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

1 Motivation

Language learners generalize beyond the surface strings they hear, but theories posit various factors on what supports this generalization. The *poverty-of-stimulus* view posits innate biases that outstrip available evidences, while statistical learning accounts emphasize how learners exploit distributional regularities in caregiver input. Filler-gap dependencies (FGDs) are a useful testbed because they span multiple constructions (e.g., wh-questions and relative clauses) and exhibit extraction-site asymmetries (e.g., subject vs. object gaps) that affect learning and processing. Yet the relevant input is hard to quantify at scale with the granularity needed to adjudicate these accounts.

In this study: (i) we present an automated, fine-grained detection tool for three FGD constructions subtyped by extraction sites; (ii) we apply it to the entire CHILDES database to characterize FGD distributions children receive and produce; and (iii) we outline how our tool can be applied to both address open questions in acquisition and to study linguistic generalization in modern language models.

2 Target Constructions and Our Detector

We present an automated detector that identifies three core spoken-English FGD families in CHILDES —matrix questions (MQs), embedded questions (EQs), and relative clauses (RCs) — and subtypes each instance by extraction site (subject vs. object vs. adjunct; plus polar and reduced variants where applicable). Our detector leverages the strengths of both constituency parsing (to extract clausal boundaries and complement types) and dependency parsing (to access head-dependent configurations informative about extraction sites). For each construction, we: (i) detect a local constituency signature (e.g., `NP` → `NP SBAR` for RCs); (ii) retrieve the wh-phrase's category (e.g., `WHNP`); (iii) infer the likely gap site using structural cues inside the clause; and (iv) validate the hypothesized label with dependency relations (e.g., `relcl`, `nsubj`, `dobj`). Combining both parsing strategies improves robustness by handling cases where either parse alone is systematically insufficient, and by reducing false detections caused by noisy and error-prone speech transcriptions.

3 Evaluation

We evaluate our detector by comparing against the manually annotated treebank from Pearl and Sprouse (2013), which contains 56,461 child-directed utterances from two CHILDES corpora, annotated with trace information.

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

Precision and recall scores are shown in Table 1. Across most categories, the detector achieves strong agreement with human judgments and yields balanced precision and recall when compared to trace-derived gold labels. Despite being imperfect, our automated detector is viable for large-scale, fine-grained detection of our target structures.

4 A Case Study on CHILDES

We apply our detector to 57 English-NA CHILDES corpora accessed via `childes-db`, yielding 2.84 millions of utterances after filtering missing ages. We summarize: (i) developmental distributions (adult input vs. child production) for all utterances;

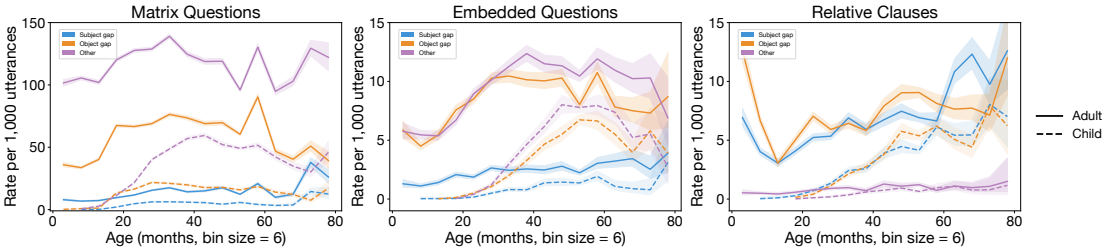

Figure 1: Adult (solid) and child (dashed) speech distributions across time by constructions and extraction sites. Uncertainty is shown as 95% Wilson intervals (wider in sparser bins).

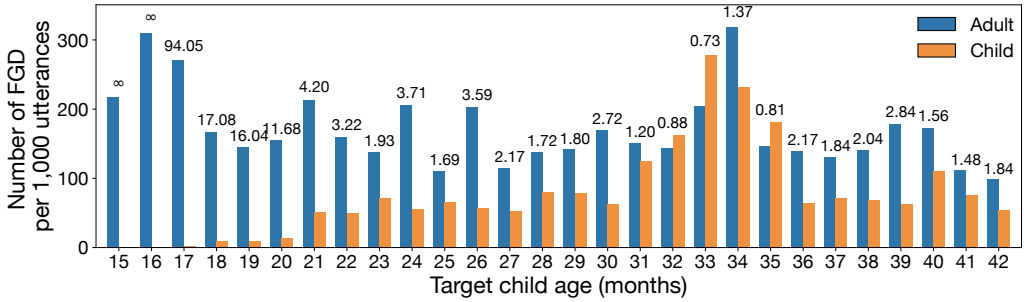

Figure 2: Per-1,000 utterance rates of all filler-gap dependency sentences received and produced by Laura.

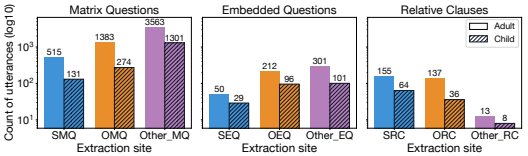

Figure 3: Number of utterances with targeted constructions received and produced by Laura in log scale.

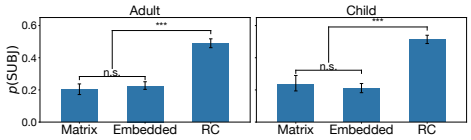

Figure 4: Cross-construction comparisons of subject share across age.

(ii) distributions of total exposures of one individual child Laura (75,740 utterances available from a longitudinal study); and (iii) extraction-site asymmetries across constructions. Three robust trends emerge: (1) matrix questions are substantially more frequent than embedded questions and relatives (Figure 1 and 3); (2) within question families, object gaps outnumber subject gaps, whereas relatives are closer to balanced (Figure 4 compares the subject share $p(\text{SUBJ}) = \frac{\#\text{SUBJ}}{\#\text{SUBJ}+\#\text{OBJ}}$ across the three constructions); and (3) child production broadly tracks adult input for question families, while relative clauses increase more gradually (Figure 1, 2, and 3). These large-scale counts offer a descriptive baseline for "what children hear" and "what children say" at a construction-by-gap-site resolution.

5 Prospective Applications

For open questions in acquisition, our detector enables frequency analyses at the level where frequency is predicted to matter (construction vs. extraction site vs. lexical combinations), supporting sharper tests of complexity-based vs. distributional accounts and role-sensitive comparisons across dependency types (Ambridge et al., 2015). For computational learners, our detector enables targeted input control for testing generalization in language models, including filtered-corpus training (Patil et al., 2024) to probe transfer across related FGDs. It also supports input-attribution analyses (Koh and Liang, 2017) that connect a model's behavior to the subsets of training data most responsible for it.

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.

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

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.