From Linear Input to Hierarchical Structure: Function Words as Statistical Cues for Language Learning

Anonymous ACL submission

Abstract

What statistical conditions support learning hierarchical structure from linear input? In this paper, we address this question by focusing on the statistical distribution of function words. Function words have long been argued to play a crucial role in language acquisition due to their distinctive distributional properties, including high frequency, reliable association with syntactic structure, and alignment with phrase boundaries. We use cross-linguistic corpus analysis to first establish that all three properties are present across 186 studied languages. Next, we use a combination of counterfactual language modeling and ablation experiments to show that language variants preserving all three properties are more easily acquired by neural learners, with frequency and structural association contributing more strongly than boundary alignment. Follow-up probing and ablation analyses further reveal that different learning conditions lead to systematically different reliance on function words, indicating that similar performance can arise from distinct internal mechanisms.[1]

1 Introduction

A central puzzle in language acquisition is how learners abstract grammatical structure from linear input. Extensive research on statistical learning argues that 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, highlighting 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). If these properties play a systematic role in learning, 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 natural texts (RQ2.1) and which properties most influence learning (RQ2.2). Furthermore, if these statistical properties shape learning, do they necessarily result in the same strategy for deploying function words during inference? (RQ3) Addressing this question would clarify whether there are multiple routes to successful acquisition and whether certain statistical cues are redundant.

We answer RQ1 by conducting a cross-linguistic analysis using data from the Universal Dependencies (UD) project (de Marneffe et al., 2021), testing whether the three distributional properties attributed to function words are universal across 186 languages in our sample. To address RQ2, in line with a growing body of work that uses neural language models as tools for theory specification and evaluation (Pearl and Sprouse, 2015; Lappin and Shieber, 2007), we take transformer models as domain-general and weakly-biased learn-

[1] Code and models will be released once accepted.

ers (Wilcox et al., 2024) and train them on counterfactual variants of natural text in which each distributional property of function words is systematically manipulated. To address RQ3, we conduct attention probing and function-word-related ablation experiments to examine how models internally represent and deploy function-word information.

We confirm that a language is most learnable when all three statistical properties of function words jointly hold, but we show that these properties do not contribute equally. Specifically, they contribute to language learnability with a clear hierarchy: lexical frequency > structural association > boundary alignment, even though all three properties are cross-linguistically robust. However, frequency alone is not sufficient: we observe a Goldilocks effect, whereby function words must be frequent enough to be reliable, yet sufficiently diverse to remain informative. Finally, our probing results suggest that similar behavioral performance does not necessarily arise from the same internal mechanisms, indicating multiple routes to successful grammatical learning by neural learners.

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). Function words are typically 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). Although these properties have been hypothesized to serve as universal cues for acquisition, their cross-linguistic robustness has not been systematically evaluated.

2.2 Function Words in Language Acquisition

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

Computational research on function words has historically received less attention, generally falling into two strands. The first strand uses computational models to simulate human acquisition phenomena. Several studies focus on syntactic categorization, demonstrating that simple distributional cues (e.g., frequent frames) are sufficient to categorize words into nouns and verbs (Mintz, 2003; Chemla et al., 2009; Gutman et al., 2015). Mintz et al. (2002); Johnson et al. (2014) found that putting function words at phrase boundaries alone yields accurate categorization or word learning. More recently, Ma and Xu (2025) attempted to replicate the experiments by Valian and Coulson (1988) with LLMs, but they found that LLMs showed limited sensitivity to marker frequency in small-scale artificial settings. The second strand probes the linguistic knowledge encoded in pretrained language models. Work by Kim et al. (2019), Ettinger (2020), and Portelance et al. (2024) has investigated whether (visual) language models retain knowledge of function words, yielding mixed results.

3 Methodology

3.1 Function Word Identification

Although function words are often treated as a closed class, the boundary between function and content words is not always clear-cut (Kayne et al., 2005)[2]. 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, 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

[2] *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)

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

cues. Numerals and quantifiers are excluded as they carry non-trivial semantic content.

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).

3.2 Training Data Construction

We construct our training corpora by manipulating text from Wikipedia.[3] We prioritize Wikipedia over simplified datasets like BabyLM (Hu et al., 2024) for several reasons. One of our core manipulations targets the alignment between function words and syntactic phrase boundaries, which requires input sentences with rich, complex phrase structures. Sequences in BabyLM are substantially shorter (14.8 vs. 25.3 tokens on average) and exhibit lower syntactic complexity (dependency locality: 3.0 vs. 3.4; Gibson, 2000). In such simplified input, phrase boundaries are often trivial, leaving little room to meaningfully disrupt boundary information. Applying our manipulations to BabyLM would thus yield language variants that are similar to the original, undermining the experimental contrast.

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.

3.3 Experimental Conditions

Inspired by previous research (Valian and Coulson, 1988; Green, 1979; Morgan et al., 1987; Christophe et al., 2008), we generate different manipulated versions of the corpus across the three properties discussed in Section 2. Content words, sentence length distributions, and total dataset size are held constant across all conditions.

High 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 the 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 funciton words 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 ∼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 deterministically by the *following* word (local bigram dependency) rather than by structural class. We first construct a one-to-one mapping between

[3] https://huggingface.co/datasets/wikimedia/wikipedia

vocabulary items plus a special unknown word and function words. When regenerating the corpus, function words are generated conditioned on the following word.
- 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.

3.4 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 averaged over 3 random seeds. Detailed hyperparameters are provided in Appendix B.

3.5 Evaluation

We do not use perplexity as an evaluation metric because our manipulations alter the distribution (entropy) of the corpora, rendering cross-condition perplexity comparisons invalid. Instead, we evaluate structural generalization using the BLiMP benchmark (Warstadt et al., 2020).

To adapt BLiMP to our manipulated languages, we apply the same text transformations (e.g., NOFUNCTION) to the test sets. We implement a rigorous filtering protocol: (1) We remove categories where the critical distinct 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 Do Function Words Share Universal Distributional Properties?

We aim to confirm that the properties identified as crucial for model learning, i.e., high-frequency, structural anchor, and phrase boundary, are linguistic universals. For this purpose, we use the UD (v2.17) treebanks covering 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 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. If word classes were uniformly distributed, the proportion of function-word types would match their token 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.

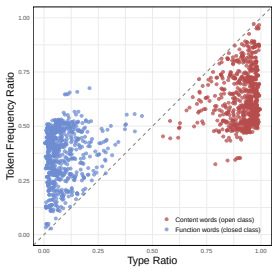
Reliable structural prediction Function words occur in predictable syntactic environments. For instance, in English, a determiner deterministically signals a noun phrase structure. To test this cross-linguistically, we analyze the *syntactic selectivity* of POS tags by modeling dependency trees as undirected graphs. We calculate the entropy of the syntactic neighbors (i.e., co-dependents) for each POS category. For any given POS tag x, the entropy of x is defined as:

$$H(x) = -\sum_{t\in\mathcal{T}} p(t|x)\log_2 p(t|x) \qquad (1)$$

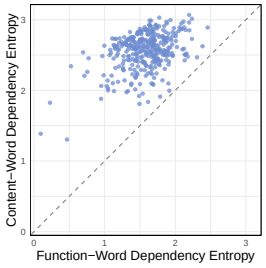
where $p(t|x)$ is the probability that a connected node has the POS tag t, given the current node has

Condition	Overall	S-V Agr	Irregular	NPI	Island	Filler.Gap	Ellipsis	Ctrl.Rais	Binding	Arg.Str	Ana.Agr
NATURALFUNCTION	67.3	65.0	88.4	56.4	55.9	64.2	80.1	70.3	72.3	69.9	93.8
NOFUNCTION	56.5 (-10.9)	54.8 (-10.3)	59.1 (-29.3)	50.1 (-6.3)	50.1 (-5.8)	42.1 (-22.1)	70.1 (-10.1)	61.6 (-8.8)	64.3 (-8.0)	58.0 (-11.9)	92.8 (-1.0)
FIVEFUNCTION	64.2 (-3.2)	57.6 (-7.4)	81.5 (-6.9)	59.6 (+3.3)	49.2 (-6.7)	73.3 (+9.1)	78.1 (-2.0)	62.1 (-8.2)	70.7 (-1.6)	63.0 (-6.9)	82.0 (-11.9)
MOREFUNCTION	58.5 (-8.9)	55.8 (-9.2)	62.9 (-25.5)	56.2 (-0.1)	49.6 (-6.3)	53.6 (-10.7)	66.5 (-13.7)	60.4 (-10.0)	66.2 (-6.1)	60.8 (-9.1)	70.6 (-23.2)
BIGRAMDEP	62.5 (-4.8)	56.9 (-8.2)	83.8 (-4.6)	54.9 (-1.5)	49.6 (-6.3)	66.5 (+2.2)	71.1 (-9.0)	62.8 (-7.6)	67.2 (-5.1)	63.0 (-6.8)	92.2 (-1.6)
RANDOMDEP	60.7 (-6.6)	56.4 (-8.6)	68.6 (-19.7)	59.4 (+3.0)	45.9 (-10.0)	62.5 (-1.7)	76.4 (-3.7)	61.1 (-9.2)	66.1 (-6.2)	61.0 (-8.9)	86.4 (-7.5)
WITHINBOUNDARY	65.7 (-1.6)	69.8 (+4.7)	79.9 (-8.4)	53.4 (-2.9)	50.2 (-5.7)	66.8 (+2.6)	73.0 (-7.1)	66.2 (-4.1)	69.3 (-2.9)	71.8 (+2.0)	85.4 (-8.5)

Table 2: BLiMP accuracy by training condition. Values are mean accuracy across 3 random seeds. Deltas indicate differences relative to the NATURALFUNCTION condition.



(a) Word frequency ratio vs. type ratio



(b) Function vs. Content word dependency entropy

Figure 1: Distributional properties of function words across languages.

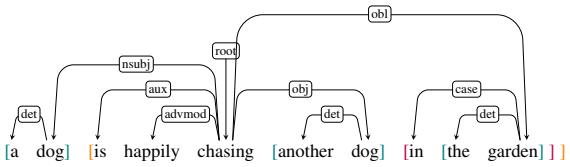


Figure 2: An example dependency tree with constituent spans highlighted using colored brackets: NP, VP, and PP. Phrase labels are theoretically neutral.

tag x. We verify this by computing the weighted average entropy for the set of function tags $\mathcal{F}$:

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

The same normalization applies to content word entropy $\bar{H}_{\mathcal{C}}$. Figure 1b illustrates the results. We observe 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 function words align with phrase boundaries. While a constituency-parsed dataset would be preferable, using one would require drastically reducing the number of languages considered.

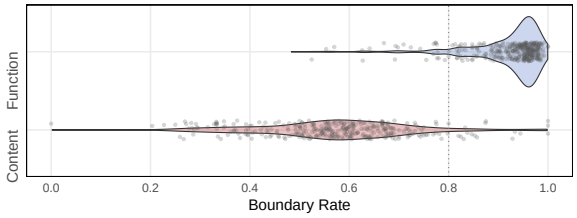


Figure 3: The boundary ratio of function and content words across languages.

We therefore opt to use UD as it provides the most comprehensive collection of multilingual syntactic data currently available, and make the following modifications. we approximate constituent spans using the yield of dependency subtrees. For each target function word, we identify the subtree governed by its syntactic head and check if the function 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 function word is immediately preceded (or followed) by another function word that

marks the constituent boundary, we also consider the target function word as a boundary marker.[4]

Figure 3 presents the results across languages. The data reveals a strong cross-linguistic tendency: the median ratio of function words occurring at phrase boundaries is 0.95 while for content words the number is only 0.58.

5 Which Properties of Function Words Support Structural Learning?

Having established that the three distributional properties of function words are robust across languages, we next ask which of these properties are most important for supporting the learnability of hierarchical structure by neural language models. We define the learnability of hierarchical structure as the ability to acquire abstract grammatical generalizations, measured by performance on the BLiMP benchmark.

Experiment Setup. We train transformer language models on the manipulated languages shown in Table 1 and evaluate them on their corresponding BLiMP benchmarks.

Results. Results are shown in Table 2. Models trained on the NATURALFUNCTION condition, which preserves all three distributional properties, achieve the highest grammatical accuracy while models trained on NOFUNCTION which lack all the three properties give the worst performance. All other conditions exhibit varying degrees of degradation. In particular, reducing function-word frequency (MOREFUNCTION) leads to the largest performance drops followed by destroying reliable structural association (RANDOMDEP). By contrast, the WITHINBOUNDARY condition shows a comparatively smaller penalty. A paired t-test did not reveal a statistically significant difference between WITHINBOUNDARY and NATURALFUNCTION ($p = 0.08$). Although function words are displaced from their canonical boundary positions, the deterministic manipulation preserves a learnable association between structure and function words, partially mitigating the loss.

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.

Category-wise effects are not uniform. Manipulations of function words have relatively little impact on NPI licensing, suggesting that some grammatical phenomena rely less on function-word distributions and may be learned from alternative cues in the input.

Summary. We show that not all distributional properties of function words contribute equally to learnability. Optimal learnability arises when function words are sufficiently frequent to be reliable, yet sufficiently diverse to support distinctions among grammatical structures.

6 How Are Function Words Represented and Deployed?

The previous experiments show that manipulating the distributional properties of function words leads to systematic differences in grammatical performance. However, does comparable performance necessarily imply that models rely on the same information or internal mechanisms? To address this doubt, we examine how models trained under different function-word distributions internally represent and deploy function-word information. If function words play a foundational role in learning, disrupting function-word information at evaluation time should selectively impair models that rely on these cues. We therefore combine attention probing and function-word ablation to assess the extent and manner in which models depend on function words after learning.

6.1 Probing LM's Attention to Function Words

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

[4] 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.

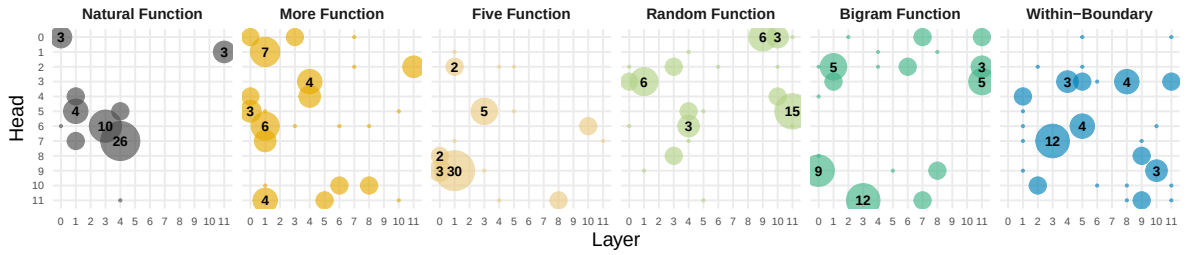
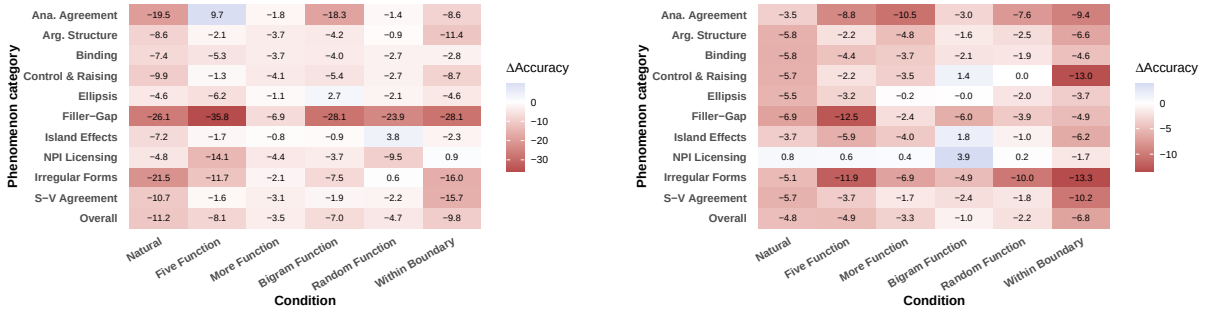


Figure 4: Distribution of dominant function heads across BLiMP categories under different training conditions (seed=53). Numbers indicate the top-5 heads and the number of BLiMP categories in which each head assigns the highest attention to function words. Results for other seeds are reported in Appendix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



(a) Function word deletion

(b) Function word masking

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

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 either frequency (MoreFunction) or structural predictability (RandomDep) fail to produce similarly concentrated representations. Attention to function words in these models is diffuse and inconsistent, suggesting that the absence of these properties prevents the formation of stable function-word representations.

The FiveFunction condition shows an intermediate pattern: despite extreme frequency, reduced lexical diversity weakens specialization, indicating that frequency alone is insufficient without differentiated structural association. Finally, although WithinBoundary achieves comparable performance to NaturalFunction in our previous experiment, its function-word attention is more sparsely distributed across heads, suggesting that similar behavioral outcomes may arise from different internal representations.

6.2 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 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, models trained on NaturalFunction,

FIVEFUNCTION, and WITHINBOUNDARY exhibit the largest performance drops under both ablations, indicating strong reliance on function-word information. In contrast, models trained on MORE-FUNCTION and RANDOMDEP show relatively small performance degradation, suggesting weaker reliance on function words.

Masking and deletion further reveal condition-specific differences. In the masking experiment, which removes function-word identity but the absolute positions of content words are preserved, WITHINBOUNDARY shows a larger performance drop than NATURALFUNCTION, indicating a greater reliance on function-word identity rather than positional cues. By contrast, models trained on BIGRAMDEP remain relatively robust to masking, consistent with the fact that strong local predictability allows learning to be supported by alternative cues (i.e., the following word).

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. In particular, disrupting frequency or reliable structural association leads to weaker reliance on function words overall. Although WITHINBOUNDARY achieves performance comparable to NATURALFUNCTION, attention probing and function-head ablation reveal a different strategy: models show more sparse function heads and rely more on function-word identity than positional information, showing that similar learning outcomes can arise from distinct internal mechanisms.

7 Discussion & Conclusion

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 refine 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, leaving a fuller account to future work. 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.[5] 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 by function words, while segmentation cues may be partially recovered from other sources such as transition statistics.

In addition, our results reveal a dissociation between performance and competence in language models: models that achieve similar performance of structural generalization do not necessarily rely on the same internal mechanisms. This finding highlights the limitations of purely behavioral evaluation and underscores the importance of probing and ablation analyses for understanding how learning proceeds.

Finally, it is important to emphasize that learnability is defined relative to the learner. We treat transformer language models as weakly biased, domain-general statistical learners and use them to re-examine claims from the acquisition literature under controlled conditions. From this view, our results clarify which distributional properties are sufficient to support grammatical learning in a general learner, without assuming human-specific constraints, and provide a foundation for future comparative work across learner types.

[5] 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 establish 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 structural generalization. English uniquely satisfies both requirements with reliable parsing tools (Stanza) and the BLiMP benchmark, enabling the rigorous controlled comparisons central to our research questions. As our cross-linguistic analysis demonstrates that the three distributional properties hold robustly across typologically diverse languages, we hypothesize that the functional hierarchy we identify (frequency > structural association > boundary alignment) reflects general learning principles rather than English-specific patterns. Nevertheless, 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, 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. 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.

Finally, as discussed in Section 7, learnability should be understood as relative to the learner rather than as an intrinsic property of the language alone. While human learners and neural language models differ in many fundamental ways (e.g., Warstadt and Bowman, 2022; Chemero, 2023) and we do not claim that the models studied here mirror human behavior, in line with Futrell and Mahowald (2025), we nevertheless treat language models as weakly biased statistical learners whose behavior can be used to clarify which distributional properties are sufficient to support grammatical learning, and to generate testable hypotheses for future work on human language acquisition, human mind, and language evolution.

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.

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.

Anthony Chemero. 2023. LLMs differ from human cognition because they are not embodied. *Nature Human Behaviour*, 7(11):1828–1829.

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*.

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.

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.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *arXiv preprint arXiv:2501.17047*.

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.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 94–126. The MIT Press, Cambridge, MA.

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.

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.

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.

Shalom Lappin and Stuart M Shieber. 2007. Machine learning theory and practice as a source of insightinto universal grammar. *Journal of Linguistics*, 43(2):393–427.

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.

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.

Lisa S Pearl and Jon Sprouse. 2015. Computational modeling for language acquisition: A tutorial with syntactic islands. *Journal of Speech, Language, and Hearing Research*, 58(3):740–753.

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.

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 and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

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.

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.

A BLiMP Categories Considered

We list the categories that are removed from our experiments because of 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

C Function Word Inventory

Function words are collected from the gold annotations of GUM and EWT with manual corrections.

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.

Model	
model_type	GPT2-small
vocab_size	32768
seeds	42,53,67
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

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.

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:

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

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

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

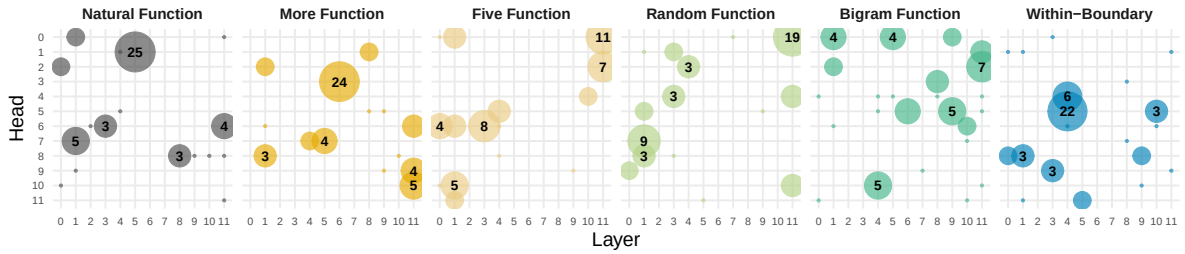
Finally, we identify the specific head (h^*, l^*) that maximizes this function attention score for each subcategory in BLiMP.

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

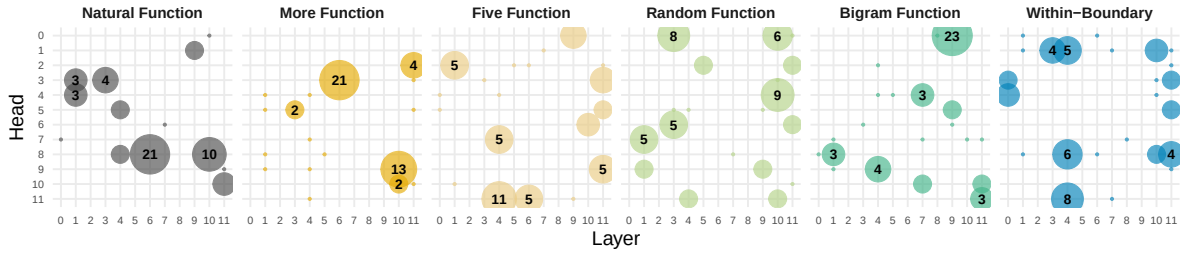
Table 6: Function word inventories used in the experiments, grouped by syntactic category.

F Function Head Distribution Results for the Other Two Random Seeds

See Figure **??**.



(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.

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

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).