On the Flatness, Non-linearity, and Branching Direction of Natural Language and Random Constituency Trees: Analyzing Structural Variation within and across Languages

Taiga Ishii and Yusuke Miyao
The University of Tokyo
{taigarana,yusuke}@is.s.u-tokyo.ac.jp

Abstract

Natural languages exhibit remarkable diversity in their syntactic structures. Previous research has investigated the cross-lingual differences in local structural features such as word order or dependency relations. However, considering structural variation within individual language, it remains unclear how such features influence the variation in the overall constituency tree structure and hence the structural variation across languages. To this end, we focus on the shape of constituency trees, analyzing the cross-lingual overlap in the distributions of flatness, non-linearity, and branching direction. While acknowledging that the findings may be influenced by the potential annotation idiosyncrasies across treebanks, the experiments quantitatively suggest that flatness and branching direction vary significantly across languages. As for non-linearity, the cross-lingual difference was relatively small, and the distributions tend to skew towards linear structures. Furthermore, comparison with randomly generated trees suggests that while phrase category and frequency information is crucial for reproducing the branching direction found in natural languages, non-linearity can be replicated reasonably well even without such information.

1 Introduction

Uncovering the universals and variations in syntactic structures across natural languages is a central challenge in computational linguistics and natural language processing. In the context of linguistic typology, differences and universal properties among languages have been extensively discussed from perspectives such as word order (Dryer, 1992; Östling, 2015; Baylor et al., 2024; Alves et al., 2023), dependency relations (Blache et al., 2016; Chen and Gerdes, 2017), morphology (Cotterell et al., 2019; Bentz et al., 2016; Bjerva and Augenstein, 2018), and phonology (Cotterell and Eisner, 2017; Bjerva and Augenstein, 2018).

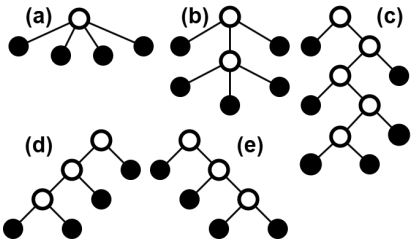

Figure 1: Example constituency trees: (a) fully flat, (b, c) fully non-linear, (d, e) fully left/right-branching

However, capturing structural variation within language remains a challenge for existing typological studies (Ponti et al., 2019). When considering within-language variations, it is not trivial how local structural features, such as word order and dependency relations, influence the variation in the overall structure of constituency trees and, consequently, relate to cross-lingual structural differences and universals.

To address this issue, we propose an approach based on the shape of constituency trees. Specifically, we quantify three features characterizing tree shape: flatness, non-linearity, and branching direction (Figure 1).[1] We then analyze the distribution of these shape features within each language and their distributional overlap across languages, using treebanks from diverse languages.

A key advantage of our approach, using tree shape features, is its potential to enable comparisons beyond natural languages. While traditional linguistic typology relies mainly on features derived from linguistic theories and often limits its scope to comparisons among natural languages (Dryer and Haspelmath, 2013), our approach allows us to investigate the statistical properties of natural language constituency trees within the broader space of all theoretically possible tree structures.

[1] Examples of English constituency trees are shown in Appendix A.

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 90–104

Tree Shape	Measure	Intuitive Description	Range
Flatness	AR	average number of internal nodes per leaf	$[0, 1]$
Non-linearity	NCE	normalized depth of max center embedding	$[0, 1]$
Branching Direction	CC, EWC, RJ	left-right diff of number of leaves at each node	$[-1, 1]$

Table 1: Overview of the tree shape measures used in this study.

Furthermore, in addition to comparing tree shape distributions across natural languages, we conduct comparisons with randomly generated tree structures. This comparative analysis aims to shed light on the fundamental question of what information (e.g., grammatical category information) is essential for characterizing the structure of natural language trees.

Our analyses are conducted on constituency treebanks from 11 diverse languages. Experimental results quantitatively show that while flatness and branching direction exhibit significant cross-lingual variation with minimal distributional overlap, the distributions of non-linearity are skewed towards the linear region, resulting in a certain degree of overlap across languages. Moreover, comparisons with random trees suggest that category information is crucial for reproducing branching direction, whereas non-linearity can be relatively well replicated even without such information. However, a key limitation is that the treebanks used in the experiments are not harmonized, meaning that the findings may be influenced by differences in the annotation schemes. Disentangling the genuine linguistic differences from potential annotation artifacts is a crucial direction for future work. The implementations of the experiments are available at `https://github.com/mynlp/tree-shape-distribution.git`.

2 Background

Cross-lingual analysis based on the shape features of dependency trees has been actively conducted. For example, (directed) dependency distance is used to investigate typological differences (Chen and Gerdes, 2017; Yadav et al., 2020) and universal tendencies like dependency distance minimization (Futrell et al., 2015; Yu et al., 2019). Previous research has also examined structural properties like clause/center embedding depth across languages (Blasi et al., 2019; Noji and Miyao, 2014) and statistical patterns such as Menzerath's law (Mačutek et al., 2017, 2021; Berdičevskis, 2021), sometimes including comparisons with random trees (Tanaka-Ishii, 2021).

In contrast, cross-lingual analysis based on the shape of constituency trees is relatively sparse compared to those on dependency trees. For example, while there are studies on the relationship between the center embedding depth and human reading time in English (van Schijndel et al., 2015), and a comparison of branching direction in English and Chinese (Zhang et al., 2022), these studies are often limited to a small number of languages. Tanaka-Ishii and Tanaka (2023) conducted an extensive analysis on various natural languages and random trees, but their work is limited to the Strahler number (Strahler, 1957), which measures the lower bounds on memory requirements for processing constituency trees.

This study aims to conduct a systematic comparative analysis specifically for constituency trees, using 3 tree shape features —flatness, non-linearity, and branching direction— across a diverse set of languages.

3 Tree Shape Measures

This section defines the tree shape measures used in this study. Following Chan et al. (2010), we analyze the tree shape of delexicalized constituency trees, where leaves are POS tags and internal nodes represent phrases. While some parsing research assumes only binary trees (Liang et al., 2007; Kim et al., 2019), we consider general n-ary tree structures after removing unary nodes and ignore phrase category labels.

We specifically examine three features: flatness, non-linearity, and branching direction (Table 1). As our interest lies in the overall shape rather than the absolute size of trees, the shape measures are normalized to mitigate the effect of tree size, i.e., the number of leaf nodes, corresponding to sentence length.

3.1 Flatness

Flatness can be interpreted as the degree of nesting within a tree. To quantify flatness, we use the "aspect ratio", a measure adapted from the one

proposed by Chan et al. (2010) as a feature for unsupervised parsing.

For a given unlabeled tree t, the aspect ratio $\mathrm{AR}(t)$ roughly corresponds to the number of internal nodes divided by the number of leaves, and is formally calculated as:

$$\mathrm{AR}(t) \equiv \begin{cases} \frac{|V(t)|-1}{|t|_L-2}, & \text{if } |t|_L > 2 \\ 0, & \text{otherwise} \end{cases}$$

where $|t|_L$ denotes the number of leaves in t, and $|V(t)|$ denotes the number of internal nodes.

This measure is designed such that it equals 0 for a fully flat tree (Figure 1 (a)) and 1 for a fully binary branching tree (Figure 1 (c, d, e)), regardless of the number of leaf nodes $|t|_L$.[2] Note that the original aspect ratio proposed by Chan et al. (2010) was simply $\frac{|V(t)|}{|t|_L}$. Our definition modifies this by subtracting offsets for normalization.

3.2 Non-linearity

Non-linearity is a key concept for discussing whether natural languages are more complex than regular languages (Chomsky, 1956), often captured by center embedding structures in trees. Center embedding also has drawn attention from a cognitive perspective, particularly concerning processing memory load (van Schijndel et al., 2015).

Prior work measured center embedding via the maximum stack depth required by a left-corner parser (van Schijndel et al., 2015; Noji and Miyao, 2014). However, this metric is problematic for purely capturing tree shape, as its inherent left-right asymmetry yields different values for flipped tree structures (Noji, 2016). Furthermore, the Strahler number, employed by Tanaka-Ishii and Tanaka (2023) to measure memory requirement lower bounds, is also not suitable for quantifying non-linearity, because it cannot distinguish between fully center-embedding binary trees (Figure 1 (c)) and fully left/right-branching ones (Figure 1 (d, e)).

We introduce a left-right symmetric center embedding depth measure, calculated via Algorithm 1. Roughly, for a node v, its center embedding depth $\mathrm{CenterEmb}_t(v)$ counts ancestors that are neither the left-most nor right-most child of their respective parent. The overall center embedding depth

Algorithm 1 Function for calculating the center embedding depth of a given node v in tree t.

function $\mathrm{CenterEmb}_t(v)$
  $c \leftarrow 0, \quad nl \leftarrow \mathrm{False}, \quad nr \leftarrow \mathrm{False}$
  while v is not root of t **do**
    if v is not the left-most child in t **then**
      $nl \leftarrow \mathrm{True}$
    if v is not the right-most child in t **then**
      $nr \leftarrow \mathrm{True}$
    if $nl \wedge nr$ **then** ▷ Current node is center embedded
      $c \leftarrow c + 1$
      $nl \leftarrow \mathrm{False}, \quad nr \leftarrow \mathrm{False}$
    $v \leftarrow$ parent of v
  return c

$\mathrm{CE}(t)$ of a tree t is the maximum $\mathrm{CenterEmb}_t$ value among the parents of leaf nodes:[3]

$$\mathrm{CE}(t) \equiv \max_{v\text{: leaf of } t} \mathrm{CenterEmb}_t(\text{parent of } v)$$

To normalize for tree size and capture tree shape purely, we define $\mathrm{NCE}(t)$ as $\mathrm{CE}(t)$ divided by the maximum possible CE value for a tree with $|t|_L$ leaves:

$$\mathrm{NCE}(t) \equiv \begin{cases} \frac{\mathrm{CE}(t)}{\lceil \frac{|t|_L-3}{2} \rceil}, & \text{if } |t|_L > 3 \\ 0, & \text{otherwise} \end{cases}$$

The denominator $\lceil \frac{|t|_L-3}{2} \rceil$ represents the maximum value achieved by fully center-embedding trees (Figure 1 (b, c)). Thus, $\mathrm{NCE}(t)$ approaches 0 for both linear (Figure 1 (d, e)) and flat (Figure 1 (a)) structures, while it approaches 1 for fully center-embedding ones (Figure 1 (b, c)).

3.3 Branching Direction

While local structural features such as word order and dependency direction differ across languages (Dryer and Haspelmath, 2013; Chen and Gerdes, 2017), it is not obvious how these local orderings affect the overall shape, particularly the directional bias, of constituency trees. To this end, we employ three branching direction measures proposed by Ishii and Miyao (2023). These measures are extensions of tree balance indices used in phylogenetics (Heard, 1992; Mooers and Heard, 1997;

[2] Since unary nodes are removed, the range of AR for the trees we analyze is $[0, 1]$. If there were unary nodes, the value could be larger than 1.

[3] Using the parent of the leaf node rather than the leaf node itself is intended to reflect phrase-level embedding, as opposed to word-level embedding. Also, as this is a purely tree shape measure, it does not consider grammatical or semantic constraints, unlike (Wilcox et al., 2019).

Rogers, 1996), adapted to capture left-right asymmetry. They are primarily calculated based on the difference in the number of leaves between the left and right subtrees at each internal node.

To calculate such difference in the number of leaves $h_t(v)$ for a node v in a non-binary tree t, the i-th child of v from the left is weighted by a position-dependent weight $w_v(i)$:

$$h_t(v) \equiv \sum_{i=0}^{|v|_C^t-1} w_v(i) \cdot |t_{v_i}|_L$$

Here, $|v|_C^t$ is the number of child nodes of v, and t_{v_i} denotes the subtree rooted at v_i. The weight $w_v(i)$ is defined as:

$$w_v(i) \equiv \begin{cases} g(i - \frac{|v|_C^t-1}{2}) \cdot \frac{1}{\lfloor |v|_C^t/2 \rfloor}, & \text{if } |v|_C^t > 1 \\ 0, & \text{otherwise} \end{cases}$$

where $g(x) \equiv \text{sign}(x) \cdot \lceil |x| \rceil$ is rounding toward infinity. The weight is symmetric, being close to 0 for central children and -1 or 1 for the outermost children. For example, if $|v|_C^t = 4$, the weights for the children from left to right are $-1, -\frac{1}{2}, \frac{1}{2}, 1$.

The three measures aggregate $h_t(v)$ differently across the tree. The range of all measures is $[-1, 1]$, where values closer to -1 indicate a tree closer to a fully left-branching tree (Figure 1 (d)), and values closer to 1 indicate a tree closer to a fully right-branching tree (Figure 1 (e)).

First, the corrected Colles index (CC) is calculated as:

$$\text{CC}(t) \equiv \frac{2}{(|t|_L - 1)(|t|_L - 2)} \cdot \sum_{v \in V(t)} h_t(v)$$

CC tends to give more weight to the branching bias (h_t) at internal nodes closer to the root. In contrast, the equal weights Colles index (EWC) normalizes the h_t at each internal node by the size of its subtree, aiming to evaluate the contribution of each node to the overall branching direction more evenly. It is calculated as:

$$\text{EWC}(t) \equiv \frac{1}{|t|_L - 2} \cdot \sum_{v \in V(t): |t_v|_L > 2} \frac{h_t(v)}{|t_v|_L - 2}$$

Finally, the Roger's J index (RJ) aggregates branching bias using only the sign of $h_t(v)$, thus evaluating the branching bias at a coarser granularity than EWC:

$$\text{RJ}(t) \equiv \frac{1}{|t|_L - 2} \cdot \sum_{v \in V(t)} \text{sign}(h_t(v))$$

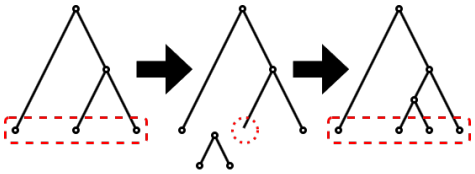

Figure 2: Example of leaf replacement in Yule model.

Algorithm 2 Parameterized Yule model to sample a single non-labeled tree.

Input: w_{len} ▷ Counts of sentence lengths
Input: $w^1_{\text{arity}}, \ldots$ ▷ Counts of arities
Input: $w^1_{\text{pos}}, \ldots$ ▷ Counts of replaced leaf indices
$t \leftarrow \text{SampleCherry}(w^1_{\text{arity}})$ ▷ Initialization
$n_{\text{lim}} \leftarrow \text{Sample}(w_{\text{len}})$ ▷ Sample length limit
while $|t|_L < n_{\text{lim}}$ **do**
  if $\text{sum}(w^{|t|_L}_{\text{pos}}) = 0 \vee \text{sum}(w^{|t|_L}_{\text{arity}}) = 0$ **then**
    Restart from initialization due to lack of statistical infomation
  else
    $i \leftarrow \text{Sample}(w^{|t|_L}_{\text{pos}})$ ▷ Sample leaf index for replacement
    $c \leftarrow \text{SampleCherry}(w^{|t|_L}_{\text{arity}})$ ▷ Sample n-ary cherry
    Replace i-th leaf with cherry c
return t

4 Generating Random Trees

To investigate what statistical information is essential for characterizing the structure of natural language trees, we perform experiments with randomly generated trees. Our methodology is to first extract different levels of statistical information from a given treebank, and then use this information to parameterize random tree models. By comparing the tree shape distributions of the generated trees with those of the original treebank, we can assess the importance of the specific statistical information used by the model.

For this purpose, we employ 6 random tree models based on 2 approaches: the Yule model and Probabilistic Context-Free Grammar (PCFG). Both approaches are hierarchical processes, but they differ primarily in that PCFG utilizes phrase category information, while the Yule model does not.

4.1 Yule Model

The Yule model (Harding, 1971; Yule, 1925; Fischer et al., 2023) is a basic model for generating unlabeled trees by starting from a single leaf node, iteratively replacing a uniformly randomly selected

leaf with a cherry until reaching a target tree size (Figure 2). Typically, a cherry refers to a single internal node tree with two leaves; however, in this study, we consider a general n-ary cherry.

To better capture natural language properties, we parameterize the Yule model using three types of empirical statistics extracted from treebanks: (1) target tree size (stopping criterion) (2) node arity, and (3) leaf replacement position.

Statistics (2) and (3) are estimated as conditional distributions w^k_{arity}, w^k_{pos} dependent on the number of candidate leaves k for the replacement at each step. The process of a Yule model parameterized with these statistics is shown in Algorithm 2. w^k_{arity} and w^k_{pos} are calculated via an inverse Yule process that collapses cherries back into single leaf nodes (Algorithm 3). Since the inverse Yule process for a given tree is generally not unique, we apply the process N times to the treebank.[4]

We compare 4 variants. Yule+arity+pos uses $w^k_{\text{arity}}, w^k_{\text{pos}}$ calculated by Algorithm 3. Yule+arity uses conditional w^k_{arity} with uniform replacement, i.e., $\forall k.w^k_{\text{pos}} = [1, \dots]$. Yule+pos uses conditional w^k_{pos} with corpus-level empirical arity distribution, i.e., $\forall k.w^k_{\text{arity}} = \hat{w}_{\text{arity}}$. Yule uses neither, employing uniform replacement and $\hat{w}_{\text{arity}}$.

4.2 Probabilistic Context-free Grammar

To analyze the role of phrase category information, we also generate random trees using Probabilistic Context-Free Grammars (PCFGs), a standard formalism in parsing (Charniak, 1996; Johnson et al., 2007; Liang et al., 2007). We employ a PCFG with rule probabilities estimated by counts of production rules in the treebank. Since controlling tree size during PCFG generation is non-trivial, we use breadth-first generation, restarting sampling if the number of bottom-most nodes exceeds the maximum tree size, i.e., number of leaves, observed in the original treebank.[5] To isolate the effect of rule frequency information inherent in the PCFG, we additionally evaluate a uniform PCFG (UPCFG) where all production rules for a given nonterminal have uniform probability.

[4]Generation is retried if the empirical distribution for k is unavailable.

[5]Breadth-first generation avoids potential traversal order biases caused by size-based cancellation in depth-first generation.

Algorithm 3 Inverse Yule process to obtain conditional empirical distribution for node arity and leaf replacement positions.

Input: $\mathcal{T}$ ▷ List of trees
Input: N ▷ Number of iteration over given trees
for $n = 1, \dots$ **do** ▷ Initialize the counts of replaced leaf indices and arities when there are n leaves
$w^n_{\text{pos}} \leftarrow [0, \dots], \quad w^n_{\text{arity}} \leftarrow [0, \dots]$
for N times **do**
for $t \in \mathcal{T}$ **do**
$t' \leftarrow t$ ▷ Just copy
while t' is not an n-ary cherry **do**
$l_{\text{cherry}} \leftarrow$ list of root nodes of n-ary cherries in t'
$v \leftarrow \text{UniformSample}(l_{\text{cherry}})$
$a \leftarrow |v|^{t'}_C$
Replace subtree ${t'}_v$ with dummy leaf
$i \leftarrow$ leaf index of replaced dummy leaf
$w^{|t'|_L}_{\text{arity}}[a] \leftarrow w^{|t'|_L}_{\text{arity}}[a] + 1$
$w^{|t'|_L}_{\text{pos}}[i] \leftarrow w^{|t'|_L}_{\text{pos}}[i] + 1$
$w^1_{\text{arity}}[|t'|_L] \leftarrow w^1_{\text{arity}}[|t'|_L] + 1$ ▷ Count the arity of root
return $w^1_{\text{pos}}, \dots, w^1_{\text{arity}}, \dots$

5 Experiments and Discussion

Datasets. In this study, we use treebanks from 11 languages: English (Penn Treebank (Marcus et al., 1993)), Chinese (Chinese Treebank (Palmer et al., 2005)), Japanese (NPCMJ), French, German, Korean, Basque, Hebrew, Hungarian, Polish, and Swedish (SPMRL (Seddah et al., 2013)).[6] It should be noted that these treebanks are not harmonized and thus annotation schemes are not identical. Following Chan et al. (2010), we focus on delexicalized constituency trees. For preprocessing, we apply the following steps to the annotated tree structures in each treebank: (1) remove null elements, (2) strip functional tags from category labels, (3) remove word tokens and treat POS tags as the new leaf nodes, and finally (4) remove unary nonterminals by concatenating their category labels, similar to (Gómez-Rodríguez and Vilares, 2018).[7] Note that we do not remove punctuation.

Furthermore, to analyze the distributions of over-

[6]NPCMJ: NINJAL Parsed Corpus of Modern Japanese (http://NPCMJ.ninjal.ac.jp/).

[7]Further details are described in Appendix B.

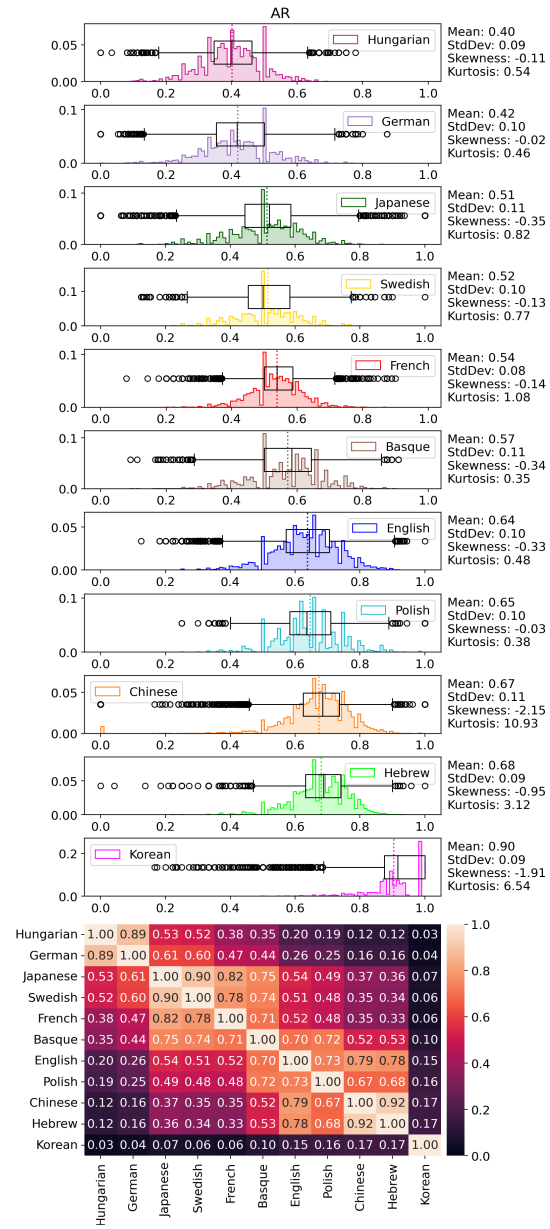

Figure 3: Distributions of AR for 11 natural language treebanks. The values in the heatmap are pairwise HIs.

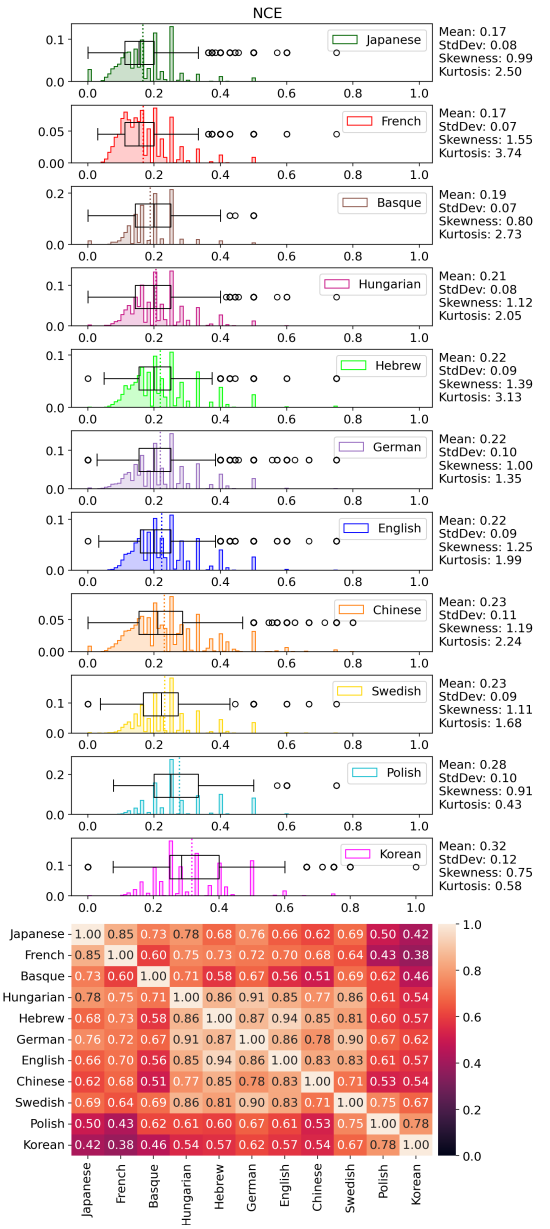

Figure 4: Distributions of NCE for 11 natural language treebanks. The values in the heatmap are pairwise HIs.

all tree shape, we use only sentences with lengths of 10 or more in the experiment. The tree shape distributions for random tree models are also calculated from this subset. This is because shorter sentences have a limited number of possible tree structures, which makes it difficult to analyze cross-lingual differences. Additionally, this length-based filtering provides a simple way to exclude typically short non-sentential fragments, e.g., "(FRAG (PU （) (VV 完) (PU ）))".[8] Table 2 shows the statistics of the preprocessed treebanks, including the number of data points and the mean number of leaf nodes. Note that the values are calculated after applying the length-based filtering. For each model and original dataset, we generate 10000 random trees for analysis.

Evaluation. For all evaluation metrics, we compute distributions as normalized histograms with 100 bins. To quantify cross-lingual differences in tree shape distributions, we use Histogram Intersection (HI). HI measures the proportion of overlap among a set of distributions, yielding a score in $[0, 1]$, where 0 indicates no overlap and 1 indicates identical distributions. This direct measure of overlap makes the score highly intuitive to inter-

[8]The example is from CTB and translates to "(finish)" in English.

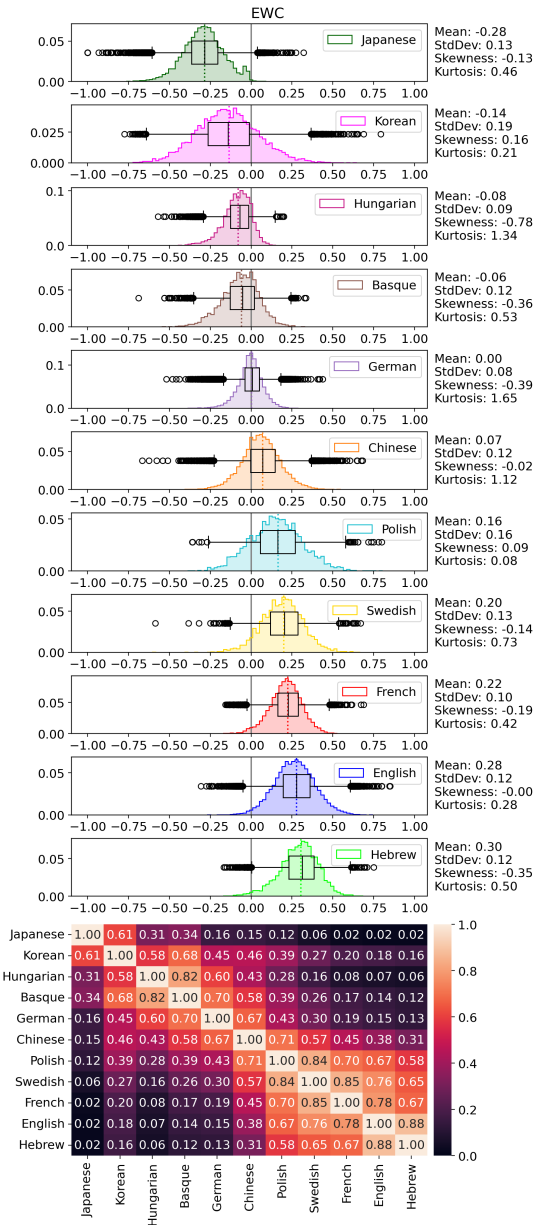

Figure 5: Distributions of EWC for 11 natural language treebanks. The values in the heatmap are pairwise HIs.

pret. Furthermore, unlike pairwise metrics such as Kullback-Leibler divergence, HI has the advantage of being applicable to multiple distributions simultaneously, which allows us to compute a single overall score across all languages.

While HI quantifies the overall overlap, to further understand the characteristics of each distribution, we also analyze its shape using standard deviation, skewness, and kurtosis. Skewness measures how a distribution is biased towards left or right; positive/negative skewness implies large part of the distribution is on the left/right-side. Kurtosis describes the sharpness of a peak of distribution and the weight of its tails compared to a normal distribution, which is defined to have kurtosis of 0; a positive value indicates a more pointed peak and heavier tails.[9]

	#Data	#Leaves
English	35.0K	25.4 ± 10.3
Chinese	14.1K	32.6 ± 18.0
Japanese	47.0K	26.8 ± 18.1
French	13.4K	32.4 ± 16.7
German	30.5K	21.8 ± 9.9
Korean	16.4K	15.3 ± 4.0
Basque	4.9K	15.8 ± 5.1
Hebrew	4.5K	27.6 ± 13.9
Hungarian	6.8K	23.6 ± 11.1
Polish	2.9K	14.4 ± 4.4
Swedish	3.5K	19.0 ± 8.4

Table 2: Statistics of the preprocessed treebanks: the number of data points and the mean ± standard deviation of the number of leaf nodes. The statistics are calculated after applying length-based filtering with a threshold 10.

	AR	NCE	CC	EWC	RJ
HI	0.03	0.33	0.03	0.02	0.02

Table 3: HI across 11 natural language treebanks.

5.1 How Are Natural Language Trees Different?

Table 3 shows the HI across all languages for each tree shape measure. Figure 3, Figure 4, and Figure 5 show the distributions for AR, NCE, EWC in each language, together with box plots within 1.5 IQR and heatmaps of the pairwise HI between languages.[10] Note that languages are sorted by the mean for each measure.

Flatness. As shown in Table 3, the HI is only 3%, indicating that flatness varies considerably across languages. However, languages differ not only in their average flatness but also in the shape of their flatness distributions. While some languages like German and Polish exhibit skewness near 0, others such as Chinese and Korean show values close to −2. Similarly, for kurtosis, Basque has a value around 0.3, whereas Chinese and Korean have much larger values, approximately 6 and 10, respectively.

We speculate two potential factors for the cross-

[9]The presence of outliers can also lead to high kurtosis.
[10]Results for CC and RJ are provided in Appendix C.

	AR	NCE	CC	EWC	RJ
Yule	0.84± 0.01	0.88± 0.02	0.64± 0.06	0.51± 0.08	0.52± 0.08
Yule+arity	0.81± 0.02	0.88± 0.03	0.61± 0.08	0.49± 0.08	0.51± 0.08
Yule+pos	0.84± 0.01	**0.90**± 0.01	0.82± 0.01	0.70± 0.05	0.72± 0.05
Yule+arity+pos	0.81± 0.02	**0.90**± 0.01	0.84± 0.03	0.65± 0.05	0.70± 0.05
UPCFG	0.27± 0.04	0.74± 0.02	0.74± 0.03	0.53± 0.06	0.58± 0.05
PCFG	**0.91**± 0.01	0.88± 0.01	**0.90**± 0.01	**0.92**± 0.01	**0.90**± 0.01

Table 4: Average and standard error of HI between each random model and its original treebank.

lingual differences in flatness. First, differences in annotation schemes across treebanks may play a role. For example, the Japanese and Hungarian treebanks used in this study do not have annotations for VPs (verb phrases) as in PTB due to the relatively free word order (Csendes et al., 2005), and phrases like PP (prepositional phrase) are annotated flatly in German treebank (Brants et al., 2004). Furthermore, we hypothesize that distinctly high AR, i.e., lower flatness, of Korean is due to its tokenization, where multiple word tokens, e.g., compound nouns, are often agglutinated into a single token (Seddah et al., 2013), reducing the number of leaves per nonterminal. This implies that, for any language, the shape of constituency trees calculated based on the number of leaves, can vary depending on the granularity of tokenization, i.e., definition of phrase size.

Non-linearity. From Table 3, we can observe that non-linearity NCE has a 33% overlap across all languages, indicating higher cross-lingual commonality compared to flatness or branching direction. Indeed, the heatmap in Figure 4 shows that pairwise HI values for NCE are generally higher than those for flatness (Figure 3) and branching direction (Figure 5).

Moreover, even Korean, which has the highest average NCE, only reaches 0.32. This suggests that natural language trees are generally quite linear among all possible trees. The skewness ranges from 0.75 to 1.55 across all languages, consistently showing relatively large positive values. This indicates that the distributions are skewed to the left, i.e., towards the more linear region.

Branching Direction. As shown in Table 3, for all branching direction measures CC, EWC, and RJ, the cross-lingual HI is very small, only 2-3%, highlighting significant variation across languages. Furthermore, Figure 5, displaying the distributions and heatmap for EWC, reveals considerable variation in branching direction even within individual languages.

For instance, based on the means, languages such as Japanese, Korean, Hungarian, and Basque tend to be left-branching. However, within each of these languages, right-branching structures are also observed. Similarly, languages such as Hebrew, English, French, Swedish, Polish, and Chinese are right-branching on average, yet they also exhibit left-branching structures internally. It is also interesting to note that the left/right-branching language group based on the mean EWC is mostly the same with that based on the sign of directional dependency distance (Chen and Gerdes, 2017) except Chinese.[11] Conversely, skewness values are close to 0 for all languages, suggesting that the distributions tend to be relatively symmetrical regardless of the language. These results suggest that even when structural variations within individual languages are taken into account, significant structural variation still emerges across languages.

5.2 How Do Random Trees Differ from Natural Language Trees?

Table 4 presents the HI between each random model and the original treebanks, averaged over languages. PCFG performs best overall, achieving nearly 90% overlap on most measures.

Yule models using non-uniform leaf replacement (Yule+pos, Yule+arity+pos) better model non-linearity and branching direction than other Yule variants. However, their lower overlap on EWC, RJ compared to CC suggests that positional information w^k_{pos} for leaf replacement becomes noisy for large k, impacting EWC, RJ that give equal weights to branches near leaves, unlike the root-focused CC. In contrast, PCFG shows consistent strength across CC, EWC, RJ, unlike UPCFG, highlighting the importance of category

[11] However, the order of language itself is not exactly the same as (Chen and Gerdes, 2017).

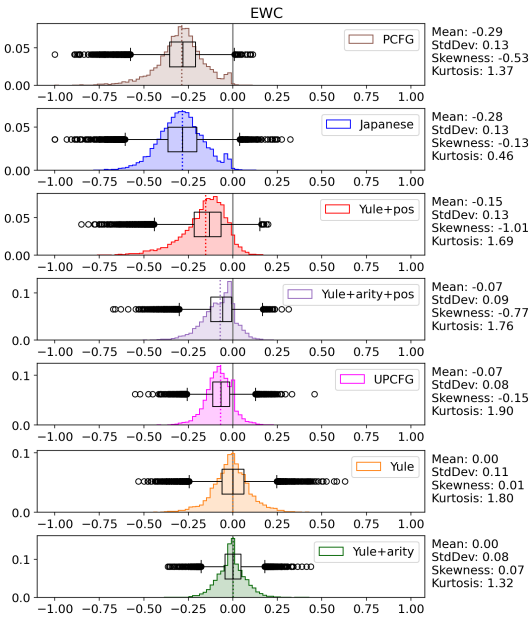

Figure 6: Distributions of EWC for Japanese treebank and its random models.

and frequency information for branching direction. Surprisingly, adding complex arity information (Yule+arity, Yule+arity+pos) degrades performance, suggesting that arity distributions can act as noise when conditioned on the number of leaves for replacement k for models that do not distinguish leaves from nonterminals. The distributions of EWC for Japanese (Figure 6) illustrate these differences; Yule and Yule+arity are mostly centered around 0 while that of the Japanese treebank is around -0.29; Yule+pos seems to capture the left-branching bias to a certain degree, but it is still skewed towards 0.0 compared to PCFG.

Notably, the basic Yule model —which only uses the corpus-level empirical arity distribution and assumes uniform leaf replacement positions— replicates non-linearity NCE well with 88% overlap, suggesting that non-linearity of natural language may be governed by a general mechanism beyond specific grammar or cognitive constraints.

6 Conclusion

We investigated structural variation of constituency trees within and across 11 languages, focusing on flatness, non-linearity, and branching direction. Analysis of the cross-lingual distributional overlap revealed that flatness and branching direction vary significantly across languages, indicating that cross-lingual differences emerge even when considering the structural variation within each language. Meanwhile, the distributions of non-linearity showed smaller cross-lingual difference and tend to skew towards linear trees.

Comparison with 6 random tree models based on the Yule model and PCFG showed that category information, accompanied by frequency statistics, are crucial for reproducing the branching direction patterns in natural language. In contrast, non-linearity was reasonably replicated even by relatively simple Yule models that lack such information, suggesting that non-linearity may be governed by more universal mechanisms independent of fine-grained grammatical details.

While this work focused on the shape of constituency trees, a key future direction is to analyze the joint distribution of overall tree shape features and local structural features, such as word order or dependency relations used in traditional linguistic typology. Such an analysis could lead to a deeper understanding of cross-lingual variations and universality in syntactic structures.

Limitations

As discussed in section 5, since this study analyzes annotated constituency trees, our experimental results can be influenced by the annotation scheme. First, while we included punctuations in the trees, they are sometimes removed in parsing (Li et al., 2020). Given that punctuation annotation methods can also differ across treebanks, investigating the impact of these annotation differences and the presence/absence of punctuation remains a task for future work. Second, as noted in section 5, the category labels annotated in the datasets used for our analysis are not consistent across all languages; for example, VP is not annotated in Japanese and Hungarian. Such difference in annotated phrase categories may also affect the analysis. Third, we did not apply any tokenization to the annotated constituency trees. However, as discussed in section 5, tokenization might affect tree shape.

Acknowledgements

This work was supported by JST SPRING Grant Number JPMJSP2108 and JSPS KAKENHI Grant Number JP24KJ0666 and JP24H00087.

References

Diego Alves, Božo Bekavac, Daniel Zeman, and Marko Tadić. 2023. Analysis of corpus-based word-order typological methods. In *Proceedings of the*

Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023), pages 36–46.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual gradient word-order typology from Universal Dependencies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 42–49.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardzic. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153.

Aleksandrs Berdičevskis. 2021. Successes and failures of Menzerath's law at the syntactic level. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 17–32.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916.

Philippe Blache, Stéphane Rauzy, and Grégoire Montcheuil. 2016. MarsaGram: an excursion in the forests of parsing trees. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2336–2342.

Damian Blasi, Ryan Cotterell, Lawrence Wolf-Sonkin, Sabine Stoll, Balthasar Bickel, and Marco Baroni. 2019. On the distribution of deep clausal embeddings: A large cross-linguistic study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3938–3943, Florence, Italy. Association for Computational Linguistics.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Res. Lang. Comput.*, 2(4):597–620.

Samuel W K Chan, Lawrence Y L Cheung, and Mickey W C Chong. 2010. Tree topological features for unlexicalized parsing. In *Coling 2010: Posters*, pages 117–125, Beijing, China. Coling 2010 Organizing Committee.

Eugene Charniak. 1996. Tree-bank grammars. In *Proceedings of the AAAI Conference on Artificial Intelligence, 13*.

Xinying Chen and Kim Gerdes. 2017. Classifying languages by dependency structure Typologies of delexicalized Universal Dependency treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 54–63.

N Chomsky. 1956. Three models for the description of language. *IEEE Trans. Inform. Theory*, 2(3):113–124.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Trans. Assoc. Comput. Linguist.*, 7:327–342.

Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged treebank. In *Text, Speech and Dialogue*, Lecture notes in computer science, pages 123–131. Springer Berlin Heidelberg, Berlin, Heidelberg.

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

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Mareike Fischer, Lina Herbst, Sophie Kersting, Annemarie Luise Kühn, and Kristina Wicke. 2023. *Tree balance indices: A comprehensive survey*. Springer International Publishing, Cham.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proc. Natl. Acad. Sci. U. S. A.*, 112(33):10336–10341.

Carlos Gómez-Rodríguez and David Vilares. 2018. Constituent parsing as sequence labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1314–1324, Stroudsburg, PA, USA. Association for Computational Linguistics.

E. F. Harding. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, 3(1):44–77.

Stephen B Heard. 1992. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution*, 46(6):1818–1826.

Taiga Ishii and Yusuke Miyao. 2023. Tree-shape uncertainty for analyzing the inherent branching bias of unsupervised parsing models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 532–547, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Johnson, Thomas L Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 139–146.

Yoon Kim, Chris Dyer, and Alexander Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jun Li, Yifan Cao, Jiong Cai, Yong Jiang, and Kewei Tu. 2020. An empirical comparison of unsupervised constituency parsing methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3278–3283, Online. Association for Computational Linguistics.

Percy Liang, Slav Petrov, Michael Jordan, and Dan Klein. 2007. The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 688–697, Prague, Czech Republic. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*.

Ján Mačutek, Radek Čech, and Marine Courtin. 2021. The Menzerath-Altmann law in syntactic structure revisited. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 65–73.

Ján Mačutek, Radek Čech, and Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 100–107.

Arne O Mooers and Stephen B Heard. 1997. Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology*, 72(1):31–54.

Hiroshi Noji. 2016. *Left-corner Methods for Syntactic Modeling with Universal Structural Constraints*. Ph.D. thesis, The Graduate University for Advanced Studies.

Hiroshi Noji and Yusuke Miyao. 2014. Left-corner transitions on dependency parsing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2140–2150, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Martha Palmer, Fu-Dong Chiou, Nianwen Xue, and Tsan-Kuang Lee. 2005. Chinese treebank 5.1 LDC2005T01U01.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Comput. Linguist.*, 45(3):559–601.

James S Rogers. 1996. Central moments and probability distributions of three measures of phylogenetic tree imbalance. *Systematic Biology*, 45(1):99–110.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, and 4 others. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182.

Arthur N Strahler. 1957. Quantitative analysis of watershed geomorphology. *Eos, Transactions American Geophysical Union*, 38(6):913–920.

Kumiko Tanaka-Ishii. 2021. Menzerath's law in the syntax of languages compared with random sentences. *Entropy*, 23(6):661.

Kumiko Tanaka-Ishii and Akira Tanaka. 2023. Strahler number of natural language sentences in comparison with random trees. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(12).

Marten van Schijndel, Brian Murphy, and William Schuler. 2015. Evidence of syntactic working memory usage in MEG data. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 79–88, Denver, Colorado. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 181–190, Florence, Italy. Association for Computational Linguistics.

Marcin Woliński. 2019. *Automatyczna analiza składnikowa języka polskiego*. Wydawnictwa Uniwersytetu Warszawskiego.

Himanshu Yadav, Ashwini Vaidya, Vishakha Shukla, and Samar Husain. 2020. Word order typology interacts with linguistic complexity: A cross-linguistic corpus study. *Cogn. Sci.*, 44(4):e12822.

Xiang Yu, Agnieszka Falenska, and Jonas Kuhn. 2019. Dependency length minimization vs. word order constraints: An empirical study on 55 treebanks. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 89–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

G. U. Yule. 1925. II.—a mathematical theory of evolution, based on the conclusions of dr. J. C. Willis, F. R. S. *Philos. Trans. R. Soc. Lond.*, 213(402-410):21–87.

Xiaohan Zhang, Shaonan Wang, Nan Lin, and Chengqing Zong. 2022. Is the brain mechanism for hierarchical structure building universal across languages? an fMRI study of Chinese and English. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7852–7861, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Stroudsburg, PA, USA. Association for Computational Linguistics.

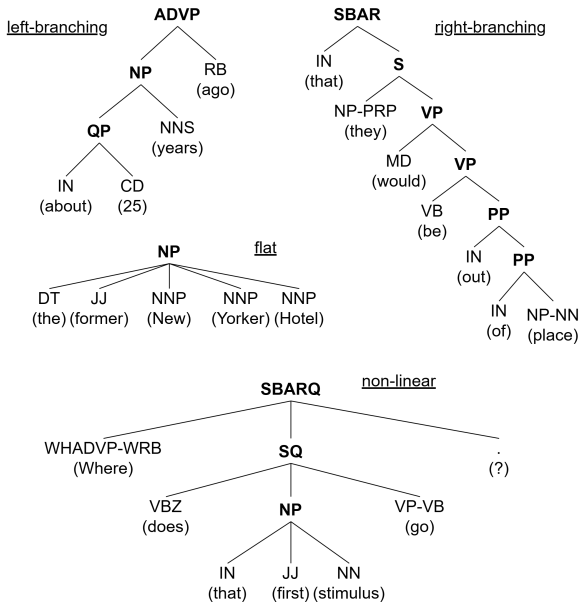

Figure 7: Examples of English constituency trees with fully left/right-branching, flat, and non-linear structures.

A Example Trees in English

Figure 7 shows examples of English constituency trees with fully left/right-branching, flat, and non-linear structures. As it is difficult to find complete sentences that exhibit these specific structures, the example trees in Figure 7 are subtrees extracted from larger constituency trees.

B Setting Details

In this study, we analyze delexicalized constituency trees, treating preterminal nodes, typically POS tags, as leaf nodes. However, the Hebrew and Polish treebanks employ specific annotation conventions that necessitate different preprocessing steps, as detailed below.

First, the Hebrew treebank features two layers of preterminals (Seddah et al., 2013). Therefore, we use the higher preterminal node as the effective leaf node in our analysis.

In the Polish treebank, the lowest-layer nonterminals (i.e., those directly dominating the preterminals) function similarly to preterminals themselves (Woliński, 2019). Unlike the Hebrew data, these lowest-layer nonterminals in Polish are sometimes nested. When these lowest-layer nonterminals are nested, we simply treat the highest ones as leaf nodes.

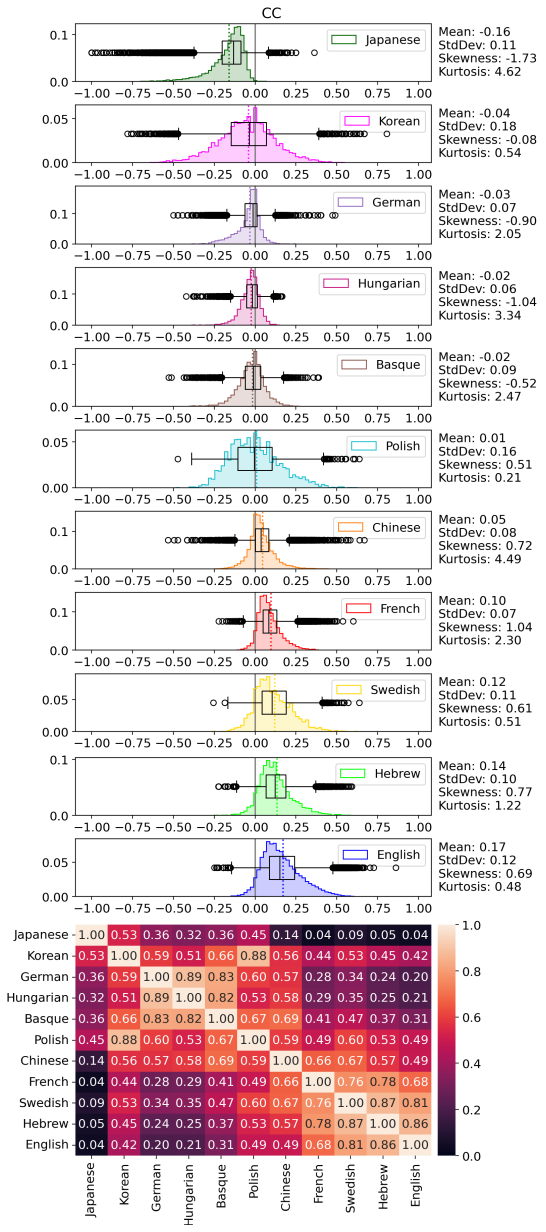

Figure 8: Distributions of CC for 11 natural language treebanks. The values in the heatmap are pairwise HIs.

C Other Results

Figure 8 and Figure 9 show the distributions for CC, RJ in each language, together with box plots within 1.5 IQR and heatmaps of the pairwise HI between languages.

Table 5 shows the HI between each random model and the original treebank for each measures.

AR	English	Chinese	Japanese	French	German	Korean	Basque	Hebrew	Hungarian	Polish	Swedish
Yule	0.93	0.87	0.75	0.78	0.83	**0.88**	0.85	0.80	0.83	0.83	0.84
Yule+arity	**0.94**	0.82	0.70	0.73	0.79	**0.88**	0.83	0.77	0.75	0.82	0.82
Yule+pos	0.93	0.87	0.75	0.78	0.83	**0.88**	0.85	0.80	0.83	0.83	0.84
Yule+arity+pos	**0.94**	0.82	0.70	0.73	0.79	**0.88**	0.83	0.77	0.75	0.82	0.82
UPCFG	0.10	0.21	0.37	0.13	0.55	0.12	0.15	0.20	0.28	0.43	0.46
PCFG	0.91	**0.90**	**0.92**	**0.93**	**0.91**	0.87	**0.89**	**0.89**	**0.95**	**0.89**	**0.92**

NCE	English	Chinese	Japanese	French	German	Korean	Basque	Hebrew	Hungarian	Polish	Swedish
Yule	**0.93**	0.85	0.83	0.87	0.89	0.94	0.64	**0.93**	0.90	0.92	**0.93**
Yule+arity	0.92	0.83	0.89	0.88	0.86	0.94	0.63	**0.93**	0.88	**0.94**	**0.93**
Yule+pos	0.91	**0.92**	0.89	0.83	**0.91**	**0.96**	**0.82**	**0.93**	0.90	0.86	**0.93**
Yule+arity+pos	0.91	0.89	0.90	0.86	0.89	**0.96**	**0.82**	**0.93**	0.89	**0.94**	**0.93**
UPCFG	0.79	0.66	0.81	0.83	0.74	0.60	0.68	0.82	0.83	0.65	0.78
PCFG	0.88	0.86	**0.92**	**0.92**	**0.91**	0.84	**0.82**	0.88	**0.96**	0.85	0.87

CC	English	Chinese	Japanese	French	German	Korean	Basque	Hebrew	Hungarian	Polish	Swedish
Yule	0.37	0.75	0.32	0.44	0.79	0.85	0.81	0.44	0.78	0.87	0.57
Yule+arity	0.30	0.69	0.17	0.33	0.86	0.85	**0.92**	0.33	0.84	**0.92**	0.49
Yule+pos	**0.87**	0.86	0.77	0.82	0.81	0.87	0.74	0.83	0.76	0.78	0.88
Yule+arity+pos	0.83	**0.87**	0.57	0.79	**0.94**	0.86	0.85	0.80	0.90	0.89	0.88
UPCFG	0.57	0.70	0.60	0.76	0.72	0.87	0.87	0.70	0.88	0.72	0.78
PCFG	0.86	0.83	**0.90**	**0.91**	0.91	**0.95**	0.85	**0.93**	**0.96**	0.88	**0.93**

EWC	English	Chinese	Japanese	French	German	Korean	Basque	Hebrew	Hungarian	Polish	Swedish
Yule	0.22	0.73	0.21	0.21	0.93	0.69	0.81	0.19	0.65	0.57	0.37
Yule+arity	0.21	0.71	0.16	0.18	0.91	0.69	0.81	0.17	0.60	0.56	0.35
Yule+pos	0.57	0.88	0.53	0.49	0.87	0.73	0.81	0.47	0.73	**0.92**	0.67
Yule+arity+pos	0.57	0.85	0.32	0.44	0.89	0.73	0.79	0.44	0.63	0.78	0.68
UPCFG	0.26	0.64	0.26	0.30	0.81	0.76	0.65	0.32	0.63	0.73	0.45
PCFG	**0.91**	**0.92**	**0.94**	**0.93**	**0.94**	**0.95**	**0.85**	**0.89**	**0.97**	0.90	**0.93**

RJ	English	Chinese	Japanese	French	German	Korean	Basque	Hebrew	Hungarian	Polish	Swedish
Yule	0.28	0.77	0.17	0.22	0.88	0.72	0.82	0.22	0.69	0.57	0.39
Yule+arity	0.27	0.76	0.13	0.20	0.88	0.71	0.81	0.21	0.68	0.56	0.38
Yule+pos	0.66	**0.93**	0.47	0.50	0.85	0.77	**0.85**	0.54	0.77	**0.90**	0.70
Yule+arity+pos	0.67	0.90	0.28	0.49	0.88	0.77	0.82	0.55	0.72	0.81	0.76
UPCFG	0.38	0.65	0.28	0.40	0.84	0.76	0.70	0.40	0.70	0.71	0.53
PCFG	**0.88**	0.91	**0.93**	**0.90**	**0.95**	**0.93**	**0.85**	**0.86**	**0.95**	0.88	**0.90**

Table 5: HI between each random model and its original treebank.

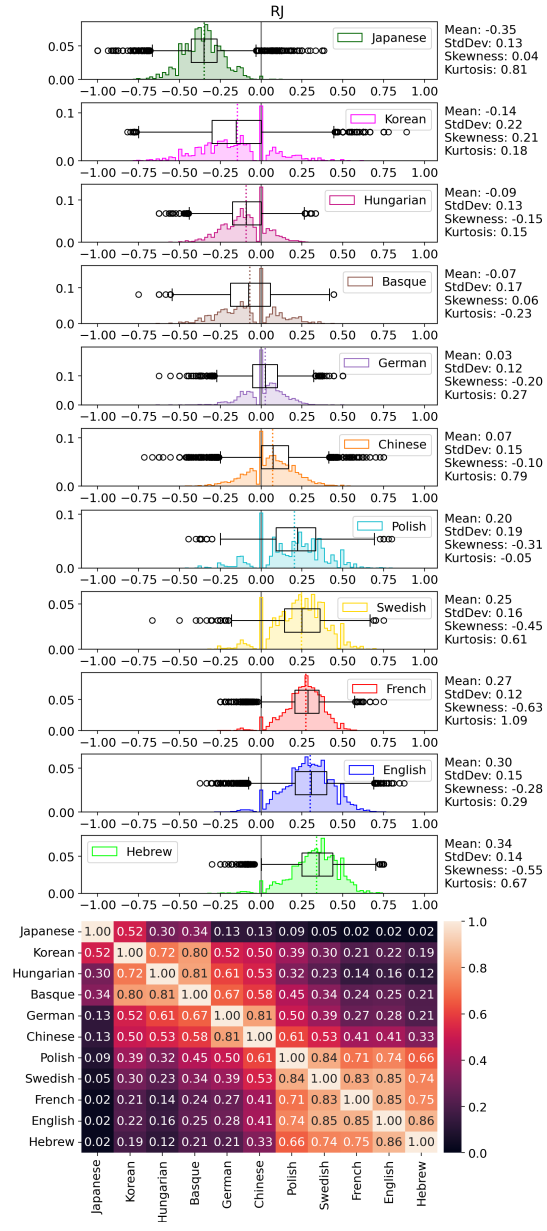

Figure 9: Distributions of RJ for 11 natural language treebanks. The values in the heatmap are pairwise HIs.