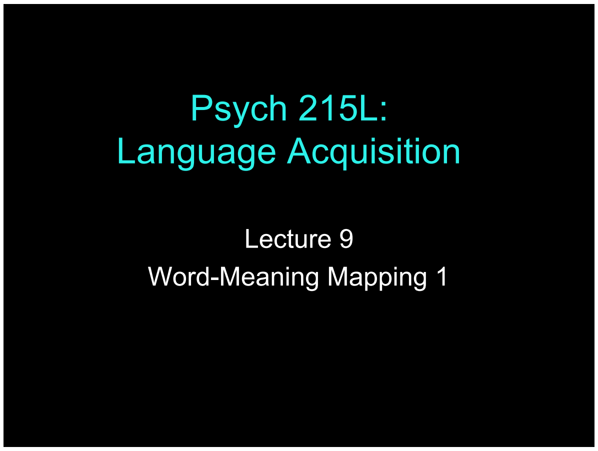
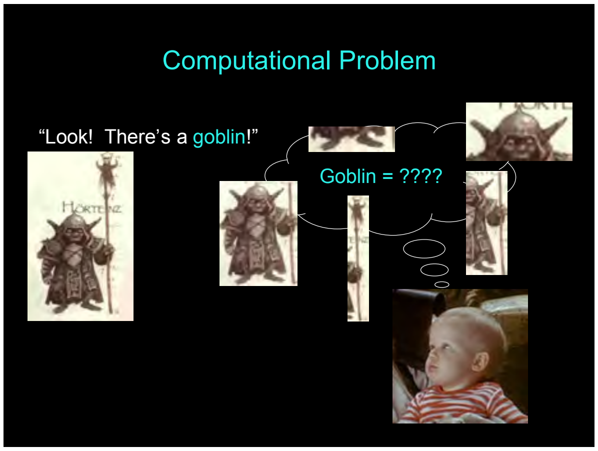


Psych 215L: Language Acquisition

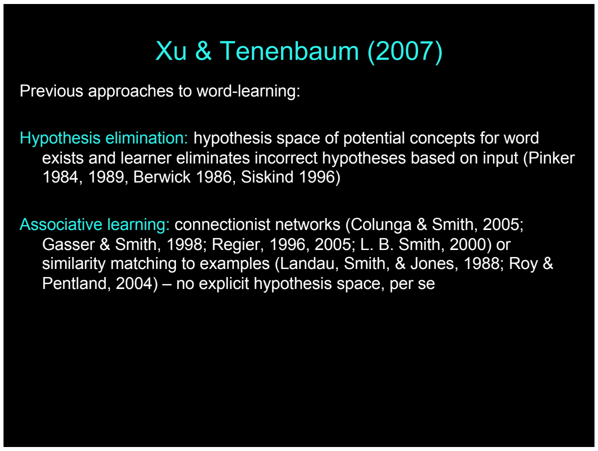
Lecture 9

Word-Meaning Mapping 1



Computational Problem

"Look! There's a goblin!"

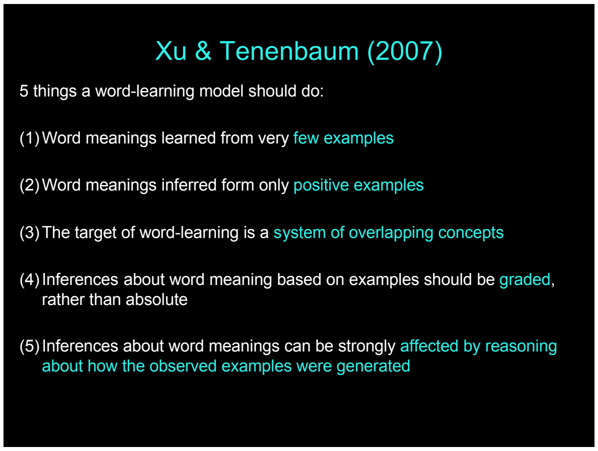


Xu & Tenenbaum (2007)

Previous approaches to word-learning:

Hypothesis elimination: hypothesis space of potential concepts for word exists and learner eliminates incorrect hypotheses based on input (Pinker 1984, 1989, Berwick 1986, Siskind 1996)

Associative learning: connectionist networks (Colunga & Smith, 2005; Gasser & Smith, 1998; Regier, 1996, 2005; L. B. Smith, 2000) or similarity matching to examples (Landau, Smith, & Jones, 1988; Roy & Pentland, 2004) – no explicit hypothesis space, per se



Xu & Tenenbaum (2007)

5 things a word-learning model should do:

(1) Word meanings learned from very few examples

(2) Word meanings inferred form only positive examples

(3) The target of word-learning is a system of overlapping concepts

(4) Inferences about word meaning based on examples should be graded, rather than absolute

(5) Inferences about word meanings can be strongly affected by reasoning about how the observed examples were generated

Xu & Tenenbaum (2007)

Approach to word learning based on rational statistical inference (ideal learner)

Hypothesis about word meanings evaluated by Bayesian probability theory

Claim: “The interaction of Bayesian inference principles with appropriately structured hypothesis spaces can explain the core phenomena listed above. Learners can rationally infer the meanings of words that label multiple overlapping concepts, from just a few positive examples. Inferences from more ambiguous patterns of data lead to more graded and uncertain patterns of generalization. Pragmatic inferences based on communicative context affect generalizations about word meanings by changing the learner’s probabilistic models.”

Ruling out unnatural extensions

dog = dog parts, front half of dog, dog spots, all spotted things, all running things, all dogs + one cat

Traditional Solutions:

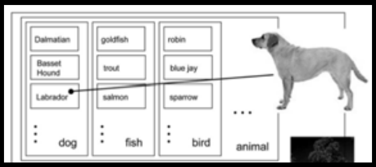
Whole Object constraint: First guess is that a label refers to a whole object, rather than part of the object (*dog parts, front half of dog*) or an attribute of the object (*dog spots*)

Taxonomic constraint (Markman 1989): First guess about an unknown label is that it applies to the taxonomic class (ex: *dog*, instead of *all running things* or *all dogs + one cat*)

The issue of overlapping hypotheses

Object-kind labels: *dog* vs. *dalmatian* vs. *animal*

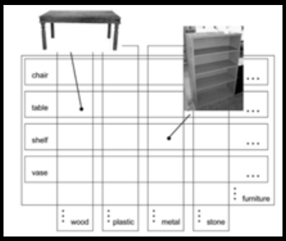
Issue: clearly overlapping labels – a dalmatian is a dog and an animal, but not all animals are dogs, and not all dogs are dalmatians. Which level does each label apply to?



The issue of overlapping hypotheses

Multiple properties potentially relevant: *shape* vs. *material*

Issue: clearly overlapping labels – which aspect is being labeled?



Traditional solutions

For object-kind labeling:

Markman (1989): learners prefer the “basic” level of categorization (*dog* over *dalmatian* or *animal*)

Remaining issue: How do learners figure out non-basic level labels? That is, how do they overcome this bias? Since concepts are overlapping, it’s not enough to learn that *dog* can label a dog. Learners must somehow figure out that *animal* is also a fine label for a dog (and a cat and a bird).

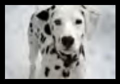
A Bayesian solution

Suspicious Coincidences:

Situation:

fep

fep



fep

fep

Suspicious: Why is no other animal or other kind of dog a *fep* if *fep* can really label any animal or any kind of dog?

Bayesian reasoning: Would expect to see other animals (or dogs) labeled as *fep* if *fep* really could mean those things. If it continues not to be used this way, this is growing support that *fep* cannot mean those things.

The Bayesian Framework

Task: Learn concept associated with word C.

X = $\{x_1, x_2, \ldots, x_n\}$ = *n* observed examples of C

Hypothesis h = pointer to some subset of entities in the domain that could be the extension of C

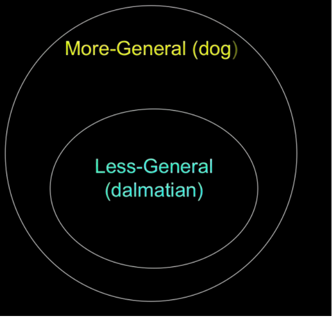
Standard Bayesian inference:
Posterior probability of hypothesis h given examples X is related to the likelihood of h producing X (p(X | h)) and the prior likelihood of h (p(h)). These are normalized against the cumulative likelihood of producing X given any hypothesis (p(X)).

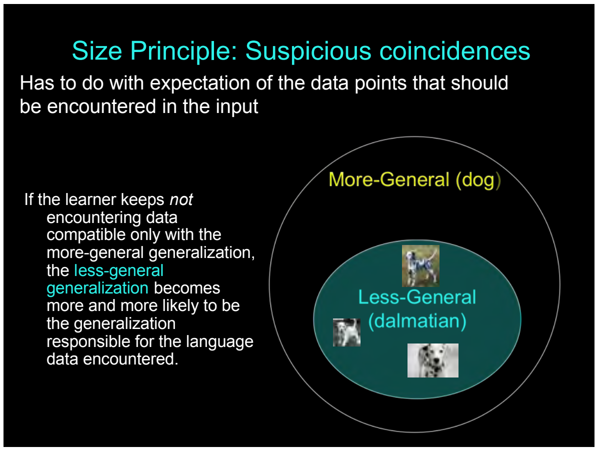
$$p(h \mid X) = \frac{p(X \mid h)p(h)}{p(X)} = \frac{p(X \mid h)p(h)}{\sum_{h' \in H} p(X \mid h')p(h')}$$

Size Principle: Suspicious coincidences

Has to do with expectation of the data points that should be encountered in the input

If more-general generalization (dog) is correct, the learner should encounter some data that can only be accounted for by the more-general generalization (like beagles or poodles). This data would be incompatible with the less-general generalization (dalmatian).





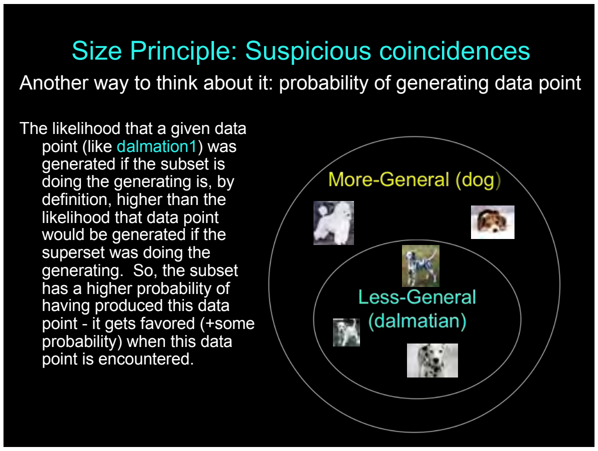
Size Principle: Suspicious coincidences

Has to do with expectation of the data points that should be encountered in the input

If the learner keeps *not* encountering data compatible only with the more-general generalization, the less-general generalization becomes more and more likely to be the generalization responsible for the language data encountered.

More-General (dog)

Less-General (dalmatian)



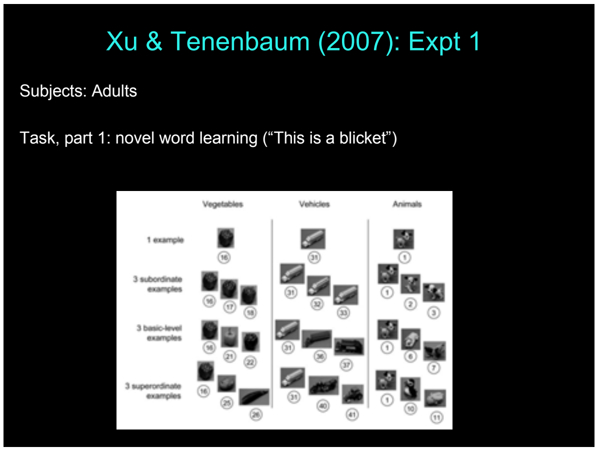
Size Principle: Suspicious coincidences

Another way to think about it: probability of generating data point

The likelihood that a given data point (like dalmation1) was generated if the subset is doing the generating is, by definition, higher than the likelihood that data point would be generated if the superset was doing the generating. So, the subset has a higher probability of having produced this data point - it gets favored (+some probability) when this data point is encountered.

More-General (dog)

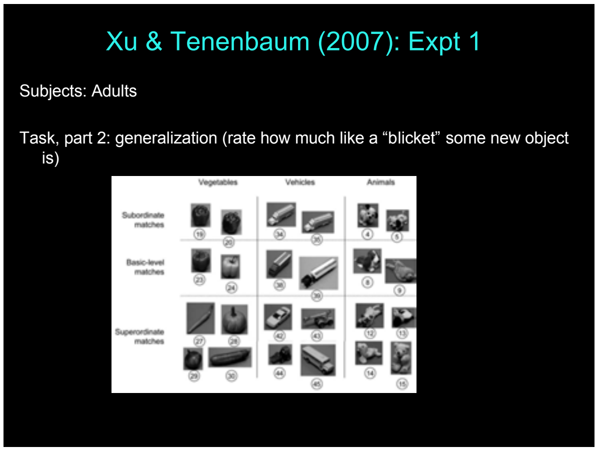
Less-General (dalmatian)



Xu & Tenenbaum (2007): Expt 1

Subjects: Adults

Task, part 1: novel word learning ("This is a blicket")



Xu & Tenenbaum (2007): Expt 1

Subjects: Adults

Task, part 2: generalization (rate how much like a "blicket" some new object is)

Xu & Tenenbaum (2007): Expt 1

Results: Clear differentiation between being shown 1 example vs. being shown 3 examples. When shown 3 examples, adults restricted their hypotheses to the most conservative meaning consistent with the examples. 1 example scenario most like being shown 3 basic-level examples, but has more uncertainty.

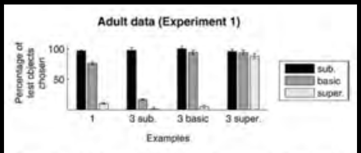
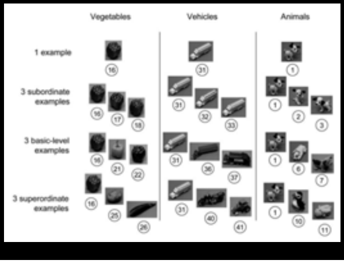


Figure 5. Adults' generalization of word meanings in Experiment 1, averaged over domain. Results are shown for each of four types of example set (one example, three subordinate [sub.] examples, three basic-level examples, and three superordinate [super.] examples). Bar height indicates the frequency with which participants generalized to new objects at various levels. Error bars indicate standard errors.

Xu & Tenenbaum (2007): Expt 2

Subjects: 3- and 4-year-old children

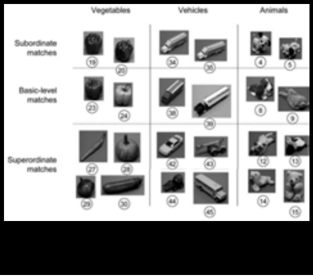
Task, part 1: novel word learning ("This is a blick/fep/dax")



Xu & Tenenbaum (2007): Expt 2

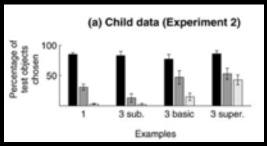
Subjects: 3- and 4-year-old children

Task, part 2: generalization (asked to help Mr.Frog identify only things that are "blicks"/ "feps"/ "daxes" from a set of new objects)



Xu & Tenenbaum (2007): Expt 2

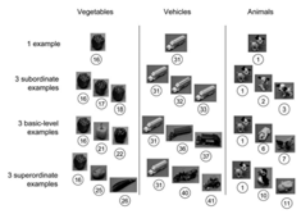
Results: Qualitatively similar behavior to adults, though distinctions are not quite as sharp (possibly due to experimental design factors: dalmatians too interesting, peppers too boring). 1-example learning shows graded judgment, but not much of a basic-level bias.



Xu & Tenenbaum (2007): Expt 3

Subjects: 3- and 4-year-old children

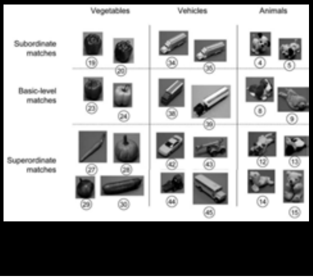
Task, part 1: novel word learning (“This is a blick/fep/dax”) with objects of more equal salience to children (dalmatians replaced by terriers, and peppers replaced by chili peppers). Also, single example labeled 3 times so same number of labeling events occur across conditions.



Xu & Tenenbaum (2007): Expt 3

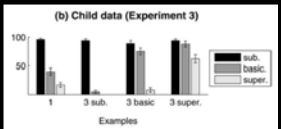
Subjects: 3- and 4-year-old children

Task, part 2: generalization (asked to tell Mr.Frog whether something was a “blick”/ “fep”/ “dax” from a set of new objects)



Xu & Tenenbaum (2007): Expt 3

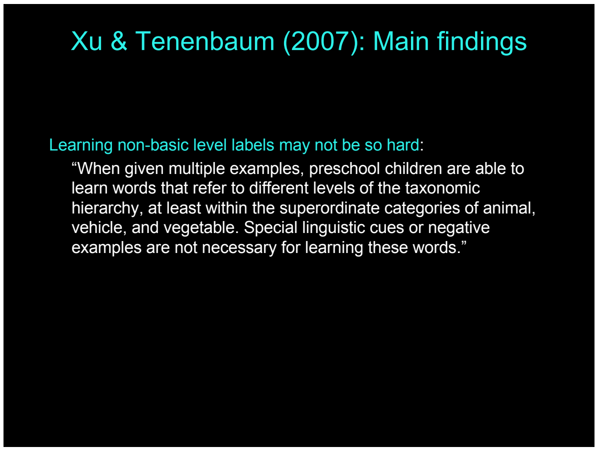
Results: Again, qualitatively similar behavior to adults, though distinctions are not quite as sharp – but still much sharper than before. 1-example learning shows graded judgment, but still not much of a basic-level bias. Surprising tendency not to completely allow superordinate labeling when given superordinate examples.



Xu & Tenenbaum (2007): Main findings

Number of examples matters:

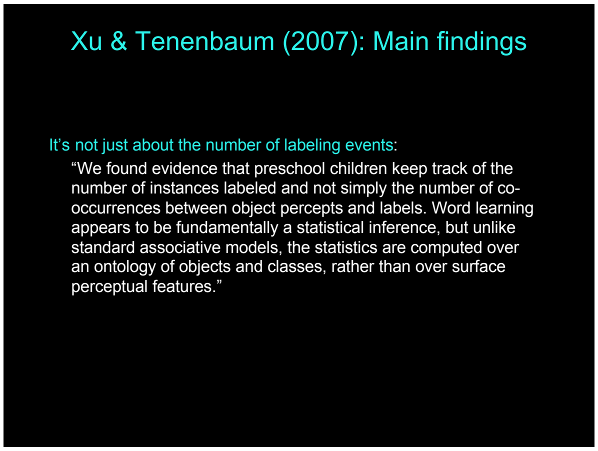
“We found that word learning displays the characteristics of a statistical inference, with both adult and child learners becoming more accurate and more confident in their generalizations as the number of examples increased…Both adult and child learners appear to be sensitive to suspicious coincidences in how the examples given for a novel word appear to cluster in a taxonomy of candidate categories to be named.”



Xu & Tenenbaum (2007): Main findings

Learning non-basic level labels may not be so hard:

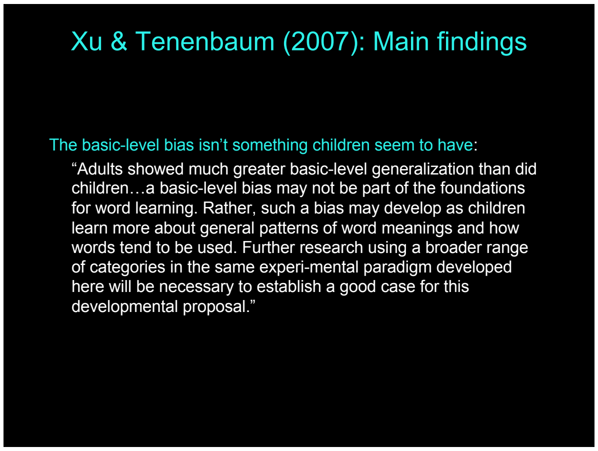
> "When given multiple examples, preschool children are able to learn words that refer to different levels of the taxonomic hierarchy, at least within the superordinate categories of animal, vehicle, and vegetable. Special linguistic cues or negative examples are not necessary for learning these words."



Xu & Tenenbaum (2007): Main findings

It's not just about the number of labeling events:

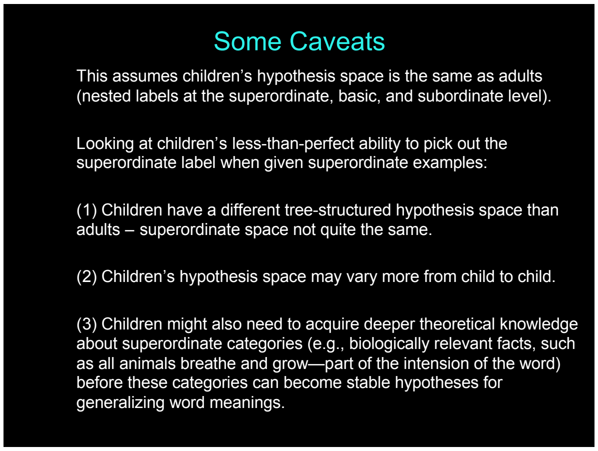
> "We found evidence that preschool children keep track of the number of instances labeled and not simply the number of co-occurrences between object percepts and labels. Word learning appears to be fundamentally a statistical inference, but unlike standard associative models, the statistics are computed over an ontology of objects and classes, rather than over surface perceptual features."



Xu & Tenenbaum (2007): Main findings

The basic-level bias isn't something children seem to have:

> "Adults showed much greater basic-level generalization than did children…a basic-level bias may not be part of the foundations for word learning. Rather, such a bias may develop as children learn more about general patterns of word meanings and how words tend to be used. Further research using a broader range of categories in the same experi-mental paradigm developed here will be necessary to establish a good case for this developmental proposal."



Some Caveats

This assumes children's hypothesis space is the same as adults (nested labels at the superordinate, basic, and subordinate level).

Looking at children's less-than-perfect ability to pick out the superordinate label when given superordinate examples:

(1) Children have a different tree-structured hypothesis space than adults – superordinate space not quite the same.

(2) Children's hypothesis space may vary more from child to child.

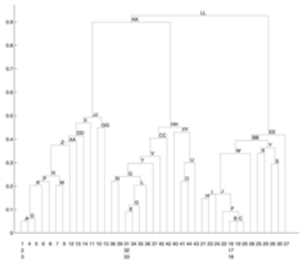
(3) Children might also need to acquire deeper theoretical knowledge about superordinate categories (e.g., biologically relevant facts, such as all animals breathe and grow—part of the intension of the word) before these categories can become stable hypotheses for generalizing word meanings.

Basic-level bias in children’s vocabularies?

“Several factors may be important in explaining the time lag between acquiring basic-level labels and acquiring subordinate- and superordinate-level labels. First, subordinate- and superordinate-level labels may require multiple examples. If each example is labeled on different occasions and spread out in time, children may forget the examples over time. Second, subordinate- and superordinate-level category labels are used much less frequently in adult speech, and so the relevant examples are harder to come by. Middle-class American parents tend to point to objects and label them with basic-level terms. Last, superordinates are often used to refer to collections (Markman, 1989), and so children may be misled by the input in interpreting these words. In our studies, we have presented children with a simplified learning situation in order to uncover the underlying inferential competence that guides them in—but is not exclusively responsible for—real-world performance.”

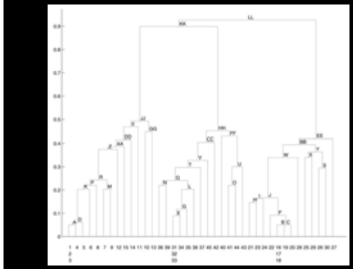
Bayesian model fit to experimental data

Hypothesis space: constructed from adult similarity judgments in experiment 1.



Bayesian model fit to experimental data

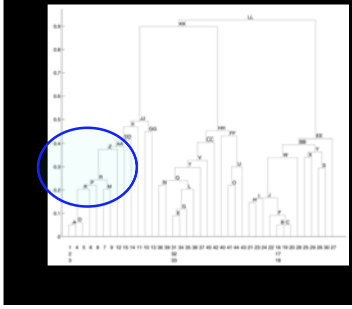
Likelihood: Approximate p(X | h) by height of cluster, which represents object distinctiveness (how dissimilar they are to neighboring cluster)



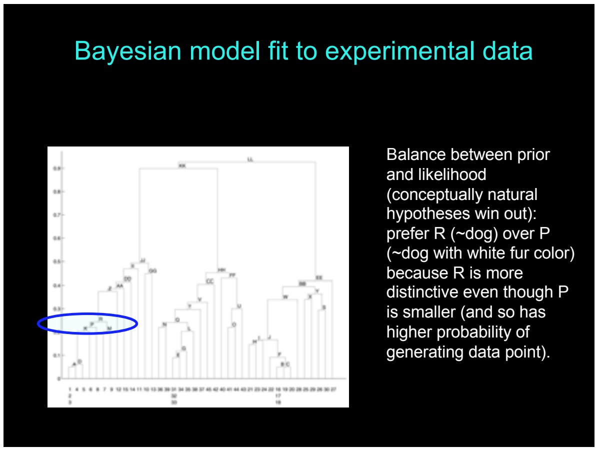
$$p(X \mid h) \propto \left[\frac{1}{\text{height}(h) + \varepsilon}\right]^{n}$$

Bayesian model fit to experimental data

Prior: Approximate p(h) as preference for cluster distinctiveness (prefer R cluster over P [which is not too much lower than R] or Z [which is not too much lower than AA]) – based on cluster branch height

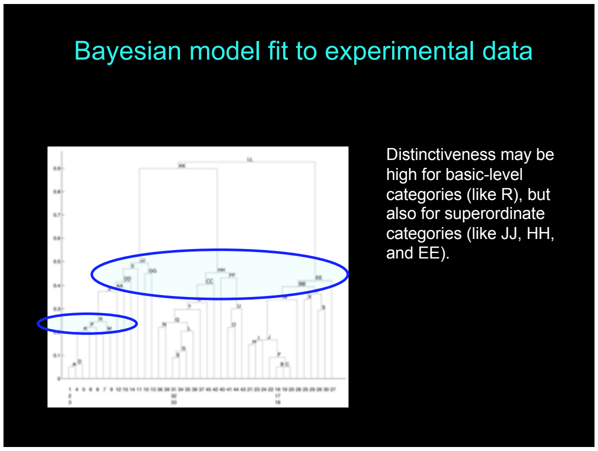


$$p(h) \propto \text{height}(\text{parent}[h]) - \text{height}(h)$$



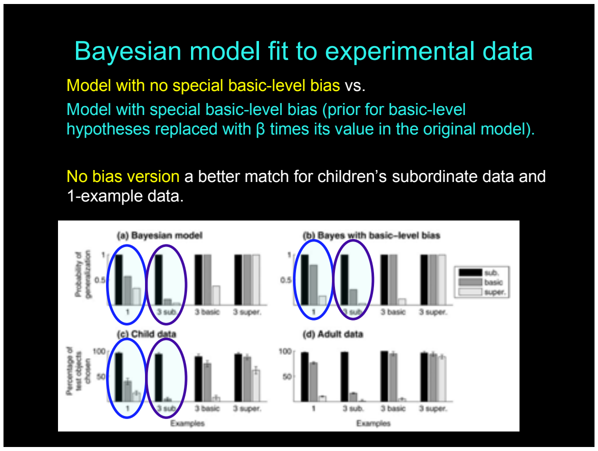
Bayesian model fit to experimental data

Balance between prior and likelihood (conceptually natural hypotheses win out): prefer R (~dog) over P (~dog with white fur color) because R is more distinctive even though P is smaller (and so has higher probability of generating data point).



Bayesian model fit to experimental data

Distinctiveness may be high for basic-level categories (like R), but also for superordinate categories (like JJ, HH, and EE).

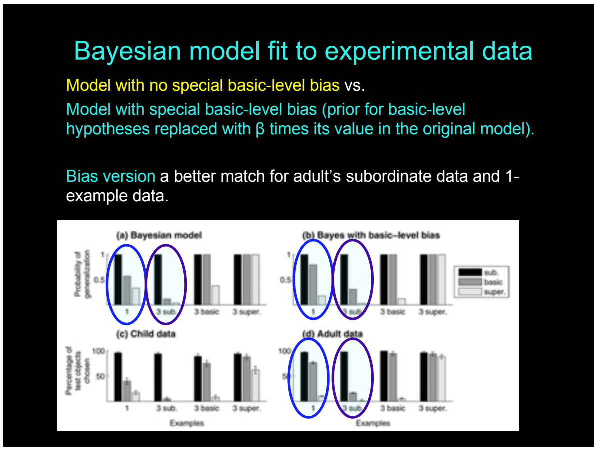


Bayesian model fit to experimental data

Model with no special basic-level bias vs.

Model with special basic-level bias (prior for basic-level hypotheses replaced with β times its value in the original model).

No bias version a better match for children's subordinate data and 1-example data.

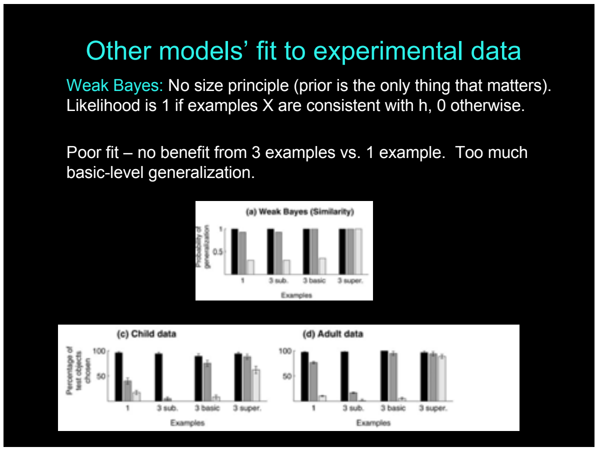


Bayesian model fit to experimental data

Model with no special basic-level bias vs.

Model with special basic-level bias (prior for basic-level hypotheses replaced with β times its value in the original model).

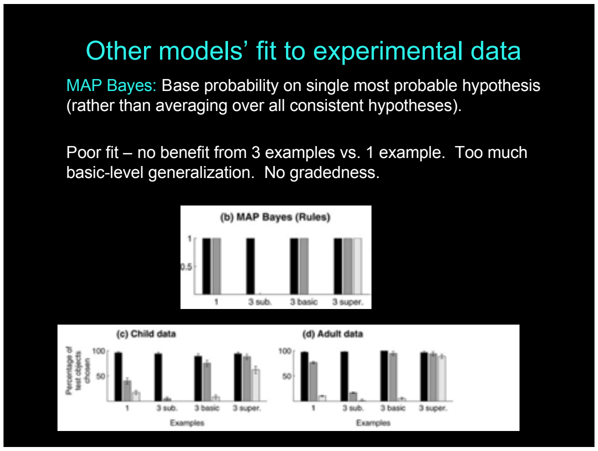
Bias version a better match for adult's subordinate data and 1-example data.



Other models’ fit to experimental data

Weak Bayes: No size principle (prior is the only thing that matters). Likelihood is 1 if examples X are consistent with h, 0 otherwise.

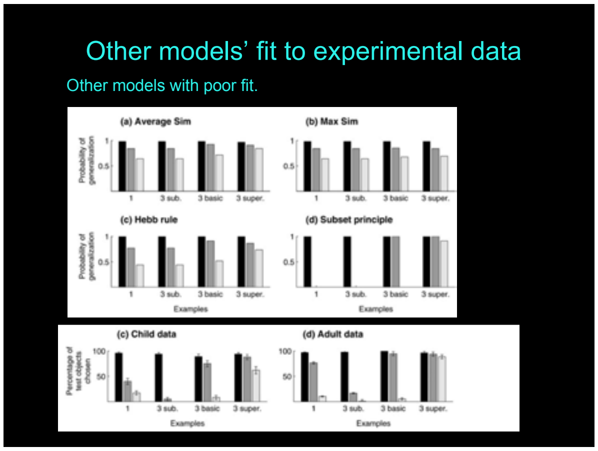
Poor fit – no benefit from 3 examples vs. 1 example. Too much basic-level generalization.



Other models’ fit to experimental data

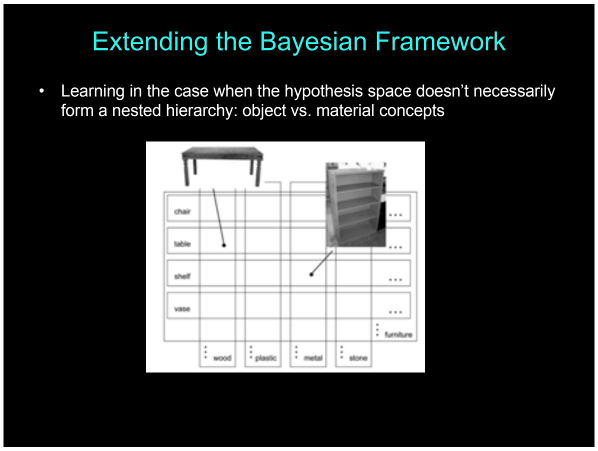
MAP Bayes: Base probability on single most probable hypothesis (rather than averaging over all consistent hypotheses).

Poor fit – no benefit from 3 examples vs. 1 example. Too much basic-level generalization. No gradedness.



Other models’ fit to experimental data

Other models with poor fit.



Extending the Bayesian Framework

- Learning in the case when the hypothesis space doesn’t necessarily form a nested hierarchy: object vs. material concepts

Extending the Bayesian Framework

- Prasada, Ferenz, and Haskell (2002):

  (1) Given a single regularly shaped entity, people tended to choose an object category.

  (2) Given a single irregularly shaped entity, people tended to choose the substance interpretation.

  (3) When people were shown multiple essentially identical entities, each with the same complex irregular shape and novel material, their preference for labeling an entity in this set switched from a substance interpretation to an object interpretation.

Extending the Bayesian Framework

- Bayesian explanation of Prasada, Ferenz, and Haskell (2002):

  (1) Each object category is organized around a prototypical shape.

  (2) Prior for object categories with regular shapes >
  Prior for substance categories >
  Prior for object categories with irregular shapes.

  (3) There are more conceptually distinct shapes that support object categories than there are material properties that support substance categories. Object-kind hypotheses are smaller than substance hypotheses = object-kinds have higher likelihood than substance-kinds. Seeing several examples of novel objects with the same name is a greater suspicious coincidence.

Extending the Bayesian Framework

- Using negative evidence: “That’s a dalmatian. It’s a kind of dog.”

  dalmatian = subset of *dog*

  Bayesian learner can treat this as conclusive evidence that *dalmatian* is a subset of *dog* and give 0 likelihood to any hypothesis where *dalmatian* is not contained within the set of *dogs*.

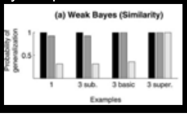
Extending the Bayesian Framework

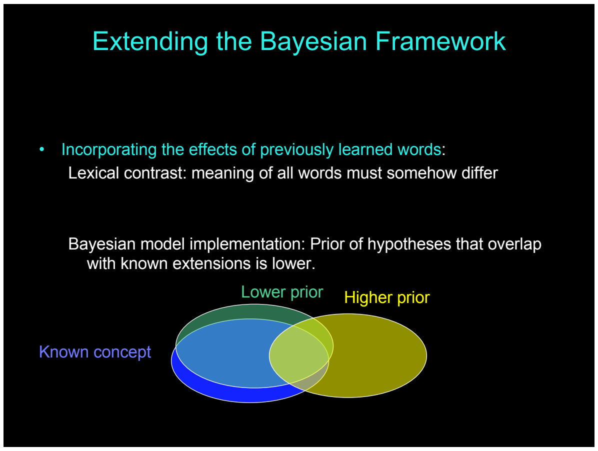
- Being sensitive to the source of the input:

  Random sampling (“teacher-driven”) vs. Non-random sampling (“learner-driven”) of subordinates

  Xu & Tenenbaum (2007 Developmental Science): Adults and pre-schoolers in teacher-driven condition made subordinate generalization while the ones in the learner-driven condition allowed basic-level generalization.

  Bayesian model explanation: Likelihood no longer reflects size principle (not a suspicious coincidence if you had control of which examples were picked), just similarity – equivalent to Weak Bayes model





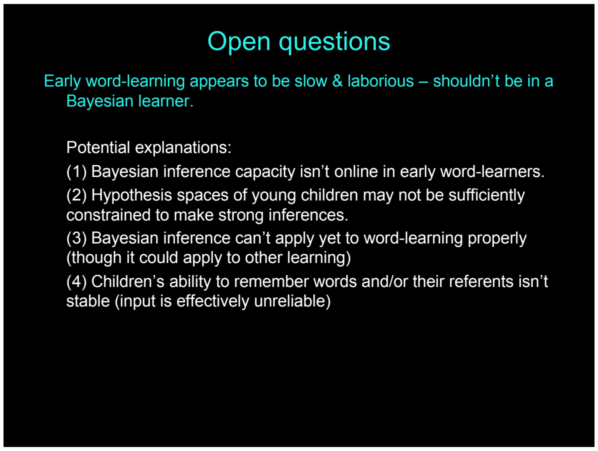
Extending the Bayesian Framework

- Incorporating the effects of previously learned words:

  Lexical contrast: meaning of all words must somehow differ

  Bayesian model implementation: Prior of hypotheses that overlap with known extensions is lower.

Lower prior
Higher prior
Known concept



Open questions

Early word-learning appears to be slow & laborious – shouldn't be in a Bayesian learner.

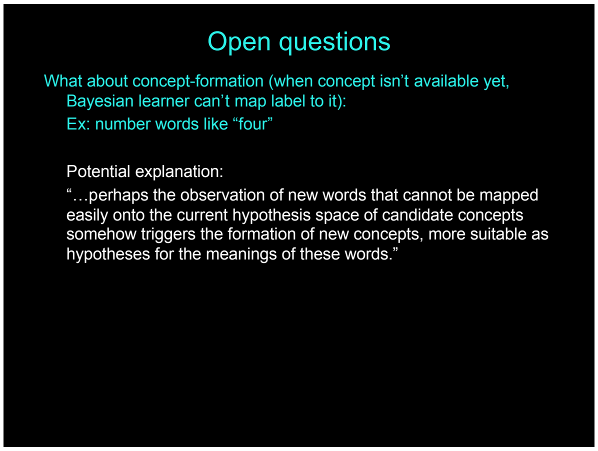
Potential explanations:

(1) Bayesian inference capacity isn't online in early word-learners.

(2) Hypothesis spaces of young children may not be sufficiently constrained to make strong inferences.

(3) Bayesian inference can't apply yet to word-learning properly (though it could apply to other learning)

(4) Children's ability to remember words and/or their referents isn't stable (input is effectively unreliable)



Open questions

What about concept-formation (when concept isn't available yet, Bayesian learner can't map label to it):

Ex: number words like "four"

Potential explanation:

"…perhaps the observation of new words that cannot be mapped easily onto the current hypothesis space of candidate concepts somehow triggers the formation of new concepts, more suitable as hypotheses for the meanings of these words."