How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech

Aditya Yedetore*[1], Tal Linzen[2], Robert Frank[3], R. Thomas McCoy*[4]
[1]Boston University, [2]New York University, [3]Yale University, [4]Princeton University
*Work done while at Johns Hopkins University

Overview

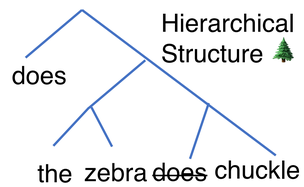
Approach: We trained LSTMs and Transformers on the type of linguistic input that children receive.
Finding: These models capture the surface statistics of the training data but fail to generalize as humans do on the hierarchically governed syntactic phenomenon of English yes-no questions.
Implications: Human-like generalization from text alone may require biases stronger than the general sequence-processing biases of standard neural networks.

Background

Syntax is driven by hierarchical structure, yet we typically encounter sentences as linear sequences of words.

Linear Input ➡️

[*does*, *the*, *zebra*, *chuckle*, …] → Hierarchical Structure 🌲



What leads kids to recognize the hierarchical nature of the languages they acquire?

Possibilities:

👶: Humans have a hierarchical inductive bias (Chomsky 1965)

📖: There is clear evidence for hierarchical structure in the input (Lewis & Elman 2001)

Classic case study in hierarchical generalization: yes/no questions

(1) Type of evidence present in a child's input:
a. Those **are** your checkers.
b. **Are** those your checkers?

Such examples are consistent with two rules:

- **HierarchicalQ**: The auxiliary at the start of a question corresponds to the **main** auxiliary of the corresponding declarative.
- **LinearQ**: The auxiliary at the start of a question corresponds to the **first** auxiliary of the corresponding declarative.

Yet: Children reliably favor the hierarchical generalization

(2) Disambiguating examples (not present in children's input)
a. The boy who **has** talked **can** read.
b. **Can** the boy who **has** talked __ read?
c. ***Has** the boy who __ talked **can** read?

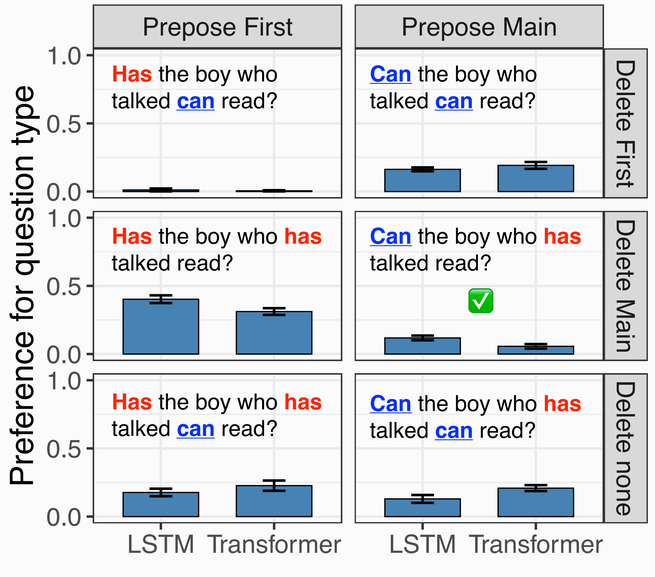
Our research question: when trained on data like children receive, will LSTMs and Transformers (learners without hierarchical biases) generalize hierarchically?
- Tests if children's input contains clear cues to hierarchical structure

Experiment 1: Relative Acceptability

- **Models 🤖:** LSTMs and Transformers
- **Training set:** 8-million-word corpus from CHILDES

- **Results**
  - **Language model quality:** Our 🤖s got a perplexity near 20; a 5-gram model baseline got 24.37
  - **General syntactic evaluation:** On the Zorro dataset of targeted syntactic evaluations each of our 🤖s scores well on at least some syntactic evaluations
  - **Yet on an evaluation of yes/no questions:** none of the 🤖s display preferences consistent with the correct, fully-hierarchical generalization.
    - Preference for question types measured by perplexity: lower perplexity = greater preference

Example Declarative: The boy who **has** talked **can** read.

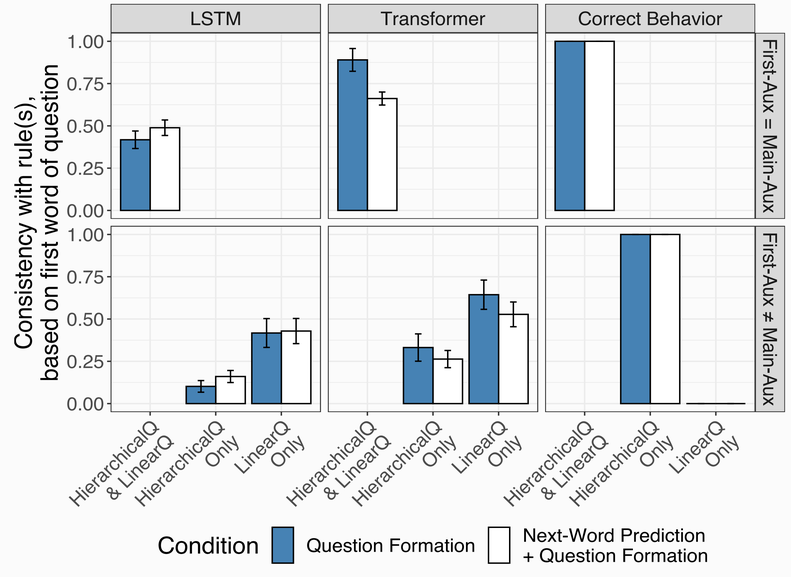


Experiment 2: Question Formation

- **Models 🤖:** LSTMs and Transformers
- **Training regimen:**
  - **Pretraining**: next-word prediction on 8-million-word corpus from CHILDES
  - **Finetuning**: transformation of declarative sentences into questions on 10,000 questions from CHILDES
    - i.e., given *he can see* our 🤖s must produce *can he see?*

Evaluation datasets:
- First-Aux = Main-Aux: examples like in (1) where **LinearQ** and **HierarchicalQ** make the same predictions
- First-Aux ≠ Main-Aux: examples like in (2) that disambiguate **LinearQ** and **HierarchicalQ**

Results
- 🤖's performed more consistently with **LinearQ** than **HierarchicalQ** when evaluated on their accuracy on the first word of the question.



Takeaways

For LSTMs and Transformers: at least when learning from text alone, LSTMs and Transformers do not display human-like language learning (🤖 + 📖 = ➡️, 🤖 ≠ 👶).

For the Poverty of the Stimulus Debate: The biases sufficient for capturing the statistical patterns in the training data are not likely sufficient for hierarchical generalization: stronger biases may be necessary (💪👶).

For the Type of Training Data: Prosody, visual information, meaning, and/or social interaction might aid hierarchical generalization (🤖 + 📖 + 👂 + 👁️ + 🌎 + 🏘️ = 🌲?)

Portions of this research were supported by the National Science Foundation (NSF) under grants BCS-2114505, BCS-1919321, and Graduate Research Fellowship Program grant no. 1746891.