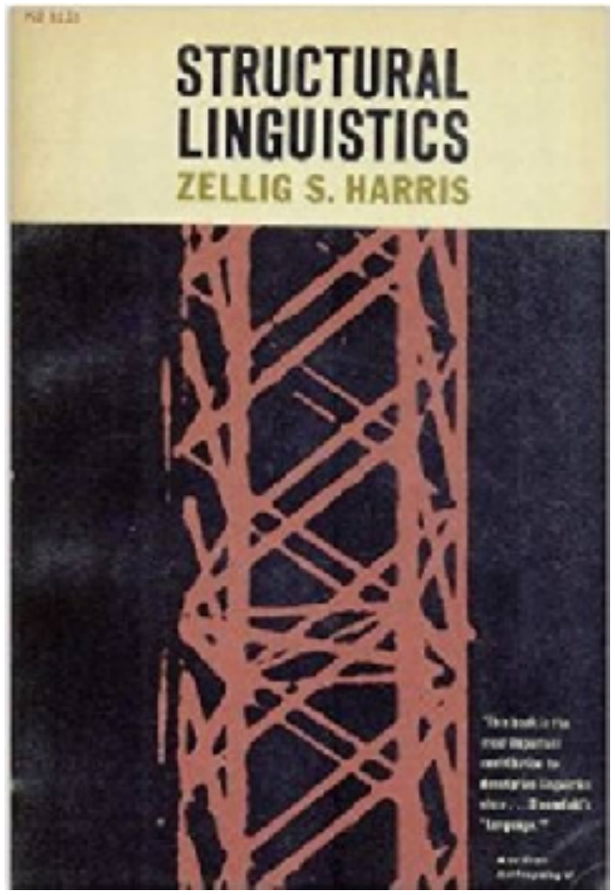
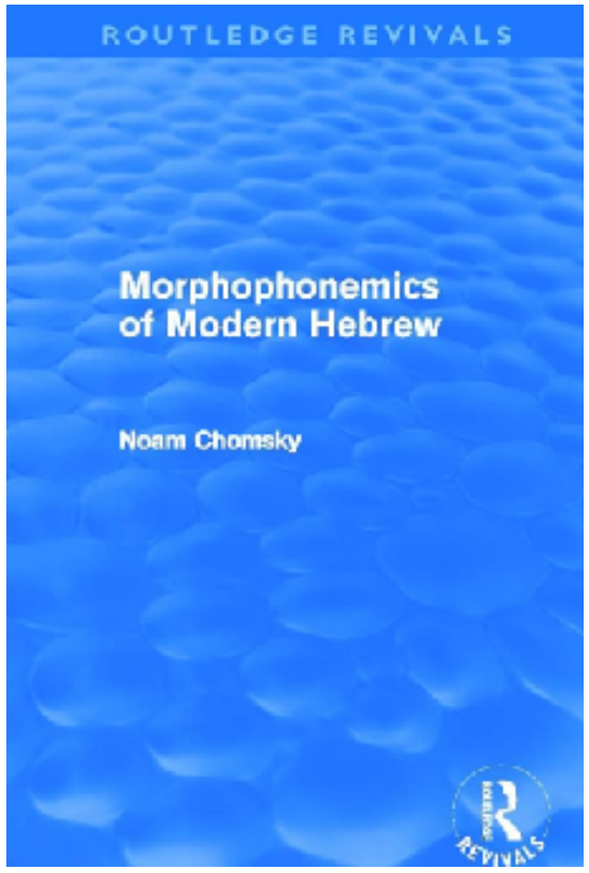
Discovering the Limit of Language Variation

Charles Yang
University of Pennsylvania

The original end-to-end systems



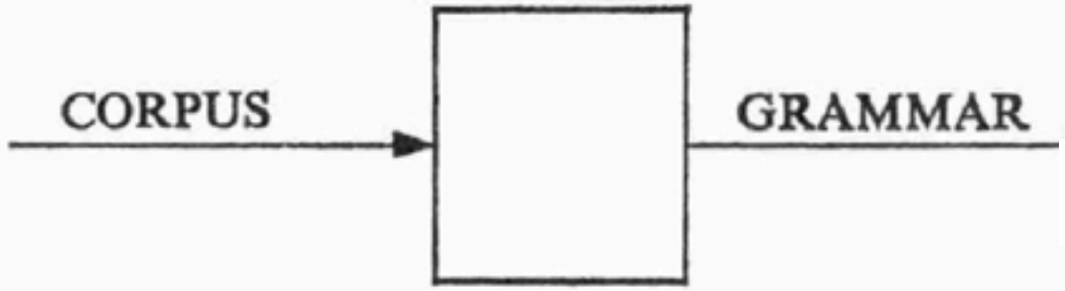


THE LOGICAL STRUCTURE OF LINGUISTIC THEORY

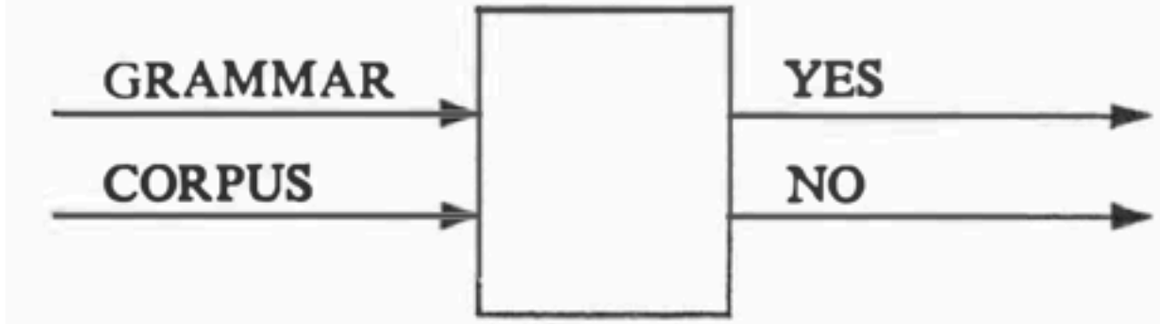
Noam Chomsky — June, 1955 - 56

Massachusetts Institute of Technology

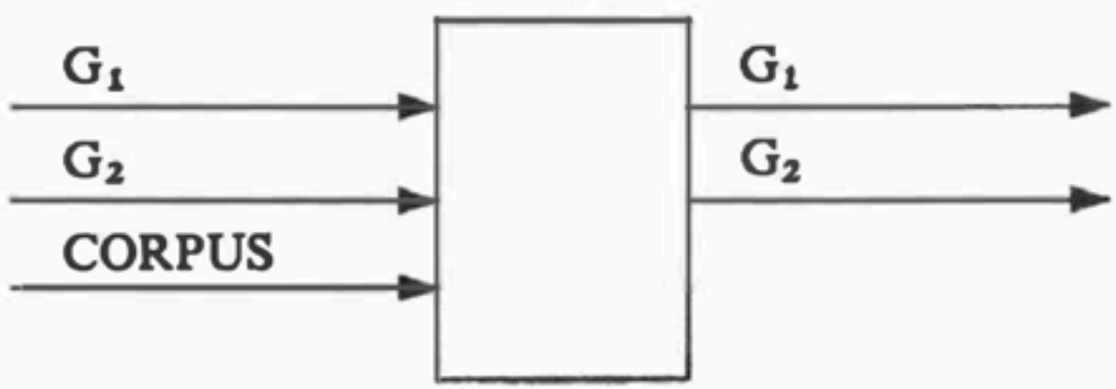
Discovery, Decision, Evaluation



The strongest requirement that could be placed on the relation between a theory of linguistic structure and particular grammars is that the theory must provide a practical and mechanical method for actually constructing the grammar, given a corpus of utterances. Let us say that such a theory provides us with a *discovery procedure*

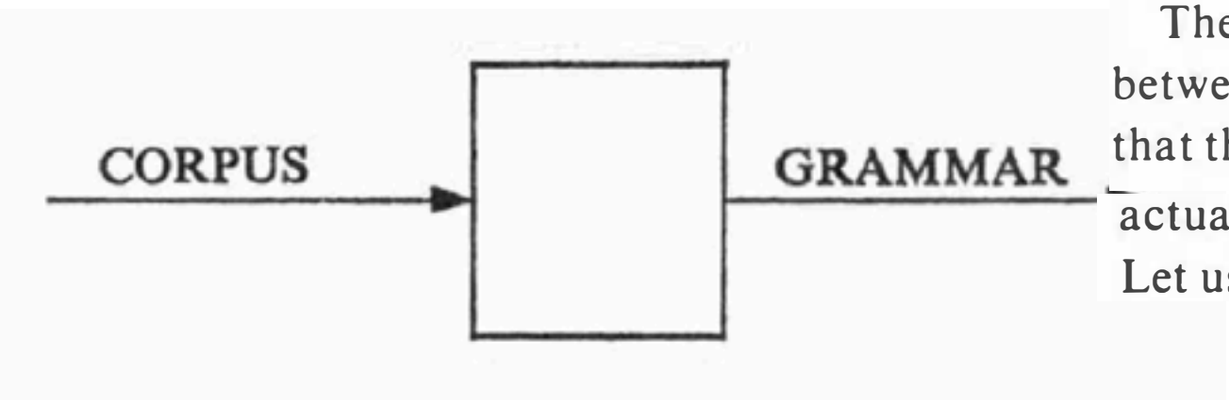


A weaker requirement would be that the theory must provide a ractical and mechanical method for determining whether or not a rammar proposed for a given corpus is, in fact, the best grammar f the language from which this corpus is drawn. Such a theory, hich is not concerned with the question of *how* this grammar was onstructed, might be said to provide a *decision procedure* for rammars.



even weaker requirement would be that given a corpus and two proposed grammars G_1 and G_2, the theory must tell us is the better grammar of the language from which the corpus awn. In this case we might say that the theory provides an *ation procedure* for grammars.

A discovery procedure: Premature

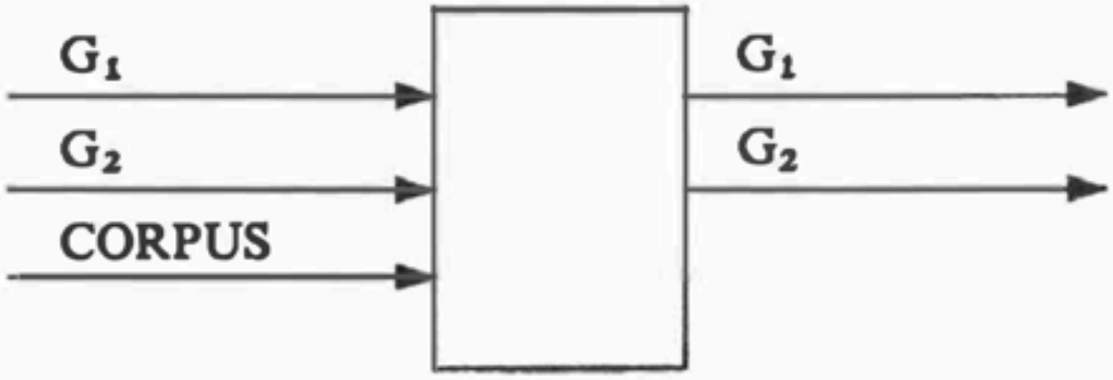


The strongest requirement that could be placed on the relation between a theory of linguistic structure and particular grammars is that the theory must provide a practical and mechanical method for actually constructing the grammar, given a corpus of utterances. Let us say that such a theory provides us with a *discovery procedure*

- "Practical and mechanical": goes without saying, interpretable.

- But it was premature:
  - No corpus: Brown Corpus 10 years away.
  - No theory of learning or computation: No programming languages.
  - No understanding of child language: Berko's Wug test (1958).

But does it help?

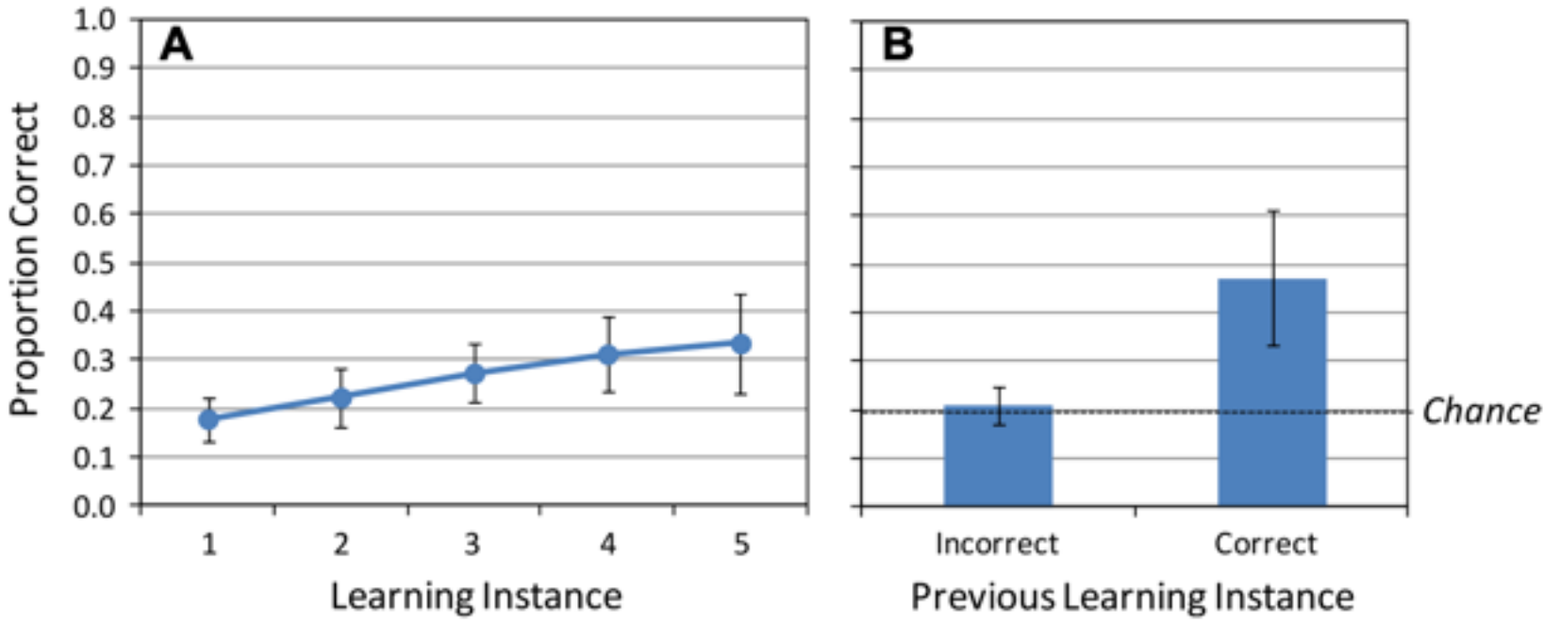


even weaker requirement would be that given a corpus and two proposed grammars G_1 and G_2, the theory must tell us is the better grammar of the language from which the corpus awn. In this case we might say that the theory provides an *ation procedure* for grammars.

- Knowing more means rejecting more, and it's not easy!
- How much of nativism are we—or Chomsky himself!—willing to put up with?
- The search for the best grammar given a corpus, defined in information-theoretic terms in LSLT, comes at a tremendous/inconceivable cost so it's no longer practical and mechanical.

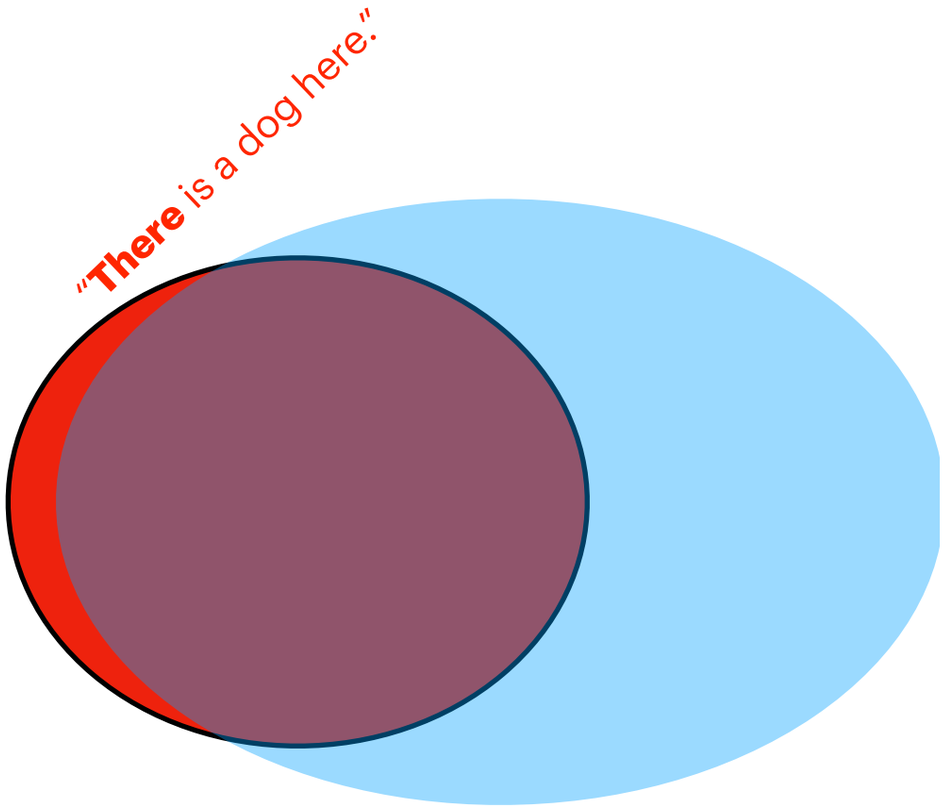
The most probable word?

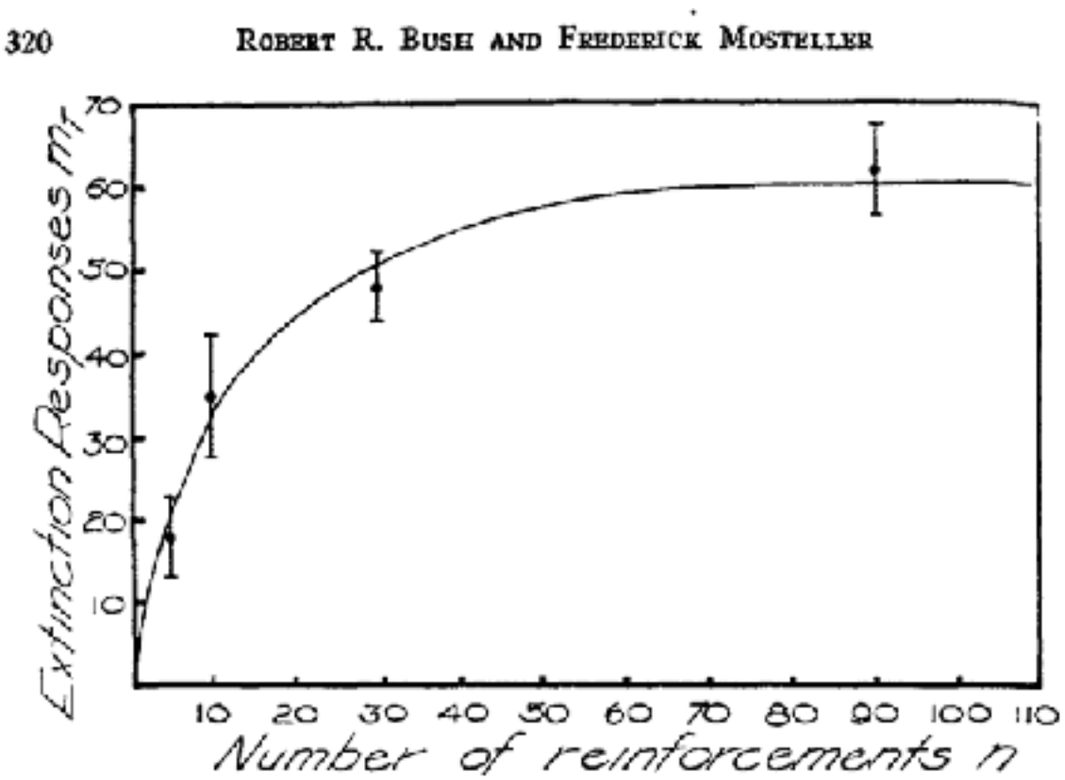
- Why can’t do we better with words?
  - A collaboration with Lila Gleitman and John Trueswell along with Jon Stevens and Christine Soh Yue.

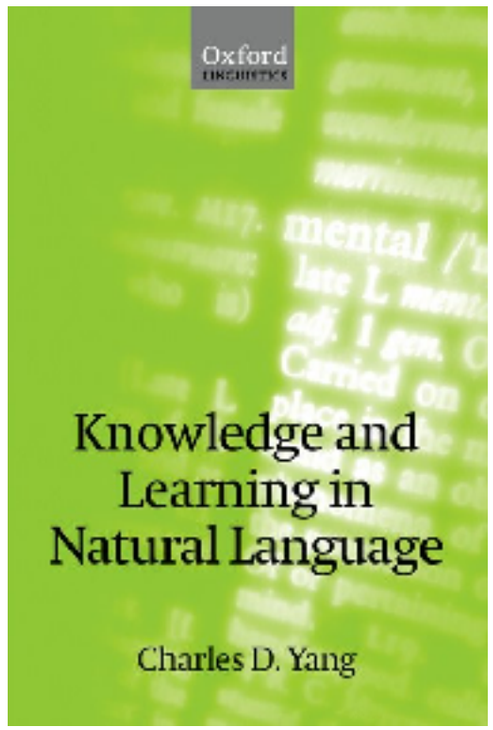


The most probable parameter?

- Principles and Parameters (Chomsky 1981)
  - S $\xrightarrow{p}$ [+pro-drop], $p = 0$ for English and $p = 1$ for Italian





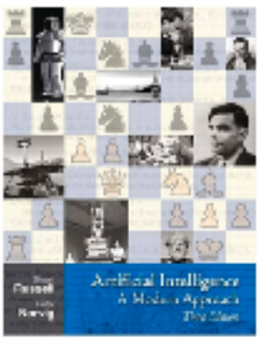


A BEHAVIORAL MODEL OF RATIONAL CHOICE

By HERBERT A. SIMON*

Traditional economic theory postulates an "economic man," who, in the course of being "economic" is also "rational."

Broadly stated, the task is to replace the global rationality of economic man with a kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist.



27.3 ARE WE GOING IN THE RIGHT DIRECTION?

In Chapter 1, we said that our goal was to build agents that *act rationally*. However, we also said that

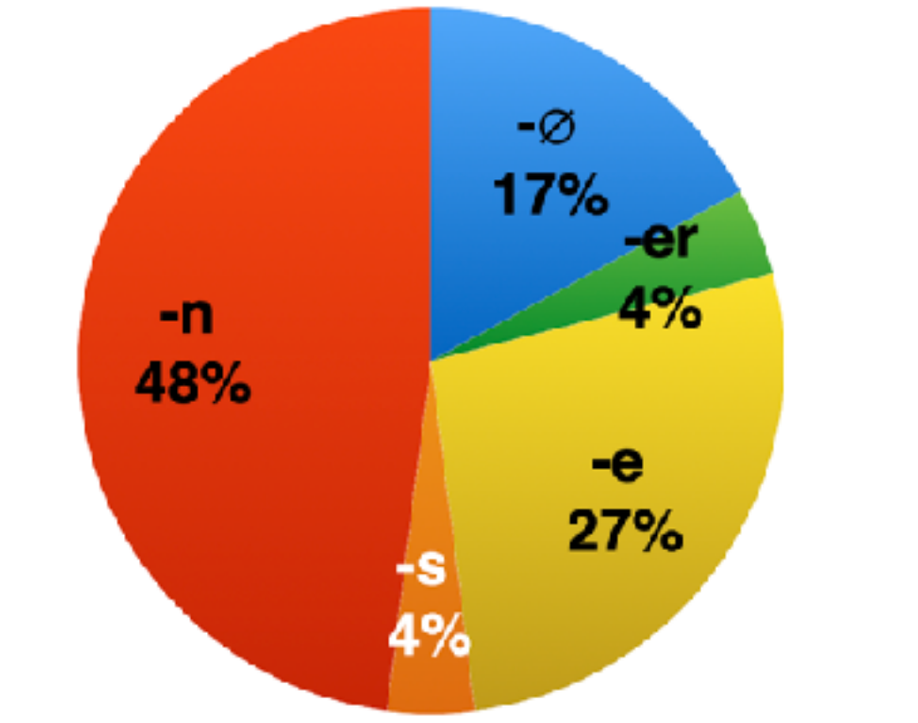
> . . . achieving perfect rationality—always doing the right thing—is not feasible in complicated environments. The computational demands are just too high. For most of the book, however, we will adopt the working hypothesis that perfect rationality is a good starting point for analysis.

> The capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problems whose solution is required for objectively rational behavior in the real world—or even for a reasonable approximation to such objective rationality.

He suggested that bounded rationality works primarily by **satisficing**—that is, deliberating only long enough to come up with an answer that is "good enough." Simon won the Nobel Prize in economics for this work and has written about it in depth (Simon, 1982). It appears to be a useful model of human behaviors in many cases. It is not a formal specification for intelligent agents, however, because the definition of "good enough" is not given by the theory. Furthermore, satisficing seems to be just one of a large range of methods used to cope with bounded resources.

Language learning must be satisficing ...

- Because “all grammars leak” (Sapir 1928).
- English past tense: some 150 irregular verbs that do not add “-ed” but children typically learn the rule “add -ed” before age 3 and "-ed" is used on new verbs ("google", "blick").



- End in “e”: add -n.
- End in “el/er/en”: do nothing.
- Is masculine: add -e.
- Is feminine: add -en.
- …
- …
- Altogether about **15%** of nouns require rote memorization.

When all rules fail

- Because sometimes there are no rules or generalization!

I stride down the street.

You strode down the street.

They have ??? down the street.

(6) Third-conjugation mid vowel stem change patterns:

sumergir 'to submerge' (no change):

pres. indic.	sumerjo	sumerges	sumerge	sumergimos	sumergís	sumergen
pres. subj.	sumerja	sumerjas	sumerja	sumerjamos	sumerjáis	sumerjan

discernir 'to distinguish' (diphthongizing):

pres. indic.	discierno	disciernes	discierne	discernimos	discernís	disciernen
pres. subj.	discierna	disciernas	discierna	discernamos	discernáis	disciernan

desvestir 'to undress' (lowering):

pres. indic.	desvisto	desvistes	desviste	desvestimos	desvestís	desvisten
pres. subj.	desvista	desvistas	desvista	desvistamos	desvistáis	desvistan

agredir 'to attack' (defective):

pres. indic.	*	*	*	agredimos	agredís	*
pres. subj.	*	*	*	*	*	*

a. je fris, tu fris, il frit, nous ???, vous ???, ils ???
I fry, you.sg fry, he fries, we ???, you.pl ???, they ???

b. je clos, tu clos, il clôt, nous ???, vous ???, ils closent
I close, you.sg close, he closes, we ???, you.pl ???, they close

do+not = don't

are + not = aren't

is + not = isn't

does + not = doesn't

...

am + not ≠ amn't

may + not ≠ mayn't

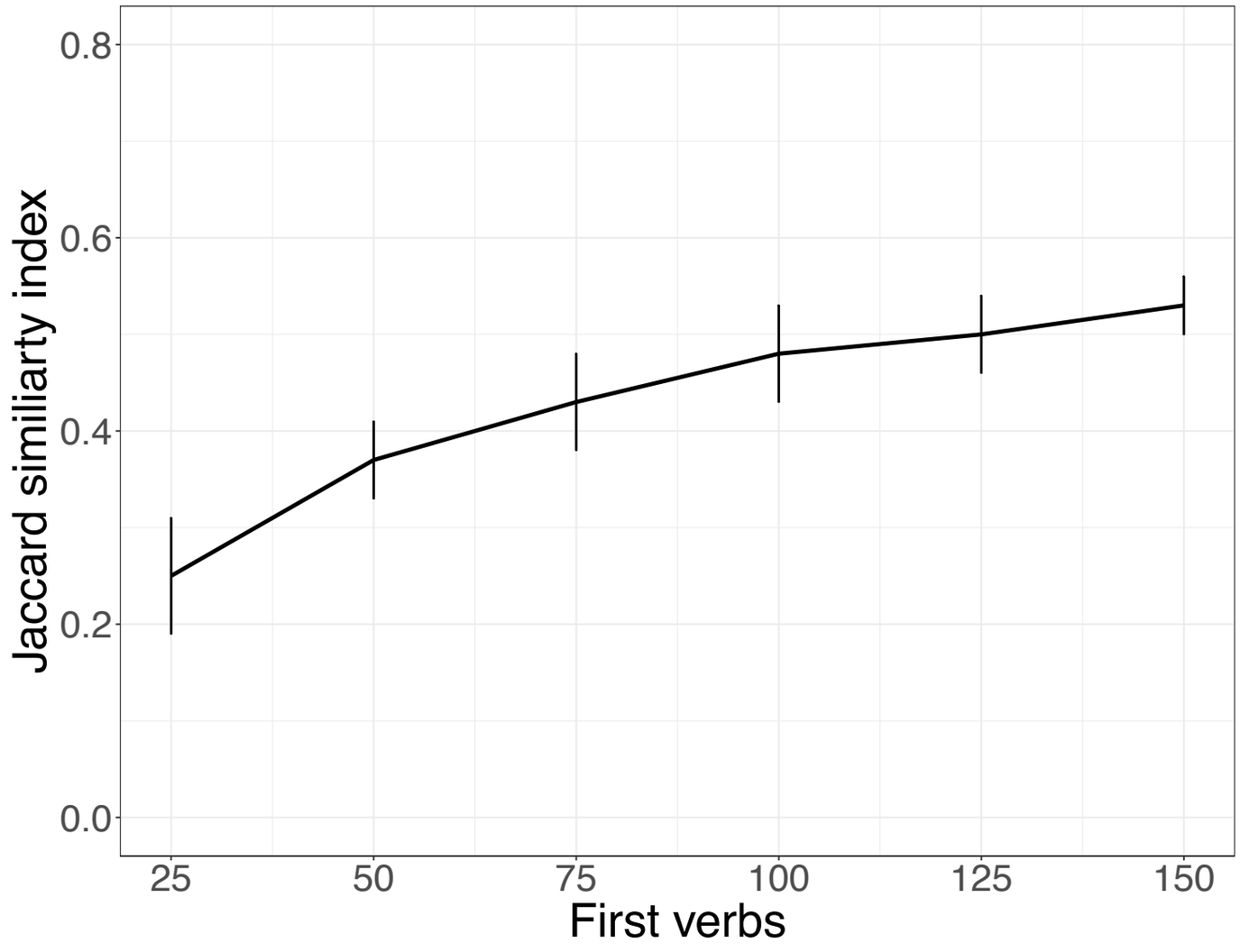
The Bronx
The Hague
*The Berlin
*The Chicago
*The Montreal

***lažu** 'I climb'
***pobežu** (or ***pobeždu**) 'I conquer'
***deržu** 'I talk rudely'
***muču** 'I stir up'
***erunžu** 'I behave foolishly'

What did John see that Bill ate *t* ?
*What did John complain that Bill ate *t* ?
*What did John quip that Bill ate *t* ?

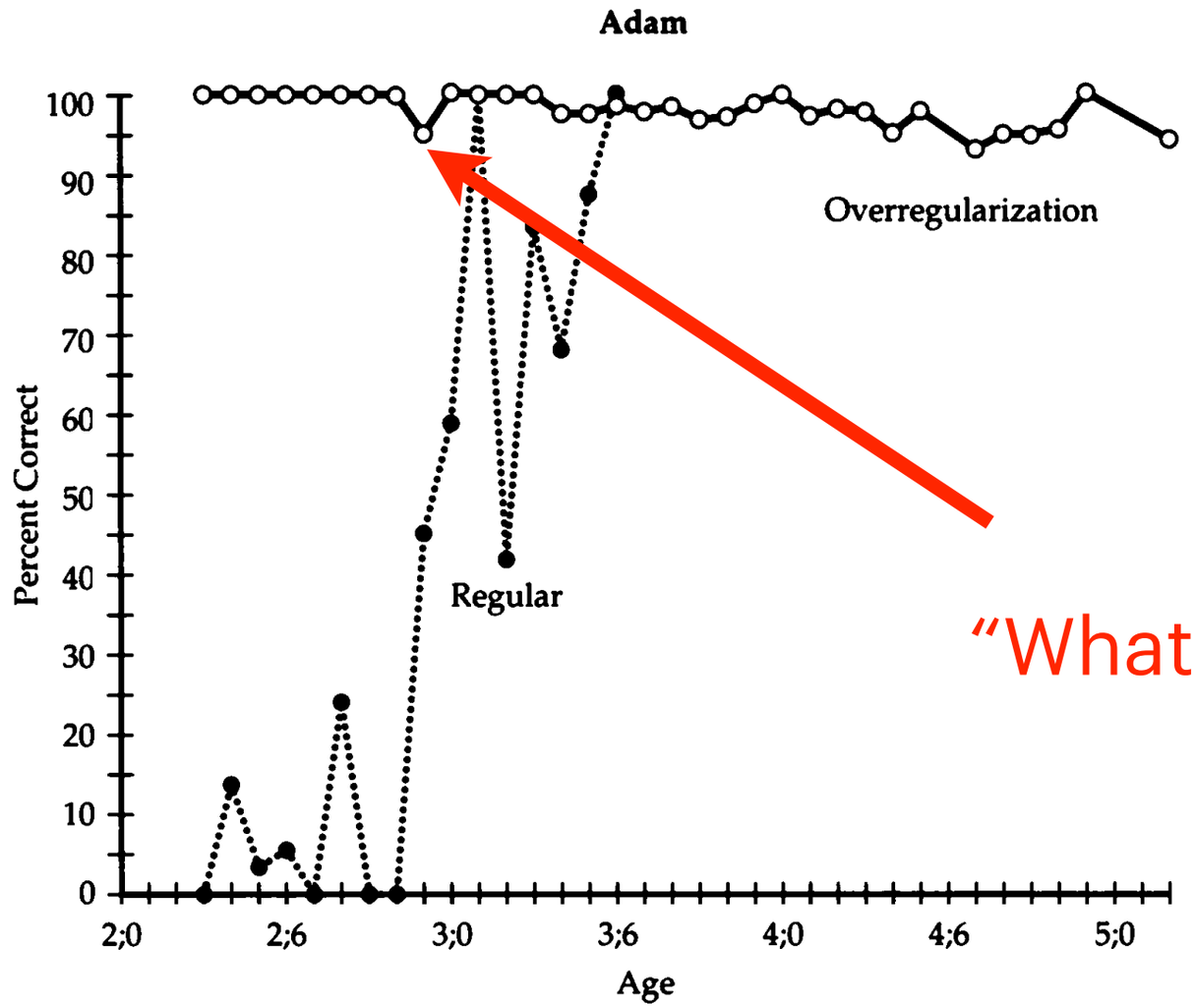
Why it must be a threshold

Jaccard similarity $\frac{|A \cap B|}{|A \cup B|}$



- "The end result is a high degree of uniformity in both the categorical and variable aspects of language production, where individual variation is reduced below the level of linguistic significance" (Labov 2012; Labov 1973).

Why it must be a threshold



"What that feeled like"?

Non-monotonic learning is extremely common in child language.

"I died him."

"Don't giggle me."

"I said you something."

"They delivered you a lot of pizzas."

In search for a model

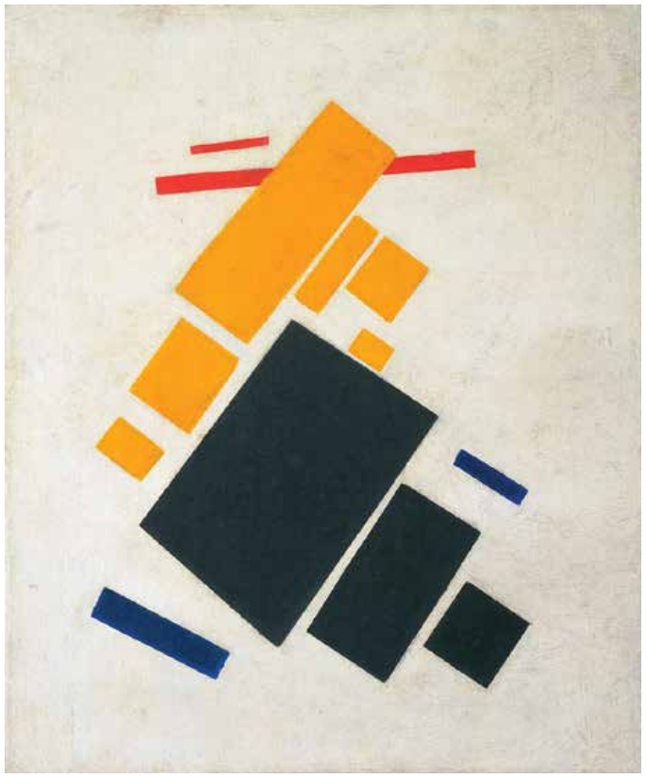
- Most/all models in psychology and cognitive science but with hyper-parameters: not learning models from the data but statistical models of the (experimental) data.

- Most/all models in ML/NLP are optimizing, in addition to many hyper-parameters:

  - Many highly frequent rules in language are not learned early, and many infrequent/absent forms in language are used robustly by children.

An Intuition: Enough is enough

- Give a set of items:
  - If *many* do X, then all do X
  - if *few* do X, then remember the few that do and stay put!
- How many is *many* or *few*?
- Learning rules with exceptions is a classic problem in cognitive science but there were no principled solutions.

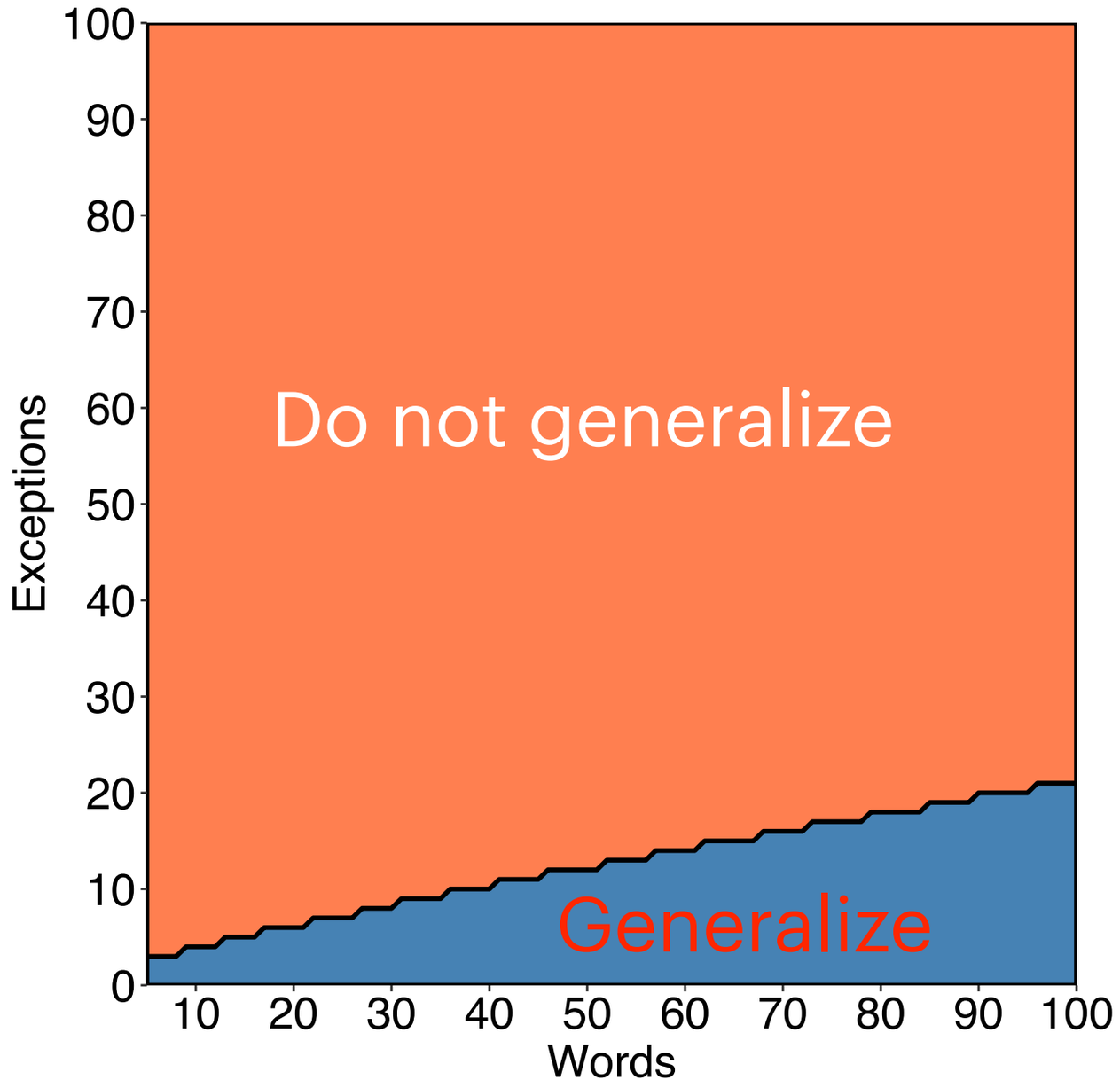


Tolerance Principle

A productive rule applicable to **N** items in the learning data cannot have more than **e** exceptions (Proof by Sam Gutmann)

N	θ_N	%
10	4	40%
20	6	30%
50	12	24%
100	21	21%
200	37	19%
500	80	16%
1000	144	14%

$$\theta_N = \frac{N}{\ln N}$$



Parameter free: no tuning and runs out-of-the-box

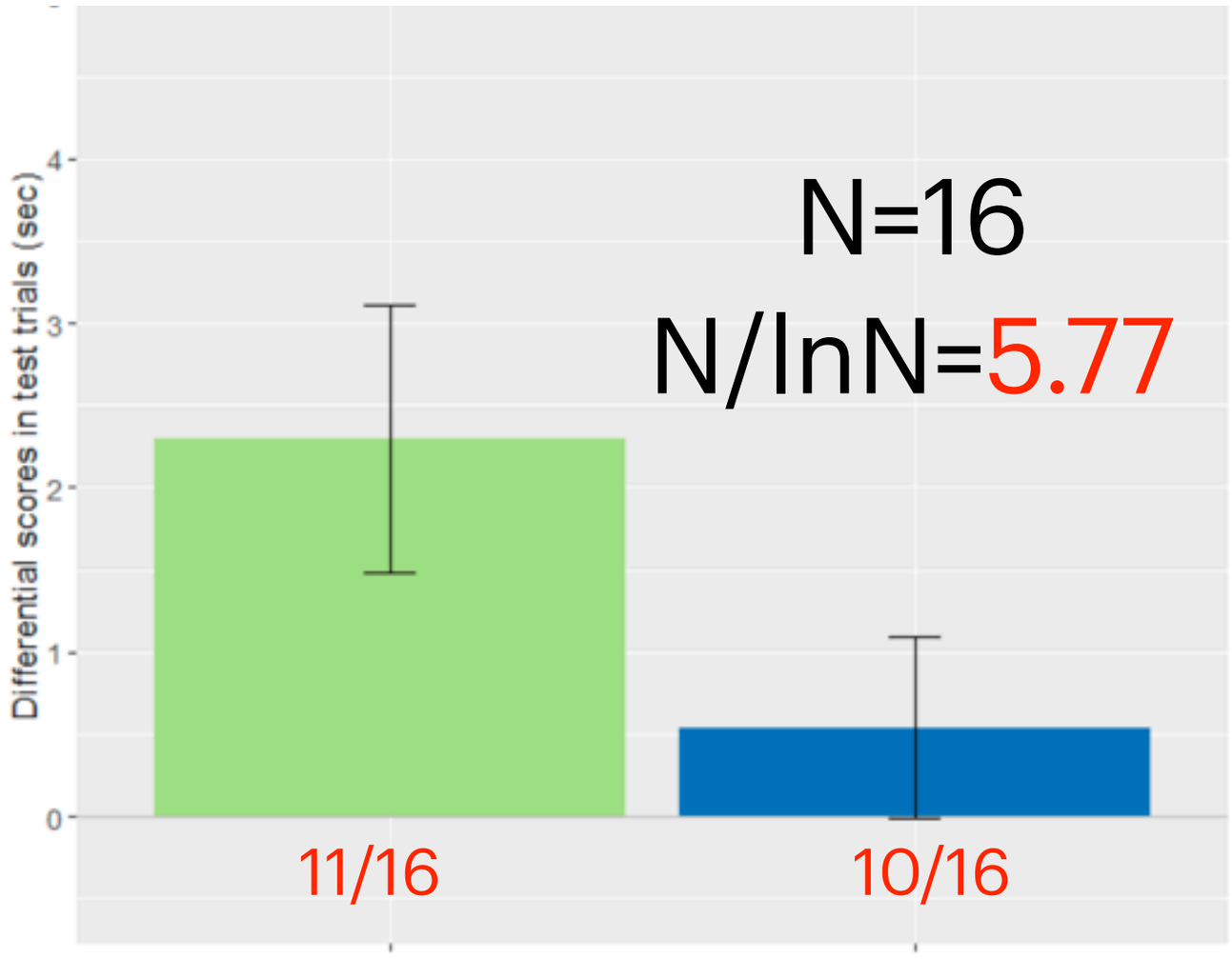
“Artificial” language (Shi & Emond 2023)

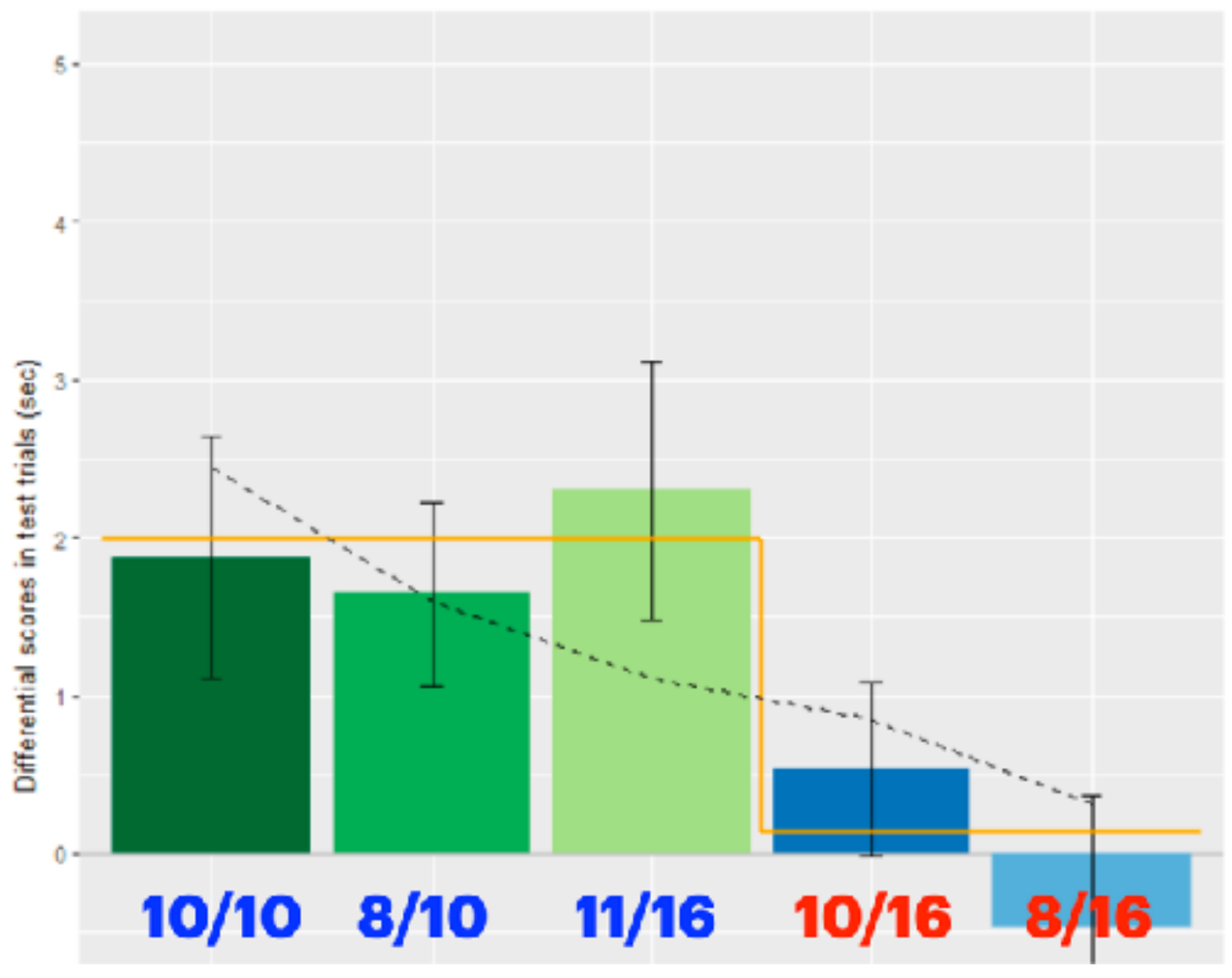
14-month-old non-Russian learning infants

“Movement” R1: ABC→BAC; R2: ABC→ACB

Machty gnutsja lukom → Gnutsja machty lukom

Machty gnutsja lukom → Machty lukom gnutsja





How it's done?

- No one has the faintest idea how this, or the $N/\ln N$ threshold, is implemented in the brain, but one should not underestimate animal's apparently numerical abilities (e.g., ant pedometer, Wittlinger et al. 2006, Gallistel and Gibbon 2000).

- Ratio tracking in infancy: Transitional Probability (Saffran, Aslin & Newport 1996) as statistical learning

  - TP(A→B) = P(AB)/P(A): how well A predicts B

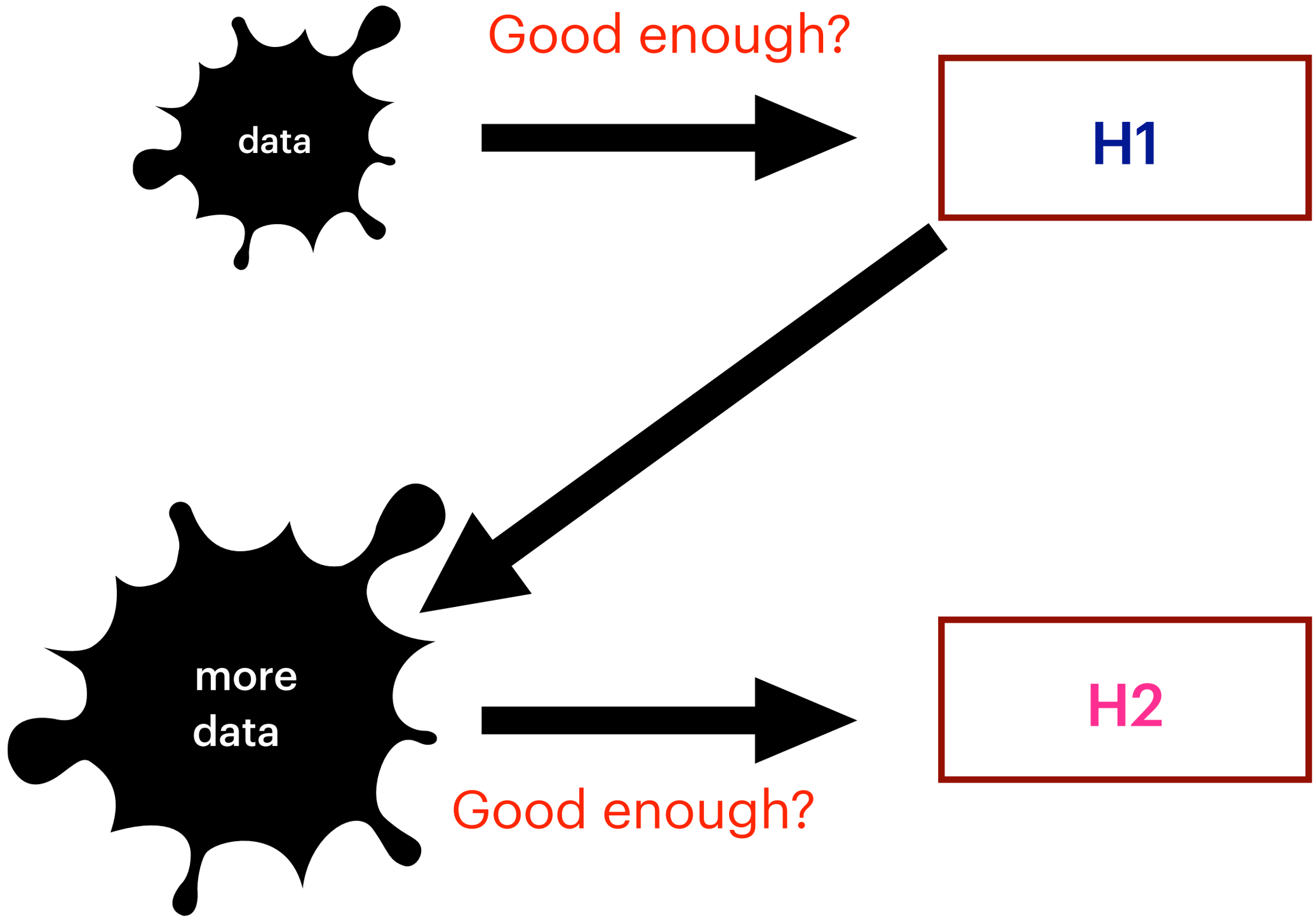
  - "Their" TP: (token frequency of a type)/(sum of token frequencies over multiple types)

  - My TP: type/type; if all types have the same token frequency, then babies may well be tracking a ratio of types.

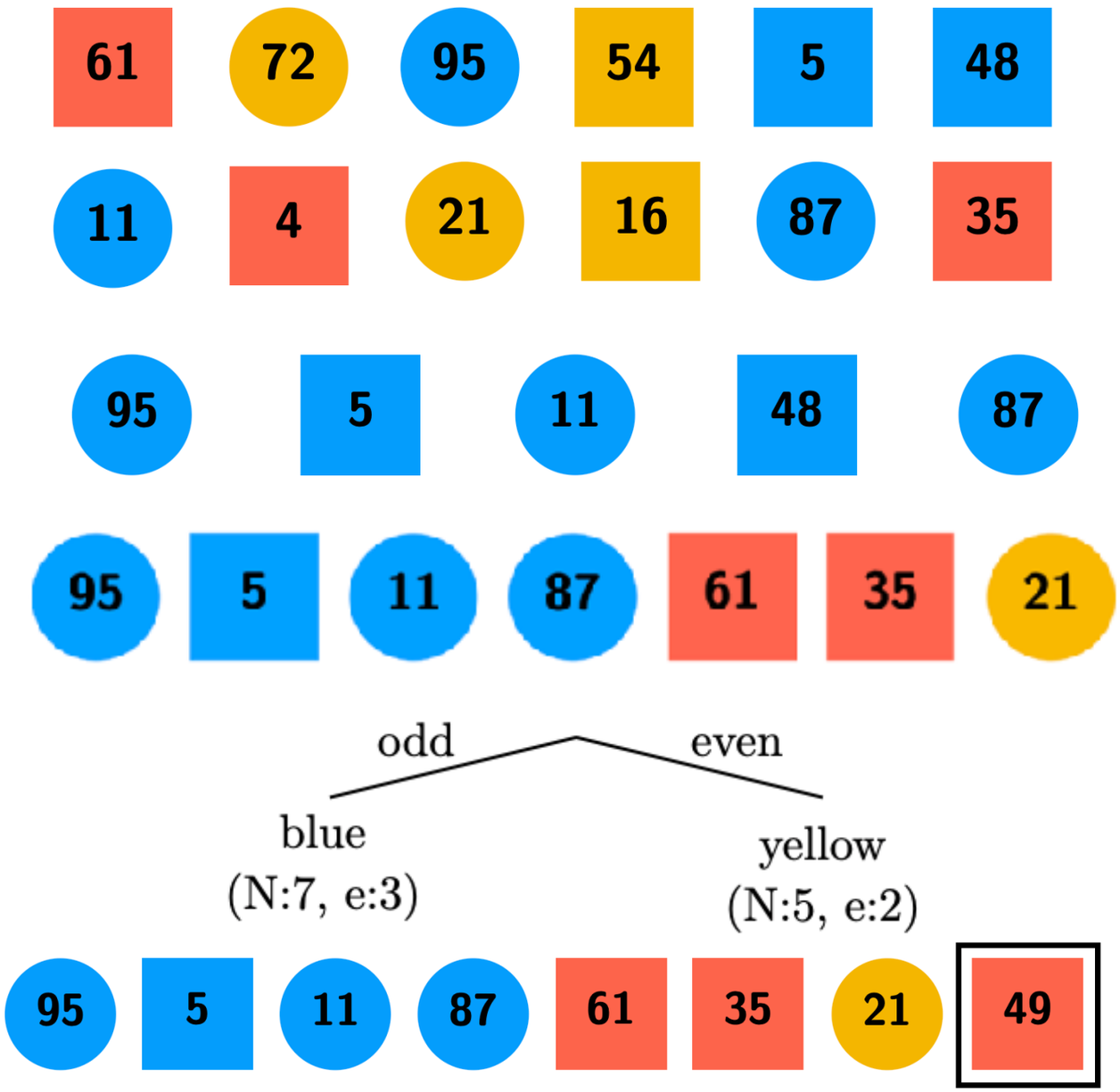
Grammar as Satisficing

- The grammar is always provisional: keep it as long as it's adequate
  - An adequate theory of hypothesis testing: no need to over-intellectualize.
- The status of rules may change as more items are learned
  - If you know 10 item, of which 7 follow a pattern, then you generalize ($\theta_{10} = 4$)
  - If you learn another 5 items, but only 2 more follows the pattern, then you don't generalize (but only memorize the 9 that do follow the rule): $\theta_{15} = 5$
  - If you learn another 5 items, all of which follow the pattern, then you generalize again: $\theta_{20} = 6$

Grammar by Abduction



Abductive categorization



Applications

- The substance of hypotheses, and the coverage criterion for the hypotheses, are entirely separated.
  - The substance of grammar, and the coverage criterion for the grammar, are entirely separated: substance-free UG.
  - You see red, not infrared. (Lila Gleitman)
    - One substantive feature at a time (Medin et al. 1987).
- Three case studies
  - The German plural suffixes.
  - Recursive structures in English and German morphosyntax.
  - Grammar vs. learning: Regularities and accidents in English morphosyntax.

German plurals

German noun plural suffixes

Suffix	Type	%
-n	169	38.4%
-null	121	27.5%
-e	80	18.2%
-en	47	10.7%
-er	15	3.4%
-s	8	1.8%
Total	440	100%

Autos, Parks, Pizzas, …, iPhones

Charles learning German plural suffix

Suffix	Type	%
-n	169	38.4%
-null	121	27.5%
-e	80	18.2%
-en	47	10.7%
-er	15	3.4%
-s	8	1.8%
Total	440	100%

Leo Corpus (CHILDES)

Sache	Sachen	F	250		
Schiene	Schienen		F	121	
Ente	Enten	F	89		
Gleis	Gleise	N	89		
Muschel	Muscheln		F	79	
Socke	Socken	F	67		
Frosch	Frosche	M	55		
Bein	Beine	N	50		
Kastanie		Kastanien		F	42
Schranke		Schranken		F	31
Bild	Bilder	N	25		
Kerze	Kerzen	F	25		
Ei	Eier	N	24		
Mensch	Menschen		M	23	
Robbe	Robben	F	22		
Affe	Affen	M	22		
Junge	Jungen	M	21		
Flugel	Flugel	M	21		
Scheibe	Scheiben		F	20	
Lowe	Lowen	M	17		
Platzchen		Platzchen		N	16
Tunnel	Tunnels	M	16		
Zwiebel	Zwiebeln		F	16	
Giraffe	Giraffen		F	16	

```
169  n
121
 80  e
 47  en
 15  er
  8  s
```

- Gender:
  - F: 152, M: 17
- Phonology:
  - e#: 151, l#: 12, r#: 6
- Try gender (better): F→n
  - F: 200, -n: 152: fails (needs 163)
- Try phonology:
  - e#: 151, -n: 151: Succeeds
- R1: e#→n. Remove 151 e# words, no exceptions

```
121
 80   e
 47   en
 18   n
 15   er
  8   s
```

- Gender:
  - M: 81, N: 40
- Phonology:
  - [l, r, n]#: 121
- Try phonology one segment: [l, r, n]#→NULL
  - [l, r, n]#: 175, NULL: 121: fails (needs 142)
- Try phonology two segments:
  - [el, er, en]#: 145, NULL: 121: Succeeds
- R2: [el, er, en]#→NULL Remove 145 words, memorize 24 exceptions

R1 and R2 are what Wiese (1996) calls “reduced syllable constraint” but they are easily discoverable

```
76  e
46  en
15  er
 7  s
```

```
12 g
11 s
 8 t
 8 n
 8 l
 7 f
 6 h
 4 m
 3 z
 2 k
 2 d
 2 b
 1 v
 1 r
 1 i
```

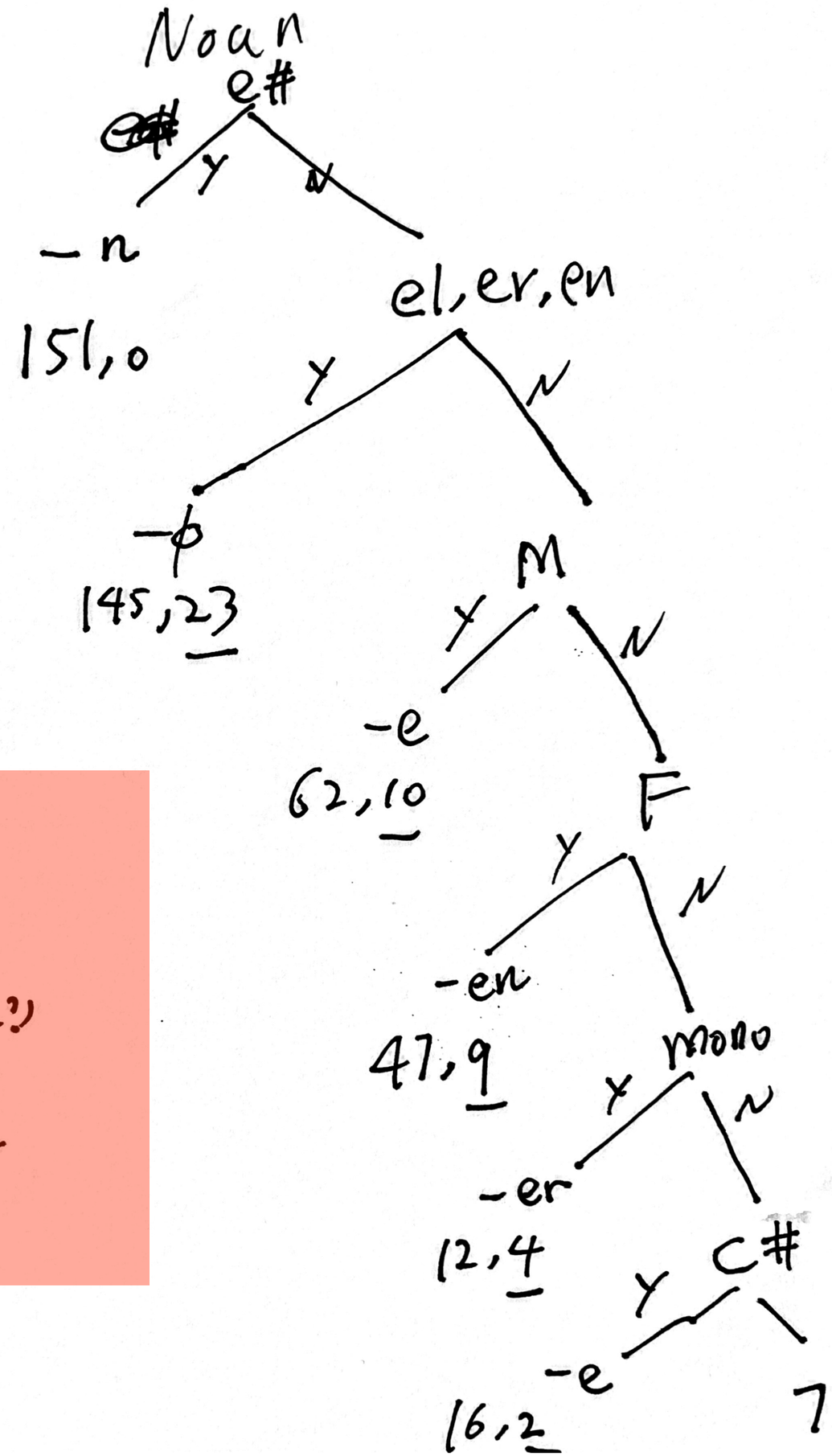
- Gender:
  - M: 52, N: 17, F: 7
- Phonology:
  - consonant → e
- Try phonology first: C#→e
  - C#: 144, e:75: fails (needs 116)
- Try gender: M→e
  - M: 62, e: 52: Succeeds (needs 47)
- R3: M→e, Remove 62 words, memorize 10 exceptions

```
39  en
24  e
12  er
 7  s
```

- Gender:
  - F: 38, N: 1
- Phonology:
  - C# → en
  - C#: 58, en: 38: Fails (needs 44)

```
14 g
13 t
 5 r
 3 n
 2 i
 1 m
 1 d
```

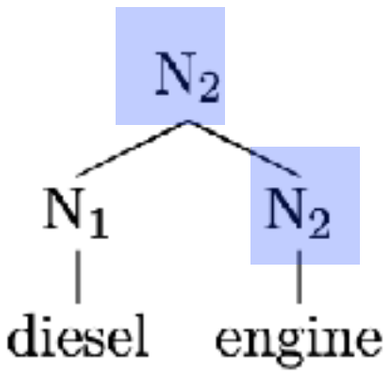
- Try gender: F→en
  - F: 47, en: 38: Succeeds (needs 35)
- R4: F→en, Remove 47 words, memorize 9 exceptions

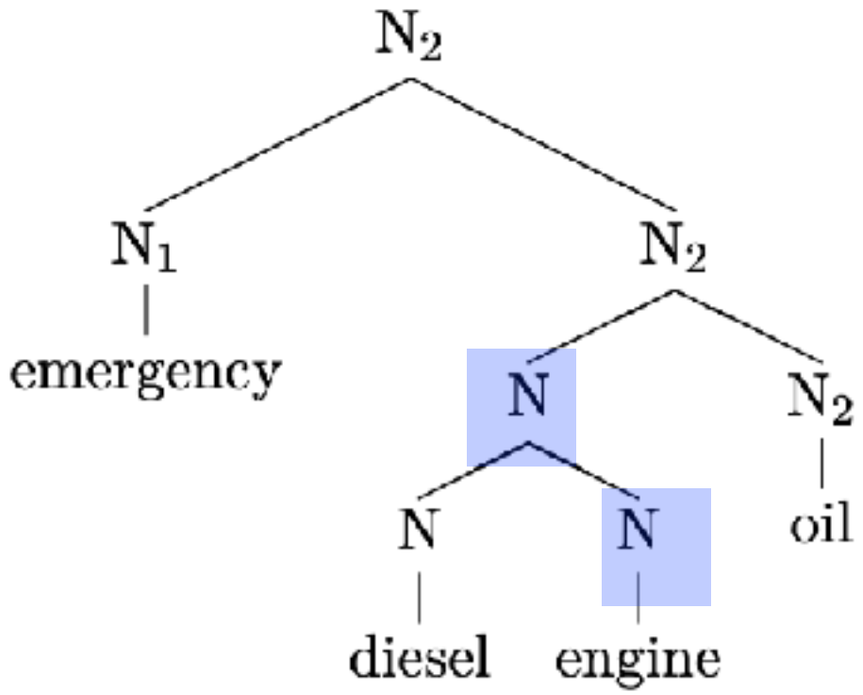

Noun
e#
y
n
-n
151,0
el, er, en
y
n
-φ
145,23
M
y
n
-e
62,10
F
y
n
-en
47,9
mono
y
n
-er
12,4
C#
y
-e
16,2
7

23+10
+9+4
+2+7(?)

= 55

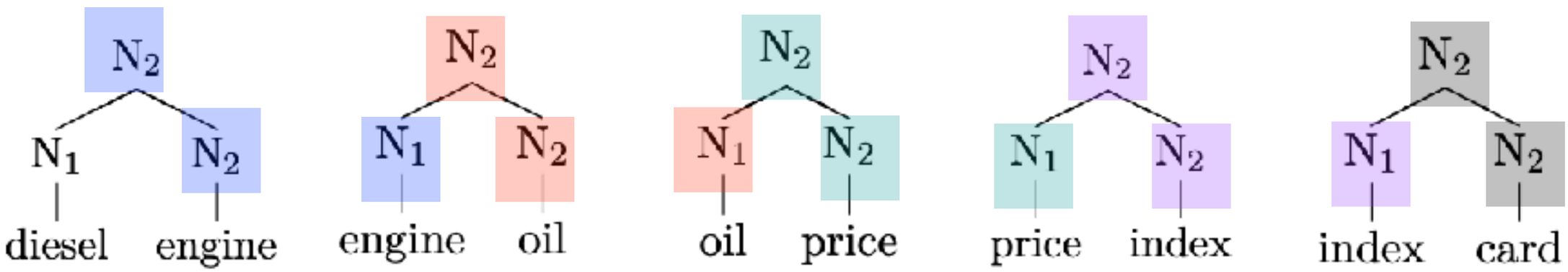
“Learning” Recursion





- Children learn NN compounding very early (Clark 1982)
  - bear hat, crash cars, doctor house, house smoke, tickle sock, snow beard, fashion car, wood friend, …
  - blood face man, change money box, dragon water bug, meat baby food, state champ man, frog party favor, …
- Just hearing n-level embedding doesn’t mean (n+1)-level embedding is also possible.

Recursion as Selection



- Recursion: [diesel engine] = [engine], [engine oil] = [oil]
  - If “engine” were the only noun in English that can appear in both positions (N_1 and N_2), then English would only allow “engine engine engine ….”
- “Learning”recursion is to learn the productive conditions under which self-embedding
- Lexical semantics probably doesn’t play a significant role.

English Noun-Noun Compounds

- CHILDES input data and extracted N_1 N_2 in adjacent positions
- Tested the 100 most frequent nouns in child English (Chicago Corpus)
  - 94 appear in the N_1 position and 95 in the N_2 position
- There is no restriction to recursive NN compounding.

Most frequent	In N_1	In N_2	Need	Productive?
50	49	49	38	Yes
100	99	99	79	Yes
200	199	196	163	Yes
300	296	289	248	Yes
500	476	455	420	Yes

Truck driver vs. Sweetheart

- Child English: picture taker, diaper changer, kid driver, house seller, raisin keeper, television tape recorder!

- Adjective-noun compounds: grand-kin, blackmail, blackboard, bluejay, redhead, redneck, redwood, whitehead, greenhouse, longhorn, longshot, shortcake, shortcut, shortstop, hardwood, softball, freestyle, mainstream, dumbbell, …

Top -er	Compound	Need	Productive?
50	49	38	Yes
100	99	79	Yes
150	149	121	Yes
200	199	163	Yes

Top adj	Compound	Need	Productive?
50	6	38	No
100	16	79	No
150	22	121	No
200	27	163	No

“Possessive” structure and recursion

- The man’s neighbor’s book
- ?*The book of the neighbor
- *The book of the neighbor of the man

. das Buch *von* dem Mann
the book of the$_{Dat}$ man
‘the man’ s book’

das Buch *von* dem Nachbarn von dem Mann
the book of the$_{dat}$ neighbor of the$_{dat}$ man
‘the man’ s neighbor’ s book’

na ren *de* linju *de* shu
that man GEN neighbor GEN book
‘that man’s neighbor’s book’

Peters/Vaters Buch
Peter’ s/father´s book

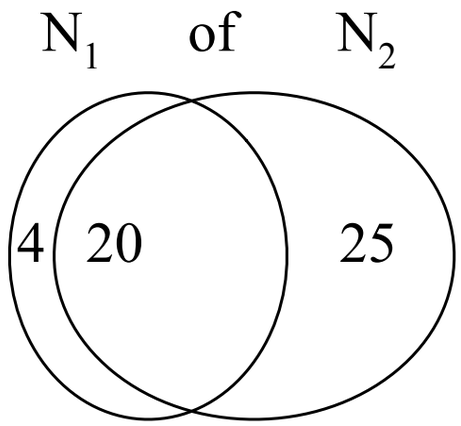
*na ren linju shu
that man neighbor book
‘that man’s neighbor’s book’

*Manns Buch
man’s book

*Vaters/ Peters Nachbars Buch
father’ s/Peter’s neighbor’s book’

2021 paper with Daoxin Li (Penn→Northwestern) and Lydia Grohe and Petra Schulz (Frankfurt)

Most frequent nouns



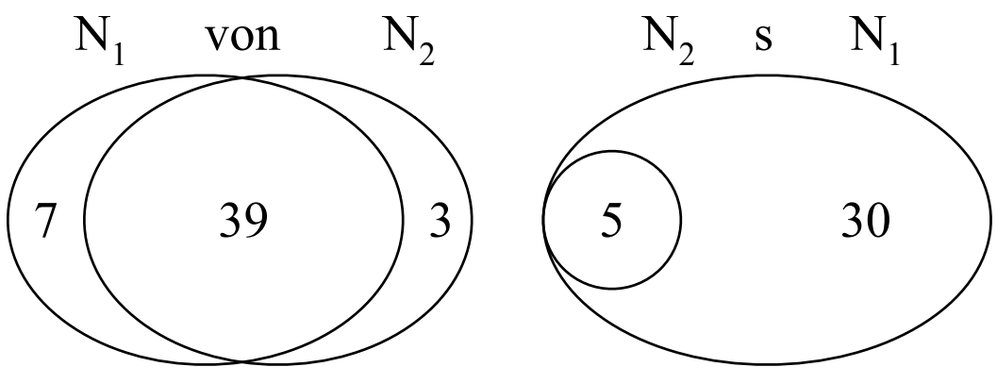
$$\frac{20}{24}$$

$N_1 \Longrightarrow N_2$

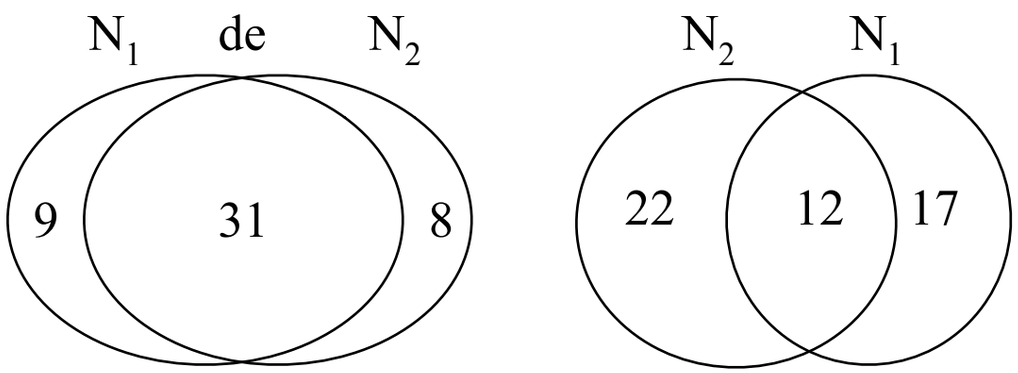
of-possessive (N_1 of N_2)

N_1 (24)	*piece*, top, *bit*, picture, name, *cup*, time, color, day head, door, *box*, way, hair, thing, mouth, book, school, room, man
N_2 (45)	cheese, cake, head, book, train, house, water, milk, box, baby, hair, car, juice, food, school, fish, hat, day, dog, man

Measure and inalienable possession



(a) The German possessives



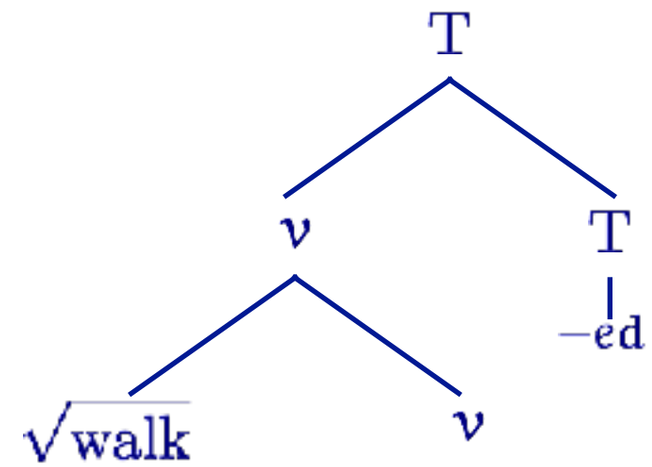
(b) The Mandarin possessives

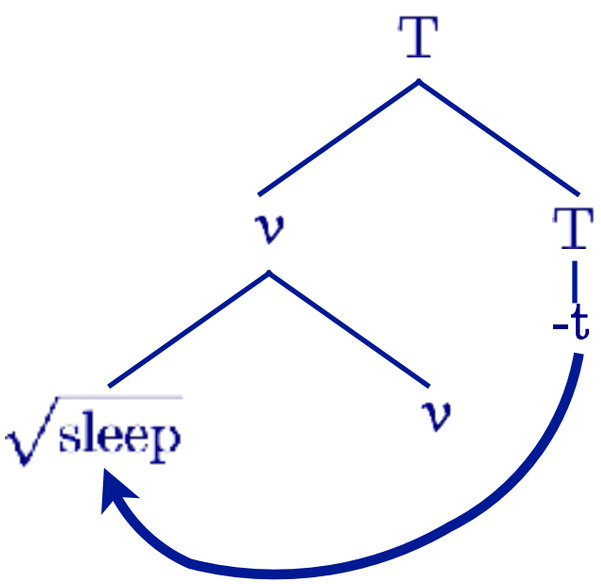
Productivity and Recursion

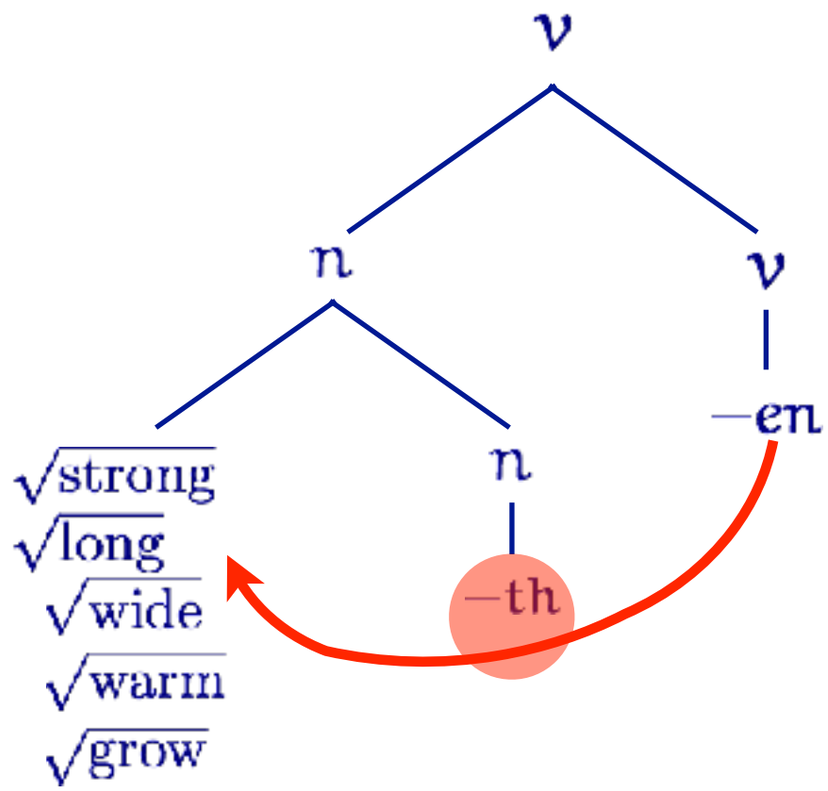
The cover of the book
The color of the cover of the book
The son of the president of the union
The name of the son of the president of the union
The top of the third inning of the fifth game of the first Word Series of the twenty aughts

UG vs. Learning

- Structural locality and productivity (Marantz 1997 followed by Bobaljik, Embick, Harley, etc.): “far” ≈ productive, and “near” ≈ lexicalized selection







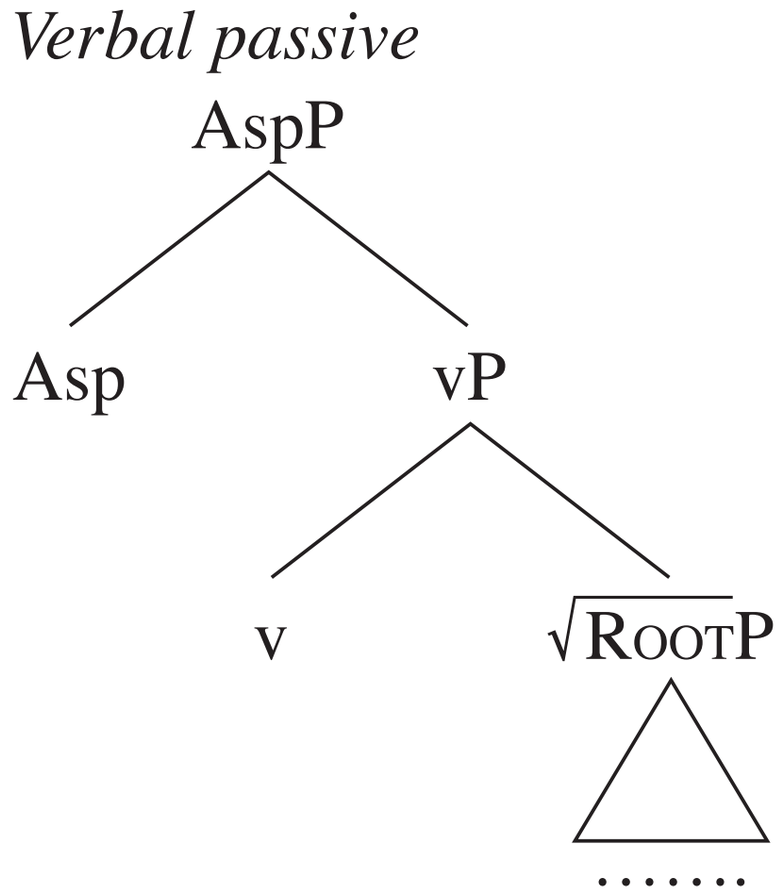
Lengthen, *warmthen, *pinken

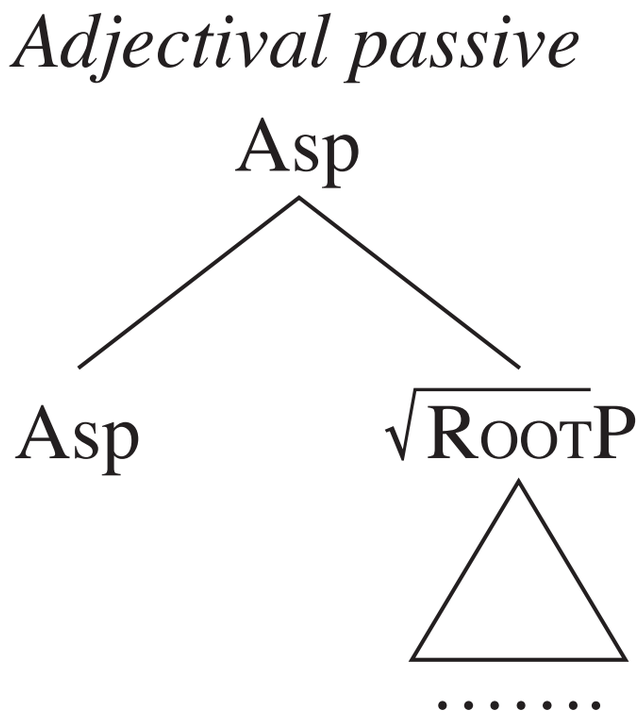
- About **50** -*en*-taking verbs in English, **45** are monosyllabic obstruent adjectives (Jackendoff & Audring 2017) and **6** are not: lengthen, strengthen, christen, hearten, threaten, frighten
  - Hypothesis formation: [monosyllabic, +obstruent, +adjective] ⟹ -en
- Hypothesis testing: Is the generalization productive?
  - At least **300** adjectives fit this description
  - Non-productive no matter how the words are sampled
  - Generalization failed so lexicalize: *warmthen, *pinken, *greaten, ? toughen
  - A list of accidents and local experiences

English Passives

- *The read book
- *The told story
- *The debated resolution
- *The mentioned bisque

- The unread book
- The untold story
- The carefully debated resolution
- The previously mentioned bisque





Adjectival Passives

- Theme analysis (Anderson, Wasow, Bresnan, Williams)
  - A recently offered deal vs. *A recently offered customer
- Sole Complement Generalization (Levin & Rappaport 1986): arguments that can serve as the sole NP complement to a verb lead to an adjectival passive
- I offered a deal vs. *I offered a customer
- I fed the baby vs. *I fed the cereal
- The recently fed baby vs. *The recent fed cereal

But

- They read a book ⇏ *the read book.
- They told a story ⇏ *the told story
- They mentioned an example ⇏ *the mentioned example
- Not even some dative verbs: e.g., showed (them) a movie ⇏ *a shown movie, shot (him) an email ⇏ *a shot email

- I googled the topic ⇏ *the googled topic
- I friended my neighbor ⇏ *the friended neighbor
- Taylor dropped an album ⇏ *the dropped album

Verbal vs. Adjectival Passive

- Top 100 most frequent transitive verbs
  - 95 have verbal passive counterpart (e.g., The pizza was eaten): productive
  - Only **5** have unambiguous adjectival counterpart (attributive usage in NPs):
    - *baked, chopped, fried, drunk, squashed:* Nowhere near the level for productive generalization
    - Productive subclasses possible: *fried, grilled, sauteed, boiled, baked, ...* ⇒ sousvided
- Adjectival passive is unproductive: need to hear it!

*read book vs. unread book

- If *read is not an adjective, how come **unread** is an adjective?
  - Only answer: **un-** is unproductive
- 64 **un-** prefixed adjectives in 5 million words of CHILDES
  - Only 10 are morphologically simplex: happy, usual, even, fair, true, real, pleasant, dead, stable, safe, able
    - This seems to a necessary condition: unred, unquick, unnice …
    - But is it sufficient?
- Top 50 adjectival passives, only 16 have an un- counterpart: **not close**
  - **Un-** is not productive: **unread book** is good because we hear it
  - advanced/unadvanced technology, missed/unmissed opportunity, recommend/unrecommeded dish, noted/unnoted scholar, …

Where we are

- The grammar needn’t be perfect but only good enough.
- The Tolerance Principle provides a measure of what counts as a real generalization, so a productive rule is constructed to account for it.
- Evaluation procedure: The child as the little linguist
- Discovery procedure: The linguist as the little child!