Can Language Models Learn Typologically Implausible Languages?

Tianyang Xu[a,b] **Tatsuki Kuribayashi**[c] **Yohei Oseki**[d]
Ryan Cotterell[a] **Alex Warstadt**[a,e]

[a]ETH Zürich, Switzerland [b]Toyota Technical Institute at Chicago, USA [c]MBZUAI, UAE
[d]The University of Tokyo, Japan [e]University of California San Diego, USA
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 exhibit intriguing correlations, often attributed to learning biases in humans. Language models (LMs) provide a scalable and naturalistic framework for studying artificial language learning—one not available in human research. We investigate how learnability varies across typologically *plausible* and *implausible* languages that closely follow the word order *universals* identified by linguistic typologists. Our study trains LMs on highly naturalistic counterfactual versions of English (head-initial) and Japanese (head-final). Compared to prior work, our datasets more precisely target the boundary between typological plausibility and implausibility. Our experiments show that LMs learn subtly implausible languages more slowly, though they eventually reach similar performance on some metrics regardless of typological plausibility. These findings suggest that LMs exhibit typologically aligned learning preferences and that certain typological patterns may emerge from 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—more generally—possible natural languages, and, moreover, to explain why some possible grammars are widely attested, but others are not. In service of that larger goal, many typological universals and tendencies have been identified (Greenberg, 1963; Barwise and Cooper, 1988; Dryer, 1992; Hyman, 2008). However, at present, we do not fully understand *why* these tendencies exist. One widespread belief is that typological patterns are caused by a learning bias; however, there is disagreement over several decades regarding whether that bias is language-specific (Culbertson and Kirby, 1981; Baker, 2009) or domain-general (Culbertson and Kirby, 2016). Moreover, other explanations exist that do not appeal to learning bias at all (Hahn et al., 2020). This debate has been difficult to resolve experimentally with human subjects because we cannot manipulate variables during the acquisition of a child's first language, leaving only the possibility of small-scale artificial language learning experiments. Language models (LMs), by contrast, have recently been proposed as convenient stand-ins for human learners, enabling large-scale and fully 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, *inter alia*), sparking some debate. Some have gone so far as to argue that LMs have little consequence for linguistic theory because they putatively can learn many possible and impossible languages with equal facility (Mitchell and Bowers, 2020; Chomsky et al., 2023; Moro et al., 2023). Empirical results, however, indicate a more nuanced picture: In counterfactual language learning experiments, Kallini et al. (2024) found that LMs do in fact struggle to learn impossible languages—e.g., languages lacking hierarchical structure—relative to similar plausible languages.

In this paper, we investigate whether statistical learning in a learner lacking an explicit innate bias toward typologically plausible languages can nonetheless exhibit such a bias. If so, this would suggest that some conceivable but unattested grammars are dispreferred due to

Transactions of the Association for Computational Linguistics, vol. 14, pp. 588–611, 2026. https://doi.org/10.1162/TACL.a.640
Action Editor: Nathan Schneider. Submission batch: 1/2025; Revision batch: 6/2025; Published 4/2026.

domain-general principles—for instance, biases arising from more general information-theoretic constraints (Someya et al., 2025)—implying that strong innate restrictions on the hypothesis space are not always necessary for LMs to develop human-like learning biases. We test this hypothesis by examining the learnability of typologically dispreferred languages that lie closer to the boundary of possibility than those studied by Kallini et al. (2024).

In this study, the typological tendencies we investigate are among 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 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 is novel. First, 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. Second, we also target fine-grained decision boundaries between typologically *plausible* and *implausible* languages by manipulating one specific grammatical property in each counterfactual corpus at a time. 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 languages' learnability under two types of LMs (autoregressive and masked) trained from scratch on our counterfactual corpora. We evaluate learnability from multiple perspectives: (i) perplexity per-token on the full corpus, (ii) preferences on minimal pairs contrasting original and counterfactual word orders, and (iii) accuracies on broad syntactic benchmarks BLiMP (Warstadt et al., 2020) and JBLiMP (Someya and Oseki, 2023). Our results show that LMs often find it harder to learn counterfactual, typologically implausible languages compared to minimally different natural languages. While Kallini et al. (2024) reached a similar conclusion using more contrived languages that do not resemble any human language, our findings extend this observation to languages that are merely *implausible*, i.e., those resembling a small minority of attested human languages. Although we cannot entirely rule out confounds introduced during the creation of counterfactual corpora, our results have important implications if they prove robust. Despite arguments that LMs are largely irrelevant to linguistics and cognitive science due to their non-human-like learning biases (Chomsky et al., 2023; Moro et al., 2023), we contend that such dismissal is premature: Whether such biases arise in a learner without explicit restrictions on its hypothesis space bears directly on long-standing debates about the necessary and sufficient conditions for human-like learning preferences. Furthermore, our findings suggest that specific learning biases hypothesized in humans *are* naturally aligned with those observed in language models.

2 Background

2.1 Typological Tendencies

What are all the conceivable types of language that humans could develop? Linguistic theory offers one possible answer: Natural language grammars can be described through the lens of formal language theory (Chomsky, 1956), and only a subset of formal languages appear compatible with the human cognitive architecture. For instance, no human language counts modulo three, even though such a pattern is trivially encodable by a finite-state automaton (Newmeyer, 2005). However, an exact formal characterization of which languages are learnable by humans continues to elude us. Even after decades of research, it is clear that human languages occupy a complex and difficult-to-define region within the space of possible grammars (Newmeyer, 2005; Chomsky and Lasnik, 1993). Nevertheless, many generalizations have been proposed about this space—for example, concerning syntactic categories (Chomsky, 1965), quantifiers (Barwise and Cooper, 1988), and phonology (Hyman, 2008). Some of these generalizations are true universals that no human language violates—such as the prohibition against counting modulo three—whereas others reflect strong statistical tendencies, describing correlations that occur far more often than would be expected by chance. Accordingly, we can distinguish between **impossible** and **implausible** languages.

In terms of statistical tendencies, Greenberg (1963) proposed a list of several dozen word order and morphological correlations based on

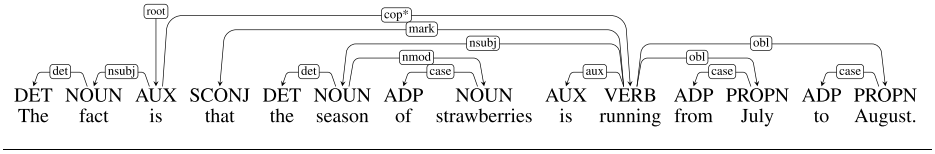
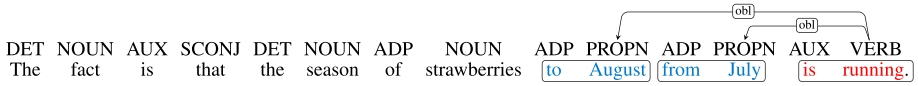
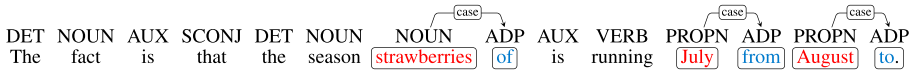
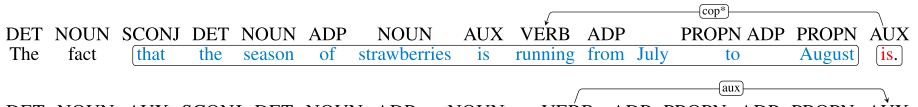
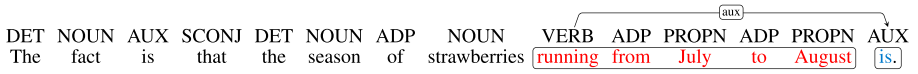
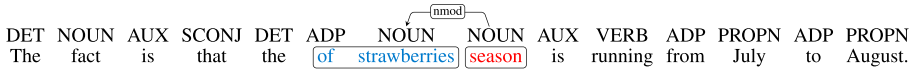
Correlation Pair	Example
Original	
<V, O>	
<Adp, NP>	
<Cop, Pred>	
<Aux, V>	
<Noun, Genitive>	

Table 1: Illustrative examples of each of our counterfactual variants of English. Head phrases are colored red, and dependent phrases are colored blue. The `cop*` notation reflects the swapped copula edge. As mentioned in Figure 5, *O* in *<V, O>* covers nominal objects of verbs, obliques and complement clauses, etc. 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.

a survey of 30 languages. For example, he observes that ‘‘[i]n languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes.’’ Building on this line of work, Dryer (1992) formulated a list of **correlation pairs**—pairs of morphosyntactic categories where each pair consists of a head category *H* and its dependent category *D*[1]—that tend to share the *same* relative ordering as the dominant order of the verb and object 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, and to those that violate them as **non-harmonic**, i.e., implausible. A subset of the correlation pairs we focus on in this paper is listed in Table 1.

2.2 Learnability of Implausible Languages

Learnability has long been proposed as a key mechanism underlying typological universals and tendencies such as word order harmony. This account has an intuitive appeal: Language evolves through cycles of reanalysis by child learners (Peyraube, 1912; Cournane, 2017), and such reanalysis tends to favor grammars that are easier to acquire, making them more frequent over generations (Kirby et al., 2008). However, there is little consensus on why some grammars are inherently easier or harder to learn, or why humans might possess a **harmonic bias** that facilitates the acquisition of harmonic grammars.

One possibility is that humans possess **language-specific biases**, of which harmonic bias is only one example. For instance, Chomsky (1965) proposed the theory of **universal grammar**, which posits that humans have an innate language acquisition device that biases the learning of certain grammars. This theory was later refined into the **principles and parameters**

[1]Syntacticians often disagree on the correct generalization that characterizes correlation pairs (Hawkins, 1983; Dryer, 1992; Kayne, 1994). Indeed, an entirely theory-neutral description is unlikely to ever emerge for these data. As suggested by our notation, the *H* elements that pattern with the verb tend to be functional 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 refer to these elements as heads and dependents, our study relies only on the existence of these correlation pairs, not on any specific theoretical analysis.

framework (Chomsky, 1981; Chomsky and Lasnik, 1993), which—most relevant to the present discussion—introduced the **head parameter**, determining whether complements appear to the left or right of their heads (Chomsky and Lasnik, 1993, p. 35). More recently, some proponents of universal grammar, including Chomsky, have recharacterized it as a domain-general bias for recursion (Hauser et al., 2002; Berwick and Chomsky, 2016), and the necessity of the head parameter has been questioned (Kayne, 2013). Nevertheless, contemporary advocates of innate language-specific biases continue to posit parameters to account for head directionality (Baker, 2009; Cinque, 2017) as well as a host of other typological patterns (Biberauer et al., 2009; Baker, 2015, *inter alia*).

An alternative to universal grammar is that **domain-general biases** are sufficient to explain harmonic bias and possibly other typological patterns. For instance, there is evidence that humans have a simplicity bias across several domains of cognition (Chater and Vitányi, 2003; Hsu et al., 2013; Stabler, 2013), and such a bias could explain the preference for harmonic languages (Culbertson and Kirby, 2016, 2022), 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). Neuroimaging also shows that typologically implausible artificial languages lead to lower activation in language centers in the brain than typologically plausible ones (Musso et al., 2003). 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; Hale and Stanojević, 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 **grammar-first** approach where a counterfactual corpus is generated from a manually specified lexicon and grammar, or a **corpus-first** approach where a naturalistic corpus is modified according to a set of rules. At the extreme end of grammar-first approaches are studies testing the learnability of different formal language classes across neural network architectures, generating data that often falls well outside the complexity class of natural language (Ebrahimi et al., 2020; DuSell and Chiang, 2022; Hao et al., 2022; Delétang et al., 2023; Borenstein et al., 2024; Someya et al., 2024).[2] A more natural variant uses probabilistic context-free grammars inspired by natural language but designed to selectively violate specific typological properties. Such studies (White and Cotterell, 2021; Kuribayashi et al., 2024) have yielded mixed evidence on whether the inductive biases of LMs align with those of humans. However, grammar-first corpora greatly simplify the challenges of language learning and processing. Naturalistic data, by contrast, exhibits a richness that is difficult to replicate in synthetic corpora.

The corpus-first approach achieves greater ecological validity by starting from a natural corpus that retains the full complexity of real data and applying controlled manipulations, often via constituency or dependency parses. A common method filters particular sentence types using parses (Jumelet and Hupkes, 2018; Warstadt, 2022; Patil et al., 2024; Misra and Mahowald, 2024), while other work applies transformation rules to parsed sentences. Wang and Eisner (2016) generate about 50,000 synthetic languages by stochastically reordering the dependents of nouns

[2]Such empirical studies differ from work that proves which languages can be recognized by LMs (Strobl et al., 2024).

and verbs in treebanks to match other languages' word order patterns, and Hahn et al. (2020) construct counterfactual dependency grammars by specifying, for each arc label, whether the dependent appears to the left or right of the head and its relative distance. While such methods yield more ecologically valid counterfactual languages, they are noisy and difficult to control: messy source data, annotation errors, and limits of linguistic annotation systems often result in counterfactual corpora containing more ungrammatical content with respect to the counterfactual grammar than the original corpus. Our study, in contrast, takes a careful corpus-first approach designed to minimize and control for such 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, i.e., implausible languages.

The Independent Variable. We manipulate word order harmony using a corpus-first approach to counterfactual corpus generation. Starting from naturally occurring corpora for harmonic languages, we systematically violate five Greenbergian correlation pairs, one at a time (§4). Each pair yields two harmonic (SVO head-initial, SOV head-final) and two non-harmonic (SVO head-final, SOV head-initial) variants, allowing controlled comparison across word order types.

The Dependent Variables. 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 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. 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 Universal Dependencies (UD) 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 corpus 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. 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 Table 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 (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 each input sentence. Word order is manipulated at the span level: Given a sentence $\boldsymbol{w} = [w_1, \ldots, w_N]$ and its dependency parse p, a word pair (w_h, w_d) where $1 \leq h, d \leq N$ with a specific dependency type is identified, and their spans $\boldsymbol{s}_h$ and $\boldsymbol{s}_d$ are defined as contiguous word sequences in $\boldsymbol{w}$ consisting of the identified word and its descendants.[3] The relative positions of $\boldsymbol{s}_h$ and $\boldsymbol{s}_d$ are then swapped. All token pairs meeting the <*H, D*> criteria in Table 3 are processed recursively (Algorithm 1. in Appendix A), with exceptions and coordination handling described 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.).

Handling Japanese-Specific Issues. Sometimes, a direct application of the English implementation to Japanese fails due to grammatical and annotation differences. For instance, Japanese UD employs a looser notion of *word* than English UD. To ensure the swapping algorithm remains both comparable and correct across languages, we introduce 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 Figure 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,

[3]In practice, span boundaries vary slightly by language and annotation; we do not always include all and only the descendants.

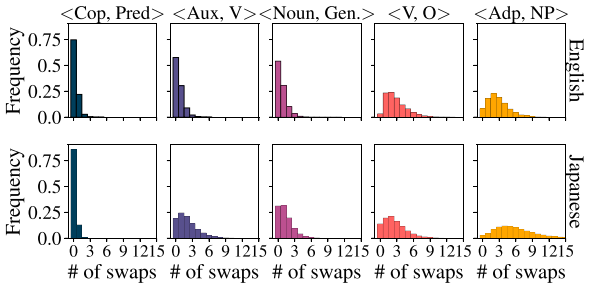


Figure 1: Histogram of the number of swaps per sentence for each counterfactual language.

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 corpus-first studies, which usually do not 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-swapping 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.*''

	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.

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

[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).

from more developmentally plausible data such as child-directed speech. We use Stanza to obtain dependency parses for every sentence in the corpora, and coarsified the obtained arc labels following Hahn et al.'s implementation, i.e., removing the fine-grained syntactic labels after the colon except for special cases. 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 non-Japanese brackets or brackets that contain English characters or numbers 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 compare the perplexity (PPL) trajectories through training epochs exhibited by the LMs on the held-out data in each language, including counterfactual ones.[7] Figure 2 shows the PPL differences (ΔPPL = $\text{PPL}_{\text{counterfactual}} - \text{PPL}_{\text{baseline}}$), where a positive value indicates worse performance in the counterfactual languages. The ΔPPL consistently converge to positive scores for GPT-2 models. They trend towards slightly positive for LTG-BERT in English, and they appear to be distributed around 0 for LTG-BERT in Japanese. In general, this suggests that counterfactual languages are more difficult to learn than the originals, or similarly difficult. Surprisingly, for Japanese, we observe negative ΔPPL in the early stages of training, but they quickly converge to positive values for GPT-2, suggesting that the counterfactual languages have shallow

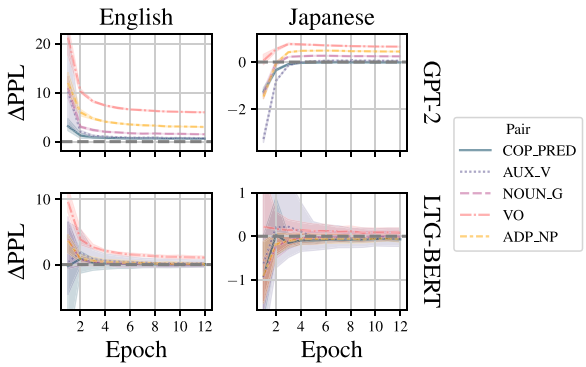


Figure 2: PPL differences between counterfactual and BASELINE LMs on their respective held-out data for the English-based (left) and Japanese-based (right) corpora. Shaded areas indicate standard deviations over three random seeds.

patterns that are easier to pick up on, but they are ultimately harder to acquire.

We find that the ORIGINAL LMs achieve slightly better PPL than BASELINE but at an approximately similar scale. Thus, we conclude that 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 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. To test the learnability difference between counterfactual and real languages, for each of ten conditions $\{\text{En}, \text{Ja}\} \times \{5 \text{ word orders}\}$,[8] we perform a paired Wilcoxon signed-rank test by comparing 72 PPL 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.[9]

[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.

[8] When we further split the conditions to distinguish the model architectures, i.e., 20 conditions of $\{\text{GPT}-2, \text{LTG-BERT}\} \times \{\text{En}, \text{Ja}\} \times \{5 \text{ word orders}\}$, and run the paired Wilcoxon signed-rank tests separately for each condition, we did not find any architecture-specific tendencies, at least through the lens of our statistical tests. Almost all conditions of English GPT-2 (5/5) and LTG-BERT (4/5) found the significance (at $\alpha = 0.05$), and all the Japanese conditions did not find the significance (0/10) (at $\alpha = 0.05$). Nevertheless, based on Figure 2, GPT-2s seemingly yielded somewhat clearer results with larger effect size and less variance than LTG-BERTs.

[9] We additionally applied the Bonferroni correction; thus, we consider $\alpha = 0.0025 (= 0.05/20)$.

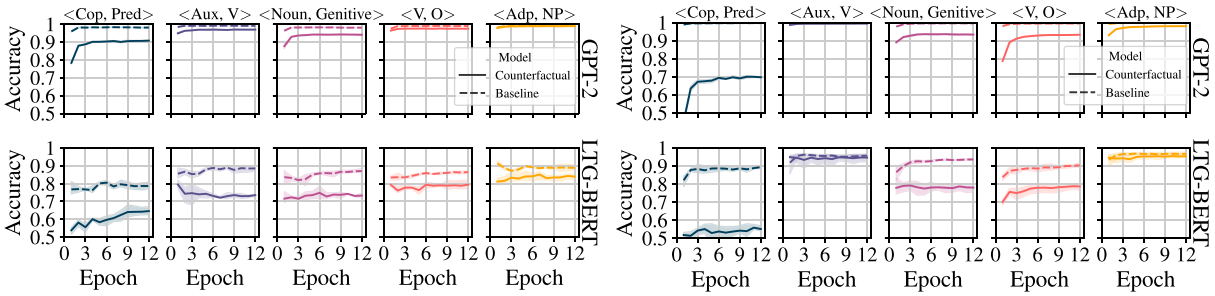


Figure 3: Performance trajectories for minimal pair comparisons targeting the counterfactual word order for counterfactual models and natural order for baseline model, for English-based LMs (left) and Japanese-based LMs (right). Shaded areas present standard deviation (SD) over three random seeds.

The null hypothesis is that the counterfactual language exhibits the same PPL scores as the real language (equal learnability of these two data for LMs). For six out of ten conditions, the real language exhibited significantly lower PPLs (at $\alpha = 0.05$) than the counterfactual one. This was the case for all English counterfactuals and the Japanese *<O, V>* counterfactual. The real language tends to be easier for LMs to learn throughout training than the counterfactual languages, while the magnitude of PPL difference was not so substantial (Figure 2).

6.2 Evaluation 2: Minimal Pair Preferences

Settings. The previous evaluation measures PPL on all the tokens in the corpus; however, many tokens are not directly related to our targeted word order change. For a more targeted evaluation, we design a forced-choice task testing whether each model prefers the correct word order given a minimal pair of 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. Following much prior work (Marvin and Linzen, 2018; Warstadt et al., 2020; Hu et al., 2020), we assess word order preferences by comparing the model's surprisal for each sentence; that is, the sentence with lower surprisal is preferred by the model. We report accuracy on this task computed for a set of minimal pairs sampled from the held-out set of Wiki-40B data.

Results. Figure 3 shows the trajectory of accuracy during LM training. While GPT-2 consistently performs better than LTG-BERT, all the counterfactual LMs prefer the correct word order over the incorrect one much more than random chance (accuracy of 0.5). This supports the conclusion that LMs ultimately generalize well to counter-Greenbergian languages and learn the counterfactual ordering pattern successfully. Nevertheless, in many settings, the BASELINE LMs yield 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. Furthermore, LMs often appear to converge more quickly on the counterfactual languages than on the original language. However, there are some exceptional cases in which counterfactual languages exhibit almost the same accuracies as the corresponding BASELINE LMs, specifically when performance is near ceiling.

Statistical Tests. We perform paired Wilcoxon signed-rank tests for each correlation pair in each language, i.e., $\{5 \text{ word orders}\} \times \{\text{En}, \text{Ja}\}$, comparing 72 accuracy scores from $\{\text{GPT2}, \text{LTG-BERT}\} \times \{3 \text{ seeds}\} \times \{12 \text{ epochs}\}$ between the counterfactual and BASELINE models. Our null hypothesis is that counterfactual and real languages are equally learnable for an LM. In all ten settings, BASELINE LMs achieved significantly higher accuracies than the counterfactual LMs at $\alpha = 0.05$; in eight cases, $p < 1\text{e}{-12}$.

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, again using the minimal pair paradigm.

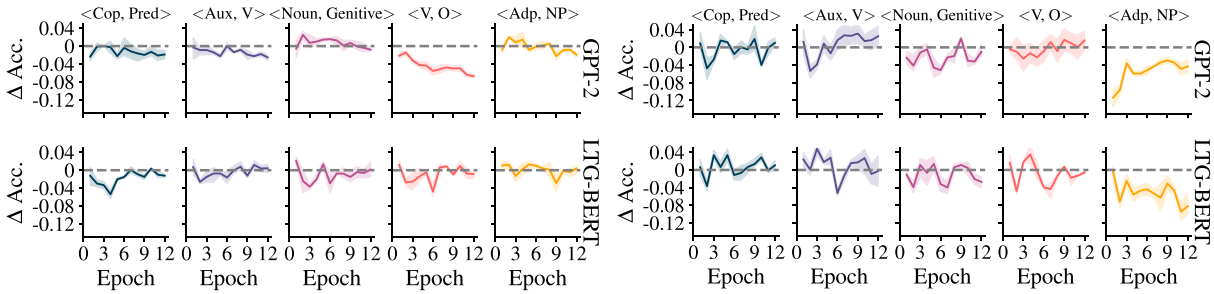


Figure 4: Performance differences between counterfactual and BASELINE LMs on BLiMP (left) and JBLiMP (right). Shaded areas indicate standard deviations over three random seeds.

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[10] 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 focus on the macro average of accuracy over the 12 BLiMP suites (or 9 JBLiMP suites). Figure 4 shows the difference in (J)BLiMP accuracy throughout training (ΔAcc. = $\text{accuracy}_{\text{counterfactual}} - \text{accuracy}_{\text{baseline}}$). The ΔAcc. scores are clustered around zero, but trend slightly negative, suggesting that counterfactual word order slightly but not drastically prevented LMs from acquiring grammatical knowledge.

Statistical Tests. We performed paired Wilcoxon signed-rank tests for each correlation pair in each source language $\{5 \text{ word orders}\} \times \{\text{En}, \text{Ja}\}$, comparing 864 accuracy scores from $\{\text{GPT2}, \text{LTG-BERT}\} \times \{3 \text{ seeds}\} \times \{12 \text{ epochs}\} \times \{12 \text{ BLiMP categories}\}$ between the counterfactual and BASELINE models. In six of ten settings, BASELINE LMs showed significantly higher BLiMP accuracies than the counterfactual ones ($\alpha = 0.05$).[11]

7 Discussion and Conclusions

Our findings show that autoregressive and masked LMs have a general learning bias—with some notable exceptions—favoring harmonic languages over the nonharmonic counterfactual languages we examined. Strikingly, the experimental results from §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. This suggests that unnatural changes to one part of the grammar can have deleterious consequences for other grammatical phenomena, aligning with earlier findings on the existence of indirect evidence in grammar learning (Warstadt, 2022; Misra and Mahowald, 2024). However, it also shows that implausible languages are still learnable for LMs at a coarse-grained level.

[10] 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.

[11] After Bonferroni correction for 30 tests across the three experiments, the conclusions generally hold. For the PPL (§6.1) and minimal pair (§6.2) experiments, results remain unchanged ($p < 0.0016 = 0.05/30$). In the BLiMP experiment (§6.3), real languages achieved higher accuracy than four counterfactual ones.

While the learning biases of LMs are interesting in their own right, we argue that they also have implications for our understanding of the possible mechanisms underlying linguistic typology. 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; Kodner et al., 2023; Milliére, 2024; McGrath et al., 2024). Here, we will not rehash all the details of this debate, but present a condensed account of how our experiments on LMs can inform ongoing debates about human language.

LMs are well-suited to provide an existence proof that a linguistic property or learning pattern can emerge without appealing to language-specific bias. Our results offer such an existence proof (with some caveats discussed below) that harmonic bias does not require language-specific bias. This is distinct from showing that language-specific bias is not the cause of harmonic bias in humans, as important differences remain between humans and LMs. Our reasoning aligns with arguments made by others (Clark and Lappin, 2011; Linzen, 2019; 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) arguably relies on domain-general learning biases, as evidenced in part by its efficacy for domains as far-reaching as vision (Dosovitskiy et al., 2021) and protein sequences (Jumper et al., 2021).[12] If one accepts this premise and accepts that our results demonstrate harmonic bias in Transformers, it follows that language-specific biases are not a necessary precondition for harmonic bias. While this may be suggestive that other hypothesized linguistic biases could also arise from domain-general learning principles, our results only speak to the sufficient conditions for harmonic bias.

More controversially, this existence proof may also increase one's credence in an explanation for harmonic bias *in humans* in terms of externally motivated domain-general biases such as simplicity bias. Anyone who previously rejected that explanation *a priori* on the assumption that it was impossible or highly unlikely that harmonic bias would arise from domain-general biases should update their priors. As with generalizing results from any model ''organism'', how much to do so depends on one's priors and how similar the model is to a human. Put another way, someone who believes that harmonic bias in humans arises from language-specific bias might have predicted that harmonic bias is something specific to humans, or learners very similar to humans. That this prediction is not borne out suggests that we should instead look for a potential common cause of harmonic bias in LMs and humans in terms of some other shared bias.

It falls to other work to better understand what such a common cause might be. Simplicity bias is a strong candidate for a domain-general bias that could lead to harmonic bias (Chater and Vitányi, 2003; Hsu et al., 2013; Kirby et al., 2008; Culbertson and Kirby, 2016). But this is only part of the story: How does the input interact with simplicity bias to give rise to harmonic bias? The Indirect Evidence Hypothesis (Perfors et al., 2011; Reali and Christiansen, 2005) holds that the existence (or absence) of one grammatical rule can influence the learner's credence in another. There is increasing evidence that indirect evidence influences grammatical generalization in model learners (Jumelet et al., 2021; Warstadt, 2022; Patil et al., 2024), and some work has already identified specific instances of indirect evidence (Misra and Mahowald, 2024). Our results suggest, for instance, that if a learner observes verbs preceding objects in a language, they may interpret that as indirect evidence that adpositions precede associated nouns. However, more targeted counterfactual language learning experiments are needed to better identify the specific causal links between word order phenomena. We must also endow model learners with different biases to determine which biases influence these indirect links.

Our findings also speak to two other issues in linguistics: First, there is the argument that LMs

[12] A counter-argument is that the Transformer architecture may have innate language-specific bias as a result of decades of trial-and-error searching for effective neural architectures for language processing. However, there is no explicit implementation of domain-specific biases, including those proposed for humans as an explanation for typology. Therefore, the null hypothesis should be that Transformers lack substantial language-specific bias. Furthermore, empirical results suggest that the bias of Transformers favors linear generalizations over hierarchical ones (Petty and Frank, 2021) and fail to systematically learn many classes of formal grammars (Delétang et al., 2023).

have minimal relevance to the study of human language because they learn possible and impossible languages with equal facility (Mitchell and Bowers, 2020; Chomsky et al., 2023; Moro et al., 2023). This argument relies on two premises that our work calls into question. The first premise is that only models that do not learn impossible languages have a role to play in linguistics. We argue that LMs are relevant to linguistics regardless of whether they show human-like learning biases. One of the primary goals of linguistic theory is to understand how different learning biases affect language learning outcomes. The domain generality and weak biases of LMs makes them a valuable model to demonstrate what learning outcomes can be observed without substantive language-specific bias.

The second premise is that LMs learn possible and impossible languages equally easily. While it is true that LMs can eventually learn counterfactual languages, they show learning preferences similar to those hypothesized by humans. Kallini et al. (2024) already demonstrated this, but they study counterfactual languages that are presumably truly unlearnable by humans, and they arguably fail to compare against minimally different counterfactual but linguistically plausible manipulations (Hunter, 2025). Our study furthers this conclusion by showing that LMs continue to show a learning bias for typologically preferred counterfactual languages closer to the boundary between plausible and implausible. Our counterfactual languages can be generated by a constituency-based naïve grammatical theory, and they could plausibly be learnable given the attestation of similar but cross-linguistically rare languages. Thus, the bias that our models demonstrate is arguably so subtle in humans as to only be a *preference* that exerts an influence on grammatical structure over generations (Kirby et al., 2008). If humans and LMs really do share such subtle biases, this increases the utility of LMs as models of human learners.

Our results speak to a second issue in linguistics: whether humans actually have a harmonic bias in the first place. As discussed in §2.2, the experimental evidence in support of this conclusion from human subjects is limited to small-scale studies on simple artificial languages. The corpus-first approach to counterfactual language creation allows us to test for the existence of harmonic bias in a naturalistic and complex domain at the scale of the input to human learners. While the existence of harmonic bias in LMs obviously does not imply its existence in humans, these converging results support the conclusion that harmonic bias is likely not a rare or outlandish property of effective language learners. Anyone who wishes to argue against earlier experimental results from human subjects (Culbertson et al., 2012; Culbertson and Newport, 2015; Culbertson et al., 2020) must now explain why LMs but not humans have a harmonic bias.

It bears mentioning other factors beside harmonic bias may still be equally (if not more) important causes of 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. 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 corpus-first 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

Mark Baker. 2015. *Case: Its Principles and Parameters*. Cambridge University Press. Number: 146.

Mark C. Baker. 2009. The macroparameter in a microparametric world. In *The Limits of Syntactic Variation*, pages 351–373.

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. https://doi.org/10.1007/978-94-009-2727-8_10

Robert C. Berwick and Noam Chomsky. 2016. *Why Only Us: Language and Evolution*. MIT Press.

Theresa Biberauer, Anders Holmberg, Ian Roberts, and Michelle Sheehan. 2009. Introduction: Parameters in minimalist theory. In *Parametric Variation: Null Subjects in Minimalist Theory*, pages 1–57. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511770784.001

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. https://doi.org/10.18653/v1/2024.acl-long.807

Nick Chater and Paul Vitányi. 2003. Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1):19–22. https://doi.org/10.1016/S1364-6613(02)00005-0

Noam Chomsky. 1956. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124. https://doi.org/10.1109/TIT.1956.1056813

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

Noam Chomsky. 1981. *Lectures on Government and Binding*. https://doi.org/10.1515/9783110884166

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*.

Guglielmo Cinque. 2017. A microparametric approach to the head-initial/head-final parameter. *Linguistic Analysis*, 41(3–4):309–366.

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. https://doi.org/10.1162/tacl_a_00589

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2024. Investigating critical period effects in language acquisition through neural language models. https://doi.org/10.1162/tacl_a_00725

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. https://doi.org/10.1016/j.cognition.2020.104392, PubMed: 32673786

Jennifer Culbertson and Simon Kirby. 2016. Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, 6. https://doi.org/10.3389/fpsyg.2015.01964

Jennifer Culbertson and Simon Kirby. 2022. Syntactic harmony arises from a domain-general learning bias. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*. Number: 44.

Jennifer Culbertson and Elissa L. Newport. 2015. Harmonic biases in child learners: In support of language universals. *Cognition*, 139:71–82. https://doi.org/10.1016/j.cognition.2015.02.007

Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. 2012. Learning biases predict a word order universal. *Cognition*, 122(3):306–329. https://doi.org/10.1016/j.cognition.2011.10.017, PubMed: 22208785

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, pages 1–54. https://doi.org/10.1162/coli_a_00402

Grégoire Delétang, 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*, 68:138–181.

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. https://doi.org/10.18653/v1/2020.findings-emnlp.384

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. https://doi.org/10.1073/pnas.1215776109, PubMed: 23071337

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. https://doi.org/10.1073/pnas.1910923117, PubMed: 31964811

John T. Hale and Miloš Stanojević. 2024. Do LLMs learn a true syntactic universal? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17106–17119, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.950

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. https://doi.org/10.1162/tacl_a_00490

Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579. https://doi.org/10.1126/science.298.5598.1569, PubMed: 12446899

John A. Hawkins. 1983. *Word Order Universals*. Academic Press, San Diego. https://doi.org/10.1016/B978-0-12-333370-4.50002-4

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. https://doi.org/10.18653/v1/2022.conll-1.7

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. https://doi.org/10.1111/tops.12005, PubMed: 23335573

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.158

Tim Hunter. 2025. Kallini et al. (2024) do not compare impossible languages with constituency-based ones. *Computational Linguistics*, 51:641–650. https://doi.org/10.1162/coli_a_00554

Larry M. Hyman. 2008. Universals in phonology. *The Linguistic Review*, 25(1–2):83–137. https://doi.org/10.1515/TLIR.2008.003

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.439

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. https://doi.org/10.18653/v1/W18-5424

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, 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 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. https://doi.org/10.1038/s41586-021-03819-2

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. https://doi.org/10.18653/v1/2024.acl-long.787

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. https://doi.org/10.1080/15475441.2005.9684215

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics*, 17(e13153). https://doi.org/10.5964/bioling.13153

Richard S. Kayne. 1994. *The Antisymmetry of Syntax*, volume 25, MIT Press.

Richard S. Kayne. 2013. Why are there no directionality parameters? *Studies in Chinese Linguistics*, 34(1).

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. https://doi.org/10.1073/pnas.0707835105, PubMed: 18667697

Jordan Kodner, Sarah Payne, and Jeffrey Heinz. 2023. Why linguistics will thrive in the 21st century: A reply to Piantadosi (2023). *arXiv preprint arXiv:2308.03228*.

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. https://doi.org/10.18653/v1/2024.acl-long.781

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. https://doi.org/10.1162/ling_a_00533

Tal Linzen. 2019. What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, 95(1):e99–e108. https://doi.org/10.1353/lan.2019.0015

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1151

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. https://doi.org/10.1177/09637214241268098, PubMed: 39949337

Raphäl Milliére. 2024. 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. https://doi.org/10.18653/v1/2024.emnlp-main.53

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. https://doi.org/10.18653/v1/2020.coling-main.451

Andrea Moro, Matteo Greco, and Stefano F. Cappa. 2023. Large languages, impossible languages and human brains. *Cortex*, 167:82–85. https://doi.org/10.1016/j.cortex.2023.07.003, PubMed: 37540953

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

Mariacristina Musso, Andrea Moro, Volkmar Glauche, Michel Rijntjes, Jürgen Reichenbach, Christian Büchel, and Cornelius Weiller. 2003. Broca's area and the language instinct. *Nature Neuroscience*, 6(7):774–781. Publisher: Nature Publishing Group US New York. https://doi.org/10.1038/nn1077, PubMed: 12819784

Frederick J. Newmeyer. 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. https://doi.org/10.1353/lan.2019.0009

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. https://doi.org/10.1162/tacl_a_00720

Andy Perfors, Joshua B. Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338. https://doi.org/10.1016/j.cognition.2010.11.001, PubMed: 21186021

Jackson Petty and Robert Frank. 2021. Transformers generalize linearly. *arXiv:2109.12036 [cs]*. ArXiv: 2109.12036.

Alain Peyraube. 1912. L'évolution des formes grammaticales. *Scientia; rivista di scienza*, 6(12):384. https://doi.org/10.3406/lgge.2002.2400

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, *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. https://doi.org/10.18653/v1/2020.acl-demos.14

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. https://doi.org/10.18653/v1/N19-1356

Florencia Reali and Morten H. Christiansen. 2005. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, 29(6):1007–1028. https://doi.org/10.1207/s15516709cog0000_28

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. https://doi.org/10.18653/v1/2023.findings-eacl.146

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

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. https://doi.org/10.18653/v1/2023.findings-eacl.117

Taiga Someya, Anej Svete, Brian DuSell, Timothy J. O'Donnell, Mario Giulianelli, and Ryan Cotterell. 2025. Information locality as an inductive bias for neural language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27995–28013, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1357

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. https://doi.org/10.63317/29pwuvvsdqdf

Edward P. Stabler. 2013. Two models of minimalist, incremental syntactic analysis. *Topics in Cognitive Science*, 5(3):611–633. https://doi.org/10.1111/tops.12031, PubMed: 23757195

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. https://doi.org/10.1162/tacl_a_00663

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. https://doi.org/10.1075/cilt.253.06ter

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.

Dingquan Wang and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more

data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505. https://doi.org/10.1162/tacl_a_00113

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. https://doi.org/10.18653/v1/2023.conll-babylm.1

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. https://doi.org/10.1162/tacl_a_00321

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

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44. https://doi.org/10.1162/ling_a_00491

Colin Wilson. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30(5):945–982. https://doi.org/10.1207/s15516709cog0000_89, PubMed: 21702842

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. https://doi.org/10.18653/v1/2020.emnlp-demos.6

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 Table 3. However, there are numerous exceptions which we discuss below.

H	UD Relation	*D*
verb	$\xrightarrow{obj}$, $\xrightarrow{iobj}$, $\xrightarrow{obl}$, $\xrightarrow{advcl}$ $\xrightarrow{cop*}$, $\xrightarrow{ccomp}$, $\xrightarrow{xcomp}$	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. The advcl label here only included nonfinite adverbial clauses.

Verbs and Objects. We construe the <*V, O*> correlation more broadly to refer to a verb on the

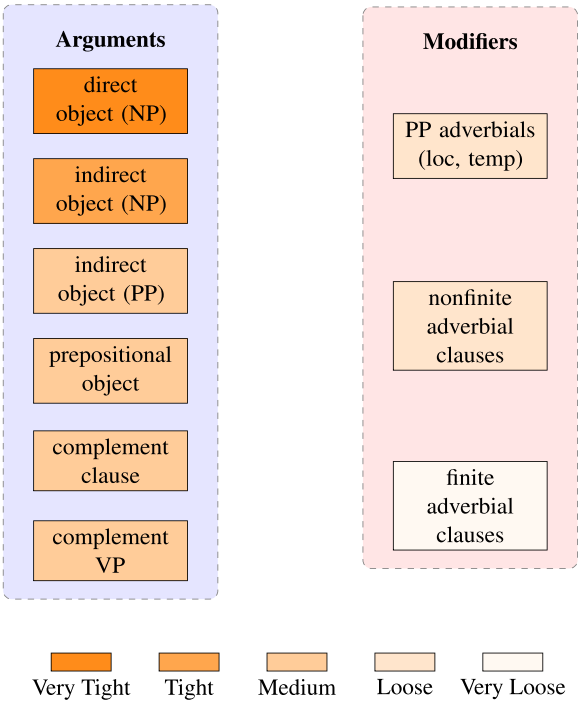


Figure 5: Illustration of different levels of tightness when classifying verbal dependents as objects.

one hand and its arguments and phrasal modifiers on the other. In linguistic theory, there is no universally agreed upon test for this notion of objecthood. 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, and prepositional phrase adverbials as ''objects'' in our implementation.

Mapping these linguistic constituents to UD relations, we use `obj`, `iobj`, `obl`, `cop`, `expl`, `advcl`, `xcomp`, and `ccomp` as dependency arc labels to identify the <*V, O*> pair. While we only regard nonfinite adverbial clauses as `advcl` in this paper, the `advcl` arc in the UD corpus corresponds to both finite & non-finite adverbial clauses. Our approach depends on identifying an `nsubj` arc 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, however, 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

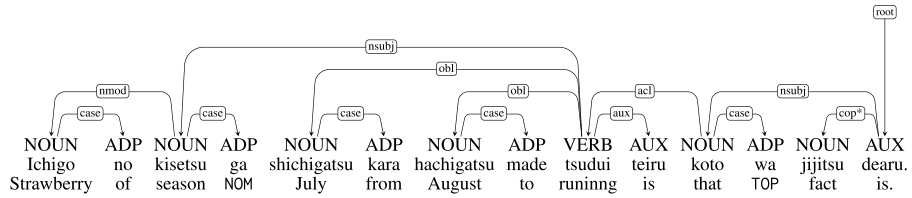
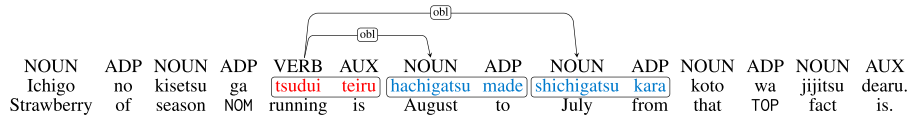
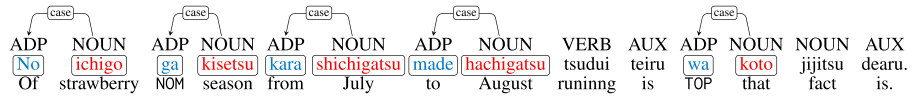
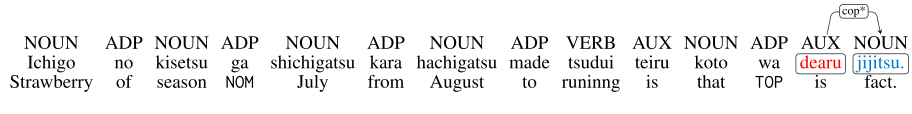
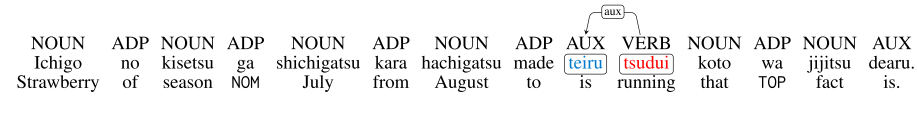
Correlation Pair	Example
Original	
<V, O>	
<Adp, NP>	
<Cop, Pred>	
<Aux, V>	
<Noun, Genitive>	

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.

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

A.3 Japanese-Specific Treatments

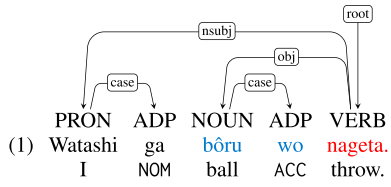
Table 4 shows examples of counterfactual variants of a Japanese sentence. The following paragraphs explain some treatments employed in modifying each word order correlation pair in the Japanese language.

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:

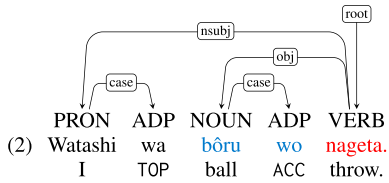
- If a word has a nsubj dependency AND the nominative case marker ‘‘ga,’’ the word is treated as a subject, and not swapped as object.
- If a word has a topicalization marker ‘‘wa,’’ the word is not swapped as object.
- The other arguments with the nsubj, obj, iobj, obl, cop, ccomp dependency are treated as an object and can be swapped (expl and xcomp are not used in the Japanese UD part).

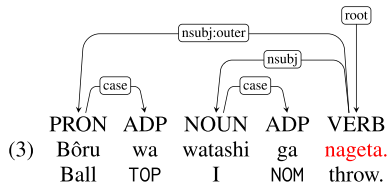
That is, unless an argument is explicitly marked as a subject or marked topic, it is regarded as an object, which is compatible with the loose definition of object employed in the English experiment.

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:



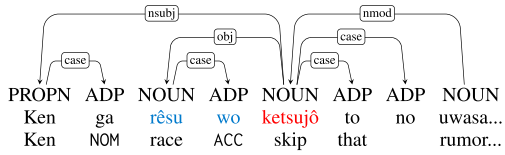
The subject is topicalized in sentence (2), and the object is topicalized in sentence (3):





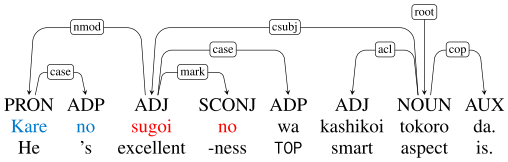
The topicalized component is typically ambiguous in terms of its grammatical case, and thus, the parser outputs were erroneous. Such a *marked* word order is beyond our interest since the Greenbergial correlations are generally on the canonical, *unmarked* word order of language. Thus, we did not modify 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.

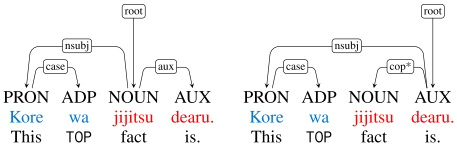
Adpositions and Noun Phrases. Japanese has a nominalizer, ‘‘-no,’’ which can convert any content word to a noun. For example, a verb ‘‘hataraku’’ (*work*) can be a noun with that nominalier ‘‘hataraku-no’’ (*working*), but such nominalization is not reflected in the PoS tag of the nominalized words. 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).