information structure and learning:
the artificiality of grammar

michael ramscar

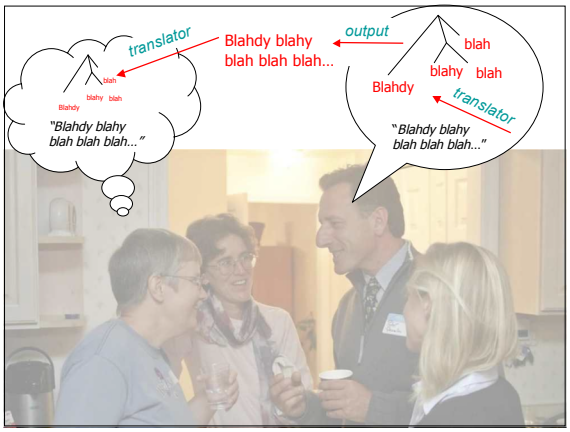
thanks to

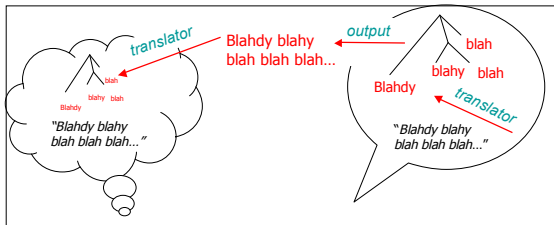
melody dye
dan yarlett
richard futrell
inbal arnon
dan jurafsky

ramscar lab

how language work?

one approach…





language as a formal code

a program

how are our thoughts encoded?

how do we learn the rules by which we turn out thoughts into a common code that allows them to be exchanged?

how do we get meaning out of the code?

is it learnable?

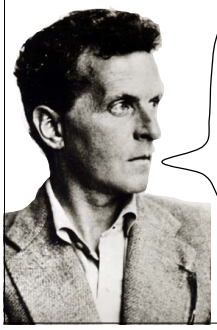
are symbolic or sub-symbolic models of this process more accurate?

form vs function?

how best to study the system

should we focus on the formal system, or the function of its components?

functions?

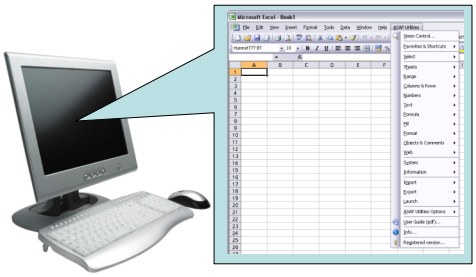


'We talk of process and states, and leave their nature undecided. Sometime perhaps we will know more about them - we think. But that is just what commits us to a particular way of looking at the matter.

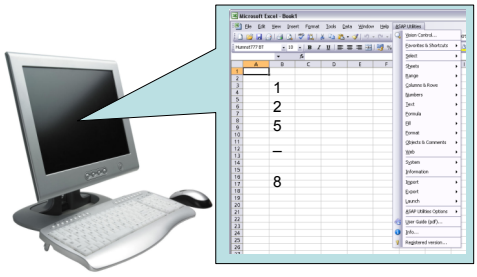
l Wittgenstein, *philosophical investigations*

a metaphor?

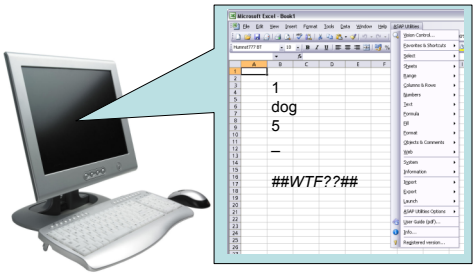
a metaphor?



a metaphor?



a metaphor?



two metaphors?

two metaphors?

communication & learning

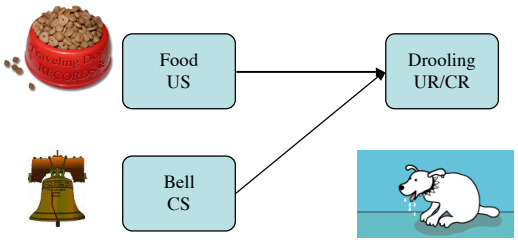
formally = *prediction* and *discrimination*

- a different program
  - speakers co-operate with listeners to help them discriminate the content of messages from possible alternatives
  - try to characterize what it is about learners and their interactions with the environment (including others) that enables them to predict and understand others

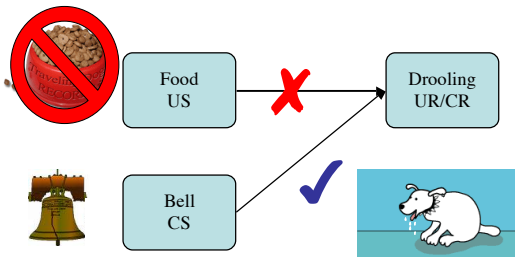
learners

- what is learning?
  - formal learning theory treats learning as a process through which information about the cues that allow environmental regularities to be *predicted* is acquired

classical conditioning



classical conditioning



how does it work?

- learning is driven by discrepancies between what is expected and what actually occurs.
- discrepancies in prediction cause:
  - value of predictive cues to be strengthened when outcomes are underpredicted,
  - weakened when overpredicted

the importance of being wrong

value of predictive cues is strengthened when outcomes are underpredicted

- most philosophers, linguists, psychologists etc., get this part:
  - A **occurs** after B, but was not fully anticipated:

    if wrong, **increase the predictive value** of B

value of predictive cues is weakened when outcomes are overpredicted

- most philosophers, linguists, psychologists etc., **don't** get this:
  - if A **does not** occur after B, but **was** anticipated

    if wrong, **reduce** the predictive value of B

"being wrong"

in error driven learning, **violation of expectation** is a powerful source of **negative evidence**

animals learn this way

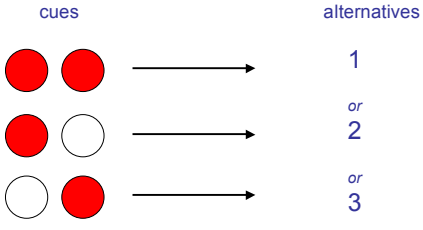
maybe people do as well?
your date doesn't show up
or a punch line

is human learning error-driven?

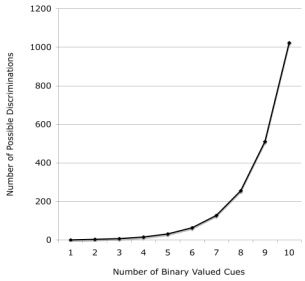
how to tell?

- cues allow predictions to be learned and made
- the amount of *discriminative information* that can be *positively* encoded in a set of S cues with V values is $(V^s)-1$

2 binary valued cues allow 3 positive outcomes to be discriminated between



amount of discrimination that can be *positively* encoded in a set of x cues is $(2^x)-1$



discrimination and coding

if one set of bits a is used to specify the what is or can be encoded in another set of bits b, a must have sufficient bits to encode b.

something about the structure of information in our world

the world is complicated

speech sounds are less so
- hard to decompose
- we can easily discriminate **between** the units that sound symbols are made up of, but not **within** them

words ("symbols") appear to comprise a vastly smaller set of bits than their possible meanings…

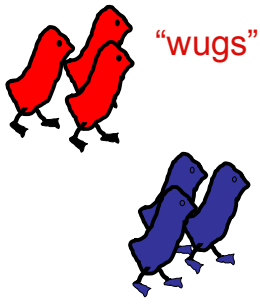
a window into the nature of symbolic learning?

if this analysis is correct, we can use our understanding of learning and information structure to analyze human learning…

…and make predictions

does the set size principle apply to symbolic coding and symbolic learning?

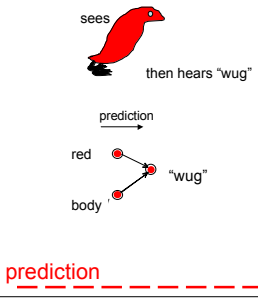
words and prediction



"wugs"

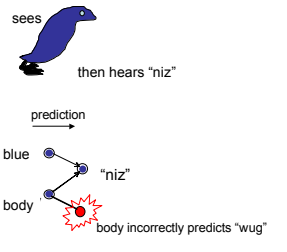
"nizes"

words and prediction



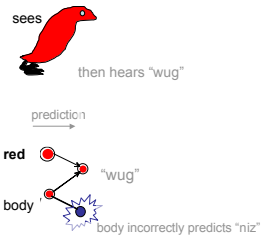
prediction

words and prediction



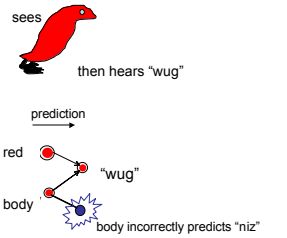
prediction

words and prediction



prediction

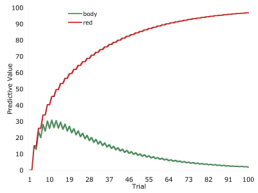
words and prediction



prediction

prediction error and discrimination learning

- overprediction will cause "body" to lose value when in competition with "red" and "blue"



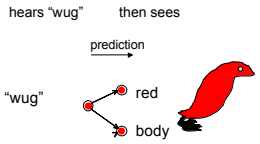
despite the misleading name, "associative models" work because they **dissociate** weak cues and not just because they associate cues with events

-- they are **discrimination** networks

symbols and set size

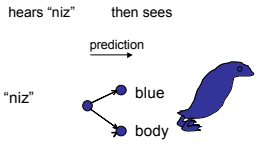
- if words are not complex cues, we can predict something interesting…

words and prediction



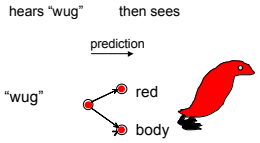
prediction

words and prediction



prediction

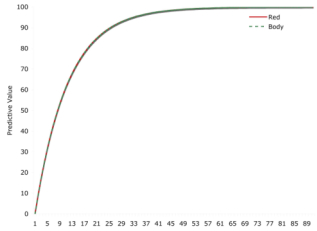
words and prediction



prediction

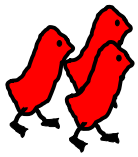
no cue competition = no discrimination

- without cue competition , learning simply tracks co-occurrences between the cues and the labels

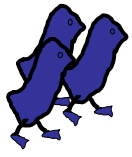


discrimination is important

75%



"wugs"

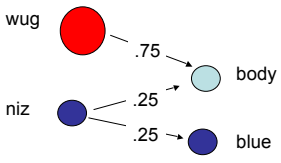


"nizes"

25%

what is this?

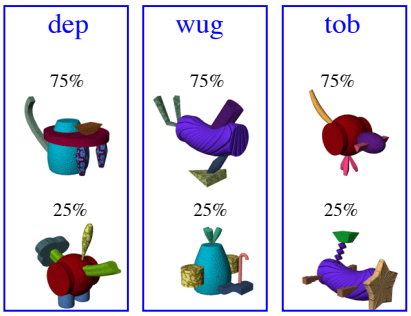
?



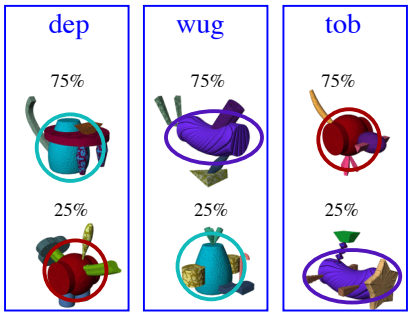
co-occurrences can mislead…

exploring this idea

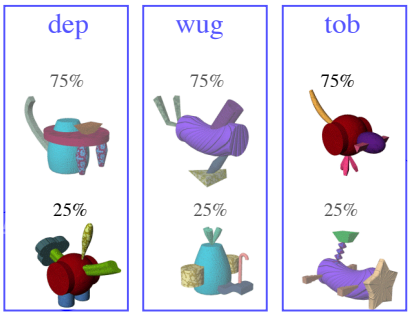
fribble categories



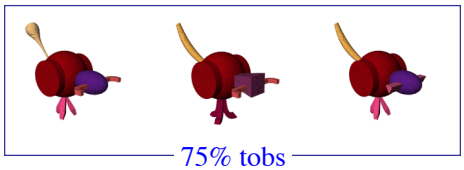
fribble categories

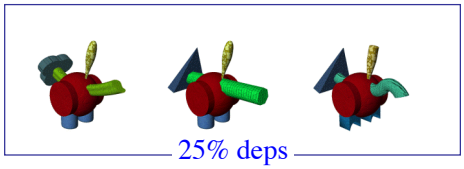


fribble categories

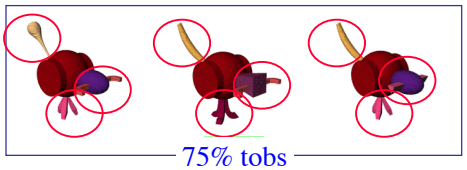


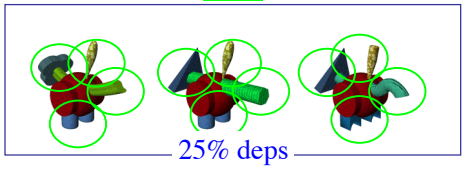
fribble categories


75% tobs

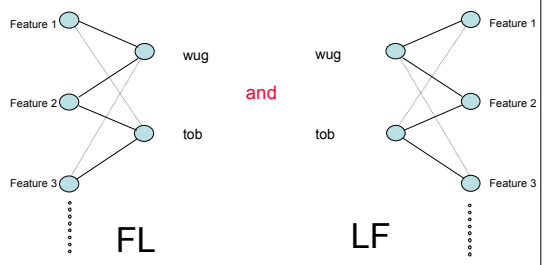

25% deps

fribble categories

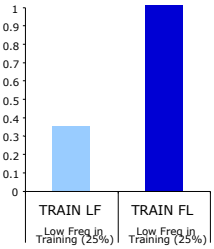

75% tobs


25% deps

2 learning models



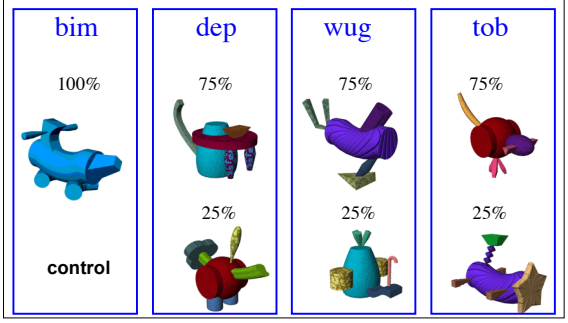
how well did the models discriminate the low frequency exemplars?



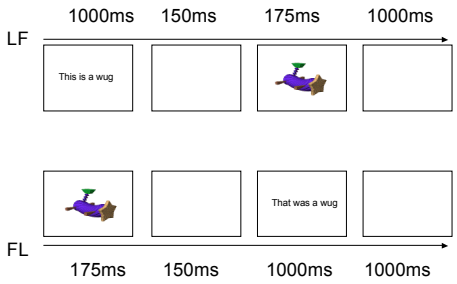
in the models, different representations are learned depending on the structure of the environment

do people do this?

fribble categories



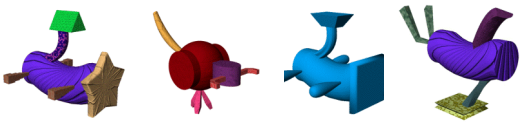
training



test 1

bim dep wug tob

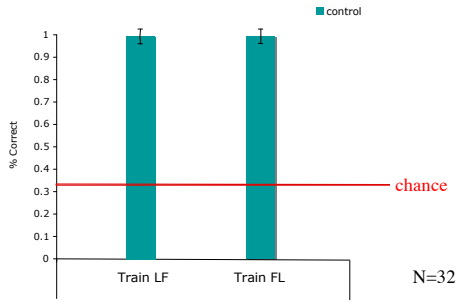
test 2



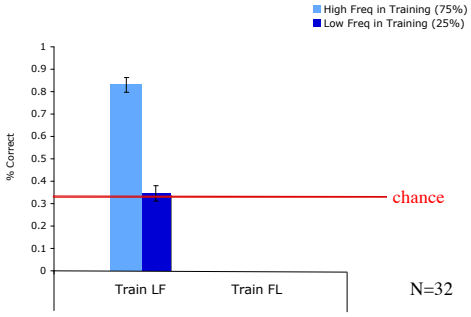
tob

what did people do?

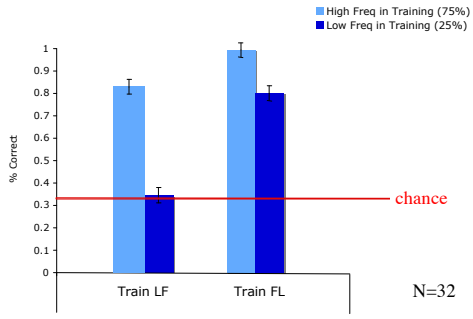
data



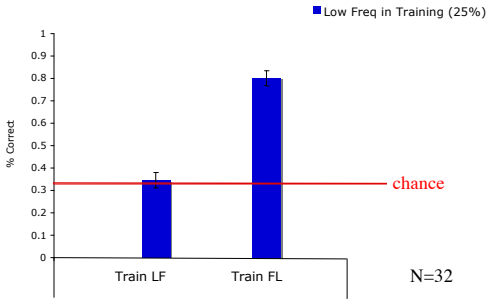
data



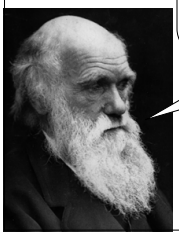
data



discrimination



what about “real” meaning, “in the wild?



I carefully followed the mental development of my small children, and was astonished to learn that soon after they had reached the age where they knew the names of all ordinary things, they appeared to be entirely incapable of giving the right name to colors… I remember quite clearly to have stated they were color-blind

Charles Darwin, 1877

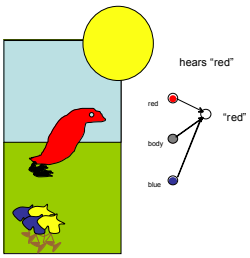
the problem

- kids learn color words late in english
  - may struggle to learn even after 1000s of training trials as late as 4
- “know” the words
  - red, yellow etc in vocabularies early
- “know” something about them
  - use them in correct part of speech
- “don’t know” the words
  - use of individual color words haphazard and interchangeable
- behavior same as blind children (gleitman & landau)…

the problem

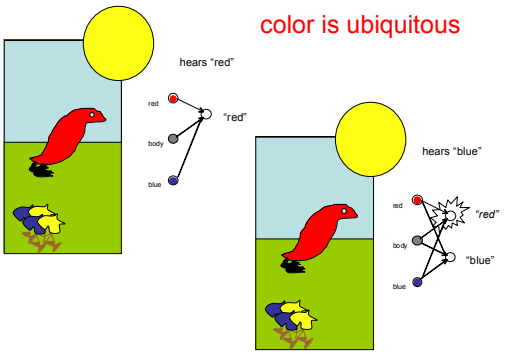
- why are color words hard to learn?
  - color is everywhere
  - surely kids will get lots of training
- is ubiquity the problem?
  - color is everywhere
  - what kind of *information* is their in the environment?

color is ubiquitous



(i) time

color is ubiquitous

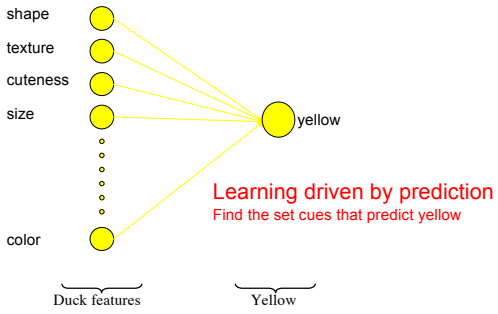


(i) time (ii) …

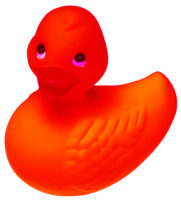
so how might children learn color adjectives?

"look at the duck… it's yellow…"

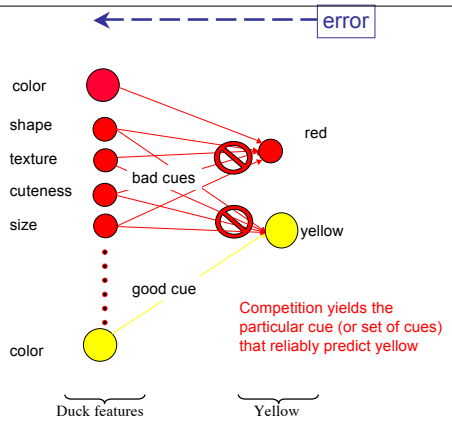
how might color adjective learning work?



how might color adjective learning work?



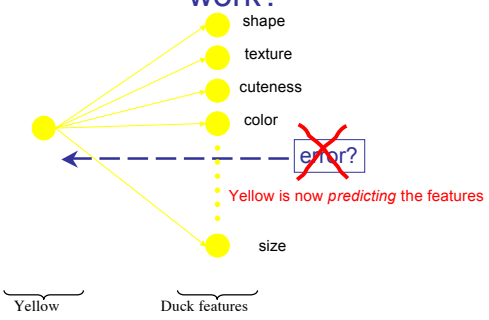
"look at the duck… it's red…"



a problem with english

"look at the yellow duck…"

how might color adjective learning work?

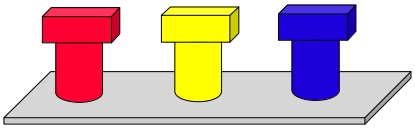


an experiment

training

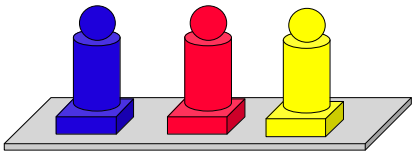
- test kids at 24 - 26 months (N=41)
  - little evidence of consistent color comprehension at this age
- 3 colors, red, blue & yellow
- test - train - test design
  - expect kids to know something: can we reinforce what they already know?

test



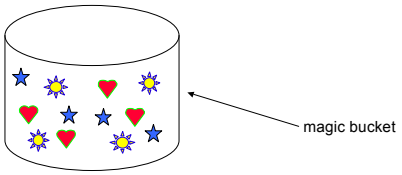
"show me the red one"

test



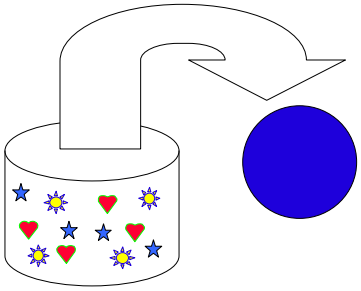
"show me the one that is blue"

train



half of the children

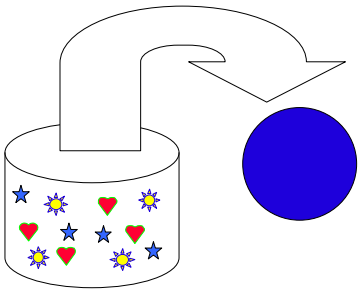
train



"this is a blue ball"

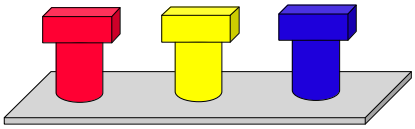
other half of the children

train



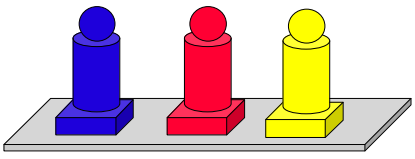
"this ball is blue"

test



"show me the red one"

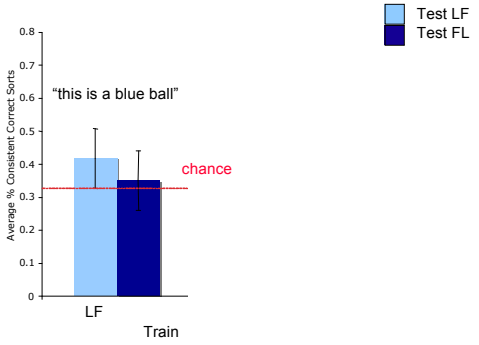
test



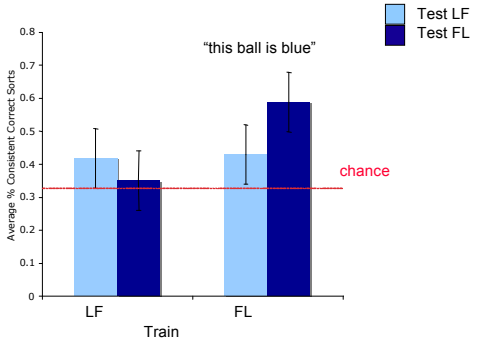
"show me the one that is blue"

results

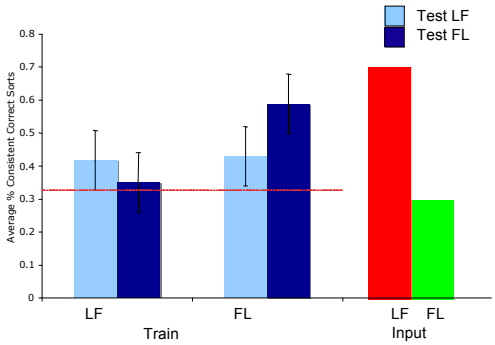
consistent color sorting



consistent color sorting



consistent color sorting



what about other words?

- feature label ordering can offer interesting insights into how concrete nouns and adjectives are learned…

  …but there is more to language than that

- can feature label ordering can help us understand how the more abstract parts of language are learned?

grammatical gender

- in english we uniformly (?) apply a single determiner to all nouns

  the chair, the dog

- in other languages, things are more complicated

  la chaise, le chien

grammatical gender

- system found in many languages
  - assigns all nouns (including inanimate ones) to noun classes, and marks neighboring words for agreement
    - In Hebrew, for example, verbs and adjectives are marked for gender.
    - in Spanish and French, articles have to agree in gender with the nouns they precede.
    - knowing a nouns' gender in gender-marking languages is essential to correct sentence construction

learning grammatical gender

- native speakers
  - rapid mastery by children
  - children and adults can use gender information to guide lexical access
- adult l2 learners
  - persistent difficulty even after extensive exposure
  - do not use gender information to guide lexical access

why?

- L2 learners have learned to segment a first language…

  … unlike children
- maybe children start off from larger units
- how might this affect the learning grammatical gender?

blocking

1a. *AB* ⇨ *X*

1b. *A* ⇨ *X*

1c. *B* ⇨ *?*

blocking

- reliable effect in animal learning

why might this harm l2 learners?

1.  ⇨ *…/ la mano /...*

1.  ⇨ */ mano /*

child

why might this harm l2 learners?

1.  ⇨ *…/ la mano /...*

1.  ⇨ */ mano /*

child

2.  ⇨ */ mano /*

2.  ⇨ *…/ la mano /*

adult l2 learner

teaching grammatical gender in an artificial language

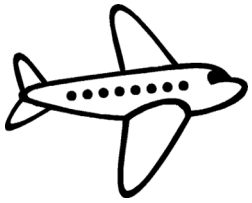
- two conditions:
  - sequence-first condition: whole sequences first and then noun labels
  - label-first condition: noun labels first and then whole sequences
- identical frequency-of-exposure

training

- participants saw a picture and heard speech
  - Each noun label 5 times
  - Each determiner-noun sequence 5 times

noun label only

hear: "slindot"



noun label only

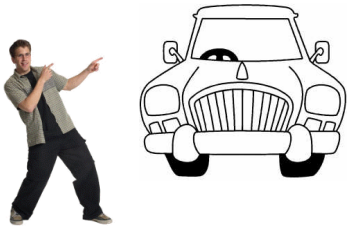
hear: "viltord"

full sequence

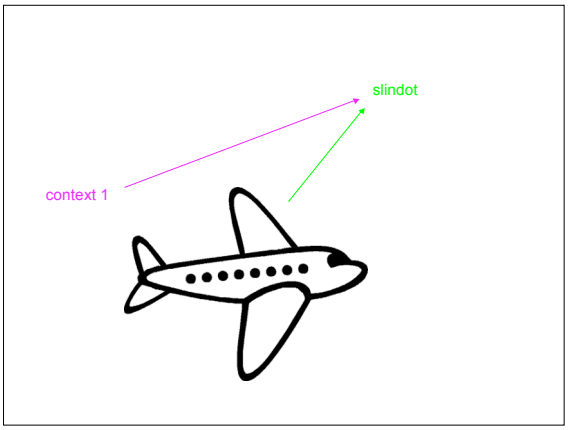
Hear: "Os ferpel een bol slindot"

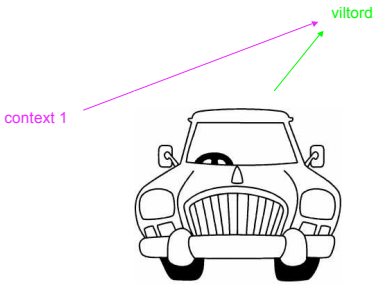
full sequence

Hear: "Os ferpel een oos viltord"

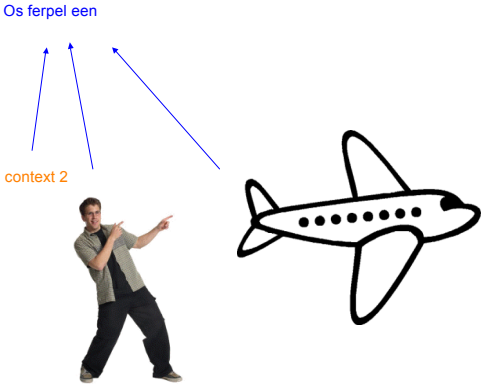


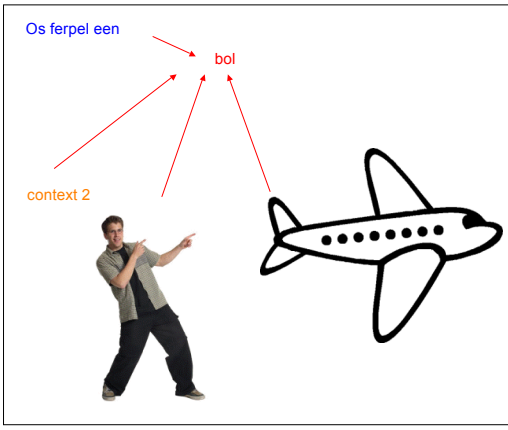
information structure for noun only

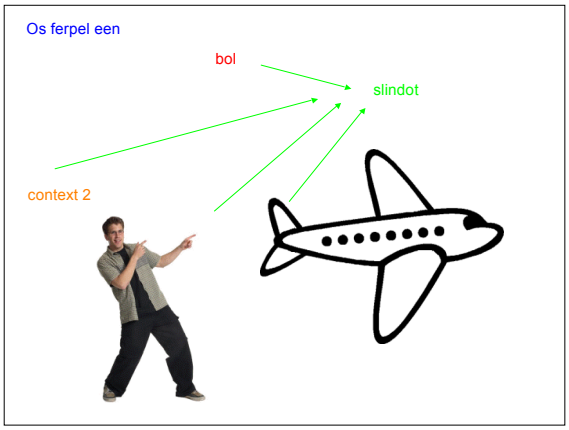


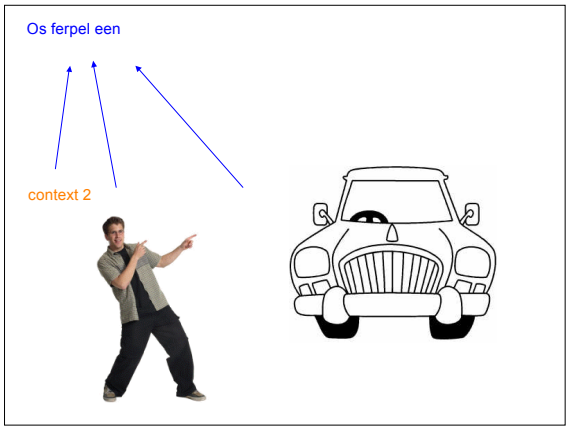


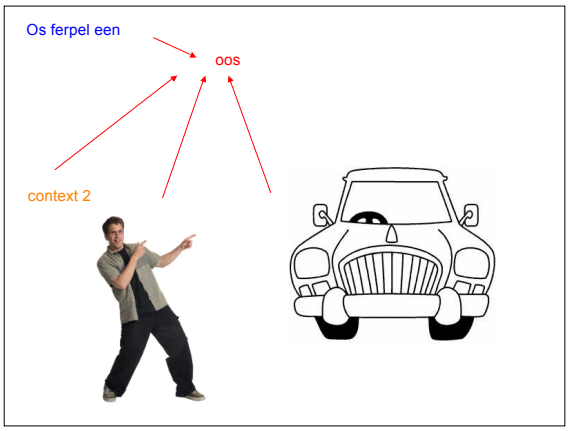
information structure for sequence

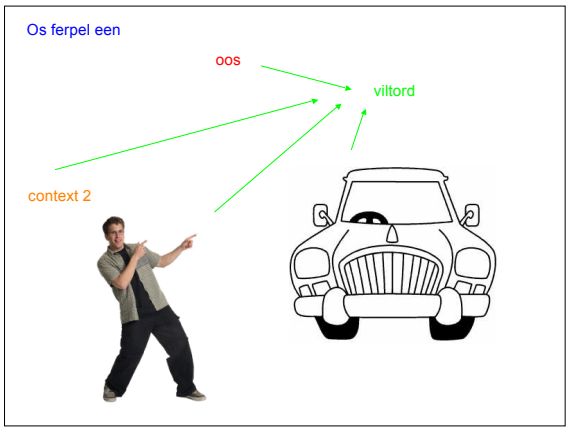


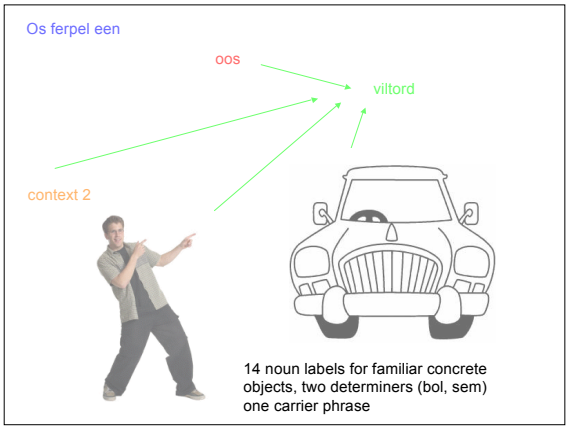












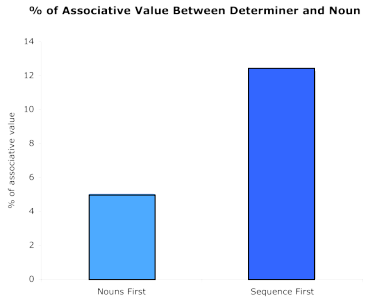
model training

only difference: order of exposure

- sequence-first condition
  1. block of whole sentences
  2. block of noun labels

- label-first condition
  1. block of noun labels
  2. block of whole sentences

modeling result



human training

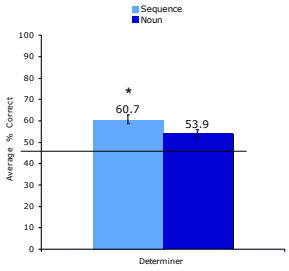
only difference: order of exposure

- sequence-first condition
  1. block of whole sentences
  2. block of noun labels
- label-first condition
  1. block of noun labels
  2. block of whole sentences

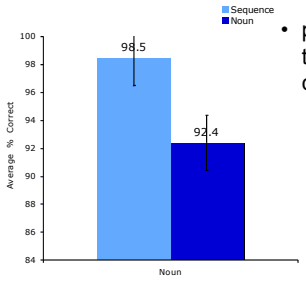
testing

- forced-choice: see a picture, hear two descriptions and choose the correct one
  - determiner trials: incorrect sentence had right label but wrong determiner
  - noun trials: incorrect sentence had right determiner but wrong label
- production: see a picture, produce a full sentence

determiner forced-choice results

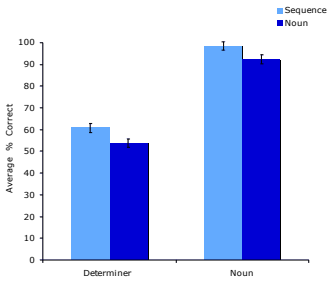


noun forced-choice results



- participants better in the sequence-first condition!

forced-choice results



wait?

- gendered determiners help?
- but…
  – isn't gender just silly, pointless stuff?

" In German… every noun has a gender, and there is no sense or system in the distribution; so the gender of each must be learned separately and by heart. There is no other way. To do this one has to have a memory like a memorandum-book. In German, a young lady has no sex, while a turnip has. Think what overwrought reverence that shows for the turnip, and what callous disrespect for the girl…:

Gretchen: Wilhelm, where is the turnip?
Wilhelm: She has gone to the kitchen.
Gretchen: Where is the accomplished and beautiful English maiden?
Wilhelm: It has gone to the opera."

Mark Twain, (1880) "The Awful German Language"

"The presence of such systems [German gender] in a human cognitive system constitutes by itself excellent testimony to the occasional nonsensibleness of the species. Not only was this system devised by humans, but generation after generation of children peaceably relearns it."

Michael Maratsos (1979)

german gender & entropy reduction

(1) Yesterday I *!!* visited the *!* Doctor

nouns are the most frequent POS

german gender & entropy reduction

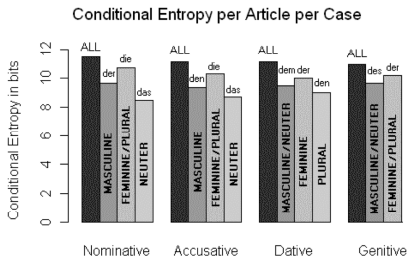
(1) Yesterday I *!!* visited the *!* Doctor

vs

(2) Gestern besuchte ich den *!* Arzt
yesterday visited I the.*MASC* *!* doctor

how to find out?

- examined the NEGRA II corpus of German newspapers (Skut et al. 1997)
- for each case, every noun immediately preceded by a definite article was extracted and counted.
- the entropy of all the nouns in each case was calculated separately, and then the conditional entropy given each type of article was calculated



how might it work?

- frequency: higher frequency nouns more likely to be encountered in sparser contexts
- Which means that information requirements for high and low frequency nouns may be different

how might it work?

- compare the information requirements for helping someone predict that Beethoven and not Mozart will be the topic of a sentence
- versus helping someone predict Villa Lobos rather than Schoenberg (C20th composers) will be the topic of a sentence.
  - if the topic of discussion is either Beethoven or Mozart , frequency and saliency alone will tend to render Villa Lobos and Schoenberg largely irrelevant.
  - if Villa Lobos or Schoenberg were to be the topic of a sentence, a cue corresponding to 20th century classical music would be incredibly informative

So how does it work?

- examined the 512 nouns that occur in the 100 conversations of the spoken callhome corpus.
- a logistic regression model predicted gender sameness for pairs of nouns based on the frequency, mutual information (a measure of how often the two nouns co-occurred with one another, controlling for frequency), and semantic similarity of each pair.

and?

- among noun pairs, overall gender sameness was predicted by two factors:
  - 1) semantic similarity (the more tightly semantically coupled, the more likely the pair was to share a gender), and
  - 2) the frequency of the words in the pairing (the lower their frequency, the more likely the pair was to share a gender).
- while there was no main effect of co-occurrence, co-occurrence did enter into a significant interaction with frequency

in other words…

- for high frequency words, likelihood of co-occurrence tended to predict *gender difference*
- for low frequency words, semantic similarity tended to predict *same gender*

this structure may remind you of something…

what if we had no gender?

- german gender classes serve to make nouns more predictable in context
- what does this mean for english, a germanic language that has largely shed noun class?

a comparison

- compared the average entropy of nouns in the negra corpus studied so far to those in the new york times gigaword corpus of english.
- entropy of english nouns after definite articles is 10.17 bits
- entropy of german nouns after gendered definite articles was 10.55 bits.
  - entropy of german nouns rises to 11.71 bits when calculated with gender information removed.

which means?

- nouns after definite articles are more diverse in german than english
  - compared type/token ratio of lemmas in the samples
    - average frequency of the german lemmas in negra is 2.12,
    - average frequency of a similar noun lemma in the gnglish sample is 4.93

which means?

- german is a more informative langauage than english

which means?

- german is a more informative langauage than english

or

- english makes use of another device t make nouns more predictable

adjectives

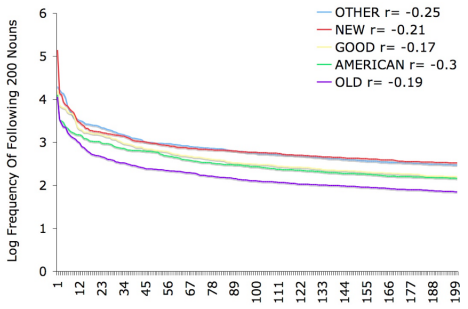
- examined the determiner-adjective-noun sequences in the negra and the nyt corpuses.
  - average frequency of a german adjective in this context was 6.66,
  - average frequency of an english adjective here was 4.08 (
    - average frequency of a german noun in this context is 2.26,
    - average frequency of an english noun is 3.36

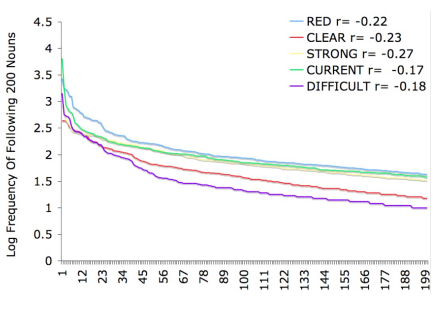
hmmm

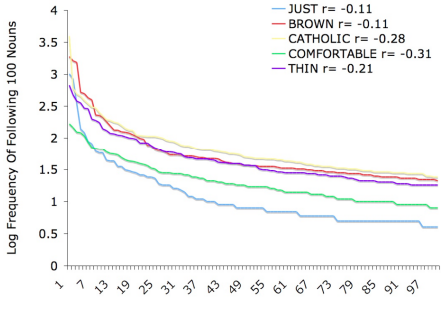
- high-frequency nouns are, by definition, less informative than low-frequency nouns (for instance
  - "doberman" is more informative than "puppy," which is more informative than "dog," see e.g., rosch, 1978).
  - one might reasonably expect adjectives to be applied more to high-frequency nouns, which are less informative and are therefore more in need of semantic augmentation than low-frequency nouns, which tend to be more specific.

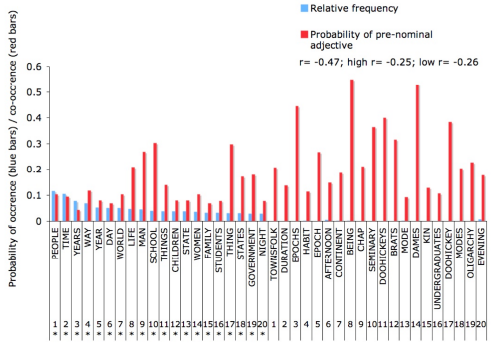
but

- if english entropy reduction is provided by adjectives, we would expect that pre-nominal adjectives would be applied more to low-frequency nouns than high-frequency nouns
- low-frequency nouns convey more information (and therefore benefit more from entropy reduction) than high-frequency nouns.









so

- gender markers harder for adults to learn…
- english genders lost during norse colonization
- german noun class function has shifted to adjectives in english?