What Exactly do Children Receive in Language Acquisition? A Case Study on CHILDES with Automated Detection of Filler-Gap Dependencies

Zhenghao Herbert Zhou[1] **William Dai**[2] **Maya Viswanathan**[2]
Simon Charlow[1,3] **R. Thomas McCoy**[1,3] **Robert Frank**[1,3]
[1]Department of Linguistics, Yale University
[2]Department of Computer Science, Yale University
[3]Wu Tsai Institute, Yale University
{herbert.zhou, william.dai, maya.viswanathan,
simon.charlow, tom.mccoy, robert.frank}@yale.edu

Abstract

Children's acquisition of filler-gap dependencies has been argued by some to depend on innate grammatical knowledge, while others suggest that the distributional evidence available in child-directed speech suffices. Unfortunately, the relevant input is difficult to quantify at scale with fine granularity, making this question difficult to resolve. We present a system that identifies three core filler-gap constructions in spoken English corpora — matrix wh-questions, embedded wh-questions, and relative clauses — and further identifies the extraction site (i.e., subject vs. object vs. adjunct). Our approach combines constituency and dependency parsing, leveraging their complementary strengths for construction classification and extraction site identification. We validate the system on human-annotated data and find that it scores well across most categories. Applying the system to 57 English CHILDES corpora, we are able to characterize children's filler-gap input and their filler-gap production trajectories over the course of development, including construction-specific frequencies and extraction-site asymmetries. The resulting fine-grained labels enable future work in both acquisition and computational studies, which we demonstrate with a case study using filtered corpus training with language models.

1 Introduction

Language speakers possess abstract linguistic knowledge: instead of memorizing the low-level distribution of language inputs we receive, we instead formulate knowledge of abstract structures that generalizes beyond specific string combinations. How do children learn such abstract linguistic knowledge from limited input, and how do they generalize across similar constructions? The *argument from the poverty of the stimulus* claims that human learners are endowed with innate inductive biases which enable them to arrive at linguistic knowledge that goes beyond what is directly evidenced by the input data (Chomsky, 1986). From an alternative perspective, statistical learning theories argue that input frequency plays a central role: learners extract distributional regularities from input and use them to build increasingly abstract representations over time (e.g., Rowland et al., 2003).

Filler-gap dependencies (FGDs) have long been a central topic in syntactic acquisition. FGDs serve as a natural testbed for linguistic generalization: do learners formulate an abstract dependency between fillers and gaps that unifies superficially different syntactic constructions (Chomsky, 1977), or do they begin with shallow, lexically anchored patterns and only later generalize across items and constructions (Diessel and Tomasello, 2005)?

Adjudicating between the poverty of stimulus argument and statistical learning theories requires a characterization of the data that learners are exposed to at each stage of learning (e.g., Pullum and Scholz, 2002; Legate and Yang, 2002). A major challenge to this enterprise stems from the granularity and scale that are required. FGDs can occur in a variety of sentence constructions (e.g., matrix and embedded wh-questions, relative clauses, clefting and pseudo-clefting, and topicalization), and the filler's extraction site (e.g., subject versus object positions) also plays an important role in both acquiring and processing FGDs (Ambridge et al., 2015). Pearl and Sprouse (2013) manually annotated two corpora from the CHILDES database of child-directed speech (MacWhinney, 2000) with trace information that indicates filler-gap movement. Hsiao et al. (2023) similarly used manually annotated data to study the distributions of types of relative clauses. However, manual annotation is infeasible as the scale of the data increases. This issue could however be overcome with an automated system that was sufficiently accurate (see Carslaw et al. (2025) for a project in this direction but for clausal embedding rather than FGDs).

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 692–706

Construction	Extraction Site	Label	Sample Sentence
Matrix wh-questions	Subject	SMQ	**Who** __ praised the student?
	Object	OMQ	**Who** did the professor praise __?
	Adjunct	AMQ	**When** did the professor praise the student __?
	Polar	PMQ	Did the professor praise the student?
	Plain/Fragment	PlainMQ	**Who**?
	Cross-clausal	CC_	**Who** did the student say [that the professor praised __]?
Embedded wh-questions	Subject	SEQ	I wonder [**who** __ praised the student].
	Object	OEQ	I wonder [**who** the professor praised __].
	Adjunct	AEQ	I wonder [**why** the professor praised the student __].
	Polar	PEQ	I wonder [**whether** the professor praised the student].
Relative clauses	Subject	SRC	The professor [**who** __ praised the student] smiled.
	Object	ORC	The professor [**who** the student praised __] smiled.
	Adjunct	ARC	The day [**when** the professor praised the student __] was memorable.
	Possessive	PRC	The professor [**whose** student won the prize] smiled.
	Subject (reduced)	SRC_reduced	The professor [(**that/who** __ was) praised by the student] smiled.
	Object (reduced)	ORC_reduced	The professor [(**that/who**) the student praised __] smiled.

Table 1: Target filler-gap constructions and subtypes by extraction site, with gap positions underlined.

In this study, we develop an automated tool for annotating FGD constructions that allows us to characterize the distribution of various FGD constructions. Our specific contributions are:

1. We present an automated detection tool, leveraging the strength of both constituency and dependency parsing, for identifying three well-studied FGD constructions, matrix wh-questions, embedded wh-questions, and relative clauses, each subtyped by extraction site.
2. We apply our tool to 57 English corpora from the CHILDES database and present descriptive statistical analyses of the FGD distribution in both child-directed speech and children's production. The resulting dataset and the detector program is released on Github.[1]
3. We also outline how large datasets of FGDs with fine-grained annotations can be fruitfully applied both to address open questions in the acquisition literature as well as to study linguistic generalization in modern language models using methods such as filtered corpus training and input attribution.

2 Target Constructions

We focus on three core **constructions**—matrix wh-questions, embedded wh-questions, and relative clauses—all of which instantiate the hallmark of filler-gap dependencies: a leftwardly displaced filler (typically a wh-phrase or relative operator) that is interpreted in a position (the gap) where no pronounced constituent appears. Within each construction, we distinguish subtypes by **extraction site** (i.e., subject vs. object vs. adjunct), as these distinctions often correlate with distributional and processing differences and therefore are relevant for characterizing learners' input. This adds another level of granularity that no previous automated approaches have achieved. Table 1 summarizes the construction types and extraction-site subtypes we target, with fillers highlighted and gap positions explicitly marked by an underline to make the dependency visible in the surface string.[2] Although polar matrix questions do not involve wh-movement, they belong to the matrix question family and could affect generalization — it is potentially of interest to analyze whether PMQs behave like matrix wh-questions, and whether and how children generalize across PMQs and other question types. We thus included PMQs as one of our target constructions.

3 Detection Algorithms

We take a hybrid approach to identifying the presence and type of FGDs that combines the strengths of constituency and dependency parsing. We use the spaCy dependency parser (Honnibal et al., 2020) and the spaCy implementation of the Berkeley Neural parser (Kitaev and Klein, 2018; Kitaev et al., 2019), a widely used constituency parser based on a self-attentive architecture. Here we present the core steps of the detection algorithms and demonstrate the necessity of using both parse types.

[1]The code base is available at: https://github.com/herbert-zhou/filler_gap_detector_childes.git

[2]We excluded certain constructions that are closely related to some shown in Table 1, such as (i) **free relatives** (e.g., *I read what the professor wrote*) and (ii) **infinitival relative clauses** (e.g., *The professor picked a time for students to come*).

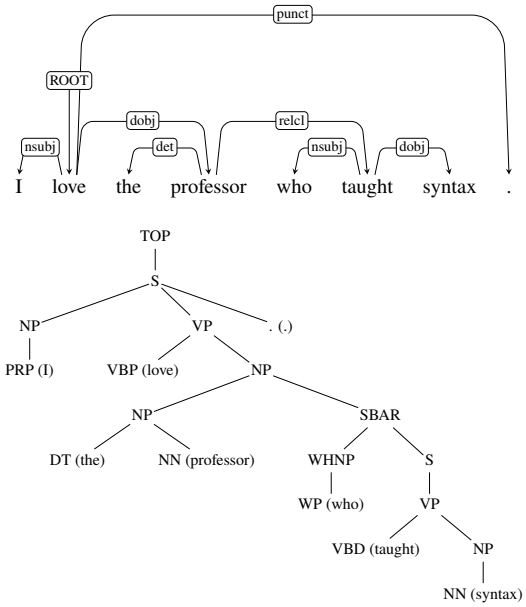

Figure 1: Dependency (top) and constituency (bottom) parses of a sentence containing a subject relative clause.

3.1 Detection Steps

Our detection algorithm consists of three detectors corresponding to the three constructions, each implementing a set of heuristics for construction detection and extraction site subtyping. Since a complex sentence can involve multiple constructions and subtypes, our algorithm applies all three detectors separately and outputs a list of all detected categories from Table 1. In this section, we illustrate core detection steps for relative clauses (see Appendix B for detailed elaboration), which can be generalized to other constructions. For the complete algorithm, we refer readers to the released code base, as it is infeasible to present it fully here.

Step 1: Core Structure Detection Starting with the constituency parse of the input sentence, illustrated in Figure 1, we recursively detect all occurrences of the local structural signature for relative clauses, `NP -> NP SBAR`. The `NP` child node denotes the noun phrase modified by the relative clause `SBAR`.

Step 2: Wh-category Identification For each detected `NP -> NP SBAR`, we look for a wh-category immediately dominated by `SBAR` and retrieve the entire span of the wh-phrase, in this case `WHNP` (*who*). For sentences where the wh-phrase is omitted, such as reduced relative clauses, we branch into a separate set of heuristics for detecting their properties using constituency parses: the absence of a wh-word makes dependency relations less robust.

Step 3: Extraction Site Inference To identify the extraction site, we look for the lowest `S` node under `SBAR` (to tolerate intervening material) and check whether there is an `NP` preceding a `VP`. If that is the case and the wh-word is `WHNP`, then it is likely that the extraction is from object position, as the subject position is occupied by another element. If instead no `NP` precedes `VP` below `S` (as in Figure 3), this suggests that extraction has taken place from subject position. Adjunct categories like `WHPP` and `WHADVP` follow similar logic of checking the existence of an `NP` at subject positions.

Step 4: Dependency Validation We verify the labels hypothesized from the constituency parse with the dependency parse. We start by verifying that the wh-word that was identified below the detected `SBAR` is dependent on the verb inside the relative clause (*taught*) with the relation `relcl`. In cases like Figure 3 where constituency heuristics labeled the sentence as containing a subject relative clause, we confirm that the wh-word has either an `nsubj` or `nsubjpass` dependency relation. Finally, we verify the relation between *taught* and the modified noun by checking if the noun that *taught* depends on is the noun in the `NP` from the constituency parse.

Additional construction-specific heuristics are added to deal with noise, parsing errors, and constructions that we intentionally omitted (see Appendix C for details). We note that there is unlikely to be a perfect solution that could deal with all edge cases, since adding more heuristics that fix one set of cases will lead to underdetection or overdetection in another set of cases.

3.2 Arguments for a hybrid approach

Constituency parsing and dependency parsing offer two ways of characterizing syntactic structure: constituency structures make clausal boundaries and complement types explicit, which is useful for detecting the construction types; dependency relations provide more direct access to head-dependent configurations that are useful for identifying extraction sites. While it is possible to analyze the majority of sentences using only one parsing strategy, there are situations where information offered by one parser is systematically insufficient and could lead to false detection. In contrast, combining information from both leads to more robust detection.

Consider the subject matrix question *Which book do I remember Mary wrote?* and its embedded question counterpart *I remember which book Mary*

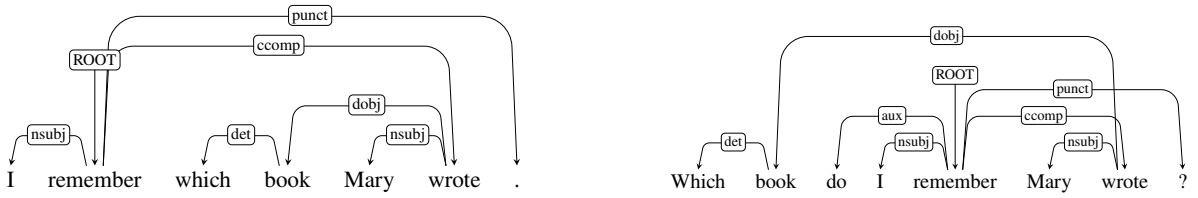

Figure 2: Dependency parses of an embedded question (left) and a matrix question (right).

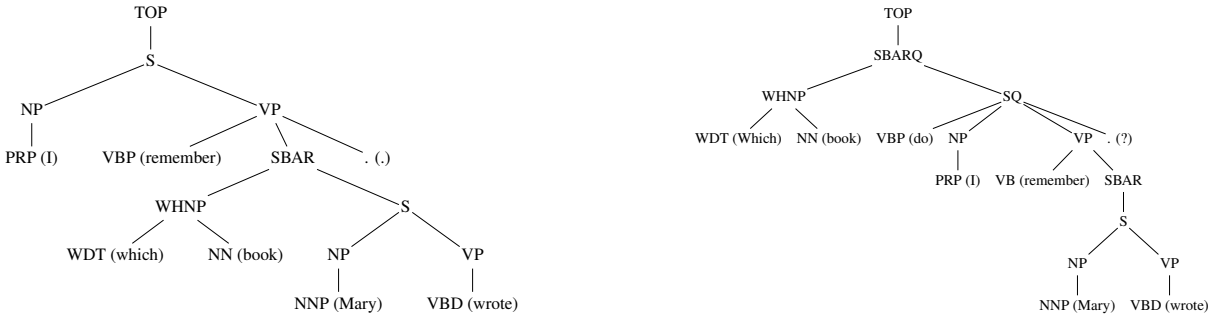

Figure 3: Constituency parses of an embedded question (left) and a matrix question (right).

wrote. As shown in Figure 2, both sentences exhibit superficially identical dependency relations (except for the extra auxiliary *do* in the matrix question). In particular, the embedded wh-clause in the matrix question is not distinguished from a genuine embedded question solely from the dependency structure — the dependency parse does not encode the scope of the wh-operator.[3] On the other hand, in Figure 3, the `VP -> VP SBAR` structure clearly marks embedded questions and the `SBARQ -> WHNP SQ` structure marks matrix questions. This is an example of utilizing additional information provided by constituency parsing to resolve the ambiguity between matrix and embedded questions that dependency parsing alone cannot adjudicate. See Appendix A for an example of how dependency parsing complements constituency parsing. An additional practical motivation is imperfections in the source data: child speech can be noisy, ungrammatical, and subject to transcription errors, which could lead to parsing errors. Combining information from both parses makes the detection more robust to noise.

4 Evaluation

To assess the effectiveness of our detection algorithm, we first conducted a small-scale, manual evaluation of the labels produced by our approach. We then compared our detection to information from a human-annotated corpus (Pearl and Sprouse, 2013).

[3]One could attempt to distinguish between these using the linear position of the wh-word, but doing so will introduce other complexities and ignores the natural structural generalization.

4.1 Manual Evaluation

As a sanity check, we first applied our detector to CHILDES and sampled 100 sentences from both child-directed speech and child speech that the detector identified as belonging to one of six core categories: the three FGD constructions, each with extraction sites of subject or object. Five human annotators with expertise in linguistics (faculty and students in linguistics) then provided binary assessments of whether each sentence contains the labeled construction. The detector was taken to have been accurate for a given sentence if a majority of the human judges responded positively to the assigned label. Results are shown in Table 2. Our detector achieved nearly perfect accuracy for child-directed speech and somewhat lower—but still fairly high—accuracy for child speech. Manual inspection of error cases indicated that child speech includes more edge cases due to ungrammatical, fragmented speech as well as transcription errors, which led to lower precision compared to child-directed speech.

4.2 Comparison with Pearl and Sprouse

We next performed a larger scale evaluation by comparing our detection results with human annotations of FGDs from Pearl and Sprouse (2013). This corpus includes annotations of 56,461 child-directed utterances from the Brown corpus (Brown, 1973) and the Valian corpus (Valian, 1991) with trace and coindexing information represented in constituency parses. This information allowed us to infer the construction and subtype information present in the

Category	Child-directed	Child speech
SMQ	1.00	0.83
OMQ	0.99	0.89
SEQ	0.95	0.95
OEQ	0.94	0.97
SRC	0.98	0.98
SRC reduced	0.98	0.92
ORC	1.00	0.91
ORC reduced	0.95	0.80

Table 2: Human annotated detector accuracy by construction and gap site (out of 100 sentences per category) for child-directed speech and child speech.

human annotations.[4]

We run our detector on the same corpus and use the human annotations as gold labels. This allows us to compute precision (how often our detector is correct when claiming a label) and recall (how often our detector claims a label given that human annotation produced that label). These results are shown in Table 3, together with the total number of sentences of each category.[5] We omitted PMQs in this comparision since they do not involve wh-movement and lack trace labels in human annotations, and we omitted PRCs since none occurred in human annotations. We also omitted reduced SRC and ORC because we found inconsistency in trace annotations.[6] All labels except for adjunct relative clauses achieved F1 scores above 0.8, with more than half above 0.9. No category had extremely unbalanced precision and recall, suggesting that our detector largely aligns with human annotations for deciding constructions and extraction sites. Given both evaluation results, we conclude that despite being imperfect, our automated detector is a viable solution for large-scale, fine-grained detection of our target structures.

5 A Case Study on CHILDES

CHILDES (MacWhinney, 2000) is the largest publicly available collection of longitudinal, naturalistic child and child-directed speech, making it uniquely suitable for quantifying the input distributions that could shed light on children's acquisition of filler-gap dependencies. By applying our detection algorithm to CHILDES, we move beyond small-scale, hand-annotated samples to get a better measure of how often different dependency types occur in both child and adult speech,[7] how these frequencies change with child age, and how they differ across extraction sites, which are known to modulate both learning and processing (see Section 7 for further discussion). These corpus-scale measurements serve two complementary goals: they provide a descriptive picture of what children hear and produce across time, and they yield construction-specific summaries that can be used to evaluate acquisition theories and to design computational studies with explicit input control, further discussed in Section 6. These goals are hard to achieve with the existing human-annotated corpora from Pearl and Sprouse (2013) since they only includes child-directed speech and cover a sparse age range.

Category	Precision (total)	Recall (total)	F1
SMQ	0.827 (712)	0.792 (716)	**0.809**
OMQ	0.908 (4950)	0.977 (4464)	**0.941**
AMQ	0.921 (1839)	0.957 (1746)	**0.939**
SEQ	0.925 (146)	0.894 (236)	**0.909**
OEQ	0.958 (642)	0.895 (1325)	**0.925**
AEQ	0.808 (339)	0.905 (924)	**0.854**
PEQ	1.000 (243)	0.905 (611)	**0.950**
SRC	0.924 (157)	0.883 (247)	**0.903**
ORC	0.901 (121)	0.810 (473)	**0.853**
ARC	0.842 (38)	0.713 (94)	0.772

Table 3: Precision (against parser labels), recall (against annotation labels), and F1 scores (bolded if > 0.8) by construction and extraction site.

We accessed the entire English-NA CHILDES database via `childes-db` (Sanchez et al., 2019), extracting 3,194,544 utterances distributed over 50,327 chat session transcripts from 57 corpora. Although target children's ages ranged up to 144 months, we decided to focus on utterances with target children in the range 3 to 80 months old since data outside this range was relatively sparse. After removing utterances missing age information, we retained 2,841,084 utterances (92.42% of total utterances) for further analyses. See Appendix E for more details of our data pre-processing.

[4] See Appendix D for details on their annotation schema and our inference method.

[5] Since our detector provides more categories (see Table 1 for the set of labels we consider) than what we inferred from tree annotations, we merged secondary categories with their main categories: e.g., those labeled as cross-clausal SMQs were treated as SMQs when computing the precision and recall.

[6] We note that for most of the categories for embedded questions and relative clauses, the total numbers of detected sentences for recall are greater than those for precision. This is because human annotations include sentences that our detector omitted by design (elaborated on in Section 2). For sentences labeled by human annotations but not detected by our detector, we manually checked and counted them as true positives if they were indeed among those omitted constructions.

[7] In the rest of this paper, we use the term *adult speech* instead of child-directed speech, although we acknowledge that not all child-directed speech is produced by adults.

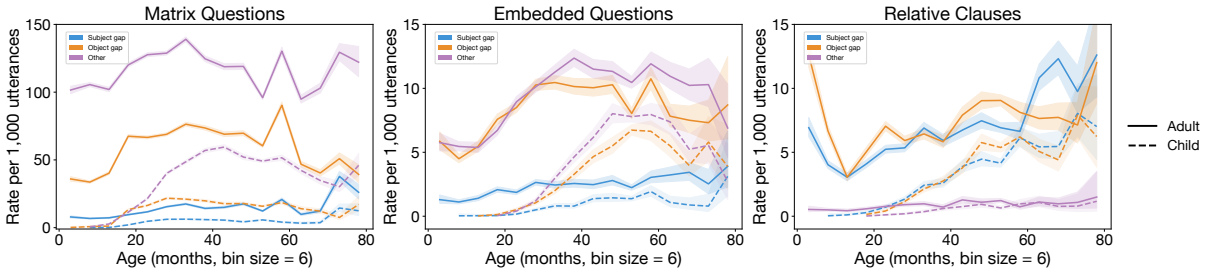

Figure 4: Adult (solid) and child (dashed) speech distributions across time by constructions and extraction sites. Uncertainty is shown as 95% Wilson intervals (wider in sparser bins).

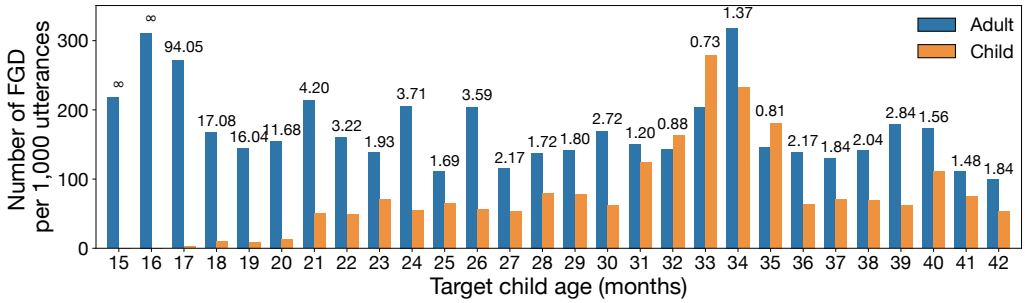

Figure 5: Per-1,000 utterance rates of all filler-gap dependency sentences received and produced by Laura.

5.1 Distributions Across Time

Understanding adult input and child output distributions over the course of development is central for evaluating acquisition theories, which differ in whether children's generalizations are driven by input frequency and in whether child production tracks input in construction-specific ways. We grouped utterances into 6-month age bins and, for each construction subtype, computed rates per 1,000 utterances for adult and child speech.

Figure 4 shows three robust patterns. First, matrix questions are roughly an order of magnitude more frequent than embedded questions and relative clauses (note the different y-axis scales), consistent with their lower structural complexity. Second, for both wh-question families, adjunct and polar questions outnumber subject and object extractions, and object extractions consistently exceed subject extractions. For relative clauses, subject and object extractions are comparable and both outnumber adjunct relatives. Third, for both wh-question families across ages, child production closely mirrors adult production in the rankings of the extraction sites. For relative clauses, subject and object RCs show a steady increase through about 55 months.

5.2 Total Exposures of One Individual Child

How much relevant input does an individual child receive over development, and what share do filler-gap constructions occupy? These questions matter for acquisition theories and for evaluating how human-like LLM training inputs are (see Section 6). As a reference point, we report one longitudinal case study as a reference for future work.

Among longitudinal English CHILDES corpora, Laura in the Braunwald corpus (Braunwald, 1971) has the largest amount of data: 75,740 adult and child utterances across 900 transcripts, spanning the age range of 15-77 months. We focus on the period of 15 to 42 months because later months are sparse. Figure 7 plots absolute counts of each FGD constructions on a log scale, with totals labeled. The resulting profile mirrors the global trends in Figure 4: matrix questions dominate embedded questions and relative clauses, and object extractions are consistently more frequent than subject extractions for wh-questions but not for relative clauses.

To contextualize these counts within overall exposure, Figure 5 shows the by-month proportion of utterances containing any FGD among all utterances, normalized as per-1,000-utterance rates. Each month is annotated with the adult versus child proportion ratio (values > 1 indicate higher FGD prevalence in adult speech). Despite substantial month-to-month variation in total utterances, FGD utterances remain a relatively small fraction in both streams. Laura's first detected FGD occurs at 17 months, and the adult versus child ratio decreases with age, dropping below 1 between 32-35 months, indicating that Laura's speech contains a higher pro-

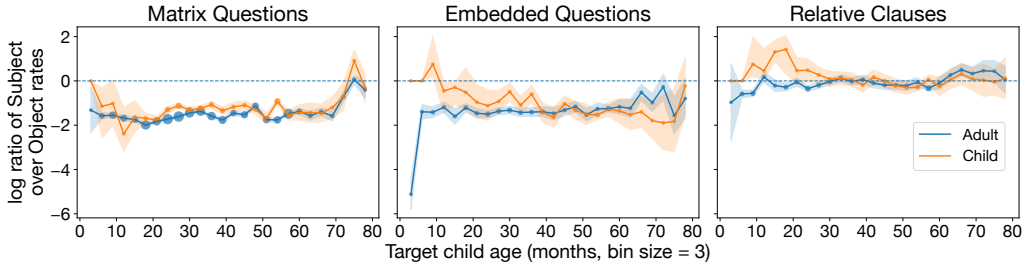

Figure 6: Subject- versus object-extracted log ratios for each construction across ages, binned by 3 months. Sizes of dot shades indicate number of total utterances within each bin.

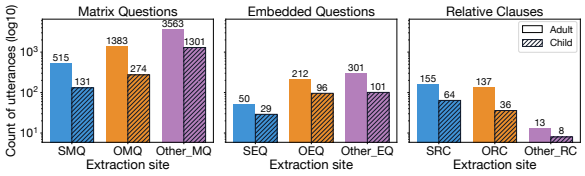

Figure 7: Number of utterances with targeted constructions received and produced by Laura in log scale.

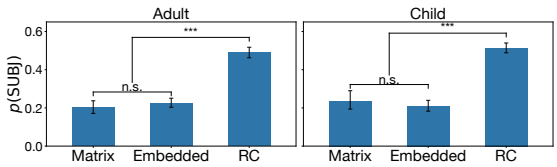

Figure 8: Cross-construction comparisons of subject share across age.

portion of FGD utterances than adult speech during that period. This increase might be taken to indicate a gradual acquisition of FGD constructions, but it is also consistent with a gradual increase in the use of sentences with higher syntactic complexity.

While longitudinal corpora cannot capture all utterances a child receives or produces, this case study provides stable *relative* distributions over constructions and extraction sites. These summaries can be interpreted as a scaled estimate of children's exposure and serve as an empirical target for acquisition theories and for computational modeling with explicit input control (Frank, 2023).

5.3 Subject versus Object Gaps

A central question in FGD acquisition is generalization: to what extent can children generalize knowledge of a construction to other related constructions? Extraction site is a natural case to test generalization. For instance, Gagliardi et al. (2016) studied whether gap sites show systematic extraction biases, and whether such biases differ by construction (see Section 7 for more discussion). We focus on subject- vs. object-extracted utterances as a first step. For each construction and age bin, we computed the log ratio of the rate-per-1,000 utterances for subject versus object extractions, $\log \frac{\#\text{SUBJ}+\epsilon}{\#\text{OBJ}+\epsilon}$, where #SUBJ denotes the number of subject-extracted utterances per 1,000 utterances. Values above 0 indicate a subject bias and values below 0 an object bias. Figure 6 shows a construction-specific profile: matrix questions are strongly object-biased in both adult input and child production; embedded questions exhibit a similar object bias; and relative clauses are closer to being balanced. Consistent with Figure 4, child trajectories broadly track adult trajectories within each family, suggesting that child speech preserves both the direction and relative strength of the input bias.

To compare biases across constructions, we computed the subject share $p(\text{SUBJ}) = \frac{\#\text{SUBJ}}{\#\text{SUBJ}+\#\text{OBJ}}$ for each age bin and each construction. We then averaged across bins that meet a minimum-count threshold, with pairwise comparison across constructions shown in Figure 8. We found no significant difference in p_{SUBJ} between matrix and embedded questions, with both being below 0.5 for adult and child speech; whereas relative clauses have a reliably higher subject share than both question types, with $p_{\text{SUBJ}} \approx 0.5$.[8] Together, these results suggest that extraction biases are not uniform across filler-gap constructions: these biases are similar across question types, but relative clauses differ from questions.

6 Linguistic Generalization in BabyLMs

6.1 Motivation

Beyond analyses of human acquisition, our detector also supports targeted tests of linguistic generalization in modern language models (LMs): to what extent do human-scale LMs generalize across related constructions, and how does the input affect such generalization? Since our detector partitions data by construction type and extraction site, it enables controlled interventions on training corpora that can help to answer such questions.

[8] See Table 6 in Appendix F for details.

Filtered Construction	# Sentences	# Tokens
Matrix questions	69,214	502,555
Embedded questions	3,765	51,930
Relative clauses	5,850	62,637

Table 4: Numbers of tokens and sentences removed from the training corpus for each construction family.

As a case study in this direction, we used the filtered corpus training paradigm, in which a model is trained on data with selected constructions removed to test whether generalization to held-out patterns depends on direct exposure or transfers across related constructions (Jumelet et al., 2021; Misra and Mahowald, 2024; Patil et al., 2024). Prior work has applied analogous ideas to filler-gap phenomena in restricted settings: Howitt et al. (2024) manipulate exposure via targeted augmentation (e.g., clefting and topicalization) and interpret improvements elsewhere as evidence for shared representations; Lan et al. (2024) extend augmentation to rarer movement phenomena (e.g., parasitic gaps and across-the-board movement) and find improved generalization; Chang et al. (2025) evaluated existing models on a suite of tests testing a broad range of constructions and reported that enriching the dataset with filler-gap dependencies leads to improvements in evaluations but does not lead to human-level performance. There has also been work characterizing the abstract causal mechanisms in LMs responsible for processing filler-gap dependencies (Baguraev et al., 2025; Desai and Nair, 2026). Our detector makes training-data interventions scalable across larger corpora, more construction families, and extraction-site subtypes, enabling more precise tests of generalization.

6.2 Experiment Setup

Corpus Filtering We conducted filtered corpus training on relatively small-scale LMs trained on child-directed language, specifically the CHILDES component from the 10-million-token version of the BabyLM dataset[9] (Warstadt et al., 2023). We used only child-directed speech as training input in order to simulate the types of sentences encountered during acquisition. This results in a total of 360,146 sentences (2,091,023 tokens). We experiment with filtering each of the three constructions, resulting in three conditions, one for each construction. The number of sentences and tokens *removed* for each filtered condition is given in Table 4. For each construction, we also produced a control dataset by removing the same number of sentences but randomly selected instead of targeting a specific construction. To ensure that the ablated and control datasets had the same number of sentences and comparable length distributions, we removed the same number of sentences in both cases and matched the removed sentences in number of tokens. This control condition allows us to attribute differences between models trained on the ablated versus control datasets specifically to the absence of the ablated construction, rather than to reduced training size or sentence length.

Evaluation Dataset To assess models' FGD knowledge, we evaluated models with 3432 synthetically constructed minimal pairs for matrix questions and 5000 synthetically constructed minimal pairs for each of embedded questions and relative clauses. Sentences in each pair differ in whether they contain a gap, and they share the same continuation after the (filled or unfilled) gap position, as shown in Table 5. We compared the probability of the continuation conditioned on the contexts of the two sentences and made binary judgments on whether the grammatical sentence has a higher continuation probability than the ungrammatical one. We constructed 15 templates and created the minimal pairs by filling in the templates with high frequency lexical items selected from the BabyLM dataset.[10] We manually checked their selectional restrictions to ensure semantic naturalness. See Appendix H for details of the evaluation dataset.

6.3 Results

For each ablated dataset and control dataset, we trained language model instances with Llama (Touvron et al., 2023) and GPT-2 (Radford et al., 2019) architectures with 15 different random seeds for each (see Appendix G for more details on training). Results for the Llama models are shown in Figure 9: compared to control models (trained on randomly filtered data), filtering each construction leads to a significant decrease in performance on the same construction, suggesting that the training input of the constructions is crucial to the acquisition of FGD knowledge. We further found that filtering matrix questions leads to significant degradation of performance for embedded questions and

[9]The dataset was accessed from `https://babylm.github.io/`

[10]We did not evaluate on subject relative clauses because the format of minimal pairs used with object relatives would not lead to a grammaticality difference in the case of subject relative clauses.

Construction × Extraction Site	Sample Evaluation Pairs
Matrix Question Object Gap	**What** will you build __ *today* *You will build __ *today* You will build *it* ***What** will you build *it*
Matrix Question Subject Gap	**Who** will __ *chase the doctor* *__ will *chase the doctor* It will *chase the doctor* ***Who** will it *chase the doctor*
Embedded Question Object Gap	I knew **what** you built __ *today* *I knew that you built __ *today* I knew that you built *it* *I knew **what** you built *it*
Embedded Question Subject Gap	I knew **who** __ *chased the doctor* *I knew that __ *chased the doctor* I knew that they *chased the doctor* *I knew **who** they *chased the doctor*
Relative Clause Object Gap	I knew the cake **that** *you made* __ *I knew that *you made* __ I knew **that** *you made it* *I knew the cake **that** *you made it*

Table 5: Example minimal pairs. Gap positions are marked by underline and continuations marked by italics.

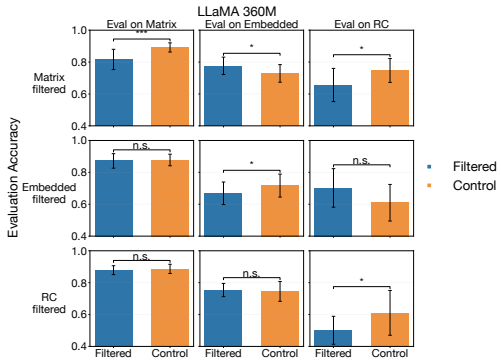

Figure 9: Cross-construction evaluations on filtered corpus training results of 15 Llama models. Evaluation accuracy means accuracy on the minimal pair tests.

relative clauses, while filtering out the other two constructions did not lead to significant degradation in the performance on other constructions. Our results support the claim in Boguraev et al. (2025) that more frequent constructions serve as sources of generalization to less frequent constructions, but not vice versa. Future studies can extend this experiment and investigate the more fine-grained generalization across extraction sites.

7 Prospective Applications

Testing human acquisition theories Work on FGD acquisition has emphasized two questions about what children learn from input. One line of research contrasts complexity-based acquisition orders with frequency-based accounts: classic proposals link earlier mastery to lower linguistic complexity (e.g., wh-pronominals before wh-sententials; general verbs before more restrictive ones; Bloom et al., 1982), whereas distributional approaches argue that acquisition is better predicted by how often children encounter the relevant patterns (e.g., wh-word-verb combinations; Rowland et al., 2003). Crucially, frequency effects are expected only when frequency is measured at the appropriate level of generalization (lexical items vs. constructions vs. subtype divisions; Ambridge et al., 2015). Relatedly, a second line asks which dimensions matter for characterizing FGDs and whether different dependency types share a common developmental profile: subject-object asymmetries can differ in frequency, timing, and processing difficulty (Ambridge et al., 2015; Atkinson et al., 2018), and direct comparisons between wh-questions and relative clauses report dissociations across ages and roles (Gagliardi et al., 2016; Sprouse et al., 2016).

Our detector makes these debates testable at scale. It enables frequency-based analyses at the construction-by-gap-site level, precisely where prior work predicts frequency should matter (Ambridge et al., 2015). It also supports large-scale comparisons of role-sensitive trajectories across dependency types. By avoiding subtype collapse, the resulting measurements allow sharper replications and more discriminative evaluations of competing acquisition accounts. Furthermore, future work could consider extracting lexical information (e.g., wh-word, head verb, modified nouns, etc.) and study lexical distributions associated with FGD constructions to test existing claims in the literature.

Generalization in computational models In addition to the filtered corpus training study presented in Section 6, our detector enables input-attribution analyses, which aim to quantify which parts of the input or which training examples most causally affect a model's prediction, typically by measuring output changes under perturbations or by computing gradient- or influence-based relevance scores (e.g., Koh and Liang, 2017; Hao, 2020; Grosse et al., 2023). Our detector enables attribution to aggregated groups of training data, linking model performance to construction-specific exposure.

Taken together, these applications illustrate how fine-grained control and measurement of filler-gap input can help distinguish learning of shallow lexical patterns from learning of abstract dependency representations, and they provide a concrete bridge between acquisition-motivated corpus analysis and modern computational modeling.

Limitations

Our detectors are imperfect due to the complexity of the data they take as input (e.g., there can be parser errors that mislead the detectors). Thus, the datasets we have extracted may not be suitable for research questions that require perfect identification of a given phenomenon. Nonetheless, we have shown that they are useful for illustrating broader statistical trends. Further, our detectors only identify some filler-gap phenomena and not others (e.g., cleftings and topicalization), and they are only defined over English. Future work could extend the approach to other phenomena and languages.

Our BabyLM filtered corpus training experiment only tested construction level corpus ablation. Future research could extend this approach by further analyzing the behavioral results and underlying mechanisms of controlled corpus filtering at different levels of granularity. One example is to filter the corpus by extraction site in order to test whether the FGD knowledge acquired for subject gaps can be generalized to object ones, and vice versa. Future work can also experiment with alternative control dataset schemes such as removing only from the set of sentences that lack certain construction labels, as well as designing more fine-grained evaluations by testing subtypes of each construction.

Acknowledgements

We would like to thank the anonymous reviewers from CoNLL and SCiL 2026 for their thoughtful comments, which helped us refine this paper. We would also like to thank Athulya Aravind for helpful discussion and suggestions. We thank the Yale Center for Research Computing for guidance and use of the research computing infrastructure, specifically the Grace cluster. Any errors are our own.

References

Ben Ambridge, Evan Kidd, Caroline F Rowland, and Anna L Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(2):239–273.

Emily Atkinson, Matthew W Wagers, Jeffrey Lidz, Colin Phillips, and Akira Omaki. 2018. Developing incrementality in filler-gap dependency processing. *Cognition*, 179:132–149.

Lois Bloom, Susan Merkin, and Janet Wootten. 1982. "wh"-Questions: Linguistic Factors That Contribute to the Sequence of Acquisition. *Child development*, pages 1084–1092.

Sasha Boguraev, Christopher Potts, and Kyle Mahowald. 2025. Causal Interventions Reveal Shared Structure Across English Filler–Gap Constructions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25032–25053.

Susan R Braunwald. 1971. Mother-child communication: The function of maternal-language input. *Word*, 27(1-3):28–50.

Roger Brown. 1973. *A first language: The early stages*. Harvard University Press, Cambridge, MA.

Iona Carslaw, Sivan Milton, Nicolas Navarre, Ciyang Qing, and Wataru Uegaki. 2025. Automatic extraction of clausal embedding based on large-scale English text data. In *Proceedings of the Society for Computation in Linguistics 2025*, pages 322–332.

Chi-Yun Chang, Xueyang Huang, Humaira Nasir, Shane Storks, Olawale Akingbade, and Huteng Dai. 2025. Mind the Gap: How BabyLMs Learn Filler-Gap Dependencies. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15060–15076.

Noam Chomsky. 1977. On Wh-movement. In *Formal Syntax*, pages 71–132. Academic Press, New York.

Noam Chomsky. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Convergence. Praeger, New York.

Atrey Desai and Sathvik Nair. 2026. Filling in the Mechanisms: How do LMs Learn Filler-Gap Dependencies under Developmental Constraints? *Preprint*, arXiv:2604.14459.

Holger Diessel and Michael Tomasello. 2005. A New Look at the Acquisition of Relative Clauses. *Language*, 81(1):1–25.

Michael C Frank. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992.

Annie Gagliardi, Tara M Mease, and Jeffrey Lidz. 2016. Discontinuous development in the acquisition of filler-gap dependencies: Evidence from 15-and 20-month-olds. *Language Acquisition*, 23(3):234–260.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying Large Language Model Generalization with Influence Functions. *Preprint*, arXiv:2308.03296.

Yiding Hao. 2020. Evaluating Attribution Methods using White-Box LSTMs. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 300–313, Online. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Katherine Howitt, Sathvik Nair, Allison Dods, and Robert Melvin Hopkins. 2024. Generalizations across filler-gap dependencies in neural language models. In *Proceedings of the 28th conference on computational natural language learning*, pages 269–279.

Yaling Hsiao, Nicola J Dawson, Nilanjana Banerji, and Kate Nation. 2023. The nature and frequency of relative clauses in the language children hear and the language children read: A developmental cross-corpus analysis of English complex grammar. *Journal of Child Language*, 50(3):555–580.

Rodney D Huddleston, Geoffrey K Pullum, and Laurie Bauer. 2002. *The Cambridge grammar of the English language*. Cambridge University Press Cambridge.

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

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

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

Julie Anne Legate and Charles D Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, 19(1-2):151–162.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk. Volume II: The Database*, 3rd edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929.

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

Lisa Pearl and Jon Sprouse. 2013. Computational models of acquisition for islands. *Experimental Syntax and Islands Effects*, pages 109–131.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Caroline F Rowland, Julian M Pine, Elena VM Lieven, and Anna L Theakston. 2003. Determinants of acquisition order in wh-questions: Re-evaluating the role of caregiver speech. *Journal of Child Language*, 30(3):609–635.

Alessandro Sanchez, Stephan C Meylan, Mika Braginsky, Kyle E MacDonald, Daniel Yurovsky, and Michael C Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, 51(4):1928–1941.

Jon Sprouse, Ivano Caponigro, Ciro Greco, and Carlo Cecchetto. 2016. Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory*, 34(1):307–344.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Virginia Valian. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition*, 40(1-2):21–81.

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

A An Example of How Dependency Parses Complement Constituency Parses

Section 3.2 illustrates an example of how constituency parsing complements dependency parsing

in distinguishing matrix questions from embedded questions. Here, we show an example of how the latter complements the former. For the object matrix question *What's your name* (with an assumed deep structure of *Your name is* __), as shown in Figure 10, the constituency parse has `SQ -> VP .` without an `NP` preceding `VP`, suggesting a subject gap. However, the dependency parse marks *name* as the `nsubj`, suggesting the correct extraction site of object position. Similar examples exist for subtyping relative clauses.

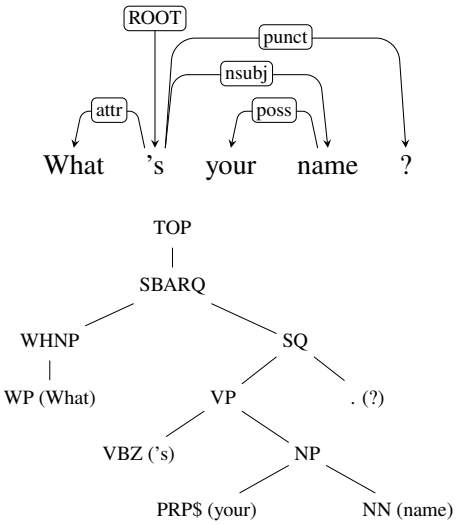

Figure 10: Dependency parse (top) and constituency parse (bottom) of object matrix question *What's your name?*

There are ambiguities on how to interpret extraction sites. For instance, one may analyze the deep structure of Figure 10 as __ *is your name*, so that this sentence is instead a subject matrix question. Such ambiguity is reflected in the inconsistent information given by the two parses. In these cases, we have to make a choice to implement the detector. Future studies can customize the heuristics to give different labeling if they choose different interpretations.

An additional practical motivation is imperfections in the source data: child speech can be noisy, ungrammatical, and subject to transcription errors, which could lead to parsing errors. Combining information from both parses makes the detection more robust to noise.

B An Elaboration of the Relative Clause Detection Algorithm

Figure 13 illustrates the decision flow for the types of relative clauses detected by the current detector.

C Additional Heuristics for Embedded Question Detection

The distinction between embedded questions and free relatives depends on the selectional properties of the matrix predicate: question-embedding predicates select interrogative complementizer phrases (CP), while predicates that do not select questions force the wh-clause to be interpreted as a free relative (e.g., Huddleston et al., 2002). For instance, *I **know** [what he ate]* contains an embedded question since the matrix verb *know* selects an interrogative CP, while the minimally different sentence *I **ate** [what he ate]* contains a relative clause since *eat* does not take an embedded question as its complement, so that this sentence is interpreted as *I ate [the thing that he ate]*. This poses a challenge to our approach to identifying such structures using constituency and dependency parsing, since the resulting representations do not distinguish these two kinds of sentences.

In order to distinguish embedded questions from free relatives, we applied a lexical filter following the application of our constituency- and dependency-based heuristics for the detection of embeddeded questions. In practice, we applied constituency and dependency parsing to the entire BabyLM 10M dataset to identify all potential sentences involving an embedded-question-like structure. We then grouped them by the matrix verb lemma and sorted the verb lemmas by frequency. For each of the top 50 verb lemmas (covering 95% of the total sentences with potential embedded question structure), we manually analyzed some example sentences containing that verb to decide if it can head embedded questions, resulting in the following list of verb lemmas: *know, see, tell, look, remember, wonder, guess, ask, say, forget, figure, understand, decide, show, watch, hear, think*. As the final step of our embedded question detection process, the detector only includes sentences with the selecting verbs in the list, and it additionally excludes sentences with *whatever, whoever, whomever, whichever, whenever*, or *wherever* in the embedded clause.

D Processing Treebank Annotations from Pearl and Sprouse

This appendix describes the process of comparing our detection results with Treebank annotations in Pearl and Sprouse (2013). Figure 11 shows a sample annotation for object matrix question *what*

should the birdie say?. Starting with manually corrected constituency parses from the Stanford parser[11], Pearl and Sprouse marked gaps by annotating a trace item `*T*-1` with an index at the gap positions, together with the type of trace `-NONE-ABAR-WH-` (indicating a wh-movement), and they annotated the filler node `WHNP-1` to coindex with the trace. Relevant to the current study, there were 8157 wh-movement traces and 1,804 relative clause traces from their annotations.

```
(ROOT
  (SBARQ
    (WHNP-1-<INANIM>-<THEME-V1>
      (WP what))
    (SQ
      (VP
        (MD should)
        (NP-<ANIM>-<AGENT-V1>
          (DT the)
          (NN birdie))
        (VP
          (VB-<V1> say)
          (NP
            (-NONE-ABAR-WH- *T*-1)))))
    (. ?)))
```

Figure 11: Sample annotation with trace (bolded) from Pearl and Sprouse (2013).

The trace type and coindexing information enable us to determine extraction sites with high precision, which neither constituency nor dependency parsing explicitly provides. To directly compare annotated sentences with our detection results, we started with the Treebank-formatted annotations and developed simple heuristics to classify sentences into our labels given the parent nodes of the filler and the gap as well as their constituent types. We illustrate with the same example in Figure 11. We first checked the parent node of the gap, in this case `NP`, as well as the filler category, in this case `WHNP`, suggesting that the wh-word moves from an argument position. Next, we checked the first sentential complement node above the filler, in this case the `SBARQ` directly under `ROOT`, suggesting a matrix question. Finally, we traced down the sister node of the filler, in this case, `SQ`. Similar to Step 3 in Section 3, we checked whether there is an `NP` preceding the `VP` node, neglecting intermediate layers that contains auxiliary categories like `MD` (for modal, *should*). In this case, the `NP` *the birdie* precedes `VP` *say*, this indicates that the subject position was already filled, suggesting an object gap.

[11]https://nlp.stanford.edu/software/lex-parser.shtml

E Processing CHILDES Data

Given the objective of separating child versus adult speech, we extracted each speaker's age information from each transcript and corrected the speaker age of each utterance's metadata to reflect speech from children that were not labeled as `Target Child`.

Since the input distribution is likely to diverge between spontaneous production and targeted activities such as book reading, we further extracted metadata for each transcript via directly accessing raw files other than `.cha` files from the TalkBank.[12] This allowed us to do fine-grained disaggregations by longitudinal versus cross-sectional studies, as well as by daily, spontaneous productions versus various elicited productions.

See the released code base for the detailed transcription-level metadata. Future studies are encouraged to further explore more fine-grained results after disaggregation by types of studies and activities.

F Supplementary Data for CHILDES Statistics

Pair	Adult Δ_{subj} [95% CI]	Child Δ_{subj} [95% CI]
MatrixQ∼EmbQ	[-0.057, 0.013] (n=25)	[-0.032, 0.102] (n=19)
MatrixQ∼RC	[-0.321, -0.256] (n=25)	[-0.309, -0.193] (n=20)
EmbQ∼RC	[-0.288, -0.239] (n=25)	[-0.322, -0.258] (n=19)

Table 6: Across-age summary of Δ_{subj} for each construction pair. Intervals are bootstrap 95% CIs; parentheses give the number of age bins passing the minimum-count filter.

To complement the pairwise significance test of how object biases compare across constructions, we also summarize direct assessment of *cross-construction* differences in Table 6. For each age bin and each pair of constructions, in addition to the subject share $p(\text{subj}) = \frac{\#\text{subject extractions}}{\#\text{subject extractions}+\#\text{object extractions}}$ described in Section 5.3, we also took $\Delta_{\text{subj}}(c1, c2) = p_{c1}(\text{subj}) - p_{c2}(\text{subj})$ (where c stands for a construction), averaging across bins that meet a minimum-count threshold. This cross-family summary shows that matrix and embedded questions are closely matched (mean $\Delta_{\text{subj}} \approx 0$), whereas relative clauses reliably differ from both question types, exhibiting a higher subject share ($\Delta_{\text{subj}} < 0$ for MatrixQ∼RC and EmbQ∼RC). Together, we conclude that extraction biases are not uniform across filler-gap constructions.

[12]https://talkbank.org/childes/access/Eng-NA/

G Additional Details for BabyLM Filtered-Corpus Training

We conducted the filtered corpus training experiment outlined in Section 6 with two model architectures, GPT-2 and Llama. The relevant model configurations and training (hyper)parameters are summarized in Table 7.

(Hyper)parameter	GPT-2	Llama
Data		
Sequence length	128	128
Eval samples	8192	8192
Model		
# Parameters	705M	360M
Hidden size	1536	1024
Intermediate size	3072	3072
# Layers	24	24
# Heads	16	8
Residual dropout	0.0	–
Attention dropout	0.0	–
Embedding dropout	0.0	–
Training		
Learning rate	2.5×10^{-4}	3.0×10^{-4}
Batch size	128	128
# Epochs	6	6
Grad. accumulation	16	8
Warmup steps	300	300
FP16	True	True

Table 7: Summary of configuration and training hyperparameters.

In addition to the previously reported Llama figure, we also have results for GPT-2, which are shown in Figure 12. Consistent with results for Llama (see Figure 9), we observe a significant decrease in performance when filtering with matrix questions and relative clauses and evaluated on the same construction. We nonetheless did not observe significant effects for embedded questions. One potential explanation is that matrix questions become the source of generalization as they have the highest frequency in the training data, compared to the other constructions. That is, it is easier to generalize from matrix questions to other constructions, but not in the other direction.

H Additional Details for the BabyLM Evaluation Dataset

To construct the evaluation dataset, we created templates for each construction type. Each template consists of three sentence components: two contexts and one continuation whose probability we

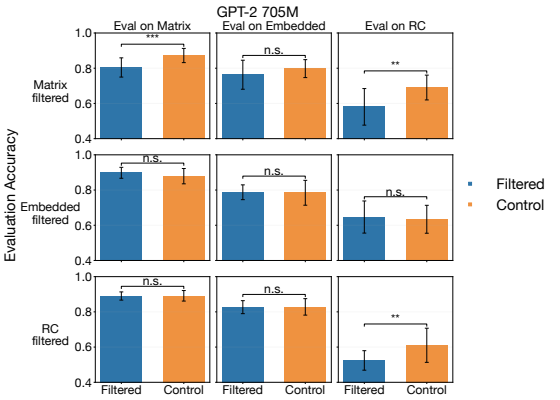

Figure 12: Cross-construction evaluations on filtered corpus training results of 15 GPT-2 models.

test following each of the contexts. In each case, the continuation is grammatical after context 1 but not after context 2. See our GitHub for all of the templates and the process for filling in the templates.

Note also that we did not include subject relative clauses, because there is ambiguity that is not caused by object relative clauses. For example in a minimal pair such as "I knew the author that wrote this book" and "I knew that wrote this book", the second sentence could be considered grammatical if we consider "that" as a demonstrative. To avoid this ambiguity, we did not include subject relative clauses in the evaluation dataset.

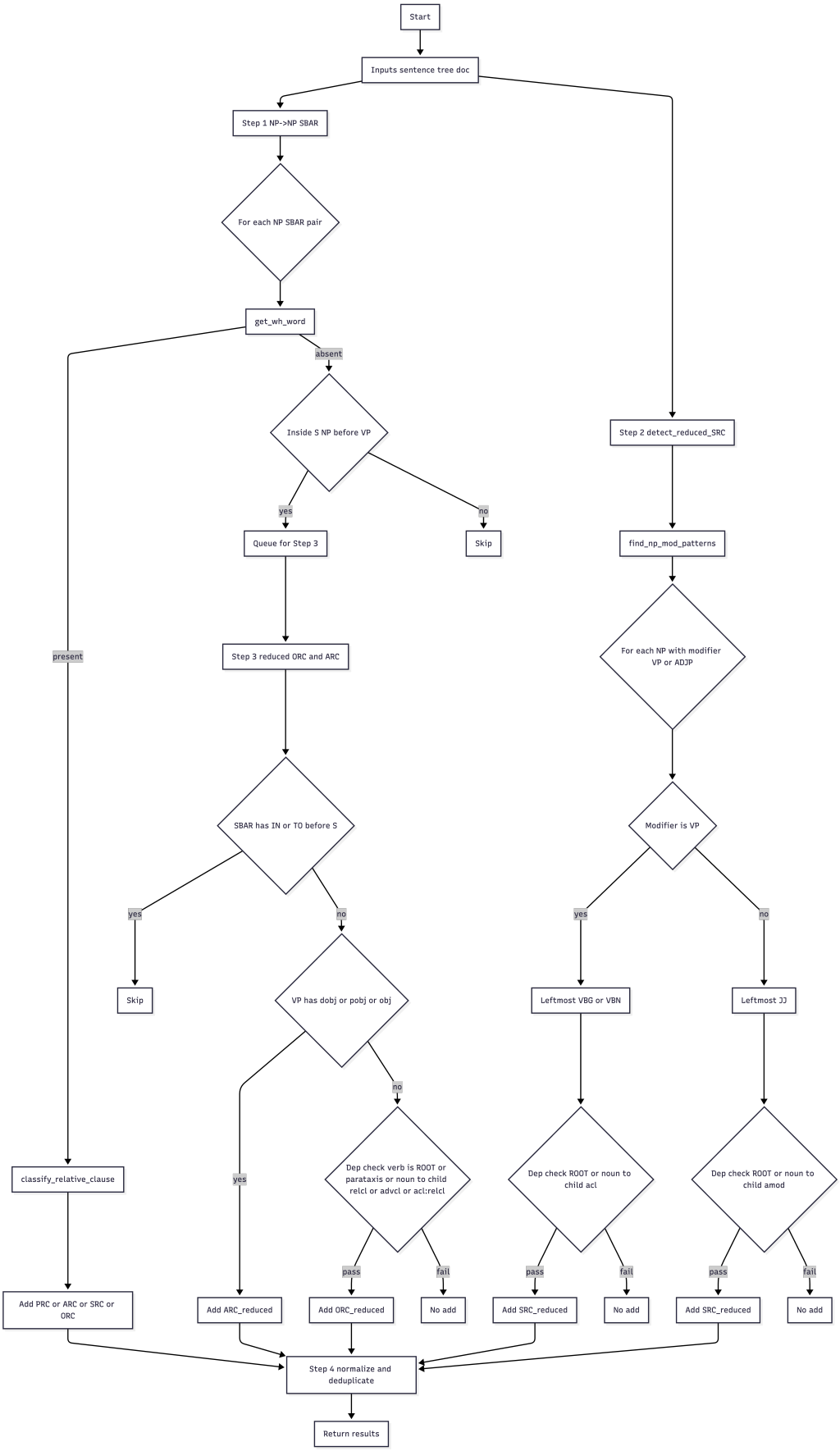

Figure 13: A flowchart for the relative clause detection process.