

# Function Words as Statistical Cues for Language Learning

Xiulin Yang Heidi Getz Ethan Gotlieb Wilcox  
Georgetown University  
{xy236, heidi.getz, ethan.wilcox}@georgetown.edu

## Abstract

What statistical properties might support learning abstract grammatical knowledge from linear input? We address this question by examining the statistical distribution of function words. Function words have been argued to aid acquisition through three distributional properties: high frequency, reliable syntactic association, and phrase-boundary alignment. We conduct a cross-linguistic corpus analysis of 186 languages, which confirms that all three properties are universal. Using counterfactual language modeling and ablation experiments on English, we show that preserving these properties facilitates acquisition in neural learners, with a Goldilocks effect: function words must be frequent enough to be reliable, yet diverse enough to remain informative to structural dependency. Probing analyses further reveal that different learning conditions produce systematically different reliance on function words.<sup>1</sup>

## 1 Introduction

A central puzzle in language acquisition is how learners abstract grammatical knowledge from linear input. Extensive research on statistical learning argues that human language learners make use of distributional information in their input, such as transition probability (Thompson and Newport, 2007) and prosodic grouping (Morgan et al., 1987; Romberg and Saffran, 2010). Within this literature, function words, such as determiners, auxiliaries, and prepositions, have been argued to play a critical role (e.g., Green, 1979; Hicks, 2006). Three properties of function words have been argued to facilitate acquisition: (i) high lexical frequency, (ii) reliable association with particular syntactic structures, and (iii) consistent position at phrase boundaries. Because of these properties, function words might

serve as anchor points for learning, cuing particular word sequences for analysis (e.g., the Anchoring Hypothesis; Valian and Coulson 1988; see also Morgan et al. 1987 on boundary-aligned cues) or giving learners a small set of high-frequency items to track (e.g., the Marker Hypothesis; Green 1979, with related evidence from Getz and Newport 2019; Zhang et al. 2015; Mintz 2006).

This literature suggests that function words contribute in important ways to statistical learning, yet two major gaps remain. First, existing studies draw conclusions about the properties of function words primarily from English or a small set of languages (e.g., Green, 1979; Kimball, 1973; Clark and Clark, 1977; Shi et al., 1998), limiting the generalization of their findings. Linguists have characterized the grammatical properties of function words in detail for thousands of languages (Haspelmath, 2005), and these analyses suggest that many of the distributional properties of function words are likely to be universal; however, this has not been confirmed through statistical analyses of diverse language corpora. If these properties play a systematic role in language acquisition in general, it is important to examine whether they are robust across languages (RQ1). Second, because these properties are often examined in isolation in simplified artificial language settings, it remains unclear whether their findings can be extended to more complex and therefore ecologically valid learning simulations (RQ2.1) and how and to what extent each property contributes to syntactic generalization (RQ2.2). Furthermore, is the role of function words as statistical cues in language learning carried over to language processing? Or are other cues used to parse sentences once grammatical representations become robust enough? (RQ3)

We answer RQ1 by conducting a cross-linguistic analysis using data from the Universal Dependencies (UD) project (de Marneffe et al., 2021), confirming the three distributional properties are uni-

<sup>1</sup>Code: [https://github.com/picol-georgetown/function\\_word](https://github.com/picol-georgetown/function_word); models: <https://huggingface.co/collections/xiulinyang/function-words>.

versal across 186 languages in our sample. To address RQ2, we take transformer language models (LMs) as domain-general and weakly-biased learners (Wilcox et al., 2024) and train them on counterfactual variants of natural text in which each distributional property of function words is systematically manipulated. We also train  $n$ -gram models as baselines for syntactic generalization. We confirm that a language is most learnable when all three statistical properties of function words jointly hold. Additionally, we observe a Goldilocks effect, whereby function words must be frequent enough to be reliable, yet sufficiently diverse to remain informative for their linguistic dependents. Finally, for RQ3, we conduct attention probing and function word ablation experiments to examine how LMs internally represent and deploy function word information. We find models trained on languages with the three function word properties develop attention heads specialized for function words and rely more on function word information during processing.

## 2 Background & Related Work

### 2.1 Defining Function Words

The distinction between function and content words is foundational in linguistic theory (Rizzi and Cinque, 2016; Abney, 1987) and language development research (Dye et al., 2019), though the boundary is not always clear-cut (Kayne et al., 2005)<sup>2</sup>. Typically, function words are characterized as a closed class with high frequency (Morgan et al., 1987), reduced prosodic prominence (Selkirk, 2014; Bögel, 2021), systematic alignment with phrase boundaries (Kimball, 1973; Christophe et al., 2008; Clark and Clark, 1977), and relatively light semantic content despite high grammatical utility (Carlson, 1983). In our modeling experiments, function words are defined purely through their phrasal alignment and high frequency. We leave cues such as semantic content and prosodic prominence as avenues for future research.

### 2.2 Function Words in Language Acquisition

Although early speech production by children usually lacks function words (Brown, 1973), evidence suggests that even in infancy, learners perceive these elements (Shi et al., 2006, 1999) and are

<sup>2</sup>The conclusion is that the number of functional elements in syntax is not easy to estimate, but at the same time that 100 would be a low estimate.”(Kayne et al., 2005, p.288)

sensitive to their distribution (Hochmann et al., 2010; Gerken et al., 2005; Christophe et al., 2008; Kedar et al., 2006). Experiments with artificial languages have shown that languages are more learnable by humans when function words are (1) more frequent than content words (Valian and Coulson, 1988); (2) reliable predictors of specific phrase structures (Green, 1979; Getz and Newport, 2019); and (3) systematically positioned at phrase boundaries (Morgan et al., 1987).

### 2.3 Computational Perspectives

Our work aligns with a growing body of research examining how manipulating training data shapes language model behavior. One line of work alters the overall statistical distribution of languages, testing whether languages that are less aligned with hypothesized grammatical principles are harder for models to learn (Kallini et al., 2024; Yang et al., 2025a; Ziv et al., 2026). Another line manipulates the proportion of certain targeted constructions, often referred to as Filtered Corpus Training (Patil et al., 2024), and has been applied to a range of phenomena, including AANN constructions (Misra and Mahowald, 2024), *let alone* (Scivetti et al., 2025), and Poverty of the Stimulus effects (Yang et al., 2026), among others (e.g., Yao et al., 2025).

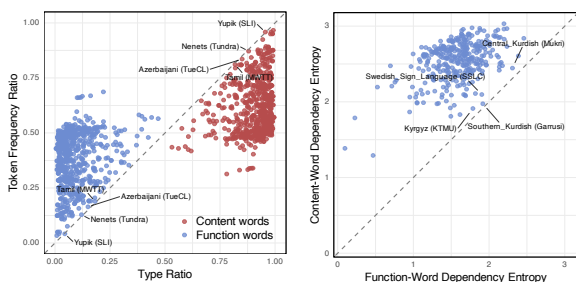
In contrast, computational work on function words remains relatively limited and falls into two strands. The first uses computational models to simulate human acquisition, showing that distributional cues such as frequent frames can support syntactic categorization (Mintz, 2003; Chemla et al., 2009; Gutman et al., 2015). Related work demonstrates that placing function words at phrase boundaries facilitates categorization and word learning (Mintz et al., 2002; Johnson et al., 2014). More recently, Ma and Xu (2025) replicate Valian and Coulson (1988) with LLMs and find limited sensitivity to marker frequency in artificial language settings. The second strand probes pretrained models, with mixed evidence on whether language or vision-LMs encode knowledge of function words (Kim et al., 2019; Ettinger, 2020; Portelance et al., 2024).

## 3 Do Function Words Share Universal Distributional Properties?

Research on function words in the field of language acquisition has been heavily skewed to English. In this section, we take a broader crosslinguistic per-

Properties	Manipulation	Example
Lexical Frequency	NOFUNCTION	dog happily chasing dog garden .
	MOREFUNCTION	thi dog wist happily chasing que dog ap tho garden .
	FIVEFUNCTION	the dog will happily chasing the dog at the garden .
	STANDARD FUNCTION	a dog is happily chasing another dog in the garden .
Structural Dependency	RANDOMDEP	by dog in happily chasing at dog by will garden .
	BIGRAMDEP	by dog in happily chasing by dog at will garden .
	PHRASEDEPENDENCY	a dog is happily chasing another dog in the garden .
Phrase boundary	WITHINBOUNDARY	a dog happily is chasing another dog the in garden .
	ATBOUNDARY	a dog is happily chasing another dog in the garden .

Table 1: An overview of manipulated languages



(a) Word frequency ratio vs. (b) Function vs. Content word type ratio

Figure 1: Distributional properties of function words across languages. Function words tend to have higher frequency-to-type ratios and lower syntactic dependency entropy than content words. The top four languages (with treebank names in parentheses) closest to the diagonal are labeled.

spective. We ask whether the properties identified as crucial for learning, i.e., high-frequency, structural predictability, and phrase boundary alignment, are linguistic universals. For this purpose, we use UD (v2.17), which includes 339 treebanks across 186 languages..

**High frequency** We quantify the frequency of function words by jointly considering their inventory size (*types*) and their usage frequency (*tokens*). Specifically, for a word class  $c$ , we define the *type ratio* as  $\frac{|\mathcal{V}_c|}{|\mathcal{V}|}$ , where  $\mathcal{V}_c$  denotes the set of unique word types belonging to class  $c$  (content or function words) and  $\mathcal{V}$  the full vocabulary. We further define the *token frequency ratio* as  $\frac{\sum_{w \in \mathcal{V}_c} \text{count}(w)}{\sum_{w \in \mathcal{V}} \text{count}(w)}$ . Following the UD convention, we define function words as closed-class items excluding NUM and treat NUM and other open-class categories as content words. We exclude non-linguistic tags (X, PUNCT, SYM).

Results are shown in Figure 1a, where each UD treebank contributes two points, corresponding to

scores of function words and content words. If word classes were uniformly distributed, the proportion of function word tokens would match their type proportion, yielding points along the diagonal. Instead, we observe a robust cross-linguistic pattern: function words occupy a small inventory size but account for a disproportionately large share of tokens, while content words fall below the diagonal across languages, reflecting larger inventories with lower token frequencies.

**Reliable structural prediction** Function words occur in predictable syntactic environments. For instance, in English, determiners occur only at the beginning of noun phrases. To test this cross-linguistically, we analyze the *syntactic selectivity* of POS tags by modeling dependency trees as undirected graphs. We calculate the dependency entropy of the syntactic neighbors (i.e., co-dependents) for each POS category. Let  $X$  and  $Y$  be nodes that are in a syntactic dependency with each other, and ranging over a syntactic tagset  $\mathcal{T}$ . We calculate the syntactic selectivity as the conditional entropy:

$$H(X | Y = y) = - \sum_{x \in \mathcal{T}} p(x|y) \log_2 p(x|y) \quad (1)$$

where  $p(x|y)$  is the probability that a connected node has the POS tag  $x$ , given the base node has tag  $y$ . Probabilities are estimated via maximum likelihood estimation from our UD corpus. We then calculate the average entropy for a set of function word tags  $\mathcal{F} \subset \mathcal{T}$ ,

$$\bar{H}_{\mathcal{F}} = \frac{\sum_{f \in \mathcal{F}} \text{Freq}(f) H(X | Y = f)}{\sum_{f \in \mathcal{F}} \text{Freq}(f)} \quad (2)$$

and likewise calculate this metric for content words  $\bar{H}_C$ . Figure 1b illustrates the results. We observe

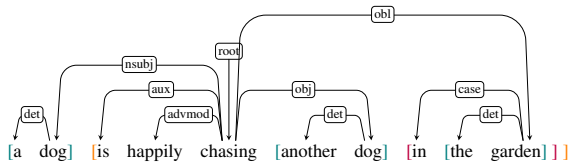


Figure 2: An example dependency tree with constituent spans highlighted using colored brackets: NP, VP, and PP. Phrase labels are theoretically neutral.

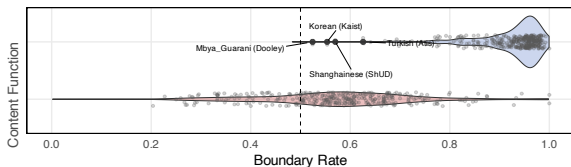


Figure 3: The boundary ratio of function and content words across languages. Function words consistently show higher boundary ratios across languages.

a consistent pattern across languages:  $\bar{H}_{\mathcal{F}}$  is always lower than  $\bar{H}_{\mathcal{C}}$ . This indicates that function words exhibit high syntactic selectivity, connecting to a restricted set of categories, whereas content words operate as high-entropy hubs with diverse connections.

**Phrase Boundary Alignment** We examine the degree to which words belonging to class  $c$  align with phrase boundaries ( $\frac{\sum_{w \in \mathcal{V}_c} \mathbf{1}\{w \text{ is at a boundary}\}}{|\mathcal{V}_c|}$ ). While a constituency-parsed dataset would be preferable, using one would require drastically reducing the number of languages considered. We therefore opt to use UD as it provides the most comprehensive collection of multilingual syntactic data currently available. We approximate constituent spans using the yield of dependency subtrees: For each target word, we identify the subtree governed by its syntactic head and check if the word appears at the subtree’s left or right periphery.

We apply two refinements to this heuristic. First, dependency trees do not perfectly map to constituent structures. For example, as shown in Figure 2, the auxiliary *is* depends on the verb *chasing*. A naive subtree yield of the verb phrase would place *is* internally (as *dog* is also a dependent to *chasing*), failing to capture its boundary status in the corresponding constituent. To mitigate such alignment errors, we exclude categories that often function as dependents of predicates (specifically PRON, PART, and AUX) and focus our analysis on ADP, DET, SCONJ, and CCONJ. For content words, we only focus on ADJ and NUM as other POSes are

all related to verbs.

Second, we adjust for nested structures, such as an NP embedded within a PP (e.g., *in [the garden]*). In standard UD, the adposition *in* and the determiner *the* both attach to the noun *garden*. A naive boundary check on the subtree of *garden* would identify *in* as the boundary, incorrectly labeling *the* as internal. To address this, we implement a relaxation rule: if a word is immediately preceded (or followed) by another word from the same function/content word class that marks the constituent boundary, we also consider the target word as a boundary marker.<sup>3</sup>

Figure 3 shows a strong cross-linguistic tendency: the median ratio of function words at phrase boundaries is 0.95 (minimum 0.55 in Korean), compared to only 0.58 for content words.

## 4 Which Properties of Function Words Support Syntactic Learning?

Having established that the three distributional properties of function words are cross-linguistically robust, we ask whether findings from human artificial language learning generalize to natural language and which of these properties matters most.

### 4.1 Training Data Construction

We construct our training corpora by manipulating text from Wikipedia.<sup>4</sup> We use Wikipedia instead of BabyLM (Hu et al., 2024) due to the lack of accurate parsers for child-directed speech (CDS), as one of our manipulations (WITHINBOUNDARY) requires reliable identification of function words and their syntactic heads.<sup>5</sup> All text is lowercased prior to training. We use Stanza (Qi et al., 2020) to obtain word-level POS tags and dependency parses for identifying function words and phrase boundaries during corpus manipulation.

<sup>3</sup>Although these adjustments reflect English-centric assumptions, they serve as a necessary approximation to maintain methodological consistency across diverse languages lacking gold-standard constituency resources.

<sup>4</sup><https://huggingface.co/datasets/wikimedia/wikipedia>

<sup>5</sup>One potential concern is that Wikipedia may exhibit a different distribution of function words compared to CDS, which could limit the generalizability of our findings to language acquisition. However, we find that Wikipedia and UD-CHILDES (Yang et al., 2025b) have comparable proportions of function words under the WITHINBOUNDARY transformation (CHILDES: 54%; Wikipedia: 55%).

## 4.2 Function Word Identification

Same as Experiment 1, we identify function words based on closed-class POS tags in UD. Specifically, we include DET (determiners), ADP (adpositions), CCONJ (coordinating conjunctions), SCONJ (subordinating conjunctions), and AUX (auxiliaries).

We explicitly exclude pronouns, degree quantifiers, and numerals. Pronouns often realize core argument positions; manipulating them would directly alter the argument structure of sentences, introducing a confound where degraded performance reflects missing arguments rather than missing structural cues. Numerals and degree quantifiers are excluded as they exhibit more variable and semi-lexical behavior across languages: they are often gradable, can function as independent predicates, and in many languages pattern more like adjectival or nominal modifiers rather than core determiners (Corver and van Riemsdijk, 2013).

To construct the inventory, we collect word types associated with the selected POS tags from the GUM (Zeldes, 2017) and EWT (Silveira et al., 2014) treebanks. After filtering items with fewer than 10 occurrences and obvious annotation errors, we obtain a final set of 116 English function words (see Appendix C).

## 4.3 Experimental Conditions

Building on prior work on distributional properties of function words (Valian and Coulson, 1988; Green, 1979; Morgan et al., 1987; Christophe et al., 2008), we generate manipulated corpus variants targeting function word properties that have been hypothesized to contribute to their status as cues for grammatical structure. To ensure comparability across conditions, content words and total dataset size are held constant; sentence length distributions are also matched across conditions (except **NOFUNCTION**). All experiments use English only (see Section 7 for discussion).

**High Lexical Frequency.** We manipulate lexical frequency by varying how token counts are distributed across function word types, while holding the total number of function word tokens constant (except in **NOFUNCTION**). Specifically, we vary the size of the function word inventory to induce variation in their lexical frequency:

- **STANDARD FUNCTION:** The natural English inventory (116 types).
- **NOFUNCTION:** A language with all function words removed.

- **FIVEFUNCTION:** A minimized inventory where all function words within a syntactic category (e.g., all determiners) are mapped to a single type (5 function word types in total).
- **MOREFUNCTION:** An expanded inventory where each natural function word is mapped to 10 distinct pseudowords based on their forms using Wuggy (Keuleers and Brysbaert, 2010), increasing the inventory size to  $\sim 1.2k$  types.

**Structural Predictability.** We manipulate the statistical dependency between function words and their context:

- **PHRASEDEPENDENCY:** The natural baseline, where function words reliably predict specific phrase structures.
- **BIGRAMDEP:** Function words are determined by the *following* word (i.e., the local bigram dependency) rather than by structural class.<sup>6</sup>
- **RANDOMDEP:** Function word location is preserved, but their identity is randomly shuffled.

**Phrase-boundary Alignment.** We manipulate the structural placement of function words:

- **ATBOUNDARY:** The natural baseline, where function words appear systematically at phrase boundaries.
- **WITHINBOUNDARY:** Function words are displaced from phrase boundaries to positions immediately adjacent to their syntactic heads. This minimizes dependency length but destroys boundary cues. This manipulation changes 55% of the location of function words in over 99% of the sentences in the training split.

As each baseline corresponds to natural English, we refer to this shared unmanipulated condition as **NATURALFUNCTION**. Examples are shown in Table 1, and the effects of each manipulation on the three properties are summarized in Appendix D.

## 4.4 Model & Training

We train GPT-2 Small models from scratch on each language variant. To ensure a fair comparison, a dedicated tokenizer is trained for each manipulated corpus to accommodate vocabulary changes. Each model is trained for 10 epochs. We report results

<sup>6</sup>We first construct a one-to-one mapping between vocabulary items plus a special unknown word and function words, keeping the number of function word types the same. When regenerating the corpus, function words are generated conditioned on the following word.

averaged over 3 random seeds. Detailed hyperparameters are provided in Appendix B. For each language, we also train a 5-gram model with Kneser-Ney smoothing using KenLM (Heafield, 2011).

#### 4.5 Evaluation

We do not use perplexity as an evaluation metric because our manipulations alter the entropy of the corpora. Instead, we evaluate syntactic generalization using the BLiMP benchmark (Warstadt et al., 2020).<sup>7</sup> Manipulated examples are shown in Table A.

To adapt BLiMP to our manipulated languages, we apply the same text transformations to the test sets. We implement a rigorous filtering protocol: (1) We remove categories where the critical word in the minimal pair is a function word (e.g., Determiner-Noun agreement), as these distinctions may be erased in conditions like NOFUNCTION. (2) We remove pairs that become identical after manipulation. (3) We apply intersection filtering: if a minimal pair is removed in any language condition, it is removed from the evaluation set for *all* models. This ensures that all models are evaluated on the exact same set of underlying syntactic phenomena. A list of excluded categories is in Appendix A.

#### 4.6 Results.

Results are shown in Table 2. Note, we focus on which distributional properties support better syntactic generalization, rather than on whether models acquire full human-like syntax. First,  $n$ -gram models consistently lag behind transformers across all conditions, and improve on BIGRAMDEP over NATURALFUNCTION. The increased scores for Transformers suggest that they capture aspects of syntactic knowledge that are not available to linear heuristic models.<sup>8</sup> Among the transformer models, NATURALFUNCTION achieves the highest accuracy, whereas NOFUNCTION performs worst, with the other conditions falling in between. A linear

<sup>7</sup>To make sure that Transformer models are able to learn the manipulated languages, we measured language model competence by asking each model to perform a 7-way forced-choice task on sentences consistent with its own training distribution vs. distractors, which tests whether it has learned the underlying grammar of its variant. The result is reported in Appendix G.

<sup>8</sup>We additionally evaluate several interaction conditions for both transformer and  $n$ -gram models (i.e., FIVEFUNCTION\_RANDOMDEP, FIVEFUNCTION\_WITHINBOUNDARY, BIGRAMDEP\_WITHINBOUNDARY). The two model classes show different sensitivity to these combined manipulations, further suggesting that transformers generalize beyond linear heuristic rules. Details are explained in Appendix H.

mixed-effects model ( $\text{acc} \sim \text{condition} + (1 \mid \text{category:phenomenon}) + (1 \mid \text{seed})$ ) shows that, relative to NATURALFUNCTION, all conditions have significant negative effects ( $p < 0.05$ ), except for FIVEFUNCTION, which is marginally worse than NATURALFUNCTION ( $p = 0.08$ ). These results suggest that disrupting any of the statistical properties of function words makes syntactic generalization harder to learn.

Additionally, we observe a Goldilocks effect in function word distributions: optimal learnability requires function words to be frequent enough to be reliable, yet sufficiently diverse to remain structurally informative. Collapsing the function word inventory into a small set of highly frequent items, as in FIVEFUNCTION, reduces the distinctiveness of these cues and leads to degraded learning despite preserving high frequency and structural association. Conversely, expanding the inventory excessively—which diversifies the inventory but reduces type frequency of function words, as in MOREFUNCTION—also harms learnability. These results suggest that successful grammatical learning requires a balance between the frequency and diversity of function words.

Effects for individual BLiMP categories are mostly negative, but not uniform. The large improvement for Filler Gap tests, especially in our FIVEFUNCTION setting, can be attributed to the fact that some of our manipulations simplify the dependencies between the filler and gap by inventories for fillers, which are function words.

**Summary.** We show that languages in which function words have all three properties are most learnable by LMs and not all distributional properties of function words contribute equally to learnability. Specifically, disrupting the structural association leads to a larger performance drop than manipulating lexical frequency or positional boundary. We also observe a Goldilocks effect in function word distribution: optimal learnability arises when function words are sufficiently frequent to be reliable, yet sufficiently diverse to support distinctions among grammatical structures.

## 5 How Are Function Words Represented and Deployed?

The previous experiments establish that function word distributions shape grammatical learning in our neural models. But do models actually rely on function word information when making gram-

Condition	Overall 5-gram	Overall Transformer	S-V Agr	Irregular	NPI	Island	Filter-Gap	Ellipsis	Ctrl.Rails	Binding	Arg-Str	Ana-Agr
NATURALFUNCTION	55.5	72.7	68.5	89.7	55.9	56.0	65.9	79.3	72.5	73.9	74.2	94.8
NOFUNCTION	54.1 (-1.4)	60.7 (-12.0)	54.0 (-14.5)	59.1 (-30.6)	50.3 (-5.6)	50.5 (-5.4)	44.3 (-21.6)	69.9 (-9.4)	63.7 (-8.8)	64.2 (-9.8)	59.4 (-14.9)	94.8 (+0.0)
FIVEFUNCTION	55.4 (-0.1)	70.9 (-1.8)	58.9 (-9.6)	89.7 (-0.1)	55.0 (-0.9)	54.9 (-1.0)	72.5 (+6.6)	77.1 (-2.2)	68.1 (-4.4)	72.7 (-1.2)	68.8 (-5.4)	95.3 (+0.5)
MOREFUNCTION	52.8 (-2.7)	69.7 (-3.0)	65.6 (-2.9)	88.7 (-1.0)	53.3 (-2.6)	56.5 (+0.6)	55.5 (-10.4)	72.6 (-6.7)	70.5 (-2.0)	73.2 (-0.8)	70.9 (-3.4)	94.0 (-0.8)
BIGRAMDEP	56.1 (+0.6)	67.4 (+5.3)	56.2 (-12.3)	85.9 (-3.8)	54.6 (-1.2)	50.6 (-5.4)	67.2 (+1.3)	70.2 (-9.1)	66.6 (-5.9)	66.7 (-7.3)	66.1 (-8.1)	93.4 (-1.4)
RANDOMDEP	53.4 (-2.1)	67.0 (-5.7)	56.0 (-12.5)	84.6 (-5.2)	53.1 (-2.8)	52.6 (-3.3)	60.8 (-5.1)	74.9 (-4.5)	65.3 (-7.2)	67.7 (-6.3)	64.1 (-10.2)	95.4 (+0.6)
WITHINBOUNDARY	54.5 (-1.0)	69.7 (-3.0)	67.1 (-1.4)	83.8 (-6.0)	47.2 (-8.6)	51.5 (-4.5)	67.8 (+1.8)	72.8 (-6.6)	70.7 (-1.8)	70.1 (-3.8)	73.4 (-0.9)	95.8 (+1.1)

Table 2: BLiMP accuracy by training condition, averaged over 3 seeds. Deltas are relative to NATURALFUNCTION.

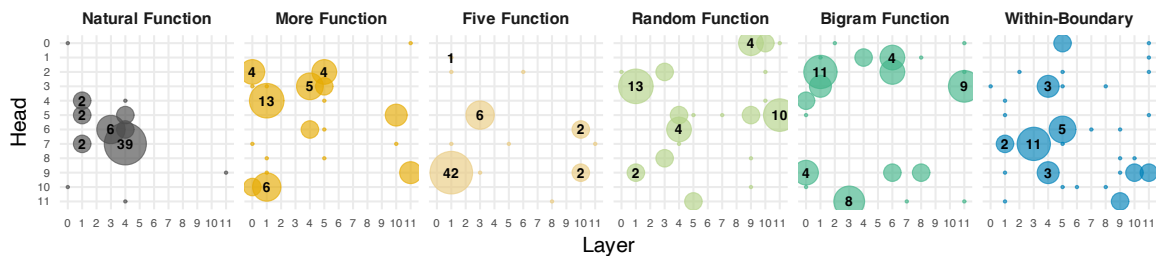


Figure 4: Distribution of dominant function heads across BLiMP categories under different training conditions (seed=53). Numbers indicate the top-5 heads and the number of BLiMP categories in which each head assigns the highest attention to function words. Results for other seeds are reported in Appendix 6.

maticity judgments? If we assume that information actively used at processing time reflects what was encoded during training, then demonstrating processing-time dependence on function words would provide convergent evidence that these distributional properties genuinely drove acquisition rather than serving as incidental correlates of other statistical patterns in the data. We therefore combine attention probing and function word ablation to test this directly.

### 5.1 Probing Function Word Attention

**Experiment Setup.** We probe attention patterns to examine whether successful learning is associated with specialized representations for function words. We apply the probing method of Aoyama and Wilcox (2025) to identify attention heads that consistently attend to function words across BLiMP categories (see Appendix E for details). Models trained without function words (NOFUNCTION) are excluded from this analysis.

**Results.** Figure 4 shows a clear divergence across conditions. Models trained with NATURALFUNCTION exhibit a highly concentrated pattern of function word attention: a small number of heads, primarily in Layers 3 and 4, account for the majority of function word attention across BLiMP categories. This pattern indicates the emergence of specialized representations tied to function words when all three distributional properties are present.

In contrast, conditions that disrupt structural predictability (RANDOMDEP or BIGRAMDEP) or phrase boundaries (WITHINBOUNDARY) do not produce similarly concentrated representations. The FIVEFUNCTION condition shows a more concentrated pattern, likely because of the limited function vocabulary ( $n=5$ ). Attention to function words in these models is diffuse, suggesting a link between how these words are represented internally and the grammatical abilities of the models.

To confirm that this observation is not due to chance, we visualize the function heads from two additional random seeds in Appendix F. We further aggregate the number of function heads across all heads and layers, and compute the entropy of this distribution across three random seeds. The results are reported in Table 10 in Appendix I. We find that NATURALFUNCTION exhibits lower mean entropy and standard deviation, indicating that function word representations are concentrated in a small number of heads and that this pattern is stable across random seeds.

### 5.2 Ablating function word Information

**Experiment Setup.** If models have learned to exploit function word information, then disrupting this information at evaluation time should selectively impair performance, depending on how strongly models rely on such cues. We therefore experiment with two types of ablations which we call

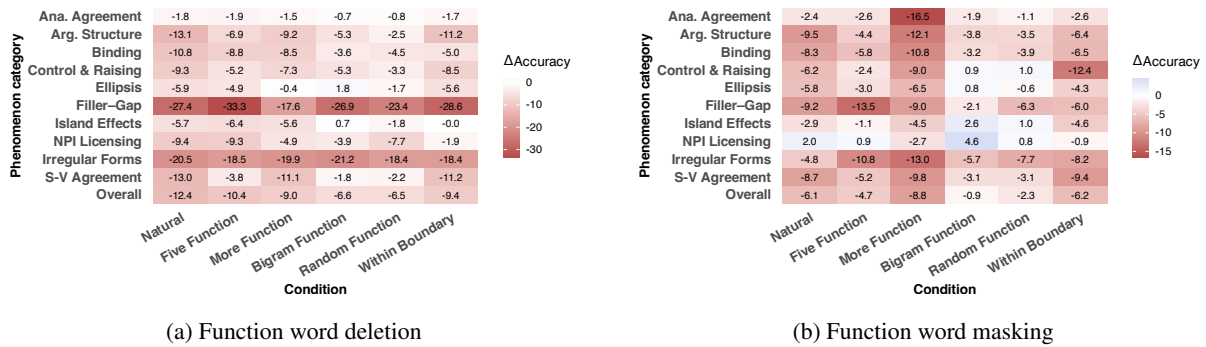


Figure 5: Difference in BLiMP accuracy before and after ablation.

function word *masking* and function word *deletion*. In function word masking, we block attention *to* and *from* function word tokens in their corresponding BLiMP benchmark during evaluation, thus turning them into content-free placeholders. In function word deletion, we remove function words entirely from the input by evaluating on **NOFUNCTION** BLiMP. Models trained without function words (**NOFUNCTION**) are excluded from this analysis.

**Results.** Results are summarized in Figure 5. Overall, function word deletion leads to a larger drop than function word masking, meaning that positional information of function words are helpful for LMs to process texts. In the function word deletion experiment, models trained on **NATURALFUNCTION** show the largest performance drop, followed by **FIVEFUNCTION** and **WITHINBOUNDARY**, indicating that these models rely heavily on function word information for making judgments. The **MOREFUNCTION** condition leads to the largest drop in the function word masking experiment, likely because the increased number of function words creates stronger associations between more diverse function words and specific structural contexts, making their identity particularly informative. In contrast, **BIGRAMDEP** and **RANDOMDEP** show the smallest performance drops for both experiments. This suggests that the association between function words and their underlying structural roles is disrupted, models become less reliant on function words.

For both experiments, we perform paired *t*-tests comparing each ablated model with its corresponding non-ablated baseline. In the deletion experiment, all conditions show significant drops ( $p < 0.01$ ). In the masking experiment, performance decreases significantly in most conditions ( $p < 0.05$ ), except for **BIGRAMDEP** and **MORE-**

**FUNCTION**. However, the masking results are less stable across runs, indicating that models do not consistently rely on function word identity under this manipulation.

**Summary.** Models trained on different language variants differ systematically in how they represent and deploy function words. Disrupting any of the three properties prevents the emergence of focused function word attention heads and decrease the reliance on function words. In particular, disrupting reliable structural association leads to weaker reliance on function words overall.

## 6 Discussion

Beyond replicating earlier findings from natural language acquisition (e.g., Valian and Coulson, 1988; Morgan et al., 1987), our computational approach extends prior artificial-language-based studies to more complex naturalistic text and refines existing function word related hypotheses. We show that not all statistical cues contribute equally to grammatical learning in neural learners. In particular, high frequency and reliable structural association exert stronger effects on both learning outcomes and internal representations than phrase-boundary alignment.

A natural follow-up question is why these properties matter and why they exhibit a graded pattern of influence. Here we offer a tentative interpretation, grounded in the prior literature on linguistic typology and language acquisition.

We first propose that these properties matter because they maximize learnability. Cognitive scientists have argued that aspects of language structure may result from the particular ways in which humans learn languages (Christiansen and Chater, 2008; Culbertson et al., 2012; Newport, 2016). According to these proposals, languages are shaped by

selectional pressures from successive generations of learners, with learnability being one key selective pressure among others (e.g., communication, compression, see Kirby et al., 2015; Kirby, 1999). Languages that are easier to learn are more likely to be transmitted successfully across generations.

With the same amount of input, language learnability emerges from the interaction between learner-internal factors and language-specific properties. Languages can evolve structural features that facilitate learning, while learners bring inductive biases that privilege such patterns, accelerating acquisition. In our experiment, we treat transformer LMs as weakly biased, domain-general statistical learners and use them to re-examine claims from the acquisition literature under controlled conditions. Our results show that languages are most learnable by neural learners when they exhibit all three function word properties. Critically, all three properties are attested in the 186 languages we examined, suggesting they represent convergent evolutionary solutions to the learnability problem. This claim about learnability aligns with studies of human learners, who tend to focus distributional analysis on function words and learn the predictable syntactic distribution of a function word more easily than the equally predictable distribution of a content word (Getz and Newport, 2019). This learning asymmetry aligns with the statistical asymmetry we observe: function words reliably predict syntactic structure, making them effective anchors for grammar learning. Therefore, both the statistical regularities in natural language and human inductive biases contribute to learning efficiency.

Second, we address why these three properties facilitate learning in the first place, and why they exhibit the observed graded pattern of influence. Following Gerken (1987); Gerken and McIntosh (1993), a long-standing proposal in the acquisition literature distinguishes two foundational challenges prior to syntactic analysis: segmentation, identifying constituent boundaries in the input, and labeling, determining the syntactic type of those constituent types.<sup>9</sup> Viewed through this lens, the three distributional properties we study may support learning in distinct ways. High frequency makes function words salient and trackable, facilitating their availability as learning cues, which

may help explain why Zipfian distributions support language learning and generalization (e.g., Lavi-Rotbain and Arnon, 2022; Wolters et al., 2024). Phrase-boundary alignment plausibly contributes to segmentation by providing positional cues in linear input. Reliable structural association, by contrast, supports labeling by consistently linking function words to particular constituent types, consistent with evidence that function words cue grammatical category learning (Mintz, 2003; Zhang et al., 2015). Consistent with this interpretation, we find that disrupting structural association leads to larger learning costs than disrupting boundary alignment, suggesting that labeling-related information plays a more central role for function words, while segmentation cues may be partially recovered from other sources, such as transition statistics. In summary, by providing the segmentation and labeling information, function words thus serve as anchors to specific syntactic structures and categories, bootstrapping grammar learning.

## 7 Conclusion

In this study, we investigated how function words contribute to syntactic learning using corpus analysis and computational simulations. Building on generalizations about function words originally discovered through linguistic analysis, we identified three distributional properties of function words (i.e., high frequency, reliable structural association, and systematic alignment with phrase boundaries) and showed that they are robustly attested across 186 languages. Using controlled manipulations and probing analyses, we demonstrated that these properties differentially support grammatical learning in transformer LMs. Languages that preserve all three properties are learned most effectively, mirroring classic findings from studies on human language learning. In addition, disrupting individual properties incurs learning costs. Among the properties we studied, reliable structural association shows stronger influences than boundary alignment and frequency, suggesting that not all statistical cues contribute equally to learning. Finally, we interpret these findings through the framework of cultural evolution: function words act as *anchors* that have been selected for during cultural transmission to maximize learnability, allowing learners to bootstrap complex syntactic structure from limited input via the segmentation and labeling information they provide.

<sup>9</sup>Related distinctions have also been noted in earlier theoretical work (Fodor and Garrett, 1967; Clark and Clark, 1977).

## Limitations

First, our modeling experiments focus exclusively on English. While we conduct a cross-linguistic analysis across 186 languages to argue for the universality of function word properties, the counterfactual modeling experiments are limited to English for methodological reasons. Our experimental design requires: (i) high-quality dependency parsing to identify phrase boundaries and implement the **WITHINBOUNDARY** manipulation, and (ii) a standardized benchmark for evaluating syntactic generalization. English uniquely satisfies both requirements with reliable parsing tools (Stanza) and the BLiMP benchmark, enabling the controlled comparisons central to our research questions. Although our cross-linguistic analysis demonstrates that the three distributional properties hold robustly across typologically diverse languages, testing whether the absolute magnitude of these effects varies with language-specific factors (e.g., word order, morphological richness) remains an important direction for future work.

Second, all of our experiments focus on word-level function categories and do not consider grammatical morphology. In many languages, like Turkish, function-like information is encoded in bound morphemes rather than free function words. UD also ignores the morphology level annotation. Investigating whether and how such morphological markers support syntactic learning and inference would be an important extension of the present study.

Third, prior work in language acquisition emphasizes the role of prosodic cues, such as stress and rhythm, in helping learners distinguish function words from content words (e.g., Morgan et al., 1987). Our experiments, however, rely exclusively on written text. It remains to be seen whether the same patterns hold when models are trained on speech input, where prosodic information may provide additional structural cues.

Fourth, the language **MOREFUNCTION** introduces potential randomness for each function word, as each pseudo-function word corresponding to a function word is randomly assigned. This design disrupts the natural predictive associations between function words and content words found in actual language. Because these pseudo-function words lack natural collocational histories, it is unclear how they would associate with particular syntactic structures in a realistic linguistic setting. However,

this does not mean that **MOREFUNCTION** is similar to **RANDOMDEP**. In **MOREFUNCTION**, the pseudo words are based on the original function word, which means that if a pseudo word appears in one context associated with one function word, it cannot appear in another context. We believe that despite this simplification, the current formulation still provides useful insights and serves as a first step toward understanding the role of function words in structured representations.

Finally, it is inherently difficult to determine whether our experimental setup, i.e., evaluating LMs’ syntactic generalization using minimal pairs, captures genuine syntactic knowledge like humans or reliance on other heuristic strategies. While BLiMP covers a broad range of linguistic phenomena and the test sentences are distinct from the training data, such evaluations cannot fully rule out alternative explanations based on surface-level cues. For example, a 5-gram model can reach above-chance accuracy as shown in our experiment. Therefore, we treat BLiMP as a useful **proxy** for syntactic generalization in LMs, while acknowledging that success on this benchmark does not imply that models acquire grammatical knowledge in the same way as humans.

## Ethical Considerations

Wikipedia data are publicly available and released under open licenses, including the GNU Free Documentation License (GFDL) and the Creative Commons Attribution–ShareAlike 4.0 License (CC BY-SA 4.0). The UD treebanks used in this study are released under the CC BY-SA 4.0 license. None of the data contain personally identifiable information. For certain UD treebanks, we exclude them due to copyright restrictions. We do not anticipate any potential risks associated with our experiments.

## Acknowledgements

We thank Elissa Newport, Najoung Kim, and members of NERT lab for their helpful discussion and suggestions. Special thanks to anonymous reviewers and AC for their constructive feedback.

## References

- Steven P Abney. 1987. *The English noun phrase in its sentential aspect*. Ph.D. thesis, Massachusetts Institute of Technology.
- Tatsuya Aoyama and Ethan Wilcox. 2025. [Language models grow less humanlike beyond phase transition](#).

- In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24938–24958, Vienna, Austria. Association for Computational Linguistics.
- Roger Brown. 1973. A first language: The early stages. In *A first language*. Harvard University Press.
- Tina Bögel. 2021. [Function words at the interface: A two-tier approach](#). *Languages*, 6(4).
- Greg N Carlson. 1983. [Marking constituents](#). In *Linguistic categories: Auxiliaries and related puzzles: Volume one: Categories*, pages 69–98. Springer.
- Emmanuel Chemla, Toben H Mintz, Savita Bernal, and Anne Christophe. 2009. [Categorizing words using ‘frequent frames’: What cross-linguistic analyses reveal about distributional acquisition strategies](#). *Developmental Science*, 12(3):396–406.
- Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024. [Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs](#). In *The Twelfth International Conference on Learning Representations*.
- Morten H Christiansen and Nick Chater. 2008. Language as shaped by the brain. *Behavioral and brain sciences*, 31(5):489–509.
- Anne Christophe, Séverine Millotte, Savita Bernal, and Jeffrey Lidz. 2008. [Bootstrapping lexical and syntactic acquisition](#). *Language and Speech*, 51(1-2):61–75.
- Herbert H. Clark and Eve V. Clark. 1977. *Psychology and Language: An Introduction to Psycholinguistics*. Harcourt Brace Jovanovich, New York.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What does BERT look at? an analysis of BERT’s attention](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- Norbert Corver and Henk van Riemsdijk. 2013. *Semi-lexical categories: The function of content words and the content of function words*, volume 59. Walter de Gruyter.
- Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. 2012. Learning biases predict a word order universal. *Cognition*, 122(3):306–329.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Cristina Dye, Yarden Kedar, and Barbara Lust. 2019. [From lexical to functional categories: New foundations for the study of language development](#). *First Language*, 39(1):9–32.
- Allyson Ettinger. 2020. [What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models](#). *Transactions of the Association for Computational Linguistics*, 8:34–48.
- Jerry A Fodor and Merrill Garrett. 1967. Some syntactic determinants of sentential complexity. *Perception & Psychophysics*, 2(7):289–296.
- LouAnn Gerken and Bonnie J McIntosh. 1993. Interplay of function morphemes and prosody in early language. *Developmental psychology*, 29(3):448.
- Louann Gerken, Rachel Wilson, and William Lewis. 2005. Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, 32(2):249–268.
- LouAnn A. Gerken. 1987. Telegraphic speaking does not imply telegraphic listening. *Papers and Reports on Child Language Development*, 26:48–55.
- Heidi R Getz and Elissa L Newport. 2019. Privileged computations for closed-class items in language acquisition. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 41.
- Thomas RG Green. 1979. The necessity of syntax markers: Two experiments with artificial languages. *Journal of Verbal Learning and Verbal Behavior*, 18(4):481–496.
- Ariel Gutman, Isabelle Dautriche, Benoît Crabbé, and Anne Christophe. 2015. Bootstrapping the syntactic bootstrapper: Probabilistic labeling of prosodic phrases. *Language Acquisition*, 22(3):285–309.
- Martin Haspelmath. 2005. *The world atlas of language structures*. Oxford University Press.
- Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.
- Jessica Peterson Hicks. 2006. *The impact of function words on the processing and acquisition of syntax*. Ph.D. thesis, Northwestern University.
- Jean-Rémy Hochmann, Ansgar D Endress, and Jacques Mehler. 2010. Word frequency as a cue for identifying function words in infancy. *Cognition*, 115(3):444–457.
- Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Leshem Choshen, Ryan Cotterell, Alex Warstadt, and Ethan Gotlieb Wilcox, editors. 2024. *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Miami, FL, USA.
- Mark Johnson, Anne Christophe, Emmanuel Dupoux, and Katherine Demuth. 2014. [Modelling function words improves unsupervised word segmentation](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

- Long Papers*), pages 282–292, Baltimore, Maryland. Association for Computational Linguistics.
- 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.
- Richard S Kayne and 1 others. 2005. *Movement and silence*, volume 10. Oxford University Press Oxford.
- Yarden Kedar, Marianella Casasola, and Barbara Lust. 2006. Getting there faster: 18- and 24-month-old infants’ use of function words to determine reference. *Child Development*, 77(2):325–338.
- Emmanuel Keuleers and Marc Brysbaert. 2010. [Wuggy: A multilingual pseudoword generator](#). *Behavior Research Methods*, 42(3):627–633.
- Najoung Kim, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R. Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. [Probing what different NLP tasks teach machines about function word comprehension](#). In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (\*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.
- John Kimball. 1973. [Seven principles of surface structure parsing in natural language](#). *Cognition*, 2(1):15–47.
- Simon Kirby. 1999. *Function, selection, and innateness: The emergence of language universals*. OUP Oxford.
- Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.
- Ori Lavi-Rotbain and Inbal Arnon. 2022. The learnability consequences of zipfian distributions in language. *Cognition*, 223:105038.
- Xiaomeng Ma and Qihui Xu. 2025. [Implicit in-context learning: Evidence from artificial language experiments](#). In *Second Conference on Language Modeling*.
- Toben H Mintz. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.
- Toben H Mintz. 2006. Finding the verbs: Distributional cues to categories available to young learners. *Action meets word: How children learn verbs*, 1:31–63.
- Toben H. Mintz, Elissa L. Newport, and Thomas G. Bever. 2002. [The distributional structure of grammatical categories in speech to young children](#). *Cognitive Science*, 26(4):393–424.
- 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.
- James L Morgan, Richard P Meier, and Elissa L Newport. 1987. Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, 19(4):498–550.
- Elissa L Newport. 2016. Statistical language learning: Computational, maturational, and linguistic constraints. *Language and Cognition*, 8(3):447–461.
- 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.
- Eva Portelance, Michael C Frank, and Dan Jurafsky. 2024. Learning the meanings of function words from grounded language using a visual question answering model. *Cognitive Science*, 48(5):e13448.
- 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.
- Luigi Rizzi and Guglielmo Cinque. 2016. Functional categories and syntactic theory. *Annual Review of Linguistics*, 2:139–163.
- Alexa R Romberg and Jenny R Saffran. 2010. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):906–914.
- Wesley Scivetti, Tatsuya Aoyama, Ethan Wilcox, and Nathan Schneider. 2025. [Unpacking Let Alone: Human-scale models generalize to a rare construction in form but not meaning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27503–27514, Suzhou, China. Association for Computational Linguistics.
- Elisabeth Selkirk. 2014. The prosodic structure of function words. In *Signal to Syntax*, pages 187–213. Psychology Press.
- Rushen Shi, James L Morgan, and Paul Allopenna. 1998. Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of child language*, 25(1):169–201.
- Rushen Shi, Janet F Werker, and Anne Cutler. 2006. Recognition and representation of function words in english-learning infants. *Infancy*, 10(2):187–198.

- Rushen Shi, Janet F Werker, and James L Morgan. 1999. Newborn infants’ sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72(2):B11–B21.
- Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.
- Susan P Thompson and Elissa L Newport. 2007. Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, 3(1):1–42.
- Virginia Valian and Seana Coulson. 1988. Anchor points in language learning: The role of marker frequency. *Journal of memory and language*, 27(1):71–86.
- 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.
- Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848.
- Lucie Wolters, Ori Lavi-Rotbain, and Inbal Arnon. 2024. Zipfian distributions facilitate children’s learning of novel word-referent mappings. *Cognition*, 253:105932.
- Xiulin Yang, Tatsuya Aoyama, Yuekun Yao, and Ethan Wilcox. 2025a. **Anything goes? a crosslinguistic study of (im)possible language learning in LMs**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26058–26077, Vienna, Austria. Association for Computational Linguistics.
- Xiulin Yang, Arianna Bisazza, Nathan Schneider, and Ethan Gotlieb Wilcox. 2026. A unified assessment of the poverty of the stimulus argument for neural language models. *arXiv preprint arXiv:2602.09992*.
- Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu, and Nathan Schneider. 2025b. **UD-English-CHILDES: A collected resource of gold and silver Universal Dependencies trees for child language interactions**. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 52–58, Ljubljana, Slovenia. Association for Computational Linguistics.
- Qing Yao, Kanishka Misra, Leonie Weissweiler, and Kyle Mahowald. 2025. Both direct and indirect evidence contribute to dative alternation preferences in language models. *arXiv preprint arXiv:2503.20850*.
- Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.
- Zhao Zhang, Rushen Shi, and Aijun Li. 2015. Grammatical categorization in mandarin-chinese-learning infants. *Language acquisition*, 22(1):104–115.
- Imry Ziv, Nur Lan, and Emmanuel Chemla. 2026. **Biasless language models learn unnaturally: How LLMs fail to distinguish the possible from the impossible**. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5393–5403, Rabat, Morocco. Association for Computational Linguistics.

## A BLiMP Categories Considered

We list the categories that are removed from our experiments because the function word is the key to making a correct judgment. The list is reported in Table 8.

## B Training Data Statistics and Training Details

Statistic	Value
Average sentence length	25.34
Total number of sentences	3,519,842
Total number of words	89,192,973

Table 3: Overall corpus statistics.

Split	Number of Words
Training set	76,007,866
Development set	108,728
Test set	217,241

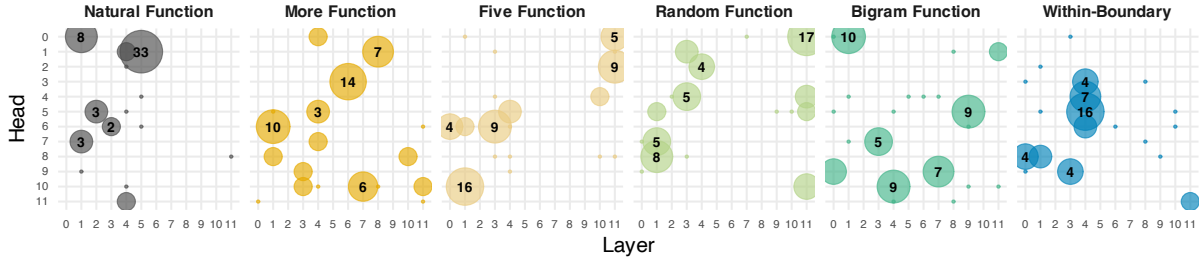
Table 4: Train, development, and test split sizes in number of words.

## C Function Word Inventory

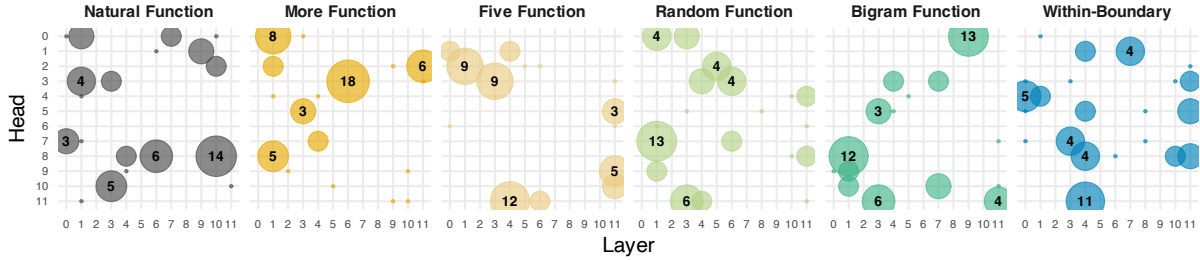
Function words are collected from the gold annotations of GUM and EWT with manual corrections. We also filter function word pieces that are less frequent than 10 times (e.g., *wilt*).

## D Manipulated Properties in Each Language

See Table 7.



(a) Random seed = 42



(b) Random seed = 67

Figure 6: Distribution of the dominant function head across BLiMP categories under different training conditions. Numbers in each subplot indicate the top-5 most frequent heads and the number of categories for which each head assigns the highest attention to function words.

Model	
model_type	GPT2-small
seeds	42, 53, 67
Tokenizer	
vocab_size	32,768
tokenization_alg	BPE
training_data	Wiki train split
Training	
num_epoch	10
batch_size	128
context_length	128
grad_acc_steps	1
weight_decay	0.1
warmup_steps	10%
lr	5e-4
lr_scheduler	linear
GPU	Tesla V100

Table 5: Model and training parameters

Category	Function Words
DET	the, this, a, an, no, all, another, each, that, any, those, these, both, every, either, neither
CCONJ	and, but, or, yet
SCONJ	that, if, although, after, whereas, while, before, as, though, until, because, since, once, whether, unless, albeit, till, whilst
AUX	will, be, had, were, being, is, would, was, do, could, are, have, been, has, did, should, might, can, does, 's, may, must, ca, am, shall, art, ar, re, ought, need
ADP	at, in, of, near, for, by, to, with, on, from, behind, into, within, despite, against, as, over, than, during, about, between, among, except, through, around, after, like, off, without, under, before, throughout, unlike, across, toward, along, above, aboard, until, upon, via, beneath, unto, beyond, per, below, amongst, till, beside, amid, onto, towards, underneath, alongside

Table 6: Function word inventories used in the experiments, grouped by syntactic category.

## E Attention Probing Details

Following [Aoyama and Wilcox \(2025\)](#), we adapt attention probing techniques originally proposed by [Clark et al. \(2019\)](#) and [Chen et al. \(2024\)](#) to autoregressive models such as GPT-2. In contrast to prior work, we focus specifically on dependency relations involving function words.

For a given attention head  $h$  at layer  $l$ , we define a head-specific probe  $f_{h,l}$  that predicts the syntactic

parent of a target word  $x_i$  by selecting the word  $x_j$  receiving the strongest attention connection with  $x_i$ , excluding self-attention:

$$f_{h,l}(x_i) = \arg \max_{j \neq i} a_{ij}^{(h,l)}, \quad (3)$$

where  $a_{ij}^{(h,l)}$  denotes the word-level attention

Language	Lexical Frequency	Structural Association	Phrase Boundary
NATURALFUNCTION	✓ high	✓ preserved	✓ aligned
NOFUNCTION	✗ zero	✗ none	✗ none
FIVEFUNCTION	✓ ✓ very high	✓ preserved	✓ aligned
MOREFUNCTION	✗ low	✓ preserved	✓ aligned
BIGRAMDEP	✓ high	✗ destroyed	✓ aligned
RANDOMDEP	✓ high	✗ destroyed	✓ aligned
WITHINBOUNDARY	✓ high	✓ preserved	✗ disrupted
FIVEFUNCTION_WITHINBOUNDARY	✓ ✓ very high	✓ preserved	✗ disrupted
FIVEFUNCTION_RANDOMDEP	✓ ✓ very high	✗ destroyed	✓ aligned
BIGRAMDEP_WITHINBOUNDARY	✓ high	✗ destroyed	✗ disrupted

Table 7: The manipulated properties in different counterfactual languages.

Phenomenon group	Sub-phenomenon
Determiner-noun agreement	determiner_noun_agreement_1
	determiner_noun_agreement_2
	determiner_noun_agreement_irregular_1
	determiner_noun_agreement_irregular_2
	determiner_noun_agreement_with_adjective_1
	determiner_noun_agreement_with_adjective_2
	determiner_noun_agreement_with_adj_irregular_1
determiner_noun_agreement_with_adj_irregular_2	
NPI licensing	matrix_question_npi_licensor_present
Quantifiers	existential_there_quantifiers_1
	existential_there_quantifiers_2
	superlative_quantifiers_1
	superlative_quantifiers_2

Table 8: BLiMP phenomenon groups and their corresponding sub-phenomena removed in our experiments (13 subcategories in total).

strength between  $x_i$  and  $x_j$ , defined as the maximum of attention from  $x_i$  to  $x_j$  and from  $x_j$  to  $x_i$  for head  $h$  at layer  $l$ .

Because we use BPE tokenization, we convert token-level attention weights to word-level attention scores. When attending *to* a split word, we sum attention weights over its constituent tokens; when attending *from* a split word, we average attention weights across its tokens.

We quantify the extent to which a given head attends to function words using a *function attention score*. Let  $F$  denote the set of function words and let  $N$  be the number of evaluated target words. The function attention score for head  $h$  at layer  $l$  is defined as:

$$S_F(h, l) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[f_{h,l}(x_i) \in F]. \quad (4)$$

Finally, we identify the specific head ( $h^*$ ,  $l^*$ ) that maximizes this function attention score for each subcategory in BLiMP.

Language	Score
NATURALFUNCTION	0.9681
NOFUNCTION	0.9748
MOREFUNCTION	0.0634
FIVEFUNCTION	0.9919
BIGRAMDEP	0.9885
RANDOMDEP	0.1602
WITHINBOUNDARY	0.9650

Table 9: Performance scores across different language conditions.

## F Function Head Distribution Results for the Other Two Random Seeds

See Figure 6.

## G Language learning results

The result is reported in Table G. These results demonstrate that for the majority of conditions, the LMs are highly competent at their respective languages, indicating that their performance drops on BLiMP to a failure in syntactic induction rather than a lack of general language modeling competence. For **RANDOMDEP** and **MOREFUNCTION**, the low scores indicate these variants are significantly harder to learn as a language.

Assigning high probability to in-domain data does not necessarily guarantee good syntactic generalization. What is interesting is that **FIVEFUNCTION** shows the highest score, meaning the language model learns such a distribution the best. However, as shown in Table 2, **FIVEFUNCTION** does not lead to the best BLiMP generalization.

## H Interaction effect of the statistical properties of function words

We further train languages `FIVEFUNCTION_WITHINBOUNDARY`, `FIVEFUNCTION_RANDOMDEP`, and `BIGRAMDEP_WITHINBOUNDARY` to examine the interaction between frequency, structure, and boundary positions of function words. For the transformer model, we use the same hyperparameters on random seed 53. We also train 5-gram models with the same settings in the main text. The results are shown in Figure 7.

First, we find that transformer models are sensitive to all three distributional properties and their interactions. Boundary alignment, frequency, and structural dependency each have clear effects on performance. The effects are largely additive in the frequency condition, but we observe a modest interaction between structure and boundary alignment, with the boundary advantage reduced under the `BIGRAMDEP` condition. In addition, randomization has a strong negative effect, particularly under natural frequency distributions compared with `FIVEFUNCTION` condition, indicating that disrupting the function word system is more harmful when distributional cues are informative.

In contrast, n-gram models exhibit minimal sensitivity to these manipulations. Performance remains largely unchanged across frequency and structural conditions, and boundary effects are weak and inconsistent. Only the randomization condition shows some variation, but the pattern differs from that observed in transformer models: a randomized language with only five function words is better than a five-function language without randomization. This contrast suggests that transformer models capture structured dependencies among function words, whereas n-gram models rely primarily on local statistical patterns.

## I Function head results

For the attention-head masking experiment, we constructed a  $12 \times 12$  layer-head frequency matrix for each condition, linearized it, and computed the entropy of the resulting distribution. The averaged results are reported in Table 10 and per-seed result in Table 11.

`NATURALFUNCTION` achieves the lowest mean entropy overall. We also observe that models trained with seed 67 consistently exhibit higher entropy across all conditions, suggesting relatively

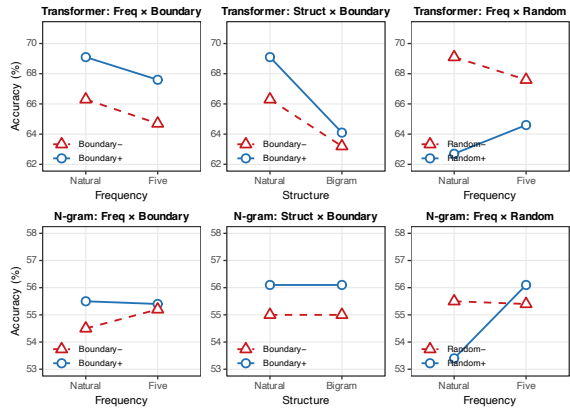


Figure 7: The interaction among the tree properties

sparser function-head representations.

Condition	Mean Entropy (bits)	SD
<code>NATURALFUNCTION</code>	<b>2.74</b>	0.861
<code>MOREFUNCTION</code>	3.53	0.263
<code>FIVEFUNCTION</code>	2.87	0.924
<code>RANDOMDEP</code>	3.64	0.156
<code>BIGRAMDEP</code>	3.60	0.257
<code>WITHINBOUNDARY</code>	4.02	0.463

Table 10: Mean entropy and standard deviation across different conditions. `NATURALFUNCTION` has the lowest mean entropy, indicating that its function head distribution is more concentrated.

Condition	Seed 42	Seed 53	Seed 67
<code>NATURALFUNCTION</code>	<b>2.47</b>	2.05	3.71
<code>MOREFUNCTION</code>	3.57	3.78	3.26
<code>FIVEFUNCTION</code>	3.42	<b>1.81</b>	3.40
<code>RANDOMDEP</code>	3.47	3.69	3.77
<code>BIGRAMDEP</code>	3.71	3.78	<b>3.31</b>
<code>WITHINBOUNDARY</code>	3.55	4.47	4.03

Table 11: Results across different random seeds.

Language	Sentence (correct)	Sentence (incorrect)
NATURALFUNCTION NOFUNCTION FIVEFUNCTION MOREFUNCTION BIGRAMDEP RANDOMDEP WITHINBOUNDARY	brian thought that rebecca was criticizing him . brian thought rebecca criticizing him . brian thought the rebecca will criticizing him . brian thought blat rebecca wam criticizing him . brian thought albeit rebecca with criticizing him brian thought since rebecca until criticizing him . brian thought rebecca was that criticizing him .	brian thought that rebecca was criticizing himself . brian thought rebecca criticizing himself . brian thought the rebecca will criticizing himself . brian thought blat rebecca wam criticizing himself . brian thought albeit rebecca with criticizing himself . brian thought since rebecca until criticizing himself . brian thought rebecca was that criticizing himself .
FIVEFUNCTION_WITHINBOUNDARY FIVEFUNCTION_RANDOMDEP BIGRAMDEP_WITHINBOUNDARY	brian thought rebecca that that criticizing him . brian thought that rebecca that criticizing him . brian thought rebecca albeit with criticizing him .	brian thought rebecca that that criticizing himself . brian thought that rebecca that criticizing himself . brian thought rebecca albeit with criticizing himself .

Table 12: Minimal pair examples in different created BLiMP benchmarks. Example sentence: *Brian thought that Rebecca was criticizing him.*