Measuring Language Development from Child-Centered Recordings

Yaya Sy[1], William N. Havard[1,2], Marvin Lavechin[1,2,3], Emmanuel Dupoux[1,2,3], Alejandrina Cristia[1]
[1]Laboratoire de Sciences Cognitives et de Psycholinguistique,
Département d'Études Cognitives, ENS, EHESS, CNRS, PSL University
[2] Cognitive Machine Learning Team, INRIA
[3] Meta AI Research, France

Introduction & Goals

- Today, **measuring child language development (CLD)** from spontaneous corpora **requires costly human labor** (describing languages, transcribing speech).
- Our proposal includes:
  (1) **a new CLD metric** based on **entropy** from a corpus of text in the relevant language
  (2) how to **derive this metric from speech** from a smaller text+speech parallel corpus

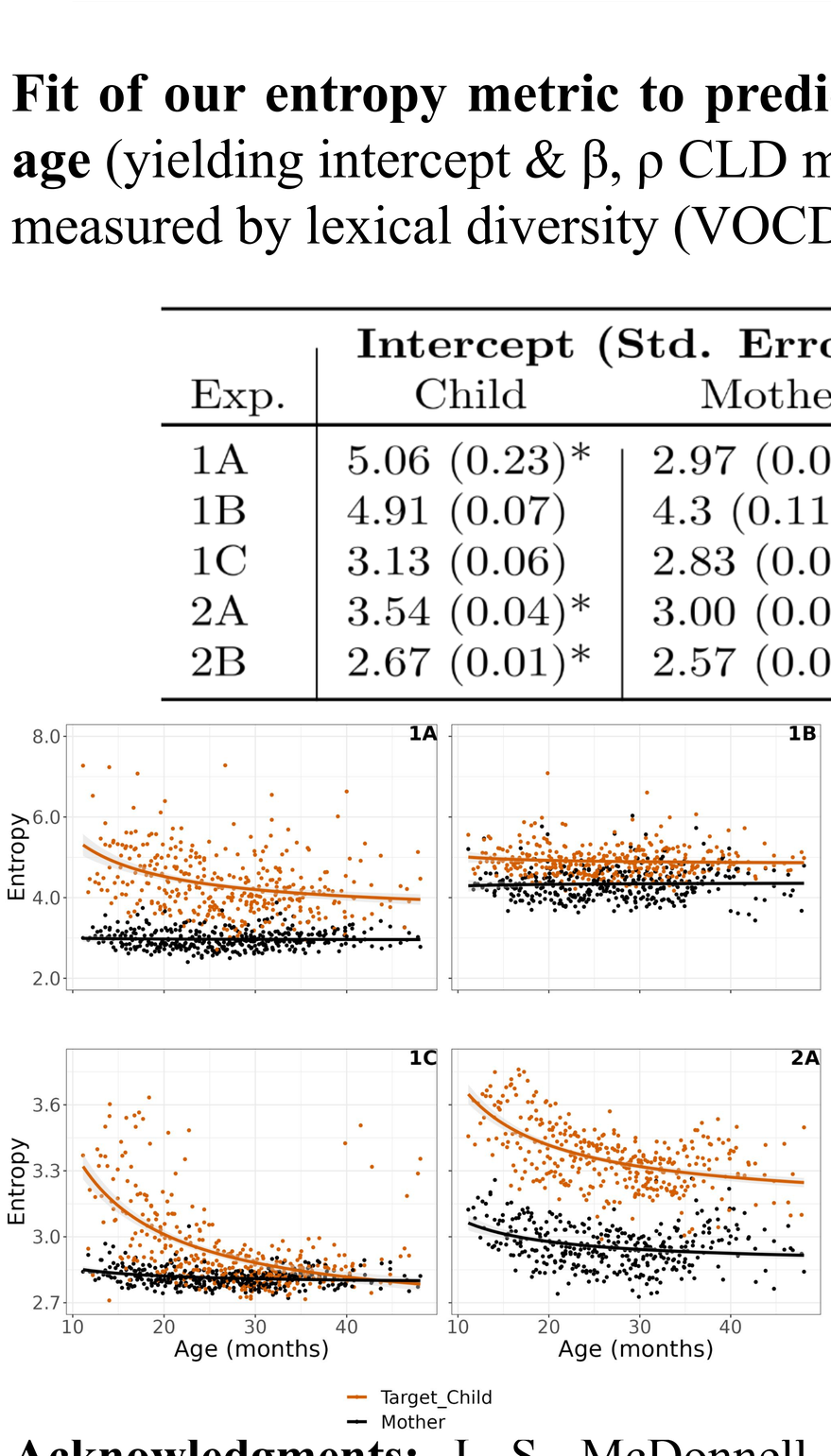
Methods & Results

Experiments: Model, data, and units for the relevant analyses. Training: LIBRI. stands for LibriSpeech; THOM. stands for Thomas. Test always drew from Providence, which contains speech & transcriptions for 6 English learning children aged 11-48 months

Exp.	Model	Model input units	Train	Test
1A	5-gram language model	phones	LIBRI. text	text
1B		HUBERT-BASE discrete clusters	Libri. audio	speech
1C				synthetic speech
2A	linear regression	speech (+ text entropies at training time)	THOM.	speech
2B			LIBRI.	

Fit of our entropy metric to predictions. Mixed-model regressions **predicting entropy from child age** (yielding intercept & β, ρ CLD metric shows how entropy correlates with children's development as measured by lexical diversity (VOCD), and morphosyntactic complexity (MLU, IPSyn)

	Intercept (Std. Error)		**β age (Std. Error)**		**ρ CLD metric**		
Exp.	Child	Mother	Child ↓	Mother ↔	VOCD	MLU	IPSyn
1A	5.06 (0.23)*	2.97 (0.07)*	−0.31 (0.06)*	0.01 (0.02)	−0.23	−0.56	−0.45
1B	4.91 (0.07)	4.3 (0.11)	0.01 (0.02)	0.01 (0.04)	0.03	−0.02	−0.01
1C	3.13 (0.06)	2.83 (0.01)	−0.1 (0.02)	−0.01 (0.00)	−0.21	−0.56	**−0.54**
2A	3.54 (0.04)*	3.00 (0.03)*	−0.1 (0.01)*	−0.02 (0.01)	**−0.27**	**−0.73**	−0.53
2B	2.67 (0.01)*	2.57 (0.02)*	0.00 (0.00)	0.01 (0.01)	0.14	−0.08	−0.16



Conclusion & Future Directions

- **Entropy** from speech **requires** within-domain data.
- Use **more ecological recordings** (long-form, from childworn devices) with **more children**
- Use a more **diverse set of languages**

LAAC-LSCP/EntropyBasedCLDMetrics

Acknowledgments: J. S. McDonnell Foundation; European Research Council (ERC, H2020) ExELang, 101001095), HPC resources from GENCI–IDRIS (Grant 2021-AD011013145). **Contact:** alecristia@gmail.com