/lan.2005.0112

Benjamini, Yoav and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to

multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57.1:289–200. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly.

Booth, Taylor L. 1969. Probabilistic representation of formal languages. In *10th Annual Symposium on Switching and Automata Theory (SWAT 1969)*, pages 74–81. https://doi.org/10.1109/SWAT.1969.17

Booth, Taylor L. and Richard A. Thompson. 1973. Applying probability measures to abstract languages. *IEEE Transactions on Computers*, C-22:442–450. https://doi.org/10.1109/T-C.1973.223746

Breiman, Leo. 1957/1960. The individual ergodic theorem of information theory (& correction). *Annals of Mathematical Statistics*, 28/31:809–811/809–810. https://doi.org/10.1214/aoms/1177706899

Brown, Roger. 1973. *A First Language: The Early Stages*. Harvard University Press. https://doi.org/10.4159/harvard.9780674732469

Casby, Michael W. 2011. An examination of the relationship of sample size and mean length of utterance for children with developmental language impairment. *Child Language Teaching and Therapy*, 27:286–293. https://doi.org/10.1177/0265659010394387

Chao, Anne and Tsung-Jen Shen. 2003. Non-parametric estimation of Shannon's index of diversity when there are unseen species in the sample. *Environmental and Ecological Statistics*, 10:429–443. https://doi.org/10.1023/A:1026096204727

Chao, Anne, Y. T. Wang, and Lou Jost. 2013. Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution*, 4:1091–1100. https://doi.org/10.1111/2041-210X.12108

Cheung, Hintat and Susan Kemper. 1992. Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, 13:53–76. https://doi.org/10.1017/S0142716400005427

Chi, Zhiyi. 1999. Statistical properties of probabilistic context-free grammars. *Computational Linguistics*, 25:131–160.

Chi, Zhiyi and Stuart Geman. 1998. Estimation of probabilistic context-free grammars. *Computational Linguistics*, 24:299–305.

Chomsky, Noam. 1956. Three models for the description of language. *IEEE Transactions on Information Theory*, 2:113–124. https://doi.org/10.1109/TIT.1956.1056813

Chung, Kai Lai. 1961. A note on the ergodic theorem of information theory. *Annals of Mathematical Statistics*, 32:612–614. https://doi.org/10.1214/aoms/1177705069

Corazza, Anna, Alberto Lavelli, and Giorgio Satta. 2013. An information-theoretic measure to evaluate parsing difficulty across treebanks. *ACM Transactions on Speech and Language Processing*, 9:7:1–7:31. https://doi.org/10.1145/2407736.2407737

Craig, Cecil C. 1928. The frequency of function of y/x. *The Annals of Mathematics*, 30:471–86. https://doi.org/10.2307/1968296

Crystal, David. 1974. A review of Brown's *A First Language*. *Journal of Child Language*, 1:289–307. https://doi.org/10.1017/S030500090000074X

D'Agostino, Ralph and Egon S. Pearson. 1973. Tests for departure from normality. Empirical results for the distributions of b^2 and $\sqrt{b^1}$. *Biometrica*, 60:613–22. https://doi.org/10.1093/biomet/60.3.613

de Marneffe, Marie Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308. https://doi.org/10.1162/coli_a_00402

Du Bois, John W. 2014. Towards a dialogic syntax. *Cognitive Linguistics*, 25:359–410. https://doi.org/10.1515/cog-2014-0024

Eisner, Jason M. 1996. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, pages 340–345. https://doi.org/10.3115/992628.992688

Evans, Nicholas and Steven C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32:429–492. https://doi.org/10.1017/S0140525X0999094X, PubMed: 19857320

Fieller, Edgar Charles. 1932. The distribution of the index in a normal bivariate population. *Biometrika*, 24:428–440.

https://doi.org/10.1093/biomet/24.3-4.428
Frazier, Lyn. 1985. Syntactic complexity. In D. Dowty, L. Karttunen, and A. Zwicky, editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*. Cambridge University Press. https://doi.org/10.1017/CBO9780511597855.005
Futtrell, Richard, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100.
Gaifman, Haim. 1965. Dependency systems versus phrase-structure systems. *Information and Control*, 8:304–337. https://doi.org/10.1016/S0019-9958(65)90232-9
Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76. https://doi.org/10.1016/S0010-0277(98)00034-1, PubMed: 9775516
Godfrey, John J., Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. https://doi.org/10.1109/ICASSP.1992.225858
Good, Irving John. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264. https://doi.org/10.1093/biomet/40.3-4.237
Gotelli, Nicholas J. and Anne Chao. 2013. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In Simon A. Levin, editor, *Encyclopedia of Biodiversity (second edition)*, volume 5. Academic Press, pages 195–211. https://doi.org/10.1016/B978-0-12-384719-5.00424-X
Grenander, Ulf. 1967. Syntax-controlled probabilities. Technical report, Division of Applied Mathematics, Brown University, Providence, RI.
Grenander, Ulf. 1976. *Lectures in Pattern Theory*, volume 1, Pattern Synthesis. Springer-Verlag. https://doi.org/10.1007/978-1-4612-6369-2_1
Hale, John. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32:101–123. https://doi.org/10.1023/A:1022492123056, PubMed: 12690827
Hale, John. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30:609–672. https://doi.org/10.1207/s15516709cog0000_64
Hausser, Jean and Korbinian Strimmer. 2009. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10:1469–1484.
Hinkley, David V. 1969. On the ratio of two correlated normal random variables. *Biometrika*, 56:635–639. https://doi.org/10.1093/biomet/56.3.635
Hutchins, Sandra E. 1972. Moments of string and derivation lengths of stochastic context-free grammars. *Information Sciences*, 4:179–191. https://doi.org/10.1016/0020-0255(72)90011-4
Jiang, Jingyang, Peng Bi, and Haitao Liu. 2019. Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing*, 46:100666. https://doi.org/10.1016/j.jslw.2019.100666
Jing, Yingqi, Paul Widmer, and Balthasar Bickel. 2021. Word order variation is partially constrained by syntactic complexity. *Cognitive Science*, 45:e13056. https://doi.org/10.1111/cogs.13056, PubMed: 34758151
Klee, T. and M. D. Fitzgerald. 1985. The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language*, 12:251–269. https://doi.org/10.1017/S0305000900006437, PubMed: 4019603
Kuich, Werner. 1970. On the entropy of context-free languages. *Information and Control*, 15:173–200. https://doi.org/10.1016/S0019-9958(70)90105-1
Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23:533–572. https://doi.org/10.1515/lingty-2019-0025
Lin, Dekang. 1996. On the structural complexity of natural language sentences. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, pages 729–733. https://doi.org/10.3115/993268.993295
Linzen, Tal and Florian Jaeger. 2014. Investigating the role of entropy in sentence processing. In *Proceedings of the Fifth Workshop on Cognitive Modelling and Computational Linguistics*, pages 10–18. https://doi.org/10.3115/v1/W14-2002

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9:151–191. https://doi.org/10.17791/jcs.2008.9.2.159

Marcus, Mitchell, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1995. *Treebank-3 LDC99T42*. Linguistic Data Consortium, Philadelphia, PA.

Marsaglia, George. 1972. Choosing a point from the surface of a sphere. *Annals of Mathematical Statistics*, 43:645–646. https://doi.org/10.1214/aoms/1177692644

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530. https://doi.org/10.3115/1220575.1220641

Miller, George A. 1955. Note on the bias of information estimates. In Henry Quastler, editor, *Information Theory in Psychology*. Free Press, pages 95–100.

Miller, Michael I. and Joseph A. O'Sullivan. 1992. Entropies and combinatorics of random branching processes and context-free languages. *IEEE Transactions on Information Theory*, 38:1292–1310. https://doi.org/10.1109/18.144710

Moscoso del Prado Martín, Fermín. 2008. A theory of reaction time distributions. *CogPrints preprint CogPrints:6310*.

Moscoso del Prado Martín, Fermín. 2014. Grammatical change begins within the word: Causal modeling of the co-evolution of Icelandic morphology and syntax. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 2657–2662.

Moscoso del Prado Martín, Fermín. 2017. Vocabulary, grammar, sex, and aging. *Cognitive Science*, 41:950–975. https://doi.org/10.1111/cogs.12367, PubMed: 28523653

Moscoso del Prado Martín, Fermín. 2025. On the distribution of cosine similarities between non-negative vectors.

Moscoso del Prado Martín, Fermín, Aleksandar Kostić, and R. Harald Baayen. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, 94:1–18. https://doi.org/10.1016/j.cognition.2003.10.015, PubMed: 15302325

Nemenman, Ilya, William Bialek, and Rob de Ruyter van Steveninck. 2004. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69:056111. https://doi.org/10.1103/PhysRevE.69.056111, PubMed: 15244887

Nemenman, Ilya, Fariel Shafee, and William Bialek. 2002, Entropy and inference, revisited. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. MIT Press, pages 471–478. https://doi.org/10.7551/mitpress/1120.003.0065

Nice, Margaret Morse. 1925. Length of sentences as a criterion of a child's progress in speech. *Journal of Educational Psychology*, 16:370–379. https://doi.org/10.1037/h0073259

Norris, John and Lourdes Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30:555–578. https://doi.org/10.1093/applin/amp044

Ostrowski, Alexander. 1937. Über die Determinanten mit überwiegender Hauptdiagonale. *Commentarii Mathematici Helvetici*, 10:69–96. https://doi.org/10.1007/BF01214284

Pakhomov, Serguei, Dustin A. Chacón, Mark Wicklund, and Jeanette Gunde. 2011. Computerized assessment of syntactic complexity in Alzheimer's disease: A case study of Iris Murdoch's writing. *Behavioral Research Methods*, 43:136–144. https://doi.org/10.3758/s13428-010-0037-9, PubMed: 21287110

Paninski, Liam. 2003. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253. https://doi.org/10.1162/089976603321780272

Parker, Matthew D. and Kent Brorson. 2005. A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language*, 25:365–376. https://doi.org/10.1177/0142723705059114

Poon, Leo L. M., Timothy Song, Roni Rosenfeld, Xudong Lin, Matthew B. Rogers, Bin Zhou, Robert Sebra, Rebecca A. Halpin, Yi Guan, Alan Twaddle, et al. 2016. Quantifying influenza virus diversity and transmission in humans. *Nature Genetics*, 48:195–200. https://doi.org/10.1038/ng.3479, PubMed: 26727660

Pullum, Geoffrey K. and Barbara C. Scholz. 2010. Recursion and the infinitude claim. In Harry van der Hulst, editor, *Recursion in Human Language (Studies in Generative Grammar)*. Mouton de Gruyter, pages 113–119. https://doi.org/10.1515/9783110219258.111

Rezaii, Neguine, Kyle Mahowald, Rachel Ryskin, Bradford Dickerson, and Edward Gibson. 2022. A syntax–lexicon trade-off in language production. *Proceedings of the National Academy of Sciences USA*, 119:e2120203119. https://doi.org/10.1073/pnas.2120203119, PubMed: 35709321

Roark, Brian, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting Mild Cognitive Impairment. In *Biological, Translational, and Clinical Language Processing*, pages 1–8. https://doi.org/10.3115/1572392.1572394

Rögnvaldsson, Eiríkur, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic parsed historical corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984.

Sánchez, Joan Andreu and José-Miguel Benedí. 1997. Consistency of stochastic context-free grammars from probabilistic estimation based on growth transformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:1052–1055. https://doi.org/10.1109/34.615455

Scarborough, Hollis S. 1990. Index of Productive Syntax. *Applied Psycholinguistics*, 11:1–22. https://doi.org/10.1017/S0142716400008262

Scarborough, Hollis S., Leslie Rescorla, Helen Tager-Flusberg, Anne E. Fowler, and Vicki Sudhalter. 1991. The relation of utterance length to grammatical complexity in normal and language-disordered groups. *Applied Psycholinguistics*, 12:23–45. https://doi.org/10.1017/S014271640000936X

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

Soule, Stephen. 1974. Entropies of probabilistic grammars. *Information and Control*, 25:57–74. https://doi.org/10.1016/S0019-9958(74)90799-2

Spokoyny, Daniel, Jeremy Irvin, and Fermín Moscoso del Prado Martín. 2016. Explicit causal connections between the acquisition of linguistic tiers: Evidence from dynamical systems modeling. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 73–81. https://doi.org/10.18653/v1/W16-1910

Sy, Yaya, William Havard, Marvin Lavechin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. Measuring language development from child-centered recordings. In *Proceedings of Interspeech 2023*, pages 4618–4622. https://doi.org/10.21437/Interspeech.2023-1569

Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale*. Klincksiek.

Theakston, Anna L., Elena V. M. Lieven, Julian M. Pine, and Caroline F. Rowland. 2004. Semantic generality, input frequency and the acquisition of syntax. *Journal of Child Language*, 31:61–99. https://doi.org/10.1017/S0305000903005956, PubMed: 15053085

Vu, Vincent Q., Bin Yu, and Robert E. Kass. 2007. Coverage-adjusted entropy estimation. *Statistics in Medicine*, 26:4039–4060. https://doi.org/10.21236/ADA472999

Wetherell, Charles Stanley. 1980. Probabilistic languages: A review and some open questions. *ACM Computing Surveys (CSUR)*, 12:361–379. https://doi.org/10.1145/356827.356829

Yngve, Victor H. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466.