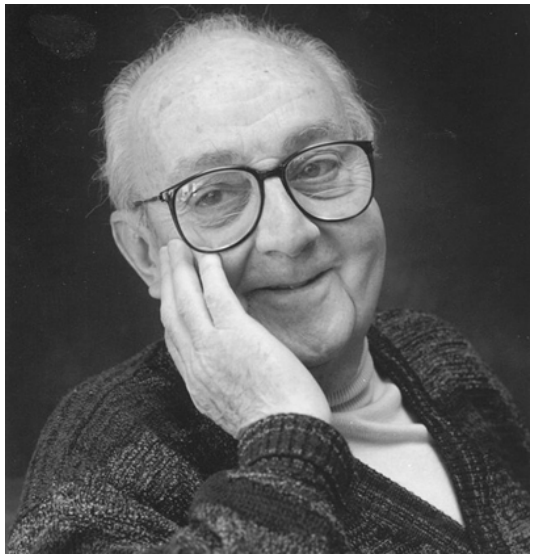
Unit 6: Bayesian Statistics

3. Language learning as Bayesian inference

5/07/2021

A quick caveat

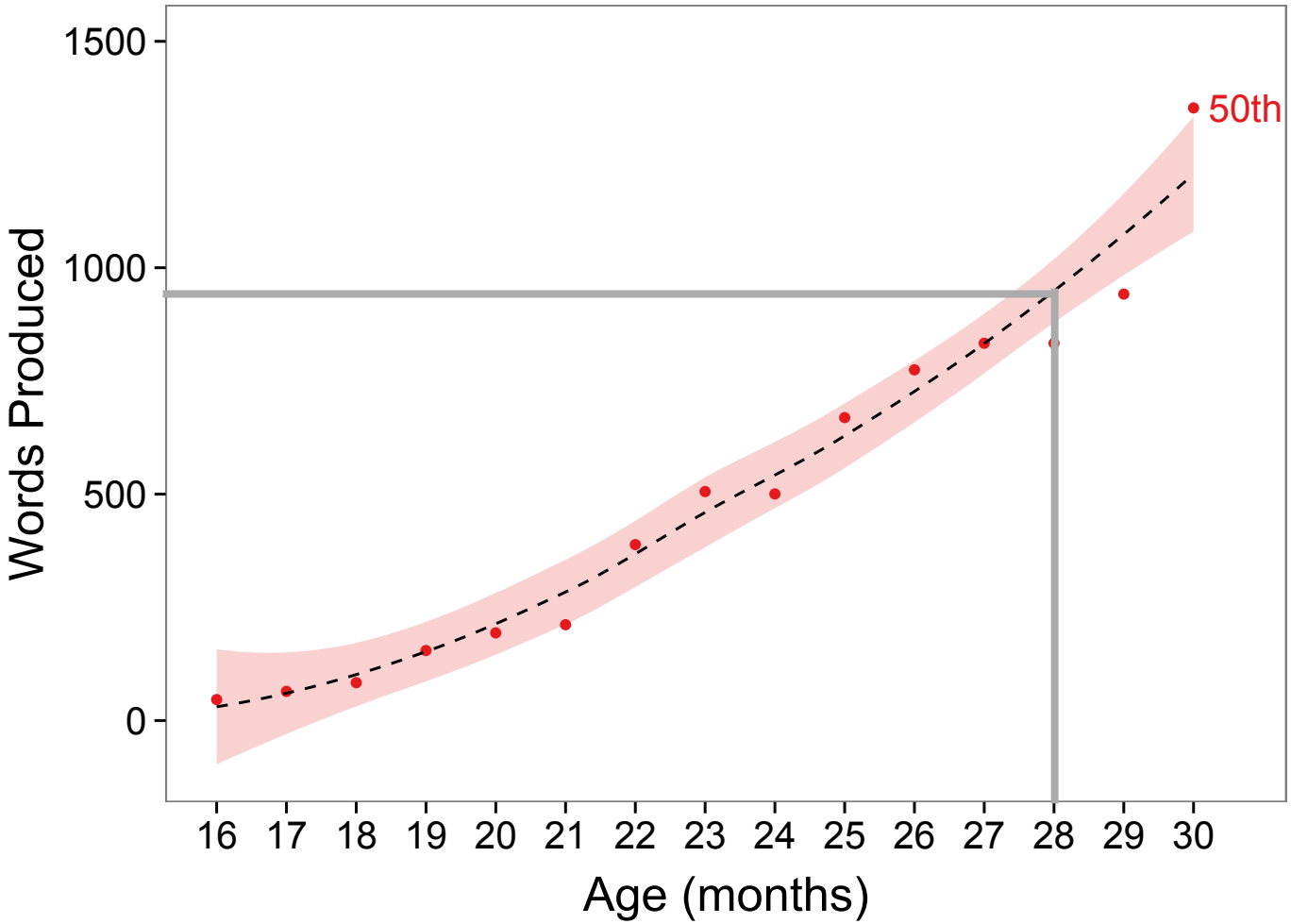
All models are wrong,
but some are useful



George Box

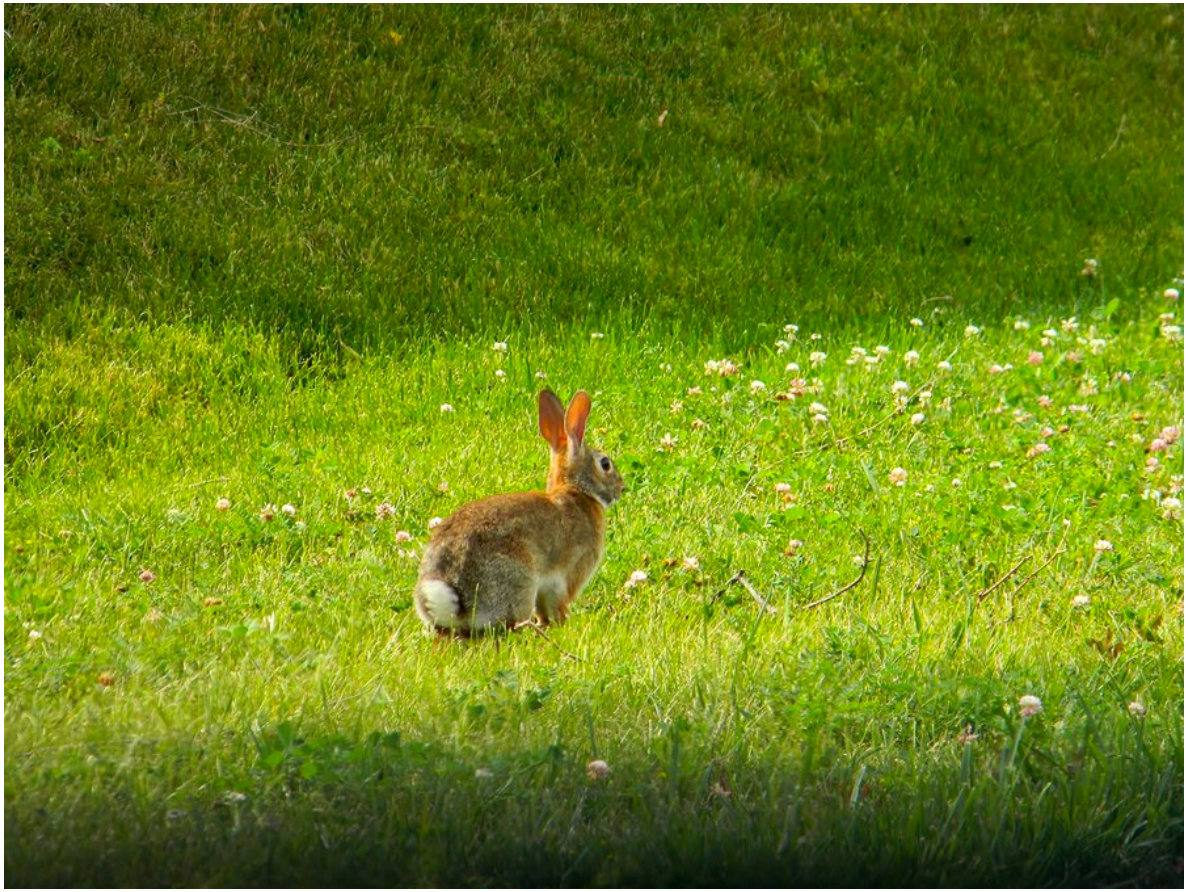
I am going to make an intentionally simplistic argument. It has flaws. But it is useful!

Language learning happens fast



Fenson et al. (2007); Mayor & Plunkett(2011)

The gavagai problem



What does “gavagai” mean?

Quine (1960)

Three big ideas

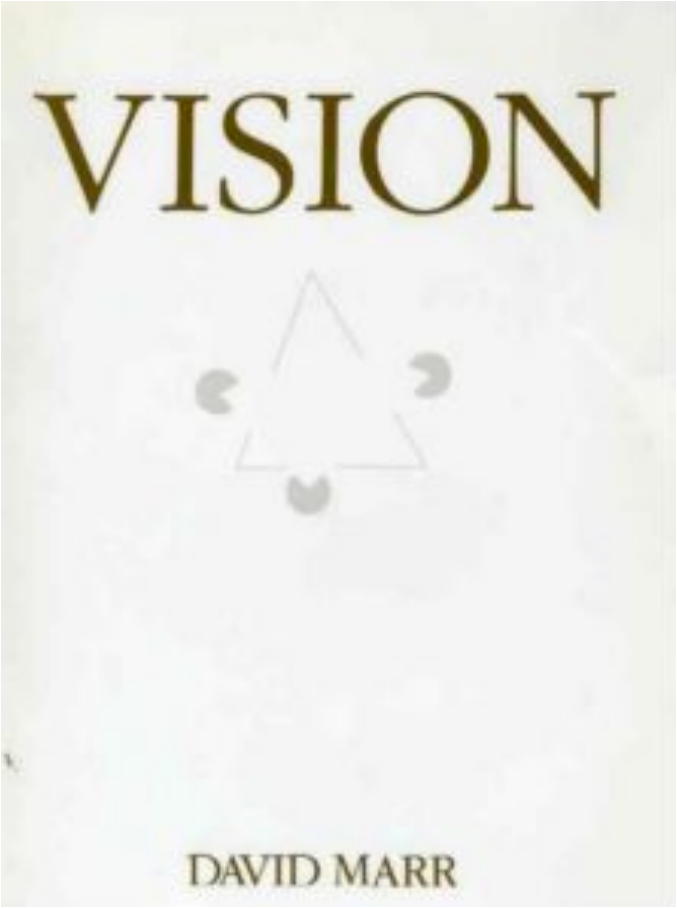
1. **Rational analysis**
   The structure of information available for solving a problem can tell you how people solve it

2. **Learning words from language *structure***
   You can learn a word's meaning by tracking the structure of its use

3. ***Language structure* reflects *intentional structure***
   You can use intentions to learn language, or language to learn intentions

Marr's levels of analysis

David Marr



Marr's levels of analysis

Computational Theory
What is the goal of the computation? What is the logic of the strategy by which it can be carried out?

Representation and algorithm
What is the representation for the input and output, and what is the algorithm for the transformation?

Hardware implementation
How can the representation and algorithm be realized physically?

Marr (1982)

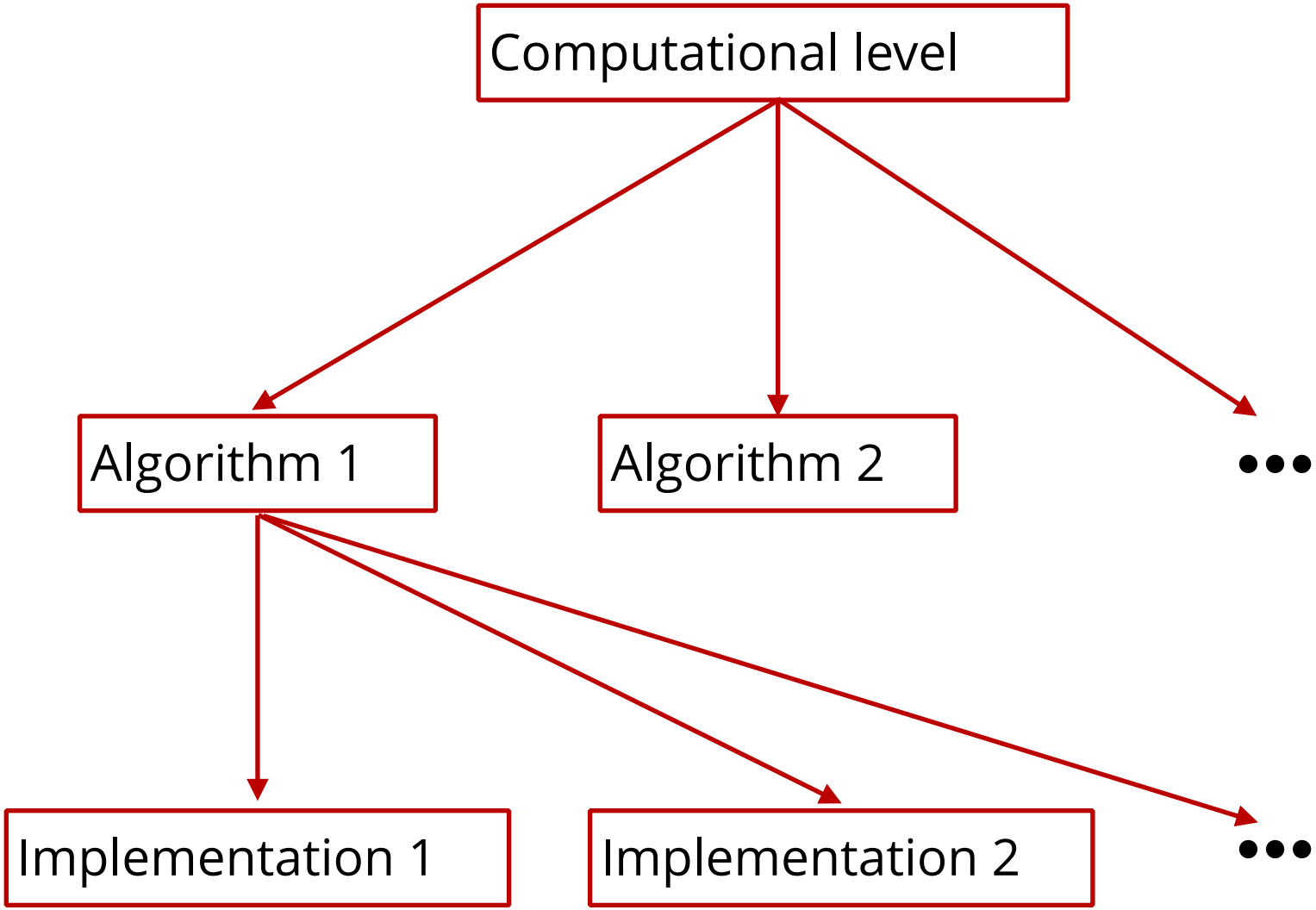
An example: The cash register

Computational level
Calculates sum of numbers using the theory of addition
So, it will be e.g commutative (order doesn't matter)

Algorithmic level
Uses fixed-point approximations and Arabic numerals
So, if numbers are too big, it will fail (unlike theory of addition)

Implementation level
Uses physical buttons and gears
So, things can fail if these break.

Each level is multiply realizable in the next lower



Constraints across levels

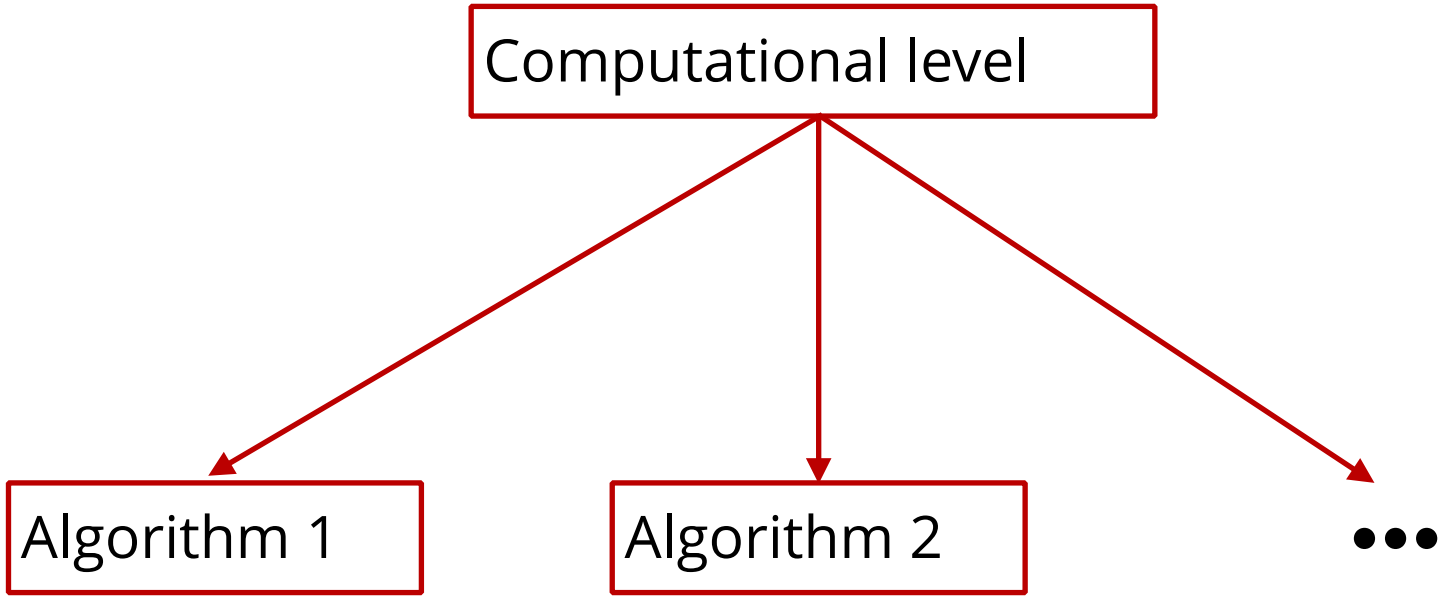
Every level constrains every other level.

But these constraints are *asymmetric*.
Higher levels constrain lower levels more.

So, you get more bang for your buck by working on higher levels.

Functionalism: Let's not worry about implementation—forget about the brain.

Cognitive psychology (approximately)



Which algorithm do people use?

Let’s go one step further…

Every level constrains every other level.

But these constraints are *asymmetric*.
Higher levels constrain lower levels more.

So, you get more bang for your buck by working on higher levels.

Rational analysis:
Let’s not worry about mechanism either.

Rational analysis

For a given computational problem, there is an *optimal solution*. Whatever it is, we have evolved to approximate it.

Figure out the optimal solution, and you'll know a lot about what people do.

"The predictions flow from the statistical structure of the environment and **not** the assumed structure of the mind."
(Anderson, 1991)

Bayes' rule: The rule for rational analysis

Posterior Probability
How likely is my *hypothesis* given the *data* I have observed

Likelihood
How likely am I to observe this *data* if my hypothesis is true

Prior
How likely is my *hypothesis* before I see any *data*

$$P(H|D) \propto P(D|H)P(H)$$

Three big ideas

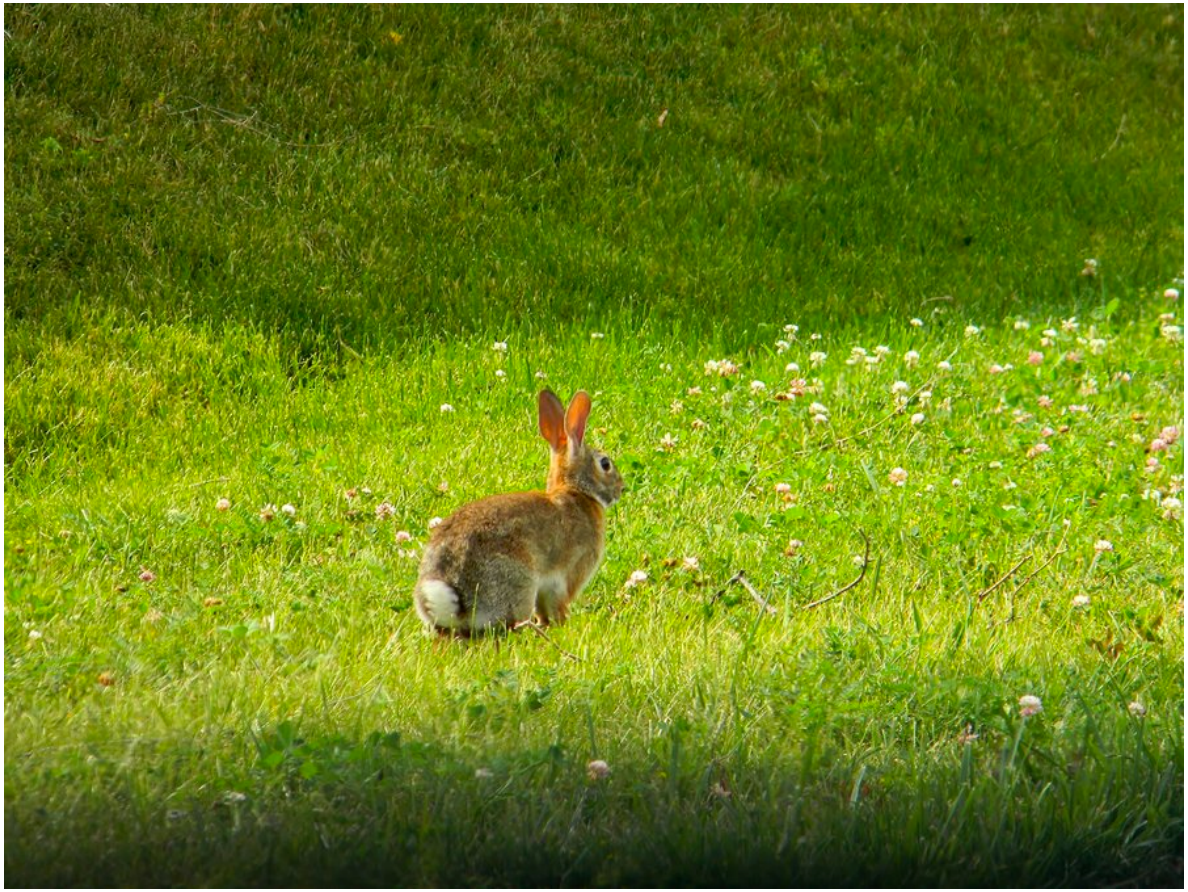
1. **Rational analysis**

   The structure of information available for solving a problem can tell you how people solve it

2. **Learning words from language *structure***

   You can learn a word's meaning by tracking the structure of its use

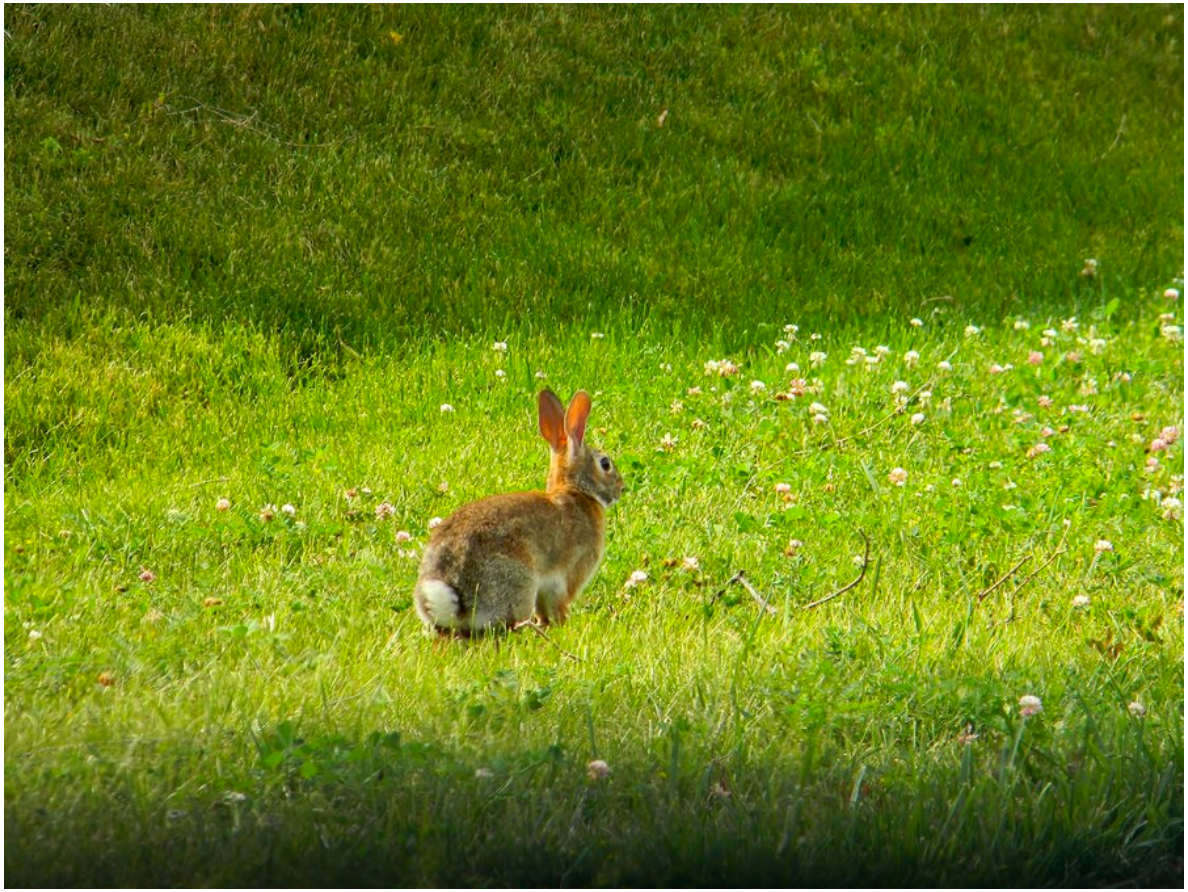
The gavagai problem



What does “gavagai” mean?

Quine (1960)

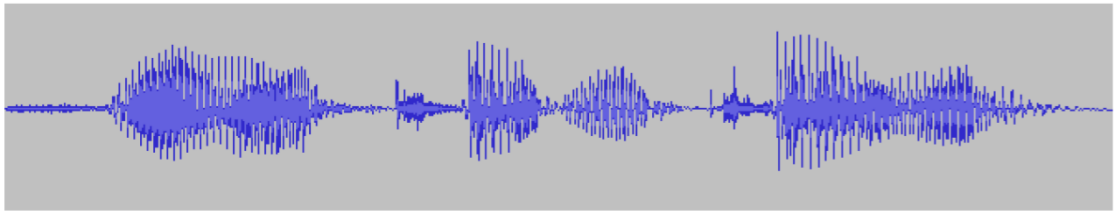
A warmup for the gavagai problem



How do you know that “gavagai” is one word?

Quine (1960)

How do you know where the words are>



He re ki tt y ki tt y

Word boundaries are not marked by silences!

But we can hear them anyway

Segmenting words by detecting dependency

olookwhataprettybaby
whataprettyshirt
ohlookatthehappybaby
itsprettylatealready
theresababycanyouseeit

But if you just heard ba, you're very like to next hear by.

P(by | ba) is high

If you just heard ty, you can't predict whether you will hear ba.

P(ba | ty) is low

Saffran, Aslin, & Newport (1996)

Segmenting words by detecting dependency

D=buladobigokudatibabuladotadupabigoku...

$$P(H|D) \propto P(D|H)P(H)$$

$$P(\text{bigoku}|D) \propto P(D|\text{bigoku})\cancel{P(\text{bigoku})}$$

$$P(\text{dobigo}|D) \propto P(D|\text{dobigo})\cancel{P(\text{dobigo})}$$

Let's assume words are arbitrary—
no word is any more likely to be in our language

Segmenting words by detecting dependency

P(bi|do) = .33

buladobigokudatibabuladotadupabigoku

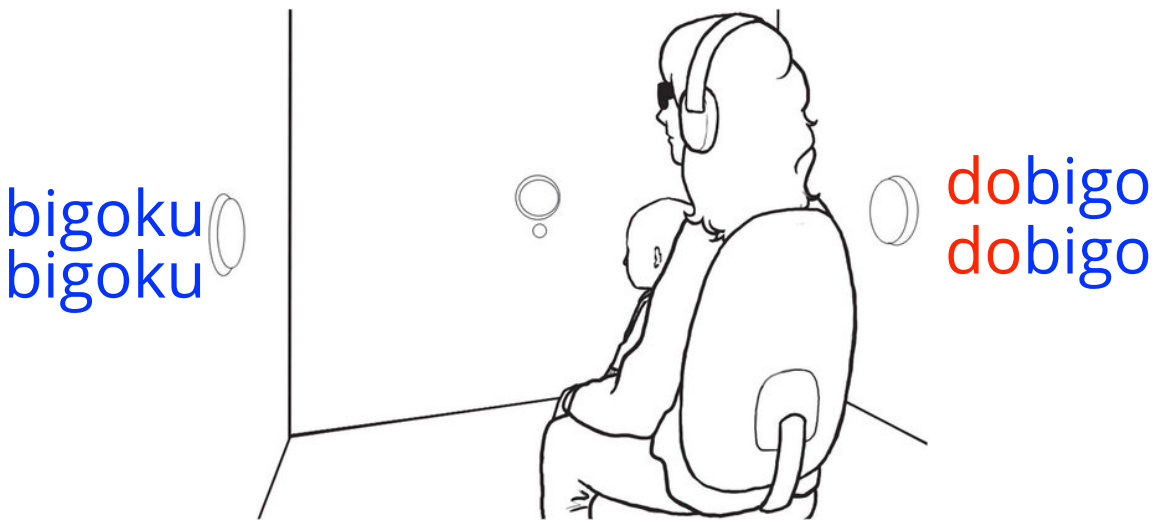
P(ku|go) = 1

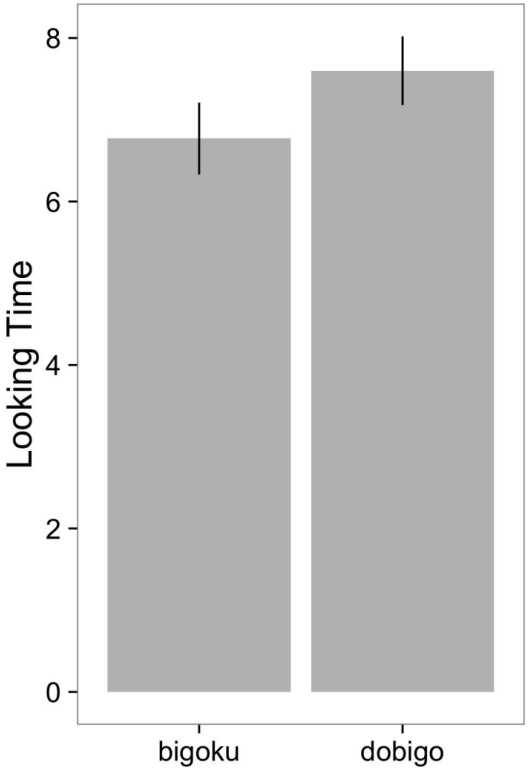
bigoku: P(ku|go) * P(go|bi) = 1

dobigo: P(go|bi) * P(bi|do) = .33

Saffran, Aslin, & Newport (1996)

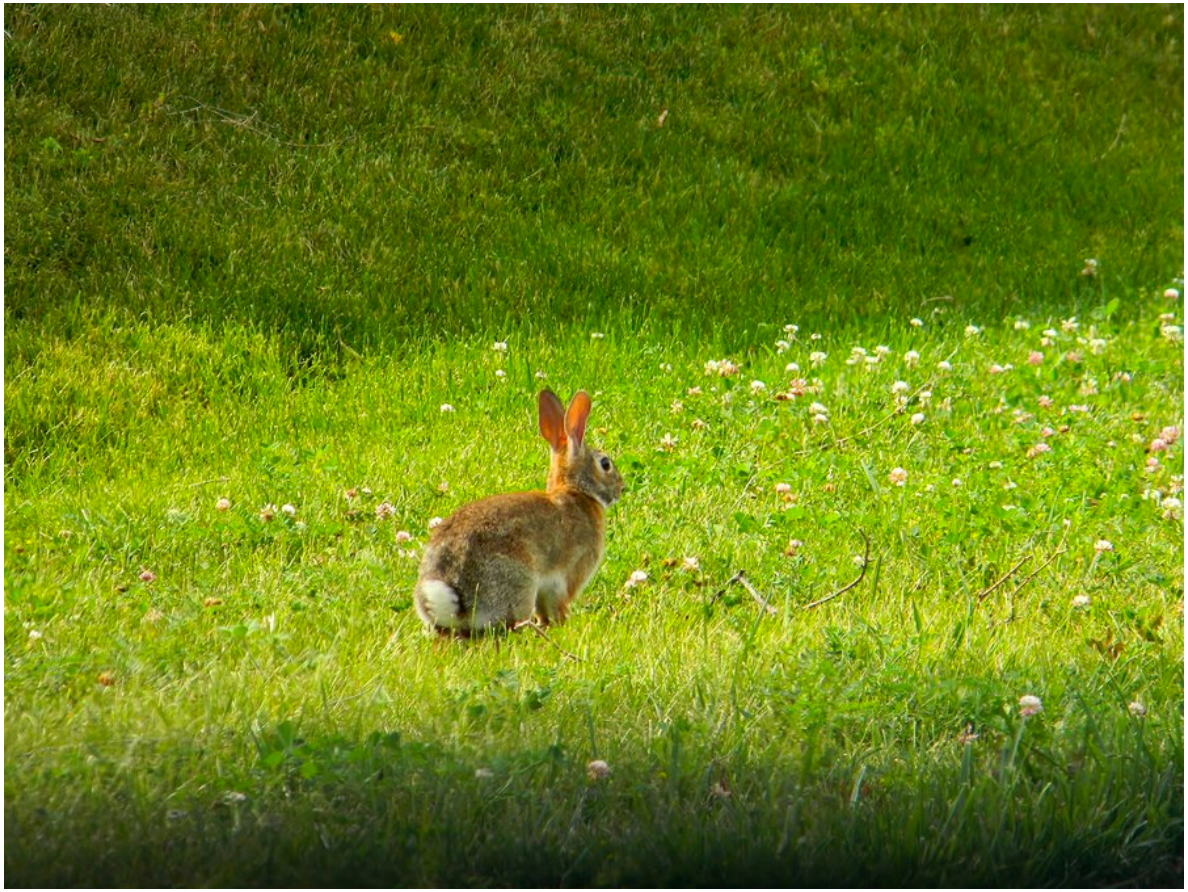
8-moth-old infants can use statistics to segment speech





Saffran, Aslin, & Newport (1996)

The gavagai problem



What does “gavagai” mean?

Quine (1960)

Conditional probability for meaning?

gavagai

gavagai

gavagai

If we often talk about the things around us, then we can use the same strategy to learn the meaning of gavagai

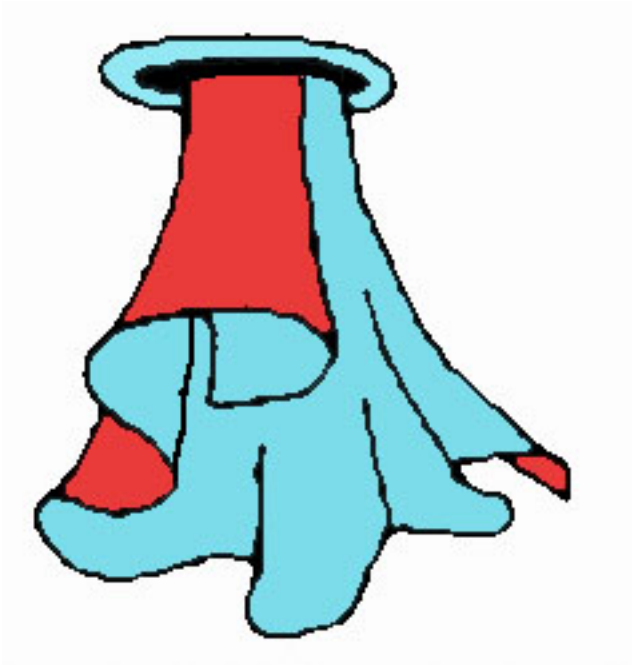
P(gavagai | ) is high

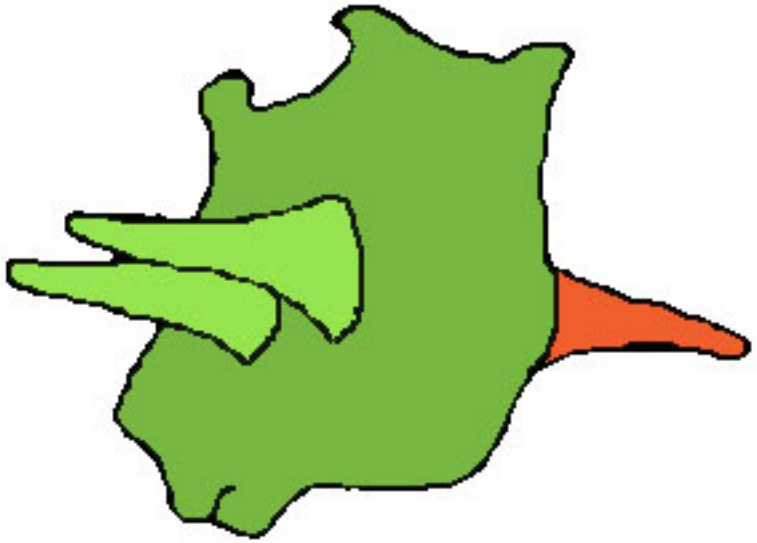
P(gavagai | ) is low

Siskind (1996)

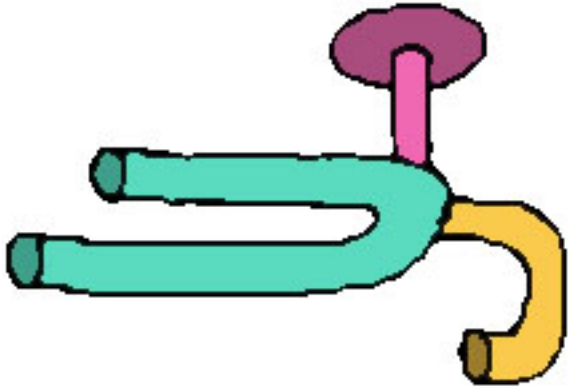
Can you learn words this way?

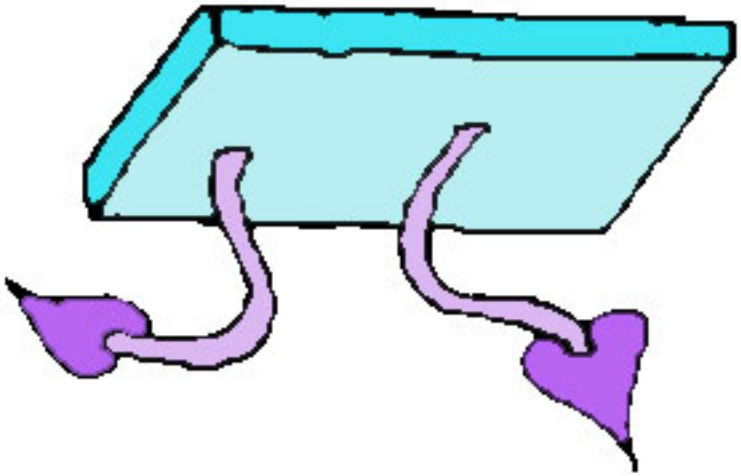
Can you learn words this way?



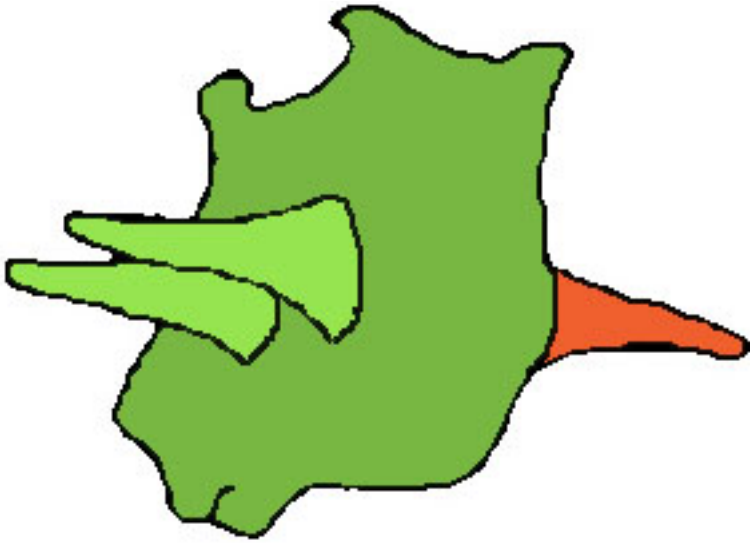


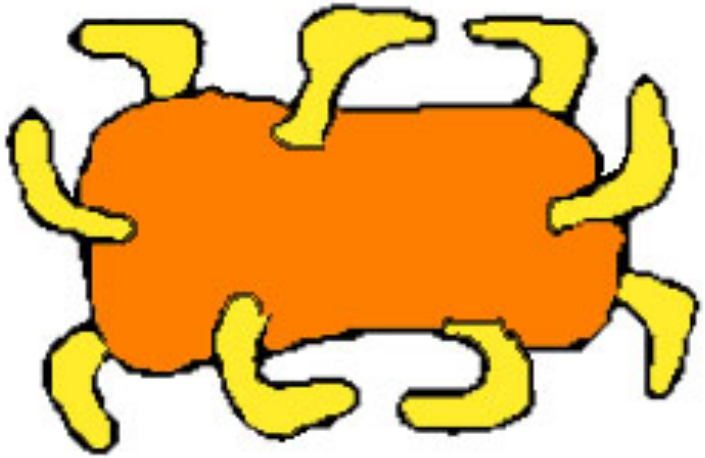
Can you learn words this way?



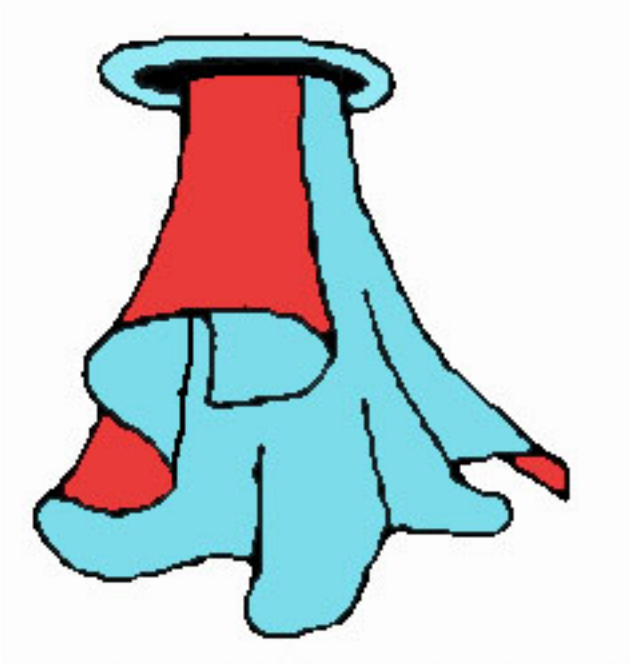


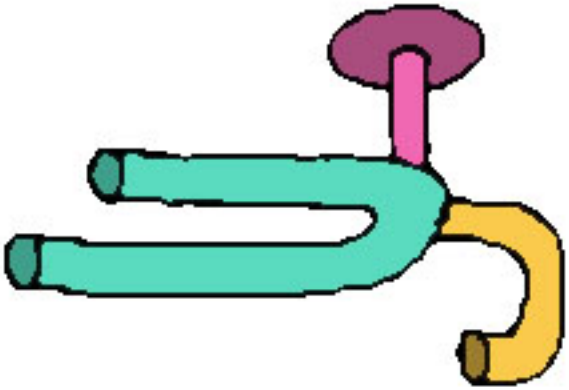
Can you learn words this way?



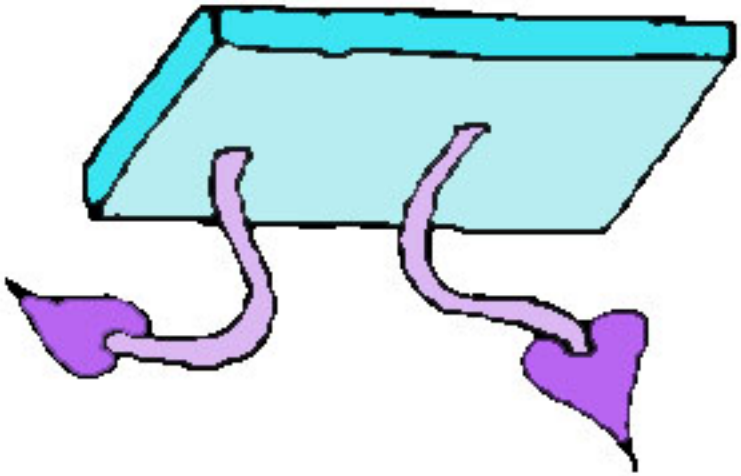


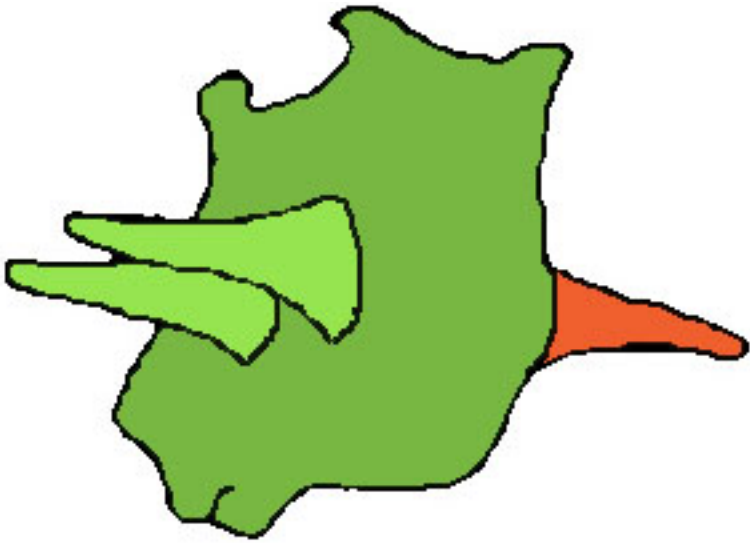
Can you learn words this way?





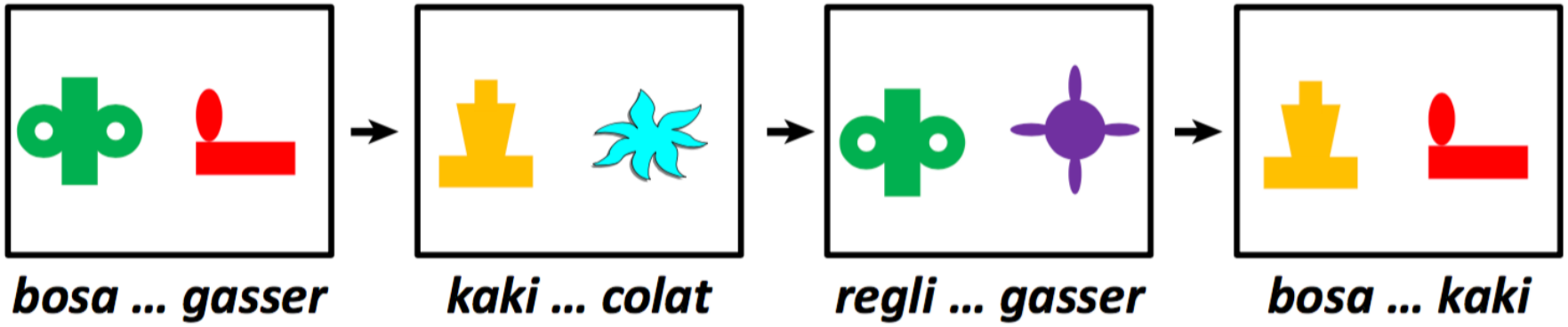
Can you learn words this way?

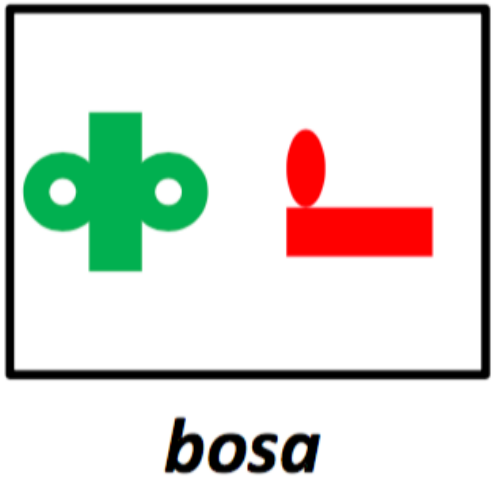


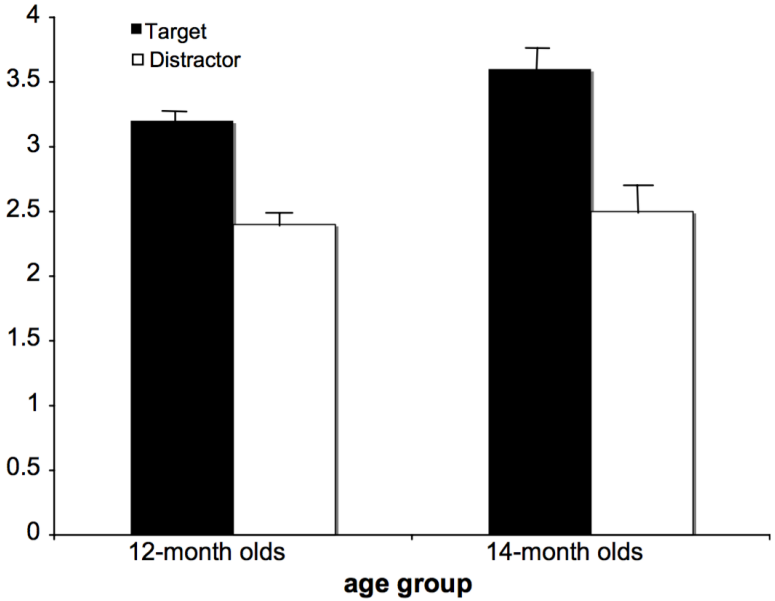


Which one is the jic?

Children can use this structure to learn words







Smith & Yu (2008)

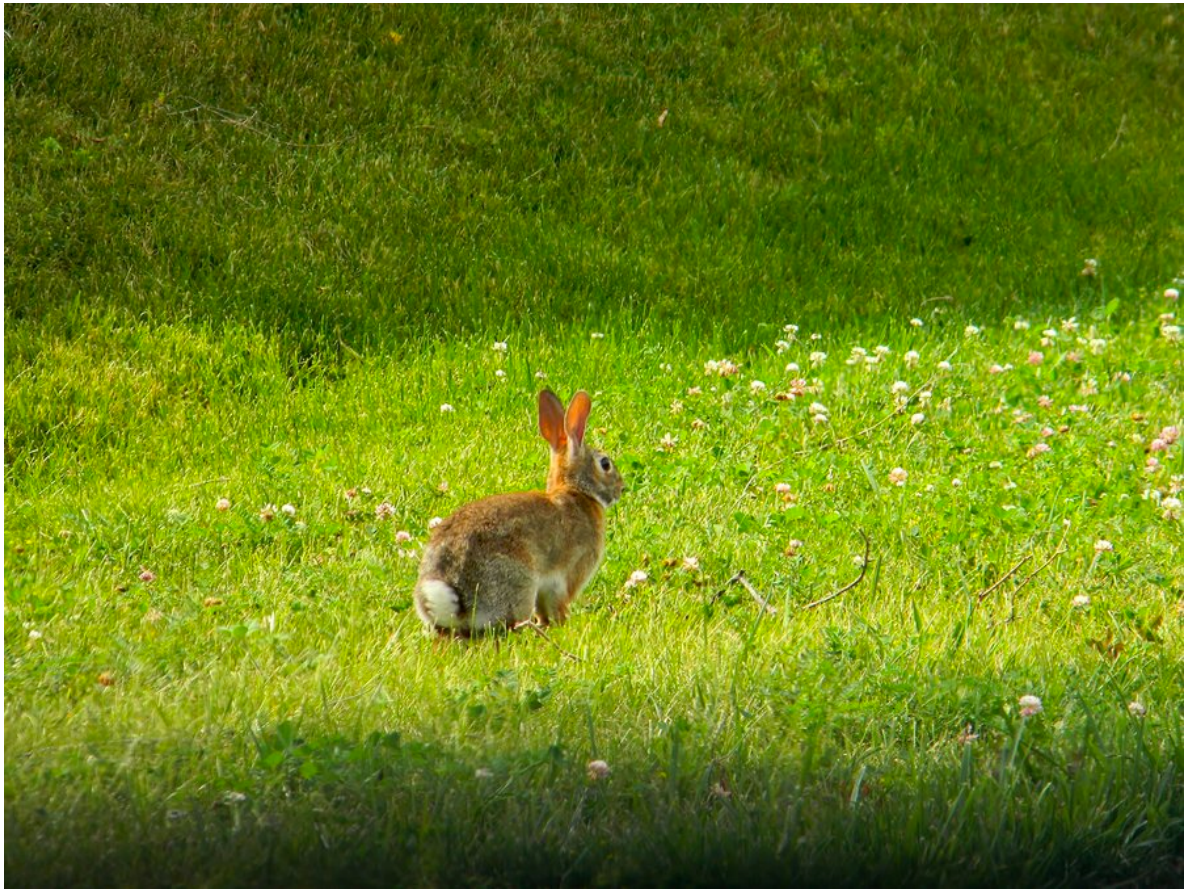
Three big ideas

1. **Rational analysis**
   The structure of information available for solving a problem can tell you how people solve it

2. **Learning words from language *structure***
   You can learn a word's meaning by tracking the structure of its use

3. ***Language structure* reflects *intentional structure***
   You can use intentions to learn language, or language to learn intentions

The gavagai problem



But what if gavagai means animal?

or Angora rabbit?

Quine (1960)

Let’s try it out

dax

dalmatian

dog

animal

dax

dax

dax

dalmatian

dog

animal

Xu & Tenenbaum (2007)

What's going on here?

$$P(H|D) \propto P(D|H)P(H)$$

P(dog| ) ∝ P( |dog)P(dog)

P(dalmation| ) ∝ P( |dalmation)P(dalmation)

What is P( |dog)?

Xu & Tenenbaum (2007)

The size principle

P( | dog)

<

P( | dalmation)

Xu & Tenenbaum (2007)

What's going on here?

$$P(H|D) \propto P(D|H)P(H)$$

P(dog| ) ∝ P( |dog)P(dog)

P(dalmation| ) ∝ P( |dalmation)P(dalmation)

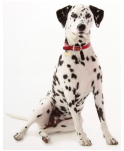
What is P(dog)? What is P(dalmation)?

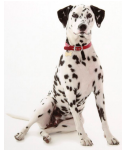
So maybe P(dog) > P(dalmation)

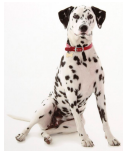
Xu & Tenenbaum (2007)

3 dalmations from the dog category? A suspicious coincidence!

$$P(H|D) \propto P(D|H)P(H)$$

P(dog|  ,  , ) ∝

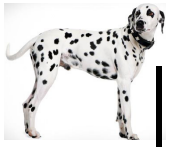
P( ,  ,  | dog)P(dog)

P(dalmation|  ,  , ) ∝

P( ,  ,  | dalmation)P(dalmation)

Xu & Tenenbaum (2007)

The size principle

P( ,  ,  | dog)

If I'm picking examples from the dog category, it's really unlikely to pick three dalmations

Xu & Tenenbaum (2007)

Let’s try it out

dax

dalmatian

dog

animal

dax

dax

dax

dalmatian

dog

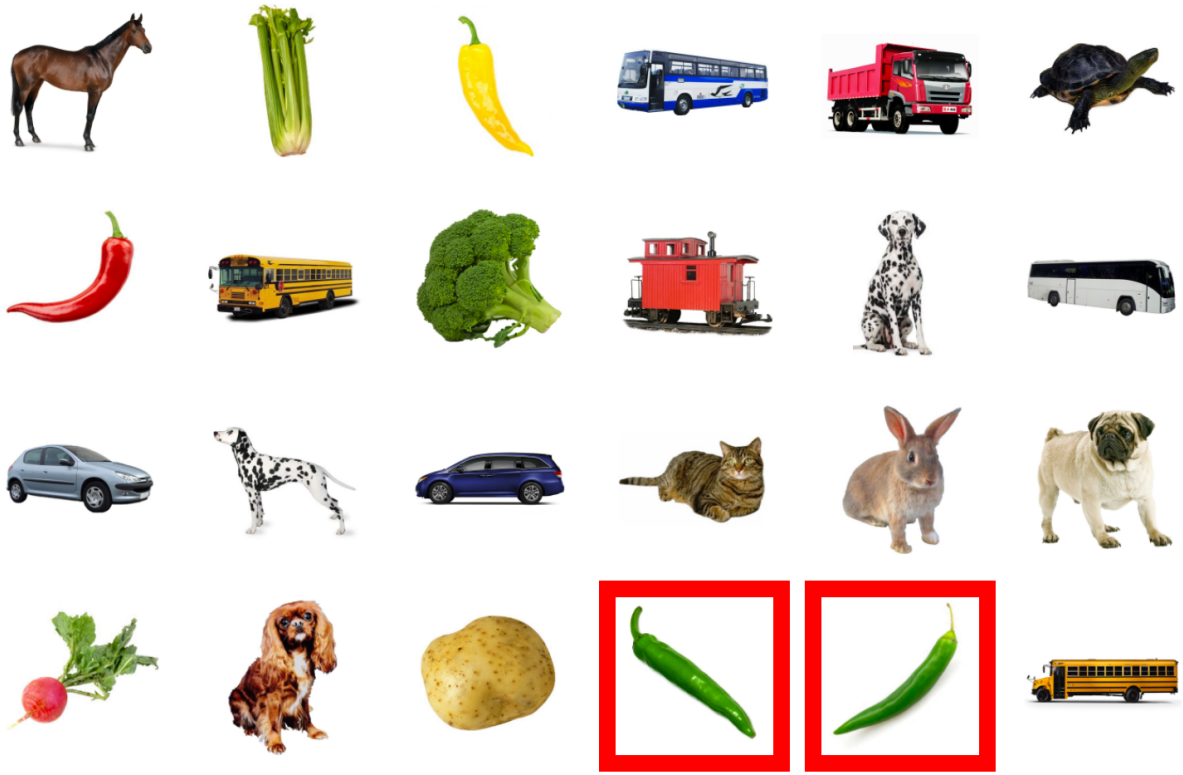
animal

Xu & Tenenbaum (2007)

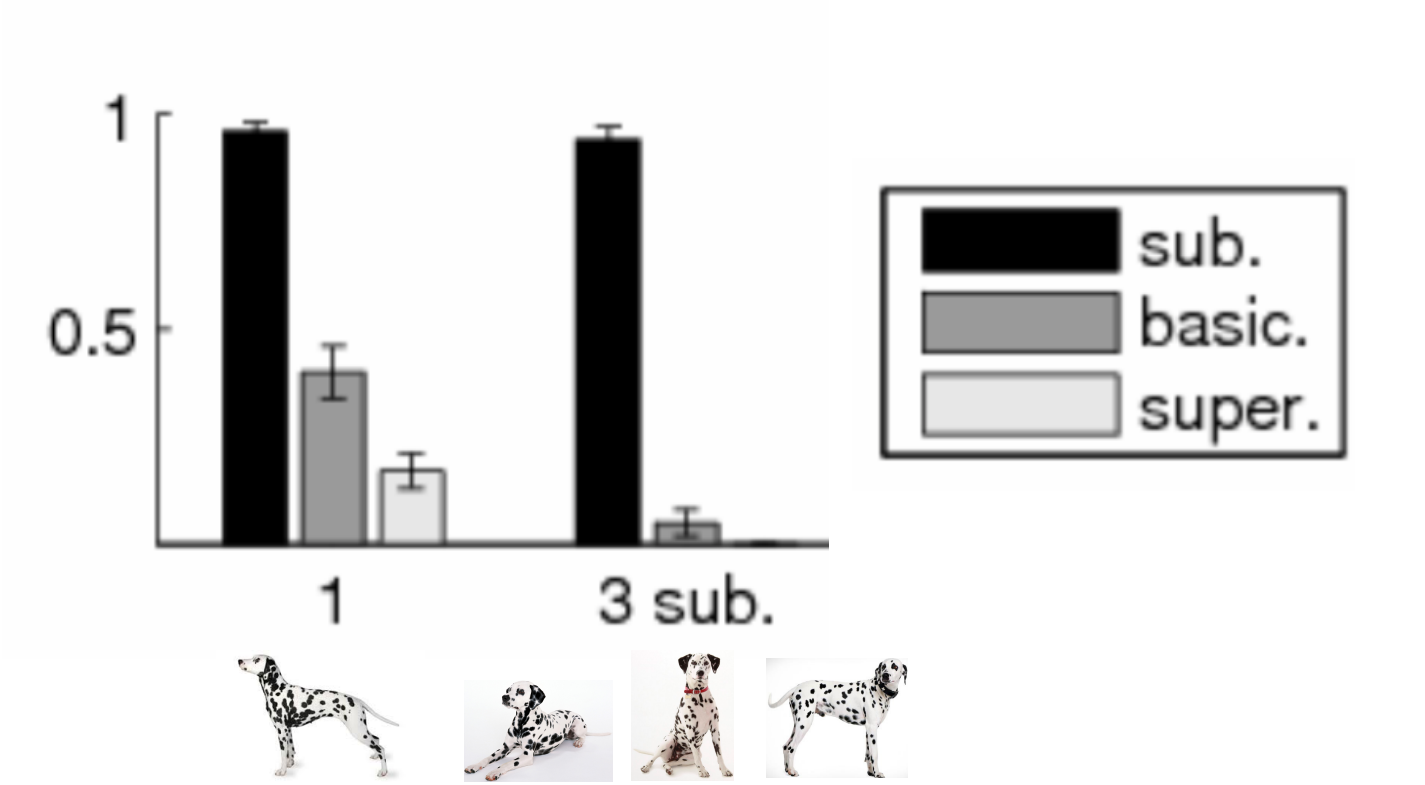
Testing the suspicious coincidence

Here are three sibs. Can you give Mr. Frog all the other sibs?

To give a sib, click on it below. When you have given all the sibs, click the Next button.

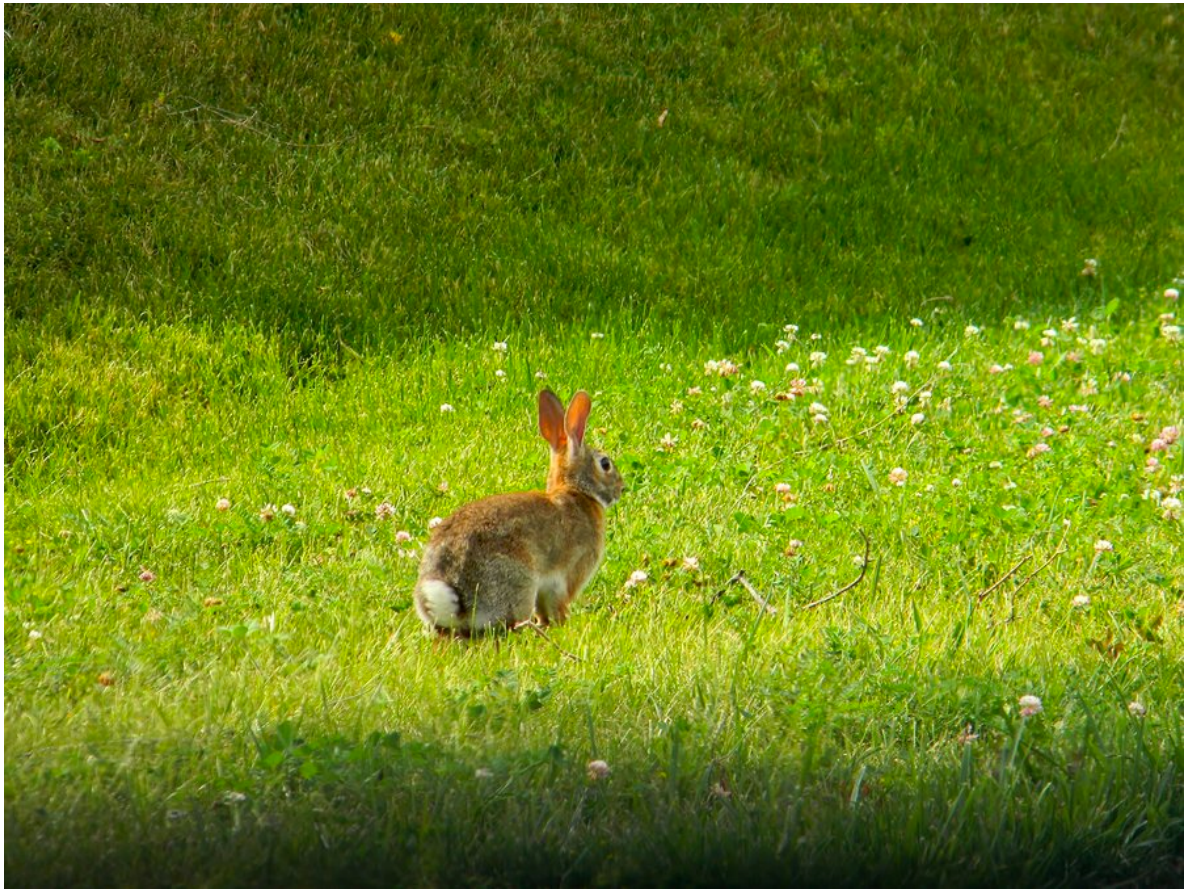


3- and 4-year-olds make this inference



Xu & Tenenbaum (2007)

The gavagai problem



But what if gavagai means rabbit ears?
or rabbit feet?

Quine (1960)

Pragmatic inference

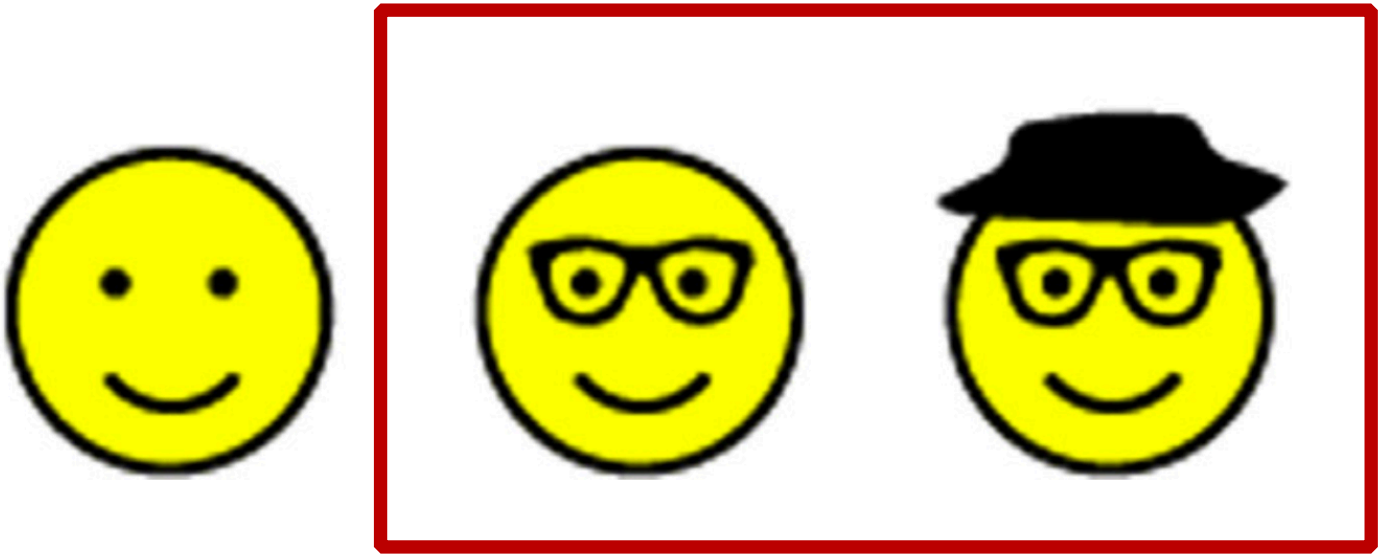
Suppose you heard me say
"My friend has glasses"

Which one is my friend?

Goodman and Frank (2016)

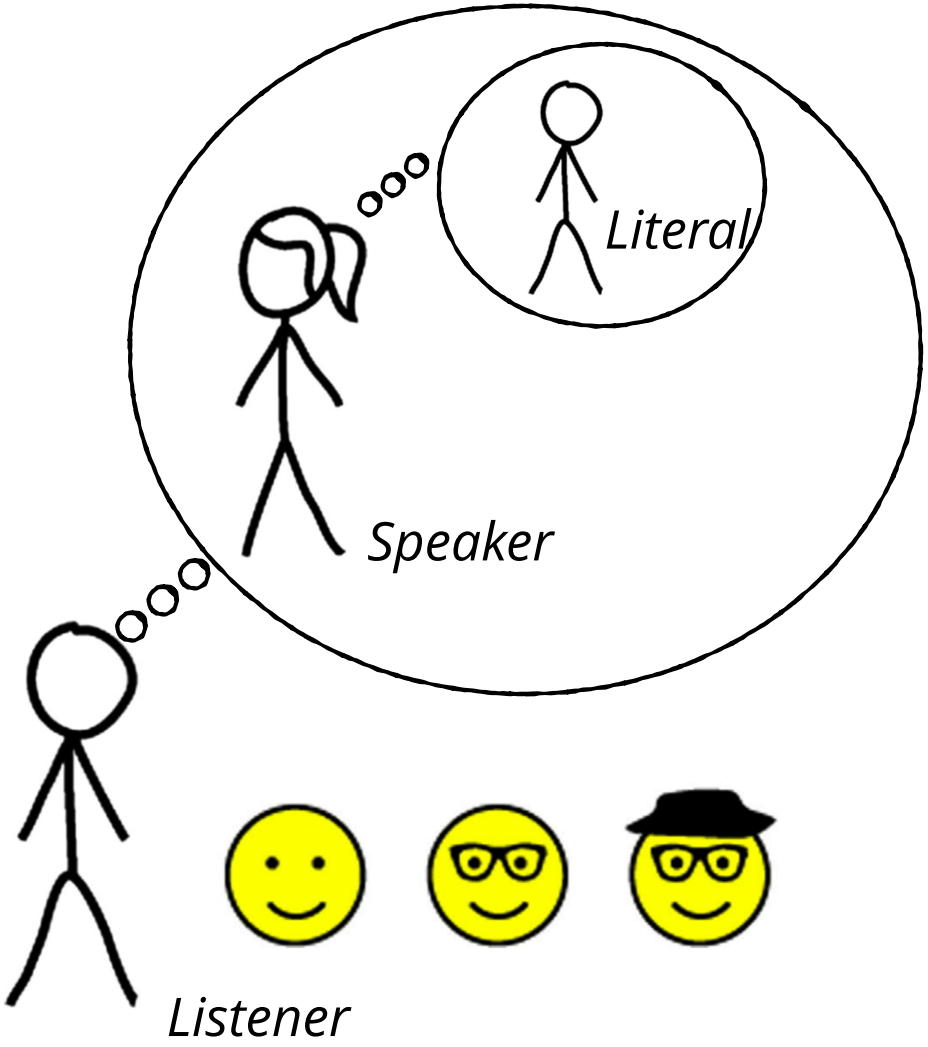
Pragmatic inference

Why not guess randomly from these two?



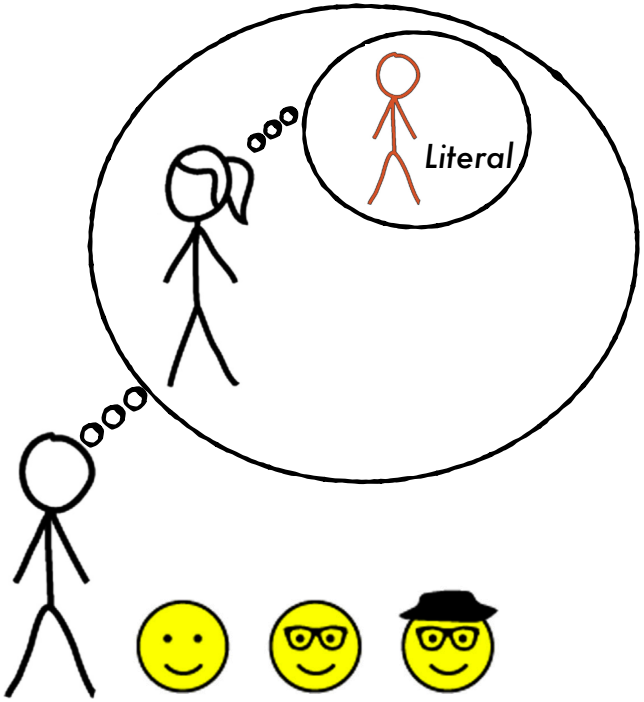
Goodman and Frank (2016)

The recursive reasoning of pragmatic inference



Goodman and Frank (2016)

The literal listener



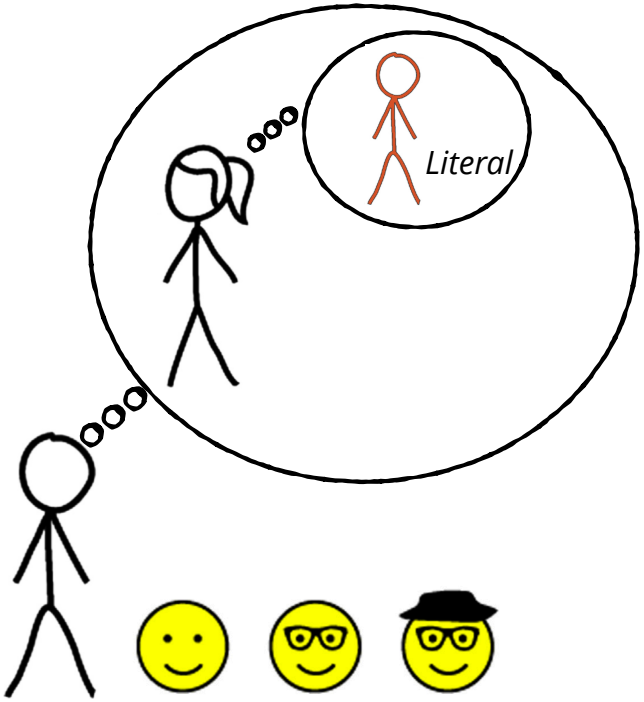
The Literal listener randomly chooses a face that matches the description

$$P(\bullet \mid word) = \frac{\delta(\bullet, word)}{\sum_{\bullet'} \delta(\bullet', word)}$$

$$\delta(\bullet, word) = \begin{cases} 1 \; if \; True \\ 0 \end{cases}$$

Goodman and Frank (2016)

Checking our intuition about the literal listener



P( | glasses) =

P( | hat) =

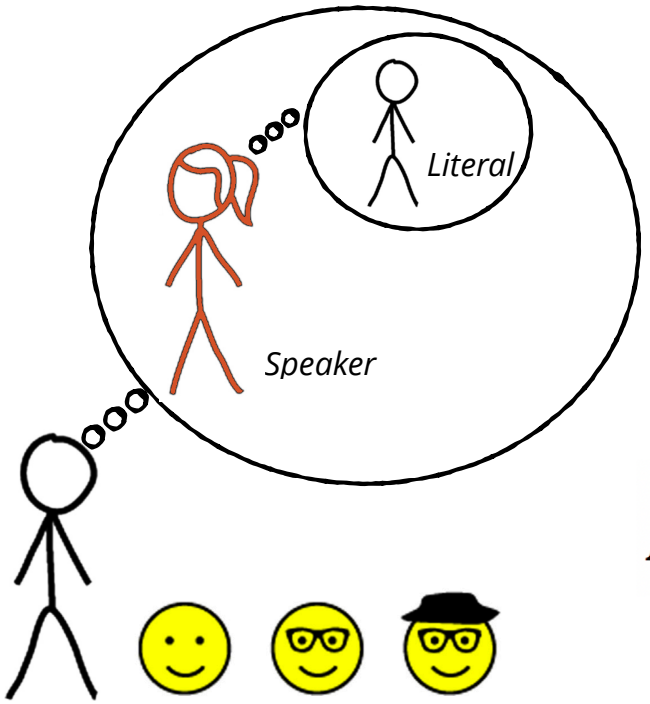
P( | hat) =

P( | glasses) =

P( | glasses) =

Goodman and Frank (2016)

The speaker



The Speaker chooses a word in proportion to informativeness to the Literal listener

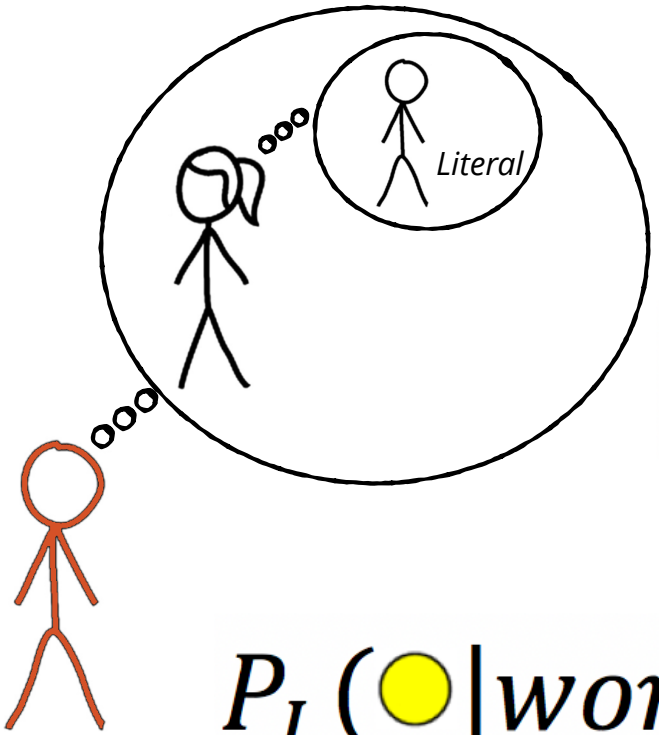
$$P_s(word|\bigcirc) \propto P_{lit}(\bigcirc|word)$$

$$P_s(glasses|\text{🤓}) \propto P_{lit}(\text{🤓}|glasses) = ½$$

$$P_s(hat|\text{🤓}) \propto P_{lit}(\text{🤓}|hat) = 1$$

Goodman and Frank (2016)

The pragmatic listener



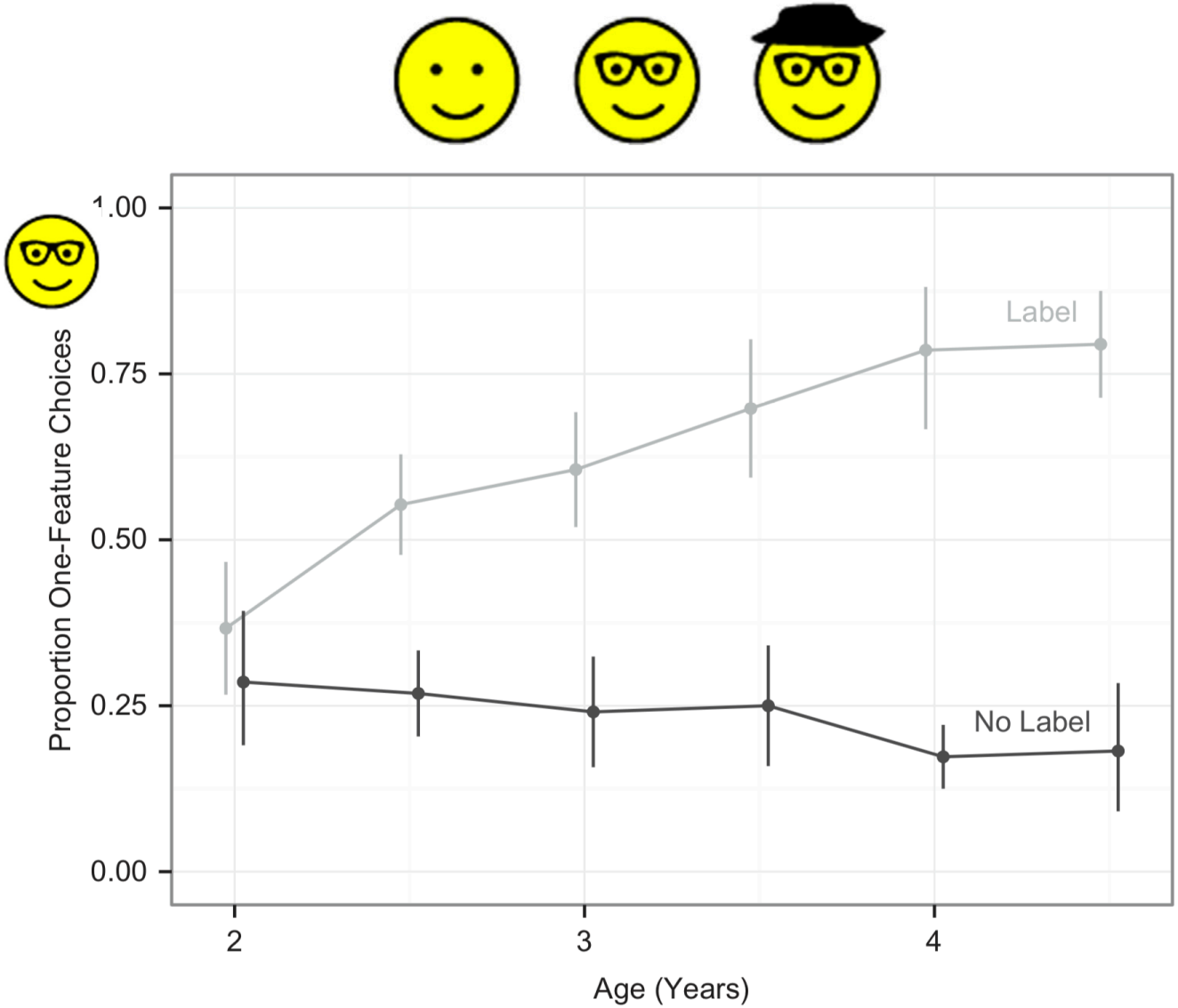
The Listener chooses a referent in proportion to how likely the Speaker is to have used that word to refer to it

$$P_s(word|\bigcirc) \propto P_{lit}(\bigcirc|word)$$

$$P_L(\bigcirc|word) \propto P_S(word|\bigcirc)P(\bigcirc)$$

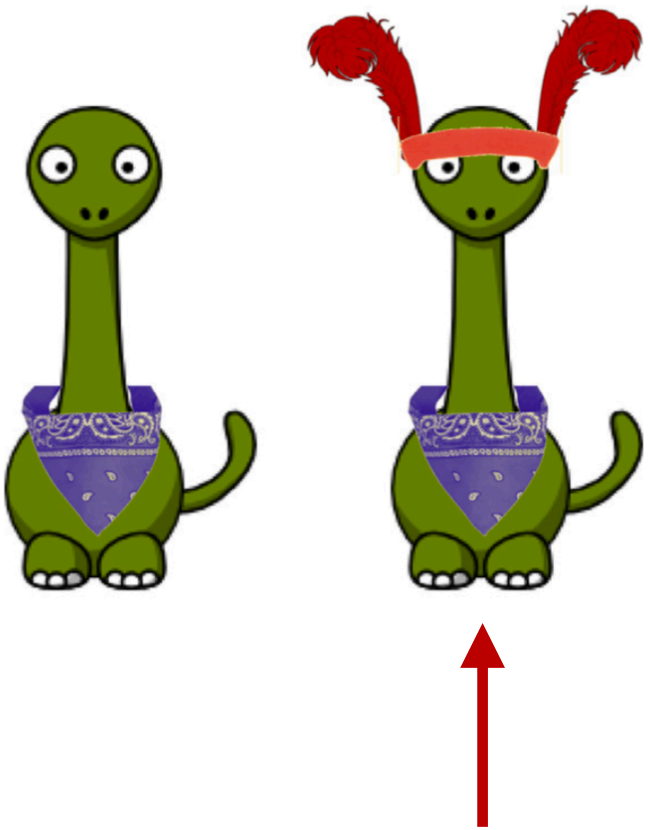
Goodman and Frank (2016)

Pragmatic inference in young children



Stiller, Goodman, and Frank (2015)

Using pragmatic inference to learn words



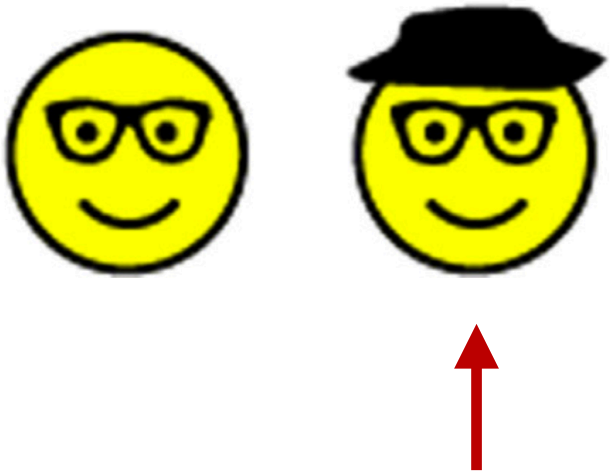
This is a dinosaur with a dax

Frank and Goodman (2014)

Using pragmatic inference to learn words

$P_L(\text{🤓}|glasses) \propto$

$P_S(glasses|\text{🤓})P(\text{🤓})$



By the power of Bayes' rule!

$P_L(glasses|\text{🤓}) \propto$

$P_S(\text{🤓}|glasses)P(glasses)$

This is a friend with a dax

Frank and Goodman (2014)

Three big ideas

1. **Rational analysis**
   The structure of information available for solving a problem can tell you how people solve it

2. **Learning words from language *structure***
   You can learn a word's meaning by tracking the structure of its use

3. ***Language structure* reflects *intentional structure***
   You can use intentions to learn language, or language to learn intentions

What went wrong here?