# Do Structural Priors Help Neural Language Models Learn Grammar? Evidence from Child-Scale Data

**Jon-Paul Cacioli**
Independent Researcher
Melbourne, Australia
synthium@hotmail.com

## Abstract

We show that structural grammatical priors produce targeted, linguistically specific effects on grammatical learning: improving filler-gap dependencies — which require long-distance hierarchical tracking — by 9–13 percentage points beyond structural regularisation alone ($d = 2.41$–$2.82$), while damaging locally cued phenomena regardless of whether the grammar is real or random. This phenomenon-specificity, revealed by a random grammar control, suggests the right question is not whether structural priors help, but for which constructions and why. We test this by augmenting BabyBERTa (7.4M parameters) with a differentiable PCFG auxiliary loss derived from Minimalist Grammar, trained on AO-CHILDES (893K sentences of child-directed speech). In a pre-registered study of 190 experimental runs spanning 7 constraint strengths, 3 data scales, 5 random seeds, and 3 independent lexicon permutations, our confirmatory hypotheses about overall accuracy and sample efficiency are falsified. However, a random grammar control ($n = 15$ runs per condition; three independent lexicon permutations) reveals that linguistically accurate category assignments specifically drive filler-gap gains: real grammar outperforms both a structurally equivalent random grammar and the no-grammar baseline, while both conditions equally damage subject-verb agreement. These results show that structural priors function as targeted interventions rather than global boosters: they help specifically the constructions, specifically long-distance dependencies, whose computational demands align with what phrase-structure representations encode. We release code and pre-registered materials.[1]

[1]Pre-registration: https://osf.io/5rz9w/. Code: https://github.com/synthiumjp/neurosym-grammar. Hypotheses were registered after smoke-testing confirmed code correctness, but before the main experimental grid was run. All results reported regardless of outcome.

## 1 Introduction

The BabyLM Challenge (Warstadt et al., 2023, 2024) has demonstrated that small language models trained on developmentally plausible data can acquire surprising amounts of grammatical knowledge. BabyBERTa (Huebner et al., 2021) achieves grammatical competence comparable to RoBERTa-base using 6,000× fewer words and 15× fewer parameters, establishing a strong baseline for what distributional learning alone can accomplish at child scale. These results raise a question central to both computational linguistics and acquisition research: can explicit structural priors improve learning beyond what distributional statistics provide?

Cognitive scientists have long debated whether grammatical knowledge can be acquired from linguistic input alone or requires innate structural biases (Chomsky, 1965; Pullum and Scholz, 2002; Clark and Lappin, 2010; Pearl, 2022). Computational models offer a way to test these claims empirically: if adding structural priors to a neural learner measurably improves grammatical generalisation, this constitutes evidence that such priors are *useful*, regardless of whether they are innate (Warstadt and Bowman, 2022). Recent work has shown that neural language models can learn aspects of filler-gap dependencies and island constraints from data alone (Wilcox et al., 2024), but these models train on orders of magnitude more data than children receive. Whether structural priors help at genuinely developmental scales remains untested.

Despite advances in neurosymbolic grammar induction (Kim et al., 2019; Yang et al., 2021; Park and Kim, 2025), developmentally plausible language modelling (Huebner et al., 2021; Warstadt et al., 2023), and recent work integrating Minimalist Grammar into BabyLM training (Chesi et al., 2024), no prior work has combined explicit grammar constraints at child-scale data volumes with a random grammar control that decomposes struc-



*Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL)*, pages 15–26

tural regularisation from linguistic content. We address this gap with four contributions:

1. A **neurosymbolic architecture** combining BabyBERTa with a differentiable PCFG auxiliary loss, where Minimalist Grammar-inspired rules shape learning through gradient-based structural supervision.

2. A **comprehensive evaluation** across 13 grammatical phenomena, 7 constraint strengths, 3 data scales (25%, 50%, 100% of AO-CHILDES), and 5 random seeds, totalling 130 experimental runs (plus 60 random grammar control runs).

3. A **random grammar control** with three independent lexicon permutations (60 additional runs) that decomposes observed effects into structural regularisation (from the CKY computation itself) versus linguistically specific content (from accurate category assignments).

4. **Falsified confirmatory hypotheses** reported transparently alongside informative exploratory findings from the random grammar analysis.

Our key finding is that the question "do structural priors help?" does not have a uniform answer across the grammar. Correct structural supervision specifically improves filler-gap dependencies beyond what structural regularisation alone provides, with large effect sizes ($d = 2.41$–$2.82$), while failing to improve — and in some cases damaging — locally cued phenomena. This phenomenon-specific pattern suggests that structural priors are most valuable for constructions requiring long-distance hierarchical tracking, the class of dependencies that motivates much of the debate around innate linguistic knowledge (Chomsky, 1965; Wilcox et al., 2024).

## 2 Method

### 2.1 Architecture

We extend BabyBERTa (Huebner et al., 2021), a small RoBERTa variant (7.4M parameters, 8 layers, 8 attention heads, 256 hidden dimensions) optimised for child-scale data. Training uses a combined loss:

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \lambda \cdot \mathcal{L}_{\text{grammar}} \quad (1)$$

where $\mathcal{L}_{\text{MLM}}$ is the standard masked language modelling objective and $\mathcal{L}_{\text{grammar}}$ is the negative log-probability of the input under a probabilistic context-free grammar (PCFG), computed via a differentiable inside algorithm. The hyperparameter $\lambda$ controls constraint strength.

The grammar loss operates through a **SoftLexicon** routing mechanism: a fixed, non-learnable matrix $M \in \mathbb{R}^{|V| \times |C|}$ maps from the model's token probability distribution to grammar category probabilities, where $M[\text{token}, \text{category}] = 1.0$ if the lexicon maps that token to that category. Concretely, given softmax output $p \in \mathbb{R}^{|V|}$, category probabilities are computed as $c = M^\top p$, yielding a distribution over grammatical categories that is then fed into the CKY inside algorithm. This matrix is determined entirely by the hand-crafted lexicon and contains no trainable parameters, ensuring that the grammar acts as a prior rather than an additional learned component. This approach contrasts with Chesi et al. (2024), who integrate Minimalist Grammar constraints into BabyLM training by modifying the gating mechanisms of RNNs rather than adding a differentiable auxiliary loss to a transformer. During training, the gradient chain flows from model logits through softmax token probabilities, through the category matrix, into the Cocke–Kasami–Younger (CKY) inside algorithm, and back, providing structural supervision without requiring discrete parsing decisions. We repurpose the CKY inside algorithm — traditionally used for parsing — as a differentiable scoring function: rather than finding the best parse, it marginalises over all possible parses to compute the total probability of the sentence under the grammar, and this scalar probability serves as the loss term.

Because the softmax produces non-zero probabilities for all tokens, gradient always flows; the lexicon determines the *direction* of this gradient by specifying which category each token is routed toward. For computational efficiency, grammar loss is computed every 4 training steps on 25% of each batch, with sentences truncated to 12 tokens (CKY complexity is $\mathcal{O}(n^3)$). This truncation is developmentally plausible: children's early utterances are predominantly short (Brown, 1973; Huebner et al., 2021), and the AO-CHILDES corpus has a median sentence length of 7 tokens.

### 2.2 Grammar

The grammar is a PCFG inspired by Minimalist Grammar (Stabler, 1997), implemented in Chom-

sky Normal Form. It comprises 17 nonterminal categories, 12 terminal categories, and 16 production rules encoding English phrase structure (e.g., TP $\rightarrow$ NP VP; VP $\rightarrow$ V NP; CP $\rightarrow$ C TP). The lexicon maps 7,230 word types to grammatical categories with associated features. The grammar directly encodes six of the thirteen evaluated phenomena: subject-verb agreement, determiner-noun agreement, argument structure, filler-gap dependencies, island effects, and local attractor configurations. The remaining seven phenomena (binding, NPI licensing, quantifiers, anaphor agreement, ellipsis, irregular forms, case) are not encoded in the grammar, providing a natural test of constraint specificity.

While inspired by Minimalist Grammar, our PCFG is a phrase-structure approximation rather than a full MG implementation: it encodes hierarchical constituent structure and basic filler-gap configurations through the CP $\rightarrow$ C TP rule, but cannot represent feature-checking, Agree relations, or the negative island constraints that require specifying where movement is prohibited. This scope is important for interpreting our results; benefits are expected only for phenomena whose structural properties fall within what the grammar can express.

### 2.3 Training Data

We use AO-CHILDES (Huebner et al., 2021), aggregating child-directed speech from the CHILDES database (MacWhinney, 2000) for children aged 1–6. The corpus contains 893,989 sentences ($\sim$5M words). To test sample efficiency, we create pregenerated random subsamples at 25% (223K sentences) and 50% (447K sentences). The grammar auxiliary loss is computed on sentences truncated to 12 sub-word tokens for computational tractability (the CKY inside algorithm is $\mathcal{O}(n^3)$). Given AO-CHILDES's mean sentence length of 7.3 subtokens (Huebner et al., 2021), this limit affects only $\sim$15–17% of training sentences, preserving the short conversational utterances characteristic of child-directed speech. The grammar constraint nonetheless generalises to full-length Zorro test items at evaluation time, indicating that the structural prior shapes learned representations rather than requiring explicit parsing of all training sentences.

### 2.4 Random Grammar Control

To distinguish linguistic content from structural regularisation, we construct a random grammar control by permuting the lexicon's word-to-category mappings. This preserves the exact category frequency distribution, the full CKY computation with identical algorithmic structure, and the same number of trainable gradient steps. It destroys only the linguistic accuracy of category assignments. Any performance difference between real and random grammar therefore isolates the contribution of linguistically accurate structural supervision, controlling for the regularisation effect of the CKY computation itself.

We use three independent permutations (seeds 99, 2026, and 31415), yielding $n = 15$ random-grammar runs per condition and tighter confidence intervals than a single permutation would provide. The three permutations produce highly consistent results (inter-permutation SD $\leq$ 2.7 pp for filler-gap; $\leq$ 0.3 pp for SV agreement), confirming that findings are not artefacts of any particular lexicon scrambling.

### 2.5 Evaluation

We evaluate using the Zorro benchmark (Huebner et al., 2021), a grammaticality judgement test suite designed for child-directed vocabulary. Zorro tests 13 phenomena across 23 paradigms using minimal pairs. We report accuracy (percentage of pairs where the model assigns higher probability to the grammatical sentence; ties count as incorrect) for each phenomenon and overall, averaging across paradigms within each phenomenon.

### 2.6 Experimental Design

We conduct 130 runs in total, with all hypotheses pre-registered on OSF after smoke testing confirmed implementation correctness and before the main experimental grid was executed. **Baselines** (5 runs): BabyBERTa trained on 100% data with 5 random seeds (42, 123, 456, 789, 1001). **Real grammar** (105 runs): 7 $\lambda$ values (0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0) $\times$ 3 data fractions (25%, 50%, 100%) $\times$ 5 seeds. **Random grammar** (60 runs): 3 independent lexicon permutations (seeds 99, 2026, 31415) $\times$ 2 $\lambda$ values (0.2, 0.5) $\times$ 2 data fractions (25%, 50%) $\times$ 5 seeds. Training steps are scaled by data fraction (25K / 50K / 100K). The $\lambda$ range spans from negligible (0.001, where the grammar gradient is present but too weak to reduce grammar

loss) to strong (1.0, where the weighted grammar term contributes more to the total loss than the MLM term).

All other hyperparameters were kept at BabyBERTa defaults. Confirmatory analyses use Bonferroni correction ($k = 7$ $\lambda$ values for H1, yielding per-test $\alpha = 0.05/7 = 0.007$; $k = 3$ pairwise comparisons for exploratory three-condition tests, yielding $\alpha = 0.05/3 = 0.017$); exploratory analyses (including the random grammar control) are labelled as such throughout.

## 3 Results

### 3.1 Overall Accuracy and Confirmatory Hypotheses

Our pre-registered primary hypothesis (H1) predicted that grammar constraints would improve overall Zorro accuracy. This hypothesis was **falsified**: the best constrained model at 100% data ($\lambda = 0.01$, 69.1%) did not significantly outperform the baseline (68.8%; $t(8) = 0.59$, $p = 0.574$, $d = 0.37$; Bonferroni-corrected $\alpha = 0.007$). Cohen's $d$ throughout uses the pooled standard deviation across seeds. Overall accuracy decreased monotonically above $\lambda = 0.01$ (Figure 1). Our sample efficiency hypothesis (H2) was also falsified: the best constrained model at 50% data (65.1%) fell significantly short of the baseline at 100% data (68.8%; $t(8) = -6.22$, $p = 0.003$, $d = -3.84$). Our specificity hypothesis (H3) predicting greater improvement for grammar-encoded phenomena was not supported at any data fraction (100%: $p = 0.074$; 25%: $p = 0.08$; 50%: $p = 0.35$).

BabyBERTa is already a strong baseline for locally cued phenomena, with subject-verb agreement accuracy near 62% and case near 94%, leaving limited headroom for improvement on these tasks. However, these null overall results mask phenomenon-specific patterns that emerge from the random grammar control analysis.

### 3.2 Filler-Gap Dependencies: Linguistic Specificity

The random grammar control reveals that filler-gap accuracy gains are *linguistically specific*: real grammar outperforms random grammar by a substantial and statistically significant margin (Figure 2).

At 50% data with $\lambda = 0.5$: real grammar achieves 91.0% filler-gap accuracy versus 82.1% for random grammar (mean across three permutations) versus 79.2% for baseline. The real–random gap of +8.8 pp is highly significant ($t(18) = 5.47$, $p < 0.001$, $d = 2.82$; Mann–Whitney $U = 75$, $p = 0.001$).[2] At 25% data with $\lambda = 0.5$: the pattern is even more striking. Real grammar scores 82.2% versus random grammar at 69.3% (mean across three permutations) versus baseline at 79.2%, yielding a real–random gap of +12.9 pp ($t(18) = 4.67$, $p < 0.001$, $d = 2.41$; Mann–Whitney $p = 0.002$, 70/75 pairs). Notably, random grammar at this setting actually *decreases* filler-gap accuracy below baseline (−9.9 pp), while real grammar improves it (+3.0 pp).

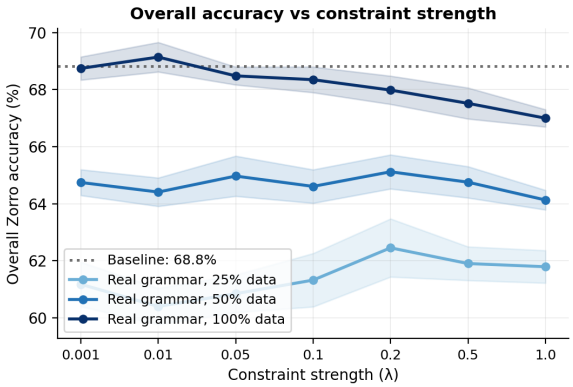


Figure 1: Overall Zorro accuracy vs. constraint strength $\lambda$ for three data fractions. The dashed line shows baseline accuracy at 100% data (68.8%). Grammar constraints do not improve overall accuracy at any $\lambda$ or data fraction, falsifying H1 and H2.

This three-condition comparison enables a principled decomposition of the grammar constraint's effect. The CKY computation itself provides some structural regularisation: random grammar at 50% data still outperforms baseline by ∼3 pp on filler-gap. However, accurate category assignments contribute substantially beyond this: approximately three-quarters of the total effect at 50% data (+8.8 pp linguistic out of +11.7 pp total, averaged across three permutations), and more than the entire effect at 25% data where random grammar actively hurts (−9.9 pp on average). As data becomes scarcer, the regularisation component diminishes while the linguistic specificity component becomes dominant.

Figure 2 plots filler-gap accuracy against $\lambda$

[2] All 75 real–random seed pairs show the real grammar exceeding random (75/75); bootstrap 95% CI for the gap: [5.4, 12.2] pp. Three independent lexicon permutations (seeds 99, 2026, 31415) yield consistent random means of 83.9%, 81.7%, and 80.8%, confirming the result is not permutation-specific. Bonferroni-corrected $\alpha = 0.017$ for three pairwise comparisons.

for both conditions. At low constraint strengths ($\lambda \leq 0.1$), real grammar performance is flat or slightly below baseline. Above this threshold, real grammar climbs steeply while random grammar diverges downward, visually confirming that the model specifically requires *correct* structural information to solve filler-gap dependencies, not merely any tree-shaped auxiliary loss.

### 3.3 Agreement, Binding, and NPI Patterns

Figure 3 shows the full pattern across focal phenomena for both grammar conditions.

**Subject-verb agreement.** Both real and random grammar reduce agreement accuracy by 8–11 pp below baseline in all conditions, and are statistically indistinguishable (all pairwise $p > 0.05$, $d \approx 0$). This symmetric damage rules out lexical interference as the cause; if misassigned categories were responsible, random grammar should damage agreement *more*. Instead, the CKY computation itself appears to divert representational capacity from the local lexical-distributional features (singular/plural morpheme correlations) that drive agreement, regardless of whether the lexicon is linguistically accurate.

**NPI licensing.** NPI licensing shows high instability across lexicon permutations. A single permutation (seed 99) suggested random grammar produced larger NPI gains than real grammar, but this does not replicate: across three permutations, random grammar averages 35.4% (barely above the 35.0% baseline) while real grammar averages 39.3%, with neither condition significantly differing ($t(18) = 0.66$, $p = 0.52$). The high per-seed variance (SD = 16.9; bimodal distribution)[3] indicates that NPI licensing is not robustly learnable from grammar constraints at this data scale, and underscores the value of multiple random controls.

**Binding.** Both grammars reduce binding accuracy below baseline at $\lambda = 0.2$ and 0.5. However, at very low constraint strengths ($\lambda = 0.001$) in our 100% data experiments, real grammar produces a substantial binding improvement (+6.2 pp) despite the grammar loss not decreasing during training. We return to this finding in §4.1.

[3]Per-seed NPI accuracy for real grammar at 50%, $\lambda = 0.5$: 20.7, 25.5, 37.5, 53.9, 58.9%.

| Condition | Fill. | SVAgr | Bind. | NPI |
|---|---|---|---|---|
| Baseline (100%) | 79.2±1.5 | **61.5**±1.4 | **66.8**±4.1 | 35.0±4.2 |
| Real (50%, λ=0.5) | **91.0**±2.0 | 52.6±1.0 | 49.7±7.7 | 39.3±16.9† |
| Rand (50%, λ=0.5) | 82.1±3.4 | 52.7±1.1 | 50.2±3.5 | **35.4**±9.3 |
| Real − Rand | +8.8* | −0.1 | −0.5 | +3.9 |

Table 1: Mean accuracy (%) ± SD on focal phenomena. Real grammar: 5 seeds. Random grammar: 15 runs across 3 independent lexicon permutations (seeds 99, 2026, 31415). *Significant: $t(18) = 5.47$, $p < 0.001$, $d = 2.82$; all 75 real–random seed pairs show real > random (87.9–93.5 vs. 75.1–86.6). †High variance reflects a bimodal distribution across seeds (20.7, 25.5, 37.5, 53.9, 58.9%); median = 37.5%. **Bold** = best per column. Fill. = Filler-gap; Bind. = Binding.

### 3.4 Training Dynamics

The grammar auxiliary loss does not impair language model training. Figure 4 shows MLM loss convergence for real and random grammar conditions at 50% data, $\lambda = 0.5$. The two conditions produce virtually identical MLM loss trajectories. Grammar loss itself, however, diverges between conditions: real grammar loss decreases during training, while random grammar loss plateaus after an initial drop, because the model cannot simultaneously satisfy the MLM objective and a grammar that assigns tokens to incorrect categories. This divergence provides direct evidence that the model extracts usable structural signal from the real grammar but not from the random grammar, even though both impose identical computational overhead.

### 3.5 Summary

Table 1 summarises the key three-condition comparisons for focal phenomena. The pattern is clear: accurate structural priors specifically benefit filler-gap dependencies (real > random > baseline) while both grammars equally damage agreement and produce comparable or reversed effects on other phenomena. The full 13-phenomenon breakdown is shown in Appendix A.

## 4 Discussion

### 4.1 Inductive Bias as Gradient Direction

An intriguing pattern emerges at very low constraint strengths. At $\lambda = 0.001$, the grammar loss remains high throughout training, meaning no explicit grammar learning occurs, yet binding accuracy improves by 6.2 pp over baseline. This dissociation between loss magnitude and behavioural effect is consistent with the PCFG objective acting

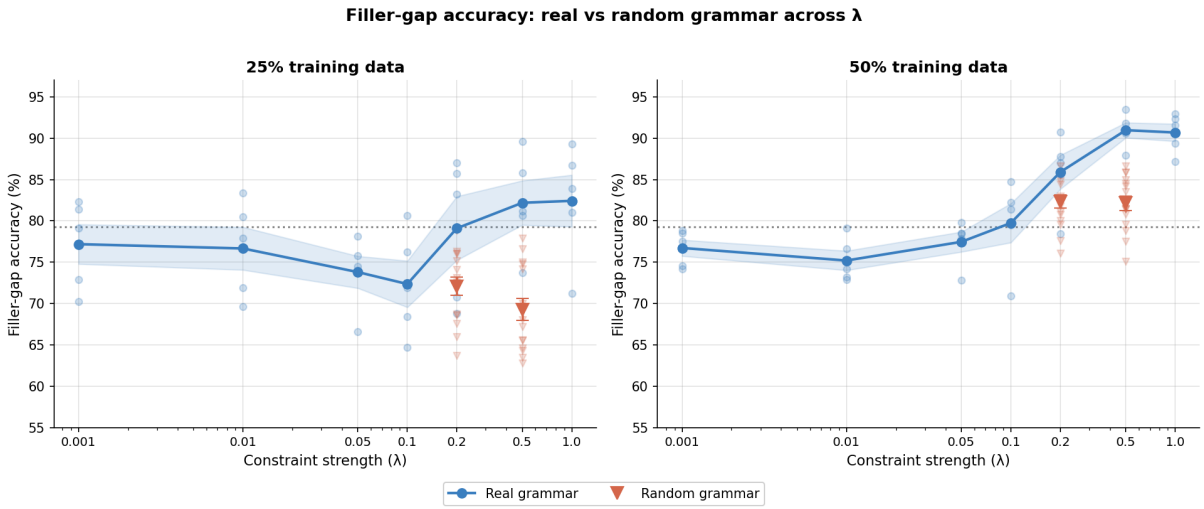


Figure 2: Filler-gap accuracy across the full $\lambda$ sweep for real grammar (circles, solid) and random grammar (triangles, red) at 25% (left) and 50% (right) training data. At low $\lambda$, both conditions perform comparably. Above $\lambda \approx 0.1$, real grammar climbs steeply while random grammar diverges downward, particularly at 25% data. Baseline (100% data) shown as dashed line.

as a *directional bias* on the gradient rather than as an optimisation target: the grammar gradient may nudge the model's representations toward hierarchical structure even when too weak to measurably reduce grammar loss.

However, we note that this observation rests on a single $\lambda$ value and a single phenomenon; we did not conduct the attention-pattern or probing analyses that would be needed to substantiate a mechanistic account. We flag it as a finding warranting dedicated investigation, particularly analysis of what changes in the model's representations at very low $\lambda$, rather than offering it as a confirmed explanation. Its consistency with accounts emphasising the role of inductive bias over explicit knowledge in acquisition (Griffiths et al., 2010; Tenenbaum et al., 2011; Yang, 2004) makes it theoretically interesting, but the empirical support here is preliminary.

### 4.2 Why Filler-Gap but Not Agreement?

The phenomenon-specific pattern has a natural linguistic interpretation rooted in the computational demands of each construction.

Filler-gap dependencies are inherently long-distance and hierarchical: they require tracking a displaced constituent across potentially unbounded clause boundaries (Wilcox et al., 2018, 2024; Howitt et al., 2024). This is precisely the kind of dependency where phrase-structure representations can encode genuine information about which constituents can be displaced and where gaps are licensed. Subject-verb agreement, by contrast, is heavily local and lexically cued; models can learn it through sequential distributional statistics without explicit hierarchical representations (Linzen et al., 2016; Gulordava et al., 2018). The CKY auxiliary loss at moderate-to-high $\lambda$ values imposes a representational trade-off. By forcing the model to represent tokens as members of abstract phrase-structure categories, it diverts capacity from the fine-grained lexical features that drive agreement. This explains why the damage is symmetric across real and random grammar.

A further dissociation — between filler-gap and island constraints — sharpens this picture. Although both are encoded in the grammar and both involve long-distance structure, filler-gap accuracy rises from 79% to 91% at 50% data with $\lambda = 0.5$, while island accuracy *decreases* from 74% to 69% under the same condition. This suggests that structural priors help the model resolve dependencies (positive licensing) but not recognise where dependencies are prohibited (negative constraints). The syntactic interpretation is straightforward: a PCFG can encode the trace position of a moved element, licensing the filler-gap dependency itself, but island constraints require specifying where movement *cannot* occur, negative conditions not naturally expressed in phrase-structure grammar without feature-passing or constraint-based machinery.

The benefit of structural priors may therefore be limited to positive structural facts encodable in the grammar fragment, while negative constraints require richer representations. Notably, the grammar

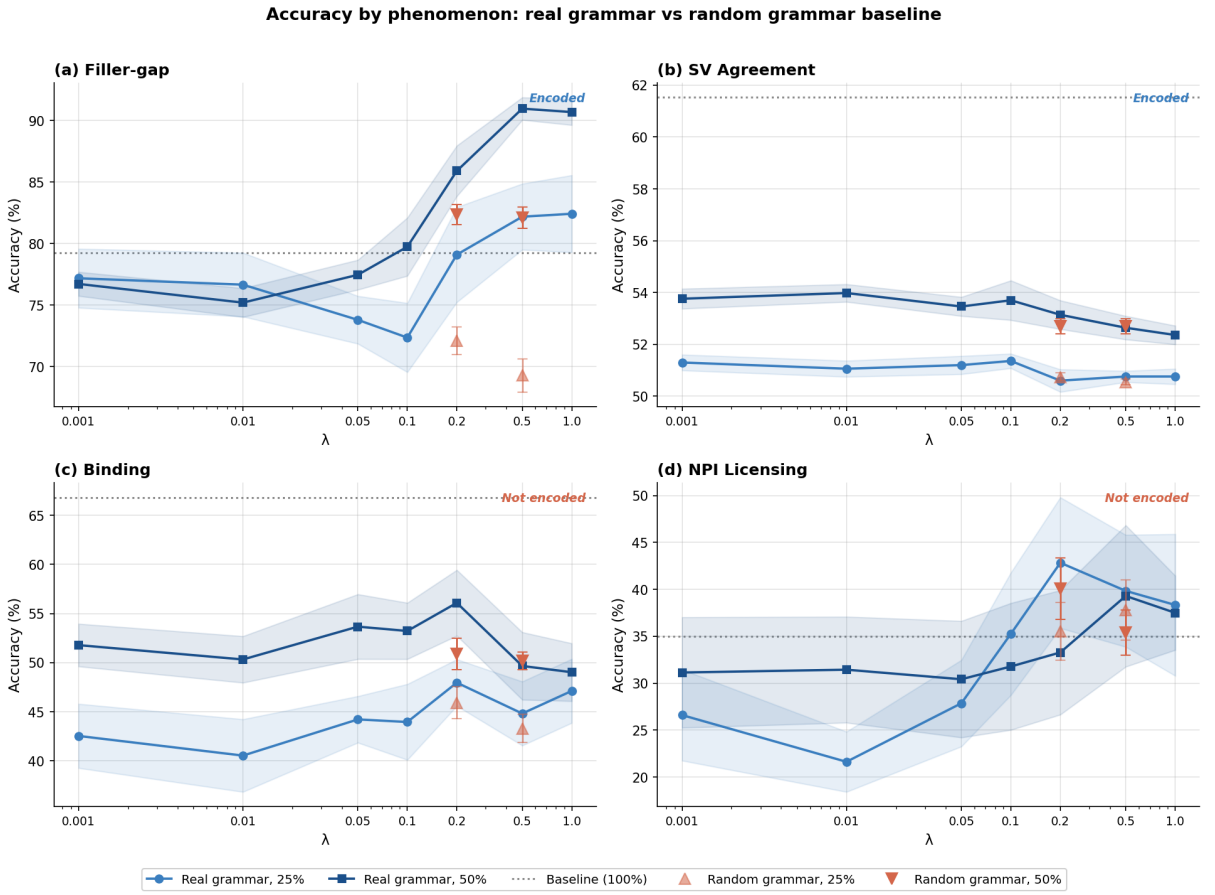


Figure 3: Accuracy by phenomenon: real grammar versus random grammar baseline across $\lambda$ at 25% (circles) and 50% (squares) training data. Dashed baseline (100% data) shown for reference. (a) Filler-gap shows strong real-grammar advantage at high $\lambda$; (b) SV agreement shows symmetric damage from both grammars; (c) Binding improves with real grammar at very low $\lambda$ despite no grammar loss decrease; (d) NPI licensing shows apparent random-grammar advantage, suggesting regularisation artefact.

loss is computed only on sentences truncated to 12 tokens, yet filler-gap improvements generalise to full-length Zorro items, suggesting that short-sentence structural supervision reshapes representations in ways that benefit long-distance processing.

### 4.3 Regularisation versus Linguistic Content

The three-permutation random grammar control enables a principled decomposition of observed effects. For filler-gap, the decomposition shifts across data scales: at 50% data, approximately three-quarters of the improvement comes from linguistically specific content (+8.8 pp) and one-quarter from structural regularisation alone (+2.9 pp above baseline); at 25% data, regularisation is not merely reduced but negative — random grammar hurts filler-gap by −9.9 pp — while linguistic content drives the entire benefit.

From a developmental perspective, this interaction has a natural interpretation as a learning-trajectory effect: under extreme input scarcity (25% data, ∼1.25M words), only linguistically accurate priors help — generic structural regularisation actively hurts. As input accumulates (50% data), regularisation begins to contribute positively, but accurate linguistic content still accounts for three-quarters of the benefit. This pattern suggests that the role of structural priors may shift during development; early acquisition, when data is most limited, would depend most critically on the accuracy of whatever structural biases are available.

These results also carry a practical warning. Single-permutation structural regularisation can produce unstable results on high-variance phenomena (as illustrated by the NPI finding in §3.3). Without multiple linguistically grounded control permutations, it would be easy to misattribute permutation-specific noise to principled linguistic content.

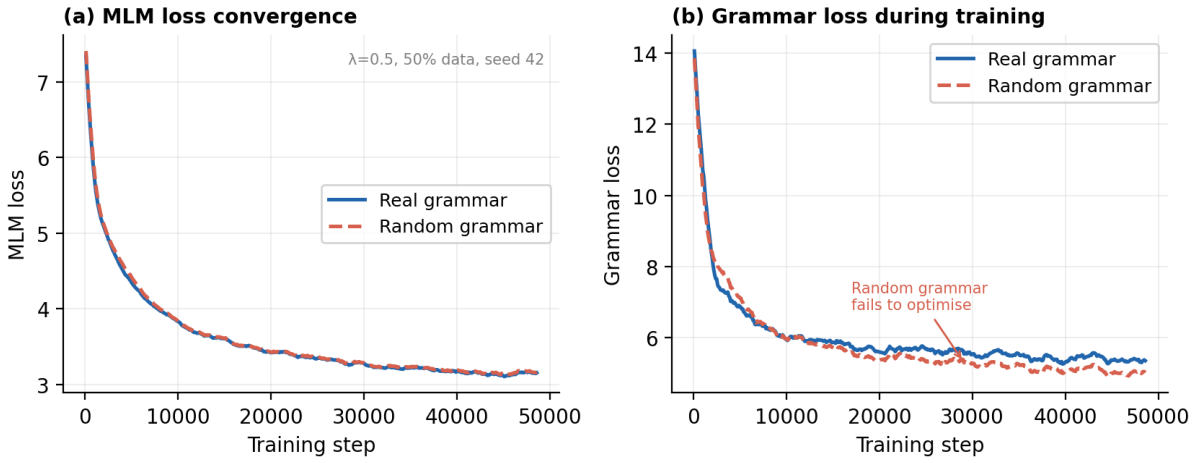


Figure 4: Training dynamics at 50% data, $\lambda = 0.5$. (a) MLM loss converges identically for real and random grammar. (b) Grammar loss decreases for real grammar but plateaus for random grammar, confirming that the model extracts usable structural signal only from the linguistically accurate lexicon.

### 4.4 Relation to Prior Work

Where the BabyLM tradition (Warstadt et al., 2023, 2024) asks *what* grammatical knowledge distributional learning can achieve, we ask *how* the learning objective changes the learning trajectory, treating grammar priors as an active intervention on the training signal. Within neurosymbolic grammar research (Kim et al., 2019; Yang et al., 2021; Park and Kim, 2025), we demonstrate that differentiable grammar integration provides targeted benefits at very small scales. Chesi et al. (2024) also integrate Minimalist Grammar constraints into BabyLM training, but by modifying RNN gating mechanisms to encode c-command and locality rather than adding a differentiable auxiliary loss to a transformer. Their eMG-RNN achieved comparable overall BLiMP scores to standard LSTMs but did not use a random grammar control, making it difficult to isolate the contribution of linguistic content from architectural regularisation.

Alternative routes for injecting structural bias include distilling from Bayesian models trained on formal languages (McCoy and Griffiths, 2025) and pre-pretraining on formal languages to impart linguistic biases (Hu et al., 2025). Our differentiable PCFG approach differs in providing continuous structural supervision during training rather than transferring structure from a separate pre-training phase. Whether these alternative routes produce the same phenomenon-specific pattern we observe is an open question.

Within the nativism/empiricism debate (Chomsky, 1965; Pullum and Scholz, 2002; Clark and Lappin, 2010; Chater and Manning, 2006), we provide empirical evidence that structural priors are *useful* for some aspects of grammar but not others. Filler-gap dependencies, which feature prominently in debates about the necessity of innate structural knowledge, are the constructions that benefit most from accurate structural supervision.

## 5 Conclusion

We tested whether explicit structural grammatical priors improve neural language model learning from child-scale data. Our pre-registered hypotheses about overall improvement and sample efficiency were falsified. However, a random grammar control with three independent lexicon permutations revealed that real grammar priors specifically improve filler-gap dependency learning beyond structural regularisation, with large effect sizes ($d = 2.41$–$2.82$). Wrong grammar damages filler-gap performance while failing to affect agreement. These phenomenon-specific effects suggest that the value of structural priors depends on the computational demands of the grammatical dependency: long-distance hierarchical constructions benefit from accurate linguistic knowledge, while locally cued phenomena do not.

Future work should test whether richer grammars encoding movement and feature-checking produce broader benefits, and whether these effects replicate cross-linguistically. It would also be valuable to investigate whether probing classifiers or attention-pattern analyses can identify the representational changes underlying the gradient-direction

effect at low $\lambda$, and whether the phenomenon-specific pattern persists in larger architectures, connecting the developmental "growing up" approach to the scalability concerns of modern language modelling.

More broadly, we propose the random grammar control as a methodological standard for neurosymbolic language model research. Without a linguistically scrambled baseline, apparent improvements from structural supervision cannot be distinguished from generic regularisation. Our results show that a single lexicon permutation may be insufficient; we recommend at least three independent permutations.

For developmental theory, our results suggest that the contribution of structural biases to language acquisition is neither uniform nor absent, but phenomenon-specific — with the strongest effects emerging precisely where input alone is most impoverished relative to the computational demands of the target construction.

## Limitations

Several limitations qualify these results. First, our grammar is a simplified PCFG fragment, not a full Minimalist Grammar. Second, we evaluate only English child-directed speech. Third, all three confirmatory hypotheses (H1–H3) were falsified; the strongest findings come from the exploratory random grammar analysis. Fourth, the bimodal per-seed distribution for NPI licensing ($\mathrm{SD} = 16.9$) indicates that at child-scale data volumes, random initialisation can dominate the effect of grammar constraints for fragile phenomena. Future work should report full seed distributions rather than means alone. Fifth, our pre-registration also specified data-derived grammar and frequency-based control conditions, which were not implemented due to scope constraints; the exploratory developmental-trajectory (H4) and error-analysis (H5) hypotheses were likewise deferred to future work.

## References


Roger Brown. 1973. *A First Language: The Early Stages*. Harvard University Press.

Nick Chater and Christopher D. Manning. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7):335–344.

Cristiano Chesi, Veronica Bressan, Matilde Barbini, Achille Fusco, Maria Letizia Piccini Bianchessi, Sofia Neri, Sarah Rossi, and Tommaso Sgrizzi. 2024. Different ways to forget: Linguistic gates in recurrent neural networks. In *Proceedings of the 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 106–117.

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

Alexander Clark and Shalom Lappin. 2010. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell.

Thomas L. Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. 2010. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364.

Kristina Gulordava, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL-HLT*, pages 1195–1205.

Kyle Howitt, Suraj Nair, Andrew Dods, and Robert M. Hopkins. 2024. Generalizations across filler-gap dependencies in neural language models. *Preprint*, arXiv:2410.18225.

Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. 2025. Between circuits and Chomsky: Pre-pretraining on formal languages imparts linguistic biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9691–9709.

Philip A. Huebner, Elior Sulem, Cynthia Fisher, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of CoNLL*, pages 624–646.

Yoon Kim, Chris Dyer, and Alexander Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of ACL*, pages 2369–2385.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum.

R. Thomas McCoy and Thomas L. Griffiths. 2025. Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *Nature Communications*, 16(1):4676.

Jiyeon Park and Kangil Kim. 2025. Probability distribution collapse in unsupervised neural grammar induction. In *Proceedings of EMNLP*, pages 33392–33403.

Lisa Pearl. 2022. Poverty of the stimulus without tears. *Language Learning and Development*, 18(4):415–454.

Geoffrey K. Pullum and Barbara C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19(1–2):9–50.

Edward Stabler. 1997. Derivational minimalism. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics*. Springer.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*. CRC Press.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjape, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan G. Wilcox, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2024. Insights from the first BabyLM Challenge: Training sample-efficient language models on a developmentally plausible corpus. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Ethan G. Wilcox, Richard Futrell, and Roger Levy. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848.

Ethan G. Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? In *Proceedings of BlackboxNLP*, pages 211–221.

Charles Yang. 2004. Universal Grammar, statistics, or both? *Trends in Cognitive Sciences*, 8(10):451–456.

Songlin Yang, Yanpeng Zhao, and Kewei Tu. 2021. PCFGs can do better: Inducing probabilistic context-free grammars with many symbols. In *Proceedings of NAACL-HLT*, pages 1487–1498.

# A Full Phenomenon Heatmap

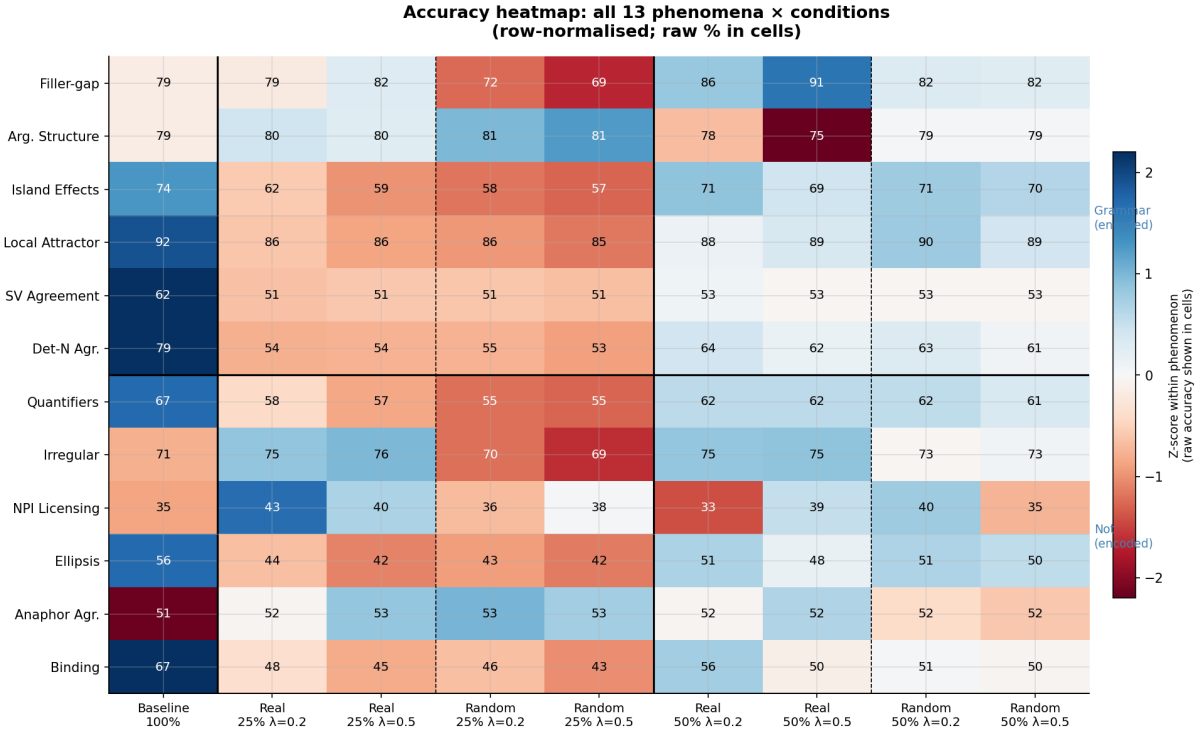


Figure 5: Accuracy heatmap across all 13 phenomena and 9 conditions (row-normalised; raw accuracy shown in cells). Conditions are organised by data fraction (25%, 50%) and constraint type (real / random), with the 100% baseline shown at left. Strong blue = high relative accuracy; strong red = low relative accuracy. The clearest cross-phenomenon contrast is between **Filler-gap** (row 1) — where Real 50%/$\lambda$=0.5 is the single darkest blue cell in the matrix (91%) — and **SV Agreement** (row 5), where all grammar-constrained conditions show uniform red regardless of whether the lexicon is real or random. This visual contrast directly reflects the paper's central finding: filler-gap improvement is linguistically specific, while SV damage is not. NPI Licensing (row 9) shows the highest condition-by-seed variance; binding and case show partial floor/ceiling effects at 25% data.

## B Grammar Construction

The PCFG was hand-crafted to capture core English phrase structure as described by Minimalist Grammar (Stabler, 1997), implemented in Chomsky Normal Form for compatibility with the CKY inside algorithm. We describe its components, construction rationale, and limitations.

**Production rules (16 rules).** The grammar encodes basic X-bar structure: sentences are TPs dominating NP subjects and VP predicates (TP $\rightarrow$ NP VP). VPs decompose into verbs with complements (VP $\rightarrow$ V NP; VP $\rightarrow$ V CP). Embedded clauses use CP $\rightarrow$ C TP, which also serves as the filler-gap licensing configuration: the complementiser position marks where a wh-element can be base-generated, with the TP providing the clause containing the gap. Determiner phrases use DP $\rightarrow$ D NP. Coordination, adjunction, and PP attachment are included as binary-branching rules.

**Terminal categories (12).** N (noun), V (verb), D (determiner), Adj (adjective), Adv (adverb), P (preposition), C (complementiser), Conj (conjunction), Pro (pronoun), Aux (auxiliary), Neg (negation), Wh (wh-word). These were chosen to cover the major lexical and functional categories relevant to the Zorro phenomena.

**Lexicon construction (7,230 entries).** Each word type in AO-CHILDES was assigned to one or more grammatical categories based on its predominant usage in child-directed speech, informed by the CHILDES %mor tier annotations. Polysemous items (e.g., "run" as N or V; "that" as D, C, or Pro) receive multiple category assignments; the SoftLexicon routing distributes probability mass across all assigned categories proportionally. Approximately 12% of word types have multiple category assignments.

**What the grammar encodes and does not encode.** The grammar directly encodes hierarchical phrase structure, basic subcategorisation (transitive vs. intransitive verbs), determiner-noun cooccurrence, and the CP $\rightarrow$ C TP configuration that licenses filler-gap dependencies. It does *not* encode: feature-checking or Agree relations (so agreement is structurally represented but not enforced), binding domains (Principle A/B/C), negative polarity licensing contexts, island constraints (which require specifying where movement is prohibited), or quantifier scope. This limitation is by design: phenomena not encoded in the grammar serve as natural controls for specificity.

## C Per-Seed Accuracy

Table 2 reports per-seed accuracy for focal phenomena at the key conditions discussed in the main text.

| Condition | Seed | Fill. | SVAgr | Bind. | NPI |
|---|---|---|---|---|---|
| Baseline | 42 | 79.8 | 60.2 | 61.7 | 29.1 |
| | 123 | 77.8 | 63.6 | 64.5 | 38.1 |
| | 456 | 77.7 | 60.1 | 71.9 | 33.6 |
| | 789 | 81.1 | 61.8 | 69.7 | 34.5 |
| | 1001 | 79.9 | 61.9 | 66.4 | 39.7 |
| Real (50%, $\lambda$=0.5) | 42 | 91.0 | 52.9 | 43.5 | 20.7 |
| | 123 | 90.6 | 52.8 | 62.6 | 25.5 |
| | 456 | 87.9 | 54.1 | 46.7 | 58.9 |
| | 789 | 93.5 | 51.5 | 50.4 | 53.9 |
| | 1001 | 91.8 | 51.9 | 45.1 | 37.5 |
| Random, lex 99 (50%, $\lambda$=0.5) | 42 | 86.6 | 54.1 | 54.0 | 37.5 |
| | 123 | 81.5 | 51.6 | 53.3 | 37.2 |
| | 456 | 84.2 | 53.8 | 52.3 | 49.9 |
| | 789 | 85.9 | 50.8 | 44.8 | 40.2 |
| | 1001 | 81.5 | 53.0 | 50.3 | 55.3 |
| Random, lex 2026 (50%, $\lambda$=0.5) | 42 | 85.9 | 53.4 | 51.3 | 27.3 |
| | 123 | 79.6 | 53.3 | 55.3 | 27.0 |
| | 456 | 80.8 | 53.3 | 50.2 | 40.7 |
| | 789 | 77.5 | 51.6 | 53.5 | 34.0 |
| | 1001 | 84.5 | 52.6 | 49.8 | 28.9 |
| Random, lex 31415 (50%, $\lambda$=0.5) | 42 | 81.4 | 53.7 | 51.1 | 24.4 |
| | 123 | 83.5 | 51.1 | 50.3 | 24.2 |
| | 456 | 85.0 | 54.0 | 47.5 | 35.2 |
| | 789 | 78.8 | 51.4 | 45.1 | 42.5 |
| | 1001 | 75.1 | 52.7 | 43.7 | 26.4 |

Table 2: Per-seed accuracy (%) on focal phenomena for key conditions. Real grammar filler-gap scores range 87.9–93.5%, with all 5 seeds above any of the 15 random grammar seeds at this condition. NPI licensing shows the bimodal pattern discussed in §3.3, with real grammar seeds splitting into low (20.7, 25.5) and high (37.5, 53.9, 58.9) clusters.