Child-Directed Language Does Not Consistently Boost Syntax Learning in Language Models

Francesca Padovani & Jaap Jumelet & Yevgen Matusevych & Arianna Bisazza | GroNLP, University of Groningen | f.padovani@rug.nl

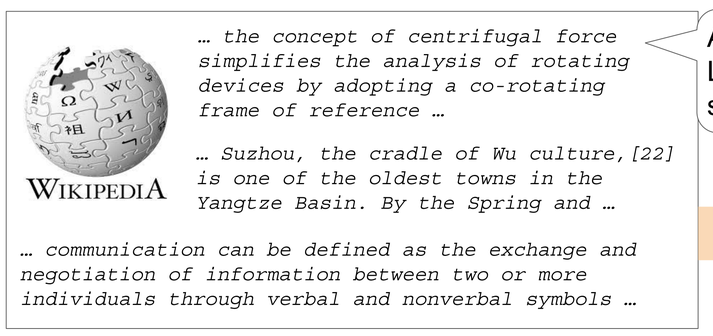
EMNLP 2025 Suzhou, China 中国苏州

Motivation

LMs have human-like linguistic skills, but are trained in non-human ways. What if we trained LMs in more **developmentally plausible** ways?

Training on **Child-Directed Language (CDL)** has been argued to **benefit syntax learning** in English LMs (Huebner et al., 2021; Salhan et al. 2024), **but …**

Are CDL benefits consistent across models, languages, and syntactic benchmarks?



We find no consistent benefit of training LMs on Child-Directed Language for syntactic learning; effects vary by model architecture, language, and benchmark, and are partly driven by lexical frequency and corpus composition.

Evaluation setup

- 2 models architectures: GPT2 (Causal LM) & RoBERTa (Masked LM)
- 3 language datasets: English, French, German
- 3 syntactic benchmarks: BLiMP, Zorro, CLAMS

MIXED RESULTS! → CHILDES ≈ Wiki; CHILDES ≳ Wiki; CHILDES ≪ Wiki

	Model	Training Data	BLiMP	Zorro	CLAMS English	CLAMS French	CLAMS German
GPT-2	CLM	CHILDES	0.61 ± 0.02	**0.76 ± 0.04**	0.60 ± 0.01	0.64 ± 0.01	0.69 ± 0.03
		Wiki	0.61 ± 0.02	0.69 ± 0.04	**0.71 ± 0.01**	**0.80 ± 0.01**	**0.81 ± 0.01**
RoBERTa	MLM	CHILDES	0.59 ± 0.03	0.66 ± 0.05	0.57 ± 0.02	0.59 ± 0.02	0.70 ± 0.01
		Wiki	0.59 ± 0.03	0.67 ± 0.03	**0.63 ± 0.01**	**0.69 ± 0.01**	**0.75 ± 0.01**

Better evaluation: Frequency Informed Testing (FIT)

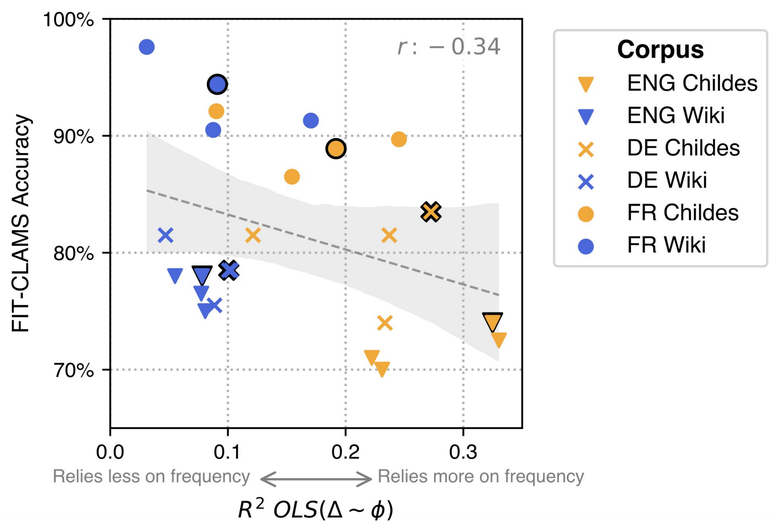
Problem: Comparing syntactic accuracy of models trained on different corpora is tricky because of different vocabularies!
Our solution: Generate minimal pairs balanced by corpus-specific lexical frequency, ensuring a fair spread of high- and low-frequency items *for each corpus*.

EN Nouns	Bin	Freq	Df
roommate, roommates	0	2	CHI
resident, residents	1	6	CHI
librarian, librarians	2	13	CHI
officer, officers	3	36	CHI
toddler, toddlers	4	90	CHI
farmer, farmers	5	264	CHI
policeman, policemen	6	380	CHI
doctor, doctors	7	656	CHI
man, men	8	2156	CHI
daddy, daddies	9	7027	CHI

EN Verbs	Bin	Freq	Long VP	Df
awaits, await	0	2	*the guests*	CHI
complains, complain	1	8	*about the noise*	CHI
arrives, arrive	2	17	*at the station*	CHI
disappears, disappear	2	42	*from the scene*	CHI
bows, bow	4	243	*to the king*	CHI
hides, hide	4	391	*from the chicken*	CHI
leaves, leave	6	1793	*the room*	CHI
sits, sit	7	4219	*in the car*	CHI
thinks, think	8	14710	*about the trip*	CHI
goes, go	9	27620	*to the new store*	CHI

Regression Analysis

How does lexical frequency in minimal pairs affect model accuracy? ⇨ Models that rely more on lexical frequency (higher fit) tend to perform worse on FIT-CLAMS.



Results of CLM models on FIT-CLAMS

As expected, LMs perform better on minimal pairs created with in-distribution lexical items.
However, **LMs trained on Wiki still outperform those trained on CHILDES** in most cases including Simple Agreement, even when controlling for frequency effects

Training	Eval. lex.	EN	FR	DE
CHILDES	CHILDES	0.63 ± 0.02	0.78 ± 0.04	0.73 ± 0.03
	Wiki	0.63 ± 0.03	0.67 ± 0.03	0.69 ± 0.04
Wiki	CHILDES	**0.72 ± 0.03**	**0.86 ± 0.02**	**0.83 ± 0.02**
	Wiki	**0.75 ± 0.02**	**0.88 ± 0.06**	**0.82 ± 0.03**

Conclusions

- CDL shows no clear benefit for syntax learning
- This holds for current modeling approaches: LMs trained in **static, non-interactive** environments, without feedback, developmental grounding, or cognitive constraints, unlike human learners
- CDL may still hold promise in interactive, situated learning environments, shifting focus toward the **communicative and contextual factors essential to language acquisition**