A Cross-Linguistic Pressure for Uniform Information Density in Word Order

Thomas Hikaru Clark[1] **Clara Meister**[2] **Tiago Pimentel**[3] **Michael Hahn**[4]
Ryan Cotterell[2] **Richard Futrell**[5] **Roger Levy**[1]
[1]MIT, USA [2]ETH Zürich, Switzerland [3]University of Cambridge, UK
[4]Saarland University, Germany [5]UC Irvine, USA
thclark@mit.edu meistecl@inf.ethz.ch tp472@cam.ac.uk
mhahn@lst.uni-saarland.de ryan.cotterell@inf.ethz.ch
rfutrell@uci.edu rplevy@mit.edu

Abstract

While natural languages differ widely in both canonical word order and word order flexibility, their word orders still follow shared cross-linguistic statistical patterns, often attributed to functional pressures. In the effort to identify these pressures, prior work has compared real and counterfactual word orders. Yet one functional pressure has been overlooked in such investigations: The uniform information density (UID) hypothesis, which holds that information should be spread evenly throughout an utterance. Here, we ask whether a pressure for UID may have influenced word order patterns cross-linguistically. To this end, we use computational models to test whether real orders lead to greater information uniformity than counterfactual orders. In our empirical study of 10 typologically diverse languages, we find that: (i) among SVO languages, real word orders consistently have greater uniformity than reverse word orders, and (ii) only linguistically implausible counterfactual orders consistently exceed the uniformity of real orders. These findings are compatible with a pressure for information uniformity in the development and usage of natural languages.[1]

1 Introduction

Human languages differ widely in many respects, yet there are patterns that appear to hold consistently across languages. Identifying explanations for these patterns is a fundamental goal of linguistic typology. Furthermore, such explanations may shed light on the cognitive pressures underlying and shaping human communication.

This work studies the *uniform information density* (UID) hypothesis as an explanatory principle for word order patterns (Fenk and Fenk, 1980; Genzel and Charniak, 2002; Aylett and Turk, 2004; Jaeger, 2010; Meister et al., 2021). The UID hypothesis posits a communicative pressure to avoid spikes in information within an utterance, thereby keeping the information profile of an utterance relatively close to uniform over time. While the UID hypothesis has been proposed as an explanatory principle for a range of linguistic phenomena, e.g., speakers' choices when faced with lexical and syntactic alternations (Levy and Jaeger, 2006), its relationship to word order patterns has received limited attention, with the notable exception of Maurits et al. (2010).

Our work investigates the relationship between UID and word order patterns, differing from prior work in several ways. We (i) use Transformer language models (LMs) (Vaswani et al., 2017) to estimate information-theoretic operationalizations of information uniformity; (ii) analyze large-scale naturalistic datasets of 10 typologically diverse languages; and (iii) compare a range of theoretically motivated counterfactual grammar variants.

Experimentally, we find that among SVO languages, the real word order has a more uniform information density than nearly all counterfactual word orders; the only orders that consistently exceed real orders in uniformity are generated using an implausibly strong bias for uniformity, at the cost of expressivity. Further, we find that counterfactual word orders that place verbs before objects are more uniform than ones that place objects before verbs in nearly every language.

Our findings suggest that a tendency for uniform information density may exist in human language,

[1] Code for reproducing our experiments is available at https://github.com/thomashikaru/word-order-uid.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1048–1065, 2023. https://doi.org/10.1162/tacl_a_00589
Action Editor: Mark-Jan Nederhof. Submission batch: 9/2022; Revision batch: 1/2023; Published 8/2023.

with two potential sources: (i) word order rules, with SVO order generally being more uniform than SOV; and (ii) choices made by speakers, who use the flexibility present in real languages to structure information more uniformly at a global level (and not only in a small number of isolated constructions).

2 Functional Pressures in Language

2.1 Linguistic Optimizations

A number of linguistic theories link cross-linguistic patterns to functional pressures. For example, both the grammatical rules of a language and speakers' choices (within the space of grammatically acceptable utterances) are posited to reflect a trade-off between effort and robustness: Shorter and simpler structures are easier to produce and comprehend, but longer and more complex utterances can encode more information (Gabelentz, 1901; Zipf, 1935; Hawkins, 1994, 2004, 2014; Haspelmath, 2008). Another such functional pressure follows from the principle of dependency length minimization (DLM), which holds that, in order to minimize working memory load during comprehension, word orders should place words in direct dependency relations close to each other (Rijkhoff, 1986, 1990; Hawkins, 1990, 1994, 2004, 2014; Grodner and Gibson, 2005; Gibson, 1998, 2000; Bartek et al., 2011; Temperley and Gildea, 2018; Futrell et al., 2020). A growing body of work has turned to informa tion theory, the mathematical theory of communication (Shannon, 1948), to formalize principles that explain linguistic phenomena (Jaeger and Tily, 2011; Gibson et al., 2019; Pimentel et al., 2021c). One such principle is that of uniform information density.

2.2 Uniform Information Density

According to the UID hypothesis, speakers tend to spread information evenly throughout an utterance; large fluctuations in the per-unit information content of an utterance can impede communication by increasing the processing load on the listener. Speakers may modulate the information profile of an utterance by selectively producing linguistic units such as optional complementizers in English (Levy and Jaeger, 2006; Jaeger, 2010). A pressure for UID in speaker choices has also been studied in specific constructions in other languages, though with mixed conclusions (Zhan and Levy, 2018; Clark et al., 2022).

Formally, the information conveyed by a linguistic signal $\boldsymbol{y}$, e.g., an utterance or piece of text, is quantified in terms of its surprisal $s(\cdot)$, which is defined as $\boldsymbol{y}$'s negative log-probability: $s(\boldsymbol{y}) \overset{\text{def}}{=} -\log p_\ell(\boldsymbol{y})$. Here, p_ℓ is the underlying probability distribution over sentences $\boldsymbol{y}$ for a language ℓ. Note that we do not have access to the true distribution p_ℓ, and typically rely on a language model with learned parameters $\boldsymbol{\theta}$ to estimate surprisal values with a second distribution $p_{\boldsymbol{\theta}}$.

Surprisal can be additively decomposed over the units that comprise a signal. Explicitly, for a signal $\boldsymbol{y}$ that can be expressed as a series of linguistic units $\langle y_1, \ldots, y_N \rangle$, where $y_n \in \mathcal{V}$ and $\mathcal{V}$ is a set vocabulary of words or morphemes, the surprisal of a unit y_n is its negative log-probability given prior context: $s(y_n) = -\log p_\ell(y_n \mid \boldsymbol{y}_{<n})$. Note that the distribution $p_\ell(\cdot \mid \boldsymbol{y}_{<n})$ has support $\overline{\mathcal{V}} \overset{\text{def}}{=} \mathcal{V} \cup \{\textsc{eos}\}$, where EOS is a designated symbol indicating the end of a sequence;[2] a valid, complete signal $\boldsymbol{y} = \langle y_1, \ldots, y_N \rangle$ has $y_N = \textsc{eos}$. The quantity $s(\boldsymbol{y})$ can thus likewise be expressed as $s(\boldsymbol{y}) = \sum_{n=1}^{N} s(y_n)$. Assuming that we have a fixed amount of information to convey and that high-surprisal items are disproportionately difficult to process,[3] it can be shown mathematically that spreading information evenly throughout a signal optimizes ease of processing for the comprehender (Levy and Jaeger, 2006; Smith and Levy, 2013; Levy, 2018; Meister et al., 2021.

While the UID hypothesis is often discussed in the context of speaker choices, it has also been presented as a general cognitive constraint that might influence reading times (Meister et al., 2021), speech duration (Pimentel et al., 2021b), and word lengths (Piantadosi et al., 2011). Selection for UID has also been discussed as a potential evolutionary pressure on language that can explain typological differences (Jaeger and Tily,

[2] This symbol allows for the global normalization of p_ℓ, i.e., a valid probability distribution over finite-length sequences $\mathcal{V}^*$ (see Du et al., 2022, for a discussion).

[3] Most empirical results (Hale, 2001; Levy, 2008; Shain et al., 2022) suggest that a word's processing effort is directly proportional to its surprisal. Yet there is also evidence of a superlinear relationship, which would imply a preference by the comprehender for UID (Meister et al., 2021; Hoover et al., 2022).

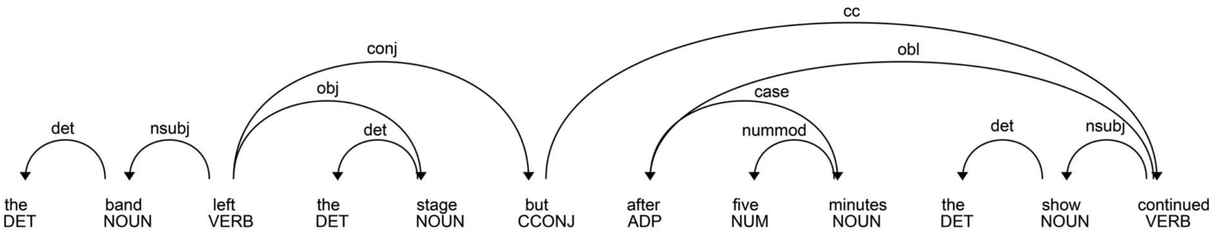

Figure 1: An example dependency tree showing syntactic relationships according UD, transformed so that function words are heads (§3.2). Arrows point from heads to dependents.

2011). Within this literature, there is not a consensus on how to formally operationalize UID. For example, Frank and Jaeger (2008) measure regression of surprisal towards a language-wide mean; Collins (2014) and Bloem (2016) consider more local changes in surprisal in their quantification of UID.

In this work, we consider three metrics for operationalizing UID (Meister et al., 2021):

$$\textsc{uid}_v(\boldsymbol{y}) \overset{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N}(s(y_n)-\mu)^2 \tag{1}$$

In Equation (1), $\textsc{uid}_v$ is the mean within-sentence variance of word surprisals, where $\mu = \frac{1}{N}\sum_{n=1}^{N} s(y_n)$ is a sentence-level mean.

$$\textsc{uid}_{lv}(\boldsymbol{y}) \overset{\text{def}}{=} \frac{1}{N-1}\sum_{n=2}^{N}(s(y_n)-s(y_{n-1}))^2 \tag{2}$$

In Equation (2), $\textsc{uid}_{lv}$ quantifies the average word-to-word change in surprisal, a more localized measure (Collins, 2014). Intuitively, this is maximized when high-surprisal words alternate with low-surprisal words, and minimized when words appear in sorted order by information content.

$$\textsc{uid}_p(\boldsymbol{y}) \overset{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} s(y_n)^k \tag{3}$$

In Equation (3), $\textsc{uid}_p$ is a power mean with $k > 1$, which disproportionately increases in the presence of larger surprisal values.[4] Note that for all of these operationalizations, lower values correspond to greater uniformity.[5]

3 Counterfactual Language Paradigm

Following prior work that has used counterfactual languages to study the functional pressures at play in word order patterns, we investigate to what degree a language's word order shows signs of optimization for UID. In this approach, a corpus of natural language is compared against a counterfactual corpus containing minimally changed versions of the same sentences, where the changes target an attribute of interest, e.g., the language's word order. For example, several studies of DLM have compared syntactic dependency lengths in real and counterfactual corpora, generated by permuting the sentences' word order either randomly (Ferrer-i-Cancho, 2004; Liu, 2008) or deterministically by applying a counterfactual grammar (Gildea and Temperley, 2010; Gildea and Jaeger, 2015; Futrell et al., 2015b, 2020). Similarly, we will compare measures of UID in real and counterfactual corpora to investigate whether real languages' word orders exhibit more uniform information density than alternative realizations.

3.1 Formal Definition

We build on the counterfactual generation procedure introduced by Hahn et al. (2020) to create parallel corpora. This procedure operates on sentences' dependency parses. Formally, a dependency parse 🌲 of a sentence $\boldsymbol{y}$ is a directed tree with one node for every word, where each word in $\boldsymbol{y}$, with the exception of a designated root word, is the child of its (unique) syntactic head; see Zmigrod et al. (2020) for a discussion of the role of the root constraint in dependency tree annotation. Each edge in the tree is annotated with the syntactic relationship between the words connected by that edge; see Figure 1 for an example. Here we use the set of dependency relations defined by the Universal Dependencies (UD) paradigm (de Marneffe et al., 2021), though

[4] This metric suggests a super-linear processing cost for surprisal.

[5] We note that, while a fully uniform language would have value 0 for $\textsc{uid}_v$ and $\textsc{uid}_{lv}$, it would not for $\textsc{uid}_p(\boldsymbol{y})$, so the metrics are not directly comparable.

```
def linearize(t, g):
    return recurse(t.root, g)
def recurse(node, g):
    y = []; dependents = node.dependents
    order = g(t, node)
    sorted(dependents, key=lambda x:
    ↪  order.index(x))
    for dep in dependents: # Add left dependents
        if order.index(dep) < order.index(node):
            y += recurse(dep, g)
    y += [node.word]       # Add node
    for dep in dependents: # Add right dependents
        if order.index(dep) > order.index(node):
            y += recurse(dep, g)
    return y
```

Figure 2: Pseudo-code to linearize a dependency tree 🌲 according to a grammar's ordering function $\mathtt{g}$. In this code, each `node` contains a `word` and its syntactic `dependents`.

we follow Hahn et al. (2020) in transforming dependency trees such that function words are treated as heads, leading to representations closer to those of standard syntactic theories; see also Gerdes et al. (2018).

Tree Linearization. While syntactic relationships are naturally described hierarchically, sentences are produced and processed as linear strings of words. Importantly, there are many ways to linearize a dependency parse 🌲's nodes into a string $\boldsymbol{y}$. Concretely, a grammar under our formalism is defined by an *ordering function* (see Kuhlmann, 2010) $\mathtt{g}(\cdot, \cdot)$ which takes as arguments a dependency parse and a specific node in it, and returns an ordering of the node and its dependents. For each node, its dependents are arranged from left to right according to this ordering; any node without dependents is trivially an ordered set on its own. This process proceeds recursively to arrive at a final ordering of all nodes in a dependency tree, yielding the final string $\boldsymbol{y}$. Pseudocode for the linearization of a tree 🌲 based on an ordering function $\mathtt{g}$ is given in Figure 2.

Simplifying Assumptions. One consequence of this formalism is that all counterfactual orders correspond to projective trees, i.e., trees with no crossing dependencies. While projectivity is a well-attested cross-linguistic tendency, human languages do not obey it absolutely (Ferrer-i-Cancho et al., 2018; Yadav et al., 2021). Within the space of projective word order interventions allowed by this formalism, the grammars which we borrow from Hahn et al. (2020) enforce two additional simplifying constraints. First, the relative positioning (left or right) between the head and dependent of a particular relation is fixed. Second, the relative ordering of different relations on the same side of a head is also fixed. We denote grammars which satisfy both constraints as *consistent*. Notably, natural languages violate both of these assumptions to varying degrees. For example, even in English—a language with relatively strict word order—adverbs can generally appear before or after their head. While these simplifications mean that the formalism cannot perfectly describe natural languages, it provides a computationally well-defined method for intervening on many features of word order. In particular, the consistent grammars of Hahn et al. (2020) are parameterized by a set of scalar weights corresponding to each possible syntactic relation; the ordering function thus reduces to sorting each head's dependents based on their weight values. Notably, Hahn et al. (2020) also introduced a method for optimizing these grammars for various objective functions by performing stochastic gradient descent on a probabilistic relaxation of the grammar formalism; we use several of these grammars (described in §3.2) in our subsequent analysis.

Creating Counterfactual Word Orderings. The above paradigm equips us with the tools necessary for systematically altering sentences' word orderings, which in turn, enables us to create counterfactual corpora. Notably, the large corpora we use in this study contain sentences as strings, not as their dependency parses. We therefore define our counterfactual grammar intervention as the output of a (deterministic) word re-ordering function $\mathtt{f} : \mathcal{Y} \to \mathcal{Y}$, where $\mathcal{Y} \overset{\text{def}}{=} \mathcal{V}^*$ is the set of all possible sentences that can be constructed using a language's vocabulary $\mathcal{V}$.[6] This function takes as input a sentence from our original language and outputs a sentence with the counterfactual word order defined by a given ordering function $\mathtt{g}$. We decompose this function into two steps:

$$\mathtt{f}(\boldsymbol{y}) = \mathtt{linearize}(\mathtt{parse}(\boldsymbol{y}), \mathtt{g}) \tag{4}$$

[6] For notational brevity, we leave the dependency of $\mathcal{V}$ on ℓ implicit as it should be clear from context.

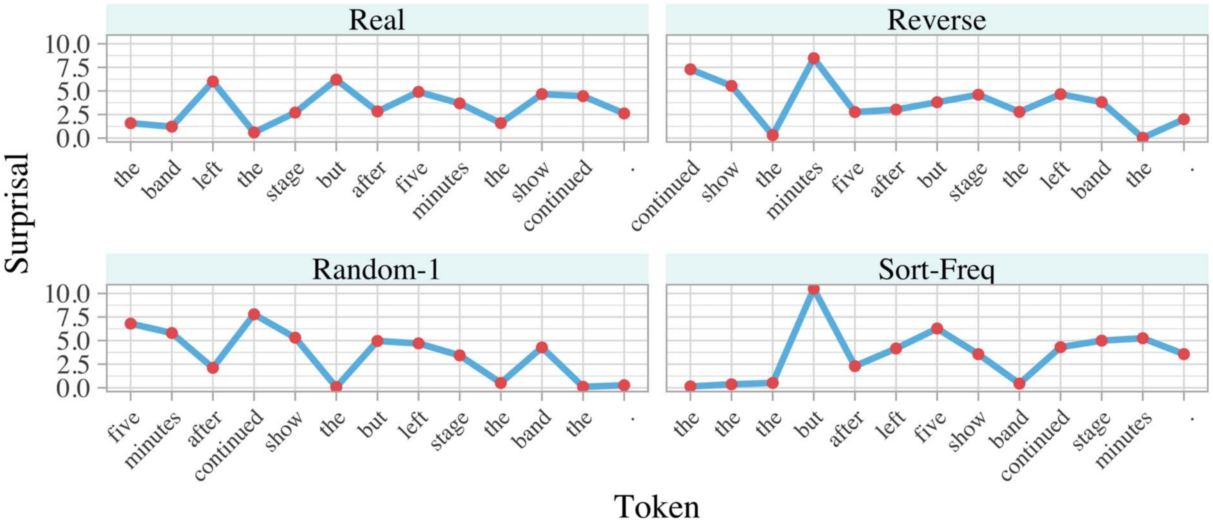

Figure 3: The same source sentence according to 4 real and counterfactual orderings.

We use a state-of-the-art parser (Straka and Straková, 2017) to implement $\texttt{parse} : \mathcal{Y} \to \mathcal{T}$ where $\mathcal{T}$ is the set of all dependency parses. Specifically, we define $\texttt{parse}(\boldsymbol{y})$ = argmax$_{🌲 \in \mathcal{T}}$ $p($🌲$\mid \boldsymbol{y})$ for a learned conditional probability distribution over possible parses $p(\cdot \mid \boldsymbol{y})$. We then obtain the linearized form of the resulting tree by supplying it and the ordering function $\texttt{g}$ to $\texttt{linearize}$, as defined above. Collectively, the outputs of this process (parallel datasets differing only in word order) are referred to as *variants*. Importantly, $\texttt{f}$ here is a deterministic function; one could instead consider $\texttt{f}$ to be probabilistic in nature, with each sentence $\boldsymbol{y}$ having a distribution over tree structures 🌲. We discuss the implications of this choice in §4.

3.2 Counterfactual Grammar Specifications

In addition to the original Real word order, we explore the following theoretically motivated counterfactual grammars for each language. Example sentences from several of these grammars are shown in Figure 3.

Consistent Approximation to Real Order. Approx is a consistent approximation to the real word order within our formalism; it uses an ordering function parameterized by weights that were fitted to maximize the likelihood of observed word orders for each language, as reported by Hahn et al. (2020). This variant captures most of the word order features of a real language while allowing for a fair comparison to deterministic counterfactual grammars that do not model the flexibility of real language. From the perspective of the UID hypothesis, we expect this variant to be less uniform that Real because it has less flexibility to accommodate speakers' choices that optimize for UID.

Consistent Random Grammars. We include variants Random$_1$ through Random$_5$, which use ordering functions parameterized by randomly assigned weights. This means that for a given random grammar, each dependency relation has a fixed direction (left or right), but that the directions of these relations lack the correlations observed in natural language (Greenberg, 1963). Random grammars with the same numerical index share weights across languages.

Consistent Grammars Optimized for Efficiency. We include two consistent grammars that are optimized for the joint objective of parseability (how much information an utterance provides about its underlying syntactic structure) and sentence-internal predictability, as reported by Hahn et al. (2020), one with OV order (Efficient-OV) and one with VO order (Efficient-VO). For example, the Efficient-OV grammar for English would give a plausible version of a consistent and efficient grammar in the counterfactual world where English has verbs after objects.

Grammars Optimized for Dependency Length Minimization. From the same work we also take consistent grammars that are optimized for

DLM, denoted as Min-Dl-Opt. While linearizations produced by these grammars are not guaranteed to minimize dependency length for any particular sentence, they minimize the expected average dependency length of a large sample of sentences in a language. In addition, we include Min-Dl-Loc, an inconsistent grammar that applies the projective dependency-length minimization algorithm of Gildea and Temperley (2007) at the sentence level, leading to sentences with minimal DL but without the constraint of consistency.

Frequency-sorted Grammars. Sort-Freq is an inconsistent grammar which orders words in a sentence from highest to lowest frequency, ignoring dependency structure altogether. We use this ordering as a heuristic baseline for which we expect UID to hold relatively strongly: Low-frequency elements, which tend to have higher surprisal even if solely from their less frequent usage (Ellis, 2002), are given more context, and thus should have smaller surprisals than if they occurred early; more conditioning context tends to reduce the surprisal of the next word (Luke and Christianson, 2016). We also test Sort-Freq-Rev, ordering words from least to most frequent, which for analogous reasons we expect to perform poorly in terms of UID. However, both of these orderings lead to massive syntactic ambiguity by introducing many string collisions—any two sentences containing the same words in different orders would be linearized identically. This eliminates word order as a mechanism for expressing distinctions in meaning, so these orders are implausible as alternatives to natural languages (Mahowald et al., 2022).

Reverse Grammar. Finally, we also include the Reverse variant, where the words in each sentence appear in the reverse order of the original. This variant preserves all pairwise distances between words within sentences and has identical dependency lengths as the original order, thus isolating the effect of linear order on information density from other potential influences. Notably, if the original language happens to be perfectly consistent, then Reverse will also satisfy consistency; in practice, this is unlikely to hold with natural languages.

3.3 UID and Counterfactual Grammars

Let $p_\ell(\boldsymbol{y})$ be the probability distribution over sentences $\boldsymbol{y}$ for a language of interest ℓ. We can define a language's UID score as the expected value of its sentences' UID scores, where we overload the $\textsc{uid}$ function to take either a sentence $\boldsymbol{y}$ or an entire language ℓ:

$$\textsc{uid}(\ell) \overset{\text{def}}{=} \sum_{\boldsymbol{y}\in\mathcal{Y}} p_\ell(\boldsymbol{y})\, \textsc{uid}(\boldsymbol{y}) \tag{5}$$

where sentence-level $\textsc{uid}$ can be $\textsc{uid}_v(\boldsymbol{y})$, $\textsc{uid}_{lv}(\boldsymbol{y})$, or $\textsc{uid}_p(\boldsymbol{y})$. In practice, we estimate this language-level UID score using a Monte-Carlo estimator, taking the mean sentence-level UID score across a held-out test set S_ℓ of sentences $\boldsymbol{y}$ in language ℓ, where we assume $\boldsymbol{y} \sim p_\ell$:

$$\widehat{\textsc{uid}}(\ell) \overset{\text{def}}{=} \frac{1}{|S_\ell|} \sum_{\boldsymbol{y}\in S_\ell} \textsc{uid}(\boldsymbol{y}) \tag{6}$$

Similarly, the expected surprisal (or Shannon entropy, H) of this language is computed as:

$$\mathrm{H}(\ell) \overset{\text{def}}{=} -\sum_{\boldsymbol{y}\in\mathcal{Y}} p_\ell(\boldsymbol{y})\, \log p_\ell(\boldsymbol{y}) \tag{7}$$

We evaluate how well a language model $p_{\boldsymbol{\theta}}$ approximates p_ℓ by its cross-entropy:

$$\mathrm{H}(p_\ell, p_{\boldsymbol{\theta}}) = -\sum_{\boldsymbol{y}\in\mathcal{Y}} p_\ell(\boldsymbol{y})\, \log p_{\boldsymbol{\theta}}(\boldsymbol{y}) \tag{8}$$

where a smaller value of H implies a better model. Again using a Monte Carlo estimator, we measure cross-entropy using the held-out test set S_ℓ:

$$\widehat{\mathrm{H}}(p_\ell, p_{\boldsymbol{\theta}}) = -\frac{1}{|S_\ell|} \sum_{\boldsymbol{y}\in S_\ell} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}) \tag{9}$$

This is simply the *mean surprisal* that the model assigns to a corpus of naturalistic data.

These computations can also be applied to counterfactual variants of a language. Let $\ell_{\mathtt{f}}$ stand for a language identical to ℓ, but where its strings have been transformed by $\mathtt{f}$; this language's distribution over sentences would be $p_{\ell_{\mathtt{f}}}(\boldsymbol{y}) = \sum_{\boldsymbol{y}'\in\mathcal{Y}} p_\ell(\boldsymbol{y}')\ \{\boldsymbol{y} = \mathtt{f}(\boldsymbol{y}')\}$. Since entropy is non-increasing over function transformations (by Jensen's inequality), it follows that:

$$\mathrm{H}(\ell) \geq \mathrm{H}(\ell_{\mathtt{f}}) \tag{10}$$

Further, if our counterfactual generation function $\mathtt{f}$ is a bijection—meaning that each input string gets mapped to a distinct output string and each output string has an input that maps to it—then we can create a second function $\mathtt{f}^{-1} : \mathcal{Y} \to \mathcal{Y}$, which would generate ℓ from $\ell_{\mathtt{f}}$. Then, the following holds:

$$\mathrm{H}(\ell) \geq \mathrm{H}(\ell_{\mathtt{f}}) \geq \mathrm{H}(\ell_{\mathtt{f}^{-1} \circ \mathtt{f}}) = \mathrm{H}(\ell) \tag{11}$$

i.e., it must be that $\mathrm{H}(\ell) = \mathrm{H}(\ell_{\mathtt{f}})$. Reversing a sentence is an example of a bijective function, and thus Equation (11) holds necessarily for the pair of REAL and REVERSE variants; the counterfactual generation procedure thus should not produce differences in mean surprisal between these variants. At the same time, bijectivity does not necessarily hold for our other counterfactual transformations and is violated to a large degree when mapping to SORT-FREQ and SORT-FREQ-REV. Thus in general, we can only guarantee Inequality 10.

Crucially, however, the transformation $\mathtt{f}$ might change the UID score of such a language, allowing us to evaluate the impact of word order on information uniformity. As a simple example, consider the language ℓ_1 that places a uniform distribution over only four strings: aw, ax, by, and bz. In this language, the first and second symbols always have 1 bit of surprisal, and the end of the string has 0 bits of surprisal. If the counterfactual language ℓ_2 is the reverse of ℓ_1, we have a uniform distribution over the strings wa, xa, yb, and zb. Here, the first symbol always has 2 bits of surprisal, and the second symbol and end of sentence always have zero bits, as their values are deterministic for a given initial symbol. While the mean surprisal per symbol is the same for ℓ_1 and ℓ_2, ℓ_1 has more uniform information density than ℓ_2.

4 Limitations

4.1 Use of Counterfactual Grammars

Real Word Orders Are not Consistent. The consistent grammars borrowed from Hahn et al. (2020) assume that the direction of each syntactic relation, as well as the relative ordering of dependents on the same side of a head, are fixed. This is not generally true of natural languages. We address this difference by including the variant APPROX as a comparison to the counterfactual variants, which are constrained by consistency, and by including REVERSE as a comparison to REAL, both of which are not constrained by consistency.

Automatic Parsing Errors. Another issue is that the dependency parses extracted for each original sentence as part of the counterfactual generation pipeline may contain parsing errors. These errors may introduce noise into the counterfactual datasets that is not present in the original sentences, and may cause deviations from the characteristics that we assume our counterfactual grammars should induce. For example, MIN-DL-LOC only produces sentences with minimized dependency length if the automatic parse is correct.

Deterministic Parsing. Finally, our counterfactual generation procedure assumes a deterministic mapping from sentences to dependency trees as one of its steps. However, multiple valid parses of sentences are possible in the presence of syntactic ambiguity. In such cases, we always select the most likely structure according to the parser, which learns these probabilities based on its training data. Therefore, this design choice could lead to underrepresentation of certain syntactic structures when applying a transformation. However, we note that the variants REAL, REVERSE, SORT-FREQ, and SORT-FREQ-REV do not depend on dependency parses and so are unaffected by this design choice.

4.2 Choice of Dataset

Properties of language can vary across genres and domains. When drawing conclusions about human language in general, no single dataset will be completely representative. Due to the amount of data required to train LMs, we use written corpora in this work, and use the term *speaker* loosely to refer to any language producer regardless of modality. To address potential concerns about the choice of dataset in this study, we conducted a supplementary analysis on a subset of languages using a different web corpus, which we report in §7.5.

4.3 Errors and Inductive Biases

Model Errors. Language model quality could impact the estimated values of our UID metrics UID_v, UID_p, and UID_{lv}. To see why, consider a model $p_{\boldsymbol{\theta}}$ that—rather than providing unbiased estimates of p_ℓ—is a smoothed interpolation between p_ℓ and the uniform distribution:

$$p_{\boldsymbol{\theta}}(y_n \mid \boldsymbol{y}_{<n}) = \lambda \, p_\ell(y_n \mid \boldsymbol{y}_{<n}) + \frac{1-\lambda}{|\overline{\mathcal{V}}|} \quad (12)$$

for $\lambda \in [0, 1]$. Here, an increase in $1-\lambda$ would lead to an increase in $\mathrm{H}(p_\ell, p_{\boldsymbol{\theta}})$, since the cross-entropy is only minimized when $p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{y}_{<n}) = p_\ell(\cdot \mid \boldsymbol{y}_{<n})$. This change, however, would be reflected as an *increase* in uniformity, e.g., a decrease in UID_v: Surprisals would be closer to uniform for smaller values of λ. Alternatively, consider the situation where a language ℓ has perfect information uniformity, i.e., where UID_v, UID_p, and UID_{lv} are their minimum possible values. The interpolation of p_ℓ with any non-uniform distribution should instead *decrease* the measured uniformity, at least with respect to UID_v and UID_{lv}.

In summary, our UID metrics could be biased either positively or negatively by the quality of our models. However, since our analysis focuses on the comparison of UID metrics between word order variants rather than their absolute value, this bias should not be a major concern. We use the same model architecture for all language–variant combinations, and so a bias in the UID metric corresponding to one combination should likewise be reflected in all of the metrics that it is compared to. Further, our results hold even when controlling for mean surprisal, as described in §6.

Inductive Biases. Because modern LMs have been developed to model natural language, they may contain subtle biases towards the properties of real word orders or of highly resourced languages. Based on Inequality (10), if two probabilistic models m_ℓ and $m_{\ell_{\mathtt{f}}}$ were to perfectly learn the true and counterfactual distributions p_ℓ and $p_{\ell_{\mathtt{f}}}$, respectively, then m_ℓ should assign approximately the same or higher mean surprisal to a corpus $\{\boldsymbol{y}^{(m)}\}_{m=1}^{M}$ from ℓ than $m_{\ell_{\mathtt{f}}}$ assigns to the counterfactual corpus from $\ell_{\mathtt{f}}$. This implies that previous results of Gildea and Jaeger (2015), Ravfogel et al. (2019), Hahn et al. (2020), and White and Cotterell (2021), which found that real corpora tend to have lower average per-word surprisal than deterministically generated counterfactual versions of the same corpora, were in fact due to the inductive bias of the learning algorithms used to estimate surprisals. There is a clear reason why the trigram model of Gildea and Jaeger (2015) would yield higher mean surprisals for counterfactual corpora: The transformation functions $\mathtt{f}$ tended to increase dependency lengths, and words in a dependent–head relation tend to have higher mutual information than other pairs of words (Futrell and Levy, 2017; Futrell et al., 2019, 2020). Hence the transformations tended to push words that are predictive of each other outside of the conditioning window of the model (see also Hahn and Xu, 2022, for similar effects). The Transformer architecture we use in this work could thus also contain biases favoring features of real language, which we attempt to control for (see §6).

5 Experimental Setup

5.1 Data

This work uses the publicly available Wiki40b dataset (Guo et al., 2020), a large text corpus derived from Wikipedia articles. We use subsets of the Wiki40b dataset in 10 languages: English, Russian, French, German, Hindi, Farsi, Vietnamese, Indonesian, Hungarian, and Turkish. The first six represent the Germanic, Slavic, Romance, Indo-Aryan, and Iranian sub-families of the Indo-European language family. The latter four belong to the Austroasiatic, Austronesian, Uralic, and Turkic language families, respectively. Turkish, Hindi, and Farsi have basic SOV word order, while the other languages have SVO order, with Hungarian being mixed (Dryer, 2013). Languages were chosen based on the amount of available data in the Wiki40b dataset, their typological properties (covering a range of families, canonical word orders, and morphological complexity), and availability of automatic dependency parsing models.

The datasets are subsampled to yield approximately 20M words in the training set of each language and approximately 1M words in the test and validation sets. We automatically generate dependency parses for all sentences using the UDPipe parser (Straka and Straková, 2017), yielding syntactic representations in the UD paradigm. We then apply each of the counterfactual orderings introduced in §3.2 to the original data to create parallel corpora for each language. Sentences are stripped of punctuation (as determined by the dependency parser's PUNCT label) and are lowercased. Periods are added back in to mark the end of sentences, regardless of what the original final

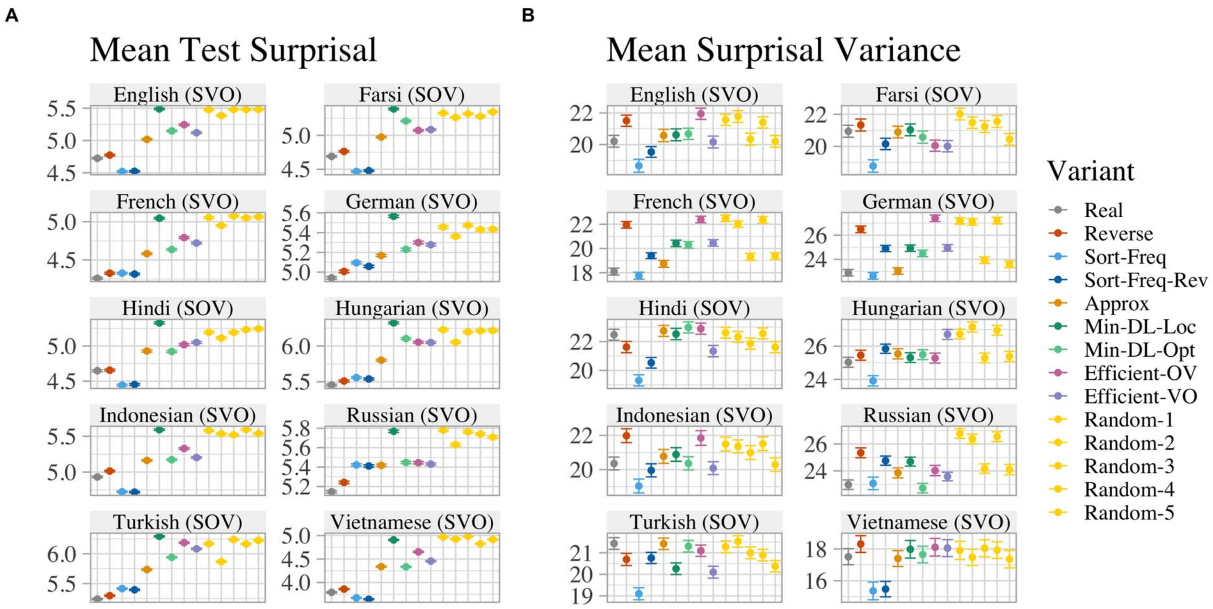

Figure 4: Mean test-set surprisal and surprisal variance of language models across real and counterfactual grammars in 10 languages. Error bars denote the 95% CI of the mean.

punctuation was. Sub-word tokenization is then applied to the corpora using a byte-pair encoding (BPE) model, trained with a fixed vocabulary size of 30K tokens and using the algorithm of Sennrich et al. (2016).[7]

5.2 Language Modeling

For each variant of each language, we train a Transformer language model (Vaswani et al., 2017) using `fairseq` (Ott et al., 2019). Models are trained on document-level inputs, with a maximum length of 512 tokens; this means that each token is predicted with the preceding material of the entire document as context. Each model is trained with early stopping, halting training after no improvement in validation loss for three epochs. The Adam optimizer was used (Kingma and Ba, 2017), with a learning rate of 0.0005, weight decay of 0.01, and dropout of 0.1. Training scripts are available in the project's GitHub repository.[1] In all of our analyses, we use the word-by-word surprisals estimated using our trained models on their corresponding held-out test sets. Note that we do not consider the designated EOS symbol in the computation of any of our UID-related metrics. In the case that a word is composed of multiple sub-word tokens, we aggregate their surprisals by summation, since surprisal decomposes additively.

6 Results

Estimates of mean per-word surprisal on the test set are in Figure 4A. Consistent with the results of Hahn et al. (2020), our trained models for nearly all counterfactual variants assign higher per-word surprisal to their respective test sets than the Real models assign to theirs. Across all 10 languages, Reverse has mean surprisal close to, but consistently slightly higher than, that of the real ordering. Sort-Freq and Sort-Freq-Rev have mean surprisals close to or below those of Real.

Estimates of mean surprisal variance (UID_v) over sentences are shown in Figure 4B. Notably, there is a dissociation between the rank order of variants according to mean surprisal and according to UID_v: Variants with similar mean surprisals did not necessarily have similar UID_v scores, and vice versa, suggesting that information uniformity and mean surprisal can vary independently of each other. Our main observations are as follows: (i) In all languages except Turkish and Hindi, our estimates of UID_v for Real are lower than those for Reverse, despite the variants' similarities in mean surprisal. (ii) As predicted, the Sort-Freq baseline has UID_v equal to or lower than that of Real. (iii) The other counterfactual variants typically

[7] All variants of the same language are tokenized using the same BPE model, trained on a sample of 100K documents from all variants; BPE tokens could not cross word boundaries for compatibility with different word orders.

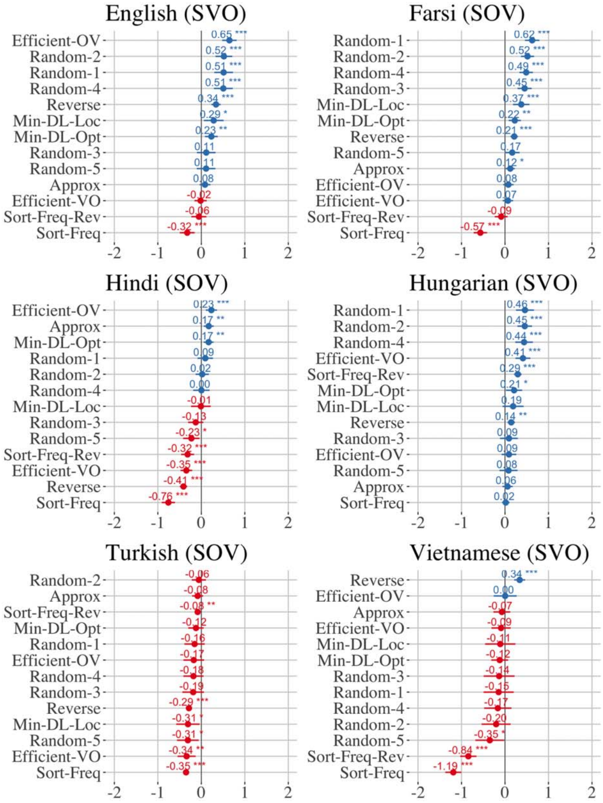

Figure 5: Linear regression coefficient estimates when predicting UID$_p$ as a function of mean surprisal, variant, and dataset size. The reference level for variant is Real, so positive coefficients (blue) indicate variants with greater UID$_p$, i.e., less uniformity, than real language.

exhibit higher UID$_v$ than Real, with the exception of mixed results for Sort-Freq-Rev. (iv) The Efficient-VO variants typically have lower UID$_v$ than Efficient-OV (with Hungarian being a noteworthy exception), which supports findings based on toy grammars showing that SVO orders are more uniform than SOV orders (Maurits et al., 2010). Crucially, these results are qualitatively similar using the UID$_{lv}$ metric (Figure 6B).

To fairly compare variants using the UID$_p$ metric, we first need to account for the fact that, unlike surprisal variance, the metric is sensitive to shifts in mean surprisal. To control for this, we fit a regression model predicting the UID$_p$ score based on three variables: The mean surprisal, the grammar variant, and the dataset size (20M, 6.6M, and 2.2M words). We train multiple language models for each language-variant combination (3 dataset sizes and 2 random seeds), resulting in 84 data points per language. We apply treatment coding to the variants, with Real as the reference level. Figure 5 shows the resulting estimates of the coefficients for each variant, where a coefficient should be positive if that variant is less uniform than Real. Qualitatively, the regression results match the results given by UID$_v$ and UID$_{lv}$: Real is more uniform than Reverse in SOV languages, Sort-Freq is the only counterfactual variant that is consistently more uniform than Real, and Efficient-VO is more uniform than Efficient-OV in most languages; the opposite is true in Hungarian and the difference is negligible in Russian.

7 Discussion

We offer a discussion of the results observed in §6, including their implications for the role of functional pressures in language.

7.1 Differences in Mean Surprisal

Across 10 typologically diverse languages, we find that Transformer LMs learn to predict data from real word orders better than data from counterfactual orders, with the exception of the Sort-Freq and Sort-Freq-Rev variants. This suggests that these LMs' inductive biases somehow favor properties of real languages, in line with previous work on other modeling architectures (Gildea and Jaeger, 2015; Ravfogel et al., 2019). This is not surprising, given that commonly used architectures and hyperparameters have been selected specifically based on their good performance on real language tasks. Unlike in n-gram models, the precise inductive bias of Transformer models that favors real word orders is not transparent and merits further study.[8]

7.2 Differences Between Real and Approx

We observe that despite the similarities between the Real and Approx variants of a given language, the latter are consistently assigned higher mean surprisal by their respective LMs. Meanwhile, the various UID metrics show similar results for Real and Approx, suggesting that the greater flexibility of Real is not responsible for UID differences in our results. This is somewhat surprising, since it may appear that such flexibility is what enables speakers' choices, which have been previously discussed as contributing to UID. However, many speaker choices that potentially impact UID, such as word choice, active versus passive voice, and optional words, are not captured by this difference in flexibility between Real and Approx.

7.3 Greater Uniformity of Real over Reverse in SVO Languages

While mean surprisal is always very close for Real and Reverse grammars, Reverse is less uniform in 8 out of 10 languages, including all SVO languages. This held across multiple operationalizations of UID, with the exception of mixed results for Hungarian, a language with considerable flexibility in word order. Thus, while both Real and Reverse orders are learned approximately equally well by language models, they differ in how uniformly they distribute information.

One key difference between Real and Reverse is that insofar as Real sentences exhibit a tendency to mention entities from the end of a given sentence close to the beginning of the next one, Reverse does not preserve this property. For example, the pair of sentences ``I like dogs. They are friendly.'' would become ``Dogs like I. Friendly are they.''; note that the distance between antecedent and pronoun is significantly increased. This feature of the Reverse raises the possibility that the uniformity patterns we observe are due to speaker choices taking cross-sentence dependencies into consideration. To minimize the influence of cross-sentence dependencies, we can consider only sentences occurring at the start of a document, which cannot refer to previous sentences. Figure 6A shows that the tendency for Real to have lower surprisal variance than Reverse still holds in this setting across most languages. This suggests that cross-sentence dependencies alone cannot fully explain the observed differences in information uniformity.

Notably, our results show that the UID preference for Real over Reverse is not consistently present in languages with basic SOV order (Turkish, Hindi, and Farsi). We propose the following explanation for this result: As argued in Maurits et al. (2010), SVO languages tend to have more uniform information density profiles than SOV languages—a finding supported by our empirical results in which Efficient-VO had lower surprisal variance than Efficient-OV in 9 out of 10 languages. Unlike the short, simple sentences of Maurits et al., however, the present study considers long and complex sentences where speaker choices have considerable opportunity to influence information uniformity, in addition to the role of basic word order. These choices include whether to use a pronoun, whether to use an active or passive construction, and what order to present a conjunction or list of items, among others. Importantly, speakers make choices conditional on the forward ordering of real language, so we expect that the choices made in an attempt to increase UID—which constitutes a non-trivial percentage of utterances (Levy and Jaeger, 2007)—would have a greater effect on UID in Real than in Reverse. In SVO languages, the effects upon UID of basic word order and speaker choices both go in the same direction: towards more uniformity. In SOV languages, these effects conflict: The basic word order is non-optimal in terms of UID, and so uniformity can theoretically be increased by a transformation to Reverse, while speaker choices are presumably already mostly optimal in Real. This may explain the heterogeneous patterning among the three SOV languages.

[8] Notably, White and Cotterell (2021) show that there is a large variation in how Transformer language models perform in toy languages with diverse word orders; they, however, do not find evidence that Transformers perform better on the most frequently occurring orders (as opposed to, e.g., OVS and VOS word orders, which are found in few languages).

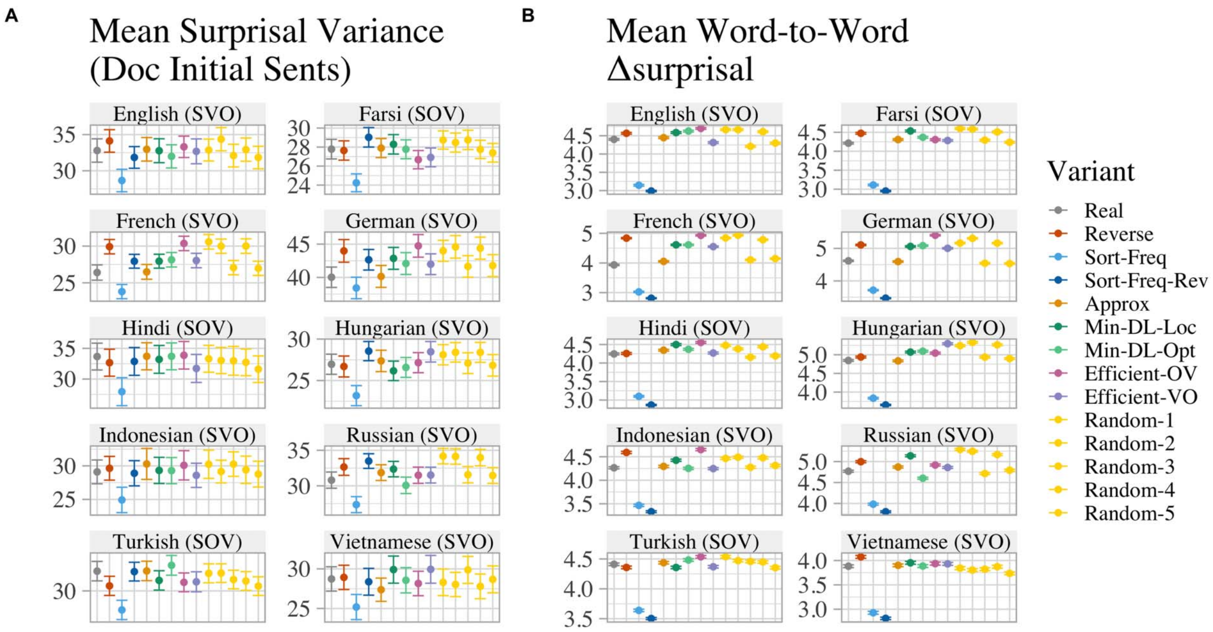

Figure 6: A. Surprisal variance ($\text{UID}_v(\boldsymbol{y})$) for document-initial sentences only. B. Mean squared word-to-word change ($\text{UID}_{lv}(\boldsymbol{y})$) in surprisal. Error bars denote 95% CI of the mean.

Furthermore, these results can potentially shed light on an important question in linguistic typology: Why are some basic word orders more common than others? According to some theories, SOV order (the most typologically common) is the most natural for expressing events with subjects and objects (Goldin-Meadow et al., 2008; Gibson et al., 2013; Futrell et al., 2015a). If these theories are correct, an evolutionary pressure on languages to shift from SOV to SVO could help account for the prevalence of SVO languages, which are nearly as common as SOV ones. A pressure for information uniformity offers one such account.

Finally, Pimentel et al. (2021a) have recently shown that the distribution of per-phone information *within words* is more uniform when analysed in reverse order than in forward order—the opposite of what we observe on our sentence-level analysis. This difference may suggest qualitatively distinct information-theoretic pressures being present at the lexical and sentential levels and is a potential topic for further study.

7.4 Other Variants

The variants designed to minimize dependency length, MIN-DL-LOC and MIN-DL-OPT, showed mixed results in terms of information uniformity compared to REAL. The random grammars fell into two groups: RANDOM$_1$, RANDOM$_2$, and RANDOM$_4$ tended to be less uniform than REAL, while RANDOM$_3$ and RANDOM$_5$ tended to be similar in uniformity to REAL. Since random grammars have fixed but uncorrelated directions of syntactic relations, these cross-linguistically consistent patterns suggest that some settings of the parameterized grammar are inherently more favorable from the perspective of UID than others.

The only counterfactual word order to consistently have a higher degree of information uniformity than the real orders was the highly constrained SORT-FREQ, which turns sentences into sorted word lists. Thus, while it appears possible to improve on real word orders' information uniformity, this comes at the cost of massive syntactic ambiguity and reduced expressivity.

7.5 Robustness to Dataset Choice

In this study, the chosen dataset (Wiki40b) contains formal writing that may not exhibit the same communicative pressures as spoken language. It is largely devoid of first and second person pronouns, interrogatives, and other features common in everyday speech; further, it may have disproportionate amounts of translationese (Koppel and Ordan, 2011). As a supplementary analysis, we repeated the experiments on the CC100 dataset (Conneau et al., 2020), using only a subset of languages due to computational constraints. This dataset is sourced from a web crawl and therefore

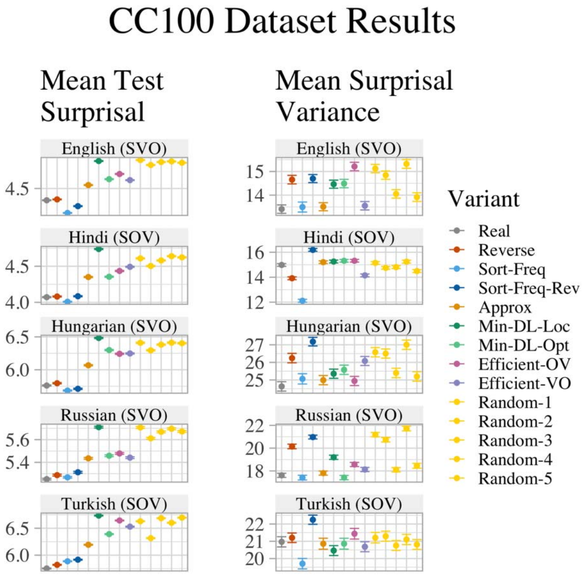

Figure 7: Surprisal mean and variance for a subset of languages on the CC100 dataset. Error bars denote 95% CI.

contains a wider range of genres and styles than Wiki40b. UID$_v$ scores for these experiments are shown in Figure 7. The results qualitatively match the patterns from the Wiki40b experiments in the following ways: (i) better UID$_v$ scores for Real than for Reverse among SVO languages, (ii) better UID$_v$ scores for Efficient-VO than Efficient-OV in most languages (with Hungarian again being an exception), and (iii) the only variant that has higher uniformity that Real across a majority of languages is Sort-Freq.

8 Conclusion

In conclusion, we have empirically demonstrated that in many languages, real word orders distribute information more uniformly than a range of counterfactual orders. The fact that this pattern holds in every SVO languages but is mixed among SOV languages lends support to the view that SVO basic word order is preferable to SOV order from the perspective of maximizing UID. We posit that there are two potential sources of optimization within a language for greater UID: Language evolution favoring word orders that produce less variance in information content, and speaker choices in favor of constructions that smooth the information profile of utterances. Our results are consistent with the UID hypothesis, and support the idea that communicative pressures (operationalized in terms of information theory) influence the structure of human language.

Acknowledgments

We thank our action editor and the anonymous reviewers for their detailed feedback on this paper. CM was supported by the Google PhD Fellowship. TP was supported by a Facebook PhD Fellowship. This work was supported by NSF grant BCS-2121074 to RPL.

References

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56. https://doi.org/10.1177/00238309040470010201, PubMed: 15298329

Brian Bartek, Richard L. Lewis, Shravan Vasishth, and Mason R. Smith. 2011. In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5):1178–1198. https://doi.org/10.1037/a0024194, PubMed: 21707210

Jelke Bloem. 2016. Testing the processing hypothesis of word order variation using a probabilistic language model. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 174–185, Osaka, Japan. The COLING 2016 Organizing Committee.

Thomas Hikaru Clark, Ethan Gotlieb Wilcox, Edward Gibson, and Roger P. Levy. 2022. Evidence for availability effects on speaker choice in the Russian comparative alternation. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*.

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, 43(5):651–681. https://doi.org/10.1007/s10936-013-9273-3, PubMed: 24077911

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Matthew S. Dryer. 2013. Order of subject, object and verb (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*.

Li Du, Lucas Torroba Hennigen, Tiago Pimentel, Clara Meister, Jason Eisner, and Ryan Cotterell. 2022. A measure-theoretic characterization of tight language models. *arXiv preprint arXiv: 2212.10502v1*. https://doi.org/10.48550/arXiv.2212.10502

Nick Ellis. 2002. Frequency effects in language processing. *Studies in Second Language Acquisition*, 24(2):143–188. https://doi.org/10.1017/S0272263102002024

August Fenk and Gertrud Fenk. 1980. Konstanz im Kurzzeitgedächtnis—Konstanz im sprachlichen Informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie*, 27(3):400–414.

Ramon Ferrer-i-Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70(5 Pt 2):056135. https://doi.org/10.1103/PhysRevE.70.056135, PubMed: 15600720

Ramon Ferrer-i-Cancho, Carlos Gómez-Rodríguez, and Juan Luis Esteban. 2018. Are crossing dependencies really scarce? *Physica A: Statistical Mechanics and its Applications*, 493:311–329. https://doi.org/10.1016/j.physa.2017.10.048

A. Frank and T. F. Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Cognitive Science Society*.

Richard Futrell, Tina Hickey, Aldrin Lee, Eunice Lim, Elena Luchkina, and Edward Gibson. 2015a. Cross-linguistic gestures reflect typological universals: A subject-initial, verb-final bias in speakers of diverse languages. *Cognition*, 136:215–221. https://doi.org/10.1016/j.cognition.2014.11.022, PubMed: 25498747

Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain. Association for Computational Linguistics.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–413. https://doi.org/10.1353/lan.2020.0024

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015b. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences, U.S.A.*, 112(33):10336–10341. https://doi.org/10.1073/pnas.1502134112, PubMed: 26240370

Richard Futrell, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. 2019. Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 3–13, Paris, France. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-7703

Georg von der Gabelentz. 1901. *Die Sprachwissenschaft, ihre Aufgaben, Methoden, und bisherigen Ergebnisse*, 2nd edition. C. H. Tauchnitz, Leipzig.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073117

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies, UDW@EMNLP 2018, Brussels,* *Belgium, November 1, 2018*, pages 66–74, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/w18-6008

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76. https://doi.org/10.1016/S0010-0277(98)00034-1, PubMed: 9775516

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126, Cambridge, MA. MIT Press.

Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407. https://doi.org/10.1016/j.tics.2019.02.003, PubMed: 31006626

Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. 2013. A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, 24(7):1079–1088. https://doi.org/10.1177/0956797612463705, PubMed: 23649563

Daniel Gildea and T. Florian Jaeger. 2015. Human languages order information efficiently. *arXiv preprint arXiv:1510.02823*. https://doi.org/10.48550/arXiv.1510.02823

Daniel Gildea and David Temperley. 2007. Optimizing grammars for minimum dependency length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191, Prague, Czech Republic. Association for Computational Linguistics.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310. https://doi.org/10.1111/j.1551-6709.2009.01073.x, PubMed: 21564213

Susan Goldin-Meadow, Wing Chee So, Asli Özyürek, and Carolyn Mylander. 2008. The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences, U.S.A.*, 105(27):9163–9168. https://doi.org/10.1073/pnas.0710060105, PubMed: 18599445

Joseph H. Greenberg. 1963. *Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements*. MIT Press, Cambridge, MA.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29(2):261–290. https://doi.org/10.1207/s15516709cog0000_7, PubMed: 21702774

Mandy Guo, Zihang Dai, Denny Vrandecic, and Rami Al-Rfou. 2020. Wiki-40b: Multilingual language model dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2020. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences, U.S.A.*, 117(5):2347–2353. https://doi.org/10.1073/pnas.1910923117, PubMed: 31964811

Michael Hahn and Yang Xu. 2022. Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality. *Proceedings of the National Academy of Sciences, U.S.A.*, 119(24):e2122604119. https://doi.org/10.1073/pnas.2122604119, PubMed: 35675428

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8. https://doi.org/10.3115/1073336.1073357

Martin Haspelmath. 2008. Parametric versus functional explanations of syntactic universals. In T. Biberauer, editor, *The Limits of Syntactic Variation*, pages 75–107. John Benjamins, Amsterdam. https://doi.org/10.1075/la.132.04has

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

John A. Hawkins. 1994. *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199252695.001.0001

John A. Hawkins. 2014. *Cross-linguistic Variation and Efficiency*. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199664993.001.0001

Jacob L. Hoover, Morgan Sonderegger, Steven T. Piantadosi, and Timothy J. O'Donnell. 2022. The plausibility of sampling as an algorithmic theory of sentence processing. *PsyArXiv preprint PsyArXiv:qjnpv*. https://doi.org/10.31234/osf.io/qjnpv

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002, PubMed: 20434141

T. Florian Jaeger and Harry Tily. 2011. On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):323–335. https://doi.org/10.1002/wcs.126, PubMed: 26302080

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Marco Kuhlmann. 2010. *Projective Dependency Structures*, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14568-1_3

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006, PubMed: 17662975

Roger Levy. 2018. Communicative efficiency, uniform information density, and the rational speech act theory. In *Proceedings for the 40th Annual Meeting of the Cognitive Science Society*, pages 684–689. https://doi.org/10.31234/osf.io/4cgxh

Roger Levy and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*, 19:849–856.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191. https://doi.org/10.17791/jcs.2008.9.2.159

Steven G. Luke and Kiel Christianson. 2016. Limits on lexical prediction during reading. *Cognitive Psychology*, 88:22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Kyle Mahowald, Evgeniia Diachek, Edward Gibson, Evelina Fedorenko, and Richard Futrell. 2022. Experimentally measuring the redundancy of grammatical cues in transitive clauses. *arXiv preprint arXiv:2201.12911*. https://doi.org/10.48550/arXiv.2201.12911

Marie-Catherine de Marneffe, 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

Luke Maurits, Dan Navarro, and Amy Perfors. 2010. Why are some word orders more common than others? A Uniform Information Density account. In *Advances in Neural Information Processing Systems*, pages 1585–1593.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic.

Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.74

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4009

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences, U.S.A.*, 108(9):3526–3529. https://doi.org/10.1073/pnas.1012551108, PubMed: 21278332

Tiago Pimentel, Ryan Cotterell, and Brian Roark. 2021a. Disambiguatory signals are stronger in word-initial positions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 31–41, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.3

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021b. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 949–962, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.3

Tiago Pimentel, Irene Nikkarinen, Kyle Mahowald, Ryan Cotterell, and Damián Blasi. 2021c. How (non-)optimal is the lexicon? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4426–4438, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.350

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of RNNs with synthetic variations of natural languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3532–3542, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1356

Jan Rijkhoff. 1986. Word order universals revisited: The principle of head proximity. *Belgian Journal of Linguistics*, 1:95–125. https://doi.org/10.1075/bjl.1.05rij

Jan Rijkhoff. 1990. Explaining word order in the noun phrase. *Linguistics*, 28(1):5–42. https://doi.org/10.1515/ling.1990.28.1.5

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger P. Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time. *PsyArXiv preprint PsyArXiv:4hyna*. https://doi.org/10.31234/osf.io/4hyna

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319. https://doi.org/10.1016/j.cognition.2013.02.013, PubMed: 23747651

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-3009

David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:1–15. https://doi.org/10.1146/annurev-linguistics-011817

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jennifer C. White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 454–463, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.38

Himanshu Yadav, Samar Husain, and Richard Futrell. 2021. Do dependency lengths explain constraints on crossing dependencies? *Linguistics Vanguard*, 7(s3). https://doi.org/10.1515/lingvan-2019-0070

Meilin Zhan and Roger Levy. 2018. Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese. 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 1997–2005, New Orleans, Louisiana. Association for Computational Linguistics. http://doi.org/10.18653/v1/N18-1181

George Kingsley Zipf. 1935. *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Houghton-Mifflin, Boston.

Ran Zmigrod, Tim Vieira, and Ryan Cotterell. 2020. Please mind the root: Decoding arborescences for dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4809–4819, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.390