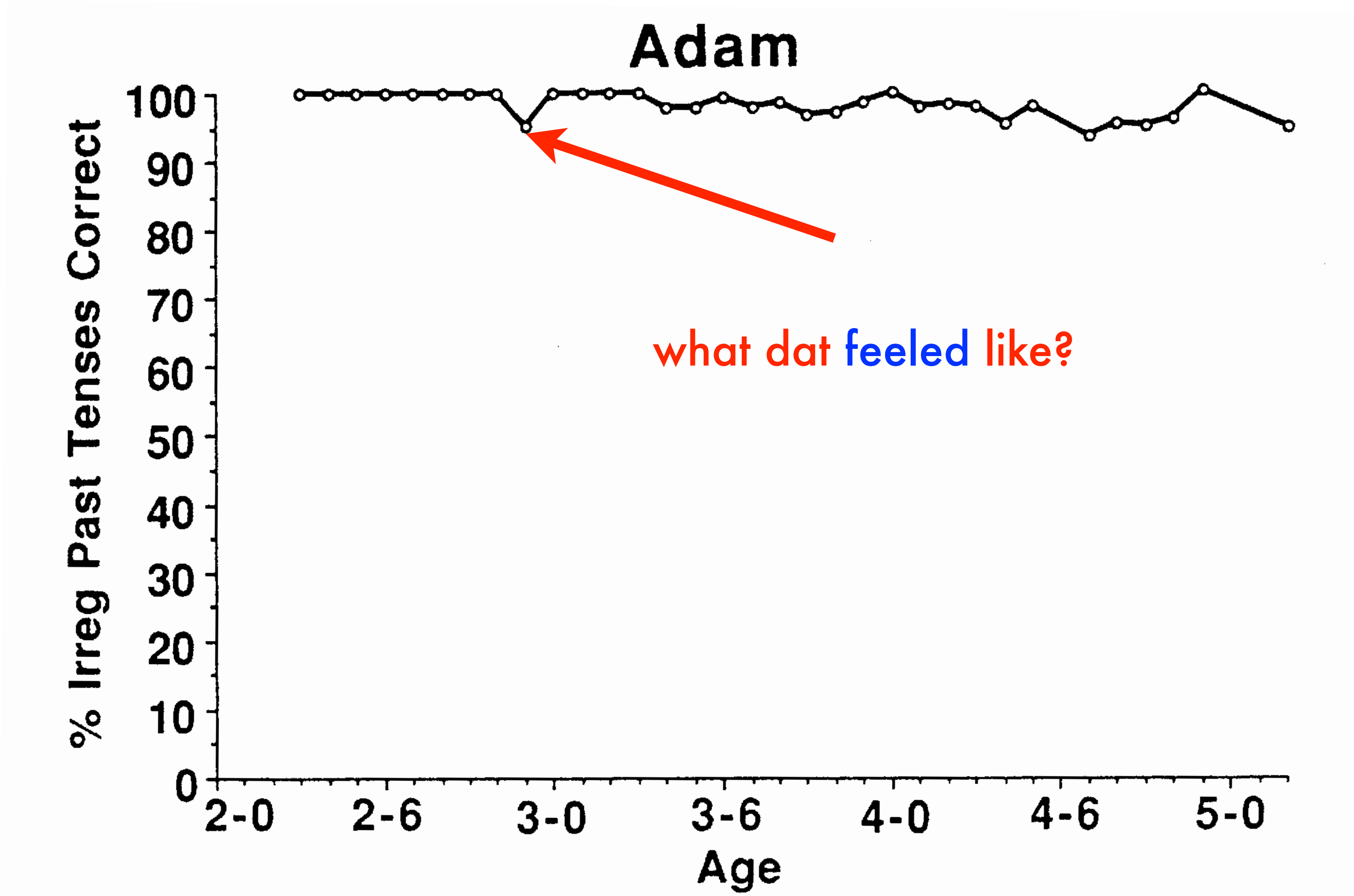
A Universal Law for Linguistic Generalization

(Why it is not optimal)

(What it has to do with Numbers)

Charles Yang
University of Pennsylvania

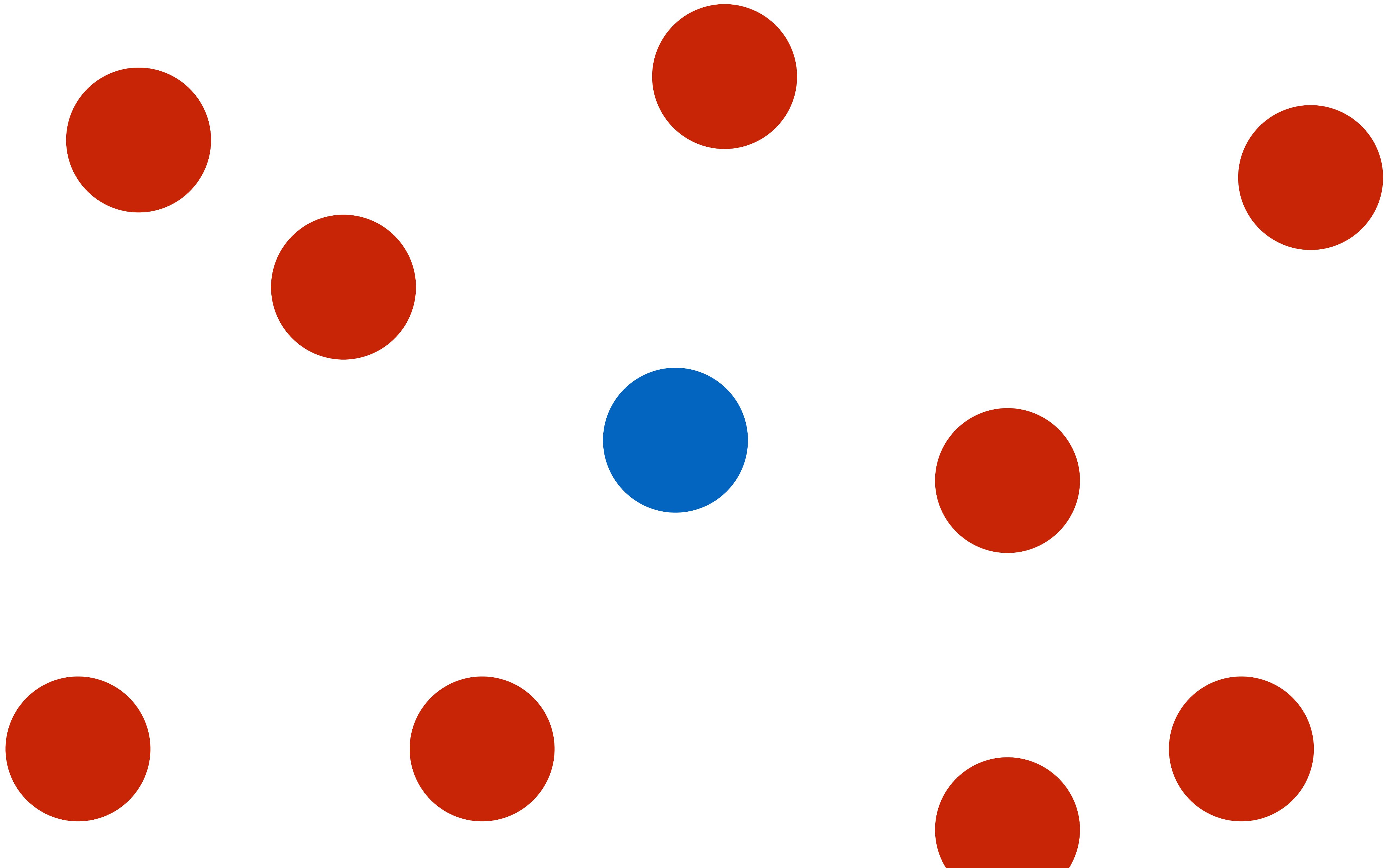
Lorentz Center 2016

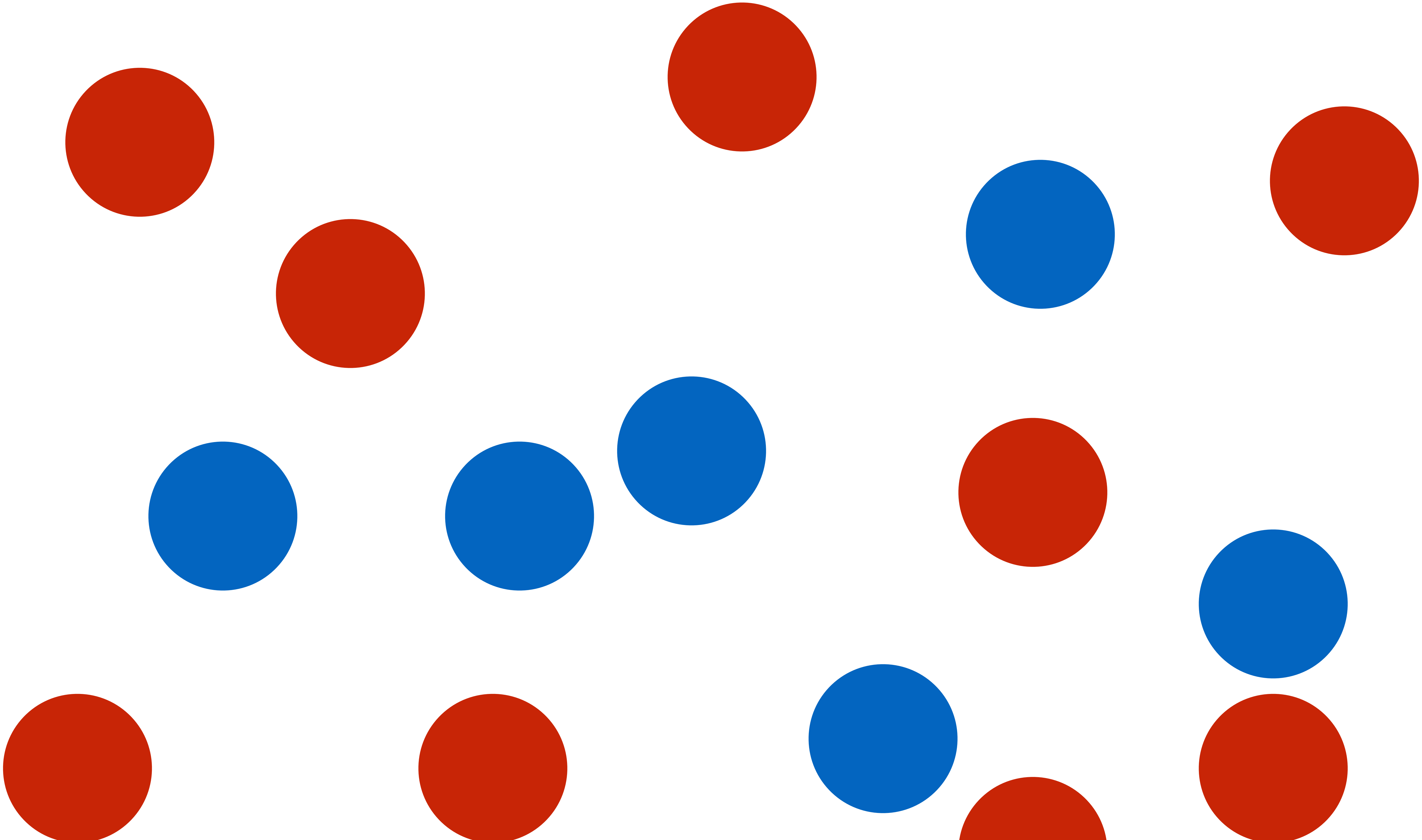


Adam
% Irreg Past Tenses Correct
100
90
80
70
60
50
40
30
20
10
0
2-0
2-6
3-0
3-6
4-0
4-6
5-0
Age
what dat feeled like?

All grammars leak

- Why does “-ed” emerge so later?
- How can a minority rule achieve productivity?
    - German noun plural suffix -s (4-5%; Marcus, Clahsen …)
- When a statistically dominate rule fails to generalize?
    - English words are overwhelmingly stress-initial (>80%; Cutler & Davis 1988) but does not fall a “quantity insensitive” system
- Where language breaks down?
    - The ineffables: *stride-strode-stridden, Russian inflection (Halle 1973), Polish singular masc. genitive (Dabrowska 2000), …





Roadmap

- A calculus for linguistic generalization (2016, MIT Press)
  - A selection of case studies: English, Polish, artificial language
- Why linguistic generalization is not optimal/Bayesian
  - Another empirical case study of English
- How linguistic generalization may be related to numbers
  - A speculation …

A decision rule

- Space?
  - Minimum Description Length (MDL; ≈Bayesian inference)
  - Difficult to identify an empirically motivated currency
- Time!
  - The organization of grammar (rules in balance with exceptions) favors **faster** systems (“third factor”: Chomsky 2005)

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

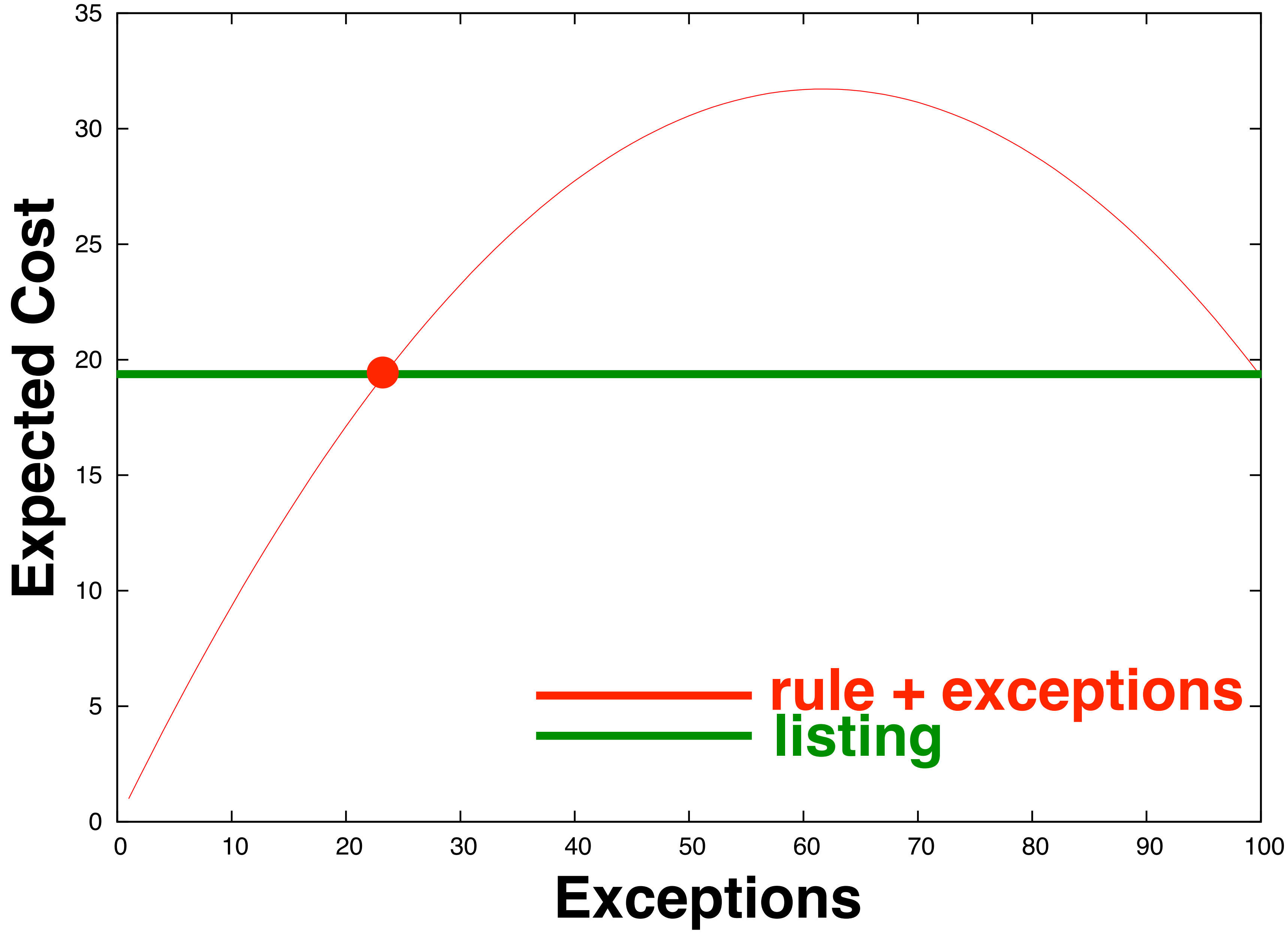
$<$

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

Tipping Point

- Let **R** be a rule applicable to **N** lexical items out of which **e** do not follow **R**. **R** is productive iff

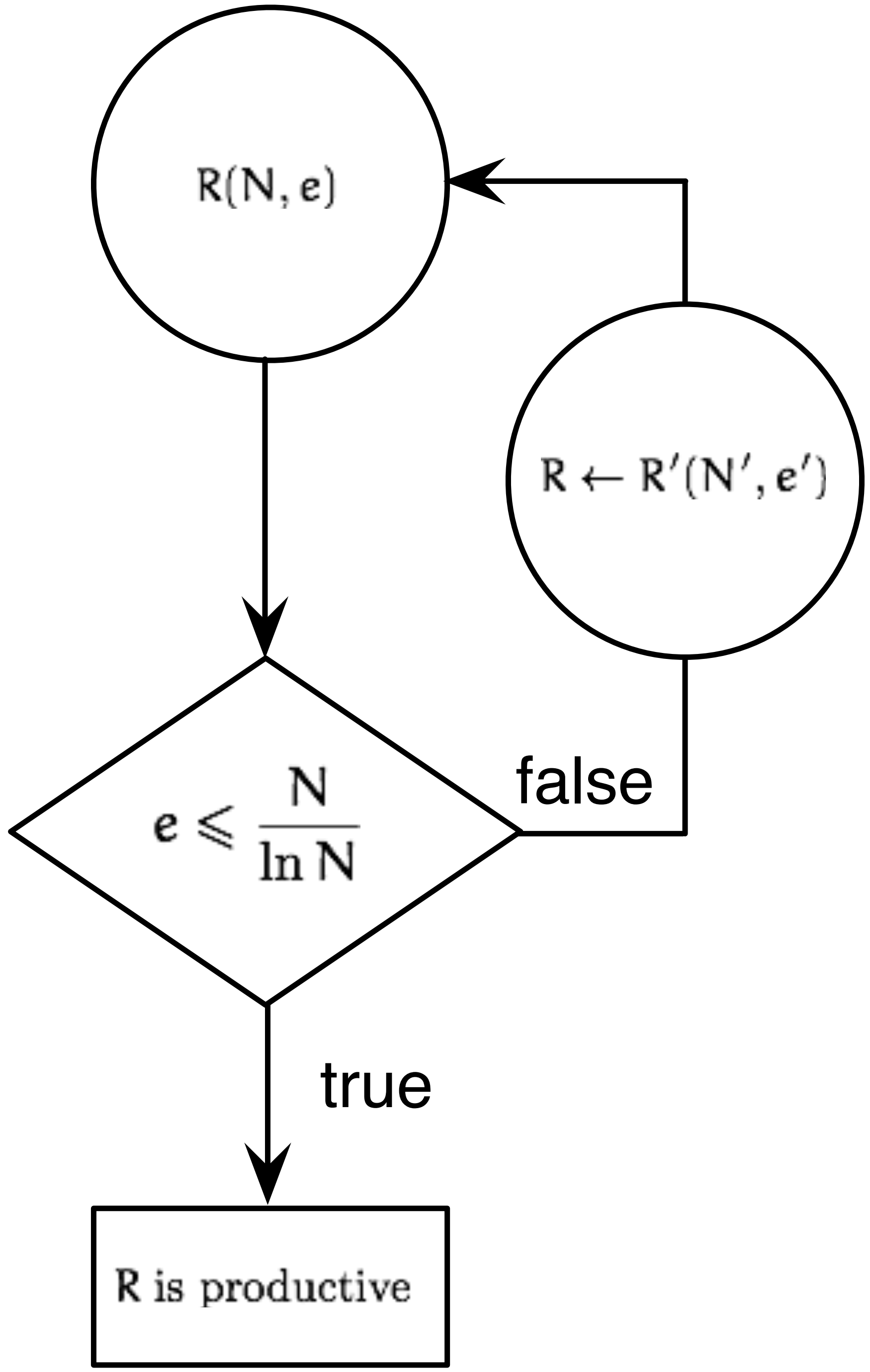
$$e \leq \theta_N := \frac{N}{\ln N}$$



N=100

Tolerance Principle

N	θ_N
10	4
20	7
50	13
100	22
200	38
500	80
1000	144
2000	263
5000	587
10000	1086



parameter/regression free

Space for individual variation

past tense rules “longitudinally”

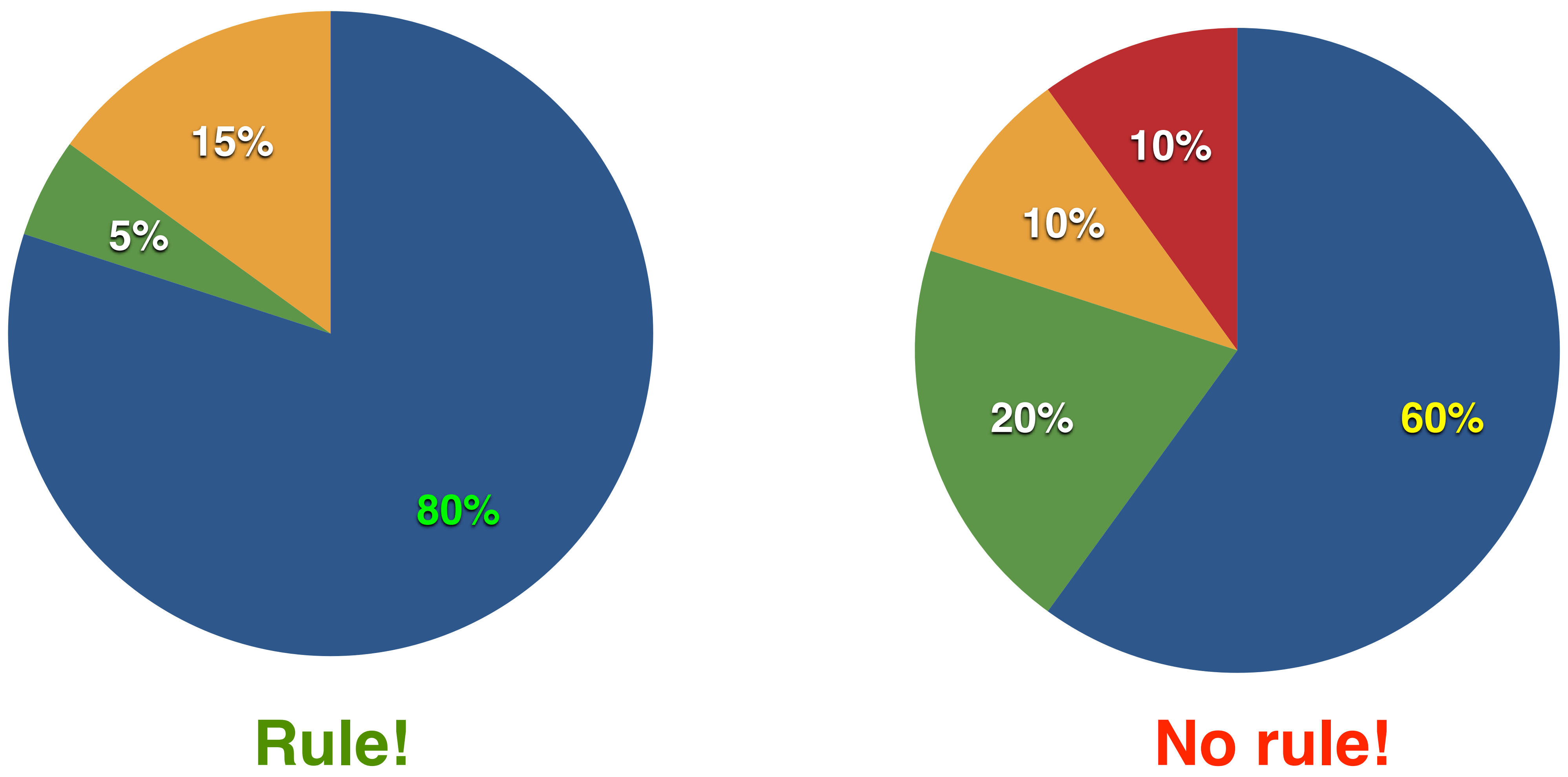
top N	sing→sang	feed→fed	fly→flew	-d	θ_N
100	—	—	**(8, 3)**	(100, 54)	22
200	**(3, 1)**	—	(10, 5)	(200, 76)	37
300	**(3, 1)**	—	(13, 8)	(300, 92)	52
500	**(5, 2)**	**(6, 3)**	(15, 10)	(500, 103)	80
800	(8, 5)	(11, 7)	(18, 13)	(800, 121)	119
1022	(8, 5)	(13, 9)	(22, 16)	**(1022, 127)**	147

6 million words of child-directed English

When **feel** became **feeled**?

- Adam: first instance of over-regularization (2;11) (feeled)
- Collect all word types and extract all verbal stems
- Adam used **N=300** verb stems, **e=57** irregulars
  - $\theta_{300}=300/\ln 300 = 53 \approx 57$
- Crucially, Adam did not learn **-ed** rule sooner: need filibuster proof majority

Collapse of Productivity



Absence of productivity:
(a) paradigmatic gaps (Halle 1973) and (b) language change

It can do minority productive rules (e.g., German plural -s; see Yang 2016)

Polish

- Polish **singular** masculine genitives take either -*a* or -*u* as suffix but neither seems to be the default based on a suite of tests (Dabrowska 2000).
  - **Plurals** take -*ow* as the default, with exceptional -*i/y* suffix

drut ‘wire’

rower ‘bike’

balon ‘balloon’

karabin ‘rifle’

autobus ‘bus’

lotos ‘lotus flower’

Polish Acquisition

- Analysis of child-directed Polish in CHILDES
- Error rates from Dabrowska (2000, 2005)

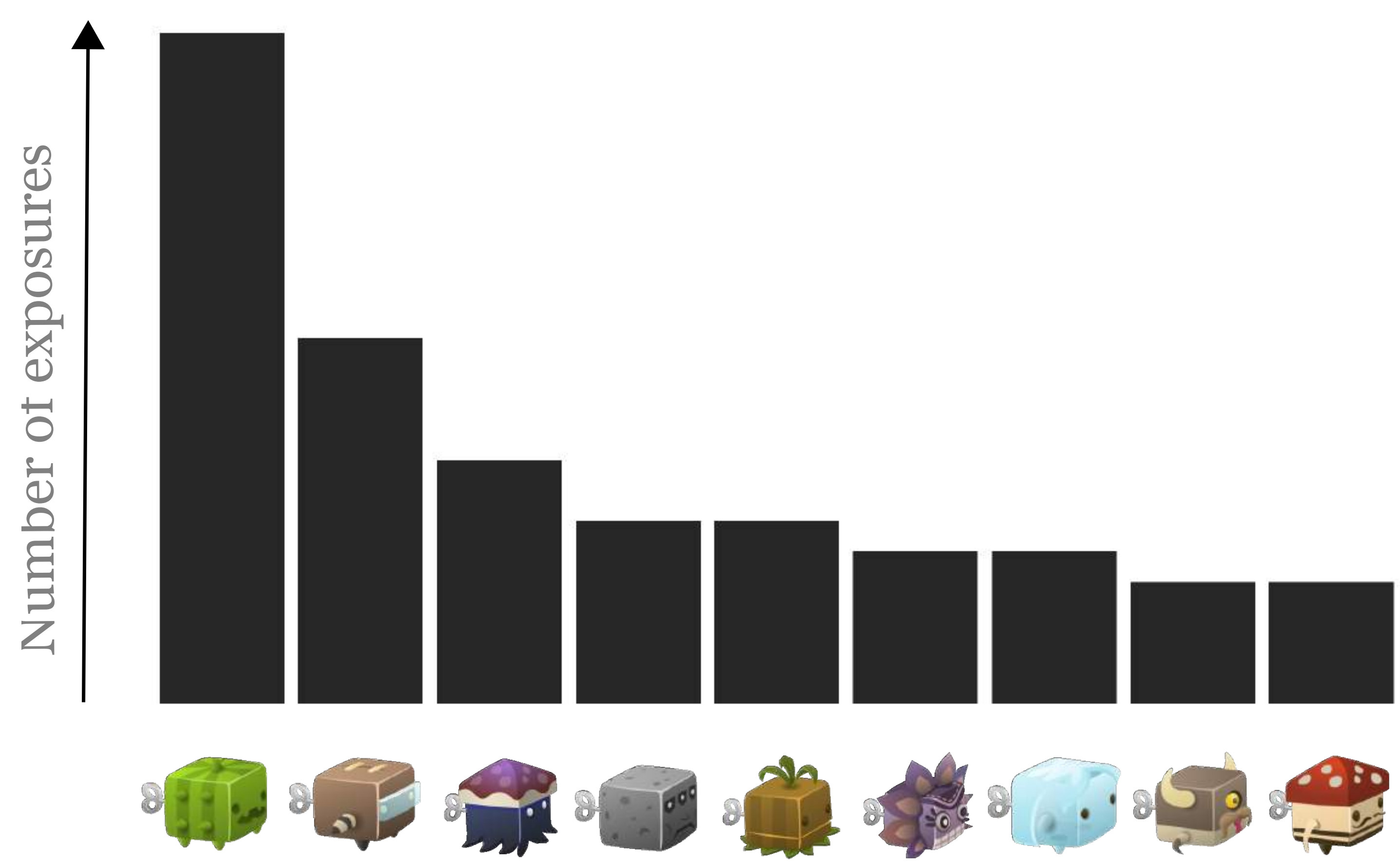
suffix	type freq.	productive?	ave. token freq.	error %
-a (sg.)	**837**	**no**	7.17	1.28%
-u (sg.)	**516**	**no**	8.8	0.24%
-ow (pl.)	**551**	**yes**	6.5	0.41%
-i/y (pl.)	**61**	**no**	11.4	**15.53%**

predict gaps and default: errors not a function of frequency (as in usage-based theories)

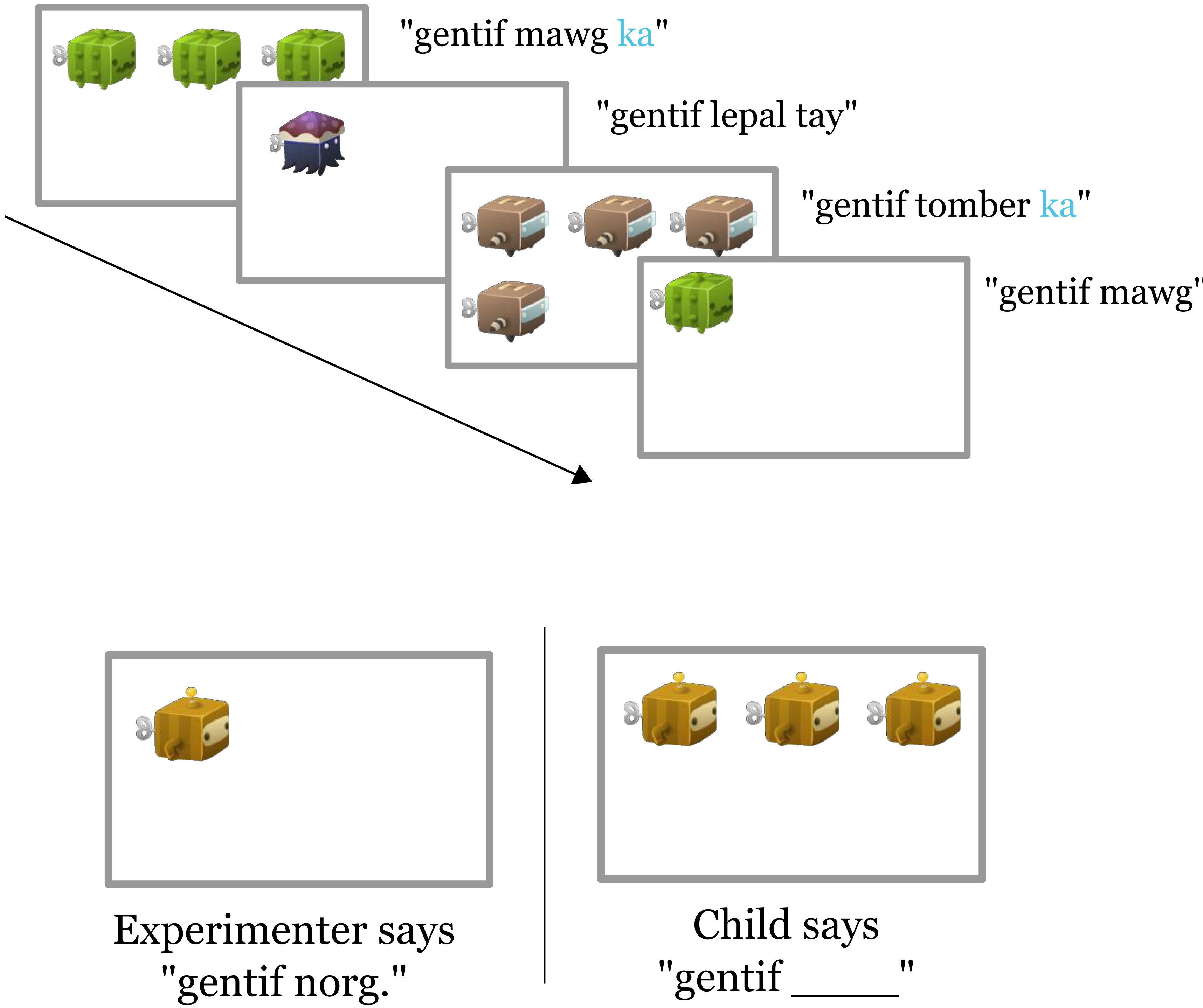
From the Lab

Schuler, Yang, and Newport (2016)

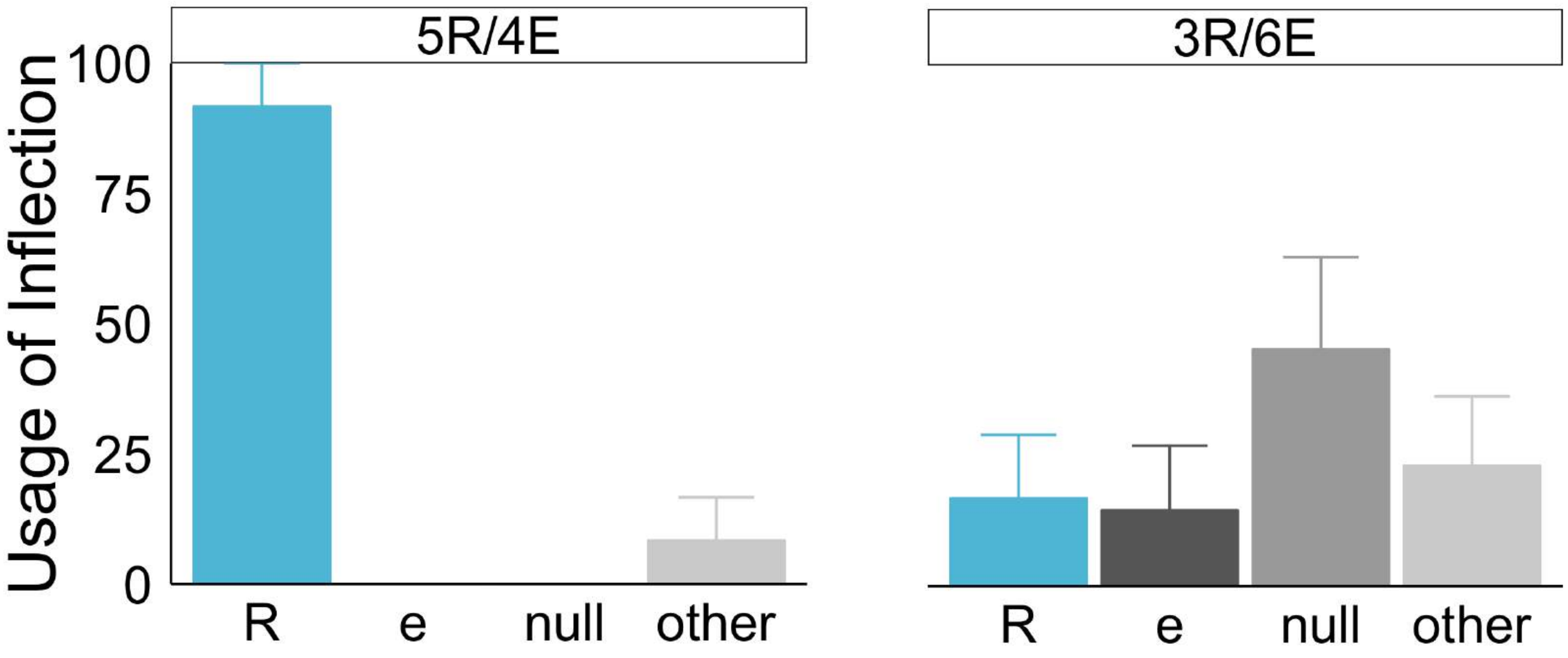
<table>
<tr><th rowspan="2">condition</th><th colspan="2"># of nouns following</th></tr>
<tr><th>Rule (R)</th><th>Exceptions (e)</th></tr>
<tr><td>5R/4e</td><td>5</td><td>4</td></tr>
<tr><td>3R/6e</td><td>3</td><td>6</td></tr>
</table>



Training and Wug test



Children are categorical!



English past tense! **Polish genitives!**

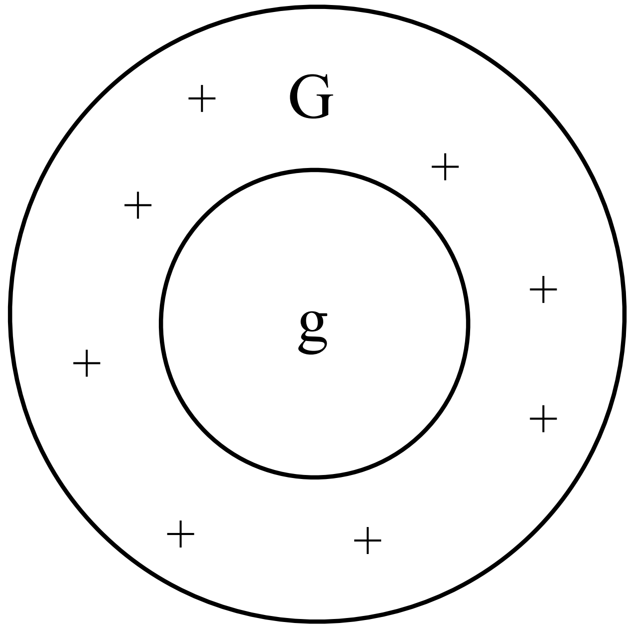
	5R/4E	3R/6E
# children using rule 100%	6	1
# children using rule <100%	1	7

No **frequency-based** regularization of -ka. Adults probability match.

Why it isn't Bayesian-optimal

I gave a book to the library.
I gave the library a book.

I donated a book to the library.
*I donated the library a book



- The target grammar is ***g***, and the learner has conjectured ***G:*** how does the learner back off to ***g***, without negative evidence?
- Hypothesis ordering (Angluin 1980, Berwick 1985)
- Indirect negative evidence (Gold1967, Chomsky 1981)
  - Pinker (1989, p40): "virtually a restatement of the original learning problem."

Bayesian Inference

- Prior probabilities: P(g) and P(G)
- Data: D
- Likelihood function: P(D|g) and P(D|G)
  - A hypothesis assigns a probabilistic distribution to the set of admissible strings
  - Unattested but admissible strings lowers P(D|G)
- Posterior probability: P(g)P(D|g) vs. P(G)P(D|G)
- Select the better hypothesis

There's another way to phrase that and that is that the ***absence of evidence is not evidence of absence****. ... Simply because you do not have evidence that something exists does not mean that you have evidence that it doesn't exist.*

Complexity

- How to determine which is bigger hypothesis (g vs. G)
- Innate hypothesis ordering: can’t be entirely true
- Computed online:
  - generally uncomputable (Osherson, Stob, & Weinstein 1986)
  - provably intractable (Luby 1993, Chickering et al. 2004, Fodor & Sakas 2011)
  - even approximation methods are NP-hard (Kwisthout et al. 2013)

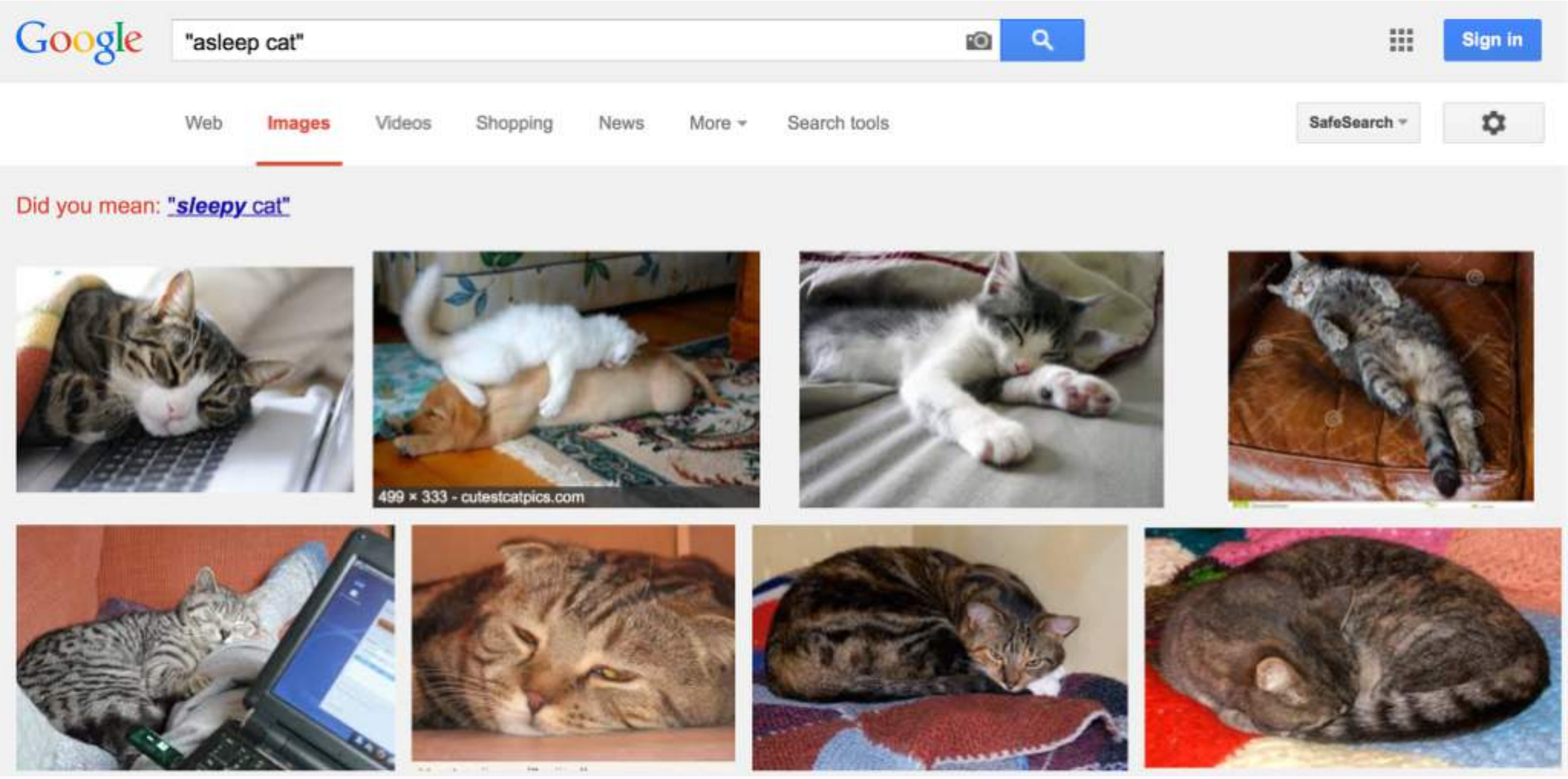
Non-optimal performance

- Perfors, Tenenbaum, & Wonnacott (2010): dative constructions via hierarchical Bayesian models
  - MLE works equally well (Villavicencio, Berwick, & Malioutov 2013)
- Frank, Goodman, & Tenenbaum (2009): Bayesian model of word learning
  - A reinforcement learning model performs better on child-directed English data (Stevens, Trueswell, Gleitman, & Yang, 2014, under review)

Optimality

- But language is inherently variant (Weinreich et al. 1968)
  - Children are capable of acquiring probabilistic/variable rules along with their structural conditionings (Labov & Roberts 1996)
  - Competing grammars in language learning and change (Kroch 1989, Yang 2000, Han, Lidz, & Mussolini 2016): Optimality predicts no change
  - Probability matching remains a challenge for Bayesian models (Suppes 1966) and straightforwardly applies to language variation (Yang 2002, 2002)
- The best still is not good enough: Morphological gaps (Polish, Russian, English, Spanish, French …)

Effectiveness: Why there are no asleep cats

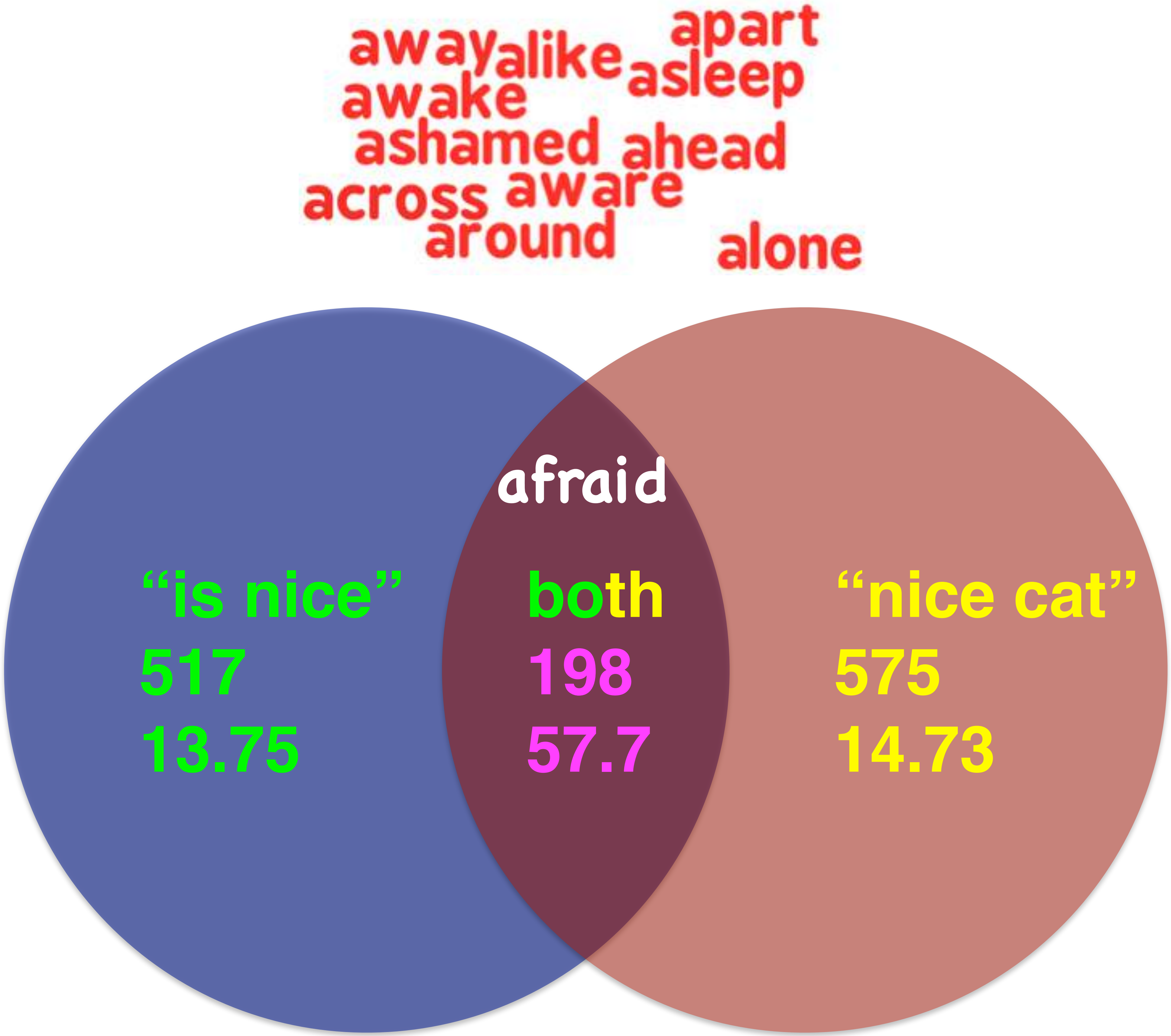


Frequency-based expectation: sufficiently frequent adjectives failing to appear as **AN**

Paraphrase (Uniqueness, Mutual Exclusivity, Principle of Contrast): "
the cat that is asleep" i**nstead of AN**

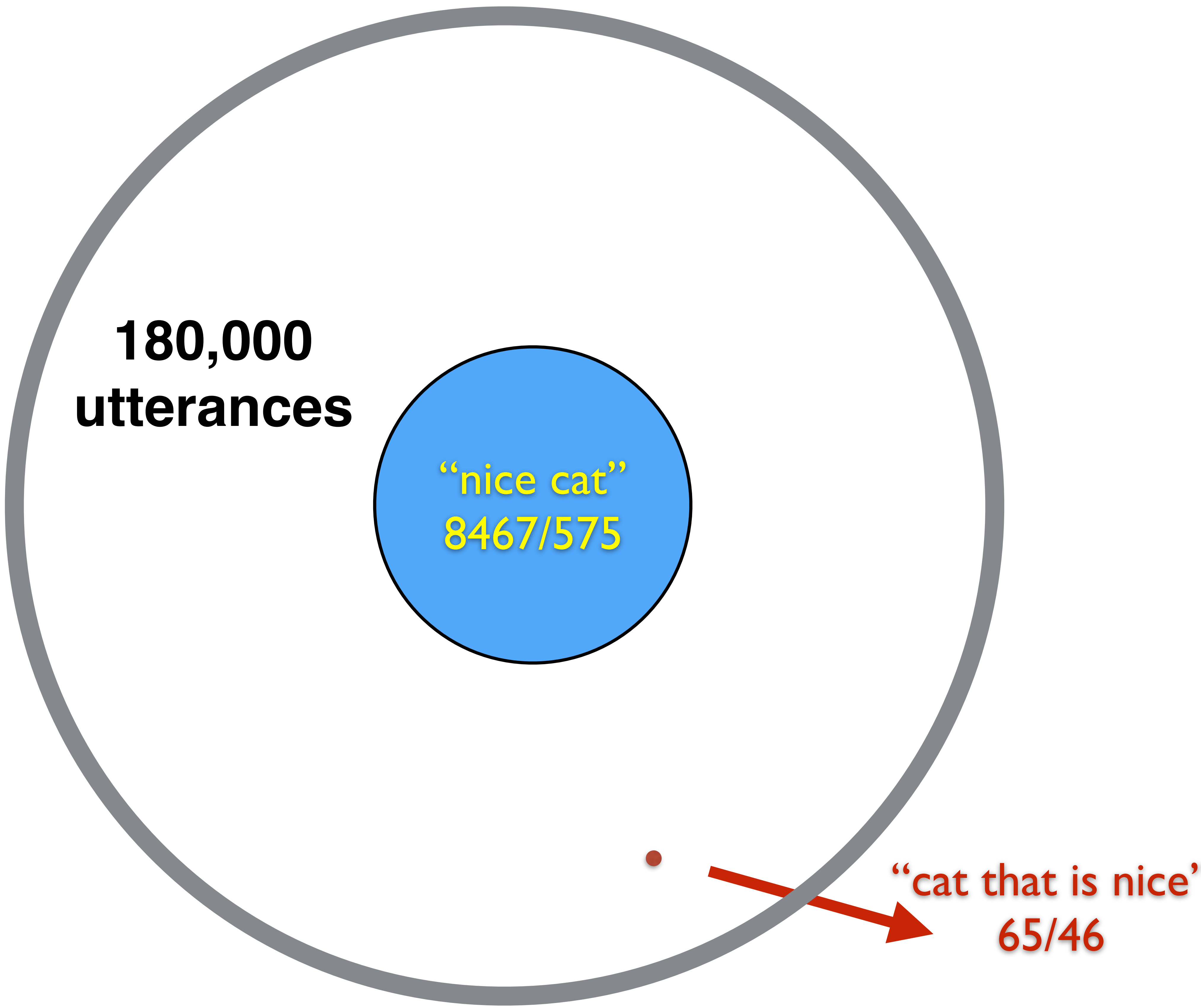
Data

- 180,000 parsed sentences from child-directed English
- 6 million words of child-directed English text



many more frequent adjectives only appeared predicatively but can appear attributively without difficulty (*careful*, *sorry*, *ready*, …)

Finding Paraphrases



440,000 words of CDS

torn

on

careful

sure

annoying

washable

handicapped

afraid

jolly

gone

interested

alike

shaped

supposed

heavier

off

wrapped

6 million words of CDS

torn

asleep

on

careful

annoying

washable

afraid

jolly

gone

interested

alike

wrapped

off

A positive solution

- The statistical distribution of a-adjective usage does not rule out the over-hypothesis: indirect negative evidence, Bayesian or otherwise, is ineffective

- Solution: Avoid over-hypothesis!

A- is the key

- Creation of new words (Salkoff 1983, *Language*)
- The tree is abud with green shoots.

  ? An abud tree is a beautiful thing to see.

- The water is afizz with bubbles.

  ? The afizz water was everywhere.

A+stem

- Larson & Marusic 2004: a**beam**, a**blaze**, a**bloom**, a**buzz**, a**cross**, a**drift**, a**fire**, a**flame**, a**fraid**, a**gape**, a**ghast**, a**gleam**, a**glitter**, a**glow**, a**ground**, a**head**, a**jar**, a**kin**, a**light**, a**like**, a**live**, a**lone**, a**miss**, a**mok**, a**muck**, a**part**, a**round**, a**shamed**, a**shore**, a**skew**, a**slant**, a**sleep**, a**stern**, a**stir**, a**tilt**, a**wake**, a**ware**, a**whirl**, a**wash,** a**way**
- Non-a-adjectives:
  - The a**bove** examples
  - The a**loof** professor
  - The a**lert** student
  - The a**stute** investor
  - The a**mazing** car

A-adjectives are not Atypical

- The teacher is present. *The present teacher
- The receptionist is out. *The out receptionist.
- The batter is up. *The up batter.
- The runner is on. *The on runner.
- The game is over. *The over game.
- The delivery is here/there. *The here/there delivery.

Prepositional Phrases

- The ball is out of sight. ?*The out of sight ball.
- The dog is behind the fence. ?*The behind the fence dog.
- The singers are at ease. ?*The at ease singers.
- The marbles are in the jar. ?*The in the jar marbles.

Special adverbs

- I was well/wide **awake** at 4am.
- The race leader is well **ahead**.
- The baby fell right/sound **asleep**.
- You can go right **ahead**.
- The guards are well **aware** (of the danger).

The cat came straight **out**.

The cat ran straight **out of the house.**

The answer was wide **off**.

The answer was wide **off the mark.**

The arrow was shot well **over**.

The arrow was shot well **over the fence.**

The ball sailed far **out**.

The ball sailed far **out of the park.**

*The car is right/straight/well **new/nice/red**.

*The politician is right/straight/well **annoying/amazing/available**.

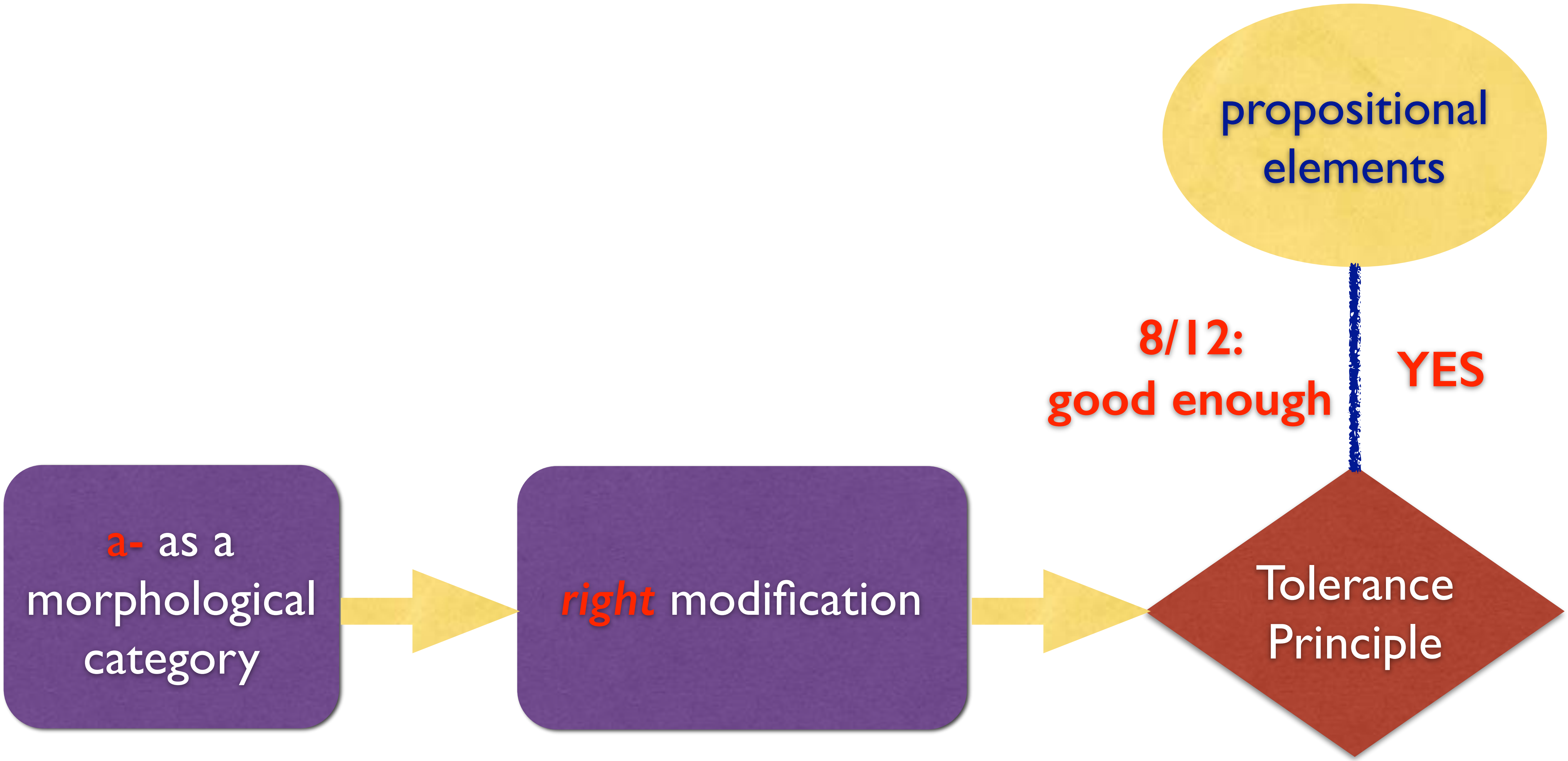
Morphological Partitioning

- Morphological condition: extracted all adjectives, removed initial schwa to obtain a phonotactically valid remainder
  - Assume children can do morphological segmentation (Brown 1970, Peters 1983)
  - Analysis of 6 million words of child directed English
- **afraid, awake, aware, ashamed, ahead, alone, apart, around, asleep, alike, away, across**
- **amazing, annoying, allergic, available, adorable, another, american, attractive, approachable, acceptable, agreeable, affectionate, adept, above, aberrant**

Adverbial Modification

- 8 out of 12 a-adjectives are ***right*** modified (3 to >100)
  - are you **wide** awake?
  - I'm **well** aware of my shortcomings
  - thank you go **right** ahead.
  - it fell **right** apart on you.
  - turn **right** around.
  - finish the book **right** away.
  - he fell **fast** asleep.
  - we are coming **right** across.
- No non-a-adjectives are ***right*** modified at all
- Numerous instances of ***right here***, ***right under the table***

It walks like a duck, quacks like a duck …



Big data may be harmful!

- Speakers reliably have judgement for a-adjectives
- 12 appear in a year's speech, 8 have **signatures**
- 51 million word corpus (SUBTLEX-US), only 28 a-adjectives
  - very unlikely that most of them have signatures
- If the child were to know many more a-adjectives, the generalization may not be valid:
  - **8/12** is fine, but **10/28** is not fine

Early productivity

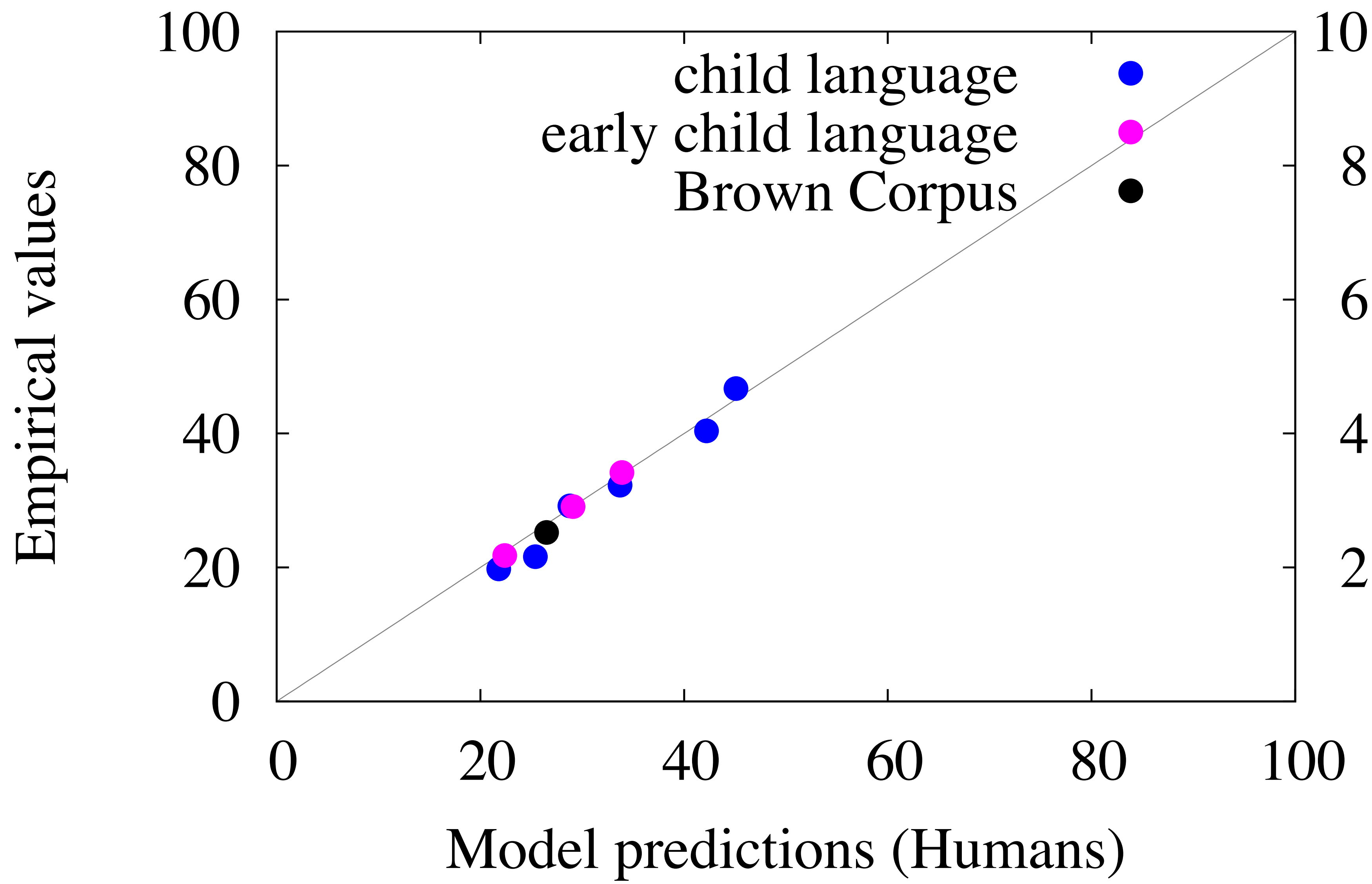
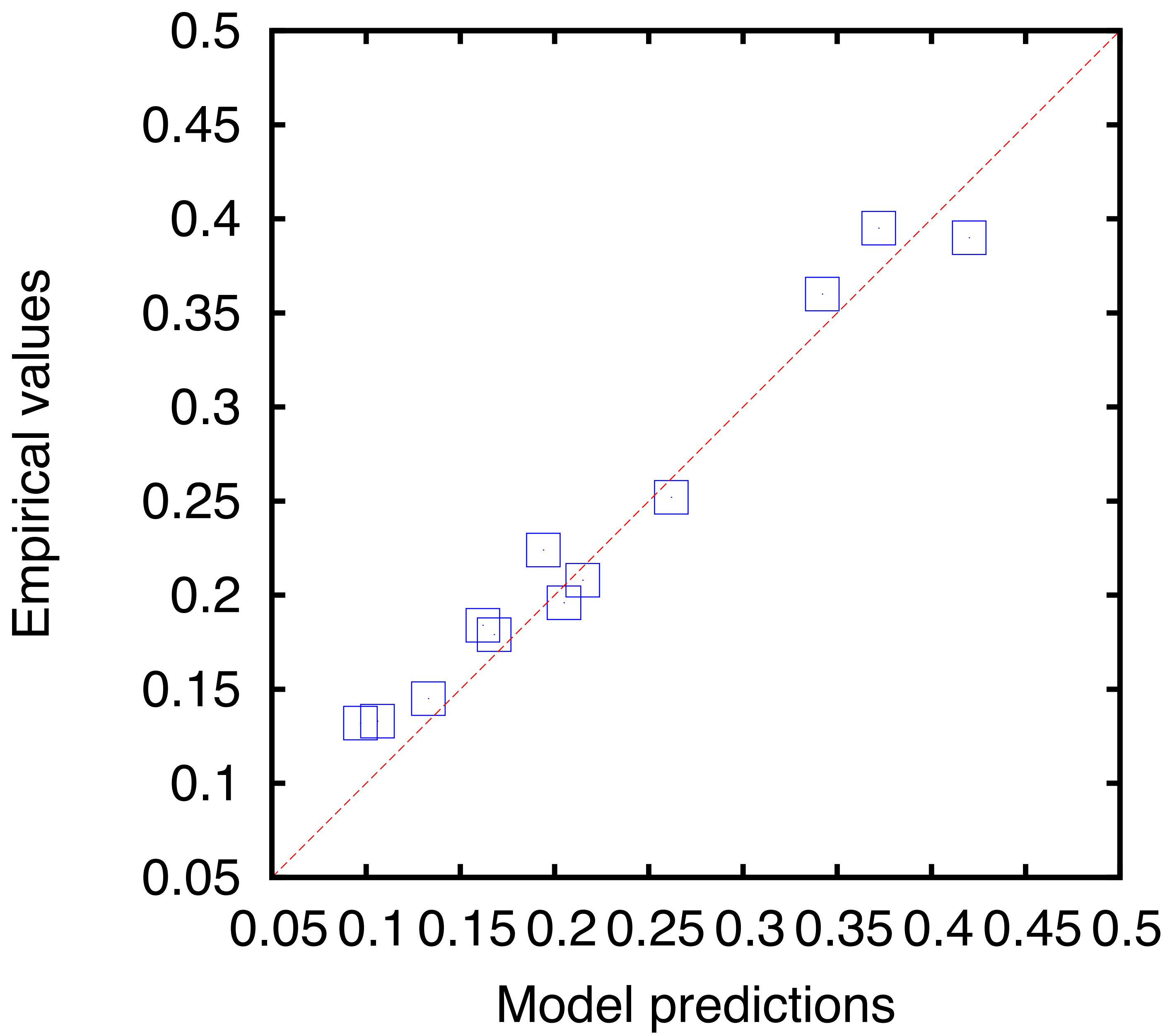


Figure 1a

- Yang (2013, *PNAS*): children use determiners (*the* and *a/n*) and nouns interchangeably, contra Tomasello (2000)
- Lin's CCC: ρ_c=0.977, 95% CI 0.925-0.993

Productivity without a Model



- David's homesign combinations: nominals with 12 predicate classes
- Goldin-Meadow & Yang (in press): ρ_c=0.975, 95% CI 0.926-0.992.

Adam's Mother

- Adam's mother:
  - 914 singular nouns, only **34**% are used with both **a/n** and **the**
- A third of the batters are observed to switch-hit:
  - Do **all** of them switch hit?

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.

Less may/must be more

- Top **50** most frequent nouns: **43** appear with both
- Top **100** most frequent nouns: **87** appear with both
- Limiting attention to the most frequent, and potentially more "important", items in language may be necessary for the successful acquisition of language (Newport, Elman)
- Current projects jointly led with Mitch Marcus are exploring and exploiting these ideas in NLP

Part III: Numbers

- Successor function: Discrete infinity
- Claim: Learning a **productive** numeral system, i.e., the language-specific morphosyntactic system that goes on in a predictable fashion, is a **sufficient** condition for inducing/learning/triggering the successor function
- Proposal: Identify the conditions under which a productive numeral system can be acquired
  - E.g., the child needs to learn the numeral word up to **N** such that a productive **rule** can be learned in the presence of **idiosyncratic** numerals

Example I: English

one two three four five six seven eight nine ten

eleven twelve thirteen fourteen fifteen sixteen seventeen

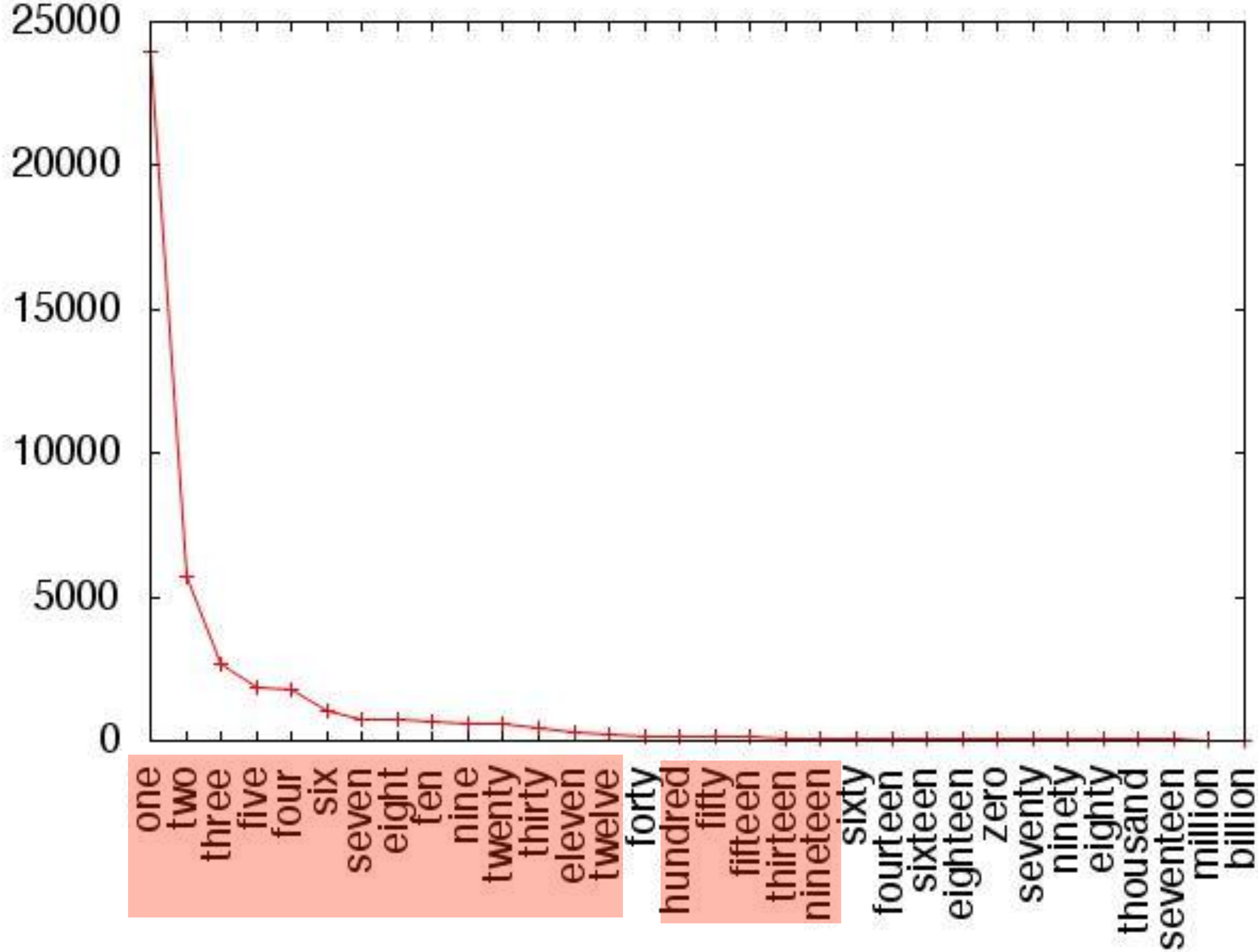
eighteen nineteen twenty 21 22 23 24 25 26 27 28 29 30

31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 50

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

e_{max}=17, min(N_{max})=73

If 1-10 are **idiosyncratic**, the child needs to learn a lot more numerals than 1-10 to have successor function even for **small** numbers


25000
20000
15000
10000
5000
0
one
two
three
five
four
six
seven
eight
ten
nine
twenty
thirty
eleven
twelve
forty
hundred
fifty
fifteen
thirteen
nineteen
sixty
fourteen
sixteen
eighteen
zero
seventy
ninety
eighty
thousand
seventeen
million
billion

Example II: Chinese

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。。。

$e_{max}=11$, $\min(N_{max})=40$

Suggestions

- Successor function may be induced/triggered by the acquisition of the productive numeral system in a specific language
  - which requires a vocabulary of certain size (not necessarily consecutive)
- There will be cross-linguistic differences
  - Including languages that may fail to trigger the successor function (Barner, Pica, Spelke)
- There will be cross-individual differences
  - Children with larger vocabulary may have an advantage

Back to language

- Language/number has the **capacity** for discrete infinity
- Different languages allow for different degrees of recursion in different domains
  - Maria’s neighbor’s friend’s house (English)
  - * Marias Nachbars Freundin Haus (German)
  - And there is Piraha …
- Proposal: a single learning system for language and numbers

Summary

- Language poses a lot of hard problems for learning theories
  - Specific linguistic details matter
- (Some) learning models are computationally complex and may not be effective
  - Traditional learnability research remains useful, and models that bridge Marrian levels should be the target
- Tightening the connection between language and numbers
  - One way to do so is to see how language learning supports number cognition

Relevant publications

(2002) *Knowledge and learning in natural language.* OUP. [Probabilistic (reinforcement) learning for UG]

(2015) Negative knowledge from positive evidence. *Language*. [On the a-adjectives and their acquisition]

(2016) *The price of linguistic productivity: How children learn to break rules of language*. MIT Press. [The Tolerance Principle and many case studies]

(2016) Schuler, Yang, and Newport. Testing the Tolerance Principle. (Artificial language under Tolerance Principle)

(Forthcoming) Rage against the machine: Evaluation Metrics in the 21st century. *Language Acquisition.* (Bayesian critique, Marr levels)

(In press) Goldin-Meadown & Yang. Statistical evidence that a child can create a combinatorial linguistic system without linguistic input:Implications for language evolution. *Neuroscience and Biobehavioral Review*

(Under review) Stevens, Trueswell, Gleitman, & Yang. Pursuit of word meanings. (Resource-limited online learning vs. cross-situational learning)

Thanks

- Susan Goldin-Meadow
- Lila Gleitman
- Mitch Marcus
- Elissa Newport
- John Trueswell

Constantine Lignos
(BBN)

Kyle Gorman
(Google Research)

Kathryn Schuler
(Georgetown)

Jon Stevens
(ZAS Berlin)