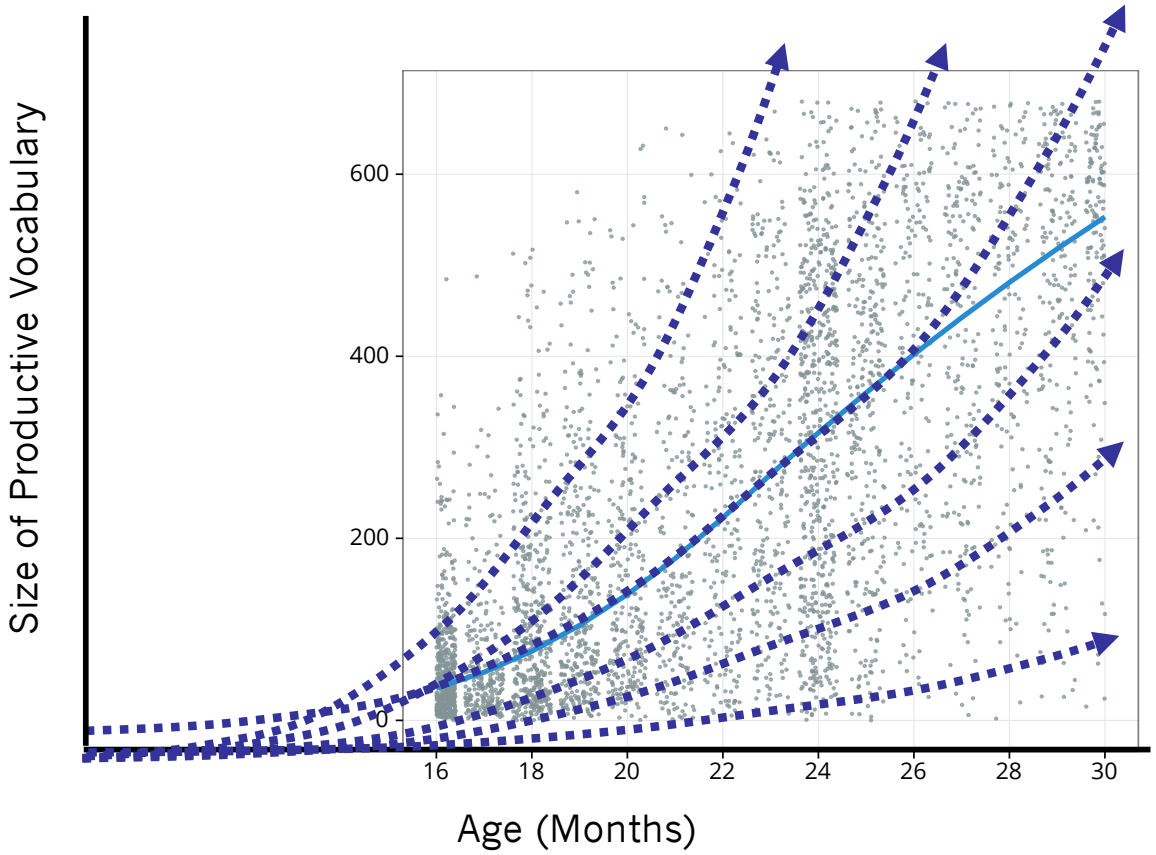
Language Learning: A Data-Driven Approach

Day 2: Digging deeper into Wordbank

Michael C. Frank

LOT Winter School

We want to understand what is **variable** and what is **consistent** in early word learning

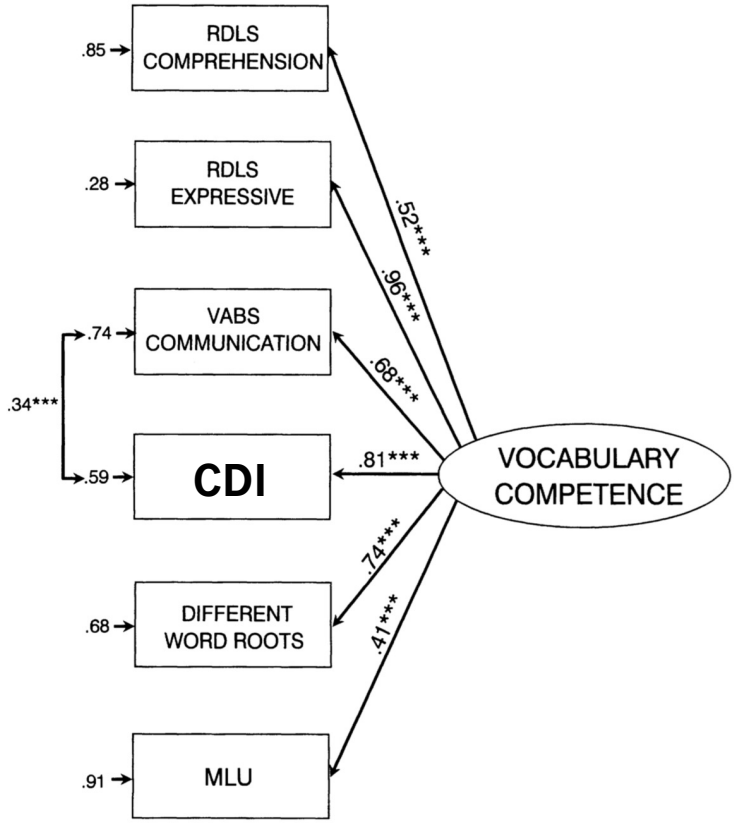


The MacArthur-Bates Communicative Development Inventory (CDI)

- Very cost-effective compared with direct assessment
- Transcripts very labor intensive
- Experimental assessment often requires lab visits, limited to a few words
- Parent report gives a holistic view of the child's language

Evaluating parent report

- CDIs are imperfect instruments (error & bias)
- Yet: summed score surprisingly reliable and valid (Fenson et al., 1994; 2007)
- Other measures of early language are very related to CDI



***$p \leq .001$

(Bornstein et al., 1998)

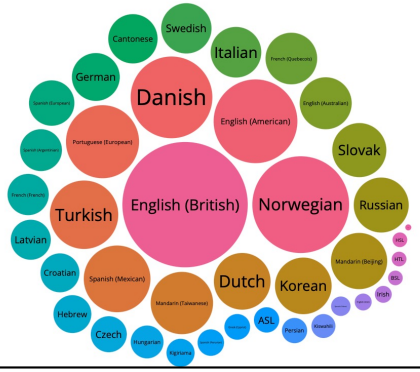
http://wordbank.stanford.edu

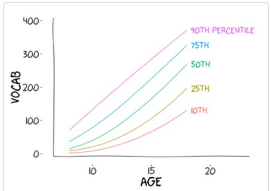
2022

Wordbank

An open database of children's vocabulary development

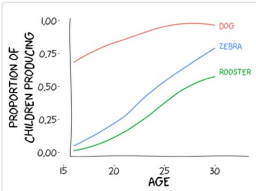
Wordbank contains data from 79,027 children and 93,258 CDI administrations, across 40 languages and 76 instruments:





Vocabulary Norms

Explore vocabulary size growth curves for various languages and demographic groups.



Item Trajectories

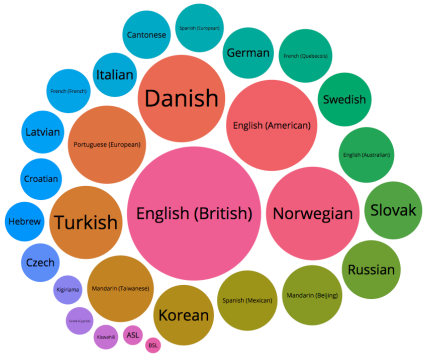
Explore trajectories of individual words, word categories, and grammar items.

2019

Wordbank

An open database of children's vocabulary development

Wordbank contains data from 75,144 children and 82,983 CDI administrations, across 29 languages and 56 instruments:

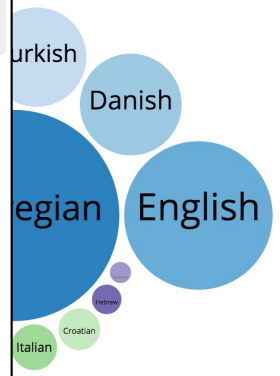


2015

Wordbank

An open database of children's vocabulary development

40,000 CDI administrations, ages and 23 instruments:



Frank et al. (2017), *Journal of Child Language*

Vocabulary Norms

This analysis shows growth curves for vocabulary size, the number of words that a child produces or understands, for different languages, forms, and measures. For some datasets, it is possible to compare growth curves across different demographic groups (birth order, ethnicity, gender, mother's education). Use a median quantile type to compare demographic groups on a single plot, or other quantile sizes to see separate curves and plots for each groups. See below for more details and an important disclaimer about clinical usage.

Language

English (American)

Form

Words & Sentences

Measure

Produces

question that parents were asked to answer

Choose Data

- [x] cross-sectional only
- [ ] normative sample

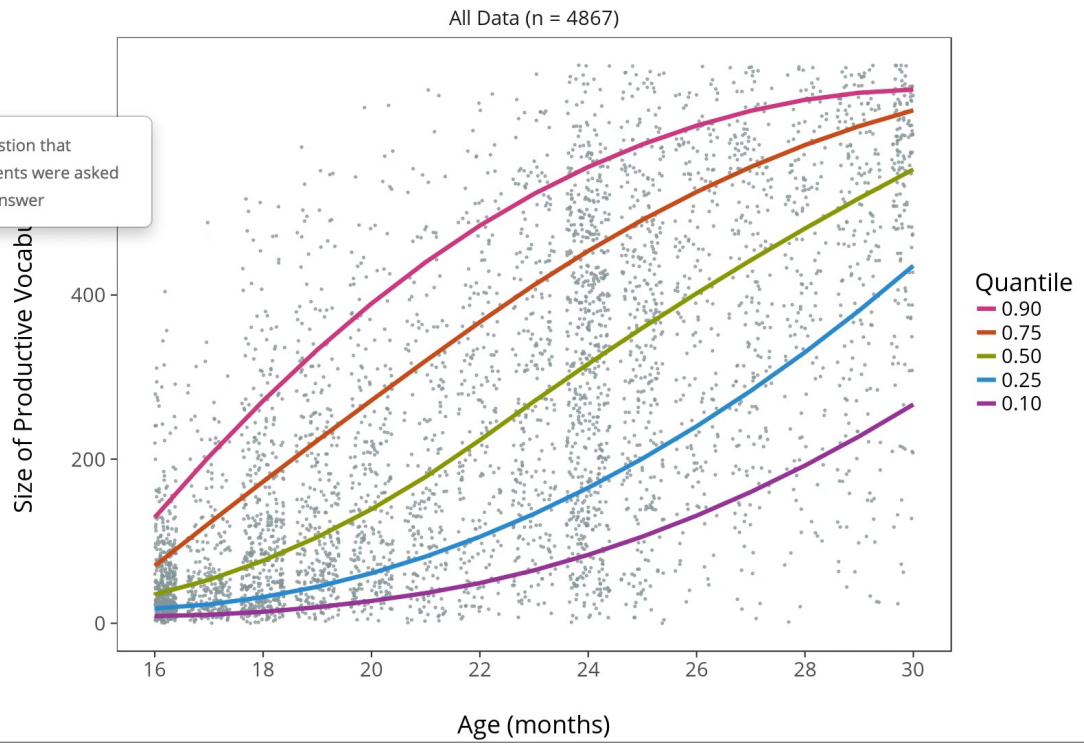
Split Variable

None

Quantiles

Standard

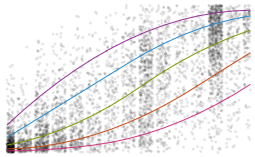
Plot | Table



Some generalizations about early language

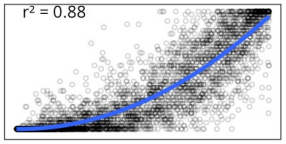
1. The **first words** are very consistent across languages and largely social in nature

2. Across children, **variability** is a constant in early language



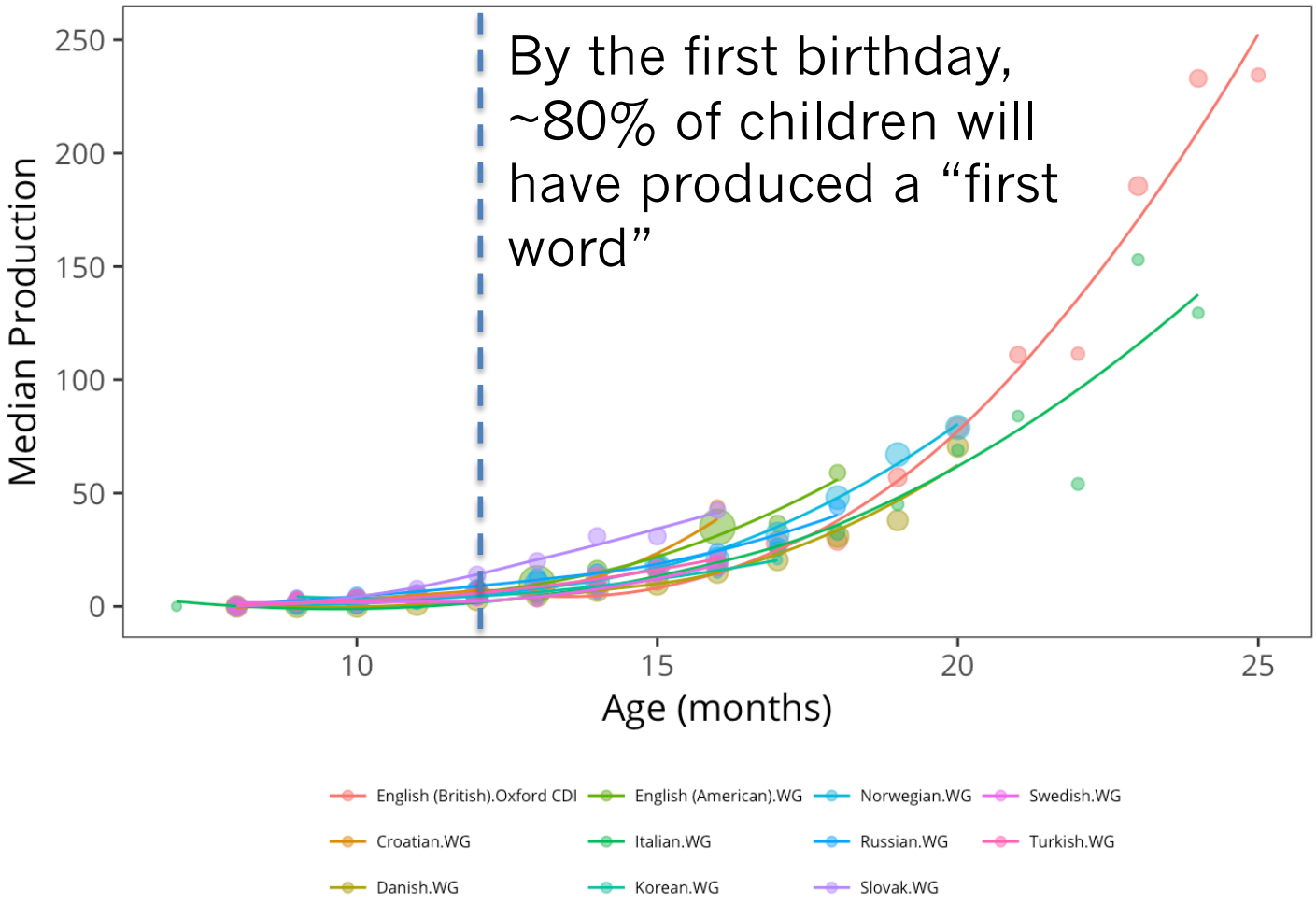
3. While there is a **noun bias** in early language, verbs/adjectives vary by language

4. The **growth of grammar** is linked to vocabulary growth



https://langcog.github.io/wordbank-book

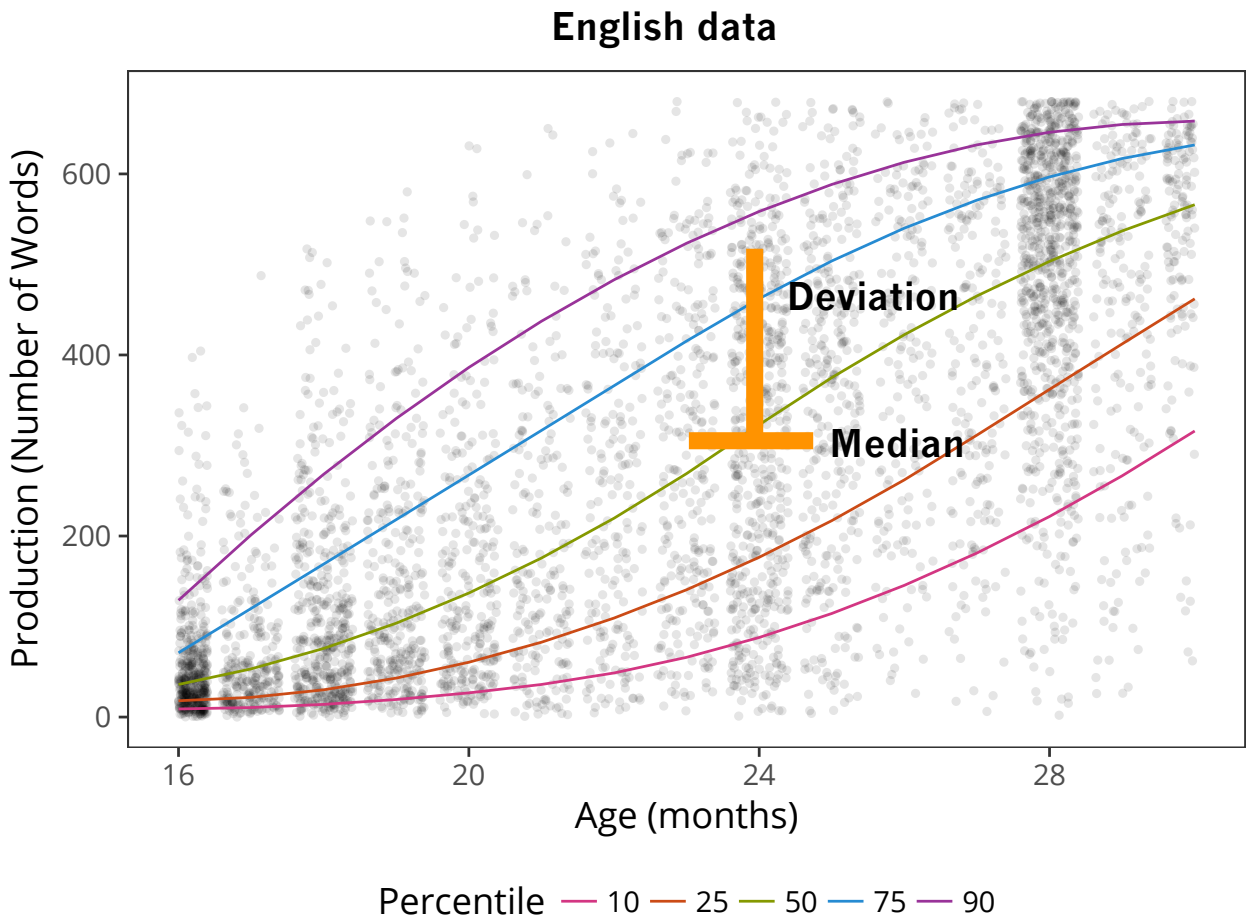
The emergence of (spoken) language



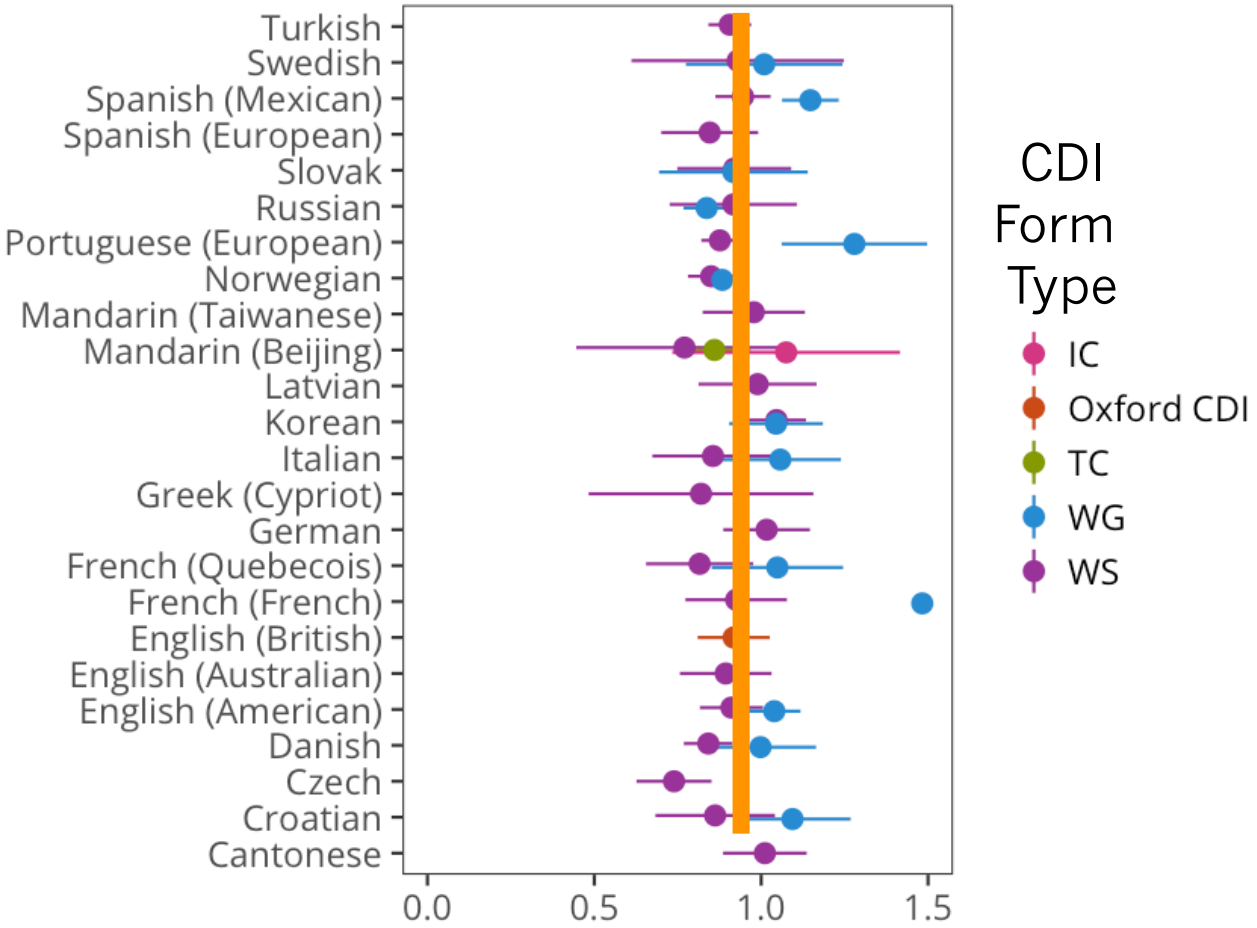
Consistency in early vocabulary

	Croatian	Danish	English (American)	Italian	Korean	Norwegian	Russian	Spanish	Swedish	Turkish
1	**grandma**	**peekaboo**	sister	**peekaboo**	**child's name**	**daddy**	**grandma**	milk	**child's name**	**mommy**
2	**mommy**	pattycake	brother	**daddy**	again	**peekaboo**	**daddy**	**bye**	**mommy**	**daddy**
3	**daddy**	**child's name**	**daddy**	**mommy**	money	**mommy**	grandpa	**yum yum**	**daddy**	**yum yum**
4	**bye**	**daddy**	**mommy**	baby food	aunt	**no**	**yum yum**	**daddy**	**peekaboo**	water
5	**child's name**	**mommy**	**child's name**	**child's name**	cracker	**child's name**	need	**mommy**	**bye**	outing
6	grandpa	**no**	bottle	**hi**	play	**hi**	can	water	bath	**child's name**
7	woof woof	**yum yum**	**peekaboo**	**grandma**	**peekaboo**	**bye**	weee	**no**	**hi**	baby food
8	cat	**hi**	**no**	water	**grandma**	pattycake	thump	ouch	**no**	up
9	vroom	**bye**	**bye**	**no**	with	**yum yum**	hide and seek	ball	lamp	clap
10	**yum yum**	ouch	bath	woof woof	food	good night	meow	tickle	look	ball

Children vary... widely!



Toddlers are all over the place – the world around!

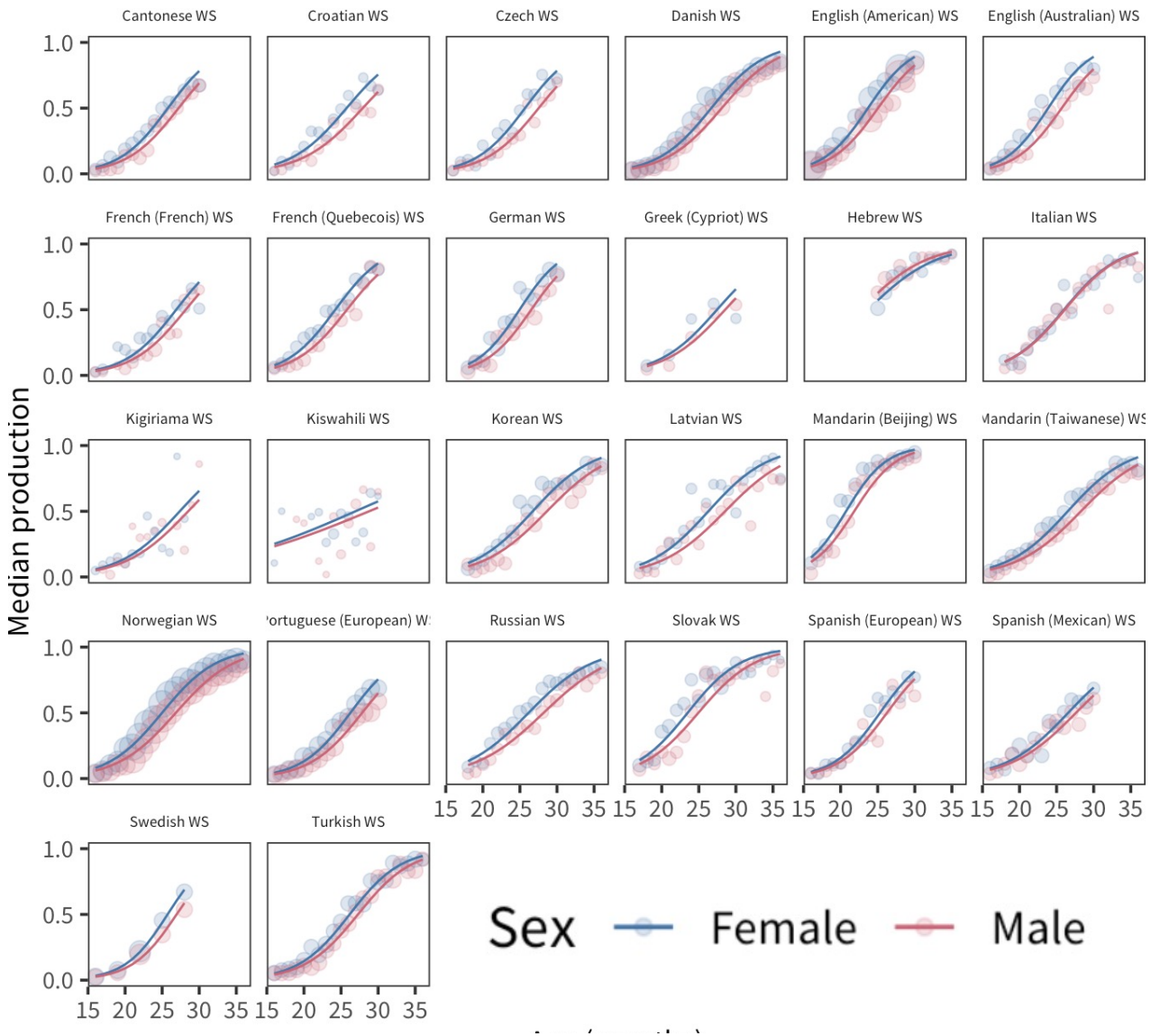


Language is variable!

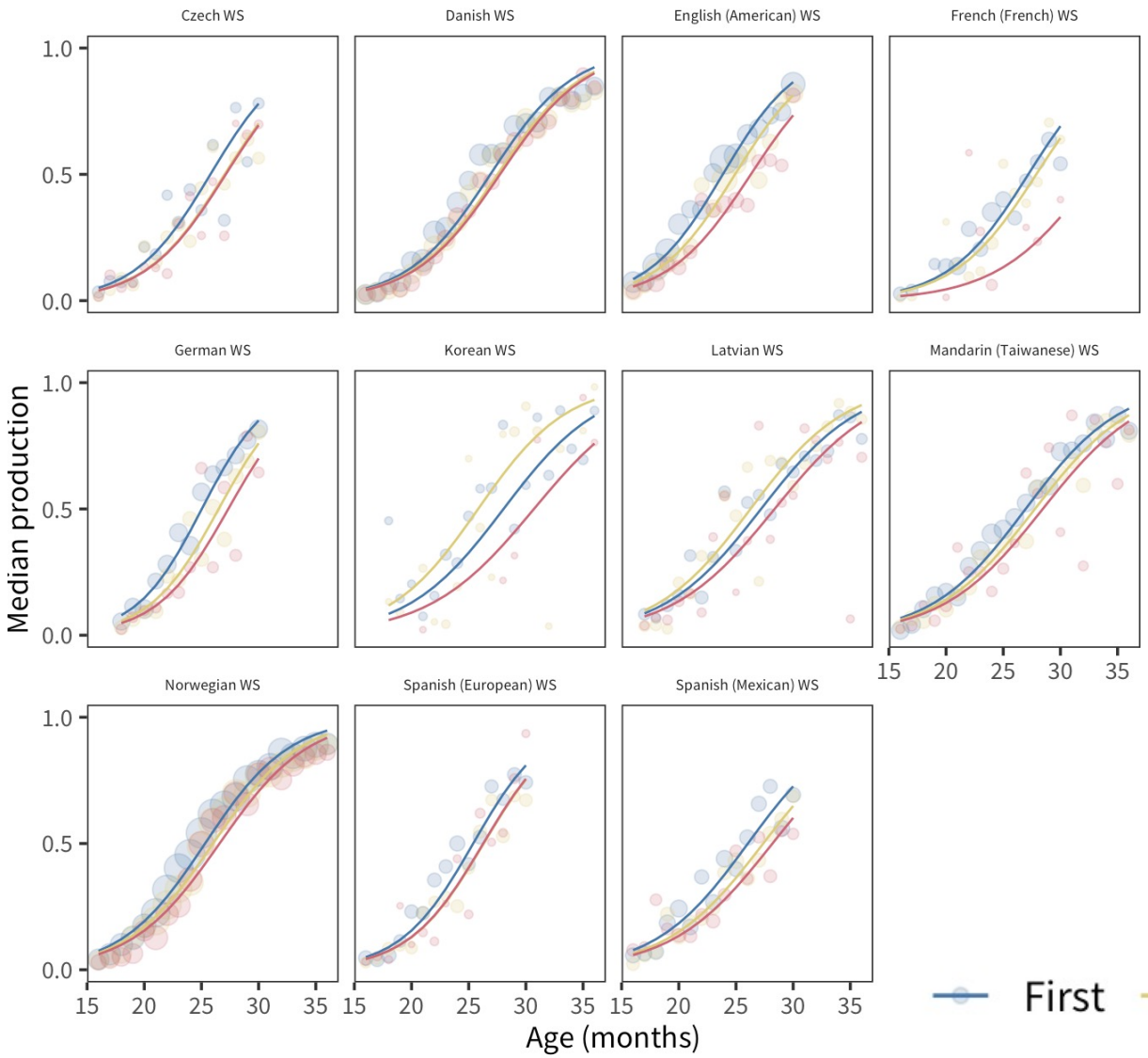
Walks – 2-3 month range

Says 200 words – 6-12 mo range

Sex differences



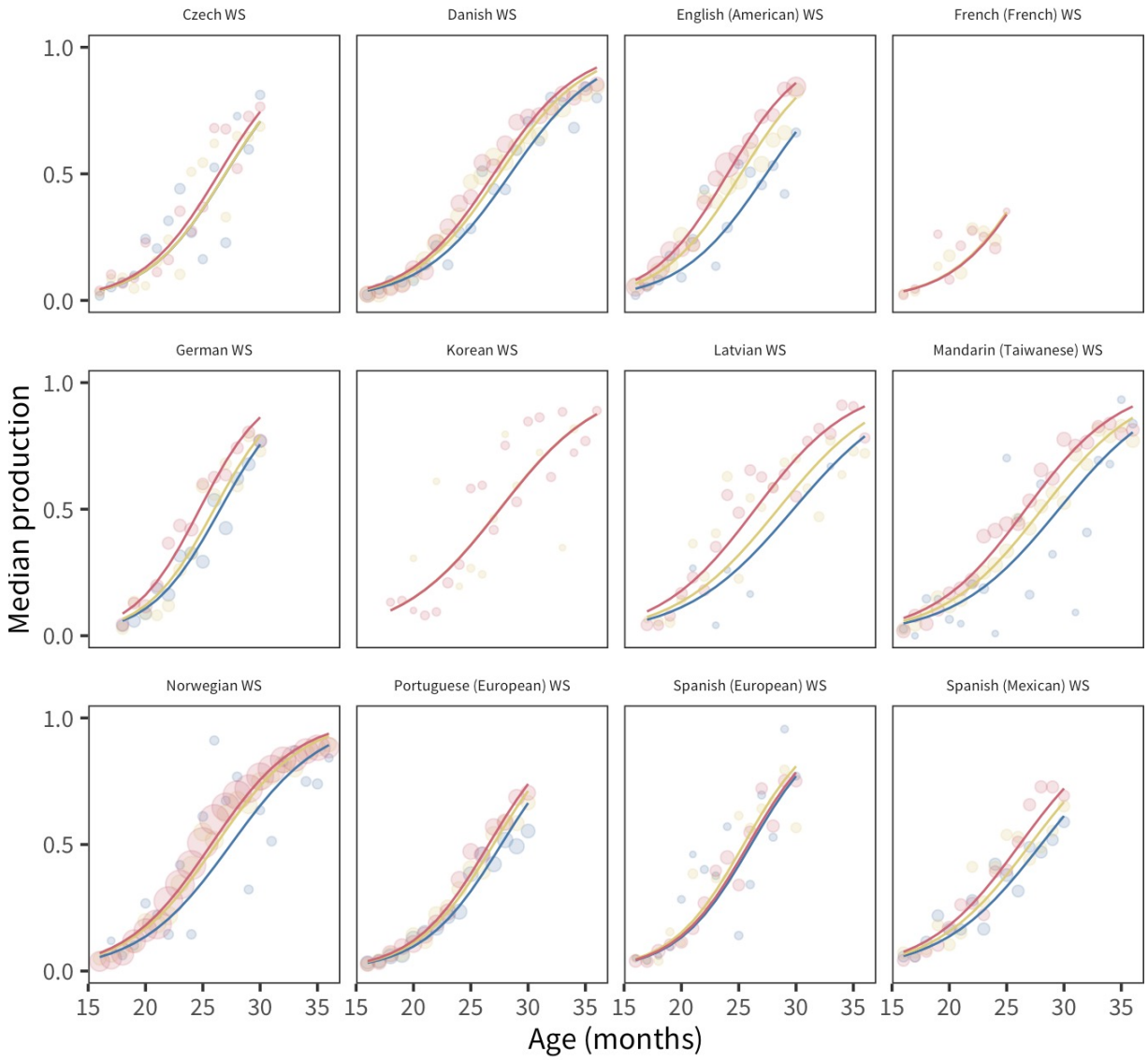
Birth order



Birth Order

First Second Third+

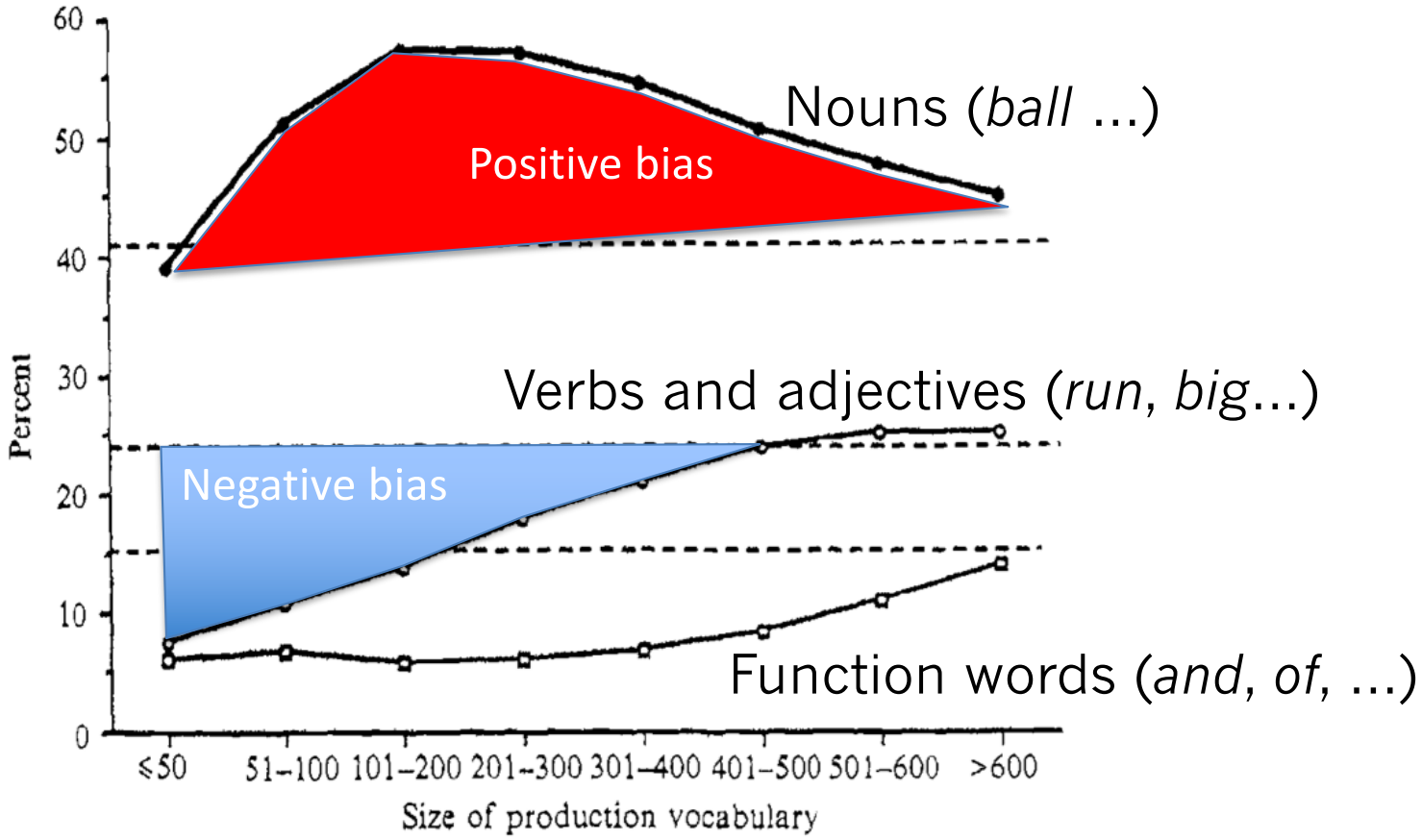
Maternal education



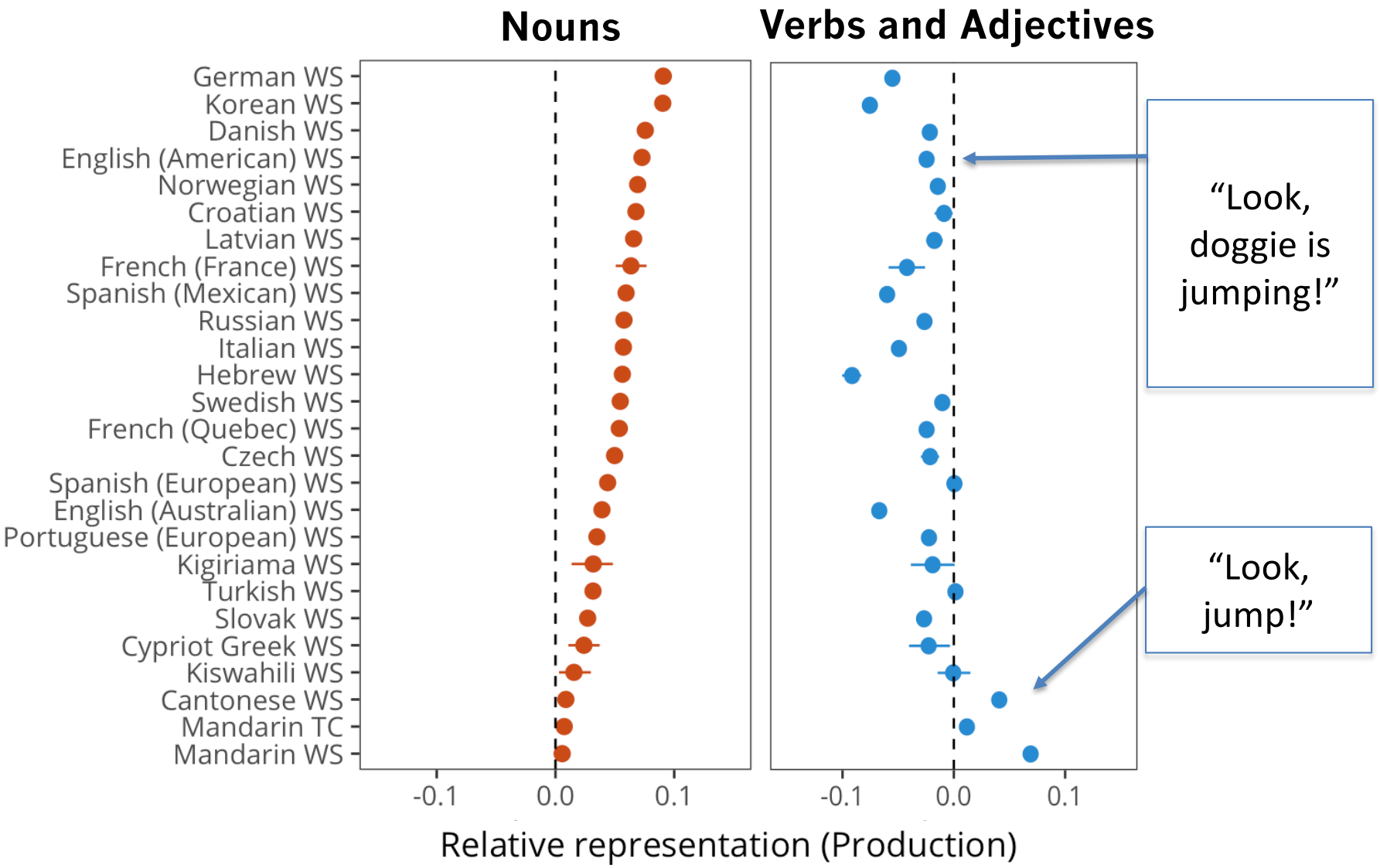
Maternal Education

- Below Secondary
- Secondary
- College and Above

A “noun bias”



Bates et al. (1994)


Nouns
Verbs and Adjectives
German WS
Korean WS
Danish WS
English (American) WS
Norwegian WS
Croatian WS
Latvian WS
French (France) WS
Spanish (Mexican) WS
Russian WS
Italian WS
Hebrew WS
Swedish WS
French (Quebec) WS
Czech WS
Spanish (European) WS
English (Australian) WS
Portuguese (European) WS
Kigiriama WS
Turkish WS
Slovak WS
Cypriot Greek WS
Kiswahili WS
Cantonese WS
Mandarin TC
Mandarin WS
-0.1
0.0
0.1
-0.1
0.0
0.1
Relative representation (Production)
"Look, doggie is jumping!"
"Look, jump!"

Sounds | Games & Routines | Body Parts | Vehicles | Animals

Food & Drink | Toys | People | Clothing | House

Time Words

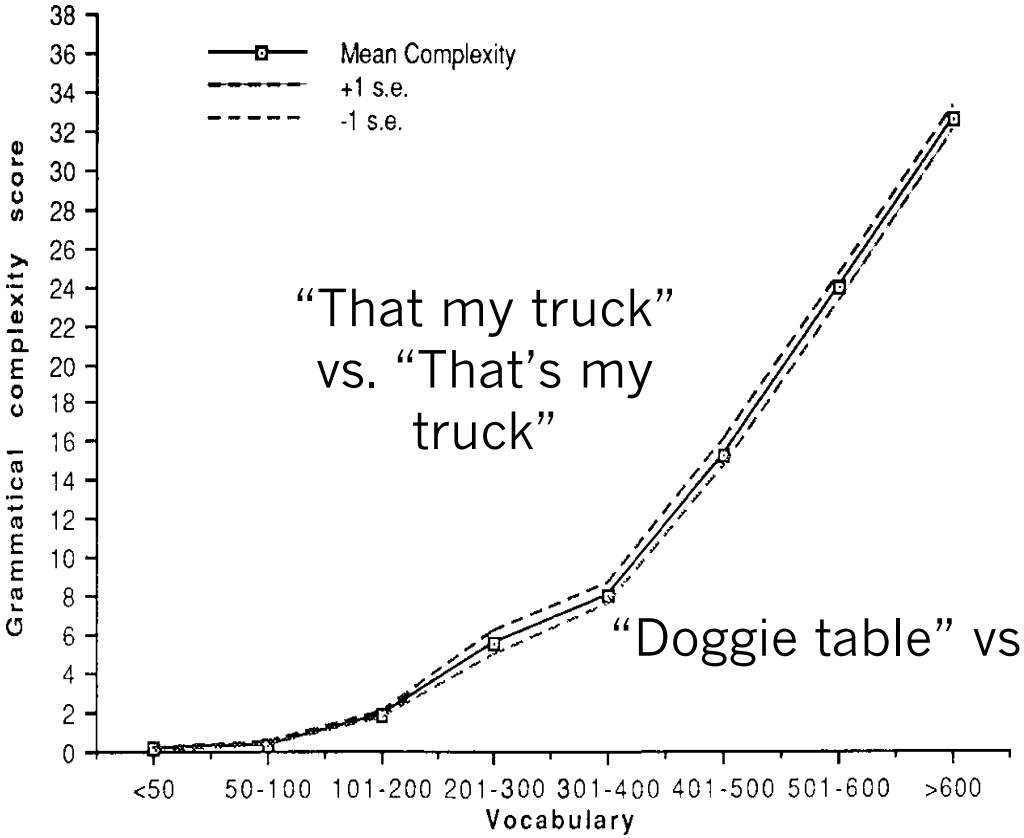
1.0 | 0.0 | 0.5 | 1.0

Vocabulary size

Madeline (33mo): “My baby really sick.” Oh yeah? What happened? “My baby spit up three times last year and seventy-nine times last night.”

(cf. Tillman & Barner, 2015)

Learning to combine words



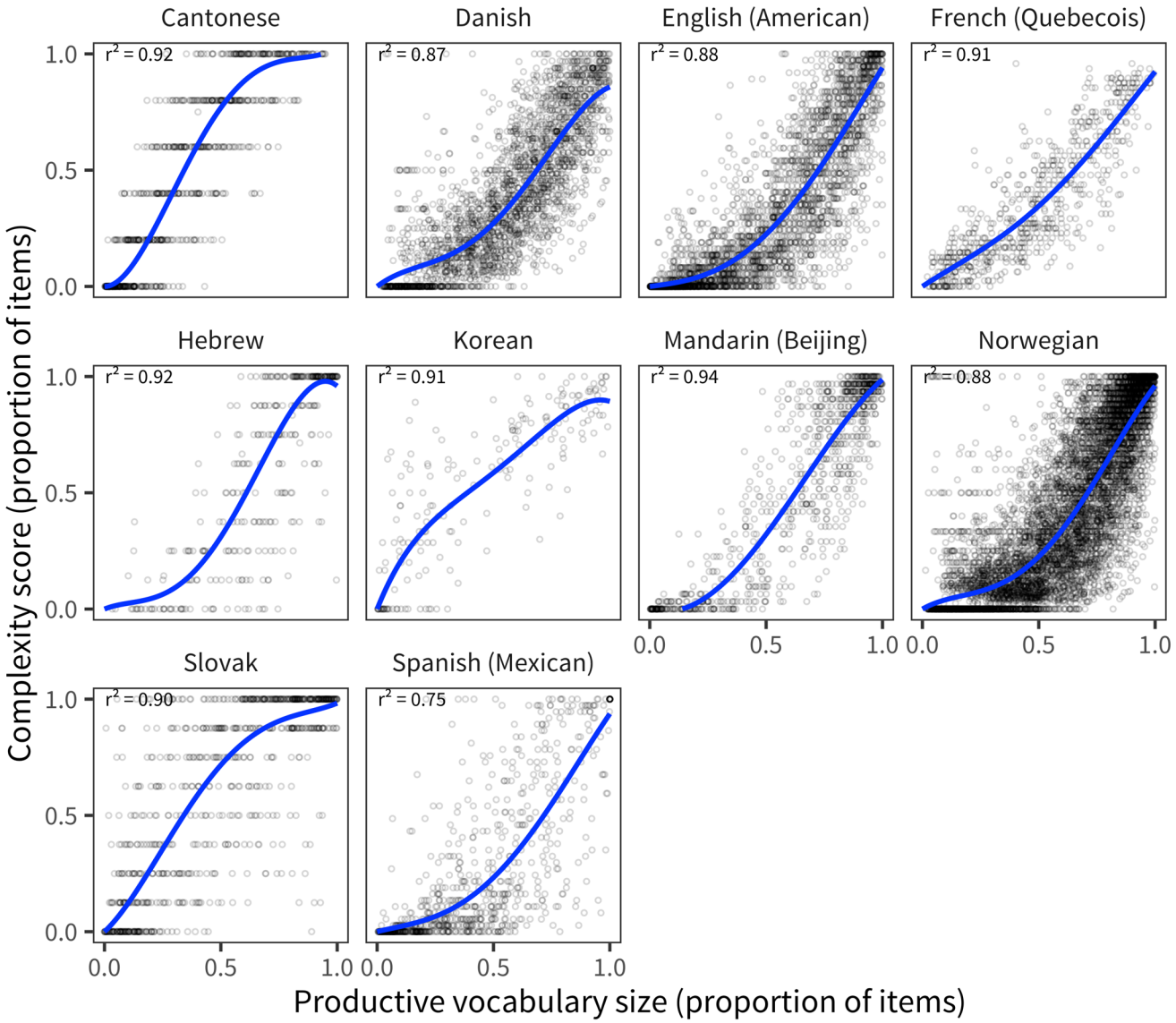
"That my truck" vs. "That's my truck"

"Baby crying" vs. "Baby crying cuz she's sad"

"Doggie table" vs. "Doggie on table"

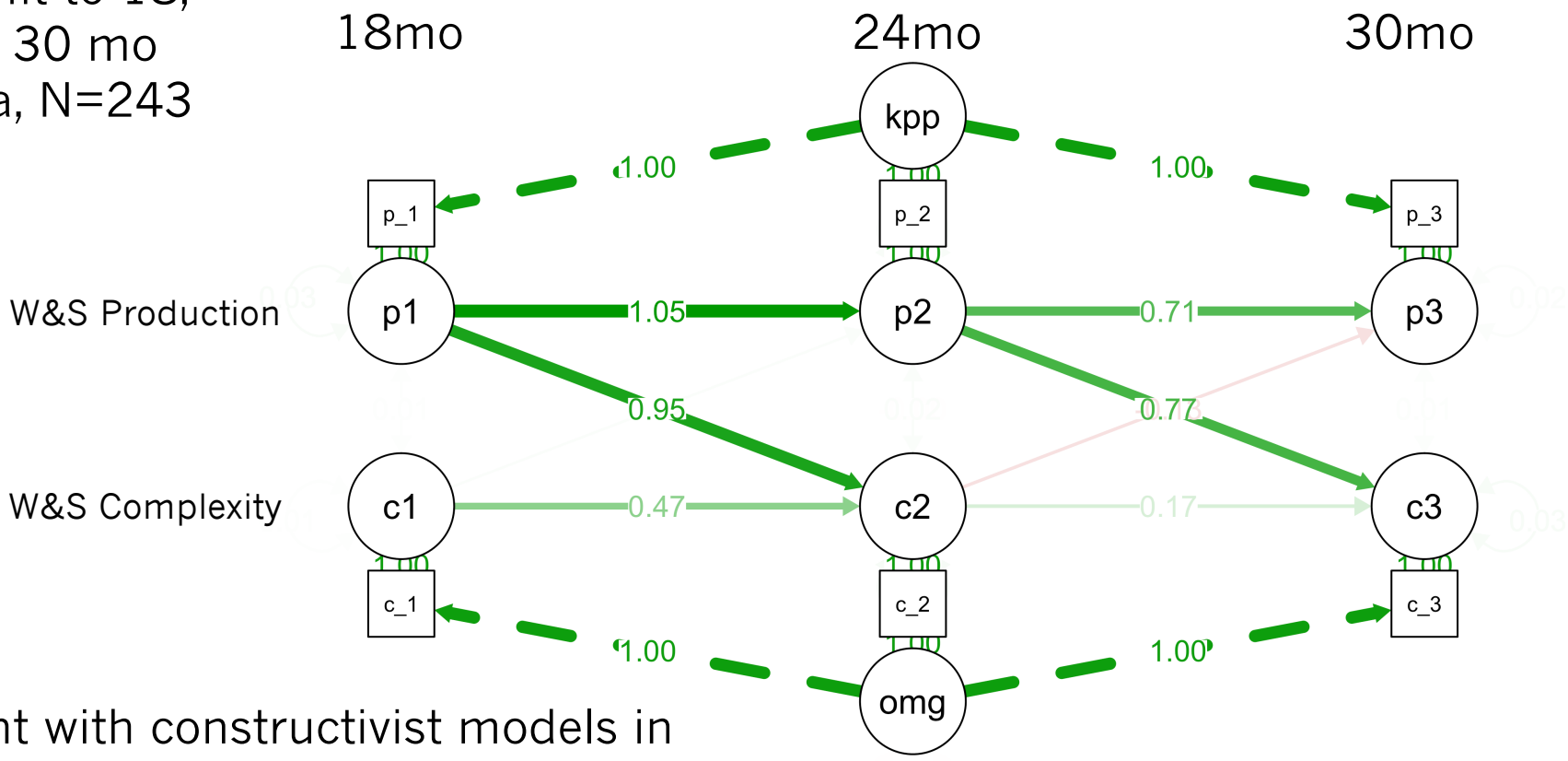
Bates & Goodman (1997)

Grammar and vocabulary



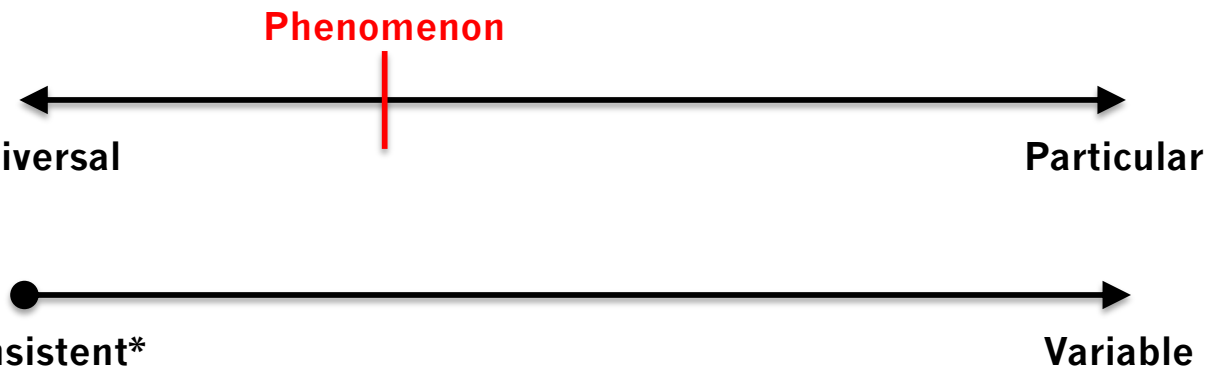
Grammar and Vocabulary

RI-CLPM fit to 18, 24, and 30 mo W&S data, N=243



Consistent with constructivist models in which grammar is generalization from individual items

What can we learn from cross-linguistic data?



Can assess variability quantitatively across phenomena

Aspects of language development that are invariant across languages and cultures

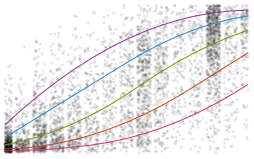
Aspects of language development that vary with context, culture, and language

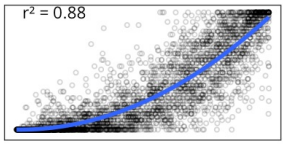
* Claims of cross-linguistic universality require a large, typologically diverse sample of languages

Variability and Consistency

Conclusions

- Early words are **consistent in content** across languages
- Children are **consistently variable** in their early language
- Grammar may emerge as a **generalization from vocabulary**
- Larger-scale datasets both reveal the need – and lay the foundations – for **quantitative theory**





Practicum: manipulating wordbank data!