What is the real benefit of using Child Directed Language for Language Modeling?

Francesca Padovani[1] & Jaap Jumelet[1] & Yevgen Matusevych[1] & Arianna Bisazza[1] | [1]GroNLP, University of Groningen | f.padovani@rug.nl

RESEARCH MOTIVATION

Child-Directed Language (CDL) has proven to be beneficial to enhance the grammatical learning of language models (LMs) when used as training data.

However, this claim is primarily supported by a **generalistic accuracy score** across various syntactic paradigms, making it difficult to pinpoint the specific aspects of CDL that contribute to this improvement.

RESEARCH QUESTIONS

1. Do these results hold across **different models**, **languages**, and more **principled evaluation datasets**?
2. How does the composition of training data (CDL vs. Wikipedia) influences model behavior?

DATASET COMPOSITION (Wikipedia vs CHILDES)

		TOTAL TOKENS	TYPE/TOKEN RATIO	AVG SENTENCE LENGTH
ENGLISH	CHILDES	4.3 M	0.0746	4.86
	WIKIPEDIA	4.3 M	0.1335	20.31
FRENCH	CHILDES	1.7 M	0.080	5.11
	WIKIPEDIA	1.7 M	0.1996	24.57
GERMAN	CHILDES	2.6 M	0.1115	4.56
	WIKIPEDIA	2.6 M	0.2253	17.42

MODELS (MLM vs CLM)

- **Masked language model (RoBERTa)**: follows previous work where CDL-trained models outperformed Wikipedia-trained models
- **Causal language model (GPT-2)**: aligns better with cognitive plausibility.

CLAMS (Mueller et al., 2020): EVALUATION BENCHMARK

Multilingual and semantically plausible minimal pair benchmark focused on **subject-verb agreement.**

Simple Agreement

the teachers are short
the teachers is short

Agreement in VP Coordinates

the manager laughs and is young
the manager laughs and are young

Agreement in Prepositional Phrases

the teacher to the side of the guard laughs
the teacher to the side of the guard laugh

Agreement in Subject Relative Clauses

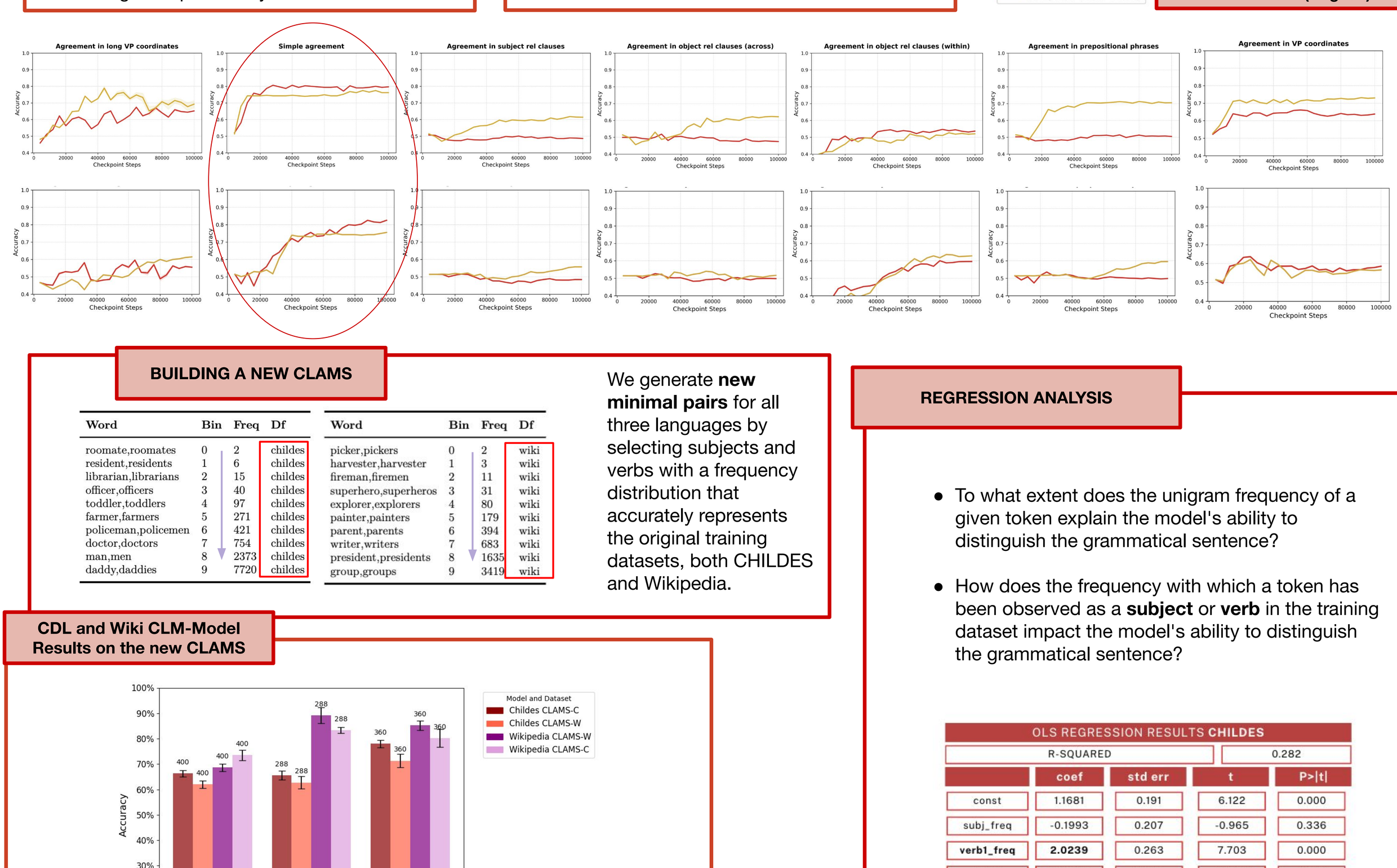
the teachers that love the parents are young
the teachers that love the parents is young

ACCURACY RESULTS on CLAMS

	ENGLISH	FRENCH	GERMAN
CLM	CHILDES 0.56 ± 0.01	CHILDES 0.68 ± 0.02	CHILDES 0.66 ± 0.01
	WIKIPEDIA 0.65 ± 0.03	**WIKIPEDIA 0.78 ± 0.02**	**WIKIPEDIA 0.79 ± 0.01**
MLM	CHILDES 0.57 ± 0.01	CHILDES 0.66 ± 0.02	CHILDES 0.67 ± 0.02
	WIKIPEDIA 0.60 ± 0.00	**WIKIPEDIA 0.76 ± 0.02**	**WIKIPEDIA 0.76 ± 0.02**

No model trained on **CDL outperforms** a model trained on **Wikipedia** in terms of overall performance, across any language or model type.

LEARNING CURVES FOR EACH PARADIGM (English)



BUILDING A NEW CLAMS

Word	Bin	Freq	Df
roomate,roomates	0	2	childes
resident,residents	1	6	childes
librarian,librarians	2	15	childes
officer,officers	3	40	childes
toddler,toddlers	4	97	childes
farmer,farmers	5	271	childes
policeman,policemen	6	421	childes
doctor,doctors	7	754	childes
man,men	8	2373	childes
daddy,daddies	9	7720	childes

Word	Bin	Freq	Df
picker,pickers	0	2	wiki
harvester,harvester	1	3	wiki
fireman,firemen	2	11	wiki
superhero,superheros	3	31	wiki
explorer,explorers	4	80	wiki
painter,painters	5	179	wiki
parent,parents	6	394	wiki
writer,writers	7	683	wiki
president,presidents	8	1635	wiki
group,groups	9	3419	wiki

We generate **new minimal pairs** for all three languages by selecting subjects and verbs with a frequency distribution that accurately represents the original training datasets, both CHILDES and Wikipedia.

CDL and Wiki CLM-Model Results on the new CLAMS

Models trained on Childes perform better on the new set of simple agreement minimal pairs derived from subjects and verbs sourced from the **CHILDES** training dataset distribution.
Similarly, **models trained on Wikipedia** perform better on the new set of minimal pairs derived from the **Wikipedia distribution** (except for English).

REGRESSION ANALYSIS

- To what extent does the unigram frequency of a given token explain the model's ability to distinguish the grammatical sentence?
- How does the frequency with which a token has been observed as a **subject** or **verb** in the training dataset impact the model's ability to distinguish the grammatical sentence?

OLS REGRESSION RESULTS CHILDES

R-SQUARED: 0.282

	coef	std err	t	P>|t|
const	1.1681	0.191	6.122	0.000
subj_freq	-0.1993	0.207	-0.965	0.336
verb1_freq	**2.0239**	0.263	7.703	0.000
verb2_freq	**-1.0099**	0.262	-3.855	0.000

OLS REGRESSION RESULTS WIKIPEDIA

R-SQUARED: 0.688

	coef	std err	t	P>|t|
const	1.7635	0.129	13.700	0.000
subj_freq	0.2269	0.129	1.758	0.080
verb1_freq	**4.2844**	0.214	20.023	0.000
verb2_freq	**-4.1178**	0.214	-19.253	0.000