Reverse-engineering language acquisition

2021-07-08 @ PAISS

Alejandrina (Alex) Cristia

Laboratoire de Sciences Cognitives et Psycholinguistique

Language Acquisition Across Cultures Team

Thanks to my team for help with the slides!

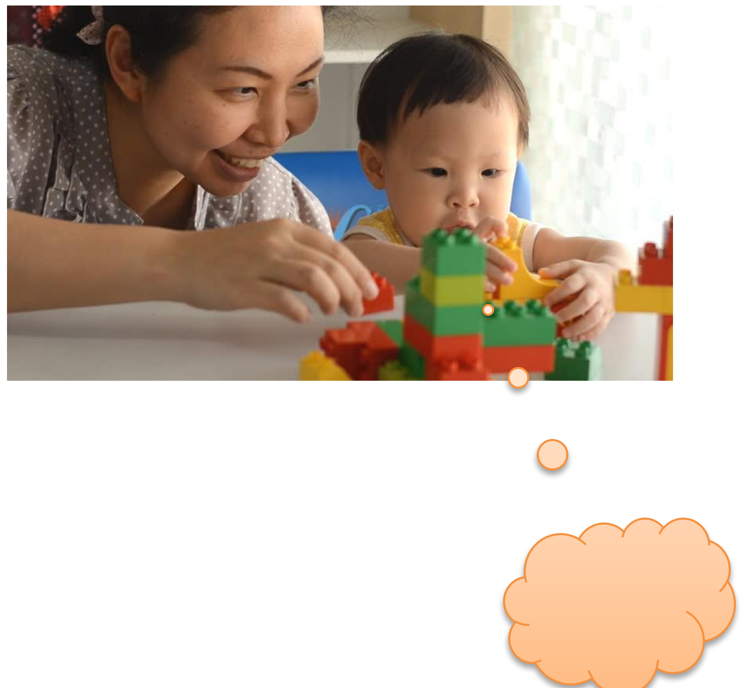
Erh, what IS language acquisition?

Which of the following are true?

Please vote TRUE=👍 ; FALSE =😯

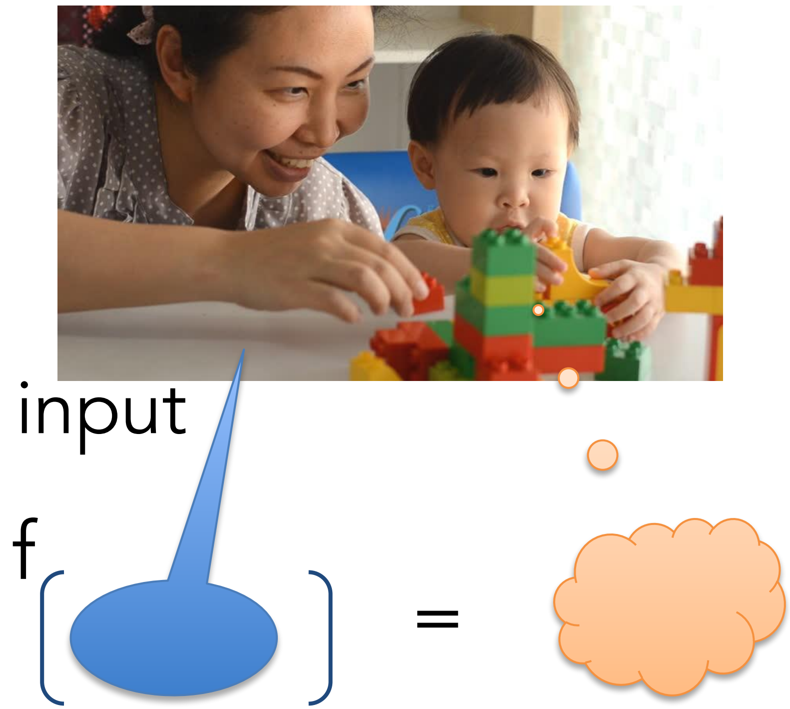
- Newborns prefer listening to their native language than to an unfamiliar language
- Newborns know their name
- By 6 months, babies know their name
- By 6 months, babies say their first word
- By 12 months, babies say their first word

A broad language acquisition theory (v 1.0)



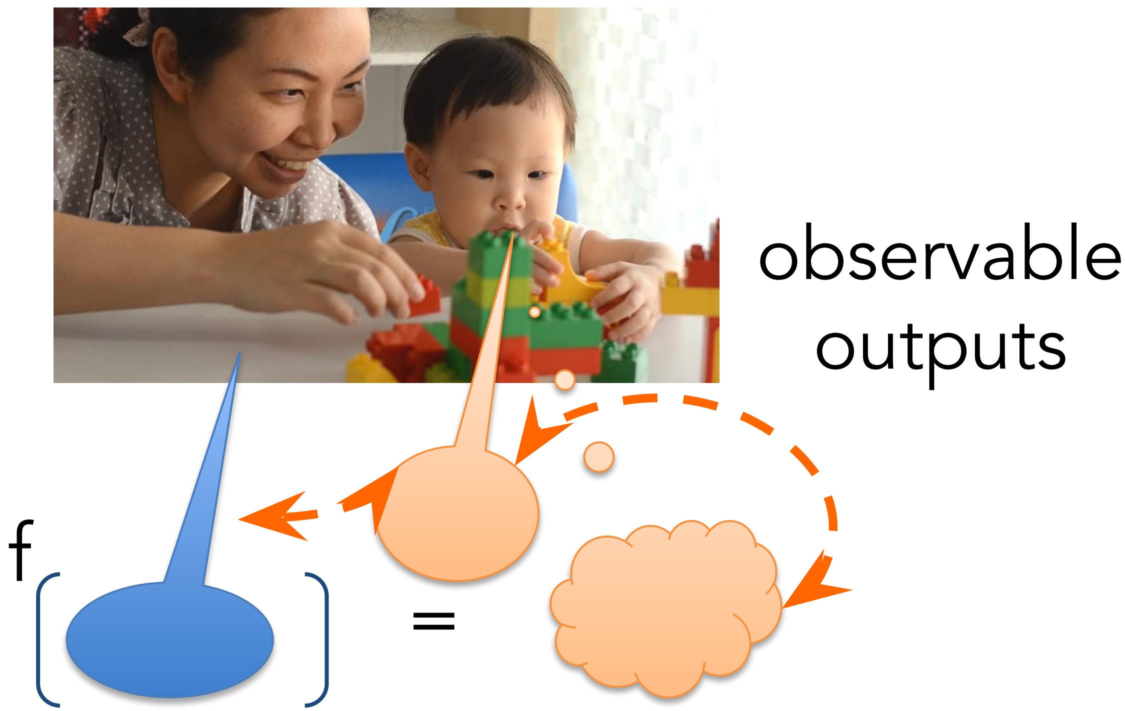
Mental representations appropriate to native language(s)

A broad language acquisition theory (v 1.0)

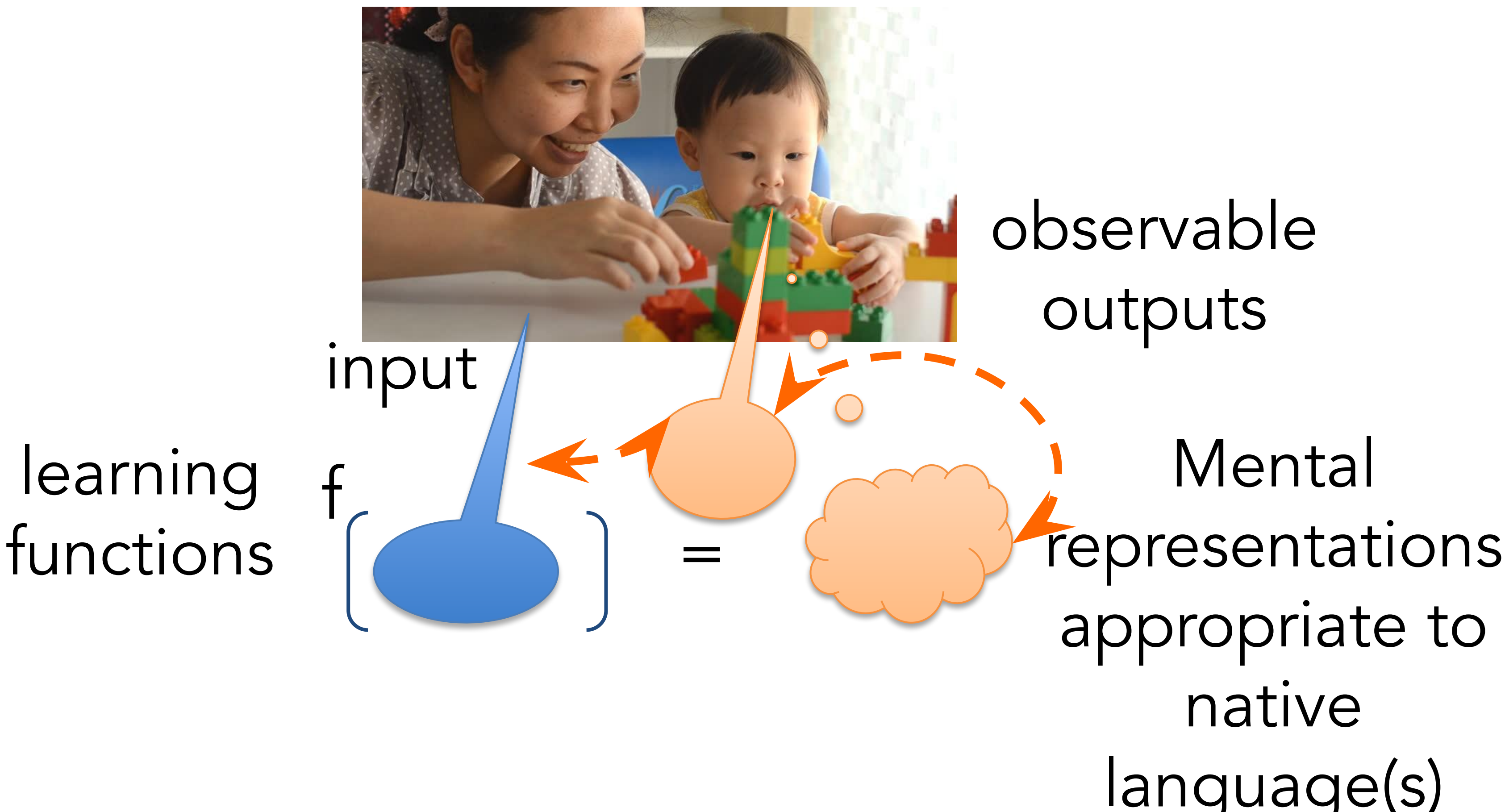


learning functions

A broad language acquisition theory (v 1.0)



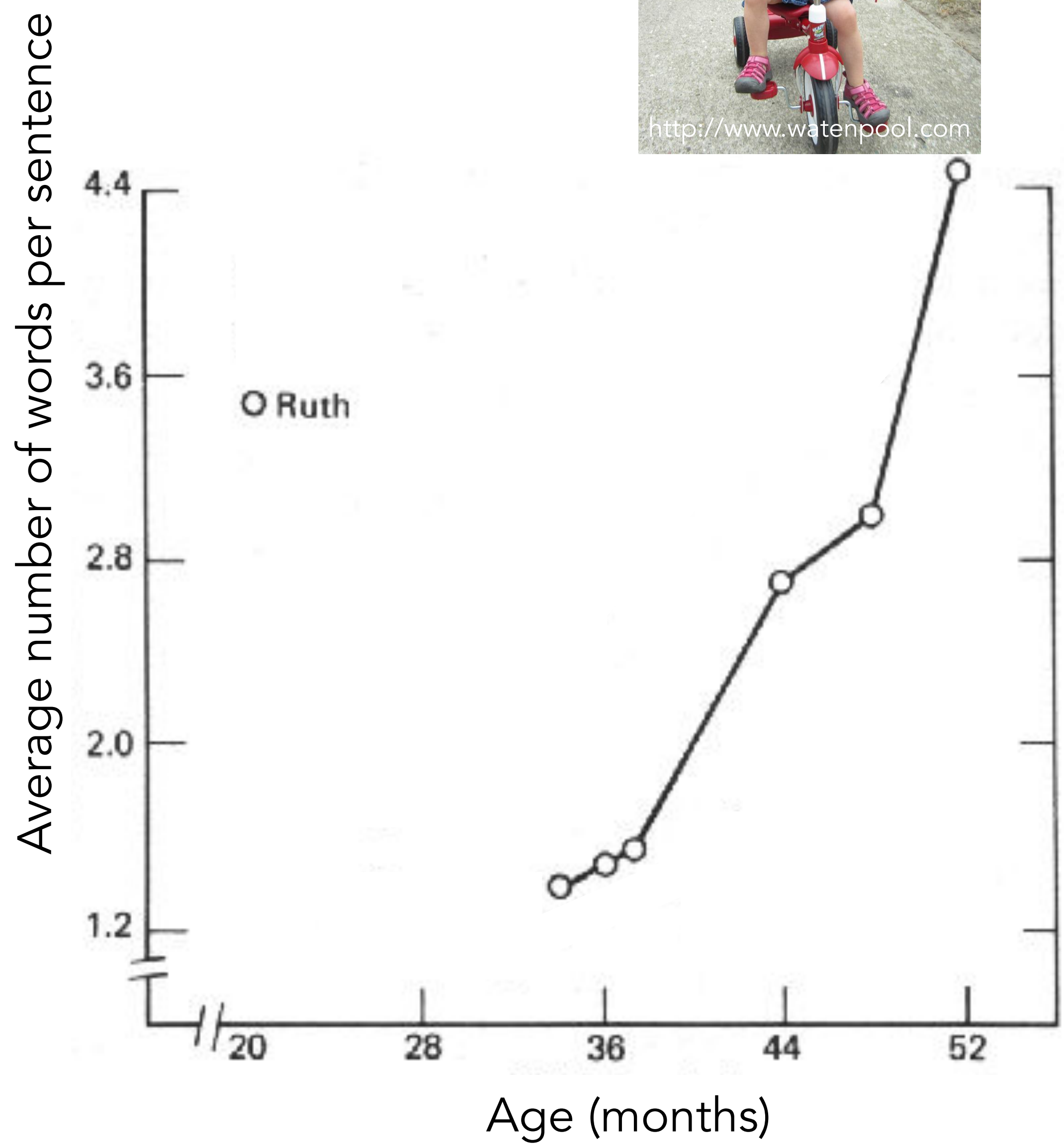
A broad language acquisition theory (v 1.0)



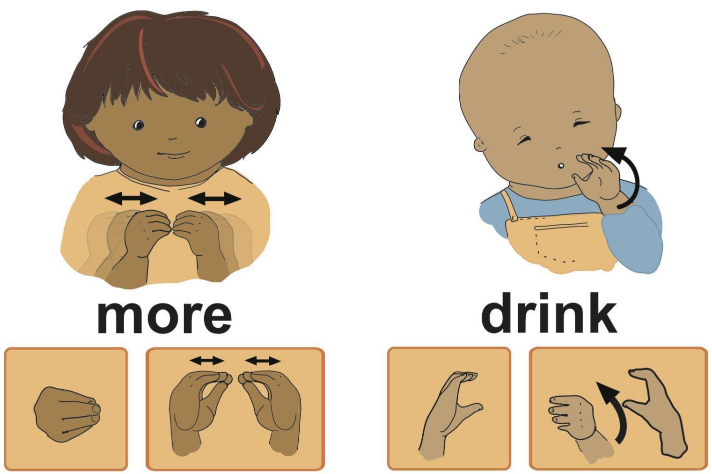
Which of the following are true?

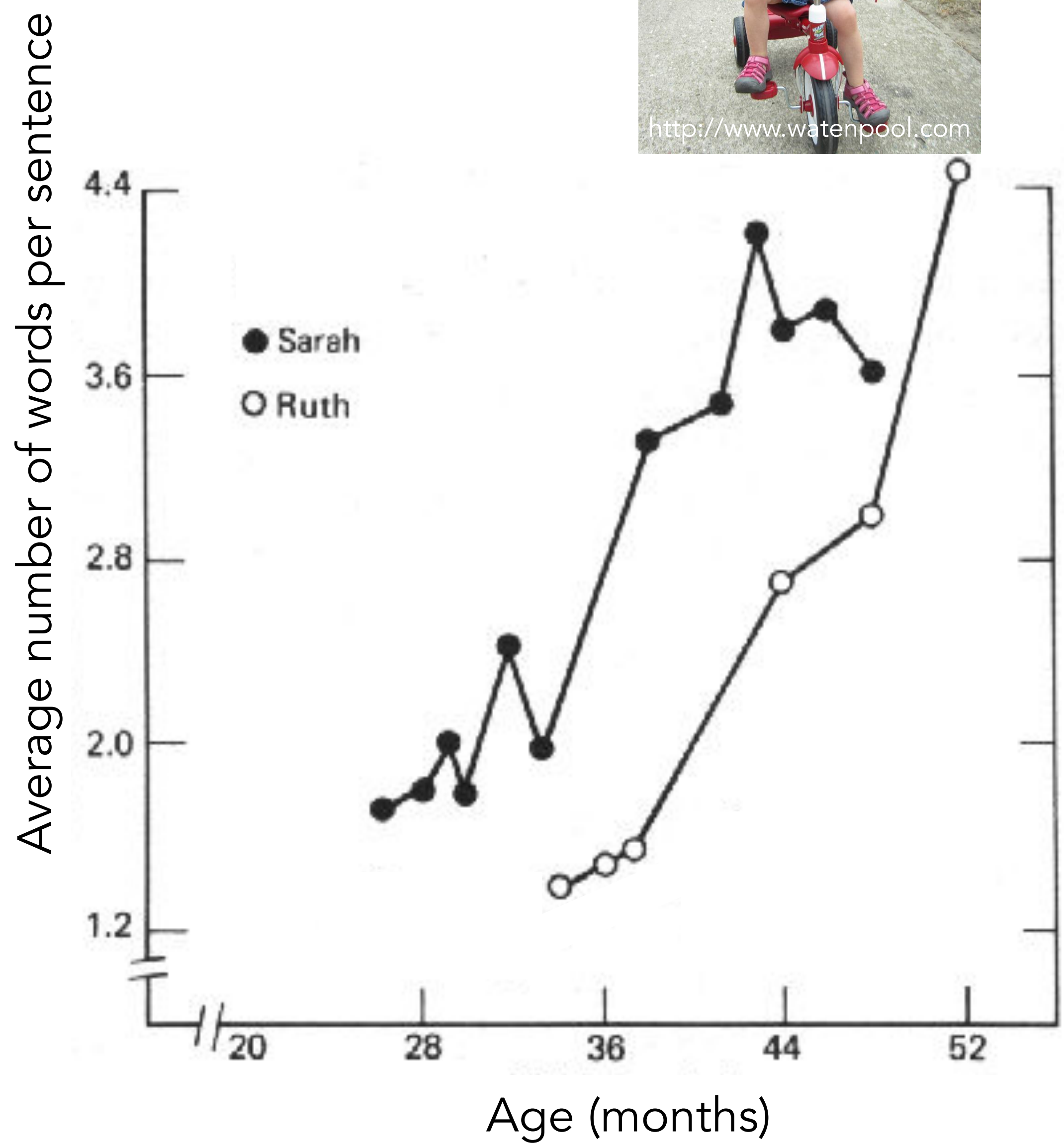
Please vote TRUE= 👍 ; FALSE = 😯

- Humans and chimpanzees share a majority of their genetic information
- In terms of their visual skills, humans and chimpanzees are more similar to each other than humans and killer whales are
- In terms of their communication system, humans and chimpanzees are more similar to each other than humans and killer whales are
- You can raise a chimpanzee to use language like human babies do

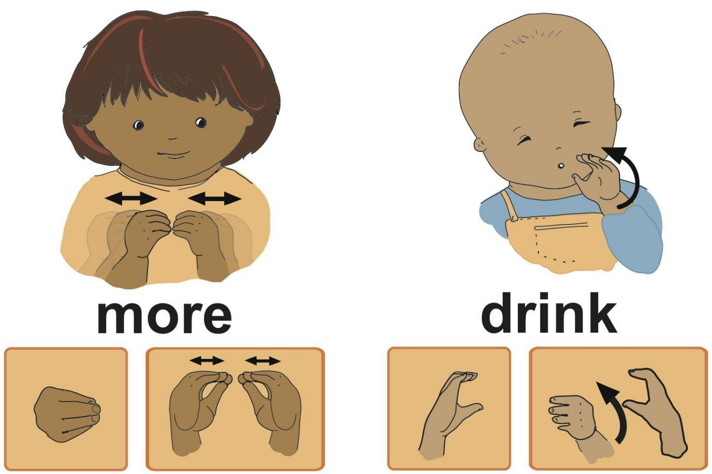


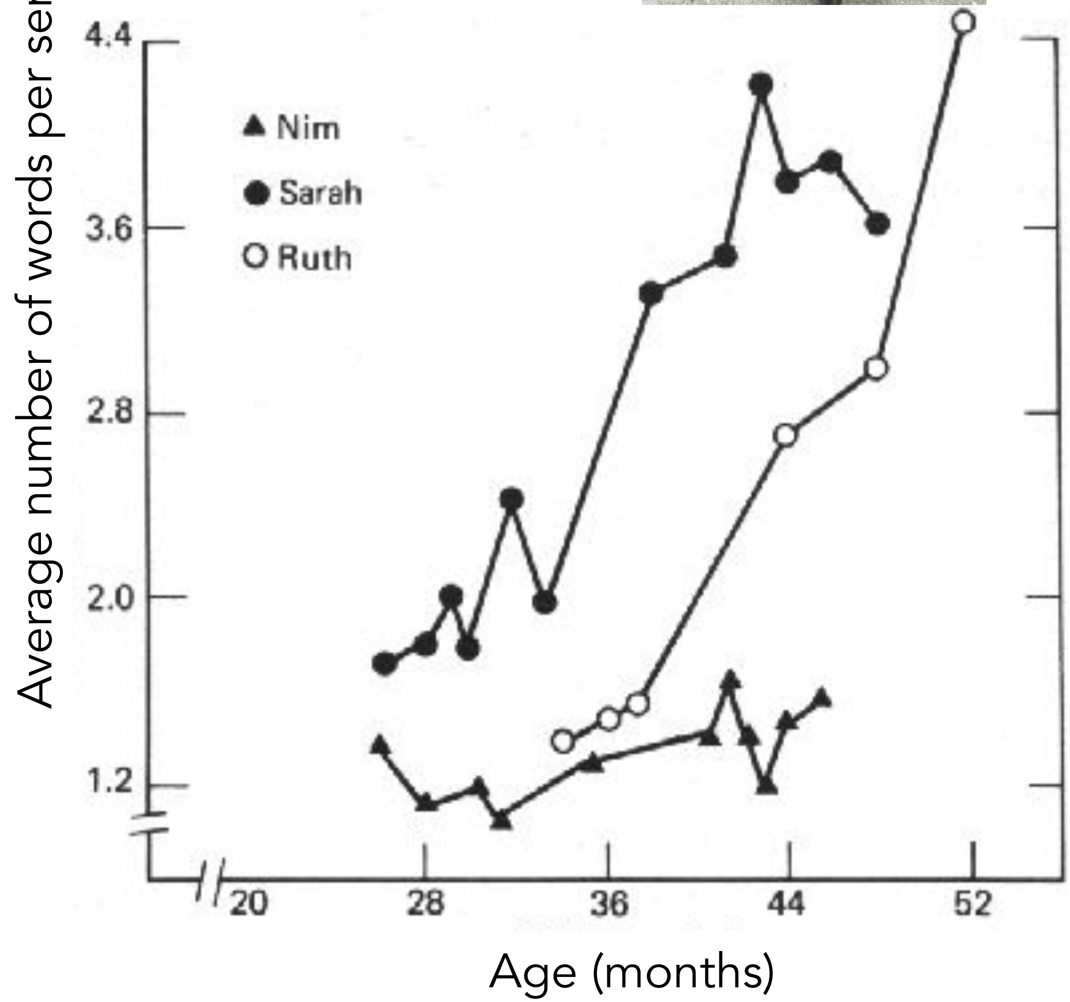
Terrace 1979 Science





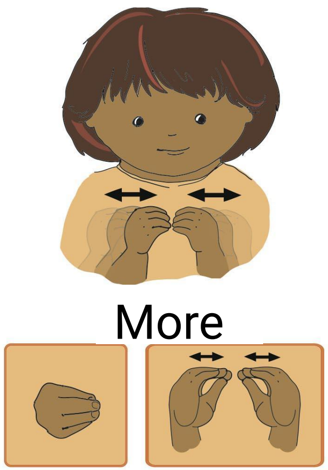
Terrace 1979 Science



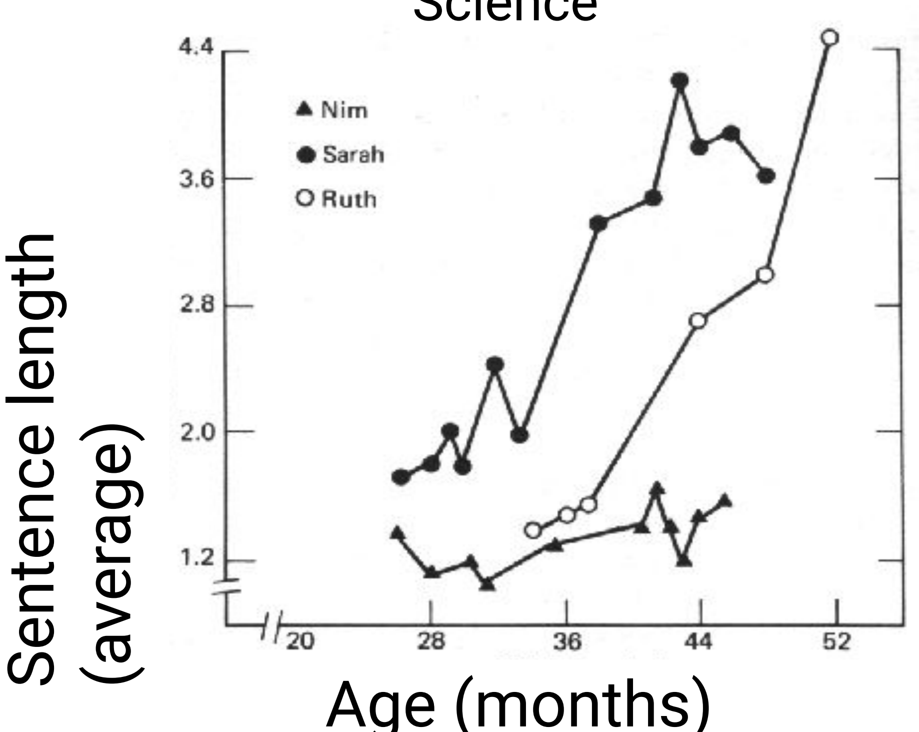


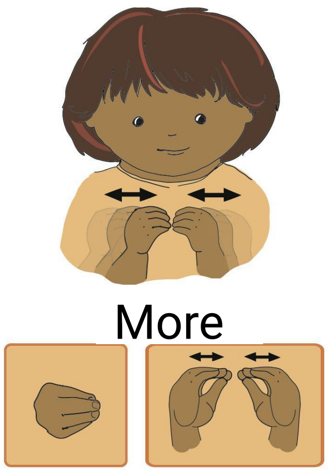
Terrace 1979 Science

Innate



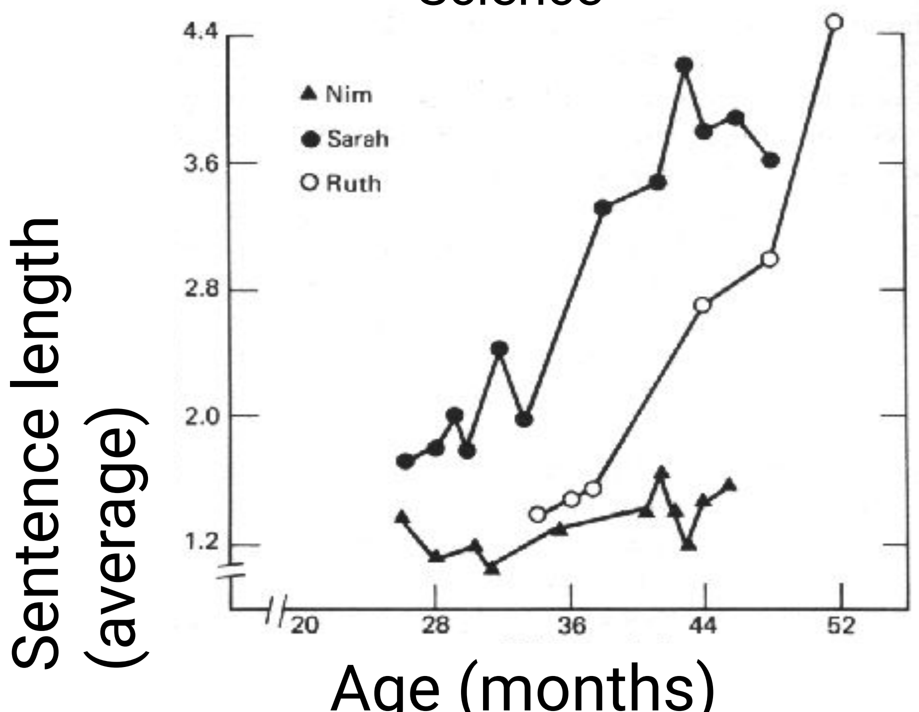
Terrace 1979
Science

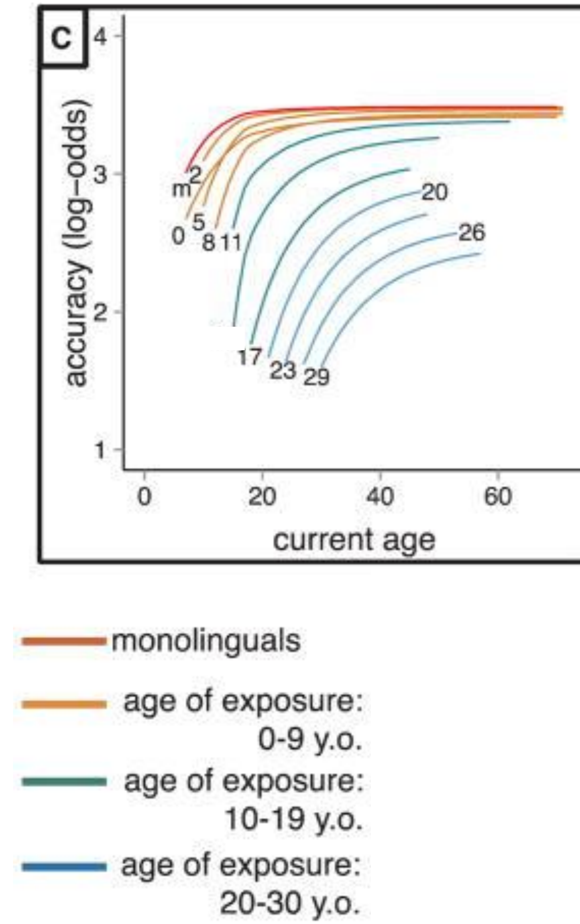




Terrace 1979
Science

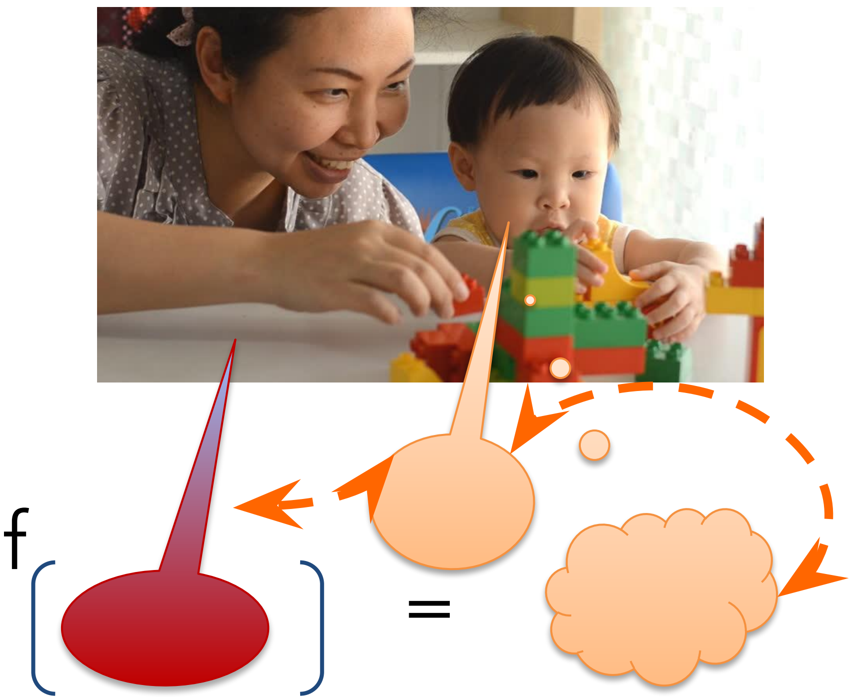
Innate & acquired





Hartshorn et al. 2018
Cognition

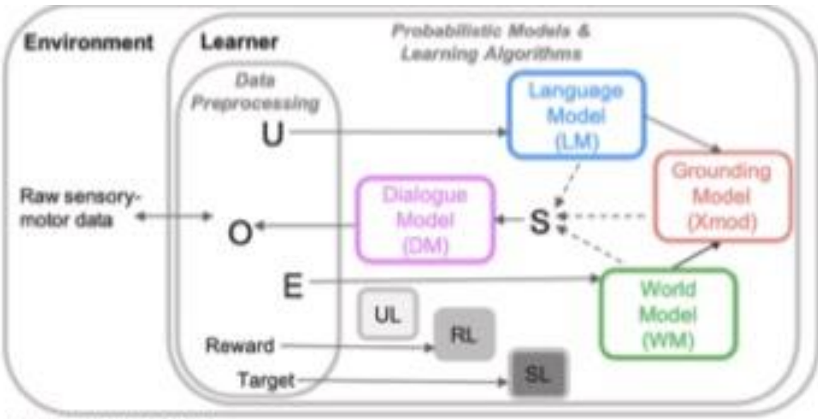
A more specific language acquisition theory (v 2.0):
Adult input "fuels" language acquisition



Adults' speech is high quality
- a stable linguistic system
- developed “theory of mind”

One on one
- topics adapted to child’s attention & abilities
- use of “Parentese”

Socio-Computational Architecture of Language Acquisition



Probabilistic Models

- **Language Model**. Estimates P(U), the probability distribution of message U.
- **World Model**. Estimates P(E), the probability of event E.
- **Grounding Model**. Estimates probabilities of association between verbal form and event (P(U,E)). Assumes that the intended meaning is accessible here-and-now.
- **Dialogue Model**. Computes the probability of communicative output O given message and current state of world S (P(O|S)). S is computed from a representation of past events and utterances.

Learning Algorithms

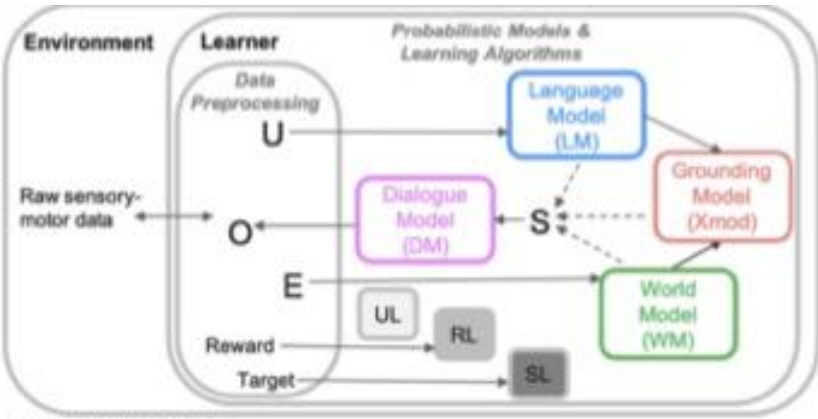
- **Unsupervised Learning (UL)**. Tries to optimize the likelihood of observing a given input (U or E). Language Models (LM) and World Models (WM) can be learned in this fashion.
- **Reinforcement Learning (RL)**. Tries to optimize the expected reward (Reward). Dialogue Models (DM) can be learned this way.
- **Supervised Learning (SL)**. Tries to minimize the discrepancy between an expected response (Target) provided by the environment and actual response O. DMs can be learned in this way.

Data Preprocessing

- **Filtering:** what sensory data counts as a language input (U), a world input (E), a Reward, a Target ?
- **Segmenting:** what are the units of the language stream (U), what is an event (E) ?
- **Routing:** is there an intended/corrective target (Target), and if so, what output O is it supposed to correct? If there is a referential act, which parts of U map to which part of E for cross modal learning?

Tsuji et al. 2021 Cognition

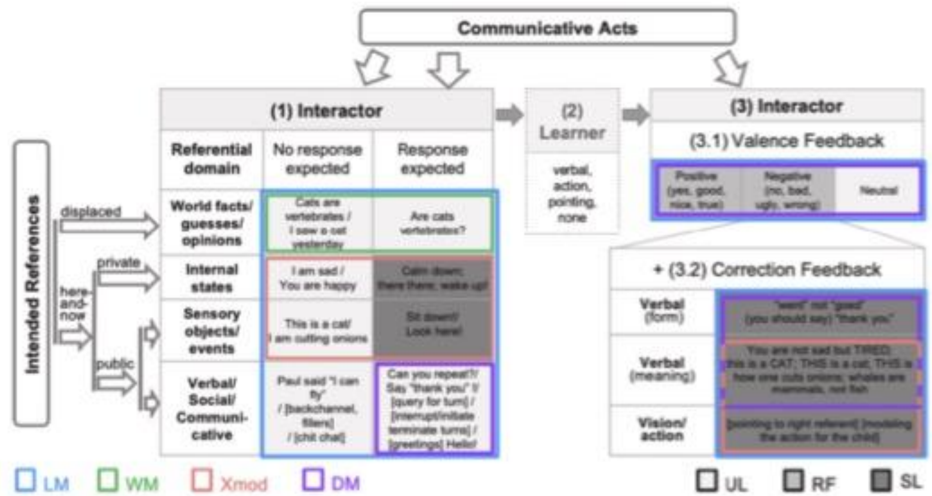
Socio-Computational Architecture of Language Acquisition



Probabilistic Models
- **Language Model**. Estimates P(U), the probability distribution of message U.
- **World Model**. Estimates P(E), the probability of event E.
- **Grounding Model**. Estimates probabilities of association between verbal form and event (P(U,E)). Assumes that the intended meaning is accessible here-and-now.
- **Dialogue Model**. Computes the probability of communicative output O given message and current state of world S (P(O|S)). S is computed from a representation of past events and utterances.

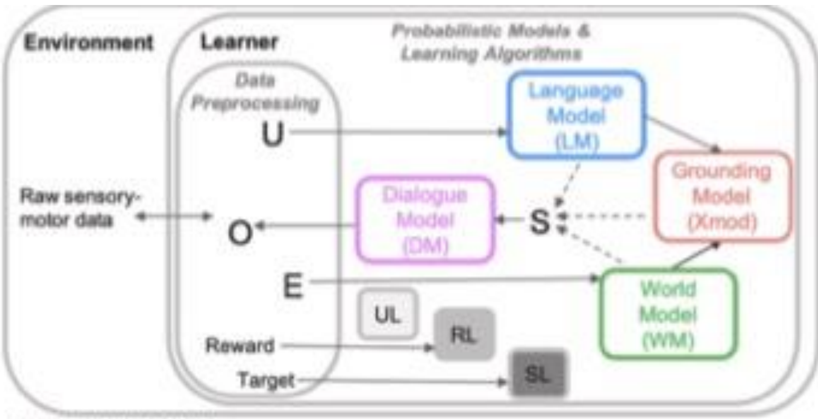
Learning Algorithms
- **Unsupervised Learning (UL)**. Tries to optimize the likelihood of observing a given input (U or E). Language Models (LM) and World Models (WM) can be learned in this fashion.
- **Reinforcement Learning (RL)**. Tries to optimize the expected reward (Reward). Dialogue Models (DM) can be learned this way.
- **Supervised Learning (SL)**. Tries to minimize the discrepancy between an expected response (Target) provided by the environment and actual response O. DMs can be learned in this way.

Data Preprocessing
- **Filtering:** what sensory data counts as a language input (U), a world input (E), a Reward, a Target ?
- **Segmenting:** what are the units of the language stream (U), what is an event (E) ?
- **Routing:** is there an intended/corrective target (Target), and if so, what output O is it supposed to correct? If there is a referential act, which parts of U map to which part of E for cross modal learning?



Tsuji et al. 2021 Cognition

Socio-Computational Architecture of Language Acquisition



Probabilistic Models

- **Language Model**. Estimates P(U), the probability distribution of message U.
- **World Model**. Estimates P(E), the probability of event E.
- **Grounding Model**. Estimates probabilities of association between verbal form and event (P(U,E)). Assumes that the intended meaning is accessible here-and-now.
- **Dialogue Model**. Computes the probability of communicative output O given message and current state of world S (P(O|S)). S is computed from a representation of past events and utterances.

Learning Algorithms

- **Unsupervised Learning (UL)**. Tries to optimize the likelihood of observing a given input (U or E). Language Models (LM) and World Models (WM) can be learned in this fashion.
- **Reinforcement Learning (RL)**. Tries to optimize the expected reward (Reward), Dialogue Models (DM) can be learned this way.
- **Supervised Learning (SL)**. Tries to minimize the discrepancy between an expected response (Target) provided by the environment and actual response O. DMs can be learned in this way.

Data Preprocessing

- **Filtering:** what sensory data counts as a language input (U), a world input (E), a Reward, a Target ?
- **Segmenting:** what are the units of the language stream (U), what is an event (E) ?
- **Routing:** is there an intended/corrective target (Target), and if so, what output O is it supposed to correct? If there is a referential act, which parts of U map to which part of E for cross modal learning?

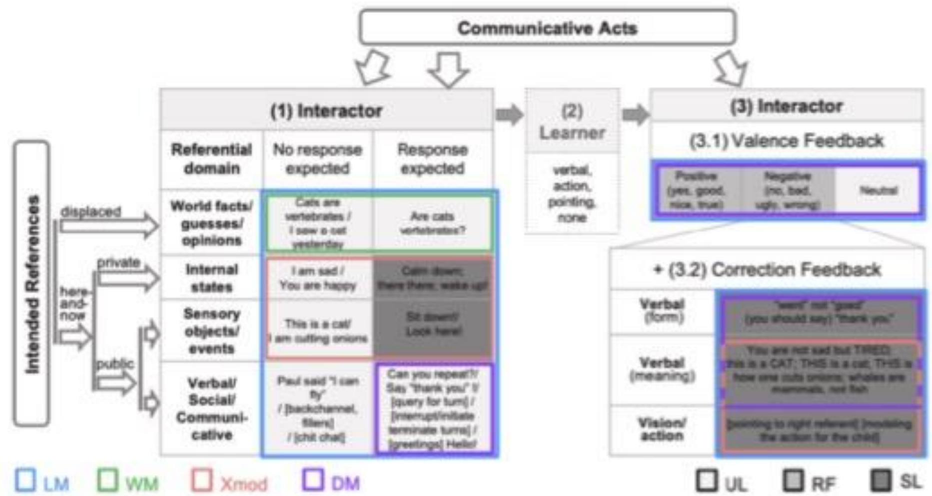
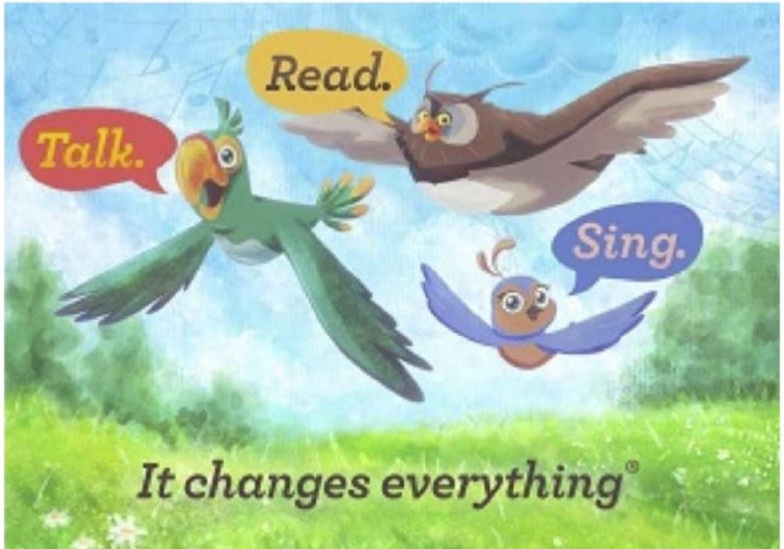


Table 1
Overview of proposed differential contributions by corpus analysts, computer modelers, and experimentalists to different research avenues.

	Algorithms	Input Data	Outcome measures	Integration
Corpus Analysis		Estimate prevalence of the various referential and event types	Measures of language output maturity	Explanations of outcome/ input relationships in infants across cultures
Computer Modeling	Implementation of probabilistic models, learning and preprocessing algorithms	Estimate of outcomes as a function of prevalence of referential/event types in the input for each combination of algorithm and preprocessing		Predictions of outcomes of interventions
Experimental Studies	Proof-of-concept of preprocessing and learning algorithms		Measure of tacit knowledge (probabilistic models of infants)	

Tsuji et al. 2021 Cognition



FIRST 5 CALIFORNIA

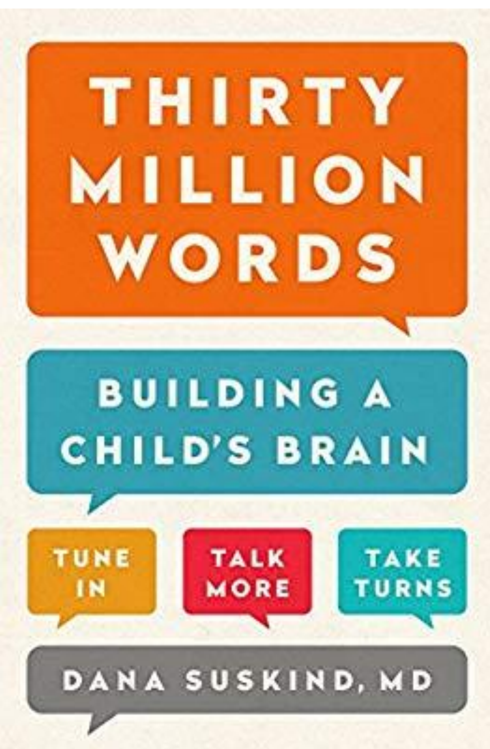
TALK WITH ME BABY

PEQUEÑOS Y VALIOSOS

UNIVISION CONTIGO

PROVIDENCE TALKS

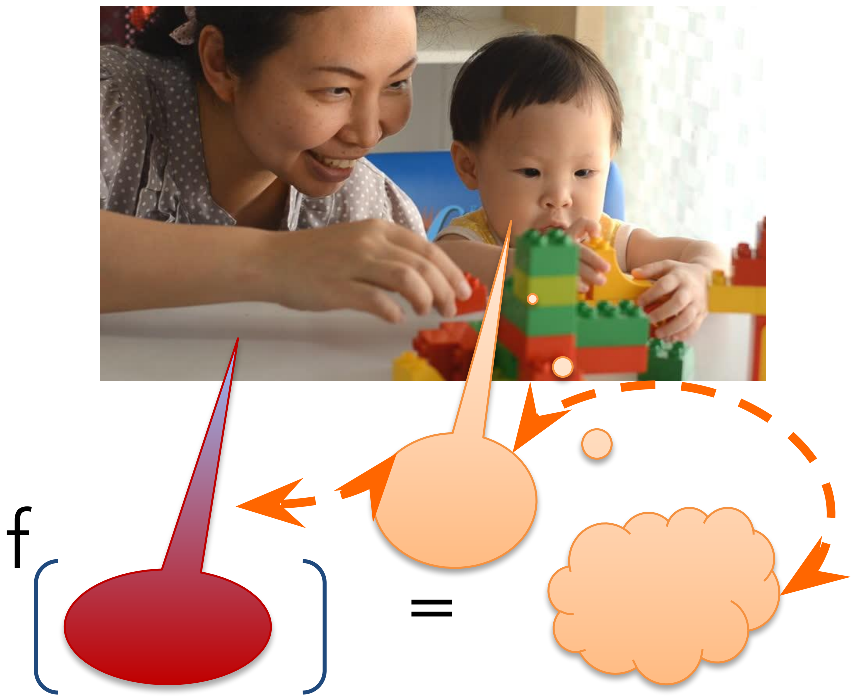
TALK TO TEACH



Thanks to Janet Bang for this selection!

The idea that

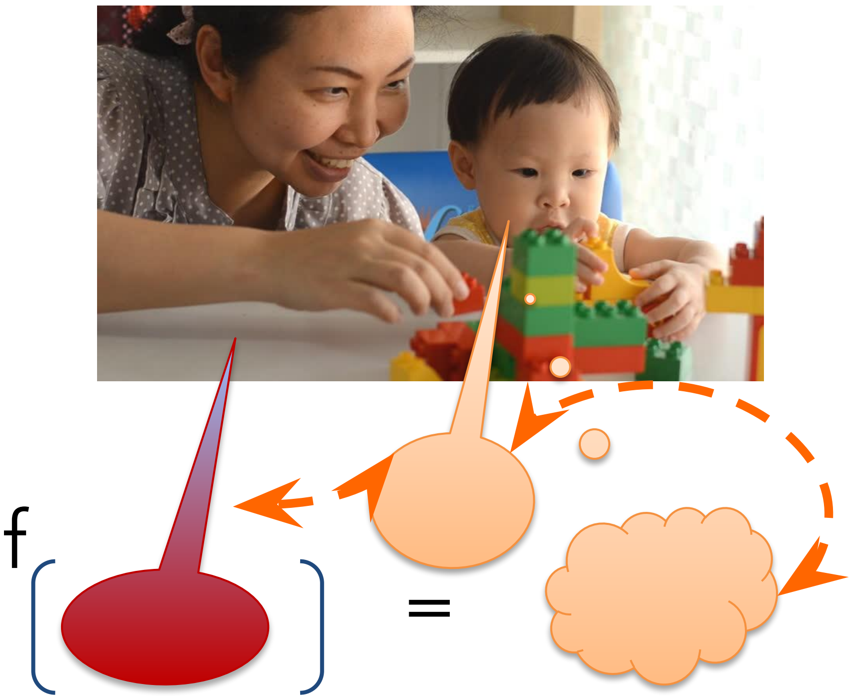
Adult input "fuels" language acquisition



is based on *evidence*

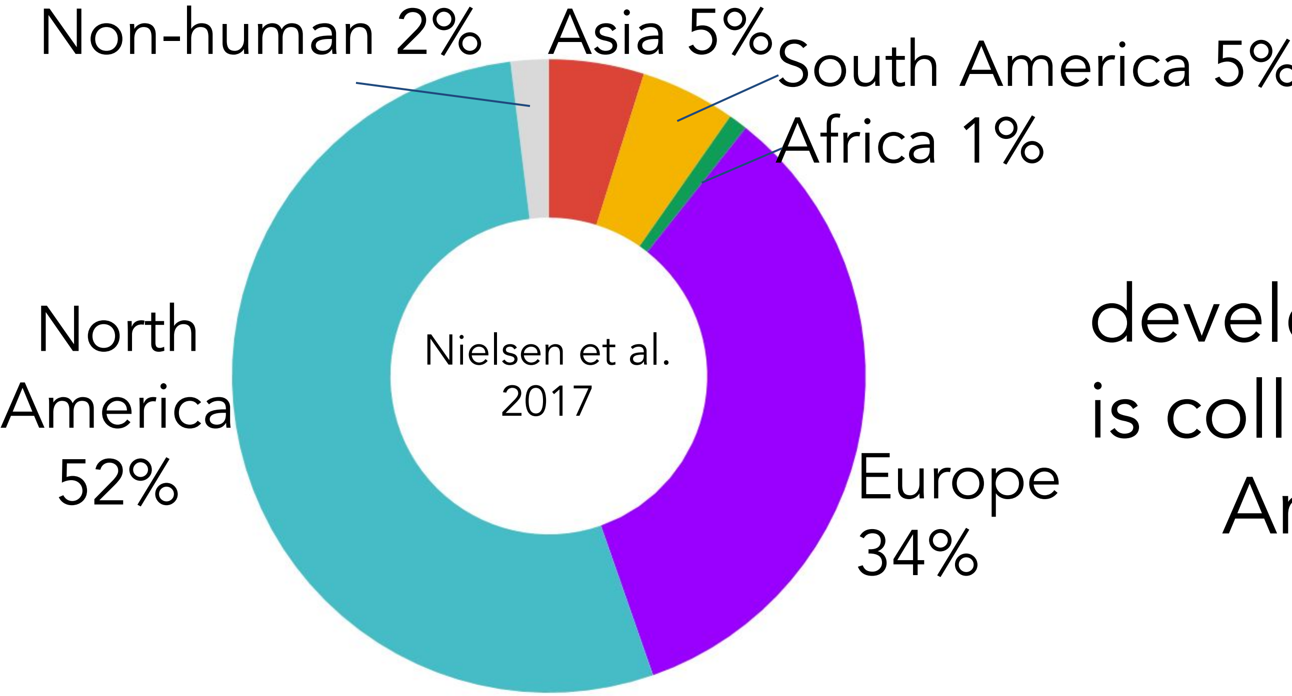
The idea that

Adult input "fuels" language acquisition

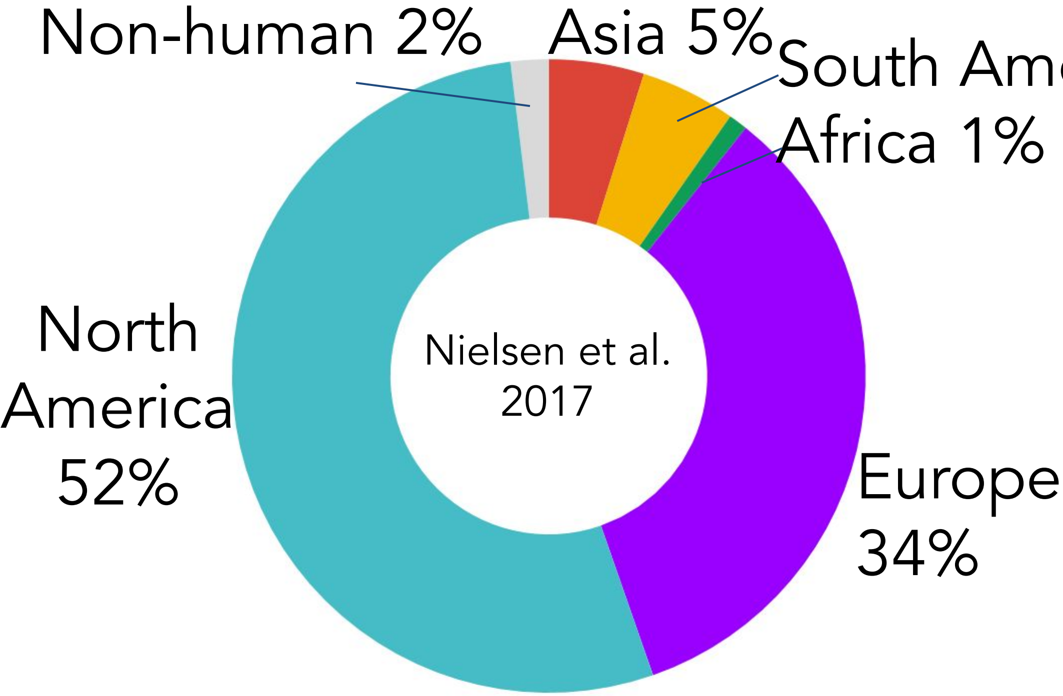


is based on *evidence*

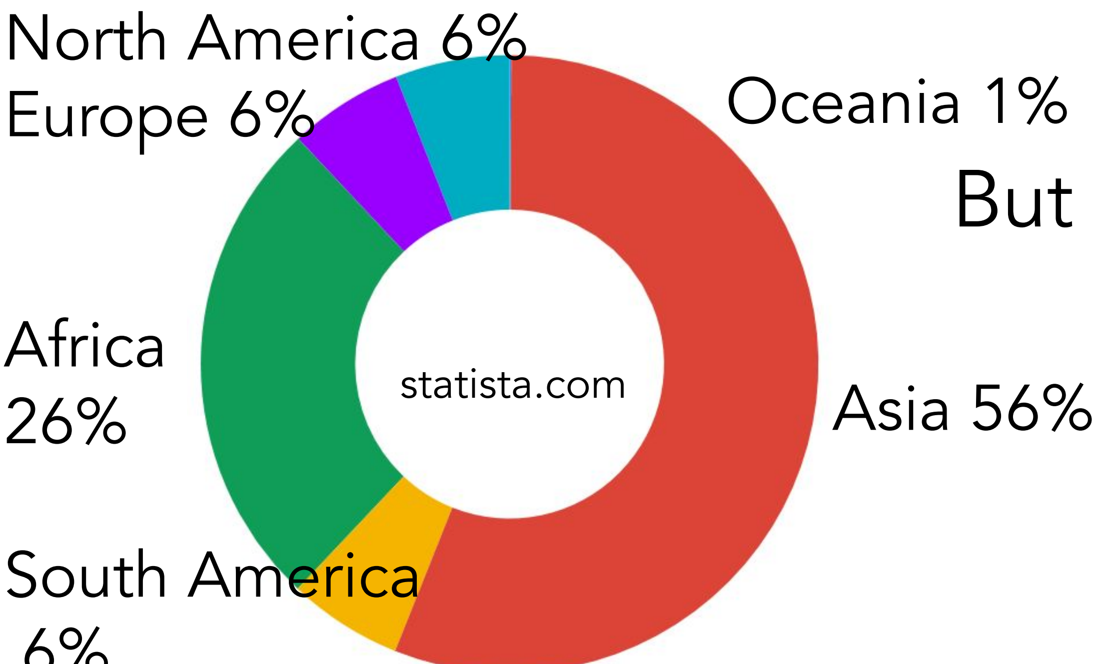
but this evidence is <u>biased</u>



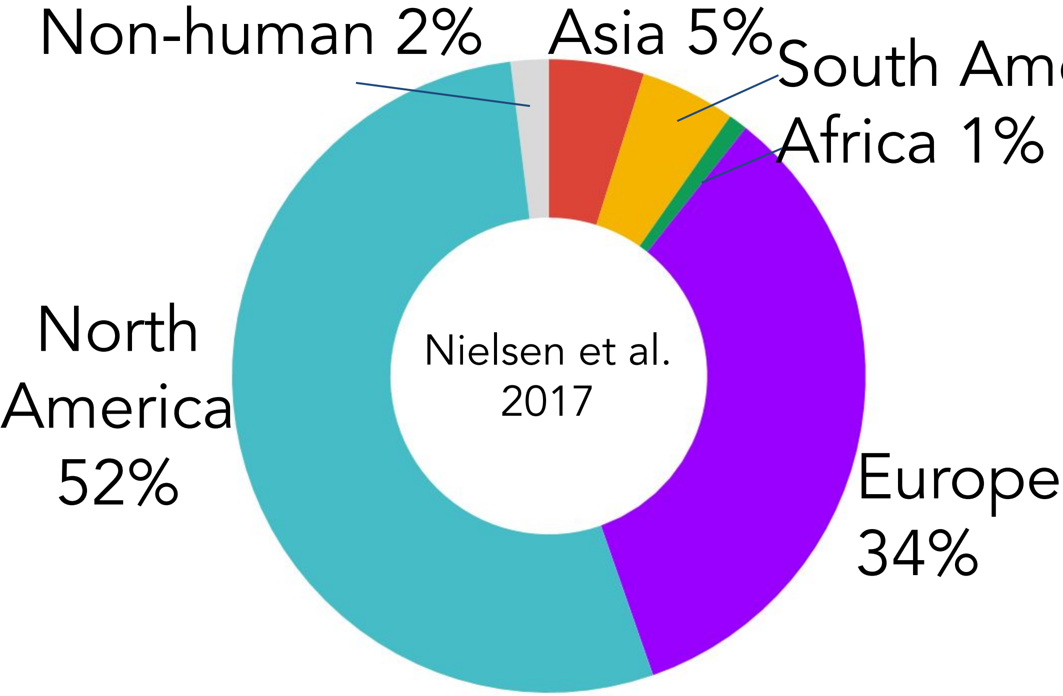
Most developmental data is collected in North America and Europe



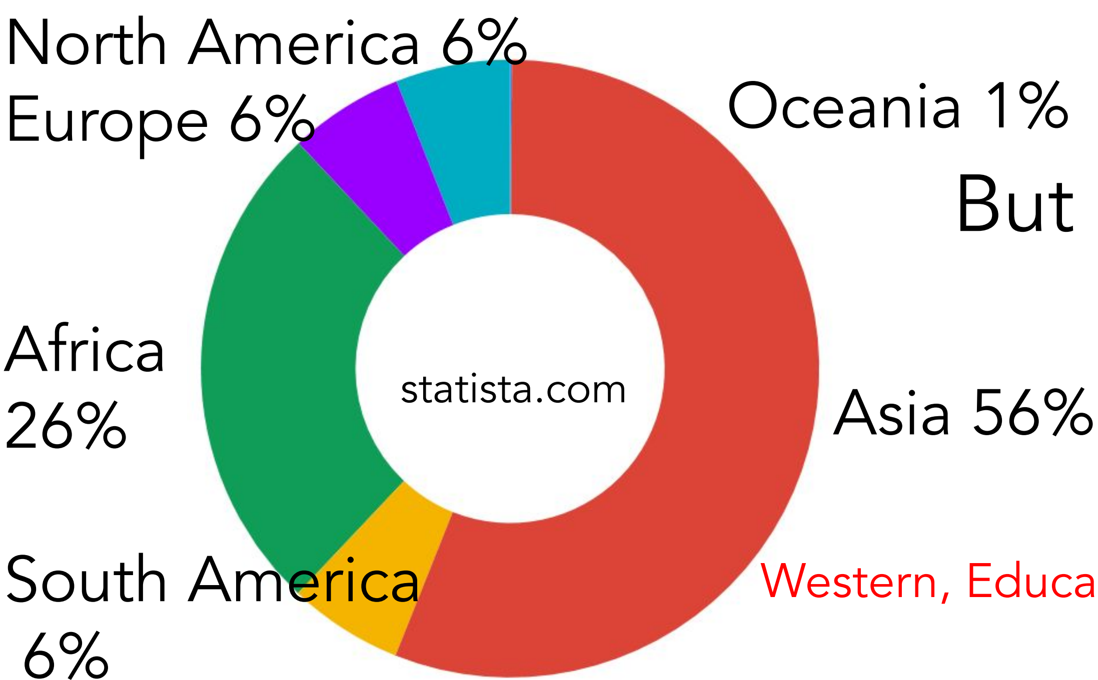
Most developmental data is collected in North America and Europe



But most children live in Asia and Africa



Most developmental data is collected in North America and Europe



But most children live in Asia and Africa

WEIRD bias=

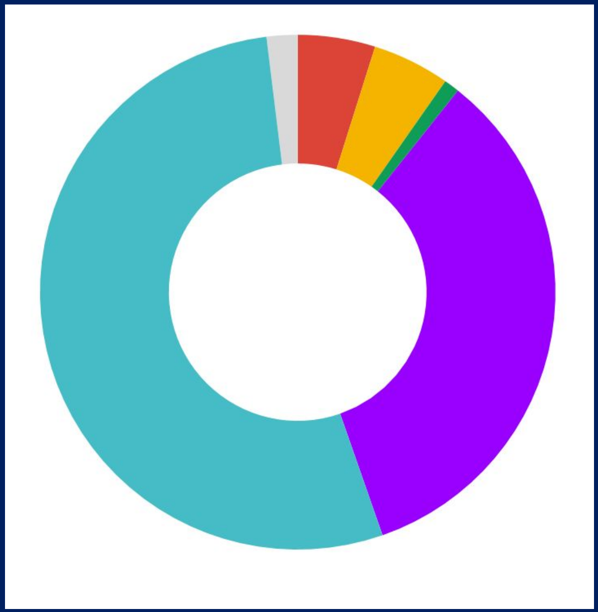
Western, Educated, Industrialized, Rich, Democratic

Heinrich et al. 2010

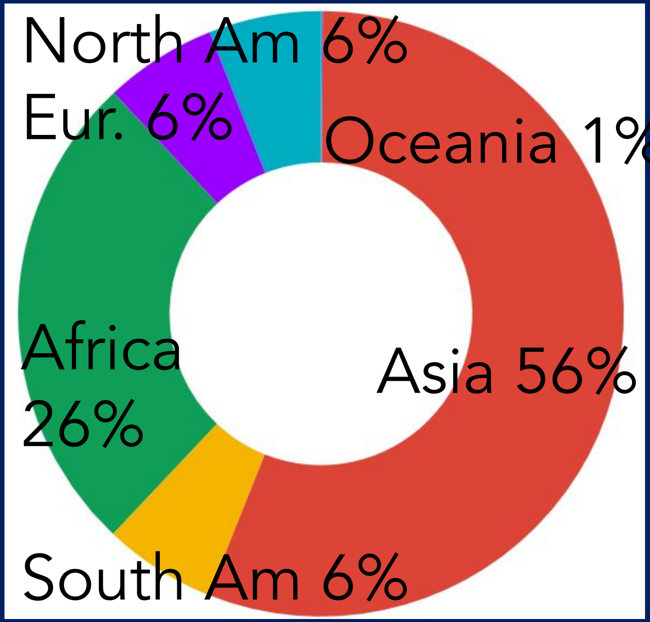
Please write in the chat where you grew up...

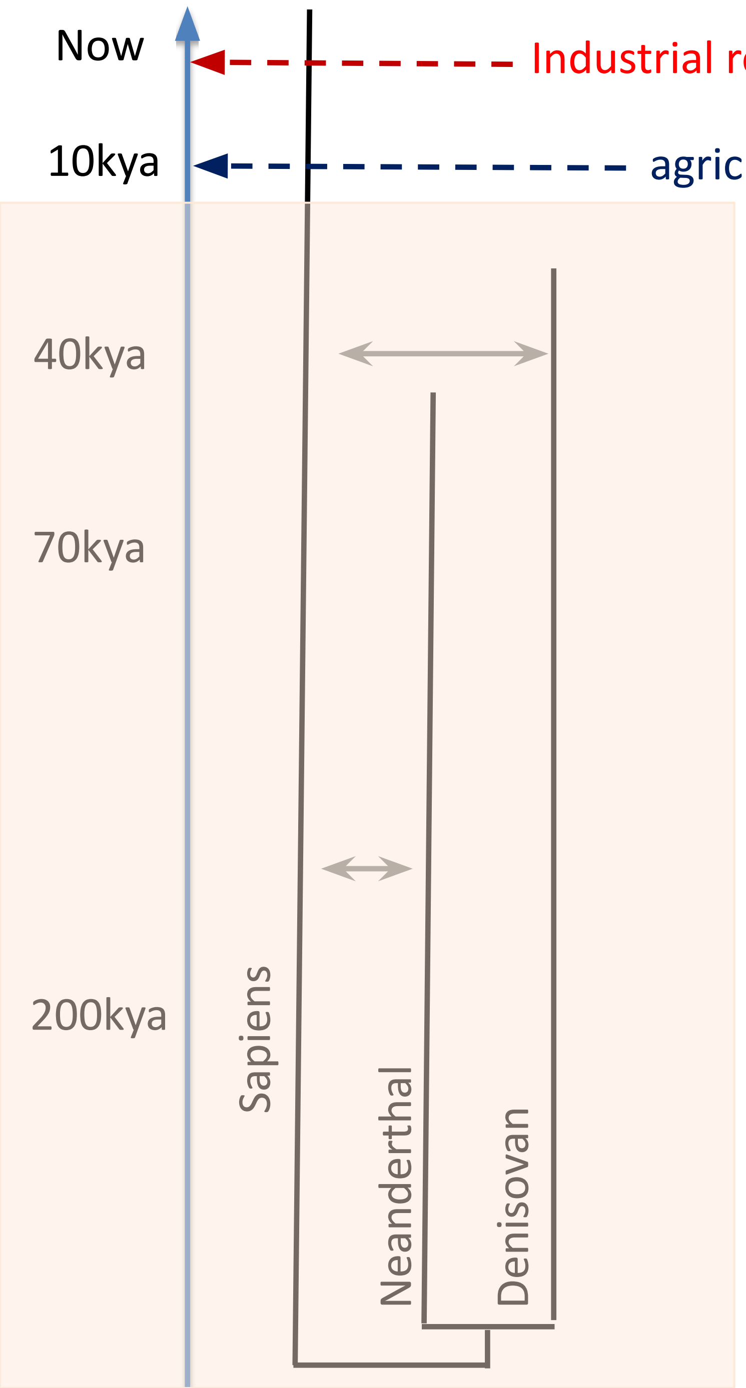
For instance, for me, that would be:
Rosario (large city), Argentina, South America

Developmental research



Developmental reality





WEIRD settings do not represent <u>natural</u> human ecology

rural
lower socioeconomic status
less formal education
greater diversity in ecological settings

Tree from Dediu & Levinson 2013, *Frontiers*
Levinson & Holler, 2014 *Phil.T.R.Soc*.

Does the WEIRD bias matter?
Comparing 'urban' & 'rural' families

industrialized
higher socioeconomic status
more formal education
fewer children
single caregiver

rural
lower socioeconomic status
less formal education
more children
shared caregiving

higher prevalence child-directed speech predicted

North-American

urban dwellers

average # children: 1.93

Statista 2021

rural

!Kung

hunter-gatherers

average # children: 4

Konner 2016

Tsimane'

hunter-farmers

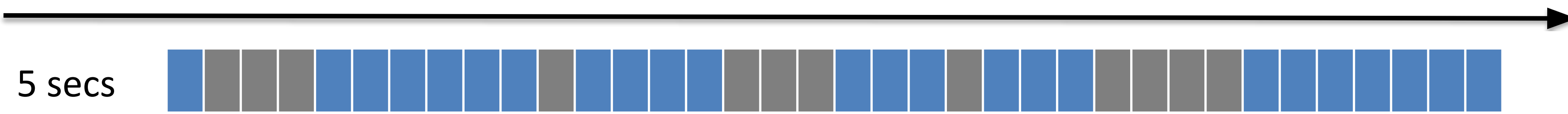
average # children: 9

Stieglitz et al. 2013

lower prevalence child-directed speech predicted

'Urban' versus 'rural' input quantities A systematic review of previous literature using behavioral observations

Most common method: “Time sampling”

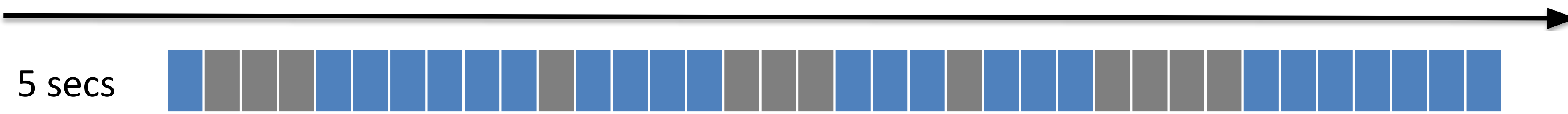


12 out of 24 →
frequency of infant-directed vocalizations is 0.50

Cristia (under review)

'Urban' versus 'rural' input quantities A systematic review of previous literature using behavioral observations

Most common method: “Time sampling”



12 out of 24 →
frequency of infant-directed vocalizations is 0.50

27 anthropology & social psychology papers

totaling 1,284 children

Dependent variable: % observations with infant-directed vocalizations
~ how frequently children are talked to in urban versus rural setting

Cristia (under review)

Write your guess in the chat!

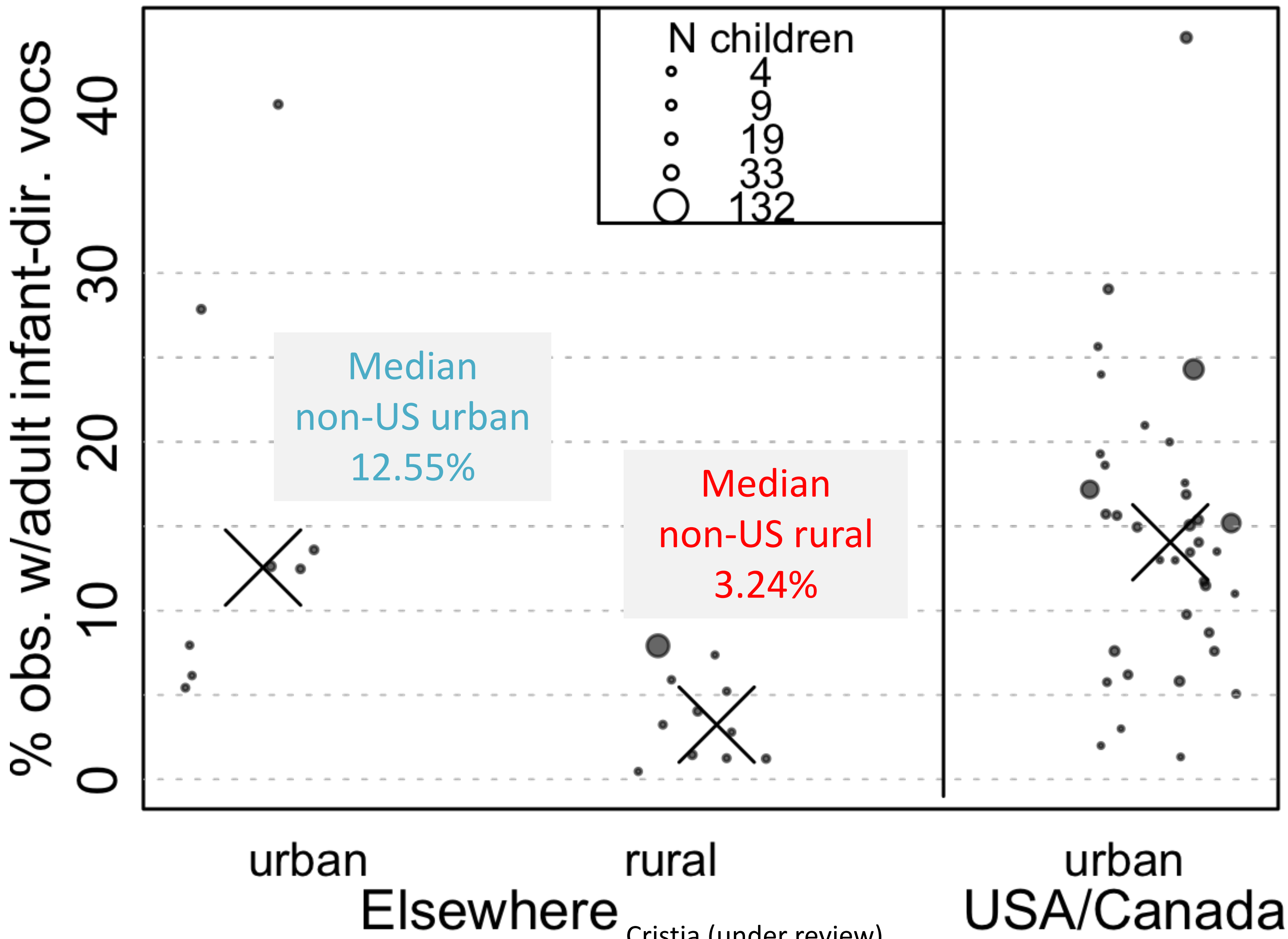
$$\frac{\text{how frequently \textbf{urban} infants get talked to}}{\text{how frequently \textbf{rural} infants get talked to}}$$

= 1 → same amount

= 1.1 → 10% more in urban than rural

= 2 → 100% more (=twice as much) in urban than rural

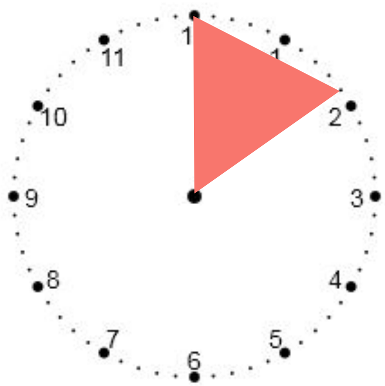
Urban/rural ratio: 3.87 (287% more)



Cristia (under review)

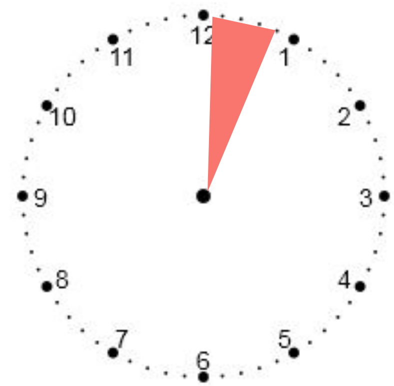
Or, converted to time…

US/non-US urban



1.5h
infant-directed
vocalizations (in a
12h awake day)

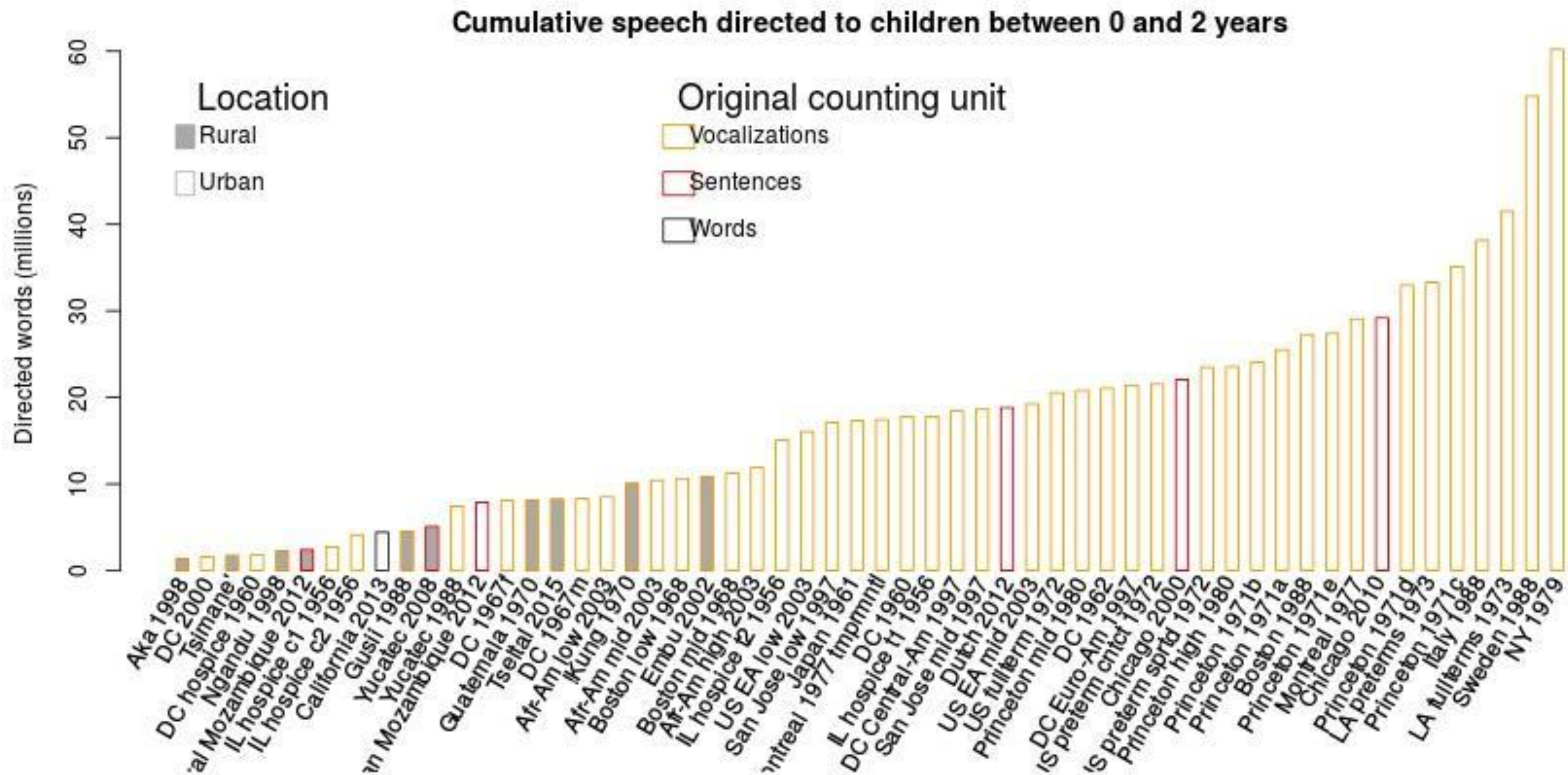
Non-urban,
non-USA



0.4h infant-directed
vocalizations (in a
12h awake day)

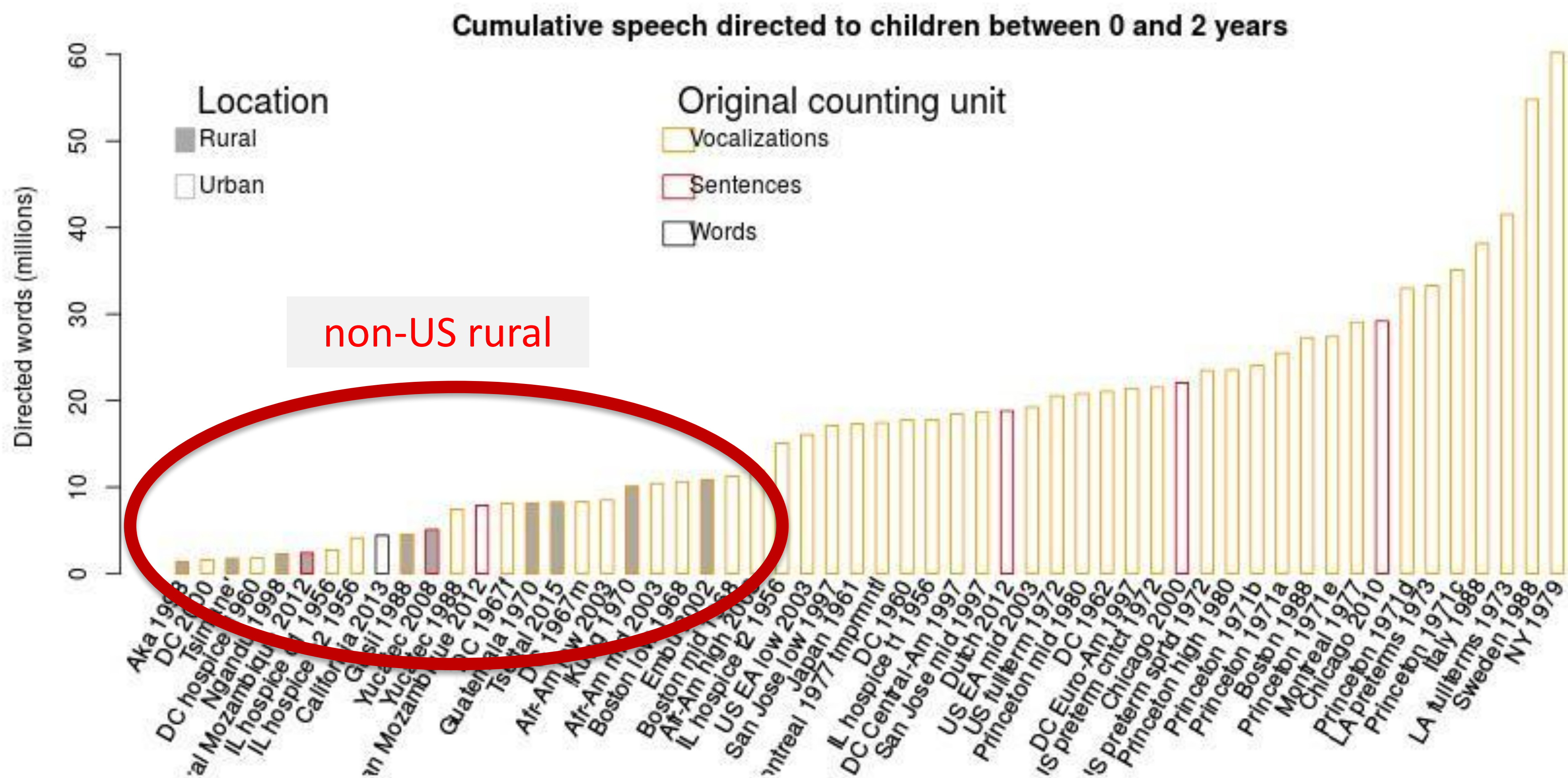
Cristia (under review)

Cross-population differences may be under-estimated xcult.shinyapps.io/vocsr/



Cristia (under review)

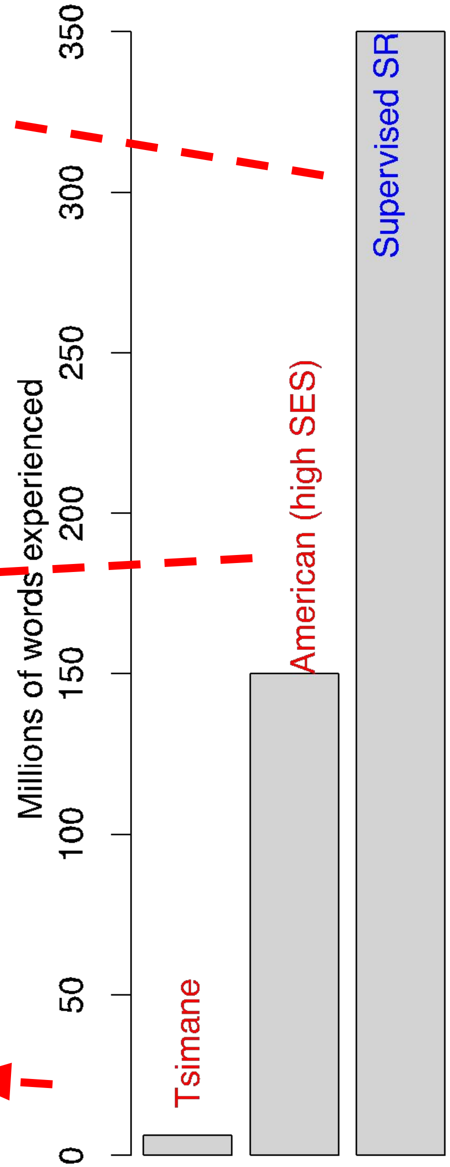
Cross-population differences may be under-estimated xcult.shinyapps.io/vocsr/



Cristia (under review)

Baby-machine comparison is even more astounding:

MS's first-pass human-level ASR transcription



humans cumulated to 10 years of age

Children **everywhere** learn to **perceive (& produce) speech** with

much less input

& supervision

than machines do

Supervised SR: Xiong et al. 2016 arXiv
American: Hart & Risley (1995)
Tsimane: Cristia et al. (2019) *Child Dev*

Wait.

Maybe this is <u>just</u> methodological variation, or differential observer effects homebank.talkbank.com

+ ecological
+ coverage

15 hours
(15$)

Casillas &
Cristia (2019)
Collabra

A day in the life…

14-hour recording centered on Natasha, aged 1 year (« key child »)
+ mother, sister, & father

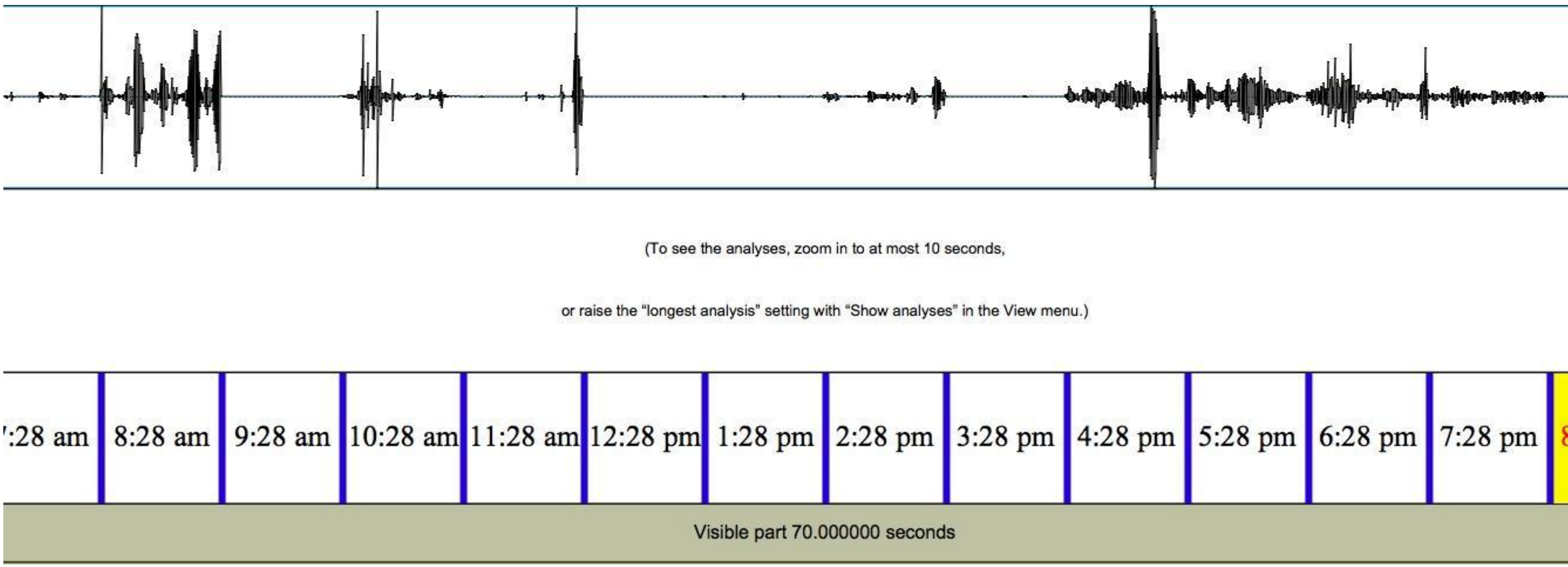
We extracted 5 seconds per hour periodically

full recording browsable at
https://sla.talkbank.org/TBB/homebank/Public/VanDam-Daylong/BN32/BN32_010007.cha

downloadable via
https://github.com/LAAC-LSCP/vandam-daylong-demo

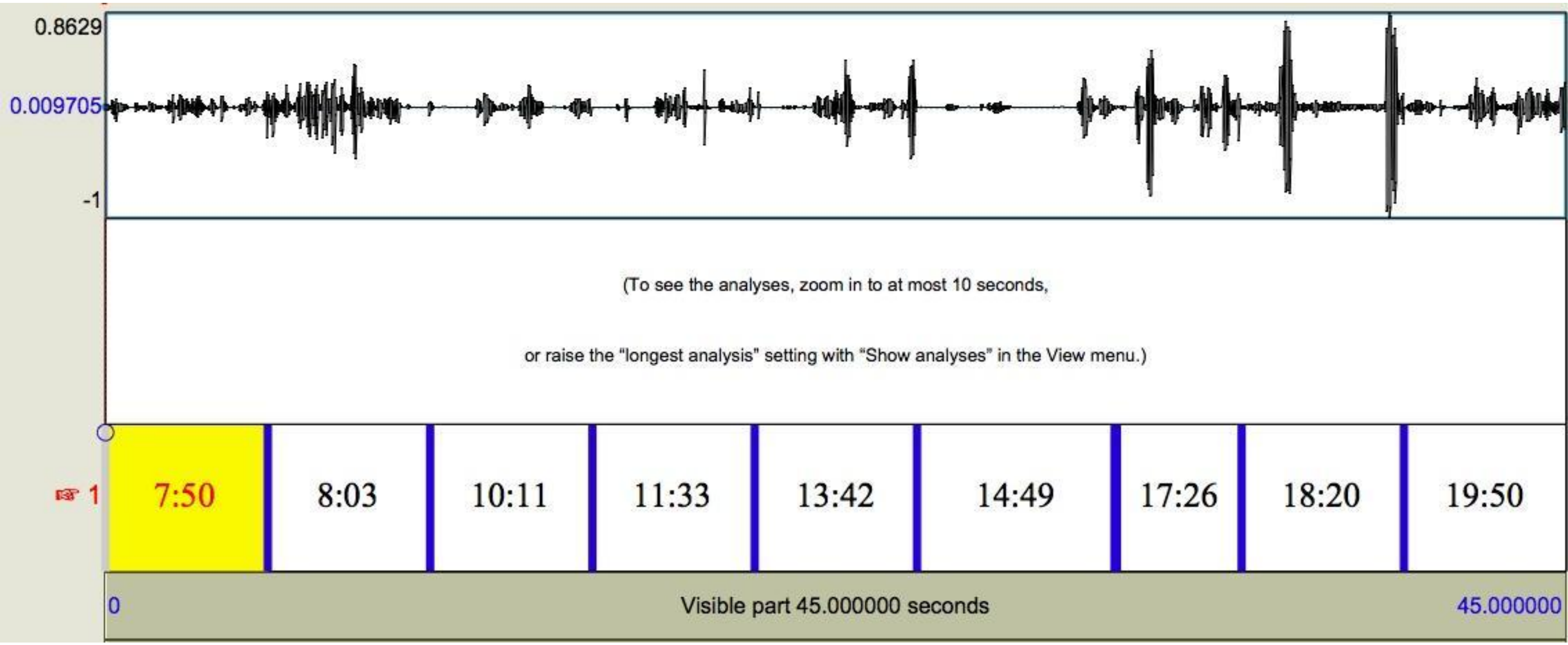
VanDam, Mark (2018). VanDam Public Daylong HomeBank Corpus. doi:10.21415/T5388S

A day in the life…



most of this child’s day is silent, so we exclude silent sections & try again…

A day in the life…



« key child » only heard a couple of times

most speech is from mother & father

sibling heard too, talking to parents (not to « key child »)

A word on long-form recordings

cheap

unobtrusive

field-work friendly

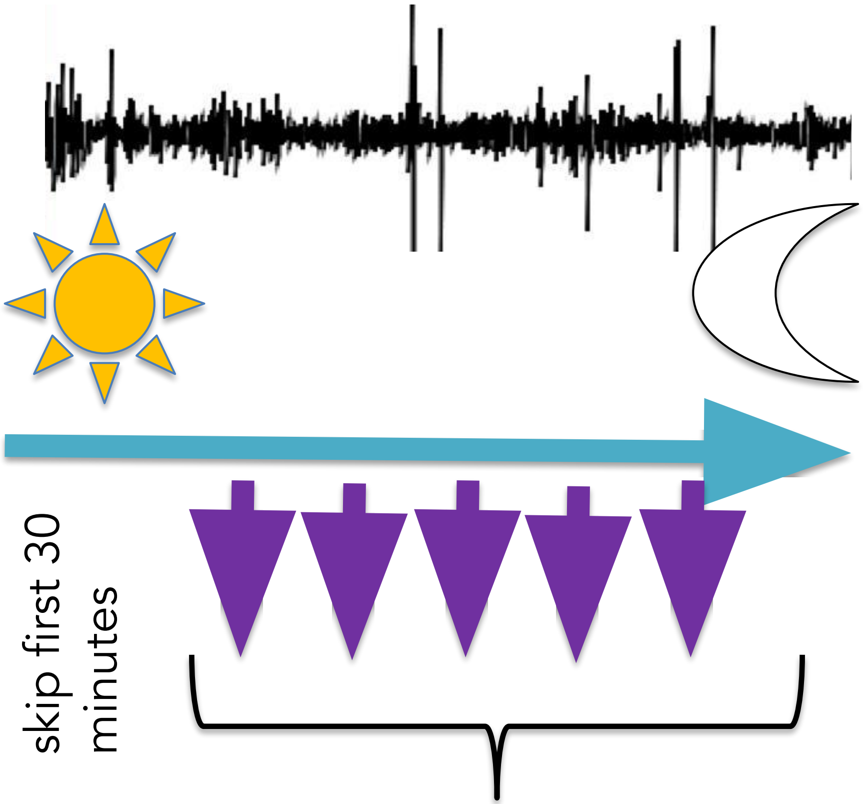
high re-use potential (anthropology, biology, economics, linguistics, etc.)

private information

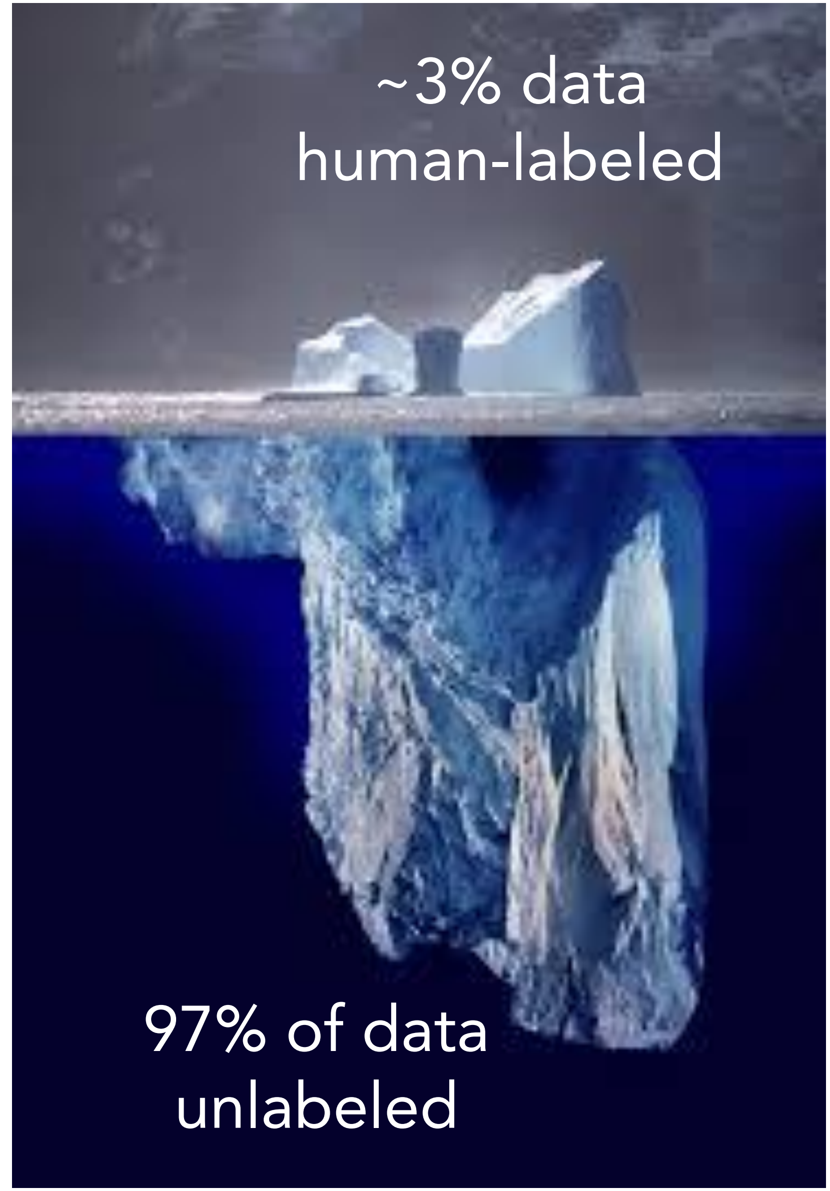
SO . MUCH . DATA

Gautheron, Rochat, & Cristia 2021 ([preprint]())

Ask me about all this!



code 15 x 2
minute clips
from 10 children



Preliminary results

overall child-directed speech quantity fairly stable across populations

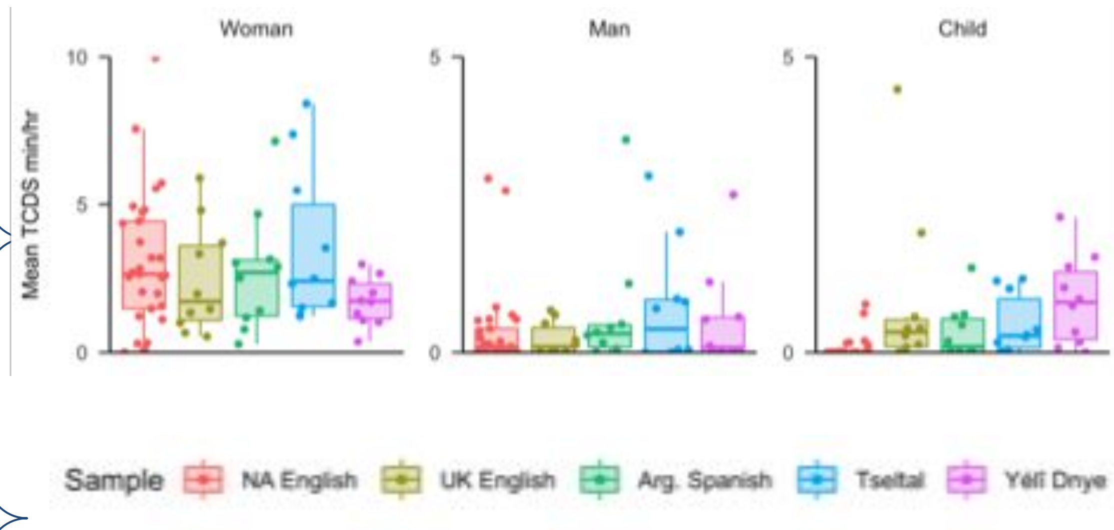
Language	TCDS rate	
NA English	3.49 (3.24; 0-10.12)	
UK English	3.69 (3.72; 1.22-7.15)	urban
Arg. Spanish	4.77 (3.19; 1.4-9.38)	
Tseltal	3.54 (3.94; 0.83-6.55)	rural
Yélî Dnye	3.13 (2.95; 1.58-6.26)	

hand-annotated data analyzed in Bunce et al. (2021)

Preliminary results

overall child-directed speech quantity
fairly stable across populations

Language	TCDS rate
NA English	3.49 (3.24; 0-10.12)
UK English	3.69 (3.72; 1.22-7.15)
Arg. Spanish	4.77 (3.19; 1.4-9.38)
Tseltal	3.54 (3.94; 0.83-6.55)
Yélî Dnye	3.13 (2.95; 1.58-6.26)



sizable **source** variation across populations

hand-annotated data analyzed
in Bunce et al. (2021)

Preliminary results

overall child-directed speech quantity fairly stable across populations

Language	TCDS rate
NA English	3.49 (3.24; 0-10.12)
UK English	3.69 (3.72; 1.22-7.15)
Arg. Spanish	4.77 (3.19; 1.4-9.38)
Tseltal	3.54 (3.94; 0.83-6.55)
Yélî Dnye	3.13 (2.95; 1.58-6.26)

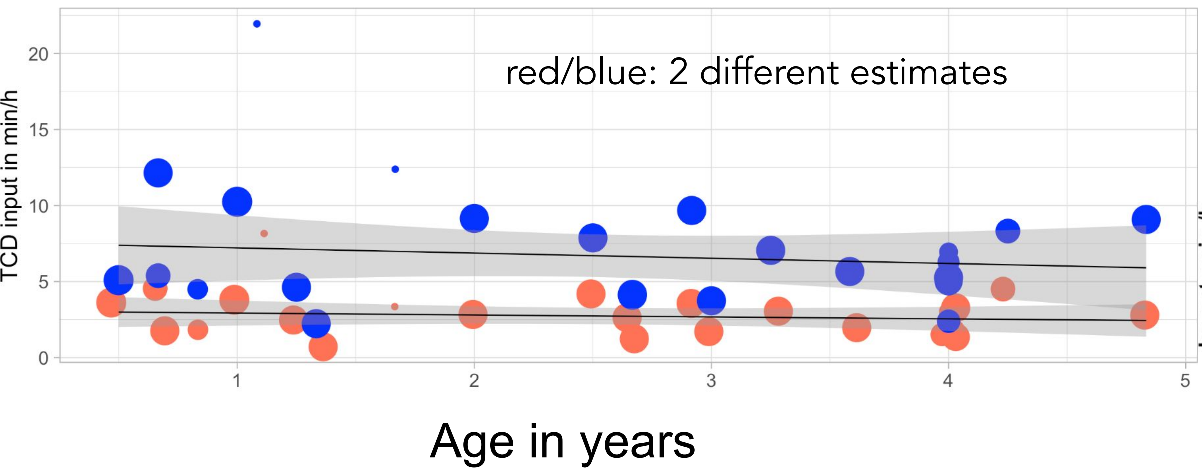
sizable **source** variation across populations

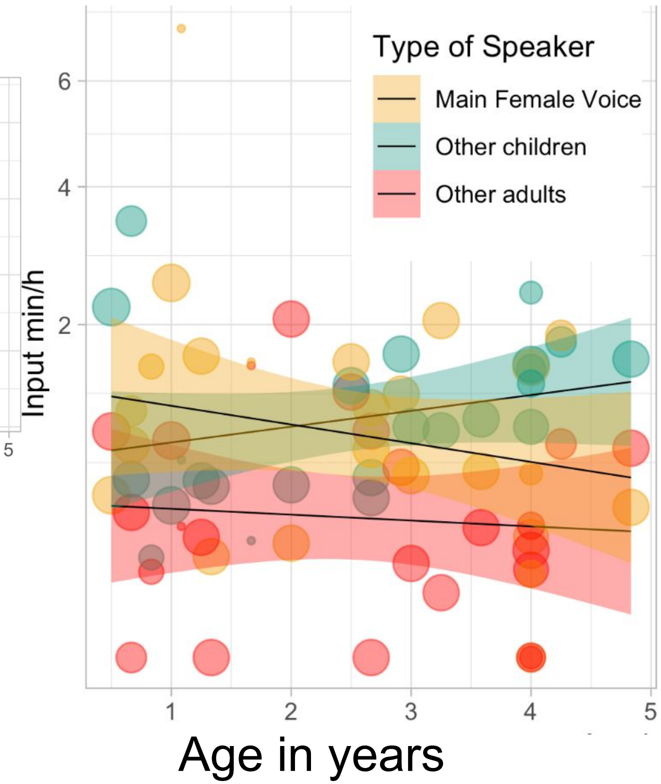
hand-annotated data analyzed in Bunce et al. (2021)

Example from hand-annotated data from the Tsimane' (hunter-horticulturalist in Lowland Bolivia)

Preliminary results

Both input quantities & sources vary a lot **across individuals**





Scaff et al. (in prep)

Interim take-home messages

Very different results when looking at

- behavioral observations (3x difference between rural and urban, up to 10x across populations)
- long-form audiorecordings (overlap between rural and urban, up to 2/4x across populations)

Technique effects
short/whispered speech missed by observers?

Observer effects
perhaps rural vs. urban families react differently to observers?

Interim take-home messages

Very different results when looking at

- behavioral observations (3x difference between rural and urban, up to 10x across populations)
- long-form audiorecordings (overlap between rural and urban, up to 2/4x across populations)

Technique effects
short/whispered speech missed by observers?

Observer effects
perhaps rural vs. urban families react differently to observers?

Tremendous individual variation!

Interim take-home messages

Very different results when looking at

- behavioral observations (3x difference between rural and urban, up to 10x across populations)
- long-form audiorecordings (overlap between rural and urban, up to 2/4x across populations)

Technique effects
short/whispered speech missed by observers?

Observer effects
perhaps rural vs. urban families react differently to observers?

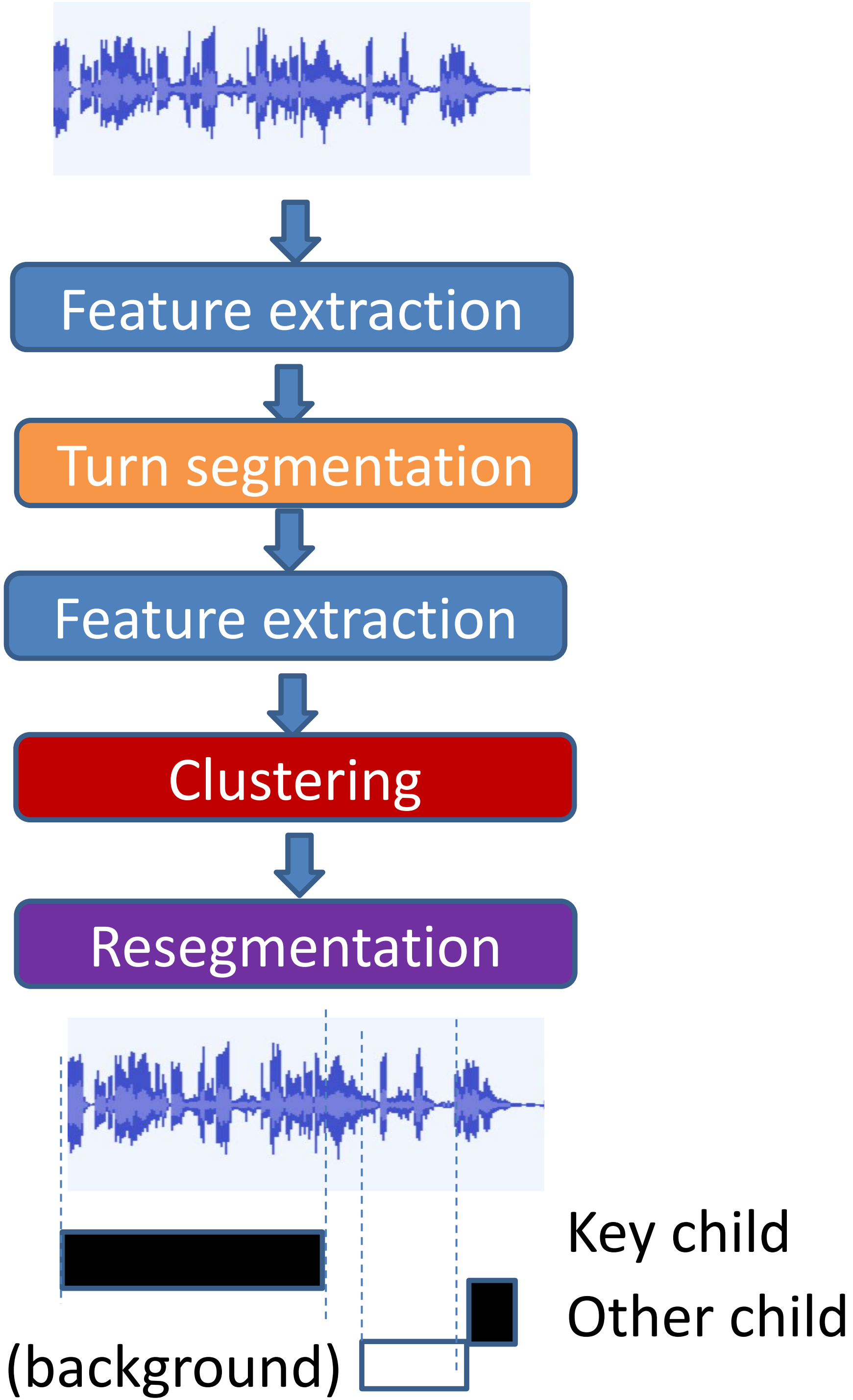
Tremendous individual variation!

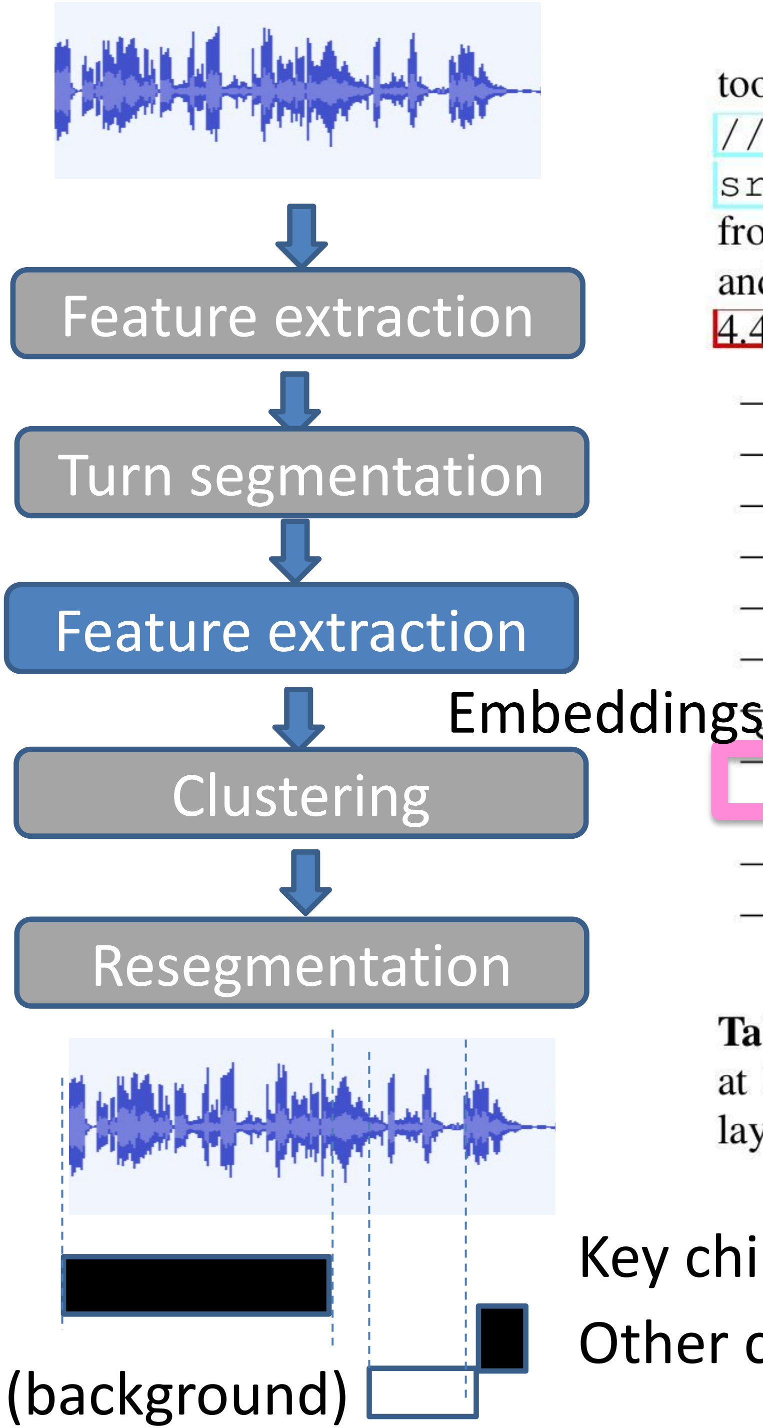
Estimation accuracy?
based on very little data!

Building classifiers to generalize to unlabeled data

Talker diarization
(who speaks when)

DIHARD 2018, 2019/2021 Interspeech


Feature extraction
Turn segmentation
Feature extraction
Clustering
Resegmentation
Key child
Other child
(background)

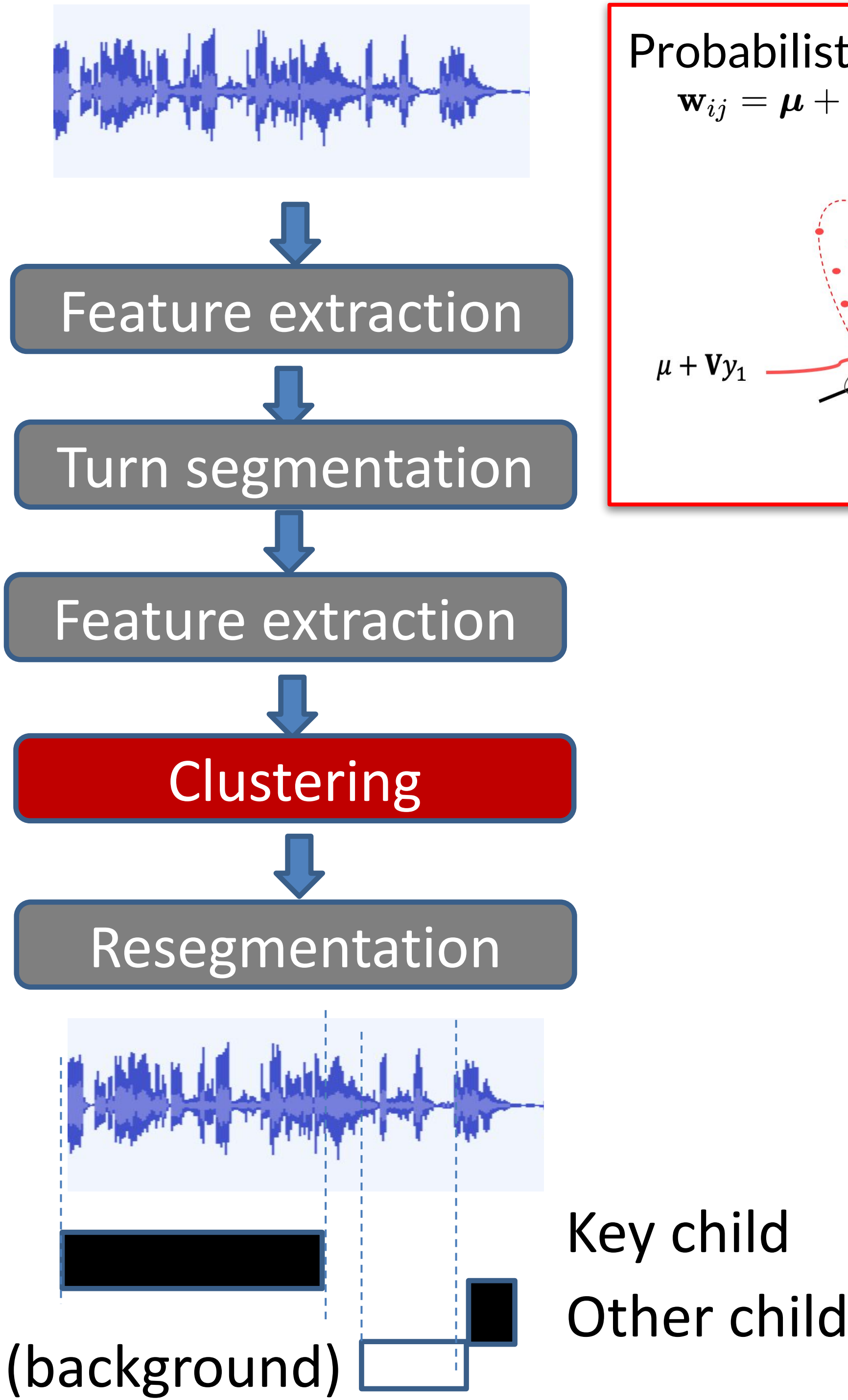


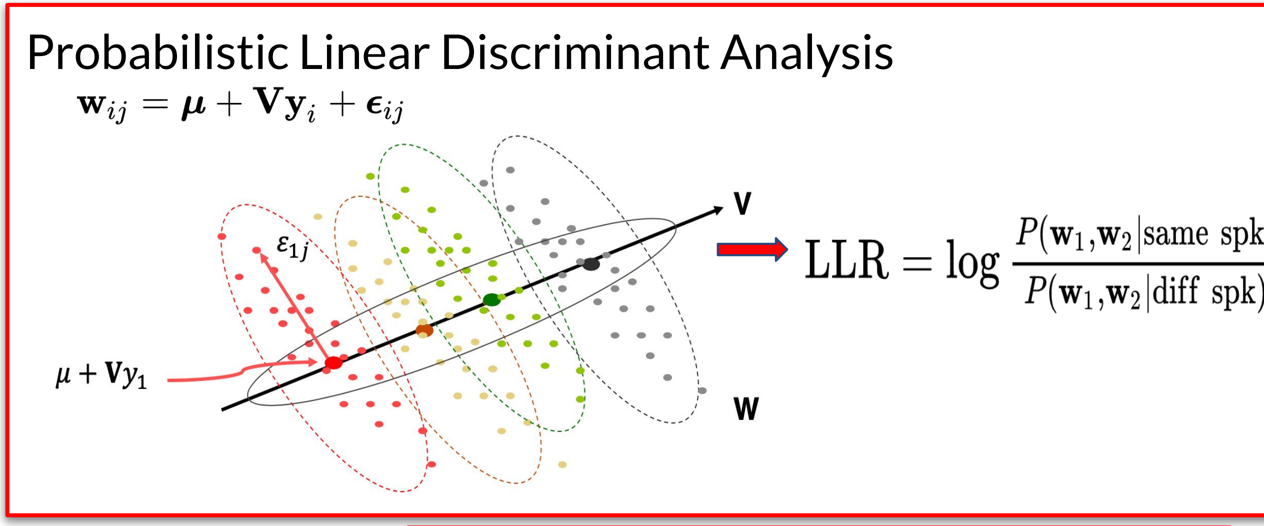
Our software framework has been made available in the Kaldi toolkit. An example recipe is in the main branch of Kaldi at https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2 and a pretrained x-vector system can be downloaded from http://kaldi-asr.org/models.html. The recipe and model are similar to the x-vector system described in Section 4.4.

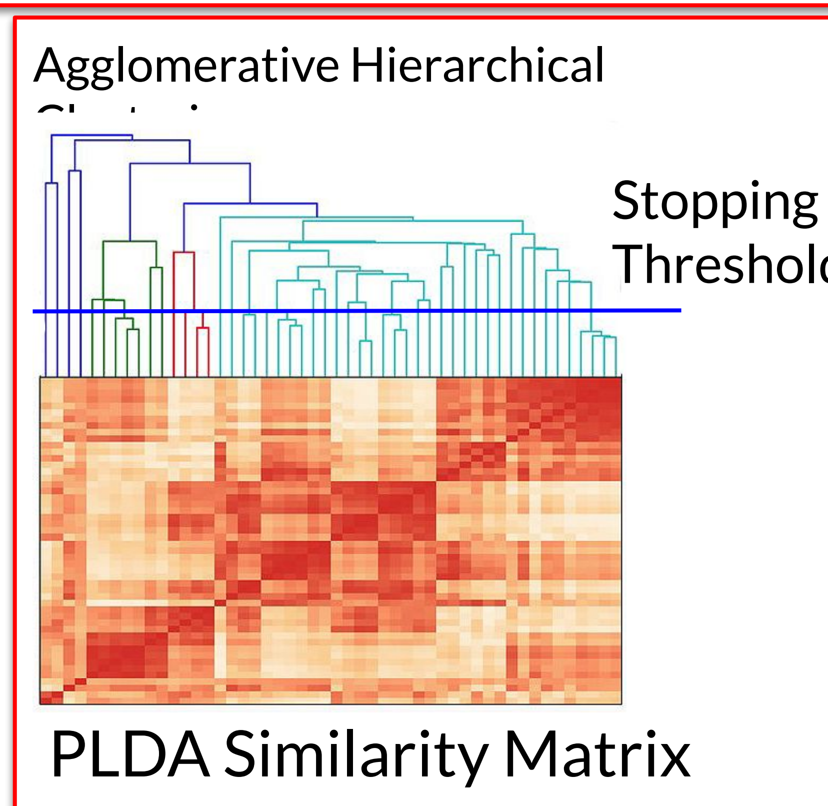
Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN

Table 1. The embedding DNN architecture. x-vectors are extracted at layer *segment6*, before the nonlinearity. The N in the softmax layer corresponds to the number of training speakers.

Snyder et al. 2018 ICASSP

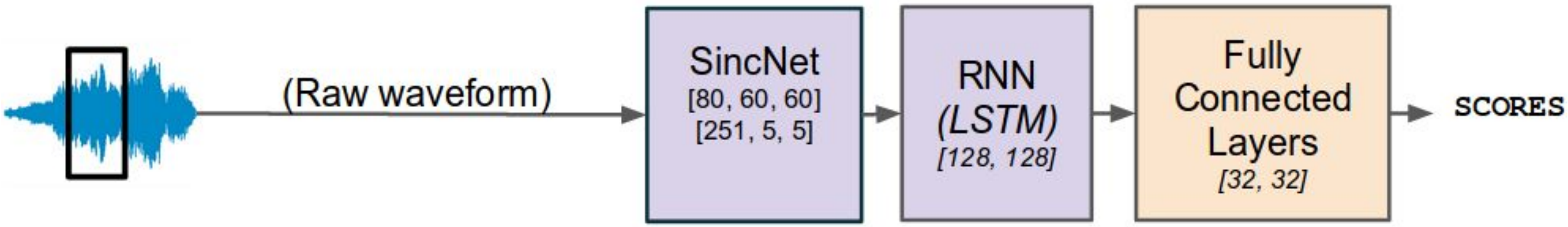






images by J. Villalba (JHU)

State of the art in voice type classification



Class	Precision	Recall	Fscore
KCHI	81.69	73.48	77.37
CHI	18.78	40.45	25.65
FEM	77.94	87.40	82.40
MAL	37.82	47.86	42.25
SPEECH	85.51	91.59	88.45
AVE	60.35	68.15	63.22

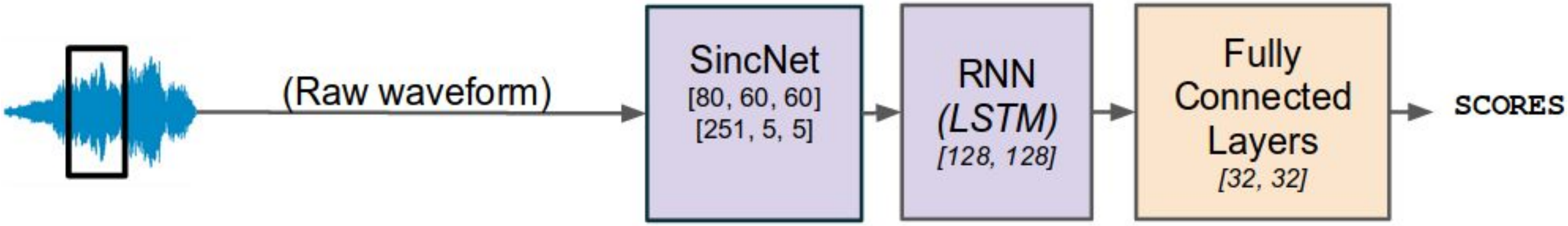
Tab 2. Performances of our model on the test set.

Class	Precision	Recall	Fscore
KCHI	62.37	76.67	68.78
CHI	46.77	25.78	33.24
FEM	70.30	57.87	63.48
MAL	39.52	46.92	42.91
SPEECH	77.03	79.89	78.43
AVE	59.20	57.42	57.37

Tab 3. Performances of our model on the held-out set.

Lavechin et al. 2020 Interspeech code

State of the art in voice type classification



Class	Precision	Recall	Fscore
KCHI	81.69	73.48	77.37
CHI	18.78	40.45	25.65
FEM	77.94	87.40	82.40
MAL	37.82	47.86	42.25
SPEECH	85.51	91.59	88.45
AVE	60.35	68.15	63.22

Tab 2. Performances of our model on the test set.

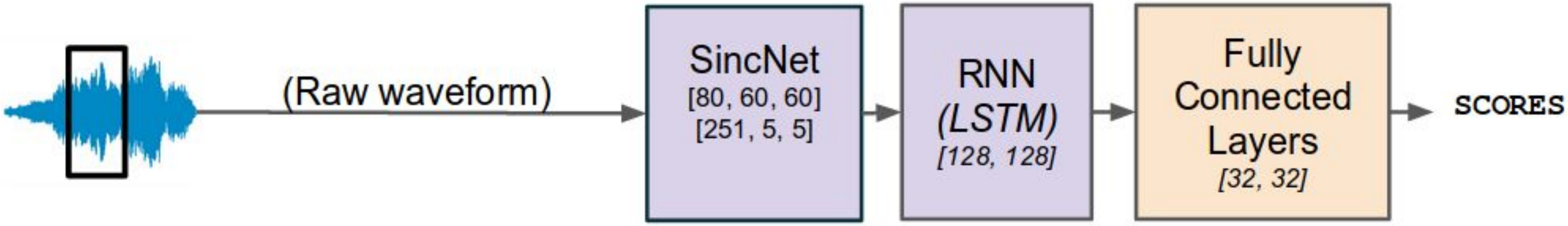
Class	Precision	Recall	Fscore
KCHI	62.37	76.67	68.78
CHI	46.77	25.78	33.24
FEM	70.30	57.87	63.48
MAL	39.52	46.92	42.91
SPEECH	77.03	79.89	78.43
AVE	59.20	57.42	57.37

Tab 3. Performances of our model on the held-out set.

OK performance on key child (wearing the device) & female adult voice

Lavechin et al. 2020 Interspeech code

State of the art in voice type classification



Class	Precision	Recall	Fscore
KCHI	81.69	73.48	77.37
CHI	18.78	40.45	25.65
FEM	77.94	87.40	82.40
MAL	37.82	47.86	42.25
SPEECH	85.51	91.59	88.45
AVE	60.35	68.15	63.22

Tab 2. Performances of our model on the test set.

Class	Precision	Recall	Fscore
KCHI	62.37	76.67	68.78
CHI	46.77	25.78	33.24
FEM	70.30	57.87	63.48
MAL	39.52	46.92	42.91
SPEECH	77.03	79.89	78.43
AVE	59.20	57.42	57.37

Tab 3. Performances of our model on the held-out set.

sad performance on other child (NOT wearing the device) & male adult voice

Lavechin et al. 2020 Interspeech [code]

(Algorithm) bias

Table 1: *Description of the BabyTrain data set. Child-centered corpora included cover a wide range of conditions (including differen languages and recording devices). ACLEW-Random is kept as a held-out data set on which LENA and our model are compared.*

Corpus	LENA-recorded?	Language	Tot. Dur.	Cumulated utterance duration				
				KCHI	OCH	MAL	FEM	UNK
			BabyTrain					
ACLEW-Starter	mostly	Mixture	1h30m	10m	5m	6m	20m	0m
Lena Lyon	yes	French	26h51m	4h33m	1h14m	1h9m	5h02m	1h0m
Namibia	no	Ju|'hoan	23h44m	1h56m	1h32m	41m	2h22m	1h01m
Paido	no	Greek, Eng., Jap.	40h08m	10h56m	0m	0m	0m	0m
Tsay	no	Mandarin	132h02m	34h07m	2h08m	10m	57h31m	28m
Tsimane	mostly	Tsimane	9h30m	37m	23m	11m	28m	0m
Vanuatu	no	Mixture	2h29m	12m	5m	5m	9m	1m
WAR2	yes	English (US)	50m	14m	0m	0m	0m	9m

~50h key child

>60h female adult

(Algorithm) bias

Table 1: *Description of the BabyTrain data set. Child-centered corpora included cover a wide range of conditions (including differen languages and recording devices). ACLEW-Random is kept as a held-out data set on which LENA and our model are compared.*

				Cumulated utterance duration				
Corpus	LENA-recorded?	Language	Tot. Dur.	KCHI	OCH	MAL	FEM	UNK
			BabyTrain					
ACLEW-Starter	mostly	Mixture	1h30m	10m	5m	6m	20m	0m
Lena Lyon	yes	French	26h51m	4h33m	1h14m	1h9m	5h02m	1h0m
Namibia	no	Ju|'hoan	23h44m	1h56m	1h32m	41m	2h22m	1h01m
Paido	no	Greek, Eng., Jap.	40h08m	10h56m	0m	0m	0m	0m
Tsay	no	Mandarin	132h02m	34h07m	2h08m	10m	57h31m	28m
Tsimane	mostly	Tsimane	9h30m	37m	23m	11m	28m	0m
Vanuatu	no	Mixture	2h29m	12m	5m	5m	9m	1m
WAR2	yes	English (US)	50m	14m	0m	0m	0m	9m

~50h key child

>60h female adult

<5h other child

<3h male adult

Building classifiers to generalize to unlabeled data

Talker diarization
(who speaks when)

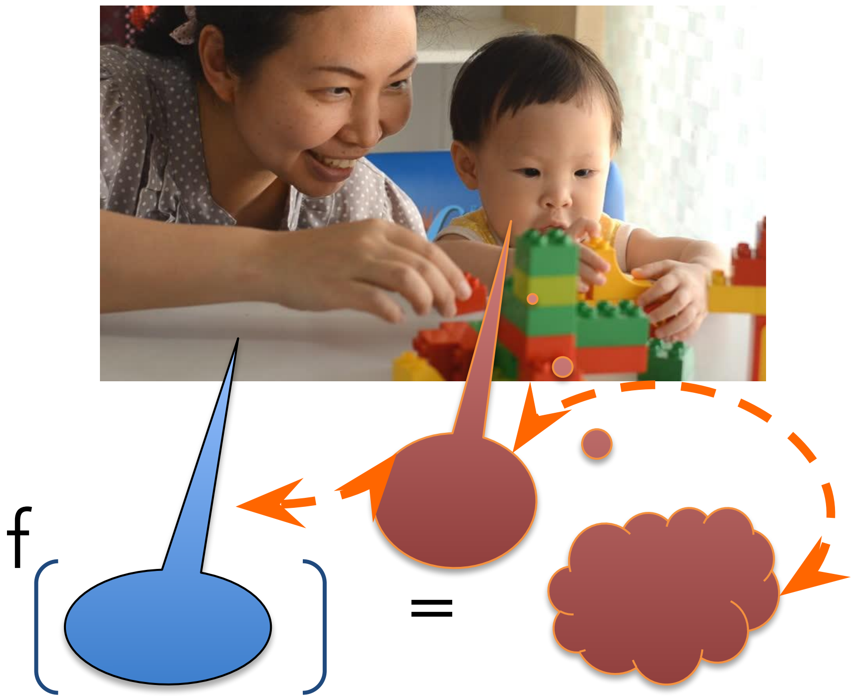
DIHARD 2018, 2019/2021 Interspeech

Addressee classification
(whom are they talking to)

ComParE 2017 Interspeech

2 classes,
no team beat the baseline

But what about acquisition outcomes?



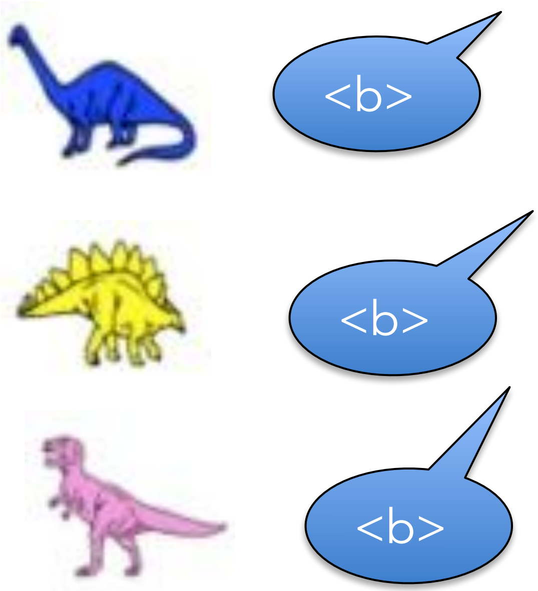
Example: categorization task with words

Perszyk & Waxman 2017 [JOVE]

Example: categorization task with words

Perszyk & Waxman 2017 [JOVE]

Example: categorization task with backward words



Perszyk & Waxman 2017 [JOVE]

Example: categorization task with lemur calls

Perszyk & Waxman 2017 [JOVE]

MetaLab Explore Data Documentation Publications Team

metalab.stanford.edu

MetaLab

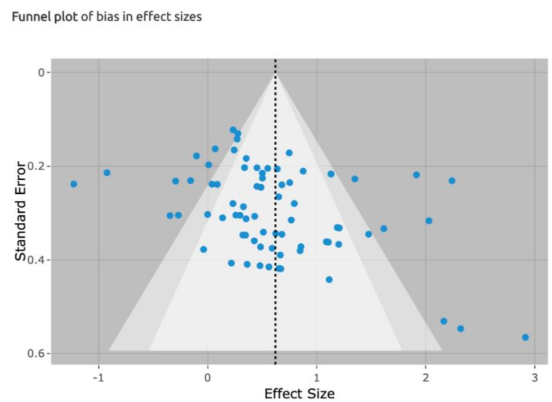
Interactive, community-augmented meta-analysis tools for cognitive development research

New: The [2020 Contribution Challenge Winners]

Explore Apps View Documentation ›

New MetaLab User? Check out Getting Started first!

The MetaLab database contains **2,496 effect sizes** from **30 meta-analyses** across two domains of cognitive development, based on data from **687 papers** and **45,244 subjects**.



Data from ~30 phenomena (including looking-while-listening)

Over 45k children represented

Domains

Early Language

How do children learn their native language?

24 meta-analyses **550 papers** **2,134 effect sizes** **38,961 subjects**

Cognitive Development

What is the nature of children's understanding?

6 meta-analyses **137 papers** **362 effect sizes** **6,283 subjects**

MetaLab | Explore Data | Documentation | Publications | Team

metalab.stanford.edu

MetaLab

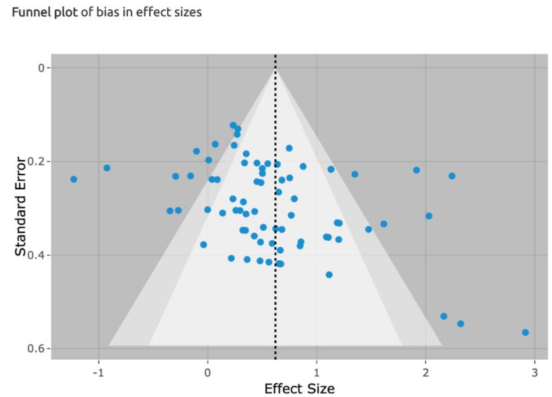
The MetaLab database contains **2,496 effect sizes** from **30 meta-analyses** across two domains of cognitive development, based on data from **687 papers** and **45,244 subjects**.

Interactive, community-augmented meta-analysis tools for cognitive development research

New: The [2020 Contribution Challenge Winners]

Explore Apps   View Documentation ›

New MetaLab User? Check out Getting Started first!



Data from ~30 phenomena (including "categorization task")

Over 45k children represented

even more biased than data discussed above!
(1 eg: 75% NorthAm, 23% Eur, 2% Asia)

Domains

Early Language

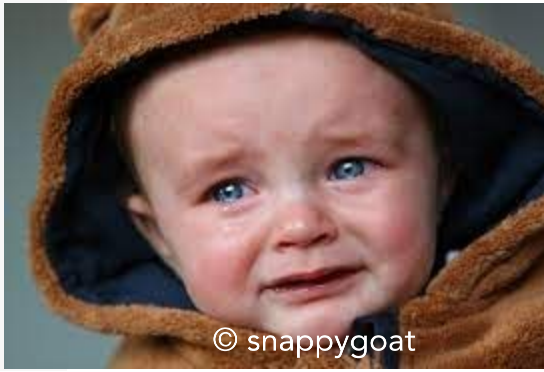
How do children learn their native language?

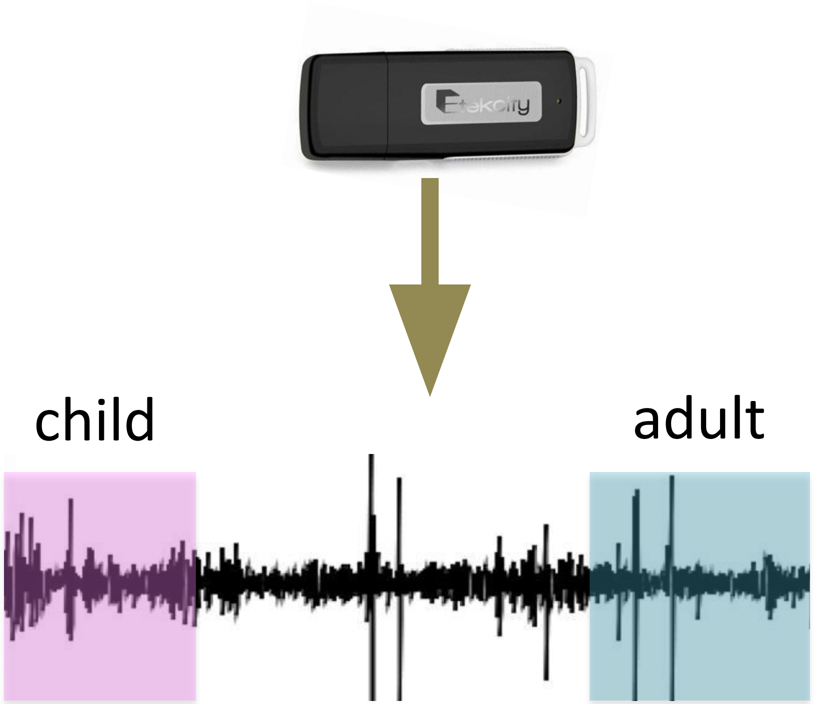
24 meta-analyses	550 papers	2,134 effect sizes	38,961 subjects

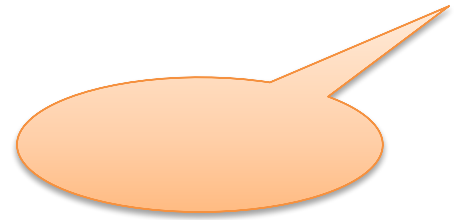
Cognitive Development

What is the nature of children's understanding?

6 meta-analyses	137 papers	362 effect sizes	6,283 subjects

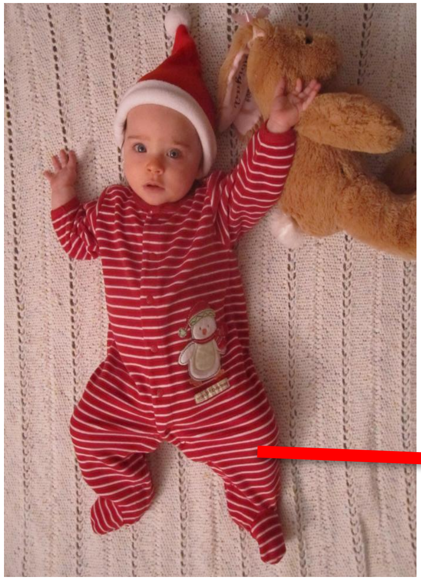


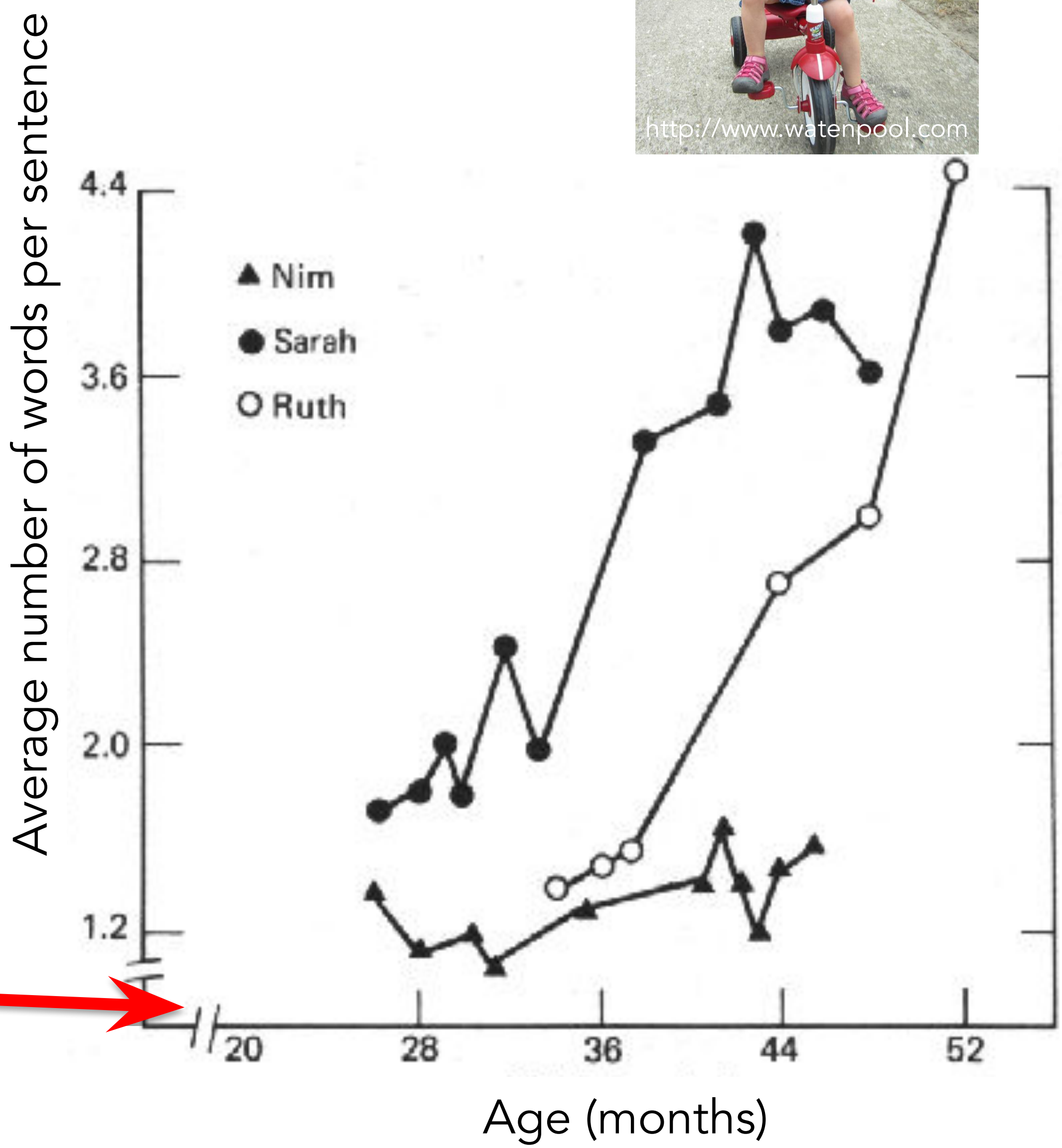




Long-form audio recordings to the rescue!

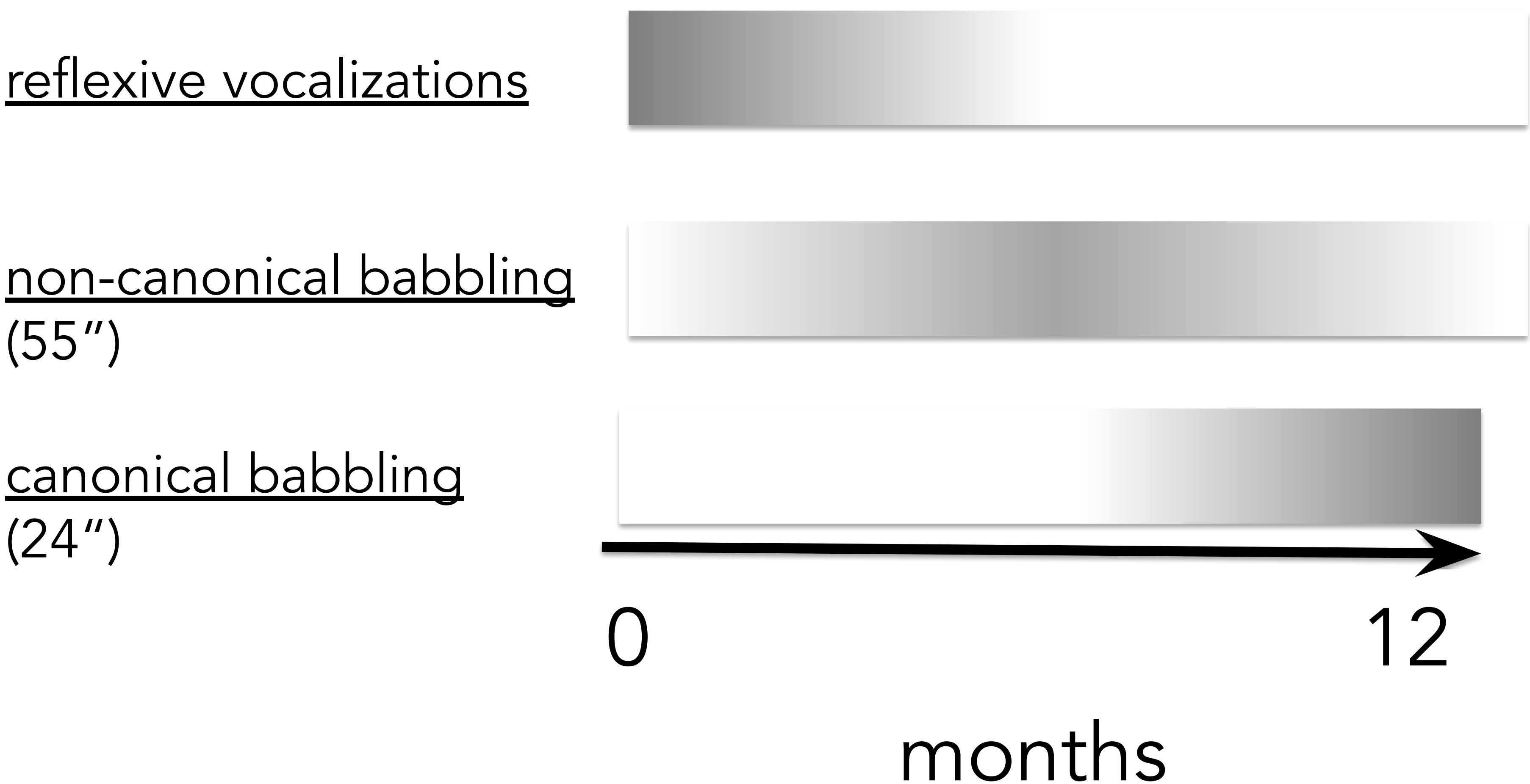
plenty
happens
before 1 year!

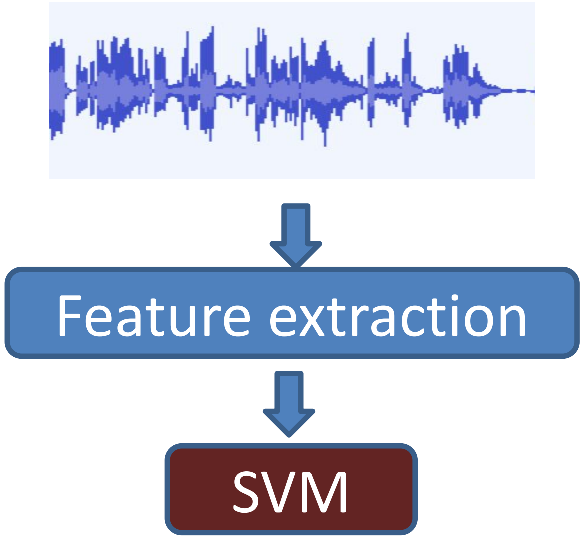


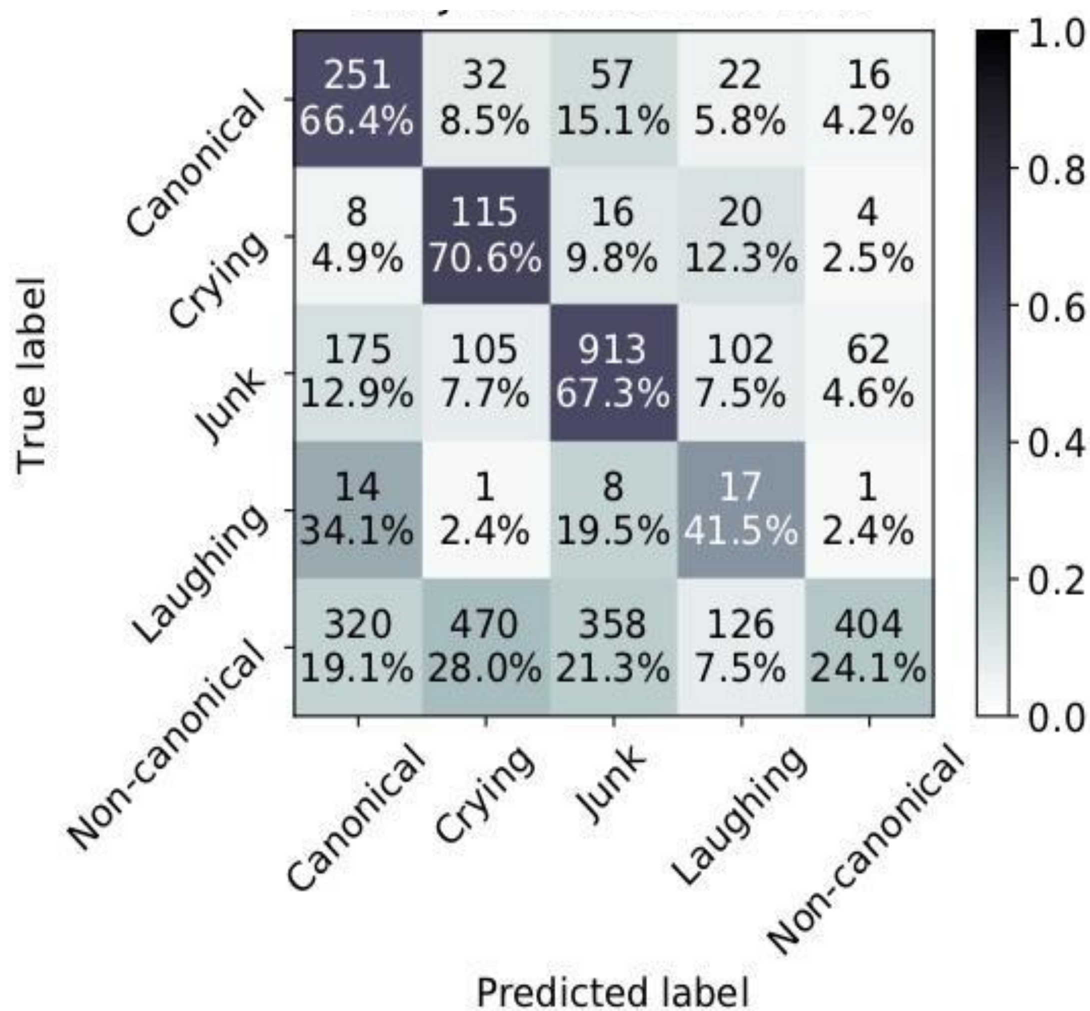


Terrace 1979 Science

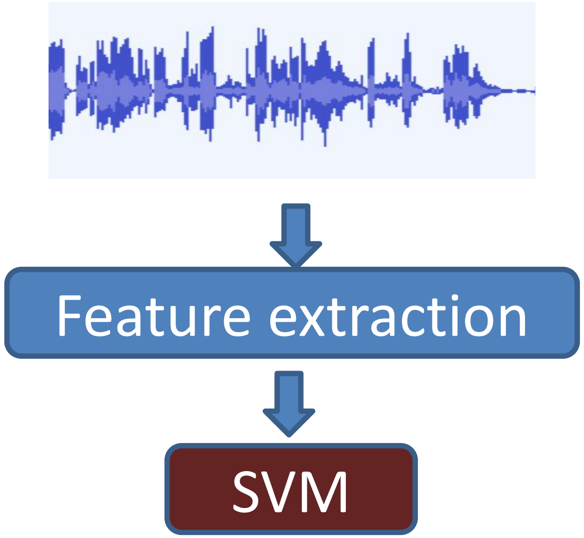
Vocalizations vary in complexity



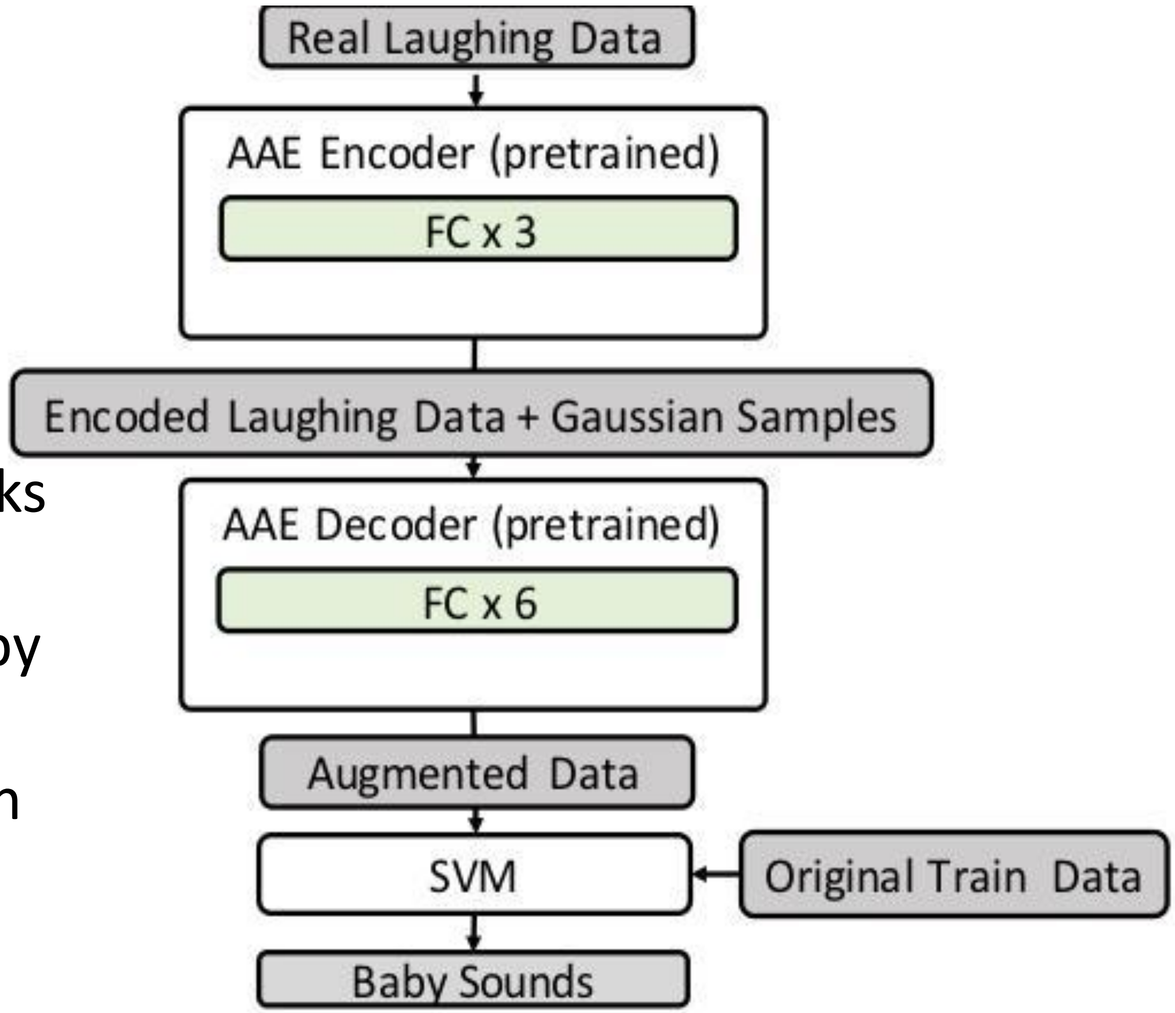

Feature extraction
SVM


True label
Predicted label
Canonical
Crying
Junk
Laughing
Non-canonical
251 66.4% | 32 8.5% | 57 15.1% | 22 5.8% | 16 4.2%
8 4.9% | 115 70.6% | 16 9.8% | 20 12.3% | 4 2.5%
175 12.9% | 105 7.7% | 913 67.3% | 102 7.5% | 62 4.6%
14 34.1% | 1 2.4% | 8 19.5% | 17 41.5% | 1 2.4%
320 19.1% | 470 28.0% | 358 21.3% | 126 7.5% | 404 24.1%
1.0
0.8
0.6
0.4
0.2
0.0

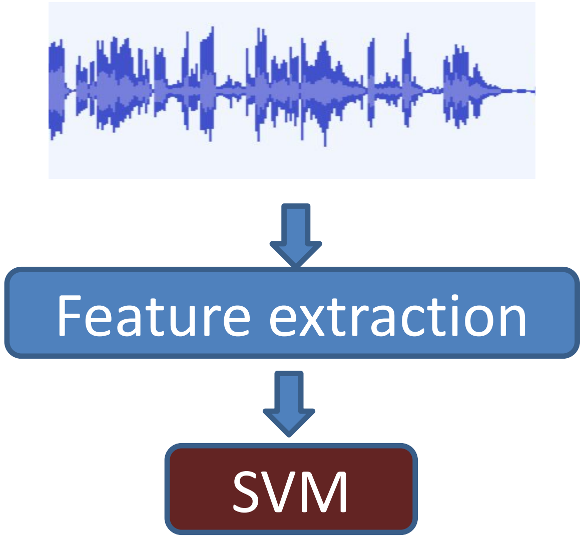
And the winner is…



"Using Attention Networks and Adversarial Augmentation for … Baby Sound Recognition", Sung-Lin Yeh ... Chi-Chun Lee

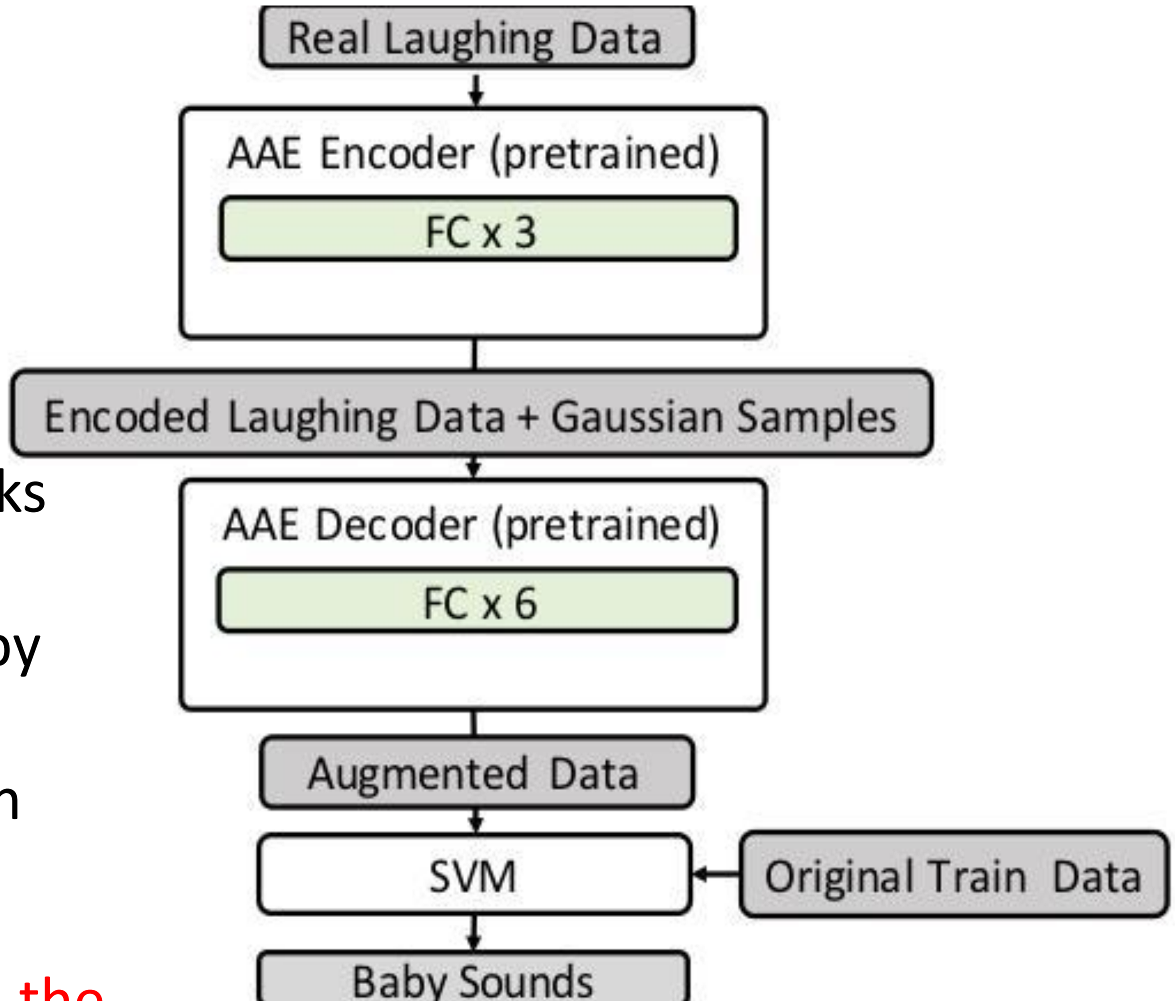


And the winner is…



"Using Attention Networks and Adversarial Augmentation for … Baby Sound Recognition", Sung-Lin Yeh ... Chi-Chun Lee

By 2% & through gains in the laughing category



Building classifiers to generalize to unlabeled data

Shamelessly stolen from Y. LeCun

Talker diarization
(who speaks when)

DIHARD 2018, 2019 Interspeech

Addressee classification
(whom are they talking to)

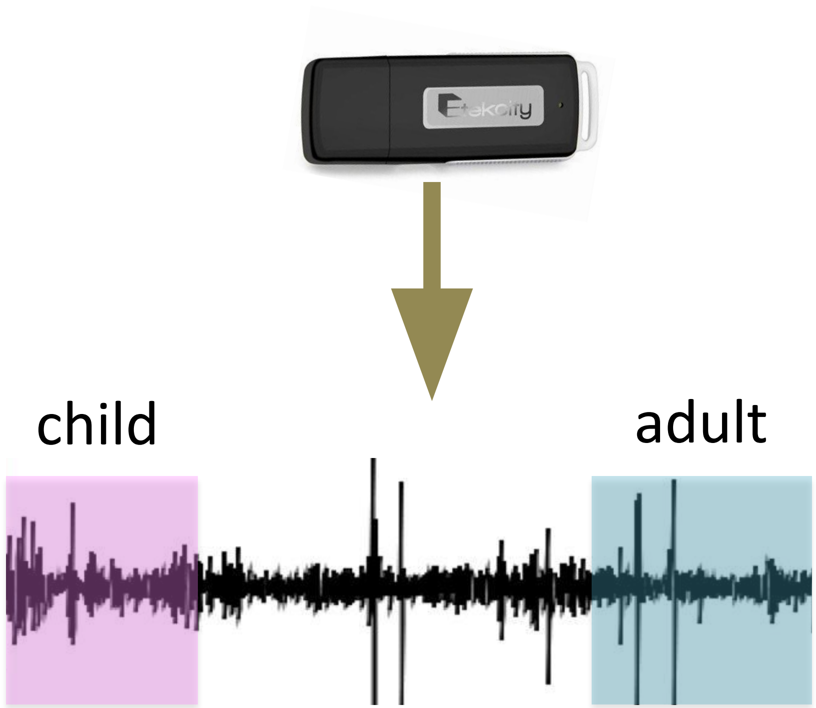
ComParE 2017 Interspeech

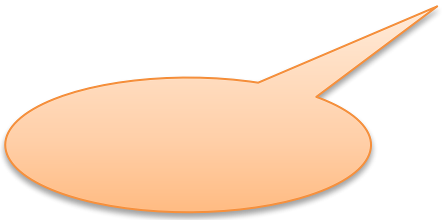
Child vocalization types
(babbling, crying, …)

ComParE 2019 Interspeech

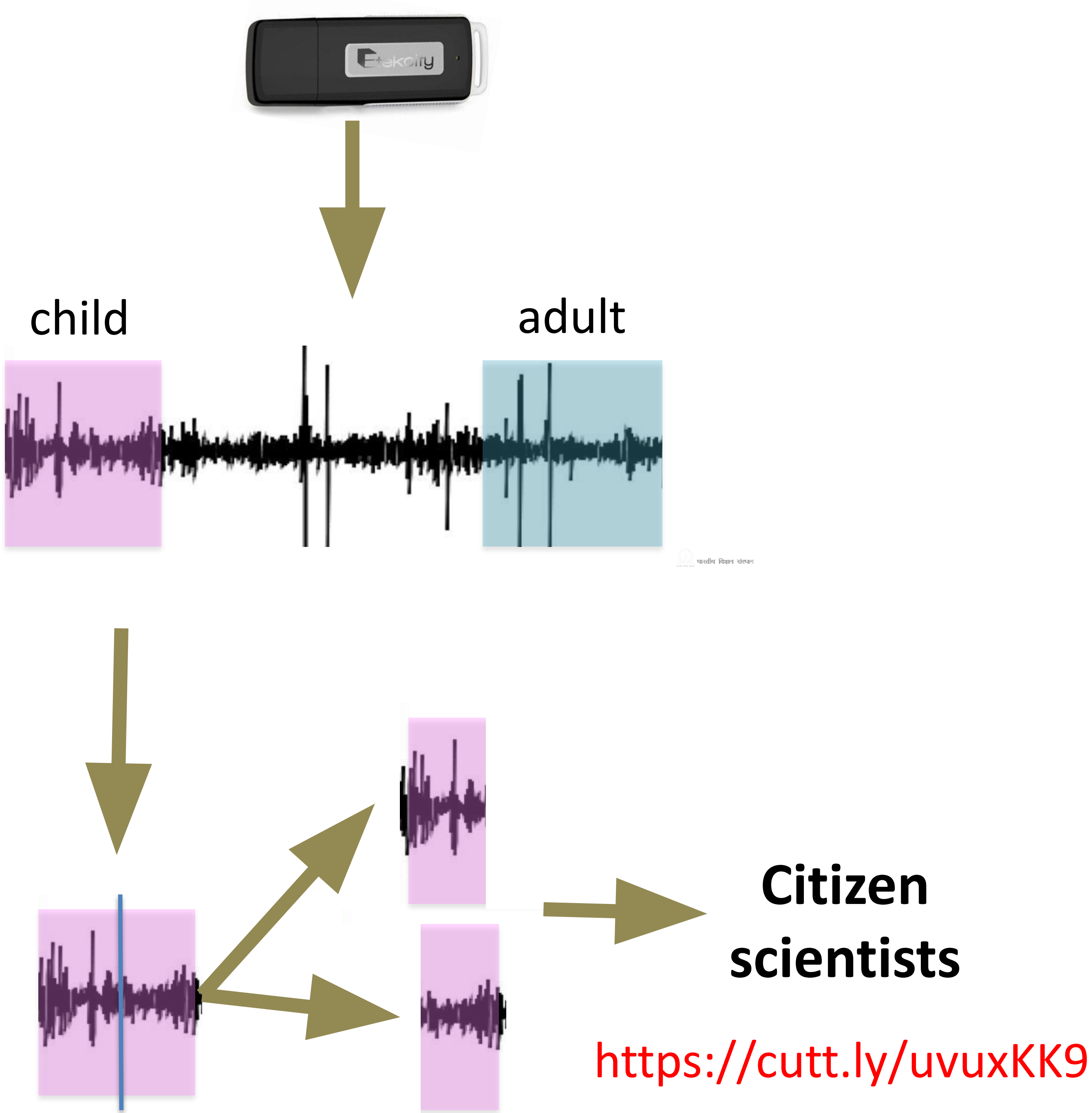
TO BE CONTINUED

NEEDED:
more work exploiting unsupervised, semi-supervised, and self-supervised classification

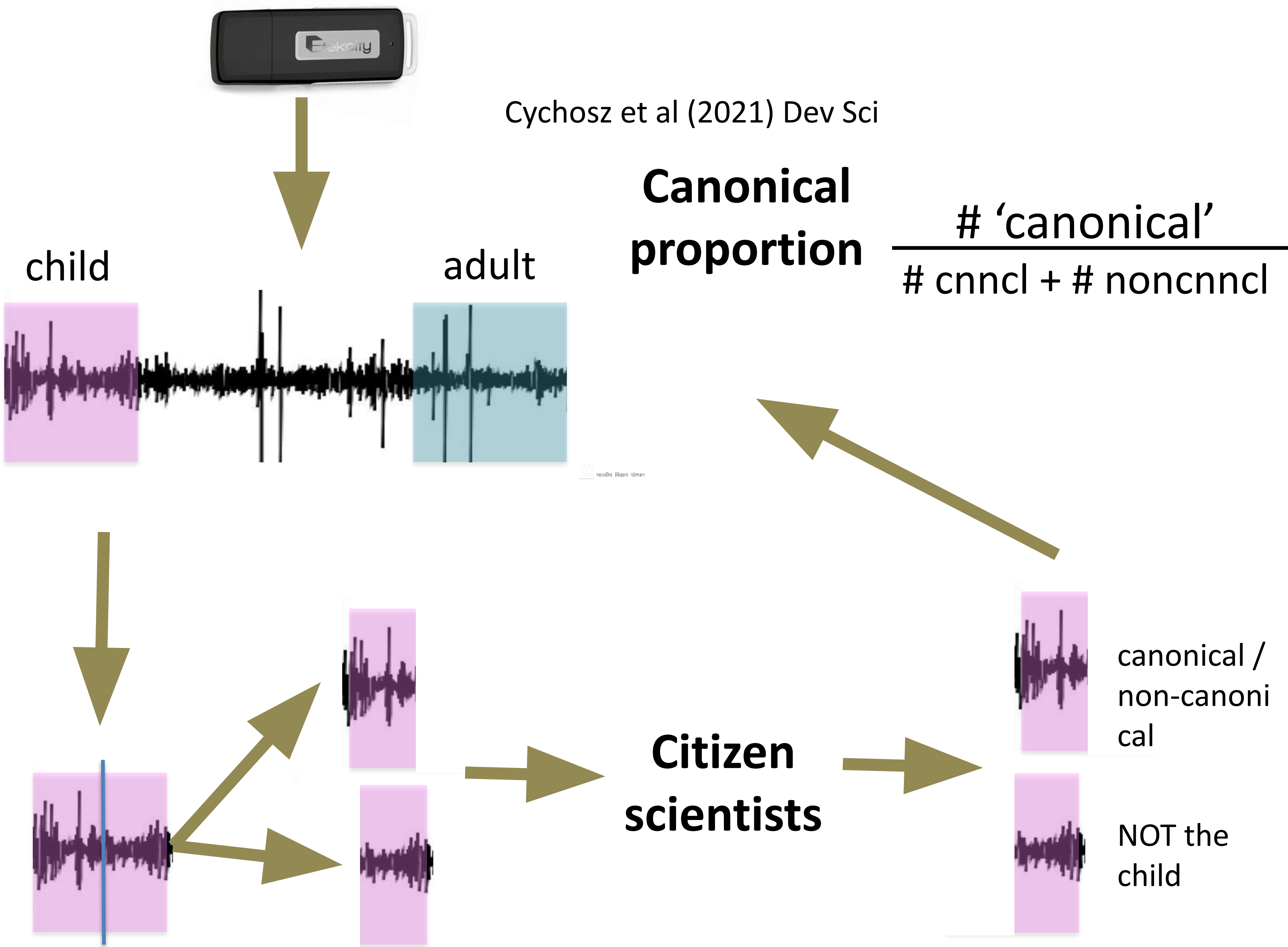


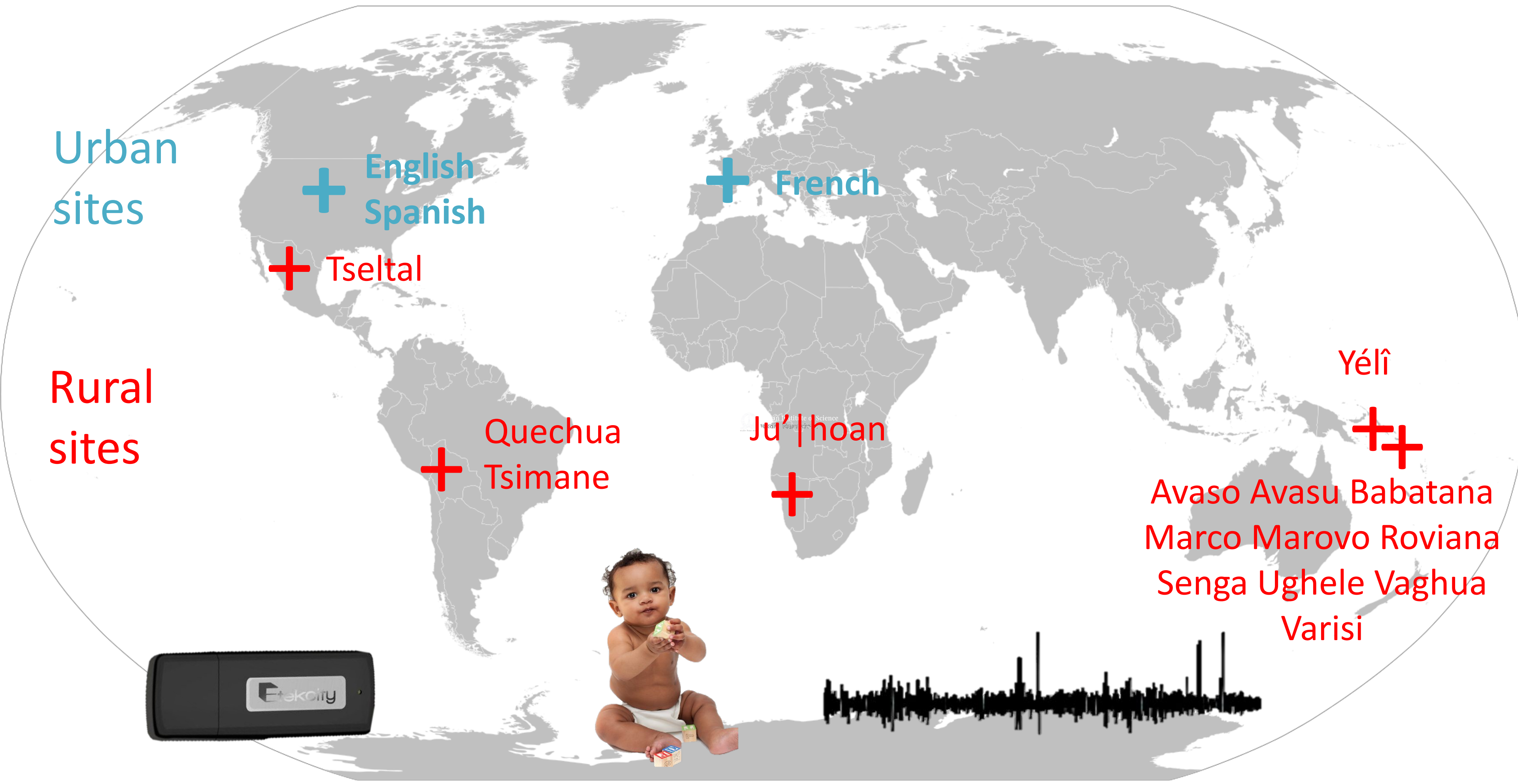


Long-form audio recordings + citizen scientists to the rescue!


child
adult
Citizen scientists
https://cutt.ly/uvuxKK9


ZOONIVERSE
child
adult
Canonical
Non-Canonical
Laughing
Crying
Junk
We ask you to classify the sound you will hear in one of these 5 categories. Select only one option.
Continue
Junk
NEED SOME HELP WITH THIS TASK?
Done & Talk
Done
FIELD GUIDE
SIGN IN REGISTER
FIELD GUIDE
Maturity of Baby Sounds
Canonical Sounds
Non Canonical Sounds
Laughing
Crying
Junk
0:00 / 0:00
TASK
Please classify this sound:
Canonical
Non-Canonical
Laughing
Crying
Junk
Done & Talk
Citizen scientists
https://cutt.ly/uvuxKK9

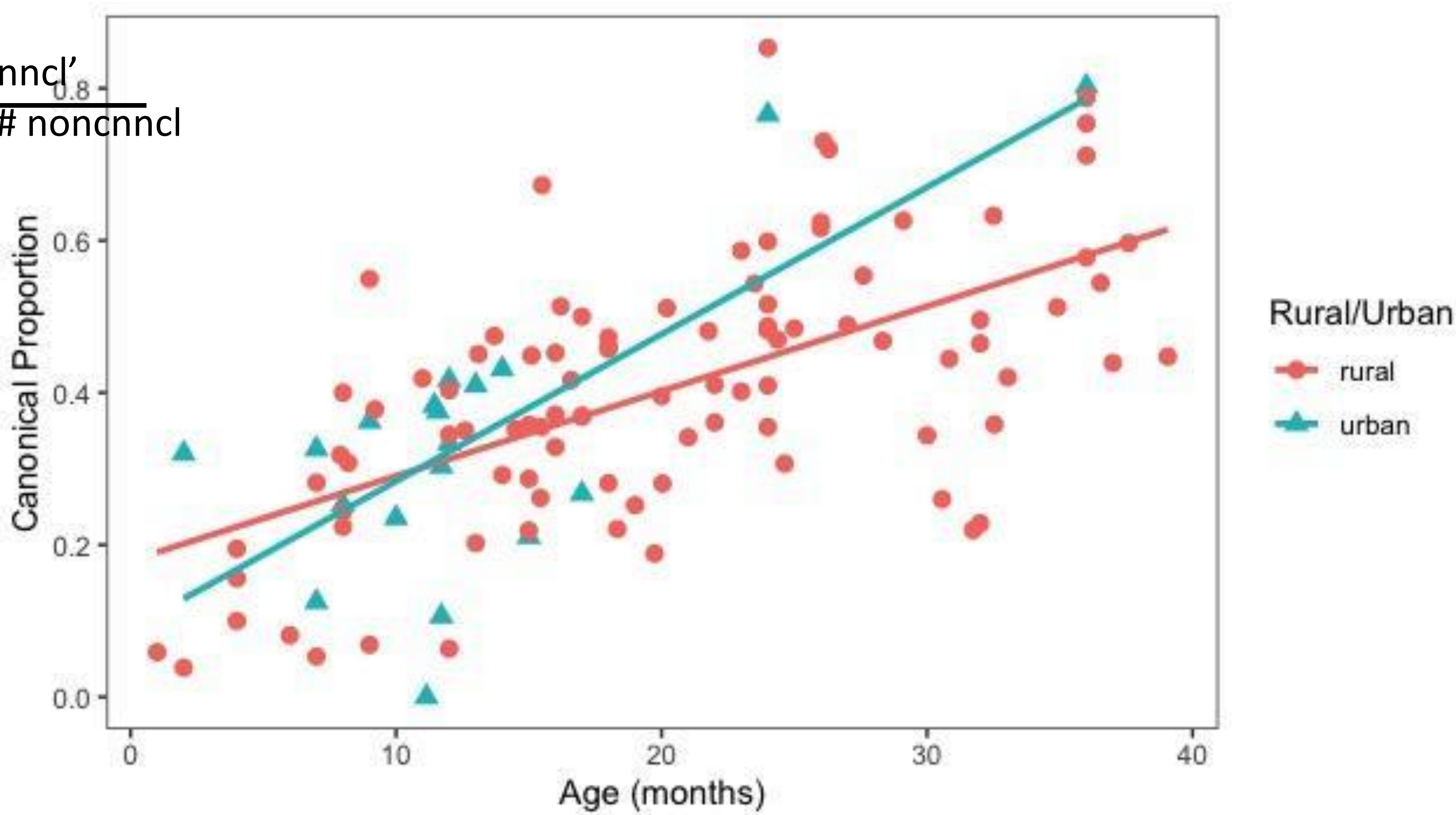

Cychosz et al (2021) Dev Sci
Canonical proportion
'canonical'
cnncl + # noncnncl
child
adult
Citizen scientists
canonical / non-canonical
NOT the child



19 children learning English, Spanish, or French in urban locations
95 learning one of 19 other languages in rural sites

Preliminary results

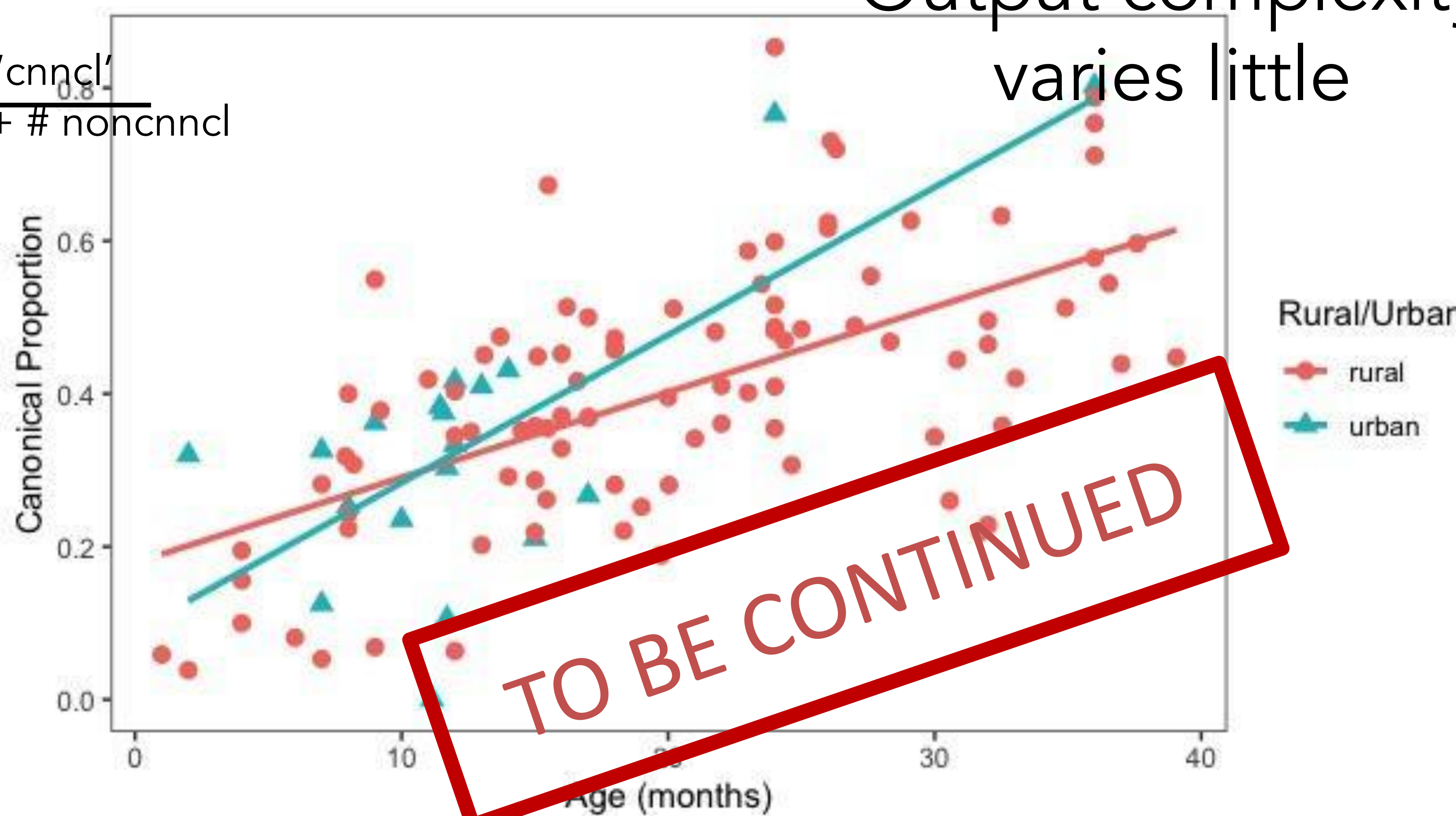
$$\frac{\#\ \text{‘cnncl’}}{\#\ \text{cnncl} + \#\ \text{noncnncl}}$$



Preliminary results

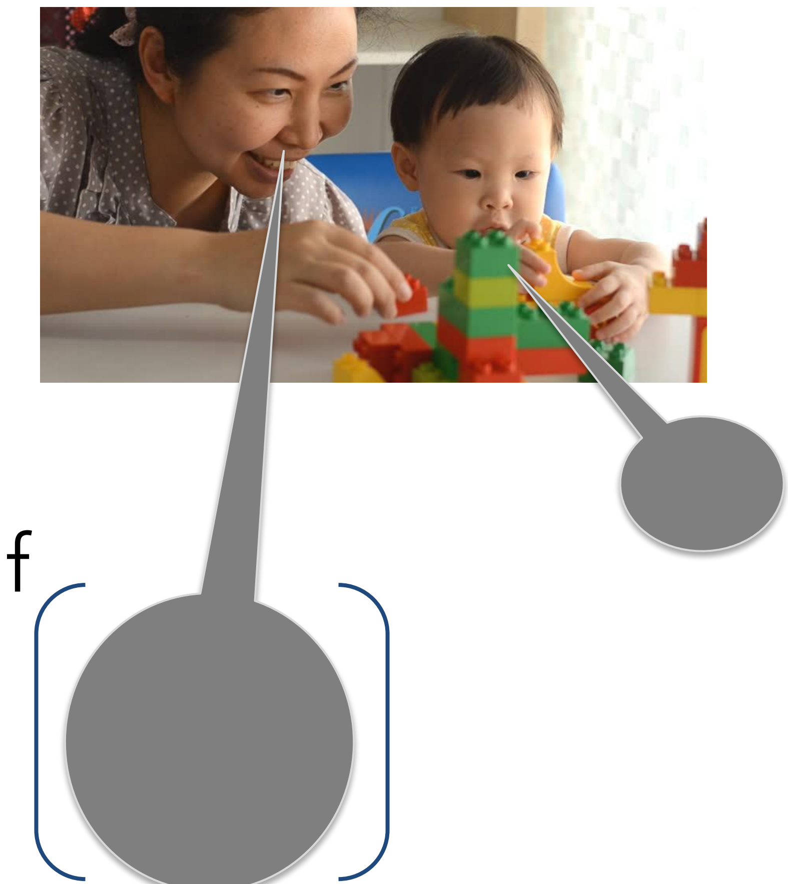
Output complexity varies little

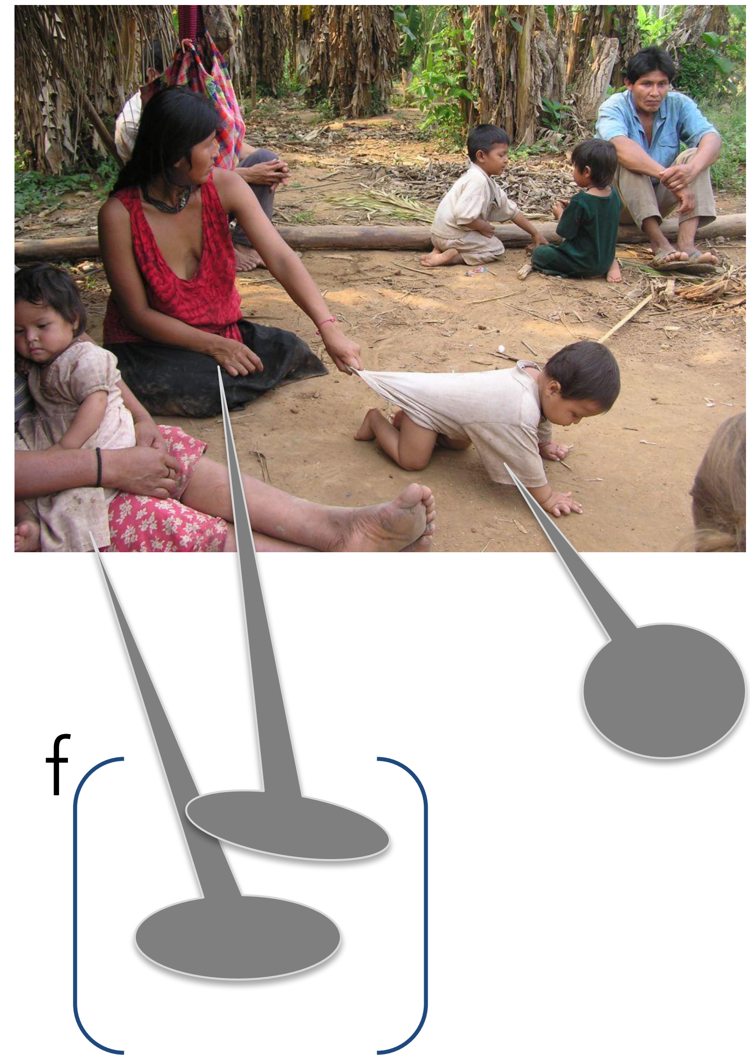
$$\frac{\#\ \text{'cnncl'}}{\#\ \text{cnncl} + \#\ \text{noncnncl}}$$



on average, fewer than 6 children per language/site

Assuming results hold, our broad language acquisition theory (v 2.1)





Assuming results hold, our broad language acquisition theory (v 2.1)

f

May infants learn from overheard speech?

f

Assuming results hold, our broad language acquisition theory (v 2.1)

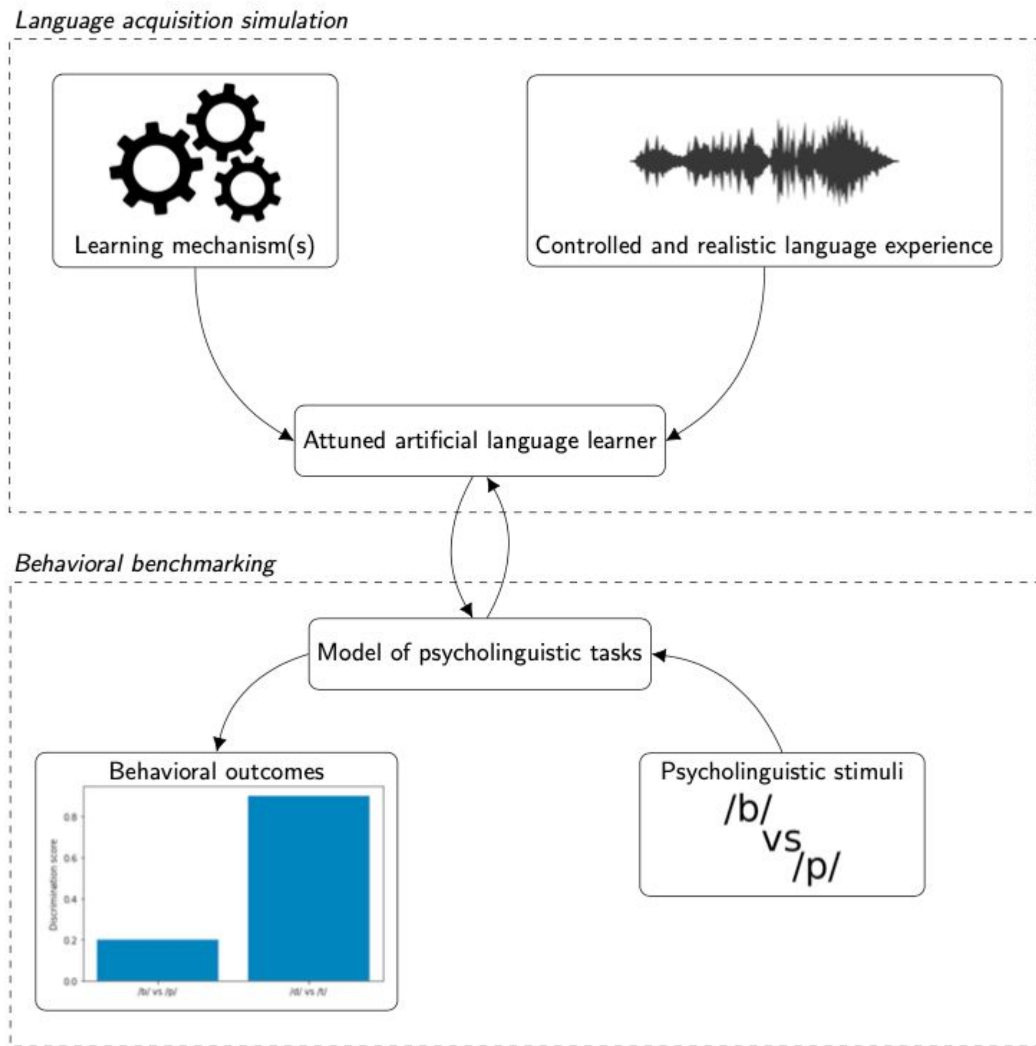
f

May infants learn from overheard speech?

Next step: Learnability properties

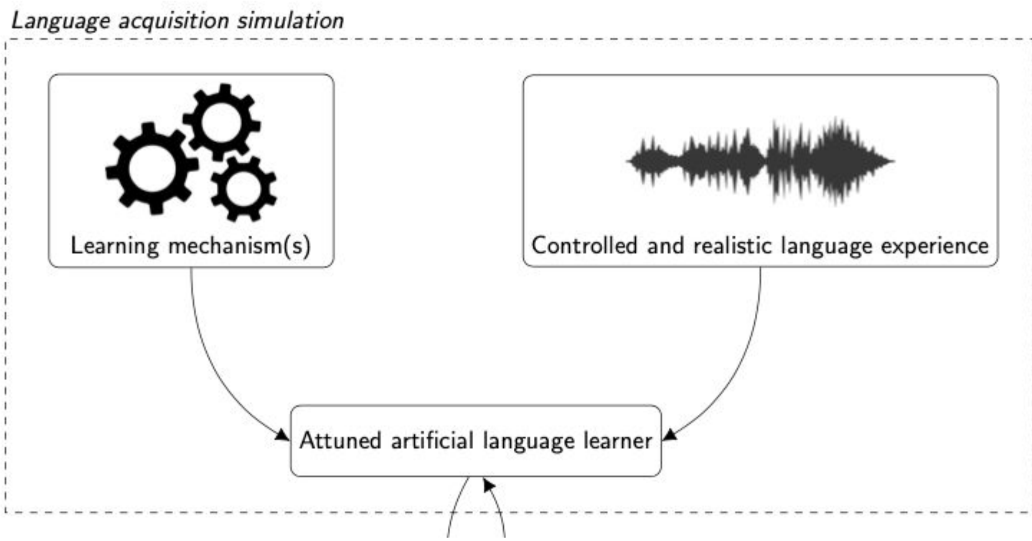
f

Reverse-engineering language acquisition: Our current proposal



Lavechin et al 2021 [preprint]

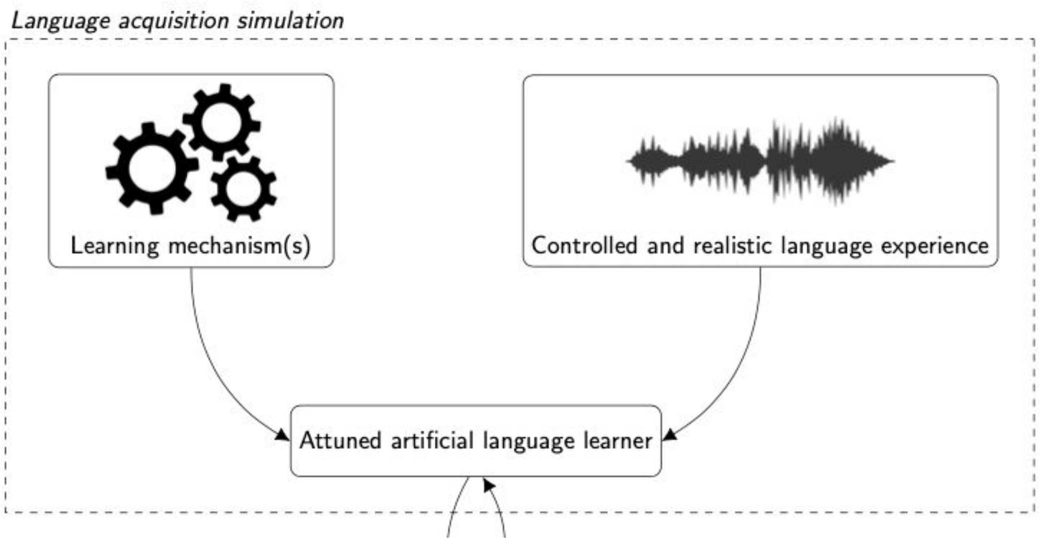
Simulating language acquisition



Lavechin et al 2021 [preprint]

Desiderata for the function

Unsupervised
Self-supervised
~~Plausible~~

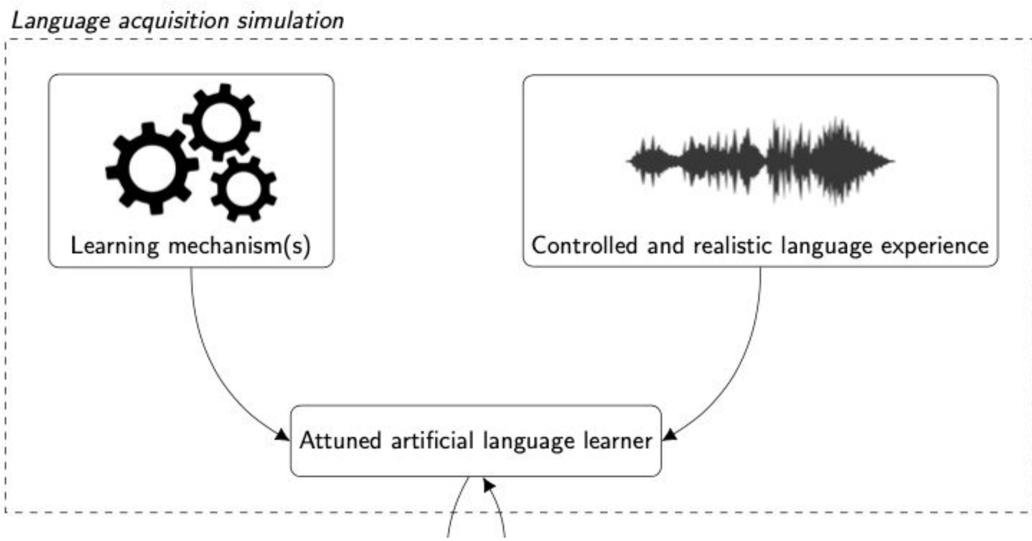


f

Lavechin et al 2021 [preprint]

Desiderata for the input

Unsupervised
Self-supervised
~~Plausible~~



Child-centered
Realistic
Controlled

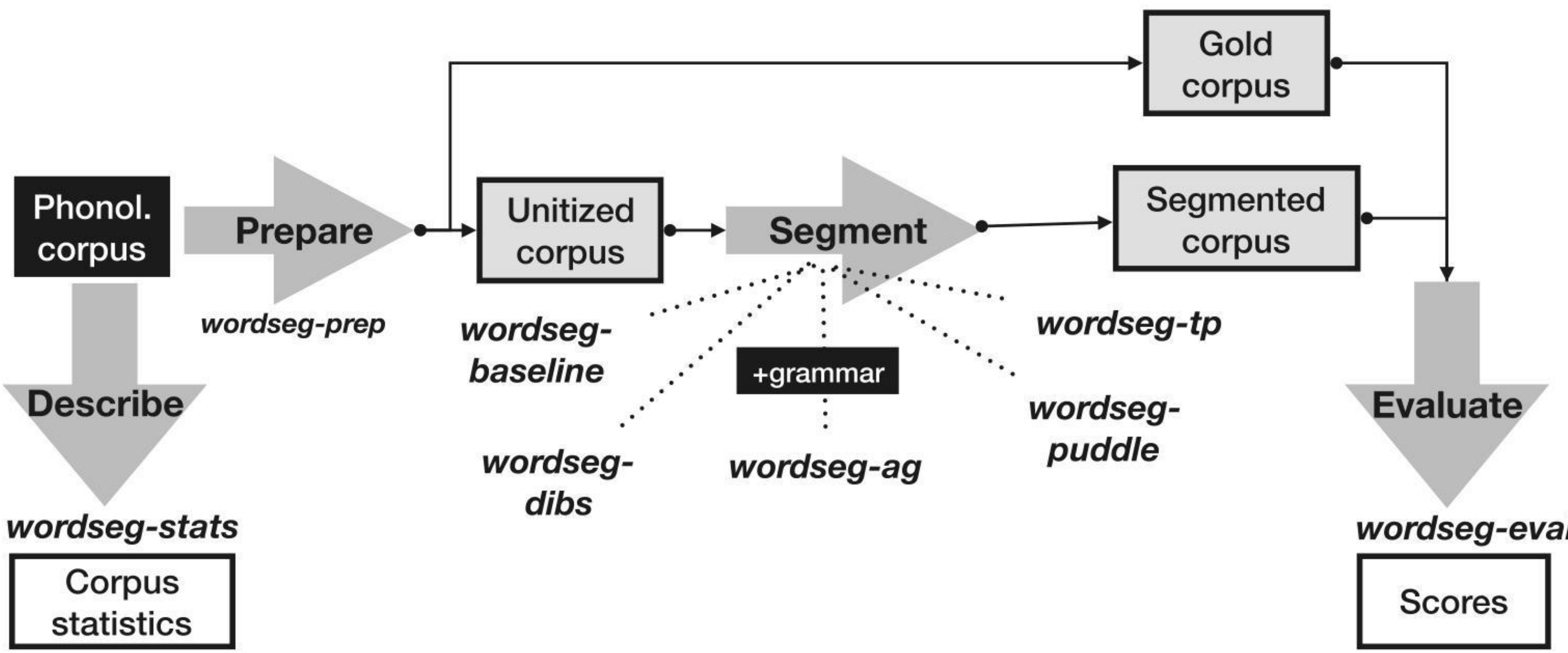
Lavechin et al 2021 [preprint]

Studying learnability properties: eg Unsupervised word segmentation

f

WordSeg Package

wordseg.readthedocs.io



Example algorithms

1. Baseline

Simplest strategies

- Every sentence is a word (**SentBase**)
- Every syllable is a word (**SyllBase**)

Lignos 2012

2. Sub-lexical

Goal is to "cut" using local cues

- Transitional Probabilities (TP)

x Absolute/Relative threshold

TP_abs
TP_rel

- Diphone-Based Segmentation **(DiBS)**

Daland + 2009; Saksida + 2016

3. Lexical

Goal is to learn a set of "minimal recombinable units"

- Adaptor Grammar **(AG)**
- Phonotactics from Utterances Determine Distributional Lexical Elements **(Puddle)**

Johnson + 2007; Monaghan + 2010

Bernard et al. 2019 Beh Res Meth (preprint)

Studying learnability properties: Unsupervised word segmentation

f( ) WordSeg Package

+

hibaby
areyouacutebaby?

Transcribed speech corpora

English may not be the best language to study learnability on…

English (and other contact/imperial languages)

Finish it, I'll be here!

He's dressed.

English may not be the best language to study learnability on…

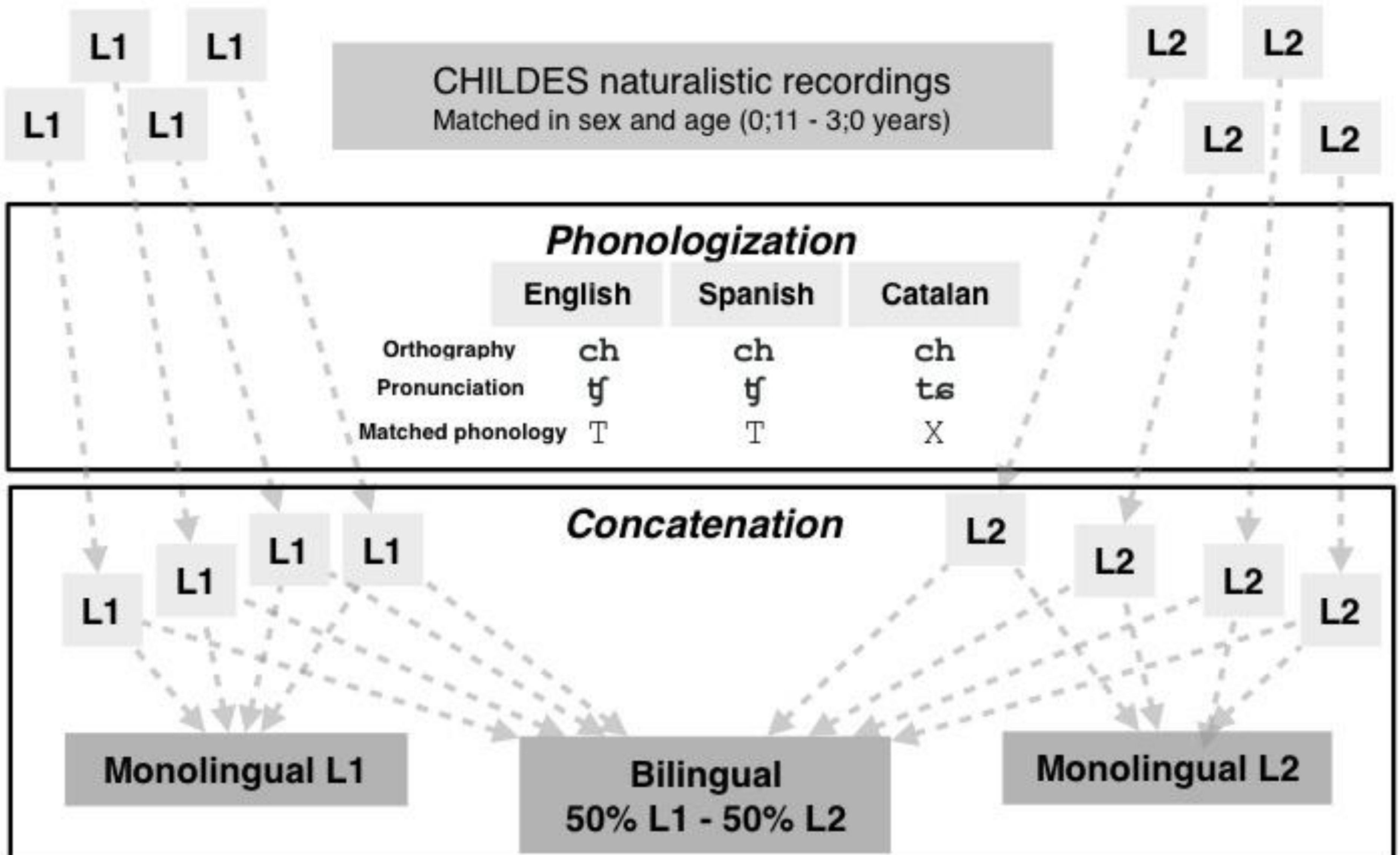
English (and other contact/imperial languages)

Inuktitut

Finish it, I'll be here! = Nungullugungai, taavanilangajualusunga!

He's dressed. = Annuraqsimajualuuman.

Creating bilingual corpora



Factors we manipulated

f

Different processing algorithms

Different languages

Monolingual versus bilingual input

Which factor had the biggest impact on performance? Guess in chat!

Different processing algorithms

f

ALGO

LANG

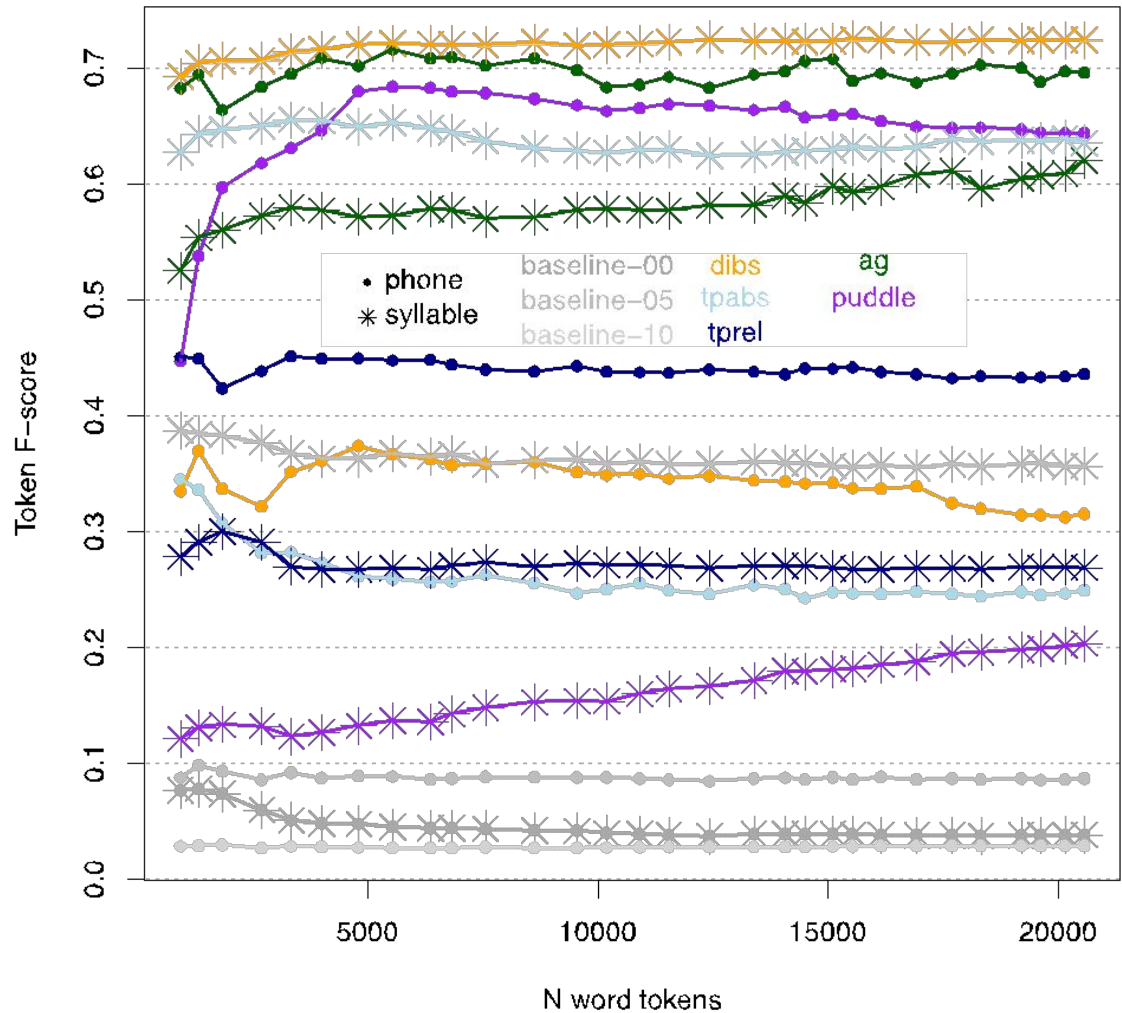
Different languages

MONO

Monolingual versus bilingual input

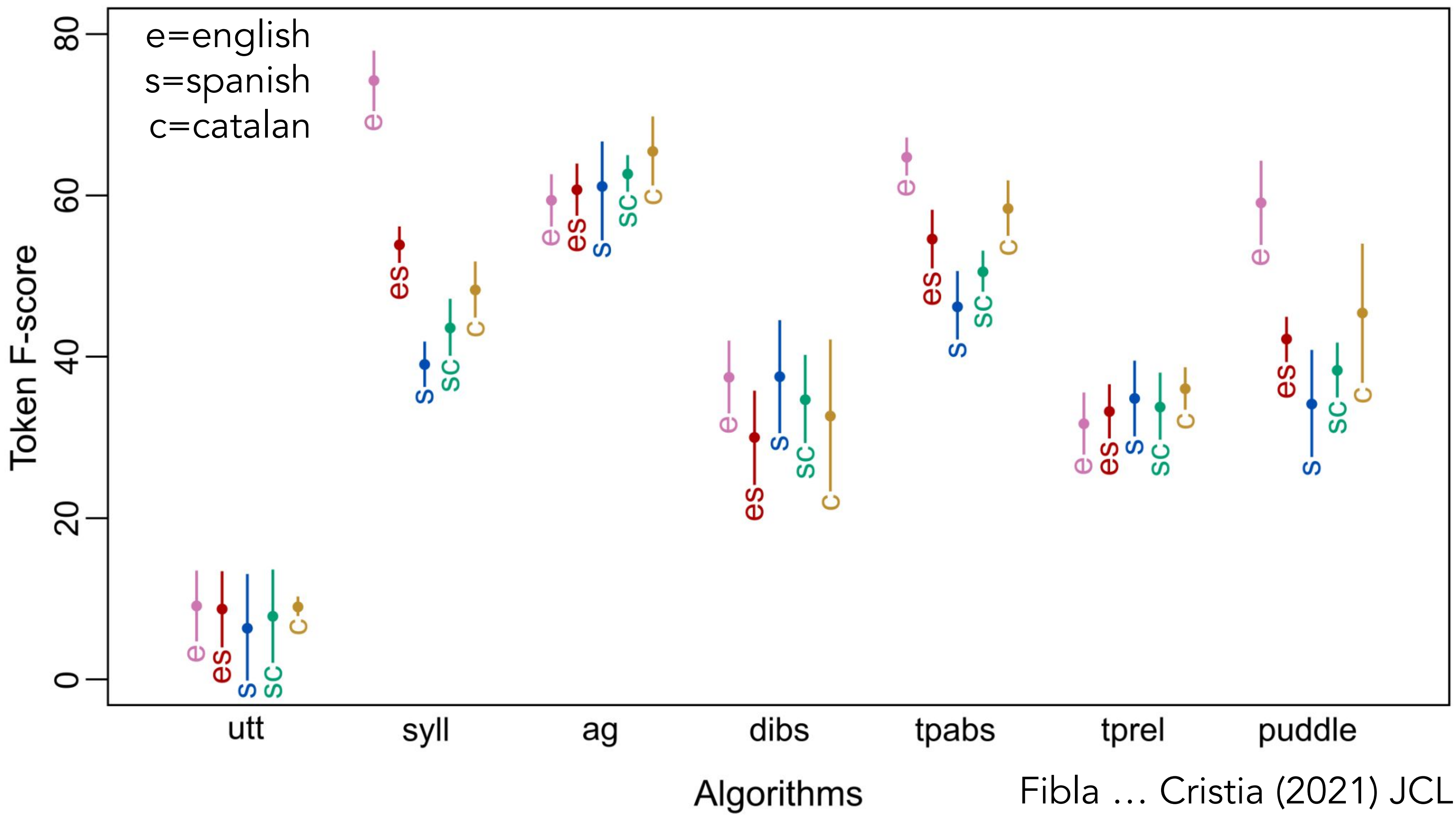
Differences between learning algorithms are enormous (40-60%)

f



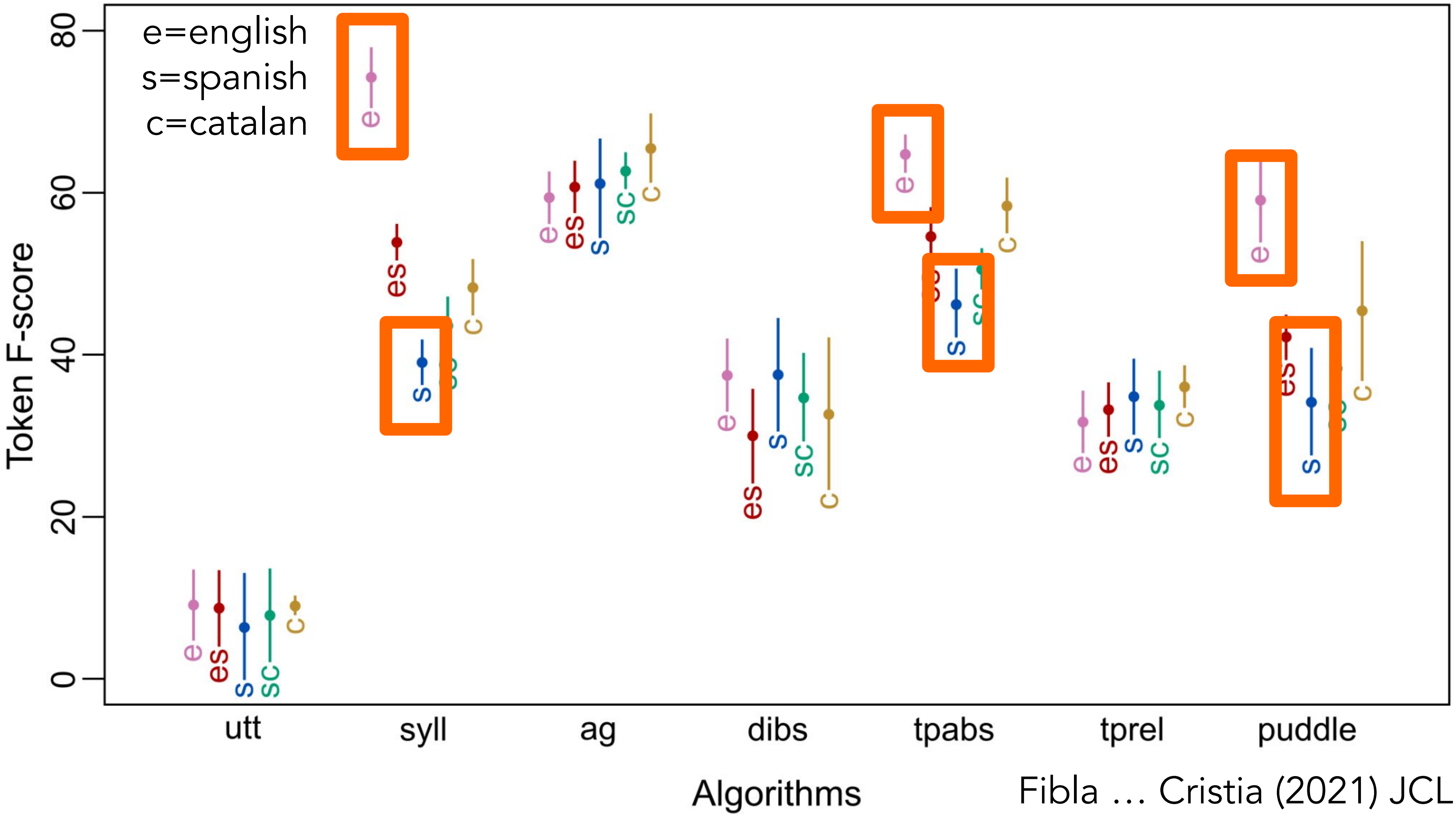
Mathieu … Cristia (2019) Beh Res Methods

Differences bet/ languages? Monolingual advantage?



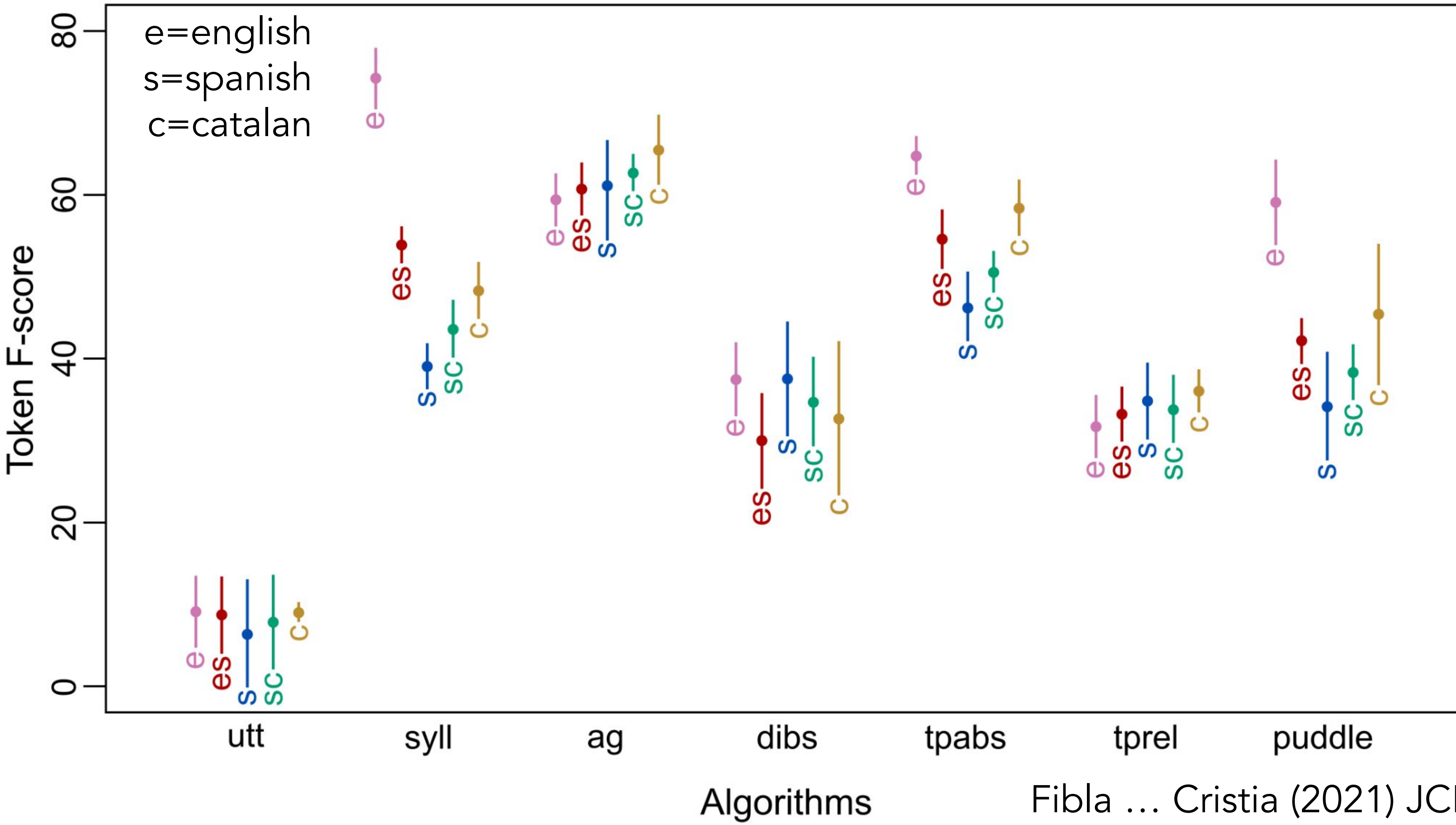
Fibla … Cristia (2021) JCL

Smaller differences bet/ languages



Fibla … Cristia (2021) JCL

Smaller differences bet/ languages No clear monolingual advantage



Fibla … Cristia (2021) JCL

Results so far

f

Differences between learning algorithms are- enormous (40-60%)

> than that between languages as a function of languages by morphological type (20%)

- Monolingual versus bilingual input (<5%)

Mathieu … Cristia (2019) Beh Res Methods

Loukatou … Cristia (2019) ACL
Fibla … Cristia (2021) JCL

Results so far

f

Differences between learning algorithms are enormous (40-60%)

> than that between languages as a function of languages by morphological type (20%)

- Monolingual versus bilingual input (<5%)

TO BE CONTINUED

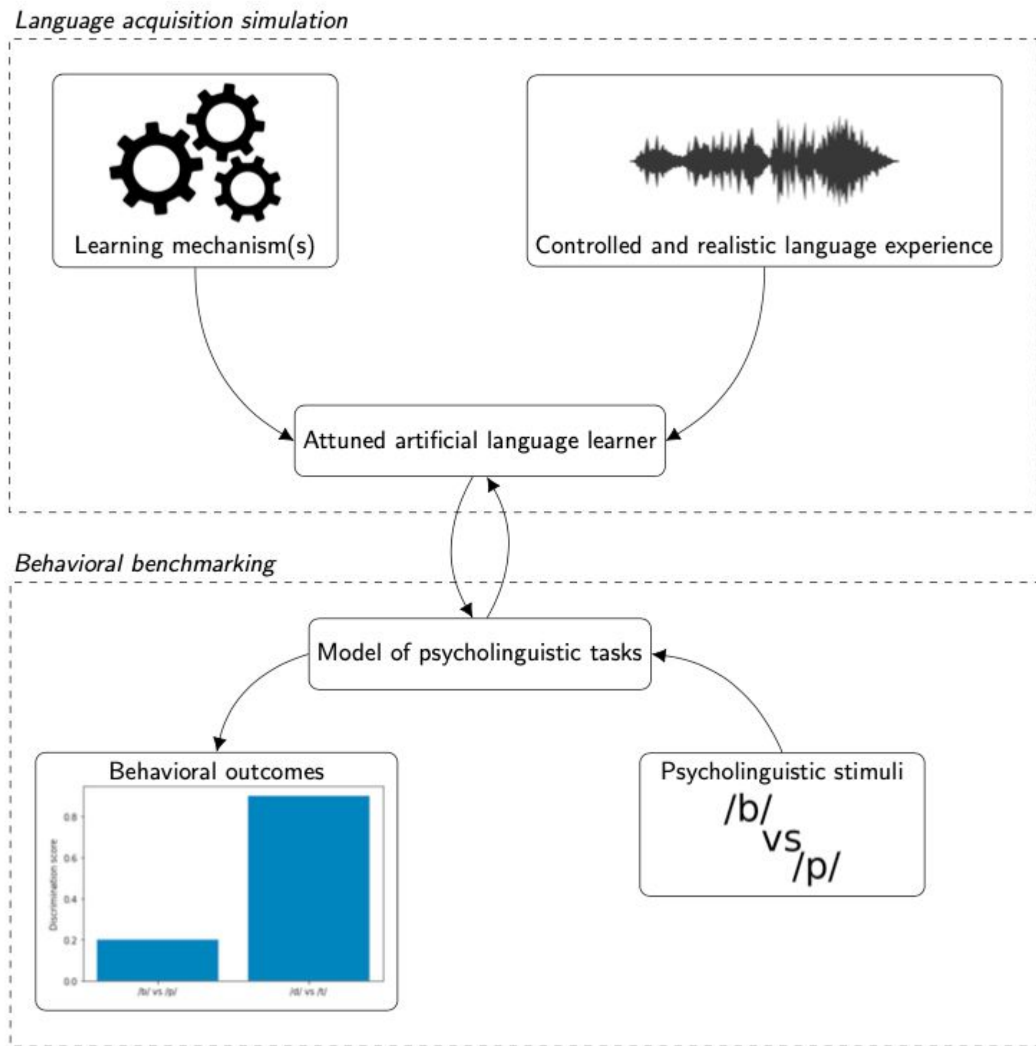
Mathieu … Cristia (2019) Beh Res Methods

Loukatou … Cristia (2019) ACL
Fibla … Cristia (2021) JCL

NEEDED:
- learnability on other levels;
- *real infant evidence*

Behavioral benchmarking

Unsupervised
Self-supervised
~~Plausible~~



Child-centered
Realistic
Controlled

Behavioral correlates that can be realistically measured at scale on humans & machines

Lavechin et al 2021 [preprint]

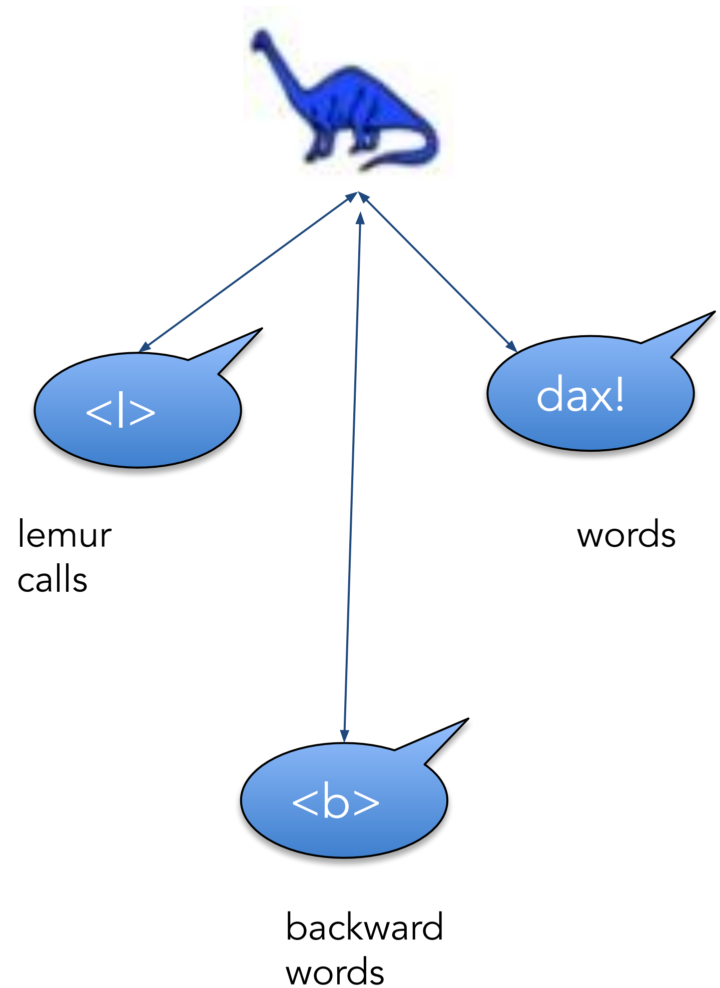
Example: categorization task with words

Perszyk & Waxman 2017 [JOVE]

Behavioral correlates in humans & machines

Sound only behaviors	Age (mo)	Task	Dataset
discriminate across rhythmically distinct languages	0	distance-based	bilingual set of stimuli
discriminate across rhythmically similar languages only if exposed to one of them	0	distance-based	bilingual set of stimuli
discriminate native and non-native consonants	6-8	distance-based	phonetically aligned clean speech
accept novel content words more easily than novel function words	6	probability-based	jabberwocky sentences
prefer native over non-native phonotactics	9	probability-based	made-up words varying in phonotactics
prefer high over low phonotactics	9	probability-based	made-up words varying in phonotactics
prefer high over low frequency content words	11	probability-based	real words varying in frequency
do not discriminate non-native consonants	12	distance-based	phonetically aligned clean speech

Cross-modal behaviors	Age (mo)	Task	Dataset
treat words and monkey calls, but not beeps or coughs, as possible labels	3	few-shot learning + distance-based	images paired with words, monkey calls, beeps or
treat words but not monkey calls as possible labels	6	few-shot learning + distance-based	images paired with words or monkey calls
treat content but not function words as possible labels	6	few-shot learning + distance-based	images paired with function words or content words
few-shot learning of new word-object pairings	9	few-shot learning + distance-based	images paired with words
treat words with native but not non-native sounds as possible labels	10	few-shot learning + distance-based	images paired with L1 words and L2 words



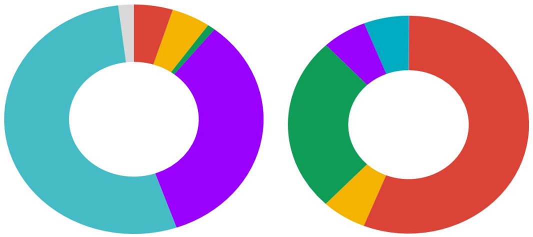
Lavechin et al 2021 [preprint]

An interdisciplinary endeavor

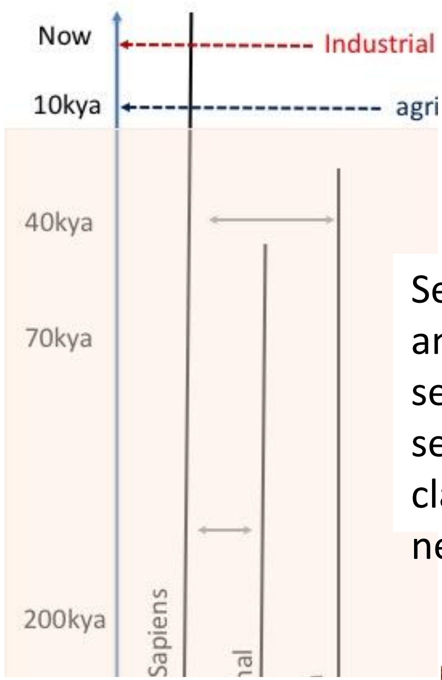
<table>
<tr><th></th><th>Algorithms</th><th>Input Data</th><th>Outcome measures</th><th>Integration</th></tr>
<tr><td>Corpus Analysis</td><td></td><td>Estimate prevalence of the various referential and event types</td><td>Measures of language output maturity</td><td rowspan="3">Explanations of outcome/input relationships in infants across cultures

Predictions of outcomes of interventions</td></tr>
<tr><td>Computer Modeling</td><td>Implementation of probabilistic models, learning and preprocessing algorithms</td><td colspan="2">Estimate of outcomes as a function of prevalence of referential/event types in the input for each combination of algorithm and preprocessing</td></tr>
<tr><td>Experimental Studies</td><td>Proof-of-concept of preprocessing and learning algorithms</td><td></td><td>Measure of tacit knowledge (probabilistic models of infants)</td></tr>
</table>

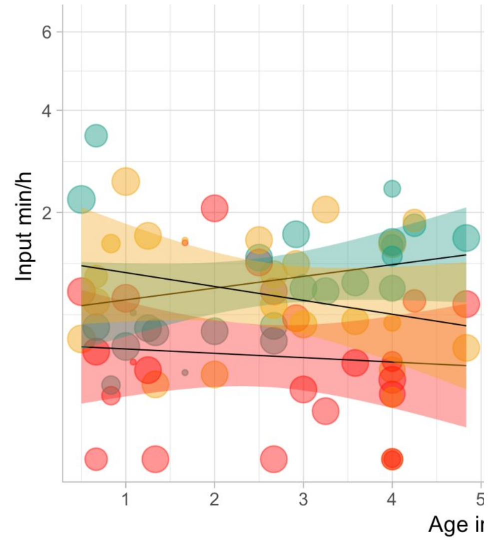
Tsuji et al. 2021 Cognition ([pdf](pdf))



All extant datasets are biased

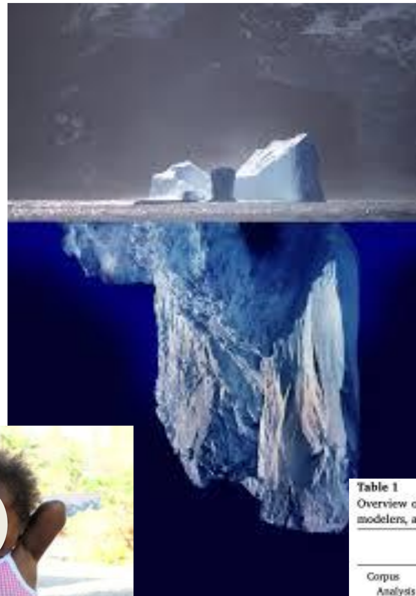


Humans evolved in a setting crucially different from that represented in those data

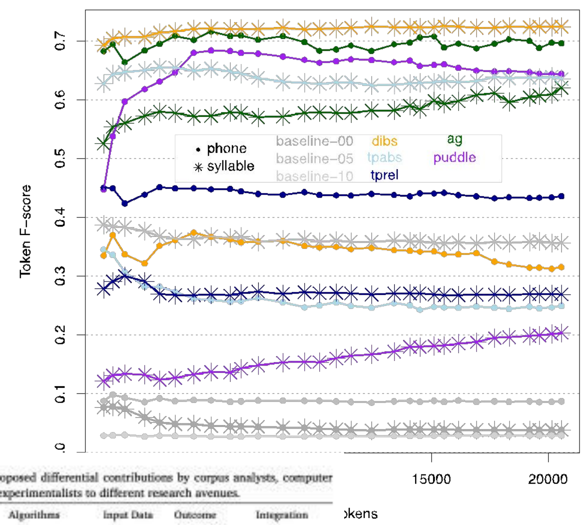


… suggest some children succeed with little directed input from adults

Semi-, un-, and self-supervised classifiers needed!



Studying learnability properties using artificial agents



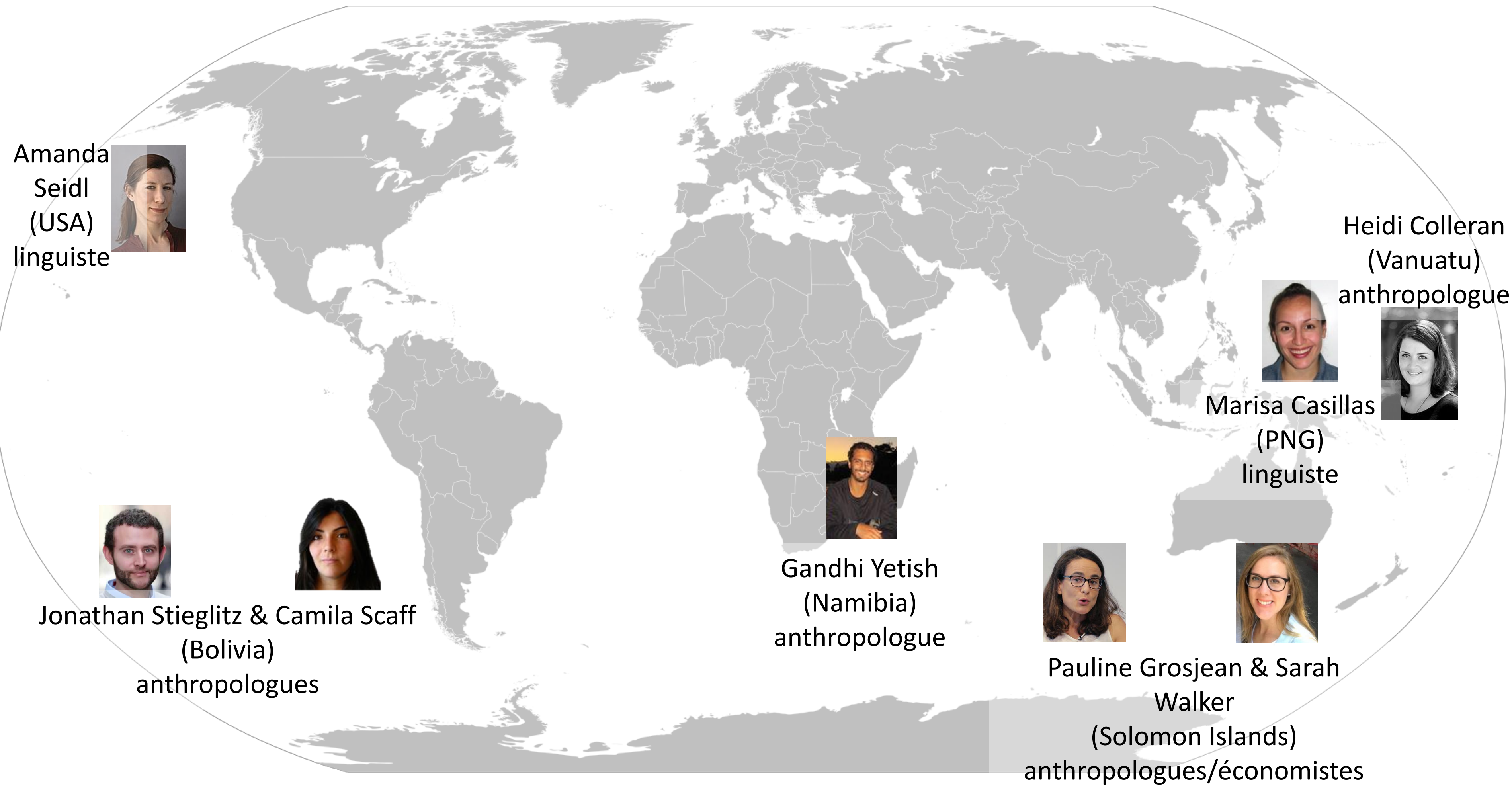
Naturalistic, massive datasets of child language…

Table 1
Overview of proposed differential contributions by corpus analysts, computer modelers, and experimentalists to different research avenues.

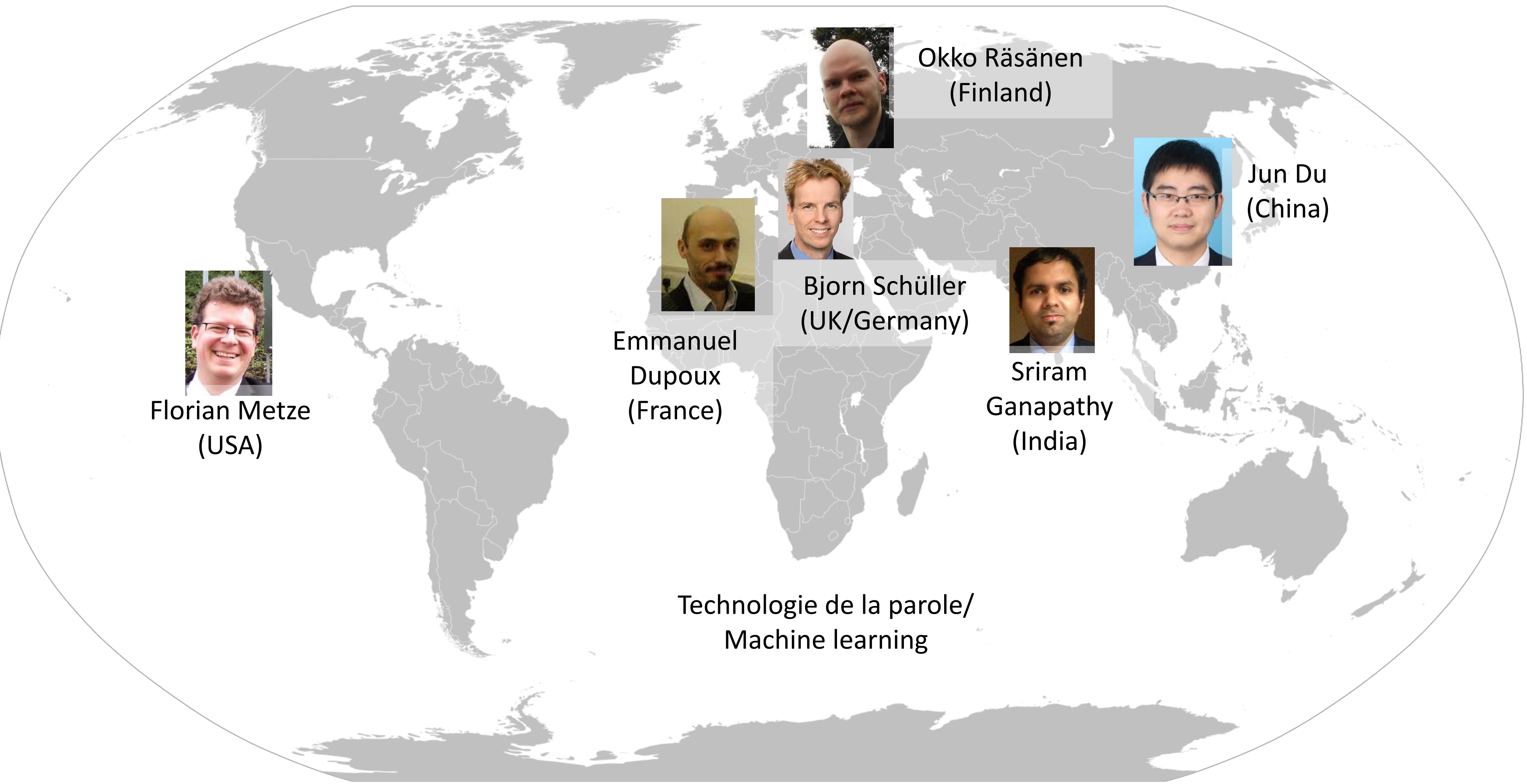
	Algorithms	Input Data	Outcome measures	Integration
Corpus Analysis		Estimate prevalence of the various referential and event types	Measures of language output maturity	Explanations of outcome/ input relationships in infants across cultures
Computer Modeling	Implementation of probabilistic models, learning and preprocessing algorithms	Estimate of outcomes as a function of prevalence of referential/event types in the input for each combination of algorithm and preprocessing		Predictions of outcomes of interventions
Experimental Studies	Proof-of-concept of preprocessing and learning algorithms		Measure of tacit knowledge (probabilistic models of infants)	

Solving this puzzle requires interdisciplinary research

If you want to go fast,
go alone.
If you want to go far,
go together



Amanda Seidl (USA) linguiste
Heidi Colleran (Vanuatu) anthropologue
Marisa Casillas (PNG) linguiste
Gandhi Yetish (Namibia) anthropologue
Jonathan Stieglitz & Camila Scaff (Bolivia) anthropologues
Pauline Grosjean & Sarah Walker (Solomon Islands) anthropologues/économistes


Okko Räsänen
(Finland)
Jun Du
(China)
Bjorn Schüller
(UK/Germany)
Emmanuel
Dupoux
(France)
Sriram
Ganapathy
(India)
Florian Metze
(USA)
Technologie de la parole/
Machine learning

Affiliated researchers

Camila Scaff
(PhD Cog Sci)
U Zurich

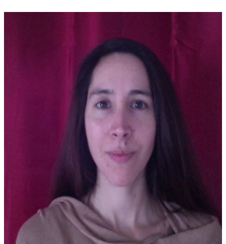
Sho Tsuji
(PhD Cog Sci)
U Tokyo

Marvin Lavechin
Machine learning
PhD student
(CIFR Facebook Artificial Intelligence Research)

Interns (summer 2021):

- Marina Drobi (Cogmaster, PMI)
- Chloé Magnier & Cédric Dubreil (SLP)
- Ninoh Da Silva (Linguistic informatics)
- Martin Frébourg (speech tech intern)

We'll be hiring!
(2021-2023)
see exelang.fr
for more info

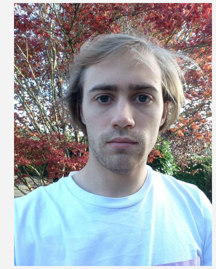


Alex Cristia
(PhD Linguistics)

Kasia Hitczenko
(PhD Linguistics)

William Havard
(PhD NLP)

Tech personnel



Lucas Gautheron
M1 Physics
Data Manager

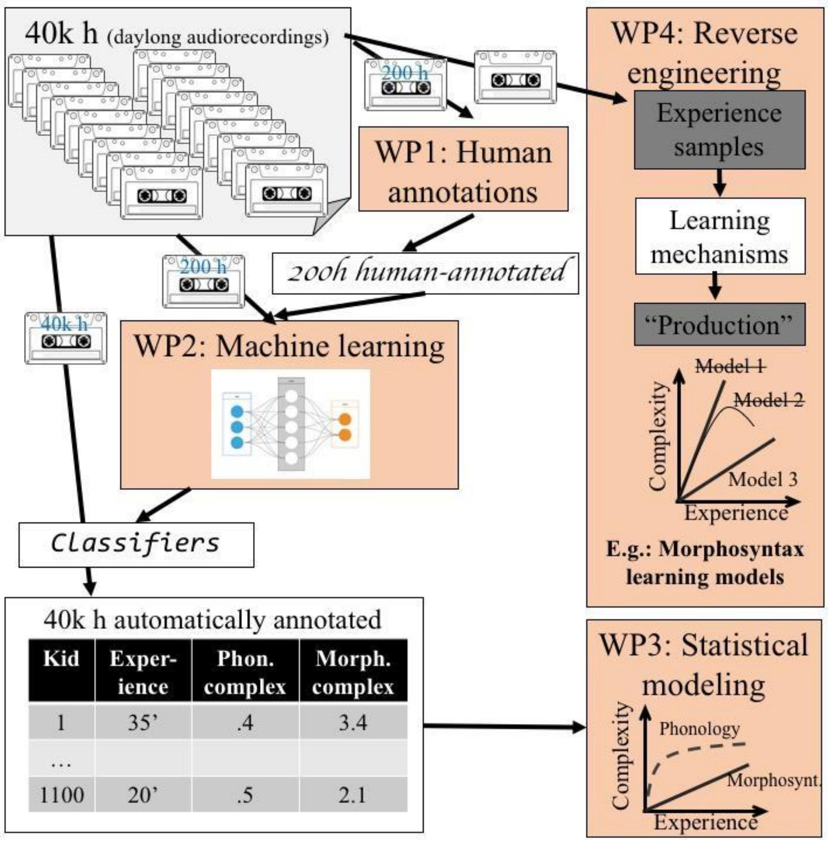
Sara Pisani
M1 Cultural Industries
Data donor advisor

Shared with Cognitive Machine Learning (CoML, INRIA)

Xuan Nga Cao
(PhD Linguistics)
Research Engineer

Catherine Urban
Admin Magician

ExELang.fr: Experience Effects on Language



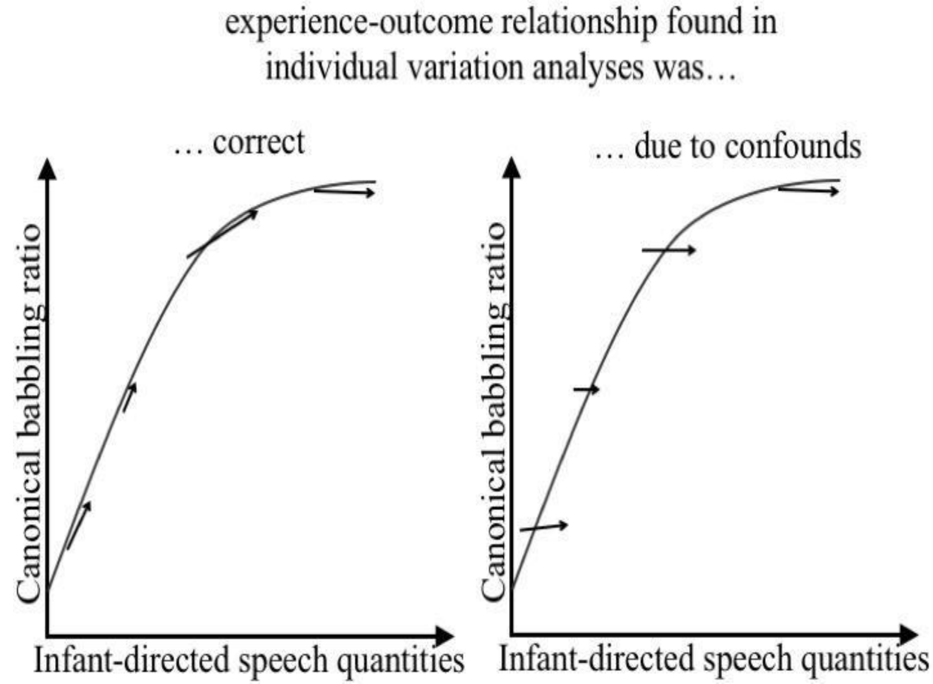
New <u>approach</u>:

Developing unsupervised language-learning models to reverse-engineer human learning

European Research Council
Established by the European Commission

ExELang.fr: Experience Effects on Language

New <u>data sets</u>:
"micro-grants"
Re-using data from randomized control trials



A potential result of predicting pre-post-intervention changes in the Randomized Control Trials' corpora. Each arrow represents data from one Randomized Control Trial (beginning of the arrow = "pre-intervention" quantities, tip =post-intervention quantities).

European Research Council
Established by the European Commission

Thanks to:

Participating families

Participating villages

Team, collaborators & colleagues

Funding agencies

And you.

alecristia@gmail.com

www.acristia.org

James S. McDonnell Foundation

ANR

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

Documentation on the systematic review
xcult.shinyapps.io/vocsr/

Sample daylong recording
https://github.com/LAAC-LSCP/vandam-daylong-demo

Zooniverse project (complete!)
https://cutt.ly/uvuxKK9

Annotation tools
sites.google.com/view/aclewdid
(*Annotations* & *Tools* tabs)

ExELang project
https://exelang.fr

Child-rearing among hunter-gatherer communities

- **Universal**
- Co-sleeping & physical contact
- Maternal primacy <1y
- Multi-age groups >1y
- Frequent breast-feeding

- **Variation**
- Non-maternal care
- Self-provisioning
- Assigned chores
- Father involvement
- Weaning age/ inter-birth interval duration

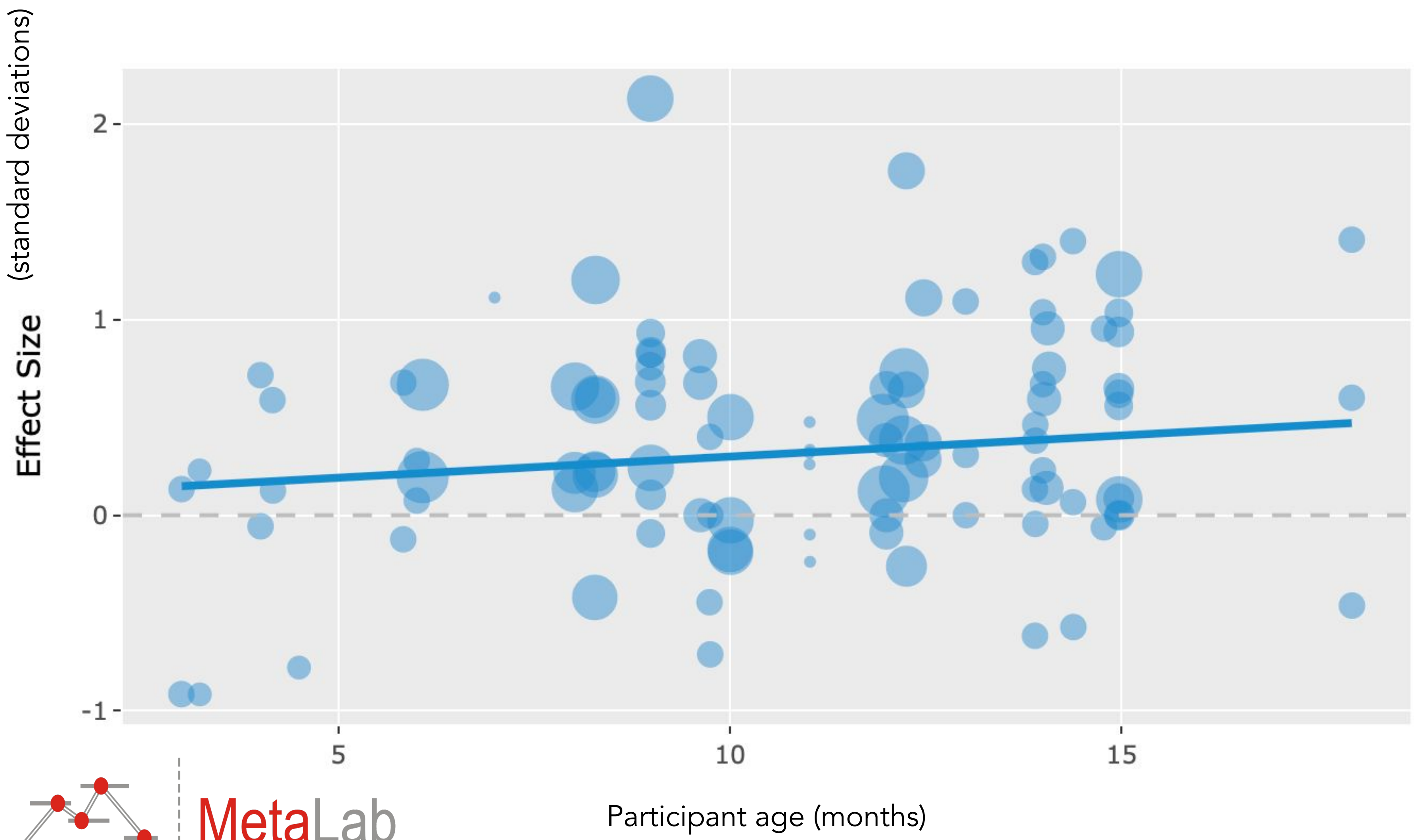
Variation in reproductive strategies

e.g. in number of children

Konner 2016

Hewlett et al. 2000

The noisy reality of infant studies



MetaLab