Language Learning: A Data-Driven Approach

Day 2: Examining transcripts with CHILDES and childes-db

Michael C. Frank

LOT Winter School

CHILDES

Child Language Data Exchange System

CHILDES is the child language component of the ***TalkBank*** system.
TalkBank is a system for sharing and studying conversational interactions.

System

*****Ground Rules*****

Contributing New Data

IRB Principles

Overviews and Introductions

Other Child Language sites

Research based on CHILDES

Child Language Diaries

Database

*****Index to Corpora*****

Browsable Database

LuCiD Toolkit

childes-db

CLAN

XML creator and ***XML Schema***

Related Software

Manuals

CHAT - ***CLAN*** - ***MOR***

Tutorial Screencasts

SLP's Guide to CLAN and ***中文***

Brian MacWhinney : ***homepage***

How to subscribe to ***Mailing Lists***

Phonology and Fonts

Phon and PhonBank

Unicode and IPA for ***Mac***

Unicode and IPA for ***Windows***

Teaching

Topics in Language Acquisition

Teaching Resources

YouTube Examples

Bibliographies

Morphology and Lexicon

Part of Speech Analysis by MOR

MRC lexical dictionary

ChildFREQ ***Site*** and ***Paper***

Special Procedures

CA analysis

Digitized video

Digitized audio

Versions

Derived Corpora and Counts

XML version of the database

Database Versioning

More Resources

Building a New Corpus

CCT Computerized Comprehension

LEAT Assessment Tool

The original “big data” for child language

Outline

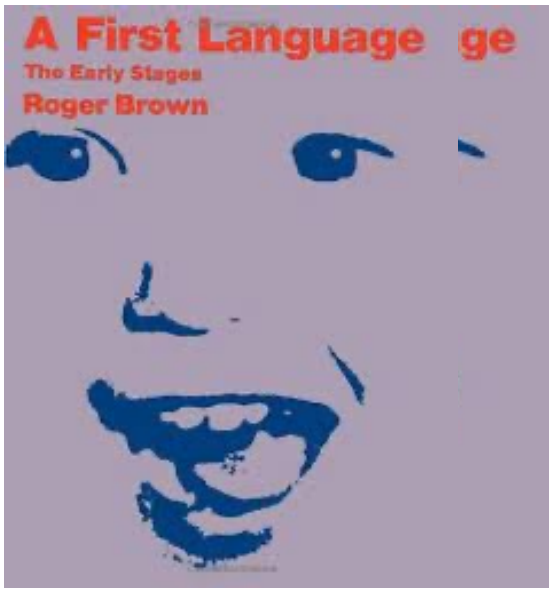
1. Introducing CHILDES
2. A case study in grammatical productivity
3. childes-db as a response to challenges of reproducibility
4. Examining function word development using childes-db

Outline

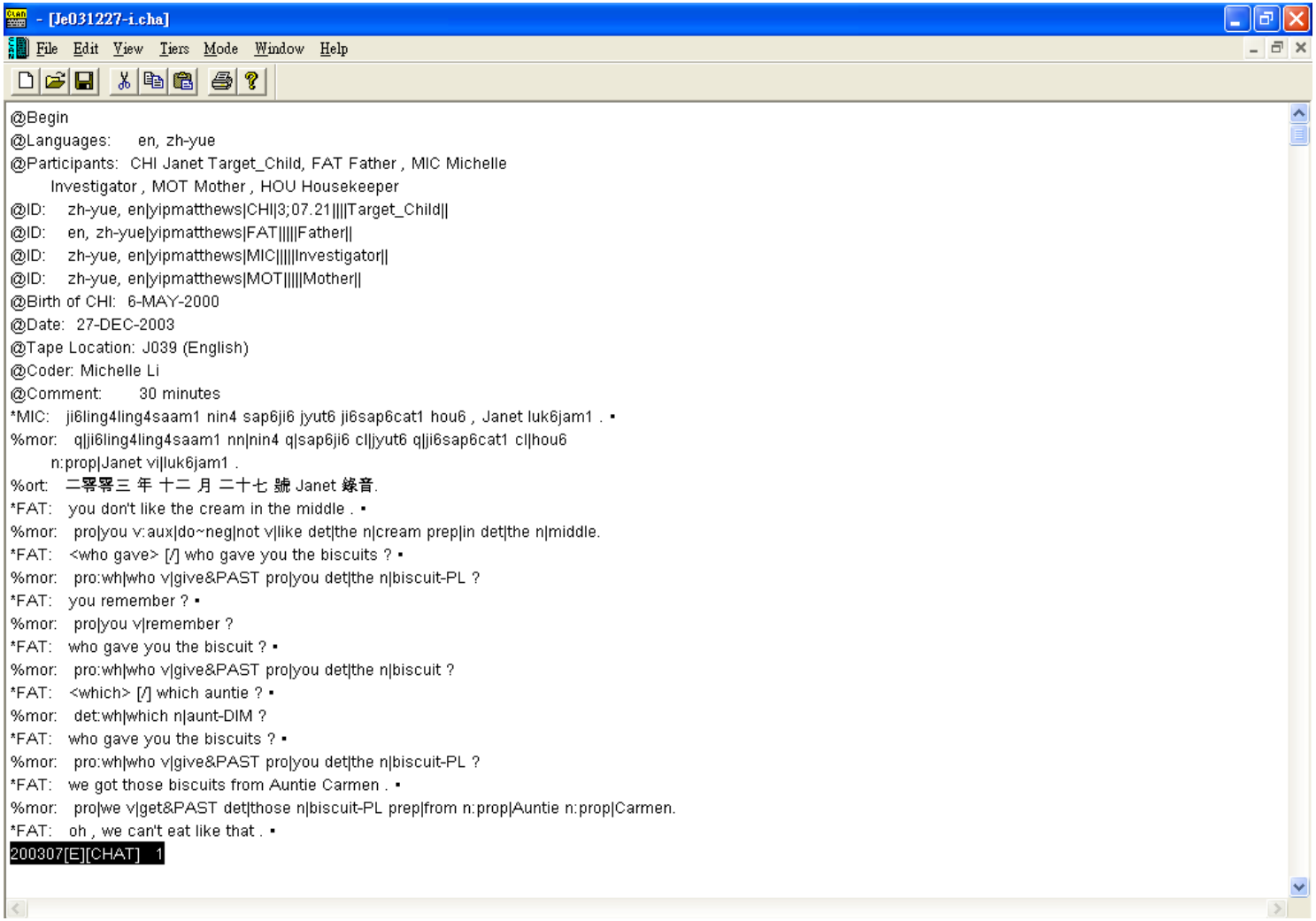
1. **Introducing CHILDES**
2. A case study in grammatical productivity
3. childes-db as a response to challenges of reproducibility
4. Examining function word development using childes-db

CHILDES - origins

- Diary studies were a critical method for early pioneers in child language, but verbatim capture difficult - anecdotes
- Growth of recording enabled exact transcripts, more and more precise data (Brown, 1973)
- Transcripts now primary method for studying child language production
- Still out of reach for automated annotation with ML (though soon?)



CHAT format


- [Je031227-i.cha]
File Edit View Tiers Mode Window Help
@Begin
@Languages: en, zh-yue
@Participants: CHI Janet Target_Child, FAT Father , MIC Michelle
Investigator , MOT Mother , HOU Housekeeper
@ID: zh-yue, en|yipmatthews|CHI|3;07.21||||Target_Child||
@ID: en, zh-yue|yipmatthews|FAT|||||Father||
@ID: zh-yue, en|yipmatthews|MIC|||||Investigator||
@ID: zh-yue, en|yipmatthews|MOT|||||Mother||
@Birth of CHI: 6-MAY-2000
@Date: 27-DEC-2003
@Tape Location: J039 (English)
@Coder: Michelle Li
@Comment: 30 minutes
*MIC: ji6ling4ling4saam1 nin4 sap6ji6 jyut6 ji6sap6cat1 hou6 , Janet luk6jam1 . •
%mor: q|ji6ling4ling4saam1 nn|nin4 q|sap6ji6 cl|jyut6 q|ji6sap6cat1 cl|hou6
n:prop|Janet v|luk6jam1 .
%ort: 二零零三 年 十二 月 二十七 號 Janet 綠音.
*FAT: you don't like the cream in the middle . •
%mor: pro|you v:aux|do~neg|not v|like det|the n|cream prep|in det|the n|middle.
*FAT: <who gave> [/] who gave you the biscuits ? •
%mor: pro:wh|who v|give&PAST pro|you det|the n|biscuit-PL ?
*FAT: you remember ? •
%mor: pro|you v|remember ?
*FAT: who gave you the biscuit ? •
%mor: pro:wh|who v|give&PAST pro|you det|the n|biscuit ?
*FAT: <which> [/] which auntie ? •
%mor: det:wh|which n|aunt-DIM ?
*FAT: who gave you the biscuits ? •
%mor: pro:wh|who v|give&PAST pro|you det|the n|biscuit-PL ?
*FAT: we got those biscuits from Auntie Carmen . •
%mor: pro|we v|get&PAST det|those n|biscuit-PL prep|from n:prop|Auntie n:prop|Carmen.
*FAT: oh , we can't eat like that . •
200307[E][CHAT] 1

Transcript

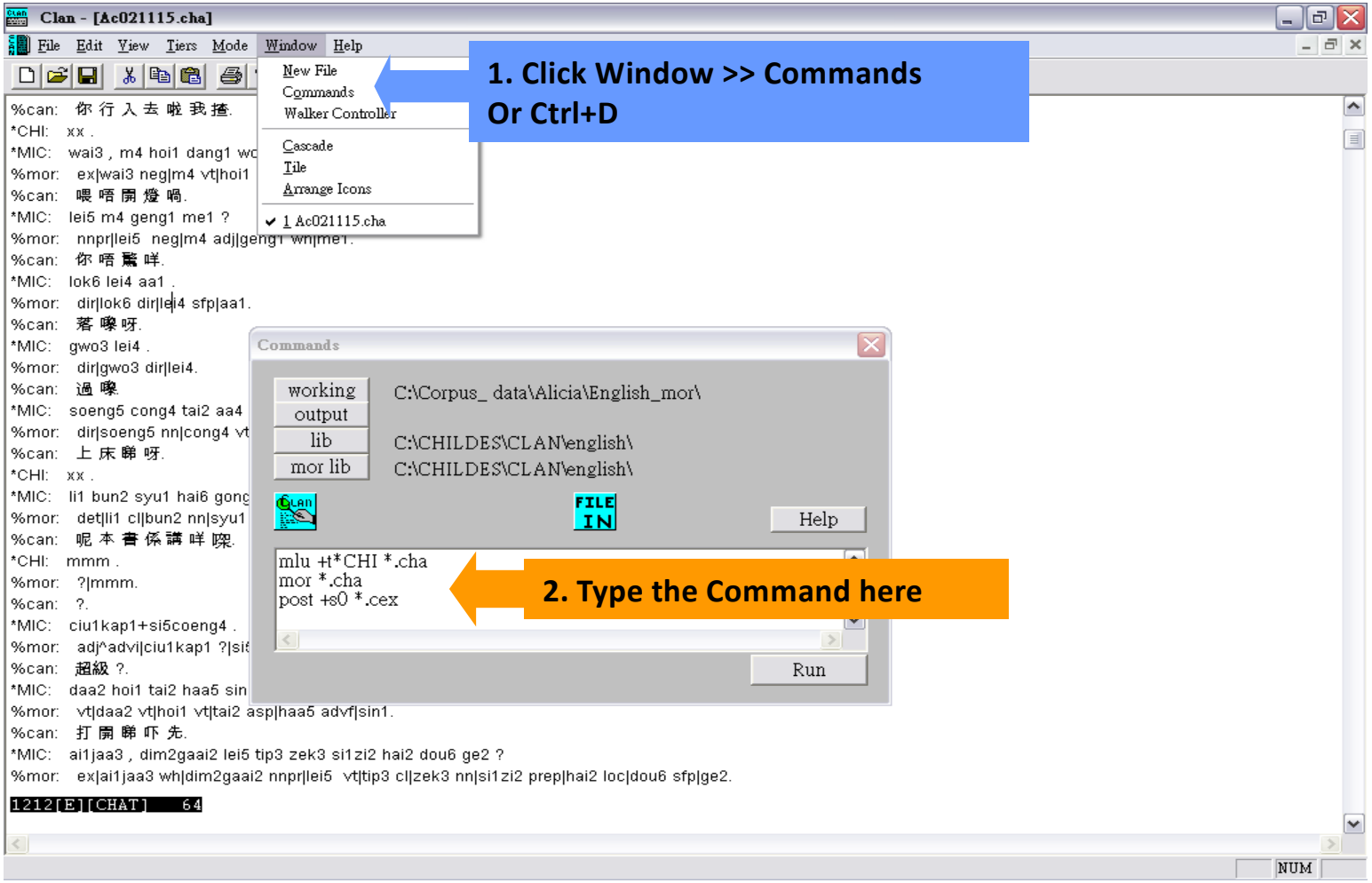
*FAT: okay , so come and sit down here . ▪
%mor: co|okay co|so v|come conj:coo|and v|sit adv|down adv:loc|here .
*FAT: shall we look at Postman_Bear first ? ▪
%mor: v:aux|shall pro|we v|look prep|at n:prop|Postman_Bear adv|first ?
*FAT: okay . ▪
%mor: co|okay .
*FAT: what's bear doing ? ▪
%mor: pro:wh|what~v:aux|be&3S n|bear v|do-PROG ?
*CHI: writing a letter , letter . ▪
%mor: v|write-PROG det|a n|letter n|letter .
*FAT: who is he writing to ? ▪
%mor: pro:wh|who v:aux|be&3S pro|he v|write-PROG prep|to ?
*CHI: friends . ▪
%mor: n|friend-PL
*FAT: why ? ▪
%mor: adv:wh|why ?
*FAT: ah , it's the first friends , so bear is writing <his letters> [//]
three letters . ▪
%mor: fil|ah pro|it~v|be&3S det|the adj|first n|friend-PL co|so n|bear v:aux|be&3S
200307[E][CHAT] 1

FAT=Father, he is saying "what's bear doing?"

CHI=children, saying "writing a letter, letter"

%mor=morphological tier, list parts of speech
"n" is NOUN, "PL" is plural, so "friends" is a plural noun

CLAN program



Beyond transcripts

- Audio and video available in some cases
- %MOR contains automatic part of speech tags
- %PHO contains phonetic transcripts for some corpora
- Talkbank (umbrella repository) contains many other corpora, including AphasiaBank, HomeBank, etc.

Outline

1. Introducing CHILDES
2. **A case study in grammatical productivity**
3. childes-db as a response to challenges of reproducibility
4. Examining function word development using childes-db

Early Grammatical Productivity

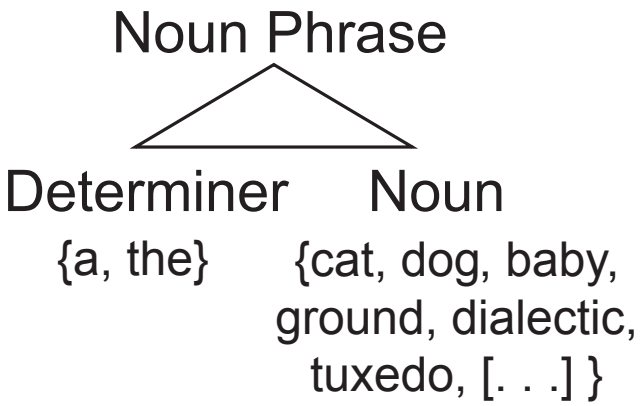
When do children have abstract grammatical categories?

Empiricist / Item-Based	Consider as Endpoints of a Continuum!	Generativist
- Later in development, after some language use (Braine, 1976) - Inductive / specific-to-general - *Products* of learning		- From the beginning of language learning (Valian, 1986; Yang, 2013) - Deductive / general-to-specific - *Supports* learning

The Case Study of Determiners

indefinite definite

- Determiners *a* and *the*
  - Distinction of definiteness (common ground, Clark & Brennan, 1991)
  - Probably requires advanced theory of mind
- Grammar licenses both “a” and “the” for count nouns
  - Child only ever heard *a boat …* does the child use *the boat*?



Empiricist / Item-Based	Generativist
No… stick to the observed uses of *boat*	**Yes…** generalize from observations of determiners used with other nouns

Measuring Grammatical Productivity

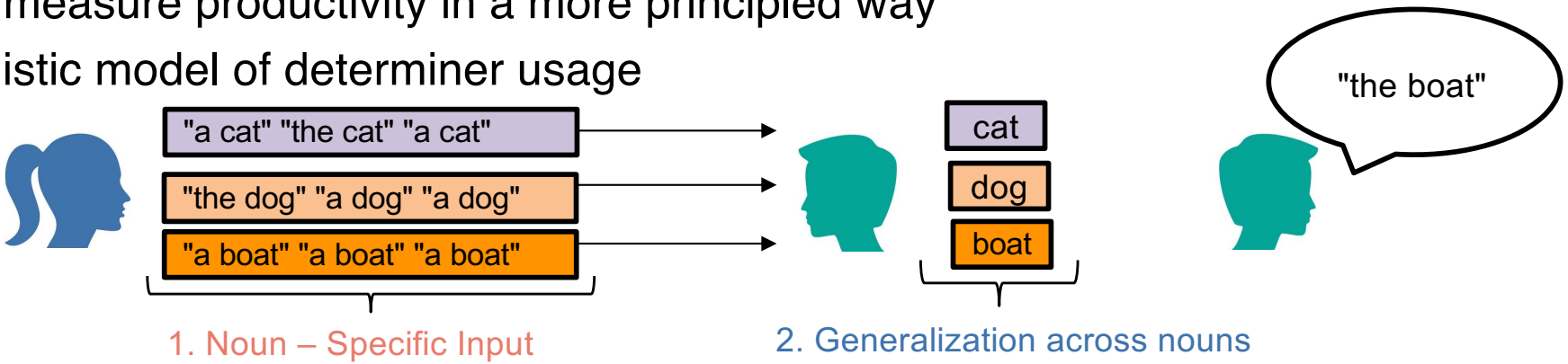
- Assess by looking for novel combinations?
- Can't without exhaustive corpora
- "Overlap statistic": proportion of nouns used with both determiners
  - Evidence for the empiricist hypothesis (Pine & Martindale, 1996)
- Overlap statistic is deeply flawed
  - Biased by sample size (Valian et al. 2009)
  - Increase with child age even for a purely imitative learner!

Overlap Statistic

$$\frac{\text{\# Unique Nouns used with "a" and "the"}}{\text{\# Unique Nouns used with "a" and/or "the"}}$$

Model: Decouple Input vs. Generalization

- Want to measure productivity in a more principled way
- Probabilistic model of determiner usage



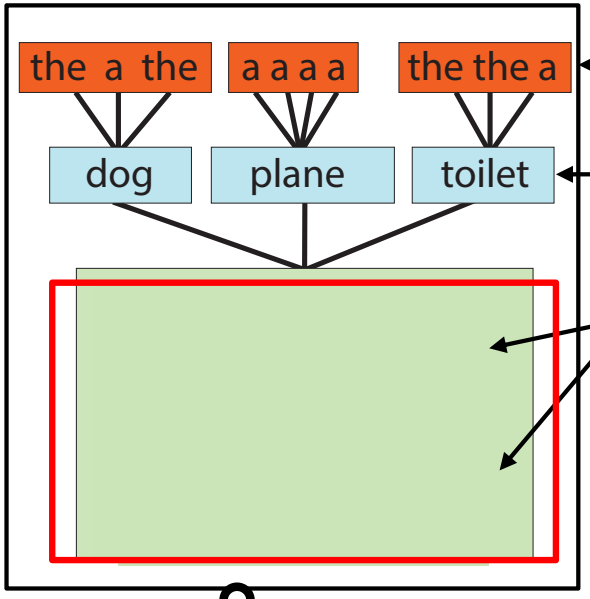
Determiner Usage with a Specific Noun (*e.g., boat*)

Adult Input	Child Production	Inference

Want to pool across all nouns... → Hierarchical Bayesian Model!

The Beta-Binomial Model: Child Productivity

Generative Model

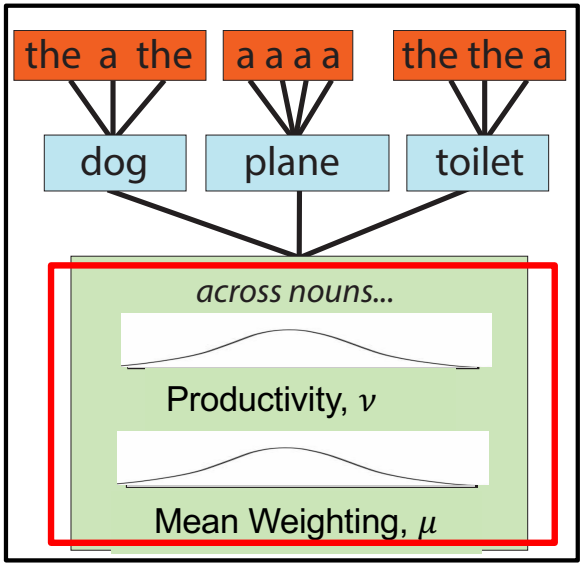


Child determiner+noun productions
the = 1, a = 0

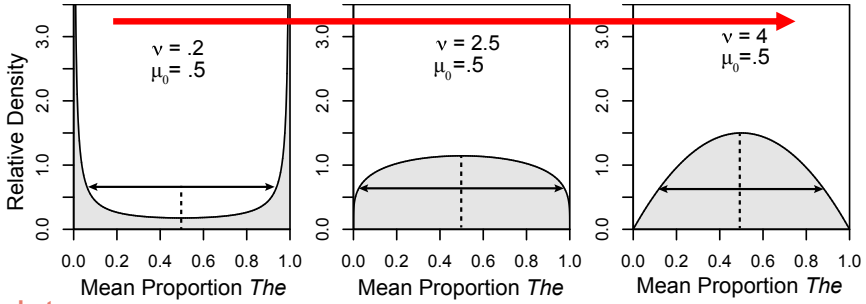
Each noun has a **determiner preference**
Binomial: μ_n between 0 and 1

Two parameters to control the distribution over determiner preferences...
$\text{Beta}(\mu\nu, [1-\mu]\nu)$ defines a distribution over *p*

Inference



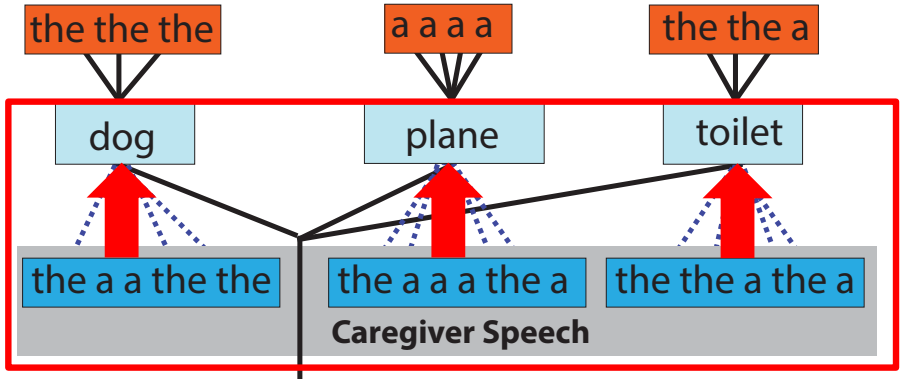
Productivity, ν = Concentration Parameter



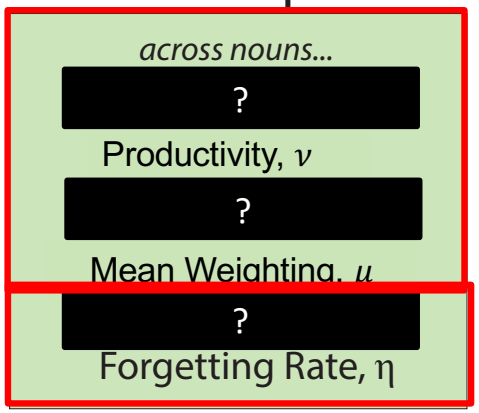
What is the **distribution over parameter values** that would generate these det+noun productions?

~Empiricist ⟷ ~Generativist

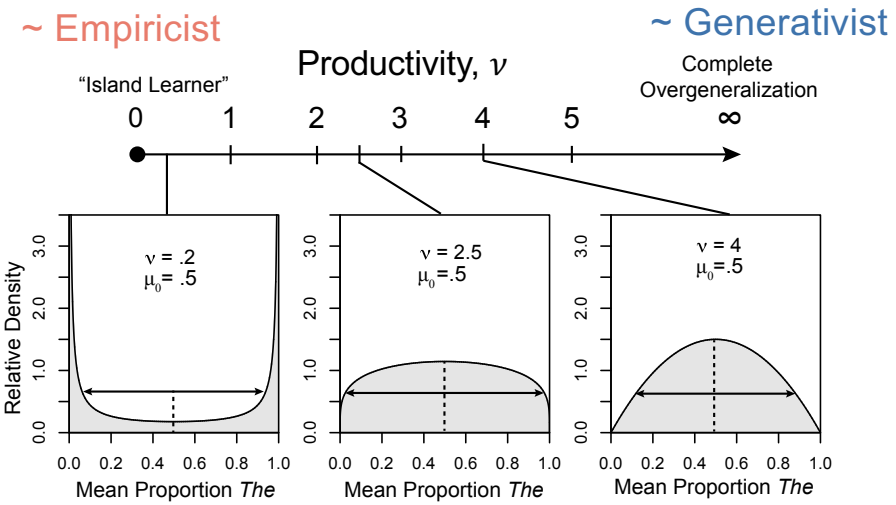
Combining Generative and Item-based Information



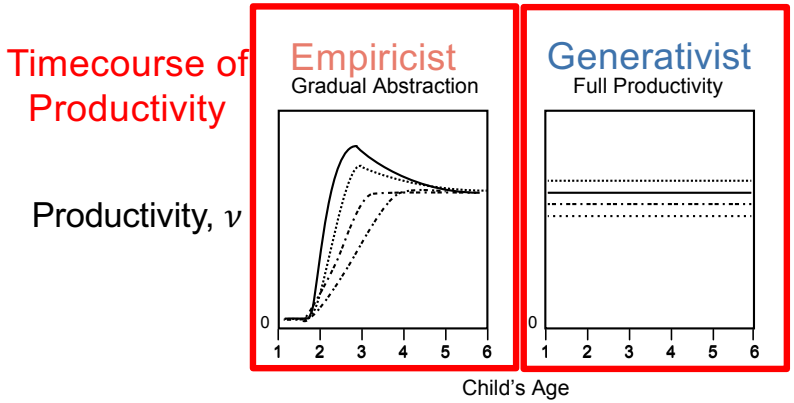
Item-based Information



Generative Knowledge

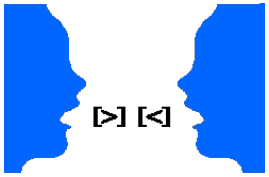


Timecourse of Productivity

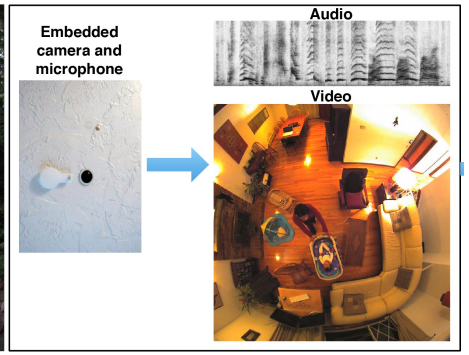


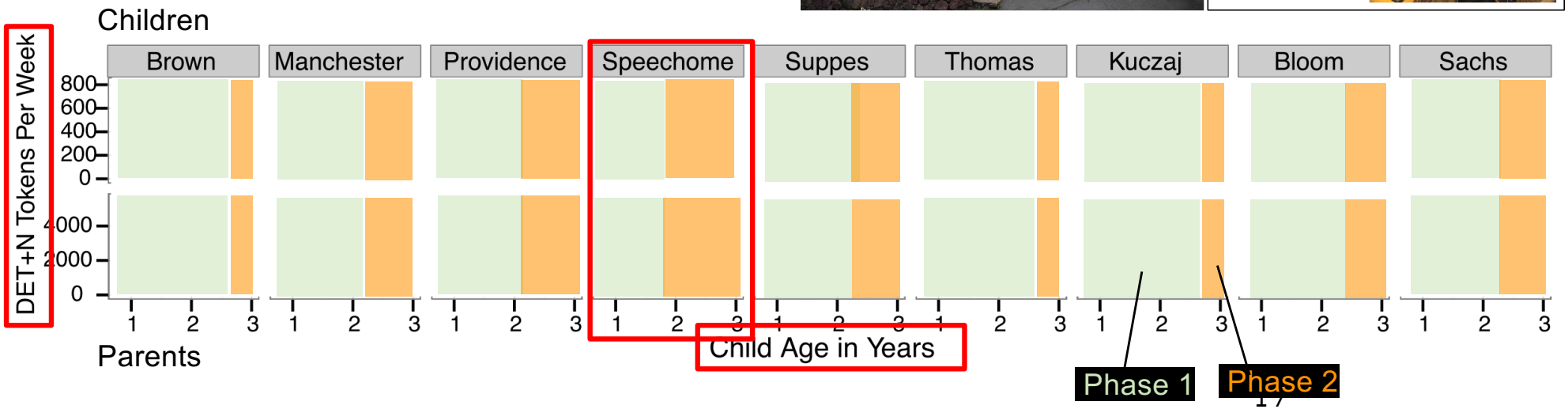
Fitting to Developmental Corpora

CHILDES:
Child Language
Data Exchange
System

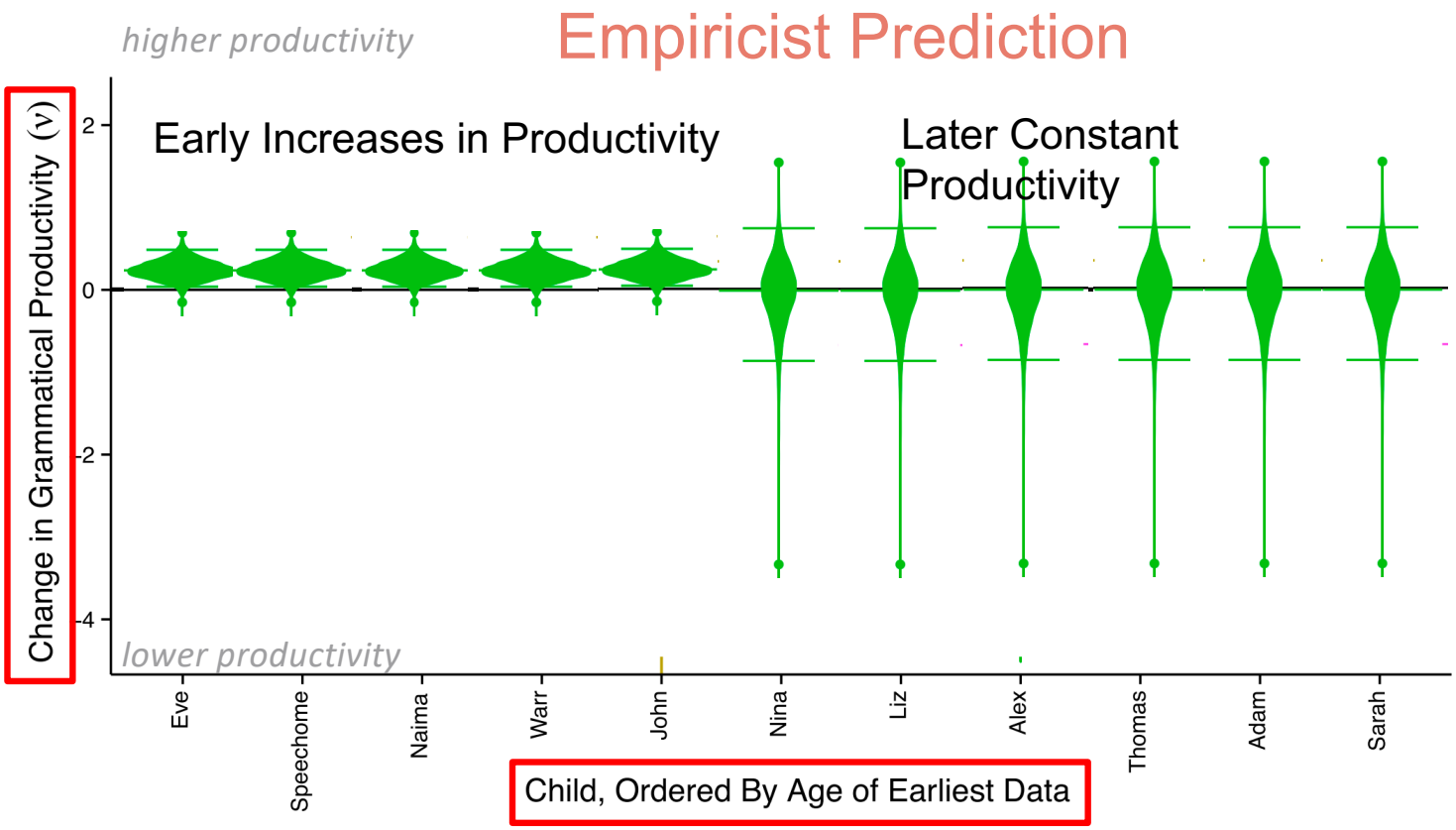


Human
Speechome
Corpus

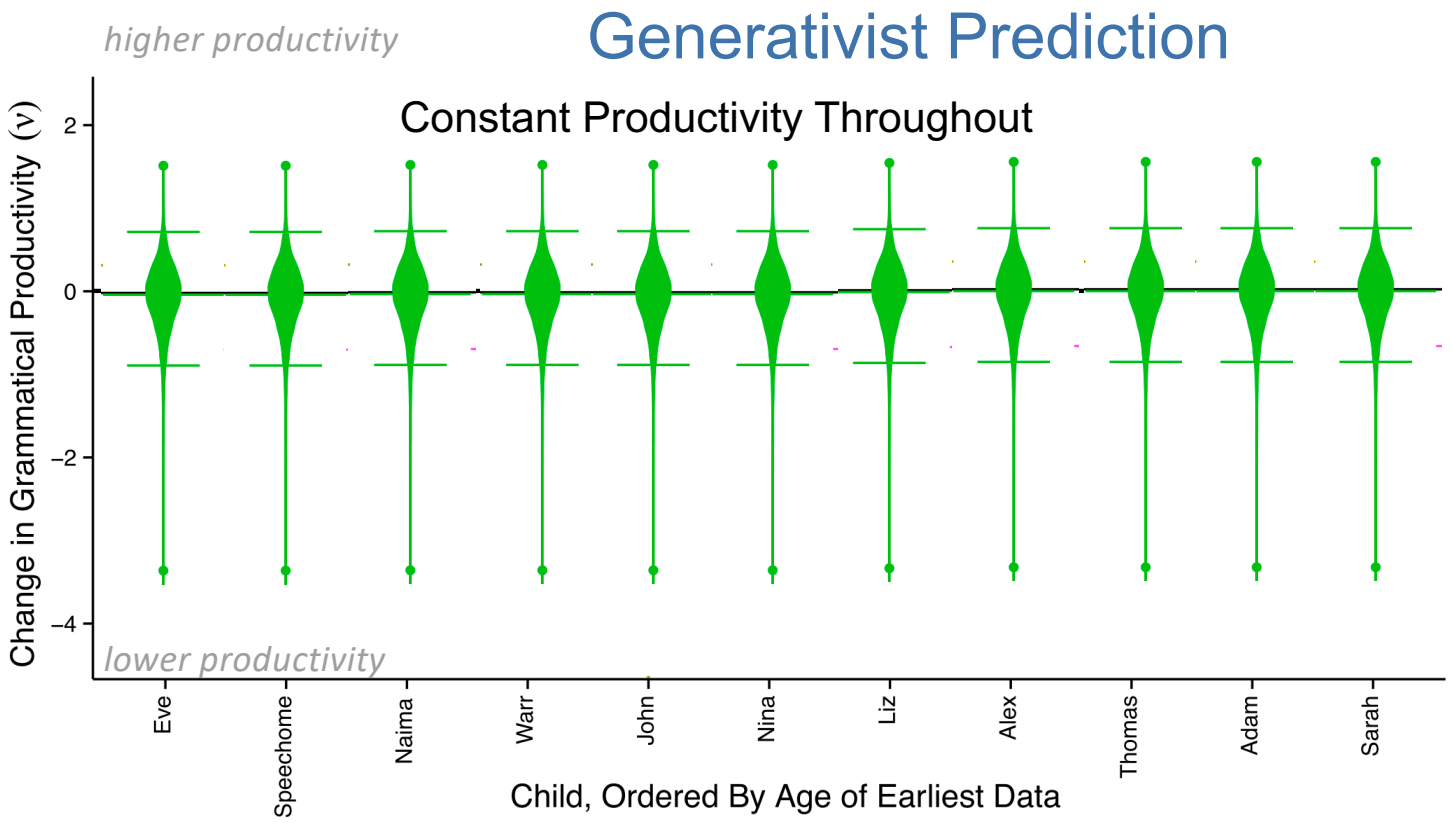




Predictions



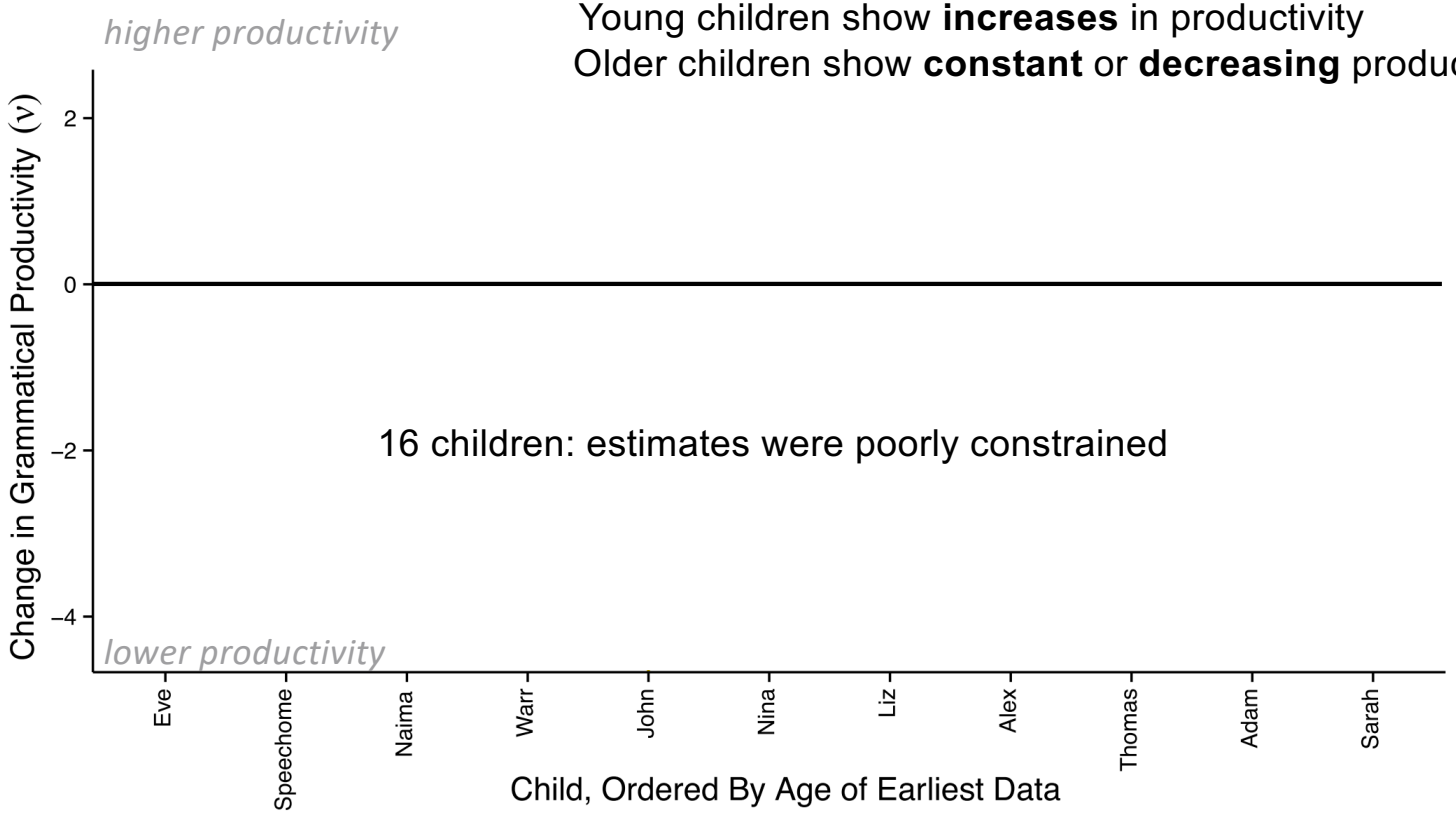
Predictions



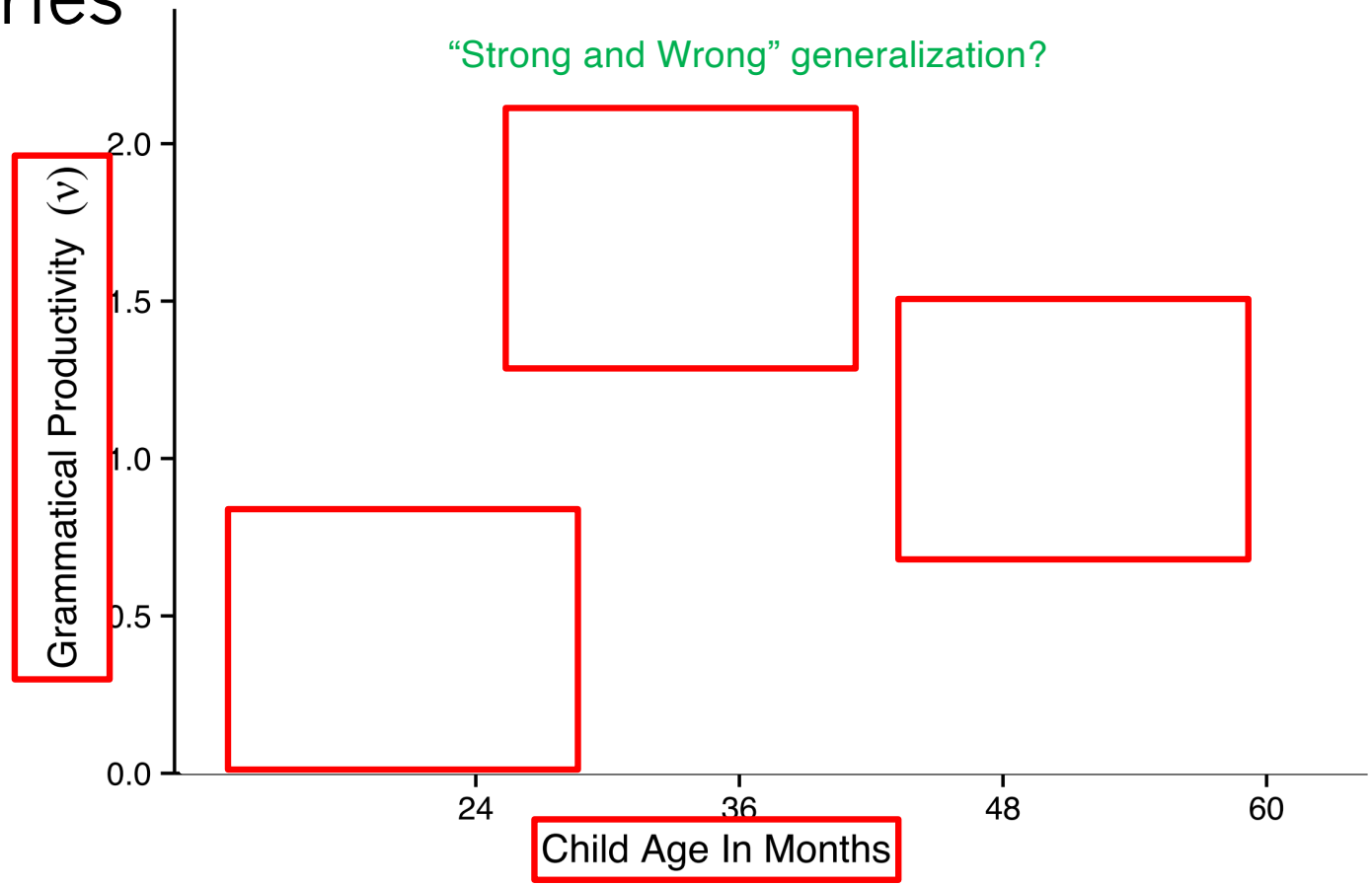
Results: Young Children Show Increasing Productivity

Consistent With Empiricist Prediction

Young children show **increases** in productivity
Older children show **constant** or **decreasing** productivity



Results: Trajectory Consistent With Item-Based Theories



Outline

1. Introducing CHILDES
2. A case study in grammatical productivity
3. **childes-db as a response to challenges of reproducibility**
4. Examining function word development using childes-db

Introducing childes-db

- ~40% of the time to prepare Meylan et al. (2017) was data extraction and preprocessing. Yuck!
  - Our solution: try many data processing choices and show whether your model holds regardless. Time-consuming!
- Gender language project (BUCLD 2016): Opportunity to separate and re-use data retrieval / preprocessing from analyses.
- Pushed on it summer + fall 2017, with a broad set of contributors
- *Behavior Research Methods*, 2019

childes-db

A flexible and reproducible interface to CHILDES

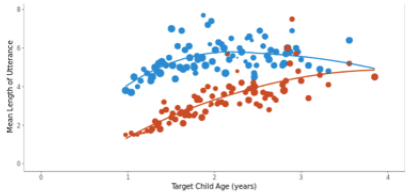
API Tutorial

Get a hands on walk-through on accessing `childes-db` through R.

```
> library(childesr)
> d_adam_prod <- get_tokens(collection = NULL,
+                           corpus = "Brown",
+                           role = "target_child",
+                           age = NULL,
+                           sex = NULL,
+                           child = "Adam",
+                           token = c("dog", "ball"))
Getting data from 1 child in 1 corpus ...
```

Visualizations

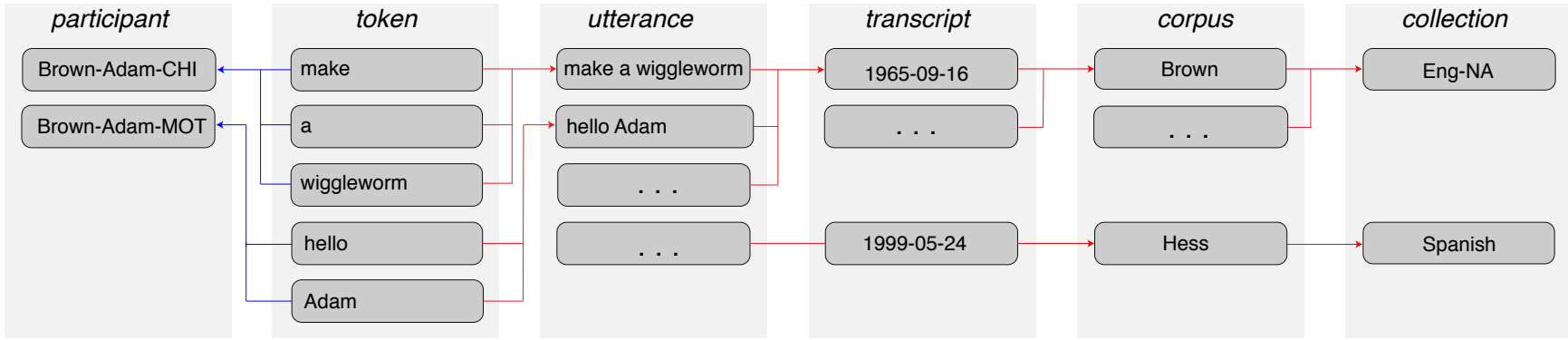
Explore the data in `childes-db` using our interactive applications.



The `childes-db` project is an open database storing CHILDES data in an easily accessible, tabular format. Researchers can now interface with CHILDES through interactive visualizations or the childesr R package.

Schema (Database Structure)

Relational database ~= hyperlinked Excel spreadsheet where you can't load the whole thing at once (60m rows)



id	gloss	replacement	stem	part_of_speech	speaker_id	utterance_id	token_order	corpus_id	transcript_id	speaker_code	speaker_name	speaker_role	target_child_id	target_child_age	target_child_name	target_child_sex	utterance_type	collection_id	collection_name	english	prefix	suffix
1	había		habe	v	1	1	1	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	have		13S PAS
2	una		un	det:art	1	1	2	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	one		f
3	vez		vez	n	1	1	3	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	turn		f
4	muy		muy	adv	1	2	1	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	very		
5	una		un	det:art	1	1	4	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	one		f
6	bien		bien	adv	1	2	2	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	well		
7	niña		niña	co:voc	1	1	5	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	child		
8	Diana		Diana	n:prop	1	2	3	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish			
9	que		que	pro:rel	1	3	1	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	that		
10	quién		quién	pro:int	1	4	1	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	question	1	Spanish	who		
11	le		le	pro:ind	1	3	2	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	him		
12	ganó		gana	v	1	4	2	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	question	1	Spanish	win		3S PRET
13	tenía		tene	v	1	3	3	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	have		13S PAS
14	mucho		mucho	adv	1	3	4	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	much		
15	la		el	det:art	2	5	1	1	1	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	the		f SG
16	miedo		miedo	n	1	3	5	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	fear		m
17	roja		rojo	adj	2	5	2	1	1	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	red		f
18	a		a	aprep	1	3	6	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	to		
19	la		el	det:art	1	3	7	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	the		f SG
20	oscuridad		oscuridad	n	1	3	8	1	2	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	darkness		f
21	la		el	det:art	1	6	1	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	the		f SG
22	roja		rojo	adj	1	6	2	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	red		f
23	en		en	prep	2	7	1	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	in		
24	eso		eso	pro:dem	2	7	2	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	that_one		
25	porqué		porqué	pro:int	1	9	1	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	question	1	Spanish	why		
26	quién		quién	pro:int	1	8	1	1	3	KAR	Karina	Adult	NULL	NULL	NULL	NULL	question	1	Spanish	who		
27	se		se	pro:refl	2	7	3	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	itself		
28	ganó		gana	v	1	8	2	1	3	KAR	Karina	Adult	NULL	NULL	NULL	NULL	question	1	Spanish	win		3S PRET
29	fue		i	v	2	7	4	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	go		3S PRET
30	la		el	det:art	2	7	5	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	the		f SG
31	porque		porque	conj	2	10	1	1	1	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	because		
32	ella		ello	pro:sub	2	10	2	1	1	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	he		f
33	luz		luz	n	2	7	6	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	light		f
34	el		el	det:art	2	11	1	1	3	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	the		m SG
35	la		la	pro:obj	2	10	3	1	1	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	she		f
36	amarillo		amarillo	adj	2	11	2	1	3	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	yellow		m
37	explicó		explica	v	2	10	4	1	1	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	explain		3S PRET
38	mejor		mejor	adj	2	10	5	1	1	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	better		
39	entonces		entonces	adv	2	12	1	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	then		
40	se		se	pro:refl	2	12	2	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	itself		
41	el		el	det:art	1	13	1	1	3	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	the		m SG
42	espantó		espanta	v	2	12	3	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	frighten		3S PRET
43	amarillo		amarillo	adj	1	13	2	1	3	KAR	Karina	Adult	NULL	NULL	NULL	NULL	declarative	1	Spanish	yellow		m
44	la		la	pro:obj	1	14	1	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	question	1	Spanish	she		f
45	tanto		tanto	adj	2	12	4	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	so_much		m
46	explicó		explica	v	1	14	2	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	question	1	Spanish	explain		3S PRET
47	mejor		mejor	adj	1	14	3	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	question	1	Spanish	better		
48	porqué		porqué	pro:int	1	15	1	1	3	KAR	Karina	Adult	NULL	NULL	NULL	NULL	question	1	Spanish	why		
49	que		que	pro:rel	2	16	1	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	that		
50	mhm		mhm	int	2	17	1	1	1	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish			
51	prendió		prende	v	2	16	2	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	ignite		3S PRET
52	porque		porque	conj	2	18	1	1	3	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	because		
53	una		un	det:art	2	16	3	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	one		f
54	éste		éste	pro:dem	2	18	2	1	3	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	this_one		
55	vela		vela	n	2	16	4	1	2	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	candle		f
56	dijo		deci	v	2	18	3	1	3	DIA	Diana	Child	NULL	NULL	NULL	NULL	declarative	1	Spanish	say		3S PRET
57	en		en	prep	1	19	1	1	1	KAR	Karina	Adult	NULL	NULL	NULL	NULL	question	1	Spanish	in		

Tabular Data for Text? That's Crazy!

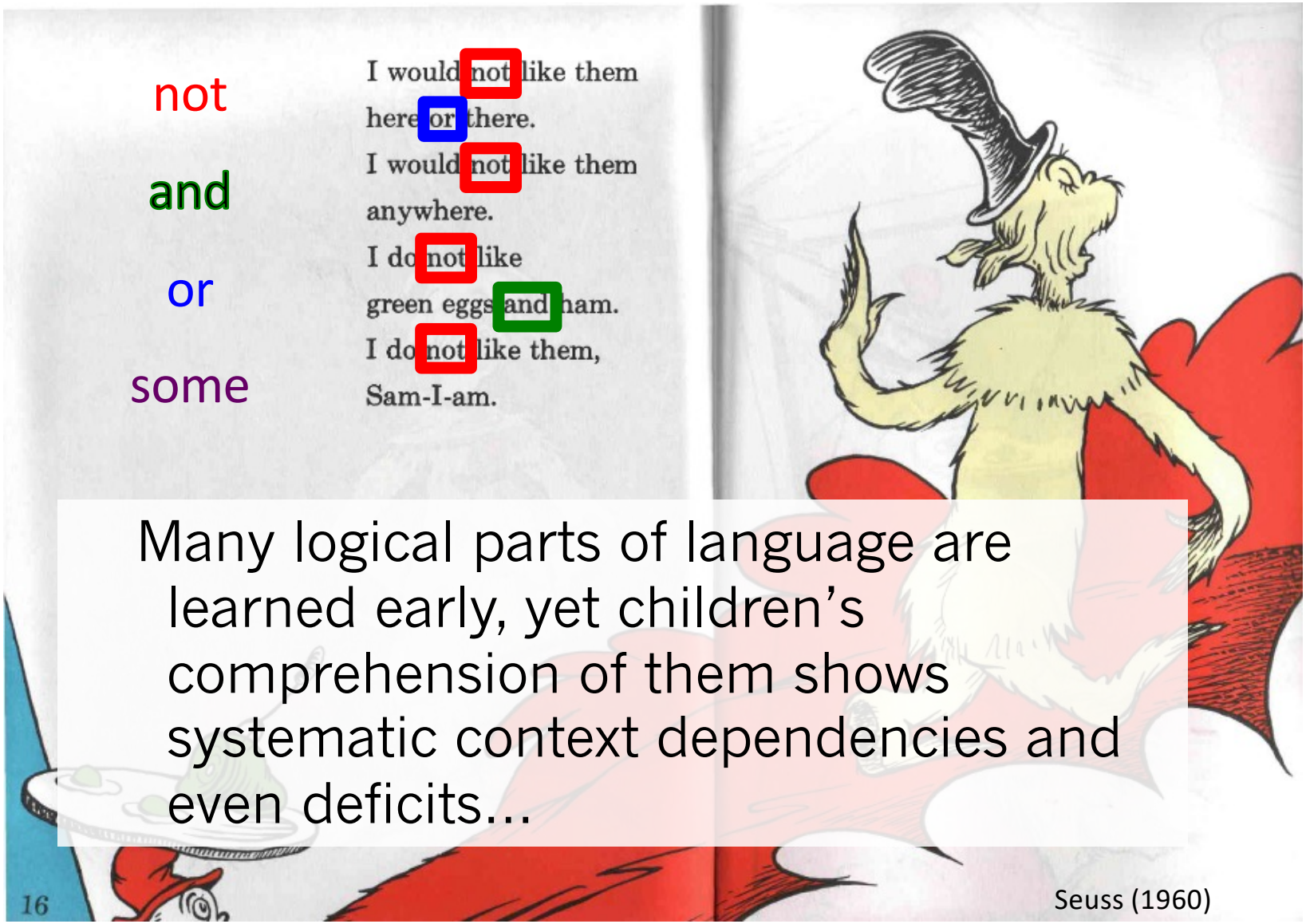
- Words = records; many pieces of information about each word
- Filtering, aggregation, merging, and counting: the components of any analyses
  - How many times did someone say "sheep"?
  - How many times did MOT say "sheep": condition on another column
- (words + annotations), (utterances + annotations)
- Disadvantages: Need to use indexes in order to do more complex queries of sequential material; extra computational overhead
  - Luckily computers are fast... and CHILDES is small

Known Limitations

- If it isn't in CHILDES, it isn't in childes-db. So far.
- CHI / MOT + FAT / Interviewer? Not all datasets
  - Use the database manuals. German MLU example.
- Doesn't cover TalkBank / AphasiaBank, etc.
- Only contains a subset of tiers (Phonbank)
- Brian MacWhinney keeps changing CHILDES, silently

Outline

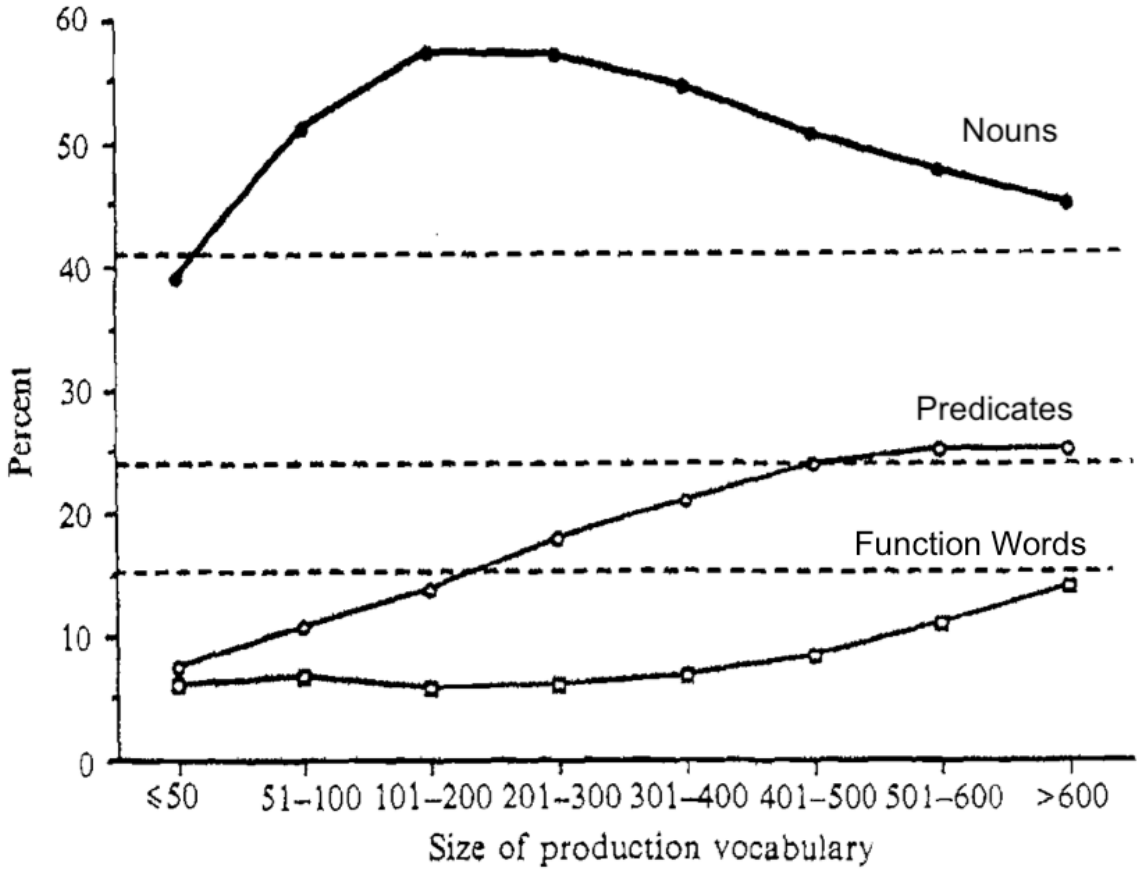
1. Introducing CHILDES
2. A case study in grammatical productivity
3. childes-db as a response to challenges of reproducibility
4. **Examining function word development using childes-db**



Many logical parts of language are learned early, yet children's comprehension of them shows systematic context dependencies and even deficits…

Seuss (1960)

Function words under-represented in early vocab



Bates et al. (1994)

Conjunction and Disjunction

- Case study for acquisition of logical meaning
- ”and” has relatively straightforward semantics
- In contrast, “or” has complex semantics and pragmatics!
- How is “or” acquired?

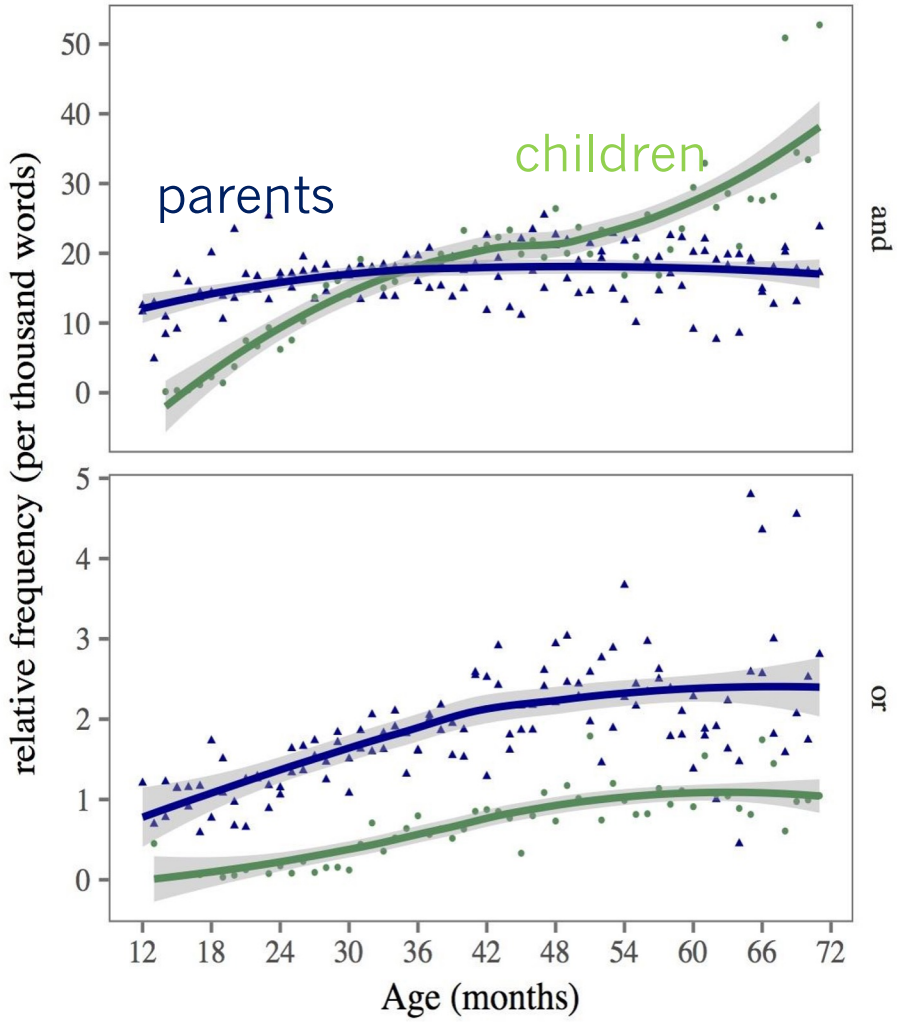
Examples of implications commonly conveyed by the use of linguistic disjunction.

Example	Implication	Label
Those above 65 or with symptoms are eligible.	⇝ including those above 65 and with symptoms.	Inclusivity
Abe plays basketball or soccer	⇝ he does not play both.	Exclusivity
I left the keys on the table or the counter.	⇝ The speaker does not know which.	Ignorance
You can use a pen or a pencil.	⇝ You can use a pen and you can use a pencil.	Free Choice

Jasbi et al. (2022)

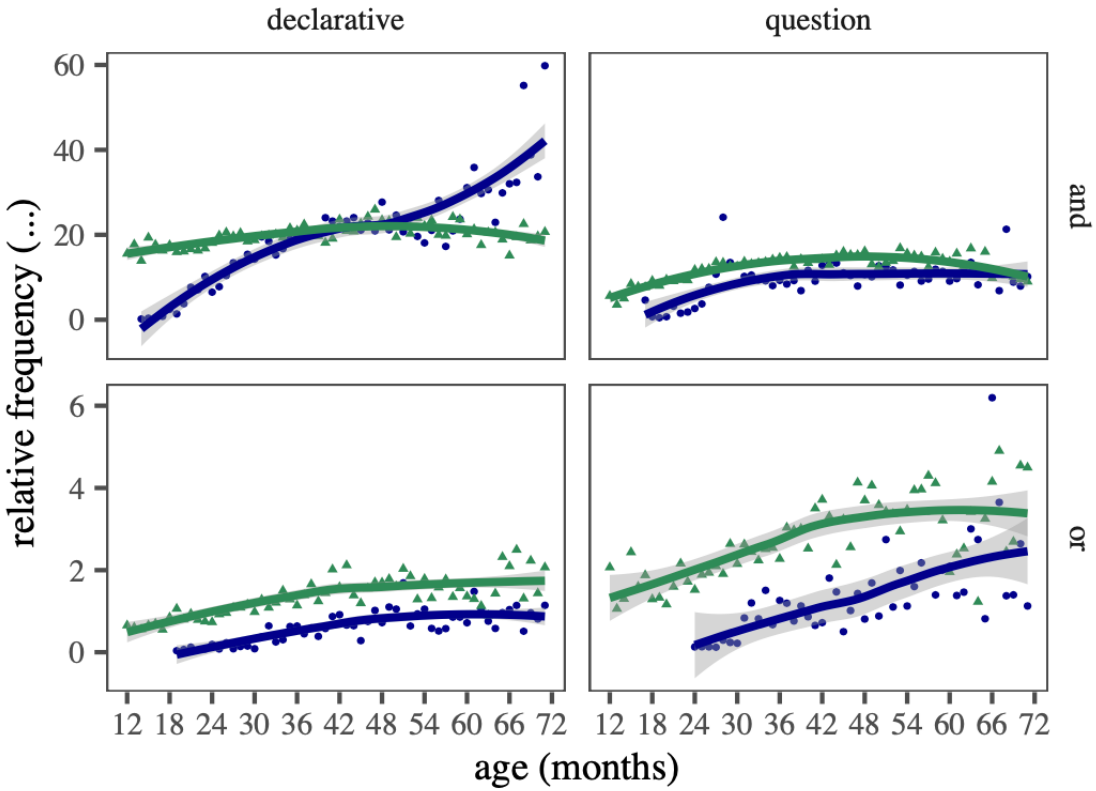
CHILDES counts

- Children produce “and” as much or more than parents
- Children lag being on the production of “or”
- What accounts for this difference?



Speech acts are important

- “and” is used more in declaratives, “or” more in questions
- Parents ask more questions than children!



Investigating contexts of use

- Hand coded 1000 uses of ”or” in CHILDES providence corpus
- Hand-coded features including
  - Interpretation (outcome)
  - Prosody
  - Logical consistency
  - Presence of negation
  - Answer type (polar question)
- Asked whether particular interpretations could be learned for a subset of items

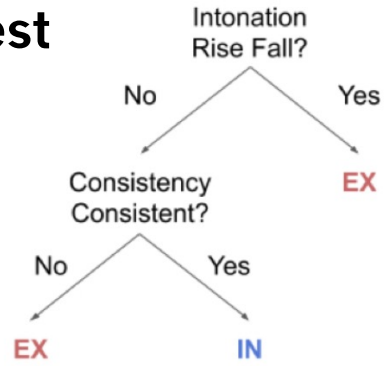
Decision trees

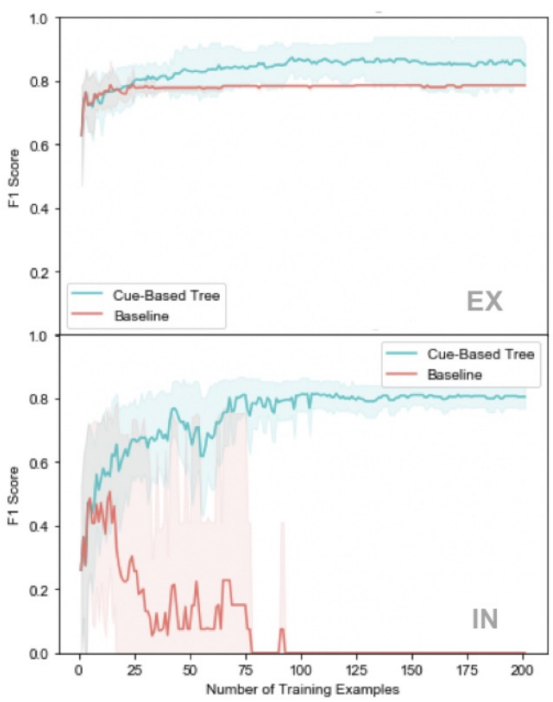
baseline

All Examples

EX

best





Could classify interpretation consistently with >80% using two features

Practicum: childes-db