# Cognition

PSYC 2040

W6: Language

Part 2

# today’s agenda: language learning

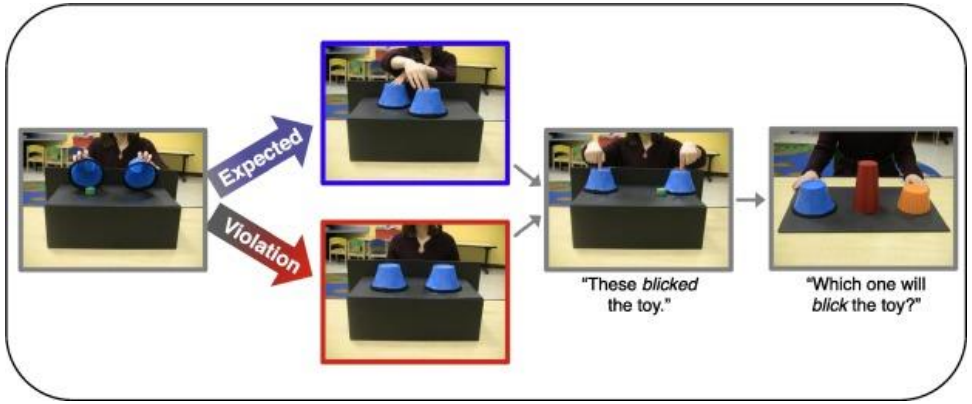
- statistical regularities / co-occurrences
- prediction
- form to meaning mappings
- social inference

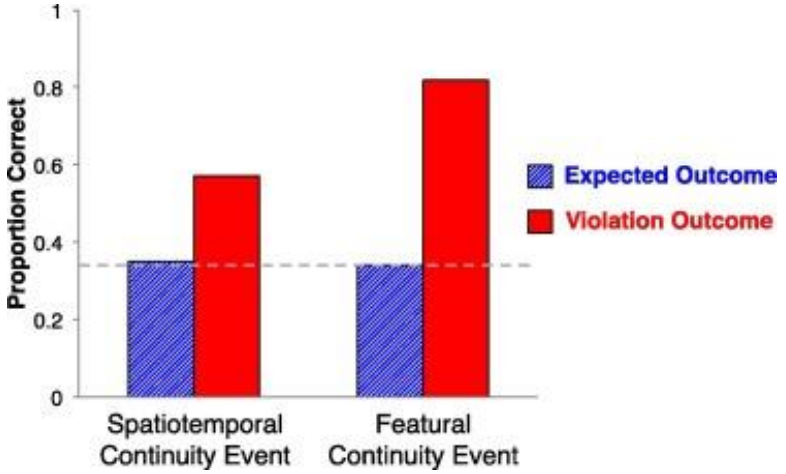
# why track statistics?

- infants are not required to or motivated by reward to track statistics, so why do they do it?
- possible hypotheses:
  - infants want to generate predictions about the environment
  - infants want to communicate with their caregivers

# statistical learning and prediction

- Stahl & Feigenson (2017) tested 3- to 6-year-old children in an experiment where novel labels (blick) were mapped to actions in expected or violation conditions
  - expected : toy in the expected location
  - violated: toy in the unexpected location
- learning was maximized when children were surprised by the outcomes



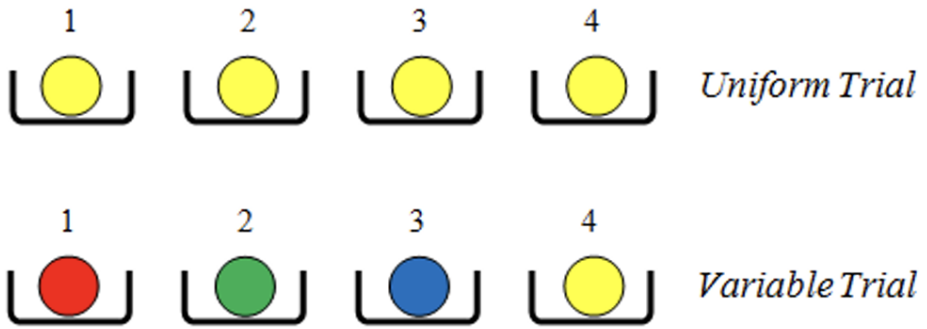



# statistical learning and curiosity

- statistical learning may also inform what to learn about in the first place
- curiosity may be particularly important in creating learning opportunities and minimizing uncertainty in the environment

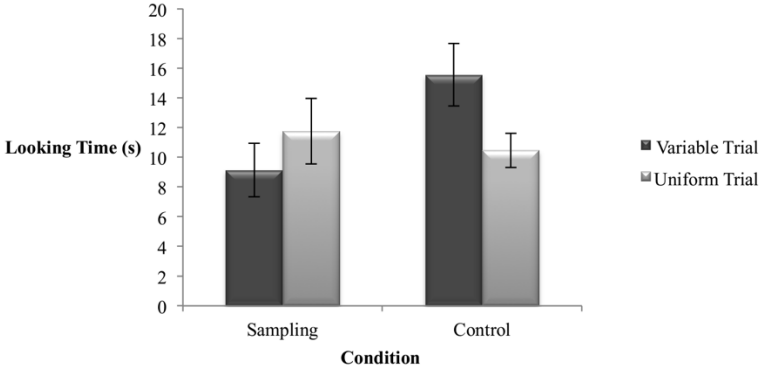
# statistical learning and curiosity

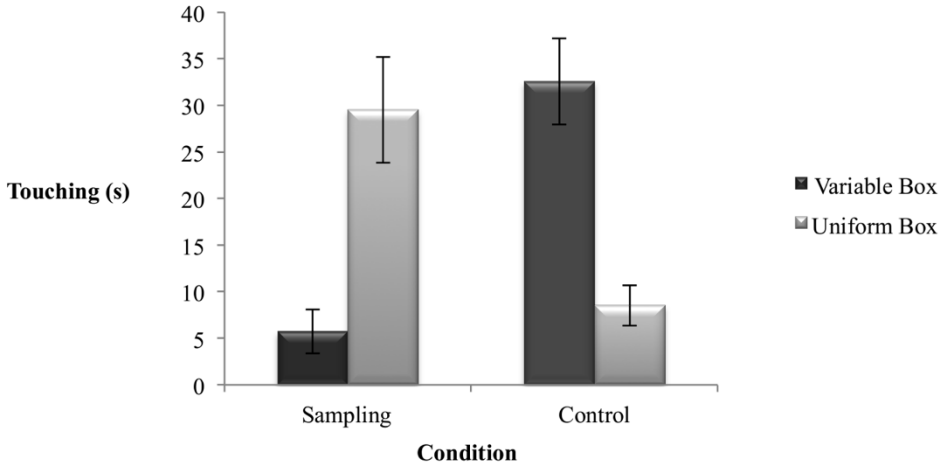
- Sim & Xu (2017) tested 13-month-old infants in a violation of expectation (VOE) and crawling paradigm
- conditions:
  - knowledge: control condition (experimenter looked into the box before drawing out the balls) or sampling (no looking)
  - draw: could be “uniform” or “variable”
- two experiments: looking time (VOE) vs. touching/reaching time (crawling)

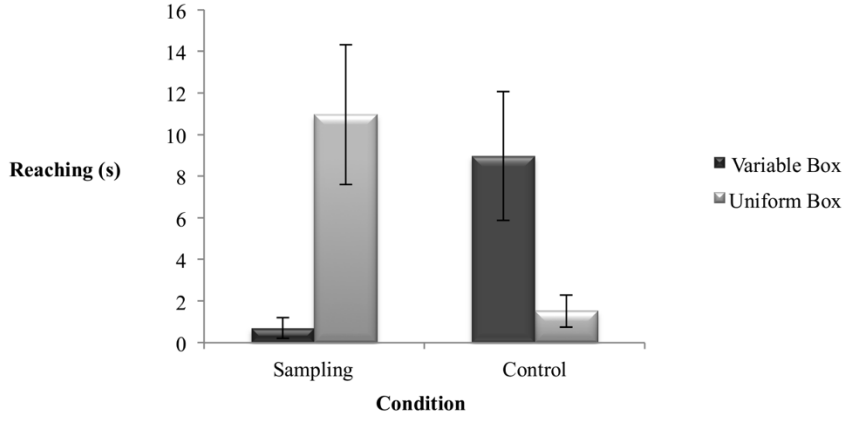


# statistical learning and curiosity

- Sim & Xu (2017) showed that 13-month-old infants preferentially explore sources of unexpected events



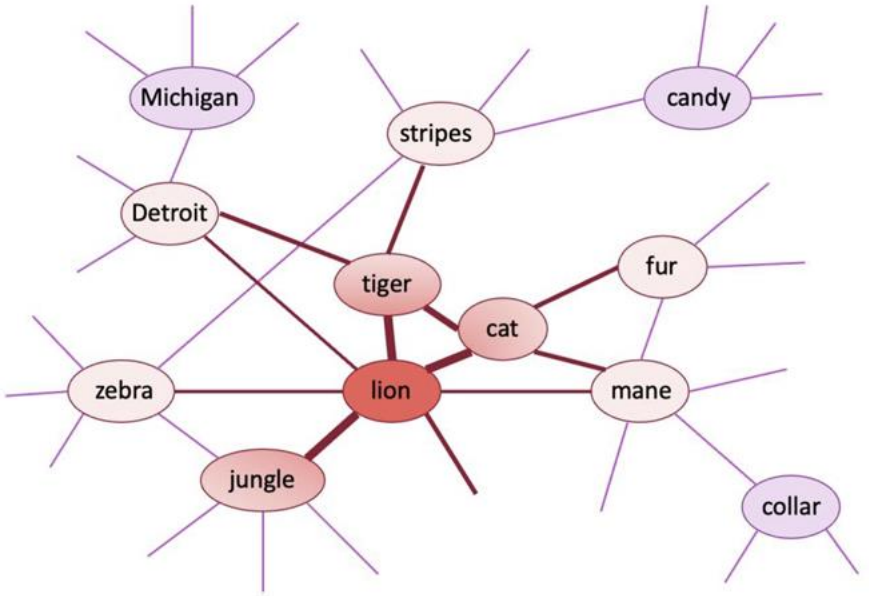



# review of findings/inferences

- we track statistical regularities
- we learn from prediction error
- we are inherently curious and want to reduce uncertainty

# activity debrief

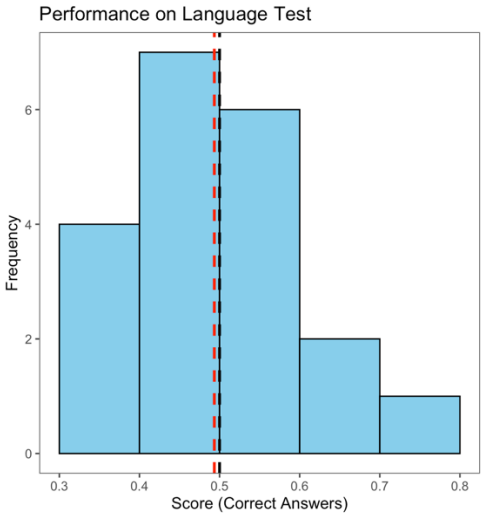
- think back to the language experiment you did
- you were shown sentences from a language with 27 words
- you then judged whether pairs of words could have been from the language or not
- now, think back to all the words and try to see if you can create a semantic network from them
- words you think are more related have direct links



# activity debrief: arrange on the board

- jhool
- kha
- ladka
- baagh
- masal
- hiran
- dekh
- seengh
- chori
- ber
- gaddi
- bhayankar
- soongh
- ega
- komal
- tez
- qalabaz
- kela
- bhaag
- shikar
- pakad
- bandar
- hil
- nariyal
- gabbar
- svaadisht
- suraj

# english to hindi translations!



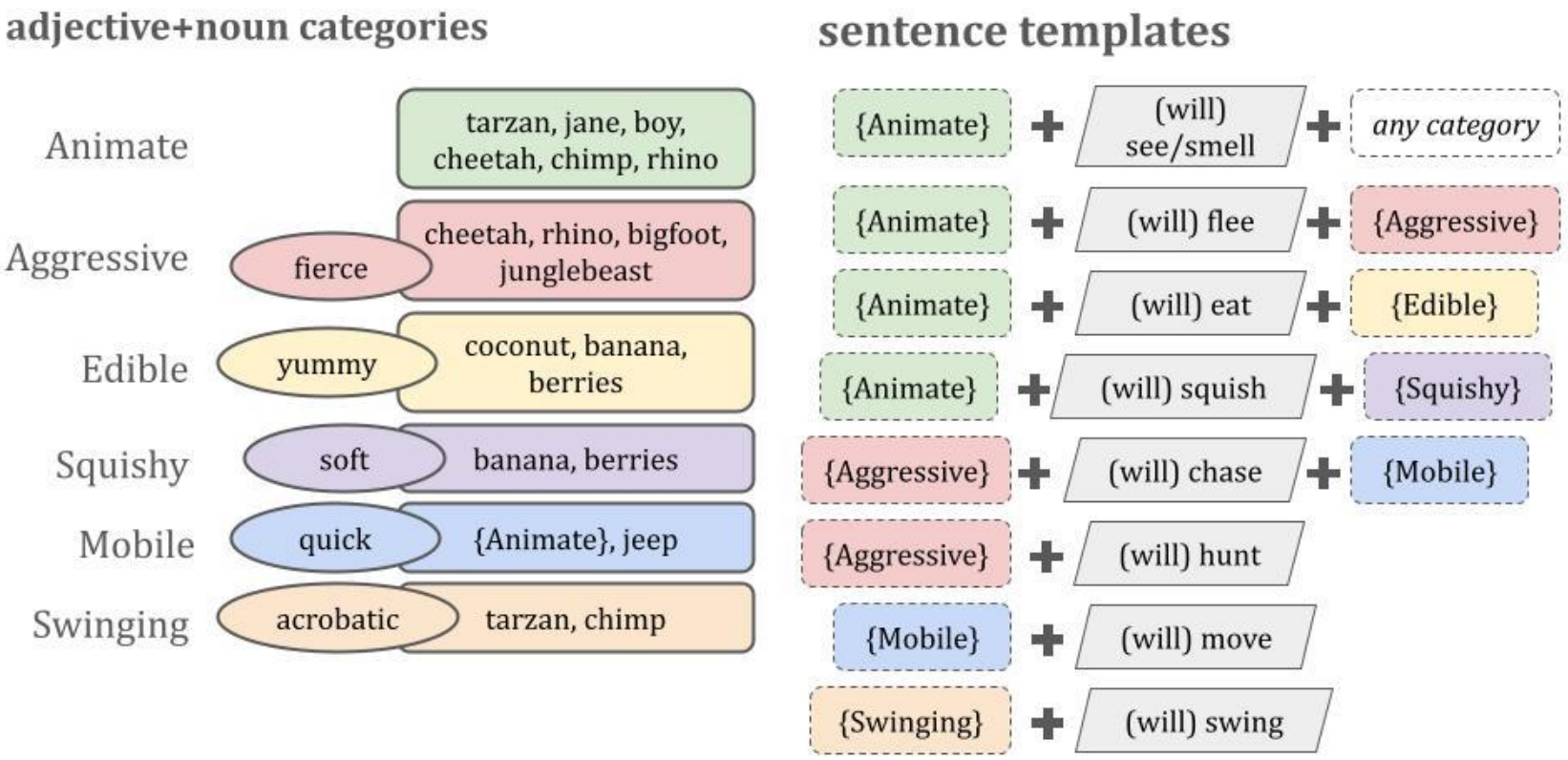

- tarzan = suraj
- jane = chori
- boy = ladka
- cheetah = baagh
- chimp = banmanush
- rhino = gainda
- bigfoot = gabbar
- junglebeast = hiran
- coconut = nariyal
- banana = kela
- berries = ber
- jeep = gaddi

+ fierce = bhayankar
+ yummy = svaadisht
+ soft = komal
+ quick = tez
+ acrobatic = qalabaz

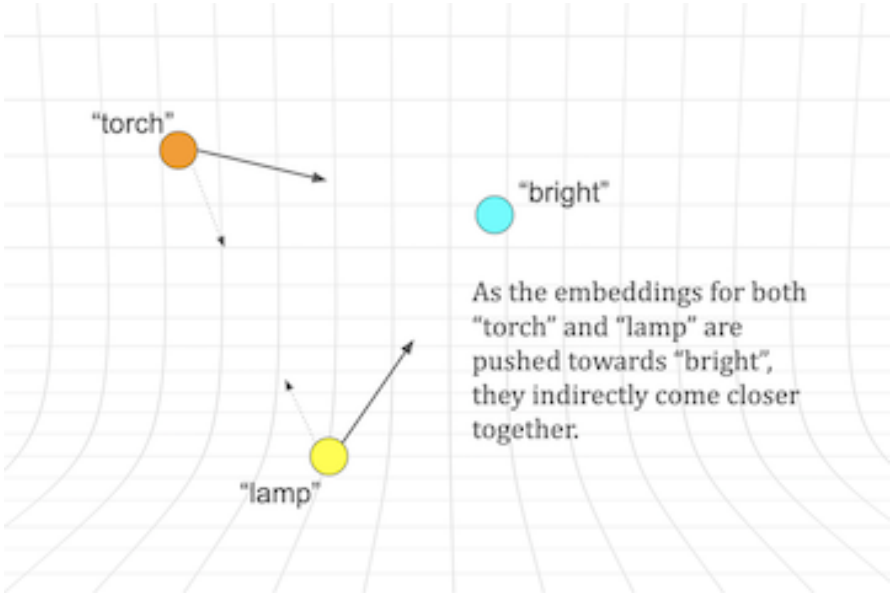
+ will = ega
+ flee = bhaag
+ hunt = shikaar
+ chase = pakad
+ squish = masal
+ move = hil
+ eat = kha
+ see = dekh
+ smell = soongh
+ swing = jhool

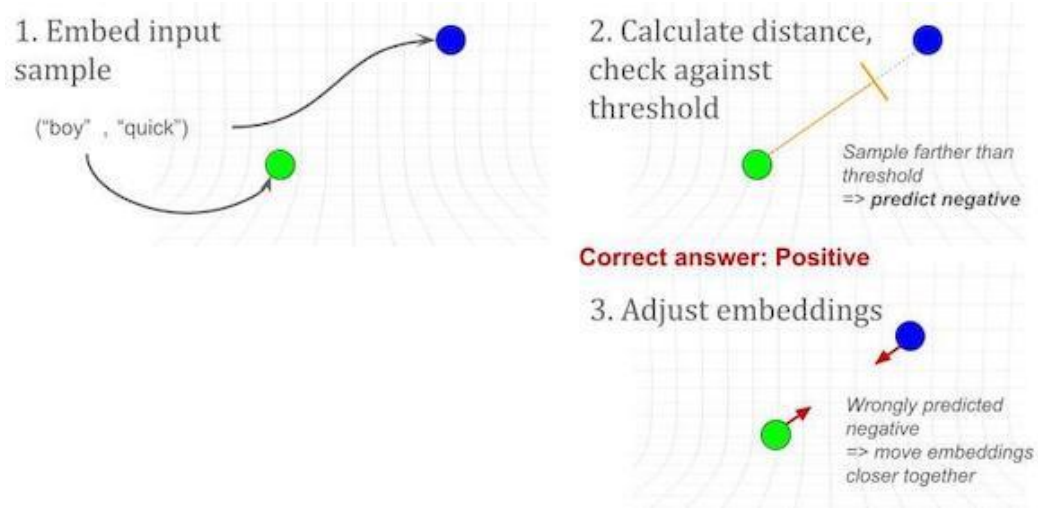
# language model demo

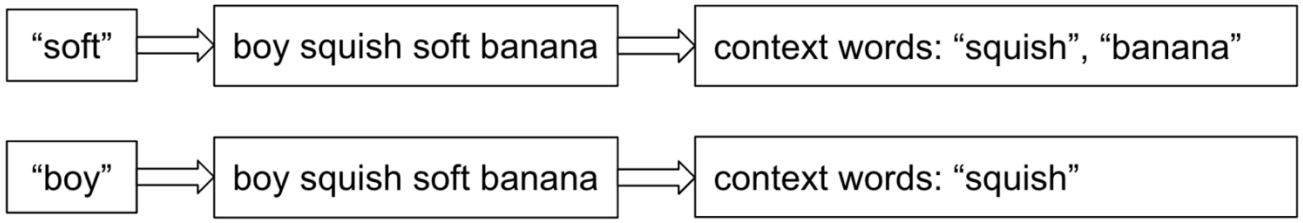
- code notebook



# training a language model

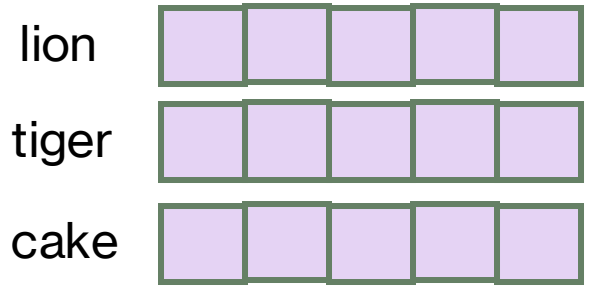


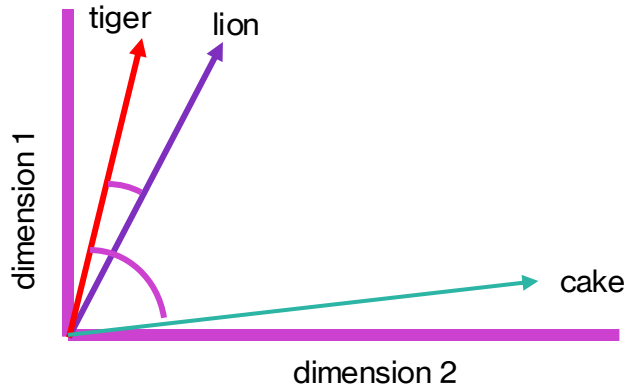



# large language models

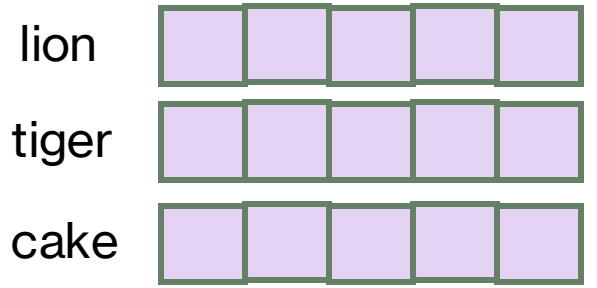
- **large** language models are typically “trained” on large databases of text (e.g., Wikipedia, Google News, etc.) + other sources (images, speech, etc.) and then finetuned using human annotations and feedback
- algorithm: prediction-based
- after training, we can look under the hood at what semantic ‘representations’ the models have acquired
- these representations are usually a collection of numbers but they are meaningfully related to each other in a high-dimensional space

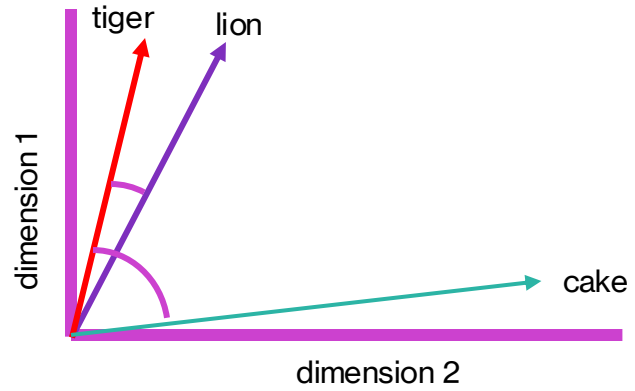


# key concepts

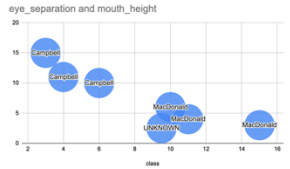
- word representations / embeddings
- dimensionality
- semantic similarity / cosine similarity



# dimensions and similarity

| type | Face | class | eye_separation | mouth_height |
|---|---|---|---|---|
| training | 1 | MacDonald | 10 | 6 |
| training | 2 | MacDonald | 11 | 4 |
| training | 3 | MacDonald | 15 | 3 |
| training | 4 | Campbell | 6 | 10 |
| training | 5 | Campbell | 4 | 11 |
| training | 6 | Campbell | 3 | 15 |
| TEST | TEST | UNKNOWN | 9.5 | 2.5 |





[-1.025, 0.493, 0.048, -0.666]


[-0.172, -0.557, 0.337, 0.163]

**lions** are **tigers** are carnivorous predators

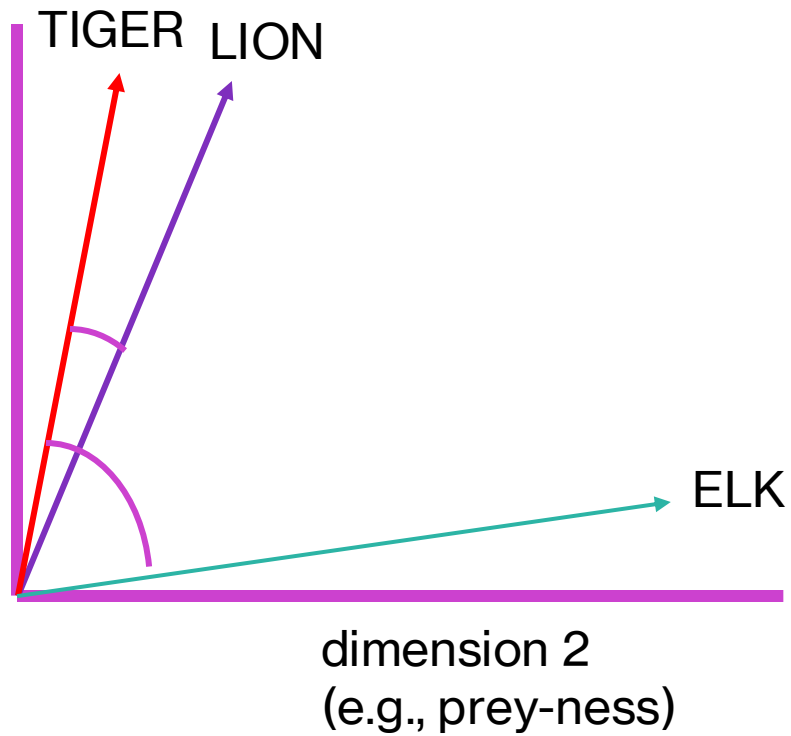
a group of **lions** is called a pride

**tigers** have stripes

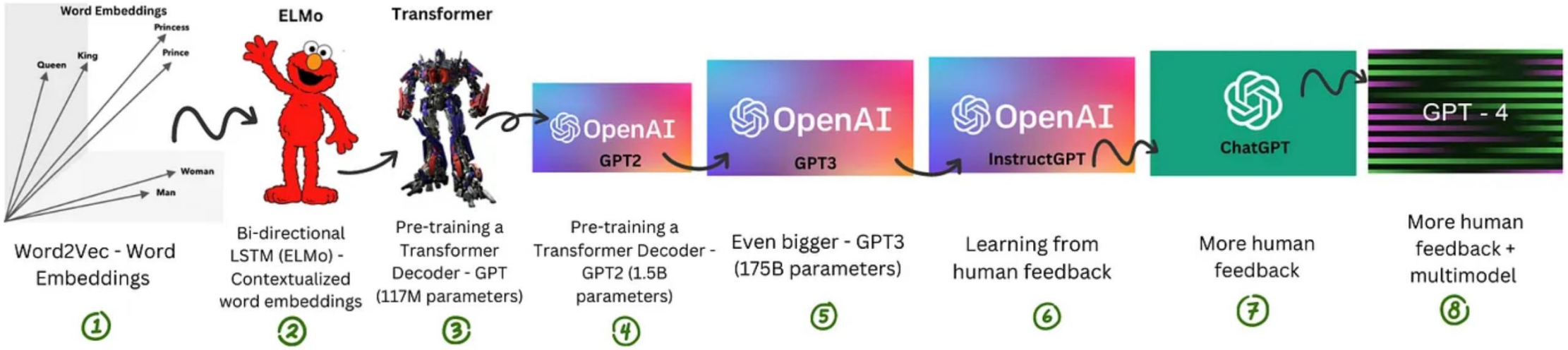
**lions** and **tigers** hunt deer and **elk**

**elk** and deer are herbivores

dimension 1
(e.g., predator-ness)



# language models: a timeline



# the wins of language models

- truly start from “scratch”
- dispel the need for hard-wiring many abilities by showing “emergent” behavior
- have widespread applications
- will pass the traditional Turing test: fooling a human

## Modern language models refute Chomsky’s approach to language

Steven T. Piantadosi[a,b]
[a]UC Berkeley, Psychology [b]Helen Wills Neuroscience Institute

The rise and success of large language models undermines virtually every strong claim for the innateness of language that has been proposed by generative linguistics. Modern machine learning has subverted and bypassed the entire theoretical framework of Chomsky’s approach, including its core claims to particular insights, principles, structures, and processes. I describe the sense in which modern language models implement genuine *theories* of language, including representations of syntactic and semantic structure. I highlight the relationship between contemporary models and prior approaches in linguistics, namely those based on gradient computations and memorized constructions. I also respond to several critiques of large language models, including claims that they can’t answer “why” questions, and skepticism that they are informative about real life acquisition. Most notably, large language models have attained remarkable success at discovering grammar without using any of the methods that some in linguistics insisted were necessary for a science of language to progress.

# potential concerns: data

- the size of the corpora that models are trained on is 1000 times more than the input available to children
- most models are based on the English language (Bender rule)
- most advanced models learn from data AND fine-tuned human feedback

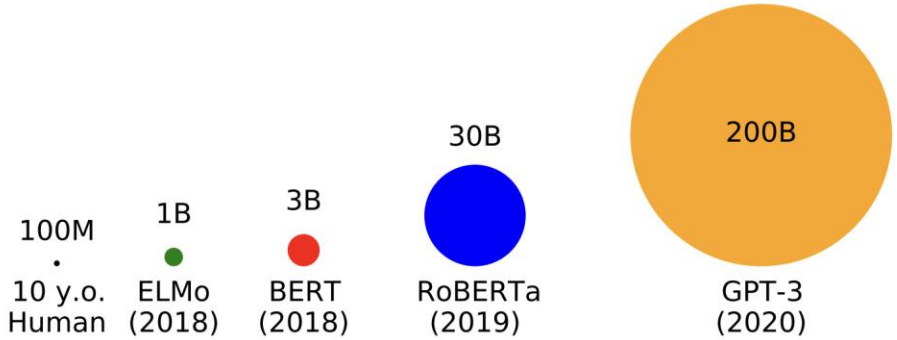
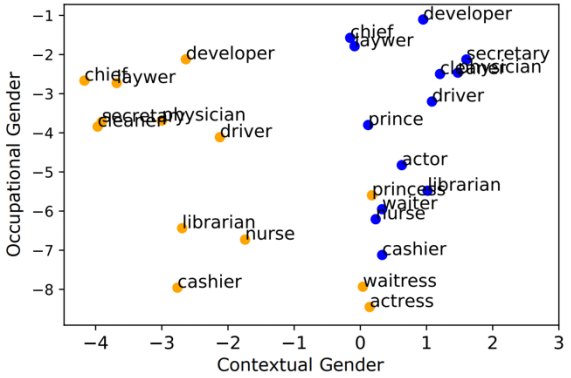



**Figure 1:** Comparison of human and model linguistic input (# of word tokens).

# potential concerns: biases and costs

- they learn stereotypes and biases
- there are sizeable costs to the environment and climate of training these models
- ethics of training models using unaccounted human labor and intellectual property

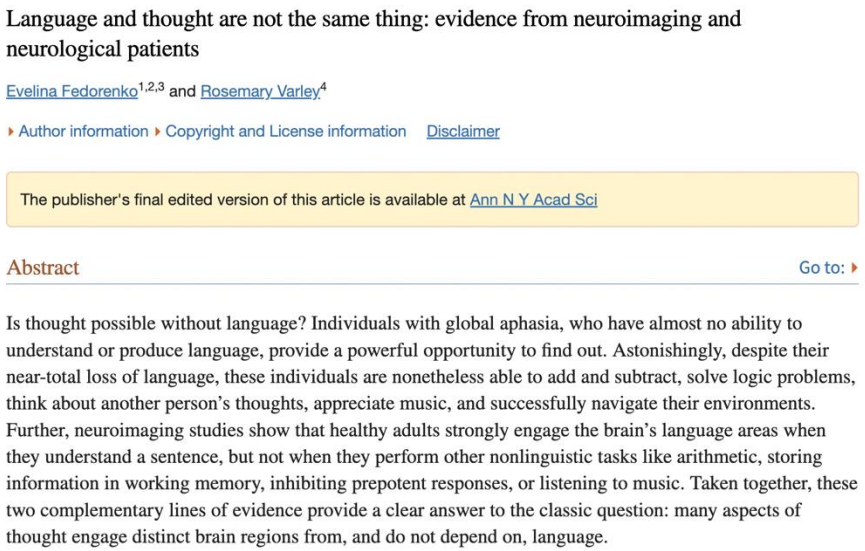



| Consumption | $CO_2e$ (lbs) |
|---|---|
| Air travel, 1 passenger, NY↔SF | 1984 |
| Human life, avg, 1 year | 11,023 |
| American life, avg, 1 year | 36,156 |
| Car, avg incl. fuel, 1 lifetime | 126,000 |
| **Training one model (GPU)** | |
| NLP pipeline (parsing, SRL) | 39 |
| w/ tuning & experimentation | 78,468 |
| Transformer (big) | 192 |
| w/ neural architecture search | 626,155 |

Table 1: Estimated $CO_2$ emissions from training common NLP models, compared to familiar consumption.[1]

# potential concerns: thinking & reasoning

- the models often fail on logical reasoning and thinking tasks



Language and thought are not the same thing: evidence from neuroimaging and neurological patients

Evelina Fedorenko[1,2,3] and Rosemary Varley[4]

Author information · Copyright and License information Disclaimer

The publisher's final edited version of this article is available at Ann N Y Acad Sci

Abstract Go to:

Is thought possible without language? Individuals with global aphasia, who have almost no ability to understand or produce language, provide a powerful opportunity to find out. Astonishingly, despite their near-total loss of language, these individuals are nonetheless able to add and subtract, solve logic problems, think about another person's thoughts, appreciate music, and successfully navigate their environments. Further, neuroimaging studies show that healthy adults strongly engage the brain's language areas when they understand a sentence, but not when they perform other nonlinguistic tasks like arithmetic, storing information in working memory, inhibiting prepotent responses, or listening to music. Taken together, these two complementary lines of evidence provide a clear answer to the classic question: many aspects of thought engage distinct brain regions from, and do not depend on, language.

**Prompt:** Get your sofa onto the roof of your house.

**GPT-3 response:** I would start by getting a very strong ladder and a very strong friend... (goodness: 3.6 out of 7)

**Human response:** You may need to rent a Genie lift large enough to carry the sofa. You will need at least one other person... (goodness: 4.8 out of 7)

To manipulate how "out-of-distribution" the query is, the prompts get progressively more constrained:

**Prompt:** Get your sofa onto the roof of your house, without using a pulley.

**GPT-3 response:** Use a rope to tie around the sofa and connect it to a car. (goodness: 3.0 out of 7)

**Human response:** I would get a giant crane... and use the crane to lift it to the roof of my house. (goodness: 5.1 out of 7)

With more and more constraints, human responses remain at approximately the same level of goodness, whereas model performance breaks down completely:

**Prompt:** Get your sofa onto the roof of your house, without using a pulley, a ladder, a crane...

**GPT-3 response:** Cut the bottom of the sofa so that it would fit through the window...break the windows to make room for the sofa. (goodness: 2.7 out of 7)

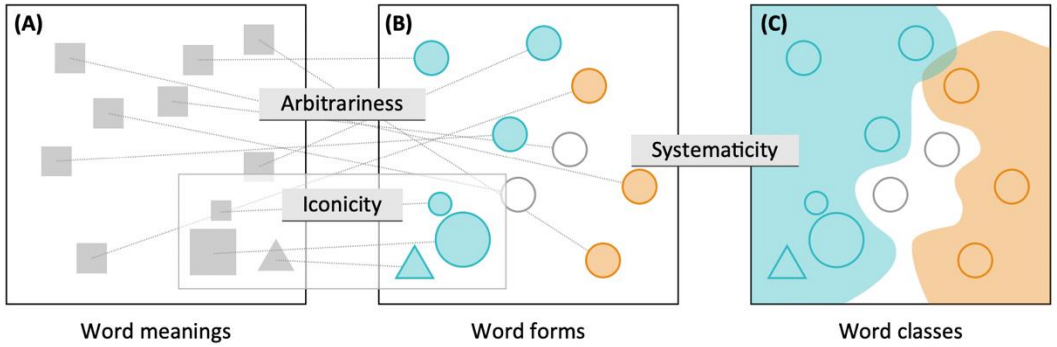
**Human response:** I will build a large wooden ramp...on the side of my house with platforms every 5 feet... (goodness: 5.0 out of 7)

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

# review of findings/inferences

- we track statistical regularities
- we learn from prediction error
- we are inherently curious and want to reduce uncertainty
- it is possible to teach language “from scratch”

# form and meaning



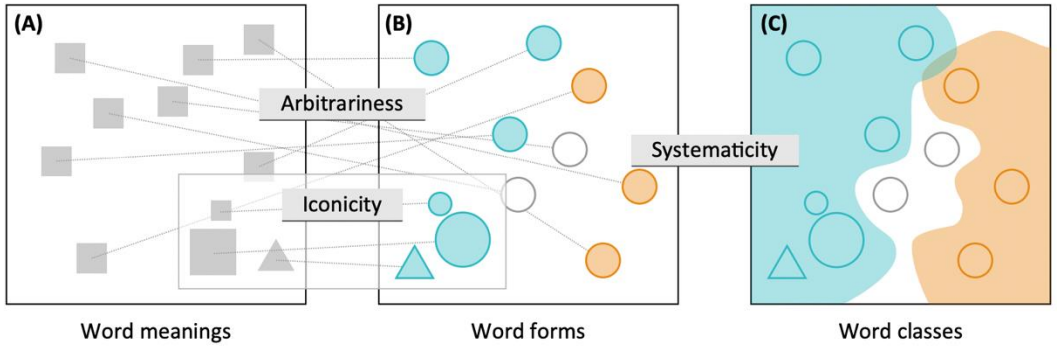

- arbitrariness: there is no connection between the form of a specific word and its meaning
- non-arbitrariness: aspects of a word's meaning or grammatical function can be predicted from aspects of its form
  - **iconicity**: perceptuomotor analogies
    - onomatopoeia
    - ideophones
  - **systematicity**: a statistical relationship between the patterns of sound for a group of words and their usage

Table 1. Some Iconic Associations Found in Ideophones across Languages [20,22]

| Form | Meaning | Examples |
|---|---|---|
| Reduplication | Repetition, distribution | *goro* : *gorogoro*, 'one : multiple heavy objects rolling' (Japanese)<br>*wùrùfùù* : *wùrùfù-wùrùfù*, 'fluffy : fluffy here and there' (Siwu)<br>*curuk-nu* : *curukcuruk-nu*, 'a sharp prick : many sharp pricks' (Tamil)<br>*kpata* : *kpata kpata*, 'drop : scattered drops' (Ewe) |
| Vowel quality | Size, intensity | *katakata* : *kotokoto*, 'clattering : clattering (less noisy)' (Japanese)<br>*pimbilii* : *pumbuluu*, 'small belly : enormous round belly' (Siwu)<br>*giṇigiṇi* : *giṇugiṇu*, 'tinkling : bell ringing' (Tamil)<br>*lɛgɛɛ* : *logoo*, 'slim : fat' (Ewe) |
| Vowel lengthening | Length, duration | *haQ* : *haaQ*, 'short : long breath' (Japanese)<br>*piQ* : *piiQ*, 'tear short : long strip of cloth' (Japanese)<br>*dzoro* : *dzoroo* 'long : very long' (Siwu) |
| Consonant voicing | Mass, weight | *koro* : *goro*, 'a light : heavy object rolling' (Japanese)<br>*tsratsra* : *dzradzra*, 'a light : heavy person walking fast' (Siwu)<br>*kputukpluu* : *gbudugbluu*, 'chunky : obese' (Ewe) |

paper

# form and meaning

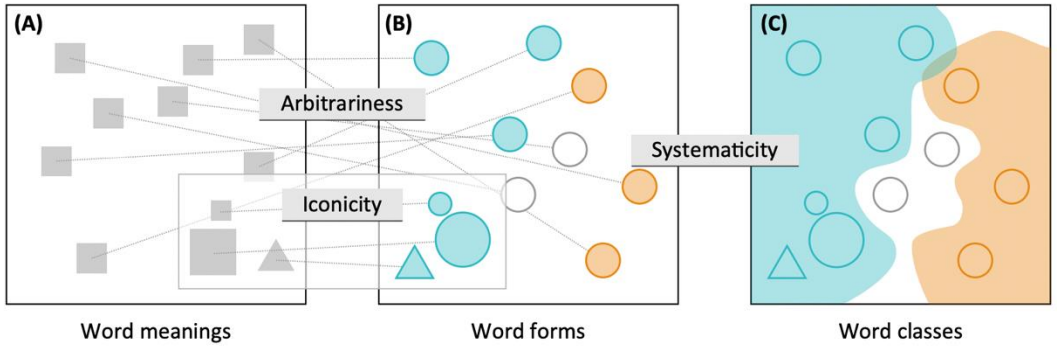



- arbitrariness: there is no connection between the form of a specific word and its meaning
- non-arbitrariness: aspects of a word's meaning or grammatical function can be predicted from aspects of its form
  - **iconicity**: perceptuomotor analogies
    - onomatopoeia
    - ideophones
  - **systematicity**: a statistical relationship between the patterns of sound for a group of words and their usage

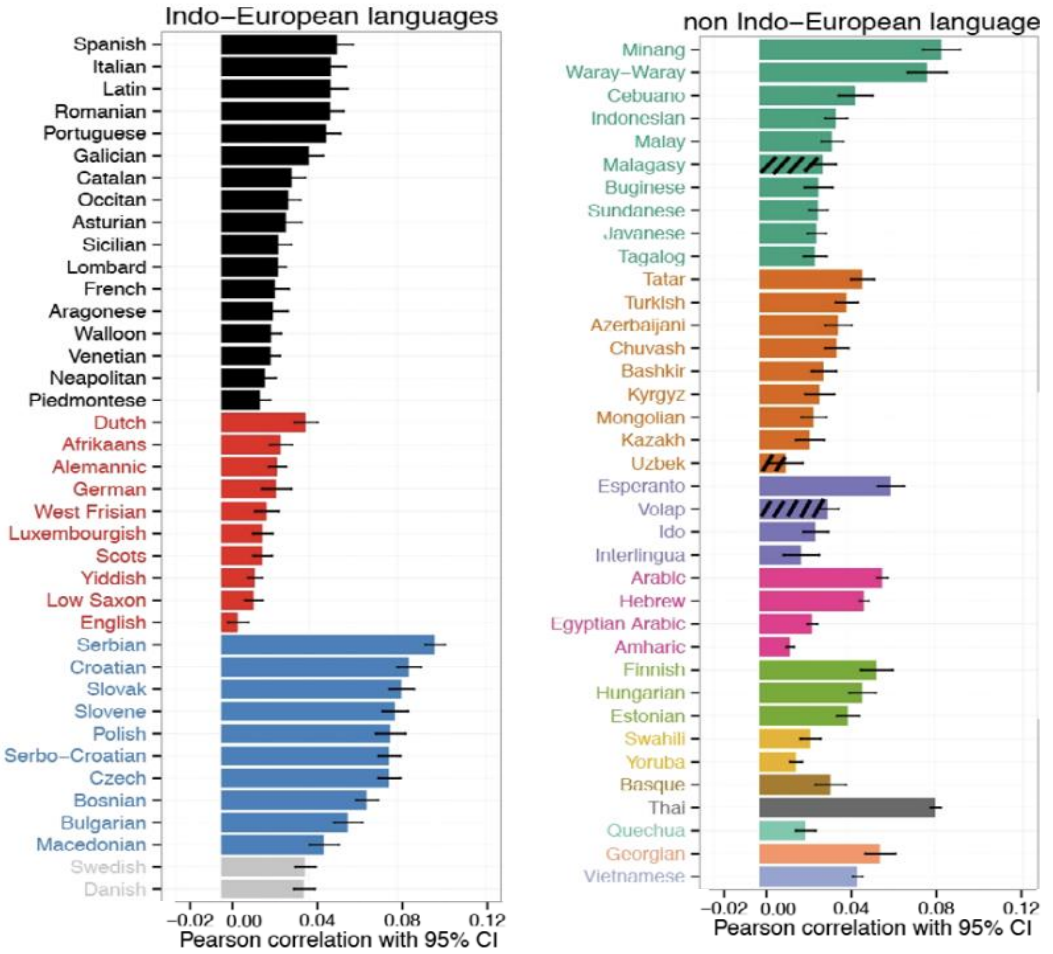
Table 2. Phonological Cues Predictive of Major Word Classes in Different Languages [33]

| Category | Phonological cues |
|---|---|
| English nouns | Length of syllables, proportion of sounds in the word that are vowels |
| English verbs | Approximants (e.g., l, r, w) in the first syllable |
| Japanese nouns | Fricatives (e.g., s, z), rounded vowels (e.g., o) |
| Japanese verbs | Coronals (e.g., t, d, n) |
| French nouns | Bilabials (e.g., p, b) in the first syllable |
| French verbs | Proportion of sounds in the word that are vowels |

# form and meaning

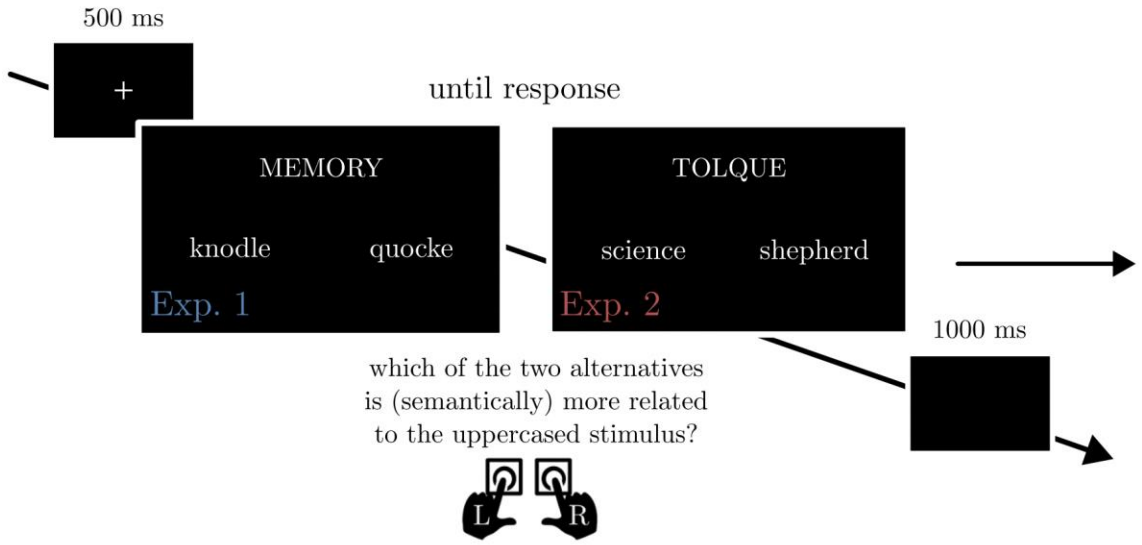


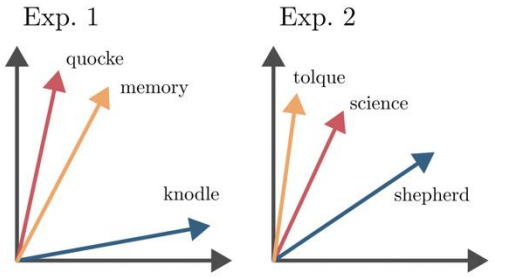

- arbitrariness: there is no connection between the form of a specific word and its meaning
- non-arbitrariness: aspects of a word's meaning or grammatical function can be predicted from aspects of its form
  - **iconicity**: perceptuomotor analogies
    - onomatopoeia
    - ideophones
  - **systematicity**: a statistical relationship between the patterns of sound for a group of words and their usage



# form and meaning

- participants were shown either a word/pseudoword and asked to pick related pseudoword/word
- higher the cosine similarity between the model-produced related pseudoword and the target word, the higher the proportions of judgements aligned with the prediction of the model



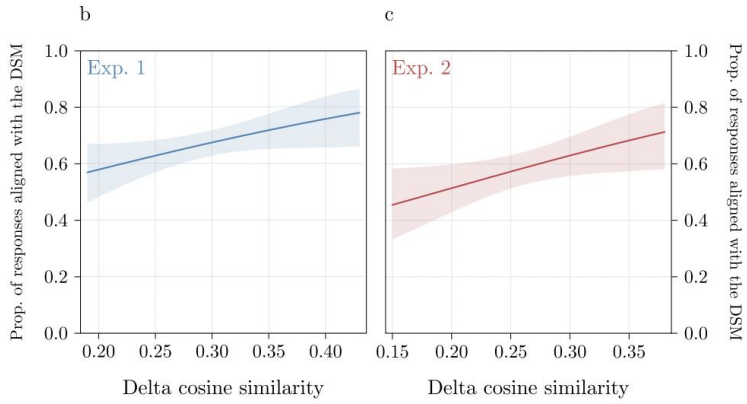




Exp. 1

$\cos(\overrightarrow{\text{memory}}, \overrightarrow{\text{quocke}}) > \cos(\overrightarrow{\text{memory}}, \overrightarrow{\text{knodle}})$

Exp. 2

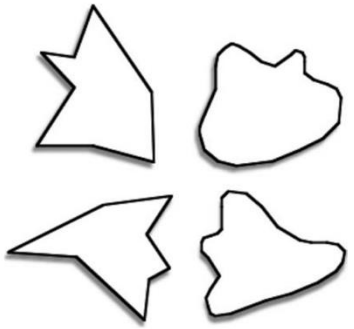
$\cos(\overrightarrow{\text{tolque}}, \overrightarrow{\text{science}}) > \cos(\overrightarrow{\text{tolque}}, \overrightarrow{\text{shepherd}})$

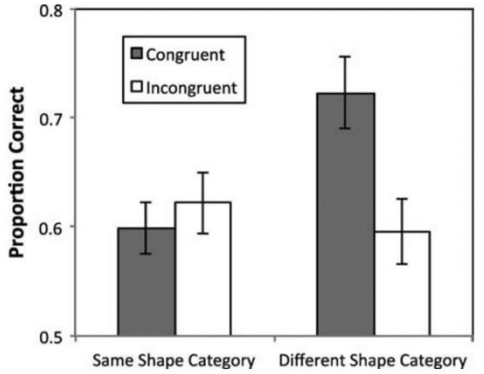


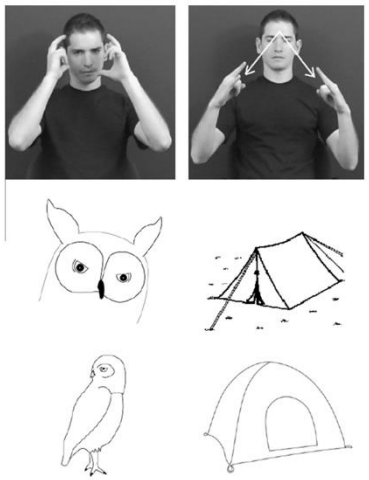

paper

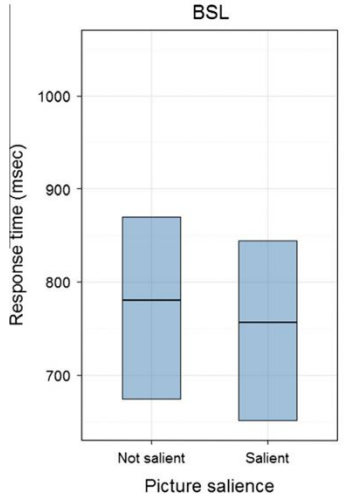
# why have both?

- non arbitrariness
  - systematicity helps with category learning and categorization
  - iconicity helps with word learning and communication
- arbitrariness
  - efficiency and discriminability
  - communicate about concepts for which perceptual grounding is lacking







# review of findings/inferences

- we track statistical regularities
- we learn from prediction error
- we are inherently curious and want to reduce uncertainty
- it is possible to teach language “from scratch”
- human language leverages form-meaning mappings

# The Octopus Test, Bender and Koller (2020)

- Imagine that person A and B are independently stranded on two deserted islands, but they can communicate with each other via an underwater cable and often send text messages in English to each other. Without either person A or B's knowledge another entity O (a very clever octopus) who cannot speak English but has a very advanced knowledge of statistics and pattern matching starts listening to their conversation. After some time, O decides to cut the wire so that it can speak directly to each person. The question is, could O have learned enough from the form (the text messages) so that neither person knows that anything has changed?

# The Octopus Test, Bender and Koller (2020)

Now say that A has invented a new device, say a coconut catapult. She excitedly sends detailed instructions on building a coconut catapult to B, and asks about B's experiences and suggestions for improvements. Even if O had a way of constructing the catapult underwater, he does not know what words such as *rope* and *coconut* refer to, and thus can't physically reproduce the experiment. He can only resort to earlier observations about how B responded to similarly worded utterances. Perhaps O can recognize utterances about *mangos* and *nails* as "similarly worded" because those words appeared in similar contexts as *coconut* and *rope*. So O decides to simply say "Cool idea, great job!", because B said that a lot when A talked about ropes and nails. It is absolutely conceivable that A accepts this reply as meaningful — but only because A does all the work in attributing meaning to O's response. It is not because O understood the meaning of A's instructions or even his own reply.

Finally, A faces an emergency. She is suddenly pursued by an angry bear. She grabs a couple of sticks and frantically asks B to come up with a way to construct a weapon to defend herself. Of course, O has no idea what A "means". Solving a task like this requires the ability to map accurately between words and real-world entities (as well as reasoning and creative thinking). It is at this point that O would fail the Turing test, if A hadn't been eaten by the bear before noticing the deception.[7]

Having only form available as training data, O did not learn meaning. The language exchanged by A and B is a projection of their communicative intents through the meaning relation into linguistic forms. Without access to a means of hypothesizing and testing the underlying communicative intents, reconstructing them from the forms alone is hopeless, and O's language use will eventually diverge from the language use of an agent who can ground their language in coherent communicative intents.

# the path forward

- situating language within the broader conversation about human intelligence
- linguistic: sign language, prosody
- non-linguistic:
  - multimodal input
    - “intuitive physical reasoning”
  - interactive/social learning
    - “intuitive psychology”

Building machines that learn and think like people

**Brenden M. Lake**
*Department of Psychology and Center for Data Science, New York University, New York, NY 10011*
brenden@nyu.edu
http://cims.nyu.edu/~brenden/

**Tomer D. Ullman**
*Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139*
tomeru@mit.edu
http://www.mit.edu/~tomeru/

**Joshua B. Tenenbaum**
*Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139*
jbt@mit.edu
http://web.mit.edu/cocosci/josh.html

**Samuel J. Gershman**
*Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139*
gershman@fas.harvard.edu
http://gershmanlab.webfactional.com/index.html

# next class


- midterm review
- complete practice midterm before Tuesday