Can Language Models Learn Typologically Implausible Languages?

Tianyang Xu[a,*] **Tatsuki Kuribayashi**[c] **Yohei Oseki**[d]
Ryan Cotterell[b] **Alex Warstadt**[e,*]
[a]Toyota Technical Institute at Chicago [b]ETH Zürich [c]MBZUAI
[d]The University of Tokyo [e]University of California San Diego
sallyxu@ttic.edu tatsuki.kuribayashi@mbzuai.ac.ae
oseki@g.ecc.u-tokyo.ac.jp rcotterell@inf.ethz.ch awarstadt@ucsd.edu

Abstract

Grammatical features across human languages show intriguing correlations often attributed to learning biases in humans. However, empirical evidence has been limited to experiments with highly simplified artificial languages, and whether these correlations arise from domain-general or language-specific biases remains a matter of debate. Language models (LMs) provide an opportunity to study artificial language learning at a large scale and with a high degree of naturalism. In this paper, we begin with an in-depth discussion of how LMs allow us to better determine the role of domain-general learning biases in language universals. We then assess learnability differences for LMs resulting from typologically *plausible* and *implausible* languages closely following the word-order "universals" identified by linguistic typologists. We conduct a symmetrical cross-lingual study training and testing LMs on an array of highly naturalistic but counterfactual versions of the English (head-initial) and Japanese (head-final) languages. Compared to similar work, our datasets are more naturalistic and fall closer to the boundary of plausibility. Our experiments show that these LMs are often slower to learn these subtly implausible languages, while ultimately achieving similar performance on some metrics regardless of typological plausibility. These findings lend credence to the conclusion that LMs do show some typologically-aligned learning preferences, and that the typological patterns may result from, at least to some degree, domain-general learning biases.

https://github.com/sally-xu-42/Typological_Universals

1 Introduction

A fundamental goal in linguistics is to elucidate the universal properties underlying attested natural languages and to explain why some conceivable grammars but not others are widely attested. Many typological universals and tendencies have been identified (Greenberg, 1963; Barwise and Cooper, 1988; Dryer, 1992; Hyman, 2008), but their causes are more elusive. There is disagreement over whether typological patterns are caused by a learning bias that is language-specific (Chomsky, 1965) or domain-general (Culbertson and Kirby, 2016), or even whether such a bias is the cause at all (Hahn et al., 2020). This debate has been difficult to resolve because we cannot manipulate variables during acquisition of a child's first language. However, language models (LMs) have recently been advocated for as a convenient model for human learners that can enable large-scale controlled experiments on language acquisition (Warstadt, 2022).

Relatedly, a lively literature on counterfactual language learning in LMs has developed (Ravfogel et al., 2019; Hahn et al., 2020; White and Cotterell, 2021; Clark et al., 2023; Kallini et al., 2024; Kuribayashi et al., 2024, i.a.), sparking some debate. Chomsky et al. (2023) criticized neural language models (LMs) as having little consequence for linguistic theory precisely because they can putatively learn both possible and impossible languages (Mitchell and Bowers, 2020). In response, Kallini et al. (2024) performed a set of experiments to test neural LMs' learnability of data with uncontroversially impossible properties as a natural language (e.g., lacking hierarchical structure), finding instead that LMs do indeed struggle with learning typologically impossible languages.

In this paper, we advance these debates by testing the learnability of typologically dispreferred languages that fall closer to the boundary of possibility. The typological tendencies we study are those famously enumerated by Greenberg (1963) and subsequently refined based on larger-scale typological studies (Dryer, 1992). For example, languages with dominant subject-verb-object (SVO) order

*Work conducted partially at ETH Zürich.

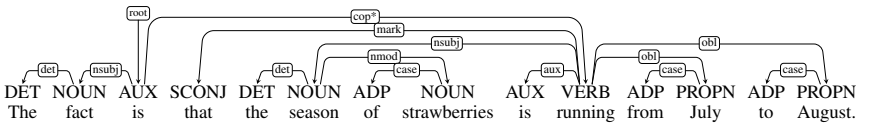
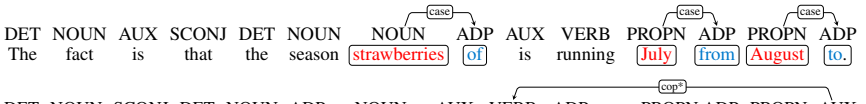
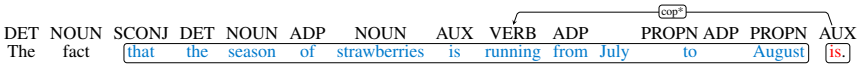
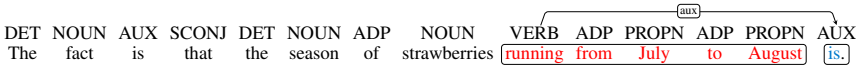
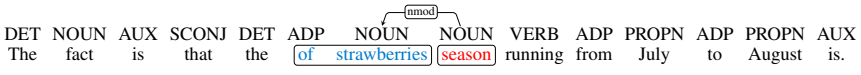
Correlation Pair	Example
Original	
<V, O>	
<Adp, NP>	
<Cop, Pred>	
<Aux, V>	
<Noun, Genitive>	

Table 1: Illustrative examples of each of our counterfactual variants of English. Head phrases are colored red, and dependent phrases are colored blue. In the *<V, O>* example, we do not swap the copula and predicate due to readability, but these elements would be swapped in the actual dataset. The *<V, O>* example demonstrates the reflective swapping ($H\ D_1\ D_2 \rightarrow D_2\ D_1\ H$) explained in §4.1.

overwhelmingly have prepositions, while subject-object-verb (SOV) languages tend to have postpositions. While previous work cited above has tested learnability of artificial languages with LMs, our approach to constructing counterfactual corpora has a unique combination of properties: We aim to maximize naturalness by manipulating pre-existing natural language corpora and by iteratively annotating the counterfactual data and identifying and correcting corner cases. We also target the decision boundary between typologically *plausible* and *implausible* languages by individually manipulating one specific grammatical property in each counterfactual corpus. Finally, we balance biases due to the source language by symmetrically applying this procedure to a head-initial language (English) and a head-final language (Japanese).

In our experiments, we test the learnability of two types of LMs (autoregressive and masked) from scratch on each of our counterfactual languages. We evaluate learnability from multiple perspectives: (i) perplexity per token on the entire corpus, (ii) preferences on minimal pairs targeting the manipulated feature; and (iii) broad syntactic tests (BLiMP, Warstadt et al., 2020; and JBLiMP, Someya and Oseki, 2023). Our experimental results show that LMs often struggle to learn counterfactual, typologically implausible languages relative to minimally different natural languages. Thus we extend the findings of Kallini et al. (2024) on *possible* vs. *impossible* languages even closer to the boundary between *plausible* vs. *implausible* languages. While we cannot entirely rule out confounds due to errors introduced in the creation of counterfactual corpora, these findings have important implications if they prove to be robust. We also argue, contra Chomsky et al. (2023) and inspired by other recent arguments (Linzen, 2019; Warstadt and Bowman, 2022; Wilcox et al., 2023; Constantinescu et al., 2024) that learnability results from LMs can have important implications for our understanding of human language: If Transformers, which lack language-specific learning biases, show a preference for typologically plausible languages, it is likely that humans have a similar learning preference as a result of domain-general learning biases. Our results tentatively support this conclusion — language-specific bias is not necessary, at least as a minimum requirement, to distinguish between typologically plausible and implausible word orders, pointing to a potential new line of evidence on a long-standing debate about the origins of linguistic typological patterns.

2 Background

2.1 Typological Tendencies

What are all the conceivable grammars that human language could have? While this might seem like

an unanswerable question, linguistic theory gives us a particular kind of answer: One of the key insights of modern linguistics is that natural language grammars can be viewed as instantiations of formal languages (Chomsky, 1956). Under this view, it becomes clear that there are conceivable classes of formal languages – for example the regular languages – to which no natural language belongs. But decades of research have shown that human languages occupy a much harder-to-define region within the high-dimensional space of possible grammars (Newmayer, 2005; Chomsky and Lasnik, 1993). Many generalizations have been made about the space of possible human languages, including generalizations about syntactic categories (Chomsky, 1965), quantifiers (Barwise and Cooper, 1988), and phonology (Hyman, 2008), to name just a few. While some of these generalizations are true universals that no human language violates—for instance, no language has rules that require counting surface positions greater than two (Newmayer, 2005)—other generalizations are merely correlations of features that occur far more frequently than if features were sampled independently at random. Thus, we can distinguish between **impossible** and **implausible** languages.

In the latter category, Greenberg (1963) proposed a list of several dozen word order and morphological correlations based on a survey of 30 languages; for example, "In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes." Subsequently, Dryer (1992) formulated a list of **correlation pairs**, that is, a list of pairs of morphosyntactic categories H and D[1] that tend to have the *same* relative ordering as the dominant order of the verb and object, respectively, across a sample of 625 languages. Following Culbertson and Newport (2015) we refer to languages (including most human languages) that follow these typological correlations as **harmonic** (i.e., plausible), while languages that violate them are **non-harmonic** (i.e., implausible). A subset of these correlation pairs that we focus on in this paper is listed in Table 1.

[1]Syntacticians disagree on the correct generalization that characterizes these correlation pairs (Hawkins, 1983; Dryer, 1992; Kayne, 1994), so there is no entirely theory-neutral description, perhaps besides "verb-patterners" and "object-patterners". As suggested by our notation, the H elements that pattern with the verb tend to be functional heads or lexical heads, while the D elements that pattern with the object tend to be phrasal arguments or dependents. For example, the adposition is the functional head of an adpositional phrase. While we tend to refer to these elements as heads and dependents, our study is predicated only on the existence of these correlation pairs, not the correct theoretical analysis.

2.2 Learnability of Implausible Languages

Learnability has long been proposed as a primary mechanism behind typological universals and tendencies such as word order harmony. This mechanism has an appealing story: Language evolves through re-analysis by child learners (Peyraube, 1912; Cournane, 2017), and re-analysis tends to favor easier-to-learn grammars, leading them to become more frequent on the scale of generations (Kirby et al., 2008). But why are some grammars harder or easier to learn in the first place? Some scholars propose that humans have **language-specific biases**. For instance, Chomsky's (1965) theory of **Universal Grammar** posited that humans have an innate language acquisition device that biases the learning of certain grammars. This theory was later refined into the theory of **Principles and Parameters** (Chomsky and Lasnik, 1993), which – most relevant to the present discussion – included a **Head Parameter** determining whether complements come to the left or right of their heads (p. 35). Other scholars favor the view that **domain-general biases** are sufficient to explain some typological patterns. For instance, humans appear to have a **simplicity bias** across several domains of cognition (Chater and Vitányi, 2003; Hsu et al., 2013), and such a bias could explain the preference for harmonic languages (Culbertson and Kirby, 2016), as harmonic grammars presumably have a shorter description length than non-harmonic ones.

From the empirical side, the evidence that human learning biases favor typologically plausible languages comes largely from artificial language learning experiments in laboratory settings. Studies of this kind have shown that humans regularize novel grammatical rules in typologically plausible ways in the domains of phonology (Wilson, 2006) and morphology (Kam and Newport, 2005; Fedzechkina et al., 2012). Most relevant to the present discussion, a harmonic learning bias in artificial language learning has been found for English-speaking adults (Culbertson et al., 2012) and children (Culbertson and Newport, 2015), as well as native speakers of cross-linguistically rare non-harmonic languages (Culbertson et al., 2020).

2.3 Counterfactual Language Paradigm

Artificial or counterfactual language learning has also been widely applied to LMs in recent years (Ravfogel et al., 2019; Hahn et al., 2020; White and Cotterell, 2021; Hopkins, 2022; Clark et al., 2023; Kallini et al., 2024; Kuribayashi et al., 2024). Whereas studies on human subjects are highly constrained by time, resources, and the limits of human attention, LMs can feasibly be trained extensively on artificial languages which can be highly complex, naturalistic, or formal. Accordingly, the design space for these types of studies is large and comes with numerous trade-offs. Specifically, we can distinguish the artificial language designs based on whether they take what we refer to as a **bottom-up** approach where a counterfactual corpus is generated from a manually specified lexicon and grammar; or a **top-down** approach where a naturalistic corpus is modified according to a set of rules.

At the extreme end of bottom-up approaches are studies that examine the learnability of different classes of formal languages for different neural network architectures, and therefore generate data potentially far outside the complexity class of natural language (Ebrahimi et al., 2020; DuSell and Chiang, 2022; Hao et al., 2022; Deletang et al., 2023; Borenstein et al., 2024; Someya et al., 2024). [2] A slightly more natural approach is to design and generate texts from probabilistic context-free grammars inspired by those of natural language but which can violate specific typological properties. These studies (White and Cotterell, 2021; Kuribayashi et al., 2024) have yielded diverging results on whether the inductive biases of LMs align with those of humans. However, bottom-up corpora massively simplify the problem of language learning and processing. Naturalistic data contains a depth of constructions, statistical patterns, and errors that cannot practically be generated using a bottom-up approach.

The top-down approach achieves greater ecological validity by taking as a starting point a corpus that includes all the complexity of natural data, and performing controlled manipulations, often using constituency or dependency parses of the data. One common approach uses parses to filter particular sentence types from a training corpus (Jumelet and Hupkes, 2018; Warstadt, 2022; Patil et al., 2024; Misra and Mahowald, 2024). Other work applies rules to parses to modify sentences. Ravfogel et al. (2019) use gold parses from the Penn Treebank (Marcus et al., 1993) to create counterfactual versions of English with different agreement marking systems and each of the six possible dominant orders of subject, object, and verb. Hahn et al. (2020) create counterfactual dependency grammars by specifying for each arc label whether the dependent goes to the left or right and how close to the head it is placed relative to its sisters. While this approach results in more ecologically valid counterfactual languages, it is also a noisy and difficult process to control. Messy source data, annotation errors, or limitations of linguistic annotation systems mean that counterfactual corpora have more ungrammatical content (relative to the counterfactual grammar) than the original corpus. Nonetheless, our study takes a top-down approach, while attempting to minimize and control for noise.

3 Experimental Design

Our experiments test whether LMs show differences in learning natural languages with harmonic word orders compared to minimally different artificial languages with non-harmonic word orders (implausible languages).

The Independent Variable: Harmonic and non-harmonic languages We manipulate word order harmony using a top-down approach to counterfactual corpus generation. We modify naturally occurring corpora for languages with harmonic word orders by violating five specific Greenbergian correlation pairs, one at a time (see §4). For each For each correlation pair, there are two types of harmonic languages (SVO with head-initial order, SOV with head-final order) represented by the natural corpora and two types of non-harmonic languages (SVO with head-final order, SOV with head-initial order) represented by counterfactual corpora.

The Dependent Variables: Measures for learnability There is no universally accepted definition or measure for learnability in the LM literature. In this study, we investigated the learnability of counter-Greenbergian languages based on the learning trajectory of the LMs as well as their final performance after a certain period of training. Given the concern that some counter-Greenbergian

[2] These empirical studies should be distinguished from theoretical studies that prove analytically which languages can be recognized by different architectures. See Strobl et al. (2024) for an overview of that line of work.

languages might eventually be *learnable* for humans, one would naturally hypothesize that these tendencies could exist due to other learning barriers, such as *learning efficiency*. Therefore, we observed the learning trajectory of the counterfactual LMs across their checkpoints. Details of our evaluation metrics and experimental results are shown in §5.

Addressing confounds: Symmetrical experimental design A key confound we try to avoid is that if we test on a fully head-initial language like English and make it head-final, the change in learnability can result from other factors than breaking the correlations, such as (a) models' learning biases towards the head direction of a language, or (b) the amount of noise we induced during counterfactual corpus generation. Our approach involves various ineliminable noise sources, including parser errors or ambiguities, punctuation removal prior to corpus editing, and the limitations of UD (universal dependencies) annotations.

We address (a) by conducting our experiments symmetrically with both a fully head-initial language and a fully head-final one. We address (b) by reporting human validation scores, identifying parser ambiguities, and creating BASELINE corpora variants that follow the same preprocessing steps of removing punctuation and lower-casing as applied to counterfactual corpora.

4 Creating Counterfactual Languages

This section describes our procedure for creating counterfactual corpora by modifying natural sentences top-down. Implementation details and examples are further provided in Appendix A.

4.1 Swapping Greenbergian Correlation Pairs

Notation We denote a correlation pair using the notation <*H, D*>, where *H* is the *Verb patterner* and is a mnemonic for *head*, and *D* is the *Object patterner* and is a mnemonic for *dependent*. We use this notation to name a type of correlation pair by its syntactic categories (e.g., <*Adp, NP*>) or to refer to a single instance of expressions belonging to the relevant categories (e.g., <*in, the house*>).

Targeted Correlation Pairs Table 1 summarizes the selected subset of Greenbergian correlation pairs identified by Dryer (1992) in our study. As shown in Tab. 3, we identify the five correlation pairs in dependency parses in the Universal Dependencies framework partly following Hahn et al. (2020). While dependency arcs are a good start for identifying instances of *H* or *D*, they only connect two words, not entire phrases, and there is no one-to-one or even many-to-one correspondence between Universal Dependencies arcs (De Marneffe et al., 2021) and Dryer's (1992) correlation pairs. For each language examined and each of the five correlation pairs, we implement a version of the swapping algorithm below to generate six distinct variants of a corpus with different word orders (Table 1).

Algorithm Overview The goal of our algorithm for creating counterfactual corpora is to swap the relative order of all instances of the relevant correlation pair within the input sentence. The word order is swapped at a span level. That is, given a sentence $\boldsymbol{w} = [w_1, ..., w_n]$ and its dependency parse p, a word pair (w_H, w_D) with a specific dependency type is first identified, and their spans $\boldsymbol{s}_H$ ($\boldsymbol{s}_D$) are determined as a continuous word sequence in $\boldsymbol{w}$ consisting of the identified word w_H (w_D) and its descendants[3] in the dependency structure, i.e., $\boldsymbol{s}_h = [w_i, \cdots, w_H, \cdots, w_j]$; here, $1 \leqslant i \leqslant H \leqslant j \leqslant n$, and w_H should be the head of the partial dependency structure of s_H. Then, the word order is swapped so that the relative position of $\boldsymbol{s}_h$ and $\boldsymbol{s}_d$ changes. All the pairs of tokens (spans) that meet the criteria of an <*H, D*>-pair in Table 3 are identified, and this span-swapping process is performed recursively (Algorithm 1 in Appendix A). Exceptions, additional conventions, and handling of coordination are covered in Appendix A.1.

Handling Multiple Pairs In the case that multiple dependent spans share the same head w_H in a sentence, we perform swapping by *reflecting* the dependents around w_H. In other words, we maintain the relative distance between *H* and *D*. In an abstractive example of swapping "*H* D_1 D_2", the swapped order becomes "D_2 D_1 *H*". In addition, since the dependency parse of a sentence exhibits a directed acyclic graph structure, and there might be nested correlation pairs, we perform a depth-first search over the sentence in our swapping algorithm (Algorithm 1).

[3] Strictly speaking, we use more different criteria to determine a word's span depending on the grammar of the language and the annotation, i.e. we do not always include all and only the descendants of the word.

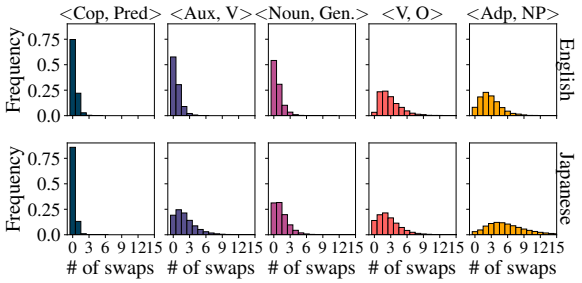


Figure 1: Histogram of the number of swaps per sentence for each counterfactual language.

	Train Data (En)			BLiMP	Train Data (Ja)			JBLiMP
Pair	Prec	Rec	Val	Val	Prec	Rec	Val	Val
<*Cop, P.*>	59.1	54.2	4.4	4.9	55.0	55.0	4.8	4.9
<*Aux, V*>	95.8	95.8	5.0	4.9	72.7	83.3	4.5	4.5
<*N., Gen.*>	80.0	80.0	4.8	5.0	81.0	81.0	4.8	4.9
<*V, O*>	74.4	73.4	4.3	4.6	85.9	81.6	4.2	4.3
<*Adp, NP*>	78.9	81.8	4.7	4.9	85.8	89.0	4.6	4.6

Table 2: Human validation results of counterfactual corpora. "Prec," "Rec," and "Val" denote precision, recall, and the averaged validation score indicated in the 5-point Likert scale.

Handling Japanese-Specific Issues Sometimes, a naive application of English implementation to Japanese does not work consistently due to differences in grammars and annotation conventions. For example, Japanese UD does not adhere to as rigid a notion of *word* as English UD. To make the swapping algorithm for both languages as similar yet correct as possible, we introduced some additional rules for the Japanese implementation (see Appendix A.3).

Statistics The frequency distributions of word-order swapping in a sentence for each correlation pair are shown in Fig. 1, which were estimated using a held-out set of LM training data (Wiki-40B). The total number of swaps from lowest to highest is <*Cop, Pred*>, <*Aux, V*>, <*Noun, Genitive*>, <*V, O*>, and <*Adp, NP*>. Henceforth, the experimental results are reported in this order to facilitate interpretation of the results.

4.2 Human Data Validation

We conduct a human validation of our counterfactual corpora at several stages to ensure the validity of our swapping algorithm and iteratively improve our swapping algorithms. Earlier iterations of validation were less formal, and resulted in changes to the swapping algorithm. Below we describe the validation of our final counterfactual corpora. While the swapping algorithm is not perfect, we believe that transparency about these flaws is an improvement over previous studies on top-down counterfactual corpora, none of which report any metric to evaluate the quality of their counterfactual corpora (Ravfogel et al., 2019; Hahn et al., 2020; Clark et al., 2023).

Quantitative Evaluation Annotators manually list all <*H, D*> pairs that should be swapped for that sentence, according to their judgment. They compare this **gold** list to the **silver** list of all <*H, D*> pairs identified by the parser and swapped by the algorithm. We then compute the precision of the silver swaps ($\frac{\#\text{correct silver}}{\#\text{silver}}$) and the recall ($\frac{\#\text{correct silver}}{\#\text{gold}}$) over the entire annotated sentences.

Qualitative Evaluation Annotators also subjectively assess the validity of each swapped sentence using a 5-point Likert scale (see Appendix C). This additional evaluation is motivated for several reasons: First, the quantitative evaluation unjustifiably favors mistakes that fail to identify a pair (which affects only recall) over mistakes where a silver pair is similar but not an exact match to a gold pair (which harms precision and recall). Second, the silver string may sometimes be correct even if the identified pairs are not, i.e., some pairs are truly subjective due to ambiguity in the sentence or inevitable underspecificity in our annotation guidelines. Third, errors can cascade, i.e., a single incorrect arc can lead to two (or more) errors arising from the words incorrectly connected and the words incorrectly *not* connected. Finally, some errors are intuitively less divergent from the counterfactual target (e.g., incorrectly resolving a prepositional phrase attachment) than others (e.g., misparsing a verb as a noun).

Annotators and Data One English native speaker and two Japanese native speakers annotated the gold word swap and the validity score for each sentence (each example was assessed by one annotator). The annotators are all authors on the paper with PhD-level training in linguistics. Our validation is mainly made on the training data for LMs (see §5), but we also conducted the qualitative evaluation part on sentences sampled from BLiMP/JBLiMP benchmarks, which are used in our LM evaluations §6.3. We sampled 120 sentences for <*V, O*> and 40 sentences for the other correlation pairs for annotation, respectively, from the respective data sources, and thus 280 sentences

are of validation target in each evaluation setting (e.g., English/Japanese LM training data).[4] Notably, these validation targets include sentences without any target of respective swapping to properly estimate the precision of the algorithm.[5]

Results Table 2 shows the results. The precision and recall of the word-sapping are typically above or near 80%, and the average validity score on a 5-point scale is above 4. Thus, we conclude that our word-swapping algorithm properly worked in most cases. In addition, the 5-Likert scale scores are generally similar between LM training data and (J)BLiMP; thus, there are no issues specifically associated with the (J)BLiMP datasets, which include more complex or rare linguistic phenomena. Though the swapping precision/recall for the <*Cop, Pred*> part was particularly low, the validity scores are high. This is due to frequent minor errors, typically in identifying the scope of the predicate in the copula construction. For example, our algorithm converted a sentence "*he was active in the rsp student wing.*" into "*he active was in the rsp student wing.*," while human annotation was "*he active in the rsp student wing was.*"

5 Model Training

Language Modeling To assess the inductive bias of both causal LMs and masked LMs, we duplicate our experiments with both GPT-2 small (Radford et al., 2019) and LTG-BERT (Samuel et al., 2023) architectures.[6] All models are trained for 12 epochs from scratch, and we examined three different random seeds for each setting. Appendix D shows additional training details.

Data We choose English and Japanese to perform our symmetrical (head initial/final→final/initial) experiments. Train, validation, and test splits consist of 100M words, 10M words, and 1M words, respectively. Token numbers are counted based on whitespace in English and MeCab (Kudo, 2005) with the ipadic dictionary in Japanese, respectively. These sentences are sampled from the English and Japanese parts of the Wiki-40B dataset (Guo et al., 2020). We choose Wikipedia data as the domain is similar to the data that the UD parsers were trained on, and thus we expect the resulting counterfactual corpora to be more accurate than would result from more developmentally plausible data such as child-directed speech.

We use Stanza to obtain dependency parses for every sentence in the corpora. To avoid erroneous swapping, we removed (i) all punctuations from English and Japanese sentences; (ii) brackets (with their inside content) from Japanese sentences, i.e., typically rubi for Japanese Kanji; and (iii) sentences with lower-cased English words from the Japanese corpus. We set two baseline models: (i) an ORIGINAL model that is trained on our 100M Wiki-40B dataset without any preprocessing or swapping, and (ii) a BASELINE model that is trained on the corpus with the preprocessing but without any swapping. Comparisons between the ORIGINAL and BASELINE models function as a check for any unintended biases from our preprocessing. Comparisons between BASELINE and the other counterfactual LMs are of primary interest in how much counterfactual word order hurts language learning.

6 Results

6.1 Evaluation 1: Perplexity

Results We first compare the perplexities (PPLs) on the held-out data achieved by the LMs in each language, including counterfactual ones (Figure 2).[7] In the final epoch, the counterfactual LMs achieved similar PPL scores to the BASELINE LMs. However, if we look at the entire learning trajectory, learning appears to be slower for the counterfactual languages. Note that the ORIGINAL LMs also achieved PPL slightly better than BASELINE but at an approximately similar scale; our preprocessing did not drastically change the language modeling task difficulty. The <*V, O*> variants tend to have slightly worse PPLs compared to BASELINE and other counterfactual languages, which might be due to the fact that <*V, O*> corpora have a large

[4] We annotated an especially large number of sentences for the <*V, O*> swap since it induced more diverse changes than the other correlation pairs.

[5] When sampling LM training data to annotate, we balanced the data in each correlation pair to have 20 sentences with no silver swaps to better estimate the precision of the algorithm. Reported precision and recall reflect the distribution in the overall corpus, not the balanced sample.

[6] LTG-BERT is a masked LM which resembles DeBERTa (He et al., 2021) with some additional optimizations. We choose this architecture as it is the basis for the model that won the BabyLM Challenge, a competition on data-efficient pretraining (Warstadt et al., 2023).

[7] We report PPL per character for the Japanese results. This is necessary because the change in word order in different Japanese variants results in different token lengths due to the lack of whitespace word boundaries in Japanese.

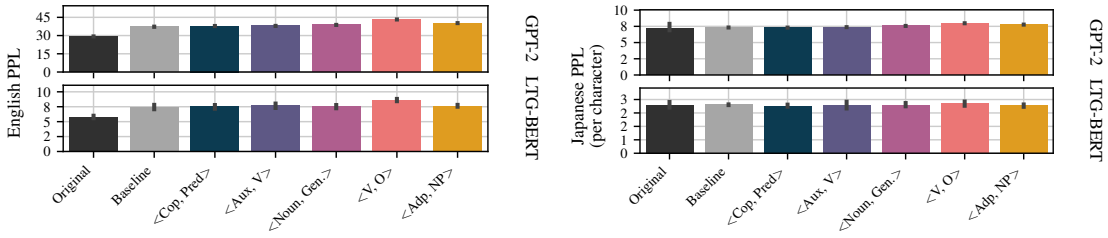


Figure 2: Final PPLs on respective heldout data for English-based counterfactual LMs (left) and Japanese-based counterfactual LMs (right). Error bars indicate standard deviation over three random seeds.

number of syntactically complex swaps (Figure 1) and relatively worse swapping validity according to our human annotation (Table 2). Thus, the performance of LMs might plausibly reflect noise in the corpus as well as the difficulty of the (intended) grammar.

Statistical Tests We perform a paired Wilcoxon signed-rank test for each correlation pair in each source language by comparing six PPL scores of $\{\text{GPT2}, \text{LTG-BERT}\} \times \{3 \text{ seeds}\}$ from the corresponding counterfactual models and those from BASELINE models. Out of the ten settings of $\{\text{En}, \text{Ja}\} \times \{5 \text{ word orders}\}$, only one setting of English *<V, O>* showed that the baseline model (real English) is significantly easier to learn ($p = 0.03 < 0.05$) than the counterfactual one. However, if we extend this analysis into learning trajectories, the statistical tests between 72 PPLs of $\{\text{GPT2}, \text{LTG-BERT}\} \times \{3 \text{ seeds}\} \times \{12 \text{ epochs}\}$ can be made, and this yields that the real English is significantly easier to learn than the counterfactual one in all the five correlation pairs ($p < 0.05$), while the real Japanese is *not* significantly easier to learn than the counterfactual one in all the five correlation pairs ($p > 0.05$).

6.2 Evaluation 2: Minimal Pair Preferences

Settings The previous evaluation measures PPL on all the tokens in the corpus; some of them are not necessarily related to our targeted word order change. For a more targeted evaluation of the learnability of counterfactual Greenbergian word ordering, we design a binary task requiring selecting the right word order given two sentences containing at least one instance of a relevant correlation pair, differing only in whether the order of the elements in each pair is correct. The task design is symmetrical between counterfactual and BASELINE LMs; the correct option follows the counterfactual word order when evaluating counterfactual LMs, and vice-versa for the BASELINE LMs. To assess word order preference we compare the predicted probability (i.e., accumulated surprisal) of each sentence; that is, the option with a higher probability, i.e., lower surprisal, is regarded as preferred by LMs. We report the accuracy in the binary task of selecting the correct word order. The sentences were sampled from the held-out set of Wiki-40B data.

Results Figure 3 shows the trajectory of accuracy during LM training. All the counterfactual LMs prefer the correct word order over the incorrect one much more than random chance (accuracy of 0.5), which leads to our conclusion that LMs generalized well to counter-Greenbergian languages and learned the counterfactual ordering pattern successfully. Nevertheless, in many settings, the BASELINE LMs yielded higher accuracies than the counterfactual ones; thus, at least through the lens of this experiment, the real languages are usually easier to learn their word order for LMs. However, this does not always appear to be the case as some counterfactual languages exhibit almost the same accuracies as the corresponding BASELINE LMs, specifically for GPT-2.

Statistical Tests We perform a paired Wilcoxon signed-rank test for each correlation pair in each source language by comparing 72 accuracy scores of $\{\text{GPT2}, \text{LTG-BERT}\} \times \{3 \text{ seeds}\} \times \{12 \text{ epochs}\}$ from the corresponding counterfactual models and those from BASELINE models. In all the ten settings of $\{\text{En}, \text{Ja}\} \times \{5 \text{ word order}\}$, the BASELINE LMs exhibited significantly higher accuracies than the counterfactual LMs ($p < 0.05$; in eight settings $p <$1e-12).

6.3 Evaluation 3: BLiMP & JBLiMP

Settings In addition to the minimal pair preference on Wiki-40B sentences (§6.2), we further evaluate LMs on specific linguistic phenomena, ranging over morphology, syntax, and semantics,

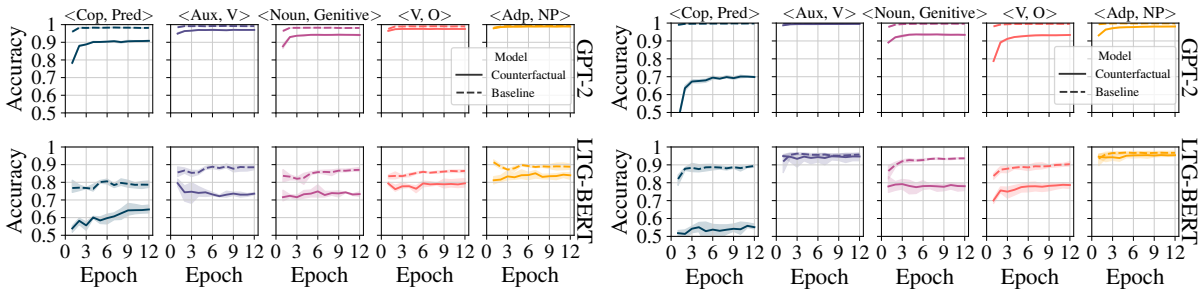


Figure 3: Performance trajectories for minimal pair comparisons targeting the counterfactual word order for counterfactual models and natural order for baseline model, for English-based LMs (left) and Japanese-based LMs (right). Shaded areas present standard deviation (SD) over three random seeds.

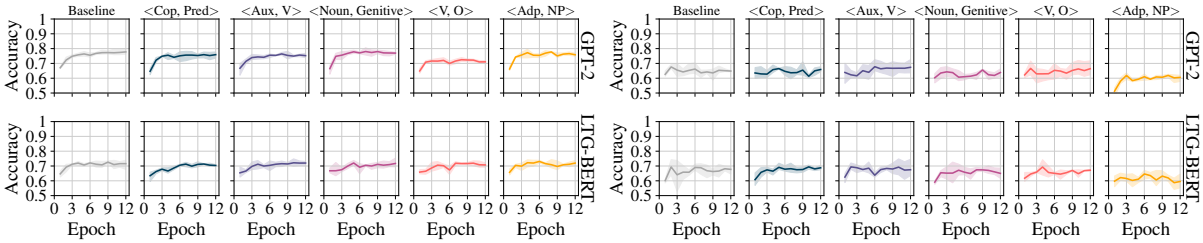


Figure 4: Performance trajectories of English-based counterfactual LMs in BLiMP (left) and Japanese-based counterfactual LMs in JBLiMP (right). Shaded areas present SDs over three random seeds.

again using the minimal pair paradigm. This evaluation tells us whether counterfactual word order has negative impacts on learning specific grammar rules not necessarily related to the swapped rule. Specifically, we test LMs on a downsampled[8] version of BLiMP (Warstadt et al., 2020) and JBLiMP (Someya and Oseki, 2023) benchmarks of minimal pairs for English and Japanese experiments, respectively. For each counterfactual language, we also create a respective counterfactual version of BLiMP and JBLiMP by applying the same word-order swapping algorithm (§4) to them. Thus, each example in counterfactual (J)BLiMP consists of a pair of grammatically correct and incorrect sentences *in the counterfactual language space*. Notably, as demonstrated in §4.2, the accuracy of the word-order swapping algorithm was generally good even in BLiMP/JBLiMP datasets; this alleviates (but does not fully eliminate) the potential concern that these counterfactual versions of benchmarks are too noisy to estimate the model's linguistic knowledge.

Results We report the macro average of accuracy over the 12 BLiMP suites (or 9 JBLiMP suites). Figure 4 shows the performance trajectory of LMs during training. The BASELINE trajectories are relatively similar or slightly better than those from counterfactual LMs, suggesting that counterfactual word order not drastically but slightly prevented LMs from acquiring grammatical knowledge.

Statistical Tests We performed a paired Wilcoxon signed-rank test for each correlation pair in each source language by comparing 864 accuracy scores of $\{\text{GPT2}, \text{LTG-BERT}\} \times \{3 \text{ seeds}\} \times \{12 \text{ epochs}\} \times \{12 \text{ BLiMP categories}\}$ from the corresponding counterfactual models and those from BASELINE models. In six of the ten settings of $\{\text{En}, \text{Ja}\} \times \{5 \text{ word orders}\}$, BASELINE LMs exhibited significantly higher BLiMP accuracies than the counterfactual ones ($p < 0.05$).[9]

7 Discussion and Conclusions

Our findings show that autoregressive and masked LMs have a consistent learning bias—with some notable exceptions—favoring harmonic languages over the nonharmonic counterfactual languages we examined. Strikingly, the experimental results

[8] We randomly sample 5 examples from each of the 67 BLiMP circuits, combine them into 12 BLiMP categories, and calculate the macro average accuracy over 12 categories.

[9] If we apply the Bonferroni correction, given that we performed statistical tests 30 times through our three experiments, the results could be more conservative, where baseline language yielded significantly higher BLiMP accuracies than four counterfactual languages ($p < 0.0016 = 0.05/30$).

from Section 6.2 show that, for sentences involving the modified grammar rule, learning trajectories of the counterfactual languages lag behind those of the original language for every counterfactual language, model, and source language we examine. The evaluations measuring PPL §6.1 and (J)BLiMP performance §6.3 are more mixed, with counterfactual languages showing significantly worse performance across training only about half the time.

While these conclusions about LMs' learning biases are interesting in their own right, their implications for debates about linguistic typology are particularly important. The role of modern LMs in linguistics and cognitive science has been a topic of much discussion and controversy (Pater, 2019; Linzen, 2019; Baroni, 2022; Warstadt and Bowman, 2022; Lan et al., 2024; Wilcox et al., 2023; Piantadosi, 2023; Katzir, 2023; Millière; McGrath et al., 2024). Here, we will not rehearse all the details of this debate, but present a condensed account of how our experiments on LMs can inform ongoing debates about human language:

Chomsky et al. (2023) publicized aspects of this debate by claiming LMs learn impossible languages, and consequently have limited relevance to the study of human language. Kallini et al. (2024) empirically test the first part of this claim in detail, showing that LMs display relative difficulty acquiring counterfactual versions of English with rules involving highly unnatural operations such as reversing strings and counting. Our study furthers Kallini et al.'s conclusions by showing that LMs continue to show a learning bias for typologically dispreferred counterfactual languages closer to the boundary between plausible and implausible. However, we also take issue against the second part of Chomsky et al.'s argument.[10]

We contend that LMs, can help answer two questions about linguistic typology and acquisition, regardless (in some cases) of whether they show human-like biases. First is the question of whether humans actually have a learning bias for harmonic languages. As discussed in §2.2, the evidence in support of this conclusion from human subjects is limited somewhat due to the small scale and simplicity of the artificial languages employed. Although LMs come with other limitations, the top-down approach to counterfactual language creation allows for naturalistic complexity and scale in the training data, providing a complementary line of evidence. Our observation of a harmonic learning bias in LMs is powerful converging evidence adding to evidence from human studies that a learning bias for harmonic languages is real. Given the LMs show this bias as well, there is less reason to doubt similar findings regarding humans.

Second is the question of whether a harmonic learning bias in humans is due to language-specific or domain-general priors. The argument here is similar to that in several prior works (Clark and Lappin, 2011; Warstadt and Bowman, 2022; Wilcox et al., 2023; Constantinescu et al., 2024; Kuribayashi et al., 2024): The Transformer architecture on which modern LMs are based (Vaswani et al., 2017) is not specifically designed for language but is highly effective for domains as far reach as language, vision (Dosovitskiy et al., 2021), and protein sequences (Jumper et al., 2021), suggesting that it relies on domain-general learning biases. Thus, if one accepts that Transformes do show a harmonic bias, it follows that language-specific biases are not necessary to observe this phenomenon at least to some degree, and that should increase our credence in an explanation in terms of externally motivated domain-general biases in humans. Furthermore, previous findings of harmonic bias in humans (e.g., Culbertson et al., 2012) might not be construed as evidence for language-specific bias in humans.

Importantly, evidence from this kind of experiment is relevant regardless of the result. If we had found that Transformers did not show a harmonic bias, it would follow that such a bias is not a necessary consequence of the domain-general biases sufficient for language learning (at least assuming they learned the counterfactual languages successfully). While it would still be possible in this counterfactual scenario that Transformers lack the relevant domain-general bias, we would nonetheless have increased our credence that humans might have some idiosyncratic learning mechanism which may well be language-specific.

It bears mentioning that even if humans do have a harmonic learning bias, other factors may still be equally if not more important to explain typological correlations. Communicative pressures are another mechanism that might explain these phenomena, and extending our methods to test this mechanism is a promising avenue for future work.

[10] Kallini et al. do claim that evidence from LMs is relevant to questions about the innate priors required for language learning (p. 14699) but do not fully spell out the argument that we give below.

Hahn et al. (2020) and Clark et al. (2023) have both found that counterfactual languages perform worse than natural languages on measures of communicative efficiency, such as dependency length and uniformity of information density. These measures can be straightforwardly applied to our counterfactual corpora, which employ both more targeted and syntactically informed manipulations than in those previous works.

We must acknowledge an important limitation that tempers the force of our conclusions: Our manual validation shows that even our relatively careful approach to counterfactual language construction leads to numerous errors arising from parser errors. Thus, it is possible that our findings may be due partially or entirely to increased noise in the counterfactual corpora, rather than inherent differences in learnability between the original and counterfactual grammars. One defense against this unsatisfying conclusion is that on the PPL evaluation the final performance of counterfactual and original LMs are mostly not significantly different, suggesting that in the limit, the counterfactual languages are largely as predictable as the originals. While it is true that the languages with the most noise according to our validity annotations, the <*V,O*> languages, show the highest PPL, this pattern does not apply across other counterfactual languages. We leave it to future work to explore alternative methods to reduce noise in naturalistic counterfactual corpora or to control for the amount of noise introduced by different forms of data manipulation.

Finally, while our study is a step forward in testing the learnability of counterfactual languages, it still leaves open many questions and avenues for future work. Our conclusions are based only on two languages, so it will be important to try to replicate these results with more SVO and SOV languages, and also on languages with inconsistent VO ordering, such as German, though this direction will require input from many domain specialists and native-speaker linguists. Future work should also study a wider variety of models as well as train models on more developmentally plausible data, such as dialogue data and child-directed speech.

To conclude, the rise of effective and efficiently trainable Transformer LMs has created the possibility of investigating the learnability of counterfactual languages at a scale and level of naturalism not possible with human subjects. Through our emphasis on a syntactically sophisticated top-down approach to counterfactual language construction and the release of our code and models, we hope our work inspires further exploration of the diverse space of possible languages and deepens our understanding of the particular subspace that human languages occupy.

References

Marco Baroni. 2022. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. In Shalom Lappin and Jean Philippe Bernardy, editors, *Algebraic Structures in Natural Language*, pages 1–16. CRC Press.

Jon Barwise and Robin Cooper. 1988. Generalized quantifiers and natural language. In Jack Kulas, James H. Fetzer, and Terry L. Rankin, editors, *Philosophy, language, and artificial intelligence: Resources for processing natural language*, pages 241–301. Springer Netherlands, Dordrecht.

Nadav Borenstein, Anej Svete, Robin Chan, Josef Valvoda, Franz Nowak, Isabelle Augenstein, Eleanor Chodroff, and Ryan Cotterell. 2024. What languages are easy to language-model? A perspective from learning probabilistic regular languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15115–15134, Bangkok, Thailand. Association for Computational Linguistics.

Nick Chater and Paul Vitányi. 2003. Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1):19–22.

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

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Noam Chomsky and Howard Lasnik. 1993. The theory of principles and parameters. In *Syntax: An International Handbook of Contemporary Research*. Walter de Gruyter.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam Chomsky: The False Promise of ChatGPT. *The New York Times*.

Alexander Clark and Shalom Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for uniform information density in word order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2024. Investigating critical period effects in language acquisition through neural language models.

Ailís Cournane. 2017. In defence of the child innovator. In Eric Mathieu and Robert Truswell, editors, *Micro-change and macro-change in diachronic syntax*, pages 10–36. Oxford University Press.

Jennifer Culbertson, Julie Franck, Guillaume Braquet, Magda Barrera Navarro, and Inbal Arnon. 2020. A learning bias for word order harmony: Evidence from speakers of non-harmonic languages. *Cognition*, 204:104392.

Jennifer Culbertson and Simon Kirby. 2016. Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, 6.

Jennifer Culbertson and Elissa L. Newport. 2015. Harmonic biases in child learners: In support of language universals. *Cognition*, 139:71–82.

Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. 2012. Learning biases predict a word order universal. *Cognition*, 122(3):306–329.

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, pages 1–54.

Gregoire Deletang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A Ortega. 2023. Neural networks and the chomsky hierarchy. In *The Eleventh International Conference on Learning Representations*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Matthew Dryer. 1992. The Greenbergian word order correlations. *Language*, 68:138 – 81.

Brian DuSell and David Chiang. 2022. Learning hierarchical structures with differentiable nondeterministic stacks. In *International Conference on Learning Representations*.

Javid Ebrahimi, Dhruv Gelda, and Wei Zhang. 2020. How can self-attention networks recognize Dyck-n languages? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4301–4306, Online. Association for Computational Linguistics.

Maryia Fedzechkina, T. Florian Jaeger, and Elissa L. Newport. 2012. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44):17897–17902.

Riki Fujihara, Tatsuki Kuribayashi, Kaori Abe, Ryoko Tokuhisa, and Kentaro Inui. 2022. Topicalization in language models: A case study on Japanese. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 851–862, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joseph H Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the Twelfth 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, 117(5):2347–2353.

Yiding Hao, Dana Angluin, and Robert Frank. 2022. Formal language recognition by hard attention transformers: Perspectives from circuit complexity. *Transactions of the Association for Computational Linguistics*, 10:800–810.

John A. Hawkins. 1983. *Word order universals*. Academic Press, San Diego.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding enhanced BERT with disentangled attention. In *International conference on learning representations*.

Mark Hopkins. 2022. Towards More Natural Artificial Languages. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 85–94, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Anne S. Hsu, Nick Chater, and Paul Vitányi. 2013. Language Learning From Positive Evidence, Reconsidered: A Simplicity-Based Approach. *Topics in Cognitive Science*, 5(1):35–55.

Larry M. Hyman. 2008. Universals in phonology. *The Linguistic Review*, 25(1-2):83–137.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin ídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Carla L. Hudson Kam and Elissa L. Newport. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2):151–195.

Roni Katzir. 2023. Why Large Language Models Are Poor Theories of Human Linguistic Cognition: A Reply to Piantadosi. *Biolinguistics*, 17(e13153).

Richard S. Kayne. 1994. *The antisymmetry of syntax*, volume 25. MIT press.

Simon Kirby, Hannah Cornish, and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Takumitsu Kudo. 2005. Mecab : Yet another part-of-speech and morphological analyzer.

Tatsuki Kuribayashi, Ryo Ueda, Ryo Yoshida, Yohei Oseki, Ted Briscoe, and Timothy Baldwin. 2024. Emergent word order universals from cognitively-motivated language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14522–14543, Bangkok, Thailand. Association for Computational Linguistics.

Nur Lan, Emmanuel Chemla, and Roni Katzir. 2024. Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, pages 1–28.

Tal Linzen. 2019. What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, 95(1):e99–e108.

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

Sam Whitman McGrath, Jacob Russin, Ellie Pavlick, and Roman Feiman. 2024. How can deep neural networks inform theory in psychological science? *Current Directions in Psychological Science*, 33(5):325–333.

Raphaël Millière. Language Models as Models of Language. In R. Nefdt, G. Dupre, and K. Stanton, editors, *The Oxford Handbook of the Philosophy of Linguistics*. Oxford University Press.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Jeff Mitchell and Jeffrey Bowers. 2020. Priorless recurrent networks learn curiously. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5147–5158, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yugo Murawaki. 2019. On the definition of Japanese word.

Frederick J. Newmayer. 2005. *Possible and probable languages: a generative perspective on linguistic typology*. Oxford University Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Hisashi Noda. 1996. *Wa to ga (Wa and ga)*. Kurosio Publishers.

Mai Omura, Aya Wakasa, and Masayuki Asahara. 2021. Word delimitation issues in UD Japanese. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 142–150, Sofia, Bulgaria. Association for Computational Linguistics.

Joe Pater. 2019. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 95(1):e41–e74.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615.

Alain Peyraube. 1912. L'évolution des formes grammaticales. *Scientia; rivista di scienza*, 6(12):384.

Steven T Piantadosi. 2023. Modern language models refute Chomsky' s approach to language. In Edward Gibson and Moshe Poliak, editors, *From fieldwork to linguistic theory: A tribute to Dan Everett*, volume 15 of *Empirically Oriented Theoretical Morphology and Syntax*, pages 353–414. Language Science Press, Berlin.

Gregory Pringle. 2016. Thoughts on the Universal Dependencies proposal for Japanese: The problem of the word as a linguistic unit.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Taiga Someya, Ryo Yoshida, and Yohei Oseki. 2024. Targeted syntactic evaluation on the Chomsky hierarchy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15595–15605, Torino, Italia. ELRA and ICCL.

Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. 2024. What formal languages can transformers express? A survey. *Transactions of the Association for Computational Linguistics*, 12:543–561.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. 2016. Universal Dependencies for Japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1651–1658, Portorož, Slovenia. European Language Resources Association (ELRA).

Kazuhiro Teruya. 2004. Metafunctional profile of the grammar of Japanese. *Language typology. A functional perspective*, pages 185–254.

Kazuhiro Teruya. 2007. *A systemic functional grammar of Japanese*. Bloomsbury Academic.

Natsuko Tsujimura. 2013. *An introduction to Japanese linguistics*. John Wiley & Sons.

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.

Alex Warstadt. 2022. *Artificial neural networks as models of human language acquisition*. PhD Thesis, New York University.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In Shalom Lappin and Jean-Philippe Bernardy, editors, *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM challenge at the 27th conference on computational natural language learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

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.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44.

Colin Wilson. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, 30(5):945–982.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Implementation Details

A.1 English Policies

Here we provide the implementation details of the swapping algorithm for each correlation pair in the case of English experiments. Unless otherwise specified in the next subsection, the same policy as English is adopted for Japanese. Generally speaking, we identify correlation pair instances using the dependency arcs in Tab. 3. However, there are numerous exceptions which we discuss below.

H	UD Relation	*D*
verb	$\xrightarrow{obj}$, $\xrightarrow{iobj}$, $\xrightarrow{obl}$ $\xrightarrow{cop*}$, $\xrightarrow{ccomp}$, $\xrightarrow{xcomp}$	object
adposition	$\xleftarrow{case}$	NP
copula verb	$\xrightarrow{cop*}$	predicate
auxiliary	$\xleftarrow{aux}$	VP
noun	$\xrightarrow{nmod}$	genitive

Table 3: Word orders of interest in Greenbergian correlation pairs and their associated Universal Dependencies, adopted mostly from Hahn et al. (2020). The asterisked *cop** is originally UD (universal dependencies) label *cop* that we changed direction (lifted) during preprocessing, according to linguistic conventions.

Verbs and Objects We construe the <*V, O*> correlation more broadly to refer to a verb on the one hand and its arguments and phrasal modifiers on the other. In linguistic theory, there is no univerally agreed upon test for this notion of objectood. To obtain a usable boundary for objects when swapping *verb* and *object*, we established five different selection criteria that identify objects with verbs based on their levels of connection, depicted in Figure 5. Each of the five criteria corresponds to a boundary, ranging from *very tight* to *very loose*, and we adopt the "loose" boundary for objects in our implementation. Under this boundary, we treat all direct and indirect objects, prepositional objects, complement clauses and complement verb phrases, prepositional phrase adverbials, and nonfinite adverbial clauses of a verb as objects in our implementation.

Mapping these linguistic constituents to UD relations, we use `obj`, `iobj`, `obl`, `cop`, `expl`, `xcomp` and `ccomp` as dependency arc labels to identify the <*V, O*> pair. Since the `ccomp` arc corresponds to both finite & non-finite adverbial clauses, our approach depends on identifying an `nsubj` arc

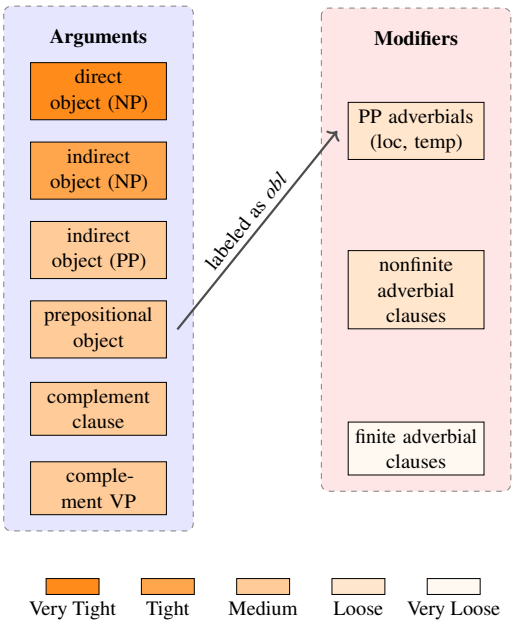


Figure 5: Illustration of different levels of tightness when classifying verbal dependents as objects.

linked to the clause's main verb to differentiate between finite and non-finite adverbial clauses. We acknowledge that using the presence of a subject as the distinguishing factor might not be the best practice, given that the distinction between these clause types does not solely depend on having a subject, but it is an effective heuristic for most cases.

Adpositions and Noun Phrases POS tags NOUN, PROPN, NUM, PRON for noun phrases and the UD arc label `case` identify the adposition and noun phrase word spans. For compound adpositions, such as "in front of", we identify multiple `case` arcs one by one and swap accordingly.

Copula and Predicate The correlation pair <*Cop, Pred*> is also included in <*V, O*> pair in our formalization. In UD, the predicate is considered the head of the `cop` arc and all VP modifiers. Following conventions in English syntax, we reverse the direction of the `cop` arc, making the copula the head of the predicate during preprocessing and transferring the VP modifiers to it before identifying both word spans using the `cop*`.

Auxiliary and Verb The <*Aux, V*> pair is identified by UD relation `aux`. We choose the associated verb phrase instead of a single verb for the word span of *V* following conventions in English syntax.

Noun and Genitive The <*Noun, Genitive*> pair is identified by UD relation `nmod`. In English, how-

ever, possessive nominal modifiers are also labeled with nmod, such as *John's book*, contrasting with *book of John*. Thus we include an additional condition on the existence of "of" between a noun and its nominal dependents to identify genitives and exclude possessives.

To identify the span associated with the *Noun*, we select all children preceding the *Noun* and connected by *nummod, compound, appos* and *flat*, and all children between the *Noun* and the genitive. This choice is a heuristic developed through trial and error across several stages of annotation.

A.2 Handling Coordination

We also adopt a set of conventions regarding cases of coordination, illustrated in the table below using the correlation pair of <*V,O*> as an example.

The first pair of rows illustrates cases where there is coordination of two dependents, which share a single head. In such cases, we treat the pair of dependents plus the conjunction as a chunk that is swapped with the head.

The second pair of rows illustrates cases where there two head–dependent pairs are coordinated. The dependency parse will have a conj arc between the two heads, and each head will have its own dependents. In such cases, we perform swapping for each head–dependent pair separately.

In the final two pairs of rows, we have two heads coordinated, with the second one having a dependent. Importantly, in the first of these pairs, the dependent is shared by both heads, while in the second, the dependent belongs only to the second head. Unfortunately, both sentences will receive the same dependency graph, so it is impossible to distinguish between these two cases. We adopt the convention that the two heads are treated as a chunk when swapping with the dependent, although this inevitably leads to incorrect swaps in cases like last example below.

Constructions	Examples
$H\ D_1$ conj D_2 D_1 conj $D_2\ H$	we are students and teachers we students and teachers are
$H_1\ D_1$ conj $H_2\ D_2$ $D_1\ H_1$ conj $D_2\ H_2$	we like cats and love dogs we cats like and dogs love
H_1 conj $H_2\ D$ $D\ H_1$ conj H_2	we sing and dance in the park we in the park sing and dance
H_1 conj $H_2\ D$ $D\ H_1$ conj H_2	we dance and play tag we tag dance and play

A.3 Japanese-Specific Treatments

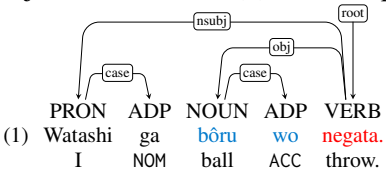
Table 4 shows examples of counterfactual variants of a Japanese sentence. The following paragraphs explain some treatments employed in modifying each word-order correlation pair in the Japanese language.

Verbs and Objects The Japanese language has a flexible word order, and the grammatical case of arguments is marked with a special marker rather than its word order (Tsujimura, 2013). However, these particles are sometimes omitted or overwritten by other particles, such as "wa" (topicalization marker; TOP) or "mo" (*also*), making the grammatical relationships ambiguous superficially and leading to erroneous parser outputs. To handle such errors, we employed several heuristic rules on top of the parser output to improve the accuracy and consistency of the word-order swapping algorithm:

- If a word has a nsubj dependency AND the nominative case marker "ga," the word is treated as a subject (i.e., the word is not swapped).
- If a word has a topicalization marker "wa," the word is not swapped.
- The other arguments with the nsubj, obj, iobj, obl, cop, expl, xcomp dependency are treated as an object (i.e., the word order is swapped).

That is, unless an argument is explicitly marked as a subject or marked topic, it is regarded as an object, which is compatible with the loose definition of object employed in the English experiment.

The second rule regarding the topicalization marker "wa" handles the topicalization phenomena. Note that the Japanese language is topic-prominent (Noda, 1996; Teruya, 2004, 2007; Fujihara et al., 2022), and a certain component of a sentence is frequently topicalized (i.e., moved to the initial part of the sentence with a special topicalization marker TOP). For example, either the subject or object of a sentence (1) can be topicalized:



The subject is topicalized in sentence (2), and the object is topicalized in sentence (3):

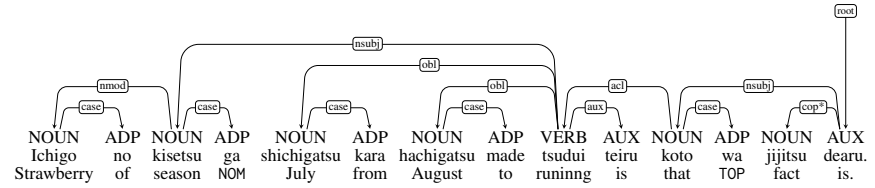
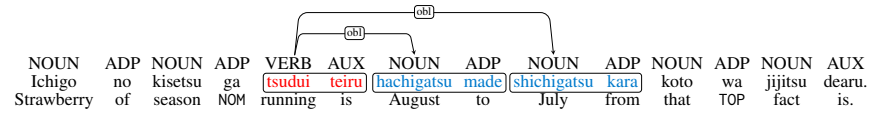
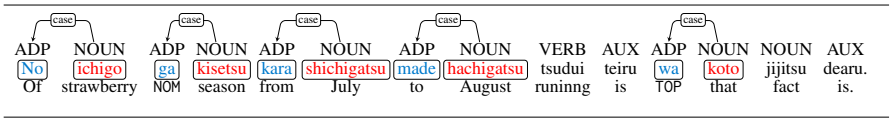
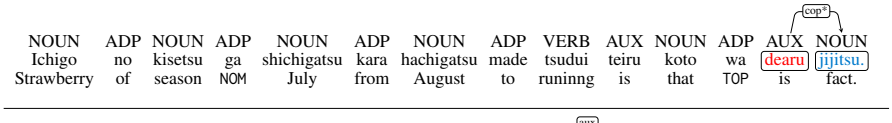
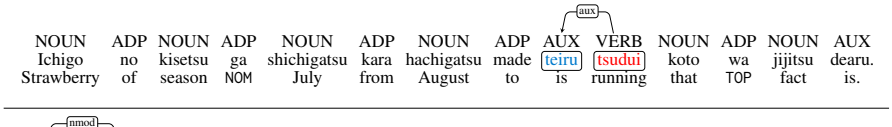
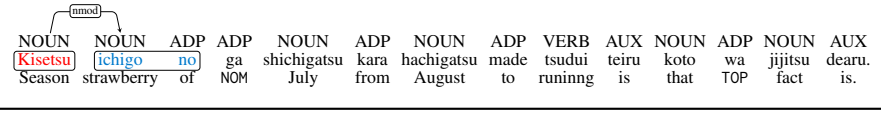
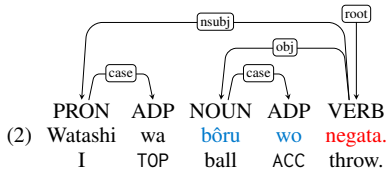
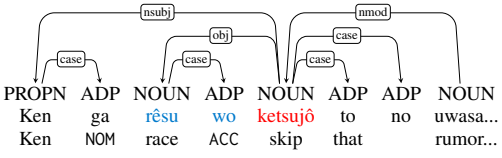
Correlation Pair	Example
Original	
<V, O>	
<Adp, NP>	
<Cop, Pred>	
<Aux, V>	
<Noun, Genitive>	

Table 4: Counterfactual examples from our variants of the Japanese language. The word span of *verb patterner* is colored red, and the word span of *object patterner* is colored blue. In the *<V, O>* example, we omit the swapping regarding the cop dependency for the purpose of explanation and brevity. The *<V, O>* example demonstrates the reflective swapping ($H\ D_1\ D_2 \rightarrow D_2\ D_1\ H$) mentioned in §4.1.



The topicalized component is typically ambiguous in terms of its grammatical case, and thus, the parser outputs were erroneous. Such a *marked* word order is beyond our interest since the Greenbergial correlations are generally on the canonical, *unmarked* word order of language. Thus, we did not modify the word order of such an explicitly topicalized word, even if it is seemingly an object of a verb. For example, the topicalized object, "Bôru wa" in Example (3), is no longer the target of *<V, O>* swapping.

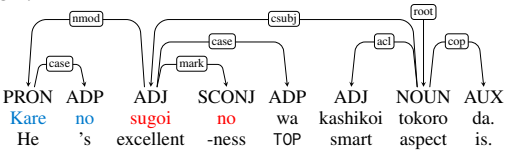
Another Japanese-specific concern is on a particular type of noun, called *sa-hen* noun, which can behave as a verb with a special conjugation verb "suru," e.g., "yôyaku" (NOUN)→"yôyaku-suru" (VERB), like the English words"summary" (NOUN)→"summar-ize" (VERB). However, the conjugation verb "suru" is sometimes omitted even when the sa-hen noun is used as a verb. Such nouns are typically annotated as NOUN with objects in the Japanese UD:



Here, "ketsujô" (*skip*) is annotated as a NOUN but can be regarded as a VERB, and the native Japanese validator indeed pointed out this should be included in the verb-object pairs. Thus, we regarded *sa-hen* nouns with either `nsubj, obj, iobj, obl, cop, expl, xcomp` dependent as verbs even when there is no conjugation verb. With this rule, in the above example, "ketsujô" is treated as a verb, and thus

the position of its object "rêsu-wo" (*race*-ACC) will be changed by the <*V, O*> swapping algorithm.

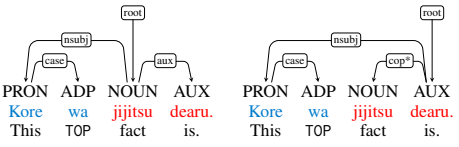
Adpositions and Noun Phrases Japanese has a nominalizer, "-no," which can convert any content word to a noun. For example, a verb "hataraku" (*work*) can be a noun with that nominalier "hataraku-no" (*working*), but such nominalization is not reflected in the PoS tag of the nominalized words. We regard the words nominalized by "-no" (tagged as SCONJ) as NOUN in this paper, and thus, the following sentence will also be a target of <*Adp, NP*> swapping even though the head of the nmod dependency "karui" is ADJ rather than NOUN:



Copula and Predicate The cop dependency is attached only to an auxiliary verb "desu" in the original Japanese UD. We increased the coverage of copula verb based on the following criteria:

- AUX of "dearu" (*is*), "denai" (*is not*), "dewanai" (*is not*), "janai" (*is not*), "rashî" (*looks/seems/sounds like*) "kamoshirenai" (*may be*) .
- VERB with "iru" (*exist*), "aru" (*exist*), or "naru" (*become*) as its lexicon.

That is, in the following example, the original annotation on the left with the copula verb "dearu" is converted into the dependency graph on the right:



Note that we only targeted the cases where the copula verb has a nsubj dependent since the corresponding construction in English, i.e., a sentence "A is B." with the omission of "A," is very rare.

Auxiliary and Verb While auxiliary words are swapped with an entire verb phrase rather than a single verb in the English implementation of <*Aux, V*> swapping, the Japanese implementation only swaps a single verb. This is because Japanese auxiliary verbs are typically analyzed as affixes, and thus separating them from the verb modifies the language beyond simply breaking the Greenbergian correlation. Taking the sentence in Table 4 as an example, the auxiliary "teiru" is moved immediately before the verb "tsudui," rather than the initial position of the sentence, regarding the whole descendants of the verb ("Ichigo no kisetsu ga shichigatsu kara hachigatsu made") in the <*Aux, V*> variant.

Noun and Genitive We identified the genitive constructions as follows:

- A nmod dependency to a noun phrase.
- The dependent has either particle of "no," "ga," or "tsu."

We exclude some exceptional constructions; for example, we did not swap the expression "X-no yô na" to be "yô X-no na." We also considered the nominalization in identifying a noun, as explained in the <*Adp, NP*> swapping.

B General Swapping Algorithm

Algorithm 1 below is the basic form of the depth-first swapping algorithm. This basic algorithm was modified to handle the specific of each language and correlation pair as described in Appendix A.

Algorithm 1 Swapping Greenbergian correlation pairs in a sentence

```
def Swap:sentence s, UD parse p, Correlation pair <X, Y>
  stack ← [root]
  visited ← set()
  while stack is not empty :
    node ← POP(stack)
    if node is not in visited :
      ADDTOVISITED(visited, node)
      for each child c of node in the parse p of s :
        if node is verb-patterner X and c is object-patterner Y :
          SWAPPAIR(node, c, s, p)
        if c is not in visited :
          PUSH(stack, c)
  return s
```

C Additional Annotation Guidelines

The 5-point Likert scale used to evaluate the validity of swapped sentences is given below:

1. All or most swaps have serious errors
2. A few serious errors or several small errors
3. A few small errors
4. A minor error or less likely but valid changes
5. Perfect

D Details on Experimental settings

Language Models All models are trained using the HuggingFace library (Wolf et al., 2020). For GPT-2 small model, sub-word tokenization is implemented by Byte-Pair Encoding (BPE) algorithm (Sennrich et al., 2016) with a vocabulary size of 32,000. For LTG-BERT, we adopted the same WordPiece tokenizer with a vocabulary size of $2^{14} = 16384$ as in the original implementation (Samuel et al., 2023), only removing special characters <*TAB*> and <*PAR*> as it doesn't apply to Wiki-40B text format.

Stanza Parsers We use Stanza (Qi et al., 2020) version 1.5.1 and 1.6.1 based on the UD 2.0 formalism (Nivre et al., 2020) for English and Japanese, respectively. For Japanese, we used a long-unit-word (LUW) parser (`https://github.com/UniversalDependencies/UD_Japanese-GSDLUW`) which is more compatible with the syntactic UD scheme (Omura et al., 2021) rather than the default, short-unit-word (SUW) parser which is better for morphological analysis (Tanaka et al., 2016; Murawaki, 2019; Pringle, 2016).