

Word Learning as Bayesian Inference

Fei Xu and Joshua B. Tenenbaum

Presented by Prutha Deshpande



Approaches to Word Learning

1. Hypothesis Elimination (deductive approach)

- Consider a hypothesis space of concept-word mappings.
- Eliminate incorrect hypotheses based on prior knowledge and observation.

2. Associative Learning

- Connectionist networks or similarity matching to examples.
- Abstract generalizations of word meanings.

- Also, word learning through observation of how words tend to be used together (looking for word clusters).

Core Phenomena of Word Learning

1. Word meanings can be learned from very few examples.
2. Word meanings can be inferred from only positive examples of what the word refers to.
3. Word meanings are complex and consist of overlapping concepts.
4. Inferences about word meanings from examples may be graded.
5. Inferences about word meanings can be affected by pragmatic reasoning given the relevant communicative context.

Evaluating Traditional Approaches

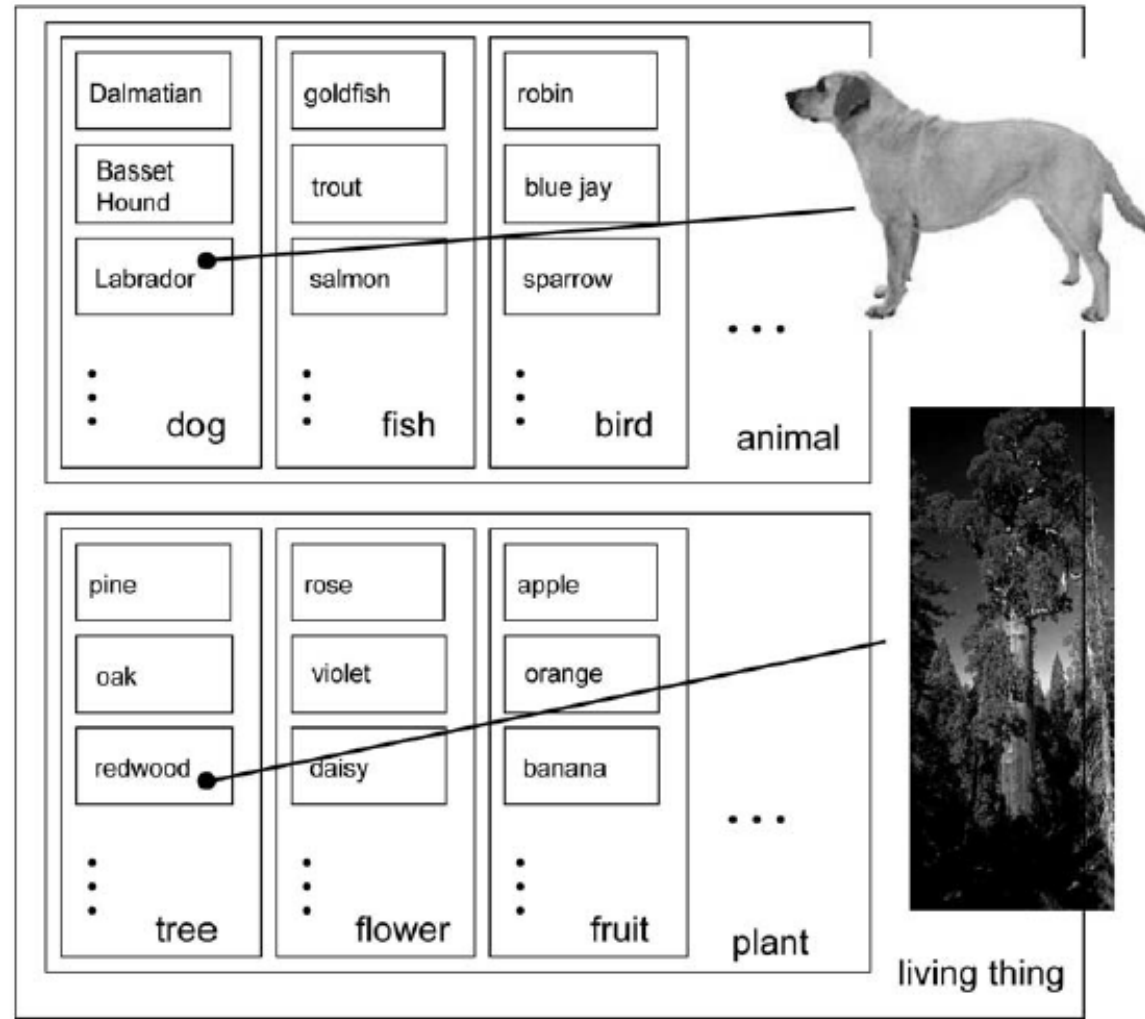
Inference about word meanings from sparse data: few and only positive examples (phenomena 1 & 2)?

Hypothesis Elimination Paradigm

Strong prior knowledge about the kinds of viable word meanings:

- Whole object constraint
- Taxonomic constraint

→ But these constraints are not sufficient to solve the problem of overlapping concepts (phenomenon 3).



Tree-structured hierarchy of taxonomical classes →
 Problem of overlapping concepts

Evaluating Traditional Approaches

Inference about word meanings from sparse data: few and only positive examples (phenomena 1 & 2)?

Associative Learning

- Assumption of implicit negative evidence
- Tuning of attentional biases

→ But these also do not solve the problem of overlapping concepts (phenomenon 3).

Suspicious Coincidence



- Bayesian approach captures the intuition of a suspicious coincidence.
 - Scores alternative hypotheses about a word's meaning according to how well they predict the observed data, as well as how they fit with the learner's prior expectations about natural meanings.
- Hypothesis elimination and associative learning cannot explain learning from suspicious coincidences.

Graded Word Learning (Phenomenon 4)

- Inference becomes more confident with multiple examples.
- The shift in confidence suggests that the initial inference does not completely rule out a hypothesis.
- This is a problem for the hypothesis elimination approach.
- The Bayesian approach can however evaluate probabilities of alternate hypotheses.



Model Framework

X = set of n observed examples of novel word C

X is drawn from domain of entities U

Each hypothesis h is a pointer to subset of entities in U that is a candidate extension for C

$$p(h | X) = \frac{p(X | h)p(h)}{p(X)} \quad (1)$$

$$= \frac{p(X | h)p(h)}{\sum_{h' \in H} p(X | h')p(h')} \quad (2)$$

Model Framework

- Generalization behavior: decision whether any given new object y belongs to the extension of C , given the observations X .

$$p(y \in C | X) = \sum_{h \in H} p(y \in C | h) p(h | X). \quad (3)$$

$$p(y \in C | X) = \sum_{h \supset y, X} p(h | X), \quad (4)$$

Probabilistic Components of the Model

Likelihood

$$p(X | h) = \left[\frac{1}{\text{size}(h)} \right]^n, \quad (5)$$

Size principle - Hypotheses with smaller extensions assign greater probability than do larger hypotheses to the same data, and they assign exponentially greater probability as the number of consistent examples increases.

→ Suspicious coincidence phenomenon

Probabilistic Components of the Model

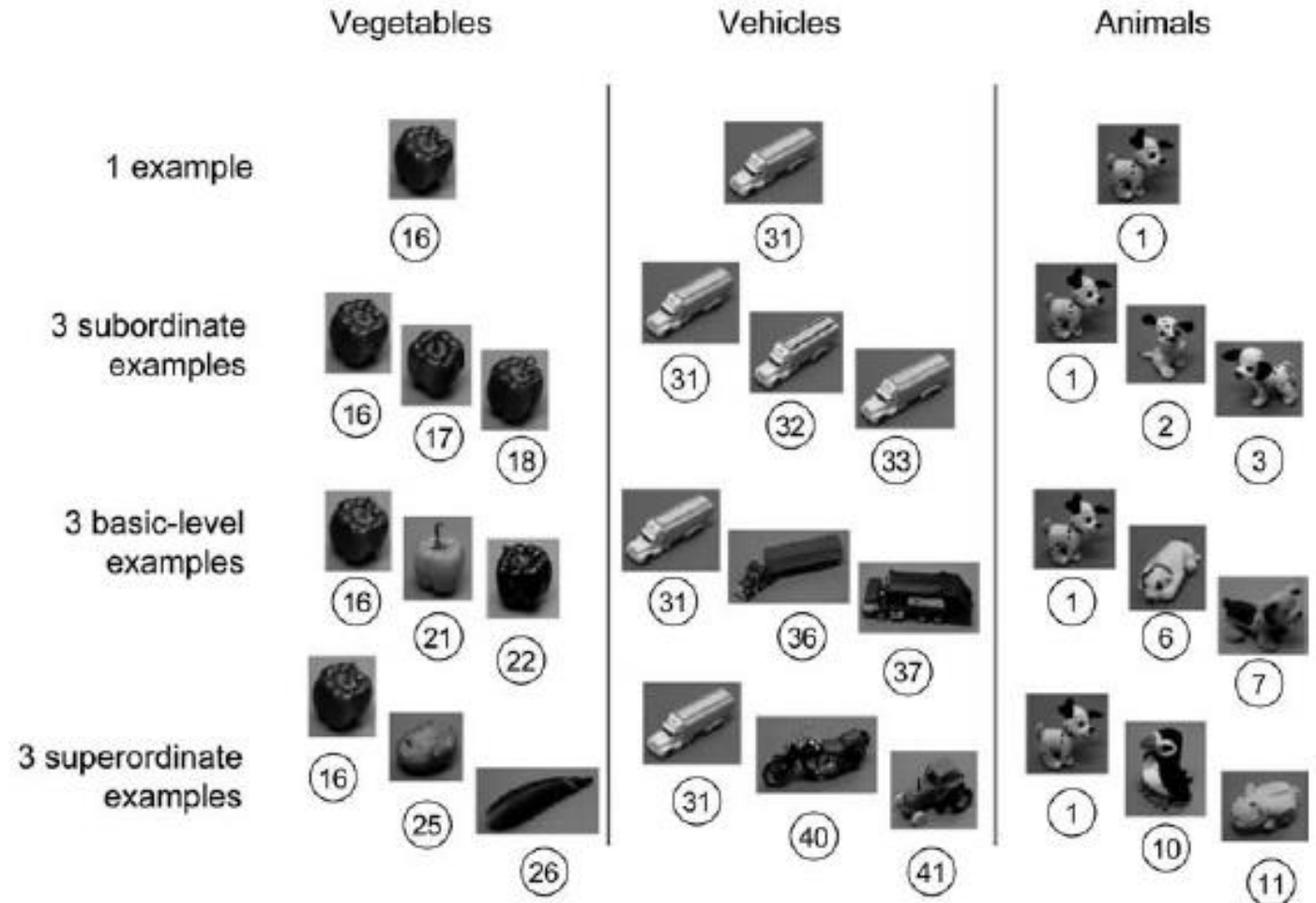
Prior

- Assumption that hypotheses correspond to nodes in a tree-structured taxonomy.
- This assigns zero prior probability to all other subsets of objects in the world that do not conform to the particular taxonomy.
- A constrained prior is required to make meaningful generalizations.
- Preference for distinctiveness in prior probabilities.


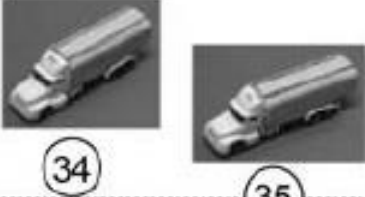
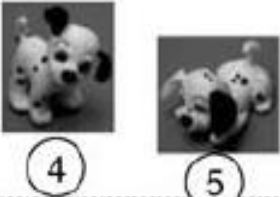
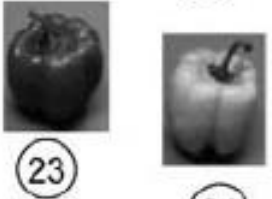

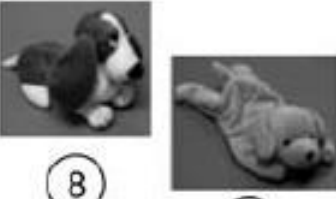
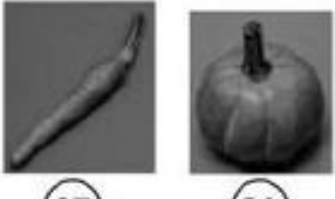
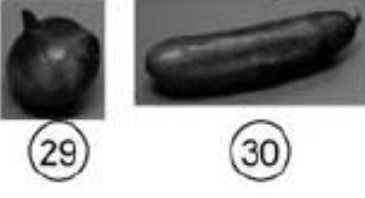

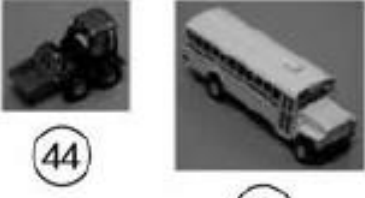
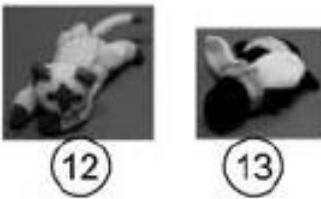
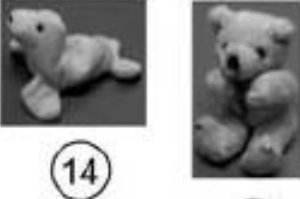
Experiment 1

- Adult word-learning
 1. Word-learning phase
 2. Similarity judgment phase
- Manipulated variables:
 1. Number of examples
 2. Range of examples

Training stimuli →



Test Stimuli →

	Vegetables	Vehicles	Animals
Subordinate matches			
Basic-level matches			
Superordinate matches	 	 	 

Adult data (Experiment 1)

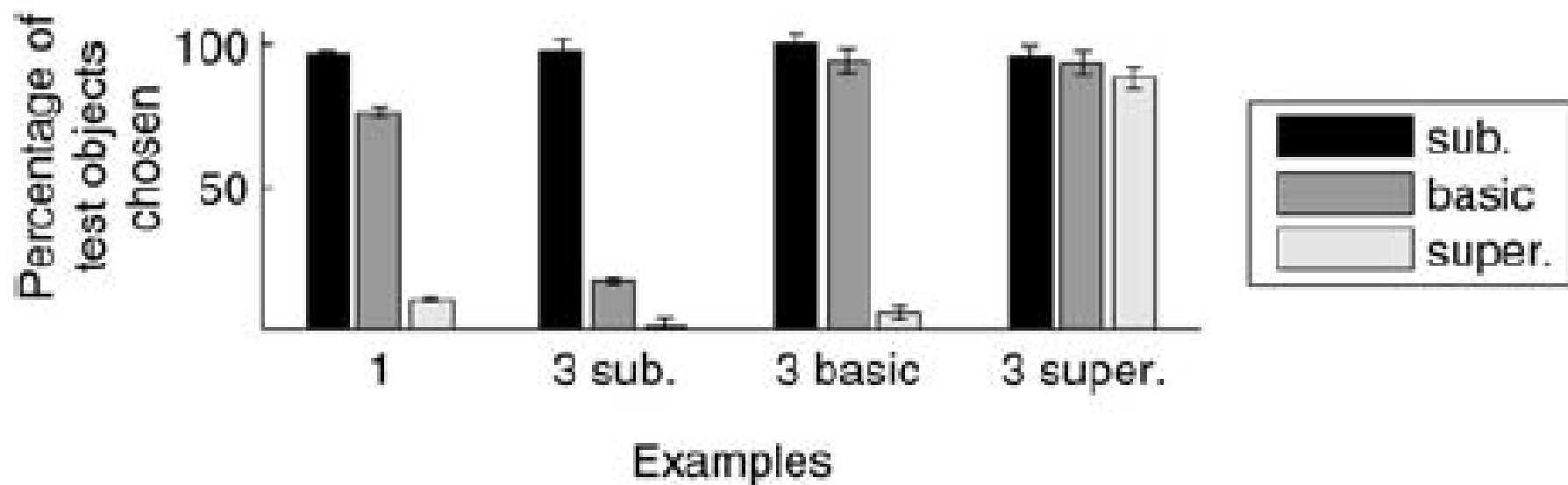
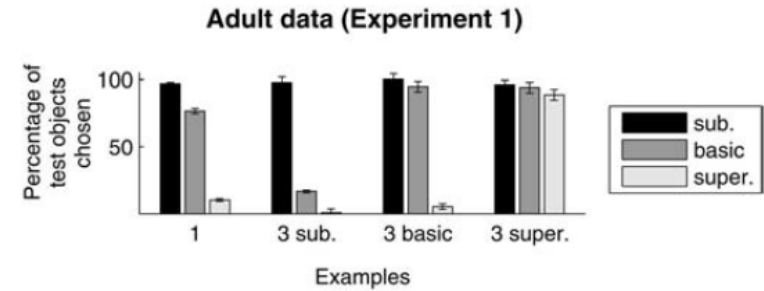


Figure 5. Adults' generalization of word meanings in Experiment 1, averaged over domain. Results are shown for each of four types of example set (one example, three subordinate [sub.] examples, three basic-level examples, and three superordinate [super.] examples). Bar height indicates the frequency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

Questions addressed with *t* tests



1. Did participants generalize further in the one-example trials compared with the three-example subordinate trials?
 - Was there a threshold in generalization at the basic level in the one-example trials and did they restrict their generalization to the subordinate level in the three-example trials?
2. Did the three-example trials differ from each other depending on the range spanned by the examples?
 - Did participants restrict their generalization to the most specific level that was consistent with the set of exemplars?

Conclusions

1. With one example, adults showed graded generalization from subordinate to basic-level to superordinate matches.
2. Adults showed a basic-level preference: They generalized to all of the other exemplars from the same basic-level category but generalized much less to the superordinate category.
3. With three examples, adults made generalizations in more of an all-or-none manner, restricting their generalizations to the most specific level that was consistent with the examples.

Experiment 2

- Word learning in children (3 – 4 year olds)
- Manipulated variables:
 1. Number of examples
 2. Range of examples
- Each child participated in a total of three trials, one from each of the three superordinate categories.



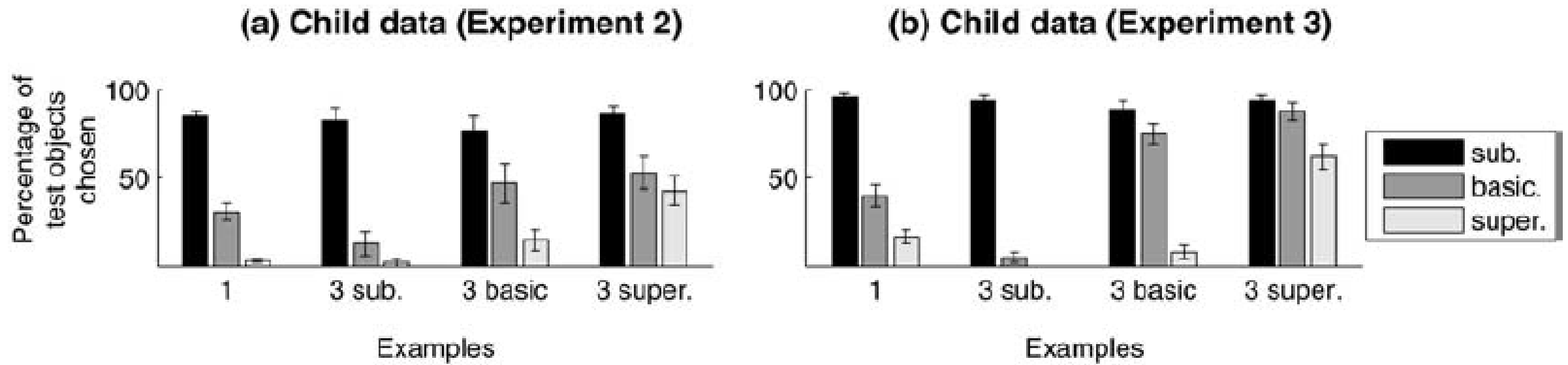


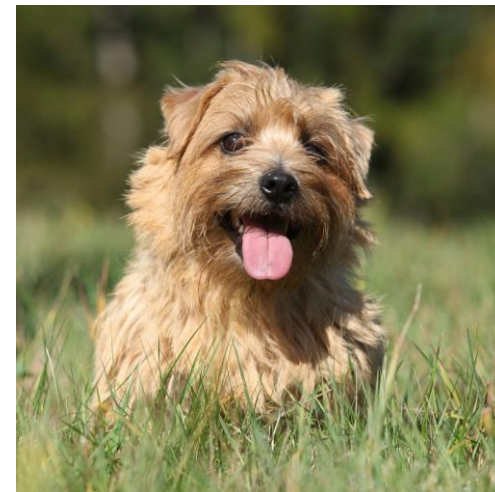
Figure 6. Children’s generalization of word meanings in Experiments 2 and 3, averaged over domain. Results are shown for each of four types of example set (one example, three subordinate [sub.] examples, three basic-level examples, and three superordinate [super.] examples). Bar height indicates the frequency with which participants generalized to new objects at various levels. Error bars indicate standard errors.

Conclusions

1. With one example, children showed graded generalization from subordinate to basic-level to superordinate matches.
2. Children did not show a basic-level preference.
3. With three examples, children modified their generalizations depending on the span of the examples, and were consistent with the most specific category that included all of the examples.

Experiment 3

- Word learning in children (3 – 4 year olds).
- Equated the number of labeling events in the one-example and three-example conditions.
- Experimenter chose 10 of the 24 target objects and asked for the child's judgment in each case.



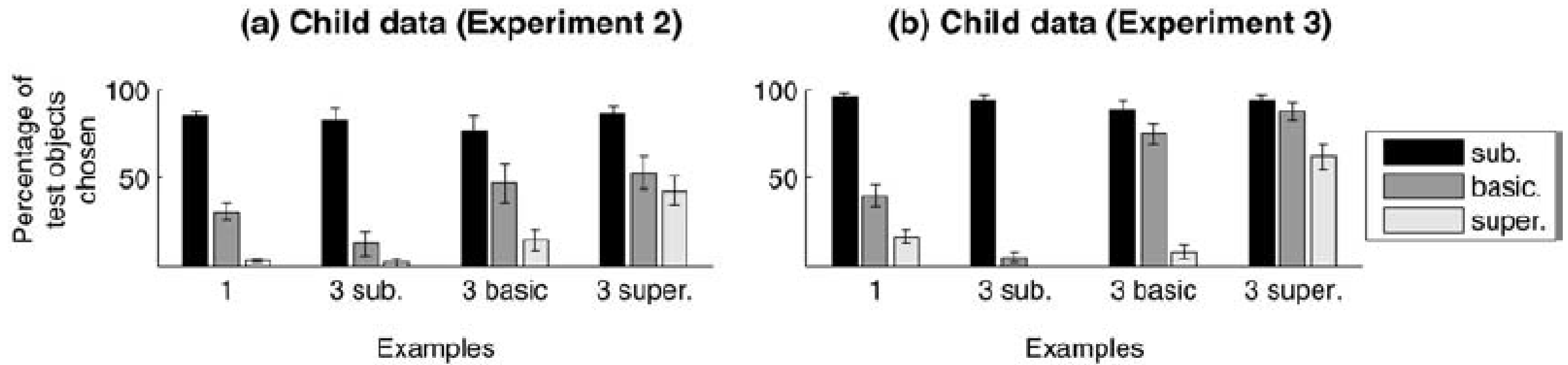


Figure 6. Children’s generalization of word meanings in Experiments 2 and 3, averaged over domain. Results are shown for each of four types of example set (one example, three subordinate [sub.] examples, three basic-level examples, and three superordinate [super.] examples). Bar height indicates the frequency with which participants generalized to new objects at various levels. Error bars indicate standard errors.

Conclusions

1. Children still showed a much lower tendency for basic-level generalization given a single example.
2. The differences in generalization between the one-example and three-example conditions of Experiment 2 persisted, even though the number of labeling events was equated across conditions.
 - This suggests that children make statistical inferences about word meanings that are computed over the number of examples labeled, not just the number of word–object pairings (associative approach).

Overall Conclusions

1. Word learning displays the characteristics of statistical inference.
 - Both adults and children sensitive to suspicious coincidences
 - Children keep track of the number of instances labeled and not simply the number of co-occurrences between object and labels
2. Adults showed much greater basic-level generalization than did children.
3. When given multiple examples, preschool children are able to learn words that refer to different levels of the taxonomic hierarchy.

Bayesian Model for Object-Kind Labels

Hypothesis Space

- Internal node – cluster of similar objects
- Height of node - pairwise dissimilarity of objects in the cluster
- Length of branch above node – how distinctive a cluster is

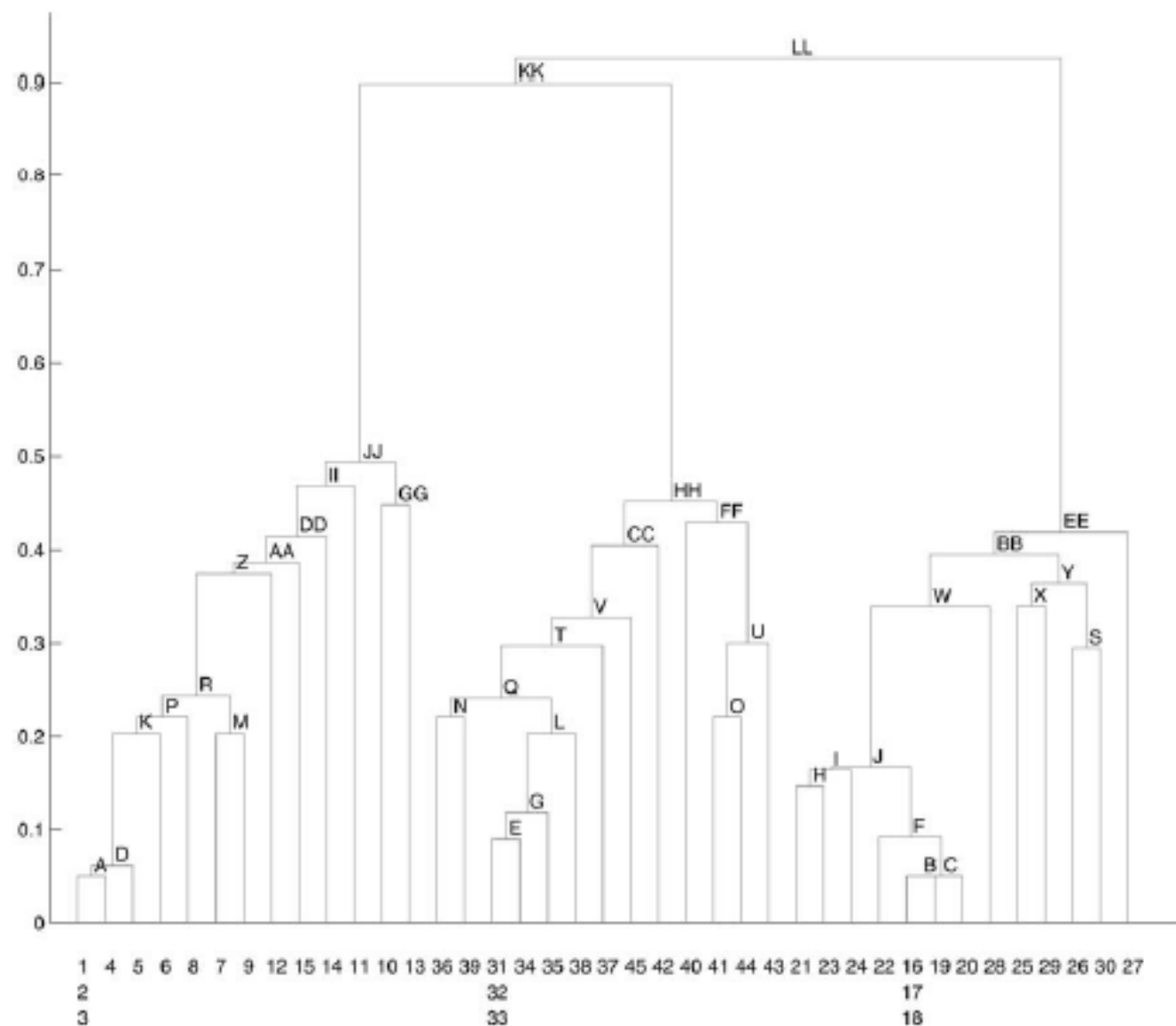


Figure 7. Hierarchical clustering of similarity judgments yields a taxonomic hypothesis space for Bayesian word learning. Letter codes refer to specific clusters (hypotheses for word meaning): vegetable (EE), vehicle (HH), animal (JJ), pepper (J), truck (T), dog (R), green pepper (F), yellow truck (G), and Dalmatian (D). The clusters labeled by other letter codes are given in the text as needed. Numbers indicate the objects located at each leaf node of the hierarchy, keyed to the object numbers shown in Figures 3 and 4. The height of a cluster, as given by the vertical axis on the left, represents the average within-cluster dissimilarity of objects within that cluster.

Numerical Values for Likelihoods and Priors

Likelihood

- A function of the size of the extension of a hypothesis

$$p(X | h) \propto \left[\frac{1}{\text{height}(h) + \epsilon} \right]^n, \quad (6)$$

Prior

- Preference for cluster distinctiveness

$$p(h) \propto \text{height}(\text{parent}[h]) - \text{height}(h). \quad (7)$$

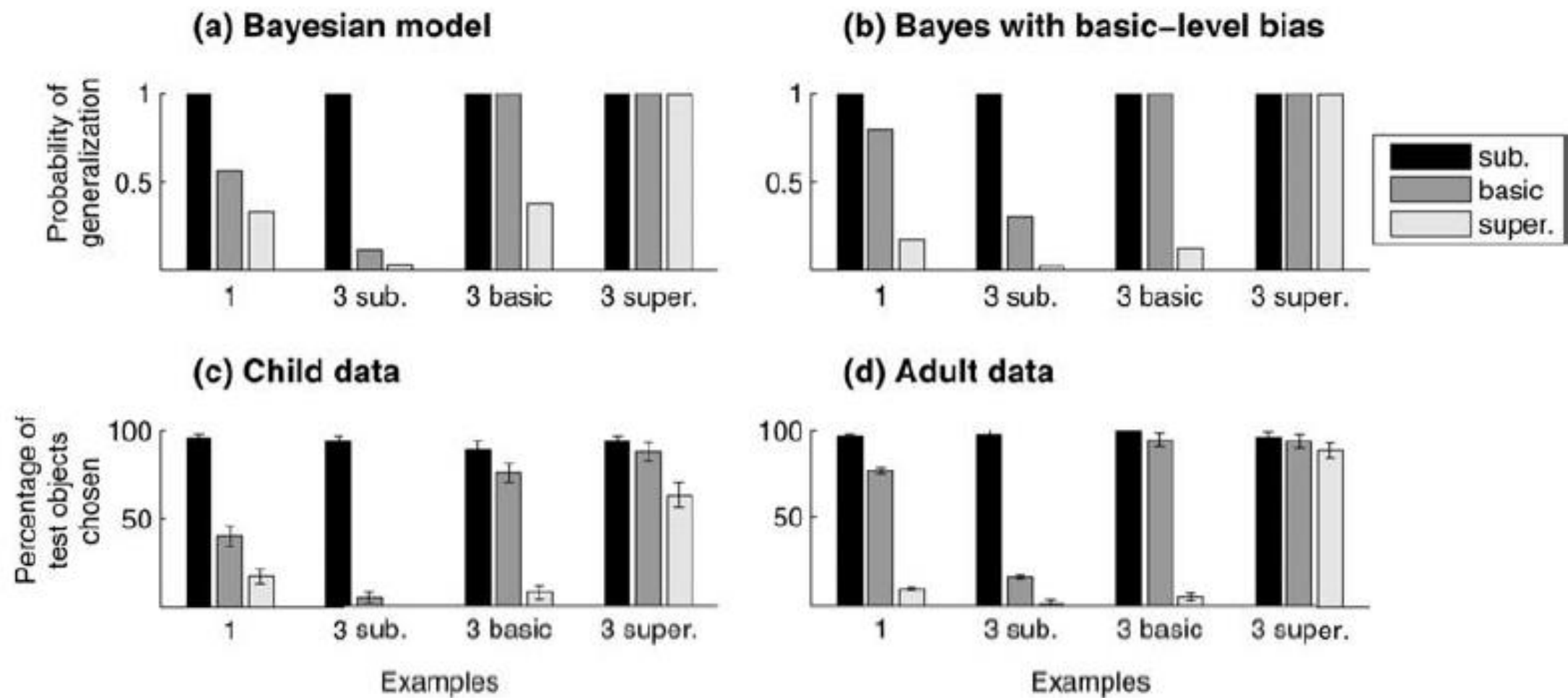


Figure 8. Predictions of the Bayesian model, both with and without a basic-level bias, compared with the data from adults in Experiment 1 and those from children in Experiment 3. Sub. = subordinate; super. = superordinate.

Comparisons with other Models

Weak Bayes

- Weighing all hypotheses strictly by their prior, instead of by the size principle.
- Reduces Bayes to a similarity-like feature matching computation.

Maximum a Posteriori Bayes (MAP)

- Basing generalization on just the single most probable hypothesis
- Reduces Bayes to an all-or-none rule like computation.

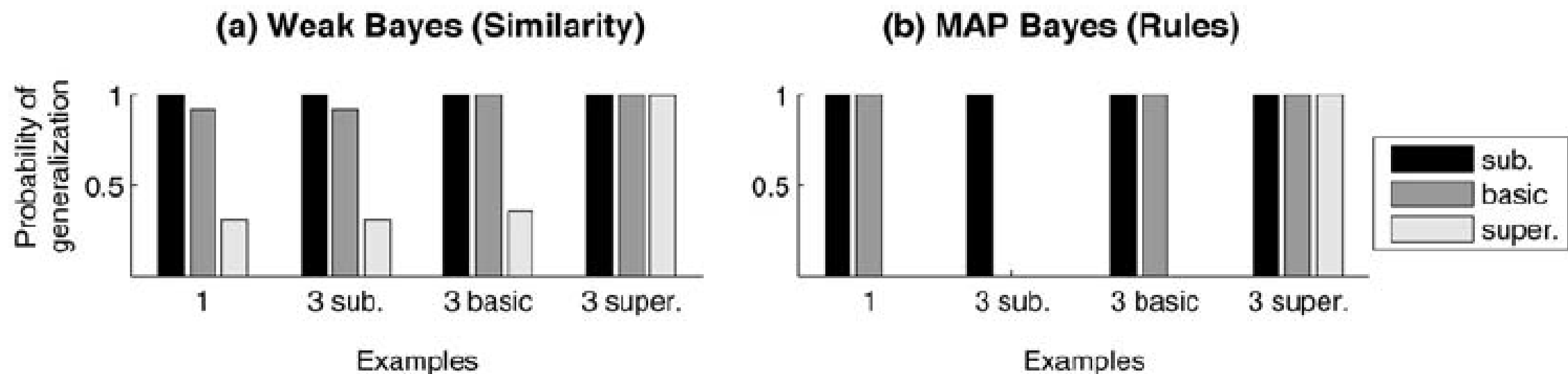


Figure 9. Predictions of two variants of the Bayesian model. (a) Without the size principle, Bayesian generalization behaves like an exemplar-similarity computation. (b) Without hypothesis averaging, Bayesian generalization follows an all-or-none, rulelike pattern. MAP Bayes = maximum a posteriori Bayes approach; sub. = subordinate; super. = superordinate.

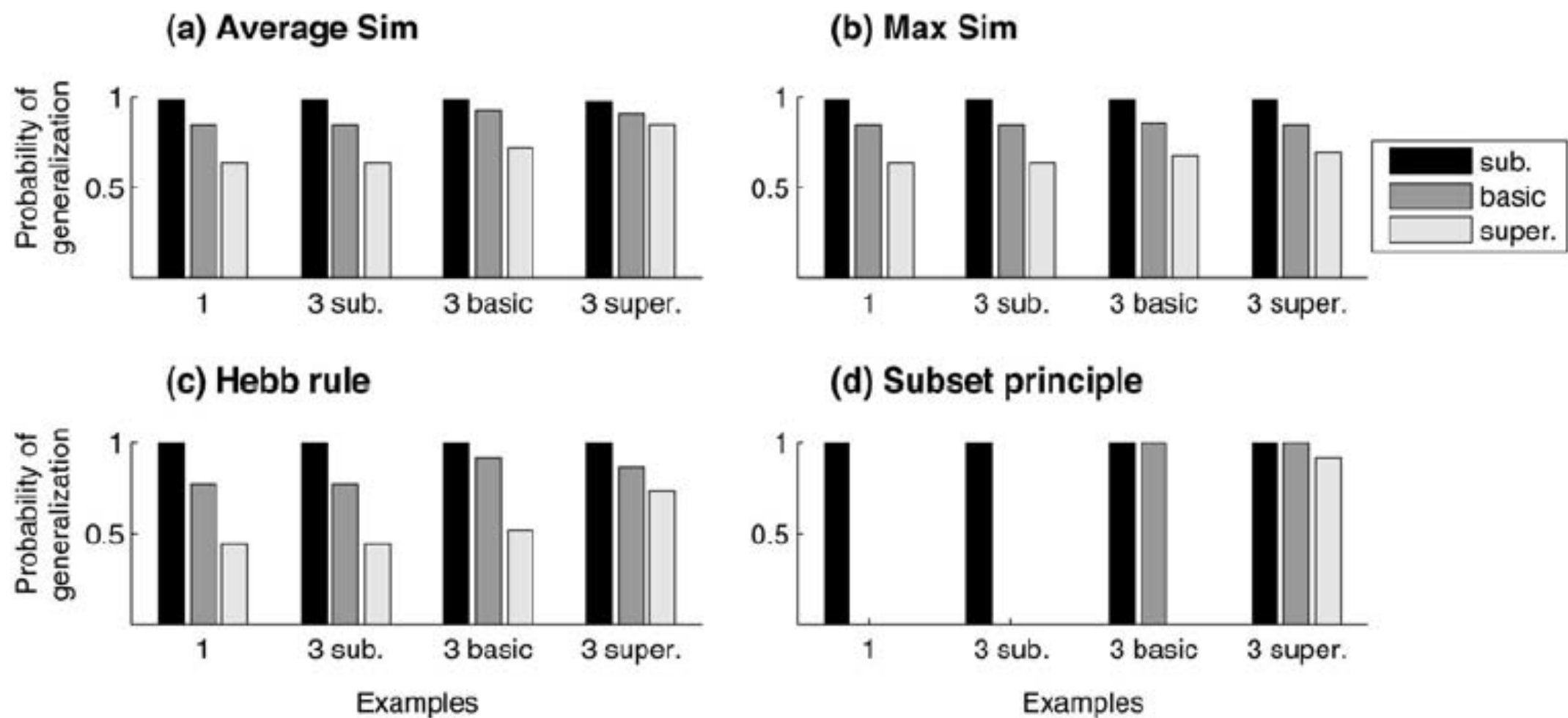


Figure 10. Predictions of four alternative, non-Bayesian models. Max = maximum; Sim = similarity; sub. = subordinate; super. = superordinate.

Extending the Bayesian Framework

1. Hypothesis space: objects and solid substances
2. Transforming the likelihood function
 - Other sources of input – negative examples and special linguistic cues
 - Theory-of-mind reasoning and sensitivity to sampling
3. Transforming prior probabilities
 - The effects of previously learned words