Cognateness, frequency, and vocabulary size

An interactive account of bilingual lexical acquisition

Gonzalo Garcia-Castro

Daniela S. Ávila-Varela

Ignacio Castillejo

Núria Sebastian-Galles

Twitter: gongcastro

GitHub: gongcastro/isp_2023_trajectories

Universitat Pompeu Fabra Barcelona

Word acquisition

- Word learning is **challenging**: variability, ambiguity
- Earliest evidence of word acquisition: **6 months** of age
- **Bilingual word acquisition** is more complex: more than one word-form per referent (*gos* → DOG ← *perro*)
- Do bilinguals fall behind? **Mixed evidence** across language pairs: English-French, Catalan-Spanish, etc.

Jusczyk and Aslin (1995), Bergelson and Swingley (2012), Hoff et al. (2012)

Linguistic distance

Bilingual toddlers learning two languages that share more **cognates** show larger vocabulary sizes

Cognates
Form-similar translation equivalents (TEs)

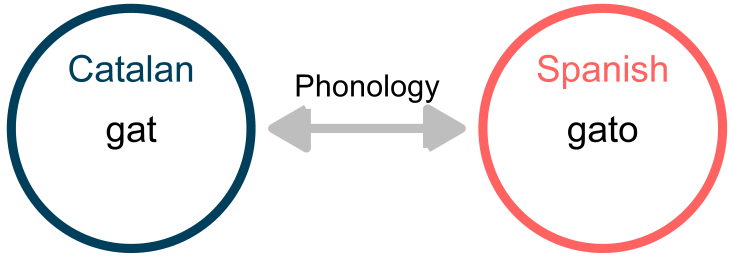
Cognate	*Non-cognate*
[cat] /ˈgat-ˈga.to/	[dog] /ˈgos-ˈpe.ro/

Cognates are acquired **earlier** than non-cognates. Why?

Floccia et al. (2018), Mitchell, Tsui, and Byers-Heinlein (2022), Bosch and Ramon-Casas (2014), Bilson et al. (2015)

Parallel activation: candidate mechanism?

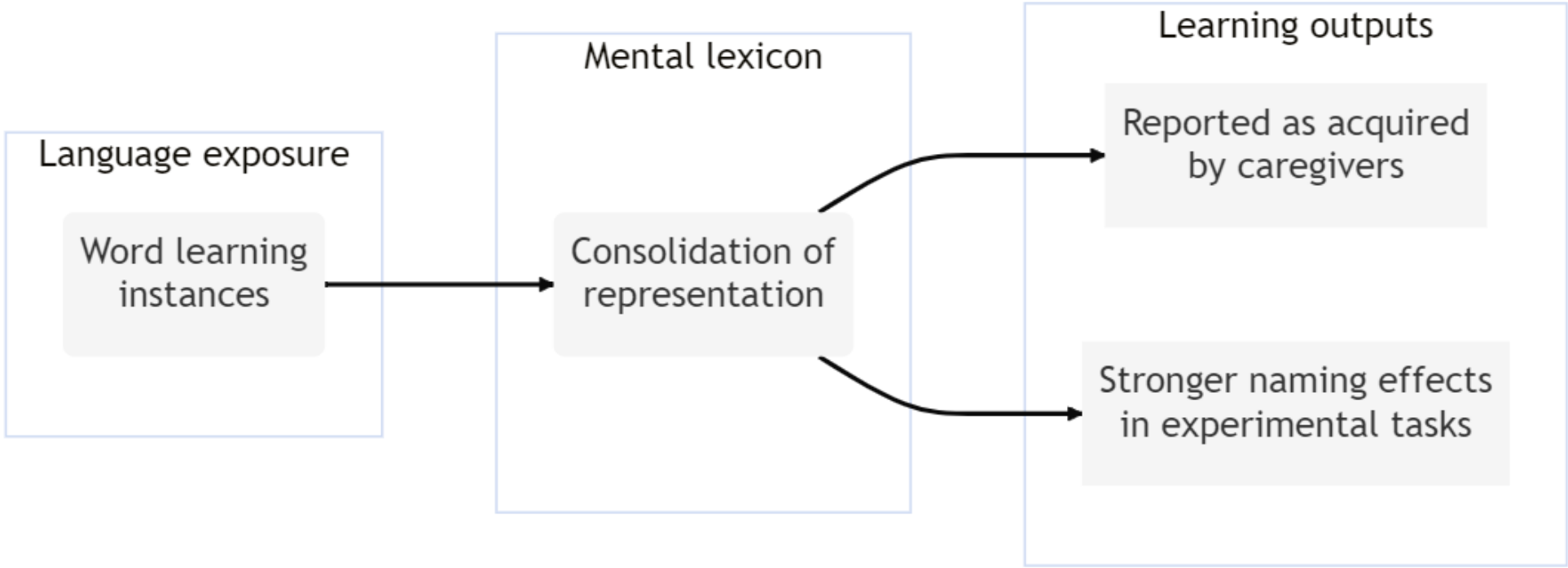
- Lexical access is **language non-selective**: both languages are co-activated, even in monolingual situations
- Cross-language activation across translation equivalents
- **Dissociation** between models of bilingual word *processing* and *word acquisition*



Spivey and Marian (1999), Costa, Caramazza, and Sebastian-Galles (2000)

Accumulator models

Word acquisition as a **continuous process** of **lexical consolidation**: accumulation of word **learning instances**

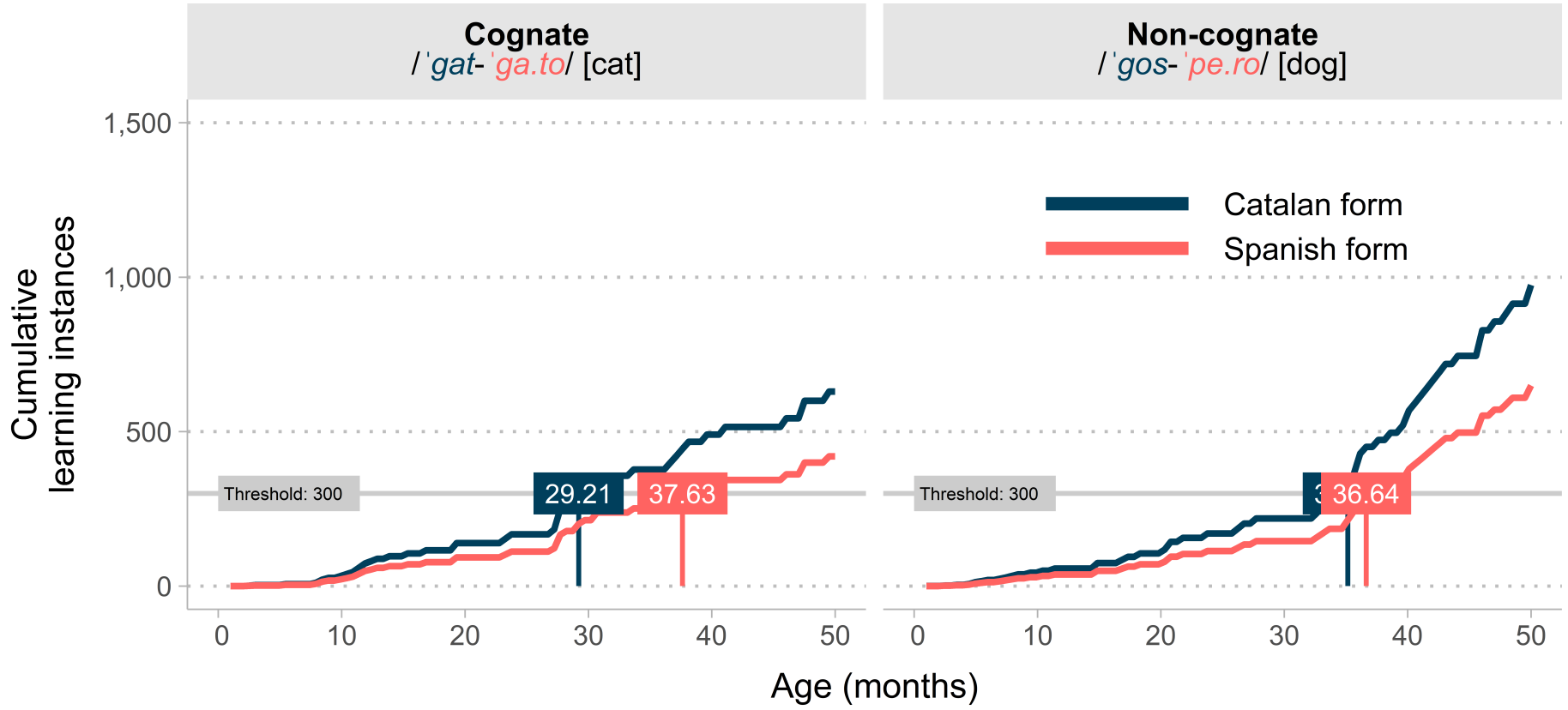


Hidaka (2013), Mollica and Piantadosi (2017)

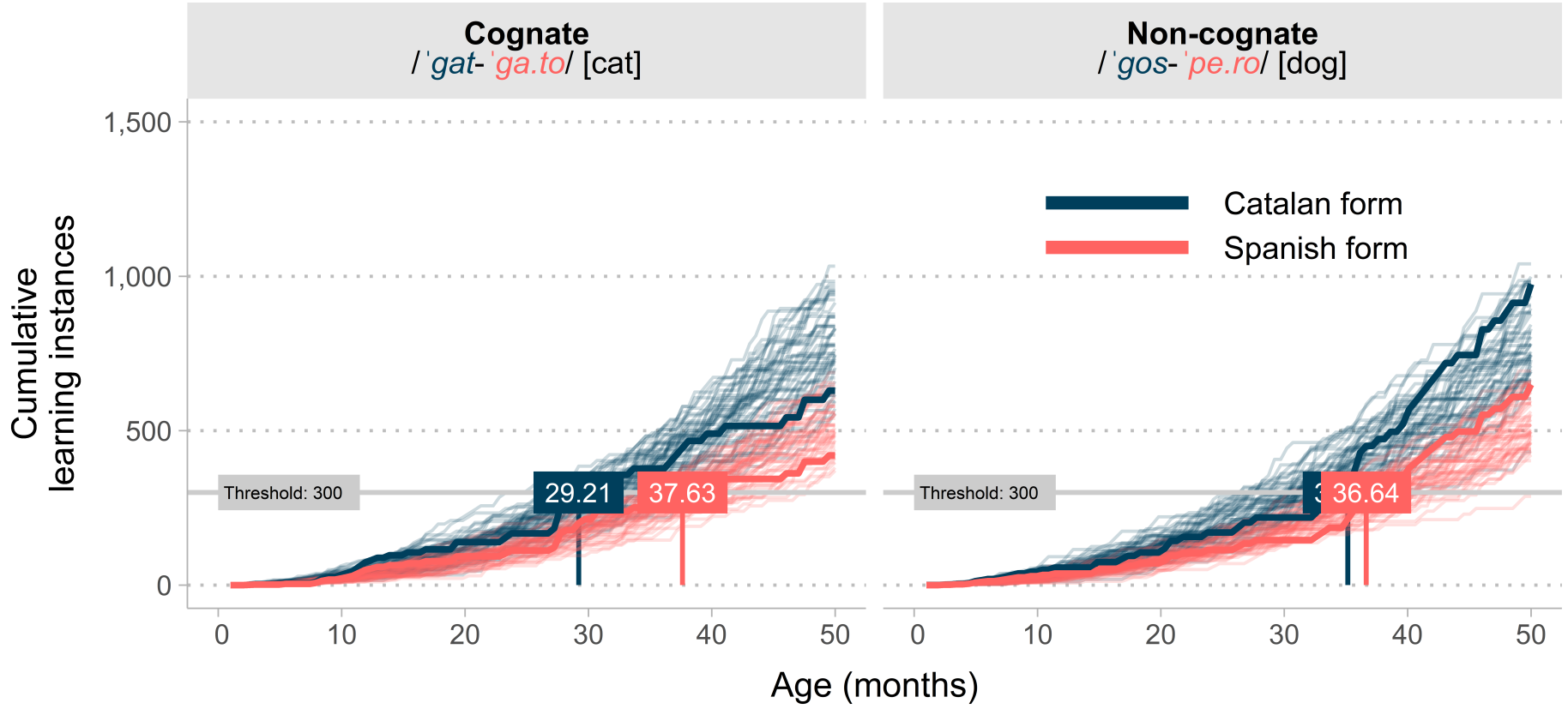
Simulating bilingual word acquisition

Simulation 1: **no parallel activation**

	Catalan	Spanish
Language exposure	60%	40%



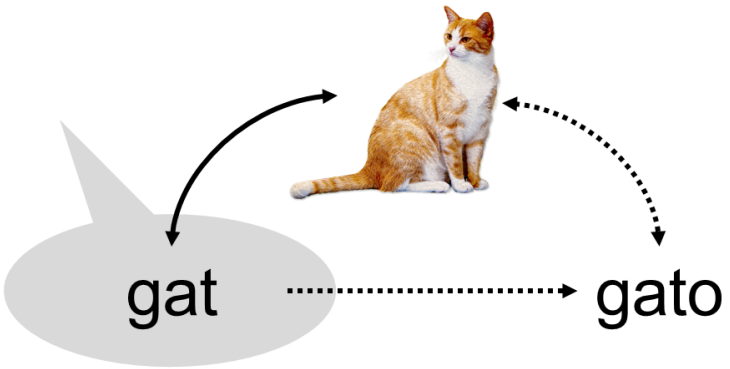
	Catalan	Spanish
Language exposure	60%	40%

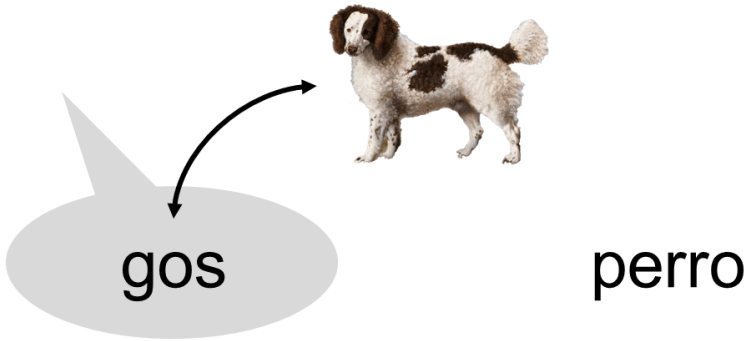


Simulation 2: **parallel activation**

Parallel activation hypothesis

Word-representations receive learning instances from their translations. This increment in learning instances is proportional to form-similarity (**cognateness**).

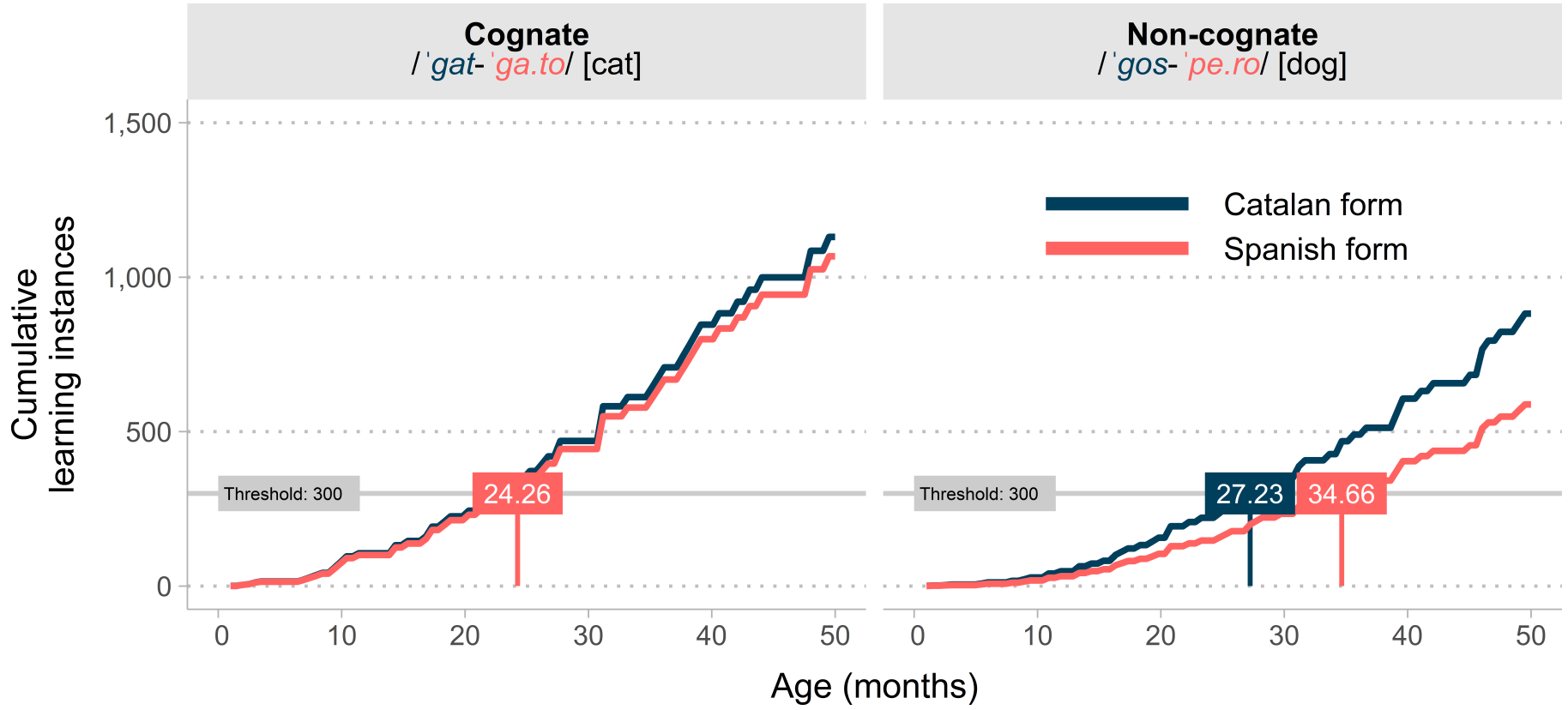




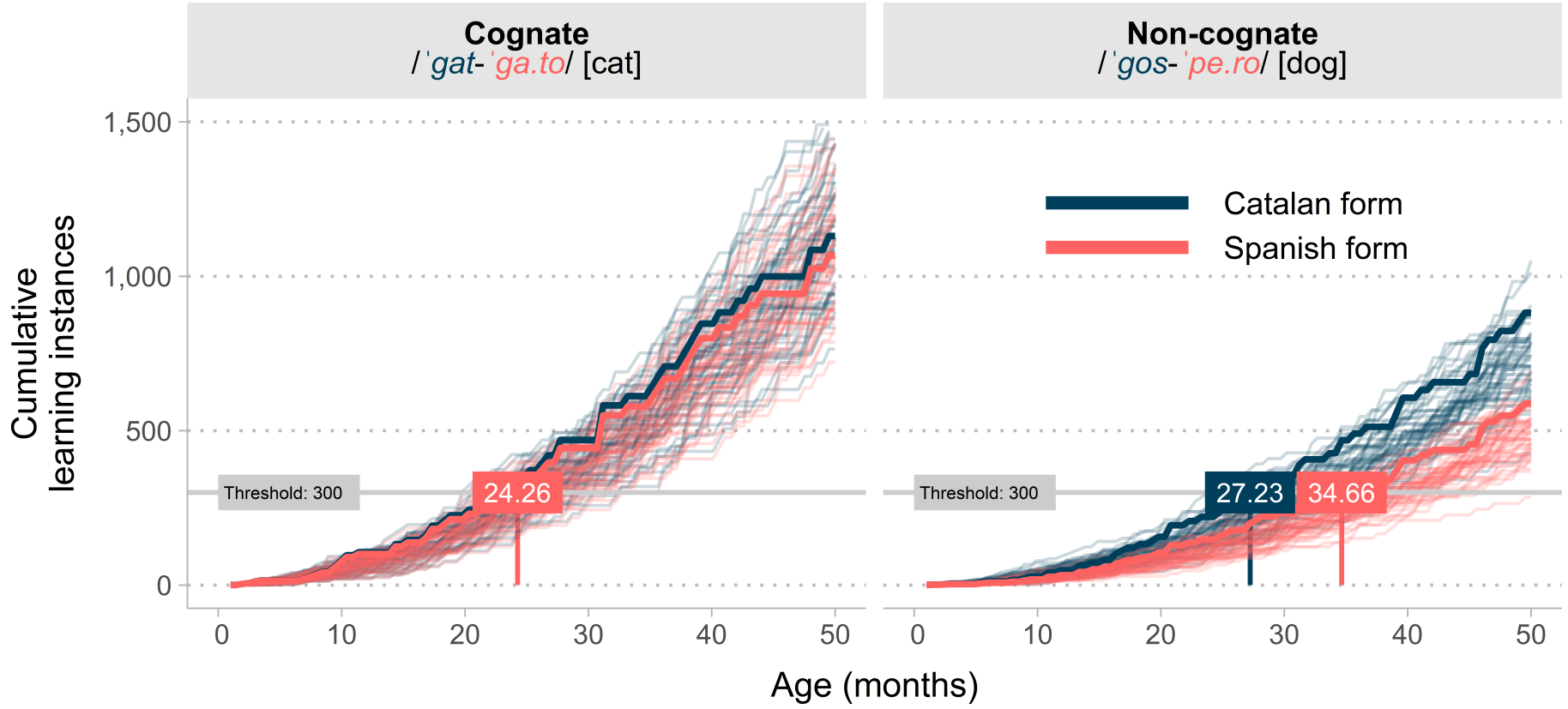
Cognate

Non-cognate

	Catalan	Spanish
Language exposure	60%	40%



	Catalan	Spanish
Language exposure	60%	40%



Testing hypotheses

Data collection

Questionnaire

Barcelona Vocabulary Questionnaire (BVQ)

- On-line, inspired by CDI
- 1,600 words: 800 Catalan and 800 Spanish (sub-lists of 500 words)
- Short-listed 302 (noun) translation pairs

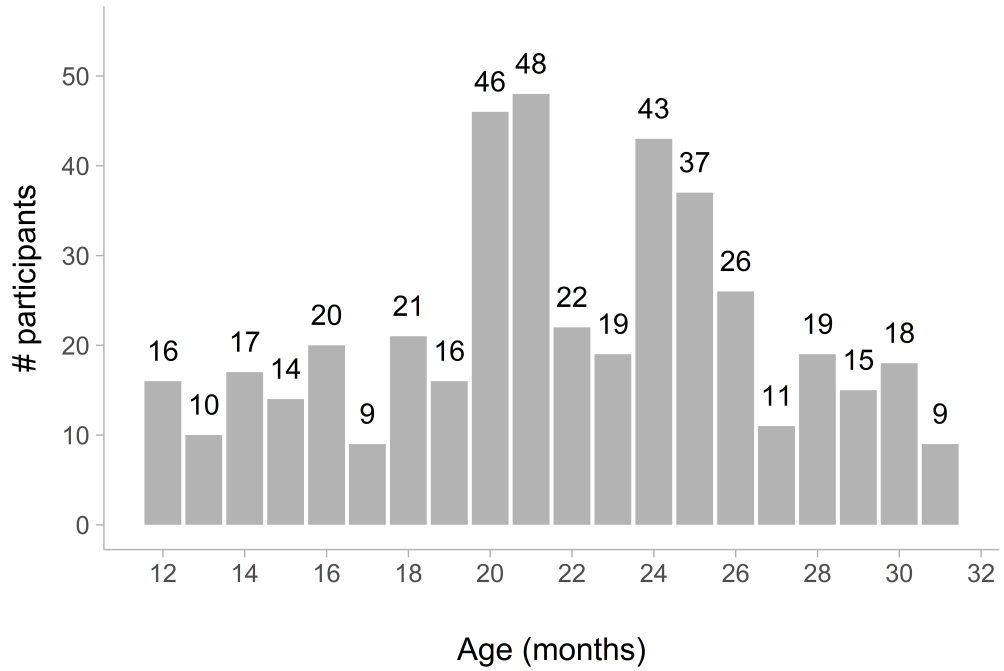
	Understands	Understands & Says
chair	[x]	[ ]
table	[ ]	[ ]
...	[ ]	[x]

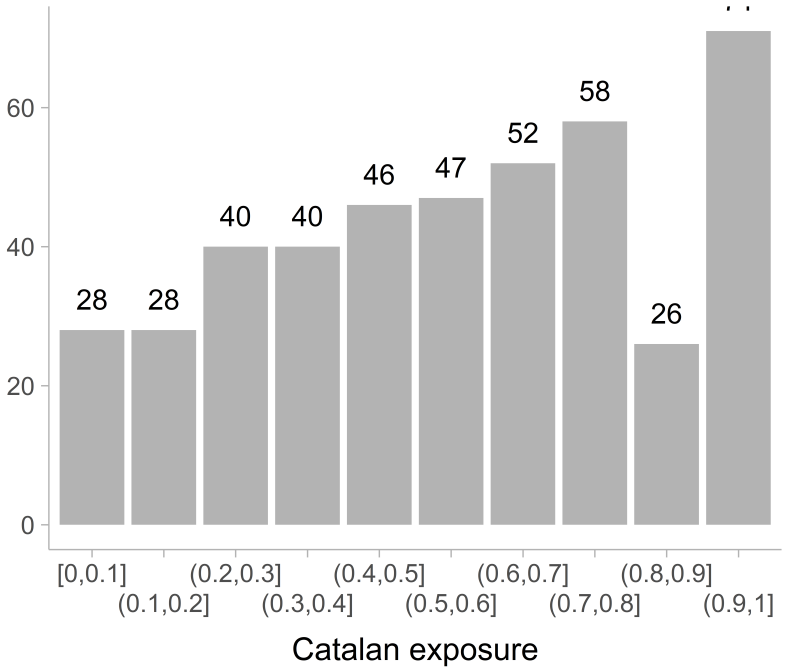
Fenson et al. (1994)

Participants

138,078 item responses from **366 Catalan–Spanish bilinguals**

1 time	2 times	3 times	4 times
312	42	8	4





Testing hypotheses

Modelling and statistical inference

Model

Multilevel, ordinal regression model:

- *No* < *Understands* < *Understands and Says*

Bayesian (`brms`/Stan): probability of parameter values

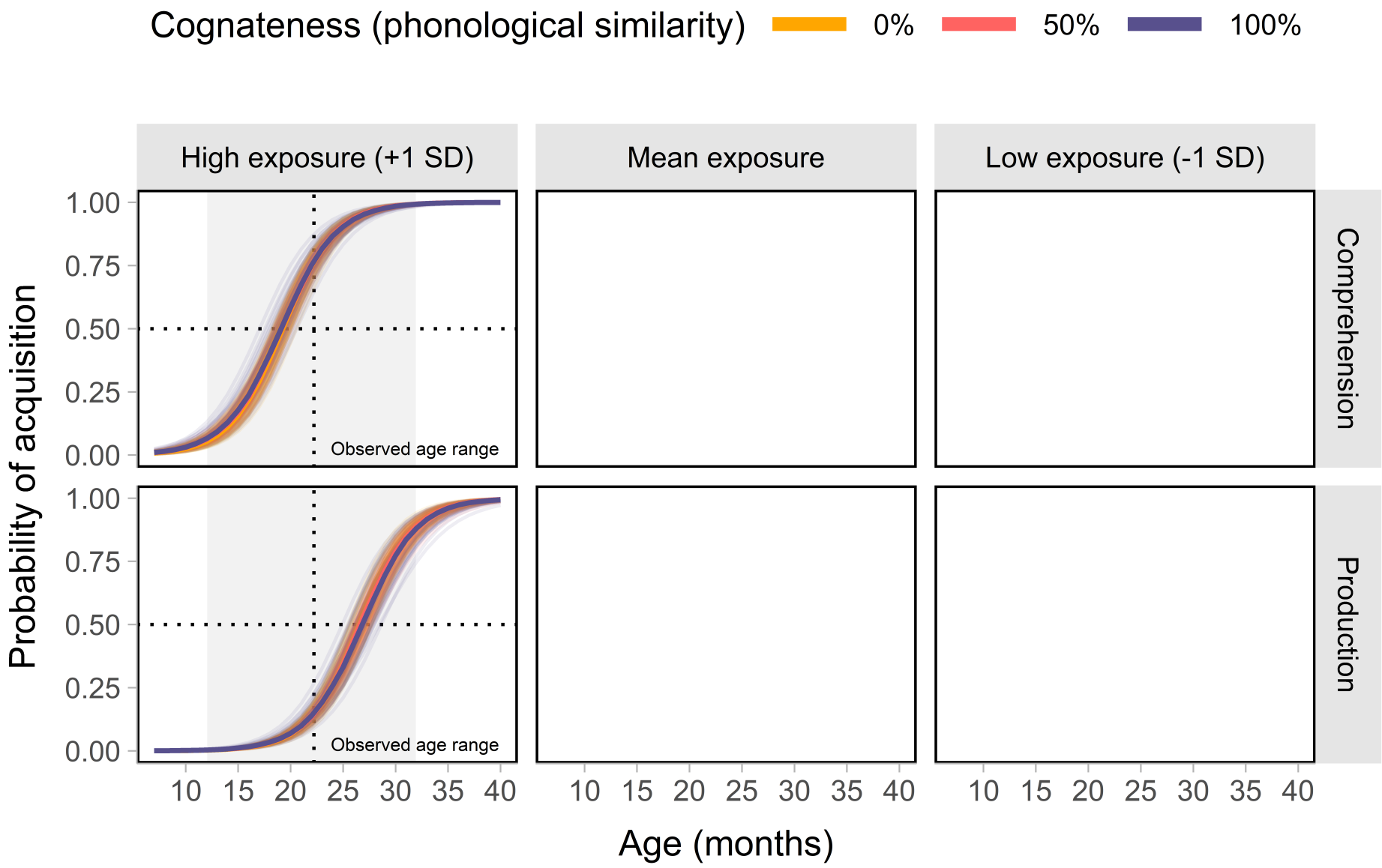
$$P(\text{model}|\text{data}) \propto P(\text{data}|\text{model}) \times P(\text{model})$$

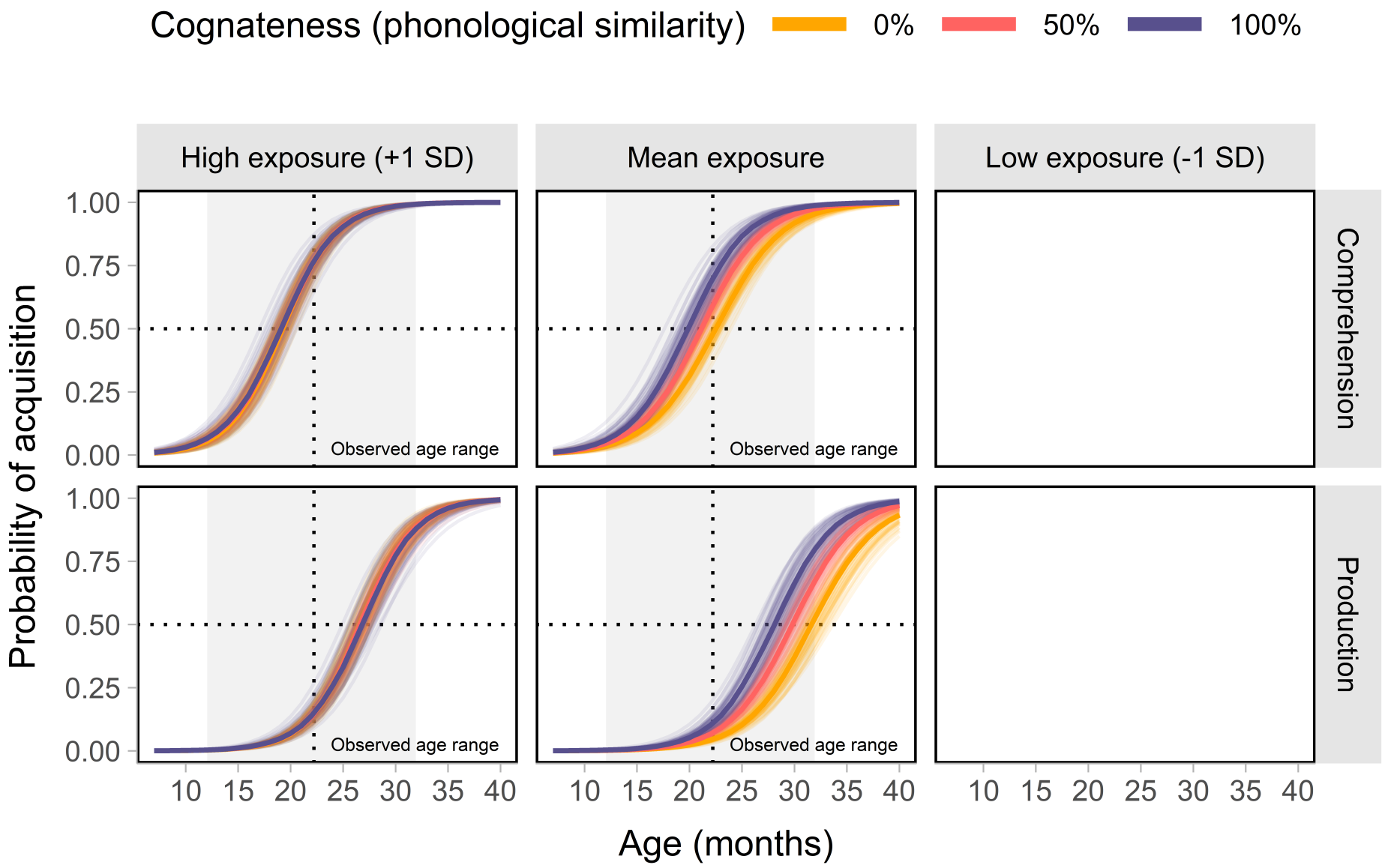
Predictors

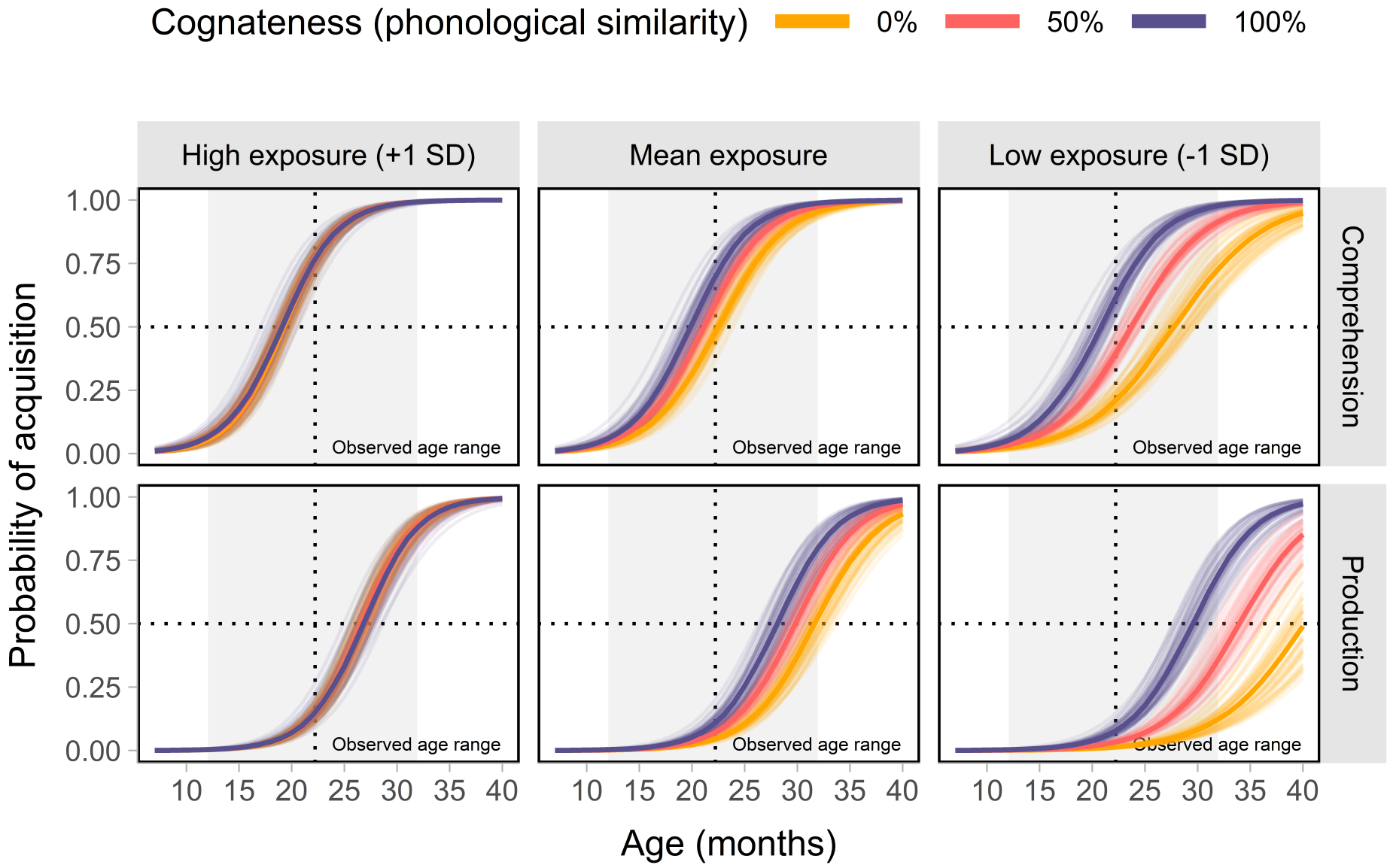
- **Age**
- **Length**: number of phonemes in word-form
- **Exposure**: lexical frequency $\times$ language exposure
- **Cognateness**: Levenshtein similarity between a word-form and its translation

Two-way and three-way **interactions** between age, exposure, and cognateness

Predictor	Estimate	95% HDI	p(H0)
Intercepts			
Comprehension and Production	0.438	[−0.5, 0.5]	0.088
Comprehension	0.936	[2.44, 0.95]	0.000
Slopes			
Age (+1 SD, 4.87, months)	0.405	[1.43, 0.45]	0.000
Exposure (+1 SD, 1.81)	0.233	[0.8, 0.27]	0.000
Cognateness (+1 SD, 0.26)	0.058	[0.06, 0.1]	0.037
Length (+1 SD, 1.56 phonemes)	−0.062	[−0.35, −0.04]	0.000
Age × Exposure	0.071	[0.16, 0.1]	0.000
Age × Cognateness	0.014	[0, 0.03]	0.985
Exposure × Cognateness	−0.057	[−0.28, −0.05]	0.000
Age × Exposure × Cognateness	−0.018	[−0.11, −0.01]	0.975


Cognateness (phonological similarity)
0%
50%
100%
High exposure (+1 SD)
Mean exposure
Low exposure (-1 SD)
Comprehension
Production
Probability of acquisition
1.00
0.75
0.50
0.25
0.00
Observed age range
10
15
20
25
30
35
40
Age (months)


Cognateness (phonological similarity)
0%
50%
100%
High exposure (+1 SD)
Mean exposure
Low exposure (-1 SD)
Comprehension
Production
Probability of acquisition
1.00
0.75
0.50
0.25
0.00
Observed age range
10
15
20
25
30
35
40
Age (months)


Cognateness (phonological similarity)
0%
50%
100%
High exposure (+1 SD)
Mean exposure
Low exposure (-1 SD)
Comprehension
Production
Probability of acquisition
1.00
0.75
0.50
0.25
0.00
Observed age range
10
15
20
25
30
35
40
Age (months)

Discussion

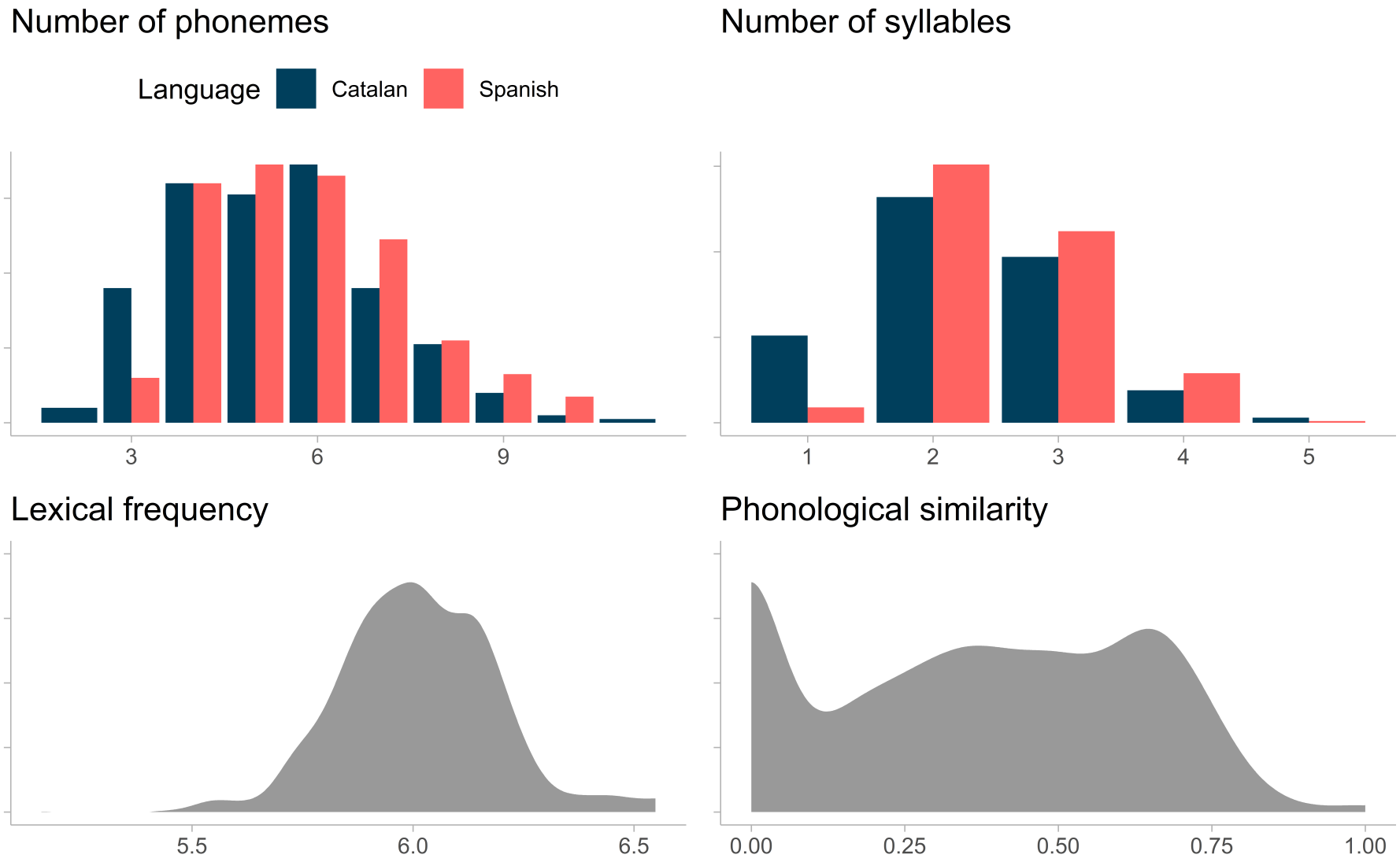
- **Cognateness facilitates word acquisition**
- Only **low-exposure** words bene t from their cognate status: less dominant language receives more facilitation
- **Parallel activation** as mechanism that boosts lexical consolidation: increment in **cumulative learning instances**
- Next steps: word-learning, formalisation

Thanks!

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
ICREA
upf. Universitat Pompeu Fabra Barcelona

Appendix

Item properties



Levenshtein similarity

Levenshtein distance: number of edits for two character strings to become identical

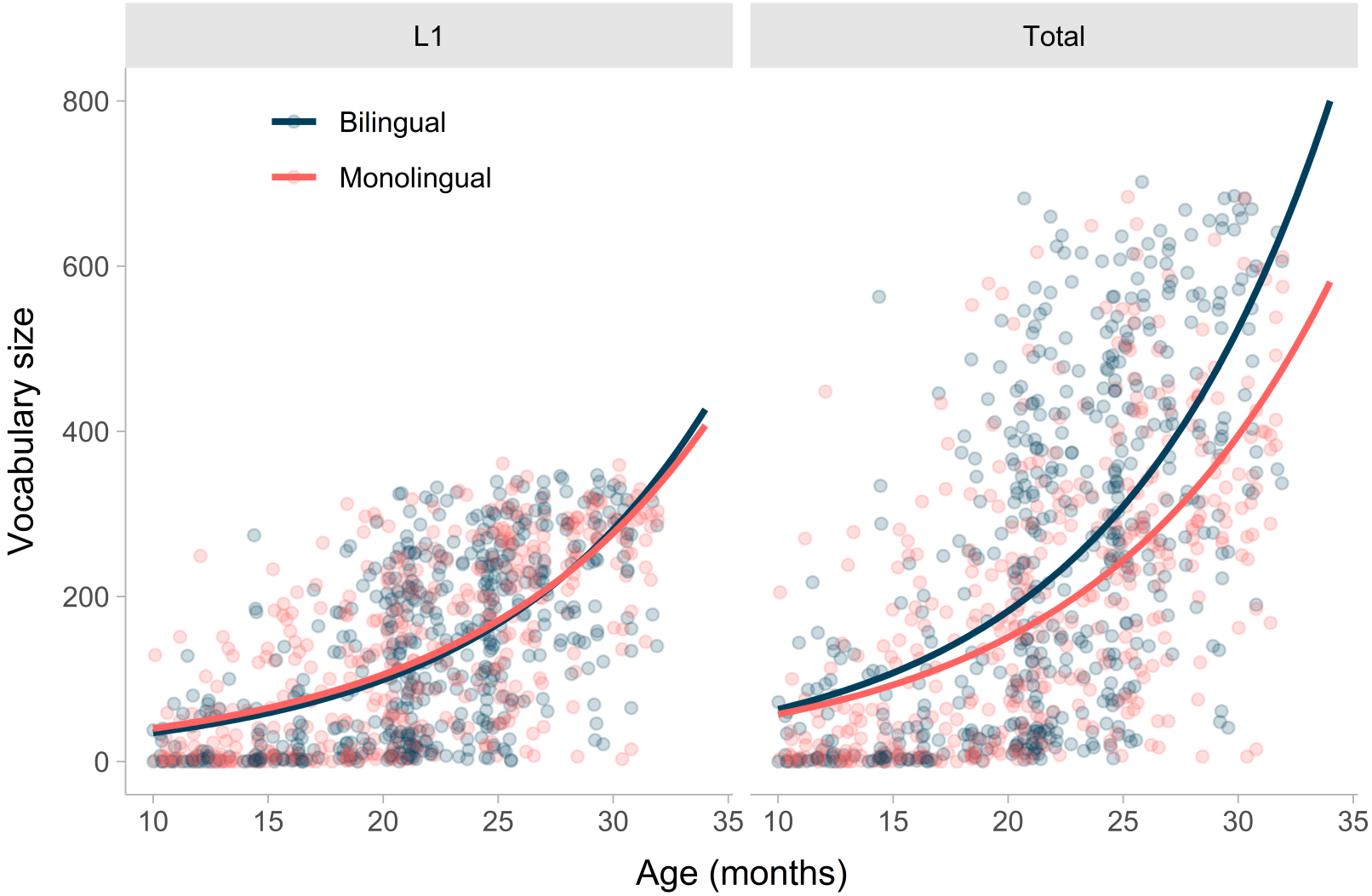
	Orthography	Phonology	String
Catalan	*porta*	/ˈpɔɾ.tə/	`pɔɾtə`
Spanish	*puerta*	/ˈpweɾ.ta/	`pweɾta`

Levenshtein similarity

$$1 - \frac{lev(A, B)}{Max(length(A), length(B))}$$

Catalan	Spanish	Levenshtein
porta (/ˈpɔɾ.tə/)	puerta (/ˈpweɾ.ta/)	0.50 (3)
taula (/ˈtaw.lə/)	mesa* (/ˈmesa/)	0.00 (5)
cotxe (/ˈkɔ.t͡ʃə/)	coche (/ˈkot͡ʃe/)	0.40 (3)
…	…	…

Vocabulary

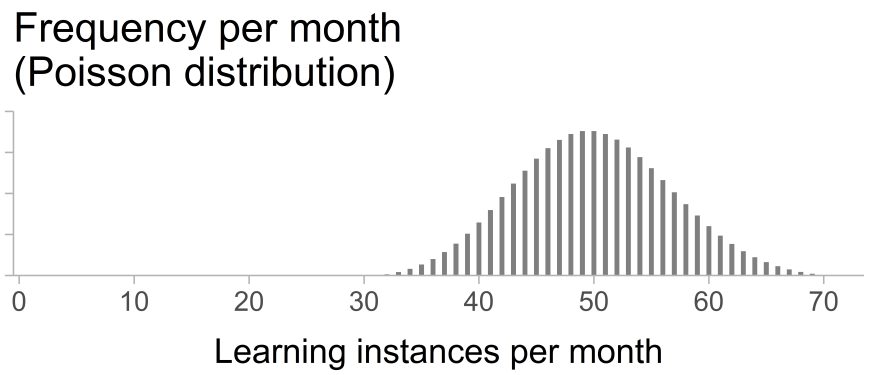

L1
Total
Bilingual
Monolingual
Vocabulary size
800
600
400
200
0
10
15
20
25
30
35
Age (months)

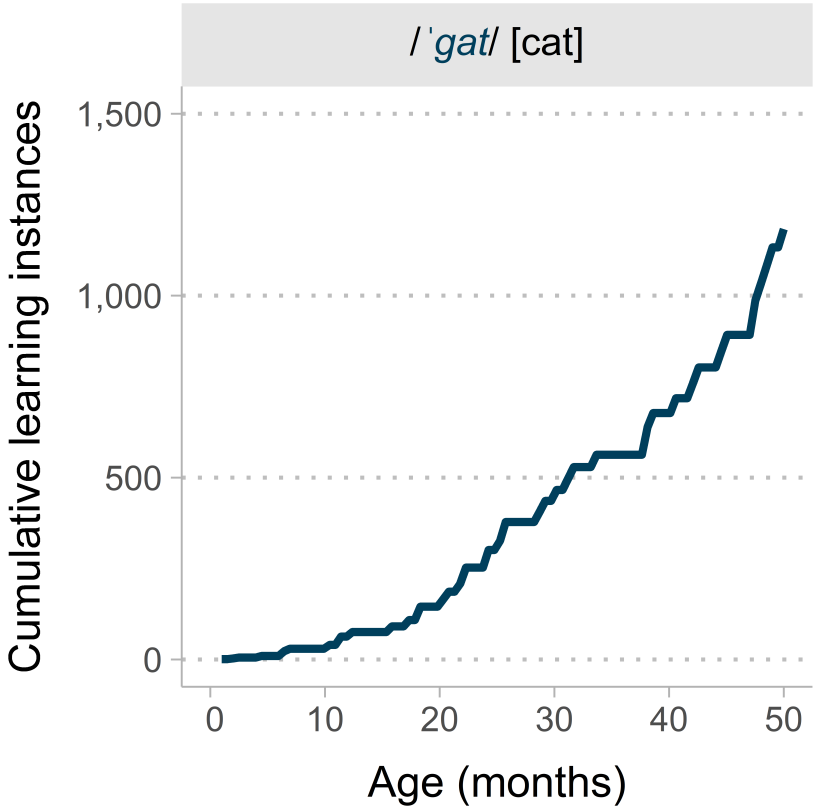
Simulation: monolinguals

For participant i and word j:

$$\text{Learning instances}_{ij} = \text{Age}_i \cdot \text{Frequency}_j$$
$$\text{Frequency}_j \sim \text{Poisson}(\lambda)$$

For simulations: $\lambda = 50$



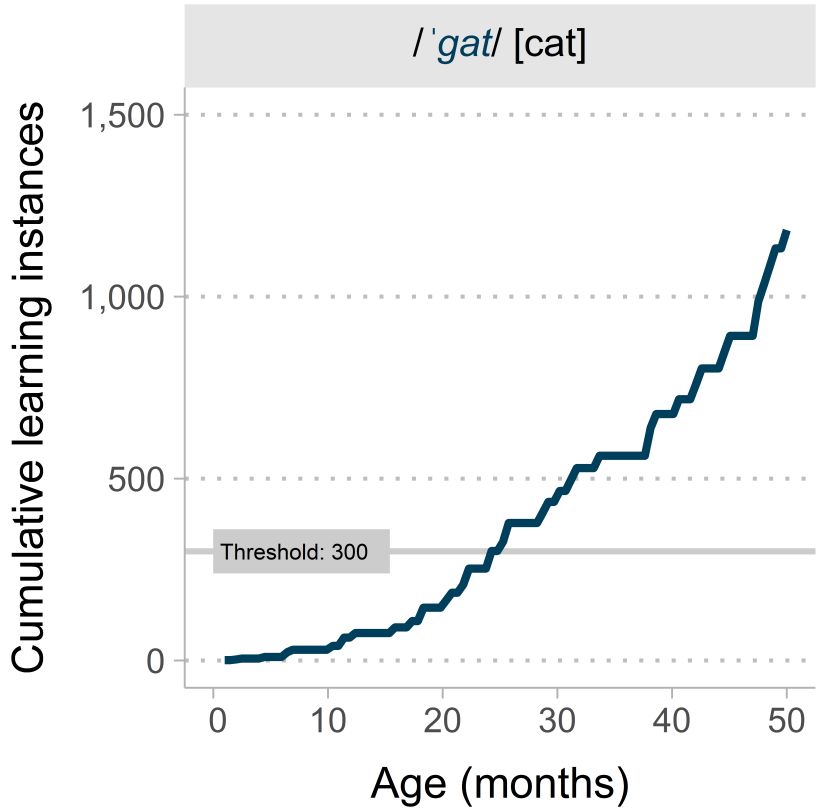


For participant i and word j:

$$\text{Learning instances}_{ij} = \text{Age}_i \cdot \text{Frequency}_j$$
$$\text{Frequency}_j \sim \text{Poisson}(\lambda)$$

For simulations:

$$\lambda = 50$$
$$\text{Threshold} = 300$$



For participant i and word j:

$$\text{Learning instances}_{ij} = \text{Age}_i \cdot \text{Frequency}_j$$

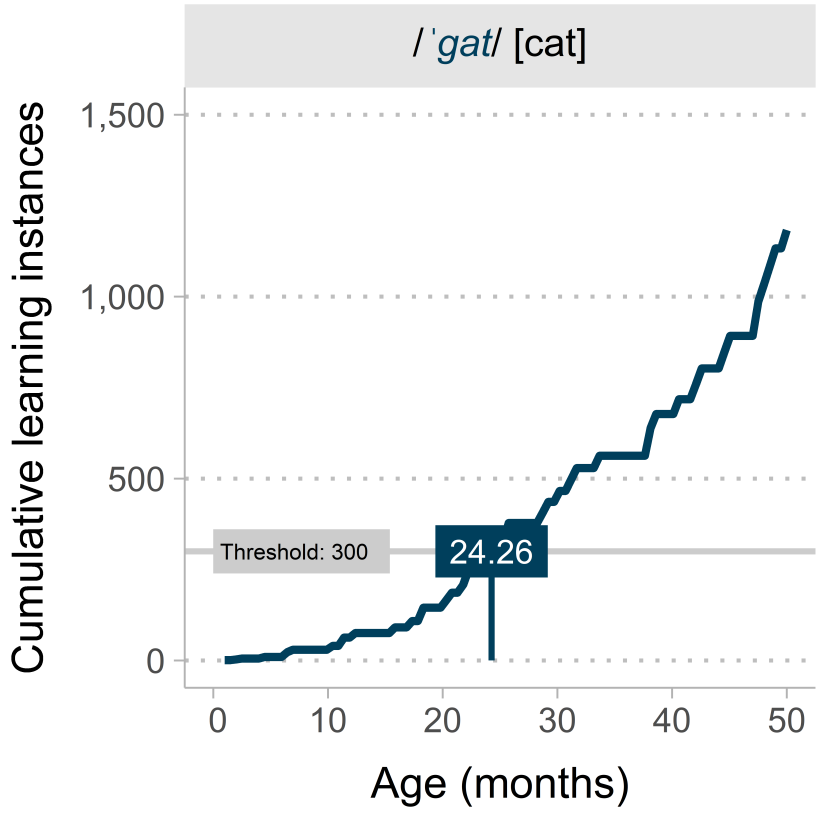
$$\text{Frequency}_j \sim \text{Poisson}(\lambda)$$

For simulations:

$$\lambda = 50$$

$$\text{Threshold} = 300$$

$$\text{Age of Acquisition}_{ij} = \text{Age}_{i\ [\text{Threshold}]}$$



For participant i and word j:

$$\text{Learning instances}_{ij} = \text{Age}_i \cdot \text{Frequency}_j$$

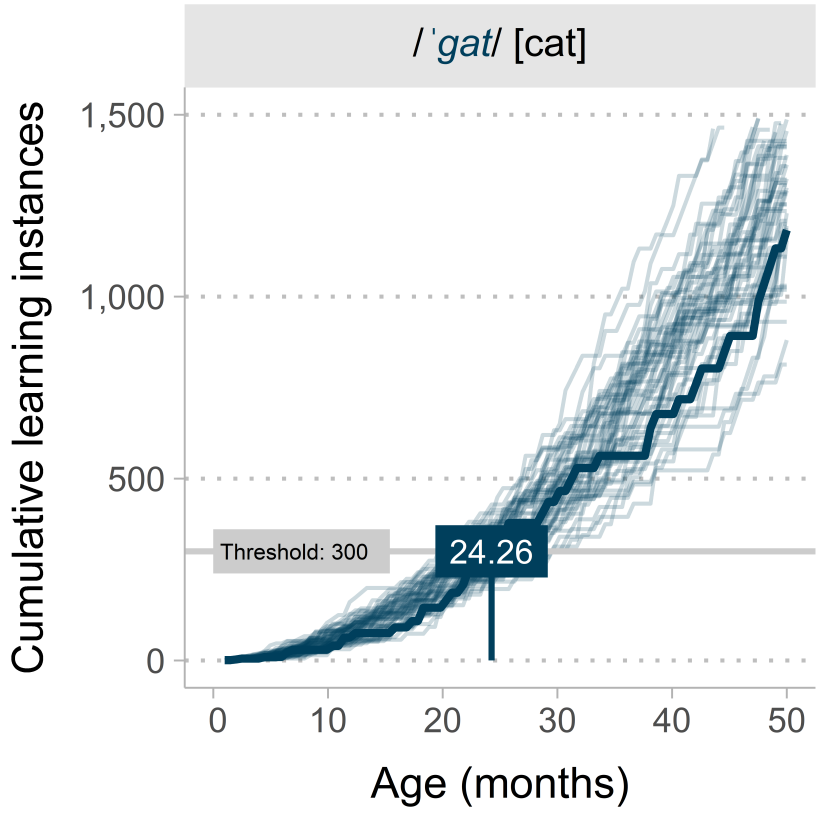
$$\text{Frequency}_j \sim \text{Poisson}(\lambda)$$

For simulations:

$$\lambda = 50$$

$$\text{Threshold} = 300$$

$$\text{Age of Acquisition}_{ij} = \text{Age}_{i\,[\text{Threshold}]}$$



Simulations: bilinguals

Including learning instances from **parallel activation**:

- **Hypothesis**: word-representations receive learning instances from their translations
- Proportional to the amount of form-similarity (**cognateness**)

Monolinguals:

$$\text{Learning instances}_{ij} = Age_i \cdot Frequency_j$$

Bilinguals:

$$\text{Learning instances}_{ij} = \text{Age}_i \cdot \text{Frequency}_j \cdot \text{Exposure}_i + (\text{Cognateness}_j \cdot \text{Learning instances}_{ij'})$$

References

Bergelson, Elika, and Daniel Swingley. 2012. "At 6–9 Months, Human Infants Know the Meanings of Many Common Nouns." *Proceedings of the National Academy of Sciences* 109 (9): 3253–58. https://doi.org/10.1073/pnas.1113380109.

Bilson, Samuel, Hanako Yoshida, Crystal D Tran, Elizabeth A Woods, and Thomas T Hills. 2015. "Semantic Facilitation in Bilingual First Language Acquisition." *Cognition* 140: 122–34.

Bosch, Laura, and Marta Ramon-Casas. 2014. "First Translation Equivalents in Bilingual Toddlers' Expressive Vocabulary: Does Form Similarity Matter?" *International Journal of Behavioral Development* 38 (4): 317–22. https://doi.org/10.1177/0165025414532559.

Costa, Albert, Alfonso Caramazza, and Nuria Sebastian-Galles. 2000. "The Cognate Facilitation Effect: Implications for Models of Lexical Access." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26: 1283–96. https://doi.org/10.1037/0278-7393.26.5.1283.

Fenson, Larry, Philip S. Dale, J. Steven Reznick, Elizabeth Bates, Donna J. Thal, Stephen J. Pethick, Michael Tomasello, Carolyn B. Mervis, and Joan Stiles. 1994. "Variability in Early Communicative Development." *Monographs of the Society for Research in Child Development* 59 (5): i–185. https://doi.org/10.2307/1166093.

Floccia, Caroline, Thomas D. Sambrook, Claire Delle Luche, Rosa Kwok, Jeremy Goslin, Laurence White, Allegra Cattani, et al. 2018. "I: Introduction." *Monographs of the Society for Research in Child Development* 83 (1): 7–29.