Why 72?

Charles Yang
University of Pennsylvania

NAACL 2018

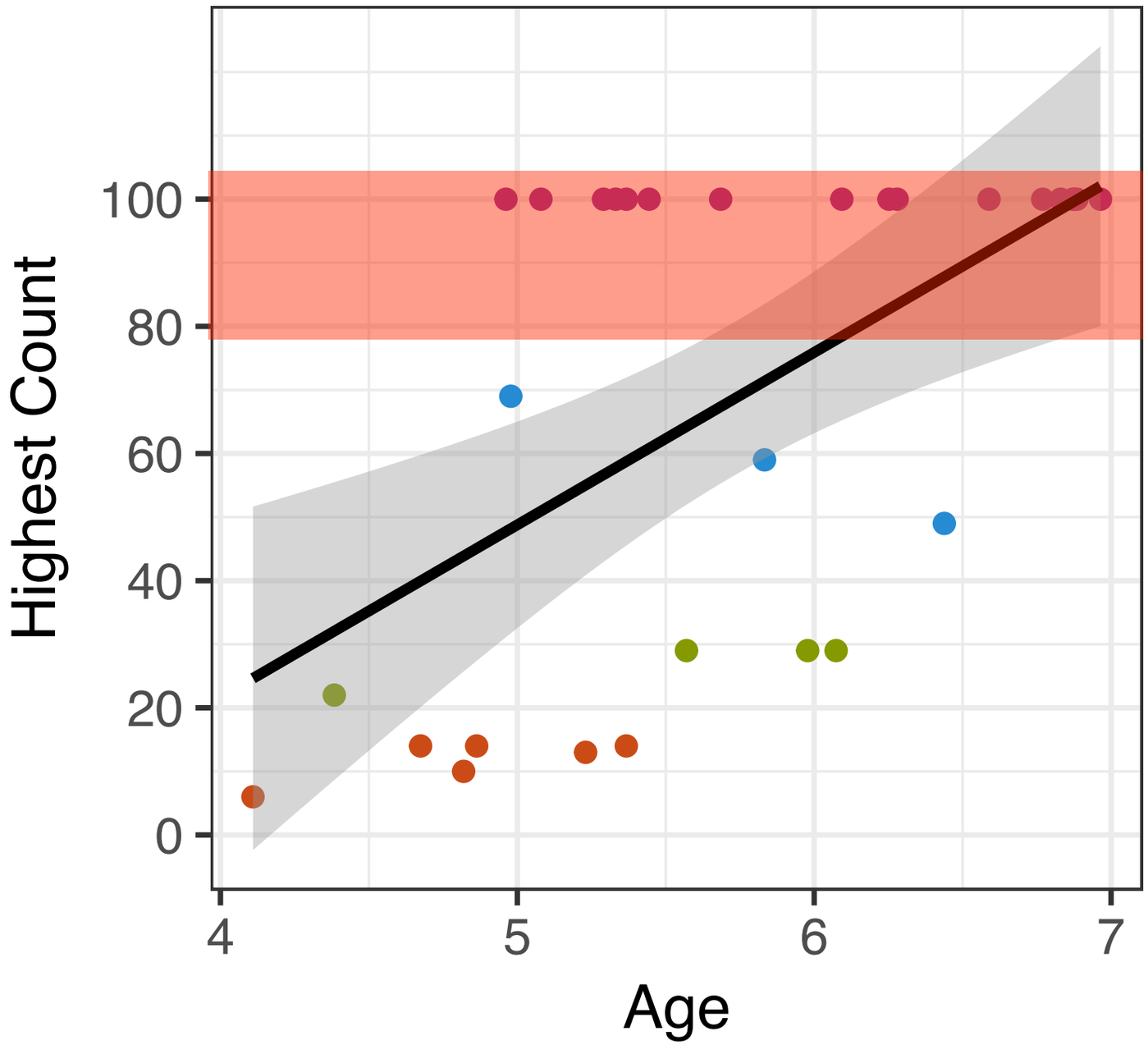
Our Family-Friendly Chair

Courtesy of Chris Callison-Burch

Age/grade[a]	$n < 10$	$10 \leqslant n < 14$	$14 \leqslant n < 20$	$20 \leqslant n < 30$	$30 \leqslant n < 72$	$72 \leqslant n < 101$	$101 \leqslant n < 201$	$201 \leqslant n$
3 years 6 months to 3 years 11 months	17	44	22	17	0	0	0	0
4 years to 4 years 5 months	0	41	35	12	12	0	0	0
4 years 6 months to 4 years 11 months	0	12	47	18	12	12	0	0
5 years to 5 years 5 months	0	6	25	13	44	13	0	0
5 years 6 months to 5 years 11 months	0	6	22	17	44	11	0	0
Kindergarten	0	7	11	30	26	4	22	0
First grade	0	0	3	14	7	21	48	7
Second grade	0	0	0	0	8	3	31	58
Third grade	0	0	0	0	0	4	25	71

Fuson, Richards, and Briars (1983)

“ … because children take off after 72.”
Karen Fuson (p.c., March 2016)



Rose Schneider (2016)

The Chinese Advantage

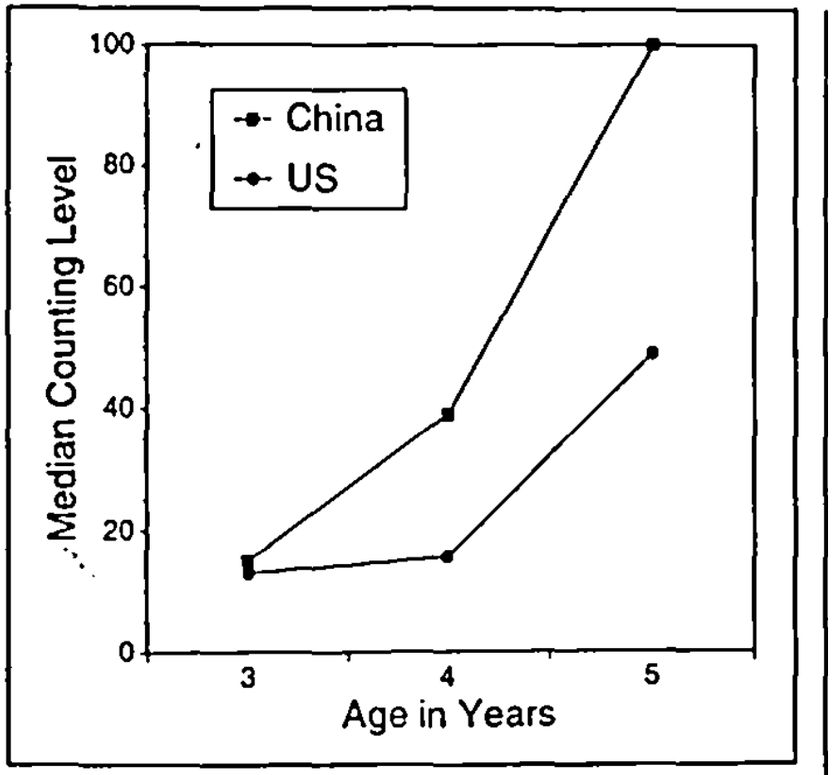


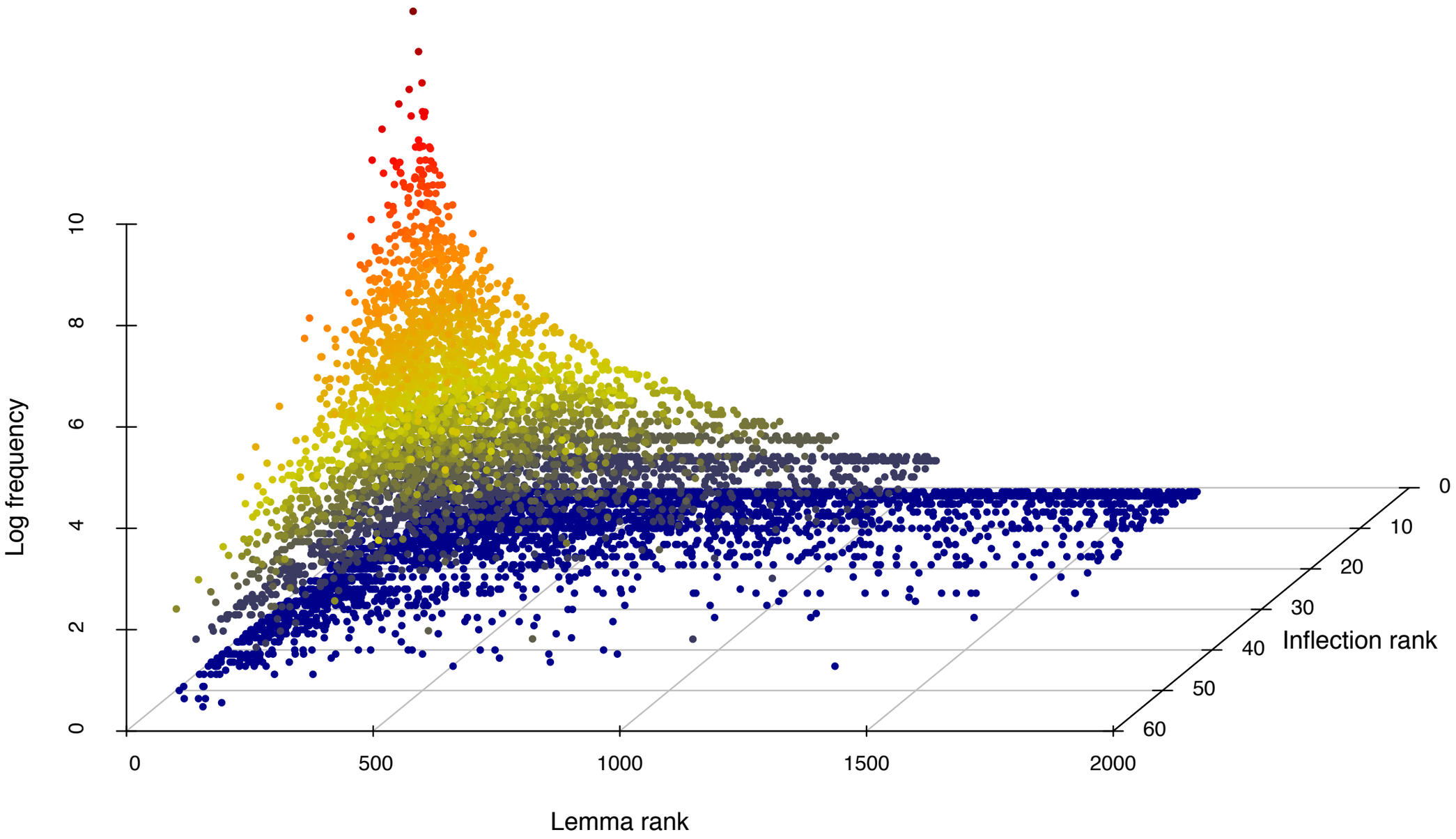
Fig. 1. Median level of abstract counting (highest number reached) by age and language. Significant differences favoring Chinese-speaking subjects were found at ages 4 and 5 years, but not at age 3.

"Four-year-olds in China made very rapid progress in generalizing number names up to 100 after they could count to approximately **40**" (Miller, Kelly, & Zhang 2005)

A Roadmap

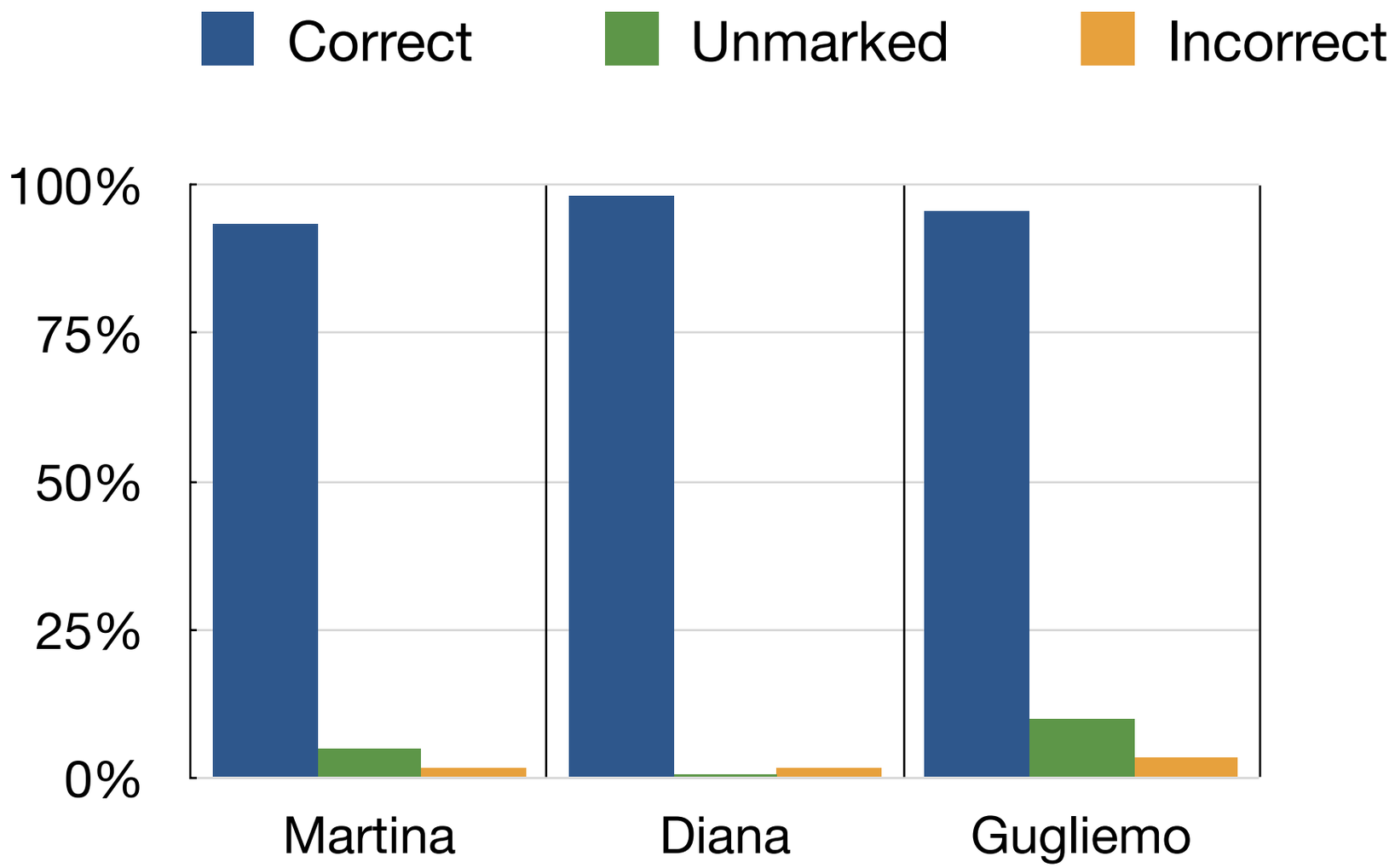
- How children learn the rules of languages
  - A proposal and an assortment of evidence
- What it can do to help develop unsupervised learning systems
  - Making the most out of very little data
- Why 72
  - Some connection between language and number

Sparsity of Input



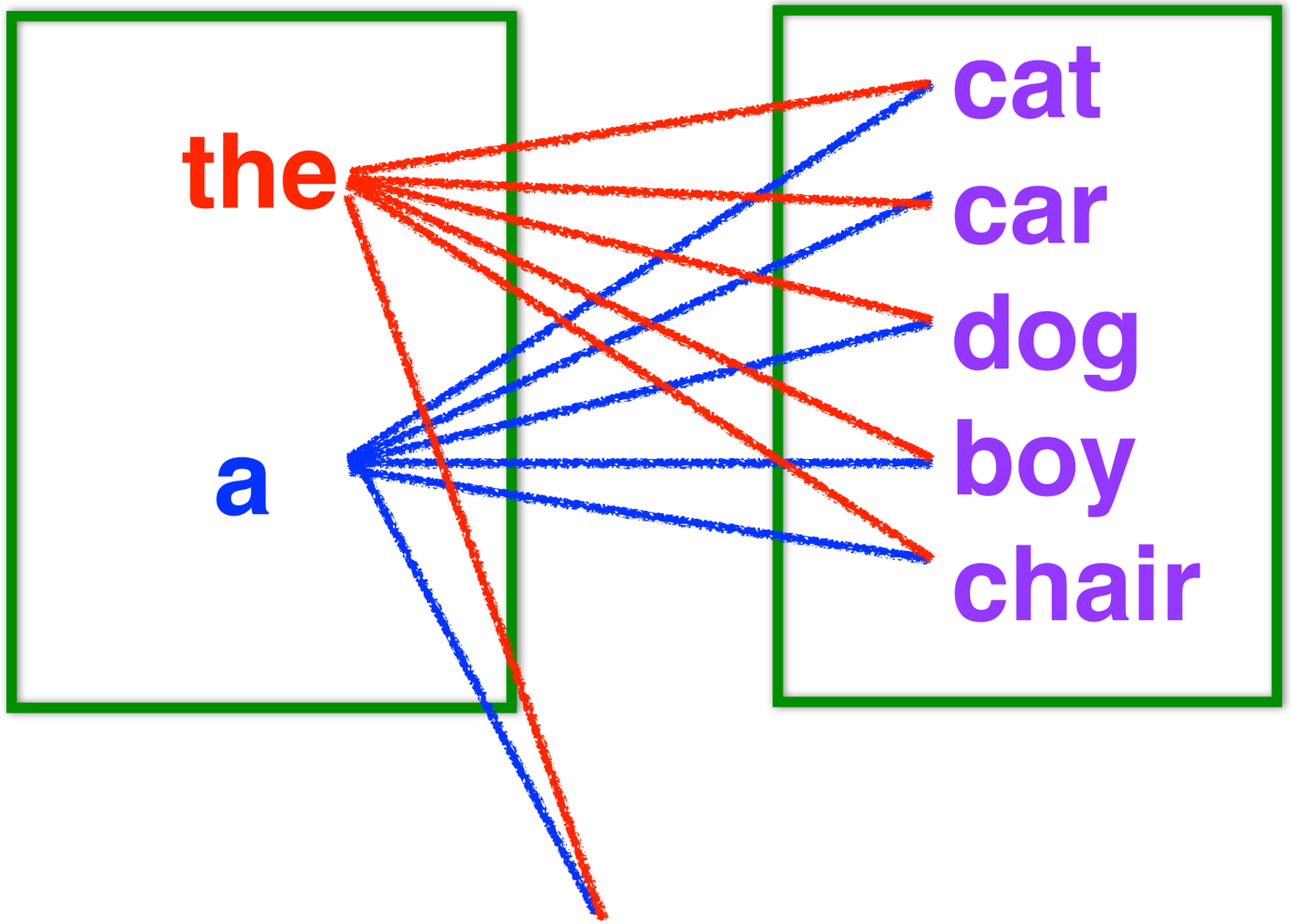
Courtesy of Erwin Chan and Constantine Lignos

Richness of Output



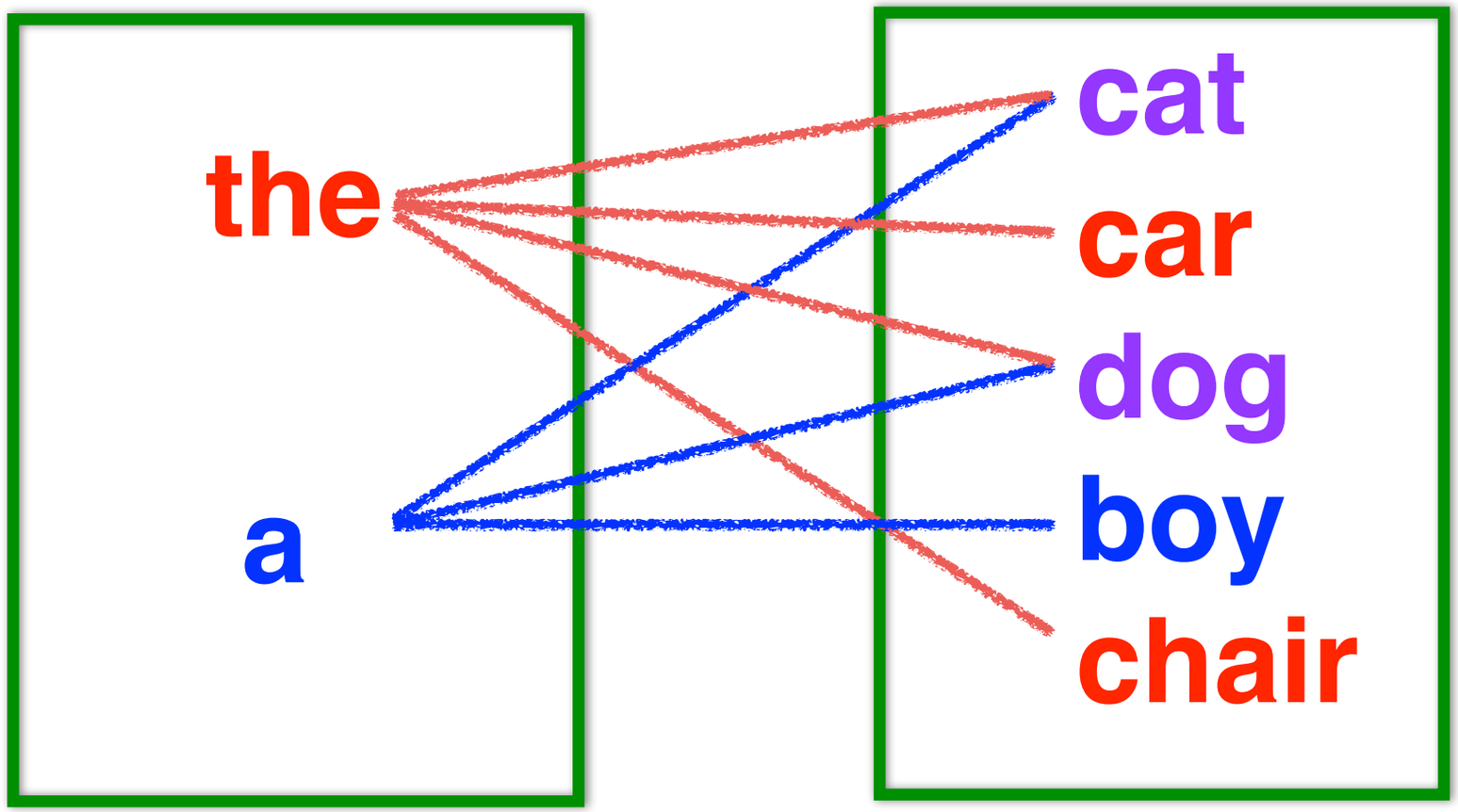
From Guasti (1992, *Language Acquisition*)

A noun ⟺ The noun



hallmark of human language

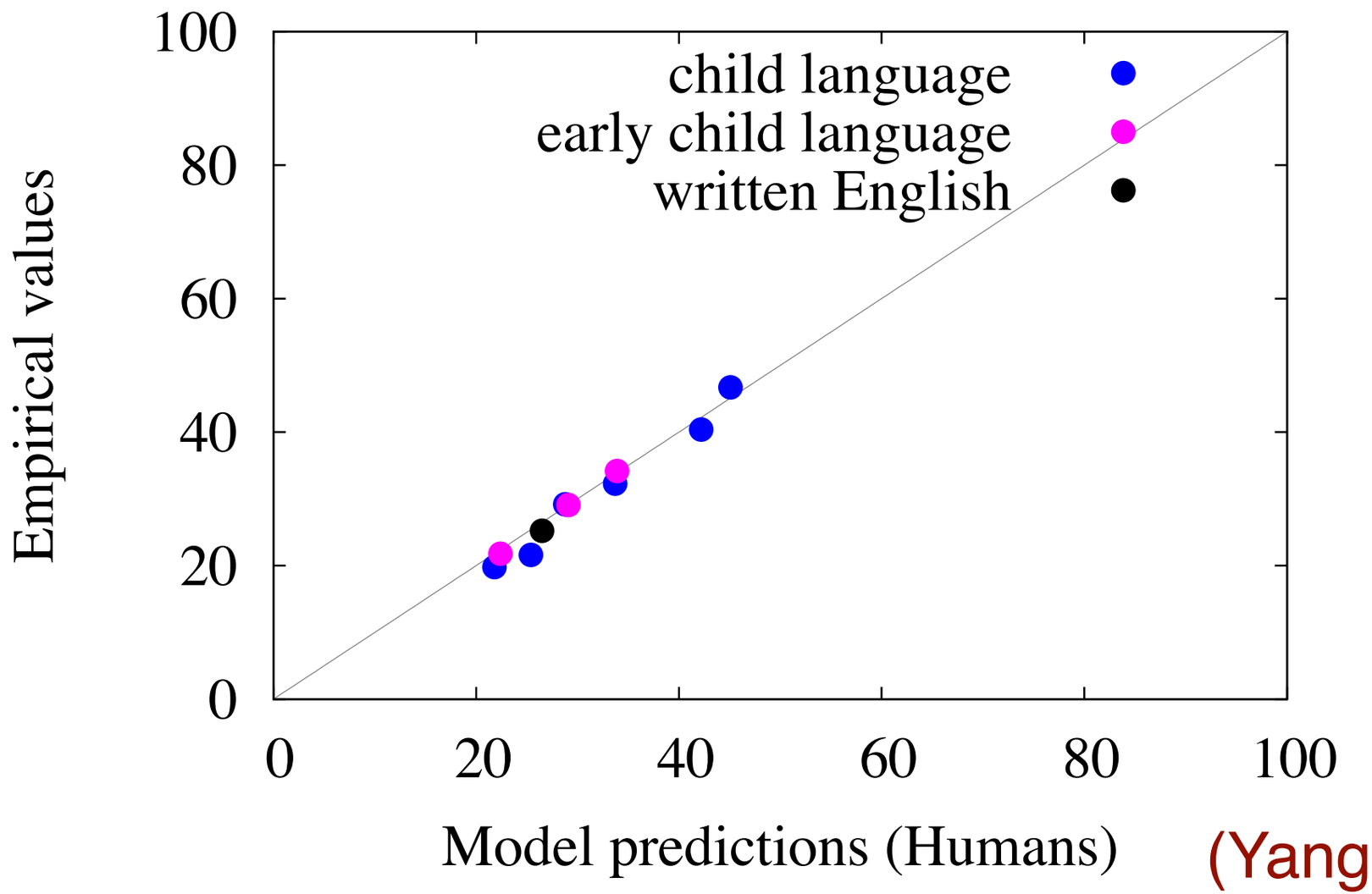
What children say



2/5

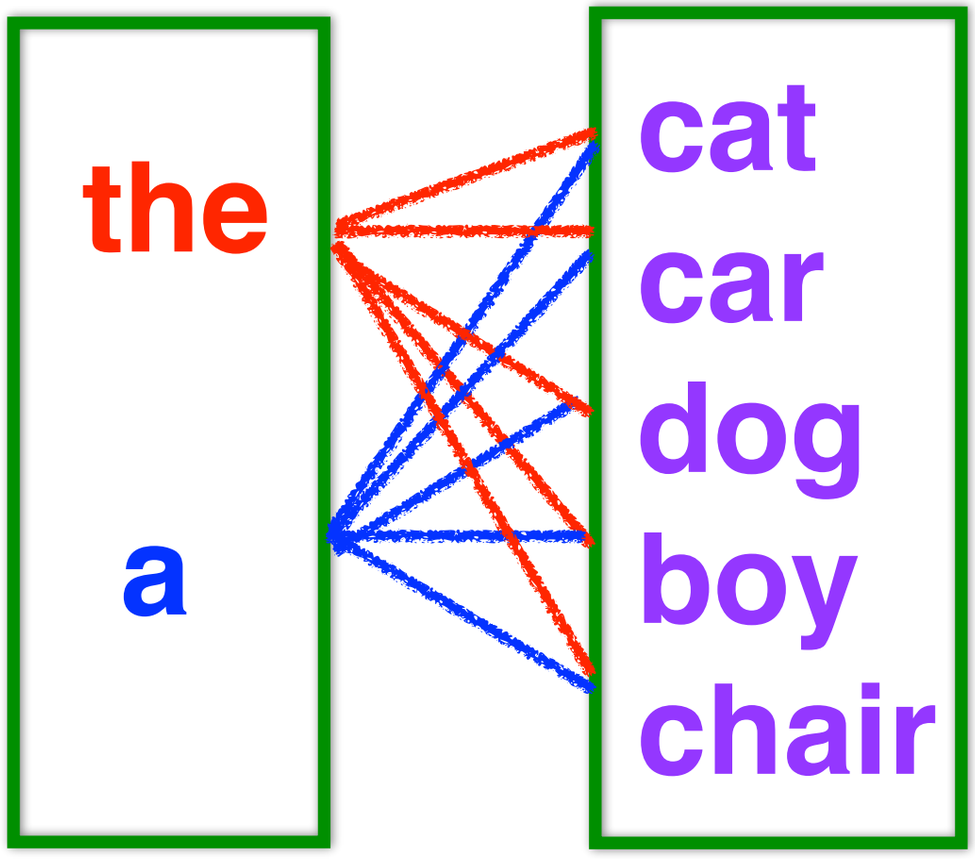
Statistics of Grammar

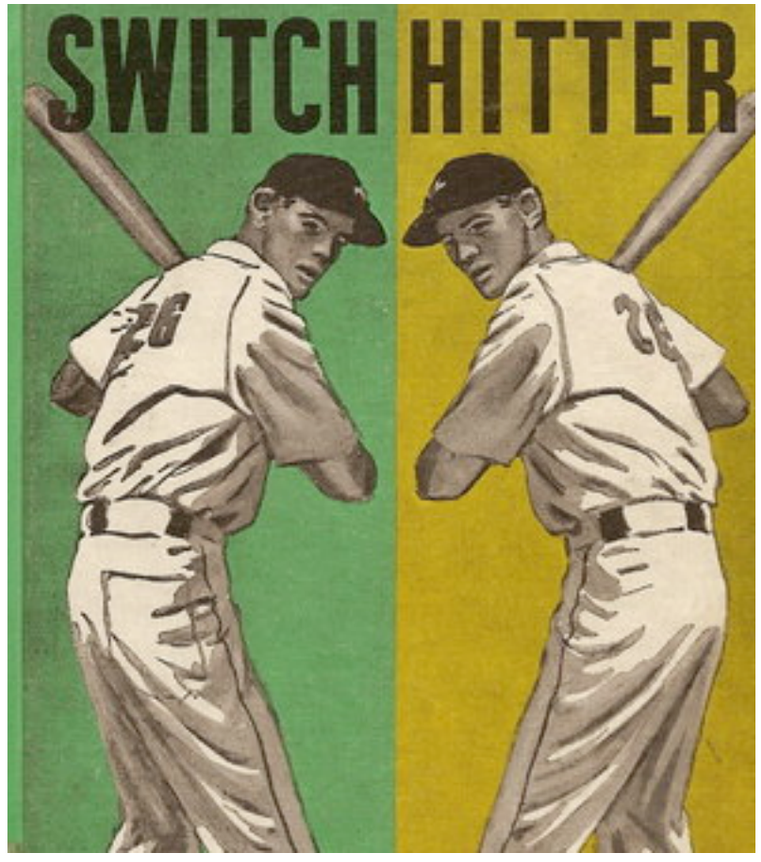
A baseline: assume independence and multiply marginal probabilities but take sparsity into account



(Yang 2013, *PNAS*)

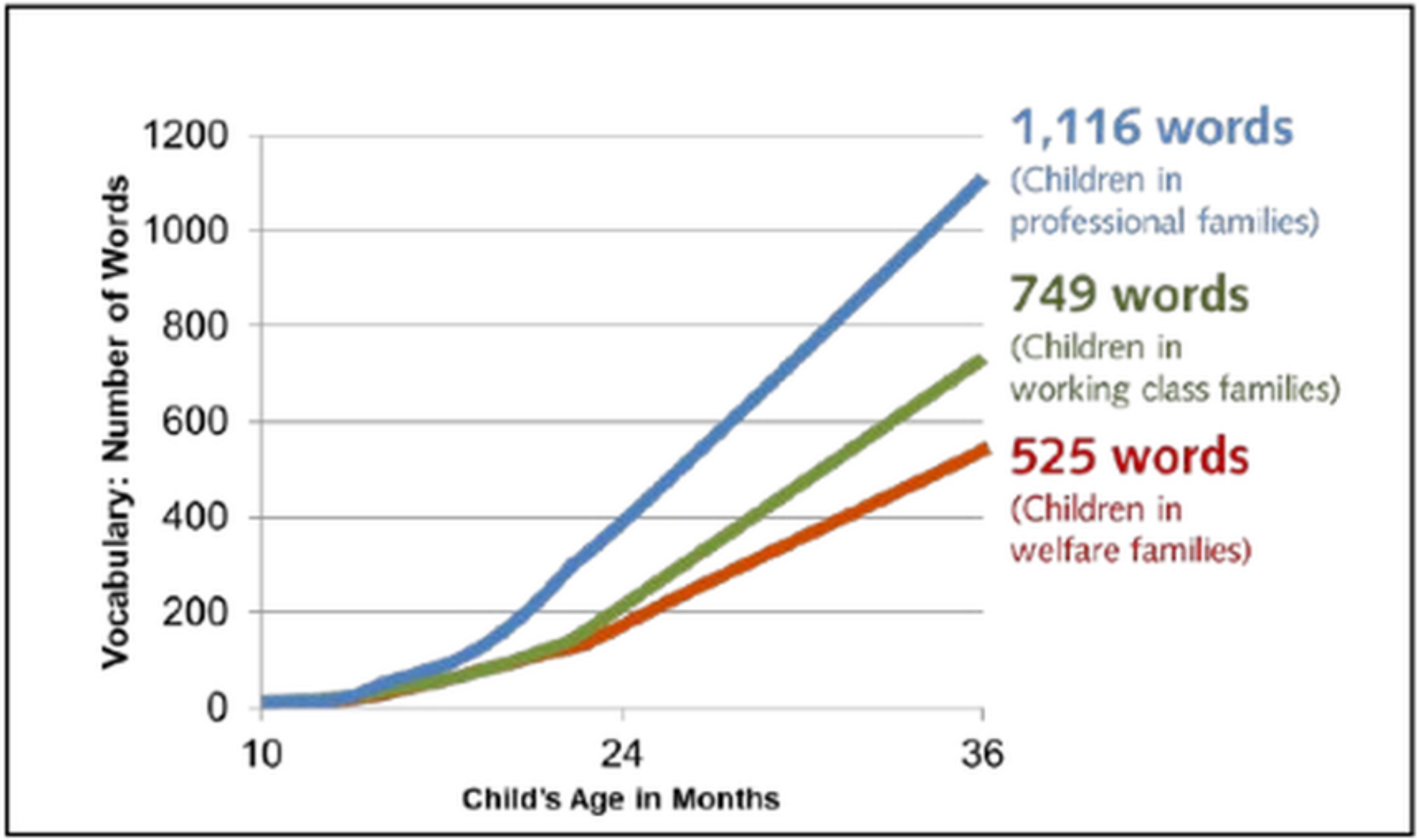
Generalization from Small Data





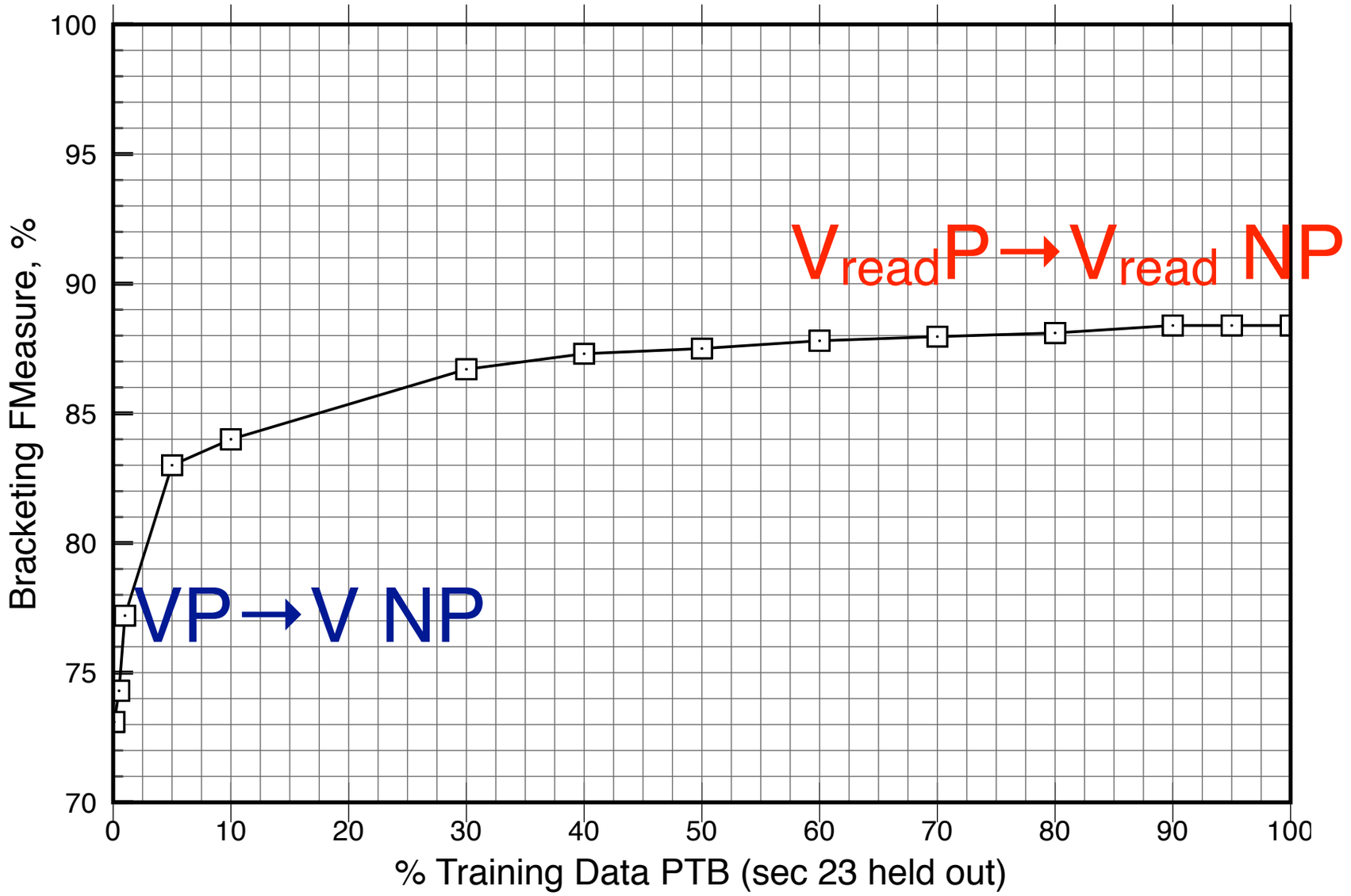
- Typical parental data
- **914** singular nouns, only **34%** are used with both **a** and **the,** but the child used them interchangeably

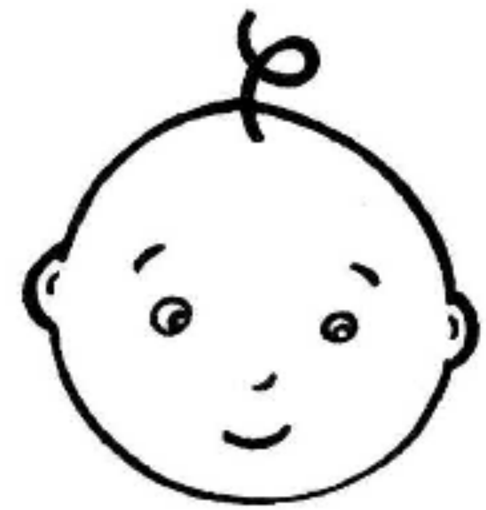
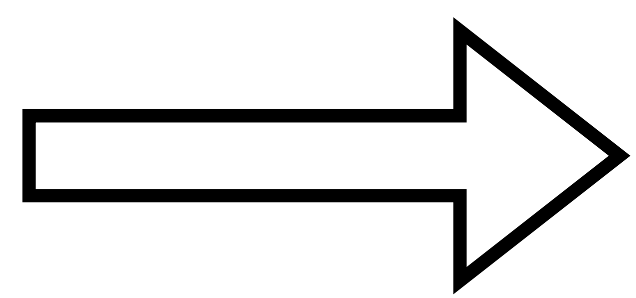
- A third of the batters are observed to switch-hit
- **All** switch-hitters?



Hart & Risley (1996) *Meaningful Differences …*Brooke Publishing.

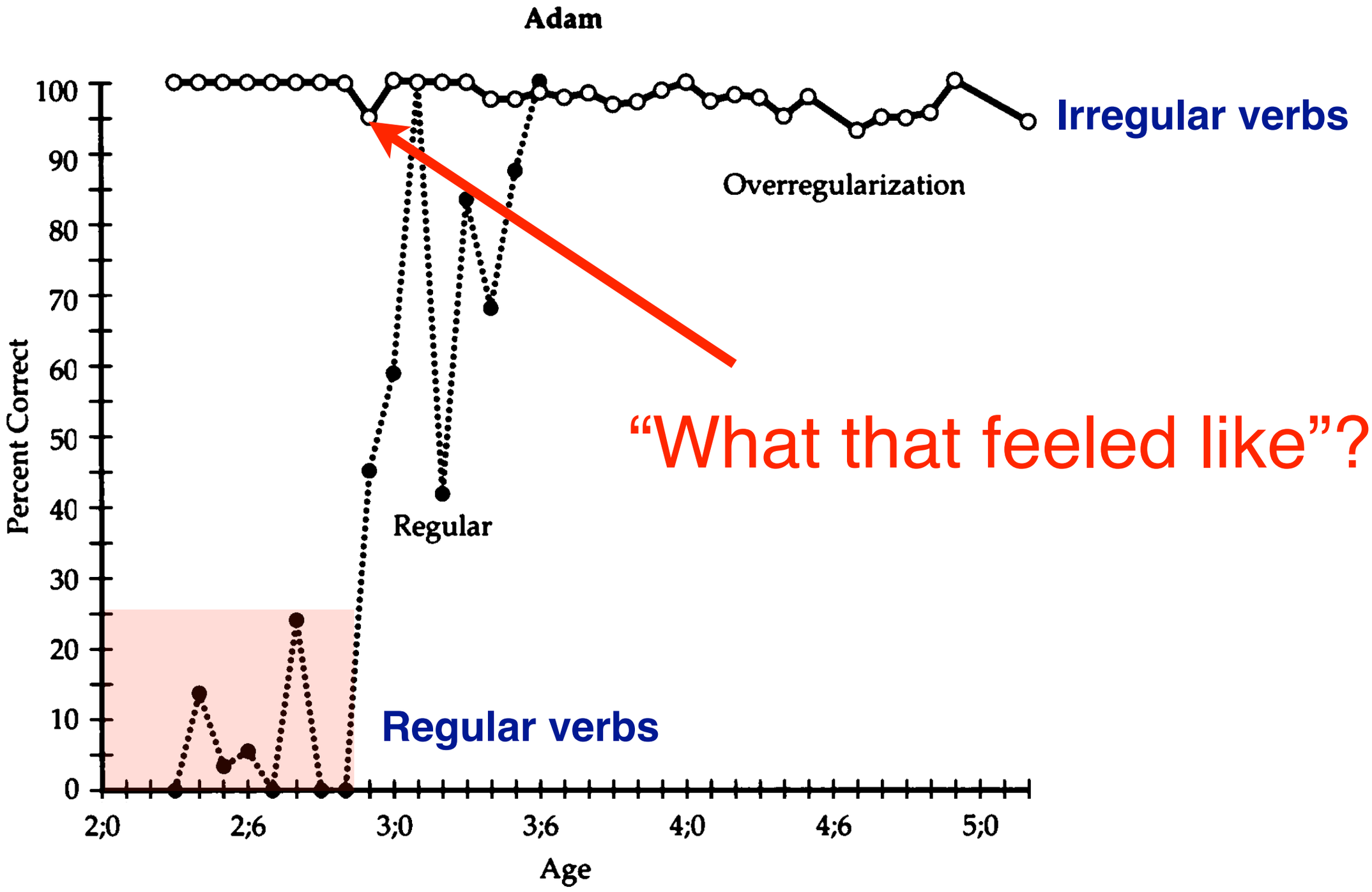
The first three minutes




right
oh
yeah
okay
go
want
yes
Laura
know
get
see
like
come
book
play
gonna
wanna
think
one
going
look
big
put
good
now
well
baby
back
Child
little
just
take
Investigator
two
name
haveto
watch
house
today
else
Daddy
color
school
truck
car
need
remember
hi
nice
open
eat
bird
bag
bad
read
help
mommy
tell
things
please
show
make
water
got
huh
kind
boy
sure
mother
girl
mean
goes
room
find
day
something
maybe
thankyou
another
red
time
sit
went
Peanuts
bottle
howcome
duck
dog
mean
let
upstairs
Joanna
poo
hear
old
door
toys
bunny
Mommy
home
orange
chair
hard
away
give
outside
happened
really
head
three
milk



G

Why so slow?

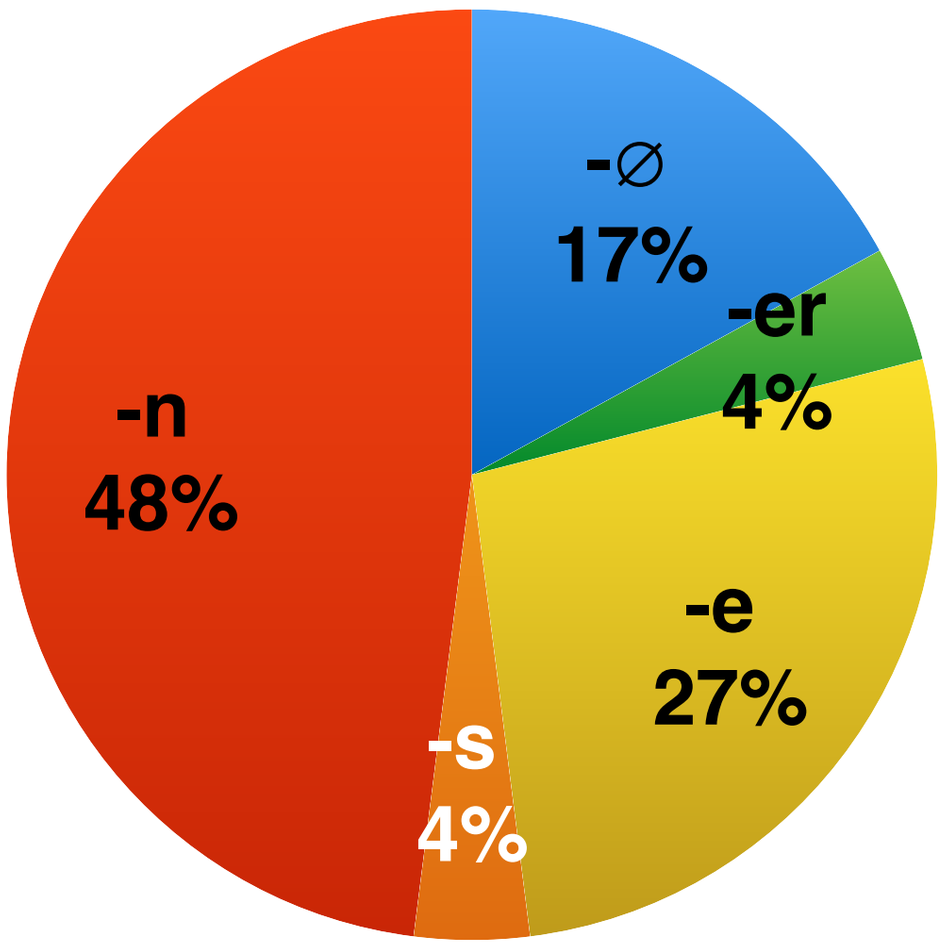


"What that feeled like"?

Pinker (1995, *An invitation to cognitive science*)

Majority doesn't rule

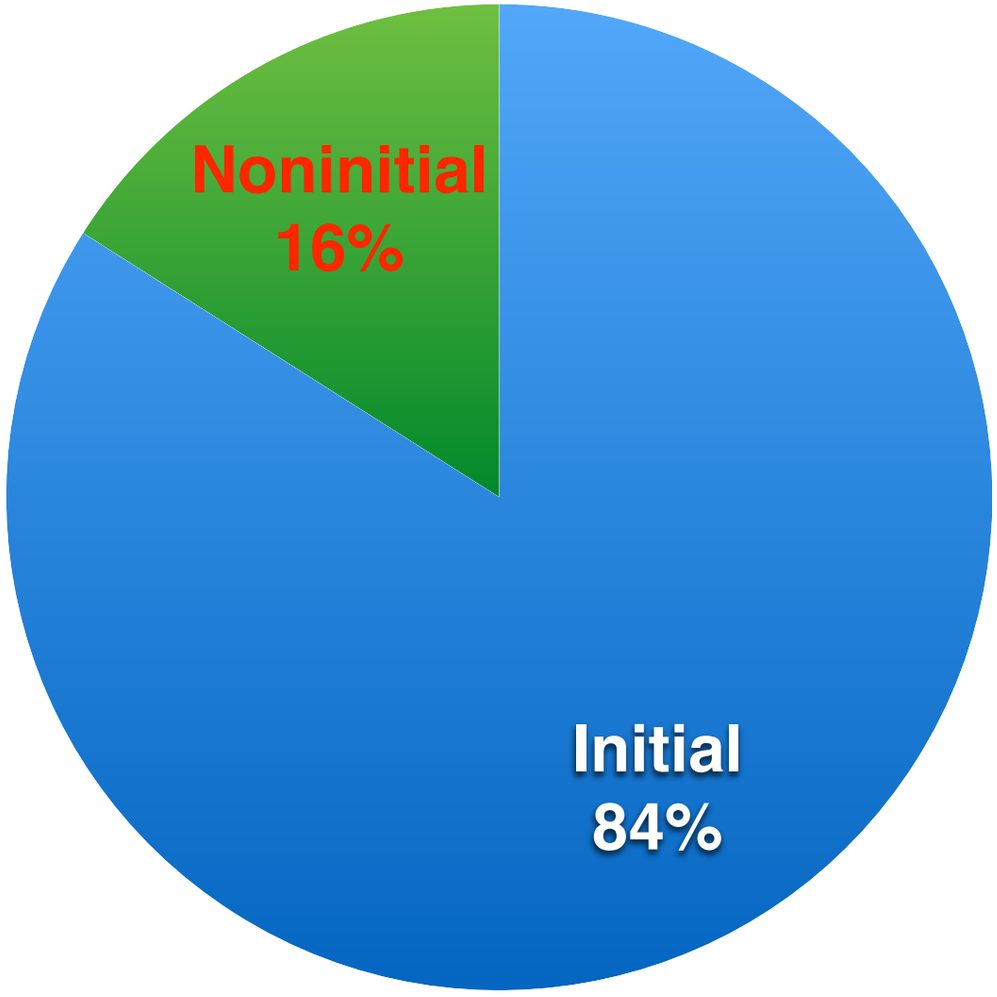
German noun plural suffixes



Autos, Parks, Pizzas, …, iPhones

Clahsen (1999, *Brain and Behavioral Sciences*)

English word stress



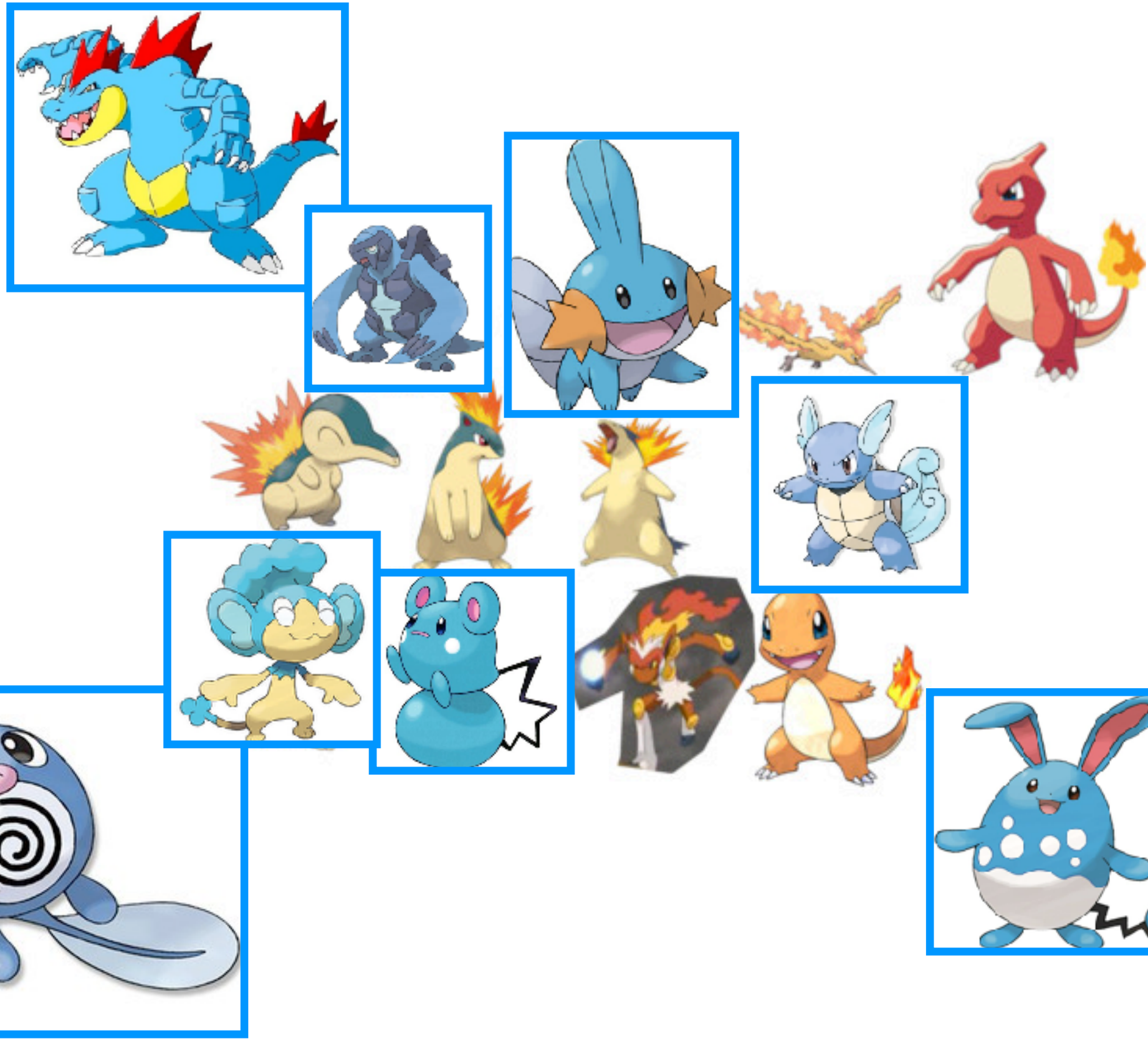
They per**mit** you to get a **per**mit

They will re**cord** a **re**cord.

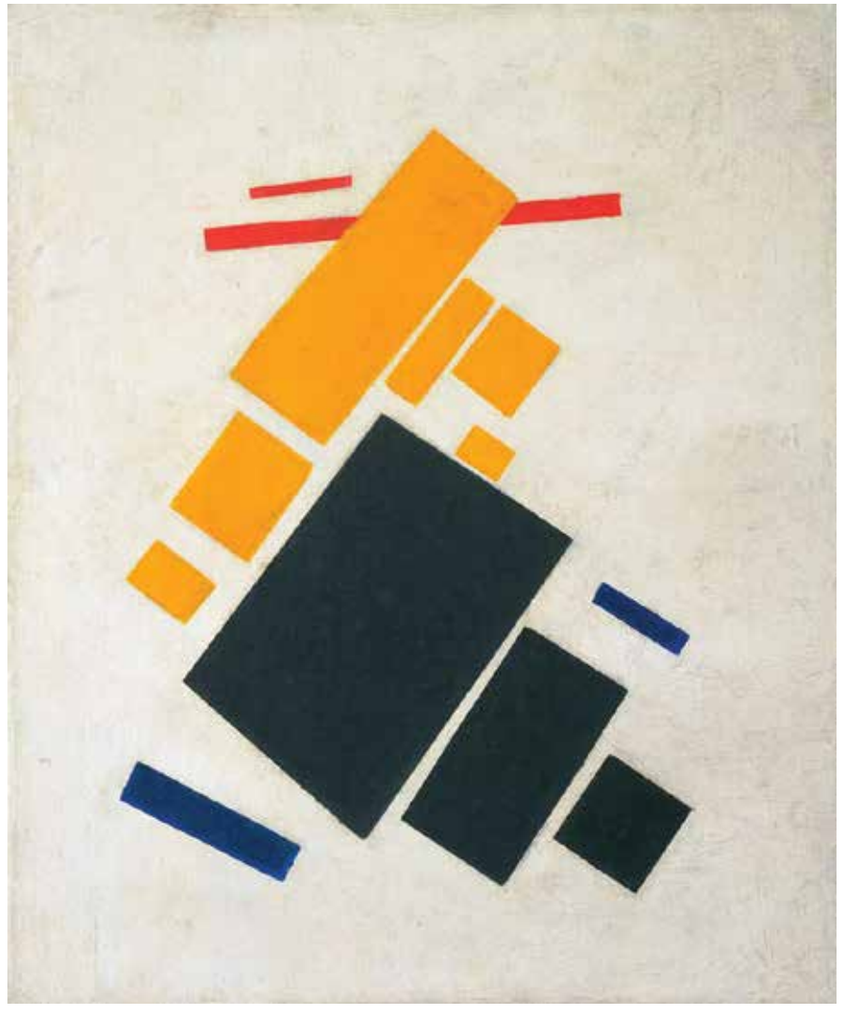
Cutler & Davis (1987, *Comp. & Speech*)

How do we learn rules?

Not so Sure?



- Give a set of items:
  - If **many** do X, then all do X
  - if ***few*** do X, then don't do X (or withhold judgment)
- How many is ***many*** or ***few***?



HOW CHILDREN LEARN TO BREAK THE RULES OF LANGUAGE

THE PRICE OF LINGUISTIC PRODUCTIVITY

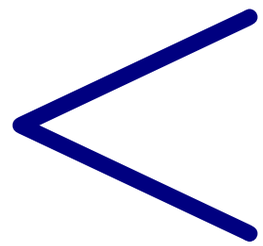
CHARLES YANG

Cambridge, MA: MIT Press

Rule + Exceptions

- Exception *1*
- Exception *2*
- Exception *3*
- ...
- Exception *e*
- Rule **(N-e)**

$$\sum_{i=1}^{e} ip_i + iP[N-e]$$



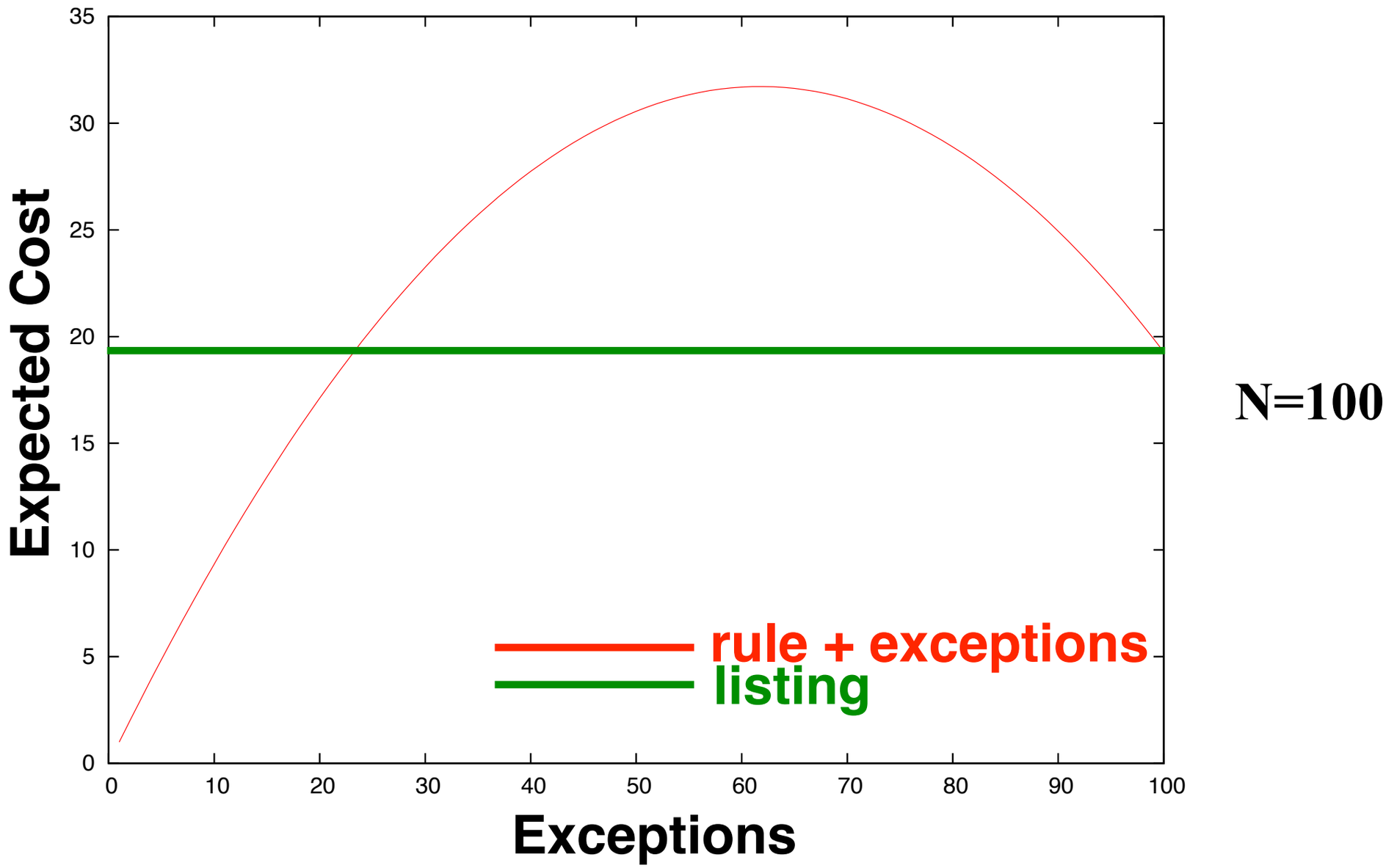
Memorize Everything

- Exception *1*
- Exception *2*
- Exception *3*
- ...
- ...
- Exception **N**

$$\sum_{i=1}^{N} ip_i$$

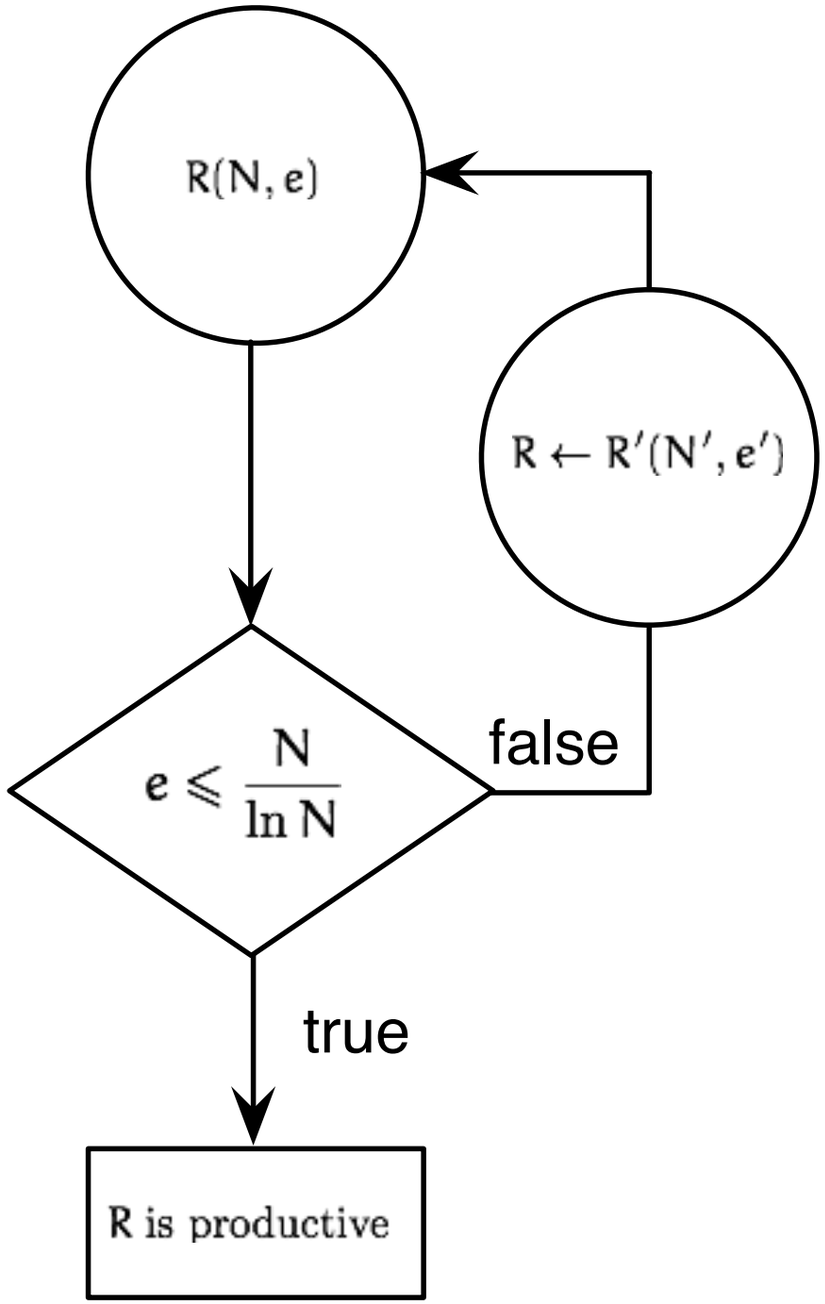
Tolerance Principle

A productive rule over **N** items cannot have more than **N/lnN** exceptions



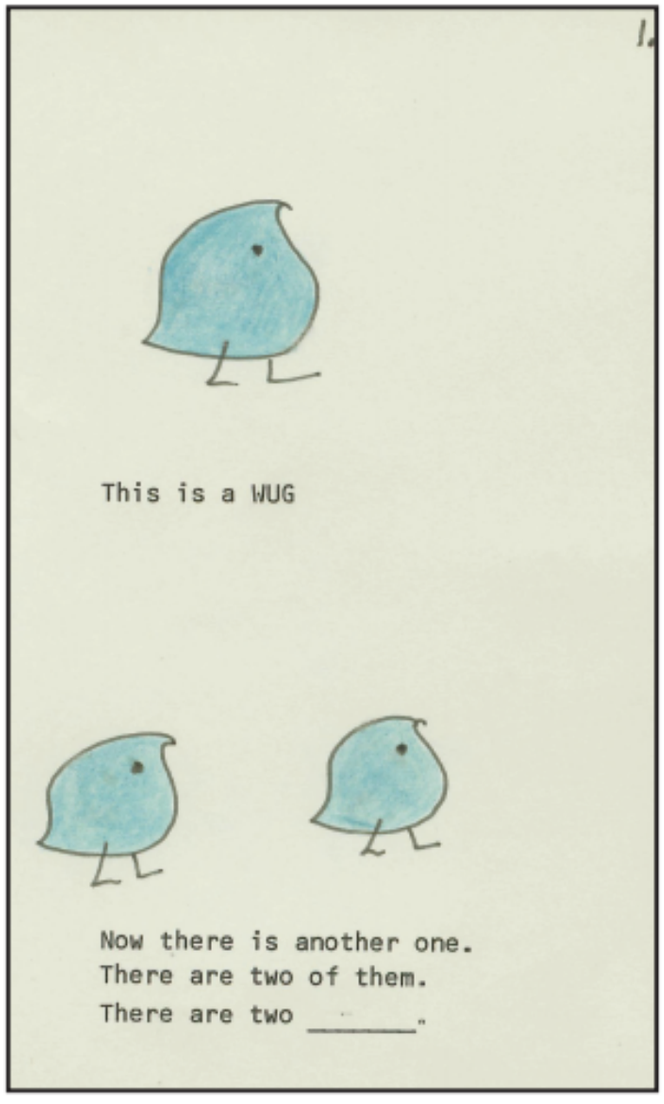
Tolerance Principle in Action

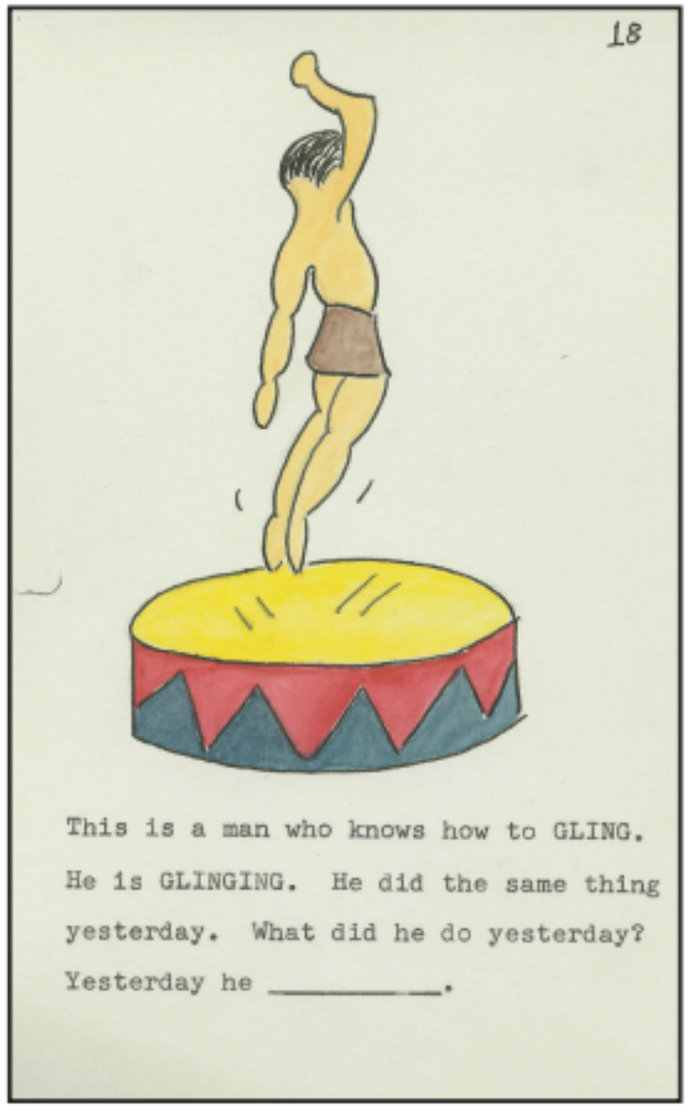
N	θ_N	%
10	4	40%
20	7	35%
50	13	26%
100	22	22%
200	38	19%
500	80	16%
1000	144	14%
2000	263	13%
5000	587	12%
10000	1086	11%

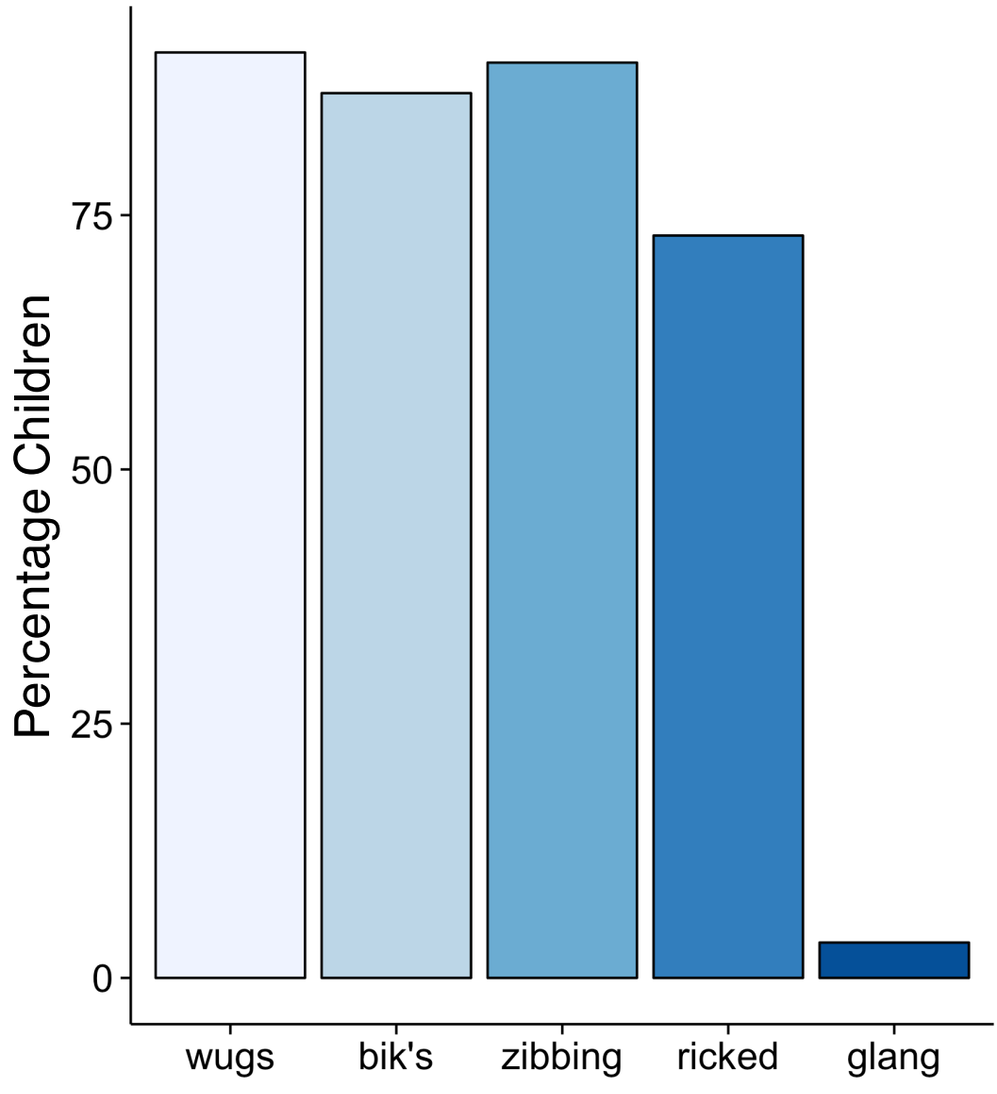


parameter free, small is better

Wug Test





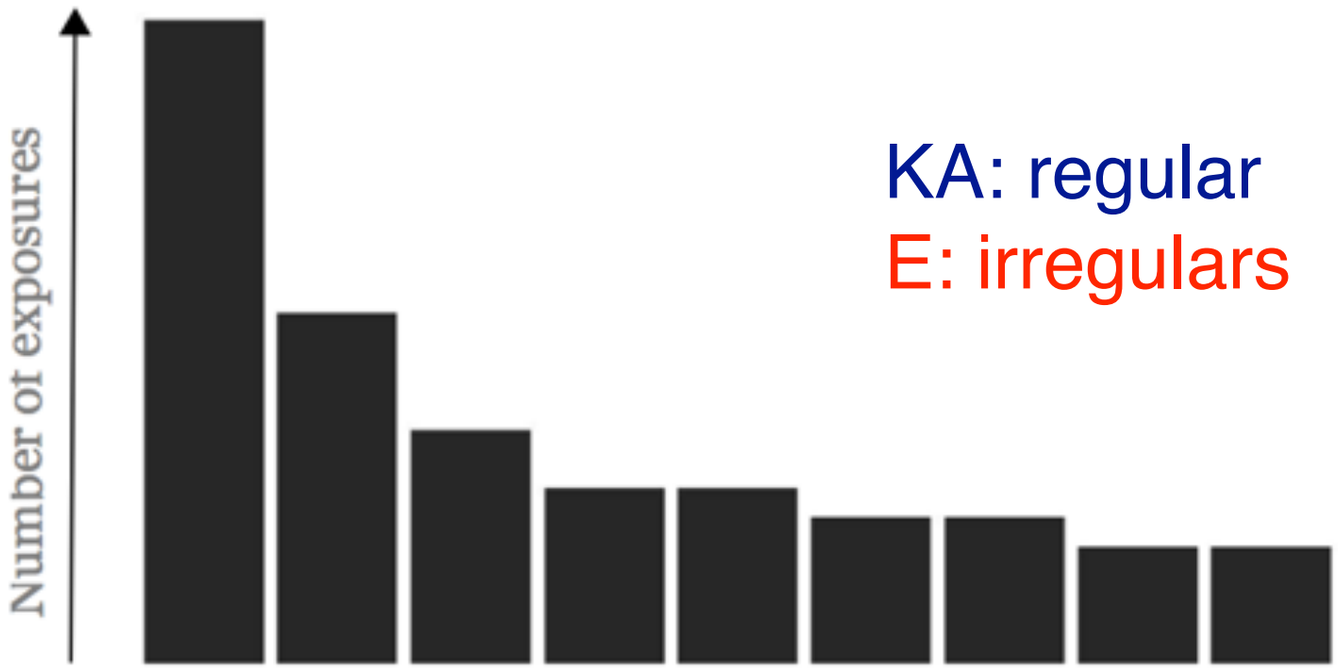


child performance

Artificial Language

Why 5/4 and 3/6?

9/ln9 = 4.2!



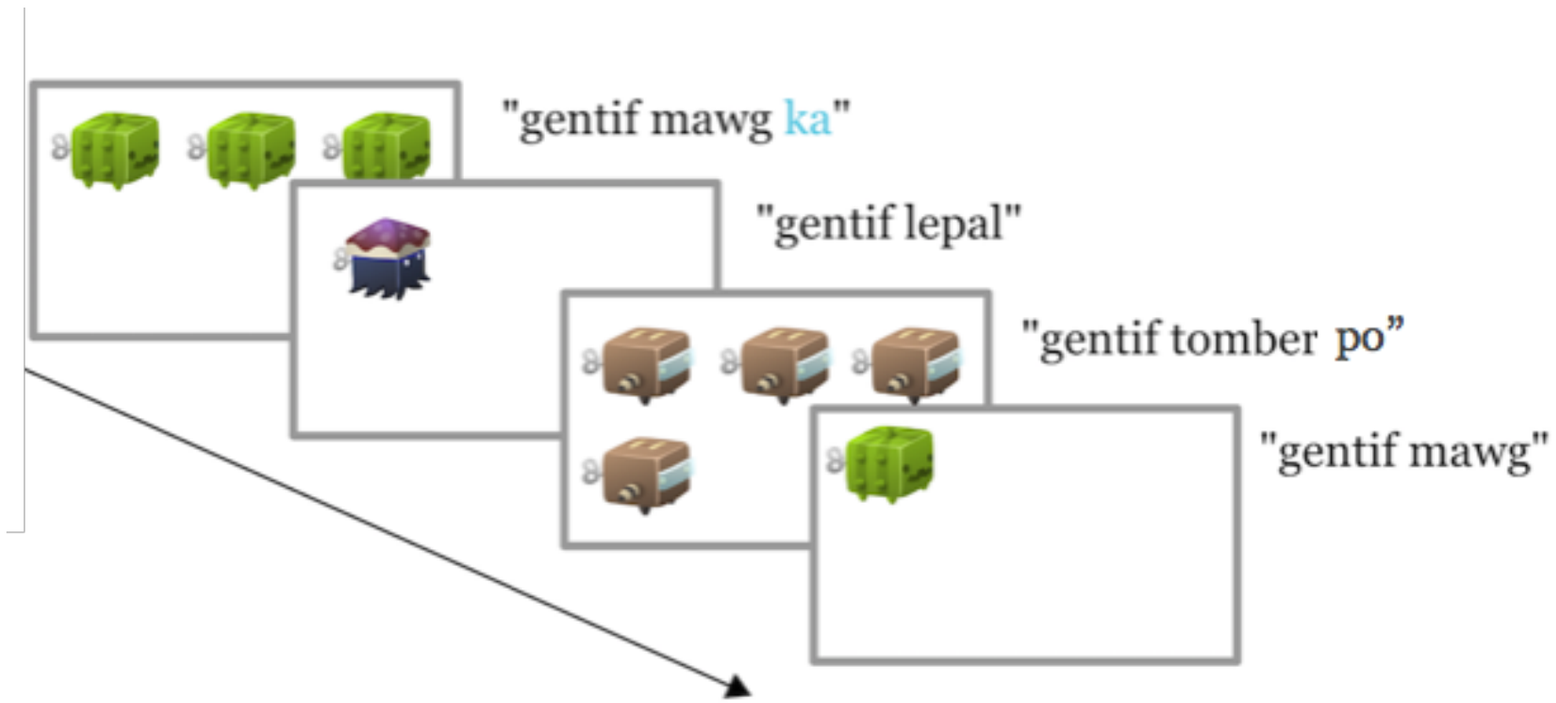
KA: regular
E: irregulars

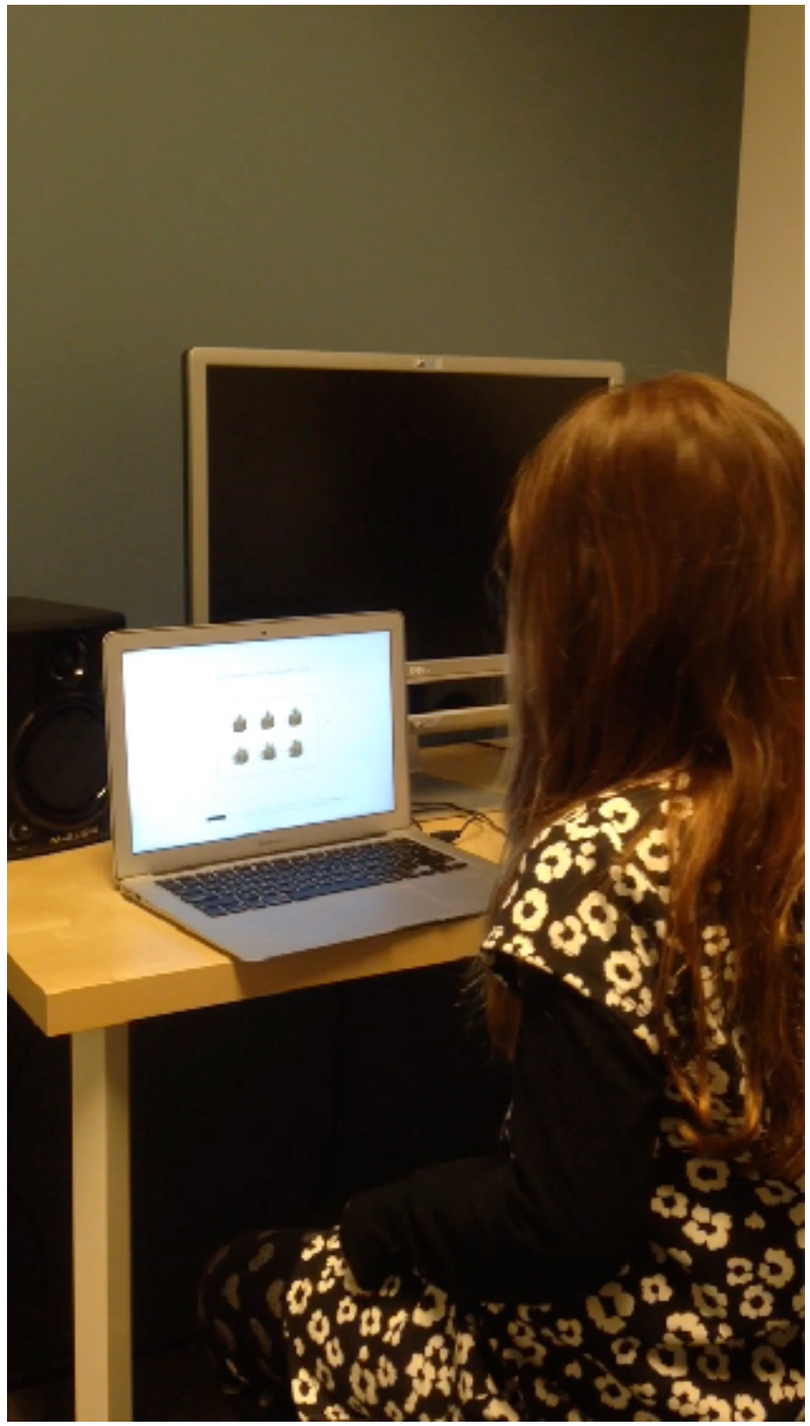
Conditions

5R4E	**ka**	**ka**	**ka**	**ka**	**ka**	po	lee	bae	tay
3R6E	**ka**	**ka**	**ka**	po	lee	bae	tay	muy	woo

Schuler, Yang, & Newport (2016, Cog. Sci. Soc.)

Training

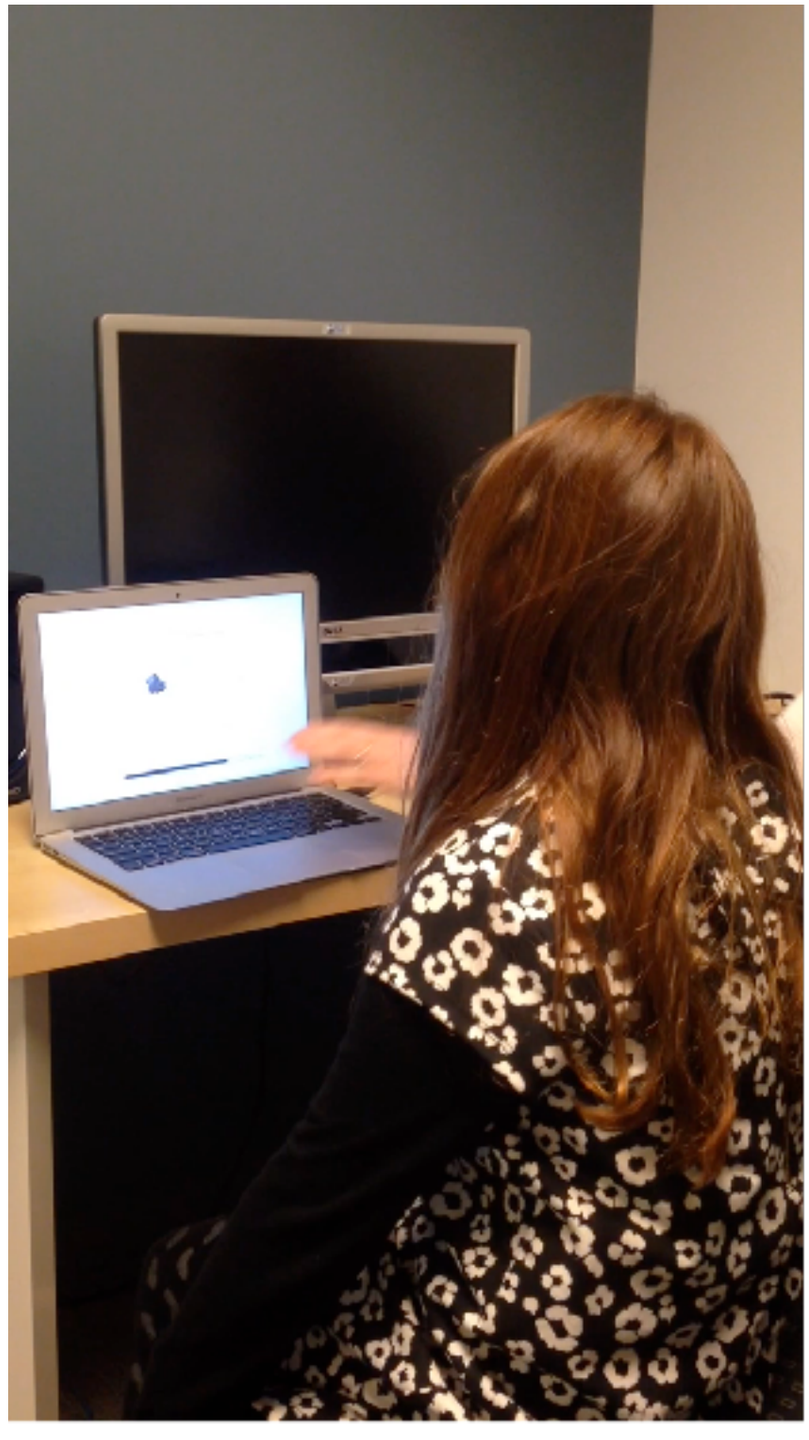




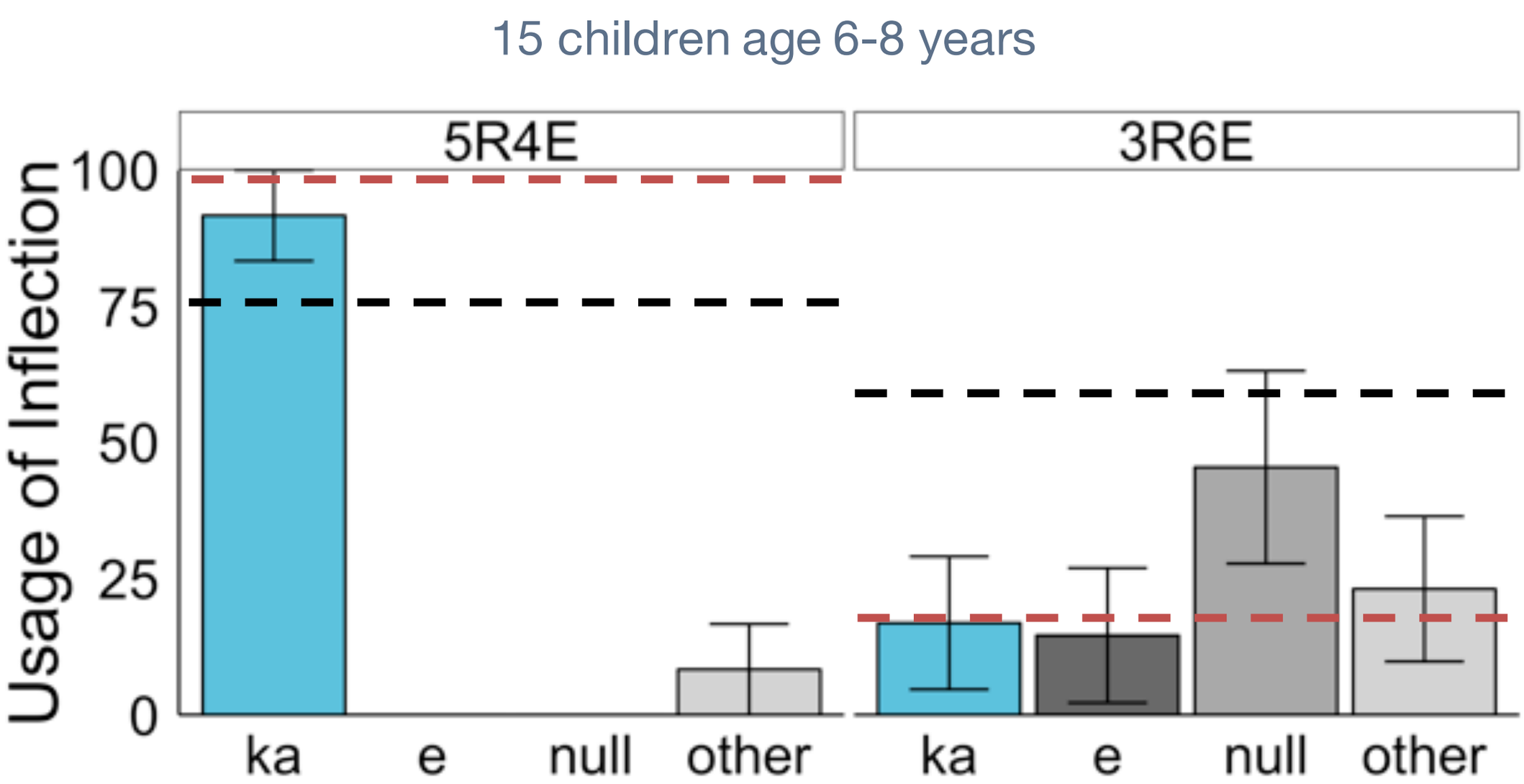
Testing

Experimenter says
"gentif norg."

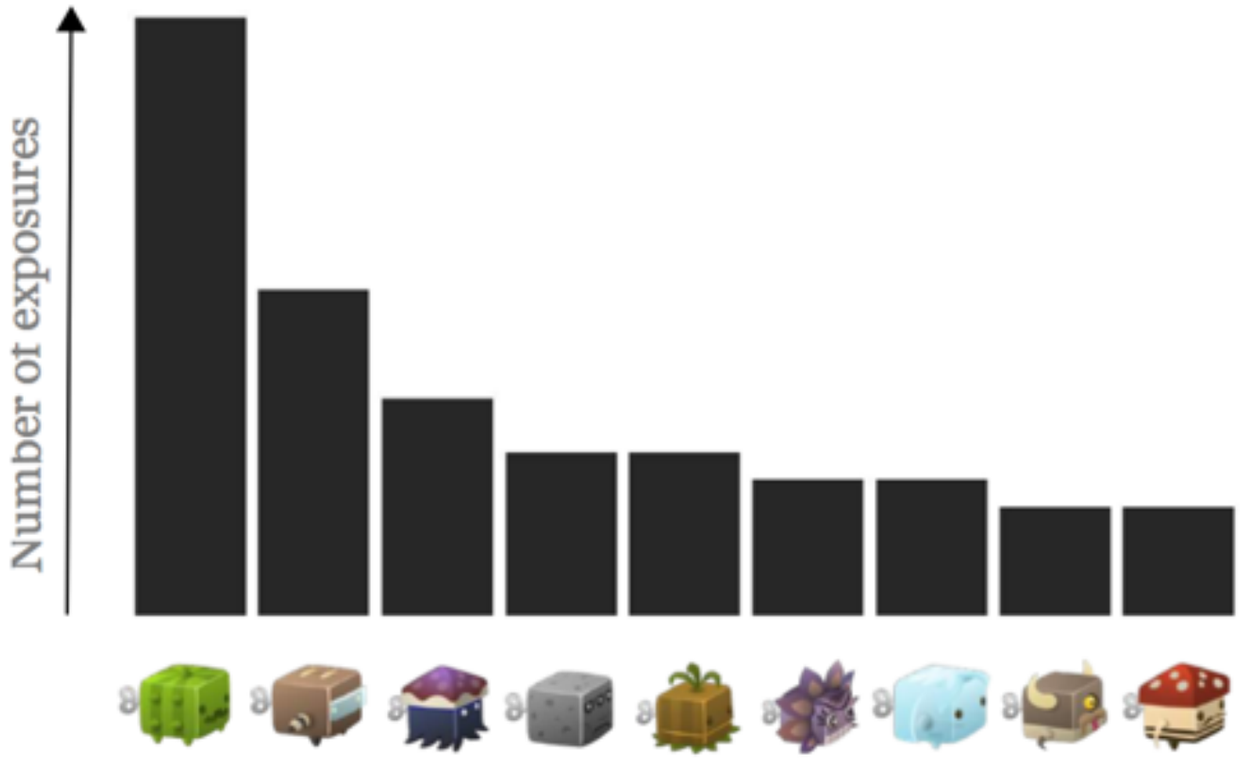
Child says
"gentif _____"



Essentially categorical!



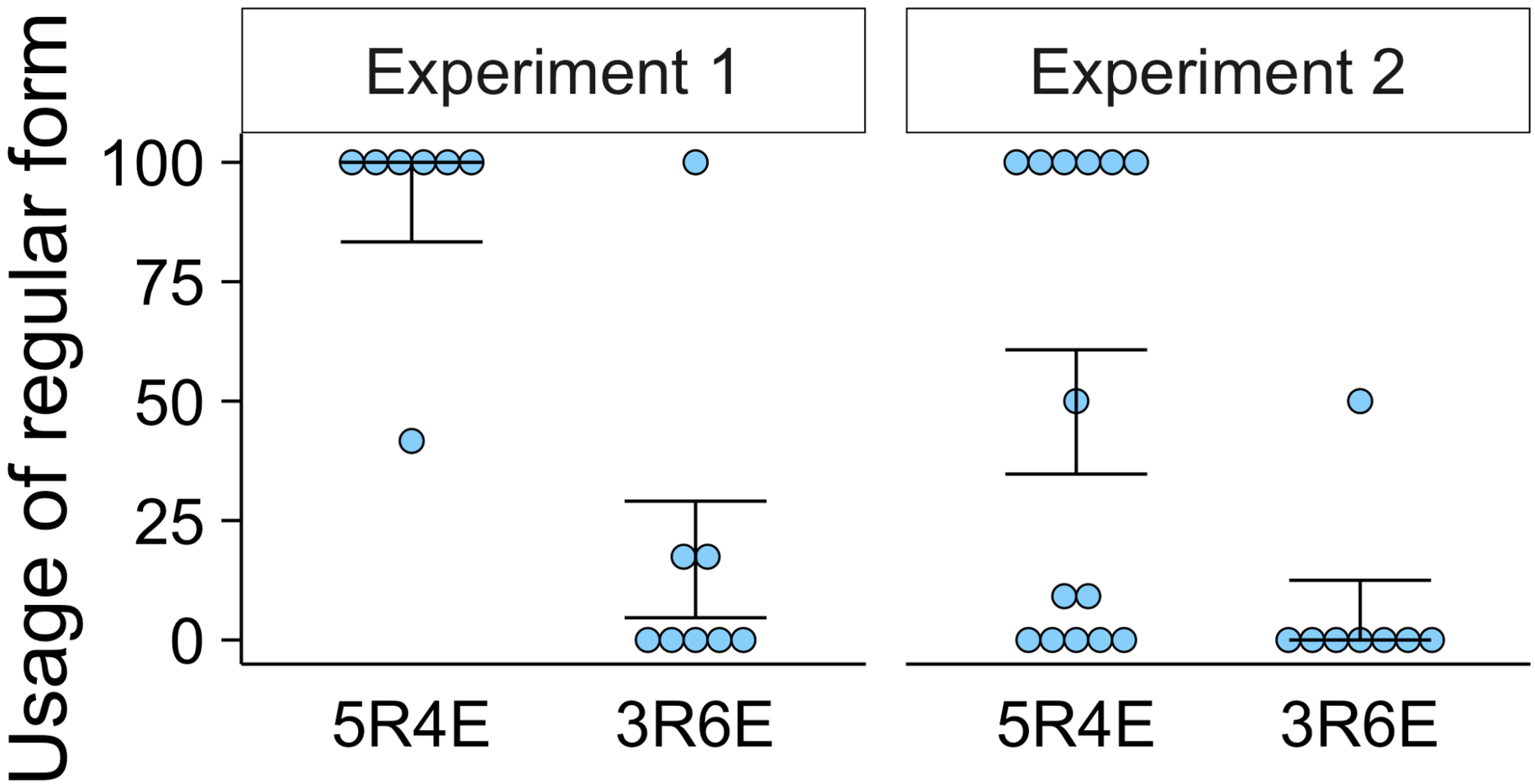
Making it harder



Conditions

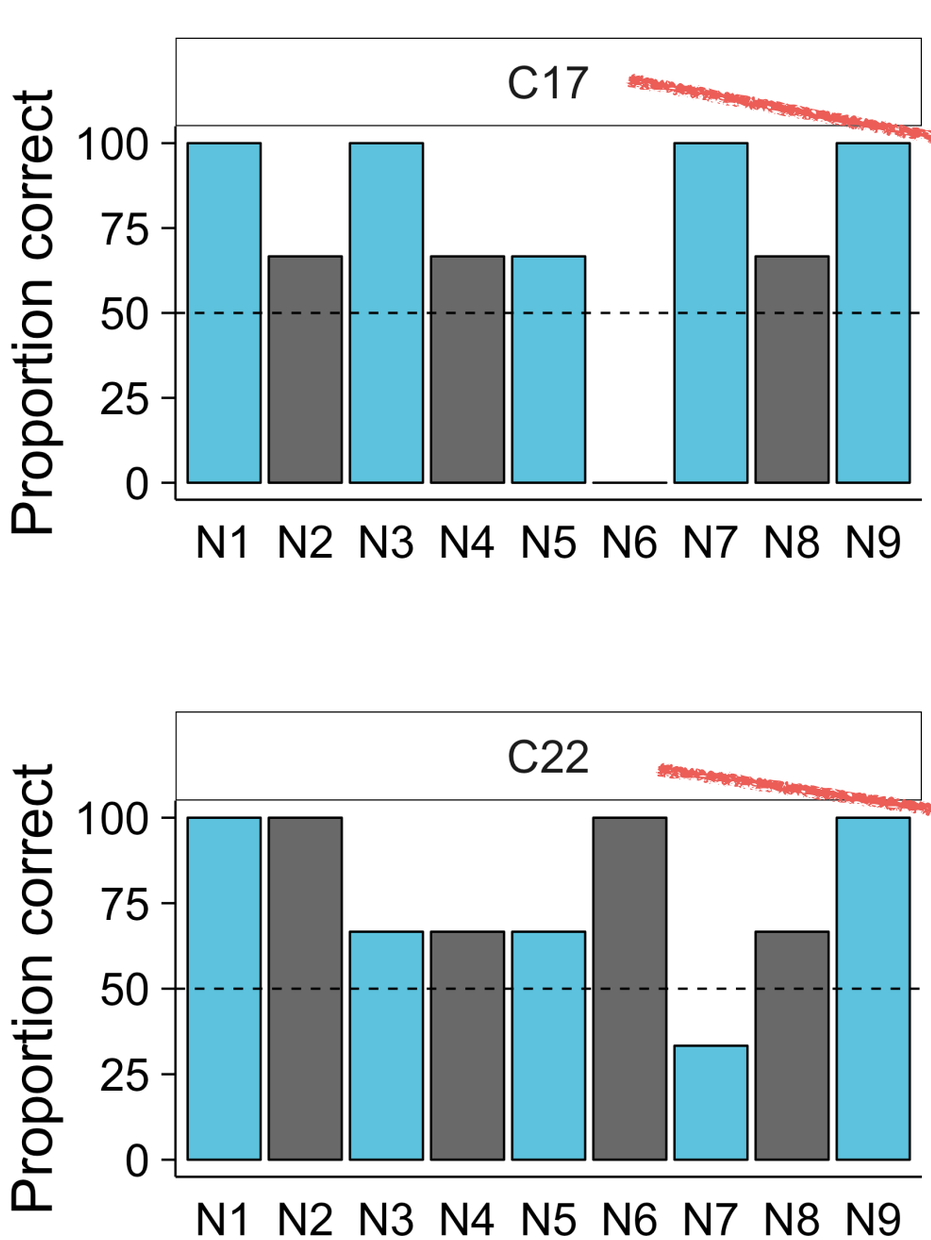
5R4E	**ka**	po	**ka**	lee	**ka**	bae	**ka**	tay	**ka**
3R6E	**ka**	po	**ka**	lee	**ka**	bae	tay	muy	woo

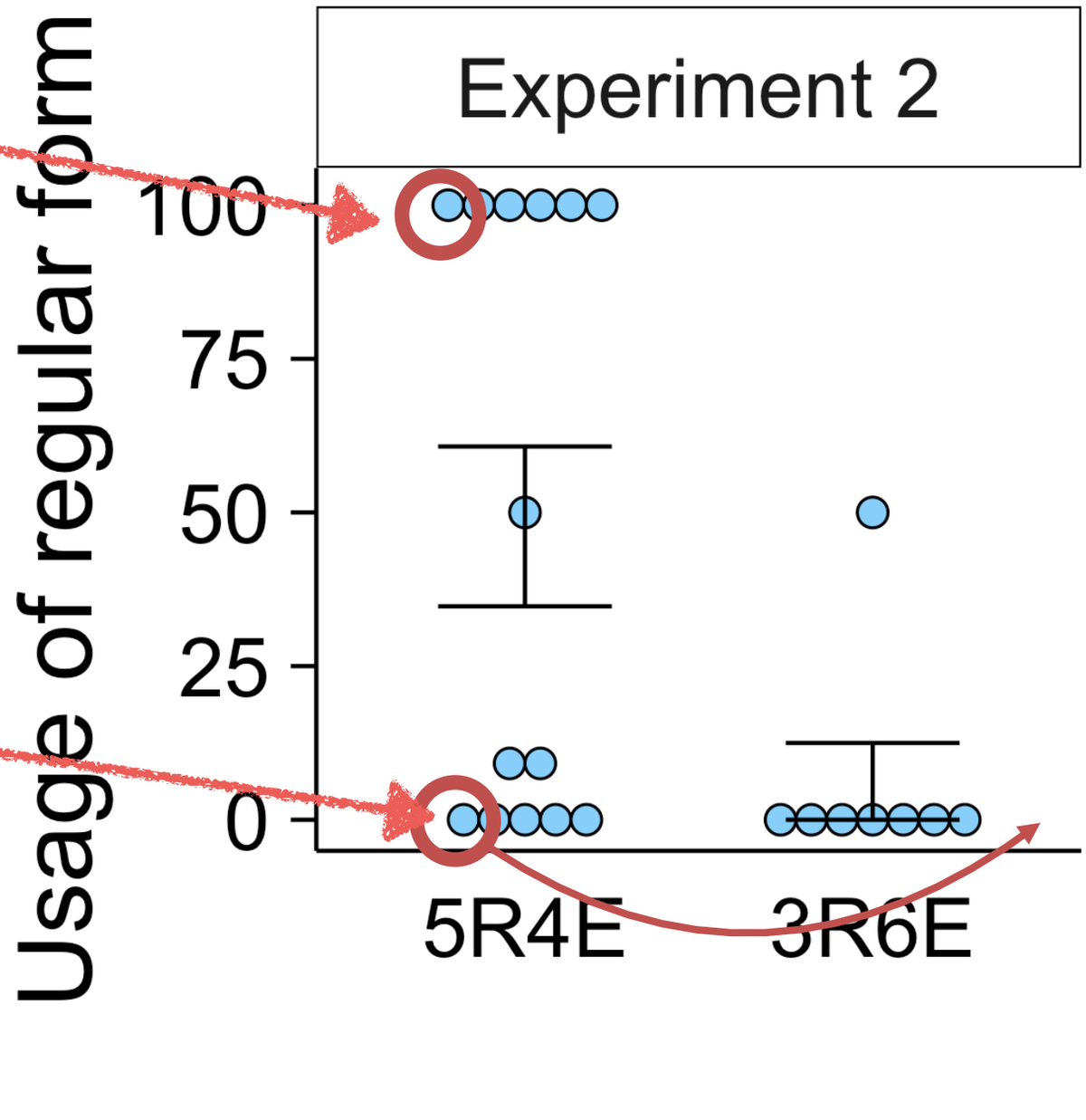
Broken?


Experiment 1
Experiment 2
Usage of regular form
100
75
50
25
0
5R4E
3R6E
5R4E
3R6E

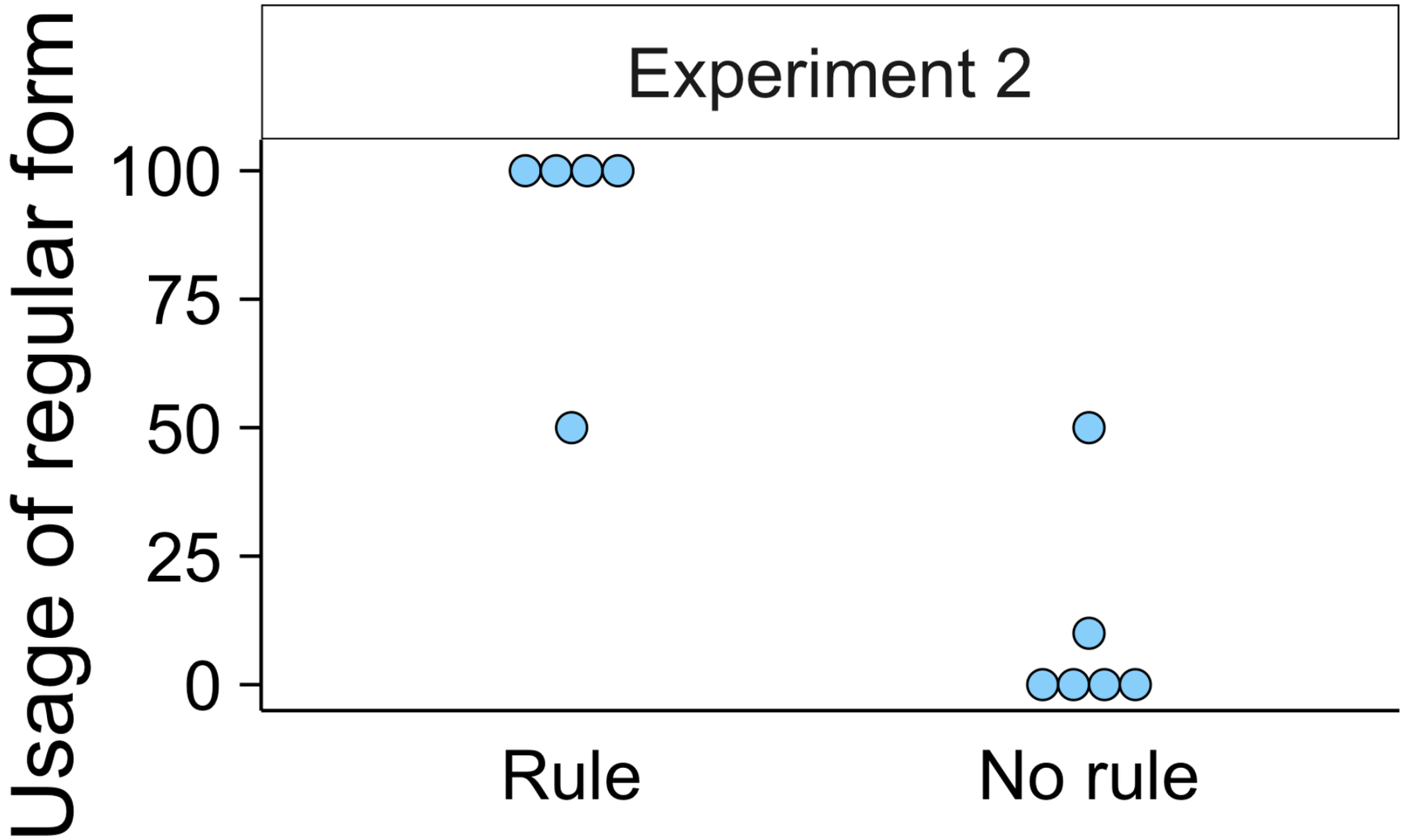
Personalized Tolerance Principle

8/ln(8) = 3.85

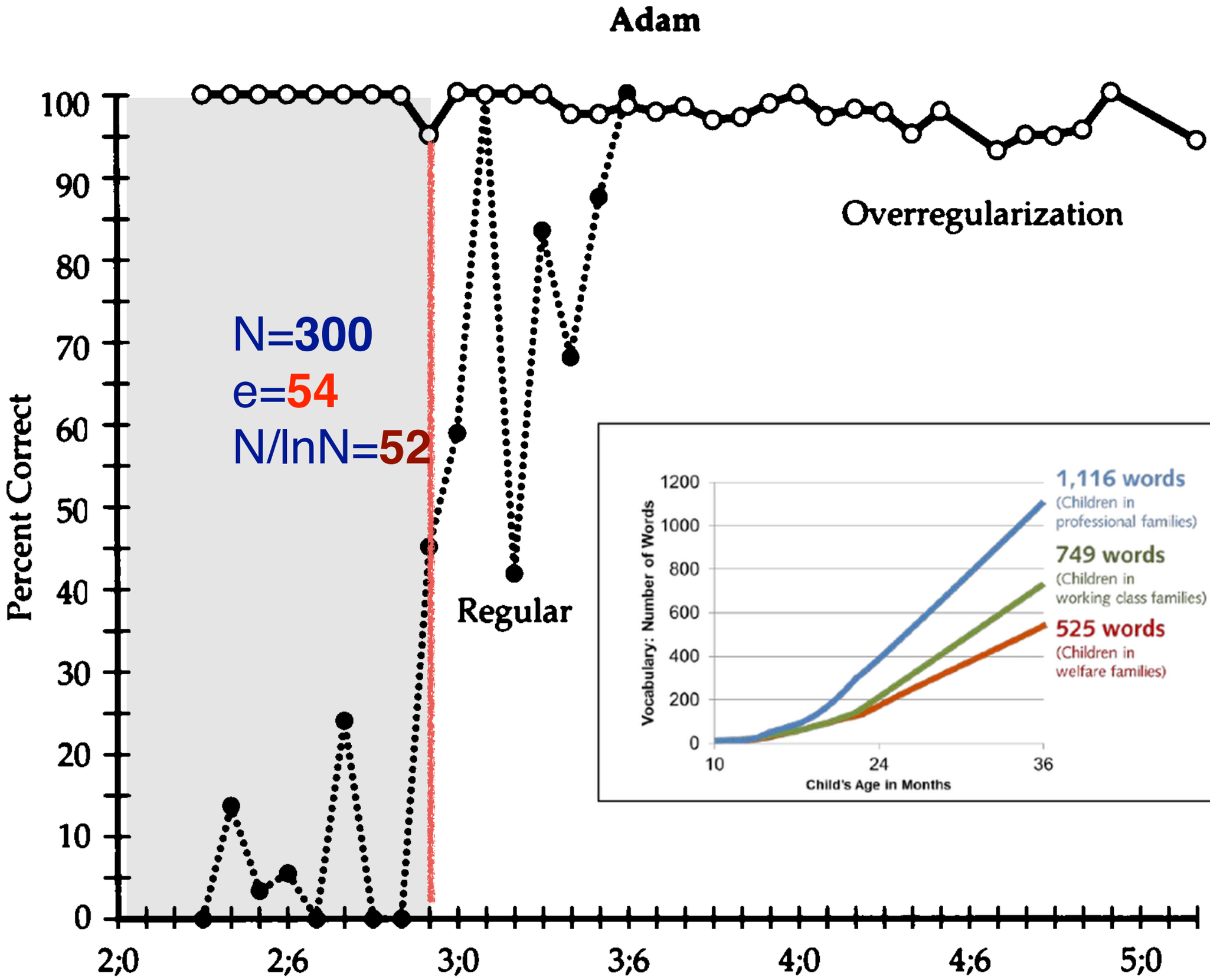




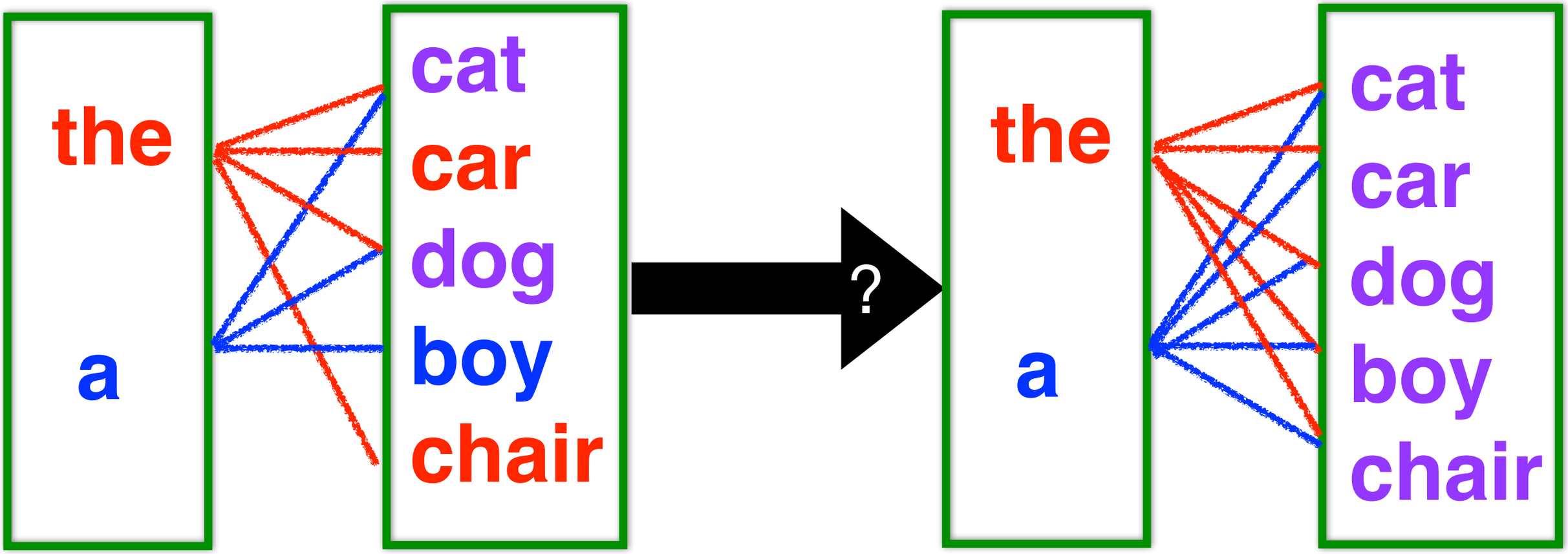
Categorical again

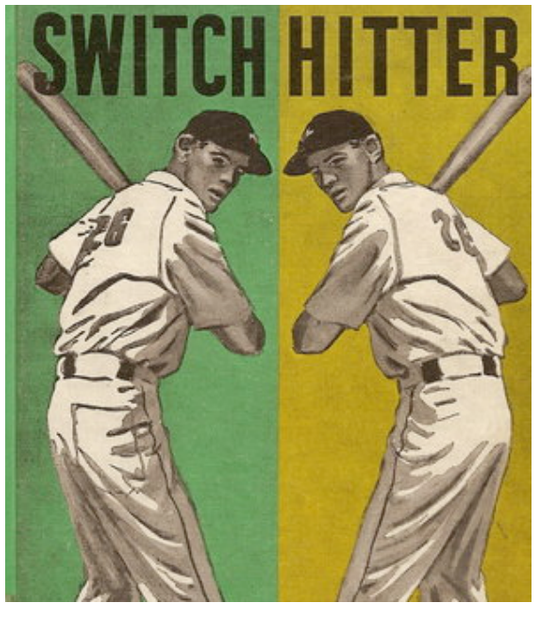


When *felt* became *feeled*?



Generalization from Small Data



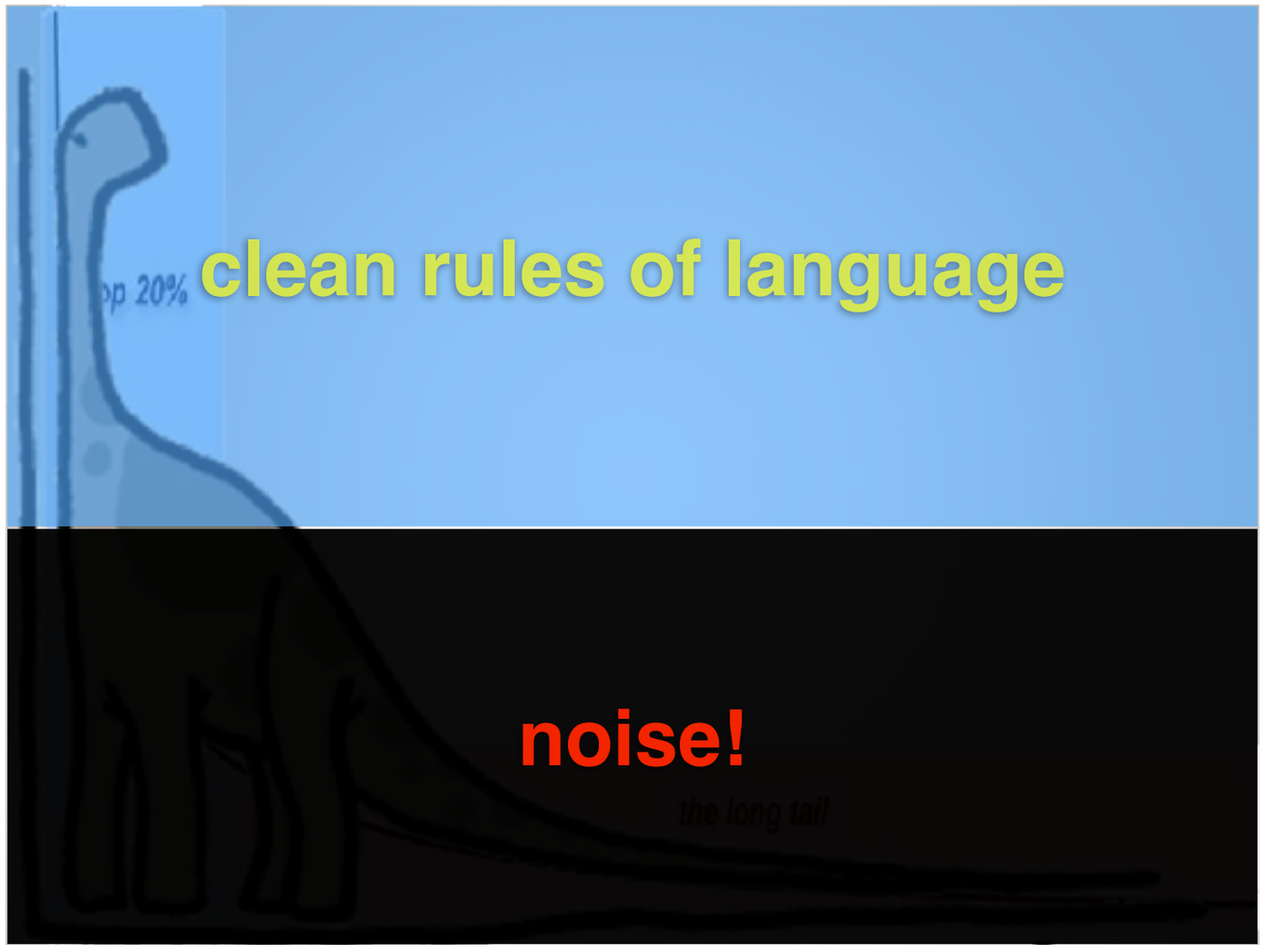


Make More out of Less

the ONION
America's Finest News Source

Study Reveals: Babies Are Stupid

Above: Despite their relatively large cranial capacities, babies such as this one are so unintelligent that they are unable to distinguish colorful plastic squeak toys from food sources.



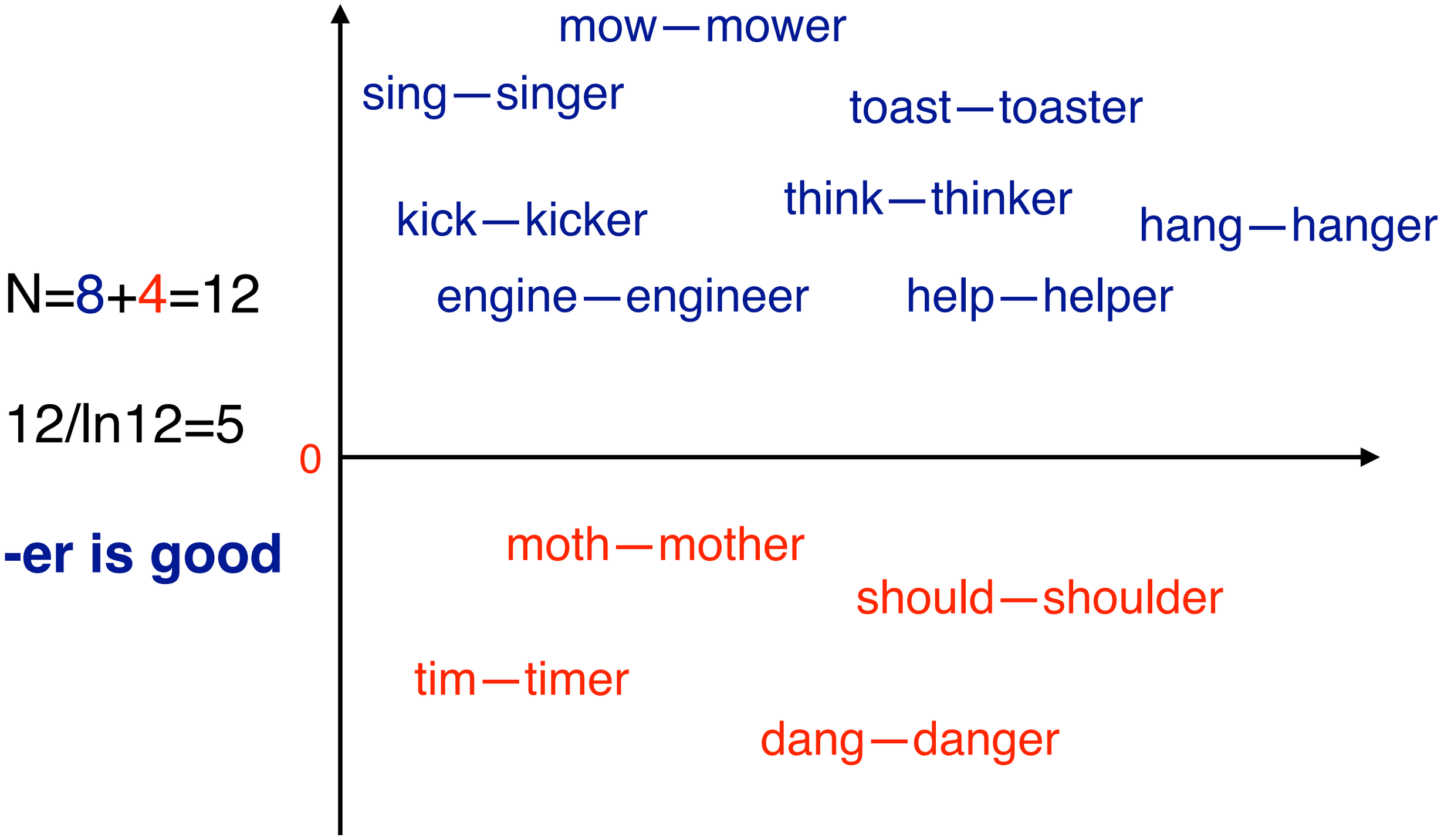
- All nouns: only **40%** appear with both
  - Top **50** nouns: **43** appear with both
  - Top **100** nouns: **87** appear with both

Unsupervised Morphology

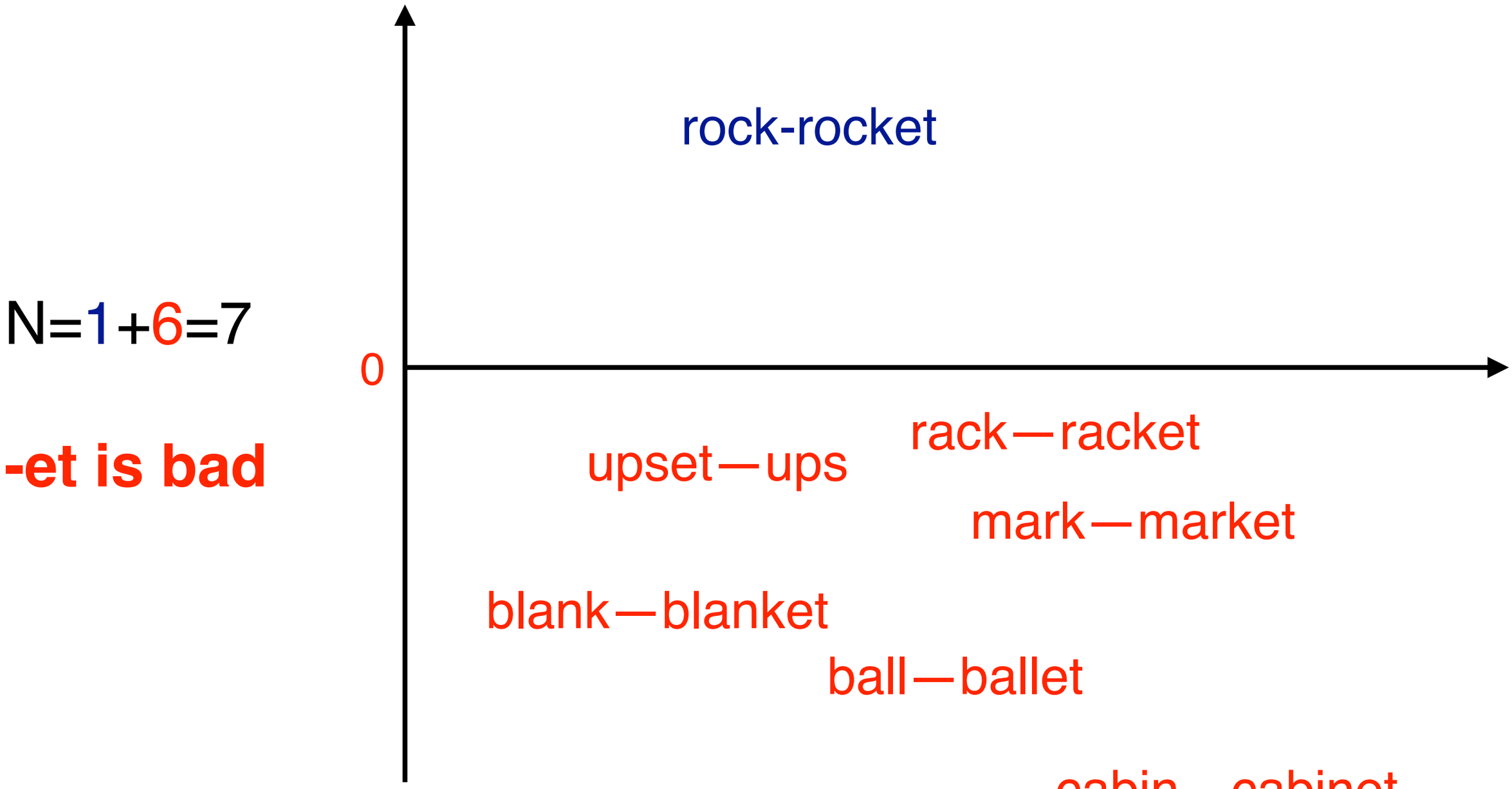
- Mitch Marcus (Penn), Lyle Ungar (Penn)
- Emily Pitler (Google), Erwin Chan (Arizona), Constantine Lignos (ISI), Hongzhi Xu (Penn)

- **er**: productive but with corner cases
  - think-think**er**, drink-drink**er**, …, corn-corn**er**, live-liv**er**
  - Young children: “sounder” = radio, “caser” = storage bin
- **et**: spurious and should be purged
  - wall-wall**et**, bull-bull**et**, ball-ball**et**, ass-…

Embedding -er in 4.5 million words



Embedding -et in 4.5 million words



Approach

- Morphology = Phonology + Semantics(e.g., Schone & Jurafsky 2001)
- A conventional distributional learner for affixes {S}
- Subject each affix (S) to Tolerance test
  - **N** is the number of word pairs related by S
  - Accept S if there are fewer than **N/lnN** negatively related words
- Segmentation using filtered {S}'

Performance

Model	Embedding	Precision	Recall	F-score
NBJ2015	129M Wikipedia	0.807	0.722	0.762
ccccc-, -ccccc +tolerance	4.5M child-directed English	**0.879**	0.719	**0.786**
Xu et al. 2018	**none**	0.810	**0.787**	0.798
Xu et al. 2018 +tolerance	17M words (Word2Vec demo)	0.840	0.770	**0.803**

Narasimhan, Barzilay, & Jaakkola (2015, *TACL*)
Xu, Marcus, Ungar, & Yang (2018, COLING)
Thanks to Hongzhi Xu

Why 72?

one two three four five six seven eight nine ten

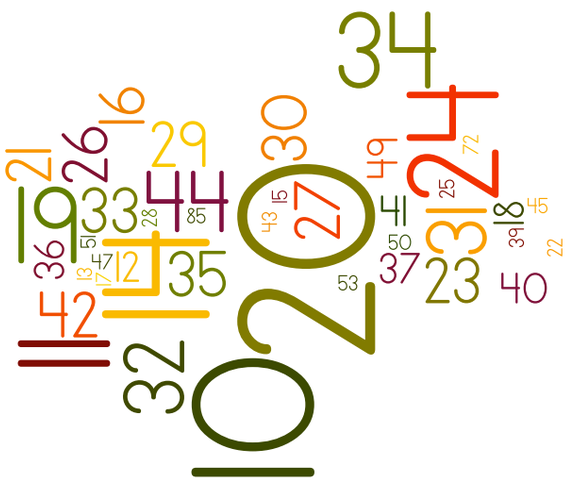
eleven twelve thirteen fourteen fifteen sixteen seventeen

eighteen nineteen twenty 21 22 23 24 25 26 27 28 29

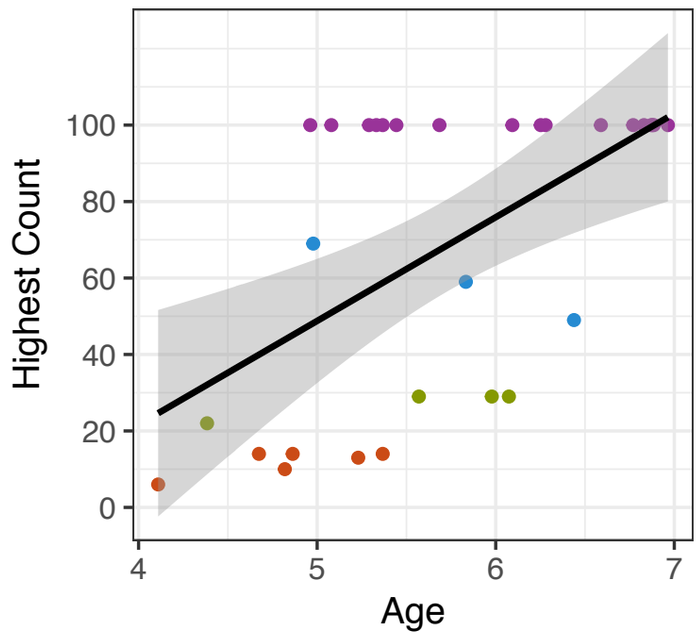
thirty 31 32 33 34 35 36 37 38 39

40 41 42 43 44 45 46 47 48 fifty

51 52 53 54 55 56 57 58 59 60 ...



$$\frac{73}{\ln 73} = 17$$



Why approximately 40?

一 二 三 四 五 六 七 八 九 十

十一 十二 十三 十四 十五 十六 十七 十八 十九 二十

二十一 二十二 二十三 二十四 二十五 。 。 。

$$\frac{42}{\ln 42} = 11$$

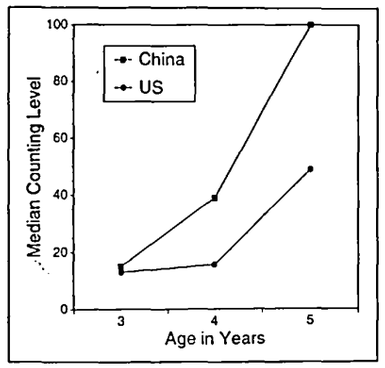
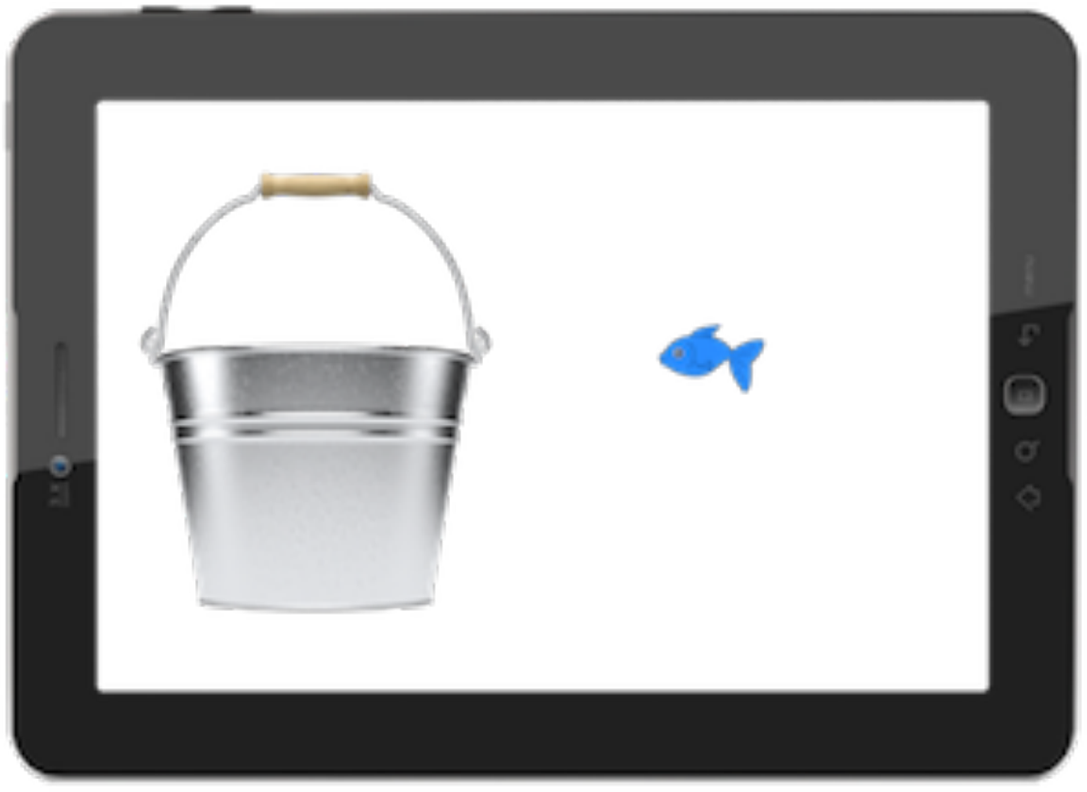


Fig. 1. Median level of abstract counting (highest number reached) by age and language. Significant differences favoring Chinese-speaking subjects were found at ages 4 and 5 years, but not at age 3.

"Four-year-olds in China made very rapid progress in generalizing number names up to 100 after they could count to approximately **40**" (Miller, Kelly, & Zhang 2005)

"Look! I have 7 fish. I'm putting 7 fish in the bucket!"



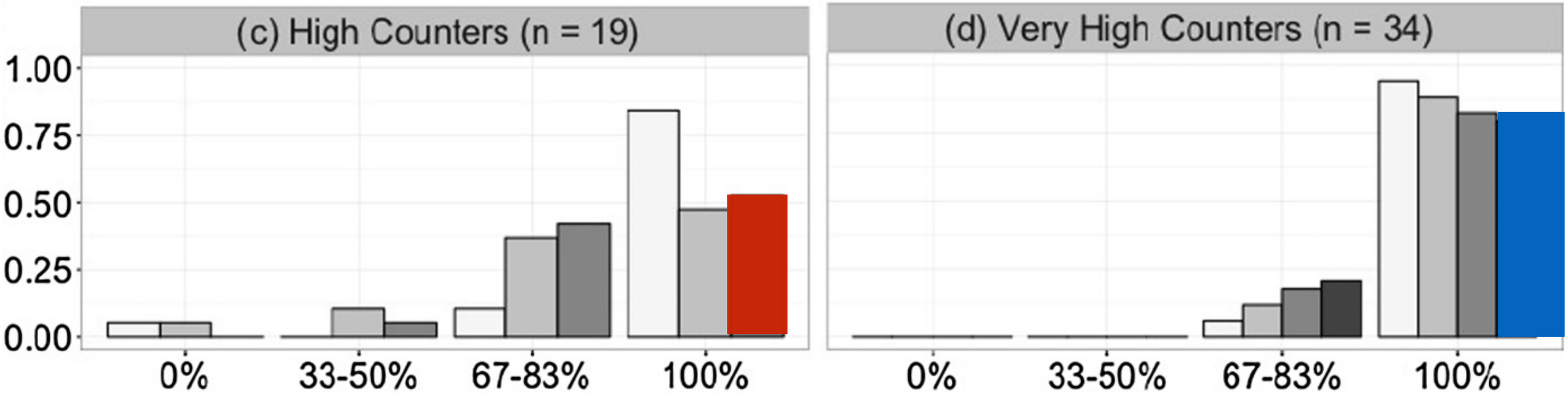
"Look!"

"Are there 8 or 9 fish in Mr. Dino's bucket?"

Productive Counting is the Key

40-79

80-100



23, 24, 28, 31, 35, 36

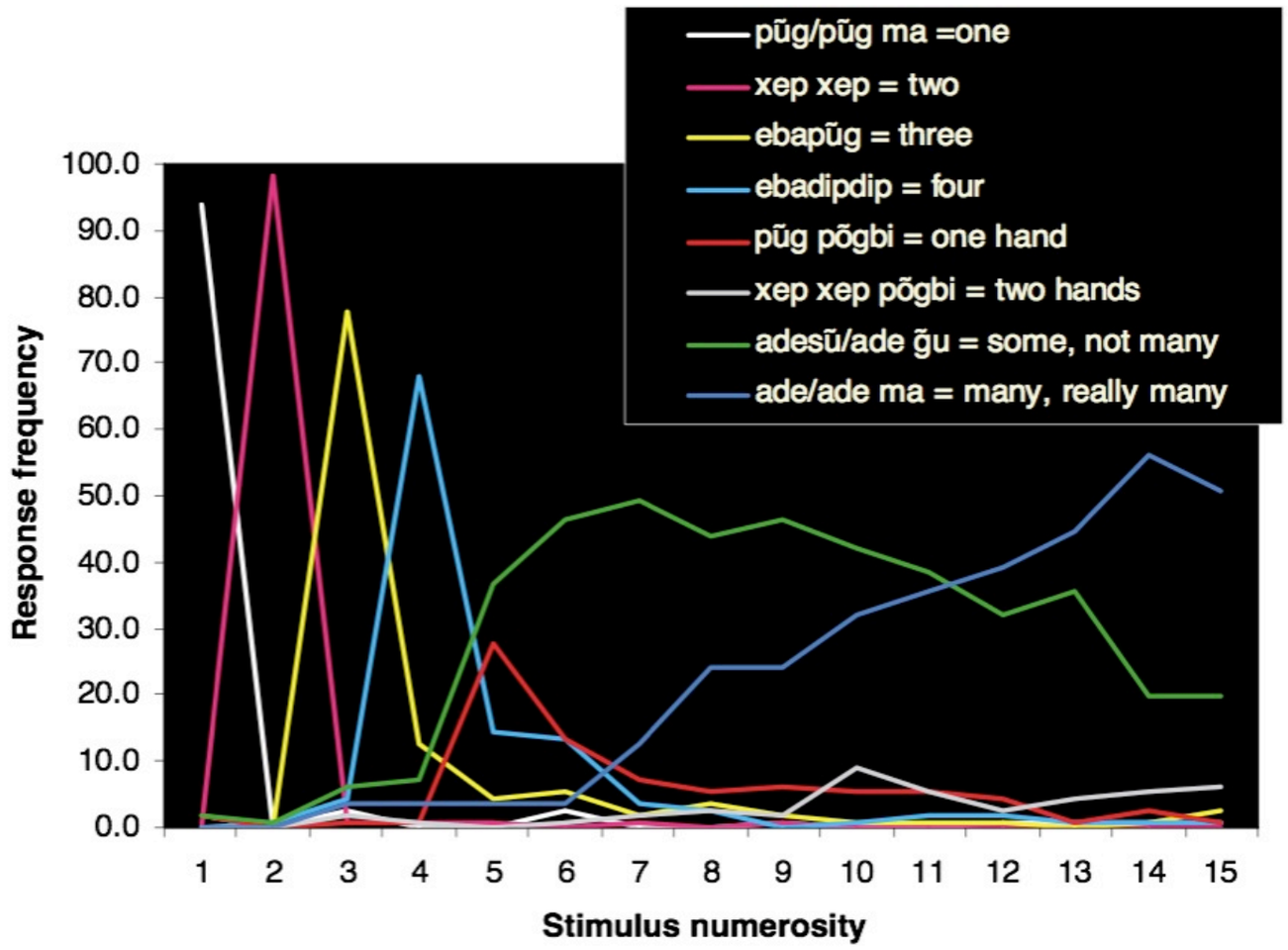
53, 57, 76, 77

Cheung et al. (2017, *Cog. Psych.*)

The Induction of Successor Function

- **Base**: 1+1=2 (infants and other animals know this innately)
- **Induction**: what's true of the **first two** elements is true of **all** elements in a **infinitely ordered** list
    - Can only do so by counting sufficiently high, in order to work out the productive **rule** of counting
- The concept of integers and discrete infinity is enabled by our linguistic ability to form rules
    - Only compelling case for Whorf (Pica et al. 2004, *Science*)

A language without productive numbers



- “the Munduruku do not have a counting routine ...By requiring an exact one-to-one pairing of objects with the sequence of numerals, counting may promote a conceptual integration of approximate number representations, discrete object representations, and the verbal code” (Pica et al. 2004)

Conclusion

- Children can find language from very small data so should we
  - Big data may even be harmful
- The cognitive processes of language acquisition may be deductively studied and contribute to the development of unsupervised NLP systems