What Exactly do Children Receive in Language Acquisition?

A Case Study on CHILDES with Automated Detection of Filler-Gap Dependency

paper (CoNLL) dataset

Zhenghao Herbert Zhou & William Dai & Maya Viswanathan & Simon Charlow & R. Thomas McCoy & Robert Frank
Department of Linguistics & Department of Computer Science Yale

TL;DR

- **Fine-grained input matters: children's filler-gap dependency (FGD) exposure differs by construction and extraction site, which shape linguistic generalization in acquisition. In the paper:**
  - ✳A detection tool for fine-grained FGD, combining constituency and dependency parsing.
  - ✳A large-scale analysis of FGD distribution on CHILDES (with dataset).
  - ✳A BabyLM filterd-corpus training study on generalization.

Motivations

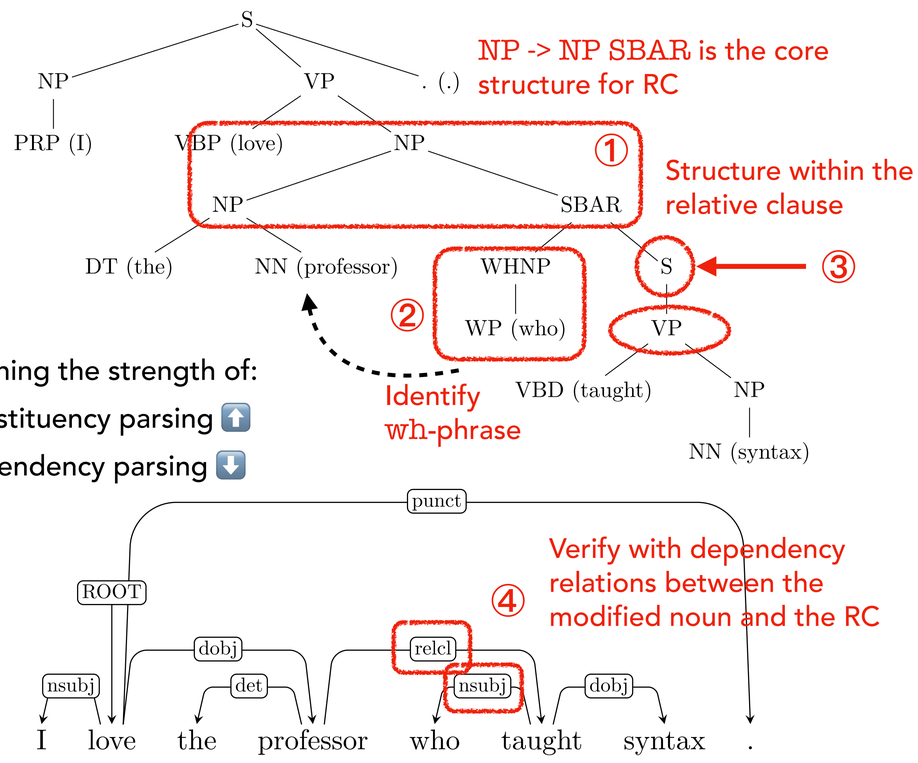
- Do children learn an abstract FGD that generalizes across constructions, or do they first learn construction-specific, lexically anchored patterns?
- We need *fine-grained, large-scale* data to answer those questions!
- Prevous hand annotations are hard to scale, and we provide an **automated, fine-grained FGD detection tool.**

Target Constructions

Matrix Questions	**Subject**: *Who* __ praised the students? **Object**: *Who* did the professor praise __? **Adjunct**: *When* did the professor praise the student __?
Embedded Questions	**Subject**: I wonder [*who* __ praised the student]. **Object**: I wonder [*who* the professor praised __]. **Adjunct**: I wonder [*why* the professor praised the student __].
Relative Clause	**Subject**: The professor [*who* __ praised the student] smiled. **Object**: The professor [*who* the student praised __] smiled. **Adjunct**: The day [*when* the professor praised the student __] was memorable.

- *For extraction site, we also detected polar questions (yes/no questions) and reduced SRC and ORC (whiz-deletion, contact RC).*

Detection Steps



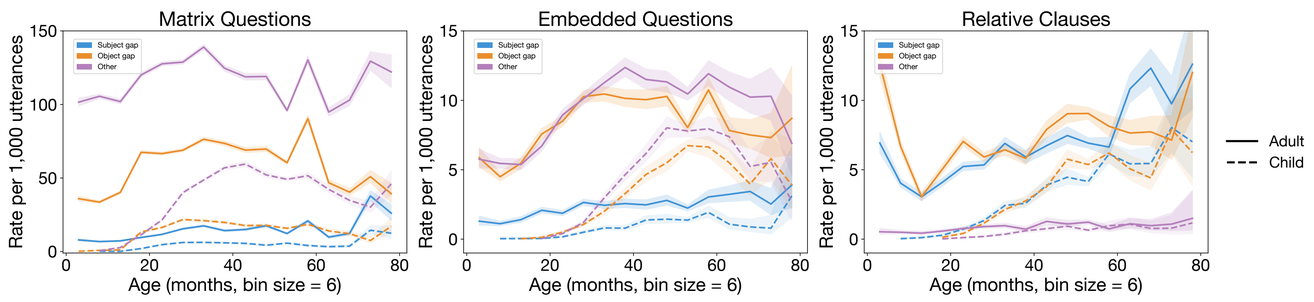
Combining the strength of:

- Constituency parsing ⬆️
- Dependency parsing ⬇️

- **Step 1: Core sturcture detector**
  - Recursively look for `NP -> NP SBAR` structure for RC;
- **Step 2: Wh-category Identification**
  - Look for a wh-catetory immediately dominated by `SBAR` and retrieve the span of the wh-pharse; here, `WHNP` (*who*);
- **Step 3: Extraction Site Inference**
  - Look for the lowest `S` node under `SBAR`, check whether there is an `NP` preceding a `VP`; here no means likely subject-extracted.
- **Step 4: Dependency Validation**
  - Verify the `relcl` relation between embedded verb and modified noun, as well as the `nsubj` relation between embedded verb and the *wh*-head.

✳ Verified against human annotated treebanks from Pearl & Sprouse (2013) and achieved > 0.85 F1 scores for most constructions.

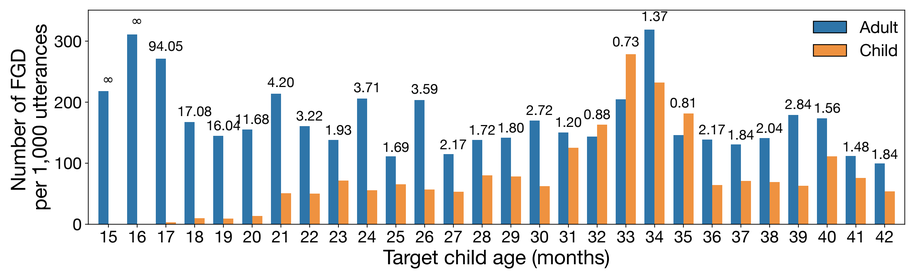
Developmental Distributions



- **Children's FGD productions broadly track adult input across age, while matrix questions are far more frequent than embedded questions and relative clauses.**

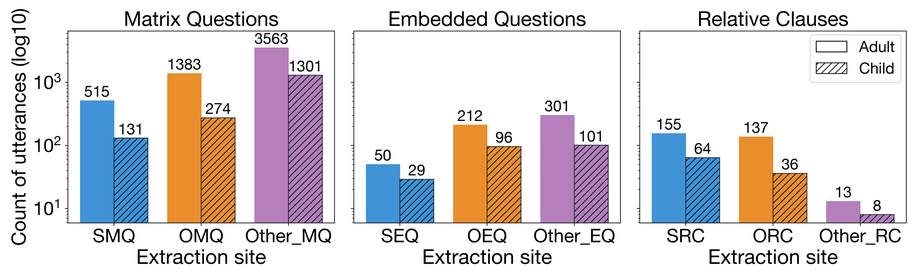
Input and Production of an Individual Child

We studied child Laura from the Braunwald corpus (75,740 utterances across 15-77 months):

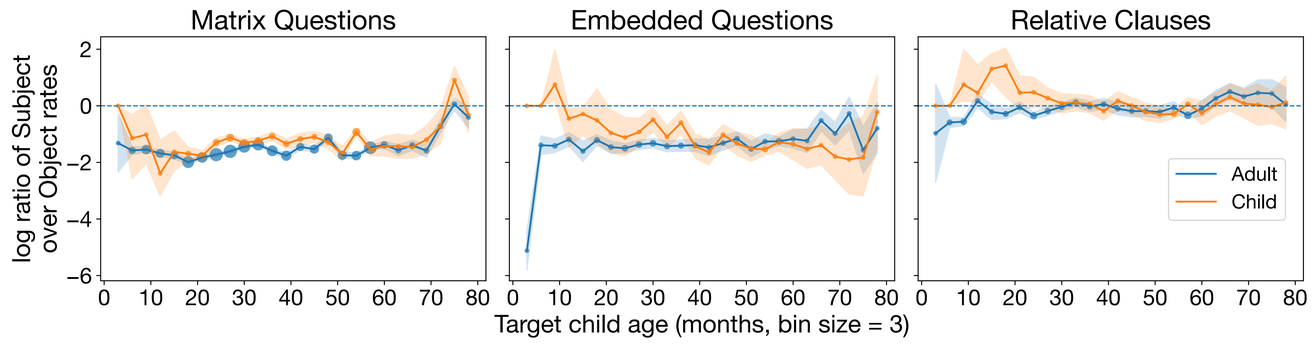


- **FGD utterances remain a small but increasing share of child speech.**

- **Matrix questions dominate both input and production, while embedded questions and relative clauses are much rarer.**



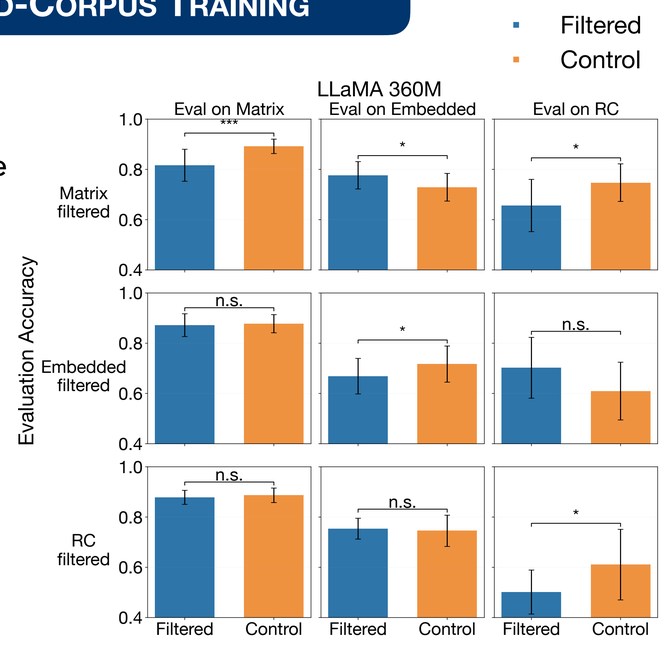
Subject vs. Object Extraction Asymmetry



- **Wh-question constructions are consistently object-biased, whereas relative clauses are much closer to balanced between subject and object extractions.**

BabyLM Filtered-Corpus Training

- **Question** = does FGD knowledge generalize across constructions?
- **Method** = filtered-corpus training: removing one FGD construction from training data and evaluate perforamcne in other constructions;
- **Results**:
  - Removing a construction hurts performance on that construction
    → direct exposure matters for FGD learning.
  - Removing matrix questions (the most frequent) weakens performance on the others, but not vice versa
    → asymmetric generalization from frequent to rarer FGD constructions.



Contact: herbert.zhou@yale.edu